Rhoads MacGuire
Data Wrangling for Capstone

The data I am using for this project comes from Kaggle and is comprised of nfl game data, including points scored by each team, the point and over/under spreads, and weather details on the game. The data set did not require a lot of cleaning, but it did require some additions and alterations to better suit my intentions. Those intentions of course are to analyse the effects of weather on the accuracy of both the point and over/under spread.

The first cleaning step I took was to delete columns that would not improve my analysis. Most of these columns contained superficial details like the name of the stadium. Others contained data that could have been useful, but were so unpopulated it threatened their validity. For example the column titled “weather_detail” claims that only 100/ 7725 NFL games in the last thirty years featured rain. That number appears to be too low and including it could show us false conclusions a more robust dataset would disprove. It is possible to gauge the effects of extreme weather with temperature and wind speed data so losing this column should not hamper the analysis.

My next step was to eliminate rows where essential information was missing. This informations included, point and over/under spreads, temperature, wind speed, and points scored by each team. I cleaned these rows using a boolean series and the .notnull() method. The result included only rows where all the data was present. While reviewing the data after this step, I remembered the 1987 NFL strike where the owners continued the season with replacement players. While this period was only a handful of weeks, I believe the resulting changes in the league warrant eliminating data from before this point. This left us with 7725 full rows of data, but some adjustments still had to be made.

For my analysis I wanted a column that was the score of the game relative to the spread. To do this, I decided to make a column, “results”, that would indicate the point difference in the game between the two teams. It would be negative if the favorite won and positive if the underdog won. This would make it easy to use in formulas with the spread column, which is always negative, to indicate the favorite. To create this column took a few steps. First, I needed to change the home and away team columns to match the ‘spread_favorite_id’ column so the ‘spread_favorite_id’ would match either the home or away team. This was complicated by the fact that historical teams, such as the Houston Oilers, were associated with the same “TEN” team id, as in the Tennessee Titans. To make the data easier to analyze, I converted all historical teams to the abbreviations of their current iterations. The names of the teams do not really matter but this was essential in further analysis. To do this I wrote a method called teamSwitch. This method takes the long hand name for a team and reduces it to their 3 or 2 letter acronym. Once the acronyms matched I made a column to show whether the home team was favored and created the ‘results’ column.

When checking my data for outliers I realized that while some values in the column, point spread, exceeded 3 standard deviations from the mean they have value because the most important factor is how the spread relates to the outcomes. Additionally, their does not seem to much value in eliminating observations from something that is so random as football.