HW1: Predicting Baseball Wins

https://github.com/d-ev-craig/DATA621_Group/tree/main/HW1

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

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_WINS	Number of wins	
TEAM_BATTING_H	Base Hits by batters (1B,2B,3B,HR)	Positive Impact on Wins
TEAM_BATTING_2B	Doubles by batters (2B)	Positive Impact on Wins
TEAM_BATTING_3B	Triples by batters (3B)	Positive Impact on Wins
TEAM_BATTING_HR	Homeruns by batters (4B)	Positive Impact on Wins
TEAM_BATTING_BB	Walks by batters	Positive Impact on Wins
TEAM_BATTING_HBP	Batters hit by pitch (get a free base)	Positive Impact on Wins
TEAM_BATTING_SO	Strikeouts by batters	Negative Impact on Wins
TEAM_BASERUN_SB	Stolen bases	Positive Impact on Wins
TEAM_BASERUN_CS	Caught stealing	Negative Impact on Wins
TEAM_FIELDING_E	Errors	Negative Impact on Wins
TEAM_FIELDING_DP	Double Plays	Positive Impact on Wins
TEAM_PITCHING_BB	Walks allowed	Negative Impact on Wins
TEAM_PITCHING_H	Hits allowed	Negative Impact on Wins
TEAM_PITCHING_HR	Homeruns allowed	Negative Impact on Wins
TEAM_PITCHING_SO	Strikeouts by pitchers	Positive Impact on Wins

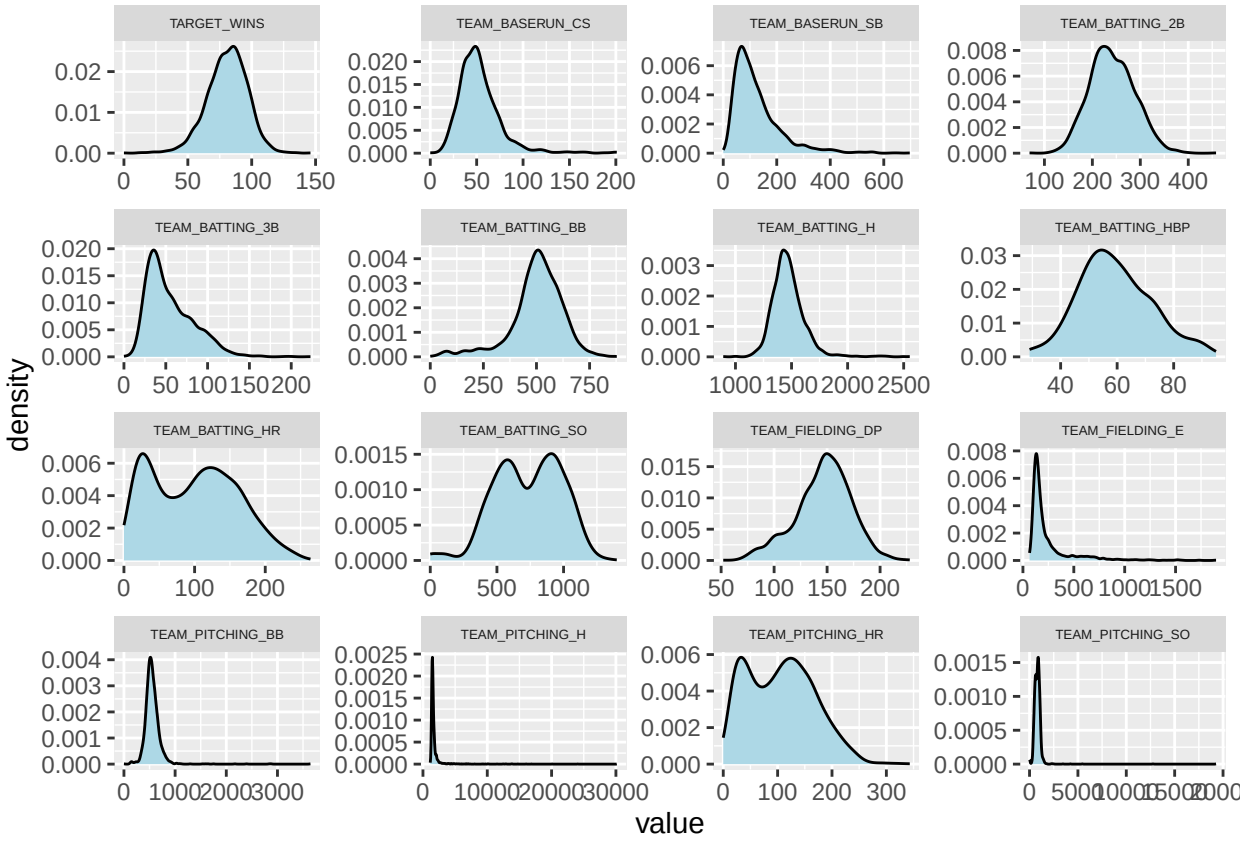
Figure 1: A caption

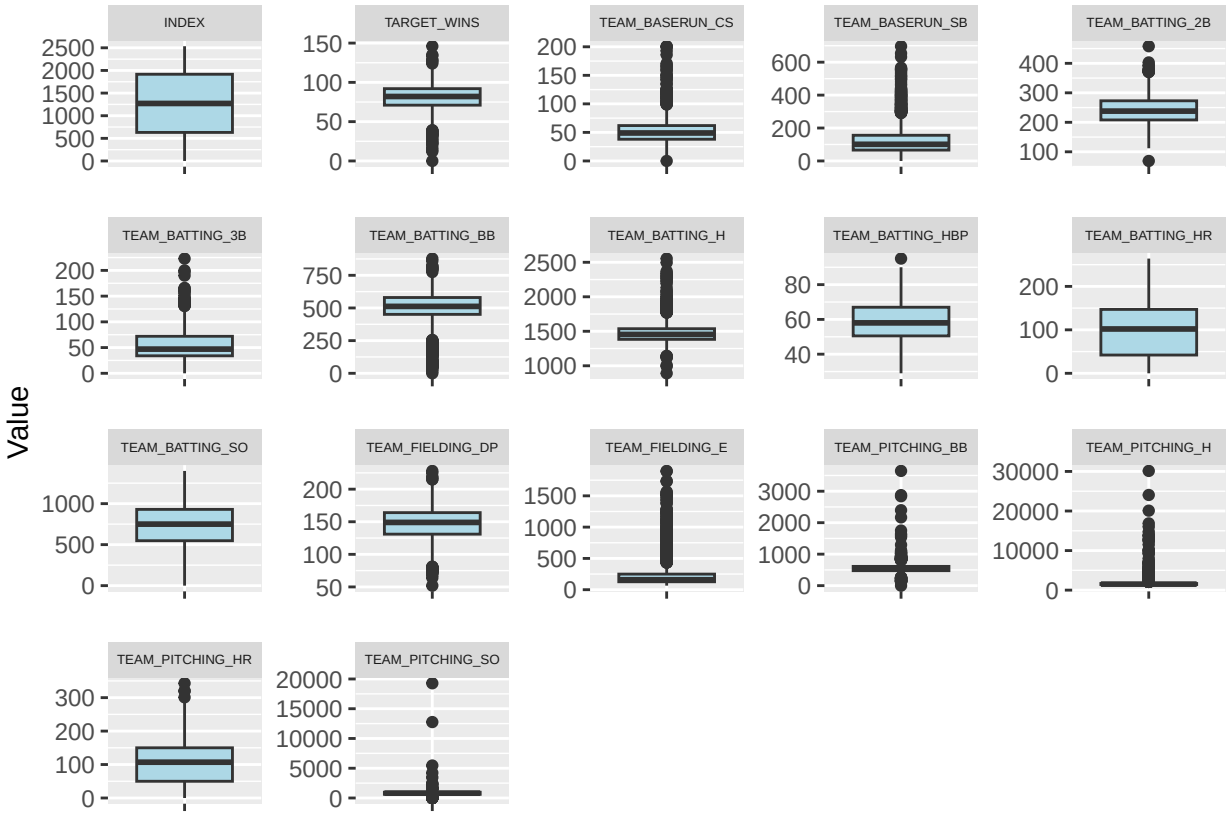
A quick look at distributions with histograms and boxplots reveal a few alarming takeaways:

- 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.




Value
INDEX
TARGET_WINS
TEAM_BASERUN_CS
TEAM_BASERUN_SB
TEAM_BATTING_2B
TEAM_BATTING_3B
TEAM_BATTING_BB
TEAM_BATTING_H
TEAM_BATTING_HBP
TEAM_BATTING_HR
TEAM_BATTING_SO
TEAM_FIELDING_DP
TEAM_FIELDING_E
TEAM_PITCHING_BB
TEAM_PITCHING_H
TEAM_PITCHING_HR
TEAM_PITCHING_SO

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

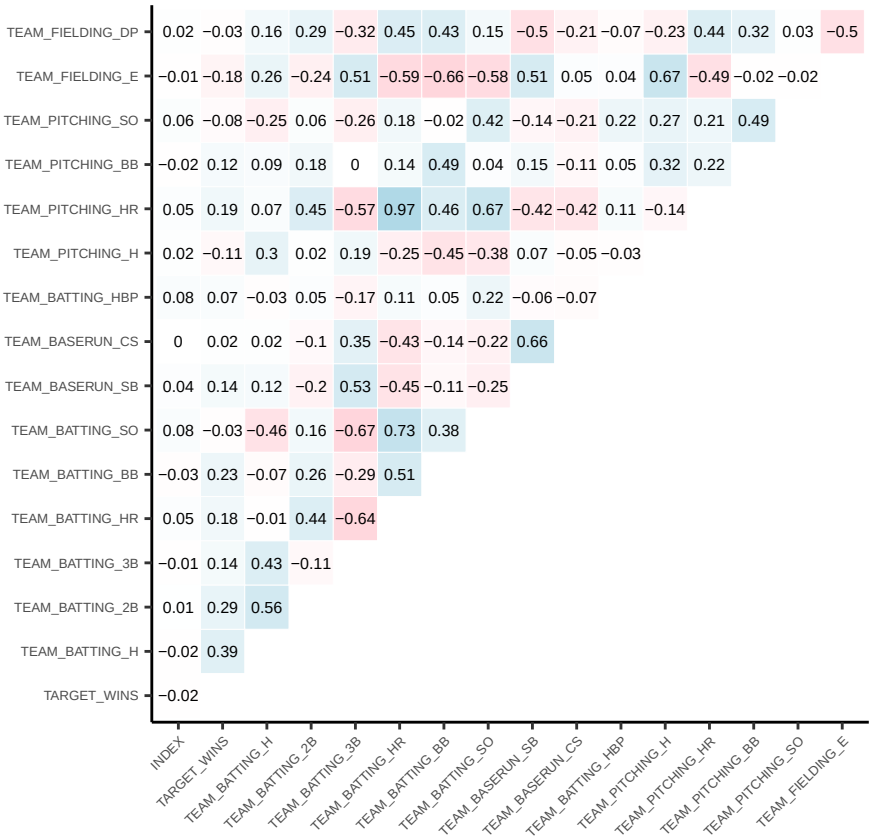
```
## Variable Missing Percentage
```

	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

	x
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

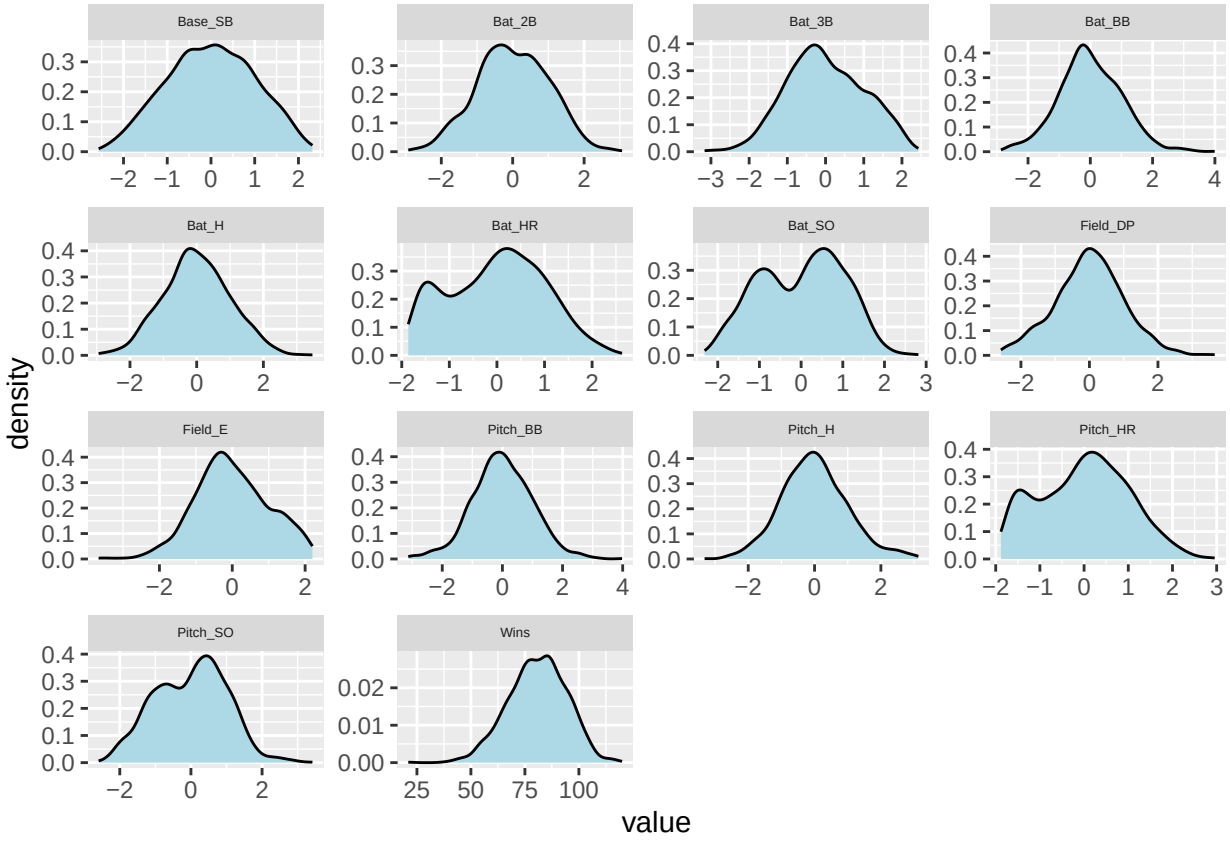
Adding Variables

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

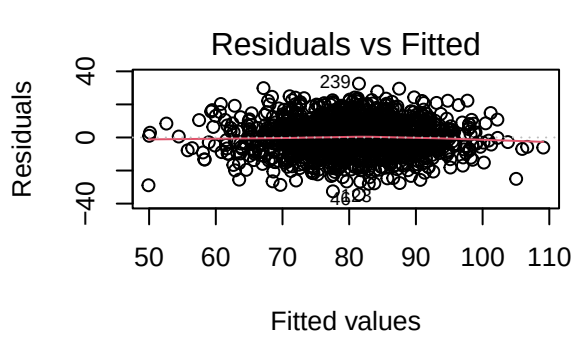
For modeling techniques, each dataset (outliers-dropped and winsorized) had a model created with all main order and interaction effects and a third model was created including only main effects for the outliers dropped data. Features were removed one at a time based on their p-values and contribution 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.

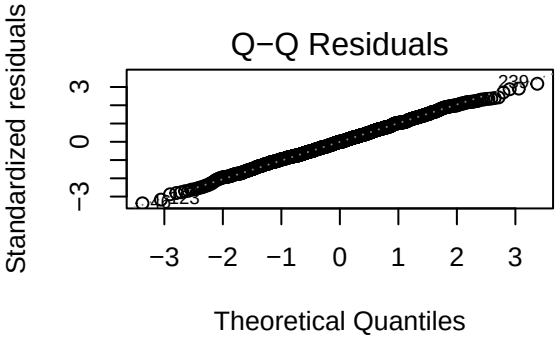
The Variance Inflation Factor was reviewed for all models. The Outliers-Dropped model had significantly lower VIF for its variables compared the Winsorized model. The Outliers-Dropped model relied heavily on main effects, while the Winsorized model used many interaction effects. A main-effects-only model 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.

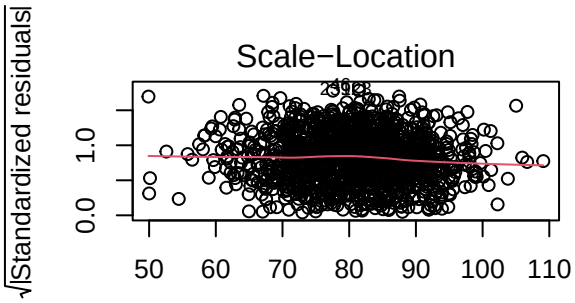
The three models achieved the following results in testing, with the Outliers Dropped model to be chosen for prediction:

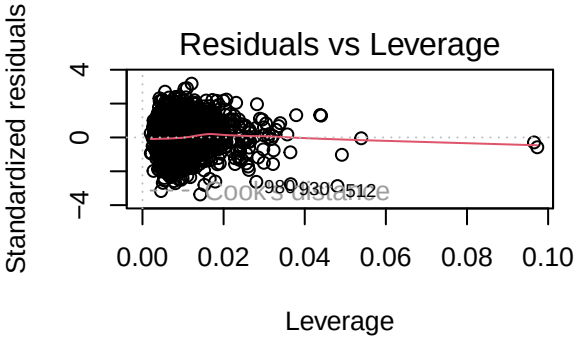
Model	RMSE	MSE	MAE	Adj_R_squared
Outliers Dropped	10.81660	116.9988	8.680498	0.3964623
Winsorized	13.12356	172.2279	10.062822	0.3268877
Main Effects	11.41008	130.1900	9.203352	0.3279413

The 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.


Residuals vs Fitted
Residuals
Fitted values
239


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
239


Scale−Location
√|Standardized residuals|
Fitted values


Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
980
930
512

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
```