CENGAGE

Camm Cochran Fry Ohlmann

Business Analytics

Descriptive • Predictive • Prescriptive

Business Analytics

Descriptive • Predictive • Prescriptive

Jeffrey D. Camm
Wake Forest University

James J. Cochran
University of Alabama

Michael J. Fry
University of Cincinnati

Jeffrey W. Ohlmann
University of Iowa

CENGAGE

Australia • Brazil • Mexico • Singapore • United Kingdom • United States

CENGAGE

***Business Analytics*, Fourth Edition**
Jeffrey D. Camm, James J. Cochran,
Michael J. Fry, Jeffrey W. Ohlmann

Senior Vice President, Higher Education & Skills Product: Erin Joyner

Product Director: Jason Fremder

Senior Product Manager: Aaron Arnsparger

Senior Content Manager: Conor Allen

Product Assistant: Maggie Russo

Marketing Manager: Chris Walz

Senior Learning Designer: Brandon Foltz

Digital Delivery Lead: Mark Hopkinson

Intellectual Property Analyst: Ashley Maynard

Intellectual Property Project Manager: Kelli Besse

Production Service: MPS Limited

Senior Project Manager, MPS Limited: Santosh Pandey

Art Director: Chris Doughman

Text Designer: Beckmeyer Design

Cover Designer: Beckmeyer Design

Cover Image: iStockPhoto.com/tawanlubfah

For product information and technology assistance, contact us at
Cengage Customer & Sales Support, 1-800-354-9706
or support.cengage.com.

For permission to use material from this text or product,
submit all requests online at
www.cengage.com/permissions.

Library of Congress Control Number: 2019921119

ISBN: 978-0-357-13178-7

Loose-leaf Edition:
ISBN: 978-0-357-13179-4

Cengage
200 Pier 4 Boulevard
Boston, MA 02210
USA

Cengage is a leading provider of customized learning solutions with employees residing in nearly 40 different countries and sales in more than 125 countries around the world. Find your local representative at **www.cengage.com**.

Cengage products are represented in Canada by Nelson Education, Ltd.

To learn more about Cengage platforms and services, register or access your online learning solution, or purchase materials for your course, visit **www.cengage.com**.

Printed in the United States of America
Print Number: 01 Print Year: 2020

Brief Contents

Contents

Available in the MindTap Reader:

About the Authors

Jeffrey D. Camm. is the Inmar Presidential Chair and Associate Dean of Business Analytics in the School of Business at Wake Forest University. Born in Cincinnati, Ohio, he holds a B.S. from Xavier University (Ohio) and a Ph.D. from Clemson University. Prior to joining the faculty at Wake Forest, he was on the faculty of the University of Cincinnati. He has also been a visiting scholar at Stanford University and a visiting professor of business administration at the Tuck School of Business at Dartmouth College.

Dr. Camm has published over 40 papers in the general area of optimization applied to problems in operations management and marketing. He has published his research in *Science*, *Management Science*, *Operations Research*, *Interfaces*, and other professional journals. Dr. Camm was named the Dornoff Fellow of Teaching Excellence at the University of Cincinnati and he was the 2006 recipient of the INFORMS Prize for the Teaching of Operations Research Practice. A firm believer in practicing what he preaches, he has served as an operations research consultant to numerous companies and government agencies. From 2005 to 2010 he served as editor-in-chief of *Interfaces*. In 2016, Professor Camm received the George E. Kimball Medal for service to the operations research profession, and in 2017 he was named an INFORMS Fellow.

James J. Cochran. James J. Cochran is Associate Dean for Research, Professor of Applied Statistics and the Rogers-Spivey Faculty Fellow at The University of Alabama. Born in Dayton, Ohio, he earned his B.S., M.S., and M.B.A. from Wright State University and his Ph.D. from the University of Cincinnati. He has been at The University of Alabama since 2014 and has been a visiting scholar at Stanford University, Universidad de Talca, the University of South Africa and Pole Universitaire Leonard de Vinci.

Dr. Cochran has published more than 40 papers in the development and application of operations research and statistical methods. He has published in several journals, including *Management Science*, *The American Statistician*, *Communications in Statistics—Theory and Methods*, *Annals of Operations Research*, *European Journal of Operational Research*, *Journal of Combinatorial Optimization*, *Interfaces*, and *Statistics and Probability Letters*. He received the 2008 INFORMS Prize for the Teaching of Operations Research Practice, 2010 Mu Sigma Rho Statistical Education Award and 2016 Waller Distinguished Teaching Career Award from the American Statistical Association. Dr. Cochran was elected to the International Statistics Institute in 2005, named a Fellow of the American Statistical Association in 2011, and named a Fellow of INFORMS in 2017. He also received the Founders Award in 2014 and the Karl E. Peace Award in 2015 from the American Statistical Association, and he received the INFORMS President's Award in 2019.

A strong advocate for effective operations research and statistics education as a means of improving the quality of applications to real problems, Dr. Cochran has chaired teaching effectiveness workshops around the globe. He has served as an operations research consultant to numerous companies and not-for-profit organizations. He served as editor-in-chief of *INFORMS Transactions on Education* and is on the editorial board of *INFORMS Journal of Applied Analytics*, *International Transactions in Operational Research*, and *Significance*.

Michael J. Fry. Michael J. Fry is Professor of Operations, Business Analytics, and Information Systems (OBAIS) and Academic Director of the Center for Business Analytics in the Carl H. Lindner College of Business at the University of Cincinnati. Born in Killeen, Texas, he earned a B.S. from Texas A&M University, and M.S.E. and Ph.D. degrees from the University of Michigan. He has been at the University of Cincinnati since 2002, where he served as Department Head from 2014 to 2018 and has been named a Lindner Research Fellow. He has also been a visiting professor at Cornell University and at the University of British Columbia.

Professor Fry has published more than 25 research papers in journals such as *Operations Research*, *Manufacturing & Service Operations Management*, *Transportation Science*, *Naval Research Logistics*, *IIE Transactions*, *Critical Care Medicine*, and *Interfaces*. He serves on editorial boards for journals such as *Production and Operations Management*, *INFORMS Journal of Applied Analytics* (formerly *Interfaces*), and *Journal of Quantitative Analysis in Sports*. His research interests are in applying analytics to the areas of supply chain management, sports, and public-policy operations. He has worked with many different organizations for his research, including Dell, Inc., Starbucks Coffee Company, Great American Insurance Group, the Cincinnati Fire Department, the State of Ohio Election Commission, the Cincinnati Bengals, and the Cincinnati Zoo & Botanical Gardens. In 2008, he was named a finalist for the Daniel H. Wagner Prize for Excellence in Operations Research Practice, and he has been recognized for both his research and teaching excellence at the University of Cincinnati. In 2019, he led the team that was awarded the INFORMS UPS George D. Smith Prize on behalf of the OBAIS Department at the University of Cincinnati.

Jeffrey W. Ohlmann. Jeffrey W. Ohlmann is Associate Professor of Business Analytics and Huneke Research Fellow in the Tippie College of Business at the University of Iowa. Born in Valentine, Nebraska, he earned a B.S. from the University of Nebraska, and M.S. and Ph.D. degrees from the University of Michigan. He has been at the University of Iowa since 2003.

Professor Ohlmann's research on the modeling and solution of decision-making problems has produced more than two dozen research papers in journals such as *Operations Research*, *Mathematics of Operations Research*, *INFORMS Journal on Computing*, *Transportation Science*, and the *European Journal of Operational Research*. He has collaborated with companies such as Transfreight, LeanCor, Cargill, the Hamilton County Board of Elections, and three National Football League franchises. Because of the relevance of his work to industry, he was bestowed the George B. Dantzig Dissertation Award and was recognized as a finalist for the Daniel H. Wagner Prize for Excellence in Operations Research Practice.

Preface

Business Analytics 4E is designed to introduce the concept of business analytics to undergraduate and graduate students. This edition builds upon what was one of the first collections of materials that are essential to the growing field of business analytics. In Chapter 1, we present an overview of business analytics and our approach to the material in this textbook. In simple terms, business analytics helps business professionals make better decisions based on data. We discuss models for summarizing, visualizing, and understanding useful information from historical data in Chapters 2 through 6. Chapters 7 through 9 introduce methods for both gaining insights from historical data and predicting possible future outcomes. Chapter 10 covers the use of spreadsheets for examining data and building decision models. In Chapter 11, we demonstrate how to explicitly introduce uncertainty into spreadsheet models through the use of Monte Carlo simulation. In Chapters 12 through 14, we discuss optimization models to help decision makers choose the best decision based on the available data. Chapter 15 is an overview of decision analysis approaches for incorporating a decision maker's views about risk into decision making. In Appendix A we present optional material for students who need to learn the basics of using Microsoft Excel. The use of databases and manipulating data in Microsoft Access is discussed in Appendix B. Appendixes in many chapters illustrate the use of additional software tools such as R, JMP Pro and Tableau to apply analytics methods.

This textbook can be used by students who have previously taken a course on basic statistical methods as well as students who have not had a prior course in statistics. *Business Analytics* 4E is also amenable to a two-course sequence in business statistics and analytics. All statistical concepts contained in this textbook are presented from a business analytics perspective using practical business examples. Chapters 2, 4, 6, and 7 provide an introduction to basic statistical concepts that form the foundation for more advanced analytics methods. Chapters 3, 5, and 9 cover additional topics of data visualization and data mining that are not traditionally part of most introductory business statistics courses, but they are exceedingly important and commonly used in current business environments. Chapter 10 and Appendix A provide the foundational knowledge students need to use Microsoft Excel for analytics applications. Chapters 11 through 15 build upon this spreadsheet knowledge to present additional topics that are used by many organizations that are leaders in the use of prescriptive analytics to improve decision making.

Updates in the Fourth Edition

The fourth edition of *Business Analytics* is a major revision. We have added online appendixes for many topics in Chapters 1 through 9 that introduce the use of R, the exceptionally popular open-source software for analytics. *Business Analytics* 4E also includes an appendix to Chapter 3 introducing the powerful data visualization software Tableau. We have further enhanced our data mining chapters to allow instructors to choose their preferred means of teaching this material in terms of software usage. We have expanded the number of conceptual homework problems in both Chapters 5 and 9 to increase the number of opportunities for students learn about data mining and solve problems without the use of data mining software. Additionally, we now include online appendixes on using JMP Pro and R for teaching data mining so that instructors can choose their favored way of teaching this material. Other changes in this edition include an expanded discussion of binary variables for integer optimization in Chapter 13, an additional example in Chapter 11 for Monte Carlo simulation, and new and revised homework problems and cases.

- **Tableau Appendix for Data Visualization.** Chapter 3 now includes a new appendix that introduces the use of the software Tableau for data visualization. Tableau is a very powerful software for creating meaningful data visualizations that can be used to display, and to analyze, data. The appendix includes step-by-step directions for generating many of the charts used in Chapters 2 and 3 in Tableau.

- **Incorporation of R.** R is an exceptionally powerful open-source software that is widely used for a variety of statistical and analytics methods. We now include online appendixes that introduce the use of R for many of the topics covered in Chapters 1 through 9, including data visualization and data mining. These appendixes include step-by-step directions for using R to implement the methods described in these chapters. To facilitate the use of R, we introduce RStudio, an open-source integrated development environment (IDE) that provides a menu-driven interface for R. For Chapters 5 and 9 that cover data mining, we introduce the use of Rattle, a library package providing a graphical-user interface for R specifically tailored for data mining functionality. The use of RStudio and Rattle eases the learning curve of using R so that students can focus on learning the methods and interpreting the output.
- **Updates for Data Mining Chapters.** Chapters 5 and 9 have received extensive updates. We have moved the Descriptive Data Mining chapter to Chapter 5 so that it is located after our chapter on Probability. This allows us to use probability concepts such as conditional probability to explain association rule measures. Additional content on text mining and further discussion of ways to measure distance between observations have been added to a reorganized Descriptive Data Mining chapter. Descriptions of cross-validation approaches, methods of addressing class imbalanced data, and out-of-bag estimation in ensemble methods have been added to Chapter 9 on Predictive Data Mining. The end-of-chapter problems in Chapters 5 and 9 have been revised and generalized to accommodate the use of a wide range of data mining software. To allow instructors to choose different software for use with these chapters, we have created online appendixes for both JMP Pro and R. JMP has introduced a new version of its software (JMP Pro 14) since the previous edition of this textbook, so we have updated our JMP Pro output and step-by-step instructions to reflect changes in this software. We have also written online appendixes for Chapters 5 and 9 that use R and the graphical-user interface Rattle to introduce topics from these chapters to students. The use of Rattle removes some of the more difficult line-by-line coding in R to perform many common data mining techniques so that students can concentrate on learning the methods rather than coding syntax. For some data mining techniques that are not available in Rattle, we show how to accomplish these methods using R code. And for all of our textbook examples, we include the exact R code that can be used to solve the examples. We have also added homework problems to Chapters 5 and 9 that can be solved without using any specialized software. This allows instructors to cover the basics of data mining without introducing any additional software. The online appendixes for Chapters 5 and 9 also include JMP Pro and R specific instructions for how to solve the end-of-chapter problems and cases using JMP Pro and R. Problem and case solutions using both JMP Pro and R are also available to instructors.
- **Additional Simulation Model Example.** We have added an additional example of a simulation model in Chapter 11. This new example helps bridge the gap in the difficultly levels of the previous examples. The new example also gives students additional information on how to build and interpret simulation models.
- **New Cases.** *Business Analytics* 4E includes nine new end-of-chapter cases that allow students to work on more extensive problems related to the chapter material and work with larger data sets. We have also written two new cases that require the use of material from multiple chapters. This helps students understand the connections between the material in different chapters and is more representative of analytics projects in practice where the methods used are often not limited to a single type.
- **Legal and Ethical Issues Related to Analytics and Big Data.** Chapter 1 now includes a section that discusses legal and ethical issues related to analytics and the use of big data. This section discusses legal issues related to the protection of data as well as ethical issues related to the misuse and unintended consequences of analytics applications.

- **New End-of-Chapter Problems.** The fourth edition of this textbook includes more than 20 new problems. We have also revised many of the existing problems to update and improve clarity. Each end-of-chapter problem now also includes a short header to make the application of the exercise more clear. As we have done in past editions, Excel solution files are available to instructors for problems that require the use of Excel. For problems that require the use of software in the data-mining chapters (Chapters 5 and 9), we include solutions for both JMP Pro and R/Rattle.

Continued Features and Pedagogy

In the fourth edition of this textbook, we continue to offer all of the features that have been successful in the first two editions. Some of the specific features that we use in this textbook are listed below.

- **Integration of Microsoft Excel:** Excel has been thoroughly integrated throughout this textbook. For many methodologies, we provide instructions for how to perform calculations both by hand and with Excel. In other cases where realistic models are practical only with the use of a spreadsheet, we focus on the use of Excel to describe the methods to be used.

- **Notes and Comments:** At the end of many sections, we provide Notes and Comments to give the student additional insights about the methods presented in that section. These insights include comments on the limitations of the presented methods, recommendations for applications, and other matters. Additionally, margin notes are used throughout the textbook to provide additional insights and tips related to the specific material being discussed.

- **Analytics in Action:** Each chapter contains an Analytics in Action article. Several of these have been updated and replaced for the fourth edition. These articles present interesting examples of the use of business analytics in practice. The examples are drawn from many different organizations in a variety of areas including healthcare, finance, manufacturing, marketing, and others.

- **DATAfiles and MODELfiles:** All data sets used as examples and in student exercises are also provided online on the companion site as files available for download by the student. DATAfiles are Excel files (or .csv files for easy import into JMP Pro and R/Rattle) that contain data needed for the examples and problems given in the textbook. MODELfiles contain additional modeling features such as extensive use of Excel formulas or the use of Excel Solver, JMP Pro, or R.

- **Problems and Cases:** With the exception of Chapter 1, each chapter contains an extensive selection of problems to help the student master the material presented in that chapter. The problems vary in difficulty and most relate to specific examples of the use of business analytics in practice. Answers to even-numbered problems are provided in an online supplement for student access. With the exception of Chapter 1, each chapter also includes at least one in-depth case study that connects many of the different methods introduced in the chapter. The case studies are designed to be more open-ended than the chapter problems, but enough detail is provided to give the student some direction in solving the cases. New to the fourth edition is the inclusion of two cases that require the use of material from multiple chapters in the text to better illustrate how concepts from different chapters relate to each other.

MindTap

MindTap is a customizable digital course solution that includes an interactive eBook, autograded exercises from the textbook, algorithmic practice problems with solutions feedback, Exploring Analytics visualizations, Adaptive Test Prep, and more! MindTap is also where instructors and users can find the online appendixes for JMP Pro and R/Rattle. All of these materials offer students better access to resources to understand the materials within the course. For more information on MindTap, please contact your Cengage representative.

WebAssign

Prepare for class with confidence using WebAssign from Cengage. This online learning platform fuels practice, so students can truly absorb what you learn – and are better prepared come test time. Videos, Problem Walk-Throughs, and End-of-Chapter problems and cases with instant feedback help them understand the important concepts, while instant grading allows you and them to see where they stand in class. Class Insights allows students to see what topics they have mastered and which they are struggling with, helping them identify where to spend extra time. Study Smarter with WebAssign.

For Students

Online resources are available to help the student work more efficiently. The resources can be accessed through **www.cengage.com/decisionsciences/camm/ba/4e**.

- **R, RStudio, and Rattle:** R, RStudio, and Rattle are open-source software, so they are free to download. *Business Analytics* 4E includes step-by-step instructions for downloading these software.
- **JMP Pro:** Many universities have site licenses of SAS Institute's JMP Pro software on both Mac and Windows. These are typically offered through your university's software licensing administrator. Faculty may contact the JMP Academic team to find out if their universities have a license or to request a complementary instructor copy at www.jmp.com/contact-academic. For institutions without a site license, students may rent a 6- or 12-month license for JMP at www.onthehub.com/jmp.
- **Data Files:** A complete download of all data files associated with this text.

For Instructors

Instructor resources are available to adopters on the Instructor Companion Site, which can be found and accessed at **www.cengage.com/decisionsciences/camm/ba/4e** including:

- **Solutions Manual:** The Solutions Manual, prepared by the authors, includes solutions for all problems in the text. It is available online as well as print. Excel solution files are available to instructors for those problems that require the use of Excel. Solutions for Chapters 5 and 9 are available using both JMP Pro and R/Rattle for data mining problems.
- **Solutions to Case Problems:** These are also prepared by the authors and contain solutions to all case problems presented in the text. Case solutions for Chapters 5 and 9 are provided using both JMP Pro and R/Rattle. Extensive case solutions are also provided for the new multi-chapter cases that draw on material from multiple chapters.
- **PowerPoint Presentation Slides:** The presentation slides contain a teaching outline that incorporates figures to complement instructor lectures.
- **Test Bank:** Cengage Learning Testing Powered by Cognero is a flexible, online system that allows you to:
    - author, edit, and manage test bank content from multiple Cengage Learning solutions,
    - create multiple test versions in an instant, and
    - deliver tests from your Learning Management System (LMS), your classroom, or wherever you want.

Acknowledgments

We would like to acknowledge the work of reviewers and users who have provided comments and suggestions for improvement of this text. Thanks to:

Rafael Becerril Arreola
University of South Carolina

Matthew D. Bailey
Bucknell University

Phillip Beaver
University of Denver

M. Khurrum S. Bhutta
Ohio University

Paolo Catasti
Virginia Commonwealth University

Q B. Chung
Villanova University

Elizabeth A. Denny
University of Kentucky

Mike Taein Eom
University of Portland

Yvette Njan Essounga
Fayetteville State University

Lawrence V. Fulton
Texas State University

Tom Groleau
Carthage College

James F. Hoelscher
Lincoln Memorial University

Eric Huggins
Fort Lewis College

Faizul Huq
Ohio University

Marco Lam
York College of Pennsylvania

Thomas Lee
University of California, Berkeley

Roger Myerson
Northwestern University

Ram Pakath
University of Kentucky

Susan Palocsay
James Madison University

Andy Shogan
University of California, Berkeley

Dothan Truong
Embry-Riddle Aeronautical University

Kai Wang
Wake Technical Community College

Ed Wasil
American University

Ed Winkofsky
University of Cincinnati

A special thanks goes to our associates from business and industry who supplied the Analytics in Action features. We recognize them individually by a credit line in each of the articles. We are also indebted to our senior product manager, Aaron Arnsparger; our Senior Content Manager, Conor Allen; senior learning designer, Brandon Foltz; digital delivery lead, Mark Hopkinson; and our senior project manager at MPS Limited, Santosh Pandey, for their editorial counsel and support during the preparation of this text.

Jeffrey D. Camm
James J. Cochran
Michael J. Fry
Jeffrey W. Ohlmann

Chapter 1

Introduction

CONTENTS

You apply for a loan for the first time. How does the bank assess the riskiness of the loan it might make to you? How does Amazon.com know which books and other products to recommend to you when you log in to their web site? How do airlines determine what price to quote to you when you are shopping for a plane ticket? How can doctors better diagnose and treat you when you are ill or injured?

You may be applying for a loan for the first time, but millions of people around the world have applied for loans before. Many of these loan recipients have paid back their loans in full and on time, but some have not. The bank wants to know whether you are more like those who have paid back their loans or more like those who defaulted. By comparing your credit history, financial situation, and other factors to the vast database of previous loan recipients, the bank can effectively assess how likely you are to default on a loan.

Similarly, Amazon.com has access to data on millions of purchases made by customers on its web site. Amazon.com examines your previous purchases, the products you have viewed, and any product recommendations you have provided. Amazon.com then searches through its huge database for customers who are similar to you in terms of product purchases, recommendations, and interests. Once similar customers have been identified, their purchases form the basis of the recommendations given to you.

Prices for airline tickets are frequently updated. The price quoted to you for a flight between New York and San Francisco today could be very different from the price that will be quoted tomorrow. These changes happen because airlines use a pricing strategy known as revenue management. Revenue management works by examining vast amounts of data on past airline customer purchases and using these data to forecast future purchases. These forecasts are then fed into sophisticated optimization algorithms that determine the optimal price to charge for a particular flight and when to change that price. Revenue management has resulted in substantial increases in airline revenues.

Finally, consider the case of being evaluated by a doctor for a potentially serious medical issue. Hundreds of medical papers may describe research studies done on patients facing similar diagnoses, and thousands of data points exist on their outcomes. However, it is extremely unlikely that your doctor has read every one of these research papers or is aware of all previous patient outcomes. Instead of relying only on her medical training and knowledge gained from her limited set of previous patients, wouldn't it be better for your doctor to have access to the expertise and patient histories of thousands of doctors around the world?

A group of IBM computer scientists initiated a project to develop a new decision technology to help in answering these types of questions. That technology is called Watson, named after the founder of IBM, Thomas J. Watson. The team at IBM focused on one aim: How the vast amounts of data now available on the Internet can be used to make more data-driven, smarter decisions. Watson is an example of the exploding field of **artificial intelligence (AI)**. Broadly speaking, AI is the use of data and computers to make decisions that would have in the past required human intelligence. Often, the computer software mimics the way we understand the human brain functions.

Watson became a household name in 2011, when it famously won the television game show, *Jeopardy!* Since that proof of concept in 2011, IBM has reached agreements with the health insurance provider WellPoint (now part of Anthem), the financial services company Citibank, Memorial Sloan-Kettering Cancer Center, and automobile manufacturer General Motors to apply Watson to the decision problems that they face.

Watson is a system of computing hardware, high-speed data processing, and analytical algorithms that are combined to make data-based recommendations. As more and more data are collected, Watson has the capability to learn over time. In simple terms, according to IBM, Watson gathers hundreds of thousands of possible solutions from a huge data bank, evaluates them using analytical techniques, and proposes only the best solutions for consideration. Watson provides not just a single solution, but rather a range of good solutions with a confidence level for each.

For example, at a data center in Virginia, to the delight of doctors and patients, Watson is already being used to speed up the approval of medical procedures. Citibank is beginning to explore how to use Watson to better serve its customers, and cancer specialists at more than a dozen hospitals in North America are using Watson to assist with the diagnosis and treatment of patients.[1]

This book is concerned with data-driven decision making and the use of analytical approaches in the decision-making process. Three developments spurred recent explosive growth in the use of analytical methods in business applications. First, technological advances—such as improved point-of-sale scanner technology and the collection of data through e-commerce and social networks, data obtained by sensors on all kinds of mechanical devices such as aircraft engines, automobiles, and farm machinery through the so-called Internet of Things and data generated from personal electronic devices—produce incredible amounts of data for businesses. Naturally, businesses want to use these data to improve the efficiency and profitability of their operations, better understand their customers, price their products more effectively, and gain a competitive advantage. Second, ongoing research has resulted in numerous methodological developments, including advances in computational approaches to effectively handle and explore massive amounts of data, faster algorithms for optimization and simulation, and more effective approaches for visualizing data. Third, these methodological developments were paired with an explosion in computing power and storage capability. Better computing hardware, parallel computing, and, more recently, cloud computing (the remote use of hardware and software over the Internet) have enabled businesses to solve big problems more quickly and more accurately than ever before.

In summary, the availability of massive amounts of data, improvements in analytic methodologies, and substantial increases in computing power have all come together to result in a dramatic upsurge in the use of analytical methods in business and a reliance on the discipline that is the focus of this text: business analytics. As stated in the Preface, the purpose of this text is to provide students with a sound conceptual understanding of the role that business analytics plays in the decision-making process. To reinforce the applications orientation of the text and to provide a better understanding of the variety of applications in which analytical methods have been used successfully, Analytics in Action articles are presented throughout the book. Each Analytics in Action article summarizes an application of analytical methods in practice.

1.1 Decision Making

It is the responsibility of managers to plan, coordinate, organize, and lead their organizations to better performance. Ultimately, managers' responsibilities require that they make strategic, tactical, or operational decisions. **Strategic decisions** involve higher-level issues concerned with the overall direction of the organization; these decisions define the organization's overall goals and aspirations for the future. Strategic decisions are usually the domain of higher-level executives and have a time horizon of three to five years. **Tactical decisions** concern how the organization should achieve the goals and objectives set by its strategy, and they are usually the responsibility of midlevel management. Tactical decisions usually span a year and thus are revisited annually or even every six months. **Operational decisions** affect how the firm is run from day to day; they are the domain of operations managers, who are the closest to the customer.

Consider the case of the Thoroughbred Running Company (TRC). Historically, TRC had been a catalog-based retail seller of running shoes and apparel. TRC sales revenues grew quickly as it changed its emphasis from catalog-based sales to Internet-based sales. Recently, TRC decided that it should also establish retail stores in the malls and downtown areas of major cities. This strategic decision will take the firm in a new direction that it hopes will complement its Internet-based strategy. TRC middle managers will therefore have to make a variety of tactical decisions in support of this strategic decision, including

[1] "IBM's Watson Is Learning Its Way to Saving Lives," Fastcompany web site, December 8, 2012; H. Landi, "IBM Watson Health Touts Recent Studies Showing AI Improves How Physicians Treat Cancer," FierceHealthcare web site, June 4, 2019.

how many new stores to open this year, where to open these new stores, how many distribution centers will be needed to support the new stores, and where to locate these distribution centers. Operations managers in the stores will need to make day-to-day decisions regarding, for instance, how many pairs of each model and size of shoes to order from the distribution centers and how to schedule their sales personnel's work time.

Regardless of the level within the firm, *decision making* can be defined as the following process:

1. Identify and define the problem.
2. Determine the criteria that will be used to evaluate alternative solutions.
3. Determine the set of alternative solutions.
4. Evaluate the alternatives.
5. Choose an alternative.

Step 1 of decision making, identifying and defining the problem, is the most critical. Only if the problem is well-defined, with clear metrics of success or failure (step 2), can a proper approach for solving the problem (steps 3 and 4) be devised. Decision making concludes with the choice of one of the alternatives (step 5).

There are a number of approaches to making decisions: tradition ("We've always done it this way"), intuition ("gut feeling"), and rules of thumb ("As the restaurant owner, I schedule twice the number of waiters and cooks on holidays"). The power of each of these approaches should not be underestimated. Managerial experience and intuition are valuable inputs to making decisions, but what if relevant data were available to help us make more informed decisions? With the vast amounts of data now generated and stored electronically, it is estimated that the amount of data stored by businesses more than doubles every two years. How can managers convert these data into knowledge that they can use to be more efficient and effective in managing their businesses?

1.2 Business Analytics Defined

What makes decision making difficult and challenging? Uncertainty is probably the number one challenge. If we knew how much the demand will be for our product, we could do a much better job of planning and scheduling production. If we knew exactly how long each step in a project will take to be completed, we could better predict the project's cost and completion date. If we knew how stocks will perform, investing would be a lot easier.

Another factor that makes decision making difficult is that we often face such an enormous number of alternatives that we cannot evaluate them all. What is the best combination of stocks to help me meet my financial objectives? What is the best product line for a company that wants to maximize its market share? How should an airline price its tickets so as to maximize revenue?

Some firms and industries use the simpler term, ***analytics****. Analytics is often thought of as a broader category than business analytics, encompassing the use of analytical techniques in the sciences and engineering as well. In this text, we use* ***business analytics*** *and* ***analytics*** *synonymously.*

Business analytics is the scientific process of transforming data into insight for making better decisions.[2] Business analytics is used for data-driven or fact-based decision making, which is often seen as more objective than other alternatives for decision making.

As we shall see, the tools of business analytics can aid decision making by creating insights from data, by improving our ability to more accurately forecast for planning, by helping us quantify risk, and by yielding better alternatives through analysis and optimization. A study based on a large sample of firms that was conducted by researchers at MIT's Sloan School of Management and the University of Pennsylvania concluded that firms guided by data-driven decision making have higher productivity and market value and increased output and profitability.[3]

[2] We adopt the definition of analytics developed by the Institute for Operations Research and the Management Sciences (INFORMS).

[3] E. Brynjolfsson, L. M. Hitt, and H. H. Kim, "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?" Thirty-Second International Conference on Information Systems, Shanghai, China, December 2011.

1.3 A Categorization of Analytical Methods and Models

Business analytics can involve anything from simple reports to the most advanced optimization techniques (methods for finding the best course of action). Analytics is generally thought to comprise three broad categories of techniques: descriptive analytics, predictive analytics, and prescriptive analytics.

Descriptive Analytics

Descriptive analytics encompasses the set of techniques that describes what has happened in the past. Examples are data queries, reports, descriptive statistics, data visualization including data dashboards, some data-mining techniques, and basic what-if spreadsheet models.

Appendix B, at the end of this book, describes how to use Microsoft Access to conduct data queries.

A **data query** is a request for information with certain characteristics from a database. For example, a query to a manufacturing plant's database might be for all records of shipments to a particular distribution center during the month of March. This query provides descriptive information about these shipments: the number of shipments, how much was included in each shipment, the date each shipment was sent, and so on. A report summarizing relevant historical information for management might be conveyed by the use of descriptive statistics (means, measures of variation, etc.) and data-visualization tools (tables, charts, and maps). Simple descriptive statistics and data-visualization techniques can be used to find patterns or relationships in a large database.

Data dashboards are collections of tables, charts, maps, and summary statistics that are updated as new data become available. Dashboards are used to help management monitor specific aspects of the company's performance related to their decision-making responsibilities. For corporate-level managers, daily data dashboards might summarize sales by region, current inventory levels, and other company-wide metrics; front-line managers may view dashboards that contain metrics related to staffing levels, local inventory levels, and short-term sales forecasts.

Data mining is the use of analytical techniques for better understanding patterns and relationships that exist in large data sets. For example, by analyzing text on social network platforms like Twitter, data-mining techniques (including cluster analysis and sentiment analysis) are used by companies to better understand their customers. By categorizing certain words as positive or negative and keeping track of how often those words appear in tweets, a company like Apple can better understand how its customers are feeling about a product like the Apple Watch.

Predictive Analytics

Predictive analytics consists of techniques that use models constructed from past data to predict the future or ascertain the impact of one variable on another. For example, past data on product sales may be used to construct a mathematical model to predict future sales. This mode can factor in the product's growth trajectory and seasonality based on past patterns. A packaged-food manufacturer may use point-of-sale scanner data from retail outlets to help in estimating the lift in unit sales due to coupons or sales events. Survey data and past purchase behavior may be used to help predict the market share of a new product. All of these are applications of predictive analytics.

Linear regression, time series analysis, some data-mining techniques, and simulation, often referred to as risk analysis, all fall under the banner of predictive analytics. We discuss all of these techniques in greater detail later in this text.

Data mining, previously discussed as a descriptive analytics tool, is also often used in predictive analytics. For example, a large grocery store chain might be interested in developing a targeted marketing campaign that offers a discount coupon on potato chips. By studying historical point-of-sale data, the store may be able to use data mining to predict which customers are the most likely to respond to an offer on discounted chips by purchasing higher-margin items such as beer or soft drinks in addition to the chips, thus increasing the store's overall revenue.

Simulation involves the use of probability and statistics to construct a computer model to study the impact of uncertainty on a decision. For example, banks often use simulation to model investment and default risk in order to stress-test financial models. Simulation is also often used in the pharmaceutical industry to assess the risk of introducing a new drug.

Prescriptive Analytics

Prescriptive analytics differs from descriptive and predictive analytics in that **prescriptive analytics** indicates a course of action to take; that is, the output of a prescriptive model is a decision. Predictive models provide a forecast or prediction, but do not provide a decision. However, a forecast or prediction, when combined with a rule, becomes a prescriptive model. For example, we may develop a model to predict the probability that a person will default on a loan. If we create a rule that says if the estimated probability of default is more than 0.6, we should not award a loan, now the predictive model, coupled with the rule is prescriptive analytics. These types of prescriptive models that rely on a rule or set of rules are often referred to as **rule-based models**.

Other examples of prescriptive analytics are portfolio models in finance, supply network design models in operations, and price-markdown models in retailing. Portfolio models use historical investment return data to determine which mix of investments will yield the highest expected return while controlling or limiting exposure to risk. Supply-network design models provide plant and distribution center locations that will minimize costs while still meeting customer service requirements. Given historical data, retail price markdown models yield revenue-maximizing discount levels and the timing of discount offers when goods have not sold as planned. All of these models are known as **optimization models**, that is, models that give the best decision subject to the constraints of the situation.

Another type of modeling in the prescriptive analytics category is **simulation optimization** which combines the use of probability and statistics to model uncertainty with optimization techniques to find good decisions in highly complex and highly uncertain settings. Finally, the techniques of **decision analysis** can be used to develop an optimal strategy when a decision maker is faced with several decision alternatives and an uncertain set of future events. Decision analysis also employs **utility theory**, which assigns values to outcomes based on the decision maker's attitude toward risk, loss, and other factors.

In this text we cover all three areas of business analytics: descriptive, predictive, and prescriptive. Table 1.1 shows how the chapters cover the three categories.

1.4 Big Data

On any given day, 500 million tweets and 294 billion e-mails are sent, 95 million photos and videos are shared on Instagram, 350 million photos are posted on Facebook, and 3.5 billion searches are made with Google.[4] It is through technology that we have truly been thrust into the data age. Because data can now be collected electronically, the available amounts of it are staggering. The Internet, cell phones, retail checkout scanners, surveillance video, and sensors on everything from aircraft to cars to bridges allow us to collect and store vast amounts of data in real time.

In the midst of all of this data collection, the term *big data* has been created. There is no universally accepted definition of big data. However, probably the most accepted and most general definition is that **big data** is any set of data that is too large or too complex to be handled by standard data-processing techniques and typical desktop software. IBM describes the phenomenon of big data through the four Vs: volume, velocity, variety, and veracity, as shown in Figure 1.1.[5]

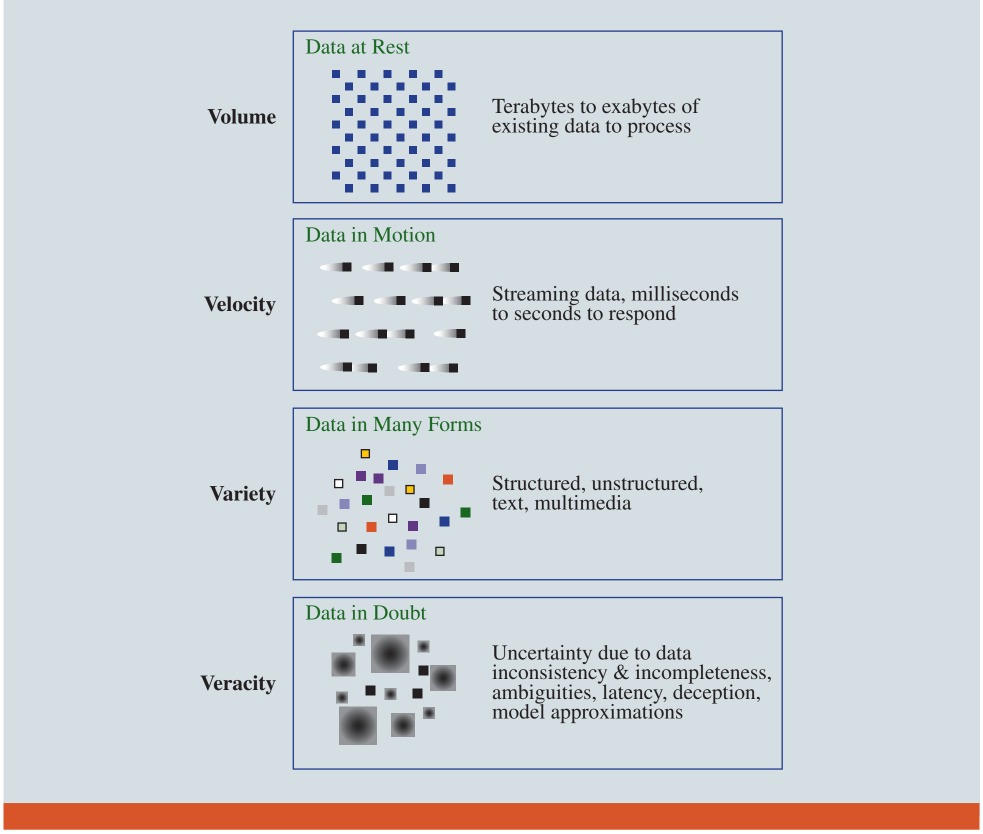
[4] J. Desjardins, "How Much Data Is Generated Each Day?" Visual Capitalist web site, April 15, 2019.

[5] IBM web site: www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg.

TABLE 1.1 Coverage of Business Analytics Topics in This Text

Chapter	Title	Descriptive	Predictive	Prescriptive
1	Introduction	●	●	●
2	Descriptive Statistics	●		
3	Data Visualization	●		
4	Probability: An Introduction to Modeling Uncertainty	●		
5	Descriptive Data Mining	●		
6	Statistical Inference	●		
7	Linear Regression		●	
8	Time Series and Forecasting		●	
9	Predictive Data Mining		●	
10	Spreadsheet Models	●	●	●
11	Monte Carlo Simulation		●	●
12	Linear Optimization Models			●
13	Integer Linear Optimization Models			●
14	Nonlinear Optimization Models			●
15	Decision Analysis			●

FIGURE 1.1 The Four Vs of Big Data


Volume
Data at Rest
Terabytes to exabytes of existing data to process
Velocity
Data in Motion
Streaming data, milliseconds to seconds to respond
Variety
Data in Many Forms
Structured, unstructured, text, multimedia
Veracity
Data in Doubt
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Source: IBM.

Volume

Because data are collected electronically, we are able to collect more of it. To be useful, these data must be stored, and this storage has led to vast quantities of data. Many companies now store in excess of 100 terabytes of data (a terabyte is 1,024 gigabytes).

Velocity

Real-time capture and analysis of data present unique challenges both in how data are stored, and the speed with which those data can be analyzed for decision making. For example, the New York Stock Exchange collects 1 terabyte of data in a single trading session, and having current data and real-time rules for trades and predictive modeling are important for managing stock portfolios.

Variety

In addition to the sheer volume and speed with which companies now collect data, more complicated types of data are now available and are proving to be of great value to businesses. Text data are collected by monitoring what is being said about a company's products or services on social media platforms such as Twitter. Audio data are collected from service calls (on a service call, you will often hear "this call may be monitored for quality control"). Video data collected by in-store video cameras are used to analyze shopping behavior. Analyzing information generated by these nontraditional sources is more complicated in part because of the processing required to transform the data into a numerical form that can be analyzed.

Veracity

Veracity has to do with how much uncertainty is in the data. For example, the data could have many missing values, which makes reliable analysis a challenge. Inconsistencies in units of measure and the lack of reliability of responses in terms of bias also increase the complexity of the data.

Businesses have realized that understanding big data can lead to a competitive advantage. Although big data represents opportunities, it also presents challenges in terms of data storage and processing, security, and available analytical talent.

The four Vs indicate that big data creates challenges in terms of how these complex data can be captured, stored, and processed; secured; and then analyzed. Traditional databases more or less assume that data fit into nice rows and columns, but that is not always the case with big data. Also, the sheer volume (the first V) often means that it is not possible to store all of the data on a single computer. This has led to new technologies like **Hadoop**—an open-source programming environment that supports big data processing through distributed storage and distributed processing on clusters of computers. Essentially, Hadoop provides a divide-and-conquer approach to handling massive amounts of data, dividing the storage and processing over multiple computers. **MapReduce** is a programming model used within Hadoop that performs the two major steps for which it is named: the map step and the reduce step. The map step divides the data into manageable subsets and distributes it to the computers in the cluster (often termed nodes) for storing and processing. The reduce step collects answers from the nodes and combines them into an answer to the original problem. Technologies like Hadoop and MapReduce, paired with relatively inexpensive computer power, enable cost-effective processing of big data; otherwise, in some cases, processing might not even be possible.

While some sources of big data are publicly available (Twitter, weather data, etc.), much of it is private information. Medical records, bank account information, and credit card transactions, for example, are all highly confidential and must be protected from computer hackers. **Data security**, the protection of stored data from destructive forces or unauthorized users, is of critical importance to companies. For example, credit card transactions are potentially very useful for understanding consumer behavior, but compromise of these data could lead to unauthorized use of the credit card or identity theft. A 2016 study of 383 companies in 12 countries conducted by the Ponemon Institute and IBM found that the average cost of a data breach is $3.86 million.[6] Companies such as Target, Anthem, JPMorgan Chase, Yahoo!, Facebook, Marriott, Equifax, and Home Depot have faced major data breaches costing millions of dollars.

The complexities of the 4 Vs have increased the demand for analysts, but a shortage of qualified analysts has made hiring more challenging. More companies are searching for **data scientists**, who know how to effectively process and analyze massive amounts of data because they are well trained in both computer science and statistics. Next we discuss three examples of how companies are collecting big data for competitive advantage.

Kroger Understands Its Customers[7] Kroger is the largest retail grocery chain in the United States. It sends over 11 million pieces of direct mail to its customers each quarter. The quarterly mailers each contain 12 coupons that are tailored to each household based on several years of shopping data obtained through its customer loyalty card program. By collecting and analyzing consumer behavior at the individual household level, and better matching its coupon offers to shopper interests, Kroger has been able to realize a far higher redemption rate on its coupons. In the six-week period following distribution of the mailers, over 70% of households redeem at least one coupon, leading to an estimated coupon revenue of $10 billion for Kroger.

MagicBand at Disney[8] The Walt Disney Company offers a wristband to visitors to its Orlando, Florida, Disney World theme park. Known as the MagicBand, the wristband contains technology that can transmit more than 40 feet and can be used to track each visitor's location in the park in real time. The band can link to information that allows Disney to better serve its visitors. For example, prior to the trip to Disney World, a visitor might be asked to fill out a survey on his or her birth date and favorite rides, characters, and restaurant table type and location. This information, linked to the MagicBand, can allow Disney employees using smartphones to greet you by name as you arrive, offer you products they know you prefer, wish you a happy birthday, have your favorite characters show up as you wait in line or have lunch at your favorite table. The MagicBand can be linked to your credit card, so there is no need to carry cash or a credit card. And during your visit, your movement throughout the park can be tracked and the data can be analyzed to better serve you during your next visit to the park.

General Electric and the Internet of Things[9] The **Internet of Things (IoT)** is the technology that allows data, collected from sensors in all types of machines, to be sent over the Internet to repositories where it can be stored and analyzed. This ability to collect data from products has enabled the companies that produce and sell those products to better serve their customers and offer new services based on analytics. For example, each day General Electric (GE) gathers nearly 50 million pieces of data from 10 million sensors on medical equipment and aircraft engines it has sold to customers throughout the world. In the case of aircraft engines, through a service agreement with its customers, GE collects data each time an airplane powered by its engines takes off and lands. By analyzing these data, GE can better predict when maintenance is needed, which helps customers avoid unplanned maintenance and downtime and helps ensure safe operation. GE can also use the data to better control how the plane is flown, leading to a decrease in fuel cost by flying more efficiently. GE spun off a new company called GE Digital 2.0 which operates as a stand-alone company focused on software that leverages IoT data. In 2018, GE announced that it would spin off a new company from its existing GE Digital business that will focus on industrial IoT applications.

Although big data is clearly one of the drivers for the strong demand for analytics, it is important to understand that, in some sense, big data issues are a subset of analytics. Many very valuable applications of analytics do not involve big data, but rather traditional data sets that are very manageable by traditional database and analytics software. The key to

[6] S. Shepard, "The Average Cost of a Data Breach," Security Today web site, July 17, 2018.

[7] Based on "Kroger Knows Your Shopping Patterns Better than You Do," Forbes.com, October 23, 2013.

[8] Based on "Disney's $1 Billion Bet on a Magical Wristband," Wired.com, March 10, 2015.

[9] Based on "G.E. Opens Its Big Data Platform," NYTimes.com, October 9, 2014; "GE Announces New Industrial IoT Software Business," Forbes web site, December 14, 2018.

analytics is that it provides useful insights and better decision making using the data that are available—whether those data are "big" or "small."

1.5 Business Analytics in Practice

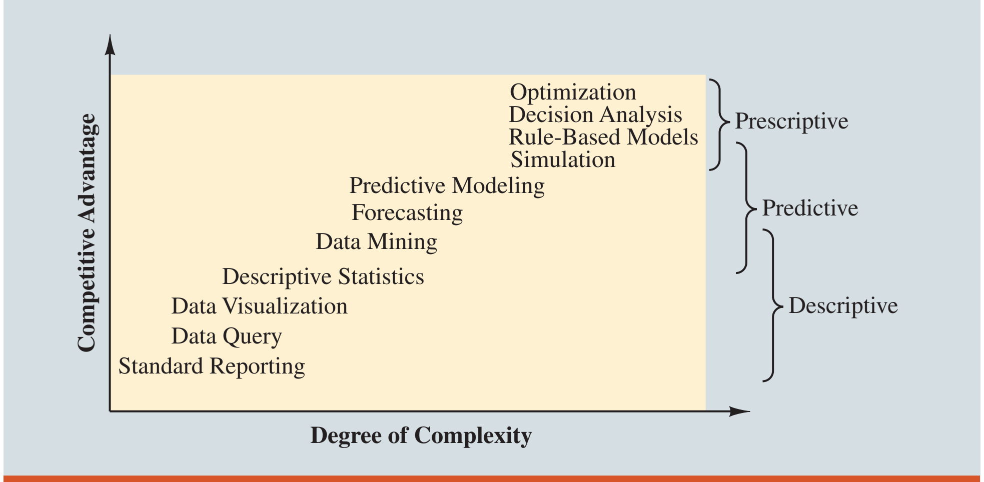
Business analytics involves tools as simple as reports and graphs to those that are as sophisticated as optimization, data mining, and simulation. In practice, companies that apply analytics often follow a trajectory similar to that shown in Figure 1.2. Organizations start with basic analytics in the lower left. As they realize the advantages of these analytic techniques, they often progress to more sophisticated techniques in an effort to reap the derived competitive advantage. Therefore, predictive and prescriptive analytics are sometimes referred to as **advanced analytics**. Not all companies reach that level of usage, but those that embrace analytics as a competitive strategy often do.

Analytics has been applied in virtually all sectors of business and government. Organizations such as Procter & Gamble, IBM, UPS, Netflix, Amazon.com, Google, the Internal Revenue Service, and General Electric have embraced analytics to solve important problems or to achieve a competitive advantage. In this section, we briefly discuss some of the types of applications of analytics by application area.

Financial Analytics

Applications of analytics in finance are numerous and pervasive. Predictive models are used to forecast financial performance, to assess the risk of investment portfolios and projects, and to construct financial instruments such as derivatives. Prescriptive models are used to construct optimal portfolios of investments, to allocate assets, and to create optimal capital budgeting plans. For example, Europcar, the leading rental car company in Europe, uses forecasting models, simulation and optimization to predict demand, assess risk, and optimize the use of its fleet. It's models are implemented via a decision support system used in nine countries in Europe and has led to higher utilization of its fleet, decreased costs, and increased profitability.[10] Simulation is also often used to assess risk in the financial sector; one example is the deployment by Hypo Real Estate International of simulation models to successfully manage commercial real estate risk.[11]

FIGURE 1.2 The Spectrum of Business Analytics



Source: Adapted from SAS.

[10] J. Guillen et al., "Europcar Integrates Forecasting, Simulation, and Optimization Techniques in a Capacity and Revenue Management System," *INFORMS Journal on Applied Analytics*, 49, no. 1 (January–February 2019).

[11] Y. Jafry, C. Marrison, and U. Umkehrer-Neudeck, "Hypo International Strengthens Risk Management with a Large-Scale, Secure Spreadsheet-Management Framework," *Interfaces* 38, no. 4 (July–August 2008).

Human Resource (HR) Analytics

A relatively new area of application for analytics is the management of an organization's human resources. The HR function is charged with ensuring that the organization (1) has the mix of skill sets necessary to meet its needs, (2) is hiring the highest-quality talent and providing an environment that retains it, and (3) achieves its organizational diversity goals. Google refers to its HR Analytics function as "people analytics." Google has analyzed substantial data on their own employees to determine the characteristics of great leaders, to assess factors that contribute to productivity, and to evaluate potential new hires. Google also uses predictive analytics to continually update their forecast of future employee turnover and retention.[12]

Marketing Analytics

Marketing is one of the fastest-growing areas for the application of analytics. A better understanding of consumer behavior through the use of scanner data and data generated from social media has led to an increased interest in marketing analytics. As a result, descriptive, predictive, and prescriptive analytics are all heavily used in marketing. A better understanding of consumer behavior through analytics leads to the better use of advertising budgets, more effective pricing strategies, improved forecasting of demand, improved product-line management, and increased customer satisfaction and loyalty. For example, Turner Broadcasting System Inc. uses forecasting and optimization models to create more-targeted audiences and to better schedule commercials for its advertising partners. The use of these models has led to an increase in Turner year-over-year advertising revenue of 186% and, at the same time, dramatically increased sales for the advertisers. Those advertisers that chose to benchmark found an increase in sales of $118 million.[13]

In another example of high-impact marketing analytics, automobile manufacturer Chrysler teamed with J.D. Power and Associates to develop an innovative set of predictive models to support its pricing decisions for automobiles. These models help Chrysler to better understand the ramifications of proposed pricing structures (a combination of manufacturer's suggested retail price, interest rate offers, and rebates) and, as a result, to improve its pricing decisions. The models have generated an estimated annual savings of $500 million.[14]

Health Care Analytics

The use of analytics in health care is on the increase because of pressure to simultaneously control costs and provide more effective treatment. Descriptive, predictive, and prescriptive analytics are used to improve patient, staff, and facility scheduling; patient flow; purchasing; and inventory control. A study by McKinsey Global Institute (MGI) and McKinsey & Company[15] estimates that the health care system in the United States could save more than $300 billion per year by better utilizing analytics; these savings are approximately the equivalent of the entire gross domestic product of countries such as Finland, Singapore, and Ireland.

The use of prescriptive analytics for diagnosis and treatment is relatively new, but it may prove to be the most important application of analytics in health care. For example, a group of scientists in Georgia used predictive models and optimization to develop personalized treatment for diabetes. They developed a predictive model that uses fluid dynamics and patient monitoring data to establish the relationship between drug dosage and drug effect at the individual level. This alleviates the need for more invasive procedures to monitor drug concentration. Then they used an optimization model that takes output from the predictive model to determine how an

[12] J. Sullivan, "How Google Is Using People Analytics to Completely Reinvent HR," Talent Management and HR web site, February 26, 2013.

[13] J. A. Carbajal, P. Williams, A. Popescu, and W. Chaar, "Turner Blazes a Trail for Audience Targeting on Television with Operations Research and Advanced Analytics," *INFORMS Journal on Applied Analytics*, 49, no. 1 (January–February 2019).

[14] J. Silva-Risso et al., "Chrysler and J. D. Power: Pioneering Scientific Price Customization in the Automobile Industry," *Interfaces* 38, no. 1 (January–February 2008).

[15] J. Manyika et al., "Big Data: The Next Frontier for Innovation, Competition and Productivity," McKinsey Global Institute Report, 2011.

individual achieves better glycemic control using less dosage. Using the models results in about a 39% savings in hospital costs, which equates to about $40,880 per patient.[16]

Supply Chain Analytics

The core service of companies such as UPS and FedEx is the efficient delivery of goods, and analytics has long been used to achieve efficiency. The optimal sorting of goods, vehicle and staff scheduling, and vehicle routing are all key to profitability for logistics companies such as UPS and FedEx.

Companies can benefit from better inventory and processing control and more efficient supply chains. Analytic tools used in this area span the entire spectrum of analytics. For example, the women's apparel manufacturer Bernard Claus, Inc. has successfully used descriptive analytics to provide its managers a visual representation of the status of its supply chain.[17] ConAgra Foods uses predictive and prescriptive analytics to better plan capacity utilization by incorporating the inherent uncertainty in commodities pricing. ConAgra realized a 100% return on its investment in analytics in under three months—an unheard of result for a major technology investment.[18]

Analytics for Government and Nonprofits

Government agencies and other nonprofits have used analytics to drive out inefficiencies and increase the effectiveness and accountability of programs. Indeed, much of advanced analytics has its roots in the U.S. and English military dating back to World War II. Today, the use of analytics in government is becoming pervasive in everything from elections to tax collection. For example, the New York State Department of Taxation and Finance has worked with IBM to use prescriptive analytics in the development of a more effective approach to tax collection. The result was an increase in collections from delinquent payers of $83 million over two years.[19] The U.S. Internal Revenue Service has used data mining to identify patterns that distinguish questionable annual personal income tax filings. In one application, the IRS combines its data on individual taxpayers with data received from banks, on mortgage payments made by those taxpayers. When taxpayers report a mortgage payment that is unrealistically high relative to their reported taxable income, they are flagged as possible underreporters of taxable income. The filing is then further scrutinized and may trigger an audit.

Likewise, nonprofit agencies have used analytics to ensure their effectiveness and accountability to their donors and clients. Catholic Relief Services (CRS) is the official international humanitarian agency of the U.S. Catholic community. The CRS mission is to provide relief for the victims of both natural and human-made disasters and to help people in need around the world through its health, educational, and agricultural programs. CRS uses an analytical spreadsheet model to assist in the allocation of its annual budget based on the impact that its various relief efforts and programs will have in different countries.[20]

Sports Analytics

The use of analytics in sports has gained considerable notoriety since 2003 when renowned author Michael Lewis published *Moneyball*. Lewis' book tells the story of how the Oakland Athletics used an analytical approach to player evaluation in order to assemble a competitive team with a limited budget. The use of analytics for player evaluation and on-field strategy is now common, especially in professional sports. Professional sports teams use analytics to assess players for the amateur drafts and to decide how much to offer players in contract negotiations;[21]

[16] E. Lee et al., "Outcome-Driven Personalized Treatment Design for Managing Diabetes," *Interfaces,* 48, no. 5 (September–October 2018).

[17] T. H. Davenport, ed., *Enterprise Analytics* (Upper Saddle River, NJ: Pearson Education Inc., 2013).

[18] "ConAgra Mills: Up-to-the-Minute Insights Drive Smarter Selling Decisions and Big Improvements in Capacity Utilization," IBM Smarter Planet Leadership Series. Available at: www.ibm.com/smarterplanet/us/en/leadership/conagra/, retrieved December 1, 2012.

[19] G. Miller et al., "Tax Collection Optimization for New York State," *Interfaces* 42, no. 1 (January–February 2013).

[20] I. Gamvros, R. Nidel, and S. Raghavan, "Investment Analysis and Budget Allocation at Catholic Relief Services," *Interfaces* 36, no. 5 (September–October 2006).

[21] N. Streib, S. J. Young, and J. Sokol, "A Major League Baseball Team Uses Operations Research to Improve Draft Preparation," *Interfaces* 42, no. 2 (March–April 2012).

professional motorcycle racing teams use sophisticated optimization for gearbox design to gain competitive advantage;[22] and teams use analytics to assist with on-field decisions such as which pitchers to use in various games of a Major League Baseball playoff series.

The use of analytics for off-the-field business decisions is also increasing rapidly. Ensuring customer satisfaction is important for any company, and fans are the customers of sports teams. The Cleveland Indians professional baseball team used a type of predictive modeling known as conjoint analysis to design its premium seating offerings at Progressive Field based on fan survey data. Using prescriptive analytics, franchises across several major sports dynamically adjust ticket prices throughout the season to reflect the relative attractiveness and potential demand for each game.

Web Analytics

Web analytics is the analysis of online activity, which includes, but is not limited to, visits to web sites and social media sites such as Facebook and LinkedIn. Web analytics obviously has huge implications for promoting and selling products and services via the Internet. Leading companies apply descriptive and advanced analytics to data collected in online experiments to determine the best way to configure web sites, position ads, and utilize social networks for the promotion of products and services. Online experimentation involves exposing various subgroups to different versions of a web site and tracking the results. Because of the massive pool of Internet users, experiments can be conducted without risking the disruption of the overall business of the company. Such experiments are proving to be invaluable because they enable the company to use trial-and-error in determining statistically what makes a difference in their web site traffic and sales.

1.6 Legal and Ethical Issues in the Use of Data and Analytics

With the advent of big data and the dramatic increase in the use of analytics and data science to improve decision making, increased attention has been paid to ethical concerns around data privacy and the ethical use of models based on data.

As businesses routinely collect data about their customers, they have an obligation to protect the data and to not misuse that data. Clients and customers have an obligation to understand the trade-offs between allowing their data to be collected and used, and the benefits they accrue from allowing a company to collect and use that data. For example, many companies have loyalty cards that collect data on customer purchases. In return for the benefits of using a loyalty card, typically discounted prices, customers must agree to allow the company to collect and use the data on purchases. An agreement must be signed between the customer and the company, and the agreement must specify what data will be collected and how it will be used. For example, the agreement might say that all scanned purchases will be collected with the date, time, location, and card number, but that the company agrees to only use that data internally to the company and to not give or sell that data to outside firms or individuals. The company then has an ethical obligation to uphold that agreement and make every effort to ensure that the data are protected from any type of unauthorized access. Unauthorized access of data is known as a data breach. Data breaches are a major concern for all companies in the digital age. A study by IBM and the Ponemon Institute estimated that the average cost of a data breach is $3.86 million.

Data privacy laws are designed to protect individuals' data from being used against their wishes. One of the strictest data privacy laws is the General Data Protection Regulation (GDPR) which went into effect in the European Union in May 2018. The law stipulates that the request for consent to use an individual's data must be easily understood and accessible, the intended uses of the data must be specified, and it must be easy to withdraw consent. The law also stipulates that an individual has a right to a copy of their data and the right "to be forgotten," that is, the right to demand that their data be erased. It is the

[22] J. Amoros, L. F. Escudero, J. F. Monge, J. V. Segura, and O. Reinoso, "TEAM ASPAR Uses Binary Optimization to Obtain Optimal Gearbox Ratios in Motorcycle Racing," *Interfaces* 42, no. 2 (March–April 2012).

responsibility of analytics professionals, indeed, anyone who handles or stores data, to understand the laws associated with the collection, storage, and use of individuals' data.

Ethical issues that arise in the use of data and analytics are just as important as the legal issues. Analytics professionals have a responsibility to behave ethically, which includes protecting data, being transparent about the data and how it was collected, and what it does and does not contain. Analysts must be transparent about the methods used to analyze the data and any assumptions that have to be made for the methods used. Finally, analysts must provide valid conclusions and understandable recommendations to their clients.

Intentionally using data and analytics for unethical purposes is clearly unethical. For example, using analytics to identify whom to target for fraud is of course inherently unethical because the goal itself is an unethical objective. Intentionally using biased data to achieve a goal is likewise inherently unethical. Misleading a client by misrepresenting results is clearly unethical.

For example, consider the case of an airline that runs an advertisement that "84% of business fliers to Chicago prefer that airline over its competitors." Such a statement is valid if the airline randomly surveyed business fliers across all airlines with a destination of Chicago. But, if for convenience, the airline surveyed only its own customers, the survey would be biased, and the claim would be misleading because fliers on other airlines were not surveyed. Indeed, if anything, the only conclusion one can legitimately draw from the biased sample of its own customers would be that 84% of that airlines' own customers preferred that airline and 16% of its own customers actually preferred another airline![23]

In her book, *Weapons of Math Destruction*, author Cathy O'Neil discusses how algorithms and models can be unintentionally biased.[24] For example, consider an analyst who is building a credit risk model for awarding loans. The location of the home of the applicant might be a variable that is correlated with other variables like income and ethnicity. Income is perhaps a relevant variable for determining the amount of a loan, but ethnicity is not. A model using home location could therefore lead to unintentional bias in the credit risk model. It is the analysts' responsibility to make sure this type of model bias and data bias do not become a part of the model.

Researcher and opinion writer Zeynep Tufecki[25] examines so-called "unintended consequences" of analytics, and particularly of machine learning and recommendation engines. Tufecki has pointed out that many Internet sites that use recommendation engines often suggest more extreme content, in terms of political views and conspiracy theories, to users based on their past viewing history. Tufecki and others theorize that this is because the machine learning algorithms being used have identified that more extreme content increases users' viewing time on the site, which is often the objective function being maximized by the machine learning algorithm. Therefore, while it is not the intention of the algorithm to promote more extreme views and disseminate false information, this may be the unintended consequence of using a machine learning algorithm that maximizes users' viewing time on the site. Analysts and decision makers must be aware of potential unintended consequences of their models, and they must decide how to react to these consequences once they are discovered.

Several organizations, including the American Statistical Association (ASA) and the Institute for Operations Research and the Management Sciences (INFORMS), provide ethical guidelines for analysts. In their "Ethical Guidelines for Statistical Practice,"[26] the ASA uses the term *statistician* throughout, but states that this "includes all practitioners of statistics and quantitative sciences—regardless of job title or field of degree—comprising statisticians at all levels of the profession and members of other professions who utilize and report statistical analyses and their applications." Their guidelines

[23] A. Barnett, "Misapplications Reviews: Newswatch," *Interfaces* 14, no. 6 (November–December 1984).

[24] C. O'Neil, *Weapons of Math Destruction, How Big Data Increases Inequality and Threatens Democracy* (New York: Crown Publishing, 2016).

[25] Z. Tufecki. "YouTube, the Great Radicalizer," *The New York Times,* March 10, 2018.

[26] Ethical Guidelines for Statistical Practice, the American Statistical Association, April 14, 2018.

state that "Good statistical practice is fundamentally based on transparent assumptions, reproducible results, and valid interpretations." More details are given in eight different sections of the guidelines and we encourage you to read and familiarize yourself with these guidelines.

INFORMS is a professional society focused on operations research and the management sciences, including analytics. INFORMS offers an analytics certification called CAP—certified analytics professional. All candidates for CAP are required to comply with the code of ethics/conduct provided by INFORMS.[27] The INFORMS CAP guidelines state, "In general, analytics professionals are obliged to conduct their professional activities responsibly, with particular attention to the values of consistency, respect for individuals, autonomy of all, integrity, justice, utility and competence." INFORMS also offers a set of Ethics Guidelines for its members, which covers ethical behavior for analytics professionals in three domains: Society, Organizations (businesses, government, nonprofit organization, and universities), and the Profession (operations research and analytics).[28] As these guidelines are fairly easy to understand and at the same time fairly comprehensive, we list them here in Table 1.2 and encourage you as a user/provider of analytics to make them your guiding principles.

TABLE 1.2 INFORMS Ethics Guidelines

Relative to Society

Analytics professionals should aspire to be:

- **Accountable** for their professional actions and the impact of their work.
- **Forthcoming** about their assumptions, interests, sponsors, motivations, limitations, and potential conflicts of interest.
- **Honest** in reporting their results, even when they fail to yield the desired outcome.
- **Objective** in their assessments of facts, irrespective of their opinions or beliefs.
- **Respectful** of the viewpoints and the values of others.
- **Responsible** for undertaking research and projects that provide positive benefits by advancing our scientific understanding, contributing to organizational improvements, and supporting social good.

Relative to Organizations

Analytics professionals should aspire to be:

- **Accurate** in our assertions, reports, and presentations.
- **Alert** to possible unintended or negative consequences that our results and recommendations may have on others.
- **Informed** of advances and developments in the fields relevant to our work.
- **Questioning** of whether there are more effective and efficient ways to reach a goal.
- **Realistic** in our claims of achievable results, and in acknowledging when the best course of action may be to terminate a project.
- **Rigorous** by adhering to proper professional practices in the development and reporting of our work.

Relative to the Profession

Analytics professionals should aspire to be:

- **Cooperative** by sharing best practices, information, and ideas with colleagues, young professionals, and students.
- **Impartial** in our praise or criticism of others and their accomplishments, setting aside personal interests.
- **Inclusive** of all colleagues, and rejecting discrimination and harassment in any form.
- **Tolerant** of well-conducted research and well-reasoned results, which may differ from our own findings or opinions.
- **Truthful** in providing attribution when our work draws from the ideas of others.
- **Vigilant** by speaking out against actions that are damaging to the profession

[27] Certified Analytics Professional Code of Ethics/Conduct. Available at www.certifiedanalytics.org/ethics.php.

[28] INFORMS Ethics Guidelines. Available at www.informs.org/About-INFORMS/Governance/INFORMS-Ethics-Guidelines.

SUMMARY

This introductory chapter began with a discussion of decision making. Decision making can be defined as the following process: (1) identify and define the problem, (2) determine the criteria that will be used to evaluate alternative solutions, (3) determine the set of alternative solutions, (4) evaluate the alternatives, and (5) choose an alternative. Decisions may be strategic (high level, concerned with the overall direction of the business), tactical (midlevel, concerned with how to achieve the strategic goals of the business), or operational (day-to-day decisions that must be made to run the company).

Uncertainty and an overwhelming number of alternatives are two key factors that make decision making difficult. Business analytics approaches can assist by identifying and mitigating uncertainty and by prescribing the best course of action from a very large number of alternatives. In short, business analytics can help us make better-informed decisions.

There are three categories of analytics: descriptive, predictive, and prescriptive. Descriptive analytics describes what has happened and includes tools such as reports, data visualization, data dashboards, descriptive statistics, and some data-mining techniques. Predictive analytics consists of techniques that use past data to predict future events or ascertain the impact of one variable on another. These techniques include regression, data mining, forecasting, and simulation. Prescriptive analytics uses data to determine a course of action. This class of analytical techniques includes rule-based models, simulation, decision analysis, and optimization. Descriptive and predictive analytics can help us better understand the uncertainty and risk associated with our decision alternatives. Predictive and prescriptive analytics, also often referred to as advanced analytics, can help us make the best decision when facing a myriad of alternatives.

Big data is a set of data that is too large or too complex to be handled by standard data-processing techniques or typical desktop software. The increasing prevalence of big data is leading to an increase in the use of analytics. The Internet, retail scanners, and cell phones are making huge amounts of data available to companies, and these companies want to better understand these data. Business analytics helps them understand these data and use them to make better decisions.

We also discussed various application areas of analytics. Our discussion focused on financial analytics, human resource analytics, marketing analytics, health care analytics, supply chain analytics, analytics for government and nonprofit organizations, sports analytics, and web analytics. However, the use of analytics is rapidly spreading to other sectors, industries, and functional areas of organizations. We concluded this chapter with a discussion of legal and ethical issues in the use of data and analytics, a topic that should be of great importance to all practitioners and consumers of analytics. Each remaining chapter in this text will provide a real-world vignette in which business analytics is applied to a problem faced by a real organization.

GLOSSARY

Artificial Intelligence (AI) The use of data and computers to make decisions that would have in the past required human intelligence.

Advanced analytics Predictive and prescriptive analytics.

Big data Any set of data that is too large or too complex to be handled by standard data-processing techniques and typical desktop software.

Business analytics The scientific process of transforming data into insight for making better decisions.

Data dashboard A collection of tables, charts, and maps to help management monitor selected aspects of the company's performance.

Data mining The use of analytical techniques for better understanding patterns and relationships that exist in large data sets.

Data query A request for information with certain characteristics from a database.

Data scientists Analysts trained in both computer science and statistics who know how to effectively process and analyze massive amounts of data.

Data security Protecting stored data from destructive forces or unauthorized users.

Decision analysis A technique used to develop an optimal strategy when a decision maker is faced with several decision alternatives and an uncertain set of future events.

Descriptive analytics Analytical tools that describe what has happened.

Hadoop An open-source programming environment that supports big data processing through distributed storage and distributed processing on clusters of computers.

Internet of Things (IoT) The technology that allows data collected from sensors in all types of machines to be sent over the Internet to repositories where it can be stored and analyzed.

MapReduce Programming model used within Hadoop that performs the two major steps for which it is named: the map step and the reduce step. The map step divides the data into manageable subsets and distributes it to the computers in the cluster for storing and processing. The reduce step collects answers from the nodes and combines them into an answer to the original problem.

Operational decisions A decision concerned with how the organization is run from day to day.

Optimization models A mathematical model that gives the best decision, subject to the situation's constraints.

Predictive analytics Techniques that use models constructed from past data to predict the future or to ascertain the impact of one variable on another.

Prescriptive analytics Techniques that analyze input data and yield a best course of action.

Rule-based model A prescriptive model that is based on a rule or set of rules.

Simulation The use of probability and statistics to construct a computer model to study the impact of uncertainty on the decision at hand.

Simulation optimization The use of probability and statistics to model uncertainty, combined with optimization techniques, to find good decisions in highly complex and highly uncertain settings.

Strategic decision A decision that involves higher-level issues and that is concerned with the overall direction of the organization, defining the overall goals and aspirations for the organization's future.

Tactical decision A decision concerned with how the organization should achieve the goals and objectives set by its strategy.

Utility theory The study of the total worth or relative desirability of a particular outcome that reflects the decision maker's attitude toward a collection of factors such as profit, loss, and risk.

Chapter 2

Descriptive Statistics

CONTENTS

ANALYTICS IN ACTION

U.S. Census Bureau

The U.S. Census Bureau is part of the Department of Commerce. The U.S. Census Bureau collects data related to the population and economy of the United States using a variety of methods and for many purposes. These data are essential to many government and business decisions.

Probably the best-known data collected by the U.S. Census Bureau is the decennial census, which is an effort to count the total U.S. population. Collecting these data is a huge undertaking involving mailings, door-to-door visits, and other methods. The decennial census collects categorical data such as the sex and race of the respondents, as well as quantitative data such as the number of people living in the household. The data collected in the decennial census are used to determine the number of representatives assigned to each state, the number of Electoral College votes apportioned to each state, and how federal government funding is divided among communities.

The U.S. Census Bureau also administers the Current Population Survey (CPS). The CPS is a cross-sectional monthly survey of a sample of 60,000 households used to estimate employment and unemployment rates in different geographic areas. The CPS has been administered since 1940, so an extensive time series of employment and unemployment data now exists. These data drive government policies such as job assistance programs. The estimated unemployment rates are watched closely as an overall indicator of the health of the U.S. economy.

The data collected by the U.S. Census Bureau are also very useful to businesses. Retailers use data on population changes in different areas to plan new store openings. Mail-order catalog companies use the demographic data when designing targeted marketing campaigns. In many cases, businesses combine the data collected by the U.S. Census Bureau with their own data on customer behavior to plan strategies and to identify potential customers. The data collected by the U.S. Census Bureau is publicly available and can be downloaded from its web site.

In this chapter, we first explain the need to collect and analyze data and identify some common sources of data. Then we discuss the types of data that you may encounter in practice and present several numerical measures for summarizing data. We cover some common ways of manipulating and summarizing data using spreadsheets. We then develop numerical summary measures for data sets consisting of a single variable. When a data set contains more than one variable, the same numerical measures can be computed separately for each variable. In the two-variable case, we also develop measures of the relationship between the variables.

2.1 Overview of Using Data: Definitions and Goals

Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation. Table 2.1 shows a data set containing information for stocks in the Dow Jones Industrial Index (or simply "the Dow") on June 25, 2019. The Dow is tracked by many financial advisors and investors as an indication of the state of the overall financial markets and the economy in the United States. The share prices for the 30 companies listed in Table 2.1 are the basis for computing the Dow Jones Industrial Average (DJI), which is tracked continuously by virtually every financial publication. The index is named for Charles Dow and Edward Jones who first began calculating the DJI in 1896.

A characteristic or a quantity of interest that can take on different values is known as a **variable**; for the data in Table 2.1, the variables are Symbol, Industry, Share Price, and

TABLE 2.1 Data for Dow Jones Industrial Index Companies

Company	Symbol	Industry	Share Price ($)	Volume
Apple	AAPL	Technology	195.57	21,060,685
American Express	AXP	Financial	123.16	2,387,770
Boeing	BA	Manufacturing	369.32	3,002,708
Caterpillar	CAT	Manufacturing	133.71	3,747,782
Cisco Systems	CSCO	Technology	56.08	25,533,426
Chevron Corporation	CVX	Chemical, Oil, and Gas	123.64	4,705,879
Disney	DIS	Entertainment	139.94	14,670,995
Dow, Inc.	DOW	Chemical, Oil, and Gas	49.69	4,002,257
Goldman Sachs	GS	Financial	196.06	1,828,219
The Home Depot	HD	Retail	204.74	3,583,573
IBM	IBM	Technology	138.36	2,797,803
Intel	INTC	Technology	46.85	16,658,127
Johnson & Johnson	JNJ	Pharmaceuticals	144.24	7,516,973
JPMorgan Chase	JPM	Banking	107.76	18,654,861
Coca-Cola	KO	Food and Drink	51.76	11,517,843
McDonald’s	MCD	Food and Drink	205.71	3,017,625
3M	MMM	Conglomerate	172.03	2,730,927
Merck	MRK	Pharmaceuticals	85.24	8,909,750
Microsoft	MSFT	Technology	133.43	33,328,420
Nike	NKE	Consumer Goods	82.62	7,335,836
Pfizer	PFE	Pharmaceuticals	43.76	26,952,088
Procter & Gamble	PG	Consumer Goods	111.72	6,795,912
Travelers	TRV	Insurance	153.13	1,295,768
UnitedHealth Group	UNH	Healthcare	247.66	3,178,942
United Technologies	UTX	Conglomerate	129.02	2,790,767
Visa	V	Financial	171.28	9,897,832
Verizon	VZ	Telecommunications	58.00	10,554,753
Walgreens Boots Alliance	WBA	Retail	52.95	8,535,442
Wal-Mart	WMT	Retail	110.72	6,104,935
ExxonMobil	XOM	Chemical, Oil, and Gas	76.27	9,722,688

Volume. An **observation** is a set of values corresponding to a set of variables; each row in Table 2.1 corresponds to an observation.

Decision variables used in optimization models are covered in Chapters 12, 13, and 14. Random variables are covered in greater detail in Chapters 4 and 11.

Practically every problem (and opportunity) that an organization (or individual) faces is concerned with the impact of the possible values of relevant variables on the business outcome. Thus, we are concerned with how the value of a variable can vary; **variation** is the difference in a variable measured over observations (time, customers, items, etc.).

The role of descriptive analytics is to collect and analyze data to gain a better understanding of variation and its impact on the business setting. The values of some variables are under direct control of the decision maker (these are often called decision variables). The values of other variables may fluctuate with uncertainty because of factors outside the direct control of the decision maker. In general, a quantity whose values are not known with certainty is called a **random variable, or uncertain variable**. When we collect data, we are gathering past observed values, or realizations of a variable. By collecting these past realizations of one or more variables, our goal is to learn more about the variation of a particular business situation.

2.2 Types of Data

Population and Sample Data

To ensure that the companies in the Dow form a representative sample, companies are periodically added and removed from the Dow. It is possible that the companies in the Dow today have changed from what is shown in Table 2.1.

Data can be categorized in several ways based on how they are collected and the type collected. In many cases, it is not feasible to collect data from the **population** of all elements of interest. In such instances, we collect data from a subset of the population known as a **sample**. For example, with the thousands of publicly traded companies in the United States, tracking and analyzing all of these stocks every day would be too time consuming and expensive. The Dow represents a sample of 30 stocks of large public companies based in the United States, and it is often interpreted to represent the larger population of all publicly traded companies. It is very important to collect sample data that are representative of the population data so that generalizations can be made from them. In most cases (although not true of the Dow), a representative sample can be gathered by **random sampling** from the population data. Dealing with populations and samples can introduce subtle differences in how we calculate and interpret summary statistics. In almost all practical applications of business analytics, we will be dealing with sample data.

Quantitative and Categorical Data

Data are considered **quantitative data** if numeric and arithmetic operations, such as addition, subtraction, multiplication, and division, can be performed on them. For instance, we can sum the values for Volume in the Dow data in Table 2.1 to calculate a total volume of all shares traded by companies included in the Dow. If arithmetic operations cannot be performed on the data, they are considered **categorical data**. We can summarize categorical data by counting the number of observations or computing the proportions of observations in each category. For instance, the data in the Industry column in Table 2.1 are categorical. We can count the number of companies in the Dow that are in the telecommunications industry. Table 2.1 shows three companies in the financial industry: American Express, Goldman Sachs, and Visa. We cannot perform arithmetic operations on the data in the Industry column.

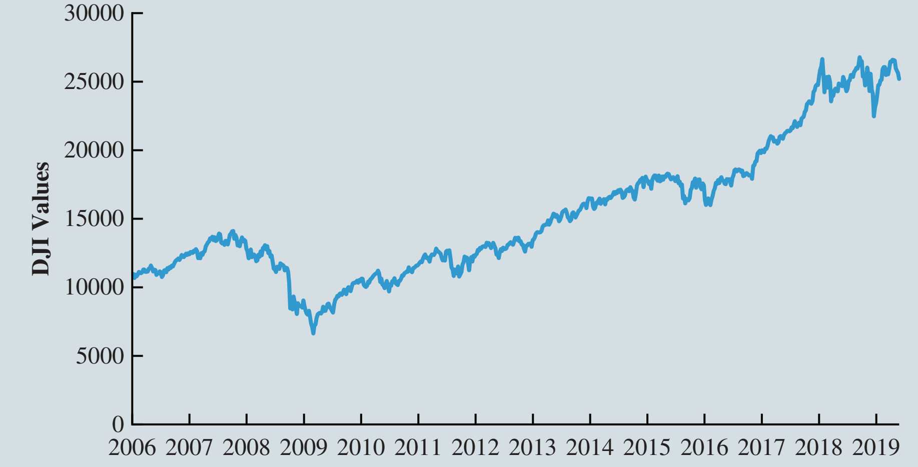
Cross-Sectional and Time Series Data

For statistical analysis, it is important to distinguish between cross-sectional data and time series data. **Cross-sectional data** are collected from several entities at the same, or approximately the same, point in time. The data in Table 2.1 are cross-sectional because they describe the 30 companies that comprise the Dow at the same point in time (June 2019). **Time series data** are collected over several time periods. Graphs of time series data are frequently found in business and economic publications. Such graphs help analysts understand what happened in the past, identify trends over time, and project future levels for the time series. For example, the graph of the time series in Figure 2.1 shows the DJI value from January 2006 to May 2019. The figure illustrates that the DJI limbed to above 14,000 in 2007. However, the financial crisis in 2008 led to a significant decline in the DJI to between 6,000 and 7,000 by 2009. Since 2009, the DJI has been generally increasing and topped 26,000 in 2019.

Sources of Data

Data necessary to analyze a business problem or opportunity can often be obtained with an appropriate study; such statistical studies can be classified as either experimental or observational. In an *experimental study*, a variable of interest is first identified. Then one or more other variables are identified and controlled or manipulated to obtain data about how these variables influence the variable of interest. For example, if a pharmaceutical firm conducts an experiment to learn about how a new drug affects blood pressure, then blood pressure is the variable of interest. The dosage level of the new drug is another variable that is hoped to have a causal effect on blood pressure. To obtain data about the effect of

FIGURE 2.1 Dow Jones Industrial Average Values Since 2006



the new drug, researchers select a sample of individuals. The dosage level of the new drug is controlled by giving different dosages to the different groups of individuals. Before and after the study, data on blood pressure are collected for each group. Statistical analysis of these experimental data can help determine how the new drug affects blood pressure.

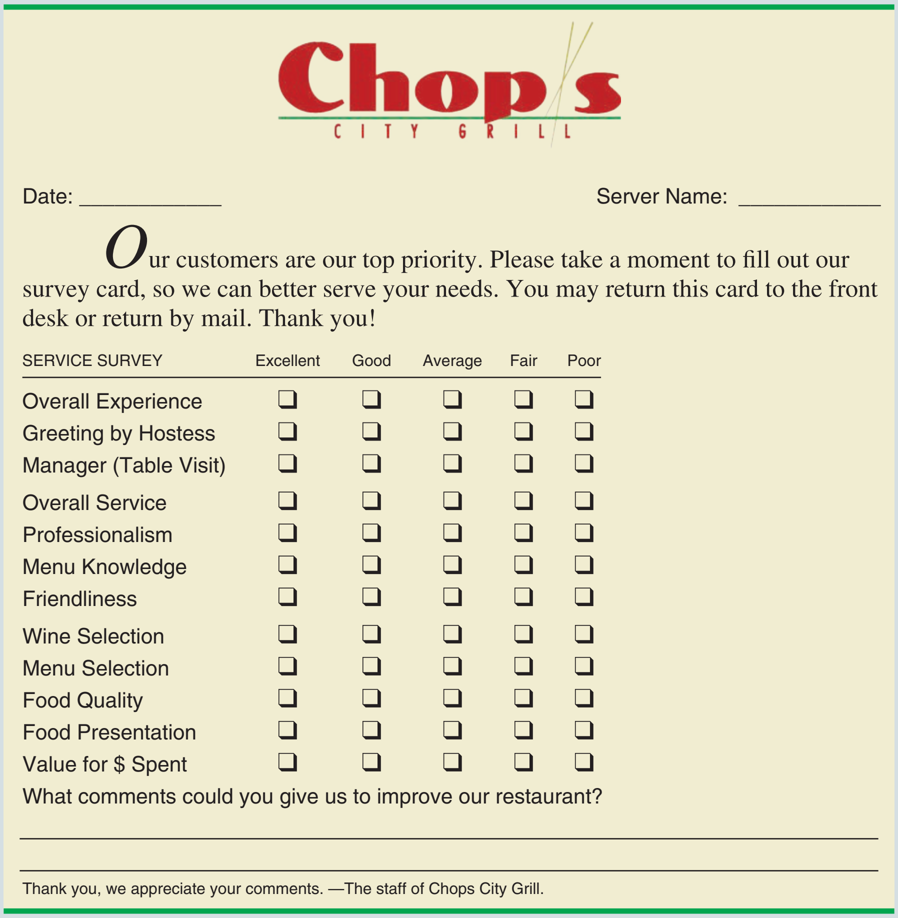
Nonexperimental, or *observational*, *studies* make no attempt to control the variables of interest. A survey is perhaps the most common type of observational study. For instance, in a personal interview survey, research questions are first identified. Then a questionnaire is designed and administered to a sample of individuals. Some restaurants use observational studies to obtain data about customer opinions with regard to the quality of food, quality of service, atmosphere, and so on. A customer opinion questionnaire used by Chops City Grill in Naples, Florida, is shown in Figure 2.2. Note that the customers who fill out the questionnaire are asked to provide ratings for 12 variables, including overall experience, the greeting by hostess, the table visit by the manager, overall service, and so on. The response categories of excellent, good, average, fair, and poor provide categorical data that enable Chops City Grill management to maintain high standards for the restaurant's food and service.

In some cases, the data needed for a particular application exist from an experimental or observational study that has already been conducted. For example, companies maintain a variety of databases about their employees, customers, and business operations. Data on employee salaries, ages, and years of experience can usually be obtained from internal personnel records. Other internal records contain data on sales, advertising expenditures, distribution costs, inventory levels, and production quantities. Most companies also maintain detailed data about their customers.

In Chapter 15 we discuss methods for determining the value of additional information that can be provided by collecting data.

Anyone who wants to use data and statistical analysis to aid in decision making must be aware of the time and cost required to obtain the data. The use of existing data sources is desirable when data must be obtained in a relatively short period of time. If important data are not readily available from a reliable existing source, the additional time and cost involved in obtaining the data must be taken into account. In all cases, the decision maker should consider the potential contribution of the statistical analysis to the decision-making process. The cost of data acquisition and the subsequent statistical analysis should not exceed the savings generated by using the information to make a better decision.

FIGURE 2.2 Customer Opinion Questionnaire Used by Chops City Grill Restaurant



Chops
CITY GRILL

Date: ____________ Server Name: ____________

Our customers are our top priority. Please take a moment to fill out our survey card, so we can better serve your needs. You may return this card to the front desk or return by mail. Thank you!

SERVICE SURVEY	Excellent	Good	Average	Fair	Poor
Overall Experience	□	□	□	□	□
Greeting by Hostess	□	□	□	□	□
Manager (Table Visit)	□	□	□	□	□
Overall Service	□	□	□	□	□
Professionalism	□	□	□	□	□
Menu Knowledge	□	□	□	□	□
Friendliness	□	□	□	□	□
Wine Selection	□	□	□	□	□
Menu Selection	□	□	□	□	□
Food Quality	□	□	□	□	□
Food Presentation	□	□	□	□	□
Value for $ Spent	□	□	□	□	□

What comments could you give us to improve our restaurant?

____

____

Thank you, we appreciate your comments. —The staff of Chops City Grill.

NOTES + COMMENTS

1. Organizations that specialize in collecting and maintaining data make available substantial amounts of business and economic data. Companies can access these external data sources through leasing arrangements or by purchase. Dun & Bradstreet, Bloomberg, and Dow Jones & Company are three firms that provide extensive business database services to clients. Nielsen and Ipsos are two companies that have built successful businesses collecting and processing data that they sell to advertisers and product manufacturers. Data are also available from a variety of industry associations and special-interest organizations.
2. Government agencies are another important source of existing data. For instance, the web site data.gov was launched by the U.S. government in 2009 to make it easier for the public to access data collected by the U.S. federal government. The data.gov web site includes over 150,000 data sets from a variety of U.S. federal departments and agencies, but many other federal agencies maintain their own web sites and data repositories. Many state and local governments are also now providing data sets online. As examples, the states of California and Texas maintain open data portals at data.ca.gov and data.texas.gov, respectively. New York City's open data web site is opendata.cityofnewyork.us and the city of Cincinnati, Ohio, is at data.cincinnati-oh.gov. In general, the Internet is an important source of data and statistical information. One can obtain access to stock quotes, meal prices at restaurants, salary data, and a wide array of other information simply by performing an Internet search.

2.3 Modifying Data in Excel

Projects often involve so much data that it is difficult to analyze all of the data at once. In this section, we examine methods for summarizing and manipulating data using Excel to make the data more manageable and to develop insights.

Sorting and Filtering Data in Excel

Excel contains many useful features for sorting and filtering data so that one can more easily identify patterns. Table 2.2 contains data on the 20 top-selling passenger-car automobiles in the United States in February 2019. The table shows the model and manufacturer of each automobile as well as the sales for the model in February 2019 and February 2018.

Figure 2.3 shows the data from Table 2.2 entered into an Excel spreadsheet, and the percent change in sales for each model from February 2018 to February 2019 has been calculated. This is done by entering the formula *=(D2-E2)/E2* in cell F2 and then copying the contents of this cell to cells F3 to F20.

Suppose that we want to sort these automobiles by February 2018 sales instead of by February 2019 sales. To do this, we use Excel's Sort function, as shown in the following steps.

Step 1. Select cells A1:F21
Step 2. Click the **Data** tab in the Ribbon
Step 3. Click **Sort** in the **Sort & Filter** group

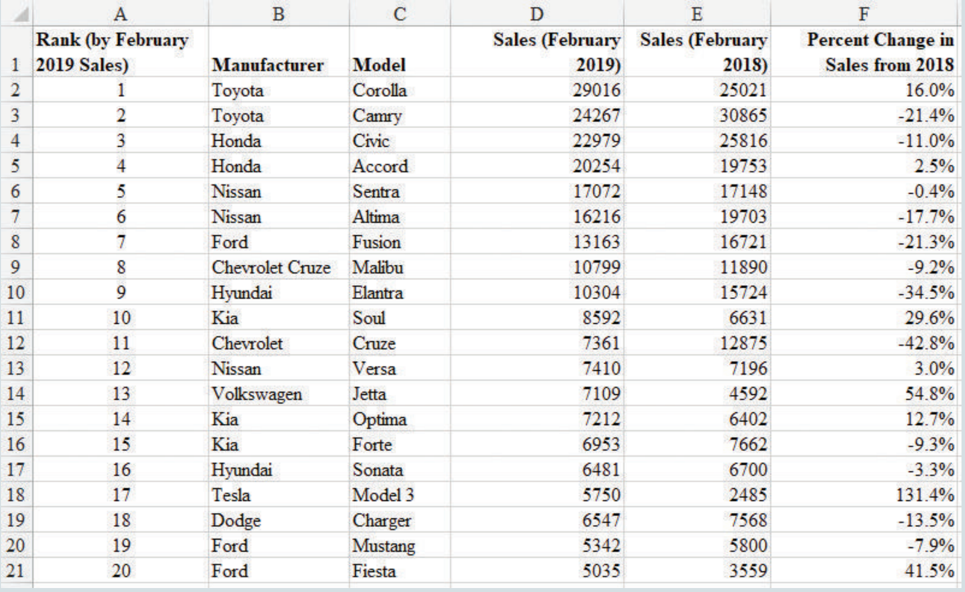
DATA*file*
Top20Cars2019

TABLE 2.2 20 Top-Selling Automobiles in United States in February 2019

Rank (by February 2019 Sales)	Manufacturer	Model	Sales (February 2019)	Sales (February 2018)
1	Toyota	Corolla	29,016	25,021
2	Toyota	Camry	24,267	30,865
3	Honda	Civic	22,979	25,816
4	Honda	Accord	20,254	19,753
5	Nissan	Sentra	17,072	17,148
6	Nissan	Altima	16,216	19,703
7	Ford	Fusion	13,163	16,721
8	Chevrolet	Malibu	10,799	11,890
9	Hyundai	Elantra	10,304	15,724
10	Kia	Soul	8,592	6,631
11	Chevrolet	Cruze	7,361	12,875
12	Nissan	Versa	7,410	7,196
13	Volkswagen	Jetta	7,109	4,592
14	Kia	Optima	7,212	6,402
15	Kia	Forte	6,953	7,662
16	Hyundai	Sonata	6,481	6,700
17	Tesla	Model 3	5,750	2,485
18	Dodge	Charger	6,547	7,568
19	Ford	Mustang	5,342	5,800
20	Ford	Fiesta	5,035	3,559

Source: *Manufacturers and Automotive News Data Center.*

FIGURE 2.3 Data for 20 Top-Selling Automobiles Entered into Excel with Percent Change in Sales from 2018



	A	B	C	D	E	F
1	Rank (by February 2019 Sales)	Manufacturer	Model	Sales (February 2019)	Sales (February 2018)	Percent Change in Sales from 2018
2	1	Toyota	Corolla	29016	25021	16.0%
3	2	Toyota	Camry	24267	30865	-21.4%
4	3	Honda	Civic	22979	25816	-11.0%
5	4	Honda	Accord	20254	19753	2.5%
6	5	Nissan	Sentra	17072	17148	-0.4%
7	6	Nissan	Altima	16216	19703	-17.7%
8	7	Ford	Fusion	13163	16721	-21.3%
9	8	Chevrolet Cruze	Malibu	10799	11890	-9.2%
10	9	Hyundai	Elantra	10304	15724	-34.5%
11	10	Kia	Soul	8592	6631	29.6%
12	11	Chevrolet	Cruze	7361	12875	-42.8%
13	12	Nissan	Versa	7410	7196	3.0%
14	13	Volkswagen	Jetta	7109	4592	54.8%
15	14	Kia	Optima	7212	6402	12.7%
16	15	Kia	Forte	6953	7662	-9.3%
17	16	Hyundai	Sonata	6481	6700	-3.3%
18	17	Tesla	Model 3	5750	2485	131.4%
19	18	Dodge	Charger	6547	7568	-13.5%
20	19	Ford	Mustang	5342	5800	-7.9%
21	20	Ford	Fiesta	5035	3559	41.5%

Step 4. Select the check box for **My data has headers**
Step 5. In the first **Sort by** dropdown menu, select **Sales (February 2018)**
Step 6. In the **Order** dropdown menu, select **Largest to Smallest** (see Figure 2.4)
Step 7. Click **OK**

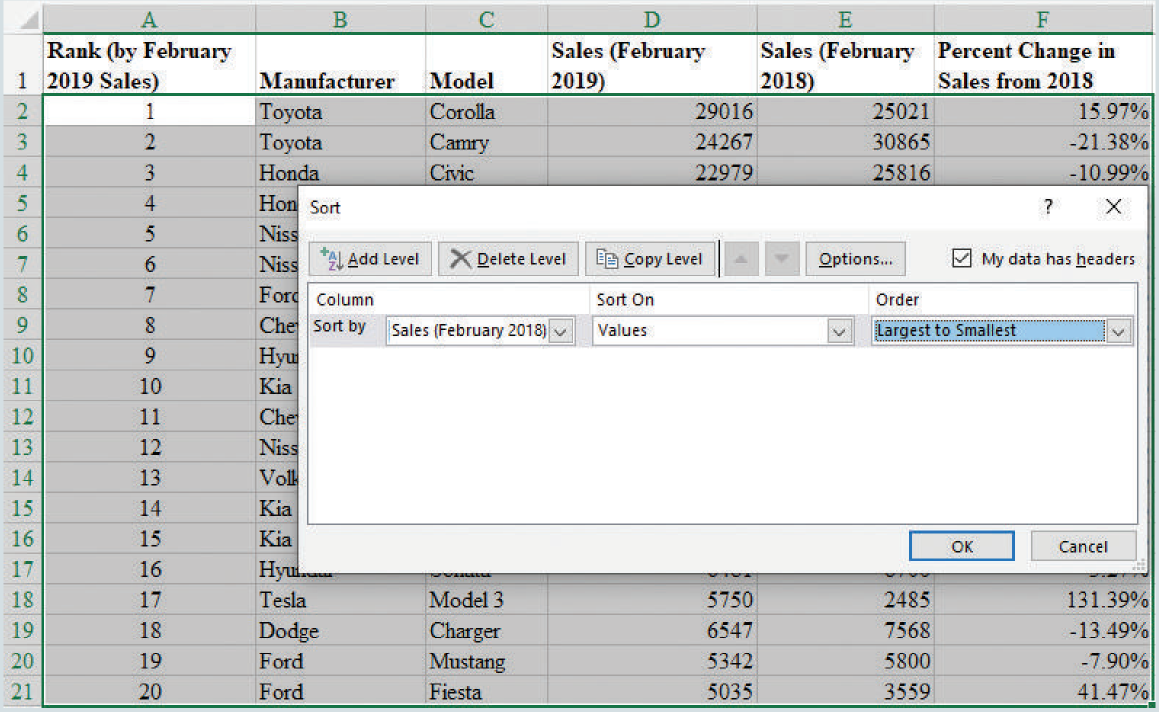
The result of using Excel's Sort function for the February 2018 data is shown in Figure 2.5. Now we can easily see that, although the Toyota Corolla was the best-selling automobile in February 2019, both the Toyota Camry and the Honda Civic outsold the Toyota Corolla in February 2018. Note that while we sorted on Sales (February 2018), which is in column E, the data in all other columns are adjusted accordingly.

Now let's suppose that we are interested only in seeing the sales of models made by Nissan. We can do this using Excel's Filter function:

Step 1. Select cells A1:F21
Step 2. Click the **Data** tab in the Ribbon
Step 3. Click **Filter** in the **Sort & Filter** group
Step 4. Click on the **Filter Arrow** ▾ in column B, next to **Manufacturer**
Step 5. If all choices are checked, you can easily deselect all choices by unchecking (**Select All**). Then select only the check box for **Nissan**.
Step 6. Click **OK**

The result is a display of only the data for models made by Nissan (see Figure 2.6). We now see that of the 20 top-selling models in February 2019, Nissan made three of them: the Altima, the Sentra, and the Versa. We can further filter the data by choosing the down arrows in the other columns. We can make all data visible again by clicking on the down arrow in column B and checking **(Select All)** and clicking **OK**, or by clicking **Filter** in the **Sort & Filter** Group again from the **Data** tab.

FIGURE 2.4 Using Excel's Sort Function to Sort the Top-Selling Automobiles Data



	A	B	C	D	E	F
1	Rank (by February 2019 Sales)	Manufacturer	Model	Sales (February 2019)	Sales (February 2018)	Percent Change in Sales from 2018
2	1	Toyota	Corolla	29016	25021	15.97%
3	2	Toyota	Camry	24267	30865	-21.38%
4	3	Honda	Civic	22979	25816	-10.99%
5	4	Hon				
6	5	Niss				
7	6	Niss				
8	7	Ford				
9	8	Che				
10	9	Hyu				
11	10	Kia				
12	11	Che				
13	12	Niss				
14	13	Volk				
15	14	Kia				
16	15	Kia				
17	16	Hyundai	Sonata			
18	17	Tesla	Model 3	5750	2485	131.39%
19	18	Dodge	Charger	6547	7568	-13.49%
20	19	Ford	Mustang	5342	5800	-7.90%
21	20	Ford	Fiesta	5035	3559	41.47%

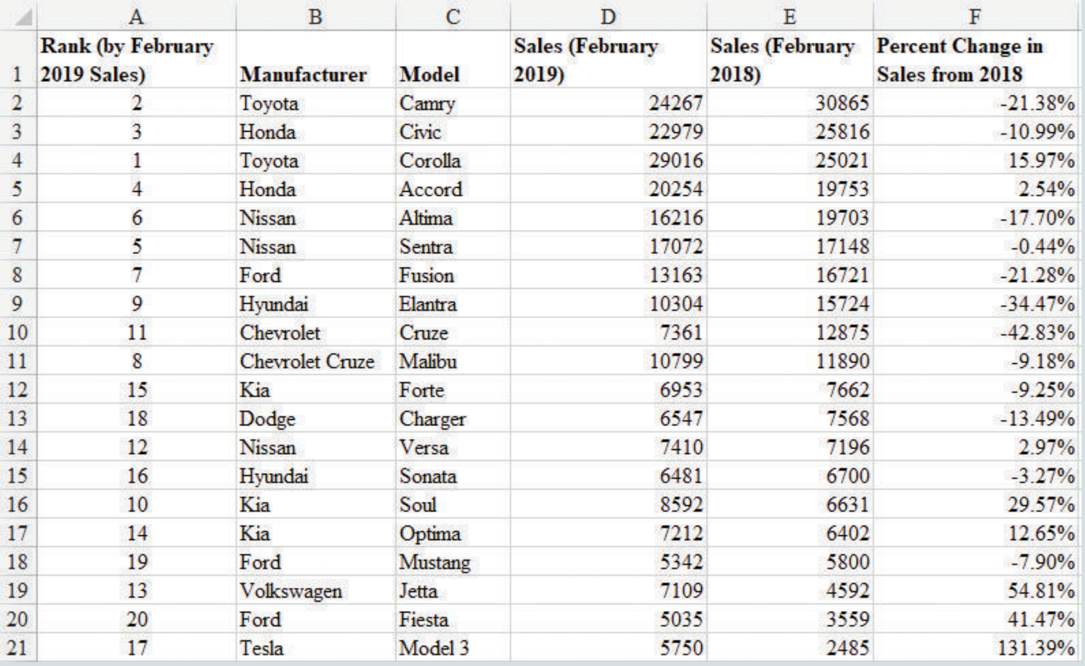
Sort

Add Level | Delete Level | Copy Level | Options... | ☑ My data has headers

	Column	Sort On	Order
Sort by	Sales (February 2018)	Values	Largest to Smallest

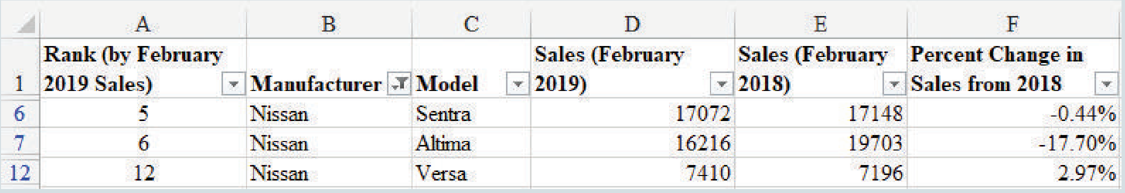
OK | Cancel

FIGURE 2.5 Top-Selling Automobiles Data Sorted by Sales in February 2018 Sales



	A	B	C	D	E	F
1	Rank (by February 2019 Sales)	Manufacturer	Model	Sales (February 2019)	Sales (February 2018)	Percent Change in Sales from 2018
2	2	Toyota	Camry	24267	30865	-21.38%
3	3	Honda	Civic	22979	25816	-10.99%
4	1	Toyota	Corolla	29016	25021	15.97%
5	4	Honda	Accord	20254	19753	2.54%
6	6	Nissan	Altima	16216	19703	-17.70%
7	5	Nissan	Sentra	17072	17148	-0.44%
8	7	Ford	Fusion	13163	16721	-21.28%
9	9	Hyundai	Elantra	10304	15724	-34.47%
10	11	Chevrolet	Cruze	7361	12875	-42.83%
11	8	Chevrolet Cruze	Malibu	10799	11890	-9.18%
12	15	Kia	Forte	6953	7662	-9.25%
13	18	Dodge	Charger	6547	7568	-13.49%
14	12	Nissan	Versa	7410	7196	2.97%
15	16	Hyundai	Sonata	6481	6700	-3.27%
16	10	Kia	Soul	8592	6631	29.57%
17	14	Kia	Optima	7212	6402	12.65%
18	19	Ford	Mustang	5342	5800	-7.90%
19	13	Volkswagen	Jetta	7109	4592	54.81%
20	20	Ford	Fiesta	5035	3559	41.47%
21	17	Tesla	Model 3	5750	2485	131.39%

FIGURE 2.6 Top-Selling Automobiles Data Filtered to Show Only Automobiles Manufactured by Nissan



	A	B	C	D	E	F
1	Rank (by February 2019 Sales)	Manufacturer	Model	Sales (February 2019)	Sales (February 2018)	Percent Change in Sales from 2018
6	5	Nissan	Sentra	17072	17148	-0.44%
7	6	Nissan	Altima	16216	19703	-17.70%
12	12	Nissan	Versa	7410	7196	2.97%

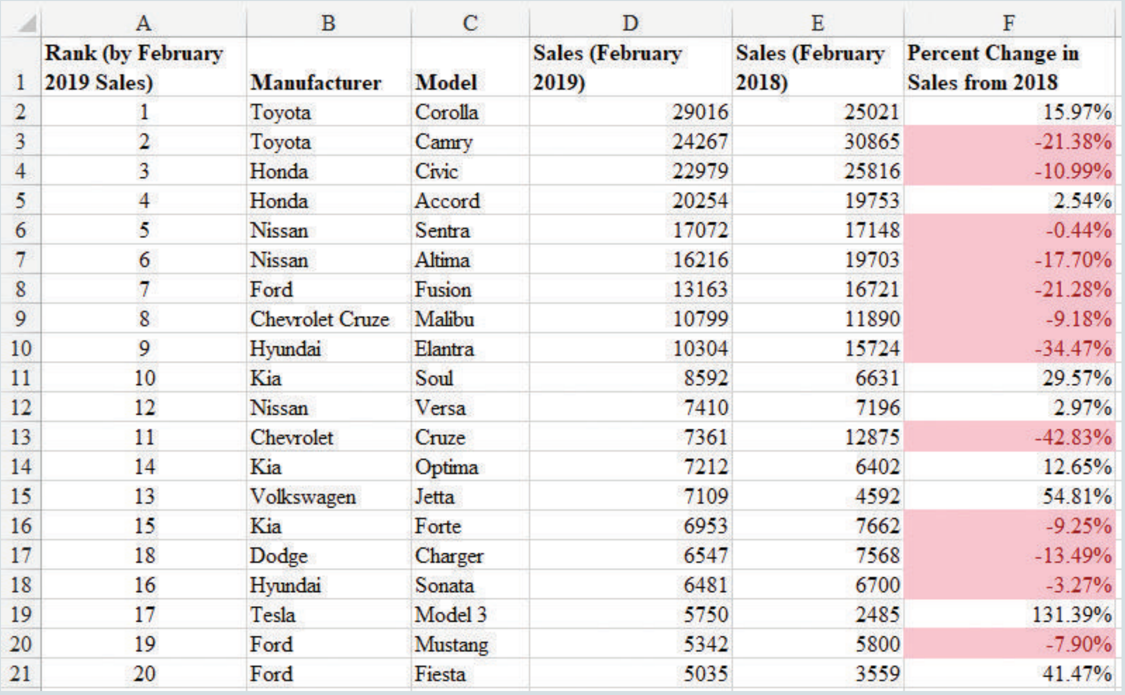
Conditional Formatting of Data in Excel

Conditional formatting in Excel can make it easy to identify data that satisfy certain conditions in a data set. For instance, suppose that we wanted to quickly identify the automobile models in Table 2.2 for which sales had decreased from February 2018 to February 2019. We can quickly highlight these models:

Step 1. Starting with the original data shown in Figure 2.3, select cells F1:F21
Step 2. Click the **Home** tab in the Ribbon
Step 3. Click **Conditional Formatting** in the **Styles** group
Step 4. Select **Highlight Cells Rules**, and click **Less Than . . .** from the dropdown menu
Step 5. Enter *0%* in the **Format cells that are LESS THAN:** box
Step 6. Click **OK**

The results are shown in Figure 2.7. Here we see that the models with decreasing sales (for example, Toyota Camry, Honda Civic, Nissan Sentra, Nissan Altima) are now

FIGURE 2.7 Using Conditional Formatting in Excel to Highlight Automobiles with Declining Sales from February 2018



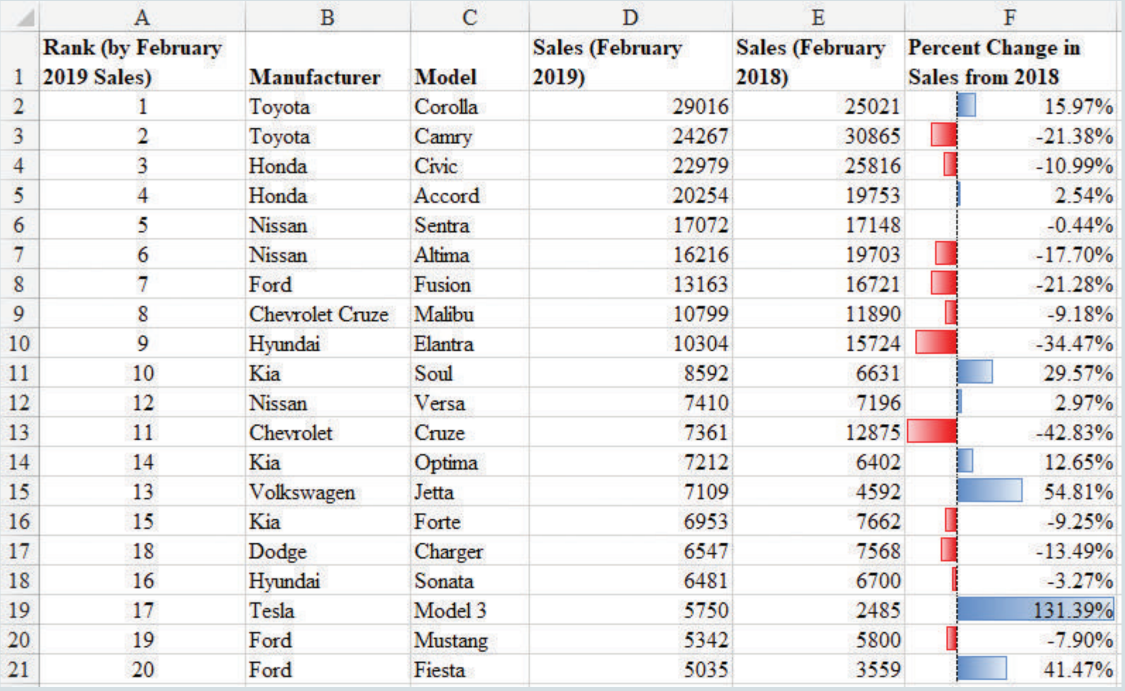
	A	B	C	D	E	F
1	Rank (by February 2019 Sales)	Manufacturer	Model	Sales (February 2019)	Sales (February 2018)	Percent Change in Sales from 2018
2	1	Toyota	Corolla	29016	25021	15.97%
3	2	Toyota	Camry	24267	30865	-21.38%
4	3	Honda	Civic	22979	25816	-10.99%
5	4	Honda	Accord	20254	19753	2.54%
6	5	Nissan	Sentra	17072	17148	-0.44%
7	6	Nissan	Altima	16216	19703	-17.70%
8	7	Ford	Fusion	13163	16721	-21.28%
9	8	Chevrolet Cruze	Malibu	10799	11890	-9.18%
10	9	Hyundai	Elantra	10304	15724	-34.47%
11	10	Kia	Soul	8592	6631	29.57%
12	12	Nissan	Versa	7410	7196	2.97%
13	11	Chevrolet	Cruze	7361	12875	-42.83%
14	14	Kia	Optima	7212	6402	12.65%
15	13	Volkswagen	Jetta	7109	4592	54.81%
16	15	Kia	Forte	6953	7662	-9.25%
17	18	Dodge	Charger	6547	7568	-13.49%
18	16	Hyundai	Sonata	6481	6700	-3.27%
19	17	Tesla	Model 3	5750	2485	131.39%
20	19	Ford	Mustang	5342	5800	-7.90%
21	20	Ford	Fiesta	5035	3559	41.47%

clearly visible. Note that Excel's Conditional Formatting function offers tremendous flexibility. Instead of highlighting only models with decreasing sales, we could instead choose **Data Bars** from the **Conditional Formatting** dropdown menu in the **Styles** Group of the **Home** tab in the Ribbon. The result of using the **Blue Data Bar Gradient Fill** option is shown in Figure 2.8. Data bars are essentially a bar chart input into the cells that shows the magnitude of the cell values. The widths of the bars in this display are comparable to the values of the variable for which the bars have been drawn; a value of 20 creates a bar twice as wide as that for a value of 10. Negative values are shown to the left side of the axis; positive values are shown to the right. Cells with negative values are shaded in red, and those with positive values are shaded in blue. Again, we can easily see which models had decreasing sales, but Data Bars also provide us with a visual representation of the magnitude of the change in sales. Many other Conditional Formatting options are available in Excel.

Bar charts and other graphical presentations will be covered in detail in Chapter 3. We will see other uses for Conditional Formatting in Excel in Chapter 3.

The **Quick Analysis** button in Excel appears just outside the bottom-right corner of a group of selected cells whenever you select multiple cells. Clicking the **Quick Analysis** button gives you shortcuts for Conditional Formatting, adding Data Bars, and other operations. Clicking on this button gives you the options shown in Figure 2.9 for **Formatting**. Note that there are also tabs for **Charts**, **Totals**, **Tables**, and **Sparklines**.

FIGURE 2.8 Using Conditional Formatting in Excel to Generate Data Bars for the Top-Selling Automobiles Data



	A	B	C	D	E	F
1	Rank (by February 2019 Sales)	Manufacturer	Model	Sales (February 2019)	Sales (February 2018)	Percent Change in Sales from 2018
2	1	Toyota	Corolla	29016	25021	15.97%
3	2	Toyota	Camry	24267	30865	-21.38%
4	3	Honda	Civic	22979	25816	-10.99%
5	4	Honda	Accord	20254	19753	2.54%
6	5	Nissan	Sentra	17072	17148	-0.44%
7	6	Nissan	Altima	16216	19703	-17.70%
8	7	Ford	Fusion	13163	16721	-21.28%
9	8	Chevrolet Cruze	Malibu	10799	11890	-9.18%
10	9	Hyundai	Elantra	10304	15724	-34.47%
11	10	Kia	Soul	8592	6631	29.57%
12	12	Nissan	Versa	7410	7196	2.97%
13	11	Chevrolet	Cruze	7361	12875	-42.83%
14	14	Kia	Optima	7212	6402	12.65%
15	13	Volkswagen	Jetta	7109	4592	54.81%
16	15	Kia	Forte	6953	7662	-9.25%
17	18	Dodge	Charger	6547	7568	-13.49%
18	16	Hyundai	Sonata	6481	6700	-3.27%
19	17	Tesla	Model 3	5750	2485	131.39%
20	19	Ford	Mustang	5342	5800	-7.90%
21	20	Ford	Fiesta	5035	3559	41.47%