Inferential Statistics

I started my investigation by looking at the Pearson correlation between every numeric independent variable and the dependent variable: yards. I created a data frame containing four columns: variable, Pearson's correlation coefficient, absolute value of Pearson's correlation coefficient, and p-value. I then selected only rows with a p-value less than 0.05 and sorted by absolute correlation coefficient. From here I could quickly see which variables had the strongest relationships with yards. Runner's acceleration, runner's distance to the defense's centroid, and yards left to the goal line were the three variables with the strongest relationships with yards. Runners with higher acceleration and runners further from the centroid of the defense gain more yards. Runners further away from the goal line also gain more yards, however this is expected because the runner can only gain as many yards as are left to the goal line. Some other variables that had a strong relationship with yards gained were runner's speed, distance to first down, runner's distance to closest defender, and combined weight of the defensive line.

I also conducted tests to see if average yards gained differed for some categories. First I looked at whether average yards gained differed by the down of the play (first down, second down, etc.). The Levene test showed that the variances of the categories were not homogenous, and I know that yards gained are not normally distributed so I used the Kruskal-Wallis H-test to determine whether median yards gained differed by category. This test was highly significant, and pos-hoc analysis shows that runs on fourth down gain significantly fewer yards than those on first, second, or third downs. Additionally, runs on third down gain fewer yards than those on first or second.

Another test was to check whether runs to the weak side of the field gained more yards than those to the strong side. The strong side of the field is defined as the side (left or right) with

more offensive linemen. Although runs to the strong side of the field gain more yards on average, an independent samples t-test shows that this difference is not significant.