INSTRUCTOR'S SOLUTIONS MANUAL

BUSINESS STATISTICS
A DECISION-MAKING APPROACH
TENTH EDITION

David F. Groebner

Boise State University

Patrick W. Shannon

Boise State University

Phillip C. Fry

Boise State University

Pearson

The author and publisher of this book have used their best efforts in preparing this book. These efforts include the development, research, and testing of the theories and programs to determine their effectiveness. The author and publisher make no warranty of any kind, expressed or implied, with regard to these programs or the documentation contained in this book. The author and publisher shall not be liable in any event for incidental or consequential damages in connection with, or arising out of, the furnishing, performance, or use of these programs.

Reproduced by Pearson from electronic files supplied by the author.

Publishing as Pearson, 330 Hudson Street, NY NY 10013

ISBN-13: 978-0-13-449642-9
ISBN-10: 0-13-449642-6

Contents

Chapter 1: The Where, Why, and How of Data Collection

Section 1.1

1.1. This application is primarily descriptive in nature. The owner wishes to develop a presentation. She will most likely use charts, graphs, tables and numerical measures to describe her data.

1.2. The graph is a bar chart. A bar chart displays values associated with categories. In this case the categories are the departments at the food store. The values are the total monthly sales (in dollars) in each department. A bar chart also typically has gaps between the bars. A histogram has no gaps and the horizontal axis represents the possible values for a numerical variable.

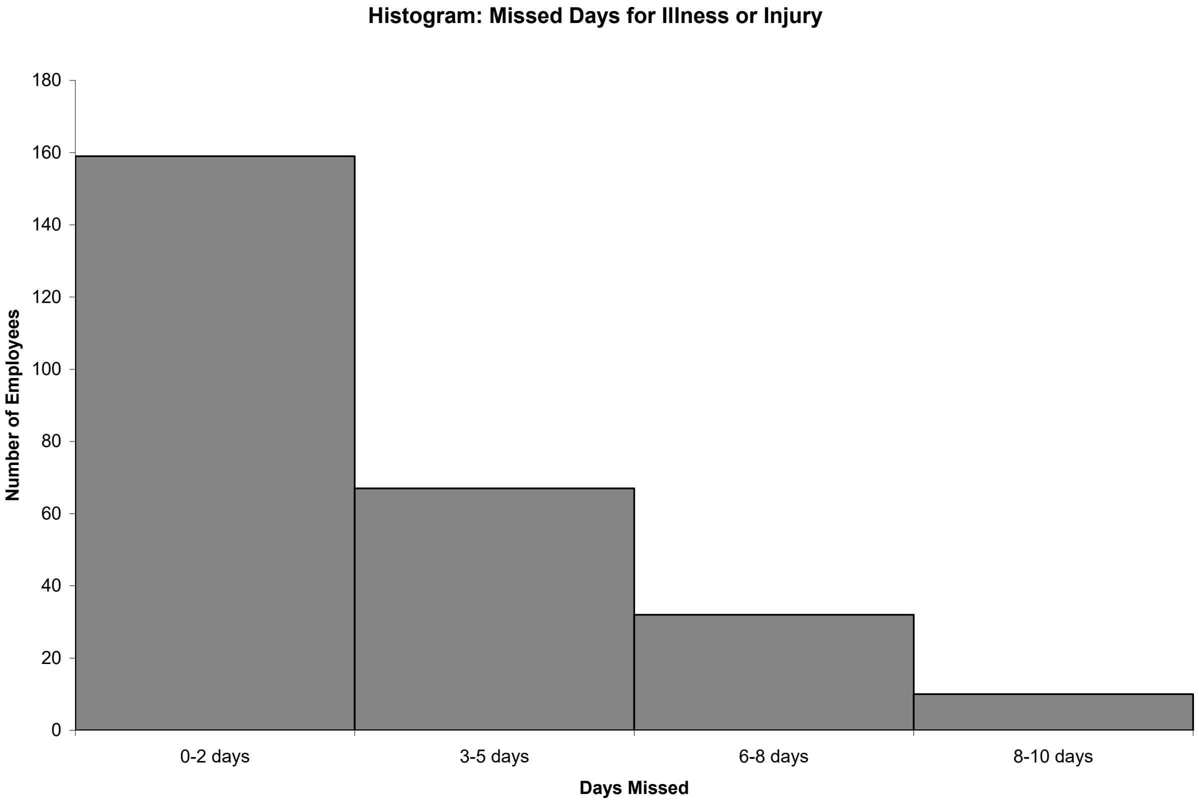
1.3. A bar chart is used whenever you want to display data that has already been categorized while a histogram is used to display data over a range of values for the factor under consideration. Another fundamental difference is that there typically are gaps between the bars on a bar chart but there are no gaps between the bars of a histogram.

1.4. Businesses often make claims about their products that can be tested using hypothesis testing. For example, it is not enough for a pharmaceutical company to claim that its new drug is effective in treating a disease. In order for the drug to be approved by the Food and Drug Administration the company must present sufficient evidence that the drug first does no harm and that it also provides an effective treatment against the disease. The claims that the drug does no harm and is an effective treatment can be tested using hypothesis testing.

1.5. The company could use statistical inference to determine if its parts last longer. Because it is not possible to examine every part that could be produced the company could examine a randomly chosen subset of its parts and compare the average life of the subset to the average life of a randomly chosen subset of the competitor's parts. By using statistical inference procedures the company could reach a conclusion about whether its parts last longer or not.

1.6. Student answers will vary depending on the periodical selected and the periodical's issue date, but should all address the three parts of the question.

1.7. The appropriate chart in this case is a histogram where the horizontal axis contains the number of missed days and the height of the bars represent the number of employees who missed each number of days



Note, there are no gaps between the bars.

1.8. Because it would be too costly, too time consuming, or practically impossible to contact every subscriber to ascertain the desired information, the decision makers at *Fortune* might decide to use statistical inference, particularly estimation, to answer its questions. By looking at a subset of the data and using the procedures of estimation it would be possible for the decision makers to arrive at values for average age and average income that are within tolerable limits of the actual values.

1.9. Student answers will vary depending on the business periodical or newspaper selected and the article referenced. Some representative examples might include estimates of the number of CEO's who will vote for a particular candidate, estimates of the percentage increase in wages for factory workers, estimates of the average dollar advertising expenditures for pharmaceutical companies in a specific year, and the expected increase in R&D expenditures for the coming quarter.

1.10. Student answers will vary. However, the examples should illustrate how statistics has been used and should clearly indicate the type of statistical analysis employed.

Section 1.2

1.11. As discussed in this section, the pet store would most likely use a written survey or a telephone survey to collect the customer satisfaction data.

1.12. A leading question is one that is designed to elicit a specific response, or one that might influence the respondent's answer by its wording. The question is posed so that the respondent believes the researcher has a specific answer in mind when the question is asked, or worded in such a way that the respondent feels obliged to provide an answer consistent with the question. For example, a question such as "Do you agree with the experts who recommend that more tax dollars be given to clean up dangerous and unhealthy pollution?" could cause respondents to provide the answer that they think will be consistent with the "experts" with whom they do not want to disagree. Leading question should be avoided in surveys because they may introduce bias.

1.13. An experiment is any process that generates data as its outcome. The plan for performing the experiment in which the variable of interest is defined is referred to as an experimental design. In the experimental design one or more factors are identified to be changed so that the impact on the variable of interest can be observed or measured.

1.14. There will likely by a high rate of nonresponse bias since many people who work days will not be home during the 9–11 AM time slot. Also, the data collectors need to be careful where they get the phone number list as some people do not have listed phones in phone books and others have no phone or only a cell phone. This may result in selection bias.

1.15. a. Observation would be the most likely method. Observers could be located at various bike routes and observe the number of riders with and without helmets. This would likely be better than asking people if they wear a helmet since the popular response might be to say yes even when they don't always do so.

b. A telephone survey to gas stations in the state. This could be a cost effective way of getting data from across the state. The respondent would have the information and be able to provide the correct price.

c. A written survey of passengers. This could be given out on the plane before the plane lands and passengers could drop the surveys in a box as they de-plane. This method would likely garner higher response rates compared to sending the survey to passengers' mailing address and asking them to return the completed survey by mail.

1.16. The two types of validity mentioned in the section are internal validity and external validity. For this problem external validity is easiest to address. It simply means the sampling method chosen will be sufficient to insure the results based on the sample will be able to be generalized to the population of all students. Internal validity would involve making sure the data gathering method, for instance a questionnaire, accurately determines the respondent's attitude toward the registration process.

1.17. This data could have been collected through a survey. Employees of the USDA could provide periodic reports of fire ant activity in their region. Also, medical reports could be used to collect data assuming people with bites had required medical attention.

1.18. There are many potential sources of bias associated with data collection. If data is to be collected using personal interviews it will be important that the interviewer be trained so that interviewer bias, arising from the way survey questions are asked, is not injected into the survey. If the survey is conducted using either a mail survey or a telephone survey then it is important to be aware of nonresponse bias from those who do not respond to the mailing or refuse to answer your calls. You must also be careful when selecting your survey subjects so that selection bias is not a problem. In order to have useful, reliable data that is representative of the true student opinions regarding campus food service, it is necessary that the data collection process be conducted in a manner that reduces or eliminates the potential for these and other sources of potential bias.

1.19. For retailers technology that scans the product UPC code at checkout makes the collection of data fast and accurate. Retailers that use such technology can automatically update their inventory records and develop an extensive collection of customer buying habits. By applying advanced statistical techniques to the data the retailer can identify relationships among purchases that might otherwise go unnoticed. Such information could enable retailers to target their advertising or even rearrange the placement of products in the store to increase sales. Manufacturing firms use bar code scanning to collect information concerning product availability and product quality. Credit card purchases are automatically tracked by the retailer and the bankcard company. In this way the credit card company is able to track your purchases and even alert you to potential fraud if purchases on your card appear to be unusual. Finally, some companies are using radio frequency identification (RFID) to track products through their supply chain, so that product delays and inventory problems can be minimized.

1.20. One advantage of this form of data gathering is the same as for mail questionnaires. That is low cost. Additional factors being speed of delivery and, with current software, with closed- ended questions, instant updating of data analysis. Disadvantages are also similar, in particular low response and potential confusion about questions. An additional factor might be the ability of competitors to “hack” into the database and analysis program.

1.21. Student answers will vary. Look for clarity of questions and to see that the issue questions are designed to gather useful data. Look for appropriate demographic questions.

1.22. Students should select some form of personal observation as the data-gathering technique. In addition, there should be a discussion of a sampling procedure with an effort made to ensure the sample randomly selected both days of the week unless daily observations are made, and randomly selected times of the day since 24 hour observation would likely be impossible. A complete answer would also address efforts to reduce the potential bias of having an observer standing in an obvious manner by the displays.

1.23. Student answers will vary. However, the issue questions should be designed to gather the desired data regarding customers’ preferences for the use of the space. Demographic questions should provide data so that the responses can be broken down appropriately so that United Fitness Center managers can determine which subset of customers have what opinion about this issue. Regarding questionnaire layout, look at neatness and answer location space. Make sure questions are properly worded, used reasonable vocabulary, and are not leading questions.

1.24. The results of the survey are based on telephone interviews with 744 adults, aged 18 and older. Students may also answer that the survey could have been conducted using a written survey via mail questionnaire or internet survey. Because telephone interviews were used to collect the survey data nonresponse biases associated with sampled adults who are not at home when phoned, or adults who refuse to participate in the survey. There is also the problem that some adults do not have a landline phone. If written surveys are used to collect the data then it is important to guard against nonresponse bias from those sampled adults who do not complete the survey There is also the problem of selection bias. In phone interviews we may miss the people who work evenings and nights. If written surveys are used we must be careful to select a representative sample of the adult population.

Section 1.3

1.25. a. Because the population is spread over a large geographical area, a cluster random sample could be selected to reduce travel costs.

b. A stratified random sample would probably be used to keep sample size as small as possible.

c. Most likely a convenience sample would be used since doing a statistical sample would be too difficult.

1.26. To determine the range of employee numbers for the first employee selected in a systematic random sample use the following:

$$\text{Part Range} = \frac{\text{Population Size}}{\text{Sample Size}} = \frac{18,000}{100} = 180$$

Thus, the first person selected will come from employees 1–180. Once that person is randomly selected, the second person will be the one numbered 180 higher than the first, and so on.

1.27. Whenever a descriptive numerical measure such as an average is calculated from the entire population it is a parameter. The corresponding measure calculated from a subset of the population, that is to say a sample, is a statistic.

1.28. Statistical sampling techniques consist of those sampling methods that select samples based on chance. Nonstatistical sampling techniques consist of those methods of selecting samples using convenience, judgment, or other nonchance processes. In convenience sampling, samples are chosen because they are easy or convenient to sample. There is no attempt to randomize the selection of the selected items. In convenience sampling not every item in the population has a random chance of being selected. Rather, items are sampled based on their convenience alone. Thus, convenience sampling is not a statistical sampling method.

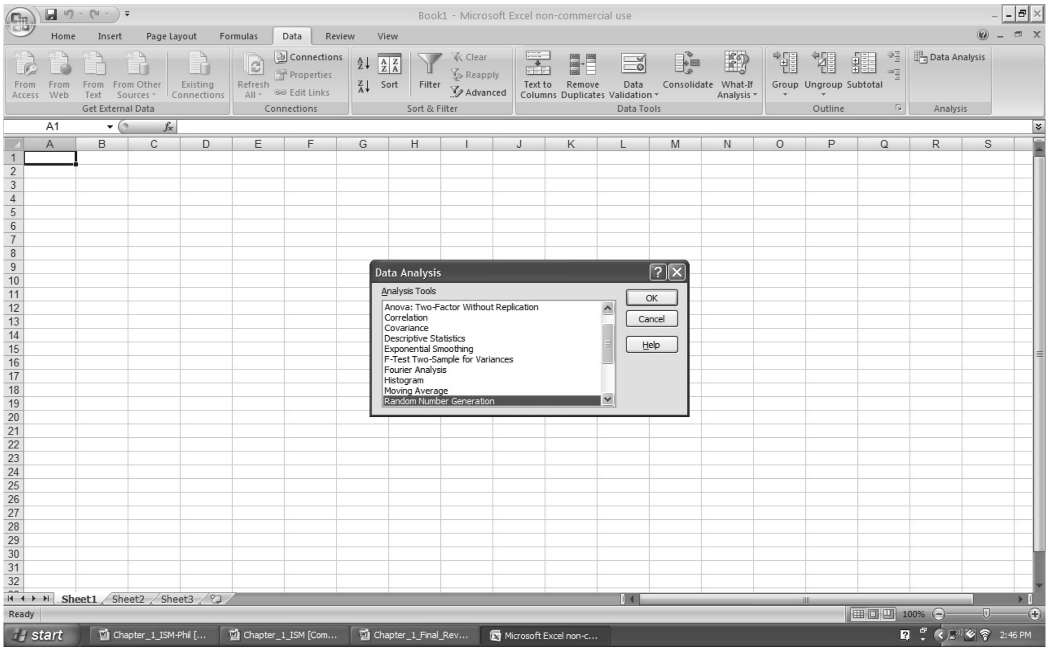
1.29. From a numbered list of all customers who own a certificate of deposit the bank would need to randomly determine a starting point between 1 and *k*, where *k* would be equal to 25000/1000 = 25. This could be done using a random number table or by having a statistical package or a spreadsheet generate a random number between 1 and 25. Once this value is determined the bank would select that numbered customer as the first sampled customer and then select every 25th customer after that until 100 customers are sampled.

1.30. A census is an enumeration of the entire set of measurements taken from the population as a whole. While in some cases, the items of interest are obtained from people such as through a survey, in many instances the items of interest come from a product or other inanimate object. For example, a study could be conducted to determine the defect rate for items made on a production line. The census would consist of all items produced on the line in a defined period of time.

1.31. Values computed from a sample are always considered statistics. In order for a value, such as an average, to be considered a parameter it must be computed from all items in the population.

1.32. In stratified random sampling, the population is divided into homogeneous groups called strata. The idea is to make all items in a stratum as much alike as possible with respect to the variable of interest thereby reducing the number of items that will need to be sampled from each stratum. In cluster sampling, the idea is to break the population into heterogeneous groups called clusters (usually on a geographical basis) such that each cluster looks as much like the original population as possible. Then clusters are randomly selected and from the cluster, individual items are selected using a statistical sampling method.

1.33. Using Excel, choose the Data tab, select Data Analysis from the Analysis Group , then Random Number Generation—shown as follows:

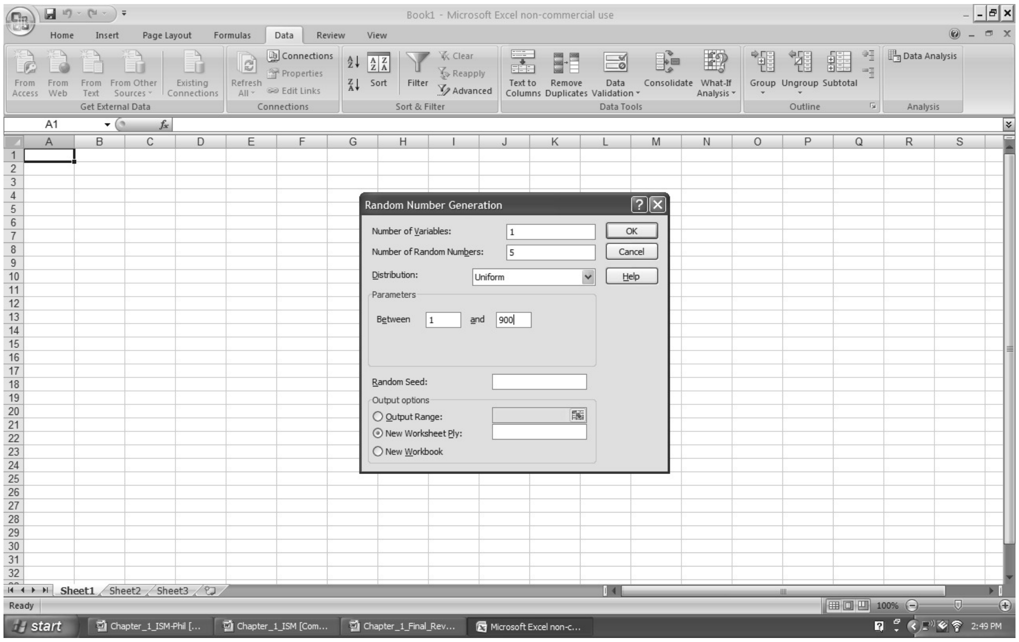


Data Analysis

Analysis Tools

- Anova: Two-Factor Without Replication
- Correlation
- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram
- Moving Average
- Random Number Generation

The next step is to complete the random number generation dialog as follows:



Random Number Generation

Number of Variables: 1

Number of Random Numbers: 5

Distribution: Uniform

Parameters

Between 1 and 900

Random Seed:

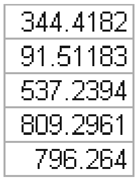
Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

The resulting random numbers generated are:



344.4182
91.51183
537.2394
809.2961
796.264

Note, the students' answers may differ since Excel generates different streams of random numbers each time it is used. Also, if the application requires integer numbers, the Decrease Decimal option can be used.

1.34. If these percentages were based on all students attending college in those years they would be parameters, if the percentages were based on a sample they would be statistics.

1.35. This is a statistic. A poll would be a sample of eligible voters rather than all eligible voters.

1.36. Solution

a. Stratified random sampling

b. Simple random sampling or possibly cluster random sampling

c. Systematic random sampling

d. Stratified random sampling

1.37. This is a statistical sample. Every employee has an equal chance of being selected using this method. In fact, this is an example of a simple random sample because every possible sample of size 50 has an equal chance of being selected.

1.38. a. Student answers will vary

b. Cluster sampling could be used to ensure that you get all types of cereal. Make each cluster the area where certain cereals are located (i.e., isle, row, shelf, etc.)

c. Cluster sampling would give you a better idea of the inventory of all types of cereal. Simple random sampling could possibly end up with only looking at 2 or 3 cereal types.

1.39. Students should choose the Data tab, select Data Analysis from the Analysis group—Random Number Generation process. Students' answers will differ since Excel generates different streams of random numbers each time it is used, but 40 random numbers should be generated from a uniform distribution with values ranging from 1 to 578. Since the application requires integer numbers, the Decrease Decimal option should be used.

1.40. a. The population should be all users of cross-country ski lots and trailheads in Colorado.

b. Several sampling techniques could be selected. Be sure that some method of ensuring randomness is discussed. In addition, some students might give greater weight to frequent users of the lots. In which case the population would really be user days rather than individual users.

c. Students using Excel should choose the Data tab, select Data Analysis from the Analysis group—Random Number Generation process. Students' answers may differ since Excel generates different streams of random numbers each time it is used. Since the application requires integer numbers, the Decrease Decimal option should be used.

1.41. a. Since there are 4,000 patient files we could give each file a unique identification number consisting of 4 digits. The first file would be given the identification number "0001." The last file would be given the identification number of "4000." By assigning each patient a number and randomly selecting the 100 numbers allows each possible sample of 100 an equal chance of being selected.

b. Either use a random number table (randomly select the starting row and column), or use a computer program, such as Microsoft Excel, which has a random number generator.

c. Since each patient is assigned a 4-digit identification number, we would need a 4-digit random number for each random number selected.

d. Answers will vary.

Section 1.4

1.42. a. Time-series
b. Cross-sectional
c. Time-series
d. Cross-sectional

1.43. Qualitative data are categories or numerical values that represent categories. Quantitative data is data that is purely numerical.

1.44. a. Ordinal—categories with defined order
b. Nominal—categories with no defined order
c. Ratio
d. Nominal—categories with no defined order

1.45. Nominal data involves placing observations in separate categories according to some measurable characteristic. Ordinal data also involves placing observations into separate categories, but the categories can be rank-ordered.

1.46. Since the circles involve a ranking from best to worst, this would be ordinal data.

1.47. a. The data are cross-sectional. The data are collected from 2,300 customers at approximately the same point in time
b. This is a ratio level, quantitative variable. The data represent a measurement of time.
c. Ordinal with a numerical value representing customers rating of level of service

1.48. a. Nominal Data
b. Ratio Data
c. Ratio Data
d. Ratio Data
e. Nominal

1.49. a. Cross-sectional
b. Time-series
c. Cross-sectional
d. Cross-sectional
e. Time-Series

1.50. Columns A–G are nominal—they are all codes
Columns H–L are ratio level.

End of Chapter Exercises

1.51. Answers will vary with the student. But a good discussion should include the following factors:

Sampling techniques and possible problems selecting a representative sample.
Determining how to develop questions to measure approval.
Structuring questions to avoid bias.
The measurement scale associated with the questions.
The fact these polls tend to develop time-series data.

1.52. Nominal data or ordinal data.

1.53. Interval or ratio data.

1.54. Ratings are typical uses of ordinal scale data. And since ratings are based on personal opinion, even though people are using the same scale, a direct comparison between the two ratings is not possible. This is a common problem when people are asked to rate an object using an ordinal scale.

1.55. Answers will vary with the student. But a good discussion should include the following factors:

Sampling techniques and possible problems selecting a representative sample.

Determining how to measure confidence.

Structuring questions to avoid bias.

The measurement scale associated with the questions.

The fact this poll is specifically intended to develop time-series data.

1.56. Answers will vary with the student.

1.57. Answers will vary with the student.

1.58. a. No because a random sample means that every item in the population has an equal chance of being selected. Individuals who do not have or use email do not have an equal chance of being included in this survey. Also, volunteer emails would not be random.

b. In this survey the biggest drawback is that only individuals with strong feelings one way or the other are apt to respond to this survey. This could lead to a great deal of bias in the results of the survey. Another big problem with a survey is nonresponse bais. Again because they are requesting viewers to write in there will be a great deal of nonresponse to this survey. I would also include in the answer that the question being asked is somewhat leading. The phrase "using too much force in routine traffic stops" implies that, in fact, force is being used which one would not expect in a routine traffic stop.

1.59. a. They would probably want to sample the salsa jars as they come off the assembly line at the plant for a specified time period. They would want to use a random sample. One method would be to take a systematic random sample. They could then calculate the percentage of the sample that had an unacceptable thickness.

b. The product is going to be ruined after testing it. You would not want to ruin the entire product that comes off the assembly line.

1.60. a. Student answers will vary but one method would be personal observation at grocery stores or another method would be to simply look at their sales. Are buyers of the energy drinks purchasing bottles or cans?

b. If using personal observation just have people at grocery stores observe people over a specified period of time and note which are selecting cans and which are selecting bottles and look at the percentages of each.

c. You would be looking at ratio data because you could have a true 0 if, for example, no one purchased bottles.

d. Depends on the way the data are collected. Sales data would be quantitative.

1.61. a. The fact that the friend has selected his favorite players means that all players did not have a chance of being selected in the sample. The sample would be biased toward the type of players the friend favors.

b. One method would be to obtain a list of all NBA players. Then assign each player a number. Then you could use Excel's random number generator to obtain a random sample of 40 players from the list.

1.62. The appropriate design would be a stratified random sampling method. Start by dividing the students into class standing (Freshman, Sophomore, Junior, and Senior). Then randomly select students from each strata.

Chapter 2: Graphs, Charts, and Tables—Describing Your Data

When applicable, the first few problems in each section will be done following the appropriate step by step procedures outlined in the corresponding sections of the chapter. Following problems will provide key points and the answers to the questions, but all answers can be arrived at using the appropriate steps.

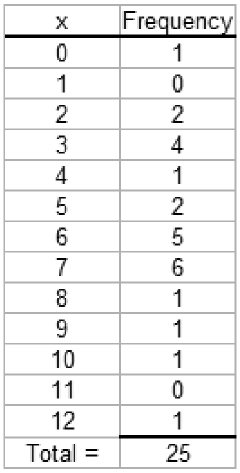
Section 2.1

2.1. Step 1: List the possible values.

The possible values for the discrete variable are 0 through 12.

Step 2: Count the number of occurrences at each value.

The resulting frequency distribution is shown as follows:



x	Frequency
0	1
1	0
2	2
3	4
4	1
5	2
6	5
7	6
8	1
9	1
10	1
11	0
12	1
Total =	25

2.2. Given $n = 2,000$, the minimum number of groups for a grouped data frequency distribution determined using the $2^k \geq n$ guideline is:

$$2^k \geq n \text{ or } 2^{11} = 2,048 \geq 2,000. \text{ Thus, use } k = 11 \text{ groups.}$$

2.3. a. Given $n = 1,000$, the minimum number of classes for a grouped data frequency distribution determined using the $2^k \geq n$ guideline is:

$$2^k \geq n \text{ or } 2^{10} = 1,024 \geq 1,000. \text{ Thus, use } k = 10 \text{ classes.}$$

b. Assuming that the number of classes that will be used is 10, the class width is determined as follows:

$$w = \frac{\text{High} - \text{Low}}{\text{Classes}} = \frac{2,900 - 300}{10} = \frac{2,600}{10} = 260$$

Then we round to the nearest 100 points giving a class width of 300.

2.4. Recall that the Ogive is produced by plotting the cumulative relative frequency against the upper limit of each class. Thus, the first class upper limit is 100 and has a relative frequency of $0.2 - 0.0 = 0.2$. The second class upper limit is 200 and has a relative frequency of $0.4 - 0.2 = 0.2$. Of course, the frequencies are obtained by multiplying the relative frequency by the sample size. As an example, the first class has a frequency of $(0.2)50 = 10$. The others follow similarly to produce the following distribution

Class	Frequency	Relative Frequency	Cumulative Relative Frequency
0 – < 100	10	0.20	0.20
100 – < 200	10	0.20	0.40
200 – < 300	5	0.10	0.50
300 – < 400	5	0.10	0.60
400 – < 500	20	0.40	1.00
500 – < 600	0	0.00	1.00

2.5. a. There are $n = 60$ observations in the data set. Using the $2^k > n$ guideline, the number of classes, *k*, would be 6. The maximum and minimum values in the data set are 17 and 0, respectively. The class width is computed to be: $w = (17-0)/66 = 2.833$, which is rounded to 3. The frequency distribution is

Class	Frequency
0–2	6
3–5	13
6–8	20
9–11	14
12–14	5
15–17	2
	Total = 60

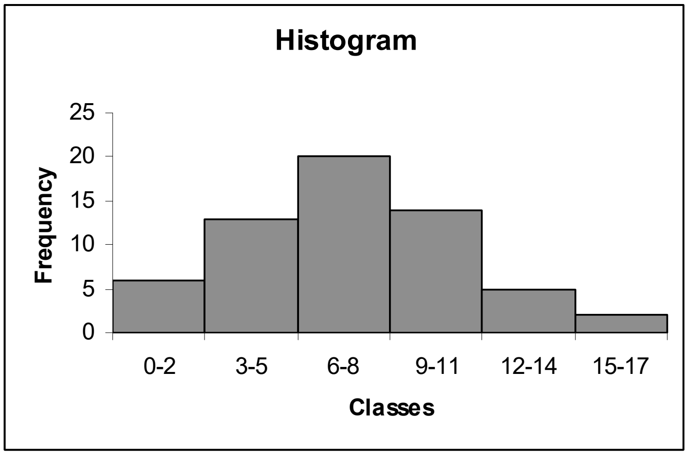
b. To construct the relative frequency distribution divide the number of occurrences (frequency) in each class by the total number of occurrences. The relative frequency distribution is shown below.

Class	Frequency	Relative Frequency
0–2	6	0.100
3–5	13	0.217
6–8	20	0.333
9–11	14	0.233
12–14	5	0.083
15–17	2	0.033
	Total = 60	

c. To develop the cumulative frequency distribution, compute a running sum for each class by adding the frequency for that class to the frequencies for all classes above it. The cumulative relative frequencies are computed by dividing the cumulative frequency for each class by the total number of observations. The cumulative frequency and the cumulative relative frequency distributions are shown below.

Class	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
0–2	6	0.100	6	0.100
3–5	13	0.217	19	0.317
6–8	20	0.333	39	0.650
9–11	14	0.233	53	0.883
12–14	5	0.083	58	0.967
15–17	2	0.033	60	1.000
	Total = 60			

d. To develop the histogram, first construct a frequency distribution (see part a). The classes form the horizontal axis and the frequency forms the vertical axis. Bars corresponding to the frequency of each class are developed. The histogram based on the frequency distribution from part (a) is shown below.



2.6. a. Proportion of days in which no shortages occurred = 1 – proportion of days in which shortages occurred $=1-0.24=0.76$.

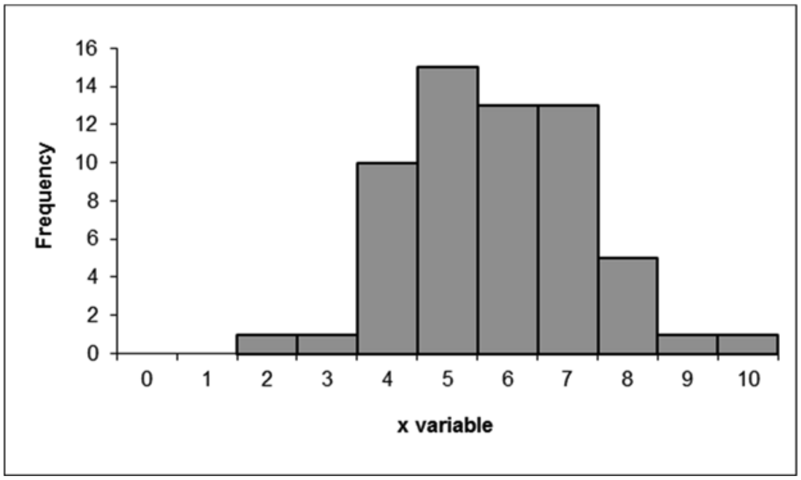
b. Less than \$20 off implies that overage was less than \$20 and the shortage was less than \$20 = (proportion of overages less \$20) – (proportion of shortages at most \$20) $=0.56-0.08=0.48$.

c. Proportion of days with less than \$40 over or at most \$20 short = Proportion of days with less than \$40 over – proportion of days with more than \$20 short $=0.96-0.08=0.88$.

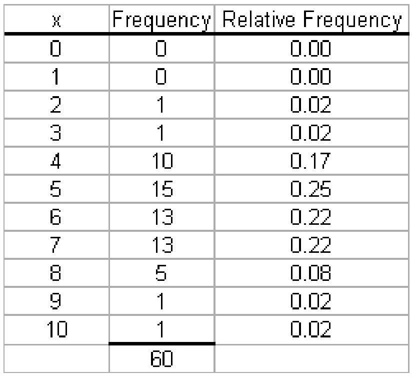
2.7. a. The data do not require grouping. The following frequency distribution is given:

x	Frequency
0	0
1	0
2	1
3	1
4	10
5	15
6	13
7	13
8	5
9	1
10	1

b. The following histogram could be developed.

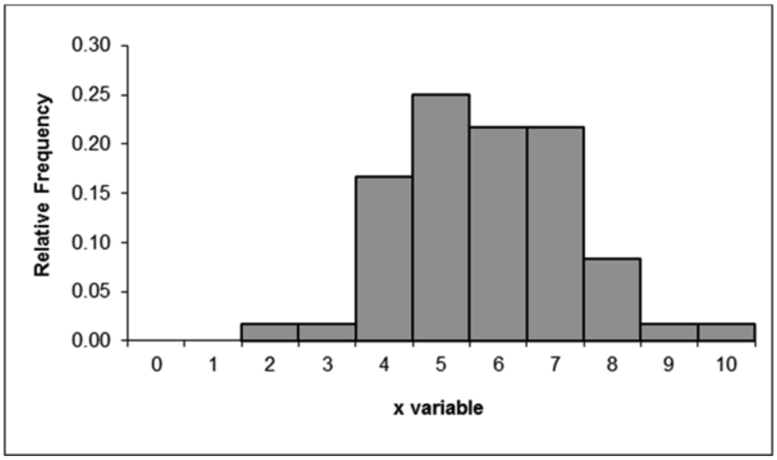


c. The relative frequency distribution shows the fraction of values falling at each value of *x*.



x	Frequency	Relative Frequency
0	0	0.00
1	0	0.00
2	1	0.02
3	1	0.02
4	10	0.17
5	15	0.25
6	13	0.22
7	13	0.22
8	5	0.08
9	1	0.02
10	1	0.02
	60	

d. The relative frequency histogram is shown below.



e. The two histograms look exactly alike since the same data are being graphed. The bars represent either the frequency or relative frequency.

2.8. a. Step 1 and Step 2: Group the data into classes and determine the class width:

The problem asks you to group the data. Using the $2^k \geq n$ guideline we get:

$2^k \geq 60$ so $2^6 \geq 60$

Class width is:

$$W = \frac{\text{Maximum} - \text{Minumum}}{\text{Number of Classes}} = \frac{10-2}{6} = 1.33$$

which we round up to 2.0

Step 3: Define the class boundaries:

Since the data are discrete, the classes are:

Class
2–3
4–5
6–7
8–9
10–11

Step 4: Count the number of values in each class:

Class	Frequency	Relative Frequency
2–3	2	0.0333
4–5	25	0.4167
6–7	26	0.4333
8–9	6	0.1000
10–11	1	0.0167

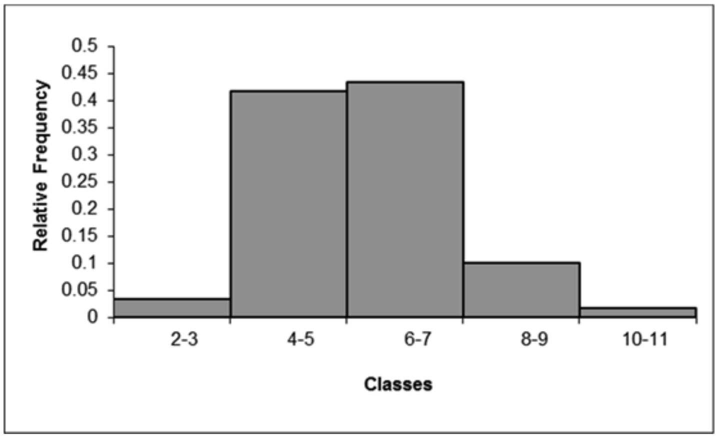
b. The cumulative frequency distribution is:

Class	Frequency	Cumulative Frequency
2–3	2	2
4–5	25	27
6–7	26	53
8–9	6	59
10–11	1	60

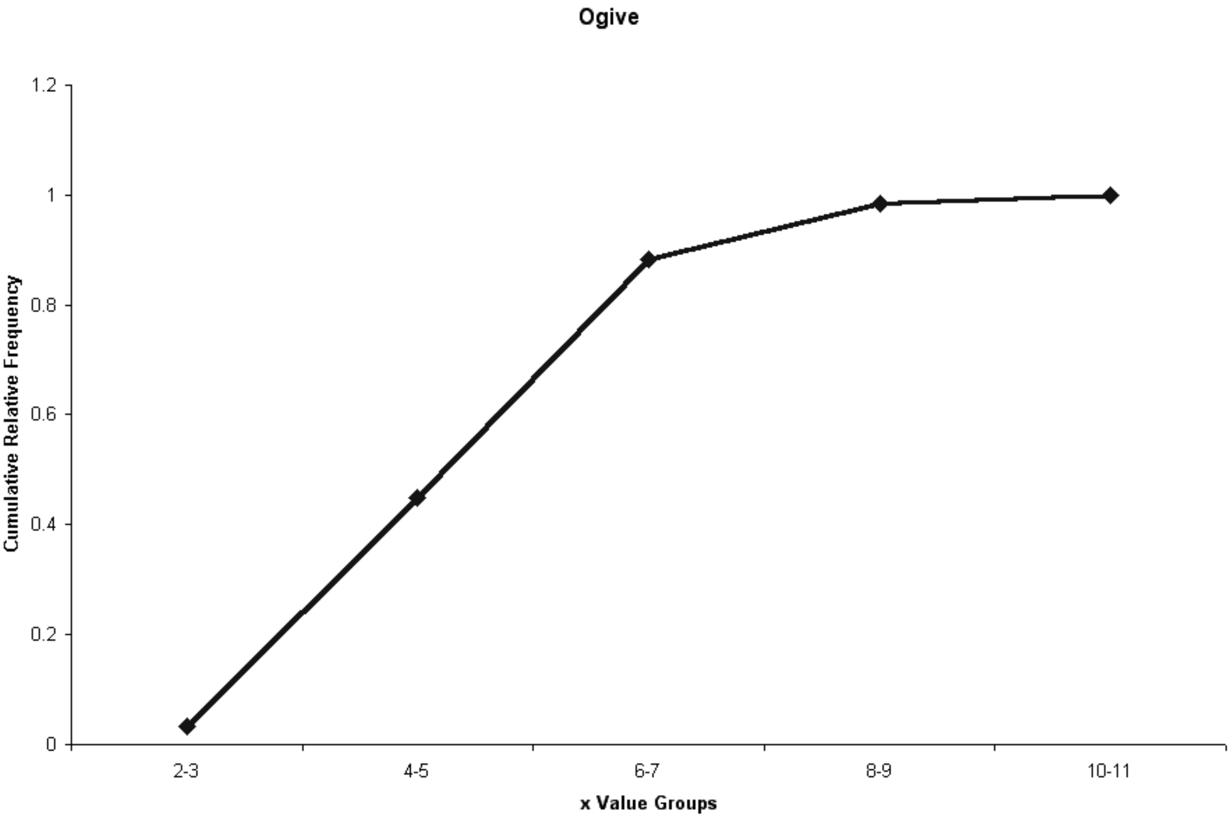
c.

			Cumulative
Class	Frequency	Relative Frequency	Relative Frequency
2–3	2	0.0333	0.0333
4–5	25	0.4167	0.4500
6–7	26	0.4333	0.8833
8–9	6	0.1000	0.9833
10–11	1	0.0167	1.000

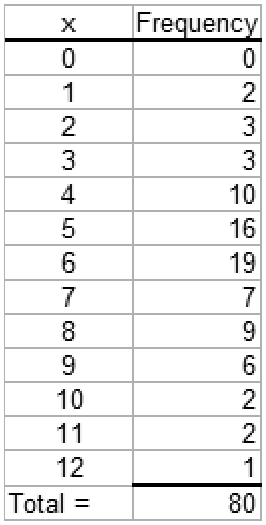
The relative frequency histogram is:



d. The ogive is a graph of the cumulative relative frequency distribution.



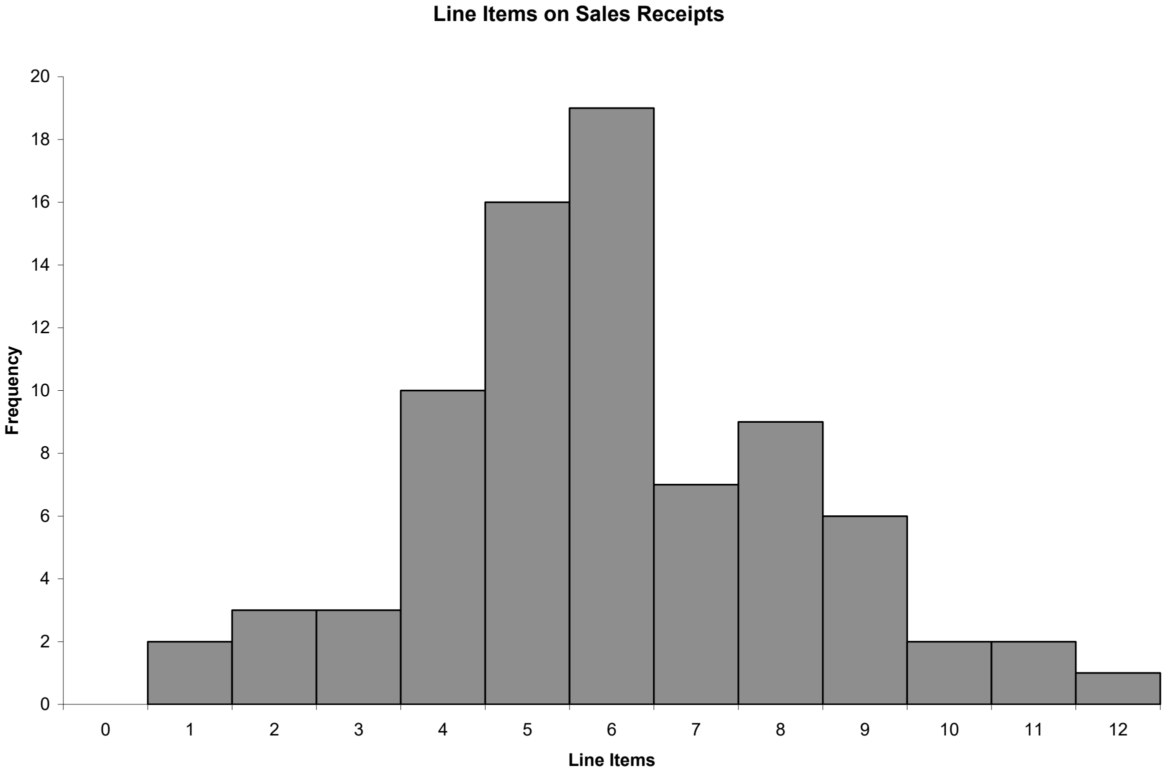
2.9. a. Because the number of possible values for the variable is relatively small, there is no need to group the data into classes. The resulting frequency distribution is:



x	Frequency
0	0
1	2
2	3
3	3
4	10
5	16
6	19
7	7
8	9
9	6
10	2
11	2
12	1
Total =	80

This frequency distribution shows the manager that most customer receipts have 4 to 8 line items.

b. A histogram is a graph of a frequency distribution for a quantitative variable. The resulting histogram is shown as follows.



2.10. a.

	Knowledge Level			
	Savvy	Experienced	Novice	Total
Online Investors	32	220	148	400
Traditional Investors	8	58	134	200
	40	278	282	600

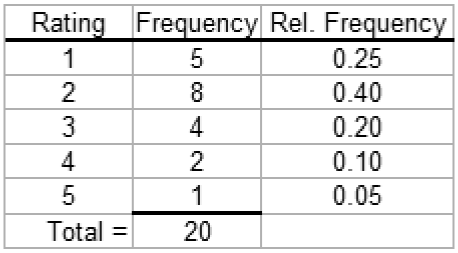
b.

	Knowledge Level		
	Savvy	Experienced	Novice
Online Investors	0.0533	0.3667	0.2467
Traditional Investors	0.0133	0.0967	0.2233

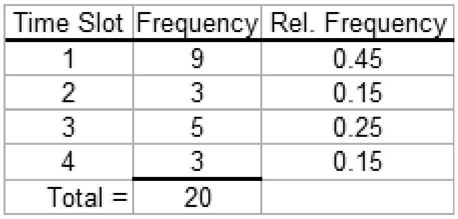
c. The proportion that were both on-line and experienced is 0.3667.

d. The proportion of on-line investors is 0.6667

2.11. a. The following relative frequency distributions are developed for the two variables:

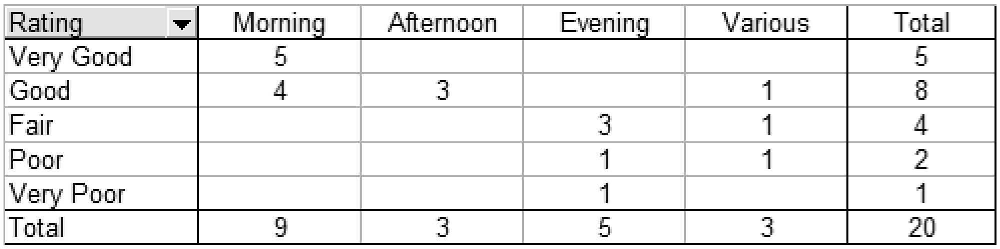


Rating	Frequency	Rel. Frequency
1	5	0.25
2	8	0.40
3	4	0.20
4	2	0.10
5	1	0.05
Total =	20	



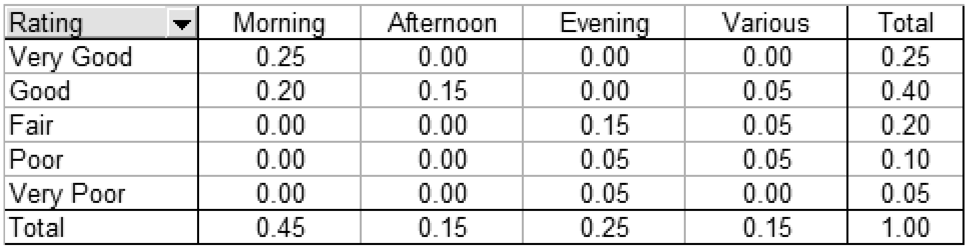
Time Slot	Frequency	Rel. Frequency
1	9	0.45
2	3	0.15
3	5	0.25
4	3	0.15
Total =	20	

b. The joint frequency distribution is a two dimensional table showing responses to the rating on one dimension and time slot on the other dimension. This joint frequency distribution is shown as follows:



Rating ▾	Morning	Afternoon	Evening	Various	Total
Very Good	5				5
Good	4	3		1	8
Fair			3	1	4
Poor			1	1	2
Very Poor			1		1
Total	9	3	5	3	20

c. The joint relative frequency distribution is determined by dividing each frequency by the sample size, 20. This is shown as follows:



Rating ▾	Morning	Afternoon	Evening	Various	Total
Very Good	0.25	0.00	0.00	0.00	0.25
Good	0.20	0.15	0.00	0.05	0.40
Fair	0.00	0.00	0.15	0.05	0.20
Poor	0.00	0.00	0.05	0.05	0.10
Very Poor	0.00	0.00	0.05	0.00	0.05
Total	0.45	0.15	0.25	0.15	1.00

Based on the joint relative frequency distribution, we see that those who advertise in the morning tend to provide higher service ratings. Evening advertisers tend to provide lower ratings. The manager may wish to examine the situation further to see why this occurs.

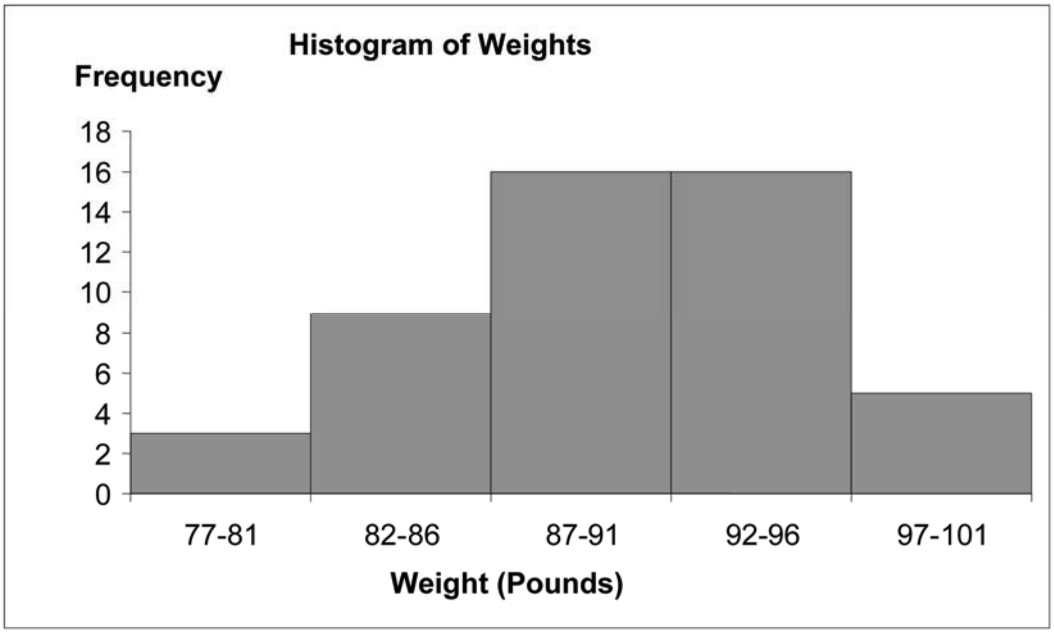
2.12. a. The weights are sorted from smallest to largest to create the data array.

77	79	80	83	84	85	86
86	86	86	86	86	87	87
87	88	88	88	88	89	89
89	89	89	90	90	91	91
92	92	92	92	93	93	93
94	94	94	94	94	95	95
95	96	97	98	98	99	101

b. Five classes having equal widths are created by subtracting the smallest observed value (77) from the largest value (101) and dividing the difference by 5 to get the width for each class (4.8 rounded to 5). Five classes of width five are then constructed such that the classes are mutually exclusive and all inclusive. Identify the variable of interest. The weight of each crate is the variable of interest. The number of crates in each class is then counted. The frequency table is shown below.

Weight (Classes)	Frequency
77–81	3
82–86	9
87–91	16
92–96	16
97–101	5
	Total = 49

c. The histogram can be created from the frequency distribution. The classes are shown on the horizontal axis and the frequency on the vertical axis. The histogram is shown below.



d. Convert the frequency distribution into relative frequencies and cumulative relative frequencies as shown below.

Weights (Classes)	Frequency	Relative Frequency	Cumulative Relative Frequency
77–81	3	0.0612	0.0612
82–86	9	0.1837	0.2449
87–91	16	0.3265	0.5714
92–96	16	0.3265	0.8980
97–101	5	0.1020	1.0000
	Total = 49		

The percentage of sampled crates with weights greater than 96 pounds is 10.20%.

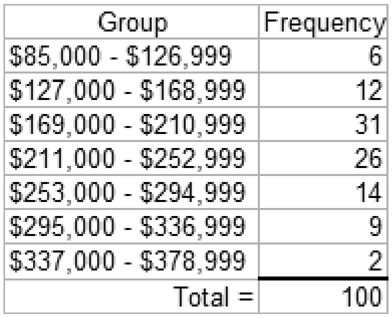
2.13. a. There are $n = 100$ values in the data. Then using the $2^k \geq n$ guideline we would need at least $k = 7$ classes.

b. Using $k = 7$ classes, the class width is determined as follows:

$$w = \frac{\text{High} - \text{Low}}{\text{Classes}} = \frac{\$376{,}644 - \$87{,}429}{7} = \frac{\$289{,}215}{7} = \$41{,}316.43$$

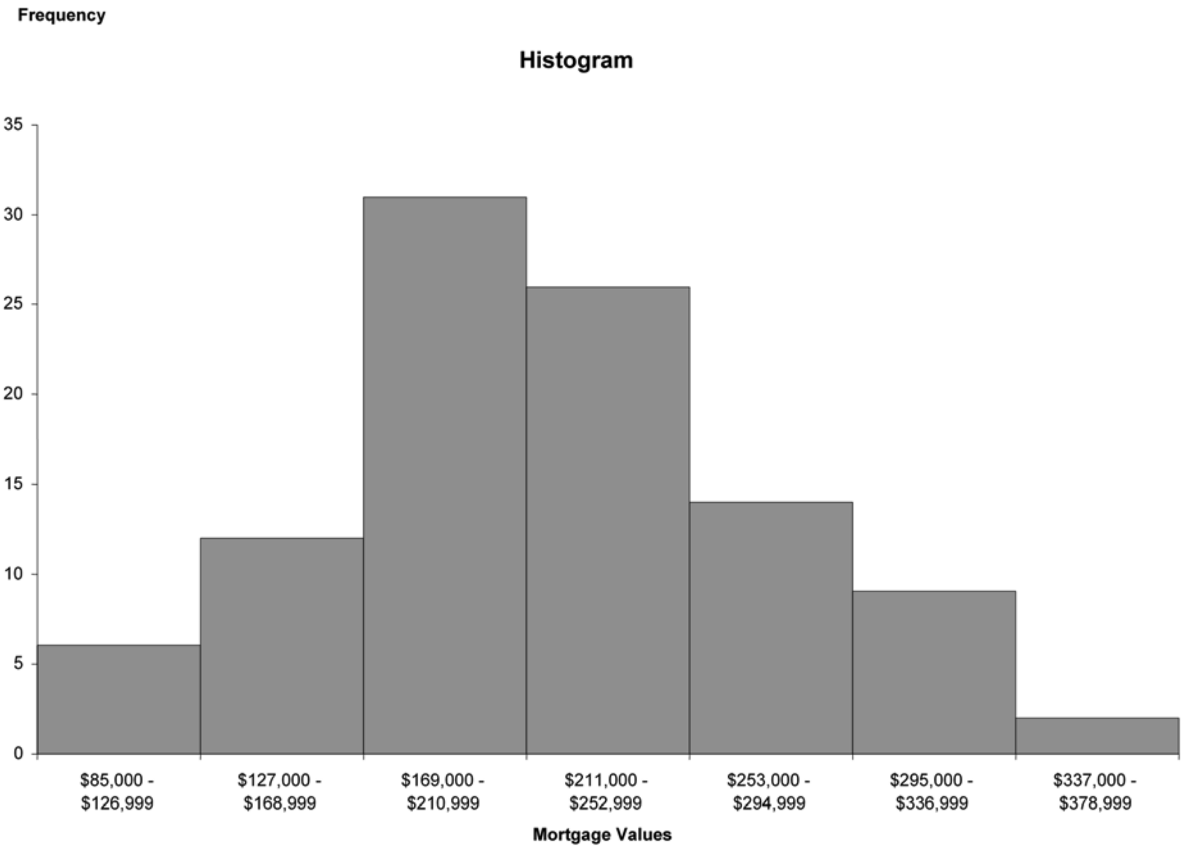
Rounding this up to the nearest $1,000, the class width is $42,000.

c. The frequency distribution with seven classes and a class width of $42,000 will depend on the starting point for the first class. This starting value must be at or below the minimum value of $87,429. Student answers will vary depending on the starting point. We have used $85,000. Care should be made to make sure that the classes are mutually exclusive and all-inclusive. The following frequency distribution is developed:



Group	Frequency
$85,000 - $126,999	6
$127,000 - $168,999	12
$169,000 - $210,999	31
$211,000 - $252,999	26
$253,000 - $294,999	14
$295,000 - $336,999	9
$337,000 - $378,999	2
Total =	100

d. The histogram for the frequency distribution in part c is shown as follows:



Interpretation should involve a discussion of the range of values with a discussion of where the major classes are located.

2.14. a. $w = \dfrac{\text{Largest} - \text{Smallest}}{\text{Number of Classes}} = \dfrac{214.4 - 112.6}{11} = 9.255 \rightarrow w = 10.$

The salaries in the first class are $(105,\ 105+10) = (105,\ 115)$. The frequency distribution follows

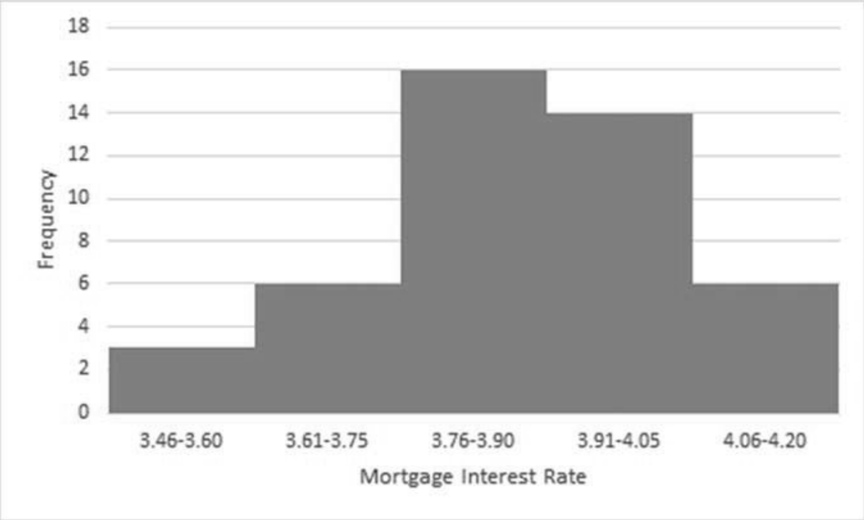
Classes	Frequency	Relative Frequency	Cumulative Relative Frequency
(105 – <115)	1	0.04	0.04
(115 – <125)	1	0.04	0.08
(125 – <135)	2	0.08	0.16
(135 – <145)	1	0.04	0.20
(145 – <155)	1	0.04	0.24
(155 – <165)	7	0.28	0.52
(165 – <175)	4	0.16	0.68
(175 – <185)	3	0.12	0.80
(185 – <195)	2	0.08	0.88
(195 – <205)	0	0.00	0.88
(205 – <215)	3	0.12	1.00

b. The data shows 8 of the 25, or 0.32 of the salaries are at least 175,000

c. The data shows 18 of the 25, or 0.72 having salaries that are at most $205,000 and a least $135,000.

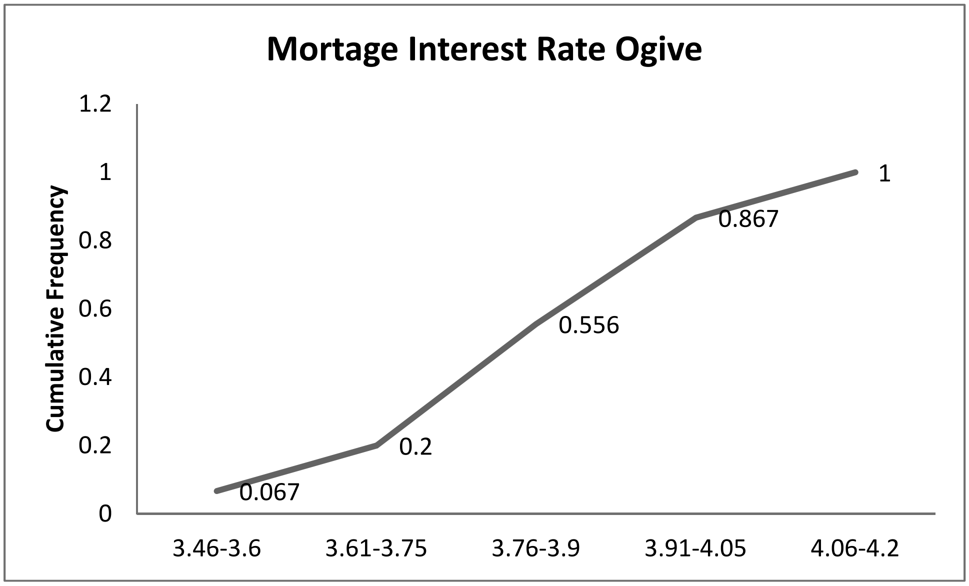
2.15. a. We are assuming mortgage rates are limited to two decimal places. Students making other assumptions will get a slightly difference histogram. We are also rounding the calculated class width to .15.

Class	Frequency	Relative Frequency
3.46–3.60	3	0.067
3.61–3.75	6	0.133
3.76–3.90	16	0.356
3.91–4.05	14	0.311
4.06–4.20	6	0.133

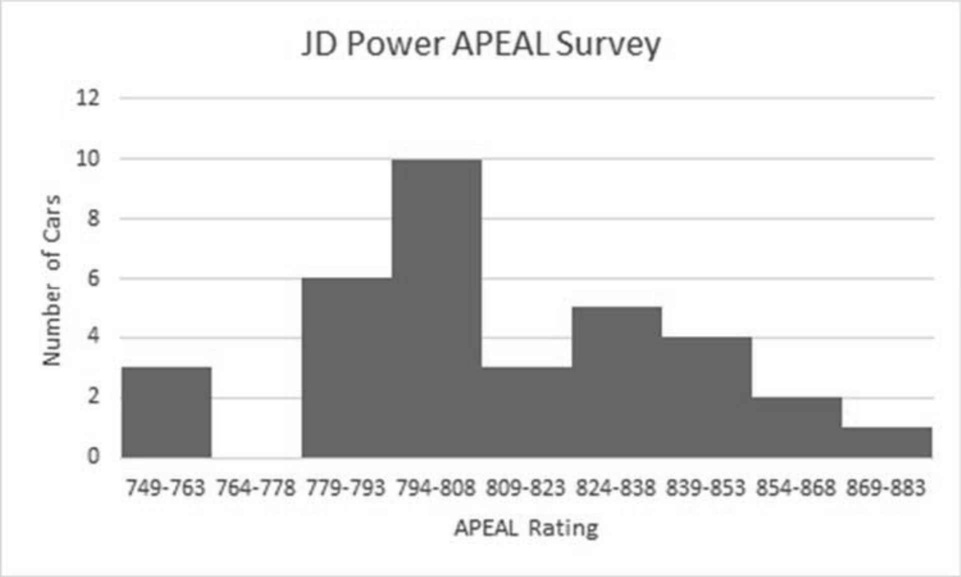


b. Proportion of rates that are at least 3.76% is the sum of the relative frequencies of the last four classes = 0.356 + 0.311 + 0.133 = 0.800

c.



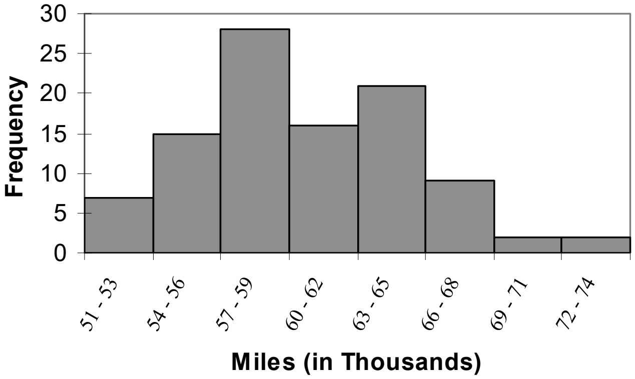
2.16. a.



b. The 2015 average is 798 which exceeds the 2008 average of 790. This could indicate that the new models are slightly more appealing on average to automobile customers, or customers could simply have reduced expectations.

2.17. a.

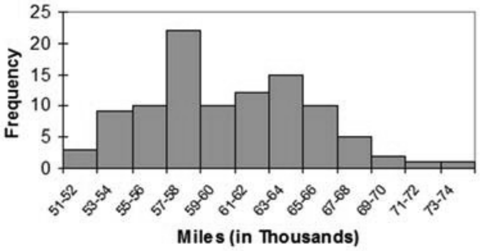
Classes	Frequency
51–53	7
54–56	15
57–59	28
60–62	16
63–65	21
66–68	9
69–71	2
72–74	2



b. The tread life of at least 50% of the tires is 60,000 or more. The top 10% is greater than 66,000 and the longest tread tire is 74,000. Additional information will vary.

c.

Classes	Frequency
51–52	3
53–54	9
55–56	10
57–58	22
59–60	10
61–62	12
63–64	15
65–66	10
67–68	5
69–70	2
71–72	1
73–74	1



Students will probably say that the 12 classes give better information because it allows you to see more detail about the number of miles the tires can go.

2.18. a. There are $n = 294$ values in the data. Then using the $2^k \geq n$ guideline we would need at least $k = 9$ classes.

b. Using $k = 9$ classes, the class width is determined as follows:

$$w = \frac{\text{High} - \text{Low}}{\text{Classes}} = \frac{32 - 10}{9} = \frac{22}{9} = 2.44$$

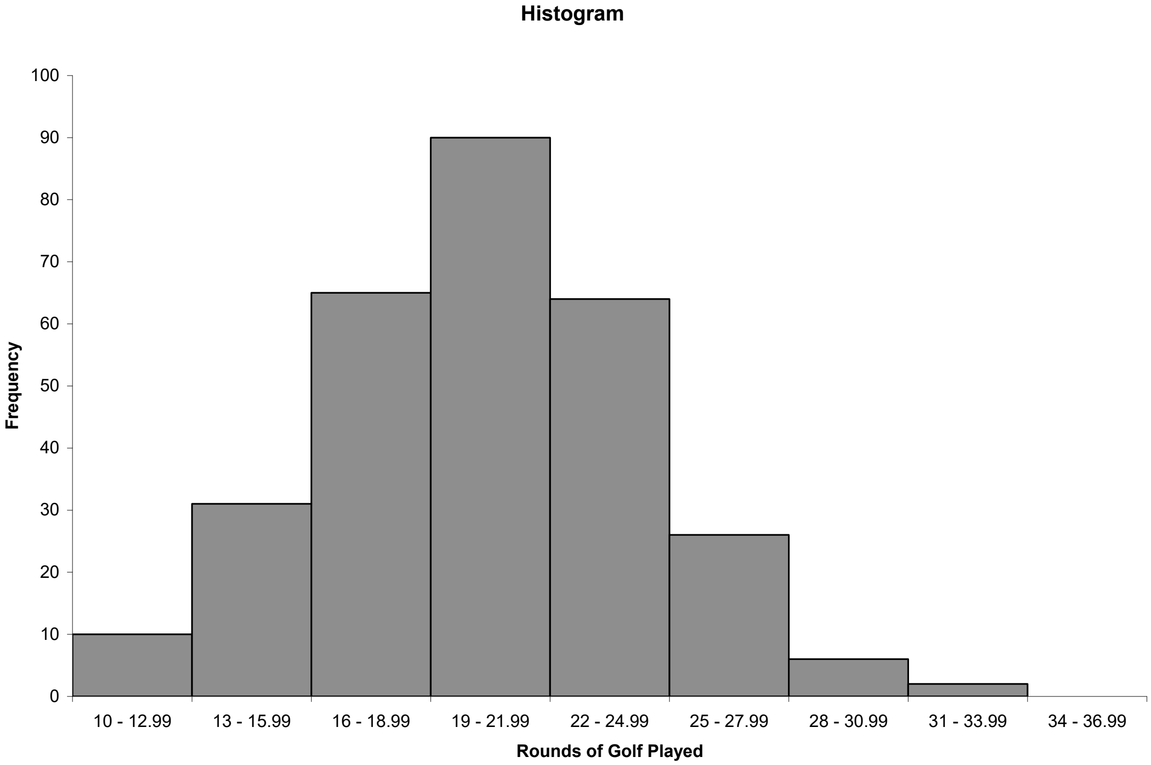
Rounding this up to the nearest 1.0, the class width is 3.0.

c. The frequency distribution with nine classes and a class width of 3.0 will depend on the starting point for the first class. This starting value must be at or below the minimum value of 10. Student answers will vary depending on the starting point. We have used 10 as it is nice round number. Care should be made to make sure that the classes are mutually exclusive and all-inclusive. The following frequency distribution is developed:

Rounds	Frequency
10, 11, 12	10
13, 14, 15	31
16, 17, 18	65
19, 20, 21	90
22, 23, 24	64
25, 26, 27	26
28, 29, 30	6
31, 32, 33	2
34, 35, 36	0
Total	294

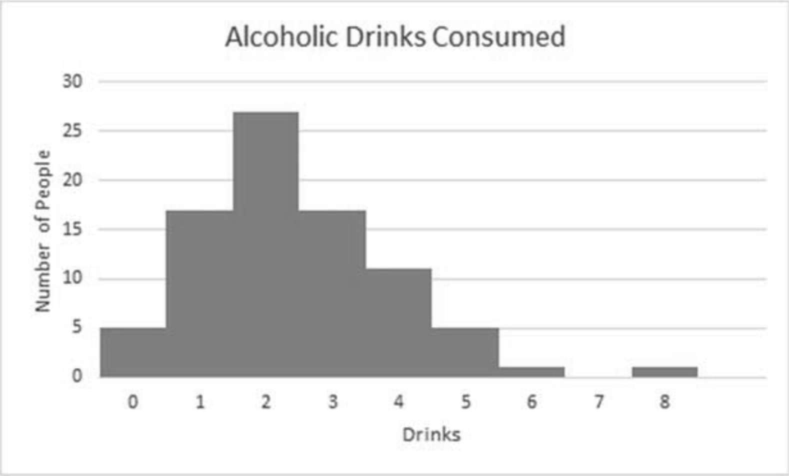
Students should recognize that by rounding the class width up from 2.44 to 3.0, and by starting the lowest class at the minimum value of 10, the 9th class is actually not needed.

Based on the results in part c, the frequency histogram is shown as follows:



The distribution for rounds of golf played is mound shaped and fairly symmetrical. It appears that the center is between 19 and 22 rounds per year, but the rounds played is quite spread out around the center.

2.19.



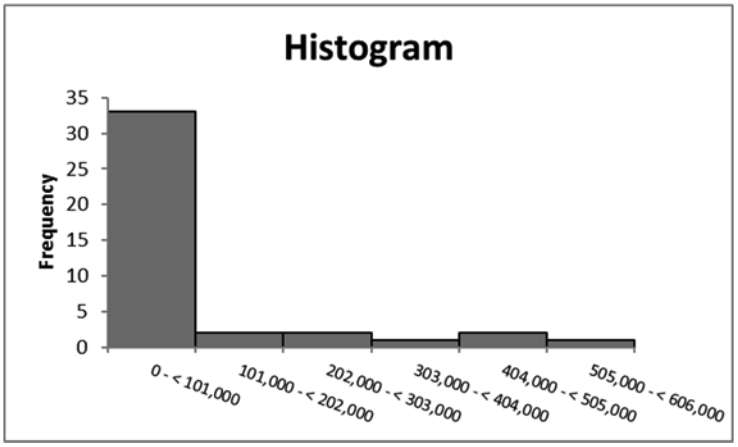
2.20. a. Using the $2^k \geq n$ guideline, the number of classes would be 6. There are 41 airlines. $2^5 = 32$ and $2^6 = 64$. Therefore, 6 classes are chosen.

b. The maximum value is 602,708 and the minimum value is 160 from the Total column. The difference is $602,708 - 160 = 602548$. The class width would be $602548/6 = 100424.67$. Rounding up to the nearest 1,000 produces a class width of 101,000.

c.

Class	Frequency
0 – < 101,000	33
101,000 – < 202,000	2
202,000 – < 303,000	2
303,000 – < 404,000	1
404,000 – < 505,000	2
505,000 – < 606,000	1

Histogram follows:



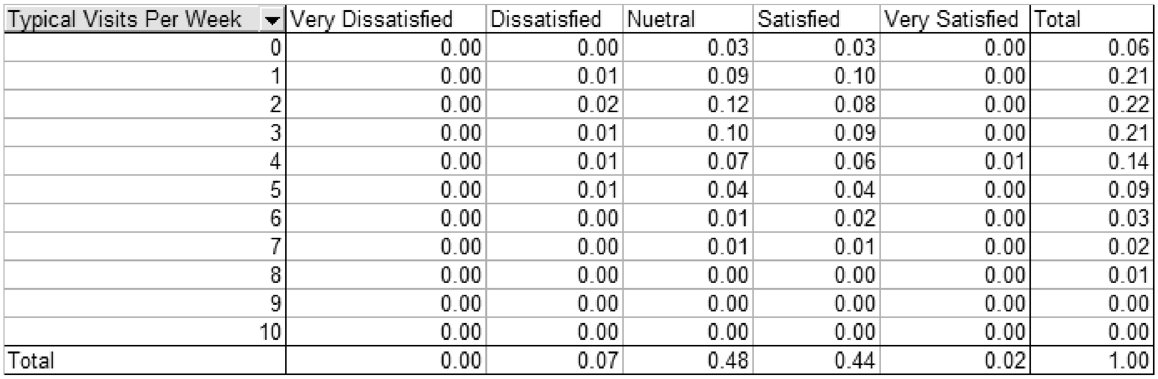
The vast majority of airlines had fewer than 101,000 monthly passengers for December 2011.

2.21. a. The frequency distribution is:

Satisfaction Level	Frequency
1 = Very Dissatisfied	1
2 = Dissatisfied	82
3 = Neutral	578
4 = Satisfied	530
5 = Very Satisfied	23
	Total = 1214

The frequency distribution shows that over 1,100 people rated the overall service as either neutral or satisfied. While only 83 people expressed dissatisfaction, the manager should be concerned that so many people were in the neutral category. It looks like there is much room for improvement.

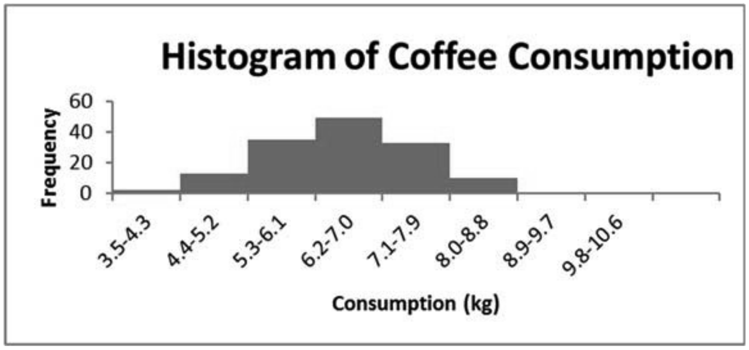
b. The joint relative frequency distribution for “Overall Service Satisfaction” and “Number of Visits Per Week” is:



Typical Visits Per Week	Very Dissatisfied	Dissatisfied	Nuetral	Satisfied	Very Satisfied	Total
0	0.00	0.00	0.03	0.03	0.00	0.06
1	0.00	0.01	0.09	0.10	0.00	0.21
2	0.00	0.02	0.12	0.08	0.00	0.22
3	0.00	0.01	0.10	0.09	0.00	0.21
4	0.00	0.01	0.07	0.06	0.01	0.14
5	0.00	0.01	0.04	0.04	0.00	0.09
6	0.00	0.00	0.01	0.02	0.00	0.03
7	0.00	0.00	0.01	0.01	0.00	0.02
8	0.00	0.00	0.00	0.00	0.00	0.01
9	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00
Total	0.00	0.07	0.48	0.44	0.02	1.00

The people who expressed dissatisfaction with the service tended to visit 5 or fewer times per week. While 38% of the those surveyed both expressed a neutral rating and visited the club between 1 and 4 times per week.

2.22. a. The histogram can be created from the frequency distribution. The classes are shown on the horizontal axis and the frequency on the vertical axis. The histogram is shown below.



The histogram shows the shape of the distribution. This histogram is showing that fewer people consume small and large quantities and that most individuals consume between 5.3 and 8.0 kg of coffee, with the highest percentage of individuals consuming between 6.2 and 7.0.

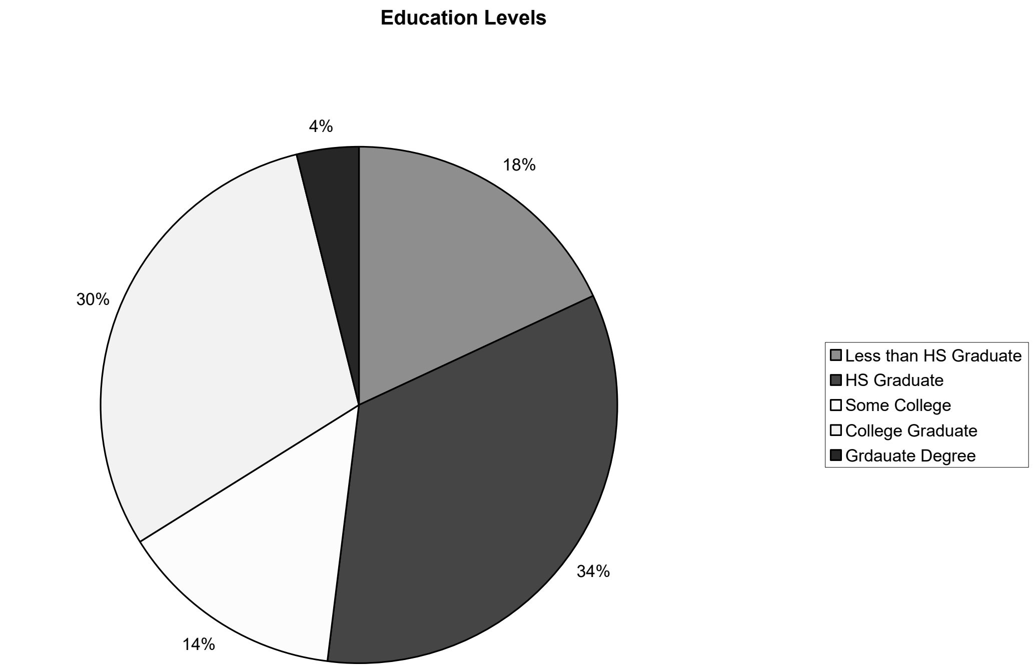
b. Convert the frequency distribution into relative frequencies and cumulative relative frequencies as shown below.

Consumption	Frequency	Relative Frequency	Cumulative Relative Frequency
3.5–4.3	2	0.0139	0.0139
4.4–5.2	13	0.0903	0.1042
5.3–6.1	35	0.2431	0.3472
6.2–7.0	49	0.3403	0.6875
7.1–7.9	33	0.2292	0.9167
8.0–8.8	10	0.0694	0.9861
8.9–9.7	1	0.0069	0.9931
9.8–10.6	1	0.0069	1

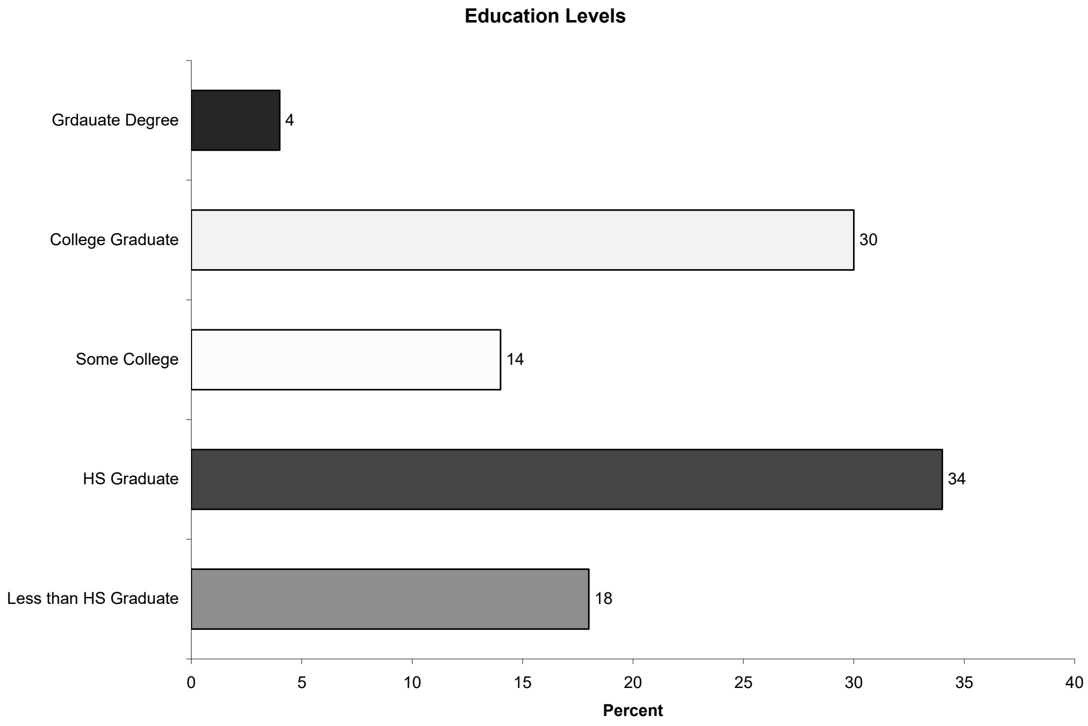
8.33% (100–91.67) of the coffee drinkers sampled consumes 8.0 kg or more annually.

Section 2.2

2.23. a. The pie chart is shown as follows:



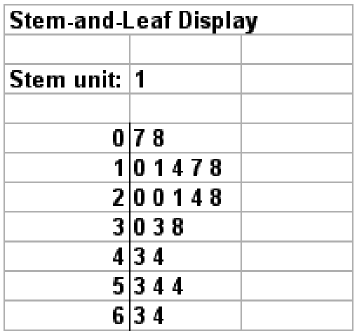
b. The horizontal bar chart is shown as follows:



2.24. Step 1: Sort the data from low to high. The lowest value is 0.7 and the highest 6.4.
Step 2: Split the values into a stem and leaf. Stem = units place leaf = decimal place

Step 3: List all possible stems from lowest to highest.

Step 4: Itemize the leaves from lowest to highest and place next to the appropriate stems.



Stem-and-Leaf Display		
Stem unit:	1	
0	7 8	
1	0 1 4 7 8	
2	0 0 1 4 8	
3	0 3 8	
4	3 4	
5	3 4 4	
6	3 4	

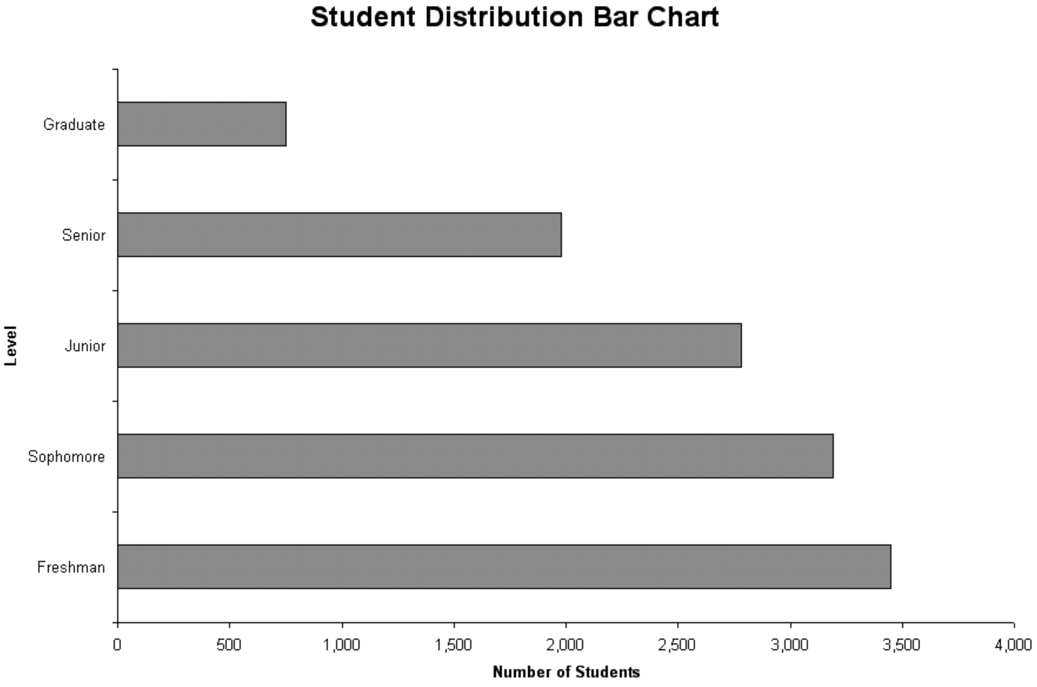
2.25. a. Step 1: Define the categories.

The categories are grade level.

Step 2: Determine the appropriate measure.

The measure is the number of students at each grade level.

Step 3: Develop the bar chart.



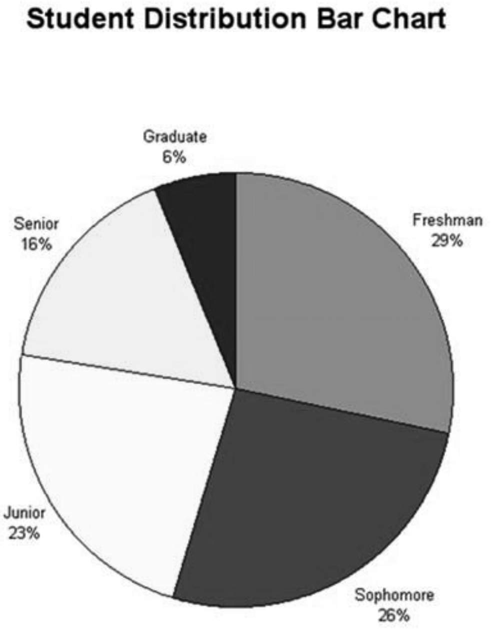
b. Step 1: Define the categories.

The categories are grade level.

Step 2: Determine the appropriate measure.

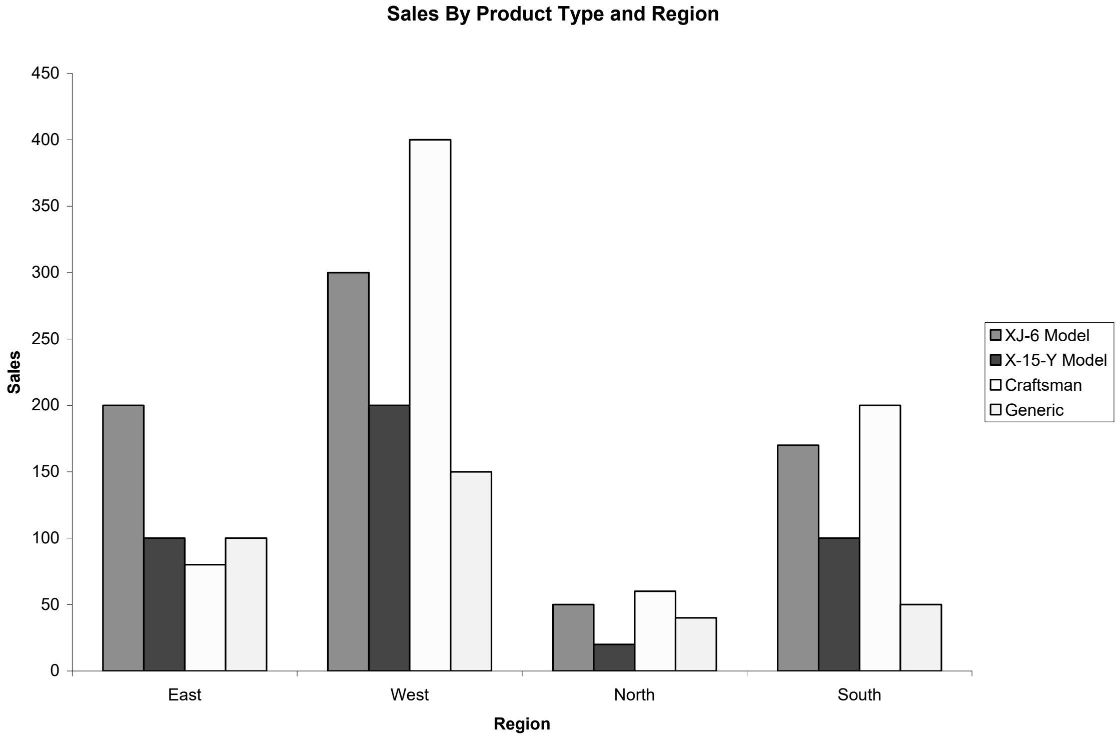
The measure is the number of students at each grade level.

Step 3: Develop the pie chart.

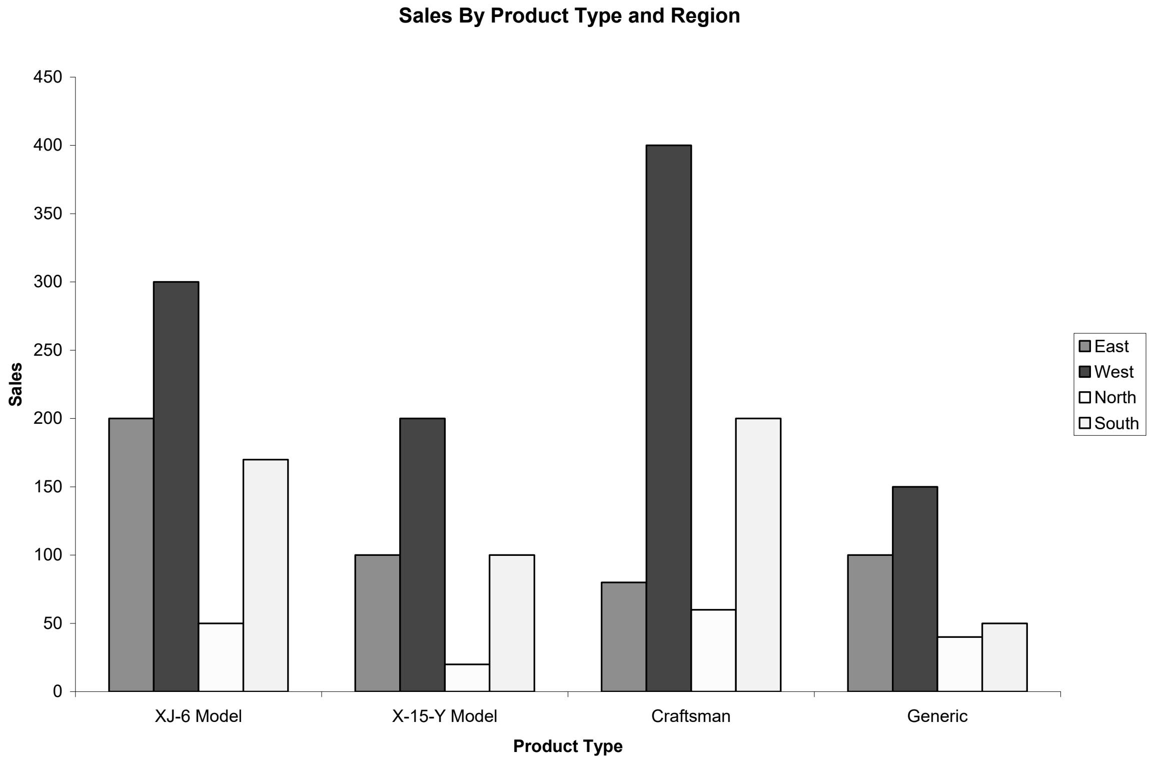


c. A case can be made for either a bar chart or pie chart. Pie charts are especially good at showing how the total is divided into parts. The bar chart is best to draw attention to specific results. In this case, a discussion might be centered on the possible attrition that takes place in the number of students between Freshman and Senior years.

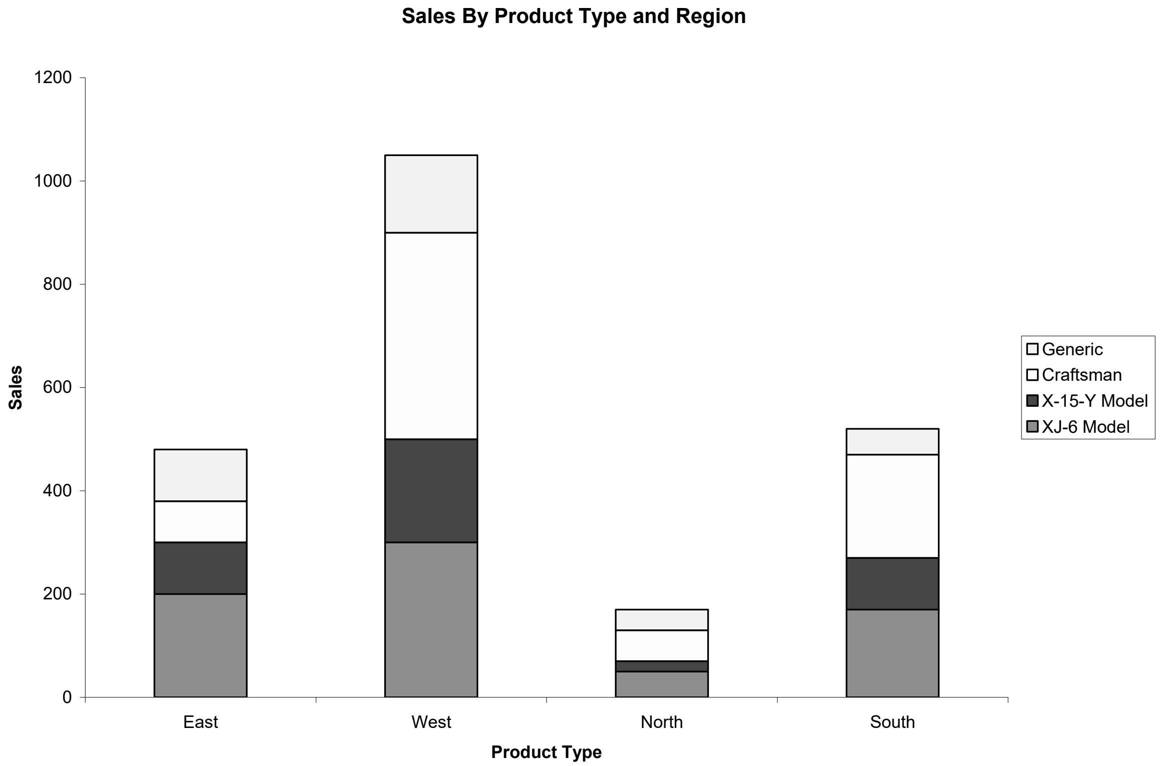
2.26. One possible bar chart is shown as follows:



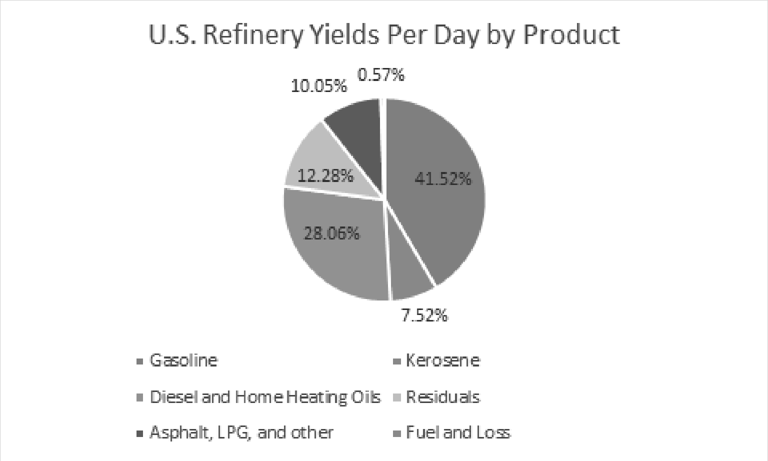
Another way to present the same data is:



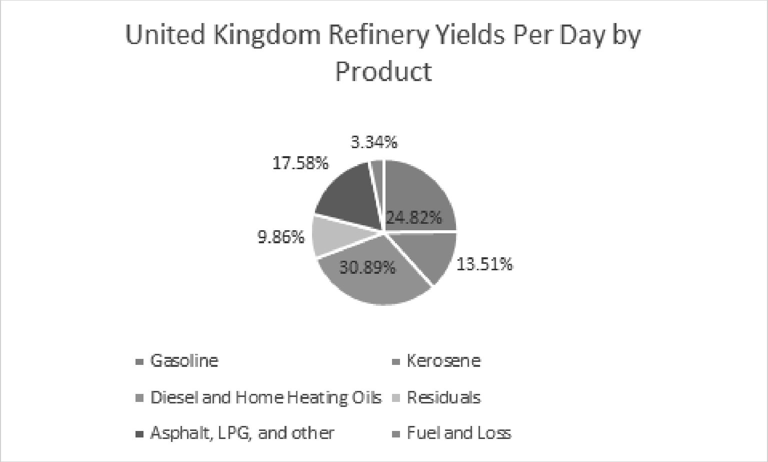
Still another possible way is called a "stacked" bar chart.



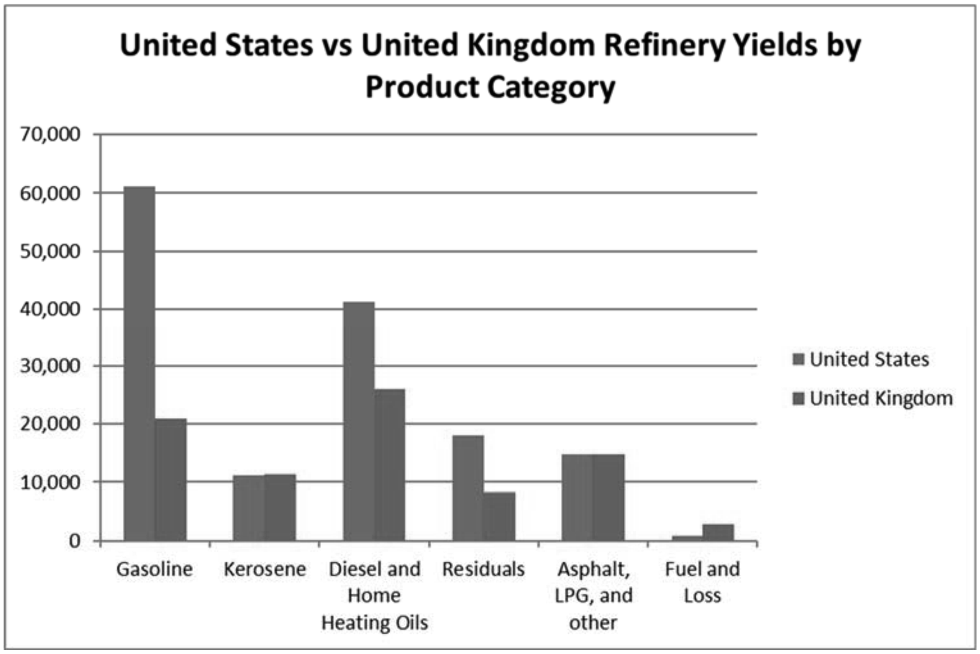
2.27. a.


U.S. Refinery Yields Per Day by Product
10.05%
0.57%
12.28%
41.52%
28.06%
7.52%
Gasoline
Kerosene
Diesel and Home Heating Oils
Residuals
Asphalt, LPG, and other
Fuel and Loss

b.

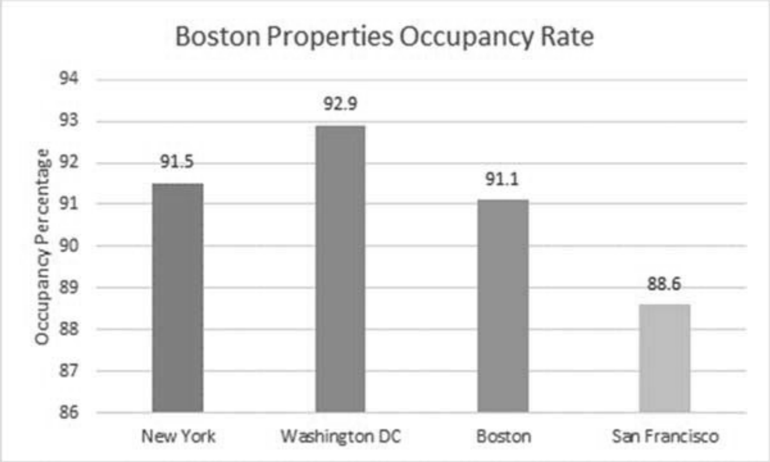

United Kingdom Refinery Yields Per Day by Product
3.34%
17.58%
24.82%
9.86%
30.89%
13.51%
Gasoline
Kerosene
Diesel and Home Heating Oils
Residuals
Asphalt, LPG, and other
Fuel and Loss

c.


United States vs United Kingdom Refinery Yields by Product Category
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
Gasoline
Kerosene
Diesel and Home Heating Oils
Residuals
Asphalt, LPG, and other
Fuel and Loss
United States
United Kingdom

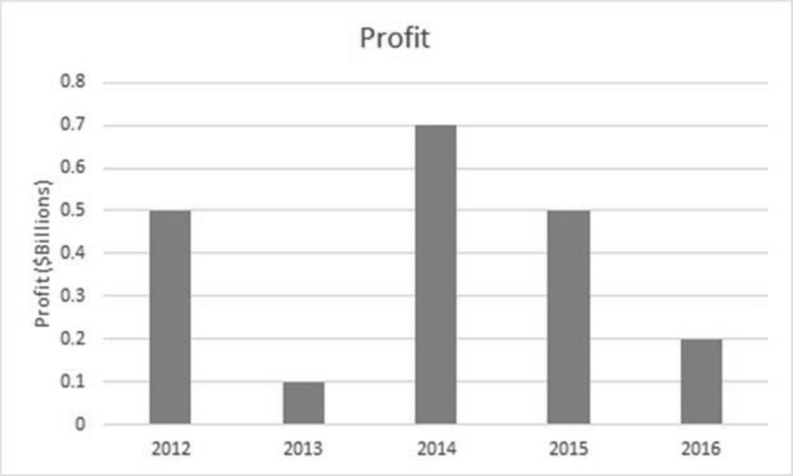
2.28. a. The pie chart would not be appropriate because pie charts are used to illustrate how a total is split between catagories. In this case, the sum of the percentages for the four cities is a meaningless number.

b.

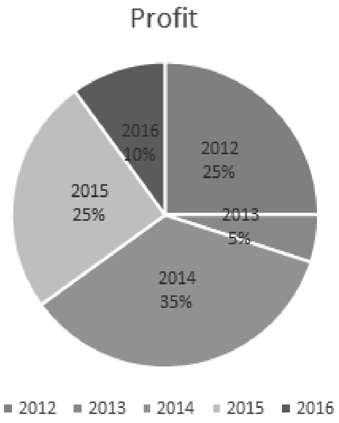


c. The bar chart makes it easier to compare percentages across cities. The pie chart is not appropriate as discussed in part a.

2.29 a.

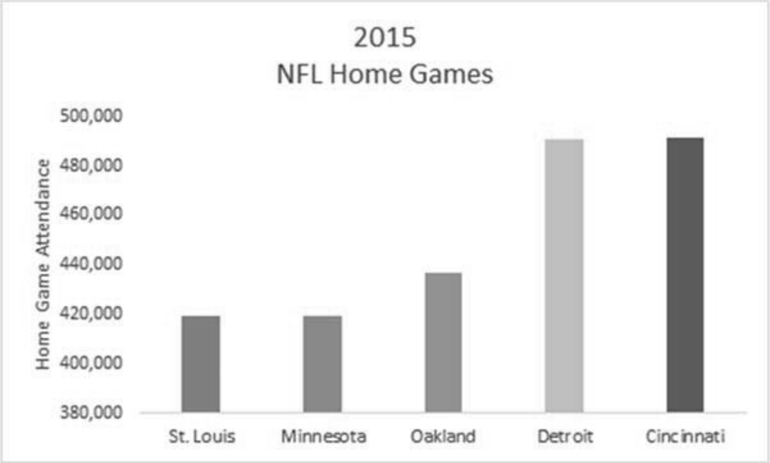


b.



c. Arguments exist for both the pie chart and the bar chart. Pie charts are especially good at showing how the total is divided into parts. The bar chart is best to draw attention to specific results. In this case, it is most likely that the historic change in profits is to be displayed. The bar chart is best at presenting time defined data.

2.30.



2.31. a. Pie charts are typically used to show how a total is divided into parts. In this case, the total of the five ratios is not a meaningful value. Thus, a pie chart showing each ratio as a fraction of the total would not be meaningful. Thus a pie chart is not the most appropriate tool. A bar chart would be appropriate.

b. Step 1: Define the categories.

The categories are the five cities where the plants are located

Step 2: Determine the appropriate measure.

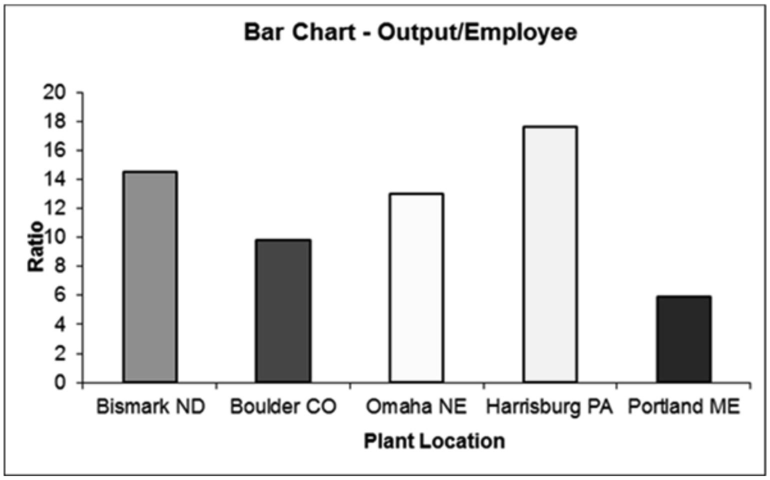
The measure of interest is the ratio of manufactured output to the number of employees at the plant.

Step 3: Develop the bar chart.

The bar chart is shown as follows:

Step 4: Interpret the results.

It appears that the number of units manufactured per employee of the plants in the Midwest is larger than in the West.



2.32. Step 1: Define the categories.

The categories are the five years, 2000, 2001…., 2004

Step 2: Determine the appropriate measure.

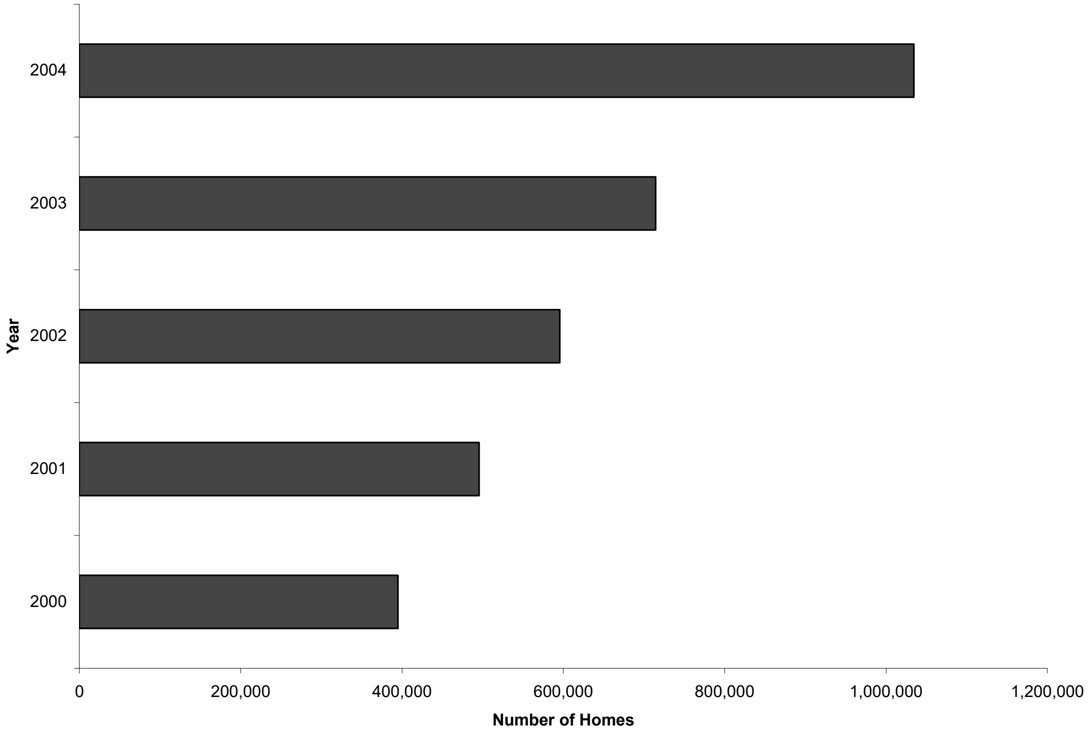
The measure of interest is the number of homes that have a value of $1 million or more.

Step 3: Develop the bar chart.

The horizontal bar chart is shown as follows:

Step 4: Interpret the results.

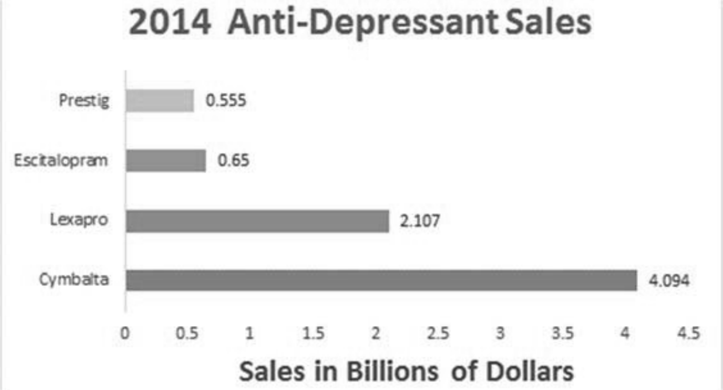
The bar chart is skewed below indicating that number of $1 Million houses was growing rapidly. It also appears that that growth is not linear.



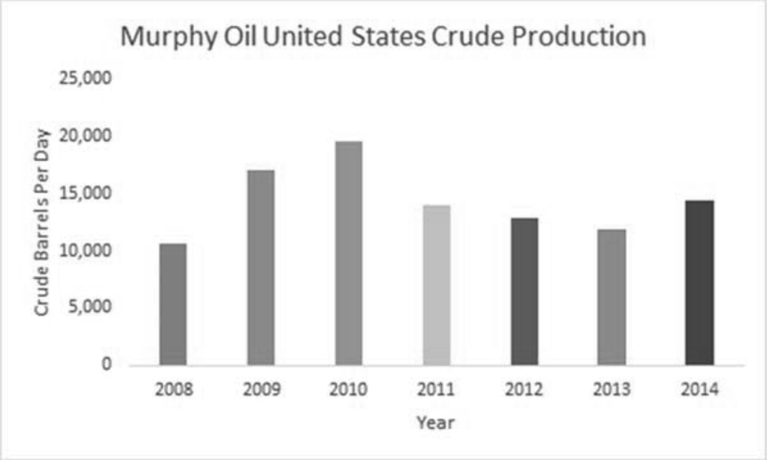
Step 4: Interpret the results.

The bar chart is skewed below indicating that number of $1 Million houses is growing rapidly. It also appears that that growth is exponential rather than linear.

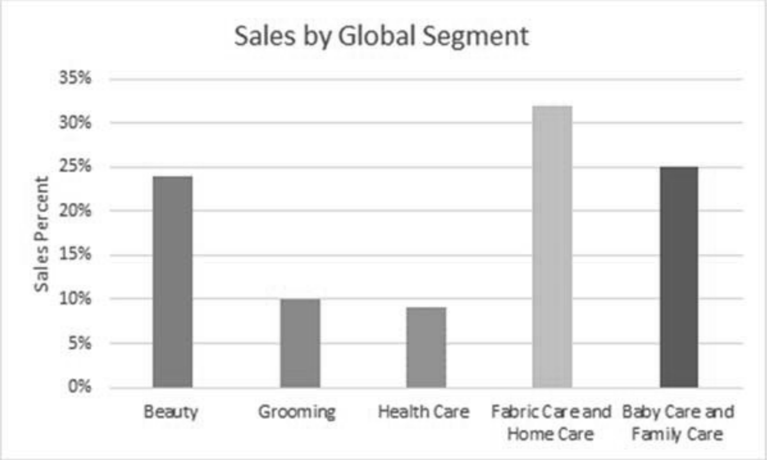
2.33. One appropriate graph for these data is a horizontal bar chart



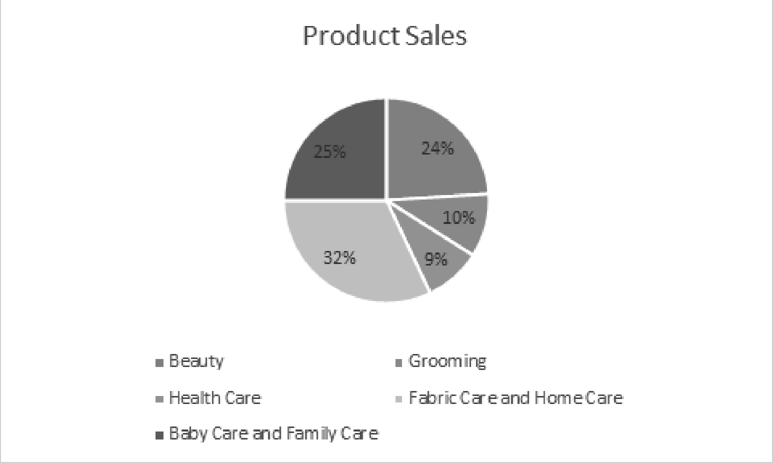
2.34.



2.35. a. The bar chart is shown below. The categories are the Global Segments and the measure for each category is the percent of total sales for the Global Segment.



b. The pie chart is shown below. The categories are the Global Segments and the measure is the proportion of each segment's total net sales.



2.36. a. The following stem and leaf diagram was created using PhStat. The stem unit is 10 and the leaf unit is 1.

Stem-and-Leaf Display for Drive-Thru Service (Seconds)

Stem unit: 10

6	8
7	1 3 4 6 9
8	3 5 8
9	0 2 3
10	3 5
11	0 6 9
12	
13	0 4 8
14	5 6 7
15	6 6
16	2
17	8
18	1

b. The most frequent speed of service is between 70 and 79 seconds.

2.37. a. The following stem-and-leaf diagram was developed using PhStat. The stem unit is 10 and the leaf unit is 1.

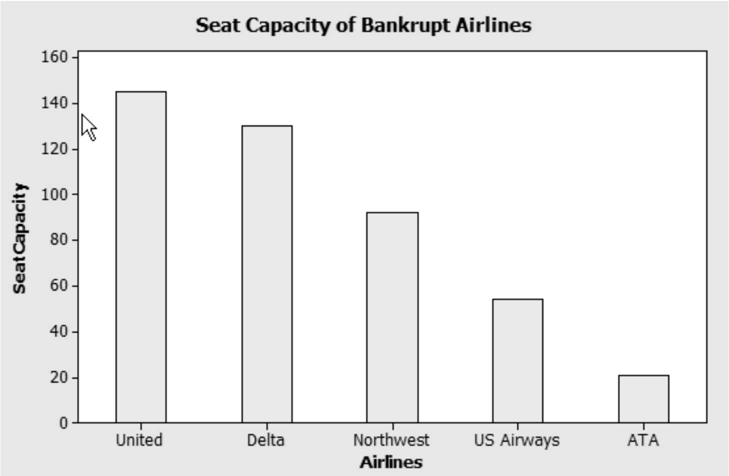
Stem-and-Leaf Display for Number of Days to Collect Payment

Stem unit: 10

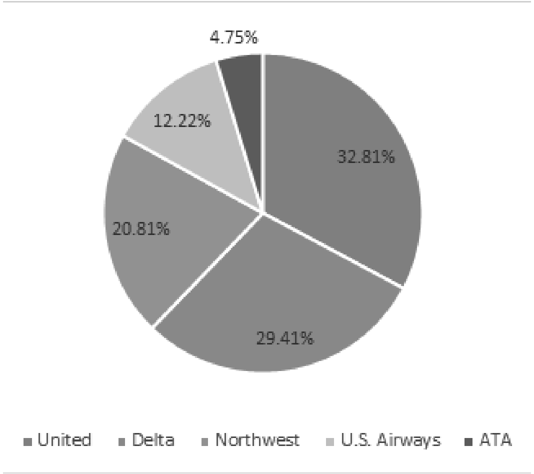
2	2 4 8 9
3	0 1 2 3 3 4 5 5 5 6 6 7 8 8 9
4	1 3 5 7 8
5	5 6 6
6	0 5 6

b. Most payments are collected in the range of 30–39 days.

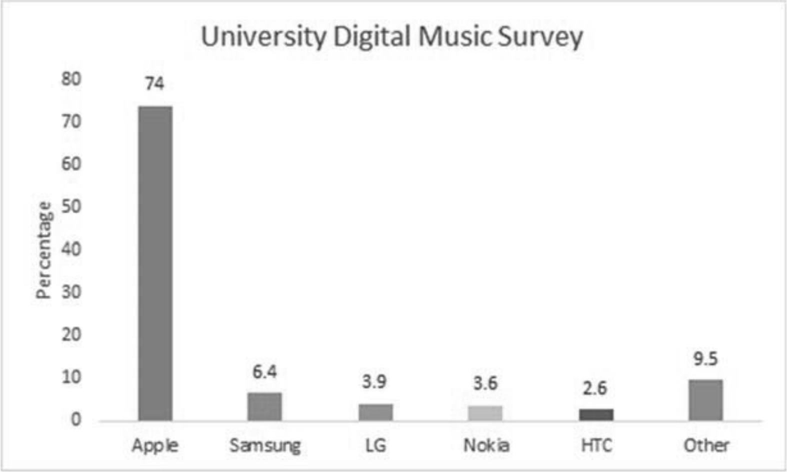
2.38. a. The bar graph is



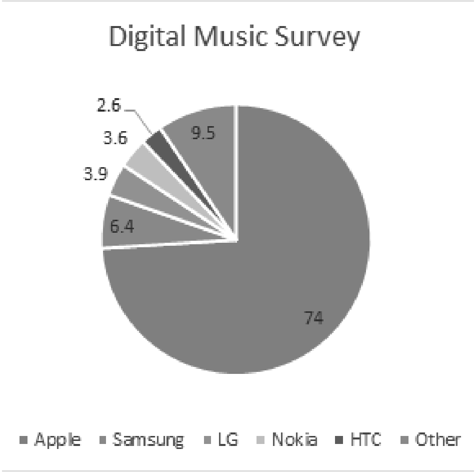
b. The percent equals the individual capacity divided by the total, e.g. United $\rightarrow$ percent $= (145/442)100\% = 32.81\%$, etc. This produces the following pie chart:



2.39. a.

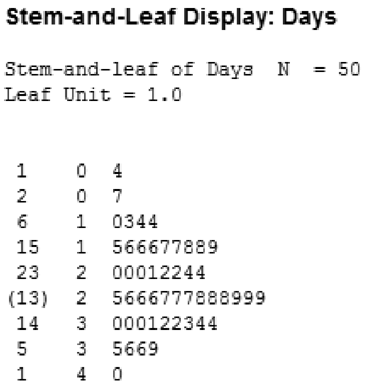


b.



c. A case can be made for either a bar chart or a pie chart. Pie charts are especially good at showing how the total is divided into parts. The bar chart is best to draw attention to specific results. In this case, a discussion might be centered on the relative large percentage attributable to Apple.

2.40.



Stem-and-Leaf Display: Days

```
Stem-and-leaf of Days  N  = 50
Leaf Unit = 1.0


  1    0  4
  2    0  7
  6    1  0344
 15    1  566677889
 23    2  00012244
(13)   2  5666777888999
 14    3  000122344
  5    3  5669
  1    4  0
```

2.41. a. A bar chart is an appropriate graph since there are two categories, males and females. A pie chart could also be used to display the data.

b. The following steps are used to construct the bar chart:

Step 1: Define the categories.

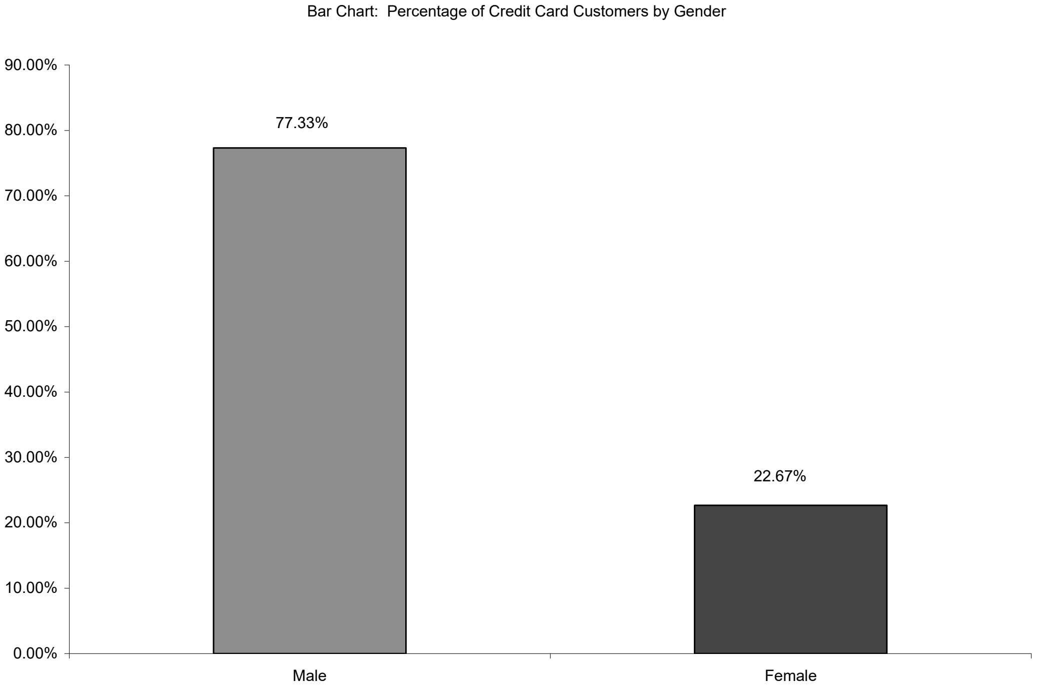
The categories are the two genders, male and female

Step 2: Determine the appropriate measure.

The measure of interest is the percentage of credit card holders who are male and female.

Step 3: Develop the bar chart using computer software such as Excel or Minitab.

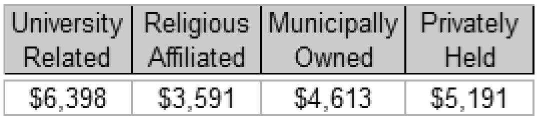
The bar chart is shown as follows:



Step 4: Interpret the results.

This shows that a clear majority of credit card holders are males (77.33%)

2.42. a. The following are the averages for each hospital computed by summing the charges and dividing by the number of charges:



University Related	Religious Affiliated	Municipally Owned	Privately Held
$6,398	$3,591	$4,613	$5,191

b. The following steps are used to construct the bar chart:

Step 1: Define the categories.

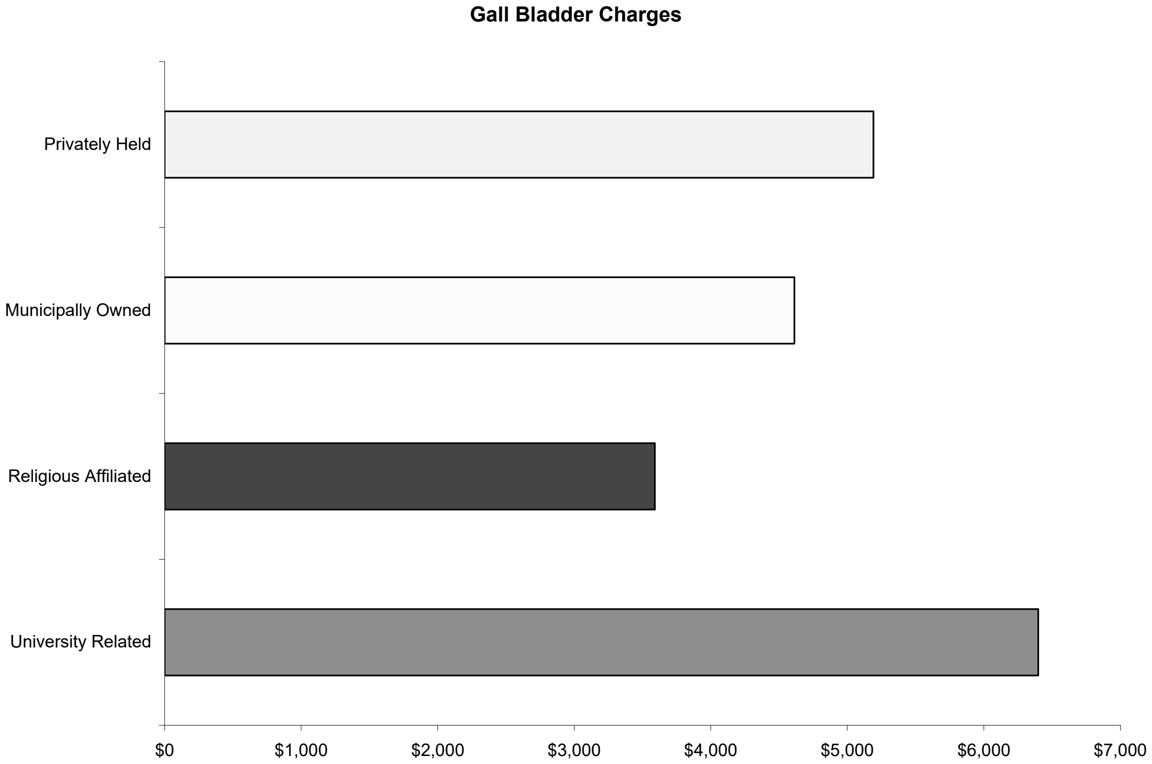
The categories are the four hospital types

Step 2: Determine the appropriate measure.

The measure of interest is the average charge for outpatient gall bladder surgery.

Step 3: Develop the bar chart using computer software such as Excel or Minitab.

The bar chart is shown as follows:

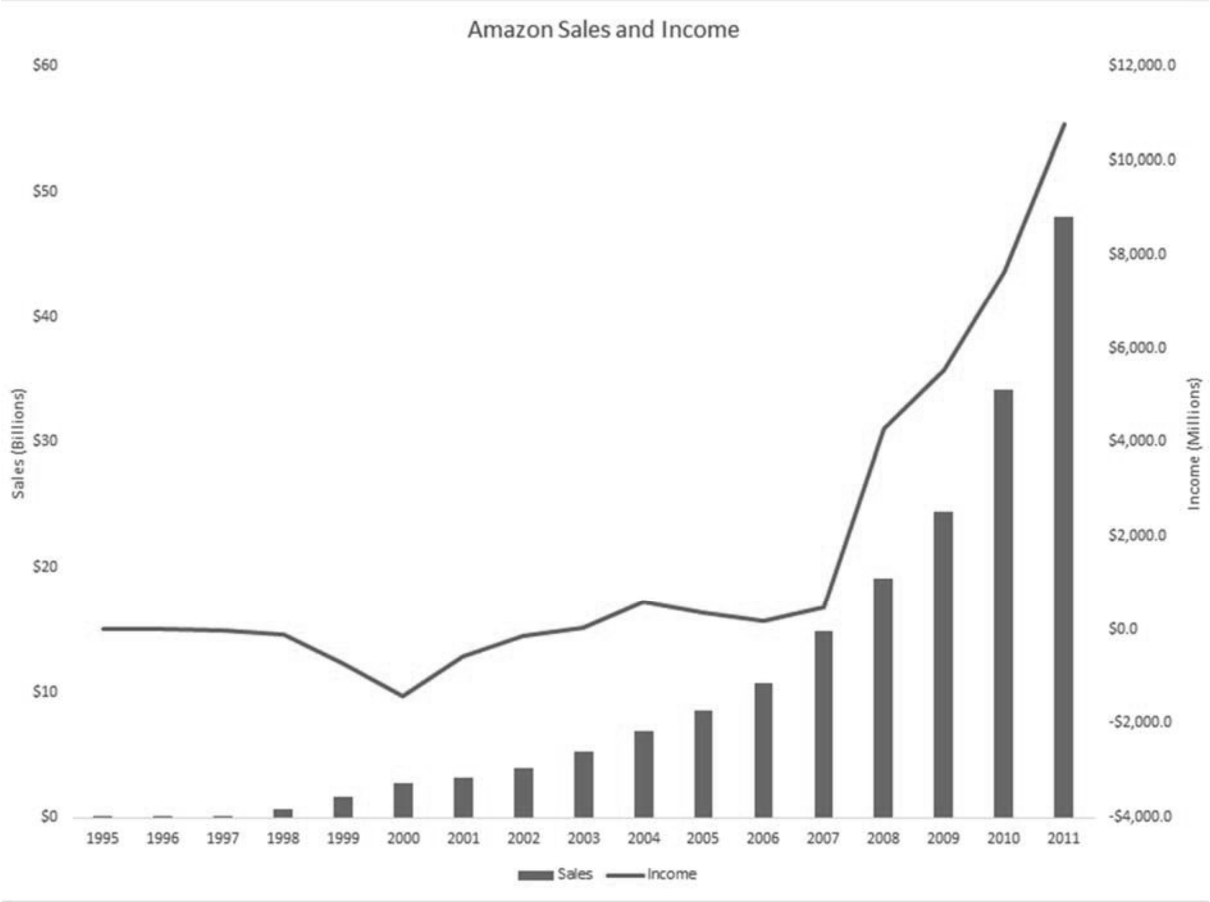


Step 4: Interpret the results.

The largest average charges occurred for gall bladder surgery appears to be in University Related hospitals and the lowest average appears to be in Religious Affiliated hospitals.

c. A pie chart is used to display the parts of a total. In this case the total charges of the four hospital types is not a meaningful number so a pie chart showing how that total is divided among the four hospital types would not be useful or appropriate.

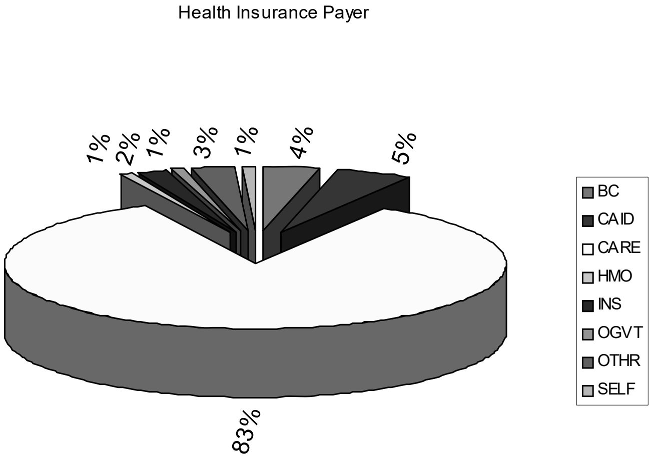
2.43. a.



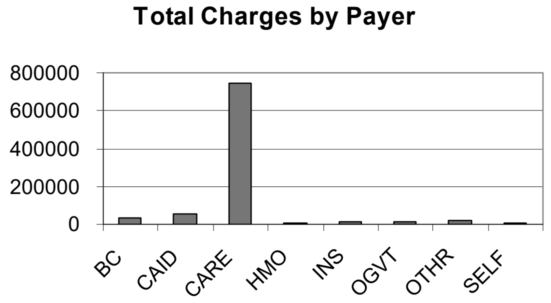
b. There appears to be a linear relationship between sales and years in which the sales were made.

c. In time period between 2000 and 2001, Amazon experienced a decrease in its losses. Prior to this time, each year produced increased losses.

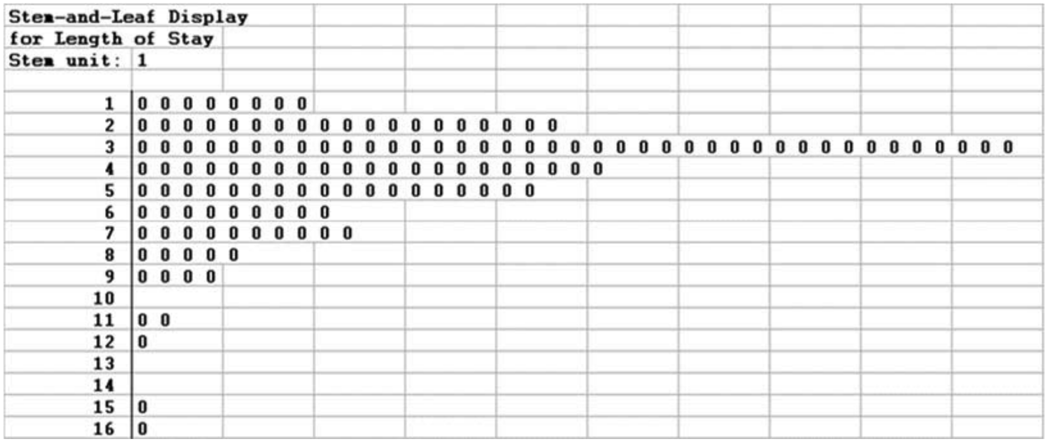
2.44. a.



b.

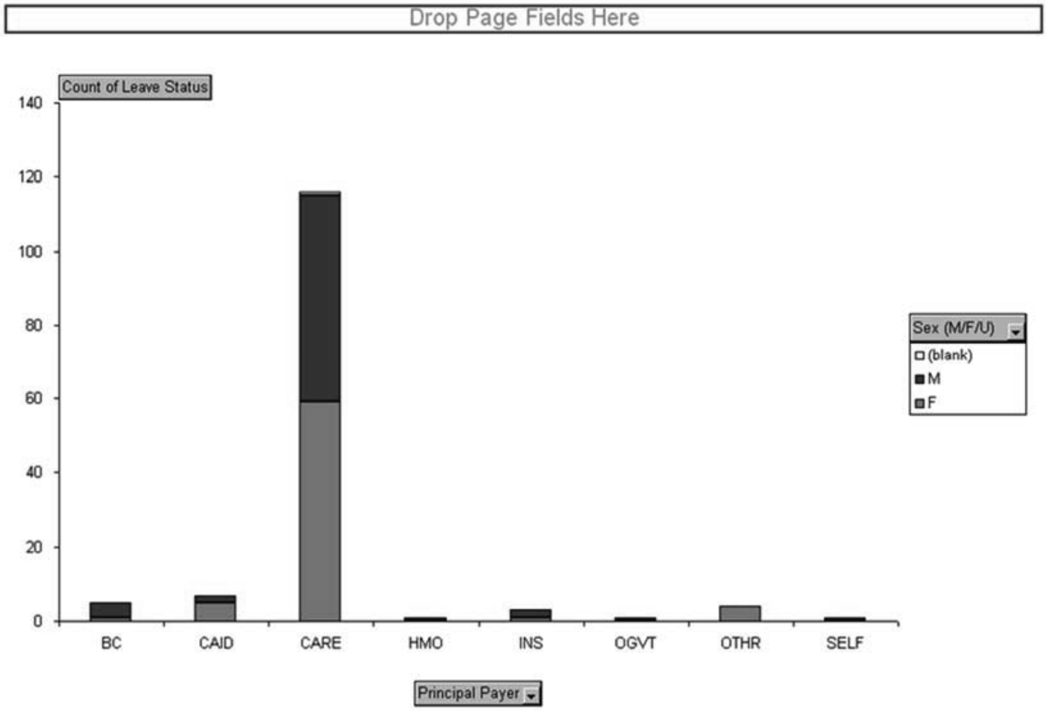


c. Using PHStat the stem & leaf diagram is shown as follows.



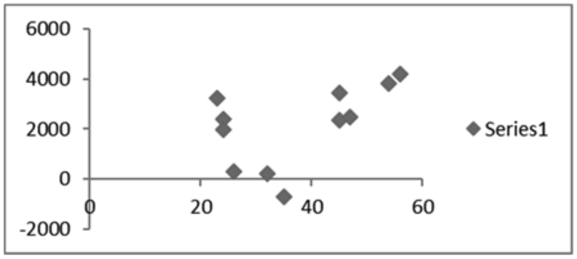
```
Stem-and-Leaf Display
for Length of Stay
Stem unit:  1

   1 | 0 0 0 0 0 0 0 0
   2 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
   3 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
   4 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
   5 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
   6 | 0 0 0 0 0 0 0 0 0
   7 | 0 0 0 0 0 0 0 0 0 0
   8 | 0 0 0 0 0
   9 | 0 0 0 0
  10 |
  11 | 0 0
  12 | 0
  13 |
  14 |
  15 | 0
  16 | 0
```

d. Excel's pivot table can be used to develop a bar chart. The chart showed is a stacked bar chart.



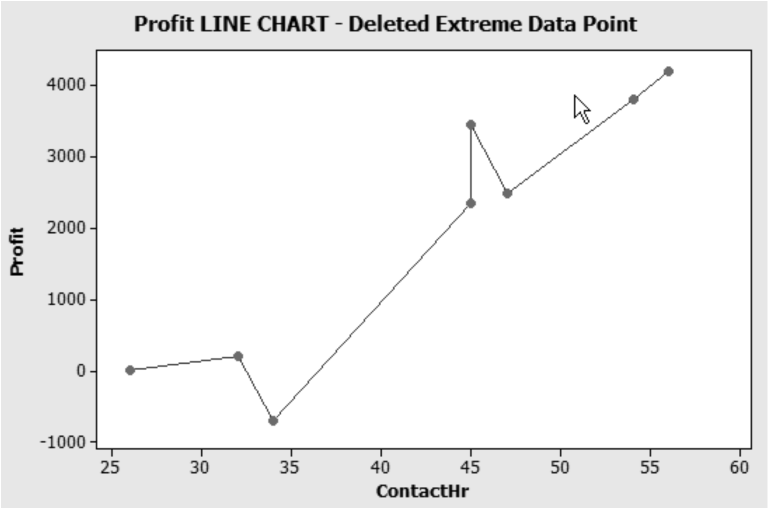
Section 2.3

2.45. a.



There appears to be a curvilinear relationship between the dependent and independent variables.

b.



Having removed the extreme data points, the relationship between dependent and independent variables seems to be linear and positive.

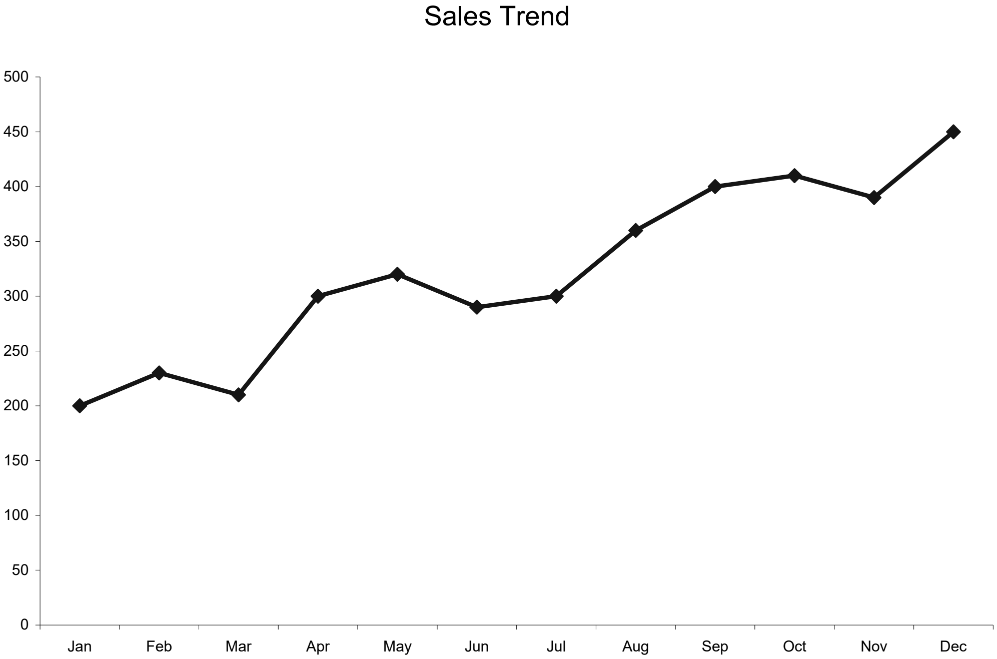
2.46. Step 1: Identify the time-series variable

The variable of interest is the monthly sales.

Step 2: Layout the Horizontal and Vertical Axis

The horizontal axis will be month and the vertical axis is sales.

Step 3: Plot the values on the graph and connect the points



The sales have trended upward over the past 12 months.

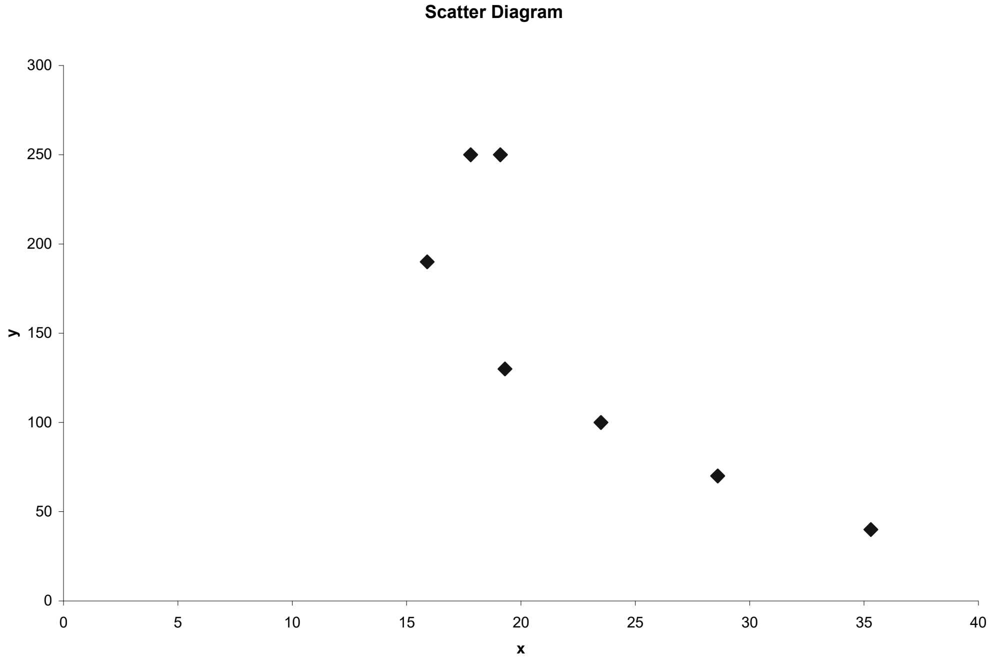
2.47. Steps 1 and 2: Identify the two variables of interest

The variables are *y* (dependent variable) and *x* (independent variable)

Step 3: Establish the scales for the vertical and horizontal axes

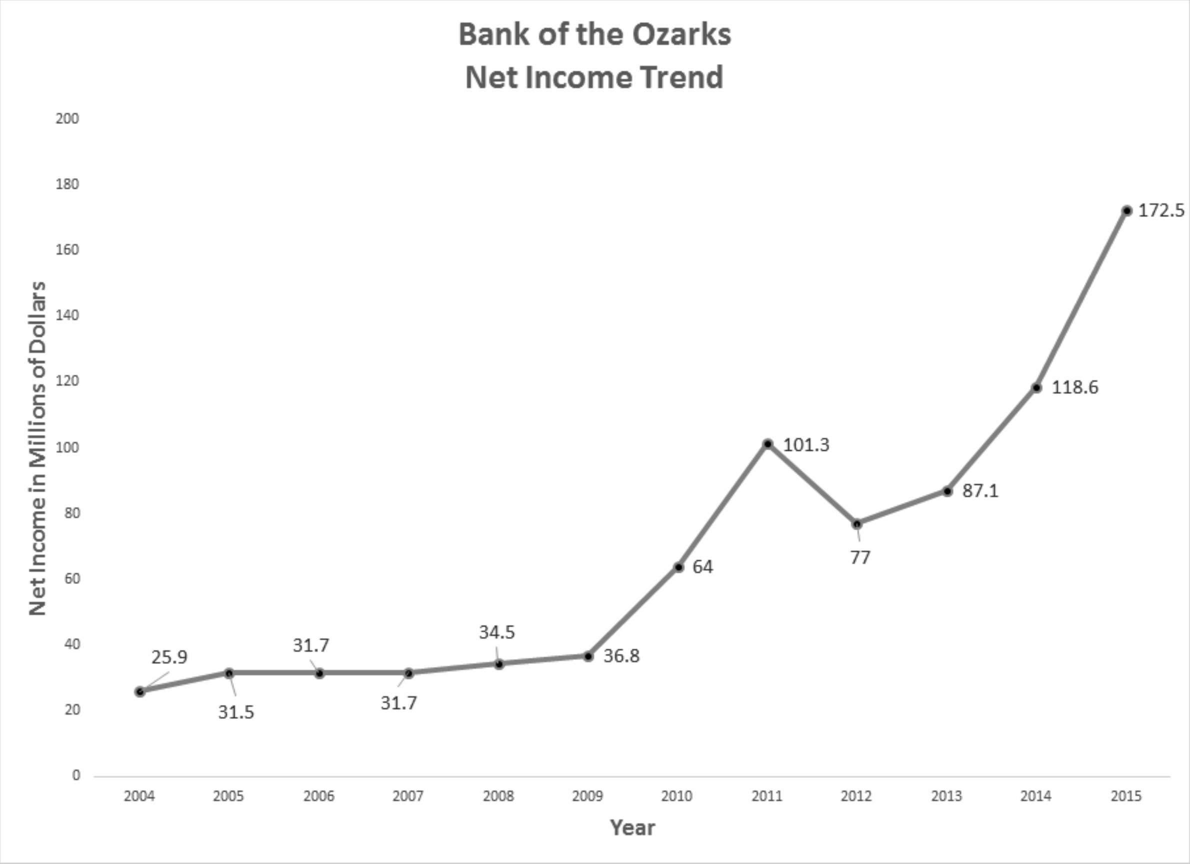
The *y* variable ranges from 40 to 250 and the *x* variable ranges from 15.9 to 35.3

Step 4: Plot the joint values for the two variables by placing a point in the *x*, *y* space shown as follows:



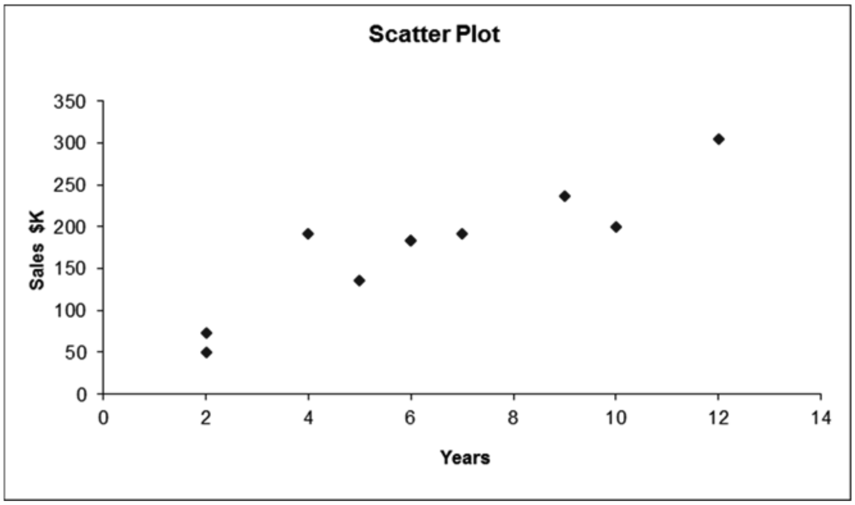
There is negative linear relationship between the two variables.

2.48. The time-series variable is Net Income ($ in millions) measured over 12 years with a maximum value of 172.5(million). The horizontal axis will have 12 time periods equally spaced. The vertical axis will start at 0 and go to a value exceeding 200. The vertical axis will also be divided into 10-unit increments. The line chart of the data is shown below.



The line chart shows that Net Income has been increasing very steadily since 2005, but have increased more sharply since 2012.

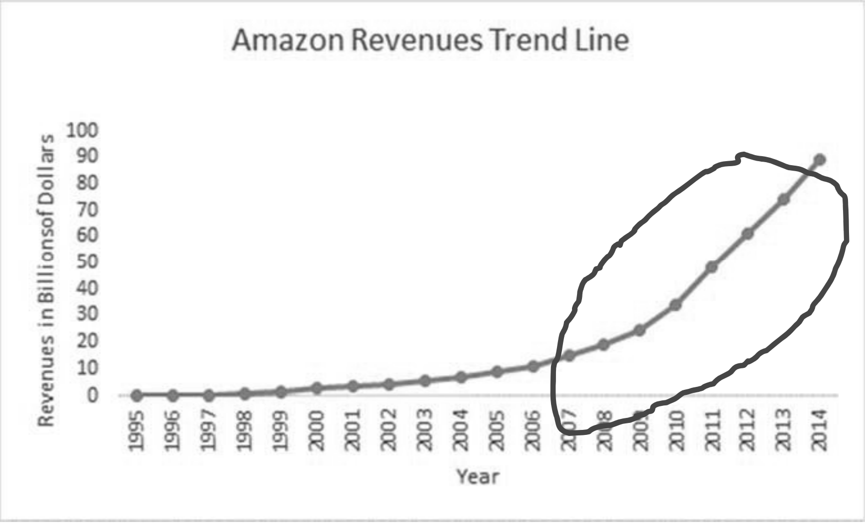
2.49. a.



b. The relationship appears to be linear and positive.

c. Note on the line plot that the Years starts at 2 years and stops at 12 years: a range of 10 years. Also the sales increase from 50 to 300 $K: a range of 250. This suggests that 250/10 = 25 $K average increase per one year increase.

2.50. a. The data supplied with the exercise begins in 2008. However, if we use sales since 1995, we get the following graph. Using only data from 2008 to 2014, the plot would include only those data that are circled.



b. Using data from 1995, the relationship appears to be curvilinear. However, the trend starting in 2014 and extending through 2014 is linear.

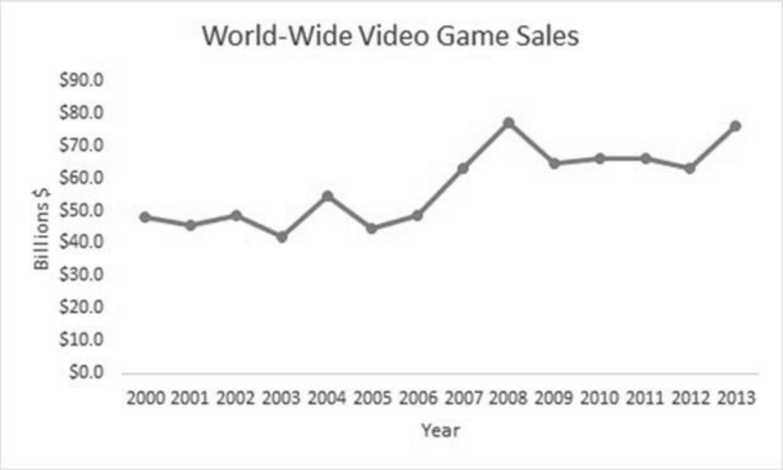
2.51. Step 1: Identify the time-series variable

The variable of interest is annual World-wide sales of video games

Step 2: Layout the Horizontal and Vertical Axis

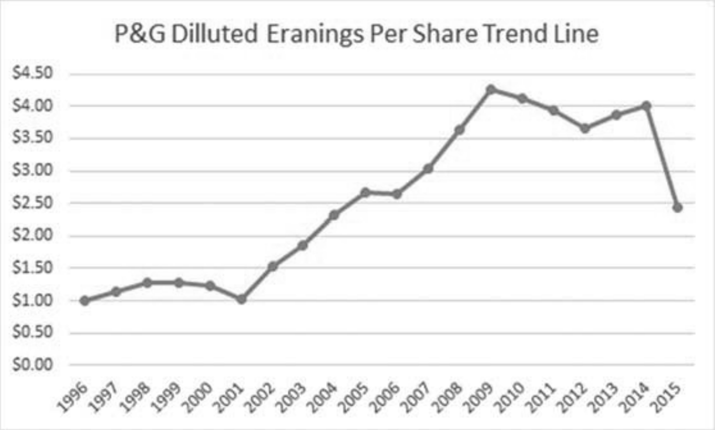
The horizontal axis will be the year and the vertical axis is sales.

Step 3: Plot the values on the graph and connect the points

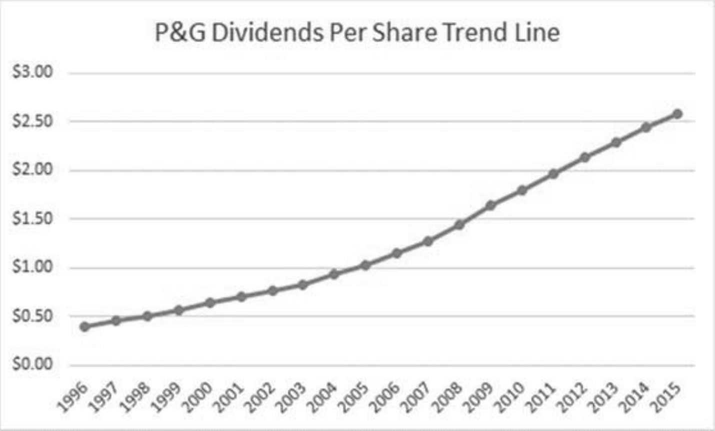


The line chart illustrates that over the 14 year period between 2000 and 2013, video game sales have grown quite steadily from just below $50 billion to over $70 billion.

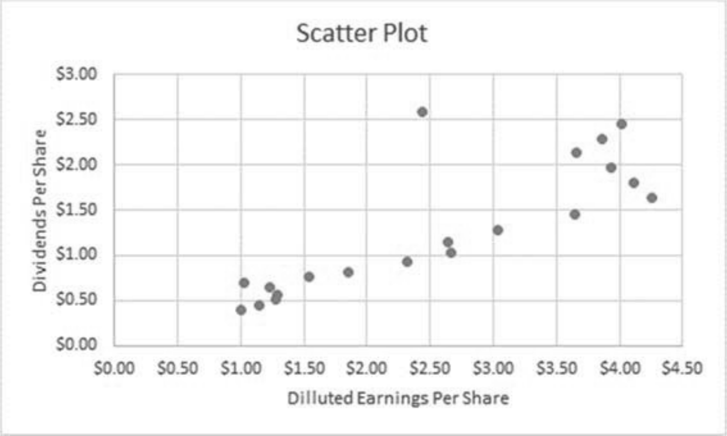
2.52. a. The time-series variable is diluted net earnings per common share measured over 20 years with a maximum value of $4.26. The horizontal axis will have 1 time periods equally spaced. The vertical axis will start at 0 and go to a value exceeding $4.26. We will use $4.50. The vertical axis will also be divided into $0.50-unit increments. The line chart of the data is shown below.



b. The time-series variable is dividends per common share measured over 20 years with a maximum value of $2.59. The horizontal axis will have 20 time periods equally spaced. The vertical axis will start at 0 and go to a value exceeding $2.59. We will use $3.00. The vertical axis will also be divided into $0.50-unit increments. The line chart of the data is shown below.

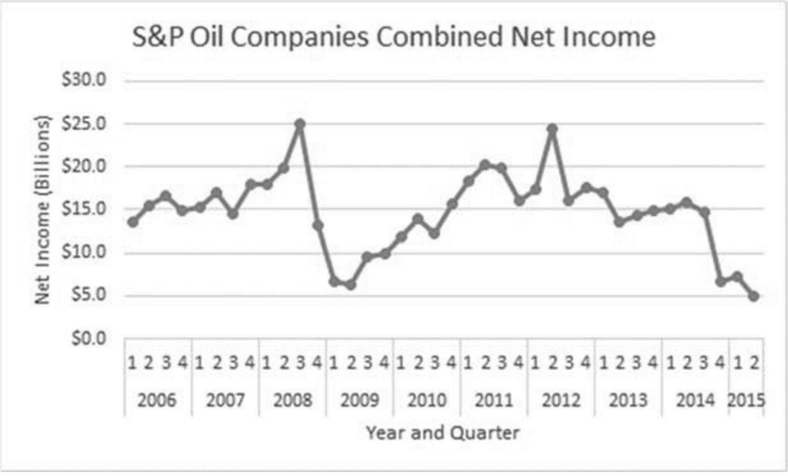


c. One variable is Diluted Net Earnings per Common Share and the other variable is Dividends per Common Share. The variable dividends per common share is the dependent (y) variable. The maximum value for each variable is \$4.26 for Diluted Net Earnings and \$2.52 for Dividends. The XY Scatter Plot is shown below.



There is a relatively strong positive relationship between the two variables, which is as one would expect. That is, one might expect to see the two variables move in the same direction.

2.53.



2.54. Steps 1: Identify the two variables of interest

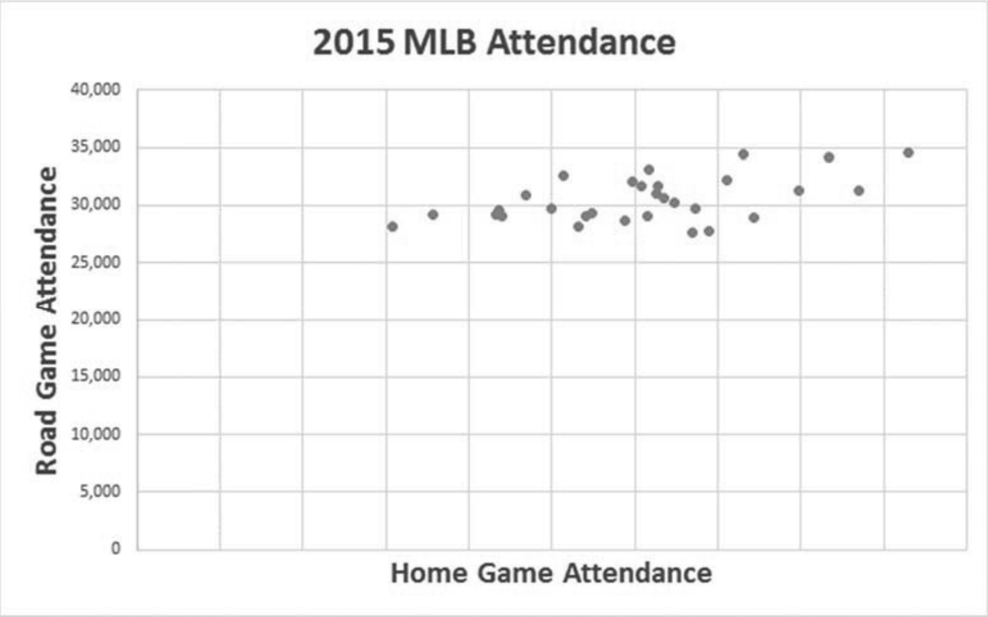
In this example, there are two variables of interest, average home attendance and average road game attendance.

Step 2: Identify the dependent and independent variables.

Either one of these variables can be selected as the dependent variable. We will select average road game attendance

Step 3: Establish the scales for the vertical and horizontal axes

Step 4: Plot the joint values for the two variables by placing a point in the x, y space shown as follows:



Based on the scatter diagram, it appears that there is a slight positive linear relationship between home and road attendance. However, the relationship is not perfect.

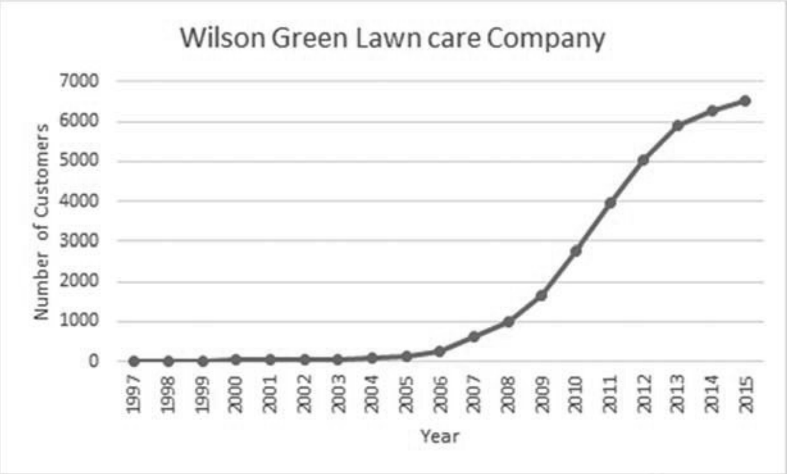
2.55. Step 1: Identify the time-series variable

The variable of interest is number of customers

Step 2: Layout the Horizontal and Vertical Axis

The horizontal axis will be the year and the vertical axis is the number of customers

Step 3: Plot the values on the graph and connect the points



Since 1995, there has been a very steep growth in the number of customers over the time span.

2.56. Step 1: Identify the two variables of interest

In this situation, there are two variables, fuel consumption per hour, the dependent variable, and passenger capacity, the independent variable.

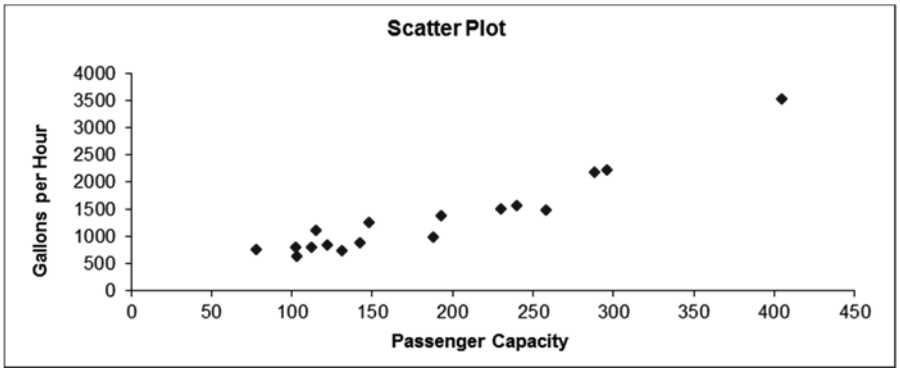
Step 2: Identify the dependent and independent variables.

The analyst is attempting to predict passenger capacity using fuel consumption per hour. Therefore, the capacity is the dependent variable and the fuel consumption per hour is the independent variable.

Step 3: Establish the scales for the vertical and horizontal axes

The *y* variable (fuel consumption) ranges from 631 to 3,529 and the *x* variable (passenger capacity) ranges from 78 to 405.

Step 4: Plot the joint values for the two variables by placing a point in the *x*, *y* space shown as follows:



Based on the scatter diagram we see there is a strong positive linear relationship between passenger capacity and fuel consumption per hour.

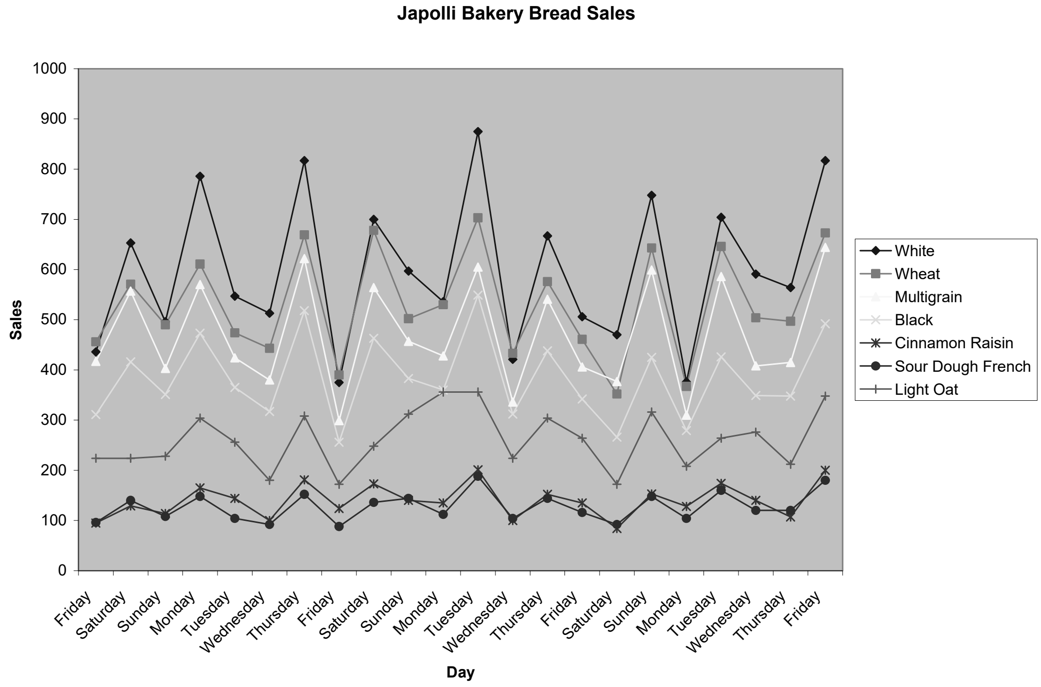
2.57. Step 1: Identify the time-series variable

In this case, there are seven variables of interest. These are the daily sales for each of the bread types

Step 2: Layout the Horizontal and Vertical Axis

The horizontal axis will be the day and the vertical axis is the number of loaves of bread that were sold.

Step 3: Plot the values on the graph and connect the points.



The graph illustrates a general pattern in the bread sales. Higher sales tend to occur for all types of bread on Saturdays, Mondays and Thursdays with Fridays typically the lowest.

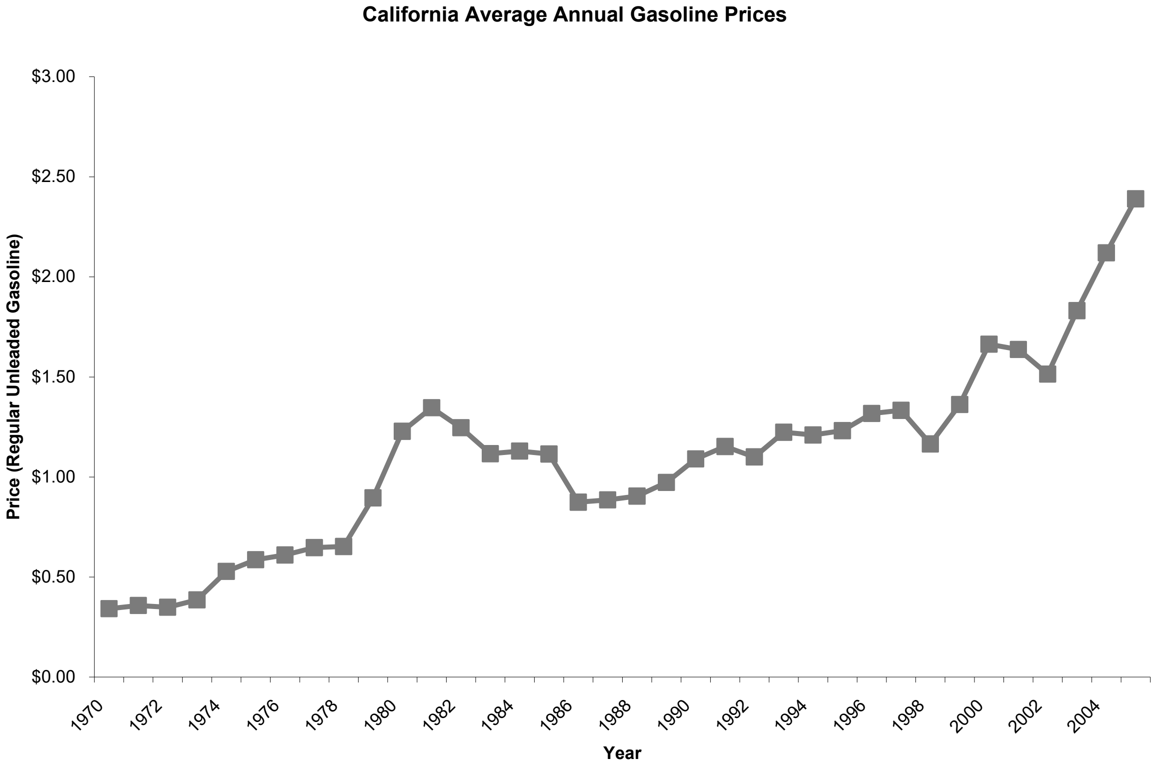
2.58. a. Step 1: Identify the time-series variable

The variable of interest is annual average price of gasoline in California

Step 2: Layout the Horizontal and Vertical Axis

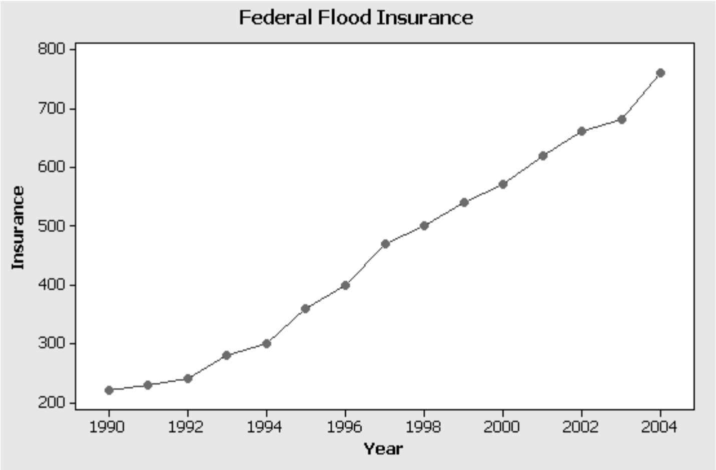
The horizontal axis will be the year and the vertical axis is average price (See Step 3)

Step 3: Plot the values on the graph and connect the points



Gasoline prices have trended upward over the 36 year period with some short periods of decline. However, prices rises have been very steep since 1999.

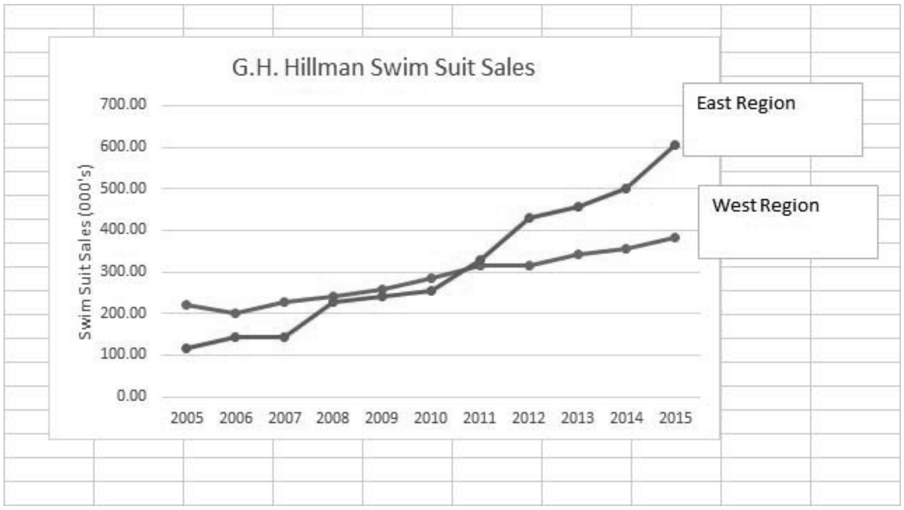
2.59. a.



b. The relationship appears to be linear and positive.

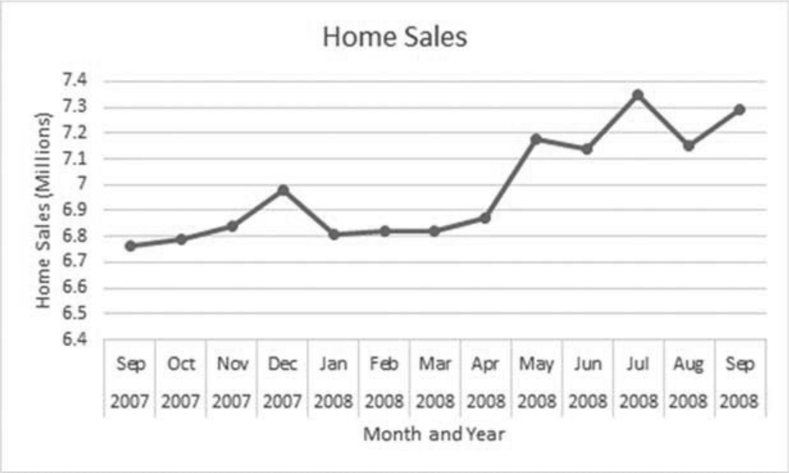
c. The average equals the sum divided by the number of data points $= 6830/15 = 455.33$.

2.60. a.

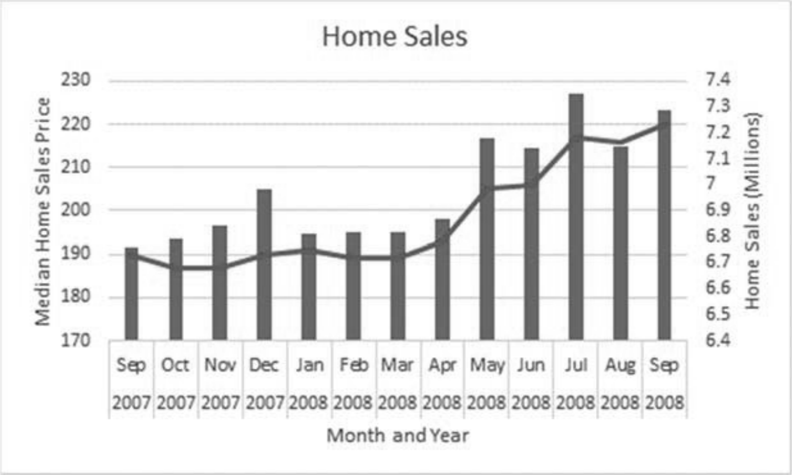


b. Both relationships seem to be linear in nature but the East region is growing faster in more recent years.

2.61. a.



b.



Note, a combo chart in Excel was used to show both data series since they have very different values.

c. It appears from the line plots that the monthly sales have been fluctuating greatly during this period, dipping in January, heading back up during the summer months and then declining again. Median sales price has shown a steady minor decline during the period.

End of Chapter Exercises

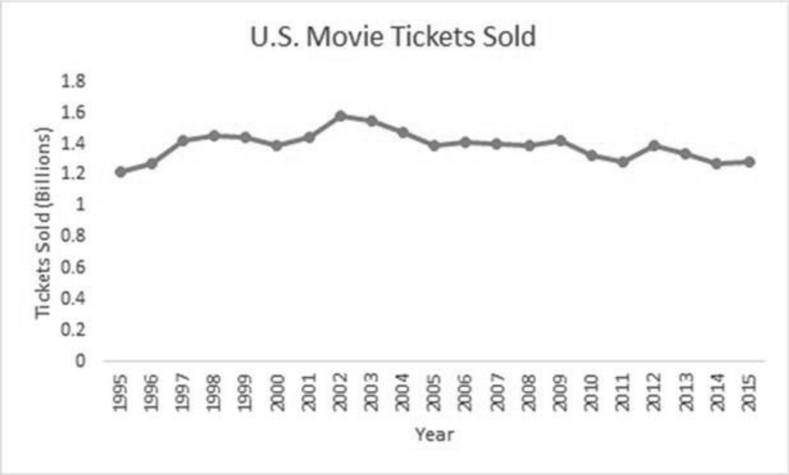
2.62. A relative frequency distribution deals with the percentage of the total observations that fall into each class rather than the number that fall into each class. Sometimes decision makers are more interested in percentages than numbers. Politicians, for instance, are often more interested in the percentage of voters that will vote for them (more than 50%) than the total number of votes they will get. Relative frequencies are also valuable when comparing distributions from two populations that have different total numbers.

2.63. Thinking in terms of the types of data discussed in chapter 1, that is nominal, ordinal, interval and ratio, bar charts are visual representations of frequency distributions constructed from nominal or ordinal data.

2.64. Pie charts are effectively used when the data set is made up of parts of a whole, and therefore each part can be converted to a percentage. For instance, if the data involves a budget, a pie chart can represent the percentage of budget each category represents. Or, if the data involves total company sales, a pie chart can be used to represent the percentage contribution to sales for each major product line.

2.65. A line chart is an effective tool to represent the relation between a dependent and an independent variable when values of the independent variable form a natural increasing sequence. In many cases this means the independent variable is a measure of time and the data is time-series data. With a scatter plot the values of the independent variable are not determined according to a preset sequence.

2.66.



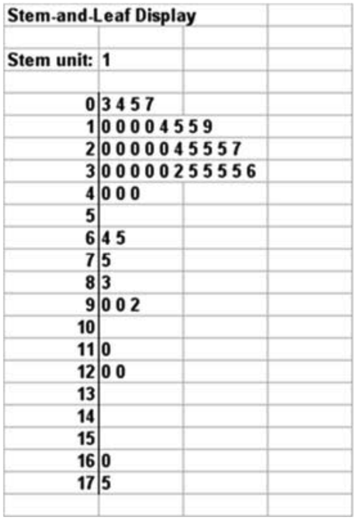
2.67. a. Using the $2^k \geq n$ guideline:

$$2^k \geq 48 = 2^6 \geq 48$$

To determine the class width, (17.5 – 0.3)/6 = 2.87 so round up to 3 to make it easier.

Classes	Frequency
0.1 to 3	27
3.1 to 6.0	9
6.1 to 9.0	6
9.1 to 12	4
12.1 to 15.0	0
15.1 to 18.0	2

b.

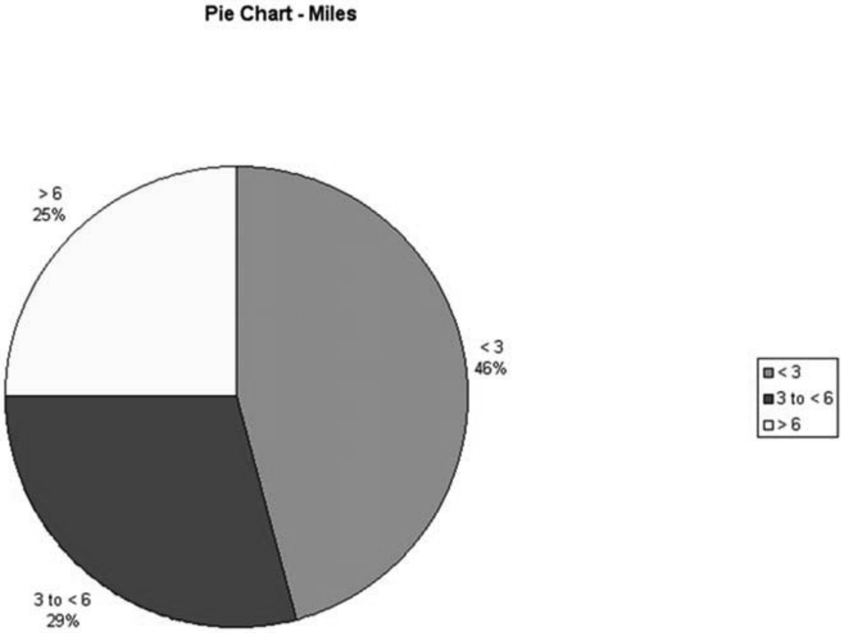


Stem-and-Leaf Display

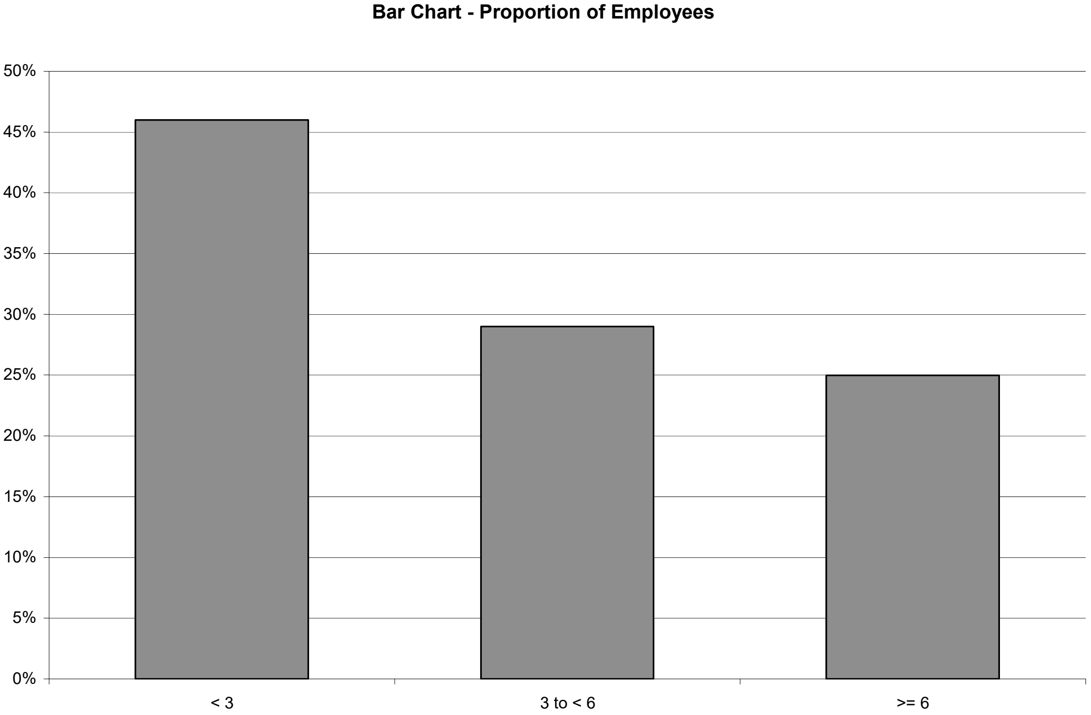
Stem unit: 1

0	3 4 5 7
1	0 0 0 0 4 5 5 9
2	0 0 0 0 0 4 5 5 5 7
3	0 0 0 0 0 2 5 5 5 5 6
4	0 0 0
5	
6	4 5
7	5
8	3
9	0 0 2
10	
11	0
12	0 0
13	
14	
15	
16	0
17	5

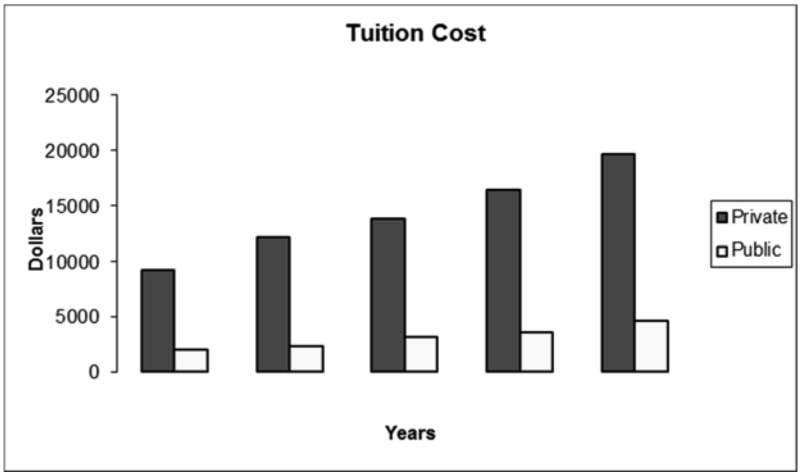
c.


Pie Chart - Miles
> 6
25%
< 3
46%
3 to < 6
29%
▨ < 3
■ 3 to < 6
□ > 6

d.

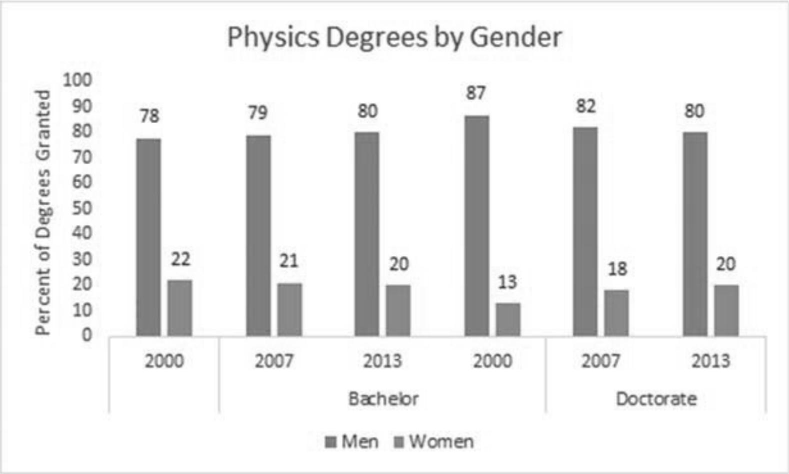

Bar Chart - Proportion of Employees
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
< 3
3 to < 6
>= 6

2.68. a.



b. Student answers will vary but should include identifying that both private and public college tuition costs have more than doubled in the 20 years of data.

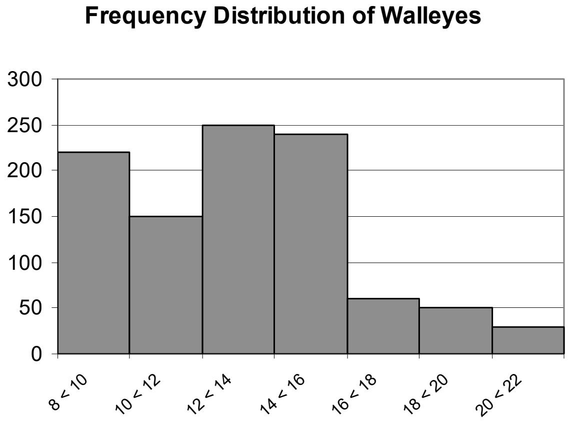
2.69. a.



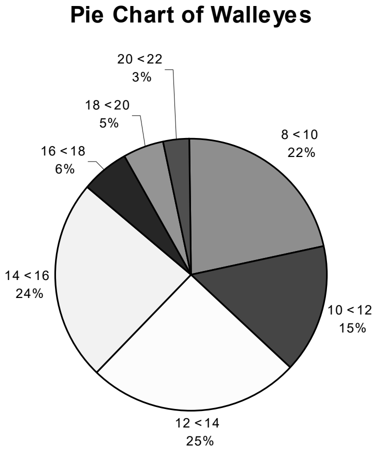
b. There has been a slight decline in the percentage of physics Bachelor degrees granted to women and an increase in percentage of Doctorate degrees granted to women.

2.70. a. The frequencies can be calculated by multiplying the relative frequency times the sample size of 1,000.

Class Length (Inches)	Frequency	Relative Frequency
8 < 10	220	0.22
10 < 12	150	0.15
12 < 14	250	0.25
14 < 16	240	0.24
16 < 18	60	0.06
18 < 20	50	0.05
20 < 22	30	0.03



b. The histogram is probably a better representation of the fish length data.

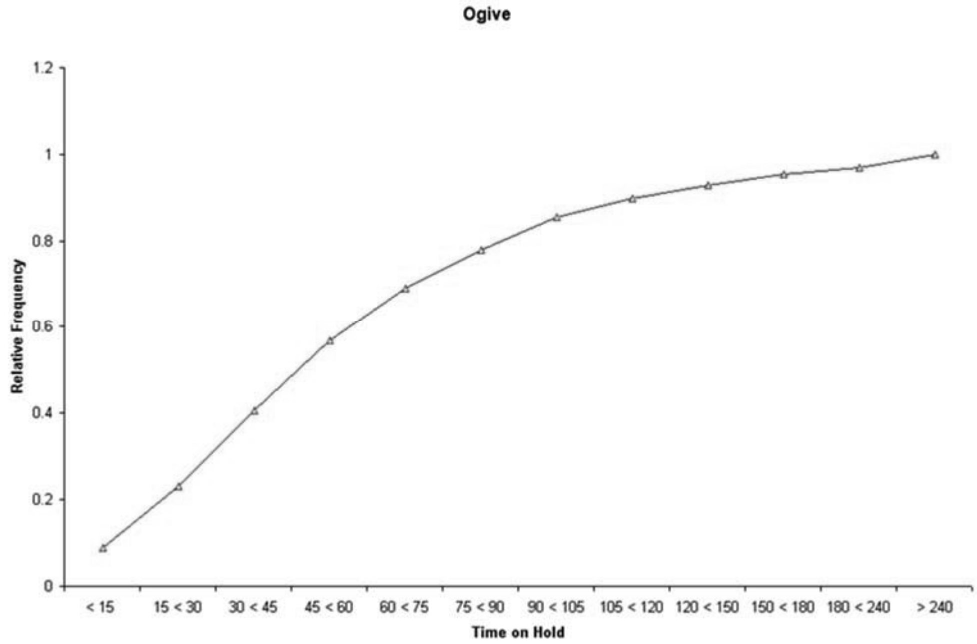


2.71. a. Based upon the following table the percent of class that hold at least 120 seconds (2 minutes) is

$$0.0311 + 0.0244 + 0.0171 + 0.0301 = 0.1029$$

Classes (in seconds)	Number	Relative Frequency
< 15	456	0.0899
15 < 30	718	0.1415
30 < 45	891	0.1756
45 < 60	823	0.1622
60 < 75	610	0.1202
75 < 90	449	0.0885
90 < 105	385	0.0759
105 < 120	221	0.0435
120 < 150	158	0.0311
150 < 180	124	0.0244
180 < 240	87	0.0171
≥ 240	153	0.0301

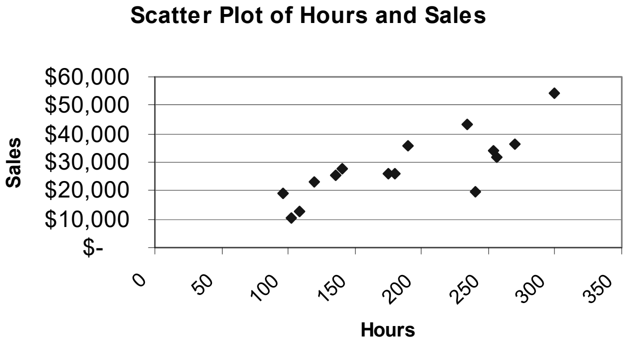
Note: For this problem the class widths are not equal.



b. The number of people who have to wait 120 seconds (2 minutes) or more is

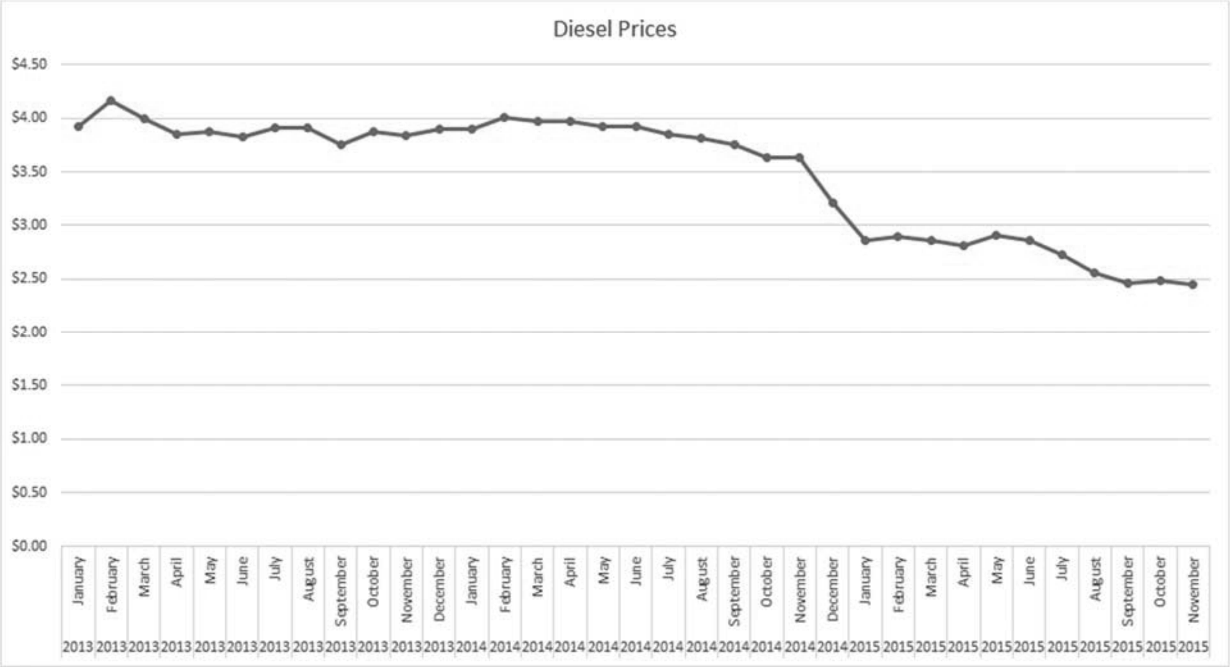
$$158 + 124 + 87 + 153 = 522 * \$30 = \$15,660 \text{ month.}$$

2.72. a. The independent variable is hours and the dependent variable is sales



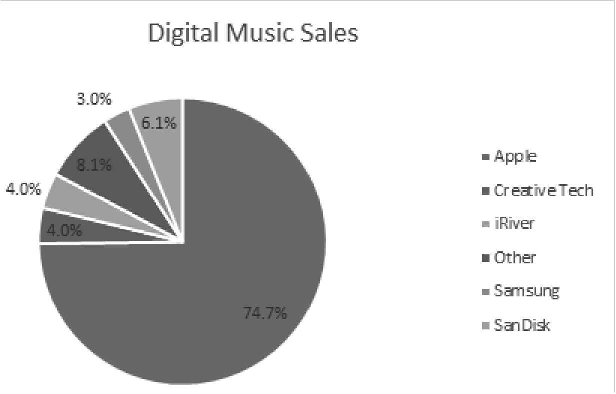
b. It appears that there is a positive linear relationship between hours worked and weekly sales. It appears that the more hours worked the greater the sales. No stores seem to be substantially different in terms of the general relationship between hours and sales.

2.73. a.



Student reports will vary but should discuss the negative trend in diesel prices over the time period given.

2.74.



2.75. a. Using the $2^k \ge n$

$$2^k \ge 100 \text{ so } 2^7 = 128$$

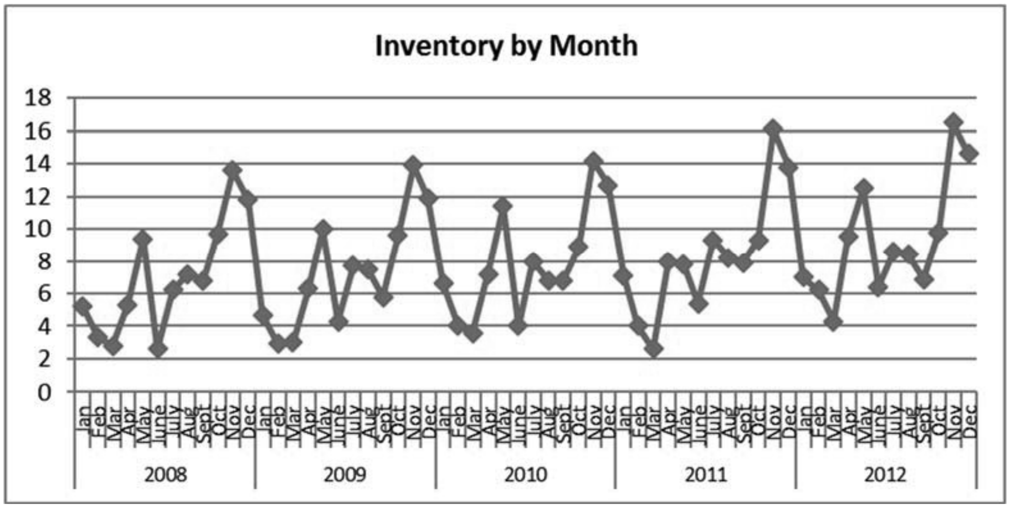
Determine the width: (310494 – 70464)/7 = 34,290. Round to 35,000

Classes	Frequency
70,000–104,999	43
105,000–139,999	34
140,000–174,999	13
175,000–209,999	5
210,000–244,999	2
245,000–274,999	1
280,000–314,999	2

b.

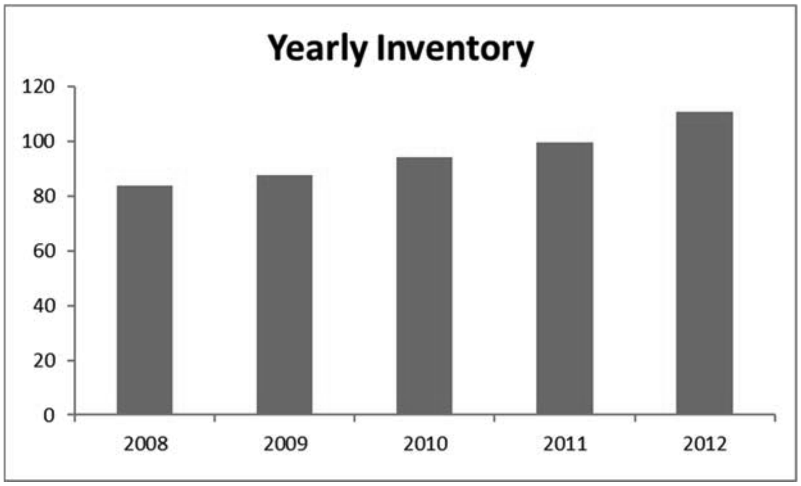
Classes	Frequency	Relative Frequency	Cumulative Relative Frequency
70,000–104,999	43	0.43	0.43
105,000–139,999	34	0.34	0.77
140,000–174,999	13	0.13	0.90
175,000–209,999	5	0.05	0.95
210,000–244,999	2	0.02	0.97
245,000–274,999	1	0.01	0.98
280,000–314,999	2	0.02	1.00

2.76. a.



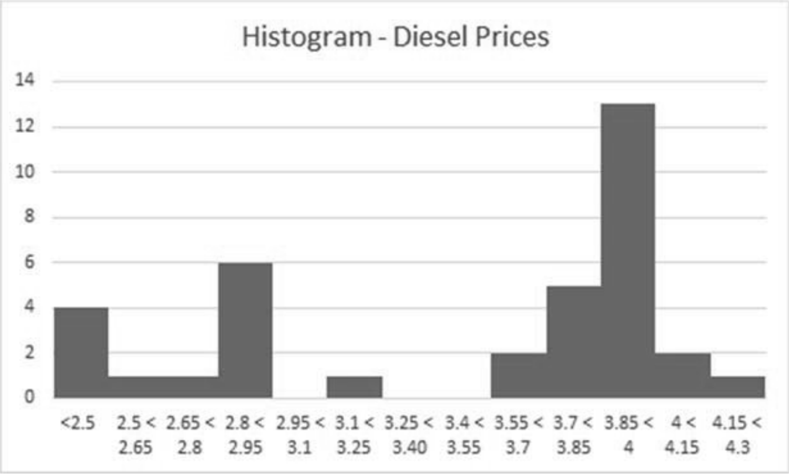
Inventory has been trending slightly up over the five years, but appears to be highly seasonal with predictable highs at certain points each year.

b.



This bar chart is effective for showing the growth in total annual inventory over the five years. However, students should keep in mind that the sum of monthly inventory does not equate to how much inventory the store had on hand at the end of the year. Students might question why the store would graph the total inventory

2.77. a.



b. Notice that three of the class interval have no observations. Since the numbers are averages taken across the United States, it is possible that the sampling technique, simply from randomness, didn't select prices in those ranges. It bears further investigation.

Chapter 3: Describing Data Using Numerical Measures

Section 3.1

3.1. The sample mean is computed using the following steps:

Step 1: Collect the sample data.

Step 2: Add the values in the sample:

$$\sum x = 74349$$

Step 3: Divide the sum by the sample size.

$$\bar{x} = \frac{74349}{15} = 4956.60$$

The quartiles are found using the following steps.

Step 1: Sort the data from low to high.

4132	4188	4209	4423	4568
4573	4983	5002	5052	5176
5310	5381	5611	5736	6005

Step 2: Determine the percentile location index.

To determine the location index for the 1st quartile (p = 25) we do the following:

i = 25/100(n) = (25/100) * 15 = 3.75.

Step 3: Find the percentile.

Because the index is not an integer it is rounded up to 4. The first quartile is the fourth value in the sorted array and is equal to 4423.

The median (or second quartile) is found by sorting the data from lowest to highest (see above). The index point, i, for the median is found by i = 1/2(n) = 1/2(15) = 7.5. Because i is not an integer it is rounded up to 8. The median is located by counting 8 values into the sorted data. The median is 5002.

The location index for the 3rd quartile is:

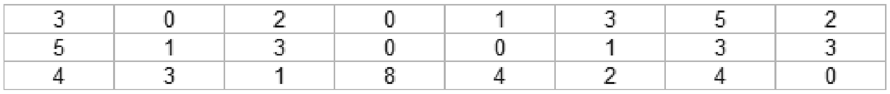
75/100(n) = 75/100(15) = 11.25.

Because i is not an integer it is rounded up to 12. The third quartile is the 12th value in the sorted array and is equal to 5381.

3.2. a. The sample mean is computed using the following steps:

Step 1: Collect the sample data.

The sample data are:



3	0	2	0	1	3	5	2
5	1	3	0	0	1	3	3
4	3	1	8	4	2	4	0

Step 2: Add the values in the population:

$$\sum x = 3 + 0 + 2 + ... + 4 + 0 = 58$$

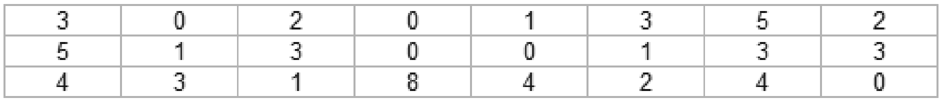
Step 3: Divide the sum by the population size.

$$\mu = \frac{\sum x}{N} = \frac{58}{24} = 2.42$$

The mean number of defects for this sample of 24 employees is 2.42.

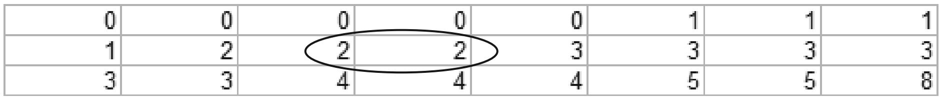
b. The median is computed using the following steps:

Step 1: Collect the population data.



3	0	2	0	1	3	5	2
5	1	3	0	0	1	3	3
4	3	1	8	4	2	4	0

Step 2: Sort the data from smallest to largest, forming a data array.



0	0	0	0	0	1	1	1
1	2	2	2	3	3	3	3
3	3	4	4	4	5	5	8

Step 3: Find the median index.

The location index for the median is determined using the following equation:

$$i = \frac{1}{2}(n)$$

For $n = 24$, the index is:

$$i = \frac{1}{2}(24) = 12$$

Step 4: Find the median

Because the index is 12 which is an integer, the median is the average of the 12th and 13th data values going from either end. These two values are 2 and 3. This the median is:

$$M_d = \frac{2+3}{2} = 2.5$$

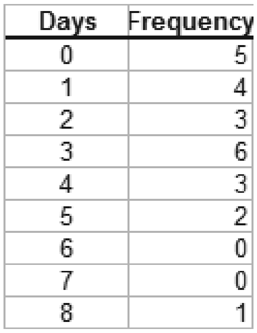
Half the data values fall below 2.5 and half the data values fall above 2.5.

c. To determine if there is a mode and what the value of the mode is, we use the following steps:

Step 1: Collect the population data.

See parts a. or b.

Step 2: Organize the data into a frequency distribution.



Days	Frequency
0	5
1	4
2	3
3	6
4	3
5	2
6	0
7	0
8	1

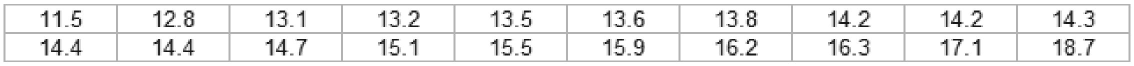
Mode = 3

3 occurs 6 times

Step 3: Determine the value that occurs most frequently.

The value 3 occurs 6 times which is the most of any value in the sample.

3.3. Step 1: Sort the data from low to high.



11.5	12.8	13.1	13.2	13.5	13.6	13.8	14.2	14.2	14.3
14.4	14.4	14.7	15.1	15.5	15.9	16.2	16.3	17.1	18.7

Step 2: Determine the quartile location index and find the quartile value.

To determine the location index for the 1st quartile ($p = 25$) we do the following:

$$i = \frac{p}{100}(n) = \frac{25}{100}(20) = 5$$

Since the index, 5, is an integer, the 1st quartile is determined by finding the average of the 5th and 6th values from the lower end of the sorted data. This is:

$$Q1 = \frac{13.5 + 13.6}{2} = 13.55$$

The location index for the 3rd quartile is:

$$i = \frac{p}{100}(n) = \frac{75}{100}(20) = 15$$

Since the index, 15, is an integer, the 3rd quartile is determined by finding the average of the 15th and 16th values from the lower end of the sorted data. This is:

$$Q3 = \frac{15.5 + 15.9}{2} = 15.7$$

Thus, the 1st quartile value is 13.55 meaning that 25 percent of the data values fall below 13.55. The third quartile is 15.7 meaning that 75% of the data values fall below 15.7.

3.4. The mean is found by adding the data values together and dividing by the number of values in the sample: 1873/16 = 117.0625.
The sorted array is shown below

51	72	78	100
101	106	116	125
128	129	130	135
139	141	153	169

The position for the first quartile, Q_1, the median, and the third quartile, Q_3 are calculated below:
Q_1 position = 25/100(16) = 4.
Q_1 is the average of the 4th and 5th values in the sorted array. Q_1 = (100 + 101)/2 = 100.5
Q_2 position = median = 50/100(16) = 8. The median is the average of the 8th and 9th values in the sorted array. Median = (125 + 128)/2 = 126.5
Q_3 position = 75/100(16) = 12.
Q_3 is the average of the 12th and 13th values in the sorted array = (135 + 139)/2 = 137.

3.5. Step 1: Sort the data from low to high.

The sorted data are shown below:

35	50	50	50
60	75	75	75
80	85	85	90
90	100	100	100
100	125	125	150

Step 2: Calculate the 25th percentile (Q_1), the 50th percentile (median), and the 75th percentile (Q_3).

The 25th percentile location is $(25/100)*20 = 5$. So Q_1 is the average of the values in the 5th and 6th position of the sorted array. Q_1 = (60 + 75)/2 = 67.5.

The median location is $(50/100)*20 = (1/2)*20 = 10$. Because the location point is an integer the median is the average of the values in the 10th and 11th location. Median = (85 + 85)/2 = 85.

The 75th percentile location is $(75/100)*20 = 15$. So Q3 is the average of the values in the 15th and 16th position of the sorted array. Q_3 = (100 + 100)/2 = 100.

Step 3: Draw the box so the ends correspond to Q_1 and Q_3.

Step 4: Draw a vertical line through the box at the median.

Step 5: Compute the upper and lower limits:

$$\text{Lower limit} = Q_1 - 1.5(Q_3 - Q_1) = 67.5 - 1.5*32.5 = 18.75$$

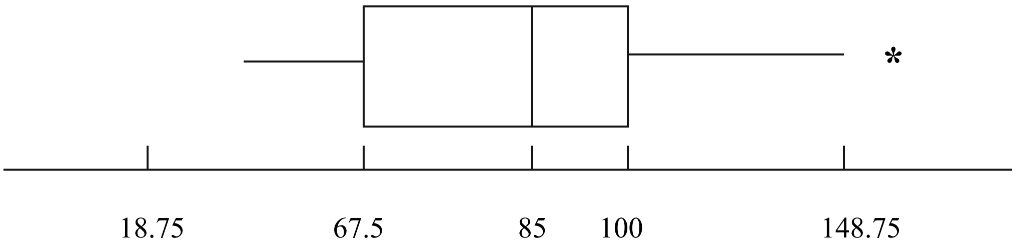
$$\text{Upper Limit} = Q_3 + 1.5(Q_3 - Q_1) = 100 + 1.5*32.5 = 148.75$$

Any value outside these limits will be labeled an outlier.

Step 6: Draw the whiskers.

Step 7: Plot the outliers. Outliers are typically indicated by an asterisk, *.

The box and whisker plot is shown below.



3.6. a. Use the following to locate the 1st and 3rd quartiles

$$i = \frac{p}{100}(n)$$

For Q_1 , $p = 25$ Then $i = \frac{25}{100}(24) = 6$.

So Q_1 is the average of the values in the 6th and 7th position of the sorted array.

$$Q_1 = (20 + 23)/2 = 21.5$$

For Q_2, $p = 50$ Then $i = 50/100(24) = 12$

So Q_2 is the average of the values in the 12th and 13th position of the sorted array.

$$Q_2 = (27 + 28)/2 = 27.5.$$

For Q_3, $p = 75$ Then $i = \frac{75}{100}(24) = 18$

So Q_3 is the average of the values in the 18th and 19th position of the sorted array.

$$Q_1 = (44 + 45)/2 = 44.5.$$

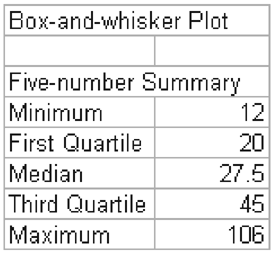
b. The 90th percentile is found using the following:

$$i = \frac{p}{100}(n)$$

When $p = 90$, we get: $i = \frac{90}{100}(24) = 21.60$

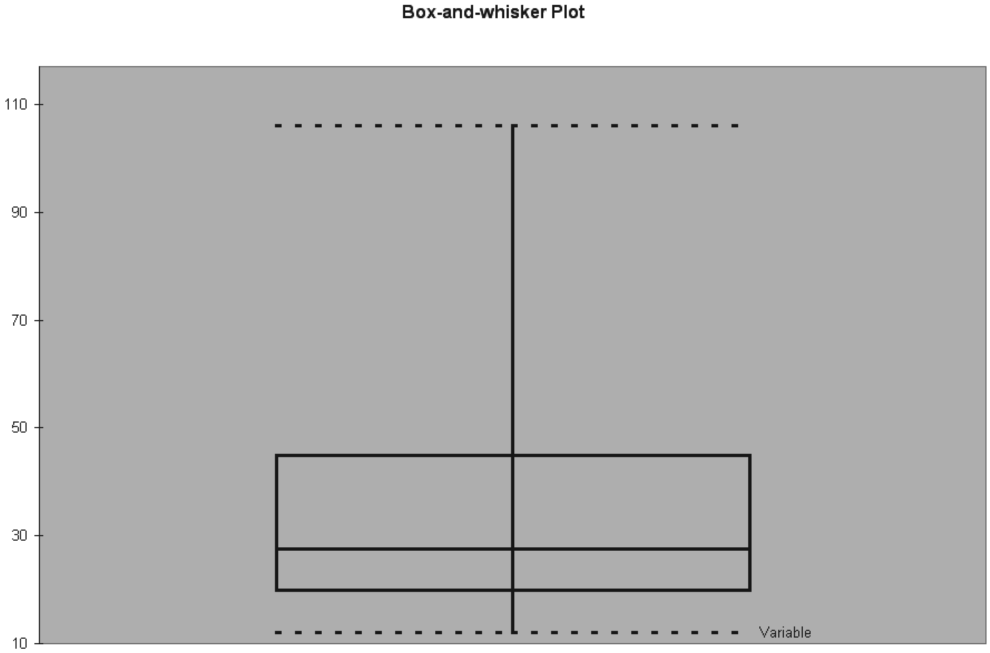
Since this is not an integer, it is rounded up to 22. This means that the 90 percentile is the 22nd (70) value in the array.

c. Using the PHStat add in to Excel:



Box-and-whisker Plot	
Five-number Summary	
Minimum	12
First Quartile	20
Median	27.5
Third Quartile	45
Maximum	106

Note that Excel calculates the quartiles in a slightly different manner than shown in the text.



d. The 20th percentile is found using $i = \frac{20}{100}(24) = 4.8$ Rounding up, this means that the 20th percentile is the 5th value from the top once the data have been arranged in numerical order. This is 19.

The 30th percentile is found using: $i = \frac{30}{100}(24) = 7.2$ Again, rounding up 30th percentile is the 8th value. This is 23.

3.7. a. The index is $(p/100)n$ = (80/100)20 = 16 = integer. Therefore, an 80th percentile is obtained by calculating the average of the 16th and 17th data value = (31.2 + 32.2)/2 = 31.7.

b. The 25th percentile: The index is $(p/100)n$ = (25/100)20 = 5 = integer. Therefore, the 25th percentile is obtained by calculating the average of the 5th and 6th data value = (12.1 + 13)/2 = 12.55. The 75th percentile: The index is $(p/100)n$ = (75/100)20 = 15 = integer. Therefore, the 75th percentile is obtained by calculating the average of the 15th and 16th data value = (26.7 + 31.2)/2 = 28.95.

c. The median is the 50th percentile. The index is $(p/100)n = (50/100)20 = 10$ = integer. Therefore, a 50th percentile is obtained by calculating the average of the 10th and 11th data value = (20.8 + 22.8)/2 = 21.8.

3.8. a.

Instrument	Final	Project	Midterm 1	Midterm 2	Homework
Weight	40	10	20	20	10

b. $\bar{x}_w = \dfrac{\sum w_i x_i}{\sum w_i} = \dfrac{40(64)+10(98)+20(67)+20(63)+10(89)}{40+10+20+20+10} = 70.30$

c. $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{(64)+(98)+(67)+(63)+(89)}{5} = 76.2$. The (unweighted) average is actually a weighted average whose weights are all equal. In this case, that would indicate that the Final would have the same influence as one of the midterms, homework, and the project. Usually a (comprehensive) final is given more weight because it evaluates the material that should be retained.

3.9. a. $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n} = 456/24 = 19$

To compute the median, rank the observations and find the average of the middle two values.

10 12 14 14 17 17 18 18 19 19 19 19
19 20 20 21 21 21 21 22 22 23 25 25

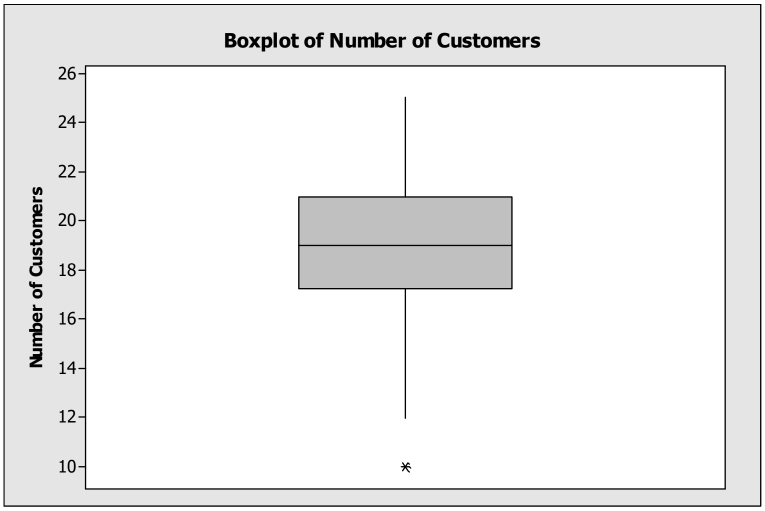
Median = (19 + 19)/2 = 19

Mode = 19

b. This data is symmetrical since the mean = median = mode

Box-and-whisker Plot

Five-number	Summary
Minimum	10
First Quartile	17.5
Median	19
Third Quartile	21
Maximum	25

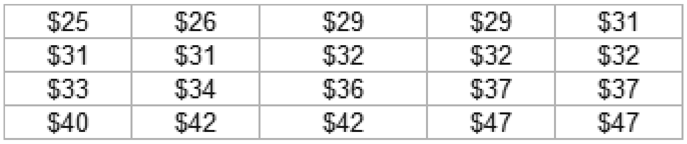


The box plot does support the idea that the distributions are symmetric although the median is not directly in the center between the Q_1 and Q_3.

3.10. a. The sample mean is:

$$\bar{x} = \frac{\sum x}{n} = \frac{693}{20} = \$34.65$$

b. The median is found by first sorting the data into order from low to high.



\$25	\$26	\$29	\$29	\$31
\$31	\$31	\$32	\$32	\$32
\$33	\$34	\$36	\$37	\$37
\$40	\$42	\$42	\$47	\$47

The position index for the median is:

$$i = \frac{1}{2}(n) = \frac{1}{2}(20) = 10$$

Since the index, 10, is an integer, the median is the average of the 10th and 11th values in the data set. Thus:

$$M_d = \frac{32+33}{2} = 32.50$$

Thus, the median value is \$32.50

c. The mode is the value that occurs most frequently. In this case there are two values that both occur three times; \$31.00 and \$32.00. Thus, there are two modes. Note, if students use Excel, only one mode will be reported; \$31.00.

d. Student paragraphs will vary but should indicate that the tip distribution is positively skewed. Also, students may point out that size of the tip income (mean \$34.65) should be viewed relative to the hours worked.

3.11. a. The weighted mean is computed using:

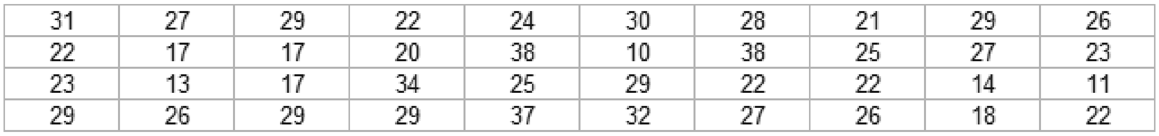
$$\bar{x}_W = \frac{\sum w_i x_i}{\sum w_i} = \frac{(7,400)(123)+(14,400)(402)+(12,300)(256)+(6,200)(109)+(3,100)(67)}{123+402+256+109+67} = 11,213.48$$

b. It is reasonable that plants with more employees would have more medical issues. The weighted average takes into account the number of employees and is a more reasonable measure of the average payments than would be an unweighted average that treats all plants as equals.

3.12. The sample mean is computed using the following steps:

Step 1: Collect the sample data.

The sample data are:



31	27	29	22	24	30	28	21	29	26
22	17	17	20	38	10	38	25	27	23
23	13	17	34	25	29	22	22	14	11
29	26	29	29	37	32	27	26	18	22

Step 2: Add the values in the sample:

$$\sum x = 31+27+29+\cdots+18+22 = 989$$

Step 3: Divide the sum by the sample size.

$$\bar{x} = \frac{989}{40} = 24.73$$

The mean number of minutes for service calls for this sample of 40 calls is 24.73 minutes.

The median is computed using the following steps:

Step 1: Collect the sample data.

See part a.

Step 2: Sort the data from smallest to largest, forming a data array.

10	11	13	14	17	17	17	18	20	21
22	22	22	22	22	23	23	24	25	25
26	26	26	27	27	27	28	29	29	29
29	29	29	30	31	32	34	37	38	38

Step 3: Find the median index.

The location index for the median is determined using the following equation:

$$i = \frac{1}{2}(n)$$

For $n = 40$, the index is:

$$i = \frac{1}{2}(40) = 20$$

Step 4: Find the median.

Because the index is 20 which is an integer, the median is the average of the 20th and 21st data values going from either end. These two values are 25 and 26. This the median is:

$$M_d = \frac{25 + 26}{2} = 25.5$$

Half the data values fall below 25.5 and half the data values fall above 25.5.

To determine if there is a mode and what the value of the mode is, we use the following steps:

Step 1: Collect the sample data.

See part a.

Step 2: Organize the data into a frequency distribution.

Minutes	Frequency
10	1
11	1
12	0
13	1
14	1
15	0
16	0
17	3
18	1
19	0
20	1
21	1
22	5
23	2

24	1
25	2
26	3
27	3
28	1
29	6
30	1
31	1
32	1
33	0
34	1
35	0
36	0
37	1
38	2
39	0
40	0

Step 3: Determine the value that occurs most frequently.

The value 29 occurs 6 times which is the most of any value in the sample. Thus the mode for these sample data is 29 minutes.

In order for data to be perfectly symmetric, the mean and median must be equal. In this case, the mean is 24.73 while the median is 25.5. Thus the data are slightly left skewed.

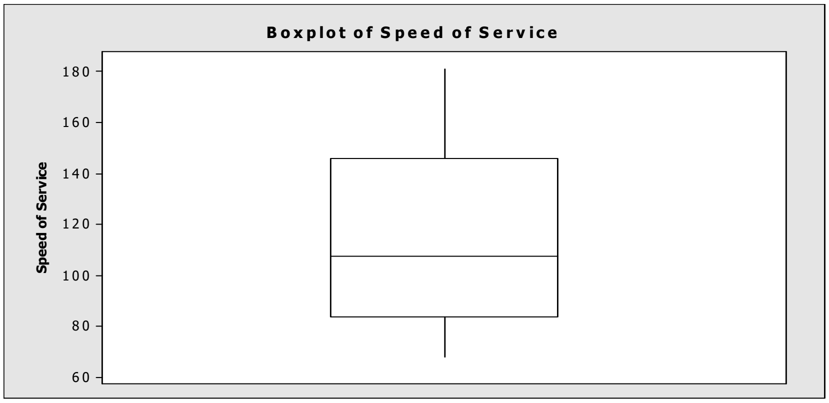
3.13. a. The sorted data are shown below.

68	85	110	146
71	88	116	147
73	90	119	156
74	92	130	156
76	93	134	162
79	103	138	178
83	105	145	181

The mean is 114.21 and the median is 107.50. Note the position of the median is (1/2)(*n*) where *n* = 28. Because the median's position is 14, the average of the 14th and 15th position in the sorted array is the median, which is (105 + 110)/2 = 107.50. The mode is the most frequently occurring value and is 156.

b. Because the mean is larger than the median the data are skewed right.

c. The box and whisker plot is shown below.



Note that the median is not in the center of the box and that the whiskers are not of equal length. The whisker going up is longer than the whisker going down which indicates that the data are skewed to the right. This supports the conclusion that the data are not symmetric, but skewed.

3.14. a. The data values are the price per square foot and the weights are the number of square feet. To compute the weighted average price per square foot the price per square foot is multiplied by the corresponding number of square feet and the products are summed.

$$(\$75)(125000)+(\$85)(37500)+\ldots+(\$110)(130000)=38,712,500$$

The weights (number of square feet) are summed:

$$125000+37500+\ldots+130000=465,000$$

The weighted mean is computed by dividing the weighted sum by the sum of the weights:

$$38,712,500\,/\,465,000=83.2527=\$83.25$$

b. The weighted mean is preferred in this case because the office buildings are of different sizes. There are more square feet at $110 per square foot than there are square feet at $45 per square foot. Therefore, the data values (price per square foot) must be weighted to reflect the differences in building size.

3.15. a. $\text{Average}=\dfrac{\sum x_i}{n}=\dfrac{11259.9}{20}=562.99$

b. The 17th observation is quite larger than the rest of the data, an *outlier*. Averages, but not medians, are highly affected by outliers. Therefore, the median would be an appropriate measure for this data set. Arranged in numerical order, the data is

400.56	464.37	474.86	475.87	511.15	528.78	531.64	533.70	538.20	545.25
558.12	564.71	567.46	588.39	589.15	606.70	610.32	625.82	632.14	912.68

Calculating the index produces (p/100)n = (50/100)20 = 10. This is an integer. The rule says to average the 10th and 11th observation = (545.25 + 558.12)/2 = 551.685.

The outlier will have the largest effect.

Deleting it, produces the average $=\dfrac{\sum x_i}{n}=\dfrac{11259.9-912.68}{20-1}=544.59.$

The observation closest to the original mean would have the least effect. That number is 564.71.

Deleting it, produces the average $=\dfrac{\sum x_i}{n}=\dfrac{11259.9-564.71}{20-1}=562.90..$

3.16. a. $\text{Average}=\dfrac{\sum x_i}{n}=\dfrac{4143.6}{25}=165.744$

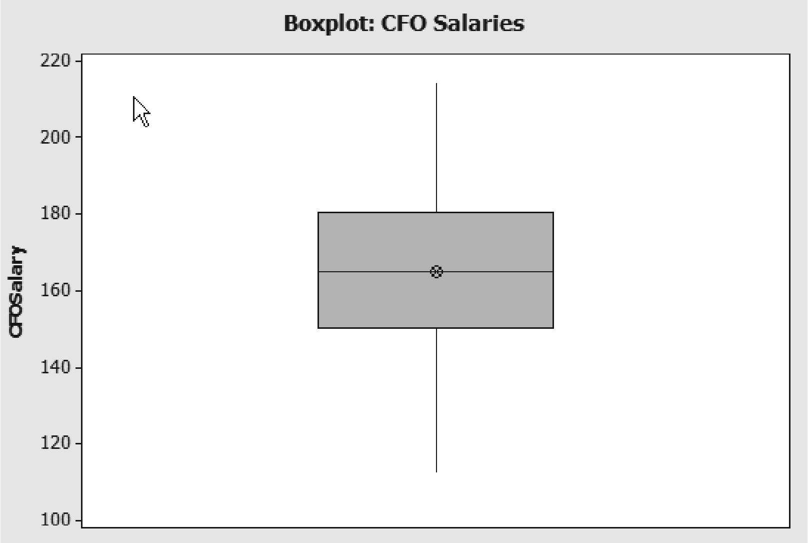
b. The ranked data are

112.6	123.9	131.0	134.0	141.9	145.4	155.2	156.5	159.3
161.3	161.9	162.7	164.9	165.8	168.3	171.2	173.1	177.4
178.8	182.0	185.8	192.0	211.1	213.1	214.4		

The median's index = (p/100)n = (50/100)25 = 12.5. The rule is to round this up to the 13th observation = 164.9. This indicates that the mean is larger than the median. The data's distribution is slightly right-skewed.

c. The first quartile's index = (25/100)25 = 6.25. The rule is to round this up to the 7th observation = 155.2.

The third quartile's index = (75/100)25 = 18.75. The rule is to round this up to the 19th observation = 178.8.



For a Boxplot:

Lower limit = $Q_1 - 1.5(Q_3 - Q_1) = 155.2 - 1.5(178.8 - 155.2) = 119.8$

Upper limit is $Q_3 + 1.5(Q_3 - Q_1) = 178.8 + 1.5(178.8 - 155.2) = 214.2$

There are no outliers shown in the plot although there are two, 214.4 and 112.6 in the data.

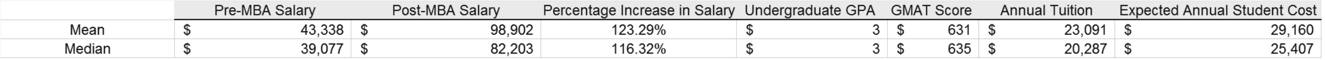
3.17. a. FDIC Commercial Banks Mean Assets = $\dfrac{\sum x_i}{n} = \dfrac{14{,}727}{5{,}410} = \$2.72 \text{ billion}$

FDIC Savings Institutions Mean Assets = $\dfrac{\sum x_i}{n} = \dfrac{1{,}074}{860} = \$1.25 \text{ billion}$

b. This shows that the mean deposits for Commercial Banks is about $1.50 billion more than the mean for Savings Institutions

c. Because these data include all banks and savings entities, the means would be considered to be parameters.

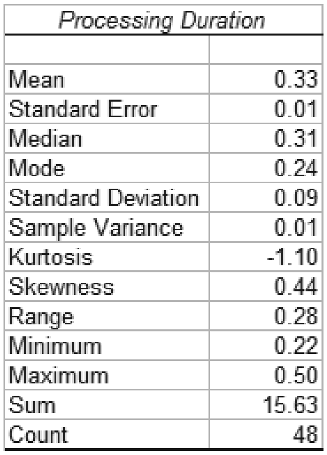
3.18.



	Pre-MBA Salary	Post-MBA Salary	Percentage Increase in Salary	Undergraduate GPA	GMAT Score	Annual Tuition	Expected Annual Student Cost
Mean	$ 43,338	$ 98,902	123.29%	$ 3	$ 631	$ 23,091	$ 29,160
Median	$ 39,077	$ 82,203	116.32%	$ 3	$ 635	$ 20,287	$ 25,407

Student reports will vary but all should report the above statistics. Comments should be made about skewness. Also, bar charts could be used to display the individual school values with mean and median values shown using call-outs on the charts. Schools with values substantially higher or lower than average might be called out on the charts or in the report.

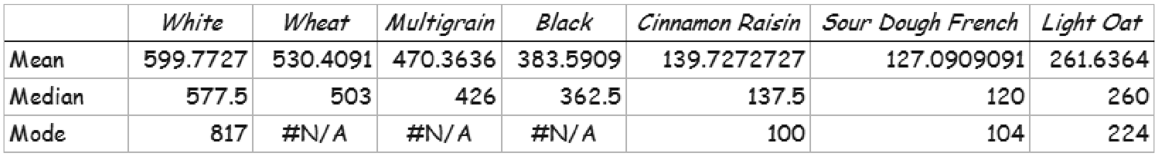
3.19. Software can be used to compute these values. For instance, using the Descriptive Statistics Tool in Excel we get the following.



Processing Duration	
Mean	0.33
Standard Error	0.01
Median	0.31
Mode	0.24
Standard Deviation	0.09
Sample Variance	0.01
Kurtosis	-1.10
Skewness	0.44
Range	0.28
Minimum	0.22
Maximum	0.50
Sum	15.63
Count	48

Thus, the mean time is .33 minutes. The median is .31 minutes and the mode is .24 minutes. The data are slightly right skewed.
The 80th percentile is computed using the Percentile function in Excel to be .40 minutes.

3.20. Excel can be used to compute the statistics shown as follows:

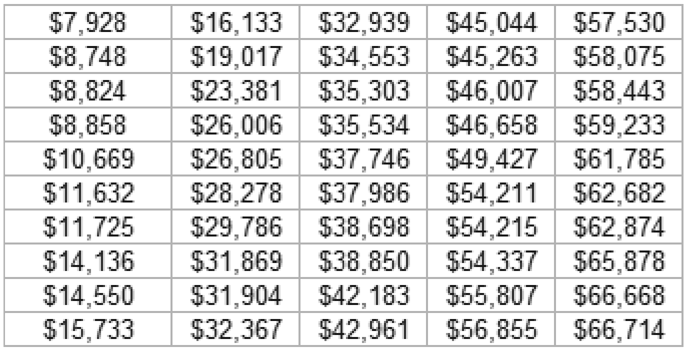


	White	*Wheat*	*Multigrain*	*Black*	*Cinnamon Raisin*	*Sour Dough French*	*Light Oat*
Mean	599.7727	530.4091	470.3636	383.5909	139.7272727	127.0909091	261.6364
Median	577.5	503	426	362.5	137.5	120	260
Mode	817	#N/A	#N/A	#N/A	100	104	224

The student reports will vary but should contain a discussion of the above statistics. Bar charts could be used to display the mean and median for the seven bread categories.

3.21. a. Using Excel's AVERAGE function, the mean index value = 108.13
Using Excel's MEDIAN function, the median value = 107.9
b. Using Excel's QUARTILE Function, the 1st Quartile = 106.8 and the 3rd Quartile = 110.4

3.22. Step 1: Sort the data from low to high.
The data, sorted by column, are:



$7,928	$16,133	$32,939	$45,044	$57,530
$8,748	$19,017	$34,553	$45,263	$58,075
$8,824	$23,381	$35,303	$46,007	$58,443
$8,858	$26,006	$35,534	$46,658	$59,233
$10,669	$26,805	$37,746	$49,427	$61,785
$11,632	$28,278	$37,986	$54,211	$62,682
$11,725	$29,786	$38,698	$54,215	$62,874
$14,136	$31,869	$38,850	$54,337	$65,878
$14,550	$31,904	$42,183	$55,807	$66,668
$15,733	$32,367	$42,961	$56,855	$66,714

Step 2: Determine the 20th percentile location index and find the 20th percentile.

To determine the location index for the 20th percentile (p = 20) we do the following:

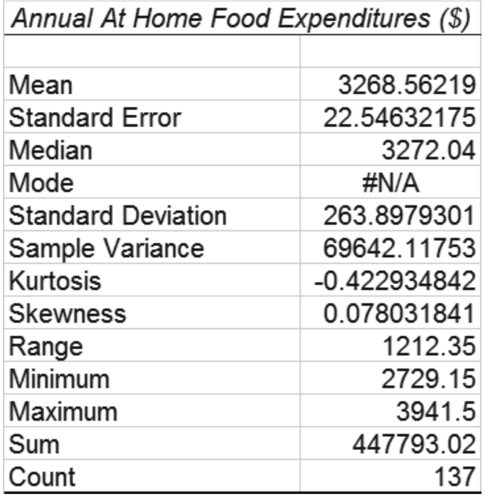
$$i = \frac{p}{100}(n) = \frac{20}{100}(50) = 10$$

Since the index, 10, is an integer, the 20th percentile is determined by finding the average of the 10th and 11th values from the lower end of the sorted data. This is:

20th Percentile = (15,733 + 16,133)/2 = 15,933

Thus, based on these sample data, the 20th percentile is determined to be $15,933. Thus, homeowners age 65 or older with incomes at or below $15,933 will be eligible for property tax relief under the current proposal.

3.23. a. Descriptive Statistics: Expenditures

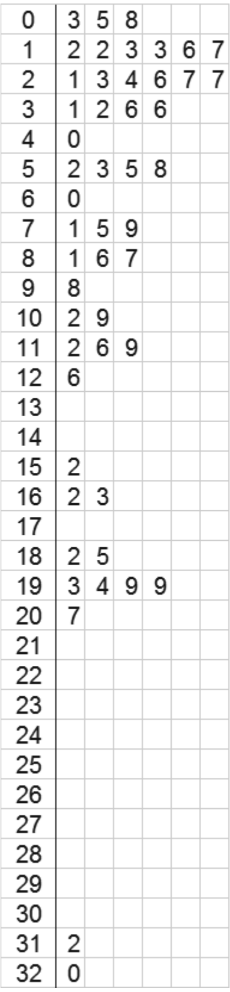


Annual At Home Food Expenditures ($)	
Mean	3268.56219
Standard Error	22.54632175
Median	3272.04
Mode	#N/A
Standard Deviation	263.8979301
Sample Variance	69642.11753
Kurtosis	-0.422934842
Skewness	0.078031841
Range	1212.35
Minimum	2729.15
Maximum	3941.5
Sum	447793.02
Count	137

b. The mean is 3,268.56 and the median is 3,272.04. Because the mean and median are close in value the data are symmetric. The box and whisker plot also shows that the data are symmetric.

c. The first quartile is approximately $3,109.76 and the third quartile is approximately $3,438.72. Thus, approximately the middle 50% of the data values are between $3,109.76 and $3,438.72.

3.24. a. The stem-and-leaf plot.



0	3	5	8			
1	2	2	3	3	6	7
2	1	3	4	6	7	7
3	1	2	6	6		
4	0					
5	2	3	5	8		
6	0					
7	1	5	9			
8	1	6	7			
9	8					
10	2	9				
11	2	6	9			
12	6					
13						
14						
15	2					
16	2	3				
17						
18	2	5				
19	3	4	9	9		
20	7					
21						
22						
23						
24						
25						
26						
27						
28						
29						
30						
31	2					
32	0					

The positive outliers and the shape of the data indicate that the distribution is right-skewed.

b. $\bar{x} = \dfrac{4429}{50} = 88.58,$ the median's index = (p/100)n = (50/100)50 = 25. Therefore, the median is the average of the 25th and 26th data values: $M_d = \dfrac{60+71}{2} = 65.5$. Since $M_d < \bar{x},$ these statistics would indicate, as did the stem-and-leaf display, that the distribution is right-skewed.

c. All evidence considered indicates a skewed distribution. The largest two observations, 312 and 320, appear to be large with respect to the rest of the data. They are outliers. The mean is unduly influenced by outliers and skewed data. Therefore, the median should be used as the measure of centrality.

Section 3.2

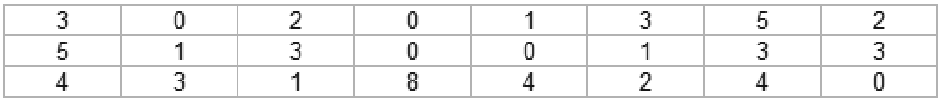
3.25. a. The range is the difference between the high value and the low value in the set of data.

Range = High – Low

Range = 8 – 0 = 8

b. The steps required to compute the sample variance are:

Step 1: Select the sample and record the data.



3	0	2	0	1	3	5	2
5	1	3	0	0	1	3	3
4	3	1	8	4	2	4	0

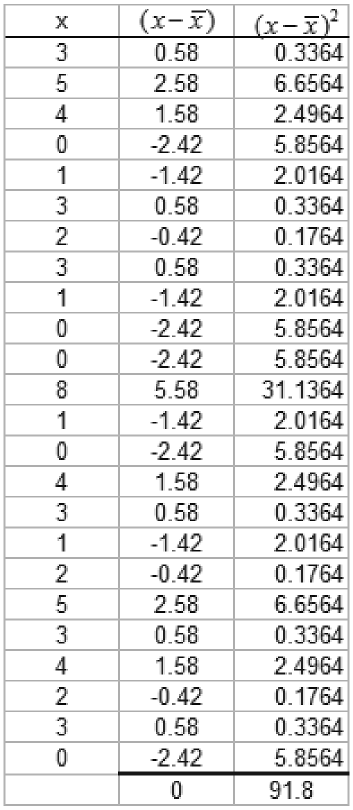
Step 2: Select the desired equation for computing the sample variance:

$$s^2 = \frac{\sum (x-\overline{x})^2}{n-1}$$

Step 3: Compute the sample mean.

$$\overline{x} = \frac{\sum x}{n} = \frac{58}{24} = 2.42$$

Step 4: Determine the sum of the squared deviations of each x and $\overline{x}$.



x	$(x-\overline{x})$	$(x-\overline{x})^2$
3	0.58	0.3364
5	2.58	6.6564
4	1.58	2.4964
0	-2.42	5.8564
1	-1.42	2.0164
3	0.58	0.3364
2	-0.42	0.1764
3	0.58	0.3364
1	-1.42	2.0164
0	-2.42	5.8564
0	-2.42	5.8564
8	5.58	31.1364
1	-1.42	2.0164
0	-2.42	5.8564
4	1.58	2.4964
3	0.58	0.3364
1	-1.42	2.0164
2	-0.42	0.1764
5	2.58	6.6564
3	0.58	0.3364
4	1.58	2.4964
2	-0.42	0.1764
3	0.58	0.3364
0	-2.42	5.8564
	0	91.8

Step 5: Compute the sample variance.

$$s^2 = \frac{\sum (x-\overline{x})^2}{n-1} = \frac{91.8}{24-1} = 3.99$$

c. The sample standard deviation is the square root of the variance.

$$s = \sqrt{\frac{\sum (x-\overline{x})^2}{n-1}} = \sqrt{3.99} = 1.998$$

3.26. a. Range $= 9 - 4 = 5$

b.

x	$x-\mu$	$(x-\mu)^2$
4	–1.83333	3.361111
6	0.166667	0.027778
9	3.166667	10.02778
4	–1.83333	3.361111
5	–0.83333	0.694444
7	1.166667	1.361111
		18.83333

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = 35/6 = 5.8333$$

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x-\mu)^2}{N} = 18.83333/6 = 3.1389$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{3.1389} = 1.7717$$

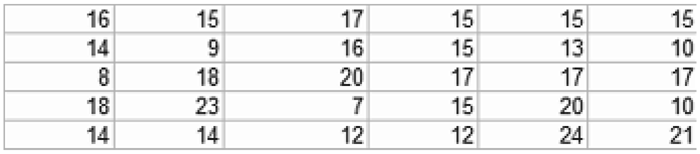
c. $$s^2 = \frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n-1} = 18.83333/(6-1) = 3.7667$$

$$S = \sqrt{S^2} = \sqrt{3.337667} = 1.9408$$

The statistics (assuming the data were a sample) will be larger than the parameters (assuming the data were a population) since the division for s and s^2 are computed using a divisor of $n-1$ rather than N.

3.27. a. The population variance is computed using the following steps. Note Equation 3.9 is used in this solution.

Step 1: Collect the data for the population.



16	15	17	15	15	15
14	9	16	15	13	10
8	18	20	17	17	17
18	23	7	15	20	10
14	14	12	12	24	21

Step 2: Select **SelectEquation 3.10** as the population variance's calculation formula.

Step 3: Calculate the population mean.

$$\mu = \frac{\sum x}{N} = \frac{457}{30} = 15.23$$

Step 4: Compute the sum of squared deviations from the mean.

$$\sum(x-\mu)^2 = (16-15.23)^2 + (15-15.23)^2 + \ldots(21-15.23)^2 = 506.24$$

Step 5: Compute the population variance.

b. The population standard deviation is the square root of the population variance.

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}} = \sqrt{\frac{506.24}{30}} = \sqrt{16.87} = 4.11$$

3.28. The sorted array is shown below

4132	4188	4209	4423	4568
4573	4983	5002	5052	5176
5310	5381	5611	5736	6005

The range is equal to the maximum value – the minimum value = 6005 – 4132 = 1873

The variance $$\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{373{,}430{,}607-(74{,}349^2)/15}{15-1} = 350{,}882.4$$

The standard deviation $=\sqrt{350{,}882.4}=592.3533$

3.29. The sorted data is shown below.

5.4	6.6	7.5	7.8
8.5	8.9	10.3	11.5
12	12.2	13	14.4

The range is equal to the maximum value minus the minimum value = 14.4 – 5.4 = 9.
The first quartile's position is $\left(25/100\right)*12=3$. Therefore Q_1 is the average of the data values in the 3rd and 4th position = (7.5 + 7.8)/2 = 7.65.
The third quartile's position is $\left(75/100\right)*12=9$. Therefore Q_3 is the average of the data values in the 9th and 10th position = (12 + 12.2)/2 = 12.1.
The IQR = $Q_3 - Q_1$ = 12.1 – 7.65 = 4.45.
The variance is 7.86 and the standard deviation is 2.8.

The variance $$\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{1{,}248.81-(118.1^2)/12}{12-1} = 7.86$$

The standard deviation $=\sqrt{7.86}=2.8$

3.30. a. First data set: Range = largest – smallest = 37 – 10 = 27, $s^2 = \frac{\sum(x_i-\overline{x})^2}{n-1} = \frac{794}{9} = 88.322,$

$s=\sqrt{s^2}=\sqrt{88.322}=9.40,$ the index for Q_1 is: $i = (p/100)n$ = (25/100)10 = 2.5. i = 3.
Therefore, Q_1 = 15. For Q_3 the index is $(p/100)n$ = (75/100)10 = 7.5. i = 8.
Therefore, Q_3 = 31. So the IQR = $Q_3 - Q_1$ = 31 – 15 = 16.

Second data set: Range = largest – smallest = 118 – 1 = 117, $s^2 = \frac{\sum(x_i-\overline{x})^2}{n-1} = \frac{10892}{9} =$ 1210.22, $s=\sqrt{s^2}=\sqrt{1210.22}=34.79$, the index for Q_1 is: $i = (p/100)n$ = (25/100)10 = 2.5. i = 3. Therefore, Q_1 = 3. For Q_3 the index is $(p/100)n$ = (75/100)10 = 7.5. i = 8.
Therefore, Q_3 = 16. So the IQR = $Q_3 - Q_1$ = 16 – 3 = 13.

b. Each of the statistics in the second data set is larger than its counterpart in the first data set except the Interquartile range. Therefore, the second data set seems to have the most spread out data.

c. First data set: Range = largest – smallest = 35 – 10 = 25, $s^2 = \frac{\sum(x_i-\overline{x})^2}{n-1} = \frac{604.22}{8} = 75.53,$

$s=\sqrt{s^2}=\sqrt{75.53}=8.69,$ the index for Q_1 is: $i = (p/100)n$ = (25/100)9 = 2.25. i = 3.

Therefore, $Q_1 = 15$. For Q_3 the index is $(p/100)n = (75/100)9 = 6.75$. $i = 7$.
Therefore, $Q_3 = 27$. So the IQR = $Q_3 - Q_1 = 27 - 15 = 12$.

Second data set: Range = largest – smallest = 18 – 1 = 17, $s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1} = \frac{437.56}{8} = 54.69$,

$s = \sqrt{s^2} = \sqrt{54.69} = 7.40$, the index for Q_1 is: $i = (p/100)n = (25/100)9 = 2.25$. $i = 3$.
Therefore, $Q_1 = 3$. For Q_3 the index is $(p/100)n = (75/100)9 = 6.75$. $i = 7$.
Therefore, $Q_3 = 16$. So the IQR = $Q_3 - Q_1 = 16 - 3 = 13$.

d. The variance for the second data set with the outlier was 1210.22 and without it was 54.69. This was the largest change. Therefore, the variance seems to be the statistic among these that is most affected by outliers.

3.31. a. Range = largest – smallest = 30 – 6 = 24, $s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1} = \frac{727.33}{14} = 51.95$,

$s = \sqrt{s^2} = \sqrt{51.95} = 7.21$, the index for Q_1 is: $i = (p/100)n = (25/100)15 = 3.75$. $i = 4$.
Therefore, $Q_1 = 12$.
For Q_3 the index is $(p/100)n = (75/100)15 = 11.25$. $i = 12$.
Therefore, $Q_3 = 24$. So the IQR = $Q_3 - Q_1 = 24 - 12 = 12$

b. Range = largest – smallest = 30 – 6 = 24, $\sigma^2 = \frac{\sum(x_i - \mu)^2}{N} = \frac{727.33}{15} = 48.49$,

$\sigma = \sqrt{48.49} = 6.96$, the index for Q_1 is: $i = (p/100)n = (25/100)15 = 3.75$. $i = 4$.
Therefore, $Q_1 = 12$.
For Q_3 the index is $(p/100)n = (75/100)15 = 11.25$. $i = 12$.
Therefore, $Q_3 = 24$. So the IQR = $Q_3 - Q_1 = 24 - 12 = 12$

c. σ^2 is smaller than s^2 by a factor of $(N - 1)/N$. σ is smaller than s by a factor of $\sqrt{(N-1)/N}$.
The range is not affected.

3.32. Students may decide on different measures but some or all of the following could be computed.
Measures of the center include:

$$\text{Mean } = \overline{x} = \frac{\sum x}{n} = \frac{674}{20} = 33.7$$

Median = 34.5

Mode = (two modes; 34 and 39 both occurring twice)

Measures of spread include:

Range = High – Low

Range = 75 – 8 = 67

Interquartile Range = $Q_3 - Q_1 = 39.5 - 25.5 = 14.0$

$$\text{Variance } = s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1} = 217.9$$

$$\text{Standard Deviation } = s = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n-1}} = 14.76$$

3.33. a. The sorted data is shown below.

35	50	50	50
60	75	75	75
80	85	85	90
90	100	100	100
100	125	125	150

The range is equal to the maximum value minus the minimum value = 150 – 35 = 115.

The first quartile's position is $\left(25/100\right)*20=5$. Therefore Q$_1$ is the average of the data values in the 5th and 6th position = (60+75)/2 = 67.5.

The third quartile's position is $\left(75/100\right)*20=15$. Therefore Q$_3$ is the average of the data values in the 15th and 16th position = (100 + 100)/2 = 100.

The IQR = Q$_3$ – Q$_1$ = 100 – 67.5 = 32.5.

The variance is 815.79 and the standard deviation is 28.56.

$$\text{The variance } = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{160{,}000-(1.700^2)/20}{20-1} = 815.79$$

$$\text{The standard deviation } = \sqrt{815.76} = 28.56$$

b. The range is computed by taking the difference between the two extreme values in a data set. That is the difference between the maximum and the minimum. The interquartile range looks at the middle 50% of the data values by taking the difference between the 75th and the 25th percentile. Both are measures of variability but the interquartile range overcomes the susceptibility of the range to being highly influenced by extreme values.

3.34.

X	$X-\overline{x}$	$(X-\overline{x})^2$
10	–6.1	37.21
19	2.9	8.41
17	0.9	0.81
19	2.9	8.41
12	–4.1	16.81
20	3.9	15.21
20	3.9	15.21
15	–1.1	1.21
16	–0.1	0.01
13	–3.1	9.61
161		112.9

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 161/10 = 16.1$$

$$S^2 = \frac{\sum_{i=1}^{n} (x-\overline{x})^2}{n-1} = 112.9/(10-1) = 12.5444$$

$$S = \sqrt{S^2} = \sqrt{12.5444} = 3.5418$$

16.1 + (1)3.5418 = 19.6418 or > 19 visits are required

3.35.

X	$X-\bar{x}$	$(X-\bar{x})^2$
32	5.9	34.81
22	–4.1	16.81
24	–2.1	4.41
27	0.9	0.81
27	0.9	0.81
33	6.9	47.61
28	1.9	3.61
23	–3.1	9.61
24	–2.1	4.41
21	–5.1	26.01
261		148.9

a. range = 33 – 21 = 12

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}=261/10=26.1$$

$$S^2=\frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n-1}=148.9/(10-1)=16.5444$$

$$S=\sqrt{S^2}=\sqrt{16.5444}=4.0675$$

the 1st quartile is equal to the 25th percentile

$i=\frac{p}{100}(n)=(25/100)(10)=2.5$ or the 3rd observation = 23

the 3rd quartile is equal to the 75th percentile

$i=\frac{p}{100}(n)=(75/100)(10)=7.5$ or the 8th observation = 28

Interquartile Range = 28 – 23 = 5

b. Student answers will vary but they should look at the number of standard deviations the mean for this school is from the U. S. mean. U.S. Mean (37.8) – This Mean (26.1) = 11.7 which is 11.7/4.0675 = 2.9 or almost 3 standard deviations from the U.S. mean. Given this, although we are working with a small sample, there appears to be evidence to suggest that the ages are lower at this university than for the U.S. Colleges and Universities as a group.

3.36. a. The mean loan balance is computed as follows:

$$\bar{x}=\frac{\sum x}{n}=\frac{\$127{,}227}{10}=\$12{,}722.7$$

b. The standard deviation is computed by finding the square root of the variance.

$$\text{Variance }=s^2=\frac{\sum(x_i-\bar{x})^2}{n-1}=20{,}462{,}362$$

Standard Deviation = $4,524

c. Student paragraphs will differ but should point out the center of the loan balance distribution is centered at $12,722 but that the loan balances are quite widely spread around the center. The standard deviation is a measure of the average deviation with a value of $4,524.

3.37. a. The range, interquartile range, variance, and standard deviation are shown below (calculated using Minitab).

```
Variable          StDev     Variance    Range      IQR
Speed of Exit     34.89     1217.14     113.00     62.25
```

The range is 113.0, the IQR is 62.25, the variance is 1217.14, and the standard deviation is 34.89.

b. No, the interquartile range looks at the middle 50% of the values so it is not affected by changes to the extreme values.

c. Adding a constant to all the data values leaves the variance unchanged.

3.38. a. We must first convert the means to measures in minutes. They are 15 + 48/60 = 15.8 and $14 + 55/60 = 14.92$. $\bar{x}_1 - \bar{x}_2 = 15.8 - 14.92 = 0.88$.

b. If s = 5 minutes, then there are 0.88/5 = 0.18 standard deviations between the two means. This seems as though it could be solely due to randomization.

c. If s = 0.25 minutes, then there are 0.88/0.25 = 3.52 standard deviations between the two means. A difference of 3.52 standard deviations is not likely due solely to randomness. We might conclude, then, that there is a real difference between the means these values represent.

3.39. a. $\mu = \dfrac{\sum x_i}{N} = \dfrac{-12.9}{11} = -1.2$

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \frac{9.722}{11} = 0.883, \quad \sigma^2 = \sqrt{\sigma^2} = \sqrt{0.883} = 0.940$$

IQR = Q3 – Q1 = –.9 – –1.6 = .7

b. Mean = –1.2 The mean is negative, indicating prices have fallen slightly (by an average of –1.2 percent per month) during the time period.

c. The standard deviation measures the average price fluctuation from month to month for the price index.

3.40. a. $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{20,261}{15} = \$1,351$

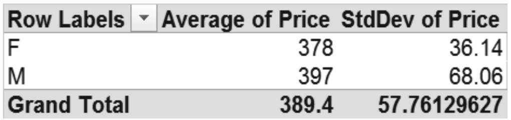
$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{297,722.9}{14} = 21,265.9, \quad s = \sqrt{s^2} = \sqrt{21,265.9} = \$145.83$$

b. The most extreme value is $1,075. Subtract the mean to determine the distance between it at the extreme value: $1,075 – $1,351 = –$276. Now divide this distance by the standard deviation:

–276/145.83 = –1.89. Tchebysheff's Theorem indicates that at least $(1 - 1/k^2)$ = $[1 - 1/(1.89)^2]$ = 0.72 of the data should be within 1.89 standard deviations of the mean. The boundaries for this are $1,351 ± 1.89(145.83) = 1,351 ± 275.62 ($1075,38 to $1626,62).

3.41. Software such as Excel or Minitab can be used to do the computations required in parts a. and b. We have used Excel's pivot table feature to provide the desired calculations.

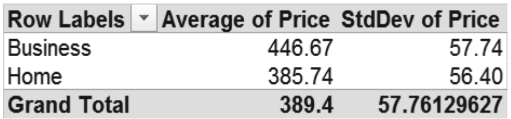
a. The mean and standard deviation for male and female customer phone purchase prices are shown as follows:



Row Labels	Average of Price	StdDev of Price
F	378	36.14
M	397	68.06
Grand Total	**389.4**	**57.76129627**

In this sample, males spent an average of $117 while females spent an average of $98 for their phones. The standard deviation for males was nearly twice that for females.

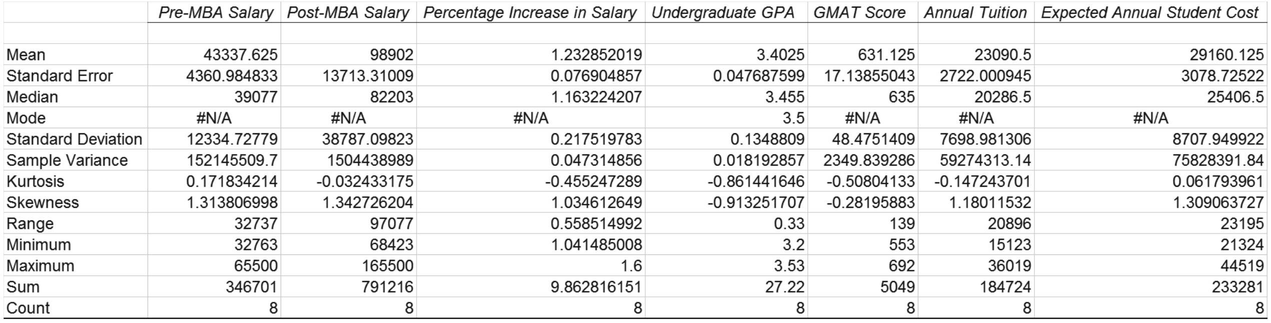
b. The mean and standard deviation for home and business use customer phone purchase prices are shown as follows:



Row Labels	Average of Price	StdDev of Price
Business	446.67	57.74
Home	385.74	56.40
Grand Total	**389.4**	**57.76129627**

In this sample, business users spent an average of $446.67 on their phone while home users spent an average of $385.74. The variation in phone costs for the two groups was about equal.

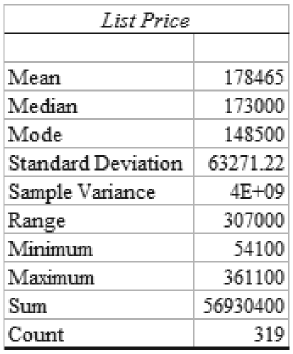
3.42. The descriptive statistics can be computed using software such as Excel or Minitab. Excel's descriptive statistics option under Data – Data Anaysis is used to provide the following results:



	Pre-MBA Salary	*Post-MBA Salary*	*Percentage Increase in Salary*	*Undergraduate GPA*	*GMAT Score*	*Annual Tuition*	*Expected Annual Student Cost*
Mean	43337.625	98902	1.232852019	3.4025	631.125	23090.5	29160.125
Standard Error	4360.984833	13713.31009	0.076904857	0.047687599	17.13855043	2722.000945	3078.72522
Median	39077	82203	1.163224207	3.455	635	20286.5	25406.5
Mode	#N/A	#N/A	#N/A	3.5	#N/A	#N/A	#N/A
Standard Deviation	12334.72779	38787.09823	0.217519783	0.1348809	48.4751409	7698.981306	8707.949922
Sample Variance	152145509.7	1504438989	0.047314856	0.018192857	2349.839286	59274313.14	75828391.84
Kurtosis	0.171834214	-0.032433175	-0.455247289	-0.861441646	-0.50804133	-0.147243701	0.061793961
Skewness	1.313806998	1.342726204	1.034612649	-0.913251707	-0.28195883	1.18011532	1.309063727
Range	32737	97077	0.558514992	0.33	139	20896	23195
Minimum	32763	68423	1.041485008	3.2	553	15123	21324
Maximum	65500	165500	1.6	3.53	692	36019	44519
Sum	346701	791216	9.862816151	27.22	5049	184724	233281
Count	8	8	8	8	8	8	8

Student reports will differ but should contain a discussion of the measures of the center and the measures of the spread.

3.43. Software such as Excel or Minitab can be used to compute the descriptive statistics in this exercise. Excel's descriptive statistics tool under Tools — Data Analysis has been used here.



List Price	
Mean	178465
Median	173000
Mode	148500
Standard Deviation	63271.22
Sample Variance	4E+09
Range	307000
Minimum	54100
Maximum	361100
Sum	56930400
Count	319

a. The population mean is:

$$\mu = \frac{\sum x}{N} = \$178,465$$

b. The population median is:

$$\tilde{\mu} = \$173,000$$

c. The range is:

R = High – Low
R = \$361,100 – \$54,100
= \$307,000

d. The population standard deviation is:

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}} = \$63,172$$

Note, when using Excel, the function STDEVP is used to find the population mean. The Descriptive Statistics tool under Tools – Data Analysis returns the sample standard deviation.

e. Student reports will vary but should include a discussion of the measures of the center and the measures of spread. Top students will think of attaching a histogram of the list prices to show graphically the distribution. They will use annotation to add the mean, median and standard deviation to the graph.

3.44. The solution was obtained using Minitab.

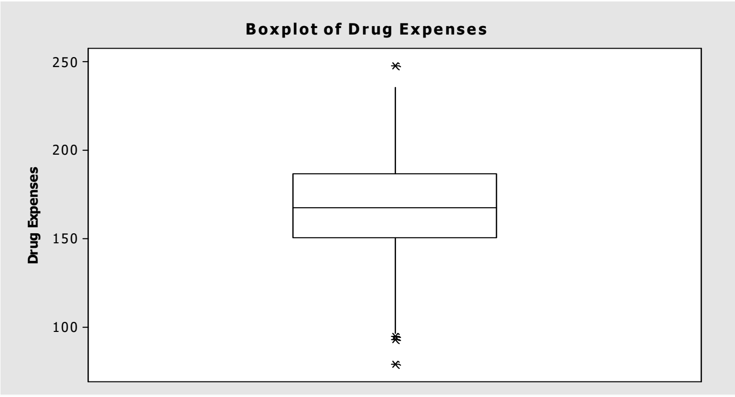
```
Variable        Mean      StDev    Variance  Minimum  Q1       Median   Q3
Drug Expenses   167.52    28.52    813.27    79.39    150.57   167.60   187.06
Variable        Maximum   Range    IQR
Drug Expenses   247.70    168.31   36.50
```

a. The mean is 167.52. The median is 167.60.

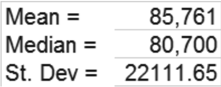
b. The range is 168.31. The variance is 813.27. The standard deviation is 28.52. The interquartile range is 36.50.

c. The box and whisker plot is shown below.

d. The mean and median are approximately equal indicating that the data are symmetrically distributed. The largest out-of pocket expense was \$247.70 and the smallest was \$79.39, resulting in a range of \$168.31. The middle 50% of out-of-pocket expenditures fell between \$150.57 and \$187.06. The standard deviation of expenditures was found to be \$28.52. There are a few extreme values or outliers in the data indicating unusually large or small expenditures.



3.45. a.

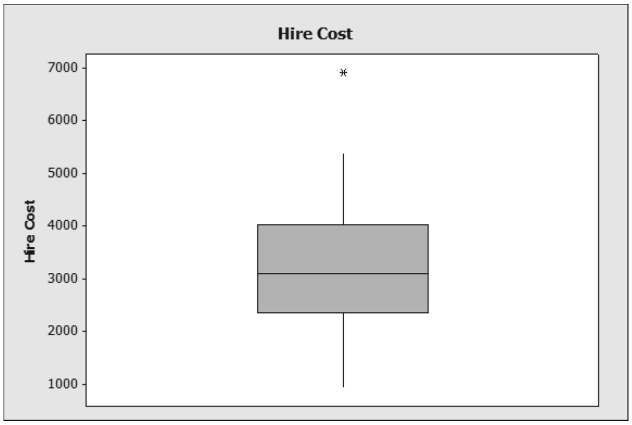


Mean =	85,761
Median =	80,700
St. Dev =	22111.65

b. 80th percentile = 110,360

c. Interquartile Range = $Q_3 - Q_1$ = 108,000 – 72,133 = 35,867

3.46. a. $s^2 = \dfrac{\sum (x_i - \overline{x})^2}{n-1} = \dfrac{68301100}{49} = 1393900,\ \ s = \sqrt{s^2} = \sqrt{1393900} = 1180.64$



b. Minitab was used to find the box plot.
The distribution seems to be skewed to the right.

c. The interquartile range is a measure of dispersion not affected by outliers.
IQR = $Q_3 - Q_1$. The index for Q_1 is: $i = (p/100)n = (25/100)50 = 12.5$, found to 13.
Therefore, $Q_1 = 2360$.
For Q_3 the index is $(p/100)n = (75/100)50 = 37.5$, round to 38.
Therefore, $Q_3 = 3984$. So the IQR = $Q_3 - Q_1 = 3984 - 2360 = 1624$

Section 3.3

3.47. a. Since the shape of the distribution is unknown, we can use Tchebysheff's theorem to determine the solution. Since $\sigma = 200$ and $\mu = 3{,}000$, then the range 2,600 to 3,400 is $\mu \pm 2(\sigma)$. According to Tchebysheff's theorem, at least 75% of the data in a distribution will fall within $\mu \pm 2(\sigma)$.

b. The range $\mu \pm 3(\sigma)$ should include at least 8/9 (89%) of the data values. That means that the range 2,400 to 3,600 should contain at least 89 percent of the data values. However, since we don't know the shape of the population, we can't say for sure what percentage will be greater than 3,600. We do know that the percentage will be less than 11% and most likely it will be considerably less.

c. The same issue is present here as in part b. We can't say with any certainty what the percentage will be. We do know that it will be less than 11 percent but we don't know how much less.

3.48. a. The standardized value is computed using:

$$z = \frac{x - \mu}{\sigma}$$

For a value, $x = 500$, the standardized value is:

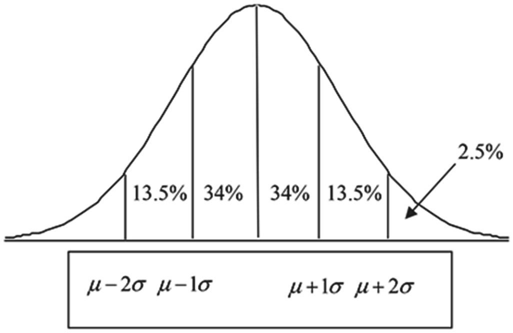
$$z = \frac{500 - 400}{50} = \frac{100}{50} = 2$$

Thus, a bulb that lasts 500 hours is 2 standard deviations higher than the population mean.

b. If the time distribution is bell shaped, the Empirical Rule can be used to determine the percentage of bulbs expected to last over 500 hours. The Empirical Rule states:

Approximately 95% of the data will fall with ±2 standard deviations from the mean.

This is shown as follows:



Thus, a bulb lasting 500 hours is two standard deviations above the mean. Only 2.5 percent of all bulbs are expected to last longer than 500 hours assuming that the distribution is approximately bell shaped.

3.49. a. The sample mean is computed by summing the data values and dividing the sum by the number of observations. The mean = 1487/14 = 106.21. The sample standard deviation is found as follows. Add the x values and square the sum = 1487^2 = 2,211,169. Divide this value by the number of observations, 14, which gives 2,211,169/14 = 157,940.643. Square each of the x values and sum those squares = 166,071. Subtract 157,940.643 from 166,071 and divide the difference by $n - 1$ or 13. The sample standard deviation is the square root of this result.

$$\sqrt{8130.3571/13} = 25.008$$

b. The coefficient of variation is the ratio of the standard deviation to the mean expressed as a percentage. Here the coefficient of variation is (25.008/106.21)*100% = 23.55%. The coefficient of variation is a measure of the relative variation in the data.

c. The range of values that should include at least 89% of the data values according to Tchebysheff's Theorem is computed as being within 3 standard deviations of the mean. Thus, the range from 31.19 to 181.24 should contain 89% of the data values. In this instance the range contains all the data values. The interval range using Tchebysheff's Theorem was conservative.

3.50. a. The coefficient of variation is computed at follows:

$$CV = \frac{\sigma}{\mu}(100)$$

Population 1 –

$$CV = \frac{\sigma}{\mu}(100)$$

$$CV = \frac{50}{700}(100) = 7\%$$

Population 2 –

$$CV = \frac{\sigma}{\mu}(100)$$

$$CV = \frac{5{,}000}{29{,}000}(100) = 17\%$$

b. Based on the coefficients of variation for the two populations, population 2 has a $CV = 17\%$ while population 1 has a $CV = 7\%$. Thus, population 2 is more variable relative to the size of the population mean than is population 1.

3.51. The coefficient of variation is used to measure the relative variability of two or more distributions. It is computed using:

$$CV = \frac{\sigma}{\mu}(100)$$

For Distribution A we get: $CV = \frac{\sigma}{\mu}(100) = \frac{100}{500}(100) = 20\%$

For Distribution B we get: $CV = \frac{\sigma}{\mu}(100) = \frac{4.0}{10.0}(100) = 40\%$

Thus, even though distribution B has a standard deviation that is only 4 percent the size of A's standard deviation, distribution B is relatively more variable because the mean of A is so much greater than the mean of B.

3.52. The standardized value is computed using:

$$z = \frac{x - \mu}{\sigma}$$

Distribution A: $z = \frac{50{,}000 - 45{,}000}{6{,}333} = 0.695$

Distribution B: $z = \frac{40 - 33.40}{4.05} = 1.63$

The smaller the *z* value, the relatively closer the *x* value is to the mean. Thus, the 50,000 value is .695 standard deviations from the mean of distribution A while the value 40 is 1.63 standard deviations from the mean of distribution B. The value from distribution A is relatively closer to its mean.

3.53. a. The standardized value is $z = \frac{800 - \bar{x}}{s} = \frac{800 - 1000}{250} = -0.80$

b. The standardized value is $z = \frac{1200 - \bar{x}}{s} = \frac{1200 - 1000}{250} = 0.80$

c. The standardized value is $z = \dfrac{1000 - \bar{x}}{s} = \dfrac{1000 - 1000}{250} = 0.00$

3.54. a. The sample mean and standard deviation for Population A were computed using Excel and are shown below.

Mean	48.33
Standard Deviation	21.87

b. The sample mean and standard deviation for Population B were computed using Excel and are shown below.

Mean	1,023.56
Standard Deviation	75.46

c. Based on the sample standard deviations Population B has greater spread than Population A.

d. The sample coefficient of variation is the ratio of the sample mean to the sample standard deviation expressed as a percentage. For Population A, the sample coefficient of variation is $(21.87/48.33)*100\% = 45.25\%$. For Population B the sample coefficient of variation is $(75.46/1{,}023.56)*100\% = 7.37\%$. The sample coefficient of variation is a measure of the relative variation in the two populations. The larger the coefficient of variation the more variable the data is relative to its mean. Because Population A has a higher sample coefficient of variation, it has the greater variation relative to its mean.

3.55. a. $\bar{x} = \dfrac{1530}{30} = 51,\ s^2 = \dfrac{\sum (x_i - \bar{x})^2}{n-1} = \dfrac{14806}{29} = 510.55,\ s = \sqrt{s^2} = \sqrt{510.55} = 22.6$

b. Therefore, $\bar{x} \pm s,\ \bar{x} \pm 2s,\ \bar{x} \pm 3s$ are, respectively, $51 \pm 22.6,\ 51 \pm 2(22.6),\ 51 \pm 3(22.6)$, i.e., (28.4, 73.6), (5.8, 96.2), and (–16.8, 118.8). There are (19/30)100% = 63.3%, of the data within (28.4, 73.6), (30/30)100% = 100%, of the data within (5.8, 96.2), (30/30)100% = 100%, of the data within (–16.8, 118.8).

c. The Empirical indicates that the percentages should be approximately 68%, 95%, and 100% in these intervals. It does seem plausible that this data came from a bell-shaped population.

3.56. a. $\bar{x} = \dfrac{2137}{30} = 71.23,\ s^2 = \dfrac{\sum (x_i - \bar{x})^2}{n-1} = \dfrac{9465.37}{29} = 326.39,\ s = \sqrt{s^2} = \sqrt{326.39} = 18.07$

b. Therefore, $\bar{x} \pm 2s,\ \bar{x} \pm 3s,\ \bar{x} \pm 4s$ are, respectively, $71.23 \pm 2(18.07),\ 71.23 \pm 3(18.07),\ 71.23 \pm 4(18.07)$, i.e., (35.09, 107.37), (17.02, 125.44), and (–1.05, 143.51).

c. There are (30/30)100% = 100%, of the data within ± 2 standard deviations. Tchebysheff's Theorem indicates that the percentages should be approximately at least $1 - (1/2)^2 = 0.75$, $1 - (1/3)^2 = 0.89$, and $1 - (1/4)^2 = 0.93$, within $\pm 2s \pm 3s \pm 4s$ respectively.

3.57. a. The sample mean and sample standard deviation of the effect times for each drug are shown below. Calculations were performed using Excel.

	Drug A	Drug B
Mean	234.75	270.92
Standard Deviation	13.92	19.90

b. Based on the sample means of the time each drug is effective, Drug B appears to be effective longer than Drug A.

c. Based on the standard deviation of effect time, Drug B exhibits a higher variability in effect time than Drug A.

d. The sample coefficient of variation is the ratio of the sample mean to the sample standard deviation expressed as a percentage. For Drug A, the sample coefficient of variation is $(13.92/234.75)*100\% = 5.93\%$. For Drug B the sample coefficient of variation is $(19.90/270.92)*100\% = 7.35\%$. The coefficient of variation is a measure of the relative variation in the effectiveness of the drugs. The larger the coefficient of variation the more variable the data is relative to its mean. Since Drug B has a higher coefficient of variation it has the greater relative spread.

3.58. a. The mean and standard deviation for the time to complete calls to automated system

	Mean	Standard Deviation
Automated	111.36	22.42

b. The mean and standard deviation for the time to complete calls to live service representatives was calculated using Excel and are shown below.

	Mean	Standard Deviation
Manual	150.86	26.84

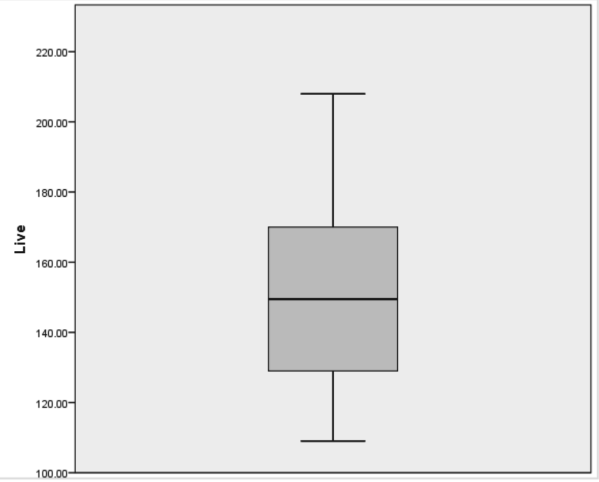
c. The coefficient of variation for the time to complete calls is shown below.

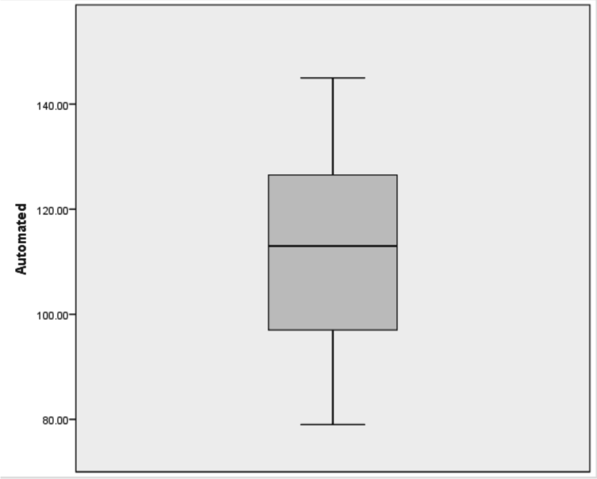
Coefficient of Variation for the automated $= (22.42/111.36)*100\% = 20.13\%$

Coefficient of Variation for the live representative $= (26.84/150.86)*100\% = 17.79\%$

Thus, the relative variation is greater for automated than for live calls.

d. A box and whisker plot for the two types of calls is shown below. Note that live calls have a higher median than automated calls. Furthermore, the third quartile for automated calls is approximately equal to the first quartile of live calls. This indicates that only about 25% of the live calls are completed within the time required to complete approximately 75% of the automated calls. The time to complete both types of calls appears to be symmetric (medians are centered in the box), however the whisker pointing up on live calls is longer than the whisker pointing down, which suggests some slight positive skewness in the time to complete live calls.





3.59. At issue is relative variability. To assess this, the proper measure is the coefficient of variation computed as follows:

For a population: $CV = \frac{\sigma}{\mu}(100)$

For a sample: $CV = \frac{s}{\overline{x}}(100)$

For the existing supplier, we treat the mean and standard deviation as population values giving:

Existing Supplier: $CV = \frac{\sigma}{\mu}(100)$

$$CV = \frac{0.078}{3.75}(100) = 2.08\%$$

New Supplier: $CV = \frac{s}{\overline{x}}(100)$

We begin by computing the mean and standard deviation from the sample data giving:

$$\overline{x} = \frac{\sum x}{n} = \frac{360.586}{20} = 18.029$$

$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{(18.018-18.029)^2 + (17.856-18.029)^2 + \ldots + (17.799-18.029)^2}{20-1}}$$

$$s = 0.135$$

Then the coefficient of variation for the new supplier is:

$$CV = \frac{s}{\overline{x}}(100)$$

$$CV = \frac{0.135}{18.029}(100) = 0.75\%$$

Student reports will differ. However, students should show the results of the coefficient of variation computations and conclude that the new supplier has the potential to produce parts with less variation than the existing supplier. The key is whether Lockheed-Martin managers believe they can effectively compare the two companies when different size products are being compared. The students should point out that the purpose of the coefficient of variation is to compare variability when the means of the two groups are different.

3.60. a. Since 2750 > 120, this would suggest that the higher income group has a much wider range of benefits bestowed than does the lower income group.

b. $CV = \frac{\sigma}{\mu}(100)$: For high income groups *CV* = (2750/8268)100 = 33; for the low income group *CV* = (120/365)100 = 33. They appear to have approximately equal relative dispersion.

3.61. The Empirical Rule can be used to determine the cut-offs for this new test. For instance, this rule states that approximately 68 percent of the data will fall within one standard deviation either side of the mean. That means that 32% will fall on either side of the mean between the mean and one standard deviation. Thus, 16 percent of the data lie outside one standard deviation from the mean. Likewise, about 95% of the data will fall within two standard deviations of the mean. Thus, 2.5 percent will fall above two standard deviations from the mean.

We start by computing the mean and standard deviation.

$$\bar{x} = \frac{\sum x}{n} = \frac{2,157}{30} = 71.9$$

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(76-71.9)^2 + (75-71.9)^2 + \ldots + (67-71.9)^2}{30-1}}$$

$$s = 10.04$$

Using the Empirical Rule:

68% within $\bar{x} \pm 1(s)$

$71.9 \pm 1(10.04)$

61.86 --------------81.94

Anyone scoring below 61.86 (rounded to 62) will be rejected without an interview.

95% within $\bar{x} \pm 2(s)$

$71.9 \pm 2(10.04)$

51.82 ----------------91.98

Anyone scoring higher than 91.98 (rounded to 92) will be sent directly to the company.

3.62. a. The coefficient of variation measures the relative standard deviation:

Public: $CV = \frac{\sigma}{\mu}(100)$ = (4500/9410)100 = 47.8.

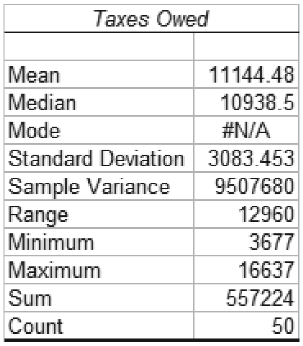
Private: $CV = \frac{\sigma}{\mu}(100)$ = (12000/32,405) = 37.0

So actually the public four-year colleges and universities have the larger relative variability.

b. The Empirical Rule suggests that virtually all of the data values are contained by $\mu \pm 3\sigma$. For public schools this range would be: $9,410 \pm 3(4500) = –\$4,090$ to \$22,910.

For private schools: –\$3,595 to \$68,405.

3.63. Student reports will vary. However, they should weave in some or all of the following statistics based on the sample data. Students will use software such as Excel and Minitab to do the computations.



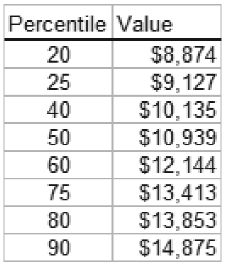
Taxes Owed	
Mean	11144.48
Median	10938.5
Mode	#N/A
Standard Deviation	3083.453
Sample Variance	9507680
Range	12960
Minimum	3677
Maximum	16637
Sum	557224
Count	50

The coefficient of variation is:

$$CV = \frac{s}{\bar{x}}(100)$$

$$CV = \frac{3{,}083.45}{11{,}144.48}(100) = 27.67\%$$

The following percentiles/quartiles have been computed:



Percentile	Value
20	\$8,874
25	\$9,127
40	\$10,135
50	\$10,939
60	\$12,144
75	\$13,413
80	\$13,853
90	\$14,875

Based on Tchebysheff's theorem, we know that at least 75 percent of CPA firms will compute a tax owed between:

$$\bar{x} \pm 2(s)$$

$$\$11{,}144.48 \pm 2(\$3{,}083.45)$$

\$4,977.58 ----------- \$17,311.38

The key is that we would expect all CPA firms to arrive at basically the same taxes owed. The fact that there is such a large variation by these experts exemplifies the difficulty that individuals have with the tax code.

3.64. a. The computer calculated statistics are $\bar{x} = 1.6848$ and $s = 0.00383$.

b. Except for random variation, the diameter of the golf balls should not be smaller than 1.682. The larger the diameter the shorter the distance when driving a golf ball. So the data should be bunched around 1.682 but not smaller. This would create a right skewed distribution where as the bell shaped distribution is symmetric.

c. $\bar{x} \pm 2s = 1.6848 \pm 2(0.00383) = (1.6771,\ 1.6925)$ contained 43 of the data points = 96% versus Tchebysheff's stipulation, at least $(1 - 1/2^2)100 = 75\%$, $\bar{x} \pm 3s = 1.6848 \pm 3(0.00383) =$ (1.6733, 1.6963) contained all 45 implying 100% compared to at least $(1 - 1/3^2)100 = 89\%$, $\bar{x} \pm 4s = 1.6848 \pm 4(0.00383) = (1.6695,\ 1.7001)$ contained all 45 implying 100% compared to at least $(1 - 1/4^2)100 = 93.75\%$.

3.65. a. Excel functions are used to compute the mean and median

mean = 17.8%

median = 16.3%

The body fat distribution appears to be right-skewed since the mean exceeds the median.

b. The standard deviation is 9.08%

c. The coefficient of variation is 9.08/17.8(100) = 51%

d. The 90th percentile = 31.12%. Since this employee has a body fat of 29, he or she is not at or above the 90th percentile.

3.66. a. The computer calculated statistics for the LAX/SLC flight were $\bar{x} = 336$ and $s = 162.8$; for the LAX/BCN flight $\bar{x} = 1997$ and $s = 203.4$.

b. For LAX/SLC $CV = \frac{s}{\bar{x}}(100) = (162.8/336)100 = 48.5\%$;

For LAX/BCN $CV = \frac{s}{\bar{x}}(100) = \left(203.4/1997\right)100 = 10.2\%$.

The SLC flight has a larger relative dispersion.

c. The mean and the standard deviation are multiplied by the same constant as each member of the data is multiplied. Therefore, $\bar{x} = 0.566\left(1997\right) = 1130.30$ and $s = 0.566(203.4) = 115.12$. The *CV* won't change since both elements of the ratio are multiplied by the same constant.

3.67. Probably the most effective way to deal with this exercise is to covert the growth rates to standardized *z*-values. To do this, we use:

$$z = \frac{x - \mu}{\sigma}$$

Excel or Minitab can be used to calculate the *z*-values. In this case, we have used Excel's STANDARDIZE function. First, we need to compute the mean and standard deviation for the growth rate in population. This is done as follows:

Mean = 1.09

Standard Deviation = 1.42

Now we calculate the *z*-values for each country and sort the countries by *z*-value. We find 3 countries that have growth rates above two standard deviation: Lebanon, Zimbabwe, and South Sudan. Only two countries have growth rates below two standard deviations, which are negative growth: Syria and Cook Islands.

End of Chapter Exercises

3.68. a. It is not possible for the sample variance to be negative. The numerator in the calculation, $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$, is the sum of squared deviations which are all positive. The denominator is one of the positive counting integers. Therefore, there ratio must also be positive.

b. The smallest a variance can be is 0. This is realized when each of the observations are of the same value.

c. The standard deviation will be larger than the variance whenever the variance is one of the numbers in the interval 0 < variance < 1.

3.69. The first interval is $\mu \pm \sigma$. It contains 68% of the data. The remainder (100 – 68 = 32) is divided between the two tails. So a half of the area in the left hand tail is 32/2. Therefore, the lower endpoint is the 16th perecentile. 68 + (32/2) of the data is to the left of the upper endpoint. It is the 84t^{h} percentile. The second interval is $\mu \pm 2\sigma$. It contains 95% of the data. The remainder (100 – 95 = 5) is divided between the two tails. So a half of the area in the left hand tail is 5/2. Therefore, the lower endpoint is the 2.5th perecentile. 95 + (5/2) of the data is to the left of the upper endpoint. It is the 97.5th percentile. The third interval is $\mu \pm 3\sigma$. It contains 100% of the data. Therefore, the lower endpoint is the 0th perecentile. The upper endpoint is the 100th percentile.

3.70. If all of the data is contained within $\mu \pm 3\sigma$, this would indicate that the largest number is approximately $\mu + 3\sigma$; the smallest number is approximately $\mu - 3\sigma$. Thus, the range is equal to $(\mu + 3\sigma) - (\mu - 3\sigma) = 6\sigma$. Therefore, $\sigma \approx \text{R}/6$. A more liberal estimate would suggest that $\mu \pm 2\sigma$ contains almost all of the data (95%). Similar reasoning gives $\sigma \approx \text{R}/4$ as an approximation.

3.71. Some problems are that it does not look at total hours taken. One student could have taken one class on campus and got an A so would have a 4.0 grade point average. Another student could have taken many hours and got all A's except one or two B's and would have lower than a 4.0 grade point average and people might conclude that the first student is a better student than the second based only upon grade point average. It also does not look at the difficulty of the classes taken. Comparing across two universities has the same problems as mentioned previously along with the fact that all universities are different and the type of classes and difficulty level of classes will be completely different. None of this is accounted for in calculating a grade point average.

3.72. The standard deviation uses all values in the data set while the range only uses two, the extreme values. In addition, the standard deviation can be used with either the Empirical Rule or Tchebysheff's theorem to indicate the percentage of the data falling within specific distances of the mean.

3.73. The mode is a useful measure of location of a set of data if the data set is large and involves nominal or ordinal data. For example, if the Labor Department is interested in the category of employment that will generate the most new jobs over the next decade, the modal class would be important. The buyer for a large department store chair would be interested in the category of shoe size most commonly bought.

3.74. a.

X	$X-\bar{x}$	$(X-\bar{x})^2$
15	1.875	3.515625
14	0.875	0.765625
16	2.875	8.265625
14	0.875	0.765625
14	0.875	0.765625
14	0.875	0.765625
13	–0.125	0.015625
8	–5.125	26.26563
12	–1.125	1.265625
9	–4.125	17.01563
7	–6.125	37.51563
17	3.875	15.01563
10	–3.125	9.765625
15	1.875	3.515625
16	2.875	8.265625
16	2.875	8.265625
210		141.75

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 261/16 = 13.125$$

To compute the median, rank the observations and compute the average of the middle two.

7 8 9 10 12 13 14 14 14 14 15 15 16 16 16 17

Median = (14 + 14)/2 = 14

Mode = 14

Range = 17 – 7 = 10

The 1st quartile is equal to the 25th percentile

$i = \frac{p}{100}n = (25/100)(16) = 4$ so the average of the 4th and 5th values.

(10 + 12)/2 = 11

The 3rd quartile is equal to the 75th percentile

$i = \frac{p}{100}n = (75/100)(16) = 12$ so the average of the 12th and 13th values

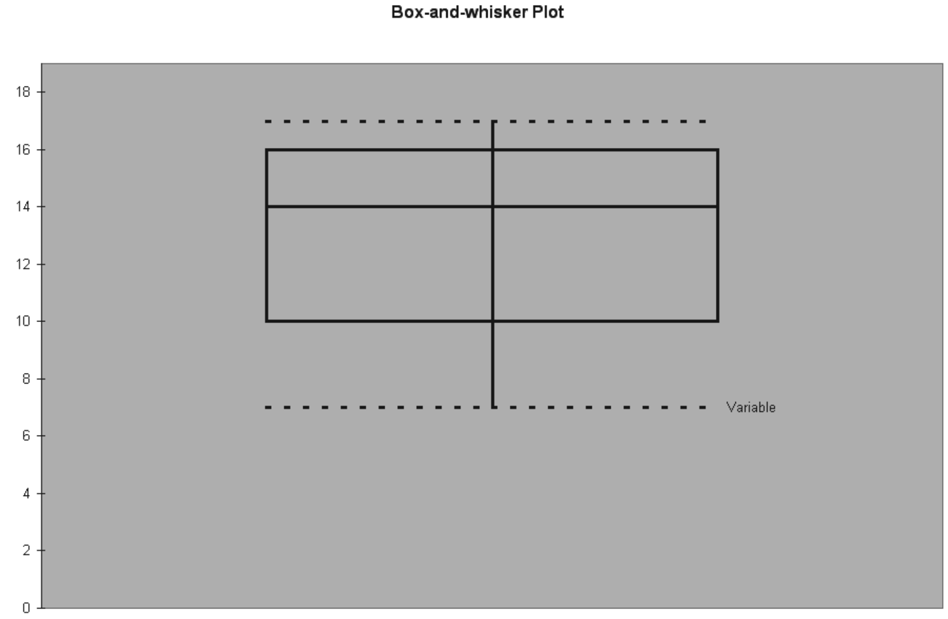
(16 + 15)/2 = 15.5

Interquartile Range = 15.5 – 11 = 4.5

$$S^2 = \frac{\sum_{i=1}^{n}(x-\overline{x})^2}{n-1} = 141.75/(16-1) = 9.45$$

$$S = \sqrt{S^2} = \sqrt{9.45} = 3.0741$$

b.



3.75. a. To determine how many standard deviations the data points are from the mean, calculate the standardized sample data: $z = \frac{x-\overline{x}}{s}$. $z = (19-28)/9 = -1$. The Empirical Rule indicates that 68% of the data is within one standard deviation from the mean. This area is half of that. So the proportion of players between 19 and 28 is 0.68/2 = 0.34.

b. To determine how many standard deviations the data points are from the mean, calculate the standardized sample data: $z = \frac{x-\overline{x}}{s}$. $z = (37-28)/9 = 1$. This is the same area as in part a. except it is on the right hand side of the mean. Thus, the proportion of players between 28 and 37 is 0.68/2 = 0.34.

c. Sixty eight percent of the players are between 19 and 37. Therefore, 100 – 68 = 32 percent are outside of this interval half of which are greater than 37. Thus, the proportion of players greater than 37 is 0.32/2 = 0.16.

3.76. a. The standardized sample data for 164 is one standard deviation from the mean. z = (164 – 218)/54 = –1. The Empirical Rule indicates that 68% of the data is within one standard deviation from the mean. One half of this is between 164 and 218 and 50% of the distribution is to the left of the mean, 218. Thus, the proportion of airfares less than 164 is 0.5 – 0.68/2 = 0.16.

b. Being a 25th percentile would mean than 0.25 of the area is to its left. The Empirical Rule says that 68% of the data is within one standard deviation of the mean. That means there is (100 – 68)/2 = 0.16 in each tail. Therefore, we need only produce a number less than one standard deviation below the mean: 218 – 1(54) = 164. So a number between 218 and 164.

c. $z = \frac{x-\overline{x}}{s} = (128-218)/54 = -1.67.$ Tchebysheff's Theorem indicates that 1 – (1/k^2)= 1 – [1/(1.67^2)] = 0.64 of the data lies within 1.67 standard deviations from the mean. Thus, the proportion beyond 1.67 standard deviations is 1 – 0.64 = 0.36. Since the distribution may not be symmetric, all of that proportion could be to the left of 128. Therefore, the largest percentile that could be attributed to an airfare of $128 is the 36st percentile.

3.77. a.

X	$X-\overline{x}$	$(X-\overline{x})^2$
229	–135.417	18,337.6736
345	–19.4167	377.0069
599	234.5833	55,029.3403
229	–135.417	18,337.6736
429	64.58333	4,171.0069
605	240.5833	57,880.3403
339	–25.4167	646.0069
339	–25.4167	646.0069
229	–135.417	18,337.6736
279	–85.4167	7,296.0069
344	–20.4167	416.8403
407	42.58333	1,813.3403
4373		183,288.9167

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 4373/12 = 364.42$$

b. $$s^2 = \frac{\sum_{i=1}^{n}(x-\overline{x})^2}{n-1} = 183,288.9167/(12-1) = 16,662.63$$

$$s = \sqrt{s^2} = \sqrt{16,662.6288} = 129.08$$

3.78.

X	$X-\overline{x}$	$(X-\overline{x})^2$
34	–5.05556	25.55864
24	–15.0556	226.6698
43	3.944444	15.55864
56	16.94444	287.1142
74	34.94444	1221.114
20	–19.0556	363.1142
19	–20.0556	402.2253
33	–6.05556	36.66975
55	15.94444	254.2253
43	3.944444	15.55864
54	14.94444	223.3364
34	–5.05556	25.55864
27	–12.0556	145.3364
34	–5.05556	25.55864
36	–3.05556	9.33642
24	–15.0556	226.6698
54	14.94444	223.3364
39	–0.05556	0.003086
703		3726.944

a. $\overline{x}=\frac{\sum_{i=1}^{n}x_i}{n}=703/18=39.0556$

b. To compute the median, rank the observations and average the middle two values.

19 20 24 24 27 33 34 34 34 36 39 43 43 54 54 55 56 74

Median = (34 + 36)/2 = 35

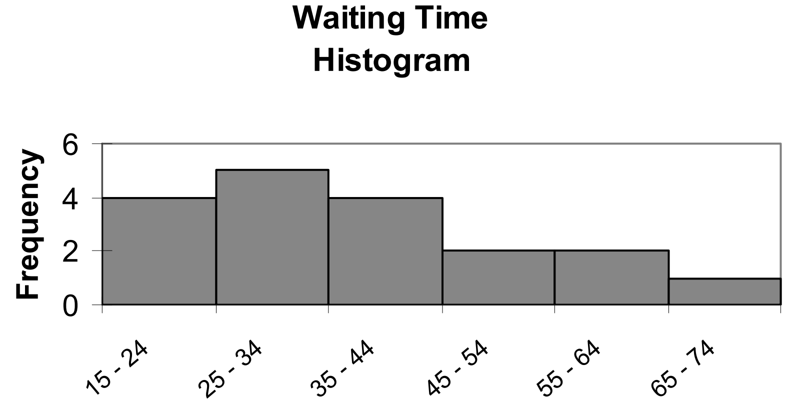
c. $S^2=\frac{\sum_{i=1}^{n}(x-\overline{x})^2}{n-1}=3726.944/(18-1)=219.232$

$S=\sqrt{S^2}=\sqrt{219.232}=14.8065$

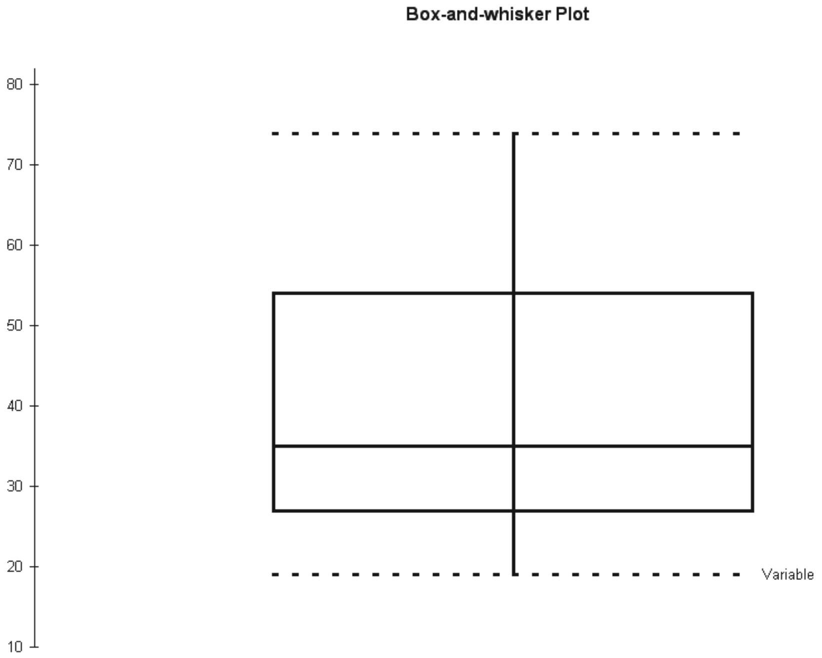
d. Use Excel’s histogram feature to create the frequency distribution.

Classes	Frequency
15–24	4
25–34	5
35–44	4
45–54	2
55–64	2
65–74	1

e. Use Excel's histogram feature to create the histogram.



f.



g. The 3rd quartile is equal to the 75th percentile

$$i = \frac{p}{100}n = (75/100)(18) = 13.25;$$

Thus, the Q3 value is 25 percent of the distance between the 13th (43) and the 14th (54) values. This is 45.75.

The minimum number of minutes the customer would have to wait is over 54 minutes

3.79. Student answers will vary but one approach would be to standardize the results for each manager

Plant 1: (810 – 700)/200 = .55 standard deviations

Plant 2: (2600 – 2300)/350 = .86 standard deviations

Plant 3: (1320 – 1200)/30 = 4 standard deviations

Based upon this the manager of Plant 3 performed far better than the other plants on a relative basis.

3.80. a. You need to calculate the coefficient of variation on each type of mutual fund.

Growth Fund:

$$CV = \frac{S}{\bar{x}}(100) = (2/8)(100) = 25\%$$

Specialized Fund:

$$CV = \frac{S}{\bar{x}}(100) = (6/18)(100) = 33.3\%$$

The specialized fund is more variable.

b. If an investor is risk averse he would invest in the less variable fund so would invest in the Growth Fund.

c. Using the empirical rule you could determine what you would expect the range of returns to be 95% of the time.

Growth Fund:

$$8 \pm 2(2)$$

$$4 - 12$$

Specialized Fund:

$$18 \pm 2(6)$$

$$6 - 30$$

The Specialized Fund appears to be the best investment.

3.81. a. Comparing only the mean bushels/acre you would say that Seed Type C produces the greatest average yield per acre. Student answers will vary but may include things such as making sure soil type is the same, make sure watering and fertilizing is the same, etc.

b. Students need to calculate the coefficient of variation for each Seed Type.

CV of Seed Type A = 25/88 = 0.2841 or 28.41%

CV of Seed Type B = 15/56 = 0.2679 or 26.79%

CV of Seed Type C = 16/100 = 0.1600 or 16%

Seed Type C shows the least relative variability.

c. Seed Type A

Approximately 68% will be within 1 standard deviation

$$88 \pm 25 = 63 \text{ to } 113$$

Approximately 95% will be within 2 standard deviations

$$88 \pm 2(25) = 38 \text{ to } 138$$

Approximately 100% will be within 3 standard deviations

$$88 \pm 3(25) = 13 \text{ to } 163$$

Seed Type B

Approximately 68% will be within 1 standard deviation

$$56 \pm 15 = 41 \text{ to } 71$$

Approximately 95% will be within 2 standard deviations

$$56 \pm 2(15) = 26 \text{ to } 86$$

Approximately 100% will be within 3 standard deviations

$$56 \pm 3(15) = 11 \text{ to } 101$$

Seed Type C

Approximately 68% will be within 1 standard deviation

$$100 \pm 16 = 84 \text{ to } 116$$

Approximately 95% will be within 2 standard deviations

$$100 \pm 2(16) = 68 \text{ to } 132$$

Approximately 100% will be within 3 standard deviations

$$100 \pm 3(16) = 52 \text{ to } 148$$

d. Student answers will vary but students should say Seed Type A because the 135 is within 2 standard deviations. Since it has higher variability there is a greater chance that it will produce135.

e. Seed type C because 115 is included within one standard deviation.

3.82. a. $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 127/8 = 15.875$

b. To compute the median, rank the observations and compute the average of the middle two.

9 12 12 13 16 16 17 32

Median = (13 + 16)/2 = 14.5

c. The mode is 12 and 16

d.

X	$X-\bar{x}$	$(X-\bar{x})^2$
13	–2.875	8.265625
32	16.125	260.0156
12	–3.875	15.01563
9	–6.875	47.26563
16	0.125	0.015625
17	1.125	1.265625
16	0.125	0.015625
12	–3.875	15.01563
127		346.875

$$S^2 = \frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n-1} = 346.875/(8-1) = 49.5536$$

$$S = \sqrt{S^2} = \sqrt{49.5536} = 7.0394$$

e. The extreme value does not affect the median or the mode. It does, however, cause the mean to increase because it gives a higher number to the sum of the numbers which causes the mean to be higher.

f. In this case the median might be a better measure since you have an extreme outlier. This would give you a better representation of future rounds of golf.

g. The top quartile would be above 75% so you would need to calculate the 75th percentile to determine the minimum number of rounds.

The 3rd quartile is equal to the 75th percentile

$$i = \frac{p}{100}n = (75/100)(8) = 6;$$

Thus, the Q_3 value is the average of the 6th (16)and 7th (17) values after the data have been arranged in numerical order.

$$Q_3 = (17 + 16)/2 = 16.5.$$

3.83. a. The number of standard deviations \$60 is from \$48.56 is $z = \frac{x - \overline{x}}{s} = \frac{60 - 48.56}{10} = 1.14.$

Tchbeyscheff's Theorem indicates that at least $1 - (1/k^2)$ of the data is contained within k standard deviations from the mean. Thus, at most $1/k^2 = 1/(1.14)^2 = 0.769$ will fall above or below 1.14 standard deviations from the mean. This proportion could all be in the right hand tail. Thus, the largest proportion of meal prices larger than \$60 is 0.769.

b. If each number is multiplied by a constant the mean and standard deviation are also multiplied by that constant. Therefore, mean = 6.46(48.56) = 313.7 CNY. Since the standard deviation is in the denominator and also has its values multiplied by 6.46 giving 64.6, there would be no change to the answer to part a.

3.84. a. Excel functions are used to calculate the required statistics.

Mean Age = 34.84 years

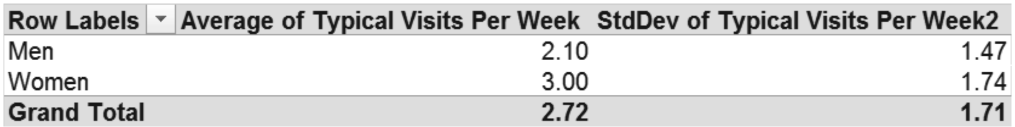
Median Age = 33 years

Based on these calculations, it appears that the age data are right skewed since the mean exceeds the median.

b. The standard deviation = 11.18 years

c. The coefficient of variation for age = $\text{st. dev./mean}*100 = 11.8/34.84*100 = 33.87\%$

d. The statistics computed by gender for typical visits per week are:

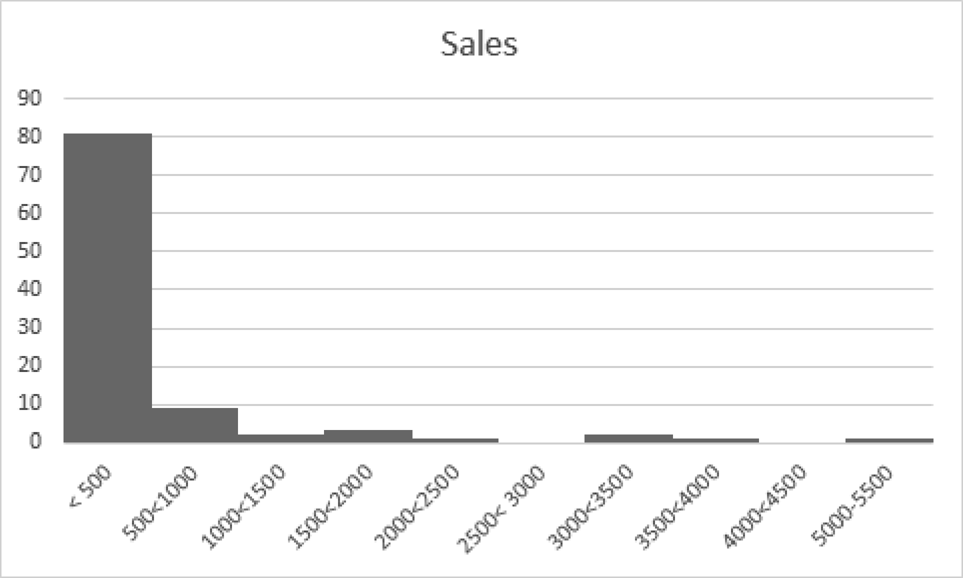


Row Labels	Average of Typical Visits Per Week	StdDev of Typical Visits Per Week2
Men	2.10	1.47
Women	3.00	1.74
Grand Total	**2.72**	**1.71**

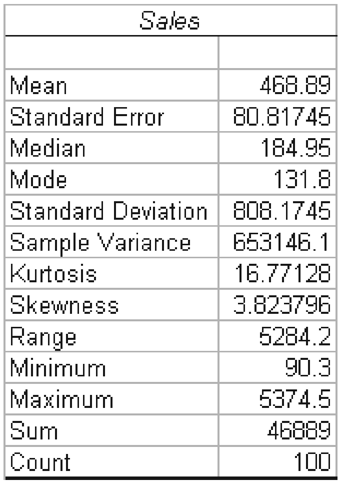
Based on these sample data, it appears that women tend to visit more frequently on average than do men.

3.85. Students should use Excel to answer this question.

a. Note, the students are free to select their own class limits. Below is one example. Students might consider having an open-end class on the upper end to better convey the data.

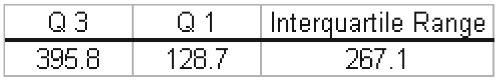


b.



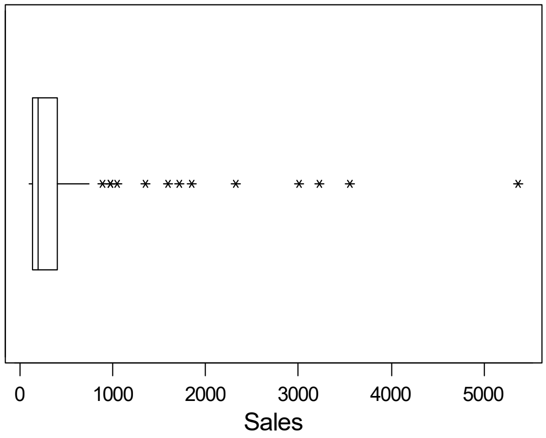
Sales	
Mean	468.89
Standard Error	80.81745
Median	184.95
Mode	131.8
Standard Deviation	808.1745
Sample Variance	653146.1
Kurtosis	16.77128
Skewness	3.823796
Range	5284.2
Minimum	90.3
Maximum	5374.5
Sum	46889
Count	100

c. Using Excel's QUARTILE function

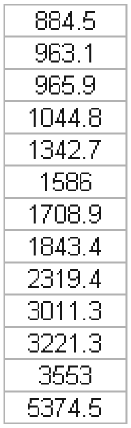


Q 3	Q 1	Interquartile Range
395.8	128.7	267.1

d.

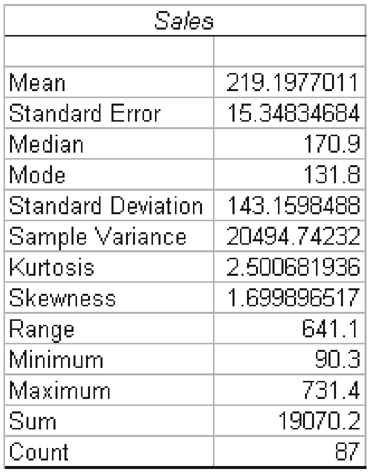


The upper limit is 395.8 + 1.5(267.1) = 796.45. Any value exceeding 796.45 will be considered an outlier. The following sales values are outliers:



884.5
963.1
965.9
1044.8
1342.7
1586
1708.9
1843.4
2319.4
3011.3
3221.3
3553
5374.5

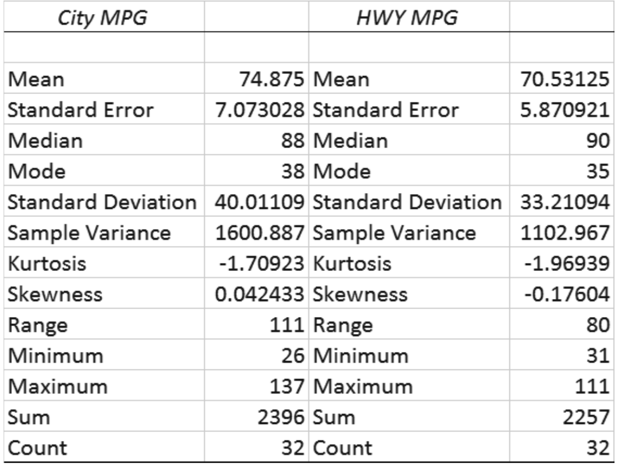
Deleting these 13 sales value, gives the following revised descriptive statistics:



Sales	
Mean	219.1977011
Standard Error	15.34834684
Median	170.9
Mode	131.8
Standard Deviation	143.1598488
Sample Variance	20494.74232
Kurtosis	2.500681936
Skewness	1.699896517
Range	641.1
Minimum	90.3
Maximum	731.4
Sum	19070.2
Count	87

e. Using Excel's PERCENTILE function, the 65th percentile is found using Excel to be $246.20

3.86. a. Use Excel's Descriptive Statistics tool to determine the sample mean and sample standard deviation for the two data sets. The results are shown below:



City MPG		*HWY MPG*	
Mean	74.875	Mean	70.53125
Standard Error	7.073028	Standard Error	5.870921
Median	88	Median	90
Mode	38	Mode	35
Standard Deviation	40.01109	Standard Deviation	33.21094
Sample Variance	1600.887	Sample Variance	1102.967
Kurtosis	-1.70923	Kurtosis	-1.96939
Skewness	0.042433	Skewness	-0.17604
Range	111	Range	80
Minimum	26	Minimum	31
Maximum	137	Maximum	111
Sum	2396	Sum	2257
Count	32	Count	32

No, the data do not support the premise that cars will get better mileage on the highway than around town. The mean for highway (70.53) is lower than the mean for city (74.87). This is a phenomena of electric cars.

b. To answer this question, calculate the coefficient of variation for each variable.

$\text{Highway CV} = 5.87 / 33,21 * 100 = 17.68\%$

City CV = 7.07/40.01 = 17.7%

City and highway speeds have the same relative variability.

3.87. a. Mean = 54.00

Standard Deviation = 3.813

b. $\bar{x} \pm 1s = 54 \pm (3.813) = (50.187,\ 57.813)$, $\bar{x} \pm 2s = (46.374,\ 61.626)$

$\bar{x} \pm 3s = (42.561,\ 65.439)$

c. The Empirical Rule indicates that 95% of the data is contained within $\bar{x} \pm 2s$. This would mean that each tail has (1 – 0.95)/2 = 0.025 of the data. Therefore, the costume should be priced at $46.37.

3.88. a.

```
Variable       StDev
California     7796
Florida        5291
```

On the basis of the standard deviations, it appears that California has the widest spectrum of salaries for MBA graduates.

b.

```
Variable       Mean     Median
California     91215    90497
Florida        68821    68403
```

c.

```
Variable       CoefVar
California     8.55
Florida        7.69
```

Based on the coefficient of variation California still has the widest spectrum of salaries for MBA graduates.

3.89. a. Mean = –.1303 Median = –.09 St. Dev. = .262

b. It means that the closing price for GE stock is an average of approximately thirteen ($0.1303) cents higher than the opening price.

3.90. Use Excel's average and standard deviation functions to determine the mean and standard deviation of each type of bread. The results are shown below.

a. White Bread has the highest average daily demand.

Bread Type	Average
White	599.77273
Wheat	530.40909
Multigrain	470.36364
Black	383.59091
Cinnamon Raisin	139.72727
Sour Dough French	127.09091
Light Oat	261.63636

b. Use Excel's Histogram feature to develop the frequency distribution for each bread type. Student answers will vary depending on number of classes selected and class widths used, but shown below are the results if you let Excel set up the bins.

White Bread	Frequency
251 – 375	1
376 – 500	5
501 – 625	7
626 – 750	5
751 – 875	4

Cinnamon Raison	Frequency
54.76 – 84.00	1
84.01 – 113.25	4
113.26 – 142.50	8
142.51 – 171.75	4
171.76 – 201.00	5

Wheat Bread	Frequency
264.26 – 352.00	1
352.01 – 439.75	3
439.76 – 527.50	8
527.51 – 615.25	4
615.26 – 703.00	6

Sour Dough French	Frequency
64 – 88	1
89 – 113	8
114 – 138	4
139 – 163	7
164 – 188	2

Multigrain Bread	Frequency
212.76 – 299.00	1
299.01 – 385.25	4
385.26 – 471.50	8
471.51 – 557.75	2
557.76 – 644.00	7

Light Oat Bread	Frequency
127 – 172	2
173 – 218	3
219 – 264	8
265 – 310	4
311 – 356	5

Black Bread	Frequency
182.76 – 256.00	1
256.01 – 329.25	5
329.26 – 402.50	7
402.51 – 475.75	6
475.76 – 549.00	3

c. White Bread has the highest standard deviation.

Bread Type	Standard Deviation
White	149.0550975
Wheat	107.0236552
Multigrain	108.0130263
Black	81.9510069
Cinnamon Raisin	33.01973698
Sour Dough French	28.68367561
Light Oat	57.89526295

d. Use Excel to calculate the coefficient of variation for each bread type. White bread has the greatest relative variability and wheat bread has the lowest relative variability.

Bread Type	Coefficient of Variation
White	24.85%
Wheat	20.18%
Multigrain	22.96%
Black	21.36%
Cinnamon Raisin	23.63%
Sour Dough French	22.57%
Light Oat	22.13%

e. Use Tchebysheff's Theorem to calculate the upper range of two standard deviations from the mean. You must use Tchebysheff's Theorem because you do not know if the data is bell-shaped.

Bread Type	Required Loaves
White	897.8829
Wheat	744.4564
Multigrain	686.3897
Black	547.4929
Cinnamon Raisin	205.7667
Sour Dough French	184.4583
Light Oat	377.4269

f. Use Excel's Pivot Table feature to calculate the average total loaves by day of week. The highest average is on day 6.

Average of Total Loaves Sold	
Day of Week	Total
1	2196
2	2947
3	2388
4	2335.5
5	2336.8
6	3116.5
7	1772
Grand Total	2512.590909

3.91. Students should use Excel or Minitab to answer questions a–c.

a.

CPA Firm	Taxes Owed	Difference
1	$16,637	–$5,077
2	$11,804	–$244
3	$8,915	$2,645
4	$9,915	$1,645
5	$14,787	–$3,227
6	$11,058	$502
7	$15,662	–$4,102
8	$13,293	–$1,733
9	$15,970	–$4,410
10	$9,103	$2,457
11	$13,223	–$1,663
12	$7,852	$3,708
13	$9,200	$2,360
14	$13,607	–$2,047
15	$13,793	–$2,233
16	$9,048	$2,512
17	$14,487	–$2,927
18	$9,409	$2,151
19	$13,342	–$1,782
20	$14,093	–$2,533
21	$12,836	–$1,276
22	$9,376	$2,184
23	$10,819	$741
24	$10,473	$1,087
25	$3,677	$7,883
26	$6,087	$5,473
27	$8,711	$2,849
28	$9,753	$1,807
29	$10,282	$1,278
30	$13,385	–$1,825
31	$11,326	$234
32	$16,183	–$4,623
33	$14,232	–$2,672
34	$8,482	$3,078
35	$16,274	–$4,714
36	$12,758	–$1,198
37	$9,411	$2,149
38	$14,632	–$3,072
39	$12,655	–$1,095
40	$6,403	$5,157
41	$11,260	$300
42	$8,478	$3,082
43	$11,586	–$26
44	$10,299	$1,261
45	$7,805	$3,755
46	$13,422	–$1,862
47	$10,628	$932
48	$9,300	$2,260
49	$5,429	$6,131
50	$6,064	$5,496

b.

Classes	Frequency
–6928.42 to –5077.00	1
–5076.99 to –3225.57	5
–3225.56 to –1374.14	11
–1374.13 to 477.29	7
477.30 to 2328.72	12
2328.73 to 4180.15	9
4180.16 to 6031.58	3
6031.59 to 7883.01	2

c.

Difference	
Mean	415.52
Standard Error	436.0660531
Median	621.5
Mode	#N/A
Standard Deviation	3083.452632
Sample Variance	9507680.132
Kurtosis	–0.501771097
Skewness	0.177525892
Range	12960
Minimum	–5077
Maximum	7883
Sum	20776
Count	50

d. The value of $11,560 would be between the 28th ($11,326) and 29th ($11,586) observation. This is found by sorting the values for the "tax owed" variable. Since $11,500 is closer to the 29th value than it is to the 28th value, we will use $i = 29$ and solve for the percentile as follows:

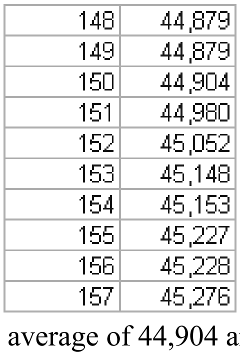
$$i = \frac{p}{100}n \text{ so } 29 = (p/100)(50) \qquad p = 58 \text{ or the 58th percentile.}$$

This shows that 58% of the tax consultants in this study showed less tax owed than did the IRS.

3.92. a. Sorting the data and determine what position 45,000 is in you can solve the percentile equation for the percentile.

$$i = \frac{p}{100}n = (75/100) * (200) = 150;$$

The 75th percentile is the average of the 150th and 151st value in the data. You can use PHStat's Stack feature under Data Preparation to reorganize the data. Then sorting the data you get the following:



148	44,879
149	44,879
150	44,904
151	44,980
152	45,052
153	45,148
154	45,153
155	45,227
156	45,228
157	45,276

Thus, the 75th percentile is the average of 44,904 and 44,980, or 44,942.

b. Using Excel's Average and Median functions you can find that

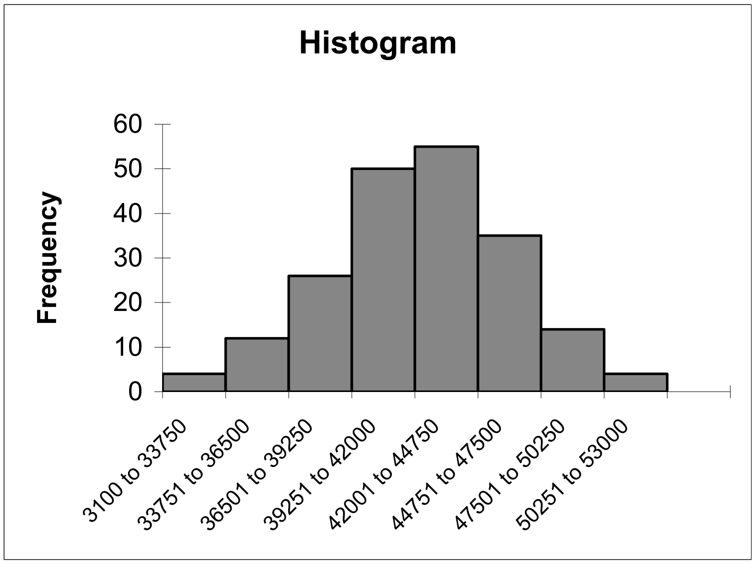
Mean = 42,261

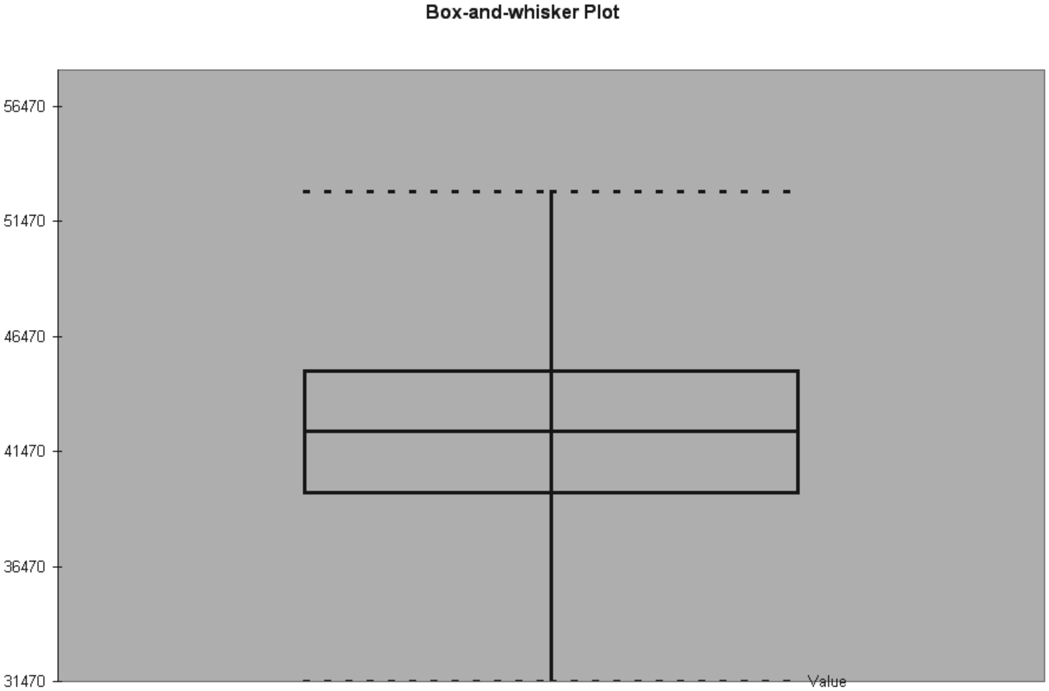
Median = 42,326

c. Using the $2^k \geq n$ guideline, the appropriate number of classes should be 8 since $2^8 = 256$: The class width is:

$$w = \frac{52,774 - 31,476}{8} = 2662.25$$

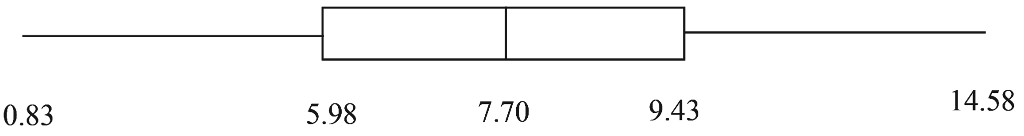
We choose to round this to 2,750 and start at 31,000 giving the following histogram:



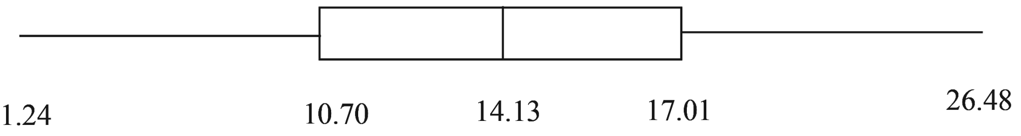


Histograms and box and whiskers plots have certain things in common. In both instances, we get an idea of how the data are distributed, where the center is, and what the shape of the distribution is and how spread out the data are. The histogram breaks the data down into classes and illustrates the actual number of values in each class where the box and whiskers plot shows the median and the interquartile range.

3.93. a. Drive Thru Box and Whisker Plot



Inside



b. Both variables are quite symmetric since Median falls about midway between Q3 and Q1 and the whiskers are about same length on each side.

c. No, The Inside sales have a median = $14.13 which is almost twice that of the Drive Thru which is $7.70.

d. Yes, The Inside sales distribution appears to be more variable since the Q3 – Q1 is greater for Inside than for Drive Thru.

e. None of the Drive-Thru sales are extreme. There is one Inside sales value that was $26.42 which was right at the end of the upper whisker for that distribution.

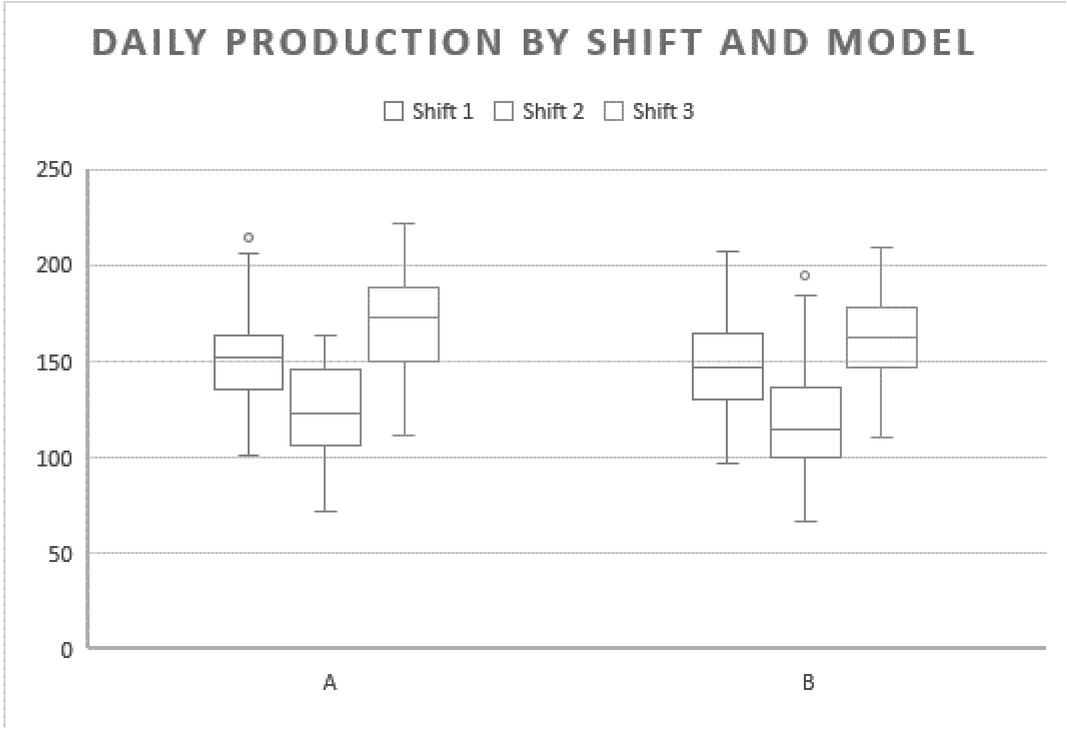
f. Drive-Thru

Min = $3.33 Q1 = $5.98 Median = $7.70 Q3 = $9.43 Max = $13.11

Inside

Min = $5.25 Q1 = $10.70 Median = $14.13 Q3 = $17.01 Max = $26.42

3.94. a.



b. There does appear to be a difference in shifts with Shift 3 having the higher median output regardless of product. Shift 2 has a lower median output than the other shifts regardless of product. Median output for product A is higher than for product B for every shift.

c. Shifts 1 and 3 seem to be fairly symmetric in both products, although an outlier is evident in Shift 1, Product A. Shift 2 seems to be left-skewed for Product A and right-skewed for Product B as denoted by the whisker lengths.

d. Shift 1 for product A has an outlier and Shift 2 for Product B has an outlier.