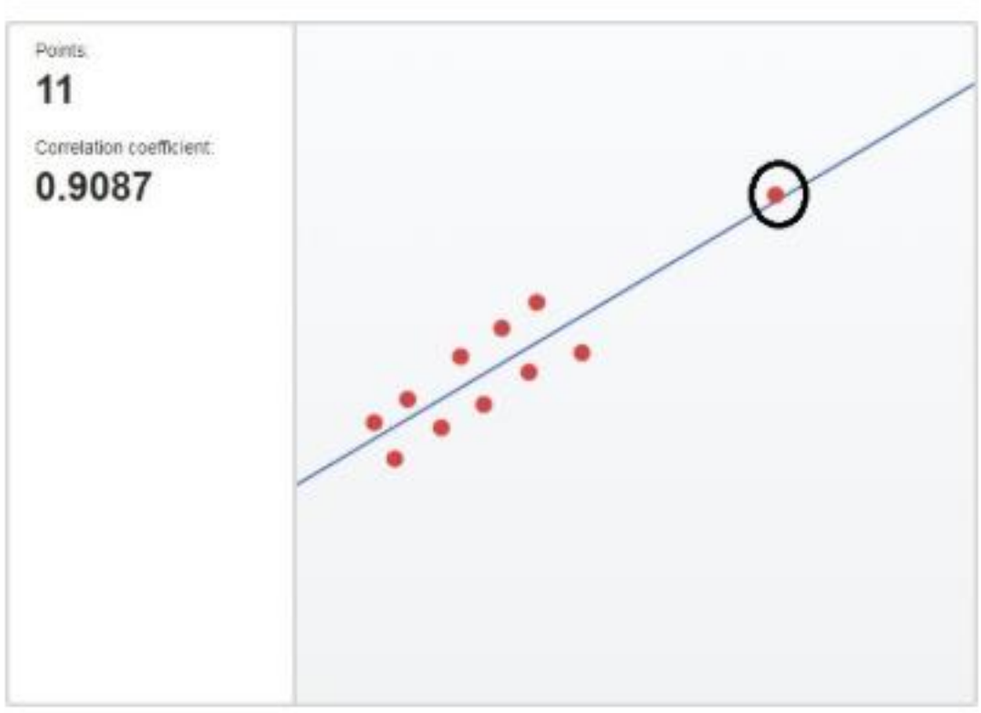
Topic 8.1: Introduction to the Lesson

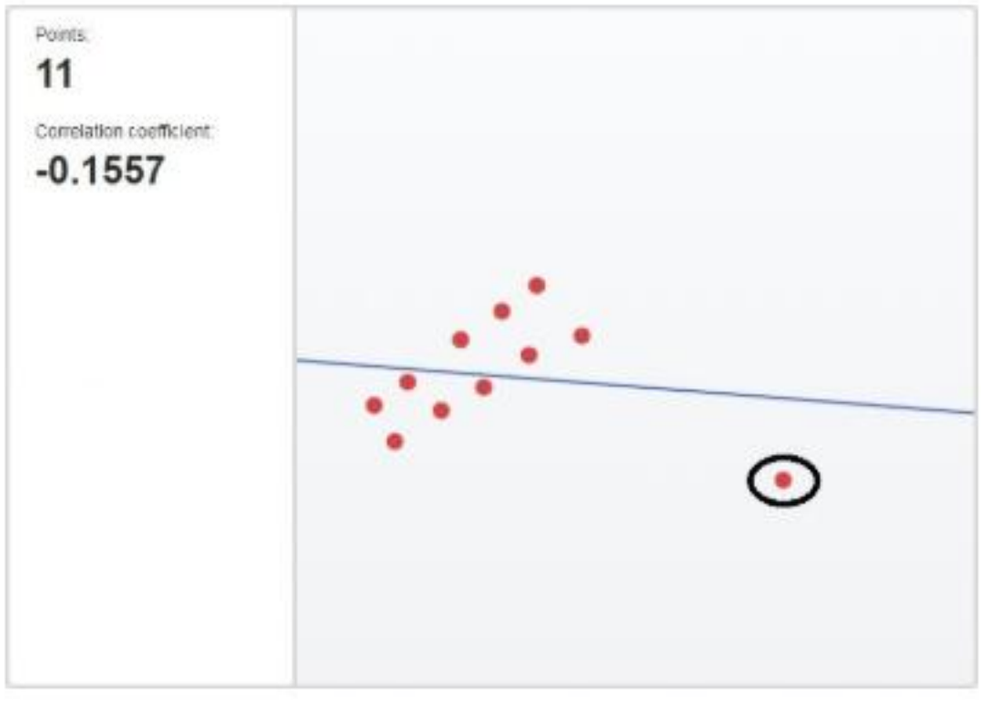
When we compute the sample average, a data point that is much larger or smaller than the rest of the data will have a large influence on the value of the average. We usually check to see if the data point was an error (in which case the value is checked for correctness) or if there is something interesting about the data point that warrants a closer look.

Likewise, we are concerned with influential data points in a regression model, which significantly alter the estimated coefficients and predicted values of a regression model with their presence (or removal).

For example, look at the three scatterplots below, specifically the data point that is circled on the second and third scatterplots (which is a data point added to the first scatterplot).



On the second scatterplot, although the data point that is circled is far away from the other data points, its presence does not drastically affect the estimated regression line.



On the third scatterplot, notice how the circled data point drastically changes the estimated regression line. In this plot, the circled data point is considered influential. We need to determine whether this data point is correct, or if there is something special about this data point that warrants further consideration.

This is a simple example using just one predictor and one response variable. In the multiple linear regression setting (with multiple predictors), it is difficult to visually detect potentially influential data points.

In this lesson, you will learn how to use residuals to identify data points that are potentially outlying. You will also learn a number of ways to measure how influential these data points are. These measures are usually based on how the presence of these data points changes the values of the estimated coefficients and predicted values in your regression model.