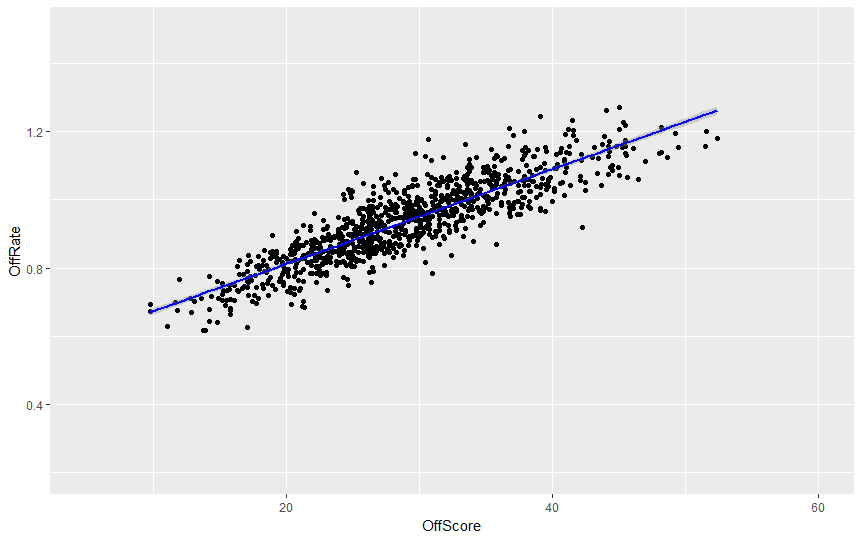
Statistical Analysis of Capstone Project

As dictated in section five of the Springboard Introduction to Data Science, this paper will cover the use of statistical analysis for the final Capstone Project. To review, this Capstone Project will determine if a set of developed formulas based on American football statistics can predict the outcomes for future games as they pertain to their spread (the handicap that is given to each game to allow for “even” betting between two teams). The main point of the project is to determine if manipulation of the formulas will give the user an indication of who to pick. As an off shoot of this outcome, the next steps are to determine if other factors can help improve the odds of success.

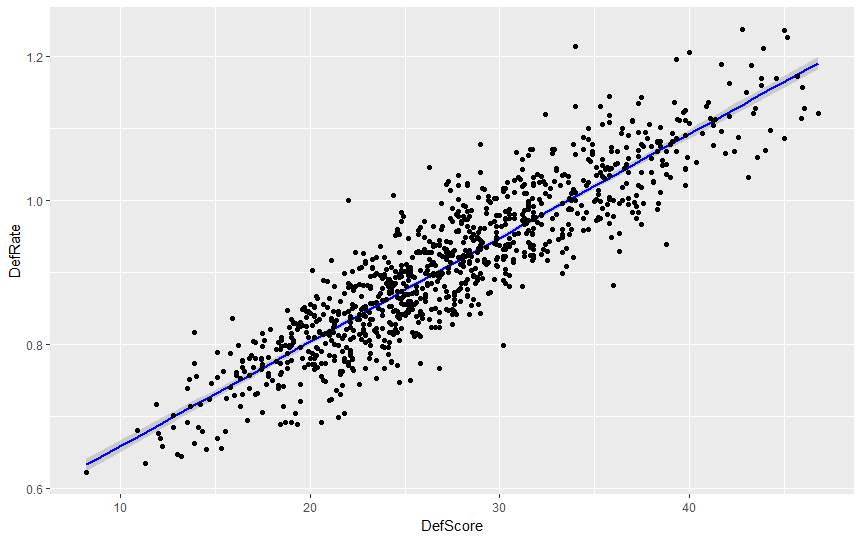
While it is believed that the main set of data (ratings, scores, outcomes, and times of games) did not need statistical analysis in the early going, there is one comparison of data that is not listed above that could help determine if the proprietary formula has viability. This would be to compare the final year ratings for each college football team for the last seven years (2010-2017) with the points per game for each team. There will be a separate comparison with the offensive rating and the points scored per game for each team and the defensive rating with the points allowed per game. Most of the information in this listing was obtained on the NCAA sports website. All told, there was around 1000 data points to analyze for each comparison.

The first test was to determine the correlation of the offensive ratings and the total scoring per team. The ggplot graph plus regression line are shown below:



This correlation looks pretty good at first glance. This does make sense because part of the rating does incorporate the number of touchdowns achieved. However, with other elements to the ratings not directly involved with scoring, this is still an encouraging step. An equation for the regression line was attempted in R, but the experience level has not been achieved yet. For now, Microsoft Excel was used to do the calculations until the skill can be learned. The regression line had a formula of y=0.0139x + 0.5323 and an R2 value of 0.7357.

The same was completed with the defensive statistics and scoring. Here is the plot for that information:



When this information was plugged into Microsoft Excel to get the regression line and R2 value, the result was an equation of y=0.0144x + 0.5149 and R2 value of .8083.

These results tell us that there is a better than average correlation between the ratings and the subsequent comparative point per game totals. But what can this be used for? Is it possible that after doing some calculation manipulations with the ratings themselves that we do not get success, could we utilize the equations do get a better result? There is also something called an Over/Under number. This number is generated as the believed combined total of points between two teams during one game. Another application would be to determine if the calculations that have been achieved here can predict if the outcome of a future came will have more or less combined points than the Over/Under number. This will also be looked at in the project (although there is less data for this than the actual lines and scores of games).