Moneyball

David Moste, Vanita Thompson, Sadia Perveen

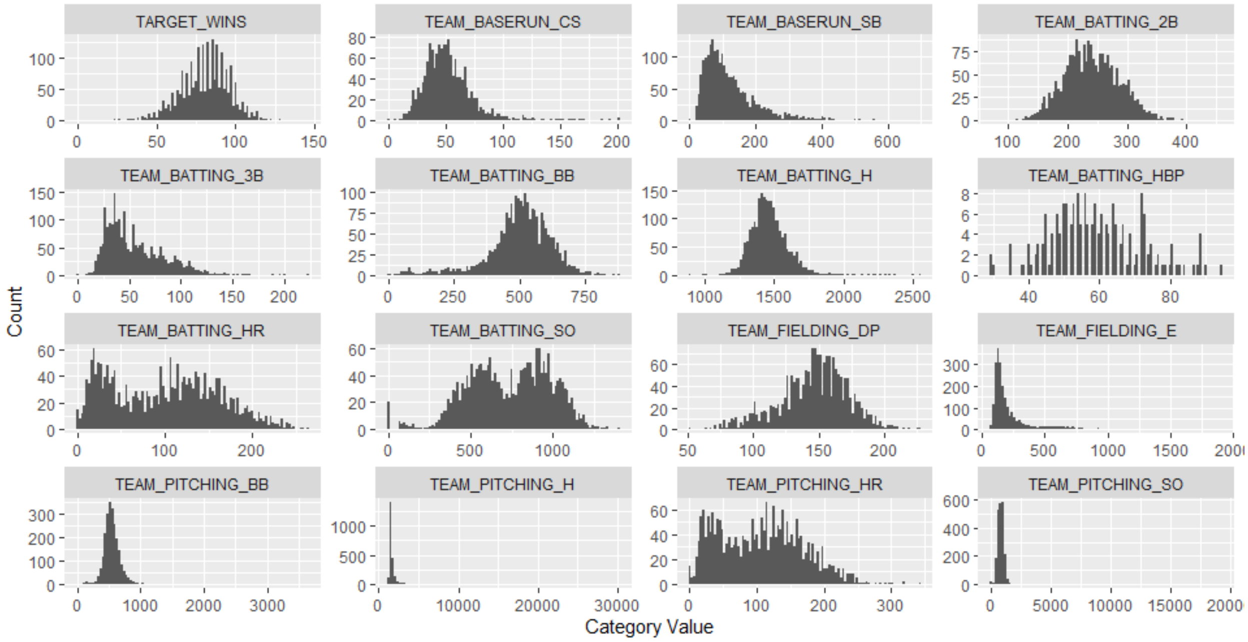
3/1/2021

Data Exploration

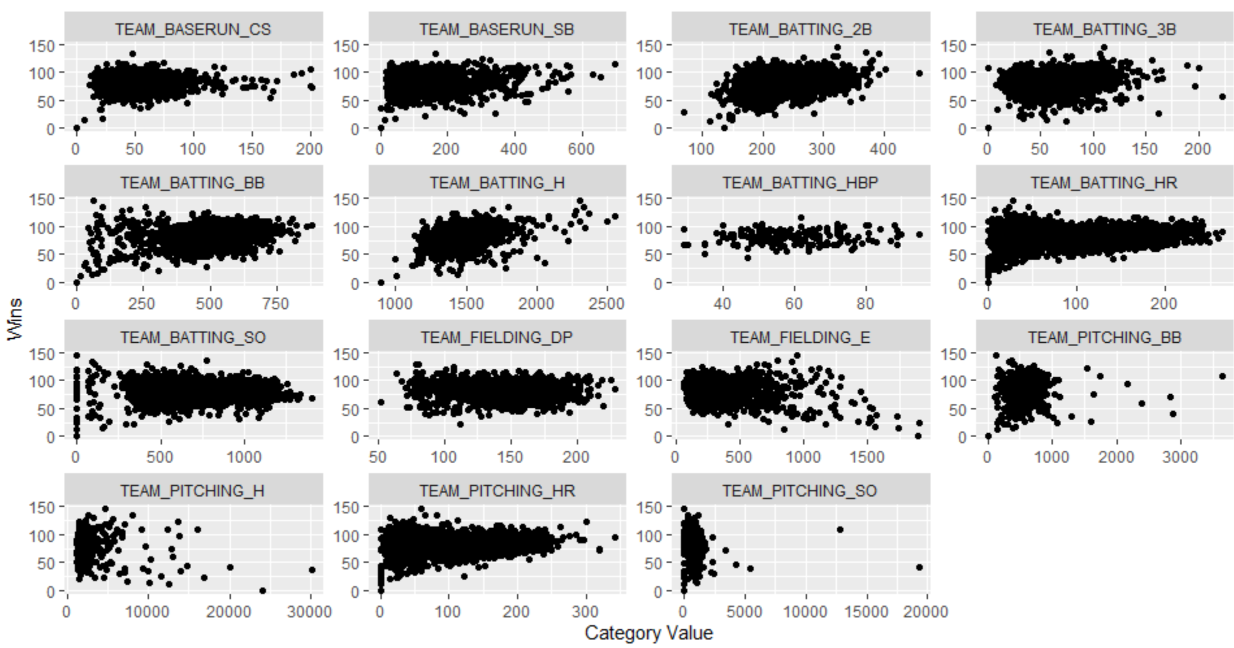
After grabbing the data, I first checked out a summary of the data to see the predictor variables provided along with their summary statistics. This also allowed me to see which predictor variables contained missing data. This summary data can be seen in the table below.

Predictor	Min	Median	Mean	Max	NAs
FIELDING_DP	52	149	146.4	228	286
FIELDING_E	65	159	246.5	1898	0
PITCHING_SO	0	813.5	817.7	19278	102
PITCHING_BB	0	536.5	553	3645	0
PITCHING_HR	0	107	105.7	343	0
PITCHING_H	1137	1518	1779	30132	0
BATTING_HBP	29	58	59.36	95	2085
BATTING_SO	0	750	735.6	1399	102
BATTING_BB	0	512	501.6	878	0
BATTING_HR	0	102	99.61	264	0
BATTING_3B	0	47	55.25	223	0
BATTING_2B	69	238	241.2	458	0
BATTING_H	891	1454	1469	2554	0
BASERUN_CS	0	49	52.8	201	772
BASERUN_SB	0	101	124.8	697	131

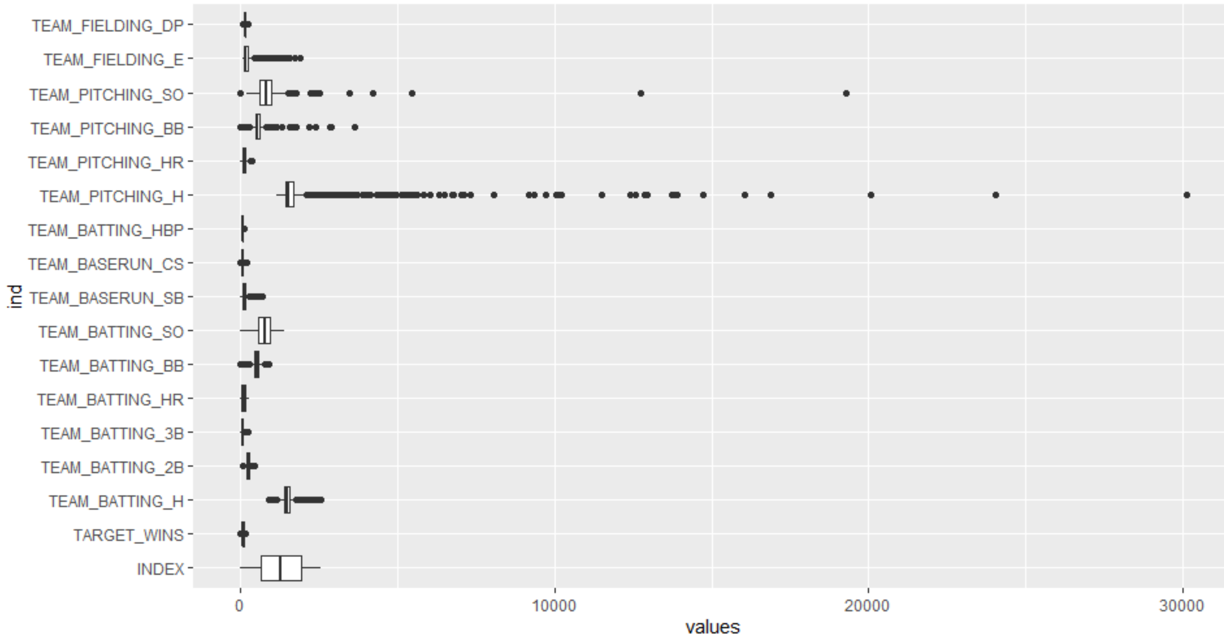
I then created three plots: a small multiples histogram, a small multiples scatterplot, and a boxplot.



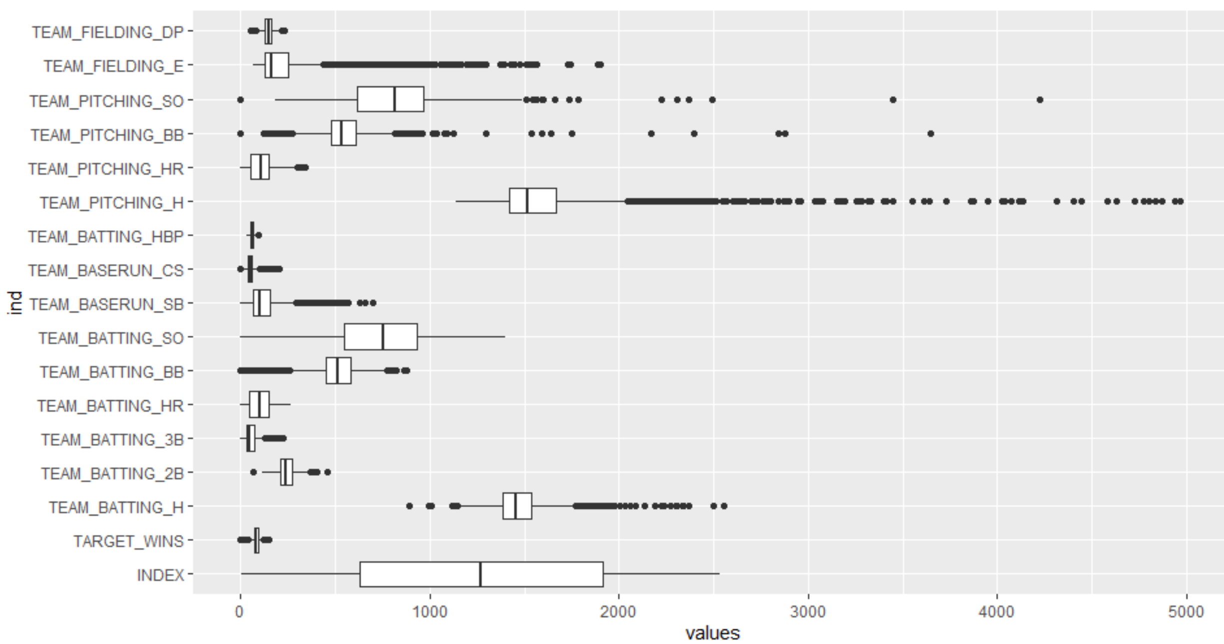
The purpose of the histogram was to get a sense of the normality of each variable. Upon looking at the histogram, it was easy to see that TEAM_BASERUN_CS, TEAM_BASERUN_SB, TEAM_BATTING_3B, TEAM_BATTING_HR, TEAM_FIELDING_E, and TEAM_PITCHING_HR were right skewed and would need to be transformed.



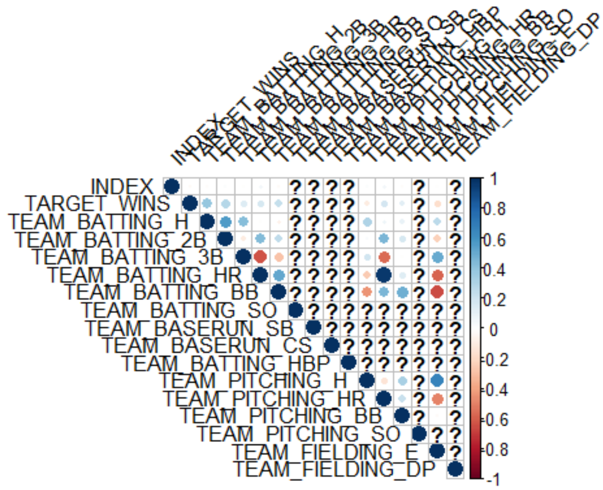
The purpose of this scatterplot was to get a sense of the relationship between each variable and TARGET_WINS. From this, you can see that no predictors have a strong negative relationship to TARGET_WINS, but TEAM_BATTING_H does seem to have a clear positive correlation.



The purpose of the boxplot was to see the data in another light and to get a sense of where there were outliers. It was easy to see at this point that TEAM_PITCHING_H contained a bunch of outliers at the top of the range.



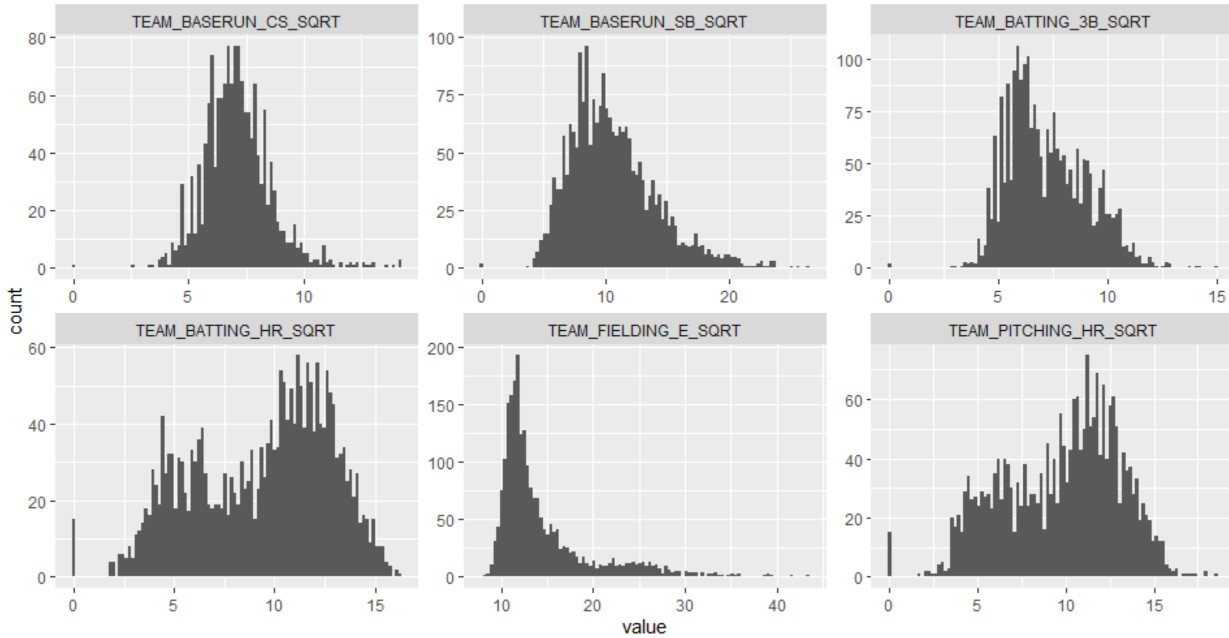
I then zoomed in on the boxplot to get a better sense of outliers in the other predictors.



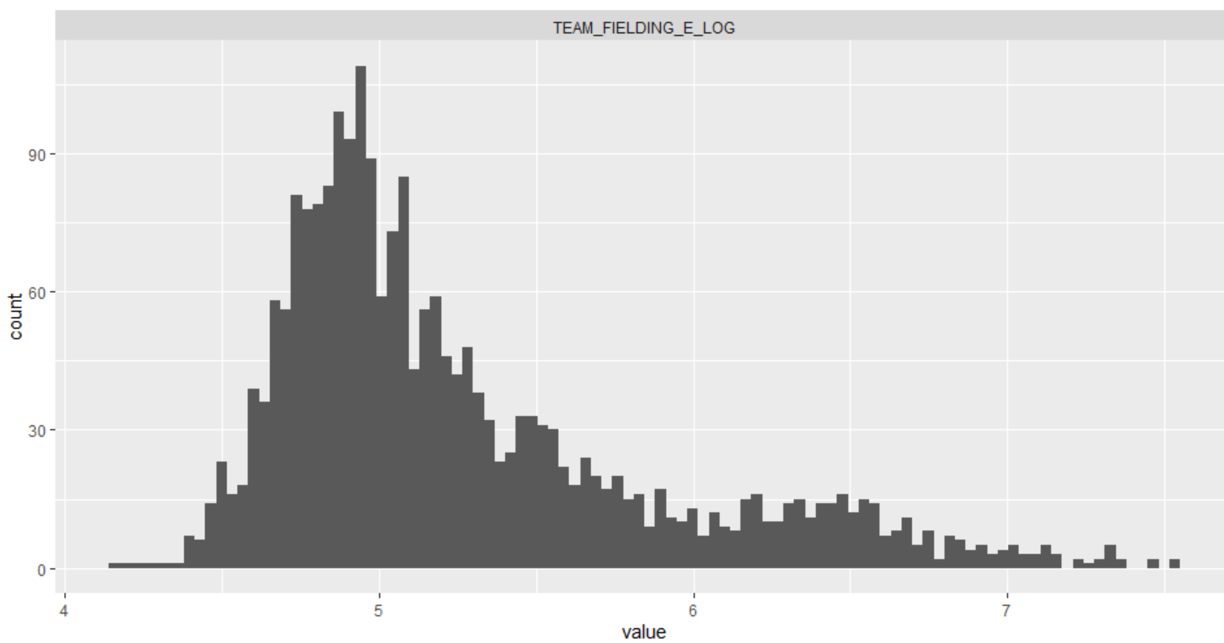
Finally, I created a correlation plot to show how different predictors are related to the target as well as each other. From this plot, it's easy to see that wins is most positively correlated to TEAM_BATTING_H and most negatively correlated to TEAM_FIELDING_E. As expected, other batting categories seem to have positive correlations as well. It is interesting to note that TEAM_PITCHING_HR has a positive correlation too, which is certainly not expected. Some other information that comes out of this visual is a strong correlation between TEAM_BATTING_HR and TEAM_PITCHING_HR and between TEAM_PITCHING_HR and TEAM_FIELDING_E as well as a strong negative correlation between TEAM_BATTING_BB and TEAM_FIELDING_E and between TEAM_FIELDING_E and TEAM_BATTING_HR.

Data Preparation

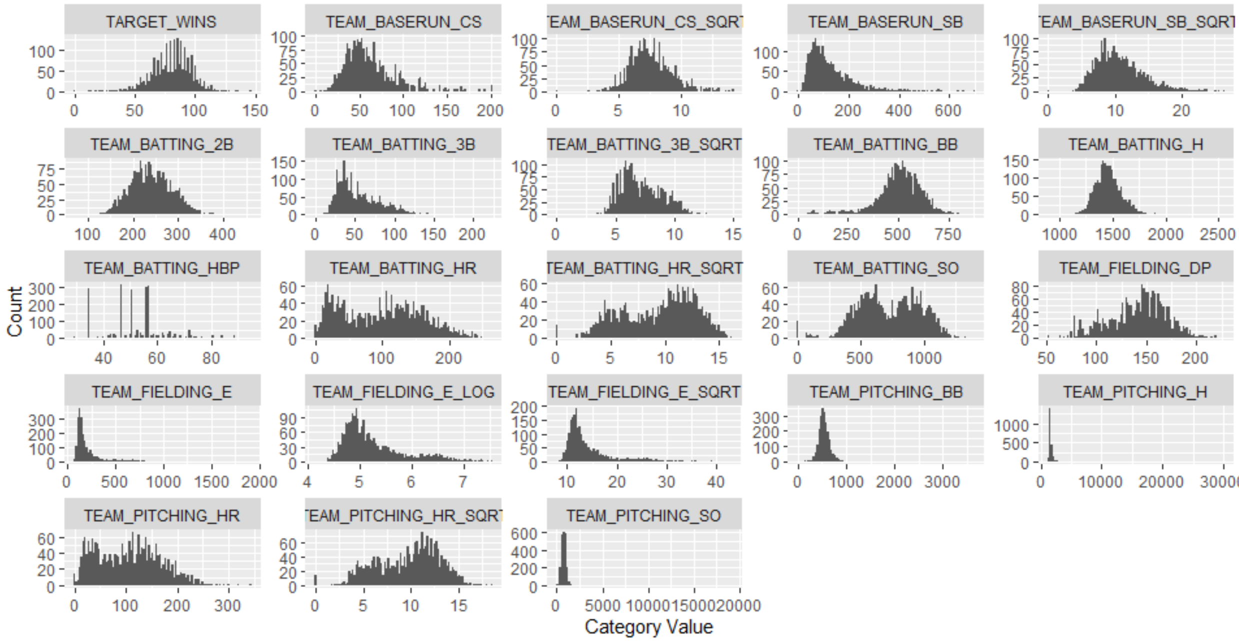
To start data preparation, I performed a few transformations. I did a square root transformation on each of the following variables to correct for their right skew : TEAM_BASERUN_CS, TEAM_BASERUN_SB, TEAM_BATTING_3B, TEAM_BATTING_HR, TEAM_FIELDING_E, and TEAM_PITCHING_HR. I ultimately chose to use a square root transformation instead of a log transformation because many of the variables had large portions of their data with values of 0. This makes log transformations a little bit less useable since you end up with -Inf values.

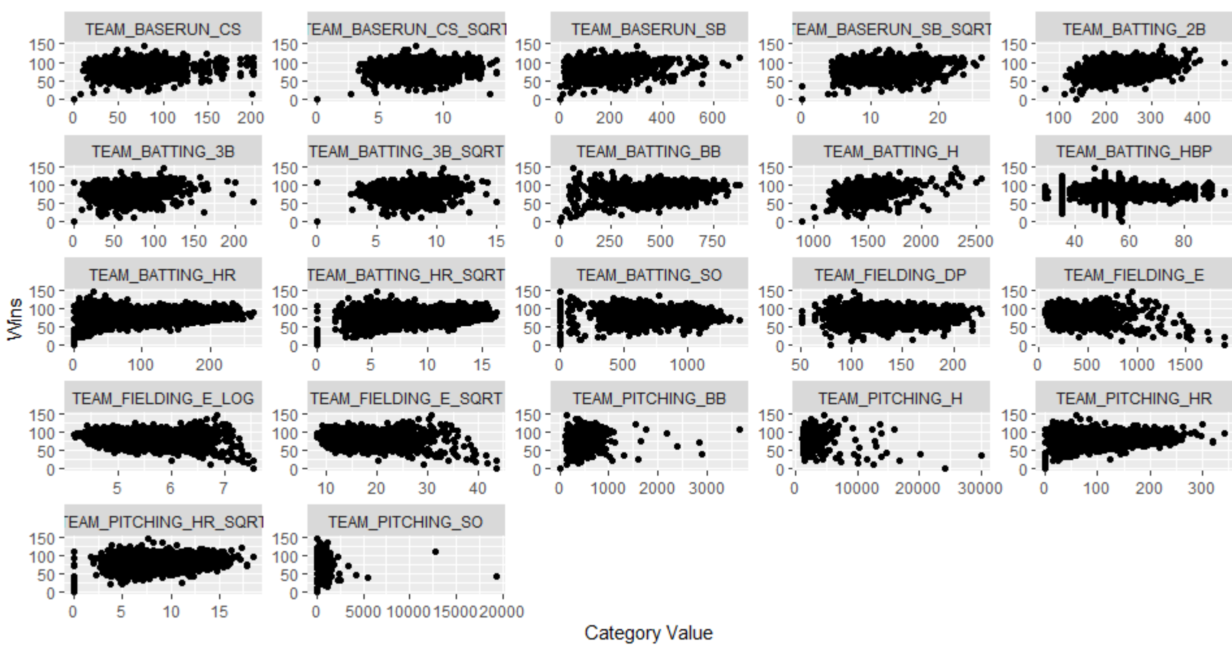


I then viewed a histogram of all the transformed predictors that I created. The histogram showed a clear bimodal distribution for TEAM_BATTING_HR and TEAM_PITCHING_HR. It also showed that TEAM_FILEDING_E was still highly right skewed. Due to this, I decided to take a log transform of TEAM_FIELDING_E to check if that would correct the skew. As can be seen below, this log transformation helped, but was not perfect.



Next, I used the MICE package to impute missing values. I used MICE to implement multiple imputations using predictive mean matching method. After imputing missing values, I created two new plots: a histogram to view normality and a scatterplot to see outliers and correlation.

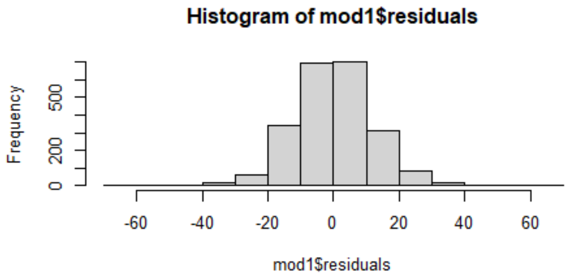


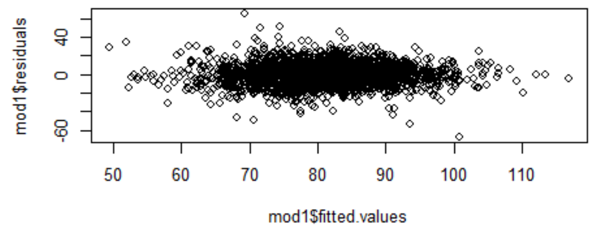


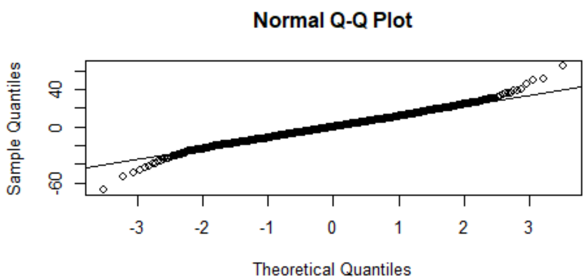
Finally, I created a new predictor, TEAM_BATTING_OB, which was meant to show how often a team got on base and I filtered a few predictors to remove extreme outliers that appeared to have some leverage.

Build Models

The first model I built was simply every variable in the data (excluding variables where I had later taken a transformation).

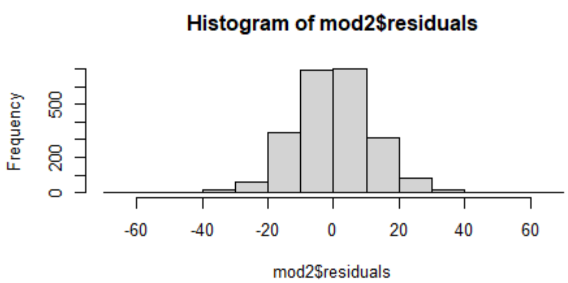


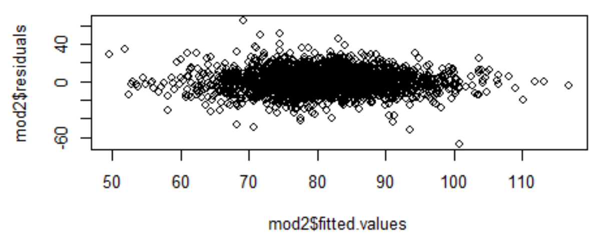


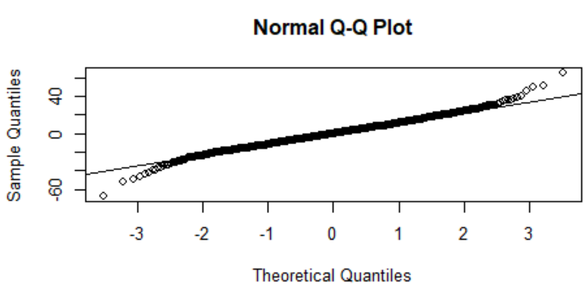


Oddly, this first model found that both TEAM_BATTING_2B and TEAM_FIELDING_DP have a negative coefficient, which suggests increasing them would decrease TARGET_WINS. On the opposite side, TEAM_BASERUN_CS_SQRT and TEAM_PITCHING_H have a positive coefficient, suggesting that they increase TARGET_WINS.

The second model I built was based off the first model, except that I iteratively removed the predictor with the highest p-value until the r-squared value was no longer increasing.

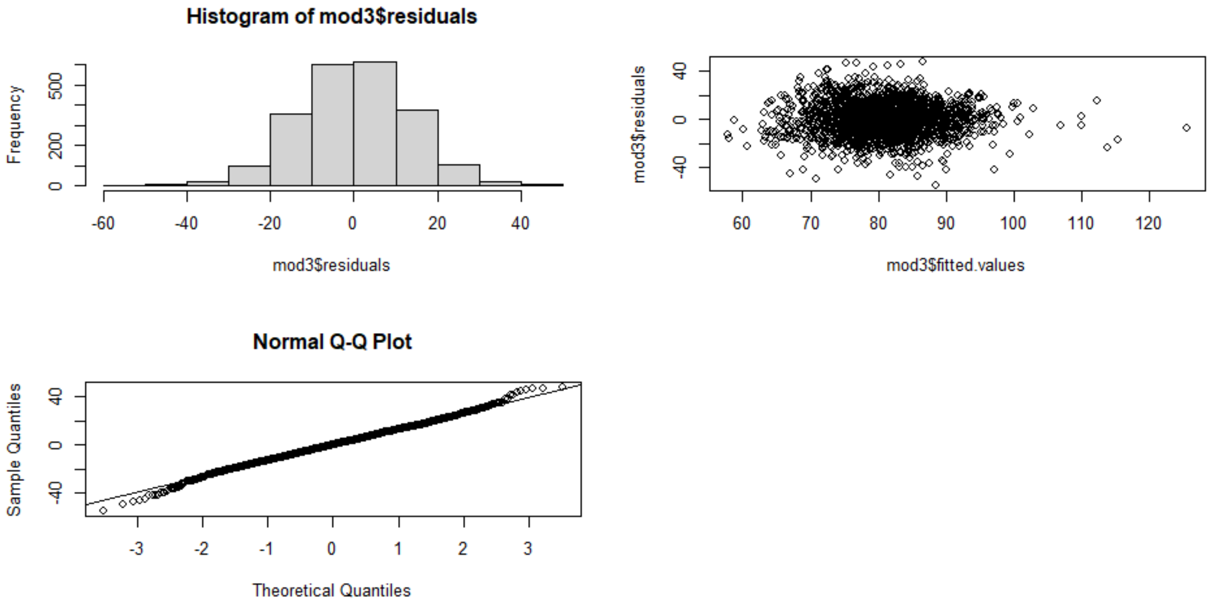






For the second model, TEAM_BATTING_2B still has a negative coefficient and TEAM_PITCHING_H still has a positive coefficient, both of which don't make a ton of immediate sense.

The final model I created was based off of the initial correlation plot I created, using the variables that had the strongest correlation (either positive or negative).



For the third model, all of the slopes make intuitive sense, but the overall fit is rather poor with an adjusted r-squared of 0.197.

Select Models

Code