

## Chapter 1: Examining Distributions

1.1 The value of the coupon is computed by subtracting the DiscPrice from the RegPrice. It is quantitative because arithmetic operations, like the average value, would make sense.

	A	B	C	D	E	F	G
1	ID	Type	Name	Item	RegPrice	DiscPrice	Value
2	1	Italian	Domo's	Pizza	20	10	10
3	2	Italian	Mama Rita's	Pizza	20	12	8
4	3	BBQ	Smokey McSween's	Barbecue	30	17	13
5	4	BBQ	Smokey Grill	Ribs	20	11	9
6	5	Mexican	Dos Amigos	Tacos	16	8	8
7	6	Mexican	Holy Guacamole	Steak fajitas	13	8	5
8	7	Seafood	Sea Grille	Shrimp platter	20	11	9

1.2 The regular price for the Smokey Grill Ribs coupon is 20, the discount price is 11.

1.3 Who: The cases are coupons, there are 7 cases. What: There are 6 variables—ID, Type, Name, Item, RegPrice, and DiscPrice. Only RegPrice and DiscPrice have units in dollars. Why: The data might be used to compare coupons to one another to see which are better. We would not want to draw conclusions about other coupons not listed.

1.4 The cases are apartments. There are 5 variables: Monthly rent-quantitative, Fitness center-categorical, Pets allowed-categorical, # of Bedrooms-quantitative, Distance to campus-quantitative.

1.5 (a) If you were interested in attending a large college, you would want to know the number of graduates. (b) If you were interested in making sure you graduate, you would want to know the graduation rate.

1.6 (a) The cases are summer jobs. (b) Variables might include: position, company, hourly wage, whether the job is on or off campus, hours per week, other answers are possible. (c) position—categorical, company-categorical, hourly wage-quantitative, on or off campus-categorical, hours per week-quantitative, other answers are possible. (d) We could use a number as a label. The reason for doing so is there could be several jobs with the same company or position that you would need to differentiate from one another. (e) Who: part (a) answer, What: part (b) and (c), Why: To compile a list of available summer jobs and possibly compare them. We would not want to draw conclusions about other jobs not listed.

1.7 (a) The cases are employees. (b) Employee identification number—label, last name—label, first name—label, middle initial—label, department—categorical, number of years—quantitative, salary—quantitative, education—categorical, age—quantitative. (c) Sample data would vary.

	A	B	C	D	E	F	G	H	I
1	EIN	Last	First	Middle	Department	Years	Salary	Education	Age
2	001	Marley	Bob	M	Sales	4	45000	some college	34
3	002	Fisher	Margeret	A	Sales	8	54000	college degree	37
4	003	Marin	Jane	E	Admin	2	39000	high school	25

1.8 Answers will vary.

1.9 (a) Quantitative. (b) Quantitative. (c) Quantitative. (d) Quantitative. (e) Categorical. (f) Categorical. For all quantitative variables, numerical summaries would be meaningful; for categorical variables, numerical summaries are NOT meaningful.

1.10 Answers will vary. 1. Rate the customer service of the restaurant—quantitative 2. Is this your first visit to our restaurant—categorical 3. If not, how many times per month do you visit our restaurant—quantitative 4. Would you recommend our restaurant to a friend—categorical 5. Do you think our dish prices are expensive, about right, inexpensive—categorical 6. Rate the taste quality of food you ate today—quantitative. For all quantitative variables numerical summaries would be meaningful, for categorical variables, numerical summaries are NOT meaningful.

1.11 Answers will vary. 1. How many hours per week do you study—quantitative, hours 2. How many nights per week do you study usually—quantitative, nights 3. Do you usually study alone or with others—categorical 3. Do you feel like you study too much, about right, not enough—categorical.

1.12 Answers and reasons will vary. Examples include: current enrollment, average time to graduate, graduation rate, job placement percentage, etc.

1.13 (a) The states are the cases. (b) The name of the state is the label variable. (c) Number of students from the state who attend college—quantitative, number of students who attend college in their home state—quantitative. (d) Answers will vary. This would tell you which states have large percentages of students that like to stay “at home” versus small percentages, which indicate students’ preference to leave home to attend college.

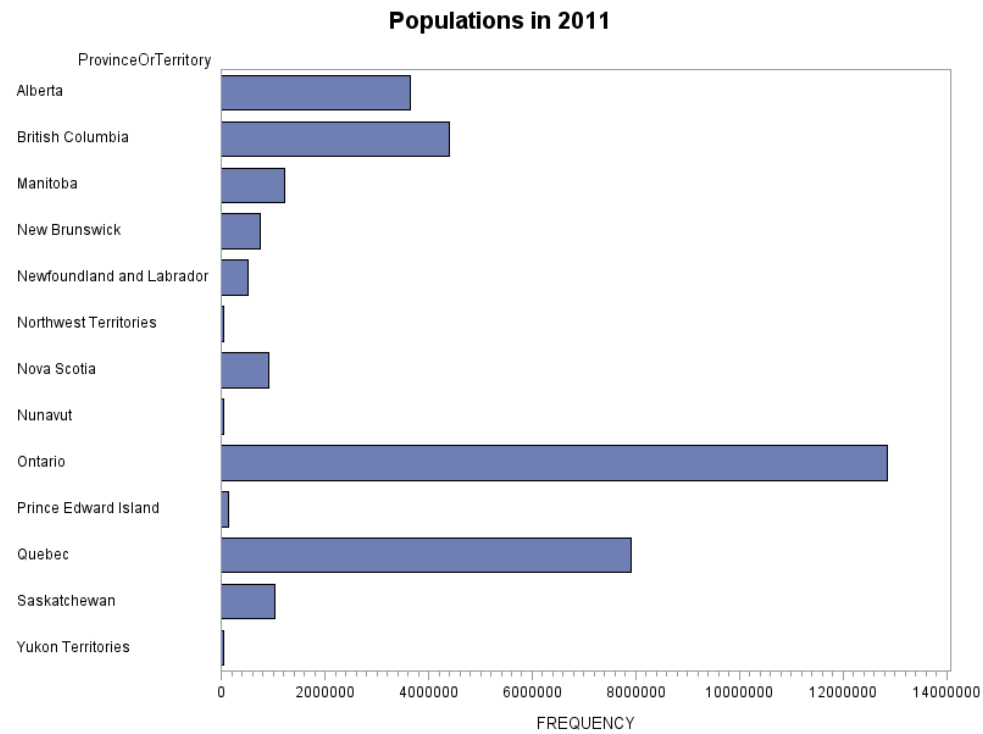
1.14 Each state could be divided as a percentage of the total of the nation’s fatalities to show state differences; the disadvantage is that states with more population would have a higher number of fatalities. Instead, each state’s fatalities could be divided by the state population to get a percentage for each state; this would be a better way to compare state-to-state rates of drunk driving fatalities.

1.15 Answers may vary. The pie chart does a better job because it shows the dominance of Google as a source, filling almost three-quarters of the pie.

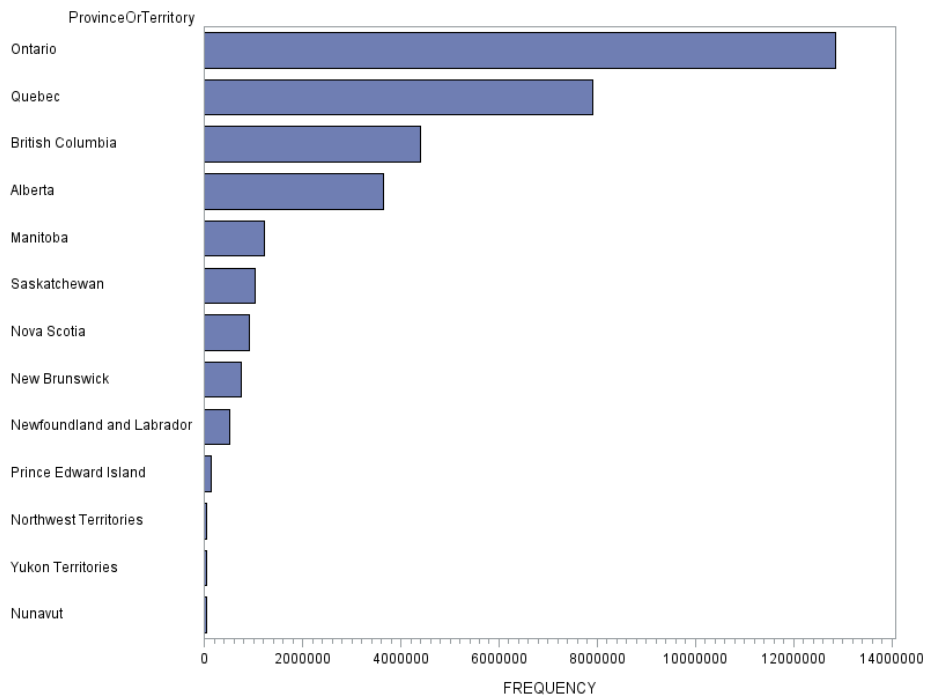
1.16 Answers may vary. It is probably a good idea to round; most of the time we just need an idea of what the data are telling us.

1.17 The Cost Centers would include, Parts and materials, Manufacturing equipment, Salaries, Maintenance, and Office lease. We need to include Office lease even though it gives more than 80%, because otherwise we would only have the top 75% according to the data. So, to get the other 5%, we need to put Office lease in, giving us 82.12% total.

1.18

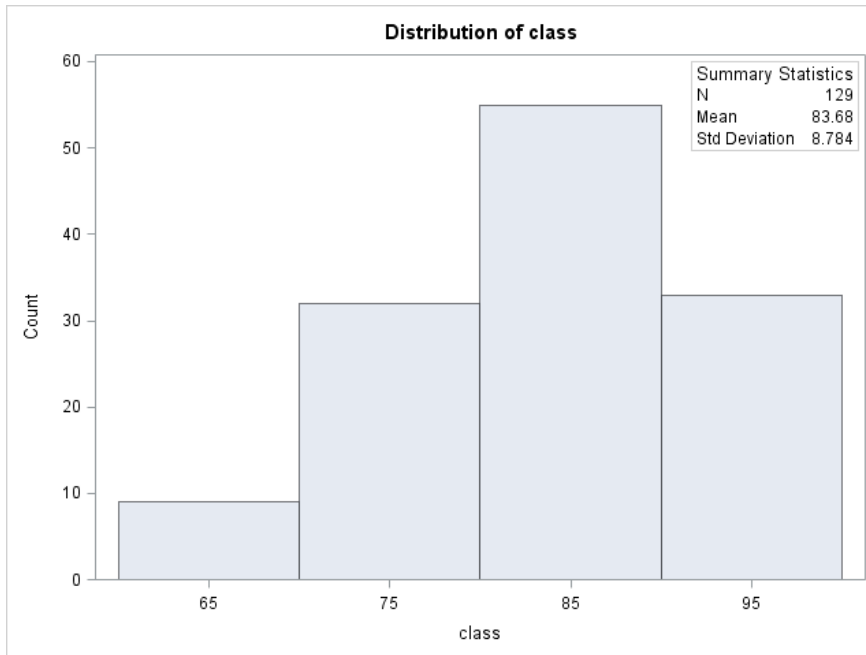


1.19 (a)



(b) Most people will prefer the Pareto because it emphasizes the largest categories.

1.20



1.21 Answers will vary. One solution is to have the highest range include 100, so  $90 < \text{score} \leq 100$ ,  $80 < \text{score} \leq 90$ , etc.

1.22 Answers will vary. One example is shown.

0	34
1	
2	2378
3	002345689
4	14579
5	
6	8

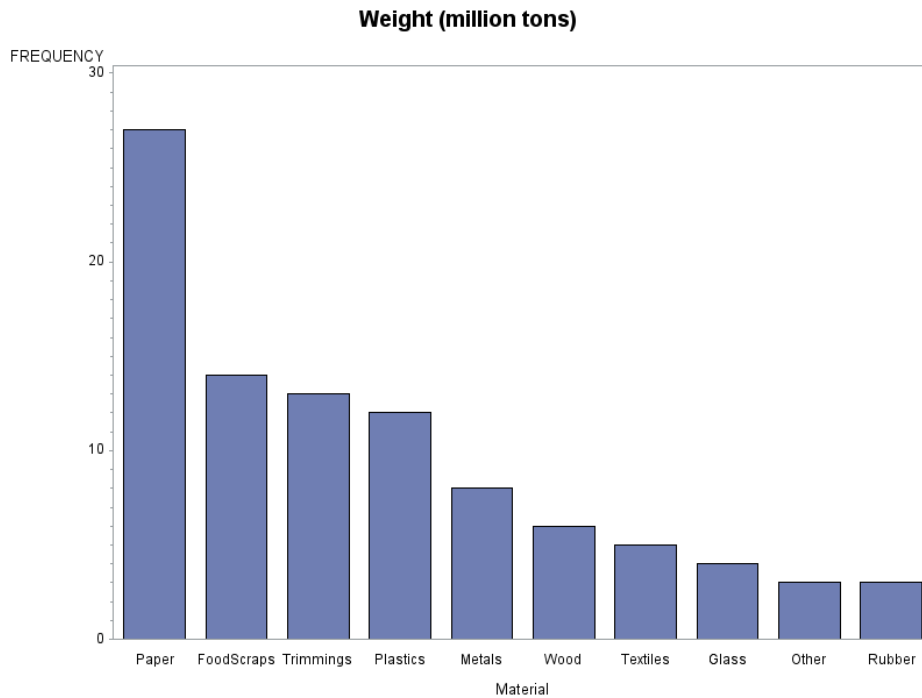
1.23 Answers will vary. One example is shown.

0	34
1	13679
2	235578
3	11256
4	157
5	45
6	8
7	2

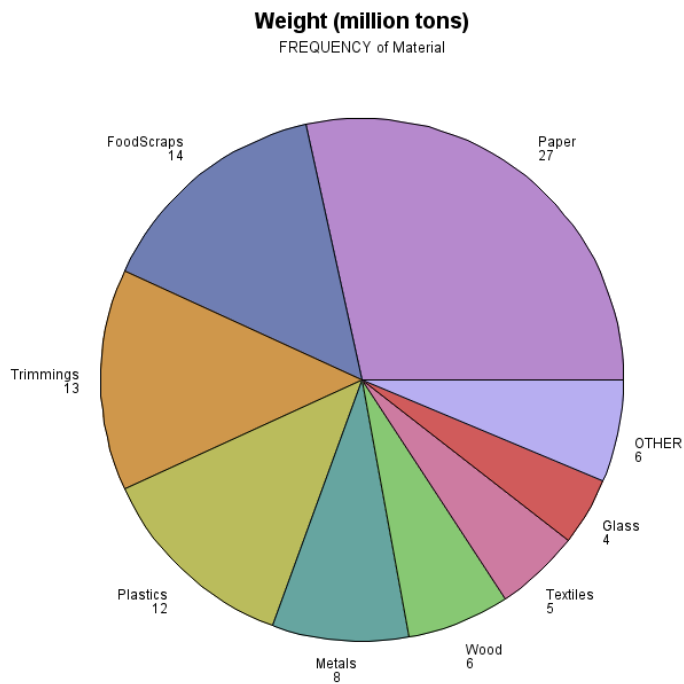
1.24 (a) T-bill interest rates were going up between 1960 and 1980, where they peaked; they have generally gone down since 1980 until now. They also have short intervals every couple of years, where they climb and fall. (b) During recessions the T-bill interest rates have generally plummeted.

1.25 (a) Histogram would be best to show. (b) Pareto chart would be the best to prioritize those characteristics that they liked best; pie chart might also be suitable. (c) A stemplot would be best because it is a small dataset; a histogram might also be suitable. (d) Pie chart is likely best in this situation to divide all the customers into groups from the whole; a Pareto or bar graph might also be suitable.

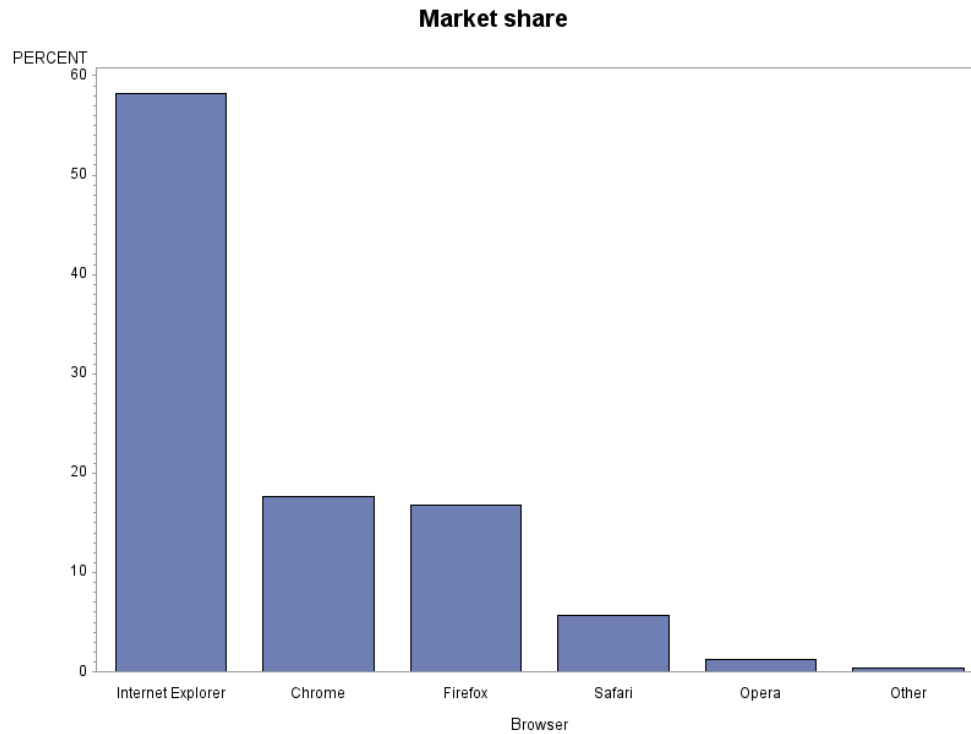
1.26 (a) The values are rounded. (b)



(c)



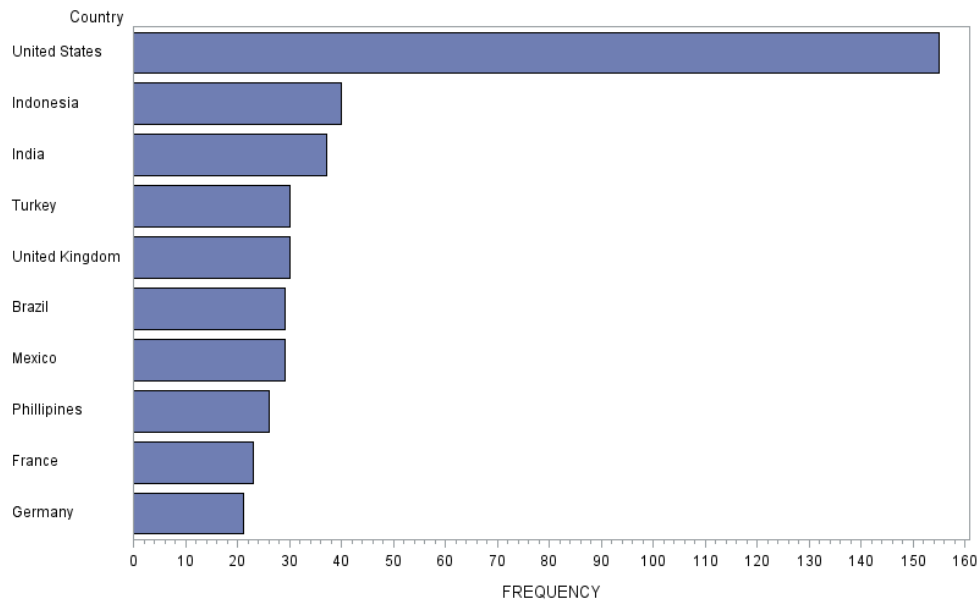
1.27 (a)



(b) Internet Explorer has by far the largest percentage of market share, followed by Chrome and Firefox. Other browsers have very little market share.

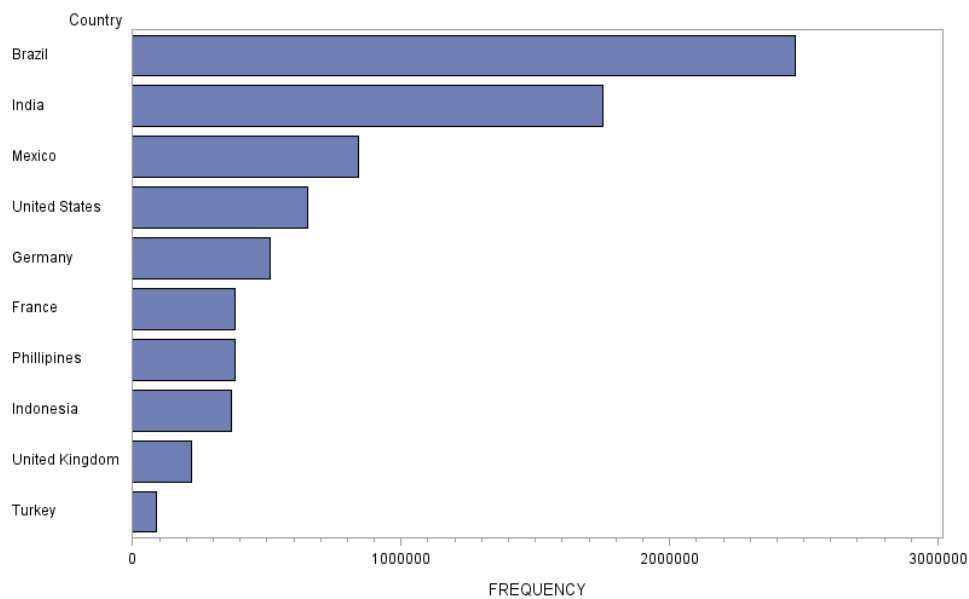
1.28 (a) Many more readers owned Brand A than Brand B. (b) A suitable measure is the percentage for each brand. Brand A is  $2942/13,376 = 0.2199$  or 21.99%. Brand B is  $192/480 = 0.4$  or 40%. Brand A is more reliable because a smaller percentage of owners of Brand A required a service call.

1.29 (a)

**Facebook users (in millions)**

(b) The United States is a clear outlier. It has 4 or 5 times as many Facebook users as the other countries, despite having a population smaller than some of the other countries. (c) The United States dominates; many other countries shown have similar amounts of Facebook users.

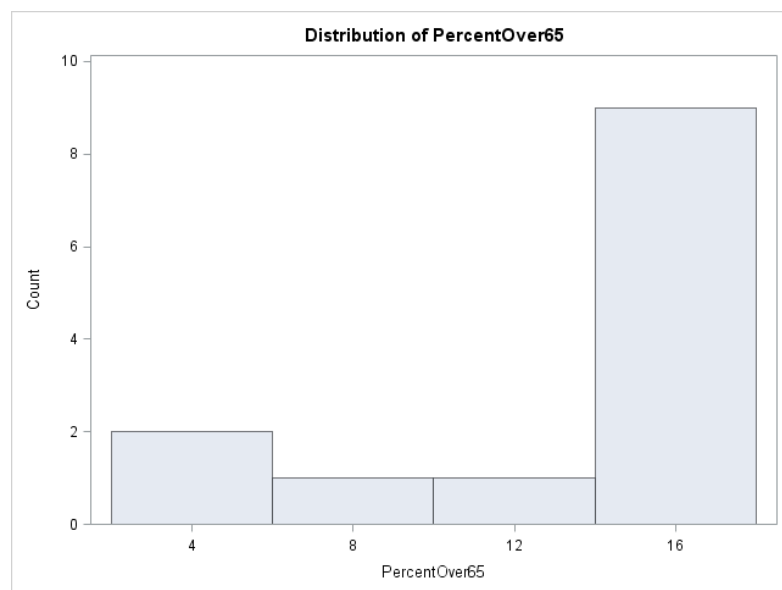
1.30 (a)

**Increase in users (in millions)**

(b) Brazil is the leading country in Facebook user growth, followed by India, then Mexico. (c) A stemplot would not be better because the data are categorical and represent the different countries better. (d) Countries with higher Facebook user growth show more online presence and would have potential for growth among online marketing and other online business ventures.

1.31 The distribution is fairly symmetric. The center is around 130 or 140. The range is between 85 and 182.

1.32 (a) Most provinces have similar percent over 65 (shown in the bar at 16 in the graph) but a few are unique and have much smaller percentages.



(b) A histogram shows the distribution amidst the various provinces. A stemplot could have also been used but likely would have been too crude.

1.33  $\bar{X} = 23.96$ .

1.34  $\bar{X} = 82.3$ .

1.35  $\bar{X} = 196.575$ .

1.36  $M = 84$ .

1.37  $M = 103.5$ .

1.38 The ordered list is: 2 4 5 5 5 5 6 6 7 8 10 11 12 13 16 17 19 19 24 25 32 38 49 53 (208).

$M = 12$ . Without the outlier the median is 11.5, with the outlier the median is 12. The outlier does not influence the median greatly.

1.39 (a)

1		34
2		00
3		7
4		



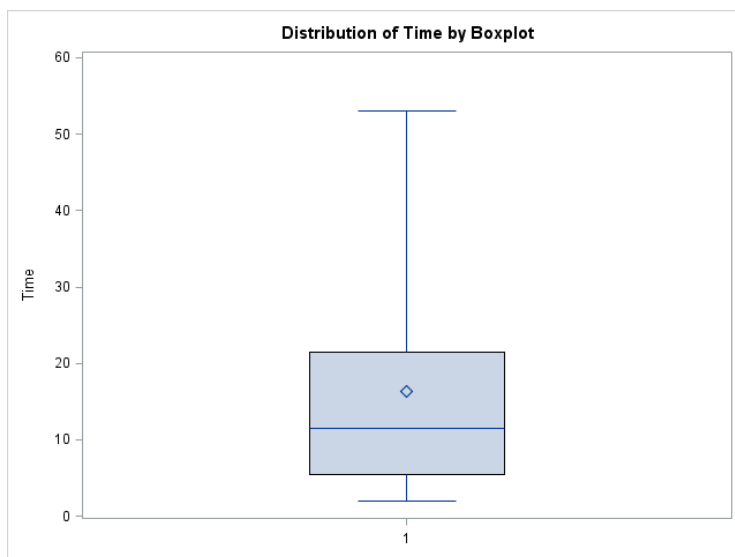
5		06
6		12456
7		8

(b) One group has 5.0 or more growth; the other group has 3.7 or less growth. (c) The mean growth rate is 4.66. Because the distribution is left-skewed, the mean is not a good measure of center. (d) The median growth rate is 5.6. Because the distribution is left-skewed, the median is a good measure of center. (e) The mean for group 1, 2.08, is much lower than the mean for group 2, 6.275. The split summaries are much better representations of the groups because there is no longer a large gap in the datasets. The gross domestic product of these countries is much better explained by the two distinct groups.

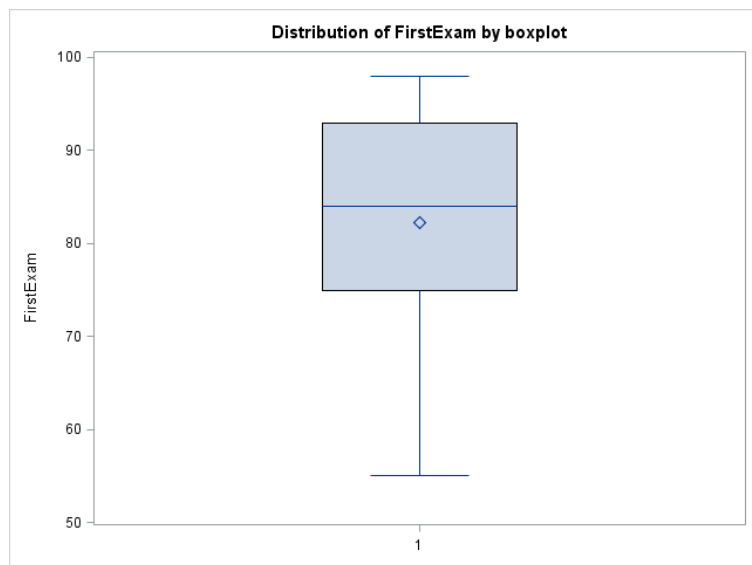
Analysis Variable : Growth for lower 5				
N	Mean	Std Dev	Minimum	Maximum
5	2.0800000	0.9628084	1.3000000	3.7000000
Analysis Variable : Growth for upper 8				
N	Mean	Std Dev	Minimum	Maximum
8	6.2750000	0.8119641	5.0000000	7.8000000

1.40 Answers will vary.

1.41 The time is right-skewed, with a long right tail. The mean is much higher than the median because of the skew. Answers will vary on preference.



1.42 A stemplot may be more helpful to see individual grades and determine possible cutoffs.



1.43 Without Suriname:  $s = 14.17$ . With Suriname:  $s = 40.77$ .

1.44 (a)

5	5
6	
7	05
8	035
9	0348

(b)  $s = 13$ . (c) The mean and standard deviation are not good numerical summaries for this dataset because the distribution is left-skewed.

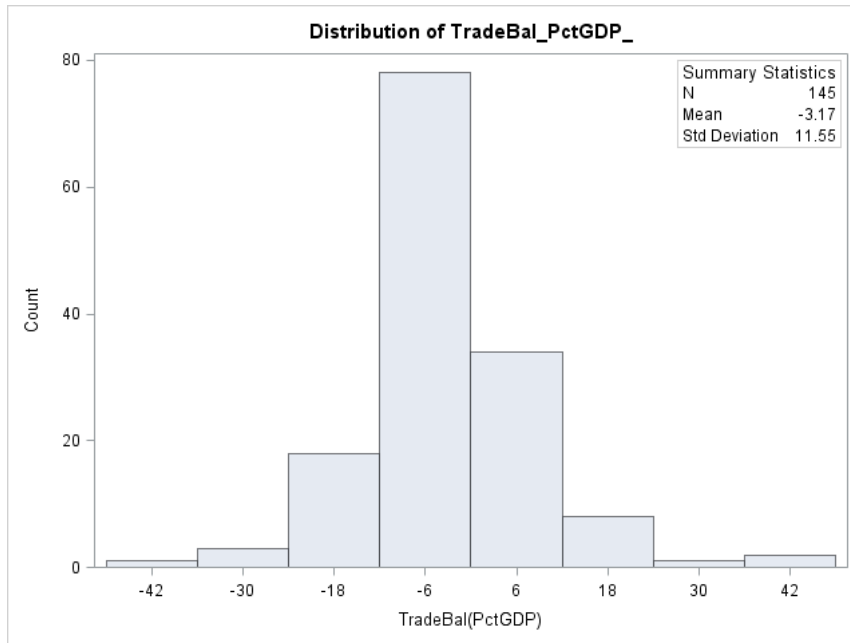
1.45 (a)  $\bar{X} = 196.575$ ,  $s = 342$ . (b)  $Min = 1$ ,  $Q1 = 54.5$ ,  $M = 103.5$ ,  $Q3 = 200$ ,  $Max = 2631$ . (c) The five-number summary is a better summary because the distribution is heavily skewed and has potential outliers.

1.46 (a)  $\bar{X} = 380,773$ ,  $s = 1,454,787$ . (b) Answers will vary. (c) Answers will vary.

1.47 (a)  $M = 27,035$ ,  $Q1 = 7103$ ,  $Q3 = 205,789$ . (c) Answers will vary.

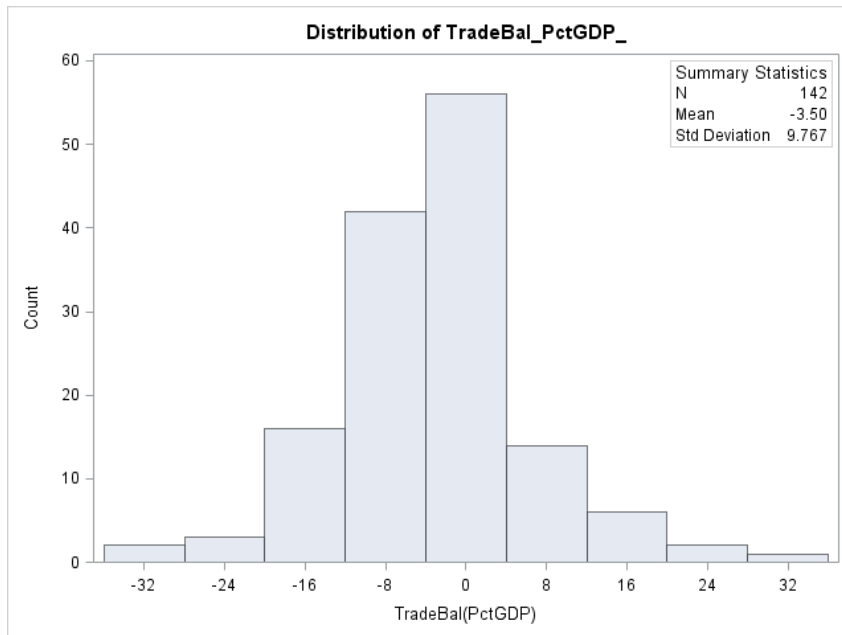
1.48 (a)  $\bar{X} = -3.173$ ,  $s = 11.554$ . (b)  $M = -3.3$ ,  $Q1 = -9.1$ ,  $Q3 = 1.0$ . (c) The distribution is symmetric; we know this because the mean and median are quite close. Also the distance between the median and the two quartiles is fairly close.

1.49 (a)



(b) Montenegro has a really low trade balance of  $-45.3$ . Kuwait,  $42.2$ , and Libya,  $40.7$ , have really high trade balances.

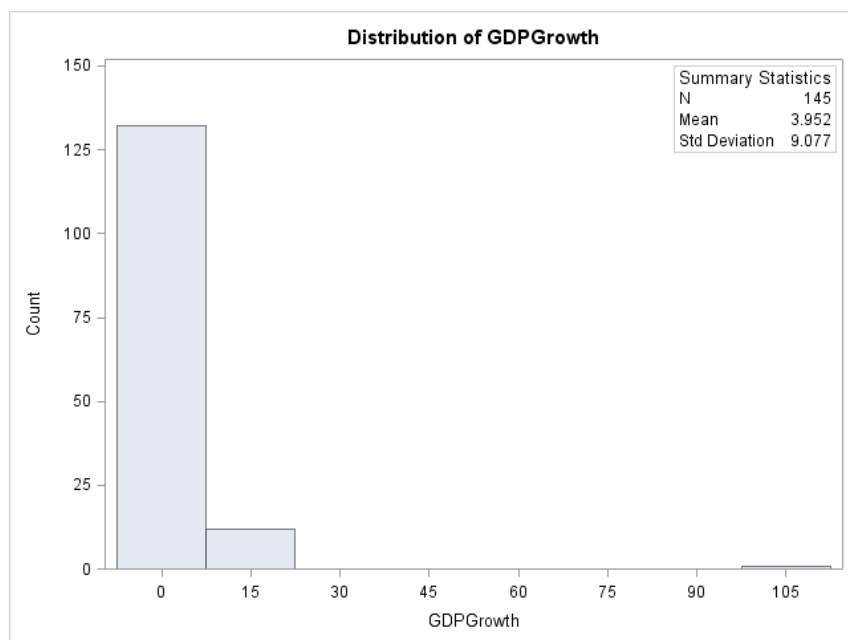
(c)



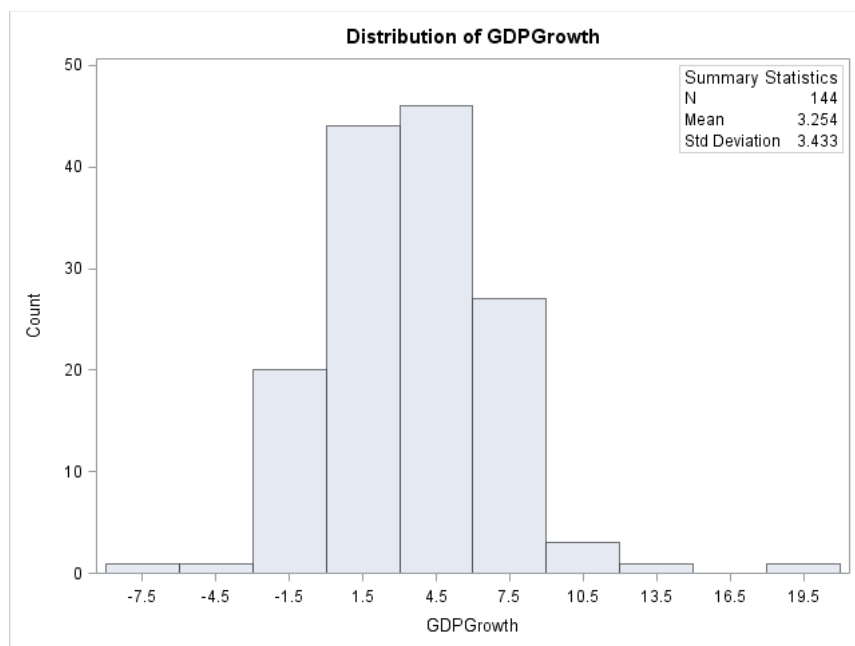
$\bar{X} = -3.50$ ,  $s = 9.767$ ,  $M = -3.3$ ,  $Q1 = -9.1$ ,  $Q3 = 0.9$ . The distribution and numerical summaries are almost identical before and after the outliers are removed.

(d) Overall, the distribution is very symmetrical, so that if some countries export a lot, there are other countries that import just as much. The mean and median trade balances are very close to 0. The outliers had almost no effect on the distribution or numerical summaries. Essentially, the outliers form longer tails on the curve.

1.50 (a) The distribution is strongly right-skewed with a very high outlier.



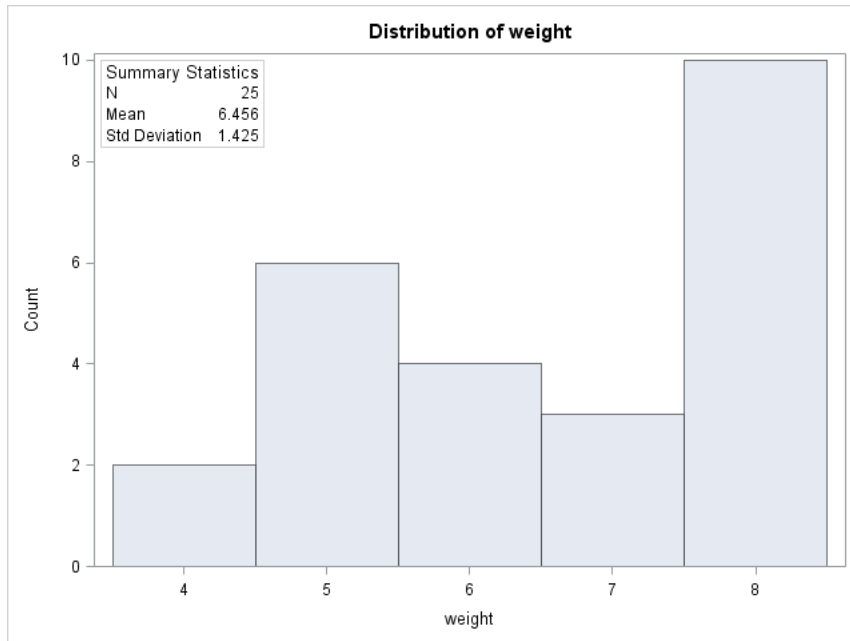
(b) Libya is the high outlier with 104.5 growth in GDP. (c) With Libya removed, the distribution is fairly symmetrical, centered at 3. The numerical summaries without Libya are:  $\bar{X} = 3.25$ ,  $s = 3.433$ ,  $M = 3.3$ ,  $Q1 = 0.85$ ,  $Q3 = 5.35$ .



(d) Most countries have positive growth in GDP, with a few having negative growth. Libya is an extreme outlier with 104.5 percent growth in GDP.

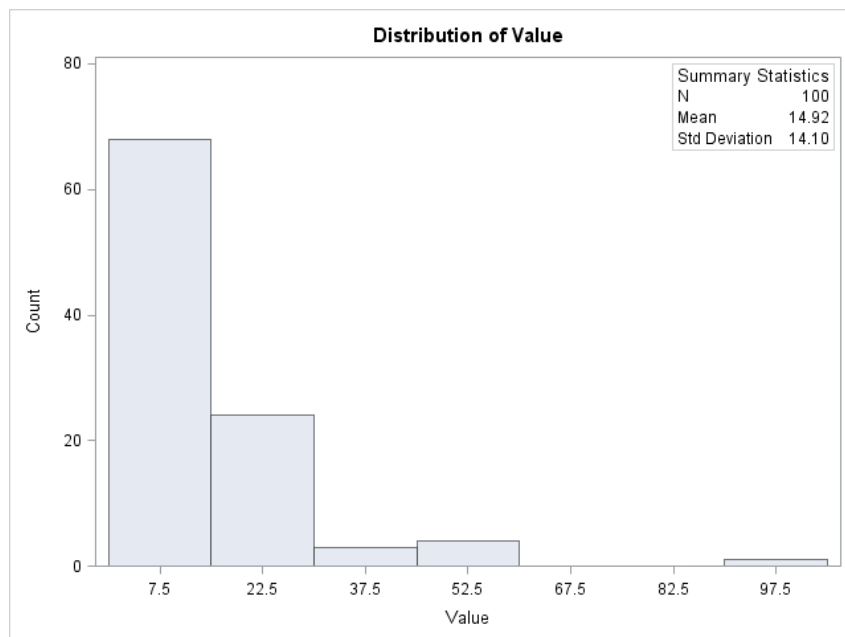
1.51 Answers will vary.

1.52 (a) Answers will vary. Because weight is quantitative and has a decent amount of observations ( $n = 25$ ), a histogram is a good choice. Mean and standard deviation are a good starting point for numerical summaries.



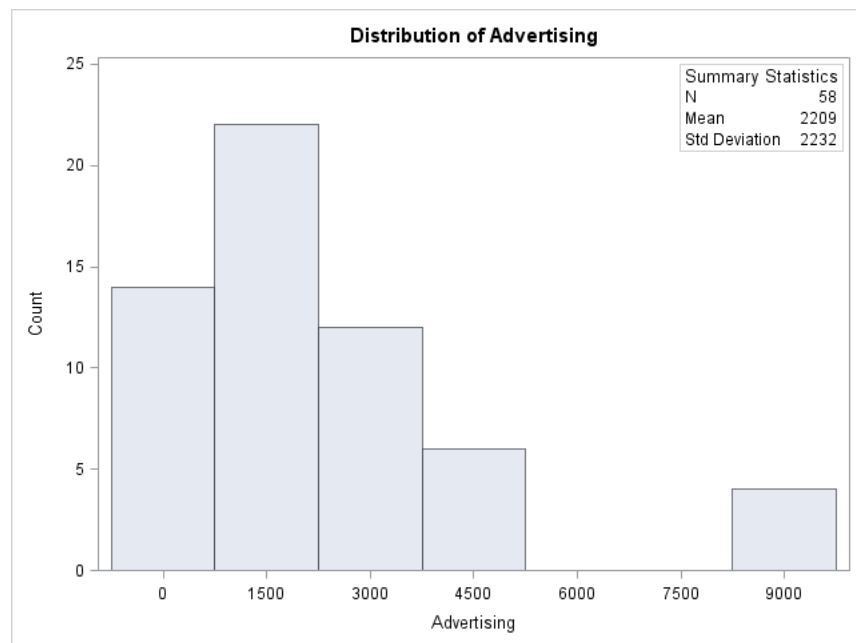
(b) Answers will vary. Now that we see the distribution is left-skewed, the choice of using the mean and standard deviation was not a good choice. Median and quartiles would have been a better choice. (c) Answers will vary. One possible break is between 5.3 and 6.0. The summaries for the groups should provide better summary statistics than the grouped data.

1.53 (a)



(b)  $\bar{X} = 14.92$ ,  $s = 14.1$ ,  $M = 9.6$ ,  $Q1 = 6.95$ ,  $Q3 = 18.05$ . (c) The distribution is strongly right-skewed, with several brands far more valuable than most others. This is shown in the numerical summaries, with 75% of brand values less than  $Q3 = 18.05$ . Additionally, the median brand value is only 9.6. The mean value is 14.92, substantially higher than the median, again indicating the skew. Thus, brands like Apple and those listed in the problem dwarf the competition.

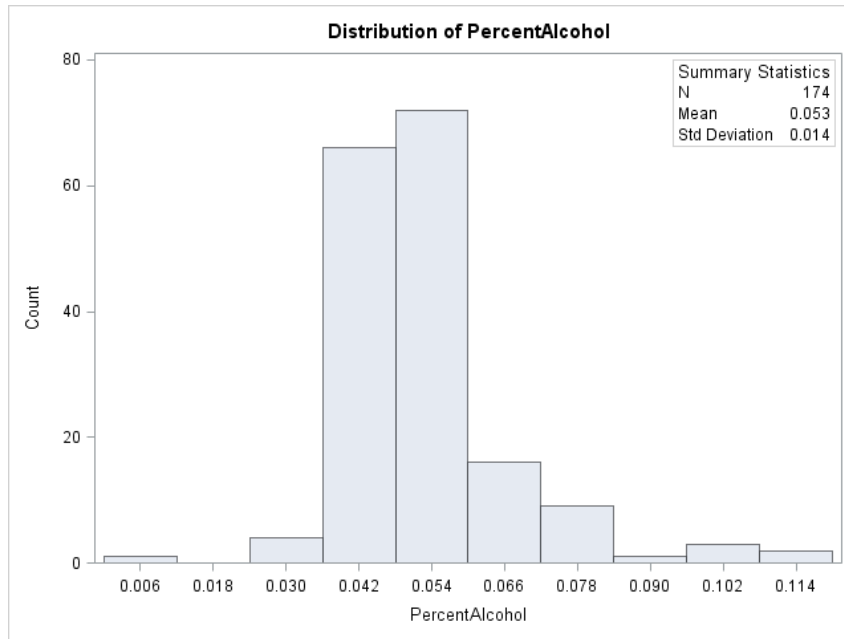
1.54 (a)



(b)  $\bar{X} = 2209$ ,  $s = 2232$ ,  $M = 1832.5$ ,  $Q1 = 772$ ,  $Q3 = 2798$ . (c) The distribution is somewhat right-skewed, but mostly due to several brands spending far more money on advertising than most others. Four companies (Gillette, L'Oréal, Pampers, and Lancome) spend more than double the amount on advertising than every other brand.

1.55 The data are right-skewed, which pull the mean, making it higher than the median.

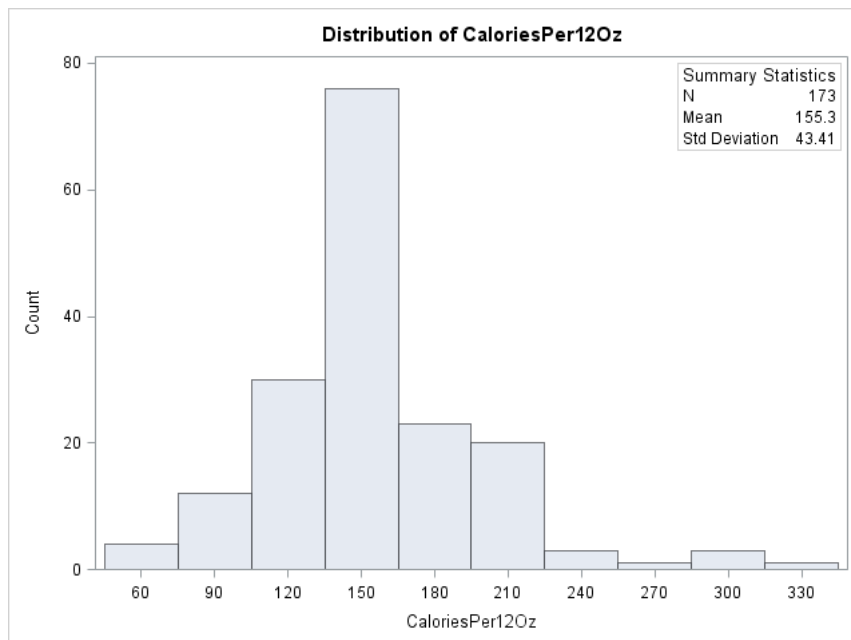
1.56 (a)  $\bar{X} = 0.053$ ,  $s = 0.014$ ,  $M = 0.0494$ ,  $Q1 = 0.045$ ,  $Q3 = 0.057$ .



(b) O'Doul's is the outlier with only 0.004 percent alcohol, it is unique because it is considered a form of non-alcoholic beer. (c) Answers will vary.

1.57 (a) With the outlier:  $\bar{X} = 0.0526$ ,  $M = 0.0494$ . Without the outlier:  $\bar{X} = 0.0529$ ,  $M = 0.0494$ . The values are nearly identical with and without the outlier. (b) With the outlier:  $s = 0.014$ ,  $Q1 = 0.045$ ,  $Q3 = 0.057$ . Without the outlier:  $s = 0.014$ ,  $Q1 = 0.045$ ,  $Q3 = 0.057$ . The values are nearly identical with and without the outlier. (c) Even though there is one outlier, its removal does not change the numerical summaries at all. This is partly due to the large sample and partly due to the fact that this outlier is not too far from the other observations, so that removing it doesn't have a huge effect on the analysis.

1.58 (a) The distribution of calories is fairly symmetric with a mean of 155.3.



(b) O'Doul's has one of the smallest amount of calories per 12 ounces, 70, but is not an outlier. (c) Answers will vary.

1.59 (a)  $Min = 8.5$ ,  $Q1 = 13.2$ ,  $M = 14.2$ ,  $Q3 = 14.8$ ,  $Max = 18.2$ . (b)  $IQR = 14.8 - 13.2 = 1.6$ ,  $Q1 - 1.5 \times IQR = 10.8$ . So, Utah with 9.5 percent and Alaska with 8.5 percent are low outliers.  $Q3 + 1.5 \times IQR = 17.2$ . So, Florida with 18.2 percent is a high outlier.

1.60 Applet, answers will vary.

1.61 Applet, answers will vary.

1.62 Applet, answers will vary.

1.63 The means and standard deviations are the same.  $\bar{X} = 7.5$ ,  $s = 2.03$ . The stemplots (rounded to 1 decimal) show very different distributions. Data A is strongly left-skewed with a couple possible low outliers; Data B is equally distributed between 5 and 9 but has one high outlier at 12.5.

Data A:

3	1
4	7
5	
6	1
7	3
8	1178
9	113

Data B:

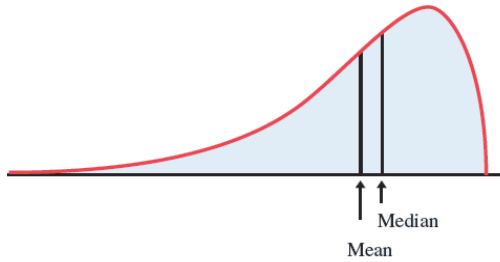
5	368
6	69
7	079
8	58
9	
10	
11	
12	5

1.64 (a)  $Min = 0.9$ ,  $Q1 = 3.0$ ,  $M = 4.95$ ,  $Q3 = 6.6$ ,  $Max = 14.7$ . (b) The mean is bigger than the median because the distribution is right-skewed.

1.65 (a)  $\bar{X} = \$100,625$ . All the employees except the owner make less than the mean.  $M = \$40,000$ . (b) The mean increases to \$105,625. The median does not change.

1.66 Answers will vary, one example is shown below. Because the distribution is left-skewed, the mean will be farther out in the long tail than the median.



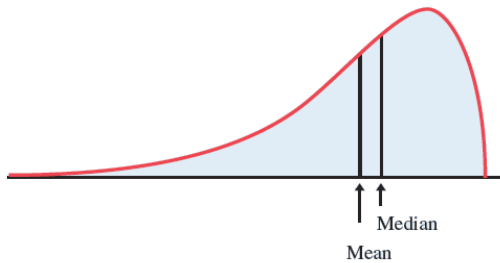


1.67 (a) Picking the same number for all four observations results in a standard deviation of 0. (b) Picking 10, 10, 20, and 20 results in the largest standard deviation = 5.77. (c) For part (a), you may pick any number as long as all observations are the same. For part (b), only one choice provides the largest standard deviation.

1.68 (a)  $\bar{X} = 16$ ,  $s = 7.51$ . (b)  $\bar{X} = 15.5$ ,  $s = 5.2$ . (c) Adding 10 more values near the mean pulled the mean halfway toward the imputed value, from 16 to 15.5. It also drastically reduced the standard deviation from 7.51 to 5.2.

1.69 The 5% trimmed mean is 12.78. The original mean was 14.92. The 5% trimmed mean is not as influenced by the large outliers as the original mean,

1.70

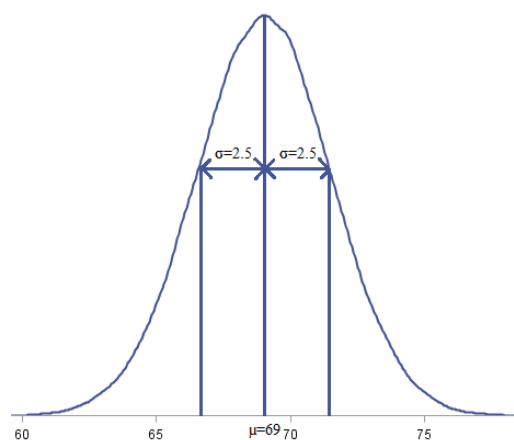


1.71 Answers will vary. Verify that the density curve is symmetric.

1.72 (a) The area under the square between 0.7 and 1 is 0.3 or 30%. (b) 0.4 or 40%. (c) 0.25 or 25%. (d) The distribution has length 1 and height 1, so the total area is also 1. (e)  $\mu = 0.5$ .

1.73 (a) The mean is at point C, the median is at point B. (b) The mean and median are both at point A. (c) The mean is at point A, the median is at point B.

1.74



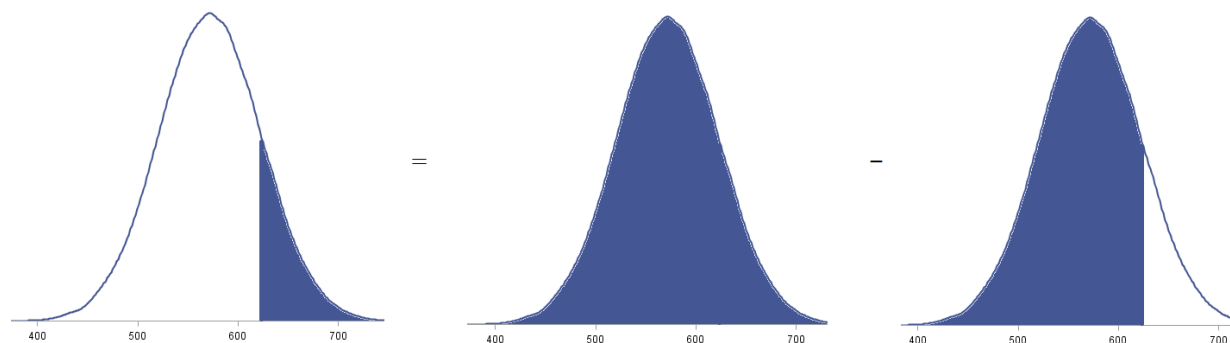
1.75 (a) 2.5%. (b) Between 64 and 74 inches. (c) 16%.

1.76 According to the rule, 95% of students will fall between  $\mu \pm 2\sigma$ . Therefore 95% of students have scores between 470 and 674.

1.77 According to the rule, 99.7% of students will fall between  $\mu \pm 3\sigma$ . Therefore 99.7% of students have scores between 419 and 725.

1.78 For Emily,  $Z = \frac{650 - 500}{100} = 1.5$ . For Michael,  $Z = \frac{28 - 18}{6} = 1.67$ . Michael has the higher standardized score.

1.79 For  $X = 620$ ,  $Z = \frac{620 - 572}{51} = 0.94$ , the proportion less than 620 is the area to the left, which is 0.8264. For the proportion greater than or equal to 620, we calculate  $1 - 0.8264 = 0.1736$ .

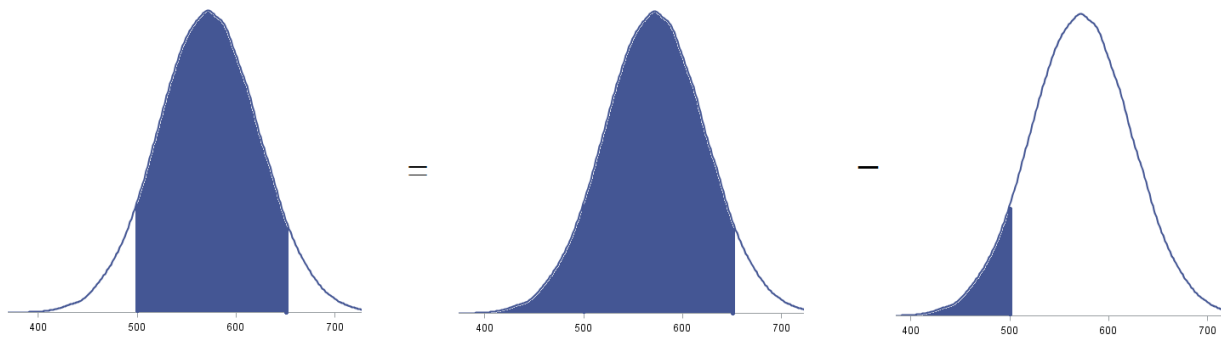


Area greater than 620 = Entire area under the Normal - Area less than 620

1.80 For  $X = 650$ ,  $Z = \frac{650 - 572}{51} = 1.53$ ; the area to the left of this is 0.9370. For  $X = 500$ ,

$Z = \frac{500 - 572}{51} = -1.41$ ; the area to the left of this is 0.0793. Subtracting gives  $0.9370 - 0.0793 = 0.8577$ .

So the proportion between 500 and 650 is 0.8577.



Area between 500 and 650 = Area to the left of 650 - Area to the left of 500

1.81 To get the top 25% of students, we need to solve for the 75th percentile. The corresponding  $Z$  is 0.67. So  $X = 572 + 51(0.67) = 606.17$ .

1.82 If 70 percent score above  $x$ ,  $x$  is the 30th percentile. The corresponding  $Z$  is  $-0.52$ . So  $x = 572 + 51(-0.52) = 545.48$ .

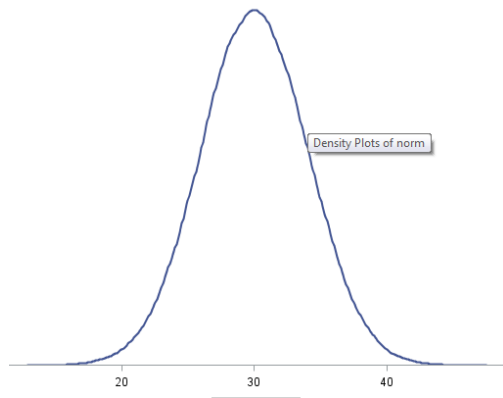
1.83 (a) The points fall below the  $45^\circ$  line; they form a straight line at first but then, near the right side, begin to increase steeply, indicating the right-skew. (b) It is likely part of a very long tail as it aligns perfectly with the curvature; see the Normal quantile plot. (c) The upper portion of the Normal quantile plot can show if a right-skew exists; specifically, if the points on the plot get steeper than a  $45^\circ$  line, this indicates a right-skew or long right tail.

1.84 (a) The points follow a  $45^\circ$  line quite closely. (b) Overall the plot indicates a Normal distribution. Only near the bottom of the plot do we see the points deviate from the  $45^\circ$  line; this should not be seen as non-Normality though, as most real datasets, although Normally distributed, will have small variations like this (usually due to random sampling).

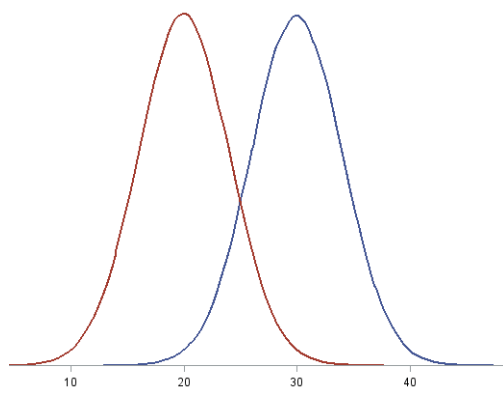
1.85 (a) This is not correct for categorical variables. A density curve is a mathematical model for the distribution of a quantitative variable. (b) This is not correct for area under the curve. The area under the curve for a density curve is always *equal to 1*. (c) A density curve cannot go below the  $x$  axis. If a variable can take only negative values, then the density curve for its distribution will still lie entirely above the  $x$  axis.

1.86 (a) This statement is not true for all distributions, only Normal distributions. The 68-95-99.7 rule applies to all Normal distributions. (b) This statement is not true for Normal distributions as they can certainly take negative values. A Normal distribution can take any value between  $-\infty$  and  $+\infty$ . (c) This statement is true for right-skewed distributions only. For a symmetric distribution, the mean will be equal to the median.

1.87 (a)

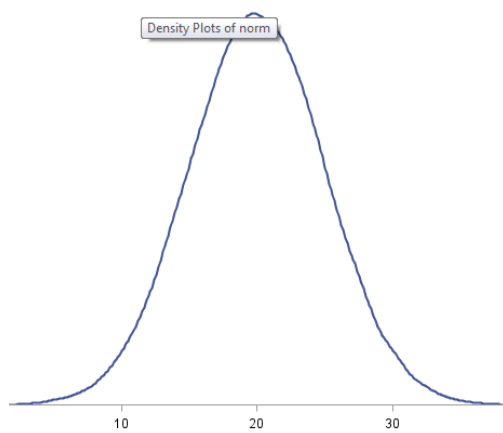


(b)

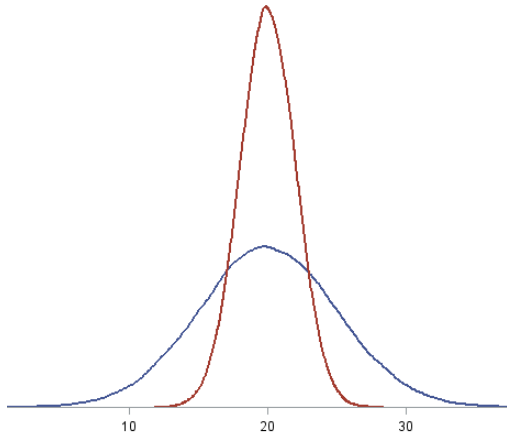


(c) The curve shifts to the left or right, but the spread remains the same.

1.88 (a)

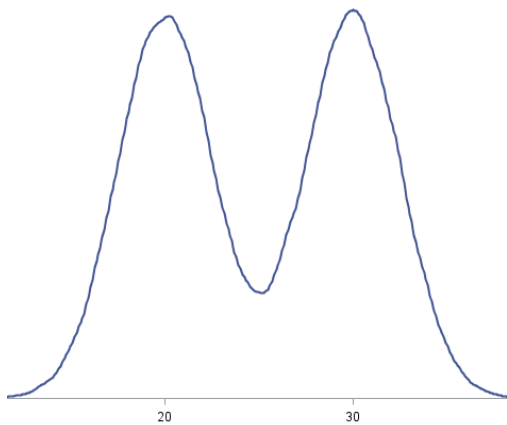


(b)

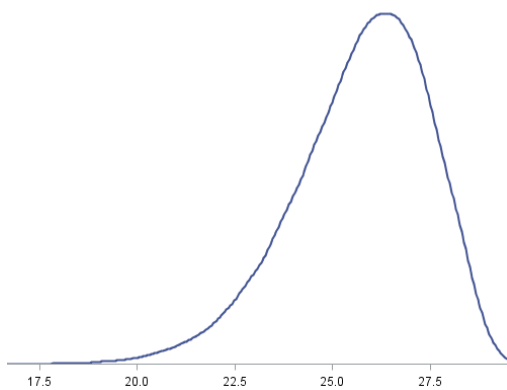


(c) The curve narrows or widens but remains centered at the same point (at the mean).

1.89 (a)



(b)



1.90 (a)  $\bar{X} = 380,773$ ,  $s = 1,454,787$ . (b) 68% should be between  $-1,074,014$  and  $1,835,560$ . 95% should be between  $-2,528,801$  and  $3,290,347$ . 99.7% should be between  $-3,983,588$  and  $4,745,134$ . (c) From the actual data, 180 out of 189 or 95.24% are within one standard deviation, 185 out of 189 or 97.88% are within two standard deviations, and 186 out of 189 or 98.41% are within three standard deviations. (d) The data are not Normally distributed because the percentages are not close to what the rule indicates.

1.91 (a) According to the rule, 68% of women speak between 5232 and 23,362 words per day, 95% of women speak between -3833 and 32,427 words per day, and 99.7% of women speak between -12,898 and 41,492 words per day. (b) This rule doesn't seem to fit because you can't speak a negative amount of words per day; therefore, the percentages don't make sense. (c) According to the rule, 68% of men speak between 5004 and 23,116 words per day, 95% of men speak between -4052 and 32,172 words per day, and 99.7% of men speak between -13,108 and 41,228 words per day. Similar to the women, the rule doesn't seem to fit because, again, you can't speak a negative amount of words per day, so the percentages don't make sense. (d) The data do show that the mean number of words spoken per day is higher for women than for men, but it is very close; in addition, the standard deviations are nearly the same, making the intervals very close. To put it in perspective, the women only speak about 1–2% more words per day on average. That small of a difference could just be due to chance.

1.92 (a) According to the rule: 68% of women speak between 8489 and 20,919 words per day, 95% of women speak between 2274 and 27,134 words per day, and 99.7% of women speak between -3941 and 33349 words per day. This rule doesn't seem to fit because you can't speak a negative amount of words per day so the percentages don't make sense. According to the rule: 68% of men speak between 7158 and 22,886 words per day, 95% of men speak between -706 and 30,750 words per day, and 99.7% of men speak between -8570 and 38,614 words per day. Similar to the women, the rule doesn't seem to fit because again you can't speak a negative amount of words per day so the percentages don't make sense. These data show that the mean number of words spoken per day is higher for men than for women, which is contrary to the conventional wisdom. (b) It is possible that women and men from Mexico speak more words per day than women and men in the United States.

1.93 (a) Using  $Z = (X - \mu)/\sigma$  or  $Z = (X - 72)/10$ , the standardized values are: -1, 2.1, -1.8, 0.4, 0.1, 2.6, -0.8, -1.7, 0.8, -0.1. (b) For the top 15% we need the 85th percentile; table A gives  $Z = 1.04$ . (c) Only two scores have standardized values bigger than 1.04, the 93 and the 98.

1.94 (a) The corresponding percentiles are: 5, 20, 55 and 85. Using Table A, the 5th percentile is  $Z = -1.64$ . The 20th percentile is  $Z = -0.84$ . The 55th percentile is  $Z = 0.13$ . The 85th percentile is  $Z = 1.04$ . For a grade of F: A Z score of -1.64 and below. For a grade of D: Z scores between -1.65 and -0.84. For a grade of C: Z scores between -0.84 and 0.13. For a grade of B: Z scores between 0.13 and 1.04. (b) Converting the Z scores from part (a) using  $X = \mu + \sigma(Z)$  or  $X = 72 + 10(Z)$ , the actual values are 55.6, 63.6, 73.3, and 82.4. (c) Answers will vary.

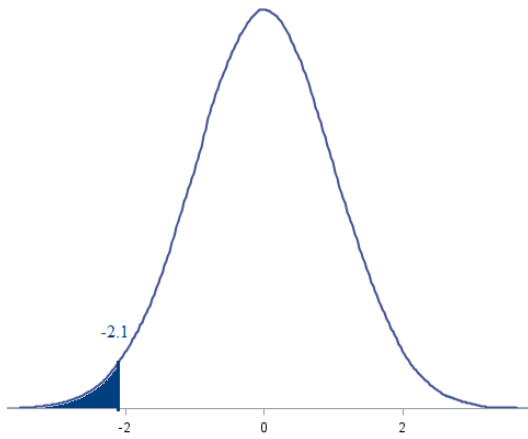
1.95 Answers will vary. The wider curve has a standard deviation of about 0.4. The narrower curve has a standard deviation about 0.2.

1.96 (a) Answers will vary. (b) Plots should be shown for each dataset in part (a). (c) The right-skewed distribution will have a Normal quantile plot with the points below the 45° line, they form a straight line at first but then near the right side begin to increase steeply, indicating the right-skew. The left-skewed distribution will have a Normal quantile plot with the points above the 45° line, they form a straight line at the top right of the graph but near the left side begin to decrease rapidly, indicating the left-skew. The symmetric and mound-shaped distribution will have a Normal quantile plot that has the data points that fall around the 45° line.

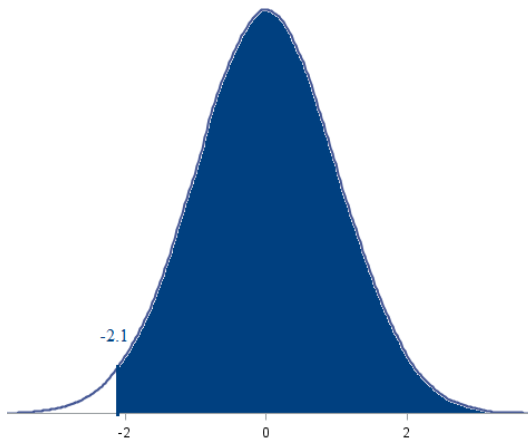
1.97 (a) Using the 95% rule we go  $2\sigma$ , or 32, from the mean in each direction.  $\mu - 2\sigma = 266 - 32 = 234$ .  $\mu + 2\sigma = 266 + 32 = 298$ . So 95% of pregnancies last between 234 and 298 days. (b) The shortest 2.5% of pregnancies fall 2 standard deviations below mean, or  $\mu - 2\sigma = 266 - 32 = 234$ . So the shortest pregnancies are 234 days or less.

1.98 Answers will vary. The uniform should have most bars around the same height, much different from a Normal distribution that will form a bell curve.

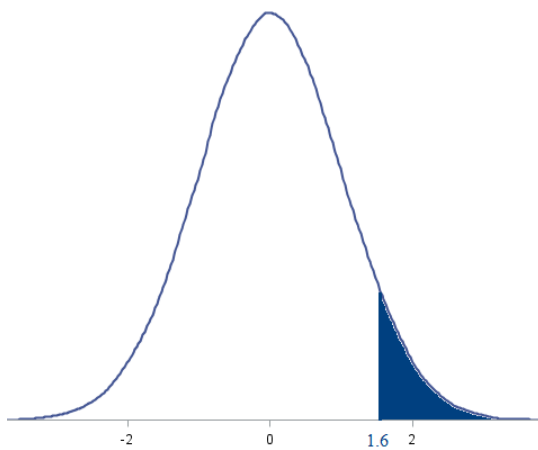
1.99 (a) 0.0179.



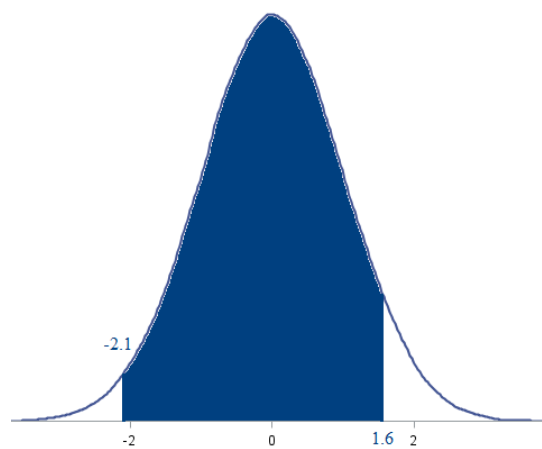
(b) 0.9821.



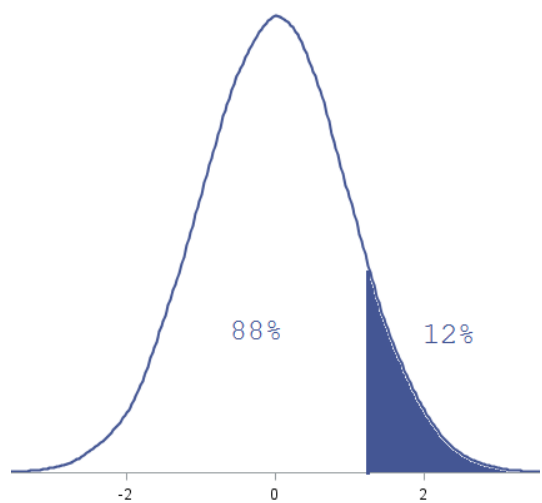
(c) 0.0548.



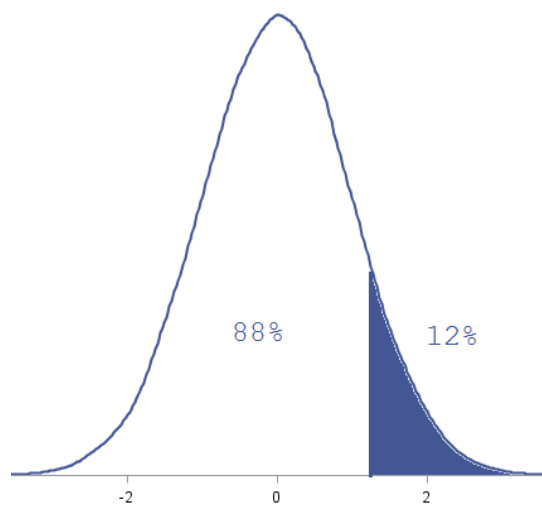
(d)  $0.9452 - 0.0179 = 0.9273$ .



1.100 (a) This is the 88th percentile; using table A gives  $Z = 1.17$  or  $1.18$ .

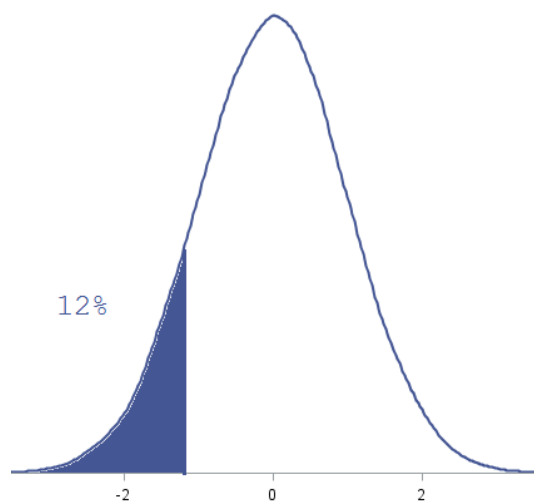


(b) This is the 88th percentile; using table A gives  $Z = 1.17$  or  $1.18$ .

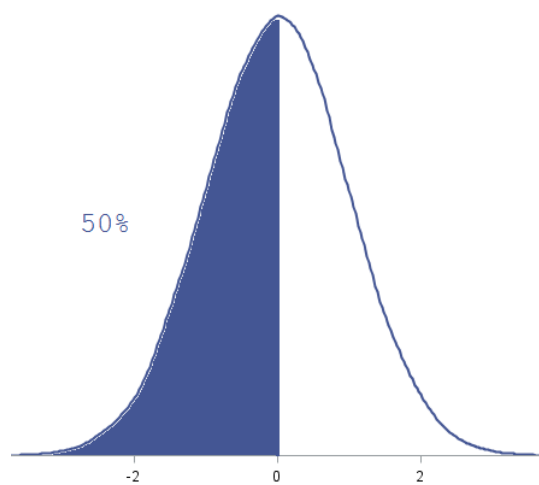


(c) This is the 12th percentile; using table A gives  $Z = -1.17$  or  $-1.18$ .





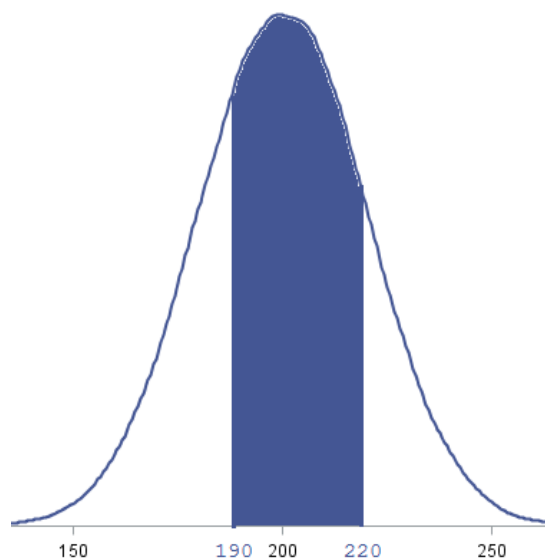
(d) This is the 50th percentile; using table A gives  $Z = 0$ .



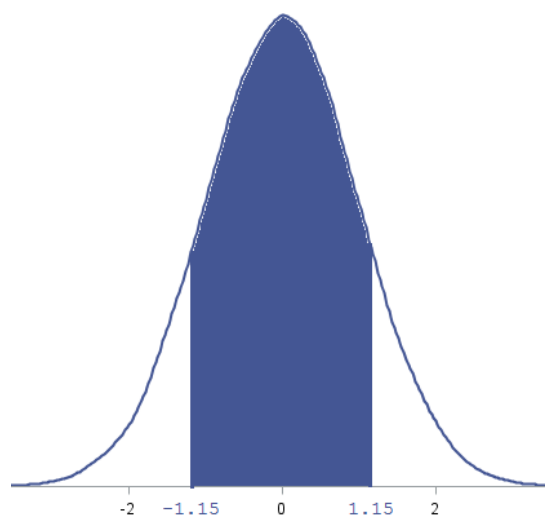
1.101 (a) For 220,  $Z = \frac{220 - 200}{20} = 1$ ; the area to the left of this is 0.8413. For 190,

$Z = \frac{190 - 200}{20} = -0.5$ ; the area to the left of this is 0.3085. Subtracting gives  $0.8413 - 0.3085 = 0.5328$ .

So the proportion between 190 and 220 is 0.5328.



(b) To solve, we need either the 12.5 or 87.5 percentile.  $200 - x$  corresponds with the 12.5 percentile  $\rightarrow Z = -1.15$ .  $200 + x$  corresponds with the 87.5 percentile  $\rightarrow Z = 1.15$ . So,  $1.15 = \frac{(200 + x) - 200}{20}$ ; solving gives  $x = 23$ .



1.102 (a)  $Z = \frac{240 - 266}{16} = -1.625$ . Using  $-1.62$  gives 0.0526 or 5.26% of pregnancies last fewer than

240 days. (The answer is 0.0516 if using  $-1.63$ .) (b) For 270,  $Z = \frac{270 - 266}{16} = 0.25$ ; the area to the left of

this is 0.5987. For 240,  $Z = \frac{240 - 266}{16} = -1.625$ ; the area to the left of this is 0.0516. Subtracting gives

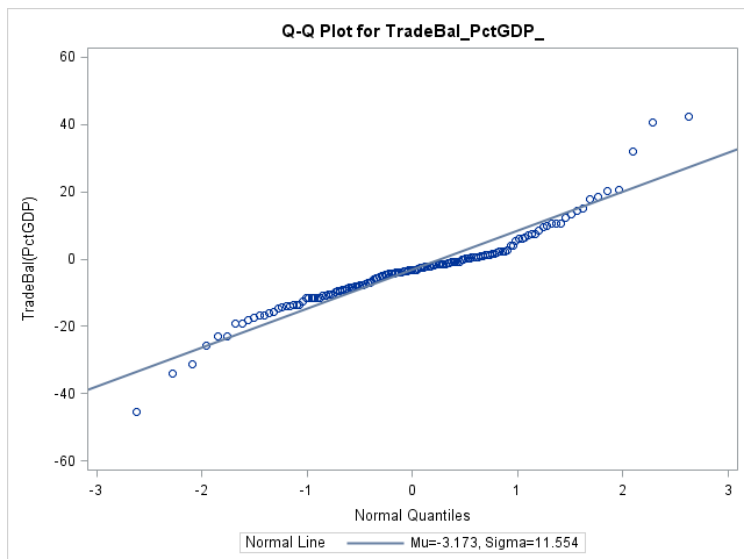
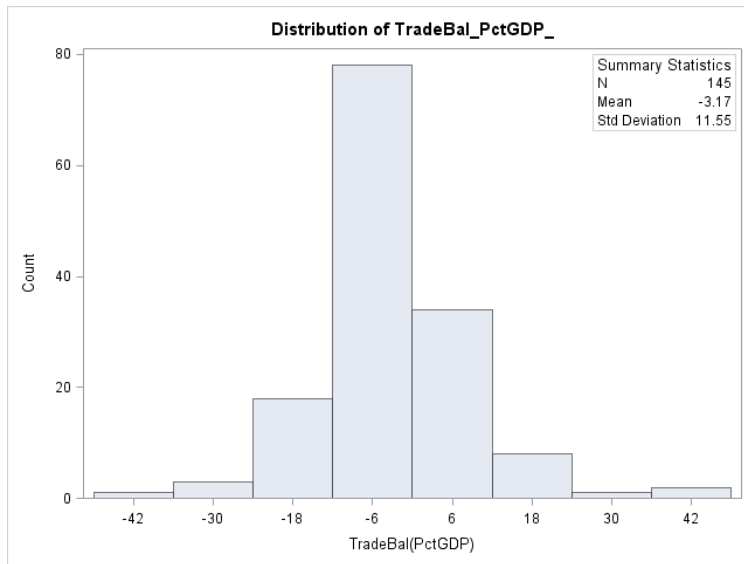
$0.5987 - 0.0516 = 0.5471$ . So 54.71% of pregnancies last between 240 and 270 days. (c) This is the 80th percentile; using table A gives  $Z = 0.84$ .  $X = \mu + \sigma(Z) = 266 + 16(0.84) = 279.44$ . The longest pregnancies last 279.44 days or longer.

1.103 (a) The area to the left of the first quartile is 25%. The corresponding  $Z$  is  $-0.67$ . The area to the left of the third quartile is 75%. The corresponding  $Z$  is  $0.67$ . (b) For the first quartile,  $X = \mu + \sigma(Z) = 266 + 16(-0.67) = 255.28$ . For the third quartile,  $X = \mu + \sigma(Z) = 266 + 16(0.67) = 276.72$ .

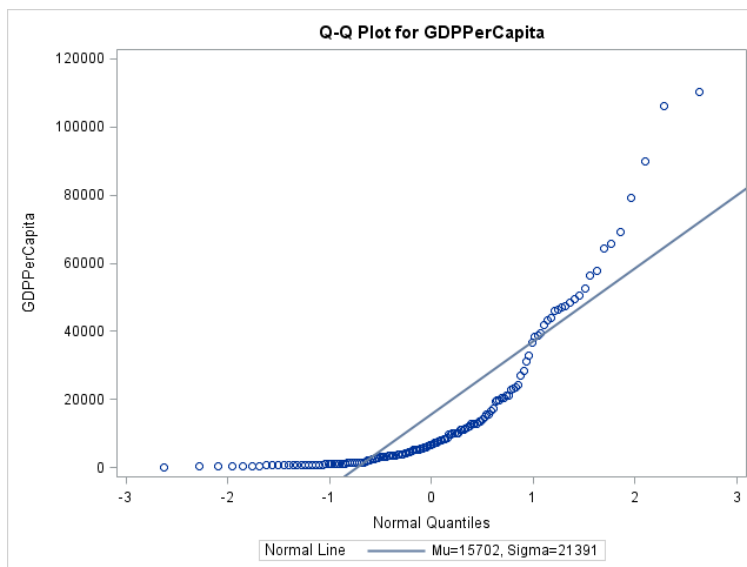
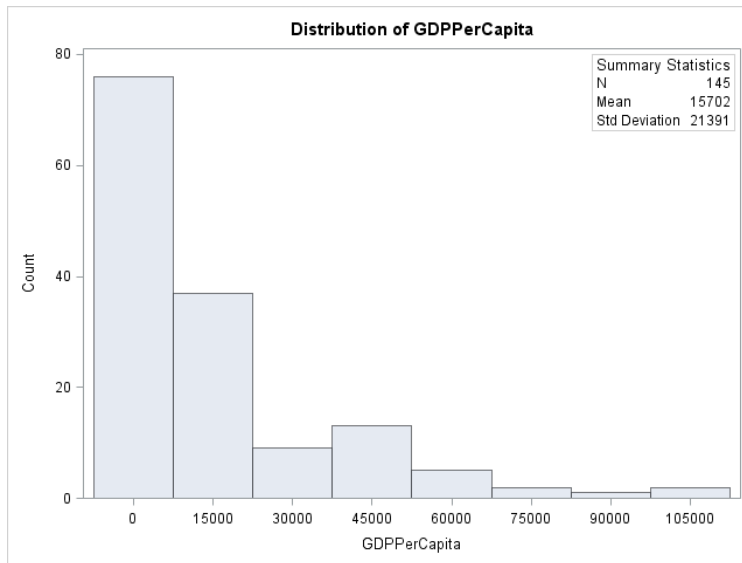
1.104 (a) For the 10th percentile, using table A gives  $Z = -1.28$ . For the 90th percentile,  $Z = 1.28$ . (b) For the first decile,  $X = \mu + \sigma(Z) = 9.12 + 0.15(-1.28) = 8.928$ . For the last decile,  $X = \mu + \sigma(Z) = 9.12 + 0.15(1.28) = 9.312$ .

1.105 Answers will vary.

1.106 While the histogram looks fairly Normal, the Normal quantile plot suggests some slight deviation from Normality, with somewhat heavy tails.

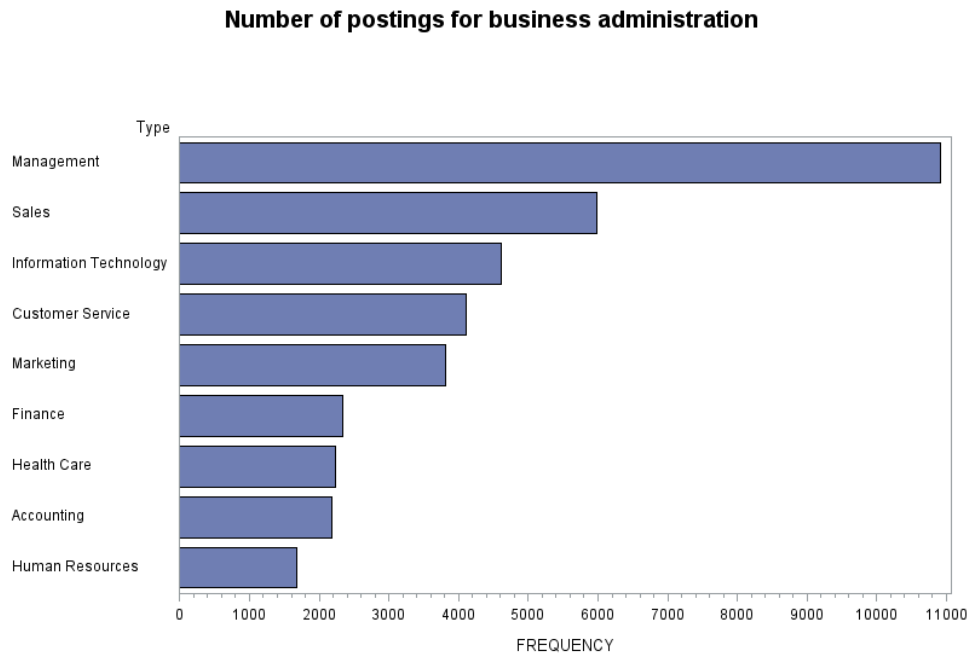


1.107 (a)



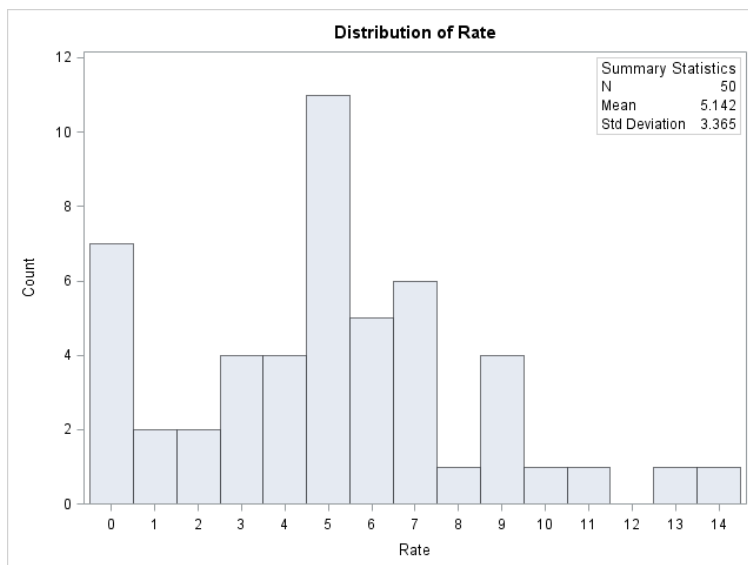
(b) The distribution is right-skewed; this is also shown in the Normal quantile plot. (c) Answers will vary.

1.108 The most prevalent job for those who have a business degree is in Management, with that category having nearly twice as much as the next closest category. Sales is the second most prevalent job available, followed by Information Technology, Customer Service, and Marketing. The limitations on using these data is that it is likely to change over time, potentially even day to day. It is also restricted to the classification specifications of the particular website which may classify jobs differently than others.



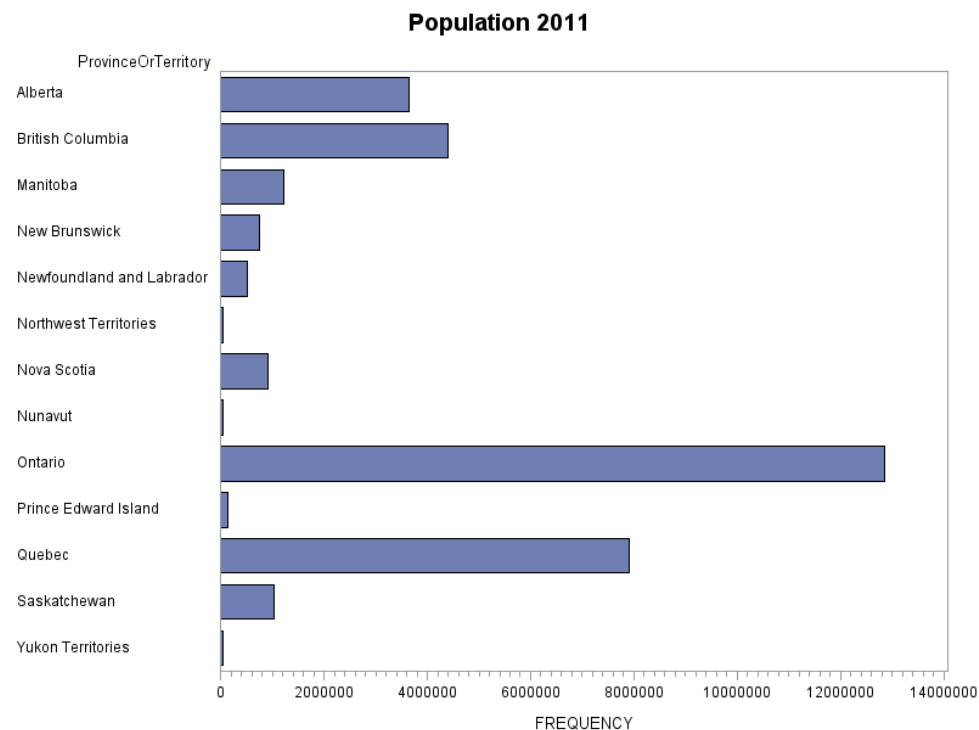
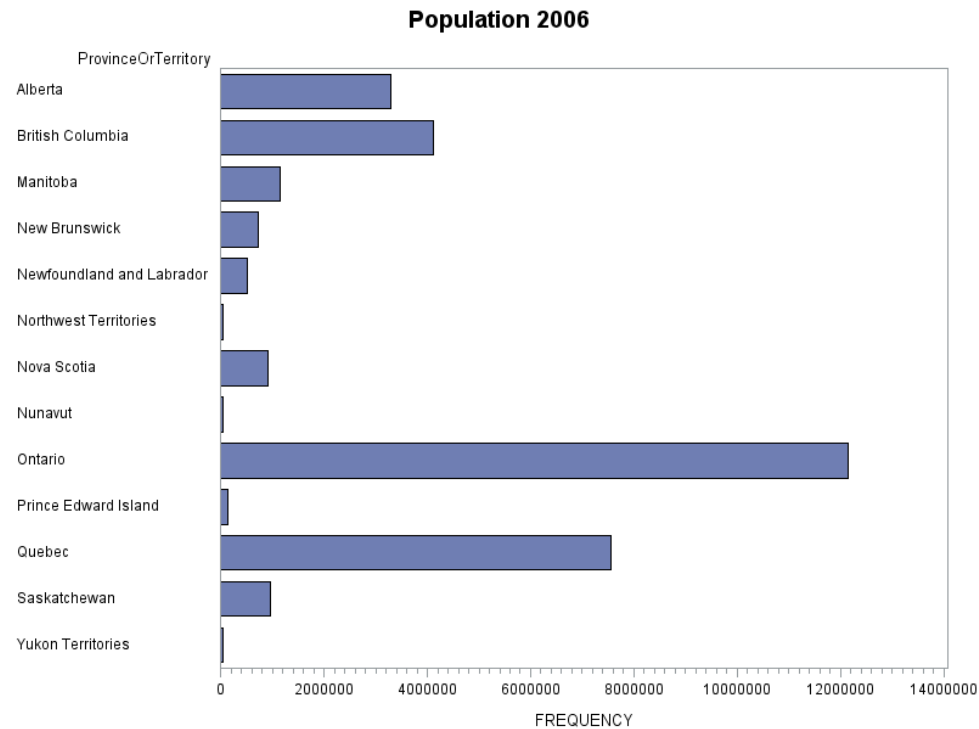
1.109 Answers will vary.

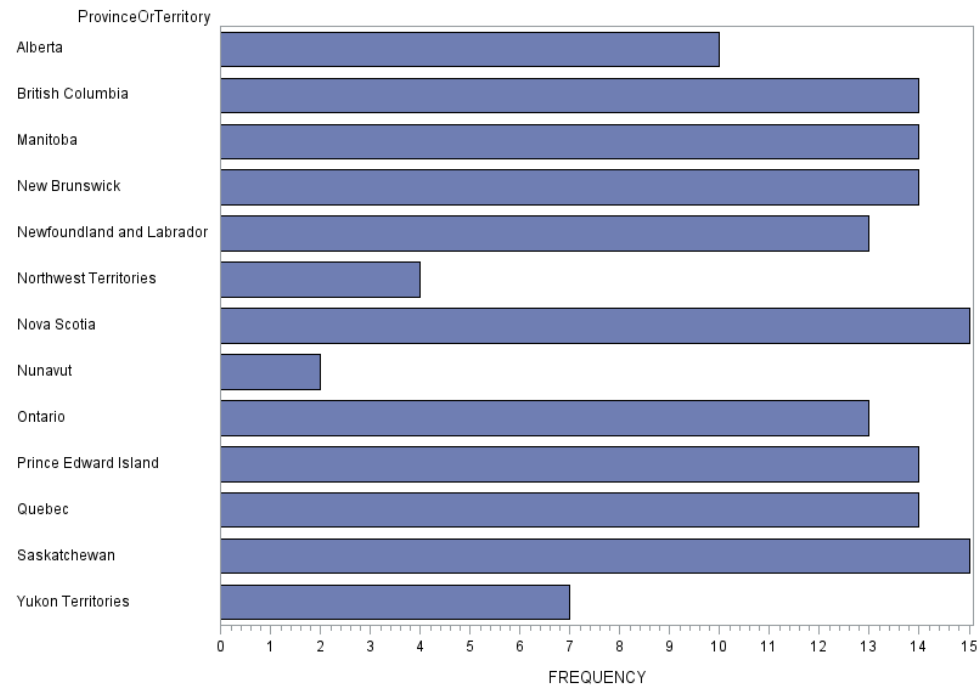
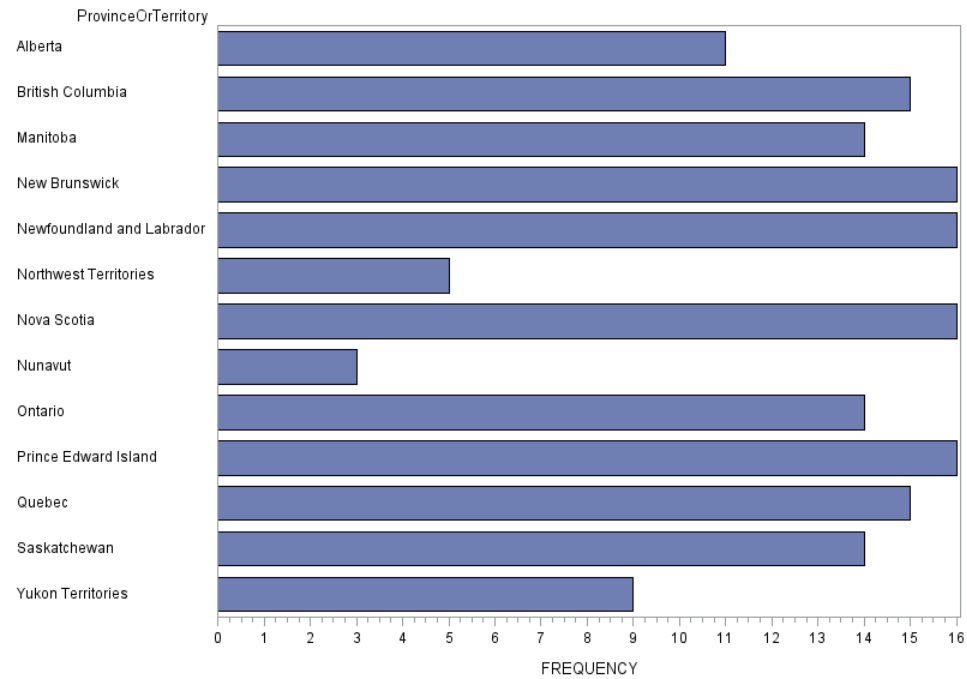
1.110 (a) In the histogram, these rates appear in the bar at the left end. In the time plot, most of the very small rates happened after 2010. In the Normal quantile plot, these rates appear near the bottom, forming a very straight line along the bottom of the graph. (b) Answers will vary; one solution is to make the classes smaller for the histogram as shown.

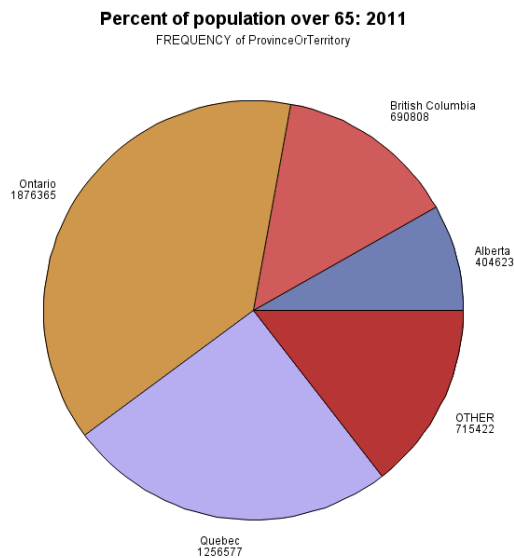
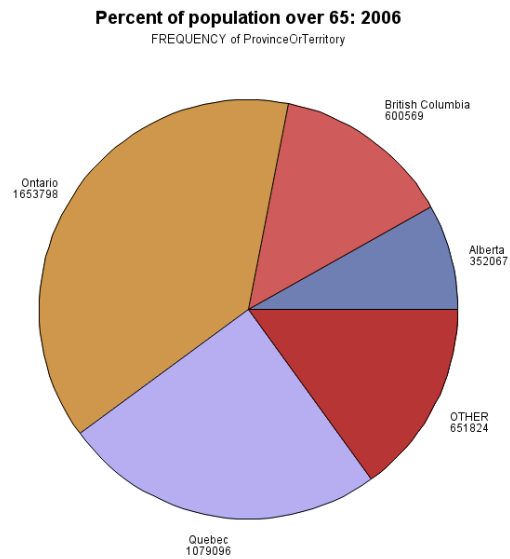


1.111 (a) Population values have not changed that much from 2006 to 2011. Ontario and Quebec have the largest populations, followed by Alberta and British Columbia. For population over 65, most regions have similar percentages of over 65 except the three territories that are most northern; similar to overall population, these numbers haven't changed much between 2006 and 2011. For percent of population over

65, four areas dominate, namely, Ontario, Quebec, British Columbia, and Alberta. Again, the numbers have not changed drastically between 2006 and 2011.



**Pecent over 65: 2006****Pecent over 65: 2011**

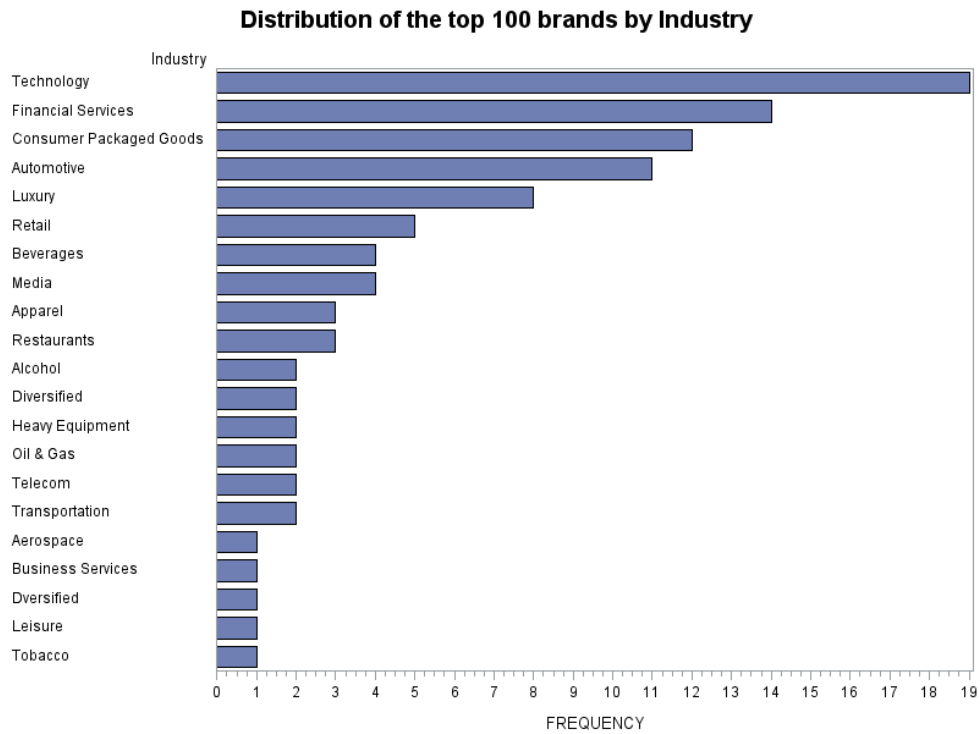


(b) Answers will vary. Most marketing techniques for targeting seniors should mention the four areas with the most population over 65: Ontario, Quebec, British Columbia, and Alberta.

1.112 (a) rank—quantitative, company name—categorical, value—quantitative, change—quantitative, revenue—quantitative, company advertising—quantitative, industry—categorical. (b) Rank is the label. (c) A case is a brand, a symbol, or images that are associated with a company.

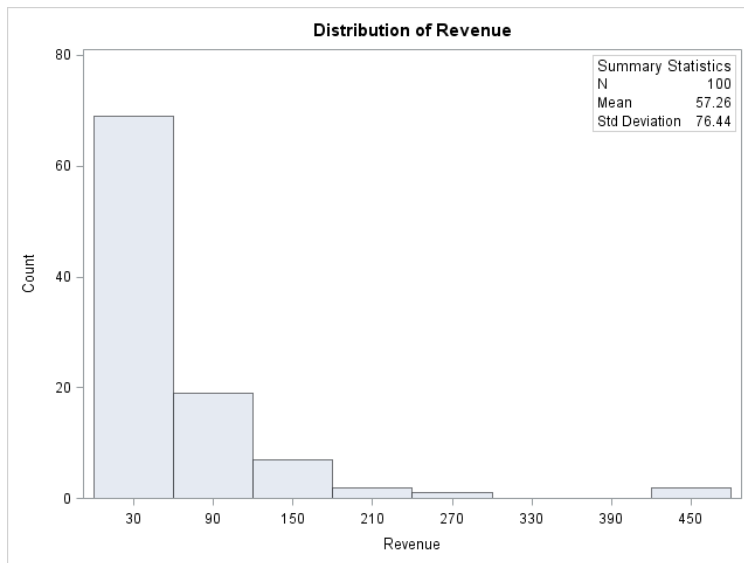
1.113 Answers will vary. The most popular industry among the top 100 brands is Technology, followed by Financial Services, Consumer Packaged Goods, Automotive, and Luxury.

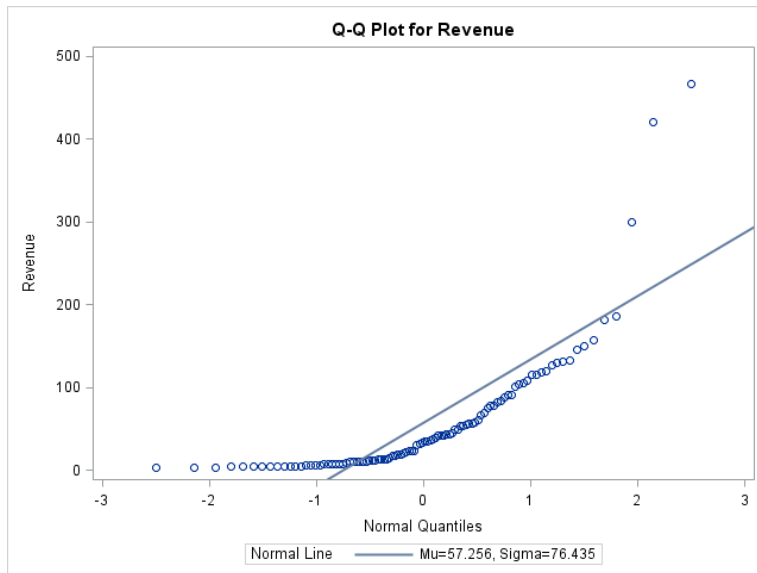




1.114 The distribution of revenue is strongly right-skewed as shown in the histogram and Normal quantile plot below. Also, as shown in both, we see that there are several large outliers.

$\bar{X} = 57.26$ ,  $s = 76.44$ ,  $M = 34.2$ ,  $Q1 = 10.15$ ,  $Q3 = 79.75$ .

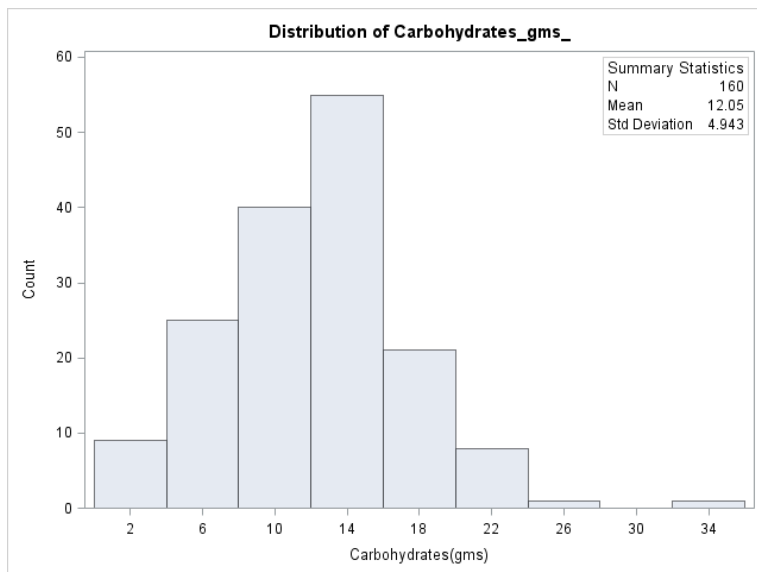


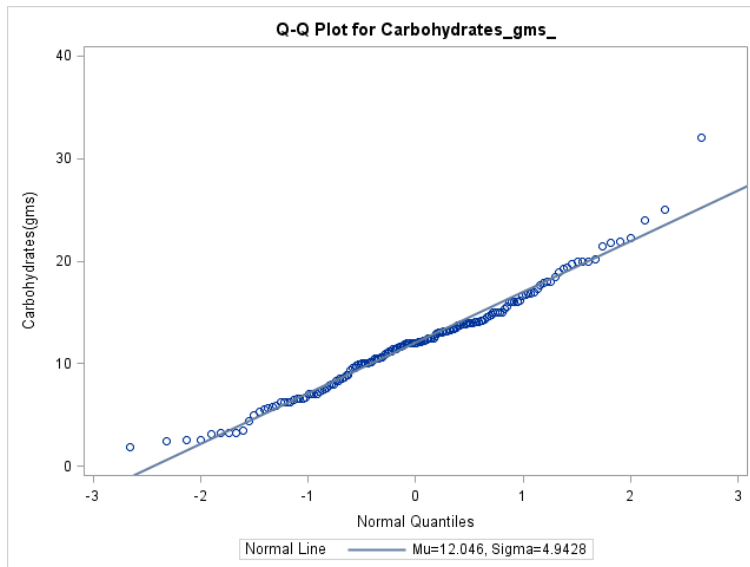


1.115 (a) brand—categorical, brewery—categorical, percent alcohol—quantitative, calories per 12 ounces—quantitative, carbohydrates in grams—quantitative. (b) Brand is the label. (c) A case is a domestic brand of beer; there are 175 cases.

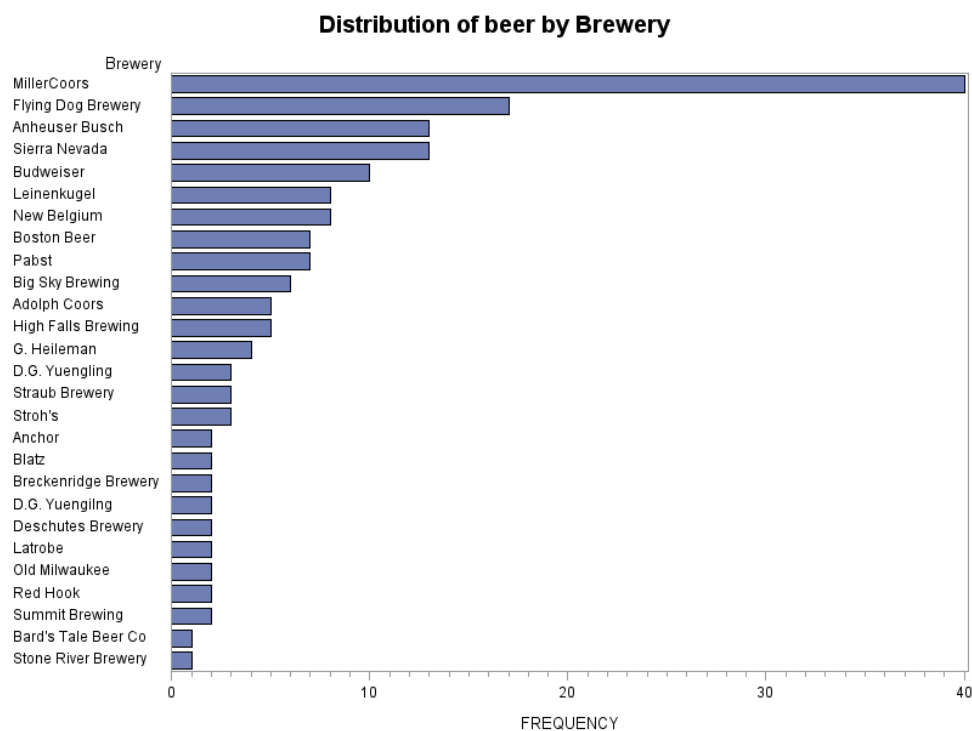
1.116 The distribution of carbohydrates is roughly Normally distributed with one possible large outlier.

$\bar{X} = 12.05$ ,  $s = 4.94$ ,  $M = 12.005$ ,  $Q1 = 8.65$ ,  $Q3 = 14.55$ .



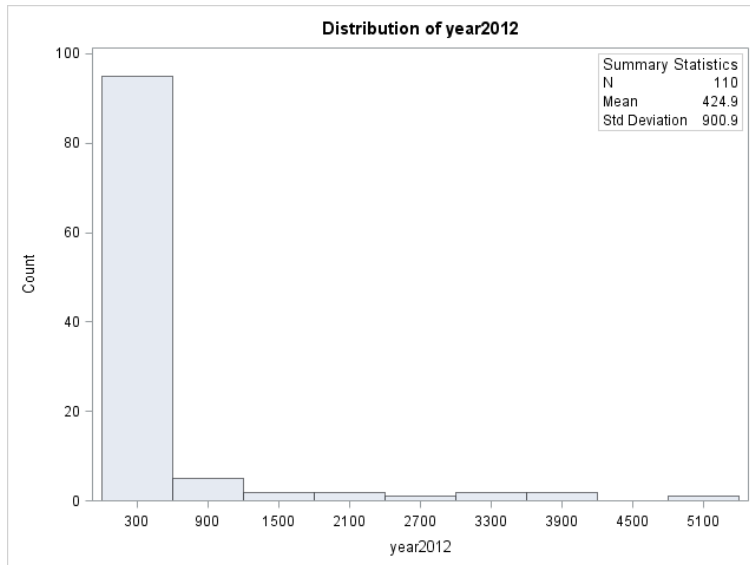


1.117 Many of the brands of beer come from the MillerCoors brewery, followed by Flying Dog Brewery, Anheuser Busch, Sierra Nevada, and Budweiser.



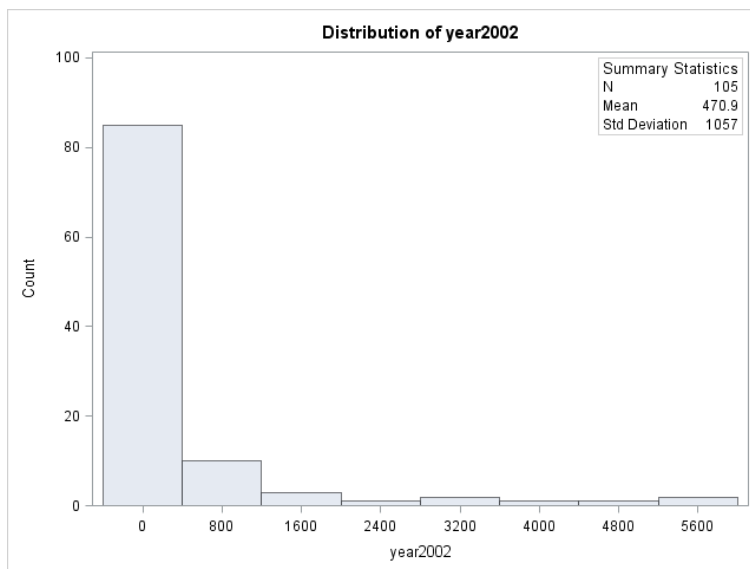
1.118  $\bar{X} = 424.9$ ,  $s = 900.9$ ,  $M = 103.5$ ,  $Q1 = 34$ ,  $Q3 = 287$ .

For 2012, the distribution is strongly right-skewed. Some countries have a huge number of incorporated companies, namely, India, United States, Canada, Japan, and Spain. 75% of countries have 287 or less incorporated companies, with the median amount at 103.5.

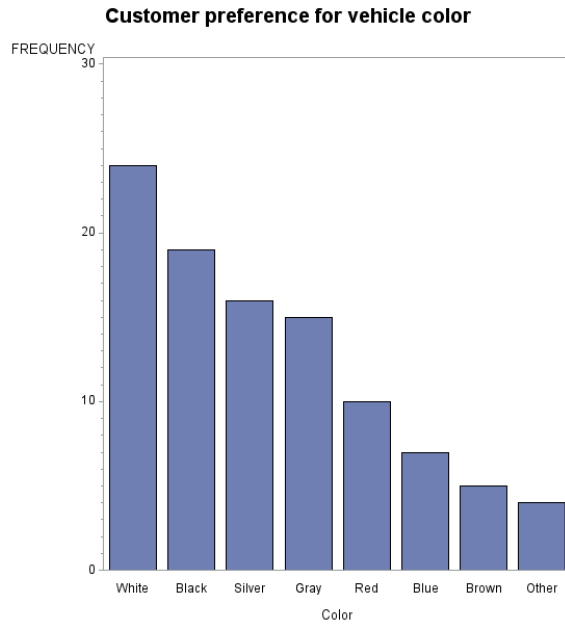


1.119  $\bar{X} = 470.9$ ,  $s = 1057$ ,  $M = 99$ ,  $Q1 = 40$ ,  $Q3 = 327$ .

For 2002, the distribution is also strongly right-skewed and has similar numerical summaries. The leading countries in 2002 are United States, India, Romania, Canada, Japan, and Spain. Romania did not appear in the top countries in incorporated companies for 2012. There are also 8 countries that have missing values. These missing values could change the summary somewhat if they have large or small amounts of incorporated companies.



1.120 White is the most popular color in 2012 for North America, followed by Black, Silver, and Gray. For marketing techniques, answers will vary.



1.121 1-c—currently there are more females in college than males. 2-b—there should be more right handed students than left handed. 3-d—height should be Normally distributed. 4-a—should be right-skewed, because some students will study much more than others.

1.122 475 is the 85th percentile; using table A gives  $Z = 1.04$ . 25 is the 15th percentile,  $Z = -1.04$ . Because it is symmetric, the mean must be half way between 25 and 475, so  $\mu = 250$ . To find the standard deviation, we solve  $1.04 = (475 - 250)/\sigma$ ,  $\sigma = 216.346$ .

1.123 Gender and automobile preference are categorical. Age and household income are quantitative.

1.124 Answers will vary.