Business Analytics 3rd edition Camm Solutions Manual

SKU: 9781337406420-SOLUTIONS

Chapter 2
Descriptive Statistics

Solutions:

1. a. Quantitative

   b. Categorical

   c. Categorical

   d. Quantitative

   e. Categorical

2. a. The top 10 countries according to GDP are listed below.

Country	Continent	GDP (millions of US$)
United States	North America	15,094,025
China	Asia	7,298,147
Japan	Asia	5,869,471
Germany	Europe	3,577,031
France	Europe	2,776,324
Brazil	South America	2,492,908
United Kingdom	Europe	2,417,570
Italy	Europe	2,198,730
Russia	Asia	1,850,401
Canada	North America	1,736,869

   b. The top 5 countries by GDP located in Africa are listed below.

Country	Continent	GDP (millions of US$)
South Africa	Africa	408,074
Nigeria	Africa	238,920
Egypt	Africa	235,719
Algeria	Africa	190,709
Angola	Africa	100,948

3. a. The sorted list of carriers appears below.

Carrier	Previous Year On-time Percentage	Current Year On-time Percentage
Blue Box Shipping	88.4%	94.8%
Cheetah LLC	89.3%	91.8%
Smith Logistics	84.3%	88.7%
Granite State Carriers	81.8%	87.6%

Super Freight	92.1%	86.8%
Minuteman Company	91.0%	84.2%
Jones Brothers	68.9%	82.8%
Honsin Limited	74.2%	80.1%
Rapid Response	78.8%	70.9%

Blue Box Shipping is providing the best on-time service in the current year. Rapid Response is providing the worst on-time service in the current year.

b. The output from Excel with conditional formatting appears below.

	A	B	C	D
1	**Carrier**	**Previous Year On-time Percentage**	**Current Year On-time Percentage**	**Change in On-time Percentage**
2	Blue Box Shipping	88.4%	94.8%	6.4%
3	Cheetah LLC	89.3%	91.8%	2.5%
4	Smith Logistics	84.3%	88.7%	4.4%
5	Granite State Carriers	81.8%	87.6%	5.8%
6	Super Freight	92.1%	86.8%	-5.3%
7	Minuteman Company	91.0%	84.2%	-6.8%
8	Jones Brothers	68.9%	82.8%	13.9%
9	Honsin Limited	74.2%	80.1%	5.9%
10	Rapid Response	78.8%	70.9%	-7.9%

c. The output from Excel containing data bars appears below.

	A	B	C	D
1	**Carrier**	**Previous Year On-time Percentage**	**Current Year On-time Percentage**	**Change in On-time Percentage**
2	Blue Box Shipping	88.4%	94.8%	6.4%
3	Cheetah LLC	89.3%	91.8%	2.5%
4	Smith Logistics	84.3%	88.7%	4.4%
5	Granite State Carriers	81.8%	87.6%	5.8%
6	Super Freight	92.1%	86.8%	-5.3%
7	Minuteman Company	91.0%	84.2%	-6.8%
8	Jones Brothers	68.9%	82.8%	13.9%
9	Honsin Limited	74.2%	80.1%	5.9%
10	Rapid Response	78.8%	70.9%	-7.9%

d. The top 4 shippers based on current year on-time percentage (Blue Box Shipping, Cheetah LLC, Smith Logistics, and Granite State Carriers) all have positive increases from the previous year and high on-time percentages. These are good candidates for carriers to use in the future.

4. a. The relative frequency of D is 1.0 – 0.22 – 0.18 – 0.40 = 0.20.

b. If the total sample size is 200 the frequency of D is 0.20*200 = 40.

c. and d.

Class	Relative Frequency	Frequency	% Frequency
A	0.22	44	22

B	0.18	36	18
C	0.40	80	40
D	0.20	40	20
Total	1.0	200	100

5. a. These data are categorical.

b.

Website	Frequency	% Frequency
FB	7	14
GOOG	14	28
WIKI	9	18
YAH	13	26
YT	7	14
Total	50	100

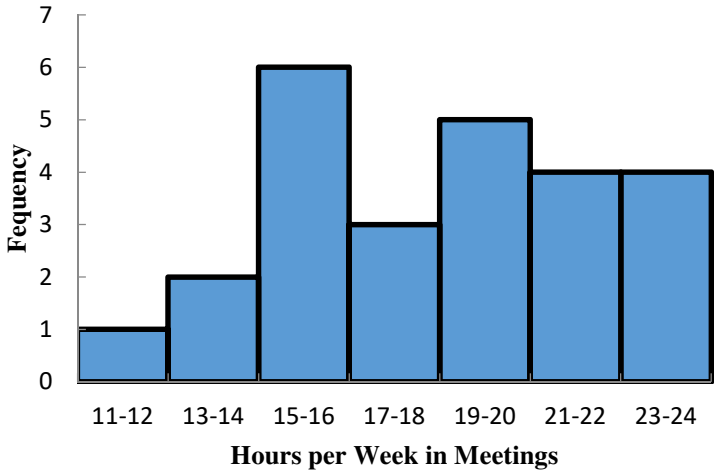
c. The most frequent most-visited-website is google.com (GOOG); second is yahoo.com (YAH).

6. a. Least = 12, Highest = 23

b.

Hours in Meetings per Week	Frequency	Percent Frequency
11-12	1	4%
13-14	2	8%
15-16	6	24%
17-18	3	12%
19-20	5	20%
21-22	4	16%
23-24	4	16%
	25	100%

c.



The distribution is slightly skewed to the left.

7. a.

Industry	Frequency	% Frequency
Bank	26	13%
Cable	44	22%
Car	42	21%
Cell	60	30%
Collection	28	14%
Total	200	100%

b. The cellular phone providers had the highest number of complaints.

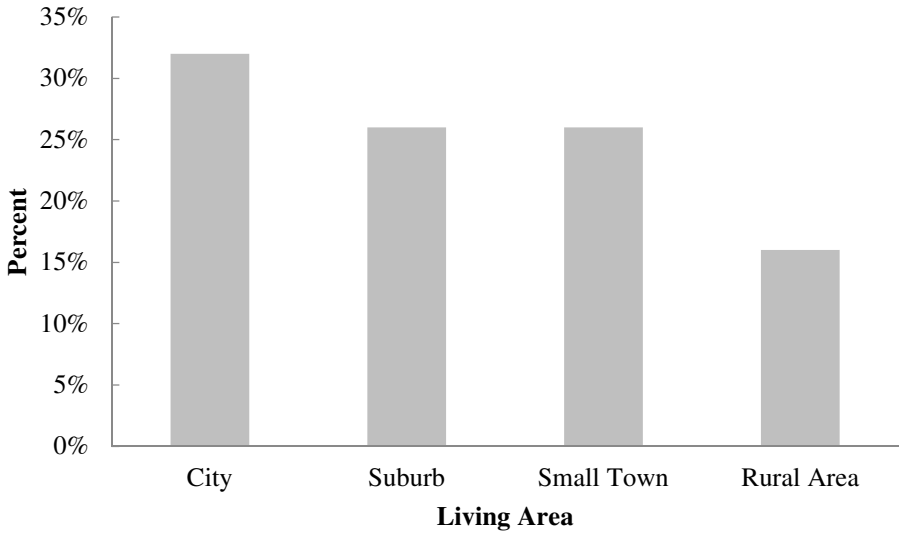
c. The percentage frequency distribution shows that the two financial industries (banks and collection agencies) had about the same number of complaints. Also, new car dealers and cable and satellite television companies also had about the same number of complaints.

8. Percent Frequency Distribution:

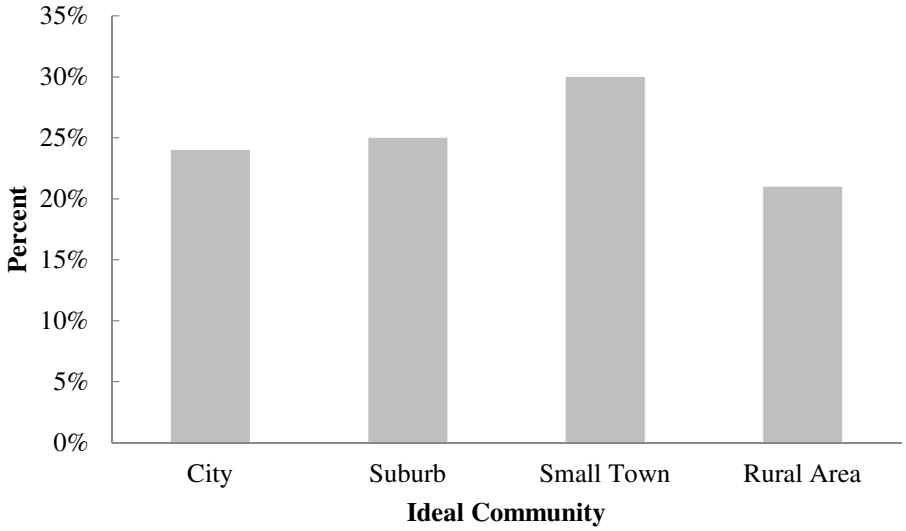
Living Area	Live Now	Ideal Community
City	32/100=32%	24/100=24%
Suburb	26/100=26%	25/100=25%
Small Town	26/100=26%	30/100=30%
Rural Area	16/100=16%	21/100=21%
Total	100%	100%

Histograms:

Where do you live now?



What do you consider the ideal community?



a. Most adults are now living in a city (32%).

b. Most adults consider the ideal community a small town (30%).

c. Changes in percentages by living area: City –8%, Suburb –1%, Small Town +4%, and Rural Area +5%.
Suburb living is steady, but the trend would be that living in the city would decline while living in small towns and rural areas would increase.

9. a.

Class	Frequency
12-14	2
15-17	8
18-20	11
21-23	10
24-26	9
Total:	40

b.

Class	Relative Frequency	Percent Frequency
12-14	0.050	5.0%
15-17	0.200	20.0%
18-20	0.275	27.5%
21-23	0.250	25.0%
24-26	0.225	22.5%
Total:	1.000	100.0%

10.

Class	Frequency	Cumulative Frequency
10-19	10	10
20-29	14	24
30-39	17	41
40-49	7	48
50-59	2	50

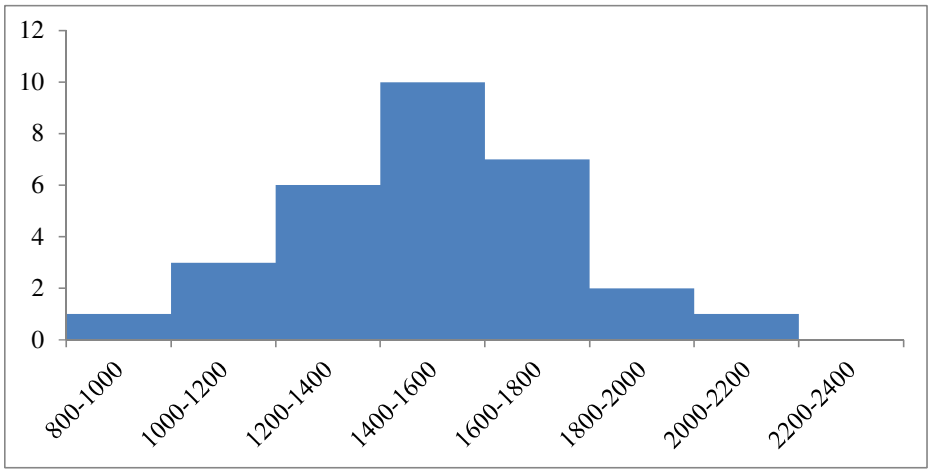
11. a – d.

Class	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
0-4	4	0.20	4	0.20
5-9	8	0.40	12	0.60
10-14	5	0.25	17	0.85
15-19	2	0.10	19	0.95
20-24	1	0.05	20	1.00
Total:	20	1.00		

e. From the cumulative relative frequency distribution, 60% of customers wait 9 minutes or less.

12. a.

Class	Frequency
800-1000	1
1000-1200	3
1200-1400	6
1400-1600	10
1600-1800	7
1800-2000	2
2000-2200	1
2200-2400	0



b. The distribution is slightly skewed to the right.

c. The most common score for students is between 1400 and 1600. No student scored above 2200, and only 3 students scored above 1800. Only 4 students scored below 1200.

13. a. Mean = $\frac{10+20+12+17+16}{5} = 15$ or use the Excel function AVERAGE.
To calculate the median, we arrange the data in ascending order:
10 12 16 17 20
Because we have $n = 5$ values which is an odd number, the median is the middle value which is 16 or use the Excel function MEDIAN.

b. Because the additional data point, 12, is lower than the mean and median computed in part a, we expect the mean and median to decrease. Calculating the new mean and median gives us mean = 14.5 and median = 14.

14. Without Excel, to calculate the 20th percentile, we first arrange the data in ascending order:
15 20 25 25 27 28 30 34
The location of the pth percentile is given by the formula $L_p = \frac{p}{100}(n+1)$
For our date set, $L_{20} = \frac{20}{100}(8+1) = 1.8$. Thus, the 20th percentile is 80% of the way between the value in position 1 and the value in position 2. In other words, the 20th percentile is the value in position 1 (15) plus 0.80 time the difference between the value in position 2 (20) and position 1 (15). Therefore, the 20th percentile is
15 + 0.80*(20-15) = 19.

We can repeat the steps above to calculate the 25th, 65th and 75th percentiles. Or using Excel, we can use the function PERCENTILE.EXC to get:
25th percentile = 21.25
65th percentile = 27.85
75th percentile = 29.5

15. Mean = $\frac{53+55+70+58+64+57+53+69+57+68+53}{11} = 59.727$ or use the Excel function AVERAGE.
To calculate the median arrange the values in ascending order
53 53 53 55 57 57 58 64 68 69 70
Because we have n = 11, an odd number of values, the median is the middle value which is 57 or use the Excel function MEDIAN.
The mode is the most often occurring value which is 53 because 53 appears three times in the data set, or use the Excel function MODE.SNGL because there is only a single mode in this data set.

16. To find the mean annual growth rate, we must use the geometric mean. First we note that

$3500=5000\left[(x_1)(x_2)\cdots(x_9)\right]$, so $\left[(x_1)(x_2)\cdots(x_9)\right]=0.700$
where x_1, x_2, … are the growth factors for years, 1, 2, etc. through year 9.
Next, we calculate $\bar{x}_g = \sqrt[n]{(x_1)(x_2)\cdots(x_n)} = \sqrt[9]{0.70} = 0.961144$.
So the mean annual growth rate is (0.961144 – 1)100% = -0.38856%

17. For the Stivers mutual fund,
$18000=10000\left[(x_1)(x_2)\cdots(x_8)\right]$, so $\left[(x_1)(x_2)\cdots(x_8)\right]=1.8$
where x_1, x_2, … are the growth factors for years, 1, 2, etc. through year 8.
Next, we calculate $\bar{x}_g = \sqrt[n]{(x_1)(x_2)\cdots(x_8)} = \sqrt[8]{1.80} = 1.07624$
So the mean annual return for the Stivers mutual fund is (1.07624 – 1)100 = 7.624%.

For the Trippi mutual fund we have:
$10600=5000\left[(x_1)(x_2)\cdots(x_8)\right]$, so $\left[(x_1)(x_2)\cdots(x_8)\right]=2.12$ and
$\bar{x}_g = \sqrt[n]{(x_1)(x_2)\cdots(x_8)} = \sqrt[8]{2.12} = 1.09848$
So the mean annual return for the Trippi mutual fund is (1.09848 – 1)100 = 9.848%.

While the Stivers mutual fund has generated a nice annual return of 7.6%, the annual return of 9.8% earned by the Trippi mutual fund is far superior.

Alternatively, we can use Excel and the function GEOMEAN as shown below:

	A	B	C	D	E
1		Stivers		Trippi	
2	Year	End of Year Value	Growth Factor	End of Year Value	Growth Factor
3	0	$10,000		$5,000	
4	1	$11,000	1.100	$5,600	1.120
5	2	$12,000	1.091	$6,300	1.125
6	3	$13,000	1.083	$6,900	1.095
7	4	$14,000	1.077	$7,600	1.101
8	5	$15,000	1.071	$8,500	1.118
9	6	$16,000	1.067	$9,200	1.082
10	7	$17,000	1.063	$9,900	1.076
11	8	$18,000	1.059	$10,600	1.071
12					
13					
14		Stivers Geometric Mean:		1.07623984	
15		Trippi Geometric Mean:		1.09847957	

	A	B	C	D	E
1		Stivers		Trippi	
2	Year	End of Year Value	Growth Factor	End of Year Value	Growth Factor
3	0	10000		5000	
4	1	11000	=B4/B3	5600	=D4/D3
5	2	12000	=B5/B4	6300	=D5/D4
6	3	13000	=B6/B5	6900	=D6/D5
7	4	14000	=B7/B6	7600	=D7/D6
8	5	15000	=B8/B7	8500	=D8/D7
9	6	16000	=B9/B8	9200	=D9/D8
10	7	17000	=B10/B9	9900	=D10/D9
11	8	18000	=B11/B10	10600	=D11/D10
12					
13					
14			Geometric Mean:	=GEOMEAN(C3:C11)	
15			Geometric Mean:	=GEOMEAN(E4:E11)	

18. a. Mean $= \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1291.5}{48} = 26.906$

b. To calculate the median, we first sort all 48 commute times in ascending order. Because there are an even number of values (48), the median is between the 24th and 25th largest values. The 24th largest value is 25.8 and the 25th largest value is 26.1.
(25.8 + 26.1)/2 = 25.95
Or we can use the Excel function MEDIAN.

c. The values 23.4 and 24.8 both appear three times in the data set, so these two values are the modes of the commute times. To find this using Excel, we must use the MODE.MULT function.

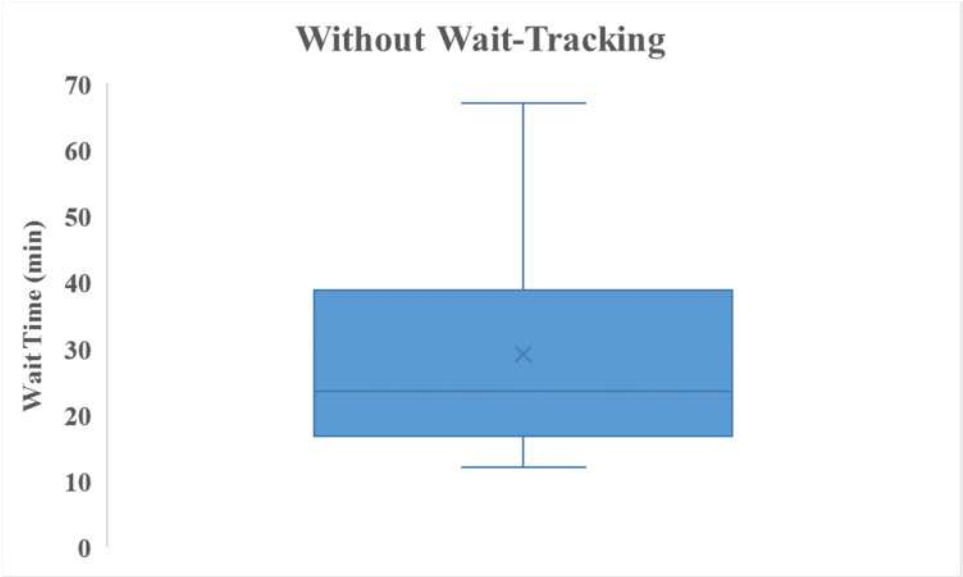
d. Standard deviation = 4.6152. In Excel, we can find this value using the function STDEV.S.
Variance = 4.6152^2 = 21.2998. In Excel, we can find this value using the function VAR.S.

e. The third quartile is the 75th percentile of the data. To find the 75th percentile without Excel, we first arrange the data in ascending order. Next we calculate $L_p = \frac{p}{100}(n+1) = L_{75} = \frac{75}{100}(48+1) = 36.75$.
In other words, this value is 75% of the way between the 36th and 37th positions. However, in our date the values in both the 36th and 37th positions are 28.5. Therefore, the 75th percentile is 28.5. Or using Excel, we can use the function PERCENTILE.EXC.

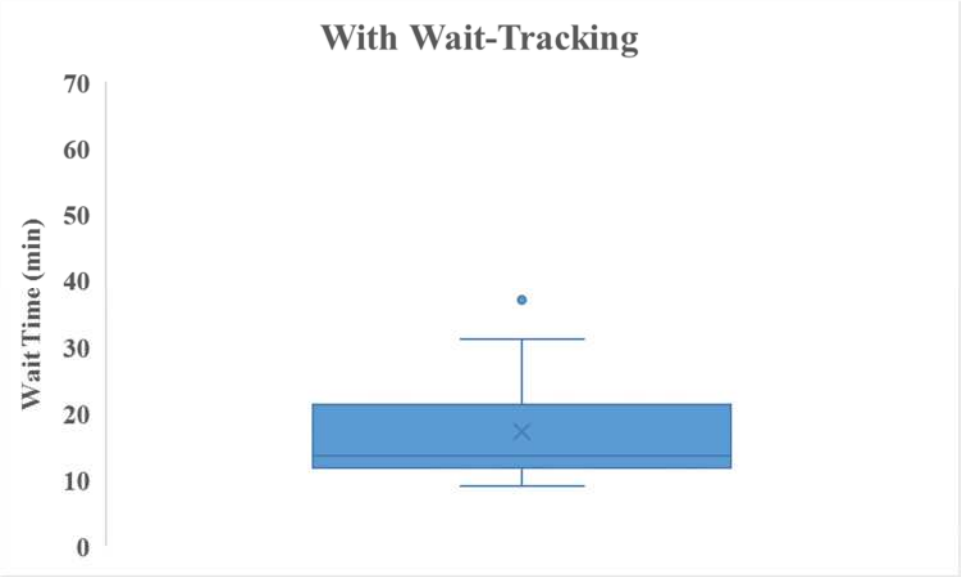
19. a. The mean waiting time for patients with the wait-tracking system is 17.2 minutes and the median waiting time is 13.5 minutes. The mean waiting time for patients without the wait-tracking system is 29.1 minutes and the median is 23.5 minutes.

b. The standard deviation of waiting time for patients with the wait-tracking system is 9.28 and the variance is 86.18. The standard deviation of waiting time for patients without the wait-tracking system is 16.60 and the variance is 275.66.

c.



d.



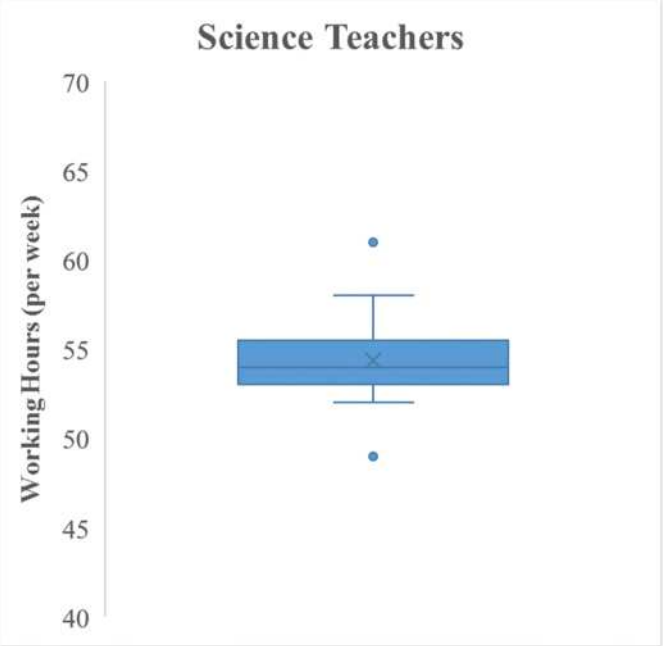
e. Wait times for patients with the wait-tracking system are substantially shorter than those for patients without the wait-tracking system. However, some patients with the wait-tracking system still experience long waits.

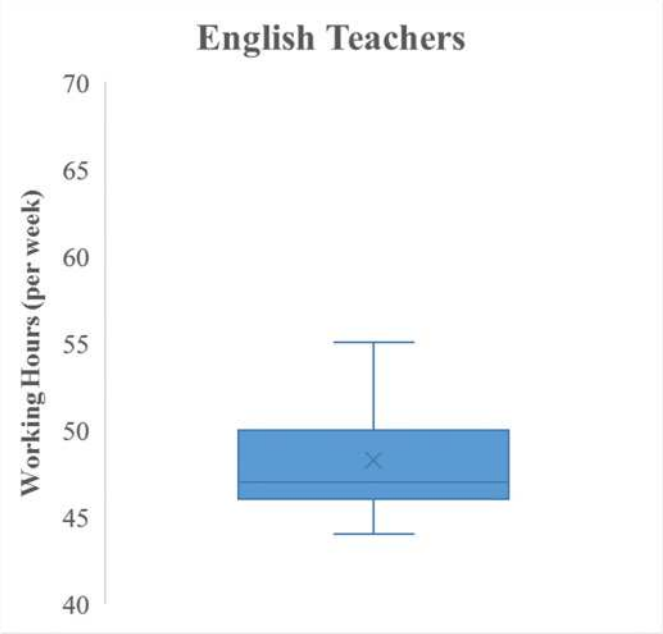
20. a. The median number of hours worked for science teachers is 54.

b. The median number of hours worked for English teachers is 47.

c.

d.





e. The box plots show that science teachers spend more hours working per week than English teachers. The box plot for science teachers also shows that most science teachers work about the same amount of hours; in other words, there is less variability in the number of hours worked for science teachers.

21. a. Recall that the mean patient wait time without wait-time tracking is 29.1 and the standard deviation of wait times is 16.6. Then the z-score is calculated as, $z = \frac{37-29.1}{16.6} = 0.48$.

b. Recall that the mean patient wait time with wait-time tracking is 17.2 and the standard deviation of wait times is 9.28. Then the z-score is calculated as, $z = \frac{37-17.2}{9.28} = 2.13$.
As indicated by the positive z–scores, both patients had wait times that exceeded the means of their respective samples. Even though the patients had the same wait time, the z–score for the sixth patient

in the sample who visited an office with a wait tracking system is much larger because that patient is part of a sample with a smaller mean and a smaller standard deviation.

c. To calculate the z-score for each patient waiting time, we can use the formula $z = \frac{x_i - \bar{x}}{s}$ or we can use the Excel function STANDARDIZE. The z–scores for all patients follow.

Without Wait-Tracking System		**With Wait-Tracking System**	
Wait Time	**z-Score**	**Wait Time**	**z-Score**
24	-0.31	31	1.49
67	2.28	11	-0.67
17	-0.73	14	-0.34
20	-0.55	18	0.09
31	0.11	12	-0.56
44	0.90	37	2.13
12	-1.03	9	-0.88
23	-0.37	13	-0.45
16	-0.79	12	-0.56
37	0.48	15	-0.24

No z-score is less than -3.0 or above +3.0; therefore, the z–scores do not indicate the existence of any outliers in either sample.

22. a. According to the empirical rule, approximately 95% of data values will be within two standard deviations of the mean. 4.5 is two standard deviation less than the mean and 9.3 is two standard deviations greater than the mean. Therefore, approximately 95% of individuals sleep between 4.5 and 9.3 hours per night.

b. $z = \frac{8-6.9}{1.2} = 0.9167$

c. $z = \frac{6-6.9}{1.2} = -0.75$

23. a. 615 is one standard deviation above the mean. The empirical rule states that 68% of data values will be within one standard deviation of the mean. Because a bell-shaped distribution is symmetric half of the remaining values will be greater than the (mean + 1 standard deviation) and half will be below (mean – 1 standard deviation). In other words, we expect that 0.5*(1 - 68%) = 16% of the data values will be greater than (mean + 1 standard deviation) = 615.

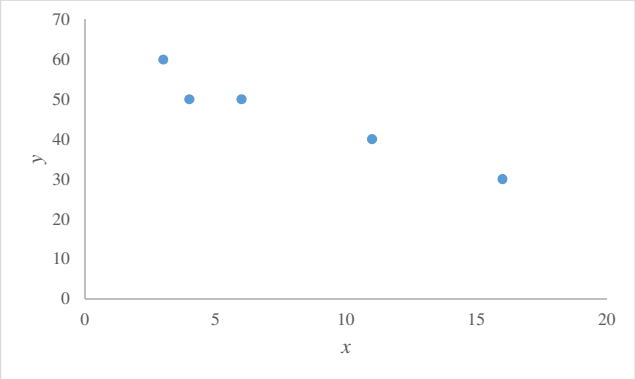
b. 715 is two standard deviations above the mean. The empirical rule states that 95% of data values will be within two standard deviations of the mean, and we expect that 0.5*(1 - 95%) = 2.5% of data values will be above two standard deviations above the mean.

c. 415 is one standard deviation below the mean. The empirical rule states that 68% of data values will be within one standard deviation of the mean, and we expect that 0.5*(1 - 68%) = 16% of data values will be below one standard deviation below the mean. 515 is the mean, so we expect that 50% of the data values will be below the mean. Therefore, we expect 50% - 16% = 36% of the data values will be between the mean and one standard deviation below the mean (between 414 and 515).

d. $z = \frac{620-515}{100} = 1.05$

e. $z = \frac{405-515}{100} = -1.10$

24. a.



b. There appears to be a negative linear relationship between the *x* and *y* variables.

c. Without Excel, we can use the calculations shown below to calculate the covariance:

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
4	50	-4	4	-16
6	50	-2	4	-8
11	40	3	-6	-18
3	60	-5	14	-70
16	30	8	-16	-128

$\bar{x} =$ 8

$\bar{y} =$ 46

$$s_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1} = \frac{-16-8-18-70-128}{4} = -60$$

Or, using Excel, we can use the COVARIANCE.S function.

The negative covariance confirms that there is a negative linear relationship between the x and y variables in this data set.

d. To calculate the correlation coefficient without Excel, we need the standard deviation for *x* and *y*: $s_x = 5.43, s_y = 11.40$. Then the correlation coefficient is calculated as:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{-60}{(5.43)(11.40)} = -0.97.$$

Or we can use the Excel function CORREL.
The correlation coefficient indicates a strong negative linear association between the *x* and *y* variables in this data set.

25. a. The scatter chart indicates that there may be a positive linear relationship between profits and market capitalization.

b. Without Excel, we can use the calculations below to find the covariance and correlation coefficient:

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
313.2	1891.9	-2468.57	-35259.75	6093826.70	1243249856.32	87041077.46
631	81458.6	-2150.77	44306.95	4625801.88	1963105961.23	-95293962.27
706.6	10087.6	-2075.17	-27064.05	4306321.16	732462715.10	56162440.18
-29	1175.8	-2810.77	-35975.85	7900415.30	1294261667.17	101119754.14
4,018.00	55188.8	1236.23	18037.15	1528270.20	325338838.31	22298108.67
959	14115.2	-1822.77	-23036.45	3322482.24	530677954.29	41990095.01
6,490.00	97376.2	3708.23	60224.55	13750986.48	3626996616.98	223326625.02
8,572.00	157130.5	5790.23	119978.85	33526789.60	14394924834.35	694705416.89
12,436.00	95251.9	9654.23	58100.25	93204200.49	3375639237.48	560913323.32
1,462.00	36461.2	-1319.77	-690.45	1741786.89	476718.98	911231.51
3,461.00	53575.7	679.23	16424.05	461356.46	269749471.38	11155745.66
854	7082.1	-1927.77	-30069.55	3716288.47	904177740.20	57967105.40
369.5	3461.4	-2412.27	-33690.25	5819035.66	1135032836.38	81269899.40
399.8	12520.3	-2381.97	-24631.35	5673770.32	606703323.37	58671077.30
278	3547.6	-2503.77	-33604.05	6268852.91	1129232068.00	84136732.35
9,190.00	32382.4	6408.23	-4769.25	41065440.67	22745730.18	-30562451.36
599.1	8925.3	-2182.67	-28226.35	4764038.47	796726743.27	61608740.10
2,465.00	9550.2	-316.77	-27601.45	100341.80	761839953.07	8743248.48
3,527.00	65917.4	745.23	28765.75	555371.12	827468465.86	21437166.03
602	13819.5	-2179.77	-23332.15	4751387.41	544389148.36	50858664.40
2,655.00	26651.1	-126.77	-10500.55	16070.06	110261516.43	1331130.81
1,455.70	21865.9	-1326.07	-15285.75	1758455.66	233654103.75	20269937.85
276	3417.8	-2505.77	-33733.85	6278871.98	1137972527.00	84529189.10
617.5	3681.2	-2164.27	-33470.45	4684054.86	1120270915.23	72439011.75
11,797.00	182109.9	9015.23	144958.25	81274412.67	21012894710.67	1306832306.01
567.6	12522.8	-2214.17	-24628.85	4902538.79	606580172.87	54532401.62
697.8	10514.8	-2083.97	-26636.85	4342921.55	709521692.00	55510332.79
634	8560.5	-2147.77	-28591.15	4612906.27	817453766.09	61407146.21
109	1381.6	-2672.77	-35770.05	7143687.40	1279496361.62	95605031.46
4,979.00	66606.5	2197.23	29454.85	4827829.60	867588283.54	64719150.12
5,142.00	53469.4	2360.23	16317.75	5570696.31	266269017.70	38513683.74
			Total	368589209.4	62647162947	3954149359

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{3954149359}{30} = 131804978.6$$

$$s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{368589209.4}{30}} = 3505.18$$

$$s_y = \sqrt{\frac{\sum(y - \bar{y})^2}{n-1}} = \sqrt{\frac{62647162947}{30}} = 45697.25$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{131804978.6}{(3505.18)(45697.25)} = 0.8229$$

Or using Excel, we use the formula = COVARIANCE.S(B2:B32,C2:C32) to calculate the covariance, which is 131804978.638. This indicates that there is a positive relationship between profits and market capitalization.

c. In the Excel file, we use the formula =CORREL(B2:B32,C2:C32) to calculate the correlation coefficient, which is 0.8229. This indicates that there is a strong linear relationship between profits and market capitalization.

26. a. Without Excel, we can use the calculations below to find the correlation coefficient:

x_i	y_i	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
7.1	7.02	0.2852	0.6893	0.0813	0.4751	0.1966
5.2	5.31	-1.6148	-1.0207	2.6076	1.0419	1.6483
7.8	5.38	0.9852	-0.9507	0.9706	0.9039	-0.9367
7.8	5.40	0.9852	-0.9307	0.9706	0.8663	-0.9170
5.8	5.00	-1.0148	-1.3307	1.0298	1.7709	1.3505
5.8	4.07	-1.0148	-2.2607	1.0298	5.1109	2.2942
9.3	6.53	2.4852	0.1993	6.1761	0.0397	0.4952
5.7	5.57	-1.1148	-0.7607	1.2428	0.5787	0.8481
7.3	6.99	0.4852	0.6593	0.2354	0.4346	0.3199
7.6	11.12	0.7852	4.7893	0.6165	22.9370	3.7605
8.2	7.56	1.3852	1.2293	1.9187	1.5111	1.7028
7.1	12.11	0.2852	5.7793	0.0813	33.3998	1.6482
6.3	4.39	-0.5148	-1.9407	0.2650	3.7665	0.9991
6.6	4.78	-0.2148	-1.5507	0.0461	2.4048	0.3331
6.2	5.78	-0.6148	-0.5507	0.3780	0.3033	0.3386
6.3	6.08	-0.5148	-0.2507	0.2650	0.0629	0.1291
7.0	10.05	0.1852	3.7193	0.0343	13.8329	0.6888
6.2	4.75	-0.6148	-1.5807	0.3780	2.4987	0.9719
5.5	7.22	-1.3148	0.8893	1.7287	0.7908	-1.1692
6.5	3.79	-0.3148	-2.5407	0.0991	6.4554	0.7999
6.0	3.62	-0.8148	-2.7107	0.6639	7.3481	2.2088
8.3	9.24	1.4852	2.9093	2.2058	8.4638	4.3208
7.5	4.40	0.6852	-1.9307	0.4695	3.7278	-1.3229
7.1	6.91	0.2852	0.5793	0.0813	0.3355	0.1652
6.8	5.57	-0.0148	-0.7607	0.0002	0.5787	0.0113
5.5	3.87	-1.3148	-2.4607	1.7287	6.0552	3.2354
7.5	8.42	0.6852	2.0893	0.4695	4.3650	1.4315
			Total	25.77407	130.0594	25.5517

$$s_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1} = \frac{25.5517}{26} = 0.9828$$

$$s_x = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}} = \sqrt{\frac{25.77407}{26}} = 0.9956$$

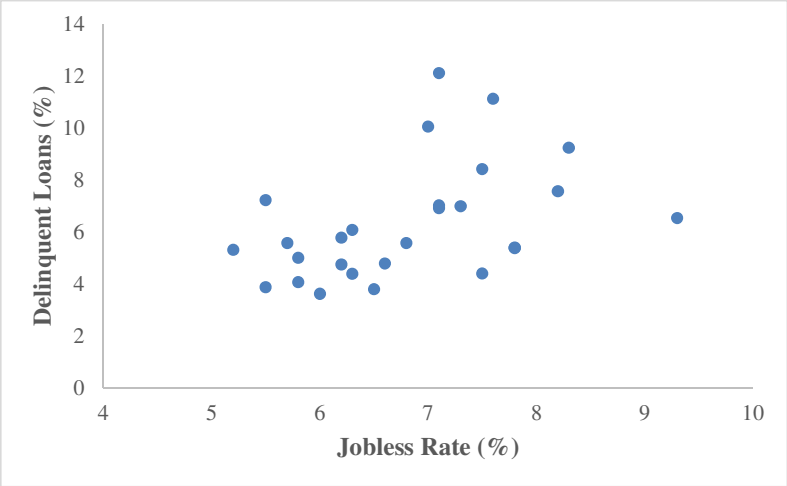
$$s_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n-1}} = \sqrt{\frac{130.0594}{26}} = 2.2366$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{0.9828}{(0.9956)(2.2366)} = 0.44$$

Or we can use the Excel function CORREL.

The correlation coefficient indicates that there is a moderate positive linear relationship between jobless rate and delinquent loans. If the jobless rate were to increase, it is likely that an increase in the percentage of delinquent housing loans would also occur.

b.



27. a. Using the Excel function COUNTBLANK we find that there is one blank response in column C (Texture) and one blank response in column F (Depth of Chocolate Flavor of the Cup). With further investigation we find that the value of texture for respondent 157 is missing, and the value of Depth of the Chocolate Flavor of the Cup for respondent 199 is missing.

b. To help us identify erroneous values, we calculate the Average, Standard Deviation, Minimum and Maximum values for each variable.

# Missing Values:	0	1	0	0	1
Average:	72.29	105.94	77.11	78.51	77.36
Standard Deviation:	20.21	447.89	19.14	66.15	17.69
Minimum:	8.40	13.00	19.00	0.67	11.00
Maximum:	100.00	6666.00	120.00	997.00	100.00

We can immediately spot some surprising values in Column C (Texture) and column E (Sweetness). Further examination identifies: the value of Texture for respondent 68 is 6666, which is outside the range for this variable; the value of sweetness for respondent 72 is 997, which is outside the range for this variable; the value of Sweetness for respondent 85 is 0.67, which is outside the range for this variable and is not an integer. Additional examination of the other responses shows that the value of Taste for respondent 90 is 8.4, which is not an integer, and the value of Creaminess of filling for respondent 197 is 120, which is outside the range for this variable.

28. a. Using the Excel function COUNTBLANK we find that one observation is missing in column B and one observation missing in column C. Additional investigation shows that the missing value in column B is for year 2016 for the Phillies and the missing value in column C is for year 2016 for the Marlins. Review of major league baseball attendance data that are available from a reliable source shows that the Phillies' attendance in 2016 was 1,915,144, which is consistent with the value of the Phillies' attendance for the observation with the missing value of season. This supports our suspicion that the value of season for this observation is 2016. Review of major league baseball attendance data that are available from a reliable source shows that the Marlins' 2016 attendance was 1,712,417. (Note that we use the reference ESPN.com, at http://www.espn.com/mlb/attendance as of May 14, 2017, as our reliable data source for comparison.)

b. To help us identify erroneous values, we calculate the average, standard deviation, minimum and maximum values for Season and Attendance (Note that we use the reference ESPN.com, at http://www.espn.com/mlb/attendance as of May 14, 2017, as our reliable data source for comparison.).

# Missing Values:	1	1
Average:	2,004	2,651,944.64
Standard Deviation:	124	2,333,151.88
Minimum:	214	-3,365,256.00
Maximum:	2,016	26,426,820.00

We immediately identify that there is an erroneous value for Season as all values should be between 2014 and 2016, but the minimum value is 214. We also identify that the minimum and maximum values for Attendance appear questionable. Review of major league baseball attendance data that are available from a reliable source shows that the Cubs' attendance in 2014 was 2,652,113, which is consistent with the value of the Cubs' attendance for the observation with the missing value of season. This supports our suspicion that the value of season for this observation is 2014.

The value for attendance for the Giants in 2016 is -3,365,256, which is unrealistic. Review of major league baseball attendance data that are available from a reliable source shows that the Giants' 2016 attendance was 3,365,256.

The value for attendance for the Cubs in 2013 is 26,426,820, which is unusually large. Review of major league baseball attendance data that are available from a reliable source shows that the Cubs' 2013 attendance was 2,642,682.

We can also sort the data in Excel by Team Name to help us identify attendance values that seem outside the norm for that team. Additional analysis of individual attendance values shows the following.

The value for attendance for the Royals in 2011 is 172,445, which is unusually small. Review of major league baseball attendance data that are available from a reliable source shows that the Royals' 2011 attendance was 1,724,450.

The value for attendance for the Marlins in 2014 is 9,732,283, which is unusually large compared to other season attendance values for the Marlins. Review of major league baseball attendance data that are available from a reliable source shows that the Marlins' 2014 attendance was 1,732,283.

The value for attendance for the Marlins in 2015 is 752,235, which is unusually small compared to other season attendance values for the Marlins. Review of major league baseball attendance data that are available from a reliable source shows that the Marlins' 2015 attendance was 1,752,235.

The value for attendance for the Orioles in 2014 is 22,464,473, which is unusually large. Review of major league baseball attendance data that are available from a reliable source shows that the Orioles' 2014 attendance was 2,464,473.

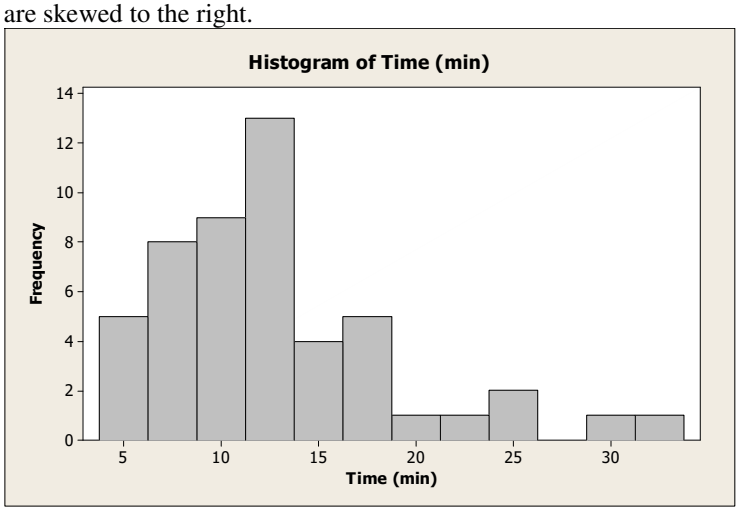
Chapter 2
Descriptive Statistics

Case Problem: Heavenly Chocolates Website Traffic

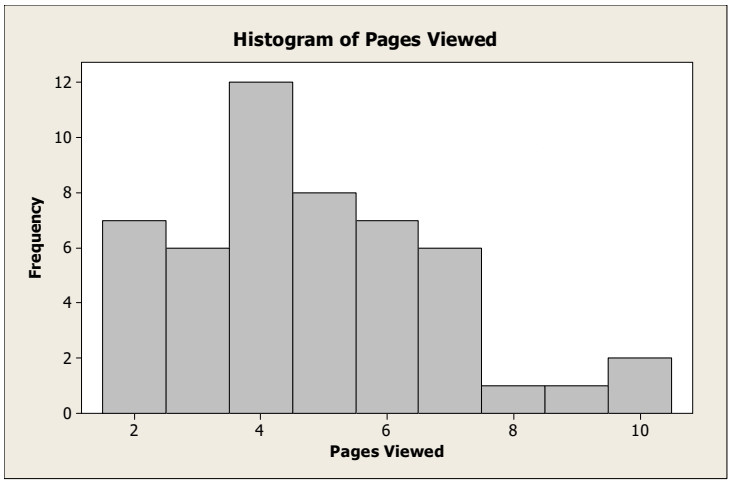
1. Descriptive statistics for the time spent on the website, number of pages viewed, and amount spent are shown below.

	Time (min)	Pages Viewed	Amount Spent ($)
Mean	12.8	4.8	68.13
Median	11.4	4.5	62.15
Standard Deviation	6.06	2.04	32.34
Range	28.6	8	140.67
Minimum	4.3	2	17.84
Maximum	32.9	10	158.51
Sum	640.5	241	3406.41

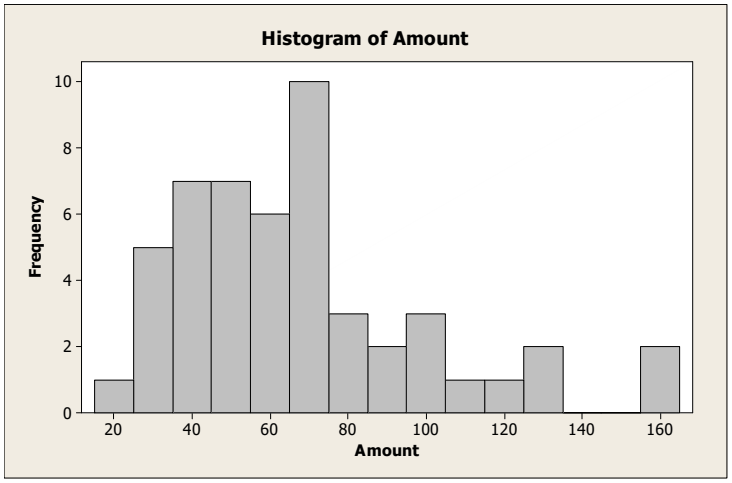
The mean time a shopper is on the Heavenly Chocolates website is 12.8 minutes, with a minimum time of 4.3 minutes and a maximum time of 32.9 minutes. The following histogram demonstrates that the data are skewed to the right.



The mean number of pages viewed during a visit is 4.8 pages with a minimun of 2 pages and a maximum of 10 pages A histogram of the number of pages viewed indicates that the data are slightly skewed to the right.



The mean amount spent for an on-line shopper is $68.13 with a minimum amount spent of $17.84 and a maximum amount spent of $158.51. The following histogram indicates that the data are skewed to the right.



2. Summary by Day of Week

Day of Week	Frequency	Total Amount Spent ($)	Average Amount Spent ($)
Sunday	5	218.15	43.63
Monday	9	813.38	90.38
Tuesday	7	414.86	59.27
Wednesday	6	341.82	56.97
Thursday	5	294.03	58.81
Friday	11	945.43	85.95
Saturday	7	378.74	54.11
Total	50	3406.41	68.13

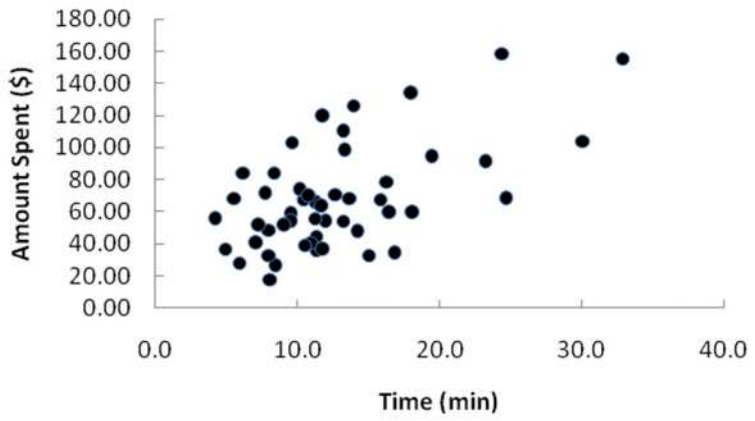
The above summary shows that Monday and Friday are the best days in terms of both the total amount spent and the averge amount spent per transaction. Friday had the most purchases (11) and the highest value for total amount spent ($945.43). Monday, with nine transactions, had the highest average amount spent per transaction ($90.38). Sunday was the worst sales day of the week in terms of number of transactions (5), total amount spent ($218.15), and average amount spent per transaction ($43.63). However, the sample size for each day of the week are very small, with only Friday having more than ten transactions. We would suggest a larger sample size be taken before recommending any specific stratgegy based on the day of week statistics.

3. Summary by Type of Browser

Browser	Frequency	Total Amount Spent ($)	Average Amount Spent ($)
Firefox	16	1228.21	76.76
Chrome	27	1656.81	61.36
Other	7	521.39	74.48

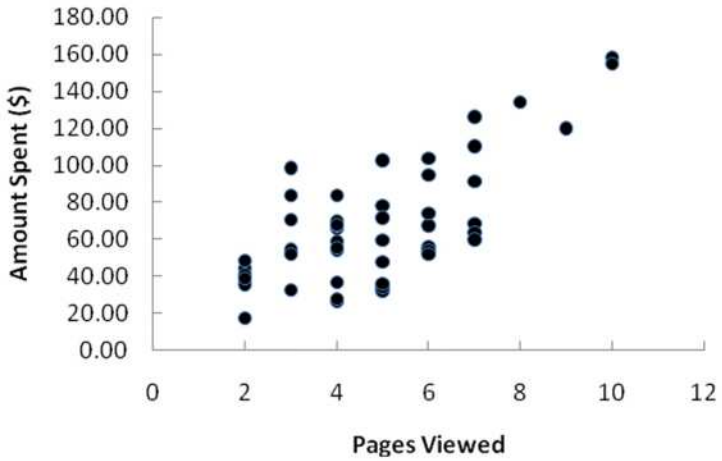
Chrome was used by 27 of the 50 shoppers (54%). But, the average amount spent spent by customers who used Chrome ($61.36) is less than the average amount spent by customers who used Firefox ($76.76) or some other type of browser ($74.48). This result would suggest targeting special promotion offers to Firefox users or users of other types of browsers. But, before recommending any specific strategies based upon the type of browser, we would suggest taking a larger smaple size.

4. A scatter diagram showing the relationship between time spent on the website and the amount spent follows:



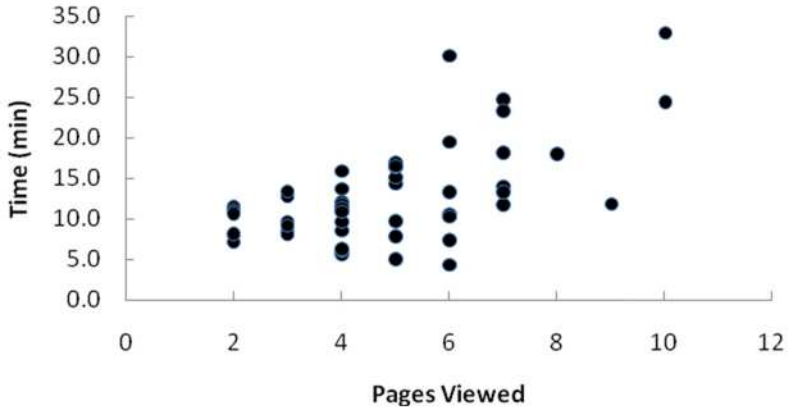
The sample correlation coefficient between these two variables is .580. The scatter diagram and the sample correlation coefficient indicate a postive relationship between time spent on the website and the total amount spent. Thus, the sample data support the conclusion that customers who spend more time on the website spend more.

5. A scatter diagram showing the relationship between the number of pages viewed and the amount spent follows:



The sample correlation coefficient between these two variables is .724. The scatter diagram and the sample correlation coefficient indicate a postive relationship between time spent on the website and the number of pages viewed. Thus, the sample data support the conclusion that customers who view more website pages spend more.

6. A scatter diagram showing the relationship between the number of pages viewed and the time spent on the website follows:



The sample correlation coefficient between these two variables is .596. The scatter diagram and the sample correlation coefficient indicate a postive relationship between the number of pages viewed and the time spent on the website.

Summary: The analysis indicates that on-line shoppers who spend more time on the company's website and/or view more website pages spend more money during their visit to the website. If Heavenly Chocolates can develop an attractive website such that on-line shoppers are willing to spend more time on the website and/or view more pages, there is a good possiblity that the company will experience greater sales. And, consideration should also be given to developing marketing strategies based upon possible differences in sales associated with the day of the week as well as differences in sales associated with the type of browser used by the customer.

Descriptive Statistics

Chapter 2

Overview of Using Data: Definitions and Goals

Overview of Using Data: Definitions and Goals (Slide 1 of 5)

- **Data**: The facts and figures collected, analyzed, and summarized for presentation and interpretation.
- **Variable**: A characteristic or a quantity of interest that can take on different values.
- **Observation**: A set of values corresponding to a set of variables.
- **Variation**: The difference in a variable measured over observations.
- **Random variable/uncertain variable**: A quantity whose values are not known with certainty.

Overview of Using Data: Definitions and Goals (Slide 2 of 5)

Table 2.1: Data for Dow Jones Industrial Index Companies

Company	Symbol	Industry	Share Price ($)	Volume
Apple	AAPL	Technology	160.47	18,997,275
American Express	AXP	Financial	91.69	2,939,556
Boeing	BA	Manufacturing	258.62	2,515,865
Caterpillar	CAT	Manufacturing	130.54	2,380,342
Cisco Systems	CSCO	Technology	33.60	9,303,117
Chevron Corporation	CVX	Chemical, Oil, and Gas	120.22	4,844,293
DuPont	DD	Chemical, Oil, and Gas	83.93	34,861,021

Overview of Using Data: Definitions and Goals (Slide 3 of 5)

Table 2.1: Data for Dow Jones Industrial Index Companies (cont.)

Company	Symbol	Industry	Share Price ($)	Volume
Disney	DIS	Entertainment	98.36	5,942,501
General Electric	GE	Conglomerate	23.19	58,639,089
Goldman Sachs	GS	Financial	236.09	7,088,445
The Home Depot	HD	Retail	163.35	4,189,197
IBM	IBM	Technology	146.54	6,372,393
Intel	INTC	Technology	39.79	15,532,818
Johnson & Johnson	JNJ	Pharmaceuticals	140.79	11,717,348
JPMorgan Chase	JPM	Banking	97.62	10,335,687

Overview of Using Data: Definitions and Goals (Slide 4 of 5)

Table 2.1: Data for Dow Jones Industrial Index Companies (cont.)

Company	Symbol	Industry	Share Price ($)	Volume
Coca-Cola	KO	Food and Drink	46.52	7,699,367
McDonald's	MCD	Food and Drink	165.40	2,379,725
3M	MMM	Conglomerate	217.75	2,150,810
Merck	MRK	Pharmaceuticals	63.22	7,028,492
Microsoft	MSFT	Technology	77.59	16,823,989
Nike	NKE	Consumer Goods	52.00	9,492,675
Pfizer	PFE	Pharmaceuticals	36.20	14,019,661
Procter & Gamble	PG	Consumer Goods	92.80	5,316,062

Overview of Using Data: Definitions and Goals (Slide 5 of 5)

Table 2.1: Data for Dow Jones Industrial Index Companies (cont.)

Company	Symbol	Industry	Share Price ($)	Volume
Travelers	TRV	Insurance	128.62	1,808,224
UnitedHealth Group	UNH	Healthcare	203.89	8,949,715
United Technologies	UTX	Conglomerate	119.36	2,026,513
Visa	V	Financial	107.54	5,979,405
Verizon	VZ	Telecommunications	48.40	14,842,814
Wal-Mart	WMT	Retail	85.98	5,851,546
ExxonMobil	XOM	Chemical, Oil, and Gas	82.96	6,444,106

Types of Data

Population and Sample Data

Quantitative and Categorical Data

Cross-Sectional and Time Series Data

Sources of Data

Types of Data (Slide 1 of 5)

Population and Sample Data:

- **Population**: All elements of interest.
- **Sample**: Subset of the population.
  - **Random sampling**: A sampling method to gather a representative sample of the population data.

Quantitative and Categorical Data:

- **Quantitative data**: Data on which numeric and arithmetic operations, such as addition, subtraction, multiplication, and division, can be performed.
- **Categorical data**: Data on which arithmetic operations cannot be performed.

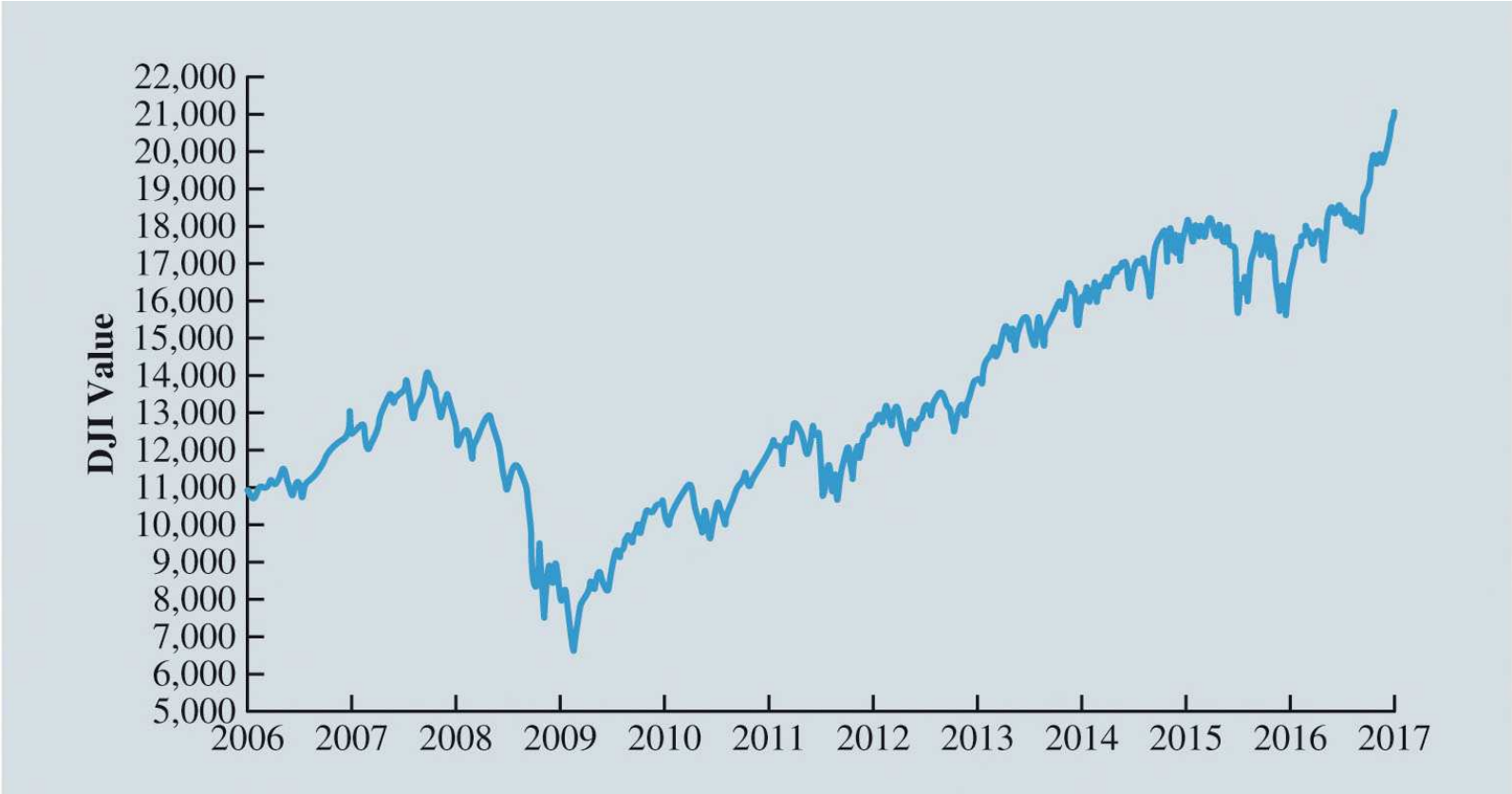
Types of Data (Slide 2 of 5)

Cross-Sectional and Time Series Data:

- **Cross-sectional data**: Data collected from several entities at the same, or approximately the same, point in time.
- **Time series data**: Data collected over several time periods.
  - Graphs of time series data are frequently found in business and economic publications.
  - Graphs help analysts understand what happened in the past, identify trends over time, and project future levels for the time series.

Types of Data (Slide 3 of 5)

Figure 2.1: Dow Jones Index Values Since 2006



Types of Data (Slide 4 of 5)

Sources of Data:

- Experimental study:
  - A variable of interest is first identified.
  - Then one or more other variables are identified and controlled or manipulated so that data can be obtained about how they influence the variable of interest.
- Nonexperimental study or observational study:
  - Makes no attempt to control the variables of interest.
  - A survey is perhaps the most common type of observational study.

Types of Data (Slide 5 of 5)

Figure 2.2: Customer Opinion Questionnaire Used by Chops City Grill Restaurant

Chops
CITY GRILL

Date: _____ Server Name: _____

Our customers are our top priority. Please take a moment to fill out our survey card, so we can better serve your needs. You may return this card to the front desk or return by mail. Thank you!

SERVICE SURVEY	Excellent	Good	Average	Fair	Poor
Overall Experience	❑	❑	❑	❑	❑
Greeting by Hostess	❑	❑	❑	❑	❑
Manager (Table Visit)	❑	❑	❑	❑	❑
Overall Service	❑	❑	❑	❑	❑
Professionalism	❑	❑	❑	❑	❑
Menu Knowledge	❑	❑	❑	❑	❑
Friendliness	❑	❑	❑	❑	❑
Wine Selection	❑	❑	❑	❑	❑
Menu Selection	❑	❑	❑	❑	❑
Food Quality	❑	❑	❑	❑	❑
Food Presentation	❑	❑	❑	❑	❑
Value for $ Spent	❑	❑	❑	❑	❑

What comments could you give us to improve our restaurant?

Thank you, we appreciate your comments. —The staff of Chops City Grill.

Modifying Data in Excel

Sorting and Filtering Data in Excel

Conditional Formatting of Data in Excel

Modifying Data in Excel (Slide 1 of 14)

Table 2.2: 20 Top-Selling Automobiles in United States in March 2011

Rank (by March 2011 Sales)	Manufacturer	Model	Sales (March 2011)	Sales (March 2010)
1	Honda	Accord	33,616	29,120
2	Nissan	Altima	32,289	24,649
3	Toyota	Camry	31,464	36,251
4	Honda	Civic	31,213	22,463
5	Toyota	Corolla/Matrix	30,234	29,623
6	Ford	Fusion	27,566	22,773
7	Hyundai	Sonata	22,894	18,935
8	Hyundai	Elantra	19,255	8,225

Modifying Data in Excel (Slide 2 of 14)

Table 2.2: 20 Top-Selling Automobiles in United States in March 2011 (cont.)

Rank (by March 2011 Sales)	Manufacturer	Model	Sales (March 2011)	Sales (March 2010)
9	Toyota	Prius	18,605	11,786
10	Chevrolet	Cruze/Cobalt	18,101	10,316
11	Chevrolet	Impala	18,063	15,594
12	Nissan	Sentra	17,851	8,721
13	Ford	Focus	17,178	19,500
14	Volkswagen	Jetta	16,969	9,196
15	Chevrolet	Malibu	15,551	17,750
16	Mazda	3	12,467	11,353

Modifying Data in Excel (Slide 3 of 14)

Table 2.2: 20 Top-Selling Automobiles in United States in March 2011 (cont.)

Rank (by March 2011 Sales)	Manufacturer	Model	Sales (March 2011)	Sales (March 2010)
17	Nissan	Versa	11,075	13,811
18	Subaru	Outback	10,498	7,619
19	Kia	Soul	10,028	5,106
20	Ford	Fiesta	9,787	0

Modifying Data in Excel (Slide 4 of 14)

Figure 2.3: Data for 20 Top-Selling Automobiles Entered into Excel with Percent Change in Sales from 2010

DATA file

Top20CarsPercent

	A	B	C	D	E	F
1	Rank (by March 2011 Sales)	Manufacturer	Model	Sales (March 2011)	Sales (March 2010)	Percent Change in Sales from 2010
2	1	Honda	Accord	33616	29120	15.4%
3	2	Nissan	Altima	32289	24649	31.0%
4	3	Toyota	Camry	31464	36251	–13.2%
5	4	Honda	Civic	31213	22463	39.0%
6	5	Toyota	Corolla/Matrix	30234	29623	2.1%
7	6	Ford	Fusion	27566	22773	21.0%
8	7	Hyundai	Sonata	22894	18935	20.9%
9	8	Hyundai	Elantra	19255	8225	134.1%
10	9	Toyota	Prius	18605	11786	57.9%
11	10	Chevrolet	Cruze/Cobalt	18101	10316	75.5%
12	11	Chevrolet	Impala	18063	15594	15.8%
13	12	Nissan	Sentra	17851	8721	104.7%
14	13	Ford	Focus	17178	19500	–11.9%
15	14	Volkswagen	Jetta	16969	9196	84.5%
16	15	Chevrolet	Malibu	15551	17750	–12.4%
17	16	Mazda	3	12467	11353	9.8%
18	17	Nissan	Versa	11075	13811	–19.8%
19	18	Subaru	Outback	10498	7619	37.8%
20	19	Kia	Soul	10028	5106	96.4%
21	20	Ford	Fiesta	9787	0	-----

Modifying Data in Excel (Slide 5 of 14)

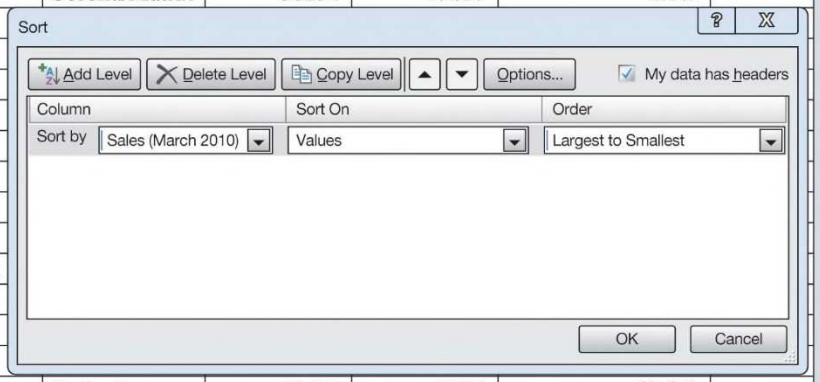
Sorting and Filtering Data in Excel:

- To sort the automobiles by March 2010 sales:
  - Step 1: Select cells A1:F21.
  - Step 2: Click the **Data** tab in the Ribbon.
  - Step 3: Click **Sort** in the **Sort & Filter** group.
  - Step 4: Select the check box for **My data has headers.**
  - Step 5: In the first **Sort by** dropdown menu, select **Sales (March 2010).**
  - Step 6: In the **Order** dropdown menu, select **Largest to Smallest.**
  - Step 7: Click **OK.**

Modifying Data in Excel (Slide 6 of 14)

Figure 2.4: Using Excel's Sort Function to Sort the Top-Selling Automobiles Data

	A	B	C	D	E	F	G
1	Rank (by March 2011 Sales)	Manufacturer	Model	Sales (March 2011)	Sales (March 2010)	Percent Change in Sales from 2010	
2	1	Honda	Accord	33616	29120	15.4%	
3	2	Nissan	Altima	32289	24649	31.0%	
4	3	Toyota	Camry	31464	36251	–13.2%	
5	4	Honda	Civic	31213	22463	39.0%	
6	5	Toyota	Corolla/Matrix	30234	29623	2.1%	
7	6	Ford					
8	7	Hyundai					
9	8	Hyundai					
10	9	Toyota					
11	10	Chevrolet					
12	11	Chevrolet					
13	12	Nissan					
14	13	Ford					
15	14	Volkswagen					
16	15	Chevrolet					
17	16	Mazda					
18	17	Nissan					
19	18	Subaru	Outback	10498	7619	37.8%	
20	19	Kia	Soul	10028	5106	96.4%	
21	20	Ford	Fiesta	9787	0	-----	



Modifying Data in Excel (Slide 7 of 14)

Figure 2.5: Top-Selling Automobiles Data Sorted by Sales in March 2010 Sales

	A	B	C	D	E	F
1	Rank (by March 2011 Sales)	Manufacturer	Model	Sales (March 2011)	Sales (March 2010)	Percent Change in Sales from 2010
2	3	Toyota	Camry	31464	36251	–13.2%
3	5	Toyota	Corolla/Matrix	30234	29623	2.1%
4	1	Honda	Accord	33616	29120	15.4%
5	2	Nissan	Altima	32289	24649	31.0%
6	6	Ford	Fusion	27566	22773	21.0%
7	4	Honda	Civic	31213	22463	39.0%
8	13	Ford	Focus	17178	19500	–11.9%
9	7	Hyundai	Sonata	22894	18935	20.9%
10	15	Chevrolet	Malibu	15551	17750	–12.4%
11	11	Chevrolet	Impala	18063	15594	15.8%
12	17	Nissan	Versa	11075	13811	–19.8%
13	9	Toyota	Prius	18605	11786	57.9%
14	16	Mazda	3	12467	11353	9.8%
15	10	Chevrolet	Cruze/Cobalt	18101	10316	75.5%
16	14	Volkswagen	Jetta	16969	9196	84.5%
17	12	Nissan	Sentra	17851	8721	104.7%
18	8	Hyundai	Elantra	19255	8225	134.1%
19	18	Subaru	Outback	10498	7619	37.8%
20	19	Kia	Soul	10028	5106	96.4%
21	20	Ford	Fiesta	9787	0	-----

Modifying Data in Excel (Slide 8 of 14)

Sorting and Filtering Data in Excel (cont.):

- Using Excel's Filter function to see the sales of models made by Toyota:
  - Step 1: Select cells A1:F21.
  - Step 2: Click the **Data** tab in the Ribbon.
  - Step 3: Click **Filter** in the **Sort & Filter** group.
  - Step 4: Click on the **Filter Arrow** in column B, next to **Manufacturer.**
  - Step 5: If all choices are checked, you can easily deselect all choices by unchecking (**Select All**). Then select only the check box for **Toyota**.
  - Step 6. Click **OK.**

Modifying Data in Excel (Slide 9 of 14)

Figure 2.6: Top Selling Automobiles Data Filtered to Show Only Automobiles Manufactured by Toyota

	A	B	C	D	E	F
1	Rank (by March 2011 Sales)	Manufacturer	Model	Sales (March 2011)	Sales (March 2010)	Percent Change in Sales from 2010
2	3	Toyota	Camry	31464	36251	–13.2%
3	5	Toyota	Corolla/Matrix	30234	29623	2.1%
13	9	Toyota	Prius	18605	11786	57.9%

Modifying Data in Excel (Slide 10 of 14)

Conditional Formatting of Data in Excel:

- Makes it easy to identify data that satisfy certain conditions in a data set.
- To identify the automobile models in Table 2.2 for which sales had decreased from March 2010 to March 2011:
  - Step 1: Starting with the original data shown in Figure 2.3, select cells F1:F21.
  - Step 2: Click on the **Home** tab in the Ribbon.
  - Step 3: Click **Conditional Formatting** in the **Styles** group.
  - Step 4: Select **Highlight Cells Rules**, and click **Less Than** from the dropdown menu.
  - Step 5: Enter *0%* in the **Format cells that are LESS THAN**: box.
  - Step 6: Click **OK.**

Modifying Data in Excel (Slide 11 of 14)

Figure 2.7: Using Conditional Formatting in Excel to Highlight Automobiles with Declining Sales from March 2010

	A	B	C	D	E	F
1	Rank (by March 2011 Sales)	Manufacturer	Model	Sales (March 2011)	Sales (March 2010)	Percent Change in Sales from 2010
2	1	Honda	Accord	33616	29120	15.4%
3	2	Nissan	Altima	32289	24649	31.0%
4	3	Toyota	Camry	31464	36251	–13.2%
5	4	Honda	Civic	31213	22463	39.0%
6	5	Toyota	Corolla/Matrix	30234	29623	2.1%
7	6	Ford	Fusion	27566	22773	21.0%
8	7	Hyundai	Sonata	22894	18935	20.9%
9	8	Hyundai	Elantra	19255	8225	134.1%
10	9	Toyota	Prius	18605	11786	57.9%
11	10	Chevrolet	Cruze/Cobalt	18101	10316	75.5%
12	11	Chevrolet	Impala	18063	15594	15.8%
13	12	Nissan	Sentra	17851	8721	104.7%
14	13	Ford	Focus	17178	19500	–11.9%
15	14	Volkswagen	Jetta	16969	9196	84.5%
16	15	Chevrolet	Malibu	15551	17750	–12.4%
17	16	Mazda	3	12467	11353	9.8%
18	17	Nissan	Versa	11075	13811	–19.8%
19	18	Subaru	Outback	10498	7619	37.8%
20	19	Kia	Soul	10028	5106	96.4%
21	20	Ford	Fiesta	9787	0	-----

Modifying Data in Excel (Slide 12 of 14)

Figure 2.8: Using Conditional Formatting in Excel to Generate Data Bars for the Top-Selling Automobiles Data

	A	B	C	D	E	F
1	Rank (by March 2011 Sales)	Manufacturer	Model	Sales (March 2011)	Sales (March 2010)	Percent Change in Sales from 2010
2	1	Honda	Accord	33616	29120	15.4%
3	2	Nissan	Altima	32289	24649	31.0%
4	3	Toyota	Camry	31464	36251	–13.2%
5	4	Honda	Civic	31213	22463	39.0%
6	5	Toyota	Corolla/Matrix	30234	29623	2.1%
7	6	Ford	Fusion	27566	22773	21.0%
8	7	Hyundai	Sonata	22894	18935	20.9%
9	8	Hyundai	Elantra	19255	8225	134.1%
10	9	Toyota	Prius	18605	11786	57.9%
11	10	Chevrolet	Cruze/Cobalt	18101	10316	75.5%
12	11	Chevrolet	Impala	18063	15594	15.8%
13	12	Nissan	Sentra	17851	8721	104.7%
14	13	Ford	Focus	17178	19500	–11.9%
15	14	Volkswagen	Jetta	16969	9196	84.5%
16	15	Chevrolet	Malibu	15551	17750	–12.4%
17	16	Mazda	3	12467	11353	9.8%
18	17	Nissan	Versa	11075	13811	–19.8%
19	18	Subaru	Outback	10498	7619	37.8%
20	19	Kia	Soul	10028	5106	96.4%
21	20	Ford	Fiesta	9787	0	-----

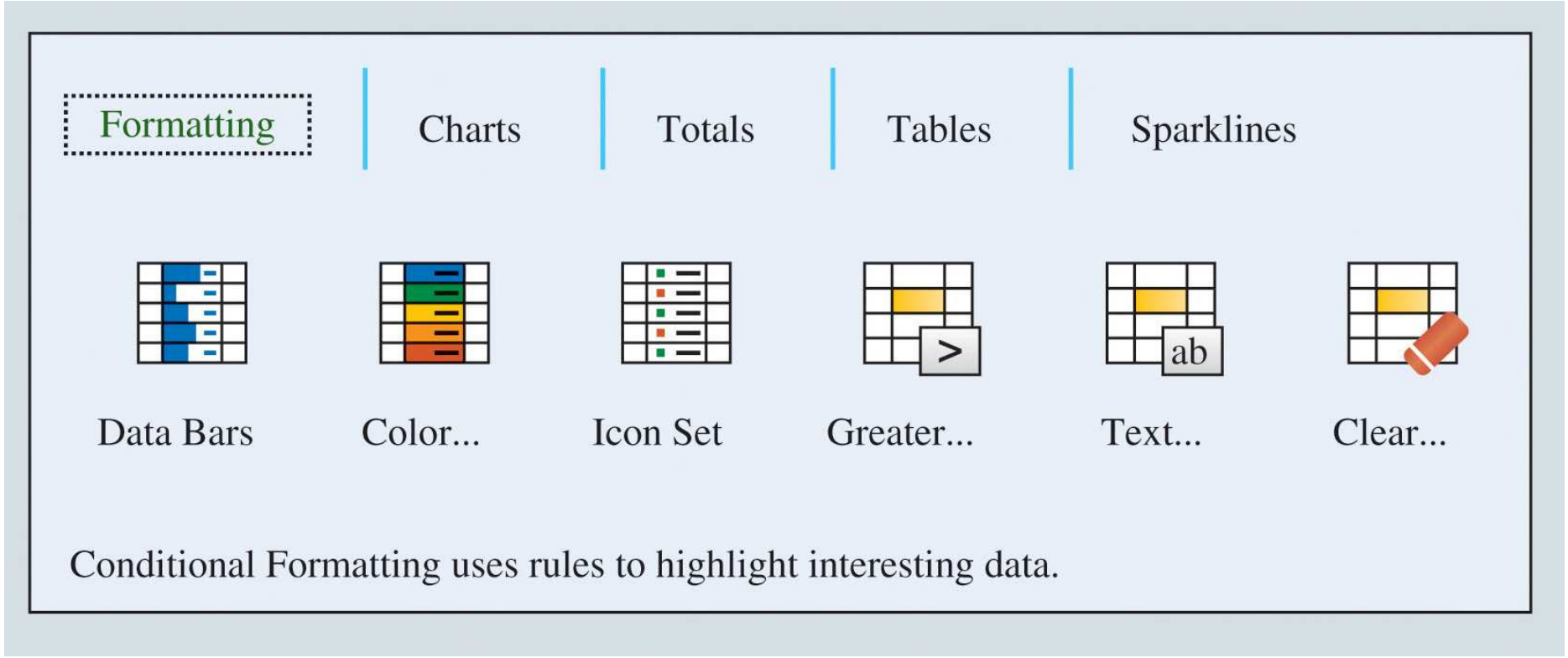
Modifying Data in Excel (Slide 13 of 14)

Conditional Formatting of Data in Excel (cont.):

- **Quick Analysis** button appears just outside the bottom-right corner of a group of selected cells.
- It provides shortcuts for Conditional Formatting, adding Data Bars, and other operations.

Modifying Data in Excel (Slide 14 of 14)

Figure 2.9 Excel Quick Analysis Button Formatting Options



Creating Distributions from Data

Frequency Distributions for Categorical Data

Relative Frequency and Percent Frequency Distributions

Frequency Distributions for Quantitative Data

Histograms

Cumulative Distributions

Creating Distributions from Data (Slide 1 of 18)

Frequency Distributions for Categorical Data:

- **Frequency distribution**: A summary of data that shows the number (frequency) of observations in each of several nonoverlapping classes.
- Typically referred to as **bins**, when dealing with distributions.

Creating Distributions from Data (Slide 2 of 18)

Table 2.3: Data from a Sample of 50 Soft Drink Purchases

Coca-Cola	Sprite	Pepsi
Diet Coke	Coca-Cola	Coca-Cola
Pepsi	Diet Coke	Coca-Cola
Diet Coke	Coca-Cola	Coca-Cola
Coca-Cola	Diet Coke	Pepsi
Coca-Cola	Coca-Cola	Dr. Pepper
Dr. Pepper	Sprite	Coca-Cola
Diet Coke	Pepsi	Diet Coke
Pepsi	Coca-Cola	Pepsi
Pepsi	Coca-Cola	Pepsi
Coca-Cola	Coca-Cola	Pepsi
Dr. Pepper	Pepsi	Pepsi
Sprite	Coca-Cola	Coca-Cola
Coca-Cola	Sprite	Dr. Pepper
Diet Coke	Dr. Pepper	Pepsi
Coca-Cola	Pepsi	Sprite
Coca-Cola	Diet Coke	

Creating Distributions from Data (Slide 3 of 18)

Table 2.4: Frequency Distribution of Soft Drink Purchases

Soft Drink	Frequency
Coca-Cola	19
Diet Coke	8
Dr. Pepper	5
Pepsi	13
Sprite	5
Total	50

- The frequency distribution summarizes information about the popularity of the five soft drinks:
  - Coca-Cola is the leader.
  - Pepsi is second.
  - Diet Coke is third.
  - Sprite and Dr. Pepper are tied for fourth.

Creating Distributions from Data (Slide 4 of 18)

Figure 2.10: Creating a Frequency Distribution for Soft Drinks Data in Excel

	A	B	C	D	E
1	**Sample Data**			**Bins**	
2	Coca-Cola	Coca-Cola		**Coca-Cola**	19
3	Diet Coke	Sprite		**Diet Coke**	8
4	Pepsi	Pepsi		**Dr. Pepper**	5
5	Diet Coke	Coca-Cola		**Pepsi**	13
6	Coca-Cola	Pepsi		**Sprite**	5
7	Coca-Cola	Sprite			
8	Dr. Pepper	Dr. Pepper			
9	Diet Coke	Pepsi			
10	Pepsi	Diet Coke			
11	Pepsi	Pepsi			
12	Coca-Cola	Coca-Cola			
13	Dr. Pepper	Coca-Cola			
14	Sprite	Diet Coke			
15	Coca-Cola	Pepsi			
16	Diet Coke	Pepsi			
17	Coca-Cola	Pepsi			
18	Coca-Cola	Coca-Cola			
19	Diet Coke	Dr. Pepper			
20	Coca-Cola	Sprite			
21	Coca-Cola	Coca-Cola			
22	Coca-Cola	Coca-Cola			
23	Sprite	Pepsi			
24	Coca-Cola	Dr. Pepper			
25	Coca-Cola	Pepsi			
26	Diet Coke	Pepsi			

Creating Distributions from Data (Slide 5 of 18)

Relative Frequency and Percent Frequency Distributions:

- **Relative frequency distribution**: A tabular summary of data showing the relative frequency for each bin.
- **Percent frequency distribution**: Summarizes the percent frequency of the data for each bin.
- Percent frequency distribution is used to provide estimates of the relative likelihoods of different values of a random variable.

Creating Distributions from Data (Slide 6 of 18)

Table 2.5: Relative Frequency and Percent Frequency Distributions of Soft Drink Purchases

Soft Drink	Relative Frequency	Percent Frequency (%)
Coca-Cola	0.38	38
Diet Coke	0.16	16
Dr. Pepper	0.10	10
Pepsi	0.26	26
Sprite	0.10	10
Total	1.00	100

Creating Distributions from Data (Slide 7 of 18)

Frequency Distributions for Quantitative Data:

- Three steps necessary to define the classes for a frequency distribution with quantitative data:
  1. Determine the number of nonoverlapping bins.
  2. Determine the width of each bin.
  3. Determine the bin limits.

APPROXIMATE BIN WIDTH

$$\frac{\text{Largest data value} - \text{smallest data value}}{\text{Number of bins}} \qquad (2.1)$$

Creating Distributions from Data (Slide 8 of 18)

Table 2.6: Year-End Audit Times (Days)

12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

Creating Distributions from Data (Slide 9 of 18)

Table 2.7: Frequency, Relative Frequency, and Percent Frequency Distributions for the Audit Time Data

Audit Times (days)	Frequency	Relative Frequency	Percent Frequency
10–14	4	0.20	20
15–19	8	0.40	40
20–24	5	0.25	25
25–29	2	0.10	10
30–34	1	0.05	5

Creating Distributions from Data (Slide 10 of 18)

Figure 2.11: Using Excel to Generate a Frequency Distribution for Audit Times Data

	A	B	C	D
1	Year-End Audit Times (in Days)			
2	12	14	19	18
3	15	15	18	17
4	20	27	22	23
5	22	21	33	28
6	14	18	16	13
7				
8				
9	Bin	Frequency		
10	14	=FREQUENCY(A2:D6,A10:A14)		
11	19	=FREQUENCY(A2:D6,A10:A14)		
12	24	=FREQUENCY(A2:D6,A10:A14)		
13	29	=FREQUENCY(A2:D6,A10:A14)		
14	34	=FREQUENCY(A2:D6,A10:A14)		

	A	B	C	D
1	Year-End Audit Times (in Days)			
2	12	14	19	18
3	15	15	18	17
4	20	27	22	23
5	22	21	33	28
6	14	18	16	13
7				
8				
9	Bin	Frequency		
10	14	4		
11	19	8		
12	24	5		
13	29	2		
14	34	1		

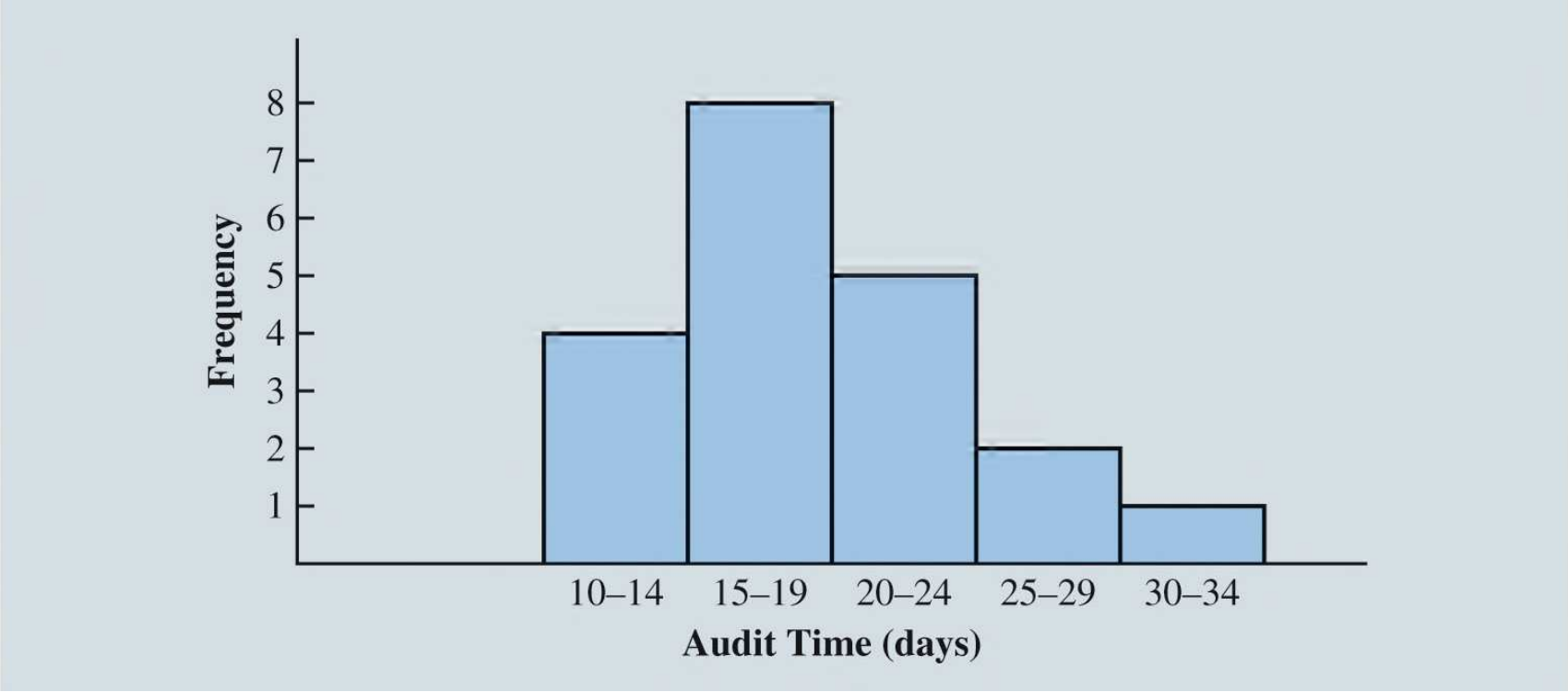
Creating Distributions from Data (Slide 11 of 18)

Histograms:

- **Histogram:** A common graphical presentation of quantitative data.
- Constructed by placing the variable of interest on the horizontal axis and the selected frequency measure (absolute frequency, relative frequency, or percent frequency) on the vertical axis.
- The frequency measure of each class is shown by drawing a rectangle whose base is the class limits on the horizontal axis and whose height is the corresponding frequency measure.

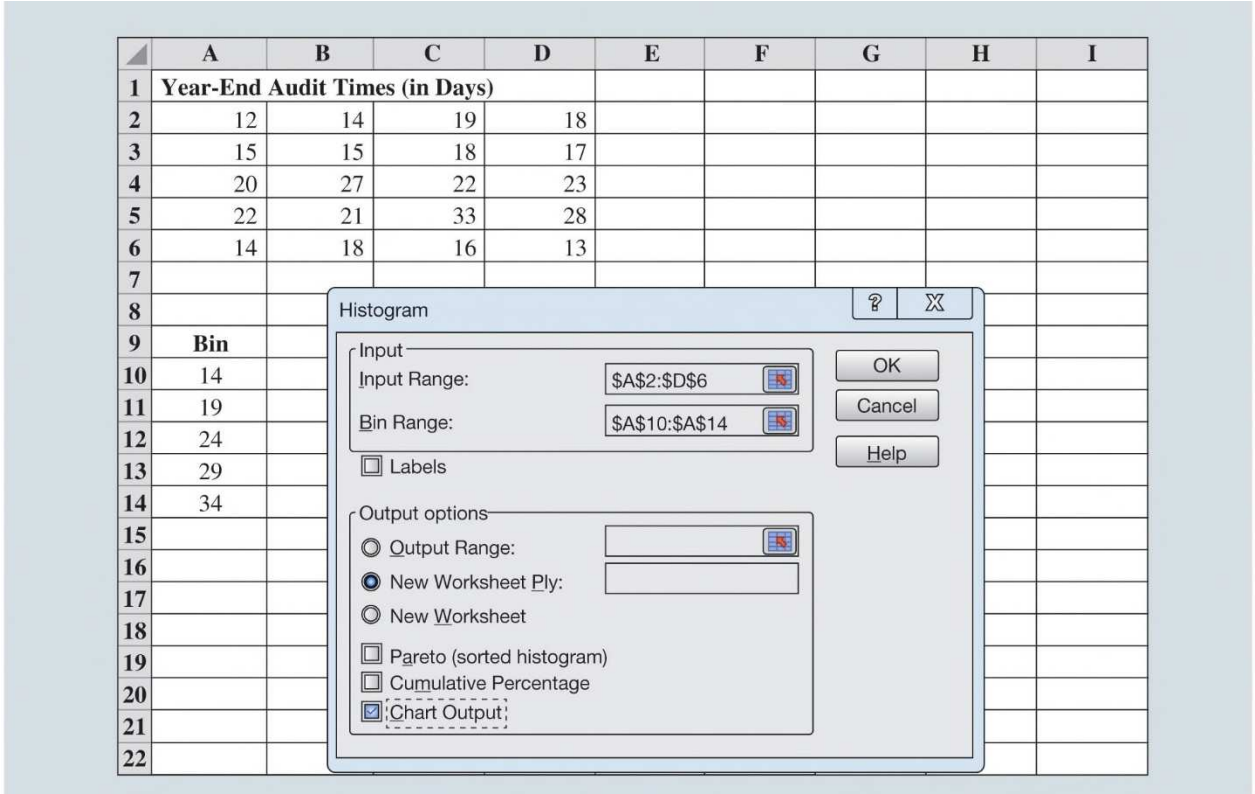
Creating Distributions from Data (Slide 12 of 18)

Figure 2.12: Histogram for the Audit Time Data



Creating Distributions from Data (Slide 13 of 18)

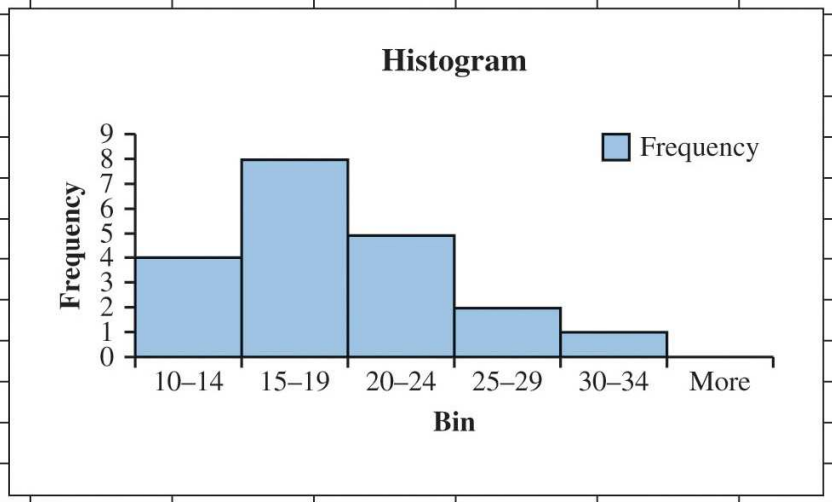
Figure 2.13: Creating a Histogram for the Audit Time Data Using Data Analysis Toolpak in Excel



Creating Distributions from Data (Slide 14 of 18)

Figure 2.14: Completed Histogram for the Audit Time Data Using Data Analysis ToolPak in Excel

	A	B	C	D	E	F	G	H	I	J
1	*Bin*	*Frequency*								
2	10–14	4								
3	15–19	8								
4	20–24	5								
5	25–29	2								
6	30–34	1								
7	More	0								
8										
9										
10										
11										
12										
13										
14										



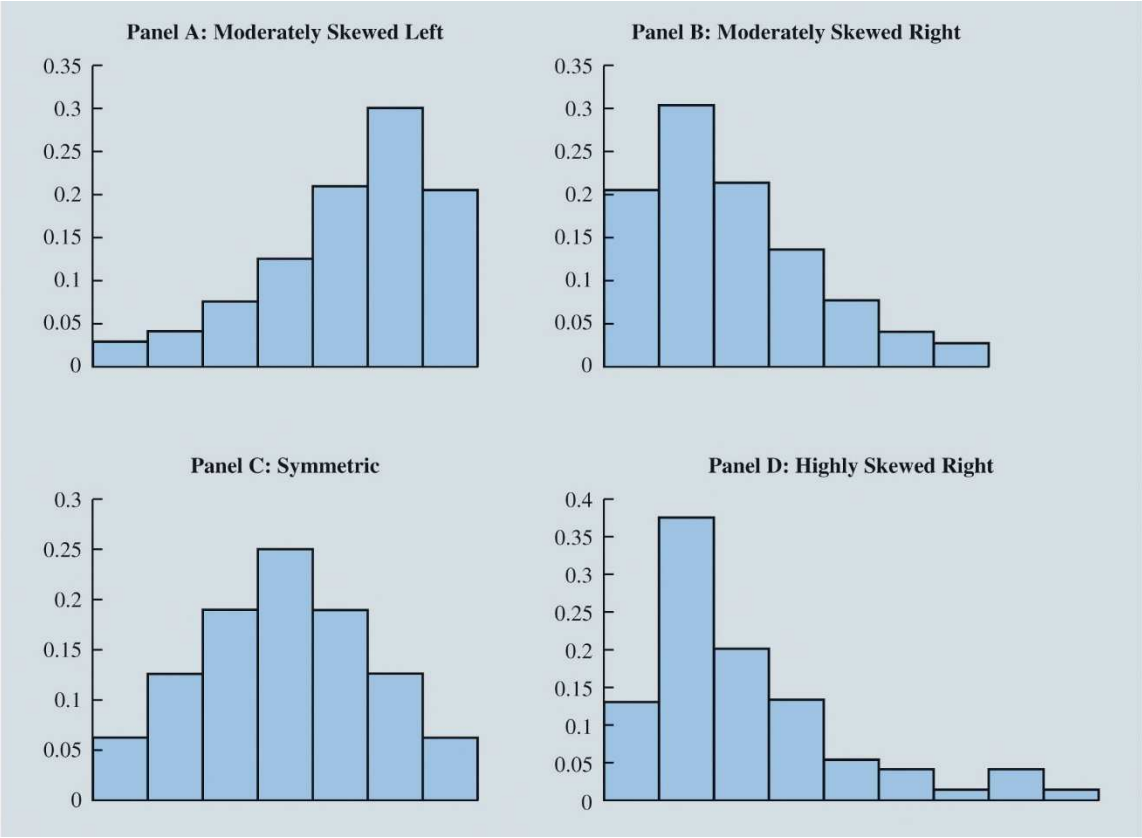
Creating Distributions from Data (Slide 15 of 18)

Histograms (cont.):

- Histograms provide information about the shape, or form, of a distribution.
- **Skewness**: Lack of symmetry.
- Skewness is an important characteristic of the shape of a distribution.

Creating Distributions from Data (Slide 16 of 18)

Figure 2.15: Histograms Showing Distributions with Different Levels of Skewness



Creating Distributions from Data (Slide 17 of 18)

Cumulative Distributions

- **Cumulative frequency distribution**: A variation of the frequency distribution that provides another tabular summary of quantitative data.
  - Uses the number of classes, class widths, and class limits developed for the frequency distribution.
  - Shows the number of data items with values less than or equal to the upper class limit of each class.

Creating Distributions from Data (Slide 18 of 18)

Table 2.8: Cumulative Frequency, Cumulative Relative Frequency, and Cumulative Percent Frequency Distributions for the Audit Time Data

Audit Time (days)	Cumulative Frequency	Cumulative Relative Frequency	Cumulative Percent Frequency
Less than or equal to 14	4	0.20	20
Less than or equal to 19	12	0.60	60
Less than or equal to 24	17	0.85	85
Less than or equal to 29	19	0.95	95
Less than or equal to 34	20	1.00	100

Measures of Location

Mean (Arithmetic Mean)

Median

Mode

Geometric Mean

Measures of Location (Slide 1 of 13)

Mean/Arithmetic Mean:

- Average value for a variable.
- The mean is denoted by $\bar{x}$.
- n = sample size.
- x_1 = value of variable *x* for the first observation.
- x_2 = value of variable *x* for the second observation.
- x_i = value of variable *x* for the ith observation.

SAMPLE MEAN

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad (2.2)$$

Measures of Location (Slide 2 of 13)

Table 2.9: Data on Home Sales in a Cincinnati, Ohio, Suburb

Home Sale	Selling Price ($)
1	138,000
2	254,000
3	186,000
4	257,500
5	108,000
6	254,000
7	138,000
8	298,000
9	199,500
10	208,000
11	142,000
12	456,250

Measures of Location (Slide 3 of 13)

Computation of Sample Mean:

- Illustration: Computation of the mean home selling price for the sample of 12 home sales.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \ldots + x_{12}}{12}$$

$$= \frac{138{,}000 + 254{,}000 + \ldots 456{,}250}{12}$$

$$= \frac{2{,}639{,}250}{12} = 219{,}937.50$$

Measures of Location (Slide 4 of 13)

Median:

- **Median:** Value in the middle when the data are arranged in ascending order.
- Middle value, for an odd number of observations.
- Average of two middle values, for an even number of observations.

Measures of Location (Slide 5 of 13)

Computation of Sample Median:

- Illustration: When the number of observations are odd,
  - Consider the class size data for a sample of five college classes:

46 54 42 46 32

  - Arrange the class size data in ascending order:

32 42 46 46 54

  - Middlemost value in the data set = 46.
  - Median is 46.

Measures of Location (Slide 6 of 13)

Computation of Sample Median:

Illustration: When the number of observations are even:

- Consider the data on home sales in Cincinnati, Ohio, Suburb (Table 2.9).
- Arrange the data in ascending order:

  108,000 138,000 138,000 142,000 186,000 199,500 208,000
  254,000 254,000 257,500 298,000 456,250

- Median = average of two middle values:

$$\text{Median} = \frac{199{,}500 + 208{,}000}{2} = 203{,}750$$

Measures of Location (Slide 7 of 13)

Mode:

- **Mode:** Value that occurs most frequently in a data set.
- Consider the class size data:

32 42 46 46 54

- Observe: 46 is the only value that occurs more than once.
- Mode is 46.
- Multimodal data: Data contain at least two modes.
- Bimodal data: Data contain exactly two modes.

Measures of Location (Slide 8 of 13)

Figure 2.16: Calculating the Mean, Median, and Modes for the Home Sales Data using Excel

	A	B	C	D	E
1	Home Sale	Selling Price ($)			
2	1	138,000		Mean:	=AVERAGE(B2:B13)
3	2	254,000		Median:	=MEDIAN(B2:B13)
4	3	186,000		Mode 1:	=MODE.MULT(B2:B13)
5	4	257,500		Mode 2:	=MODE.MULT(B2:B13)
6	5	108,000			
7	6	254,000			
8	7	138,000			
9	8	298,000			
10	9	199,500			
11	10	208,000			
12	11	142,000			
13	12	456,250			

	A	B	C	D	E
1	Home Sale	Selling Price ($)			
2	1	138,000		Mean:	$ 219,937.50
3	2	254,000		Median:	$ 203,750.00
4	3	186,000		Mode 1:	$ 138,000.00
5	4	257,500		Mode 2:	$ 254,000.00
6	5	108,000			
7	6	254,000			
8	7	138,000			
9	8	298,000			
10	9	199,500			
11	10	208,000			
12	11	142,000			
13	12	456,250			

Measures of Location (Slide 9 of 13)

Geometric Mean:

- **Geometric mean:** A measure of location that is calculated by finding the *n*th root of the product of *n* values
- Used in analyzing growth rates in financial data.
- Sample geometric mean:

SAMPLE GEOMETRIC MEAN

$$\bar{x}_g = \sqrt[n]{(x_1)(x_2)\cdots(x_n)} = [(x_1)(x_2)\cdots(x_n)]^{1/n} \qquad (2.3)$$

Measures of Location (Slide 10 of 13)

Table 2.10: Percentage Annual Returns and Growth Factors for the Mutual Fund Data:

- Illustration: Consider the percentage annual returns and **growth factors** for the mutual fund data over the past 10 years.
- We will determine the mean rate of growth for the fund over the 10-year period.

Measures of Location (Slide 11 of 13)

Table 2.10: Percentage Annual Returns and Growth Factors for the Mutual Fund Data (cont.)

Year	Return (%)	Growth Factor
1	−22.1	0.779
2	28.7	1.287
3	10.9	1.109
4	4.9	1.049
5	15.8	1.158
6	5.5	1.055
7	−37.0	0.630
8	26.5	1.265
9	15.1	1.151
10	2.1	1.021

Measures of Location (Slide 12 of 13)

Computation of Geometric Mean:

- Solution:
  - Product of the growth factors:

$$\$100\left[(0.779)(1.287)(1.109)\left(1.049\right)(1.158)(1.055)(0.630)(1.265)(1.151)(1.021)\right]$$

$$=\$100\left(1.335\right)=\$133.45$$

  - Geometric mean of the growth factors:

$$\overline{x}_g = \sqrt[10]{1.335} = 1.029.$$

  - Conclude that annual returns grew at an average annual rate of $\left(1.029-1\right)100\%$ or 2.9%.

Measures of Location (Slide 13 of 13)

Figure 2.17: Calculating the Geometric Mean for the Mutual Fund Data Using Excel

	A	B	C	D
1	Year	Return (%)	Growth Factor	
2	1	–22.1	0.779	
3	2	28.7	1.287	
4	3	10.9	1.109	
5	4	4.9	1.049	
6	5	15.8	1.158	
7	6	5.5	1.055	
8	7	–37.0	0.630	
9	8	26.5	1.265	
10	9	15.1	1.151	
11	10	2.1	1.021	
12				
13	Geometric Mean:		1.029	
14				

Measures of Variability

Range

Variance

Standard Deviation

Coefficient of Variation

Measures of Variability (Slide 1 of 10)

Table 2.11: Annual Payouts for Two Different Investment Funds

Year	Fund A ($)	Fund B ($)
1	1,100	700
2	1,100	2,500
3	1,100	1,200
4	1,100	1,550
5	1,100	1,300
6	1,100	800
7	1,100	300
8	1,100	1,600
9	1,100	1,500
10	1,100	350
11	1,100	460

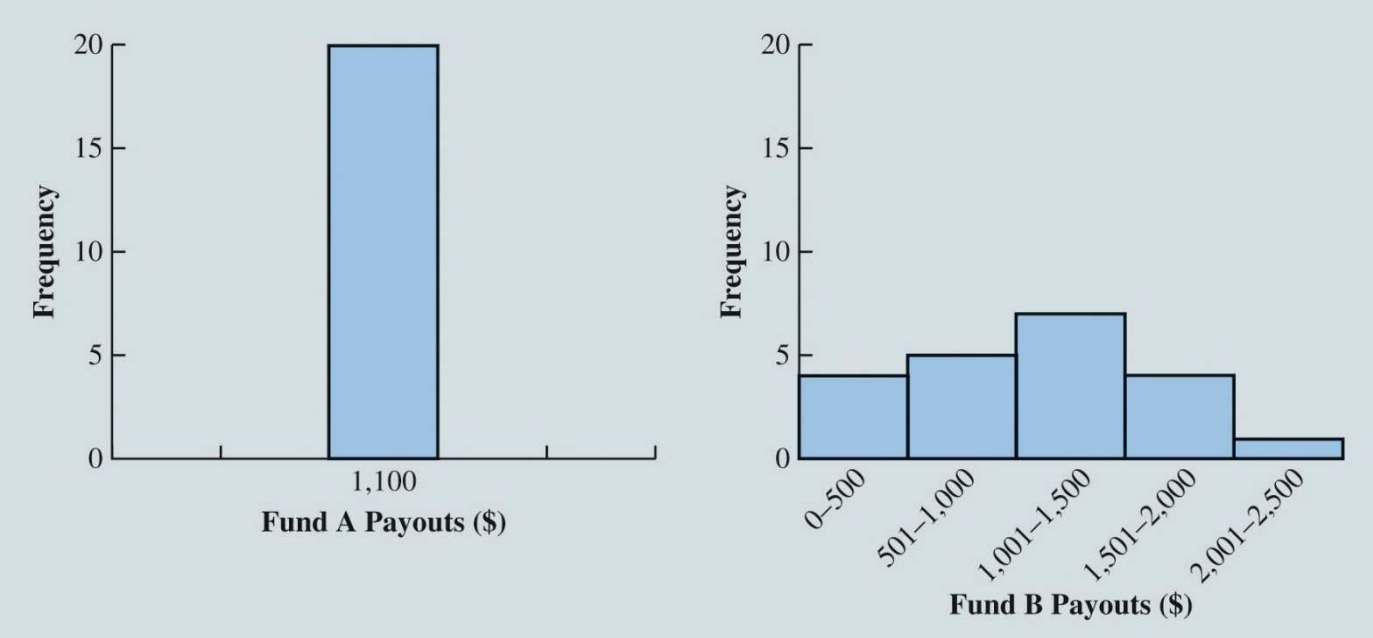
Measures of Variability (Slide 2 of 10)

Table 2.11: Annual Payouts for Two Different Investment Funds (cont.)

Year	Fund A ($)	Fund B ($)
12	1,100	890
13	1,100	1,050
14	1,100	800
15	1,100	1,150
16	1,100	1,200
17	1,100	1,800
18	1,100	100
19	1,100	1,750
20	1,100	1,000
Mean	1,100	1,100

Measures of Variability (Slide 3 of 10)

Figure 2.18: Histograms for Payouts of Past 20 Years from Fund A and Fund B



Measures of Variability (Slide 4 of 10)

Computation of Range:

Range:

- The **range** can be found by subtracting the smallest value from the largest value in a data set.
- Illustration: Consider the data on home sales in a Cincinnati, Ohio, suburb.
  - Largest home sales price: $456,250.
  - Smallest home sales price: $108,000.

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$
$$= \$456,250 - \$108,000$$
$$= \$348,250$$

- Drawback: Range is based on only two of the observations and thus is highly influenced by extreme values.

Measures of Variability (Slide 5 of 10)

Variance:

- **Variance** is a measure of variability that utilizes all the data.
- It is based on the deviation about the mean, which is the difference between the value of each observation (x_i) and the mean.
- The deviations about the mean are squared while computing the variance.

SAMPLE VARIANCE

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} \qquad (2.4)$$

- Population variance: $\sigma^2 = \dfrac{\Sigma\left(x_i - \mu\right)^2}{N}$.

Measures of Variability (Slide 6 of 10)

Table 2.12: Computation of Deviations and Squared Deviations About the Mean for the Class Size Data

Number of Students in Class (x_i)	Mean Class Size ($\bar{x}$)	Deviation About the Mean ($x_i - \bar{x}$)	Squared Deviation About the Mean $(x_i - \bar{x})^2$
46	44	2	4
54	44	10	100
42	44	22	4
46	44	2	4
32	44	−12	144
		0	256
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

- Computation of Sample Variance: $s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{256}{4} = 64.$

Measures of Variability (Slide 7 of 10)

Standard Deviation:

- **Standard deviation** is the positive square root of the variance.
- Measured in the same units as the original data.

SAMPLE STANDARD DEVIATION

$$s = \sqrt{s^2} \qquad (2.5)$$

- For population, $\sigma = \sqrt{\sigma^2}$.

Measures of Variability (Slide 8 of 10)

Figure 2.19: Calculating Variability Measures for the Home Sales Data in Excel

	A	B	C	D	E
1	Home Sale	Selling Price ($)			
2	1	138000		Mean:	=AVERAGE(B2:B13)
3	2	254000		Median:	=MEDIAN(B2:B13)
4	3	186000		Mode 1:	=MODE.MULT(B2:B13)
5	4	257500		Mode 2:	=MODE.MULT(B2:B13)
6	5	108000			
7	6	254000		Range:	=MAX(B2:B13)-MIN(B2:B13)
8	7	138000		Variance:	=VAR.S(B2:B13)
9	8	298000		Standard Deviation:	=STDEV.S(B2:B13)
10	9	199500			
11	10	208000		Coefficient of Variation:	=E9/E2
12	11	142000			
13	12	456250		85th Percentile:	=PERCENTILE.EXC(B2:B13,0.85)
14					
15				1st Quartile:	=QUARTILE. EXC(B2:B13,1)
16				2nd Quartile:	=QUARTILE. EXC(B2:B13,2)
17				3rd Quartile:	=QUARTILE. EXC(B2:B13,3)
18					
19				IQR:	=E17-E15

	A	B	C	D	E
1	Home Sale	Selling Price ($)			
2	1	138000		Mean:	$ 219,937.50
3	2	254000		Median:	$ 203,750.00
4	3	186000		Mode 1:	$ 138,000.00
5	4	257500		Mode 2:	$ 254,000.00
6	5	108000			
7	6	254000		Range:	$ 348,250.00
8	7	138000		Variance:	9037501420
9	8	298000		Standard Deviation:	$ 95,065.77
10	9	199500			
11	10	208000		Coefficient of Variation:	43.22%
12	11	142000			
13	12	456250		85th Percentile:	$ 305,912.50
14					
15				1st Quartile:	$ 139,000.00
16				2nd Quartile:	$ 203,750.00
17				3rd Quartile:	$ 256,625.00
18					
19				IQR:	$ 117,625.00

Measures of Variability (Slide 9 of 10)

Coefficient of Variation:

- The **coefficient of variation** is a descriptive statistic that indicates how large the standard deviation is relative to the mean.
- Expressed as a percentage.

COEFFICIENT OF VARIATION

$$\left(\frac{\text{Standard deviation}}{\text{Mean}} \times 100\right)\% \qquad (2.6)$$

Measures of Variability (Slide 10 of 10)

Computation of Coefficient of Variation:

Illustration:

- Consider the class size data:

  46 54 42 46 32

- Mean, $\bar{x} = 44.$
- Standard deviation, $s = 8.$
- Coefficient of variation = $\left(\frac{8}{44} \times 100 \right)\% = 18.2\%.$

Analyzing Distributions

Percentiles

Quartiles

z-Scores

Empirical Rule

Identifying Outliers

Box Plots

Analyzing Distributions (Slide 1 of 15)

Percentiles:

- A **percentile** is the value of a variable at which a specified (approximate) percentage of observations are below that value.
- The *p*th percentile tells us the point in the data where:
  - Approximately *p* percent of the observations have values less than the *p*th percentile.
  - Approximately $(100 - p)$ percent of the observations have values greater than the *p*th percentile.

Location of the *p*th Percentile

$$L_p = \frac{p}{100}(n + 1) \qquad (2.7)$$

Analyzing Distributions (Slide 2 of 15)

Illustration:

- To determine the 85th percentile for the home sales data in Table 2.9:

1. Arrange the data in ascending order:

   108,000 138,000 138,000 142,000 186,000 199,500
   208,000 254,000 254,000 257,500 298,000 456,250

2. Compute $L_{85} = \frac{p}{100}(n+1) = \left(\frac{85}{100}\right)(12+1) = 11.05.$

3. The interpretation of $L_{85} = 11.05$ is that the 85th percentile is 5% of the way between the value in position 11 and value in position 12.

Analyzing Distributions (Slide 3 of 15)

Illustration (cont.):

- To determine the 85th percentile for the home sales data in Table 2.9.
  - The value in the 11th position is 298,000.
  - The value in the 12th position is 456,250.
  - $305,912.50 represents the 85th percentile of the home sales data:

$$\begin{aligned} 85\text{th percentile} &= 298{,}000 + 0.05(456{,}250 - 298{,}000) \\ &= 298{,}000 + 0.05(158{,}250) \\ &= 305{,}912.50 \end{aligned}$$

Analyzing Distributions (Slide 4 of 15)

Quartiles:

- **Quartiles:** When the data is divided into four equal parts:
  - Each part contains approximately 25% of the observations.
  - Division points are referred to as quartiles.

Q_1 = first quartile, or 25th percentile.

Q_2 = second quartile, or 50th percentile (also the median).

Q_3 = third quartile or 75th percentile.

- The difference between the third and first quartiles is often referred to as the **interquartile range**, or IQR.

Analyzing Distributions (Slide 5 of 15)

z-Scores:

- The **z-score** measures the relative location of a value in the data set.
- Helps to determine how far a particular value is from the mean relative to the data set's standard deviation.
- Often called the standardized value.

Analyzing Distributions (Slide 6 of 15)

z-Scores (cont.):

- If $x_1, x_2, \ldots, x_n$ is a sample of n observations:

z-SCORE

$$z_i = \frac{x_i - \bar{x}}{s} \qquad (2.8)$$

where

z_i = the z-score for x_i
$\bar{x}$ = the sample mean
s = the sample standard deviation

Analyzing Distributions (Slide 7 of 15)

Table 2.13: *z*-Scores for the Class Size Data

No. of Students in Class (x_i)	Deviation About the Mean ($x_i - \bar{x}$)	z-Score $\left(\frac{x_i - \bar{x}}{s}\right)$
46	2	2/8 = 0.25
54	10	10/8 = 1.25
42	–2	−2/8 = −0.25
46	2	2/8 = 0.25
32	–12	−12/8 = −1.50

- For class size data, $\bar{x} = 44$ and $s = 8$.
  - For observations with a value $>$ mean, z-score > 0.
  - For observations with a value $<$ mean, z-score < 0.

Analyzing Distributions (Slide 8 of 15)

Figure 2.20: Calculating *z*-Scores for the Home Sales Data in Excel

	A	B	C
1	Home Sale	Selling Price ($)	*z*-Score
2	1	138000	=STANDARDIZE(B2,B15,B16)
3	2	254000	=STANDARDIZE(B3,B15,B16)
4	3	186000	=STANDARDIZE(B4,B15,B16)
5	4	257500	=STANDARDIZE(B5,B15,B16)
6	5	108000	=STANDARDIZE(B6,B15,B16)
7	6	254000	=STANDARDIZE(B7,B15,B16)
8	7	138000	=STANDARDIZE(B8,B15,B16)
9	8	298000	=STANDARDIZE(B9,B15,B16)
10	9	199500	=STANDARDIZE(B10,B15,B16)
11	10	208000	=STANDARDIZE(B11,B15,B16)
12	11	142000	=STANDARDIZE(B12,B15,B16)
13	12	456250	=STANDARDIZE(B13,B15,B16)
14			
15	Mean:	=AVERAGE(B2:B13)	
16	Standard Deviation:	=STDEV.S(B2:B13)	

	A	B	C
1	Home Sale	Selling Price ($)	*z*-Score
2	1	138,000	–0.862
3	2	254,000	0.358
4	3	186,000	–0.357
5	4	257,500	0.395
6	5	108,000	–1.177
7	6	254,000	0.358
8	7	138,000	–0.862
9	8	298,000	0.821
10	9	199,500	–0.215
11	10	208,000	–0.126
12	11	142,000	–0.820
13	12	456,250	2.486
14			
15	Mean:	$ 219,937.50	
16	Standard Deviation:	$ 95,065.77	

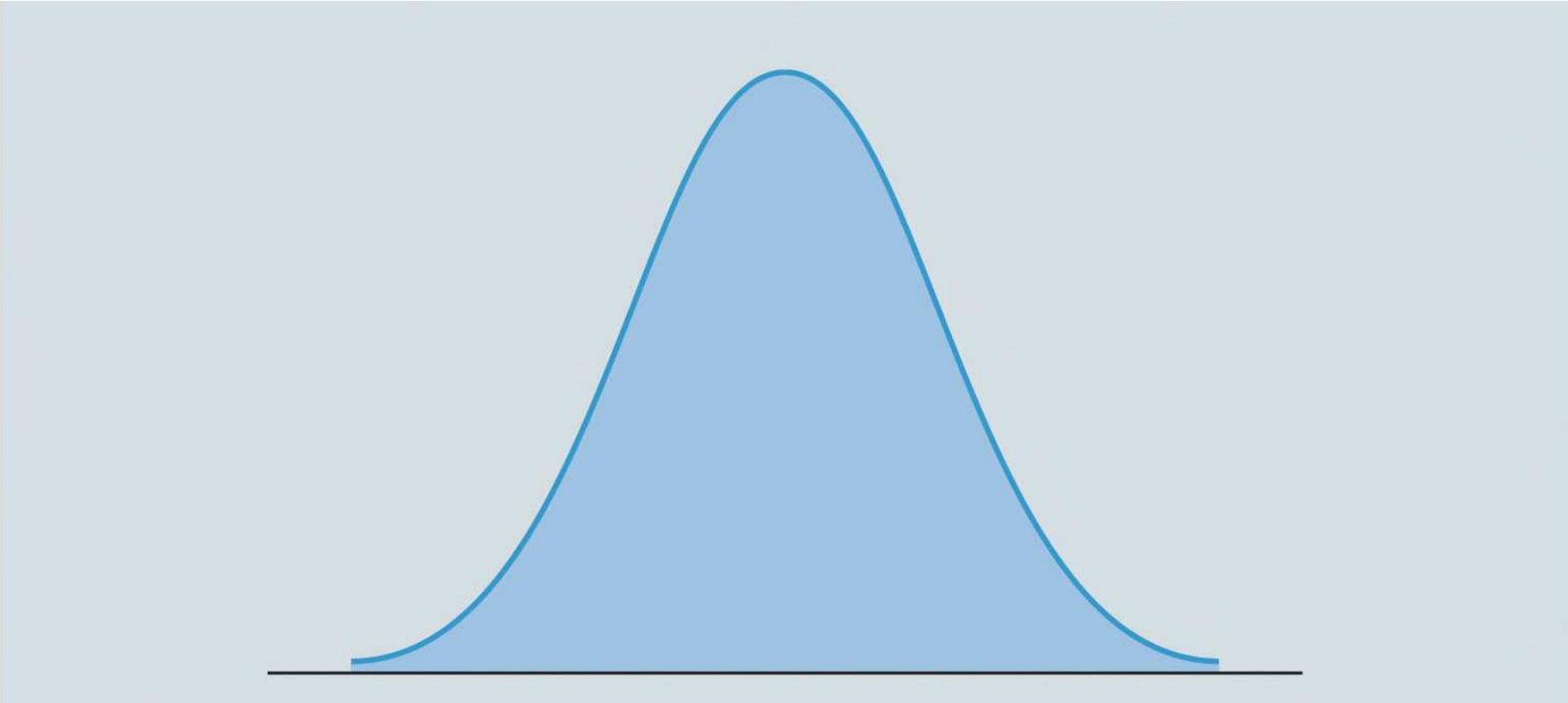
Analyzing Distributions (Slide 9 of 15)

Empirical Rule:

- When the distribution of data exhibits a symmetric bell-shaped distribution (as shown in Figure 2.21), the **empirical rule** can be used to determine the percentage of data values that are within a specified number of standard deviations of the mean.
- For data having a bell-shaped distribution:
  - Approximately 68% of the data values will be within 1 standard deviation.
  - Approximately 95% of the data values will be within 2 standard deviations.
  - Almost all the data values will be within 3 standard deviations.

Analyzing Distributions (Slide 10 of 15)

Figure 2.21: A Symmetric Bell-Shaped Distribution



Analyzing Distributions (Slide 11 of 15)

Identifying Outliers:

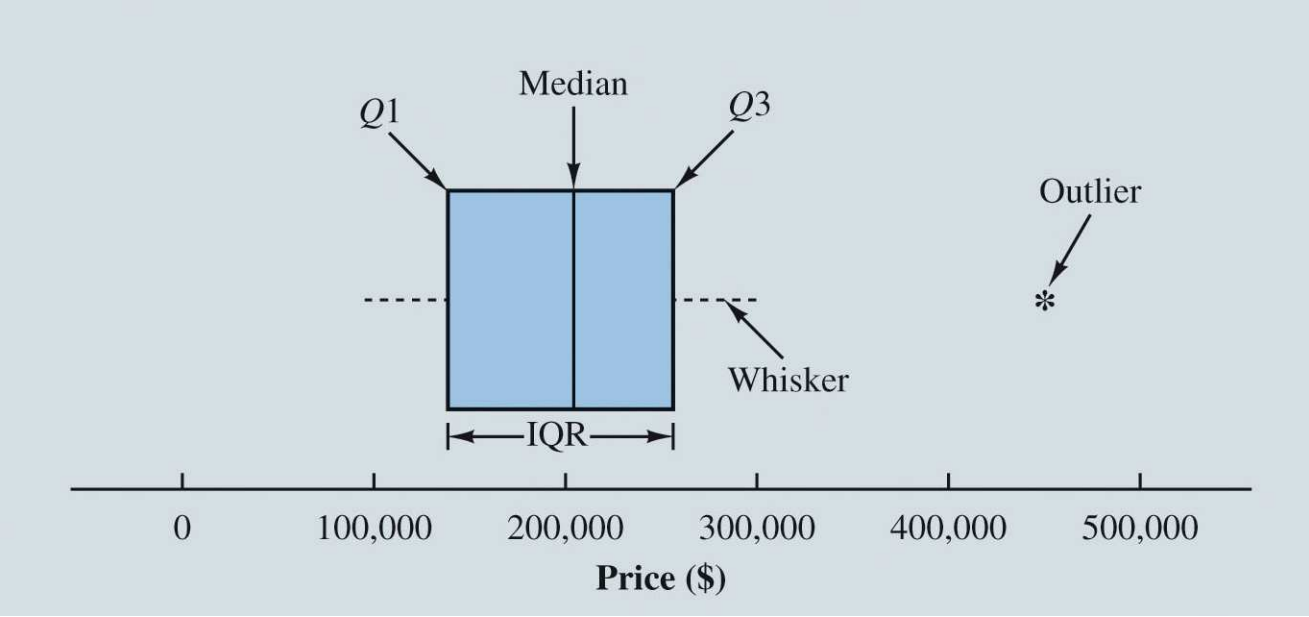
- **Outliers**: Extreme values in a data set.
- They can be identified using standardized values (*z*-scores).
- Any data value with a *z*-score less than –3 or greater than +3 is an outlier.
- Such data values can then be reviewed to determine their accuracy and whether they belong in the data set.

Analyzing Distributions (Slide 12 of 15)

Box Plots:

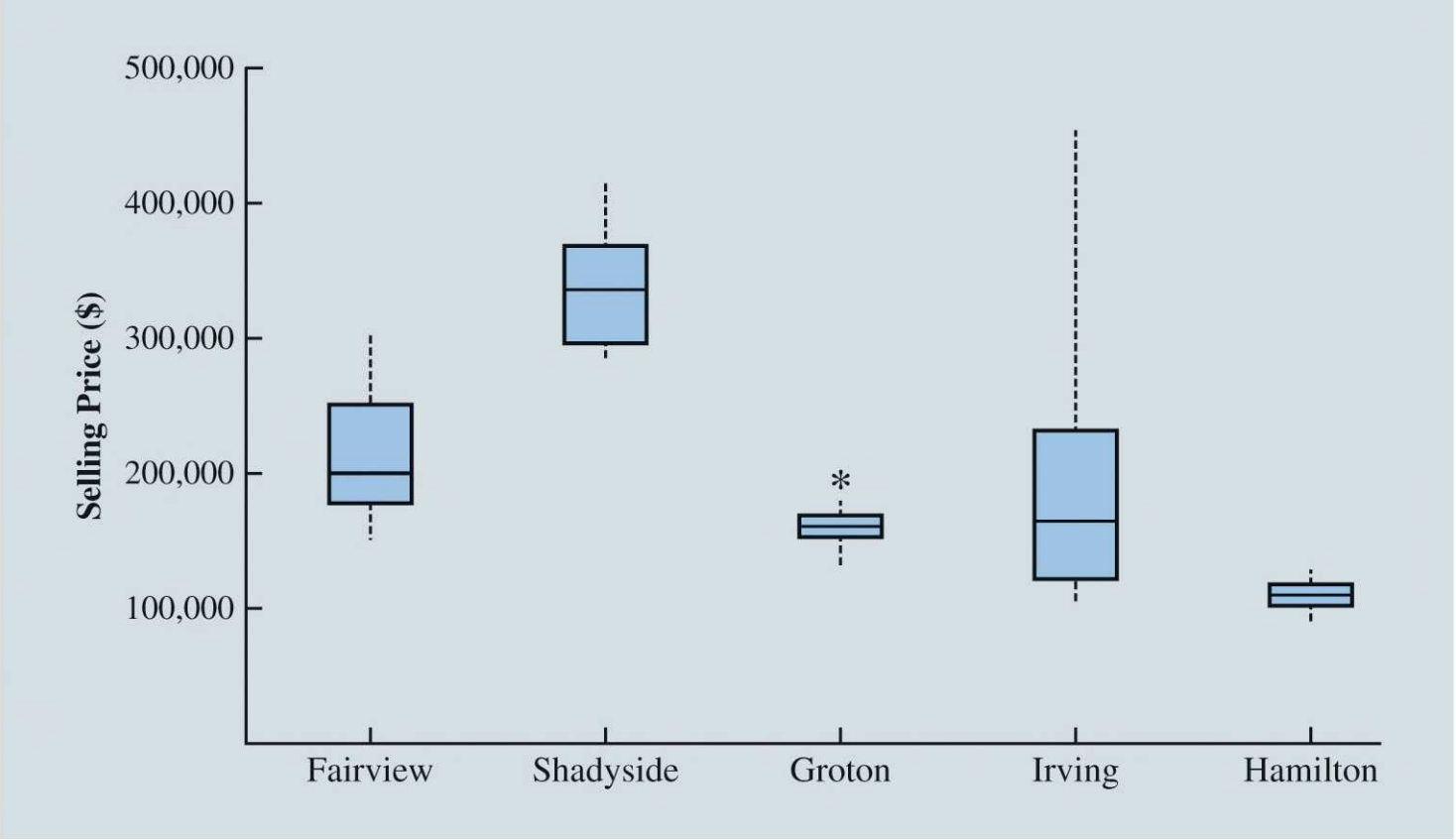
- A **box plot** is a graphical summary of the distribution of data.
- Developed from the quartiles for a data set.

Figure 2.22: Box Plot for the Home Sales Data



Analyzing Distributions (Slide 13 of 15)

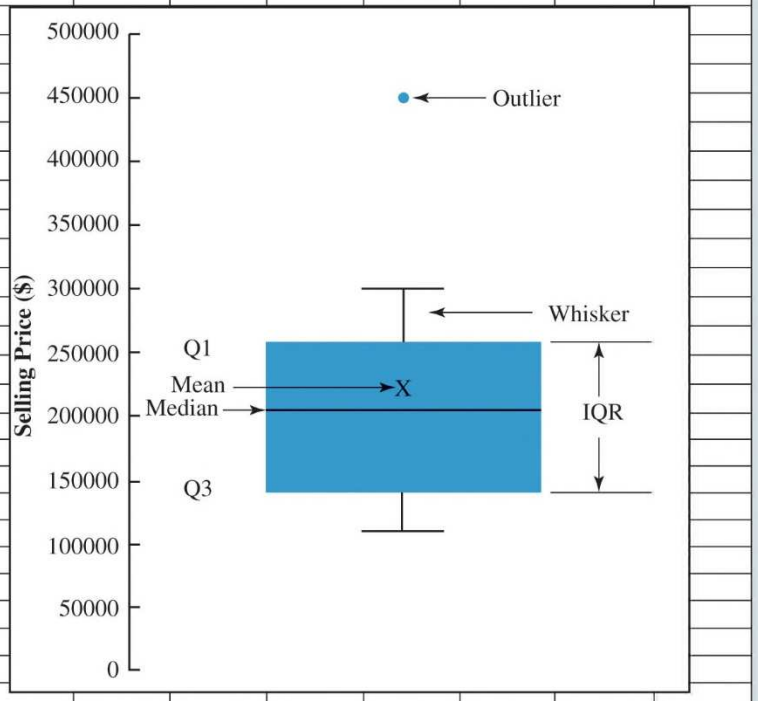
Figure 2.23: Box Plots Comparing Home Sale Prices in Different Communities



Analyzing Distributions (Slide 14 of 15)

Figure 2.24: Box Plot Created in Excel for Home Sales Data

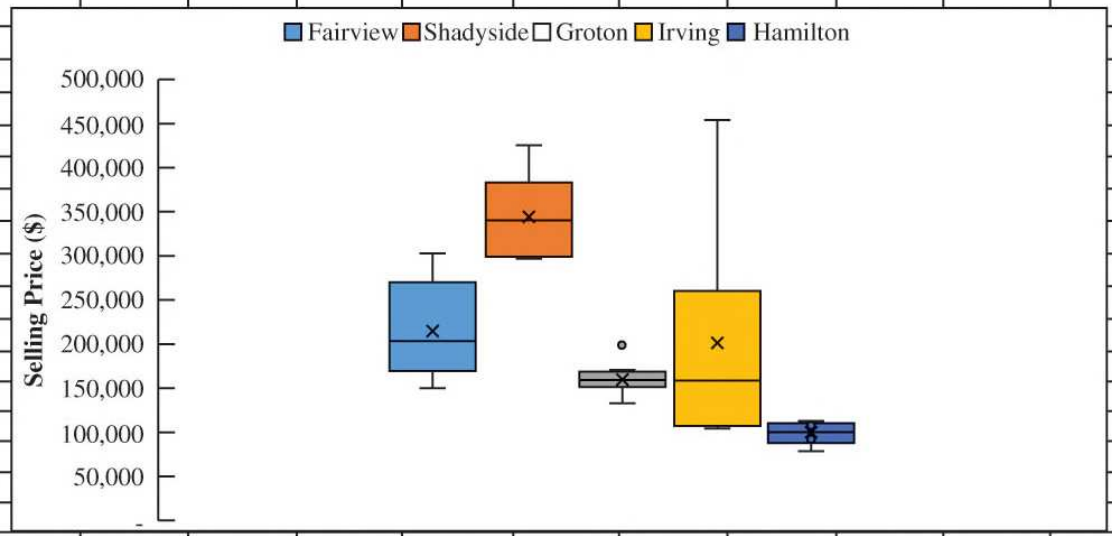
	A	B
1	Home Sale	Selling Price ($)
2	1	138000
3	2	254000
4	3	186000
5	4	257500
6	5	108000
7	6	254000
8	7	138000
9	8	298000
10	9	199500
11	10	208000
12	11	142000
13	12	456250



Analyzing Distributions (Slide 15 of 15)

Figure 2.25: Box Plots for Multiple Variables Created in Excel

	A	B	C	D	E	F
1		Fairview	Shadyside	Groton	Irving	Hamilton
2	Selling Prices	302,000	336,000	152,000	201,000	102,000
3		265,000	398,000	158,000	365,000	108,000
4		280,000	378,000	154,000	115,000	88,000
5		220,000	298,000	170,000	105,000	111,000
6		149,000	425,000	132,000	225,000	105,000
7		155,000	344,000	164,000	115,000	87,000
8		198,000	302,000	198,000	108,000	95,000
9		187,000	300,000	158,000	218,000	111,000
10		208,000	298,000	149,000	454,000	98,000
11		174,000	342,000	165,000	103,000	78,000
12						
13						
14						
15						
16						
17						
18						



Measures of Association Between Two Variables

Scatter Charts

Covariance

Correlation Coefficient

Measures of Association Between Two Variables (Slide 1 of 11)

Table 2.14: Data for Bottled Water Sales at Queensland Amusement Park for a Sample of 14 Summer Days

High Temperature (°F)	Bottled Water Sales (cases)
78	23
79	22
80	24
80	22
82	24
83	26
85	27
86	25
87	28
87	26
88	29
88	30
90	31
92	31

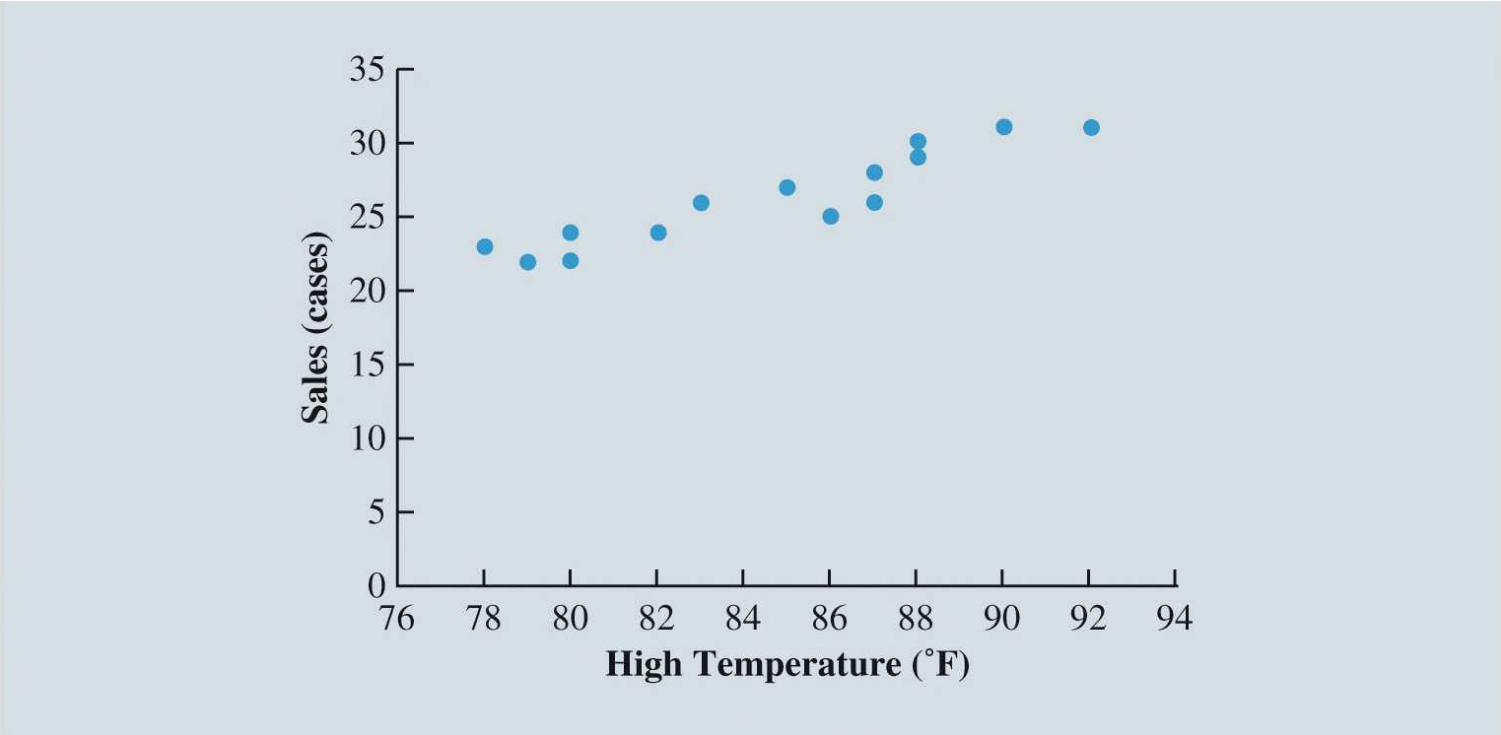
Measures of Association Between Two Variables (Slide 2 of 11)

Scatter Charts:

- A **scatter chart** is a useful graph for analyzing the relationship between two variables.
- The scatter chart in Figure 2.26 is an example of a positive relationship, because when one variable (high temperature) increases, the other variable (sales of bottled water) generally also increases.
- The scatter chart also suggests that a straight line could be used as an approximation for the relationship between high temperature and sales of bottled water.

Measures of Association Between Two Variables (Slide 3 of 11)

Figure 2.26: Chart Showing the Positive Linear Relation Between Sales and High Temperatures



Measures of Association Between Two Variables (Slide 4 of 11)

Covariance:

- **Covariance** is a descriptive measure of the linear association between two variables:

Sample Covariance

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \quad (2.9)$$

- Population covariance, $\sigma_{xy} = \frac{\Sigma(x_i - \mu_x)\Sigma(y_i - \mu_y)}{N}$.

Measures of Association Between Two Variables (Slide 5 of 11)

Table 2.15: Sample Covariance Calculations for Daily High Temperature and Bottled Water Sales at Queensland Amusement Park

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	78	23	−6.6	−3.3	21.78
	79	22	−5.6	−4.3	24.08
	80	24	−4.6	−2.3	10.58
	80	22	−4.6	−4.3	19.78
	82	24	−2.6	−2.3	5.98
	83	26	−1.6	−0.3	0.48
	85	27	0.4	0.7	0.28
	86	25	1.4	−1.3	−1.82
	87	28	2.4	1.7	4.08
	87	26	2.4	−0.3	−0.72
	88	29	3.4	2.7	9.18
	88	30	3.4	3.7	12.58
	90	31	5.4	4.7	25.38
	92	31	7.4	4.7	34.78
Totals	1,185	368	0.6	−0.2	166.42

$\bar{x} = 84.6$

$\bar{y} = 26.3$

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{166.42}{14 - 1} = 12.8$$

Measures of Association Between Two Variables (Slide 6 of 11)

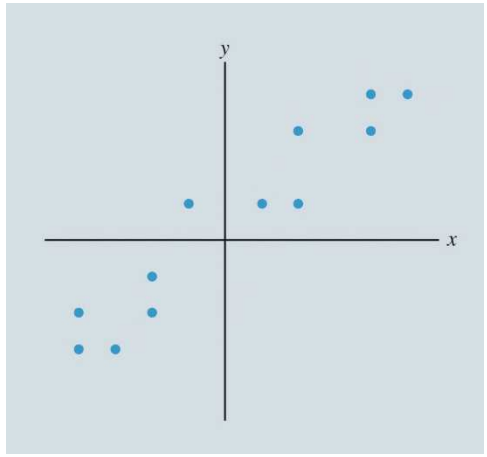
Figure 2.27: Calculating Covariance and Correlation Coefficient for Bottled Water Sales Using Excel

	A	B
1	High Temperature (°F)	Bottled Water Sales (cases)
2	78	23
3	79	22
4	80	24
5	80	22
6	82	24
7	83	26
8	85	27
9	86	25
10	87	28
11	87	26
12	88	29
13	88	30
14	90	31
15	92	31
16		
17	Covariance:	=COVARIANCE.S(A2:A15,B2:B15)
18	Correlation Coefficient:	=CORREL(A2:A15,B2:B15)

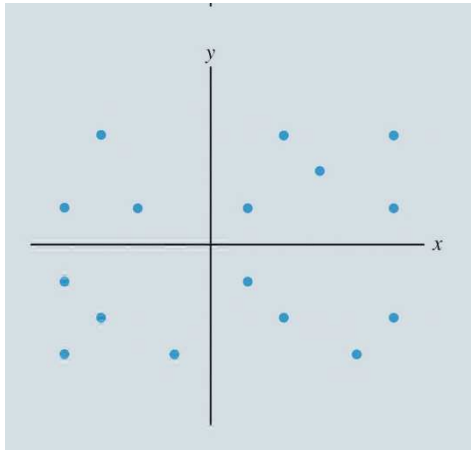
	A	B
1	High Temperature (°F)	Bottled Water Sales (cases)
2	78	23
3	79	22
4	80	24
5	80	22
6	82	24
7	83	26
8	85	27
9	86	25
10	87	28
11	87	26
12	88	29
13	88	30
14	90	31
15	92	31
16		
17	Covariance:	12.80
18	Correlation Coefficient:	0.93

Measures of Association Between Two Variables (Slide 7 of 11)

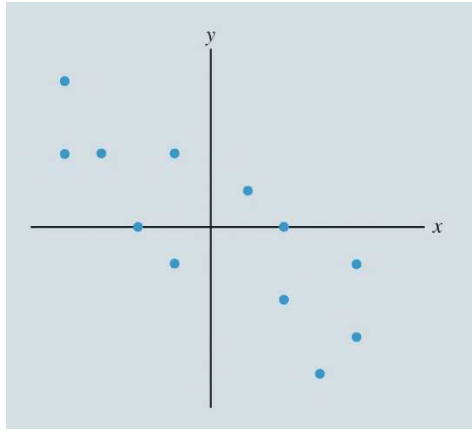
Figure 2.28: Scatter Diagrams and Associated Covariance Values for Different Variable Relationships



(a)
s_{xy} **Positive:**
(*x* and *y* are positively linearly related)



(b)
s_{xy} **Approximately 0:**
(*x* and *y* are not linearly related)



(c)
s_{xy} **Negative:**
(*x* and *y* are negatively linearly related)

Measures of Association Between Two Variables (Slide 8 of 11)

Correlation Coefficient:

- The **correlation coefficient** measures the relationship between two variables.
- Not affected by the units of measurement for *x* and *y*.

Sample Correlation Coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \qquad (2.10)$$

where

r_{xy} = sample correlation coefficient
s_{xy} = sample covariance
s_x = sample standard deviation of x
s_y = sample standard deviation of y

Measures of Association Between Two Variables (Slide 9 of 11)

Interpretation of Correlation Coefficient:

$-1 \leq r \leq +1$

***r* value**	**Relationship between the *x* and *y* variables**
< 0	Negative linear
Near 0	No linear relationship
> 0	Positive linear

Measures of Association Between Two Variables (Slide 10 of 11)

Computation of Correlation Coefficient:

Illustration:

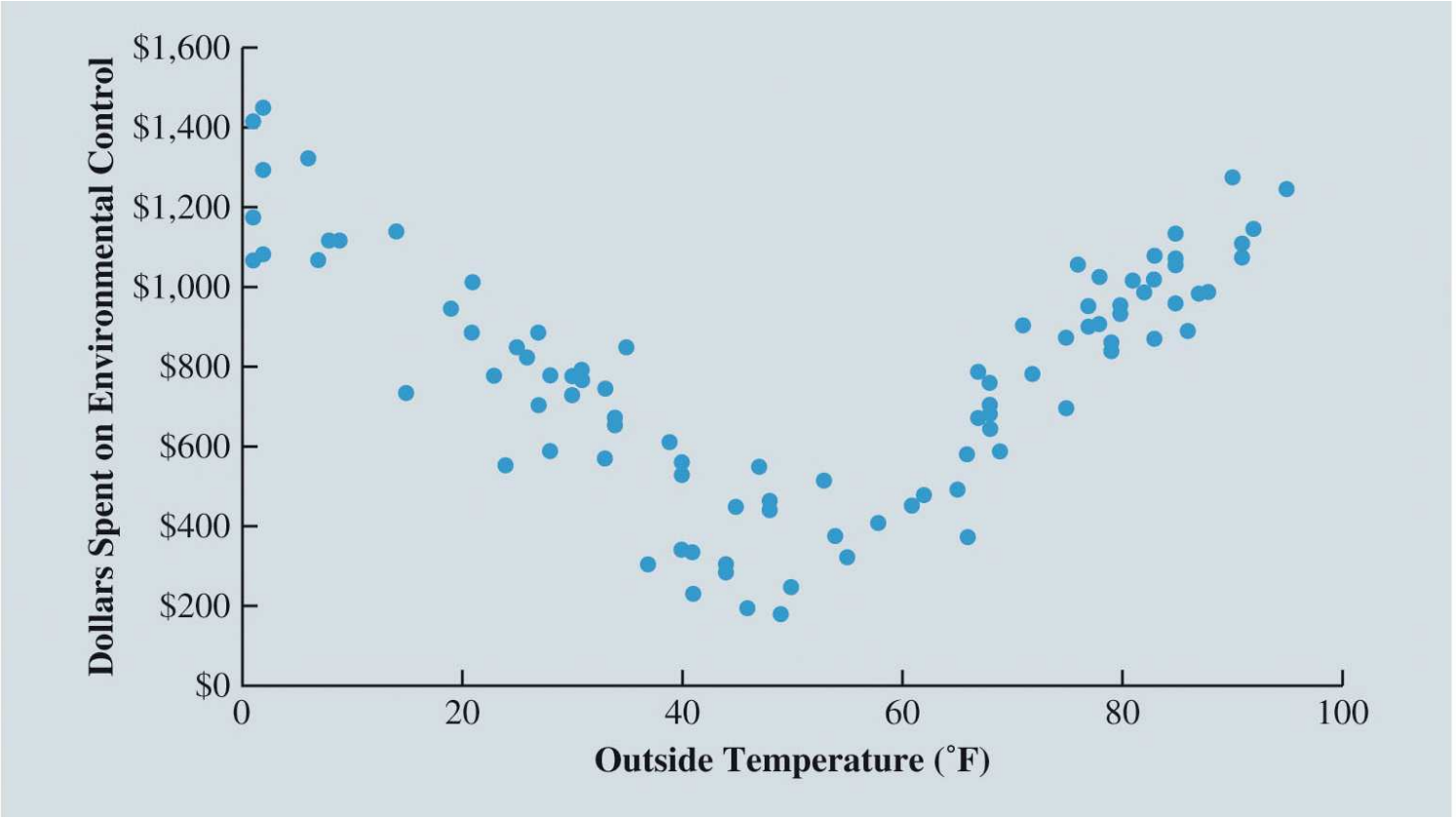
- To determine the sample correlation coefficient for bottled water sales at Queensland Amusement Park:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{12.8}{(4.36)(3.15)} = 0.93$$

- There is a very strong linear relationship between high temperature and sales.

Measures of Association Between Two Variables (Slide 11 of 11)

Figure 2.29: Example of Nonlinear Relationship Producing a Correlation Coefficient Near Zero



Data Cleansing

Missing Data

Blakely Tires

Identification of Erroneous Outliers and other Erroneous Values

Variable Representation

Data Cleansing (Slide 1 of 11)

Missing Data:

- Data sets commonly include observations with missing values for one or more variables.
- In some cases missing data naturally occur; these are called **legitimately missing data.**
  - Generally, no remedial action is taken for legitimately missing data.
- In other cases missing data occur for different reasons; these are called **illegitimately missing data.**
  - The primary options for addressing such missing data are:
    1. To discard observations (rows) with any missing values.
    2. To discard any variable (column) with missing values.
    3. To fill in missing entries with estimated values.
    4. To apply a data-mining algorithm that can handle missing values.

Data Cleansing (Slide 2 of 11)

Missing Data (cont.):

- **Missing completely at random** (MCAR): The tendency for an observation to be missing the value for some variable is entirely random; whether data are missing does not depend on either the value of the missing data or the value of any other variable in the data.
- **Missing at random** (MAR): The tendency for an observation to be missing a value for some variable is related to the value of some other variable(s) in the data.
- **Missing not at random** (MNAR): The tendency for the value of a variable to be missing is related to the value that is missing.
- **Imputation**: The systematic replacement of missing values with values that seem reasonable.

Data Cleansing (Slide 3 of 11)

Blakely Tires:

- A U.S. producer of automobile tires wants to learn about the conditions of its tires on automobiles in Texas.
- The data obtained includes the position of the tire on the automobile, age of the tire, mileage on the tire, and depth of the remaining tread on the tire.
- Begin assessing the quality of these data by determining which (if any) observations have missing values (see Figure 2.30).

Data Cleansing (Slide 4 of 11)

Figure 2.30: Portion of Excel Spreadsheet Showing Number of Missing Values for Variables in *TreadWear* Data

	A	B	C	D	E	F	G	H	I	J
1	ID Number	Position on Automobile	Life of Tire (Months)	Tread Depth	Miles			Life of Tire (Months)	Tread Depth	Miles
2	13391487	LR	58.4	2.2	2805		# of Missing Values	0	0	1
3	21678308	LR	17.3	8.3	39371					
4	18414311	RR	16.5	8.6	13367					
5	19778103	RR	8.2	9.8	1931					
6	16355454	RR	13.7	8.9	23992					
7	8952817	LR	52.8	3.0	48961					
8	6559652	RR	14.7	8.8	4585					

Data Cleansing (Slide 5 of 11)

Blakely Tires (cont.):

- Sort all of Blakely's data on Miles from smallest to largest value to determine which observation is missing its value of this variable.

Figure 2.31: Portion of Excel Spreadsheet Showing *TreadWear* Data Sorted on Miles from Lowest to Highest Value

	A	B	C	D	E	F	G	H	I	J
1	ID Number	Position on Automobile	Life of Tire (Months)	Tread Depth	Miles			Life of Tire (Months)	Tread Depth	Miles
2	15890813	LF	16.1	8.6	206		# of Missing Values	0	0	1
3	15890813	LR	16.1	8.6	206					
4	15890813	RF	16.1	8.6	206					
455	9306585	RR	45.4	4.1	107237					
456	9306585	LF	45.4	4.1	107237					
457	3354942	LF	17.1	8.5						

Data Cleansing (Slide 6 of 11)

Figure 2.32: Portion of Excel Spreadsheet Showing *TreadWear* Data Sorted from Lowest to Highest by ID Number

54	3121851	LR	17.1	8.4	21378
55	3121851	RR	17.1	8.4	21378
56	3121851	RF	17.1	8.4	21378
57	3121851	LF	17.1	8.5	21378
58	3354942	LF	17.1	8.5	
59	3354942	RF	21.4	7.7	33254
60	3354942	RR	21.4	7.8	33254
61	3354942	LR	21.4	7.7	33254
62	3374739	RR	73.3	0.2	57313
63	3574739	RF	73.3	0.2	57313
64	3574739	LF	73.3	0.2	57313
65	3574739	LR	73.3	0.2	57313

Data Cleansing (Slide 7 of 11)

Identification of Erroneous Outliers and other Erroneous Values:

- Examining the variables in the data set by use of summary statistics, frequency distributions, bar charts and histograms, z-scores, scatter plots, correlation coefficients, and other tools can uncover data-quality issues and outliers.
- Many software ignore missing values when calculating various summary statistics.
- If missing values in a data set are indicated with a unique value (such as 9999999), these values may be used by software when calculating various summary statistics.
- Both cases can result in misleading values for summary statistics.
- Many analysts prefer to deal with missing data issues prior to using summary statistics to attempt to identify erroneous outliers and other erroneous values in the data.

Data Cleansing (Slide 8 of 11)

Figure 2.33: Portion of Excel Spreadsheet Showing the Mean and Standard Deviation for Each Variable in the *TreadWear* Data

	A	B	C	D	E	F	G	H	I	J
1	ID Number	Position on Automobile	Life of Tire (Months)	Tread Depth	Miles			Life of Tire (Months)	Tread Depth	Miles
2	80441	LR	19.0	8.1	37419		# of Missing Values	0	0	1
3	80441	LF	19.0	8.1	37419		Mean	23.80	7.68	25440.22
4	80441	RR	19.0	8.2	37419		Standard Deviation	31.82	2.62	23600.21
5	80441	RF	19.0	8.1	37419					
6	95990	RR	8.6	9.7	5670					
7	95990	LR	8.6	9.7	5670					
8	95990	LF	8.6	9.7	5670					

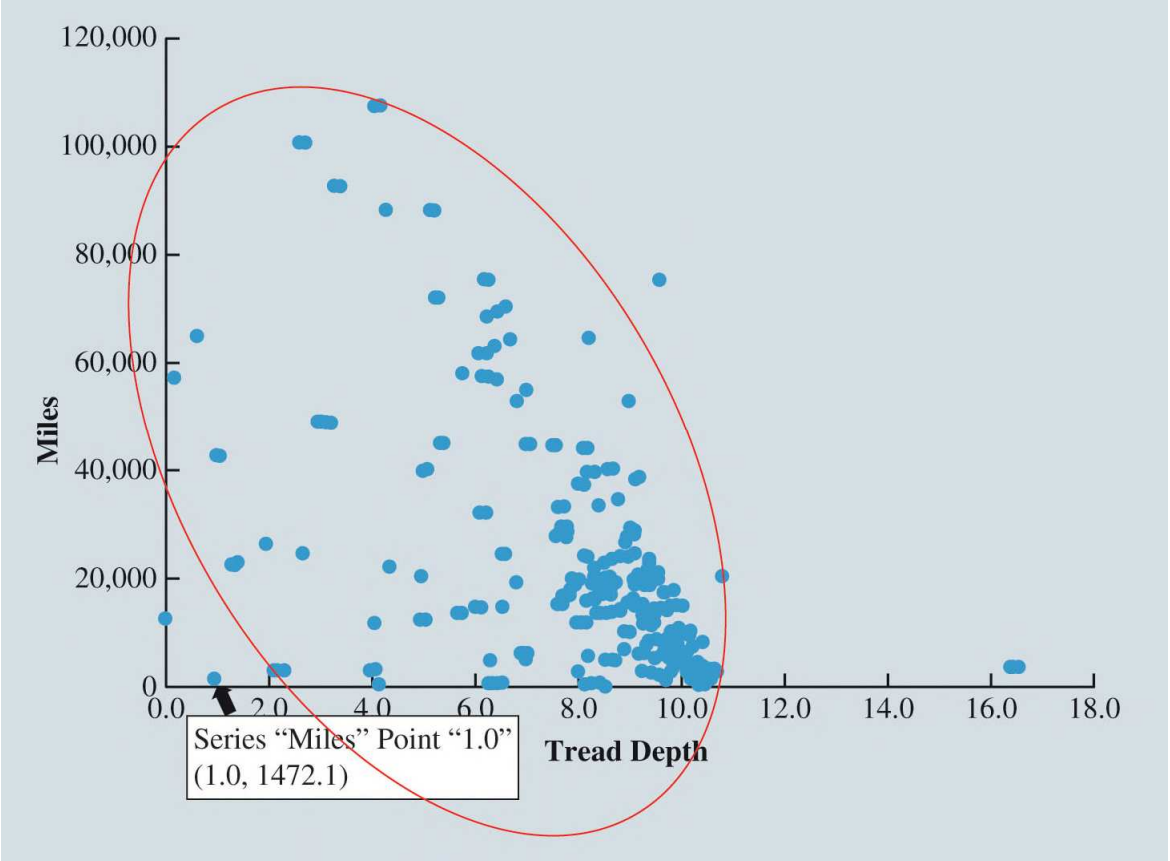
Data Cleansing (Slide 9 of 11)

Figure 2.34: Portion of Excel Spreadsheet Showing the *TreadWear* Data Sorted on Life of Tires (Months) from Lowest to Highest Value

	A	B	C	D	E	F	G	H	I	J
1	ID Number	Position on Automobile	Life of Tire (Months)	Tread Depth	Miles			Life of Tire (Months)	Tread Depth	Miles
2	9091771	RF	1.8	10.8	2917		# of Missing Values	0	0	1
3	9091771	RR	1.8	10.7	2917		Mean	23.80	7.68	25440.22
4	9091771	LF	1.8	10.7	2917		Standard Deviation	31.82	2.62	23600.21
5	7712178	LF	2.1	10.7	2186		Minimum	1.8		
6	7712178	RR	2.1	10.7	2186		Maximum	601.0		
452	3574739	RR	73.3	0.2	57313					
453	3574739	LF	73.3	0.2	57313					
454	3574739	LR	73.3	0.2	57313					
455	3574739	LR	73.3	0.2	57313					
456	2122934	LR	111.0	9.3	21000					
457	8696859	LR	601.0	2.0	26129					

Data Cleansing (Slide 10 of 11)

Figure 2.35: Scatter Diagram of Tread Depth and Miles for the *TreadWear* Data



Data Cleansing (Slide 11 of 11)

Variable Representation:

- In many data-mining applications, it may be prohibitive to analyze the data because of the number of variables recorded.
- **Dimension reduction** is the process of removing variables from the analysis without losing crucial information.
- A critical part of data mining is determining how to represent the measurements of the variables and which variables to consider.
- Often data sets contain variables that, considered separately, are not particularly insightful but that, when appropriately combined, result in a new variable that reveals an important relationship.