DATA 621 - Homework 1

2022-09-10

Problem Statement and Goals

Our objective is to make a linear regression model that can predict how many wins a baseball team will have in a season based on certain metrics. The variables we have been provided theoretically have positive or negative effects on the total number of wins. We will be exploring this in depth in our research to figure out which variables are correlated the most strongly with the wins, as well as finding out if some of the variables can be consolidated using known conventional baseball-stats algorithms like SABER.

Data Exploration

Viewing Data

Upon first glance, the data contains 17 columns. The index column will be ignored for analysis purposes, and so that leaves the other 16. `TARGET_WINS` is the variable we want to investigate with regards to how well it is correlated with the other columns. To give some context, every row represents a baseball team and its performance during a particular season. `TARGET_WINS` is the number of wins, and each column after that represents a particular metric for the season. For example, `TEAM_BATTING_H` represents how many base hits by batters occurred for that team during the season. `TEAM_PITCHING_E` represents how many times an opposing team made a pitching mistake during the season. In general, there are four categories of feature types:

- Batting
- Baserunning
- Pitching
- Fielding

```
##   TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
##  Min.   :  0.00   Min.   : 891   Min.   : 69.0   Min.   :  0.00
##  1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0   1st Qu.: 34.00
##  Median : 82.00   Median :1454   Median :238.0   Median : 47.00
##  Mean   : 80.79   Mean   :1469   Mean   :241.2   Mean   : 55.25
##  3rd Qu.: 92.00   3rd Qu.:1537   3rd Qu.:273.0   3rd Qu.: 72.00
##  Max.   :146.00   Max.   :2554   Max.   :458.0   Max.   :223.00
##
##  TEAM_BATTING_HR  TEAM_BATTING_BB TEAM_BATTING_SO  TEAM_BASERUN_SB
##  Min.   :  0.00   Min.   :  0.0   Min.   :   0.0   Min.   :  0.0
##  1st Qu.: 42.00   1st Qu.:451.0   1st Qu.: 548.0   1st Qu.: 66.0
##  Median :102.00   Median :512.0   Median : 750.0   Median :101.0
##  Mean   : 99.61   Mean   :501.6   Mean   : 735.6   Mean   :124.8
##  3rd Qu.:147.00   3rd Qu.:580.0   3rd Qu.: 930.0   3rd Qu.:156.0
##  Max.   :264.00   Max.   :878.0   Max.   :1399.0   Max.   :697.0
##                                   NA's   :102      NA's   :131
##  TEAM_BASERUN_CS TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR
##  Min.   :  0.0   Min.   :29.00    Min.   : 1137   Min.   :  0.0
##  1st Qu.: 38.0   1st Qu.:50.50    1st Qu.: 1419   1st Qu.: 50.0
```

```
##  Median : 49.0   Median :58.00    Median : 1518   Median :107.0
##  Mean   : 52.8   Mean   :59.36    Mean   : 1779   Mean   :105.7
##  3rd Qu.: 62.0   3rd Qu.:67.00    3rd Qu.: 1682   3rd Qu.:150.0
##  Max.   :201.0   Max.   :95.00    Max.   :30132   Max.   :343.0
##  NA's   :772     NA's   :2085
##  TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E  TEAM_FIELDING_DP
##  Min.   :   0.0   Min.   :    0.0   Min.   :  65.0   Min.   : 52.0
##  1st Qu.: 476.0   1st Qu.:  615.0   1st Qu.: 127.0   1st Qu.:131.0
##  Median : 536.5   Median :  813.5   Median : 159.0   Median :149.0
##  Mean   : 553.0   Mean   :  817.7   Mean   : 246.5   Mean   :146.4
##  3rd Qu.: 611.0   3rd Qu.:  968.0   3rd Qu.: 249.2   3rd Qu.:164.0
##  Max.   :3645.0   Max.   :19278.0   Max.   :1898.0   Max.   :228.0
##                   NA's   :102                        NA's   :286
```

From the above summary, we can see that that target variable is roughly normally distributed, with a mean of total wins around 80 games. This makes intuitive sense, as a standard season is 162 games, we would expect that the average number of wins would be roughly half of this value.

There are a few columns which appear to have outliers, particularly TEAM_PITCHING_H, and we will investigate those in depth throughout our data exploration and data preparation steps.

NA exploration

As can be seen below, some of the columns have missing values. Contextually, this can be possible because not every metric must have a value- for example it is possible that an entire season can be played without a batter being hit by the pitch. However it is less likely that an entire season can be played without any strikeouts by batters. We did some research and came up with ways to address each of these issues- more on that later.

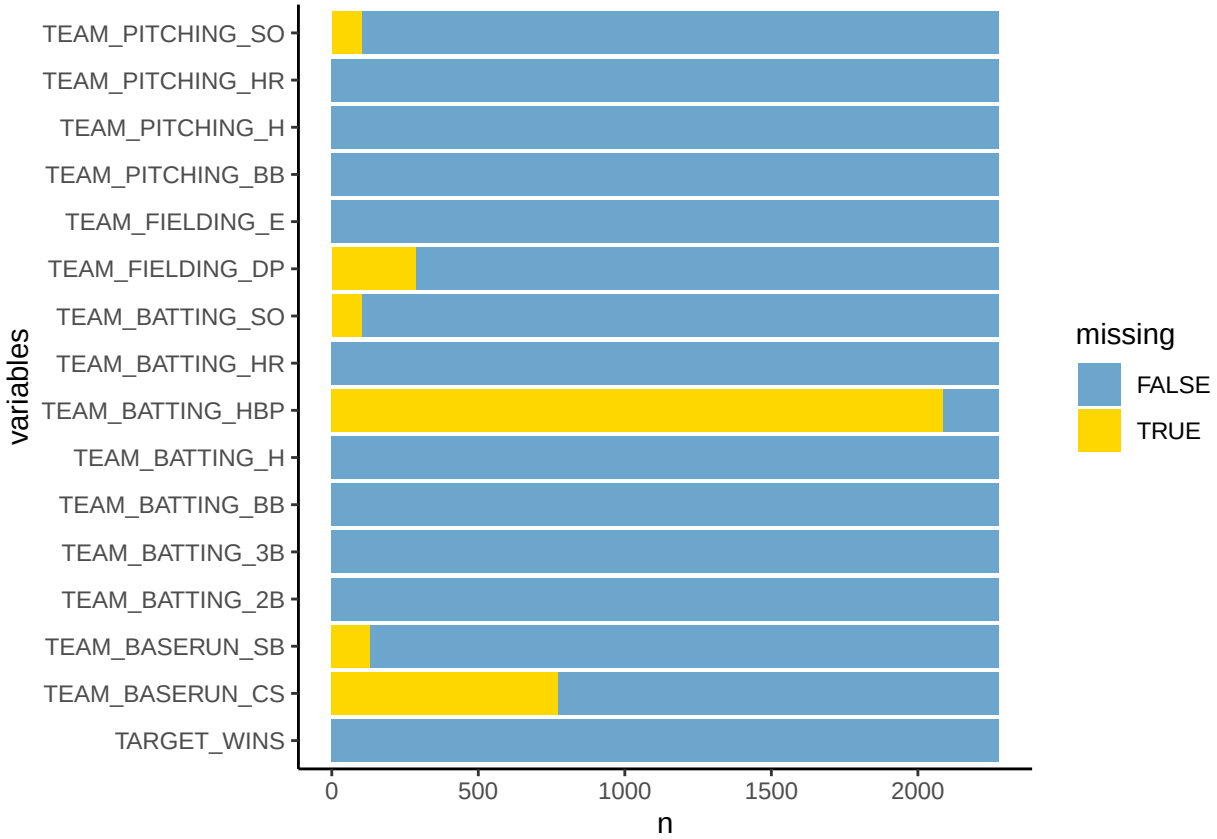


Figure 1: Barplot of number of missing values for each predictor.

Outliers

Another question we had was one of outliers- some of the values were way too high to be realistic of a season of baseball - such as one team having over 20,000 strikeouts.

Below we can see very quickly that some variables have extreme outliers.

```
## Warning: Removed 3478 rows containing non-finite values (stat_boxplot).
```

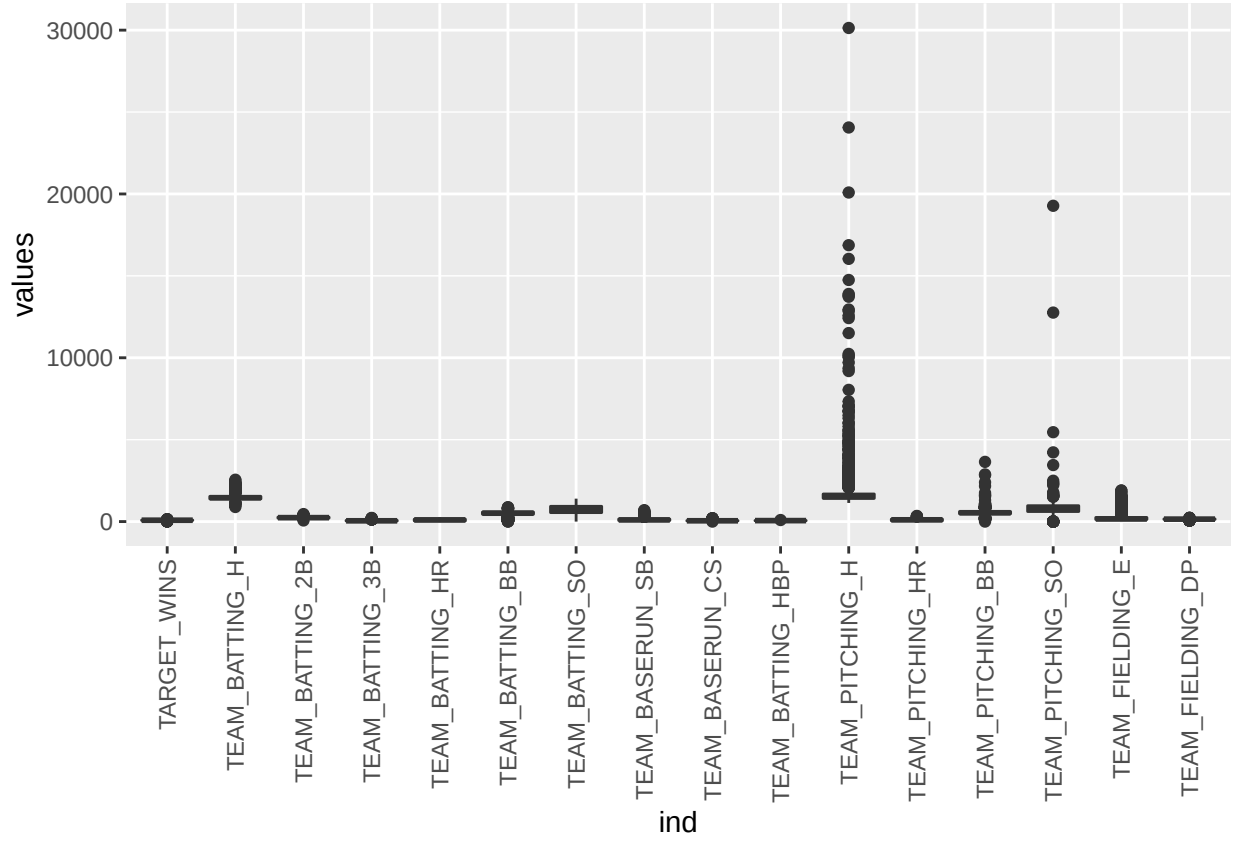


Figure 2: Boxplots for all variables.

From the chart above, we can see that the most problematic features include `TEAM_PITCHING_H`, `TEAM_PITCHING_BB`, and `TEAM_PITCHING_SO`. Where available we will employ cutoffs based on third party reference data such as baseball-almanac.com. If there is no available data, we will use other logical imputation methods to replace the outliers with reasonable values more fit to the data.

Data Skew

It's important to understand the distributions of each feature. Optimally, we would want to see normal distributions in order to create an effective regression model.

A histogram for each of the variables is provided in Figure 3.

```
## Warning: Removed 3478 rows containing non-finite values (stat_bin).
```

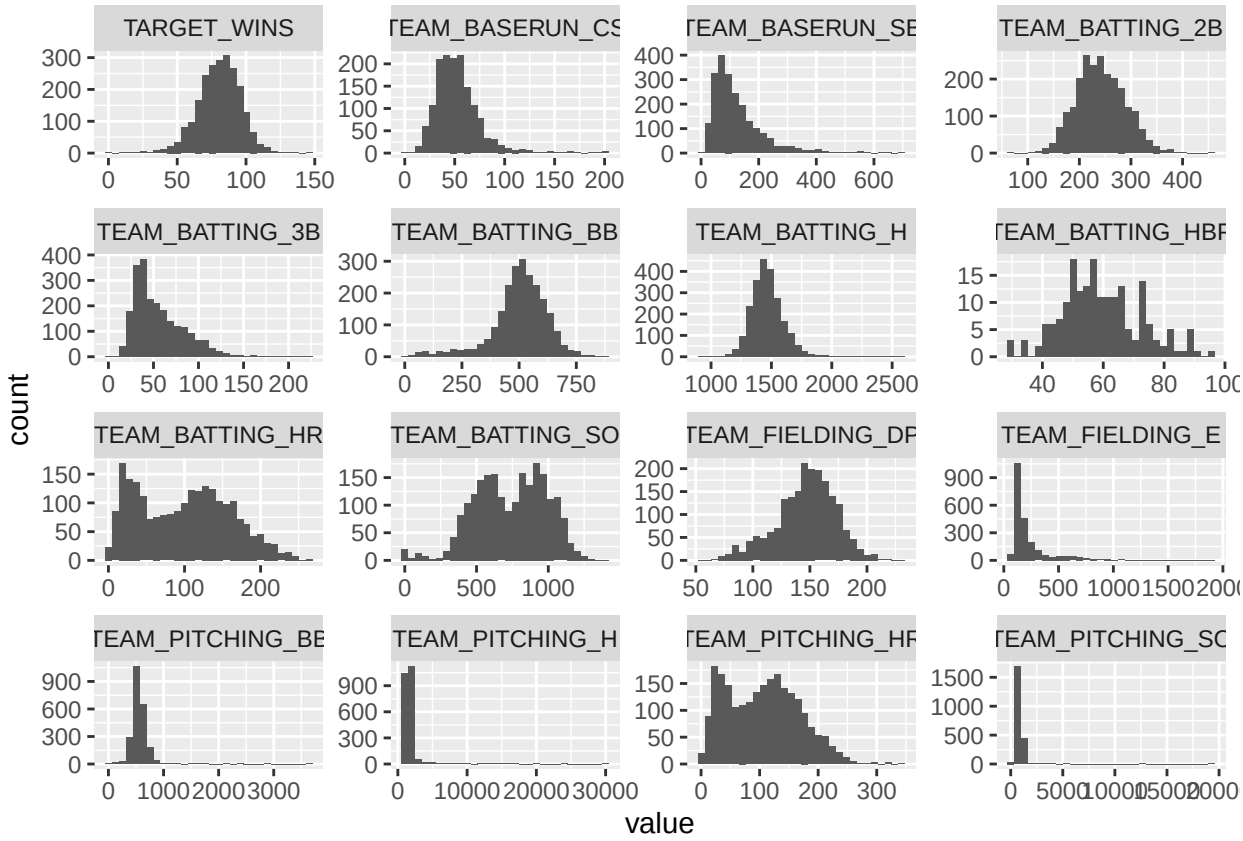


Figure 3: Histograms for all of the variables.

We can see that some of the variables are skewed to the right or the left like `TEAM_PITCHING_SO`. Some of them even have more than one spike (bimodal) like `TEAM_PITCHING_H`. We will also handle these individually in the Data Preparation portion.

While some columns exhibit these abnormalities, it is worth noting that the majority of features will not need to be addressed with transformations. As mentioned before, our target feature is very well normally distributed.

Initial Correlation

This is an initial exploration of how the variables correlate with wins. In the chart below we can see that some of these variables correlate as we would expect with the number of wins - such as `TEAM_BATTING` correlating positively with wins. However some of them did not make sense- like `TEAM_PITCHING_SO` having a negative correlation with wins. We made this chart to get a general idea of how each variable related to the number of wins.

In this initial exploration it is clear that the outliers in some of the variables are affecting the lines of best fit. When we handle them properly, as well as impute the missing data, these lines will likely change.

```
## Warning: Removed 3478 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 3478 rows containing missing values (geom_point).
```

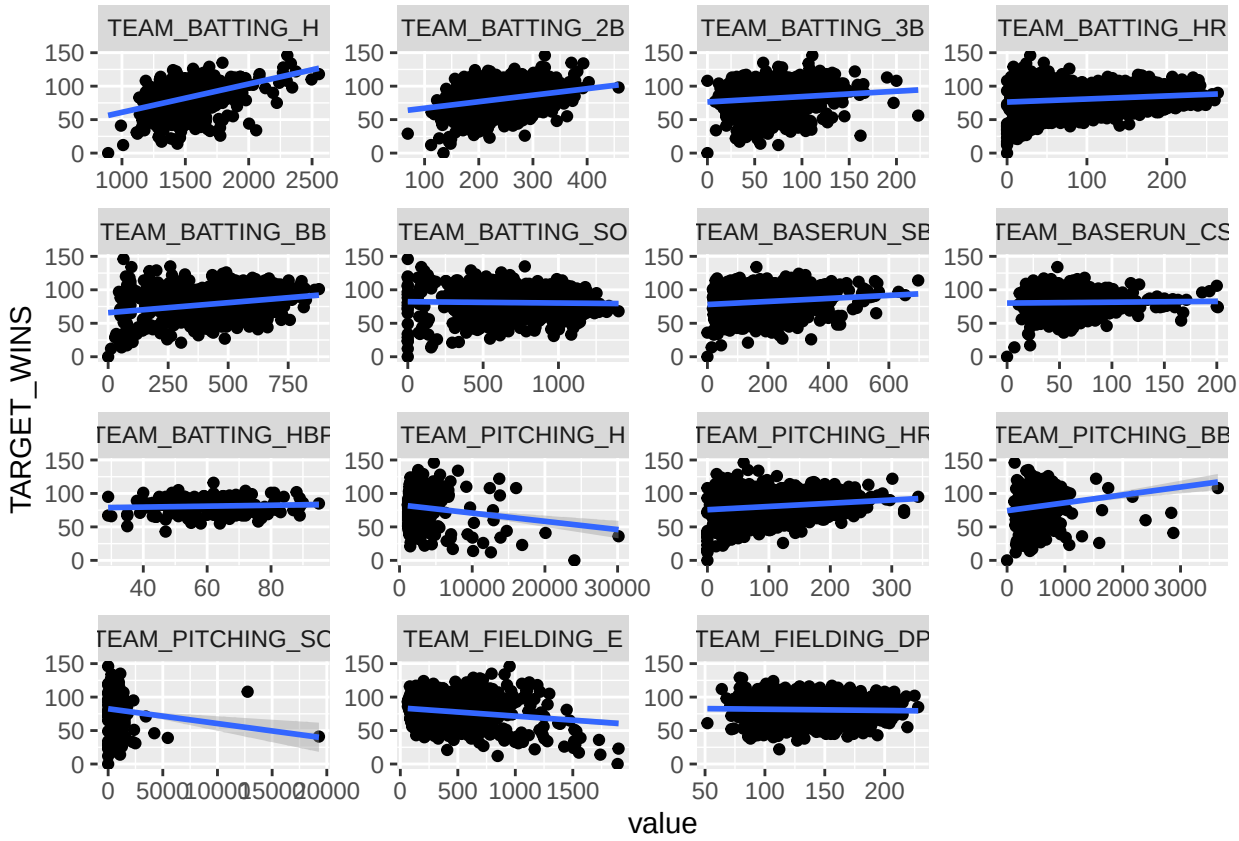


Figure 4: Correlation plot against the target variable for each predictor.

Examining Feature Multicollinearity

Finally, it is imperative to understand which features are correlated with each other in order to address and avoid multicollinearity within our models. By using a correlation plot, we can visualize the relationships between certain features.

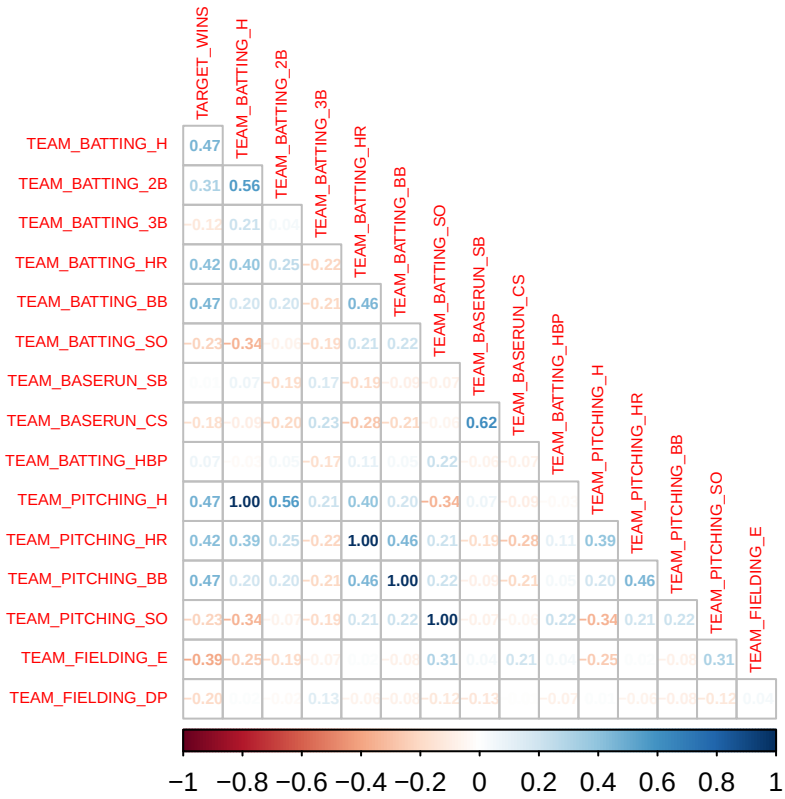


Figure 5: Feature correlation plot.

From the above correlation plot, we notice that there are a few features which exhibit very strong positive correlation. In particular:

- `TEAM_PITCHING_H` & `TEAM_BATTIING_H` == 1.0 correlation
- `TEAM_PITCHING_HR` & `TEAM_BATTING_HR` == 1.0 correlation
- `TEAM_PITCHING_BB` & `TEAM_BATTING_BB` == 1.0 correlation
- `TEAM_PITCHING_SO` & `TEAM_BATTING_SO` == 1.0 correlation

However, we must consider that these initial correlation values could be influenced by the fact that missing values and outliers have yet to be addressed. In later sections we will revisit this chart to determine final correlation values prior to model development.

Data Preperation

Renaming Column Names

Keeping column names short and readable is important in order to practice "table hygiene". Therefore, new column names were generated and are shown on Table XX.

Original Column Name	New Column Name
TARGET_WINS	target
TEAM_BATTING_H	bat_h
TEAM_BATTING_2B	bat_2b
TEAM_BATTING_3B	bat_3b
TEAM_BATTING_HR	bat_hr
TEAM_BATTING_BB	bat_bb

Original Column Name	New Column Name
TEAM_BATTING_HBP	bat_hbp
TEAM_BATTING_SO	bat_so
TEAM_BASERUN_CS	bas_cs
TEAM_FIELDING_E	f_e
TEAM_FIELDING_DP	f_dp
TEAM_PITCHING_BB	p_bb
TEAM_PITCHING_H	p_h
TEAM_PITCHING_HR	p_hr
TEAM_PITCHING_SO	p_so

Table 1: Renamed columns

Dealing with Missing Values

As shown in section 1, there are 6 features that have missing values:

- Strikeouts by batters (5%): Should use median or regression model for imputation
- Stolen bases (6%): Stolen bases weren't tracked officially until 1887, which means some of the missing data could be from 1871-1886. These values could be imputed.
- Caught stealing (34%): Stolen bases weren't tracked officially until 1887, so some of the missing data could be from 1871-1886. These values could be imputed.
- Batter hit by pitch (92%): This predictor will be removed from the analysis as too many of its values are missing.
- Strikeouts by pitchers (4%): Should use median or regression model for imputation
- Double plays (12%): Should use median or regression model for imputation

Tabachnick and Fidell

In general, imputations by the means/medians is acceptable if the missing values only account for 5% of the sample. Peng et al.(2006) However, should the degree of missing values exceed 20% then using these simple imputation approaches will result in an artificial reduction in variability due to the fact that values are being imputed at the center of the variable's distribution.

Our team decided to employ another technique to handle the missing values: Multiple Regression Imputation using the MICE package.

The MICE package in R implements a methodology where each incomplete variable is imputed by a separate model. Alice points out that plausible values are drawn from a distribution specifically designed for each missing datapoint. Many imputation methods can be used within the package. The one that was selected for the data being analyzed in this report is PMM (Predictive Mean Matching), which is used for quantitative data.

Van Buuren explains that PMM works by selecting values from the observed/already existing data that would most likely belong to the variable in the observation with the missing value. The advantage of this is that it selects values that must exist from the observed data, so no negative values will be used to impute missing data.Not only that, it circumvents the shrinking of errors by using multiple regression models. The variability between the different imputed values gives a wider, but more correct standard error. Uncertainty is inherent in imputation which is why having multiple imputed values is important. Not only that. Marshall et al. 2010 points out that:

"Another simulation study that addressed skewed data concluded that predictive mean matching 'may be the preferred approach provided that less than 50% of the cases have missing data. . . '

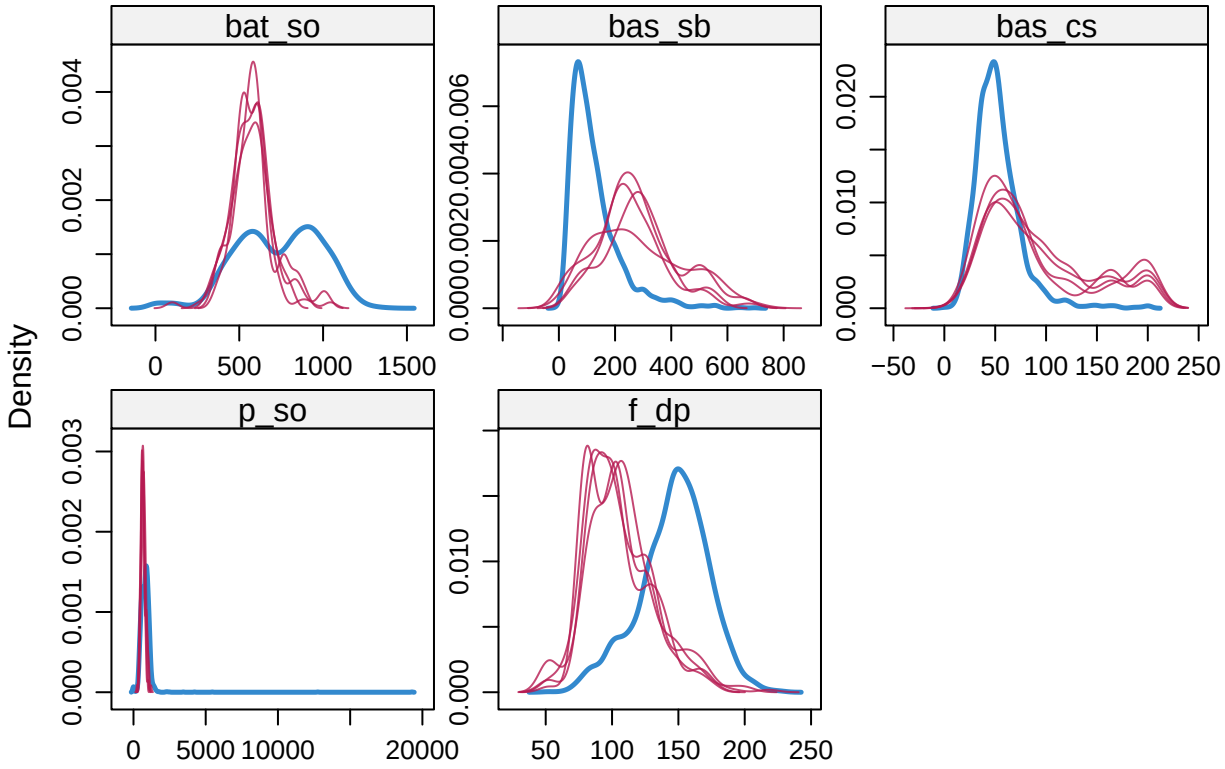


Figure 6: Density plots for variables containing missing data.

Following use of the MICE package, we can visualize the distributions of the imputed versus existing data points as shown on Figure 6. The density of the imputed data for each imputed dataset is shown in magenta. The density of the observed data is shown in blue. For the MICE algorithm, the number of multiple imputations was set to five. The imputed distribution for `bas_sb` and `p_so` look close to the original data distribution which is good. The imputed data distributions for the other variables do not match so closely to the original data. Reasons include:

- Some of the variables are bimodal in nature (which is why in `bas_cs` for example, there is bimodality in the imputed distributions).
- 34% of the data for `bas_cs` is missing, which is above 5%, while the missing data for `p_so` only makes up 4% of the total amount of missing data for that predictor.
- 12% of the data for `f_dp` is missing, which is above 5%, while the missing data for `p_so` only makes up 4% of the total amount of missing data for that predictor.

Analysis of Outliers

Several predictors contained outliers that contradicted with existing baseball statistics or fell out of an "acceptable" range given the feature's inherent distribution. These features are:

- bat_h: The most hits by team in a season is 1783. Therefore, any values above 1,783 were replaced with the median for the predictor (Source).
- p_h: We could not find any suitable statistics from outside sources for this feature. However, we can apply interquartile outlier analysis. By analyzing a given feature, those datapoints which fall above or below an "acceptable" range can be identified given the features inherent distribution.

- p_so: The record for most strikeouts in a season is 1595. Anything above this should be removed or imputed (Source).
- f_e: The record for most errors in a season is 886. Anything above this should be removed or imputed (Source).
- p_bb: We could not find any suitable statistics from outside sources for this feature. However, we can apply interquartile outlier analysis. By analyzing a given feature, those datapoints which fall above or below an "acceptable" range can be identified given the features inherent distribution.

After replacing the above outliers, we can visualize the improved distributions by use of a boxplot.

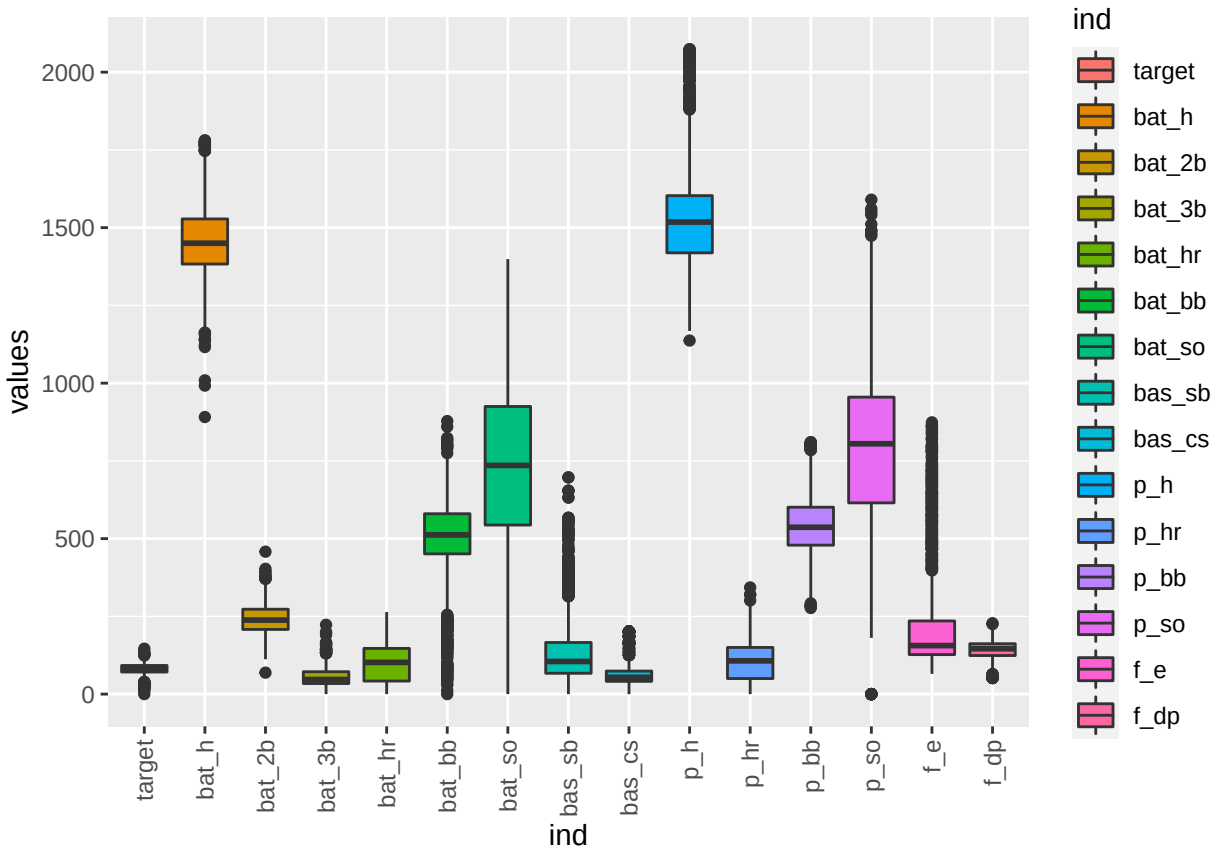


Figure 7: Updated distributions after outlier analysis and imputing NA Values

While there are still outliers present in the dataset, particularly for bas_sb and f_e, we can see a large improvement from before. All features are wihin the range 0-2500. We can attempt to further deal with outliers should the need arise, but for now we will accept this distribution.

Box-Cox Transformation for skewed variables

Based on the previous distribution plot (using histograms) we noticed that a select group of columns exhibited non-normal skew. In particular, the following columns showed signs of left-skew:

- bat_3b
- bas_sb
- bas_cs
- f_e
- p_bb
- p_h

```
## Warning: Removed 903 rows containing non-finite values (stat_bin).
```

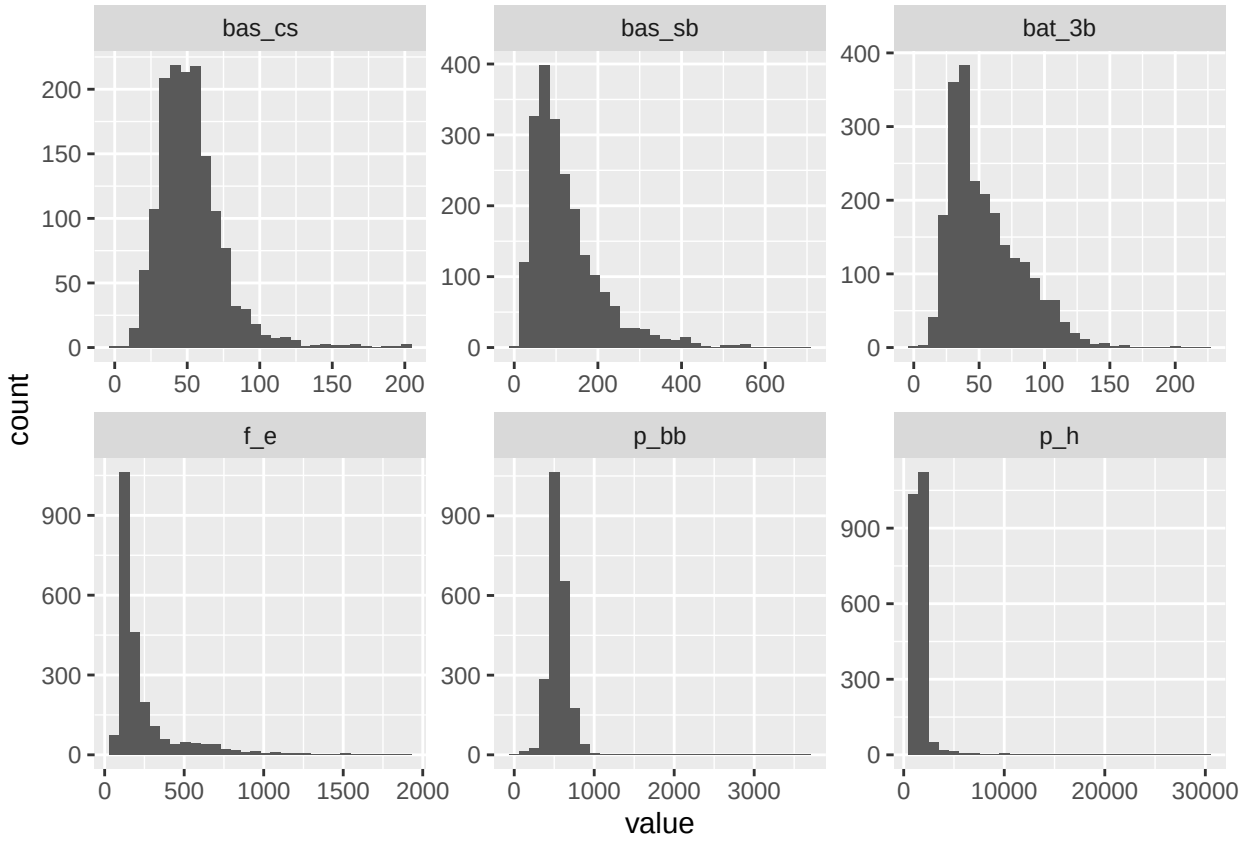


Figure 8: Skewed variables.

In order to address this skewness and attempt to normalize these features for future modeling, we will employ box-cox transformations. Because some of these values include 0, we will need to replace any zero values with infintesimmaly small, non-zero values.

The λ's that were used to transform the skewed variables are shown on Table 2.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

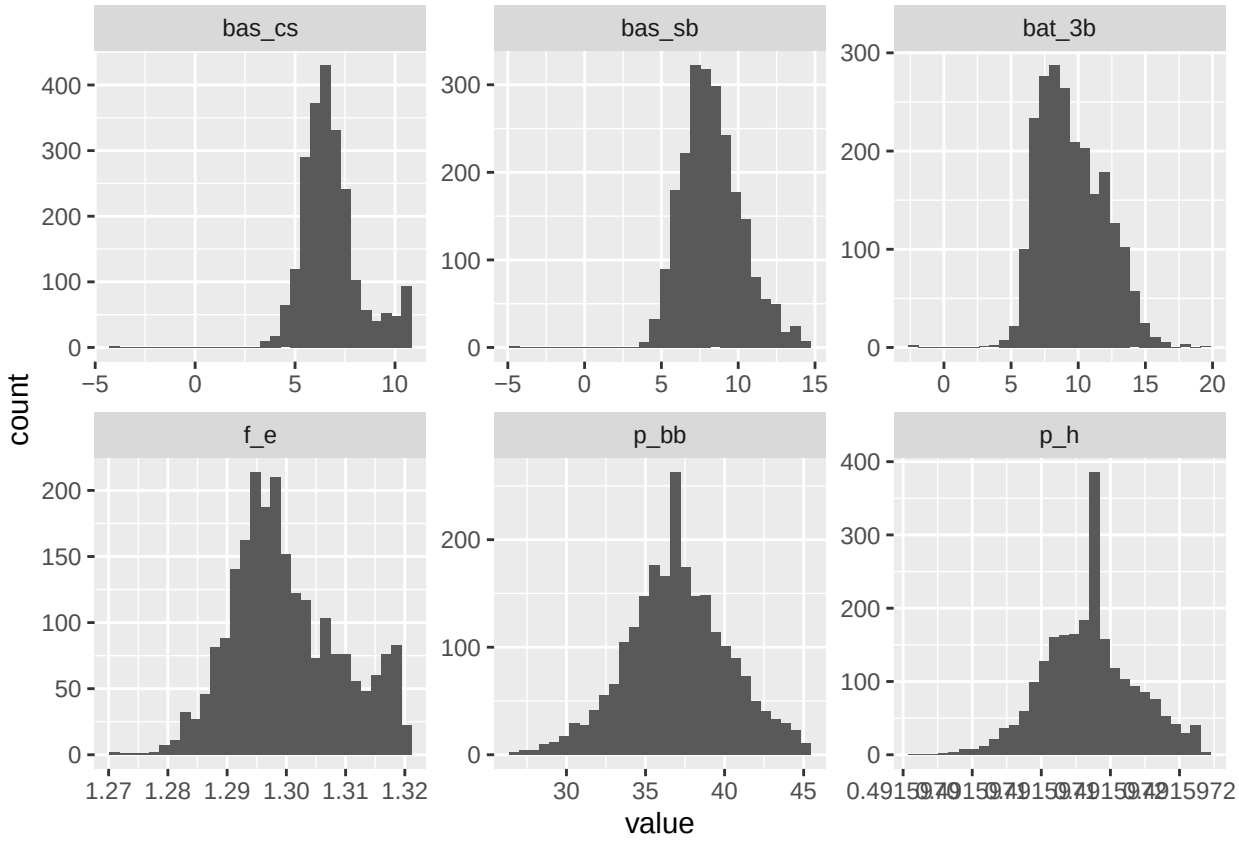


Figure 9: Histograms for transformed variables.

Column Name	λ
bat_3b	0.400
bas_sb	0.220
bas_cs	0.232
f_e	-0.753
p_bb	0.460
p_h	-2.034

Table 2: λ's for skewed variables.

As we can see from the above, the boxcox transformations on the selected features performed extremely well. We can see that all features included now exhibit normal or near-normal distributions around their respective centers.

Dealing with Bimodal Variables

Bimodal distributions in data are interesting, in that they represent features which actually contain multiple (2) inherent systems resulting in separated distributional peaks. Figure XX shows that bimodality is present in `bat_so`, `p_hr`, `bat_hr`. While a Box-Cox transformation could have been undertaken in order to transform the bimodal variables to a normal distribution. However, this throws away important information that is inherent in the bimodal variable itself. The fact that the variable is bimodal in the first place is essentially ignored, and the predicted values in the linear multiple regression model will not reflect this bimodality.

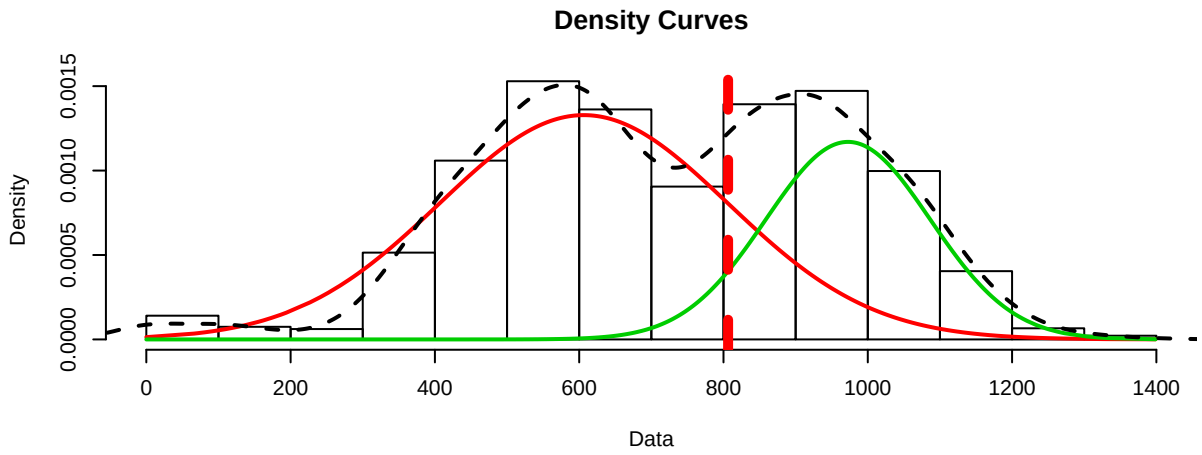
Our approach to solving this is to create dummy variables representing which side of the local minimum each datapoint falls with respect to it's original bimodal distribution. First, two histograms were fit to

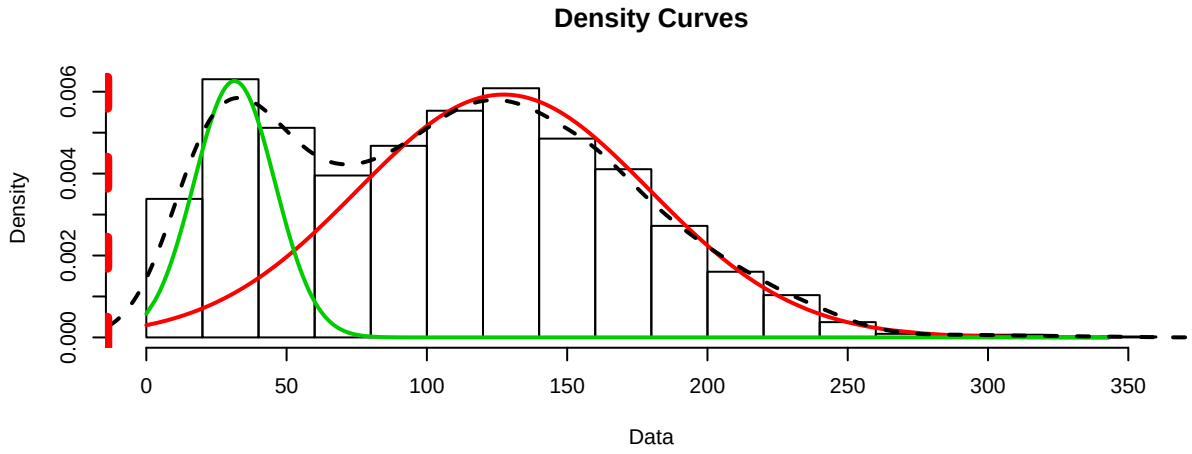
these variables using the `mixtools` package. Then, the intersection point between the two histograms was determined by solving for c. Where

$$c = \frac{\mu_2\sigma_1^2 - \sigma_2(\mu_1\sigma_2 + \sigma_1\sqrt{(\mu_1 - \mu_2)^2 + 2(\sigma_1^2 - \sigma_2^2)log\frac{\sigma_1}{\sigma_2}})}{\sigma_1^2 - \sigma_2^2}$$

Where μ_1 and σ_1 are the mean and standard deviation for the left distribution and ν_2 and σ_2 are the mean and standard deviation for the right distribution.

A new variable was created for each bimodal predictor, where any observed values below c would be assigned a value of 0, while any observed values above c would be assigned a value of 1. For example, c for `bat_so` was calculated to be 806.39. `bi_bat_so` is a new dummy variable that was created where any values above 806.39 in the original `bat_so` data were assigned a value of 0, while values below 806.39 were assigned a value of 1. The λ's for the three bimodal variables are shown in Table XX. The counts for the unique values are shown in each dummy variable are shown on the barcharts on Figure XXX.





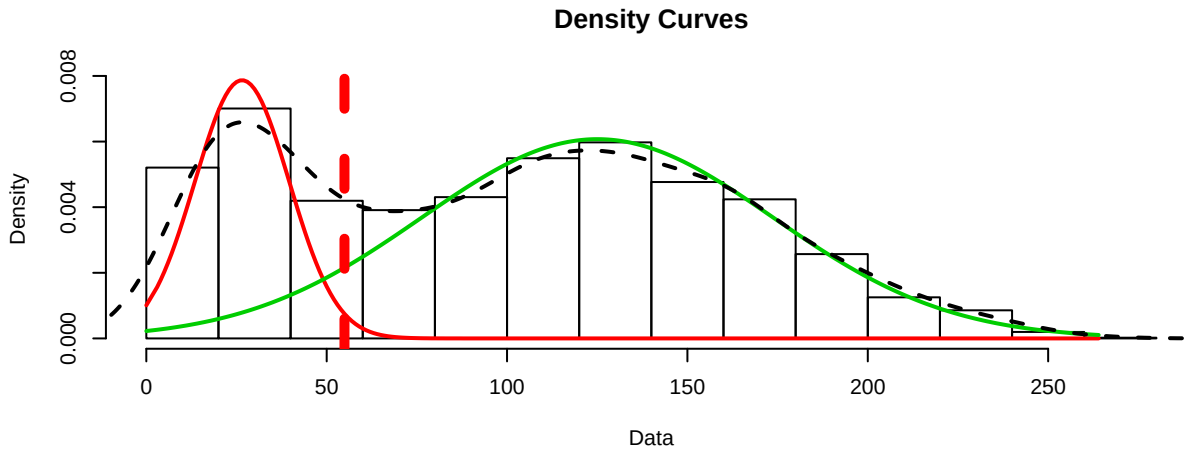


Figure 10: Density curves for each bimodal predictor with two normal distributions fit to each peak.

Column Name	μ_1	μ_2	σ_1	σ_2	c	Count of 0's
bat_so	606.31	972.61	199.88	114.06	806.38	969

Column Name	μ_1	μ_2	σ_1	σ_2	c	Count of 0's
p_hr	31.43	127.37	14.39	52.08	60.93	1602
bat_hr	26.55	125.06	13.10	48.72	54.93	1583

Table 3: Summary of bimodal dummy variable generation

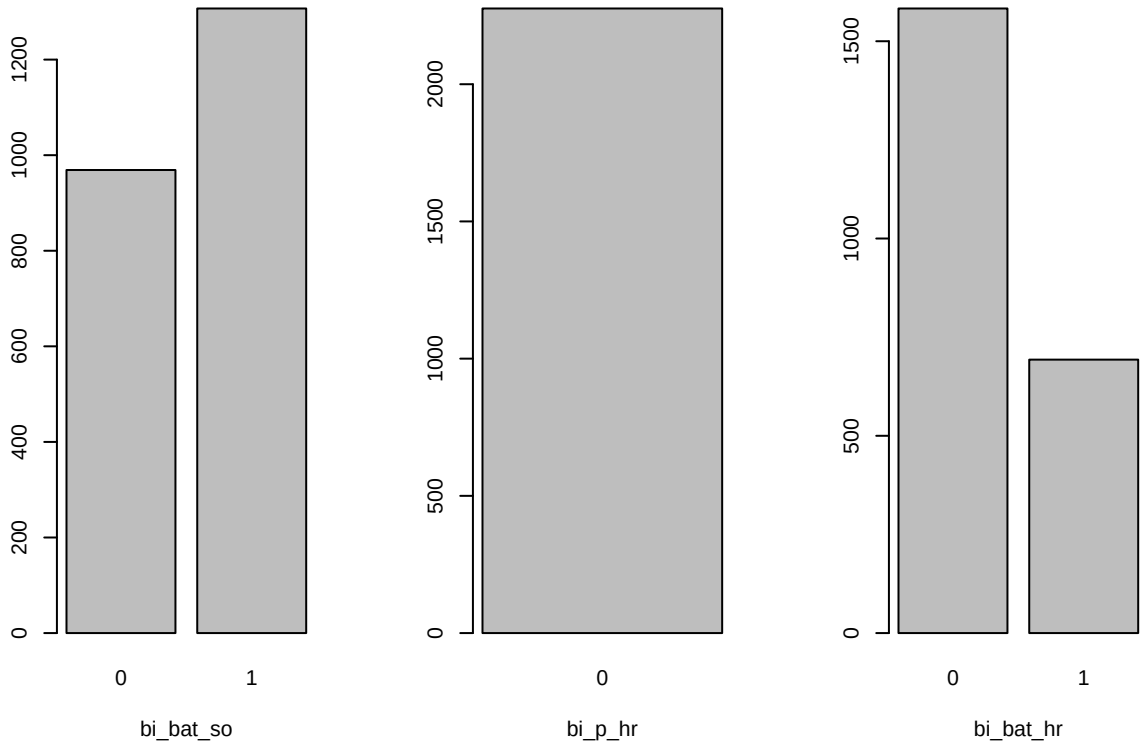


Figure 11: Bar graphs for each of the bimodal dummy variables. 0 *represents the amount of observations for the original variable where the value was above c, while* 1 *represents the amount of observations below c*

Saber Model

Finally, we would like to employ outside analysis in order to engineer new, potentially powerful features. Popularized in the movie "Moneyball", the SABERMETRICS model for baseball analysis includes a feature known as BsR (base runs). This statistic estimates the amount of runs a team should score.

(see http://tangotiger.net/wiki_archive/Base_Runs.html for more information). The formula for constructing this metric is as follows:

$$BSR = AB/(B + A) + C$$

where:

$$A = TEAM\_BATTING\_1B + TEAM\_BATTING\_2B + TEAM\_BATTING\_3B + TEAM\_BATTING\_BB$$

$$B = 1.02(1.4TEAM\_TOTAL\_BASES - 0.6TEAM\_BATTING\_H + 0.1TEAM\_BATTING\_BB)$$

$$C = TEAM\_BATTING\_HR$$

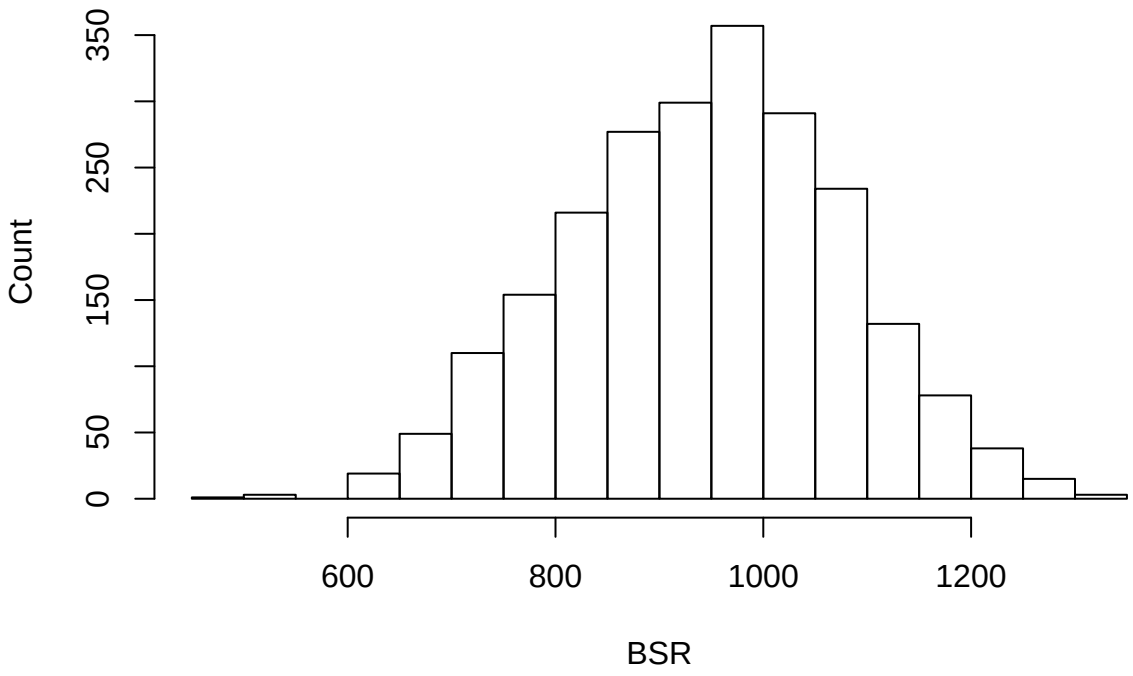


Figure 12: Histogram of BSR Predictor

Reviewing the correlations

After performing multiple cleaning and imputation steps, we would like to visualize again the correlations between features and their target, as well as between features themselves.

```
## Warning in cor(cor_numeric): the standard deviation is zero
```

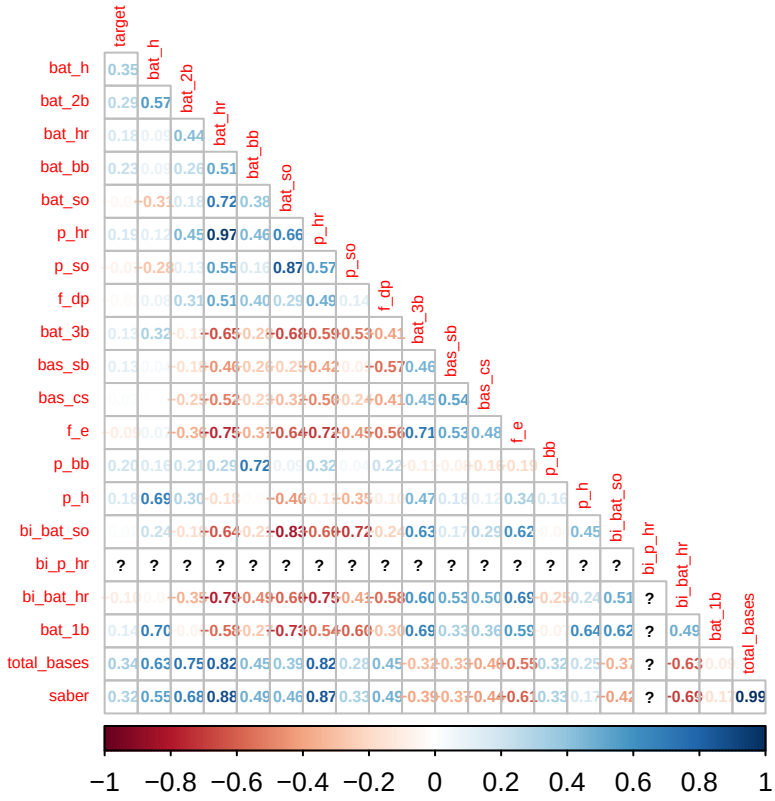


Figure 13: Feature correlation plot after data preparation

These correlation values make much more sense than before. We can see that features no longer have 1.0 correlations, which in general are highly unlikely to occur naturally. The new most correlated (and least correlated) features are as follows:

- p_hr & bat_hr (0.97): This is an interested correlation, as we would not have intitially expected the amount of homeruns allowed to be correlated with the number of homeruns achieved from a team. However, one could make the argument that a team which focuses on offense would similarly be lacking in defense.
- bat_1b & bat_so (-0.73): These features are negatively correlated, which makes intuititve sense. If a team has many players making it to base, then conversely we would expect that this team would have less strikeouts at bat.
- bat_so & p_so (0.87): These features intuitively should not have such high correlation. Similar to above, we would not expect the performance of batter strikeouts to have any relationship to the performance of pitching strikouts on the same team.

```
## Warning in predict.lm(model, newdata = new_data_frame(list(x = xseq)), se.fit =
## se, : prediction from a rank-deficient fit may be misleading
```

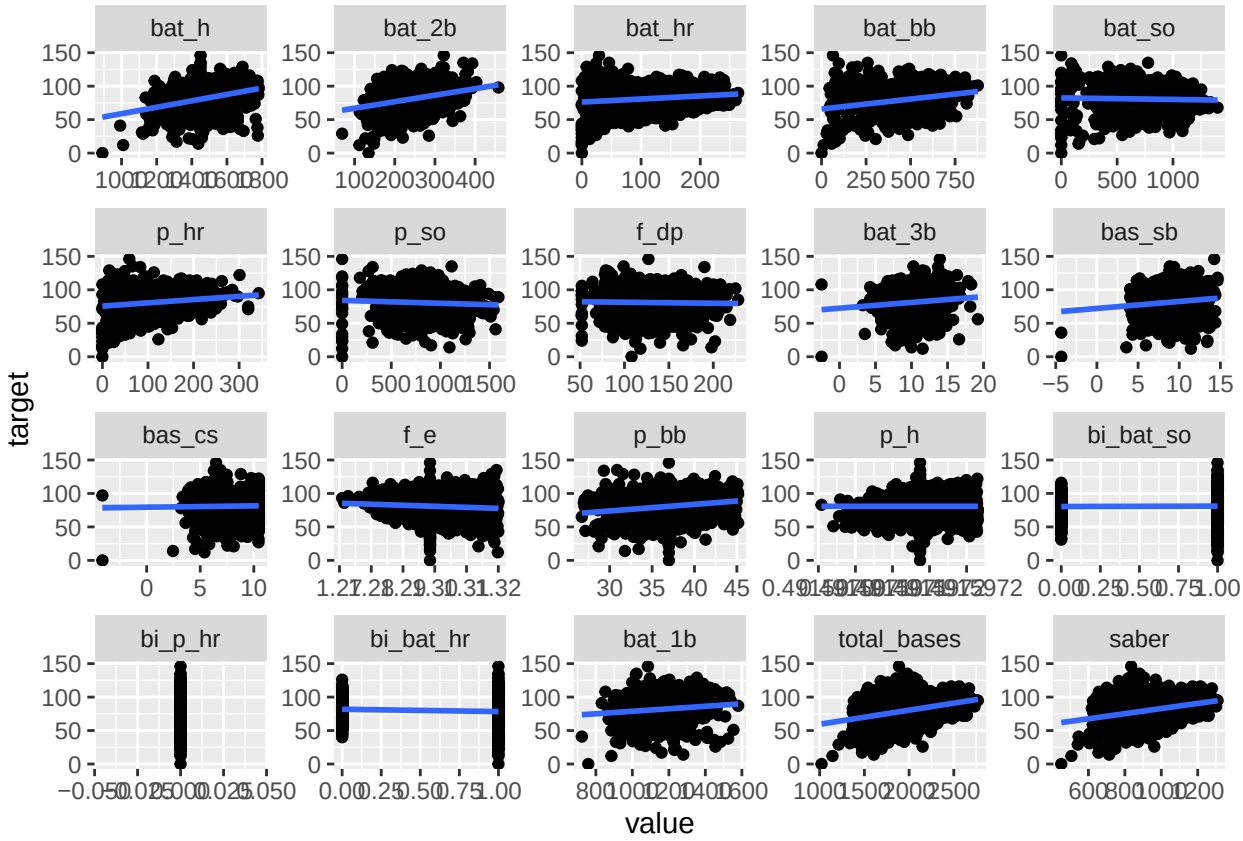


Figure 14: Target correlation plot after cleaning.

After applying all transformations and imputations, we can see that the feature correlation with the target variable has also improved. Features predicted to have positive correlations (as provided by the assignment guide) do tend to have positive correlations. Similarly, features with expected negative correlations behave as described. This provides us some level of validation as we take the next steps with model building.

Build Models

Examine base model, no transformations, no engineering

Our first model (Base model) will use all of the initially provided columns, after cleaning and imputation. We will use the results of this model to understand a baseline for our future model development.

```
##
## Call:
## lm(formula = target ~ ., data = base)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -70.757  -8.133   0.296   8.305  76.441
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.925e+02  7.438e+01   9.311  < 2e-16 ***
## bat_h        1.826e-02  3.888e-03   4.696 2.82e-06 ***
## bat_2b       2.799e-02  8.674e-03   3.227 0.001270 **
## bat_hr       7.147e-02  2.849e-02   2.508 0.012211 *
```

```
## bat_bb       3.987e-02  4.593e-03   8.681  < 2e-16 ***
## bat_so      -1.911e-02  4.195e-03  -4.557 5.48e-06 ***
## p_h                 NA         NA      NA       NA
## p_hr         1.550e-02  2.456e-02   0.631 0.528020
## p_bb        -5.142e-01  1.492e-01  -3.446 0.000579 ***
## p_so         1.494e-03  3.230e-03   0.462 0.643773
## f_dp        -1.033e-01  1.323e-02  -7.808 8.78e-15 ***
## bat_3b       1.422e+00  2.047e-01   6.945 4.93e-12 ***
## bas_sb       1.848e+00  2.048e-01   9.027  < 2e-16 ***
## bas_cs       1.393e-01  2.522e-01   0.552 0.580954
## f_e         -5.056e+02  5.677e+01  -8.906  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.38 on 2262 degrees of freedom
## Multiple R-squared:  0.2827, Adjusted R-squared:  0.2786
## F-statistic: 68.58 on 13 and 2262 DF,  p-value: < 2.2e-16
```

Based on the above output, we can see that this model performs relateively poorly against the training data. However, as this is our base model, we will assess the performance of all future models against this value. Moving forward, if we can lift the Adjusted r^2 to above 0.3, we will consider it a general improvement.

Evaluate SABER model

The next model we would like to evaluate is the SABER model. Here we will use all original features, and additionally we will include the engineered SABER metrics. Hopefully we will see a lift in performance after utilizing these industry-derived features.

```
##
## Call:
## lm(formula = target ~ ., data = complete_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -69.471  -8.181   0.385   8.274  75.444
##
## Coefficients: (4 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.029e+02  7.506e+01   9.365  < 2e-16 ***
## bat_h       -2.447e-01  4.883e-02  -5.012 5.82e-07 ***
## bat_2b      -1.192e-01  2.865e-02  -4.159 3.31e-05 ***
## bat_hr      -8.353e-01  1.701e-01  -4.912 9.67e-07 ***
## bat_bb       2.190e-02  5.660e-03   3.869 0.000113 ***
## bat_so      -2.251e-02  4.672e-03  -4.818 1.55e-06 ***
## p_hr        -1.240e-02  2.498e-02  -0.497 0.619539
## p_so         4.658e-03  3.286e-03   1.418 0.156404
## f_dp        -1.164e-01  1.346e-02  -8.648  < 2e-16 ***
## bat_3b       1.017e+00  2.194e-01   4.637 3.73e-06 ***
## bas_sb       1.951e+00  2.080e-01   9.380  < 2e-16 ***
## bas_cs       2.414e-01  2.520e-01   0.958 0.338136
## f_e         -5.180e+02  5.766e+01  -8.984  < 2e-16 ***
## p_bb        -3.387e-01  1.522e-01  -2.226 0.026093 *
## p_h                 NA         NA      NA       NA
## bi_bat_so    3.147e-01  1.103e+00   0.285 0.775411
## bi_p_hr             NA         NA      NA       NA
```

```
## bi_bat_hr    3.209e+00  1.407e+00   2.281 0.022668 *
## bat_1b              NA         NA      NA       NA
## total_bases         NA         NA      NA       NA
## saber        5.535e-01  1.026e-01   5.396 7.54e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.3 on 2259 degrees of freedom
## Multiple R-squared:  0.2925, Adjusted R-squared:  0.2875
## F-statistic: 58.36 on 16 and 2259 DF,  p-value: < 2.2e-16
```

As expected, we did see a lift in perfomance after including SABER metrics. However, the lift was hardly significant. We are still below 0.3 Adjusted R^2.

SABER reduced

Here we will test out a more parsimonious version of the above SABER model. In the spirit of simplifying the model for human use and understanding, we will select only the features that have high significance from the above SABER model. Additionally, we will exlude any features which were included as part of the construction of SABER, in order to reduce inherent multicollinearity.

```
##
## Call:
## lm(formula = target ~ ., data = sab_reduced)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.289  -8.323   0.419   8.599  66.396
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.805e+02  6.843e+01   7.022 2.88e-12 ***
## saber        5.223e-02  3.177e-03  16.443  < 2e-16 ***
## bi_bat_hr    3.002e-02  1.087e+00   0.028    0.978
## f_e         -3.466e+02  5.174e+01  -6.699 2.64e-11 ***
## bas_sb       2.130e+00  1.911e-01  11.142  < 2e-16 ***
## f_dp        -1.041e-01  1.354e-02  -7.688 2.21e-14 ***
## bat_so      -2.097e-02  1.632e-03 -12.855  < 2e-16 ***
## bat_bb       2.767e-02  2.809e-03   9.850  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.56 on 2268 degrees of freedom
## Multiple R-squared:  0.2611, Adjusted R-squared:  0.2589
## F-statistic: 114.5 on 7 and 2268 DF,  p-value: < 2.2e-16
```

While the Adjusted R^2 has been slightly reduced to 0.26, we have also significantly reduced the complexity of the model. This provides value in itself, as the model can be more easily distributed to players and coaches.

Step AIC

Step AIC works by deselecting features that negatively affect the AIC, which is a criterion similar to the R-squared. It selects the model with not only the best AIC score but also a model with less predictors than the full model, since the full model may have predictors that do not contribute or negatively contribute to the model's performance. The direction for the Step AIC algorithm was set to `both`, because this implements both forward and backward elimination in order to decide if a predictor negatively affects the model's performance.

```
##
## Call:
## lm(formula = target ~ bat_h + bat_2b + bat_hr + bat_bb + bat_so +
##     p_bb + f_dp + bat_3b + bas_sb + f_e, data = base)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.990  -8.093   0.307   8.321  75.870
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.868e+02  7.298e+01   9.410  < 2e-16 ***
## bat_h        1.818e-02  3.829e-03   4.747 2.19e-06 ***
## bat_2b       2.846e-02  8.567e-03   3.322 0.000909 ***
## bat_hr       8.689e-02  9.999e-03   8.689  < 2e-16 ***
## bat_bb       3.778e-02  3.987e-03   9.476  < 2e-16 ***
## bat_so      -1.755e-02  2.213e-03  -7.927 3.49e-15 ***
## p_bb        -4.578e-01  1.358e-01  -3.370 0.000764 ***
## f_dp        -1.029e-01  1.319e-02  -7.800 9.34e-15 ***
## bat_3b       1.463e+00  2.004e-01   7.302 3.91e-13 ***
## bas_sb       1.897e+00  1.933e-01   9.813  < 2e-16 ***
## f_e         -5.018e+02  5.581e+01  -8.991  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.37 on 2265 degrees of freedom
## Multiple R-squared:  0.2822, Adjusted R-squared:  0.2791
## F-statistic: 89.06 on 10 and 2265 DF,  p-value: < 2.2e-16
```

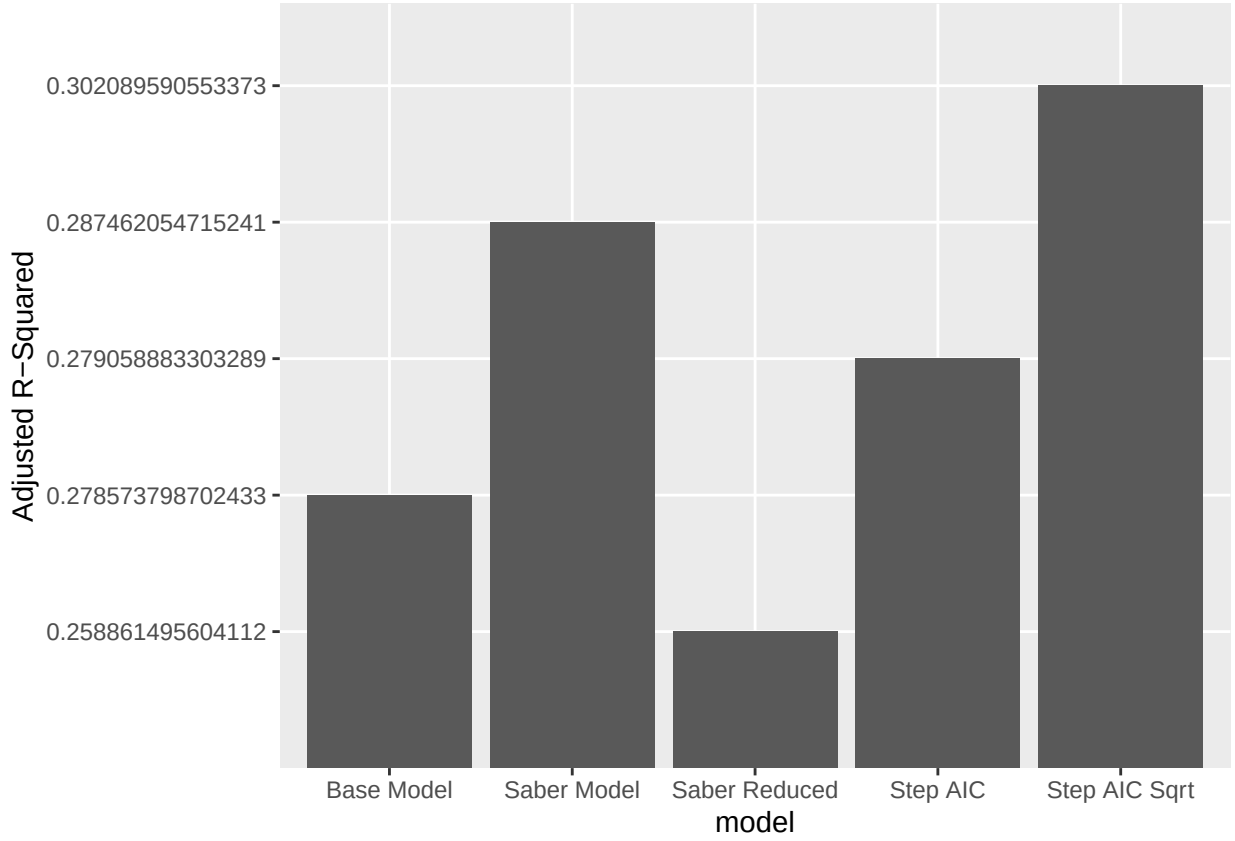
Square Root Step AIC

The following model was generated using the same AIC methodology, except that the `target` variable was square rooted.

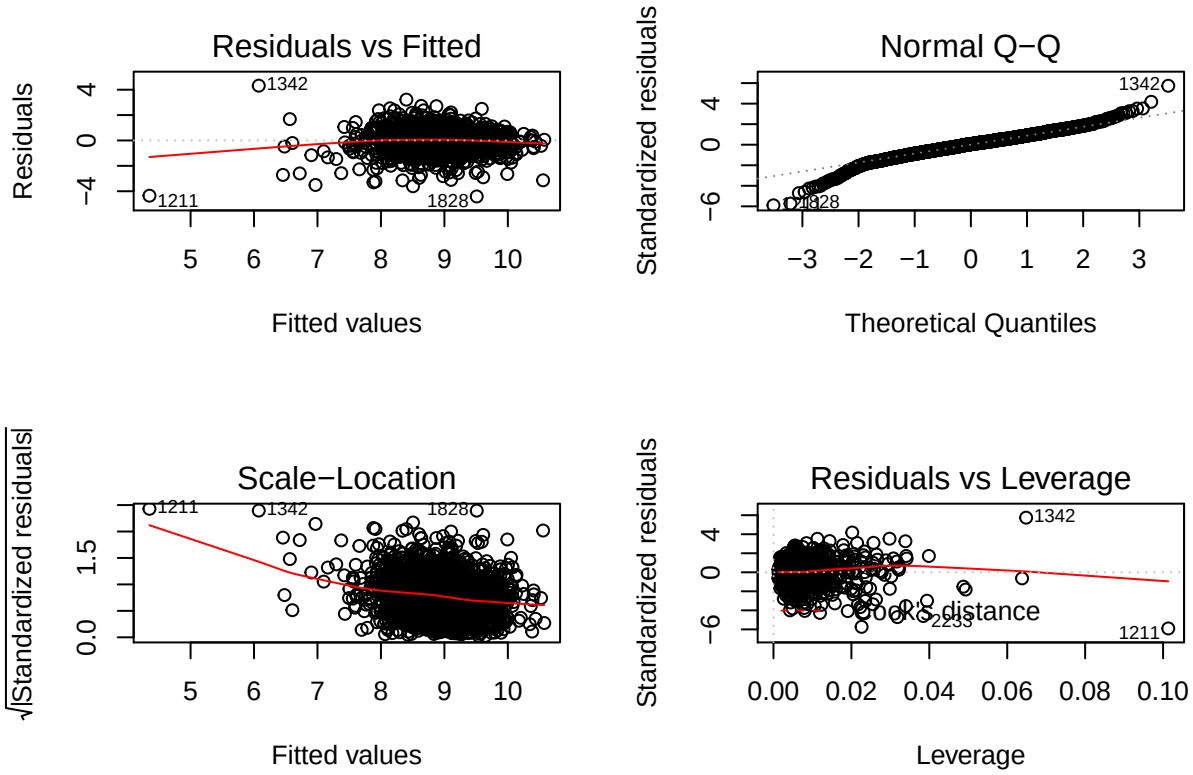
```
##
## Call:
## lm(formula = sqrt(target) ~ bat_h + bat_2b + bat_hr + bat_bb +
##     bat_so + f_dp + bat_3b + bas_sb + bas_cs + f_e + p_bb + bi_bat_hr +
##     saber, data = complete_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4104 -0.4493  0.0444  0.4612  4.3176
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.320e+01  4.261e+00  10.139  < 2e-16 ***
## bat_h       -1.490e-02  2.789e-03  -5.344 1.00e-07 ***
## bat_2b      -7.579e-03  1.654e-03  -4.584 4.82e-06 ***
## bat_hr      -5.282e-02  9.910e-03  -5.331 1.08e-07 ***
## bat_bb       1.496e-03  2.962e-04   5.051 4.74e-07 ***
## bat_so      -8.563e-04  1.292e-04  -6.630 4.19e-11 ***
## f_dp        -6.502e-03  7.865e-04  -8.267 2.32e-16 ***
## bat_3b       6.274e-02  1.269e-02   4.945 8.18e-07 ***
## bas_sb       1.159e-01  1.200e-02   9.664  < 2e-16 ***
```

```
## bas_cs       2.759e-02  1.467e-02   1.880 0.060191 .
## f_e         -2.895e+01  3.261e+00  -8.877  < 2e-16 ***
## p_bb        -2.724e-02  8.113e-03  -3.358 0.000799 ***
## bi_bat_hr    1.745e-01  8.219e-02   2.123 0.033867 *
## saber        3.404e-02  5.874e-03   5.796 7.75e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7779 on 2262 degrees of freedom
## Multiple R-squared:  0.3061, Adjusted R-squared:  0.3021
## F-statistic: 76.75 on 13 and 2262 DF,  p-value: < 2.2e-16
```

Model Selection



The model that is ultimately chosen for this analysis is Step AIC Square Root. We were able to increase over the base model by 3%. AIC is a measure of multicollinearity so the selection process parsed out variables that were highly colinear with other variables, giving us a model that has the lowest AIC values based on a select number of predictors. This is important because this model needs to be used and understood by professionals in the industry; the step AIC model ensures that only the most prominent features are included.



The QQ plot shows that the data is centered in the middle, but there is significant amount of residuals in the middle of the distribution. This is known as the "thin tail" phenomenon. Normal distributions with "thin tails" correspond to the first quantiles occurring at larger than expected values and the last quantiles occurring at less than expected values. Notice that the "thin tailed" Q-Q plot is a reflection of a "fat tailed" Q-Q plot, which is the opposite phenomenon, across the X-Y diagonal. The Residuals vs. Fitted, and scale-location plots show that the residuals are mostly centered around the zero line. However there is some skewness as a result of outliers that are marked numerically on the plot itself. The Residuals vs. Leverage plot shows that there are no residual values that exceed Cook's distance, which is good.

Important Metrics for Step AIC Square Root Model Important metrics for the Step AIC Square Root model are:

- R-squared: 0.3027
- F-statistic: 71.54
- RSS: 1366.88
- MSE: 0.6
- RMSE: 0.774

Note #### Predictions on Evaluation Set

The predictions were generated using the evaluation set on the Square Root Step AIC model. These predictions are provided in the `predictions.csv` file.