HW1: Predicting Baseball Wins

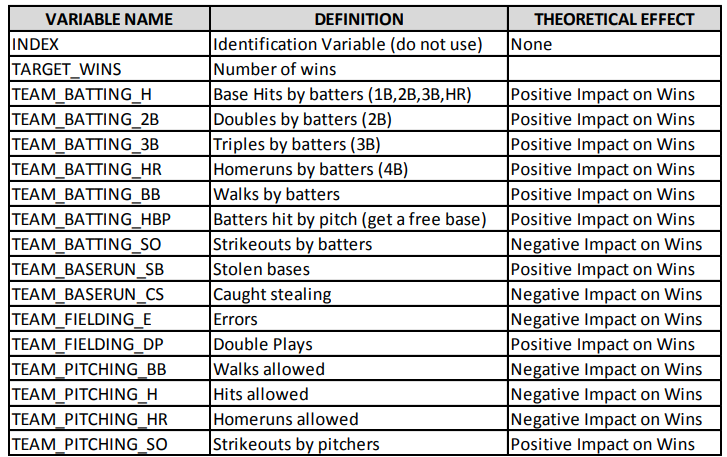
Selina Noori, Gavriel Steinmitz-Silber, John Cruz, Shaya Engelman, Daniel Craig

2024-02-23

## Data Exploration

   This data set describes baseball team statistics between the years of 1871 to 2006. The dataset contains 2,276 quantitative observations, documenting pitching, batting, and fielding performances across 17 variables. A quick explanation of each variable is below with their expected impact on predicting wins for a baseball team. All variables were numeric.

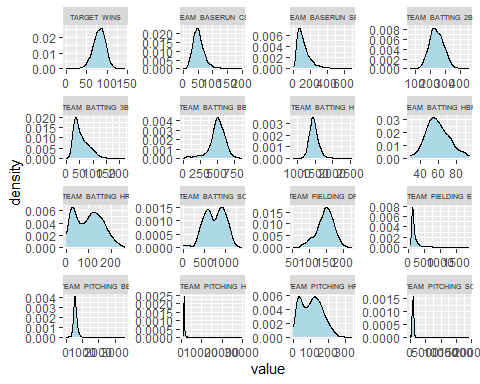
### Variable Summary

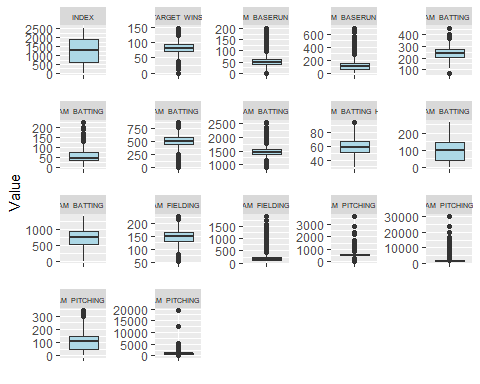


   A quick look at distributions with histograms and boxplots reveal a few alarming takeaways:

* TEAM\_FIELDING\_E: Numerous severe outliers
* TEAM\_PITCHING\_BB: Numerous severe outliers
* TEAM\_PITCHING\_H: Numerous severe outliers
* TEAM\_PITCHING\_SO: Numerous severe outliers
* TEAM\_BATTING\_H: Some severe outliers
* TEAM\_BASERUN\_SB: Some outliers
* TEAM\_BATTING\_3B Some outliers

   Outliers are detrimental to a model’s ability to predict due to their over-centralizing nature and weight a multiple linear regression model attributes to those observations when predicting. In particular, TEAM\_FIELDING\_E, TEAM\_PITCHING\_BB, TEAM\_PITCHING\_H, and TEAM\_PITCHING\_SO are heavily skewed with long tails due to these outliers.  
   Bi-modal distributions typically mean that basic mean or median imputation can introduce more bias into the dataset for missing values. Variables TEAM\_BATTING\_HR, TEAM\_BATTING\_SO, TEAM\_PITCHING\_HR all have bi-modal distributions. Observing the boxplots reinforces the significant number of outliers present in the data.





### Missing Data & Zero Values

   Upon further inspection of the data, many records contained 0’s instead of NA’s as recorded metrics, which were judged by the analysts as unreasonable values. There was also skepticism in whether the outlier values were reasonable or should be treated as errors. Zero values were replaced by NAs for imputation. Most columns were not missing data, two columns in particular stand out. TEAM\_BASERUN\_CS is missing 772 or 33.9% of values. TEAM\_BATTING\_HBP is missing 2085 or 91.6% of of values.  
   The threshold for observation removal was set at 50%. If an observation was missing values for 50% or more of its variables, the observation would be removed. No rows were missing more than 50% of their values and none removed. The threshold for variable removal was set at 25%. TEAM\_BASERUN\_CS and TEAM\_BATTING\_HBP both exceeded with missing values at 33.96% and 91.6%. Both had little correlation to TARGET\_WINS, although had moderate correlation to other variables as will be seen in the next section.

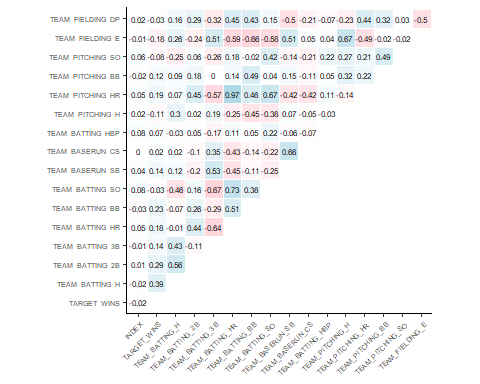
|  | missing\_counts |
| --- | --- |
| INDEX | 0 |
| TARGET\_WINS | 0 |
| TEAM\_BATTING\_H | 0 |
| TEAM\_BATTING\_2B | 0 |
| TEAM\_BATTING\_3B | 0 |
| TEAM\_BATTING\_HR | 0 |
| TEAM\_BATTING\_BB | 0 |
| TEAM\_BATTING\_SO | 102 |
| TEAM\_BASERUN\_SB | 131 |
| TEAM\_BASERUN\_CS | 772 |
| TEAM\_BATTING\_HBP | 2085 |
| TEAM\_PITCHING\_H | 0 |
| TEAM\_PITCHING\_HR | 0 |
| TEAM\_PITCHING\_BB | 0 |
| TEAM\_PITCHING\_SO | 102 |
| TEAM\_FIELDING\_E | 0 |
| TEAM\_FIELDING\_DP | 286 |

## Variable Missing Percentage

|  | x |
| --- | --- |
| INDEX | 0.000000 |
| TARGET\_WINS | 0.000000 |
| TEAM\_BATTING\_H | 0.000000 |
| TEAM\_BATTING\_2B | 0.000000 |
| TEAM\_BATTING\_3B | 0.000000 |
| TEAM\_BATTING\_HR | 0.000000 |
| TEAM\_BATTING\_BB | 0.000000 |
| TEAM\_BATTING\_SO | 4.481547 |
| TEAM\_BASERUN\_SB | 5.755712 |
| TEAM\_BASERUN\_CS | 33.919156 |
| TEAM\_BATTING\_HBP | 91.608084 |
| TEAM\_PITCHING\_H | 0.000000 |
| TEAM\_PITCHING\_HR | 0.000000 |
| TEAM\_PITCHING\_BB | 0.000000 |
| TEAM\_PITCHING\_SO | 4.481547 |
| TEAM\_FIELDING\_E | 0.000000 |
| TEAM\_FIELDING\_DP | 12.565905 |

### Correlation

   Correlations between TARGET\_WINS and the other variables are generally weak, with the strongest being with TEAM\_BATTING\_H with a positive 39% rating as expected. Notable negative correlations were limited to TEAM\_PITCHING\_E at -18% and TEAM\_PITCHING\_H at -11%. Surprisingly, TEAM\_PITCHING\_HR was very slightly positively correlated to target wins at 19% which was unexpected. At a glance, the overall correlations of a team’s batting related metrics are stronger than the metrics expected to be related to a negative effect. This may suggest that baseball play rewards batting more than not making errors or decreasing the enemy team’s abilities after a certain amount.



## Data Preparation

   The four main categories that preparation targeted were zero values, missing data, outliers, and skewness. Before any transformations, the training dataset was split on a 70/30 ratio to create a test data set to test models on. TEAM\_BASERUN\_CS and TEAM\_BATTING\_HBP are both dropped due to crossing a threshold of 25% missing data used as a general benchmark for removal. Zero values were replaced with NA (Not Applicable) values for imputation. Any observations that passed a threshold of 50% missing data were dropped as imputation is unreliable with so little data. A BoxCox transformation, centering, and scaling were all performed to help reduce the effect of outliers.

   Outliers were dealt with in two ways for testing. Many of the outliers break historical records and could be considered errors, but without contact with those that gathered the data this cannot be confirmed. If treated as errors per historical records, a large portion of data (roughly 30%) would be dropped. Instead, two methods were used to diminish impact of outliers. The first method was to drop values greater or smaller than 1.5 times the Interquartile Range (IQR) of data. The IQR is the distance between the 25th and 75th percentiles of a data’s distribution, effectively holding the majority of observations within it. The second method was to Winsorize values outside 1.5 times the IQR by replacing the outlier with a value in the 5th or 95th percentile of the distribution. Imputation for the dataset where outliers were dropped used mean imputation, except for columns TEAM\_BATTING\_HR, TEAM\_BATTING\_SO, TEAM\_PITCHING\_HR, and TEAM\_PITCHING\_SO where median imputation was used due to their less-normal behavior. Median imputation should deviate less from a distribution when non-normal.

| Variable | Count\_Of\_Drops |
| --- | --- |
| Bat\_H\_Drop | 0 |
| Bat\_2B\_Drop | 0 |
| Bat\_3B | 20 |
| Bat\_HR | 0 |
| Bat\_BB | 0 |
| Base\_SB\_Drop | 77 |
| Pitch\_H\_Drop | 0 |
| Pitch\_HR\_Drop | 3 |
| Pitch\_BB\_Drop | 0 |
| Pitch\_SO\_Drop | 0 |
| Field\_E\_Drop | 145 |
| Field\_DP\_Drop | 0 |

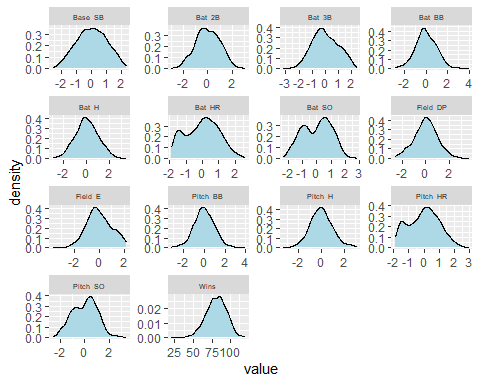
### Adding Variables

   The style of baseball play has changed a lot over time. For example, strikeouts are far more common today than they were many years ago. It follows that the raw number of batting strikeouts a team has is not especially insightful. Rather, what might be insightful is how a team’s batting strikeouts compare to their pitching strikeouts. As such, we create four such ratios. These variables and their expected relationship to Wins are:

1. TEAM\_BATTING\_H\_TEAM\_PITCHING\_H\_RATIO - Positive Correlation
2. TEAM\_BATTING\_HR\_TEAM\_PITCHING\_HR\_RATIO - Positive Correlation
3. TEAM\_BATTING\_BB\_TEAM\_PITCHING\_BB\_RATIO - Positive Correlation
4. TEAM\_BATTING\_SO\_TEAM\_PITCHING\_SO\_RATIO - Positive Correlation

### BoxCox Transformation

   To help manage the skewness introduced into the data from the outliers, a BoxCox transformation was introduced. It helped signficiantly. The distributions of the transformed data can be seen below for comparison with the original histograms.



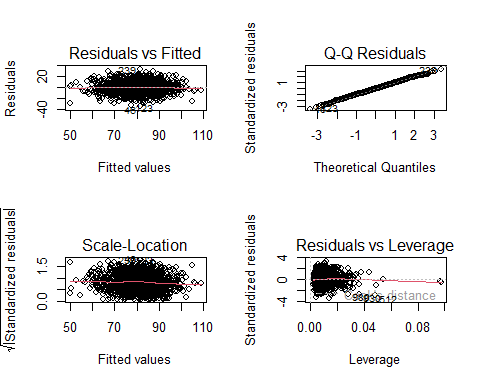
## Data Modeling

   For modeling techniques, each dataset (outliers-dropped and winsorized) had a model created with all main order and interaction effects and a third model was created including only main effects for the outliers dropped data. Features were removed one at a time based on their p-values and contribution to the model’s predictive ability. Only features holding a significance of .01 or lower were kept to control for family-wise error as testing many features for significance can result in false positives. When checking for significant effects from variables, the more of them that are checked the higher the chances are that a false positive is detected. If an interaction was signficiant, but dependent variables were insignficant, the dependent variables were still kept in the model.  
   The Variance Inflation Factor was reviewed for all models. The Outliers-Dropped model had significantly lower VIF for its variables compared the Winsorized model. The Outliers-Dropped model relied heavily on main effects, while the Winsorized model used many interaction effects. A main-effects-only model was also made and performed roughly as well as the Winsorized model. Multi-colinearity was only apparent in the Winsorized model with high VIF values.

   The three models achieved the following results in testing, with the Outliers Dropped model to be chosen for prediction:

| Model | RMSE | MSE | MAE | Adj\_R\_squared |
| --- | --- | --- | --- | --- |
| Outliers Dropped | 10.81660 | 116.9988 | 8.680498 | 0.3964623 |
| Winsorized | 13.12356 | 172.2279 | 10.062822 | 0.3268877 |
| Main Effects | 11.41008 | 130.1900 | 9.203352 | 0.3279413 |
| The final formula used was: |  |  |  |  |

   The Outliers Dropped model served to have the best metrics across all RMSE, MSE, MAE, and R-Squared by a significant margin. The F-Statistic was significantly high at 72.26 confirming that the model predicts better than an intercept only model with no variables. The Residuals vs Fitted plot seems relatively okay with a slight oblong nature although some relationship is likely missing. The QQ Plot follows normality well. The Scale v Location plot is not perfectly horizontal and shares the same oblong nature as the Residauls vs Fitted plot. This suggests that homoscedasticiy is relatively intact with equal variance between Residuals and Fitted observations. The Residuals vs Leverage plot highlights 2 - 3 points that altered predictions significantly. Considering the original data and its significant number of outliers, this is quite respectable as it only leverages the model slightly. Due to the highest accuracy in testing and its lower VIF values, the Outliers Dropped model was chosen for prediction.



### Coefficients Discussion

   Ultimately, the relationships between wins and the used predictors are below. Surprisingly TEAM\_FIELDING\_DP had a negative relationship with Wins. It may be because a double play implies that many batters were able to get on bases the negative effect of poor pitching is not adequately captured. Another surprising relationship is TEAM\_BATTING\_2B having a negative effect on Wins. It is odd that both double play related metrics have a negative relationship with wins. It may be a point to bring up with data collectors for further insight. It is also interesting to note that scoring related variables are prevalent, yet defensive variables are unimpactful. Exact metrics can be seen below.

| Variable | Defintion | Model\_Effect |
| --- | --- | --- |
| TEAM\_FIELDING\_DP | Double Plays | Moderately Negative |
| TEAM\_FIELDING\_E | Errors | Heavily Negative |
| TEAM\_BASERUN\_SB | Stolen Bases | Moderatly Positive |
| TEAM\_BATTING\_3B | Triples by Batters | Moderately Positive |
| TEAM\_BATTING\_BB | Walks by batters | Moderately Positive |
| TEAM\_BATTING\_HR | Homeruns by batters | Moderately Positive |
| TEAM\_BATTING\_H | Base Hits by batters | Moderatly Positive |
| TEAM\_BATTING\_SO | Strikeouts by batters | Moderately Negative |
| TEAM\_BATTING\_2B | Doubles by batters | Moderately Negative |

| Variable\_Interaction | Model\_Effect |
| --- | --- |
| TEAM\_BASERUN\_SB \* TEAM\_BATTING\_HR | Moderately Negative |
| TEAM\_FIELDING\_E \* TEAM\_BATTING\_2B | Moderatly Positive |
| TEAM\_FIELDING\_E \* TEAM\_BATTING\_SO | Moderatly Positive |
| TEAM\_FIELDING\_DP \* TEAM\_BATTING\_BB | Slightly Positive |

##   
## Call:  
## lm(formula = Wins ~ Field\_DP + Field\_E + Base\_SB + Bat\_3B + Bat\_BB +   
## Bat\_HR + Bat\_H + Bat\_SO + Bat\_2B + Bat\_HR + Bat\_SO + Bat\_HR \*   
## Base\_SB + Bat\_2B \* Field\_E + Bat\_SO:Field\_E + Bat\_BB:Field\_DP,   
## data = trainBC\_outs\_drop)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -34.460 -7.113 0.033 7.036 32.538   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 81.5560 0.3805 214.318 < 2e-16 \*\*\*  
## Field\_DP -3.4298 0.3566 -9.619 < 2e-16 \*\*\*  
## Field\_E -7.6425 0.5171 -14.780 < 2e-16 \*\*\*  
## Base\_SB 2.5898 0.3598 7.198 1.01e-12 \*\*\*  
## Bat\_3B 2.5732 0.4726 5.445 6.16e-08 \*\*\*  
## Bat\_BB 3.1704 0.3279 9.669 < 2e-16 \*\*\*  
## Bat\_HR 2.7791 0.6308 4.406 1.14e-05 \*\*\*  
## Bat\_H 4.6155 0.5346 8.633 < 2e-16 \*\*\*  
## Bat\_SO -3.7710 0.5820 -6.480 1.29e-10 \*\*\*  
## Bat\_2B -1.5194 0.4858 -3.127 0.001801 \*\*   
## Base\_SB:Bat\_HR -3.2052 0.3569 -8.982 < 2e-16 \*\*\*  
## Field\_E:Bat\_2B 1.5063 0.3066 4.913 1.01e-06 \*\*\*  
## Field\_E:Bat\_SO 2.4864 0.3812 6.523 9.75e-11 \*\*\*  
## Field\_DP:Bat\_BB 0.8646 0.2434 3.552 0.000395 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.3 on 1336 degrees of freedom  
## Multiple R-squared: 0.4129, Adjusted R-squared: 0.4071   
## F-statistic: 72.26 on 13 and 1336 DF, p-value: < 2.2e-16

## Field\_DP Field\_E Base\_SB Bat\_3B Bat\_BB   
## 1.597333 3.397168 1.644661 2.837420 1.366056   
## Bat\_HR Bat\_H Bat\_SO Bat\_2B Base\_SB:Bat\_HR   
## 5.054706 3.631286 4.095865 2.998605 1.628836   
## Field\_E:Bat\_2B Field\_E:Bat\_SO Field\_DP:Bat\_BB   
## 1.523223 1.364257 1.173820

### Details for Other Models

## \*\*\*\*Main Effects Summary\*\*\*\*\*

##   
## Call:  
## lm(formula = Wins ~ Field\_E + Field\_DP + Base\_SB + Bat\_2B + Bat\_3B +   
## Bat\_3B + Bat\_H + Bat\_BB + Bat\_SO, data = trainBC\_outs\_drop)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -45.265 -7.267 0.224 7.166 38.324   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 80.7793 0.3000 269.301 < 2e-16 \*\*\*  
## Field\_E -6.7237 0.5165 -13.017 < 2e-16 \*\*\*  
## Field\_DP -3.1904 0.3699 -8.626 < 2e-16 \*\*\*  
## Base\_SB 1.6254 0.3469 4.686 3.07e-06 \*\*\*  
## Bat\_2B -2.8624 0.4980 -5.748 1.12e-08 \*\*\*  
## Bat\_3B 2.5041 0.4808 5.208 2.20e-07 \*\*\*  
## Bat\_H 5.4531 0.5054 10.790 < 2e-16 \*\*\*  
## Bat\_BB 4.1030 0.3377 12.150 < 2e-16 \*\*\*  
## Bat\_SO -1.7489 0.5058 -3.457 0.000562 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.02 on 1341 degrees of freedom  
## Multiple R-squared: 0.3258, Adjusted R-squared: 0.3218   
## F-statistic: 81 on 8 and 1341 DF, p-value: < 2.2e-16

##   
## \*\*\*\*\*Main Effects VIF\*\*\*\*\*

## Field\_E Field\_DP Base\_SB Bat\_2B Bat\_3B Bat\_H Bat\_BB Bat\_SO   
## 2.963047 1.502333 1.336159 2.754473 2.567243 2.836511 1.266466 2.704870

##   
## \*\*\*\*\*Winsor Summary\*\*\*\*\*

##   
## Call:  
## lm(formula = Wins ~ Pitch\_SO + Bat\_SO\_Pitch\_SO\_Ratio + Field\_E +   
## Base\_SB + Bat\_3B + Bat\_H + Bat\_BB + Pitch\_H + Pitch\_HR +   
## Pitch\_BB + Bat\_H \* Bat\_HR\_Pitch\_HR\_Ratio + Bat\_2B \* Field\_E +   
## Bat\_HR \* Bat\_BB + Bat\_HR \* Base\_SB + Pitch\_SO \* Field\_E +   
## Pitch\_BB \* Bat\_SO\_Pitch\_SO\_Ratio + Field\_DP + Bat\_2B \* Bat\_H\_Pitch\_H\_Ratio +   
## Pitch\_SO \* Field\_E + Bat\_H:Bat\_HR\_Pitch\_HR\_Ratio + Bat\_BB\_Pitch\_BB\_Ratio,   
## data = trainBC\_winsor)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -52.115 -7.588 0.130 7.384 46.584   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 80.5137 0.4440 181.332 < 2e-16 \*\*\*  
## Pitch\_SO -2.5353 0.5565 -4.556 5.62e-06 \*\*\*  
## Bat\_SO\_Pitch\_SO\_Ratio 1.9659 1.0864 1.810 0.070559 .   
## Field\_E -9.9037 0.7222 -13.713 < 2e-16 \*\*\*  
## Base\_SB 3.9857 0.4233 9.415 < 2e-16 \*\*\*  
## Bat\_3B 3.0927 0.5582 5.541 3.52e-08 \*\*\*  
## Bat\_H 1.3136 1.3174 0.997 0.318857   
## Bat\_BB 2.2141 1.3440 1.647 0.099685 .   
## Pitch\_H 7.5841 2.3065 3.288 0.001031 \*\*   
## Pitch\_HR 4.7216 2.5808 1.830 0.067510 .   
## Pitch\_BB 0.2959 1.2829 0.231 0.817593   
## Bat\_HR\_Pitch\_HR\_Ratio -2.1905 1.2357 -1.773 0.076492 .   
## Bat\_2B -0.5592 0.4942 -1.132 0.257984   
## Bat\_HR -2.1899 2.9201 -0.750 0.453419   
## Field\_DP -3.1115 0.3898 -7.981 2.77e-15 \*\*\*  
## Bat\_H\_Pitch\_H\_Ratio 92.4201 82.6393 1.118 0.263586   
## Bat\_BB\_Pitch\_BB\_Ratio -87.5453 82.6512 -1.059 0.289665   
## Bat\_H:Bat\_HR\_Pitch\_HR\_Ratio -2.4118 0.5201 -4.637 3.83e-06 \*\*\*  
## Field\_E:Bat\_2B 1.4224 0.4634 3.069 0.002183 \*\*   
## Bat\_BB:Bat\_HR 1.2380 0.3867 3.201 0.001397 \*\*   
## Base\_SB:Bat\_HR -3.5796 0.4174 -8.576 < 2e-16 \*\*\*  
## Pitch\_SO:Field\_E 3.8399 0.4196 9.150 < 2e-16 \*\*\*  
## Bat\_SO\_Pitch\_SO\_Ratio:Pitch\_BB 1.2908 0.3494 3.694 0.000228 \*\*\*  
## Bat\_2B:Bat\_H\_Pitch\_H\_Ratio -1.9179 0.4140 -4.633 3.90e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.94 on 1571 degrees of freedom  
## Multiple R-squared: 0.4273, Adjusted R-squared: 0.419   
## F-statistic: 50.97 on 23 and 1571 DF, p-value: < 2.2e-16

##   
## \*\*\*\*\*Winsor VIF\*\*\*\*\*

## Pitch\_SO Bat\_SO\_Pitch\_SO\_Ratio   
## 3.268851 12.456323   
## Field\_E Base\_SB   
## 5.829806 1.886325   
## Bat\_3B Bat\_H   
## 3.480034 19.398901   
## Bat\_BB Pitch\_H   
## 20.190454 59.462034   
## Pitch\_HR Pitch\_BB   
## 73.854193 18.395013   
## Bat\_HR\_Pitch\_HR\_Ratio Bat\_2B   
## 16.929277 2.729682   
## Bat\_HR Field\_DP   
## 94.556136 1.484475   
## Bat\_H\_Pitch\_H\_Ratio Bat\_BB\_Pitch\_BB\_Ratio   
## 76332.272682 76354.389524   
## Bat\_H:Bat\_HR\_Pitch\_HR\_Ratio Field\_E:Bat\_2B   
## 4.547342 2.552010   
## Bat\_BB:Bat\_HR Base\_SB:Bat\_HR   
## 1.793095 1.459457   
## Pitch\_SO:Field\_E Bat\_SO\_Pitch\_SO\_Ratio:Pitch\_BB   
## 1.641103 2.806071   
## Bat\_2B:Bat\_H\_Pitch\_H\_Ratio   
## 2.507611