M1: Data Assignment

Jason Merten

Data Exploration:

The task for this assignment was to examine the Moneyball dataset which has a total of 2276 rows and 16 columns of data of various statistics for baseball teams from 1871 to 2006. Of the 16 columns of data, there are 6 columns that contain missing (NA) values: **BATTING_SO**, **BASERUN_SB**, **BASERUN_CS**, **BATTING_HBP**, **PITCHING_SO**, and **FIELDING_DP**. There are two columns that I'd like to point out before getting further into the data exploration and preparation: **BASERUN_CS** and **BATTING_HBP**. Both columns have 772 and 2085 missing values, respectfully, and should be removed from further use in our model due to the large amount (33% and 91%) of missing data. For the purposes of this assignment, I filled in the missing data using normal distributions. The remainder of the data exploration will be covered as we talk through the remaining columns of data that contain missing values.

Data Preparation:

BATTING_SO: Batting Strikeouts contains 102 missing values (~4%) with no extreme outliers. Most of the data sits between 548 and 930 with a median of 750. The data also appears to generally follow a normal distribution (with the minor exceptions that there's a small trough near the median and a small increase near 0). These facts led me to believe that the best way to handle the missing values would be to fill the data with values from a normal distribution fitted to the sample data's mean and standard deviation. After creating a new flag variable and filling in the missing data, the new data appears to fit well within the original data set and all missing values are no longer present.

BASERUN_SB: The Stolen Bases data contains 131 missing values (~6%) and has 22 data points outside the 99th quantile. Looking at the data minus the outliers, the data looked to follow a gamma distribution which aided in filling in the missing data. After the missing data was filled in, all outliers were reduced to the 99th quantile. Before the outliers were modified, I created a new variable outlier_TEAM_BASERUN_SB, to annotate that the values were altered.

PITCHING_SO: The Pitching Strikeouts data contains 102 missing values (~5%) and 22 points that sit outside of the 99th quantile which needed to be reduced due to the impact on the centrality of the data. After handling the outliers and creating the flag variable, the data appeared to be following a normal distribution which I randomly sampled to fill in the missing data. Comparing the original data summary with the new summary, the filled data doesn't seem to affect the means of centrality too much.

FIELDING_DP: The Fielding Double Plays data contains 286 missing values (~10%), no significant outliers, and appears to be following a normal distribution. After creating a flag variable, the missing variables were filled using a random sampling of a normal distribution fitted to the sample data. Plotting the filled variables with the originals, the filled variables appear to be well integrated and random.