Moneyball

ADEC 7320.02

Silas Selfe

11/9/2021

Data Exploration

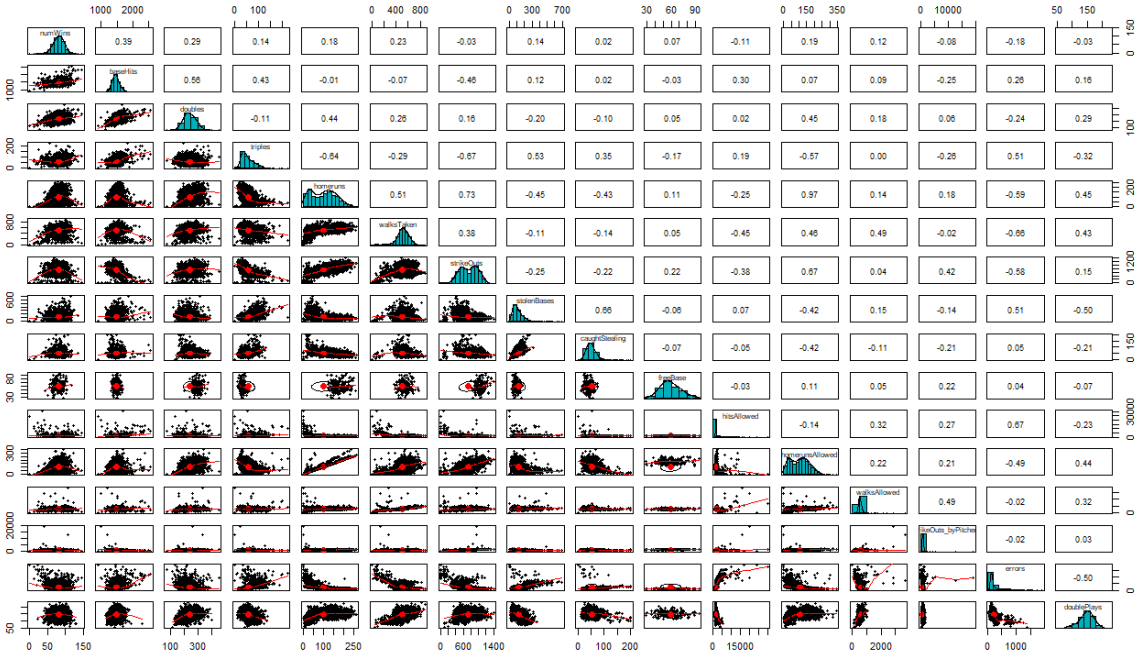
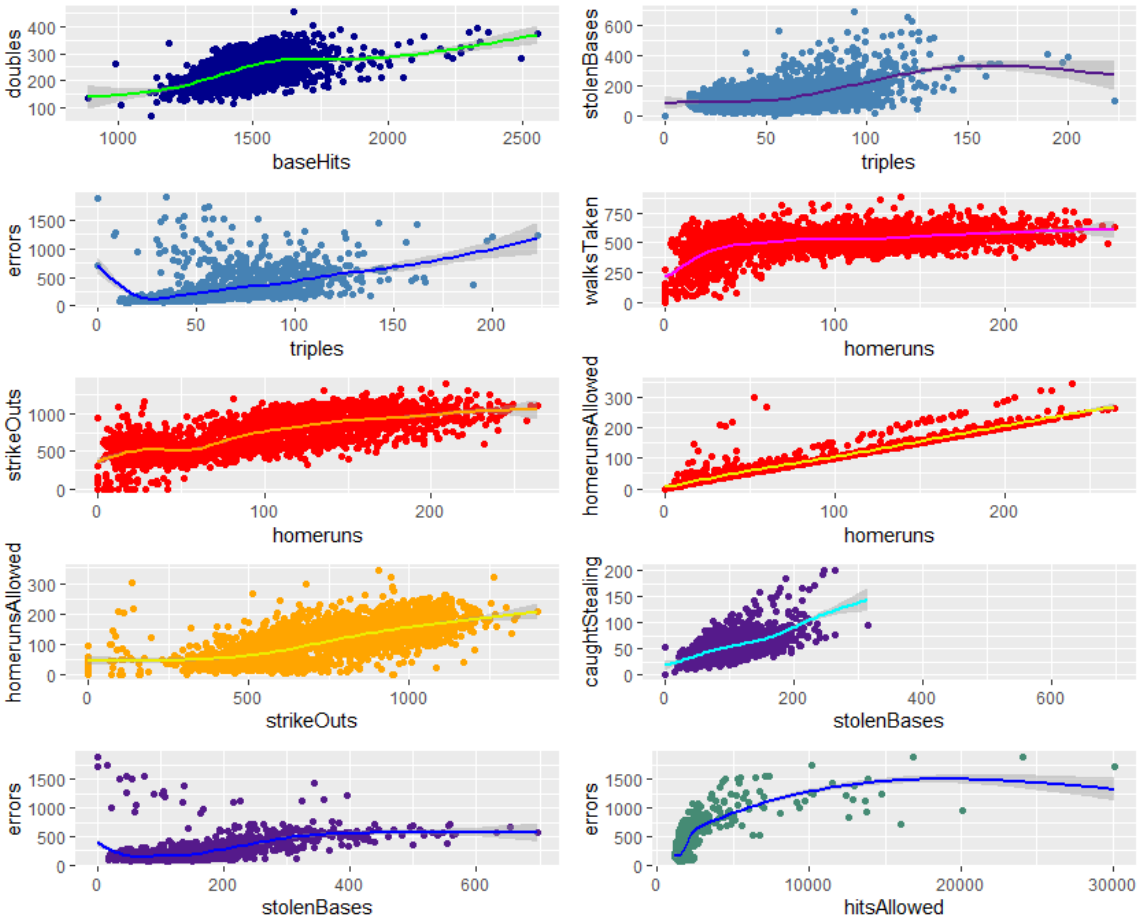


Table 1: Variables with Positive Correlation (> 0.50)

Compared Variable	V1	V2	V3
baseHits	doubles	NA	NA
triples	stolenBases	errors	NA
homeruns	walksTaken	strikeOuts	homerunsAllowed
strikeOuts	homerunsAllowed	NA	NA
stolenBases	caughtStealing	errors	NA
hitsAllowed	errors	NA	NA

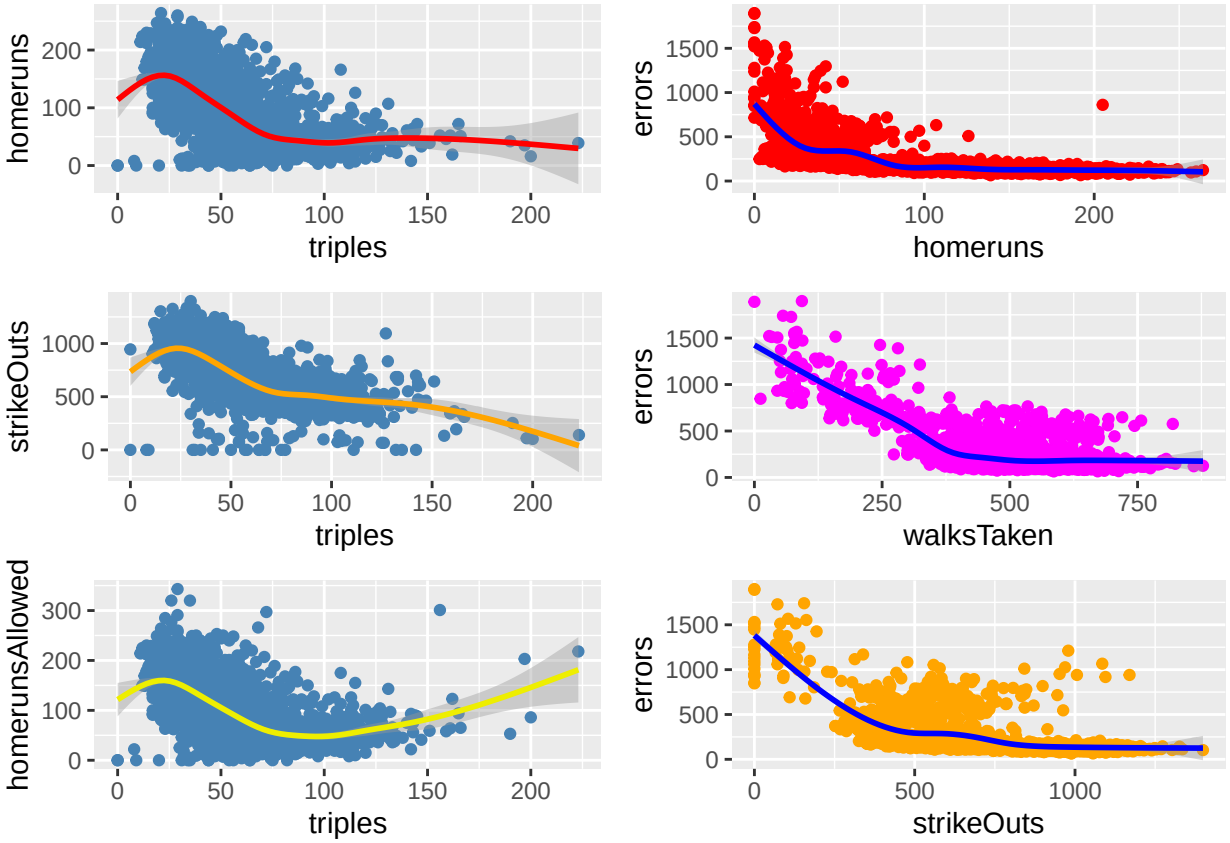


The variables with positive correlation greater than 0.5 seem to follow their general distribution patterns seen in the variable plots on page 2. The variable that most apparently disrupts these patterns is errors, which is seen immediately by the outliers it produces in plots [1,3] and [1,4].

While some plots are somewhat frivolous, the ones that are not raise interesting questions. [1,3] Does having a batter on third result in a pressured outfield making more errors? [1,2] If a pitcher has a faster pitch, does this result in more Strike Outs? If so, when batters hit a faster pitch, does it translate into more Home Runs? And most important is the question being answered, how can this data predict the number of wins for a given team?

Table 2: Variables with Negative Correlation (< 0.50)

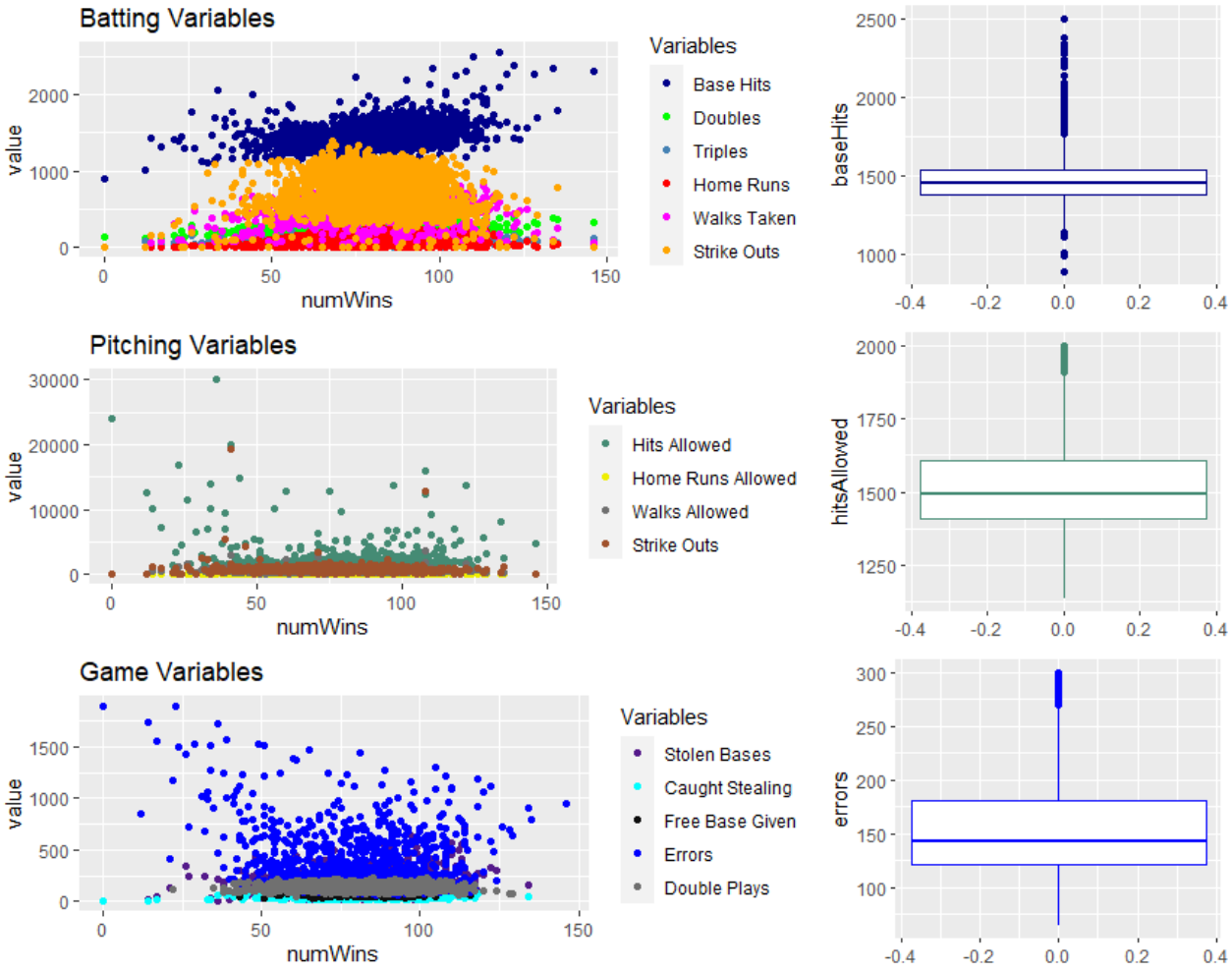
Compared Variable	V1	V2	v3
triples	homeruns	strikeOuts	homerunsAllowed
homeruns	errors	NA	NA
walksTaken	errors	NA	NA
strikeOuts	errors	NA	NA



What is interesting in these negatively correlated variables is that "triples" is correlated with the only three variables that take a binomial distribution. It could follow a similar sort of logic discussed on page 3. Plot [2,1] shows that there are fewer triples when there are more strikeouts, so a fast throwing pitcher is bound to get more strike outs, but batters who make contact with this pitch are more likely to hit deep into the outfield. This hypotheses seems to hold weight by looking at [1,1] where a similar correlation is seen which suggests that a faster pitch results in more strikeouts, while also allowing further hits, which result in homeruns and triples.

The "errors" variable also negatively contributes to the variables seen on the right. It is possible that in [1,2], if there are fewer homeruns, there is a slower pitcher that allows more batters to hit into the infield, creating opportunity for errors to arise. In the same spirit, in [2,2], if more batters are walked, the ball is not frequently in play which reduces the opportunity for errors. The same logic is followed with [3,2] as a strikeout almost entirely removes the opportunity for error.

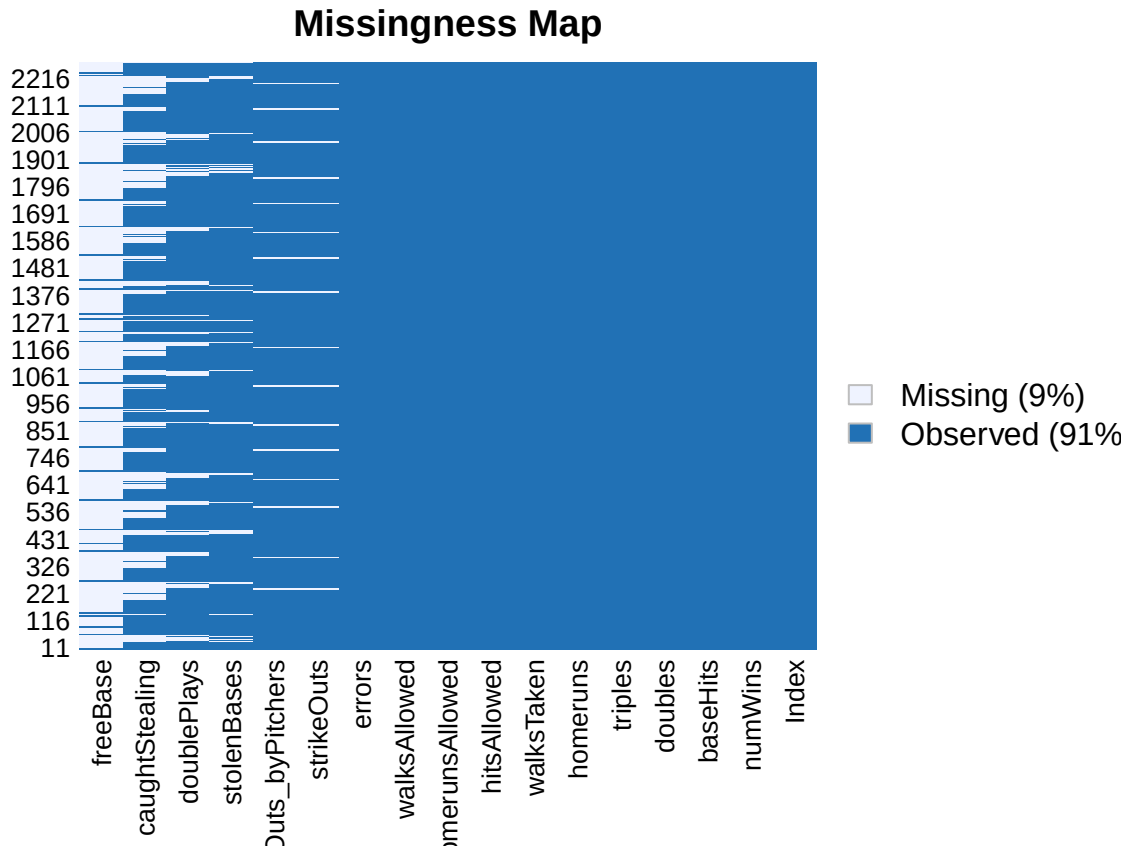
Batting v Pitching v Game Variables




```
##     baseHits     hitsAllowed        errors
##  Min.   : 891   Min.   : 1137   Min.   :  65.0
##  1st Qu.:1383   1st Qu.: 1419   1st Qu.: 127.0
##  Median :1454   Median : 1518   Median : 159.0
##  Mean   :1469   Mean   : 1779   Mean   : 246.5
##  3rd Qu.:1537   3rd Qu.: 1682   3rd Qu.: 249.2
##  Max.   :2554   Max.   :30132   Max.   :1898.0
```

These comparisons illustrate what appear to be key determinants in the number of games a team wins. It should come as no surprise that winning and losing seemingly come down to a teams ability to hit the ball and how well they are able to prevent their opponent from doing so.

Data Preparation



Since the variable "***freebase***" shows no significant correlation with other variables, it will be removed from the modeling dataset. The missing values for the variables "***double plays***" and "***stolen bases***" are replaced by their corresponding mean value since they follow a normal distribution and similarly, for "***strikeouts***" and "***strikeouts by pitcher***" they're replaced by their median values as their distributions aren't neatly normal.

For the "***caught stealing***" variable, its missing values–determined by a dataset with all NAs removed–are replaced by the observed value of "***stolen bases***" multiplied by 0.6115212, which was found to be the mean proportion of getting caught stealing.

New variables were also created. Among them are the probability of a hit being a home run, the proportion in which a team hits compared to the amount of hits they allow, the proportion of hits per error, an estimated amount of times a team would change from batting to fielding in a season, etc. etc.

Much time was spent here. An attempt to calculate the On Base Percentage was made, as well as the Slugging Percentage. The variables were manipulated heavily in an attempt to minimize the squared error and tighten the models fit. The most useful variable that was created was the "***totHits***" variable that simply summed up all hits for a team. As is seen in the models below, not many created variables contributed in a significant way to any of the models.

Models

Model 1

```
##
## Call:
## lm(formula = numWins ~ totHits + strikeOuts + hitsAllowed + hit_prob +
##     errors + error_prob + stolenBases + homer_prob, data = mdlData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.923  -8.882   0.314   8.541  57.684
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.853e+01  5.487e+00  10.666  < 2e-16 ***
## totHits      3.025e-02  1.994e-03  15.173  < 2e-16 ***
## strikeOuts  -8.496e-03  2.079e-03  -4.086 4.54e-05 ***
## hitsAllowed -2.382e-03  4.979e-04  -4.785 1.82e-06 ***
## hit_prob    -2.027e+01  3.716e+00  -5.454 5.47e-08 ***
## errors      -1.954e-02  5.870e-03  -3.329 0.000885 ***
## error_prob  -3.758e+01  1.238e+01  -3.037 0.002419 **
## stolenBases  5.436e-02  4.256e-03  12.771  < 2e-16 ***
## homer_prob   5.501e+01  1.898e+01   2.899 0.003785 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.1 on 2246 degrees of freedom
## Multiple R-squared:  0.2737, Adjusted R-squared:  0.2712
## F-statistic: 105.8 on 8 and 2246 DF,  p-value: < 2.2e-16
```

The variables in this model were chosen with the intent of keeping each variably significant to its prediction. The coefficients of these variables all make sense. To clarify, the variable ***hit_prob*** depicts how many hits a team has for every hit of their opponent.

Model 2

```
##
## Call:
## lm(formula = numWins ~ totHits + baseHits + doubles + triples +
##     walksTaken + strikeOuts + stolenBases + hitsAllowed + homerunsAllowed +
##     walksAllowed + errors + homer_prob + error_prob, data = mdlData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.206  -8.640   0.053   8.295  64.078
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.495e+01  7.177e+00   4.870 1.19e-06 ***
## totHits          7.526e-02  5.244e-02   1.435   0.1514
## baseHits        -4.438e-02  5.415e-02  -0.820   0.4125
## doubles         -9.513e-02  5.663e-02  -1.680   0.0931 .
## triples          4.248e-02  5.496e-02   0.773   0.4397
## walksTaken       1.280e-02  5.802e-03   2.207   0.0274 *
## strikeOuts      -2.625e-03  2.361e-03  -1.112   0.2663
## stolenBases      4.586e-02  4.586e-03  10.000  < 2e-16 ***
## hitsAllowed     -2.339e-03  5.791e-04  -4.040 5.53e-05 ***
## homerunsAllowed  1.053e-02  2.517e-02   0.418   0.6757
## walksAllowed    -3.085e-03  4.142e-03  -0.745   0.4564
## errors           5.409e-03  5.039e-03   1.073   0.2832
## homer_prob      -8.699e+01  9.561e+01  -0.910   0.3630
## error_prob      -8.253e+01  1.234e+01  -6.688 2.85e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.03 on 2241 degrees of freedom
## Multiple R-squared:  0.2835, Adjusted R-squared:  0.2793
## F-statistic: 68.21 on 13 and 2241 DF,  p-value: < 2.2e-16
```

This model includes more variables in an attempt to make its predictions a closer fit to the actual values. While its R_squared value increases slightly, its variables lose their significance. Many of its coefficients also do not make sense, yet where they don't, they're close to zero. This may be caused by the model attempting to overfit because some of its variables count the same information more than once. This model will not be kept.

Model 3

```
##
## Call:
## lm(formula = numWins ~ baseHits + doubles + triples + homeruns +
##     walksTaken + strikeOuts + stolenBases + hitsAllowed + homerunsAllowed +
##     walksAllowed + errors + hits_inning + hit_prob + homer_prob +
##     error_prob, data = mdlData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.262  -8.555   0.148   8.156  66.530
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.675e+01  9.155e+00   8.383  < 2e-16 ***
## baseHits         2.901e-02  4.904e-03   5.915 3.82e-09 ***
## doubles          8.548e-03  1.053e-02   0.812 0.416950
## triples          1.575e-01  1.840e-02   8.560  < 2e-16 ***
## homeruns         1.958e-01  5.741e-02   3.410 0.000661 ***
## walksTaken       3.181e-02  6.224e-03   5.110 3.49e-07 ***
## strikeOuts      -1.193e-03  2.451e-03  -0.487 0.626419
## stolenBases      4.816e-02  4.531e-03  10.629  < 2e-16 ***
## hitsAllowed     -8.577e-04  6.216e-04  -1.380 0.167770
## homerunsAllowed -1.430e-01  3.201e-02  -4.467 8.31e-06 ***
## walksAllowed    -1.692e-02  4.503e-03  -3.757 0.000177 ***
## errors          -3.134e-02  6.998e-03  -4.478 7.91e-06 ***
## hits_inning      3.949e-03  7.313e-02   0.054 0.956942
## hit_prob        -4.687e+01  5.822e+00  -8.051 1.32e-15 ***
## homer_prob       6.726e+01  9.835e+01   0.684 0.494153
## error_prob      -3.978e+01  1.362e+01  -2.921 0.003520 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.84 on 2239 degrees of freedom
## Multiple R-squared:  0.3046, Adjusted R-squared:    0.3
## F-statistic: 65.39 on 15 and 2239 DF,  p-value: < 2.2e-16
```

This model attempts to remove the overfitting by reducing the amount of times a variable and its transformed values are included. The coefficients on this model make sense, however not all variables contribute significantly to the overall model. There are variables that don't contribute, but the model has a tighter fit to the actual values which is seen by its R_squared value.

Select Models

The model to be selected is Model 1. This is chosen in sacrifice of the slightly better R_squared value seen in Model 3 because all of its variables contribute significantly to the model, reducing the opportunity for multi-collinearity to arise, and also because it has a larger F-statistic. This reinforces the significance of the variables used by showing their joint effect on the model. As seen in the predictions of the three models, all predict similar values, however with such a loosely fitted model there is plenty of room for improvement in its overall accuracy. Seen below are the first 25 predicted values from all three models.

```
##    teamINDEX mypred1 mypred2 mypred3
## 1          9      66      68      67
## 2         10      67      69      69
## 3         14      74      75      74
## 4         47      89      86      85
## 5         60      78      70      85
## 6         63      77      67      76
## 7         74      80      76      74
## 8         83      75      72      73
## 9         98      71      74      73
## 10       120      75      73      73
## 11       123      75      73      72
## 12       135      84      84      84
## 13       138      81      83      83
## 14       140      82      80      80
## 15       151      80      79      78
## 16       153      79      79      78
## 17       171      73      72      72
## 18       184      81      81      81
## 19       193      66      66      64
## 20       213      92      90      89
## 21       217      84      83      83
## 22       226      86      84      84
## 23       230      79      81      82
## 24       241      75      73      73
## 25       291      85      83      82
```