Data Analysis For Business Decisions

A Laboratory Manual

Second Edition

Andres Fortino

Data Analysis for Business Decisions

LICENSE, DISCLAIMER OF LIABILITY, AND LIMITED WARRANTY

By purchasing or using this book and its companion files (the "Work"), you agree that this license grants permission to use the contents contained herein, but does not give you the right of ownership to any of the textual content in the book or ownership to any of the information, files, or products contained in it. *This license does not permit uploading of the Work onto the Internet or on a network (of any kind) without the written consent of the Publisher.* Duplication or dissemination of any text, code, simulations, images, etc. contained herein is limited to and subject to licensing terms for the respective products, and permission must be obtained from the Publisher or the owner of the content, etc., in order to reproduce or network any portion of the textual material (in any media) that is contained in the Work.

Mercury Learning and Information ("MLI" or "the Publisher") and anyone involved in the creation, writing, production, accompanying algorithms, code, or computer programs ("the software"), and any accompanying Web site or software of the Work, cannot and do not warrant the performance or results that might be obtained by using the contents of the Work. The author, developers, and the Publisher have used their best efforts to insure the accuracy and functionality of the textual material and/or programs contained in this package; we, however, make no warranty of any kind, express or implied, regarding the performance of these contents or programs. The Work is sold "as is" without warranty (except for defective materials used in manufacturing the book or due to faulty workmanship).

The author, developers, and the publisher of any accompanying content, and anyone involved in the composition, production, and manufacturing of this work will not be liable for damages of any kind arising out of the use of (or the inability to use) the algorithms, source code, computer programs, or textual material contained in this publication. This includes, but is not limited to, loss of revenue or profit, or other incidental, physical, or consequential damages arising out of the use of this Work.

The sole remedy in the event of a claim of any kind is expressly limited to replacement of the book and only at the discretion of the Publisher. The use of "implied warranty" and certain "exclusions" vary from state to state, and might not apply to the purchaser of this product.

Companion files also available for downloading from the publisher by writing to info@merclearning.com.

Data Analysis for Business Decisions

A Laboratory Manual

Second Edition

Andres Fortino, PhD

Mercury Learning and Information
Dulles, Virginia
Boston, Massachusetts
New Delhi

Publisher: David Pallai
Mercury Learning and Information
22841 Quicksilver Drive
Dulles, VA 20166
info@merclearning.com
www.merclearning.com
1-800-232-0223

A. Fortino. *Data Analysis for Business Decisions: A Laboratory Manual 2/E.*
ISBN: 978-1-68392-592-7

Library of Congress Control Number: 2020942356
202122321 Printed on acid-free paper in the United States of America.

Dedicated to Kathleen
for her patience and support

Contents

Preface

This laboratory manual was written for business analysts who wish to increase their skills in conducting statistical analysis of data sets to support business decision-making. Most of the exercises use Excel, today's most common analysis tool. They range from the most basic descriptive statistical techniques to more advanced techniques, such as multivariate linear regression and forecasting.

Advanced exercises cover inferential statistics for continuous variables (t-Test) and categorical variables (Chi-square), as well as A/B testing. The manual ends with techniques to deal with the analysis of text data (text data mining) and tools to manage the analysis of large data sets (Big Data) using Excel. A set of cases is provided to assist the analyst to improving their data visualization skills.

On the Companion Files

The exercises require access to the data sets used in analyzing the cases. They may be accessed on companion disc. with this book or for downloading by writing to the publisher at info@merclearning.com. A file folder *Lab Data* has all the files referenced in the exercises. A zip file *Lab Data.zip*, found in the same repository, can be downloaded to make data available on a local drive. The solution folders within each exercise folder contain some illustrative charts and tables as well as solution spreadsheets. All of the figures (including those in full color) are on the companion files for enlargement and easy-viewing. The analysis techniques presented in each chapter have short companion videos you may use to understand the ideas further. The video lessons may be found on the companion disc. If you wish to stream the video rather than download it, there is a document on the disc with links to all the companion videos to be found on a streaming service. The companion files are also available for download from the publisher by writing to info@merclearning.com.

Acknowledgments

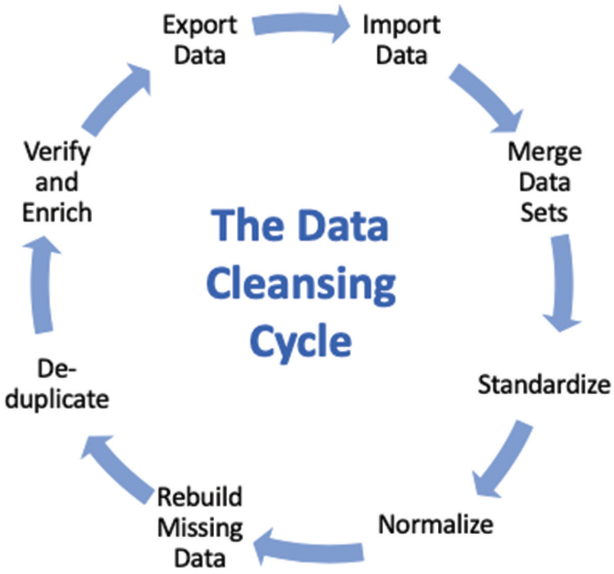
Practical books such as this, that are full of cases, are created by many years of trying them out on students until you get them right. It's a matter of keep changing the exercises and trying things out until they seem to work, and in the end, they help people learn. I wish to thank the legion of students who were very patient with me and helped me perfect these cases. Both my graduate students at the NYU School of Professional Studies and the many American Management Association professionals who attended my AMA seminars deserve my gratitude.

I also wish to thank my colleague, Nicole Morgenstern, for taking a chance with me at AMA. Thank you, Nicole, for sponsoring this work and running interference for me. My thanks to my graduate student, Karen pey-rong Hong, who did a superb job updating all the exercises to the latest version of Excel. The entire team of editors and artists at Mercury Learning was terrific and has my gratitude. A special thanks to Jim Walsh, my editor, who kept asking for more and more and helped shape an excellent book. In the end, it paid off, Jim. Finally, to my loving and patient wife, Kathleen, who not only labored over the manuscript by copyediting, but provided much-needed advice. You were always right, dear.

Dr. Andres Fortino
June 2020

CHAPTER 1

SHAPING AND CLEANING DATA



In the first set of exercises, we will look at the importance of shaping and cleaning data files. The initial image in this chapter shows the Data Cleansing Cycle with many activities, starting with importing the data; merging the data sets; standardizing and normalizing data; rebuilding missing data; de-duplicating; and last, verifying and enriching the data set. The object is to produce a data set in Excel in what is called a flat-file format. When expressed in that format, the first row of the table must contain all the variable names (with none missing); every row is of the same nature, and there are no empty rows or columns. All other rows and columns outside of the table area should be clear of data. Once in that format, the table is ready for analysis, and we can safely apply many of the Excel analytic tools.

The source of the data table varies; sometimes we extract it from a DBMS using the SQL language using queries. Other times we may obtain a comma-separated values file (with a *.CSV* extension), or a formatted text file (with a *.TXT* extension), or we may have scraped it from an HMTL formatted Web page. In Analysis Case 1.1, we explore loading and shaping data files from several sources. We study how long it takes to load data files of different sizes, including some huge files that tax the limit of Excel. They not only take a long time to load, but they are unwieldy to analyze.

Once we practice loading data in various formats, we explore cleaning it in Analysis Case 1.2. We practice using a small data file that contains several errors that need to be corrected. You are directed to the original data to find the original values. The exercise allows you to utilize many tools in Excel that make the data cleaning process efficient.

The whole process of scraping, uploading, cleaning, annotating, and shaping the data file is referred to as data wrangling. Many studies have shown that this process is not only tedious but can take up to 80% of the overall time needed to perform the analysis. But it is critical for success in the analysis. The more skilled you are in the use of cleaning and shaping tools, and the smarter you are in their use, the sooner you will start the analysis, and the less time you will need to find answers.

Analysis Case 1.1 – Shaping the Data File

1. Using the Lab Data set provided, open the *Analysis Case 1.1* folder in it, and find the file *ORDERS.csv* (1.8 MB file with 8,400 records). (The data set was made available courtesy of Tableau, Inc.)

2. Open *ORDERS.csv* using Excel.

3. Excel will automatically recognize the .CSV format and open the file with no further work on your part.

4. *ORDERS* is a comma-separated values file.

5. Save the file as *ORDERS.xlsx* in the *Analysis Case 1* folder.

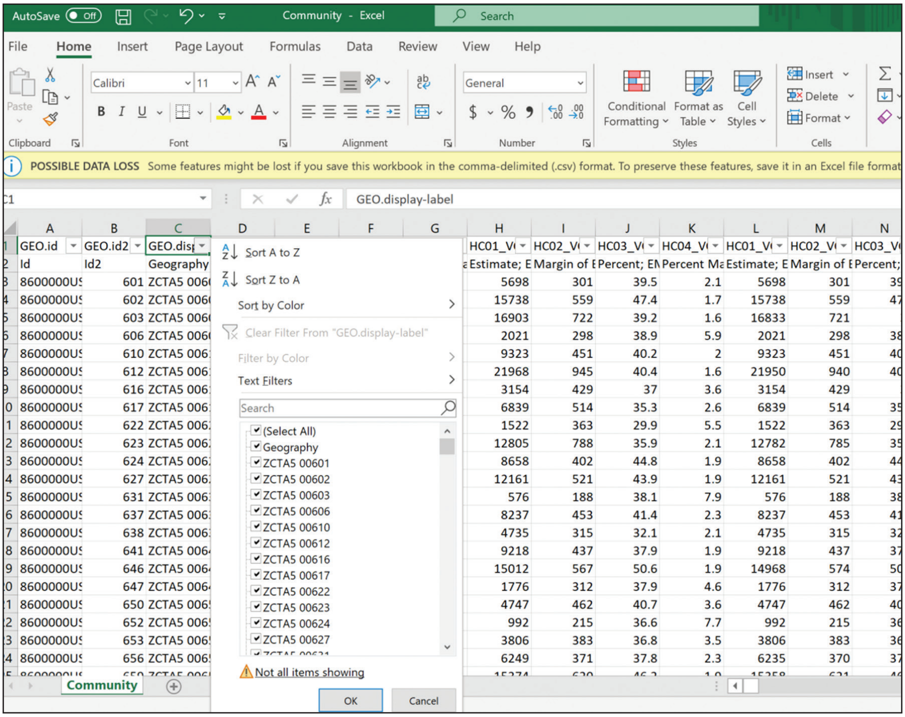
6. We will explore the difficulties with scraping, opening, and working with large data files in Excel. Consider the following four data files found in the *Analysis Case 1.1* folder (Table 1.1). Each file is progressively larger and more difficult to open in Excel than the next.

Table 1.1 Characteristics of the data files used to exemplify the lag in loading data demonstrated in this case.

Name	Size (MB)	Rows	Columns	Source	Description
ORDERS.csv	1.8	8,400	22	Company	Office supplies orders
Community.csv	70	376,000	551	US Census	2013 ACS census file
Courses.csv	73	631,139	21	MIT	edX 2013 MOOC Courses
BankComplaints.csv	306	753,324	18	US FTC	Bank complaints to the FTC

7. Using the Lab Data set and in the *Analysis Case 1.1* folder, find the file *Community.csv* (70 MB file with 376,000 records, 551 variables).

8. Note how long it takes to load. Add filters to the top row, and then filter column D to a "0" value only. Note how long it takes Excel to execute these commands due to the large number of rows in the file. (The filter function is found in the Data ribbon.) Figure 1.1 shows the *Community.csv* file opened in Excel.



POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format

C1 fx GEO.display-label

☑ (Select All)
☑ Geography
☑ ZCTA5 00601
☑ ZCTA5 00602
☑ ZCTA5 00603
☑ ZCTA5 00606
☑ ZCTA5 00610
☑ ZCTA5 00612
☑ ZCTA5 00616
☑ ZCTA5 00617
☑ ZCTA5 00622
☑ ZCTA5 00623
☑ ZCTA5 00624
☑ ZCTA5 00627

	A	B	C	H	I	J	K	L	M	N
1	GEO.id	GEO.id2	GEO.disp	HC01_V	HC02_V	HC03_V	HC04_V	HC01_V	HC02_V	HC03_V
2	Id	Id2	Geography	Estimate; E	Margin of E	Percent; EN	Percent Ma	Estimate; E	Margin of E	Percent;
3	8600000US	601	ZCTA5 006	5698	301	39.5	2.1	5698	301	39
4	8600000US	602	ZCTA5 006	15738	559	47.4	1.7	15738	559	47
5	8600000US	603	ZCTA5 006	16903	722	39.2	1.6	16833	721	
6	8600000US	606	ZCTA5 006	2021	298	38.9	5.9	2021	298	38
7	8600000US	610	ZCTA5 006	9323	451	40.2	2	9323	451	40
8	8600000US	612	ZCTA5 006	21968	945	40.4	1.6	21950	940	40
9	8600000US	616	ZCTA5 006	3154	429	37	3.6	3154	429	
10	8600000US	617	ZCTA5 006	6839	514	35.3	2.6	6839	514	35
11	8600000US	622	ZCTA5 006	1522	363	29.9	5.5	1522	363	29
12	8600000US	623	ZCTA5 006	12805	788	35.9	2.1	12782	785	35
13	8600000US	624	ZCTA5 006	8658	402	44.8	1.9	8658	402	44
14	8600000US	627	ZCTA5 006	12161	521	43.9	1.9	12161	521	43
15	8600000US	631	ZCTA5 006	576	188	38.1	7.9	576	188	38
16	8600000US	637	ZCTA5 006	8237	453	41.4	2.3	8237	453	41
17	8600000US	638	ZCTA5 006	4735	315	32.1	2.1	4735	315	32
18	8600000US	641	ZCTA5 006	9218	437	37.9	1.9	9218	437	37
19	8600000US	646	ZCTA5 006	15012	567	50.6	1.9	14968	574	50
20	8600000US	647	ZCTA5 006	1776	312	37.9	4.6	1776	312	37
21	8600000US	650	ZCTA5 006	4747	462	40.7	3.6	4747	462	40
22	8600000US	652	ZCTA5 006	992	215	36.6	7.7	992	215	36
23	8600000US	653	ZCTA5 006	3806	383	36.8	3.5	3806	383	36
24	8600000US	656	ZCTA5 006	6249	371	37.8	2.3	6235	370	37

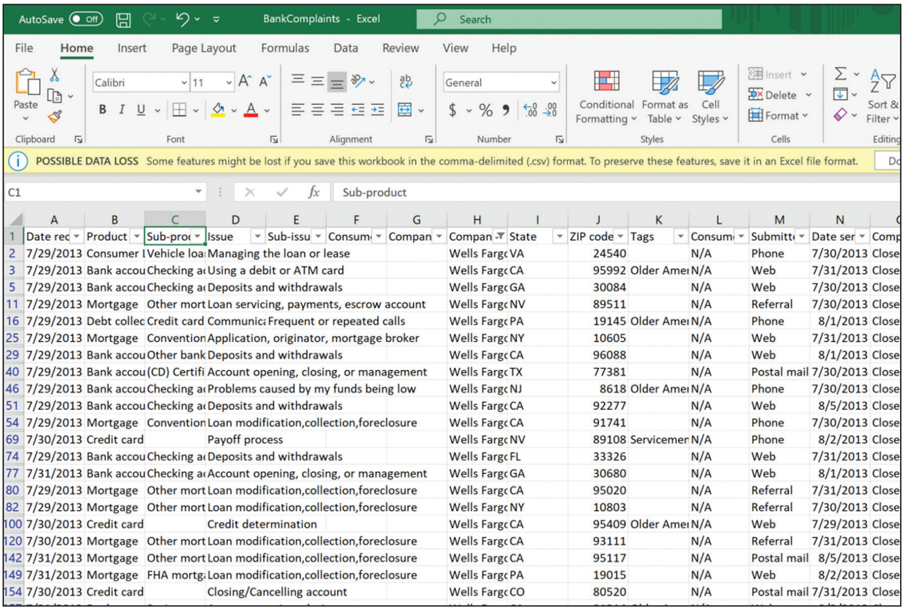
FIGURE 1.1 Exercise to demonstrate that the entire file was loaded and how long it takes to execute a function when the data file is relatively small.

9. Using the Lab Data set and in the *Analysis Case 1.1* folder, find the file *Courses.csv* (73 MB file with 631,139 records, 21 variables). (The data set was made available courtesy of the Harvard Dataverse Project.)

10. Note how long it takes to load. Add filters to the top row. Sort the data by column T. Note how long it takes to perform this task. Filter column T to a "1" value only.

WARNING: The next file to be studied is so large that, if you do not have enough memory on your computer, it may lock up Excel, so be patient when loading and be ready to reboot your machine if the file does not load or the program locks up.

11. Using the Lab Data set and in the *Analysis Case 1.1* folder, find the file *BankComplaints.csv* (306 MB file with 753,324 records, 18 variables). The file is very large because one column contains the full text of the complaints, which may run to several paragraphs each. (The data set was made available courtesy of the U.S. Government Department of Consumer Affairs.)



POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

C1 | *fx* Sub-product

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Date rec	Product	Sub-proc	Issue	Sub-issu	Consum	Compan	Compan	State	ZIP code	Tags	Consum	Submitt	Date ser	Comp
2	7/29/2013	Consumer l	Vehicle loa	Managing the loan or lease				Wells Fargc	VA	24540		N/A	Phone	7/30/2013	Close
3	7/29/2013	Bank accou	Checking ac	Using a debit or ATM card				Wells Fargc	CA	95992	Older Amer	N/A	Web	7/31/2013	Close
5	7/29/2013	Bank accou	Checking ac	Deposits and withdrawals				Wells Fargc	GA	30084		N/A	Web	7/30/2013	Close
11	7/29/2013	Mortgage	Other mort	Loan servicing, payments, escrow account				Wells Fargc	NV	89511		N/A	Referral	7/30/2013	Close
16	7/29/2013	Debt collec	Credit card	Communica	Frequent or repeated calls			Wells Fargc	PA	19145	Older Amer	N/A	Phone	8/1/2013	Close
25	7/29/2013	Mortgage	Convention	Application, originator, mortgage broker				Wells Fargc	NY	10605		N/A	Web	7/31/2013	Close
29	7/29/2013	Bank accou	Other bank	Deposits and withdrawals				Wells Fargc	CA	96088		N/A	Web	8/1/2013	Close
40	7/29/2013	Bank accou	(CD) Certifi	Account opening, closing, or management				Wells Fargc	TX	77381		N/A	Postal mail	7/30/2013	Close
46	7/29/2013	Bank accou	Checking ac	Problems caused by my funds being low				Wells Fargc	NJ	8618	Older Amer	N/A	Phone	7/30/2013	Close
51	7/29/2013	Bank accou	Checking ac	Deposits and withdrawals				Wells Fargc	CA	92277		N/A	Web	8/5/2013	Close
54	7/29/2013	Mortgage	Convention	Loan modification,collection,foreclosure				Wells Fargc	CA	91741		N/A	Phone	7/30/2013	Close
69	7/30/2013	Credit card		Payoff process				Wells Fargc	NV	89108	Servicemen	N/A	Phone	8/2/2013	Close
74	7/29/2013	Bank accou	Checking ac	Deposits and withdrawals				Wells Fargc	FL	33326		N/A	Web	7/31/2013	Close
77	7/31/2013	Bank accou	Checking ac	Account opening, closing, or management				Wells Fargc	GA	30680		N/A	Web	8/1/2013	Close
80	7/29/2013	Mortgage	Other mort	Loan modification,collection,foreclosure				Wells Fargc	CA	95020		N/A	Referral	7/31/2013	Close
82	7/29/2013	Mortgage	Other mort	Loan modification,collection,foreclosure				Wells Fargc	NY	10803		N/A	Referral	7/30/2013	Close
100	7/30/2013	Credit card		Credit determination				Wells Fargc	CA	95409	Older Amer	N/A	Web	7/29/2013	Close
120	7/30/2013	Mortgage	Other mort	Loan modification,collection,foreclosure				Wells Fargc	CA	93111		N/A	Referral	7/31/2013	Close
142	7/31/2013	Mortgage	Other mort	Loan modification,collection,foreclosure				Wells Fargc	CA	95117		N/A	Postal mail	8/5/2013	Close
149	7/31/2013	Mortgage	FHA mortg	Loan modification,collection,foreclosure				Wells Fargc	PA	19015		N/A	Web	8/2/2013	Close
154	7/30/2013	Credit card		Closing/Cancelling account				Wells Fargc	CO	80520		N/A	Postal mail	7/31/2013	Close

FIGURE 1.2 Exercise to demonstrate that the entire file was loaded and how long it takes to execute a function when the data file is relatively large.

12. Note how long it takes to load. Add filters to the top row. Sort the data by column D. Note how long it takes to perform this task. Then, filter column H to a “Wells Fargo” value only (Figure 1.2).

13. Excel, as a tool, does not always handle large data sets well. We will work with the large data files by sampling them and analyzing the samples in Chapter 14.

Analysis Case 1.2 – Cleaning the Data File

A drug manufacturer has collected drug test data on 178 patients. We suspect that the data has transcription problems. (There were errors when entering data into the computer from the experiment notes.)

1. Using an Internet connection and Web browser, navigate to *http://bit.ly/2zoUVqz*.
2. Scrape the data and paste it into a new Excel spreadsheet. Note that all the data is in one column. The data dictionary is available at *https://bit.ly/2HymiXr*.
3. Use the "Text to Columns" function under the Data ribbon to distribute the data into their respective columns.

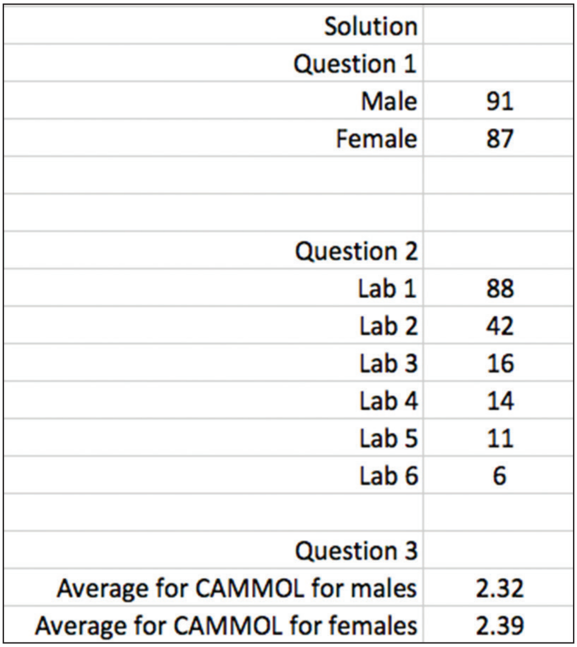
Or obtain the data by this other method:

4. Using the Lab Data set and in the *Analysis Case 1.1* folder, find the file *calciumgood.txt*. The data dictionary for this file is *calcium.txt*. (The data set was made available by permission of John P. Holcomb, Jr.)
5. Open *calciumgood.txt* using Excel.
6. Employ the Excel function to import a data file under the Data set of functions.
7. The data is in columns, but there are no column titles. Use the data dictionary to add column titles so all the variables are labeled. Note that the file is now in a flat-file format with columns as variables and rows as records.
8. Save the file as *calcium.xlsx*.
9. The file contains many errors. Clean it by looking at it and correcting these errors. (For example, some of the numbers in the SEX column are coded as 22 instead of 2. You can fix that easily. Fixing data coded as 12 is harder.) If you need to refer to the original data that was collected, use this link to access the original observations: *http://academic.csuohio.edu/holcombj/clean/bigtable.htm*.

10. When you feel reasonably sure you have a clean data file, answer the following questions by using Excel:

> *How many men and how many women were in the study? Sort by gender and compute a sum for each group.*
>
> *Were the tests evenly distributed over the labs? Sort by lab type and compute subtotals by lab type.*
>
> *Are the calcium levels for the males above or below the average for the females in the test? Sort by gender and use the AVG function to average the CAMMOL columns for each sex.*

11. Check your results against the following solutions (Figure 1.3). Keep cleaning the data until you have found all the errors.

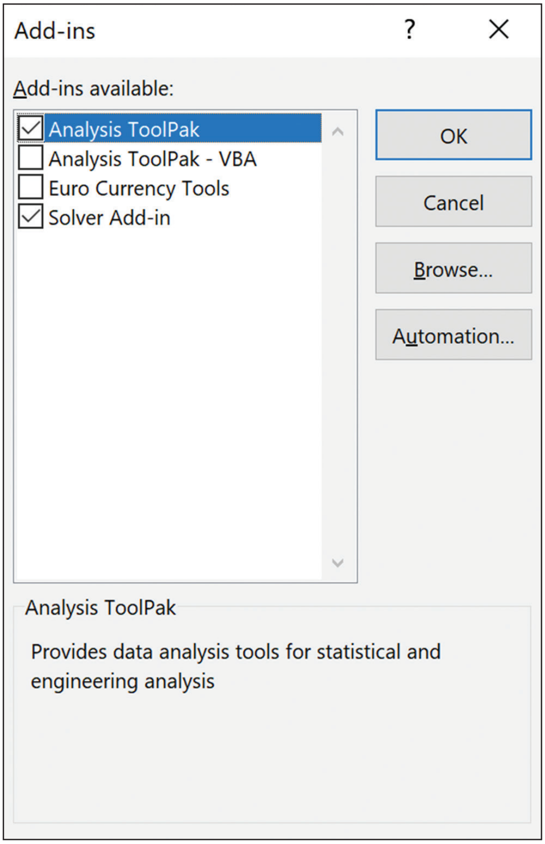


Solution	
Question 1	
Male	91
Female	87
Question 2	
Lab 1	88
Lab 2	42
Lab 3	16
Lab 4	14
Lab 5	11
Lab 6	6
Question 3	
Average for CAMMOL for males	2.32
Average for CAMMOL for females	2.39

FIGURE 1.3 Analysis results computed after cleaning the data file.

CHAPTER 2

INSTALLING THE ANALYSIS TOOLPAK



Add-ins

Add-ins available:

- ☑ Analysis ToolPak
- ☐ Analysis ToolPak - VBA
- ☐ Euro Currency Tools
- ☑ Solver Add-in

Analysis ToolPak

Provides data analysis tools for statistical and engineering analysis

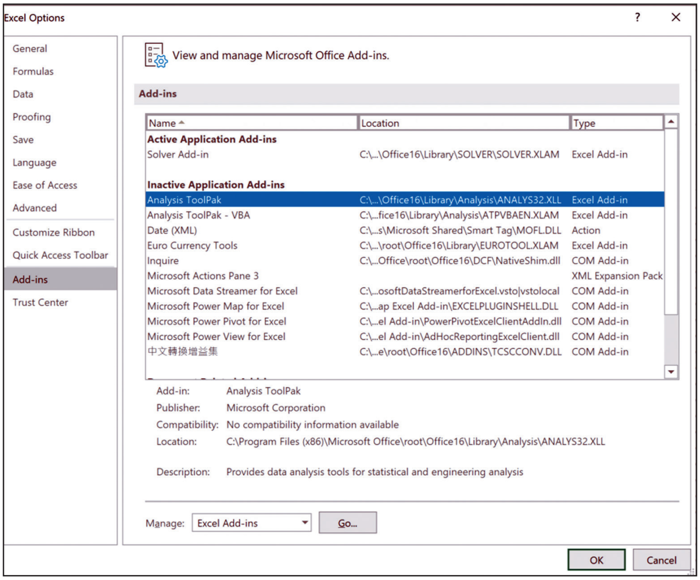
The techniques in this book are practiced using Microsoft Excel. It turns out that Excel has a wealth of advanced analysis tools, from regression to inferential tools, tucked in a hidden tool kit called the Analysis ToolPak. It's there in Excel already, nothing to install, you just have to activate it. We make use of many of these advanced tools in the exercises throughout the book, so it is good to activate it early on before tackling the advanced exercises.

The first exercise guides the reader through the activation process. Note that the activation varies by operating system (PC or Mac) and Excel version type. At the end of this chapter, there is a simple exercise to ensure the Analysis ToolPak is active and readily available.

Analysis Case 2.1 – Excel Analysis ToolPak

Installation and Activation

1. Launch Excel.
2. If using Microsoft Windows, click the Office button logo (or "File" in 2010) in the upper-left corner of the window.
3. On the PC version, click "Options" on the bottom of the pop-up. Click "Add-ins" in the Excel Options pop-up. Select "Analysis ToolPak," and then click "Go." You should see a dialog screen as shown in Figure 2.1.



View and manage Microsoft Office Add-ins.

Add-ins

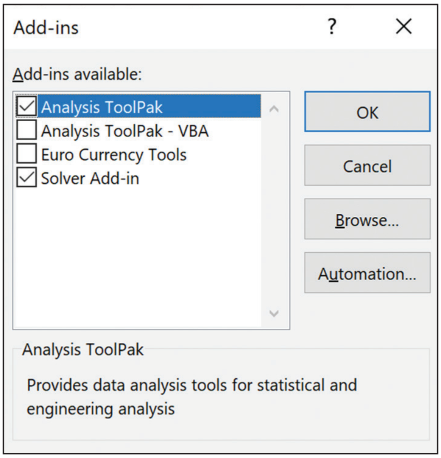
Name	Location	Type
Active Application Add-ins		
Solver Add-in	C:\...\Office16\Library\SOLVER\SOLVER.XLAM	Excel Add-in
Inactive Application Add-ins		
Analysis ToolPak	C:\...\Office16\Library\Analysis\ANALYS32.XLL	Excel Add-in
Analysis ToolPak - VBA	C:\...fice16\Library\Analysis\ATPVBAEN.XLAM	Excel Add-in
Date (XML)	C:\...s\Microsoft Shared\Smart Tag\MOFL.DLL	Action
Euro Currency Tools	C:\...\root\Office16\Library\EUROTOOL.XLAM	Excel Add-in
Inquire	C:\...Office\root\Office16\DCF\NativeShim.dll	COM Add-in
Microsoft Actions Pane 3		XML Expansion Pack
Microsoft Data Streamer for Excel	C:\...osoftDataStreamerforExcel.vsto|vstolocal	COM Add-in
Microsoft Power Map for Excel	C:\...ap Excel Add-in\EXCELPLUGINSHELL.DLL	COM Add-in
Microsoft Power Pivot for Excel	C:\...el Add-in\PowerPivotExcelClientAddIn.dll	COM Add-in
Microsoft Power View for Excel	C:\...el Add-in\AdHocReportingExcelClient.dll	COM Add-in
中文轉換增益集	C:\...e\root\Office16\ADDINS\TCSCCONV.DLL	COM Add-in

Add-in: Analysis ToolPak
Publisher: Microsoft Corporation
Compatibility: No compatibility information available
Location: C:\Program Files (x86)\Microsoft Office\root\Office16\Library\Analysis\ANALYS32.XLL

Description: Provides data analysis tools for statistical and engineering analysis

FIGURE 2.1 The Excel wizard showing the Analysis ToolPak prior to making it active.

4. In the Add-Ins pop-up screen, place a check mark in the box next to "Analysis ToolPak" on the list of available add-ins (Figure 2.2). Then click OK.



Add-ins available:

☑ Analysis ToolPak
□ Analysis ToolPak - VBA
□ Euro Currency Tools
☑ Solver Add-in

Analysis ToolPak

Provides data analysis tools for statistical and engineering analysis

FIGURE 2.2 The second wizard screen in activating the Analysis ToolPak showing it as active.

5. You should now see Data Analysis as a selection in the Analysis group under the Data ribbon in Excel (Figure 2.3).

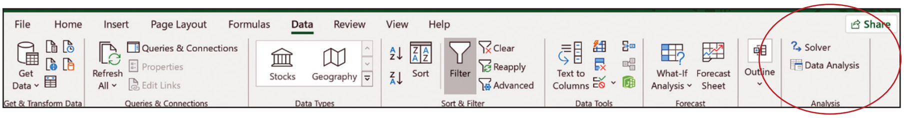


FIGURE 2.3 The Excel Data ribbon after installation showing the activated Analysis ToolPak now appearing as a "Data Analysis" button.

On a Mac, the Analysis ToolPak is only available in the Excel 2016 version, not in earlier versions. To activate it on a Mac, open a spreadsheet and, in the "Tools" option (Figure 2.4) on the main menu, select "Excel Add-ins" and then "Analysis ToolPak" from the wizard pop-up box to activate (Figure 2.5).

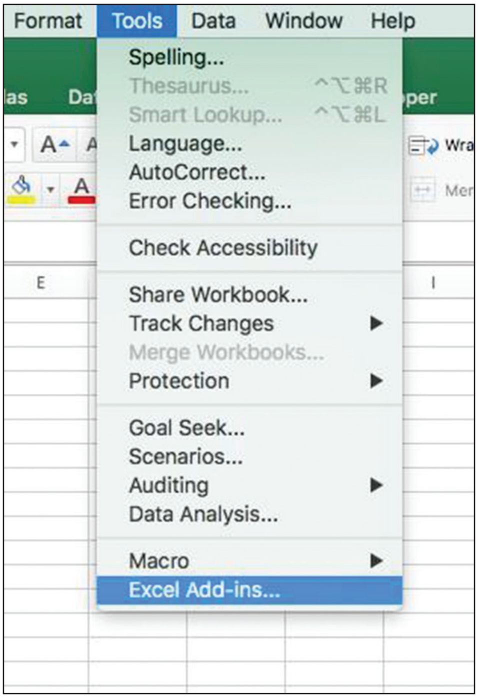
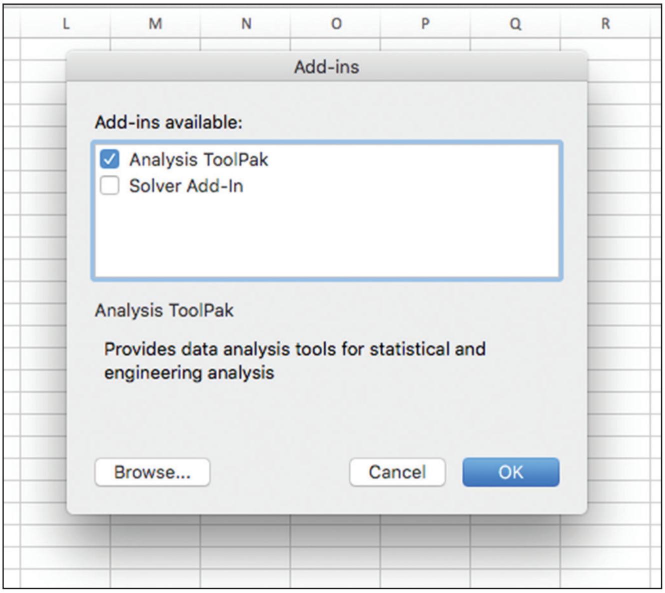


FIGURE 2.4 Location of the "Add-ins" function for the Mac version of Excel.



Add-ins available:

☑ Analysis ToolPak
□ Solver Add-In

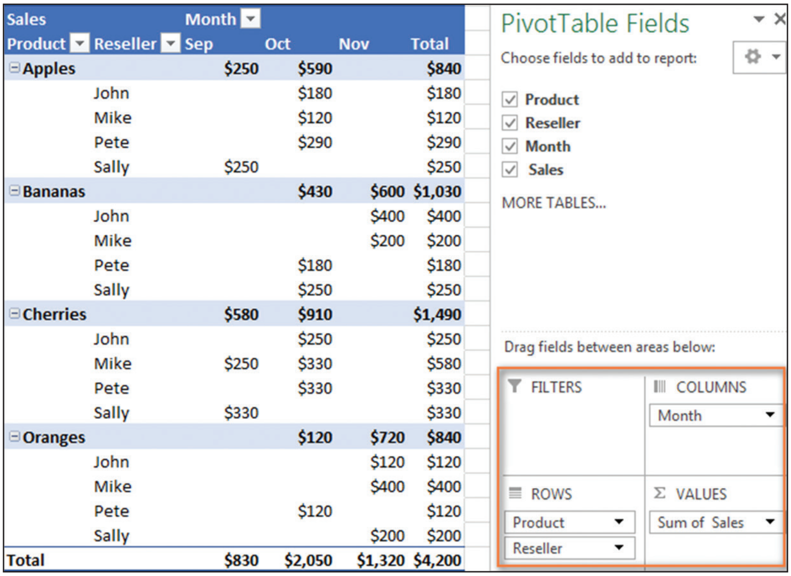
Analysis ToolPak

Provides data analysis tools for statistical and engineering analysis

FIGURE 2.5 The Mac version of Excel Add-ins wizard screen showing the Analysis ToolPak being activated.

CHAPTER 3

DESCRIPTIVE STATISTICS



Sales		Month ▾			
Product ▾	Reseller ▾	Sep	Oct	Nov	Total
⊟ **Apples**		**$250**	**$590**		**$840**
	John		$180		$180
	Mike		$120		$120
	Pete		$290		$290
	Sally	$250			$250
⊟ **Bananas**			**$430**	**$600**	**$1,030**
	John			$400	$400
	Mike			$200	$200
	Pete		$180		$180
	Sally		$250		$250
⊟ **Cherries**		**$580**	**$910**		**$1,490**
	John		$250		$250
	Mike	$250	$330		$580
	Pete		$330		$330
	Sally	$330			$330
⊟ **Oranges**			**$120**	**$720**	**$840**
	John			$120	$120
	Mike			$400	$400
	Pete		$120		$120
	Sally			$200	$200
Total		**$830**	**$2,050**	**$1,320**	**$4,200**

PivotTable Fields

Choose fields to add to report:

☑ **Product**

☑ **Reseller**

☑ **Month**

☑ **Sales**

MORE TABLES...

Drag fields between areas below:

FILTERS

COLUMNS: Month ▾

ROWS: Product ▾, Reseller ▾

Σ VALUES: Sum of Sales ▾

Descriptive statistics deal with the past and the present: what happened? They differ from predictive analytics, which deals very much with the future: what might happen? Or with assurance or results, inferential statistics: are we sure, or is it the result of some random event? And descriptive statistics are a summarization of numerical (averages, sums, extreme) or categorical variables (tabulation). This also differs from summarizing text data: what are people saying? We tackle that in Chapter 13.

This technique answers the business question: "How many are there, how much, and how do they compare?"

The primary tool for descriptive statistics will be the five-point summaries (available in the ToolPak). We cover quartiles as well as averages and medians, maximum, minimum, and measure so the spread of the data, variances, and interquartile ranges.

This chapter also introduces two other very useful tools: (a) the creation of box plots as a summarization and visualization tool for numeric variables; and (b) the use of Pivot Tables as a tabulation tool for categorical variables.

In this chapter, we begin the practice of offering two types of exercises: basic exercises for beginners as well as additional and more challenging exercises for advanced students. If you are a beginner, getting through the basic exercises for each tool is a good start. For more advanced students, we offer additional exercises that are more challenging. All are encouraged to try them out as well.

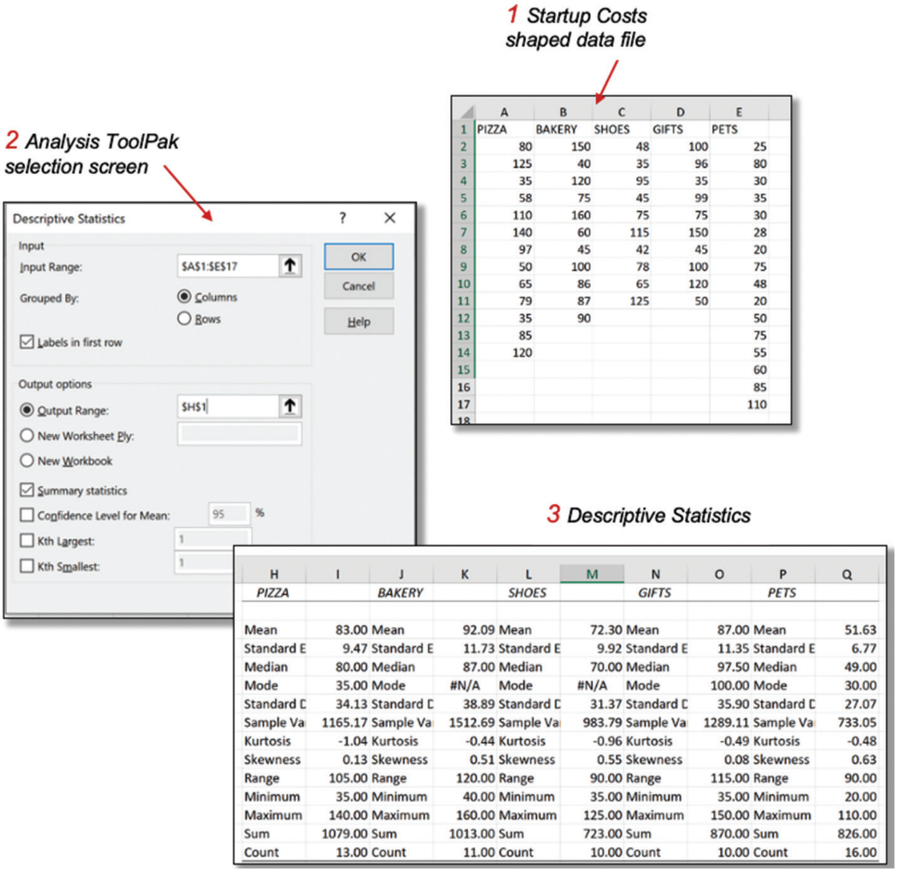
Analysis Case 3.1 – Descriptive Statistics

Five-Point Summaries

1. Using the Lab Data set provided, open the *Analysis Case 3* folder in it, and find the file *StartupCosts.xlsx.*
2. Open *StartupCosts.xlsx* using Excel.
3. We are going to answer these questions:

   > *Which type of startup business has the best characteristics?*
   >
   > *What are the descriptive statistics for a group of businesses to be able to compare them?*

4. Following our practice of not changing raw data, select the data only (leaving out the data dictionary at the bottom of the table and making sure to copy the headings), and then file and copy it as the shaped file in a new spreadsheet. Label this spreadsheet *StartUpStats*.
5. Using the Analysis ToolPak, select "Descriptive Statistics." Enter the entire range of data, including the column labels, as the range. Make sure to click the "labels in the first row" box. Put the results in another spreadsheet or somewhere in the same sheet as the shaped file. Select only the "Summary Statistics" box.
6. Change the formatting of the statistics to Numbers and two decimals. Put the column labels on top of the stats numbers and the row labels. Delete every other row label to leave only the statistics.
7. Compute Q1 (first quartile) and Q3 (third quartile) at the bottom of each column in the statistics to add these important numbers.


1 Startup Costs shaped data file
2 Analysis ToolPak selection screen
3 Descriptive Statistics

	A	B	C	D	E
1	PIZZA	BAKERY	SHOES	GIFTS	PETS
2	80	150	48	100	25
3	125	40	35	96	80
4	35	120	95	35	30
5	58	75	45	99	35
6	110	160	75	75	30
7	140	60	115	150	28
8	97	45	42	45	20
9	50	100	78	100	75
10	65	86	65	120	48
11	79	87	125	50	20
12	35	90			50
13	85				75
14	120				55
15					60
16					85
17					110
18					

Descriptive Statistics

Input

Input Range: $A$1:$E$17

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Output options

● Output Range: $H$1

○ New Worksheet Ply:

○ New Workbook

☑ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

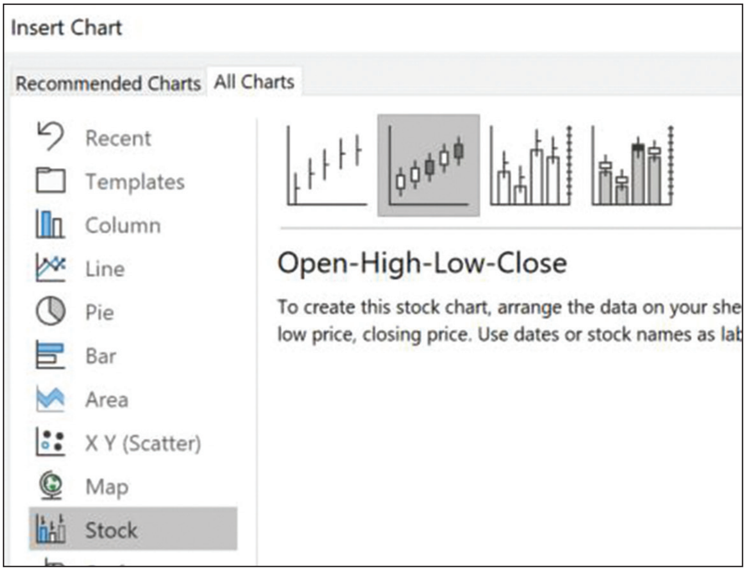
□ Kth Smallest: 1

H	I	J	K	L	M	N	O	P	Q
PIZZA		BAKERY		SHOES		GIFTS		PETS	
Mean	83.00	Mean	92.09	Mean	72.30	Mean	87.00	Mean	51.63
Standard E	9.47	Standard E	11.73	Standard E	9.92	Standard E	11.35	Standard E	6.77
Median	80.00	Median	87.00	Median	70.00	Median	97.50	Median	49.00
Mode	35.00	Mode	#N/A	Mode	#N/A	Mode	100.00	Mode	30.00
Standard D	34.13	Standard D	38.89	Standard D	31.37	Standard D	35.90	Standard D	27.07
Sample Va	1165.17	Sample Va	1512.69	Sample Va	983.79	Sample Va	1289.11	Sample Va	733.05
Kurtosis	-1.04	Kurtosis	-0.44	Kurtosis	-0.96	Kurtosis	-0.49	Kurtosis	-0.48
Skewness	0.13	Skewness	0.51	Skewness	0.55	Skewness	0.08	Skewness	0.63
Range	105.00	Range	120.00	Range	90.00	Range	115.00	Range	90.00
Minimum	35.00	Minimum	40.00	Minimum	35.00	Minimum	35.00	Minimum	20.00
Maximum	140.00	Maximum	160.00	Maximum	125.00	Maximum	150.00	Maximum	110.00
Sum	1079.00	Sum	1013.00	Sum	723.00	Sum	870.00	Sum	826.00
Count	13.00	Count	11.00	Count	10.00	Count	10.00	Count	16.00

FIGURE 3.1 Steps in using the Analysis ToolPak to obtain the Descriptive Statistics and a 5-point summary for a data set.

8. Create a five-point summary table of MEDIAN, Q1, MAXIMUM, MINIMUM, Q3 by copying the appropriate elements for the previous table (Figure 3.1).

9. Enter a set of arbitrary dates above each column to be able to use the stock table chart. Select the dates and the five-point summary data for all five startups and insert a chart. Select the (Open, High, Low, Close) stock chart (Figure 3.2).



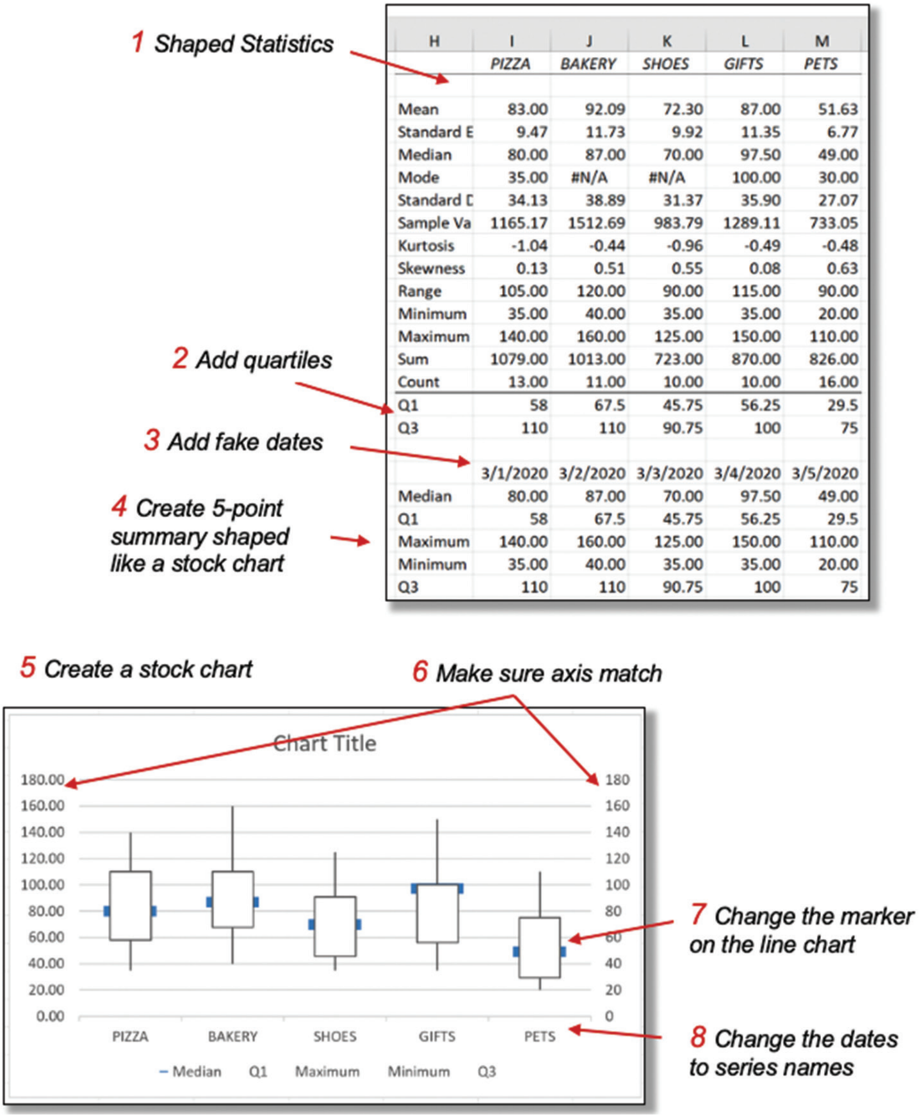
Open-High-Low-Close

To create this stock chart, arrange the data on your she
low price, closing price. Use dates or stock names as lab

FIGURE 3.2 Selecting the appropriate stock-chart type for constructing box-plot charts.

10. Edit the chart to make both axes the same (both maximum at 180). Now, we will change the x-axis to replace the dates with the names of the startups. Right-click on the chart to Select Data (NOT Format Axis) and change the settings in the Horizontal (Category) Axis Labels. Edit the median line by changing the median to a dash with a width of 30 to make it visible (Figure 3.3).

11. From the resulting box plots, decide which type of startup seems most favorable.



1 Shaped Statistics

2 Add quartiles

3 Add fake dates

4 Create 5-point summary shaped like a stock chart

H	I	J	K	L	M
	PIZZA	*BAKERY*	*SHOES*	*GIFTS*	*PETS*
Mean	83.00	92.09	72.30	87.00	51.63
Standard E	9.47	11.73	9.92	11.35	6.77
Median	80.00	87.00	70.00	97.50	49.00
Mode	35.00	#N/A	#N/A	100.00	30.00
Standard D	34.13	38.89	31.37	35.90	27.07
Sample Va	1165.17	1512.69	983.79	1289.11	733.05
Kurtosis	-1.04	-0.44	-0.96	-0.49	-0.48
Skewness	0.13	0.51	0.55	0.08	0.63
Range	105.00	120.00	90.00	115.00	90.00
Minimum	35.00	40.00	35.00	35.00	20.00
Maximum	140.00	160.00	125.00	150.00	110.00
Sum	1079.00	1013.00	723.00	870.00	826.00
Count	13.00	11.00	10.00	10.00	16.00
Q1	58	67.5	45.75	56.25	29.5
Q3	110	110	90.75	100	75
	3/1/2020	3/2/2020	3/3/2020	3/4/2020	3/5/2020
Median	80.00	87.00	70.00	97.50	49.00
Q1	58	67.5	45.75	56.25	29.5
Maximum	140.00	160.00	125.00	150.00	110.00
Minimum	35.00	40.00	35.00	35.00	20.00
Q3	110	110	90.75	100	75

5 Create a stock chart

6 Make sure axis match

7 Change the marker on the line chart

8 Change the dates to series names

FIGURE 3.3 Completing the construction of a box-plot diagram for a data set using the stock-chart format.

Analysis Case 3.2 – Additional Analysis Case Using the ORDERS File

1. Using the Lab Data set and in the *Analysis Case* 3 folder, find the file *ORDERS.xlsx.*

2. Open *ORDERS.xlsx* using Excel.

3. We are going to answer these questions:

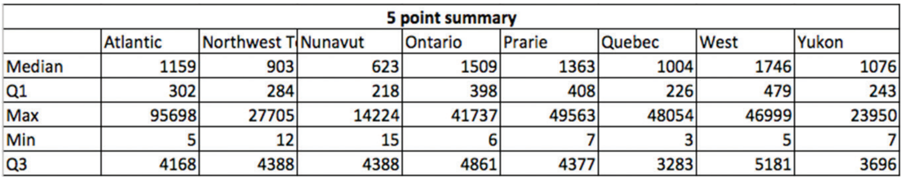
   Which regions had the best average sales for the year?

   What are the descriptive statistics for each sales region? Compare them.

4. Create descriptive statistics of sales by region and create a boxplot diagram.

5. Use a Pivot Table to create a tabulation of sales by region, making sure to sum sales by ORDERDATE to have a large Pivot Table as a result for the next step.

6. Create a descriptive statistics table using Data > Analysis ToolPak (Figure 3.4):



5 point summary								
	Atlantic	Northwest T	Nunavut	Ontario	Prarie	Quebec	West	Yukon
Median	1159	903	623	1509	1363	1004	1746	1076
Q1	302	284	218	398	408	226	479	243
Max	95698	27705	14224	41737	49563	48054	46999	23950
Min	5	12	15	6	7	3	5	7
Q3	4168	4388	4388	4861	4377	3283	5181	3696

FIGURE 3.4 Resulting 5-point summary for the *ORDERS* file using the Analysis ToolPak Descriptive Statistics function.

7. Then, generate the box plots (follow instructions from earlier in this exercise on generating box plots) (Figure 3.5):

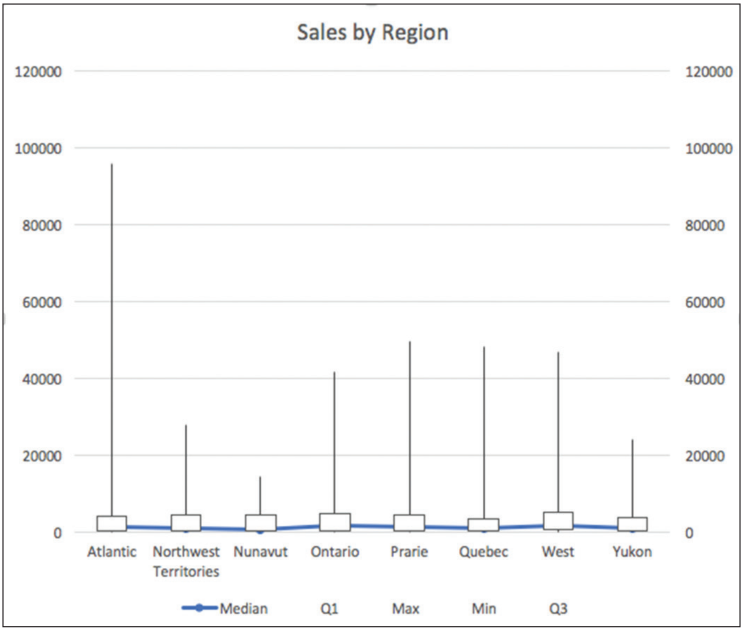


FIGURE 3.5 The completed box-plot chart for the *ORDERS* file case.

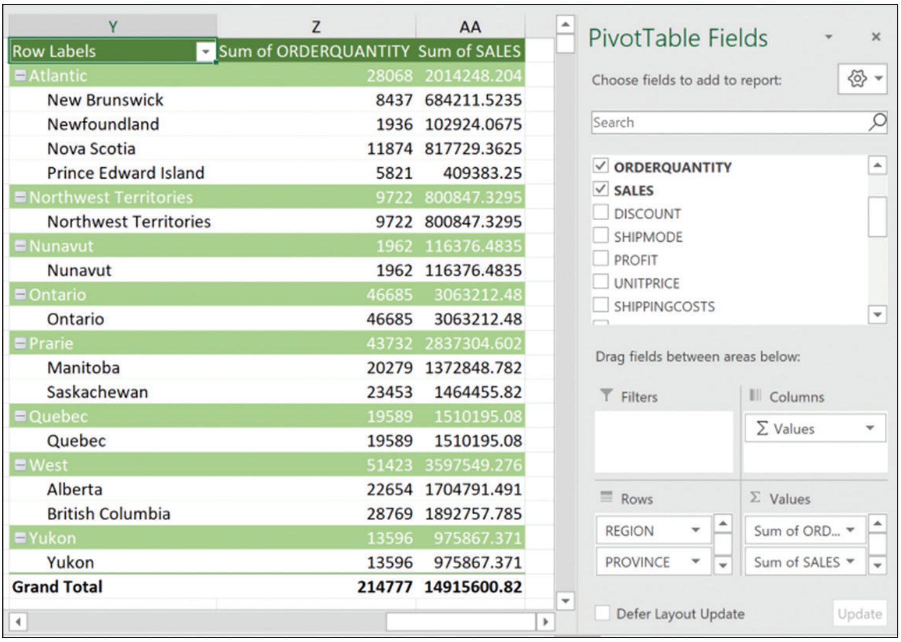
Analysis Case 3.3 – Descriptive Statistics

Tabulation and Pivot Tables

1. Using the Lab Data set provided, open the *Analysis Case 3* folder and find the file *ORDERS.xlsx*.
2. Open *ORDERS.xlsx* using Excel.
3. We are going to answer this question:

   Which province has the largest number of big sales? (Tabulate order quantity and sales by region and province. Cross-tabulate profit by province and customer segment.)

4. Following our practice of not changing the raw data, select the entire file and copy it as the shaped file in a new spreadsheet, and label the tab *Summary*. Create a table with the name *ORDERS*.
5. Using a Pivot Table, generate answers to the question in Step 3.
6. Tabulate order quantity and sales by region and province (Figure 3.6):



Row Labels	Sum of ORDERQUANTITY	Sum of SALES
⊟Atlantic	28068	2014248.204
New Brunswick	8437	684211.5235
Newfoundland	1936	102924.0675
Nova Scotia	11874	817729.3625
Prince Edward Island	5821	409383.25
⊟Northwest Territories	9722	800847.3295
Northwest Territories	9722	800847.3295
⊟Nunavut	1962	116376.4835
Nunavut	1962	116376.4835
⊟Ontario	46685	3063212.48
Ontario	46685	3063212.48
⊟Prarie	43732	2837304.602
Manitoba	20279	1372848.782
Saskachewan	23453	1464455.82
⊟Quebec	19589	1510195.08
Quebec	19589	1510195.08
⊟West	51423	3597549.276
Alberta	22654	1704791.491
British Columbia	28769	1892757.785
⊟Yukon	13596	975867.371
Yukon	13596	975867.371
Grand Total	**214777**	**14915600.82**

PivotTable Fields

Choose fields to add to report:

Search

☑ ORDERQUANTITY
☑ SALES
□ DISCOUNT
□ SHIPMODE
□ PROFIT
□ UNITPRICE
□ SHIPPINGCOSTS

Drag fields between areas below:

Filters

Columns: Σ Values

Rows: REGION, PROVINCE

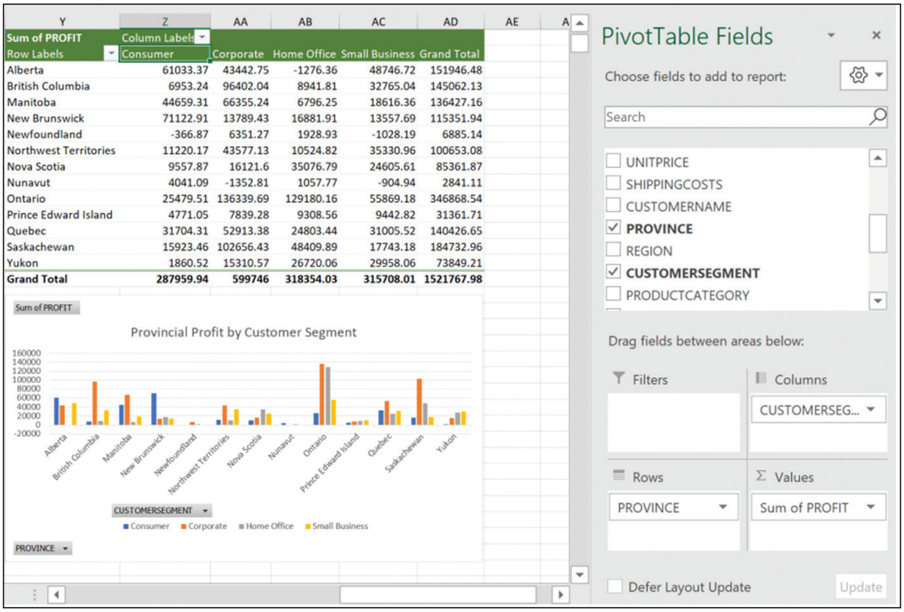
Values: Sum of ORD..., Sum of SALES

□ Defer Layout Update

Update

FIGURE 3.6 Pivot table settings to tabulate order quantity and sales by region and province in the *ORDERS* data set.

7. Tabulate profit by province (independent variable, rows) and customer segment (dependent variables, columns) (Figure 3.7). Which province appears to have the most profits across customer categories?



Sum of PROFIT	Column Labels				
Row Labels	Consumer	Corporate	Home Office	Small Business	Grand Total
Alberta	61033.37	43442.75	-1276.36	48746.72	151946.48
British Columbia	6953.24	96402.04	8941.81	32765.04	145062.13
Manitoba	44659.31	66355.24	6796.25	18616.36	136427.16
New Brunswick	71122.91	13789.43	16881.91	13557.69	115351.94
Newfoundland	-366.87	6351.27	1928.93	-1028.19	6885.14
Northwest Territories	11220.17	43577.13	10524.82	35330.96	100653.08
Nova Scotia	9557.87	16121.6	35076.79	24605.61	85361.87
Nunavut	4041.09	-1352.81	1057.77	-904.94	2841.11
Ontario	25479.51	136339.69	129180.16	55869.18	346868.54
Prince Edward Island	4771.05	7839.28	9308.56	9442.82	31361.71
Quebec	31704.31	52913.38	24803.44	31005.52	140426.65
Saskachewan	15923.46	102656.43	48409.89	17743.18	184732.96
Yukon	1860.52	15310.57	26720.06	29958.06	73849.21
Grand Total	**287959.94**	**599746**	**318354.03**	**315708.01**	**1521767.98**

PivotTable Fields

Choose fields to add to report:

□ UNITPRICE
□ SHIPPINGCOSTS
□ CUSTOMERNAME
☑ **PROVINCE**
□ REGION
☑ **CUSTOMERSEGMENT**
□ PRODUCTCATEGORY

Drag fields between areas below:

Filters:

Columns: CUSTOMERSEG...

Rows: PROVINCE

Values: Sum of PROFIT

□ Defer Layout Update

FIGURE 3.7 Pivot table settings to tabulate profit by province and customer segment in the *ORDERS* data set.

8. Which customer categories appear to be the most profitable across all provinces?

Now, let's compute outliers.

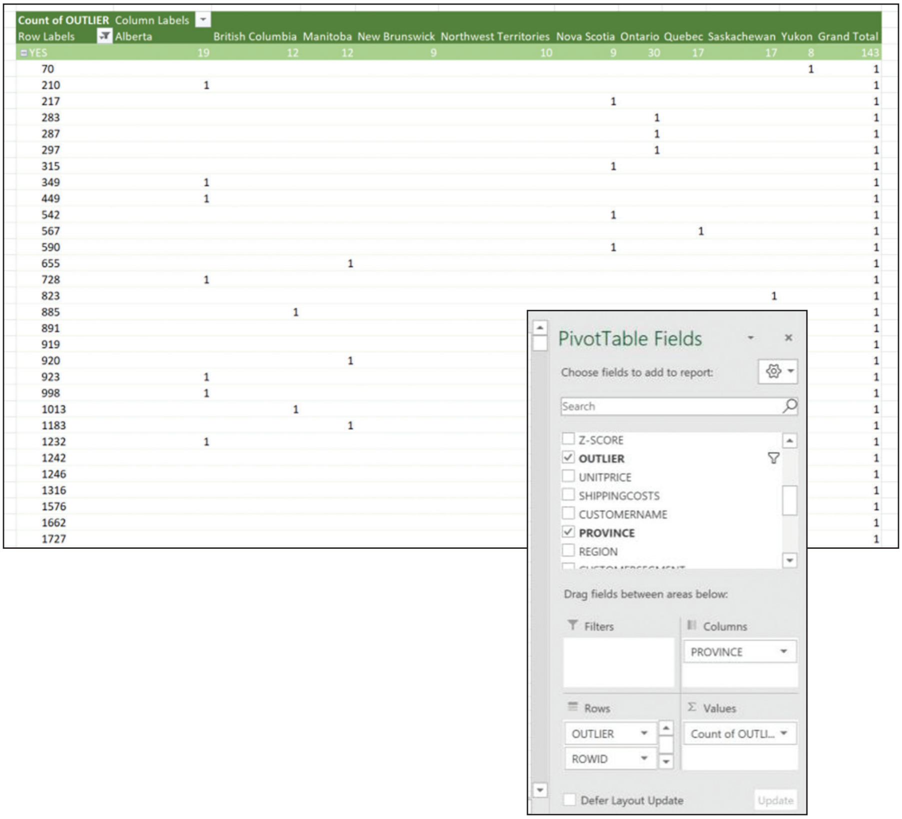
9. Insert an empty column next to PROFIT. Insert the z-score of the PROFIT column into this new variable. Use the STANDARDIZE function on profit. Label it ZSCORE.

10. Sort the table by z-score to identify big winners (z-score > 3) and big losers (z-score < 3).

11. Create another column and code the outliers into YES (ABS(z-score) = or > 3.0), or NO (ABS(z-score) < 3.0). Label it OUTLIER.

12. Using a Pivot Table, create a summary table of the number of the outliers by z-score code and identify the province that has the highest number of sales outliers (Figure 3.8). Make sure to use a filter of OUTLIER=YES to only count the outliers.

13. Which province appears to have the largest number of sales ABS(z-score) > 3.0?



Count of OUTLIER	Column Labels										
Row Labels	Alberta	British Columbia	Manitoba	New Brunswick	Northwest Territories	Nova Scotia	Ontario	Quebec	Saskachewan	Yukon	Grand Total
⊟YES	19	12	12	9	10	9	30	17	17	8	143
70										1	1
210	1										1
217						1					1
283							1				1
287							1				1
297							1				1
315						1					1
349	1										1
449	1										1
542						1					1
567								1			1
590						1					1
655			1								1
728	1										1
823									1		1
885		1									1
891											1
919											1
920			1								1
923	1										1
998	1										1
1013		1									1
1183			1								1
1232	1										1
1242											1
1246											1
1316											1
1576											1
1662											1
1727											1

PivotTable Fields

Choose fields to add to report:

Search

☐ Z-SCORE
☑ **OUTLIER**
☐ UNITPRICE
☐ SHIPPINGCOSTS
☐ CUSTOMERNAME
☑ **PROVINCE**
☐ REGION

Drag fields between areas below:

Filters

Columns: PROVINCE

Rows: OUTLIER, ROWID

Σ Values: Count of OUTLI...

☐ Defer Layout Update

Update

FIGURE 3.8 Using a Pivot Table tabulation technique to discover which region has the greatest number of outliers in ORDERS.

Analysis Case 3.4 – Additional Case Using Titanic Data

1. Using the Lab Data set and in the *Analysis Case 3* folder, find the file *Titanic.xlsx*.

2. Open *Titanic.xlsx* using Excel.

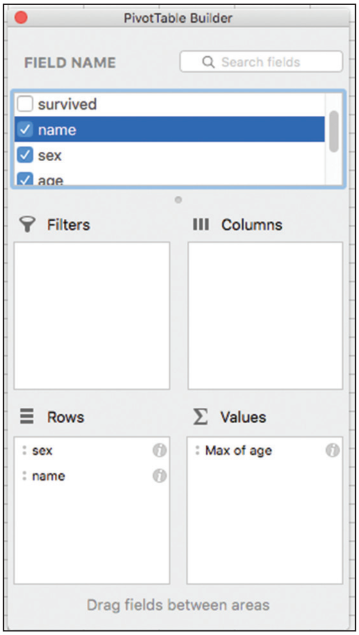
3. We are going to answer these questions:

   Were male passengers older or younger, on average, than female passengers?

   What are the descriptive statistics for each gender? Compare them. Repeat for each class of passenger and compare.

4. Create a Pivot Table of the Titanic Passenger data (Figure 3.9).

5. Create a summary by age. Do a tabulation by age and enter the maximum age for males and females. Do a sub-summary under each gender of the passenger name. Now you have two lists, one above the other, of passenger ages sorted by gender.



□ survived
☑ name
☑ sex
☑ age

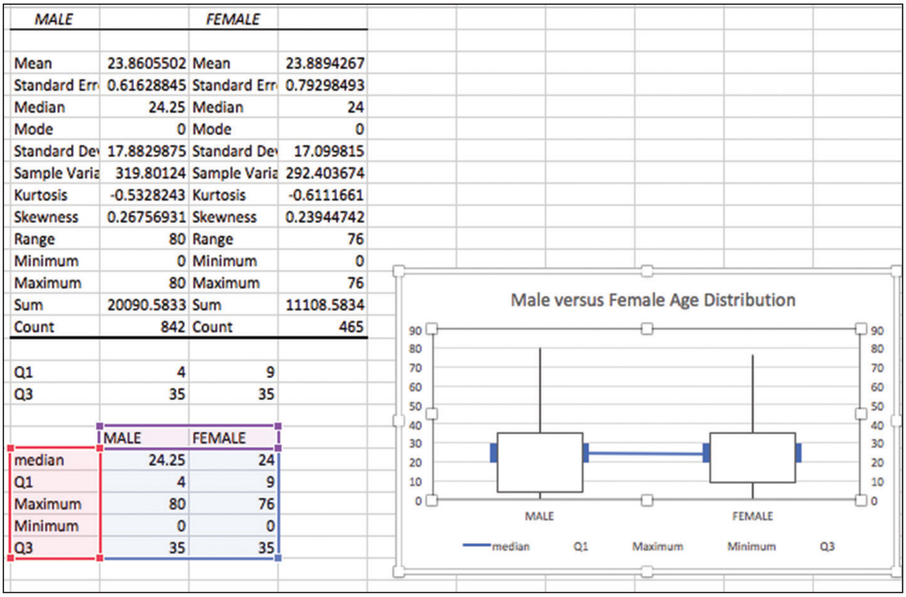
Rows: sex, name

Values: Max of age

FIGURE 3.9 Pivot Table configuration to summarize passengers by gender and subcategory of names to create two lists by gender.

6. Scrape each list and paste it into a new sheet under MALE and FEMALE columns.

7. Obtain the descriptive statistics of age by gender (labeled sex in the file) and create a box-plot diagram (Figure 3.10).



MALE		FEMALE	
Mean	23.8605502	Mean	23.8894267
Standard Err	0.61628845	Standard Err	0.79298493
Median	24.25	Median	24
Mode	0	Mode	0
Standard Dev	17.8829875	Standard Dev	17.099815
Sample Varia	319.80124	Sample Varia	292.403674
Kurtosis	-0.5328243	Kurtosis	-0.6111661
Skewness	0.26756931	Skewness	0.23944742
Range	80	Range	76
Minimum	0	Minimum	0
Maximum	80	Maximum	76
Sum	20090.5833	Sum	11108.5834
Count	842	Count	465
Q1	4	9	
Q3	35	35	

	MALE	FEMALE
median	24.25	24
Q1	4	9
Maximum	80	76
Minimum	0	0
Q3	35	35

FIGURE 3.10 Final box plots summarizing passenger’s ages compared by gender.

Analysis Case 3.5 – Additional Case Using SFO Airport Survey Data

1. Use the latest SFO Airport ACQ Survey data downloaded from *https://www.flysfo.com/media/customer-survey-data*.

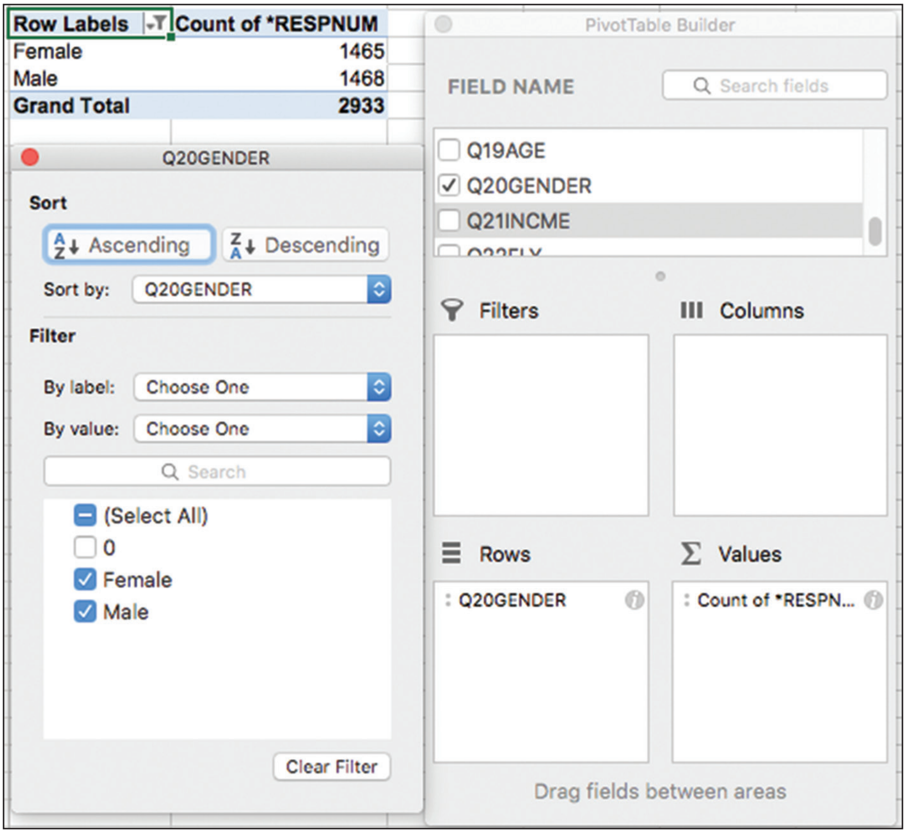
2. Or using the Lab Data set provided in the *Analysis Case 3* folder, find the file *2016_SFO_Customer_Survey_Data.xls*. (This data set was made available courtesy of the San Francisco Airport.)

3. Open the data dictionary file and have it available to consult as you work with the data.

4. Open *2016_SFO_Customer_Survey_Data.xls* using Excel.

5. Save the reshaped data file to your computer as *SFOAirport Survey2016.xlsx*.

6. Tabulate the number of men and women taking the survey (Figure 3.11). Is the difference between men and women larger than 10% or about evenly distributed? Make sure to use the filter option on the row categories to only select Men and Women counts.



Row Labels	Count of *RESPNUM
Female	1465
Male	1468
Grand Total	**2933**

Q20GENDER

Sort

Ascending Descending

Sort by: Q20GENDER

Filter

By label: Choose One

By value: Choose One

☑ (Select All)
□ 0
☑ Female
☑ Male

Clear Filter

PivotTable Builder

FIELD NAME

□ Q19AGE
☑ Q20GENDER
□ Q21INCME

Filters

Columns

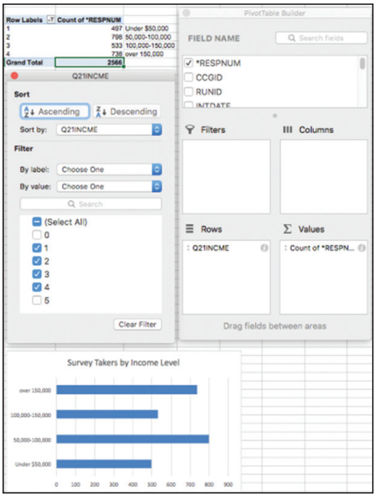
Rows: Q20GENDER

Values: Count of *RESPN...

Drag fields between areas

FIGURE 3.11 Tabulation of SFO survey takers by gender.

7. Tabulate the distribution of income levels among respondents (Figure 3.12).



Row Labels	Count of *RESPNUM	
1	497	Under $50,000
2	798	50,000-100,000
3	533	100,000-150,000
4	738	over 150,000
Grand Total	2566	

Sort by: Q21INCME

By label: Choose One

By value: Choose One

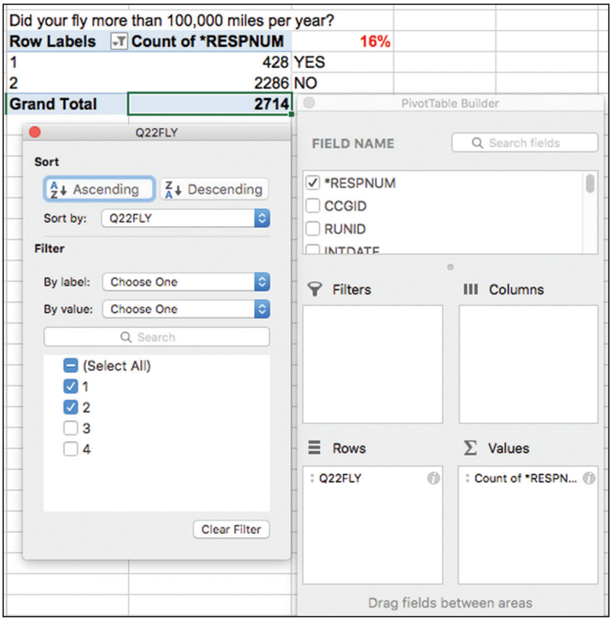
□ 0
☑ 1
☑ 2
☑ 3
☑ 4
□ 5

Rows: Q21INCME

Values: Count of *RESPN...

FIGURE 3.12 Tabulation of income distribution for SFO survey takers.

8. Tabulate the number of frequent flyer passengers who use the airport on average (Figure 3.13).



Did your fly more than 100,000 miles per year?

Row Labels	Count of *RESPNUM		16%
1	428	YES	
2	2286	NO	
Grand Total	2714		

Sort by: Q22FLY

By label: Choose One

By value: Choose One

☑ 1
☑ 2
□ 3
□ 4

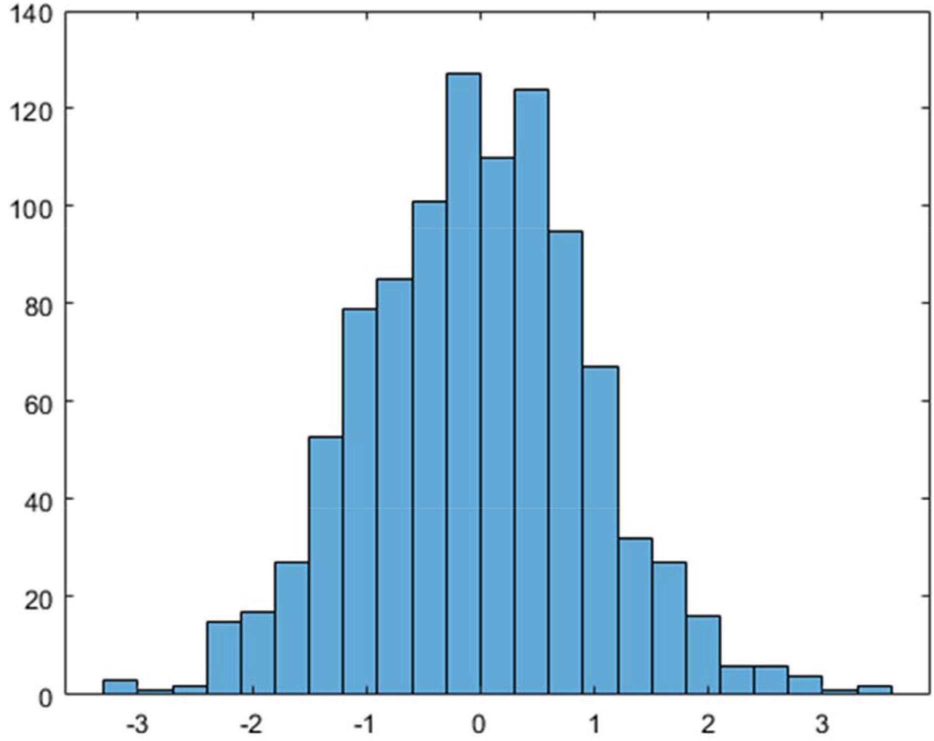
Rows: Q22FLY

Values: Count of *RESPN...

FIGURE 3.13 Tabulation to compute ratio of frequent flyers to total airport users.

CHAPTER 4

HISTOGRAMS



This chapter also introduces a potent visualization tool for numerical (continuous) variables: the histogram. We want to visualize how a column of numbers in our table is distributed. Do we have a lot of large numbers? A lot of small numbers? Or are they roughly distributed about the mean or average? The tool to visualize this distribution of values is a histogram.

This technique answers the business question: "How is this variable distributed?"

Roughly speaking, we will divide the entire number line for the largest to the smallest value into intervals, count how many of the data points in our column fall into each range, and display the bar graph of the count in each interval. We have now categorized the numerical variable into bins. The bins could be equal in size, or they could differ in size; we could have a lot of bins or a few bins, but binning we must do.

There is a function in Excel called FREQUENCY, but using it is complicated. Fortunately, there is a tool in the Analysis ToolPak called histogram that will automate many of the tasks and do a decent job. Always ask for the chart to be generated. And make sure to modify the resulting bar graph so the bars touch each other. That will make it look distinctively like a histogram and not just a bar graph. It should end up looking like the picture on the first page of the chapter.

As with the previous chapter, we offer two types of exercises: basic exercises for beginners as well as additional and more challenging exercises for advanced students. If you are a beginner, getting through the basic exercises for each tool is a good start. For more advanced students, we offer additional exercises that are more challenging. All are encouraged to try them out as well.

Analysis Case 4.1 – Histograms

Frequency Distributions

1. Using the Lab Data set provided, open the *Analysis Case 4* folder and find the file *ORDERS.xlsx*.
2. Open *ORDERS.xlsx* using Excel.

3. We are going to answer these questions:

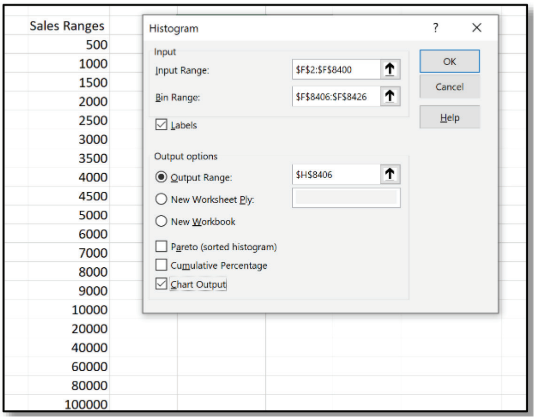
> *Is our sales volume made up of mostly low-priced products, or do we sell even amounts of products across all product prices?*
>
> *Does any range of product pricing account for the bulk of our revenue, or is our revenue evenly distributed among low-priced, mid-priced, and high-priced products?*

4. Following our practice of not changing the raw data, select the entire file, copy it as the shaped file in a new spreadsheet, and label the tab *ORDERS working table*.

5. At the bottom of the Revenue column, compute the MAX of SALES and the MIN of revenue.

6. Create a range of bins for the histogram in $500 increments from the MIN to the MAX in a set of cells next to the shaped table. Select histogram divisions that seem reasonable to you (start with a value of 500 and use even increments from there).

7. Use the histogram function of the Analysis ToolPak or the FREQUENCY function to create the histogram (Figure 4.1).



Sales Ranges
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
6000
7000
8000
9000
10000
20000
40000
60000
80000
100000

Histogram

Input

Input Range: $F$2:$F$8400

Bin Range: $F$8406:$F$8426

☑ Labels

Output options

● Output Range: $H$8406

○ New Worksheet Ply:

○ New Workbook

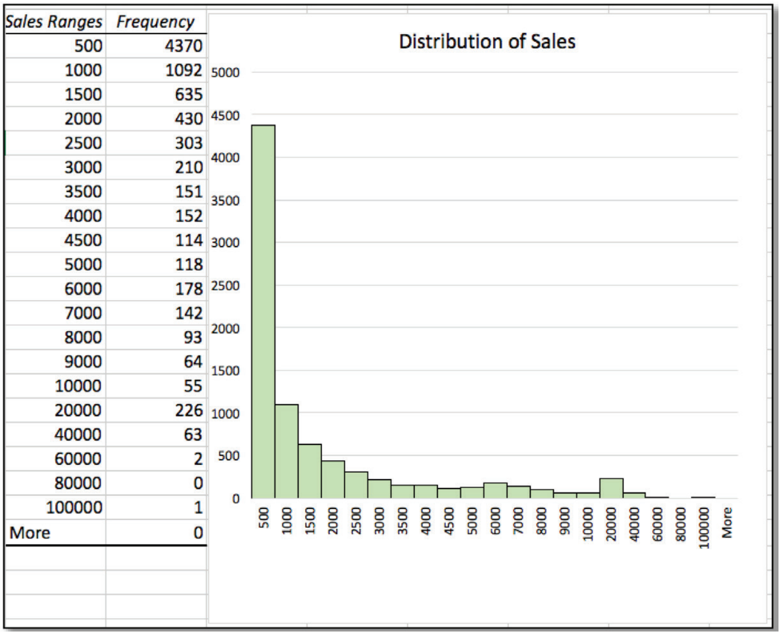
☐ Pareto (sorted histogram)

☐ Cumulative Percentage

☑ Chart Output

FIGURE 4.1 Illustration of the use of the Excel Frequency function in the Analysis ToolPak to create a histogram of sales.

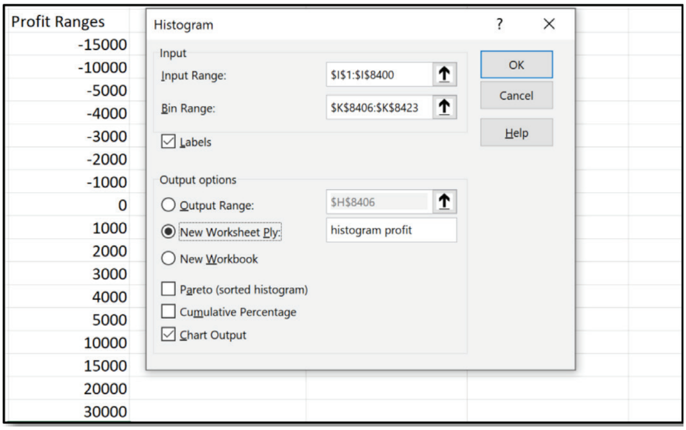
8. What can you say about how the distribution looks (Figure 4.2)?



Sales Ranges	Frequency
500	4370
1000	1092
1500	635
2000	430
2500	303
3000	210
3500	151
4000	152
4500	114
5000	118
6000	178
7000	142
8000	93
9000	64
10000	55
20000	226
40000	63
60000	2
80000	0
100000	1
More	0

FIGURE 4.2 Completed sales histogram.

9. Compute a histogram of profit (Figure 4.3):



Profit Ranges
-15000
-10000
-5000
-4000
-3000
-2000
-1000
0
1000
2000
3000
4000
5000
10000
15000
20000
30000

Histogram

Input

Input Range: $I$1:$I$8400

Bin Range: $K$8406:$K$8423

☑ Labels

Output options

○ Output Range: $H$8406

● New Worksheet Ply: histogram profit

○ New Workbook

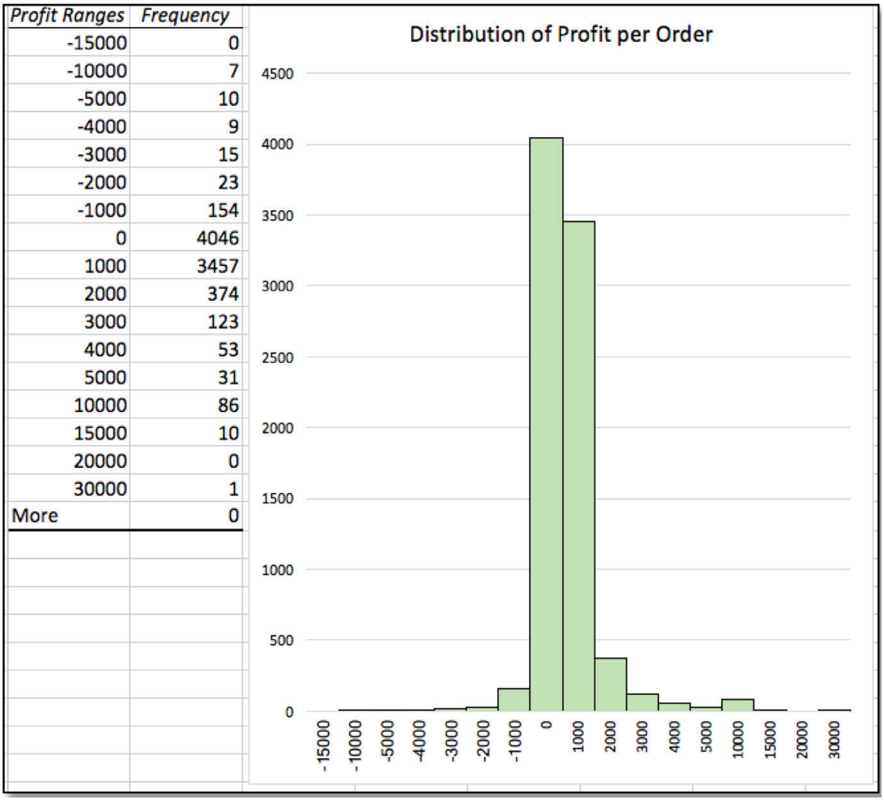
□ Pareto (sorted histogram)

□ Cumulative Percentage

☑ Chart Output

FIGURE 4.3 Illustration of the use of the Excel Frequency function in the Analysis ToolPak to create a histogram of profit.

10. What can you say about the distribution of profits per sale (Figure 4.4)?



Profit Ranges	*Frequency*
-15000	0
-10000	7
-5000	10
-4000	9
-3000	15
-2000	23
-1000	154
0	4046
1000	3457
2000	374
3000	123
4000	53
5000	31
10000	86
15000	10
20000	0
30000	1
More	0

FIGURE 4.4 Completed profit histogram.

Analysis Case 4.2 – Additional Case Using Titanic Data

1. Using the Lab Data set and in the *Analysis Case 4* folder, find the file *Titanic.xlsx*.

2. Open *Titanic.xlsx* using Excel.

3. We are going to answer these questions:

> *What is the age distribution of all passengers?*
>
> *Is there a difference in age distributions by passenger class?*

4. Copy the Titanic data table to a new spreadsheet and label the tab *Passengers*.

5. Sort the table by age and delete all rows without an age entry.

6. Create a histogram of passenger age from the Titanic Passenger data.

7. Sort the table by passenger class. Use the same bin range on the three age ranges by class and compare the histograms. Use ten-year bins. What differences do you see (Figure 4.5)?

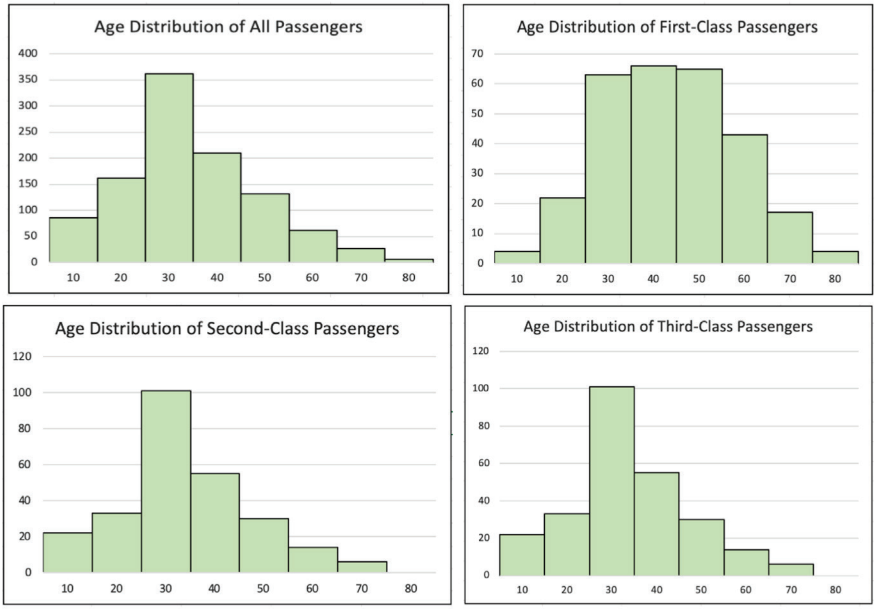


FIGURE 4.5 Histograms of age distribution in ten-year ranges for all passengers on the Titanic, also showing First-, Second-, and Third-class passenger age distributions.

8. Now repeat using five-year ranges. How would you describe the differences (Figure 4.6)?

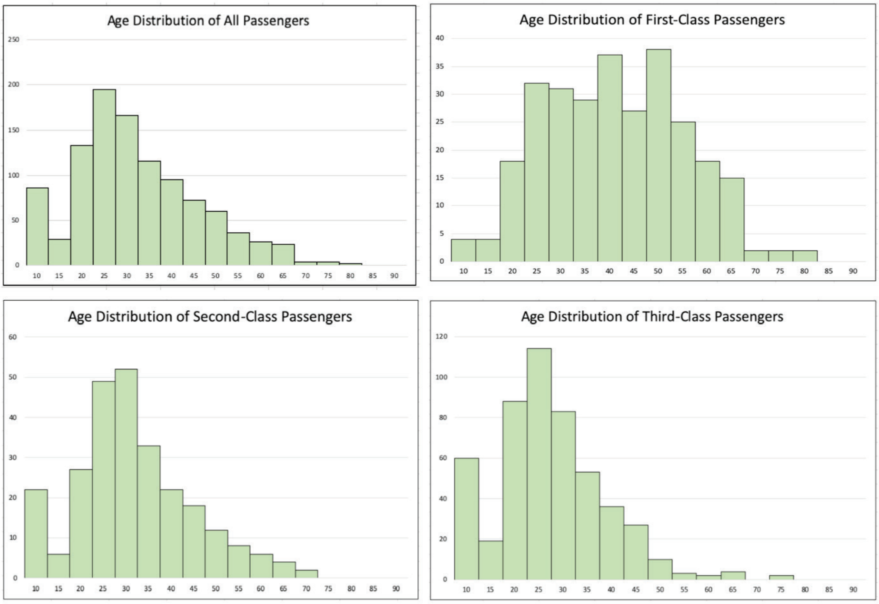
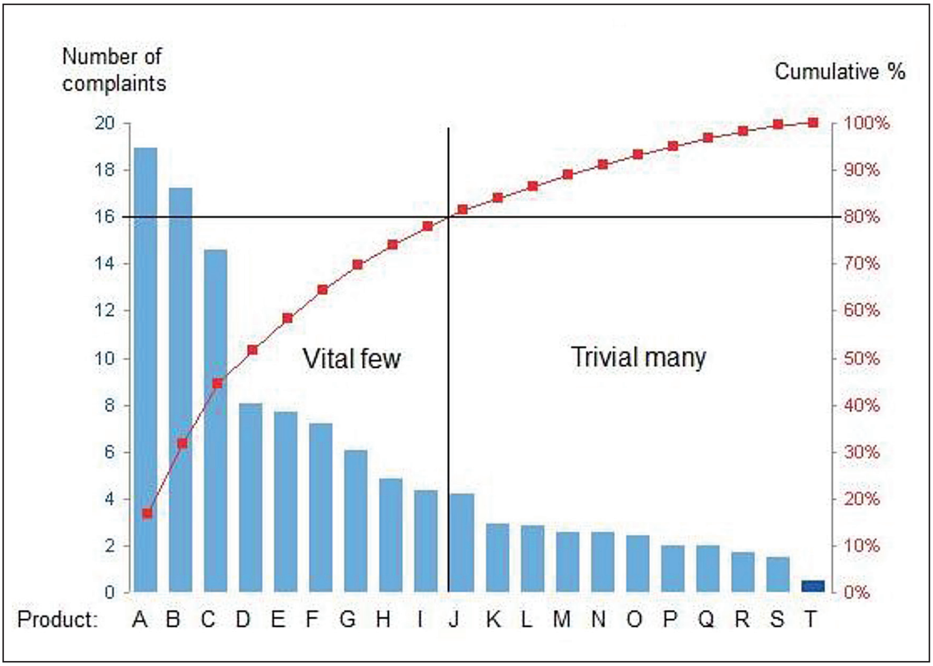


FIGURE 4.6 Histograms of age distribution in five-year ranges for all passengers on the Titanic, also showing First-, Second-, and Third-class passenger age distributions.

CHAPTER 5

PARETO ANALYSIS



This chapter also introduces another powerful business analysis technique for a continuous variable: the Pareto analysis. Consider two columns in a data table, one of them being categories and the other some numerical value associated with those categories. We can ask the following question: "which group of categories, when taken together, contribute the most to the total; for example, considering all the cities in a particular country, which contribute the most to the overall population?" To put it another way, in which urban centers does the majority of the population live? The answer involves computing which of the categories contribute to 80% of the total.

This technique answers the business question: "Which . . . are the most important, are the most populous, bring in the most revenue, are the largest, etc.?"

No function in Excel will produce this directly; it must be computed by hand. Fortunately, the algorithm is simple, and Analysis Case 5.1 demonstrates it, so you can readily reproduce it for other circumstances. Although there is a Pareto analysis included in the histogram function of the Analysis ToolPak, it does not readily apply to most situations, so follow the process in the first exercise. And practice by doing the rest of the Chapter 5 exercises.

Analysis Case 5.1 – Pareto Analysis

Which Are the Most Important?

1. Using the Lab Data set provided, open the *Analysis Case 5* folder in it, and find the file *WDIAnnotatedData.xlsx*.

2. Open *WDIAnnotatedData.xlsx* using Excel.

3. We are going to answer this question:

   Which countries contribute the most to the global Internet user population? (We will compute it for 2012, the latest year for which we have data.)

4. We are going to need population and Internet users per 100 for all countries for the year 2012

5. Following our practice of not changing the raw data, select the appropriate columns and rows, create a shaped file in a new spreadsheet, and label the tab *Pareto*.

6. Delete rows at the bottom of the table that appear after the last individual country (Zimbabwe).

7. Once you have a shaped file with *Country Name, Population,* and *Internet Users per 100*, save the file.

8. Create a variable in an empty column to the right and label it *Internet Users*. Compute and enter into this new column the product of *Population* and *Internet Users per 100* to get the total number of Internet users in each country.

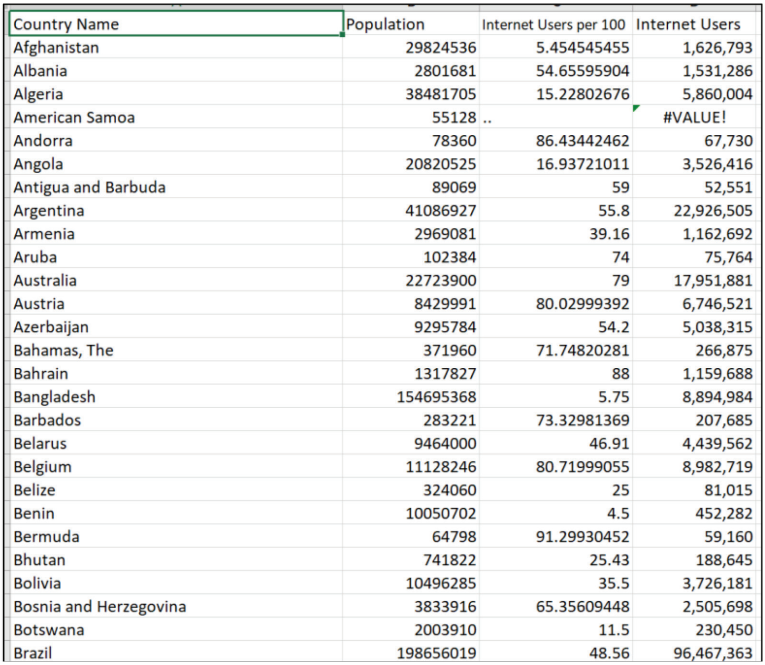
9. Compute the sum total of Internet users and populations and enter it at the bottom of each respective column.

   What is the ratio of Internet users to the total population in the world?

   Are you surprised? (You may need to reformat the cells to easily read the numbers.)

10. Sort the data file by *Internet Users* (Figure 5.1).

11. Create a variable on an empty column on the right and enter a computed variable of percent Internet users for each country with respect to the total number of Internet users for the world. Make the variable a percent with no decimals so it can be read easily.



Country Name	Population	Internet Users per 100	Internet Users
Afghanistan	29824536	5.454545455	1,626,793
Albania	2801681	54.65595904	1,531,286
Algeria	38481705	15.22802676	5,860,004
American Samoa	55128	..	#VALUE!
Andorra	78360	86.43442462	67,730
Angola	20820525	16.93721011	3,526,416
Antigua and Barbuda	89069	59	52,551
Argentina	41086927	55.8	22,926,505
Armenia	2969081	39.16	1,162,692
Aruba	102384	74	75,764
Australia	22723900	79	17,951,881
Austria	8429991	80.02999392	6,746,521
Azerbaijan	9295784	54.2	5,038,315
Bahamas, The	371960	71.74820281	266,875
Bahrain	1317827	88	1,159,688
Bangladesh	154695368	5.75	8,894,984
Barbados	283221	73.32981369	207,685
Belarus	9464000	46.91	4,439,562
Belgium	11128246	80.71999055	8,982,719
Belize	324060	25	81,015
Benin	10050702	4.5	452,282
Bermuda	64798	91.29930452	59,160
Bhutan	741822	25.43	188,645
Bolivia	10496285	35.5	3,726,181
Bosnia and Herzegovina	3833916	65.35609448	2,505,698
Botswana	2003910	11.5	230,450
Brazil	198656019	48.56	96,467,363

FIGURE 5.1 World Bank data set shaped and ready for analysis.

12. On another empty column to the right, create a computed variable where you enter the cumulative distribution function (CDF) results.

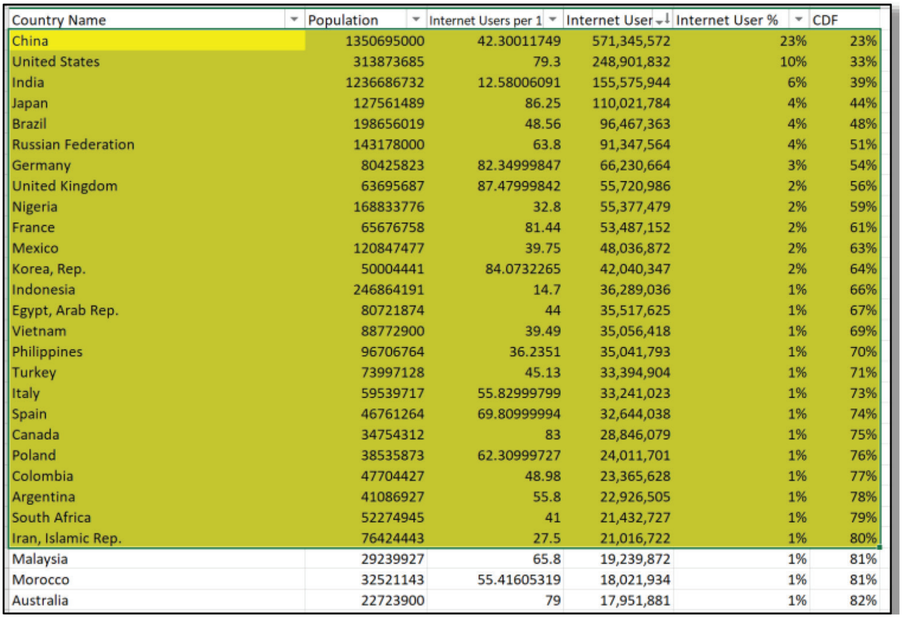
13. Highlight the top rows up to the 80% cumulative result (Figure 5.2).

14. Answer these questions:

> *How many countries does 80% represent? What percentage are they of all countries in the world? Does it fit the 80/20 rule?*

15. Plot the CDF for the first 50 countries. What do you notice? What does it tell you?

16. We see that 25 out of 126 countries (or 20%) contribute 80% of the world's total Internet users. In this case, the 80/20 rule works well (Figure 5.3).



Country Name	Population	Internet Users per 1	Internet User	Internet User %	CDF
China	1350695000	42.30011749	571,345,572	23%	23%
United States	313873685	79.3	248,901,832	10%	33%
India	1236686732	12.58006091	155,575,944	6%	39%
Japan	127561489	86.25	110,021,784	4%	44%
Brazil	198656019	48.56	96,467,363	4%	48%
Russian Federation	143178000	63.8	91,347,564	4%	51%
Germany	80425823	82.34999847	66,230,664	3%	54%
United Kingdom	63695687	87.47999842	55,720,986	2%	56%
Nigeria	168833776	32.8	55,377,479	2%	59%
France	65676758	81.44	53,487,152	2%	61%
Mexico	120847477	39.75	48,036,872	2%	63%
Korea, Rep.	50004441	84.0732265	42,040,347	2%	64%
Indonesia	246864191	14.7	36,289,036	1%	66%
Egypt, Arab Rep.	80721874	44	35,517,625	1%	67%
Vietnam	88772900	39.49	35,056,418	1%	69%
Philippines	96706764	36.2351	35,041,793	1%	70%
Turkey	73997128	45.13	33,394,904	1%	71%
Italy	59539717	55.82999799	33,241,023	1%	73%
Spain	46761264	69.80999994	32,644,038	1%	74%
Canada	34754312	83	28,846,079	1%	75%
Poland	38535873	62.30999727	24,011,701	1%	76%
Colombia	47704427	48.98	23,365,628	1%	77%
Argentina	41086927	55.8	22,926,505	1%	78%
South Africa	52274945	41	21,432,727	1%	79%
Iran, Islamic Rep.	76424443	27.5	21,016,722	1%	80%
Malaysia	29239927	65.8	19,239,872	1%	81%
Morocco	32521143	55.41605319	18,021,934	1%	81%
Australia	22723900	79	17,951,881	1%	82%

FIGURE 5.2 Pareto analysis of 2012 world GDP data showing the countries that contribute the most to the world's economy.

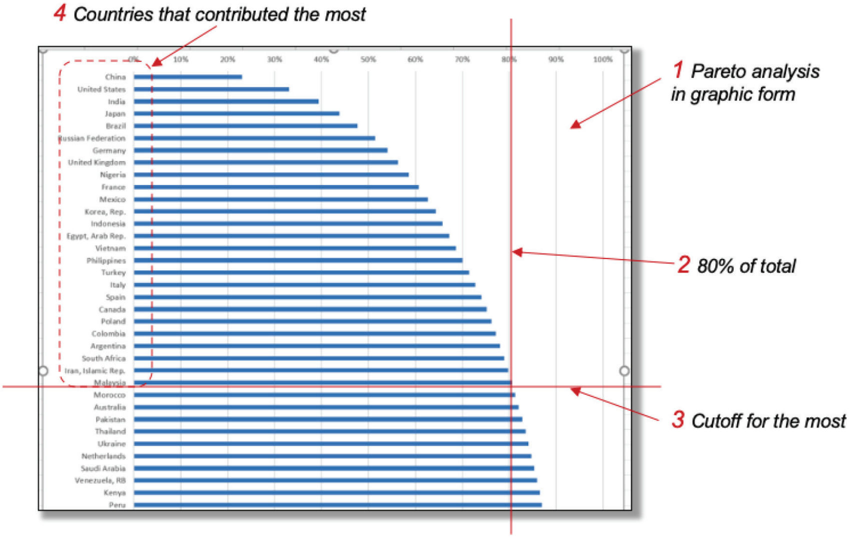


FIGURE 5.3 Details of how the Pareto chart is constructed.

Analysis Case 5.2 – Additional Case Using MOVIES Data

1. Let's do an additional exercise to see another way to interpret a Pareto chart.
2. Using the Lab Data set, open the *Analysis Case 5* folder and find the file *Movies.xlsx*.
3. Open *Movies.xlsx* using Excel.
4. Following our practice of not changing the raw data, select all rows and columns and copy them into the buffer.
5. Paste the data into a new spreadsheet and label the tab *Working Data*.
6. We are going to answer this question:

   Which movies contributed the most revenue to the industry?
7. Let's use total revenue as our criteria.
8. Sort the data file by total revenue.
9. Create total revenue at the bottom of the column.
10. Create a variable on an empty column on the right and enter a computed variable of percent revenue of each movie with respect to the total of all movies. Make the variable a percent with no decimals so it can be read easily.
11. On another empty column to the right, create a computed variable where you enter the cumulative distribution results.
12. Highlight the top rows up to the 80% cumulative result.
13. Plot the CDF (Figure 5.4). What do you notice? What does it tell you?

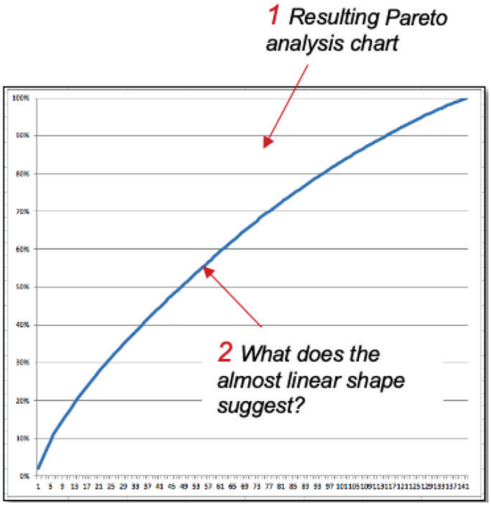
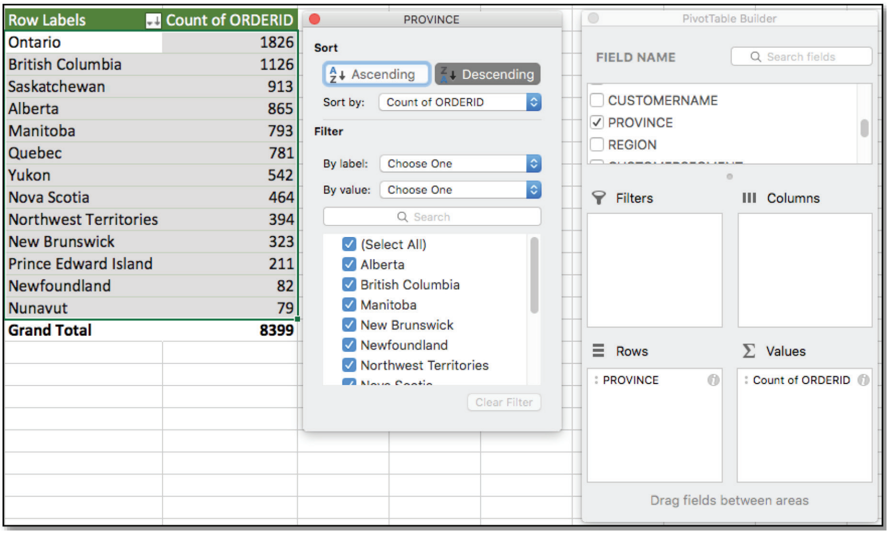


FIGURE 5.4 Illustration of a situation where the Pareto analysis leads to the conclusion that there is no significant few.

Analysis Case 5.3 – Additional Case Using ORDERS

1. Using the Lab Data set and in the *Analysis Case* 5 folder, find the file *ORDERS.xlsx*.
2. Open *ORDERS.xlsx* using Excel.
3. We are going to answer this question:

   Which provinces generated the most orders?
4. Create a Pivot Table of orders by province to tabulate the amount for each.
5. Sort the Pivot Table provinces by the greatest number of orders first (Figure 5.5).
6. Note the computed total of all orders.



Row Labels	Count of ORDERID
Ontario	1826
British Columbia	1126
Saskatchewan	913
Alberta	865
Manitoba	793
Quebec	781
Yukon	542
Nova Scotia	464
Northwest Territories	394
New Brunswick	323
Prince Edward Island	211
Newfoundland	82
Nunavut	79
Grand Total	**8399**

PROVINCE

Sort

Sort by: Count of ORDERID

Filter

By label: Choose One

By value: Choose One

☑ (Select All)
☑ Alberta
☑ British Columbia
☑ Manitoba
☑ New Brunswick
☑ Newfoundland
☑ Northwest Territories

PivotTable Builder

FIELD NAME

□ CUSTOMERNAME
☑ PROVINCE
□ REGION

Rows: PROVINCE

Values: Count of ORDERID

FIGURE 5.5 How to set up the Pivot Table to tabulate orders by province.

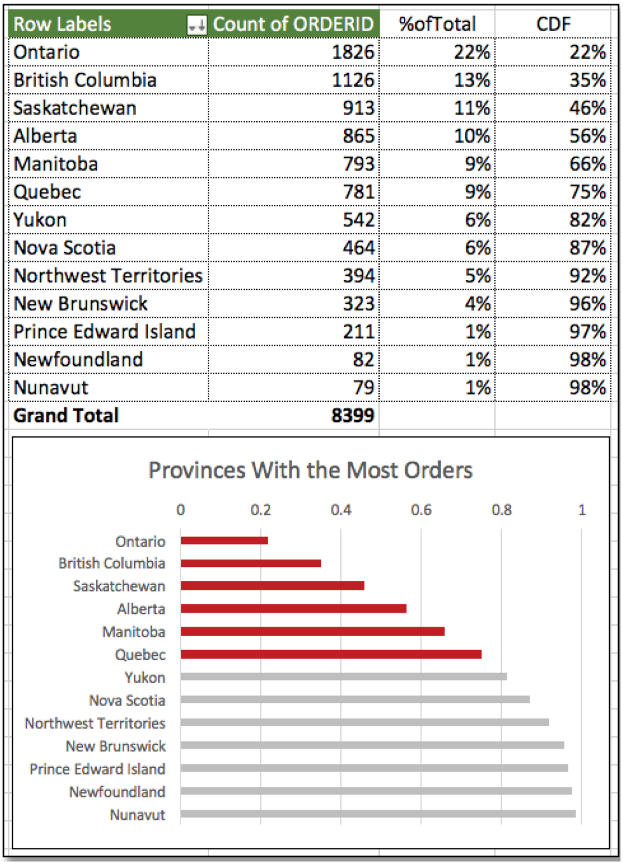
7. Let's create a Pareto chart using the count of orders by province.

8. Create a new variable next to the Pivot Table and enter the percentage of each province's orders into it.

9. Then, compute a second new variable as the cumulative distribution function.

10. Select the provinces that account for 80% of all orders. Is that 20% of all provinces? Do you see a knee in the curve (the region where the curve goes from steep increase to where it starts to flatten out)?

11. Can you identify the provinces with the "most" orders (Figure 5.6)?



Row Labels	Count of ORDERID	%ofTotal	CDF
Ontario	1826	22%	22%
British Columbia	1126	13%	35%
Saskatchewan	913	11%	46%
Alberta	865	10%	56%
Manitoba	793	9%	66%
Quebec	781	9%	75%
Yukon	542	6%	82%
Nova Scotia	464	6%	87%
Northwest Territories	394	5%	92%
New Brunswick	323	4%	96%
Prince Edward Island	211	1%	97%
Newfoundland	82	1%	98%
Nunavut	79	1%	98%
Grand Total	**8399**		

FIGURE 5.6 A Pareto analysis showing which provinces contribute the most orders to the total.

Analysis Case 5.4 – Additional Case Using SFO Airport Survey Data

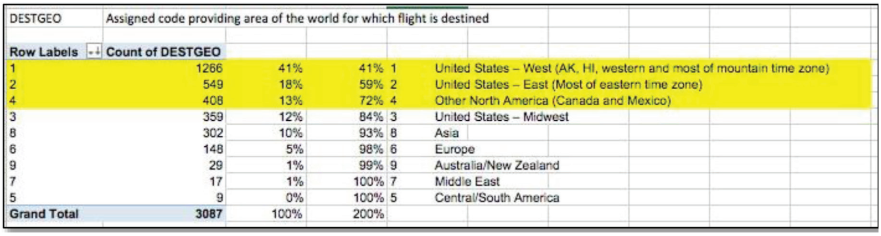
1. Download the latest SFO Airport ACQ Survey data from *https://www.flysfo.com/media/customer-survey-data*.
2. Or using the Lab Data set provided in the *Analysis Case* 5 folder, find the file *2016_SFO_Customer_Survey_Data*.xls.
3. Open the data dictionary file and have it available to consult as you work with the data.

4. Open the data file using Excel.

5. We are going to answer the following questions:

> *What are the most frequent destinations of travelers using SFO? What are the most frequent reasons passengers are traveling?*
>
> *What are passengers' most frequent complaints about airport cleanliness?*

6. For the first question, create a Pivot Table by DESTGEO and tabulate the number of comments by comment code, filter out the no comment (code = 0) rows, and sort rows by most frequent count. Then, perform a Pareto analysis to discover the most frequent destination (Figure 5.7).

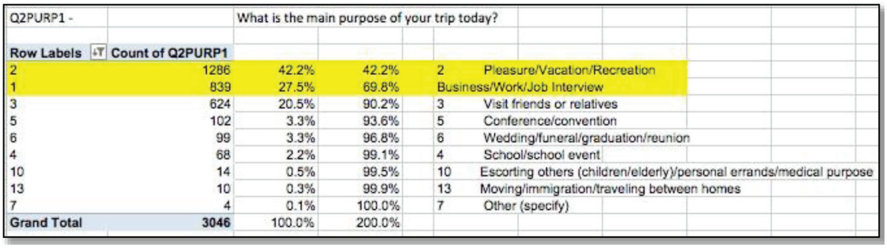


DESTGEO	Assigned code providing area of the world for which flight is destined				
Row Labels	Count of DESTGEO				
1	1266	41%	41%	1	United States – West (AK, HI, western and most of mountain time zone)
2	549	18%	59%	2	United States – East (Most of eastern time zone)
4	408	13%	72%	4	Other North America (Canada and Mexico)
3	359	12%	84%	3	United States – Midwest
8	302	10%	93%	8	Asia
6	148	5%	98%	6	Europe
9	29	1%	99%	9	Australia/New Zealand
7	17	1%	100%	7	Middle East
5	9	0%	100%	5	Central/South America
Grand Total	3087	100%	200%		

FIGURE 5.7 Using a Pivot Table to tabulate count of passengers by DESTGEO and the resulting Pareto analysis to determine the most frequent destinations.

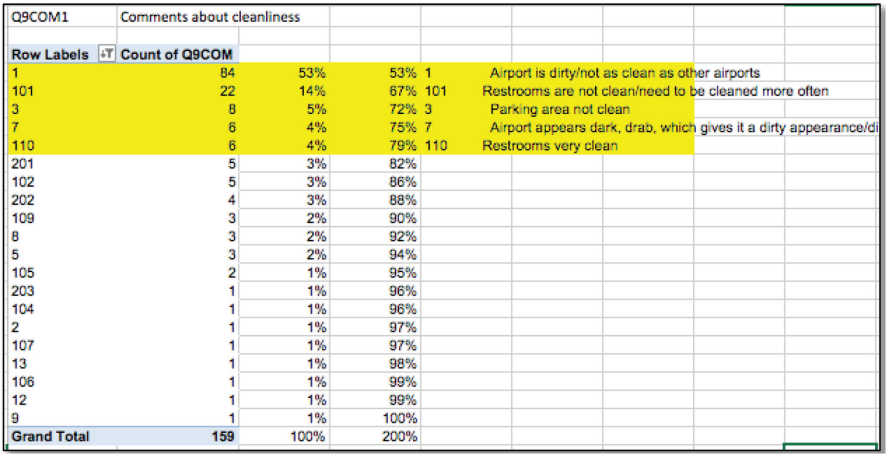
7. For the next question, create a Pivot Table by Q2 comments, Q2COMM, and tabulate the number of comments by comment code; filter out the nocomment (code = 0) rows; and sort rows by most frequent count. Then, perform a Pareto analysis to discover the most frequent reasons for their trips.

8. For the last question, create a Pivot Table by Q9 comments, Q9COM1, and tabulate the number of comments by comment code; filter out the no comment (code = 0) rows; and sort rows by most frequent count. Then, perform a Pareto analysis to discover the most frequent complaints about cleanliness (Figure 5.8).



Q2PURP1 -		What is the main purpose of your trip today?			
Row Labels	Count of Q2PURP1				
2	1286	42.2%	42.2%	2	Pleasure/Vacation/Recreation
1	839	27.5%	69.8%	Business/Work/Job Interview	
3	624	20.5%	90.2%	3	Visit friends or relatives
5	102	3.3%	93.6%	5	Conference/convention
6	99	3.3%	96.8%	6	Wedding/funeral/graduation/reunion
4	68	2.2%	99.1%	4	School/school event
10	14	0.5%	99.5%	10	Escorting others (children/elderly)/personal errands/medical purpose
13	10	0.3%	99.9%	13	Moving/immigration/traveling between homes
7	4	0.1%	100.0%	7	Other (specify)
Grand Total	3046	100.0%	200.0%		

FIGURE 5.8 Pivot Table and Pareto analysis to discover the most frequent type of airport visitor.



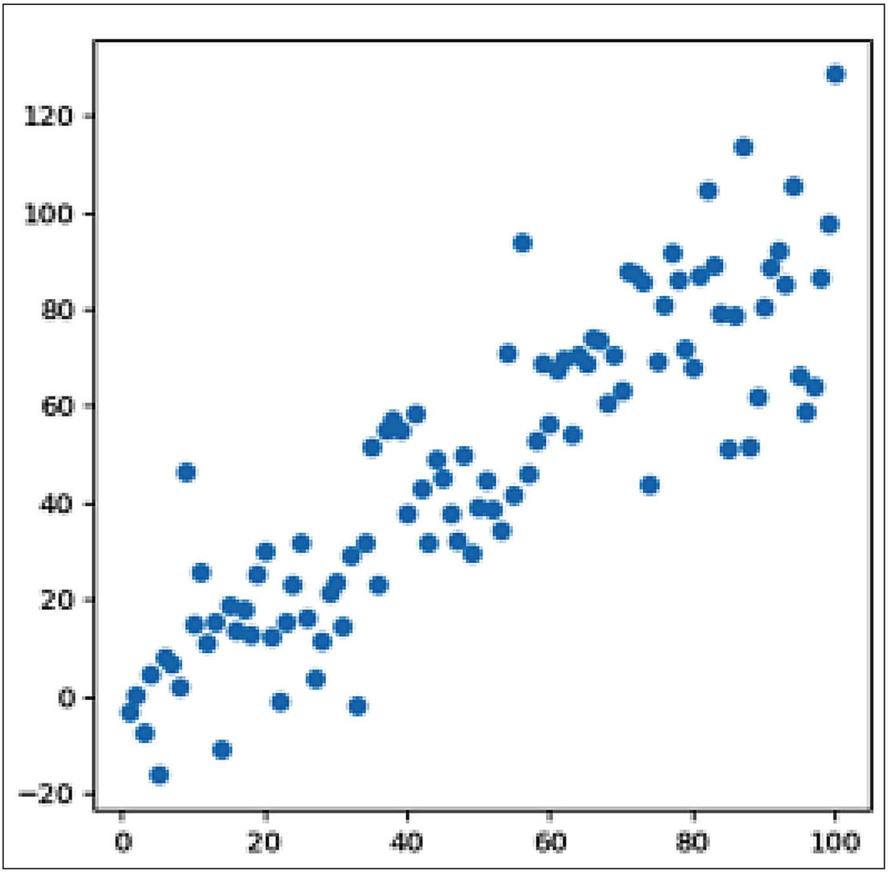
Q9COM1	Comments about cleanliness				
Row Labels	Count of Q9COM				
1	84	53%	53%	1	Airport is dirty/not as clean as other airports
101	22	14%	67%	101	Restrooms are not clean/need to be cleaned more often
3	8	5%	72%	3	Parking area not clean
7	6	4%	75%	7	Airport appears dark, drab, which gives it a dirty appearance/di
110	6	4%	79%	110	Restrooms very clean
201	5	3%	82%		
102	5	3%	86%		
202	4	3%	88%		
109	3	2%	90%		
8	3	2%	92%		
5	3	2%	94%		
105	2	1%	95%		
203	1	1%	96%		
104	1	1%	96%		
2	1	1%	97%		
107	1	1%	97%		
13	1	1%	98%		
106	1	1%	99%		
12	1	1%	99%		
9	1	1%	100%		
Grand Total	159	100%	200%		

FIGURE 5.9 Pareto analysis to discover the most frequent complaints about cleanliness.

9. Make sure to refer to the data dictionary (you may need to use the Word version) to capture the meanings of the codes to annotate your results (Figure 5.9).

CHAPTER 6

SCATTER PLOTS



Chapter 6 introduces a business data analysis technique to compare two continuous variables. Consider three columns in a data table, one of them a categorical variable and the other two being some numerical values related to each other. We will use these two variables to select which category optimizes the other two, and we can ask the following question: "Which category is best along these two dimensions?" We produce a powerful decision graphic called a 2X2 diagram. We create it such that the coordinates of the "best" categories always show up in the upper right-hand quadrant of the scatter plot.

Consider an example. We want to buy a car and have many different models and brands from which to choose. Those are our categories. We compile a detailed table of many features for these choices. We then select two of the numerical columns in the table, such as gas mileage and cost, which will help us choose the "best" option. We create a scatter plot of gas mileage (x-axis) and cost (y-axis) and label each data point on our graph with each category (brand). Manipulating the display and making sure we plot the cost variable in reverse order (low cost is better, so it must show up in the upper right-hand quadrant), we produce a decision tool to show us the "best" choices.

This technique answers the business question: "Which are the best choices?"

As in previous chapters, we demonstrate the technique in the first exercise and allow for more challenging work in subsequent exercises.

Analysis Case 6.1 – Scatter Plots and 2x2 Analysis

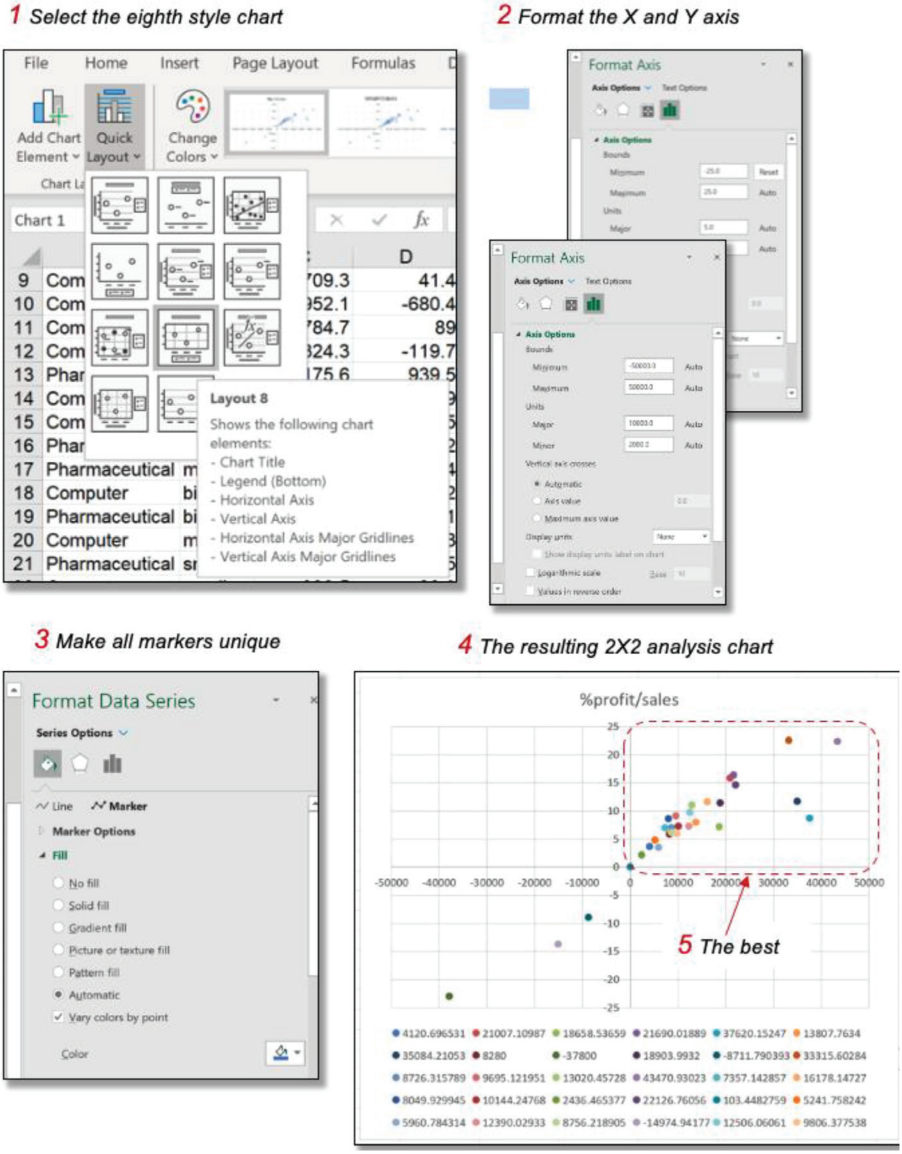
Which Are the "Best"?

1. Using the Lab Data set provided, open the *Analysis Case 6* folder and find the file *Companies.xlsx*.
2. Open *Companies.xlsx* using Excel.

3. We are going to answer this question:

   Which companies have the best profit per employee and profit per sales ratios?

4. Following our practice of not changing the raw data, select the entire file, copy it as the shaped file into a new spreadsheet, and label the tab *Best Rates*.
5. Shape the file such that profit/employee and profit/sales are in two columns next to each other.
6. Select the two data columns and insert a scatter plot chart (the first one in the list).
7. Compute the maximum and minimum of each column in cells below the columns. This gives us rough extremes for our 2x2 plot.
8. Under the Chart ribbon in the "Charts Quick Layouts," select the one with both vertical and horizontal grid lines.
9. Right-click on the x-axis, and using "Format Axis" change the range to go from +50,000 to −50,000 in major units of 50,000. This gives us only one major division.
10. Repeat for the y-axis to go from −25 to +25 in major units of 25.
11. Click on one of the data points and, making sure that all data points are selected, right-click to select "Format Data Series."
12. Under "Marker Fill," select "Vary color by point." See Figure 6.1 for an outline of the process and the finished product.



1 Select the eighth style chart

Layout 8

Shows the following chart elements:
- Chart Title
- Legend (Bottom)
- Horizontal Axis
- Vertical Axis
- Horizontal Axis Major Gridlines
- Vertical Axis Major Gridlines

2 Format the X and Y axis

Format Axis

Axis Options

Bounds

Minimum -25.0 Reset

Maximum 25.0 Auto

Units

Major 5.0 Auto

Format Axis

Axis Options

Bounds

Minimum -50000.0 Auto

Maximum 50000.0 Auto

Units

Major 10000.0 Auto

Minor 2000.0 Auto

Vertical axis crosses

● Automatic

○ Axis value 0.0

○ Maximum axis value

Display units None

□ Show display units label on chart

□ Logarithmic scale Base 10

□ Values in reverse order

3 Make all markers unique

Format Data Series

Series Options

Marker Options

Fill

○ No fill

○ Solid fill

○ Gradient fill

○ Picture or texture fill

○ Pattern fill

● Automatic

☑ Vary colors by point

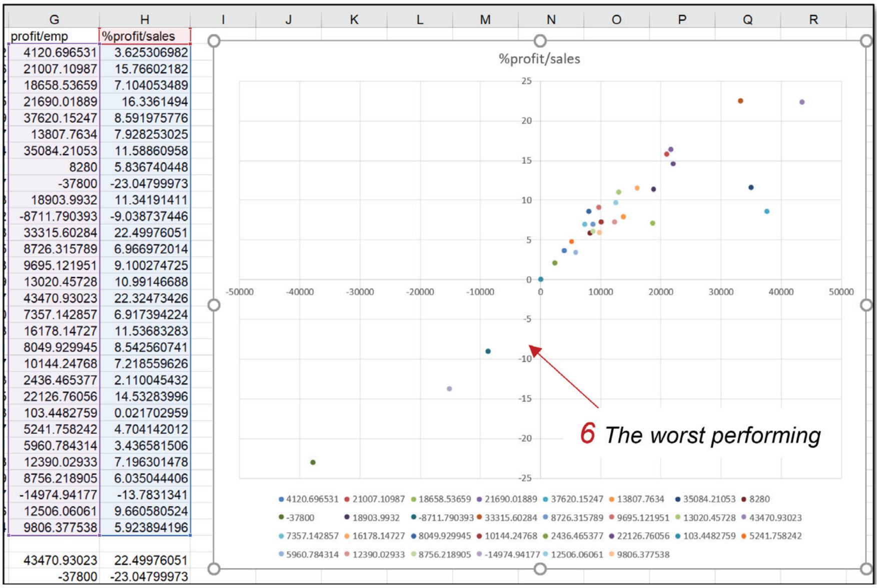
4 The resulting 2X2 analysis chart

5 The best

FIGURE 6.1 Steps in converting a scatter plot into a decision support 2X2 chart to determine the “best.”

13. Now let's look for the reverse (Figure 6.2):

Which are the worst-performing companies?



G	H
profit/emp	%profit/sales
4120.696531	3.625306982
21007.10987	15.76602182
18658.53659	7.104053489
21690.01889	16.3361494
37620.15247	8.591975776
13807.7634	7.928253025
35084.21053	11.58860958
8280	5.836740448
-37800	-23.04799973
18903.9932	11.34191411
-8711.790393	-9.038737446
33315.60284	22.49976051
8726.315789	6.966972014
9695.121951	9.100274725
13020.45728	10.99146688
43470.93023	22.32473426
7357.142857	6.917394224
16178.14727	11.53683283
8049.929945	8.542560741
10144.24768	7.218559626
2436.465377	2.110045432
22126.76056	14.53283996
103.4482759	0.021702959
5241.758242	4.704142012
5960.784314	3.436581506
12390.02933	7.196301478
8756.218905	6.035044406
-14974.94177	-13.7831341
12506.06061	9.660580524
9806.377538	5.923894196
43470.93023	22.49976051
-37800	-23.04799973

FIGURE 6.2 Results of converting a scatter plot into a decision support 2X2 chart to determine the "worst performing."

Analysis Case 6.2 – Additional Case Using World Bank Data

1. Using the Lab Data set and in the *Analysis Case 6* folder, find the file *WDIAnnotatedData.xlsx*.

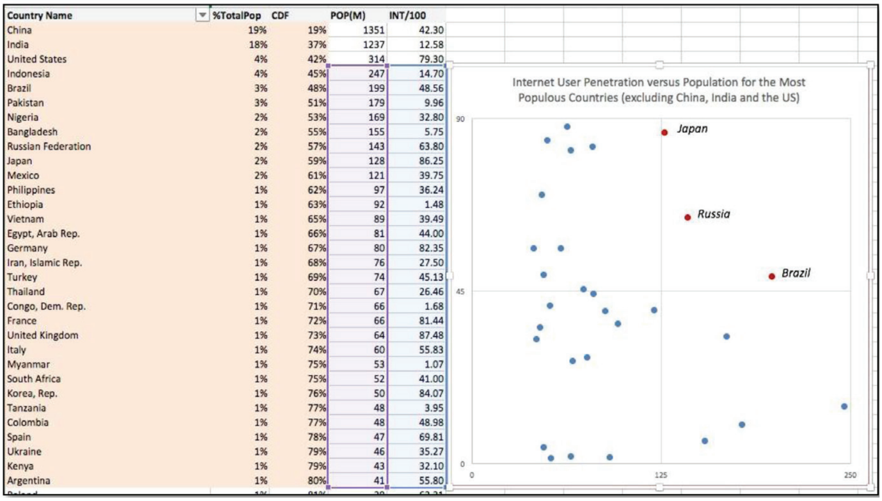
2. Open *WDIAnnotatedData.xlsx* using Excel.

3. We are going to answer these questions:

Which of the most populous countries have the largest Internet penetration (exclude the top three countries, China, India, and the U.S.)?

Which of the emerging economies (E7 countries) have the best Internet penetration and the largest population?

4. We are going to need population and Internet users per 100 for all countries for the year 2012.
5. Create a Pareto chart of the most populous countries.
6. Prepare a 2x2 chart of Internet users per 100 and population in millions to determine the "best" countries (Figure 6.3).

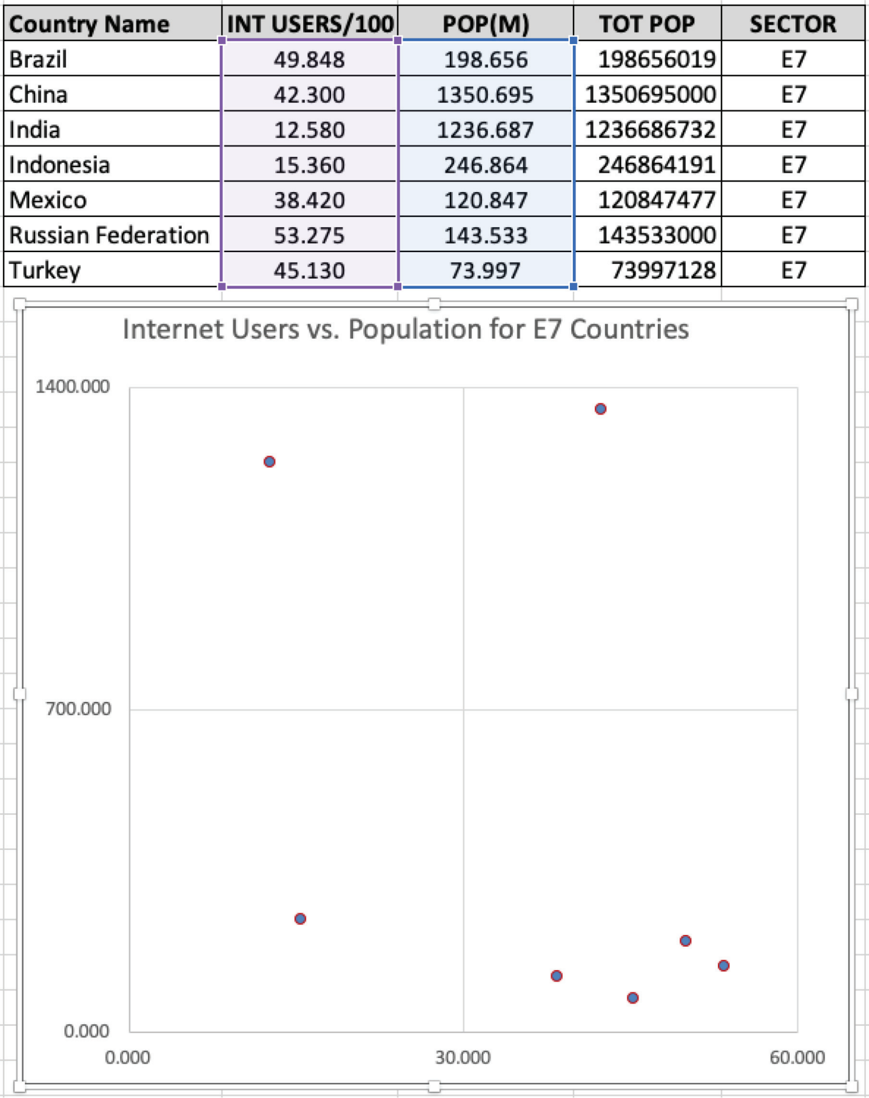


Country Name	%TotalPop	CDF	POP(M)	INT/100
China	19%	19%	1351	42.30
India	18%	37%	1237	12.58
United States	4%	42%	314	79.30
Indonesia	4%	45%	247	14.70
Brazil	3%	48%	199	48.56
Pakistan	3%	51%	179	9.96
Nigeria	2%	53%	169	32.80
Bangladesh	2%	55%	155	5.75
Russian Federation	2%	57%	143	63.80
Japan	2%	59%	128	86.25
Mexico	2%	61%	121	39.75
Philippines	1%	62%	97	36.24
Ethiopia	1%	63%	92	1.48
Vietnam	1%	65%	89	39.49
Egypt, Arab Rep.	1%	66%	81	44.00
Germany	1%	67%	80	82.35
Iran, Islamic Rep.	1%	68%	76	27.50
Turkey	1%	69%	74	45.13
Thailand	1%	70%	67	26.46
Congo, Dem. Rep.	1%	71%	66	1.68
France	1%	72%	66	81.44
United Kingdom	1%	73%	64	87.48
Italy	1%	74%	60	55.83
Myanmar	1%	75%	53	1.07
South Africa	1%	75%	52	41.00
Korea, Rep.	1%	76%	50	84.07
Tanzania	1%	77%	48	3.95
Colombia	1%	77%	48	48.98
Spain	1%	78%	47	69.81
Ukraine	1%	79%	46	35.27
Kenya	1%	79%	43	32.10
Argentina	1%	80%	41	55.80

Internet User Penetration versus Population for the Most Populous Countries (excluding China, India and the US)
90
45
0
0
125
250
Japan
Russia
Brazil

FIGURE 6.3 Elements of converting a scatter plot into a 2X2 chart to determine which countries "have the largest population with the largest internet penetration."

7. Use the Internet to discover which countries are considered the "emerging economies"—the so-called "E7 countries."
8. Prepare a 2x2 chart of Internet users per 100 and population in millions to determine the "best" E7 country or countries (Figure 6.4).
9. Since Excel does not allow you to label the countries directly, the way to identify the data point is by its coordinates and referring back to the table. Can you tell which country fits the "best" criteria?



Country Name	INT USERS/100	POP(M)	TOT POP	SECTOR
Brazil	49.848	198.656	198656019	E7
China	42.300	1350.695	1350695000	E7
India	12.580	1236.687	1236686732	E7
Indonesia	15.360	246.864	246864191	E7
Mexico	38.420	120.847	120847477	E7
Russian Federation	53.275	143.533	143533000	E7
Turkey	45.130	73.997	73997128	E7

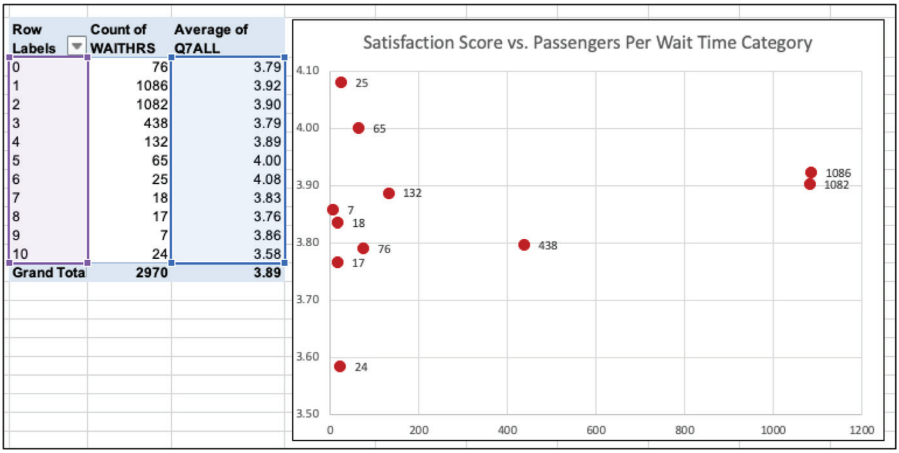
FIGURE 6.4 Steps in converting a scatter plot into a 2X2 chart to determine the E7 countries "with the largest population and with the largest Internet penetration."

Analysis Case 6.3 – Additional Case Using SFO Airport Survey Data

1. Use the latest SFO Airport ACQ Survey data downloaded from *https://www.flysfo.com/media/customer-survey-data*.
2. Or using the Lab Data set provided in the *Analysis Case 6* folder, find the file *2016_SFO_Customer_Survey_Data.xls*.
3. Open the data dictionary file and have it available to consult as you work with the data.
4. Open the data file using Excel.
5. We are going to answer the following question:

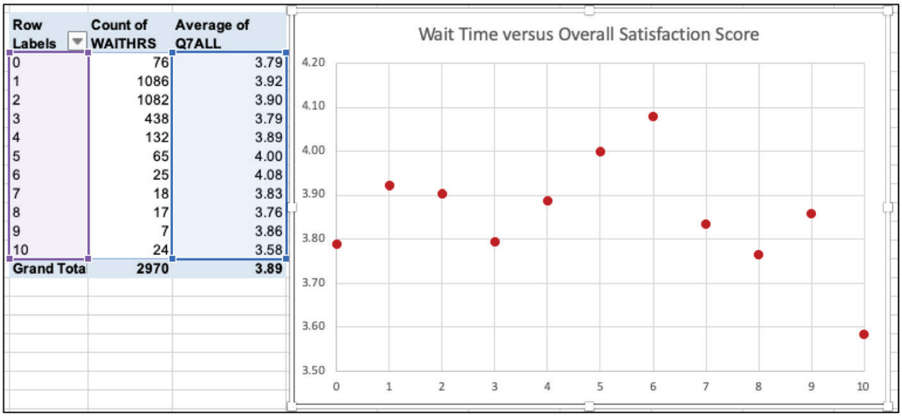
   Is there a relationship between how long a passenger waited between flights and his/her overall satisfaction score?

6. Select the waiting time between flights and the overall satisfaction score (consult the data dictionary).
7. Transfer the two columns of data to another spreadsheet.
8. Normalize the wait time to hours (divide by 60).
9. Make sure all columns are in numeric format and sort the columns by wait time. Delete all rows with nonnumerical data.
10. Create a column coding the wait time in hours to an integer value. Round it up using the ROUND function.
11. Create a Pivot Table and tabulate the average satisfaction score and the total number of passengers in each whole hour by wait time categories.
12. Create scatter plots of wait time versus customer satisfaction scores and the total number of customers in each category versus satisfaction scores. Can you answer the question now (Figures 6.5 and 6.6)?



Row Labels	Count of WAITHRS	Average of Q7ALL
0	76	3.79
1	1086	3.92
2	1082	3.90
3	438	3.79
4	132	3.89
5	65	4.00
6	25	4.08
7	18	3.83
8	17	3.76
9	7	3.86
10	24	3.58
Grand Tota	2970	3.89

FIGURE 6.5 Scatter plot of how long a passenger waited between flights and their overall satisfaction score.

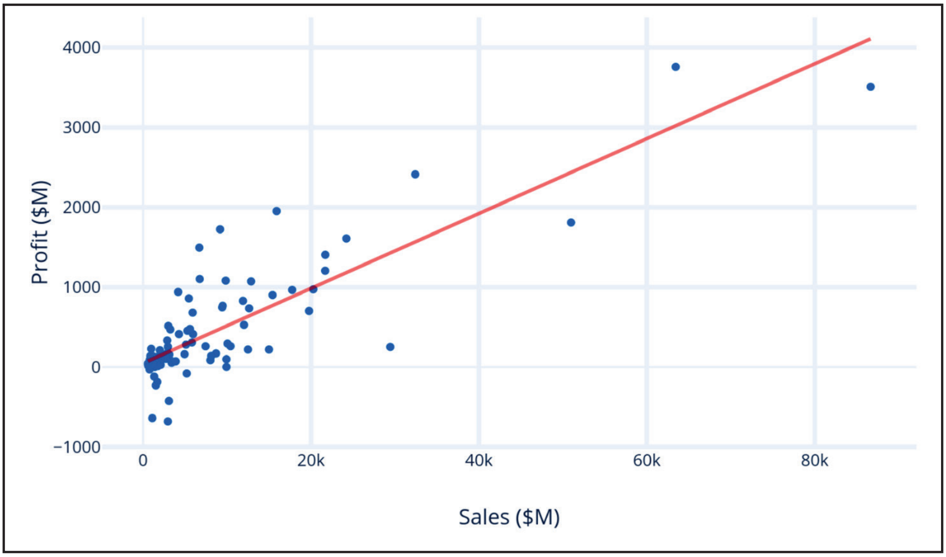


Row Labels	Count of WAITHRS	Average of Q7ALL
0	76	3.79
1	1086	3.92
2	1082	3.90
3	438	3.79
4	132	3.89
5	65	4.00
6	25	4.08
7	18	3.83
8	17	3.76
9	7	3.86
10	24	3.58
Grand Tota	2970	3.89

FIGURE 6.6 Scatter plot of wait-time category and overall satisfaction score.

CHAPTER 7

CORRELATION AND LINEAR REGRESSION



Chapter 7 introduces predictive analytics and the use of a simple linear regression model. Consider two columns in a data table of numerical data. We may ask the question: "how related are these two variables to each other?" We compute the correlation coefficient for that. If we find a strong relationship, then we can further ask: "Can we use one variable to predict the other?" The predicting variable (usually plotted on the x axis) is sometimes called the predictor or independent variable. The variable being predicted (plotted on the y axis) is also known as the independent variable. We will make extensive use of the Analysis ToolPak functions of Correlation and Regression for these exercises.

Consider an example. We know the gross annual sales of all the stores in our franchise. We also know the annual profit from each store. Are these two related? How strongly related? And could we somehow use the data to build a simple linear model that would generate the annual predicted profit from a store if I knew its annual sales? In Chapter 9, we will extend this model to time series, where the x variable is a date, and we will be able to perform trend analysis and build forecasts. And in Chapter 8 we will use many more x variables (multiple input variables) to predict sales in a multivariate regression model.

This technique answers the business questions: "How are two numerical variables related?" and "Can we use one variable to predict another?"

As in previous chapters, we demonstrate the technique in the first exercise and allow for more challenging work in subsequent exercises.

Analysis Case 7.1 – Correlation and Linear Regression

How Are the Variables Related?

1. Using the Lab Data set provided, open the *Analysis Case 7* folder and find the file *Companies.xlsx*.
2. Open *Companies.xlsx* using Excel.

3. We are going to answer these questions:

   Is there a relationship between assets and profit/ sales for this group of companies?

   Is there a relationship between sales and profits?

4. Following our practice of not changing the raw data, select the entire file and copy it as the shaped file into a new spreadsheet and label the tab *Correlation Data*.

5. Use the Analysis ToolPak to compute the correlation matrix between all the continuous variables.

6. Select all the continuous variable data columns. Make sure to select the row headers as well.

7. Go to the "Data" ribbon and select "Data Analysis" and then "Correlation" from the Analysis ToolPak menu (Figure 7.1).

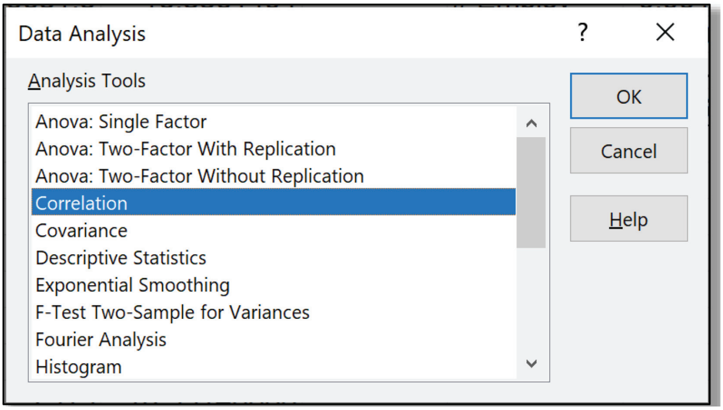
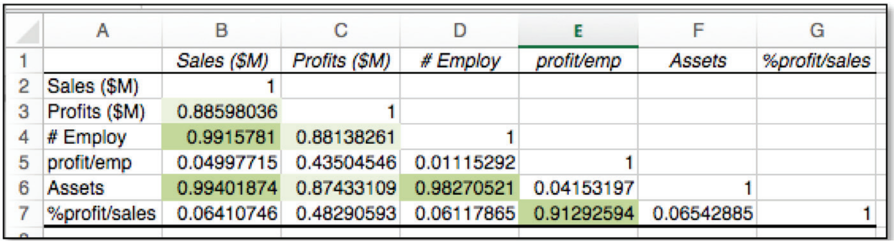


FIGURE 7.1 Analysis ToolPak wizard screen showing location of the Correlation function.

8. Accept the data selection, make sure you analyze by columns, and then place the results in another spreadsheet. Make sure to check the "Labels in First Row" box.

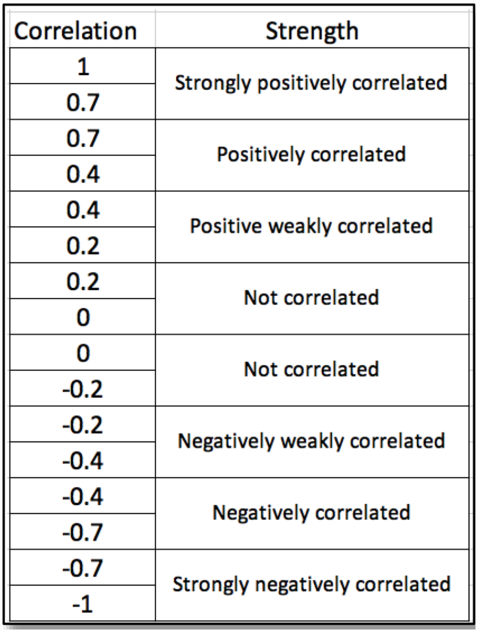
9. Which rows are correlated, and which do not appear to be correlated? Is this something you can explain (Figure 7.2)?



	A	B	C	D	E	F	G
1		*Sales ($M)*	*Profits ($M)*	*# Employ*	*profit/emp*	*Assets*	*%profit/sales*
2	Sales ($M)	1					
3	Profits ($M)	0.88598036	1				
4	# Employ	0.9915781	0.88138261	1			
5	profit/emp	0.04997715	0.43504546	0.01115292	1		
6	Assets	0.99401874	0.87433109	0.98270521	0.04153197	1	
7	%profit/sales	0.06410746	0.48290593	0.06117865	0.91292594	0.06542885	1

FIGURE 7.2 Correlation between assets and profit/sales for this group of companies.

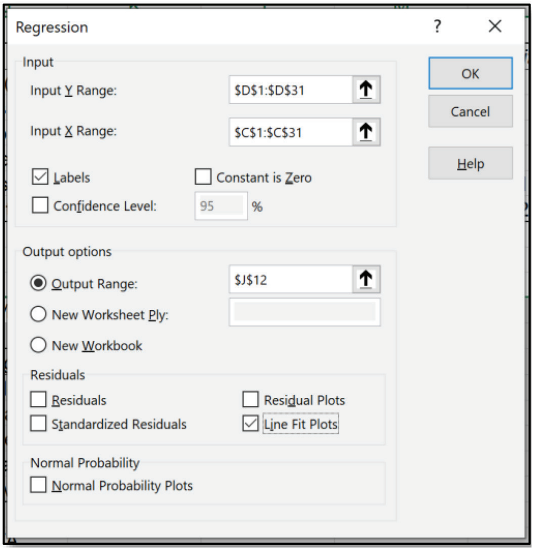
10. If you are wondering how we interpret correlations, here is a chart that might help (Figure 7.3). This is how we speak about various levels of correlation. Be mindful these are not hard and fast ranges but working definitions.



Correlation	Strength
1	Strongly positively correlated
0.7	
0.7	Positively correlated
0.4	
0.4	Positive weakly correlated
0.2	
0.2	Not correlated
0	
0	Not correlated
-0.2	
-0.2	Negatively weakly correlated
-0.4	
-0.4	Negatively correlated
-0.7	
-0.7	Strongly negatively correlated
-1	

FIGURE 7.3 Categories of correlation strength (approximate).

11. Now let's compute the linear relationship between sales and profit. Go to the "Data" ribbon and select "Data Analysis" and then "Regression" from the Analysis ToolPak menu (Figure 7.4).



Input

Input Y Range: $D$1:$D$31

Input X Range: $C$1:$C$31

☑ Labels
□ Constant is Zero

□ Confidence Level: 95 %

Output options

● Output Range: $J$12

○ New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals
□ Residual Plots

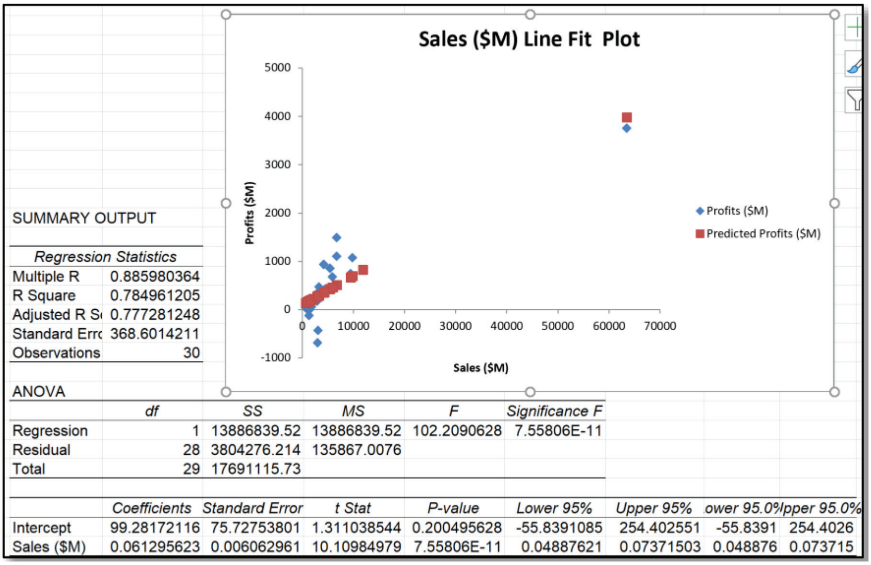
□ Standardized Residuals
☑ Line Fit Plots

Normal Probability

□ Normal Probability Plots

FIGURE 7.4 Analysis ToolPak wizard setting to perform a linear regression analysis.

12. What is the linear regression between sales and profits (Figure 7.5)?



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.885980364
R Square	0.784961205
Adjusted R Sq	0.777281248
Standard Erro	368.6014211
Observations	30

ANOVA

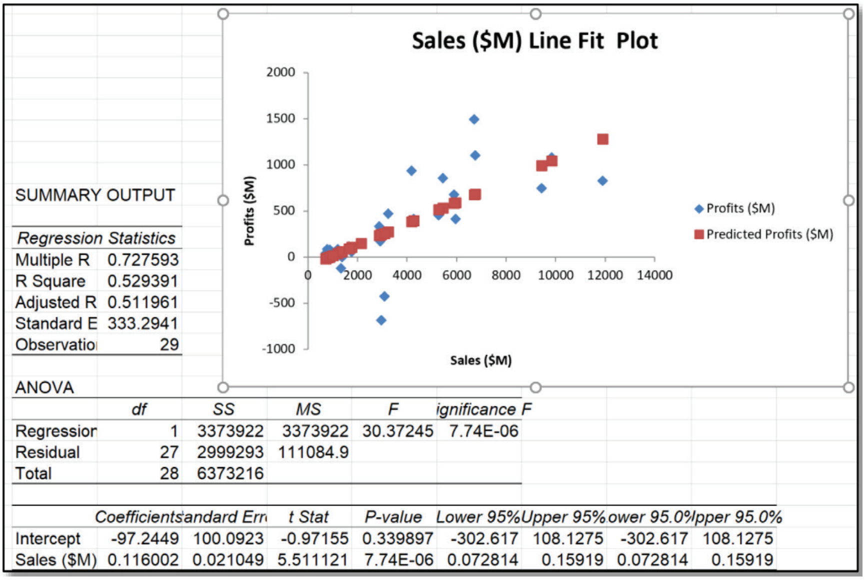
	df	SS	MS	F	Significance F
Regression	1	13886839.52	13886839.52	102.2090628	7.55806E-11
Residual	28	3804276.214	135867.0076		
Total	29	17691115.73			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	lpper 95.0%
Intercept	99.28172116	75.72753801	1.311038544	0.200495628	-55.8391085	254.402551	-55.8391	254.4026
Sales ($M)	0.061295623	0.006062961	10.10984979	7.55806E-11	0.04887621	0.07371503	0.048876	0.073715

FIGURE 7.5 Computed regression analysis to predict profit from the sales data.

13. Predict the average profit for a $5,000 sale.

14. Note the data point by itself at the extreme right. That is an outlier. To view the rest of the data in greater detail, remove this outlier row of data and create a new chart (Figure 7.6). This spreads the remaining data points, and the shape of the data can be seen more clearly.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.727593
R Square	0.529391
Adjusted R	0.511961
Standard E	333.2941
Observatioı	29

ANOVA

	df	SS	MS	F	ignificance F
Regressior	1	3373922	3373922	30.37245	7.74E-06
Residual	27	2999293	111084.9		
Total	28	6373216			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	Ipper 95.0%
Intercept	-97.2449	100.0923	-0.97155	0.339897	-302.617	108.1275	-302.617	108.1275
Sales ($M)	0.116002	0.021049	5.511121	7.74E-06	0.072814	0.15919	0.072814	0.15919

FIGURE 7.6 Computed regression analysis to predict profit from the sales data with the outlier removed.

Analysis Case 7.2 – Additional Case Using ORDERS

1. Using the Lab Data set and in the *Analysis Case 7* folder, find the file *ORDERS.xlsx*.

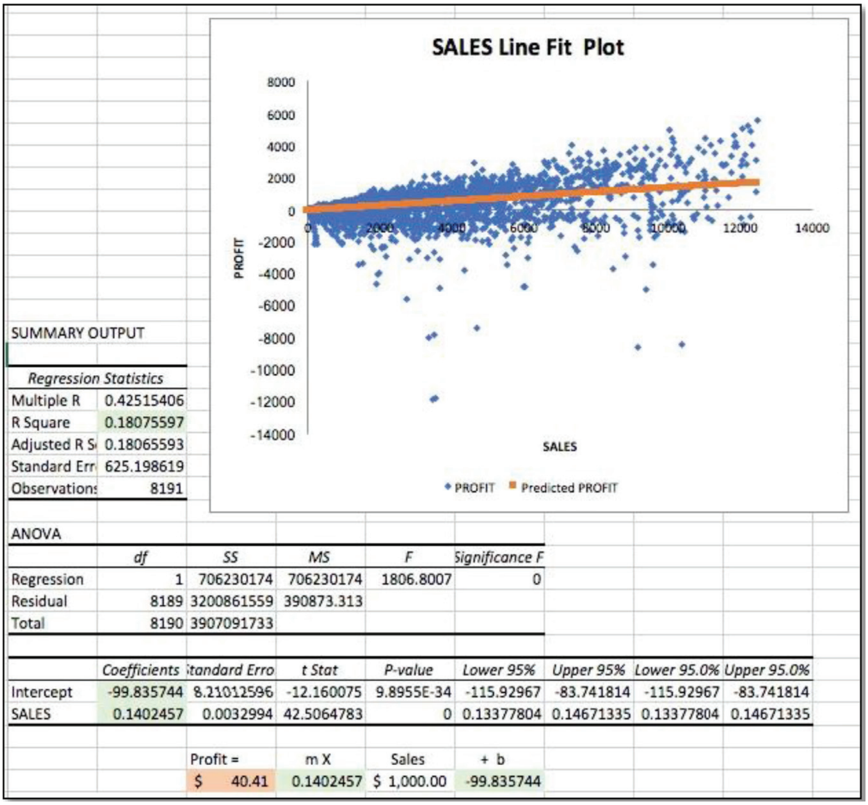
2. Open *ORDERS.xlsx* using Excel.

3. We are going to answer these questions:

   What is a linear regression model of sales to predict profit?

   How much average profit do we predict for a $5,000 sale?

4. Using the Analysis ToolPak compute the linear regression predicting PROFIT from SALES (Figure 7.7). Be sure to remove the outliers.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.42515406
R Square	0.18075597
Adjusted R S	0.18065593
Standard Err	625.198619
Observations	8191

ANOVA

	df	SS	MS	F	Significance F
Regression	1	706230174	706230174	1806.8007	0
Residual	8189	3200861559	390873.313		
Total	8190	3907091733			

	Coefficients	Standard Erro	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-99.835744	8.21012596	-12.160075	9.8955E-34	-115.92967	-83.741814	-115.92967	-83.741814
SALES	0.1402457	0.0032994	42.5064783	0	0.13377804	0.14671335	0.13377804	0.14671335

Profit =	m X	Sales	+ b
$ 40.41	0.1402457	$ 1,000.00	-99.835744

FIGURE 7.7 Computed regression analysis to predict profit from the sales data using the *ORDERS* data.

Analysis Case 7.3 – Additional Case Using SFO Airport Survey Data

1. Use the latest SFO Airport ACQ Survey data downloaded from *https://www.flysfo.com/media/customer-survey-data*.

2. Or using the Lab Data set provided in the *Analysis Case 7* folder, find the file *2016_SFO_Customer_Survey_Data.xls*.

3. Open the data dictionary file and have it available to consult as you work with the data.

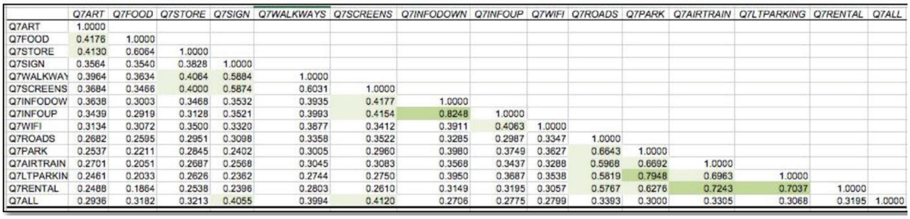
4. Open the data file using Excel.

5. We are going to answer this question:

   Which survey question scores for individual areas of the airport correlate with each other and which correlate with the overall score?

6. Use all the scores in the Q7 columns and compute the correlation matrix using the Analysis ToolPak (Figure 7.8). There is a slight problem with the data in that "0" and "6" responses are not indicative of poor or excellent ratings, but removing these scores does not give us enough working rows, so we must take the correlations with some misgivings. At least it gives us an indicator, as imperfect as it may be.

7. Which sub-questions appear to be correlated to the overall score and to each other?

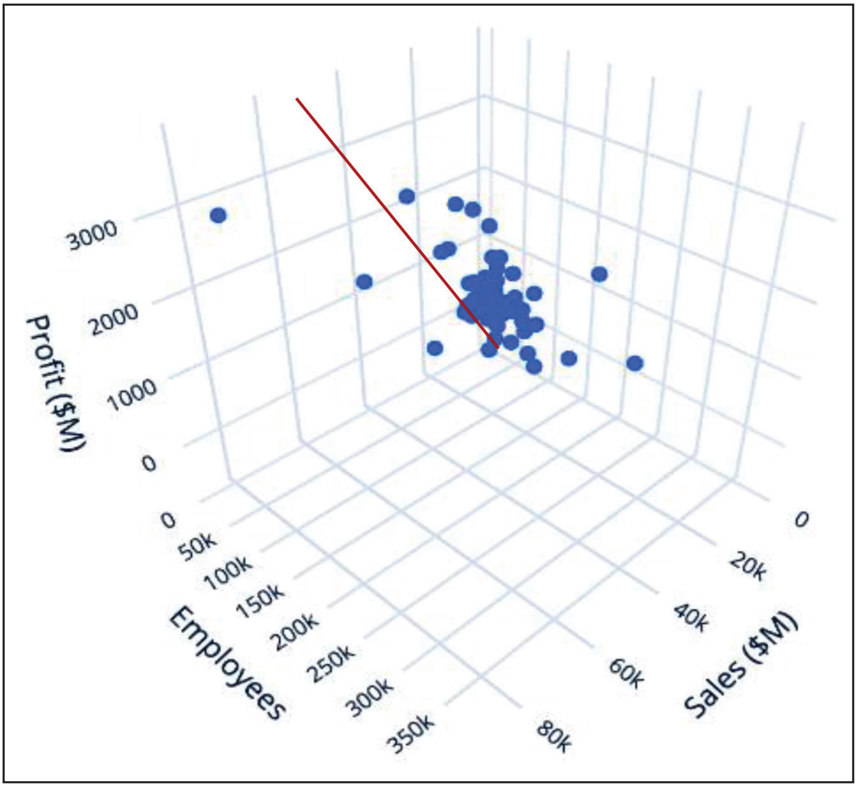


	Q7ART	*Q7FOOD*	*Q7STORE*	*Q7SIGN*	*Q7WALKWAYS*	*Q7SCREENS*	*Q7INFODOWN*	*Q7INFOUP*	*Q7WIFI*	*Q7ROADS*	*Q7PARK*	*Q7AIRTRAIN*	*Q7LTPARKING*	*Q7RENTAL*	*Q7ALL*
Q7ART	1.0000														
Q7FOOD	0.4176	1.0000													
Q7STORE	0.4130	0.6064	1.0000												
Q7SIGN	0.3564	0.3540	0.3828	1.0000											
Q7WALKWAY	0.3964	0.3634	0.4064	0.5884	1.0000										
Q7SCREENS	0.3684	0.3466	0.4000	0.5874	0.6031	1.0000									
Q7INFODOW	0.3638	0.3003	0.3468	0.3532	0.3935	0.4177	1.0000								
Q7INFOUP	0.3439	0.2919	0.3128	0.3521	0.3993	0.4154	0.8248	1.0000							
Q7WIFI	0.3134	0.3072	0.3500	0.3320	0.3877	0.3412	0.3911	0.4063	1.0000						
Q7ROADS	0.2682	0.2595	0.2951	0.3098	0.3358	0.3522	0.3285	0.2987	0.3347	1.0000					
Q7PARK	0.2537	0.2211	0.2845	0.2402	0.3005	0.2960	0.3980	0.3749	0.3627	0.6643	1.0000				
Q7AIRTRAIN	0.2701	0.2051	0.2687	0.2568	0.3045	0.3083	0.3568	0.3437	0.3288	0.5968	0.6692	1.0000			
Q7LTPARKIN	0.2461	0.2033	0.2626	0.2362	0.2744	0.2750	0.3950	0.3687	0.3538	0.5819	0.7948	0.6963	1.0000		
Q7RENTAL	0.2488	0.1864	0.2538	0.2396	0.2803	0.2610	0.3149	0.3195	0.3057	0.5767	0.6276	0.7243	0.7037	1.0000	
Q7ALL	0.2936	0.3182	0.3213	0.4055	0.3994	0.4120	0.2706	0.2775	0.2799	0.3393	0.3000	0.3305	0.3068	0.3195	1.0000

FIGURE 7.8 Correlation matrix of scores for customer satisfaction with individual areas of the airport to each other and to the overall satisfaction score.

CHAPTER 8

MULTIVARIATE REGRESSION



Chapter 8 extends the work of Chapter 7 into the advanced predictive analytic topic of multivariate regression. Consider as before several columns of numerical data from our data table. Now we have not just one predicting (x) independent variable, but several, and still only one predicted (y) dependent or outcome variable. We may ask the question: "How related are all these variables to each other?" To answer that question, we compute the correlation coefficient matrix and see which input variables are most correlated to the intended outcome variable. We will use in our model only input (x) variables, which are most correlated to the outcome variable (y). This action is called variable reduction to the most important ones. Using only those that have a strong correlation to the intended y-variable, then we can further ask: "Can we use all those input x-variables to predict the outcome y-variable?" We will make extensive use of the Analysis ToolPak functions of Correlation and Regression for these exercises.

Consider an example. We know the gross annual sales of all the stores in our franchise, the square footage of each store, the amount of advertising we spend for each store, the inventory carried per store, and the size of the market. We also know the annual profit from each store, as we did in Chapter 7. Are these all related? How strongly are they related? Further, can I somehow use all the input columns for the data table to build a sophisticated linear regression model that will generate the annual predicted profit from a store given levels of inventory, sales, size of the store, advertising budget, market, and competition levels? In the next chapter, we will extend this model to time series, where the x variable is a date, and we will be able to perform trend analysis and build forecasts.

This technique answers the business questions: "How are many input numerical variables related to a numerical outcome variable?" and "Can we use many of the input variables to predict the outcome?"

As in previous chapters, we demonstrate the technique in the first exercise and allow for more challenging work in subsequent exercises.

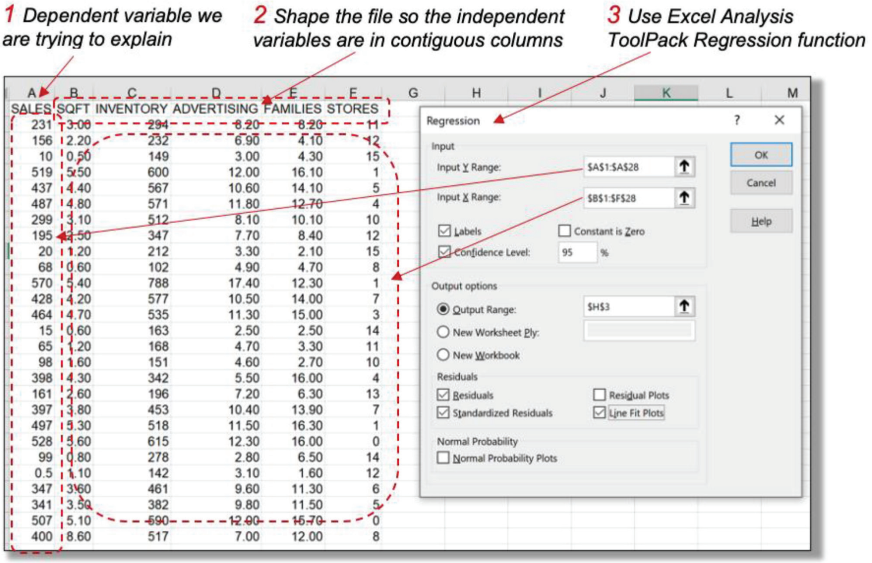
Analysis Case 8.1 – Multivariate Regression

Predictive Modeling

1. Using the Lab Data set provided, open the *Analysis Case 8* folder and find the file *Franchises.xlsx*.
2. Open *Franchises.xlsx* using Excel.
3. We are going to answer these questions:

   > *What factors affect sales in the franchises, and how?*
   >
   > *Can we create a model so we design a store size given certain demographics that allow us to achieve a certain level of sales?*

4. Following our practice of not changing the raw data, select the entire file and copy it as the shaped file in a new spreadsheet, and label the tab *Model*.
5. Make sure the file is properly shaped by having the dependent variable (SALES) to the left of the independent variables (SQFT, INVENTORY, ADVERTISING, FAMILIES, STORES), and make sure that the independent variable columns are contiguous. Excel regression analysis requires this.
6. Invoke the Analysis ToolPak and select the “Regression” function. For the y-variable, use the SALES data (make sure to include the column header and checkmark the “My data has headers” box). For the x-variable, use all the data in the independent variable columns (also include headers). You will want to checkmark the “My data has headers” box.
7. You can put the results on another spreadsheet or right next to our datatable. For now, select a cell next to the table (Figure 8.1).



1 Dependent variable we are trying to explain

2 Shape the file so the independent variables are in contiguous columns

3 Use Excel Analysis ToolPack Regression function

SALES	SQFT	INVENTORY	ADVERTISING	FAMILIES	STORES
231	3.00	294	8.20	8.20	11
156	2.20	232	6.90	4.10	12
10	0.50	149	3.00	4.30	15
519	5.50	600	12.00	16.10	1
437	4.40	567	10.60	14.10	5
487	4.80	571	11.80	12.70	4
299	3.10	512	8.10	10.10	10
195	2.50	347	7.70	8.40	12
20	1.20	212	3.30	2.10	15
68	0.60	102	4.90	4.70	8
570	5.40	788	17.40	12.30	1
428	4.20	577	10.50	14.00	7
464	4.70	535	11.30	15.00	3
15	0.60	163	2.50	2.50	14
65	1.20	168	4.70	3.30	11
98	1.60	151	4.60	2.70	10
398	4.30	342	5.50	16.00	4
161	2.60	196	7.20	6.30	13
397	3.80	453	10.40	13.90	7
497	5.30	518	11.50	16.30	1
528	5.60	615	12.30	16.00	0
99	0.80	278	2.80	6.50	14
0.5	1.10	142	3.10	1.60	12
347	3.60	461	9.60	11.30	6
341	3.50	382	9.80	11.50	5
507	5.10	590	12.00	15.70	0
400	8.60	517	7.00	12.00	8

Regression

Input

Input Y Range: $A$1:$A$28

Input X Range: $B$1:$F$28

☑ Labels

□ Constant is Zero

☑ Confidence Level: 95 %

Output options

● Output Range: $H$3

○ New Worksheet Ply:

○ New Workbook

Residuals

☑ Residuals

□ Residual Plots

☑ Standardized Residuals

☑ Line Fit Plots

Normal Probability

□ Normal Probability Plots

FIGURE 8.1 Steps in a Multivariate Linear Regression analysis to predict SALES from the other franchise store factors.

8. Analyze the results.

9. How much of the variation in the SALES data can be explained by the model? (See *R-squared.*)

10. How confident are you in the validity of this model? (See *Significance F.*)

11. What are the relationships between SALES and the other factors? (See *Coefficients.*)

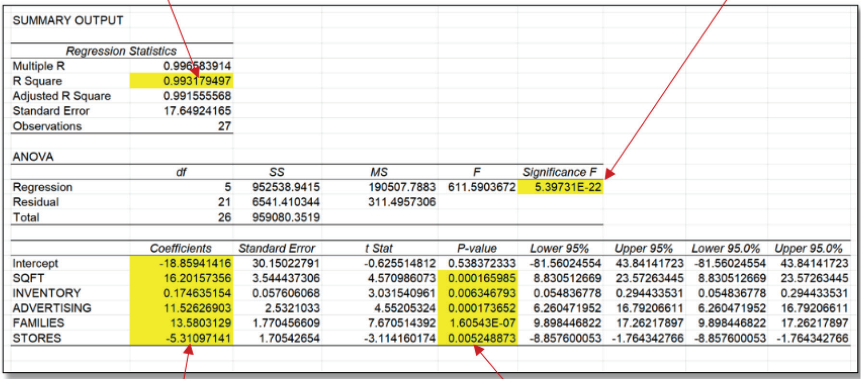
12. Build a linear model using the intercept and the coefficients under the data on the spreadsheet.

13. How confident are we on each coefficient? (See *P-value.*)

14. What does the model predict the sales for the first store will be, and how far off is it compared to actual data (Figures 8.2 and 8.3)?

1 All these x variables combined explain 99% of the variations in SALES

2 Since this is less than .05 we are 95% confident that we have a good model to explain SALES



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.996583914
R Square	0.993179497
Adjusted R Square	0.991555568
Standard Error	17.64924165
Observations	27

ANOVA

	df	SS	MS	F	Significance F
Regression	5	952538.9415	190507.7883	611.5903672	5.39731E-22
Residual	21	6541.410344	311.4957306		
Total	26	959080.3519			

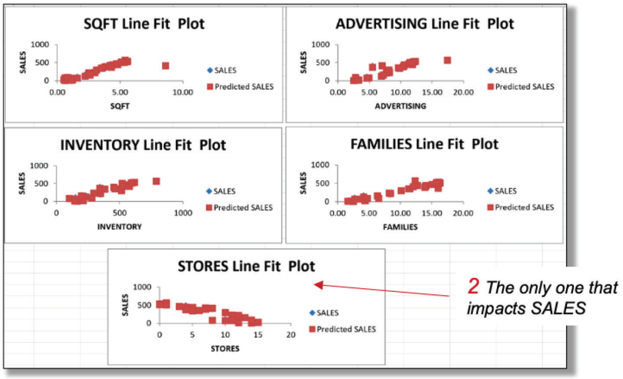
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-18.85941416	30.15022791	-0.625514812	0.538372333	-81.56024554	43.84141723	-81.56024554	43.84141723
SQFT	16.20157356	3.544437306	4.570986073	0.000165985	8.830512669	23.57263445	8.830512669	23.57263445
INVENTORY	0.174635154	0.057606068	3.031540961	0.006346793	0.054836778	0.294433531	0.054836778	0.294433531
ADVERTISING	11.52626903	2.5321033	4.55205324	0.000173652	6.260471952	16.79206611	6.260471952	16.79206611
FAMILIES	13.5803129	1.770456609	7.670514392	1.60543E-07	9.898446822	17.26217897	9.898446822	17.26217897
STORES	-5.31097141	1.70542654	-3.114160174	0.005248873	-8.857600053	-1.764342766	-8.857600053	-1.764342766

4 Use these coefficients to code a predictive equation for SALES

3 Since all of these are less than .05 we are 95% confident that they all make a strong contribution to SALES

SALES = -18.9 + SQFT *16 .2 + INVENTORY *..18 + ADVERTISING *11.5 + FAMILIES *13.6 - STORES *5.3

FIGURE 8.2 Steps in creating a linear regression model to predict SALES.



2 The only one that impacts SALES

1 A quick review of these graphical displays reveals that as you increase all these variables the SALES should increase, except in those regions where there are more competitors (STORES). Which makes sense—note the outlier.

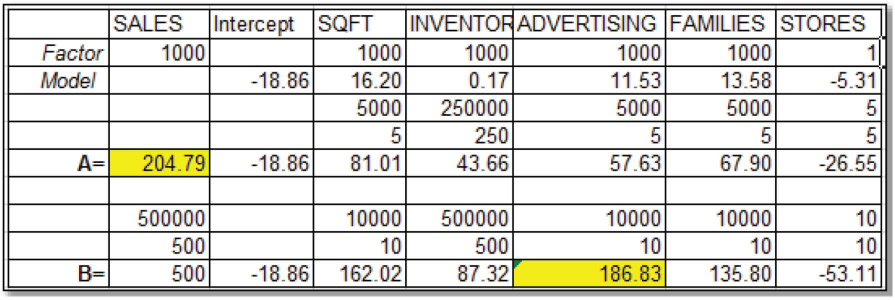
FIGURE 8.3 Line fit plots of SALES against each of the other variables to visualize dependencies.

15. Let's use the model to do some analysis. We are going to answer this question (Figure 8.4, Answer A):

> *We just opened a store in a neighborhood with 5,000 families. The store is 5,000 square feet. We are planning to spend $5,000 a month on advertising. We carry $250,000 in inventory, and there are five competing stores in the neighborhood. What are the projected sales? (Hint: Use the information in the data dictionary to normalize the variables properly and use the model equation.)*

16. Let's do it again in reverse. Build a linear regression model to predict the level of advertising needed when given all the other parameters as input. Then answer the following question (Figure 8.4, Answer B):

> *We want to open a 10,000 square-foot store and realize $500,000 a month in sales in a neighborhood with 10,000 families. We are planning to spend $10,000 a month in advertising. We carry $500,000 in inventory, and there are 10 competing stores in the neighborhood. How much should we spend monthly on advertising to realize our expected sales? (Hint: Use the model equation in reverse.)*



	SALES	Intercept	SQFT	INVENTOR	ADVERTISING	FAMILIES	STORES
Factor	1000		1000	1000	1000	1000	1
Model		-18.86	16.20	0.17	11.53	13.58	-5.31
			5000	250000	5000	5000	5
			5	250	5	5	5
A=	204.79	-18.86	81.01	43.66	57.63	67.90	-26.55
	500000		10000	500000	10000	10000	10
	500		10	500	10	10	10
B=	500	-18.86	162.02	87.32	186.83	135.80	-53.11

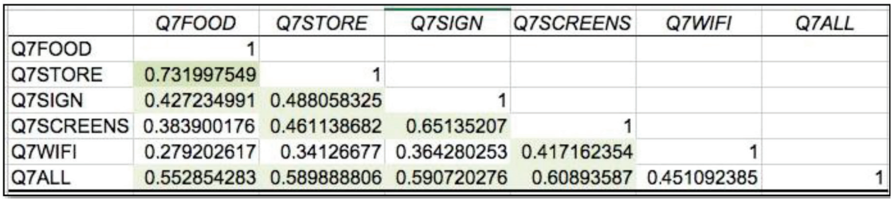
FIGURE 8.4 Results of implementing the prediction models and using them to predict SALES (A) and ADVERTISING budget (B).

Analysis Case 8.2 – Additional Case Using SFO Airport Survey Data

1. Use the latest SFO Airport ACQ Survey data downloaded from *https://www.flysfo.com/media/customer-survey-data*.
2. Or use the Lab Data set provided in the Analysis Case 8 folder and find the file *2016_SFO_Customer_Survey_Data.xls*.
3. Open the data dictionary file and have it available to consult as you work with the data.
4. Open the data file using Excel.
5. We are going to answer this question:

   > *Create a multivariate regression model to predict the overall score Q7ALL someone will give based on the selected sub-question scores (Q7FOOD, Q7STORE, Q7SIGN, Q7SCREENS, and Q7WIFI).*

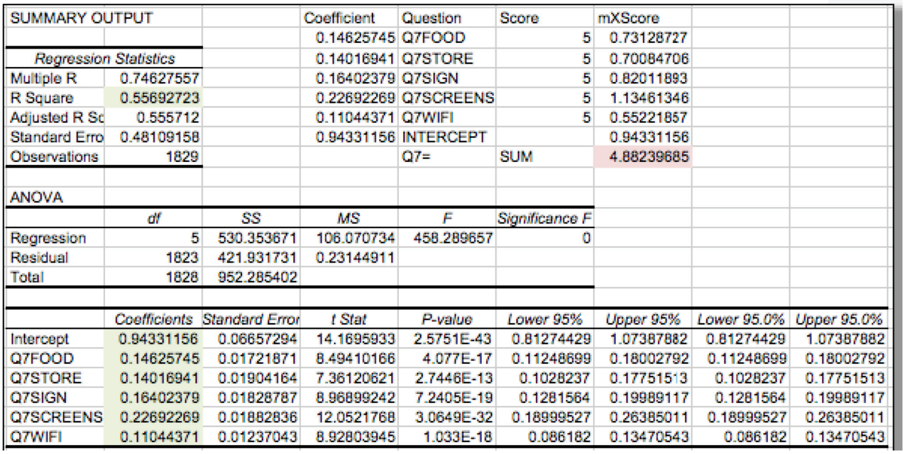
6. To remove the problem of "0" and "6" scores, which should not be entered into a predictive model, delete all rows that have either a "0" or a "6" in any one column. Hopefully, this leaves enough response rows to make a good model. The annotated table should leave behind more than 50% of the original rows.
7. Which sub-questions appear to be correlated to the overall score and to each other?
8. Create a multivariate linear regression model (MLR) to predict the Q7Overall score given the scores in the other sub-questions (Figure 8.5).



	Q7FOOD	Q7STORE	Q7SIGN	Q7SCREENS	Q7WIFI	Q7ALL
Q7FOOD	1					
Q7STORE	0.731997549	1				
Q7SIGN	0.427234991	0.488058325	1			
Q7SCREENS	0.383900176	0.461138682	0.65135207	1		
Q7WIFI	0.279202617	0.34126677	0.364280253	0.417162354	1	
Q7ALL	0.552854283	0.589888806	0.590720276	0.60893587	0.451092385	1

FIGURE 8.5 Correlation matrix of customer satisfaction scores for individual areas of the airport to each other and to the overall satisfaction score.

9. Test your model by predicting the Q7Overall score when someone gives all “5s” and all “1s” in the sub-questions and see how close to 5 and 1 you come in each case (Figure 8.6).



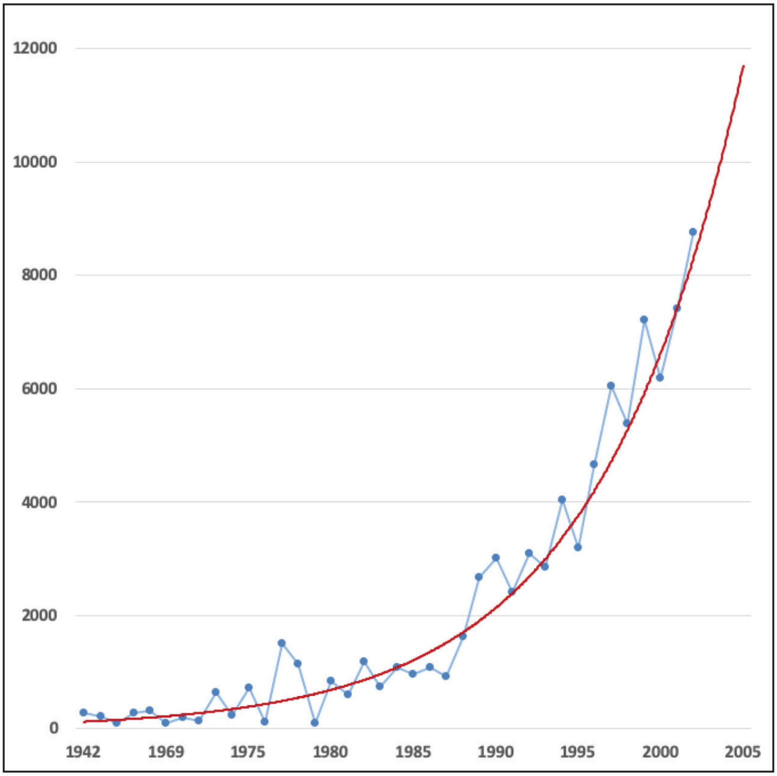
SUMMARY OUTPUT			Coefficient	Question	Score	mXScore		
			0.14625745	Q7FOOD	5	0.73128727		
Regression Statistics			0.14016941	Q7STORE	5	0.70084706		
Multiple R	0.74627557		0.16402379	Q7SIGN	5	0.82011893		
R Square	0.55692723		0.22692269	Q7SCREENS	5	1.13461346		
Adjusted R Sc	0.555712		0.11044371	Q7WIFI	5	0.55221857		
Standard Erro	0.48109158		0.94331156	INTERCEPT		0.94331156		
Observations	1829			Q7=	SUM	4.88239685		
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	5	530.353671	106.070734	458.289657	0			
Residual	1823	421.931731	0.23144911					
Total	1828	952.285402						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.94331156	0.06657294	14.1695933	2.5751E-43	0.81274429	1.07387882	0.81274429	1.07387882
Q7FOOD	0.14625745	0.01721871	8.49410166	4.077E-17	0.11248699	0.18002792	0.11248699	0.18002792
Q7STORE	0.14016941	0.01904164	7.36120621	2.7446E-13	0.1028237	0.17751513	0.1028237	0.17751513
Q7SIGN	0.16402379	0.01828787	8.96899242	7.2405E-19	0.1281564	0.19989117	0.1281564	0.19989117
Q7SCREENS	0.22692269	0.01882836	12.0521768	3.0649E-32	0.18999527	0.26385011	0.18999527	0.26385011
Q7WIFI	0.11044371	0.01237043	8.92803945	1.033E-18	0.086182	0.13470543	0.086182	0.13470543

FIGURE 8.6 Multivariate linear regression model to predict overall customer satisfaction from satisfaction scores in selected areas of the airport operation.

10. Note that the “P-value” of all coefficients is very, very small (less than .05), which means that all sub-question scores are significant and highly related to the overall score.

CHAPTER 9

FORECASTING AND TIME SERIES



Chapter 9 extends the work of Chapter 7 and applies it to time series analysis. Consider the situation where one of the data columns in our table is a date variable. We can use that as the x-axis in a scatter plot with the y-axis being one or more numeric columns of data, and each column will show up as a series. The resulting plot ceases to be a scatter plot, and now we call it a time series. We can then model each time series plot with a regression model. The simplest is a linear model, and it is plotted as a trendline. The Excel plotting function will also give us the equation of the trendline, which we can use to interpolate or extrapolate (also called a forecast). Excel allows us to extend the trendline graphically into future periods, which is very useful. Sometimes when we observe the resulting trendline, we see it does not fit very well. Excel provides other regression models, such as exponential and moving averages, which might give a better fit. We exercise all these options in the exercises in the chapter.

This technique answers the business questions: "What is the trendline for this time series?" and "Can we forecast into future time periods? What is the nature of the trend: linear, exponential, or some other form?"

As in previous chapters, we demonstrate the technique in the first exercise and allow for more challenging work in subsequent exercises.

Analysis Case 9.1 – Forecasting and Time Series

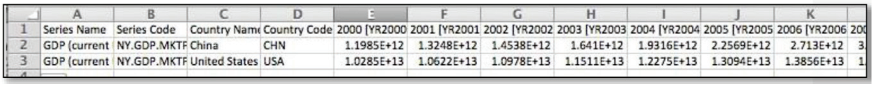
Predicting Trends and Future Values

1. We can use linear regression and other forms of curve fitting to create forecasting models.
2. Using the Lab Data set, and in the *Analysis Case* 9 folder, find the file *WDISelectedData.xlsx*.
3. Open *WDISelectedData.xlsx* using Excel.

4. We are going to answer these questions:

   What is the growth rate of the Chinese GDP and the growth rate of the U.S. GDP? When do you project the Chinese GDP to catch up with the U.S. GDP?

5. Following our practice of not changing the raw data, we are going to create a shaped file for our analysis in a new spreadsheet. Select the title row for the file and enter it into a new spreadsheet. Repeat for the GDP data rows for the U.S. and China. You should have a simple table (Figure 9.1):



	A	B	C	D	E	F	G	H	I	J	K	
1	Series Name	Series Code	Country Nam	Country Code	2000 [YR2000	2001 [YR2001	2002 [YR2002	2003 [YR2003	2004 [YR2004	2005 [YR2005	2006 [YR2006	200
2	GDP (current	NY.GDP.MKTF	China	CHN	1.1985E+12	1.3248E+12	1.4538E+12	1.641E+12	1.9316E+12	2.2569E+12	2.713E+12	3.
3	GDP (current	NY.GDP.MKTF	United States	USA	1.0285E+13	1.0622E+13	1.0978E+13	1.1511E+13	1.2275E+13	1.3094E+13	1.3856E+13	1

FIGURE 9.1 Scraped and cleaned data for the U.S. vs. China GDP analysis.

6. To create a properly shaped file, delete the first two columns and the column with country abbreviations.

7. Replace the dates on the top row with simple dates (2000, 2001, 2002, etc.).

8. Create another two rows with the country data normalized by dividing by 1 trillion (1,000,000,000,000). This gives us numbers we can easily recognize.

9. Swap China and the U.S. so the U.S. is in the first row of data (so the colors of the resulting bars are semantically correct, and the colors associated with each country are what the viewer is expecting).

10. Select the data rows and insert a bar chart into the spreadsheet. Enter the titles of each series and put the dates into the x-axis.

11. Move the legend on the chart to the bottom.

12. Now let's create a predictive model.

13. Right-click on the U.S. data points on the chart (make sure all data points for the U.S. series are selected). Click on "Add a

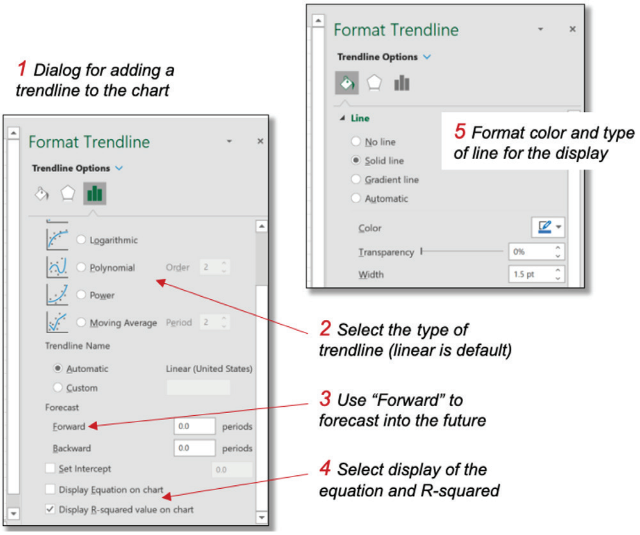
Trendline." Make sure also to click to have the R-squared factor added. You will get a linear model (Figure 9.2).

14. Repeat for China. Make it a linear model as well. What is the R-squared factor for each linear model? Acceptable? Does the linear model look like a good fit for China's GDP trend?

15. Click on the data for China once again and add another trendline. Select an exponential model for this new trendline. Make sure to ask for R-squared, and change the color of this trendline to red. Does it fit the data better?

16. Now let's use these as forecasts. Click on the U.S. linear model and select "Format Trendline." In the middle of the dialog box, you can add a forecast. Note that the data is in years. That indicates that the "periods" in a forecast will be "years." In the trendline dialog box, find the "periods" section and enter five periods (years in this case). Repeat for the Chinese exponential model. Where do the lines cross? What is the answer to our question?

17. Repeat for the Chinese GDP linear model and see what sort of five-year forecast it creates (Figure 9.3).



1 Dialog for adding a trendline to the chart

Format Trendline

Trendline Options

○ Logarithmic
○ Polynomial Order 2
○ Power
○ Moving Average Period 2

Trendline Name
● Automatic Linear (United States)
○ Custom

Forecast
Forward 0.0 periods
Backward 0.0 periods
□ Set Intercept 0.0
□ Display Equation on chart
☑ Display R-squared value on chart

2 Select the type of trendline (linear is default)

3 Use "Forward" to forecast into the future

4 Select display of the equation and R-squared

Format Trendline

Trendline Options

Line
○ No line
● Solid line
○ Gradient line
○ Automatic

Color
Transparency 0%
Width 1.5 pt

5 Format color and type of line for the display

FIGURE 9.2 Using dialogs to add and format the trend line.

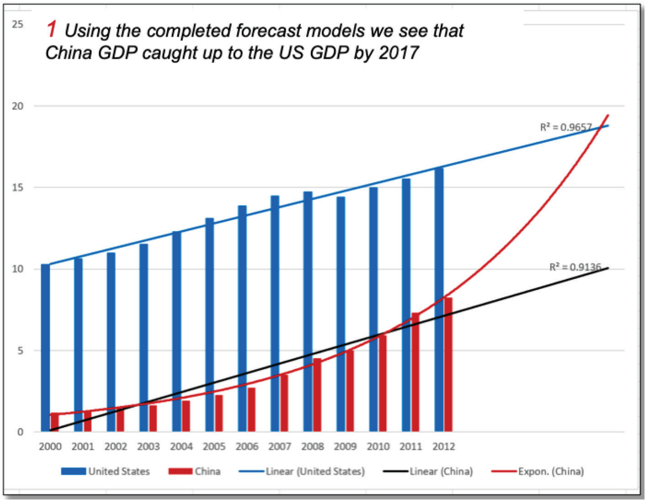


FIGURE 9.3 Steps in creating forecasts to compare U.S. and Chinese GPD and analyze crossover.

Analysis Case 9.2 – Additional Case Using ORDERS

1. Using the Lab Data set and in the *Analysis Case 9* folder, find the file *ORDERS.xlsx*.

2. Open *ORDERS.xlsx* using Excel.

3. We are going to answer these questions:

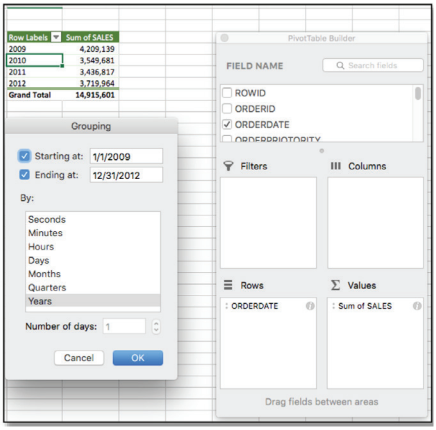
   What is the yearly sales forecast for the past four years?

   What is the overall sales trend?

   Can we identify a province that is contributing the most to that trend?

4. Create a Pivot Table of the entire ORDERS table. Select ORDERDATE as the rows and tabulate the sum of the SALES variable. Now, every order date for each sale is considered to be a category, so the Pivot Table has thousands of rows. We can aggregate them by year (or by month or by month and year) by using a function in Pivot Tables that does that. In the Pivot Table “Format” ribbon, find the “Group by” function. With any one of the dates

selected in the Pivot Table, invoke the function and select "group dates by Year." You should get a short table with four years and the sum of sales for all stores for all years (Figure 9.4).



Row Labels	Sum of SALES
2009	4,209,139
2010	3,549,681
2011	3,436,817
2012	3,719,964
Grand Total	**14,915,601**

Grouping

☑ Starting at: 1/1/2009

☑ Ending at: 12/31/2012

By:

Seconds
Minutes
Hours
Days
Months
Quarters
Years

Number of days: 1

PivotTable Builder

FIELD NAME

☐ ROWID
☐ ORDERID
☑ ORDERDATE
☐ ORDERPRIOTORITY

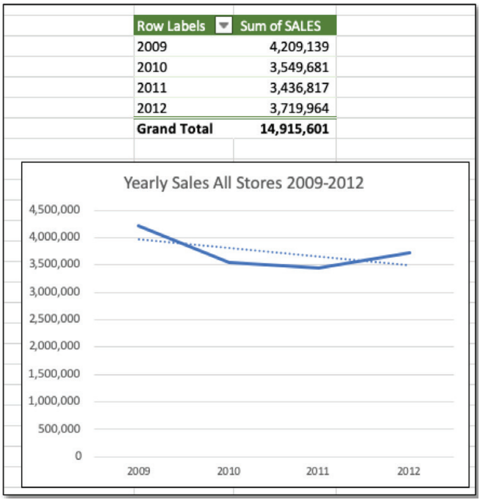
Rows: ORDERDATE

Values: Sum of SALES

FIGURE 9.4 Pivot Table parameters to compute the sales forecast for the last four years of total sales.

5. Create a line chart of the yearly sales totals and add a trendline (Figure 9.5).

   What does the trendline tell you about overall sales?



Row Labels	Sum of SALES
2009	4,209,139
2010	3,549,681
2011	3,436,817
2012	3,719,964
Grand Total	**14,915,601**

FIGURE 9.5 Plot of four years of total sales with added trendline.

6. There is a downward trend in sales. We can do a little digging and discover which province is responsible, or if all provinces are having a yearly loss in sales.

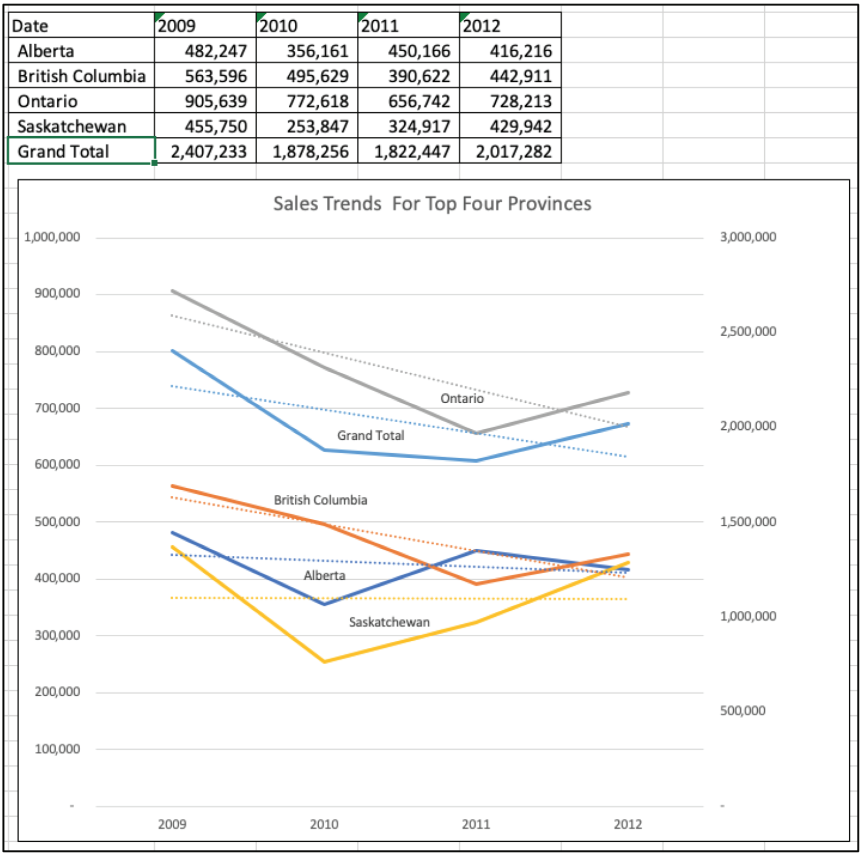
7. Let's create a Pivot Table with yearly sales (rows) for each province (columns). This would be a very busy chart, so we can limit the table to those provinces that have the most sales. Yes, that means performing a Pareto analysis. Fortunately, we did that in Analysis Case 5. Refer to that exercise. There were six out of the 12 provinces that contributed the most to annual sales. Let's restrict it even further by selecting the top four contributors (Alberta, British Columbia, Ontario, and Saskatchewan). We can do this by selecting the filter on columns and then selecting only these four provinces.

8. It's not easy to create a line chart with so many series, as the table is displayed, so we have to copy the table to another spreadsheet. But when you paste it in, use "Paste Special" and checkmark the "Transpose rows and columns" box. This will paste it in a format that will make it easier to create the proper line charts.

9. Create line charts of all sales for the top four provinces from 2009 to 2012. Put total sales on a separate axis so the province trends are easily observed.

10. Add trendlines to all series (Figure 9.6).

    Which province seems to be contributing the most to the total sales decline trend?



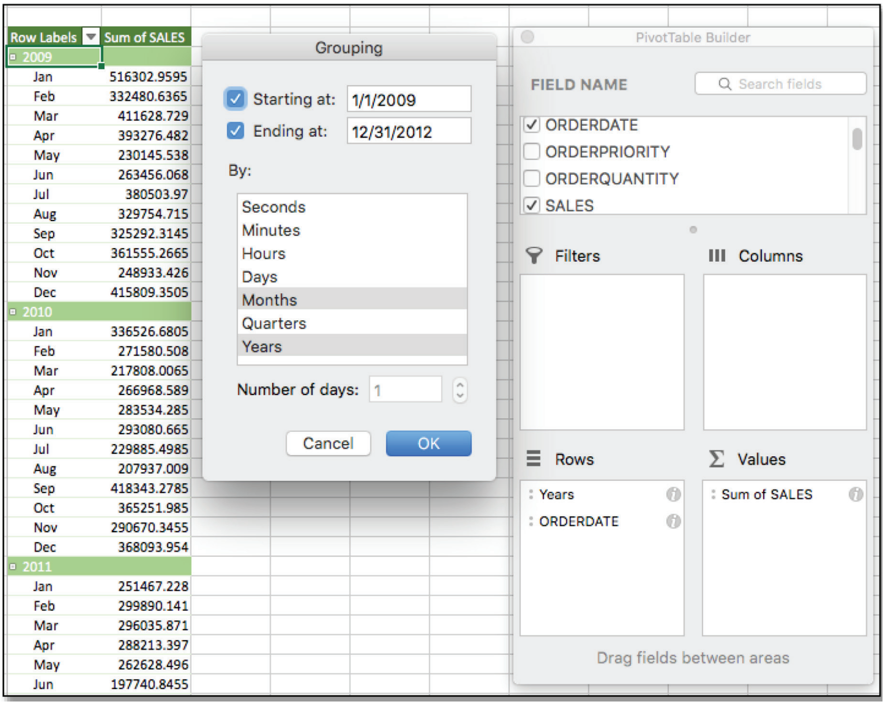
Date	2009	2010	2011	2012
Alberta	482,247	356,161	450,166	416,216
British Columbia	563,596	495,629	390,622	442,911
Ontario	905,639	772,618	656,742	728,213
Saskatchewan	455,750	253,847	324,917	429,942
Grand Total	2,407,233	1,878,256	1,822,447	2,017,282

FIGURE 9.6 Chart showing the sales trends by province.

11. Perhaps displaying the monthly sales would be more useful and give us more insight.

12. Repeat the process but now group the ORDERDATE rows by Month and Year.

13. Note that now the years and months are part of the table (Figure 9.7).



Row Labels	Sum of SALES
2009	
Jan	516302.9595
Feb	332480.6365
Mar	411628.729
Apr	393276.482
May	230145.538
Jun	263456.068
Jul	380503.97
Aug	329754.715
Sep	325292.3145
Oct	361555.2665
Nov	248933.426
Dec	415809.3505
2010	
Jan	336526.6805
Feb	271580.508
Mar	217808.0065
Apr	266968.589
May	283534.285
Jun	293080.665
Jul	229885.4985
Aug	207937.009
Sep	418343.2785
Oct	365251.985
Nov	290670.3455
Dec	368093.954
2011	
Jan	251467.228
Feb	299890.141
Mar	296035.871
Apr	288213.397
May	262628.496
Jun	197740.8455

Grouping

☑ Starting at: 1/1/2009

☑ Ending at: 12/31/2012

By: Months, Years

Number of days: 1

PivotTable Builder

FIELD NAME

☑ ORDERDATE
☐ ORDERPRIORITY
☐ ORDERQUANTITY
☑ SALES

Filters

Columns

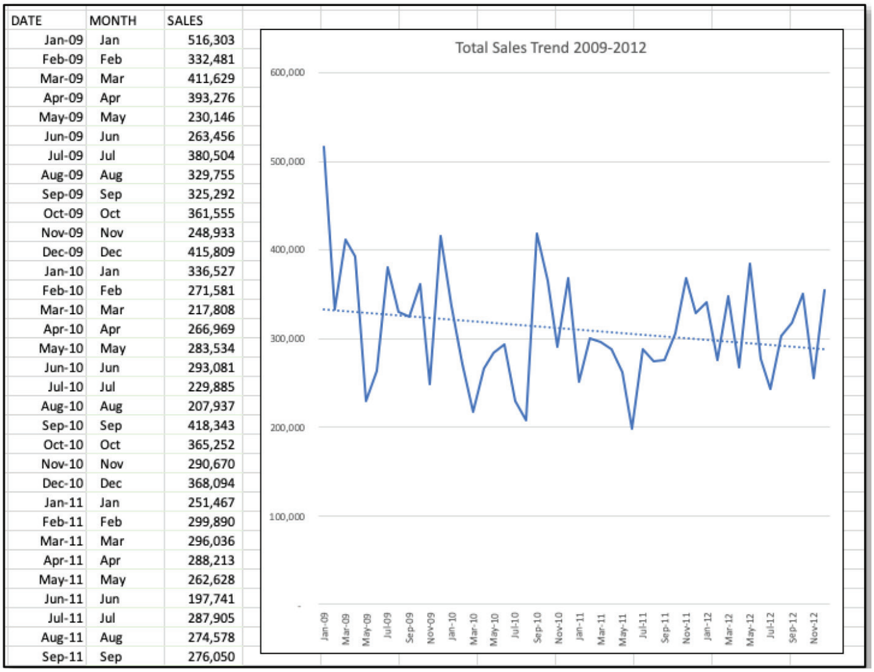
Rows: Years, ORDERDATE

Values: Sum of SALES

FIGURE 9.7 Pivot Table configuration for tabulating SALES by month.

14. Create a table for the line chart. Paste the two columns from the Pivot Table to another sheet, remove the date rows, and add a month-year variable. Then add a trendline (Figure 9.8).

15. We can see the same downward trend here. But what else do you notice?

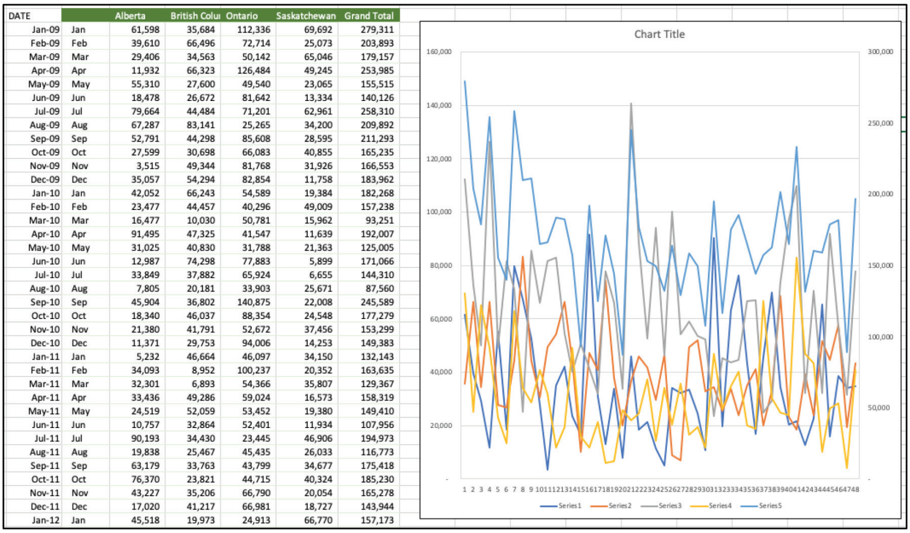


DATE	MONTH	SALES
Jan-09	Jan	516,303
Feb-09	Feb	332,481
Mar-09	Mar	411,629
Apr-09	Apr	393,276
May-09	May	230,146
Jun-09	Jun	263,456
Jul-09	Jul	380,504
Aug-09	Aug	329,755
Sep-09	Sep	325,292
Oct-09	Oct	361,555
Nov-09	Nov	248,933
Dec-09	Dec	415,809
Jan-10	Jan	336,527
Feb-10	Feb	271,581
Mar-10	Mar	217,808
Apr-10	Apr	266,969
May-10	May	283,534
Jun-10	Jun	293,081
Jul-10	Jul	229,885
Aug-10	Aug	207,937
Sep-10	Sep	418,343
Oct-10	Oct	365,252
Nov-10	Nov	290,670
Dec-10	Dec	368,094
Jan-11	Jan	251,467
Feb-11	Feb	299,890
Mar-11	Mar	296,036
Apr-11	Apr	288,213
May-11	May	262,628
Jun-11	Jun	197,741
Jul-11	Jul	287,905
Aug-11	Aug	274,578
Sep-11	Sep	276,050

FIGURE 9.8 Trendline of monthly SALES data.

16. There is a lot of variability to the monthly data. In addition, there is no discernible cyclicality to the data. So, this chart is not as useful as the simpler yearly data in discovering trends.

17. And if we tried the same analysis for the top four provinces compared to total sales but monthly, the chart gets even more confusing (Figure 9.9). One must produce both types of charts to see which one yields greater insight. Given that it was not insightful, this last chart does not need to be further formatted with chart title, series names, axes titles, and so on.



DATE		Alberta	British Colu	Ontario	Saskatchewan	Grand Total
Jan-09	Jan	61,598	35,684	112,336	69,692	279,311
Feb-09	Feb	39,610	66,496	72,714	25,073	203,893
Mar-09	Mar	29,406	34,563	50,142	65,046	179,157
Apr-09	Apr	11,932	66,323	126,484	49,245	253,985
May-09	May	55,310	27,600	49,540	23,065	155,515
Jun-09	Jun	18,478	26,672	81,642	13,334	140,126
Jul-09	Jul	79,664	44,484	71,201	62,961	258,310
Aug-09	Aug	67,287	83,141	25,265	34,200	209,892
Sep-09	Sep	52,791	44,298	85,608	28,595	211,293
Oct-09	Oct	27,599	30,698	66,083	40,855	165,235
Nov-09	Nov	3,515	49,344	81,768	31,926	166,553
Dec-09	Dec	35,057	54,294	82,854	11,758	183,962
Jan-10	Jan	42,052	66,243	54,589	19,384	182,268
Feb-10	Feb	23,477	44,457	40,296	49,009	157,238
Mar-10	Mar	16,477	10,030	50,781	15,962	93,251
Apr-10	Apr	91,495	47,325	41,547	11,639	192,007
May-10	May	31,025	40,830	31,788	21,363	125,005
Jun-10	Jun	12,987	74,298	77,883	5,899	171,066
Jul-10	Jul	33,849	37,882	65,924	6,655	144,310
Aug-10	Aug	7,805	20,181	33,903	25,671	87,560
Sep-10	Sep	45,904	36,802	140,875	22,008	245,589
Oct-10	Oct	18,340	46,037	88,354	24,548	177,279
Nov-10	Nov	21,380	41,791	52,672	37,456	153,299
Dec-10	Dec	11,371	29,753	94,006	14,253	149,383
Jan-11	Jan	5,232	46,664	46,097	34,150	132,143
Feb-11	Feb	34,093	8,952	100,237	20,352	163,635
Mar-11	Mar	32,301	6,893	54,366	35,807	129,367
Apr-11	Apr	33,436	49,286	59,024	16,573	158,319
May-11	May	24,519	52,059	53,452	19,380	149,410
Jun-11	Jun	10,757	32,864	52,401	11,934	107,956
Jul-11	Jul	90,193	34,430	23,445	46,906	194,973
Aug-11	Aug	19,838	25,467	45,435	26,033	116,773
Sep-11	Sep	63,179	33,763	43,799	34,677	175,418
Oct-11	Oct	76,370	23,821	44,715	40,324	185,230
Nov-11	Nov	43,227	35,206	66,790	20,054	165,278
Dec-11	Dec	17,020	41,217	66,981	18,727	143,944
Jan-12	Jan	45,518	19,973	24,913	66,770	157,173

FIGURE 9.9 Trend analysis of monthly SALES data for the top four provinces.

Analysis Case 9.3 – Additional Case Using MOVIES

1. Let's do an additional exercise to see how to gain insight from trendlines.

   What are the trends in revenue in the movie industry from 1937 to 2012?

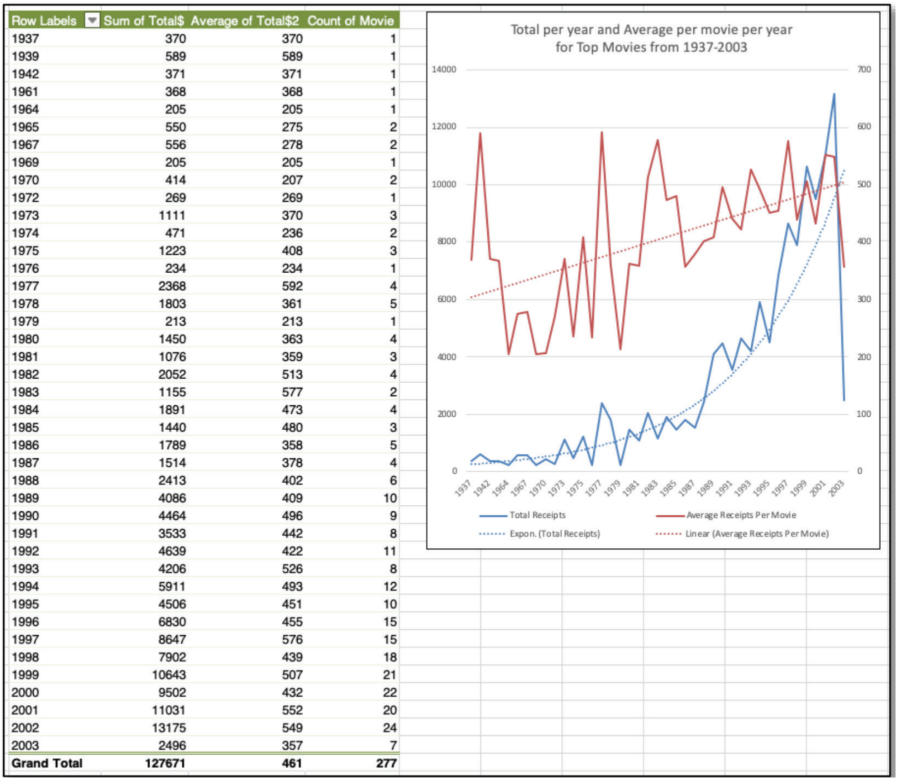
2. Using the Lab Data set and in the *Analysis Case* 9 folder, find the file *Movies.xlsx*.

3. Open *Movies.xlsx* using Excel.

4. Create a Pivot Table of the sum of total receipts, average receipts per movie, and count of movies per year (Figure 9.10).

5. Notice that total receipts seem to be increasing exponentially, but that this may be because the number of movies used to compute receipts drastically changes over the years. The average revenue per movie also seems to have increased, but at a more

linear rate. This may be due to inflation. (Are the receipts over the years normalized to inflation?) All of these things must be considered.



Row Labels	Sum of Total$	Average of Total$2	Count of Movie
1937	370	370	1
1939	589	589	1
1942	371	371	1
1961	368	368	1
1964	205	205	1
1965	550	275	2
1967	556	278	2
1969	205	205	1
1970	414	207	2
1972	269	269	1
1973	1111	370	3
1974	471	236	2
1975	1223	408	3
1976	234	234	1
1977	2368	592	4
1978	1803	361	5
1979	213	213	1
1980	1450	363	4
1981	1076	359	3
1982	2052	513	4
1983	1155	577	2
1984	1891	473	4
1985	1440	480	3
1986	1789	358	5
1987	1514	378	4
1988	2413	402	6
1989	4086	409	10
1990	4464	496	9
1991	3533	442	8
1992	4639	422	11
1993	4206	526	8
1994	5911	493	12
1995	4506	451	10
1996	6830	455	15
1997	8647	576	15
1998	7902	439	18
1999	10643	507	21
2000	9502	432	22
2001	11031	552	20
2002	13175	549	24
2003	2496	357	7
Grand Total	**127671**	**461**	**277**

FIGURE 9.10 Trend analysis for total yearly sales and for yearly average.

Analysis Case 9.4 – Additional Forecasting Case Study

1. Let's use our knowledge of forecasting and predictive models to analyze a real business situation.

2. Case study:

   You own a business in New York State. You have a 34,000-square-foot building, which you heat with propane. Last year, you used

22,000 gallons of propane. Rather than pay the spot price for propane each time the company fills up your tank, the gas company proposes a future-buy contract where you agree to a set price per gallon for a portion of your expected use next heating season. If the price goes up, you are protected by the set price. The downside is, if the price goes down, you will lose on the possible savings. But you have been burned before by large price fluctuations, and you are seriously considering this offer. The gas company offers to sell you 18,000 gallons under four possible contract options:

A. *Pay for all 18,000 gallons up front at the rate of $2.10 per gallon.*

B. *Pay a deposit of 20 percent up front and the rest of the 18,000 gallons on ten months of equal payments for $2.29 per gallon.*

C. *Pay a $50 fee up front and, for 18,000 gallons, pay ten equal payments over ten months for $2.49 per gallon.*

D. *Avoid taking out a contract at all and pay monthly at the spot price, which fluctuates every day.*

3. You may use the actual weekly cost of propane for the last two years compiled by the NYSRDA government agency of the state of New York. The data file is in the *Analysis Cases* folder under the *Analysis Case* 9 folder and labeled *PropanePrices.xlsx*.

4. The business question:

> *Which of the four contract options is most economically advantageous?*

3. You may want to use forecasting to predict the cost of propane six months in the future (the middle of winter) to be able to project what the savings might be (Figure 9.11).

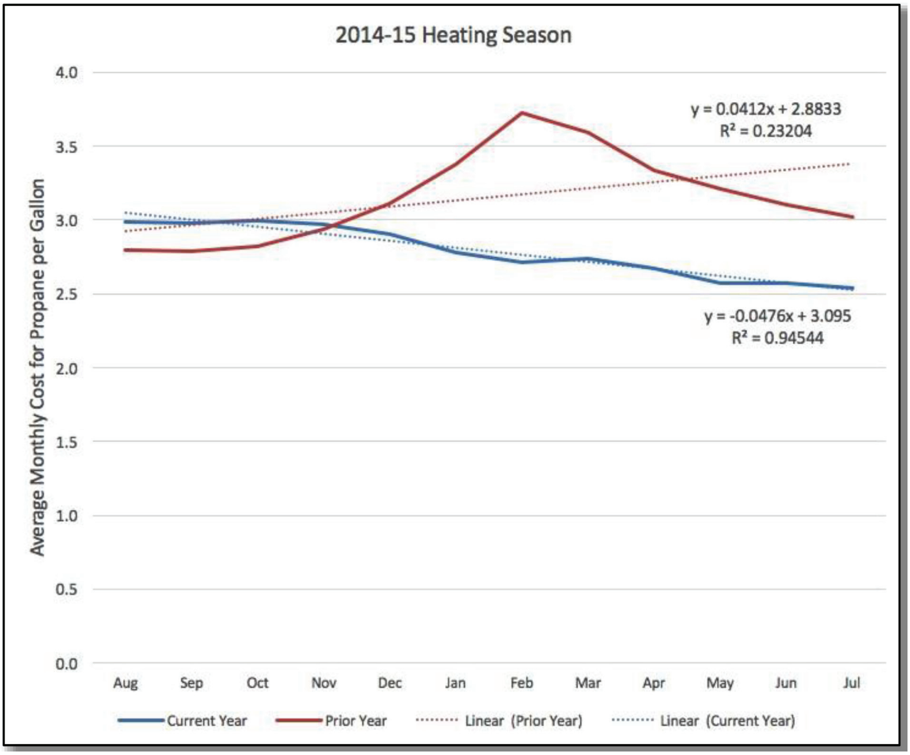
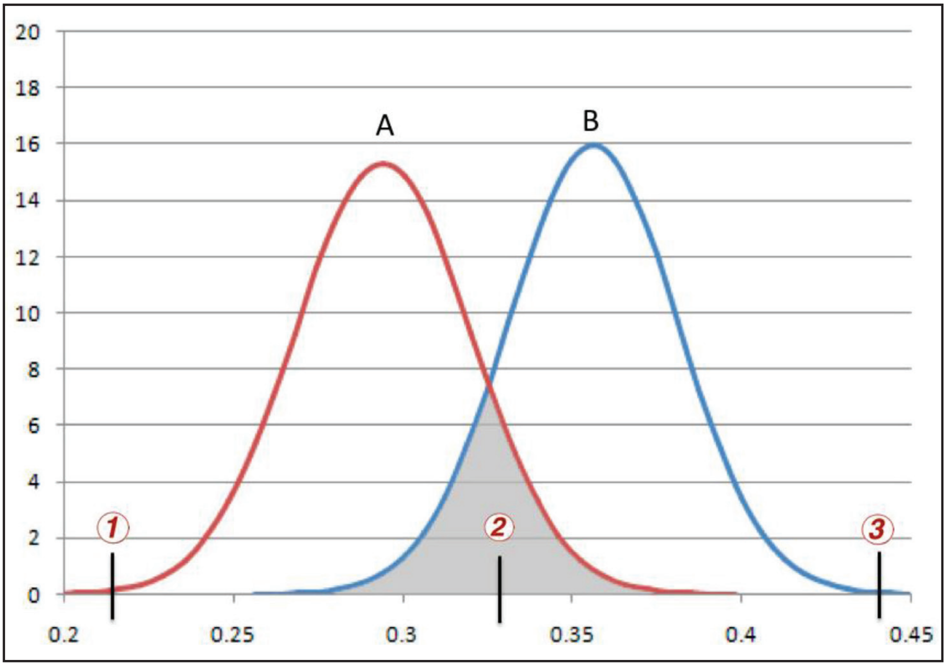


FIGURE 9.11 Trend analysis to compare two heating seasons.

CHAPTER 10

INFERENTIAL STATISTICS



In Chapter 10, we do something entirely different. Consider the situation where one of two columns of data is the status of a person before and after watching a movie. The one column of data is the rating they give the film before they see it (from what they heard of the movie), and the other rating is their opinion after they watched the movie. We want to compare the ratings from before and after; we want to say something about how similar or dissimilar the average ratings are before and after watching the movie. We will use a t-Test to say something about the difference between the means of the two sets of ratings. The t-Test tells us how likely it would be that we would make an error if we knew a score and ascribed it as belonging to one population when it really belonged to the other. The degree of certainty will depend on the overlap on the distributions of the two sets of scores. If we want to be 95% confident that the means are statistically significantly different, we want that overlap (*p-value*) of the two distributions to be less than 5% (for business purposes, we consider that alpha level sufficient). We can't be totally sure, but we infer a certain level of certainty from the analysis of the data using this t-Test.

This technique is instrumental in testing hypothetical outcomes, and it is used in research based on the scientific method—running experiments. In business, we use it to assure ourselves that averages are statistically significantly different or not, as the case may be. This technique answers the business question: "Are the means of two related distributions statistically significantly different or not?"

As in previous chapters, we demonstrate the technique in the first exercise and allow for more challenging work in subsequent exercises. We also demonstrate and allow practice on the ANOVA technique that checks for differences in means in more than paired sets.

Analysis Case 10.1 – Assurance of Results

Inferential Statistics

1. Using the *Lab Data* set provided, open the *Analysis Case 10* folder in it, and find the file *StartupCosts.xls*.

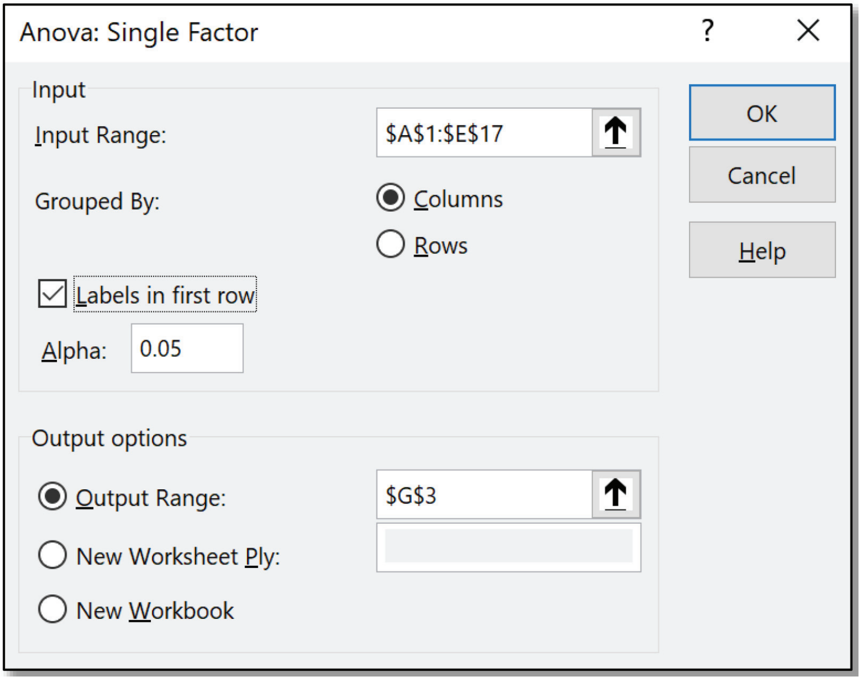
2. Open *StartupCosts.xls* using Excel.

3. We are going to answer this question:

> *Each group's startup costs vary widely, and the average startup costs vary from one type of store to another. Are the differences in the average real, or are they different purely by chance?*

4. Following our practice of not changing the raw data, we are going to create a shaped file for our analysis in a new spreadsheet. Select the data and enter it into a new spreadsheet in the workbook under a tab called *ANOVA*.

5. Use the Analysis ToolPak to compute the "ANOVA Single Factor" test. Make sure to select all the data, including the column headers (Figure 10.1).



Anova: Single Factor

Input

Input Range: $A$1:$E$17

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Alpha: 0.05

Output options

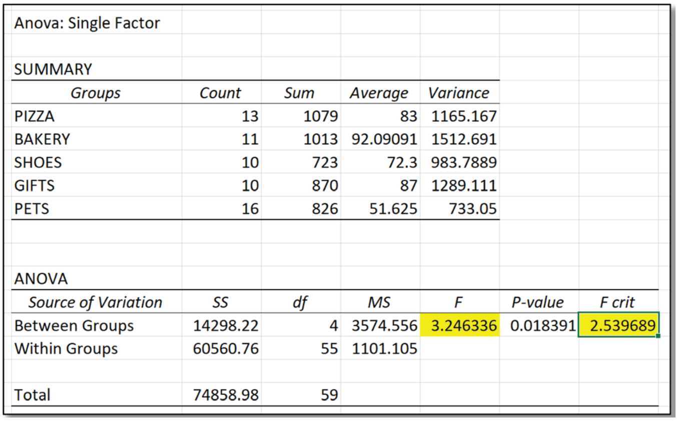
● Output Range: $G$3

○ New Worksheet Ply:

○ New Workbook

FIGURE 10.1 Analysis ToolPak configuration for an ANOVA inferential analysis of startup cost data.

6. Compare the *F value* to the *F critical value* (Figure 10.2) and determine if the means are the same or not. (In this case, there is at least one pair of means that is different.)



Anova: Single Factor

SUMMARY

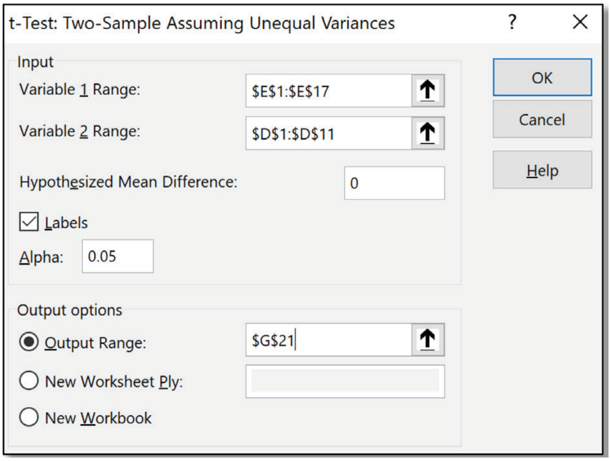
Groups	Count	Sum	Average	Variance
PIZZA	13	1079	83	1165.167
BAKERY	11	1013	92.09091	1512.691
SHOES	10	723	72.3	983.7889
GIFTS	10	870	87	1289.111
PETS	16	826	51.625	733.05

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	14298.22	4	3574.556	3.246336	0.018391	2.539689
Within Groups	60560.76	55	1101.105			
Total	74858.98	59				

FIGURE 10.2 Results of the ANOVA analysis of *StartupCosts* data.

7. Use the Analysis ToolPak to compute the t-Test comparison of pairs of means (Figure 10.3). Use the "Two-Sample Assuming Unequal Variances" function. Make sure to select two columns of data, including the column headers. Compare *GIFTS* and *PETS* startups.



t-Test: Two-Sample Assuming Unequal Variances

Input

Variable 1 Range: $E$1:$E$17

Variable 2 Range: $D$1:$D$11

Hypothesized Mean Difference: 0

☑ Labels

Alpha: 0.05

Output options

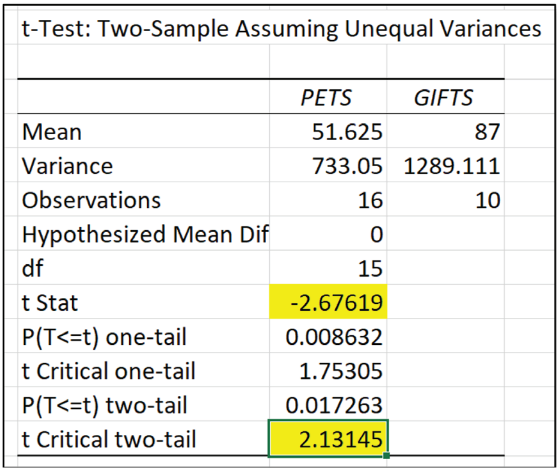
● Output Range: $G$21

○ New Worksheet Ply:

○ New Workbook

FIGURE 10.3 Analysis ToolPak configuration for t-Test inferential analysis to compare the means for startup costs for two different types of stores.

8. Compare the *t-stat* value to the *t critical* value for two tails (Figure 10.4) and determine if the means are the same or not. (In this case, *t-stat* > *t critical,* so the means are significantly different.)



t-Test: Two-Sample Assuming Unequal Variances

	PETS	*GIFTS*
Mean	51.625	87
Variance	733.05	1289.111
Observations	16	10
Hypothesized Mean Dif	0	
df	15	
t Stat	-2.67619	
P(T<=t) one-tail	0.008632	
t Critical one-tail	1.75305	
P(T<=t) two-tail	0.017263	
t Critical two-tail	2.13145	

FIGURE 10.4 The results of the t-Test analysis.

Analysis Case 10.2 – Additional Case Using FOOTBALLTRIALS

1. Using the Lab Data set and in the *Analysis Case 10* folder, find the file *FootballTrials.xls*.

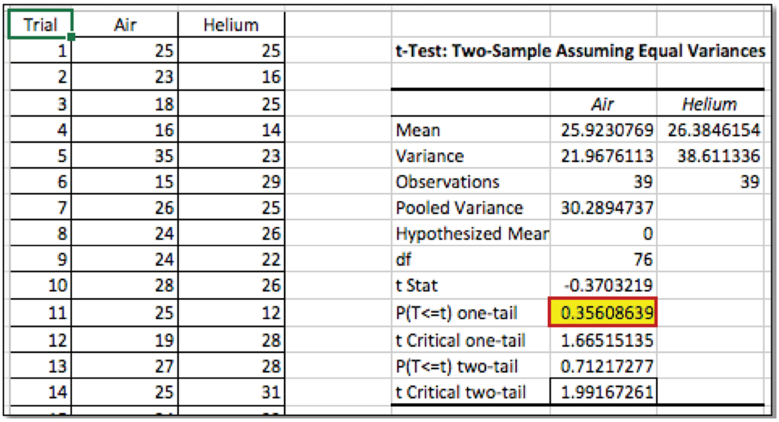
2. Open *FootballTrials.xls* using Excel.

3. Be sure to consult the data dictionary and the description of the case study.

4. We are going to answer this question:

   Does it make a difference in the distance a kicked football travels if it is filled with air or filled with helium?

5. Use the Analysis ToolPak and compute the t-Test to compare the means of the air-filled football trials to the helium-filled football trials (Figure 10.5). (Is the p-value above or below the .05 criterion for a 95% confidence that the means are different?)



Trial	Air	Helium
1	25	25
2	23	16
3	18	25
4	16	14
5	35	23
6	15	29
7	26	25
8	24	26
9	24	22
10	28	26
11	25	12
12	19	28
13	27	28
14	25	31

t-Test: Two-Sample Assuming Equal Variances

	Air	*Helium*
Mean	25.9230769	26.3846154
Variance	21.9676113	38.611336
Observations	39	39
Pooled Variance	30.2894737	
Hypothesized Mear	0	
df	76	
t Stat	-0.3703219	
P(T<=t) one-tail	0.35608639	
t Critical one-tail	1.66515135	
P(T<=t) two-tail	0.71217277	
t Critical two-tail	1.99167261	

FIGURE 10.5 Results of the t-Test analysis of air-filled and helium-filled football sets of kicks.

6. What is the null hypothesis in this case? Did we prove the null hypothesis? Is this good in this case?

7. How would you answer that question?

Analysis Case 10.3 – Additional Exercise Using SFO Airport Survey Data

1. Use the latest SFO Airport ACQ Survey data downloaded from *https://www.flysfo.com/media/customer-survey-data*.

2. Or using the Lab Data set provided in the *Analysis Case 10* folder, find the file *2016_SFO_Customer_Survey_Data.xls.*

3. Open the data dictionary and have it available to consult as you work with the data.

4. We are going to answer several questions using contingency analysis:

> *Do frequent flyer customers (fly > 100,000 miles/year) and regular flyers feel differently about airport cleanliness, safety, and overall rating?*

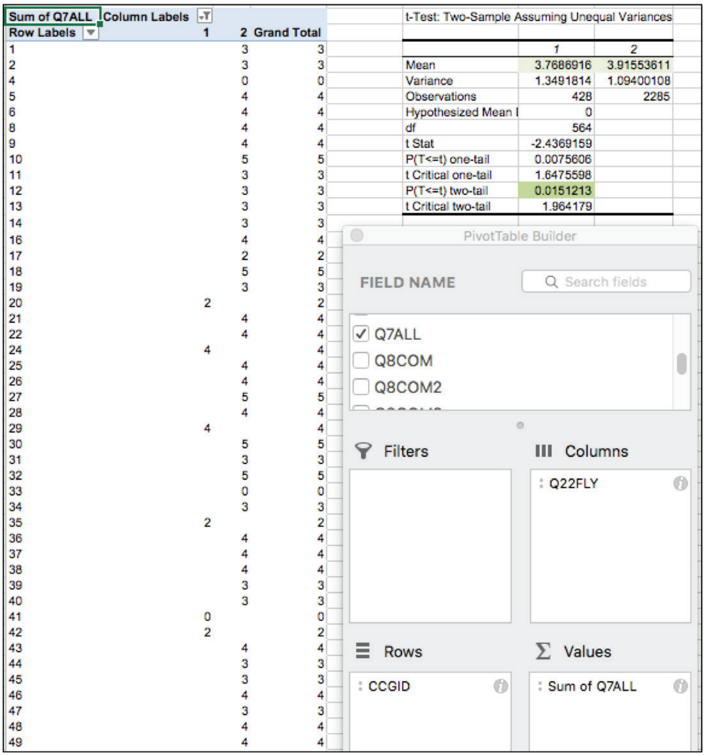
5. Open the data file in Excel.

6. Create a Pivot Table of all IDs as rows by Q21FLY as columns. Enter the ratings for each row in the data set by putting the sum of Q7ALL in the Results box. This yields two columns for the t-Test comparison (Figure 10.6).

7. Make sure to remove all but category 1 (frequent fliers) and category 2 (regular flyers) from the columns using the column filter. Perform a t-Test analysis with unequal variances.

8. This should give us an answer to the difference between how frequent flyers feel compared to regular customers. Do the two groups feel differently? Are the means significantly different?

9. Since before performing the t-Test, we did not know which average of the two would be higher, we should accept the two-tailed result.



Sum of Q7ALL	Column Labels		
Row Labels	1	2	Grand Total
1		3	3
2		3	3
4		0	0
5		4	4
6		4	4
8		4	4
9		4	4
10		5	5
11		3	3
12		3	3
13		3	3
14		3	3
16		4	4
17		2	2
18		5	5
19		3	3
20	2		2
21		4	4
22		4	4
24	4		4
25		4	4
26		4	4
27		5	5
28		4	4
29	4		4
30		5	5
31		3	3
32		5	5
33		0	0
34		3	3
35	2		2
36		4	4
37		4	4
38		4	4
39		3	3
40		3	3
41	0		0
42	2		2
43		4	4
44		3	3
45		3	3
46		4	4
47		3	3
48		4	4
49		4	4

t-Test: Two-Sample Assuming Unequal Variances

	1	2
Mean	3.7686916	3.91553611
Variance	1.3491814	1.09400108
Observations	428	2285
Hypothesized Mean I	0	
df	564	
t Stat	-2.4369159	
P(T<=t) one-tail	0.0075606	
t Critical one-tail	1.6475598	
P(T<=t) two-tail	0.0151213	
t Critical two-tail	1.964179	

PivotTable Builder

FIELD NAME

Search fields

☑ Q7ALL
☐ Q8COM
☐ Q8COM2

Filters

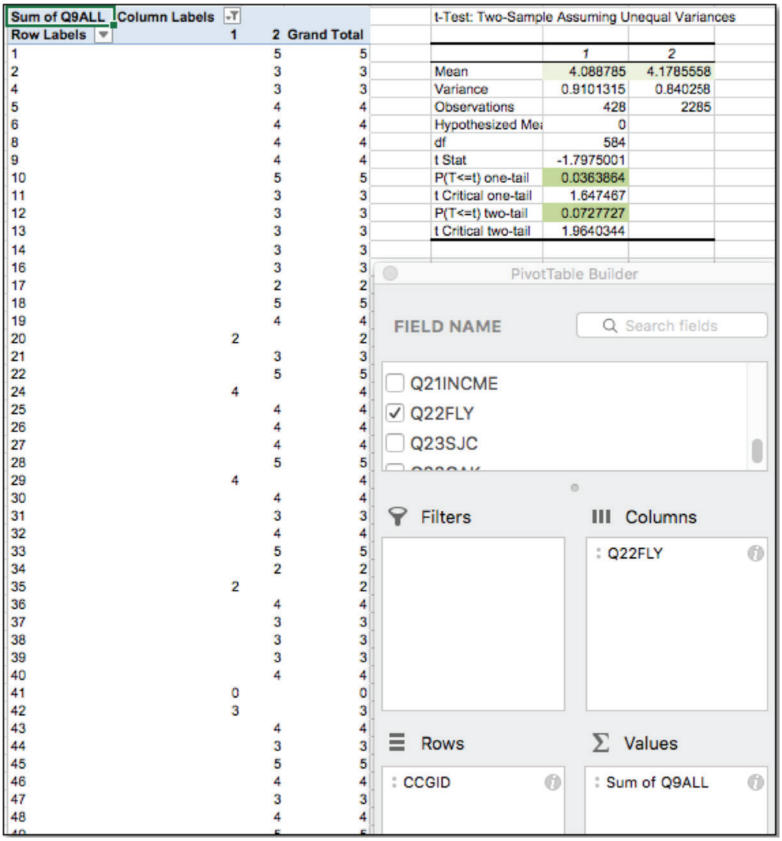
Columns: Q22FLY

Rows: CCGID

Values: Sum of Q7ALL

FIGURE 10.6 Pivot Table parameters and results of the t-Test analysis of overall customer satisfaction comparing regular and frequent flyer customers.

10. Create a Pivot Table of all IDs as rows by Q21FLY as columns (Figure 10.7). Enter the ratings for each row in the data set by putting the sum of Q9ALL (cleanliness) in the Results box. This yields two columns for the t-Test comparison. Make sure to remove all but category 1 (frequent fliers) and category 2 (regular flyers) from the columns using the column filter. Perform a t-Test analysis with unequal variances. This yields an answer to the difference between how frequent flyers feel about cleanliness compared to regular customers. Since before performing the test, we did not know which mean of the two would be higher, we should accept the two-tailed result.



Sum of Q9ALL	Column Labels		
Row Labels	1	2	Grand Total
1		5	5
2		3	3
4		3	3
5		4	4
6		4	4
8		4	4
9		4	4
10		5	5
11		3	3
12		3	3
13		3	3
14		3	3
16		3	3
17		2	2
18		5	5
19		4	4
20	2		2
21		3	3
22		5	5
24	4		4
25		4	4
26		4	4
27		4	4
28		5	5
29	4		4
30		4	4
31		3	3
32		4	4
33		5	5
34		2	2
35	2		2
36		4	4
37		3	3
38		3	3
39		3	3
40		4	4
41	0		0
42	3		3
43		4	4
44		3	3
45		5	5
46		4	4
47		3	3
48		4	4

t-Test: Two-Sample Assuming Unequal Variances

	1	2
Mean	4.088785	4.1785558
Variance	0.9101315	0.840258
Observations	428	2285
Hypothesized Mea	0	
df	584	
t Stat	-1.7975001	
P(T<=t) one-tail	0.0363864	
t Critical one-tail	1.647467	
P(T<=t) two-tail	0.0727727	
t Critical two-tail	1.9640344	

PivotTable Builder

FIELD NAME

□ Q21INCME
☑ Q22FLY
□ Q23SJC

Filters

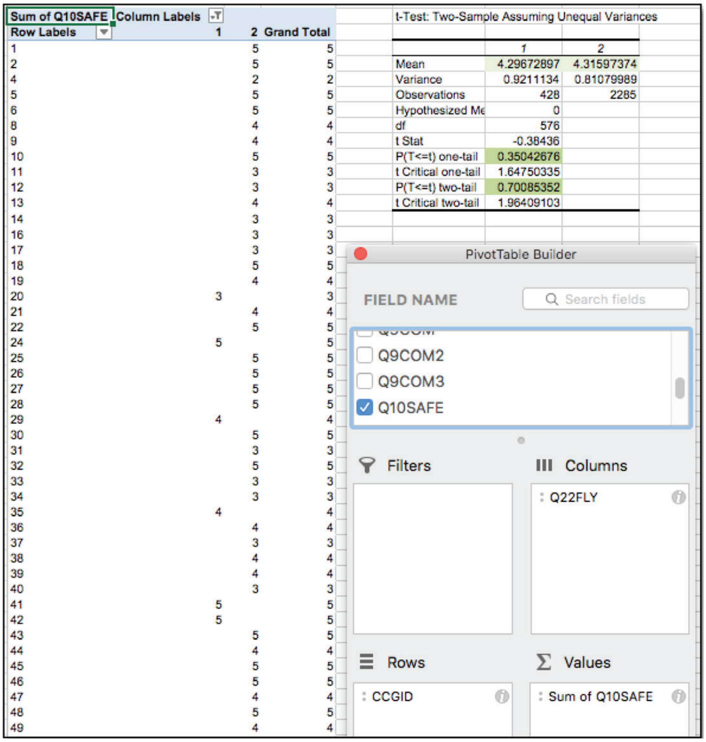
Columns: Q22FLY

Rows: CCGID

Values: Sum of Q9ALL

FIGURE 10.7 Pivot Table parameters and results of the t-Test analysis of customer satisfaction with airport cleanliness comparing regular and frequent flyer customers.

11. Create a Pivot Table of all IDs as rows by Q21FLY as columns (Figure 10.8). Enter the ratings for each row in the data set by putting the sum of Q10ALL in the Results box. This yields two columns for the t-Test comparison. Make sure to remove all but category 1 (frequent fliers) and category 2 (regular flyers) from the columns using the column filter. Perform a t-Test analysis with unequal variances. This should give us an answer to the difference between how frequent flyers feel about safety compared to regular customers. Since before performing the test, we did not know which mean of the two would be higher, we should accept the two-tailed result. Notice that in this case, the t-Test fails the two-tailed result but passes the one-tailed result, forcing us to accept the null hypothesis and conclude that the means are not significantly different.



Sum of Q10SAFE	Column Labels		
Row Labels	1	2	Grand Total
1		5	5
2		5	5
4		2	2
5		5	5
6		5	5
8		4	4
9		4	4
10		5	5
11		3	3
12		3	3
13		4	4
14		3	3
16		3	3
17		3	3
18		5	5
19		4	4
20	3		3
21		4	4
22		5	5
24	5		5
25		5	5
26		5	5
27		5	5
28		5	5
29	4		4
30		5	5
31		3	3
32		5	5
33		3	3
34		3	3
35	4		4
36		4	4
37		3	3
38		4	4
39		4	4
40		3	3
41	5		5
42	5		5
43		5	5
44		4	4
45		5	5
46		5	5
47		4	4
48		5	5
49		4	4

t-Test: Two-Sample Assuming Unequal Variances

	1	2
Mean	4.29672897	4.31597374
Variance	0.9211134	0.81079989
Observations	428	2285
Hypothesized Me	0	
df	576	
t Stat	-0.38436	
P(T<=t) one-tail	0.35042676	
t Critical one-tail	1.64750335	
P(T<=t) two-tail	0.70085352	
t Critical two-tail	1.96409103	

PivotTable Builder

FIELD NAME

Search fields

☐ Q9COM2

☐ Q9COM3

☑ Q10SAFE

Filters

Columns: Q22FLY

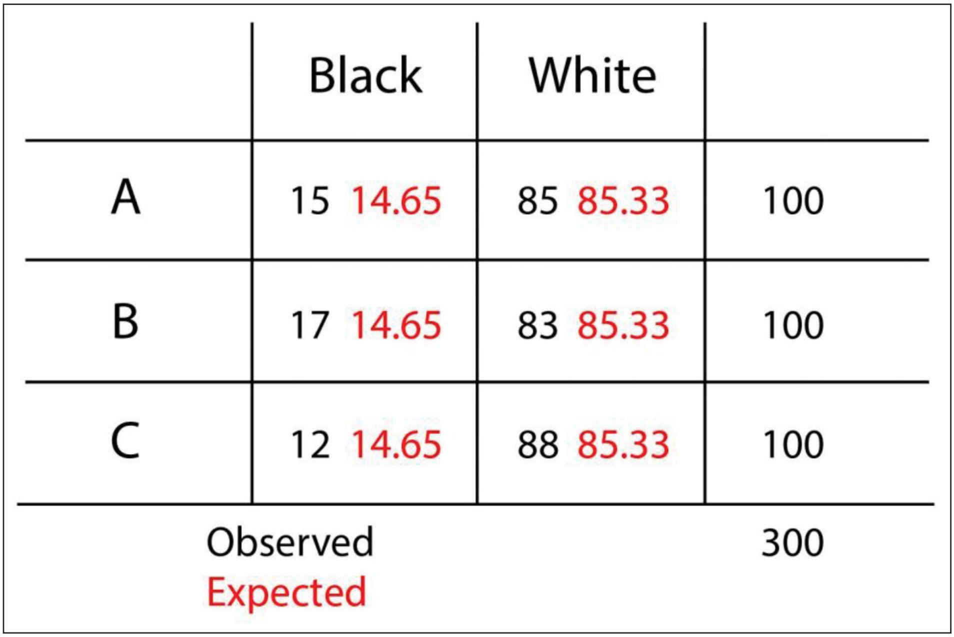
Rows: CCGID

Values: Sum of Q10SAFE

FIGURE 10.8 Pivot Table parameters and results of the t-Test analysis of customer satisfaction with airport safety comparing regular and frequent flyer customers.

CHAPTER 11

CONTINGENCY ANALYSIS



	Black	White	
A	15 14.65	85 85.33	100
B	17 14.65	83 85.33	100
C	12 14.65	88 85.33	100
Observed Expected			300

In Chapter 11, we continue our work of inferential statistical analysis. Now we work with categorical variables. Consider the situation where we have two categorical tables, say of the passengers on the Titanic. One of two columns of categorical data is the survival status of a passenger, the other their gender. First, we practice cross-tabulating two variables to see if gender made a difference in who survived. That's a contingency table. In this chapter, we go further. We have been told that in this disaster, the crew followed the Law of The Sea and that there was a concerted effort, a bias, to place women on lifeboats ahead of men. We notice that the ratios indeed show a higher survival rate for women. But then we ask: was this by chance or can we show that there was an underlying bias to put women on the boats? For that, we need to employ the inferential analysis technique of computing chi-squared for this contingency table. If the *p-value* resulting from the analysis is less than .05 (5%), we infer that the outcome variable *survived* is not independent of the input variable *gender* and that there was a bias.

It is not easy to set up a chi-squared analysis for a crosstab (contingency table) in Excel. So we practice using chi-squared calculations freely available on the Internet. We present many situations in the exercises where we wish to test for independence.

As in Chapter 10, this technique is instrumental in testing hypothetical outcomes, and it is used in research based on the scientific method—running experiments, but for categorical outcomes. In business, we use it to assure ourselves that averages are statistically significantly different or not, as the case may be. If we want to be 95% confident that the variables are dependent, we are looking for the resulting *p-value* from the chi-squared analysis to be less than 5% (for business purposes, we consider that alpha level sufficient). We can't be 100% sure there was a bias, but we infer a certain level of certainty from the analysis of the data using the alpha = .05 test level.

This technique answers the business question: "Are two categorical variables independent, or is there an underlying bias or relationship between them?"

As in previous chapters, we demonstrate the technique in the first exercise and allow for more challenging work in subsequent exercises.

Analysis Case 11.1 – Contingency Analysis and Chi-Squared

Test of Variable Independence

1. In this case study, we will use contingency tables and the test of independence of those variables to investigate the Titanic disaster more closely. Suppose you are an investigative journalist writing a story on the Titanic and hear from a few people that the crew of the Titanic followed the "law of the sea" in saving passenger lives. What is the law of the sea, and how would you show that the crew heroically followed it?

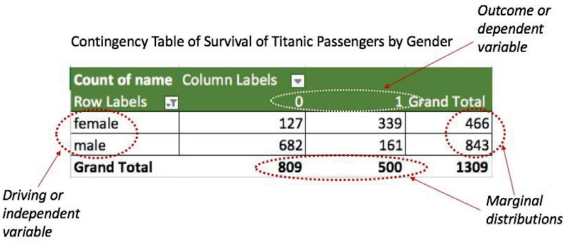
2. The Law of the Sea states that "women and children should be saved first in a disaster." (See *https://en.wikipedia.org/wiki/Women_and_children_first.*) Did this happen on the Titanic? We are going to investigate it by answering the following questions about the Titanic disaster:

   Did women survive at a higher rate than men?

   Did children survive at a higher rate than adults?

   Did the passenger class of a passenger make a difference in his/her survival rate?

3. We are going to use contingency analysis, which in Excel is basically performed with Pivot Tables. This is what it looks like (Figure 11.1):



Contingency Table of Survival of Titanic Passengers by Gender

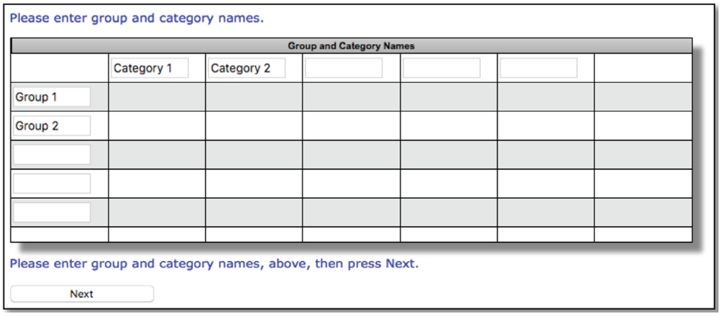
Count of name	Column Labels		
Row Labels	0	1	Grand Total
female	127	339	466
male	682	161	843
Grand Total	809	500	1309

Outcome or dependent variable
Driving or independent variable
Marginal distributions

FIGURE 11.1 Elements of a contingency table.

4. Excel does provide a formula to compute chi-squared and the p-value derived from it, but it requires the creation of the expected matrix from the contingency table, which is complicated and tedious.

5. We are going to use a free chi-squared computation calculator found on the Internet, which makes it much easier. We have a calculator that does the analysis for up to a 5x5 table: *www.socscistatistics.com/tests/chisquare2/Default2.aspx* (Figure 11.2). For up to a 10x10 table, you can use *www.quantpsy.org/chisq/chisq.htm*.



Please enter group and category names.

Group and Category Names

	Category 1	Category 2				
Group 1						
Group 2						

Please enter group and category names, above, then press Next.

FIGURE 11.2 A Web-based 5X5 chi-square calculator.

6. The process is first to use a Pivot Table to create the contingency table of the two categorical variables in question. Then we compute the ratios we are interested in. Next, we ensure

that the differences in the ratios we see are real and not due to chance by performing the chi-squared test. The variables are "dependent" if the p-value is less than .05, which means that we are at least 95% confident that the differences are real.

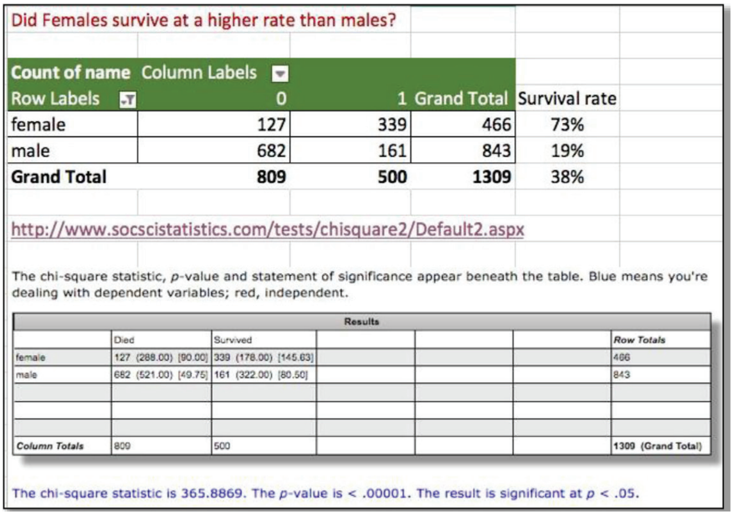
7. Using the Lab Data set provided, open the *Analysis Case 11* folder in it, and find the file *Titanic.xlsx*.

8. Open *Titanic.xlsx* with Excel.

9. To answer the first question, create a Pivot Table of gender (variable sex in the table) versus survival. Compute the ratios to see if women survived at a higher rate than men.

10. Enter the elements of the Pivot Table in the 5X5 online calculator to ensure that the difference in survival rates was not due to chance. What do you see?

11. The chi-squared statistic can best be described as the standardized deviation of observed data from expected data. Our concern is not with the actual number but with the probability that the results we observe are due to chance, given by the p-value (Figure 11.3).



Did Females survive at a higher rate than males?

Count of name	Column Labels			
Row Labels	0	1	Grand Total	Survival rate
female	127	339	466	73%
male	682	161	843	19%
Grand Total	**809**	**500**	**1309**	38%

http://www.socscistatistics.com/tests/chisquare2/Default2.aspx

The chi-square statistic, *p*-value and statement of significance appear beneath the table. Blue means you're dealing with dependent variables; red, independent.

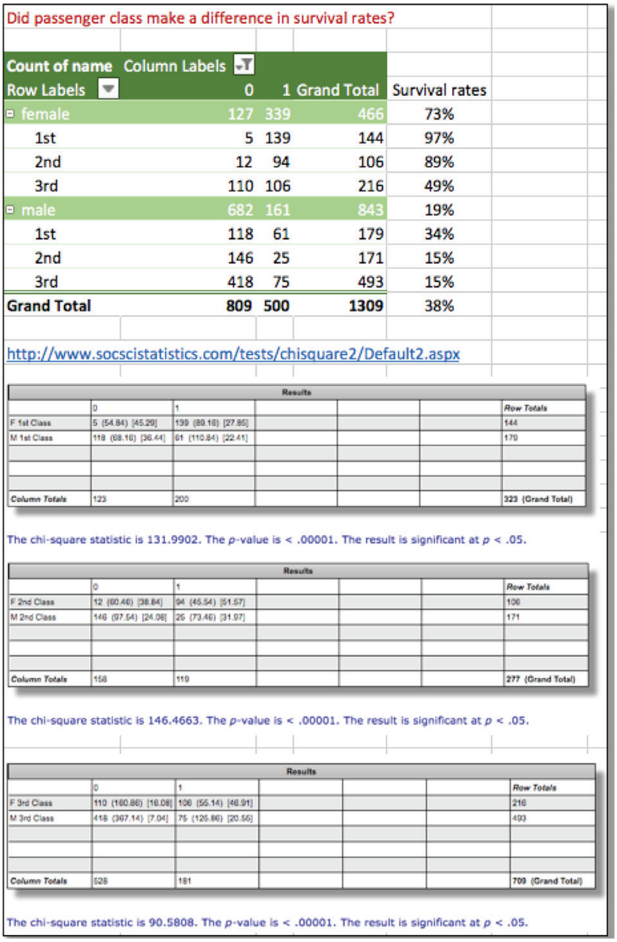
Results						
	Died	Survived				*Row Totals*
female	127 (288.00) [90.00]	339 (178.00) [145.63]				466
male	682 (521.00) [49.75]	161 (322.00) [80.50]				843
Column Totals	809	500				1309 (Grand Total)

The chi-square statistic is 365.8869. The *p*-value is < .00001. The result is significant at *p* < .05.

FIGURE 11.3 Computing a contingency table using a Pivot Table and performing a chi-square analysis for independence using a Web-based calculator.

12. To answer the next question, create a Pivot Table of gender (variable sex in the table) versus survival (Figure 11.4). Compute the ratios to see if women survived at a higher rate than men and add a subcategory of the *pclass* variable (passenger class).

13. Again, use the online calculator inputting the elements of the Pivot Table to ensure that the differences in survival rates by passenger class were not due to chance. What do you see?



Did passenger class make a difference in survival rates?

Count of name	Column Labels			
Row Labels	0	1	Grand Total	Survival rates
⊟ female	127	339	466	73%
1st	5	139	144	97%
2nd	12	94	106	89%
3rd	110	106	216	49%
⊟ male	682	161	843	19%
1st	118	61	179	34%
2nd	146	25	171	15%
3rd	418	75	493	15%
Grand Total	**809**	**500**	**1309**	38%

http://www.socscistatistics.com/tests/chisquare2/Default2.aspx

Results

	0	1				*Row Totals*
F 1st Class	5 (54.84) [45.29]	139 (89.16) [27.85]				144
M 1st Class	118 (68.16) [36.44]	61 (110.84) [22.41]				179
Column Totals	123	200				323 (Grand Total)

The chi-square statistic is 131.9902. The *p*-value is < .00001. The result is significant at *p* < .05.

Results

	0	1				*Row Totals*
F 2nd Class	12 (60.46) [38.84]	94 (45.54) [51.57]				106
M 2nd Class	146 (97.54) [24.08]	25 (73.46) [31.97]				171
Column Totals	158	119				277 (Grand Total)

The chi-square statistic is 146.4663. The *p*-value is < .00001. The result is significant at *p* < .05.

Results

	0	1				*Row Totals*
F 3rd Class	110 (160.86) [16.08]	106 (55.14) [46.91]				216
M 3rd Class	418 (367.14) [7.04]	75 (125.86) [20.56]				493
Column Totals	528	181				709 (Grand Total)

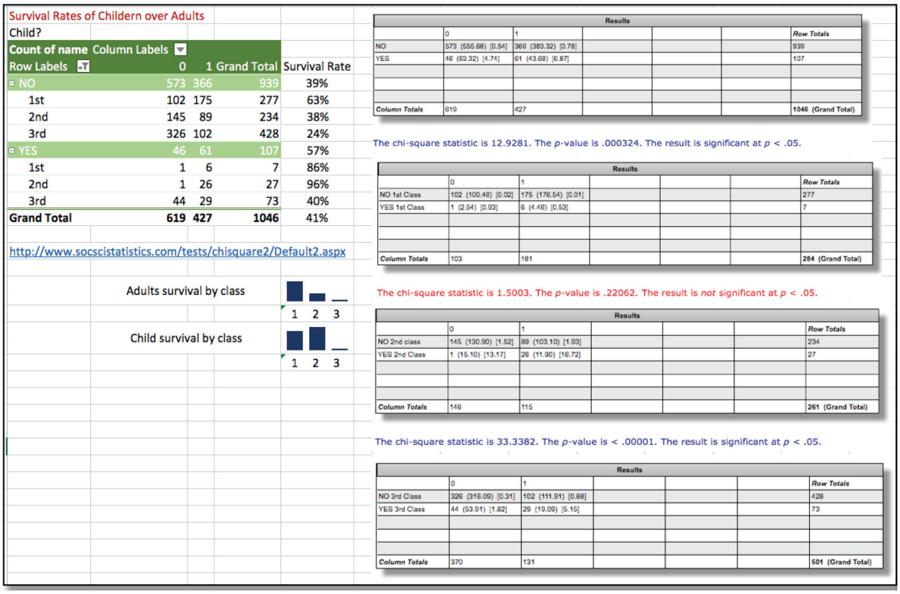
The chi-square statistic is 90.5808. The *p*-value is < .00001. The result is significant at *p* < .05.

FIGURE 11.4 Computing a contingency table using a Pivot Table and performing a chi-square analysis using a Web-based calculator for a Titanic passenger survival analysis.

14. To answer the last question, we need to add a new variable based on age, which we can call CHILD (YES or NO) using an IF function (=IF(E2>14, IF(E2="NA"," ","NO"),"YES")). We will use 14 years of age as a cutoff for children, as that was the acceptable age of childhood 100 years ago. Note that we want to have a blank for any person whose age is not given in the table (N/A). That way we can filter blanks out when we create the contingency table.

15. Create a Pivot Table of child versus survive (Figure 11.5). Compute the ratios to see if children survived at a higher rate than adults. Also, add the subcategory *pclass* to see if there was a survival bias based on the passenger class.

16. Use the online calculator to ensure that the differences in survival rates were not due to chance. Note the addition of sparklines to visualize the rates of survival by class. What do you see?



Survival Rates of Childern over Adults

Child?

Count of name	Column Labels			
Row Labels	0	1	Grand Total	Survival Rate
⊟ NO	573	366	939	39%
1st	102	175	277	63%
2nd	145	89	234	38%
3rd	326	102	428	24%
⊟ YES	46	61	107	57%
1st	1	6	7	86%
2nd	1	26	27	96%
3rd	44	29	73	40%
Grand Total	619	427	1046	41%

http://www.socscistatistics.com/tests/chisquare2/Default2.aspx

Results						
	0	1				*Row Totals*
NO	573 (555.68) [0.54]	366 (383.32) [0.78]				939
YES	46 (63.32) [4.74]	61 (43.68) [6.87]				107
Column Totals	619	427				1046 (Grand Total)

The chi-square statistic is 12.9281. The *p*-value is .000324. The result is significant at *p* < .05.

Results						
	0	1				*Row Totals*
NO 1st Class	102 (100.46) [0.02]	175 (176.54) [0.01]				277
YES 1st Class	1 (2.54) [0.93]	6 (4.46) [0.53]				7
Column Totals	103	181				284 (Grand Total)

The chi-square statistic is 1.5003. The *p*-value is .22062. The result is *not* significant at *p* < .05.

Results						
	0	1				*Row Totals*
NO 2nd class	145 (130.90) [1.52]	89 (103.10) [1.93]				234
YES 2nd Class	1 (15.10) [13.17]	26 (11.90) [16.72]				27
Column Totals	146	115				261 (Grand Total)

The chi-square statistic is 33.3382. The *p*-value is < .00001. The result is significant at *p* < .05.

Results						
	0	1				*Row Totals*
NO 3rd Class	326 (316.09) [0.31]	102 (111.91) [0.88]				428
YES 3rd Class	44 (53.91) [1.82]	29 (19.09) [5.15]				73
Column Totals	370	131				501 (Grand Total)

FIGURE 11.5 Computing a contingency table using a Pivot Table and performing a chi-square analysis using a Web-based calculator for a Titanic passenger survival analysis based on passenger class.

Analysis Case 11.2 – Additional Case Using SFO Airport Survey Data

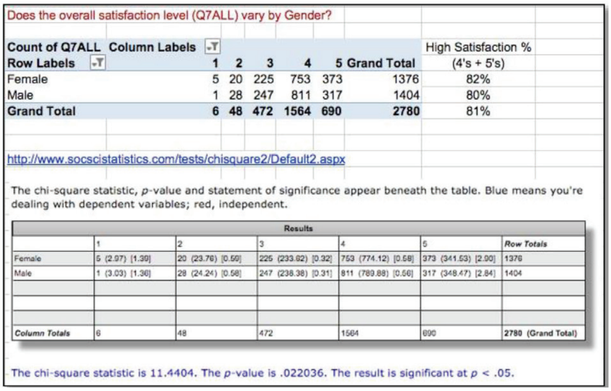
1. Use the latest SFO Airport ACQ Survey data downloaded from *www.flysfo.com/media/customer-survey-data*.
2. Or using the Lab Data set provided in the *Analysis Case 11* folder, find the file *2016_SFO_Customer_Survey_Data.xls*.
3. Open the data dictionary and have it available to consult as you work with the data.
4. We are going to answer several questions using contingency analysis:

    How does the overall satisfaction level with the SFO Airport vary by gender?

    How does overall satisfaction vary by income level?

    How does overall satisfaction vary by age?

5. Open the data file using Excel.
6. Create a Pivot Table of Q7ALL by Q22GENDER (Figure 11.6). Make sure to remove the "0" and "6" response rows for Q7 and the blank and "3" responses for Q22. Perform a chi-squared test to ensure that the variables are dependent. Check the percentage approval using the sum of the numbers of the "4" and "5" ratings.



Does the overall satisfaction level (Q7ALL) vary by Gender?

Count of Q7ALL	Column Labels						High Satisfaction %
Row Labels	1	2	3	4	5	Grand Total	(4's + 5's)
Female	5	20	225	753	373	1376	82%
Male	1	28	247	811	317	1404	80%
Grand Total	6	48	472	1564	690	2780	81%

http://www.socscistatistics.com/tests/chisquare2/Default2.aspx

The chi-square statistic, *p*-value and statement of significance appear beneath the table. Blue means you're dealing with dependent variables; red, independent.

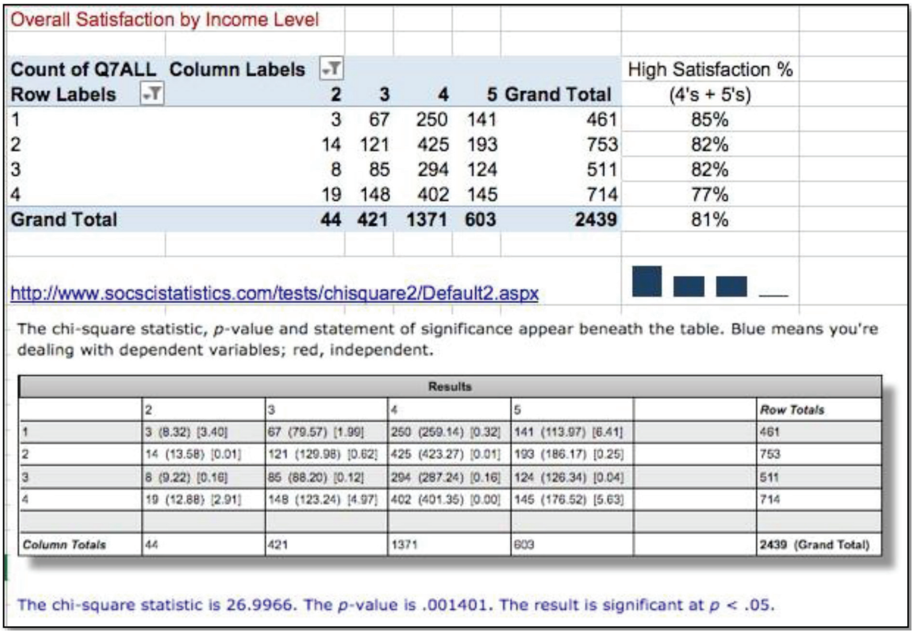
Results

	1	2	3	4	5	*Row Totals*
Female	5 (2.97) [1.39]	20 (23.76) [0.59]	225 (233.62) [0.32]	753 (774.12) [0.58]	373 (341.53) [2.90]	1376
Male	1 (3.03) [1.36]	28 (24.24) [0.58]	247 (238.38) [0.31]	811 (789.88) [0.56]	317 (348.47) [2.84]	1404
Column Totals	6	48	472	1564	690	2780 (Grand Total)

The chi-square statistic is 11.4404. The *p*-value is .022036. The result is significant at $p < .05$.

FIGURE 11.6 Computing the contingency table using a Pivot Table and performing a chi-square analysis using a Web-based calculator for a Titanic passenger survival analysis based on passenger class.

7. Create a Pivot Table of Q7ALL by Q21INCOME (Figure 11.7). Make sure to remove the "0" and "6" response rows for Q7 and the blank and "3" responses for Q21. Perform a chi-squared test to assure that the variables are dependent. Check the percentage approval using the sum of the number of the "4" and "5" ratings. Note the addition of a sparkline graph to visualize the differences across income levels.



Overall Satisfaction by Income Level

Count of Q7ALL	Column Labels					High Satisfaction %
Row Labels	2	3	4	5	Grand Total	(4's + 5's)
1	3	67	250	141	461	85%
2	14	121	425	193	753	82%
3	8	85	294	124	511	82%
4	19	148	402	145	714	77%
Grand Total	44	421	1371	603	2439	81%

http://www.socscistatistics.com/tests/chisquare2/Default2.aspx

The chi-square statistic, *p*-value and statement of significance appear beneath the table. Blue means you're dealing with dependent variables; red, independent.

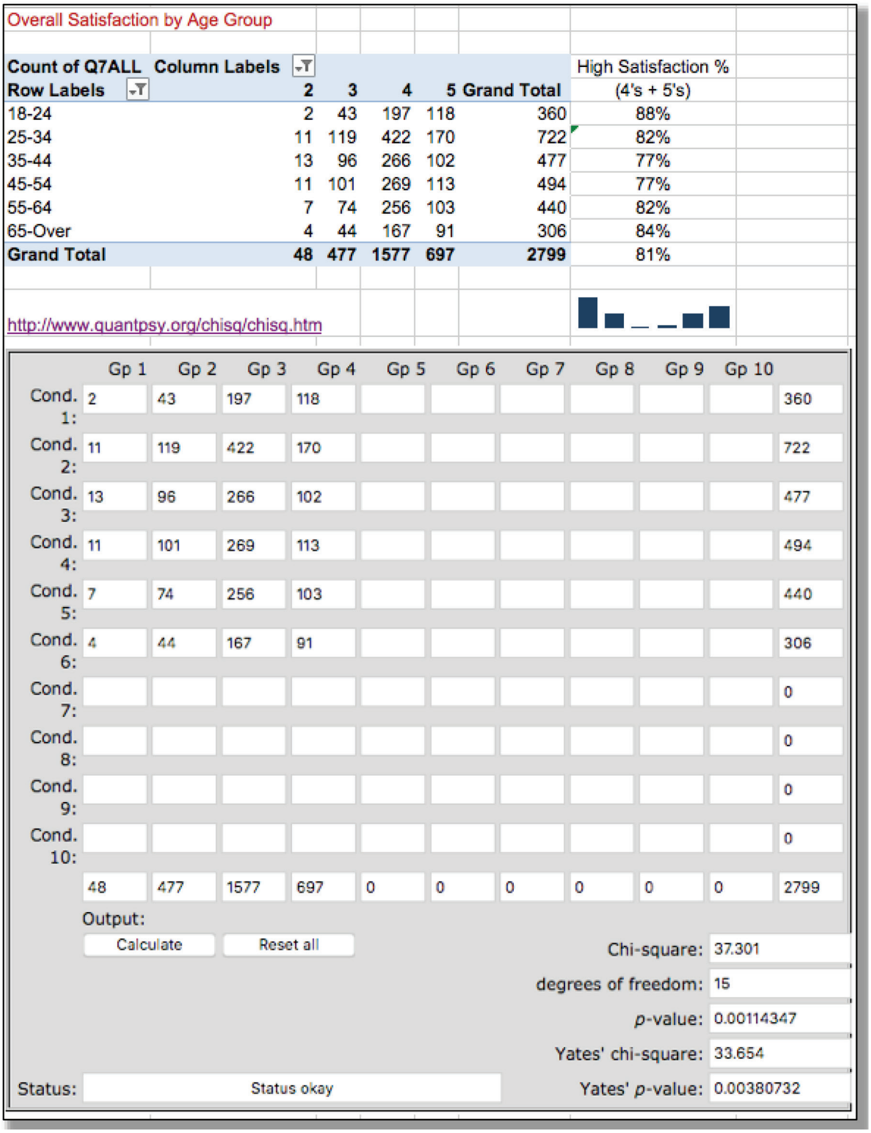
Results

	2	3	4	5		*Row Totals*
1	3 (8.32) [3.40]	67 (79.57) [1.99]	250 (259.14) [0.32]	141 (113.97) [6.41]		461
2	14 (13.58) [0.01]	121 (129.98) [0.62]	425 (423.27) [0.01]	193 (186.17) [0.25]		753
3	8 (9.22) [0.16]	85 (88.20) [0.12]	294 (287.24) [0.16]	124 (126.34) [0.04]		511
4	19 (12.88) [2.91]	148 (123.24) [4.97]	402 (401.35) [0.00]	145 (176.52) [5.63]		714
Column Totals	44	421	1371	603		2439 (Grand Total)

The chi-square statistic is 26.9966. The *p*-value is .001401. The result is significant at $p < .05$.

FIGURE 11.7 Computing a contingency table using a Pivot Table and performing a chi-square analysis using a Web-based calculator for a passenger satisfaction analysis based on passenger income level.

8. Create a Pivot Table of Q7ALL by Q22GENDER (Figure 11.8). Make sure to remove the "0" and "6" response rows for Q7 and the blank and "3" responses for Q22. Perform a chi-squared test to ensure that the variables are dependent. Check the percentage approval using the sum of the number of the "4" and "5" ratings. Also, note that we had to use a different online calculator for the chi-squared statistic since we have more than five levels for one of the variables (*www.quantpsy.org/chisq/chisq.htm*).



Overall Satisfaction by Age Group

Count of Q7ALL	Column Labels					High Satisfaction %
Row Labels	2	3	4	5	Grand Total	(4's + 5's)
18-24	2	43	197	118	360	88%
25-34	11	119	422	170	722	82%
35-44	13	96	266	102	477	77%
45-54	11	101	269	113	494	77%
55-64	7	74	256	103	440	82%
65-Over	4	44	167	91	306	84%
Grand Total	**48**	**477**	**1577**	**697**	**2799**	81%

http://www.quantpsy.org/chisq/chisq.htm

	Gp 1	Gp 2	Gp 3	Gp 4	Gp 5	Gp 6	Gp 7	Gp 8	Gp 9	Gp 10	
Cond. 1:	2	43	197	118							360
Cond. 2:	11	119	422	170							722
Cond. 3:	13	96	266	102							477
Cond. 4:	11	101	269	113							494
Cond. 5:	7	74	256	103							440
Cond. 6:	4	44	167	91							306
Cond. 7:											0
Cond. 8:											0
Cond. 9:											0
Cond. 10:											0
	48	477	1577	697	0	0	0	0	0	0	2799

Output:

Chi-square: 37.301

degrees of freedom: 15

p-value: 0.00114347

Yates' chi-square: 33.654

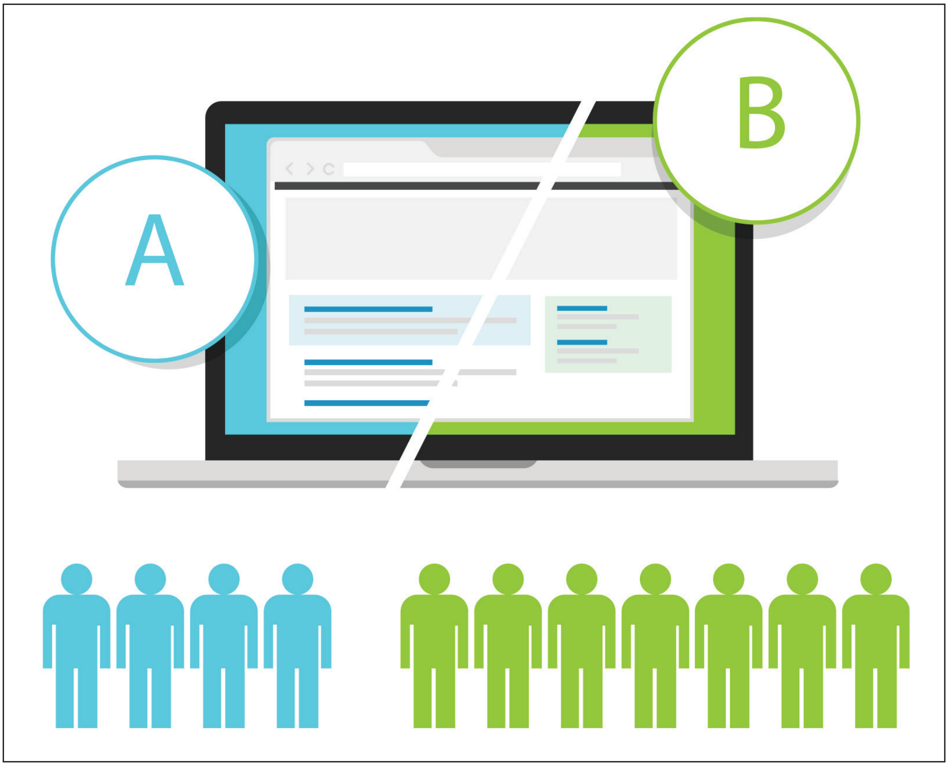
Yates' *p*-value: 0.00380732

Status: Status okay

FIGURE 11.8 Computing a contingency table using a Pivot Table and performing a chi-square analysis using a Web-based calculator for a passenger satisfaction analysis based on passenger age bracket.

CHAPTER 12

A/B TESTING



In Chapter 12, we continue our work of inferential statistical analysis for categorical variables started in Chapter 11. In this case we work with two categorical variables, each having one of two outcomes. We set up a simple one-factor two-level experiment. We make one change, say to a Website, before the change we call it B, and after the change we call it A. On the changed Web page, we have some desired outcome (the visitor pushes the buy button, or moves on to the next Web age, or donates, or votes). We are testing whether the change from B to A made a difference in the response rate. The first variable is the label of the test, either test A or test B. The second variable is the outcome (yes or no) for each visitor. We set up the contingency table on responses of each version of the Website and perform a chi-squared test to see if the difference in the ratios is significantly different from chance, to see if there is a bias toward one or the other of the two Website designs. The exercises present a ready-made template to compute the contingency table and perform the chi-squared test.

This technique answers the business questions: “Is there a significant difference between the outcomes of testing two versions of a proposed change?”

As in previous chapters, we demonstrate the technique in the first exercise and allow for more challenging work in subsequent exercises. We also provide a template for the easy calculation of the chi-squared test and the resulting p-value.

Analysis Case 12.1 – Design and Analysis of Trials

A/B Testing

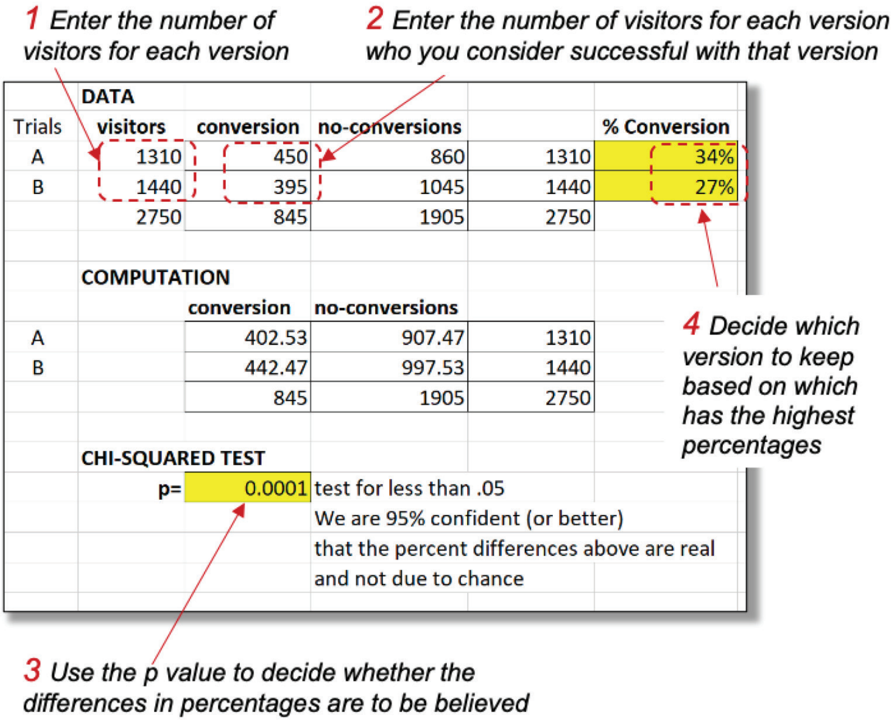
1. Using the *Lab Data* set provided, open the *Analysis Case 12* folder in it, and find the file *TestOfSignificanceForA/BTest.xls*.
2. Open *TestOfSignificanceForA/BTest.xls* using Excel.
3. Use the chi-squared test to decide which Web page changes to keep.

4. Enter the data in the appropriate places in the spreadsheet.

5. Let's analyze the following A/B data (Figure 12.1):

> *You made a significant change to the Website (Version A) from the original (Version B). In the last three days, there were 2,750 visitors to the Website, with 1,310 using Version A and 1,440 using Version B. Of those using Version A, 450 yielded positive results. Of those using Version B, 395 yielded positive results.*

6. Are you ready to switch to Version A? Why or why not?



1 Enter the number of visitors for each version

2 Enter the number of visitors for each version who you consider successful with that version

	DATA				
Trials	visitors	conversion	no-conversions		% Conversion
A	1310	450	860	1310	34%
B	1440	395	1045	1440	27%
	2750	845	1905	2750	
	COMPUTATION				
		conversion	no-conversions		
A		402.53	907.47	1310	
B		442.47	997.53	1440	
		845	1905	2750	
	CHI-SQUARED TEST				
	p=	0.0001	test for less than .05		
			We are 95% confident (or better)		
			that the percent differences above are real		
			and not due to chance		

4 Decide which version to keep based on which has the highest percentages

3 Use the p value to decide whether the differences in percentages are to be believed

FIGURE 12.1 Steps in using the *TestforSignificanceCalculatorFor A/B test* calculator.

Analysis Case 12.2 – Additional Case Using ORDERS

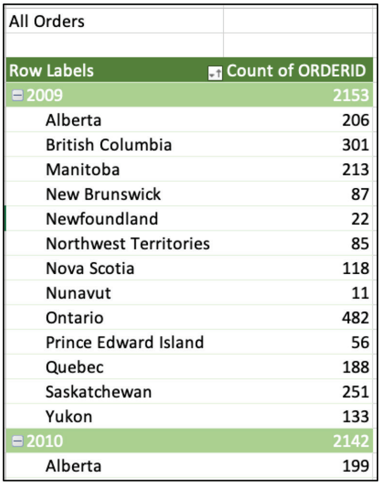
1. Using the *Lab Data* set and in the *Analysis Case 12* folder, find the file *ORDERS.xlsx*.

2. Open *ORDERS.xlsx* using Excel.

3. We are going to answer this question:

   > *In 2012, a policy was put in place to give bonuses to managers of provincial stores that increased the number of profitable orders by 10% in 2011. Was that policy change effective?*

4. Create a Pivot Table of the entire ORDERS table (Figure 12.2). Select ORDERDATE as the rows and group by year. Add PROVINCE under ORDERDATE in the columns of the Pivot Table. Tabulate the count of ORDERID to get a count of all the orders by year and by province.

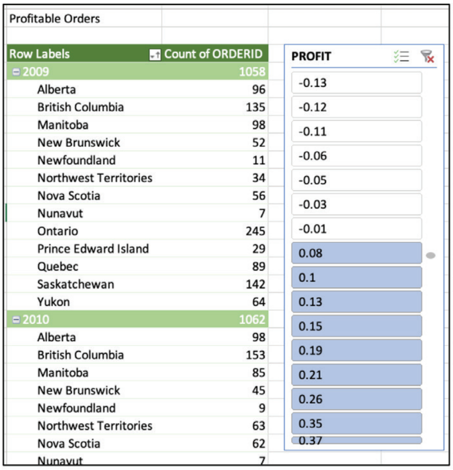


All Orders

Row Labels	Count of ORDERID
⊟ 2009	2153
Alberta	206
British Columbia	301
Manitoba	213
New Brunswick	87
Newfoundland	22
Northwest Territories	85
Nova Scotia	118
Nunavut	11
Ontario	482
Prince Edward Island	56
Quebec	188
Saskatchewan	251
Yukon	133
⊟ 2010	2142
Alberta	199

FIGURE 12.2 Pivot Table of number of orders by year and by province.

5. Create another Pivot Table as in the previous process but add a “slicer” (found on the Pivot Table “Format” ribbon) to only show the results of PROFIT for each order above 0 (in other words, only count orders that were profitable) (Figure 12.3).

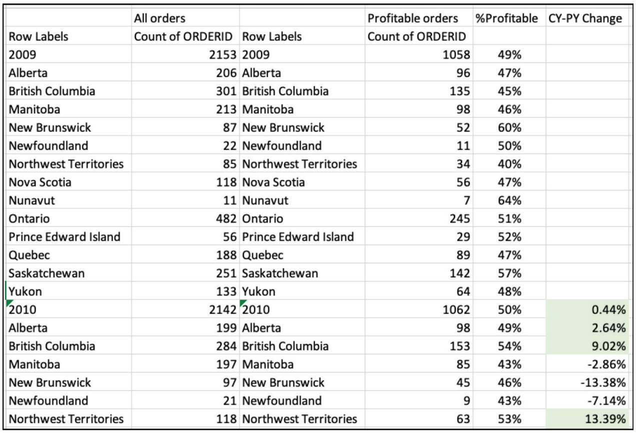


Profitable Orders

Row Labels	Count of ORDERID
2009	1058
Alberta	96
British Columbia	135
Manitoba	98
New Brunswick	52
Newfoundland	11
Northwest Territories	34
Nova Scotia	56
Nunavut	7
Ontario	245
Prince Edward Island	29
Quebec	89
Saskatchewan	142
Yukon	64
2010	1062
Alberta	98
British Columbia	153
Manitoba	85
New Brunswick	45
Newfoundland	9
Northwest Territories	63
Nova Scotia	62
Nunavut	7

FIGURE 12.3 Pivot Table of number of orders by year and by province filtered for profitable orders.

6. Copy and paste both lists side by side as values into another sheet. Compute the percent of orders that were profitable and enter into it another contiguous column (Figure 12.4). Also, in another column, compute the percent difference from one year to the next. We are looking for provinces with a difference greater than 10% between 2012 and 2011.

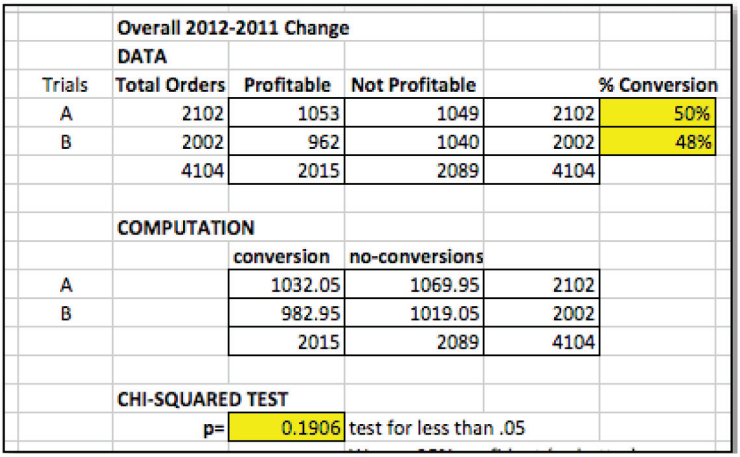


	All orders		Profitable orders	%Profitable	CY-PY Change
Row Labels	Count of ORDERID	Row Labels	Count of ORDERID		
2009	2153	2009	1058	49%	
Alberta	206	Alberta	96	47%	
British Columbia	301	British Columbia	135	45%	
Manitoba	213	Manitoba	98	46%	
New Brunswick	87	New Brunswick	52	60%	
Newfoundland	22	Newfoundland	11	50%	
Northwest Territories	85	Northwest Territories	34	40%	
Nova Scotia	118	Nova Scotia	56	47%	
Nunavut	11	Nunavut	7	64%	
Ontario	482	Ontario	245	51%	
Prince Edward Island	56	Prince Edward Island	29	52%	
Quebec	188	Quebec	89	47%	
Saskatchewan	251	Saskatchewan	142	57%	
Yukon	133	Yukon	64	48%	
2010	2142	2010	1062	50%	0.44%
Alberta	199	Alberta	98	49%	2.64%
British Columbia	284	British Columbia	153	54%	9.02%
Manitoba	197	Manitoba	85	43%	-2.86%
New Brunswick	97	New Brunswick	45	46%	-13.38%
Newfoundland	21	Newfoundland	9	43%	-7.14%
Northwest Territories	118	Northwest Territories	63	53%	13.39%

FIGURE 12.4 Computation of percent difference in orders by province from current year (CY) compared to prior year (PY).

7. Note that the overall CY-PY difference between 2012 and 2011 is only 2%, and Quebec and PEI were the only two that had percent changes greater than 10%. Of the two, Quebec is more important, since it had ten times more orders than PEI.

8. Use the *TestForSignificanceofABTest.xlsx* worksheet found in the *Lab Files* folder under the *Tools* folder to analyze the A/B data (Figure 12.5). Compute the chi-squared statistic for the A/B test for overall orders and for Quebec. We see that, in the case of Quebec, the change was significant. Overall, the 2% change was not significant.



	Overall 2012-2011 Change				
	DATA				
Trials	Total Orders	Profitable	Not Profitable		% Conversion
A	2102	1053	1049	2102	50%
B	2002	962	1040	2002	48%
	4104	2015	2089	4104	
	COMPUTATION				
		conversion	no-conversions		
A		1032.05	1069.95	2102	
B		982.95	1019.05	2002	
		2015	2089	4104	
	CHI-SQUARED TEST				
	p=	0.1906	test for less than .05		

FIGURE 12.5 Using A/B testing to test for significance of the CY/PY profit change for the Quebec province.

Analysis Case 12.3 – Additional Case Using SFO Airport Survey Data

1. Use the latest SFO Airport ACQ Survey data downloaded from *www.flysfo.com/media/customer-survey-data.*

2. Or using the *Lab Data* set provided in the *Analysis Case 12* folder, find the file *2016_SFO_Customer_Survey_Data.xls*.

3. Make sure also to download the previous year's data and data dictionary.

4. For the purposes of this lab, and in case the Website is not reachable, you may use the data sets in the Lab Files folder in the Lab Data set under the Raw Data folder. Use the 2016 and 2015 data sets.

5. We are going to answer two questions:

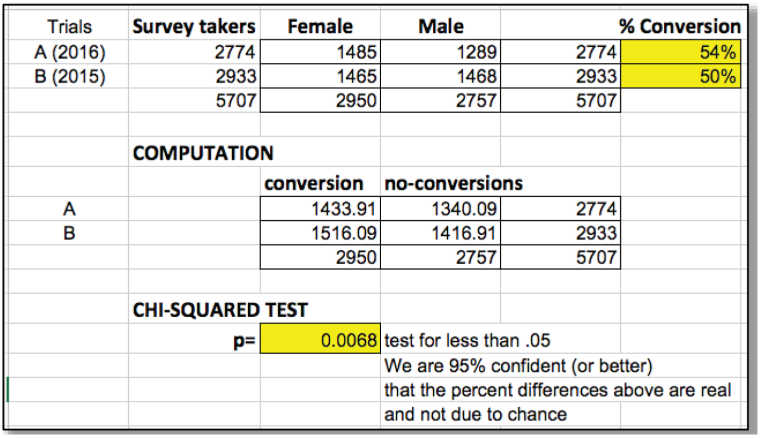
   Given that SFO launched a marketing campaign in 2016 to attract more female customers, did the campaign succeed?

   Given that the airport increased its cleaning staff and the frequency of cleaning in 2016, did the efforts yield greater customer satisfaction?

6. Open the data files using Excel.

7. To answer the first questions, create Pivot Tables to tabulate the ratio of male to female customers in the latest year (or 2016) and prior year (or 2015).

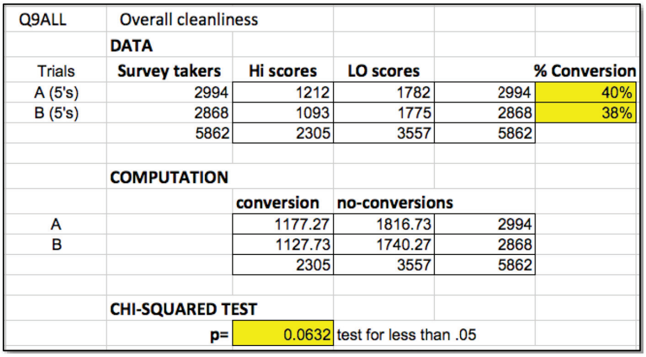
8. Using the *TestForSignificanceofABTest.xlsx* worksheet found in the *Lab Files* folder under the *Tools* folder, analyze the A/B data (Figure 12.6). Was the change significant? Could you state with confidence that the campaign succeeded?



Trials	**Survey takers**	**Female**	**Male**		**% Conversion**
A (2016)	2774	1485	1289	2774	54%
B (2015)	2933	1465	1468	2933	50%
	5707	2950	2757	5707	
	COMPUTATION				
		conversion	**no-conversions**		
A		1433.91	1340.09	2774	
B		1516.09	1416.91	2933	
		2950	2757	5707	
	CHI-SQUARED TEST				
	p=	0.0068	test for less than .05		
			We are 95% confident (or better)		
			that the percent differences above are real		
			and not due to chance		

FIGURE 12.6 A/B test results showing the significant change of the 2015/2016 marketing campaign to please female airport customers.

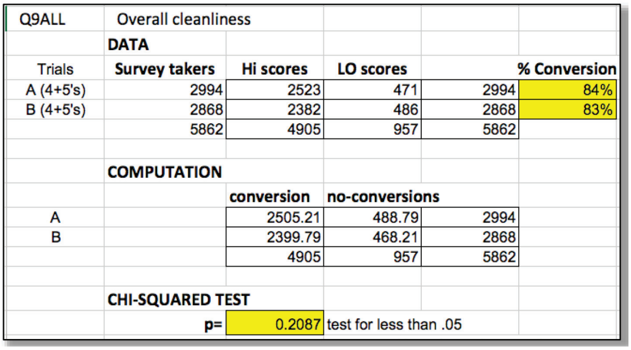
9. To answer the second question, repeat the process on Q9ALL on cleanliness (Figure 12.7). Here we will consider success if the increased number of passengers rated cleanliness a "5." Did the percentage of "5" ratings significantly increase from one year to the next under the new cleaning regime? Since p > .05, the answer is no.



Q9ALL	Overall cleanliness				
	DATA				
Trials	Survey takers	Hi scores	LO scores		% Conversion
A (5's)	2994	1212	1782	2994	40%
B (5's)	2868	1093	1775	2868	38%
	5862	2305	3557	5862	
	COMPUTATION				
		conversion	no-conversions		
A		1177.27	1816.73	2994	
B		1127.73	1740.27	2868	
		2305	3557	5862	
	CHI-SQUARED TEST				
	p=	0.0632	test for less than .05		

FIGURE 12.7 A/B test results showing significant difference in satisfaction by the 2015/2016 cleaning improvement campaign to please "highly satisfied" (5s only) airport customers.

10. What if we changed the condition to an increase in the number of customers who gave a rating of "4" or "5" from one year to the next versus a "1," "2," or "3"? Did that change the outcome of the analysis? Again, since p > .05, the answer is no (Figure 12.8).

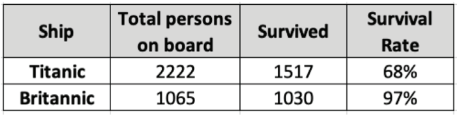


Q9ALL	Overall cleanliness				
	DATA				
Trials	Survey takers	Hi scores	LO scores		% Conversion
A (4+5's)	2994	2523	471	2994	84%
B (4+5's)	2868	2382	486	2868	83%
	5862	4905	957	5862	
	COMPUTATION				
		conversion	no-conversions		
A		2505.21	488.79	2994	
B		2399.79	468.21	2868	
		4905	957	5862	
	CHI-SQUARED TEST				
	p=	0.2087	test for less than .05		

FIGURE 12.8 A/B test results showing no significant change by the 2015/2016 cleaning improvement campaign in pleasing "satisfied" (4s and 5s) airport customers.

Analysis Case 12.4 – Additional Analysis Cases Using Titanic Data

1. The Titanic had a sister ship, the Britannic, built almost identically. Due to the Titanic disaster, the shipbuilders made the Britannic ship safer in case of disaster. Unfortunately, the Britannic sank almost in the same way as the Titanic (though from a human-made disaster—a torpedo during WWI—and not an iceberg, but she sank nevertheless). The survival rates, however, were dramatically different. The data may be obtained from *www.titanicfacts.net/* (Figure 12.9).



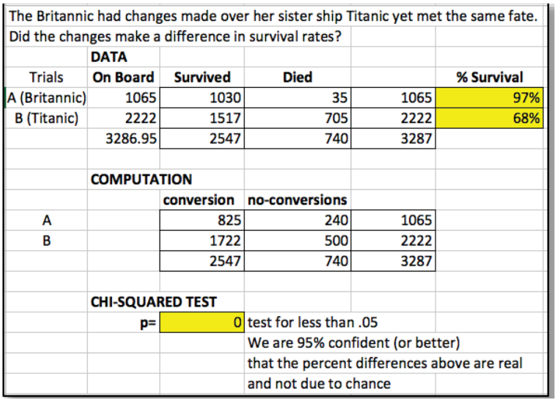
Ship	Total persons on board	Survived	Survival Rate
Titanic	2222	1517	68%
Britannic	1065	1030	97%

FIGURE 12.9 Survival statistics of the Titanic and Britannic transatlantic ships.

2. The question we want to answer is:

   Did the changes make a significant difference in survival rates?

3. We set up the analysis as an A/B test and, using the *TestForSignificanceOfABTest.xlsx* tool, we can compute a p-value to guide our answer (Figure 12.10).



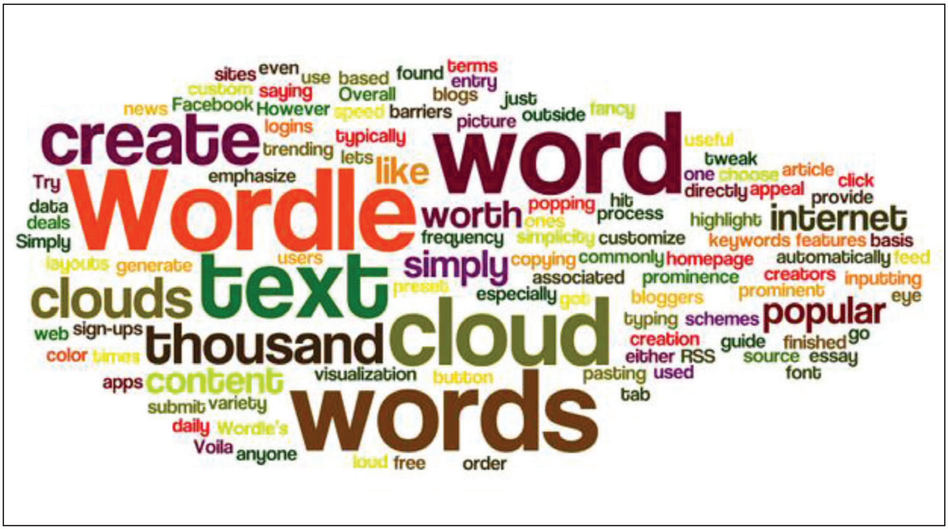
The Britannic had changes made over her sister ship Titanic yet met the same fate.					
Did the changes make a difference in survival rates?					
	DATA				
Trials	On Board	Survived	Died		% Survival
A (Britannic)	1065	1030	35	1065	97%
B (Titanic)	2222	1517	705	2222	68%
	3286.95	2547	740	3287	
	COMPUTATION				
		conversion	no-conversions		
A		825	240	1065	
B		1722	500	2222	
		2547	740	3287	
	CHI-SQUARED TEST				
	p=	0	test for less than .05		
			We are 95% confident (or better)		
			that the percent differences above are real		
			and not due to chance		

FIGURE 12.10 A/B test results showing that the Britannic survival rate was statistically significantly higher than the Titanic's.

4. The answer is: the changes made a significant difference.

CHAPTER 13

TEXT ANALYTICS



Chapter 13 helps you explore an exciting new area of data analysis: text analytics. New tools have recently been developed to extract meaning from unstructured text data easily. That data may come from open-ended responses in surveys, or tweets, emails, or Facebook postings. It could be a database of contracts or a collection of books in electronic form. Some of the tools are functions in Excel. We use the COUNTIF function to estimate the sentiment analysis in product reviews. We also use open-source Web-based text analytic tools to create word clouds and perform simple word frequency analysis to extract some underlying meaning from text.

To exemplify the techniques, we load the text files of five travel books, amounting to over one million words of text, and perform some fundamental analysis. The situation is analogous to extracting meaning from a corpus of Facebook postings, email logs, or Twitter feeds.

This technique answers the business question: "What are they saying?"

As in previous chapters, we demonstrate the technique in the first exercise and allow for more challenging work in subsequent exercises.

Analysis Case 13.1 – Unstructured Text Analysis I

1. Using the Lab Data set provided, open the *Lab Files* folder and in the *Analysis Case 13* folder find these text files:

   InnocentsAbroadMarkTwain.txt

   MagellanVoyagesAnthonyPiagafetta.txt

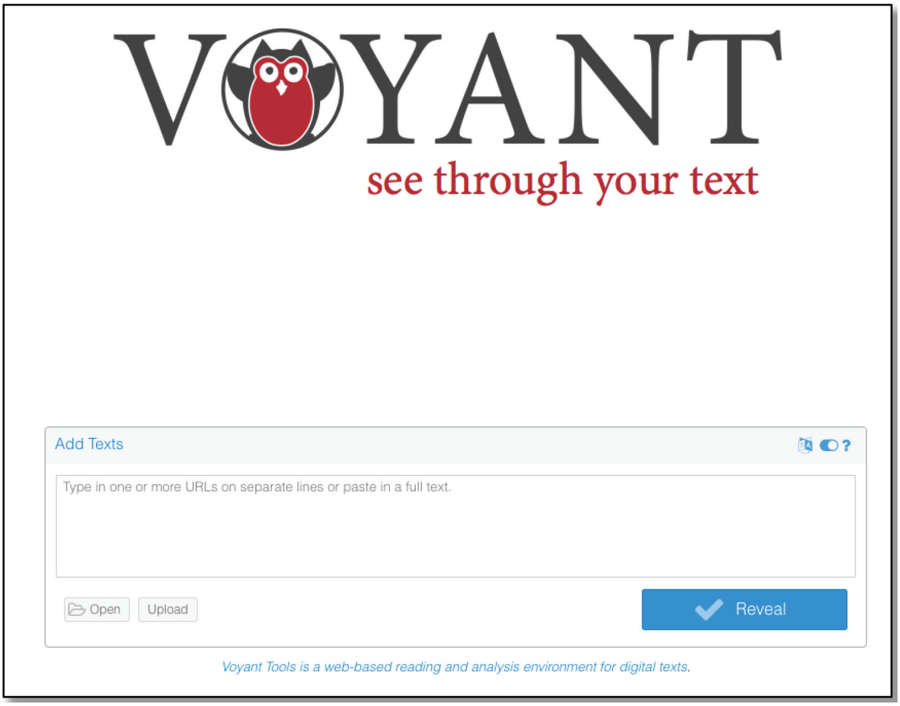
   TheAlhambraWashingtonIrving.txt

   TravelsOfMarcoPolo.txt

   VoyageOfTheBeagleDarwin.txt

2. Use a Web browser with access to the Internet.

3. Load the Voyant text analysis program found at *https://voyant-tools.org/* (Figure 13.1)

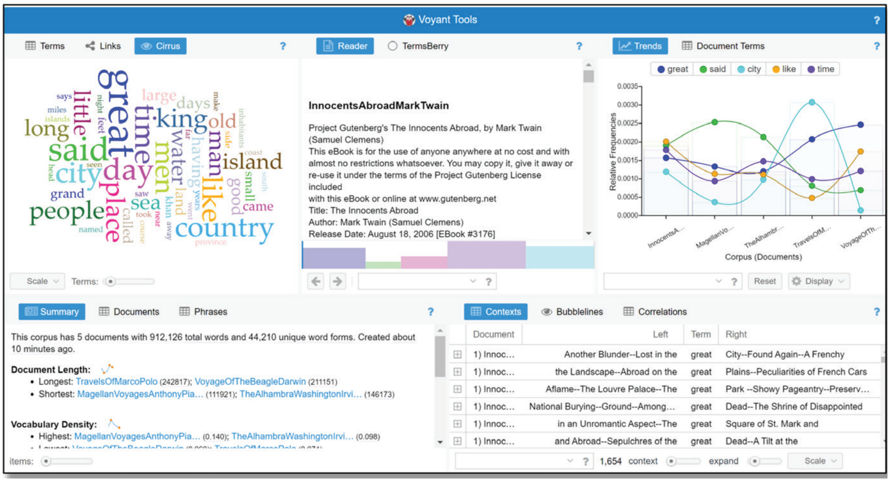


Type in one or more URLs on separate lines or paste in a full text.

Voyant Tools is a web-based reading and analysis environment for digital texts.

FIGURE 13.1 Web-based text analytic tool data entry screen.

4. Load all five texts into the corpus for analysis (Figure 13.2). Use the resulting analysis to explore the texts. Notice the resulting word cloud, a very popular analysis tool for text data.



InnocentsAbroadMarkTwain

Project Gutenberg's The Innocents Abroad, by Mark Twain (Samuel Clemens)
This eBook is for the use of anyone anywhere at no cost and with almost no restrictions whatsoever. You may copy it, give it away or re-use it under the terms of the Project Gutenberg License included
with this eBook or online at www.gutenberg.net
Title: The Innocents Abroad
Author: Mark Twain (Samuel Clemens)
Release Date: August 18, 2006 [EBook #3176]

This corpus has 5 documents with 912,126 total words and 44,210 unique word forms. Created about 10 minutes ago.

Document Length:

- Longest: TravelsOfMarcoPolo (242817); VoyageOfTheBeagleDarwin (211151)
- Shortest: MagellanVoyagesAnthonyPia... (111921); TheAlhambraWashingtonIrvi... (146173)

Vocabulary Density:

- Highest: MagellanVoyagesAnthonyPia... (0.140); TheAlhambraWashingtonIrvi... (0.098)

Document	Left	Term	Right
1) Innoc...	Another Blunder--Lost in the	great	City--Found Again--A Frenchy
1) Innoc...	the Landscape--Abroad on the	great	Plains--Peculiarities of French Cars
1) Innoc...	Aflame--The Louvre Palace--The	great	Park --Showy Pageantry--Preserv...
1) Innoc...	National Burying--Ground--Among...	great	Dead--The Shrine of Disappointed
1) Innoc...	in an Unromantic Aspect--The	great	Square of St. Mark and
1) Innoc...	and Abroad--Sepulchres of the	great	Dead--A Tilt at the

FIGURE 13.2 Results of analyzing one million words of text in a corpus of five travel books.

5. In the center of the display, there is a search box. Use it to search for the term "volcano."

   Which traveler saw volcanoes?

6. Search for "island."

   Which traveler seems to have more references to islands?

7. Repeat for the word "sea."
8. Do these results make sense?
9. Explore by searching for more single words and word combinations on your own.
10. Look for other text visualization tools at: *https://voyant-tools.org/docs/#!/guide/tools*.

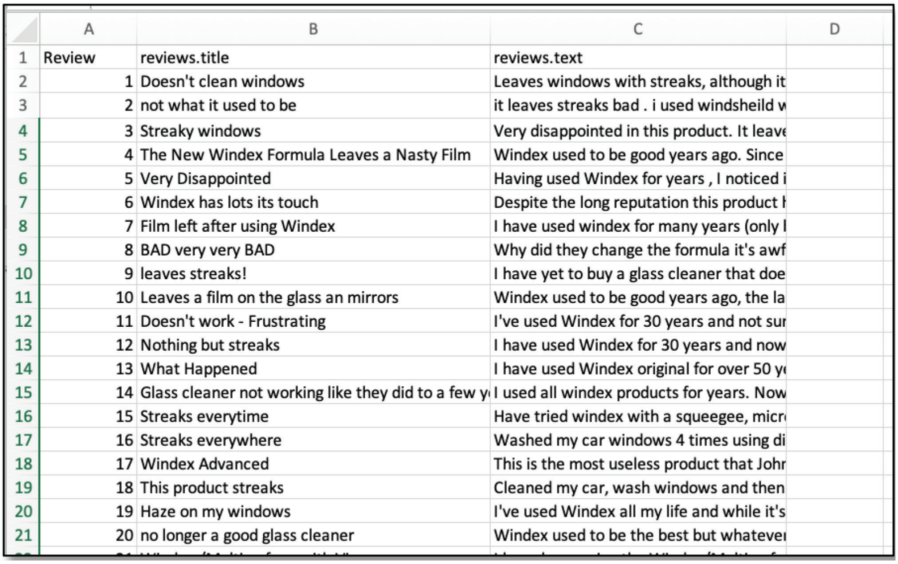
Analysis Case 13.2 – Unstructured Text Analysis II

Using Excel for Sentiment Analysis

1. We will use the Excel COUNTIF function to generate an analysis of text fields for the purposes of performing customer sentiment analysis.
2. We will answer this question:

   Do customers feel positively or negatively about a product?

3. Using the Lab Data set provided, open the *Lab Files* folder and in the *Analysis Case 13* folder find the *Product Reviews.xlxs* data file.
4. Navigate to the Product Reviews worksheet. Scrape the rows for the Windex brand (709–1056). Paste the rows into a new worksheet and label the worksheet *Windex*. Scrape the top row of the *Product Review* worksheet with the variable names and insert them in the top row of the newly created worksheet.
5. Scrape the Windex *reviews.text* and *reviews.title* columns. Paste the columns into a new worksheet and label it *Windex Reviews*. Move the reviews.text column to column C. Add the title of Reviews to column A and inset consecutive numbers down the column. Notice that there are 348 reviews. You should have something like what is shown in Figure 13.3:



	A	B	C	D
1	Review	reviews.title	reviews.text	
2	1	Doesn't clean windows	Leaves windows with streaks, although it	
3	2	not what it used to be	it leaves streaks bad . i used windsheild w	
4	3	Streaky windows	Very disappointed in this product. It leave	
5	4	The New Windex Formula Leaves a Nasty Film	Windex used to be good years ago. Since	
6	5	Very Disappointed	Having used Windex for years , I noticed i	
7	6	Windex has lots its touch	Despite the long reputation this product h	
8	7	Film left after using Windex	I have used windex for many years (only I	
9	8	BAD very very BAD	Why did they change the formula it's awf	
10	9	leaves streaks!	I have yet to buy a glass cleaner that doe	
11	10	Leaves a film on the glass an mirrors	Windex used to be good years ago, the la	
12	11	Doesn't work - Frustrating	I've used Windex for 30 years and not sur	
13	12	Nothing but streaks	I have used Windex for 30 years and now	
14	13	What Happened	I have used Windex original for over 50 ye	
15	14	Glass cleaner not working like they did to a few ye	I used all windex products for years. Now	
16	15	Streaks everytime	Have tried windex with a squeegee, micro	
17	16	Streaks everywhere	Washed my car windows 4 times using di	
18	17	Windex Advanced	This is the most useless product that John	
19	18	This product streaks	Cleaned my car, wash windows and then	
20	19	Haze on my windows	I've used Windex all my life and while it's	
21	20	no longer a good glass cleaner	Windex used to be the best but whatever	

FIGURE 13.3 Extracted product reviews for the Windex product.

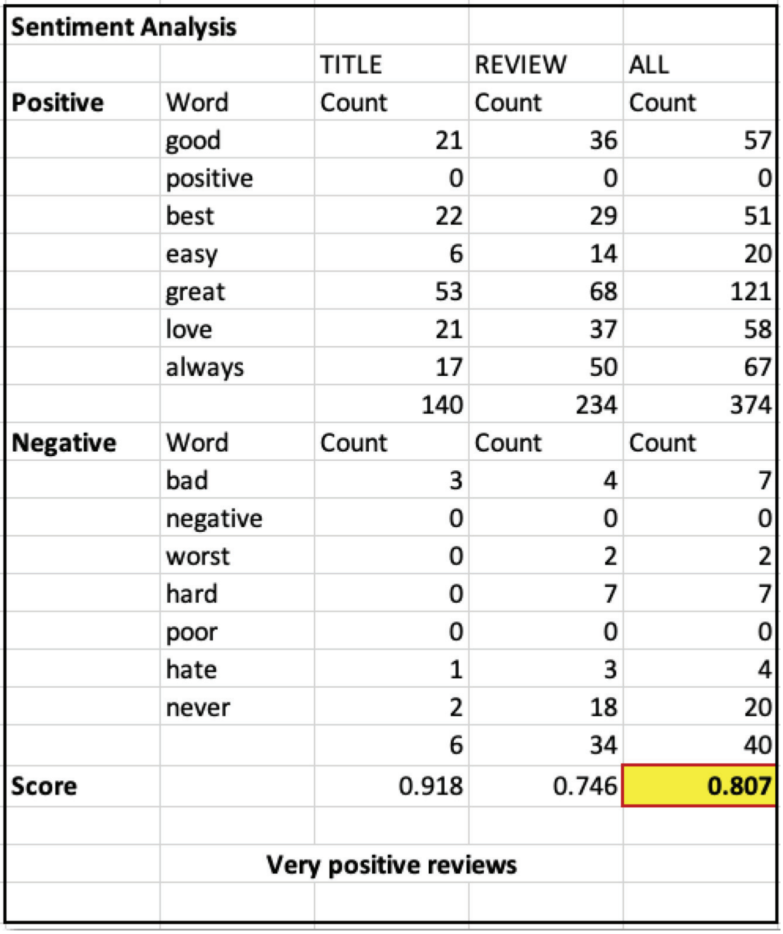
6. We will use both the title text and the review text columns to assess sentiment.

7. Create a list of words expressing likes for the product (e.g., "good," "love," etc.) and enter them in one column in the spreadsheet near to the data set. Create another list of words expressing dislike for the product (it could be the opposite of the "good" words, e.g., "bad," "hate," etc.).

8. Use the COUNTIF function to count all occurrences of your words for liking and disliking in both the review and title text columns. The form of the COUNTIF function should be something like this for each column:

```
=COUNTIF(B$2:B$348,"*good*")
```

9. Tabulate and summarize the results as in Figure 13.4. Score the sentiment of the reviews by subtracting total dislikes from the total likes and dividing by the total count. If the number is positive, then we can conclude people feel “good” about the product; if negative, we can conclude otherwise. The closer the number is to +1, the more they love the product, and vice versa.



Sentiment Analysis				
		TITLE	REVIEW	ALL
Positive	Word	Count	Count	Count
	good	21	36	57
	positive	0	0	0
	best	22	29	51
	easy	6	14	20
	great	53	68	121
	love	21	37	58
	always	17	50	67
		140	234	374
Negative	Word	Count	Count	Count
	bad	3	4	7
	negative	0	0	0
	worst	0	2	2
	hard	0	7	7
	poor	0	0	0
	hate	1	3	4
	never	2	18	20
		6	34	40
Score		0.918	0.746	**0.807**
	Very positive reviews			

FIGURE 13.4 Sentiment analysis of the Windex product from text comments in the reviews using the Excel COUNTIF function.

10. Keep in mind these are very approximate numbers (not very exact or scientific answers) and very dependent on the list of words we count. But it gives a good overall business answer.

11. Now copy the text in columns B2:C348 (both title and text) into the computer buffer. We are going to compute a word cloud and compare it to the COUNTIF score.

12. Use the Voyant tool explored earlier in this chapter and create a word cloud and do a word frequency analysis. Be sure to remove product-specific words from the STOPLIST (such as Windex, product, reviews). See Figure 13.5. Do the sentiment analysis computed with COUNTIF and what you see in the word cloud match?

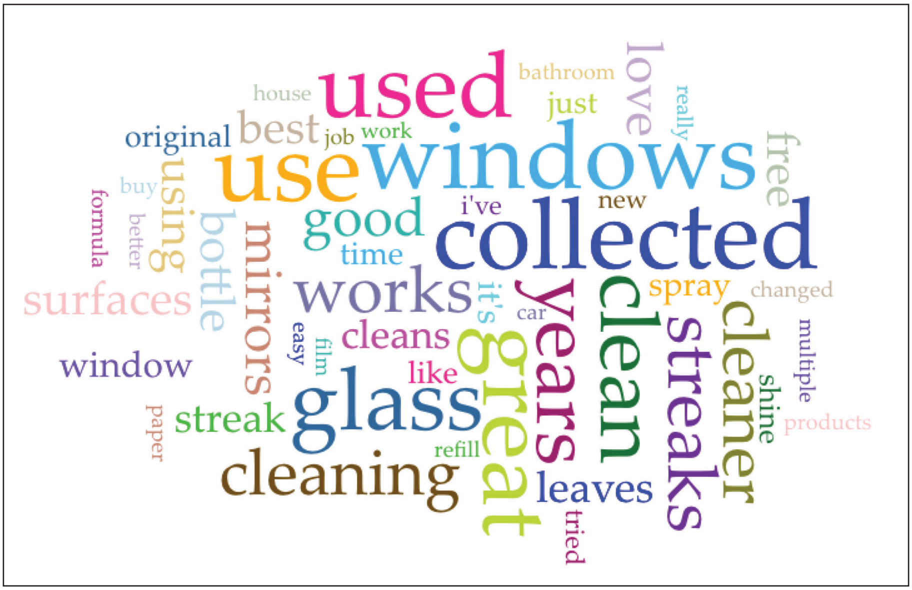
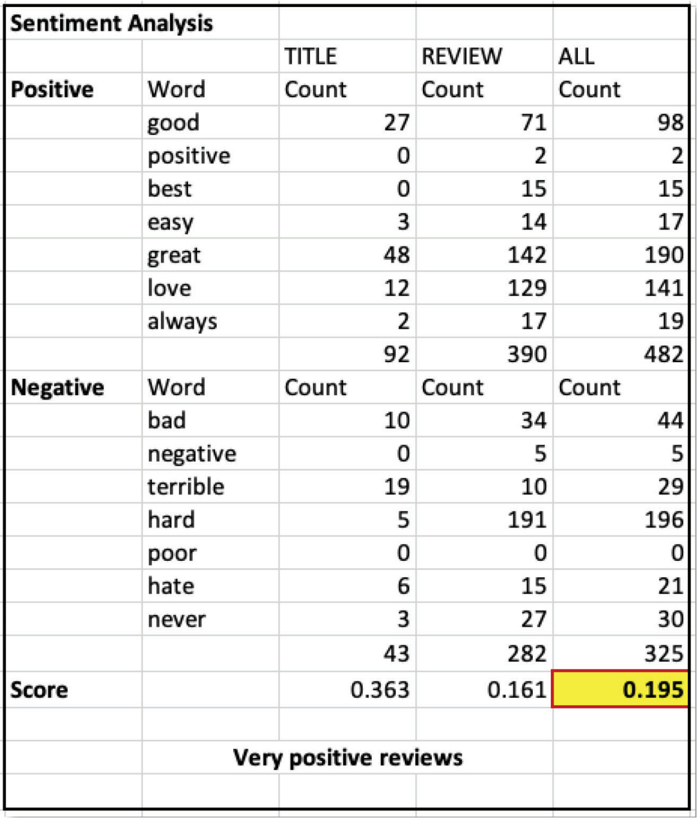


FIGURE 13.5 Sentiment analysis using a word cloud analysis of the Windex product from text comments in the product reviews.

13. Repeat steps 4–9 for the brand Rubbermaid. The results are mixed, still positive but not by much. The computed sentiment score is shown in Figure 13.6.



Sentiment Analysis				
		TITLE	REVIEW	ALL
Positive	Word	Count	Count	Count
	good	27	71	98
	positive	0	2	2
	best	0	15	15
	easy	3	14	17
	great	48	142	190
	love	12	129	141
	always	2	17	19
		92	390	482
Negative	Word	Count	Count	Count
	bad	10	34	44
	negative	0	5	5
	terrible	19	10	29
	hard	5	191	196
	poor	0	0	0
	hate	6	15	21
	never	3	27	30
		43	282	325
Score		0.363	0.161	**0.195**
		Very positive reviews		

FIGURE 13.6 Sentiment analysis using the Excel COUNTIF function of the Rubbermaid product from text comments extracted from the product reviews.

14. Use the Voyant tool explored earlier in this chapter and create a word cloud and do a word frequency analysis. Don't forget to add common words to the STOPLIST (mop, Rubbermaid, bottle, etc.). Do the sentiment analyses computed with COUNTIF and the word cloud match? (Figure 13.7.)

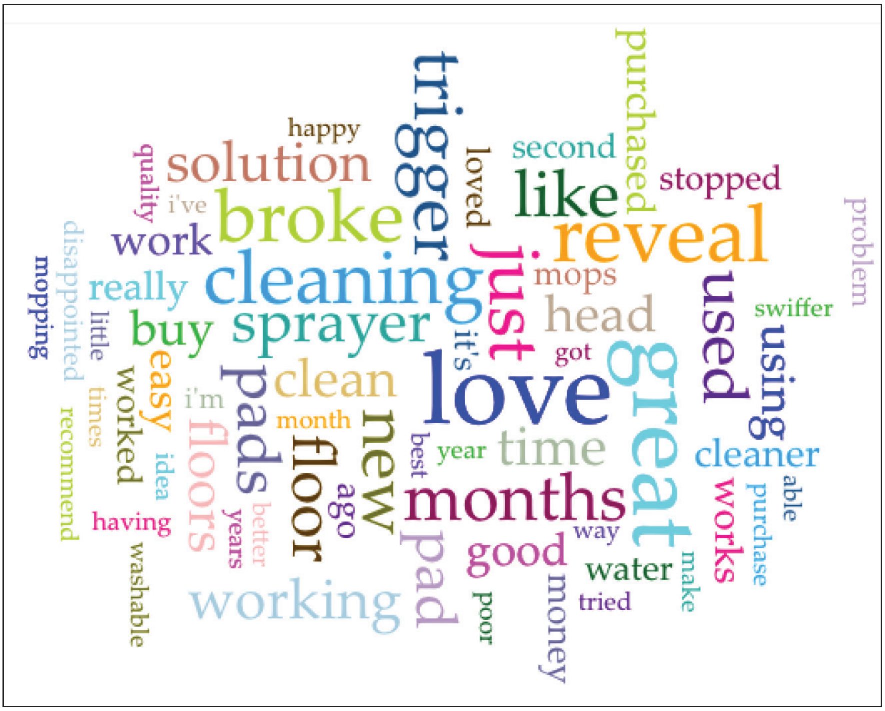
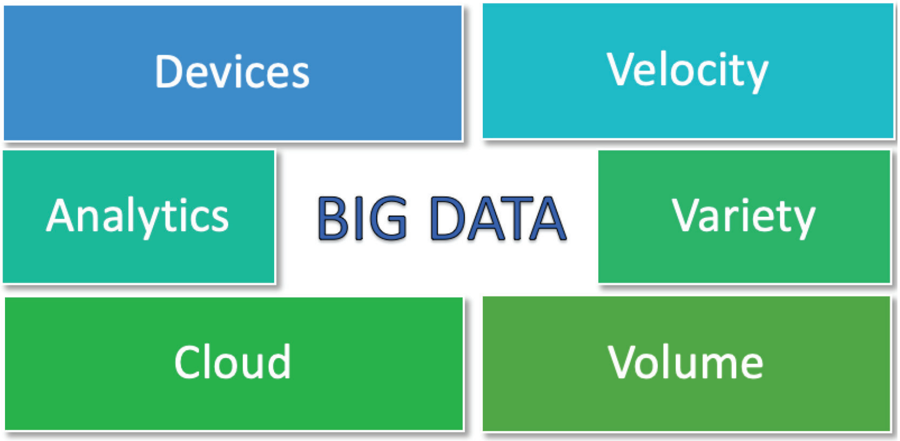


FIGURE 13.7 Sentiment analysis using a word cloud analysis of the Rubbermaid product from text comments in the product reviews.

CHAPTER 14

ANALYZING BIG DATA SETS



Chapter 14 presents techniques useful when dealing with data sets too large to load into Excel. We want to use all the valuable techniques developed in Chapters 3–12, but if only we can get the table into a spreadsheet. One useful way is to randomly sample the too-big-to-fit table and analyze the sampled table made up of the sampled rows. Excel has a randomization function, and we could use it to extract the sample rows. But, wait, we can't get the table into Excel! So we must use a different tool to perform the sampling. This is where we bring in the use of the R program. There is an exercise in this chapter where you are guided on how to set up and use R to extract a meaningful sample of rows for a large data set. You are also shown how to compute how many rows your sample you will need to obtain statistically significant results using the sample table. Once the sample rows are extracted, Excel may be used to get useful answers from the skills learned in earlier chapters.

This technique answers the business question: "How do we work with tables too large to load into Excel?"

As in previous chapters, we demonstrate the technique in the first exercise and allow for more challenging work in subsequent exercises.

Analysis Case 14.1 – Big Data Analysis

Using Sampling to Work with Large Data Files

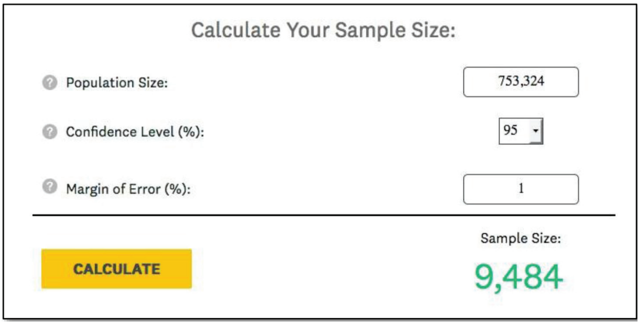
1. The premise of this exercise is that we wish to use Excel as our analysis tool but are aware of its limitations with respect to very large files. Typically, it's not the number of variables that is the problem but rather the number of rows, as we saw from Analysis Case 1. Let's say we have a very large data file, hundreds of megabytes consisting of hundreds of thousands or perhaps millions of rows. How do we use Excel in this case when we can't load the entire file in a spreadsheet? The answer is a trade-off. We are willing to accept a slight decrease in accuracy in our statistical results for the convenience of using Excel for the analysis.

2. The technique is to randomly sample the large (or big data) file and obtain a random sample of manageable rows of data. We will first use one tool to compute an adequate sample size, and then we will use another tool to sample the original file. We will use a free Web-based tool to compute sample size, and then we will use a free cloud-based program, RStudio, to extract a random sample.

Table 14.1 Characteristics of the data files used to demonstrate the sampling of large data sets.

Name	Size (MB)	Rows	Columns	Source	Description
ORDERS.csv	1.8	8,400	22	Company	Office supplies orders
Community.csv	70	376,000	551	US Census	2013 ACS census file
Courses.csv	73	631,139	21	MIT	edX 2013 MOOC Courses
BankComplaints.csv	306	753,324	18	US FTC	Bank complaints to the FTC

3. First, let's compute an adequate sample size. The entire file is our population. For example, we wish to have .95% confidence in our statistical analysis using our sample and to have no more than a 1% margin of error in our results (these are very typical parameters in business). Let's take the 306 MB *BankComplaints.csv* big data file with 753,324 rows. Using an online sample size calculator found at *www.surveymonkey.com/mp/sample-size-calculator/*, we see that we will need a random sample of 9,484 rows to achieve our desired level of accuracy and margin of error (Figure 14.1).



Calculate Your Sample Size:

Population Size: 753,324

Confidence Level (%): 95

Margin of Error (%): 1

Sample Size: 9,484

FIGURE 14.1 Using an online sample size calculator.

4. As an additional exercise use the online calculator to compute the necessary number of random rows in the other sample files for various accuracy levels in the following table (Table 14.2). Note that the rightmost column has the answer!

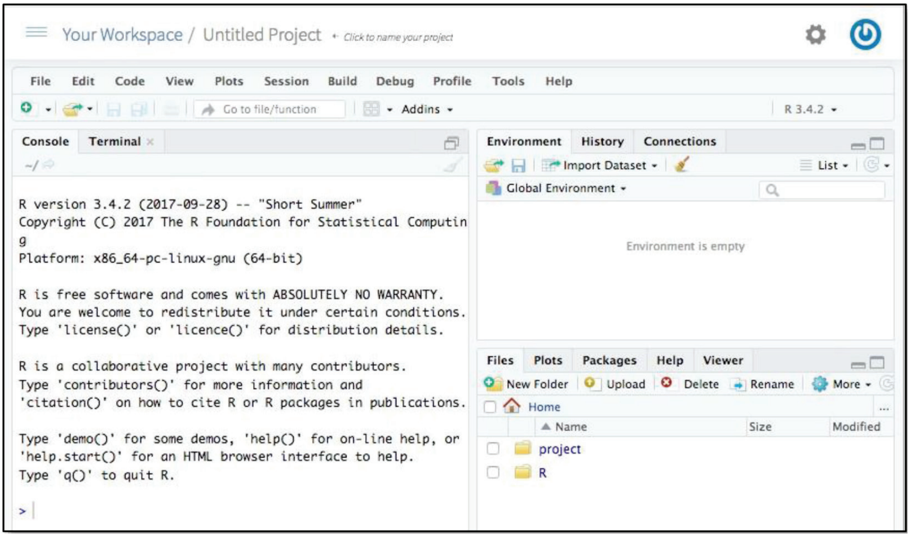
Table 14.2 Computed elements of the sampling of data sets under study.

Name	Size (MB)	Population Rows	Confidence Level %	Margin of Error %	Random Sample Rows
ORDERS.csv	1.8	8,400	95	1	4,482
Community.csv	70	376,000	95	1	9,365
Community.csv	70	376,000	99	1	15,936
Courses.csv	73	631,139	95	1	9,461
Courses.csv	73	631,139	95	2	2,394

5. We will now use a popular free cloud version of the R program: RStudio Cloud. (If you feel ambitious, download and install RStudio on your computer so you will have a permanently installed sample extraction tool for future use. Otherwise, proceed to learn the technique with the cloud version.)

6. Navigate to *https://rstudio.cloud/*, create a free account, and then proceed to the next step.

7. In RStudio Cloud, create a new project. The typical RStudio interface appears. Note the ">_" prompt in the lower-left-hand corner of the screen. It should be blinking, waiting for your R commands. The resulting screen in your browser should look something like this (Figure 14.2):

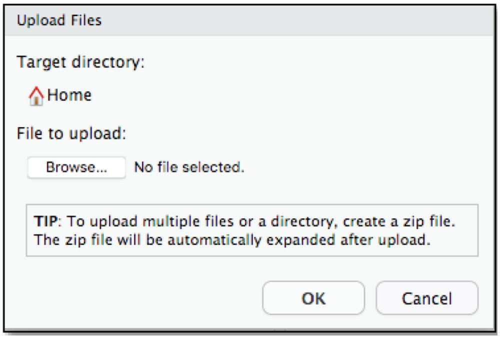


Your Workspace / Untitled Project
Click to name your project
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 3.4.2
Console
Terminal
~/
Environment History Connections
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder Upload Delete Rename More
Home
Name Size Modified
project
R

```
R version 3.4.2 (2017-09-28) -- "Short Summer"
Copyright (C) 2017 The R Foundation for Statistical Computin
g
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

FIGURE 14.2 Interface screen of RStudio Cloud cloud-based platform.

8. First, we will upload all the files we will be sampling.
9. Using the Lab Data set provided, open the *Analysis Cases* folder and then open the *Analysis Case 14* folder within it, and find the files *ORDERS.csv*, *Courses.csv*, and *Community.csv*.
10. Click on the "Files" tab in the lower-right-hand pane of the RStudio desktop on your browser. Then, click "Upload" in the new row. You will get the following interface (Figure 14.3):



Upload Files

Target directory:

Home

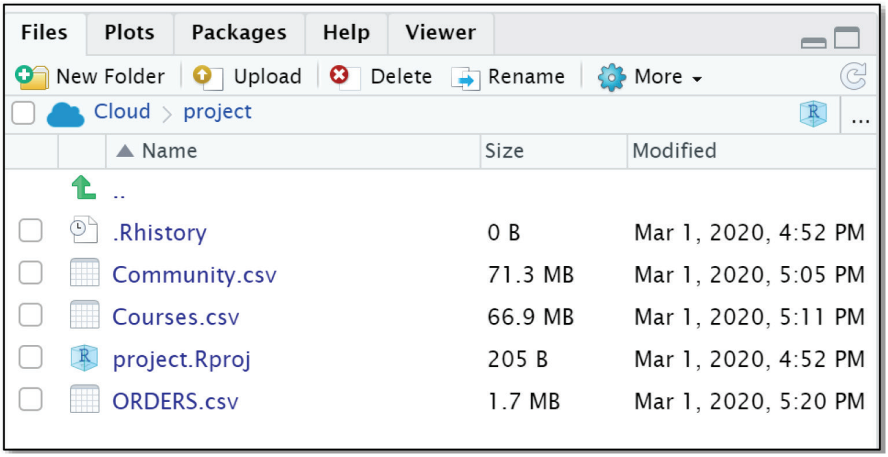
File to upload:

Browse... No file selected.

TIP: To upload multiple files or a directory, create a zip file.
The zip file will be automatically expanded after upload.

FIGURE 14.3 Details of using the RStudio Cloud tool to upload files to the Web to be analyzed.

11. Click the “Browse” button and upload each of the three files, one after the other. Be patient, as some of the larger files take some time to upload. When done, the “File” area in the upper-right-hand screen should like this (Figure 14.4):



	Name	Size	Modified
	..		
□	.Rhistory	0 B	Mar 1, 2020, 4:52 PM
□	Community.csv	71.3 MB	Mar 1, 2020, 5:05 PM
□	Courses.csv	66.9 MB	Mar 1, 2020, 5:11 PM
□	project.Rproj	205 B	Mar 1, 2020, 4:52 PM
□	ORDERS.csv	1.7 MB	Mar 1, 2020, 5:20 PM

FIGURE 14.4 Screen of the uploaded data files ready to be processed with an R script.

12. We will start with sampling the smaller file (*ORDERS*) and then move on to the larger files.

13. In the upper-left-hand panel, pull down the “File” > “Open” function and select the *ORDERS.csv* file from the list. That loads the file into the workspace (note the “Source” panel now appears and has information about the file).

14. Drop down to the lower-left-hand panel and click in front of the “>_”cursor. It should start blinking, ready for your command.

15. We will enter the following sets of commands one after the other:

```
> set.seed(123)
> Y <- read.csv(“ORDERS.csv”)
> View(Y)
> index <- sample (1:nrow(Y), 4482)
> Z <- Y[index, ]
> View(Z)
> write.csv(Z,'Z.csv')
```

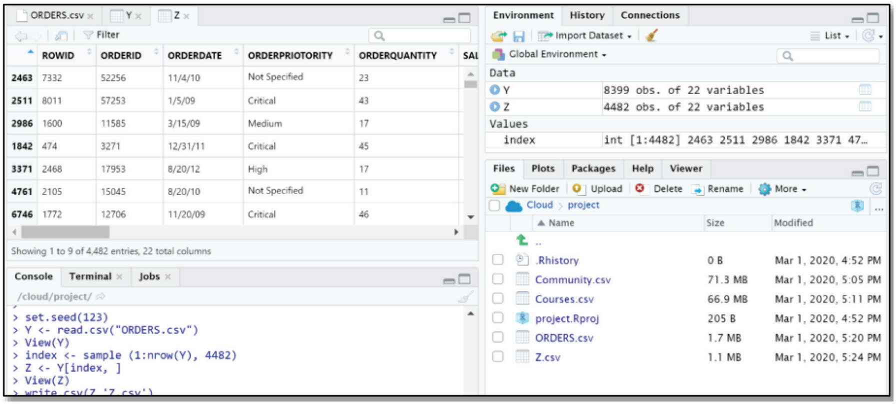
16. Make sure to enter the random number of rows required (4482), but without a comma, or the comma will be interpreted as a part of the command and not as part of the number.

17. We are using "Y" and "Z" as temporary containers for our data.

18. Note that the "Source" upper-left-hand panel shows the original data in table form (the result of the "View" command).

19. Also, note that the upper-right-hand panel shows two files in the workspace, Y and Z, and their characteristics. Note that Y has the original set of rows, 8,399, and Z has the sample rows, 4,482. The random sampling was done with the "sample" command.

20. We outputted the sample rows to the Z file, and the program wrote it out to the disk as *Z.csv*. Now the lower-right-hand panel has that file in the directory (Figure 14.5).



	ROWID	ORDERID	ORDERDATE	ORDERPRIOTORITY	ORDERQUANTITY	SAL
2463	7332	52256	11/4/10	Not Specified	23	
2511	8011	57253	1/5/09	Critical	43	
2986	1600	11585	3/15/09	Medium	17	
1842	474	3271	12/31/11	Critical	45	
3371	2468	17953	8/20/12	High	17	
4761	2105	15045	8/20/10	Not Specified	11	
6746	1772	12706	11/20/09	Critical	46	

Showing 1 to 9 of 4,482 entries, 22 total columns

```
> set.seed(123)
> Y <- read.csv("ORDERS.csv")
> View(Y)
> index <- sample (1:nrow(Y), 4482)
> Z <- Y[index, ]
> View(Z)
> write.csv(Z,'Z.csv')
```

```
Data
 Y        8399 obs. of 22 variables
 Z        4482 obs. of 22 variables
Values
  index   int [1:4482] 2463 2511 2986 1842 3371 47…
```

Name	Size	Modified
..		
.Rhistory	0 B	Mar 1, 2020, 4:52 PM
Community.csv	71.3 MB	Mar 1, 2020, 5:05 PM
Courses.csv	66.9 MB	Mar 1, 2020, 5:11 PM
project.Rproj	205 B	Mar 1, 2020, 4:52 PM
ORDERS.csv	1.7 MB	Mar 1, 2020, 5:20 PM
Z.csv	1.1 MB	Mar 1, 2020, 5:24 PM

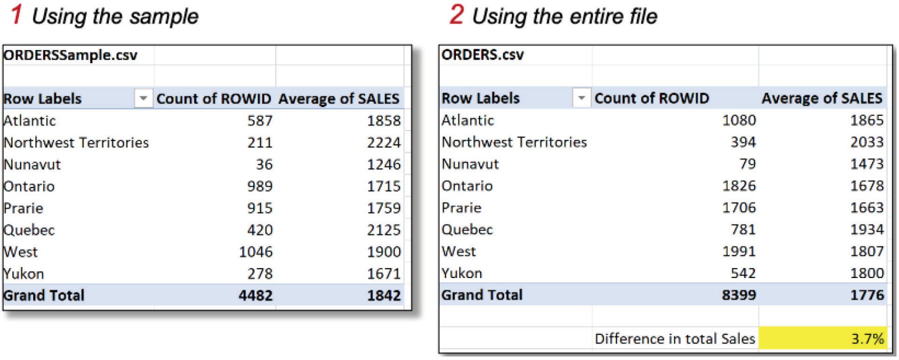
FIGURE 14.5 RStudio Cloud interface screen showing the data file (upper left), R script (lower left), details of the input and output data files (upper right), and files in a directory (lower right).

21. Now we need to download the file from the cloud directory to our computer. You will want to checkmark the box next to the *Z.csv* file. In the lower-right-hand panel, click on the "More" icon (it looks like a blue gear). Select "Export" and follow the directions to download the file to your desktop for now. You should rename the file *ORDERSSample.csv* as you save it. (It is important to note that we only used Y and Z as temporary, easy-to-use containers.)

22. To check on our work, we will compute some results using both the original population and the sample rows and compare.

23. Open *ORDERS.csv* and *ORDERSSample.csv.* Notice that the sample data set contains a new column (at the extreme left) that identifies each sample row uniquely (a random number). You need to label that column (for example, "SAMPLEID").

24. Using pivot tables, tabulate the total sales by region for both files. Compare the results from both tables (Figure 14.6). Compute the difference between the total population and the sample. You will find it to be well within the 5% margin of error.



1 Using the sample

ORDERSSample.csv

Row Labels	Count of ROWID	Average of SALES
Atlantic	587	1858
Northwest Territories	211	2224
Nunavut	36	1246
Ontario	989	1715
Prarie	915	1759
Quebec	420	2125
West	1046	1900
Yukon	278	1671
Grand Total	**4482**	**1842**

2 Using the entire file

ORDERS.csv

Row Labels	Count of ROWID	Average of SALES
Atlantic	1080	1865
Northwest Territories	394	2033
Nunavut	79	1473
Ontario	1826	1678
Prarie	1706	1663
Quebec	781	1934
West	1991	1807
Yukon	542	1800
Grand Total	**8399**	**1776**
	Difference in total Sales	3.7%

FIGURE 14.6 Comparison of the same analysis using the entire file and the sample showing less than 5% error difference.

25. Note that whereas the computed total from the sampled file is quite accurate when compared to that computed using the entire original file, there is a much wider error in the individual

regional results, especially for those regions with fewer rows. If you repeat for the PROFIT variable rather than SALES, you will see a much wider variation.

Repeat these steps using the two other data files as additional exercises.

26. Repeat the process for the *Community.csv* and *Courses.csv* files for a 95% confidence level and a 2% margin of error. Compute the summary of one of the variables for both the total population and the sampled files and compare.

Analysis Case 14.2 – Additional Case Using the BankComplaints Big Data File

1. You will find that if you try to load the *BankComplaints.csv* 300 MB file in RStudio Cloud, it will give you an error. The free cloud version only allows smaller files to load. One solution is to get a paid subscription and continue, but if we are only using R for its easy sampling capability, it may pay to stay with free versions of RStudio (or find yourself some other way to sample very large data files).

2. Our proposal is for you to install RStudio on your PC or Mac computer. Then you can use the techniques of the previous exercise as they are given. (The interface to RStudio is identical, so just follow the instructions given, except now you can load a 300 MB or 3 GB or whatever size file you need to sample.)

3. As a first step, locate the free RStudio program on the Internet and download and install. You may obtain it here: *www.rstudio.com/products/rstudio/download/*.

4. Once installed, try it out on the 300 MB *BankComplaints.csv* file. Compute the number of random rows to select for an adequate sample for a 95% confidence level and a 1% margin of error (Table 14.3).

Table 14.3 Computed parameters of the sampling of the data set under study.

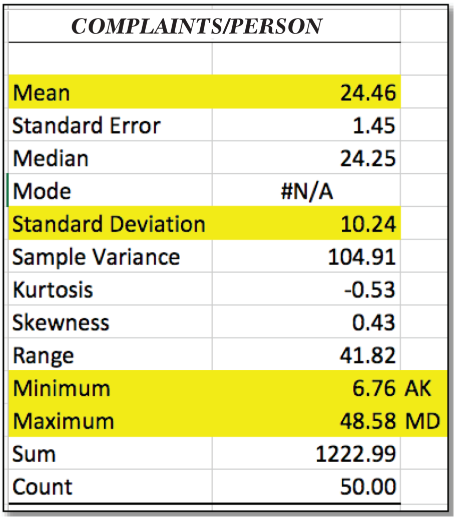
Name	Size (MB)	Population Rows	Confidence Level %	Margin of Error %	Random Sample Rows
BankComplaints.csv	306	753,324	95	1	4,484

5. Use the R commands given earlier to sample the file and save it as *BankComplaintsSample.csv*. (Make sure to use the correct file name in the commands.)

6. Use the file of samples to tabulate the percentage of complaints by state to discover the states with the most and the least complaints.

7. Add the size of the population of each state and normalize the complaintsper million residents of each state. Get the states with the least and the most complaints per capita. Compute other descriptive statistics of this variable.

8. Using the Analysis ToolPak, get summary descriptive statistics (Figure 14.7).

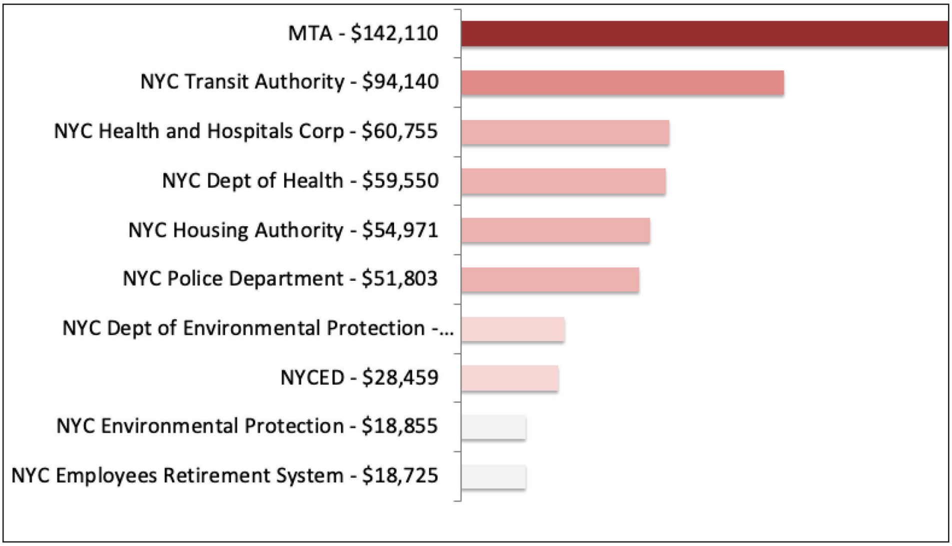


COMPLAINTS/PERSON		
Mean	24.46	
Standard Error	1.45	
Median	24.25	
Mode	#N/A	
Standard Deviation	10.24	
Sample Variance	104.91	
Kurtosis	-0.53	
Skewness	0.43	
Range	41.82	
Minimum	6.76	AK
Maximum	48.58	MD
Sum	1222.99	
Count	50.00	

FIGURE 14.7 Descriptive statistics of the sample extracted from the *BankComplaints.csv* data file.

CHAPTER 15

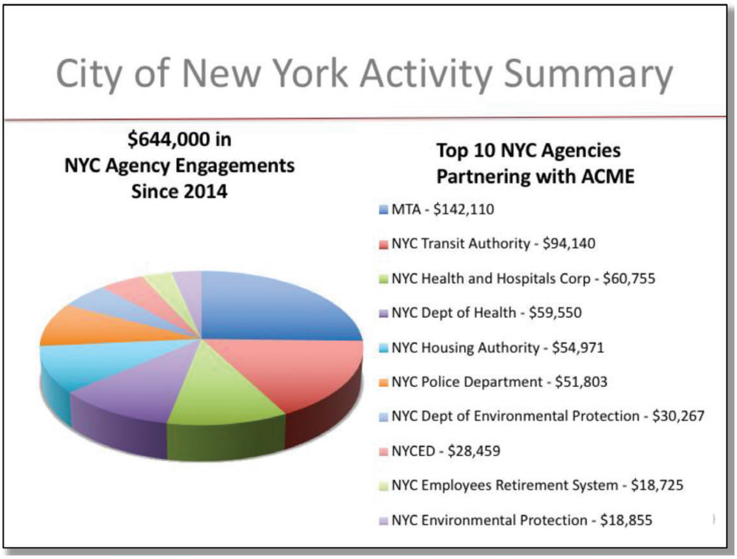
DATA VISUALIZATION



The Case

Consider being called into a meeting to review a pitch to be used by your company for a customer meeting. This company has been doing business with the City of New York for many years, and they want to reinforce the partnership they have with the City. Part of the pitch is to demonstrate how good a vendor they have been to the City by showing how much business various city agencies have done with your company. The sales department intends to use the following slide (Figure 15.1) to demonstrate the state of the relationship as a vendor. It is your job to analyze and suggest changes to the slide to make the case more compelling.

It is the purpose of this lab to analyze the slide and data visual and create a new slide. We will analyze the slide along six major dimensions: Story, Signs, Purpose, Perception, Methods, and Charts. These six dimensions are described at the start of each of the six labs in this chapter. We will use a template with a checklist of appropriate questions that need to be answered during the analysis.



City of New York Activity Summary

$644,000 in NYC Agency Engagements Since 2014

Top 10 NYC Agencies Partnering with ACME

- 🟦 MTA - $142,110
- 🟥 NYC Transit Authority - $94,140
- 🟩 NYC Health and Hospitals Corp - $60,755
- 🟪 NYC Dept of Health - $59,550
- 🟦 NYC Housing Authority - $54,971
- 🟧 NYC Police Department - $51,803
- 🟦 NYC Dept of Environmental Protection - $30,267
- 🟥 NYCED - $28,459
- 🟩 NYC Employees Retirement System - $18,725
- 🟪 NYC Environmental Protection - $18,855

FIGURE 15.1 Case study slide to be analyzed and modified and made more compelling.

The Analysis Template

We will use the analysis template provided as follows to analyze and improve charts, or data visualization, along the six dimensions outlined. The six dimensions are discussed in detail at the start of each of the six cases that follow. We are using the tool as a guide to the eye and to improve your discrimination. Using the analysis template and the associated questions produces an imperfect indicator, but it should be sufficient to guide you in making improvements to any chart. We hope that by using the tool multiple times, you will then begin to internalize the questions and, eventually, you will not need to use the tool.

To use the tool, consider answering the question associated with each principle. Consider each question and then use the following rough criteria: answer YES if the chart fulfills the question for the most part (>70%) and NO if the chart is deficient in that question (<70%). Then elaborate: what would you do to improve the chart in this dimension? This process requires a rough pass-fail judgment on the issue. It's not perfect, and it is not meant to create an exact measurement of the visual's perfection. It is meant to develop and refine your power of discrimination when analyzing and improving charts: what looks good and what does not.

Story

Create a Visual Story – *Is the point of the visual very clear?*

Make It a Prop – *Has the visual been simplified and focused?*

Emulate Legendary Storytellers – *Are past masters and the basic charts that they pioneered emulated?*

Signs

Signs – *Is the use of signs and symbols appropriate?*

Communication – *Is the signal-to-noise ratio high?*

Function – *Is the chart functionally informational rather than beautiful art?*

Purpose

Need – *Does the chart fulfill organizational information needs?*

Audience – *Does the chart allow for audience biases, needs, and journeys?*

Frame – *Does the visual answer a well-framed analytical question?*

Perception

Seeing – *Does the eye of the viewer focus on the most important point being made?*

Mind – *Have the principles of the Gestalt psychology of perception been thoughtfully employed in the visual?*

Quality – *Does the visual inform the viewer and dispel his ignorance?*

Method

Color – *Is color used judiciously and sparsely?*

Chart junk – *Is the visual clear of unnecessary visual elements not leading to a clear point being* made?

Title – *Does the title of the chart convey the point being made with the chart?*

Charts

Right Chart – *Does the type of chart being used match the level of judgment required?*

Selection – *Does the chart type used match the business question being answered?*

Tables – *Are referenceable visuals (tables) readable with appropriate conditional formatting and thumbnail graphs used for emphasis?*

Analysis Case 15.1 – Story

Create a Visual Story – *Is the point of the visual very clear?*

As the saying goes, a picture is worth a thousand words. Does your visual make a clear point that would take a lot of words to convey? What is the point of the visual? Does your audience get the point? Is your story clear?

Make It a Prop – *Has the visual has been simplified and focused?*

Famous chart maker Stephen Few admonishes us that numbers have an important story to tell. Our numbers rely on us to give them a compelling voice. A data visual should not tell the whole story but be a prop to be used by the storyteller. Charts support the storyteller. Charts are not the whole story. Charts are a complement for the storyteller by summarizing complex data in a single image.

Emulate Legendary Storytellers – *Are past masters and the basic charts that they pioneered emulated?*

We have a rich data visualization heritage; we stand on the shoulders of giants, as it were. They are master storytellers who invented and used data visuals as their props. The question is, can you improve your visual by emulating famous chart makers? Are your visuals rooted in their iconic chart exemplars? For example, see the word of folks like John Snow with his London Cholera Map; Charles Minard and Napoleon's March on Moscow; Hans Rosling who invented Gapminder and moving bubble diagrams; Joseph Priestley and his Chart of Biography and his New Chart of History; and finally, Florence Nightingale and her radar charts of the Crimean War.

Case Analysis

Review the case study and the chart to be improved. Use this exercise to practice analyzing a chart along the Story dimension. Make sure to use the analysis questions from the template, repeated as follows. In the end, after you have gone through the three principles, you will be able to answer:

> *What is wrong with the visual being analyzed along these dimensions?*
>
> *What can be done to improve the visual along all these dimensions?*

1. Create a Visual Story – Is the point of the visual very clear?
2. Make It a Prop – Has the visual been simplified and focused?
3. Emulate Legendary Storytellers

Analysis Case 15.2 – Signs

Give Them a Sign – *Is the use of signs and symbols appropriate?*

People need a strong sign to be able to make good decisions. Are we using signs and symbols properly in our visuals? Does the visual use cultural cues properly? Do we transgress any cultural conventions? Is our audience expecting the symbolism we use in our chart, or are they surprised and confused by it? The science of sign making has three parts: (a) the signifier, the intended meaning; (b) the signified or significant, which is the symbol or icon that stands in for the signifier (for example, a "dog" is represented by a "picture of a dog"); and (c) the sign, the combination that makes up our understanding.

It's Like a Communication System – *Is the signal-to-noise ratio high?*

Sign making for our charts is part of setting up a communication system. We must ask, as with any communication system, does our visual send a strong, unmistakable signal? Will the receiver, our audience, be able to decode it? Or is there too much noise, and what can we do to reduce it? In an effective presentation, the audience gets your point even in the presence of noise.

Design for Function – *Is the chart functionally informational rather than beautiful art?*

We should ask ourselves as we create our charts: does the chart inform or entertain? We should avoid having our chart be prized and classified as beautiful art, especially if it fails to inform. We are talking about functionally over making the chart pretty. Avoid frilliness. Have we sacrificed clarity to make our chart pleasing to the eye? We should strive to inform and not to create a chart that appeals to emotions, but rather to reason leading to good decisions.

Case Analysis

Review the case study and the chart to be improved. Use this exercise to practice analyzing a chart along the Sign dimension. Make sure to use the analysis questions from the template, repeated as follows. In the end, after you have gone through the three principles, you will be able to answer:

> *What is wrong with the visual being analyzed along these dimensions?*
>
> *What can be done to improve the visual along all these dimensions?*

1. Signs – Is the use of signs and symbols appropriate?
2. Communication – Is the signal-to-noise ratio high?
3. Function – Is the chart functionally informational rather than beautiful art?

Analysis Case 15.3 – Purpose

Consider the Information Need – *Does the chart fulfill organizational information needs?*

We use our presentation and its embedded graphs to fulfill the needs of the requester and of the organization. We should make sure our visuals fulfill an organizational, informational need. We should only be bringing data that is vital to the organization and its mission. We should also consider if the requester will be satisfied with this level of information and the news you bring. And most important, does the visual help them make a decision? In other words, does it educate your requester and audience sufficiently to satisfy their needs?

Consider the Audience – *Does the chart allow for audience biases, needs, and journeys?*

Together with the audience's information needs, we should consider all other aspects of our viewers and listeners. What are their biases, what journey are they on, what will they do with the information? Then we must match the visual style to the audience's biases, needs, and journey. The visual must take into consideration their point of view and account for their biases, education, and training. It should help them with their journey to make their numbers. Any mismatch would introduce noise, perhaps confuse, and the audience would miss the point you are making.

Answer Well-Framed Analytical Questions – *Does the visual answer a well-framed analytical question?*

In the end, to satisfy the information need, we must present our results of answering well-framed analytical questions stemming from those needs. The analytical questions were posed and answered as part of the analysis process. The creation of the communication set of visuals is not the time to discover or search for the answers. We select, out of the many analytical questions we used to inform ourselves of the answers, those few that are the most important. They contain the key evidence, the facts, that support our conclusions. Those key facts must pop out of our visuals as clearly evident.

Case Analysis

Review the case study and the chart to be improved. Use this exercise to practice analyzing a chart along the Purpose dimension. Make sure to use the analysis questions from the template, repeated as follows. In the end, after you have gone through the three principles, you will be able to answer:

> *What is wrong with the visual being analyzed along these dimensions?*
>
> *What can be done to improve the visual along all these dimensions?*

1. Need – Does the chart fulfill organizational information needs?
2. Audience – Does the chart allow for audience biases, needs, and journeys?
3. Frame – Does the visual answer a well-framed analytical question?

Analysis Case 15.4 – Perception

Use the Eye-Brain System of Seeing – *Does the eye of the viewer focus on the most important point being made?*

The eye is attracted unconsciously to strong focal points in images. Therefore, you should decide which are the most important things you want your viewers to focus on and make them stand out when they first view the chart. They can use Gestalt principles, color theory, employment of the right chart, removing chart junk, and other methods to assist in focusing. You have to guide their viewing so they get the point, almost unconsciously.

Employ the Gestalt Principles of Perception – *Have the principles of the Gestalt psychology of perception been thoughtfully employed in the visual?*

One way to assure that we are guiding the eye to the most important elements of the chart we have designed is to use the principles of the Gestalt psychology of perception. Have the principles been used to the greatest advantage? The most important principle for our purposes is to make sure the visual has good figure/ground differences. Does the main point visually stand out? Then we can consider secondary effects such as asking if the grouping has been used to best effect. Has connectedness been used effectively? Has flow been used properly?

Design with Quality – *Does the visual inform the viewer and dispel his ignorance?*

A quality chart is one that is full of information; in other words, our visuals must be "alive." To be "alive," a visual must provide a service. The service is the resolution of a tension the viewer brings to the chart. They want to know. If the chart informs, the tension is relieved, and the viewer then "knows." If the viewer is still puzzled over the information after viewing, the chart does not have the living quality of informing and releasing the tension of ignorance.

Case Analysis

Review the case study and the chart to be improved. Use this exercise to practice analyzing a chart along the Perception dimension. Make sure to use the analysis questions from the template, repeated as follows. In the end, after you have gone through the three principles, you will be able to answer:

> *What is wrong with the visual being analyzed along these dimensions?*
>
> *What can be done to improve the visual along all these dimensions?*

1. Seeing – Does the eye of the viewer focus on the most important point being made?
2. Mind – Have the principles of the Gestalt psychology of perception been thoughtfully employed in the visual?
3. Quality – Does the visual inform the viewer and dispel his ignorance?

Analysis Case 15.5 – Method

Use Color Effectively – *Is color used judiciously and sparsely?*

Color should be used judiciously. It really enhances the chart. But if you use Excel, for example, the program chooses which colors to display based on some internal formula. That does not always lead to the best color combination. You must also use color sparsely. Most of your chart should be in black and white. Color should be left for those elements you want to use to draw the viewer's attention. For example, you could gray out the axis somewhat to make it fade into the background. Does your use of color grab the viewer's attention? Color that is used should also be semantically correct.

Remove All Chart Junk – *Is the visual is clear of unnecessary visual elements not leading to a clear point being* made?

Chart junk is anything in your visual that detracts from the viewer's comprehension of our charts. Just as we must declutter when we prepare our family home to be sold, you should clear up the viewing space in a visual as much as possible. Be ruthless. Use the redo and undo function in your chart maker. Put the feature in and then delete it. Go back and forth, putting it in and taking it out repeatedly to determine if valuable information would be lost if it was not there. Is there a simpler way to make that point? Remove unnecessary or confusing visual elements in charts and graphs. Any markings and visual elements can be called chart junk if they are not part of the minimum set of visuals necessary to communicate the information understandably.

Tell the Story with the Title – *Does the title of the chart convey the point being made with the chart?*

Consultants at McKinsey and Company, the world-famous consulting powerhouse, have a wonderful practice in titling their presentation slides. The title of the slide (or even the visual, but not both) makes the business point of the graph explicitly. It tells the viewer what the graph means right in the title. They don't let the viewers try to figure it out for themselves. They may draw an incorrect conclusion. You need to tell them, right there, at the top of your slide. Be succinct but express it as a complete thought; a phrase will suffice. Charts should also have direct labeling for series. Avoid using a legend that makes the viewer's eye go back and forth and get lost as they try to get the point. Tell them directly what each chart feature is.

Case Analysis

Review the case study and the chart to be improved. Use this exercise to practice analyzing a chart along the Methods dimension. Make sure to use the analysis questions from the template, repeated as follows. In the end, after you have gone through the three principles, you will be able to answer:

> *What is wrong with the visual being analyzed along these dimensions?*
>
> *What can be done to improve the visual along all these dimensions?*

1. Color – Is color used judiciously and sparsely?
2. Chart junk – Is the visual clear of unnecessary visual elements not leading to a clear point being made?
3. Title – Does the title of the chart convey the point being made with the chart?

Analysis Case 15.6 – Charts

Use the Right Chart – *Does the type of chart being used match the level of judgment required?*

Researchers have developed a useful scale for elementary perceptual tasks, which we apply to decide the type of graph that should be used for the level of accuracy desired. This functional scale tells us that the higher the encoding methods, the more accurate the comparison it facilitates. Linear charts such as bar charts and line charts and scatter plots afford the highest level of accurate comparison. Charts like pie charts are good for gross but not accurate comparisons. They afford other types of judgment such as contributions to the whole. The least accurate, but still useful, are color scales used in coloring a map with data as in GIS plots. This should guide us to match the type of graph to use for the question being answered. Using this scale, we can ask: (a) is the right chart being used for the intended purpose? and (b) does the visual use the right level of encoding for the level of accurate judgment desired?

Select the Chart Type Effectively – *Does the chart type used match the business question being answered?*

Consider answering a few basic questions in the use of the chart type that was selected: (a) are the charts answering the right business question? (b) does the chart match the business question being presented? and (c) have the four basic chart types been used properly (bar, line, pie, scatter)? Each chart type is best suited to presenting answers to certain types of questions. For example, a Pareto bar chart is excellent for showing the 80/20 percent contributions for certain contributors. Pie charts are best for contributions to a whole. Match the point you are trying to make to the type of chart that suits that type of information best.

Enhance Table Data for Emphasis – *Are referenceable visuals (tables) readable with appropriate conditional formatting and thumbnail graphs used for emphasis?*

When using tables in our visuals, we must ask: are we using the table to analyze or to tell? Is the use of referenceable or glanceable visuals appropriate? Also, we must ask if the referenceable visual has been designed to be legible and readable. Do the tables have enough white space? Has shading of the table been used appropriately and sparingly? Has the emphasis been appropriately added with conditional formatting? And have thumbnail graphs such as sparklines been employed to add insight?

Case Analysis

Review the case study and the chart to be improved. Use this exercise to practice analyzing a chart along the Charts dimension. Make sure to use the analysis questions from the template, repeated as follows. In the end, after you have gone through the three principles, you will be able to answer:

> *What is wrong with the visual being analyzed along these dimensions?*
>
> *What can be done to improve the visual along all these dimensions?*

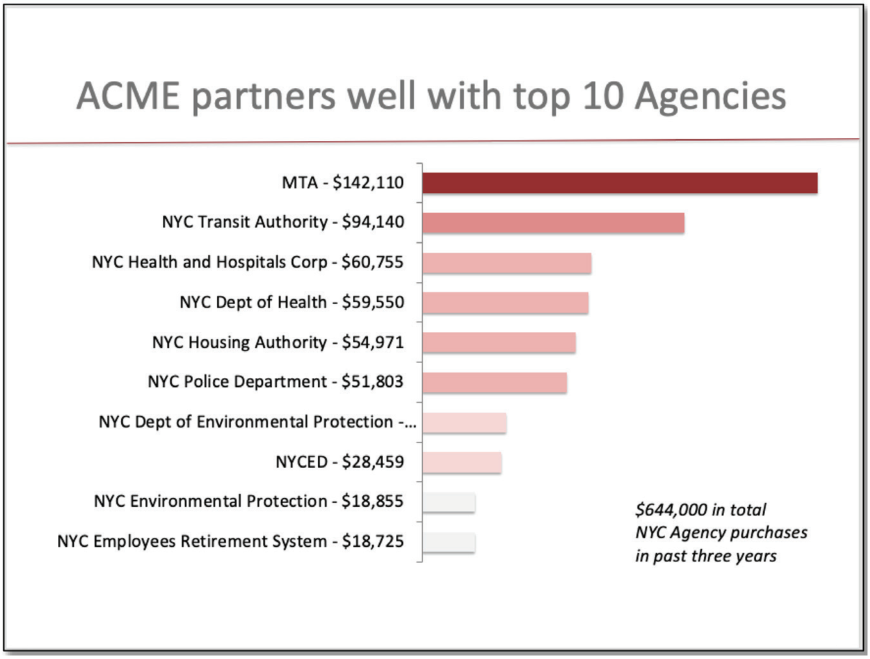
1. Right Chart – Does the type of chart being used match the level of judgment required?
2. Selection – Does the chart type used match the business question being answered?
3. Tables – Are referenceable visuals (tables) readable with appropriate conditional formatting and thumbnail graphs used for emphasis?

Analysis Case 15.7 – Putting It All Together

Create a new chart you will propose to your company to be used in place of the chart they were going to use. Do this without looking at the solution chart generated by an expert. Make sure your revised chart does the following:

1. It tells a good story.
2. It can be used as a prop.
3. It emulates traditional charts made by experts.
4. It makes a good sign.
5. It sends a strong signal.
6. It is functional, not artistic.
7. It fulfills the organizational need.
8. It allows for audience biases and journey.
9. It answers a well-framed analytical question.
10. It focuses the eye of the viewer on the most important elements.
11. It employs proper Gestalt perception principles.
12. It dispels viewer ignorance.
13. It uses color appropriately.
14. It is free of chart junk.
15. It has an appropriate title.
16. It supports the level of judgment needed.
17. It matches the business question being answered.
18. It is glanceable with appropriate conditional formatting and thumbnail graphs for emphasis, if there is a table.

A Possible Solution



ACME partners well with top 10 Agencies

$644,000 in total
NYC Agency purchases
in past three years

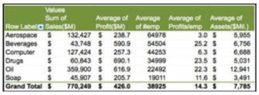
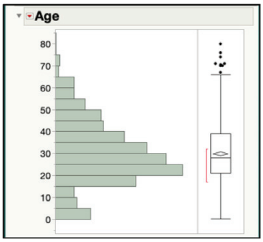
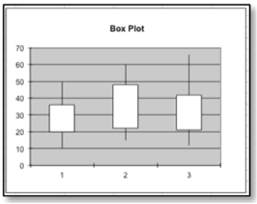
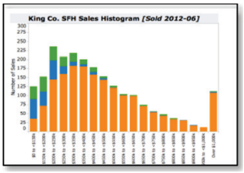
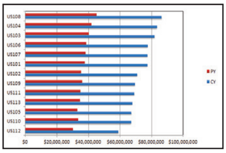
FIGURE 15.2 Case study slide modified and made more compelling.

CHAPTER 16

SUMMARY OF ANALYSIS TECHNIQUES

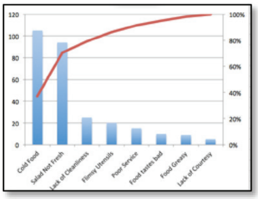
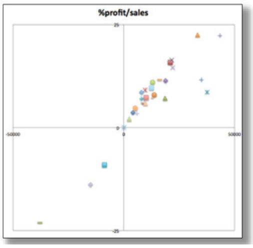
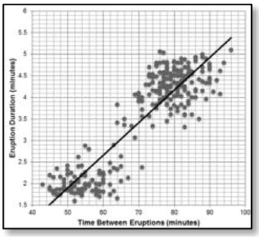
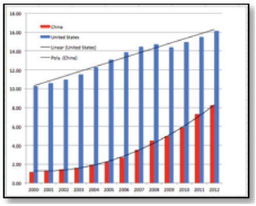
Business Questions and Analysis Techniques for Addressing Them

Basic Techniques

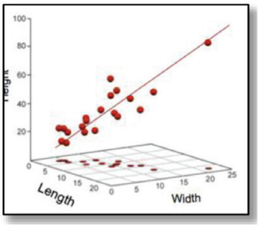
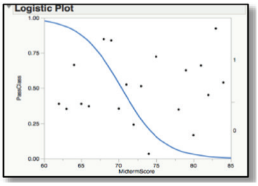
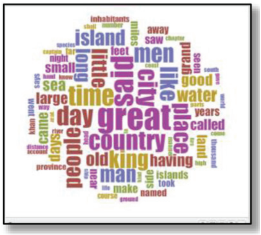
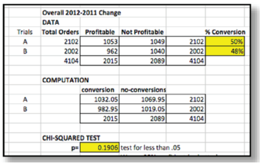
Type of Business Question	Types of Variables	Best Analysis Technique	Examples	Graphic
How many? **How much?**	Continuous, distributed over categories	Descriptive statistics, averages, time series	How much revenue did we receive last year by customer type and sales territory? What was the average sale per salesperson last month?	
Is the data widely dispersed? **Are there any outliers?**	Continuous	Quartiles, Z-score, 5-Point Summary	Are there any sales personnel who had outstanding sales last quarter? Did any employee score above all others in the leadership test?	
How is the data distributed?	Categorical and Continuous	Box Plots	Range of monthly sales by salesperson over CY compared to PY. Range of expenses by category over last quarter.	
How is the data distributed?	Continuous	Histogram	How are the salaries of our employees distributed? What is the distribution of housing prices in the county?	
Which are the significant categories?	Categorical, nominal	Bar Chart	Which products had the highest customer ratings? Rank all companies by their capitalization.	

Row Labels	Values Sum of Sales($M)	Average of Profit($M)	Average of #emp	Average of Profits/emp	Average of Assets($Mil.)
Aerospace	$ 132,427	$ 238.7	64978	3.0	$ 5,955
Beverages	$ 43,748	$ 590.9	54504	25.2	$ 6,756
Computer	$ 127,424	$ 257.3	44253	6.3	$ 6,688
Drugs	$ 60,843	$ 690.1	34999	23.5	$ 5,031
Oil	$ 359,900	$ 616.9	22492	22.3	$ 12,941
Soap	$ 45,907	$ 205.7	19011	11.6	$ 3,491
Grand Total	**$ 770,249**	**$ 426.0**	**38925**	**14.3**	**$ 7,785**

Intermediate Techniques

Type of Business Question	Types of Variables	Best Analysis Technique	Examples	Graphic
Who or which are the most important?	Continuous	Pareto	Which customers contributed to most of our revenue last year? Which are the most prevalent complaints?	
Which are the best choices?	Continuous 2 variables	2 X 2 Analysis	Which are the best vendor products for lowest cost and highest performance? Which employees have the best performance in sales and leadership?	
Are the factors related?	Continuous 2 variables compared	Scatter plot, correlation analysis	Can we use one factor as an indicator of what another variable will do?	
What are the trends? **Forecasts?**	Continuous, time series	Time series, regression, forecasts, line charts	Have our sales been increasing year after year? What will our sales be in the next three months?	

Advanced Techniques

Type of Business Question	Types of Variables	Best Analysis Technique	Examples	Graphic
Can we predict an outcome?	A model with many continuous input variables	Multivariate regression	Can we create a model to design a store given certain input parameters to achieve a certain level of sales?	
Can we predict a binary outcome?	A continuous variable predicting one of two categorical outcomes	Logistic regression	Given a customer response on a continuous variable question, can we tell whether the customer is male or female?	
What are people saying?	Unstructured text	Text analysis, word frequency analysis	Can we analyze 10,000 emails to find customer sentiment and tie it to dates?	
Are the factors related?	Categorical, two variables compared	Contingency analysis and Chi-Squared test	Can we use gender as an indicator of preference for our product for this group of customers?	

Overall 2012-2011 Change

DATA

Trials	Total Orders	Profitable	Not Profitable		% Conversion
A	2102	1053	1049	2102	50%
B	2002	962	1040	2002	48%
	4104	2015	2089	4104	

COMPUTATION

		conversion	no-conversions	
A		1032.05	1069.95	2102
B		982.95	1019.05	2002
		2015	2089	4104

CHI-SQUARED TEST

p= 0.1906 test for less than .05

Data Sources

1. **Data set *ORDERS.CSV***. This data set was made available courtesy of Tableau, Inc. as Open Source and is derived from their *Sample-Superstore* training dataset. The data set may be found at: *https://community.tableau.com/servlet/JiveServlet/downloadBody/1236-102-2-15278/Sample%20-%20Superstore.xls*.

2. **Data set *BankComplaints.csv***. This data set was made available courtesy of the U.S. Government Department of Consumer Affairs as part of their Consumer Complaint Database, a publicly available unrestricted data set. The data set can be found at: *https://catalog.data.gov/dataset/consumer-complaint-database*.

3. **Data set *SFOCustomerSurvey.csv***. This data set was made available courtesy of the San Francisco Airport as an Open Data set available at *https://data.sfgov.org/browse?q=sfo*.

4. **Data set *Football.csv***. This data set was made available by permission of Michael B. Lafferty, the author of the original news article: Lafferty, M. B., (1993, November 21), "OSU scientists get a kick out of sports controversy," *The Columbus Dispatch*, B7. The data set may be found at: *https://www3.nd.edu/~busiforc/handouts/Data%20and%20Stories/t%20test/Helium%20Footballs/Helium%20Football%20Data.html*.

5. **Data set *courses.csv.*** This data set was made available courtesy of the Harvard Dataverse Project, under DVN/26147_2014, HarvardX, publisher: Harvard Dataverse, title: HarvardX Person-Course Academic Year 2013 De-Identified dataset, version 3.0, UNF = {UNF:6:WSoYmsP5KeX2t/6g2JiEuw==, year: 2014,version: V11, doi: 10.7910/DVN/26147, URL: *https://doi.org/10.7910/DVN/26147*.

6. **Dataset *calcium.csv***. This data set was made available by permission of John P. Holcomb, Jr., PhD, Cleveland State University. The data set may be found at: *https://academic.csuohio.edu/holcombj/clean/cleaningassignment.htm*.

INDEX