Chapter 41

Summarizing data by using histograms and Pareto charts

Questions answered in this chapter:

- People often say that a picture is worth a thousand words. Can I use Excel to create a picture (called a *histogram*) that summarizes the values in a data set?
- What are some common shapes of histograms?
- What can I learn by comparing histograms from different data sets?
- How do I create a Pareto chart?

The ability to summarize a large data set is important. The three tools used most often to summarize data in Microsoft Excel are histograms, descriptive statistics, and PivotTables. In this chapter, I discuss the use of histograms for summarizing data. Prior to Excel 2016, you used the Data Analysis add-in to create histograms. The histograms created by means of the Data Analysis add-in did not update when new data was added. Excel 2016 provides new capabilities to create beautiful histograms that automatically update to include new data. I cover descriptive statistics in Chapter 42, "Summarizing data by using descriptive statistics," and PivotTables in Chapter 43, "Using PivotTables and slicers to describe data."

Answers to this chapter's questions

People often say that a picture is worth a thousand words. Can I use Excel to create a picture (called a *histogram*) that summarizes the values in a data set?

A histogram is a commonly used tool to summarize data. Essentially, a histogram tells you how many *observations* (another term for data points) fall in various ranges of values. For example, a histogram created from monthly Cisco stock returns might show how many monthly returns Cisco had from 0 percent through 10 percent, 11 percent through 20 percent, and so on. The ranges in which you group data are referred to as *bin ranges*.

Let's look at how to construct and interpret histograms that summarize the values of monthly returns for Cisco and GM stock in the years 1990–2000. You'll find this data (and returns for other stocks) in the file Stock.xlsx. Figure 41-1 shows a subset of the data (in the Stockprices worksheet). During March 1990 (row 52), for example, Cisco stock increased in value by 1.075 percent.

© 2016 Microsoft Corporation. All Rights Reserved.

	B	C	D	E	F
48					
49			min	-0.24032043	-0.2025
50			max	0.276619107	0.33898
51	Microsoft	GE	Intel	GM	CSCO
52	0.121518984	0.040485829	0.037267081	0.022284122	0.01075
53	0.047404062	-0.00389105	-0.05389221	-0.03542234	0.01064
54	0.258620679	0.083515622	0.221518993	0.115819208	0.04211
55	0.04109589	0.005444646	-0.02590674	-0.02056555	0.07071
56	-0.125	0.034296028	-0.05319149	-0.02099738	-0.0377
57	-0.07518797	-0.13438046	-0.25	-0.1313673	-0.0294
58	0.024390243	-0.11338709	-0.00374532	-0.08805031	-0.0909
59	0.011904762	-0.04587156	0.007518797	0.013793103	0.31111
60	0.13333334	0.052884616	0.119402982	0.013605442	0.33898
61	0.041522492	0.057260275	0.026666667	-0.05821918	0.13608
62	0.303986698	0.115468413	0.188311681	0.054545455	0.30362
63	0.057324842	0.070468754	0.043715846	0.100689657	-0.0427
64	0.022891566	0.023897059	-0.02094241	-0.0443038	-0.1295
65	-0.06713781	0.016157989	0.053475935	-0.05298013	0.22051
66	0.108585857	0.09908127	0.131979689	0.217482507	0.08403
67	-0.06890661	-0.0420712	-0.16591929	-0.05507246	-0.0543
68	0.078899086	-0.01013514	0.010752688	-0.02453988	0.28689
69	0.159863949	0.022184301	0.053191491	-0.03396227	0.15605
70	0.043988269	-0.06664441	-0.14646465	-0.01644737	-0.0964

F41xx01

FIGURE 41-1 Monthly stock returns.

F41xx01: This figure shows monthly stock returns during the 1990s.

When constructing histograms with Excel, you can let Excel define the bin ranges or you can define the bin ranges yourself. If Excel defines the bin ranges, you could end up with weird-looking bin ranges, such as –12.53 percent to 4.52 percent. For this reason, I prefer to define the ranges myself.

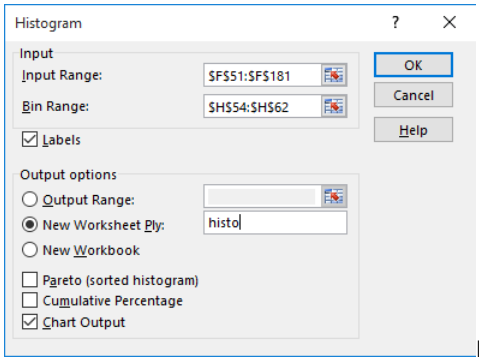
A good way to start defining bin ranges for a histogram (you can think of defining bin ranges as setting boundaries) is to divide the range of values (between the smallest and largest) into 8 to 15 equally spaced categories. All the monthly returns for Cisco are from –30 percent through 40 percent, so I chose bin range boundaries of –30 percent, –20 percent, –10 percent, 0 percent, and so on up to 40 percent.

To create bin ranges, I first enter **CSCO, 0.4, 0.3, 0.2,…, –0.2, –0.3** (the boundaries of the bin ranges) in cells H54:H62. Next, on the Data tab, in the Analysis group, I click Data Analysis to open the Data Analysis dialog box. The dialog box lists the functions of the Analysis ToolPak, which contains many of the statistical capabilities in Excel.

> **Note** If the Data Analysis command doesn't appear on the Data tab, click the File tab, click Options, and then click Add-Ins in the left pane. In the Manage box, click Excel Add-Ins, and then click Go. In the Add-Ins dialog box, select Analysis ToolPak (the first choice, not Analysis ToolPak - VBA), and then click OK. Now you can access the Analysis ToolPak functions by clicking Data Analysis in the Analysis group on the Data tab.

By clicking Histogram in the Data Analysis dialog box (and then clicking OK), you open the Histogram dialog box shown in Figure 41-2.

© 2016 Microsoft Corporation. All Rights Reserved.



F41xx02

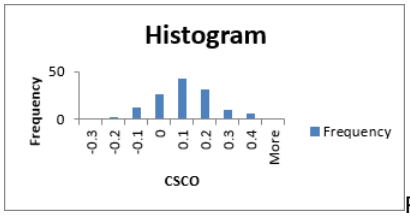
FIGURE 41-2 The Histogram dialog box for the Cisco histogram.

F41xx02: This figure shows the settings to create a histogram for Cisco monthly returns.

Here's how to fill in the dialog box as it's shown:

- Select the input range (F51:F181). (To select the range F51:F181, you can select cell F51 and then press Ctrl+Shift+Down arrow. This takes you to the bottom of the column.) This range includes all the data you want to use to create the histogram. I included the label *CSCO* from cell F51 because when you do not include a label in the first row, the x-axis of the histogram is often labeled with a number, which can be confusing.
- Select the bin range (H54:H62), which includes the boundaries of the bin ranges. Excel creates bins of –30 percent through –20 percent, –20 percent through –10 percent, and so on up to 30–40 percent.
- Check the Labels option because the first rows of both the bin range and the input range contain labels.
- In the Output Options, select New Worksheet Ply to create the histogram in a new worksheet (named *histo*).
- Select Chart Output, or Excel will not create a histogram.

Click OK in the Histogram dialog box. The Cisco histogram will look like the one shown in Figure 41-3.



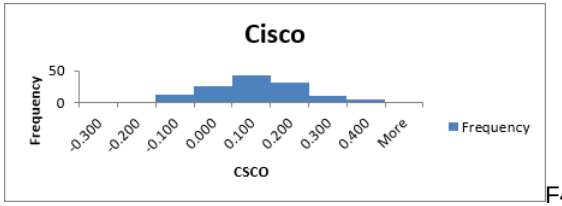
F41xx03

FIGURE 41-3 The Cisco histogram created by using an Excel Analysis ToolPak function.

© 2016 Microsoft Corporation. All Rights Reserved.

F41xx03: This figure shows the Cisco histogram with gaps.

When you create the histogram, you'll see gaps between the bars. To remove these gaps, right-click any bar on the graph and choose Format Data Series. On the Format Data Series pane, in the Series Options section, drag Gap Width to 0%. You might also see that a label does not appear for each bar. If all the labels do not appear, widen the graph by selecting it and dragging a side handle (circle shape), where your cursor changes to two arrows. You can also reduce the font size to make a hidden label appear. To reduce the font size, right-click the graph axis (the text you want to change), and then click Font. In the Font dialog box, change the font size to 5, and click OK. You can also change the title of the chart by selecting the text and entering the title you want. After I made some of these changes, the histogram appears as its shown in Figure 41-4.

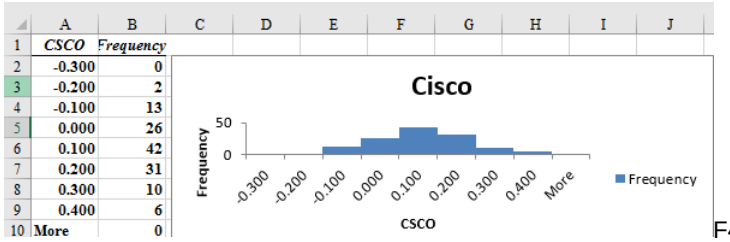


F41xx04

FIGURE 41-4 Changing the format of different elements in the chart.

F41xx04: This figure shows the histogram with no gaps and a smaller font size.

Notice that Cisco returns are most likely between 0 and 10 percent per month, and the height of the bars drops off as the graph moves away from the tallest bar. When you create the histogram, you also obtain the bin-range frequency summary shown in Figure 41-5.



F41xx05

FIGURE 41-5 The Cisco bin-range frequencies.

F41xx05: This figure shows the bin-range count and finished histogram.

From the bin-range frequencies, you can learn, for example, that for two months, Cisco's return was greater than –30 percent and less than or equal to –20 percent; for 13 months, the monthly return was greater than –20 percent and less than or equal to –10 percent.

If you add new Cisco monthly returns or even if you just modify the existing monthly returns, your histogram will not change unless you rerun the Data Analysis histogram procedure.

Excel 2016 provides an easy option to create better-looking histograms that automatically update

© 2016 Microsoft Corporation. All Rights Reserved.

with new data. To illustrate the use of the new Excel 2016 histogram chart, please open the file Iqtemp.xlsx (in this chapter's Templates folder), which contains a sample of 1,083 sixth-grade students' IQs. (See Figure 41-6.)

	E
4	IQ
5	95
6	105
7	93
8	103
9	103
10	129
11	95
12	98
13	94
14	106
15	102
16	96
17	112
18	106
19	106

F41xx06

FIGURE 41-6 The IQ data.

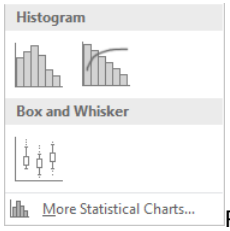
F41xx06: The IQ data used for creating an Excel 2016 histogram.

After selecting the range E4:E1177, use Ctrl+T to set up the new data as a table (leave My Table Has Headers checked in the Create Table dialog box). This ensures that as we add new data, our histogram changes. Next, select all the data (including cell E4), and then, on the Insert tab, in the Charts group, click the drop-down arrow for the Insert Statistic Chart option (see Figure 41-7), and then select Histogram, as shown in Figure 41-8.

F41xx07

FIGURE 41-7 The Insert Statistic Chart icon.

F41xx07: This figure shows the Insert Statistical Chart icon



F41xx08

FIGURE 41-8 The statistical chart options.

© 2016 Microsoft Corporation. All Rights Reserved.

F41xx08 This figure shows the new statistical charts included in Excel 2016.

You now obtain a handsome histogram. From the Design tab, you can choose from several options that change the histogram's appearance. In the Chart Styles group, I made the selection (third from the left) that shows how many data points fall in each range. (See Figure 41-9.)

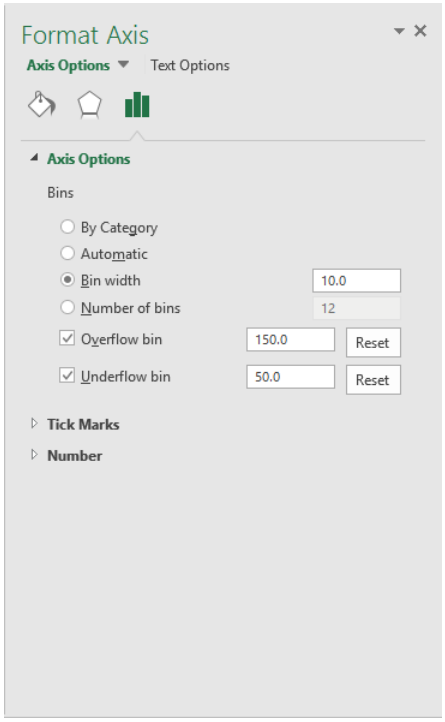
F41xx09

FIGURE 41-9 The IQ histogram.

F41xx09: This figure shows a histogram of IQs.

By right-clicking on the axis and choosing Format Axis, you can change (as shown in Figure 41-10) the definition of the bin ranges and set the lower limit for the first bin and upper limit for the last bin.

© 2016 Microsoft Corporation. All Rights Reserved.



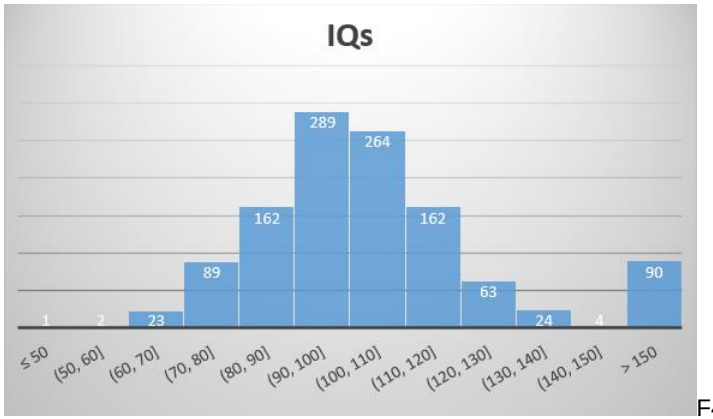
F41xx10

FIGURE 41-10 Changing the bin ranges.

F41xx10: This figure shows changes in the definition of the bin ranges.

I selected Underflow Bin and set the lower boundary of the first bin at 50. For Overflow Bin, I set the upper boundary of the last bin at 150. I also selected Bin Width and set the width of each bin equal to 10 (you might need to scroll to the right). By clicking Number (to expand that section), I could have changed the format of the axis (to Currency, for instance, for monetary data.) After making these changes, the histogram appears as shown in Figure 41-11.

© 2016 Microsoft Corporation. All Rights Reserved.



F41xx11

FIGURE 41-11 The histogram with updated bin ranges.

F41xx11: This figure shows the histogram with updated bin ranges.

We find, for example, that 90 students have IQs greater than 150. A very smart group indeed! Note that if you add more data (say 20 people with IQs of 55) you will find that the histogram automatically updates to include the new data.

What are some common shapes of histograms?

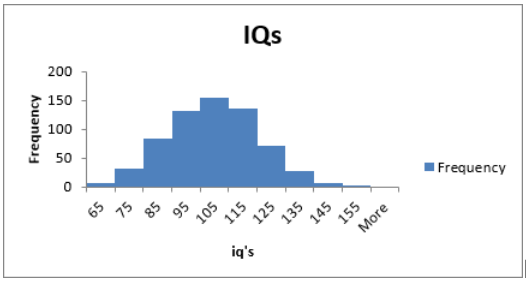
For most data sets, a histogram created from the data will be classified as one of the following:

- Symmetric
- Skewed right (positively skewed)
- Skewed left (negatively skewed)
- Multiple peaks

Let's look at each type in more detail. See the file Skewexamples.xlsx.

- **Symmetric distribution** A histogram is *symmetric* if it has a single peak and looks approximately the same to the left of the peak as to the right of the peak. Test scores (such as IQ tests) are often symmetric. For example, the histograms of IQs (see cell Z42) might look like Figure 41-12. Notice that the height of the bars one bar away from the peak bar are approximately the same, the height of the bars two bars away from the peak bar are approximately the same, and so on. The bar labeled *105* represents all people with an IQ greater than 95 and less than or equal to 105, the bar labeled *65* represents all people having an IQ less than or equal to 65, and so on. Also note that the Cisco monthly returns are approximately symmetric.

© 2016 Microsoft Corporation. All Rights Reserved.

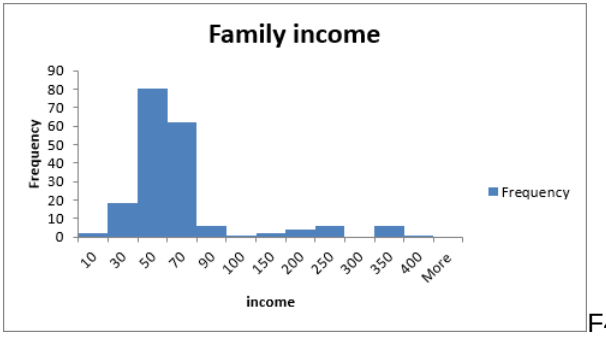


F41xx12

FIGURE 41-12 A symmetric histogram.

F41xx12: This figure shows a symmetric histogram of IQs.

- **Skewed right** (positively skewed) A histogram is *skewed right* (positively skewed) if it has a single peak and the values of the data set extend much farther to the right of the peak than to the left of the peak. Many economic data sets (such as family or individual income) exhibit a positive skew. Figure 41-13 (see cell T24) shows an example of a positively skewed histogram created from a sample of family incomes.



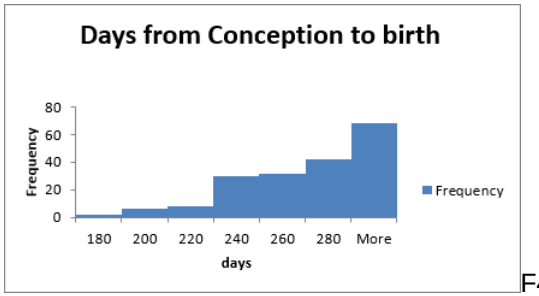
F41xx13

FIGURE 41-13 A positively skewed histogram created from family-income data.

F41xx13: This figure shows a positively skewed histogram for family income.

- **Skewed left** (negatively skewed) A histogram is *skewed left* (negatively skewed) if it has a single peak and the values of the data set extend much farther to the left of the peak than to the right of the peak. Days from conception to birth are negatively skewed. An example is shown in cell Q7 and Figure 41-14. The height of each bar represents the number of pregnant women whose time from conception to birth fell in the given bin range. For example, two women gave birth fewer than 180 days after conception.

© 2016 Microsoft Corporation. All Rights Reserved.

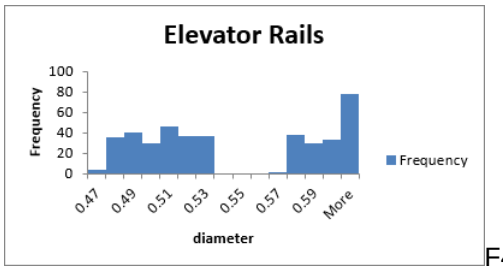


F41xx14

FIGURE 41-14 A negatively skewed histogram of data plotting days from conception to birth.

F41xx14: This figure shows a negatively skewed histogram of days from conception to birth.

- **Multiple peaks** When a histogram exhibits *multiple peaks*, it usually means that data from two or more populations are being graphed together. For example, suppose the diameter of elevator rails produced by two machines yields the histogram shown in Figure 41-15. (See cell Q11 in the file Twinpeaks.xlsx.)



F41xx15

FIGURE 41-15 A multiple-peak histogram.

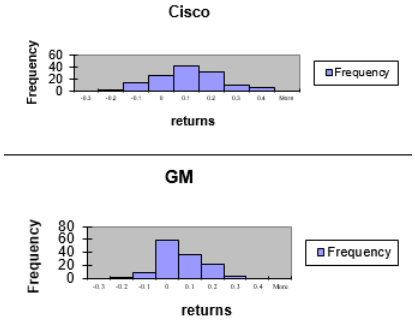
F41xx15: This figure shows a histogram with two peaks.

In this histogram, the data clusters into two separate groups. In all likelihood, each group of data corresponds to the elevator rails produced by one of the machines. If you assume that the diameter you want for an elevator rail is 0.55 inches, you can conclude that one machine is producing elevator rails that are too narrow, whereas the other machine is producing elevator rails that are too wide. You should follow up with your interpretation of this histogram by constructing a histogram that charts the elevator rails produced by each machine. This example shows why histograms are a powerful tool in quality control.

What can I learn by comparing histograms from different data sets?

Analysts are often asked to compare different data sets. For example, you might be asked how the monthly returns on GM and Cisco stock differ. To answer a question such as this, you can construct a histogram for GM by using the same bin ranges as for Cisco and then place one histogram above the other, as shown in Figure 41-16. See the Histograms worksheet of file Stock.xlsx.

© 2016 Microsoft Corporation. All Rights Reserved.



F41xx16

FIGURE 41-16 Using histograms that include the same bin ranges to compare different data sets.

F41xx16: This figure shows a comparison of Cisco and GM histograms.

By comparing these two histograms, you can draw two important conclusions:

- Typically, Cisco performed better than GM. You know this because the highest bar for Cisco is one bar to the right of the highest bar for GM. Also, the Cisco bars extend farther to the right than the GM bars.
- Cisco had more *variability*, or spread about the mean, than GM. Note that GM's peak bar contains 59 months, whereas Cisco's peak bar contains only 41 months. This shows that for Cisco, more of the returns are outside the bin that represents the most likely Cisco return. Cisco returns are more spread out than GM returns.

In Chapter 42, we'll use descriptive statistics and Boxplots to look at more details about the differences between the monthly returns on Cisco and GM.

How do I create a Pareto chart?

A *Pareto chart* is a type of chart that contains bars and line graphs. Individual values are portrayed in descending order by bars, and the cumulative total is represented by the line. Pareto charts are often used to illustrate the famous *80-20 rule* that was first discovered by the great Italian economist Vilfredo Pareto (1848–1923). The *Pareto rule* emphasizes the importance of few items in explaining a total. For example:

- 20 percent of products generate 80 percent of profits.
- 20 percent of people have 80 percent of the income.
- 80 percent of all technical support calls result from 20 percent of all possible problems.
- 20 percent of all websites get 80 percent of the hits.

To illustrate the creation of a Pareto chart with Excel 2016, open up the file Paretotemp.xlsx (from this chapter's Templates folder), which gives the revenue from each of a company's 100 products.

© 2016 Microsoft Corporation. All Rights Reserved.

(See Figure 41-17.)

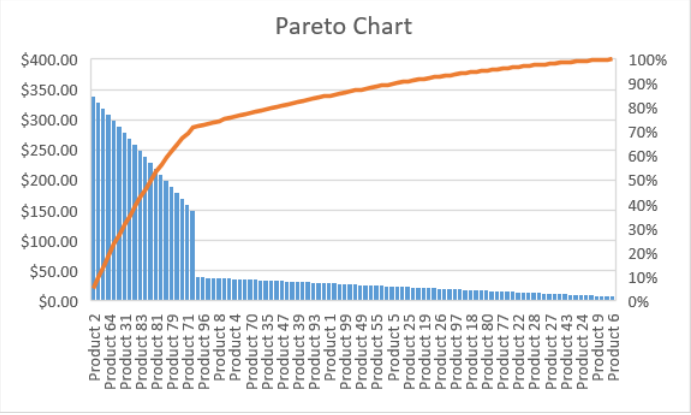
	E	F
3	**Product**	**Revenue**
4	**Product 1**	**$30.00**
5	**Product 2**	**$340.00**
6	**Product 3**	**$11.60**
7	**Product 4**	**$37.20**
8	**Product 5**	**$25.20**
9	**Product 6**	**$8.40**
10	**Product 7**	**$38.00**
11	**Product 8**	**$38.40**
12	**Product 9**	**$9.60**
13	**Product 10**	**$29.20**
14	**Product 11**	**$14.80**
15	**Product 12**	**$10.00**
16	**Product 13**	**$22.40**
17	**Product 14**	**$29.60**

F41xx17

FIGURE 41-17 Data for a Pareto chart.

F41xx17: The data in this figure will be used to create a Pareto chart.

After selecting the data (cell range E3:F103), I chose the Insert Statistic Chart icon from the Insert tab (in the Charts group) and then chose Pareto, the second Histogram option. (See Figure 41-8, earlier) We obtain the Pareto chart shown in Figure 41-18. The products are now listed in order of descending sales. The line represents the cumulative percentage of sales generated by the products. We observe that our 10 best-selling products generate about 80 percent of the sales. Of course, if we made the source data a table, then new data would automatically be incorporated in the chart.



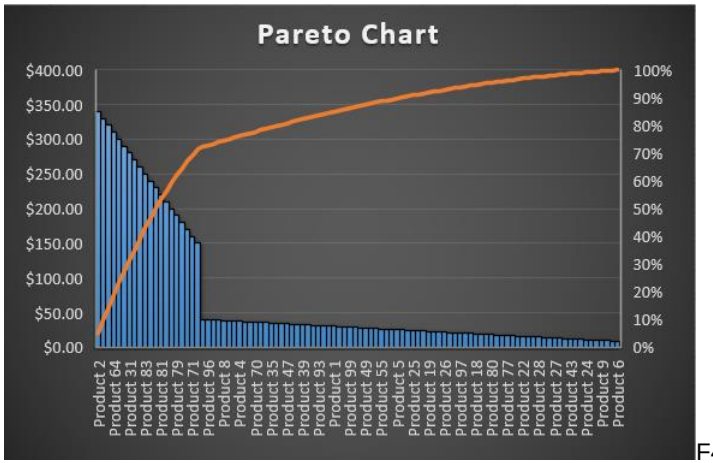
F41xx18

FIGURE 41-18 A Pareto chart.

F41xx18: This figure shows a Pareto chart.

© 2016 Microsoft Corporation. All Rights Reserved.

After clicking the Pareto chart to select it, I chose the fifth option from the Design Tab (in the Chart Styles group) and obtained the chart shown in Figure 41-19.



F41xx19

FIGURE 41-19 A Pareto chart created from the Design tab.

F41xx19: This figure shows a Pareto chart created from the Design tab.

Problems

1. Use the data in Stock.xlsx to construct histograms for monthly returns for GE and Intel.
2. Use the data in Historicalinvest2009.xlsx to create histograms for annual returns on stocks and bonds. Then compare the annual returns of these stocks and bonds.
3. You are given (in the file Deming.xlsx) the measured diameter (in inches) for 500 rods produced by Rodco, as reported by the production foreman. A rod is considered acceptable if it is at least 1 inch in diameter. In the past, the diameter of the rods produced by Rodco has followed a symmetric histogram. Do the following:
   - Construct a histogram of these measurements.
   - Comment on any unusual aspects of the histogram.

   Can you guess what might have caused any unusual aspects of the histogram? Hint: One of quality-guru Edwards Deming's 14 points is to "Drive out fear."
4. The file Unemployment.xlsx contains monthly US unemployment rates. Create a histogram. Are the unemployment rates symmetric or skewed?
5. The file Teams.xlsx contains runs scored by major league baseball teams during a season. Create a histogram. Are the runs scored symmetric or skewed?

© 2016 Microsoft Corporation. All Rights Reserved.

6. The file NFLpoints.xlsx contains points scored by NFL teams during a season. Create a histogram. Are points scored symmetric or skewed?
7. Using the data in the file Problem7data.xlsx, create a histogram that summarizes the heights of American men.
8. The data in the file Problem8data.xlsx contains the points scored by each Division I NCAA football team during the 2015 season. Create a histogram to summarize this data. Does the data appear to be symmetric?
9. The data in the file Problem9data.xlsx contains income of families in Smalltown, USA. Create a Pareto chart to summarize the family incomes.

© 2016 Microsoft Corporation. All Rights Reserved.