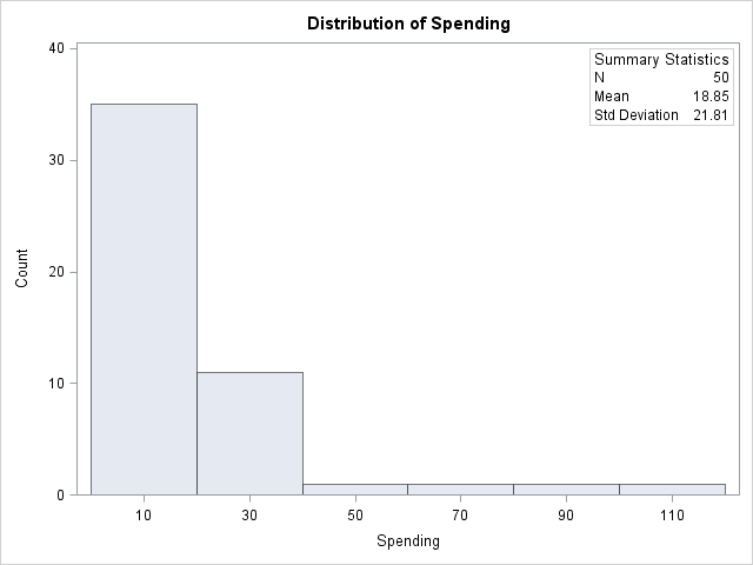
Chapter 2: Examining Relationships

2.1 (a) The cases are employees. (b) The label could be the employee's name or ID. (c) Answers will vary. (d) The explanatory variable is how much sleep they get; the response is how effectively they work.

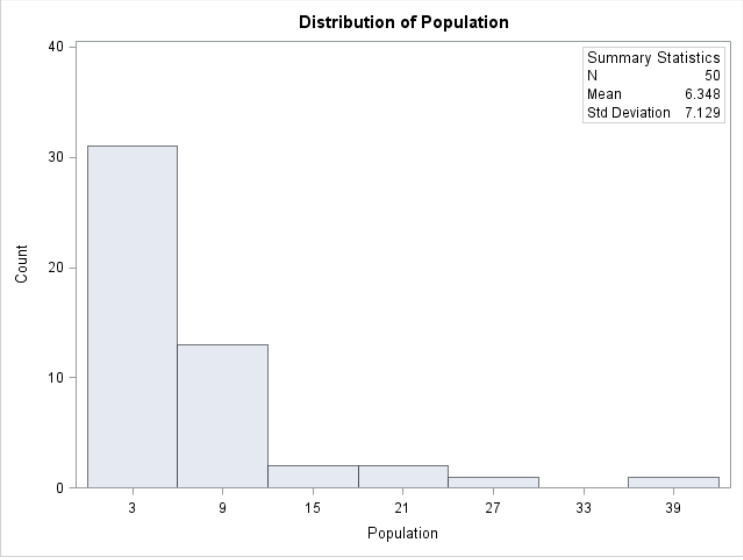
2.2 (a) The cases are the individual orders; the variables are the size and the price. (b) The size is the explanatory variable; it explains the price. Price is the response. (c) The cases are the individual orders; the variables are the ounces and the price. Ounces is the explanatory variable and price is the response.

2.3 State is the label; all other variables are quantitative.

2.4 (a)

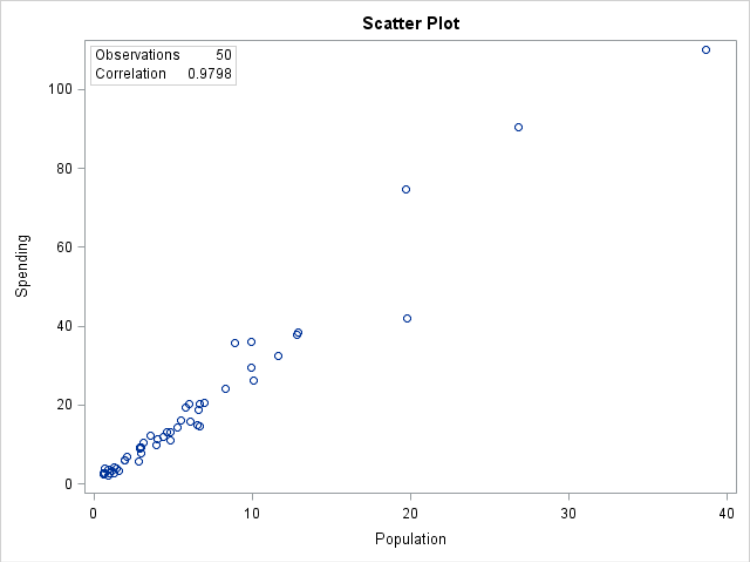


(b)

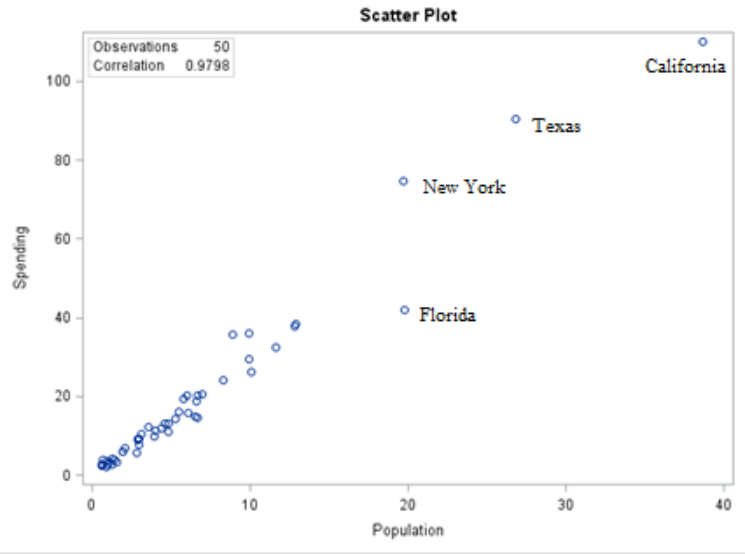


(c) Both spending and population are strongly right-skewed. The mean for spending is 18.85 with a 21.81 standard deviation. The mean for population is 6.348 with a standard deviation of 7.129.

2.5 (a)



(b)

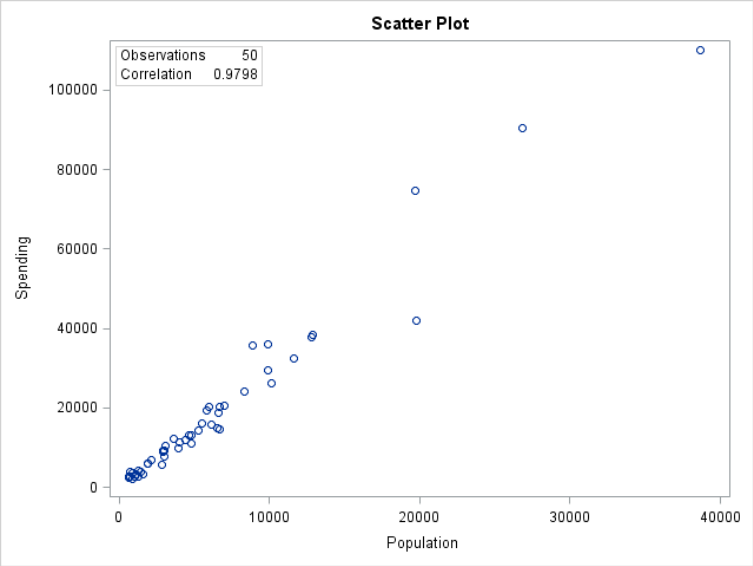


2.6 (a)

State	Spending×1000	Population×1000
Alabama	13200	4800
Alaska	3800	700
Arizona	14700	6700
Arkansas	9300	3000
California	110100	38700
Colorado	14400	5300
Connecticut	12100	3600
Delaware	3500	900

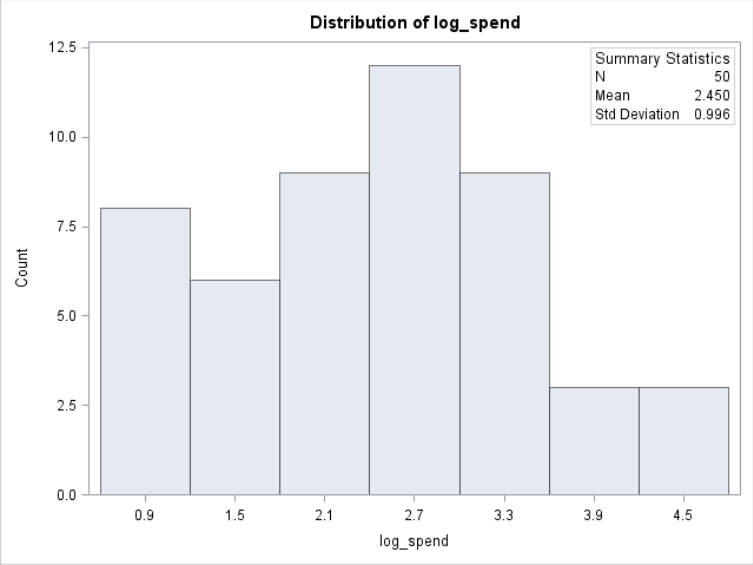
Florida	42000	19800
Georgia	26200	10100
Hawaii	3900	1400
Idaho	3300	1600
Illinois	38300	12900
Indiana	18700	6600
Iowa	10300	3100
Kansas	9400	2900
Kentucky	12000	4400
Louisiana	13000	4600
Maine	2800	1300
Maryland	20300	6000
Massachusetts	20200	6700
Michigan	36100	9900
Minnesota	16200	5500
Mississippi	7800	3000
Missouri	15800	6100
Montana	2700	1000
Nebraska	6100	1900
Nevada	5800	2800
New Hampshire	4100	1300
New Jersey	35700	8900
New Mexico	7000	2100
New York	74600	19700
North Carolina	29500	9900
North Dakota	2700	700
Ohio	32400	11600
Oklahoma	10000	3900
Oregon	11300	4000
Pennsylvania	37700	12800
Rhode Island	3300	1100
South Carolina	11000	4800
South Dakota	2100	900
Tennessee	14900	6500
Texas	90500	26800
Utah	9000	2900
Vermont	2500	600
Virginia	24000	8300
Washington	20500	7000
West Virginia	6000	1900
Wisconsin	19300	5800
Wyoming	2600	600

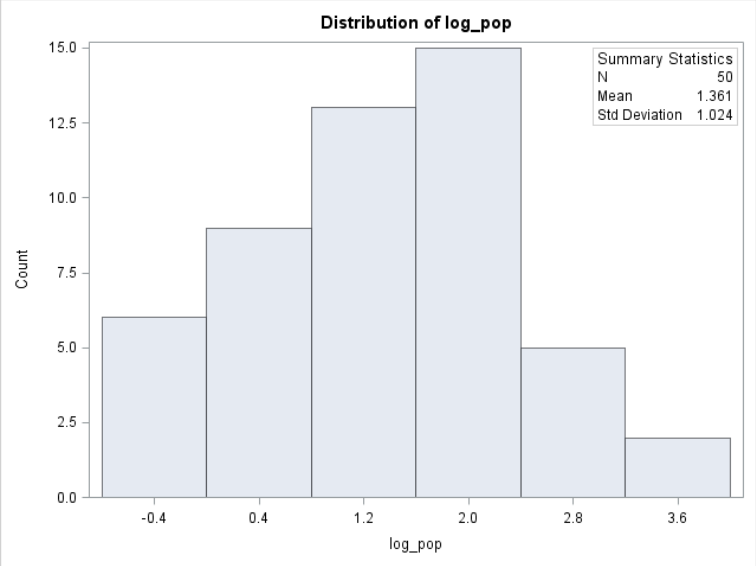
(b)



(c) The scatterplots are identical, just with the units changed.

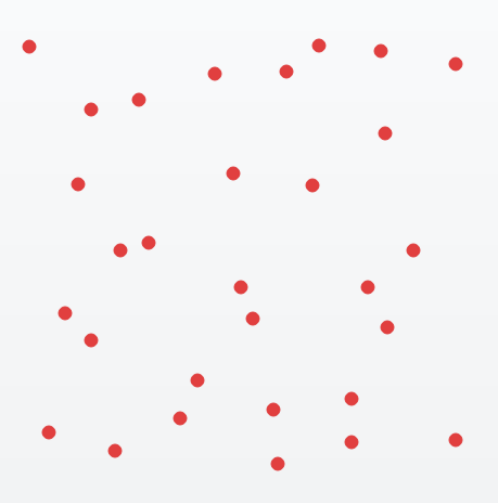
2.7 The skew is gone from both distributions. Both are close to symmetrical, with a single peak in the middle and roughly bell shaped.



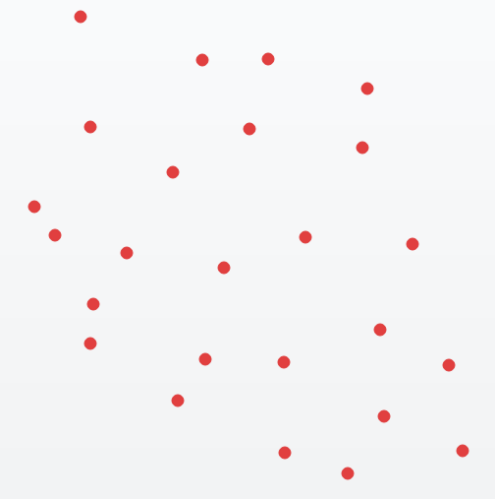


2.8 (a) If two variables are negatively associated, then low values of one variable are associated with **high** values of the other variable. (b) A **scatterplot** can used to examine the relationship between two variables. (c) The response goes on the ***y*** axis, the explanatory goes on the ***x*** axis.

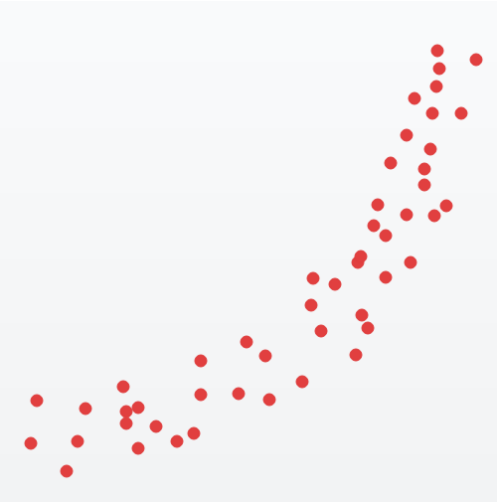
2.9 (a) No apparent relationship.



(b) A weak negative linear relationship.

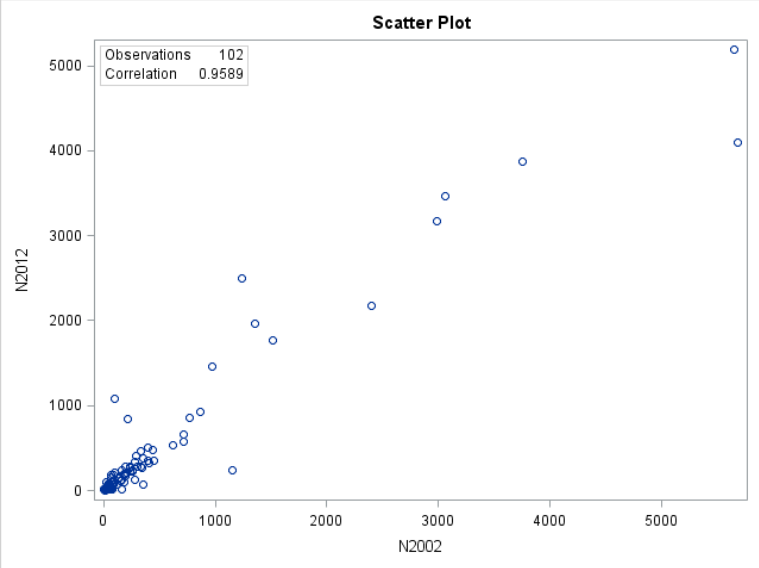


(c) A strong positive relationship that is not linear.



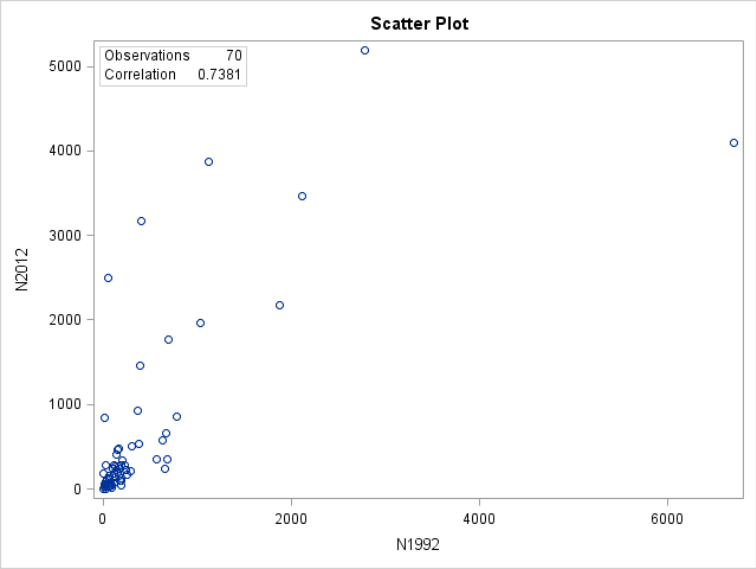
(d) A more complicated relationship. Answers will vary, below is an example of two distinct populations with separate relationships plotted together.

2.10 (a) The data for year 2002 would be the explanatory variable, the data for year 2012 would be the response. We would expect the 2002 data to explain, and possibly cause, changes in the 2012 data. (b)



(c) The form is linear; the direction is positive; the strength is very strong. (d) India and the United States appear to be outliers and have much larger values for both years than other countries.

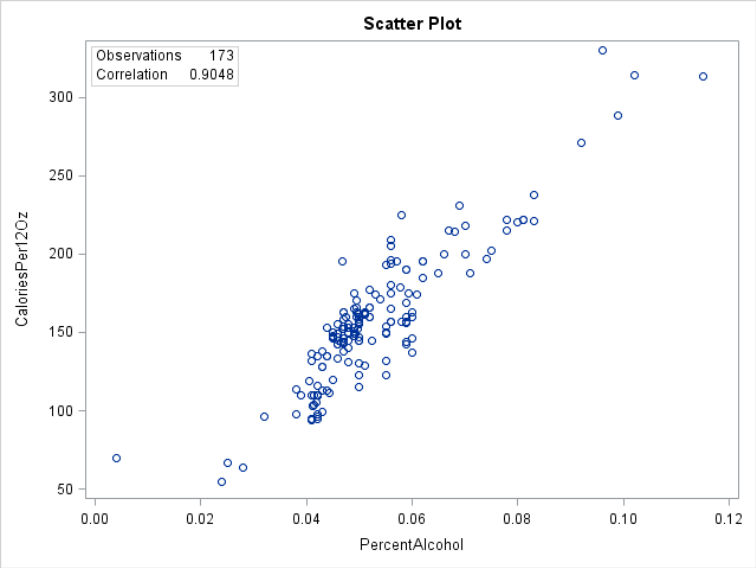
2.11 We expect the relationship between 2012 and 1992 to be weaker because the time difference is larger. (a) The data for year 1992 would be the explanatory variable; the data for year 2012 would be the response. We would expect the 1992 data to explain, and possibly cause, changes in the 2012 data. (b)



(c) The form is roughly linear; the direction is positive; the strength is moderate. (d) United States is the only outlier with a much larger value for the year 1992 than most other countries.

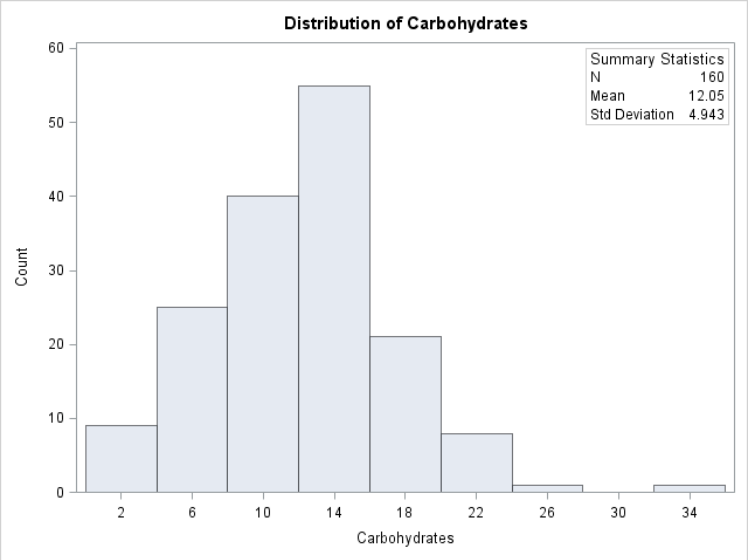
2.12 (a) The two variables calories and percent alcohol have fairly symmetric distributions with one potential outlier, O'Doul's.

(b)

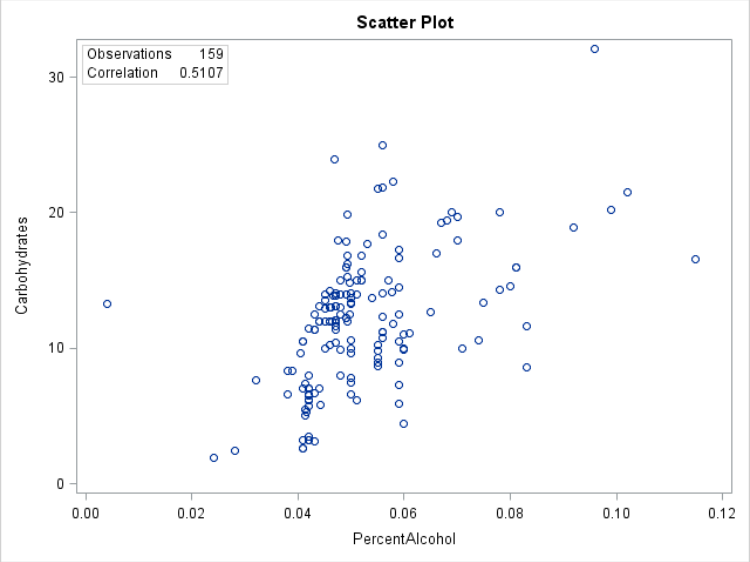


(c) The form is linear; the direction is positive; the strength is very strong. (d) O'Doul's could be a potential outlier; it has a very small percent alcohol value.

2.13 (a) From 1.156, percent alcohol is somewhat right-skewed. Carbohydrates, shown below, is fairly symmetric.

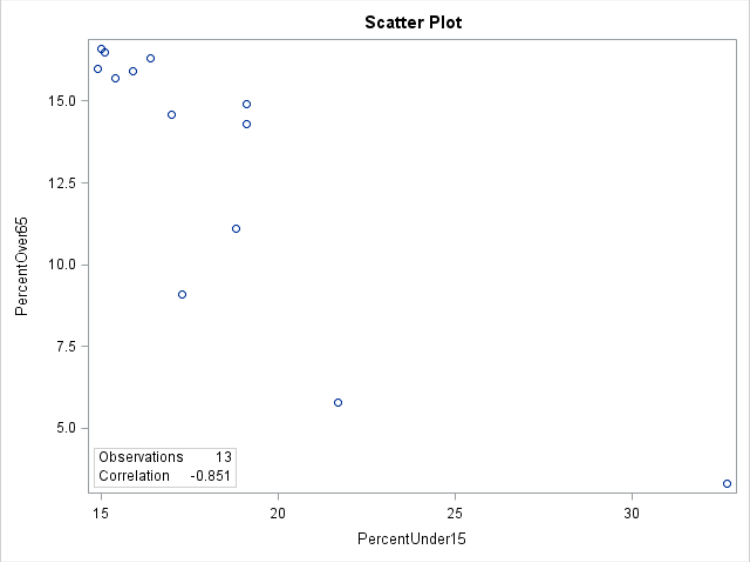


(b)



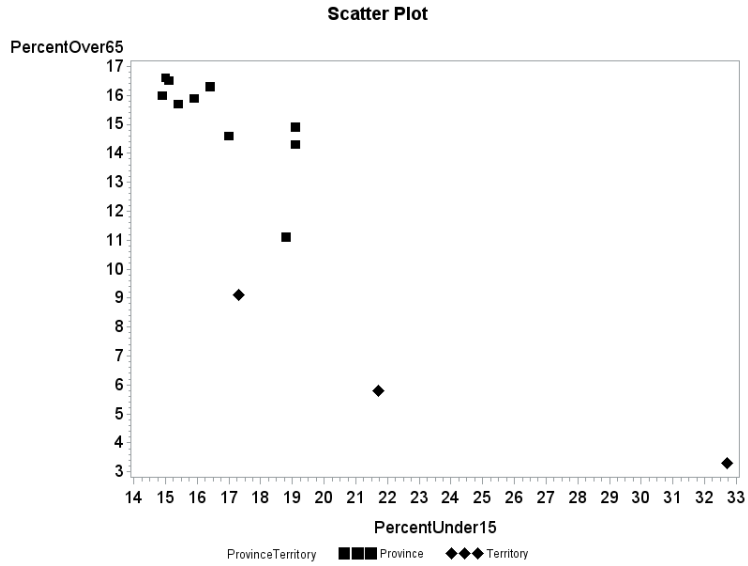
(c) The form is somewhat linear; the direction is positive; the strength is weak. (d) O'Doul's could be a potential outlier; it has a very small percent alcohol value. Sierra Nevada Bigfoot could also be a potential outlier; it has a very high amount of carbohydrates.

2.14 (a)



(b) The form is fairly linear, the direction is negative, the strength is strong.

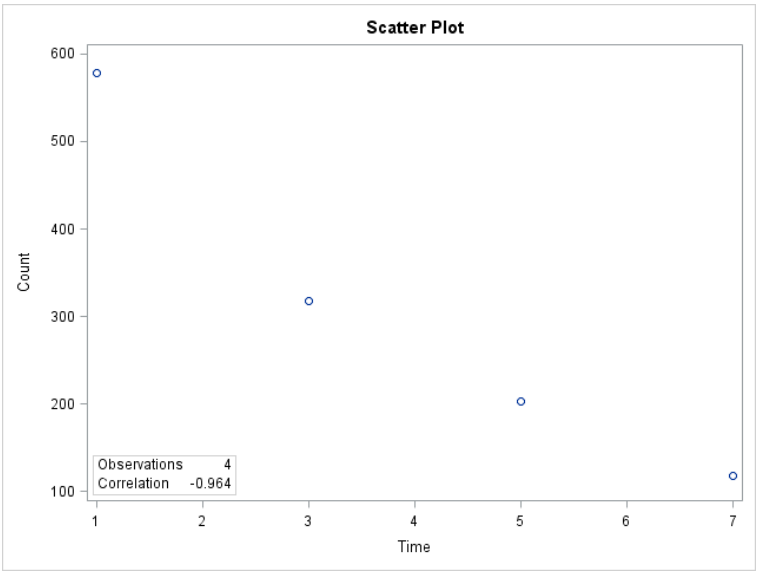
2.15 (a)



(b) The three territories have smaller percentages of the population over 65 than any of the provinces. Additionally two of the three territories have larger percentages of the population under 15 than any of the provinces.

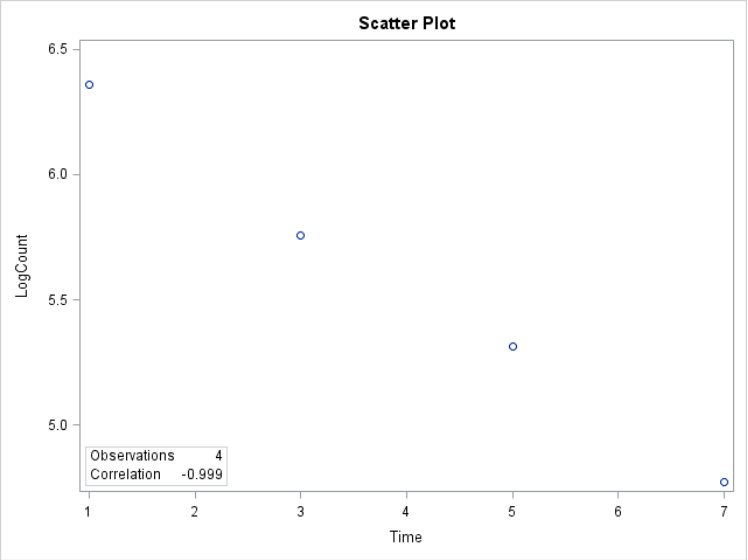
2.16 (a) The explanatory variable is the time spent on your pages. The response variable is the amount of their purchases. We would expect the time spent to explain the amount spent. (b) Both variables are quantitative. (c) Answers will vary. It is likely the association is positive because the more time they spend on your pages indicates they are likely successful during their shopping and should spend more. (d) Answers will vary.

2.17 (a) We would expect time to explain the count, so time should be on the *x* axis.



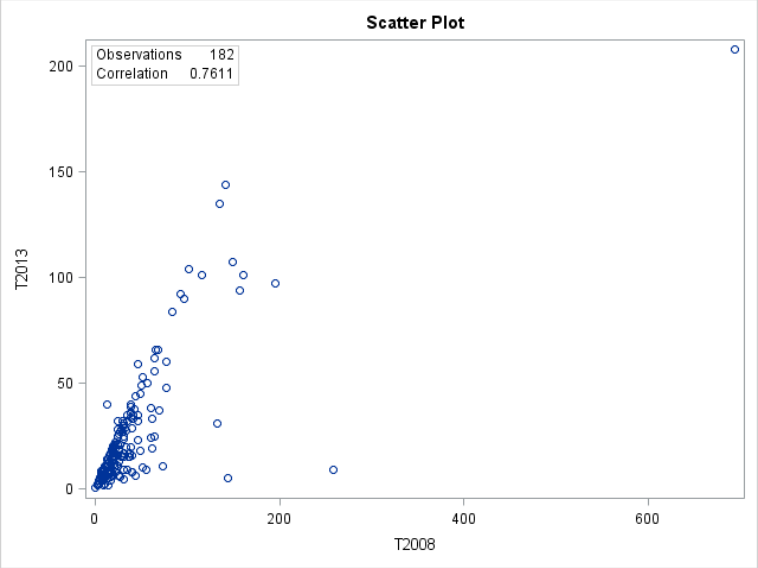
(b) As time increases, the count goes down. (c) The form is curved; the direction is negative; the strength is very strong. (d) The first data point at time 1 is somewhat of an outlier because it doesn't line up as well as the other times do. (e) A curve might fit the data better than a simple linear trend.

2.18 (a)



(b) As time increases, logcount goes down. (c) The form is linear; the direction is negative; the strength is extremely strong. (d) There are no outliers. (e) The relationship is very linear, almost a perfect line.

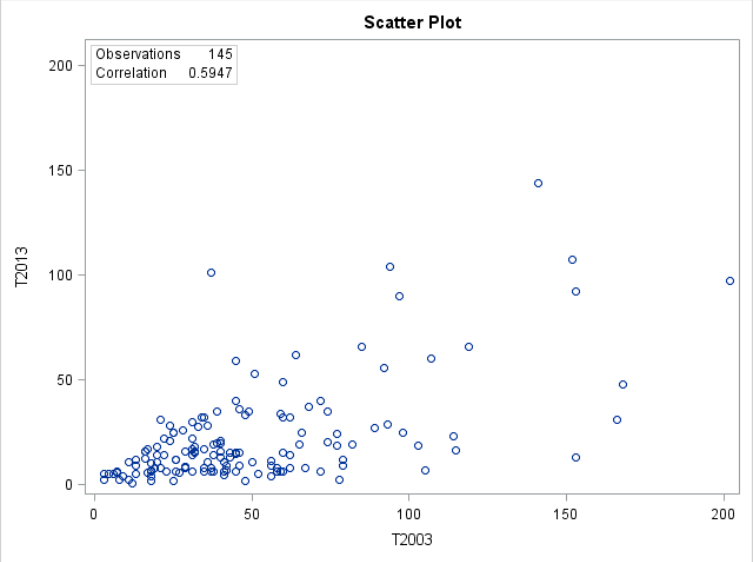
2.19 (a) 2008 data should explain the 2013 data. (b)



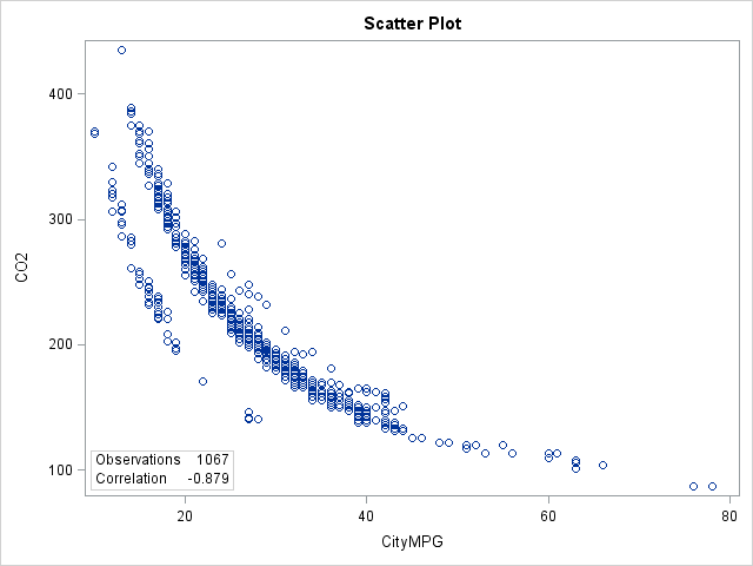
(c) There are 182 points; some of the data for 2008 are missing. (d) The form is somewhat linear; the direction is positive; the strength is moderate. (e) Suriname is an outlier for both 2008 and 2013. (f) The relationship is somewhat linear, though there are observations that don't follow the linear trend well.

2.20 (a) The 2003 data should explain the 2013 data. Here there are only 145 data points because some of the data for 2003 are missing. The form is somewhat linear; the direction is positive; the strength is weak to moderate. There are a few semi-outlying observations but nothing that seems drastic. The relationship is not extremely linear as there is quite a bit of scatter throughout the plot. (b) The relationship between the 2008 and 2013 times is stronger than the relationship between the 2003 and 2013 times. This is likely

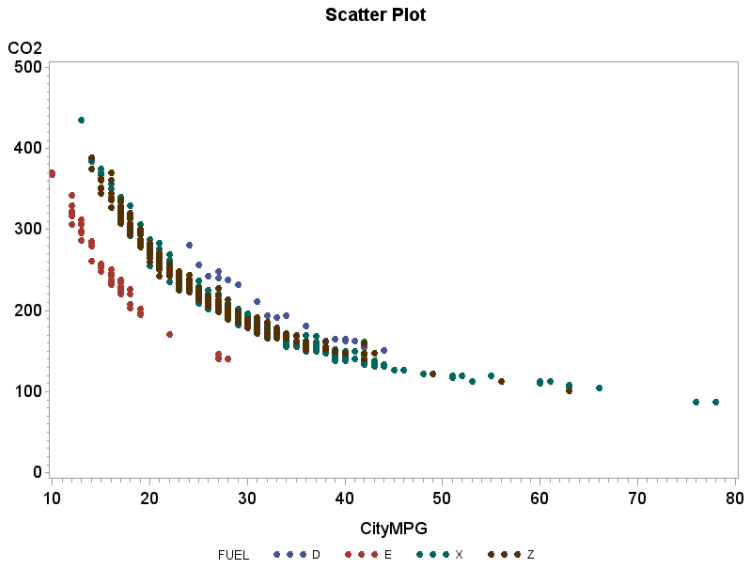
because the start times slowly change over time, so we would expect bigger differences between the 2003 and 2013 times and hence not as strong of a relationship as we saw in the 2008 and 2013 times.



2.21 There is a negative relationship between City MPG and CO_2 emissions; better City MPG is associated with lower CO_2 emissions. The relationship, however, is not linear but curved. There also seems to be two distinct lines or groups. This relationship is very similar to what we found in Example 2.7 when using highway MPG, with the patterns seen in the plot nearly identical to the form we saw in Example 2.7.



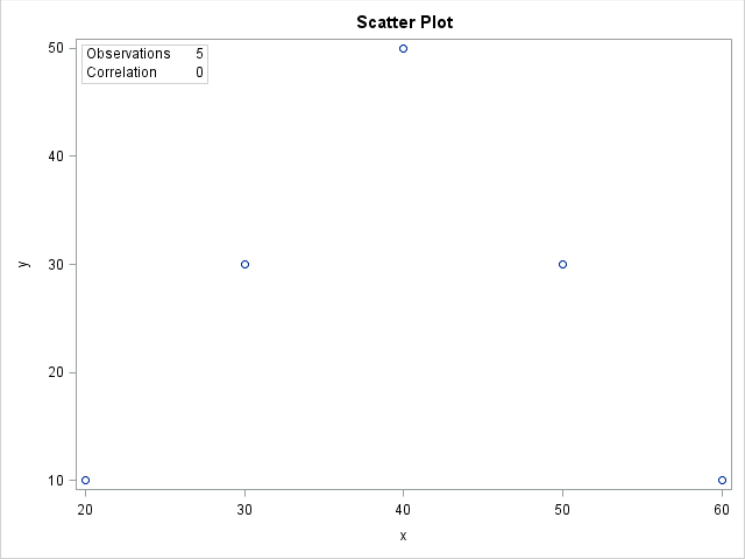
2.22 Each fuel type has a curved relationship between City MPG and CO_2 emissions. Furthermore, the curves for the 4 fuel types are very similar, although some curves for some fuel types are shifted because they provide either better or worse City MPG than other types. However, the emissions for the 4 fuel types are all very similar and all within the same range.



2.23 $r = 0.9798$.

2.24 (a) $r = 0.9798$. (b) It did not change. (c) Changing the units has no effect on the correlation.

2.25 (a)



(b) The relationship between x and y is very strong but it is not linear; it has a curved relationship or parabola. (c) $r = 0$. (d) The correlation is only good for measuring the strength of a linear relationship.

2.26 (a) $r = 1$. (b) $r = 1$.

2.27 (a) $r = 0.9589$. (b) Yes, there is a very strong linear relationship between the 2002 and 2012 data.

2.28 $r = 0.7381$. Overall, yes, because there is a moderate linear relationship between the 1992 and 2012 data. However, there is one outlier that doesn't fit the pattern. The correlation for the 1992 and 2012 data is not as strong because the data are not as linear as they were for the 2002 and 2012 data.

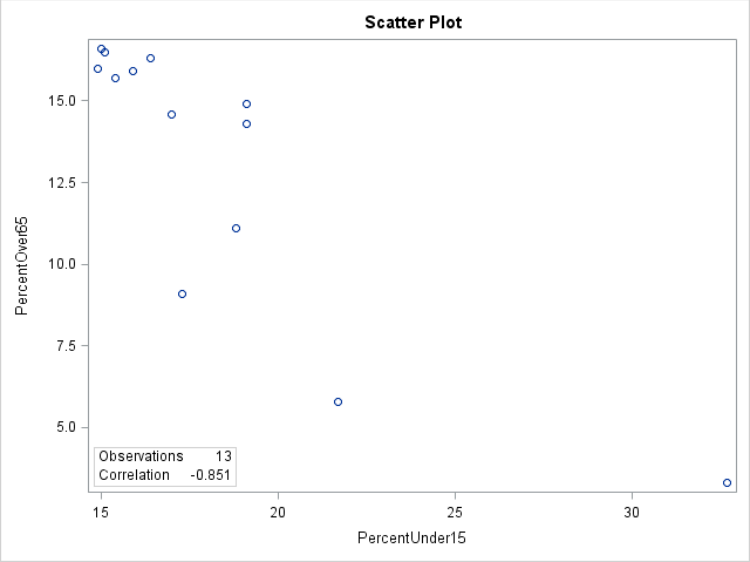
2.29 (a) Yes, the relationship between time and count is quite strong because the data points form a nice curve. (b) $r = -0.964$. (c) The correlation is not a good numerical summary because the data form a curve, not a line, a transformation is needed to get a better relationship.

2.30 (a) The relationship between time and log of the counts is very strong; the data points follow nearly a perfect line. (b) $r = -0.999$. (c) The correlation gives a very good numerical summary of the relationship because the data show a very nice linear form. (d) The correlation wasn't bad before the transformation (–0.964), but it is much better after the transformation (–0.999). A high correlation doesn't automatically mean a linear fit, especially when we see a curve in the scatterplot. Here a transformation gave us a much better fit, straighter line, and an even higher correlation. (e) The correlation by itself isn't enough to explain a relationship. If we had just calculated the correlation without looking at the scatterplot before transforming, we would have thought we had a very nice linear relationship when, in fact, a curve via the transformation provided a much better description of the actual relationship, and yielded a much higher correlation, and thus a better fit.

2.31 (a) $r = 0.9048$. (b) Yes, the relationship between percent alcohol and calories is quite linear, so the correlation gives a good numerical summary of the relationship.

2.32 (a) $r = 0.9077$. (b) We might expect outliers to always drastically change the correlation but as this example illustrates, that is not always the case. Here removing the outlier O'Doul's didn't change the correlation much at all. It really depends on where the outlier falls in the linear relationship as to how much it may or may not affect the correlation. Additionally, the number of observations does play some role, there are so many observations in this particular dataset that removing one that is only somewhat outlying doesn't change much.

2.33 (a)



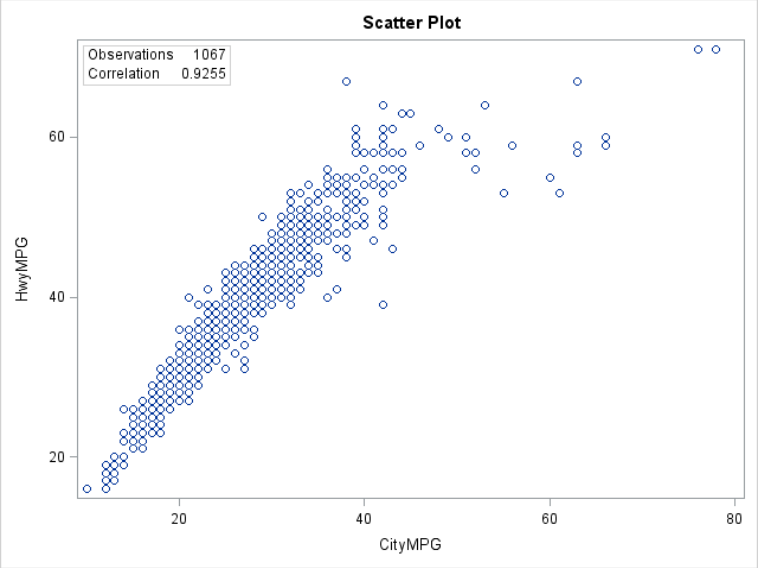
(b) $r = -0.851$. (c) No, although the relationship is mostly linear, there is an outlier, Nunavut, with a high percent of under 15 and a very low percent over 65.

2.34 (a) Yes, Nunavut is an outlier with a high percent of under 15 and a very low percent over 65. (b) $r = -0.780$. Nunavut was actually helping improve the linear relationship between percent under 15 and percent over 65. Without Nunavet, the correlation went down a little bit, from –0.851 to –0.780.

2.35 (a) $r = 0.9808$. (b) The correlation went up from 0.9798 before taking the logs to 0.9808 thereafter. Though the correlation went up a little bit, the log didn't help much with the explanation of the data.

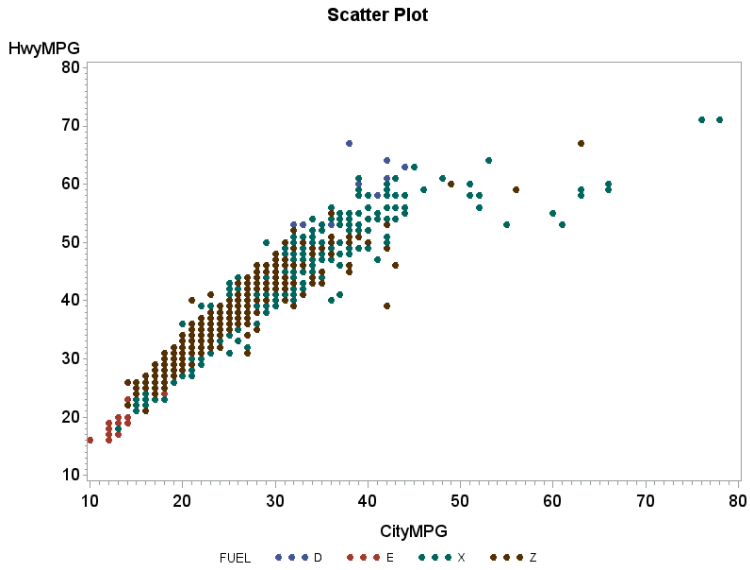
2.36 (a) $r = 0.9742$. (b) $r = 0.9753$. (c) There is very little difference between the correlations before and after removing the potential outliers. Similarly, using the log transformation had very little effect as well, both before and after removing the potential outliers. As long as the outliers follow the pattern of the linear trend, they have little effect on the correlation when removed.

2.37 (a)



(b) The relationship is somewhat linear but may also be slightly curved. Hwy MPG and City MPG increase together. (c) $r = 0.9255$. (d) The correlation is a decent numerical summary because the data are somewhat linear, but a curve may provide a better description of the relationship.

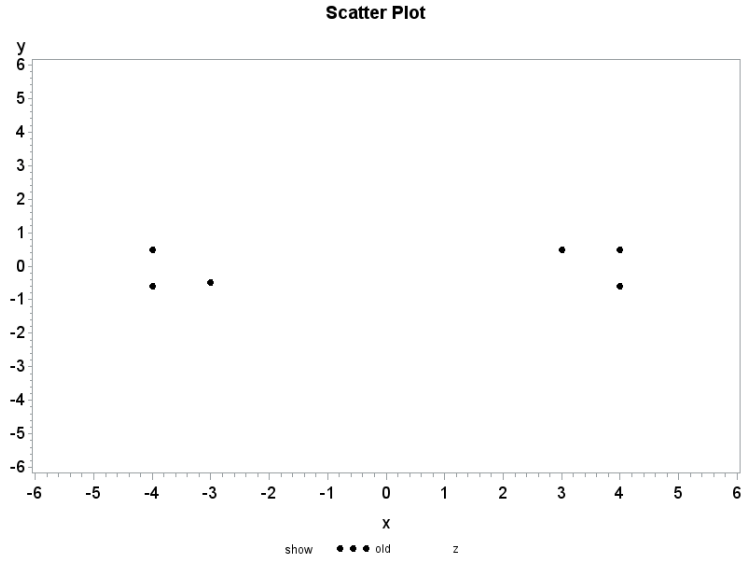
2.38 (a)



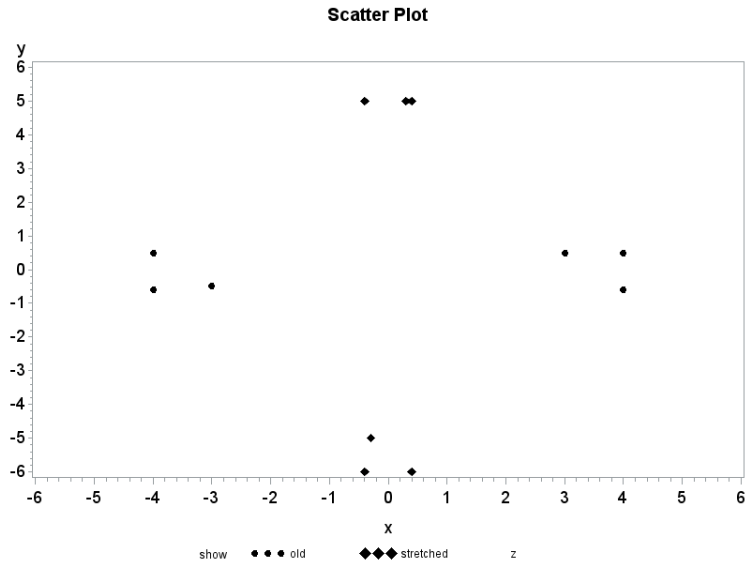
(b) The distinction between fuel types is difficult to see, especially because the relationship between City MPG and Hwy MPG is consistent regardless of fuel type. It is apparent that some fuel types get much better MPG than other fuel types, regardless of City vs. Hwy (type E has the smallest MPGs and types X and Z some of the largest), but overall the pattern is consistent across all fuel types. (c) For fuel type D: $r = 0.9382$. For fuel type E: $r = 0.9560$. For fuel type X: $r = 0.8992$. For fuel type Z: $r = 0.9351$. Although the correlations vary somewhat, they are all very similar in their description of the relationship between City MPG and Hwy MPG.

2.39 Applet, answers will vary.

2.40 (a)



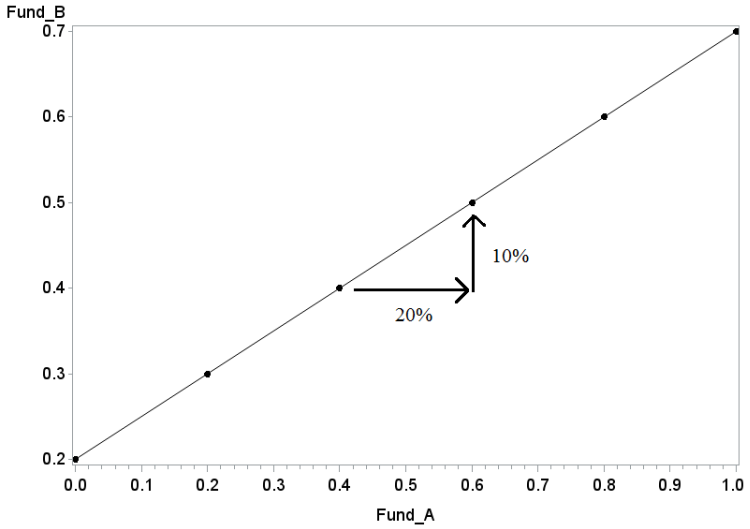
(b)



(c) The correlation between x and y is 0.2531. The correlation is still 0.2531 for x^* and y^*. By multiplying and dividing by 10 we are essentially just changing the units for x and y, which we know does not change the straight line relationship as measured by the correlation.

2.41 The magazine report is wrong because they are interpreting a correlation close to 0 as a negative association rather than no association. Answers will vary. "A new study shows no linear association (relationship) between how much a company pays their CEO and how well their stock performs."

2.42 A correlation measures the strength of a linear relationship; or, that is to say, the relationship between Fund A and Fund B is consistent along a line. It doesn't mean they have to change by the same amount. So as long as Fund A moves 20% and Fund B move 10% consistently, up or down, you will still remain on the same line.

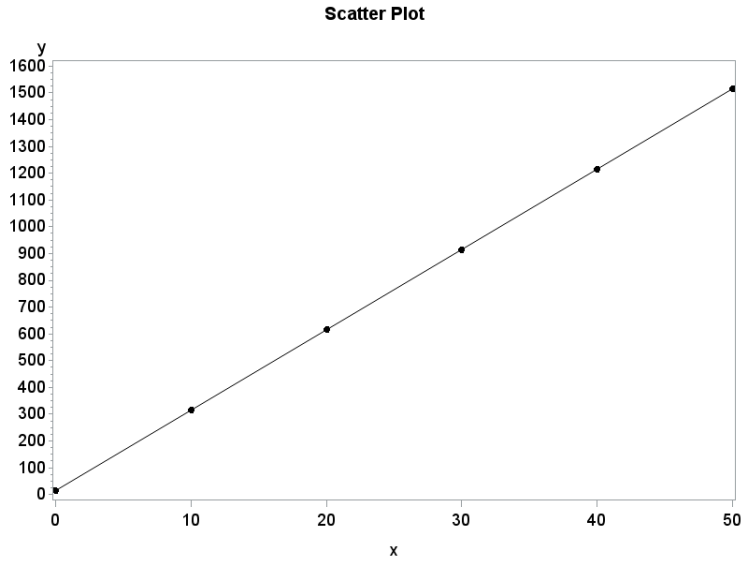


2.43 (a) The correlation is not dependent on order and remains the same between two variables regardless of order. (b) A correlation is reserved for quantitative data; because color is categorical, it cannot have any correlation. (c) A correlation can never exceed 1, which indicates a perfect linear relationship.

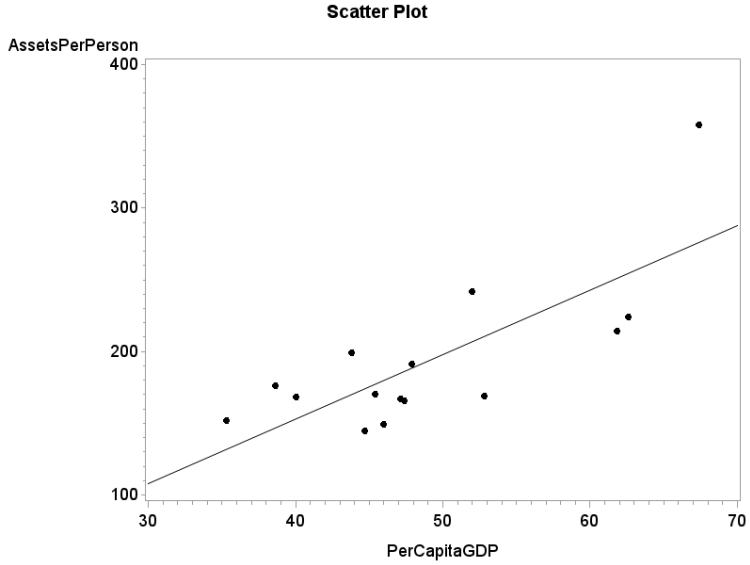
2.44 The estimated net assets per capita is about \$160,000. The prediction error = observed y – predicted y = 170 – 160 = \$10,000.

2.45 There are 7 (one is just barely above the line) positive prediction errors and 8 negative prediction errors.

2.46 (a) 30. (b) 15. (c) For $x = 10$, $y = 15 + 30(10) = 315$. For $x = 20$, $y = 15 + 30(20) = 615$. For $x = 30$, $y = 15 + 30(30) = 915$. (d)



2.47 (a) $b_1 = 4.4999$, $b_0 = -27.1682$. Locations will vary depending on software used. (b)



(c) and (d)

Country	Predicted	Prediction Error
United Kingdom	169.927	29.0728
Australia	186.127	−20.1268
United States	188.377	2.6232
Singapore	152.828	15.1724
Canada	177.127	−7.1270
Switzerland	276.125	81.8753
Netherlands	206.826	35.1737
Japan	146.528	29.4723
Denmark	254.525	−30.5252

France	179.827	−30.8270
Germany	173.977	−28.9771
Belgium	184.777	−17.7768
Sweden	210.426	−41.4263
Spain	131.678	20.3219
Ireland	250.925	−36.9253

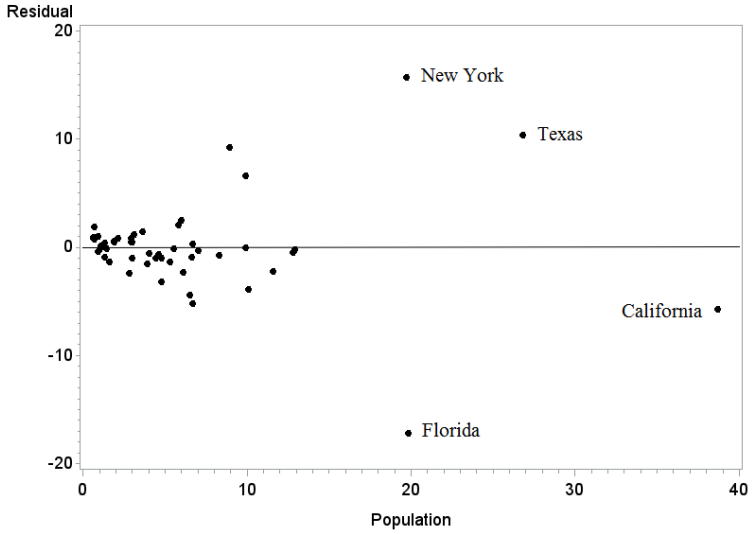
2.48 (a) $r^2 = 35.52\%$. (b) $\hat{y} = 6.083 + 1.707x$. (c) For $x = 1.75$, $\hat{y} = 6.083 + 1.707(1.75) = 9.07$. We could have given this value immediately because it is $\overline{y}$.

2.49 Applet, answers will vary.

2.50 (a) For $x = 26.8$, $\hat{y} = –0.16251 + 2.99713(26.8) = 80.16$. (b) Residual $= y – \hat{y} = 90.5 – 80.16 = 10.34$. (c) Texas is further from the regression line because it has a bigger residual in magnitude (absolute value).

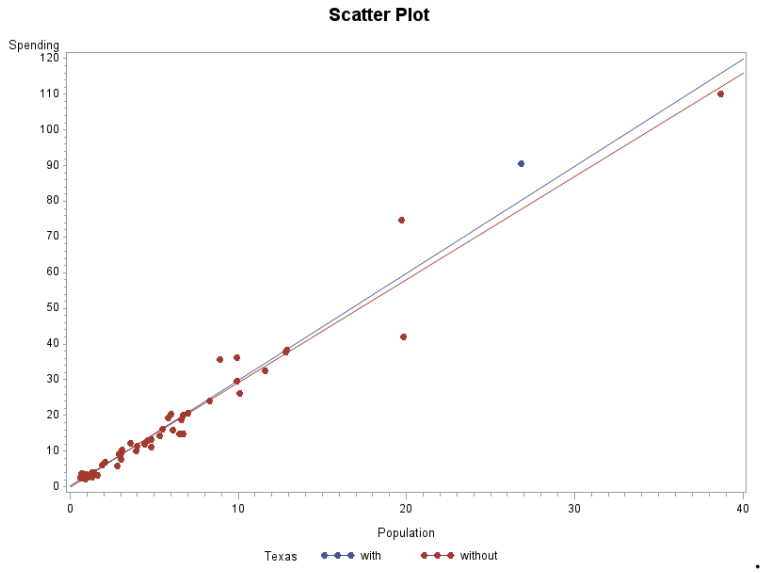
2.51 The residuals sum to –0.01. This is due to rounding error.

2.52 (a)



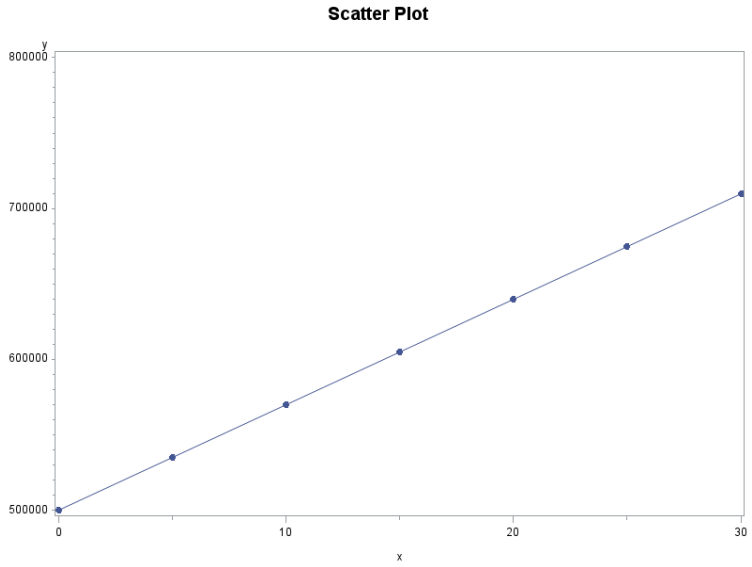
(b) It is easy to identify the points because the scatterplot and residual plot both have the *x* variable, population, on the *x* axis.

2.53 The lines are very similar, with and without Texas. Texas is not an influential observation.



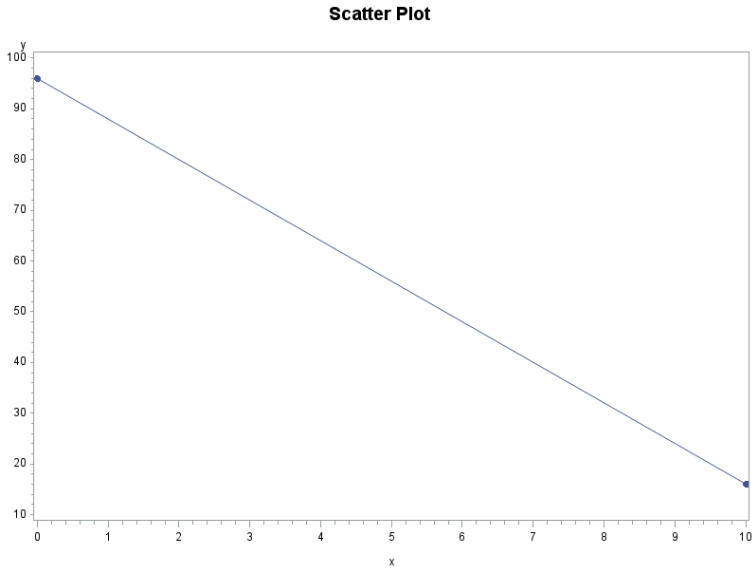
2.54 $y = 25 + 1.12x$.

2.55 (a)



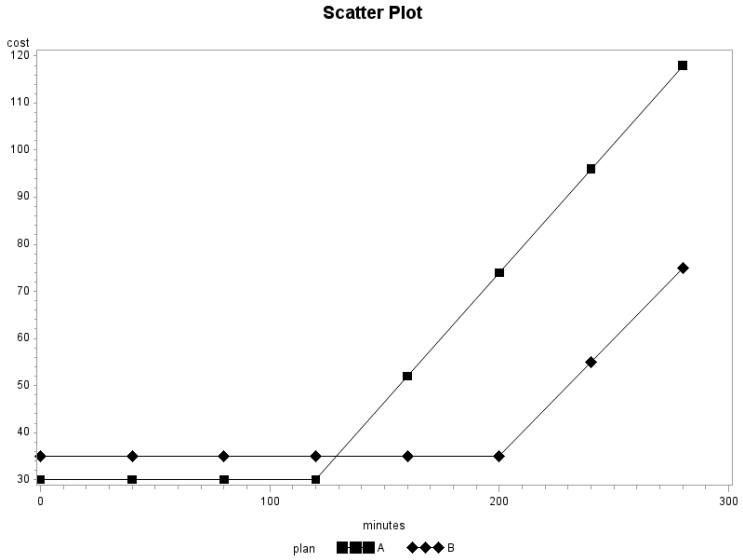
(b) For $x = 15$, $y = 500{,}000 + 7000(15) = \$605{,}000$. (c) $y = 500{,}000 + 10000x$.

2.56 (a) $y = 96 - 8x$. The slope is –8. (b)



(c) You should not predict for 25 weeks. For $x = 25$, $y = 96 – 8(25) = –104$. This is unreasonable because you can't have a negative amount of DVD players in your inventory (unless considering backordering), you will run out of DVD players after just 12 weeks.

2.57 (a)

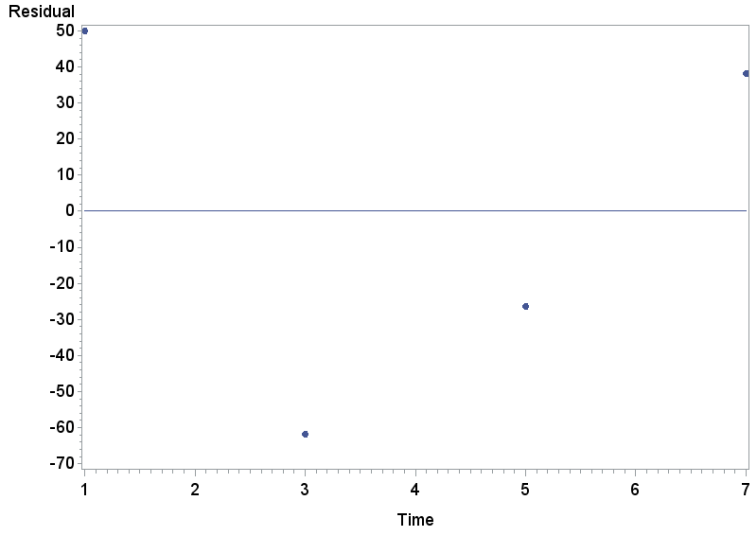


(b) Plan B gets cheaper after 130 minutes.

2.58 (a) $\hat{y} = 54.218 + 0.913x$. (b) For $x = 278$, $\hat{y} = 54.218 + 0.913x = 54.218 + 0.913(278) = 308.032$. (c) Residual $= y – \hat{y} = 332 – 308.032 = 23.968$.

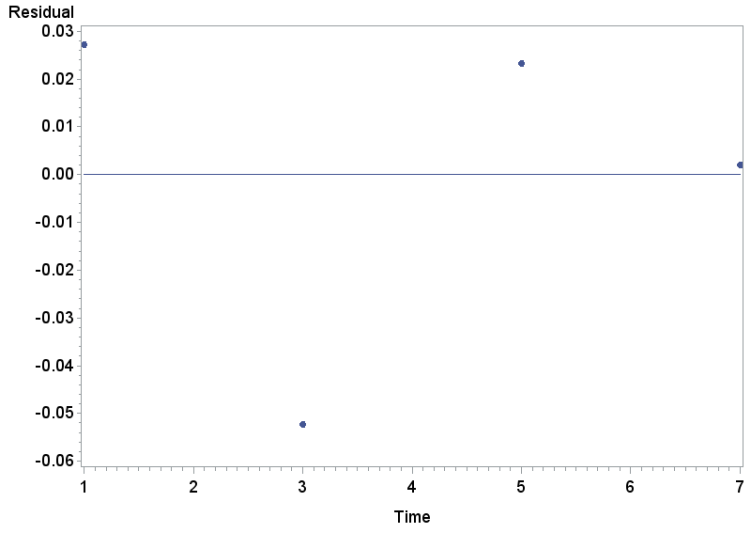
2.59 (a) $\hat{y} = 267.83 + 0.878x$. (b) For $x = 205$, $\hat{y} = 267.83 + 0.878x = 267.83 + 0.878(205) = 447.82$. (c) Residual $= y – \hat{y} = 332 – 447.82 = –115.82$. The prediction using the 2002 data is much better because the relationship was much stronger for the 2002 data than it is in the 1992 data, which produces a fairly large residual.

2.60 (a) $\hat{y} = 602.8 - 74.7x$. (b) For $x = 1$, $\hat{y} = 602.8 - 74.7(1) = 528.1$. For $x = 3$, $\hat{y} = 602.8 - 74.7(3) = 378.7$. For $x = 5$, $\hat{y} = 602.8 - 74.7(5) = 229.3$. For $x = 7$, $\hat{y} = 602.8 - 74.7(7) = 79.9$. (c) For $x = 1$, residual $= y - \hat{y} = 578 - 528.1 = 49.9$. For $x = 3$, residual $= y - \hat{y} = 317 - 378.7 = -61.7$. For $x = 5$, residual $= y - \hat{y} = 203 - 229.3 = -26.3$. For $x = 7$, residual $= y - \hat{y} = 118 - 79.9 = 38.1$. (d)



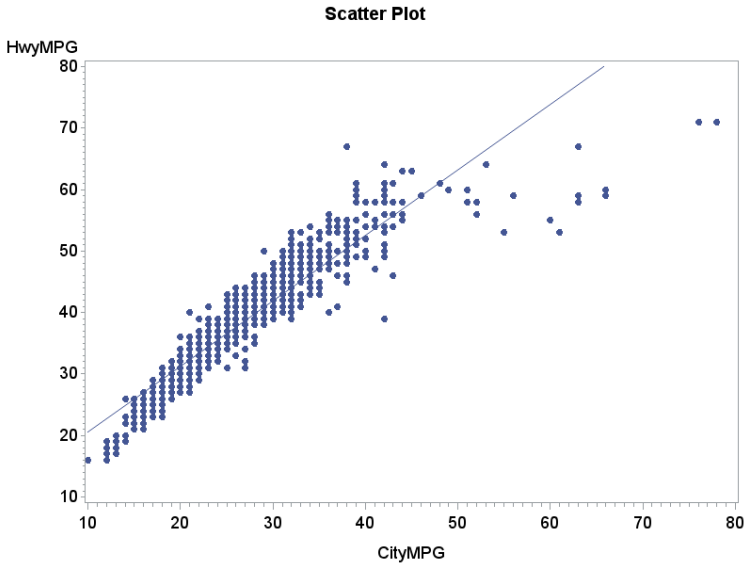
(e) The residual plot shows a curve, suggesting that the linear model we used is not appropriate.

2.61 (a) $\hat{y} = 6.59306 - 0.26062x$. (b) For $x = 1$, $\hat{y} = 6.59306 - 0.26062(1) = 6.3324$. For $x = 3$, $\hat{y} = 6.59306 - 0.26062(3) = 5.8112$. For $x = 5$, $\hat{y} = 6.59306 - 0.26062(5) = 5.29$. For $x = 7$, $\hat{y} = 6.59306 - 0.26062(7) = 4.7687$. (c) For $x = 1$, residual $= y - \hat{y} = 6.3596 - 6.3324 = 0.0271$. For $x = 3$, residual $= y - \hat{y} = 5.7589 - 5.8112 = -0.0523$. For $x = 5$, residual $= y - \hat{y} = 5.3132 - 5.29 = 0.0232$. For $x = 7$, residual $= y - \hat{y} = 4.7707 - 4.7687 = 0.0020$. (d)

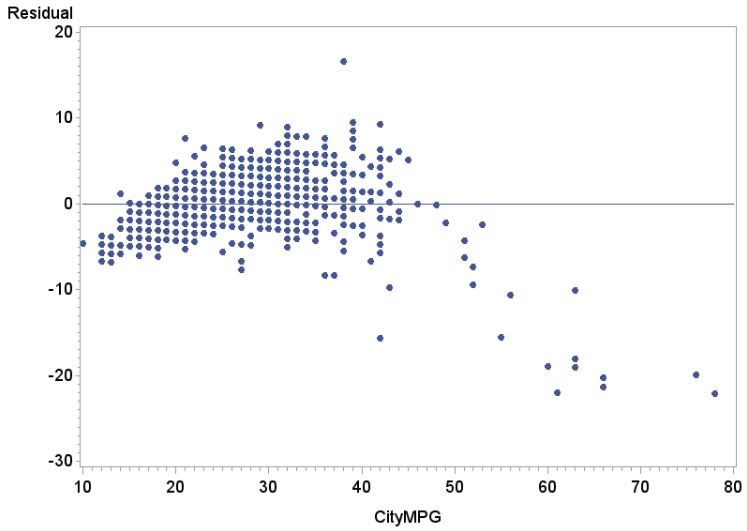


2.62 (a) $\hat{y} = 9.93714 + 1.06613x$. (b) For $x = 42$, $\hat{y} = 9.93714 + 1.06613(42) = 54.7146$. (c) Residual = $y - \hat{y} = 38 - 54.7146 = -16.7146$.

2.63 (a) The vehicles with high City MPG don't follow the regression line; rather, they have a much lower Hwy MPG than the regression line would predict.



(b) For the vehicles with high City MPG, all of the residuals are negative, creating a curve in the plot, suggesting a possible transformation is necessary.

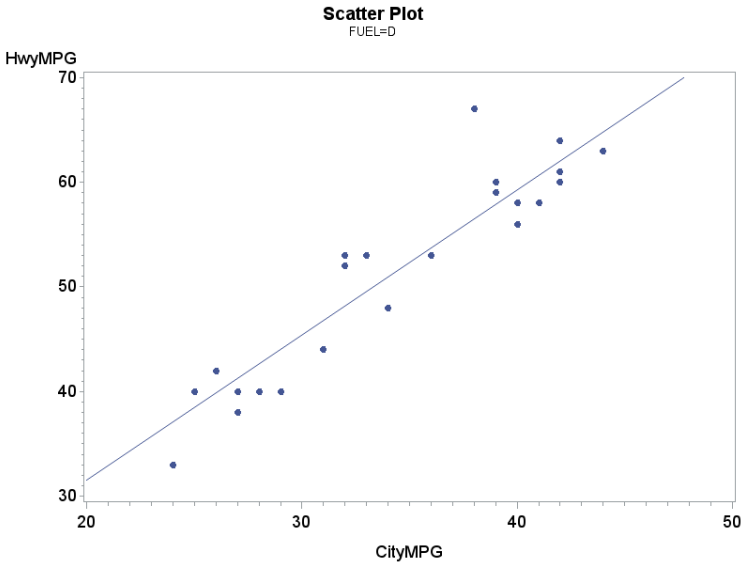


(c) Because the hybrid vehicles have an electric motor in addition to the conventional motor, which is intended to improve City MPG, we would expect them to have a much better City MPG than expected, which is why their residuals fall so far below the residuals for the conventional motor vehicles. (d) Three Toyota Prius models and two Toyota Camry Hybrid models likely are hybrids.

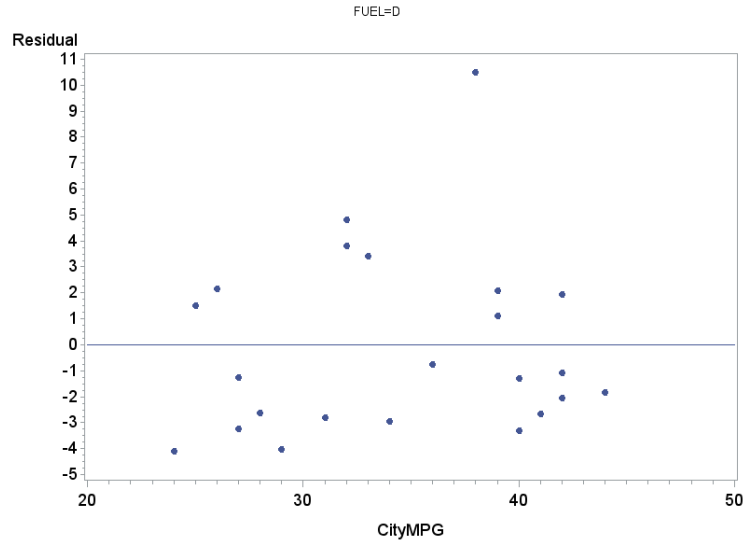
2.64

For Fuel Type D:

$\hat{y}$ = 3.78194 + 1.3877x. For x = 42, $\hat{y}$ = 3.78194 + 1.3877(42) = 62.06534. Residual = $y - \hat{y}$ = 38 – 62.06534 = –24.06534. The scatterplot shows a very strong linear regression between City MPG and Hwy MPG for vehicles with fuel type D.

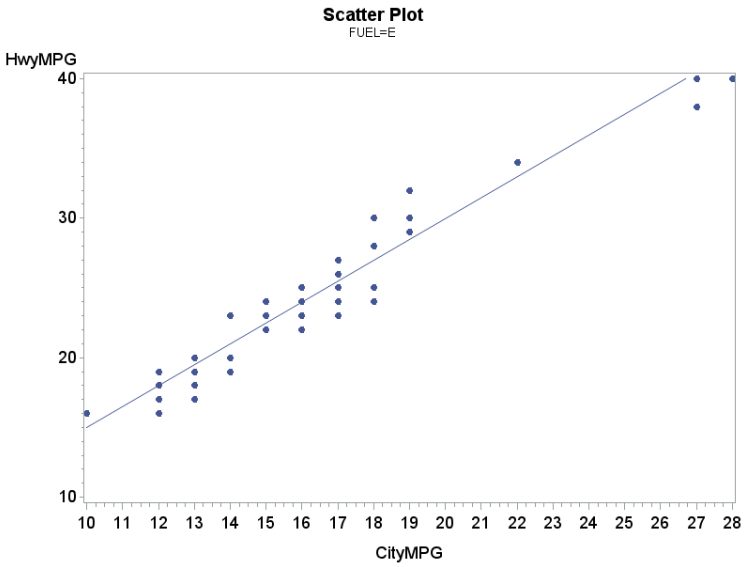


Overall, the residual plot looks good, showing a random scattering of points with one possible outlier.

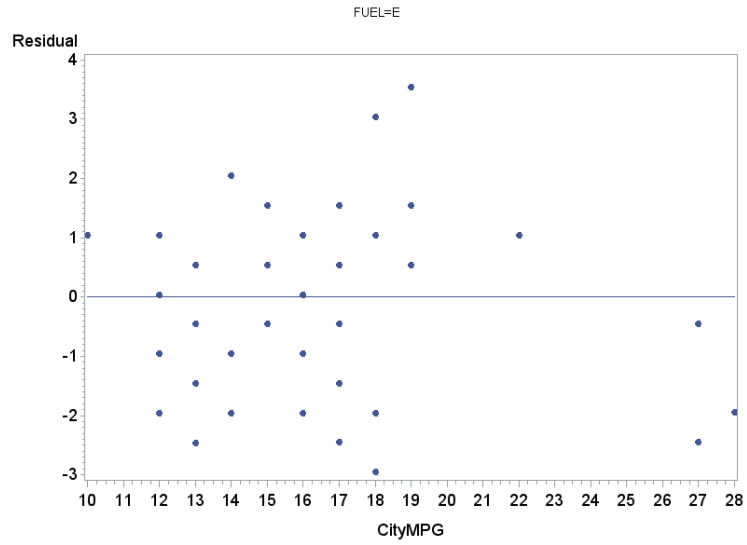


For Fuel Type E:

$\hat{y}$ = –0.02945 + 1.4991x. For x = 42, $\hat{y}$ = –0.02945 + 1.4991(42) = 62.93275. Residual = $y - \hat{y}$ = 38 – 62.93275 = –24.93275. The scatterplot shows a very strong linear regression between City MPG and Hwy MPG for vehicles with fuel type E, however, there are a few vehicles that have much higher MPG for both city and highway than the rest of the vehicles.



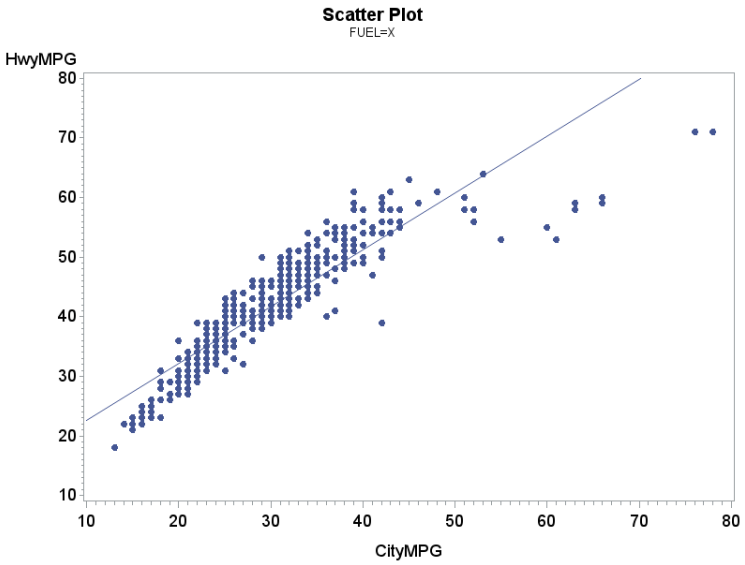
Overall, the residual plot looks good, showing a random scattering of points. We do again see the small group of outliers with much larger MPGs than the other vehicles.



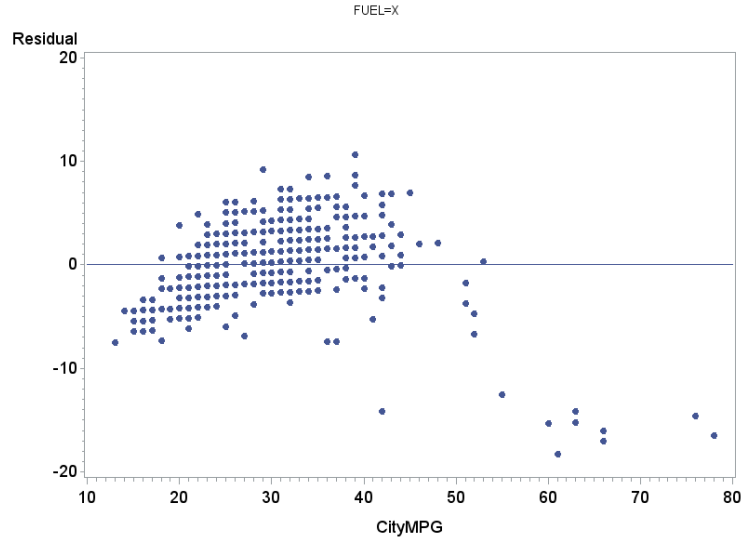
Looking at the data we find that the group of outliers are the FOCUS FF vehicles, which are flex fuel cars; this accounts for their much larger MPG for both city and highway.

For Fuel Type X:

$\hat{y} = 13.14242 + 0.95305x$. For $x = 42$, $\hat{y} = 13.14242 + 0.95305(42) = 53.17052$. Residual $= y - \hat{y} = 38 - 54.7146 = -15.17052$. For fuel type X, the scatterplot shows that the vehicles with high City MPG don't follow the regression line; rather, they have a much lower Hwy MPG than the regression line would predict. These are the same hybrids we found in 2.63.



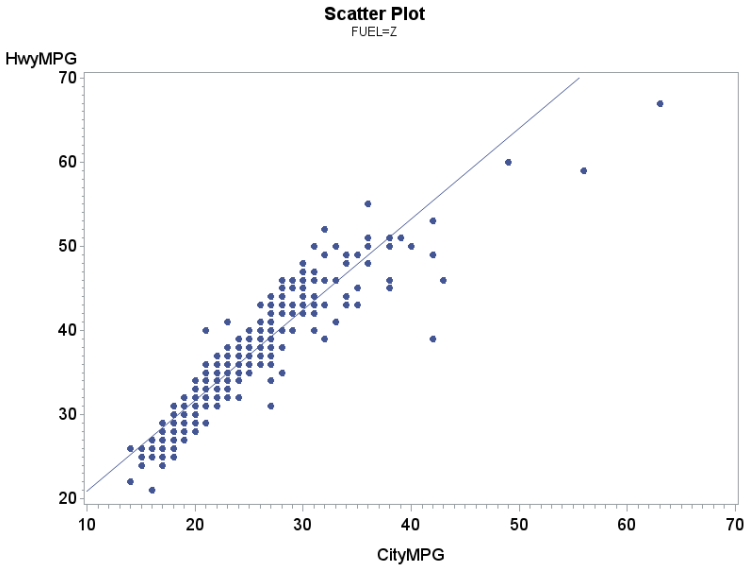
The residual plots show that for the vehicles with high City MPG, all of the residuals are negative, creating a curve in the plot suggesting that a possible transformation is necessary.



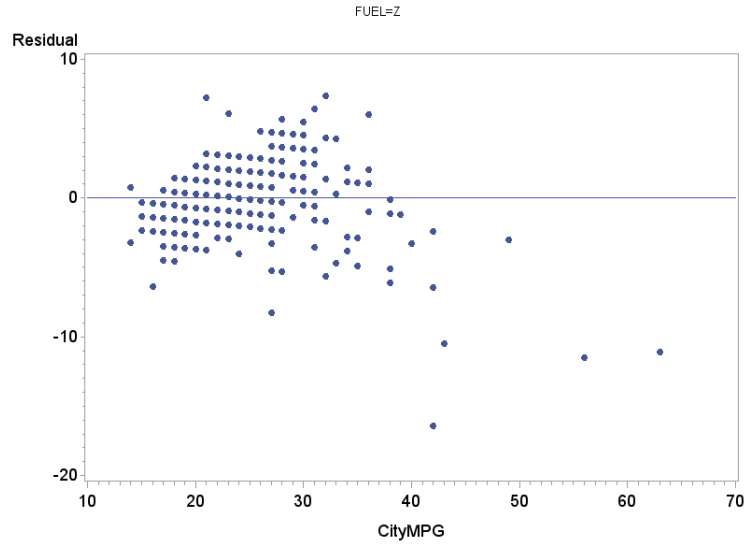
Because the hybrid vehicles have an electric motor in addition to the conventional motor, which is intended to improve City MPG, we would expect them to have a much better City MPG than expected, which is why their residuals fall so far below the residuals for the conventional motor vehicles.

For Fuel Type Z:

$\hat{y} = 10.13307 + 1.07858x$. For $x = 42$, $\hat{y} = 10.13307 + 1.07858(42) = 55.43343$. Residual $= y - \hat{y} = 38 - 54.7146 = -17.43343$. For fuel type X, the scatterplot shows that the vehicles with high City MPG don't follow the regression line; rather, they have a much lower Hwy MPG than the regression line would predict. Again, these are the same hybrids we found in 2.63.

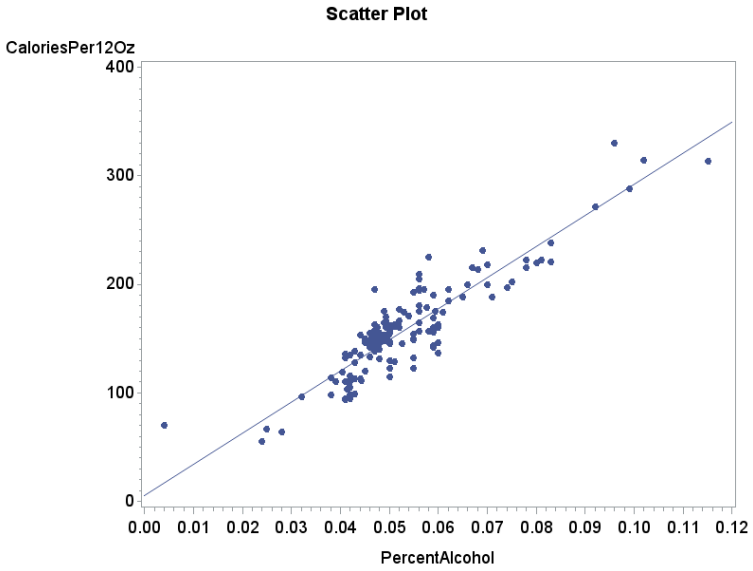


The residual plot show that for the vehicles with high City MPG, all of the residuals are negative, creating a curve in the plot suggesting that a possible transformation is necessary.



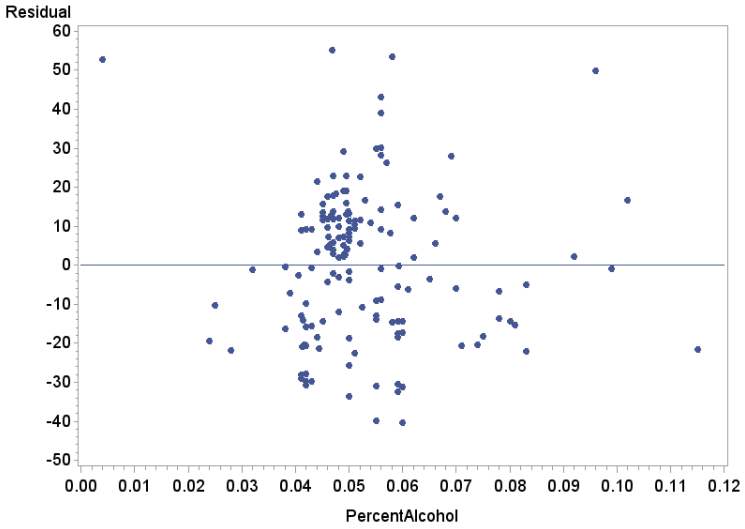
Because the hybrid vehicles have an electric motor in addition to the conventional motor, which is intended to improve City MPG, we would expect them to have a much better City MPG than expected, which is why their residuals fall so far below the residuals for the conventional motor vehicles.

2.65 (a) $\hat{y} = 5.74804 + 2858.85x$. (b)



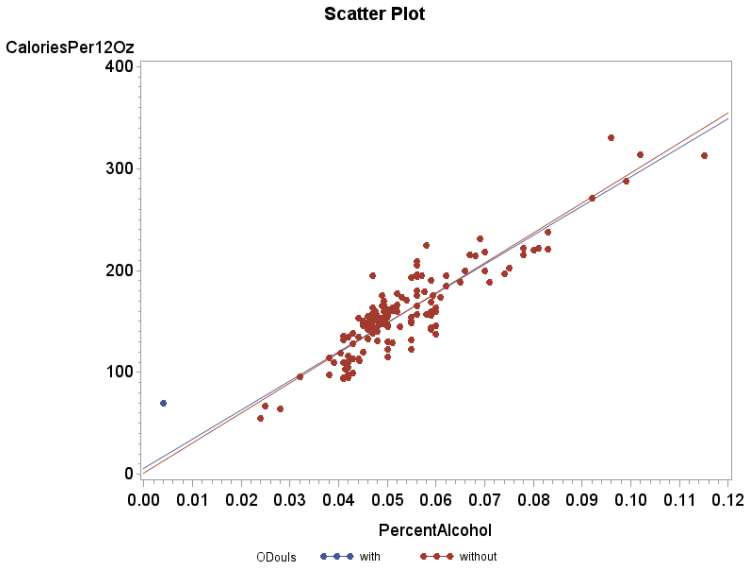
2.66 (a) For $x = 0.052$, $\hat{y} = 5.74804 + 2858.85(0.052) = 154.4082$. (b) Residual $= y - \hat{y} = 160 - 154.4082 = 5.5918$.

2.67 (a)



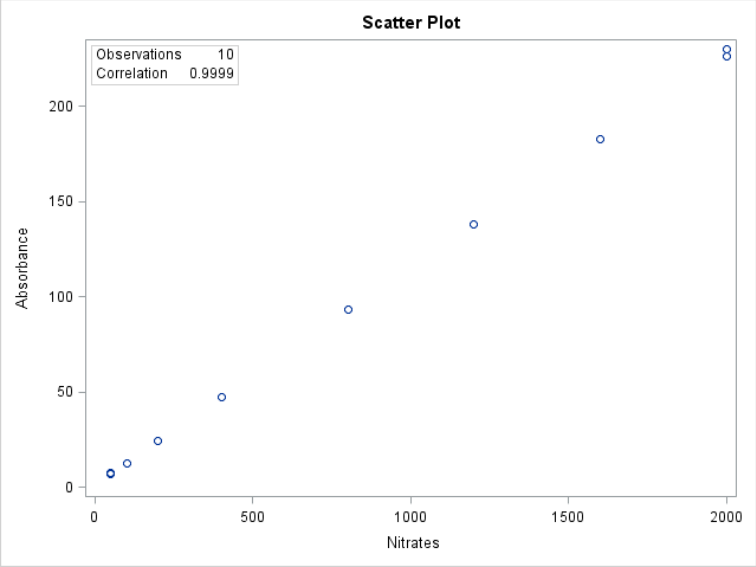
(b) The residual plot looks fairly random, although there is one potential low outlier with a very small percent alcohol. (c) There is nothing unusual about the location of New Belgium Fat Tire, it is right in the middle of the plot among many other similar brands of beer.

2.68 (a) $\hat{y} = 0.96047 + 2944.06159x$. (b)



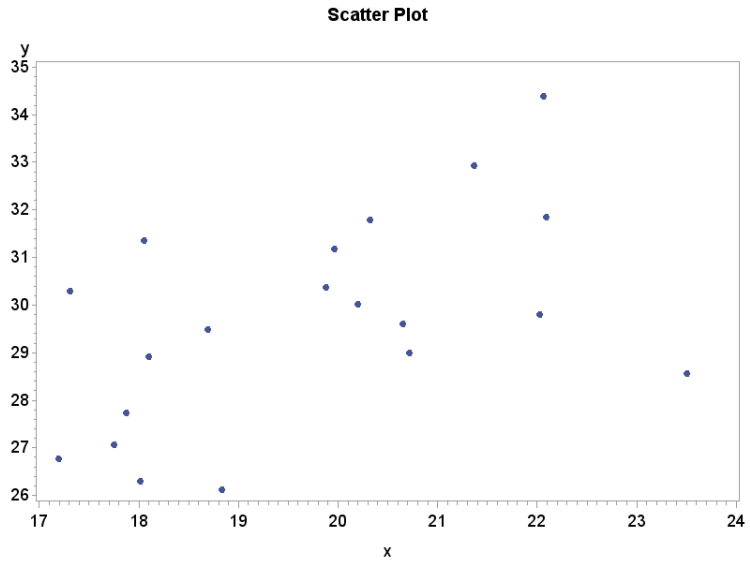
(c) No, O'Doul's is not influential, the regression lines with and without it are nearly identical.

2.69 (a) The correlation is 0.9999, so the calibration does not need to be repeated.

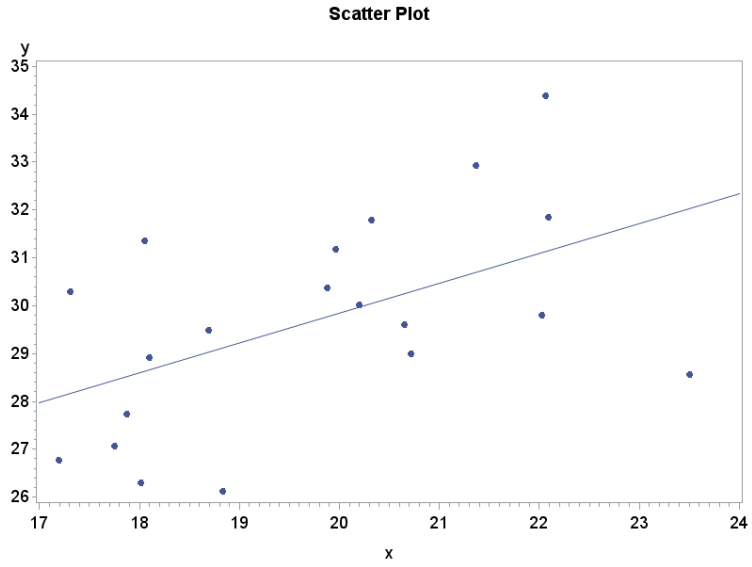


(b) $\hat{y} = 1.65709 + 0.1133x$. For $x = 500$, $\hat{y} = 1.65709 + 0.1133(500) = 58.30709$. Because the relationship is so strong, $r = 0.9999$, we would expect our predicted absorbance to be very accurate.

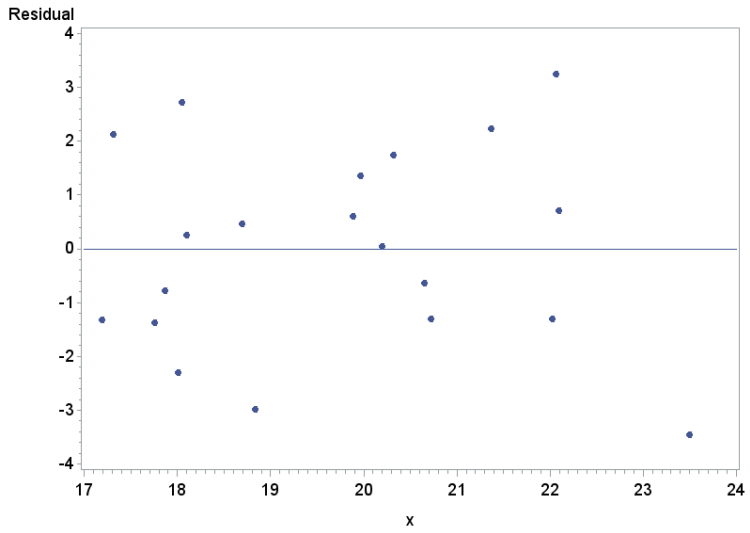
2.70 (a) There seems to be a weak positive linear relationship between y and x.



(b) $\hat{y} = 17.38036 + 0.62332x$.

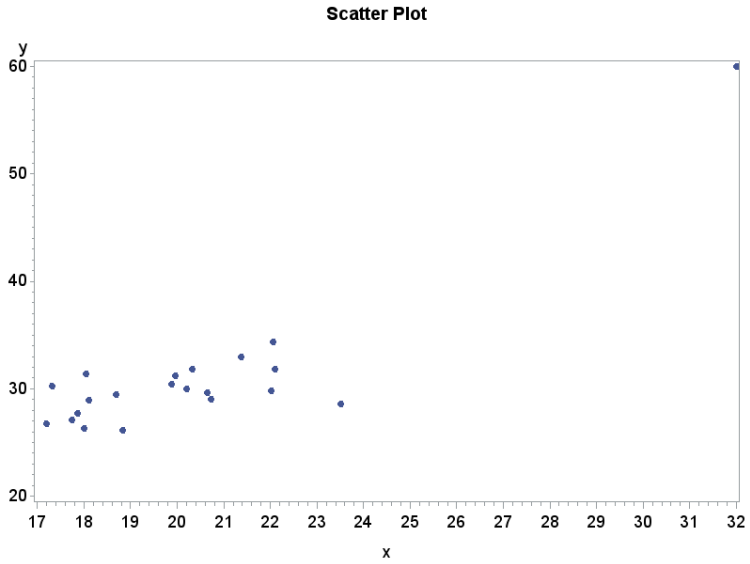


(c)

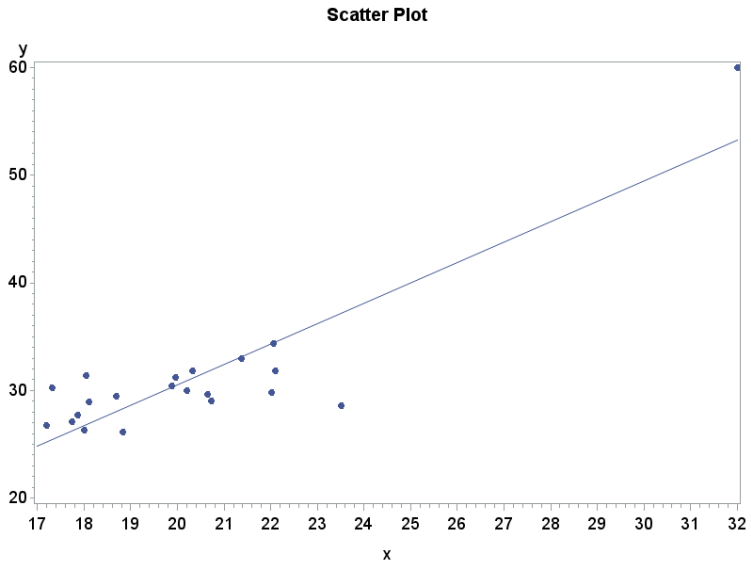


(d) $r^2 = 0.2737$ or 27.37%. (e) The x variable only accounts for 27.37% of the variation in y, so the relationship is fairly weak. The residual plot shows a random scattering suggesting a good fit, albeit weak.

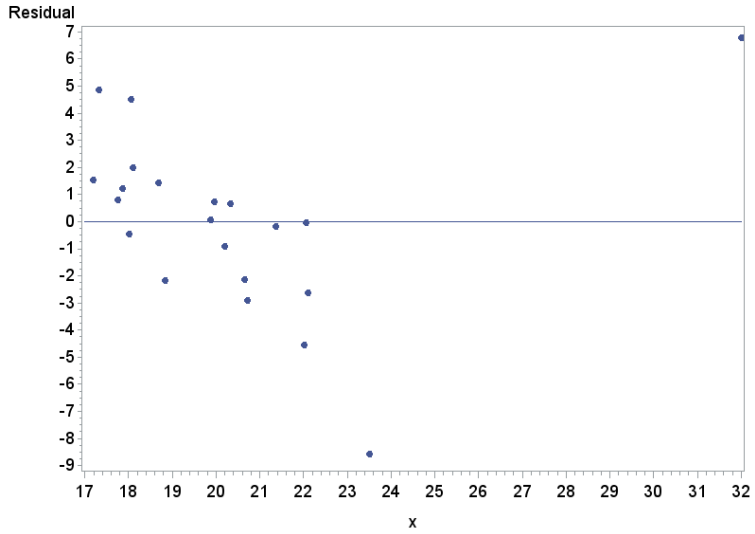
2.71 (a) There seems to be a weak positive linear relationship between y and x, but with one extreme outlier with a very high x-value.



(b) $\hat{y} = -7.28789 + 1.89089x$.

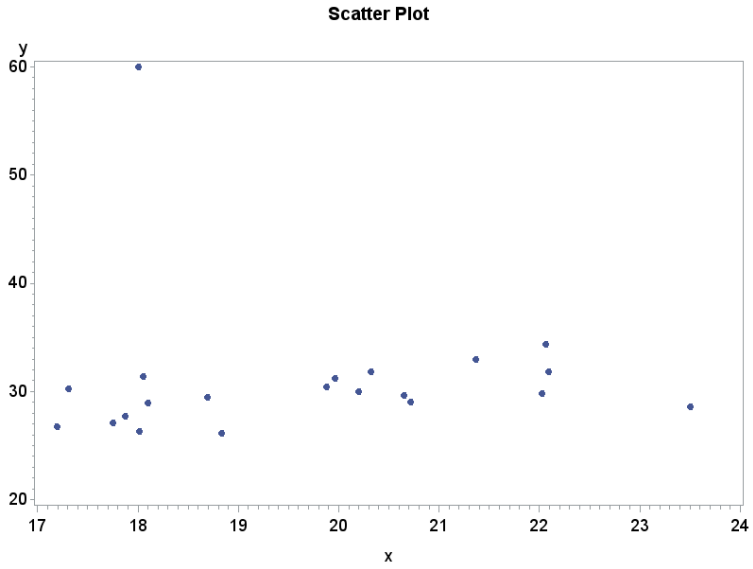


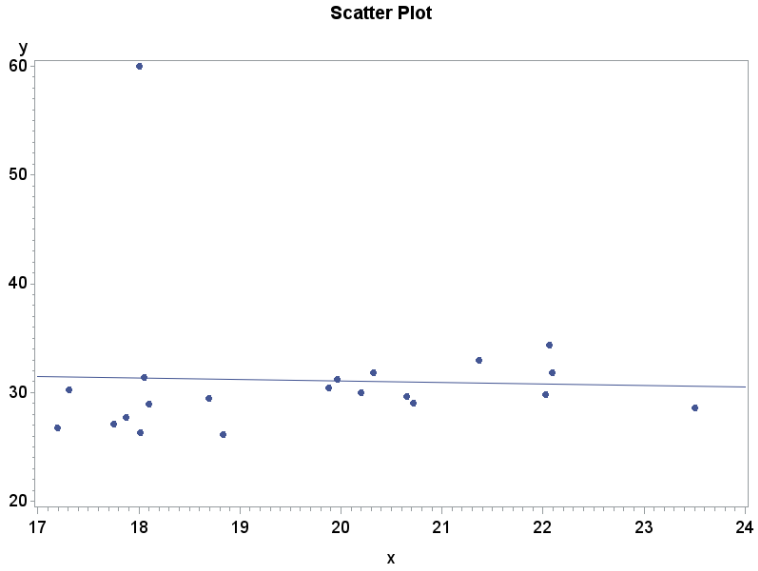
(c)

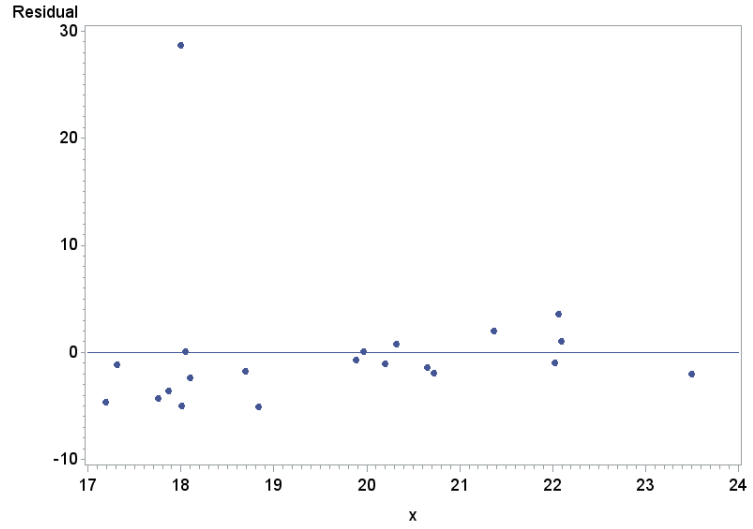


(d) $r^2 = 0.7715$ or 77.15% (e) Although the *x* variable accounts for 77.15% of the variation in *y*, there is a very high outlier for *x*, which pulls the regression line unnaturally. This is seen in the jump of the R-square value from 27% up to 77%, indicating that this observation is very influential in the analysis. This is also demonstrated by the systematic pattern in both the scatterplot and the residual plot, with most of the data points forming a line except for the outlier.

2.72 (a) There seems to be a weak positive linear relationship between *y* and *x* but with one extreme outlier with a very high *y*-value. $\hat{y} = 33.70526 - 0.13152x$. $r^2 = 0.0012$ or 0.12%. Here the outlier completely eliminated all evidence of a regression line. The *x* variable accounts for 0.12% of the variation in *y*, meaning there is no linear relationship at all. However, we know this is wrong because it is mostly due to the outlier unnaturally twisting the regression line.


Scatter Plot
y
60
50
40
30
20
17
18
19
20
21
22
23
24
x


Scatter Plot
y
60
50
40
30
20
17
18
19
20
21
22
23
24
x



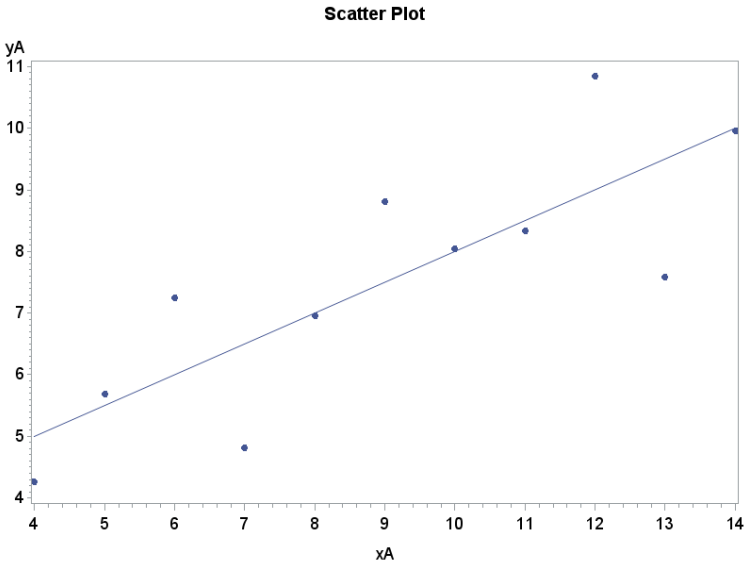
(b) In exercise 2.71, the high x outlier drastically increased the relationship between y and x, increasing the r^2 from 27% to 77%. In this exercise, 2.72, the y outlier drastically decreased the relationship between y and x, changing the r^2 from 27% to essentially 0%. This demonstrates that a single outlier can be very influential and can mislead our interpretation of the relationship between y and x if not careful.

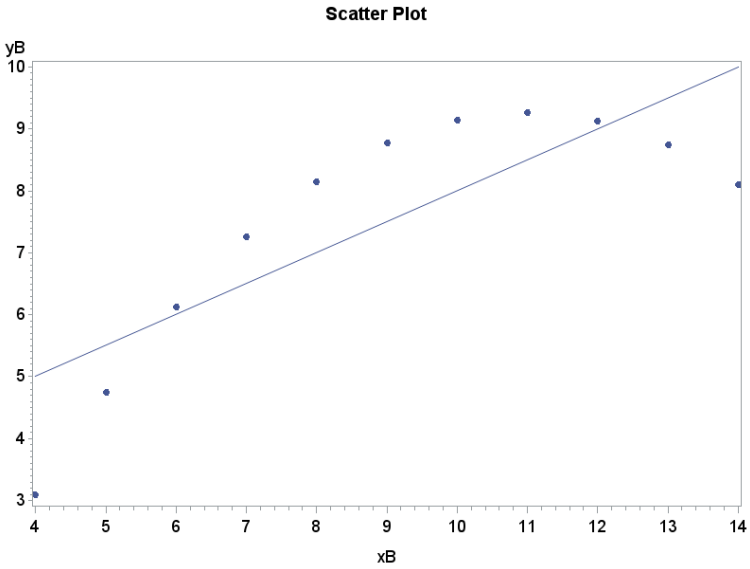
2.73 Applet, answers will vary.

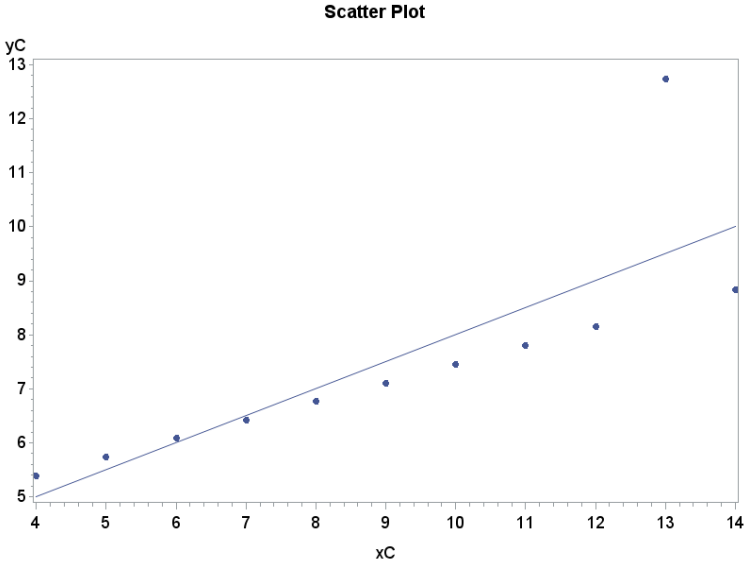
2.74 Applet, answers will vary.

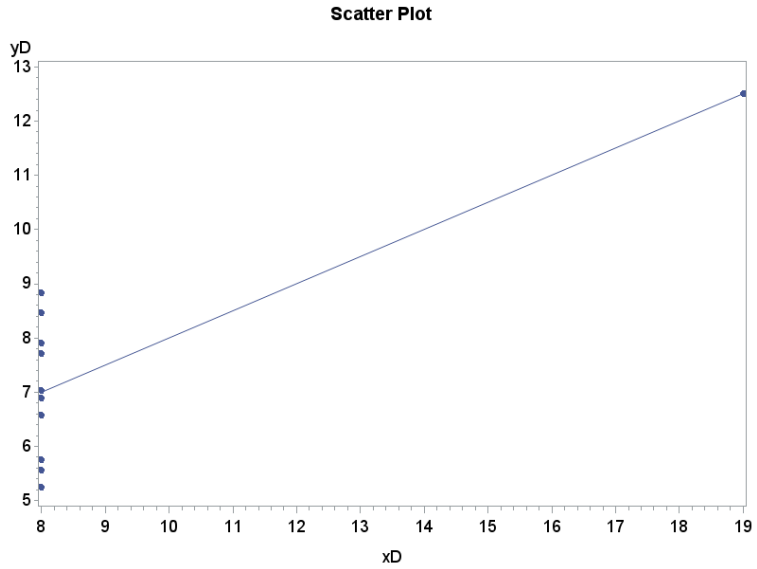
2.75 r = sqrt(0.49) = 0.7, but the value should be negative because the relationship was negative as described, so $r = -0.7$.

2.76 (a) The correlations and regression lines for all four datasets are essentially the same: $r = 0.82$ and $\hat{y} = 3 + 0.5x$. For $x = 10$, $\hat{y} = 3 + 0.5(10) = 8$. (b)



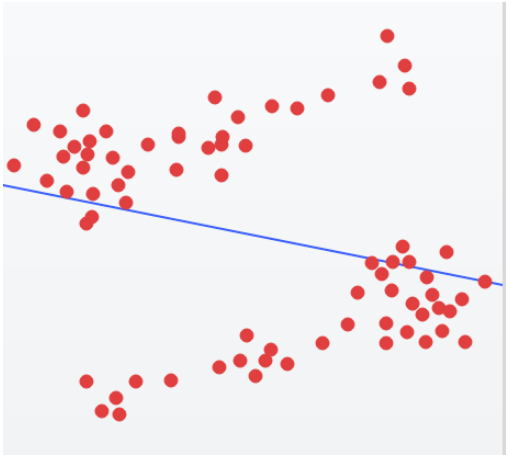

Scatter Plot
yB
10
9
8
7
6
5
4
3
4
5
6
7
8
9
10
11
12
13
14
xB


Scatter Plot
yC
13
12
11
10
9
8
7
6
5
4
5
6
7
8
9
10
11
12
13
14
xC



(c) For Data A, the regression line is a reasonable fit for the data. For Data B, there is obviously a curve in the scatterplot and a transformation is needed before a regression should be run. For Data C, this is a perfect relationship but with one outlier, which is very influential; a regression would not be valid for this data. For Data D, there is no relationship between *y* and *x* at all; only the extreme *x* outlier even makes the regression line possible but it is definitely inappropriate to use regression on this dataset.

2.77



2.78 Answers will vary. One example is a good manager that encourages his or her employees. This could both cause the employees to do better in their job and help them have more self-esteem.

2.79 No, the condition of the patient is a possible lurking variable because larger hospitals likely have more resources to treat more seriously injured or ill patients, which also explains why they stay longer.

2.80 The lurking variable is the size of the fire. If the fire is large, it likely requires more firefighters but will also likely cause more damage.

2.81 (a) Whether the relationship is negative or positive does not tell us anything as to whether or not there is causation. (b) A lurking variable can be categorical. (c) It is actually impossible for all the residuals to be negative. Even if many of the residuals are negative, this tells us nothing regarding the

relationship (positive or negative) of the variables. We need to look at the slope to determine if the relationship is positive or negative.

2.82 (a) An outlier with an extreme *x*-value can be very close to the line, which would make its residual quite small. In fact, many *x* outliers unnaturally pull the regression line toward themselves so that often their residual isn't large. (b) Extrapolating is predicting outside the range of the data; because 2.5 is within the range of the *x*-values (between 0 and 5), this is not extrapolation. (c) This is just wrong, high correlation does **not** imply causation.

2.83 If we predicted just next year's sales, the prediction would probably be reasonable assuming there were no major changes in the company. However, we should not try to predict sales 5 years from now; there is no guarantee the sales will still follow a straight line, as there could be significant changes in the company in the next 5 years.

2.84 It is likely that as people get older, and more experienced, their salaries go up, increasing the overall average for all workers. And yet, as these same workers age they will eventually move to an older age group, where it is possible they would not be making as much on average as this higher age group's income, thus pulling down the average for their new group. Also, because they were likely one of the highest paid in their previous age group (because they were one of the oldest), when they move to the new age group the average of the previous age group also goes down. So, although their pay went up, raising the overall average, each group's average goes down.

2.85 Answers will vary. A simple example is: a married man may be willing to work more (for various reasons), which raises his income.

2.86 Although the data are linear during the summer months, the relationship is very likely to change during the fall and winter.

2.87 Answers will vary. One plausible explanation is the opposite relationship, that being heavy causes the switch to artificial sweeteners instead of sugar.

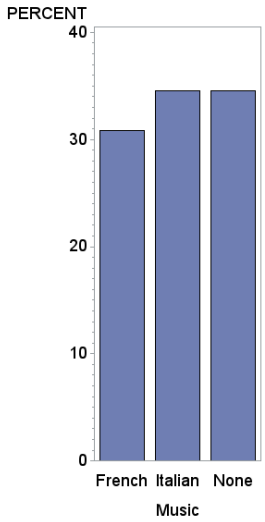
2.88 The explanatory variable is the herbal tea usage; the response is the health and cheerfulness of the residents. Answers will vary. One example of a lurking variable is the interaction between the college students and the residents, which causes the better attitudes and potential health gains.

2.89 Answers will vary. Parents' education or income, socioeconomic status, etc.

2.90 Answers will vary. Wealth or income of the households is a good example, or other environmental factors. For example, it is plausible that lower income neighborhoods are closer to power lines but may also have other negative environmental factors, such as worse water quality, etc.

2.91

Music	Frequency	Percent
French	75	30.86
Italian	84	34.57
None	84	34.57



2.92 Answers will vary.

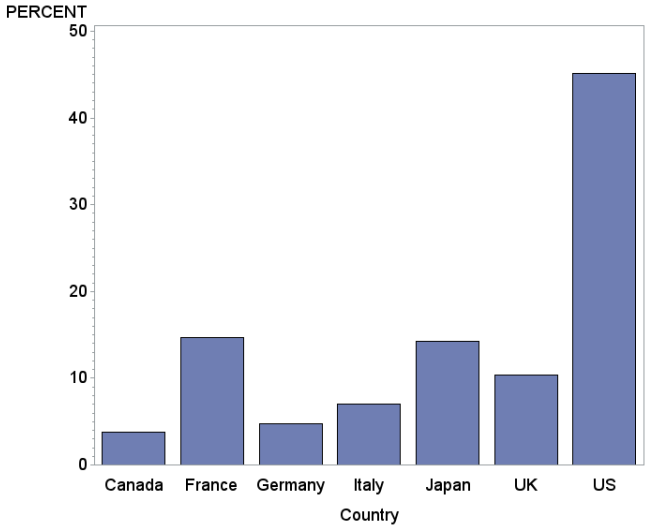
2.93 (a)

Table of FieldOfStudy by Country

FieldOfStudy	Canada	France	Germany	Italy	Japan	UK	US	Total
SocBusLaw	64	153	66	125	259	152	878	1697
SciMathEng	35	111	66	80	136	128	355	911
ArtsHum	27	74	33	42	123	105	397	801
Educ	20	45	18	16	39	14	167	319
Other	30	289	35	58	97	76	272	857
Total	176	672	218	321	654	475	2069	4585

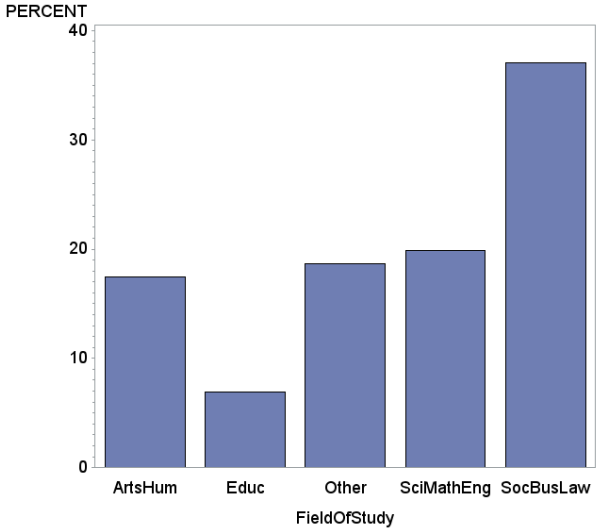
(b)

Country	Frequency	Percent
Canada	176	3.84
France	672	14.66
Germany	218	4.75
Italy	321	7
Japan	654	14.26
UK	475	10.36
US	2069	45.13



(c)

FieldOfStudy	Frequency	Percent
ArtsHum	801	17.47
Educ	319	6.96
Other	857	18.69
SciMathEng	911	19.87
SocBusLaw	1697	37.01



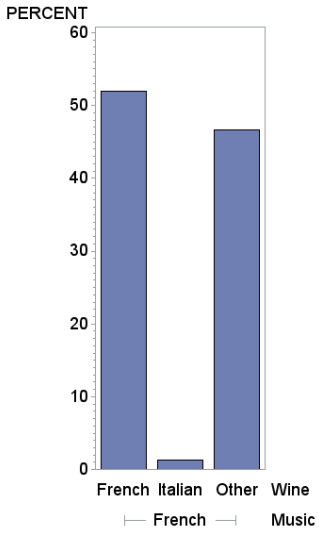
2.94 (a)

Wine	Frequency
French	39
Italian	1
Other	35
Total	75

(b)

Wine	Frequency	Percent
French	39	52
Italian	1	1.33
Other	35	46.67

(c)



(d) Yes, the percent went up from 35.7% to 52%.

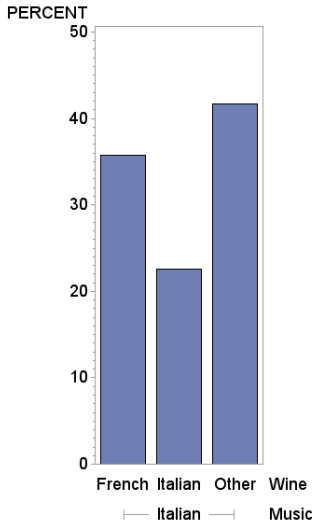
2.95 (a)

Wine	Frequency
French	30
Italian	19
Other	35
Total	84

(b)

Wine	Frequency	Percent
French	30	35.71
Italian	19	22.62
Other	35	41.67

(c)



(d) Yes, the percent went up from 13.1% to 22.62%.

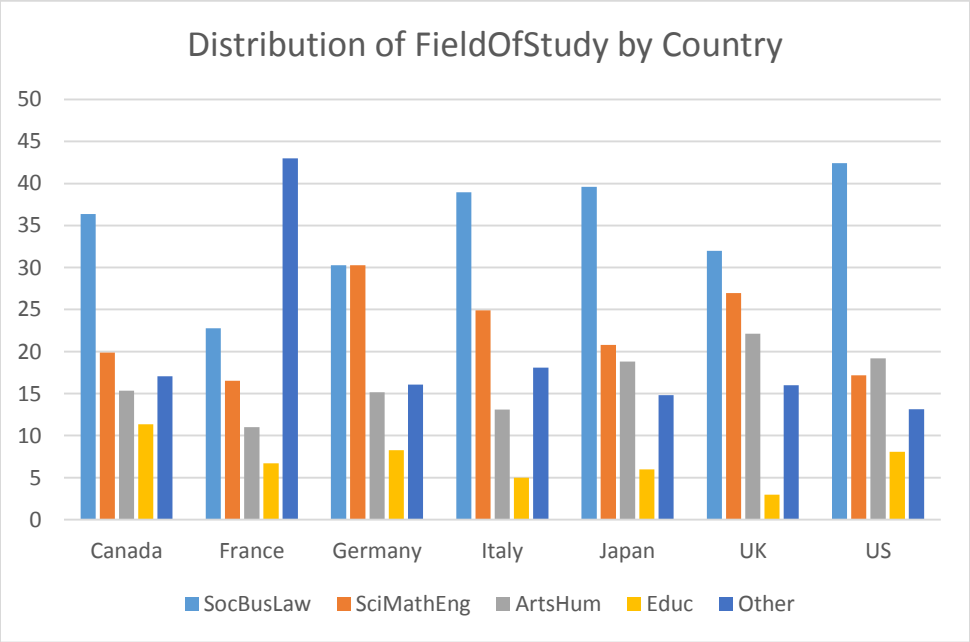
2.96 When French music is playing there is less Italian wine purchased and more French wine purchased, with almost no change in the amount of Other wine purchased. However, when Italian music is playing there is less Other wine purchased and more Italian wine purchased, with almost no change in the amount of French wine purchased.

2.97 (a) **Conditional Distribution of FieldOfStudy given Country**

Table of FieldOfStudy by Country

FieldOfStudy	Canada	France	Germany	Italy	Japan	UK	US
SocBusLaw	36.36	22.77	30.28	38.9	39.6	32	42.4
SciMathEng	19.89	16.52	30.28	24.9	20.8	27	17.2
ArtsHum	15.34	11.01	15.14	13.1	18.81	22.1	19.2
Educ	11.36	6.7	8.26	4.98	5.96	2.95	8.07
Other	17.05	43.01	16.06	18.1	14.83	16	13.2
Total	100	100	100	100	100	100	100

(b)

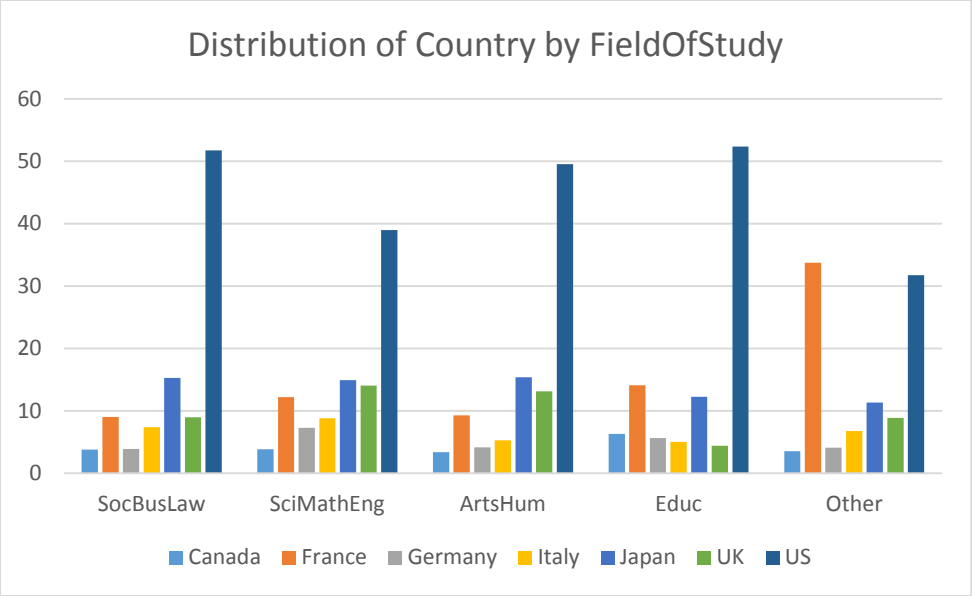


(c) The graph shows that most countries have similarities in their distributions of students among fields of study. Most countries have the most students in Social sciences, business, law, followed second by Science, math, engineering. France, however, is unique as it has a huge percentage in Other, much more than the other countries shown. Also, the UK has an extremely low percentage in Education.

2.98 The US has the most in any field, except Other where France has slightly more students. Other answers will vary.

Conditional Distribution of Country given FieldOfStudy

Table of Country by FieldOfStudy								
FieldOfStudy	**Canada**	**France**	**Germany**	**Italy**	**Japan**	**UK**	**US**	**Total**
SocBusLaw	3.77	9.02	3.89	7.37	15.26	8.96	51.7	100
SciMathEng	3.84	12.18	7.24	8.78	14.93	14.1	39	100
ArtsHum	3.37	9.24	4.12	5.24	15.36	13.1	49.6	100
Educ	6.27	14.11	5.64	5.02	12.23	4.39	52.4	100
Other	3.5	33.72	4.08	6.77	11.32	8.87	31.7	100



2.99 (a) The first approach looks at the distribution of fields of studies within each country, suggesting the popularity of fields among each country. The second approach is somewhat biased or misleading because of different population sizes, so that bigger countries will generally have more students in each field of study. (b) Answers will vary. Most students will likely find the first approach more meaningful. (c) Answers will vary. The first approach shows popularity of fields for each country; the second approach shows which countries have the largest number of students in each field of study.

2.100 63/2100 = 0.03 or 3% of Hospital A's patients died. 16/800 = 0.02 or 2% of Hospital B's patients died.

2.101 (a) For patients in poor condition, 57/1500 = 0.038 or 3.8% of Hospital A's patients died; 8/200 = 0.04 or 4% of Hospital B's patients died. (b) For patients in good condition, while 6/600 = 0.01 or 1% of Hospital A's patients died, 8/600 = 0.0133 or 1.33% of Hospital B's patients died. (c) The percentage of deaths for both conditions is lower for Hospital A, so recommend Hospital A. (d) Because Hospital A had so many more patients in poor condition (1500) compared to good condition patients (600), its overall percentage is mostly representing poor-condition patients, who have a high death rate. Similarly, Hospital B had very few patients in poor condition (200) compared to good condition patients (600), so its overall percentage is mostly representing good-condition patients, who have a low death rate, making their overall percentage lower.

2.102 Overall, only 31% of banks offer RDC. Of those that offer RDC, almost half, 48%, have an asset size of 201 or more. Conversely, of those that don't offer RDC, over half, 59%, have an asset size of under 100.

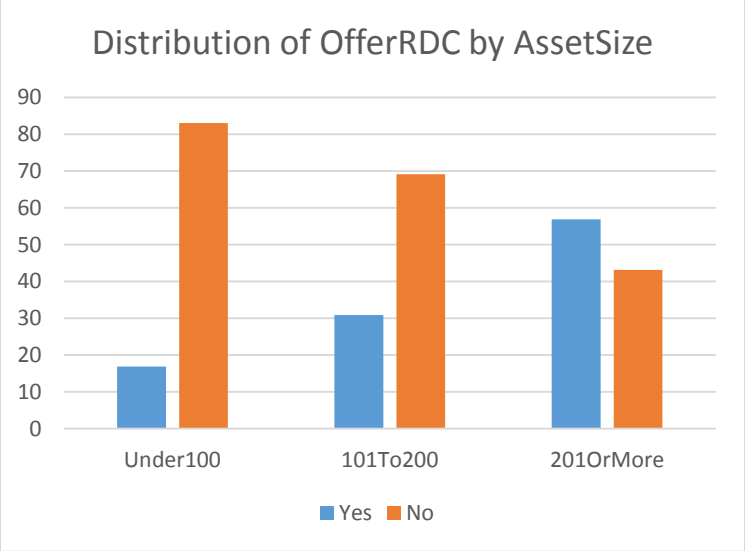
Conditional Distribution of AssetSize given OfferRDC

	OfferRDC	
AssetSize	**Yes**	**No**
Under100	26.92	58.75
101To200	25.21	25.1
201OrMore	47.86	16.16
Total	100	100

Generally speaking, banks that offer RDC are more likely to have larger asset sizes while the banks that don't offer RDC are more likely to have smaller asset sizes. Put differently, as asset size increases it is more likely that the bank will offer RDC.

Conditional Distribution of OfferRDC given AssetSize

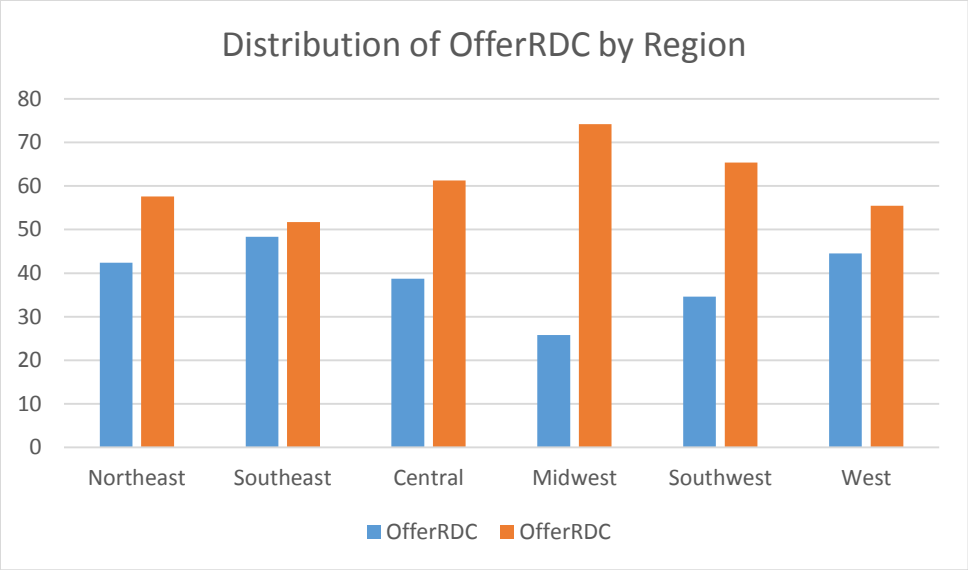
	OfferRDC		
AssetSize	**Yes**	**No**	**Total**
Under100	16.94	83.06	100
101To200	30.89	69.11	100
201OrMore	56.85	43.15	100



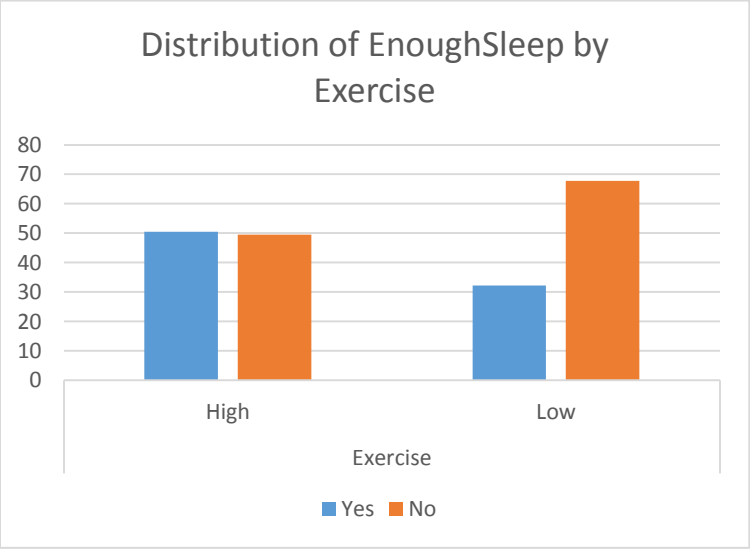
2.103 Only 37% of all banks offer RDC. Regions with high percentages of banks offering RDC are: Southeast (48.31%), West (44.53%), and Northeast (42.42%). Midwest (25.82%) has a low percentage of banks offering RDC.

Conditional Distribution of OfferRDC given Region

	OfferRDC		
Region	**Yes**	**No**	**Total**
Northeast	42.42	57.58	100
Southeast	48.31	51.69	100
Central	38.69	61.31	100
Midwest	25.82	74.18	100
Southwest	34.62	65.38	100
West	44.53	55.47	100



2.104 (a) For high exercisers, 151/299 = 0.505 or 50.5% get enough sleep and 148/299 = 0.495 or 49.5% do not. (b) For low exercisers, 115/357 = 0.322 or 32.2% get enough sleep and 242/357 = 0.678 or 67.8% do not. (c) Mosaic not shown; below is a conditional distribution showing the percentages.



(d) Those who are high exercisers are more likely to get enough sleep than those who are low exercisers.

2.105 (a) For those who get enough sleep, 151/266 = 0.568 or 56.8% are high exercisers and 115/266 = 0.432 or 43.2% are low exercisers. (b) For those who don't get enough sleep, 148/390 = 0.379 or 37.9% are high exercisers and 242/390 = 0.621 or 62.1% are low exercisers. (c) Those who get enough sleep are more likely to be high exercisers than those who don't get enough sleep. (d) Answers will vary.

2.106 (a)

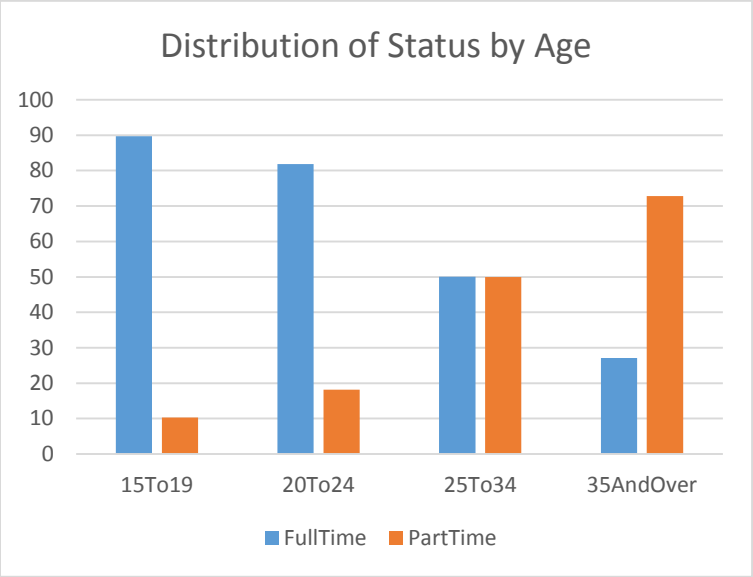
Age	Fulltime
15To19	30.55
20To24	47.23
25To34	15.35
35AndOver	6.87
Total	100

(b)

Age	PartTime
15To19	7.34
20To24	21.97
25To34	32.07
35AndOver	38.61
Total	100

(c) Full-time students tend to be in younger age groups, with over 75% under 25. Part-time students tend to be in older age groups, with over 70% 25 or older.

2.107 (a) For Age 15 to 19: 89.7% are Full-time and 10.3% are Part-time. For Age 20 to 24: 81.82% are Full-time and 18.18% are Part-time. For Age 25 to 34: 50.06% are Full-time and 49.94% are Part-time. For Age 35 and Over: 27.15% are Full-time and 72.85% are Part-time. (b)



(c) Mosaic not shown. (d) Students aged 15–24 are much more likely to be Full-time, while students aged 35 and Over and more likely to be Part-time. Students aged 25–34 are about equally likely to be Full- or Part-time students. (e) Because there are only 2 categories for Status, if we are given the percentage of Full-time students, the percentage of Part-time students must be 100% minus the percentage for Full-time. (f) Both are valid descriptions; it mostly depends on what condition the student(s) you are interested in is. If we are interested in a particular age group, the current analysis likely has more meaning, whereas if we are interested in a particular status, the previous analysis has more meaning.

2.108 (a)

Lied	Male	Female	Total
Yes	5057	5966	11023
No	4145	5719	9864
Total	9202	11685	20887

(b) Answers will vary. Some percentages are shown below.

Conditional Distribution of Lied given Gender

Lied	Male	Female
Yes	54.96	51.06
No	45.04	48.94
Total	100	100

(c) The percentages for both males and females are quite similar. For the males, about 55% admitted that they lied whereas for the females, 51% admitted that they had lied. Males maybe be slightly more willing to admit that they lied than females. Note: this doesn't mean that they are more likely to lie, just that in this study a higher percentage reported that they had lied than their female counterparts.

2.109 There were 21,140 students total; 20,032 agree and 1,108 disagree; 11,358 female and 9,782 male. 96% of females and 93% of males agreed that trust and honesty are essential. A slightly higher percentage of females said that trust and honesty are essential than males but overall the results were quite similar for both males and females.

Conditional Distribution of TrustEssential given Gender

TrustEssential	Male	Female
Yes	93.00	96.28
No	7.00	3.72
Total	100	100

2.110 (a)

	ClassSize							
CourseLevel	**1To9**	**10To19**	**20To29**	**30To39**	**40To49**	**50To99**	**100OrMore**	**Total**
1	202	659	917	241	70	99	123	2311
2	190	370	486	307	84	109	134	1680
3	150	387	314	115	96	186	53	1301
4	146	256	190	83	67	64	17	823
Total	688	1672	1907	746	317	458	327	6115

(b) **Marginal Distribution of Course Level**

CourseLevel	Percent
1	37.79
2	27.47
3	21.28
4	13.46
Total	100

(c) **Marginal Distribution of Class Size**

ClassSize	Percent
1To9	11.25
10To19	27.34
20To29	31.19
30To39	12.20
40To49	5.18
50To99	7.49
100OrMore	5.35
Total	100

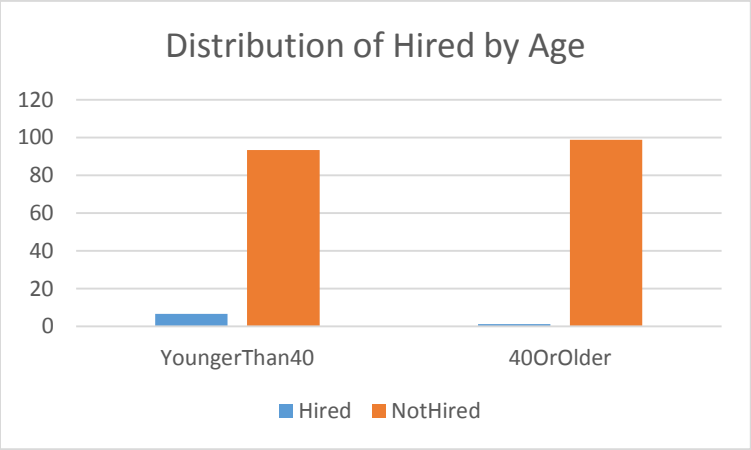
(d) **Conditional Distribution of Class size by Course Level**

	ClassSize							
CourseLevel	**1To9**	**10To19**	**20To29**	**30To39**	**40To49**	**50To99**	**100OrMore**	**Total**
1	8.74	28.52	39.68	10.43	3.03	4.28	5.32	100
2	11.31	22.02	28.93	18.27	5.00	6.49	7.98	100
3	11.53	29.75	24.14	8.84	7.38	14.30	4.07	100
4	17.74	31.11	23.09	10.09	8.14	7.78	2.07	100

(e) As course level increases, frequency decreases. The most common class sizes are 10 to 19 and 20 to 29; these two account for more than half of all classes. For all course levels, again the two most common class sizes are 10 to 19 and 20 to 29. Courses level 4 are more likely to have really small class sizes in the 1 to 9 size. Conversely, lower course levels are more likely to have massive class sizes of 100 or more. Interestingly enough, courses level 3 have by far the most classes in the 50 to 99 range, nearly 2 to 3 times as frequent as the other course levels.

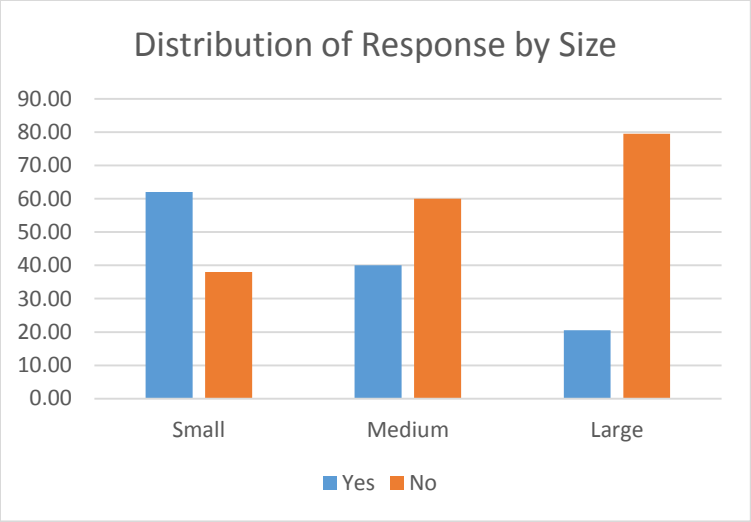
2.111 (a) For younger than 40: 6.6% were hired, 93.4% were not. For 40 or older: 1.18% were hired, 98.82% were not.

(b)



(c) The percentage of hired is greater for the younger than 40 group; the company looks like it is discriminating. (d) Education could be different among groups, making them more or less qualified.

2.112 (a) 355/600 = 0.5917 or 59.17% did not respond. (b) For small companies, 76/200 or 38% did not respond. For medium companies, 120/200 or 60% did not respond. For large companies, 159/200 or 79.5% did not respond. As size increases, so does nonresponse. (c)



2.113 (a) 27,792 never married; 65,099 married; 11,362 widowed; 13,749 divorced. (b) Joint distribution and marginal distributions

Percent	NeverMarried	Married	Widowed	Divorced	Total
18To24	10.26	1.84	0.02	0.14	12.3
25To39	8.03	15.44	0.15	2.12	25.7
40To64	4.43	29.68	2.09	7.35	43.5
65AndOver	0.83	8.21	7.37	2.04	18.5
Total	23.55	55.17	9.63	11.65	100

(c)

Conditional Distribution of Age given Marital Status

Percent	NeverMarried	Married	Widowed	Divorced
18To24	43.58	3.33	0.2	1.19
25To39	34.08	27.99	1.56	18.18
40To64	18.8	53.8	21.68	63.09
65AndOver	3.54	14.88	76.56	17.54
Total	100	100	100	100

Conditional Distribution of Marital Status given Age

Percent	NeverMarried	Married	Widowed	Divorced	Total
18To24	83.7	15	0.16	1.13	100
25To39	31.19	60	0.58	8.23	100
40To64	10.17	68.16	4.79	16.88	100
65AndOver	4.52	44.48	39.93	11.07	100

(d) Not shown. (e) More than half of women are married; of that group, age 40 to 64 is the most common, followed by 25 to 39. Almost 25% never married, but most of that group is represented by younger age groups. Widowed and Divorced have relatively small percentages across the board, though the 65 and Over group is most likely to be widowed and the 40 to 64 group is most likely to be divorced.

2.114 (a) The percentages are shown in the table below. The vast majority, 83.7%, of women aged 18 to 24 have never been married and less than 2% combined have been widowed or divorced. Whereas most of the women aged 40 to 64 are married, 68.16%, only 10.17% have never married and a fair number, over 20% combined, have been widowed or divorced.

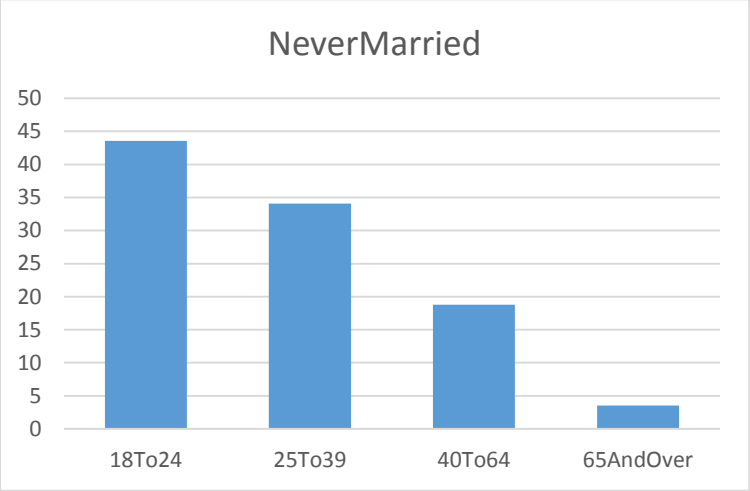
Conditional Distribution of Marital Status for Age Groups 18to24 and 40to64

Percent	NeverMarried	Married	Widowed	Divorced	Total
18To24	83.7	15	0.16	1.13	100
40To64	10.17	68.16	4.79	16.88	100

(b) The percentages along with the bar graph are shown below. The vast majority of women that have never married are between 18 and 39 years old, these two age groups account for more than 75% of women in this category. Your magazine should target these two younger age groups.

Conditional Distribution of Age for Women Never Married

Percent	NeverMarried
18To24	43.58
25To39	34.08
40To64	18.8
65AndOver	3.54
Total	100



2.115 33,748 never married; 64,438 married; 2,968 widowed; 9,964 divorced. More than half of men are married; of that group, age 40 to 64 is the most common, followed by 25 to 39 and 65 and Over. More than 30% never married, very few of which are 65 and Over. Fewer than 3% of men are widowed, and the vast majority are 65 and Over. About 9% are divorced, two-thirds in the 40 to 64 age group.

Joint and marginal distributions

Percent	NeverMarried	Married	Widowed	Divorced	Total
18To24	12.16	1.12	0.01	0.06	13.3
25To39	11.42	14.43	0.07	1.61	27.5
40To64	6.18	31.18	0.68	5.98	44
65AndOver	0.62	11.26	1.91	1.32	15.1
Total	30.37	57.99	2.67	8.97	100

Conditional Distribution given Marital Status

Percent	NeverMarried	Married	Widowed	Divorced
18To24	40.03	1.93	0.2	0.63
25To39	37.59	24.88	2.63	17.96
40To64	20.35	53.77	25.61	66.71
65AndOver	2.03	19.42	71.56	14.69
Total	100	100	100	100

Conditional Distribution given Age

Percent	NeverMarried	Married	Widowed	Divorced	Total
18To24	91.14	8.4	0.04	0.43	100
25To39	41.48	52.41	0.26	5.85	100
40To64	14.04	70.82	1.55	13.59	100
65AndOver	4.08	74.55	12.65	8.72	100

2.116 (a)

Gender	Admit	Deny	Total
Male	490	210	700
Female	280	220	500
Total	770	430	1200

(b) 490/700 or 70% of male applicants are admitted, 280/500 or 56% of female applicants are admitted. (c) For Business, 80% of male applicants are admitted and 90% of female applicants are admitted. For Law, 10% of male applicants are admitted and 33% of female applicants are admitted. (d) There are six times as many men that apply to business school than to law school, while the number of women that apply to each school are much closer, but with more women applying to law school. Because the business school has much higher admittance rates, it makes the overall percentage for men much higher when combined with the law school applicants. Similarly, because the law school admittance rates are much worse, the fact that more women applied to law school makes their overall percentage look much worse when combined with the business school applicants.

2.117 Answers will vary. One example is shown below.

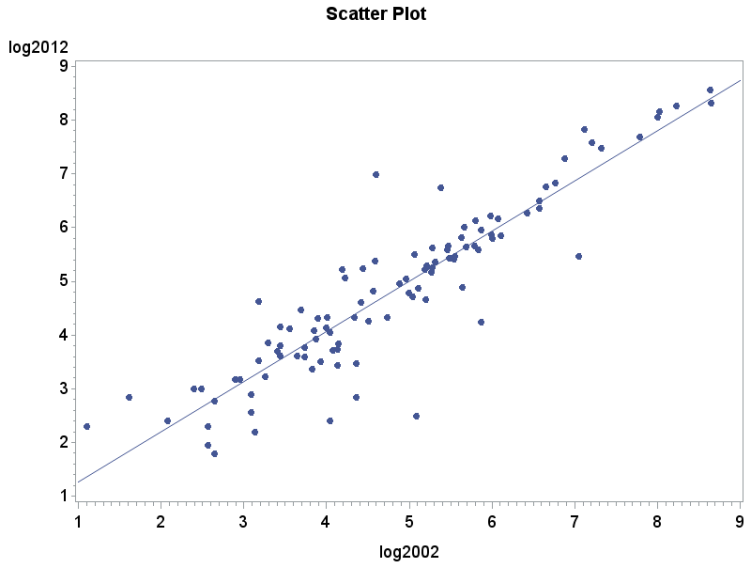
Smokers	Overweight	Not
Early Death	30	260
No	110	1400

Non Smokers	Overweight	Not
Early Death	100	100
No	1000	1200

Combined	Overweight	Not
Early Death	130	360
No	1110	2600

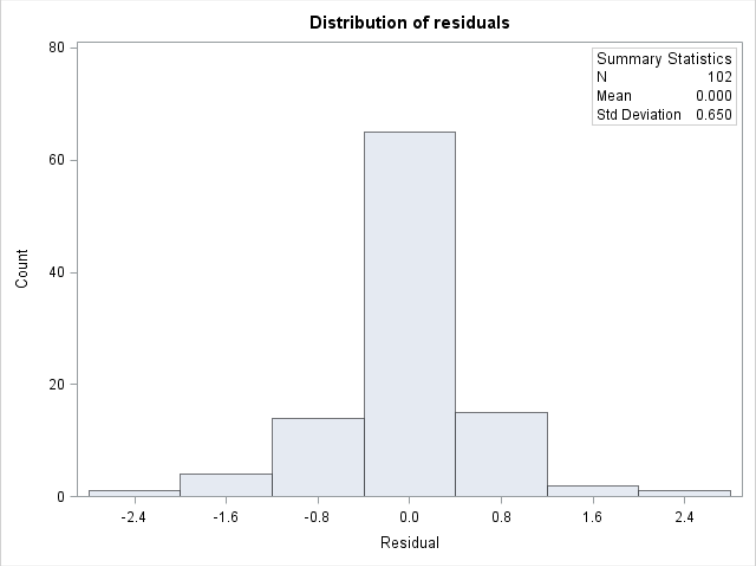
2.118 Answers will vary. For example let $a = 30$, then $b = 30$, $c = 40$, and $d = 20$.

2.119 (a) The log 2002 data (explanatory) should explain the log 2012 data (response). (b) There is a strong positive linear relationship.

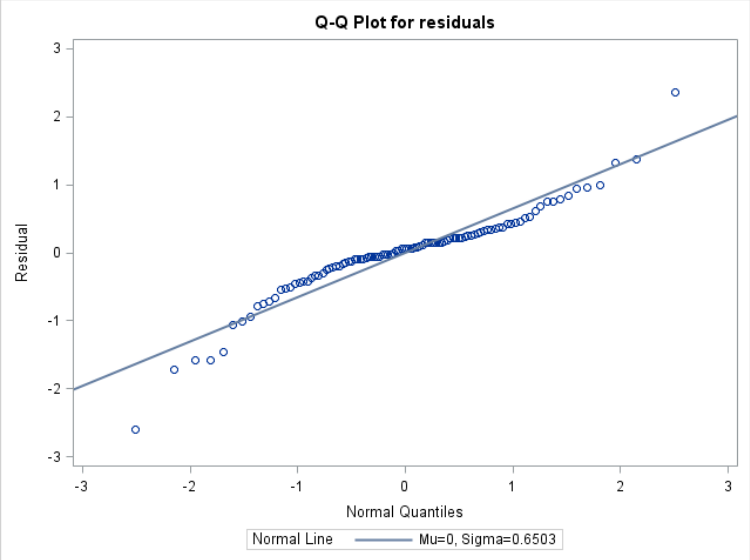


(c) $\hat{y} = 0.33695 + 0.93406x$. (d) $\hat{y} = 5.5935$, residual = 0.2116. (e) $r = 0.9133$. (f) Answers will vary.

2.120 (a)

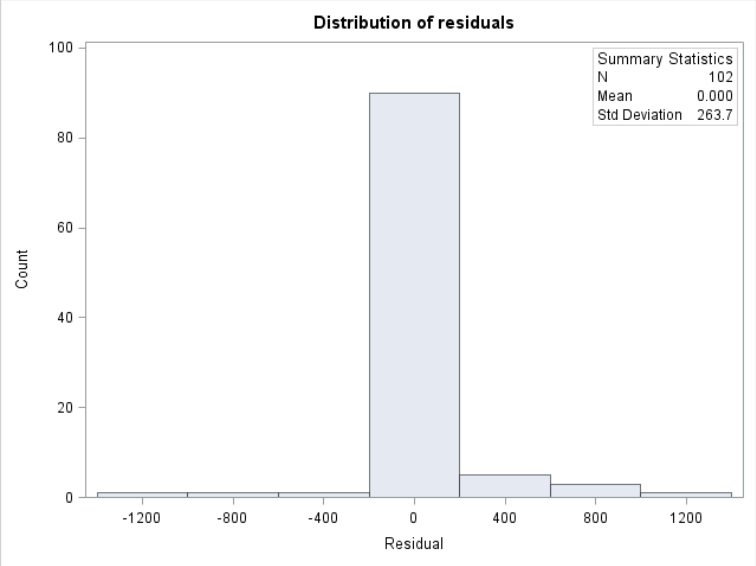


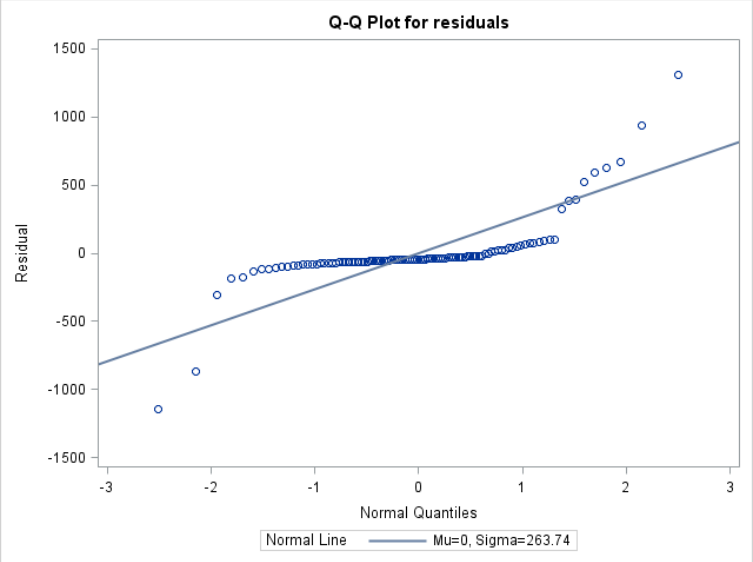
(b)



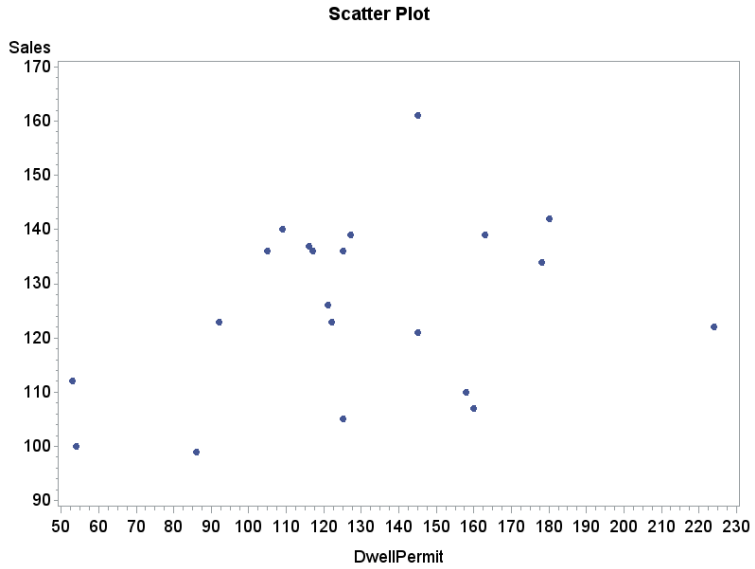
(c) The histogram shows a roughly Normal distribution, however, the Normal quantile plot shows that the distribution has slightly heavy tails and possibly a few outliers.

(d) With the log transformation, the residuals were much closer to being Normally distributed than we see here with the untransformed data. The histogram and Normal quantile plot show very long tails in the residuals creating a non-Normal distribution. As such, we should prefer the log transformed data.

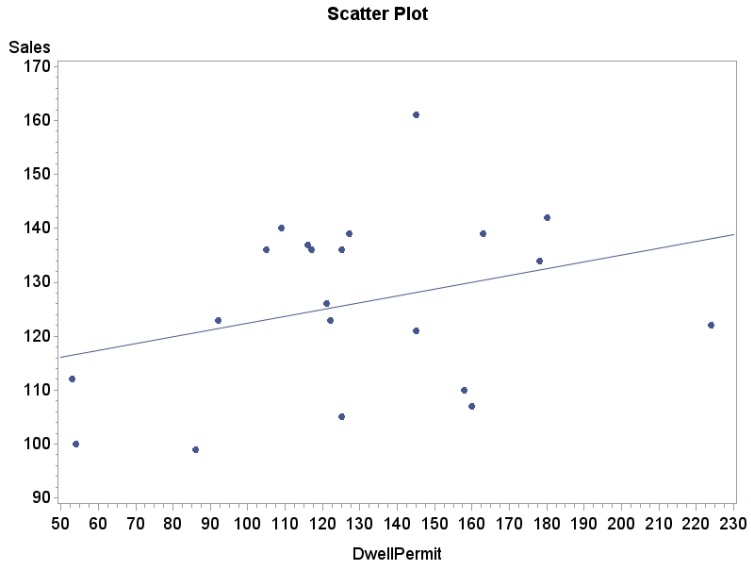




2.121 (a) There is a weak positive linear relationship. There is one high *X* outlier and one high *Y* outlier.

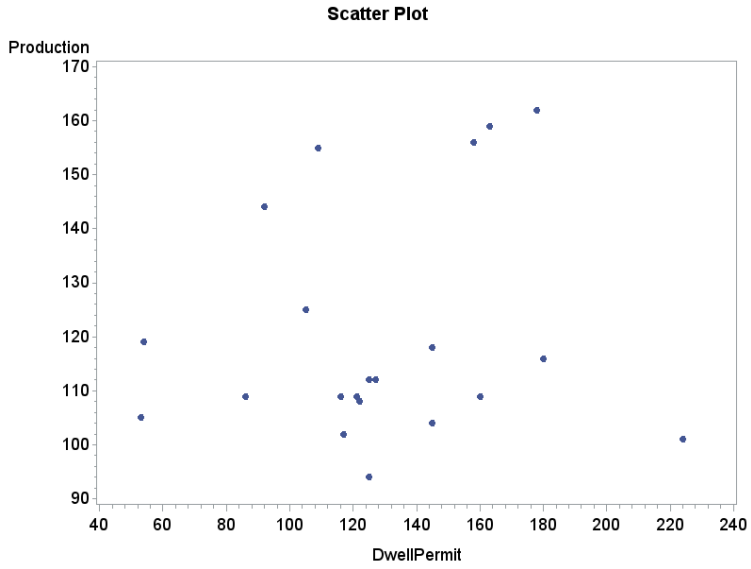


(b) $\hat{y}$ = 109.82043 + 0.12635*x*.

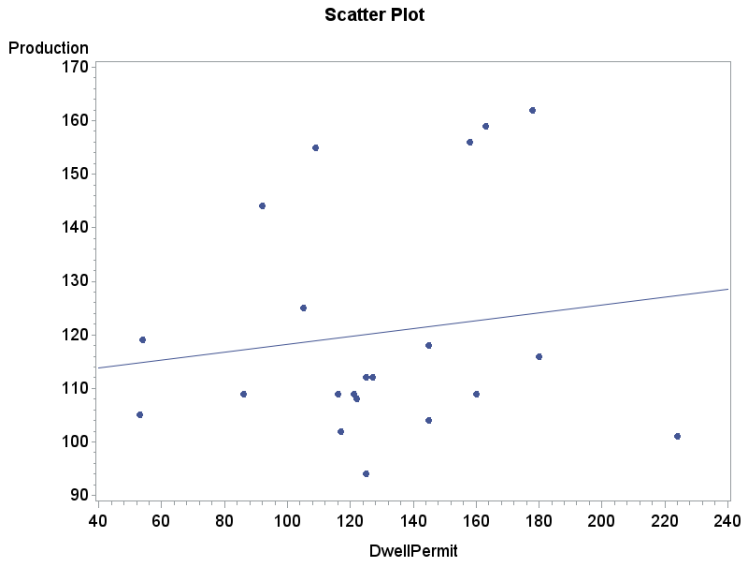


(c) $\hat{y}$ = 109.82043 + 0.12635(160) = 130.0361. (d) Residual = 107 – 130.0361 = –23.0361. (e) R-square = 10.27%.

2.122 (a) There is no real linear relationship between production and dwelling permits.

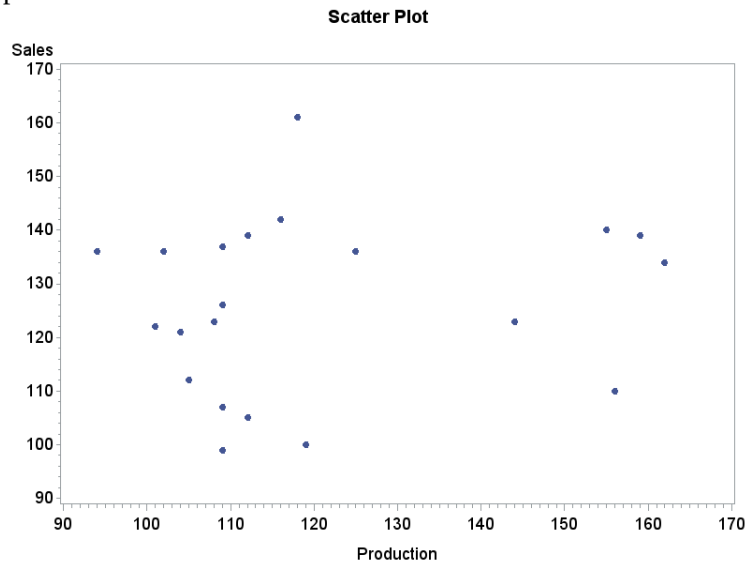


(b) $\hat{y}$ = 110.95813 + 0.07315*x*.

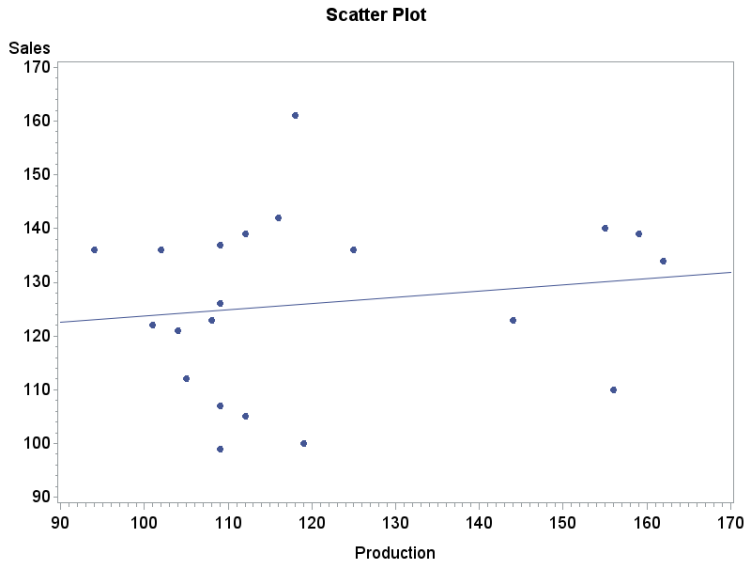


(c) $\hat{y}$ = 110.95813 + 0.07315(160) = 122.6626. (d) Residual = 109 – 122.6626 = –13.6626. (e) R-square = 1.99%. This is much smaller than the R-square value in the previous exercise. It should be noted, however, that neither variable, production or sales, had a very strong relationship to the number of permits issued for new dwellings.

2.123 (a) There is little to no relationship. There are four or five potential *X* outliers with much larger production values than the rest of the observations.



(b) $\hat{y}$ = 112.25678 + 0.11496*x*.



(c) $\hat{y}$ = 112.25678 + 0.11496(125) = 126.6262. (d) Residual = 136 – 126.6262 = 9.3738. (e) R-square = 2.29%. This R-square is also quite small, suggesting sales is not strongly related to production. There doesn't seem to be any strong relationships between any of the variables in this data.

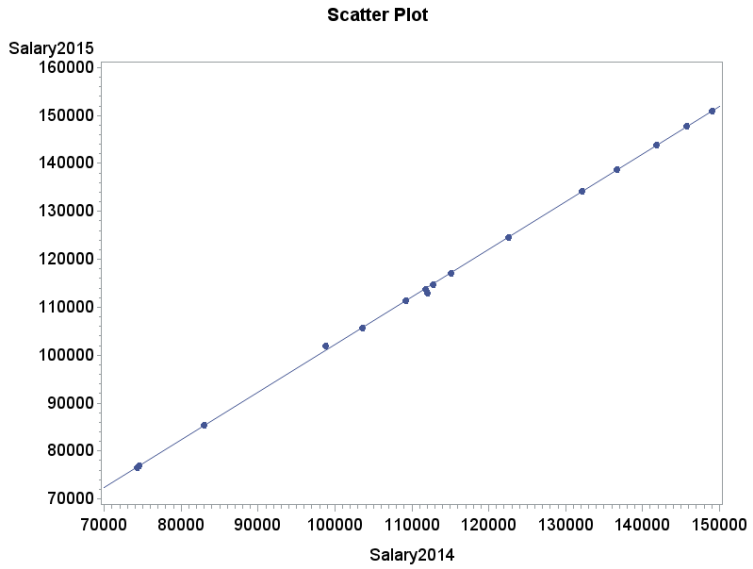
2.124 (a) There is a strong relationship between salary and year for this person. (b) 98.32%. If we judged only by the R-square value, this would indicate a very strong linear relationship, however, judging from the plot, there is an evident curve in the plot, suggesting possibly the need for a transformation.

2.125 (a) The residual plot shows that the data are not linear; a curve would provide a much better fit. (b) By zooming in on the residuals, the residual plot emphasizes the curve that could potentially be overlooked in the scatterplot.

2.126 (a) In both plots we see the strong relationship between year and salary, however, with the log transformed salary, the relationship is much improved and more linear, as shown by the tightness of the data points to the regression line. The curvature that was evident before the transformation is no longer noticeable, suggesting a much better fit. (b)The residual plot of the transformed data looks good, random, as the curvature is no longer present.

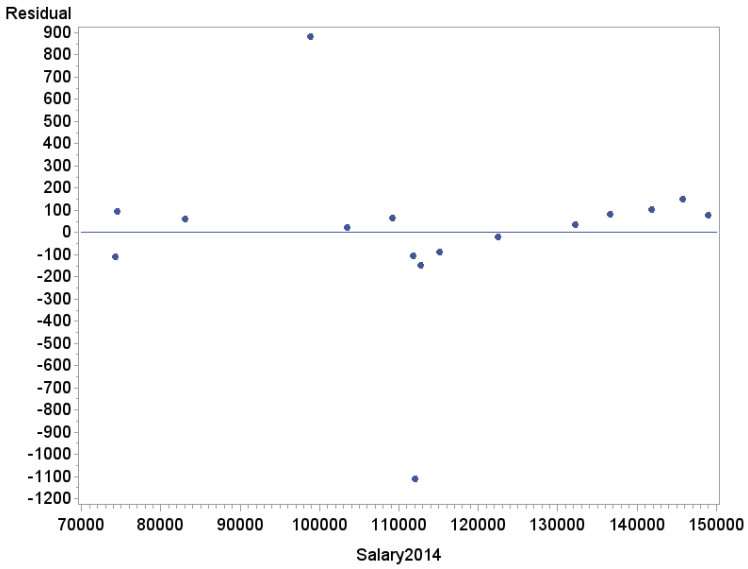
2.127 (a) The regression line is: $\hat{y}$ = 41.25263 + 3.93308*x*. So the prediction is: $\hat{y}$ = 41.25263 + 3.93308(25) = $139,579.63. (b) The regression line is: $\hat{y}$ = 3.86751 + 0.04832*x*. So the prediction is: $\hat{y}$ = 3.86751 + 0.04832(25) = 5.07551, or $160,053.80. (c) The log prediction is better because the data are curved. (d) Even if r^2 is high, this doesn't mean a linear fit is appropriate. If the data follow a curve, a transformation is needed and should give an even higher r^2. (e) Graphs can show you trends that numerical summaries cannot.

2.128 (a)

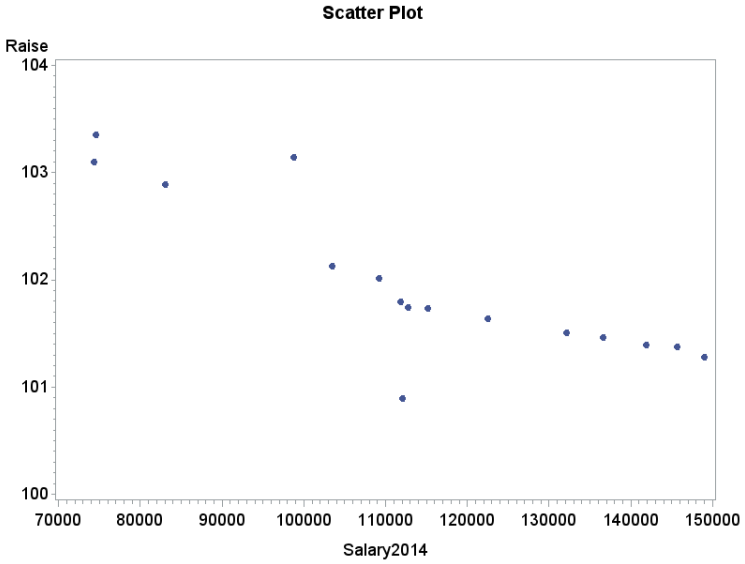


(b) The relationship is linear, positive, and very strong. (c) R-square = 0.9997 or 99.97%.

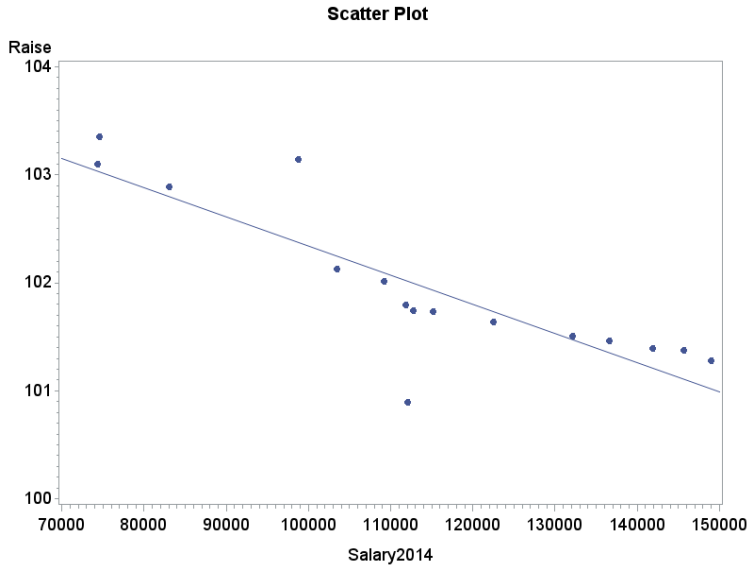
2.129 (a) $\hat{y}$ = 2990.4 + 0.99216*x*. (b) The residual plot shows two *Y* outliers, one high and one low.



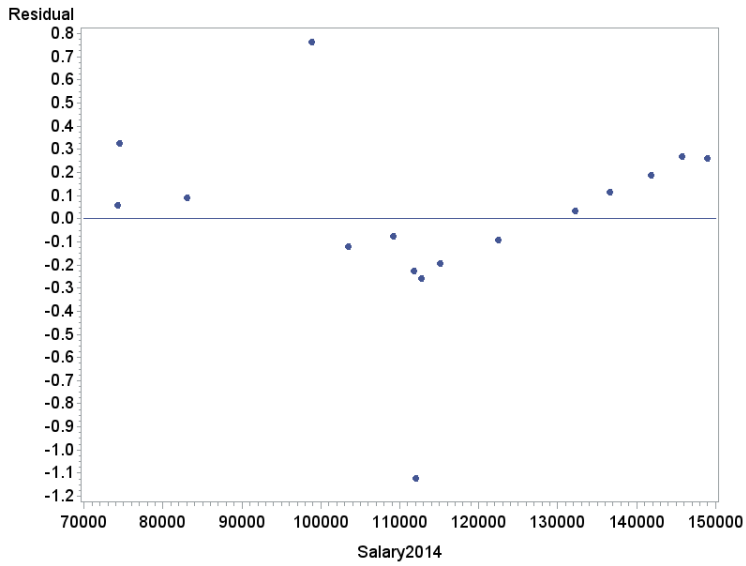
2.130 (a) There is a strong linear relationship between the raise the salary in 2014–2015; as salary increases, the percentage raise decreases.



(b) $\hat{y} = 105.04546 - 0.00002704x$.



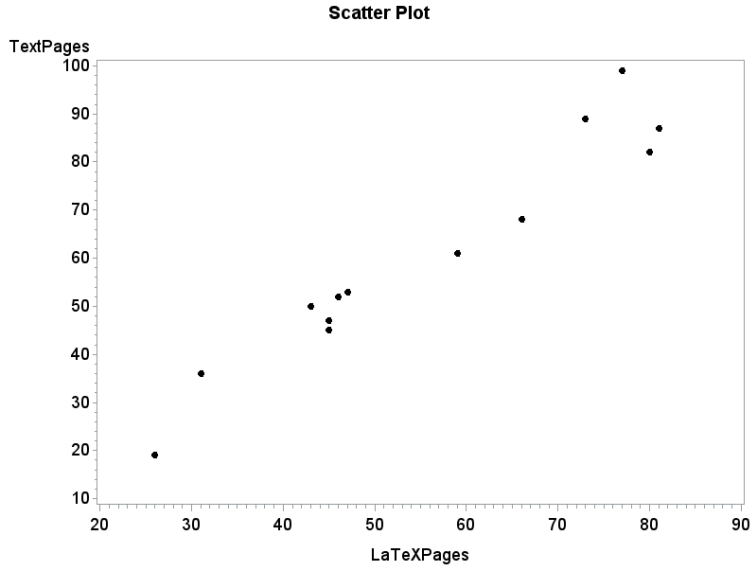
(c) The residual plot shows a definite trend as well as two strong *Y* outliers. One individual got a much larger raise than normal and another got a much smaller raise than normal. Additionally the data points in the residual plot don't follow the center line; this is probably due to the 3 observations on the left which are all above the line. It looks like they are unnaturally influencing the regression by pulling the regression line toward themselves. Without these three, you can see where the regression line "should" be.



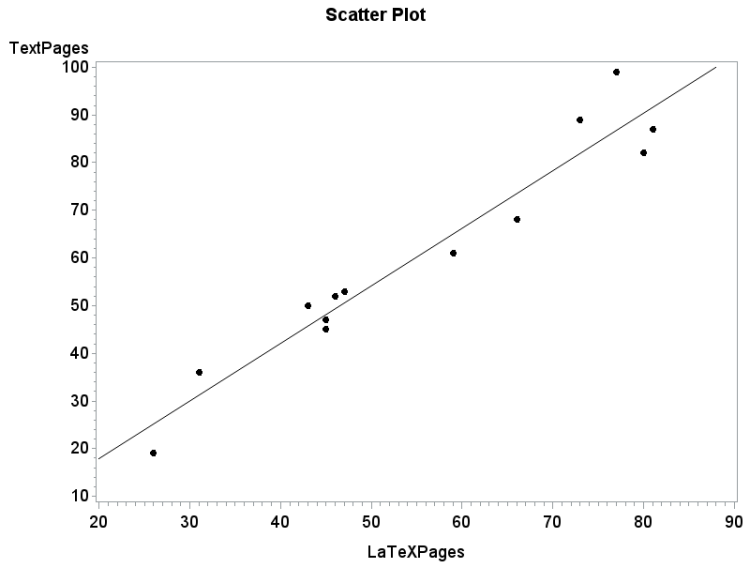
(d) As seen in the residual plot, the 3 observations with low salaries do look like they are influencing the regression line due to their large raises, and pulling or skewing the relationship we are observing. This is definite evidence that those with lower salaries are being given a greater percentage raise compared to the rest of the observations.

2.131 Graduation rates can be different based on the difficulty of programs and/or how good the incoming students are. The residual, or difference between actual graduation rate and predicted graduation rate, is better because it shows if a program is doing better or worse than what is expected given the other variables regarding the incoming students.

2.132 (a) There appears to be a strong positive linear relationship between the number of pages in the final version and the number of pages in the LaTeX files.



(b) $\hat{y} = -6.20176 + 1.2081x$.

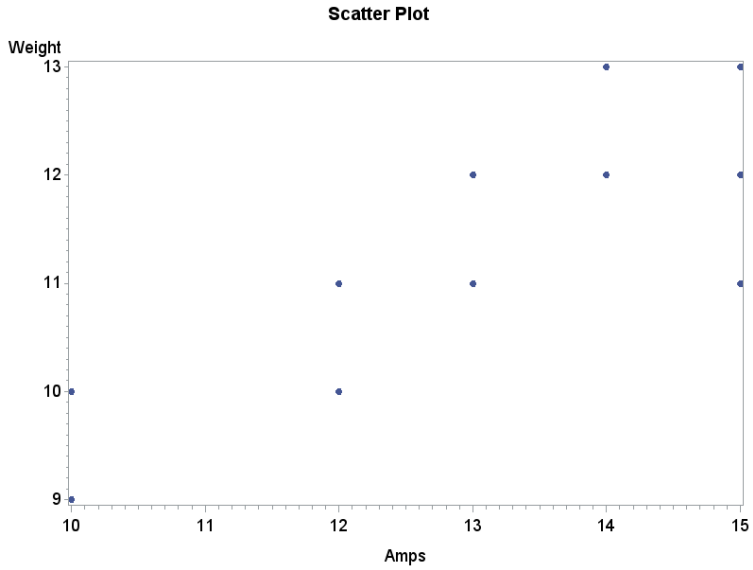


(c) $\hat{y} = -6.20176 + 1.2081(62) = 68.7$. (d) We used the number of pages in the LaTeX files to predict the number of pages in the final version of the text. The relationship was very strong and the regression slope gives us an estimate to predict how many pages we expect in the final version for each page in the LaTeX file. Using this with the regression equation, once we have the total number of LaTeX pages for the new edition, we can easily predict how many pages we believe the final version of the new edition will be.

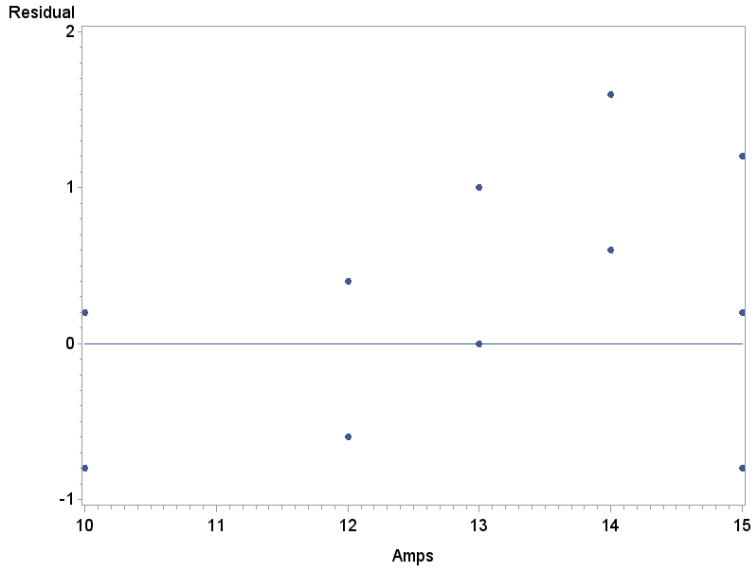
2.133 Answers will vary.

2.134 Answers will vary.

2.135 (a) Higher amps means a bigger motor and more weight. (b) As amps increase, so does weight.



(c) $\hat{y} = 5.8 + 0.4x$. $r^2 = 45.71\%$. (d) For every 1 amp increase, weight increases by 0.4 pounds. (e) 2.5 amps. (f) Yes, there is a slight curvature in the residual plot, suggesting that at the highest amp level, weights may go down somewhat.



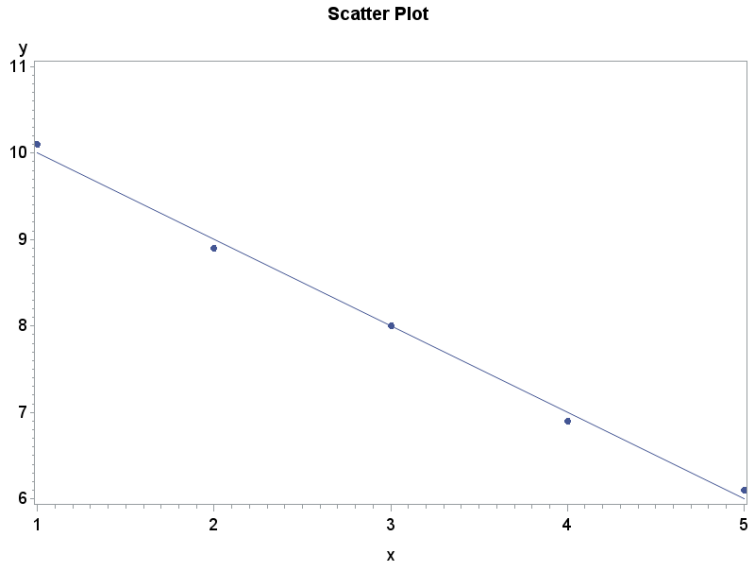
2.136 (a) $r = 0.67612$. (b)

Amps	N Obs	Mean
10	2	9.5
12	4	10.5
13	2	11.5
14	2	12.5
15	9	11.556

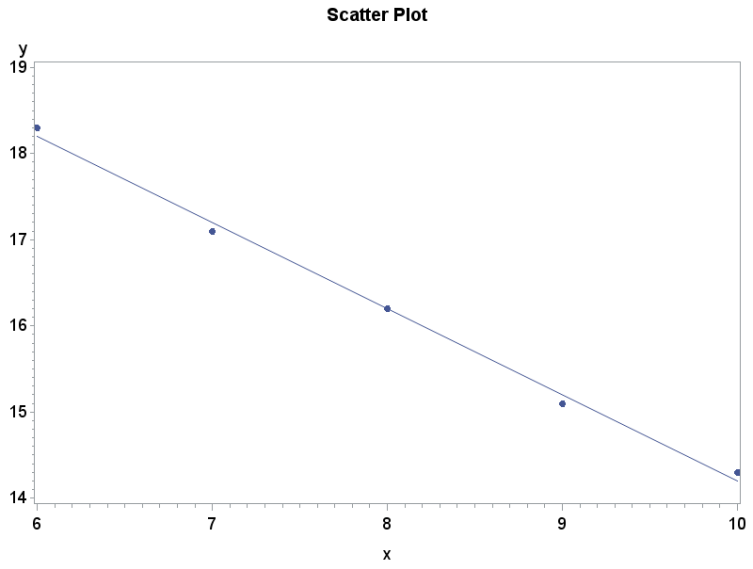
(c) $r = 0.87639$. The correlation for the average weights and amps is greater than the correlation between the individual weights and amps.

2.137 A correlation measures the strength of a linear relationship, or, that is to say, the relationship between Fund A and Fund B is consistent along a straight line. It doesn't mean they have to change by the same amount or a slope of 1. So as long as Fund A moves 20% and Fund B moves 10% consistently, up or down, you will still remain on the same regression line, and they will remain perfectly correlated.

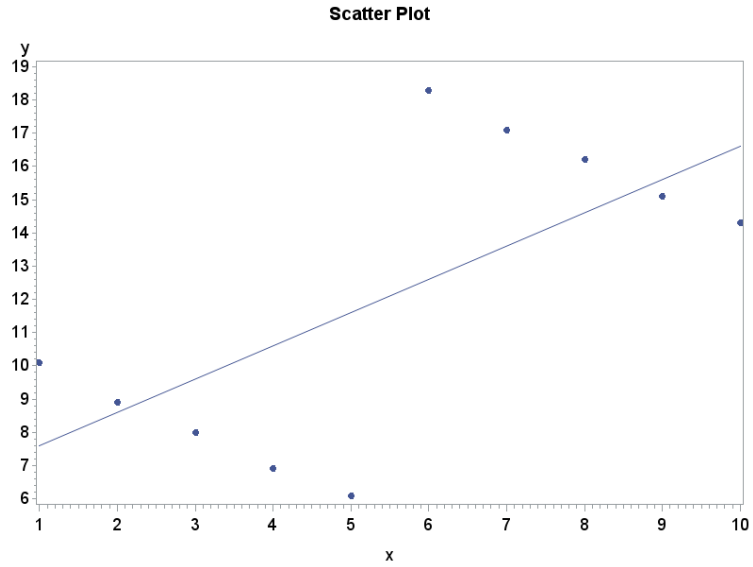
2.138 (a) $\hat{y} = 11 - 1x$. There is a strong negative linear relationship between y and x.



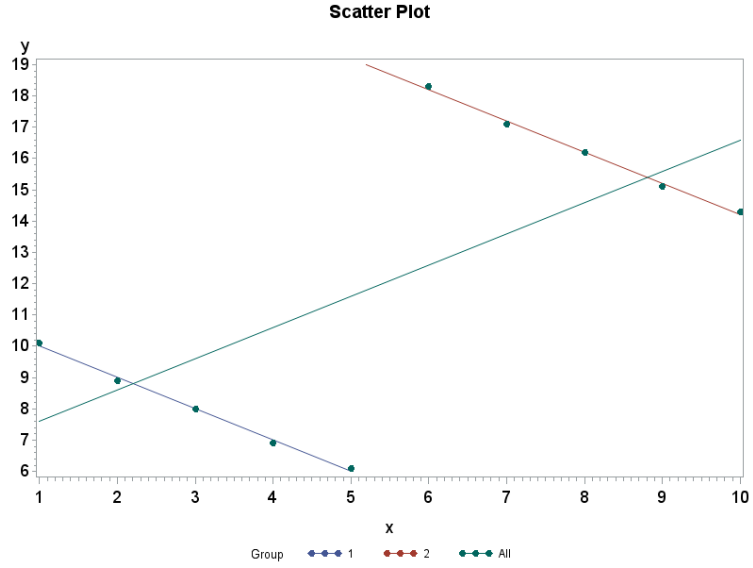
(b) $\hat{y} = 24 - 1x$. There is a strong negative linear relationship between y and x.



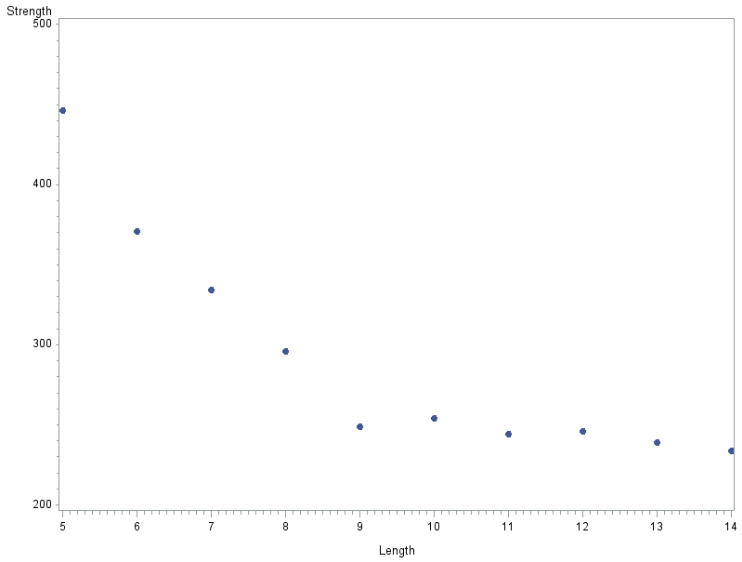
(c) $\hat{y} = 6.6 + 1x$.



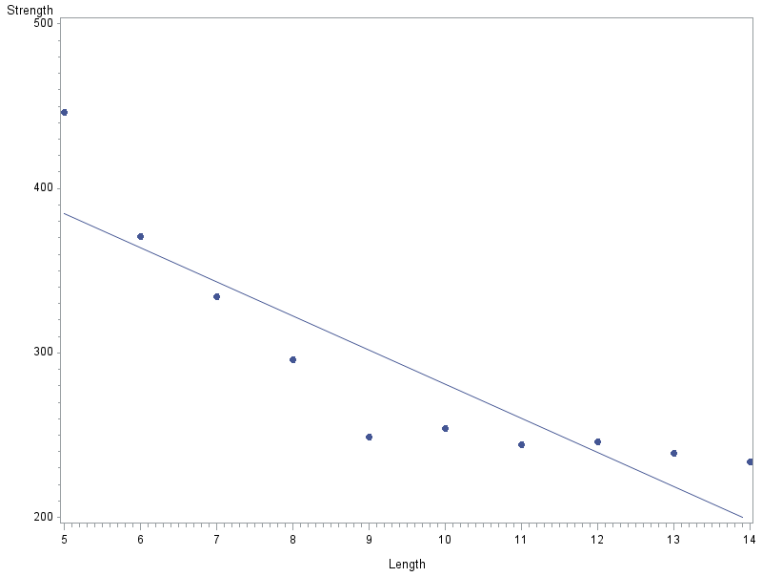
(d) The linear relationship for several groups can reverse direction when the data are combined into a single group; this is likely due to a lurking variable. As shown below, both groups 1 and 2 individually have negative relationships between *y* and *x* but once they were combined the relationship between *y* and *x* was positive.



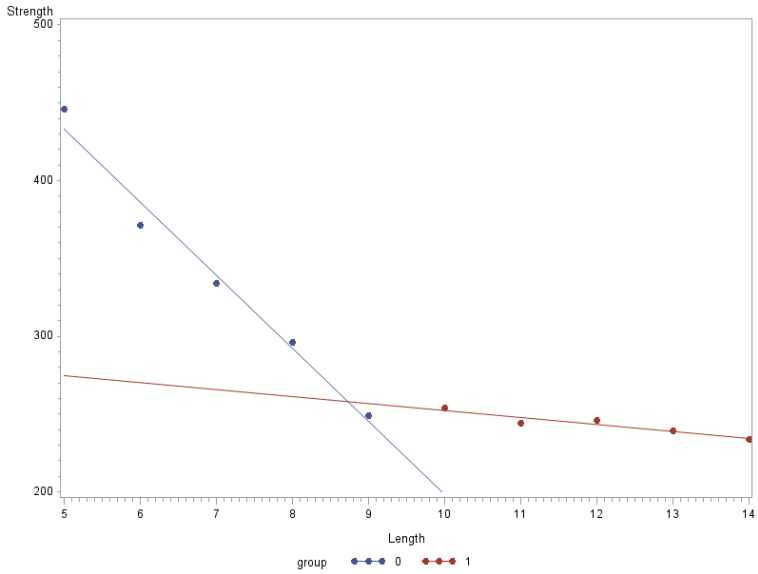
2.139 (a)



(b) The strength decreases with length until 9 inches, then levels off. There are no outliers. (c) The line does not adequately describe the relationship because the relationship changes after length 9 inches.

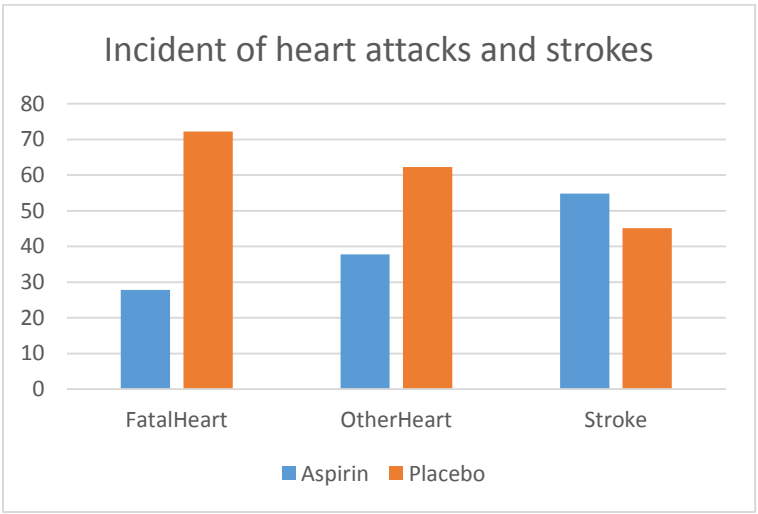


(d) The two lines adequately explain the data. Ask the wood expert what happens at 10 inches.



2.140 As shown in the table and the bar graph below, a lower percentage of aspirin takers had either fatal heart attacks or other heart attacks than then those taking a placebo. However, a greater percentage of aspirin takers had strokes than those taking the placebo. It seems as if taking aspirin may help reduce the incident of heart attacks but may contribute to the incident of strokes. Answers will vary on whether or not the study provides evidence that aspirin actually reduces heart attacks. It is possible there are lurking variables that we have not accounted for.

	Group	
Outcome	**Aspirin**	**Placebo**
FatalHeart	27.78	72.22
OtherHeart	37.72	62.28
Stroke	54.84	45.16



2.141 (a) Smokers 76.12%, Nonsmokers 68.58%.

	Smoker		
Outcome	**Yes**	**No**	**Total**
Dead	139	230	369
Alive	443	502	945
Total	582	732	1314

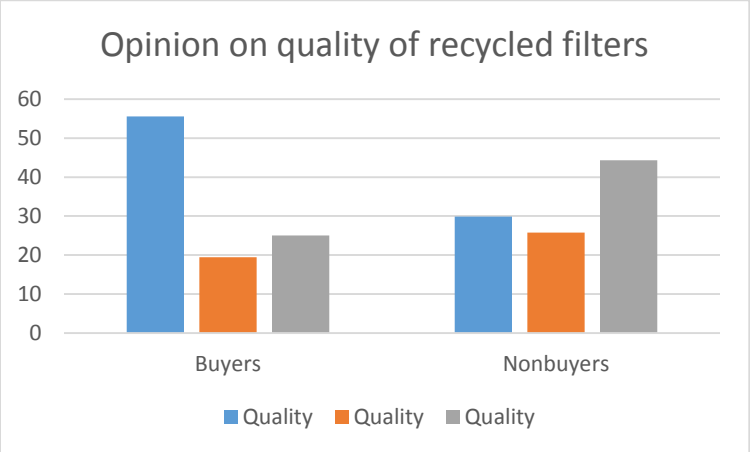
(b) Age 18 to 44 alive: Smokers 93.4%, Nonsmoker 96.18%. Age 45 to 64 alive: Smokers 68.16%, Nonsmokers 73.87%. Age 65 and Over alive: Smokers 14.29%, Nonsmokers 14.51%. (c) The percentage of smokers are 45.86% (18 to 44), 55.18% (45 to 64), 20.25% (65 and Over). This confirms the authors' explanation.

2.142 (a) The distribution shown below shows a split between what people feel about the quality of the recycled filters.

Quality	**Frequency**	**Percent**
Higher	49	36.84
Same	32	24.06
Lower	52	39.1

(b) The conditional distributions are shown in the table below. Buyers of the filters are much more likely to think the quality of the recycled filter is higher, while in contrast, Nonbuyers are much more likely to think the quality of the recycled filter is lower. It is plausible that using the filters may cause more favorable opinions.

	Quality			
Group	**Higher**	**Same**	**Lower**	**Total**
Buyers	55.56	19.44	25	100
Nonbuyers	29.9	25.77	44.33	100



Some specific teaching recommendations:

1. Introduce density functions through a Uniform distribution. Without evidence to the contrary, students may begin to think that all distributions are Normal. Some examples include the (pseudo-) random number generator on a computer, the distribution of numbers drawn in a lottery, and so on. These distributions also have added simplicity because area = (length) × (width).
2. Use distributions students can relate to. Heights of adult women, SAT scores, and so on are approximately Normal. The daily returns on stocks tend to be rather Normal (but with "fatter" tails than a true Normal distribution). Some examples of distributions that are *not* Normal might be outage length for your local power company (most are short, but in the event of a major storm, outages can last for days), major league batting averages, grades on a (easy) test, and so on.
3. Use the 68-95-99.7 rule as a guide. Building intuition with this rule (and sketching the distributions under consideration) can help students determine if their answer on a calculation is "reasonable."
4. Any value on any distribution can be "standardized" with a z-score, but not all distributions are Normal.

Some LaunchPad resources:
StatClips (and Associated Whiteboards):
The Normal Distribution
Snapshots:
Normal Distributions
StatBoards:
Density Curves
Finding a Value Given a Proportion
Stepped Out Tutorials:
Density Curves
Applet:
Normal Curve

CHAPTER 2: EXAMINING RELATIONSHIPS

Having dealt with methods for describing a single variable, we turn to relationships among several variables. At the level of *PSBE*, that means mostly relationships between two variables. That a relationship between two variables can be strongly affected by other ("lurking") variables is, however, one of the chapter's themes. Note the new vocabulary (explanatory and response variables) in the chapter introduction, as well as the reiteration of basic strategies for data analysis. Statistical model building is certainly one of the primary research tools of the social sciences, physical sciences, and engineering; building those models requires an understanding of the structure not only of each variable under consideration, but the relationships among them.

Correlation and regression are traditionally messy subjects based on opaque "computing formulas" that are in turn based on sums of squares (these formulas do not appear in the text, as they add nothing to conceptual understanding). *PSBE* asks that students have a "two-variable statistics" calculator that will give them the correlation and the slope and intercept of the least-squares regression line from keyed-in data. This liberates the instructor—we can give reasonably realistic problems and concentrate on intelligent use rather than awful arithmetic. Do remember that data input and editing can be frustrating on a calculator, so reserve large problems for computer software.

The descriptive methods in this chapter, like those in Chapter 1, correspond to formal inference procedures presented later in the text. Many texts delay the descriptive treatment of correlation and regression until inference in these settings can also be presented. There are, we think, good reasons *not* to do this. By carefully describing data first, we emphasize the separate status and greater generality of data analysis. There are many data sets for which inference procedures do not apply—data for the 50 states, for example. Fitting a least-squares line is a general procedure, while using such a line to give a 95% prediction interval requires additional assumptions that are not always valid. In addition, students become accustomed to examining data *before* proceeding to formal inference, an important principle of good statistical practice. Finally, correlation and regression are so important that they should certainly appear in a first course, even if you do not have the time to discuss formal inference in these settings.

2.1 Scatterplots

Using graphs should be comfortable by now. Constructing scatterplots is a relatively easy task (but tedious without software for all but small data sets.) Interpreting the plots takes some practice. In the classroom, build instruction on examples and stress that common sense and some understanding of the data are necessary to do a good job of description. Computers can make the plots, but people are needed to describe them. Again, the general rule is to look for overall patterns and deviations from them. Patterns such as clusters and positive and negative association are useful in many cases but can lead to distorted descriptions when imposed in situations where they do not apply.

Some specific teaching recommendations:

1. Once again, emphasize context in labeling graphs. "Y" and "X" as axis labels may be "mathematically" correct, but do nothing to help a viewer make connections and conclusions about the subject at hand.
2. Just as we discuss shape, center, variability and outliers when describing the distribution of a single numeric variable, in this setting we have form, direction, strength, and outliers. All too often, students want to describe a scatterplot using "shape" terms such as Normal. Distribution shapes cannot be seen in a scatterplot.
3. Sometimes which variable is explanatory and which is a response may be unclear (or unnecessary, if you simply want to know if a relationship exists). In such cases, either variable can be used as the "X" or "Y" in the scatterplot.
4. Transforming one (or both) variables can straighten a plot. In the next two sections, we discuss correlation and linear regression. Both assume a "straight line" relationship between the two variables.
5. Adding a categorical (lurking) variable can help explain the relationship further. See the last two exercises in this section.

Some LaunchPad Resources:

Applets:
    Two-Variable Statistical Calculator
StatBoards:
    Creating and Interpreting Scatterplots
Stepped Out Tutorial:
    Scatterplots

2.2 Correlation.

Correlation is presented before regression in part because it does not require the explanatory-response

distinction. This also allows us to give a meaningful formula for the regression slope using the correlation. Students should have a calculator that gives *r* from keyed-in data. You can therefore use the somewhat messy formula for *r* as a basis for explaining how correlation behaves (and ties back to *z*-scores as seen in Chapter 1), but avoid using it for computation. The standardized versions of the variables translate the relationship with a center at the origin. With this in mind, thinking about how a linear relationship uses the quadrants helps to motivate correlation geometrically.

Some specific teaching recommendations:

1. Emphasize that correlation makes sense only for numeric variables. The term is often misused to indicate any type of association. You cannot do the necessary arithmetic to compute a correlation with categorical variables.
2. Emphasize the need to plot data before calculating a correlation. Correlation measures strength only for linear relationships. A correlation near 0 may mean there is no relationship between the variables, or it could mean that there is a very strong curved relationship. See Exercise 2.25.
3. Because correlation is based on z-scores, it has no units (but slopes do!).
4. Correlation is not resistant to outliers. A simple example like the following proves the point sufficiently.

X	1	2	3	4	5	6
Y	2.5	3.0	3.8	3.5	3.3	3.8

These values have $r = 0.754$. Add a new data point at (12, 10). The correlation increases to 0.932. If the new data point is at (12, 3), the correlation decreases to 0.120.

Some LaunchPad Resources:

Snapshots Videos:
Correlation and Causation
StatBoards:
Computing a Correlation
Stepped Out Tutorials:
Correlation
Applets:
Two-Variable Statistical Calculator
Correlation and Regression

2.3 Least-Squares Regression

The background to regression isn't always clear to students, so don't skip over it: We'd like to draw the *best* line through the points on our scatterplot; to do this, we need an explicit statement of what we mean by "best." Most students will intuitively agree that the line should go through the "middle" of the points. The least squares idea agrees with this, and has other desirable properties. Least squares isn't terribly natural. At this point, just say that it's the most common way to fit a line. The concepts of "outlier" and "influential observation" are important; if you use the example given above or something similar in discussing correlation, this notion should be clear. An observation is influential if removing it would move the regression line. This is clearly a matter of degree. More advanced statistical methods include numerical measures of influence. I've defined "outlier" broadly to keep things simple for students—they only have to look for isolated extreme points in any direction. That's a matter of degree also. Outliers in *y* have large residuals; outliers in *x* are usually influential – either to correlation, regression, or both.

Some specific teaching recommendations:

1. Slopes have a sign, a number, and a unit in context. A slope tells us that "price increases about $80 per square foot," in the relationship between asking price of a house and its size. Don't let students fall into the trap of saying purely that "as size increases so does price." This type of interpretation totally omits any discussion about how much the price increases with size.
2. Residuals are not intuitive to students. Emphasize that these are the "leftover" errors (unexplained variability) in the model, and have a direct relationship to r^2, which measures the amount of variation in the response that is explained by the model. Try to avoid ideas more related to inference (such as the residuals having a $N(0, \sigma)$ distribution) and concentrate on them having a "random" pattern with constant variation around the $e = 0$ line.
3. The sign of a residual indicates how it relates to the regression line. Questions like "If we regress salary on years of experience, would you rather have a positive or negative residual?" usually help make that distinction. If the residual is 0, you are paid exactly according to the model. If your residual is positive, you are being paid "better than average" for your experience.

Some LaunchPad resources:

StatClips Videos (and related examples):
Regression: Introduction and Motivation
Snapshots:
Introduction to Regression
StatBoards:
Fitting the Least-Squares Regression Line
Calculating and Plotting Residuals
Stepped Out Tutorials:
Regression 2
Regression Residuals 2
Applets:
Two-Variable Statistical Calculator
Correlation and Regression

2.4 Cautions about Correlation and Regression

As calculations have become automated, interpretive ideas become a more important part of basic instruction. "Correlation Is Not Causation" is one big idea here. Others include the dangers of extrapolation (weather forecasters must do it, for example); your company's "forecast" profits might become major losses if a competitor brings out a slightly better (or cheaper) product than what you have. Lurking variables also are revisited. The moral of this section is that one must think carefully about making conclusions from a regression. There is no substitute for plotting the data.

Some LaunchPad resources:

StatBoards:
Beware Extrapolation!

2.5 Relations in Categorical Data

Some would argue that this section can be skipped without compromising the continuity of course material and that skills learned in this chapter are not essential to success in later chapters (indeed, it was starred as "optional" in prior versions of this text). However, there are at least four arguments

that spending time on this material is a good investment. First, this material deals with categorical data, which is typically central to the work of many social scientists, as well as "product satisfaction" surveys, just to name one business application. Second, there is a direct connection between this material and that in Section 8.2 ("Comparing Two Proportions"). Third, understanding that conditional and marginal distributions each convey different (and sometimes conflicting) information about relationships between the two variables is part of what constitutes basic statistical literacy—we want students to think carefully about numbers they encounter in the media; consideration of these distributions can ease the discussion of conditional probability to come in Section 4.3 (if you plan to cover that material). Finally, some of the ideas described in this chapter parallel the coming discussions of experimental design and sampling in Chapter 3. For example, the notion of a "lurking variable" arises (again) in discussions of Simpson's paradox. You can also foreshadow the idea of independence to be seen in Chapters 4 and 9. Still, if you prefer to wait until later in the course (perhaps pairing it with Chapter 9, "Inference for Categorical Data"), that is fine.

The computations required in this section are minimal. Percents and proportions are the numerical summaries. On the other hand, some very important ideas are presented. Judgment is required to select what percents to calculate. We have found it helpful to ask students to focus on "what we know" first. For example, the following three questions are all different: (1) What proportion of employees are male managers? (2) What proportion of employees who are male are managers? (3) What proportion of employees who are managers are male? Question #1 is not a conditional probability—here we are being asked about those employees who are male and managers. Questions #2 and #3 are both conditional—in Question #2 we first know that the employee is male, while in Question #3 we first know that the employee is a manager. Wording of these questions can be subtle, so it is important to have your students think about what the question is asking. You might point out that "key words" such as "if," "when," and "given" signal a conditional question. Mention to your students that these key words are not always in the question and suggest that they think about what we know first to determine what the conditioning situation is.

Some specific teaching recommendations:

1. Motivate ideas with an interesting example. Consider using one or two interesting examples to describe all the material in this chapter. A well-known and interesting example is provided by considering admission rates for men and women applicants to six academic departments at Berkeley for Fall 1973 (P.J. Bickel, E.A. Hammel, and J.W. O'Connell (1975). "Sex Bias in Graduate Admissions: Data from Berkeley." *Science* 187(4175): 398–404). Although this example is quite old, it comes up in many discussions of Simpson's paradox (and can be used to review all of the material in this section). You can use these data to illustrate bar graphs (again), compute the marginal distribution of sex, and the marginal distribution of admission status. You can then examine the conditional distributions of admission, given sex. Finally, you can use these data to illustrate Simpson's paradox:

   When we examine the table "collapsed" on department, men are admitted at a higher rate than women:

	Number of Applicants	**% Admitted**
Men	8,442	**44%**
Women	4,321	35%

   However, when we look at each department individually, this isn't the case. In fact, in most departments, women have a higher admission rate than men.

The "lurking variable," or explanation, is that men and women chose to apply to different kinds of departments— most men applied to departments that had higher admission rates (for both sexes), while most women applied to departments with lower admission rates (for both sexes).

	Men		Women	
Department	**Applicants**	**% Admitted**	**Applicants**	**% Admitted**
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	272	6%	341	**7%**

Some LaunchPad resources:

StatBoards:
- Marginal Distributions
- Conditional Distributions
- Graphing a Two-Way Table

Stepped Out Tutorials:
- Two-Way Tables
- Simpson's Paradox in Two-Way Tables

EESEE Case Studies:
- Surviving the Titanic
- On Time Flights
- Does Smoking Improve Survival?
- Cancer and Power Lines

CHAPTER 3: PRODUCING DATA

Note that, if you prefer, the material in this chapter may be covered before Chapter 1 with no loss of understanding.

This is a relatively short chapter with a lot of ideas and little numerical work. Students find the essentials quite easy, but they are very important. This chapter isn't mathematics, but it is core content for statistics. Weaknesses in data production account for most erroneous conclusions in statistical studies. The message is that production of good data requires careful planning. Random digits (Table B) are used to select simple random samples and to assign units to treatments in an experiment. There are numerous examples that can serve as the basis for classroom discussion.

The chapter also has a secondary purpose: The use of chance in random sampling and randomized comparative experiments motivates the study of chance behavior in Chapter 4.

3.1 Sources of Data

This short section gives an overview of different data "collection" methods, including using available information collected by other sources. One item not really discussed here is the limitation of using available data. You must think carefully about those limitations (go back to the "5 Ws") before deciding to use such. Even "reliable" sources such as the FBI crime statistics database have been