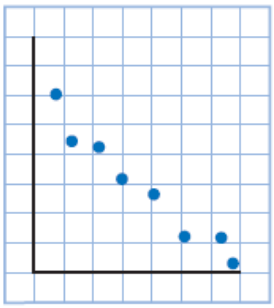
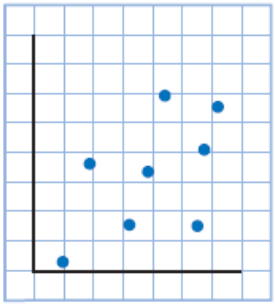
Solutions

1. Match each scatter plot with its correct correlation coefficient.

Correlation coefficients:
$-0.97, -0.56, 0.56, 0.97$



This graph shows a strong negative linear correlation, so the most likely correlation coefficient is -0.97



This graph shows a moderate positive linear correlation, so the most likely correlation coefficient is 0.56

2. The table shows a trucker's distance from home over time.

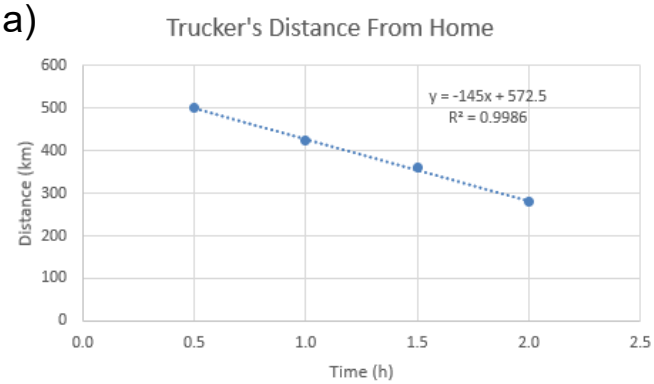
a) Use graphing technology to create a scatter plot of the data.

b) Determine the strength of linear correlation between these variables.

c) Perform a linear regression. Interpret the meaning of the equation of the line of best fit.

Time, t (h)	Distance, d (km)
0.5	500
1.0	425
1.5	360
2.0	280

a)



b) From inspection there looks to be a very strong, negative linear correlation. An r^2 value of 0.9986 gives an r-value of 0.9993 which backs this up.

c) The linear regression gives an equation of y = -145x + 572.5 where y is distance in km, and x is time in hours. The equation implies that the trucker started at a distance of 572.5 km from home (when x = 0) and that for every hour that passes the trucker is 145 km closer to home than before.

3. Children's self-esteem is positively correlated with their level of achievement.

a) Suggest a cause and effect relationship that could account for the results.

b) What reverse cause and effect relationship could also account for the results?

a) As self-esteem increases, so does their level of achievement.

b) As their level of achievement increases, so does their self-esteem.

4. Characterize each of the relationships. The independent variable is listed first.

a) Computer sales are negatively correlated with the unemployment rate.

b) The price of gas is positively correlated with the performance of a football team.

c) Running speed is positively correlated with heart rate.

a) As a drop in computer sales is more likely to be linked to an increase in unemployment this relationship can be characterized as reverse cause and effect.

b) There isn't any obvious connection between gas prices and the performance of a football team so the relationship can be characterized as accidental.

c) As you run quicker your heart rate increases, so it is reasonable to characterize this relationship as cause and effect.

5. a) Create a scatter plot of population versus time. Call 2007 year 0. Describe the correlation.

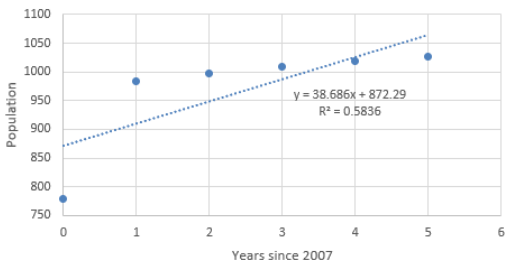
b) Perform a linear regression. Interpret the equation of the line of best fit.

c) Create a residual plot. Does this appear to be a good linear model? Explain.

The table shows student population data for a new high school. The school wants to project the school's population growth over time.

Year	Population
2007	778
2008	984
2009	998
2010	1010
2011	1018
2012	1026

a)

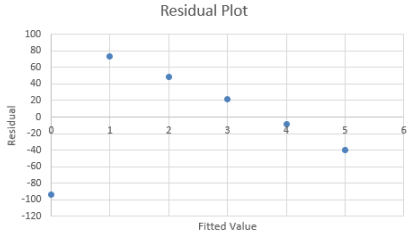


The correlation is a strong, positive linear one, except for the point (0,778).

b) The linear regression gives an equation of y = 38.686x + 872.29. The r^2 value is 0.5836 gives an r-value of 0.764 which is a moderate/strong positive linear correlation. The equation implies a starting population of 872 in 2007 and an increase of 38.686 students per year.

c)

Year	Population	Predicted	Residual
0	778	872.29	-94.29
1	984	910.976	73.024
2	998	949.662	48.338
3	1010	988.348	21.652
4	1018	1027.034	-9.034
5	1026	1065.72	-39.72



The residual for (0,778) is the furthest from the line of best fit and causes the pattern of the other residuals. The model is not a good fit.

6. Grade 12 was not offered until 2008.

a) How does this information affect the correlational study? Does it make sense to remove the 2007 datum? Explain.

b) Repeat the analysis with the outlier removed. Compare the two models.

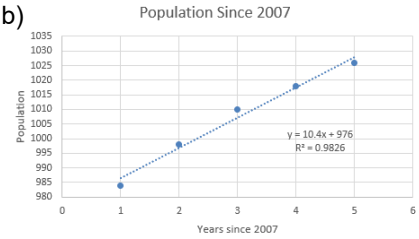
c) Use both models to predict the school's population in 2016. Which model should the principal rely on and why?

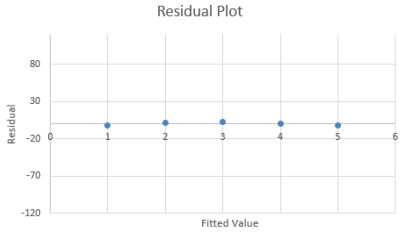
The table shows student population data for a new high school. The school wants to project the school's population growth over time.

Year	Population
2007	778
2008	984
2009	998
2010	1010
2011	1018
2012	1026

a) An outlier can have a significant impact on a linear regression if we only have a small number of data points. Because G12 was not offered until 2008 the data for 2007 only has 3 grades worth of students instead of 4. It is acceptable to ignore this datum.

b)





The linear regression gives an equation of $y = 10.4x + 976$. The r^2 value is 0.9826 gives an r-value of 0.991 which is a very strong positive linear correlation. The equation implies a starting population of 976 in 2007 and an increase of 10.4 students per year. The residuals are all close to the residual line which adds further weight to this being a very good fit.

c) The original model gives a population of 1220, the new model gives a population of 1070. As the new model is a much better fit, I would advise the principal to plan for a population of 1070.

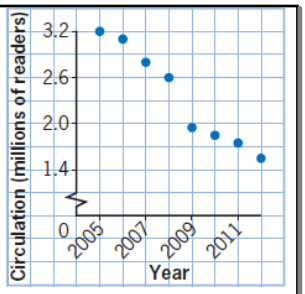
7. a) Describe the trend in sales.

b) Is there evidence of a hidden variable?

c) When do you think the newspaper raised its price from \$1 to \$1.50? Explain.

d) Explain how the price change represents a hidden variable in this correlation.

The graph shows a newspaper's annual average circulation data.



a) The trend is that is a decreasing circulation over time.

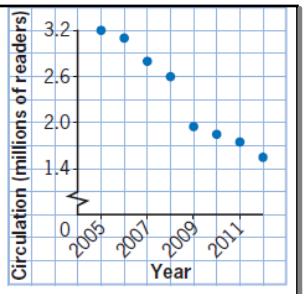
b) There are two different downward trends (2005 to 2008, and 2009 to 2012), so there could be a hidden variable that is fragmented pattern.

c) It is likely that the newspaper raised its price in 2009 which caused the sudden drop in circulation.

d) The price change represents a hidden variable because of the fragmentation in the downward trend.

8. How does the vertical scale in the newspaper circulation graph distort the linear trend?

The graph shows a newspaper's annual average circulation data.



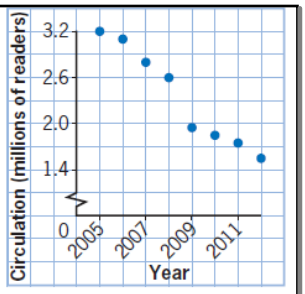
The break in the vertical scale exaggerates the downward trend. If the full scale was there it would not look quite as dramatic.

9. Suppose this graph were published with the headline "Newspaper circulation in free fall."

a) Explain how this title is biased.

b) Write an unbiased title for this graph.

The graph shows a newspaper's annual average circulation data.



a) This headline is biased because the language used is not neutral. The use of "free fall" is sensationalizing the situation.

b) Newspaper circulation in decline.

The table shows a group of students' grade 12 calculus marks and first year university marks. Half took a summer prep course.

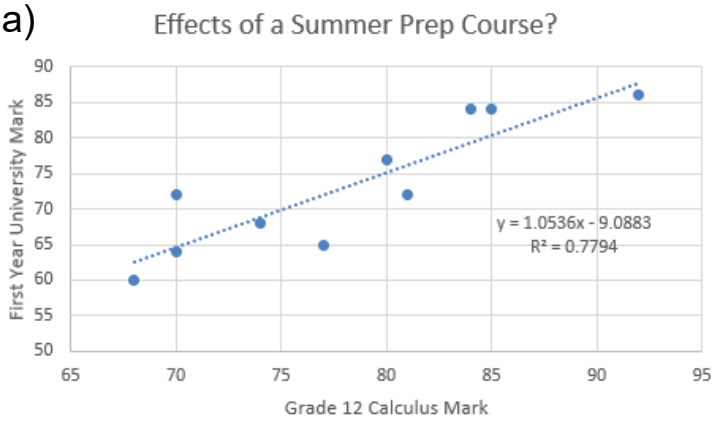
Yes Prep.		No Prep.	
Grade 12	First Year	Grade 12	First Year
80	77	77	65
70	72	74	68
92	86	68	60
84	84	70	64
85	84	81	72

10. a) Create a scatter plot that compares first year marks to grade 12 marks.

b) Perform a linear regression. Interpret the correlation coefficient.

c) Was the summer prep course helpful?

a)



b) The r^2 value is 0.7794 giving an r-value of 0.883. This would imply a strong, positive linear correlation.

c) It is unclear that the summer prep course helped. The hidden variable is likely obscured by the linear correlation.

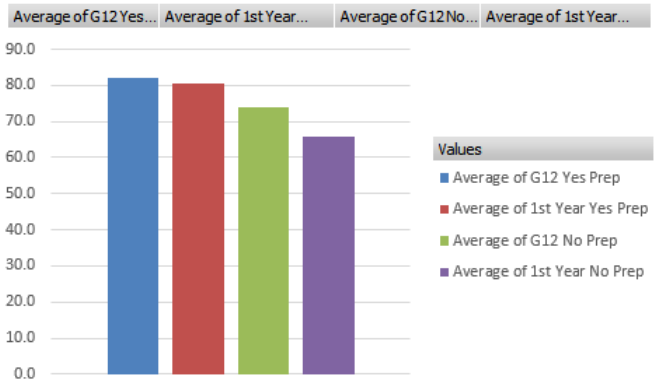
The table shows a group of students' grade 12 calculus marks and first year university marks. Half took a summer prep course.

11. a) Create a contingency table and side-by-side box plots or a pivot table and pivot chart to compare the two groups.

b) Use summary statistics to determine if the summer prep course is helpful.

G12 Yes Prep	1st Year Yes Prep	G12 No Prep	1st Year No Prep
80	77	77	65
70	72	74	68
92	86	68	60
84	84	70	64
85	84	81	72

Average of G12 Yes Prep	Average of 1st Year Yes Prep	Average of G12 No Prep	Average of 1st Year No Prep
82.2	80.6	74.0	65.8



b) The pivot table shows that for those students who took the summer prep course they had a higher mean G12 calculus mark (82.2 compared to 74). They also had a higher mean first year university mark (80.6 compared to 65.8). It would appear that the summer prep course is helpful.

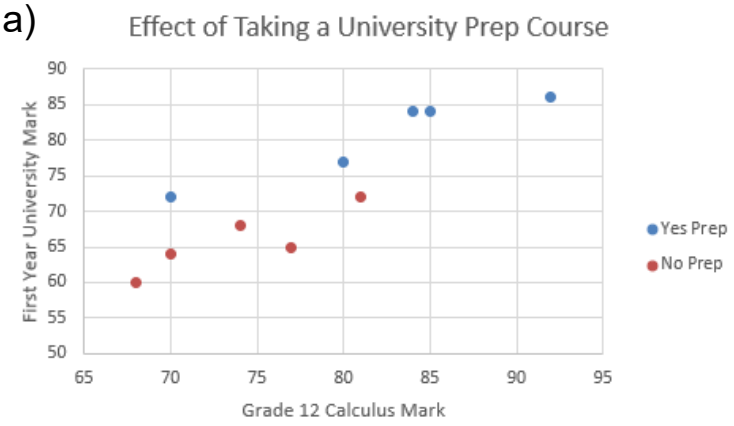
The table shows a group of students' grade 12 calculus marks and first year university marks. Half took a summer prep course.

Yes Prep.		No Prep.	
Grade 12	First Year	Grade 12	First Year
80	77	77	65
70	72	74	68
92	86	68	60
84	84	70	64
85	84	81	72

12. a) Use a bubble plot or a legend attribute to compare the two groups.

b) Does the graph show that the prep course is helpful? Explain.

a)



b) By looking at the graph we can see that the students that took the prep course (blue) did better than those that did not (red). This confirms that taking the prep course is useful.