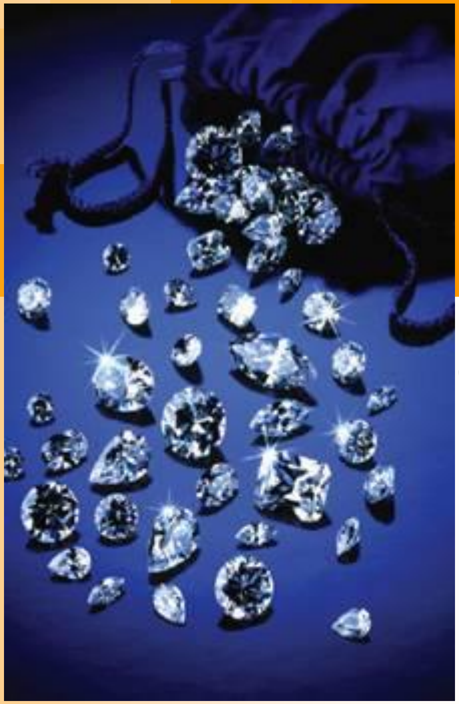
Describing Data: Displaying and Exploring Data



Chapter 4

McGraw-Hill/Irwin

Learning Objectives

LO1 Construct and interpret a *dot plot*.
LO2 Construct and describe a *stem-and-leaf display*.
LO3 Identify and compute measures of position.
LO4 Construct and analyze a *box plot*.
LO5 Compute and describe the *coefficient of skewness*.
LO6 Create and interpret a scatterplot.
LO7 Develop and explain a *contingency table*.

LO1 Construct and interpret a *dot plot*.

Dot Plots

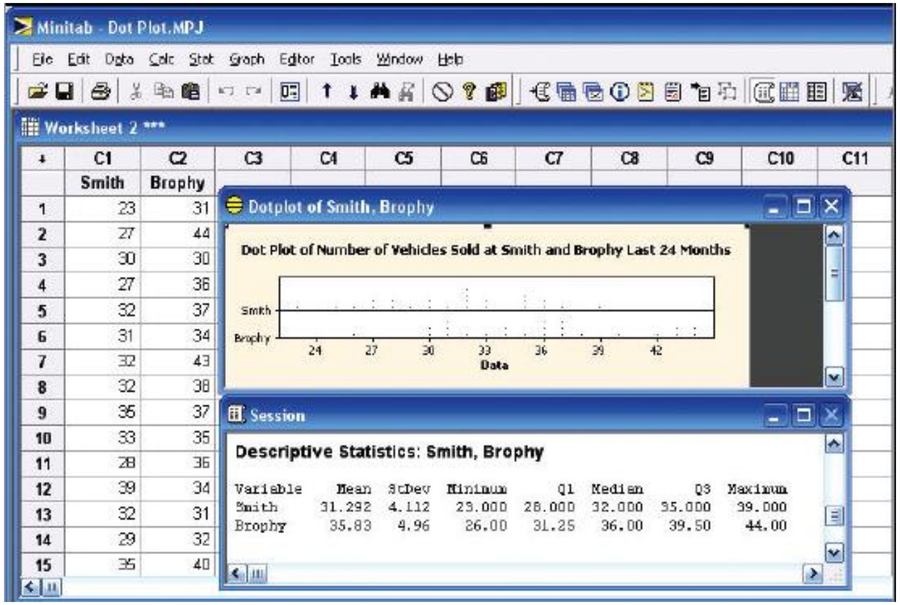
- A **dot plot** groups the data as little as possible and the identity of an individual observation is not lost.
- To develop a dot plot, each observation is simply displayed as a dot along a horizontal number line indicating the possible values of the data.
- If there are identical observations or the observations are too close to be shown individually, the dots are "piled" on top of each other.

EXAMPLE

Reported below are the number of vehicles sold in the last 24 months at Smith Ford Mercury Jeep, Inc., in Kane, Pennsylvania, and Brophy Honda Volkswagen in Greenville, Ohio. Construct dot plots and report summary statistics for the two small-town Auto USA lots.

Smith Ford Mercury Jeep, Inc.									
23	27	30	27	32	31	32	32	35	33
28	39	32	29	35	36	33	25	35	37
26	28	36	30						

Brophy Honda Volkswagen									
31	44	30	36	37	34	43	38	37	35
36	34	31	32	40	36	31	44	26	30
37	43	42	33						


Minitab - Dot Plot.MPJ
File Edit Data Calc Stat Graph Editor Tools Window Help
Worksheet 2 ***
Dotplot of Smith, Brophy
Dot Plot of Number of Vehicles Sold at Smith and Brophy Last 24 Months
Smith
Brophy
24
27
30
33
36
39
42
Data
Session

↓	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
	Smith	Brophy									
1	23	31									
2	27	44									
3	30	30									
4	27	36									
5	32	37									
6	31	34									
7	32	43									
8	32	38									
9	35	37									
10	33	35									
11	28	36									
12	39	34									
13	32	31									
14	29	32									
15	35	40									

```
Descriptive Statistics: Smith, Brophy

Variable    Mean  StDev  Minimum      Q1  Median      Q3  Maximum
Smith     31.292  4.112   23.000  28.000  32.000  35.000   39.000
Brophy     35.83   4.96    26.00   31.25   36.00   39.50    44.00
```

LO2 Construct and interpret a *stem and leaf plot*.

Stem-and-Leaf

- **Stem-and-leaf display** is a statistical technique to present a set of data. Each numerical value is divided into two parts. The leading digit(s) becomes the stem and the trailing digit the leaf. The stems are located along the vertical axis, and the leaf values are stacked against each other along the horizontal axis.
- Two disadvantages to organizing the data into a frequency distribution:
  (1) The exact identity of each value is lost
  (2) Difficult to tell how the values within each class are distributed.

EXAMPLE

Listed in Table 4–1 is the number of 30-second radio advertising spots purchased by each of the 45 members of the Greater Buffalo Automobile Dealers Association last year. Organize the data into a stem-and-leaf display. Around what values do the number of advertising spots tend to cluster? What is the fewest number of spots purchased by a dealer? The largest number purchased?

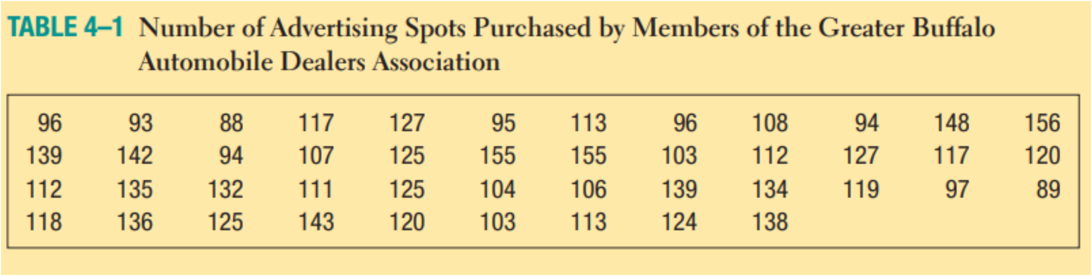
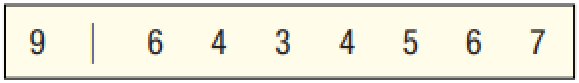
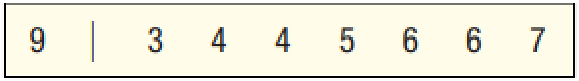


TABLE 4–1 Number of Advertising Spots Purchased by Members of the Greater Buffalo Automobile Dealers Association

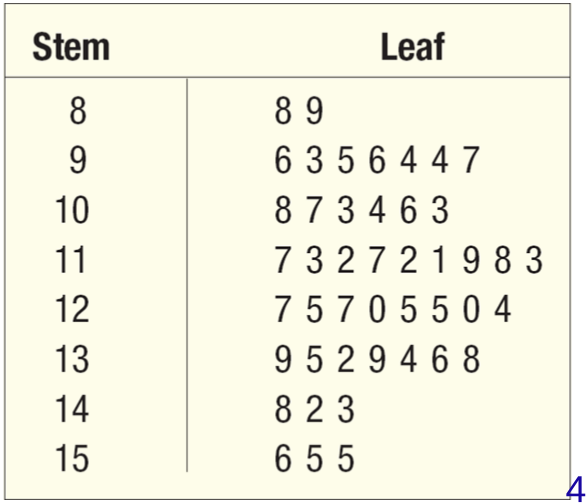
96	93	88	117	127	95	113	96	108	94	148	156
139	142	94	107	125	155	155	103	112	127	117	120
112	135	132	111	125	104	106	139	134	119	97	89
118	136	125	143	120	103	113	124	138			



9	6	4	3	4	5	6	7

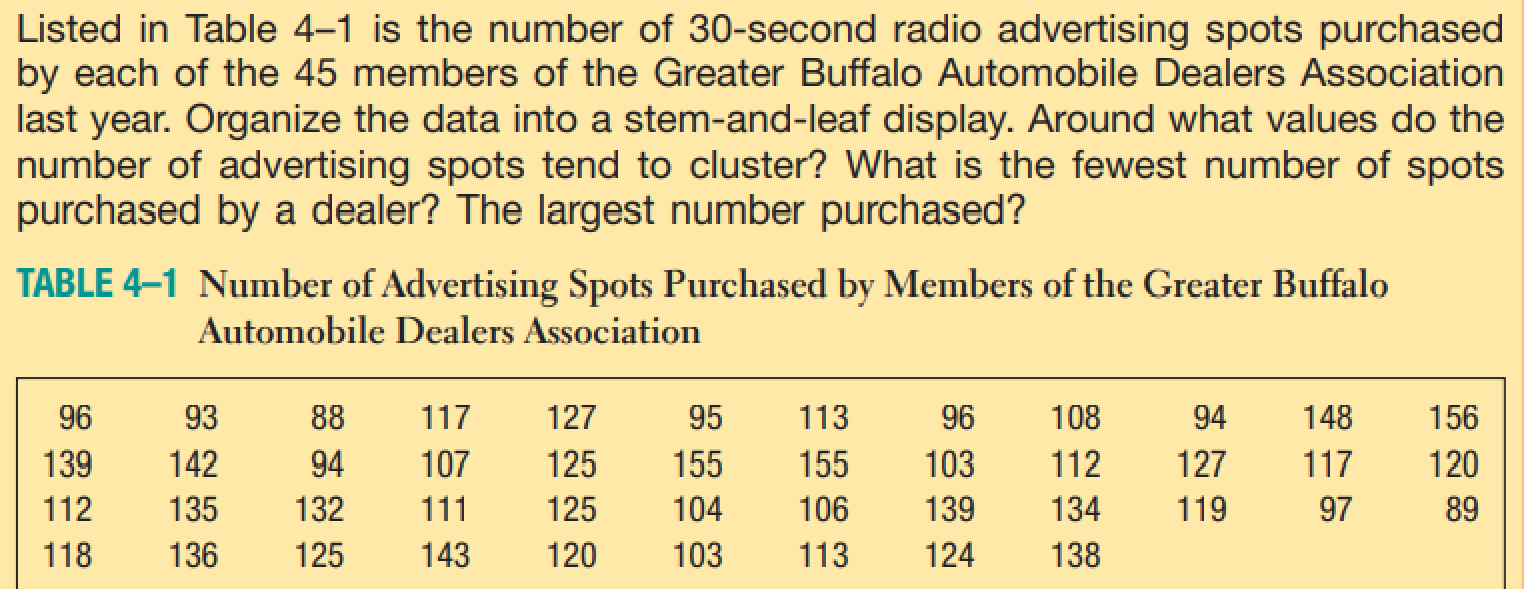


9	3	4	4	5	6	6	7



Stem	Leaf
8	8 9
9	6 3 5 6 4 4 7
10	8 7 3 4 6 3
11	7 3 2 7 2 1 9 8 3
12	7 5 7 0 5 5 0 4
13	9 5 2 9 4 6 8
14	8 2 3
15	6 5 5

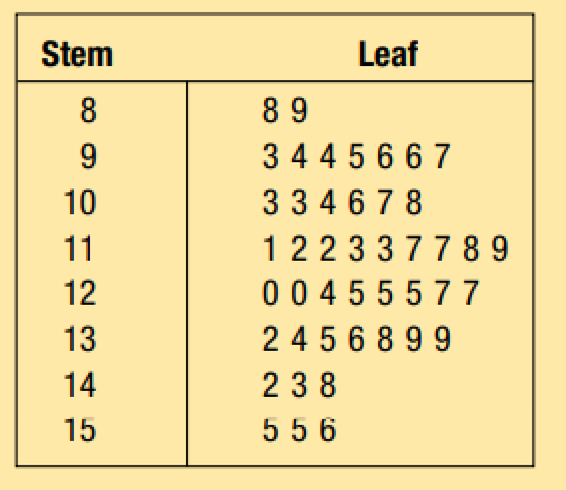
Stem-and-Leaf



Listed in Table 4–1 is the number of 30-second radio advertising spots purchased by each of the 45 members of the Greater Buffalo Automobile Dealers Association last year. Organize the data into a stem-and-leaf display. Around what values do the number of advertising spots tend to cluster? What is the fewest number of spots purchased by a dealer? The largest number purchased?

TABLE 4–1 Number of Advertising Spots Purchased by Members of the Greater Buffalo Automobile Dealers Association

96	93	88	117	127	95	113	96	108	94	148	156
139	142	94	107	125	155	155	103	112	127	117	120
112	135	132	111	125	104	106	139	134	119	97	89
118	136	125	143	120	103	113	124	138			



Stem	Leaf
8	8 9
9	3 4 4 5 6 6 7
10	3 3 4 6 7 8
11	1 2 2 3 3 7 7 8 9
12	0 0 4 5 5 5 7 7
13	2 4 5 6 8 9 9
14	2 3 8
15	5 5 6

LO3 Identify and compute measures of position.

Quartiles, Deciles and Percentiles

- The standard deviation is the most widely used measure of dispersion.
- Alternative ways of describing spread of data include determining the *location* of values that divide a set of observations into equal parts.
- These measures include **quartiles, deciles,** and **percentiles.**
- To formalize the computational procedure, let L_p refer to the location of a desired percentile. So if we wanted to find the 33rd percentile we would use L_{33} and if we wanted the median, the 50th percentile, then L_{50}.

LOCATION OF A PERCENTILE

$$L_p = (n + 1)\frac{P}{100} \qquad \textbf{[4–1]}$$

- The number of observations is *n*, so if we want to locate the median, its position is at (*n* + 1)/2, or we could write this as (*n* + 1)(*P*/100), where *P* is the desired percentile
  - □ n - the number of observations
  - P - desired percentile

LO3

Percentiles - Example

EXAMPLE

Listed below are the commissions earned last month by a sample of 15 brokers at Salomon Smith Barney's Oakland, California, office.

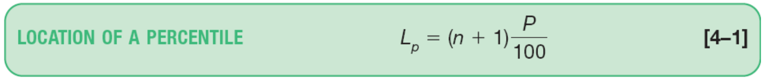
\$2,038 \$1,758 \$1,721 \$1,637 \$2,097 \$2,047 \$2,205 \$1,787
\$2,287 \$1,940 \$2,311 \$2,054 \$2,406 \$1,471 \$1,460

Locate the median, the first quartile, and the third quartile for the commissions earned.

Step 1: Organize the data from lowest to largest value

\$1,460 \$1,471 \$1,637 \$1,721 \$1,758 \$1,787 \$1,940
\$2,038 \$2,047 \$2,054 \$2,097 \$2,205 \$2,287 \$2,311
\$2,406

Step 2: Compute the first and third quartiles. Locate L_{25} and L_{75} using:



LOCATION OF A PERCENTILE $L_p = (n + 1)\frac{P}{100}$ [4–1]

$$L_{25} = (15+1)\frac{25}{100} = 4 \qquad L_{75} = (15+1)\frac{75}{100} = 12$$

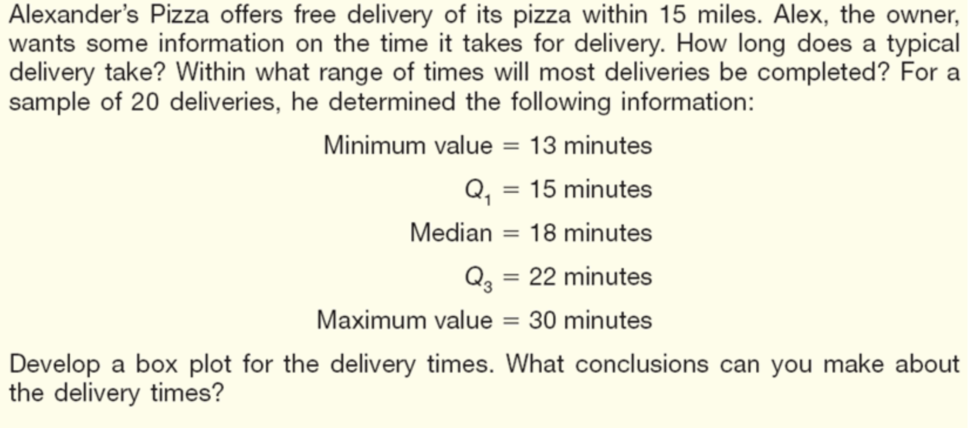
Therefore, the first and third quartiles are located at the 4th and 12th positions, respectively

$$L_{25} = \$1{,}721$$

$$L_{75} = \$2{,}205$$

LO4 Construct and analyze a *box plot*.

Boxplot - Example



Alexander's Pizza offers free delivery of its pizza within 15 miles. Alex, the owner, wants some information on the time it takes for delivery. How long does a typical delivery take? Within what range of times will most deliveries be completed? For a sample of 20 deliveries, he determined the following information:

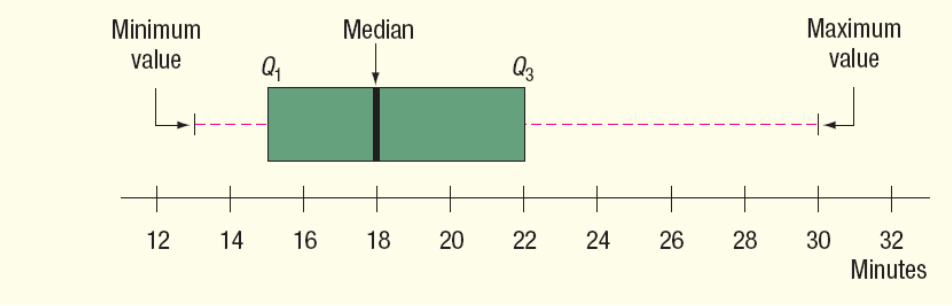
$$\begin{aligned} \text{Minimum value} &= 13 \text{ minutes} \\ Q_1 &= 15 \text{ minutes} \\ \text{Median} &= 18 \text{ minutes} \\ Q_3 &= 22 \text{ minutes} \\ \text{Maximum value} &= 30 \text{ minutes} \end{aligned}$$

Develop a box plot for the delivery times. What conclusions can you make about the delivery times?

Step1: Create an appropriate scale along the horizontal axis.

Step 2: Draw a box that starts at *Q1 (15 minutes) and ends at Q3 (22* minutes). Inside the box we place a vertical line to represent the median (18 minutes).

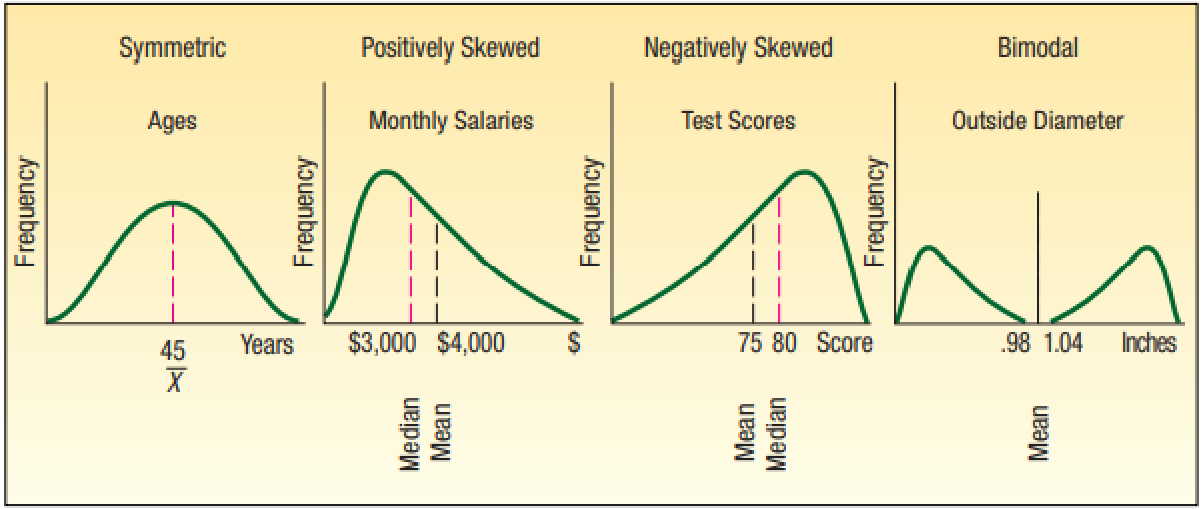
Step 3: Extend horizontal lines from the box out to the minimum value (13 minutes) and the maximum value (30 minutes).



LO5 Compute and understand the *coefficient of skewness*.

Skewness

- Another characteristic of a set of data is the shape.
- There are four shapes commonly observed: **symmetric, positively skewed, negatively skewed, bimodal**.



- The coefficient of skewness can range from -3 up to 3.
  - □ A value near -3, indicates considerable negative skewness.
  - □ A value such as 1.63 indicates moderate positive skewness.
  - □ A value of 0, which will occur when the mean and median are equal, indicates the distribution is symmetrical and that there is no skewness present.

PEARSON'S COEFFICIENT OF SKEWNESS

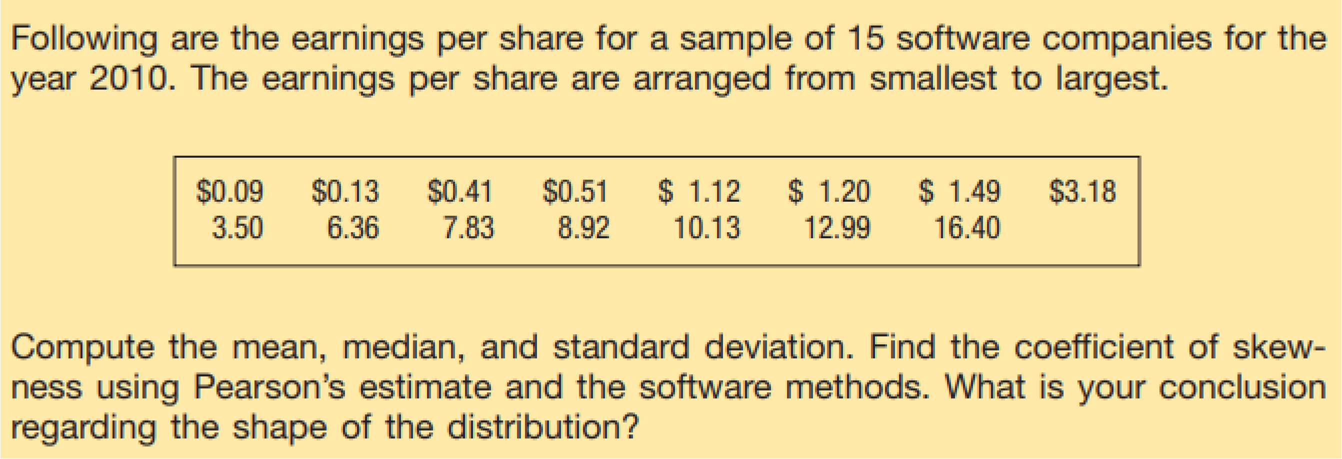
$$sk = \frac{3(\bar{X} - \text{Median})}{s} \qquad [4\text{–}2]$$

SOFTWARE COEFFICIENT OF SKEWNESS

$$sk = \frac{n}{(n-1)(n-2)}\left[\sum\left(\frac{X-\bar{X}}{s}\right)^3\right] \qquad [4\text{–}3]$$

LO5

Skewness – An Example



Following are the earnings per share for a sample of 15 software companies for the year 2010. The earnings per share are arranged from smallest to largest.

$0.09	$0.13	$0.41	$0.51	$ 1.12	$ 1.20	$ 1.49	$3.18
3.50	6.36	7.83	8.92	10.13	12.99	16.40	

Compute the mean, median, and standard deviation. Find the coefficient of skewness using Pearson's estimate and the software methods. What is your conclusion regarding the shape of the distribution?

Step 1: Compute the Mean

$$\overline{X} = \frac{\sum X}{n} = \frac{\$74.26}{15} = \$4.95$$

Step 2 : Compute the Standard Deviation

$$s = \sqrt{\frac{\Sigma\left(X - \overline{X}\right)^2}{n-1}} = \sqrt{\frac{(\$0.09 - \$4.95)^2 + \ldots + (\$16.40 - \$4.95)^2)}{15-1}} = \$5.22$$

Step 3 : Find the Median

The middle value in the set of data, arranged from smallest to largest is 3.18

Step 3 : Compute the Skewness

$$sk = \frac{3(\overline{X} - Median)}{s} = \frac{3(\$4.95 - \$3.18)}{\$5.22} = 1.017$$

LO6 Create and interpret a scatter plot.

Describing Relationship between Two Variables

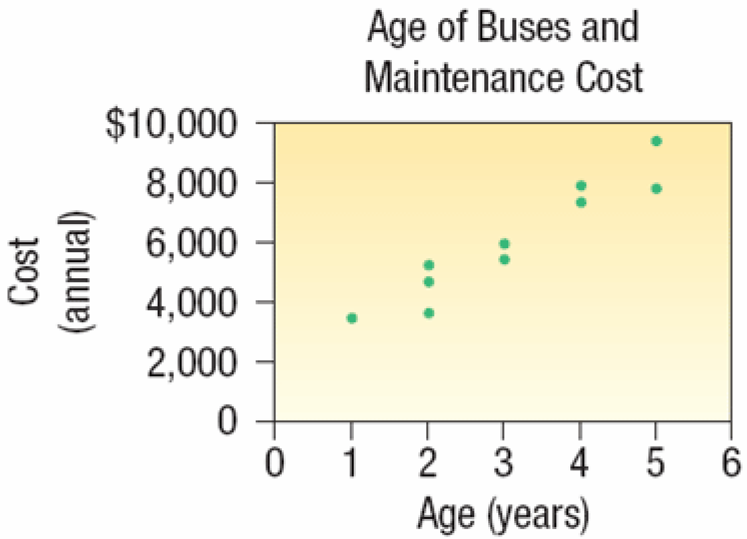
- When we study the relationship between two variables we refer to the data as **bivariate.**

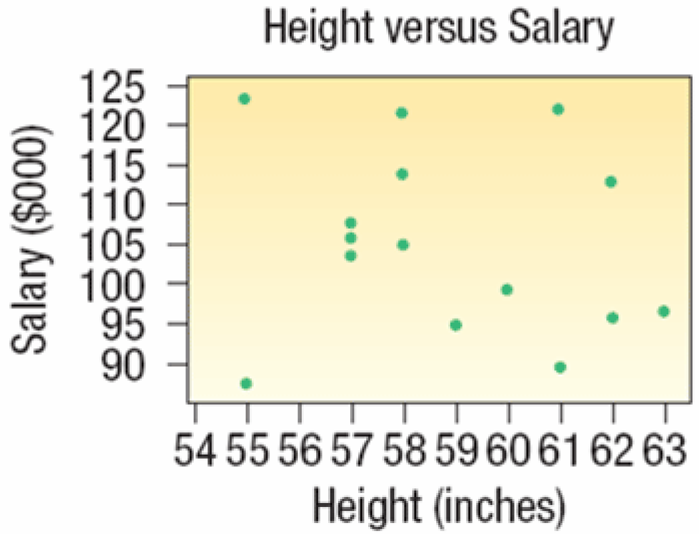
- One graphical technique we use to show the relationship between variables is called a **scatter diagram.**

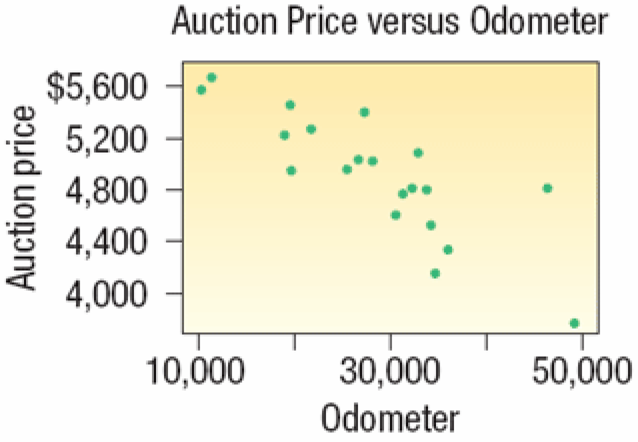
- To draw a scatter diagram we need two variables. We scale one variable along the horizontal axis (*X*-axis) of a graph and the other variable along the vertical axis (*Y*-axis).

LO6

Describing Relationship between Two Variables – Scatter Diagram Examples





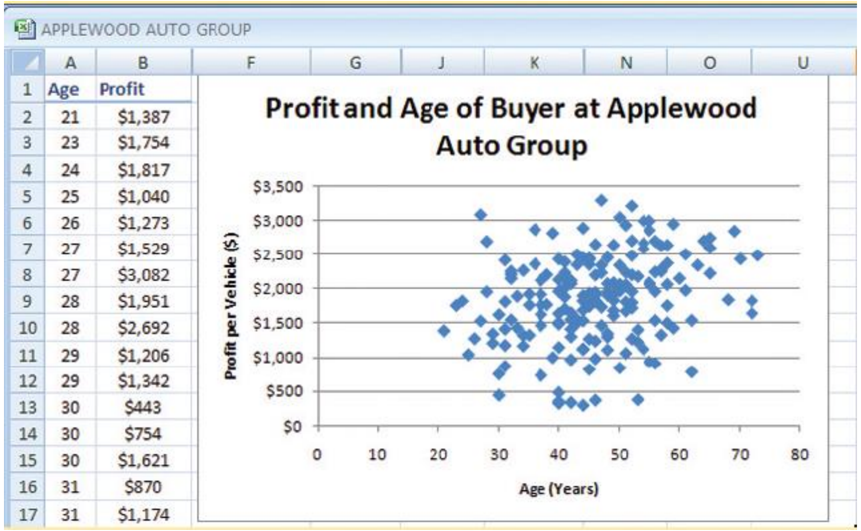


Describing Relationship between Two Variables – Scatter Diagram Excel Example

In Chapter 2 data from Auto USA was presented. We gathered information concerning several variables, including the profit earned from the sale of 180 vehicles sold last month. In addition to the amount of profit on each sale, one of the other variables is the age of the purchaser.

Is there a relationship between the profit earned on a vehicle sale and the age of the purchaser?

Would it be reasonable to conclude that the more expensive vehicles are purchased by older buyers?

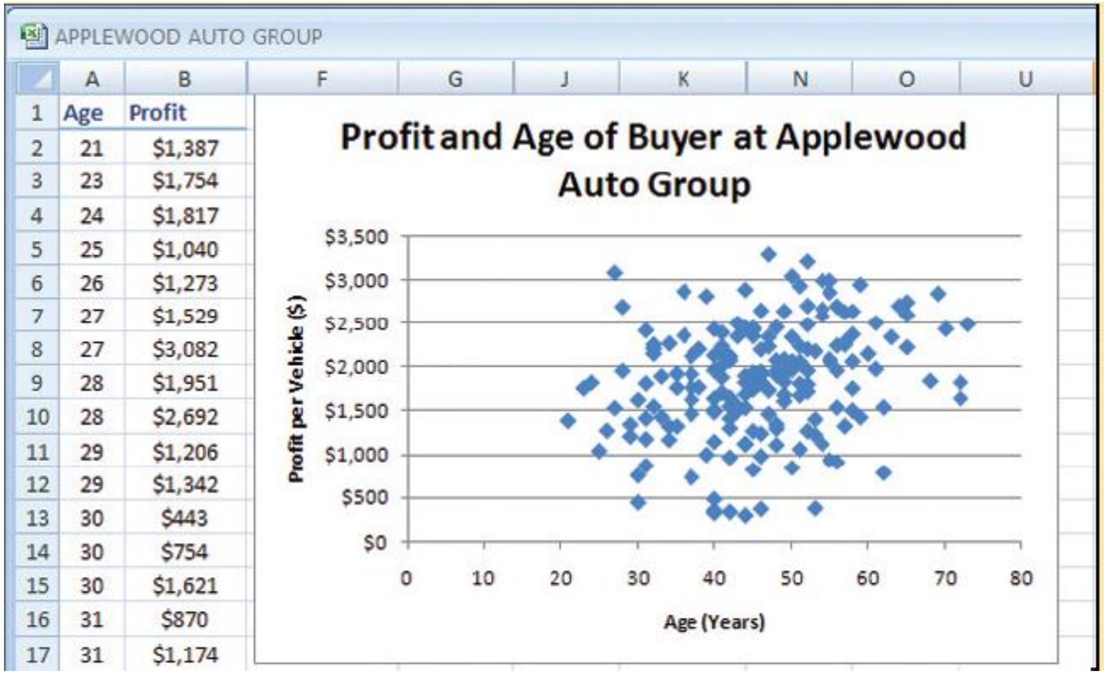


APPLEWOOD AUTO GROUP

	A	B
1	Age	Profit
2	21	$1,387
3	23	$1,754
4	24	$1,817
5	25	$1,040
6	26	$1,273
7	27	$1,529
8	27	$3,082
9	28	$1,951
10	28	$2,692
11	29	$1,206
12	29	$1,342
13	30	$443
14	30	$754
15	30	$1,621
16	31	$870
17	31	$1,174

LO6

Describing Relationship between Two Variables – Scatter Diagram Excel Example



	A	B
1	Age	Profit
2	21	$1,387
3	23	$1,754
4	24	$1,817
5	25	$1,040
6	26	$1,273
7	27	$1,529
8	27	$3,082
9	28	$1,951
10	28	$2,692
11	29	$1,206
12	29	$1,342
13	30	$443
14	30	$754
15	30	$1,621
16	31	$870
17	31	$1,174

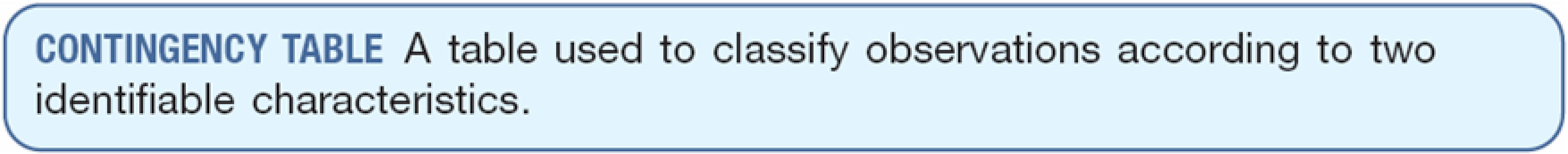
The scatter diagram shows a rather weak positive relationship between the two variables.

It does not appear there is much relationship between the vehicle profit and the age of the buyer.

LO7 Develop and explain a *contingency table*.

Contingency Tables

- A scatter diagram requires that both of the variables be at least **interval scale**.
- What if we wish to study the relationship between two variables when one or both are **nominal** or **ordinal scale**? In this case we tally the results in a **contingency table.**



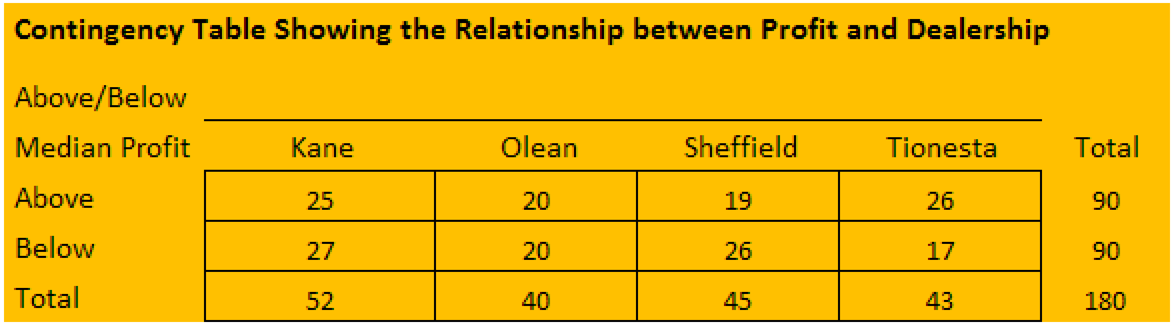
CONTINGENCY TABLE A table used to classify observations according to two identifiable characteristics.

Examples:

1. Students at a university are classified by gender and class rank.
2. A product is classified as acceptable or unacceptable and by the shift (day, afternoon, or night) on which it is manufactured.
3. A voter in a school bond referendum is classified as to party affiliation (Democrat, Republican, other) and the number of children that voter has attending school in the district (0, 1, 2, etc.).

Contingency Tables – An Example

There are four dealerships in the Apple wood Auto group. Suppose we want to compare the profit earned on each vehicle sold by the particular dealership. To put it another way, is there a relationship between the amount of profit earned and the dealership? The table below is the cross-tabulation of the raw data of the two variables.



Contingency Table Showing the Relationship between Profit and Dealership

Above/Below Median Profit	Kane	Olean	Sheffield	Tionesta	Total
Above	25	20	19	26	90
Below	27	20	26	17	90
Total	52	40	45	43	180

From the contingency table, we observe the following:

1. From the Total column on the right, 90 of the 180 cars sold had a profit above the median and half below. From the definition of the median this is expected.
2. For the Kane dealership 25 out of the 52, or 48 percent, of the cars sold were sold for a profit more than the median.
3. The percent profits above the median for the other dealerships are 50 percent for Olean, 42 percent for Sheffield, and 60 percent for Tionesta.