SOLUTIONS: Chapter 3—Displaying and Describing Quantitative Data

1. **Statistics in business.** Answers will vary.

2. **Statistics in business, part 2.** Answers will vary.

3. **University tuition.** Shape—the distribution is approximately symmetric with a single peak, making it unimodal. Centre—the centre of any distribution not perfectly symmetric is best represented by the median, the exact middle data point when the data set is ordered either in ascending or descending order. For a data set that is approximately symmetric with a clear peak, the centre can be identified visually as being in the interval represented by the peak, in this case, the interval for $5000–$6000. The distribution is centred in the interval between $5000 and $6000 so it can be approximated at $5500. Spread—the spread is determined from the range of data, low to high, or $8000–$2000 = approximately $6000. The exact range cannot be determined from the histogram because the intervals or bins do not represent the exact data points. There are no outliers or other unusual features in this distribution. It can be pointed out that most of the tuitions lie between $4000 and $6000, which includes 50 out of 66 institutions.

4. **Gas prices.** Shape—the distribution is bimodal with a peak around 130 and another at 140-145. Centre—the centre of any distribution not perfectly symmetric is best represented by the median, the exact middle data point when the data set is ordered either in ascending or descending order. There are 36 data points representing the 36 monthly prices. The centre of the data would be the average of the 18th and 19th data points (18 data points on either side of the median). It can be estimated which bin contains the centre value by adding up the values in each bin. For example, the first bin ($1.00 to $1.05) contains 1 data point. The next bin ($1.05 to $1.10) contains no values and the next bin ($1.10 to $1.15) contains 1 values. The total so far is 1+0+1 values = 2. The 18th data point has not yet been reached. Continue accumulating until the $1.25 to $1.30 bin, which gets the total up to 14 values. The 18th and 19th values will be in the next bin, so the median is in the bin ($1.30 to $1.35) at about $1.33or $1.34. Spread—the spread is determined from the range of data, low to high, or $1.55–$1.00 = approximately $0.55. The exact range cannot be determined from the histogram because the intervals or bins do not represent the exact data points. There is one outlier on the low end and a gap between it and the remaining data points.

5. **Credit card charges.**

   a. Shape—the distribution is clearly skewed to the right. Centre—it is more difficult to determine visually the centre of a skewed distribution. The centre of a skewed distribution is best represented by the median, the exact middle data point when the data set is ordered either in ascending or descending order. There are 5000 data points representing the 5000 charge customers. The centre of the data would be just to the right of the 2500th data point. It can be estimated which bin contains the median value by adding up the values in each bin. For example, the first bin (-$1000 to -$500) contains about 10 data points. The next bin ((-$500 to $0) contains slightly more, about 15 data points (the exact number is not important in this estimation of the median value) and the next bin ($0 to $5000) contains about 810 values. The next bin ($5000 to $10,000) contains about 720 values. The next bin ($1000 to $1500) contains about 830 values. The total so far is 10+15+810+720+830 = 2385 values. The 2500th data point has not yet been reached. The $1500 to $2000 bin contains a large number of values (about 750 values) which means the 2500th value is contained in that interval. Therefore, the centre of the distribution is estimated to be between $1500 and $2000. Spread—the spread is determined from the range of data, low to high, or $5000–(–$1000) = approximately $6000. The exact range cannot be determined from the histogram because the intervals or bins do not represent the exact data points. There are no outliers. Unusual features—it can be pointed out that there are a few negative values that represent customers that received more credits than charges in the month; therefore, the charge shows up as negative on the histogram.

   b. The mean will be larger than the median because the distribution is right skewed. The median represents the exact middle number whereas the mean is an average of all data points, including the data points with higher values represented by the right tail or right skewed shape. The mean will be pulled toward the tail with the higher values. The median is always the centre value whether the distribution is symmetric or skewed.

c. The median is a more appropriate measure of the centre of the distribution because it represents the middle or typical value. The mean has been pulled toward the right or a higher value due to the skewed shape and therefore is not an accurate representation of the centre.

6. Vineyards.

a. Shape—the distribution is skewed to the right. Centre—the centre of a skewed distribution is best represented by the median, the exact middle data point when the data set is ordered either in ascending or descending order. There are 36 data points representing the 36 wineries. The centre of the data would be close to the 18th value (between the 18th and 19th values). It can be estimated which bin contains the median value by adding up the values in each bin. The 0–30 acre bin contains 15 data points and the next bin (30–60 acres) contains about 13 data points. The total so far is 15+30 values = 45. The median value has to be contained in the 30 to 60 bin because that interval contains the 16th through 30th data points. Spread—the spread is determined from the range of data, high value minus low value, or 240–0 = approximately 240. Unusual features—there is a possible outlier in the 240–270 acre interval.

b. The mean is expected to be larger in a right skewed distribution because the larger values affect the mean, pulling it towards the right tail of the distribution.

c. Because the distribution is skewed, the median is a better representation of the centre of the distribution.

7. Mutual funds. Shape—the distribution is approximately symmetric with a single peak, making it unimodal. Centre—the centre of the distribution is close to the single peak but to be sure, the exact number of values and the median should be determined. The bin values can be added up as follows: 1+3+0+7+18+14+8+4+3+2+2+1+1 = 64. The median occurs to the right of the 32nd data point. Twenty-nine data points are contained in the first five bins. The 32nd data point is therefore contained in the 6th bin representing approximately 10%. Spread—the spread is determined from the range of data, high minus low, or 90%–(–15%) = approximately 105%. The two highest bins at 80% and 85% contain outliers.

8. Car discounts. Shape—the distribution has two peaks, making it bimodal. One mode occurs near $800 and the other mode occurs near $1600. The two modes could be due to that fact that men and women receive different discounts that are mixed together in this data set. To examine the differences, separate histograms should be created for female discounts and male discounts. Centre—the centre of the distribution is the median value which would be just past the 50th data point. The bin values can be added up as follows: 2+3+7+11+14+8 = 45. Forty-five data points are contained in the first six bins. The median occurs in the next bin ($1200), which contains 11 data points. The 50th data point is therefore contained in the bin representing $1200–$1400. For a bimodal distribution, the centre is usually between the two peaks. Spread—the spread is determined from the range of data, high minus low, or $2400–0 = approximately $2400. There are no outliers or other unusual features.

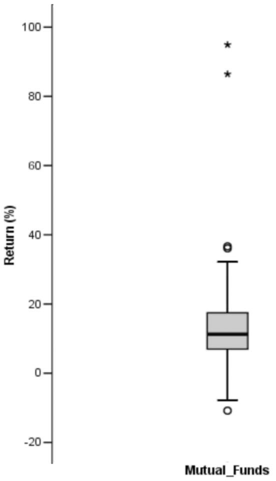
9. Mutual funds, part 2.

a. Five-Number Summary (Quartile calculations may differ slightly when using different software.)

Min	1st Qtr	Median	3rd Qtr	Max
–10.820	6.965	11.275	17.440	94.940

b. Centre can be represented by the median, which is 11.275%. The spread can be represented by the maximum—minimum values, which is 94.940–(–10.820) = 105.76%. The IQR (3rd quartile–1st quartile values) summarizes the spread by focusing on the middle 50% of the data. For this data set, the IQR = 17.440–6.965 = 10.475%. Another measure of spread is the standard deviation but this measure is reserved for symmetric distributions. The shape of the distribution can be determined from the histogram in Exercise 7. The main distribution is fairly symmetric but there are high outliers, meaning that it is not appropriate to use standard deviation as a measure of spread. In addition, due to the outliers, it is also not accurate to represent the data by a simple calculation of maximum—minimum values. For this distribution, it is better to represent the spread using the IQR value.

c. Boxplot



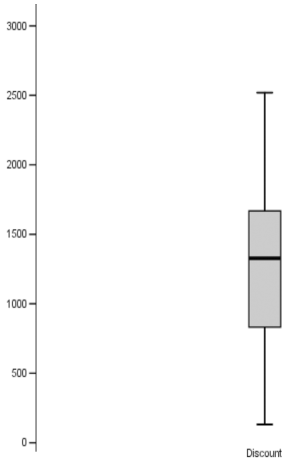
d. The histogram clearly shows the outlier values as well as the skewness. Depending on the software package used to create a boxplot, the outliers may be identified using special symbols as shown above.

10. Car discounts, part 2.

a. Five-Number Summary (Quartile calculations may differ slightly when using different software.)

Min	1st Qtr	Median	3rd Qtr	Max
$131.00	$831.50	$1327.50	$1668.00	$2520.00

b. Boxplot



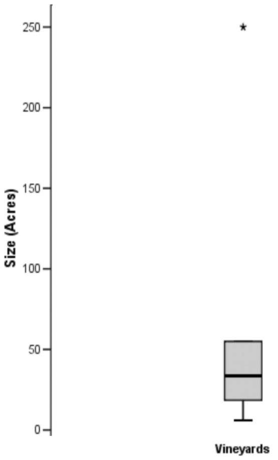
c. The histogram shows that the distribution is bimodal (2 peaks).

11. Vineyards, part 2.

a. The distribution can be described as skewed to the right. Symmetry can be determined by comparing the mean and the median. The mean is 46.50 and the median is 33.50. The mean is much larger than the median indicating a right skew (the higher values are pulling the mean value higher than the median). In addition, symmetry can be determined by comparing the difference between the first quartile and the median and the third quartile and the median. If the distribution is symmetric, these values should be fairly equal. In this summary, the median–Q1 = 33.50–8.50 = 15 compared to Q3–median = 55–33.50 = 21.5. The right side of the distribution is wider than the left which indicates a right skew. Finally, the maximum value of 250 is very high compared to the median of 33.50 while the minimum value of 6 compared to the median of 33.50 is a much smaller number also indicating a skew to the right.

b. Yes, there is one high outlier at 250. This is clearly shown in the histogram from Exercise 6.

c. The boxplot shows the outlier at 250, but without the data set the length of the upper whisker going to the upper fence (limit for the outliers) cannot be determined.

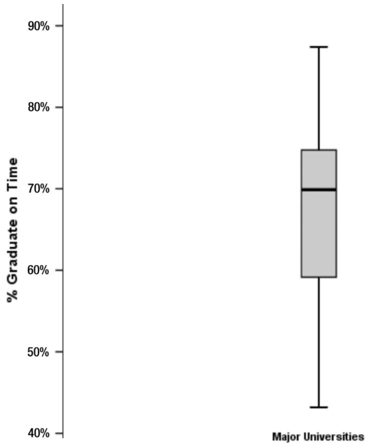


12. **Graduation.**

   a. The distribution can be described at least as roughly symmetric. Symmetry can be determined by comparing the mean and the median. The mean is 68.35 and the median is 69.90. These values are fairly close, indicating at least a roughly symmetric distribution. In addition, symmetry can be determined by comparing the difference between the first quartile and the median and the third quartile and the median. If the distribution is symmetric, these values should be fairly equal. In this summary, the median–Q1 = 69.90–59.15 = 10.75 compared to Q3–median = 74.75–69.90 = 4.85. The left side of the distribution is slightly wider than the right but not by a large margin. Finally, the maximum value of 87.40 compared to the median of 69.90 is similar to the minimum value of 43.20 compared to the median of 69.90 although again the left side is wider.

   b. Outliers can be determined mathematically by adding the term 1.5*IQR to Q3 to see if any data values fall above or by subtracting 1.5*IQR from Q1 to see if any data values fall below. In this case, IQR = Q3–Q1 = 74.75–59.15 = 15.6. 1.5*IQR = 1.5*15.6 = 23.4. To check for high outliers, add 1.5*IQR to Q3 = 23.4 + 74.75 = 98.15. The maximum data point falls below that value so there are no high outliers. To check for low outliers, subtract 1.5*IQR from Q1 = 59.15–23.4 = 35.75. The minimum data point is above this value so there are no low outliers.

   c. Boxplot



13. **Vineyards, again.** The stem and leaf plot has intervals of 20 acres (the 0 interval includes single digits and teens up to but not including 20). The distribution is shown to be clearly right skewed as seen before in the boxplot and histogram. Most of the data points end in either 0 or 5 indicating that the measurements may have been rounded.

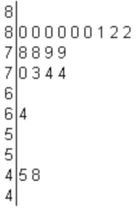
```
24|0
22|
20|
18|
16|
14|0
12|0
10|0
 8|0
 6|0 2 9
 4|0 0 5 3 5 5
 2|0 2 5 7 8 9 0 1 2 5 5 6 8
 0|6 8 0 0 0 0 1 5 7
```

14. Gas prices, again. The stem-and-leaf display has intervals of five cents, since each stem is split into two parts. For example, the teens are split into the low teens ($1.10–$1.14) and the high teens ($1.15–$1.19). The distribution looks similar to the histogram in Exercise 4.

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

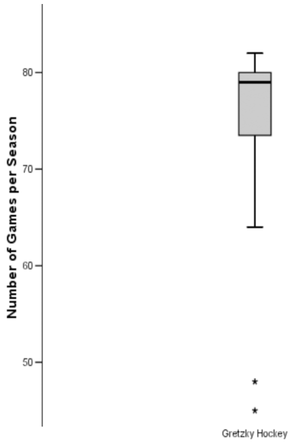
15. Hockey.

a. Stemplot: The stem and leaf plot has split stems representing 0–4 and 5–9.



```
8|
8|000000122
7|8899
7|0344
6|
6|4
5|
5|
4|58
4|
```

b. Boxplot:

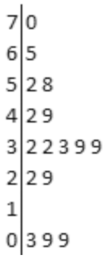


c. The distribution of the number of games played per season by Wayne Gretzky is skewed to the left toward lower values and has low outliers. The median value is 79 games and the range is 37 games.

d. The outliers identified at 45 and 48 games could have been caused by injuries. The season with 64 games has a gap to both lower and higher values. Most of his seasons were played with games totalling in the 70s and 80s.

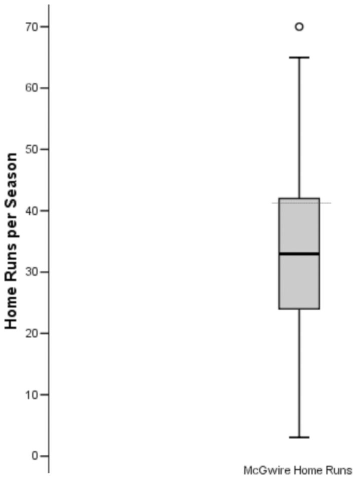
16. Baseball: Mark McGwire

a. Stemplot**:**



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

b. Boxplot**:**



c. The distribution of Mark McGwire's home runs is somewhat symmetric with typical season home runs in the 30s. McGwire had three seasons when he hit fewer than 10 home runs. With the exception of those values, the total number of home runs per season was between 22 and 70. The season with 70 home runs is identified as an outlier; however, it is only slightly higher than the season with 65 home runs.

d. The three low values are not officially identified as outliers but they can be considered to be unusual compared to the productivity of the rest of the seasons. Injury may have caused the low number of home runs.

17. Gretzky returns.

a. The distribution is skewed therefore the median is used to describe the centre.

b. The mean should be pulled toward the tail of the distribution, in this case, toward the lower values.

c. The chart displayed is not a histogram. It is a time series plot using bars to represent each data point. A histogram would arrange the data into bin intervals rather than displaying the number of games over time.

18. Baseball: Mark McGwire, again.

a. The distribution is fairly symmetric so the mean can be used to represent the centre.

b. The median value is 36.

c. The mean value should be close to the median because the distribution is fairly symmetric. The mean should be slightly higher due to the high outlier.

d. The chart displayed is not a histogram. It is a time series plot using bars to represent each data point. A histogram would arrange the data into bin intervals rather than displaying the number of games over time.

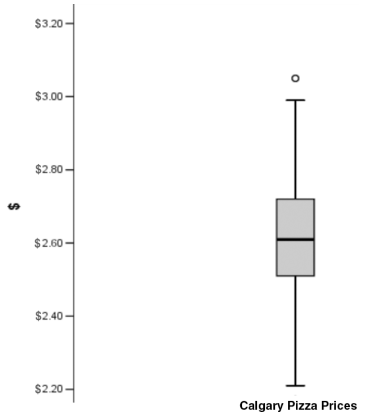
19. Pizza prices.

a. Calgary five-number summary (Quartile calculations may differ slightly when using different software.)

Min	1st Qtr	Median	3rd Qtr	Max
$2.21	$2.51	$2.61	$2.72	$3.05

b. Range = max–min = $3.05–$2.21 = $0.84; IQR = Q3–Q1 = $2.72–$2.51 = $0.21

c. Boxplot:



d. This distribution is fairly symmetric with a high outlier at $3.05. The mean is $2.62 and the standard deviation is $0.156.

e. There is one observation classified as an outlier at $3.05 per frozen pizza.

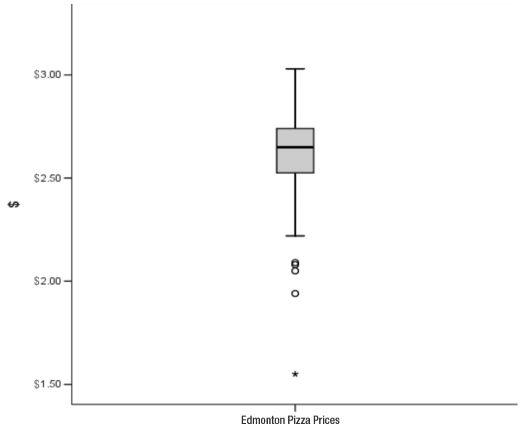
20. Pizza prices, part 2.

a. Edmonton five-number summary (Quartile calculations may differ slightly when using different software.)

Min	1st Qtr	Median	3rd Qtr	Max
$1.55	$2.525	$2.65	$2.74	$3.03

b. Range = max–min = $3.03–$1.55 = $1.48; IQR = Q3–Q1 = $2.74–$2.525 = $0.215

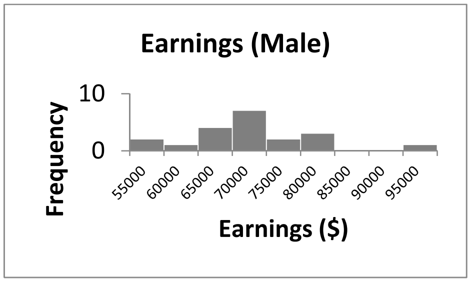
c. Boxplot:



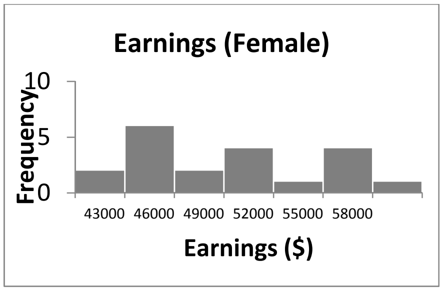
d. The main part of the distribution is approximately symmetric with four low outliers identified. The median price is $2.65 and the IQR is $0.22. Because of the outliers, the best representation of the centre of the distribution is the median price. The middle half of the prices fell in the range of $2.53 to $2.74.

e. There are four low outliers. All but five of the prices were above $2.20.

21. **Canadian yearly earnings.** A report for a given data set should include summaries and graphs with analysis. The histogram for males is approximately symmetric, with an outlier corresponding to Calgary.



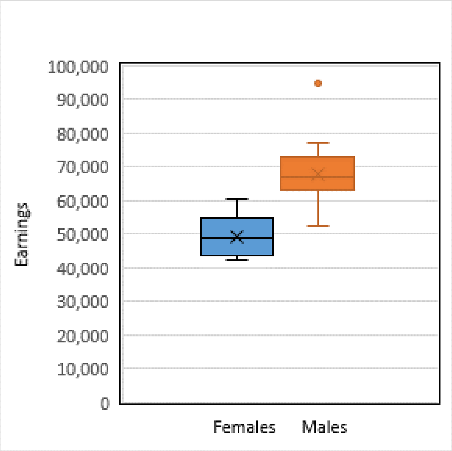
The histogram for females is slightly right-skewed, also with the highest average in Calgary.



The average across all CMAs is nearly $20 000 higher for males. Note that this does not mean that males earn $20 000 more than females for the Canadian population of full-time full-year workers; that would require a weighted average, taking into account the size of each CMA. However, Statistics Canada has this information from the Labour Force Surveys.

	Females	*Males*
Mean	48190	67785
SD	5795	9427
Minimum	42300	52600
Q1	44000	63450
Median	48700	67100
Q3	53650	71700
Maximum	60300	94900

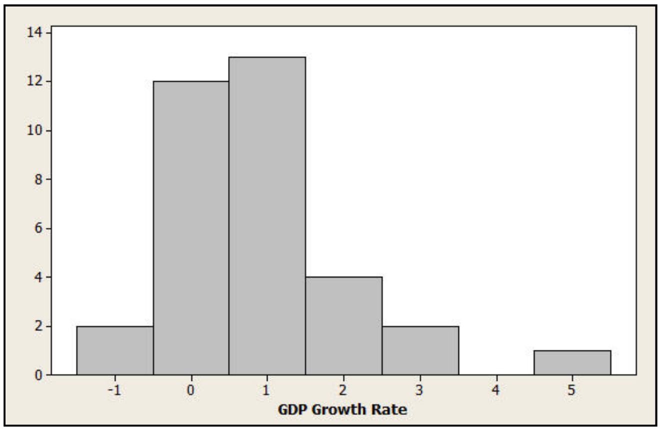
The boxplots below show the comparison of females and males across the 20 CMAs very clearly.

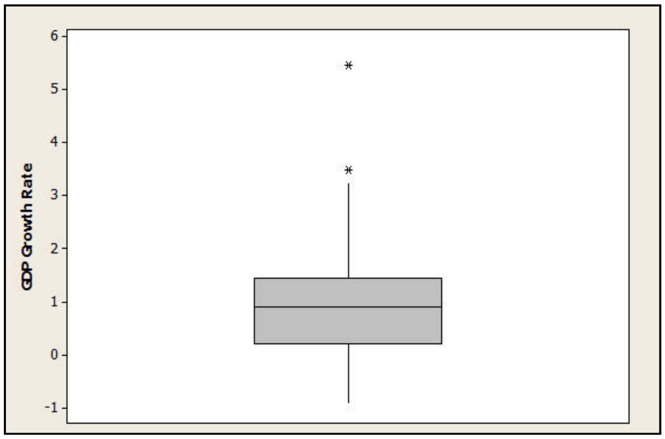


22. GDP Growth. A report for a given data set should include summaries and graphs with analysis. Five-number summary (Quartile calculations may differ slightly when using different software.)

Min	1st Qtr	Median	3rd Qtr	Max
–0.892	0.216	0.903	1.444	5.454

Range = max–min = 5.454 – (–0.892) = 6.35%; IQR = Q3 – Q1 = 1.444 – 0.216 = 1.23%



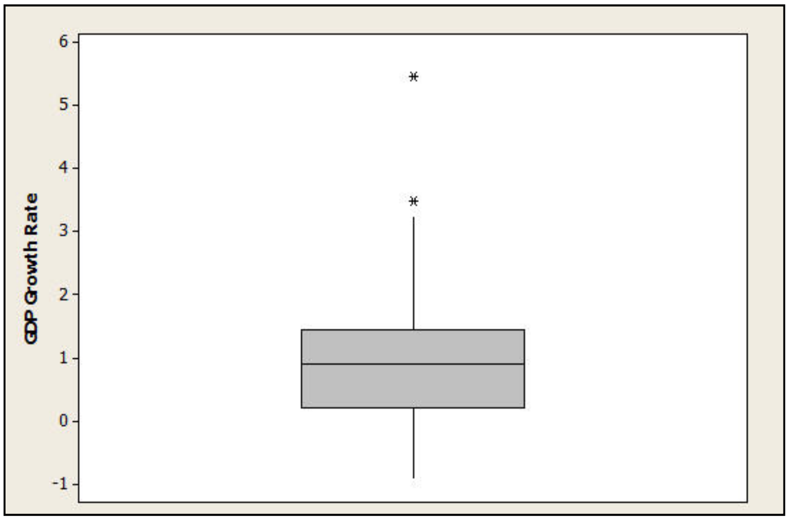


Including all countries, the mean growth rate is about 1.0% with a standard deviation of 1.26%. The histogram shows that most countries cluster around the mean. However, there are possibly three outliers (Korea, Chile, and the Slovak Republic). Not all might be identified as such depending on the software used for analysis. We may want to set them aside, or we may prefer reporting the median (0.9%) and IQR (1.2%) as summaries.

23. Start-up.

a. Range: max–min = 6796–185 = 1611 yards

b. The middle 50% of the distribution lies between Quartile 1 (5585.75 yards) and Quartile 3 (6131 yards). (Quartile calculations may differ slightly when using different software).



c. Summary statistics should include the representation of the centre, in this case, the mean (5893 yards) because the distribution is approximately symmetric. In addition, a measure of the spread for this data set is the standard deviation (386.6 yards), chosen also because the distribution is approximately symmetric.

d. Shape—the distribution of the lengths of all of the British Columbia golf courses is roughly symmetric and unimodal. Centre—the mean is 5893 yards, approximately 5900 yards. Spread—represented by the standard deviation of 386.6 yards.

24. Real estate.

a. Range: max–min = 5228–672 = 4556 sq. ft.

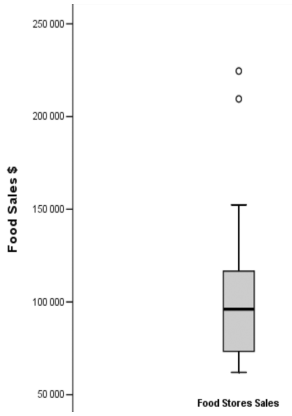
b. The middle 50% of the distribution lies between Quartile 1 (1342 sq. ft.) and Quartile 3 (2223 sq. ft.). (Quartile calculations may differ slightly when using different software).

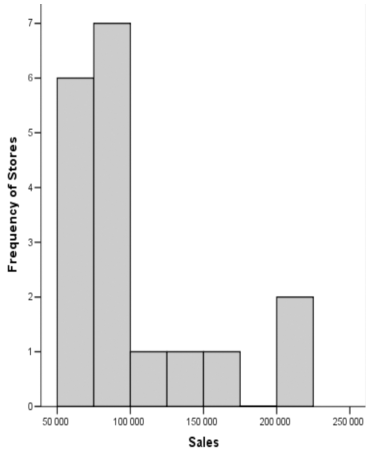
c. Summary statistics should include the representation of the centre, in this case, the median (1675 sq. ft.) because the distribution is skewed to the right. In addition, a measure of the spread for this data set is the IQR (Q3–Q1 = 2223–1342 = 881 sq. ft.), chosen also because the distribution is right skewed.

d. Shape—the distribution of the sizes of all of the houses in a specific area is skewed to the right and unimodal. Centre—the median is 1675 sq. ft. Spread—represented by the IQR (881 sq. ft.) and a range of 4556, from 672 sq. ft. (minimum) to 5228 sq. ft. (maximum).

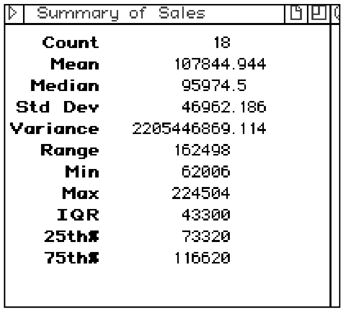
25. Food sales.

a. Suitable displays of a single quantitative variable are either a histogram or a boxplot. Both are shown here:





b. The median for the data set is $95 974.50 and the mean is $107 844.94. The mean is pulled toward the higher values because the distribution is right skewed and the higher values have the effect of increasing the mean value.



```
Summary of Sales
    Count          18
     Mean      107844.944
   Median       95974.5
  Std Dev       46962.186
 Variance  2205446869.114
    Range      162498
      Min       62006
      Max      224504
      IQR       43300
    25th%       73320
    75th%      116620
```

c. The median does a better job of depicting typical store sales because the distribution is skewed with high outliers.

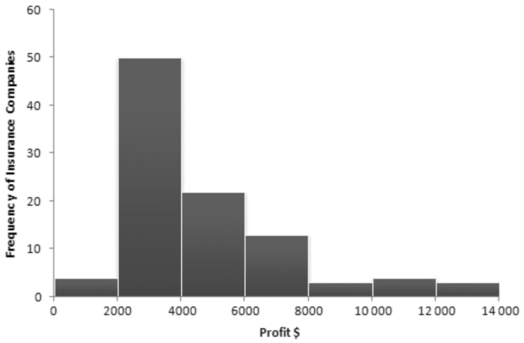
d. Standard deviation = $46 962.19 and the IQR = $43 300(Quartile calculations may differ slightly using different software)

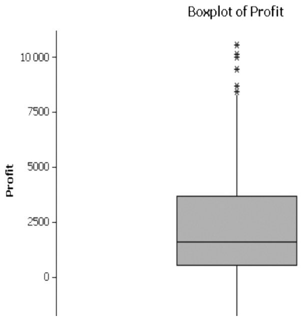
e. The IQR is a better measure of the spread for a skewed distribution with outliers. The standard deviation is affected by outliers.

f. If the outliers were removed, the mean would decrease in value because the higher numbers would not be included in the calculation. The standard deviation would also decrease, indicating a smaller spread when the high values are excluded. The median and IQR would remain relatively unaffected because the calculations are not affected by outliers unless there are a large number of them.

26. Insurance profits.

a. Suitable displays of a single quantitative variable are either a histogram or a boxplot. Both are shown here.





b. The median for the data set is $1645 and the mean is $2562. The mean is pulled toward the higher values because the distribution is right skewed and the higher values have the effect of increasing the mean value.

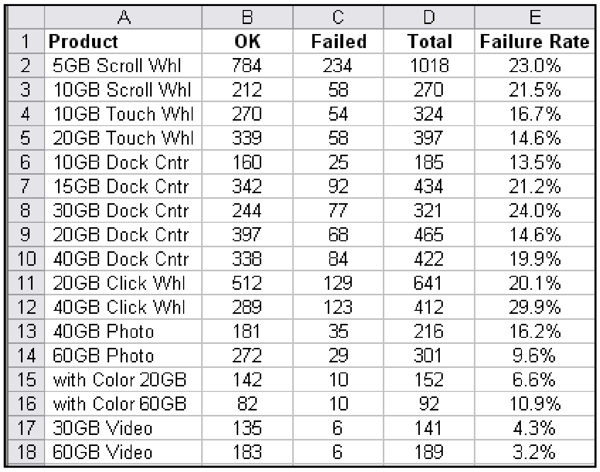
c. The median is a better indicator of the centre of the distribution because the distribution is skewed.

d. The standard deviation is $2683.10; the IQR is $3059.50. (Quartile calculations may differ slightly when using different software.)

e. The IQR is resistant to the effects of outliers.

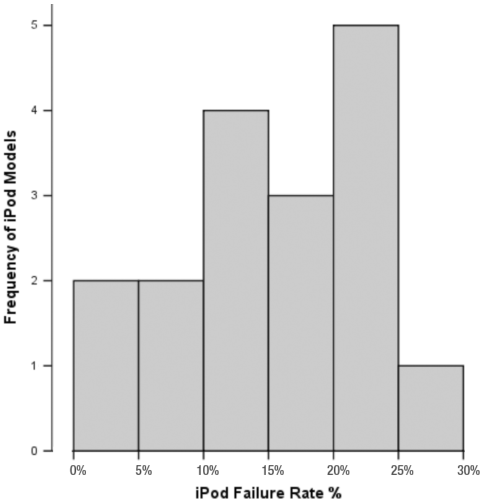
f. The mean and standard deviation would decrease. The median and IQR would be relatively unaffected.

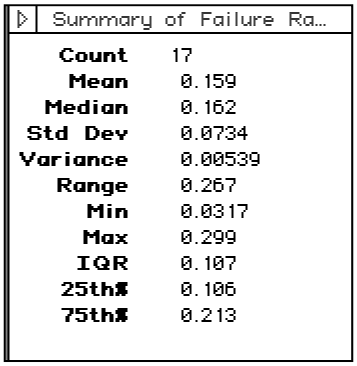
27. Ipod failures. The failure rate is calculated by dividing the number failed by the total number of iPods (both failed and OK) for each model. The result is then multiplied by 100% to get a rate.



	A	B	C	D	E
1	**Product**	**OK**	**Failed**	**Total**	**Failure Rate**
2	5GB Scroll Whl	784	234	1018	23.0%
3	10GB Scroll Whl	212	58	270	21.5%
4	10GB Touch Whl	270	54	324	16.7%
5	20GB Touch Whl	339	58	397	14.6%
6	10GB Dock Cntr	160	25	185	13.5%
7	15GB Dock Cntr	342	92	434	21.2%
8	30GB Dock Cntr	244	77	321	24.0%
9	20GB Dock Cntr	397	68	465	14.6%
10	40GB Dock Cntr	338	84	422	19.9%
11	20GB Click Whl	512	129	641	20.1%
12	40GB Click Whl	289	123	412	29.9%
13	40GB Photo	181	35	216	16.2%
14	60GB Photo	272	29	301	9.6%
15	with Color 20GB	142	10	152	6.6%
16	with Color 60GB	82	10	92	10.9%
17	30GB Video	135	6	141	4.3%
18	60GB Video	183	6	189	3.2%

The median value for the failure rate for all 17 models is 16.2%. An appropriate graphical display of the distribution of a single quantitative variable is either a histogram or a boxplot.





```
▷ Summary of Failure Ra...

   Count    17
    Mean     0.159
  Median     0.162
 Std Dev     0.0734
Variance     0.00539
   Range     0.267
     Min     0.0317
     Max     0.299
     IQR     0.107
   25th%     0.106
   75th%     0.213
```

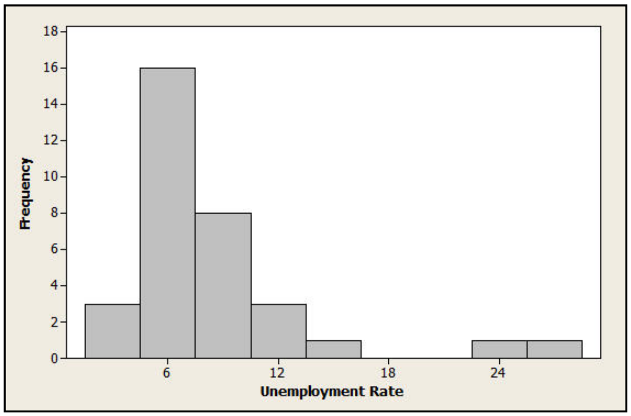
The distribution is left skewed. The centre is best represented by the median at 16.2% and the spread is best represented by the IQR which is 10.7%. (Quartile calculations may differ slightly when using different software.) The middle 50% of failure rates are between 10.6% and 21.3%. The lowest value or best failure rate is 3.2% for the 60 GB Video model, and the worst failure rate is 29.9% for the 40 GB Click Wheel.

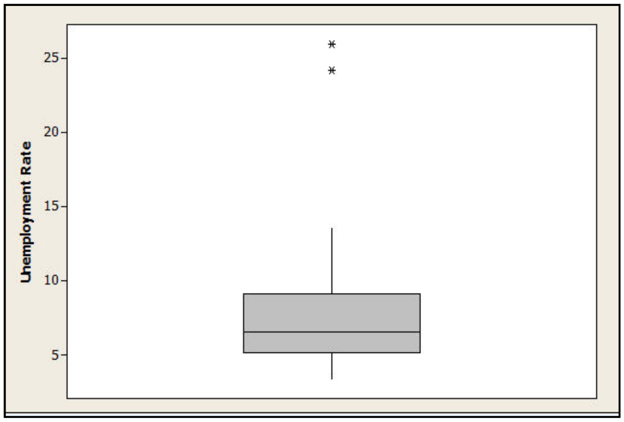
28. Unemployment. Both the histogram and boxplot show that the distribution is unimodal, skewed to the right with two high outliers (Greece and Spain)..

Five-Number Summary (Quartile calculations may differ slightly using different software)

Min	1^{st} Qtr	Median	3^{rd} Qtr	Max
3.467	5.4	6.567	8.7	25.933

The median rate for these 33 countries is 6.6%. The lowest is Korea at 3.5% and highest is Greece at 25.9%. The middle 50% of these countries have rates between 5.4% and 8.7%. The IQR is 3.3%. Range = max – min = 25.9 – (3.5) = 22.4%.





29. Sales. In describing side-by-side boxplots, there are two important things to mention—the description of each data distribution and the comparison of the distributions to each other. Location #1 is roughly symmetric with some high outliers, one exceeding $320 000. The median sales value is approximately $240 000and the minimum sales value is approximately $160 000. Location #2 is also roughly symmetric with high outliers close to $150 000to $180 000. The median sales value is approximately $110 000and the minimum sales value below $100 000. Location #1 clearly has higher sales than Location #2 in every week except for the high outlier in Location #2.

30. Sales, part 2. In describing side-by-side boxplots, there are two important things to mention—the description of each data distribution and the comparison of the distributions to each other. The distribution of stores in Saskatchewan (SK) is right skewed with several high outliers. The median sales value is approximately $115 000 and the minimum sales value is approximately $85 000. High outliers extend up to about $220 000. The distribution of stores in Manitoba (MB) is also right skewed with several outliers. The median sales value is approximately $100 000 and the minimum sales value is approximately $65 000. High outliers extend up to about $170 000. Overall, the stores in SK generally have higher sales than the stores in MB. The median value in SK is higher than the third quartile in MB. The lowest performing store in SK was higher than nearly 25% of the stores in MB.

31. Gas prices, part 2.

- **a.** It is obvious that gas prices have steadily increased over the three-year period from 2010–2012. The data spread has also increased over these three years. The 2010 distribution of prices is skewed to the left with several low outliers. Starting in 2011, the distributions have become increasingly skewed to the right. There is a high outlier in 2012 that is close to the upper fence.
- **b.** The evidence for prices showing more volatility would be prices showing a greater spread and a larger IQR. The year with the greatest spread and IQR is 2012.

32. Fuel economy. Cars with four cylinders generally get better gas mileage than cars with six cylinders, which generally get better gas mileage than cars with eight cylinders. In addition, consistency can be determined by less spread in the distribution. The most compact distribution with little variation is for the eight cylinder cars. Four-cylinder cars vary a great deal in their fuel economy, the lowest value being close to 22 mpg and the highest value close to 37 mpg. A typical value is represented by the median which is close to 32 mpg for four-cylinder cars, close to 21 mpg for six-cylinder cars and close to 17 mpg for eight-cylinder cars. The IQR represents the middle 50% of the values so we can say that four-cylinder cars typically get 28–34 mpg, six-cylinder cars typically get 18–22 mpg, and eight-cylinder cars get 16–19 mpg. There appears to be only one data point close to 21 mpg representing five cylinders.

33. Wine prices.

- **a.** Identify the highest value for the three regions—occurs in Seneca Lake.
- **b.** Identify the lowest value for the three regions—occurs in Seneca Lake.
- **c.** To answer the question about which wines are generally more expensive, look at the IQR box and determine which region has the middle 50% of its prices higher than the others. That region is clearly Keuka Lake.
- **d.** Cayuga Lake and Seneca Lake vineyards have approximately the same price at about $200 per case. The middle 50% of prices for these two regions are also similar, from approximately $150 to $220 per case. A typical Keuka Lake vineyard case has a much higher price of about $260 and the middle 50% values are between $240 and $280 per case with one outlier at approximately $170 per case. Seneca Lake vineyard prices are the most varied and include both the lowest and the highest prices for all three regions (from $100 to $300).

34. Ozone.

- **a.** Identify the highest value for all of the months—occurs in both March and April at approximately 430.
- **b.** The largest IQR representing the middle 50% of values is shown by the length of the box which is the largest for February at approximately 50.
- **c.** The smallest range is represented by the smallest distance from low to high values which looks like August with a range of less than 50.
- **d.** January had median levels slightly lower than June, with values at 340 and 350 respectively. June's ozone levels were more consistent, represented by the compact distribution. January's ozone range was 300 to 400 while June's range was 310 to 380. June has both low and high outlier values.
- **e.** Generally, ozone levels rose during the winter and were highest in the spring, then fell throughout the summer months and were lowest in the fall. Ozone levels were consistent in the summer (represented by the compact distribution) and became more variable in the fall and most variable in winter (represented by the expanded distribution). Ozone levels then became more consistent in the spring with levels starting to drop.

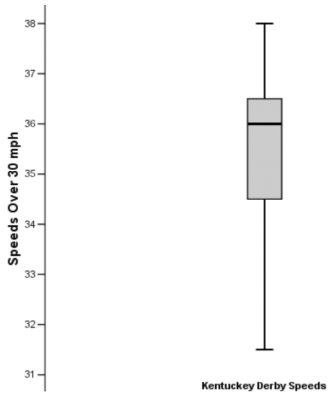
35. Derby speeds.

a. The median speed is the speed at the middle of the data values where 50% run slower and 50% run faster. The values represent the percentage of Kentucky Derby winners that have run slower than the cutoff of 48 kph. The data set already represents a percentage of the total distribution so the median and quartiles can be determined by looking up the percentage values on the y-axis. The 50% value can be determined at the 50% mark on the y-axis and moving over to the plotted points to approximately 58 kph (find 50% on the y-axis, move straight over to the right and identify the value on the plot on the x-axis).

b. The quartiles are found in a similar way. Find the winning speed representing 25% for the first quartile: Q1 = approximately 55.5 kph. Find the winning speed representing 75% for the third quartile: Q3 = approximately 58.5 kph.

c. Range = values that represent 0% to 100% = 50 to 60 kph = 10 kph. The IQR is 75% value–25% value = 58.5 kph–55.5 kph = 3 kph.

d. Boxplot**:**



e. The distribution of speeds is skewed to the left. The lowest speed is close to 50 kph mph and the fastest speed is about 60 kph. The median speed is approximately 58 kph. Twenty five percent of the speeds are above 58.5 kph mph and 75% of winning speeds are above 55.5 kph. Only a small percent of winners had speeds below 53 kph. Without the actual data set, the boxplot is constructed from the five-number summary and therefore fences and outliers other than the maximum and minimum cannot be determined.

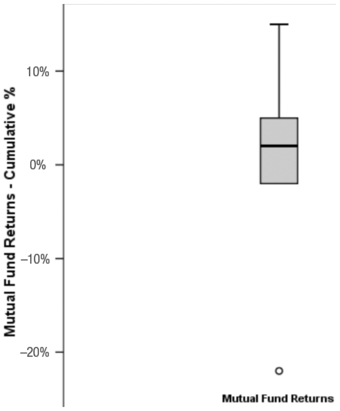
36. Mutual funds, part 3.

a. The data set already represents a percentage of the total distribution so the median and quartiles can be determined by looking up the percentage values on the y-axis. This type of cumulative frequency graph (ogive) makes this possible. The 50% value can be determined at the 50% mark on the y-axis and moving over to the plotted points to approximately 2% return (find 50% on the y-axis, move straight over to the right and identify the value on the plot on the x-axis).

b. The quartiles are found in a similar way. Find the returns representing 25% for the first quartile: Q1 = approximately –2%. The return representing 75% for the third quartile: Q3 = approximately 5%.

c. Range = values that represent 0% to 100% = –22% to 15% = 37%. The IQR is 75% value–25% value = –2%–5% = 7%.

d. The boxplot is created only from the five-number summary therefore cannot determine the fence values and multiple outliers which need the actual data points for calculation.



37. Test scores.

a. The highest mean score can be determined from looking at the shape of the test histograms. Class 1 is symmetric with a mean in the 60 interval. Class 2 is somewhat symmetric with the centre also being in the 60 interval. Class 3 is left skewed with most of the data points to the right of a score of 60 so the mean for Class 3 would be the highest.

b. The same logic can be applied to find the median score. For symmetric distributions, the mean is the same as the median and close to the median for approximately symmetric distributions. Most of the data points for Class 3 are to the right of a score of 60 therefore the median for Class 3 would be the highest.

c. For symmetric and approximately symmetric distributions, the mean and the median are almost the same. That applies to both Class 1 and Class 2. Class 3 will have a mean pulled toward its distribution tail which is to the left for a left-skewed distribution. Therefore, the difference between the mean and the median will be greatest for Class 3.

d. The smallest standard deviation represents the smallest data spread from the mean of the distribution. Because Class 1 is symmetric and clustered more tightly around the centre (about 60), it will have the smallest standard deviation.

e. The smallest IQR represents the most condensed data in the middle 50% region. Class 1 probably has the smallest IQR because it is the most symmetric, however, without the actual scores, it is impossible to calculate the exact IQRs.

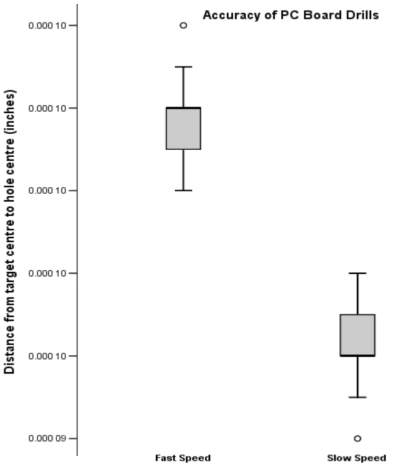
38 Test scores, again.

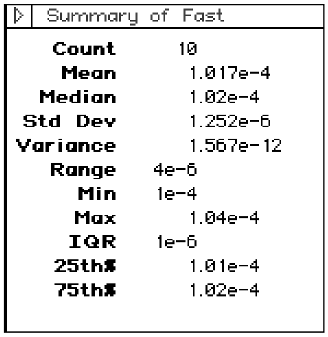
a. Class 3 has the middle 50% of the data (IQR) nearly above the median of the other two classes. Clearly Class 3 has the overall higher scores.

b. Class 1 has a symmetric distribution and Class 2 has a roughly symmetric distribution. Both are unimodal. Class 3 has a distribution of scores that is skewed to the left and also unimodal.

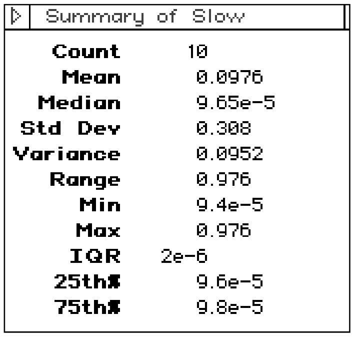
c. Class A is Class 1; Class B is Class 2; and Class C is Class 3.

39. Quality control. In order to analyze the data, it is appropriate to create summaries separately for Fast and Slow data sets. In addition, create side-by-side boxplots for comparison of distributions.





```
▷ Summary of Fast
     Count      10
      Mean        1.017e-4
    Median        1.02e-4
   Std Dev        1.252e-6
  Variance        1.567e-12
     Range    4e-6
       Min    1e-4
       Max        1.04e-4
       IQR    1e-6
     25th%        1.01e-4
     75th%        1.02e-4
```

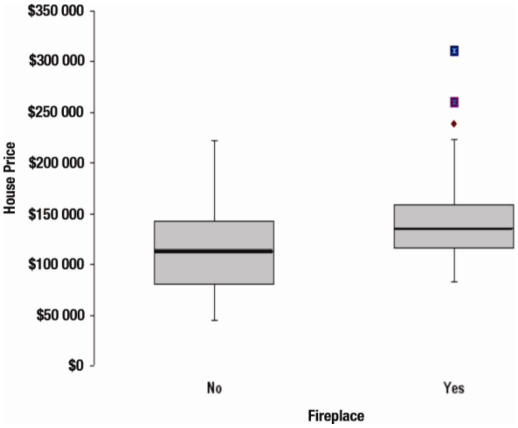


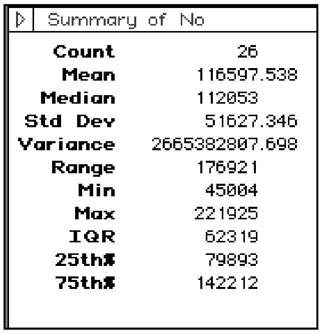
```
▷ Summary of Slow
     Count      10
      Mean        0.0976
    Median        9.65e-5
   Std Dev        0.308
  Variance        0.0952
     Range        0.976
       Min        9.4e-5
       Max        0.976
       IQR    2e-6
     25th%        9.6e-5
     75th%        9.8e-5
```

The slow drilling data contains an extremely high outlier indicating that one hole was drilled almost a centimetre away from the centre of the target. If this data point is correct, the engineers should investigate the slow speed drilling process closely for any extreme, intermittent inaccuracy. The outlier is so extreme that no graphical display can show the distributions in a meaningful way if that data point is included. The outlier can be removed in order to look at the remaining data points. However, it should be noted that it is never good scientific practice to eliminate a data point because it doesn't fit the distribution. All outliers should be investigated to determine if they are a result of human error, input error, or some problems with the measurements that need to be addressed. It seems apparent that the entry of 0.975 600 0 was in error and was meant to be 0.000 009 576. This should be investigated and, if found to be true, the summary statistics should be recalculated.

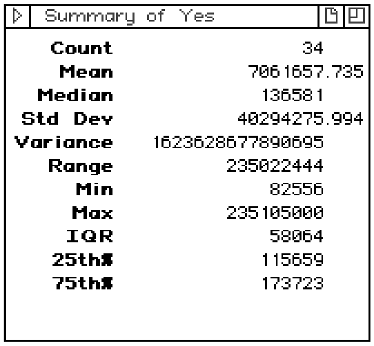
But with the outlier removed, the slow drilling process is shown to be more accurate. The entire distribution with the exception of the extreme outlier not pictured lies well below the fast distribution. The greatest distance from the target for the slow drilling process is 0.000 098 centimetres, which is more accurate than the smallest distance for the fast drilling process, 0.000 100 centimetres.

40. Fire sale. In order to analyze the data, it is appropriate to create summaries separately for the No Fireplace and Fireplace data sets. In addition, create side-by-side boxplots for comparison of distributions.





```
Summary of No
    Count          26
     Mean   116597.538
   Median      112053
  Std Dev    51627.346
 Variance 2665382807.698
    Range      176921
      Min       45004
      Max      221925
      IQR       62319
    25th%       79893
    75th%      142212
```



```
Summary of Yes
    Count                   34
     Mean          7061657.735
   Median            136581
  Std Dev         40294275.994
 Variance  1623628677890695
    Range         235022444
      Min             82556
      Max         235105000
      IQR             58064
    25th%            115659
    75th%            173723
```

The house prices with a fireplace contain a seemingly impossible high outlier at $235 105 000. This price is beyond the maximum selling price of a house in 2006. The outlier can be removed in order to look at the remaining data points. However, it should be noted that it is never good scientific practice to eliminate a data point because it doesn't fit the distribution. All outliers should be investigated to determine if they are a result of human error, input error, or some problems with the measurements that need to be addressed. It seems apparent that the entry of 235 105 000 was in error and was meant to be 235 105. This should be investigated and, if found to be true, the summary statistics should be recalculated.

But with the outlier removed, the side-by-side boxplot shows that the prices for houses with fireplaces are generally higher than those without fireplaces. The median price for houses with fireplaces is close to $140 000 compared to a value close to $110 000 for houses without. The spread of sale prices for house without fireplaces is much greater than houses with fireplaces. There are only three houses with fireplaces that are more expensive than the most expensive house without a fireplace.

41. Customer database.

a. The mean of 54.41 is meaningless. The data set is categorical, not quantitative.

b. Typically, the mean and standard deviation are influenced by outliers and skewness, however, once again, these results are meaningless in this instance because the values are categorical.

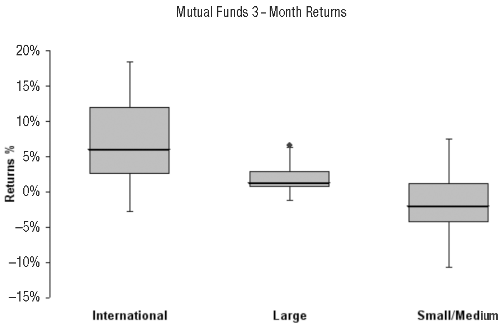
c. No. Summary statistics are only appropriate for quantitative data.

42 CEOs.

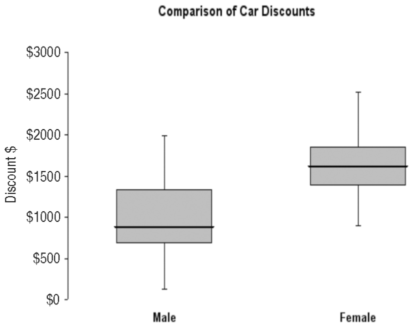
a. Industry codes are categories. Therefore, it is not appropriate to create a histogram of the data. The graph shown is a bar graph.

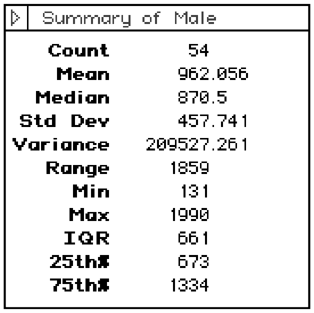
b. Histograms require a single quantitative variable for analysis. The complaints bar graph simply identifies which codes have the largest number of complaints. It is not appropriate to consider the distribution of this data set.

43. Mutual funds types. Over the three-month period, International Funds generally outperformed the other two types of mutual funds. Almost half of the International Funds outperformed all the funds in the other two categories. U.S. Domestic Large Cap Funds did better than U.S. Domestic Small/Mid Cap Funds in general. Large Cap funds had the least variation of the three types.

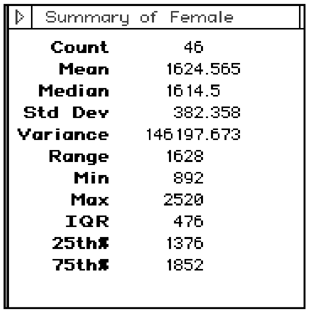


44. Car discounts, part 3. In general, women received larger discounts than men. The median discount for women (approximately $1750) was higher than the third quartile of the men's discounts (approximately $1400). The smallest discount received by a woman was larger than the median of the male discounts.





```
Summary of Male
    Count        54
     Mean        962.056
   Median        870.5
  Std Dev        457.741
 Variance        209527.261
    Range        1859
      Min        131
      Max        1990
      IQR        661
    25th%        673
    75th%        1334
```



```
Summary of Female
    Count        46
     Mean        1624.565
   Median        1614.5
  Std Dev        382.358
 Variance        146197.673
    Range        1628
      Min        892
      Max        2520
      IQR        476
    25th%        1376
    75th%        1852
```

45. Houses for sale.

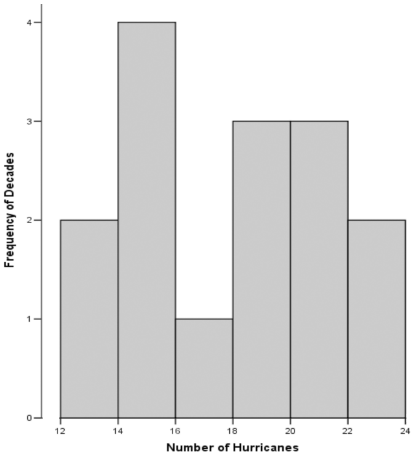
a. Even though MLS ID numbers are categorical identifiers, they are assigned sequentially, so this graph can be analyzed as shown. Most if not almost all of the houses listed a long time ago have sold and are no longer listed. The older numbers are represented by the lower values giving the distribution the left skewed shape.

b. Although some information could be gathered from this graph, a histogram is generally not an appropriate display for categorical data. The MLS ID numbers are categorical identifiers.

46. Zip codes. Zip codes are numbers but actually are categorical data, identifying the mailing codes for certain regions. They are assigned sequentially, so this graph does contain some useful information. The leading digit gives a rough East-to-West placement in the United States. The company has fewest customers in the Northeast. A bar chart using the leading digit would be a more appropriate display.

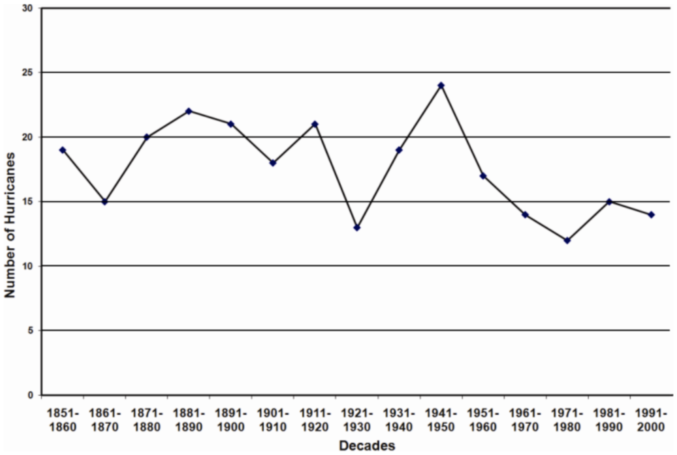
47. Hurricanes.

a. Histogram:



b. The distribution is fairly uniform and appears somewhat right skewed. There do not appear to be any outliers.

c. Timeplot:

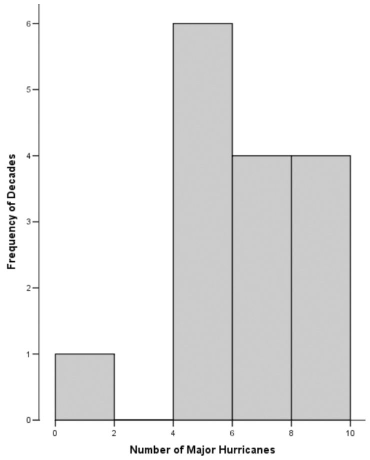


d. The timeplot does not support the claim that the number of hurricanes has increased in recent decades. There was a peak in the 1940s but numbers have decreased since that time.

48. Hurricanes, part 2.

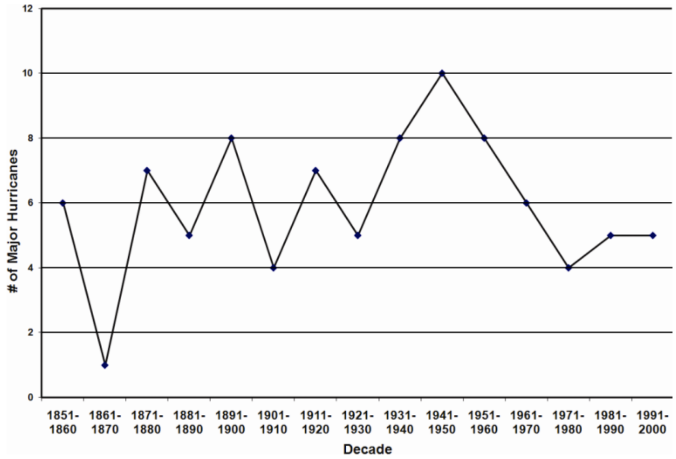
a. Histogram:

b.



c. The distribution is fairly uniform from 4 to 10 hurricanes with one outlying decade with only one major hurricane.

d. Timeplot:



e. Most decades have five to eight major hurricanes with no obvious trend over time. This does not support the scientists' claim—at least not up to the year 2000.

49. Productivity study. Questions include: What is plotted on the x-axis? If it is time, what are the units? Months? Years? Decades? How is productivity measured?

50. Productivity study revisited. Questions include: What is plotted on the x-axis and y-axis? What are the units? Since productivity and wages have different units, how does the graph make sense? We are unable to compare the two variables.

51. Real estate, part 2. To answer this question, the values have to be standardized and the *z*-scores compared according to the formula: $z = \frac{(y - \bar{y})}{s}$ where y is the value being compared to $\bar{y}$, which is the mean value, and s, which is the standard deviation. The house that sells for \$400 000 has a *z*-score of (400 000–167 900)/77 158 = 3.01. The house with 4000 sq. ft. of living space has a *z*-score of (4000–1819)/663 = 3.29. The value of 3.29 is further away from the centre of a normal distribution than 3.01 and, therefore, is the more unusual value.

52. University tuition. To answer this question, the values have to be standardized and the *z*-scores compared according to the formula: $z = \frac{(y - \bar{y})}{s}$ where y is the value being compared to $\bar{y}$, which is the mean value, and s, which is the standard deviation. The Canadian student tuition of $3000 has a *z*-score of (3000–5072)/939 = –2.21. The International student tuition of $7500 has a *z*-score of (7500–14 427)/3758 = –1.84. The value further from the centre of the distribution would be the Canadian student tuition, making it more unusual.

53. Food consumption. To answer this question, the values have to be standardized and the *z*-scores compared according the formula: $z = \frac{(y - \bar{y})}{s}$ where y is the value being compared to $\bar{y}$, which is the mean value, and s, which is the standard deviation. The mean and standard deviation need to be calculated from the given data set. For meat consumption, the U.S. *z*-score = (267.30–181.031)/53.077 = 1.625 and the Ireland *z*-score = (194.26–181.031)/53.077 = 0.25. Therefore, the U.S. has a more remarkable meat consumption than Ireland because the *z*-score is higher, indicating that it is further from the distribution mean.

To answer this question, find the *z*-scores separately for meat and alcohol for each country. For meat consumption, the U.S. *z*-score = 1.625 and the Ireland *z*-score = 0.25. For alcohol consumption, the U.S. *z*-score = (26.36 – 26.778)/10.469 = –0.04 and the Ireland *z*-score = (55.80 – 26.778)/10.469 = 2.77. The total for the U.S. is 1.625 + (–0.04) = 1.585. The total for Ireland is 0.25 + 2.77 = 3.02. Ireland has higher overall consumption for meat and alcohol combined.

54. World bank.

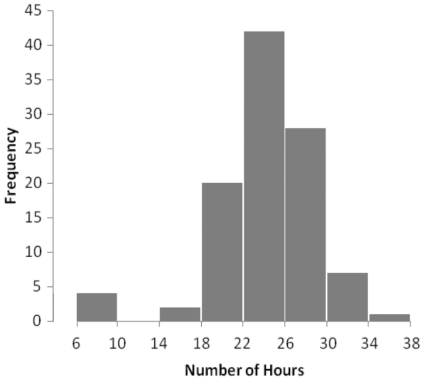
a. To answer this question, the values have to be standardized and the *z*-scores compared.

***z*-score**	**Procedures**	**Time**	**Cost**	**Total**
Spain	(10–7.9)/2.9 = 0.724	(47–27.9)/19.6 = 0.974	(15.1–14.2)/12.9 = 0.07	1.768
Guatemala	(11–7.9)/2.9 = 1.069	(26–27.9)/19.6 = –0.010	(47.3–14.2)/12.9 = 2.566	3.625
Fiji	(8–7.9)/2.9 = 0.034	(46–27.9)/19.6 = 0.923	(25.3–14.2)/12.9 = 0.860	1.817

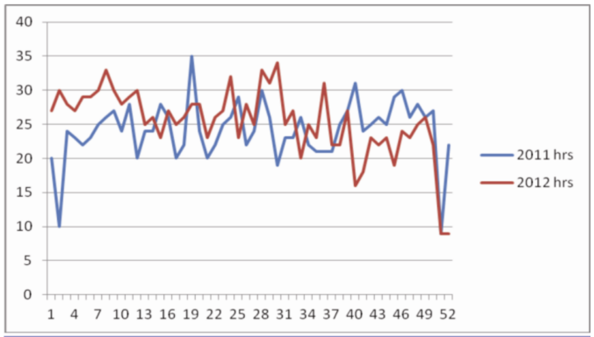
b. Spain has the lowest total and therefore is a better environment to start a business.

55. Personal fitness trainers.

a. The histogram is symmetric with outliers at the left end. There are four weeks with 10 or fewer hours. It is likely that these are vacation weeks or weeks during the December holiday season. Most weeks have between 18 and 30 hours of client training.



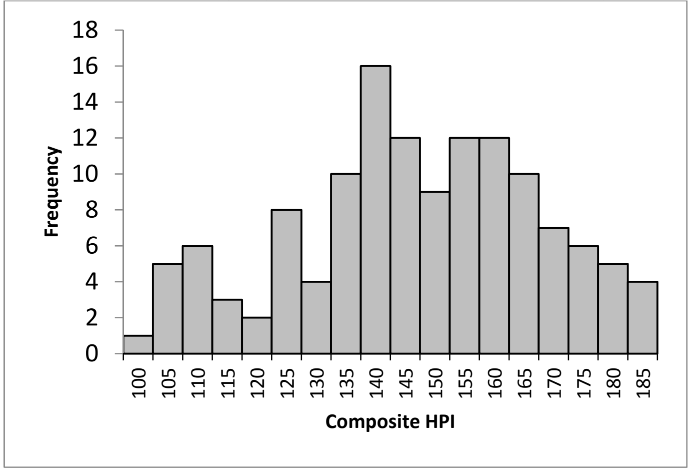
b. The time series plot shows that hours are relatively stable throughout the year, except for expected slowdowns in late December and early January. Other drops likely correspond to vacation time. Both years show very similar work patterns.

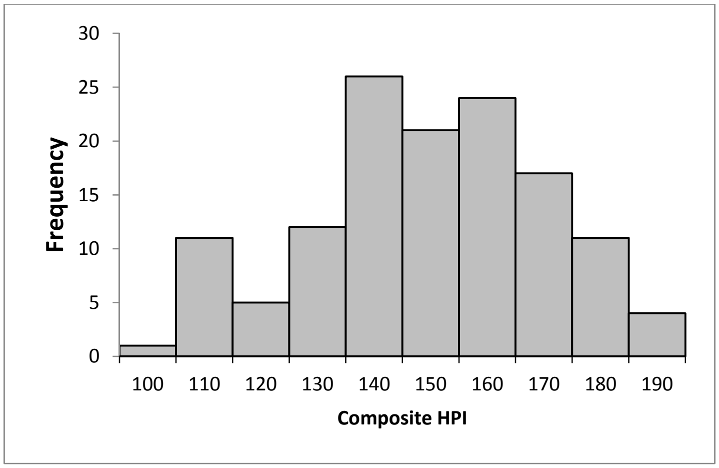


c. The time series plot is more appropriate to show the how training hours change during the year.

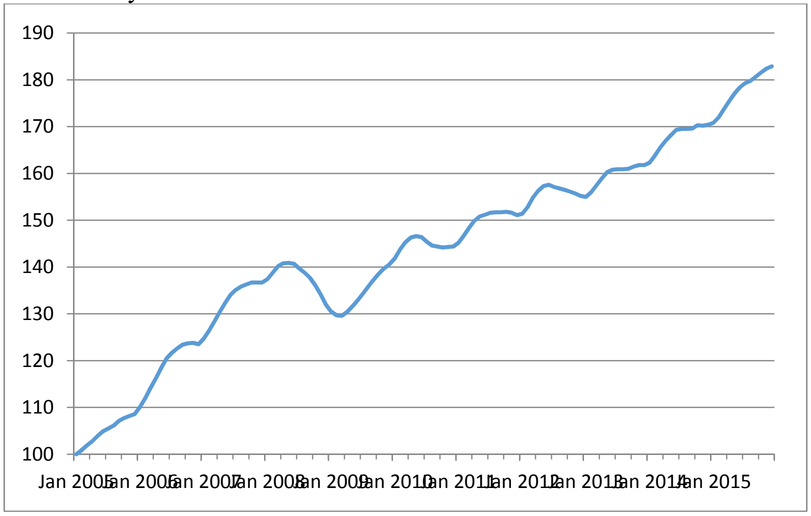
56. Canadian house prices.

a. Two versions of the histogram (one using a bin width of 5, the other using a bin width of 10), show that the distribution is slightly skewed to the left and possibly bimodal.





b. The timeplot shows that housing prices have been increasing quite steadily over the eight-year period, except for the dip in mid-2008 to mid-2009, at the height of the world economic slowdown. Prices began to recover by mid-2009.



The timeplot is more appropriate here it because it demonstrates more of the structure of the data; that is, how the data change over time.

57. Canadian unemployment rate.

a. The histogram shows that the distribution is possibly skewed to the right and definitely bimodal.

b. The trend over time is clearer in the time series plot.

c. The time series plot is more appropriate because it demonstrates more of the structure of the data, how the data change over time.

d. Unemployment decreased steadily from approximately 7.5%–8.0% in 2003 to just below 6.0% at the start of 2008. Then the rate increased sharply over the next year (2009 was when the major effects of the world economic meltdown were felt), rising to 8.5% before beginning another steady decline through 2012, to a level of 7.0–7.5%. It has remained at about 7.0% since then, with a slight drop in early 2015.

58. Mutual fund performance.

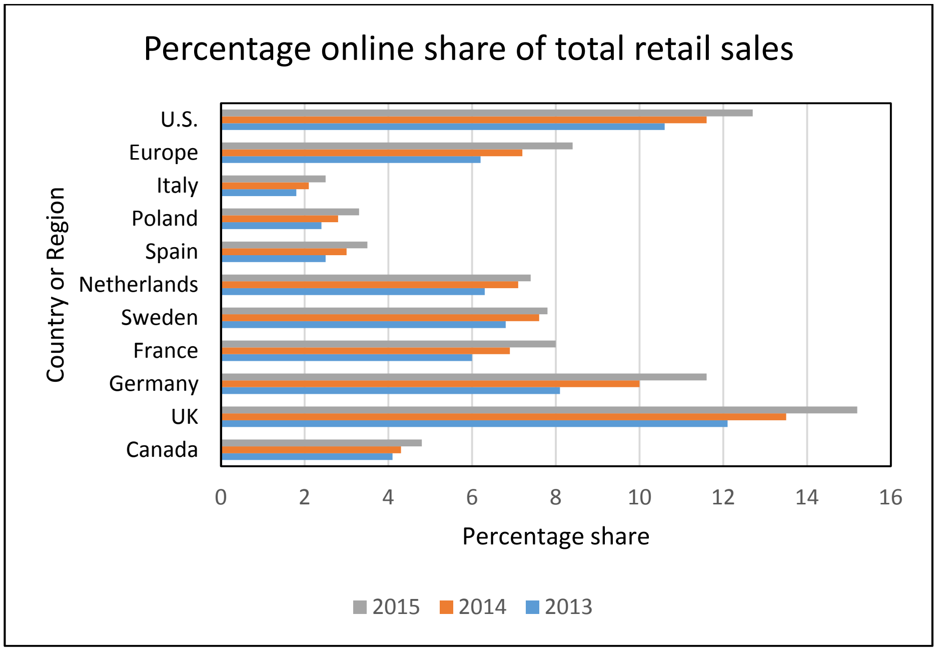
a. The distribution of returns is unimodal and symmetric with one very low outlier.

b. There is little trend over time but the variability decreased in the final few years of this series.

c. With returns fairly stable over time, both plots are appropriate and display important information about the data. The histogram describes the distribution of the overall data while the time series plot shows the stability over time with the exception of the outlier.

d. Returns have no apparent trend over time but do fluctuate sometimes between –10% and 10%. Stability seems to increase in the more recent years with less range in fluctuations. There is one low outlier below –20% corresponding to the market crash of October 1987.

59. Online retailing. A clustered horizontal bar chart allows for comparison across years for each country or region, and comparisons across countries. Other good graphs are possible.



60. Assets.

a. The distribution is severely skewed and has a number of extraordinarily large outliers. Using the mean as the centre would give a highly inflated value. Using the median would ignore the effect of the outliers. It is hard to know what is meant by centre. Similarly, the spread could be either underestimated or overestimated depending on how the outliers and skewness are treated.

b. A transformation (log or square root) is recommended.

61. Assets, again.

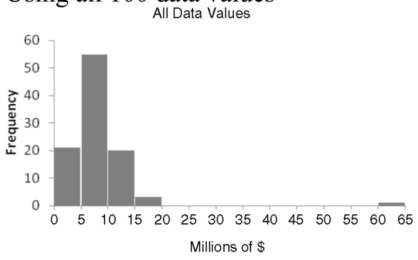
a. The log transform does a better job of reducing the skewness and the unusual nature of the outliers.

b. The company has assets of $100 billion.

Mini Case Study Project—Canadian CEO Top 100

Data have been converted to millions of dollars for graphs and calculations

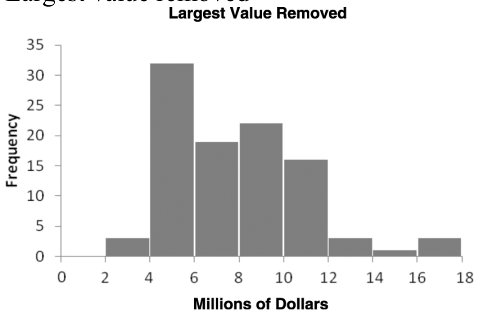
a. Using all 100 data values



Mean	8.384
SD	6.195
Min	3.892
Q1	5.242
Median	7.168
Q3	9.931
Max	68.811
IQR	4.689
Lower fence	0
Upper fence	16.965

There is only one outlier, at $68.811 million (outliers are values exceeding $16.965 million).

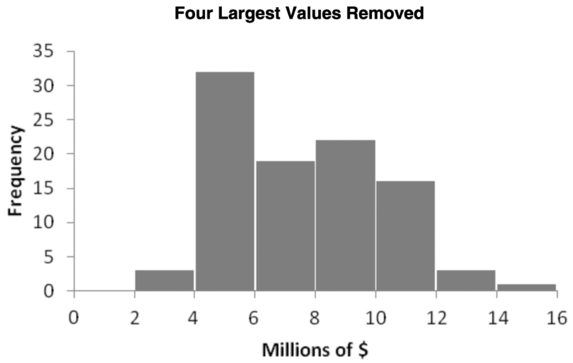
b. Largest value removed



Mean	7.845
SD	3.058
Min	3.892
Q1	5.242
Median	7.168
Q3	9.920
Max	16.679
IQR	4.678
Lower fence	0
Upper fence	16.937

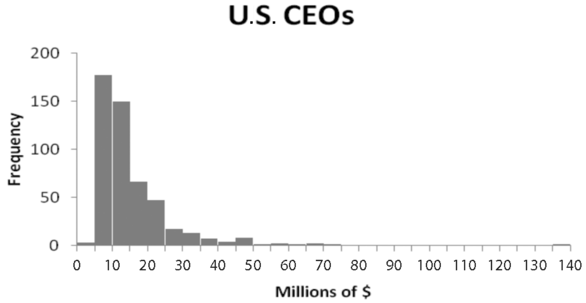
The distribution looks right-skewed. The mean has dropped from $8.38 million to $7.85 million; the SD is cut in half.

c. Largest four values removed



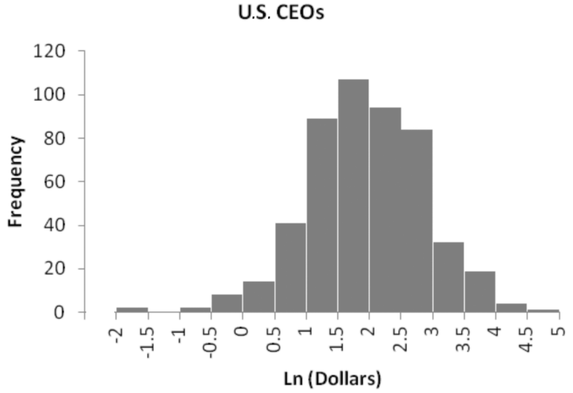
The distribution is still right-skewed. (Note: Skewness is a property of the entire distribution, not just an effect of a few outliers.)

d. Five hundred U.S. CEOs show a similarly right-skewed distribution with an extreme outlier.



Mean	10.476
SD	11.462
Min	0
Q1	3.895
Median	6.968
Q3	13.363
Max	131.190
IQR	9.468
Lower fence	0
Upper fence	27.565

e. After applying the natural logarithm of U.S. CEO pay the distribution looks very symmetric, but with a small number of outliers on the left-hand tail.



Mini Case Study Project—Hotel Occupancy Rates

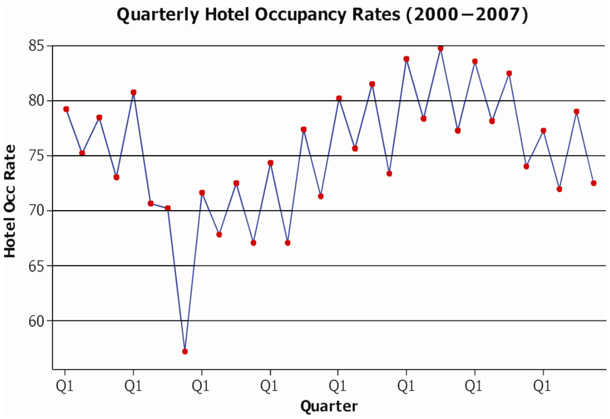
Report:

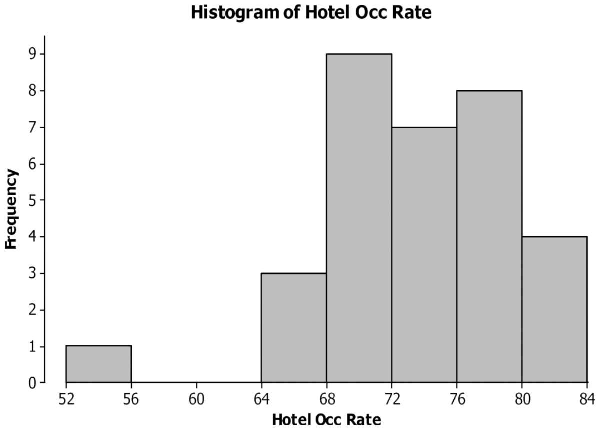
The time series plot shows a seasonal pattern in Honolulu hotel occupancy rates that peak in the first and third quarters and dip in the second and fourth quarters. On all plots and charts an unusual observation (an outlier) is evident. This occurs during the fourth quarter of 2001, likely caused by the drop in travel resulting from the September 11 terrorist attacks. Consequently, it should not be taken account in future planning. Since that quarter there has been a general upward trend until around early 2006. This pattern is most likely related to the underlying economic (business) cycle.

Descriptive Statistics: Hotel Occupancy Rate

Variable	N	N*	Mean	Mean	StDev	Minimum	Q1	Median	Q3
Hotel Occ Rate	32	0	75.26	1.04	5.89	57.20	71.73	75.45	79.15

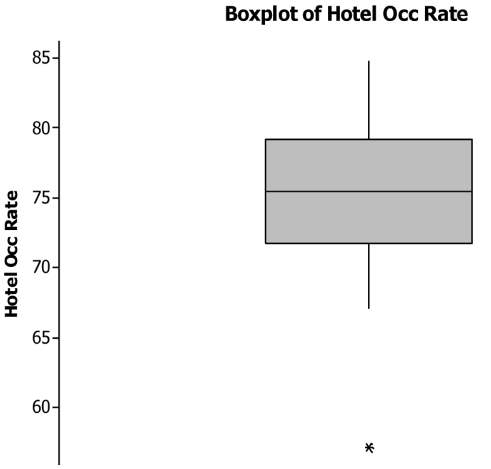
Variable	Maximum
Hotel Occ Rate	84.77





Stem-and-Leaf Display: Hotel Occ Rate

```
5|7¶
5|¶
6|¶
6|¶
6|¶
6|777¶
6|¶
7|00111¶
7|2233¶
7|4455¶
7|777¶
7|88899¶
8|001¶
8|233¶
8|4¶
```



Mini Case Study Project—Value and Growth Stock Returns

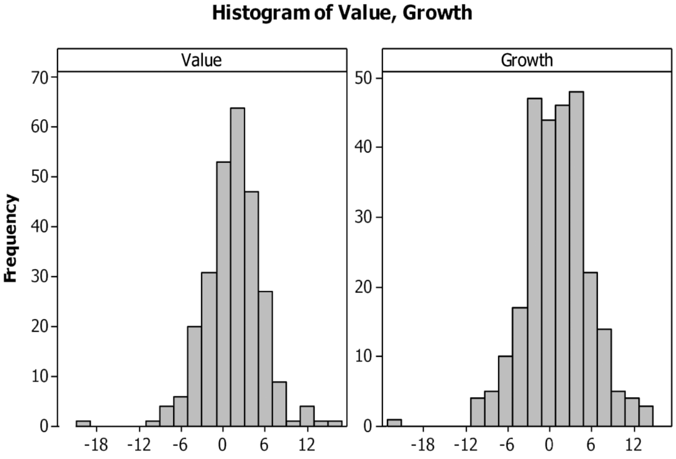
Report:

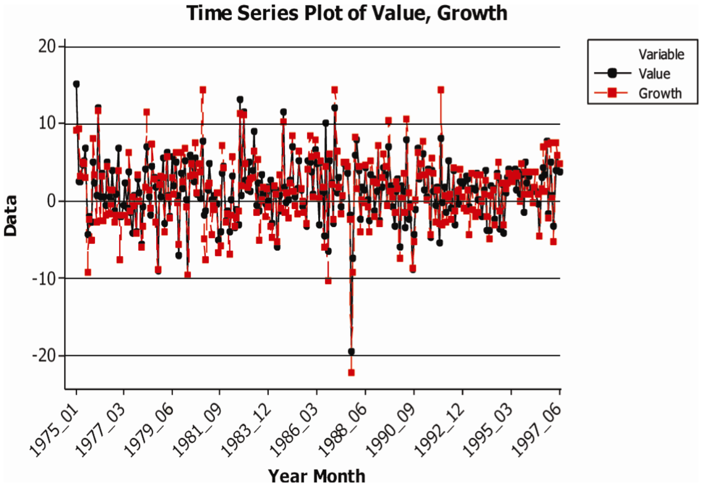
Value stocks have a higher mean and median return than growth stocks. In addition, the standard deviation is lower for value stocks compared to growth stocks. Historically, value stocks have resulted in a higher return on average and have exhibited less volatility compared to growth stocks. Based on the histograms, the distribution of returns for growth stocks is slightly more symmetric (after excluding the unusually low return seen in each group) than those for value stocks. In addition, the third quartile is higher for growth stocks. It is not clear which stock is a better investment since these data are aggregate (investments in individual stocks will be more variable) and historical performance does not necessarily predict future performance.

Descriptive Statistics: Value, Growth

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Value	270	0	1.439	0.243	3.988	–19.451	–0.734	1.448	3.800
Growth	270	0	1.245	0.283	4.644	–22.085	–1.743	1.312	3.858

Variable	Maximum
Value	15.242
Growth	14.480


Histogram of Value, Growth
Frequency
Value
Growth


Time Series Plot of Value, Growth
Data
Variable
Value
Growth
1975_01
1977_03
1979_06
1981_09
1983_12
1986_03
1988_06
1990_09
1992_12
1995_03
1997_06
Year Month