**Statistical Question**

Does a team with a better passing offense or rushing offense have an advantage?

**Results**

**Histograms**

From looking at the histograms of the variables for passing yards per game, rushing yards per game, passing attempts, rushing attempts, quarterback rating, and result, the only variable that showed a clear difference in proportion was result. This indicated that the team represented by 0, which was the home team, had a higher proportion of wins than the team represented by 1, which was the away team.

The histograms of wins by passing rank and wins by rushing rank indicated exactly what I would have expected. Teams with higher ranked passing and rushing offenses for that particular season had more wins on average than teams with lower ranks.

The histogram for number of advantages in a game showed a majority of teams have advantages in either 2 or 3 of the 5 variables we were looking at. This tells us that there are not normally teams with advantages in all 5 variables, which means the match ups are normally pretty even.

**PMF**

For the PMF I investigated how much passing and rushing offensive reliance has changed from 2002-2016. I found that the distributions have changed slightly. However, further analysis would have to be done to answer that specific question.

**Correlations**

For correlation, I investigated the relationship between passing yards per game and rushing yard per game first. Surprisingly, there was very little correlation here. I conclude that rushing yards plays only a small role, if any, in predicting passing yards, or vice versa.

I also investigated passing yards per game and quarterback rating. These were obviously correlated because passing yards is part of the quarterback rating calculation. However, it wasn’t an exact correlation because there are a lot of other factor that go into both passing yards and quarterback rating.

**Hypothesis Testing**

All null hypotheses were rejected aside from quarterback rating. For quarterback rating, the winning percentage for teams with an advantage in quarterback rating was not significantly higher than 50%. For passing attempts advantage, the winning percentage was significantly different than 50%, but it was in the other direction. Dividing the p value by 2, we still are able to reject the null hypothesis and we can say that teams with higher passing attempts had a significantly lower winning percentage than 50%. As for advantages in the other variables, their winning percentages were all significantly higher than 50%. Teams with rushing attempts advantages had the highest winning percentage of 60.40%. This would make sense as teams with the lead typically run the ball more often as a strategy to run the clock down.

**Multiple Logistic Regression Model**

The model I created predicted the probability of team 2 winning based on the advantages they had. It was able to predict the correct result 63.45% of the time. From the probabilities for specific advantages, I saw that simply having an advantage in passing alone or rushing alone does not give a team an advantage. The highest probability for winning came from different combinations of both passing and rushing advantages.

**Reflection**

After writing the last paragraph of this report, I realized that predicting whether or not team 2 won probably influenced the model. Team 2 was the away team. If I switched the 0s to 1s in the variables used in the model, it would predict the probability of the Team 1 winning, which would be the probability of the home team winning. From the initial histograms, I saw that home teams won more often than away teams. Therefore, I would expect different results for this model if I was predicting a home team win.

If I were to further this analysis, I think I would also want to figure in the amount of the advantage. For example, a yards per game advantage of 100 should probably be weighted more than a yards per game advantage of 10. Beyond this, I feel that I should have also split my data into training and testing sets to better evaluate my model.

As for areas where I don’t fully understand or am lacking, assumptions are definitely the biggest grey area for me. Thankfully I had a logistic regression problem for this project, so the result was easier to interpret for me. However, evaluating for normality, determining outliers, etc. are things I need to work on to get a better understanding. Not just judging whether data is normal and if there are outliers, but the actual calculations and what to do if the data is not normal and if there are outliers.