BUSINESS ANALYTICS

Communicating with Numbers

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

Published by McGraw Hill LLC, 1325 Avenue of the Americas, New York, NY 10019. Copyright ©2023 by McGraw Hill LLC. All rights reserved. Printed in the United States of America. Previous edition ©2021.

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: *Eric Weber*
Product Developer: *Ryan McAndrews*
Executive Marketing Manager: *Harper Christopher*
Content Project Managers: *Pat Frederickson/Jamie Koch/George Theofanopoulos*
Buyer: *Sandy Ludovissy*
Designer: *Matt Diamond*
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

Names: Jaggia, Sanjiv, 1960- author. | Kelly, Alison (Professor of
economics) author. | Lertwachara, Kevin, author. | Chen, Leida, author.
Title: Business analytics : communicating with numbers / Sanjiv Jaggia,
Alison Kelly, Kevin Lertwachara, Leida Chen.
Description: Second edition. | New York, NY : McGraw Hill LLC, [2023] |
Includes index. | Audience: Ages 18+
Identifiers: LCCN 2021047250 (print) | LCCN 2021047251 (ebook) | ISBN
9781264302802 (hardcover) | ISBN 9781265901295 (ebook)
Subjects: LCSH: Decision making—Mathematical models—Juvenile literature.
| Industrial management—Statistical methods—Computer
programs—Juvenile literature. | Decision making—Computer
programs—Juvenile literature.
Classification: LCC HD30.23 .J344 2023 (print) | LCC HD30.23 (ebook) |
DDC 658.4/033—dc23
LC record available at https://lccn.loc.gov/2021047250
LC ebook record available at https://lccn.loc.gov/2021047251

The Internet addresses listed in the text were accurate at the time of publication. The inclusion of a website does not indicate an endorsement by the authors or McGraw Hill LLC, and McGraw Hill LLC does not guarantee the accuracy of the information presented at these sites.

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

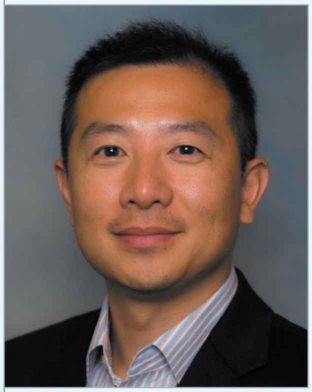
Alison Kelly is a professor of economics at Suffolk University in Boston. Dr. Kelly holds a Ph.D. from Boston College and is a Chartered Financial Analyst (CFA®). Dr. Kelly has published in a wide variety of academic journals and has co-authored three successful textbooks in business statistics and business 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 Leida 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.

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.

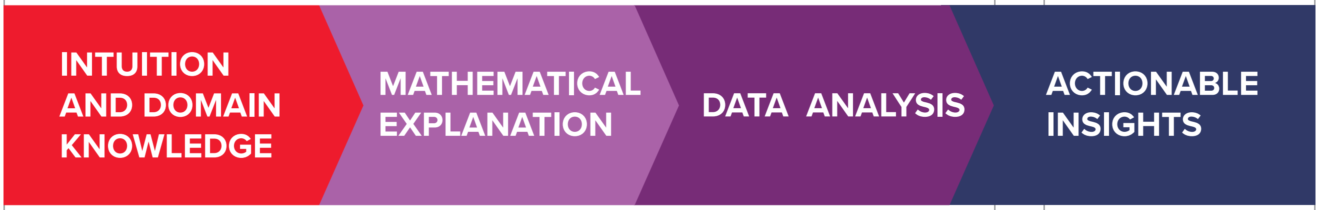
> *"This is by far the best book I have come across. It is easy to follow, very practical, and the examples are rich in detail."*
>
> **Cary Caro,** *Xavier University of Louisiana*

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

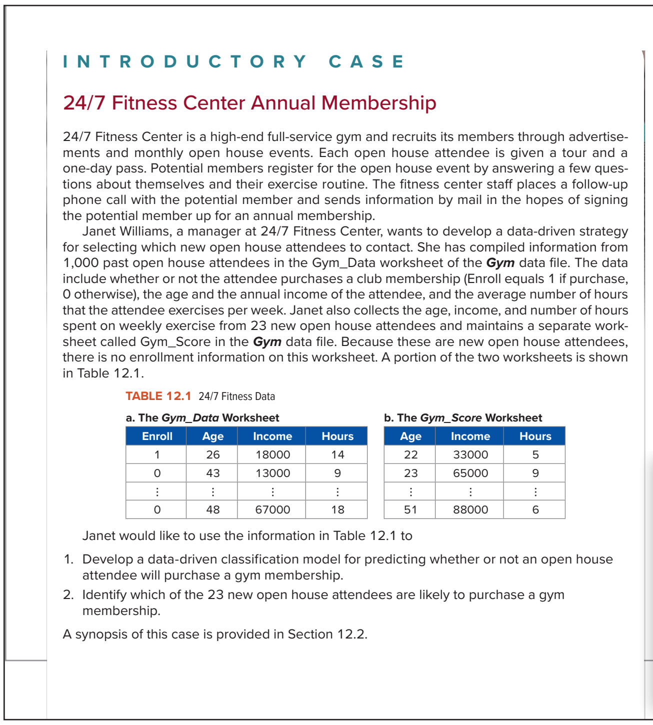
DATA ANALYSIS

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*



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

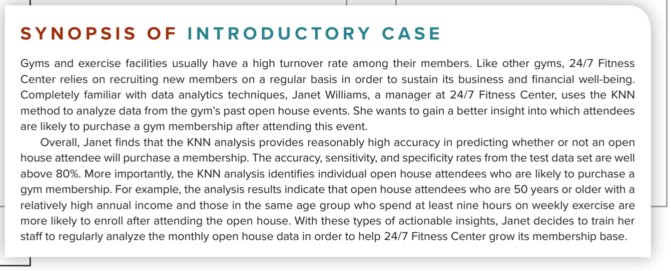
Age	Income	Hours
22	33000	5
23	65000	9
⋮	⋮	⋮
51	88000	6

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.

> *"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*



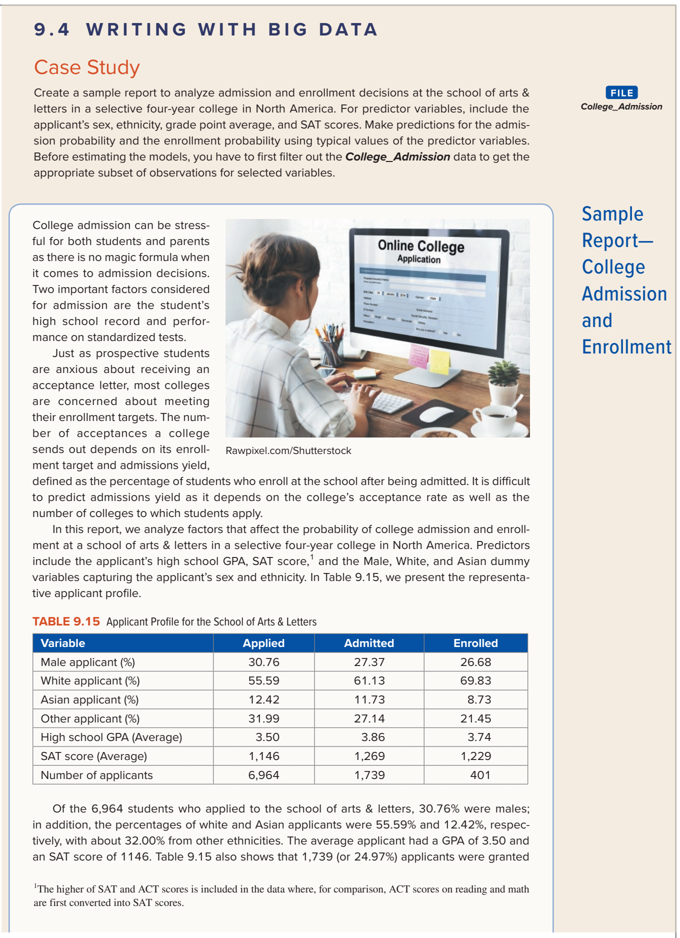
SYNOPSIS OF INTRODUCTORY CASE

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 KNN 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

Case Study

Create a sample report to analyze admission and enrollment decisions at the school of arts & letters in a selective four-year college in North 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

College 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 North 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.

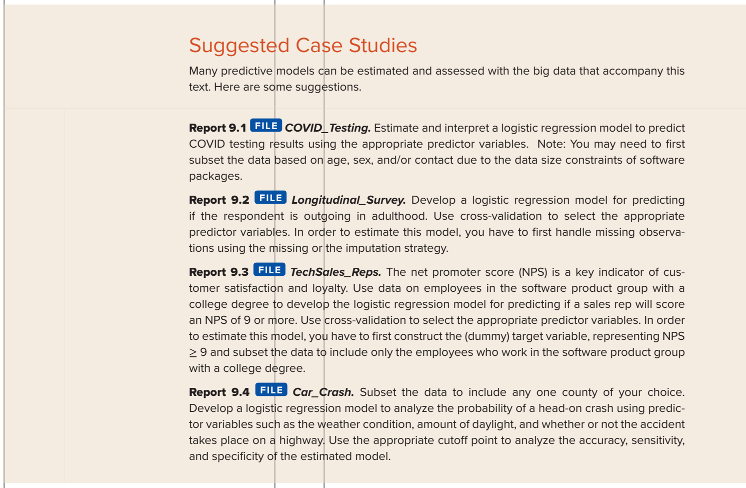
TABLE 9.15 Applicant Profile for the School of Arts & Letters

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
Number 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 Case 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 COVID testing results using the appropriate predictor variables. Note: 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. Use 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 (NPS) is a key indicator of customer satisfaction and loyalty. Use 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 NPS of 9 or more. Use cross-validation to select the appropriate predictor variables. In order to estimate this model, you have to first construct the (dummy) target variable, representing NPS ≥ 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. Use 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

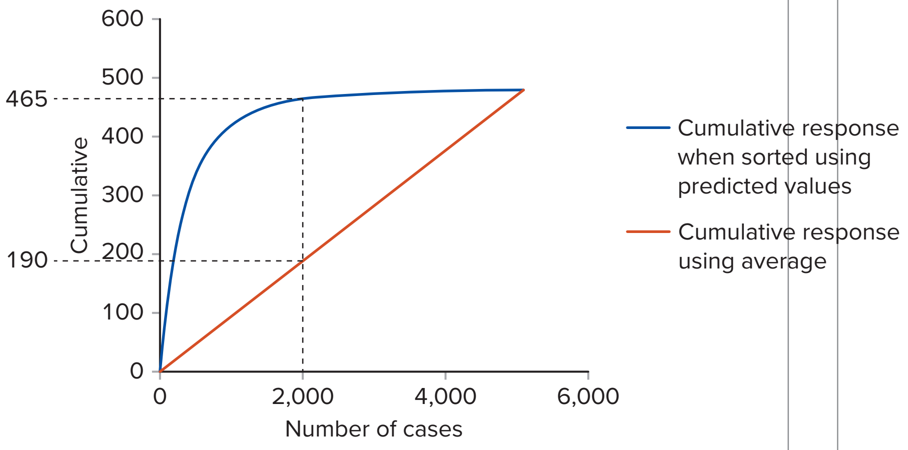


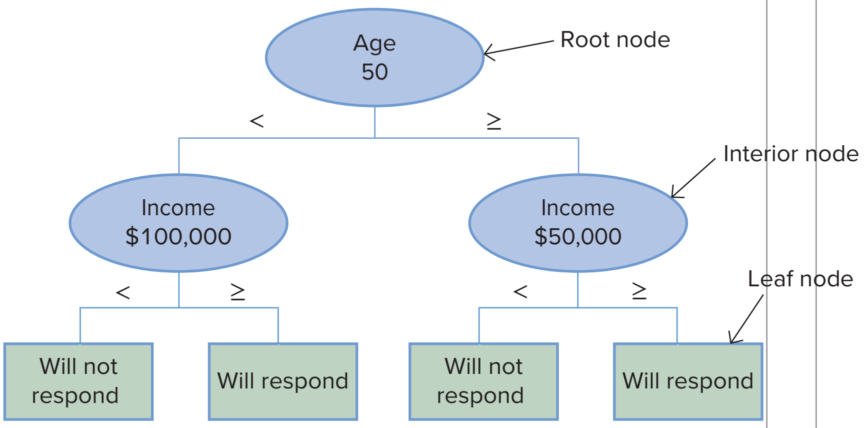
TABLE 11.14 Prediction Performance Measures

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

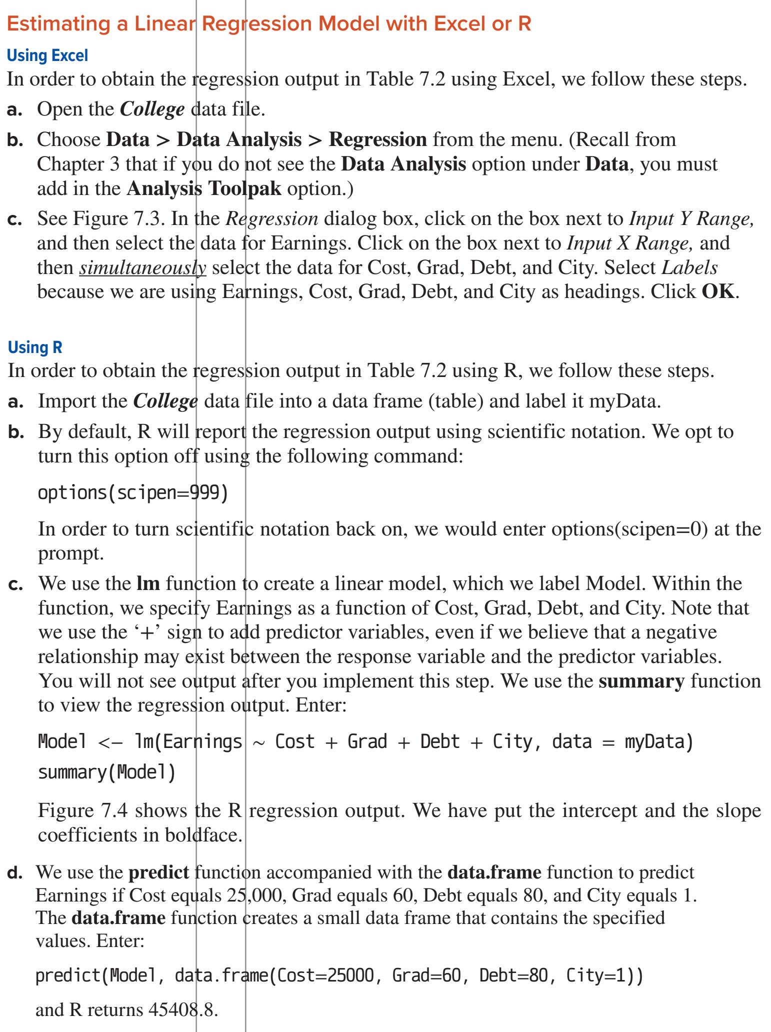


> *"I strongly applaud that this text offers a holistic approach to data analysis and places the emphasis on communicating with data. The Big Data and Data Mining sections seem promising with coverage of topics of the latest models and methods."*
>
> **Hao Chen,** *UW Platteville*

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*

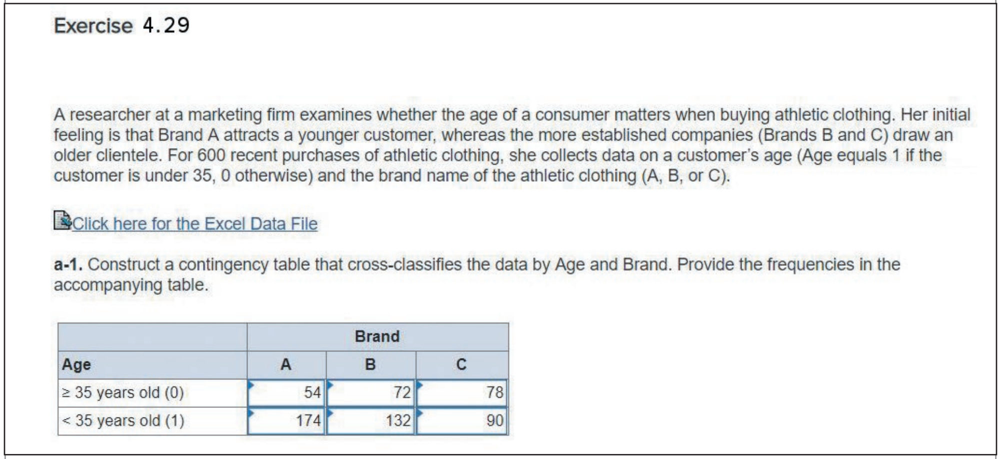
> *"I love the Excel/R examples and I love that the exercises and examples are after each section."*
>
> **Edie Schmidt,** *Nova Southeastern University*

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.*



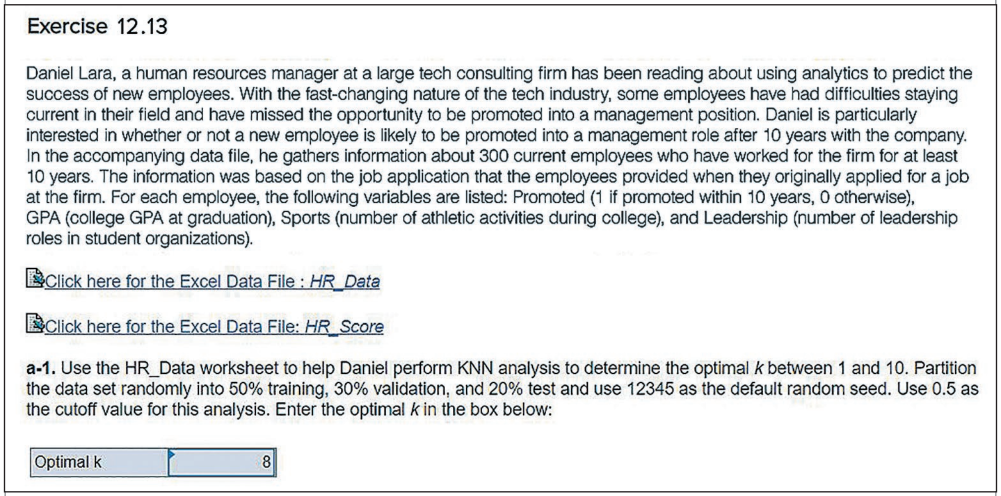
Exercise 4.29

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.

Instructors: Student Success Starts with You

Tools to enhance your unique voice

Want to build your own course? No problem. Prefer to use an OLC-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

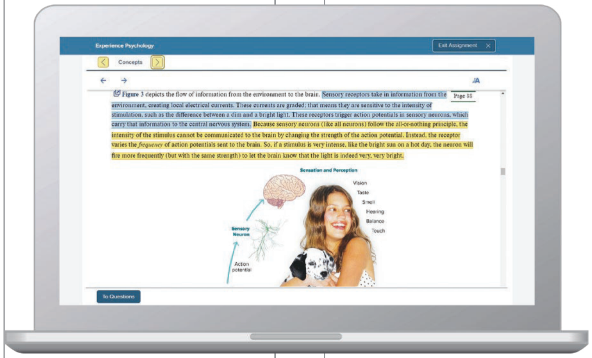


Figure 3 depicts the flow of information from the environment to the brain. Sensory receptors take in information from the environment, creating local electrical currents. These currents are graded; that means they are sensitive to the intensity of stimulation, such as the difference between a dim and a bright light. These receptors trigger action potentials in sensory neurons, which carry that information to the central nervous system. Because sensory neurons (like all neurons) follow the all-or-nothing principle, the intensity of the stimulus cannot be communicated to the brain by changing the strength of the action potential. Instead, the receptor varies the *frequency* of action potentials sent to the brain. So, if a stimulus is very intense, like the bright sun on a hot day, the neuron will fire more frequently (but with the same strength) to let the brain know that the light is indeed very, very bright.

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. Learn 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 LMS 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 affordable, 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 Learning that Fits You

Effective tools for efficient studying

Connect is designed to help you be more productive with simple, flexible, 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 Learning Assignments when it's convenient, even if you're offline. 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 email 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 found in Connect.

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.

Additionally, we would like to thank Samuel Frame, Janet Fraser, Rene Ordonez, Wendi Sun, Zhaowei Wang, and the team at Aptara for their help with developing and reviewing supplemental material.

And finally, special thanks to the wonderful and supportive team at McGraw Hill: Ryan McAndrews, Product Developer; Thomas Finn, Content Development Editor; Noelle Bathurst, Senior Portfolio Manager; Harper Christopher, Executive Marketing Manager; Pat Frederickson, Core Content Project Manager; Jamie Koch and George Theofanopoulos, Assessment Content Project Managers; Matt Diamond and Laurie Entringer, Designers; Eric Weber, Portfolio Manager; Kevin Moran, Director of Digital Content; and Tim Vertovec, Vice President BEC Portfolio.

BRIEF CONTENTS

APPENDIXES

CONTENTS

BUSINESS ANALYTICS

Communicating with Numbers

1 Introduction to Business Analytics

LEARNING OBJECTIVES

After reading this chapter, you should be able to:

LO **1.1** **Explain the importance of business analytics, data privacy, and data ethics.**

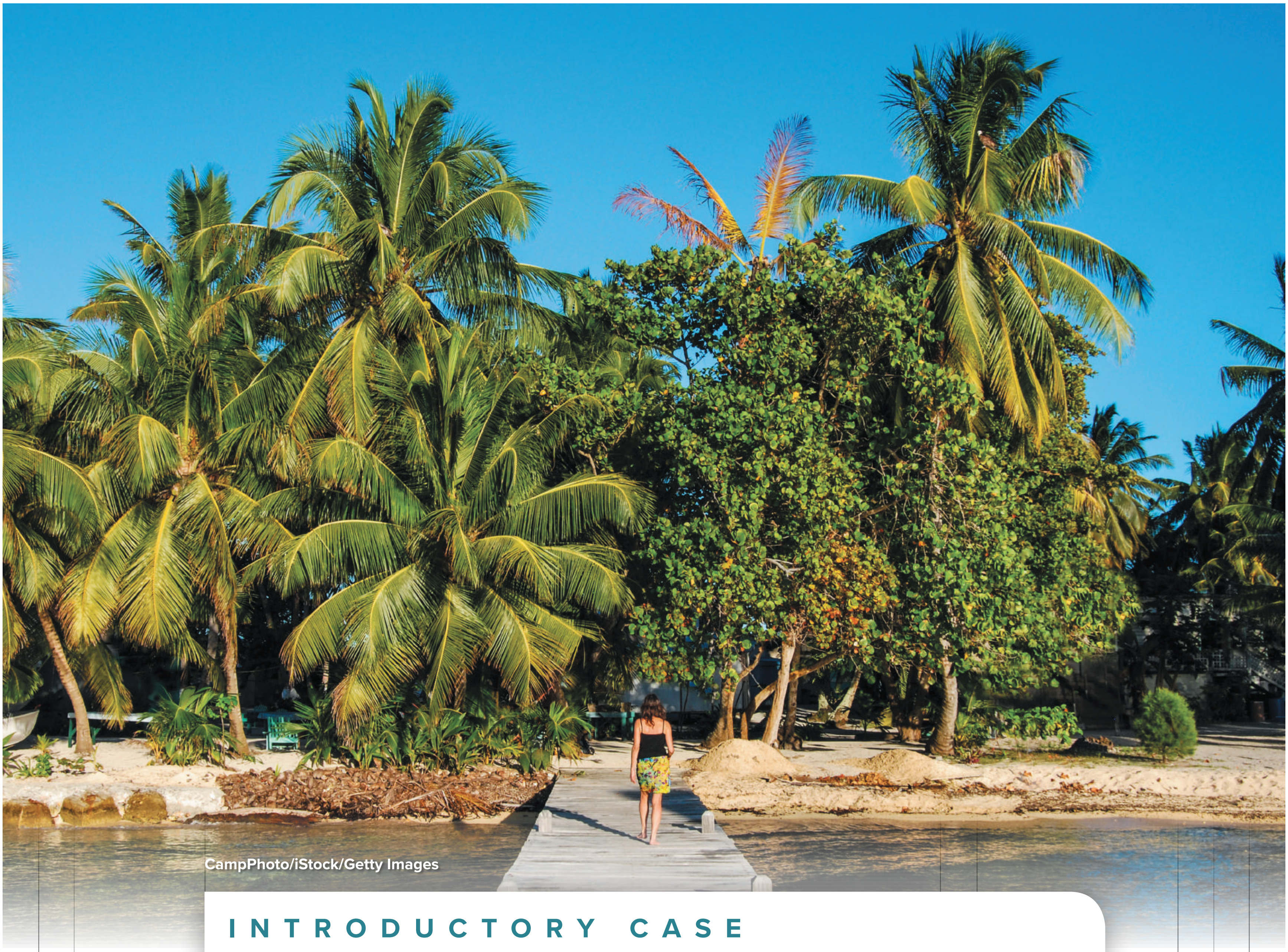
LO **1.2** **Explain the various types of data.**

LO **1.3** **Describe variables and types of measurement scales.**

LO **1.4** **Describe different data sources and file formats.**

In just about any contemporary human activity, the analysis of large amounts of data, under the umbrella of business or data analytics, helps us make better decisions. Managers, consumers, bankers, sports enthusiasts, politicians, and medical professionals are increasingly turning to data to boost a company's revenue, deepen customer engagement, find better options on consumer products, prevent threats and fraud, assess riskiness of loans, succeed in sports and elections, provide better diagnoses and cures for diseases, and so on. In the broadest sense, business analytics involves the methodology of extracting information and knowledge from data.

In this chapter, we will provide an overview of business analytics. Using real-world cases, we will highlight both the opportunities and the ethical issues emerging from the use and misuse of data. An important first step for studying business analytics is to understand data. We will describe various types of data and measurement scales of variables that will later help us choose appropriate statistical and computational models. Finally, we will discuss how we can take advantage of data sources that are publicly available and review some standard formats that people use to store and disseminate data.



CampPhoto/iStock/Getty Images

INTRODUCTORY CASE

Vacation in Belize

After graduating from a university in southern California, Emily Hernandez is excited to go on a vacation in Belize with her friends. There are different airlines that offer flights from southern California to Belize City. Emily prefers a direct flight from Los Angeles but is also worried about staying within her budget. Other, less expensive, options would mean making one or even two additional stops en route to Belize City. Once arriving at Belize City, she plans to take a sea ferry to one of the resort hotels on Ambergris Caye Island. Purchasing a vacation package from one of these hotels would be most cost-effective, but Emily wants to make sure that the hotel she chooses has all the amenities she wants, such as an early check-in option, a complimentary breakfast, recreational activities, and sightseeing services. She also wants to make sure that the hotel is reputable and has good customer reviews.

Emily starts researching her options for flights and hotels by searching for deals on the Internet. She has organized and kept meticulous records of the information she has found so that she can compare all the options. She would like to use this information to:

1. Find a flight that is convenient as well as affordable.
2. Choose a reputable hotel that is priced under $200 per night.

A synopsis of this case is provided at the end of Section 1.2.

LO 1.1

Explain the importance of business analytics, data privacy, and data ethics.

1.1 OVERVIEW OF BUSINESS ANALYTICS

Data and analytics capabilities have made a leap forward in recent years and have changed the way businesses make decisions. In the broadest sense, business analytics (also referred to as data analytics) involves the methodology of extracting information and knowledge from data to improve a company's bottom line and enhance the consumer experience. At the core, business analytics benefits companies by developing better marketing strategies, deepening customer engagement, enhancing efficiency in procurement, uncovering ways to reduce expenses, identifying emerging market trends, mitigating risk and fraud, etc. 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.

Business analytics is a broad topic, encompassing statistics, computer science, and information systems with a wide variety of applications in marketing, human resource management, economics, finance, health, sports, politics, etc. Unlike data science that is focused on advanced computer algorithms, business analytics tends to focus more on the analysis of the available data.

Raw data do not offer much value or insights. The analysis of data has been greatly simplified because of the improved computational power and the availability of sophisticated algorithms. However, in order to extract value from data, we need to be able to understand the business context, ask the right questions from the data, identify appropriate statistical and computational models, and communicate information into verbal and written language. It is important to note that numerical results are not very useful unless they are accompanied with clearly stated actionable business insights.

BUSINESS ANALYTICS

Business analytics combines qualitative reasoning with quantitative tools to identify key business problems and translate data analysis into decisions that improve business performance.

There are different types of analytics techniques designed to extract value from data that can be grouped into three broad categories: descriptive analytics, predictive analytics, and prescriptive analytics.

- **Descriptive analytics** refers to gathering, organizing, tabulating, and visualizing data to summarize "*what has happened?*" Examples of descriptive analytics include financial reports, public health statistics, enrollment at universities, student report cards, and crime rates across regions and time. Descriptive analytics is often referred to as **business intelligence (BI)**, which provides organizations and their users with the ability to access and manipulate data interactively through reports, dashboards, applications, and visualization tools. The descriptive information can be presented in a number of formats including written reports, tables, graphs, and maps.
- **Predictive analytics** refers to using historical data to predict "*what could happen in the future?*" Analytical models help identify associations between variables, and these associations are used to estimate the likelihood of a specific outcome. Examples of predictive analytics include identifying customers who are most likely to respond to specific marketing campaigns, admitted students who are likely to enroll, credit card transactions that are likely to be fraudulent, or the incidence of crime at certain regions and times.
- **Prescriptive analytics** refers to using optimization and simulation algorithms to provide advice on "*what should we do?*" It explores several possible actions and suggests a course of action. Examples include providing advice on scheduling employees' work hours and adjusting supply level in order to meet customer

demand, selecting a mix of products to manufacture, choosing an investment portfolio to meet a financial goal, or targeting marketing campaigns to specific customer groups on a limited budget.

The three categories of business analytics can also be viewed according to the level of sophistication and business values they offer. For many people, using predictive analytics to predict the future is more valuable than simply summarizing data and describing what happened in the past. In addition, predictive techniques tend to require more complex modeling and analysis tools than most descriptive techniques. Likewise, using prescriptive techniques to provide actionable recommendations could be more valuable than predicting a number of possible outcomes in the future. Turning data-driven recommendations into action also requires thoughtful consideration and organizational commitment beyond developing descriptive and predictive analytical models.

Figure 1.1 categorizes business analytics into three stages of development based on its value and the level of organizational commitment to data-driven decision making. Most chapters of this text can also be grouped within each of the three analytics categories as shown in Figure 1.1.

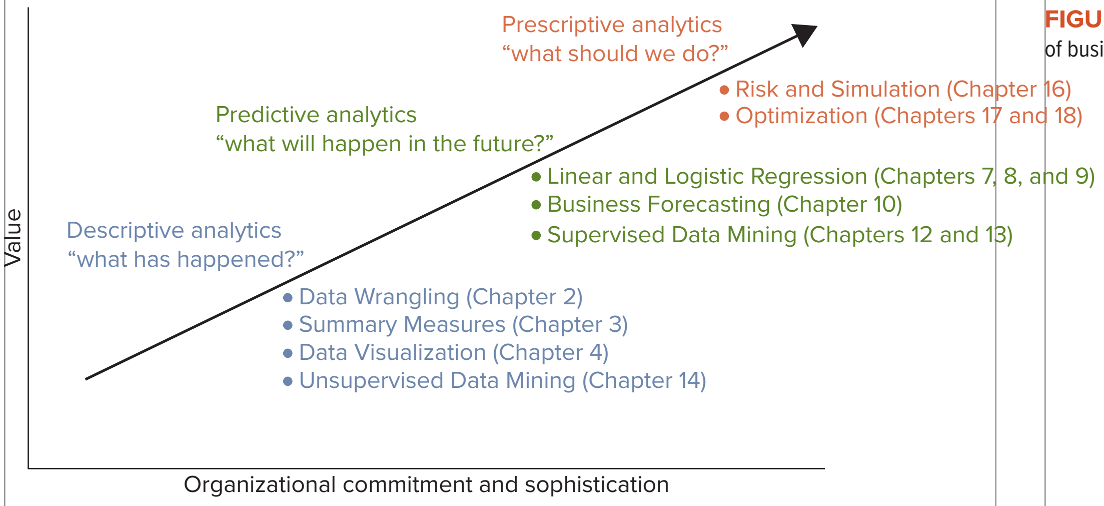

Prescriptive analytics
"what should we do?"
• Risk and Simulation (Chapter 16)
• Optimization (Chapters 17 and 18)
Predictive analytics
"what will happen in the future?"
• Linear and Logistic Regression (Chapters 7, 8, and 9)
• Business Forecasting (Chapter 10)
• Supervised Data Mining (Chapters 12 and 13)
Descriptive analytics
"what has happened?"
• Data Wrangling (Chapter 2)
• Summary Measures (Chapter 3)
• Data Visualization (Chapter 4)
• Unsupervised Data Mining (Chapter 14)
Value
Organizational commitment and sophistication

FIGURE 1.1 Three stages of business analytics

In addition to the chapters presented in Figure 1.1, Chapters 1 (Introduction to Business Analytics), 5 (Probability and Probability Distributions), 6 (Statistical Inference), 11 (Introduction to Data Mining), and 15 (Spreadsheet Modeling) cover prerequisite knowledge to the other topics in the text. Regardless of the analysis techniques we use, the overall goal of analytics is to improve business decision making. Essential to this goal is the ability to communicate the insights and value of data. Throughout the text, students will not only learn to conduct data analysis but also to tell an impactful story conveyed in written form to those who may not know detailed statistical and computational methods.

Important Business Analytics Applications

Data and analytics permeate our daily lives. We are often unaware of the impact that analytics has on even the most mundane activities, such as buying clothes, checking e-mails, interacting on social media, and watching a TV show. We will now highlight some of the important applications where the data-driven approach has been effectively used to replace and/or complement the traditional decision-making process that relies heavily on a few experts and their subjective judgment. This approach has led to more accurate, often overlooked, and unexpected findings, which have morphed into competitive advantages for the companies.

The Gap, Inc., once acclaimed as the company that "dictated how America dressed" by *The New York Times*, experienced disappointing sales as the fast fashion industry got increasingly crowded with major international competitors such as Zara, H&M, and Uniqlo. To turn the company around, Gap's CEO revolutionized the company by using analytics to identify loyal customers, match products to customers, enhance customer satisfaction, and manage product rebuys. Gap also combines the trends identified in data with real-time store sales to quickly bring more relevant products to market.

Netflix, one of the largest content-streaming companies, transformed the entertainment industry with its innovative business model. One of the engines that propelled its success was a sophisticated analytics tool called CineMatch, a recommendation system for movies and TV shows. CineMatch was so crucial to Netflix's ability to sustain and grow its customer base that the company launched a $1 million competition challenging the contestants to improve the CineMatch recommendation algorithm.

The Oakland Athletics, referred to as the A's, used analytics to build a playoff-bound baseball team; their story was depicted in the movie *Moneyball.* Traditionally, professional baseball recruiters relied heavily on subjective measures, such as body type, speed, and projected confidence, for scouting players. Recognizing this shortcoming, the A's took advantage of saber-metrics—statistical analysis of baseball records—to assess hidden worth in undervalued players. Despite being one of the poorest-funded teams in Major League Baseball, the A's were able to pick talented players cheaply, making them one of the most successful franchises in 2002 and 2003.

The Cancer Moonshot program, sponsored by the U.S. National Cancer Institute, aims to develop a cure for cancer. The program provides medical researchers with access to a vast amount of cancer patient data. By analyzing biopsy reports, treatment plans, and recovery rates of patients, researchers can study trends in how certain cancer proteins interact with different treatments and recommend the most promising treatment plan for the individual patient.

Data Privacy and Data Ethics

Data analytics allows companies to effectively target and understand their customers, but it also carries greater responsibility for understanding big data ethics. As companies capitalize on the data collected from their customers and constituents, they must also realize that there are enormous risks involved in using these data, especially in the forms of data security and privacy. This is especially the case when companies begin monetizing their data externally for purposes different from those for which the data were initially collected. In this section, we summarize the ever-evolving principles and guidelines that integrate data privacy and data ethics in data processing activities.

Data Privacy

Data Privacy, also referred to as information privacy, is a branch of data security related to the proper collection, usage, and transmission of data. Its concerns revolve around (a) how data are legally collected and stored; (b) if and how data are shared with third parties; and (c) how data collection, usage, and transmission meet all regulatory obligations.

Key principles of data privacy include

- Confidentiality. It is important that customer data and identity remain private. Should sensitive information be shared, especially medical and financial information, it must be done with utmost confidentiality.
- Transparency. It is important that data-processing activities and automated decisions are transparent. The risks, as well as social, ethical, and societal consequences, must be clearly understood by customers.
- Accountability. It is important that the data collection company has established a reflective, reasonable, and systematic use and protection of customer data.

There must be protection against unauthorized or unlawful processing and against accidental loss or destruction of the data.

The following examples serve as chilling reminders of the data privacy risks that may be encountered when using big data.

Marriott International disclosed a massive data breach in 2018 that affected up to 500 million guests. Private data accessed by the hackers included names, addresses, phone numbers, e-mail addresses, passport numbers, and encrypted credit card details. Similar attacks have been reported for Adobe in 2013, eBay in 2014, LinkedIn in 2016, Equifax in 2017, etc. Data breaches like these can lead to identity theft and/or financial fraud.

Cambridge Analytica, a political data firm hired by former President Donald Trump's 2016 election campaign, harvested private information from over 50 million Facebook profiles to create personality models for voters, which were later used to create digital ads to influence voter behaviors. Facebook has since tightened its policies on third-party access to user profiles, but the incident created enormous backlash against the social media company.

Target, like other retail stores, tracks consumer shopping habits based on the time of shopping, use of digital/paper coupons, purchase of brand name/generic, etc. These data are analyzed to find what consumers are likely to purchase. A Minneapolis father's privacy was invaded when he learned that his daughter was pregnant because he started receiving coupons for baby products from Target. Apparently, the daughter's buying habits had triggered a high "pregnancy prediction" score.

Cases involving data breach and misuse have given rise to data privacy laws that regulate how data are collected and used. These regulations also prescribe the necessary safeguards for data security and privacy protection of citizens. For example, in the United States, the U.S. Privacy Act contains important rights and restrictions on data maintained by government agencies. In the healthcare industry, the U.S. Health Insurance Portability and Accountability Act (HIPAA) outlines regulations for medical data security and privacy to ensure the confidentiality of patients.

In addition to the federal laws, state and local governments have also enacted data privacy legislations that are specific to their jurisdiction. The California Consumer Privacy Act and New York's Stop Hacks and Improve Electronic Data Security (SHIELD) Act are among the most comprehensive data privacy laws passed by the state governments. In Europe, the landmark General Data Protection Regulation (GDPR) is a comprehensive data privacy law that gives European Union citizens control over their personal data. In addition, the GDPR provides a unified data privacy framework for businesses and organizations that operate within the European Union. Similar regulations have also been enforced in other countries.

Data Ethics

It is important to note that the collection, usage, and transmission of data can also have profound impact on the well-being of individuals. **Data ethics** is a branch of ethics that studies and evaluates moral problems related to data. Its concerns revolve around evaluating whether data are being used for doing the right thing for people and society.

Key principles of data ethics include

- Human first. It is important that the human being stays at the center and human interests always outweigh institutional and commercial interests.
- No biases. It is important that the algorithms employed do not absorb unconscious biases in a population and amplify them in the analysis.

The following examples raise ethical questions about the use and misuse of big data.

Social Media. The Netflix documentary *The Social Dilemma* depicts how social media are designed to create addiction and manipulate human behavior for profit. Infinite scrolling, push notifications, and snippets of news and other trivia keep unsuspecting users addicted to their screens, which enables social media companies to maximize their advertising revenue. It is claimed that users' brains are manipulated and even rewired by algorithms that are designed to get their attention and make them buy things, including buying into distorted facts about people and society.

Beauty Contest. In 2016, computer algorithms were employed to select 44 beauty contest winners from 6,000 submitted photos from over 100 countries. The selections raised eyebrows because only a handful of winners were nonwhite. Only one person of color was selected, even though most photo submissions came from Africa and India.

DATA PRIVACY AND DATA ETHICS

- Data privacy is a branch of data security related to the proper collection, usage, and transmission of data.
- Data ethics is a branch of ethics that studies and evaluates moral problems related to data.

In this section, we provided an overview of business analytics. Using real-world examples, we highlighted both the opportunities and the ethical issues emerging from the use and misuse of data. The diverse nature of analytics cases requires deep understanding of data and data types. In the remaining part of this chapter, we will describe the various types of data and measurement scales that help us choose appropriate techniques for analyzing data. We will also discuss how we can take advantage of data sources that are publicly available and review some standard formats that people use to store and disseminate data.

LO 1.2

1.2 TYPES OF DATA

Explain the various types of data.

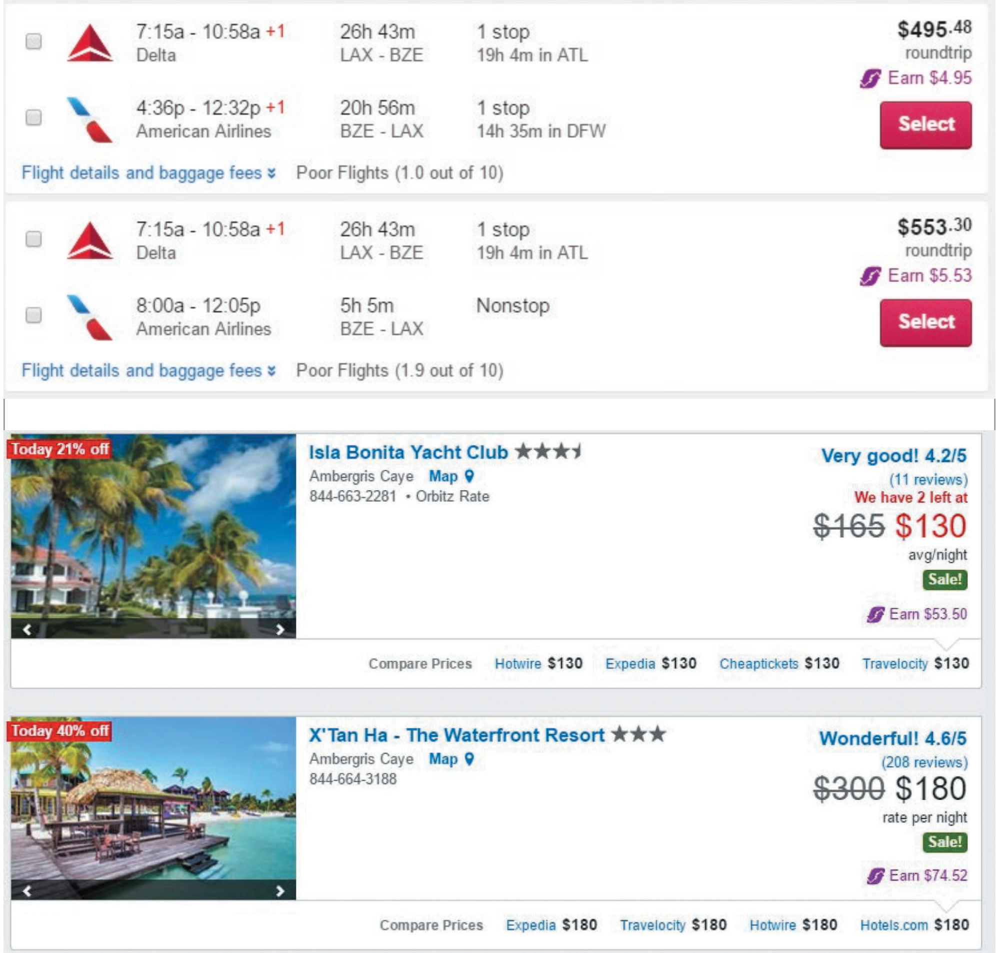
Every day, we use many kinds of data from various sources to help make decisions. An important first step for making decisions is to find the right data and prepare them for the analysis. In general, **data** are compilations of facts, figures, or other contents, both numerical and nonnumerical. Data of all types and formats are generated from multiple sources. We often find a large amount of data at our disposal. However, we also derive insights from relatively small data sets, such as from consumer focus groups, marketing surveys, or reports from government agencies.

Data that have been organized, analyzed, and processed in a meaningful and purposeful way become **information**. We use a blend of data, contextual information, experience, and intuition to derive **knowledge** that can be applied and put into action in specific situations.

DATA, INFORMATION, AND KNOWLEDGE

Data are compilations of facts, figures, or other contents, both numerical and nonnumerical. Information is a set of data that are organized and processed in a meaningful and purposeful way. Knowledge is derived from a blend of data, contextual information, experience, and intuition.

In the introductory case, Emily is looking for flights from Los Angeles International Airport (LAX) to Belize City (BZE), as well as hotels in Belize. An online search on Orbitz.com yields a total of 1,420 combinations of flight schedules. She also finds information on 19 hotels that are within her budget. Figure 1.2 shows a portion of the airfare and hotel information that Emily finds on the Internet.



☐ 7:15a - 10:58a +1 Delta — 26h 43m LAX - BZE — 1 stop 19h 4m in ATL

☐ 4:36p - 12:32p +1 American Airlines — 20h 56m BZE - LAX — 1 stop 14h 35m in DFW

$495.48 roundtrip

Earn $4.95

Flight details and baggage fees ¥ Poor Flights (1.0 out of 10)

☐ 7:15a - 10:58a +1 Delta — 26h 43m LAX - BZE — 1 stop 19h 4m in ATL

☐ 8:00a - 12:05p American Airlines — 5h 5m BZE - LAX — Nonstop

$553.30 roundtrip

Earn $5.53

Flight details and baggage fees ¥ Poor Flights (1.9 out of 10)

Today 21% off

Isla Bonita Yacht Club ★★★

Ambergris Caye Map

844-663-2281 • Orbitz Rate

Very good! 4.2/5

(11 reviews)

We have 2 left at

~~$165~~ $130

avg/night

Sale!

Earn $53.50

Compare Prices Hotwire $130 Expedia $130 Cheaptickets $130 Travelocity $130

Today 40% off

X'Tan Ha - The Waterfront Resort ★★★

Ambergris Caye Map

844-664-3188

Wonderful! 4.6/5

(208 reviews)

~~$300~~ $180

rate per night

Sale!

Earn $74.52

Compare Prices Expedia $180 Travelocity $180 Hotwire $180 Hotels.com $180

FIGURE 1.2 Screen shot of airfare and hotel search results from Orbitz.com

Source: Orbitz.com

Before we analyze the information that Emily has gathered, it is important to understand different types of data. In this section, we focus on the data categorizations.

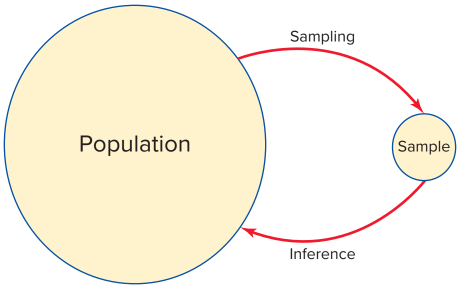
Sample and Population Data

There are several ways to categorize data depending on how they are collected, their format, and specific values they represent. In most instances, it is not feasible to collect data that consist of all items of interest—the **population**—in a statistical problem. Therefore, a subset of data—a **sample**—is used for the analysis. We rely on sampling because we are unable to use population data for two main reasons.

- **Obtaining information on the entire population is expensive.** Suppose we are interested in the average lifespan of a Duracell AAA battery. It would be incredibly expensive to test each battery. And moreover, in the end, all batteries would be dead and the answer to the original question would be useless.
- **It is impossible to examine every member of the population.** Consider how the monthly unemployment rate in the United States is calculated by the Bureau of Labor Statistics (BLS). Is it reasonable to assume that the the BLS contacts each individual in the labor force and asks whether or not he or she is employed? Given that there are over 160 million individuals in the labor force, this task would be impossible to complete in 30 days. Instead, the BLS conducts a monthly sample survey of about 60,000 households to measure the extent of unemployment in the United States.

Figure 1.3 depicts the flow of information between the population and a sample. Consider, for example, a 2016 Gallup survey that found that only 50% of millennials plan to stay at their current job for more than a year. We use this sample result, called a **sample statistic**, in an attempt to estimate the corresponding unknown **population parameter**. In this case, the parameter of interest is the percentage of *all* millennials who plan to be with their current job for more than a year.

FIGURE 1.3 Population versus sample



POPULATION VERSUS SAMPLE

A population consists of all items of interest in a statistical problem. A sample is a subset of the population. We analyze sample data and calculate a sample statistic to make inferences about the unknown population parameter.

In the introductory case, Emily is working with sample data. The population would have consisted of all the airlines and hotels, some of which may not even have shown up in an online search.

Cross-Sectional and Time Series Data

Sample data are generally collected in one of two ways. **Cross-sectional data** refer to data collected by recording a characteristic of many subjects at the same point in time, or without regard to differences in time. Subjects might include individuals, households, firms, industries, regions, and countries.

Table 1.1 is an example of a cross-sectional data set. It lists the team standings for the National Basketball Association's Eastern Conference at the end of the 2018–2019 season. The eight teams may not have ended the season precisely on the same day and time, but the differences in time are of no relevance in this example. Other examples of cross-sectional data include the recorded grades of students in a class, the sale prices of single-family homes sold last month, the current price of gasoline in different cities in the United States, and the starting salaries of recent business graduates from the University of Connecticut.

TABLE 1.1 2018–2019 NBA Eastern Conference

Team name	Wins	Losses	Winning percentage
Milwaukee Bucks	60	22	0.732
Toronto Raptors*	58	24	0.707
Philadelphia 76ers	51	31	0.622
Boston Celtics	49	33	0.598
Indiana Pacers	48	34	0.585
Brooklyn Nets	42	40	0.512
Orlando Magic	42	40	0.512
Detroit Pistons	41	41	0.500

*The Toronto Raptors won their first NBA title during the 2018–2019 season.

Time series data refer to data collected over several time periods focusing on certain groups of people, specific events, or objects. Time series data can include hourly, daily, weekly, monthly, quarterly, or annual observations. Examples of time series data include the hourly body temperature of a patient in a hospital's intensive care unit, the daily price of General Electric stock in the first quarter of 2021, the weekly exchange rate between the U.S. dollar and the euro over the past six months, the monthly sales of cars at a dealership in 2021, and the annual population growth rate of India in the last decade. In these examples, temporal ordering is relevant and meaningful.

Figure 1.4 shows a plot of the national homeownership rate in the U.S. from 2000 to 2018. According to the U.S. Census Bureau, the national homeownership rate in the first quarter of 2016 plummeted to 63.6% from a high of 69.4% in 2004. An explanation for the decline in the homeownership rate is the stricter lending practices caused by the housing market crash in 2007 that precipitated a banking crisis and deep recession. This decline can also be attributed to home prices outpacing wages in the sample period.

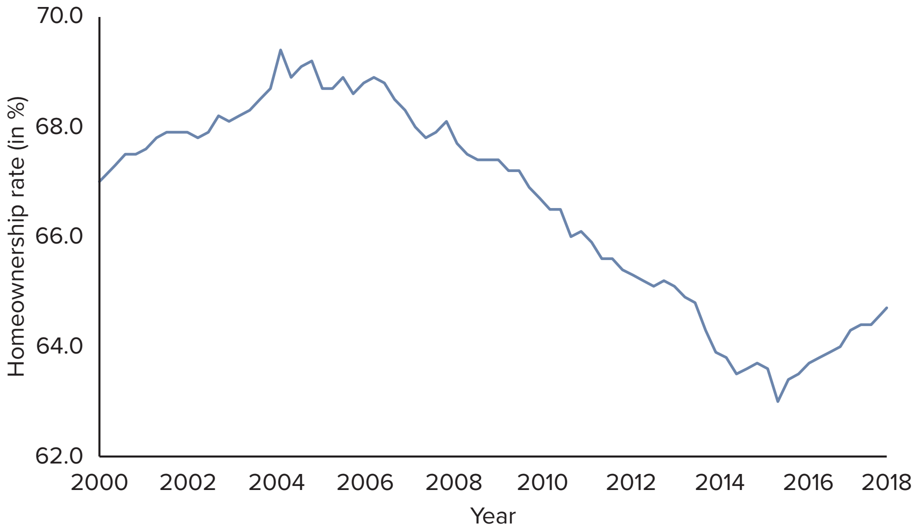


FIGURE 1.4 Homeownership rate (in %) in the U.S. from 2000 through 2018

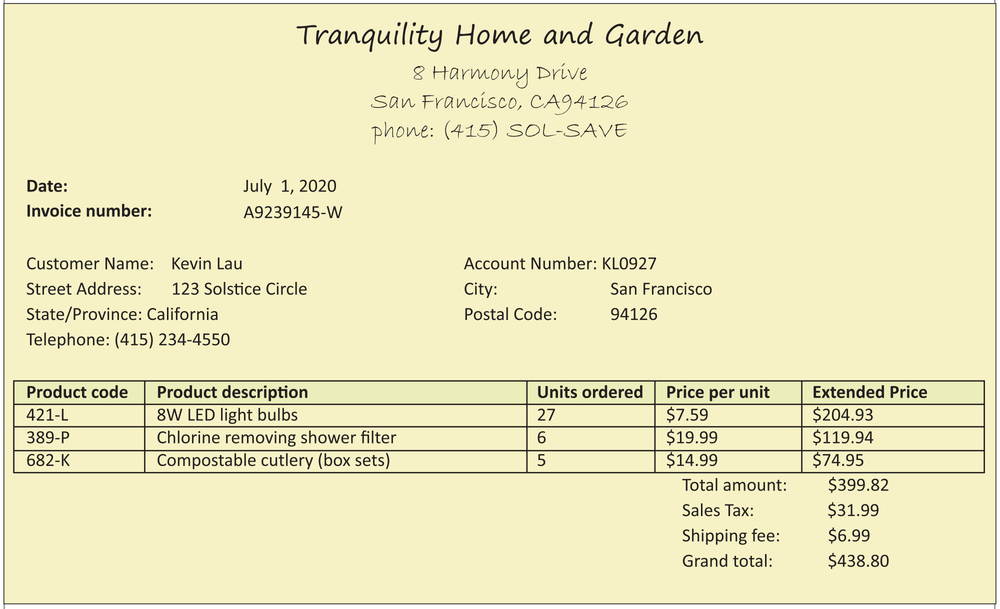
Structured and Unstructured Data

When you think of data, the first image that probably pops in your head is lots of numbers and perhaps some charts and graphs. In reality, data can come in multiple forms. For example, information exchange in social networking websites such as Facebook, LinkedIn, and Twitter also constitute data. In order to better understand the various forms of data, we make a distinction between structured and unstructured data.

Generally, **structured data** reside in a predefined, row-column format. We use spreadsheet or database applications (refer to Section 2.1) to enter, store, query, and analyze structured data. Examples of structured data include numbers, dates, and groups of words and numbers, typically stored in a tabular format. Structured data often consist of numerical information that is objective and is not open to interpretation.

Point-of-sale and financial data are examples of structured data and are usually designed to capture a business process or transaction. Examples include the sale of retail products, a money transfer between bank accounts, and the student enrollment in a university course. When individual consumers buy products from a retail store, each transaction is captured into a record of structured data.

Consider the sales invoice shown in Figure 1.5. Whenever a customer places an order like this, there is a predefined set of data to be collected, such as the transaction date, shipping address, and the units of product being purchased. Even though a receipt or an invoice may not always be presented in rows and columns, the predefined structure allows businesses and organizations to translate the data on the document into a row-column format.



Tranquility Home and Garden

8 Harmony Drive
San Francisco, CA94126
phone: (415) SOL-SAVE

Date: July 1, 2020
Invoice number: A9239145-W

Customer Name: Kevin Lau
Street Address: 123 Solstice Circle
State/Province: California
Telephone: (415) 234-4550

Account Number: KL0927
City: San Francisco
Postal Code: 94126

Product code	Product description	Units ordered	Price per unit	Extended Price
421-L	8W LED light bulbs	27	$7.59	$204.93
389-P	Chlorine removing shower filter	6	$19.99	$119.94
682-K	Compostable cutlery (box sets)	5	$14.99	$74.95

Total amount:	$399.82
Sales Tax:	$31.99
Shipping fee:	$6.99
Grand total:	$438.80

FIGURE 1.5 A sample invoice from a retail transaction

As we will see in Example 1.1, the flight information collected for the introductory case may not fit precisely in a tabular format, but because of the structured nature of the data, they can easily be summarized into rows and columns.

For decades, companies and organizations relied mostly on structured data to run their businesses and operations. Today, with the advent of the digital age, both structured and unstructured data are used for making business decisions.

Unlike structured data, **unstructured data** (or unmodeled data) do not conform to a predefined, row-column format. They tend to be textual (e.g., written reports, e-mail messages, doctor's notes, or open-ended survey responses) or have multimedia contents (e.g., photographs, videos, and audio data). Even though these data may have some implied structure (e.g., a report title, an e-mail's subject line, or a time stamp on a photograph), they are still considered unstructured as they do not conform to a row-column model required in most database systems. Social media data such as Twitter, YouTube, Facebook, and blogs are examples of unstructured data.

Both structured and unstructured data can be either **human-generated** or **machine-generated**. For structured data, human-generated data include information on price, income, retail sales, age, gender, etc., whereas machine-generated data include information from manufacturing sensors (rotations per minute), speed cameras (miles per hour), web server logs (number of visitors), etc. For unstructured data, human-generated data include texts of internal e-mails, social media data, presentations, mobile phone conversations, text message data, and so on, whereas machine-generated data include satellite images, meteorological data, surveillance video data, traffic camera images, and others.

Big Data

Nowadays, businesses and organizations generate and gather more and more data at an increasing pace. The term **big data** is a catch-phrase, meaning a massive volume of both structured and unstructured data that are extremely difficult to manage, process, and analyze using traditional data-processing tools. Despite the challenges, big data present great opportunities to gain knowledge and business intelligence with potential game-changing impacts on company revenues, competitive advantage, and organizational efficiency.

More formally, a widely accepted definition of big data is “high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation” (www.gartner.com). The three characteristics (the three Vs) of big data are:

- **Volume:** An immense amount of data is compiled from a single source or a wide range of sources, including business transactions, household and personal devices, manufacturing equipment, social media, and other online portals.
- **Velocity:** In addition to volume, data from a variety of sources get generated at a rapid speed. Managing these data streams can become a critical issue for many organizations.
- **Variety:** Data also come in all types, forms, and granularity, both structured and unstructured. These data may include numbers, text, and figures as well as audio, video, e-mails, and other multimedia elements.

In addition to the three defining characteristics of big data, we also need to pay close attention to the veracity of the data and the business value that they can generate. **Veracity** refers to the credibility and quality of data. One must verify the reliability and accuracy of the data content prior to relying on the data to make decisions. This becomes increasingly challenging with the rapid growth of data volume fueled by social media and automatic data collection. **Value** derived from big data is perhaps the most important aspect of any analytics initiative. Having a plethora of data does not guarantee that useful insights or measurable improvements will be generated. Organizations must develop a methodical plan for formulating business questions, curating the right data, and unlocking the hidden potential in big data.

Big data, however, do not necessarily imply complete (population) data. Take, for example, the analysis of all Facebook users. It certainly involves big data, but if we consider all Internet users in the world, Facebook users are only a very large sample. There are many Internet users who do not use Facebook, so the data on Facebook do not represent the population. Even if we define the population as pertaining to those who use online social media, Facebook is still one of many social media portals that consumers use. And because different social media are used for different purposes, data collected from these sites may very well reflect different populations of Internet users; this distinction is especially important from a strategic business standpoint. Therefore, Facebook data are simply a very large sample.

In addition, we may choose not to use big data in its entirety even when they are available. Sometimes it is just inconvenient to analyze a very large data set as it is computationally burdensome, even with a modern, high-capacity computer system. Other times, the additional benefits of working with big data may not justify the associated costs. In sum, we often choose to work with relatively smaller data sets drawn from big data.

STRUCTURED, UNSTRUCTURED, AND BIG DATA

Structured data are data that reside in a predefined, row-column format, while unstructured data do not conform to a predefined, row-column format. Big data is a term used to describe a massive volume of both structured and unstructured data that are extremely difficult to manage, process, and analyze using traditional data-processing tools. Big data, however, do not necessarily imply complete (population) data.

In this text, we focus on traditional statistical and data mining methods applied to structured data. Sophisticated tools to analyze unstructured data are beyond the scope of this text.

EXAMPLE 1.1

In the introductory case, Emily is looking for roundtrip flight schedules offered by different airlines from Los Angeles, California, to Belize City. Once arriving at Belize City, she plans to purchase a vacation package from one of the resort hotels on Ambergris Caye Island. She has compiled information on flight schedules and hotel options from search results on Orbitz.com. Her initial search, conducted in March 2019, yields 1,420 flight schedules and 19 hotels. In its current format, Emily is overwhelmed by the amount of data and knows that she needs to refine her search. She would like to focus on flights that are convenient (short duration) and relatively inexpensive. For hotels, she would like an affordable hotel (priced under $200 per night) with good reviews (above-average review on a 5-point scale). Given Emily's preferences, summarize the online information in tabular form.

SOLUTION:

We first search for roundtrip flight schedules where priority is given to short flight times and low prices. Even though the flight information is not in a perfect row-column configuration, the structured nature of the data allows us to summarize the information in a tabular format. The four most relevant options are presented in Panel A of Table 1.2. Given these options, it seems that the American/Delta choice might be the best for her. Similarly, once we refine the search to include only hotels that fall within her budget of under $200 per night and have an average consumer rating above 4 on a 5-point scale, the number of hotel options declines from 19 to five. These five options are presented in Panel B of Table 1.2. In this case, her choice of hotel is not clear-cut. The hotel with the highest rating is also the most expensive (X'Tan Ha – The Waterfront), while the least expensive hotel has the lowest rating (Isla Bonita Yacht Club) and the rating is based on only 11 reviews. Emily will now base her final hotel choice on online reviews. These reviews constitute unstructured data that do not conform to a row-column format.

TABLE 1.2 Tabular Format of Flight Search and Hotel Results

Panel A. Flight Search Results

		Los Angeles to Belize City			Belize City to Los Angeles		
Airlines	Price	Departure time	Duration	Number of stops	Departure time	Duration	Number of stops
Delta/American	$495.48	7:15am	26h 43m	1	4:36pm	20h 56m	1
Delta/American	$553.30	7:15am	26h 43m	1	8:00am	5h 5m	0
United/Delta	$929.91	6:55am	6h 58m	1	10:30am	5h 2m	0
American/Delta	$632.91	7:00am	7h 41m	1	10:30am	5h 2m	0

Panel B. Hotel Search Results

Hotel under $200	Average rating	Number of reviews	Price per night
Isla Bonita Yacht Club	4.2	11	$130
X'Tan Ha – The Waterfront	4.6	208	$180
Mata Rocks Resort	4.2	212	$165
Costa Blu	4.5	8	$165
Blue Tang Inn	4.2	26	$179

SYNOPSIS OF INTRODUCTORY CASE

David Schulz Photography/Shutterstock

Emily Hernandez is excited to go on vacation in Belize with her friends. She decides to search on Orbitz.com for flights that are convenient as well as affordable. She also wants to find a reputable hotel in Ambergris Caye Island that is priced under $200 per night. Although she finds 1,420 options for the flight, she focuses on morning flights with priority given to lower prices and shorter flight times.

There are four airline options that she finds suitable. With a price of $495.48, the Delta/American option is the cheapest, but the flight time is over 20 hours each way. The United/Delta option offers the shortest flight times; however, it comes with a price tag of $929.91. Emily decides on the American/Delta option, which seems most reasonable with a price of $632.91, a flight time of under eight hours to Belize, and only five hours on the return. In regard to hotels, Bonita Yacht Club offers the cheapest price of $130, but it comes with a relatively low rating of 4.2.

In addition, Emily is concerned about the credibility of the rating as it is based on only 11 reviews. Although not the cheapest, Emily decides to go with X'Tan Ha – The Waterfront. It has the highest rating of 4.6 and the price still falls within her budget of under $200. In these reviews, Emily consistently finds key phrases such as "great location," "clean room," "comfortable bed," and "helpful staff." Finally, the pictures that guests have posted are consistent with the images published by the resort on its website.

EXERCISES 1.2

Applications

1. According to recent estimates, annually, the average American spends $583 on alcohol and $1,100 on coffee.
   a. Describe the relevant population.
   b. Are the estimates based on sample or population data?

2. Many people regard video games as an obsession for youngsters, but, in fact, the average age of a video game player is 35 years old. Is the value 35 likely the actual or the estimated average age of the population? Explain.

3. An accounting professor wants to know the average GPA of the students enrolled in her class. She looks up information on Blackboard about the students enrolled in her class and computes the average GPA as 3.29. Describe the relevant population.

4. Recent college graduates with an engineering degree continue to earn high salaries. An online search revealed that the average annual salary for an entry-level position in engineering is $65,000.
   a. What is the relevant population?
   b. Do you think the average salary of $65,000 is computed from the population? Explain.

5. Research suggests that depression significantly increases the risk of developing dementia later in life. Suppose that in a study involving 949 elderly persons, it was found that 22% of those who had depression went on to develop dementia, compared to only 17% of those who did not have depression.
   a. Describe the relevant population and the sample.
   b. Are the numbers 22% and 17% associated with the population or a sample?

6. Go to www.zillow.com and find the sale price of 20 single-family homes sold in Las Vegas, Nevada, in the last 30 days. Structure the data in a tabular format and include the sale price, the number of bedrooms, the square footage, and the age of the house. Do these data represent cross-sectional or time series data?

7. Go to www.finance.yahoo.com to get the current stock quote for Home Depot (ticker symbol = HD). Use the ticker symbol to search for historical prices and create a table that includes the monthly adjusted close price of Home Depot stock for the last 12 months. Do these data represent cross-sectional or time series data?

8. Go to *The New York Times* website at www.nytimes.com and review the front page. Would you consider the data on the page to be structured or unstructured? Explain.
9. Conduct an online search to compare small hybrid vehicles (e.g., Toyota Prius, Ford Fusion, Chevrolet Volt) on price, fuel economy, and other specifications. Do you consider the search results structured or unstructured data? Explain.
10. Find Under Armour's annual revenue from the past 10 years. Are the data considered structured or unstructured? Explain. Are they cross-sectional or time series data?
11. Ask 20 of your friends about their online social media usage, specifically whether or not they use Facebook, Instagram, and Snapchat; how often they use each social media portal; and their overall satisfaction of each of these portals. Create a table that presents this information. Are the data considered structured or unstructured? Are they cross-sectional or time series data?
12. Ask 20 of your friends whether they live in a dormitory, a rental unit, or other form of accommodation. Also find out their approximate monthly lodging expenses. Create a table that uses this information. Are the data considered structured or unstructured? Are they cross-sectional or time series data?
13. Go to the U.S. Census Bureau website at www.census.gov and search for the most recent median household income for Alabama, Arizona, California, Florida, Georgia, Indiana, Iowa, Maine, Massachusetts, Minnesota, Mississippi, New Mexico, North Dakota, and Washington. Do these data represent cross-sectional or time series data? Comment on the regional differences in income.

LO 1.3

1.3 VARIABLES AND SCALES OF MEASUREMENT

Describe variables and types of measurement scales.

For any work related to business analytics, we invariably focus on people, firms, or events with particular characteristics. When a characteristic of interest differs in kind or degree among various observations (records), then the characteristic can be termed a **variable**. Marital status and income are examples of variables because a person's marital status and income vary from person to person. Variables are further classified as either **categorical** (qualitative) or **numerical** (quantitative). The observations of a categorical variable represent categories, whereas the observations of a numerical variable represent meaningful numbers. For example, marital status is a categorical variable, whereas income is a numerical variable.

For a categorical variable, we use labels or names to identify the distinguishing characteristic of each observation. For instance, a university may identify each student's status as either at the undergraduate or the graduate level, where the education level 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), IT firm (hardware, software, cloud), and course grade (A, B, C, D, F). It is important to note that categories are often converted into numerical codes for purposes of data processing, which we will discuss in Chapter 2.

For a numerical variable, we use numbers to identify the distinguishing characteristic of each observation. Numerical variables, in turn, are either discrete or continuous. A **discrete variable** assumes a countable number of values. Consider the number of children in a family or the number of points scored in a basketball game. We may observe values such as 3 children in a family or 90 points being scored in a basketball game, but we will not observe fractions such as 1.3127 children or 92.4724 scored points. The values that a discrete variable assumes need not be whole numbers. For example, the price of a stock for a particular firm is a discrete variable. The stock price may take on a value of $20.37 or $20.38, but it cannot take on a value between these two points.

A **continuous variable** is characterized by uncountable values within an interval. Weight, height, time, and investment return are all examples of continuous variables. For example, an unlimited number of values occur between the weights of 100 and 101 pounds, such as 100.3, 100.625, 100.8342, and so on. In practice, however, continuous variables are often measured in discrete values. We may report a newborn's weight (a continuous variable) in discrete terms as 6 pounds 10 ounces and another newborn's weight in similar discrete terms as 6 pounds 11 ounces.

CATEGORICAL AND NUMERICAL VARIABLES

A variable is a general characteristic being observed on a set of people, objects, or events, where each observation varies in kind or degree.

- The observations of a categorical variable assume names or labels.
- The observations of a numerical variable assume meaningful numerical values. A numerical variable can be further categorized as either discrete or continuous. A discrete variable assumes a countable number of values, whereas a continuous variable is characterized by uncountable values.

EXAMPLE 1.2

In the introductory case, Emily has conducted an online search on airfares and hotels for her planned vacation to Ambergris Caye Island. She has summarized the hotel information in Panel B of Table 1.2. Determine which of the included variables are categorical or numerical and, if numerical, determine if they are discrete or continuous.

SOLUTION:

The hotel variable is categorical because the observations—the names—are merely labels. On the other hand, the average rating, the number of reviews, and the price per night are numerical variables because the observations are all meaningful numbers. Note that the average rating is continuous because it is characterized by uncountable values within the 0 to 5 interval. The number of reviews and the price per night (measured in $) are discrete variables because they can only assume a countable number of values.

The Measurement Scales

In order to choose the appropriate techniques for summarizing and analyzing variables, we need to distinguish between the different measurement scales. The observations for any variable can be classified into one of four major measurement scales: nominal, ordinal, interval, or ratio. Nominal and ordinal scales are used for categorical variables, whereas interval and ratio scales are used for numerical variables. We discuss these scales in ascending order of sophistication.

The Nominal Scale

The **nominal scale** represents the least sophisticated level of measurement. If we are presented with nominal observations, all we can do is categorize or group them. The observations differ merely by name or label. Table 1.3 lists the 30 publicly owned companies, as of February 2019, that comprise the Dow Jones Industrial Average (DJIA). The DJIA is a stock market index that shows how these U.S.-based companies have traded during a standard trading session in the stock market. Table 1.3 also indicates where stocks of these companies are traded: on either the National Association of Securities Dealers Automated Quotations (Nasdaq) or the New York Stock Exchange (NYSE). These observations are classified as nominal scale because we are simply able to group or categorize them. Specifically, only five stocks are traded on the Nasdaq, whereas the remaining 25 are traded on the NYSE.

Often, we substitute numbers for the particular categorical characteristic or trait that we are grouping. For instance, we might use the number 0 to show that a company's stock is traded on the Nasdaq and the number 1 to show that a company's stock is traded on the NYSE. One reason why we do this is for ease of exposition; always referring to the National Association of Securities Dealers Automated Quotations, or even the Nasdaq, can be awkward and unwieldy.

TABLE 1.3 Companies of the DJIA and Exchange Where Stock Is Traded

Company	Exchange	Company	Exchange
3M (MMM)	NYSE	Johnson & Johnson (JNJ)	NYSE
American Express (AXP)	NYSE	JPMorgan Chase (JPM)	NYSE
Apple (AAPL)	Nasdaq	McDonald’s (MCD)	NYSE
Boeing (BA)	NYSE	Merck (MRK)	NYSE
Caterpillar (CAT)	NYSE	Microsoft (MFST)	Nasdaq
Chevron (CVX)	NYSE	Nike (NKE)	NYSE
Cisco (CSCO)	Nasdaq	Pfizer (PFE)	NYSE
Coca-Cola (KO)	NYSE	Procter & Gamble (PG)	NYSE
Disney (DIS)	NYSE	Travelers (TRV)	NYSE
DowDupont (DWDP)	NYSE	United Health (UNH)	NYSE
ExxonMobil (XOM)	NYSE	United Technologies (UTX)	NYSE
Goldman Sachs (GS)	NYSE	Verizon (VZ)	NYSE
Home Depot (HD)	NYSE	Visa (V)	NYSE
IBM (IBM)	NYSE	Wal-Mart (WMT)	NYSE
Intel (INTC)	Nasdaq	Walgreen (WBA)	Nasdaq

The Ordinal Scale

Compared to the nominal scale, the **ordinal scale** reflects a stronger level of measurement. With ordinal observations, we are able to both categorize and rank them with respect to some characteristic or trait. The weakness with ordinal observations is that we cannot interpret the difference between the ranked observations because the actual numbers used are arbitrary. For example, consider customer service ratings for a call center as excellent (5 stars), very good (4 stars), good (3 stars), fair (2 stars), or poor (1 star). We summarize the categories and their respective ratings in Table 1.4.

TABLE 1.4 Customer Service Ratings

Category	Rating
Excellent	5
Very good	4
Good	3
Fair	2
Poor	1

In Table 1.4, the number attached to excellent (5 stars) is higher than the number attached to good (3 stars), indicating that the response of excellent is preferred to good. However, we can easily redefine the ratings, as we show in Table 1.5.

TABLE 1.5 Redefined Customer Service Ratings

Category	Rating
Excellent	100
Very good	80
Good	70
Fair	50
Poor	40

In Table 1.5, excellent still receives a higher number than good, but now the difference between the two categories is 30 points (100 – 70), as compared to a difference of 2 points (5 – 3) when we use the first classification. In other words, differences between categories are meaningless with ordinal observations.

As mentioned earlier, observations of a categorical variable are typically expressed in words but are coded into numbers for purposes of data processing. When summarizing the results of a categorical variable, we typically count the number of observations that fall into each category or calculate the percentage of observations that fall into each category. However, with a categorical variable, we are unable to perform meaningful arithmetic operations, such as addition and subtraction.

The Interval Scale

With observations that are measured on the **interval scale**, we are able to categorize and rank them as well as find meaningful differences between them. The Fahrenheit scale for temperatures is an example of an interval-scaled variable. Not only is 60 degrees Fahrenheit hotter than 50 degrees Fahrenheit, the same difference of 10 degrees also exists between 90 and 80 degrees Fahrenheit.

The main drawback of an interval-scaled variable is that the value of zero is arbitrarily chosen; the zero point of an interval-scaled variable does not reflect a complete absence of what is being measured. No specific meaning is attached to 0 degrees Fahrenheit other than to say it is 10 degrees colder than 10 degrees Fahrenheit. With an arbitrary zero point, meaningful ratios cannot be constructed. For instance, it is senseless to say that 80 degrees is twice as hot as 40 degrees; in other words, the ratio 80/40 has no meaning.

The Ratio Scale

The **ratio scale** represents the strongest level of measurement. The ratio scale has all the characteristics of the interval scale as well as a true zero point, which allows us to interpret the ratios between observations. The ratio scale is used in many business applications. Variables such as sales, profits, and inventory levels are expressed on the ratio scale. A meaningful zero point allows us to state, for example, that profits for firm A are double those of firm B. Variables such as weight, time, and distance are also measured on a ratio scale because zero is meaningful.

Unlike nominal- and ordinal-scaled variables (categorical variables), arithmetic operations are valid on interval- and ratio-scaled variables (numerical variables). In later chapters, we will calculate summary measures, such as the mean, the median, and the variance, for numerical variables; we cannot calculate these measures for categorical variables.

MEASUREMENT SCALES

The observations for any variable can be classified into one of four major measurement scales: nominal, ordinal, interval, or ratio.

- Nominal: Observations differ merely by name or label.
- Ordinal: Observations can be categorized and ranked; however, differences between the ranked observations are meaningless.
- Interval: Observations can be categorized and ranked, and differences between observations are meaningful. The main drawback of the interval scale is that the value of zero is arbitrarily chosen.
- Ratio: Observations have all the characteristics of an interval-scaled variable as well as a true zero point; thus, meaningful ratios can be calculated.

Nominal and ordinal scales are used for categorical variables, whereas interval and ratio scales are used for numerical variables.

FILE
Tween_Survey

EXAMPLE 1.3

The owner of a ski resort two hours outside Boston, Massachusetts, is interested in serving the needs of the "tween" population (children aged 8 to 12 years old). He believes that tween spending power has grown over the past few years, and he wants their skiing experience to be memorable so that they want to return. At the end of last year's ski season, he asked 20 tweens the following four questions.

- Q1. On your car drive to the resort, which music streaming service was playing?
- Q2. On a scale of 1 to 4, rate the quality of the food at the resort (where 1 is poor, 2 is fair, 3 is good, and 4 is excellent).
- Q3. Presently, the main dining area closes at 3:00 pm. What time do you think it should close?
- Q4. How much of your own money did you spend at the lodge today?

A portion of their responses is shown in Table 1.6. Identify the scale of measurement for each variable used in the survey. Given the tween responses, provide suggestions to the owner for improvement.

TABLE 1.6 Tween Responses to Resort Survey

Tween	Question 1	Question 2	Question 3	Question 4
1	Apple Music	4	5:00 pm	20
2	Pandora	2	5:00 pm	10
⋮	⋮	⋮	⋮	⋮
20	Spotify	2	4:30 pm	10

SOLUTION:

- Q1. Responses for music streaming service are nominal because the observations differ merely in label. Twelve of the 20 tweens, or 60%, listened to Spotify. If the resort wishes to contact tweens using this means, then it may want to direct its advertising dollars to this streaming service.
- Q2. Food quality responses are on an ordinal scale because we can both categorize and rank the observations. Eleven of the 20 tweens, or 55%, felt that the food quality was, at best, fair. Perhaps a more extensive survey that focuses solely on food quality would reveal the reason for their apparent dissatisfaction.
- Q3. Closing time responses are on an interval scale. We can say that 3:30 pm is 30 minutes later than 3:00 pm, and 6:00 pm is 30 minutes later than 5:30 pm; that is, differences between observations are meaningful. The closing time responses, however, have no apparent zero point. We could arbitrarily define the zero point at 12:00 am, but ratios are still meaningless. In other words, it makes no sense to form the ratio 6:00 pm/3:00 pm and conclude that 6:00 pm is twice as long a time period as 3:00 pm. A review of the closing time responses shows that the vast majority (19 out of 20) would like the dining area to remain open later.
- Q4. The tweens' responses with respect to their own money spent at the resort are on a ratio scale. We can categorize and rank observations as well as calculate meaningful differences. Moreover, because there is a natural zero point, valid ratios can also be calculated. Seventeen of the 20 tweens spent their own money at the lodge. It does appear that the discretionary spending of this age group is significant. The owner would be wise to cater to some of their preferences.

EXERCISES 1.3

Applications

14. Which of the following variables are categorical and which are numerical? If the variable is numerical, then specify whether the variable is discrete or continuous.
    a. Points scored in a football game.
    b. Racial composition of a high school classroom.
    c. Heights of 15-year-olds.

15. Which of the following variables are categorical and which are numerical? If the variable is numerical, then specify whether the variable is discrete or continuous.
    a. Colors of cars in a mall parking lot.
    b. Time it takes each student to complete a final exam.
    c. The number of patrons who frequent a restaurant.

16. In each of the following scenarios, define the type of measurement scale.
    a. A kindergarten teacher marks whether each student is a boy or a girl.
    b. A ski resort records the daily temperature during the month of January.
    c. A restaurant surveys its customers about the quality of its waiting staff on a scale of 1 to 4, where 1 is poor and 4 is excellent.

17. In each of the following scenarios, define the type of measurement scale.
    a. An investor collects data on the weekly closing price of gold throughout the year.
    b. An analyst assigns a sample of bond issues to one of the following credit ratings, given in descending order of credit quality (increasing probability of default): AAA, AA, BBB, BB, CC, D.
    c. The dean of the business school at a local university categorizes students by major (i.e., accounting, finance, marketing, etc.) to help in determining class offerings in the future.

18. In each of the following scenarios, define the type of measurement scale.
    a. A meteorologist records the amount of monthly rainfall over the past year.
    b. A sociologist notes the birth year of 50 individuals.
    c. An investor monitors the daily stock price of BP following the 2010 oil disaster in the Gulf of Mexico.

19. FILE ***Major.*** A professor records the majors of her 30 students. The data accompanying this exercise contain the relevant information.
    a. What is the measurement scale of the Major variable?
    b. Summarize the results in tabular form.
    c. What information can be extracted from the data?

20. FILE ***DOW.*** The accompanying data set contains information on the 30 companies that comprise the Dow Jones Industrial Average (DJIA). For each company, the data set lists the year that it joined the DJIA, its industry, and its stock price (in $) as of February 15, 2019.
    a. What is the measurement scale of the Industry variable?
    b. What is the measurement scale of the Year variable? What are the strengths of this type of measurement scale? What are its weaknesses?
    c. What is the measurement scale of the Price variable? What are the strengths of this type of measurement scale?

21. FILE ***Retailer.*** An online retail company is trying to predict customer spending in the first three months of the year. Brian Duffy, the marketing analyst of the company, has compiled a data set on 200 existing customers that includes sex (Sex: Female/Male), annual income in 1,000s (Income), age (Age, in years), and total spending in the first three months of the year (Spending).
    a. Which of the above variables are categorical and which are numerical?
    b. What is the measurement scale of each of the above variables?

22. FILE ***Vacation.*** Vacation destinations often run on a seasonal basis, depending on the primary activities in that location. Amanda Wang is the owner of a travel agency in Cincinnati, Ohio. She has compiled a data set of the number of vacation packages (Vacation) that she has sold over the last 12 years.
    a. What is the measurement scale of the Year variable? What are the strengths of this type of measurement scale? What are its weaknesses?
    b. What is the measurement scale of the Quarter variable? What is a weakness of this type of measurement scale?
    c. What is the measurement scale of the Vacation variable? What are the strengths of this type of measurement scale?

1.4 DATA SOURCES AND FILE FORMATS

LO **1.4**

Describe different data sources and file formats.

The explosion in the field of statistics and data analytics is partly due to the growing availability of vast amounts of data and improved computational power. Many experts believe that 90% of the data in the world today was created in the last two years alone. Not surprisingly, businesses continue to grapple with how to best ingest, understand, and operationalize large volumes of data.

We access much of the data in this text by simply using a search engine like Google. These search engines direct us to data-providing sites. For instance, searching for economic data leads you to the Bureau of Economic Analysis (http://bea.gov), the Bureau of Labor Statistics (http://www.bls.gov), the Federal Reserve Economic Data (https://research.stlouisfed.org), and the U.S. Census Bureau (http://www.census.gov). These websites provide data on inflation, unemployment, gross domestic product (GDP), and much more, including useful international data. Similarly, excellent world development indicator data are available at http://data.worldbank.org.

The Wall Street Journal, *The New York Times*, *USA Today*, *The Economist*, *Business Week*, *Forbes*, and *Fortune* are all reputable publications that provide all sorts of data. We would like to point out that all of these data sources represent only a small portion of publicly available data. In this text, we have compiled a number of big data sets, based on online data sources, that are integrated throughout the text.

As people work more and more in collaboration with one another, a need usually arises for an ability to exchange information between different parties. Formatting data in an agreed-upon or standardized manner is important for allowing other people to understand the data contained in a file. There are many standards for file formats. For example, a text file can be organized into rows and columns to store in a table. Two common layouts for simple text files are a fixed-width format and a delimited format. In addition to text files, we can use a markup language to provide a structure to data. Three widely used markup languages are eXtensible Markup Language (XML), HyperText Markup Language (HTML), and JavaScript Object Notation (JSON). We now provide an overview of these formats and markup languages.

Fixed-Width Format

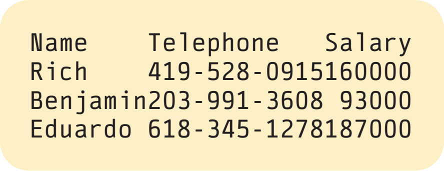
In a data file with a **fixed-width format** (or fixed-length format), each column starts and ends at the same place in every row. The actual data are stored as plain text characters. Consider the information in Table 1.7. It shows the first name, telephone number, and annual salary for three individuals.

TABLE 1.7 Sample Data for Format Illustration

Name	Telephone	Salary
Rich	419-528-0915	160000
Benjamin	203-991-3608	93000
Eduardo	618-345-1278	187000

The information in Table 1.7 can be organized into a fixed-width format as shown in Figure 1.6. The first, second, and third columns of Figure 1.6 are defined to have column widths of 8, 12, and 7 characters, respectively. Every observation or record has the exact same column widths. The fixed-width file has the simplicity of design where specific data can be found at the exact same location for every record. This can help speed up record search when the data set is very large. Furthermore, because only raw data are stored, fixed-width files tend to be significantly smaller in size compared to other data formats such as XML that include data labels and tags. However, the number of characters of each column (the column width) needs to be predetermined. In Figure 1.6, the Name column is predefined to have at most 8 characters; any names with more than 8 characters will be truncated. Finally, at times, the columns seem to run into each other. For these reasons, other formats are more popular.

FIGURE 1.6 A fixed-width file format



```
Name    Telephone   Salary
Rich    419-528-0915160000
Benjamin203-991-3608 93000
Eduardo 618-345-1278187000
```

Delimited Format

Another widely used file format to store tabular data is a **delimited format**. In Figure 1.7, we show the information in Table 1.7 in a delimited format, where each piece of data is separated by a comma.

FIGURE 1.7 A comma-separated value (csv) file format

```
Name,Telephone,Salary
Rich,419-528-0915,160000
Benjamin,203-991-3608,93000
Eduardo,618-345-1278,187000
```

In a delimited format, a comma is called a delimiter, and the file is called a comma-delimited or comma-separated value (csv) file. Sometimes, other characters such as semicolons are used as delimiters. In a delimited file, each piece of data can contain as many characters as applicable. For example, unlike the fixed-width file shown in Figure 1.6, a comma-separated value file does not limit a person's name to only eight characters.

Fixed-width and delimited files usually include plain text data that can be opened in most text editing software such as Microsoft Word and Notepad in Microsoft Windows, TextEdit in Apple's Mac, and online tools such as Google Docs.

eXtensible Markup Language

The **eXtensible Markup Language (XML)** is a simple language for representing structured data. XML is one of the most widely used formats for sharing structured information between computer programs, between people, and between computers and people. It uses markup tags to define the structure of data. Using the information from Table 1.7, Figure 1.8 shows an example of data coded in XML format.

FIGURE 1.8 An XML file format

```
<Data>
<Person>
     <Name>Rich</Name>
     <Telephone>419-528-915</Telephone>
     <Salary>160000</Salary>
</Person>
<Person>
     <Name>Benjamin</Name>
     <Telephone>203-991-608</Telephone>
     <Salary>93000</Salary>
</Person>
<Person>
     <Name>Eduardo</Name>
     <Telephone>618-345-278</Telephone>
     <Salary>187000</Salary>
</Person>
</Data>
```

Each piece of data is usually enclosed in a pair of 'tags' that follow specific XML syntax. For example, a telephone number starts with an opening tag (<Telephone>) and ends with a closing tag (</Telephone>). The XML code is case-sensitive; therefore, <Telephone> and <telephone> would indicate two different pieces of information. The tags in Figure 1.8 are not based on any predefined standard. The XML language allows each user to define his or her own tags and document structure, but XML tag names should be self-explanatory. The XML file format is designed to support readability. This makes the XML file format especially suitable for transporting data between computer applications without losing the meanings of the data. However, due to the additional labels and tags, XML data files tend to be much larger in size than fixed-width and delimited data files, making downloading and parsing more time-consuming and computationally intensive.

HyperText Markup Language

Like XML, the **HyperText Markup Language (HTML)** is a mark-up language that uses tags to define its data in web pages. The key distinction between XML and HTML is that XML tells us or computer applications what the data are, whereas HTML tells the web browser how to display the data. Using the information from Table 1.7, Figure 1.9 shows an example of data coded in HTML.

FIGURE 1.9 An HTML file format

```
<table>
  <tr>
    <th>Name</th>
    <th>Telephone</th>
    <th>Salary</th>
  </tr>
  <tr>
    <td>Rich</td>
    <td>419-528-0915</td>
    <td>160000</td>
  </tr>
  <tr>
    <td>Benjamin</td>
    <td>203-991-3608</td>
    <td>93000</td>
  </tr>
  <tr>
    <td>Eduardo</td>
    <td>618-345-1278</td>
    <td>187000</td>
  </tr>
</table>
```

Tags such as <table> are used to provide structure for textual data, such as headings, paragraphs, and tables. In Figure 1.9, the opening <table> and closing </table> tags indicate the beginning and completion of a table. Unlike XML where users can define their own markup tags, HTML tags conform to standards maintained by organizations such as the World Wide Web Consortium (W3C). Web browsers such as Google Chrome and Safari are designed to interpret the HTML code that follows these standards. For example, the tag <th> is understood by web browsers as a table heading. In our example, there are three columns and headings (i.e., Name, Telephone, and Salary). The <tr> tag, on the other hand, defines a row, and the <td> tag defines each cell within a row. Unlike XML, HTML is case-insensitive.

JavaScript Object Notation

The **JavaScript Object Notation (JSON)** has become a popular alternative to XML in recent years as open data sharing has grown in popularity. JSON is a standard for transmitting human-readable data. Originally a subset of the JavaScript syntax, JSON is currently a data standard supported by a wide range of modern programming languages such as C, Java, and Python. Using the information from Table 1.7, Figure 1.10 shows an example of data coded in JSON format.

FIGURE 1.10 A JSON file format

```
{
  “Person”:    [
     {
       “Name”: “Rich”,
       “Telephone”: “419-528-0915”,
       “Salary”: “160000”
     },
     {
       “Name”: “Benjamin”,
       “Telephone”: “203-991-3608”,
       “Salary”: “93000”
     },
     {
       “Name”: “Eduardo”,
       “Telephone”: “618-345-1278”,
       “Salary”: “187000”
     }
           ]
}
```

The JSON format is self-explanatory just as the XML format is, but it offers several advantages over the XML format. First, the JSON format is not as verbose as the XML format, making data files smaller in size. The difference in size is especially noticeable for very large data sets. Second, the JSON format supports a wide range of data types not readily available in the XML format. Finally, parsing JSON data files is faster and less resource intensive. For these reasons, the JSON format has become a widely adopted standard for open data sharing.

DATA FILE FORMATS AND MARKUP LANGUAGES

There are many standards for data file formats. Two common layouts for simple text files are the fixed-width format and the delimited format.

- With a fixed-width format, each column has a fixed width and starts and ends at the same place in every row.
- With a delimited format, each column is separated by a delimiter such as a comma. Each column can contain as many characters as applicable.

Markup languages also provide a structure to data. Three widely used languages are the eXtensible Markup Language (XML), the HyperText Markup Language (HTML), and the JavaScript Object Notation (JSON).

- XML is a simple text-based markup language for representing structured data. It uses user-defined markup tags to specify the structure of data.
- HTML is a simple text-based markup language for displaying content in web browsers.
- JSON is a standard for transmitting human-readable data in compact files.

EXERCISES 1.4

Applications

23. A used car salesperson recently sold two Mercedes, three Toyota, six Ford, and four Hyundai sedans. He wants to record the sales data.
    a. Organize the data into a fixed-width format (eight characters for brand name and four characters for the number of cars sold).
    b. Organize the data in a delimited format.
    c. Code the data in XML format.
    d. Code the data in HTML table format.
    e. Code the data in JSON format.

24. Last year, Oracle hired three finance majors from the local university. Robert Schneider started with a salary of $56,000, Chun Zhang with $52,000, Sunil Banerjee with $58,000, and Linda Jones with $60,000. Oracle wants to record the hiring data.
    a. Organize the data into a fixed-width format (10 characters for first name, 10 characters for last name, and six characters for salary).
    b. Organize the data in a delimited format.
    c. Code the data in XML format.
    d. Code the data in HTML table format.
    e. Code the data in JSON format.

25. The following table lists the population, in millions, in India and China, the two most populous countries in the world, for the years 2013 through 2017.

Year	India	China
2013	1278.56	1357.38
2014	1293.86	1364.27
2015	1309.05	1371.22
2016	1324.17	1378.67
2017	1339.18	1386.40

    a. Organize the data into a fixed-width format (four characters for Year, eight characters for India, and eight characters for China).
    b. Organize the data in a delimited format.

26. The following table lists the top five countries in the world in terms of their happiness index on a 10-point scale, and their corresponding GDP per capita as reported by the United Nations in 2017.

Country	Happiness	GDP
Finland	7.769	45670
Denmark	7.600	57533
Norway	7.544	75295
Iceland	7.494	73060
Netherlands	7.488	48754

    a. Organize the data into a fixed-width format (11 characters for Country, 10 characters for Happiness, and six characters for GDP).
    b. Organize the data in a delimited format.

27. The following three students were honored at a local high school for securing admissions to prestigious universities.

Name	University
Bridget	Yale
Minori	Stanford
Matthew	Harvard

    a. Organize the data into a fixed-width format (10 characters for Name and 10 characters for University).
    b. Organize the data in a delimited format.
    c. Code the data in XML format.
    d. Code the data in HTML table format.
    e. Code the data in JSON format.

28. According to *Forbes,* Michael Trout of the Los Angeles Angels, with earnings of $39 million, was the highest-paid player in baseball in 2019. Bryce Harper of the Philadelphia Phillies ranked second at $36.5 million. The Boston Red Sox pitcher David Price was baseball's third-highest-paid player at $32 million.
    a. Code the data on the player name, baseball team, and salary in the XML format.
    b. Repeat part a using the HTML table format.
    c. Repeat part a using the JSON format.

29. According to *U.S. News and World Report,* a statistician was the best business profession in 2019 with 12,600 projected jobs and a median salary of $84,060. A mathematician was the second-best profession with 900 projected jobs and a median salary of $103,010. Interestingly, both of these career paths are related to data analytics.
    a. Code the information on the profession, projected jobs, and median salary in XML format.
    b. Repeat part a using the HTML table format.
    c. Repeat part a using the JSON format.

1.5 WRITING WITH BIG DATA

As mentioned at the beginning of this chapter, data are not very useful unless they are converted into insightful information that is clearly articulated in written or verbal language. As such, an important aspect of business analytics is to communicate with numbers rather than focus on number crunching. In this and subsequent chapters, we include a sample report based on observations and analysis of data to convey the information in written form. These reports are intended for a nontechnical audience who may not be familiar with the details of the statistical and computational methods. Consider the following case study and accompanying report based on music popularity data.

Case Study

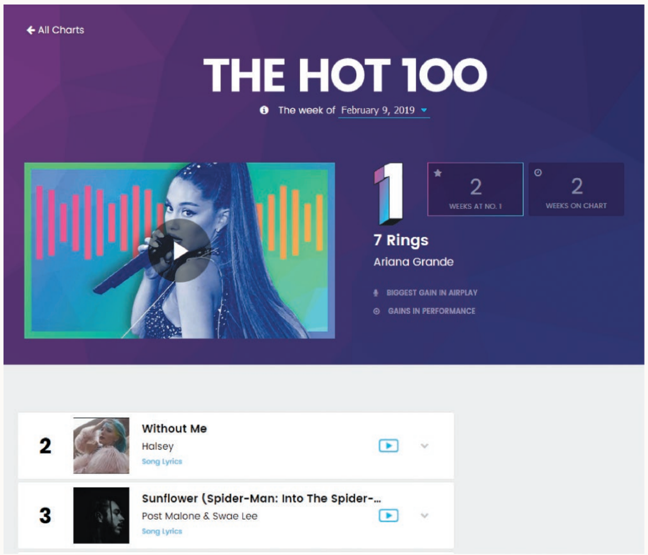
Since 1940, *Billboard* magazine has published a variety of weekly music popularity charts. Today, the magazine uses a combination of sales volume, airplay on radio, digital downloads, and online streams in order to determine the chart rankings. One of the most highly watched *Billboard* charts is the Hot 100, which provides a weekly ranking of the top 100 music singles. Each entry on the Hot 100 chart lists the current rank, last week's rank, highest position, and the number of weeks the song has been on the chart. Other *Billboard* charts rank music by genre such as Pop, Rock, R&B, and Latin or rank the popularity of music albums and artists.

Maya Alexander is a reporter for her university's newspaper, *The Campus Gazette*. She wants to launch a new Film & Music column that will include commentary, summarized data, and statistics on music popularity. In an effort to convince her editor to give her this assignment, Maya researches and evaluates what *The Campus Gazette* might be able to use from the *Billboard* website (http://www.billboard.com).

Sample Report— Billboard Charts

Music is an integral part of campus life. A new Film & Music column in *The Campus Gazette* is likely to be very popular among the campus audience. The *Billboard* website publishes weekly data on music popularity charts ranging from the top 100 singles (Hot 100) to digital sales by music genre. We can summarize these numerical data for music genres that are most popular among college students and publish them in the new Film & Music column. The popularity charts on the *Billboard* website are coded in the HTML data format and can be easily imported into a table. For example, the Hot 100 chart shown in Figure 1.11 coded as an HTML table can be readily downloaded into a table in a text document.

FIGURE 1.11 *Billboard*'s Hot 100 chart



THE HOT 100

The week of February 9, 2019

1
2 WEEKS AT NO. 1
2 WEEKS ON CHART
7 Rings
Ariana Grande
BIGGEST GAIN IN AIRPLAY
GAINS IN PERFORMANCE

2 Without Me
Halsey
Song Lyrics

3 Sunflower (Spider-Man: Into The Spider-...
Post Malone & Swae Lee
Song Lyrics

Source: Billboard.com

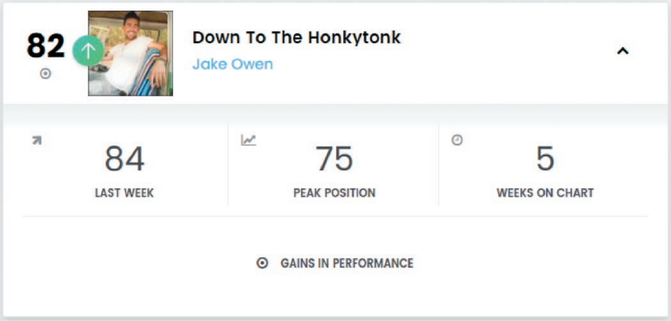
Our readers may also be interested in following the top music singles by genre. For example, Table 1.8 is an example of a summary table listing the top five singles in Country, Pop, and Rock, which is easily compiled from three different *Billboard* charts. This summary table can be complemented with written commentaries based on information provided by the popularity charts such as the highest chart position and the number of weeks on the chart for each song.

TABLE 1.8 Top Five Hot Singles by Music Genre

	Country		Pop		Rock	
Rank	Song	Artist	Song	Artist	Song	Artist
1	Tequila	Dan + Shay	Without Me	Halsey	High Hopes	Panic! At the Disco
2	Speechless	Dan + Shay	Thank U, Next	Ariana Grande	Natural	Imagine Dragons
3	Meant To Be	Bebe Rexha	High Hopes	Panic! At the Disco	Broken	lovelytheband
4	Beautiful Crazy	Luke Combs	East Side	benny blanco	Bad Liar	Imagine Dragons
5	Girl Like You	Jason Aldean	Sunflower	Post Malone	Harmony Hall	Vampire Weekend

According to the most recent sales and market performance, the *Billboard* website also organizes songs and albums into categories such as "Gains in Performance," "Biggest Gain in Streams," and "Biggest Gain in Digital Sales." These songs and albums have not yet reached the top five positions, but their rankings are quickly moving up on the popularity charts. *The Campus Gazette* can establish itself as a music trendsetter on campus by introducing our readers to these up-and-coming songs and albums in a new Film & Music column. Figure 1.12 shows an example of an up-and-coming single on the Hot 100 chart. Similar to the popularity charts, these music categories are formatted using the HTML standard on the *Billboard* website and can be readily imported into a text document. We can create commentaries on selected songs and albums from these lists to introduce up-and-coming music to our readers.

FIGURE 1.12 Up-and-coming music on *Billboard* charts



82

Down To The Honkytonk

Jake Owen

84 LAST WEEK

75 PEAK POSITION

5 WEEKS ON CHART

GAINS IN PERFORMANCE

Source: Billboard.com

Suggested Case Studies

As discussed in the chapter, data from an endless number of online sources are available for us to explore and investigate. Here are some suggested case studies using online, publicly available data.

Report 1.1. Finland is the happiest country in the world, according to the 2018 Happiness Index Report by the United Nations (http://www.worldhappiness.report). In fact, several Scandinavian countries have consistently held the top spots among the 156 countries included in the annual Happiness Index Report in the past several years. Visit the Happiness Index website, explore, and write a report based on the current data provided on the website.

Report 1.2. Millions of tourists visit Yosemite National Park in California each year. Stunning waterfalls, giant redwood trees, and spectacular granite rock formations are among the main attractions at the iconic park. However, the winding roads leading to the Yosemite Valley may be closed occasionally due to severe weather conditions. Visit a weather forecast website such as http://www.weather.com and explore the weather data around Yosemite Park. Write a report to advise a tourist planning a visit.

Report 1.3. A novel coronavirus, SARS-CoV-2, caused a worldwide pandemic of COVID-19 disease. Many public health organizations gathered and provided COVID-19 data for scientists, epidemiologists, and the general public to study the spread of the disease. Visit a public health website such as the World Health Organization (https://covid19.who.int/) or the U.S. Centers for Disease Control and Prevention (https://www.cdc.gov/) and explore the COVID-19 data. Write a report related to the pandemic in your geographical area.