NFL Capstone Inferential Statistical Analysis

When it came to the inferential statistical analysis of the dataset, I not only wanted to look at the previously intriguing plots from the Data Story but also some of the more standard statistics people focus on especially yards and points. For stats like those, I felt t-tests were the best to use because I would be comparing continuous variables (that could go very high) with discrete, binary variables (win or loss). The test that was used the most throughout my analysis was chi-squared which also compared continuous variables with the discrete and binary variables but on a smaller scale. With those tests, I dug deeper with filtering around the median and seeing how that changed the p-values as well as also comparing with a t-test. Lastly, for certain continuous variables, I used a Pearson correlation test to see if there was any predictive power between them. For all tests, the null hypothesis was that the statistic had no impact on a team's chance of winning while the alternative hypothesis was that the statistic had an impact on a team's chance of winning; alpha was set at 0.05. With all of that out of the way, here are some of my findings from the inferential statistics analysis.

For the stats I focused on in the Data Story, many tended to be statistically significant. The t-tests for Road Rushing Attempts vs. Road Win as well as Road Time of Possession vs. Road Win each had very low p-values. Others such as Road Passing Attempts vs. Home Win, Road Sacks Given Up vs. Home Win and Home Goal To Go Successes vs. Home Win also showed low p-values in all the tests they underwent including t-tests, chi-squared and chi-squared filtered above and below the median. All therefore were statistically significant and had impacts on a team's chance of winning.

However, Road Passing First Downs vs. Road Win was proven not to be statistically significant as all tests yielded a high p-value meaning passing for First Downs did not have an impact on the away team's chance of winning. When it came to Home INT TDs vs. Home Win as well as Road Fourth Down Successes vs. Home Win, both proved statistically significant with low p-values when it came to t-tests and standard chi-squared but differed when they were filtered by medians. They still showed low p-values above the median but yielded exceedingly high p-values below the median. This may be that because of how infrequently these events occur. With many data points being around 0 or 1 below the median, these statistics would not be significant and therefore not have an impact on the Home team winning. For both of those cases, they may be statistically significant overall but are not quite as impactful as other stats.

One of the more interesting things I found when it came to standard, popular statistics is although rushing and total yards did impact their individual teams' chances of winning, passing yards did not. The t-tests yielded high p-values, which really challenges this idea of how the NFL has become a big passing league. While passing attempts may have an impact, overall passing yards is not statistically significant or impactful meaning teams may want to avoid completely focusing on their passing game. Along those same lines, the Pearson correlation tests between Road Total Yards and Home Total Yards as well as Road Points and Home Points resulted in high p-values which means they are not statistically significant. Because they are not predictive of each other, these stats also negate the idea of a shootout.

For turnovers, both interceptions and fumbles are statistically significant and impact either team's chances (both the team giving it away and the team taking it away) of winning overall especially when it came to t-tests and standard chi-squared tests. For interceptions there is also a negative correlation between interceptions thrown from each team with low p-values. This means that they are statistically significant and can be predictive of each other. Seemingly this relationship would strengthen the idea of a sloppy game, however fumbles tell a different story. First off, when the chi-squared test was filtered by medians, the p-values were very high below the median for lost fumbles, which may tie into what happened with Home INT TDs and Road Fourth Down Successes. Because certain games may have zero or one fumble lost, they may not be statistically significant or have an impact on a team winning. Also, the p-value was high for Road Fumbles Lost and Home Fumbles Lost when they were put through a Pearson correlation test, which means they were not statistically significant, predictive or promote the idea of a sloppy game. Interceptions appear to be more consistently impactful on a team winning compared to fumbles.

Overall the inferential statistics told a mixed story about what stats are significant and which are not. Most of the ones that were focused on in the Data Story did appear to impact the chances of a team winning outside of Road Passing First Downs as well as the below the medians of Home INT TDs and Road Fourth Down Successes. The thing that stood out the most to me, however, was how passing yards, for either team, was not statistically significant. As everyone talks about how focused the NFL has become on the passing game, it actually appears that running the ball has more of an impact on a team winning the game. And couple that with the lack of significance for correlations between Total Yards as well as Total Points and this image of shootout NFL games seems less clear. When it comes to turnovers, both interceptions and fumbles are statistically significant but the former shows more consistency along with more predictive power between away and home teams. This leads more to the idea of a sloppy game, which could be a more likely outcome when teams get pass-happy. Couple this with the Data Story and winning in the NFL looks to be a lot less about the high-powered passing games that many assume about the league.