PEARSON BUSINESS ANALYTICS SERIES

FOURTH EDITION

EVEN YOU CAN LEARN STATISTICS and ANALYTICS

An Easy to Understand Guide

DAVID M. LEVINE | DAVID F. STEPHAN

FREE SAMPLE CHAPTER |

Even You Can Learn Statistics and Analytics

Fourth Edition

An Easy to Understand Guide to Statistics and Analytics

David M. Levine

David F. Stephan

PEARSON

Boston • Columbus • New York • San Francisco • Amsterdam • Cape Town
Dubai • London • Madrid • Milan • Munich • Paris • Montreal • Toronto • Delhi • Mexico City
São Paulo • Sidney • Hong Kong • Seoul • Singapore • Taipei • Tokyo

Editor-in-Chief: Mark L. Taub
Acquisitions Editor: Kim Spenceley
Development Editor: Chris Zahn
Managing Editor: Sandra Schroeder
Project Editor: Mandie Frank
Production Manager: Remya Divakaran/codeMantra
Copy Editor: Kitty Wilson
Indexer: Timothy Wright
Proofreader: Donna Mulder
Designer: Chuti Prasertsith
Compositor: codeMantra

ScoutAutomatedPrintCode

ISBN-13: 978-013-765476-5
ISBN-10: 0-13-765476-6

Library of Congress Control Number: 2021947626

Pearson's Commitment to Diversity, Equity, and Inclusion

Pearson is dedicated to creating bias-free content that reflects the diversity of all learners. We embrace the many dimensions of diversity, including but not limited to race, ethnicity, gender, socioeconomic status, ability, age, sexual orientation, and religious or political beliefs.

Education is a powerful force for equity and change in our world. It has the potential to deliver opportunities that improve lives and enable economic mobility. As we work with authors to create content for every product and service, we acknowledge our responsibility to demonstrate inclusivity and incorporate diverse scholarship so that everyone can achieve their potential through learning. As the world's leading learning company, we have a duty to help drive change and live up to our purpose to help more people create a better life for themselves and to create a better world.

Our ambition is to purposefully contribute to a world where:

- Everyone has an equitable and lifelong opportunity to succeed through learning.
- Our educational products and services are inclusive and represent the rich diversity of learners.
- Our educational content accurately reflects the histories and experiences of the learners we serve.
- Our educational content prompts deeper discussions with learners and motivates them to expand their own learning (and worldview).

While we work hard to present unbiased content, we want to hear from you about any concerns or needs with this Pearson product so that we can investigate and address them.

- Please contact us with concerns about any potential bias at https://www.pearson.com/report-bias.html.

Credits

Cover	ZinetroN/Shutterstock
Unnumbered Figure 3-1 – Unnumbered Figure 3-3	Microsoft Corporation
Unnumbered Figure 5-1 – Unnumbered Figure 5-3	
Figure 6-2	
Figure 6-3	
Figure 8-2 – Figure 8-5	
Unnumbered Figure 8-1	
Unnumbered Figure 8-2	
Figure 9-1 – Figure 9-3	
Figure 9-5	
Figure 9-6	
Figure 10-3	
Figure 11-1	
Figure 12-1 – Figure 12-3	
Figure 12-5 – Figure 12-7	
Figure E-1 – Figure E-5	
Unnumbered Figure E-1	
Unnumbered Figure E-2	
Figure 13-5	JMP Statistical Discovery LLC
Figure 13-6	

To our wives and our children,
and in loving memory of our parents

This page intentionally left blank

Table of Contents

Acknowledgments

We would especially like to thank the staff at Pearson: Kim Spenceley for making this fourth edition a reality, Kitty Wilson for her copy editing, Lori Lyons and Mandie Frank for their work in the production of this text.

We have sought to make the contents of this book as clear, accurate, and error-free as possible. We invite you to make suggestions or ask questions about the content if you think we have fallen short of our goals in any way. Please email your comments to authors@davidlevinestatistics.com and include the hashtag #EYCLSA4 in the subject line of your message.

About the Authors

David M. Levine and **David F. Stephan** are part of a writing team known for their series of business statistics textbooks that include *Basic Business Statistics, Business Statistics: A First Course*, and *Statistics for Managers Using Microsoft Excel*. In long teaching careers at Baruch College, both were known for their classroom innovations, with Levine being honored with a Presidential Excellence Award for Distinguished Teaching Award and Stephan granted the privilege to design and develop the College's first computer-based classroom. Both are active members of the Data, Analytics and Statistics Instruction SIG of the Decision Sciences Institute.

Levine is Professor Emeritus of Information Systems at Baruch College. He is nationally recognized innovator in business statistics education and is also the coauthor of *Applied Statistics for Engineers and Scientists Using Microsoft Excel and Minitab*. Levine is also the author or coauthor of four books about statistical quality management: *Statistics for Six Sigma Green Belts and Champions, Six Sigma for Green Belts and Champions, Design for Six Sigma for Green Belts and Champions,* and *Quality Management*, 3rd Edition. He has published articles in various journals, including *Psychometrika, The American Statistician, Communications in Statistics, Multivariate Behavioral Research, Journal of Systems Management, Quality Progress,* and *The American Anthropologist*, and has given numerous talks at American Statistical Association, Decision Sciences Institute, and Making Statistics More Effective in Schools of Business conferences.

During his more than 20 years at Baruch College, **Stephan** devised techniques for teaching computer applications such as Microsoft Excel in a business context and developed future-forward courses that explored the effects of emerging digital technologies. He also served as the associate director of a U.S. Department of Education FIPSE project that successfully integrated interactive media into classroom instruction for the humanities.. Stephan is also the developer of PHStat, the statistics add-in for Microsoft Excel distributed by Pearson Education.

This page intentionally left blank

Introduction

The *Even You Can Learn Statistics and Analytics* Owner's Manual

In today's world, understanding statistics and analytics is more important than ever before. *Even You Can Learn Statistics and Analytics: An Easy to Understand Guide to Statistics and Analytics* teaches you the basic concepts that provide you with the knowledge to apply statistics and analytics in your life. You will also learn the most commonly used statistical methods and have the opportunity to practice those methods while using Microsoft Excel.

Please read the rest of this introduction so that you can become familiar with the distinctive features of this book. To download files that support your learning of statistics, visit the website for this book at www.informit.com.

Mathematics Is Always Optional!

Never mastered higher mathematics—or generally fearful of math? Not to worry, because in *Even You Can Learn Statistics and Analytics*, you will find that every concept is explained in plain English, without the use of higher mathematics or mathematical symbols. However, if you *are* interested in the mathematical foundations behind statistics, *Even You Can Learn Statistics and Analytics* includes **Equation Blackboards**, stand-alone sections that present the equations behind statistical methods and complement the main material.

Learning with the Concept-Interpretation Approach

Even You Can Learn Statistics and Analytics uses a **Concept-Interpretation** approach to help you learn statistics and analytics:

- A **CONCEPT**, a plain language definition that uses no complicated mathematical terms.
- An **INTERPRETATION**, that fully explains the concept and its importance to statistics. When necessary, these sections also include common misconceptions about the concept as well as the common errors people can make when trying to apply the concept.

For simpler concepts, an **EXAMPLES** section lists real-life examples or applications of the statistical concepts. For more involved concepts, **WORKED-OUT PROBLEMS** provide complete solutions to statistical problems—including actual spreadsheet results—that illustrate how you can apply the concepts to other problems.

Practicing Statistics While You Learn Statistics

To help you learn statistics, you should always review the worked-out problems that appear in this book. As you review them, you can practice what you have just learned by using the optional **SPREADSHEET SOLUTION** sections.

Spreadsheet Solution sections enable you to use Microsoft Excel as you learn statistics. If you don't want to practice your spreadsheet skills, you can examine the spreadsheet results that appear throughout the book. Many spreadsheet results are available as files that you can download for free through the InformIT website, www.informit.com. Please visit the website for this book at www.informit.com to access these bonus materials.

Spreadsheet program users will also benefit from Appendix D and Appendix E, which help teach you more about spreadsheets as you learn statistics.

And if technical issues or instructions have ever confounded your using Microsoft Excel in the past, check out Appendix A, which details the technical configuration issues you might face and explains the conventions used in all technical instructions that appear in this book.

In-Chapter Aids

As you read a chapter, look for the following icons for extra help:

Important Point icons highlight key definitions and explanations.

File icons identify the downloadable files that enable you to examine the data in selected problems.

Interested in the mathematical foundations of statistics? Then look for the Interested in Math? icons throughout the book. But remember, you can skip any or all of the math sections without losing any comprehension of the statistical methods presented, because math is always optional in this book!

End-of-Chapter Features

At the end of most chapters of *Even You Can Learn Statistics and Analytics,* you can find the following features, which you can review to reinforce your learning.

Important Equations

The **Important Equations** sections present all of the important equations discussed in the chapter. You can use these lists for reference and later study even if you have skipped over the Equation Blackboards and "interested in math" passages.

One-Minute Summaries

Each **One-Minute Summary** is a quick review of the significant topics in the chapter in outline form. When appropriate, the summaries also help guide you to make the right decisions about applying statistics to the data you seek to analyze.

Test Yourself

The **Test Yourself** sections offer a set of short-answer questions and problems that enable you to review and test yourself (with answers provided) to see how much you have retained of the concepts presented in a chapter.

Summary

Even You Can Learn Statistics and Analytics can help you whether you are taking a formal course in data analysis, brushing up on your knowledge of statistics for a specific analysis, or need to learn about analytics. If you have questions about this book, feel free to contact the authors via email at authors@davidlevinestatistics.com and include the hashtag #EYCLSA4 in the subject line of your email.

This page intentionally left blank

Presenting Data in Tables and Charts

Tables and charts are ways of summarizing categorical and numerical variables that can help you present information effectively. In this chapter, you will learn the appropriate types of tables and charts to use for each type of variable.

2.1 Presenting Categorical Variables

You present a categorical variable by first sorting values according to the categories of the variable. Then you place the count, amount, or percentage (part of the whole) of each category into a summary table or into one of several types of charts.

The Summary Table

CONCEPT A two-column table in which category names are listed in the first column and the counts, amounts, or percentages of values are listed in a second column. Sometimes, additional columns present the same data in more than one way (for example, as counts and percentages).

EXAMPLE A restaurant owner records the entrées ordered by guests during the Friday-to-Sunday weekend period. The data recorded can be presented using a summary table.

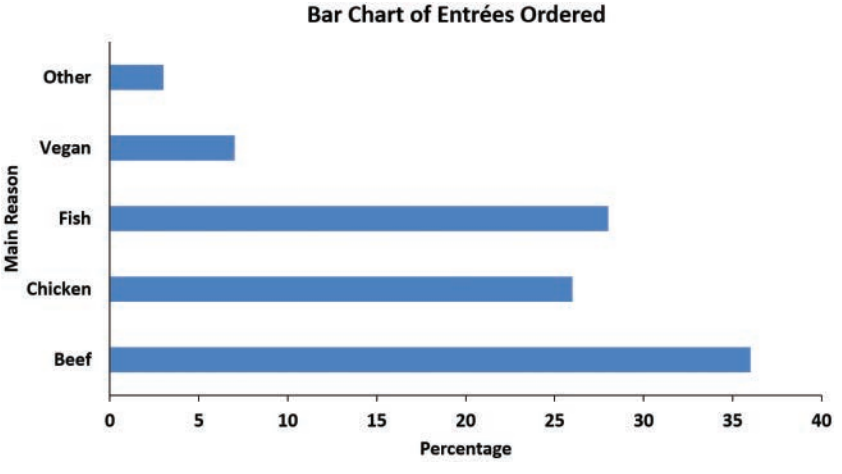
Entrée Ordered	Percentage
Beef	36
Chicken	26
Fish	28
Vegan	7
Other	3

INTERPRETATION Summary tables enable you to see the big picture about a set of data. In this example, you can conclude that most customers will order beef, chicken, or fish. Very few will order either vegan or other entrées.

The Bar Chart

CONCEPT A chart containing rectangles ("bars") in which the length of each bar represents the count, amount, or percentage of responses of one category.

EXAMPLE The data of the summary table that the previous concept uses can be visualized using a percentage bar chart.



INTERPRETATION A bar chart better presents the point that beef entrée is the single largest category of entrée ordered. For most people, scanning a bar chart is easier than scanning a column of numbers in which the numbers are unordered, as they are in the previous summary table.

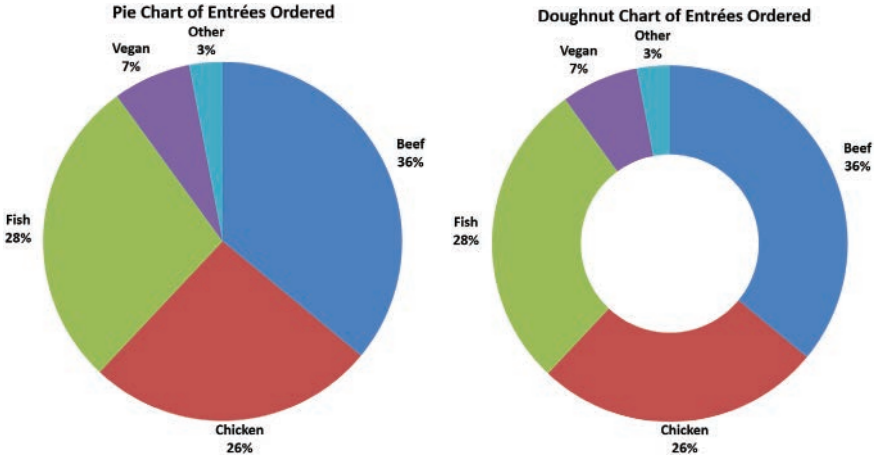
The Pie Chart and the Doughnut Chart

CONCEPT

Pie: A circle chart in which wedge-shaped areas—pie slices—represent the count, amount, or percentage of each category, and the entire circle ("pie") represents the total.

Doughnut: A circle chart in which parts of the circumference represent the count, amount, or percentage of each category, and the entire circumference represents the total.

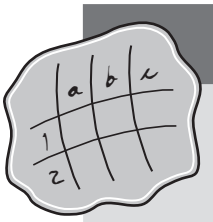
EXAMPLE The following pie and doughnut charts visualize the summary table data that the two preceding concepts use.



INTERPRETATION A pie chart or a doughnut chart enables you to see how the various categories contribute to the whole. In the example charts, you can see that chicken and fish entrées make up about half of all entrées ordered and that beef is the entrée most ordered.

In recent years, doughnut charts have become preferred over pie charts. The *area* of pie "slices" can be misperceived, making the pie slice seem larger or smaller than the percentage of the whole that the slice represents. In contrast, doughnut charts focus attention on the lengths of each arc, which are easier to compare and accurately reflect the percentage of the whole.

Note that pie and doughnut charts do not enable you to as easily compare categories as a bar chart does. On the other hand, bar charts are less useful for understanding parts of a whole. The restaurant owner who recorded the entrée selections likely will want to compare categories and understand how each category contributes to the whole. Therefore, that person might use both a bar chart and a pie or doughnut chart to visualize the collected data.



spreadsheet solution

Bar, Pie, and Doughnut Charts

Chapter 2 Bar, **Chapter 2 Pie**, and **Chapter 2 Doughnut** present the preceding bar, pie, and doughnut charts, respectively. Experiment with each chart by entering your own values in column B of each worksheet that contains a chart.

Best Practices

Sort your summary table data by the values in the second column before you create a chart. This will enable you to create a chart that fosters comparisons. For a bar chart, arrange values from smallest to largest value if you want the longest bar to appear at the top of the chart; otherwise, sort the values from largest to smallest.

Reformat charts created by software to eliminate unwanted gridlines and legends or to change the text font and size of titles and axis labels.

How-Tos

Chart Tip CT1 (see Appendix D) explains how to sort data in a summary table.

Chart Tip CT2 lists common chart-reformatting commands.

Chart Tip CT3 lists the general steps for creating charts.

The Pareto Chart

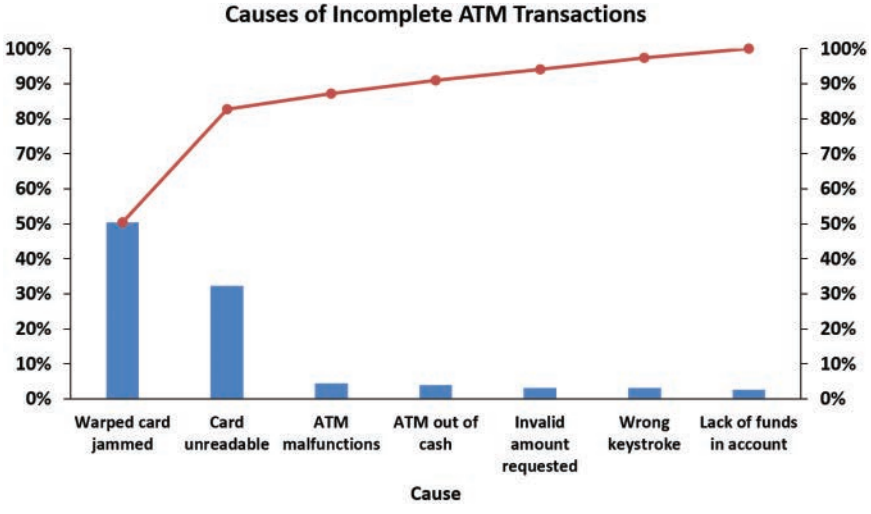
CONCEPT A special type of bar chart that presents the counts, amounts, or percentages of the categories, in descending order left to right, and also contains a superimposed plotted line that represents a running cumulative percentage.

EXAMPLE

Causes of Incomplete ATM Transactions

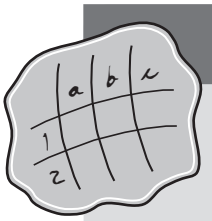
Cause	Frequency	Percentage
ATM malfunctions	32	4.42%
ATM out of cash	28	3.87%
Invalid amount requested	23	3.18%
Lack of funds in account	19	2.62%
Card unreadable	234	32.32%
Warped card jammed	365	50.41%
Wrong keystroke	23	3.18%
Total	724	100.00%

Source: Data extracted from A. Bhalla, "Don't Misuse the Pareto Principle," *Six Sigma Forum Magazine*, May 2009, pp. 15–18.



This Pareto chart uses the data of the table that immediately precedes it to highlight the causes of incomplete ATM transactions.

INTERPRETATION When you have many categories, a Pareto chart enables you to focus on the most important categories by visually separating the *vital few* from the *trivial many* categories. For the incomplete ATM transactions data, the Pareto chart shows that two categories, warped card jammed and card unreadable, account for more than 80% of all defects and that those two categories combined with the ATM malfunctions and ATM out of cash categories account for more than 90% of all defects.



spreadsheet solution

Pareto Charts

Chapter 2 Pareto contains an example of a Pareto chart. Experiment with this chart by typing your own set of values—in descending order—in column B, rows 2 through 11. (Do not alter the entries in row 12 or columns C and D.)

How-To

Chart Tip CT4 (see Appendix D) summarizes how to create a Pareto chart.

Two-Way Table

CONCEPT A table that presents the counts or percentages of responses for two categorical variables. In a two-way table, the categories of one of the variables form the rows of the table, while the categories of the second variable form the columns. The last row of a two-way table contains column totals, and the last column of such a table contains the row totals. Two-way tables are also known as cross-classification or cross-tabulation tables.

EXAMPLES This two-way table tallies entrées ordered by guests during the Friday-to-Sunday weekend period by sex.

		Sex		
		Female	**Male**	**Total**
Entrée Ordered	**Beef**	64	80	144
	Chicken	53	51	104
	Fish	72	40	112
	Vegan	8	20	28
	Other	3	9	12
	Total	200	200	400

Two-way tables can be formatted to show grand total percentages or row or column percentages.

Grand Total Percentages Table

		Sex		
		Female	Male	Total
Entrée Ordered	Beef	16.00%	20.00%	36.00%
	Chicken	13.25%	12.75%	26.00%
	Fish	18.00%	10.00%	28.00%
	Vegan	2.00%	5.00%	7.00%
	Other	0.75%	2.25%	3.00%
	Total	50.00%	50.00%	100.00%

Row Percentages Table

		Sex		
		Female	Male	Total
Entrée Ordered	Beef	44.44%	55.56%	100.00%
	Chicken	50.96%	49.04%	100.00%
	Fish	64.29%	35.71%	100.00%
	Vegan	28.57%	71.43%	100.00%
	Other	25.00%	75.00%	100.00%
	Total	50.00%	50.00%	100.00%

Column Percentages Table

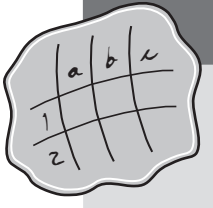
		Sex		
		Female	Male	Total
Entrée Ordered	Beef	32.00%	40.00%	36.00%
	Chicken	26.50%	25.50%	26.00%
	Fish	36.00%	20.00%	28.00%
	Vegan	4.00%	10.00%	7.00%
	Other	1.50%	4.50%	3.00%
	Total	100.00%	100.00%	100.00%

INTERPRETATION The simplest two-way table contains a row variable that has two categories and a column variable that has two categories. This creates a table that has two rows and two columns in its inner part (see the table on the next page). Each inner cell represents the count or percentage of a pairing, or cross-classifying, of categories from each variable.

		Column Variable		
		First Column Category	**Second Column Category**	**Total**
Row Variable	**First Row Category**	Count or percentage for first row and first column categories	Count or percentage for first row and second column categories	Total for first row category
	Second Row Category	Count or percentage for second row and first column categories	Count or percentage for second row and second column categories	Total for second row category
	Total	Total for first column category	Total for second column category	Overall total

Two-way tables reveal the combination of values that occurs most often in data. In the example, the tables reveal that males are more likely to order beef than females and that females are more likely to order fish.

PivotTables create worksheet summary tables from sample data and provide a good way of creating two-way tables from sample data. Advanced Technique AT1 in Appendix E discusses how to create such tables.



spreadsheet solution

Two-Way Tables

Chapter 2 Two-Way contains the counts of the download and call-to-action button variables as a simple two-way table.

Chapter 2 Two-Way PivotTable contains the counts of the entrée ordered and sex variables summarized in a two-way table that is an Excel PivotTable as well as PivotTables formatted to show grand total, row, and column percentage.

How-To

Advanced Technique ADV1 in Appendix E summarizes how to create a two-way table that is a PivotTable.

2.2 Presenting Numerical Variables

You present numerical variables by first establishing groups that represent separate ranges of values and then placing each value into the proper group. Then you create tables that summarize the groups by frequency (count) or percentage and use the table as the basis for creating charts such as a histogram, which this chapter explains.

The Frequency and Percentage Distribution

CONCEPT A table of grouped numerical data that contains the names of each group in the first column, the counts (frequencies) of each group in the second column, and the percentages of each group in the third column. This table can also appear as a two-column table that shows either the frequencies or the percentages.

EXAMPLE Consider the following data table, which presents the average ticket cost (in U.S. $) for each NBA team during a recent season.

NBA Ticket Cost

Team	Average Ticket Cost	Team	Average Ticket Cost
Atlanta	143	Miami	187
Boston	234	Milwaukee	153
Brooklyn	212	Minnesota	107
Charlotte	89	New Orleans	48
Chicago	251	New York	285
Cleveland	135	Oklahoma City	199
Dallas	124	Orlando	127
Denver	152	Philadelphia	197
Detroit	135	Phoenix	61
Golden State	463	Portland	119
Houston	177	Sacramento	198
Indiana	130	San Antonio	195
L.A. Clippers	137	Toronto	180
L.A. Lakers	444	Utah	78
Memphis	104	Washington	138

Source: Data extracted from "The Most Expensive NBA Teams to See Live," https://bit.ly/3rvSAah.

The following frequency and percentage distribution summarizes these data using 10 groupings from 0 to under 50 to 450 to under 500.

Average Ticket Cost	Frequency	Percentage
0 to under 50	1	3.33%
50 to under 100	3	10.00%
100 to under 150	11	36.67%
150 to under 200	9	30.00%
200 to under 250	2	6.67%
250 to under 300	2	6.67%
300 to under 350	0	0%
350 to under 400	0	0%
400 to under 450	1	3.33%
450 to under 500	1	3.33%
	30	100.00%

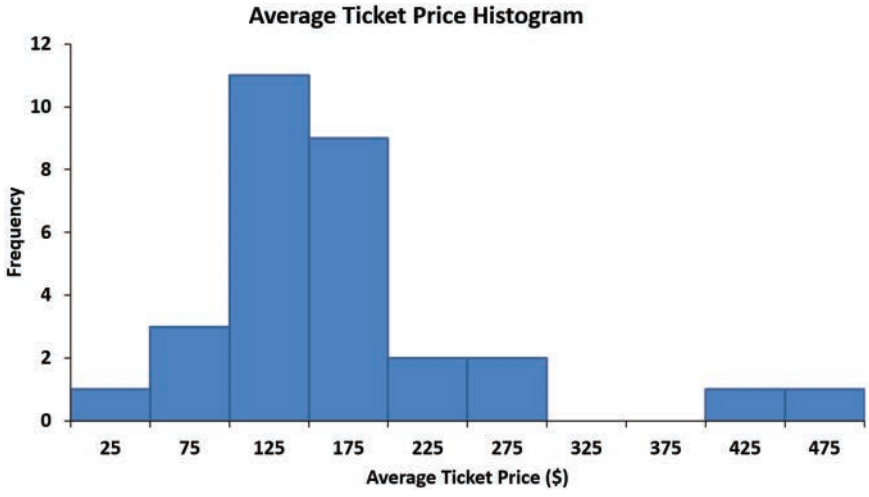
INTERPRETATION Frequency and percentage distributions enable you to quickly determine differences among the many groups of values. In this example, you can quickly see that most of the average ticket costs are between $100 and $300 and that very few average ticket costs are either below $50 or above $200.

You need to be careful in forming distribution groups because the ranges of the groups affect how you perceive the data. For example, had you grouped the average ticket costs into only two groups, below $150 and $150 and above, you would not be able to see any pattern in the data.

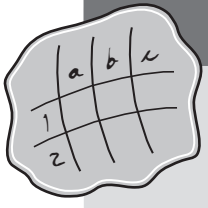
Histogram

CONCEPT A special bar chart for grouped numerical data in which the groups are represented as individual bars on the horizontal *X* axis and the frequencies or percentages for each group are plotted on the vertical *Y* axis. In a histogram, in contrast to a bar chart of categorical data, no gaps exist between adjacent bars.

EXAMPLE The following histogram presents the average ticket cost data of the preceding example. The value below each bar (25, 75, 125, 175, 225, 275, 325, 375, 425, and 475) is the **midpoint**—the approximate middle value for the group the bar represents. As with the frequency and percentage distributions, you can quickly see that very few average ticket prices are above $275.



INTERPRETATION A histogram reveals the overall shape of the frequencies in the groups. A histogram is considered symmetric if each side of the chart is an approximate mirror image of the other side. The histogram of this example has more values in the lower portion than in the upper portion, so it is considered to be non-symmetric, or *skewed*.



spreadsheet solution

Frequency Distributions and Histograms

Chapter 2 Histogram contains a frequency distribution and histogram for the average ticket cost (in U.S. $) for each NBA team during a recent season. Experiment with this chart by entering different values in column B, rows 3 through 12 of the Histogram worksheet.

How-Tos

Advanced Technique ADV2 in Appendix E and Chart Tip CT5 in Appendix D discuss how you can create frequency distributions and histograms.

The Time-Series Plot

CONCEPT A chart in which each point represents the value of a numerical variable at a specific time. By convention, the *X* axis (the horizontal axis) always represents units of time, and the *Y* axis (the vertical axis) always represents units of the variable.

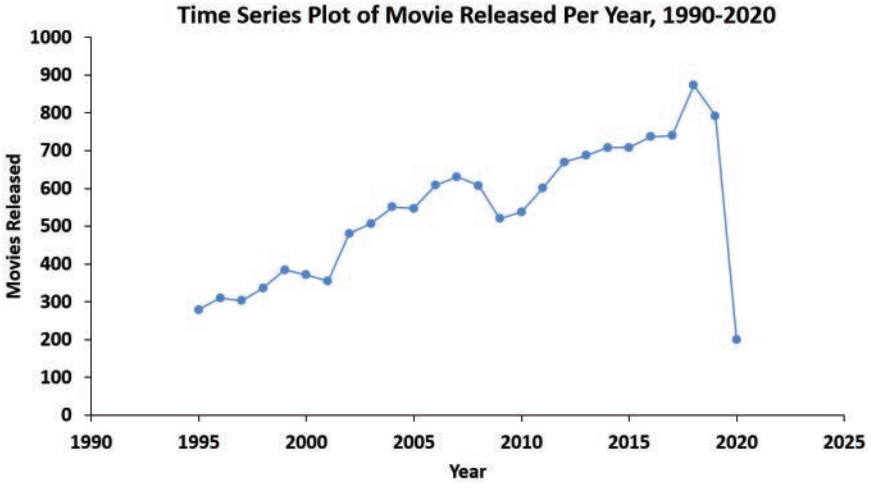
EXAMPLE Consider the following data table, which presents the number of domestic movie releases from 1990 to 2020.

Movie Releases

Year	Movies Released	Year	Movies Released
1990	224	2006	608
1991	244	2007	631
1992	234	2008	607
1993	258	2009	520
1994	254	2010	538
1995	279	2011	601
1996	310	2012	669
1997	303	2013	687
1998	336	2014	708
1999	384	2015	708
2000	371	2016	737
2001	355	2017	740
2002	480	2018	873
2003	507	2019	792
2004	551	2020	200
2005	547		

Source: Data extracted from "Domestic Yearly Box Office," https://www.boxofficemojo.com/year/.

The following time-series plot visualizes these data.



INTERPRETATION Time-series plots can reveal patterns over time—patterns that you might not see when looking at a long list of numerical values. In this example, the plot reveals that, overall, there was a general increase in the number of movies released between 1990 and 2019. Before the steep drop in 2020 caused by the COVID-19 pandemic, the number of movies released in the preceding 30 years had increased fourfold.

The Scatter Plot

CONCEPT A chart that plots the values of two numerical variables for each observation. In a scatter plot, the *X* axis (the horizontal axis) always represents units of one variable, and the *Y* axis (the vertical axis) always represents units of the second variable.

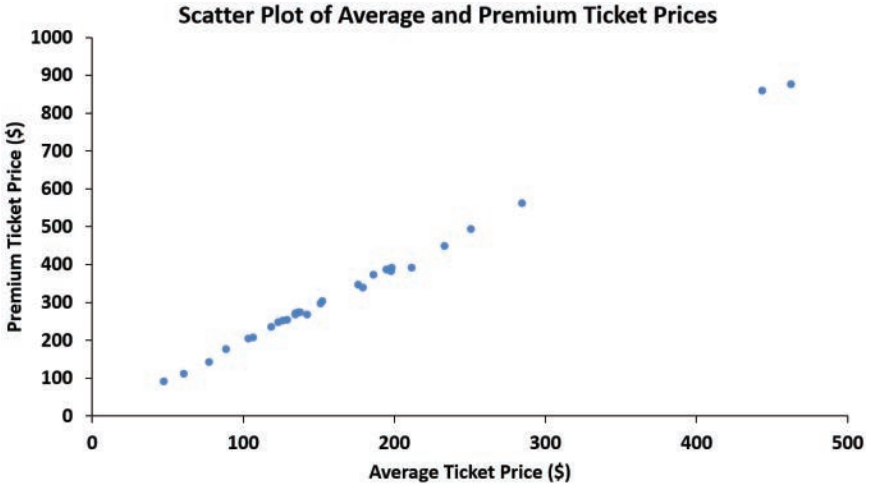
EXAMPLE Consider the following data table, which presents the average ticket cost (in U.S. $) and the premium ticket cost (in U.S. $) for each NBA team during a recent season.

NBA Ticket Cost

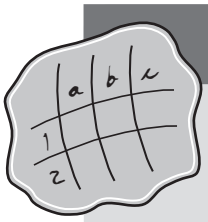
Team	Average Ticket Cost	Premium Ticket Cost
Atlanta	143	267
Boston	234	448
Brooklyn	212	391
Charlotte	89	173
Chicago	251	493
Cleveland	135	268
Dallas	124	245
Denver	152	296
Detroit	135	266
Golden State	463	874
Houston	177	346
Indiana	130	252
L.A. Clippers	137	271
L.A. Lakers	444	857
Memphis	104	203
Miami	187	371
Milwaukee	153	301
Minnesota	107	204
New Orleans	48	89
New York	285	561
Oklahoma City	199	390
Orlando	127	249

Team	Average Ticket Cost	Premium Ticket Cost
Philadelphia	197	383
Phoenix	61	110
Portland	119	233
Sacramento	198	380
San Antonio	195	384
Toronto	180	338
Utah	78	142
Washington	138	271

The following scatter plot visualizes these data.



INTERPRETATION A scatter plot helps reveal patterns in the relationship between two numerical variables. The scatter plot for these data reveals a strong positive linear (straight-line) relationship between the average ticket cost and the cost of a premium ticket. Based on this relationship, you can conclude that the average ticket cost is a useful predictor of the premium ticket cost. (Chapter 10 more fully discusses using one numerical variable to predict the value of another numerical variable.)



spreadsheet solution

Time-Series and Scatter Plots

Chapter 2 Time-Series contains the time-series plot for the domestic movie releases from 1990 to 2020. Experiment with this plot by entering different values in column B, rows 2 through 32.

Chapter 2 Scatter Plot contains the scatter plot for the NBA ticket cost data. Experiment with this scatter plot by entering different data values in columns B and C, rows 2 through 31.

How-Tos

Chart Tip CT6 (in Appendix D) discusses how you can create time-series plots.

Chart Tip CT7 (in Appendix D) discusses how you can create scatter plots.

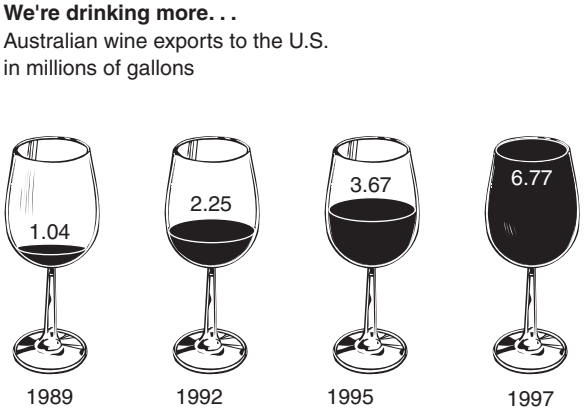
2.3 "Bad" Charts

So-called "good" charts, such as the charts presented so far in this chapter, help visualize data in ways that aid understanding. However, in the modern world, you can easily find examples of "bad" charts that obscure or confuse the data. Such charts include elements or practices known to impede understanding or fail to apply properly the techniques that this chapter discusses.

CONCEPT A "bad" chart fails to clearly present data in a useful and undistorted manner.

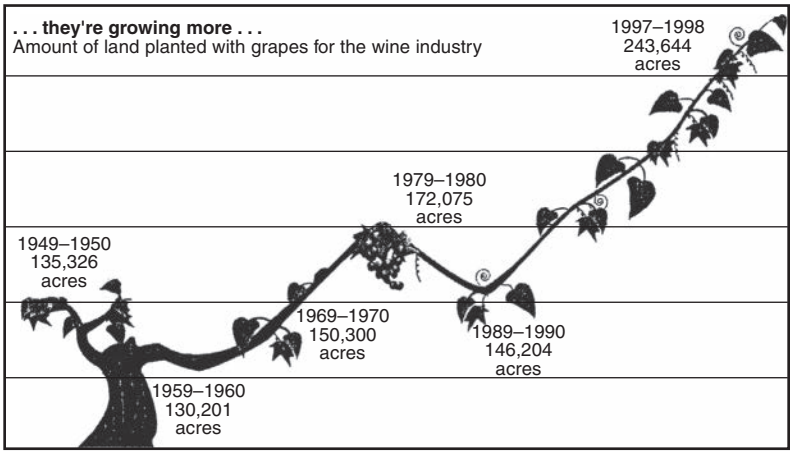
INTERPRETATION Using pictorial symbols obscures the data and can create a false impression in the mind of the reader, especially if the pictorial symbols are representations of three-dimensional objects. In Example 1, the wine glasses fail to reflect that the 1992 data (2.25 million gallons) is a bit more than twice the 1.04 million gallons for 1989. In addition, the spaces between the wine glasses falsely suggest equal-sized time periods and obscure the trend in wine exports. (Hint: Plot the data as a time-series chart to discover the actual trend.)

EXAMPLE 1: Australian Wine Exports to the United States.



Example 2 combines the inaccuracy of using a picture (grape vine) with the error of having unlabeled and improperly scaled axes. A missing *X* axis prevents the reader from immediately seeing that the 1997–1998 value is misplaced. By the scale of the graph, that data point should be closer to the rest of the data. A missing *Y* axis prevents the reader from getting a better sense of the rate of change in land planted through the years. Other problems also exist. Can you spot at least one more? (Hint: Compare the 1949–1950 data to the 1969–1970 data.)

EXAMPLE 2: Amount of Land Planted with Grapes for the Wine Industry.



When producing your own charts, use these guidelines:

- Always choose the simplest chart that can present your data.
- Always supply a title.
- Always label every axis.
- Avoid unnecessary decorations or illustrations around the borders or in the background.
- Avoid the use of fancy pictorial symbols to represent data values.
- Avoid 3D versions of bar and pie charts.
- If the chart contains axes, always include a scale for each axis.
- When charting non-negative values, the scale on the vertical axis should begin at zero.

One-Minute Summary

To choose an appropriate table or chart type, begin by determining whether your data are categorical or numerical.

If your data are categorical:

- Determine whether you are presenting one or two variables.
- If one variable, use a summary table, bar chart, pie chart, or doughnut chart. If emphasizing the *vital few* from the *trivial many*, use a Pareto chart.
- If two variables, use a two-way table.

If your data are numerical:

- If charting one variable, use a frequency and percentage distribution with or without a histogram.
- If charting two variables, if the time order of the data is important, use a time-series plot; otherwise, use a scatter plot.

Test Yourself

Short Answers

1. Which of the following graphical presentations is not appropriate for categorical data?
   a. Pareto chart
   b. scatter plot

c. bar chart
d. pie chart

2. Which of the following graphical presentations is not appropriate for numerical data?

a. histogram
b. pie chart
c. time-series plot
d. scatter plot

3. A type of histogram in which the categories are plotted in the descending rank order of the magnitude of their frequencies is called a:

a. bar chart
b. pie chart
c. scatter plot
d. Pareto chart

4. Which of the following would best show that the total of all the categories sums to 100%?

a. pie chart
b. histogram
c. scatter plot
d. time-series plot

5. The basic principle behind the _____ is the capability to separate the vital few categories from the trivial many categories.

a. scatter plot
b. bar chart
c. Pareto chart
d. pie chart

6. When studying the simultaneous responses to two categorical variables, you should construct a:

a. histogram
b. pie chart
c. scatter plot
d. cross-classification table

7. In a cross-classification table, the number of rows and columns:

a. must always be the same
b. must always be two
c. must add to 100%
d. None of the above.

Answer True or False:

8. Histograms are used for numerical data, whereas bar charts are suitable for categorical data.
9. A website monitors customer complaints and organizes these complaints into six distinct categories. Over the past year, the company has received 534 complaints. One possible graphical method for representing these data is a Pareto chart.
10. A website monitors customer complaints and organizes these complaints into six distinct categories. Over the past year, the company has received 534 complaints. One possible graphical method for representing these data is a scatter plot.
11. A social media website collected information on the age of its customers. The youngest customer was 5, and the oldest was 96. To study the distribution of the age of its customers, the company should use a pie chart.
12. A social media website collected information on the age of its customers. The youngest customer was 5, and the oldest was 96. To study the distribution of the age of its customers, the company can use a histogram.
13. A website wants to collect information on the daily number of visitors. To study the daily number of visitors, it can use a pie chart.
14. A website wants to collect information on the daily number of visitors. To study the daily number of visitors, it can use a time-series plot.
15. A professor wants to study the relationship between the number of hours a student studied for an exam and the exam score achieved. The professor can use a time-series plot.
16. A professor wants to study the relationship between the number of hours a student studied for an exam and the exam score achieved. The professor can use a bar chart.
17. A professor wants to study the relationship between the number of hours a student studied for an exam and the exam score achieved. The professor can use a scatter plot.
18. If you wanted to compare the percentage of items that are in a particular category as compared to other categories, you should use a pie chart, not a bar chart.

Fill in the Blank:

19. To evaluate two categorical variables at the same time, a _____ should be developed.
20. A _____ is a vertical bar chart in which the rectangular bars are constructed at the boundaries of each class interval.
21. A _____ chart should be used when you are primarily concerned with the percentage of the total that is in each category.
22. A _____ chart should be used when you are primarily concerned with comparing the percentages in different categories.

23. A _____ should be used when you are studying a pattern between two numerical variables.
24. A _____ should be used to study the distribution of a numerical variable.
25. You have measured your pulse rate daily for 30 days. A _____ plot should be used to study the pulse rate for the 30 days.
26. You have collected data from your friends concerning their favorite soft drink. You should use a _____ chart to study the favorite soft drink of your friends.
27. You have collected data from your friends concerning the time it takes to get ready to leave their house in the morning. You should use a _____ to study this variable.

Answers to Test Yourself Short Answers

1. b
2. b
3. d
4. a
5. c
6. d
7. d
8. True
9. True
10. False
11. False
12. True
13. False
14. True
15. False
16. False
17. True
18. False
19. two-way table
20. histogram
21. pie chart
22. bar chart
23. scatter plot
24. histogram
25. time-series plot
26. bar chart, pie chart, or Pareto chart
27. histogram

Problems

1. A Pew Research Center survey studied the key issues for employed adults who have been working at home some or all of the time. The following three summary tables present the results of that survey.

Feeling Motivated to Do Their Work	Percentage
Very Difficult	7%
Somewhat Difficult	29%
Somewhat Easy	31%
Easy	34%

Doing Work Without Interruptions	Percentage
Very Difficult	8%
Somewhat Difficult	24%
Somewhat Easy	37%
Easy	31%

Having an Adequate Workspace	Percentage
Very Difficult	4%
Somewhat Difficult	19%
Somewhat Easy	31%
Easy	47%

For each table

a. Construct a bar chart and a pie or doughnut chart.
b. Which graphical method do you think best presents these data?
c. What conclusions can you reach concerning how employed adults who have been working at home some or all of the time feel about being motivated to do their work?
d. What conclusions can you reach concerning how employed adults who have been working at home some or all of the time feel about doing work without interruptions?
e. What conclusions can you reach concerning how employed adults who have been working at home some or all of the time feel about having an adequate workspace?
f. What differences in the responses among the three issues exist?

2. Market researchers for a telecommunications company have summarized data collected about the payment methods customers use in the following summary table.

Payment Method	Frequency
Bank transfer (automatic)	1,212
Credit card (automatic)	1,191
Electronic check	2,243
Mailed check	871
Total	5,517

a. Using this table construct a bar chart and a pie or doughnut chart.
b. Which graphical method do you think best presents these data?
c. What conclusions can you reach about customer payment methods?

3. Medication errors are a serious problem in hospitals. The following summary table presents the root causes of pharmacy errors at a hospital during a recent time period.

Reason for Failure	Frequency
Additional instructions	16
Dose	23
Drug	14
Duplicate order entry	22
Frequency	47
Omission	21
Order not discontinued when received	12
Order not received	52
Patient	5
Route	4
Other	8

a. Construct a Pareto chart for these data.
b. Discuss the "vital few" and "trivial many" reasons for the root causes of pharmacy errors.

4. Students who attend a regional university located in a small town are known to favor the local independent pizza restaurant. A national chain of pizza restaurants looks to open a store in that town and conducts a survey of students who attend that university to determine pizza preferences. The following two-way table summarizes the survey variables store type and sex, based on the responses of a sample of 220 students.

		Sex	
		Female	Male
Store Type	Local	74	71
	National	19	56

a. Construct a two-way table that displays grand total percentages.
b. Construct a two-way table that displays row percentages.
c. Construct a two-way table that displays column percentages.

d. What conclusions can you reach from the tables constructed in parts (a) through (c)?

e. Which table do you think is most useful in reaching the conclusions in your part (d) answer?

5. Churning, the loss of customers to a competitor, is a problem for all companies, especially telecommunications companies. Market researchers for a telecommunications company collect data from 5,517 customers of the company. Data collected for each customer includes whether the customer churned during the last month, the sex of the customer, whether the customer is a senior citizen, and whether the customer uses paperless billing. The following three summary tables summarize these survey variables.

		Churn	
		No	Yes
Sex	Female	1,858	883
	Male	1,903	873

		Churn	
		No	Yes
Senior Citizen	No	3,142	1,285
	Yes	619	471

		Churn	
		No	Yes
Paperless Billing	No	1,394	398
	Yes	2,367	1,358

For each table

a. Construct a two-way table that displays grand total percentages

b. Construct a two-way table that displays row percentages.

c. Construct a two-way table that displays column percentages.

d. What conclusions can you reach from the tables constructed in parts (a) through (c)?

e. Which table do you think is most useful in reaching the conclusions in your part (d) answer?

Domestic Beer

6. The file **Domestic Beer** contains the percentage alcohol, number of calories per 12 ounces, and number of carbohydrates (in grams) per 12 ounces for 157 of the best-selling domestic beers in the United States.

(Data extracted from "Find Out How Many Calories in Beer?" https://www.beer100.com/beer-calories.)

a. Construct a frequency distribution and a percentage distribution for percentage alcohol, number of calories per 12 ounces, and number of carbohydrates per 12 ounces (in grams).

b. Construct a histogram for percentage alcohol, number of calories per 12 ounces, and number of carbohydrates per 12 ounces (in grams).

c. Construct three scatter plots: percentage alcohol versus calories, percentage alcohol versus carbohydrates, and calories versus carbohydrates.

d. What conclusions can you reach about the percentage alcohol, number of calories per 12 ounces, and number of carbohydrates per 12 ounces (in grams)?

Super Bowl Ads

7. The **Super Bowl Ads** file contains the average ratings of 57 ads from the 2021 NFL Super Bowl broadcast. (Data extracted from T. Schad, "Rocket mortgage ads dominate Ad Meter," *USA Today*, February 9, 2021, p. 4B.)

a. Construct a histogram based on these data.

b. What conclusions can you reach concerning Super Bowl ad ratings?

8. The **Big Mac Starbucks** file contains the cost (in U.S. $) of a McDonald's Big Mac sandwich and a Starbucks tall latte in 11 world cities.

Big Mac Starbucks

City	Big Mac	Starbucks Tall Latte
Moscow	2.29	4.35
Johannesburg	2.53	2.18
Hong Kong	2.87	4.60
Bangkok	3.85	2.60
Dubai	4.08	4.29
Buenos Aires	4.22	2.14
London	4.32	3.58
New York	5.09	4.30
Paris	5.37	4.30
Toronto	4.38	3.15
Zurich	6.89	5.94

Source: Data extracted from "How Much a Big Mac Costs Around the World," *Business Insider*, https://businessinsider.com/mcdonalds-big-mac-price-around-the-world-2018-5, and "The Starbucks Index 2019," https://www.finder.com/starbucks.index.

a. Construct a scatter plot from these data.

b. What conclusions can you reach about the relationship between the cost of a McDonald's Big Mac and a Starbucks tall latte in these 11 world cities?

9. The **Potter Movies** file contains the first weekend gross (in $millions) and the total domestic gross (in $millions) for the eight movies in the Harry Potter film series.

Potter Movies

Title	First Weekend	Total Domestic
Sorcerer's Stone	90.295	317.871
Chamber of Secrets	88.357	262.233
Prisoner of Azkaban	93.687	249.758
Goblet of Fire	102.335	290.201
Order of the Phoenix	77.108	292.137
Half-Blood Prince	77.836	302.089
Deathly Hallows Part I	125.017	296.132
Deathly Hallows Part II	169.189	381.193

Source: Data extracted from "Box Office History for Harry Potter Movies," https://www.the-numbers.com/movies/franchise/Harry-Potter.

a. Construct a scatter plot from these data.

b. What conclusion can you reach about the relationship between the first weekend and total domestic grosses?

10. The **UHDTV Wholesale Sales** file contains the U.S. wholesale sales of Ultra HDTVs (in $millions) from 2013 to 2019.

UHDTV Wholesale Sales

Year	Wholesale Sales
2013	310
2014	2,238
2015	7,673
2016	12,932
2017	13,400
2018	14,300
2019	14,900

Source: Data extracted from "4K Ultra HD TVs wholesale sales revenue in the United States from 2013 to 2019," https://www.statista.com/statistics/643511/4k-ultra-hdtv-wholesale-sales-in-us/.

a. Construct a time-series plot of the U.S. Ultra HDTV wholesale sales from 2013 to 2019.

b. What pattern does the plot reveal?

c. If you were asked to predict U.S. Ultra HDTV wholesale sales for 2020, what would you predict?

11. The **MLB Salaries** file contains the average MLB baseball player salaries (in $millions) for the years 2003 through 2020.

MLB Salaries

Year	Average MLB Salary	Year	Average MLB Salary
2003	2.37	2012	3.21
2004	2.31	2013	3.39
2005	2.48	2014	3.69
2006	2.70	2015	3.84
2007	2.82	2016	4.38
2008	2.93	2017	4.45
2009	3.00	2018	4.41
2010	3.01	2019	4.80
2011	3.10	2020	4.43

Source: Data extracted from https://statista.com/statistics/23621/mean-salary-of-players-in-major-league-baseball (no longer available).

a. Construct a time-series plot of the average MLB baseball player salaries for the years 2003 through 2020.

b. What pattern does the plot reveal?

c. If you were asked to predict the average MLB baseball player salary for 2021, what would you predict?

Answers to Test Yourself Problems

1. b. If you are more interested in determining which category of feeling motivated to do their job response occurs most often, then the bar chart is preferred. If you are more interested in seeing the distribution of the entire set of categories, then either the pie chart or the doughnut chart is preferred.

c. Respondents are about equally likely to feel that it is easy, somewhat easy, or somewhat difficult to feel motivated to do their job.

d. Respondents are about equally likely to feel that it is somewhat easy or somewhat difficult to do work without interruption.

e. Respondents are most likely to feel that it is easy to have adequate workspace.

f. They feel that it is easier to have adequate workspace than to feel motivated to do work or to work without interruption.

2. b. If you are more interested in determining which category of payment method used occurs most often, then the bar chart is preferred. If you are more interested in seeing the distribution of the entire set of categories, either the pie chart or doughnut chart is preferred.
   c. Respondents are most likely to pay by electronic check and least likely to pay by mailed check.
3. b. The most important categories of medication errors are orders not received and frequency followed by dose, duplicate order entry, and omission.
4. a. through c.

		Sex		
		Female	Male	Grand Total
Store Type	Local	33.64%	32.27%	65.91%
	National	8.64%	25.45%	34.09%
	Grand Total	42.28%	57.72%	100.00%

		Sex		
		Female	Male	Grand Total
Store Type	Local	51.03%	48.97%	100.00%
	National	25.33%	74.67%	100.00%
	Grand Total	42.27%	57.73%	100.00%

		Sex		
		Female	Male	Grand Total
Store Type	Local	79.57%	55.91%	65.91%
	National	20.43%	44.09%	34.09%
	Grand Total	100.00%	100.00%	100.00%

5. a. through c.
Sex and Churn

		Churn		
		No	Yes	Grand Total
Sex	Female	33.68%	16.01%	49.69%
	Male	34.49%	15.82%	50.31%
	Grand Total	68.17%	31.83%	100.00%

		Churn		
		No	Yes	Grand Total
Sex	Female	67.79%	32.21%	100.00%
	Male	68.55%	31.45%	100.00%
	Grand Total	68.17%	31.83%	100.00%

		Churn		
		No	Yes	Grand Total
Sex	Female	51.21%	50.59%	49.68%
	Male	48.79%	49.41%	50.32%
	Grand Total	100.00%	100.00%	100.00%

d. There is very little difference between males and females in churning.

e. Row percentages are more valuable because this table compares males and females.

Senior Citizen and Churn

		Churn		
		No	Yes	Grand Total
Senior Citizen	No	56.95%	23.29%	79.24%
	Yes	11.22%	8.54%	19.76%
	Grand Total	68.17%	31.83%	100.00%

		Churn		
		No	Yes	Grand Total
Senior Citizen	No	70.97%	29.03%	100.00%
	Yes	56.79%	43.21%	100.00%
	Grand Total	68.17%	31.83%	100.00%

		Churn		
		No	Yes	Grand Total
Senior Citizen	No	83.54%	73.17%	80.24%
	Yes	16.46%	26.83%	19.76%
	Grand Total	100.00%	100.00%	100.00%

d. Senior citizens are much less likely to churn.

e. Row percentages are more valuable because this table compares senior citizens and non-senior citizens.

Paperless Billing and Churn

		Churn		
		No	Yes	Grand Total
Paperless Billing	No	25.27%	7.21%	32.48%
	Yes	42.90%	24.61%	67.51%
	Grand Total	68.17%	31.62%	100.00%

		Churn		
		No	Yes	Grand Total
Paperless Billing	No	77.79%	22.21%	100.00%
	Yes	63.54%	36.46%	100.00%
	Grand Total	68.17%	31.83%	100.00%

		Churn		
		No	Yes	Grand Total
Paperless Billing	No	37.06%	22.67%	32.48%
	Yes	62.94%	77.33%	67.52%
	Grand Total	100.00%	100.00%	100.00%

d. Those who use paperless billing are more likely to churn than those who do not use paperless billing.

e. Row percentages are more valuable because this table best helps to compare those with and without paperless billing.

6. c. The alcohol percentage is concentrated between 4% and 6%, with more between 4% and 5%. The calories are concentrated between 140 and 160. The carbohydrates are concentrated between 12 and 15. There are outliers in the percentage of alcohol in both tails. The outlier in the lower tail is due to the nonalcoholic beer O'Doul's. The outlier in the upper tail is around 11.5%. A few beers have high calorie counts near 330 and carbohydrates as high as 32. A strong positive relationship exists between percentage of alcohol and calories and between calories and carbohydrates, and there is a moderately positive relationship between percentage alcohol and carbohydrates.

7. b. The ad ratings are fairly symmetrical, with many of the ad scores between 5 and 6. Very few ratings are below 4.5 or above 7.
8. b. There is a weak relationship between the cost of a McDonald's Big Mac and the cost of a Starbucks tall latte in various cities.
9. b. There is a moderately positive relationship between the U.S. gross and the first weekend gross for Harry Potter movies.
10. b. Ultra HDTV sales rose dramatically from 2013 to 2016 but leveled off after that.
    c. Somewhere between 15 and 16 million.
11. b. There has been a very strong linear increase in the salaries.
    c. Because there was a decrease in 2020, the prediction is that the average salary in 2021 will be less than $5 million.

References

1. Beninger, J. M., and D. L. Robyn. 1978. "Quantitative Graphics in Statistics," *The American Statistician*, 32: 1–11.
2. Berenson, M. L., D. M. Levine, K. A. Szabat, and D. F. Stephan. *Basic Business Statistics: Concepts and Applications*, 15th edition. Hoboken, NJ: Pearson Education, 2023.
3. Levine, D., D. Stephan, and K. Szabat. *Statistics for Managers Using Microsoft Excel*, 9th edition. Boston: Pearson Education, 2021.
4. Tufte, E. R. *The Visual Display of Quantitative Information*, 2nd edition. Cheshire, CT: Graphics Press, 2002.
5. Tufte, E. R. *Visual Explanations*. Cheshire, CT: Graphics Press, 1997.

Index

Numerics

A

B

C

D

E

F-G

H

I

J-K-L

M

N

O

P

Q-R

S

T

U-V

W

X-Y-Z