NFL Capstone - Statistics

For the statistics section, I wanted to take a closer look at the relationship between career AV and the various college passing and rushing stats that I have available. I wanted to do that by using statsmodels' multiple linear regression tool. What I did was dedicated the career AV column as my response column (y) and included all of the college passing and rushing stats columns as explanatory columns (X).

The summary of the model told me that the R-squared value was 0.061, which right off the bat was not very optimistic. What this means is that all of these different college stats only explain around 6.1% of the variance in career AV, which is not exactly robust.

The p-values for the individual X values were mostly very high as well. What this tells us is that the higher the p-value, the lower the statistical significance to the model of said X value the specific p-value is associated with. The X values with the highest p-values were: passing attempts, adjusted passing yards/attempt and passing touchdowns. This is somewhat curious because one would think that there would be a clearly positive relationship between career AV and these stats (Which would lead to lower p-values), but that is not the case here. I decided to then rerun the model and take out some of the X values with higher p-values.

This time, I ran the model without including passing yards/attempt, adjusted passing yards/attempt, passing touchdowns and rushing yards. The R-squared for this model returned 0.060, which was slightly lower than our last model. I decided to leave the passing attempts variable in there to see what would happen. It turns out that the p-value jumped from .84 to .94! Some of the other notably high p-values for this iteration of the model were pass completions, passing completion percentage, passing yards and interceptions.

I then repeated the model once again, this time taking out all of the variables except pass yards, rushing attempts, rushing yards/attempt and rushing touchdowns. This time, the R-squared value was even lower (Around .059) and I also started seeing the p-values drop fairly significantly. The only one that was out of hand this time was the rushing yards/attempt p-value.

I ran the model one last time and took out rushing yards/attempt. What I had left was passing yards, rushing attempts and rushing touchdowns. Passing yards returned a p-value of 0.000, rushing attempts returned 0.011 and rushing touchdowns returned 0.010 - All of which were within the threshold of statistical significance established at the beginning. The R-squared this time also dropped even more to 0.057.

Our model suggests that passing yards, rushing attempts and rushing touchdowns are the most statistically significant variables to include in the regression. However, the variables only explain around 5.7% of the total variance in career AV. What this means is that there is a lot more variance that can only be explained by other variables that we do not have access to here. I'm also fairly surprised at the variables determined to be statistically significant. I'm left wondering why are total college passing yards more significant to this model than interceptions or passing touchdowns for example? These results are a little confounding and I'm hoping to get to understand them a bit better moving forward.