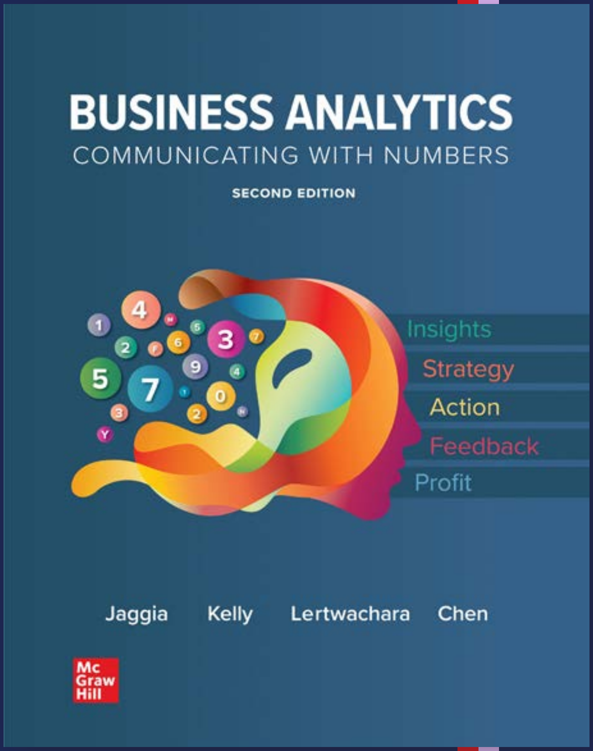
McGraw Hill connect®

# Business Analytics, Communicating with Numbers, 2e

Jaggia | Kelly | Lertwachara | Chen



## Table of Contents

# The McGraw Hill Series in Operations and Decision Sciences

## SUPPLY CHAIN MANAGEMENT

Bowersox, Closs, Cooper, and Bowersox
**Supply Chain Logistics Management**
*Fifth Edition*

Johnson
**Purchasing and Supply Management**
*Sixteenth Edition*

Simchi-Levi, Kaminsky, and Simchi-Levi
**Designing and Managing the Supply Chain: Concepts, Strategies, Case Studies**
*Fourth Edition*

Stock and Manrodt
**Fundamentals of Supply Chain Management**

## PROJECT MANAGEMENT

Larson and Gray
**Project Management: The Managerial Process**
*Eighth Edition*

## SERVICE OPERATIONS MANAGEMENT

Bordoloi, Fitzsimmons, and Fitzsimmons
**Service Management: Operations, Strategy, Information Technology**
*Tenth Edition*

## MANAGEMENT SCIENCE

Hillier and Hillier
**Introduction to Management Science: A Modeling and Case Studies Approach with Spreadsheets**
*Sixth Edition*

## BUSINESS RESEARCH METHODS

Schindler
**Business Research Methods**
*Fourteenth Edition*

## BUSINESS FORECASTING

Keating and Wilson
**Forecasting and Predictive Analytics**
*Seventh Edition*

## BUSINESS SYSTEMS DYNAMICS

Sterman
**Business Dynamics: Systems Thinking and Modeling for a Complex World**

## OPERATIONS MANAGEMENT

Cachon and Terwiesch
**Operations Management**
*Third Edition*

Cachon and Terwiesch
**Matching Supply with Demand: An Introduction to Operations Management**
*Fourth Edition*

Jacobs and Chase
**Operations and Supply Chain Management**
*Sixteenth Edition*

Jacobs and Chase
**Operations and Supply Chain Management: The Core**
*Sixth Edition*

Schroeder and Goldstein
**Operations Management in the Supply Chain: Decisions and Cases**
*Eighth Edition*

Stevenson
**Operations Management**
*Fourteenth Edition*

Swink, Melnyk, and Hartley
**Managing Operations Across the Supply Chain**
*Fourth Edition*

## BUSINESS STATISTICS

Bowerman, Drougas, Duckworth, Froelich, Hummel, Moninger, and Schur
**Business Statistics and Analytics in Practice**
*Ninth Edition*

Doane and Seward
**Applied Statistics in Business and Economics**
*Seventh Edition*

Doane and Seward
**Essential Statistics in Business and Economics**
*Third Edition*

Lind, Marchal, and Wathen
**Basic Statistics for Business and Economics**
*Tenth Edition*

Lind, Marchal, and Wathen
**Statistical Techniques in Business and Economics**
*Eighteenth Edition*

Jaggia and Kelly
**Business Statistics: Communicating with Numbers**
*Fourth Edition*

Jaggia and Kelly
**Essentials of Business Statistics: Communicating with Numbers**
*Second Edition*

## BUSINESS ANALYTICS

Jaggia, Kelly, Lertwachara, and Chen
**Business Analytics: Communicating with Numbers**
*Second Edition*

## BUSINESS MATH

Slater and Wittry
**Practical Business Math Procedures**
*Fourteenth Edition*

Slater and Wittry
**Math for Business and Finance: An Algebraic Approach**
*Second Edition*

# BUSINESS ANALYTICS

## Communicating with Numbers

Sanjiv Jaggia
*California Polytechnic State University*

Kevin Lertwachara
*California Polytechnic State University*

Alison Kelly
*Suffolk University*

Leida Chen
*California Polytechnic State University*

Mc
Graw
Hill

BUSINESS ANALYTICS, SECOND EDITION

Some ancillaries, including electronic and print components, may not be available to customers outside the United States.

This book is printed on acid-free paper.

1 2 3 4 5 6 7 8 9 LWI 27 26 25 24 23 22

ISBN 978-1-264-30280-2 (bound edition)
MHID 1-264-30280-0 (bound edition)
ISBN 978-1-265-89710-9 (loose-leaf edition)
MHID 1-265-89710-7 (loose-leaf edition)

Portfolio Manager: *Charles Synovec*
Product Developer: *Ryan McAndrews*
Executive Marketing Manager: *Harper Christopher*
Content Project Managers: *Pat Frederickson/Jamie Koch*
Buyer: *Sandy Ludovissy*
Designer: *Beth Blech*
Content Licensing Specialist: *Lorraine Buczek*
Cover Image: *Nietz_studio/Shutterstock*
Compositor: *Straive*

All credits appearing on page or at the end of the book are considered to be an extension of the copyright page.

**Library of Congress Cataloging-in-Publication Data**

mheducation.com/highered

*Dedicated to our families*

# ABOUT THE AUTHORS

## Sanjiv Jaggia

Courtesy Sanjiv Jaggia

Sanjiv Jaggia is a professor of economics and finance at California Polytechnic State University in San Luis Obispo. Dr. Jaggia holds a Ph.D. from Indiana University and is a Chartered Financial Analyst (CFA®). He enjoys research in statistics and econometrics applied to a wide range of business disciplines. Dr. Jaggia has published several papers in leading academic journals and has co-authored three successful textbooks in business statistics and business analytics. His ability to communicate in the classroom has been acknowledged by several teaching awards. Dr. Jaggia resides in San Luis Obispo with his wife and daughter. In his spare time, he enjoys cooking, hiking, and listening to a wide range of music.

## Alison Kelly

Courtesy Alison Kelly

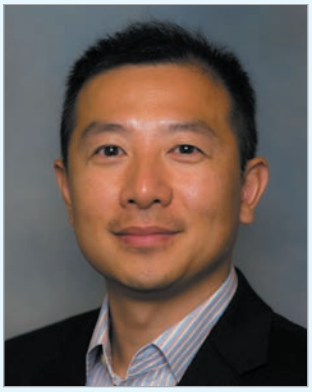
Alison Kelly is a professor of economics at Suffolk University in Boston. Dr. Kelly holds a Ph.D. from Boston College and is a Chartered Financial Analyst (CFA®). Dr. Kelly has published in a wide variety of academic journals and has co-authored three successful textbooks in business statistics and business analytics. Her courses in applied statistics and econometrics are well received by students as well as working professionals. Dr. Kelly resides in Hamilton, Massachusetts, with her husband, daughter, and son. In her spare time, she enjoys exercising and gardening.

## Kevin Lertwachara

Teresa Cameron/Frank Gonzales/ California Polytechnic State University

Kevin Lertwachara is a professor of information systems at California Polytechnic State University in San Luis Obispo. Dr. Lertwachara holds a Ph.D. in Operations and Information Management from the University of Connecticut. Dr. Lertwachara's research focuses on technology-based innovation, electronic commerce, health care informatics, and business analytics and his work has been published in scholarly books and leading academic journals. He teaches business analytics at both the undergraduate and graduate levels and has received several teaching awards. Dr. Lertwachara resides in the central coast of California with his wife and three sons. In his spare time, he coaches his sons' soccer and futsal teams.

## Leida Chen



Courtesy of l eida Chen

Leida Chen is a professor of information systems at California Polytechnic State University in San Luis Obispo. Dr. Chen earned a Ph.D. in Management Information Systems from the University of Memphis. His research and consulting interests are in the areas of business analytics, technology diffusion, and global information systems. Dr. Chen has published over 50 research articles in leading information systems journals, over 30 articles and book chapters in national and international conference proceedings and edited books, and a book on mobile application development. He teaches business analytics at both the undergraduate and graduate levels. In his spare time, Dr. Chen enjoys hiking, painting, and traveling with his wife and son to interesting places around the world.

# Making Data Analytics Relevant to Students and Businesses

Data and analytics capabilities have made a leap forward in recent years and have changed the way businesses make decisions. The explosion in the field is partly due to the growing availability of vast amounts of data, improved computational power, and the development of sophisticated algorithms. More than ever, colleges and universities need a curriculum that emphasizes business analytics, and companies need data-savvy professionals who can turn data into insights and action.

We wrote *Business Analytics: Communicating with Numbers* from the ground up to prepare students to understand, manage, and visualize the data; apply the appropriate analysis tools; and communicate the findings and their relevance. The text seamlessly threads the topics of data wrangling, descriptive analytics, predictive analytics, and prescriptive analytics into a cohesive whole.

Experiential learning opportunities have been proven effective in teaching applied and complex subjects such as business analytics. In this text, we provide a holistic analytics process, including dealing with real-life data that are not necessarily "clean" and/or "small." Similarly, we stress the importance of effective storytelling and help students develop skills in articulating the business value of analytics by communicating insights gained from a nontechnical standpoint.

## Continuing Key Features

The second edition of *Business Analytics* reinforces and expands six core features that were well-received in the first edition.

**Holistic Approach to Data Analytics**

**Integrated Introductory Cases**

**Integration of Microsoft Excel®, Analytic Solver, and R**

**Writing with Big Data**

**Emphasis on Data Mining**

**McGraw Hill's Connect®**

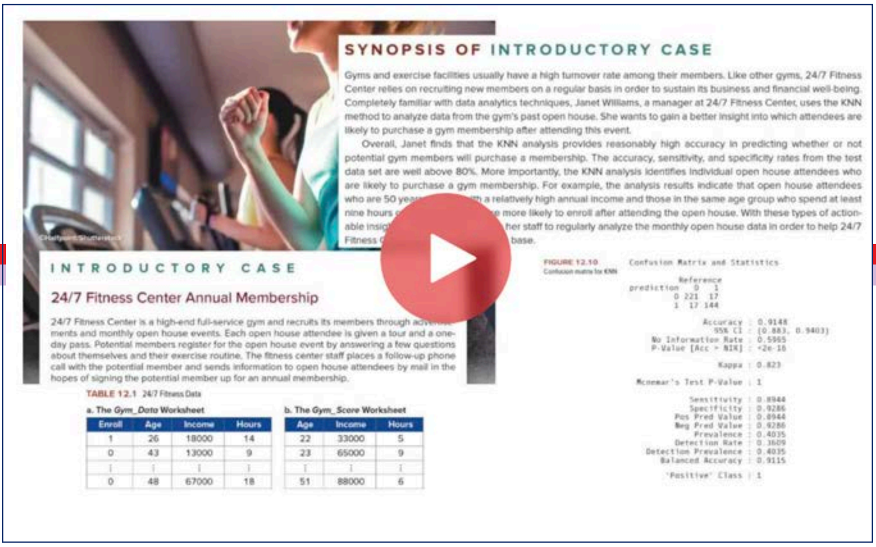
## Features New to the Second Edition

In the second edition of *Business Analytics,* we have made substantial revisions that meet the current needs of the instructors teaching the course and the companies that require the relevant skillset. These revisions are based on the feedback of reviewers and users of our first edition. The greatly expanded coverage of the text gives instructors the flexibility to select the topics that best align with their course objectives.

We cannot possibly list all the improvements made in the second edition. In addition to five new chapters, we have made useful edits in every chapter, including new subsections, examples, computer instructions, data sets, and exercises that incorporate current events such as the recent COVID-19 pandemic. Some of the major improvements are as follows.

- Chapter 1 includes a new subsection on data privacy and data ethics, highlighting the ethical issues emerging from the use and misuse of data.
- Chapter 2 on data wrangling is now based exclusively on Microsoft Excel® and R; Analytic Solver, the Excel add-in, is no longer used in this chapter.
- Chapters 3 and 4 focus on summary measures and data visualization, respectively. These topics were combined into a unified Chapter 3 in the first edition. The expanded coverage allows us to discuss subsetted means in Chapter 3, revealing valuable insights in the data. Also, given the growing popularity of Tableau for its attractive output, versatility, and ease of use, the appendix of Chapter 4 introduces Tableau and provides the detailed instructions for replicating the figures created with Excel and R in the chapter.
- Chapters 5, 6, and 7 no longer require statistical tables. The exercises related to probability distributions, statistical inference, and regression analysis are solved exclusively with Excel and R.
- Chapter 7 on regression analysis includes a new subsection on confidence and prediction intervals regarding the response variable.
- Chapter 9 focuses exclusively on logistic regression models. In the first edition, logistic models were combined with regression models with interaction variables and were used for nonlinear relationships. The exclusive chapter allows us to include topics such as interpreting an odds ratio and assessing model performance with imbalanced data. We also introduce several new applications.
- Chapter 15 is a new chapter on spreadsheet modeling—a widely used tool for business planning and decision support. We discuss the techniques for developing useful spreadsheet models for a wide range of business problems, conducting what-if analysis, and detecting spreadsheet model errors.

- Chapters 16, 17, and 18 emphasize prescriptive analytics, which is an important category of business analytics. Expanding a single chapter in the first edition into three allows a comprehensive coverage of the relevant topics in prescriptive analytics. Chapter 16 provides a more in-depth treatment of risk analysis and simulation. It uses random variables to model risk and uncertainty and applies Monte Carlo simulation models to assess risk and uncertainty in a wide variety of applications. There are two chapters designated for optimization. In Chapter 17, we formulate and solve maximization and minimization linear programming problems and describe special cases and potential issues in linear programming. Chapter 18 discusses specialized linear and integer programming techniques in important business and nonbusiness applications as well as introduces nonlinear programming optimization.
- We include COVD-19 testing data in our Big Data sets. The new data set contains a sample of over 1 million observations that include clinical symptoms, patient demographics, and the testing results released by the Israeli Ministry of Health. The new data set has been incorporated into the Writing with Big Data section throughout the text.



## SYNOPSIS OF INTRODUCTORY CASE

Gyms and exercise facilities usually have a high turnover rate among their members. Like other gyms, 24/7 Fitness Center relies on recruiting new members on a regular basis in order to sustain its business and financial well-being. Completely familiar with data analytics techniques, Janet Williams, a manager at 24/7 Fitness Center, uses the KNN method to analyze data from the gym's past open house. She wants to gain a better insight into which attendees are likely to purchase a gym membership after attending this event.

Overall, Janet finds that the KNN analysis provides reasonably high accuracy in predicting whether or not potential gym members will purchase a membership. The accuracy, sensitivity, and specificity rates from the test data set are well above 80%. More importantly, the KNN analysis identifies individual open house attendees who are likely to purchase a gym membership. For example, the analysis results indicate that open house attendees who are 50 yea[illegible]th a relatively high annual income and those in the same age group who spend at least nine hours [illegible]e more likely to enroll after attending the open house. With these types of actionable insig[illegible] her staff to regularly analyze the monthly open house data in order to help 24/7 Fitness [illegible] base.

©Halfpoint/Shutterstock

## INTRODUCTORY CASE

### 24/7 Fitness Center Annual Membership

24/7 Fitness Center is a high-end full-service gym and recruits its members through advertisements and monthly open house events. Each open house attendee is given a tour and a one-day pass. Potential members register for the open house event by answering a few questions about themselves and their exercise routine. The fitness center staff places a follow-up phone call with the potential member and sends information to open house attendees by mail in the hopes of signing the potential member up for an annual membership.

TABLE 12.1 24/7 Fitness Data

a. The *Gym_Data* Worksheet

| Enroll | Age | Income | Hours |
|---|---|---|---|
| 1 | 26 | 18000 | 14 |
| 0 | 43 | 13000 | 9 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 0 | 48 | 67000 | 18 |

b. The *Gym_Score* Worksheet

| Age | Income | Hours |
|---|---|---|
| 22 | 33000 | 5 |
| 23 | 65000 | 9 |
| ⋮ | ⋮ | ⋮ |
| 51 | 88000 | 6 |

FIGURE 12.10 Confusion matrix for KNN

```
Confusion Matrix and Statistics

          Reference
prediction   0   1
         0 221  17
         1  17 144

               Accuracy : 0.9148
                 95% CI : (0.883, 0.9403)
    No Information Rate : 0.5965
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.823

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8944
            Specificity : 0.9286
         Pos Pred Value : 0.8944
         Neg Pred Value : 0.9286
             Prevalence : 0.4035
         Detection Rate : 0.3609
   Detection Prevalence : 0.4035
      Balanced Accuracy : 0.9115

       'Positive' Class : 1
```

**Click below to watch a video for the author:**

Business Analytics Communicating with Numbers, 2e

## Unique Key Features

The pedagogy of *Business Analytics* reinforces and expands six core features that were well-received in the first edition. Countless reviewers have added their feedback and direction to ensure we have built a product that we believe addresses the needs of the market.

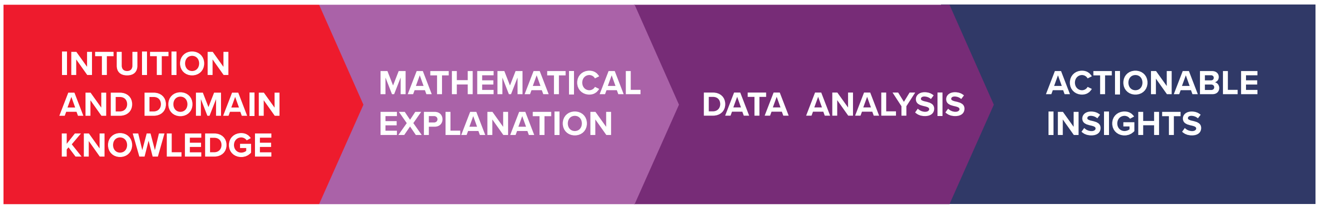
> *“This is by far the best book I have come across. It is easy to follow, very practical, and the examples are rich in detail.”*
>
> **Cary Caro,** *Xavier University of Louisiana*

### Holistic Approach to Data Analytics

Business analytics is a very broad topic consisting of statistics, computer science, and management information systems with a wide variety of applications in business areas including marketing, HR management, economics, accounting, and finance.

The text offers a holistic approach to business analytics, combining qualitative reasoning with quantitative tools to identify key business problems and translate analytics into decisions that improve business performance.

> *“I can’t agree with the approach. . . to data analytics more. I have been looking for a textbook like this.”*
>
> **Jahyun Goo,** *Florida Atlantic University*


INTUITION AND DOMAIN KNOWLEDGE

MATHEMATICAL EXPLANATION

DATA ANALYSIS

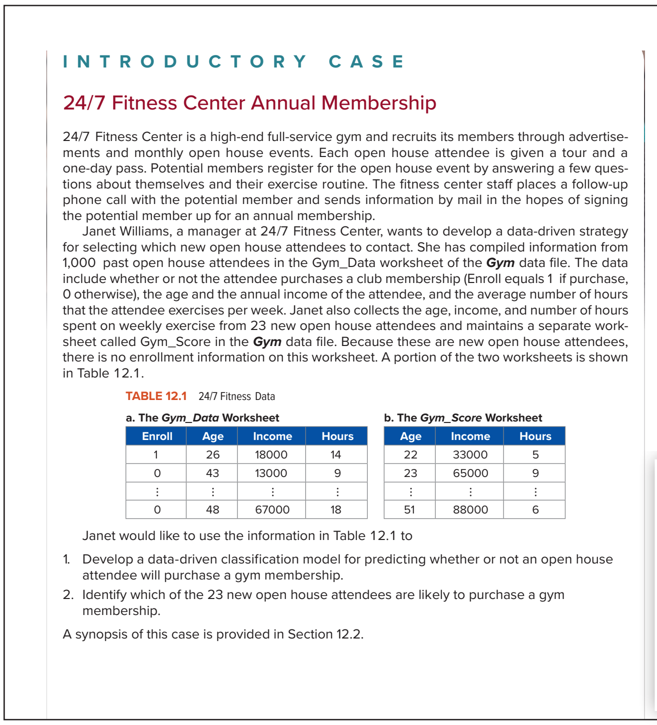
ACTIONABLE INSIGHTS

### Integrated Introductory Case

Each chapter opens with a real-life case study that forms the basis for several examples within the chapter. The questions included in the examples create a roadmap for mastering the most important learning outcomes within the chapter. A synopsis of each chapter’s introductory case is presented once the questions pertaining to the case have been answered.

> *“I think the case studies are excellent. They are varied yet practical and what students will see in the business world.”*
>
> **Ben Williams,** *University of Denver*

> *“I love everything I see! I love the application demonstrated through the case study in each chapter. . .and examples to help students apply the material.”*
>
> **Kristin Pettey,** *Southwestern College–Kansas*


INTRODUCTORY CASE

24/7 Fitness Center Annual Membership

24/7 Fitness Center is a high-end full-service gym and recruits its members through advertisements and monthly open house events. Each open house attendee is given a tour and a one-day pass. Potential members register for the open house event by answering a few questions about themselves and their exercise routine. The fitness center staff places a follow-up phone call with the potential member and sends information by mail in the hopes of signing the potential member up for an annual membership.

Janet Williams, a manager at 24/7 Fitness Center, wants to develop a data-driven strategy for selecting which new open house attendees to contact. She has compiled information from 1,000 past open house attendees in the Gym_Data worksheet of the ***Gym*** data file. The data include whether or not the attendee purchases a club membership (Enroll equals 1 if purchase, 0 otherwise), the age and the annual income of the attendee, and the average number of hours that the attendee exercises per week. Janet also collects the age, income, and number of hours spent on weekly exercise from 23 new open house attendees and maintains a separate worksheet called Gym_Score in the ***Gym*** data file. Because these are new open house attendees, there is no enrollment information on this worksheet. A portion of the two worksheets is shown in Table 12.1.

TABLE 12.1 24/7 Fitness Data

**a. The *Gym_Data* Worksheet**

| Enroll | Age | Income | Hours |
|---|---|---|---|
| 1 | 26 | 18000 | 14 |
| 0 | 43 | 13000 | 9 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 0 | 48 | 67000 | 18 |

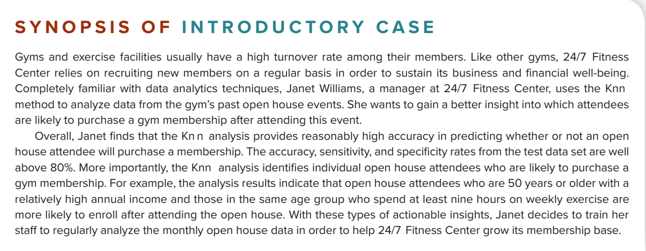
**b. The *Gym_Score* Worksheet**

| Age | Income | Hours |
|---|---|---|
| 22 | 33000 | 5 |
| 23 | 65000 | 9 |
| ⋮ | ⋮ | ⋮ |
| 51 | 88000 | 6 |

Janet would like to use the information in Table 12.1 to

1. Develop a data-driven classification model for predicting whether or not an open house attendee will purchase a gym membership.
2. Identify which of the 23 new open house attendees are likely to purchase a gym membership.

A synopsis of this case is provided in Section 12.2.

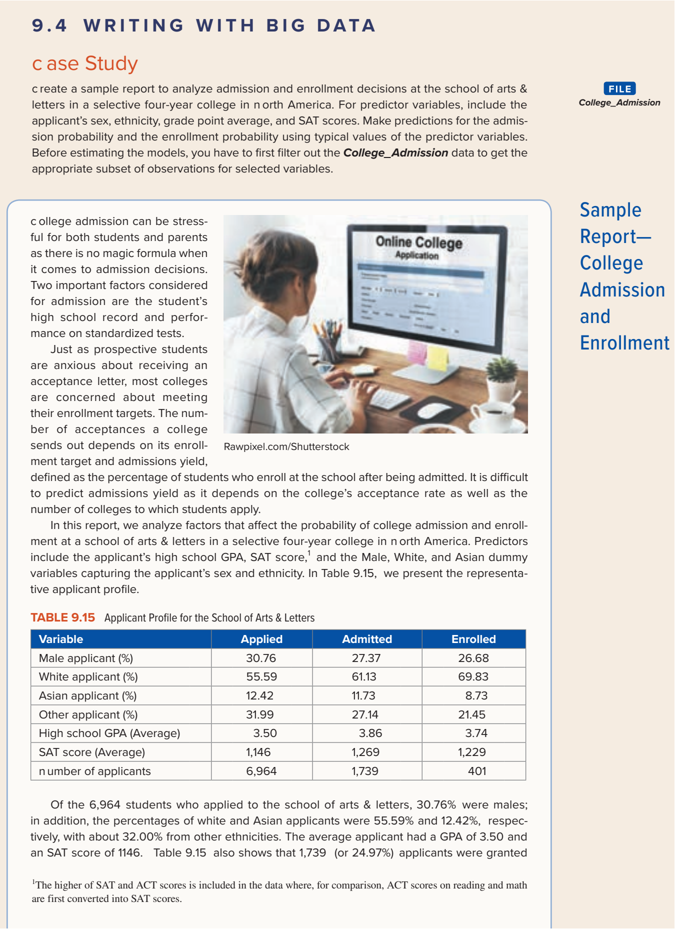

SYNOPSIS OF INTRODUCTORY CASE

Gyms and exercise facilities usually have a high turnover rate among their members. Like other gyms, 24/7 Fitness Center relies on recruiting new members on a regular basis in order to sustain its business and financial well-being. Completely familiar with data analytics techniques, Janet Williams, a manager at 24/7 Fitness Center, uses the Knn method to analyze data from the gym’s past open house events. She wants to gain a better insight into which attendees are likely to purchase a gym membership after attending this event.

Overall, Janet finds that the Kn n analysis provides reasonably high accuracy in predicting whether or not an open house attendee will purchase a membership. The accuracy, sensitivity, and specificity rates from the test data set are well above 80%. More importantly, the Knn analysis identifies individual open house attendees who are likely to purchase a gym membership. For example, the analysis results indicate that open house attendees who are 50 years or older with a relatively high annual income and those in the same age group who spend at least nine hours on weekly exercise are more likely to enroll after attending the open house. With these types of actionable insights, Janet decides to train her staff to regularly analyze the monthly open house data in order to help 24/7 Fitness Center grow its membership base.

## Writing with Big Data

A distinctive feature of *Business Analytics* is access to select big data sets with relevance to numerous applications to which students can relate. In most chapters, we have a designated section where we use these big data sets to help introduce problems, formulate possible solutions, and communicate the findings, based on the concepts introduced in the chapter. Using a sample report, our intent is to show students how to articulate the business value of analytics by communicating insights gained from a nontechnical standpoint.



### 9.4 WRITING WITH BIG DATA

c ase Study

c reate a sample report to analyze admission and enrollment decisions at the school of arts & letters in a selective four-year college in n orth America. For predictor variables, include the applicant's sex, ethnicity, grade point average, and SAT scores. Make predictions for the admission probability and the enrollment probability using typical values of the predictor variables. Before estimating the models, you have to first filter out the ***College_Admission*** data to get the appropriate subset of observations for selected variables.

FILE *College_Admission*

Sample Report—College Admission and Enrollment

c ollege admission can be stressful for both students and parents as there is no magic formula when it comes to admission decisions. Two important factors considered for admission are the student's high school record and performance on standardized tests.

Just as prospective students are anxious about receiving an acceptance letter, most colleges are concerned about meeting their enrollment targets. The number of acceptances a college sends out depends on its enrollment target and admissions yield, defined as the percentage of students who enroll at the school after being admitted. It is difficult to predict admissions yield as it depends on the college's acceptance rate as well as the number of colleges to which students apply.

Rawpixel.com/Shutterstock

In this report, we analyze factors that affect the probability of college admission and enrollment at a school of arts & letters in a selective four-year college in n orth America. Predictors include the applicant's high school GPA, SAT score,[1] and the Male, White, and Asian dummy variables capturing the applicant's sex and ethnicity. In Table 9.15, we present the representative applicant profile.

TABLE 9.15 Applicant Profile for the School of Arts & Letters

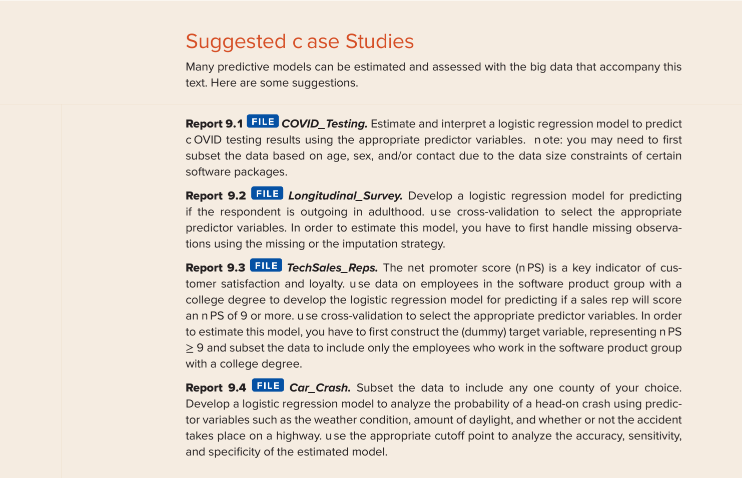
| Variable | Applied | Admitted | Enrolled |
|---|---|---|---|
| Male applicant (%) | 30.76 | 27.37 | 26.68 |
| White applicant (%) | 55.59 | 61.13 | 69.83 |
| Asian applicant (%) | 12.42 | 11.73 | 8.73 |
| Other applicant (%) | 31.99 | 27.14 | 21.45 |
| High school GPA (Average) | 3.50 | 3.86 | 3.74 |
| SAT score (Average) | 1,146 | 1,269 | 1,229 |
| n umber of applicants | 6,964 | 1,739 | 401 |

Of the 6,964 students who applied to the school of arts & letters, 30.76% were males; in addition, the percentages of white and Asian applicants were 55.59% and 12.42%, respectively, with about 32.00% from other ethnicities. The average applicant had a GPA of 3.50 and an SAT score of 1146. Table 9.15 also shows that 1,739 (or 24.97%) applicants were granted

[1] The higher of SAT and ACT scores is included in the data where, for comparison, ACT scores on reading and math are first converted into SAT scores.

*"End of chapter material is excellent ("Writing with Big Data"). . ."*

**Kevin Brown,** *Asbury University*



Suggested c ase Studies

Many predictive models can be estimated and assessed with the big data that accompany this text. Here are some suggestions.

**Report 9.1** FILE ***COVID_Testing.*** Estimate and interpret a logistic regression model to predict c OVID testing results using the appropriate predictor variables. n ote: you may need to first subset the data based on age, sex, and/or contact due to the data size constraints of certain software packages.

**Report 9.2** FILE ***Longitudinal_Survey.*** Develop a logistic regression model for predicting if the respondent is outgoing in adulthood. u se cross-validation to select the appropriate predictor variables. In order to estimate this model, you have to first handle missing observations using the missing or the imputation strategy.

**Report 9.3** FILE ***TechSales_Reps.*** The net promoter score (n PS) is a key indicator of customer satisfaction and loyalty. u se data on employees in the software product group with a college degree to develop the logistic regression model for predicting if a sales rep will score an n PS of 9 or more. u se cross-validation to select the appropriate predictor variables. In order to estimate this model, you have to first construct the (dummy) target variable, representing n PS ≥ 9 and subset the data to include only the employees who work in the software product group with a college degree.

**Report 9.4** FILE ***Car_Crash.*** Subset the data to include any one county of your choice. Develop a logistic regression model to analyze the probability of a head-on crash using predictor variables such as the weather condition, amount of daylight, and whether or not the accident takes place on a highway. u se the appropriate cutoff point to analyze the accuracy, sensitivity, and specificity of the estimated model.

*"The TOC includes all the major areas needed for a foundational level of knowledge and the added value of teaching how to communicate with the information garnered will make a strong textbook."*

**Roman Rabinovich,** *Boston University*

## Emphasis on Data Mining

Data mining is one of the most sought-after skills that employers want college graduates to have. It leverages large data sets and computer power to build predictive models that support decision making. In addition to three comprehensive chapters devoted to linear and logistic regression models, and a chapter on business forecasting, the text includes four exclusive chapters on data mining. These include detailed analysis of both supervised and unsupervised learning, covering relevant topics such as principle component analysis, *k*-nearest neighbors, naïve Bayes, classification and regression trees, ensemble trees, hierarchical and *k*-means clustering, and association rules. Each chapter offers relatable real-world problems, conceptual explanations, and easy-to-follow computer instructions. There are more than 200 exercises in these four exclusive chapters.

**FIGURE 11.4** The cumulative lift chart

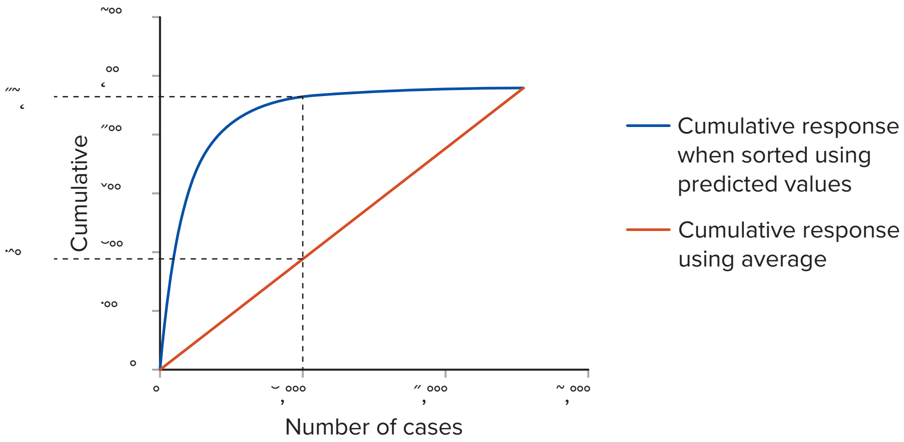


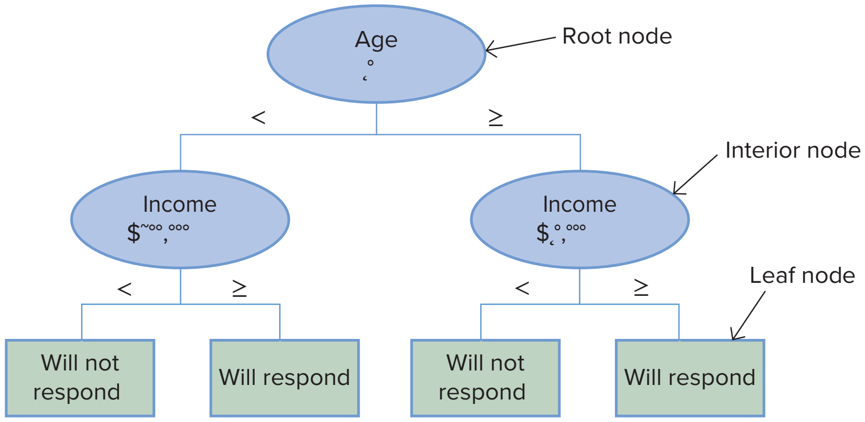
**TABLE 11.14** Prediction Performance Measures

| Performance measure | Model 1 | Model 2 |
|---|---|---|
| RMSE | 171.3489 | 174.1758 |
| ME | 11.2530 | 12.0480 |
| MAD | 115.1650 | 117.9920 |
| MPE | −2.05% | −2.08% |
| MAPE | 15.51% | 15.95% |

## Four Chapters on Data Mining

- introduction to Data Mining
- Supervised Data Mining: *k*-nearest neighbors and naïve Bayes
- Supervised Data Mining: Decision Trees
- Unsupervised Data Mining

**FIGURE 13.1** A simplified decision tree

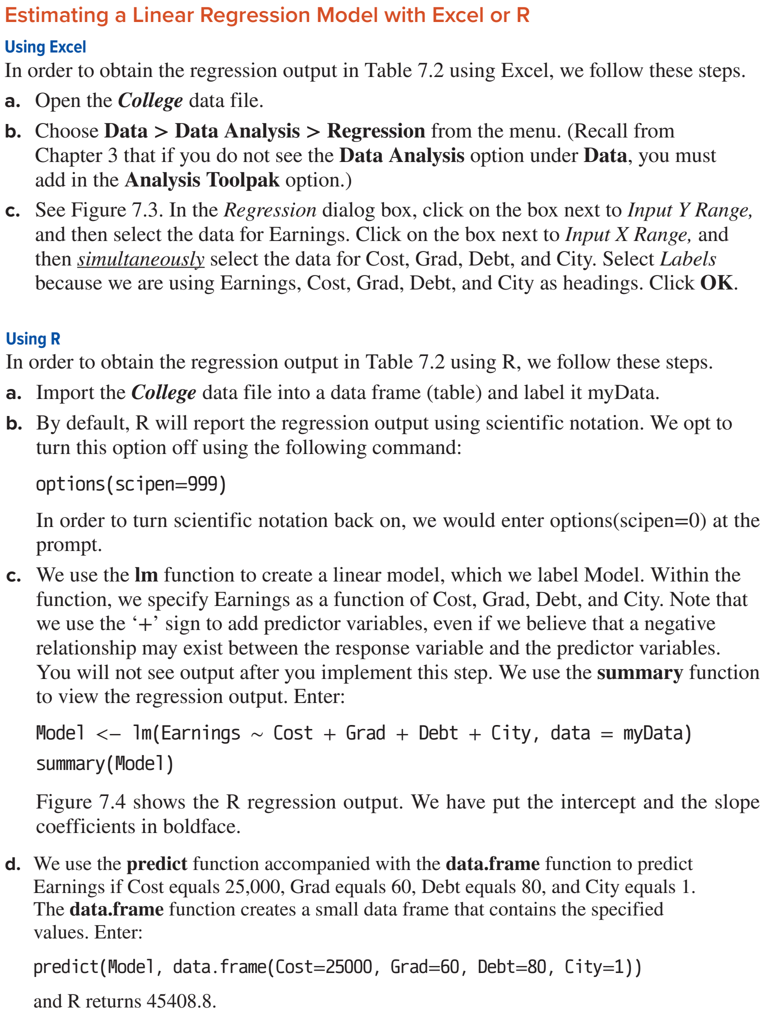


> *"I strongly applaud that this text offers a holistic approach to data analysis and places the emphasis on communicating with data. The Big Data and Data Mining sections seem promising with coverage of topics of the latest models and methods."*
>
> **Hao Chen,** *UW Platteville*

## Computer Software

The text includes hands-on tutorials and problem-solving examples featuring Microsoft Excel, Analytic Solver (an Excel add-in software for data mining analysis), as well as R (a powerful software that merges the convenience of statistical packages with the power of coding). The text includes one chapter dedicated exclusively to spreadsheet modeling and problem solving using Microsoft Excel.

Throughout the text, students learn to use the software to solve real-world problems and to reinforce the concepts discussed in the chapters. Students will also learn how to visualize and communicate with data using charts and infographics featured in the software.



### Estimating a Linear Regression Model with Excel or R

#### Using Excel

In order to obtain the regression output in Table 7.2 using Excel, we follow these steps.

**a.** Open the ***College*** data file.

**b.** Choose **Data > Data Analysis > Regression** from the menu. (Recall from Chapter 3 that if you do not see the **Data Analysis** option under **Data**, you must add in the **Analysis Toolpak** option.)

**c.** See Figure 7.3. In the *Regression* dialog box, click on the box next to *Input Y Range,* and then select the data for Earnings. Click on the box next to *Input X Range,* and then *simultaneously* select the data for Cost, Grad, Debt, and City. Select *Labels* because we are using Earnings, Cost, Grad, Debt, and City as headings. Click **OK**.

#### Using R

In order to obtain the regression output in Table 7.2 using R, we follow these steps.

**a.** Import the ***College*** data file into a data frame (table) and label it myData.

**b.** By default, R will report the regression output using scientific notation. We opt to turn this option off using the following command:

```
options(scipen=999)
```

In order to turn scientific notation back on, we would enter options(scipen=0) at the prompt.

**c.** We use the **lm** function to create a linear model, which we label Model. Within the function, we specify Earnings as a function of Cost, Grad, Debt, and City. Note that we use the '+' sign to add predictor variables, even if we believe that a negative relationship may exist between the response variable and the predictor variables. You will not see output after you implement this step. We use the **summary** function to view the regression output. Enter:

```
Model <- lm(Earnings ~ Cost + Grad + Debt + City, data = myData)
summary(Model)
```

Figure 7.4 shows the R regression output. We have put the intercept and the slope coefficients in boldface.

**d.** We use the **predict** function accompanied with the **data.frame** function to predict Earnings if Cost equals 25,000, Grad equals 60, Debt equals 80, and City equals 1. The **data.frame** function creates a small data frame that contains the specified values. Enter:

```
predict(Model, data.frame(Cost=25000, Grad=60, Debt=80, City=1))
```

and R returns 45408.8.

> *"I love that Excel and R are integrated into this chapter. Using these analytic programs for prescriptive analytics fits my student learning objectives and keeps the most prominent analytic tools at the forefront of learning."*
>
> **John Branner,** *Cape Fear Community College*

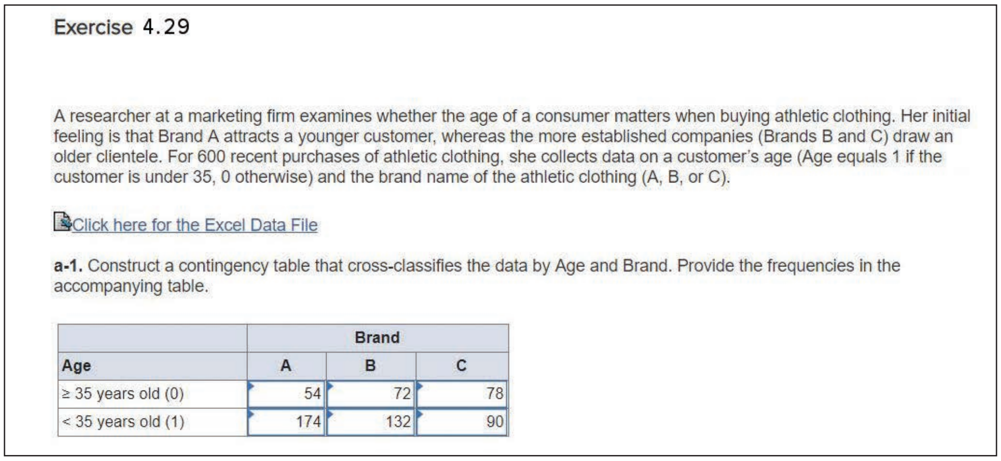
> *"I love the Excel/R examples and I love that the exercises and examples are after each section."*
>
> **Edie Schmidt,** *Nova Southeastern University*

### Exercises and McGraw Hill's Connect®

Every chapter contains dozens of applied examples from all walks of life, including business, economics, sports, health, housing, the environment, polling, psychology, and more.

We also know the importance of ancillaries—like the Instructor's Solution Manual (ISM)—and the technology component, specifically Connect. As we write *Business Analytics,* we are simultaneously developing these components with the hope of making them seamless with the text itself.

We know from experience that these components cannot be developed in isolation. For example, we review every Connect exercise as well as evaluate rounding rules and revise tolerance levels. Given the extremely positive feedback from users of our *Business Statistics* texts, we follow the same approach with *Business Analytics.*



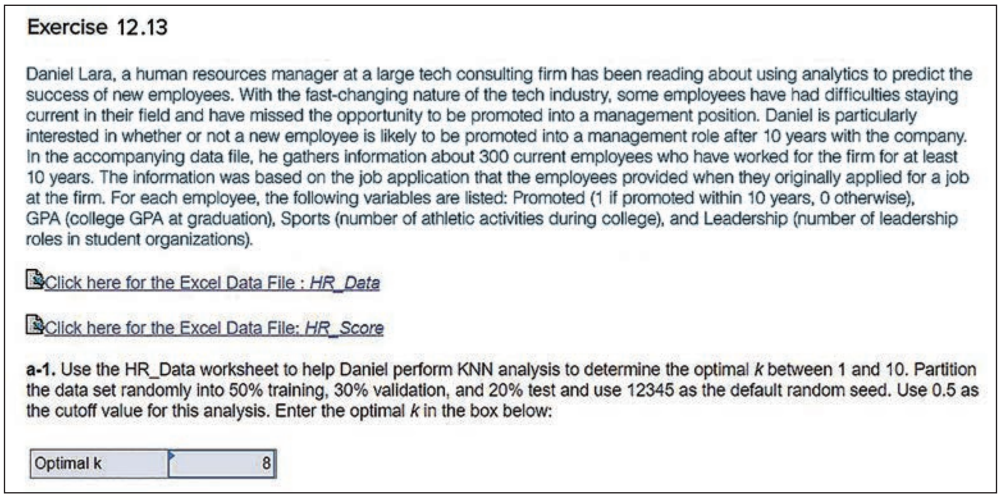
Exercise 4.29

A researcher at a marketing firm examines whether the age of a consumer matters when buying athletic clothing. Her initial feeling is that Brand A attracts a younger customer, whereas the more established companies (Brands B and C) draw an older clientele. For 600 recent purchases of athletic clothing, she collects data on a customer's age (Age equals 1 if the customer is under 35, 0 otherwise) and the brand name of the athletic clothing (A, B, or C).

Click here for the Excel Data File

**a-1.** Construct a contingency table that cross-classifies the data by Age and Brand. Provide the frequencies in the accompanying table.

| | Brand | | |
|---|---|---|---|
| **Age** | **A** | **B** | **C** |
| ≥ 35 years old (0) | 54 | 72 | 78 |
| < 35 years old (1) | 174 | 132 | 90 |



Exercise 12.13

Daniel Lara, a human resources manager at a large tech consulting firm has been reading about using analytics to predict the success of new employees. With the fast-changing nature of the tech industry, some employees have had difficulties staying current in their field and have missed the opportunity to be promoted into a management position. Daniel is particularly interested in whether or not a new employee is likely to be promoted into a management role after 10 years with the company. In the accompanying data file, he gathers information about 300 current employees who have worked for the firm for at least 10 years. The information was based on the job application that the employees provided when they originally applied for a job at the firm. For each employee, the following variables are listed: Promoted (1 if promoted within 10 years, 0 otherwise), GPA (college GPA at graduation), Sports (number of athletic activities during college), and Leadership (number of leadership roles in student organizations).

Click here for the Excel Data File : *HR_Data*

Click here for the Excel Data File: *HR_Score*

**a-1.** Use the HR_Data worksheet to help Daniel perform KNN analysis to determine the optimal *k* between 1 and 10. Partition the data set randomly into 50% training, 30% validation, and 20% test and use 12345 as the default random seed. Use 0.5 as the cutoff value for this analysis. Enter the optimal *k* in the box below:

| | |
|---|---|
| Optimal k | 8 |

> *"Exercise questions are well designed and presented, showing an excellent flow for student learning."*
>
> **Chaodong Han,** *Towson University*

## Integrated Excel

**The power of Microsoft Excel meets the power of Connect.** In this new assignment type, called ***Integrated Excel,*** Excel opens seamlessly inside Connect with no need to upload or download any additional files or software. Instructors choose their preferred auto-graded solution, with options for formula accuracy ONLY or either formula or solution value.

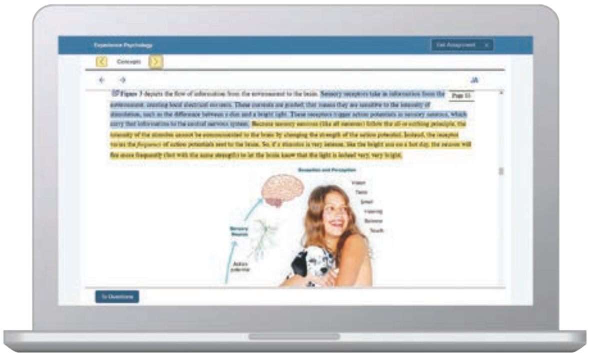
McGraw Hill connect®

# Instructors: Student Success Starts with you

## Tools to enhance your unique voice

Want to build your own course? n o problem. Prefer to use an OI C-aligned, prebuilt course? Easy. Want to make changes throughout the semester? Sure. And you'll save time with Connect's auto-grading too.

**65%**
**Less Time Grading**



Laptop: McGraw Hill; Woman/dog: George Doyle/Getty Images

## Study made personal

incorporate adaptive study resources like SmartBook® 2.0 into your course and help your students be better prepared in less time. I earn more about the powerful personalized learning experience available in SmartBook 2.0 at **www.mheducation.com/highered/connect/smartbook**

## Affordable solutions, added value

Make technology work for you with I MS integration for single sign-on access, mobile access to the digital textbook, and reports to quickly show you how each of your students is doing. And with our inclusive Access program you can provide all these tools at a discount to your students. Ask your McGraw Hill representative for more information.

Padlock: Jobalou/Getty Images

## Solutions for your challenges

A product isn't a solution. Real solutions are affordabl , reliable, and come with training and ongoing support when you need it and how you want it. Visit **www.supportateverystep.com** for videos and resources both you and your students can use throughout the semester.

Checkmark: Jobalou/Getty Images

SUPPORT AT every step

# **Students:** Get l earning that Fits you

## Effective tools for efficient study

Connect is designed to help you be more productive with simple, fl xible, intuitive tools that maximize your study time and meet your individual learning needs. Get learning that works for you with Connect.

## Study anytime, anywhere

Download the free ReadAnywhere app and access your online eBook, SmartBook 2.0, or Adaptive I earning Assignments when it’s convenient, even if you’re offli . And since the app automatically syncs with your Connect account, all of your work is available every time you open it. Find out more at **www.mheducation.com/readanywhere**

> ***“I really liked this app—it made it easy to study when you don't have your text-book in front of you.”***
>
> - Jordan Cunningham, Eastern Washington University

Calendar: owattaphotos/Getty Images

## Everything you need in one place

your Connect course has everything you need—whether reading on your digital eBook or completing assignments for class, Connect makes it easy to get your work done.

## Learning for everyone

McGraw Hill works directly with Accessibility Services Departments and faculty to meet the learning needs of all students. Please contact your Accessibility Services Office and ask them to emai accessibility@mheducation.com, or visit **www.mheducation.com/about/accessibility** for more information.

Top: Jenner Images/Getty Images, Left: Hero Images/Getty Images, Right: Hero Images/Getty Images

# Resources for instructors and Students

## Instructor Library

The Connect Instructor Library is your repository for additional resources to improve student engagement in and out of class. You can select and use any asset that enhances your course. The Connect Instructor Library includes:

- Instructor's Manual
- Instructor's Solutions Manual
- Test Bank
- Data Sets
- PowerPoint Presentations
- Digital Image Library

## R Package

R is a powerful software that merges the convenience of statistical packages with the power of coding. It is open source as well as cross-platform compatible and gives students the flexibility to work with large data sets using a wide range of analytics techniques. The software is continuously evolving to include packages that support new analytical methods. In addition, students can access rich online resources and tap into the expertise of a worldwide community of R users. In Appendix C, we introduce some fundamental features of R and also provide instructions on how to obtain solutions for many solved examples in the text.

As with other texts that use R, differences between software versions are likely to result in minor inconsistencies in analytics outcomes in algorithm-rich Chapters 11, 12, 13, and parts of Chapter 14. In these chapters, the solved examples and exercise problems are based on R version 3.5.3 on Microsoft Windows. In order to replicate the results with newer versions of R, we suggest a line of code in these chapters that sets the random number generator to the one used on R version 3.5.3.

## Analytic Solver

The Excel-based user interface of Analytic Solver reduces the learning curve for students allowing them to focus on problem solving rather than trying to learn a new software package. The solved examples and exercise problems are based on the 2021 version of Analytic Solver Desktop. Newer versions of Analytic Solver will likely produce the same analysis results but may have a slightly different user interface.

Analytic Solver can be used with Microsoft Excel for Windows (as an add-in), or "in the cloud" at **AnalyticSolver.com** using any device (PC, Mac, tablet) with a web browser. It offers comprehensive features for prescriptive analytics (optimization, simulation, decision analysis) and predictive analytics (forecasting, data mining, text mining). Its optimization features are upward compatible from the standard Solver in Excel. If interested in having students get low-cost academic access for class use, instructors should send an email to support@solver.com to get their course code and receive student pricing and access information as well as their own access information.

## Student Resources

Students have access to data files, tutorials, and detailed progress reporting within Connect. Key textbook resources can also be accessed through the Additional Student Resources page: **www.mhhe.com/JaggiaBA1e**.

# McGraw Hill Customer Care Contact information

At McGraw Hill, we understand that getting the most from new technology can be challenging. That's why our services don't stop after you purchase our products. You can e-mail our product specialists 24 hours a day to get product training online. Or you can search our knowledge bank of frequently asked questions on our support website.

For customer support, call 800-331-5094 or visit **www.mhhe.com/support**. One of our technical support analysts will be able to assist you in a timely fashion.

## Remote Proctoring & Browser-Locking Capabilities

Mc Graw Hill connect + proctorio

New remote proctoring and browser-locking capabilities, hosted by Proctorio within Connect, provide control of the assessment environment by enabling security options and verifying the identity of the student.

Seamlessly integrated within Connect, these services allow instructors to control students' assessment experience by restricting browser activity, recording students' activity, and verifying students are doing their own work.

Instant and detailed reporting gives instructors an at-a-glance view of potential academic integrity concerns, thereby avoiding personal bias and supporting evidence-based claims.

# ACKNOWLEDGMENTS

We would like to thank those who reviewed and provided valuable feedback for this and the previous edition of the text:

Sung K. Ahn, Washington State University
Triss Ashton, Tarleton State University
Anteneh Ayanso, Brock University
Matthew Bailey, Bucknell University
Palash Beram, Saint Louis University
John Branner, Cape Fear Community College
Matthew Brenn, University of Akron
Paul Brooks, Virginia Commonwealth University
Kevin Brown, Asbury University
Kevin Caskey, SUNY–New Paltz
Paolo Catasti, Virginia Commonwealth University
Michael Cervetti, University of Memphis
Xiaohui Chang, Oregon State University
Jimmy Chen, Bucknell University
Jen-Yi Chen, Cleveland State University
Rex Cheung, San Francisco State University
Alan Chow, University of South Alabama
Matthew Dean, University of Southern Maine
Alisa DiSalvo, St. Mary's University–Twin Cities
Mark Dobeck, Cleveland State University
Joan Donohue, University of South Carolina
Kathy Enget, University at Albany–SUNY
Kathryn Ernstberger, Indiana University Southeast
Jeffrey Forrest, Slippery Rock University
Robertas Gabrys, University of Southern California
Ridvan Gedik, University of New Haven
Roger Grinde, University of New Hampshire
Thomas Groleau, Carthage College
Babita Gupta, California State University–Monterey Bay
Serina Al Haddad, Rollins College
Regina Halpin, Auburn University
Chaodong Han, Towson University
Fady Harfoush, Loyola University Chicago
Dan Harpool, University of Arkansas–Little Rock
Michael Higgins, Kansas State University
Paul Holmes, Ashland University
Ping-Hung Hsieh, Oregon State University
Kuang-Chung Hsu, University of Central Oklahoma
Jason Imbrogno, University of Northern Alabama
Marina Johnson, Montclair State University
Jerzy Kamburowski, University of Toledo
Reza Kheirandish, Clayton State University
Esther Klein, St. Francis College
Bharat Kolluri, University of Hartford
Jun Liu, Georgia Southern University
Mohammad Merhi, Indiana University
Jacob Miller, Southern Utah University
Jamie Minas, Ithaca College
Sinjini Mitra, California State University–Fullerton
Kyle Moninger, Bowling Green State University
Rex Moody, Angelo State University
Ebrahim Mortaz, Pace University
Kristin Pettey, Southwestern College
Sharma Pillutla, Towson University
Daniel Power, University of Northern Iowa
Zbigniew Przasnyski, Loyola Marymount University
Roman Rabinovich, Boston University
Michael Ratajczyk, Saint Mary's College
R. Christopher L. Riley, Delta State University
John Robertson, Arkansas State University
Tim Rodgers, Colorado State University
Wendy Roth, Georgia State
Leslie Rush, University of Hawaii–West Oahu
Avijit Sarkar, University of Redlands
Edie Schmidt, Nova Southeastern University
Dmitriy Shaltayev, Christopher Newport University
Mike Shurden, Lander University
Pearl Steinbuch, Boston University
Alicia Strandberg, Villanova University School of Business

David Taylor, Sacred Heart University
Stan Taylor, California State University–Sacramento
Pablo Trejo, Miami Dade College
Nikhil Varaiya, San Diego State University
Timothy Vaughan, University of Wisconsin–Eau Claire
Fen Wang, Central Washington University
Ben Williams, University of Denver
Rick Wilson, Oklahoma State University
Liang Xu, Slippery Rock University of Pennsylvania
Yong Xu, Ferris State University
Guisen Xue, Kent State University
Oliver Yi, San Jose State University
Kanghyun Yoon, University of Central Oklahoma
John Yu, Saint Joseph's University
Jay Zagorsky, Boston University

We would also like to thank Christopher Chiang, Jenna Eisenman, Matthew Hawke, Brandon J. Meeker, Tiadora Miric, and Aidan Nevens for their outstanding research assistance.

And finally, special thanks to the wonderful and supportive team at McGraw Hill: Ryan McAndrews, Product Developer; Thomas Finn, Content Development Editor; Noelle Bathurst, Portfolio Manager; Harper Christopher, Executive Marketing Manager; Sarah Thomas, Market Development Manager; Pat Frederickson, Core Content Project Manager; Jamie Koch, Assessment Content Project Manager; Beth Blech and Laurie Entringer Designers; Chuck Synovec, Portfolio Manager; Kevin Moran, Director of Digital Content; and Tim Vertovec, Managing Director.

# BRIEF CONTENTS

## APPENDIXES

# CONTENTS

# 2 Data Management and Wrangling

## LEARNING OBJECTIVES

**After reading this chapter, you should be able to:**

LO **2.1** **Describe the key concepts related to data management.**

LO **2.2** **Inspect and explore data.**

LO **2.3** **Apply data preparation techniques to handle missing values and to subset data.**

LO **2.4** **Transform numerical variables.**

LO **2.5** **Transform categorical variables.**

Data wrangling is the process of retrieving, cleansing, integrating, transforming, and enriching data to support analytics. It is often considered one of the most critical and time-consuming steps in an analytics project. In this chapter, we focus on four key tasks during the data wrangling process: data management, data inspection, data preparation, and data transformation.

We first provide an overview of data management. Although data management is primarily the responsibility of the information technology group, understanding the relevant concepts, data structure, and data retrieval technologies is essential to the success of analytics professionals. After obtaining relevant data, most analytics professionals spend a considerable amount of time inspecting, cleansing, and preparing the data for subsequent analysis. These tasks often involve counting and sorting relevant variables to review basic information and potential data quality issues.

For data preparation, we discuss two commonly performed tasks: handling missing values and subsetting data. We then examine two strategies for handling missing values: omission and imputation. Finally, we focus on data transformation techniques for both numerical and categorical variables. For numerical variables, common data transformation techniques include binning, creating new variables, and rescaling. For categorical variables, common data transformation techniques include reducing categories, creating dummy variables, and creating category scores.

Yuliia Mazurkevych/Shutterstoick

## INTRODUCTORY CASE

### Gaining Insights into Retail Customer Data

Organic Food Superstore is an online grocery store that specializes in providing organic food products to health-conscious consumers. The company offers a membership-based service that ships fresh ingredients for a wide range of chef-designed meals to its members' homes. Catherine Hill is a marketing manager at Organic Food Superstore. She has been assigned to market the company's new line of Asian-inspired meals. Research has shown that the most likely customers for healthy ethnic cuisines are college-educated millennials (born on or after 1/1/1982 and before 1/1/2000 ).

In order to spend the company's marketing dollars efficiently, Catherine wants to focus on this target demographic when designing the marketing campaign. With the help of the information technology (IT) group, Catherine has acquired a representative sample that includes each customer's identification number (CustID), sex (Sex), race (Race), birthdate (BirthDate), whether the customer has a college degree (College), household size (HouseholdSize), zip code (ZipCode), annual income (Income), total spending in 2020 (Spending2020), total spending in 2021 (Spending2021 ), total number of orders during the past 24 months (NumOfOrders), number of days since the last order (DaysSinceLast), the customer's rating on the last purchase (Satisfaction), and the channel through which the customer was originally acquired (Channel). Table 2.1 shows a portion of the ***Customers*** data set.

**TABLE 2.1** A Sample of Organic Food Superstore Customers

| CustID | Sex | Race | BirthDate | . . . | Channel |
|---|---|---|---|---|---|
| 1530016 | Female | Black | 12/16/1986 | . . . | SM |
| 1531136 | Male | White | 5/9/1993 | . . . | TV |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 1579979 | Male | White | 7/5/1999 | . . . | SM |

FILE
*Customers*

Catherine will use the ***Customers*** data set to:

1. Identify Organic Food Superstore's college-educated millennial customers.
2. Compare the profiles of female and male college-educated millennial customers.

A synopsis of this case is provided at the end of Section 2.3.

LO 2.1

Describe the key concepts related to data management.

## 2.1 DATA MANAGeMeNT

**Data wrangling** is the process of retrieving, cleansing, integrating, transforming, and enriching data to support subsequent data analysis. This valuable process focuses on transforming the raw data into a format that is more appropriate and easier to analyze. The objectives of data wrangling include improving data quality, reducing the time and effort required to perform analytics, and helping reveal the true intelligence in the data.

Data wrangling is an essential part of business analytics. As mentioned in Chapter 1, the increasing volume, velocity, and variety of data compel organizations to spend a large amount of time and resources in garnering, cleaning, and organizing data before performing any analysis. As the amount of data grows, the need and difficulties involving data wrangling increase. In practice, the inability to clean and organize big data is among the primary barriers preventing organizations from taking full advantage of business analytics.

### DATA WRANGLING

Data wrangling is the process of retrieving, cleansing, integrating, transforming, and enriching data to support subsequent data analysis.

Analytics professionals can no longer rely solely on the corporate IT department for data retrieval and preparation. According to interviews and expert estimates, analytics professionals spend up to 95% of their time in the mundane task of collecting and preparing unruly data, before analytics can be applied (*Open Data Science,* August 6, 2019). As such, analytics professionals have to become more self-reliant and possess the necessary skills for data wrangling as well as performing data analysis. This practice allows organizations to address business problems much more quickly and make better-informed decisions. At the same time, the self-service model requires analytics professionals to have a broader skill set than just statistical and data mining techniques.

We first provide an overview of **data management**. In a very broad sense, data management is the process that an organization uses to acquire, organize, store, manipulate, and distribute data. Organizations today have a plethora of data created and stored using different, often incompatible, technologies. For the past few decades, most organizations have adopted the database approach for storing and managing data. This has tremendously improved the efficiency and effectiveness of the data management process and ultimately the quality of data. A **database** is a collection of data logically organized to enable easy retrieval, management, and distribution of data.

The most common type of database used in organizations today is the **relational database**. A relational database consists of one or more logically related data files, often called tables or relations. Each table is a two-dimensional grid that consists of rows (also called records or tuples) and columns (also called fields or attributes). A column (e.g., sex of a customer, price of a product, etc.) contains a characteristic of a physical object (e.g., products or places), an event (e.g., business transactions), or a person (e.g., customers, students). A collection of related columns makes up a record, which represents an object, event, or person. A software application for defining, manipulating, and managing data in databases is called a **database management system** (DBMS). Popular DBMS packages include Oracle, IBM DB2, SQL Server, MySQL, and Microsoft Access.

### DATA MANAGEMENT

Data management is the process that an organization uses to acquire, organize, store, manipulate, and distribute data.

The most common type of database (a collection of data) is the relational database. A relational database consists of one or more logically related data tables, where each data table is a two-dimensional grid that consists of rows and columns.

## Data Modeling: The Entity-Relationship Diagram

To understand how and where data can be extracted, one needs to understand the structure of the data, also known as the data model. **Data modeling** is the process of defining the structure of a database. Relational databases are modeled in a way to offer great flexibility and ease of data retrieval.

An **entity-relationship diagram (ERD)** is a graphical representation used to model the structure of the data. An **entity** is a generalized category to represent persons, places, things, or events about which we want to store data in a database table. A single occurrence of an entity is called an **instance**. In most situations, an instance is represented as a record in a database table. For example, Claire Johnson is an instance of a CUSTOMER entity, and organic oatmeal is an instance of a PRODUCT entity. Each entity has specific characteristics called attributes or fields, which are represented as columns in a database table. Customers' last names and product descriptions are examples of attributes in the CUSTOMER and PRODUCT database tables, respectively.

### Entity Relationships

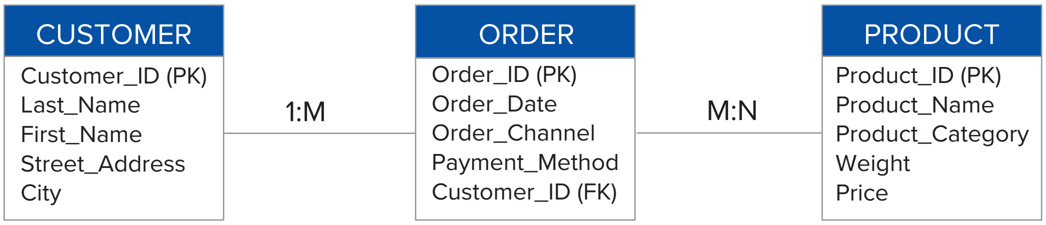
Two entities can have a one-to-one (1:1), one-to-many (1:M), or many-to-many (M:N) **relationship** with each other that represents certain business facts or rules. A 1:1 relationship is less common than the other two types. In a business setting, we might use a 1:1 relationship to describe a situation where each department can have only one manager, and each manager can only manage one department.

Recall the Organic Food Superstore from the introductory case. Figure 2.1 shows an ERD for the store's database that illustrates examples of 1:M and M:N relationships. The diagram shows three entities: CUSTOMER, ORDER, and PRODUCT. The relationship between CUSTOMER and ORDER entities is 1:M because one customer can place many orders over time, but each order can only belong to one customer. The relationship between ORDER and PRODUCT is M:N because an order can contain many products and the same product may appear in many orders.

In Figure 2.1, each entity is represented in a rectangular-shaped box in which attributes of the entity are listed. For each entity, there is a special type of attribute called **primary key** (PK), which is an attribute that uniquely identifies each instance of the entity. For example, Customer_ID is the primary key for the CUSTOMER entity because each customer would have a unique ID number. Because the primary key attribute uniquely identifies each instance of the entity, it is often used to create a data structure called an index for fast data retrieval and searches.

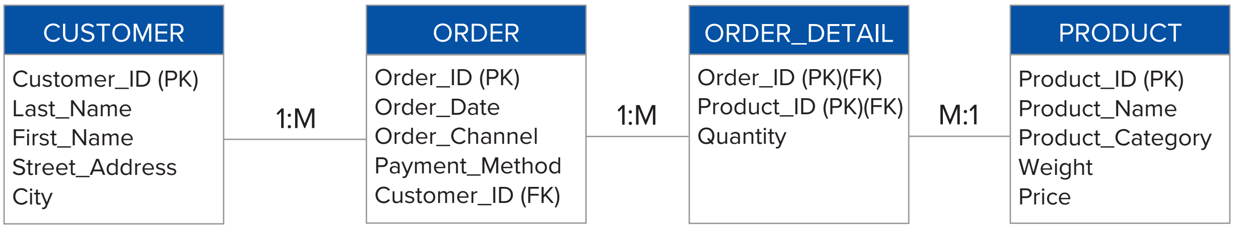
Some entities (e.g., ORDER) have another special type of attribute called **foreign key** (FK). A foreign key is defined as a primary key of a related entity. Because Customer_ID is the primary key of the CUSTOMER entity, which shares a relationship with the ORDER entity, it is considered a foreign key in the ORDER entity. A pair of the primary and foreign keys is used to establish the 1:M relationship between two entities. By matching the values in the Customer_ID fields of the CUSTOMER and ORDER entities, we can quickly find out which customer placed which order. As this example shows, the primary key field belongs to the table that is on the one side of the relationship, whereas the foreign key field belongs to the table that is on the many side of the relationship.

**FIGURE 2.1** Example of an entity relationship diagram



The ERD in Figure 2.1 is not yet complete as it is missing the ordered products and purchase quantities in the ORDER entity. Storing these data in the ORDER entity is not appropriate as one does not know in advance how many products are going to be in each order, therefore making it impossible to create the correct number of attributes. To resolve this issue, we simply create an intermediate entity, ORDER_DETAIL, between the ORDER and PRODUCT entities, as shown in Figure 2.2. As a result, the M:N relationship is decomposed into two 1:M relationships. An order has many detailed line items, but each line item can only belong to one order. While a product may appear in many orders and order lines, an order line can only contain one product.

**FIGURE 2.2** An expanded entity relationship diagram



In the ORDER_DETAIL entity, two attributes, Order_ID and Product_ID, together create a unique identifier for each instance. In this situation, Order_ID and Product_ID are referred to as a **composite primary key**, which is a primary key that consists of more than one attribute. We use a composite primary key when none of the individual attributes alone can uniquely identify each instance of the entity. For example, neither Order_ID nor Product_ID alone can uniquely identify each line item of an order, but a combination of them can uniquely identify each line item. Because both Order_ID and Product_ID are primary keys of other entities related to ORDER_DETAIL, they also serve as foreign keys. By matching the primary and foreign key pair, the system can quickly find out which order contains a line item for a particular product.

### THE ENTITY-RELATIONSHIP DIAGRAM (ERD)

An entity-relationship diagram (ERD) is a schematic used to illustrate the structure of the data.

- An entity is a generalized category to represent persons, places, things, or events.
- A relationship between entities represents certain business facts or rules. The types of relationships include one-to-one (1:1), one-to-many (1:M), and many-to-many (M:N).
- A primary key is an attribute that uniquely identifies each instance of an entity, whereas a foreign key is the primary key of a related entity. A composite primary key is a primary key that contains more than one attribute.

The data model represented in an ERD can be converted into database tables. Based on the ERD in Figure 2.2, Table 2.2 shows various tables that can be created using Organic Food Superstore sample data. Can you find out which customer placed an order for organic sweet potato on October 15, 2021? Did this customer order any other products? By matching the primary and foreign keys of the CUSTOMER, ORDER,

ORDER_DETAIL, and PRODUCT tables, we can establish relationships among these tables. With these relationships, we can extract useful information from multiple tables. For example, using his customer ID (i.e., 1531136), we learn that James Anderson was the customer who placed an order on October 15, 2021, using his mobile phone and paying for the order with his PayPal account. In addition, we can also see that he purchased organic sweet potato and organic oatmeal.

**TABLE 2.2** Database Tables for Organic Food Superstore

**a) CUSTOMER table**

| Customer_ID | Last_Name | First_Name | Street_Address | City | . . . |
|---|---|---|---|---|---|
| 1530016 | Johnson | Claire | 532 Main Street | Los Angeles | . . . |
| 1531136 | Anderson | James | 1322 Cary S treet | Los Angeles | . . . |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 1532160 | Smith | Terry | 663 Johnson Ave. | Los Angeles | . . . |

**b) ORDER table**

| Order_ID | Order_Date | Order_Channel | Payment_Method | . . . | Customer_ID |
|---|---|---|---|---|---|
| 1484001 | 09/12/2021 | Web | Credit/Debit Card | . . . | 1530016 |
| 1484212 | 3/24/2021 | Web | Credit/Debit Card | . . . | 1530016 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 1482141 | 10/15/2021 | Mobile | paypal | . . . | 1531136 |

**c) ORDER_DETAIL table**

| Order_ID | Product_ID | Quantity |
|---|---|---|
| 1484001 | 4378 | 1 |
| 1482141 | 4305 | 1 |
| ⋮ | ⋮ | ⋮ |
| 1482141 | 4330 | 2 |

**d) PRODUCT table**

| Product_ID | Product_Name | Product_Category | Weight | Price | . . . |
|---|---|---|---|---|---|
| 4305 | Organic Oatmeal | Cereals | 2 | 2.49 | . . . |
| 4330 | Organic Sweet potato | produce | 1 | 1.39 | . . . |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 4378 | Gluten-Free Bread | Bakery | 1.5 | 6.99 | . . . |

## Data Retrieval in the Database Environment

Once data are stored in a relational database, we can retrieve them using database queries, which are requests to access data from a database. The most popular query language used today is **Structured Query Language (SQL)**. SQL is a language for manipulating data in a relational database using relatively simple and intuitive commands. While a comprehensive discussion of SQL is beyond the scope of this text, we briefly demonstrate how simple SQL statements can be used to retrieve data from a database.

The basic structure of a SQL statement is relatively simple and usually consists of three keywords: Select, From, and Where. The Select keyword is followed by the names of attributes we want to retrieve. The From keyword specifies the tables from which to retrieve the data. We usually want to retrieve data based on selection criteria specified in the Where clause. The following SQL statement retrieves first and last names of the customers who live in Los Angeles from the CUSTOMER table.

```
Select Last_Name, First_Name
From CUSTOMER
Where City = “Los Angeles”
```

While simple queries like the previous one are useful, we often need to compile data from multiple tables and apply more than one selection criterion. For example, we may want to retrieve the customer names, order IDs, and order quantities of organic sweet potato purchases on October 15, 2021. The following SQL query returns this information.

```
Select CUSTOMER.Last_Name, CUSTOMER.First_Name, ORDER.Order_ID,
   ORDER_DETAIL.Quantity
From CUSTOMER, ORDER, ORDER_DETAIL, PRODUCT
Where PRODUCT.Product_Name = “Organic Sweet Potato”
and ORDER.Order_Date = “10/15/2021”
and CUSTOMER.Customer_ID = ORDER.Customer_ID
and ORDER.Order_ID = ORDER_DETAIL.Order_ID
and ORDER_DETAIL.Product_ID = PRODUCT.Product_ID
```

Because the Select clause specifies the attributes to retrieve and these attributes may come from different tables, we use the *table_name.attribute_name* format to tell the DBMS from which table the attributes are retrieved (e.g., *CUSTOMER.Last_Name* refers to the last name attribute in the CUSTOMER table). The From clause lists all the tables that have the attributes to be retrieved and the attributes used in the selection criteria. The Where clause lists multiple conditions and links them using the “and” keyword. The last three conditions match the values of the primary and foreign key pairs based on the relationships depicted in Figure 2.2.

### STRUCTURED QUERY LANGUAGE (SQL)

In a relational database, data can be retrieved using Structured Query Language (SQL) statements, whose basic structure consists of the Select, From, and Where keywords. SQL statements specify the attributes, tables, and criteria the retrieved data must meet.

While the SQL commands in the previous two examples provide the flexibility for data retrieval, many modern DBMS packages such as Microsoft Access, Oracle, and SQL Server offer a graphical interface option called Query by Example (QBE), where the user constructs a database query by dragging and dropping the query components (e.g., fields, tables, conditions, etc.) into a form. The DBMS then translates user selections into SQL statements to retrieve data.

## Data Warehouse and Data Mart

Although the relational database environment provides businesses with an efficient and effective way to manage data, the proliferation of isolated databases maintained by different business functional areas makes it difficult to analyze data across departments and business units. This phenomenon has also given rise to data redundancy and inconsistency. As a result, many organizations have developed an enterprise data warehouse, which offers an integrated, accurate “single version of truth” to support decision making.

An enterprise data warehouse or **data warehouse** is a central repository of data from multiple functional areas within an organization. One of its primary purposes is to support managerial decision making, and, therefore, data in a data warehouse are usually organized around subjects such as sales, customers, or products that are relevant to business decision making. In order to integrate data from different databases generated by various business functional areas, an extraction, transformation, and load (ETL) process is undertaken to retrieve, reconcile, and transform data into a consistent format, and then load the final data into a data warehouse. A data warehouse provides a historical and comprehensive view of the entire organization.

As you can imagine, the volume of the data in a data warehouse can become very large, very quickly. While a data warehouse integrates data across the entire organization, a **data mart** is a small-scale data warehouse or a subset of the enterprise data warehouse that focuses on one particular subject or decision area. For example, a data mart can be designed to support the marketing department for analyzing customer behaviors, and it contains only the data relevant to such analyses.

The structure of a data mart conforms to a multidimensional data model called a **star schema**, which is a specialized relational database model. Figure 2.3 displays the star schema for Organic Food Superstore. In the star schema, there are two types of tables: dimension and fact tables. The **dimension table** describes business dimensions of interest such as customer, product, location, and time. The **fact table** contains facts about the business operation, often in a quantitative format.

### DATA WAREHOUSE AND DATA MART

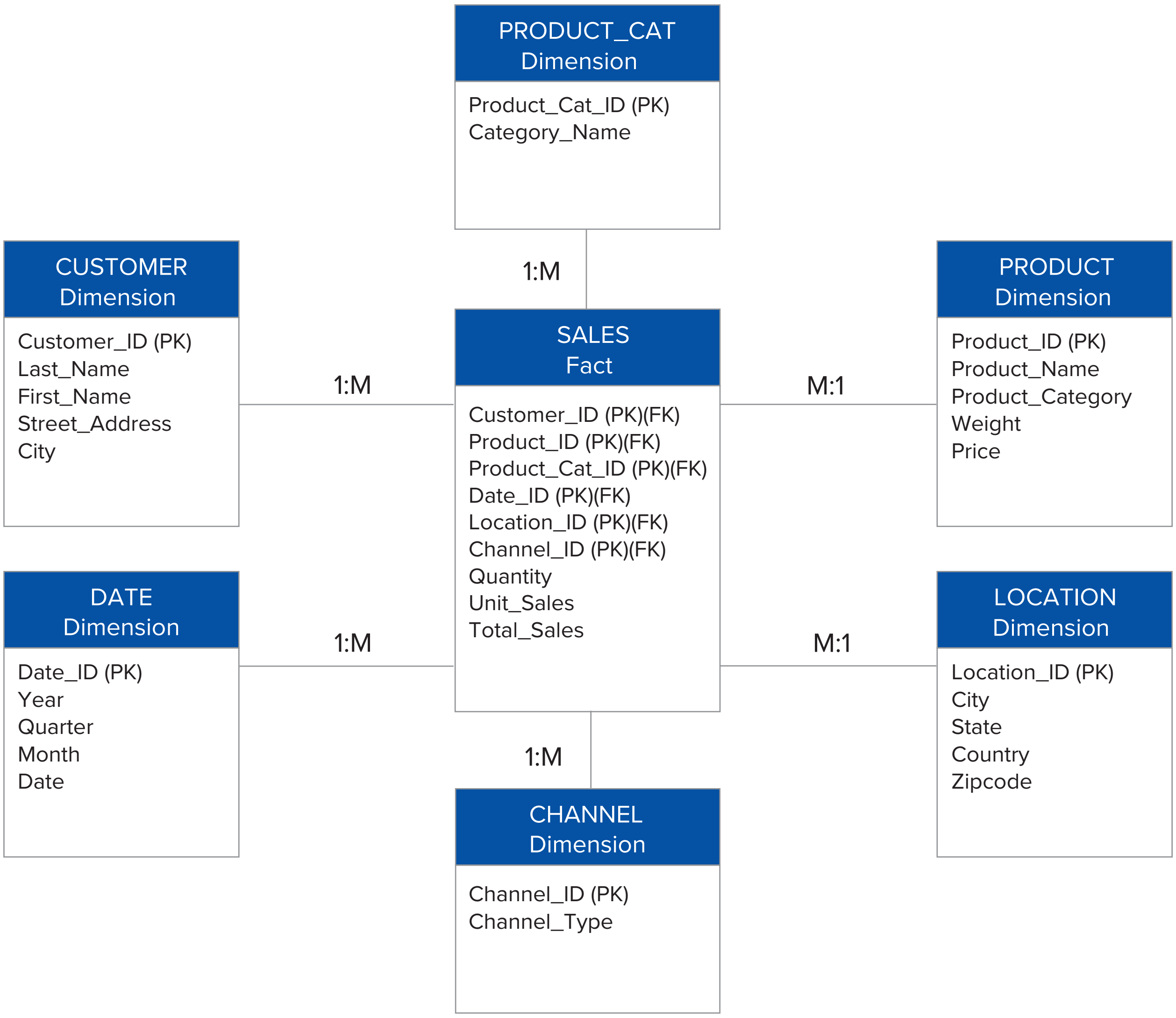
A data warehouse is a central repository of data from multiple functional areas within an organization to support managerial decision making. Analytics professionals tend to acquire data from data marts, which are small-scale data warehouses that only contain data that are relevant to certain subjects or decision areas. Data in a data mart are organized using a multidimensional data model called a star schema, which includes dimension and fact tables.

Each of the dimension tables has a 1:M relationship with the fact table. Hence, the primary keys of the dimension tables are also the foreign keys in the fact table. At the same time, the combination of the primary keys of the dimension tables forms the composite primary key of the fact table. The fact table is usually depicted at the center surrounded by multiple dimension tables forming the shape of a star. In reality, it is not uncommon to see multiple fact tables sharing relationships with a group of dimension tables in a data mart.

One of the key advantages of the star schema is its ability to “slice and dice” data based on different dimensions. For example, in the data model shown in Figure 2.3, sales data can be retrieved based on who the customer is; which product or product category is involved; the year, quarter, or month of the order; where the customer lives; or through which channel the order was submitted simply by matching the primary keys of the dimension tables with the foreign keys of the fact table.

In recent years, new forms of databases to support big data have emerged. The most notable is the NoSQL or “Not Only SQL” database. The NoSQL database is a non-relational database that supports the storage of a wide range of data types including structured, semistructured, and unstructured data. It also offers the flexibility, performance, and scalability needed to handle extremely high volumes of data. Analytics professionals will likely see NoSQL databases implemented alongside relational databases to support organizations’ data needs in today’s environment.

**FIGURE 2.3** Star schema of a data mart for Organic Food Superstore



# Exercises 2.1

## Applications

1. Which of the following statements correctly describe the data wrangling process? Select all that apply. Explain if incorrect.
   a. Data wrangling is the process of retrieving, cleansing, integrating, transforming, and enriching data.
   b. Data wrangling is the process of defining and modeling the structure of a database to represent real-world events.
   c. The objectives of data wrangling include improving data quality and reducing the time and effort required to perform analytics.
   d. Data wrangling focuses on transforming the raw data into a format that is more appropriate and easier to analyze.
2. Which of the following statements about entity-relationship diagrams (ERDs) are correct? Select all that apply. Explain if incorrect.
   a. An entity usually represents persons, places, things, or events about which we want to store data.
   b. A foreign key is an attribute that uniquely identifies each instance of the entity.
   c. A composite key is a key that consists of more than one attribute.
   d. A relationship between entities represents certain business facts or rules.
3. Which of the following statements correctly identify and describe the key elements of a relational database? Select all that apply. Explain if incorrect.

a. A table in a relational database is a two-dimensional grid that contains actual data.

b. A field or a column represents a characteristic of a physical object, an event, or a person.

c. A relational database includes software tools for advanced data visualization.

d. A tuple or a record in a table represents a physical object, an event, or a person.

4. Which of the following statements best describes what a foreign key is? Select all that apply. Explain if incorrect.

   a. It is an attribute that uniquely identifies each instance of the entity.

   b. It is a primary key that consists of multiple attributes.

   c. It is the primary key of a related database table.

   d. It is a single occurrence of an entity.

5. Which type of relationship–one-to-one (1:1), one-to-many (1:M), or many-to-many (M:N)–do the following business rules describe?

   a. One manager can supervise multiple employees, and one employee may report to multiple managers.

   b. A business department has multiple employees, but each employee can be assigned to only one department.

   c. A company can have only one CEO, and each CEO can work for only one company.

   d. An academic adviser can work with multiple students, while each student is assigned to only one adviser.

   e. A golf course offers a membership to many members, and a golfer can potentially sign up for a membership at multiple golf courses.

   f. A soccer team consists of multiple players, while an individual player can play for only one team at a time.

   g. A national healthcare management company operates a number of medical facilities across the country. Each medical facility primarily serves one regional area. A patient can choose to visit any medical facility for treatments and will be billed by the healthcare management company centrally. What are the relationships between the healthcare management company and its medical facilities and between the medical facilities and patients?

6. Which of the following statements correctly describes the benefits of Structured Query Language (SQL)? Select all that apply. Explain if incorrect.

   a. SQL can be used to manipulate structured, semi-structured, and unstructured data.

   b. SQL commands allow users to select data based on multiple selection criteria.

   c. SQL can be used to compile data from multiple tables.

   d. SQL commands are relatively simple and intuitive.

7. Which of the following statements about data warehouses and data marts are correct? Select all that apply. Explain if incorrect.

   a. A data warehouse is a subset of the enterprise database that focuses on one particular subject or decision area.

   b. The dimension table describes business dimensions of interest, such as customer, product, location, and time, while the fact table contains facts about the business operation.

   c. A star schema represents a multidimensional data model.

   d. A data warehouse is the central repository of data from multiple departments within a business enterprise to support managerial decision making.

## 2.2 DATA INSpeCTION

LO 2.2

Inspect and explore data.

Once the raw data are extracted from the database, data warehouse, or data mart, we usually review and inspect the data set to assess data quality and relevant information for subsequent analysis. In addition to visually reviewing data, counting and sorting are among the very first tasks most data analysts perform to gain a better understanding and insights into the data. Counting and sorting data help us verify that the data set is complete or that it may have missing values, especially for important variables. Sorting data also allows us to review the range of values for each variable. We can sort data based on a single variable or multiple variables.

In Example 2.1, we demonstrate how to use counting and sorting features in Excel and R to inspect and gain insights into the data. While these features also allow us to detect missing values, we discuss the treatment of missing values in Section 2.3.

## EXAMPLE 2.1

FILE
*Gig*

BalanceGig is a company that matches independent workers for short-term engagements with businesses in the construction, automotive, and high-tech industries. The 'gig' employees work only for a short period of time, often on a particular project or a specific task. A manager at BalanceGig extracts the employee data from their most recent work engagement, including the hourly wage (HourlyWage), the client's industry (Industry), and the employee's job classification (Job). A portion of the ***Gig*** data set is shown in Table 2.3.

**TABLE 2.3** Gig Employee Data

| EmployeeID | HourlyWage | Industry | Job |
|---|---|---|---|
| 1 | 32.81 | Construction | Analyst |
| 2 | 46.00 | Automotive | engineer |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 604 | 26.09 | Construction | Other |

The manager suspects that data about the gig employees are sometimes incomplete, perhaps due to the short engagement and the transient nature of the employees. She would like to find the number of missing observations for the HourlyWage, Industry, and Job variables. In addition, she would like information on the number of employees who (1) worked in the automotive industry, (2) earned more than $30 per hour, and (3) worked in the automotive industry and earned more than $30 per hour. Finally, the manager would like to know the hourly wage of the lowest- and the highest-paid employees at the company as a whole and the hourly wage of the lowest- and the highest-paid accountants who worked in the automotive and the tech industries.

Use counting and sorting functions in Excel and R to find the relevant information requested by the manager, and then summarize the results.

**Important:** Due to different fonts and type settings, copying and pasting Excel or R functions from this text directly into Excel or R may cause errors. When such errors occur, you may need to replace special characters such as quotation marks and parentheses or delete extra spaces in the functions.

### SOLUTION:

### Using Excel

**a.** Open the ***Gig*** data file. Note that the employee data are currently sorted by their employee ID in column A. Scroll to the end of the data set and note that the last record is in row 605. With the column heading in row 1, the data set has a total of 604 records.

**b.** We use two Excel functions, **COUNT** and **COUNTA**, to inspect the number of values in each column. The **COUNT** function counts the number of cells that contain numeric values and, therefore, can only apply to the EmployeeID and HourlyWage variables. The **COUNTA** function counts the number of cells that are not empty and is applicable to all four variables. Because HourlyWage is a numerical variable, we can enter either =COUNT(B2:B605) or =COUNTA(B2:B605) in an empty cell to count the number of values for HourlyWage. We get 604 values, implying that there are no missing values. Similarly, we enter =COUNTA(C2:C605) and =COUNTA(D2:D605) in empty cells to count the number of values for the Industry (column C) and Job (column D) variables. Because these two variables are non-numerical, we use **COUNTA** instead of **COUNT**. Verify that the number of records for Industry and Job are 594 and 588, respectively, indicating that there are 10 and 16 blank or missing values, respectively, in these two variables.

**c.** To count the number of employees in each industry, we use the **COUNTIF** function. Entering =COUNTIF(C2:C605,"=Automotive") in an empty cell will show that 190 of the 604 employees worked in the automotive industry. Similarly, entering =COUNTIF(B2:B605,">30") in an empty cell will show that 536 employees earned more than $30 per hour. Note that the first parameter in the **COUNTIF** function is the range of cells to be counted, and the second parameter specifies the selection criterion. Other logical operators such as >=, <, <=, and <> (not equal to) can also be used in the **COUNTIF** function.

**d.** To count the number of employees with multiple selection criteria, we use the **COUNTIFS** function. For example, entering =COUNTIFS(C2:C605, "=Automotive", B2:B605,">30") in an empty cell will show that 181 employees worked in the automotive industry and earned more than $30 per hour. Additional data ranges and selection criteria can be added in corresponding pairs. The >=, <, <=, and <> operators can also be used in the **COUNTIFS** function.

**e.** To sort all employees by their hourly wage, highlight cells A1 through D605. From the menu, click **Data > Sort** (in the Sort & Filter group). Make sure that the *My data has headers* checkbox is checked. Select HourlyWage for the *Sort by* option and choose the *Smallest to Largest* (or ascending) order. Click **OK**.

At the top of the sorted list, verify that there are three employees with the lowest hourly wage of $24.28. To sort data in descending order, repeat step e but choose the *Largest to Smallest* (or descending) order. Verify that the highest hourly wage is $51.00.

**f.** To sort the data based on multiple variables, again highlight cells A1:D605 and go to **Data > Sort**. Choose Industry in the *Sort by* option and the *A to Z* (or ascending) order. Click the *Add Level* button and choose Job in the *Then by* option and the *A to Z* order. Click the *Add Level* button again and choose HourlyWage in the second *Then by* option and the *Smallest to Largest* order. Click **OK**. We see that the lowest- and the highest-paid accountants who worked in the automotive industry made $28.74 and $49.32 per hour, respectively.

Similarly, sorting the data by industry in descending order (*Z to A*) and then by job classification and hourly wage in ascending order reveals that the lowest- and the highest-paid accountants in the Tech industry made $36.13 and $49.49 per hour, respectively.

**g.** To re-sort the data set to its original order, again highlight cells A1:D605 and go to **Data > Sort**. Select each of the *Then by* rows and click the *Delete Level* button. Choose EmployeeID in the *Sort by* option and the *Smallest to Largest* order.

### Using R

Before following all R instructions, make sure that you have read Appendix C ("Getting Started with R"). We assume that you have downloaded R and RStudio and that you know how to import an Excel file. Throughout the text, our goal is to provide the simplest way to obtain the relevant output. We denote all function names in **boldface** and all options within a function in *italics*.

**a.** Import the ***Gig*** data file into a data frame (table) and label it myData. Keep in mind that the R language is case sensitive.

**b.** We use the **dim** function in R to count the number of observations and variables. Verify that the R output shows 604 observations and four variables. Enter:

```
dim(myData)
```

**c.** Two common functions to display a portion of data are **head** and **View**. The **head** function displays the first few observations in the data set, and the **View**

function (case sensitive) displays a spreadsheet-style data viewer where the user can scroll through rows and columns. Verify that the first employee in the data set is an analyst who worked in the construction industry and made $32.81 per hour. Enter:

```
head(myData)
View(myData)
```

**d.** R stores missing values as *NA*, and we use the **is.na** function to identify the observations with missing values. R labels observations with missing values as "TRUE" and observations without missing values as "FALSE." In order to inspect the Industry variable for missing values, enter:

```
is.na(myData$Industry)
```

The R result displays a list of logical values indicating whether a value is missing (TRUE) or present (FALSE) for the Industry variable.

**e.** For a large data set, having to look through all observations is inconvenient. Alternately, we can use the **which** function together with the **is.na** function to identify "which" observations contain missing values. The following command identifies 10 observations by row number as having a missing value in the Industry variable. Verify that the first observation with a missing Industry value is in row 24. Enter:

```
which (is.na(myData$Industry))
```

**f.** To inspect the 24th observation, we specify row 24 in the myData data frame. Enter:

```
myData[24,]
```

Note that there are two elements within the square bracket, separated by a comma. The first element identifies a row number (also called row index), and the second element after the comma identifies a column number (also called column index). Leaving the second element blank will display all columns. To inspect an observation in row 24 and column 3, we enter `myData[24,3]`. In a small data set, we can also review the missing values by scrolling to the specific rows and columns in the data viewer produced by the **View** function. As mentioned earlier, the treatment of missing values is discussed in Section 2.3.

**g.** To identify and count the number of employees with multiple selection criteria, we use the **which** and **length** functions. In the following command, we identify which employees worked in the automotive industry with the **which** function and count the number of these employees using the **length** function. The double equal sign (==), also called the equality operator, is used to check whether the industry is automotive. In R, text characters such as 'Automotive' are enclosed in quotation marks. Enter:

```
length(which(myData$Industry=='Automotive'))
```

We can also use the >, >=, <, <=, and != (not equal to) operators in the selection criteria. For example, using the following command, we can determine the number of employees who earn more than $30 per hour. Enter:

```
length(which(myData$HourlyWage > 30))
```

Note that there are 190 employees in the automotive industry and there are 536 employees who earn more than $30 per hour.

**h.** To count how many employees worked in a particular industry and earned more than a particular wage, we use the **and** operator (&). The following command shows that 181 employees worked in the automotive industry and earned more than $30 per hour. Enter:

```
length(which(myData$Industry==‘Automotive’ & myData$HourlyWage >
30))
```

**i.** We use the **order** function to sort the observations of a variable. In order to sort the HourlyWage variable and store the ordered data set in a new data frame called sortedData1, enter:

```
sortedData1 <- myData[order(myData$HourlyWage),]
View(sortedData1)
```

The **View** function shows that the lowest and highest hourly wages are $24.28 and $51.00, respectively. By default, the sorting is performed in ascending order. To sort in descending order, enter:

```
sortedData1 <- myData[order(myData$HourlyWage, decreasing = TRUE),]
```

**j.** To sort data by multiple variables, we specify the variables in the **order** function. The following command sorts the data by industry, job classification, and hourly wage, all in ascending order, and stores the ordered data in a data frame called sortedData2. Enter:

```
sortedData2 <- myData[order(myData$Industry, myData$Job,
myData$HourlyWage),]
View(sortedData2)
```

The **View** function shows that the lowest-paid accountant who worked in the automotive industry made $28.74 per hour.

**k.** To sort the data by industry and job classification in ascending order and then by hourly wage in descending order, we insert a minus sign in front of the hourly wage variable. Verify that the highest-paid accountant in the automotive industry made $49.32 per hour. Enter:

```
sortedData3 <- myData[order(myData$Industry, myData$Job,
-myData$HourlyWage),]
View(sortedData3)
```

**l.** The industry and job classification variables are non-numerical. As a result, to sort the data by industry in descending order and then by job classification and hourly wage in ascending order, we use the **xtfrm** function, which converts non-numerical values into integers, with the minus sign in front of the Industry variable. Enter:

```
sortedData4 <- myData[order(-xtfrm(myData$Industry), myData$Job,
myData$HourlyWage),]
View(sortedData4)
```

The **View** function reveals that the lowest- and the highest-paid accountants in the technology industry made $36.13 and $49.49 per hour, respectively.

**m.** To sort the data by industry, job, and hourly wage, all in descending order, we use the *decreasing* option in the **order** function. Verify that the highest-paid sales representative in the technology industry made $48.87. Enter:

```
sortedData5 <- myData[order(myData$Industry, myData$Job,
myData$HourlyWage, decreasing = TRUE),]
View(sortedData5)
```

**n.** To export the sorted data from step m as a comma-separated values (csv) file, we use the **write.csv** function. Verify that the exported file is in the default folder on your computer (e.g., the Documents on Microsoft Windows). Other data frames in R can be exported using a similar statement. Enter:

```
write.csv(sortedData5,“sortedData5.csv”)
```

**Summary**

- There are a total of 604 records in the data set. There are no missing values in the HourlyWage variable. The Industry and Job variables have 10 and 16 missing values, respectively.
- 190 employees worked in the automotive industry, 536 employees earned more than $30 per hour, and 181 employees worked in the automotive industry and earned more than $30 per hour.
- The lowest and the highest hourly wages in the data set are $24.28 and $51.00, respectively. The three employees who had the lowest hourly wage of $24.28 all worked in the construction industry and were hired as Engineer, Sales Rep, and Accountant, respectively. Interestingly, the employee with the highest hourly wage of $51.00 also worked in the construction industry in a job type classified as Other.
- The lowest- and the highest-paid accountants who worked in the automotive industry made $28.74 and $49.32 per hour, respectively. In the technology industry, the lowest- and the highest-paid accountants made $36.13 and $49.49 per hour, respectively. Note that the lowest hourly wage for an accountant is considerably higher in the technology industry compared to the automotive industry ($36.13 > $28.74).

There are many ways to count and sort data to obtain useful insights. To gain further insights, students are encouraged to experiment with the ***Gig*** data using different combinations of counting and sorting options than the ones used in Example 2.1.

## EXERCISES 2.2

### Mechanics

8. FILE ***Exercise_2.8.*** The accompanying data file contains two numerical variables, $x_1$ and $x_2$.
   a. For $x_2$, how many of the observations are equal to 2?
   b. Sort $x_1$ and then $x_2$, both in ascending order. After the variables have been sorted, what is the first observation for $x_1$ and $x_2$?
   c. Sort $x_1$ and then $x_2$, both in descending order. After the variables have been sorted, what is the first observation for $x_1$ and $x_2$?
   d. Sort $x_1$ in ascending order and $x_2$ in descending order. After the variables have been sorted, what is the first observation for $x_1$ and $x_2$?
   e. How many missing values are there in $x_1$ and $x_2$?

9. FILE ***Exercise_2.9.*** The accompanying data file contains three numerical variables, $x_1$, $x_2$, and $x_3$.
   a. For $x_1$, how many of the observations are greater than 30?
   b. Sort $x_1$, $x_2$, and then $x_3$ all in ascending order. After the variables have been sorted, what is the first observation for $x_1$, $x_2$, and $x_3$?
   c. Sort $x_1$ and $x_2$ in descending order and $x_3$ in ascending order. After the variables have been sorted, what is the first observation for $x_1$, $x_2$, and $x_3$?
   d. How many missing values are there in $x_1$, $x_2$, and $x_3$?

10. FILE ***Exercise_2.10.*** The accompanying data file contains three numerical variables, $x_1$, $x_2$, and $x_3$, and one categorical variable, $x_4$.
   a. For $x_4$, how many of the observations are less than three?
   b. Sort $x_1$, $x_2$, $x_3$, and then $x_4$ all in ascending order. After the variables have been sorted, what is the first observation for $x_1$, $x_2$, $x_3$, and $x_4$?

c. Sort $x_1$, $x_2$, $x_3$, and then $x_4$ all in descending order. After the variables have been sorted, what is the first observation for $x_1$, $x_2$, $x_3$, and $x_4$?

d. How many missing values are there in $x_1$, $x_2$, $x_3$, and $x_4$?

e. How many observations are there in each category in $x_4$?

## Applications

11. FILE ***SAT.*** The accompanying data file lists the average writing and math SAT scores for the 50 states as well as the District of Columbia, Puerto Rico, and the U.S. Virgin Islands for the year 2017 as reported by the College Board.

    a. Sort the data by writing scores in descending order. Which state has the highest average writing score? What is the average math score of that state?

    b. Sort the data by math scores in ascending order. Which state has the lowest average math score? What is the average writing score of that state?

    c. How many states reported an average math score higher than 600?

    d. How many states reported an average writing score lower than 550?

12. FILE ***Fitness.*** A social science study conducts a survey of 418 individuals . The accompanying data file shows how often they exercise (Exercise), marital status (Married: Yes/No), and annual income (Income).

    a. Sort the data by annual income. Of the 10 highest income earners, how many of them are married and always exercise?

    b. Sort the data by marital status and exercise, both in descending order. How many of the individuals who are married and exercise sometimes earn more than $110,000 per year?

    c. How many missing values are there in each variable?

    d. How many individuals are married and unmarried?

    e. How many married individuals always exercise? How many unmarried individuals never exercise?

13. FILE ***Spend.*** A company conducts a survey of 500 consumers. The accompanying data file shows their home ownership (OwnHome: Yes/No), car ownership (OwnCar: Yes/No), annual household spending on food (Food), and annual household spending on travel (Travel).

    a. Sort the data by home ownership, car ownership, and the travel spending all in descending order. How much did the first customer on the ordered list spend on food?

    b. Sort the data only by the travel spending amount in descending order. Of the 10 customers who spend the most on traveling, how many of them are homeowners? How many of them are both homeowners and car owners?

    c. How many missing values are there in each variable?

    d. How many customers are homeowners?

    e. How many customers are homeowners but do not own a car?

14. FILE ***Demographics.*** The accompanying data file shows 890 individuals' income (Income in $1,000s), age, sex (F = female, M = male), and marital status (Married: Y = yes, N = no).

    a. Count the number of males and females in the data.

    b. What percentages of males and females are married?

    c. Of the 10 individuals with the highest income, how many are married males.

    d. What are the highest and the lowest incomes of males and females?

    e. What are the highest and lowest incomes of married and unmarried males?

15. FILE ***Admission.*** College admission is a competitive process where, among other things, the SAT and high school GPA scores of students are evaluated to make an admission decision. The accompanying data file contains the admission decision (Decision: Admit/Deny), SAT score, Female (Yes/No), and high school GPA (HSGPA) for 1,230 students .

    a. Count the number of male and female students.

    b. What percentages of male and female students are admitted?

    c. Of the 10 students with the highest HSGPA, how many are males?

    d. Of the 10 students with the lowest SAT, how many are females?

    e. What are the highest and the lowest SAT scores of admitted male and female students?

## 2.3 DATA pRepARATION

LO **2.3**

Apply data preparation techniques to handle missing values and to subset data.

Once we have inspected and explored data, we can start the data preparation process. In this section, we examine two important data preparation tasks: handling missing values and subsetting data. As mentioned in Section 2.2, there may be missing values in the key variables that are crucial for subsequent analysis. Moreover, most data analysis projects focus only on a portion (subset) of the data, rather than the entire data set; or sometimes the objective of the analysis is to compare two subsets of the data.

## Handling Missing Values

It is common to find missing values in data sets both large and small. This issue can lead to significant reduction in the number of usable observations for the analysis. For example, in a data set with 20 variables, if 5% of the values, spread randomly across the observations and variables, are missing, then potentially only $(1 - 0.05)^{20} = 0.3585$, or 35.85%, of the observations would be complete and usable. The reduction in the sample size not only reduces statistical power, it can also introduce a bias when the data are not missing at random. Understanding why the values are missing is the first step in the treatment of missing values.

Sometimes data are missing because the respondents decline to provide the information due to its sensitive nature (e.g., race, sexual orientation, economic status, etc.). In these cases, missing values are usually not distributed randomly across observations but tend to concentrate within one or more subgroups. For example, research has shown that male respondents often skip questions related to depression in a survey.

In other cases, data values are missing because some of the items do not apply to every respondent. For instance, patients who are still in the hospital do not have values in the discharge date column. Missing values can also be caused by human errors, sloppy data collection, and equipment failures.

Because missing values are often unavoidable in real life, it is important to learn how to handle observations with missing values. There are two common strategies for dealing with missing values: **omission** and **imputation**. The omission strategy, also called complete-case analysis, recommends that observations with missing values be excluded from the analysis. This approach is appropriate when the amount of missing values is small and are expected to be  distributed randomly across observations.

The imputation strategy replaces missing values with some reasonable imputed values. The most commonly used imputation strategy for numerical variables is the simple mean imputation where the missing values are replaced with the mean (average) values across relevant observations. For example, if the annual household income for an observation is missing, we replace the missing value with the mean household income across all observations or across a homogenous group (e.g., households with the same zip code).

Simple mean imputation is easy to implement and allows observations with missing values to be included in the analysis. However, if a large number of missing values need to be imputed, simple mean imputation will likely distort the relationships among variables. For example, the total square footage of a house tends to have a positive relationship with the value of the house. If a data set contains many missing total square footage values and these missing values are replaced with mean total square footage of the rest of the houses in the data set, then the relationship between total square footage and house value will likely be distorted. More advanced imputation techniques such as regression mean imputation that better preserve the relationships among variables can be used in these cases. Advanced imputation techniques are beyond the scope of this text.

In the case of categorical variables, the most frequent category is often used as the imputed value. For instance, if some values of sex are missing in a data set, we might replace these missing values with the predominant sex category among the rest of the observations. For categorical variables, an "Unknown" category may be created to signify missing values. This approach is especially useful if the data are missing for a reason; the fact that the values are missing for these observations may suggest certain patterns and relationships in the data.

In addition to the omission and imputation strategies, other approaches may be considered when handling missing values. If the variable that has many missing values is deemed unimportant or can be represented using a proxy variable that does not have missing values, the variable may be excluded from the analysis. Finally, some analytics techniques such as decision trees (discussed in Chapter 10) are robust and can be applied to data sets even with the inclusion of missing values.

### HANDLING MISSING VALUES

There are two common strategies for dealing with missing values.

- The omission strategy recommends that observations with missing values be excluded from subsequent analysis.
- The imputation strategy recommends that the missing values be replaced with some reasonable imputed values. For numerical variables, it is common to use mean imputation. For categorical variables, it is common to impute the most predominant category.

In Example 2.2, we demonstrate how to handle missing values with the omission and imputation strategies using Excel and R.

## EXAMPLE 2.2

FILE
*Restaurant_Reviews*

Sarah Johnson, the manager of a local restaurant, has conducted a survey to gauge customers' perception about the eatery. Each customer rated the restaurant on its ambience, cleanliness, service, and food using a scale of 1 (lowest) to 7 (highest). Table 2.4 displays a portion of the survey data.

**TABLE 2.4** Restaurant Reviews

| RecordNum | Ambience | Cleanliness | Service | Food |
|---|---|---|---|---|
| 1 | 4 | 5 | 6 | 4 |
| 2 | 6 | 6 |  | 6 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 150 | 3 | 5 | 6 | 7 |

Sarah notices that there are a number of missing values in the survey. Use the ***Restaurant_Reviews*** data to first detect the missing values. Then use both omission and imputation strategies to handle the missing values.

**SOLUTION:**

**Using Excel**

**a.** Open the ***Restaurant_Reviews*** data file.

**b.** We use the Conditional Formatting feature in Excel to detect missing values. Select the data range B2:D151. Choose **Home > Conditional Formatting > New Rule. . .**. Select *Format only cells that contain* as the rule type. Select *Format only cells with: Blanks* as the rule description. Click on the *Format* button. Select the *Fill* tab and pick any color of your choice. Click **OK**. Click **OK** again. Now the cells with missing values are highlighted in your chosen color.

**c.** Here, we use the **COUNTBLANK** function to detect the number of missing values for each observation. Create a new column with the column heading "# of Missing Values" in column F. Enter the formula =COUNTBLANK(B2:E2) in cell F2. Fill the range F3:F151 with the formula in F2. This formula shows "0" when none of the variables in a record have a missing value. For the observations with missing values, the formula shows the number of variables with missing values. Verify that observations 2, 13, 26, and 100 each has one missing value, and observation 134 has two missing values.

**d.** To count the total number of missing values in the entire data set, enter the formula =SUM(F2:F151) in cell F152. The result shows that there is a total of six missing values.

**e.** To identify variables with missing values, enter the formula =COUNTBLANK(B2:B151) in cell B152. Fill the range C152:E152 with the formula in B152. The result shows that the Ambience (column B), Cleanliness (column C), and Food (column E) variables each has one missing value, and the Service variable (column D) has three missing values.

**f.** We start with the omission strategy by removing observations with missing values. Select rows 3, 14, 27, 101, and 135 while holding down the Ctrl button. Right-click on one of the selected rows and choose Delete from the pop-up menu. Verify that the five observations have been deleted and that the final data set includes 145 complete observations. The transformed data set is ready for data analysis.

**g.** We now use the simple mean imputation strategy to handle the missing values. Choose **Undo Delete** to restore the rows deleted in step f. Enter the headings “MeanAmbience”, “MeanCleanliness”, “MeanService”, and “MeanFood” in cells G1, H1, I1, and J1, respectively. To compute the mean ambience rating, enter the formula =AVERAGE(B2:B151) in cell G2. Excel’s **AVERAGE** function ignores the missing values when computing the mean values. Fill the range H2:J2 with the formula in G2 to compute the means for the Cleanliness, Service, and Food ratings. Verify that the mean ratings are 4.43, 5.46, 5.97, and 5.08 when they are rounded to two decimal places. To impute the missing Ambience ratings, select the range B2:B151. Choose **Home > Find & Select > Replace**. In the *Find and Replace* dialog box, leave the *Find what:* value empty and enter 4.43 as the *Replace with:* value. Click the *Replace All* button. Verify that the Ambience rating for observation 134 is now 4.43. Repeat the same procedure for the Cleanliness, Service, and Food ratings. Again, the transformed data set is ready for data analysis.

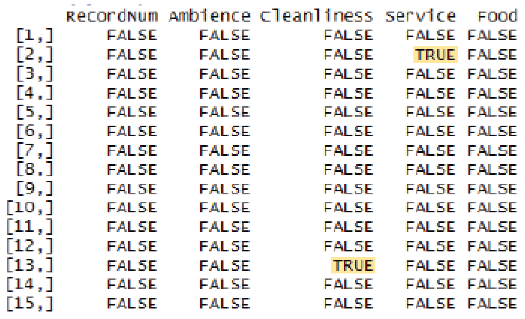
### Using R

**a.** Import the ***Restaurant_Reviews*** data file into a data frame (table) and label it myData.

**b.** To detect missing values in a data set, we use the **is.na** function. Recall from Example 2.1 that missing values are labelled as *NA* in R, and the **is.na** function returns TRUE if a missing value is detected and FALSE otherwise. Enter:

```
is.na(myData)
```

Figure 2.4 displays a portion of the output, where we have highlighted the missing values. The service rating for observation 2 and cleanliness rating for observation 13 are labeled TRUE because they are missing. The rest of the values in Figure 2.4 are labeled FALSE because they are not missing.



```
      RecordNum Ambience Cleanliness Service  Food
 [1,]     FALSE    FALSE       FALSE   FALSE FALSE
 [2,]     FALSE    FALSE       FALSE    TRUE FALSE
 [3,]     FALSE    FALSE       FALSE   FALSE FALSE
 [4,]     FALSE    FALSE       FALSE   FALSE FALSE
 [5,]     FALSE    FALSE       FALSE   FALSE FALSE
 [6,]     FALSE    FALSE       FALSE   FALSE FALSE
 [7,]     FALSE    FALSE       FALSE   FALSE FALSE
 [8,]     FALSE    FALSE       FALSE   FALSE FALSE
 [9,]     FALSE    FALSE       FALSE   FALSE FALSE
[10,]     FALSE    FALSE       FALSE   FALSE FALSE
[11,]     FALSE    FALSE       FALSE   FALSE FALSE
[12,]     FALSE    FALSE       FALSE   FALSE FALSE
[13,]     FALSE    FALSE        TRUE   FALSE FALSE
[14,]     FALSE    FALSE       FALSE   FALSE FALSE
[15,]     FALSE    FALSE       FALSE   FALSE FALSE
```

**FIGURE 2.4**
R output for detecting missing values

To detect missing values, say, in the Service variable, we enter:

```
is.na(myData$Service)
```

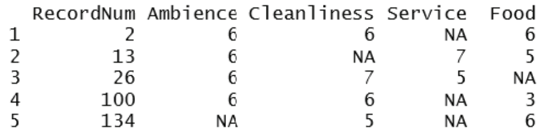
**c.** If we have a large data set, using the **is.na** function to detect missing values can be cumbersome. Alternatively, we can use the **complete.cases** function to identify the rows in the data frame or cases that are complete. Recall that leaving the second element, after the comma, blank will display all columns. Enter:

```
myData[complete.cases(myData), ]
```

**d.** Because our data are mostly complete, listing all the complete cases may not be convenient. Instead, we can use the *not* operator (**!** character) with the **complete.cases** function to identify observations with missing values. The **!** character in front of the **complete.cases** function identifies individual rows that are not complete (or cases with missing values). Enter:

```
myData[!complete.cases(myData), ]
```

Figure 2.5 shows observations 2, 13, 26, 100, and 134 have missing values.



```
  RecordNum Ambience Cleanliness Service  Food
1         2        6           6      NA     6
2        13        6          NA       7     5
3        26        6           7       5    NA
4       100        6           6      NA     3
5       134       NA           5      NA     6
```

**FIGURE 2.5** Observations with missing values

**e.** To implement the omission strategy, we use the **na.omit** function to remove observations with missing values and store the resulting data set in the omissionData data frame. The **View** function displays the updated data. Enter:

```
omissionData <- na.omit(myData)
View(omissionData)
```

R creates a new data frame, omissionData, that contains 145 complete cases. Verify that there are no missing values in the new data set. This data set is ready for data analysis.

**f.** To implement the simple mean imputation strategy, we start with the original data frame, **myData**. We then calculate the average value using the **mean** function. The option **na.rm = TRUE** ignores the missing values when calculating the average values. In order to compute the average values for the Ambience and Service variables, enter:

```
ambienceMean <- mean(myData$Ambience, na.rm = TRUE)
serviceMean <- mean(myData$Service, na.rm = TRUE)
```

Verify that the means for the Ambience and Service variables are 4.42953 and 5.965986, respectively.

**g.** To impute the missing values in the Ambience and Service variables, we again use the **is.na** function to identify the missing values and replace them with the means calculated in step f. Enter:

```
myData$Ambience[is.na(myData$Ambience)] <- ambienceMean
myData$Service[is.na(myData$Service)] <- serviceMean
```

Students are encouraged to calculate the mean and impute missing values in the Cleanliness and Food variables and inspect the resulting data set to make sure that the missing values have been replaced by the average ratings. The resulting data set is then ready for data analysis.

Another important data preparation task involves the treatment of extremely small or large values, referred to as outliers. In the presence of outliers, it is preferred to use the median instead of the mean to impute missing values. Both Excel and R allow easy imputation with the median computed by using the **MEDIAN(*data range*)** function in Excel or the **median** function in R. Refer to Chapter 3 for a detailed discussion on outliers and the median.

## Subsetting

The process of extracting portions of a data set that are relevant to the analysis is called **subsetting**. It is commonly used to pre-process the data prior to analysis. For example, a multinational company has sales data for its global operations, and it creates a subset of sales data by country and performs analysis accordingly. For time series data, which are data indexed in time order, we may choose to create subsets of recent observations and observations from the distant past in order to analyze them separately. Subsetting can also be used to eliminate unwanted data such as observations that contain missing values, low-quality data, or outliers. Sometimes, subsetting involves excluding variables instead of observations. For example, we might remove variables that are irrelevant to the problem, variables that contain redundant information (e.g., property value and property tax or employee's age and experience), or variables with excessive amounts of missing values.

SUBSETTING

Subsetting is the process of extracting parts of a data set that are of interest to the analytics professional.

Subsetting can also be performed as part of descriptive analytics that helps reveal insights in the data. For example, by subsetting student records into groups with various academic performance levels, we may be able to identify high-achieving students and relevant attributes that contribute to their success. Similarly, by comparing subsets of medical records with different treatment results, we may identify potential contributing factors of success in a treatment.

Table 2.5 shows important summary measures from two subsets of medical records of tuberculosis treatments administered in a developing country. Here, subsets 1 and 2 represent successful and unsuccessful treatments, respectively.

**TABLE 2.5** Summary Data of Tuberculosis Treatment

| Summary measures | Successful treatments | Unsuccessful treatments |
|---|---|---|
| % of Male patients | 64.3% | 12.7% |
| Average education Level | 3.8 | 2.1 |
| % of patients with Good Incomes | 92.7% | 28.1% |

Note that the sex, education level (1: lowest; 5: highest), and income of the patients differ considerably between the two subsets. Not surprisingly, male patients, especially those with higher education and income levels, have better success with tuberculosis treatment than female patients with lower education and income levels. This simple analysis highlights the importance of contributing factors in tuberculosis control efforts.

In Example 2.3, we demonstrate how to use subsetting functions in Excel and R to select or exclude variables and/or observations from the original data set.

## EXAMPLE 2.3

FILE
***Customers***

In the introductory case, Catherine Hill wants to gain a better understanding of Organic Food Superstore's customers who are college-educated millennials, born on or after 1/1/1982 and before 1/1/2000. She feels that sex, household size, income, total spending in 2021, total number of orders in the past 24 months, and channel through which the customer was acquired are useful for her to create a profile of these customers. Use Excel and R to first identify college-educated millenial customers in the ***Customers*** data file. Then, create subsets of female and male college-educated millenial customers. The synopsis that follows this example provides a summary of the results.

**SOLUTION:**

### Using Excel

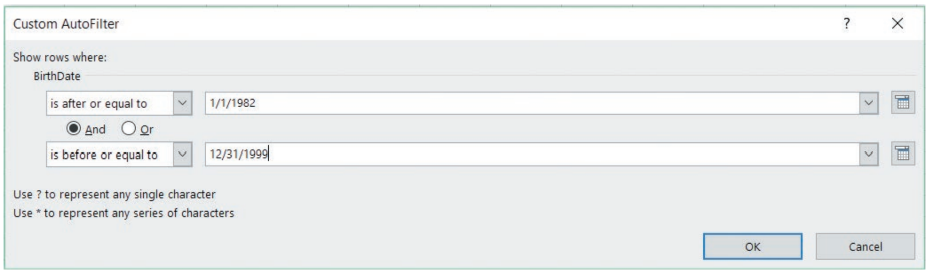
**a.** Open the ***Customers*** data file.

**b.** We first filter the data set to include only college-educated millennials. Select the data range A1:N201. From the menu choose **Home > Sort & Filter > Filter**. The column headings (A1:N1) will turn into drop-down boxes.

**c.** Click on the drop-down box in E1 (College). Uncheck *(Select all)*, then check the box next to *Yes*. Click **OK**. This step shows only those customers who have a college degree (Yes) by hiding those who don't (No) in the data set.

**d.** Click on the drop-down box in D1 (BirthDate). Select **Date filters > Between**. See Figure 2.6. In the *Custom AutoFilter* dialog box, enter 1/1/1982 next to the *is after or equal to* box or select the date from the calendar object. Select *And* and enter 12/31/1999 next to the *is before or equal to* box or select the date from the calendar object. Click **OK**. The data set now only displays college-educated millennials who were born between 1982 and 2000.

**FIGURE 2.6** Excel's AutoFilter dialog box



Show rows where:

BirthDate

is after or equal to 1/1/1982

● And ○ Or

is before or equal to 12/31/1999

Use ? to represent any single character

Use * to represent any series of characters

Microsoft Corporation

**e.** Select the entire filtered data that are left in the worksheet. Copy and paste the filtered data to a new worksheet. Verify that the new worksheet contains 59 observations of college-educated millennials. Rename the new worksheet as *College-Educated Millennials*.

**f.** We now exclude the variables that are not relevant to the current analysis. In the *College-Educated Millennials* worksheet, select cell A1 (CustID). From the menu choose **Home > Delete > Delete Sheet Columns** to remove the CustID column. Repeat this step for the Race, BirthDate, College, ZipCode, Spending2020, DaysSinceLast, and Satisfaction columns from the data set.

**g.** To subset the college-educated millennials data by sex, select column A. From the menu choose **Home > Sort & Filter > Sort A to Z**. If prompted, select *Expand the selection* in the *Sort Warning* dialog box and click *Sort*. The observations are now sorted by sex in alphabetic order. The female customer records are followed by male customer records.

**h.** Create two new worksheets and assign the worksheet names *Female* and *Male*. Copy and paste the female and male customer records, including the column headings, to the new *Female* and *Male* worksheets, respectively. Table 2.6 shows a portion of the results.

**TABLE 2.6** College-Educated Millennial Customers

**a) Female College-Educated Millennials**

| Sex | HouseholdSize | Income | Spending2021 | NumOfOrders | Channel |
|---|---|---|---|---|---|
| Female | 5 | 53000 | 241 | 3 | SM |
| Female | 3 | 84000 | 153 | 2 | Web |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Female | 1 | 52000 | 586 | 13 | Referral |

**b) Male College-Educated Millennials**

| Sex | HouseholdSize | Income | Spending2021 | NumOfOrders | Channel |
|---|---|---|---|---|---|
| Male | 5 | 94000 | 843 | 12 | TV |
| Male | 1 | 97000 | 1028 | 17 | Web |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Male | 5 | 102000 | 926 | 10 | SM |

### Using R

**a.** Import the ***Customers*** data file into a data frame (table) and label it myData.

**b.** To select college-educated millennials, we first select all customers with a college degree. Recall that the double equal sign (==) is used to check whether the College value is "Yes." Enter:

```
college <- myData[myData$College=='Yes', ]
```

**c.** We now use the Birthdate variable to select the millennials who were born between 1982 and 2000. R usually imports the date values as text characters, and, therefore, we first need to convert the BirthDate variable into the date data type using the **as.Date** function. The option format = "%m/%d/%Y" reads the values in the BirthDate variable into the mm/dd/yyyy format. For example, in order for R to read dates into the format 01/13/1990, enter:

```
college$BirthDate <- as.Date(college$BirthDate, format =
"%m/%d/%Y")
```

Other common date formats include "%Y-%m-%d", "%b %d, %Y", and "%B %d, %Y" that will read dates specified as 1990-01-13, Jan 13, 1990, and January 13, 1990, respectively.

**d.** We also use the **as.Date** function to specify the cutoff dates, January 1, 1982, and December 31, 1999, before using them as selection criteria for selecting the millennials in our data. Enter:

```
cutoffdate1 <- as.Date("01/01/1982", format = "%m/%d/%Y")
cutoffdate2 <- as.Date("12/31/1999", format = "%m/%d/%Y")
millenials <- college[college$BirthDate >= cutoffdate1 &
college$BirthDate <= cutoffdate2, ]
```

Verify that the millennials data frame contains 59 college-educated millennials.

**e.** To include only the Sex, HouseholdSize, Income, Spending2021, NumOfOrders, and Channel variables in the millenials data frame, we specify the column indices of these variables using the **c** function. Enter:

```
subset1 <- millenials[ , c(2,6,8,10,11,14)]
```

Alternately, we can create a new data frame by specifying the names of the variables to include. Enter:

```
subset2 <- millenials[ , c("Sex", "HouseholdSize", "Income",
"Spending2021", "NumOfOrders", "Channel")]
```

Note that subset1 and subset2 data are identical.

**f.** R imports non-numerical variables such as Sex and Channel as text characters. Before further subsetting and examining the data, we convert Sex and Channel into categorical variables (called factors in R) by using the **as.factor** function. Enter:

```
subset1$Sex <- as.factor(subset1$Sex)
subset1$Channel <- as.factor(subset1$Channel)
```

To verify that the Channel variable has been converted into a factor or a categorical variable, enter:

```
is.factor(subset1$Channel)
```

This command returns TRUE if the variable is a factor, and FALSE otherwise.

**g.** To create two subsets of data based on Sex, we use the **split** function. Enter:

```
sex <- split(subset1, subset1$Sex)
```

The sex data frame contains two subsets: Female and Male. We can now access and view the Female and Male subsets. Enter:

```
sex$Female
sex$Male
View(sex$Female)
View(sex$Male)
```

Verify that there are 21 female college-educated millennials and 38 male college-educated millennials. Your results should be similar to Table 2.6.

**h.** In some situations, we might simply want to subset data based on data ranges. For example, we use the following statement to subset data to include observations 1 to 50 and observations 101 to 200. Enter:

```
dataRanges <- myData[c(1:50, 101:200),]
```

## SYNOPSIS OF INTRODUCTORY CASE

hbpictures/Shutterstock

Catherine Hill has been assigned to help market Organic Food Superstore's new line of Asian-inspired meals. In order to understand the potential target market for this product, Catherine subsetted the data that contain a representative sample of the company's customers to include only college-educated millennials. She also partitioned the data set into two subsets based on sex to compare the profiles of female and male college-educated millennials.

The differences between the female and male customers have given Catherine some ideas for the marketing campaign that she is about to design. For example, the data show that an overwhelming portion of the male customers were acquired through social media ads, while female customers tend to be enticed by web ads or referrals. She plans to design and run a series of social media ads about the new product line with content that targets male customers. For female customers, Catherine wants to focus her marketing efforts on web banner ads and the company's referral program.

Furthermore, as the male customers seem to place more frequent but smaller orders than female customers do, Catherine plans to work with her marketing team to develop some cross-sell and upsell strategies that target male customers. Given the fact that the company's male college-educated millennial customers tend to be high-income earners, Catherine is confident that with the right message and product offerings, her marketing team will be able to develop strategies for increasing the total spending of these customers.

## EXERCISES 2.3

### Mechanics

16. The following table contains three variables and five observations with some missing values.

| $x_1$ | $x_2$ | $x_3$ |
|---|---|---|
| 248 | 3.5 | |
| 124 | 3.8 | 55 |
| 150 | | 74 |
| 196 | 4.5 | 32 |
| | 6.2 | 63 |

a. Handle the missing values using the omission strategy. How many observations remain in the data set and have complete cases?

b. Handle the missing values using the simple mean imputation strategy. How many missing values are replaced? What are the means of $x_1$, $x_2$, and $x_3$?

17. FILE *Exercise_2.17.* The accompanying data set contains four variables, $x_1$, $x_2$, $x_3$, and $x_4$.

a. Subset the data set to include only observations that have a date on or after May 1, 1975, for $x_3$. How many observations are in the subset data?

b. Split the data set based on the binary values for $x_4$. What are the average values for $x_1$ for the two subsets?

18. FILE *Exercise_2.18.* The accompanying data set contains five variables, $x_1$, $x_2$, $x_3$, $x_4$, and $x_5$.

a. Subset the data set to include only $x_2$, $x_3$, and $x_4$. How many missing values are there in the three remaining variables?

b. Remove all observations that have "Own" as the value for $x_2$ and those that have values lower than 150 for $x_3$. How many observations remain in the data set? What are the average values for $x_3$ and $x_4$?

19. FILE *Exercise_2.19.* The accompanying data set contains five variables, $x_1$, $x_2$, $x_3$, $x_4$, and $x_5$. There are missing values in the data set.

a. Which variables have missing values?

b. Which observations have missing values?

c. How many missing values are in the data set?

d. Handle the missing values using the omission strategy. How many observations remain in the data set and have complete cases?

20. FILE *Exercise_2.20.* The accompanying data set contains five variables, $x_1$, $x_2$, $x_3$, $x_4$, and $x_5$. There are missing values in the data set. Handle the missing values using the simple mean imputation strategy for numerical variables and the predominant category strategy for categorical variables.

a. How many missing values are there for each variable?

b. What are the values for imputing missing values in $x_1$, $x_2$, $x_3$, $x_4$, and $x_5$?

21. FILE ***Exercise_2.21.*** The accompanying data set contains five variables, $x_1$, $x_2$, $x_3$, $x_4$, and $x_5$.

a. Are there missing values for $x_1$? If so, impute the missing values using the mean value of $x_1$. After imputation, what is the mean of $x_1$?

b. Are there missing values for $x_2$? If so, impute the missing values using the mean value of $x_2$. After imputation, what is the mean of $x_2$?

c. If there are missing values in $x_4$, impute the missing values using the median value of $x_4$. [Hint: Use the **MEDIAN(*data range*)** function in Excel or the **median** function in R.] After imputation, what is the median of $x_4$?

22. FILE ***Exercise_2.22.*** The accompanying data set contains five variables, $x_1$, $x_2$, $x_3$, $x_4$, and $x_5$. There are missing values in the data set.

a. Which variables have missing values?

b. Which observations have missing values?

c. How many missing values are in the data set?

d. Handle the missing values using the omission strategy. How many observations remain in the data set and have complete cases?

23. FILE ***Exercise_2.23.*** The accompanying data set contains four variables, $x_1$, $x_2$, $x_3$, and $x_4$. There are missing values in the data set.

a. Subset the data set to include only $x_1$, $x_2$, and $x_3$.

b. Which variables have missing values?

c. Which observations have missing values?

d. How many missing values are in the data set?

e. Handle the missing values using the omission strategy. How many observations remain in the data set and have complete cases?

f. Split the data set based on the categories of $x_2$. How many observations are in each subset?

24. FILE ***Exercise_2.24.*** The accompanying data set contains seven variables, $x_1$, $x_2$, $x_3$, $x_4$, $x_5$, $x_6$, and $x_7$. There are missing values in the data set.

a. Remove variables $x_2$, $x_6$, and $x_7$ from the data set. Which of the remaining variables have missing values?

b. Which observations have missing values?

c. How many missing values are in the data set?

d. If there are missing values in $x_1$, replace the missing values with "Unknown." How many missing values were replaced?

e. Handle the missing values for numerical variables using the imputation strategy. If there are missing values in $x_3$, impute the missing values using the mean value of $x_3$. If there are missing values in $x_4$, impute the missing values using the median value of $x_4$. If there are missing values in $x_5$, impute the missing values using the mean value of $x_5$. What are the average values of $x_3$, $x_4$, and $x_5$ after imputation?

f. Remove observations that have the value "F" for $x_1$ and values lower than 1,020 for $x_4$. How many observations remain in the data set?

## Applications

25. FILE ***Population.*** The US Census Bureau records the population for the 50 states each year. The accompanying data file contains these data for the years 2010 to 2018.

a. Create two subsets of the state population data: one with 2018 population greater than or equal to 5 million and one with 2018 population less than 5 million. How many observations are in each subset?

b. In the subset of states with 5 million or more people, remove the states with over 10 million people. How many states were removed?

26. FILE ***Travel_Plan.*** Jerry Stevenson is the manager of a travel agency. He wants to build a model that can predict whether or not a customer will travel within the next year. He has compiled the accompanying data file that contains the following variables: whether the individual has a college degree (College), whether the individual has a credit card (CreditCard), annual household spending on food (FoodSpend in \$), annual income (Income in \$), and whether the customer has plans to travel within the next year (TravelPlan: 1 = have travel plans; 0 = do not have travel plans).

a. Are there any missing values in the data set? If there are, which variables have missing values? How many missing values are there in the data set?

b. Use the omission strategy to handle missing values. How many observations are removed due to missing values?

27. FILE ***Travel_Plan.*** Refer to the previous exercise for a description of the problem and data set. Based on his past experience, Jerry knows that whether the individual has a credit card or not has nothing to do with his or her travel plans and would like to remove this variable. Remove the variable CreditCard from the data set.

a. To better understand his customers with high incomes, Jerry wants to create a subset of the data that only includes customers with annual incomes higher than \$75,000 and who plan to travel within the next year. Subset the data to build the list of customers who meet these criteria. How many observations are in this subset?

b. Return to the original data set. Use the imputation strategy to handle missing values. If there are missing values for the FoodSpend variable, impute the missing values using the mean of the variable. If there are missing values for the Income variable, impute the missing values using the median of the variable. What are the average values of FoodSpend and Income after imputation?

28. FILE ***Football_Players.*** Denise Lau is an avid football fan and religiously follows every game of the National Football League. During the season, she meticulously keeps a record of how each quarterback has played throughout the season. Denise is making a presentation at the local NFL fan club about these quarterbacks. The accompanying data file contains the data that Denise has recorded, with the following variables: the player's name (Player), team's name (Team), completed passes (Comp), attempted passes (Att), completion percentage (Pct), total yards thrown (Yds), average yards per attempt (Avg), yards thrown per game (Yds/G), number of touch downs (TD), and number of interceptions (Int).

   a. Are there any missing values in the data set? If there are, which variables have missing values? Which observations have missing values? How many missing values are there in the data set?

   b. Use the omission strategy to handle missing values. How many observations are removed due to missing values?

29. FILE ***Football_Players.*** Refer to the previous exercise for a description of the data set. Denise feels that, for her presentation, it would remove some biases if the player names and team names are suppressed. Remove these variables from the data set.

   a. Denise also wants to remove outlier cases where the players have less than five touchdowns or more than 20 interceptions. Remove these observations from the data set. How many observations were removed from the data?

   b. Return to the original data set. Use the imputation strategy to handle missing values. If there are missing values for Comp, Att, Pct, Yds, Avg, or Yds/G, impute the missing values using the mean of the variable. If there are missing values for TD or Int, impute the missing values using the median of the variable. What are the average values of Comp, Att, Pct, Yds, Avg, and Yds/G after imputation?

30. FILE ***Salaries.*** Ian Stevens is a human resource analyst working for the city of Seattle. He is performing a compensation analysis of city employees. The accompanying data set contains three variables: Department, Job Title, and Hourly Rate (in $). A few hourly rates are missing in the data.

   a. Split the data set into a number of subsets based on Department. How many subsets are created?

   b. Which subset contains missing values? How many missing values are in that data set?

   c. Use the imputation strategy to replace the missing values with the mean of the variable. What is the average hourly rate for each subset?

31. FILE ***Stocks.*** Investors usually consider a variety of information to make investment decisions. The accompanying data file contains a sample of large publicly traded corporations and their financial information. Relevant information includes stock price (Price), dividend as a percentage of share price (Dividend), price to earnings ratio (PE), earnings per share (EPS), book value, lowest and highest share prices within the past 52 weeks (52 wk low and 52 wk high), market value of the company's shares (Market cap), and earnings before interest, taxes, depreciation, and amortization (EBITDA in $ billions). As the price to earnings ratio is often considered a better assessment of stock valuation than stock price or earnings per share alone, the financial analyst would like to remove Price and EPS from the data set. Remove these variables from the data set.

   a. The financial analyst is most interested in companies with a higher book value than market cap. Remove all observations that have a lower book value than market cap. How many observations are left in the data set?

   b. The financial analyst also wants to remove companies with price to earnings ratios that are higher than 25. How many observations are left in the data set after those companies are removed?

32. FILE ***Stocks.*** Refer to the previous exercise for a description of the problem and data set.

   a. Are there any missing values in the data set? If there are, which variables have missing values? Which observations have missing values? How many missing values are there in the data set?

   b. Use the omission strategy to handle missing values. How many observations are removed due to missing values?

   c. Return to the original data set. If there are missing values for Price, Dividend, Book Value, 52 wk low, or 52 wk high, replace the missing value with "M," which stands for "Missing." If there are missing values for PE, EPS, Market cap, or EBITDA, use the imputation strategy to replace the missing values with the median of the variable. What are the imputed values for the variables with missing data?

33. FILE ***Longitudinal_Partial.*** The accompanying data file contains the data from the National Longitudinal Survey (NLS), which follows over 12,000 individuals in the United States over time. Variables in this analysis include the following information on individuals: Urban (1 if lives in urban area, 0 otherwise), Siblings (number of siblings), White (1 if white, 0 otherwise), Christian (1 if Christian, 0 otherwise), FamilySize, Height, Weight (in pounds), and Income (in $). Remove Height and Weight from the data set.

   a. How many individuals live in an urban area and have more than three members in their households?

   b. Subset the data to create two data sets, one that includes observations who live in an urban area and have incomes over $40,000 and one that includes the rest of the observations. How many observations are in each subset?

34. FILE ***Longitudinal_Partial.*** Refer to the previous exercise for a description of the data set.

    a. Are there any missing values in the data set? If there are, which variables have missing values? Which observations have missing values? How many missing values are there in the data set?

    b. Use the omission strategy to handle missing values. How many observations are removed due to missing values?

    c. Return to the original data set. If there are missing values for Height or Weight, use the imputation strategy to replace the missing values with the mean of the variable. If there are missing values for Siblings, FamilySize, or Income, replace the missing values with the median of the variable. What are the imputed values of the variables with missing data?

## 2.4 TRANSFORMING NuMeRICAL DATA

LO 2.4

Transform numerical variables.

**Data transformation** is the process of converting data from one format or structure to another. It is performed to meet the requirements of statistical and data mining techniques used for the analysis. Examples of transforming numerical data include transforming an individual's date of birth to age, combining height and weight to create body mass index, calculating percentages, or converting values to natural logarithms. Sometimes it makes sense to group a vast range of numerical values into a small number of "bins." For example, we might want to arrange numerical ages into age intervals such as 20 to 39, 40 to 59, and 60 to 80, or convert dates into seasons such as fall, winter, spring, and summer.

DATA TRANSFORMATION

Data transformation is the process of converting data from one format or structure to another.

In this section, we describe techniques for transforming numerical variables into categorical values (or binning) and mathematical transformations of numerical variables. Section 2.5 explores techniques for transforming categorical variables.

### Binning

**Binning** is the process of transforming numerical variables into categorical variables by grouping the numerical values into a small number of groups or bins. It is important that the bins are consecutive and nonoverlapping so that each numerical value falls into one, and only one, bin. For example, we might want to transform income values into three groups: below $50,000, between $50,000 and $100,000, and above $100,000. The three income groups can be labeled as "low," "medium," and "high" or "1," "2," and "3." Binning can be an effective way to reduce noise in the data if we believe that all observations in the same bin tend to behave the same way. For example, the transformation of the income values into three groups makes sense when we are more interested in a person's earning power (low, medium, or high) rather than the actual income value.

BINNING

Binning is a common data transformation technique that converts numerical variables into categorical variables by grouping the numerical values into a small number of bins.

An important advantage of binning is that it reduces the noise in the data often due to minor observation errors. For example, with binning, outliers in the data (e.g., individuals with extremely high income, perhaps recorded incorrectly) will be part of the last bin and, therefore, will not distort subsequent data analysis. Binning is also useful in categorizing observations and meeting the categorical data requirements of some data mining analytics techniques such as naïve Bayes (discussed in Chapter 12).

In addition to binning numerical values according to user-defined boundaries, bins are also often created to have equal intervals. For example, we can create bins that represent an interval of 10 degrees in temperature or 10 years in age. We can also create bins of equal counts, where individual bins have the same number of observations. For example, by binning a class of 200 students into 10 equal-size groups based on their grades, we can find out the relative academic standing of the students. Students in the bin with the highest grades represent the top 10% of the class.

In Example 2.4, we demonstrate how to use Excel and R to create bins with equal counts, equal intervals, and user-defined intervals.

## EXAMPLE 2.4

FILE
*Customers*

In order to better understand her customers, Catherine Hill would like to perform the RFM analysis, a popular marketing technique used to identify high-value customers. RFM stands for recency, frequency, and monetary. The RFM ratings can be created from the DaysSinceLast (recency), NumOfOrders (frequency), and Spending2021 (monetary) variables.

Following the 80/20 business rule (i.e., 80% of your business comes from 20% of your best customers), for each of the three RFM variables, Catherine would like to bin customers into five equal-size groups, with 20% of the customers included in each group. Each group is also assigned a score from 1 to 5, with 5 being the highest. Customers with the RFM rating of 555 are considered the most valuable customers to the company.

In addition to the RFM binning, Catherine would like to bin the Income variable into five equal intervals. Finally, she would like to start a tiered membership system where different services and rewards are offered to customers depending on how much they spent in 2021. She would like to assign the bronze membership status to customers who spent less than $250, silver membership status to those who spent $250 or more but less than $1,000, and the gold membership status to those who spent $1,000 or more.

Use Excel and R to bin variables according to Catherine's specifications. Summarize the results.

**SOLUTION:**

**Using Excel**

**a.** Open the ***Customers*** data file.

**b.** To create the recency score, we need to first transform the variable DaysSinceLast to reverse the order of the data because the fewer the number of days since the last purchase, the greater is the recency score. Create a new column with the column heading DaysSinceLastReverse in column O. Enter the formula =L2*(−1) in cell O2. Verify that the value of cell O2 is −101. Fill the range O3:O201 with the formula in O2.

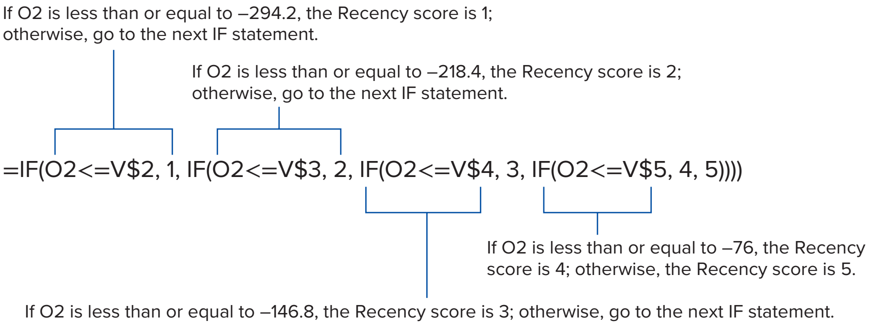
**c.** Because the bins represent the 20th, 40th, 60th, 80th, and 100th percentiles, we first use the **PERCENTILE** function to identify the range for each bin. Enter the headings RecencyRange, FrequencyRange, and MonetaryRange in cells V1, W1, and X1, respectively. The **PERCENTILE** function requires two inputs. The first input lists the cell range of the numeric observations, and the second input specifies the percentile using a decimal number between 0 and 1. For example, 0.20 represents the 20th percentile. Enter the formula =PERCENTILE(O2:O201, 0.20) in cell V2. Verify that the resulting value is −294.2. This means that 20% of the DaysSinceLastReverse observations are less than or equal to this value. Similarly, enter the formulas =PERCENTILE(O2:O201, 0.40), =PERCENTILE(O2:O201, 0.60), and =PERCENTILE(O2:O201, 0.80) in cells V3, V4, and V5, respectively, to identify the boundary values for the 40th, 60th, and 80th percentiles. Repeat this process for the FrequencyRange and MonetaryRange columns. Note that the percentile boundaries for the frequency can be found using NumOfOrders, which is in column K, and the boundaries for the monetary value can be found using Spending2021, which is in column J. Verify that the resulting boundary values match those in Table 2.7.

**TABLE 2.7** Percentile Boundary Values for Recency, Frequency, and Monetary Value

| | RecencyRange | FrequencyRange | MonetaryRange |
|---|---|---|---|
| 20th percentile | −294.2 | 4 | 300 |
| 40th percentile | −218.4 | 7 | 552.6 |
| 60th percentile | −146.8 | 11 | 765.2 |
| 80th percentile | −76 | 16 | 1044.4 |

**d.** Enter the column headings Recency, Frequency, and Monetary in cells P1, Q1, and R1. To bin the numeric values into Recency scores, we use the **IF** function. An **IF** function has the following structure: =IF(*a logical test resulting in TRUE or FALSE, result if the logical test is TRUE, result if the logical test is FALSE*). Because there are five Recency bins, we need to use multiple **IF** statements, or nested IFs. Figure 2.7 provides an illustration of how the nested **IF** function works in this example. Enter the formula =IF(O2<=V$2, 1, IF(O2<=V$3, 2, IF(O2<=V$4, 3, IF(O2<=V$5, 4, 5)))) in cell P2. Fill the range P3:P201 with the formula in P2. For the Frequency scores, enter the formula =IF(K2<=W$2, 1, IF(K2<=W$3, 2, IF(K2<=W$4, 3, IF(K2<=W$5, 4, 5)))) in cell Q2. Fill the range Q3:Q201 with the formula in Q2. For the Monetary scores, enter the formula =IF(J2<=X$2, 1, IF(J2<=X$3, 2, IF(J2<=X$4, 3, IF(J2<=X$5, 4, 5)))) in cell R2. Fill the range R3:R201 with the formula in R2. Verify that the Recency, Frequency, and Monetary scores for the first customer are 4, 1, and 1, respectively.

**FIGURE 2.7** Illustration of Excel's Nested IF Functions



If O2 is less than or equal to −294.2, the Recency score is 1; otherwise, go to the next IF statement.

If O2 is less than or equal to −218.4, the Recency score is 2; otherwise, go to the next IF statement.

=IF(O2<=V$2, 1, IF(O2<=V$3, 2, IF(O2<=V$4, 3, IF(O2<=V$5, 4, 5))))

If O2 is less than or equal to −76, the Recency score is 4; otherwise, the Recency score is 5.

If O2 is less than or equal to −146.8, the Recency score is 3; otherwise, go to the next IF statement.

It is important to note that the resulting bins may not include exactly the same number of observations due to the fact that multiple observations may hold the same values and that those values happen to be the boundary values for the bins.

Excel's **PERCENTILE** function generates the boundary values for the bins that would best achieve equal-size groups.

**e.** To create the RFM score for each customer, we create a new column with the column heading RFM in cell S1. Enter =CONCATENATE(P2, Q2, R2) in cell S2. The **CONCATENATE** function merges multiple text values into one cell. Recall that cells P2, Q2, and R2 represent the recency, frequency, and monetary indices, respectively, for the first customer. This formula creates a 3-digit RFM score for the first customer; verify that the first customer has an RFM score of 411. Fill the range S2:S201 with the formula in S2.

**f.** We now bin the Income variable into 5 groups with equal intervals. Enter the column heading IncomeRange in cell Y1. The interval of the bins can be found using the formula =(MAX(H2:H201)-MIN(H2:H201))/5, which results in 27,200. Excel's **MAX** and **MIN** functions retrieve the maximum and minimum values of a group of values, respectively. Dividing the range of the values, which is computed as the difference between the maximum and minimum values, by 5 produces the interval value for each bin. To obtain the four boundary values for the five equal-interval bins, enter the formulas =MIN(H2:H201)+(MAX(H2:H201)-MIN(H2:H201))/5, =MIN(H2:H201)+(MAX(H2:H201)-MIN(H2:H201))/5*2, =MIN(H2:H201)+(MAX(H2:H201)-MIN(H2:H201))/5*3, and =MIN(H2:H201)+(MAX(H2:H201)-MIN(H2:H201))/5*4 in cells Y2, Y3, Y4, and Y5, respectively. Verify that the boundary values are 58,200, 85,400, 112,600, and 139,800.

**g.** We now bin the income values using the IF function. Each bin includes values greater than the lower boundary value and up to and including the higher boundary value. Enter the column heading Binned_Income in cell T1 and the formula =IF(H2<=Y$2, 1, IF(H2<=Y$3, 2, IF(H2<=Y$4, 3, IF(H2<=Y$5, 4, 5)))) in cell T2. Fill the range T2:T201 with the formula in T2. Verify that the income bins for the first three customers are 1, 3, and 2.

**h.** To assign customers to the tiered membership system that Catherine designed, we have to first define tiered membership. In order to describe the range of spending for each membership tier, enter the values from Table 2.8 in cells Z1:AA4.

**TABLE 2.8** Lookup Table

| Column Z | Column AA |
|---|---|
| Spending | Membership |
| 0 | Bronze |
| 250 | Silver |
| 1000 | Gold |

**i.** Add the column heading Membership_Tier in cell U1. We use the **VLOOKUP** function to bin the spending values into user-defined categories, or membership tiers in this example. The **VLOOKUP** function has four inputs: (1) the lookup value (a member's spending in 2021 in column J); (2) a lookup or reference table (ranges of spending and corresponding membership tiers in cells Z1:AA4, without the column headings); (3) column number in the lookup table that contains the output (the output, Bronze, Silver, or Gold, is in the second column in the lookup table); and (4) whether we want to look up a value within a range (TRUE) or find an exact match (FALSE).

Enter =VLOOKUP(J2, $Z$2:$AA$4, 2, TRUE) in cell U2. This function assigns the first customer to a membership tier according to the spending in 2021. Note that the dollar signs in $Z$2:$AA$4 ensure that when the formula is copied to another cell in column U the references to cells Z2:AA4 remain unchanged. Fill the range U3:U201 with the formula in U2 to assign a membership tier to each customer. Verify that the membership tier for the first customer is Bronze.

### Using R

**a.** Import the ***Customers*** data file into a data frame and label it myData.

**b.** To create the recency score, we first transform the variable DaysSinceLast to reverse the order of the data because the fewer the number of days since the last purchase, the greater is the recency score. Create a new variable called DaysSinceLastReverse by multiplying DaysSinceLast by −1. We use the **as.numeric** function to ensure that the DaysSinceLastReverse variable is a numerical type. Enter:

```
myData$DaysSinceLastReverse <- as.numeric(myData$DaysSinceLast
* -1)
```

**c.** We now create five equal-sized bins for DaysSinceLastReverse (recency), NumOfOrders (frequency), and Spending2021 (monetary). As the bins represent the 20th, 40th, 60th, 80th, and 100th percentiles, we first use the **quantile** function to find the range for each bin and store and then view the ranges in an object called recencyBins. Enter:

```
recencyBins <- quantile(myData$DaysSinceLastReverse, probs=seq(0, 1,
by=0.20))
recencyBins
```

Figure 2.8 shows the ranges for the five equal-sized bins.

**FIGURE 2.8** Ranges for the recency bins

```
    0%      20%      40%      60%     80%    100%
−360.0   −294.2   −218.4   −146.8   −76.0    −6.0
```

We repeat similar commands for the number of orders (frequencyBins) and the Spending2021 (monetaryBins) variables. Enter:

```
frequencyBins <- quantile(myData$NumOfOrders, probs=seq(0, 1,
by=0.20))
monetaryBins <- quantile(myData$Spending2021, probs=seq(0, 1,
by=0.20))
```

Note that if we were to create 10 equal-sized bins, we would change the *probs* option to `probs=seq(0, 1, by=0.10)`.

**d.** We use the **cut** function to bin the data. The *breaks* option of the **cut** function specifies the ranges of the bins created in step c. The *labels* option assigns a label to each bin. The right=TRUE option ensures that the intervals are open on the left and closed on the right. The include.lowest=TRUE option, when used along with the right=TRUE option, includes the smallest value in the first bin. Enter:

```
myData$Recency <- cut(myData$DaysSinceLastReverse,
breaks=recencyBins,labels=c(“1”, “2”, “3”, “4”, “5”), include.
lowest=TRUE, right=TRUE)
myData$Frequency <- cut(myData$NumOfOrders, breaks=frequencyBins,
labels=c(“1”, “2”, “3”, “4”, “5”), include.lowest=TRUE,
right=TRUE)
```

```
myData$Monetary <- cut(myData$Spending2021, breaks=monetaryBins,
labels=c("1", "2", "3", "4", "5"), include.lowest=TRUE, right=TRUE)
```

The prior commands assign numbers 1 to 5 to the equal-sized bins for the three RFM variables. The RFM indices are stored in three new variables, Recency, Frequency, and Monetary. If we were to create and assign labels to 10 equal-sized bins, we would use `labels=c("1", "2", "3", "4", "5", "6", "7", "8", "9", "10")`. Other labels can also be assigned to the bins as appropriate.

e. To create the RFM score, we combine the three RFM indices using the **paste** function, and then use the **head** function to view the first few observations. Enter:

```
myData$RFM <- paste(myData$Recency, myData$Frequency,
myData$Monetary)
head(myData$RFM)
```

Verify that the first observation of the RFM variable is 411.

f. We now bin the Income variable into 5 groups with equal intervals using the **cut** function. The *breaks* option specifies 5 bins with equal intervals. We assign numbers 1 (lowest) to 5 (highest) to the bins and then use the **head** function to view the first few observations. Enter:

```
myData$BinnedIncome <- cut(myData$Income, breaks=5, labels=
c("1", "2", "3", "4", "5"), include.lowest=TRUE)
head(myData$BinnedIncome)
```

Verify that the first observation of the BinnedIncome variable is 1. To create a different number of bins, change the **breaks** value (e.g., `breaks = 3` will create three bins).

g. We use the **levels** and **cut** functions to display the ranges of the 5 intervals created in step f. Enter:

```
levels(cut(myData$Income, breaks=5))
```

Verify that the first interval is (3.09e+04,5.82e+04] or from $30,900 up to, and including, $58,200.

h. To find out the number of observations that belong to each bin, use the **table** function. Enter:

```
table(myData$BinnedIncome)
```

Verify that the first bin, or the lowest income category, has 67 customers.

i. To create the membership tiers or user-defined bins proposed by Catherine, we again use the **cut** function. Recall that customers who spent less than $250 are assigned to the Bronze membership, those who spent $250 or more but less than $1,000 receive the Silver membership, and those who spent $1,000 or more receive the Gold membership. We use the *breaks* option to specify user-defined ranges. The Inf keyword assigns any values above $1,000 to the last bin. We then use the **head** and **View** functions to view the output in various formats. Enter:

```
myData$MembershipTier <- cut(myData$Spending2021, breaks =
c(0, 250, 1000, Inf),labels = c("Bronze", "Silver", "Gold"))
head(myData$MembershipTier)
View(myData)
```

Verify that the membership tier for the first customer is Bronze.

**Summary**

Table 2.9 shows a portion of the ***Customers*** data that now includes variables that have been binned according to Catherine's specifications. The first customer has an RFM score of 411, which suggests that this customer purchased from Organic Food Superstore quite recently but made very few purchases during the last 24 months and spent very little money in 2021. Not surprisingly, this customer also has the least desirable bronze membership. In addition, in terms of income, the majority of customers belong to the first three bins; only nine customers have income in the two highest income groups.

**TABLE 2.9** Customers Data with Binned Variables

| CustID | . . . | RFM | Binned income | Membership tier |
|---|---|---|---|---|
| 1530016 | . . . | 411 | 1 | Bronze |
| 1531136 | . . . | 244 | 3 | Silver |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 1579979 | . . . | 434 | 3 | Silver |

Students are encouraged to experiment with other binning options with the ***Customers*** data file. For example, experiment with binning Income into two categories, High and Low, or redesign the membership program with four or five tiers. There are many real-life applications where numerical variables should be converted into categories, and the number of categories should be determined appropriately according to the context of the analysis. For example, to convert students' exam scores into categories, we might use two bins (Pass and Fail) or five bins (A, B, C, D, and F grades).

## Mathematical Transformations

As discussed earlier, data transformation is an important step in bringing out the information in the data set, which can then be used for further data analysis. In addition to binning, another common approach is to create new variables through mathematical transformations of existing variables. For example, to analyze diabetes risk, doctors and dieticians often focus on body mass index (BMI), which is calculated as weight in kilograms divided by squared height in meters, rather than focusing on either weight or height alone. Similarly, in order to analyze trend, we often transform raw data values into percentages.

Sometimes data on variables such as income, firm size, and house prices are highly skewed; skewness is discussed in Section 3.2 of Chapter 3. For example, according to a Federal Reserve report, the richest 1% of families in the U.S. controlled 30.8% of all households' wealth in the second quarter of 2020 (*Forbes*, October 8, 2020). The extremely high (or low) values of skewed variables significantly inflate (or deflate) the average for the entire data set, making it difficult to detect meaningful relationships with skewed variables. A popular mathematical transformation that reduces skewness in data is the natural logarithm transformation. Logarithm transformations are used extensively in regression analysis; a detailed discussion can be found in Chapter 8.

Another common data transformation involves calendar dates. Statistical software usually stores date values as numbers. For example, in R, date objects are stored as the number of days since January 1, 1970, using negative numbers for earlier dates. For example, January 31, 1970, has a value of 30, and December 15, 1969, has a value of −17. Excel implements a similar approach to store date values but uses a reference value of 1 for January 1, 1900. Transformation of date values is often performed to help bring useful information out of the data. A retail company might convert customers' birthdates into ages in order to examine the differences in purchase behaviors across age groups.

Similarly, by subtracting the airplane ticket booking date from the actual travel date, an airline carrier can identify last-minute travelers, who may behave very differently from early planners.

Sometimes transforming date values into seasons helps enrich the data set by creating relevant variables to support subsequent analyses. For example, by extracting and focusing on the months in which gym members first joined the health club, we may find that members who joined during the summer months are more interested in the aquatic exercise programs, whereas those who joined during the winter months are more interested in the strength-training programs. This insight can help fitness clubs adjust their marketing strategies based on seasonality.

Example 2.5 demonstrates how to use Excel and R to perform the following mathematical transformations: (1) compute the percentage difference between two values, (2) perform a logarithm transformation, and (3) extract information from date values.

## EXAMPLE 2.5

FILE
*Customers*

After a closer review of her customers, Catherine Hill feels that the difference and the percentage difference between a customer's 2020 and 2021 spending may be more useful to understanding the customer's spending patterns than the yearly spending values. Therefore, Catherine wants to generate two new variables that capture the year-to-year difference and the percentage difference in spending. She also notices that the income variable is highly skewed, with most customers' incomes falling between $40,000 and $100,000, with only a few very-high-income earners. She has been advised to transform the income variable into natural logarithms, which will reduce the skewness of the data.

Catherine would also like to convert customer birthdates into ages as of January 1, 2022, for exploring differences in purchase behaviors of customers across age groups. Finally, she would like to create a new variable that captures the birth month of the customers so that seasonal products can be marketed to these customers during their birth month.

Use Excel and R to transform variables according to Catherine's specifications.

**SOLUTION:**

**Using Excel**

**a.** Open the ***Customer*** data file.

**b.** Create the column heading SpendingDiff in cell O1. Enter the formula =J2 - I2 in cell O2. Verify that the resulting value in cell O2 is −46. Fill the range O3:O201 with the formula in O2.

**c.** Create the column heading PctSpendingDiff in cell P1. Enter the formula = (J2 – I2)/I2 in cell P2. In the drop-down menu of the Home menu tab, change from *General* to *Percentage* (%) with *Decimal places* equal to 2. Verify that the value in cell P2 is −16.03%. Fill the range P3:P201 with the formula in P2.

**d.** Create the column heading IncomeLn in cell R1. The **LN** function provides a natural logarithm transformation. Enter the formula =LN(H2) in cell R2. Verify that the value in cell R2 is 10.8780. Fill the range R3:R201 with the formula in R2.

**e.** Create the column heading Age in cell S1. The **YEARFRAC** function calculates the difference in years between two dates, and the **INT** function displays only the integer portion of the value and discards the decimal places. Enter the formula =INT(YEARFRAC(D2, "01-01-2022")) in cell S2. Verify that the customer's age as of January 1, 2022, is 35 years. Fill the range S3:S201 with the formula in S2.

**f.** Create the column heading BirthMonth in cell T1. The **MONTH** function extracts the month element from a date value. Enter the formula =MONTH(D2) in cell T2. Verify that the value in cell T2 is 12. Fill the range T3:T201 with the formula in T2.

Table 2.10 shows a portion of the data that includes the five transformed variables. Other useful Excel functions related to date values include **DAY**, **YEAR**, **WEEKDAY**, **TODAY**, and **NOW**. The **DAY** and **YEAR** functions extract the date and the year elements, respectively. The **WEEKDAY** function identifies the weekday as an integer value from 1 to 7 (1 for Sunday and 7 for Saturday). Finally, =TODAY() and =NOW() return the values of the current date and current time, respectively.

**TABLE 2.10** Five Transformed Variables

| CustID | . . . | SpendingDiff | PctSpendingDiff | IncomeLn | Age | BirthMonth |
|---|---|---|---|---|---|---|
| 1530016 | . . . | −46 | −16.03% | 10.8780 | 35 | 12 |
| 1531136 | . . . | −384 | −31.30% | 11.4511 | 28 | 5 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 1579979 | . . . | −154 | −14.26% | 11.5327 | 22 | 7 |

### Using R

**a.** Import the ***Customer*** data file into a data frame (table) and label it myData.

**b.** We find the spending difference, and then use the **head** function to view the first few observations. Enter:

```
myData$SpendingDiff <- myData$Spending2021 - myData$Spending2020
head(myData$SpendingDiff)
```

Verify that the first observation of the SpendingDiff variable is −46.

**c.** We create the percentage spending difference and round it to two decimal places using the **round** function. We then place the “%” sign using the **paste** function and use the **head** function to view the first few observations. Enter:

```
myData$PctSpendingDiff <- round((myData$SpendingDiff /
myData$Spending2020)*100, digits = 2)
myData$PctSpendingDiff <- paste(myData$PctSpendingDiff, “%”)
head(myData$PctSpendingDiff)
```

Verify that the first observation of the PctSpendingDiff variable is −16.03%.

**d.** We use the **log** function for the natural logarithm transformation, and then use the **head** function to view the first few observations. Enter:

```
myData$IncomeLn <- log(myData$Income)
head(myData$IncomeLn)
```

Verify that the first observation of the IncomeLn variable is 10.87805. The IncomeLn values are slightly different from those in Table 2.10 because Table 2.10 is formatted to show only four decimal places. For the base 10 logarithm transformation, use the **log10** function in place of the **log** function.

**e.** To calculate a customer’s age as of January 1, 2022, we first need to convert the Birthdate variable into the date values and create a new variable for the January 1, 2022, date. Enter:

```
myData$BirthDate <- as.Date(myData$BirthDate, format = "%m/%d/%Y")
endDate <- as.Date("01/01/2022", format = "%m/%d/%Y")
```

**f.** We use the **difftime** function to find out the number of days between the customer's birthdate and January 1, 2022. By dividing the difference in days by 365.25, we account for the leap years (by using 365.25 instead of 365) and obtain the difference in years. We use the **as.numeric** function to ensure that the Age variable has a numerical type. Finally, we use the **floor** function to remove the decimal places so that the age of a customer is an integer and the **head** function to view the first few observations. Enter:

```
myData$Age <- difftime(endDate, myData$BirthDate)/365.25
myData$Age <- as.numeric(myData$Age)
myData$Age <- floor(myData$Age)
head(myData$Age)
```

Verify that the first customer's age as of January 1, 2022, is 35 years.

**g.** We use the **months** function to extract the month name from the Birthdate variable, the **match** function to convert month names (January to December) to numbers (1 to 12), and the **head** function to view the first few observations. Enter:

```
myData$BirthMonth <- months(myData$BirthDate)
myData$BirthMonth <- match(myData$BirthMonth, month.name)
head(myData$BirthMonth)
```

Verify that the first customer's birthday is in month 12 (December).

**h.** We use the **View** function to display a spreadsheet-style data. The output should be consistent with Table 2.10. Enter:

```
View(myData)
```

Other useful date-related functions include **weekdays** and **format**. The **weekdays** function returns the day of the week; for example > `weekdays(as.Date("2000-12-25"))` returns "Monday". The **format** function returns the specified element of a date value; for example, > `format(as.Date("2000-12-25"), "%Y")` returns the year element "2000". Besides the "%Y" parameter, "%d" and "%m" specify the date and month elements. The `Sys.Date()` and `Sys.time()` functions return the current date and time values, respectively.

Another common transformation for numerical data is rescaling, which is performed when the variables in a data set are measured using different scales. For example, annual income measured in dollars typically ranges from thousands to millions, whereas the number of children typically contains values in low single digits. The variability in measurement scales can place undue influence on larger-scale variables, resulting in inaccurate outcomes. Therefore, it is commonplace to rescale the data using either standardization or normalization, especially in data mining techniques; a detailed discussion on such transformations can be found in Chapter 11.

## EXERCISES 2.4

### Mechanics

35. FILE ***Exercise_2.35.*** The accompanying data file contains three variables, $x_1$, $x_2$, and $x_3$, and six observations.

    a. Bin the values of $x_1$ into two equal-size groups. Label the groups with numbers 1 ( lower values) and 2 (higher values). What is the average value of $x_1$ for group 1? (Hint: Sort the data by group number before calculating the average.)

    b. Bin the values of $x_2$ into three equal interval groups. Label the groups with numbers 1 ( lowest values) to 3 (highest values). How many observations are assigned to group 1?

    c. Bin the values of $x_3$ into the following two groups: $\leq 50$ and $> 50$. Label the groups with numbers 1 ( lower values) and 2 (higher values). How many observations are assigned to group 2?

36. FILE ***Exercise_2.36.*** The accompanying data file contains three variables, $x_1$, $x_2$, and $x_3$.

a. Bin the values of $x_1$ into three equal-size groups. Label the groups with numbers 1 (lowest values) to 3 (highest values). How many observations are assigned to group 1?

b. Bin the values of $x_2$ into three equal-interval groups. Label the groups with numbers 1 (lowest values) to 3 (highest values). How many observations are assigned to group 2?

c. Bin the values of $x_3$ into the following three groups: $< 50{,}000$, between 50,000 and 100,000, and $> 100{,}000$. Label the groups with numbers 1 (lowest values) to 3 (highest values). How many observations are assigned to group 1?

37. FILE ***Exercise_2.37.*** The accompanying data file contains three variables, $x_1$, $x_2$, and $x_3$.

a. Bin the values of $x_1$ into three equal-size groups. Label the groups with "low" (lowest values), "medium," and "high" (highest values). How many observations are assigned to group medium?

b. Bin the values of $x_2$ into three equal-interval groups. Label the groups with "low" (lowest values), "medium," and "high" (highest values). How many observations are assigned to group high?

c. Bin the values of $x_3$ into the following three groups: $< 20$, between 20 and 30, and $> 30$. Label the groups with "low" (lowest values), "medium," and "high" (highest values). How many observations are assigned to group low?

38. FILE ***Exercise_2.38.*** The accompanying data file contains three variables, $x_1$, $x_2$, and $x_3$.

a. Bin the values of $x_1$, $x_2$, and $x_3$ into five equal-size groups. Label the groups with numbers 1 (lowest) to 5 (highest).

b. Combine the group labels of $x_1$, $x_2$, and $x_3$ to create a score like the RFM score described in Example 2.4. How many observations have the score "431"? How many observations have the score "222"?

39. The following table contains two variables and five observations.

| $x_1$ | $x_2$ |
|---|---|
| 248 | 350 |
| 124 | 148 |
| 150 | 130 |
| 196 | 145 |
| 240 | 180 |

a. Create a new variable called "Sum" that contains the sum of the values of $x_1$ and $x_2$ for each observation. What is the average value of Sum?

b. Create a new variable called "Difference" that contains the absolute difference between the values of $x_1$ and $x_2$ for each observation. What is the average value of Difference?

40. FILE ***Exercise_2.40.*** The accompanying data file contains three variables, $x_1$, $x_2$, and $x_3$.

a. Create a new variable called "Difference" that contains the difference between the values of $x_1$ and $x_2$ for each observation (i.e., $x_2 - x_1$). What is the average difference?

b. Create a new variable called "PercentDifference" that contains the percent difference between the values of $x_1$ and $x_2$ for each observation [i.e., $(x_2 - x_1)/x_1$]. What is the average percent difference?

c. Create a new variable called "Log" that contains the natural logarithms for $x_3$. What is the average logarithm value?

41. FILE ***Exercise_2.41.*** The accompanying data file contains three variables, $x_1$, $x_2$, and $x_3$.

a. Create a new variable called "Difference" that contains the difference between the values of $x_1$ and $x_2$ for each observation (i.e., $x_2 - x_1$). What is the average difference?

b. Create a new variable called "PercentDifference" that contains the percent difference between the values of $x_1$ and $x_2$ for each observation [i.e., $(x_2 - x_1)/x_1$]. What is the average percent difference?

c. Create a new variable called "Log" that contains the natural logarithms for $x_3$. Bin the logarithm values into five equal-interval groups. Label the groups using numbers 1 (lowest values) to 5 (highest values). How many observations are in group 2?

42. FILE ***Exercise_2.42.*** The accompanying data file contains two variables, *Date1* and *Date2*.

a. Create a new variable called "DifferenceInYear" that contains the difference between *Date1* and *Date2* in year for each observation. What is the average difference in year? (Hint: Use the **YEARFRAC** function if you are using Excel to complete this problem.)

b. Create a new variable called "Month" that contains the month values extracted from *Date1*. What is the average month value?

c. Bin the month values in four equal-interval groups. Label the groups using numbers 1 (lowest values) to four (highest values). Which group has the highest number of observations?

## Application

43. FILE ***Population.*** The U.S. Census Bureau records the population for the 50 states each year. The accompanying data file contains these data for the years 2010 to 2018.

a. Bin the 2017 population values into four equal-size groups. Label the groups using numbers 1 (lowest values) to 4 (highest values). How many states are assigned to group 4?

b. Bin the 2018 population values into four equal-interval groups. Label the groups using numbers 1 (lowest values) to 4 (highest values). How many states are assigned to group 2? Compare the groups in parts a and b. Which states are in higher groups for the 2018 population than for the 2017 population?

c. Bin the 2018 population values into the following three groups: < 1,000,000, between 1,000,000 and 5,000,000, and > 5,000,000. Label the groups using numbers 1 ( lowest values) to 3 (highest values). How many observations are assigned to group 2?

44. **FILE** ***Population.*** Refer to the previous exercise for a description of the data set.

a. Create a new variable called "Difference" that contains the difference between the 2018 population and the 2017 population for each state (i.e., 2018 population − 2017 population). What is the average difference?

b. Create a new variable called "PercentDifference" that contains the percent difference between the 2017 and 2018 population values for the states [i.e., (2018 population − 2017 population)/2017 population]. What is the average percent difference?

c. Create a new variable called "Log" that contains the natural logarithms for the 2018 population values for the states. Bin the logarithm values into five equal-interval groups. Label the groups using numbers 1 ( lowest values) to 5 (highest values). How many observations are in group 2?

d. Create a new variable called "SquareRoot" that contains the square root of the 2018 population values for the states. Bin the square root values into five equal-interval groups. Label the groups using numbers 1 ( lowest values) to 5 (highest values). How many observations are in group 2?

e. Compare the groups in parts c and d. Are the groupings the same or different?

45. **FILE** ***Credit_Cards.*** Greg Metcalf works for a national credit card company, and he is performing a customer value analysis on a subset of credit card customers. In order to perform the RFM analysis on the customers, Greg has compiled the accompanying data file that contains the dates of the last transaction (LastTransactionDate), total number of transactions in the past two years (Frequency), and total spending during the past two years (Spending).

a. Greg wants to calculate the number of days between January 1, 2022, and the last transaction date. Create a new variable "DaysSinceLast" that contains the number of days since the last transaction. (Hint: Use the **DATEDIF** function if you are using Excel to complete this problem.) What is the average number of days since the last purchase for all the customers?

b. Create the RFM scores for each customer. How many customers have an RFM score of 555? What is their average spending?

c. Create a new variable called "LogSpending" that contains the natural logarithms for the total spending during the past two years. Bin the logarithm values into five equal-interval groups. Label the groups using numbers 1 ( lowest values) to 5 (highest values). How many observations are in group 2?

d. Create a new variable called "AverageOrderSize" that contains the average spending per order. This is calculated by dividing total spending (Spending) by total number of transactions (Frequency) in the past two years. Bin the values of AverageOrderSize into five equal-interval groups. Label the groups using numbers 1 ( lowest values) to 5 (highest values). How many observations are in group 2?

e. Compare the groups in parts c and d. Are the groupings the same or different?

46. **FILE** ***Game_Players.*** TurboX is an online video game company that makes three types of video games: action, role play, and sports. It is interested in understanding its millennial customers. By combining the data from its customer database and a customer survey, TurboX compiled the accompanying data file that has the following variables: the player's satisfaction with the online game purchase experience (Satisfaction), the enjoyment level of the game played (Enjoyment), whether the player will recommend the game to others (Recommend), which type of game the player played (Type), total spending on games last year (SpendingLastYear), total spending on games this year (SpendingThisYear), and the date of birth of the player (BirthDate).

a. Bin the total spending on games last year into four equal-size groups. Label the groups using numbers 1 ( lowest values) to 4 (highest values). How many customers are assigned to group 4?

b. Bin the total spending on games this year into four equal-interval groups. Label the groups using numbers 1 (lowest values) to 4 (highest values). How many customers are assigned to group 3?

c. Bin the total spending on games this year into the following three groups: < 250, between 250 and 500, and > 500. Label the groups using numbers 1 ( lowest values) to 3 (highest values). How many observations are assigned to group 2?

d. Create a new variable called "Difference" that contains the difference between this year's and last year's spending on games for the players (i.e., SpendingThisYear − SpendingLastYear). What is the average difference?

e. Create a new variable called "PercentDifference" that contains the percent difference between this year's and last year's spending on games for the players [i.e., (SpendingThisYear − SpendingLastYear)/SpendingLastYear]. What is the average percent difference?

f. Create a new variable "Age" that contains the players' ages as of January 1, 2022. What is the average age of the players?

g. Create a new variable "BirthMonth" that contains the players' birth month extracted from their dates of birth. Which month is the most frequent birth month?

47. FILE ***Engineers.*** Erin Thomas, an HR manager of an engineering firm, wants to perform an analysis on the data about the company's engineers. The variables included in the data are date of birth (BirthDate), personality type according to the Myers-Briggs Personality assessment (Personality), annual salary (Salary), level of the position (Level), and number of professional certificates achieved (Certificates).

a. Create a new variable "Age" that contains the engineers' ages as of January 1, 2022. What is the average age of the engineers?

b. Bin the age values into three equal-size groups. Label the groups using numbers 1 ( lowest age values) to 3 (highest age values). How many observations are in group 3?

c. Bin the annual salary values into four equal interval groups. Label the groups using numbers 1 (lowest salary values) to 4 (highest salary values). How many engineers are assigned to group 4?

d. Bin the number of professional certificates achieved into the following three groups: < 2, between 2 and 4, and over 4. Label the groups "Low," "Medium," and "High." How many engineers are in the "High" group?

48. FILE ***Patients.*** Jerry Stevenson is the manager of a medical clinic in Scottsdale, AZ. He wants to analyze patient data to identify high-risk patients for cardiovascular diseases. From medical literature, he learned that the risk of cardiovascular diseases is influenced by a patient's age, body mass index (BMI), amount of exercise, race, and education level. Jerry has compiled the accompanying data file with the following variables for his clinic's patients: race (Race), education level (Education), body weight in kilograms (Weight), height in meters (Height), date of birth (BirthDate), and number of minutes of exercise per week (Exercise).

a. Create a new variable called "BMI" that contains the body mass index of the patients. BMI is calculated as weight in kilograms/(height in meters)$^2$. What is the average BMI of the patients?

b. Create a new variable "Age" that contains the patients' ages as of January 1, 2022. What is the average age of the patients?

c. Bin the patients' ages into five equal-size groups. Label the groups using numbers 1 ( youngest) to 5 (oldest). How many patients are in group 4?

d. Bin the patients' total minutes of exercise per week into five equal-size groups. Label the groups using numbers 1 ( highest values) to 5 (lowest values). How many patients are in group 5?

e. Bin the patients' BMI into five equal-size groups. Label the groups using numbers 1 ( lowest values) to 5 (highest values). How many patients are in group 1?

f. Create a risk score for each patient by concatenating the group numbers obtained in parts c, d, and e. How many patients are in the risk group of 555?

## 2.5 TRANSFORMING CATeGORICAL DATA

LO 2.5

Transform categorical variables.

As discussed in Chapter 1, we use labels or names to identify the distinguishing characteristics of a categorical variable. For instance, a firm may identify each customer as either a male or a female. Here, the sex of a customer is a categorical variable representing two categories. Categorical variables can also be defined by more than two categories. Examples include marital status (single, married, widowed, divorced, separated) and the performance of a manager (excellent, good, fair, poor). Recall that we use nominal and ordinal measurement scales to represent categorical variables. In the above examples, the measurement scales for marital status and performance of a manager are nominal and ordinal, respectively.

While categorical variables are known to represent less sophisticated levels of measurement, they are often the most important variables in the analysis. For example, the sex of a customer may contain the most useful information on the customer's spending behavior. Categorical data do, however, present unique challenges in data analysis. As many analysis techniques are limited in their abilities to handle categorical data directly, steps to simplify or transform categorical data into numerical formats are often performed prior to analysis. In this section, we discuss three common approaches for transforming categorical data: category reduction, dummy variables, and category scores.

## Category Reduction

Sometimes nominal or ordinal variables come with too many categories. This presents a number of potential problems. First, variables with too many categories make the analytical model overly complex because, unlike a single parameter of a numerical variable, several parameters associated with the categories of a categorical variable must be analyzed. Second, if a variable has some categories that rarely occur, it is difficult to capture the impact of these categories accurately. In addition, a relatively small sample may not contain any observations in certain categories, creating errors when the analytical model is later applied to a larger data set with observations in all categories. Third, if one category clearly dominates in terms of occurrence, the categorical variable will fail to make a positive impact because modeling success is dependent on being able to differentiate among the observations.

An effective strategy for dealing with these issues is category reduction, where we collapse some of the categories to create fewer nonoverlapping categories. Determining the appropriate number of categories often depends on the data, context, and disciplinary norms, but there are a few general guidelines.

The first guideline states that categories with very few observations may be combined to create the "Other" category. For example, in a data set that contains the demographic data about potential customers, if many zip code categories only have a few observations, it is recommended that an "Other" category be created for these observations. The rationale behind this approach is that a critical mass can be created for this "Other" category to help reveal patterns and relationships in data.

Another guideline states that categories with a similar impact may be combined. For example, when studying public transportation ridership patterns, one tends to find that the ridership levels remain relatively stable during the weekdays and then change drastically for the weekends. Therefore, we may combine data from Monday through Friday into the "Weekdays" category and Saturday and Sunday into the "Weekends" category to simplify data from seven to only two categories.

Example 2.6 demonstrates how to use Excel and R for category reduction.

FILE
*Customers*

### EXAMPLE 2.6

After gaining some insights from the ***Customers*** data set, Catherine would like to analyze race. However, in its current form, the data set would limit her ability to perform a meaningful analysis given the large number of categories of the race variable; plus some categories have very few observations. As a result, she needs to perform a series of data transformations to prepare the data for subsequent analysis. Use Excel and R to create a new category called Other that represents the two least-frequent categories.

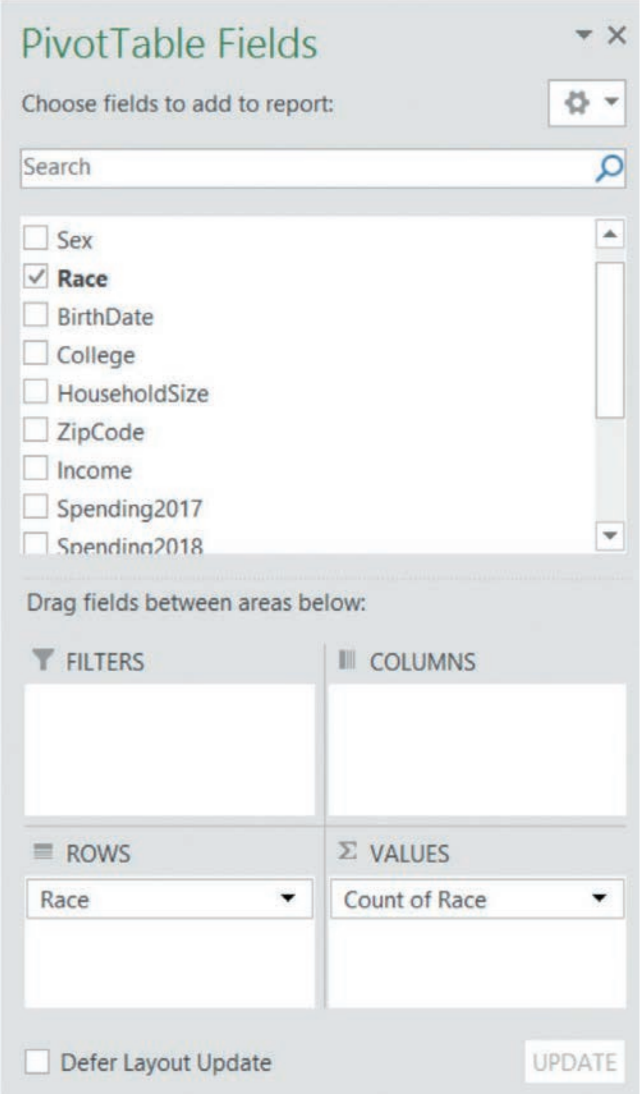
**SOLUTION:**

**Using Excel**

**a.** Open the ***Customers*** data file.

**b.** We use a pivot table to display the frequency for each race category. A pivot table allows us to summarize raw data in a more manageable way. Select cell P1 and choose **Insert > PivotTable**. In the *Create PivotTable* dialog box, we choose the *Select a table or range* option and specify the *Table/Range* value to be Customers!$A$1:$N$201, which refers to the entire ***Customers*** worksheet. We then select the *Existing Worksheet* option and specify the *Location* value to be Customers!$P$1. This means that the pivot table will start in cell P1 of the ***Customers*** worksheet. Click **OK**. The *PivotTable Fields* pane appears, as shown in Figure 2.9. First, drag the Race field to the *ROWS* box, and then drag the Race field to the *VALUES* box. The *ROWS* box lists the race categories, whereas the *VALUES* box lists the frequency of each race category.

**FIGURE 2.9** PivotTable Fields pane



PivotTable Fields

Choose fields to add to report:

Search

□ Sex
☑ Race
□ BirthDate
□ College
□ HouseholdSize
□ ZipCode
□ Income
□ Spending2017
□ Spending2018

Drag fields between areas below:

FILTERS

COLUMNS

ROWS: Race

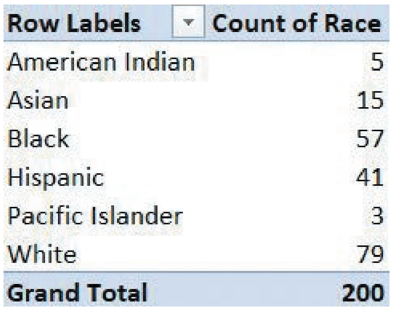
VALUES: Count of Race

□ Defer Layout Update

Microsoft Corporation

Figure 2.10 shows the pivot table that you should see on the Worksheet. It shows that American Indian and Pacific Islander are the two least-frequent categories with 5 and 3 observations, respectively.

**FIGURE 2.10** Pivot table for race



| Row Labels | Count of Race |
|---|---|
| American Indian | 5 |
| Asian | 15 |
| Black | 57 |
| Hispanic | 41 |
| Pacific Islander | 3 |
| White | 79 |
| **Grand Total** | **200** |

c. We now combine the two least-frequent race categories into the Other category using the **IF** function. In the ***Customers*** worksheet, enter the column heading NewRace in cell O1. In cell O2, enter the formula =IF(OR(C2=“American Indian”, C2=“Pacific Islander”), “Other”, C2). The **OR** function returns TRUE if the results of any of the logical tests are TRUE. The formula states that if the race category in cell C2 is either American Indian or Pacific Islander, then change the race to Other; otherwise, record the current race category in cell C2. Fill the range O3:O201 with the formula in O2. Verify that the 19th record represents the first customer in the Other category.

**Using R**

a. Import the ***Customers*** data file into a data frame (table) and label it myData.

b. First, we inspect the frequency of each Race category to identify the two least-frequent categories. Enter:

```
table(myData$Race)
```

The table shows that American Indians and Pacific Islanders are the two least-frequent categories with only five and three observations, respectively.

c. We use the **ifelse** function to create a new variable called NewRace that uses the Other category to represent American Indians and Pacific Islanders. Enter:

```
myData$NewRace <- ifelse(myData$Race %in% c(“American Indian”,
“Pacific Islander”), “Other”, myData$Race)
```

Note that the **ifelse** function evaluates the values in the Race variable, and if the value is either American Indian or Pacific Islander, it replaces it with Other; the original race value is retained otherwise.

d. We use the **table** function again to verify that the Other category has eight observations. Enter:

```
table(myData$NewRace)
```

e. We use the **View** function to display spreadsheet-style data. Enter:

```
View(myData)
```

Verify that the 19th customer is the first in the Other category.

## Dummy Variables

In many analytical models, such as regression models discussed in later chapters, categorical variables must first be converted into numerical variables. For other models, dealing with numerical data is often easier than categorical data because it avoids the complexities of the semantics pertaining to each category of the variable. A **dummy variable**, also referred to as an indicator or a binary variable, is commonly used to describe two categories of a variable. It assumes a value of 1 for one of the categories and 0 for the other category, referred to as the reference or the benchmark category. For example, we can define a dummy variable to categorize a person’s sex using 1 for male and 0 for female, using females as the reference category. Dummy variables do not suggest any ranking of the categories and, therefore, without any loss of generality, we can define 1 for female and 0 for male, using males as the reference category. All interpretation of the results is made in relation to the reference category.

### A DUMMY VARIABLE

A dummy variable, also referred to as an indicator or a binary variable, takes on values of 1 or 0 to describe two categories of a categorical variable.

Oftentimes, a categorical variable is defined by more than two categories. For example, the mode of transportation used to commute may be described by three categories: Public Transportation, Driving Alone, and Car Pooling. Given $k$ categories of a variable, the general rule is to create $k - 1$ dummy variables, using the last category as reference. For the mode-of-transportation example, we need to define only two dummy variables. Suppose we define two dummy variables $d_1$ and $d_2$, where $d_1$ equals 1 for Public Transportation, 0 otherwise, and $d_2$ equals 1 for Driving Alone, 0 otherwise. Here, Car Pooling, the reference category, is indicated when $d_1 = d_2 = 0$. Therefore, adding the third dummy variable for Car Pooling would create information redundancy; certain analytical models cannot even be estimated with $k$ dummy variables.

Example 2.7 shows how to create dummy variables using Excel and R.

## EXAMPLE 2.7

FILE
*Customers*

For the new Asian-inspired meal kits, Catherine feels that understanding the channels through which customers were acquired is important to predict customers' future behaviors. In order to include the Channel variable in her predictive model, Catherine needs to convert the Channel categories into dummy variables. Because web banner ads are probably the most common marketing tools used by Organic Food Superstore, she plans to use the Web channel as the reference category and assess the effects of other channels in relation to the Web channel. Use Excel and R to create the relevant dummy variables for the Channel variable.

**SOLUTION:**

**Using Excel**

**a.** Open the ***Customers*** data file.

**b.** Using the Web channel as the reference category, we create three dummy variables for the Channel variable because there are four categories. Enter the variable names Channel_Referral, Channel_SM, and Channel_TV in cells O1, P1, and Q1, respectively.

**c.** Enter the formula =IF(N2="Referral", 1, 0) in cell O2. Fill the range O3:O201 with the formula in O2. Enter the formula =IF(N2="SM", 1, 0) in cell P2. Fill the range P3:P201 with the formula in P2. Enter the formula =IF(N2="TV", 1, 0) in cell Q2. Fill the range Q3:Q201 with the formula in Q2. Verify that the dummy variable values for the first observation are 0, 1, and 0, respectively.

**Using R**

**a.** Import the ***Customers*** data file into a data frame (table) and label it myData.

**b.** We use the **ifelse** function to create a dummy variable for individual categories in the Channel variable. The **ifelse** function evaluates the categories in the Channel variable, and, for example, if the category is Referral, then the function assigns a 1 to the new Channel_Referral variable, and 0 otherwise.Two other dummy variables, Channel_SM and Channel_TV, are created similarly. Note that we leave out the last channel, Web, as it is a reference category. Enter:

```
myData$Channel_Referral <- ifelse(myData$Channel == "Referral", 1, 0)
myData$Channel_SM <- ifelse(myData$Channel == "SM", 1, 0)
myData$Channel_TV <- ifelse(myData$Channel == "TV", 1, 0)
```

**c.** We use the **View** function to display spreadsheet-style data. Enter:

```
View(myData)
```

Verify that the dummy variable values for the first observation are 0, 1, and 0, respectively.

## Category Scores

Finally, another common transformation of categorical variables is to create category scores. This approach is most appropriate if the data are ordinal and have natural, ordered categories. For example, in customer satisfaction surveys, we often use ordinal scales, such as very dissatisfied, somewhat dissatisfied, neutral, somewhat satisfied, and very satisfied, to indicate the level of satisfaction. While the satisfaction variable is categorical, the categories are ordered. In such cases, we can recode the categories numerically using numbers 1 through 5, with 1 being very dissatisfied and 5 being very satisfied. This transformation allows the categorical variable to be treated as a numerical variable in certain analytical models. With this transformation, we need not convert a categorical variable into several dummy variables or to reduce its categories. For an effective transformation, however, we assume equal increments between the category scores, which may not be appropriate in certain situations.

Example 2.8 shows how to convert a categorical variable into category scores using Excel and R.

### EXAMPLE 2.8

FILE
*Customers*

For the new Asian-inspired meal kits, Catherine wants to pay attention to customer satisfaction. As the customer satisfaction ratings represent ordinal data, she wants to convert them to category scores ranging from 1 (Very Dissatisfied) to 5 (Very Satisfied) to make the variable more readily usable in predictive models. Use Excel and R to create category scores for the Satisfaction variable.

**SOLUTION:**

**Using Excel**

**a.** Open the ***Customers*** data file.

**b.** Enter the column heading Satisfaction_Score in cell O1. Enter the formula =IF(M2="Very Satisfied", 5, IF(M2="Somewhat Satisfied", 4, IF(M2="Neutral", 3, IF(M2="Somewhat Dissatisfied", 2, 1)))) in cell O2. Fill the range O3:O201 with the formula in O2. The scores are now ordered based on the degree to which the customer is satisfied with Organic Food Superstore's service. See Example 2.4 for further information on the nested IF statement. Verify that the first four satisfaction scores are 1, 3, 5, and 1, respectively.

### Using R

**a.** Import the ***Customers*** data file into a data frame (table) and label it myData.

**b.** We use the **ifelse** function, in a nested format, to create category scores for the Satisfaction variable. Enter:

```
myData$Satisfaction_Score <- ifelse(myData$Satisfaction ==
"Very Dissatisfied", 1, ifelse(myData$Satisfaction == "Somewhat
Dissatisfied", 2, ifelse(myData$Satisfaction == "Neutral", 3,
ifelse(myData$Satisfaction == "Somewhat Satisfied", 4, 5))))
```

Note that the **ifelse** function evaluates the values in the Satisfaction variable, and if the value is Very Dissatisfied, the function assigns a 1 to the new Satisfaction_Score variable. Because it is a nested format, if the value is not Very Dissatisfied but is Somewhat Dissatisfied, the function assigns a 2, and so on. If the values in the Satisfaction variable are none of the first four scores, the function assigns 5 to the Satisfaction_Score variable.

**c.** We use the **View** function to display the spreadsheet-style data. Enter:

```
View(myData)
```

Verify that the first four satisfaction scores are 1, 3, 5, and 1, respectively.

## EXERCISES 2.5

### Mechanics

49. The following table has three variables and six observations.

| Sex | Income | Decision |
|---|---|---|
| Female | 95000 | Approve |
| Male | 65000 | Approve |
| Female | 55000 | Need More Information |
| Male | 72000 | Reject |
| Male | 58000 | Approve |
| Male | 102000 | Approve |

a. Convert Sex into dummy variables. Use the most frequent category as the reference category. Which category is the reference category?

b. Transform Decision into dummy variables. Use the most frequent category as the reference category. Which category is the reference category?

c. Transform the Decision values into category scores where Approve = 1, Reject = 2, and Need More Information = 3. How many observations have a category score of 2?

50. FILE ***Exercise_2.50.*** The accompanying data file contains three variables, $x_1$, $x_2$, and $x_3$.

a. The variable $x_1$ contains six categories ranging from “A” to “F.” Reduce the number of categories to five by combining the two least-frequent categories. Name the new category “Other.” How many observations are in the “Other” category?

b. The variable $x_2$ contains six categories ranging from “A” to “F.” This variable is ordinal, meaning that the categories are ordered. “A” represents the lowest level, whereas “F” represents the highest level. Replace the category names with category scores ranging from 1 ( lowest) to 6 (highest). What is the average category score for $x_2$?

c. The variable $x_3$ contains four unordered categories. To facilitate subsequent analyses, we need to convert $x_3$ into dummy variables. How many dummy variables should be created? Create the dummy variables using Category1 as the reference category.

51. FILE ***Exercise_2.51.*** The accompanying data file contains two variables, Birthdate and LoanDecision.

a. LoanDecision contains three unordered categories. To facilitate subsequent analyses, we need to convert LoanDecision into dummy variables. How many dummy variables should be created? Create the dummy variables using “Need more information” as the reference category.

b. Create a new variable based on LoanDecision. The new variable should have only two categories: “Approve” and “Not approve.” The “Not approve” category combines the “Reject” and “Need more information” categories. How many observations are in the “Not approve” category?

52. FILE ***Exercise_2.52.*** The accompanying data file contains two variables, $x_1$ and $x_2$.

a. The variable $x_1$ contains three categories ranging from “Low” to “High.” Convert the category names into category scores (i.e., 1 = “Low”, 2 = “Medium”, and 3 = “High”). How many observations have a category score of 3?

b. Reduce the number of categories in $x_2$ by combining the three least-frequent categories. Name the new category "Other". How many observations are in the "Other" category?

c. Convert the new $x_2$ into dummy variables. How many dummy variables should be created? Create the dummy variables using the "Other" category as the reference category.

53. FILE ***Exercise_2.53.*** The accompanying data file contains three variables, $x_1$, $x_2$, and $x_3$.

a. The variable $x_1$ contains three categories: S, M, and L. Convert the category names into category scores (i.e., S = 1, M = 2, and L = 3). How many observations have a category score of 3?

b. The variable $x_2$ contains two categories: Yes and No. Transform $x_2$ into an appropriate number of dummy variables. How many dummy variables should be created?

c. The variable $x_3$ contains four categories: A, B, C, and D. Reduce the number of categories by combining the two least-frequent categories into a new category E. How many observations are in the E category?

## Applications

54. FILE ***Home_Loan.*** Consider the accompanying data file that includes information about home loan applications. Variables on each application include the application number (Application), whether the application is conventional or subsidized by the federal housing administration (LoanType), whether the property is a single-family or multifamily home (PropertyType), and whether the application is for a first-time purchase or refinancing (Purpose).

a. Are the variables LoanType, PropertyType, and Purpose nominal or ordinal data? Why?

b. Which categories are the most frequent categories for LoanType, PropertyType, and Purpose?

c. To facilitate subsequent analyses, transform LoanType, PropertyType, and Purpose into dummy variables. Use the most frequent categories of the variables as the reference categories. Which categories of LoanType, PropertyType, and Purpose are the reference categories? How many dummy variables are created in total?

55. FILE ***Shipment.*** A manager of a local package delivery store believes that too many of the packages were damaged or lost. She extracts a sample of 75 packages to create the accompanying data file with the following variables: the package number (Package), the status of the package (Status: Delivered, Damaged, or Lost), the delivery type (Delivery: Standard, Express, or Same day), and the size of the package (Size: S, M, L, and XL).

a. Transform the Delivery variable into dummy variables. Use the most frequent category as the reference category. How many dummy variables should be created? Which category of Delivery is the reference category?

b. Combine the two least-frequent categories in the Size variable into a new category called Other. How many observations are there in the new category?

c. Replace the category names in the Status variable with scores 1 ( Lost), 2 (Damaged), or 3 (Delivered). What is the average status score of the 75 packages?

56. FILE ***Technician.*** After each thunderstorm, a technician is assigned to do a check on cellular towers in his or her service area. During one of the visits, the technician creates the accompanying data file that contains the tower number (Tower), the unit model (Model: A or B), and the extent of the damage to the unit (Damage: None, Minor, Partial, Severe, and Total for a total loss).

a. Transform the Model variable into dummy variables. How many dummy variables should be created?

b. Transform the Damage variable into categorical scores ranging from 4 (Total) to 3 (Severe), 2 (Partial), 1 ( Minor), and 0 (None). What is the average damage score of the cell towers?

57. FILE ***Game_Players.*** Refer to Exercise 2.46 for a description of the problem and data set.

a. The variable Satisfaction contains five ordered categories: Very Dissatisfied, Dissatisfied, Neutral, Satisfied, and Very Satisfied. Replace the category names with scores ranging from 1 ( Very Dissatisfied) to 5 (Very Satisfied). What is the average satisfaction score of the players?

b. The variable Enjoyment contains five ordered categories: Very Low, Low, Neutral, High, and Very High. Replace the category names with scores ranging from 1 ( Very Low) to 5 (Very High). What is the average enjoyment score of the players?

c. The variable Recommend contains five ordered categories: Definitely Will Not, Will Not, Neutral, Will, and Definitely Will. Replace the category names with scores ranging from 1 ( Definitely Will Not) to 5 (Definitely Will). What is the average recommendation score of the players?

d. The variable Type contains three unordered game categories: Action, Role Play, and Sports. To facilitate subsequent analyses, transform Type into dummy variables. Use the least frequent category as the reference category. Which category is the reference category? How many dummy variables should be created?

58. FILE ***Engineers.*** Refer to Exercise 2.47 for a description of the problem and data set.

a. The variable Personality contains four unordered personality types: Analyst, Diplomat, Explorer, and Sentinel. To facilitate subsequent analyses, Erin needs to convert this variable into dummy variables. How many dummy variables should be created? Create the dummy variables using Analyst type as the reference category. How many observations are in the reference category?

b. The variable Level contains three ordered position levels: Engineer I (lowest), Engineer II, and Senior Engineer (highest). Replace the level names with scores ranging from 1 ( lowest) to 3 (highest). What is the average level score of the engineers?

59. FILE ***Patients.*** Refer to Exercise 2.48 for a description of the problem and data set.

a. The variable Race contains five unordered categories: American Indian, Asian/Pacific Islander, Hispanic, Non-Hispanic Black, and Non-Hispanic White. Reduce the number of categories to four by combining the two least frequent categories. Name the new category “Other.” How many observations are in the “Other” category?

b. Transform the Race variable with the new categories into dummy variables. Use the most frequent race category in the data as the reference category. Which category is the reference category? How many dummy variables should be created?

c. The variable Education contains four ordered categories: HS (lowest educational attainment level), Some College, College, and Graduate (highest educational attainment level). Replace the category names with category scores ranging from 1 (lowest) to 4 (highest). What is the average category score for Education?

## 2.6 WRITING WITH BIG DATA

FILE
*TechSales_Reps*

### Case Study

Cassius Weatherby is a human resources manager at a major technology firm that produces software and hardware products. He would like to analyze the net promoter score (NpS) of sales professionals at the company. The NpS measures customer satisfaction and loyalty by asking customers how likely they are to recommend the company to others on a scale of 0 (unlikely) to 10 (very likely). This measure is an especially important indicator for the company’s software business as a large percentage of the sales leads come from customer referrals. Cassius wants to identify relevant factors that are linked with the NpS that a sales professional receives. These insights can help the company make better hiring decisions and develop a more effective training program.

With the help of the company’s IT group, a data set with over 20,000 records of sales professionals is extracted from the enterprise data warehouse. The relevant variables include the product line to which the sales professional is assigned, age, sex, the number of years with the company, whether the sales professional has a college degree, personality type based on the Myers-Briggs personality assessment, the number of professional certificates acquired, the average score from the 360-degree annual evaluation, base salary, and the average NpS received. Cassius is tasked with inspecting and reviewing the data and preparing a report for the company’s top management team.

### Sample Report—Evaluation of Net Promoter Scores

The net promoter score (NpS) is a key indicator of customer satisfaction and loyalty. It measures how likely a customer would be to recommend a product or company to others. Because our software line for business relies heavily on customer referrals to generate sales leads, the NpS that our sales professionals receive is a key indicator of our company’s future success.

dizain/Shutterstock

From a total of about 20,000 records of sales professionals, we select only the sales professionals in the software product group and divide them into two categories: those with an average NpS below nine and those with an average NpS of nine or ten. When a customer gives a sales professional an NpS of nine or ten, the customer is considered "enthusiastically loyal," meaning that they are very likely to continue purchasing from us and refer their colleagues to our company. Based on the NpS categorization, we then divide the sales professionals into two categories: those with zero to three professional certificates and those with four or more professional certificates. Table 2.11 shows the results. Of the 12,130 sales professionals in the software product group, we find that 65.57% have earned less than four professional certificates, whereas 34.43% have earned four or more. However, there appears to be a link between those with four or more professional certificates and NpS values. For those who received an NpS of nine or ten, we find that 62.60% have earned at least four professional certificates. Similarly, for those who received an NpS of below nine, we find that 73.00% earned less than four professional certificates.

**TABLE 2.11** Sales Professionals by the Number of Certificates and NPS Value

| Number of certificates | Full sample ($n$ = 12,130) | NPS < 9 ($n$ = 9,598) | NPS ≥ 9 ($n$ = 2,532) |
|---|---|---|---|
| 0 to 3 | 65.57% | 73.00% | 37.40% |
| 4 or more | 34.43% | 27.00% | 62.60% |

Although this might simply suggest that high-achieving employees tend to be self-motivated to earn professional certificates, we also believe that sales professionals with sufficient technical knowledge can effectively communicate and assist their customers in finding technology solutions, which will lead to increased customer satisfaction and loyalty. Our training and development program must place a greater emphasis on helping the employees earn relevant certifications and acquire necessary technical knowledge.

Based on NpS categorization, we then divide the sales professionals into categories based on personality type. Table 2.12 shows the results. In addition to professional certification, we find that personality types are linked with NpS values. Among the four personality types, Diplomats and explorers account for 72.69% of all the sales professionals in the software group. However, when we divide the employees based on the NpS values, these two personality types account for 91.63% of the group with an average NpS of nine or ten, whereas they account for only 67.69% for the below nine NpS group.

**TABLE 2.12** Sales Professionals by Personality Type and NPS Value

| Myers-Briggs Personality Type | Full sample ($n$ = 12,130) | NPS < 9 ($n$ = 9,598) | NPS ≥ 9 ($n$ = 2,532) |
|---|---|---|---|
| Analyst | 12.13% | 14.47% | 3.24% |
| Diplomat | 35.62% | 33.07% | 45.30% |
| explorer | 37.07% | 34.62% | 46.33% |
| Sentinel | 15.19% | 17.84% | 5.13% |

We also examined NpS variations by other variables such as age, sex, education attainment, sales, and commission but did not find considerable differences in NpS categorization. Other variables such as salary and the tenure of the employee with the company are not included in our initial analysis.

Based on the insights from this analysis, we request that the company appoint an analytics task force to conduct a more comprehensive analysis of sales professionals. We strongly suggest that the analysis focus on professional certification and personality, among relevant factors for determining the NpS value. At a minimum, two goals of the task force should include making recommendations on (1) a redesign of our training and development program to focus on helping employees acquire relevant professional certificates and (2) the efficacy of using personality types as part of the hiring decision.

## Suggested Case Studies

Data wrangling is a crucial step in any data analytics project. The data inspection, preparation, and transformation techniques discussed in this chapter can be applied to many data sets. Here are some suggestions using the big data that accompany this text as well as data that are easily accessed from the internet.

**Report 2.1** FILE ***Car_Crash***. Subset the data set based on the location, day of the week, type of collision, and lighting condition. Compare these subsets of data to find interesting patterns. Can you identify any links between crash fatality and the aforementioned variables? Are there any missing values? Which strategy should you use to handle the missing values? Because many of the variables are categorical, you should consider transforming them into dummy variables prior to the analysis.

**Report 2.2** FILE ***House_Price***. Subset the data based on variables, such as number of bedrooms, number of bathrooms, home square footage, lot square footage, and age of the house. Which variables can be removed when predicting house prices? Are there any variables that display a skewed distribution? If there are, perform logarithm transformations for these variables. Would it make sense to transform some of the numeric values into categorical values using the binning strategy? Is equal size or equal interval binning strategy more appropriate in these situations?

**Report 2.3** FILE ***Longitudinal_Survey***. Subset the data based on age, sex, or race. Are there any missing values in the data? Which strategy should you use to handle the missing values? Consider if any new variables can be created using the existing variables. explore the opportunities of transforming numeric variables through binning and transforming categorical variables by creating dummy variables.

**Report 2.4** FILE ***TechSales_Reps***. Consider whether data distribution skewness exists in some of the numeric variables, and if it does, determine how to transform the data into a less skewed distribution. perform data subsetting and use simple summary measures such as averages and frequency counts to find out if any differences exist across subsets.

**Report 2.5** To examine the impact of the pandemic of COVID-19 disease on the economy, visit the u.S. Department of Labor's website on unemployment insurance weekly claim data (https://oui.doleta.gov/unemploy/claims.asp) to obtain the national and state unemployment data for 2020. Write a report discussing the impact of the pandemic, which began in the u.S. in March 2020, on unemployment. Consider integrating the unemployment data with population data from the u.S. Census Bureau (www.census.gov) and/or infection data from the u.S. Centers for Disease Control and prevention (www.cdc.gov).

# 9 Logistic Regression

## LEARNING OBJECTIVES

**After reading this chapter, you should be able to:**

LO **9.1** **Estimate and interpret linear and logistic regression models.**

LO **9.2** **Calculate and interpret odds and accuracy.**

LO **9.3** **Use cross-validation to assess classification performance.**

As discussed in Chapters 7 and 8, regression analysis is one of the most widely used techniques in predictive analytics. It is used to capture the relationship between two or more variables and to predict the outcome of a response variable based on several predictor variables. For example, we use regression analysis to capture the impact of a firm's increase in its marketing expenditure on sales or predict the price of a house based on its size and location.

So far, we have used regression models where the response variable is numerical. We extend the analysis to estimate and interpret binary choice (classification) models where the response variable is a dummy variable. The linear probability model and the logistic regression model are used, for example, to predict the probability, or the odds, that an open house attendee will join a health club based on the attendee's age, income, and marital status.

There is no universal goodness-of-fit measure for binary choice models to assess how well the model fits the data. For example, unlike a linear regression model, we cannot assess a binary choice model based on its coefficient of determination $R^2$. For such models we discuss commonly used performance measures including the accuracy sensitivity, and specificity rates. Finally, we employ cross-validation techniques where we partition the original sample into a training set to build the model and a validation set to evaluate the model.

Alex Mit/Shutterstock

## INTRODUCTORY CASE

### Spam Detection System

Electronic mail (e-mail) is one of the most widely used Internet services. It is a convenient method of communication that is fast, cheap, and accessible. While e-mails can greatly benefit productivity, they are susceptible to unsolicited, unwanted, and possibly virus-infested spam messages from illegitimate e-mail addresses.

Peter Derby works as a cybersecurity analyst at a private equity firm. His colleagues at the firm have been overwhelmed by large-scale spam e-mails. Peter would like to implement an effective spam detection system on the company's e-mail server. In particular, he has been tasked to create spam filters that detect spam e-mails and stop them from getting into e-mail inboxes. He analyzes a sample of 500 spam and legitimate e-mails with the following relevant variables: spam (1 if spam, 0 otherwise), the number of recipients, the number of hyperlinks, and the number of characters in the message. A portion of the data is shown in Table 9.1.

FILE
*Spam*

**TABLE 9.1** Spam Data ($n = 500$)

| Record | Spam | Recipients | Hyperlinks | Characters |
|---|---|---|---|---|
| 1 | 0 | 19 | 1 | 47 |
| 2 | 1 | 17 | 11 | 68 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 500 | 1 | 13 | 2 | 32 |

Peter would like to use the information in Table 9.1 to:

1. Develop a logistic regression model for spam detection.
2. Assess the performance of the estimated model.
3. Make spam predictions for specific predictor variable values.

A synopsis of this case is provided at the end of Section 9.2.

LO 9.1

# 9.1 THE LInEAR PROBABILITy MODEL AnD THE LOGISTIc REGRESSIOn MODEL

Estimate and interpret linear and logistic regression models.

So far we have considered regression models where dummy (binary) variables are used only as predictor variables. In this section, we analyze **binary choice (classification) models** where the response variable is a binary variable. The consumer choice literature is replete with applications such as whether or not to buy a house, join a health club, or go to graduate school. At the firm level, managers make decisions such as whether or not to run a marketing campaign, restructure debt, or approve a loan.

In the above applications, the response variable is binary, where one of the choices (classes) can be designated as 1 and the other as 0. It is common to refer to the class designated as 1 as the target or positive class and the class designated as 0 as the nontarget or negative class. Usually, this choice can be related to a host of factors—the predictor variables. For instance, whether or not a family buys a house depends on predictor variables such as household income, mortgage rates, and so on.

## The Linear Probability Model

Consider a linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$, where $y$ is a binary variable with an expected value, conditional on the predictor variables, equal to $\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$. Because $y$ is a discrete random variable with only two possible outcomes, its conditional expected value also equals $0 \times P(y = 0) + 1 \times P(y = 1) = P(y = 1)$, or simply $p$. In other words, the probability of success $p$ is a linear function of the predictor variables; that is, $p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$. The linear regression model applied to a binary response variable is called the **linear probability model (LPM)**.

THE LINEAR PROBABILITY MODEL

The linear probability model is specified as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$, where $y$ assumes a 1 or 0 value.

- Predictions with this model are made by $\hat{p} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$, where $\hat{p}$ is the predicted probability of success and $b_0, b_1, b_2, \ldots, b_k$ are the coefficient estimates.
- It is advisable to use unrounded coefficient estimates for making predictions.

### EXAMPLE 9.1

FILE
*Mortgage*

The Great Recession has forced financial institutions to be extra stringent in granting mortgage loans. Thirty recent mortgage applications are obtained to analyze the mortgage approval rate. The response variable $y$ equals 1 if the mortgage loan is approved, 0 otherwise. It is believed that approval depends on the percentage of the down payment $x_1$ and the percentage of the income-to-loan ratio $x_2$. Table 9.2 shows a portion of the data.

**TABLE 9.2** Mortgage Application Data

| $y$ | $x_1$ | $x_2$ |
|---|---|---|
| 1 | 16.35 | 49.94 |
| 1 | 34.43 | 56.16 |
| ⋮ | ⋮ | ⋮ |
| 0 | 17.85 | 26.86 |

**a.** Estimate and interpret the linear probability model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$.

**b.** Predict the loan approval probability for an applicant with a 20% down payment and a 30% income-to-loan ratio. What if the down payment was 30%?

**SOLUTION:**

**a.** Table 9.3 shows a portion of the regression results. The estimated regression equation is $\hat{p} = -0.8682 + 0.0188x_1 + 0.0258x_2$. With $p$-values of 0.012 and 0.000, respectively, both predictor variables exert a positive and statistically significant influence on loan approval, at a 5% level. Also, holding the income-to-loan ratio constant, $b_1 = 0.0188$ implies that a 1-percentage-point increase in down payment increases the approval probability by 0.0188, or by 1.88 percentage points. Similarly, holding down payment constant, a 1-percentage-point increase in the income-to-loan ratio increases the approval probability by 0.0258, or by 2.58 percentage points.

**TABLE 9.3** The Linear Probability Model Results for Example 9.1

| | Coefficients | Standard error | *t* stat | *p*-value |
|---|---|---|---|---|
| Intercept | −0.8682 | 0.2811 | −3.089 | 0.005 |
| Down payment ($x_1$) | 0.0188 | 0.0070 | 2.695 | 0.012 |
| Income-to-loan ratio ($x_2$) | 0.0258 | 0.0063 | 4.107 | 0.000 |

**b.** Using unrounded coefficient estimates, the predicted loan approval probability for an applicant with a 20% down payment and a 30% income-to-loan ratio is $\hat{p} = -0.8682 + 0.0188 \times 20 + 0.0258 \times 30 = 0.2836$. Similarly, the predicted loan approval probability with a down payment of 30% is $\hat{p} = -0.8682 + 0.0188 \times 30 + 0.0258 \times 30 = 0.4720$. In other words, as down payment increases by 10 percentage points, the predicted probability of loan approval increases by 0.188 (= 0.4720 − 0.2836), which is essentially the estimated slope, 0.0188, multiplied by 10. The estimated slope coefficient for the percentage of income-to-loan variable can be interpreted similarly.

Although it is easy to estimate and interpret the linear probability model, it has some shortcomings. The major shortcoming is that it can produce predicted probabilities that are greater than 1 or less than 0. For instance, for a down payment of 60%, with the same income-to-loan ratio of 30%, the model predicts a loan approval probability of $\hat{p} = -0.8682 + 0.0188 \times 60 + 0.0258 \times 30 = 1.04$, a probability greater than 1. Similarly, for a down payment of 4%, the model predicts a negative probability, $\hat{p} = -0.8682 + 0.0188 \times 4 + 0.0258 \times 30 = -0.02$. Furthermore, the linearity of the relationship may also be questionable. For instance, we would expect a big increase in the probability of loan approval if the applicant makes a down payment of 30% instead of 20%. This increase in probability is likely to be much smaller if the same 10-percentage-point increase in down payment is from 60% to 70%. The linear probability model cannot differentiate between these two scenarios. For these reasons, we introduce the logistic regression model, which is a more appropriate probability model for binary response variables.

## The Logistic Regression Model

As mentioned above, the major shortcoming of the LPM is that for certain values of the predictor variables, the predicted probability can be outside the [0,1] interval. For meaningful analysis, we would like a nonlinear specification that constrains the predicted probability between 0 and 1.

The probability of success, $p$, for the **logistic regression model** is specified as

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}$$

where $\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k) = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}$ and $e \approx 2.718$. The logistic specification ensures that the predicted probability is between 0 and 1 for all values of the predictor variables.

The logistic regression model cannot be estimated with standard ordinary least squares (OLS) procedures. Instead, we rely on the method of **maximum likelihood estimation (MLE)**. While the MLE of the logistic regression model is not supported by Excel, it can easily be estimated with most statistical packages, including Analytic Solver and R.

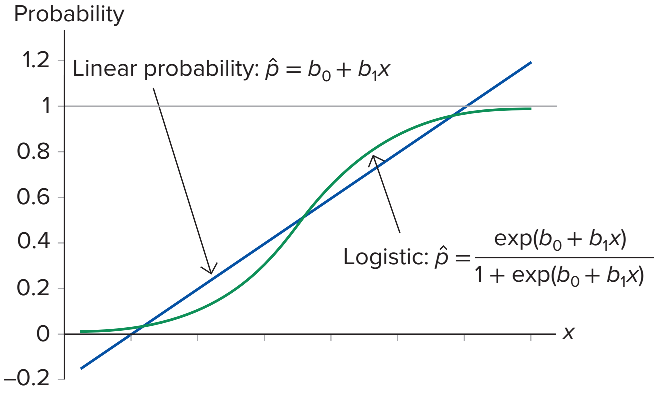
### THE LOGISTIC REGRESSION MODEL

The probability of success, $p$, for the logistic regression model is specified as $p = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}$.

- Predictions with this model are made by $\hat{p} = \frac{\exp(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k)}{1 + \exp(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k)}$, where $\hat{p}$ is the predicted probability of success and $b_0, b_1, b_2, \ldots, b_k$ are the coefficient estimates.
- It is advisable to use unrounded coefficient estimates for making predictions.

For illustration, let the binary response variable $y$ be influenced by a single predictor variable $x$. Figure 9.1 highlights the relationship between the predicted probability $\hat{p}$ and $x$ for the linear probability model and the logistic regression model, given $b_1 > 0$. Note that in the linear probability model, the probability falls below 0 for small values of $x$ and exceeds 1 for large values of $x$. The probabilities implied by the logistic regression model, however, are always constrained in the [0,1] interval.

**FIGURE 9.1** Predicted probabilities for the linear probability model and the logistic regression model, with $b_1 > 0$



It is important to be able to interpret the regression coefficients of the logistic regression model. In a linear probability model, the interpretation of a regression coefficient is straightforward. For instance, if the estimated linear probability model is $\hat{p} = -0.20 + 0.03x$, it implies that for every 1-unit increase in $x$, the predicted probability $\hat{p}$ increases by 0.03. We note that $\hat{p}$ increases by 0.03, whether $x$ increases from 10 to 11 or from 20 to 21.

Now consider the estimated logistic regression model, $\hat{p} = \frac{\exp(-2.10 + 0.18x)}{1 + \exp(-2.10 + 0.18x)}$. Because the regression coefficient $b_1 = 0.18$ is positive, we can infer that $x$ exerts a positive influence on $\hat{p}$. However, the exact impact based on the estimated regression coefficient is not obvious. A useful method to interpret the estimated coefficient is to highlight the changing impact of $x$ on $\hat{p}$. For instance, given $x = 10$, we compute the predicted probability as $\hat{p} = \frac{\exp(-2.10 + 0.18 \times 10)}{1 + \exp(-2.10 + 0.18 \times 10)} = 0.4256$. Similarly, for $x = 11$, the predicted probability is $\hat{p} = 0.4700$. Therefore, as $x$ increases by one unit from 10 to 11, the predicted probability increases by 0.0444 (= 0.4700 − 0.4256). However, the increase in $\hat{p}$ will not be the same if $x$ increases from 20 to 21. We can show that $\hat{p}$ increases from 0.8176 when $x = 20$ to 0.8429 when $x = 21$, for a smaller increase of 0.0253. Note that when $x$ is relatively large, its reduced influence on the predicted probability is consistent with the depiction of the logistic probabilities in Figure 9.1.

## EXAMPLE 9.2

**FILE**
*Mortgage*

Let's revisit Example 9.1.

**a.** Estimate and interpret the logistic regression model for the loan approval outcome $y$ based on the applicant's percentage of down payment $x_1$ and the income-to-loan ratio $x_2$.

**b.** For an applicant with a 30% income-to-loan ratio, predict loan approval probabilities with down payments of 20% and 30%.

**SOLUTION:**

**a.** We report the logistic regression results in Table 9.4. (Instructions for estimating the logistic regression model with Analytic Solver and R are provided shortly.)

**TABLE 9.4** Logistic Regression Results for Example 9.2

| | Estimate | Std. error | *z* value | *P* (>\|z\|) |
|---|---|---|---|---|
| (Intercept) | −9.3671 | 3.1958 | −2.931 | 0.003 |
| Down payment ($x_1$) | 0.1349 | 0.0640 | 2.107 | 0.035 |
| Income-to-loan ratio ($x_2$) | 0.1782 | 0.0646 | 2.758 | 0.006 |

As in the case of the linear probability model, both predictor variables exert a positive and statistically significant influence on loan approval at a 5% level, given positive estimated coefficients and $p$-values of 0.035 and 0.006, respectively. (In maximum likelihood estimation, the significance tests are valid only with large samples. Consequently, we conduct the $z$-test, in place of the usual $t$-test, to evaluate the statistical significance of a coefficient. The last column includes the $p$-values for a two-tailed test.)

**b.** Using unrounded coefficient estimates, the predicted probability is computed as

$$\hat{p} = \frac{\exp(-9.3671 + 0.1349x_1 + 0.1782x_2)}{1 + \exp(-9.3671 + 0.1349x_1 + 0.1782x_2)}.$$

The predicted loan approval probability with $x_1 = 20$ and $x_2 = 30$ is

$$\hat{p} = \frac{\exp(-9.3671 + 0.1349 \times 20 + 0.1782 \times 30)}{1 + \exp(-9.3671 + 0.1349 \times 20 + 0.1782 \times 30)} = 0.2103.$$

Similarly, the predicted loan approval probability with $x_1 = 30$ and $x_2 = 30$ is 0.5065. Note that, given the income-to-loan ratio of 30%, the predicted probability increases by 0.2962 (= 0.5065 − 0.2103) when the down payment increases from 20% to 30%. For the same income-to-loan ratio of 30%, it can be shown that the increase in the predicted probability is only 0.0449 (= 0.9833 − 0.9384) when the down payment increases from 50% to 60%.

## Using Analytic Solver and R to Estimate the Logistic Regression Model

FILE
*Mortgage*

Analytic Solver and R are useful when estimating the logistic regression model. We use the Mortgage data to replicate the results for Example 9.2.

### Using Analytic Solver

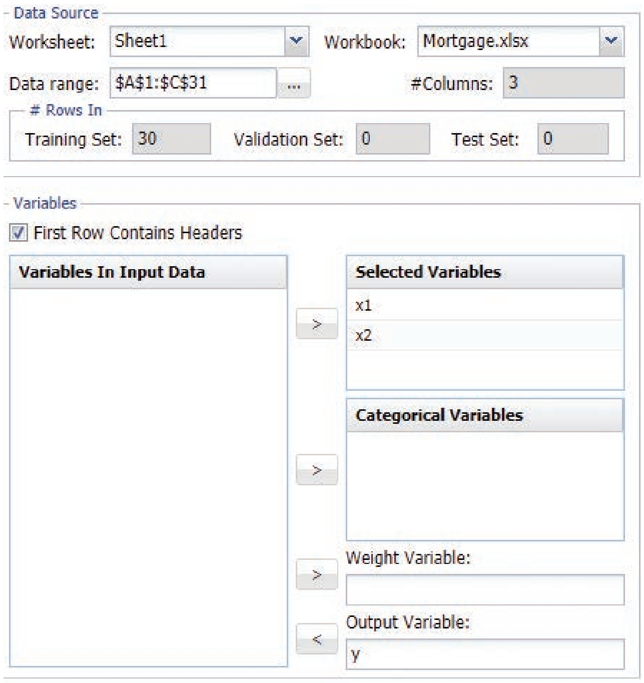
Before following the Analytic Solver instructions, make sure that you have read Appendix B ("Getting Started with Excel and Excel Add-Ins").

**a.** Open the ***Mortgage*** data file.

**b.** Choose **Data Mining > Classify > Logistic Regression** from the menu.

**c.** See Figure 9.2. Click on the ellipsis ⋯ next to the *Data range* and highlight cells A1:C31. Make sure that the box preceding *First Row Contains Headers* is checked. The *Variables in Input Data* box will populate. Select and move variables $x_1$ and $x_2$ to *Selected Variables* box and *y* to *Output Variable* box. Accept other defaults and click *Next*.

**FIGURE 9.2** Logistic regression dialog box



Data Source

Worksheet: Sheet1 Workbook: Mortgage.xlsx

Data range: $A$1:$C$31 #Columns: 3

# Rows In

Training Set: 30 Validation Set: 0 Test Set: 0

Variables

☑ First Row Contains Headers

Variables In Input Data

Selected Variables
x1
x2

Categorical Variables

Weight Variable:

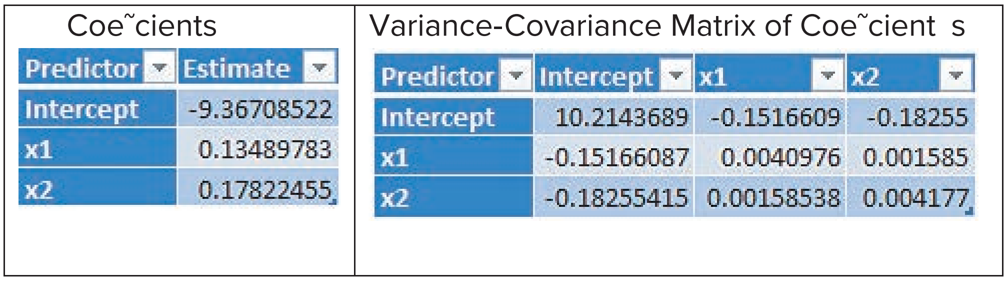
Output Variable: y

Source: Microsoft Excel

**d.** Check *Variance-Covariance Matrix* under *Regression: Display*; this is used to compute the standard errors for the significance tests. Accept other defaults and click *Next*. Accept the defaults in the *Scoring* tab and click *Finish*.

**e.** Analytic Solver will produce a lot of output in separate worksheets. Figure 9.3 shows the relevant portion from the LogReg_Output worksheet.

**FIGURE 9.3** Analytic Solver's relevant output for Example 9.2



Coefficients

| Predictor | Estimate |
|---|---|
| Intercept | -9.36708522 |
| x1 | 0.13489783 |
| x2 | 0.17822455 |

Variance-Covariance Matrix of Coefficient s

| Predictor | Intercept | x1 | x2 |
|---|---|---|---|
| Intercept | 10.2143689 | -0.1516609 | -0.18255 |
| x1 | -0.15166087 | 0.0040976 | 0.001585 |
| x2 | -0.18255415 | 0.00158538 | 0.004177 |

Source: Microsoft Excel

Note that the coefficient estimates are the same as those reported in Table 9.4. These coefficient estimates can be used to find the predicted probabilities at specific $x_1$ and $x_2$ values. In order to find the standard errors, we take the positive square root of the diagonal elements of the variance-covariance matrix. For example, the standard error of $b_1 = \sqrt{0.0040976} = 0.0640$, which is the same as in Table 9.4. The $z$-value is calculated as $z = \frac{b_1}{se(b_1)} = \frac{0.1349}{0.0640} = 2.107$.

### Using R

FILE
*Mortgage*

**a.** Import the ***Mortgage*** data file into a data frame (table) and label it myData.

**b.** We use the **glm** function to construct a logistic regression model object, which is a generalized version of the **lm** function; we label this object as Logistic_Model. Within the function, we specify the response and the predictor variables, the binomial *family* option to denote a logistic regression model (default for binomial), and the data frame. Like the linear regression model, we use the **summary** function to view the output. Enter:

```
Logistic_Model <- glm(y ~ x1 + x2, family = binomial, data = myData)
summary(Logistic_Model)
```

Note that the results are the same as those reported in Table 9.4.

**c.** We use the **predict** function to find predicted loan probabilities when $x_1 = 20$ and $x_2 = 30$, and when $x_1 = 30$ and $x_2 = 30$. Within the function we specify type =“response”, to compute predicted probabilities. Enter:

```
predict(Logistic_Model, data.frame(x1=c(20, 30), x2=30), type = “response”)
```

R returns: 0.2104205 and 0.5066462.

Note that in Example 9.2, we predicted the probabilities as 0.2103 and 0.5065; the differences are due to rounding. Other probabilities can be found similarly.

## EXERCISES 9.1

### Mechanics

1. Consider a binary response variable $y$ and a predictor variable $x$ that varies between 0 and 4. The linear probability model is estimated as $\hat{y} = -1.11 + 0.54x$.
   a. Compute the estimated probability for $x = 2$ and $x = 3$.
   b. For what values of $x$ is the estimated probability negative or greater than 1?

2. Consider a binary response variable $y$ and a predictor variable $x$. The following table contains the parameter estimates of the linear probability model (LPM) and the logistic regression model, with the associated $p$-values shown in parentheses.

| Variable | LPM | Logistic |
|---|---|---|
| Intercept | −0.72<br>(0.04) | −6.20<br>(0.04) |
| $x$ | 0.05<br>(0.06) | 0.26<br>(0.02) |

   a. Test for the significance of the intercept and the slope coefficients at the 5% level in both models.
   b. What is the predicted probability implied by the linear probability model for $x = 20$ and $x = 30$?
   c. What is the predicted probability implied by the logistic regression model for $x = 20$ and $x = 30$?

3. Consider a binary response variable $y$ and two predictor variables $x_1$ and $x_2$. The following table contains the parameter estimates of the linear probability model (LPM) and the logistic regression model, with the associated $p$-values shown in parentheses.

| Variable | LPM | Logistic |
|---|---|---|
| Intercept | −0.40<br>(0.03) | −2.20<br>(0.01) |
| $x_1$ | 0.32<br>(0.04) | 0.98<br>(0.06) |
| $x_2$ | −0.04<br>(0.01) | −0.20<br>(0.01) |

   a. Comment on the significance of the variables.
   b. What is the predicted probability implied by the linear probability model for $x_1 = 4$ with $x_2$ equal to 10 and 20?
   c. What is the predicted probability implied by the logistic regression model for $x_1 = 4$ with $x_2$ equal to 10 and 20?

4. Using 30 observations, the following output was obtained when estimating the logistic regression model.

| | Estimate | Std. error | *z* value | *P*(>\|*z*\|) |
|---|---|---|---|---|
| Intercept | −0.188 | 0.083 | 2.27 | 0.024 |
| *x* | 3.852 | 1.771 | 2.18 | 0.030 |

   a. What is the predicted probability when $x = 0.40$?
   b. Is $x$ significant at the 5% level?

5. FILE ***Exercise_9.5.*** The accompanying data file contains 20 observations on the binary response variable $y$ along with the predictor variables $x_1$ and $x_2$.
   a. Estimate the linear probability model to compute $\hat{y}$ for $x_1 = 12$ and $x_2 = 8$.
   b. Estimate the logistic regression model to compute $\hat{y}$ for $x_1 = 12$ and $x_2 = 8$.

## Applications

6. FILE ***Purchase.*** Annabel, a retail analyst, has been following Under Armour, Inc., the pioneer in the compression-gear market. Compression garments are meant to keep moisture away from a wearer's body during athletic activities in warm and cool weather. Annabel believes that the Under Armour brand attracts a younger customer relative to other companies. The accompanying file includes data on the age of the customers and whether or not they purchased Under Armour (Purchase; 1 for purchase , 0 otherwise).
   a. Estimate the linear probability model using Under Armour as the response variable and Age as the predictor variable.
   b. Compute the predicted probability of an Under Armour purchase for a 20-year-old customer and a 30-year-old customer.
   c. Test Annabel's belief that the Under Armour brand attracts a younger customer, at the 5% level.

7. FILE ***Purchase.*** Refer to the previous exercise for a description of the data set. Estimate the logistic regression model using Under Armour as the response variable and Age as the predictor variable.
   a. Compute the predicted probability of an Under Armour purchase for a 20-year-old customer and a 30-year-old customer.
   b. Test Annabel's belief that the Under Armour brand attracts a younger customer, at the 5% level.

8. FILE ***Parole.*** Parole boards use risk assessment tools to determine an individual's likelihood of returning to crime. It has been found that older people are less likely to re-offend than younger ones. In addition, once released on parole, women are not likely to re-offend. A sociologist collects data on 20 individuals who were released on parole two years ago. She notes if the parolee committed another crime over the last two years (Crime equals 1 if crime committed, 0 otherwise), the parolee's age at the time of release, and the parolee's sex (Male equals 1 if male , 0 otherwise). The accompanying file includes relevant data.
   a. Estimate the linear probability model where crime depends on age and the parolee's sex.
   b. Are the results consistent with the claims of other studies with respect to age and the parolee's sex?
   c. Predict the probability of a 25-year-old male parolee committing another crime; repeat the prediction for a 25-year-old female parolee.

9. FILE ***Parole.*** Refer to the previous exercise for a description of the data set.
   a. Estimate the logistic regression model where crime depends on age and the parolee's sex.
   b. Are the results consistent with the claims of other studies with respect to age and the parolee's sex?
   c. Predict the probability of a 25-year-old male parolee committing another crime; repeat the prediction for a 25-year-old female parolee.

10. FILE ***Health_Insurance.*** A significant proportion of Americans do not have health insurance, especially those on the lower end of the economic spectrum. The accompanying data file includes information on insurance coverage (1 for coverage, 0 for no coverage) for 30 working individuals. Also included are the percentage of the premium paid by the employer and the individual's income (in $1,000s).
    a. Estimate the linear probability model for insurance coverage with premium percentage and income used as the predictor variables.
    b. Consider an individual with an income of $60,000. What is the probability that she has insurance coverage if her employer contributes 50% of the premium? What if her employer contributes 75% of the premium?

11. FILE ***Health_Insurance.*** Refer to the previous exercise for a description of the data set. Estimate the logistic regression model where insurance coverage depends on premium percentage and income. Consider an individual with an income of $60,000. What is the probability that she has insurance coverage if her employer contributes 50% of the premium? What if her employer contributes 75% of the premium?

12. FILE ***CFA.*** The Chartered Financial Analyst (CFA) designation is the de facto professional certification for the financial industry. Historically, the pass rate is higher for those with work experience and a good college GPA. The accompanying data file includes information on 263 current employees who took the CFA Level I exam last year, including the employee's success on the exam (1 for pass, 0 for fail), the employee 's college GPA, and years of work experience.
    a. Estimate the linear probability model to predict the probability of passing the CFA Level I exam for a candidate with a college GPA of 3.80 and five years of experience.
    b. Estimate the logistic regression model to predict the probability of passing the CFA Level I exam for a

candidate with a college GPA of 3.80 and five years of experience.

13. FILE ***STEM.*** Several studies have reported lower participation in the science, technology, engineering, and mathematics (STEM) careers by female and minority students. The accompanying data file includes information on whether the student has applied to a STEM field (1 if S TEM, 0 otherwise), whether or not the student is female (1 if female , 0 otherwise), white (1 if white , 0 otherwise), and Asian (1 if Asian, 0 otherwise). Also included in the survey is the information on the student's high school GPA and the SAT scores.
    a. Estimate and interpret the logistic regression model using STEM as the response variable, and GPA, SAT, White, Female, and Asian as the predictor variables.
    b. Find the predicted probability that a white male student will apply to a STEM field with GPA = 3.4 and SAT = 14 00. Find the corresponding probabilities for an Asian male and a male who is neither white nor Asian.
    c. Find the predicted probability that a white female student will apply to a STEM field with GPA = 3.4 and SAT = 1400. Find the corresponding probabilities for an Asian female and a female who is neither white nor Asian.

14. FILE ***Complication.*** A logistic regression is estimated to analyze the probability of complications for male patients resulting from a serious infection. Predictor variables include the patient's weight, age, and whether he is diabetic (Diabetes equals 1 if diabetic, 0 otherwise). The accompanying data file includes information on 260 males who had tested positive for a serious infection.
    a. Estimate the logistic regression model using Complication as the response variable and Weight, Age, and Diabetes as the predictor variables.
    b. Find the predicted probability of complications for a 60-year-old diabetic male with weight equal to 160 pounds.
    c. Find the corresponding probability for a nondiabetic male.

15. FILE ***Subscription.*** Amanda Chen is the manager of a subscription firm that sells products on a scheduled time basis. She would like to analyze the impact on subscription due to a recent marketing initiative where customers are offered a varying one-time percentage discount (Discount). The response variable Subscribe equals 1 if the customer subscribes and 0 otherwise. Other predictor variables used for the analysis are the customers' age and sex. The accompanying file contains relevant data.
    a. Estimate the logistic regression model. Which predictor variables are statistically significant at the 5% level?
    b. Predict the subscription probabilities for 30-year-old male and female customers who receive a discount of 10 %.
    c. Find the corresponding subscription probabilities for 50-year-old male and female customers who receive a discount of 10%.

16. FILE ***Default.*** Automobiles, including cars and light trucks, are the most held nonfinancial assets among Americans, which are often financed through loans. Peter Firsov works in the car loan division of a major commercial bank. He wants to use a data-driven strategy to identify borrowers who are likely to default on the automobile loan. The accompanying file includes historical data for 400 customers including information on whether the customer defaulted and the corresponding loan-to-value ratio (LTV in %), FICO credit score (FICO), and age. Consider a logistic regression model for default using the predictor variables LTV, FICO, and Age.
    a. Estimate the logistic regression model to find the predicted probability of default given the average values of LTV, FICO, and Age.
    b. Extend the model to also include Age-squared to find the corresponding predicted probability of default.
    c. Is the Age-squared variable statistically significant at the 5% level?

17. FILE ***Membership.*** The manager is concerned about the attrition rate at her gym. She would like to identify the profile of loyal members who retain the membership for at least one year. The accompanying data file contains information on whether the member has been loyal (Loyal = 1 if the member stayed at the gym for at least one year, 0 otherwise). Also included are the member's age and income (in $1,000s) and whether he/she joined on a single or a family plan.
    a. Estimate the logistic regression model using Loyal as the response variable and Age, Income, and Single (equals 1 i f on a single plan, 0 otherwise) as the predictor variables.
    b. Predict the probability of being loyal for a 40- and a 60-year-old member with income of $80,000 and on a single plan.
    c. Predict the corresponding probabilities for a member on a family plan.

## 9.2 ODDS AnD AccuRAcy RATE

In Section 9.1, we discussed binary choice models where the response variable is a dummy variable. Consider a binary choice model with predictor variables $x_1, x_2, \ldots, x_k$ where $b_j$ is the estimated coefficient for $x_j$. For ease of exposition, we refer to $b_j$ as the estimated coefficient for both linear and logistic regressions. In the linear probability model (LPM), $b_j$ measures the partial effect of $x_j$. In other words, it measures the change in the predicted probability $\hat{p}$ for a one-unit increase in $x_j$, holding the other predictor variables constant. In the logistic regression model, however, the interpretation of $b_j$ is not straightforward. While $b_j$ conveys whether the relationship between $x_j$ and $\hat{p}$ is positive or negative, it does not imply a unique partial effect of $x_j$ on $\hat{p}$.

Recall from Chapter 8 that a useful way to communicate the influence of a predictor variable in nonlinear models is to plot the simulated values of this variable against the predicted value of the response variable, while holding the other variables constant. This approach is illustrated for binary choice models in Example 9.3.

## EXAMPLE 9.3

**FILE**
*Spam*

The objective outlined in the introductory case is to detect spam based on the number of recipients, the number of hyperlinks, and the number of characters for each e-mail. Use the data in Table 9.1 for the following analysis.

**a.** Estimate the linear probability model and the logistic regression model for spam detection.

**b.** Use simulations to highlight the partial effect of Recipients on the predicted spam probability, while holding Hyperlinks and Characters fixed at their sample means.

**SOLUTION:**

**a.** Table 9.5 shows the results of the estimated LPM and the logistic regression model.

**TABLE 9.5** Estimated Models for Spam Detection

| Variable | LPM | Logistic |
|---|---|---|
| constant | −0.1408<br>(0.094) | −3.8243*<br>(0.000) |
| Recipients | 0.0158*<br>(0.000) | 0.1075*<br>(0.001) |
| Hyperlinks | 0.0879*<br>(0.000) | 0.5133*<br>(0.000) |
| characters | −0.0020*<br>(0.006) | −0.0141*<br>(0.004) |

Notes: Parameter estimates are followed with the *p*-values in parentheses;
* represents significance at the 5% level.

For both models, Recipients and Hyperlinks are positively related with Spam, whereas Characters is negatively related with Spam. Also, given the small *p*-values in both models, all predictor variables are statistically significant at the 5% level.

**b.** For illustration, we highlight the partial effect of Recipients on the predicted probability by varying it between 10 and 50 while holding Hyperlinks and Characters at their average sample values of 6.226 and 58.602, respectively. Table 9.6 shows a portion of the predicted probabilities.

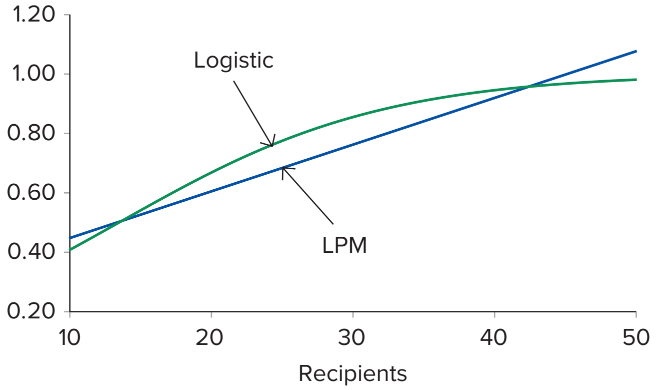
**TABLE 9.6** Predicted Spam Probabilities for Example 9.3

| Recipients | LPM | Logistic |
|---|---|---|
| 10 | 0.4462 | 0.4060 |
| 20 | 0.6039 | 0.6670 |
| 30 | 0.7616 | 0.8545 |
| 40 | 0.9193 | 0.9451 |
| 50 | 1.0770 | 0.9806 |

Note that unlike LPM, the logistic regression model ensures that the predicted probabilities are in the [0, 1] interval. As discussed earlier, the regression coefficients are easy to interpret for LPM but not for logistic regression. For LPM, the predicted spam probability from a 10-point increase in Recipients is 0.1577, irrespective of whether the increase is from 10 to 20 (= 0.6039 − 0.4462) or from 40 to 50. This impact can also be inferred from the LPM coefficient of 0.0158 multiplied by 10; the slight difference is due to rounding. For the logistic model, however, the coefficient of 0.1075 is not easy to interpret except that it implies a positive influence of Recipients on the predicted probability. Table 9.6 shows that as the number of recipients increases from 10 to 20, the probability increases by 0.2610 (= 0.6670 − 0.4060). The same 10-point increase in Recipients from 40 to 50 increases the probability by only 0.0355.

It is informative to use simulations to highlight the partial effect of a predictor variable. In fact, it is better to communicate the effect with a graph. Figure 9.4 shows a plot of the predicted spam probabilities in Table 9.6.

**FIGURE 9.4** Predicted spam probabilities for Example 9.3



The graph for the logistic regression model is relatively steep initially before tapering off, which implies that any increase in Recipients at lower values raises the predicted probability by more than the same increase in Recipients at higher values. This result, however, holds true only when Hyperlinks and Characters are fixed at their average values.

It is important to note that while simulations are informative, they vary depending on the values of the predictor variables. The predicted spam probabilities will change if Hyperlinks and/or Characters are fixed at values other than their sample means. In other words, there is no unique way to highlight the partial effect with logistic regression. Students are encouraged to plot the predicted probabilities by holding Hyperlinks and Characters constant at some non-mean values. They are also encouraged to highlight the partial effect of Hyperlinks and Characters, one at a time, while holding the remaining two predictor variables constant at some reasonable values.

## The Odds of an Event

LO 9.2

Calculate and interpret odds and accuracy.

As discussed, the estimated coefficient of a logistic regression model does not allow us to determine the partial effect of a predictor variable on the probability. It is often preferable to interpret logistic regression coefficients in terms of odds rather than probabilities.

**Odds** are defined as the ratio of the probability of success and the probability of failure. In other words, odds $= \frac{p}{1-p}$, where $p = P(y = 1)$ and the class designated as 1 is the target or positive class. Note that probability ranges between 0 and 1, whereas odds range between 0 and infinity.

The odds metric is especially popular in gambling, sports, and epidemiology. In sports, for example, if the odds of winning are 3, also referred to as 3 to 1 odds, it implies that the team has 3 out of 4 chances of winning. We can use odds to compute the probability of winning as $p = \frac{\text{odds}}{1 + \text{odds}} = \frac{3}{4} = 0.75$. Conversely, if we are given $p = 0.75$, we can compute odds $= \frac{0.75}{0.25} = 3$. The transformation from probability to odds and odds to probability is monotonic, implying that if one increases, the other one increases too. For example, if the odds of winning increase to 4, then the probability goes up to $p = \frac{4}{5} = 0.80$. In a similar vein, if the probability of infection goes up from 0.20 to 0.30, then the odds of infection increase from odds $= \frac{0.20}{0.80} = 0.25$ to odds $= \frac{0.30}{0.70} = 0.43$.

### THE RELATIONSHIP BETWEEN ODDS AND PROBABILITIES

- Odds are computed as $\frac{p}{1-p}$, where $p$ is the probability of success.
- We can use odds to compute the probability of success as $p = \frac{\text{odds}}{1 + \text{odds}}$.

So, why do we care about odds in the context of logistic regression? As it turns out, it is easy to interpret the regression coefficients in terms of odds. Let us first derive the odds for a logistic regression model with just one predictor variable; we can easily extend it to include multiple predictor variables. Recall that the predicted probability is computed as $\hat{p} = \frac{\exp(b_0 + b_1x)}{1 + \exp(b_0 + b_1x)}$.

Therefore,

$$1 - \hat{p} = 1 - \frac{\exp(b_0 + b_1x)}{1 + \exp(b_0 + b_1x)} = \frac{1 + \exp(b_0 + b_1x) - \exp(b_0 + b_1x)}{1 + \exp(b_0 + b_1x)}$$

$$= \frac{1}{1 + \exp(b_0 + b_1x)}.$$

The odds can now be derived as odds $= \frac{\hat{p}}{1-\hat{p}} = \exp(b_0 + b_1x)$. Note that the natural log of odds, also referred as the logit, is a linear function of $x$; that is, $\ln(\text{odds}) = b_0 + b_1x$.

Recall that the odds equation resembles the exponential regression model discussed in Chapter 8. The only difference is that the relationship of $x$ is with the odds rather than with the response variable. This implies that $b_1 \times 100$ is the approximate percentage change in the odds when $x$ increases by one unit. The exact percentage change is calculated as $(\exp(b_1) - 1) \times 100$.

### THE PARTIAL EFFECT OF A PREDICTOR VARIABLE ON ODDS

Consider a logistic regression model with predictor variables $x_1, x_2, \ldots, x_k$, where $b_j$ is the estimated coefficient for $x_j$. We interpret $(\exp(b_j) - 1) \times 100$ as the percentage change in the odds when $x_j$ increases by one unit, holding the other predictor variables constant.

Example 9.4 elaborates on this interpretation.

## EXAMPLE 9.4

FILE
*Spam*

Revisit the estimated logistic regression model estimated in Example 9.3. Interpret the partial effect of each predictor variable on the odds of spam.

**SOLUTION:** Table 9.5 shows that the estimated coefficients for Recipients, Hyperlinks, and Characters are 0.1075, 0.5133, and −0.0141, respectively. To interpret the partial effect on the odds, we derive $(\exp(b_j) - 1) \times 100$ for each predictor variable. For Recipients, it is computed as $(\exp(0.1075) - 1) \times 100 = 11.35$, suggesting that the odds of spam increase by 11.35% for every one-unit increase in Recipients, holding the other variables constant.

To elaborate, in Table 9.7, we show the predicted probabilities and the resulting odds when Recipients equals 20 and 21, and Hyperlinks and Characters equal their sample means.

**TABLE 9.7** Computation of Odds for Example 9.4

| Recipients | Probability, $p$ | Odds, $\frac{p}{1-p}$ |
|---|---|---|
| 20 | 0.6670 | 2.0033 |
| 21 | 0.6905 | 2.2307 |

Note that the percentage change in the odds as Recipients increases by one unit can be calculated as $\left(\frac{2.2307 - 2.0033}{2.0033}\right) \times 100 = 11.35\%$. This is identical to the one derived from the regression coefficient. This percentage change would be the same if Recipients increaseds by one unit, say, from 40 to 41.

Similarly, we can derive $(\exp(b_j) - 1) \times 100$ for Hyperlinks and Characters, suggesting that the odds of spam increase by 67.07% for every one-unit increase in Hyperlinks and decrease by 1.40% for every one-unit increase in Characters, holding the other variables constant.

**Note:** The interpretation of the estimated coefficient for a dummy variable in a logistic regression is interesting. Suppose, in the analysis of purchase probability, the regression coefficient for a female dummy variable is 0.038. We can infer that the purchase odds for females are 3.87% higher than for males because $(\exp(0.038) - 1) \times 100 = 3.87$.

## Accuracy of Binary Choice Models

There is no universal goodness-of-fit measure for binary choice models to assess how well the model fits the data. The residual analysis is meaningless because the residual, defined as $e = y - \hat{y}$, equals $1 - \hat{y}$ for $y = 1$ and $-\hat{y}$ for $y = 0$. Therefore, unlike in the case of linear regression, we cannot rely on the goodness-of-fit measures such as the standard error of the estimate $s_e$, the coefficient of determination $R^2$, or adjusted $R^2$ to assess the model.

It is common to assess the performance of binary choice models based on the accuracy rate, defined as the percentage of correctly classified observations. Remember that the response variable is binary, equaling 1 for the target or positive class and 0 for the nontarget or negative class. Therefore, to compute the accuracy rate, we first translate the predicted probabilities in the sample into binary predictions. Using a default cutoff of 0.5, we make the binary prediction $\hat{y}$ as 1 if $\hat{p} \geq 0.5$ and 0 if $\hat{p} < 0.5$, where $\hat{p}$ is the predicted probability. The accuracy rate is then calculated as the number of correct predictions divided by the total number of predictions.

### THE ACCURACY RATE FOR BINARY CHOICE MODELS

Using the default cutoff of 0.5, we make binary predictions $\hat{y}$ as 1 for $\hat{p} \geq 0.5$ and 0 for $\hat{p} < 0.5$ to compute the accuracy rate.

$$\text{Accuracy Rate} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100$$

## EXAMPLE 9.5

FILE
*Spam*

Revisit the spam example to compare the accuracy rates of the estimated LPM and the estimated logistic regression model.

**SOLUTION:** We estimate LPM and the logistic regression model as before. To compute the accuracy rates, we find the predicted spam probabilities for the sample values of Recipients, Hyperlinks, and Characters and convert them into binary predictions. For the first sample value, we have 19 recipients, 1 hyperlink, and 47 characters. We use these values along with unrounded regression coefficients to find the predicted spam probabilities as:

$$\text{LPM:} \quad \hat{p} = -0.1408 + 0.0158 \times 19 + 0.0879 \times 1 - 0.0020 \times 47 = 0.1520$$

$$\text{Logistic:} \quad \hat{p} = \frac{\exp(-3.8243 + 0.1075 \times 19 + 0.5133 \times 1 - 0.0141 \times 47)}{1 + \exp(-3.8243 + 0.1075 \times 19 + 0.5133 \times 1 - 0.0141 \times 47)} = 0.1266$$

Because the predicted values for both models are less than 0.5, their corresponding binary predictions $\hat{y}$ equal 0. Predictions for other sample observations are computed similarly; see Table 9.8 for a portion of these predictions.

**TABLE 9.8** Computing the Accuracy Rates for Example 9.5

| | | | Prediction | | Binary Prediction | |
|---|---|---|---|---|---|---|
| Recipients | Hyperlinks | Characters | LPM | Logistic | LPM | Logistic |
| 19 | 1 | 47 | 0.1520 | 0.1266 | 0 | 0 |
| 17 | 11 | 68 | 0.9574 | 0.9364 | 1 | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 13 | 2 | 32 | 0.1756 | 0.1357 | 0 | 0 |

We find that out of 500 sample records, the binary predictions match the spam values in 394 and 397 of the records for LPM and the logistic regression model, respectively. Therefore, the accuracy rate is 78.80% for LPM and 79.40% for the logistic regression model. Based on the accuracy rate, the logistic regression model is the preferred predictive model.

### Using R to Compute Accuracy Rates

We replicate the results for Example 9.5. In Analytic Solver, you estimate the logistic model as shown in Section 9.1 and then look up the accuracy rate in the LogReg_Training Score worksheet.

**a.** Import the ***Spam*** data into a data frame (table) and label it myData.

FILE
*Spam*

**b.** We use the **lm** and **glm** functions to estimate the linear and the logistic regression models. Enter:

```
Linear_Model <- lm(Spam ∼ Recipients + Hyperlinks + Characters, data =
myData)
Logistic_Model <- glm(Spam ∼ Recipients + Hyperlinks + Characters,
family = binomial, data = myData)
```

**c.** We use the **predict** function to compute the predicted probabilities. Enter:

```
pHatLin <- predict(Linear_Model)
pHatLog <- predict(Logistic_Model, type = “response”)
```

**d.** We use the **ifelse** function to convert probabilities into binary, 1 or 0, values, using a cutoff of 0.5. Enter:

```
yHatLin <- ifelse(pHatLin >= 0.5, 1,0)
yHatLog <- ifelse(pHatLog >= 0.5, 1,0)
```

**e.** We want to find the proportion of times yHatLin, or yHatLog, equals Spam. Recall from Chapter 2 that the double equal sign (==) is used to compare two values; if they are the same, the operator returns 1, and 0 otherwise. We use the **mean** function to compute the proportion of correctly classified observations, which we multiply by 100 to get a percentage. Enter:

```
100*mean(myData$Spam == yHatLin)
100*mean(myData$Spam == yHatLog)
```

R returns 78.8 for the linear probability model and 79.4 for the logistic regression model.

## SYNOPSIS OF INTRODUCTORY CASE

y H Lim/Alamy Stock Photo

The recent upsurge in the volume of unwanted e-mails has created a huge problem on the Internet. For e-mails to be an effective tool for communication, a robust spam filter that detects spam and stops it from getting into people's inboxes is essential. To estimate the probability of spam, the linear probability model and the logistic regression model were explored using a sample of 500 e-mails, of which 258 were spam, and the remaining 242 were legitimate. Predictor variables included the number of recipients for each e-mail and the number of hyperlinks and characters in the message. In both models, the predictor variables were found to be statistically significant at the 5% level. The number of recipients and the number of hyperlinks were positively associated with spam, whereas the number of characters was negatively associated with spam.

Several scenarios were created to gauge the partial effects of the predictor variables in the logistic regression model. It was found, for example, that a 10-unit increase in the number of recipients from 10 to 20 resulted in a much greater increase in the spam probability than the same increase from 40 to 50. To gain further insights, the predictor variables were related to odds, defined as the ratio of the probabilities of spam and no spam. It was found that the odds of spam increased by 11.35% for every one-unit increase in the number of recipients, holding the other variables fixed. Similarly, the odds increased by 67.07% for every one-unit increase in the number of hyperlinks and decreased by 1.40% for every one-unit increase in the number of characters. Finally, the logistic regression model was found to be superior because it made correct predictions 79.40% of the time as compared to 78.80% for the linear model.

## EXERCISES 9.2

### Mechanics

18. FILE ***Exercise_9.18.*** The accompanying data file contains 20 observations for a binary response variable $y$ along with the predictor variables $x_1$ and $x_2$.
    a. Estimate the logistic regression model to compute the probability and the odds when $x_1 = 12$ and $x_2 = 8$.
    b. What is the percentage change in the odds when $x_1$ increases by one unit, holding $x_2$ constant?

19. FILE ***Exercise_9.19.*** The accompanying data file contains 20 observations for a binary response variable $y$ along with the predictor variables $x_1$ and $x_2$.
    a. Estimate the linear probability model and the logistic regression model.
    b. Compute the accuracy rates of both models.
    c. Use the preferred model to compute $\hat{y}$ for $x_1 = 60$ and $x_2 = 18$ .

### Applications

20. FILE ***Default.*** The accompanying file contains historical data for 400 customers related to automobile loans. The information includes whether the customer defaulted on the loan (Default = 1 for default, 0 otherwise) and the corresponding loan-to-value ratio (LTV in %), FICO credit score (FICO), and customer age.
    a. Estimate the logistic regression model to find the probability and odds of default given LTV = 70, FICO = 600, and Age = 50.
    b. Find the corresponding probability and odds if FICO = 700.
    c. Interpret the percentage change in the default odds when each predictor variable increases by one unit, holding the other variables constant.

21. FILE ***Complication.*** A logistic regression model is estimated to analyze the probability of complications for male patients resulting from a serious infection. Predictor variables include the patient's weight and age and whether he is diabetic (Diabetes equals 1 if diabetic, 0 otherwise). The accompanying data file includes information on 260 male patients who had tested positive for a serious infection.

   a. Estimate the logistic regression model to find the odds of complications for a 60-year-old diabetic patient with a weight of 180 pounds.

   b. Find the corresponding odds if the patient is not diabetic.

   c. What is the percentage difference in the odds for a diabetic patient compared to a nondiabetic patient, holding the other variables constant?

22. FILE ***Subscription.*** Use the accompanying data file to analyze subscription (Subscribe equals 1 if a customer subscribes, 0 otherwise). Also included in the file are the percentage discount (Discount) and the customer's age and sex.

   a. Estimate the logistic regression model to predict the odds of subscription for 50-year-old male and female customers who receive a discount of 10%.

   b. What is the percentage difference in the subscription odds for a male customer compared to a female customer, holding discount and age constant?

23. FILE ***Membership.*** Consider the accompanying data file to estimate the logistic regression model for predicting loyalty (Loyal equals 1 if the member stayed at the gym for at least one year, 0 otherwise). Predictor variables include the member's age and income (in $1,000s) and whether he/she joined on a single plan (Single equals 1 if on a single plan, 0 otherwise).

   a. Use the estimated model to predict the loyalty odds for a 50-year-old member with income of $80,000 and on a single plan.

   b. Predict the corresponding odds for a member on a family plan.

   c. Interpret the percentage change in the loyalty odds when each predictor variable increases by one unit, holding the other variables constant.

24. FILE ***Interview.*** A recent study analyzed the influence of looks on a candidate's chances of getting called for an interview. Interestingly, more interviews were granted to plain-looking women than to attractive women. The exact opposite was true for men, who benefited from the beauty premium. An experiment was performed on 120 college graduates where the response variable, Call, equaled Yes if the applicant was called for an interview, No otherwise. Predictor variables included GPA, Male (1 for male , 0 otherwise), Looks (1 for good looks, 0 otherwise), and the interaction between Male and Looks. The accompanying file includes relevant data.

   a. Estimate the logistic regression model to predict the odds of interview for male and female applicants with a GPA of 3.5 and with good looks.

   b. Find the corresponding odds of interview for male and female applicants without good looks.

25. FILE ***Divorce.*** Divorce has become an increasingly prevalent part of American society. In general, the acceptability of divorce is higher for younger adults and those who are not very religious. In a survey, 200 American adults were asked about their opinion on divorce (Acceptable equals 1 if morally acceptable, 0 otherwise). Predictor variables include religiosity (Religious equals 1 if very religious, 0 otherwise) and age . The accompanying file contains relevant data.

   a. Estimate the logistic regression model to predict the odds of acceptability for a very religious 50-year-old adult.

   b. What is the percentage difference in the acceptability for a very religious adult compared to a nonreligious adult, holding age constant?

26. FILE ***Admit.*** Unlike small selective colleges that pay close attention to personal statements, large state universities rely primarily on the applicant's grade point average (GPA) and scores on the SAT test for making admission decisions. The accompanying file includes data for 120 applicants for college admission (Admit equals 1 if admitted, 0 otherwise) along with the applicant's GPA and SAT scores.

   a. Estimate and interpret the appropriate linear probability model and the logistic regression model.

   b. Compute the accuracy rates of both models.

   c. Use the preferred model to predict the probability of admission for an applicant with GPA $= 3.0$ and SAT $= 1400$. What if GPA $= 4.0$?

27. FILE ***Assembly.*** Assembly-line work it is not suited for everybody because it is tedious and repetitive. A production manager would like to predict whether a newly hired worker will stay in the job for at least one year (Stay equals 1 if a new hire stays for at least one year, 0 otherwise). Predictor variables include age, sex (Female equals 1 if female , 0 otherwise), and whether the worker has worked on an assembly line before (Assembly equals 1 if work ed before, 0 otherwise). The accompanying file includes data for 32 assembly-line workers.

   a. Estimate and interpret the linear probability model and the logistic regression model where being on the job one year later depends on Age, Female, and Assembly.

   b. Compute the accuracy rates of both models.

   c. Use the preferred model to predict the probability that a 45-year-old female who has not worked on an assembly line before will still be on the job one year later. What if she has worked on an assembly line before?

28. FILE ***Complication.*** Use the accompanying data file to analyze the probability of complications for male patients resulting from a serious infection. Predictor variables include the patient's weight and age and whether he is diabetic (Diabetes equals 1 if diabetic, 0 otherwise).

a. Estimate the appropriate linear probability and logistic regression models and compute the accuracy rates of both models.

b. Use the preferred model to predict the probability of complications for a 60-year-old diabetic patient with a weight of 180 pounds.

29. FILE ***Default.*** Use the accompanying data file to analyze the default probability for automobile loans. Predictor variables for default include loan-to-value ratio (LTV in %), FICO credit score (FICO), and customer age.

a. Estimate the appropriate linear probability and logistic regression models and compute their accuracy rates.

b. Use the preferred model to predict the default probability with LTV = 80, FICO = 650, and Age = 40.

30. FILE ***Membership.*** Use the accompanying data file to analyze loyalty (Loyal equals 1 if the member stayed at the gym for at least one year, 0 otherwise). Also included in the file are the member's age and income (in $1,000s) and whether he/she joined on a single plan (Single equals 1 if on a single plan, 0 otherwise).

a. Estimate the logistic regression model for loyalty, using age and income as the predictor variables, and compute the accuracy rate.

b. Extend the model to include the membership plan (single or family) and compute the resulting accuracy rate.

c. Use the preferred model to predict the loyalty probability for a 50-year-old member with income of $80,000; consider both single and family plans if you pick the model in part b.

31. FILE ***Subscription.*** Use the accompanying data file to analyze subscription (Subscribe equals 1 if a customer subscribes, 0 otherwise). Also included in the data file are the percentage discount (Discount) and the customer's age and sex.

a. Estimate the logistic regression model for subscription using discount and age as the predictor variables and compute the accuracy rate.

b. Extend the model to include sex and compute the resulting accuracy rate.

c. Use the preferred model to predict the subscription probability for a 50-year-old customer with a discount of 10%; consider both male and female customers if you pick the model in part b.

## 9.3 cROSS-VALIDATIOn OF BIn ARy cHOIcE MODELS

LO 9.3

Use cross-validation to assess classification performance.

In Section 9.2, we used the accuracy rate to assess the performance of classification models. This performance measure was calculated for the sample that was also used in estimation. Sometimes, a model may perform very well with the estimation sample but then perform poorly when making out-of-sample predictions.

Recall from Chapter 8 a useful method to assess the predictive power of a model is to test it on a data set not used in estimation. We use cross-validation to assess models by partitioning the data into a training set to build (train) the model and a validation set to evaluate (validate) it. Although we can assess the performance in the training set, this assessment may be overly optimistic. A validation set provides an independent assessment by exposing the model to unseen data.

In this section, we will first implement cross-validation using the accuracy rate as a performance measure. Later we will extend the analysis to include two more performance measures. Chapter 11 provides a comprehensive treatment of all performance measures used in classification models.

### The Holdout Cross-Validation Method

The holdout method is easy to implement. We start by partitioning the data into two independent and mutually exclusive data sets—the training set, and the validation set. There is no rule as to how the sample data should be partitioned. We generally use random draws when partitioning the data. For illustration purposes, however, we will use the first 75% of the observations for training and the remaining 25% for validation. The steps for the holdout method are:

1. Partition the sample data into two parts, labeled training set and validation set.
2. Use the training set to estimate competing models.

3. Use these estimates to make predictions in the validation set.
4. Calculate the accuracy rates and select the model with the highest value.

Example 9.6 illustrates the holdout method.

## EXAMPLE 9.6

FILE
*Spam*

The objective outlined in the introductory case is to detect spam based on the number of recipients, the number of hyperlinks, and the number of characters for each e-mail.

**a.** Use the holdout method to compare the accuracy rates of two competing logistic regression models for spam, using the first 375 observations for training and the remaining 125 observations for validation. Model 1 uses Recipients, Hyperlinks, and Characters as predictor variables, whereas Model 2 drops the predictor variable(s) found to be statistically insignificant in Model 1.

**b.** Re-estimate the preferred model with all 500 observations to predict the probability of spam if Recipients, Hyperlinks, and Characters are 20, 5, and 60, respectively.

**SOLUTION:**

**a.** We use the training set with 375 observations to estimate two logistic regression models; see Table 9.9 for the estimates. Note that in Model 1, the variable Characters is not statistically significant at the 5% level and is, therefore, dropped in Model 2. (Instructions for conducting the holdout method with Analytic Solver and R are provided shortly.)

**TABLE 9.9** Estimates of the Spam Data ($n = 375$)

| Variable | Model 1 | Model 2 |
|---|---|---|
| constant | −5.1778*<br>(0.000) | −5.8045*<br>(0.000) |
| Recipients | 0.1765*<br>(0.000) | 0.1806*<br>(0.000) |
| Hyperlinks | 0.5473*<br>(0.000) | 0.5402*<br>(0.000) |
| characters | −0.0104<br>(0.074) | nA |

Notes: Parameter estimates are followed with the *p*-values in parentheses; NA denotes not applicable;
* represents significance at the 5% level.

To compute the accuracy rates of the estimated logistic regression models, we first compute the predicted probabilities of the models in the validation set; a portion of the calculations is shown in Table 9.10.

**TABLE 9.10** Analysis of the Holdout Method, Example 9.6

| Record | Spam | Prediction | | Binary Prediction | |
|---|---|---|---|---|---|
| | | Model 1 | Model 2 | Model 1 | Model 2 |
| 376 | 0 | 0.4734 | 0.5367 | 0 | 1 |
| 377 | 0 | 0.6963 | 0.7037 | 1 | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 500 | 1 | 0.1070 | 0.0850 | 0 | 0 |

By converting the predicted probabilities into binary predictions and comparing them with the actual spam data, we find that the accuracy rates for Model 1 and Model 2 are 68.00% and 65.60%, respectively. Based on these rates, we infer that Model 1 is superior to Model 2 (68.00% > 65.60%).

**b.** We re-estimate Model 1 with all 500 observations to predict the spam probability with the number of recipients, hyperlinks, and characters as 20, 5, and 60, respectively, as:

$$\widehat{\text{Spam}} = \frac{\exp(-3.8243 + 0.1075 \times 20 + 0.5133 \times 5 - 0.0141 \times 60)}{1 + \exp(-3.8243 + 0.1075 \times 20 + 0.5133 \times 5 - 0.0141 \times 60)} = 0.51.$$

## Using Analytic Solver  and R for the Holdout Method

FILE
*Spam*

As mentioned earlier, we generally use random draws for partitioning the sample data into the training and the validation sets, which can easily be implemented in Analytic Solver and R. For illustration, we will use the latter part of the sample data for validation. We use the ***Spam*** data to replicate the results in Example 9.6 for Model 1; results for Model 2 can be derived similarly.

### Using Analytic Solver

**a.** Open the ***Spam*** data file. In column F of the data file, create a variable called Flag with the letter T for the first 375 observations and the letter V for the remaining 125 observations; here T and V denote training and validation, respectively.

**b.** From the menu, choose **Data Mining** > **Partition** > **Standard Partition**.

Specify the *Data range* by highlighting cells A1:F501. Select and move variables Spam, Recipients, Hyperlinks, and Characters to the *Selected Variables* box. Although there is an option to pick up rows randomly, for *Partitioning Options*, we select *Use partition variable*. Select and move Flag to the *Use partition variable* box. Click *OK*. The STDPartition worksheet now contains partitioned data with 375 observations in the training set and 125 observations in the validation set.

**c.** Make sure the STDPartition worksheet is active, then choose **Data Mining** > **Classify** > **Logistic Regression**. Select and move variables Recipients, Hyperlinks, and Characters to the *Selected Variables* box and Spam to the *Output Variable* box. Click *Finish*. In the LogReg_ValidationScore worksheet, you will find Accuracy (%correct) equal to 68. This is the same as derived for Model 1 in Example 9.6.

### Using R

**a.** Import the ***Spam*** data into a data frame (table) and label it myData.

**b.** We partition the sample into training and validation sets, labeled TData and VData, respectively. Enter:

```
TData <- myData[1:375,]
VData <- myData[376:500,]
```

**c.** We use the training set, TData, to estimate Model 1. Enter:

```
Model1 <- glm(Spam ~ Recipients+Hyperlinks+Characters, family=binomial,
data = TData)
```

**d.** We use the estimates to make predictions for VData and then convert them into a binary prediction. Finally, we compute the accuracy rate in the validation set. Enter:

```
pHat1 <- predict(Model1, VData, type="response")
yHat1 <- ifelse(pHat1 >= 0.5, 1,0)
100*mean(VData$Spam == yHat1)
R returns: 68
```

This is the same as derived for Model 1 in Example 9.6.

## The *k*-Fold Cross-Validation Method

Recall from Chapter 8 that the holdout method is sensitive to how the data are partitioned. Often it is preferable to use the *k*-fold cross-validation method, where we partition the data into *k* subsets, and the one that is left out in each iteration is the validation set. In other words, we perform the holdout method *k* times and use the average of the performance measures for model selection. Note that the greater the *k*, the greater will be the reliability and the greater will be its computational cost. When *k* equals the sample size, the resulting method is also called the leave-one-out cross-validation method, where you leave out just one observation for the validation set.

Example 9.7 illustrates the *k*-fold cross-validation method.

### EXAMPLE 9.7

**FILE**
***Spam***

Revisit the ***Spam*** data to apply the *k*-fold cross validation method, with $k = 4$, to compare the predictability of the two logistic regression models discussed in Example 9.6.

**SOLUTION:** We assess both models four times with the validation set formed by the observations 376–500, 251–375, 126–250, and 1–125, respectively. Each time the training set includes the remaining observations. In Table 9.11, we present the accuracy rates for each validation set and the average of the four accuracy rates.

**TABLE 9.11** The Accuracy Rates for *k*-Fold Cross-Validation with $k = 4$

| Observations in the validation set | Model 1 | Model 2 |
|---|---|---|
| 376–500 | 68.0% | 65.6% |
| 251–375 | 76.8% | 76.8% |
| 126–250 | 82.4% | 80.0% |
| 1–125 | 84.0% | 80.8% |
| Average accuracy | 77.8% | 75.8% |

The average accuracy rate for Model 1 is 77.8%, compared to 75.8% for Model 2. Consistent with the results of the holdout method, we conclude that Model 1 is superior for making predictions.

The *k*-fold cross-validation method is implemented by using the holdout method *k* times and taking the average of the *k* performance measures. The only thing we change in each iteration is the composition of the training and the validation sets. In Analytic Solver, it involves changing the flag variable in step a and in R it involves changing TData and VData in step b.

**Note:** In Appendix 9.1, we show how the *caret* package in R can be used to easily implement the *k*-fold cross-validation method. The package uses random partitioning of the data, as opposed to fixed partitioning used in this section.

## Other Performance Measures and Imbalanced Data

Accuracy is one of the most common metrics used to evaluate the performance of binary choice models. Although the accuracy measure is useful, we may want to assess how well the model predicts the target class and the nontarget class. Therefore, in addition to accuracy, it is important to report the percentage of correctly classified observations for both the target class and the nontarget class, referred to as sensitivity and specificity.

### SENSITIVITY AND SPECIFICITY

- Sensitivity is the proportion of target class cases that are classified correctly.
- Specificity is the proportion of nontarget class cases that are classified correctly.

### EXAMPLE 9.8

FILE
*Spam*

In Example 9.6, we used the holdout method to compare two logistic regression models using the first 375 observations for training and the remaining 125 observations for validation. Recall that Model 1 uses Recipients, Hyperlinks, and Characters as predictor variables, whereas Model 2 uses only Recipients and Hyperlinks. Revisit the data to compare the models in terms of accuracy, sensitivity, and specificity.

**SOLUTION:** The performance measures for the two logistic regression models, computed in the validation set, are shown in Table 9.12. (Analytic Solver reports the sensitivity and specificity measures, along with the accuracy rate, in the LogReg_ValidationScore worksheet; the R instructions are provided shortly.)

**TABLE 9.12** Performance Measures for Example 9.8

| Measure | Model 1 | Model 2 |
|---|---|---|
| Accuracy | 68.00% | 65.60% |
| Sensitivity | 72.13% | 67.21% |
| Specificity | 64.06% | 64.06% |

We prefer Model 1 because in addition to its higher accuracy, also reported in Example 9.6, it also has higher sensitivity; specificity for both models is the same.

### Using R to Compute Sensitivity and Specificity

We derive the results for Model 1 in Example 9.8.

FILE
*Spam*

**a.** Import the ***Spam*** data into a data frame (table) and label it myData.

**b.** We partition the sample into training and validation sets. Enter:

```
TData <- myData[1:375,]
VData <- myData[376:500,]
```

**c.** We estimate Model 1 using the training set. Enter:

```
Model1 <- glm(Spam ~ Recipients+Hyperlinks+Characters, family=binomial,
data = TData)
```

**d.** We make predictions for VData and then convert them into binary predictions. Enter:

```
pHat1 <- predict(Model1, VData, type = “response”)
yHat1 <- ifelse(pHat1 >= 0.5, 1,0)
```

**e.** We create binary values to identify true positives when the target class is correctly predicted and true negatives when the nontarget class is correctly predicted. Enter:

```
yTP1 <- ifelse(yHat1 == 1 & VData$Spam == 1, 1, 0)
yTN1 <- ifelse(yHat1 == 0 & VData$Spam == 0, 1, 0)
```

**f.** Finally, we compute accuracy, sensitivity, and specificity. Enter:

```
100*mean(VData$Spam == yHat1)
100*(sum(yTP1)/sum(VData$Spam==1))
100*(sum(yTN1)/sum(VData$Spam==0))
```

R returns 68, 72.13, and 64.06 for accuracy, sensitivity, and specificity, respectively.

### Imbalanced Data

The relevance of sensitivity and specificity is especially true in applications where the response variable has many ones and a few zeros, or many zeros and a few ones. The data set for such applications is referred to as imbalanced. With imbalanced data, the percentage of correctly classified observations can be high even when the model predicts the less likely outcome poorly. There is some degree of imbalance in all data sets, but a severe imbalance is challenging for assessing performance. Applications of severely imbalanced data include fraud detection, default detection, insurance claim detection, and rare disease detection.

Consider a logistic regression model applied for fraud detection, where the response variable equals 1 for fraud (target class) and 0 otherwise (nontarget class). Here, we are interested primarily in identifying fraudulent cases where the number of fraudulent cases accounts for a very small percentage (say 1%) of the total number of cases. A model that classifies all the nonfraudulent cases correctly but misses most of the fraudulent cases is not very useful despite having an extremely high accuracy rate.

The relevance of sensitivity and specificity depends on the related misspecification costs. If the misspecification costs due to incorrectly predicting the target class are high (e.g., not correctly predicting fraudulent cases), then the default cutoff of 0.50 for converting predictions into binary predictions may not be appropriate. Sometimes, it is preferable to use a cutoff that equals the proportion of observations in the target class. Continuing with the fraud detection example, we set the cutoff value to equal the proportion of fraudulent cases (say, 0.01). This improves the sensitivity of the model, even though the accuracy may be compromised. Example 9.9 elaborates on this approach.

**FILE**
*Depression*

### EXAMPLE 9.9

The transition from high school to college can be daunting for young adults. For some students, college anxieties turn into clinical depression with possible thoughts of suicide. Early identification and intervention are critical for reducing the adverse effects of depression. The human resource department at a community college has developed a logistic regression model to identify students who are likely to develop clinical depression based on their GPA, attendance in the freshman year, and sex. Table 9.13 shows a portion of the data on depression (Depression equals 1 for clinical depression, 0 otherwise), GPA, attendance (in %), and sex (Female equals 1 for female, 0 otherwise).

**TABLE 9.13** Depression Data ($n = 300$)

| Student | Depression | GPA | Attendance | Female |
|---|---|---|---|---|
| 1 | 0 | 2.5 | 54 | 1 |
| 2 | 1 | 1.9 | 31 | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 300 | 0 | 1.5 | 75 | 1 |

**a.** Use the holdout method, with the default cutoff of 0.50, to compute accuracy, sensitivity, and specificity with the first 225 observations for training and the remaining 75 observations for validation.

**b.** Compute the corresponding performance measures using the cutoff equal to the proportion of students with clinical depression in the training set.

**c.** Which cutoff is preferred in this application? Explain.

**SOLUTION:**

**a.** The performance measures with the default cutoff of 0.50 are shown in the second column of Table 9.14. Note that the model has excellent accuracy and specificity, but with sensitivity of 0.50, the model does not do a very good job in identifying students who are likely to develop clinical depression.

**TABLE 9.14** Performance Measures for Example 9.9

| Measure | Cutoff = 0.50 | Cutoff = 0.0711 |
|---|---|---|
| Accuracy | 94.67% | 86.67% |
| Sensitivity | 50.00% | 75.00% |
| Specificity | 97.18% | 87.32% |

**b.** The performance measures with the cutoff equal to 0.0711 (the proportion of students with clinical depression in the training set) are shown in the third column of Table 9.14. Note that sensitivity has jumped from 50% to 75%.

**c.** In this application, where early intervention is critical, improving the probability of identifying students who are likely to develop clinical depression is highly desirable even though it comes at the expense of lower accuracy and specificity. So, the cutoff equal to the proportion of students with clinical depression in the training set is preferred.

**Note:** We make minor tweaks to the computer instructions for part b. In Analytic Solver, when estimating the logistic regression model after partitioning the data, replace the default value of 0.5 in the *Success Probability Cutoff* box with 0.0711. The corresponding performance measures will get populated in the LogReg_ValidationScore worksheet. In R, when making binary predictions, replace the command `yHat <- ifelse(pHat >= 0.5, 1,0)` with `yHat <- ifelse(pHat >= mean(TData$Depression), 1,0)`. The rest of the script remains the same.

# EXERCISES 9.3

## Mechanics

32. FILE ***Exercise_9.32.*** The accompanying data file contains 40 observations for a binary response variable $y$ along with the predictor variables $x_1$ and $x_2$. Use the holdout cross-validation method to compare the accuracy rates of the linear probability model with the logistic regression model using the first 30 observations for training and the remaining 10 observations for validation.

33. FILE ***Exercise_9.33.*** The accompanying data file contains 100 observations for a binary response variable $y$ along with the predictor variables $x_1$ and $x_2$.
    a. Use the holdout cross-validation method to compare the accuracy rates of two logistic regression models for $y$, using the first 75 observations for training and the remaining 25 observations for validation. For predictor variable(s), Model 1 uses $x_1$ and Model 2 uses $x_1$ and $x_2$.
    b. Re-estimate the preferred model with all 100 observations to predict the probability of success when $x_1 = 25$ and $x_2 = 50$.

34. FILE ***Exercise_9.34.*** The accompanying data file contains 100 observations for a binary response variable $y$ along with the predictor variables $x_1$ and $x_2$. Use the $k$-fold cross-validation method with $k = 4$ to compare the accuracy rates of two logistic regression models for $y$. For predictor variable(s), Model 1 uses $x_1$ and Model 2 uses $x_1$ and $x_2$.

35. FILE ***Exercise_9.35.*** The accompanying data file contains 100 observations for a binary response variable $y$ along with the predictor variables $x_1$ and $x_2$. Use the holdout method, with the first 75 observations for training and the remaining 25 observations for validation, to compute and interpret accuracy, sensitivity, and specificity of the logistic regression model for $y$.

## APPLICATIONS

36. FILE ***Subscription.*** Consider the accompanying data file to predict subscription (Subscribe equals 1 if the customer subscribes, 0 otherwise). Predictor variables include the percentage discount (Discount) and the customer's age and sex.
    a. Use the holdout method to compare the accuracy rates of the linear probability model (Model 1) and the logistic regression model (Model 2) using the first 200 observations for training and the remaining 100 observations for validation.
    b. Use the $k$-fold cross-validation method to compare the accuracy rate of the models using $k = 3$.

37. FILE ***Purchase.*** Consider the accompanying data file to predict Under Armour purchases (Purchase; 1 for purchase, 0 otherwise) based on a customer's age.
    a. Use the holdout method to compare the accuracy rates of the linear probability model (Model 1) and the logistic regression model (Model 2) using the first 20 observations for training and the remaining 10 observations for validation.
    b. Use the $k$-fold cross-validation method to compare the accuracy rate of the models using $k = 3$.

38. FILE ***Interview.*** Consider the accompanying data file for predicting an interview call (Yes or No). Use the $k$-fold cross-validation method, with $k = 4$, to compare the accuracy rates of two logistic regression models. Predictor variables for Model 1 include GPA, Male (1 for male, 0 otherwise), and Looks (1 for good looks, 0 otherwise). Model 2 also includes the interaction between Male and Looks.

39. FILE ***Divorce.*** Consider the accompanying data file to analyze how people view divorce (Acceptable equals 1 if morally acceptable, 0 otherwise) based on age and religiosity (Religious equals 1 if very religious, 0 otherwise). Use the $k$-fold cross-validation method, with $k = 4$, to compare the accuracy rates of two logistic regression models for divorce. Model 1 uses Age and Religious as predictor variables, whereas Model 2 also includes the interaction between Age and Religious.

40. FILE ***Membership.*** Consider the accompanying data file to estimate the logistic regression model for predicting loyalty (Loyal equals 1 if the member stayed at the gym for at least one year, 0 otherwise). Predictor variables include the member's age and income (in $1,000s) and whether he/she joined on a single plan (Single equals 1 if on a single plan, 0 otherwise). Use the holdout method, using the first 150 observations for training and the remaining 50 observations for validation, to calculate and interpret the accuracy, sensitivity, and specificity measures.

41. FILE ***Admit.*** Consider the accompanying data file to estimate the logistic regression model for predicting college admission (Admit equals 1 if admitted, 0 otherwise). Predictor variables include the applicant's grade point average (GPA) and scores on the SAT test. Use the holdout method, using the first 90 observations for training and the remaining 30 observations for validation, to calculate and interpret the accuracy, sensitivity, and specificity measures.

42. FILE ***Complication.*** Use the accompanying data file to estimate the logistic regression model for predicting the probability of complications for male patients resulting from a serious infection. Predictor variables include the patient's weight and age and whether he is diabetic (Diabetes equals 1 if diabetic, 0 otherwise).
    a. Use the holdout method, with the cutoff of 0.50, to compute accuracy, sensitivity, and specificity with the first 180 observations for training and the remaining 60 observations for validation.
    b. Compute the corresponding performance measures using the cutoff equal to the proportion of patients with complications in the training set.
    c. Which cutoff is preferred in this application? Explain.

43. FILE ***Default.*** Use the accompanying data file to estimate the logistic regression model for predicting the probability of loan default (Default equals 1 for default, 0 otherwise). Predictor variables include loan-to-value ratio (LTV in %), FICO credit score (FICO), and customer age.
    a. Use the holdout method, using the cutoff of 0.50, to compute accuracy, sensitivity, and specificity with the first 300 observations for training and the remaining 100 observations for validation.
    b. Compute the corresponding performance measures using the cutoff equal to the proportion of loans with default in the training set.
    c. Which cutoff is preferred in this application? Explain.

## 9.4 WRITING WITH BIG DATA

### c ase Study

c reate a sample report to analyze admission and enrollment decisions at the school of arts & letters in a selective four-year college in n orth America. For predictor variables, include the applicant's sex, ethnicity, grade point average, and SAT scores. Make predictions for the admission probability and the enrollment probability using typical values of the predictor variables. Before estimating the models, you have to first filter out the ***College_Admission*** data to get the appropriate subset of observations for selected variables.

**FILE**
***College_Admission***

### Sample Report— College Admission and Enrollment

Rawpixel.com/Shutterstock

c ollege admission can be stressful for both students and parents as there is no magic formula when it comes to admission decisions. Two important factors considered for admission are the student's high school record and performance on standardized tests.

Just as prospective students are anxious about receiving an acceptance letter, most colleges are concerned about meeting their enrollment targets. The number of acceptances a college sends out depends on its enrollment target and admissions yield, defined as the percentage of students who enroll at the school after being admitted. It is difficult to predict admissions yield as it depends on the college's acceptance rate as well as the number of colleges to which students apply.

In this report, we analyze factors that affect the probability of college admission and enrollment at a school of arts & letters in a selective four-year college in n orth America. Predictors include the applicant's high school GPA, SAT score,[1] and the Male, White, and Asian dummy variables capturing the applicant's sex and ethnicity. In Table 9.15, we present the representative applicant profile.

**TABLE 9.15** Applicant Profile for the School of Arts & Letters

| Variable | Applied | Admitted | Enrolled |
|---|---|---|---|
| Male applicant (%) | 30.76 | 27.37 | 26.68 |
| White applicant (%) | 55.59 | 61.13 | 69.83 |
| Asian applicant (%) | 12.42 | 11.73 | 8.73 |
| Other applicant (%) | 31.99 | 27.14 | 21.45 |
| High school GPA (Average) | 3.50 | 3.86 | 3.74 |
| SAT score (Average) | 1,146 | 1,269 | 1,229 |
| n umber of applicants | 6,964 | 1,739 | 401 |

Of the 6,964 students who applied to the school of arts & letters, 30.76% were males; in addition, the percentages of white and Asian applicants were 55.59% and 12.42%, respectively, with about 32.00% from other ethnicities. The average applicant had a GPA of 3.50 and an SAT score of 1146. Table 9.15 also shows that 1,739 (or 24.97%) applicants were granted

[1] The higher of SAT and ACT scores is included in the data where, for comparison, ACT scores on reading and math are first converted into SAT scores.

admission, of which 401 (23.06%) decided to enroll. As expected, the average GPA and SAT scores of admitted applicants are higher than those who applied and those who enrolled, but to a lesser extent.

Two logistic regression models are estimated using the same predictor variables, one for predicting the admission probability and the other for predicting the enrollment probability. The entire pool of 6,964 applicants is used for the first regression, whereas 1, 739 admitted applicants are used for the second regression. The results are presented in Table 9.16.

**TABLE 9.16** Logistic Regressions for College Admission and Enrollment

| Variable | Admission | Enrollment |
|---|---|---|
| c onstant | −17.5732*<br>(−37.41) | 7.2965*<br>(8.48) |
| Male dummy variable | 0.0459<br>(0.61) | −0.1433<br>(−1.05) |
| White dummy variable | −0.3498*<br>(−4.43) | 0.7653*<br>(5.15) |
| Asian dummy variable | −0.4140*<br>(−3.57) | −0.0074<br>(−0.03) |
| High school GPA | 2.7629*<br>(25.74) | −1.4265*<br>(−7.17) |
| SAT score | 0.0056*<br>(20.93) | −0.0028*<br>(−5.99) |
| | | |
| Accuracy (%) | 81 | 77 |
| n umber of observations | 6,964 | 1,739 |

*Notes:* Parameter estimates are in the top half of the table with the *z*-statistics given in parentheses; * represents significance at the 5% level. Accuracy (%) measures the percentage of correctly classified observations.

With accuracy rates of 81% and 77%, respectively, both models do a good job with predicting probabilities. It seems that the sex of the applicant plays no role in the admission or enrollment decisions. Interestingly, both white and Asian applicants have a lower probability of admission than those from other ethnicities. A higher admission rate for underrepresented applicants is consistent with the admission practices at colleges that believe that diversity enriches the educational experience for all. As expected, quality applicants, in terms of both GPA and SAT, are pursued for admission.

On the enrollment side, admitted applicants who are white are more likely to enroll than all other admitted applicants. c onsider the case of a representative male applicant with a GPA of 3.8 and an SAT score of 1300. For a white male, the predicted probabilities of admission and enrollment are 47% and 24%, respectively. The corresponding probabilities are 45% and 1 3%, respectively, for Asians, and 55% and 13%, respectively , for all other ethnicities.

The lower admission yield for underrepresented applicants is noteworthy. Perhaps the college should explore the reasons for the low yield and also find ways to raise it. Finally, admitted applicants with high GPA and high SAT scores are less likely to enroll at this college. This is not surprising because academically strong applicants have many offers, which lowers the probability that an applicant will accept the admission offer of a particular college.

## Suggested c ase Studies

Many predictive models can be estimated and assessed with the big data that accompany this text. Here are some suggestions.

**Report 9.1** FILE ***COVID_Testing.*** Estimate and interpret a logistic regression model to predict c OVID testing results using the appropriate predictor variables. n ote: you may need to first subset the data based on age, sex, and/or contact due to the data size constraints of software packages.

**Report 9.2** FILE ***Longitudinal_Survey.*** Develop a logistic regression model for predicting if the respondent is outgoing in adulthood. u se cross-validation to select the appropriate predictor variables. In order to estimate this model, you have to first handle missing observations using the missing or the imputation strategy.

**Report 9.3** FILE ***TechSales_Reps.*** The net promoter score (n PS) is a key indicator of customer satisfaction and loyalty. u se data on employees in the software product group with a college degree to develop the logistic regression model for predicting if a sales rep will score an n PS of 9 or more. u se cross-validation to select the appropriate predictor variables. In order to estimate this model, you have to first construct the (dummy) target variable, representing n PS $\geq$ 9 and subset the data to include only the employees who work in the software product group with a college degree.

**Report 9.4** FILE ***Car_Crash.*** Subset the data to include any one county of your choice. Develop a logistic regression model to analyze the probability of a head-on crash using predictor variables such as the weather condition, amount of daylight, and whether or not the accident takes place on a highway. u se the appropriate cutoff point to analyze the accuracy, sensitivity, and specificity of the estimated model.

## APPEnDIX 9.1 The c aret Package in R for the *k*-fold c ross-Validation Method

We introduced the *caret* package in Appendix 8.1 for implementing the *k*-fold cross-validation method. Recall that this package uses random draws for data partitioning and, therefore, the results will vary slightly from the results in Section 9.3 that were based on fixed draws. The *caret* package is used extensively in Chapters 12 and 13.

### Assessing the Logistic Regression Model

For illustration, we use the ***Spam*** data that was used to assess Model 1 and Model 2 in Example 9.7.

FILE
***Spam***

**A.** Import the ***Spam*** data into a data frame (table) and label it myData.

**B.** Install and load the *caret* package (and the e1071 package, if necessary). Enter:

```
install.packages("caret")
library(caret)
```

**C.** We use the **trainControl** and the **train** functions. For options within the **trainControl** function, we use *method* to specify the sampling method (here denoted as "cv" for cross-validation) and *number* to indicate the number of folds. Enter:

```
myControl <- trainControl(method = "cv", number = 4)
```

**D.** Before estimating the logistic regression model, we must convert the binary response variable from numeric type into factor type so that R treats it as a categorical variable with two classes. We use the **as.factor** function to accomplish this task. Enter:

```
myData$Spam <- as.factor(myData$Spam)
```

**E.** Within the **train** function, we specify the model, and then set the following options: *data* (to indicate the data frame), *method* (here “glm” for logistic regression model), and *trControl* (to indicate the variable defined when using the **trainControl** function). Enter:

```
Model1 <- train(Spam ~ Recipients + Hyperlinks + Characters, data =
myData, trControl = myControl, method = “glm”, family = binomial,
metric = “Accuracy”)
Model1
Model2 <- train(Spam ~ Recipients + Hyperlinks, data = myData,
trControl = myControl, method = “glm”, family = binomial, metric =
“Accuracy”)
Model2
```

R reports the accuracy of 78.59% for Model 1 and 76.60% for Model 2 (results will vary). These values are fairly close but not identical to those reported in Table 9.11. Note: Rerunning the R script will again give you a slightly different result because of random partitioning. As in Example 9.7, we find that Model 1 is better for making predictions.

We can implement a four-fold cross-validation that is repeated five times by modifying the earlier function as:

```
myControl <- trainControl(method = “repeatedcv”, number = 4, repeats = 5).
```