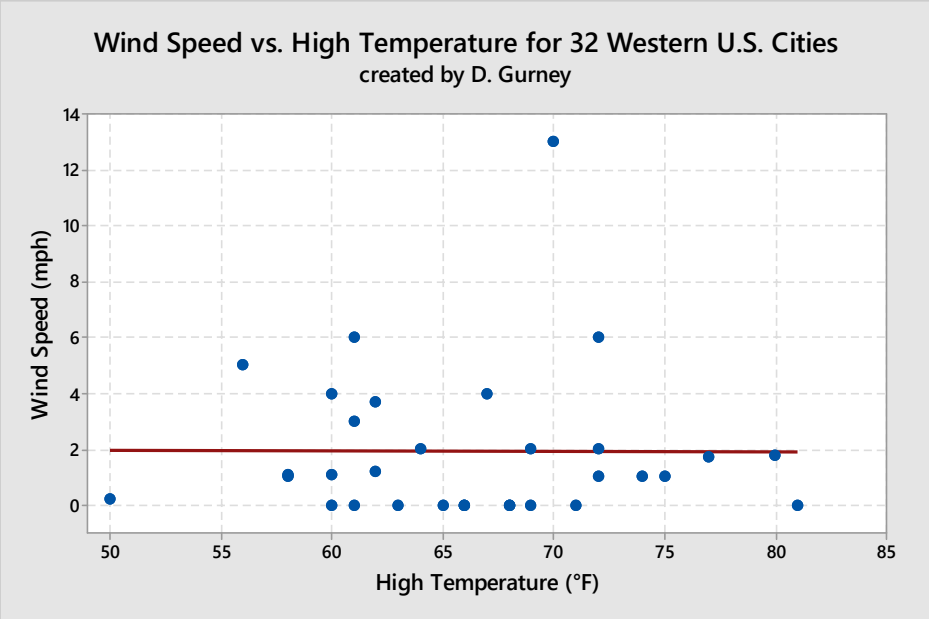
Scatter Diagrams Analysis

We will always be constructing scatter diagrams with the regression line included. When analyzing a scatter diagram, you should first say whether there is a positive or negative association and whether the association is strong, moderate or weak.

Write out the equation of the regression line and give the value of R-square, the coefficient of determination. R-square tells us what percentage of variation in the y-variable (the variable on the vertical axis) is determined by the regression line. You should state this in terms of the y-variable on the scatter diagram.

Finally, note the location of outlier points and the location of influential observations. Unless all the points of the scatter diagram are on the regression line, there will be at least one outlier. To have an influential observation, there must be a large horizontal gap in the scatter diagram between a few points and most (more than 80%) of the other points.

Example 1

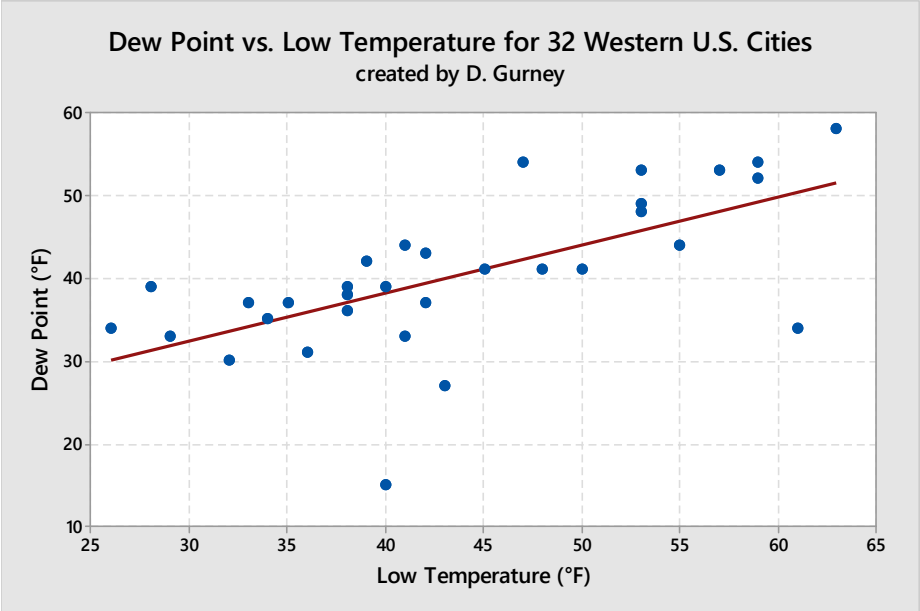


The scatter diagram of wind speed versus high temperature shows very weak association. The equation of the regression line is

$$y = 2.095 - 0.00247x.$$

The R-square value is less than 0.001, which means that less than 0.1% of the variation in the wind speed is explained by the regression line. There is one outlier at about (70, 13) and an influential observation at about (50, 0).

Example 2

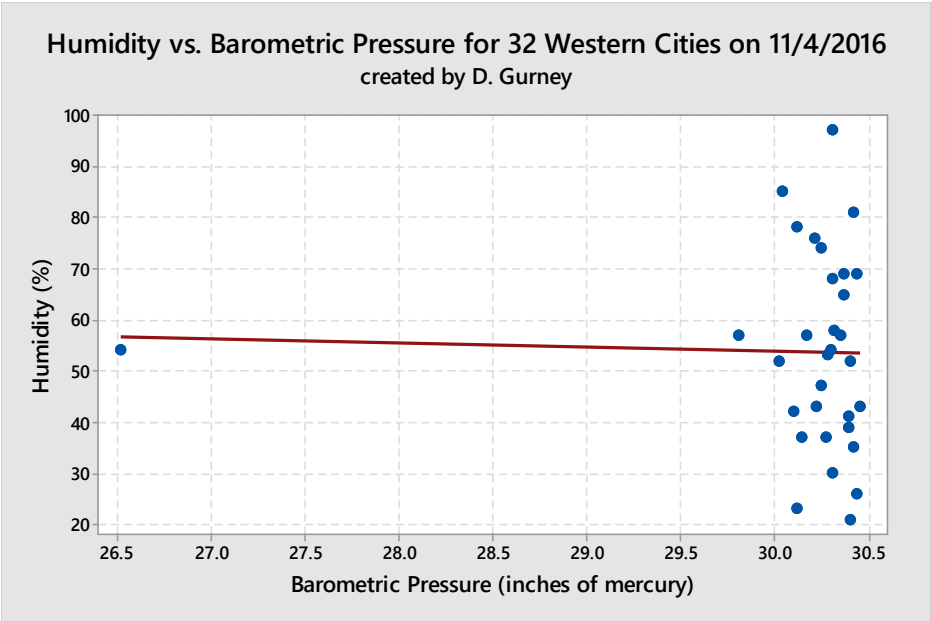


The scatter diagram of dew point versus low temperature shows a moderate positive association. The equation of the regression line is

$$y = 14.96 + 0.5810x.$$

The R-square value is 0.414, which means about 41.4% of the variation in the dew point is explained by the regression line. There is one outlier at about (40, 16). There are no influential observations.

Example 3

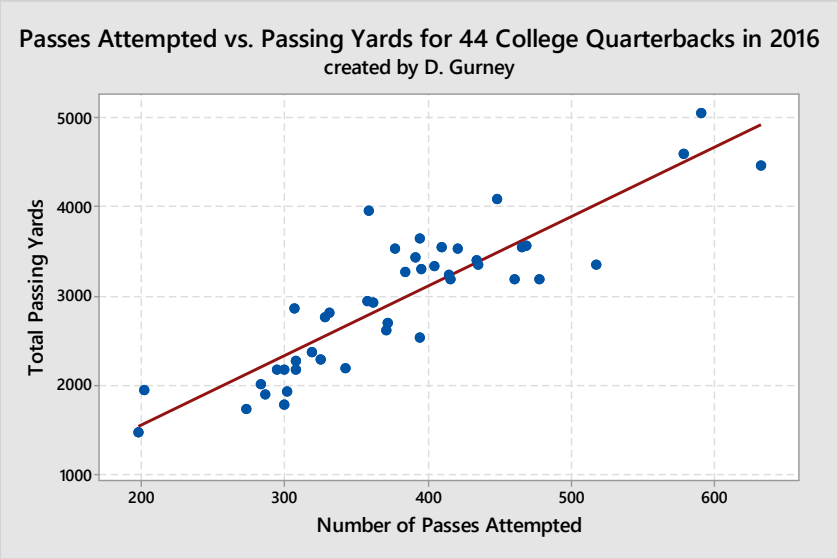


The scatter diagram of humidity versus barometric pressure shows very little association. The equation of the regression line is

$$y = 78.1 - 0.806x$$

Notice that since the barometric pressure does not change much, the regression line drops very little over the values shown even though the slope of the regression line is not that small in absolute value. There is an influential observation at about (26.5, 54). There is one outlier at about (30.3, 97).

Example 4

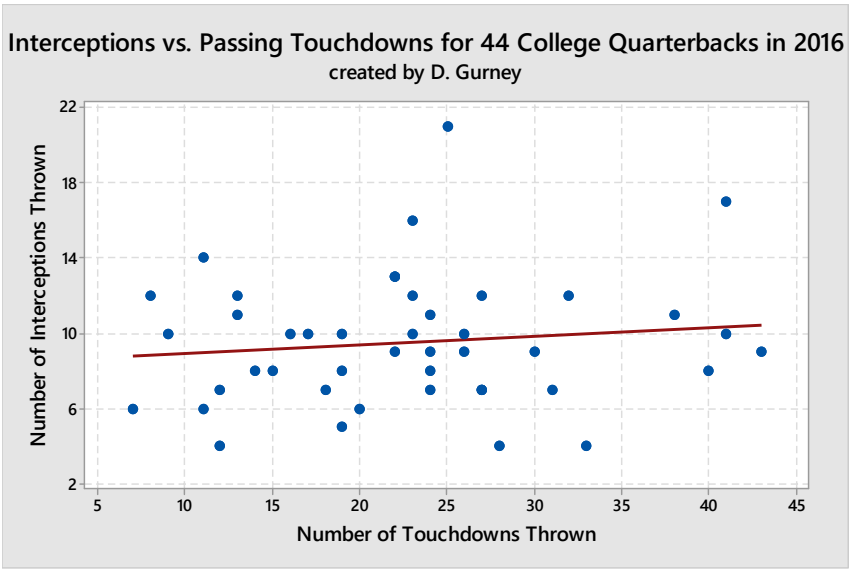


The scatter diagram of total passing yards versus attempts shows a moderate positive association. The equation of the regression line is

$$y = 8.4 + 7.764x.$$

The R-square value is 0.777, which means that about 77.7% of the variation in the passing yards is determined by the regression line. There is one outlier at about (360, 4000). The points at about (200, 1500) and (205, 2000) on the left side and at about (580, 4600), (590, 5000) and (630, 4500) on the right side are influential observations.

Example 5

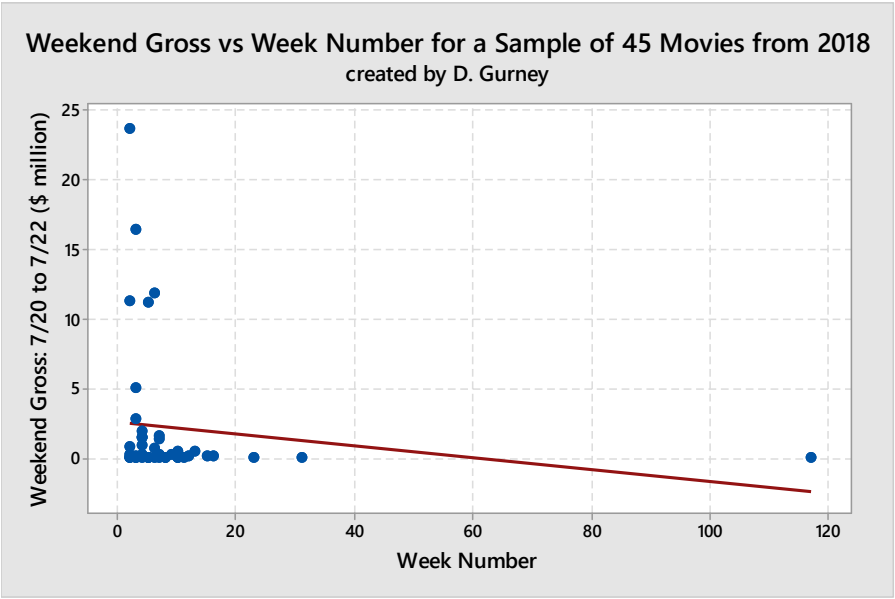


The scatter diagram of interceptions versus passing touchdowns shows a weak positive association. The equation of the regression line is

$$y = 8.516 + 0.04432x.$$

The R-square value is 0.014, which means than about 1.4% of the variation in the number of interceptions is determined by the regression line. There is one outlier at about (25, 21). There are no influential observations.

Example 6

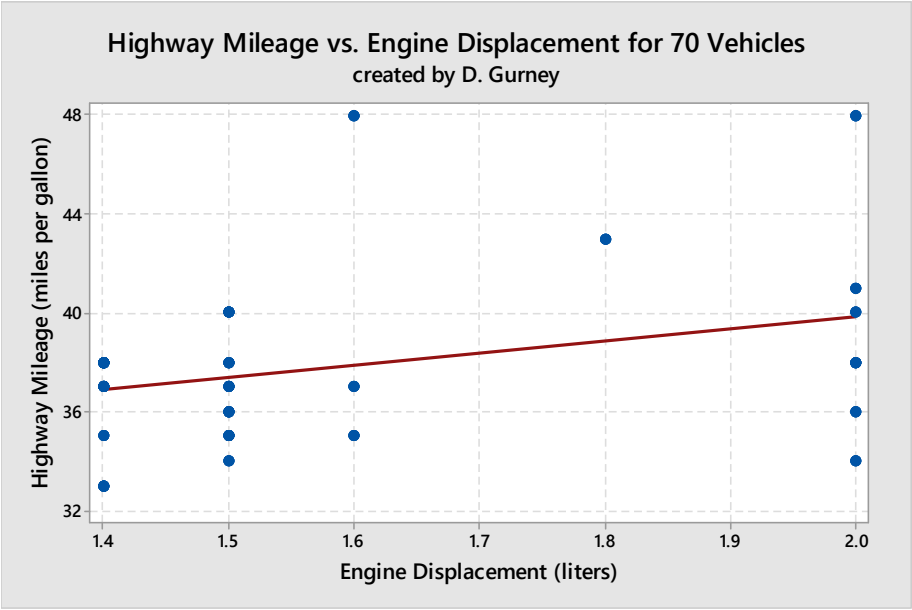


Notice that most of the data points are on the far left side of the graph. The graph shows a very weak negative association. The equation of the regression line is

$$y = 2.547 - 0.4247x.$$

The R-square value is approximately 0.022, which means that about 2.2% of the variation in the weekend gross is explained by the regression line. There is one outlier at about (2, 24), and there is an influential observation at about (118, 0).

Example 7

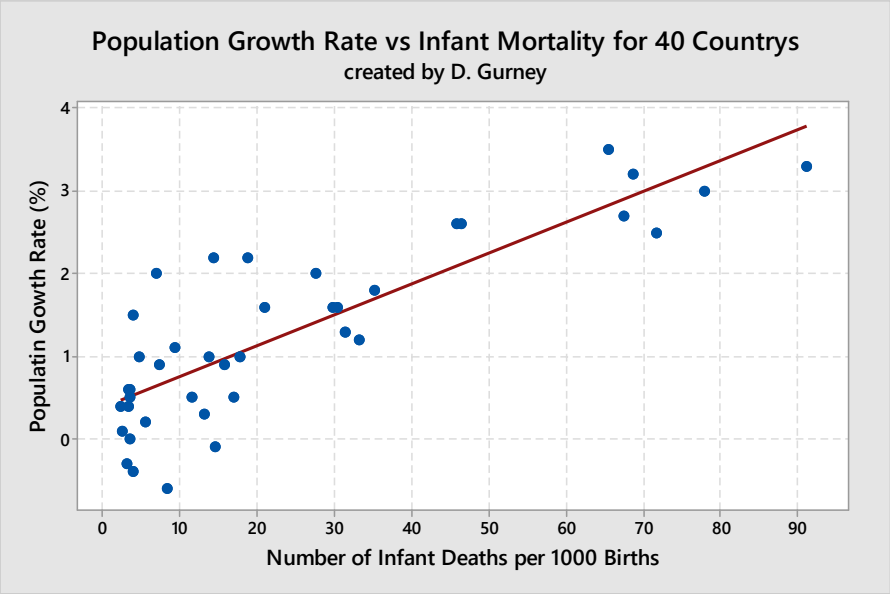


The scatter diagram of highway mileage versus engine displacement shows a weak positive association. The equation of the regression line is

$$y = 29.91 + 4.977x.$$

The R-square value is approximately 0.116, which means that about 11.6% of the variation in the mileage is explained by the regression line. There is one outlier at about (1.6, 48). There are no influential observations.

Example 8

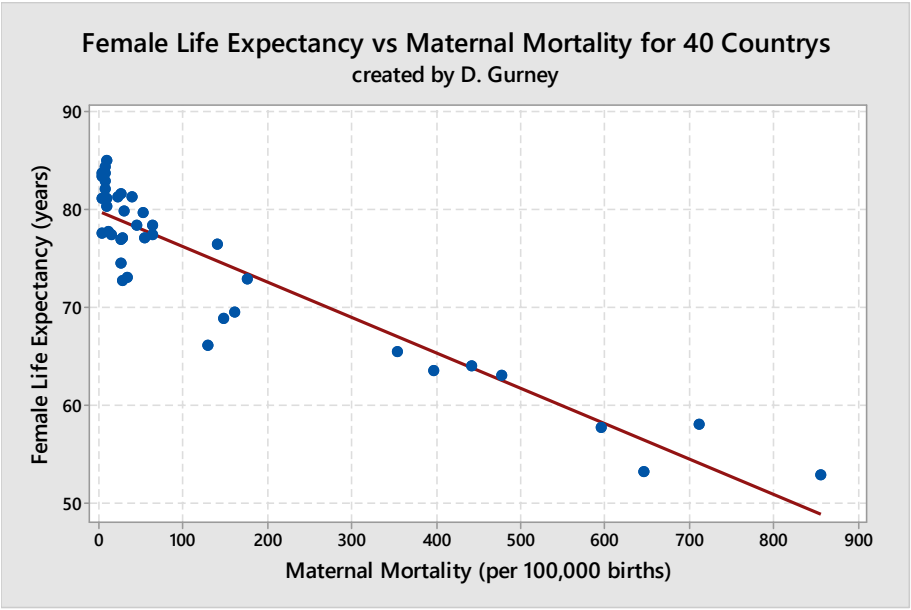


The scatter diagram of population growth rate versus infant mortality shows a moderate positive association. The equation of the regression line is

$$y = 0.3819 + 0.03738x.$$

The R-square value is approximately 0.699, which means that about 69.9% of the variation in the growth rate is explained by the regression line. There is one outlier at about (8, 2). All points to the right of the vertical line through x = 60 are influential observations.

Example 9



The scatter diagram of female life expectancy versus maternal mortality shows a fairly strong negative association. The equation of the regression line is

$$x = 79.88 - 0.03626x.$$

The R-square values is approximately 0.856, which means that about 85.6% of the variation in the life expectancy is explained by the regression line. There is one outlier at about (120, 66). All points to the right of the vertical line through x = 500 are influential observations.