CHAPTER 3

Visualizing and Exploring Data

Laborant/Shutterstock.com

Learning Objectives

After studying this chapter, you will be able to:

- Create Microsoft Excel charts.
- Determine the appropriate chart to visualize different types of data.
- Sort a data set in an Excel spreadsheet.
- Apply the Pareto Principle to analyze data.
- Use the Excel *Autofilter* to identify records in a database meeting certain characteristics.
- Explain the science of statistics and define the term *statistic*.
- Construct a frequency distribution for both discrete and continuous data.
- Construct a relative frequency distribution and histogram.
- Compute cumulative relative frequencies.
- Find percentiles and quartiles for a data set.
- Construct a cross-tabulation (contingency table).
- Use PivotTables to explore and summarize data.
- Use PivotTables to construct a cross-tabulation.
- Display the results of PivotTables using PivotCharts.

Converting data into information to understand past and current performance is the core of descriptive analytics and is vital to making good business decisions. Techniques for doing this range from plotting data on charts, extracting data from databases, and manipulating and summarizing data. In this chapter, we introduce a variety of useful techniques for descriptive analytics.

Data Visualization

The old adage "A picture is worth 1000 words" is probably truer in today's information-rich environment than ever before. In Chapter 1 we stated that data visualization is at the core of modern business analytics. **Data visualization** is the process of displaying data (often in large quantities) in a meaningful fashion to provide insights that will support better decisions. Making sense of large quantities of disparate data is necessary not only for gaining competitive advantage in today's business environment but also for surviving in it. Researchers have observed that data visualization improves decision-making, provides managers with better analysis capabilities that reduce reliance on IT professionals, and improves collaboration and information sharing.

Raw data are important, particularly when one needs to identify accurate values or compare individual numbers. However, it is quite difficult to identify trends and patterns, find exceptions, or compare groups of data in tabular form. The human brain does a surprisingly good job processing visual information—if presented in an effective way. Visualizing data provides a way of communicating data at all levels of a business and can reveal surprising patterns and relationships. For many unique and intriguing examples of data visualization, visit the Data Visualization Gallery at the U.S. Census Bureau Web site, www.census.gov/dataviz/.

EXAMPLE 3.1 Tabular versus Visual Data Analysis

Figure 3.1 shows the data in the Excel file *Monthly Product Sales*. We can use the data to determine exactly how many units of a certain product were sold in a particular month, or to compare one month to another. For example, we see that sales of product A dropped in February, specifically by 6.7% (computed by the Excel formula = 1 – B3/B2). Beyond such calculations, however, it is difficult to draw big picture conclusions.

Figure 3.2 displays a chart of monthly sales for each product. We can easily compare overall sales of different products (Product C sells the least, for example), and identify trends (sales of Product D are increasing), other patterns (sales of Product C is relatively stable while sales of Product B fluctuates more over time), and exceptions (Product E's sales fell considerably in September).

Data visualization is also important both for building decision models and for interpreting their results. For example, recall the demand-prediction models in Chapter 1 (Examples 1.9 and 1.10). To identify the appropriate model to use, we would normally have to collect and analyze data on sales demand and prices to determine the type of relationship (linear or nonlinear, for example) and estimate the values of the parameters in the model. Visualizing the data will help to identify the proper relationship and use the appropriate data analysis tool. Furthermore, complex analytical models often yield complex results. Visualizing the results often helps in understanding and gaining insight about model output and solutions.

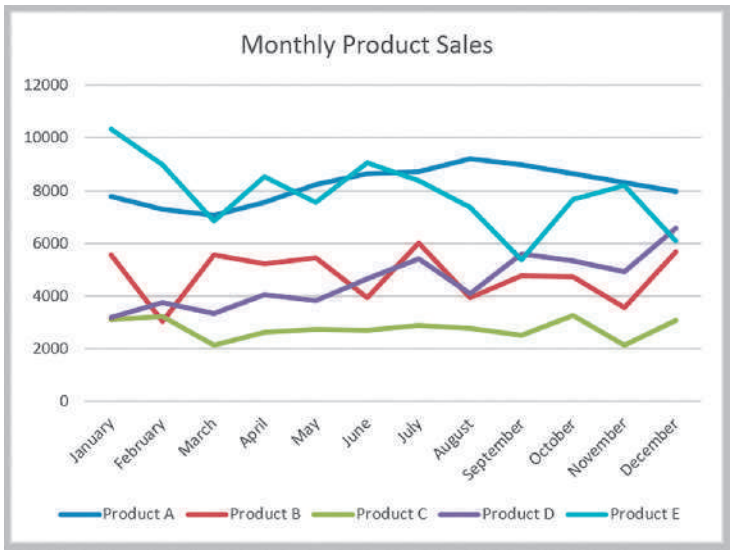
Figure 3.1

Monthly Product Sales Data

	A	B	C	D	E	F
1	Month	Product A	Product B	Product C	Product D	Product E
2	January	7792	5554	3105	3168	10350
3	February	7268	3024	3228	3751	8965
4	March	7049	5543	2147	3319	6827
5	April	7560	5232	2636	4057	8544
6	May	8233	5450	2726	3837	7535
7	June	8629	3943	2705	4664	9070
8	July	8702	5991	2891	5418	8389
9	August	9215	3920	2782	4085	7367
10	September	8986	4753	2524	5575	5377
11	October	8654	4746	3258	5333	7645
12	November	8315	3566	2144	4924	8173
13	December	7978	5670	3071	6563	6088

Figure 3.2

Visualization of *Monthly Product Sales* Data



Dashboards

Making data visible and accessible to employees at all levels is a hallmark of effective modern organizations. A **dashboard** is a visual representation of a set of key business measures. It is derived from the analogy of an automobile's control panel, which displays speed, gasoline level, temperature, and so on. Dashboards provide important summaries of key business information to help manage a business process or function. Dashboards might include tabular as well as visual data to allow managers to quickly locate key data. Figure 3.3 shows a simple dashboard for the product sales data in Figure 3.1 showing monthly sales for each product individually, sales of all products combined, total annual sales by product, a comparison of the last two months, and monthly percent changes by product.

Tools and Software for Data Visualization

Data visualization ranges from simple Excel charts to more advanced interactive tools and software that allow users to easily view and manipulate data with a few clicks, not only on computers, but on iPads and other devices as well. In this chapter we discuss basic tools available in Excel. In Chapter 10, we will see several other tools used in data mining applications that are available with the Excel add-in, *XLMiner*, that is used in this book.

Figure 3.3
Dashboard for Product Sales

While we will only focus on Excel-based tools in this book, you should be aware of other options and commercial packages that are available. In particular, we suggest that you look at the capabilities of Tableau (www.tableausoftware.com) and IBM's Cognos software (www.cognos10.com). Tableau is easy to use and offers a free trial.

Creating Charts in Microsoft Excel

Microsoft Excel provides a comprehensive charting capability with many features. With a little experimentation, you can create very professional charts for business analyses and presentations. These include vertical and horizontal bar charts, line charts, pie charts, area charts, scatter plots, and many other special types of charts. We generally do not guide you through every application but do provide some guidance for new procedures as appropriate.

Certain charts work better for certain types of data, and using the wrong chart can make it difficult for the user to interpret and understand. While Excel offers many ways to make charts unique and fancy, naive users often focus more on the attention-grabbing aspects of charts rather than their effectiveness of displaying information. So we recommend that you keep charts simple, and avoid such bells and whistles as 3-D bars, cylinders, cones, and so on. We highly recommend books written by Stephen Few, such as *Show Me the Numbers* (Oakland, CA: Analytics Press, 2004) for additional guidance in developing effective data visualizations.

To create a chart in Excel, it is best to first highlight the range of the data you wish to chart. The Excel Help files provide guidance on formatting your data for a particular type of chart. Click the *Insert* tab in the Excel ribbon (Figure 3.4). From the *Charts* group, click the chart type, and then click a chart subtype that you want to use. Once a basic chart is created, you may use the options in the *Design* and *Format* tabs within the *Chart Tools* tabs to customize your chart (Figure 3.5). In the *Design* tab, you can change the type of chart, data included in the chart, chart layout, and styles. The *Format* tab provides various formatting options. You may also customize charts easily by right-clicking on elements of the chart or by using the *Quick Layout* options in the *Chart Layout* group within the *Chart Tools Design* tab.

You should realize that up to 10% of the male population are affected by color blindness, making it difficult to distinguish between different color variations. Although we generally display charts using Excel's default colors, which often, unfortunately, use red, experts suggest using blue-orange palettes. We suggest that you be aware of this for professional and commercial applications.

Figure 3.4

Excel *Insert* Tab

Figure 3.5

Excel *Chart Tools*

Column and Bar Charts

Excel distinguishes between vertical and horizontal bar charts, calling the former **column charts** and the latter **bar charts**. A *clustered column chart* compares values across categories using vertical rectangles; a *stacked column chart* displays the contribution of each value to the total by stacking the rectangles; and a *100% stacked column chart* compares the percentage that each value contributes to a total. Column and bar charts are useful for comparing categorical or ordinal data, for illustrating differences between sets of values, and for showing proportions or percentages of a whole.

EXAMPLE 3.2 Creating Column Charts

The Excel file *EEO Employment Report* provides data on the number of employees in different categories broken down by racial/ethnic group and gender (Figure 3.6). We will construct a simple column chart for the various employment categories for all employees. First, highlight the range C3:K6, which includes the headings and data for each category. Click on the *Column Chart* button and then on the first chart type in the list (a clustered column chart). To add a title, click on the *Add Chart Elements* button in the *Design* tab ribbon. Click on "Chart Title" in the chart and change it to "EEO Employment Report—Alabama." The names of the data series can be changed by clicking on the *Select Data* button in the *Data* group of the *Design* tab. In the *Select Data Source* dialog (see Figure 3.7), click on "Series1" and then the *Edit* button. Enter the name of the data series, in this case "All Employees." Change the names of the other data series to "Men" and "Women" in a similar fashion. You can also change the order in which the data series are displayed on the chart using the up and down buttons. The final chart is shown in Figure 3.8.

Be cautious when changing the scale of the numerical axis. The heights or lengths of the bars only accurately reflect the data values if the axis starts at zero. If not, the relative sizes can paint a misleading picture of the relative values of the data.

	A	B	C	D	E	F	G	H	I	J	K
1	Equal Employment Opportunity Commission Report - Number Employed in State of Alabama, 2006										
2											
3	Racial/Ethnic Group and Gender	Total Employment	Officials &	Professionals	Technicians	Sales Workers	Office & Clerical	Craft Workers	Operatives	Laborers	Service Workers
4	ALL EMPLOYEES	632,329	60,258	80,733	39,868	62,019	67,014	61,322	120,810	68,752	71,553
5	Men	349,353	41,777	39,792	19,848	23,727	11,293	55,853	84,724	44,736	27,603
6	Women	282,976	18,481	40,941	20,020	38,292	55,721	5,469	36,086	24,016	43,950
7											
8	WHITE	407,545	51,252	67,622	28,830	41,091	44,565	45,742	67,555	26,712	34,176
9	Men	237,516	36,536	34,842	16,004	17,756	7,656	42,699	50,537	17,802	13,684
10	Women	170,029	14,716	32,780	12,826	23,335	36,909	3,043	17,018	8,910	20,492
11											
12	MINORITY	224,784	9,006	13,111	11,038	20,928	22,449	15,580	53,255	42,040	37,377
13	Men	111,837	5,241	4,950	3,844	5,971	3,637	13,154	34,187	26,934	13,919
14	Women	112,947	3,765	8,161	7,194	14,957	18,812	2,426	19,068	15,106	23,458

Figure 3.6

Portion of *EEO Employment Report* Data

Figure 3.7

Select Data Source Dialog

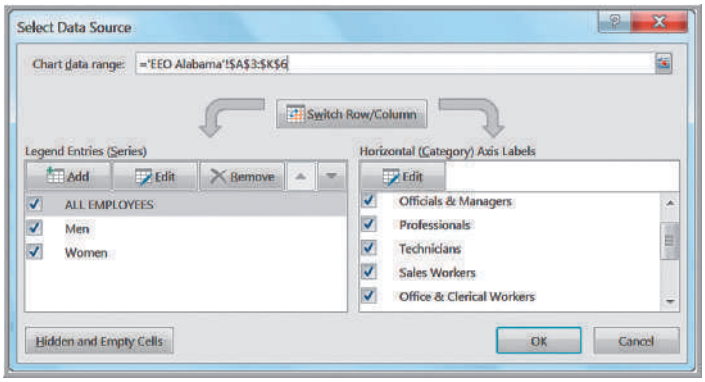
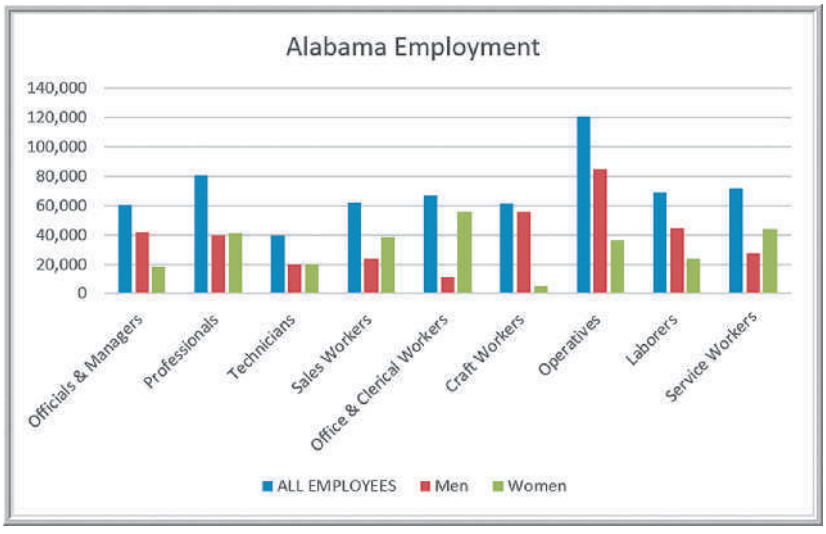
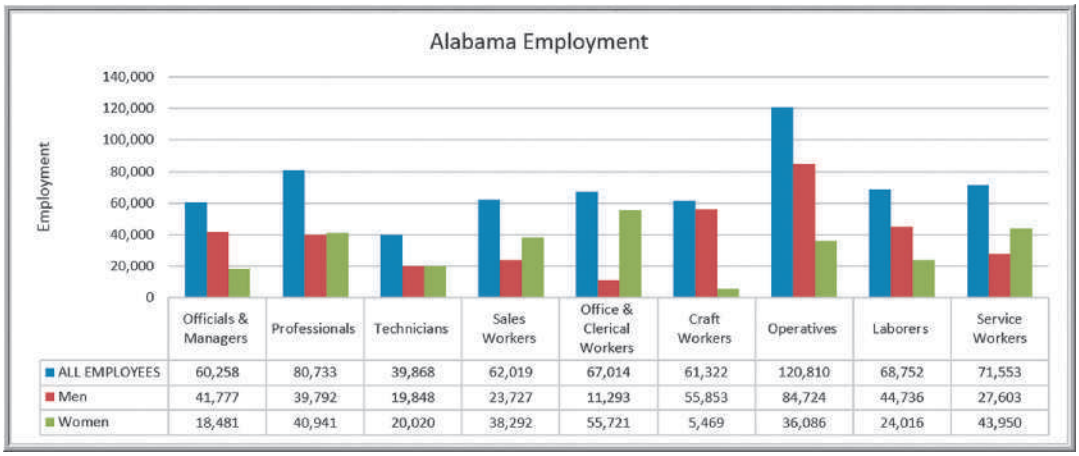


Figure 3.8

Column Chart for Alabama Employment Data





	Officials & Managers	Professionals	Technicians	Sales Workers	Office & Clerical Workers	Craft Workers	Operatives	Laborers	Service Workers
ALL EMPLOYEES	60,258	80,733	39,868	62,019	67,014	61,322	120,810	68,752	71,553
Men	41,777	39,792	19,848	23,727	11,293	55,853	84,724	44,736	27,603
Women	18,481	40,941	20,020	38,292	55,721	5,469	36,086	24,016	43,950

Figure 3.9

Alternate Column Chart Format

Data Labels and Data Tables Chart Options

Excel provides options for including the numerical data on which charts are based within the charts. Data labels can be added to chart elements to show the actual value of bars, for example. Data tables can also be added; these are usually better than data labels, which can get quite messy. Both can be added from the *Add Chart Element* Button in the *Chart Tools Design* tab, or also from the *Quick Layout* button, which provides standard design options. Figure 3.9 shows a data table added to the Alabama Employment chart. You can see that the data table provides useful additional information to improve the visualization.

Line Charts

Line charts provide a useful means for displaying data over time, as Example 3.3 illustrates. You may plot multiple data series in line charts; however, they can be difficult to interpret if the magnitude of the data values differs greatly. In that case, it would be advisable to create separate charts for each data series.

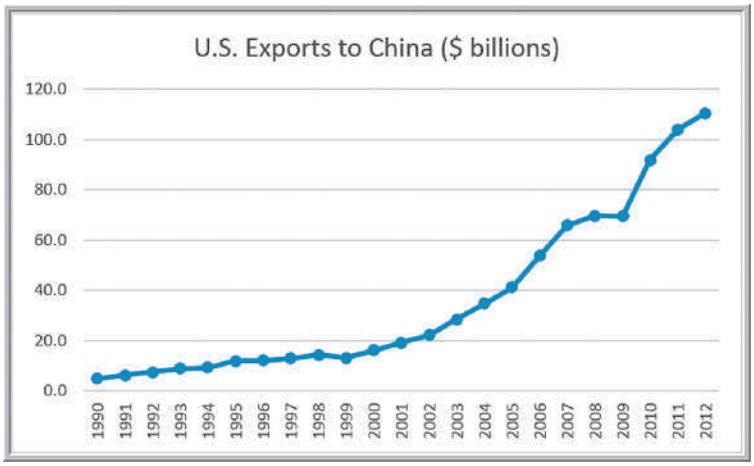
EXAMPLE 3.3 A Line Chart for China Export Data

Figure 3.10 shows a line chart giving the amount of U.S. exports to China in billions of dollars from the Excel file *China Trade Data.* The chart clearly shows a significant rise in exports starting in the year 2000, which began to level off around 2008.

Pie Charts

For many types of data, we are interested in understanding the relative proportion of each data source to the total. A **pie chart** displays this by partitioning a circle into pie-shaped areas showing the relative proportion. Example 3.4 provides one application.

Figure 3.10
Chart with Data Labels and Data Table



EXAMPLE 3.4 A Pie Chart for Census Data

Consider the marital status of individuals in the U.S. population in the Excel file *Census Education Data*, a portion of which is shown in Figure 3.11. To show the relative proportion in each category, we can use a pie chart, as shown in Figure 3.12. This chart uses a layout option that shows the labels associated with the data as well as the actual proportions as percentages. A different layout that shows both the values and/or proportions can also be chosen.

Data visualization professionals don't recommend using pie charts. For example, contrast the pie chart in Figure 3.12 with the column chart in Figure 3.13 for the same data. In the pie chart, it is difficult to compare the relative sizes of areas; however, the bars in the column chart can easily be compared to determine relative ratios of the data. If you do use pie charts, restrict them to small numbers of categories, always ensure that the numbers add to 100%, and use labels to display the group names and actual percentages. Avoid three-dimensional (3-D) pie charts—especially those that are rotated—and keep them simple.

Area Charts

An **area chart** combines the features of a pie chart with those of line charts. Area charts present more information than pie or line charts alone but may clutter the observer's mind with too many details if too many data series are used; thus, they should be used with care.

EXAMPLE 3.5 An Area Chart for Energy Consumption

Figure 3.14 displays total energy consumption (billion Btu) and consumption of fossil fuels from the Excel file *Energy Production & Consumption.* This chart shows that although total energy consumption has grown since 1949, the relative proportion of fossil fuel consumption has remained generally consistent at about half of the total, indicating that alternative energy sources have not replaced a significant portion of fossil-fuel consumption.

Scatter Chart

Scatter charts show the relationship between two variables. To construct a scatter chart, we need observations that consist of pairs of variables. For example, students in a class might have grades for both a midterm and a final exam. A scatter chart would show whether high or low grades on the midterm correspond strongly to high or low grades on the final exam or whether the relationship is weak or nonexistent.

Figure 3.11

Portion of *Census Education Data*

	A	B	C	D	E	F	G
1	Census Education Data						
2		Not a High School Grad	High School Graduate	Some College No Degree	Associate's Degree	Bachelor's Degree	Advanced Degree
18	Marital Status						
19	Never Married	4,120,320	7,777,104	4,789,872	1,828,392	5,124,648	2,137,416
20	Married, spouse present	15,516,160	36,382,720	18,084,352	8,346,624	19,154,432	9,523,712
21	Married, spouse absent	1,847,880	2,368,024	1,184,012	465,392	670,712	301,136
22	Separated	1,188,090	1,667,010	842,715	336,165	405,240	165,780
23	Widowed	5,145,683	4,670,488	1,765,010	556,657	977,544	475,195
24	Divorced	2,968,680	7,003,040	3,806,000	1,674,640	2,340,690	1,217,920

Figure 3.12

Pie Chart for Marital Status

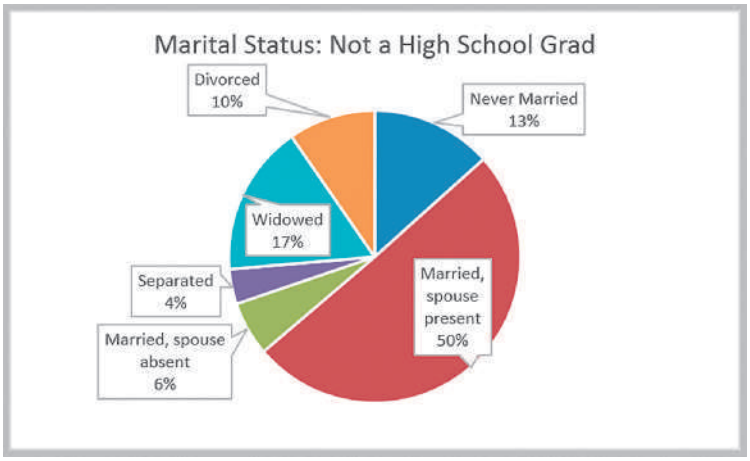


Figure 3.13

Alternative Column Chart for Marital Status: Not a High School Grad

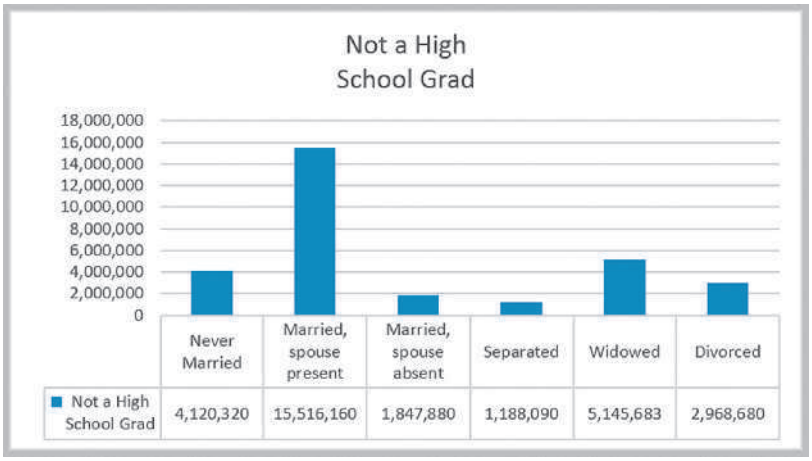
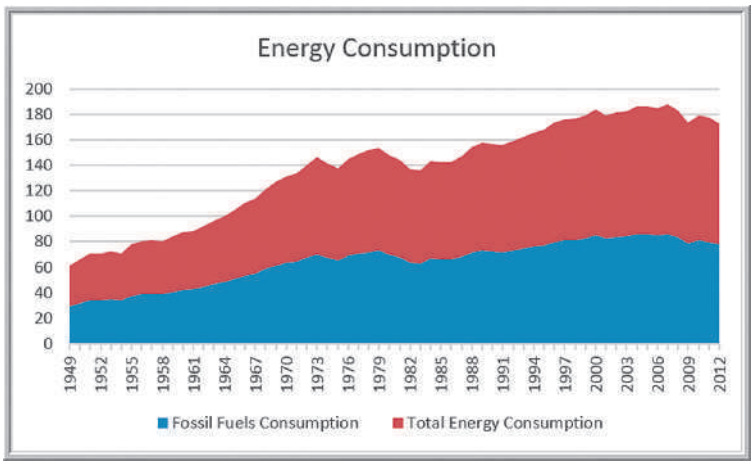


Figure 3.14

Area Chart for Energy Consumption



EXAMPLE 3.6 A Scatter Chart for Real Estate Data

Figure 3.15 shows a scatter chart of house size (in square feet) versus the home market value from the Excel file *Home Market Value.* The data clearly suggest that higher market values are associated with larger homes.

Bubble Charts

A **bubble chart** is a type of scatter chart in which the size of the data marker corresponds to the value of a third variable; consequently, it is a way to plot three variables in two dimensions.

EXAMPLE 3.7 A Bubble Chart for Comparing Stock Characteristics

Figure 3.16 shows a bubble chart for displaying price, P/E (price/earnings) ratio, and market capitalization for five different stocks on one particular day in the Excel file *Stock Comparisons*. The position on the chart shows the price and P/E; the size of the bubble represents the market cap in billions of dollars.

Figure 3.15

Scatter Chart of House Size versus Market Value

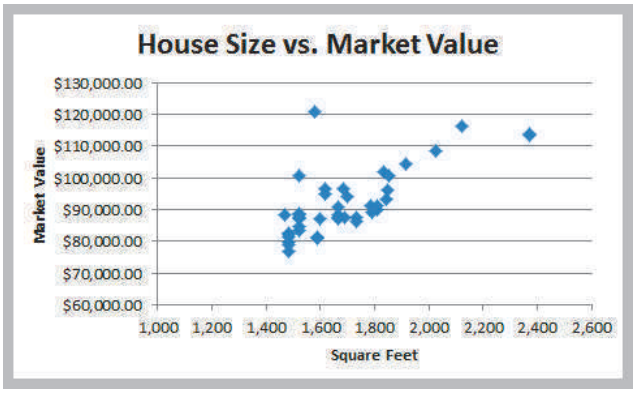
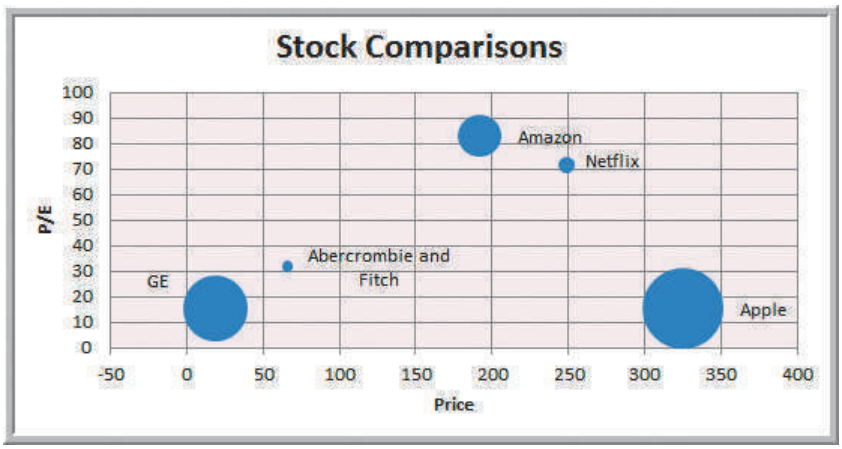


Figure 3.16

Bubble Chart for *Stock Comparisons*



Miscellaneous Excel Charts

Excel provides several additional charts for special applications. These additional types of charts (including bubble charts) can be selected and created from the *Other Charts* button in the Excel ribbon. These include the following:

- A **stock chart** allows you to plot stock prices, such as the daily high, low, and close. It may also be used for scientific data such as temperature changes.
- A **surface chart** shows 3-D data.
- A **doughnut chart** is similar to a pie chart but can contain more than one data series.
- A **radar chart** allows you to plot multiple dimensions of several data series.

Geographic Data

Many applications of business analytics involve geographic data. For example, problems such as finding the best location for production and distribution facilities, analyzing regional sales performance, transporting raw materials and finished goods, and routing vehicles such as delivery trucks involve geographic data. In such problems, data mapping can help in a variety of ways. Visualizing geographic data can highlight key data relationships, identify trends, and uncover business opportunities. In addition, it can often help to spot data errors and help end users understand solutions, thus increasing the likelihood of acceptance of decision models. Companies like Nike use geographic data and information systems for visualizing where products are being distributed and how that relates to demographic and sales information. This information is vital to marketing strategies. The use of prescriptive analytic models in combination with data mapping was instrumental in the success of Procter & Gamble Company's North American Supply Chain study, which saved the company in excess of $200 million dollars per year.[1] We discuss this application in Chapter 15.

[1]J. Camm et al., "Blending OR/MS, Judgment and GIS: Restructuring P&G's Supply Chain," *Interfaces*, 27, 1 (1997): 128–142.

Geographic mapping capabilities were introduced in Excel 2000 but were not available in Excel 2002 and later versions. These capabilities are now available through Microsoft MapPoint 2010, which must be purchased separately. MapPoint is a geographic data-mapping tool that allows you to visualize data imported from Excel and other database sources and integrate them into other Microsoft Office applications. For further information, see http://www.microsoft.com/mappoint/en-us/home.aspx.

Other Excel Data Visualization Tools

Microsoft Excel offers numerous other tools to help visualize data. These include data bars, color scales, and icon sets; sparklines, and the camera tool. We will describe each of these in the following sections.

Data Bars, Color Scales, and Icon Sets

These options are part of Excel's *Conditional Formatting* rules, which allow you to visualize different numerical values through the use of colors and symbols. Excel has a variety of standard templates to use, but you may also customize the rules to meet your own conditions and styles. We encourage you to experiment with these tools.

EXAMPLE 3.8 Data Visualization through Conditional Formatting

Data bars display colored bars that are scaled to the magnitude of the data values (similar to a bar chart) but placed directly within the cells of a range. Figure 3.17 shows data bars applied to the data in the *Monthly Product Sales* worksheet. Highlight the data in each column, click the *Conditional Formatting* button in the *Styles* group within the *Home* tab, select *Data Bars*, and choose the fill option and color.

Color scales shade cells based on their numerical value using a color palette. This is another option in the *Conditional Formatting* menu. For example, in Figure 3.18 we use a green-yellow-red color scale, which highlights cells containing large values in green, small values in red, and middle values in yellow. The darker the green, the larger the value; the darker the red, the smaller the value. For intermediate values, you can see that the colors blend together. This provides a quick way of identifying the largest and smallest product-month sales values. Color-coding of quantitative data is commonly called a **heatmap**. We will see another application of a heatmap in Chapter 14.

Finally, Icon Sets provide similar information using various symbols such as arrows or stoplight colors. Figure 3.19 shows an example.

Figure 3.17

Example of Data Bars

	A	B	C	D	E	F
1	Month	Product A	Product B	Product C	Product D	Product E
2	January	7792	5554	3105	3168	10350
3	February	7268	3024	3228	3751	8965
4	March	7049	5543	2147	3319	6827
5	April	7560	5232	2636	4057	8544
6	May	8233	5450	2726	3837	7535
7	June	8629	3943	2705	4664	9070
8	July	8702	5991	2891	5418	8389
9	August	9215	3920	2782	4085	7367
10	September	8986	4753	2524	5575	5377
11	October	8654	4746	3258	5333	7645
12	November	8315	3566	2144	4924	8173
13	December	7978	5670	3071	6563	6088

Figure 3.18

Example of Color Scales

	A	B	C	D	E	F
1	Month	Product A	Product B	Product C	Product D	Product E
2	January	7792	5554	3105	3168	10350
3	February	7268	3024	3228	3751	8965
4	March	7049	5543	2147	3319	6827
5	April	7560	5232	2636	4057	8544
6	May	8233	5450	2726	3837	7535
7	June	8629	3943	2705	4664	9070
8	July	8702	5991	2891	5418	8389
9	August	9215	3920	2782	4085	7367
10	September	8986	4753	2524	5575	5377
11	October	8654	4746	3258	5333	7645
12	November	8315	3566	2144	4924	8173
13	December	7978	5670	3071	6563	6088

Figure 3.19

Example of Icon Sets

	A	B	C	D	E	F
1	Month	Product A	Product B	Product C	Product D	Product E
2	January	7792	5554	3105	3168	10350
3	February	7268	3024	3228	3751	8965
4	March	7049	5543	2147	3319	6827
5	April	7560	5232	2636	4057	8544
6	May	8233	5450	2726	3837	7535
7	June	8629	3943	2705	4664	9070
8	July	8702	5991	2891	5418	8389
9	August	9215	3920	2782	4085	7367
10	September	8986	4753	2524	5575	5377
11	October	8654	4746	3258	5333	7645
12	November	8315	3566	2144	4924	8173
13	December	7978	5670	3071	6563	6088

Sparklines

Sparklines are graphics that summarize a row or column of data in a single cell. Sparklines were introduced by Edward Tufte, a famous expert on visual presentation of data. He described sparklines as "data-intense, design-simple, word-sized graphics." Excel has three types of sparklines: line, column, and win/loss. Line sparklines are clearly useful for time-series data, while column sparklines are more appropriate for categorical data. Win-loss sparklines are useful for data that move up or down over time. They are found in the *Sparklines* group within the Insert menu on the ribbon.

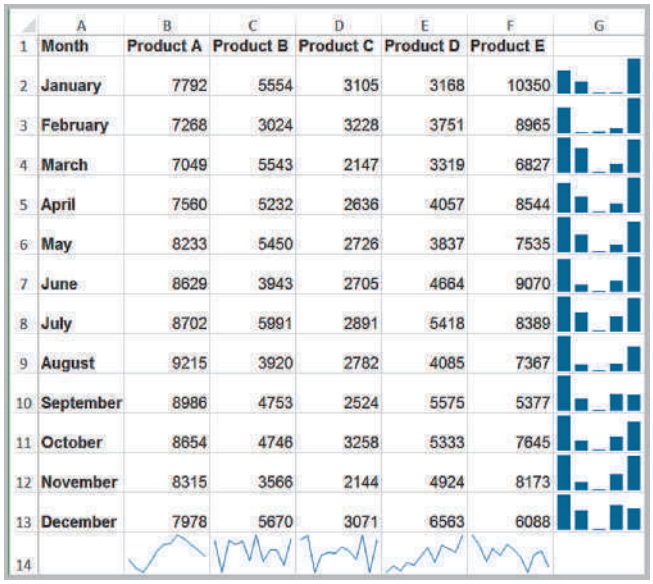
EXAMPLE 3.9 Examples of Sparklines

We will again use the *Monthly Product Sales* data. Figure 3.20 shows line sparklines in row 14 for each product. In column G, we display column sparklines, which are essentially small column charts. Generally you need to expand the row or column widths to display them effectively. Notice, however, that the lengths of the bars are not scaled properly to the data; for example, in the first one, products D and E are roughly one-third the value of Product E yet the bars are not scaled correctly. So be careful when using them.

Figure 3.21 shows a modified worksheet in which we computed the percentage change from 1 month to the next for products A and B. The win-loss sparklines in row 14 show the patterns of sales increases and decreases, suggesting that product A has a cyclical pattern while product B changed in a more random fashion. If you click on any cell containing a sparkline, the *Sparkline Tools Design* tab appears, allowing you to customize colors and other options.

Figure 3.20

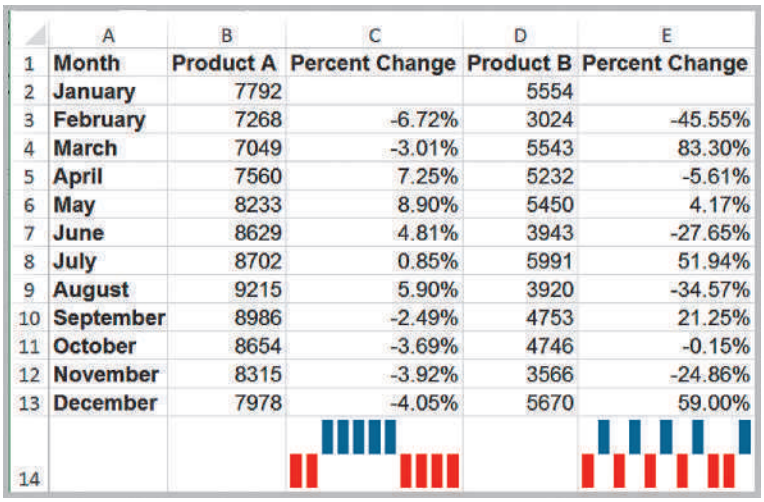
Line and Column Sparklines



	A	B	C	D	E	F	G
1	Month	Product A	Product B	Product C	Product D	Product E	
2	January	7792	5554	3105	3168	10350	
3	February	7268	3024	3228	3751	8965	
4	March	7049	5543	2147	3319	6827	
5	April	7560	5232	2636	4057	8544	
6	May	8233	5450	2726	3837	7535	
7	June	8629	3943	2705	4664	9070	
8	July	8702	5991	2891	5418	8389	
9	August	9215	3920	2782	4085	7367	
10	September	8986	4753	2524	5575	5377	
11	October	8654	4746	3258	5333	7645	
12	November	8315	3566	2144	4924	8173	
13	December	7978	5670	3071	6563	6088	
14							

Figure 3.21

Win-Loss Sparklines



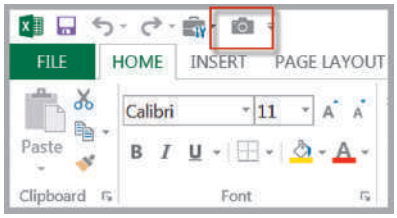
	A	B	C	D	E
1	Month	Product A	Percent Change	Product B	Percent Change
2	January	7792		5554	
3	February	7268	-6.72%	3024	-45.55%
4	March	7049	-3.01%	5543	83.30%
5	April	7560	7.25%	5232	-5.61%
6	May	8233	8.90%	5450	4.17%
7	June	8629	4.81%	3943	-27.65%
8	July	8702	0.85%	5991	51.94%
9	August	9215	5.90%	3920	-34.57%
10	September	8986	-2.49%	4753	21.25%
11	October	8654	-3.69%	4746	-0.15%
12	November	8315	-3.92%	3566	-24.86%
13	December	7978	-4.05%	5670	59.00%
14					

Excel Camera Tool

A little-known feature of Excel is the camera tool. This allows you to create live pictures of various ranges from different worksheets that you can place on a single page, size them, and arrange them easily. They are simply linked pictures of the original ranges, and the advantage is that as any data are changed or updated, the camera shots are also. This is particularly valuable for printing summaries when you need to extract data from multiple worksheets, consolidating PivotTables (introduced later in this chapter) onto one page, or for creating dashboards when the tables and charts are scattered across multiple worksheets. To use the camera too, first add it to the *Quick Access Toolbar* (the set of buttons above the ribbon). From the *File* menu, choose *Options* and then *Quick Access Toolbar*. Choose *Commands*, and then *Commands Not in the Ribbon*. Select *Camera* and add it. It will then appear as shown in Figure 3.22. To use it, simply highlight a range of cells

Figure 3.22

Excel Camera Tool Button



(if you want to capture a chart, highlight a range of cells surrounding it), click the camera tool button and then click the location where you want to place the picture. You may size the picture just like any other Microsoft Excel object. We will illustrate this tool later in the chapter when we discuss PivotTables.

Data Queries: Tables, Sorting, and Filtering

Managers make numerous queries about data. For example, in the *Purchase Orders* database (Figure 1.3), they might be interested in finding all orders from a certain supplier, all orders for a particular item, or tracing orders by order data. To address these queries, we need to sort the data in some way. In other cases, managers might be interested in extracting a set of records having certain characteristics. This is termed *filtering* the data. For example, in the *Purchase Orders* database, a manager might be interested in extracting all records corresponding to a certain item.

Excel provides a convenient way of formatting databases to facilitate analysis, called *Tables*.

EXAMPLE 3.10 Creating an Excel Table

We will use the *Credit Risk Data* file to illustrate an Excel table. First, select the range of the data, including headers (a useful shortcut is to select the first cell in the upper left corner, then click *Ctrl+Shift+down arrow*, and then *Ctrl+Shift+right arrow*). Next, click *Table* from the *Tables* group on the *Insert* tab and make sure that the box for *My Table Has Headers* is checked. (You may also just select a cell within the table and then click on *Table* from the *Insert* menu. Excel will choose the table range for you to verify.) The table range will now be formatted and will continue automatically when new data are entered. Figure 3.23 shows a portion of the result. Note that the rows are shaded and that each column header has a drop-down arrow to filter the data (we'll discuss this shortly). If you click within a table, the *Table Tools Design* tab will appear in the ribbon, allowing you to do a variety of things, such as change the color scheme, remove duplicates, change the formatting, and so on.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Credit Risk Data											
2												
3	Loan Purpo	Checkir	Savin	Months Customer	Months Employ	Gend	Marital Stat	Age	Hous	Years	J	Credit Ri
4	Small Appliance	$0	$739	13	12	M	Single	23	Own	3	Unskilled	Low
5	Furniture	$0	$1,230	25	0	M	Divorced	32	Own	1	Skilled	High
6	New Car	$0	$389	19	119	M	Single	38	Own	4	Management	High
7	Furniture	$638	$347	13	14	M	Single	36	Own	2	Unskilled	High
8	Education	$963	$4,754	40	45	M	Single	31	Rent	3	Skilled	Low
9	Furniture	$2,827	$0	11	13	M	Married	25	Own	1	Skilled	Low
10	New Car	$0	$229	13	16	M	Married	26	Own	3	Unskilled	Low
11	Business	$0	$533	14	2	M	Single	27	Own	1	Unskilled	Low
12	Small Appliance	$6,509	$493	37	9	M	Single	25	Own	2	Skilled	High
13	Small Appliance	$966	$0	25	4	F	Divorced	43	Own	1	Skilled	High
14	Business	$0	$989	49	0	M	Single	32	Rent	2	Management	High

Figure 3.23

Portion of *Credit Risk Data* Formatted as an Excel Table

An Excel table allows you to use table references to perform basic calculations, as the next example illustrates.

EXAMPLE 3.11 Table-Based Calculations

Suppose that in the *Credit Risk Data* table, we wish to calculate the total amount of savings in column C. We could, of course, simply use the function SUM(C4:C428). However, with a table, we could use the formula =SUM(Table1[Savings]). The table name, Table1, can be found (and changed) in the *Properties* group of the *Table Tools Design* tab. Note that Savings is the name of the header in column C. One of the advantages of doing this is that if we add new records to the table, the calculation will be updated automatically, and we don't have to change the range in the formula or get a wrong result if we forget to. As another example, we could find the number of home owners using the function =COUNTIF(Table1[Housing], "Own").

If you add additional records at the end of the table, they will automatically be included and formatted, and if you create a chart based on the data, the chart will automatically be updated if you add new records.

Sorting Data in Excel

Excel provides many ways to sort lists by rows or column or in ascending or descending order and using custom sorting schemes. The sort buttons in Excel can be found under the *Data* tab in the *Sort & Filter* group (see Figure 3.24). Select a single cell in the column you want to sort on and click the "AZ down arrow" button to sort from smallest to largest or the "AZ up arrow" button to sort from largest to smallest. You may also click the *Sort* button to specify criteria for more advanced sorting capabilities.

EXAMPLE 3.12 Sorting Data in the *Purchase Orders* Database

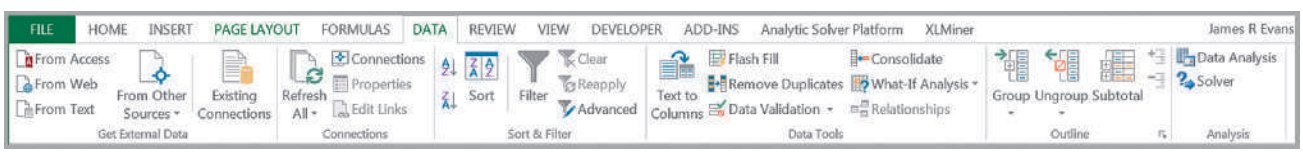
In Chapter 1 (Figure 1.3), we introduced a data set for purchase orders for an aircraft-component manufacturer. Suppose we wish to sort the data by supplier. Click on any cell in column A of the data (but not the header cell A3) and then the "AZ down" button in the *Data* tab. Excel will select the entire range of the data and sort by name of supplier in column A, a portion of which is shown in Figure 3.25. This allows you to easily identify the records that correspond to all orders from a particular supplier.

Pareto Analysis

Pareto analysis is a term named after an Italian economist, Vilfredo Pareto, who, in 1906, observed that a large proportion of the wealth in Italy was owned by a relatively small proportion of the people. The Pareto principle is often seen in many business situations. For example, a large percentage of sales usually comes from a small percentage of customers, a large percentage of quality defects stems from just a couple of sources, or a large percentage of inventory value corresponds to a small percentage of items. As a result, the Pareto principle is also often called the "80–20 rule," referring to the generic situ-

Figure 3.24

Excel Ribbon *Data* Tab



	A	B	C	D	E	F	G	H	I	J
1	Purchase Orders									
2										
3	Supplier	Order No.	Item No.	Item Description	Item Cost	Quantity	Cost per order	A/P Terms (Months)	Order Date	Arrival Date
4	Alum Sheeting	Aug11002	1243	Airframe fasteners	$ 4.25	10,000	$ 42,500.00	30	08/08/11	08/14/11
5	Alum Sheeting	Sep11002	5417	Control Panel	$ 255.00	406	$ 103,530.00	30	09/01/11	09/10/11
6	Alum Sheeting	Sep11008	1243	Airframe fasteners	$ 4.25	9,000	$ 38,250.00	30	09/05/11	09/12/11
7	Alum Sheeting	Oct11016	1243	Airframe fasteners	$ 4.25	10,500	$ 44,625.00	30	10/10/11	10/17/11
8	Alum Sheeting	Oct11022	4224	Bolt-nut package	$ 3.95	4,500	$ 17,775.00	30	10/15/11	10/20/11
9	Alum Sheeting	Oct11026	5417	Control Panel	$ 255.00	500	$ 127,500.00	30	10/20/11	10/27/11
10	Alum Sheeting	Oct11028	5634	Side Panel	$ 185.00	150	$ 27,750.00	30	10/25/11	11/03/11
11	Alum Sheeting	Oct11036	5634	Side Panel	$ 185.00	140	$ 25,900.00	30	10/29/11	11/04/11
12	Durrable Products	Aug11008	7258	Pressure Gauge	$ 90.00	100	$ 9,000.00	45	08/25/11	08/28/11
13	Durrable Products	Sep11009	7258	Pressure Gauge	$ 90.00	120	$ 10,800.00	45	09/05/11	09/09/11
14	Durrable Products	Sep11027	1369	Airframe fasteners	$ 4.20	15,000	$ 63,000.00	45	09/25/11	09/30/11
15	Durrable Products	Sep11031	1369	Airframe fasteners	$ 4.20	14,000	$ 58,800.00	45	09/27/11	10/03/11

Figure 3.25

Portion of *Purchase Orders* Database Sorted by Supplier Name

ation in which 80% of some output comes from 20% of some input. A Pareto analysis relies on sorting data and calculating the cumulative percentage of the characteristic of interest.

EXAMPLE 3.13 Applying the Pareto Principle

The Excel file *Bicycle Inventory* lists the inventory of bicycle models in a sporting goods store (see columns A through F in Figure 3.26).[2] To conduct a Pareto analysis, we first compute the inventory value of each product by multiplying the quantity on hand by the purchase cost; this is the amount invested in the items that are currently in stock. Then we sort the data in decreasing order of inventory value and compute the percentage of the total inventory value for each product and the cumulative percentage. See columns G through I in Figure 3.26. We see that about 75% of the inventory value is accounted for by less than 40% (9 of 24) of the items. If these high-value inventories aren't selling well, the store manager may wish to keep fewer in stock.

	A	B	C	D	E	F	G	H	I
1	Bicycle Inventory								
2									
3	Product Category	Product Name	Purchase Cost	Selling Price	Supplier	Quantity on Hand	Inventory Value	Percentage	Cumulative %
4	Road	Runroad 5000	$450.95	$599.99	Run-Up Bikes	5	$ 2,254.75	11.2%	11.2%
5	Road	Runroad 1000	$250.95	$350.99	Run-Up Bikes	8	$ 2,007.60	10.0%	21.1%
6	Road	Elegant 210	$281.52	$394.13	Bicyclist's Choice	7	$ 1,970.64	9.8%	30.9%
7	Road	Runroad 4000	$390.95	$495.99	Run-Up Bikes	5	$ 1,954.75	9.7%	40.6%
8	Mtn.	Eagle 3	$350.52	$490.73	Bike-One	5	$ 1,752.60	8.7%	49.3%
9	Road	Classic 109	$207.49	$290.49	Bicyclist's Choice	7	$ 1,452.43	7.2%	56.5%
10	Hybrid	Eagle 7	$150.89	$211.46	Bike-One	9	$ 1,358.01	6.7%	63.3%
11	Hybrid	Tea for Two	$429.02	$609.00	Simpson's Bike Supply	3	$ 1,287.06	6.4%	69.7%
12	Mtn.	Bluff Breaker	$375.00	$495.00	The Bike Path	3	$ 1,125.00	5.6%	75.2%
13	Mtn.	Eagle 2	$401.11	$561.54	Bike-One	2	$ 802.22	4.0%	79.2%
14	Leisure	Breeze LE	$109.95	$149.95	The Bike Path	5	$ 549.75	2.7%	81.9%
15	Children	Runkidder 100	$50.95	$75.99	Run-Up Bikes	10	$ 509.50	2.5%	84.5%
16	Mtn.	Jetty Breaker	$455.95	$649.95	The Bike Path	1	$ 455.95	2.3%	86.7%
17	Leisure	Runcool 3000	$85.95	$135.99	Run-Up Bikes	5	$ 429.75	2.1%	88.9%
18	Children	Coolest 100	$69.99	$97.98	Bicyclist's Choice	6	$ 419.94	2.1%	91.0%
19	Mtn.	Eagle 1	$410.01	$574.01	Bike-One	1	$ 410.01	2.0%	93.0%
20	Children	Green Rider	$95.47	$133.66	Simpson's Bike Supply	4	$ 381.88	1.9%	94.9%
21	Leisure	Breeze	$89.95	$130.95	The Bike Path	4	$ 359.80	1.8%	96.7%
22	Leisure	Blue Moon	$75.29	$105.41	Simpson's Bike Supply	4	$ 301.16	1.5%	98.2%
23	Leisure	Supreme 350	$50.00	$70.00	Bicyclist's Choice	3	$ 150.00	0.7%	98.9%
24	Children	Red Rider	$15.00	$25.50	Simpson's Bike Supply	8	$ 120.00	0.6%	99.5%
25	Leisure	Starlight	$100.47	$140.66	Simpson's Bike Supply	1	$ 100.47	0.5%	100.0%
26	Hybrid	Runblend 2000	$180.95	$255.99	Run-Up Bikes	0	$ -	0.0%	100.0%
27	Road	Twist & Shout	$490.50	$635.70	Simpson's Bike Supply	0	$ -	0.0%	100.0%
28						Total	$ 20,153.27		

Figure 3.26

Pareto Analysis of *Bicycle Inventory*

[2]Based on Kenneth C. Laudon and Jane P. Laudon, *Essentials of Management Information Systems*, 9th ed. (Upper Saddle River, NJ: Prentice Hall, 2011).

Filtering Data

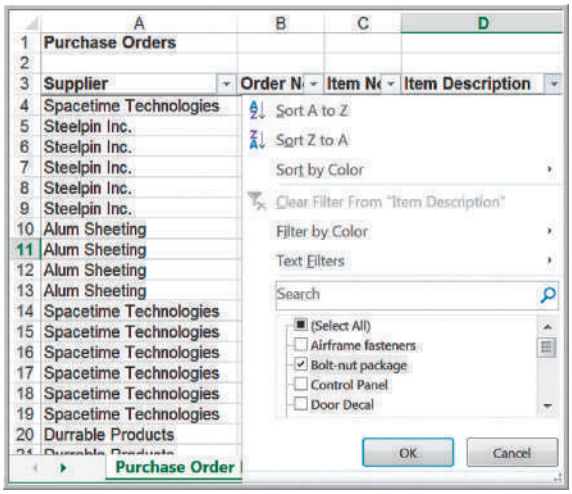
For large data files, finding a particular subset of records that meet certain characteristics by sorting can be tedious. Excel provides two filtering tools: *AutoFilter* for simple criteria and *Advanced Filter* for more complex criteria. These tools are best understood by working through some examples.

EXAMPLE 3.14 Filtering Records by Item Description

In the *Purchase Orders* database, suppose we are interested in extracting all records corresponding to the item Bolt-nut package. First, select any cell within the database. Then, from the Excel *Data* tab, click on *Filter* in the *Sort & Filter* group. A dropdown arrow will then be displayed on the right side of each header column. Clicking on one of these will display a drop-down box. These are the options for filtering on that column of data. Click the one next to the Item Description header. Uncheck the box for *Select All* and then check the box corresponding to the Bolt-nut package, as shown in Figure 3.27. Click the *OK* button, and the Filter tool will display only those orders for this item (Figure 3.28). Actually, the filter tool does not extract the records; it simply hides the records that don't match the criteria. However, you can copy and paste the data to another Excel worksheet, Microsoft Word document, or a PowerPoint presentation, for instance. To restore the original data file, click on the drop-down arrow again and then click *Clear filter from "Item Description."*

Figure 3.27

Selecting Records for Bolt-Nut Package



	A	B	C	D	E	F	G	H	I	J
1	Purchase Orders									
2										
3	Supplier	Order N	Item N	Item Description	Item Co	Quanti	Cost per orde	A/P Terms (Months	Order Dat	Arrival Dat
6	Steelpin Inc.	A0123	4312	Bolt-nut package	$ 3.75	4,250	$ 15,937.50	30	08/25/11	09/01/11
9	Steelpin Inc.	A0207	4312	Bolt-nut package	$ 3.75	4,200	$ 15,750.00	30	09/01/11	09/10/11
10	Alum Sheeting	A0223	4224	Bolt-nut package	$ 3.95	4,500	$ 17,775.00	30	10/15/11	10/20/11
19	Spacetime Technologies	A1222	4111	Bolt-nut package	$ 3.55	4,200	$ 14,910.00	25	09/15/11	10/15/11
25	Spacetime Technologies	A1444	4111	Bolt-nut package	$ 3.55	4,250	$ 15,087.50	25	09/20/11	10/10/11
26	Spacetime Technologies	A1445	4111	Bolt-nut package	$ 3.55	4,200	$ 14,910.00	25	09/25/11	10/25/11
27	Spacetime Technologies	A1449	4111	Bolt-nut package	$ 3.55	4,600	$ 16,330.00	25	10/05/11	10/19/11
29	Durrable Products	A1457	4569	Bolt-nut package	$ 3.50	3,900	$ 13,650.00	45	10/05/11	10/10/11
35	Spacetime Technologies	A3467	4111	Bolt-nut package	$ 3.55	4,800	$ 17,040.00	25	09/05/11	09/20/11
36	Spacetime Technologies	A5689	4111	Bolt-nut package	$ 3.55	4,585	$ 16,276.75	25	09/10/11	09/30/11
43	Steelpin Inc.	B0445	4312	Bolt-nut package	$ 3.75	4,150	$ 15,562.50	30	09/03/11	09/11/11

Figure 3.28

Filter Results for Bolt-Nut Package

EXAMPLE 3.15 Filtering Records by Item Cost

In this example, suppose we wish to identify all records in the *Purchase Orders* database whose item cost is at least $200. First, click on the drop-down arrow in the Item Cost column and position the cursor over *Numbers Filter*. This displays a list of options, as shown in Figure 3.29. Select *Greater Than Or Equal To . . .* from the list. This brings up a *Custom AutoFilter* dialog (Figure 3.30) that allows you to specify up to two specific criteria using "and" and "or" logic. Enter 200 in the box as shown and then click *OK*. The tool will display all records having an item cost of $200 or more.

AutoFilter creates filtering criteria based on the type of data being filtered. For instance, in Figure 3.29 we see that the *Number Filters* menu list includes numerical criteria such as "equals," "does not equal," and so on. If you choose to filter on Order Date or Arrival Date, the AutoFilter tools will display a different *Date Filters* menu list for filtering that includes "tomorrow," "next week," "year to date," and so on.

The AutoFilter can be used sequentially to "drill down" into the data. For example, after filtering the results by Bolt-nut package in Figure 3.28, we could then filter by order date and select all orders processed in September.

Figure 3.29

Selecting Records for Item Cost Filtering

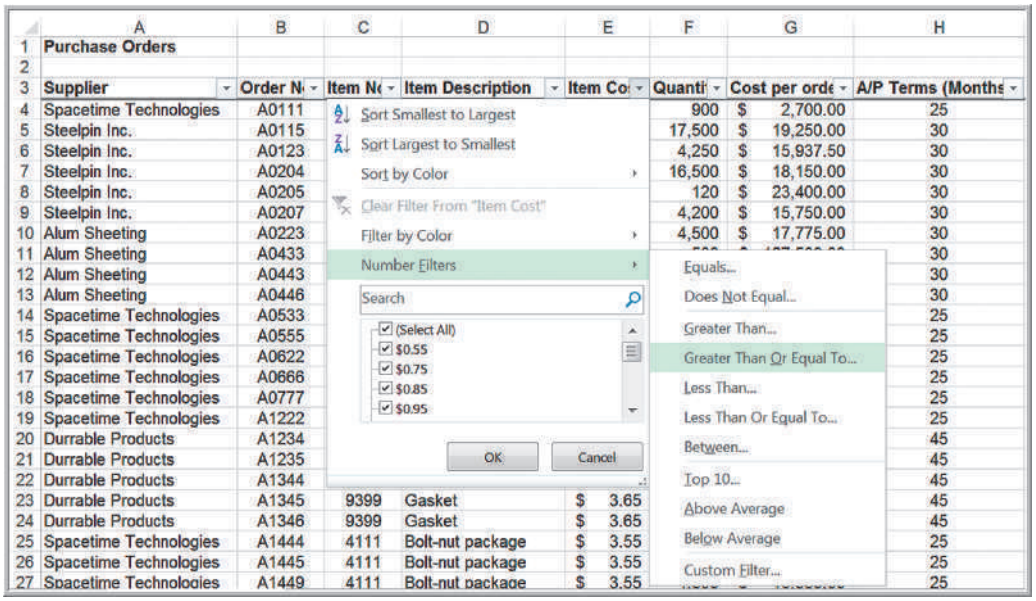
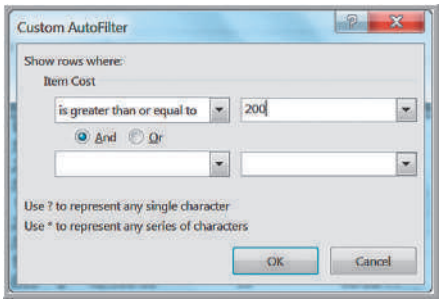


Figure 3.30

Custom AutoFilter Dialog



Analytics in Practice: Discovering the Value of Data Analysis at Allders International[3]

Allders International specializes in duty-free operations with 82 tax-free retail outlets throughout Europe, including shops in airports and seaports and on cross-channel ferries. Like most retail outlets, Allders International must track masses of point-of-sale data to assist in inventory and product-mix decisions. Which items to stock at each of its outlets can have a significant impact on the firm's profitability. To assist them, they implemented a computer-based data warehouse to maintain the data. Prior to doing this, they had to analyze large quantities of paper-based data. Such a manual process was so overwhelming and time-consuming that the analyses were often too late to provide useful information for their decisions. The data warehouse allowed the company to make simple queries, such as finding the performance of a particular item across all retail outlets or the financial performance of a particular outlet, quickly and easily. This allowed them to identify which inventory items or outlets were underperforming. For instance, a Pareto analysis of its product lines (groups of similar items) found that about 20% of the product lines were generating 80% of the profits. This allowed them to selectively eliminate some of the items from the other 80% of the product lines, which freed up shelf space for more profitable items and reduced inventory and supplier costs.

Ermek/Shutterstock.com

Statistical Methods for Summarizing Data

Statistics, as defined by David Hand, past president of the Royal Statistical Society in the UK, is *both the science of uncertainty and the technology of extracting information from data*.[4] Statistics involves collecting, organizing, analyzing, interpreting, and presenting data. A **statistic** is a summary measure of data. You are undoubtedly familiar with the concept of statistics in daily life as reported in newspapers and the media: baseball batting averages, airline on-time arrival performance, and economic statistics such as the Consumer Price Index are just a few examples.

Statistical methods are essential to business analytics and are used throughout this book. Microsoft Excel supports statistical analysis in two ways:

1. With statistical functions that are entered in worksheet cells directly or embedded in formulas
2. With the Excel *Analysis Toolpak* add-in to perform more complex statistical computations. We wish to point out that Excel for the Mac does not support the *Analysis Toolpak*. Some of these procedures are available in the free

[3]Based on Stephen Pass, "Discovering Value in a Mountain of Data," *OR/MS Today,* 24, 5, (December 1997): 24–28. (*OR/MS Today* was the predecessor of Analytics magazine.)
[4]David Hand, "Statistics: An Overview," in Miodrag Lovric, Ed., *International Encyclopedia of Statistical Science*, Springer Major Reference; http://www.springer.com/statistics/book/978-3-642-04897-5, p. 1504.

	A	B	C	D	E	F	G	H	I	J
1	Purchase Orders									
2										
3	Supplier	Order No.	Item No.	Item Description	Item Cost	Quantity	Cost per order	A/P Terms (Months)	Order Date	Arrival Date
4	Hulkey Fasteners	Aug11001	1122	Airframe fasteners	$ 4.25	19,500	$ 82,875.00	30	08/05/11	08/13/11
5	Alum Sheeting	Aug11002	1243	Airframe fasteners	$ 4.25	10,000	$ 42,500.00	30	08/08/11	08/14/11
6	Fast-Tie Aerospace	Aug11003	5462	Shielded Cable/ft.	$ 1.05	23,000	$ 24,150.00	30	08/10/11	08/15/11
7	Fast-Tie Aerospace	Aug11004	5462	Shielded Cable/ft.	$ 1.05	21,500	$ 22,575.00	30	08/15/11	08/22/11
8	Steelpin Inc.	Aug11005	5319	Shielded Cable/ft.	$ 1.10	17,500	$ 19,250.00	30	08/20/11	08/31/11
9	Fast-Tie Aerospace	Aug11006	5462	Shielded Cable/ft.	$ 1.05	22,500	$ 23,625.00	30	08/20/11	08/26/11
10	Steelpin Inc.	Aug11007	4312	Bolt-nut package	$ 3.75	4,250	$ 15,937.50	30	08/25/11	09/01/11
11	Durrable Products	Aug11008	7258	Pressure Gauge	$ 90.00	100	$ 9,000.00	45	08/25/11	08/28/11
12	Fast-Tie Aerospace	Aug11009	6321	O-Ring	$ 2.45	1,300	$ 3,185.00	30	08/25/11	09/04/11

Figure 3.31

Portion of *Purchase Orders* Database

edition of StatPlus:mac LE (www.analystsoft.com). A more complete version, StatPlus:mac Pro, can also be purchased. Some significant differences, however, exist in the tools between the Excel and Mac versions.

We use both statistical functions and the *Analysis Toolpak* in many examples.

Descriptive statistics refers to methods of describing and summarizing data using tabular, visual, and quantitative techniques. In the remainder of this chapter, we focus on some tabular and visual methods for analyzing categorical and numerical data; in the next chapter, we discuss quantitative measures.

Frequency Distributions for Categorical Data

A **frequency distribution** is a table that shows the number of observations in each of several nonoverlapping groups. Categorical variables naturally define the groups in a frequency distribution. For example, in the *Purchase Orders* database (see Figure 3.31), orders were placed for the following items:

Airframe fasteners
Bolt-nut package
Control Panel
Door Decal
Electrical Connector
Gasket
Hatch Decal
Machined Valve
O-Ring
Panel Decal
Pressure Gauge
Shielded Cable/ft.
Side Panel

To construct a frequency distribution, we need only count the number of observations that appear in each category. This can be done using the Excel COUNTIF function.

EXAMPLE 3.16 Constructing a Frequency Distribution for Items in the *Purchase Orders* Database

First, list the item names in a column on the spreadsheet. We used column A, starting in cell A100, below the existing data array. It is important to use the exact names as used in the data file. To count the number of orders placed for each item, use the function =COUNTIF($D$4:$D$97, *cell_reference*), where *cell_reference* is the cell containing the item name, our cell A101. This is shown in Figure 3.32. The resulting frequency distribution for the items is shown in Figure 3.33. Thus, the company placed 14 orders for Airframe fasteners and 11 orders for the Bolt-nut package. We may also construct a column chart to visualize these frequencies, as shown in Figure 3.34. We might wish to sort these using Pareto analysis to gain more insight into the order frequency.

Figure 3.32

Using the COUNTIF Function to Construct a Frequency Distribution

	A	B
100	**Item Description**	**Frequency**
101	Airframe fasteners	=COUNTIF($D$4:$D$97,A101)
102	Bolt-nut package	=COUNTIF($D$4:$D$97,A102)
103	Control Panel	=COUNTIF($D$4:$D$97,A103)
104	Door Decal	=COUNTIF($D$4:$D$97,A104)
105	Electrical Connector	=COUNTIF($D$4:$D$97,A105)
106	Gasket	=COUNTIF($D$4:$D$97,A106)
107	Hatch Decal	=COUNTIF($D$4:$D$97,A107)
108	Machined Valve	=COUNTIF($D$4:$D$97,A108)
109	O-Ring	=COUNTIF($D$4:$D$97,A109)
110	Panel Decal	=COUNTIF($D$4:$D$97,A110)
111	Pressure Gauge	=COUNTIF($D$4:$D$97,A111)
112	Shielded Cable/ft.	=COUNTIF($D$4:$D$97,A112)
113	Side Panel	=COUNTIF($D$4:$D$97,A113)

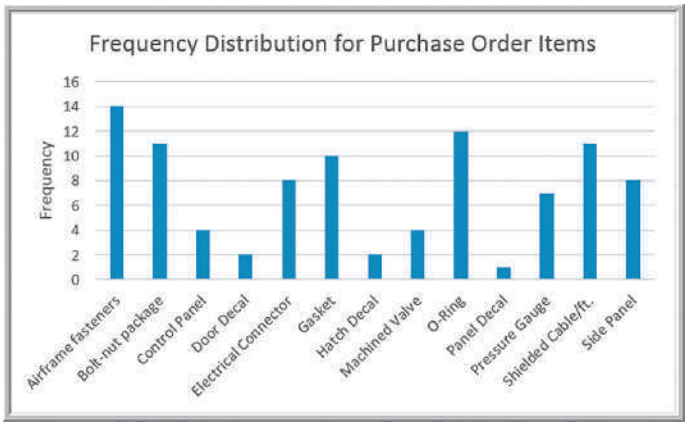
Figure 3.33

Frequency Distribution for Items Purchased

	A	B
100	**Item Description**	**Frequency**
101	Airframe fasteners	14
102	Bolt-nut package	11
103	Control Panel	4
104	Door Decal	2
105	Electrical Connector	8
106	Gasket	10
107	Hatch Decal	2
108	Machined Valve	4
109	O-Ring	12
110	Panel Decal	1
111	Pressure Gauge	7
112	Shielded Cable/ft.	11
113	Side Panel	8

Figure 3.34

Column Chart for Frequency Distribution of Items Purchased



Relative Frequency Distributions

We may express the frequencies as a fraction, or proportion, of the total; this is called the **relative frequency**. If a data set has n observations, the relative frequency of category i is computed as

$$\text{relative frequency of category } i = \frac{\text{frequency of category } i}{n} \tag{3.1}$$

We often multiply the relative frequencies by 100 to express them as percentages. A **relative frequency distribution** is a tabular summary of the relative frequencies of all categories.

Figure 3.35

Relative Frequency Distribution for Items Purchased

	A	B	C
100	Item Description	Frequency	Relative Frequency
101	Airframe fasteners	14	0.1489
102	Bolt-nut package	11	0.1170
103	Control Panel	4	0.0426
104	Door Decal	2	0.0213
105	Electrical Connector	8	0.0851
106	Gasket	10	0.1064
107	Hatch Decal	2	0.0213
108	Machined Valve	4	0.0426
109	O-Ring	12	0.1277
110	Panel Decal	1	0.0106
111	Pressure Gauge	7	0.0745
112	Shielded Cable/ft.	11	0.1170
113	Side Panel	8	0.0851
114	Total	94	1.0000

EXAMPLE 3.17 Constructing a Relative Frequency Distribution for Items in the *Purchase Orders* Database

The calculations for relative frequencies are simple. First, sum the frequencies to find the total number (note that the sum of the frequencies must be the same as the total number of observations, *n*). Then divide the frequency of each category by this value. Figure 3.35 shows the relative frequency distribution for the purchase order items. The formula in cell C101, for example, is =B101/$B$114 . You then copy this formula down the column to compute the other relative frequencies. Note that the sum of the relative frequencies must equal 1.0. A pie chart of the frequencies is sometimes used to show these proportions visually, although it is more appealing for a smaller number of categories. For a large number of categories, a column or bar chart would work better.

Frequency Distributions for Numerical Data

For numerical data that consist of a small number of discrete values, we may construct a frequency distribution similar to the way we did for categorical data; that is, we simply use COUNTIF to count the frequencies of each discrete value.

EXAMPLE 3.18 Frequency and Relative Frequency Distribution for A/P Terms

In the *Purchase Orders* data, the A/P terms are all whole numbers 15, 25, 30, and 45. A frequency and relative frequency distribution for these data is shown in Figure 3.36. A bar chart showing the proportions, or relative frequencies, in Figure 3.37, clearly shows that the majority of orders had accounts payable terms of 30 months.

Excel Histogram Tool

A graphical depiction of a frequency distribution for numerical data in the form of a column chart is called a **histogram**. Frequency distributions and histograms can be created using the *Analysis Toolpak* in Excel. To do this, click the *Data Analysis* tools button in the

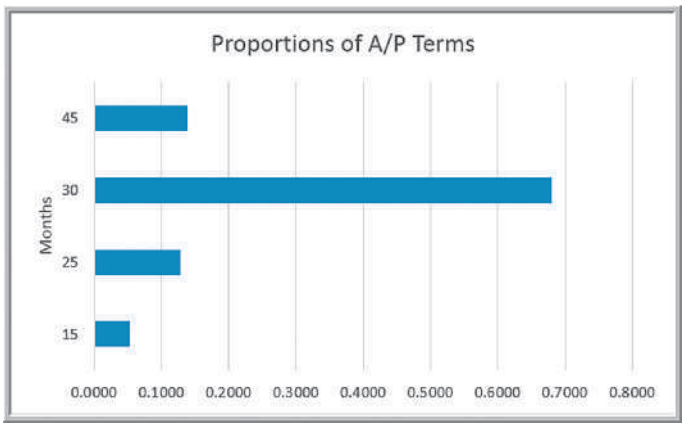
Figure 3.36

Frequency and Relative Frequency Distribution for A/P Terms

	A	B	C
117	A/P Terms	Frequency	Relative Frequency
118	15	5	0.0532
119	25	12	0.1277
120	30	64	0.6809
121	45	13	0.1383
122	Total	94	1.0000

Figure 3.37

Bar Chart of Relative Frequencies of A/P Terms



Analysis group under the *Data* tab in the Excel menu bar and select *Histogram* from the list. In the dialog box (see Figure 3.38), specify the *Input Range* corresponding to the data. If you include the column header, then also check the *Labels* box so Excel knows that the range contains a label. The *Bin Range* defines the groups (Excel calls these "bins") used for the frequency distribution. If you do not specify a *Bin Range,* Excel will automatically determine bin values for the frequency distribution and histogram, which often results in a rather poor choice. If you have discrete values, set up a column of these values in your spreadsheet for the bin range and specify this range in the *Bin Range* field. We describe how to handle continuous data shortly. Check the *Chart Output* box to display a histogram in addition to the frequency distribution. You may also sort the values as a Pareto chart and display the cumulative frequencies by checking the additional boxes.

EXAMPLE 3.19 Using the *Histogram* Tool

We will create a frequency distribution and histogram for the A/P Terms variable in the *Purchase Orders* database. Figure 3.39 shows the completed histogram dialog. The input range includes the column header as well as the data in column H. We defined the bin range below the data in cells H99:H103 as follows:

Months
15
25
30
45

If you check the *Labels* box, it is important that both the *Input Range* and the *Bin Range* have labels included in the first row. Figure 3.40 shows the results from this tool.

For numerical data that have many different discrete values with little repetition or are continuous, a frequency distribution requires that we define by specifying

1. the number of groups,
2. the width of each group, and
3. the upper and lower limits of each group.

Figure 3.38

Histogram Tool Dialog

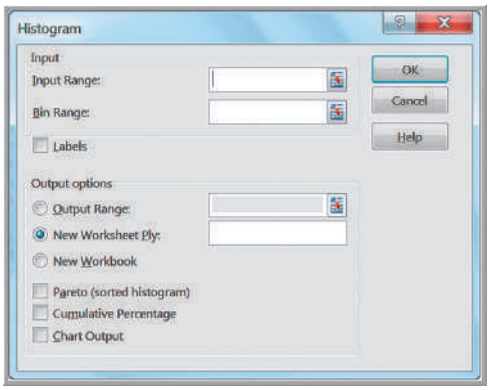


Figure 3.39

Histogram Dialog for A/P Terms Data

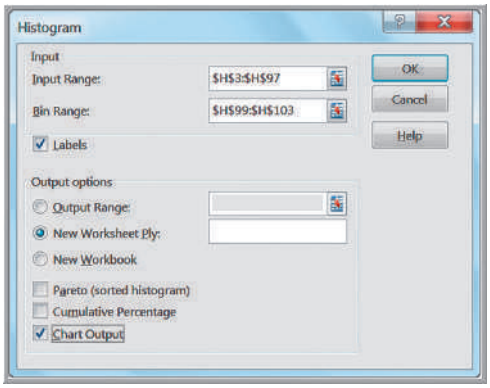
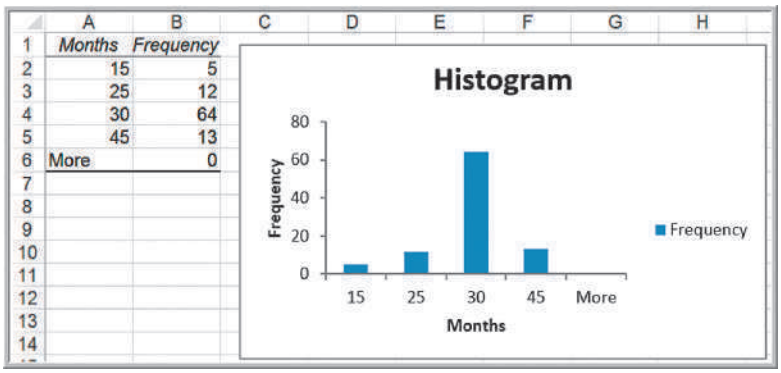


Figure 3.40

Excel Frequency Distribution and Histogram for A/P Terms



It is important to remember that the groups may not overlap, so that each value is counted in exactly one group.

You should define the groups after examining the range of the data. Generally, you should choose between 5 to 15 groups, and the range of each should be equal. The more data you have, the more groups you should generally use. Note that with fewer groups, the group widths will be wider. Wider group widths provide a "coarse" histogram. Sometimes you need to experiment to find the best number of groups to provide a useful visualization of the data. Choose the lower limit of the first group (LL) as a whole number smaller than the minimum data value and the upper limit of the last group (UL) as a whole number

larger than the maximum data value. Generally, it makes sense to choose nice, round whole numbers. Then you may calculate the group width as

$$\text{group width} = \frac{\text{UL} - \text{LL}}{\text{number of groups}} \tag{3.2}$$

EXAMPLE 3.20 Constructing a Frequency Distribution and Histogram for Cost per Order

In this example, we apply the Excel *Histogram* tool to the Cost per order data in column G of the *Purchase Orders* database. The data range from a minimum of $68.75 to a maximum of $127,500. You can find this either by using the MIN and MAX functions or simply by sorting the data. To ensure that all the data will be included in some group, it makes sense to set the lower limit of the first group to $0 and the upper limit of the last group to $130,000. Thus, if we select 5 groups, using equation (3.2) the width of each group is ($130,000 − 0)/5 = $26,000; if we choose 10 groups, the width is ($130,000 − 0)/10 = $13,000. We select 5 groups. Doing so, the bin range is specified as

Upper Group Limit
$ 0.00
$ 26,000.00
$ 52,000.00
$ 78,000.00
$104,000.00
$130,000.00

This means that the first group includes all values less than or equal to $0; the second group includes all values greater than $0 but less than or equal to $26,000, and so on. Note that the groups do not overlap because the lower limit of one group is strictly greater than the upper limit of the previous group. We suggest using the header "Upper Group Limit" for the bin range to make this clear. In the spreadsheet, this bin range is entered in cells G99:G105. The *Input Range* in the *Histogram* dialog is G4:G97. Figure 3.41 shows the results. These results show that the vast majority of orders were for $26,000 or less and fall rapidly beyond this value. Selecting a larger number of groups might help to better understand the nature of the data. Figure 3.42 shows results using 10 groups. This shows that a higher percentage of orders were for $13,000 or less than were between $13,000 and $26,000.

Figure 3.41

Frequency Distribution and Histogram for Cost per Order (5 Groups)

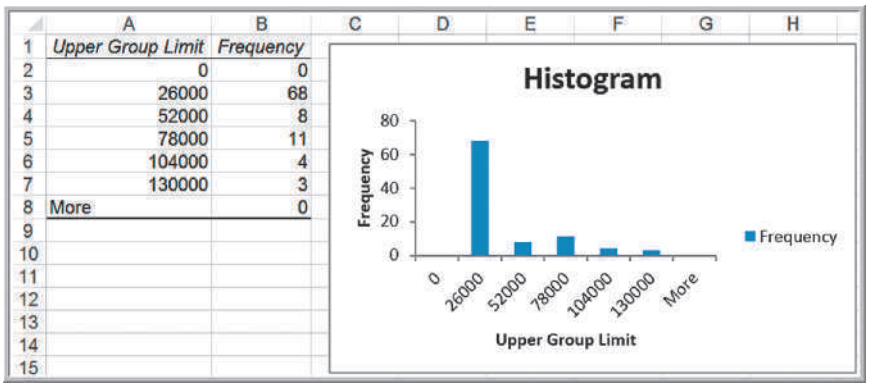
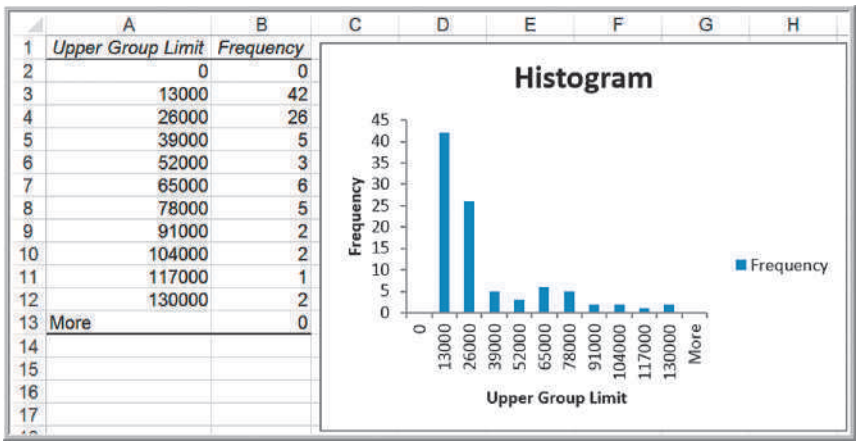


Figure 3.42

Frequency Distribution and Histogram for Cost per Order (10 Groups)



	A	B
1	*Upper Group Limit*	*Frequency*
2	0	0
3	13000	42
4	26000	26
5	39000	5
6	52000	3
7	65000	6
8	78000	5
9	91000	2
10	104000	2
11	117000	1
12	130000	2
13	More	0

One limitation of the Excel *Histogram* tool is that the frequency distribution and histogram are not linked to the data; thus, if you change any of the data, you must repeat the entire procedure to construct a new frequency distribution and histogram.

Cumulative Relative Frequency Distributions

For numerical data, we may also compute the relative frequency of observations in each group. By summing all the relative frequencies at or below each upper limit, we obtain the cumulative relative frequency. The **cumulative relative frequency** represents the proportion of the total number of observations that fall at or below the upper limit of each group. A tabular summary of cumulative relative frequencies is called a **cumulative relative frequency distribution**.

EXAMPLE 3.21 Computing Cumulative Relative Frequencies

Figure 3.43 shows the relative frequency and cumulative relative frequency distributions for the Cost per order data using 10 groups. The relative frequencies are computed using the same approach as in Example 3.17—namely, by dividing the frequency by the total number of observations (94). In column D, we set the cumulative relative frequency of the first group equal to its relative frequency. Then we add the relative frequency of the next group to the cumulative relative frequency. For, example, the cumulative relative frequency in cell D3 is computed as =D2+C3 = 0.000+0.447 = 0.447; the cumulative relative frequency in cell D4 is computed as =D3+C4 = 0.447+0.277 = 0.723, and so on. (Values shown are rounded to three decimal places.) Because relative frequencies must be between 0 and 1 and must add up to 1, the cumulative frequency for the last group must equal 1.

Figure 3.44 shows a chart for the cumulative relative frequency, which is called an **ogive**. From this chart, you can easily estimate the proportion of observations that fall below a certain value. For example, you can see that slightly more than 70% of the data fall at or below $26,000, about 90% of the data fall at or below $78,000, and so on.

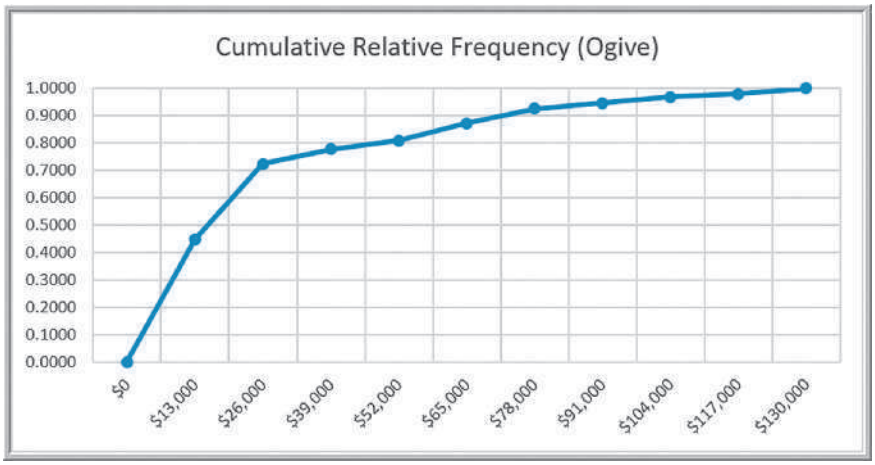
Figure 3.43

Cumulative Relative Frequency Distribution for Cost per Order Data

	A	B	C	D
1	*Upper Group Limit*	*Frequency*	*Relative Frequency*	*Cumulative Relative Frequency*
2	0	0	0.0000	0.0000
3	13000	42	0.4468	0.4468
4	26000	26	0.2766	0.7234
5	39000	5	0.0532	0.7766
6	52000	3	0.0319	0.8085
7	65000	6	0.0638	0.8723
8	78000	5	0.0532	0.9255
9	91000	2	0.0213	0.9468
10	104000	2	0.0213	0.9681
11	117000	1	0.0106	0.9787
12	130000	2	0.0213	1.0000
13	More	0	0.0000	1.0000
14	Total	94		

Figure 3.44

Ogive for Cost per Order



Percentiles and Quartiles

Data are often expressed as *percentiles* and *quartiles*. You are no doubt familiar with percentiles from standardized tests used for college or graduate school entrance examinations (SAT, ACT, GMAT, GRE, etc.). Percentiles specify the percent of other test takers who scored at or below the score of a particular individual. Generally speaking, the ***k*th percentile** is a value at or below which at least k percent of the observations lie. However, the way by which percentiles are calculated is not standardized. The most common way to compute the kth percentile is to order the data values from smallest to largest and calculate the rank of the kth percentile using the formula

$$\frac{nk}{100} + 0.5 \qquad \textbf{(3.3)}$$

where n is the number of observations. Round this to the nearest integer, and take the value corresponding to this rank as the kth percentile.

EXAMPLE 3.22 Computing Percentiles

In the *Purchase Orders* data, we have $n = 94$ observations. The rank of the 90th percentile ($k = 90$) for the Cost per order data is computed as $94(90)/100 + 0.5 = 85.1$, or, rounded, 85. The 85th ordered value is \$74,375 and is the 90th percentile. This means that 90% of the costs per order are less than or equal to \$74,375, and 10% are higher.

Statistical software use different methods that often involve interpolating between ranks instead of rounding, thus producing different results. The Excel function PERCENTILE.INC(*array, k*) computes the *k*th percentile of data in the range specified in the *array* field, where *k* is in the range 0 to 1, inclusive.

EXAMPLE 3.23 Computing Percentiles in Excel

To find the 90th percentile for the Cost per order data in the *Purchase Orders* data, use the Excel function PERCENTILE. INC(G4:G97,0.9). This calculates the 90th percentile as $73,737.50, which is different from using formula (3.3).

Excel also has a tool for sorting data from high to low and computing percentiles associated with each value. Select *Rank and Percentile* from the *Data Analysis* menu and specify the range of the data in the dialog. Be sure to check the *Labels in First Row* box if your range includes a header in the spreadsheet.

EXAMPLE 3.24 Excel *Rank and Percentile* Tool

A portion of the results from the *Rank and Percentile* tool for the Cost per order data are shown in Figure 3.45. You can see that the Excel value of the 90th percentile that we computed in Example 3.22 as $74,375 is the 90.3rd percentile value.

Quartiles break the data into four parts. The 25th percentile is called the *first quartile*, Q_1; the 50th percentile is called the *second quartile*, Q_2; the 75th percentile is called the *third quartile*, Q_3; and the 100th percentile is the *fourth quartile*, Q_4. One-fourth of the data fall below the first quartile, one-half are below the second quartile, and three-fourths are below the third quartile. We may compute quartiles using the Excel function QUARTILE.INC(*array, quart*), where *array* specifies the range of the data and *quart* is a whole number between 1 and 4, designating the desired quartile.

Figure 3.45

Portion of *Rank and Percentile* Tool Results

	A	B	C	D
1	*Point*	*Cost per order*	*Rank*	*Percent*
2	74	$127,500.00	1	100.00%
3	62	$121,000.00	2	98.90%
4	71	$110,000.00	3	97.80%
5	16	$103,530.00	4	96.70%
6	73	$ 96,750.00	5	95.60%
7	1	$ 82,875.00	6	94.60%
8	67	$ 81,937.50	7	93.50%
9	82	$ 77,400.00	8	92.40%
10	54	$ 76,500.00	9	91.30%
11	80	$ 74,375.00	10	90.30%
12	68	$ 72,250.00	11	89.20%
13	20	$ 65,875.00	12	88.10%
14	65	$ 64,500.00	13	87.00%
15	28	$ 63,750.00	14	86.00%

EXAMPLE 3.25 Computing Quartiles in Excel

For the Cost per order data in the *Purchase Orders* database, we may use the Excel function =QUARTILE.INC(G4:G97,*k*), where *k* ranges from 1 to 4, to compute the quartiles. The results are as follows:

$k = 1$	First quartile	$6,757.81
$k = 2$	Second quartile	$15,656.25
$k = 3$	Third quartile	$27,593.75
$k = 4$	Fourth quartile	$127,500.00

We may conclude that 25% of the order costs fall at or below $6,757.81; 50% fall at or below $15,656.25; 75% fall at or below $27,593.75, and 100% fall at or below the maximum value of $127,500.

We can extend these ideas to other divisions of the data. For example, *deciles* divide the data into 10 sets: the 10th percentile, 20th percentile, and so on. All these types of measures are called **data profiles**, or **fractiles**.

Cross-Tabulations

One of the most basic statistical tools used to summarize categorical data and examine the relationship between two categorical variables is cross-tabulation. A **cross-tabulation** is a tabular method that displays the number of observations in a data set for different subcategories of two categorical variables. A cross-tabulation table is often called a **contingency table**. The subcategories of the variables must be mutually exclusive and exhaustive, meaning that each observation can be classified into only one subcategory, and, taken together over all subcategories, they must constitute the complete data set. Cross-tabulations are commonly used in marketing research to provide insight into characteristics of different market segments using categorical variables such as gender, educational level, marital status, and so on.

EXAMPLE 3.26 Constructing a Cross-Tabulation

Let us examine the *Sales Transactions* database, a portion of which is shown in Figure 3.46. Suppose we wish to identify the number of books and DVDs ordered by region. A cross-tabulation will have rows corresponding to the different regions and columns corresponding to the products. Within the table we list the count of the number in each pair of categories. A cross-tabulation of these data is shown in Table 3.1. Visualizing the data as a chart is a good way of communicating the results. Figure 3.47 shows the differences between product and regional sales. It is somewhat difficult to directly count the numbers of observations easily in an Excel data file; however, an Excel tool called a PivotTable makes this easy. PivotTables are introduced in the next section.

Expressing the results as percentages of a row or column makes it easier to interpret differences between regions or products, particularly as the totals for each category differ. Table 3.2 shows the percentage of book and DVD sales within each region; this is computed by dividing the counts by the row totals and multiplying by 100 (in Excel, simply divide the count by the total and format the result as a percentage by clicking the % button in the *Number* group within the *Home* tab in the ribbon). For example, we see that although more books and DVDs are sold in the West region than in the North, the relative percentages of each product are similar, particularly when compared to the East and South regions.

Figure 3.46

Portion of *Sales Transactions* Database

	A	B	C	D	E	F	G	H
1	Sales Transactions: July 14							
2								
3	Cust ID	Region	Payment	Transaction Code	Source	Amount	Product	Time Of Day
4	10001	East	Paypal	93816545	Web	$20.19	DVD	22:19
5	10002	West	Credit	74083490	Web	$17.85	DVD	13:27
6	10003	North	Credit	64942368	Web	$23.98	DVD	14:27
7	10004	West	Paypal	70560957	Email	$23.51	Book	15:38
8	10005	South	Credit	35208817	Web	$15.33	Book	15:21
9	10006	West	Paypal	20978903	Email	$17.30	DVD	13:11
10	10007	East	Credit	80103311	Web	$177.72	Book	21:59
11	10008	West	Credit	14132683	Web	$21.76	Book	4:04
12	10009	West	Paypal	40128225	Web	$15.92	DVD	19:35
13	10010	South	Paypal	49073721	Web	$23.39	DVD	13:26

Table 3.1

Cross-Tabulation of Sales Transaction Data

Region	Book	DVD	Total
East	56	42	98
North	43	42	85
South	62	37	99
West	100	90	190
Total	261	211	472

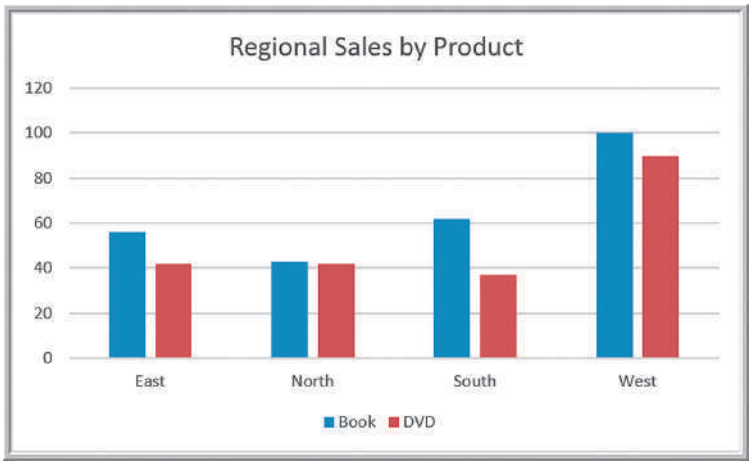
Table 3.2

Percentage Sales of Products within Each Region

Region	Book	DVD	Total
East	57.1%	42.9%	100.0%
North	50.6%	49.4%	100.0%
South	62.6%	37.4%	100.0%
West	52.6%	47.4%	100.0%

Figure 3.47

Chart of Regional Sales by Product



Exploring Data Using PivotTables

Excel provides a powerful tool for distilling a complex data set into meaningful information: **PivotTables** (yes, it is one word!). PivotTables allows you to create custom summaries and charts of key information in the data. PivotTables can be used to quickly create cross-tabulations and to drill down into a large set of data in numerous ways.

To apply PivotTables, you need a data set with column labels in the first row, similar to the data files we have been using. Select any cell in the data set and choose *PivotTable* from the *Tables* group under the *Insert* tab and follow the steps of the wizard. Excel first asks you to select a table or range of data; if you click on any cell within the data matrix before inserting a PivotTable, Excel will default to the complete range of your data. You may either put the PivotTable into a new worksheet or in a blank range of the existing worksheet. Excel then creates a blank PivotTable, as shown in Figure 3.48.

In the *PivotTable Field List* on the right side of Figure 3.48 is a list of the fields that correspond to the headers in the data file. You select which ones you want to include, either as row labels, column labels, values, or what is called a Report Filter. You should first decide what types of tables you wish to create—that is, what fields you want for the rows, columns, and data values.

EXAMPLE 3.27 Creating a PivotTable

Let us create a cross-tabulation of regional sales by product, as we did in the previous section. If you drag the field *Region* from the *PivotTable Field List* in Figure 3.48 to the *Row Labels* area, the field *Product* into the *Column Labels* area, and any of the other fields, such as *Cust ID,* into the *Values* area, you will create the PivotTable shown in Figure 3.49. However, the sum of customer ID values (the default) is meaningless; we simply want a count of the number of records in each category. Click the *Analyze* tab, and then in the *Active Field* group and choose *Field Settings*. You will be able to change the summarization method in the PivotTable in the *Value Field Settings* dialog shown in Figure 3.50. Selecting *Count* results in the PivotTable shown in Figure 3.51, which is the cross-tabulation we showed in Table 3.1. The *Value Field Settings* options in Figure 3.50 include other options, such as Average, Max, Min, and other statistical measures that we introduce in the next chapter. It also allows you to format the data properly (for example, currency or to display a fixed number of decimals) by clicking on the *Number Format* button.

Figure 3.48

Blank PivotTable

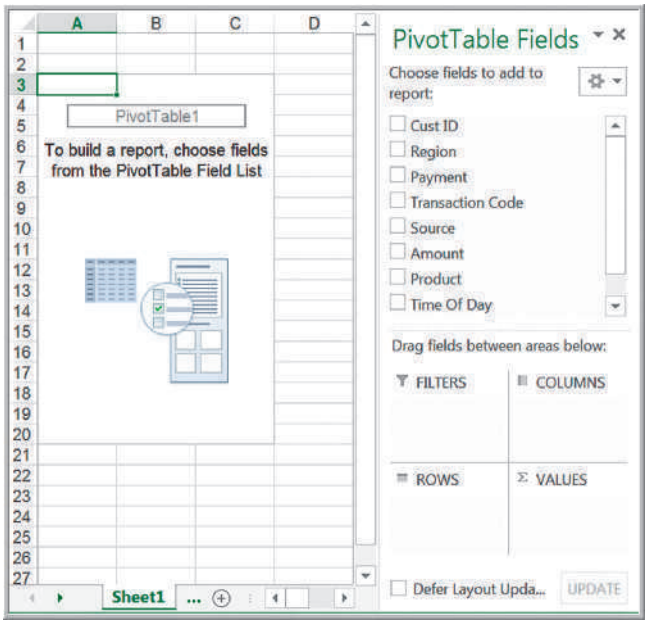


Figure 3.49

Default PivotTable for Regional Sales by Product

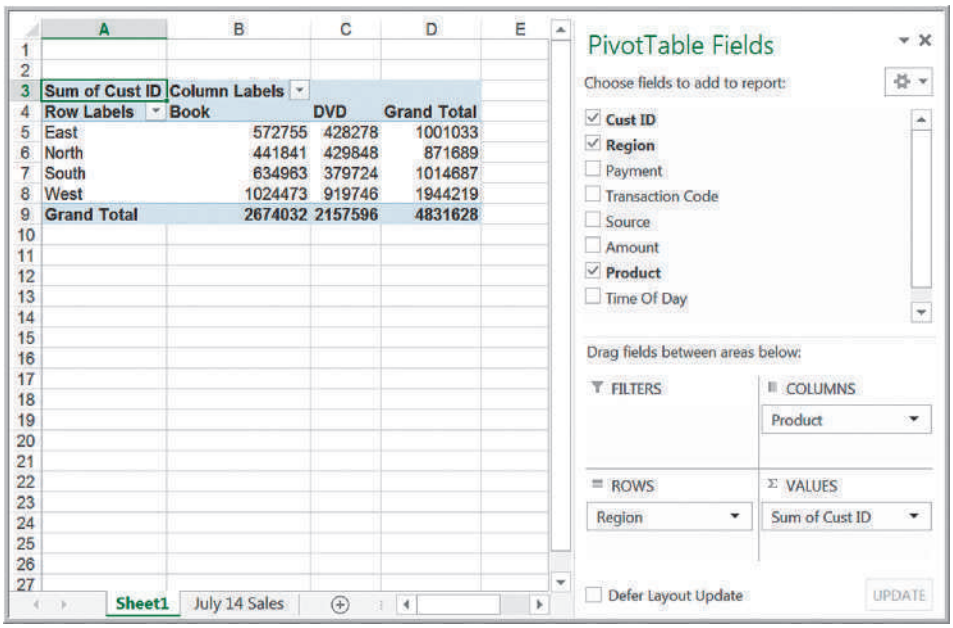


Figure 3.50

Value Field Settings Dialog

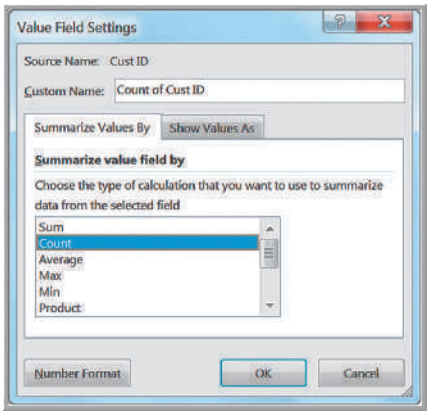


Figure 3.51

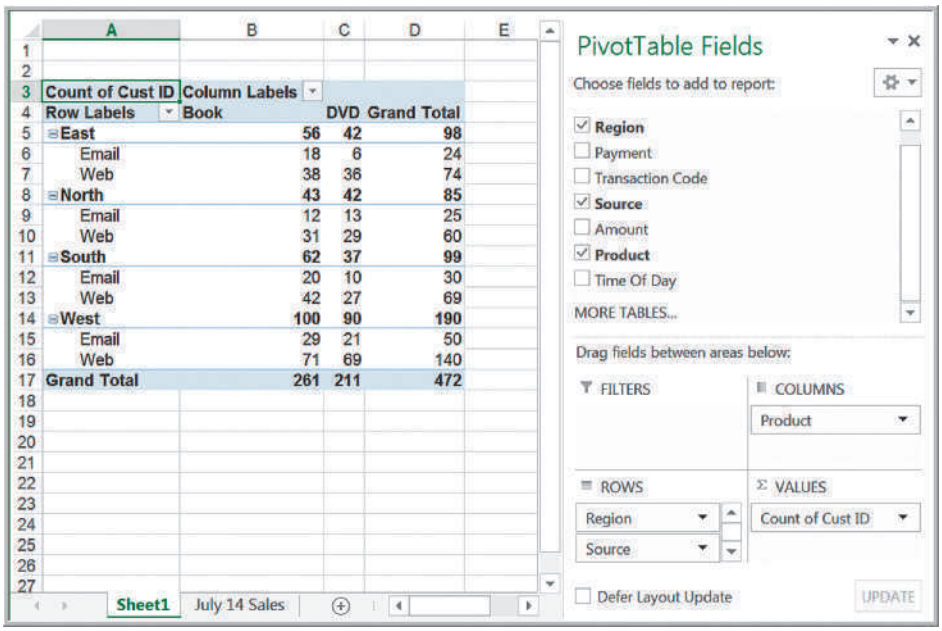
PivotTable for Count of Regional Sales by Product

Count of Cust ID	Column Labels		
Row Labels	Book	DVD	Grand Total
East	56	42	98
North	43	42	85
South	62	37	99
West	100	90	190
Grand Total	261	211	472

The beauty of PivotTables is that if you wish to change the analysis, you can simply uncheck the boxes in the *PivotTable Field List* or drag the field names to different areas. You may easily add multiple variables in the fields to create different views of the data. For example, if you drag the *Source* field into the *Row Labels* area, you will create the

Figure 3.52

PivotTable for Sales by Region, Product, and Order Source



Count of Cust ID	Column Labels		
Row Labels	Book	DVD	Grand Total
East	56	42	98
Email	18	6	24
Web	38	36	74
North	43	42	85
Email	12	13	25
Web	31	29	60
South	62	37	99
Email	20	10	30
Web	42	27	69
West	100	90	190
Email	29	21	50
Web	71	69	140
Grand Total	261	211	472

PivotTable shown in Figure 3.52. This shows a count of the number of sales by region and product that is also broken down by how the orders were placed—either by e-mail or on the Web.

Dragging a field into the *Report Filter* area in the *PivotTable Field* list allows you to add a third dimension to your analysis. Example 3.28 illustrates this. You may create other PivotTables without repeating all the steps in the Wizard. Simply copy and paste the first table. The best way to learn about PivotTables is simply to experiment with them.

EXAMPLE 3.28 Using the PivotTable Report Filter

Going back to the cross-tabulation PivotTable of regional sales by product, drag the *Payment* field into the *Report Filter* area. This places payment in row 1 of the PivotTable and allows you to break down the cross-tabulation by type of payment, as shown in Figure 3.53. Click on the drop-down arrow in row 1, and you can choose to display a cross-tabulation for one of the different payment types, Credit or Paypal. Figure 3.54 shows the results for credit-card payments, which accounted for 299 of the total number of transactions.

PivotCharts

Microsoft Excel provides a simple one-click way of creating **PivotCharts** to visualize data in PivotTables. To display a PivotChart for a PivotTable, first select the PivotTable. From the *Analyze* tab, click on *PivotChart*. Excel will display an *Insert Chart* dialog that allows you to choose the type of chart you wish to display.

Figure 3.53

PivotTable Filtered by Payment Type

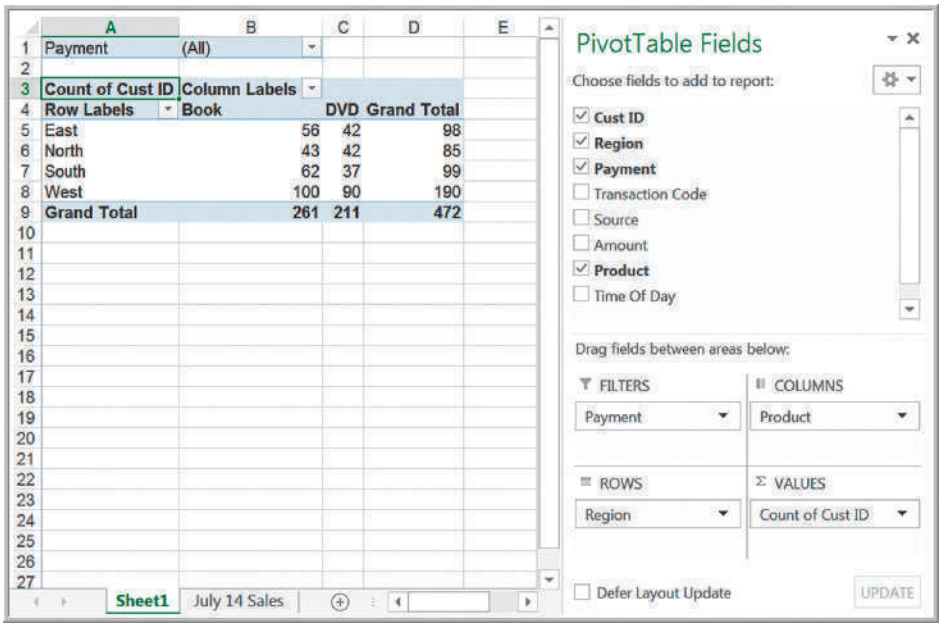


Figure 3.54

Cross-Tabulation PivotTable for Credit-Card Transactions

Payment	Credit		
Count of Cust ID	Column Labels		
Row Labels	Book	DVD	Grand Total
East	40	34	74
North	21	29	50
South	44	17	61
West	54	60	114
Grand Total	159	140	299

EXAMPLE 3.29 A PivotChart for Sales Data

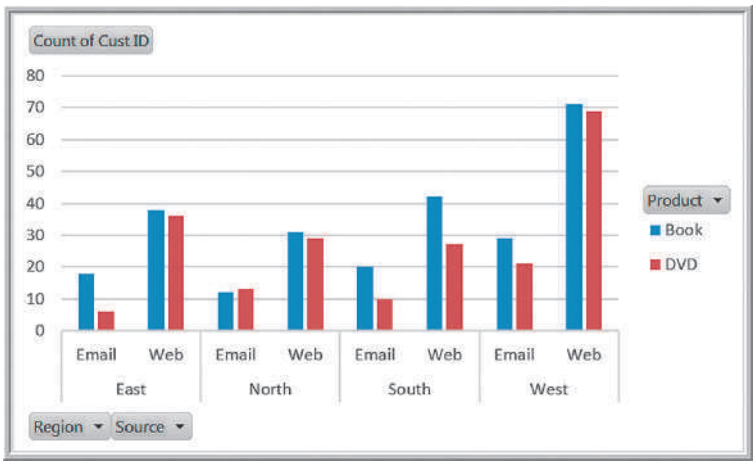
For the PivotTable shown in Figure 3.52, we choose to display a column chart from the *Insert Chart* dialog. Figure 3.55 shows the chart generated by Excel. By clicking on the drop-down buttons, you can easily change the data that are displayed by filtering the data. Also, by clicking on the chart and selecting the *PivotChart Tools Design* tab, you can switch the rows and columns to display an alternate view of the chart or change the chart type entirely.

Slicers and PivotTable Dashboards

Excel 2010 introduced **slicers**—a tool for drilling down to "slice" a PivotTable and display a subset of data. To create a slicer for any of the columns in the database, click on the PivotTable and choose *Insert Slicer* from the *Analyze* tab in the *PivotTable Tools* ribbon.

Figure 3.55

PivotChart for Sales by Region, Product, and Order Source



EXAMPLE 3.30 Using Slicers

For the PivotTable, we created in Figure 3.51 for the count of regional sales by product, let us insert a slicer for the source of the transaction as shown in Figure 3.56. In this case, we choose Source as the slicer. This results in the slicer window shown in Figure 3.57. If you click on one of the source buttons, Email or Web, the PivotTable reflects only those records corresponding to that source. In Figure 3.57, we now have a cross-tabulation only for e-mail orders.

Figure 3.56

Insert Slicers Window

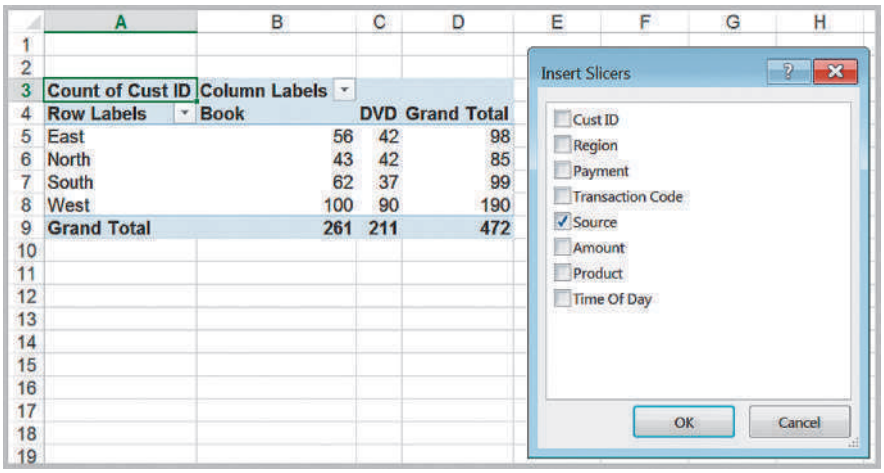
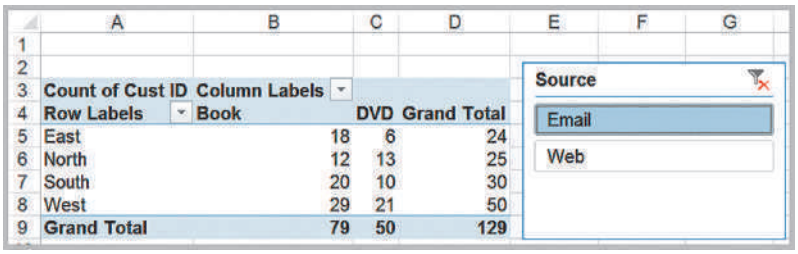
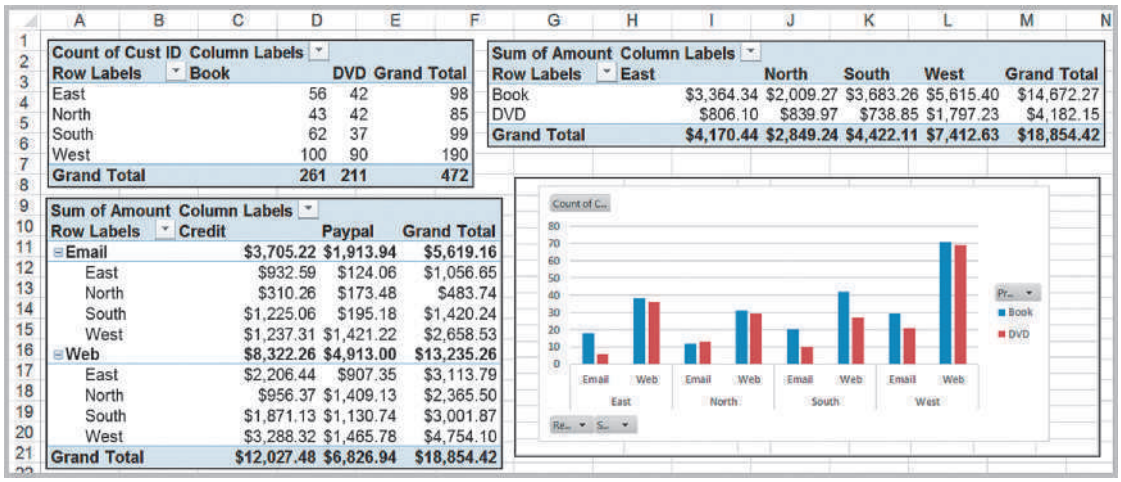


Figure 3.57

Cross-Tabulation Sliced by E-mail





Count of Cust ID	Column Labels		
Row Labels	Book	DVD	Grand Total
East	56	42	98
North	43	42	85
South	62	37	99
West	100	90	190
Grand Total	261	211	472

Sum of Amount	Column Labels				
Row Labels	East	North	South	West	Grand Total
Book	$3,364.34	$2,009.27	$3,683.26	$5,615.40	$14,672.27
DVD	$806.10	$839.97	$738.85	$1,797.23	$4,182.15
Grand Total	$4,170.44	$2,849.24	$4,422.11	$7,412.63	$18,854.42

Sum of Amount	Column Labels		
Row Labels	Credit	Paypal	Grand Total
Email	$3,705.22	$1,913.94	$5,619.16
East	$932.59	$124.06	$1,056.65
North	$310.26	$173.48	$483.74
South	$1,225.06	$195.18	$1,420.24
West	$1,237.31	$1,421.22	$2,658.53
Web	$8,322.26	$4,913.00	$13,235.26
East	$2,206.44	$907.35	$3,113.79
North	$956.37	$1,409.13	$2,365.50
South	$1,871.13	$1,130.74	$3,001.87
West	$3,288.32	$1,465.78	$4,754.10
Grand Total	$12,027.48	$6,826.94	$18,854.42

Figure 3.58

Camera-Based Dashboard

Finally, we introduced the Excel camera tool earlier in this chapter. This is a useful tool for creating PivotTable-based dashboards. If you create several different PivotTables and charts, you can easily use the camera tool to take pictures of them and consolidate them onto one worksheet. In this fashion, you can still make changes to the PivotTables and they will automatically be reflected in the camera shots. Figure 3.58 shows a simple dashboard created using the camera tool for the *Sales Transactions* database.

Analytics in Practice: Driving Business Transformation with IBM Business Analytics[5]

Founded in the 1930s and headquartered in Ballinger, Texas, Mueller is a leading retailer and manufacturer of pre-engineered metal buildings and metal roofing products. Today, the company sells its products directly to consumers all over the southwestern United States from 35 locations across Texas, New Mexico, Louisiana, and Oklahoma.

Historically, Mueller saw itself first and foremost as a manufacturer; the retail aspects of the business were a secondary focus. However, in the early 2000s, the company decided to shift the focus of its strategy and become much more retail-centric—getting closer to its end-use customers and driving new business through a better understanding of their needs. To achieve its transformation objective, the company needed to communicate its retail strategy to employees across the organization.

As Mark Lack, Manager of Strategy Analytics and Business Intelligence at Mueller, explains: "The transformation from pure manufacturing to retail-led manufacturing required a more end-customer-focused approach to sales. We wanted a way to track how successfully our sales teams across the country were adapting to this new strategy, and identify where improvements could be made."

To keep track of sales performance, Mueller worked with IBM to deploy IBM® Cognos® Business Intelligence. The IBM team helped Mueller apply technology to its balanced scorecard process for strategy management in Cognos Metric Studio.

(*continued*)

[5]"Mueller builds a customer-focused business," IBM Software, Business Analytics, © IBM Corporation, 2013.

By using a common set of KPIs, Mueller can easily identify the strengths and weaknesses of all of its sales teams through sales performance analytics. "Using Metric Studio in Cognos Business Intelligence, we get a clear picture of each team's strategy performance," says Mark Lack. "Using sales performance insights from Cognos scorecards, we can identify teams that are hitting their targets, and determine the reasons for their success. We can then share this knowledge with underperforming teams, and demonstrate how they can change their way of working to meet their targets.

"Instead of just trying to impose or enforce new ways of working, we are able to show sales teams exactly how they are contributing to the business, and explain what they need to do to improve their metrics. It's a much more effective way of driving the changes in behavior that are vital for business transformation."

Recently, IBM Business Analytics Software Services helped Mueller upgrade to IBM Cognos 10. With the new version in place, Mueller has started using a new feature called Business Insight to empower regional sales managers to track and improve the performance of their sales teams by creating their own personalized dashboards.

"Static reports are a good starting point, but people don't enjoy reading through pages of data to find the information they need," comments Mark Lack. "The new version of Cognos gives us the ability to create customized interactive dashboards that give each user immediate insight into their own specific area of the business, and enable them to drill down into the raw data if they need to. It's a much more intuitive and compelling way of using information."

Mueller now uses Cognos to investigate the reasons why some products sell better in certain areas, which of its products have the highest adoption rates, and which have the biggest margins. Using these insights, the company can adapt its strategy to ensure that it markets the right products to the right customers—increasing sales.

By using IBM SPSS® Modeler to mine enormous volumes of transactional data, the company aims to reveal patterns and trends that will help to predict future risks and opportunities, as well as uncover unseen problems and anomalies in its current operations. One initial project with IBM SPSS Modeler aims to help Mueller find ways to reduce its fuel costs. Using SPSS Modeler, the company is building a sophisticated statistical model that will automate the process of analyzing fuel transactions for hundreds of vehicles, drivers and routes.

"With SPSS Modeler, we will be able to determine the average fuel consumption for each vehicle on each route over the course of a week," says Mark Lack. "SPSS will automatically flag up any deviations from the average consumption, and we then drill down to find the root cause. The IBM solution helps us to determine if higher-than-usual fuel transactions are legitimate—for example, a driver covering extra miles—or the result of some other factor, such as fraud."

Key Terms

Area chart
Bar chart
Bubble chart
Column chart
Contingency table
Cross-tabulation
Cumulative relative frequency
Cumulative relative frequency distribution
Dashboard
Data profile (fractile)
Data visualization
Descriptive statistics
Doughnut chart
Frequency distribution
Histogram
*k*th percentile
Line chart
Ogive
Pareto analysis
Pie chart
PivotChart
PivotTables
Quartile
Radar chart
Relative frequency
Relative frequency distribution
Scatter chart
Slicers
Sparklines
Statistic
Statistics
Stock chart
Surface chart

Problems and Exercises

1. Create a line chart for the closing prices for all years, and a stock chart for the high/low/close prices for August 2013 in the Excel file *S&P 500*.
2. The Excel file *Traveler* contains the months of a year and the number of travelers that arrive by flight in the morning (AM) and the evening (PM). Prepare a line chart showing the number of AM and PM travelers for each month.
3. The Excel file *Facebook Survey* provides data gathered from a sample of college students. Create a scatter diagram showing the relationship between Hours online/week and Friends.
4. The Excel file *Sales* contain list of the products in different regions. Sort the list of products in ascending order of the sales volume in Asia. Arrange the regions (from left to right) in ascending order for the sales volume of Product 5 and determine which region has the highest sales.
5. Create a bubble chart for the first five colleges in the Excel file *Colleges and Universities* for which the *x*-axis is the Top 10% HS, *y*-axis is Acceptance Rate, and bubbles represent the Expenditures per Student.
6. The Excel file *Expenditure* shows the spending of a country on various sports during a particular year. Create a pie chart and determine the percentage of total spending on tennis.
7. The Excel file *Internet Usage* provides data about users of the Internet. Construct stacked bar charts that will allow you to compare any differences due to age or educational attainment and draw any conclusions that you can. Would another type of charts be more appropriate?
8. The Excel file *McDonald's* contains the monthly sales data of their burgers in a year. Construct the histogram and predict which type of burger has the highest sale.
9. In the Excel file *Banking Data*, apply the following data visualization tools:
   a. Use data bars to visualize the relative values of Median Home Value.
   b. Use color scales to visualize the relative values of Median Household Wealth.
   c. Use an icon set to show high, medium, and low bank balances, where high is above $30,000, low is below $10,000, and medium is anywhere in between.
10. Apply three different colors of data bars to lunch, dinner, and delivery sales in the Excel file *Restaurant Sales* to visualize the relative amounts of sales. Then sort the data (hint: use a custom sort) by the day of the week beginning on Sunday. Compare the nonsorted data with the sorted data as to the information content of the visualizations.
11. For the *Store and Regional Sales* database, apply a four-traffic light icon set to visualize the distribution of the number of units sold for each store, where green corresponds to at least 30 units sold, yellow to at least 20 but less than 30, red to at least 10 but less than 20, and black to below 10.
12. For the Excel file *Closing Stock Prices*,
    a. Apply both column and line sparklines to visualize the trends in the prices for each of the four stocks in the file.
    b. Compute the daily change in the Dow Jones index and apply a win/loss sparkline to visualize the daily up or down movement in the index.
13. Convert the *Store and Regional Sales* database to an Excel table. Use the techniques described in Example 3.11 to find:
    a. the total number of units sold
    b. the total number of units sold in the South region
    c. the total number of units sold in December
14. Convert the *Purchase Orders* database to an Excel table. Use the techniques described in Example 3.11 to find:
    a. the total cost of all orders
    b. the total quantity of airframe fasteners purchased
    c. the total cost of all orders placed with Manley Valve.
15. The Excel file *Economic Poll* provides some demographic and opinion data on whether the economy is moving in the right direction. Convert this data into an Excel table, and filter the respondents who are homeowners and perceive that the economy is not moving in the right direction. What is the distribution of their political party affiliations?

16. The total runs scored by 30 players in a test cricket match in the year 2011 were recorded to determine which score was the highest and which the lowest. The runs are:

423, 369, 387, 411, 393, 394, 371, 377, 389, 409, 392, 408, 431, 401, 363, 391, 405, 382, 400, 381, 399, 415, 428, 422, 396, 372, 410, 419, 386, 390

Construct the frequency distribution table and calculate relative frequency.

17. Sort the data in the Excel file *Automobile Quality* from highest to lowest number of problems per 100 vehicles using the sort capability in Excel.

18. In the *Purchase Orders* database, conduct a Pareto analysis of the Cost per order data. What conclusions can you reach?

19. Use Excel's filtering capability to (1) extract all orders for control panels, (2) all orders for quantities of less than 500 units, and (3) all orders for control panels with quantities of less than 500 units in the *Purchase Orders* database.

20. In the *Sales Transactions* database, use Excel's filtering capability to extract all orders that used PayPal, all orders under $100, and all orders that were over $100 and used a credit card.

21. The Excel file *Credit Risk Data* provides information about bank customers who had applied for loans.[6] The data include the purpose of the loan, checking and savings account balances, number of months as a customer of the bank, months employed, gender, marital status, age, housing status and number of years at current residence, job type, and credit-risk classification by the bank.

a. Compute the combined checking and savings account balance for each record in the database. Then sort the records by the number of months as a customer of the bank. From examining the data, does it appear that customers with a longer association with the bank have more assets? Construct a scatter chart to validate your conclusions.

b. Apply Pareto analysis to draw conclusions about the combined amount of money in checking and savings accounts.

c. Use Excel's filtering capability to extract all records for new-car loans. Construct a pie chart showing the marital status associated with these loans.

d. Use Excel's filtering capability to extract all records for individuals employed less than 12 months. Can you draw any conclusions about the credit risk associated with these individuals?

22. The Excel sheet *Engagement* contains the number of rings sold each day of the week in a jewelry store chain in different cities across India. Use sparklines to summarize the data.

23. Use the *Histogram* tool to construct a frequency distribution of lunch sales amounts in the *Restaurant Sales* database.

24. A community health-status survey obtained the following demographic information from the respondents:

Age	Frequency
18 to 29	297
30 to 45	743
46 to 64	602
65 +	369

Compute the relative frequency and cumulative relative frequency of the age groups.

25. Construct frequency distributions and histograms for the numerical data in the Excel file *Cell Phone Survey*. Also, compute the relative frequencies and cumulative relative frequencies.

26. Use the *Histogram* tool to develop a frequency distribution and histogram with six bins for the age of individuals in the Excel file *Credit Risk Data*. Compute the relative and cumulative relative frequencies and use a line chart to construct an ogive.

27. Use the *Histogram* tool to develop a frequency distribution and histogram for the number of months as a customer of the bank in the Excel file *Credit Risk Data*. Use your judgment for determining the number of bins to use. Compute the relative and cumulative relative frequencies and use a line chart to construct an ogive.

28. Construct frequency distributions and histograms using the Excel *Histogram* tool for the Gross Sales and Gross Profit data in the Excel file *Sales Data*. First let Excel automatically determine the number of bins

[6]Based on Efraim Turban, Ramesh Sharda, Dursun Delen, and David King, *Business Intelligence: A Managerial Approach*, 2nd ed. (Upper Saddle River, NJ: Prentice Hall, 2011).

and bin ranges. Then determine a more appropriate set of bins and rerun the *Histogram* tool.

29. The Excel sheet *Sampling* contains the responses on a scale of 1 to 5 from consumers regarding a product. Construct a cluttered pivot table, and show the sampling data in the histogram.

30. Find the 20th and 80th percentiles of home prices in the Excel file *Home Market Value.*

31. Find the 10th and 90th percentiles and 1st, 2nd, and 3rd quartiles for the combined amounts of checking and savings accounts in the Excel file *Credit Risk Data.*

32. Construct cross-tabulations of Gender versus Carrier and Type versus Usage in the Excel file *Cell Phone Survey*. What might you conclude from this analysis?

33. Using the data in the Excel sheet *Hardware Store*, construct a pivot table and calculate the percentage of sales , the total revenue generated in the month of March and the percentage of sales for the month of August.

34. Use PivotTables to construct a cross-tabulation for marital status and housing type in the Excel file *Credit Risk Data*. Illustrate the results on a PivotChart.

35. Create a PivotTable to find the average amount of travel expenses for each sales representative in the Excel file Travel Expenses. Illustrate your results with a PivotChart.

36. Use PivotTables to find the number of loans by different purposes, marital status, and credit risk in the Excel file *Credit Risk Data*. Illustrate the results on a PivotChart.

37. Use PivotTables to find the number of sales transactions by product and region, total amount of revenue by region, and total revenue by region and product in the *Sales Transactions* database.

38. Create a PivotTable for the data in the Excel file *Weddings* to analyze the wedding cost by type of payor and value rating. What conclusions do you reach?

39. The Excel File *Rin's Gym* provides sample data on member body characteristics and gym activity. Create PivotTables to find:

a. a cross-tabulation of gender and body type versus BMI classification

b. average running times, run distance, weight lifting days, lifting session times, and time spent in the gym by gender.

Summarize your conclusions.

40. Create useful dashboards for each of the following databases. Use appropriate charts and layouts (for example, Explain why you chose the elements of the dashboards and how a manager might use them.

a. *President's Inn*

b. *Restaurant Sales*

c. *Store and Regional Sales*

d. *Peoples Choice Bank*

41. A marketing researcher surveyed 92 individuals, asking them if they liked a new product concept or not. The results are shown below:

	Yes	No
Male	30	50
Female	6	6

Convert the data into percentages. Then construct a chart of the counts and a chart of the percentages. Discuss what each conveys visually and how the different charts may lead to different interpretations of the data.

Case: Drout Advertising Research Project

The background for this case was introduced in Chapter 1. For this part of the case, use appropriate charts to visualize the data. Summarize the data using frequency distributions and histograms for numerical variables, cross-tabulations, and other appropriate applications of PivotTables to break down the data and develop useful insights. Add your findings to the report you started for the case in Chapter 1.