CUNY DATA 621 - Business Analytics and Data Mining

Homework 1 - Moneyball

Walt Wells, 2018

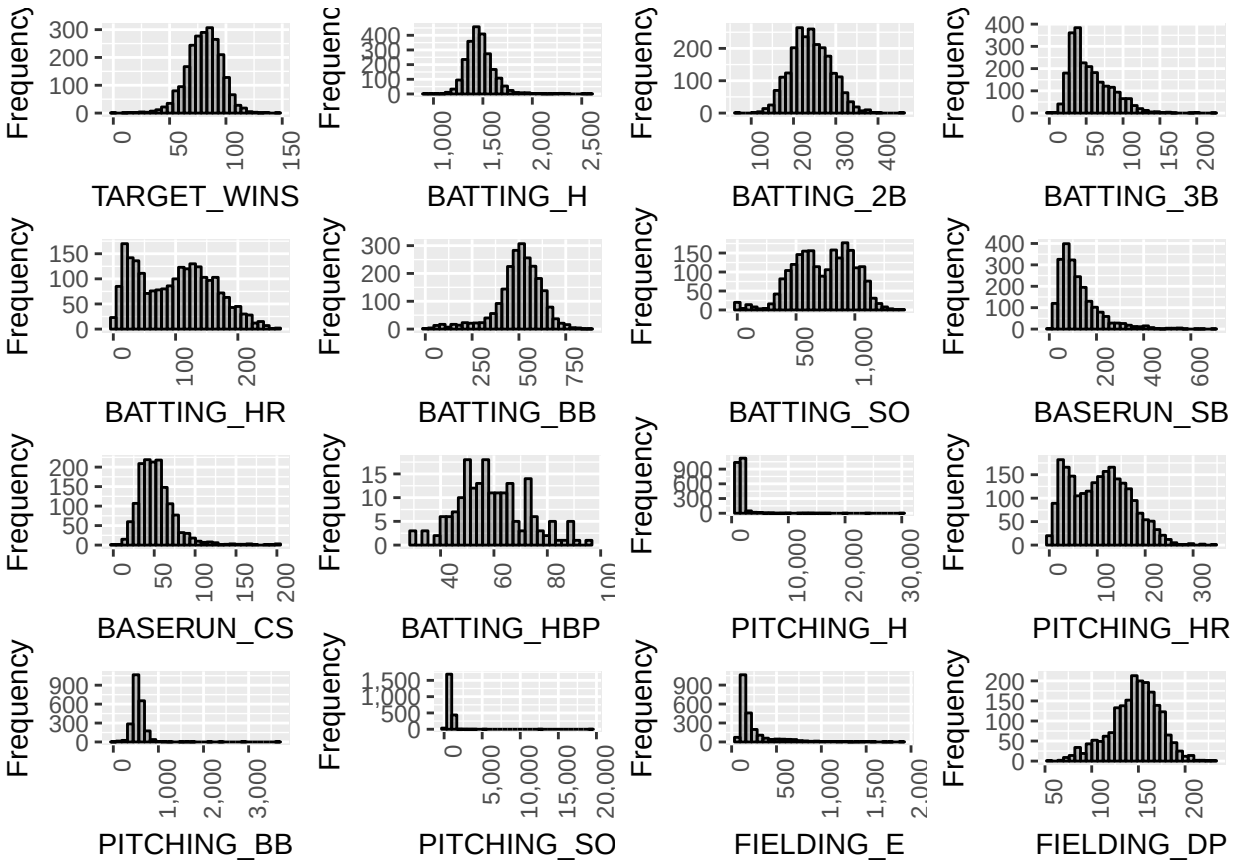
1. DATA EXPLORATION (25 Points)

Below we'll display a few basic EDA techniques to gain insight into our baseball dataset.

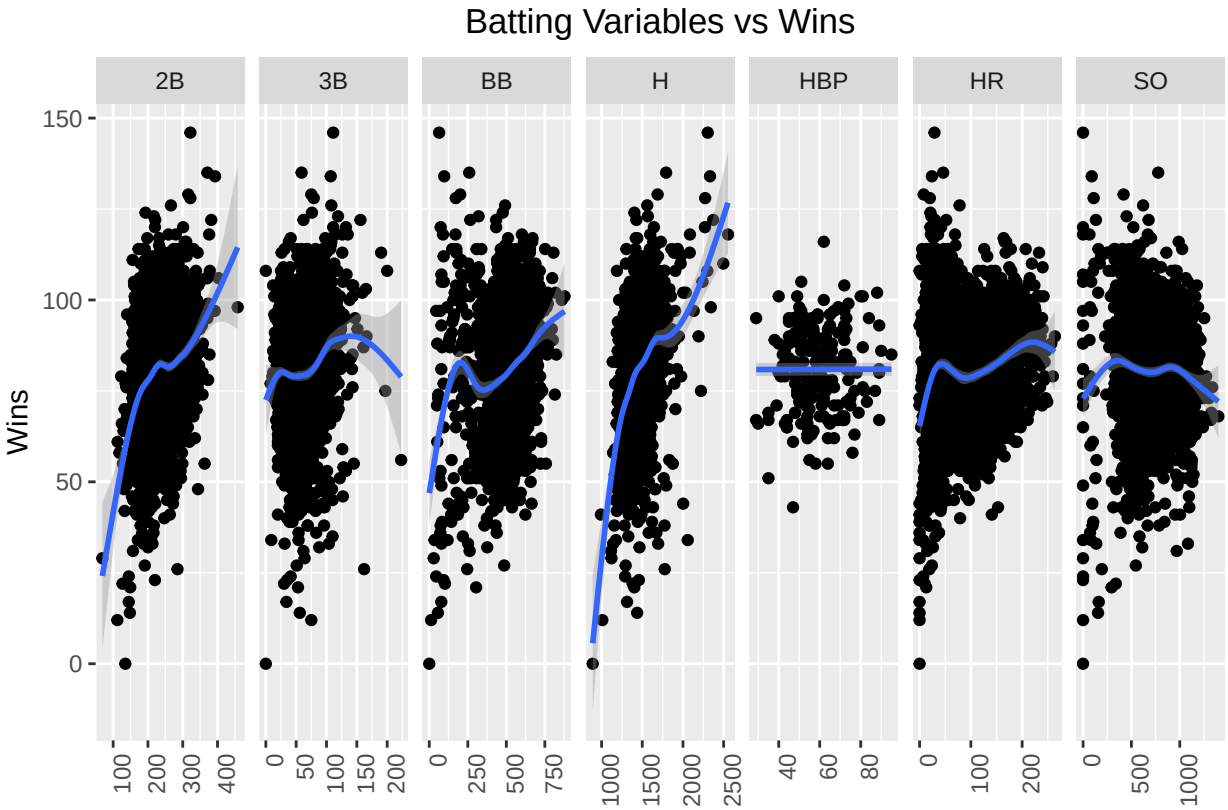
Basic Statistics

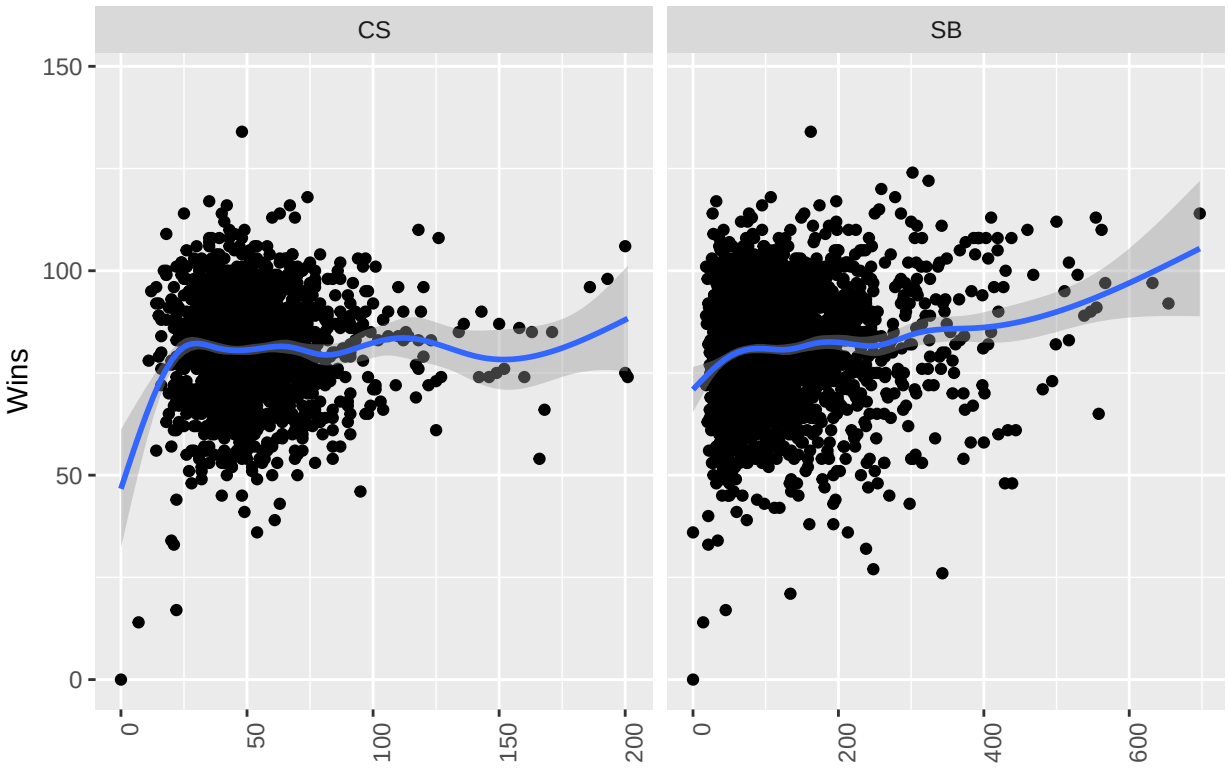
The data is 144.6 Kb in size. There are 2,276 rows and 16 columns (features). Of all 16 columns, 0 are discrete, 16 are continuous, and 0 are all missing. There are 3,478 missing values out of 36,416 data points.

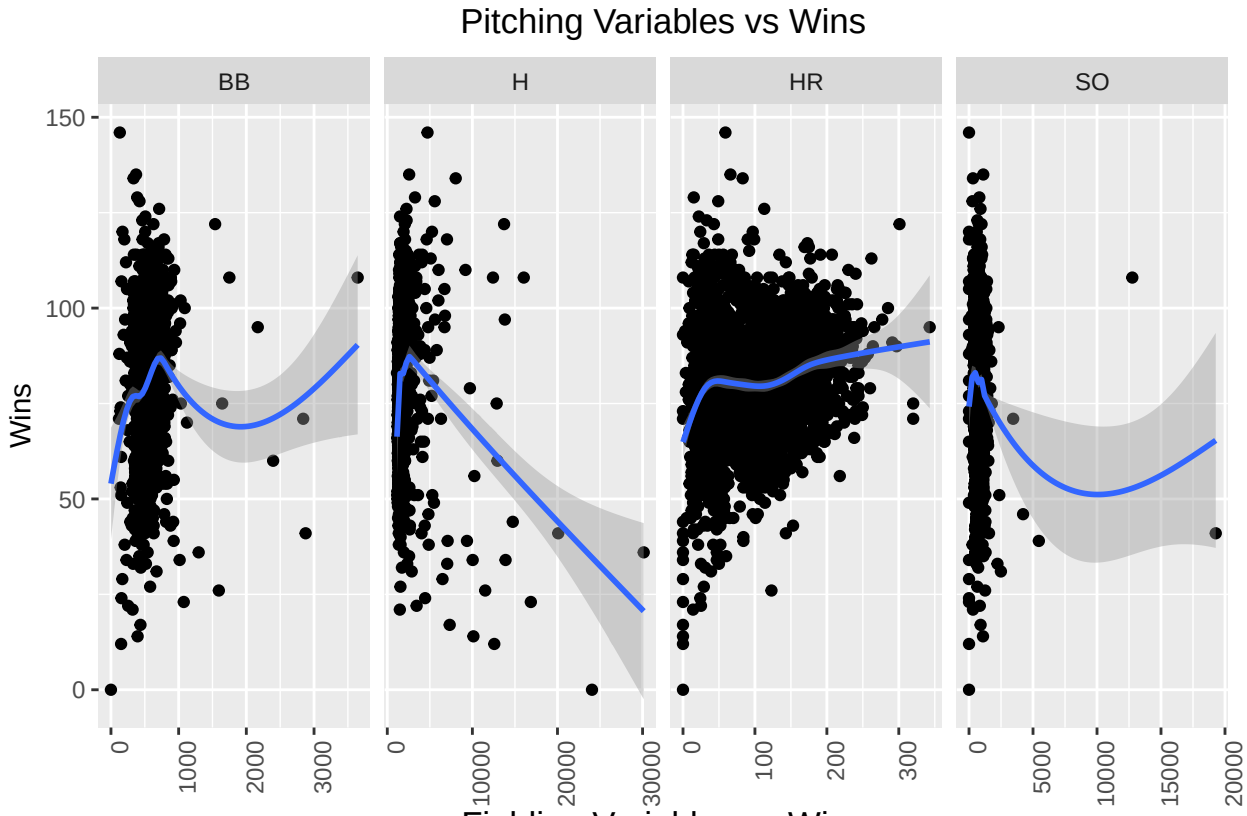
Histogram of Variables

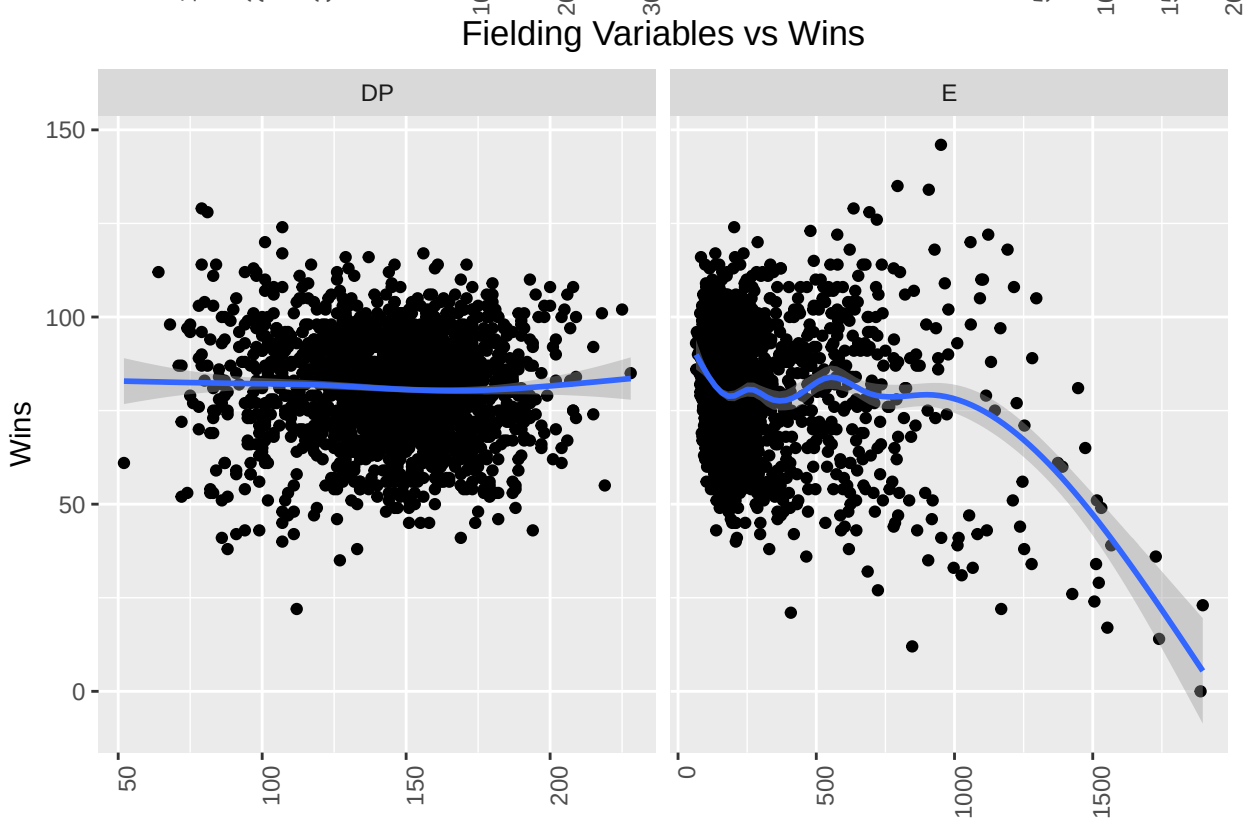


Scatterplots of Each Variable Vs Target Wins





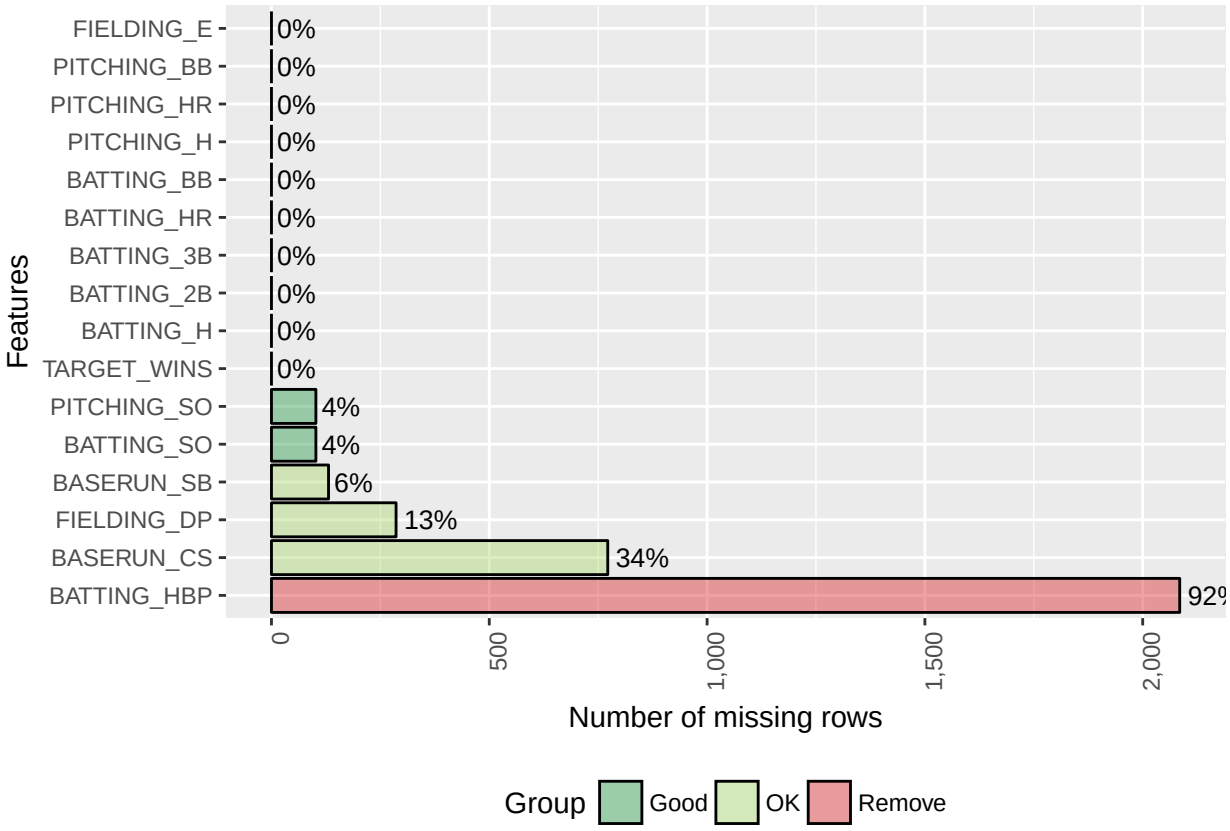

Pitching Variables vs Wins
BB
H
HR
SO
Wins
150
100
50
0
0
1000
2000
3000
0
10000
20000
30000
0
100
200
300
0
5000
10000
15000
20000


Fielding Variables vs Wins
DP
E
Wins
150
100
50
0
50
100
150
200
0
500
1000
1500

2. DATA PREPARATION (25 Points)

First, let's explore our missing values.

Missing Values



As we can see from our chart, we have a number of missing values. We'll use Median imputation for CS, SB, and DP. Since HBP has 92% missing values, we will remove that entirely. Interestingly, Pitching and Batting SO are missing in the same observations, so once we have transformed our data with the above imputations, we will remove the 4% of cases that have that characteristic.

Feature Creation

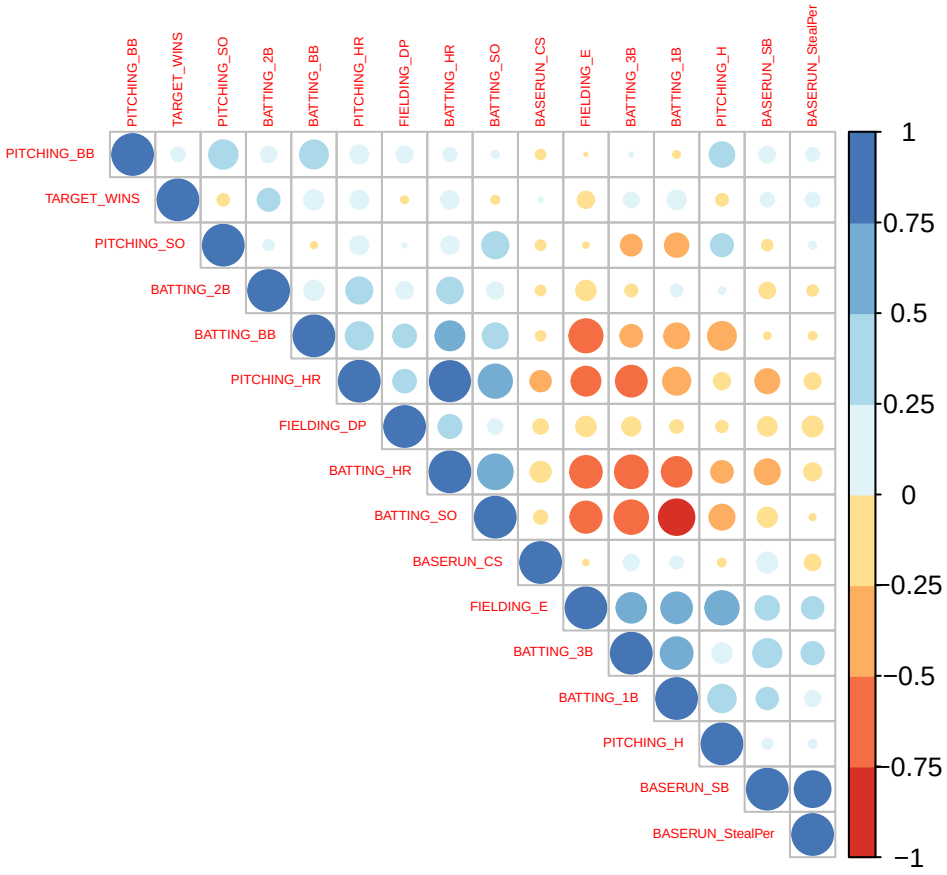
We will also create two new variables:

- BATTING_1B = BATTING_H - BATTING_HR - BATTING_3B -BATTING_2B
- BASERUN_StealPer = BASERUN_SB / (BASERUN_SB + BASERUN_CS))

Once we have created our BATTING_1B, we will remove BATTING_H from the model.

Variable Correlation

Here we explore the correlation of variables in our prepared dataset.



Transformation

Some initial exploration was done to transform our dependant and independant variables using log, sqrt, and box-cox, but no obvious gains were made. It was therefore decided not attempt to apply transformation methods at this time.

3. BUILD MODELS (25 Points)

We built and compared a few different models. These are outlined below. For each, we checked assumptions necessary for regression.

RawData_Linear_Model_AllVar

This model is essentially the raw training dataset and all the variables included. As a result, the lm method in r removes all of the NAs that we prepared using the methods outlines in section 2. We'll use this as our roughest baseline.

PrepData_Linear_Model_AllVar

This model uses our prepared training dataset and creates a linear model that utilizes all the variables.

PrepData_Linear_Model_StepSelect

This method uses the linear model above, but does backwards feature selection using the StepAIC method in R. It ultimately retains 12 of the original 15 variables.

PrepData_Poly_Model_AllVar

This method uses our prepared training dataset, and creates a polynomial fit to the 1-4 power for each of the 15 variables. This creates 61 coefficients (intercept + 15 variables * 4 powers) to fit.

PrepData_Poly_Model_StepSelect

Similar to what we did with the PrepData_Linear_Model_StepSelect, this model takes the PrepData_Poly_Model_AllVar and does backwards feature selection over it. This method eliminates 17 of the 61 variables, giving us 44 features.

OutlierRM_Poly_Model_StepSelect

This is a refined version of the PrepData_Poly_Model_StepSelect model. We performed additional diagnostics on that model, found and removed leverage points and outliers, then refit the Polynomial model and performed backwards feature selection again. After two iterations, this is the model we end up with.

4. SELECT MODELS (25 Points)

Below please find a table showing the R2, MSE, F-statistic, Number of Variables (K), Number of Observations (N), and number of observations in the original training set that were excluded from the model.

name	rsquared	mse	f	k	n	RemovedObservations
RawData_Linear_Model_AllVar	0.5501165	65.68529	14.26597	15	175	2101
PrepData_Linear_Model_AllVar	0.3139215	164.33285	65.79701	15	2157	119
PrepData_Linear_Model_StepSelect	0.3133432	164.47136	82.13971	12	2160	116
PrepData_Poly_Model_AllVar	0.4498569	131.77296	28.78335	60	2112	164
PrepData_Poly_Model_StepSelect	0.4472962	132.38630	40.06913	43	2129	147
OutlierRM_Poly_Model_StepSelect	0.4616202	126.61126	44.41879	41	2124	152

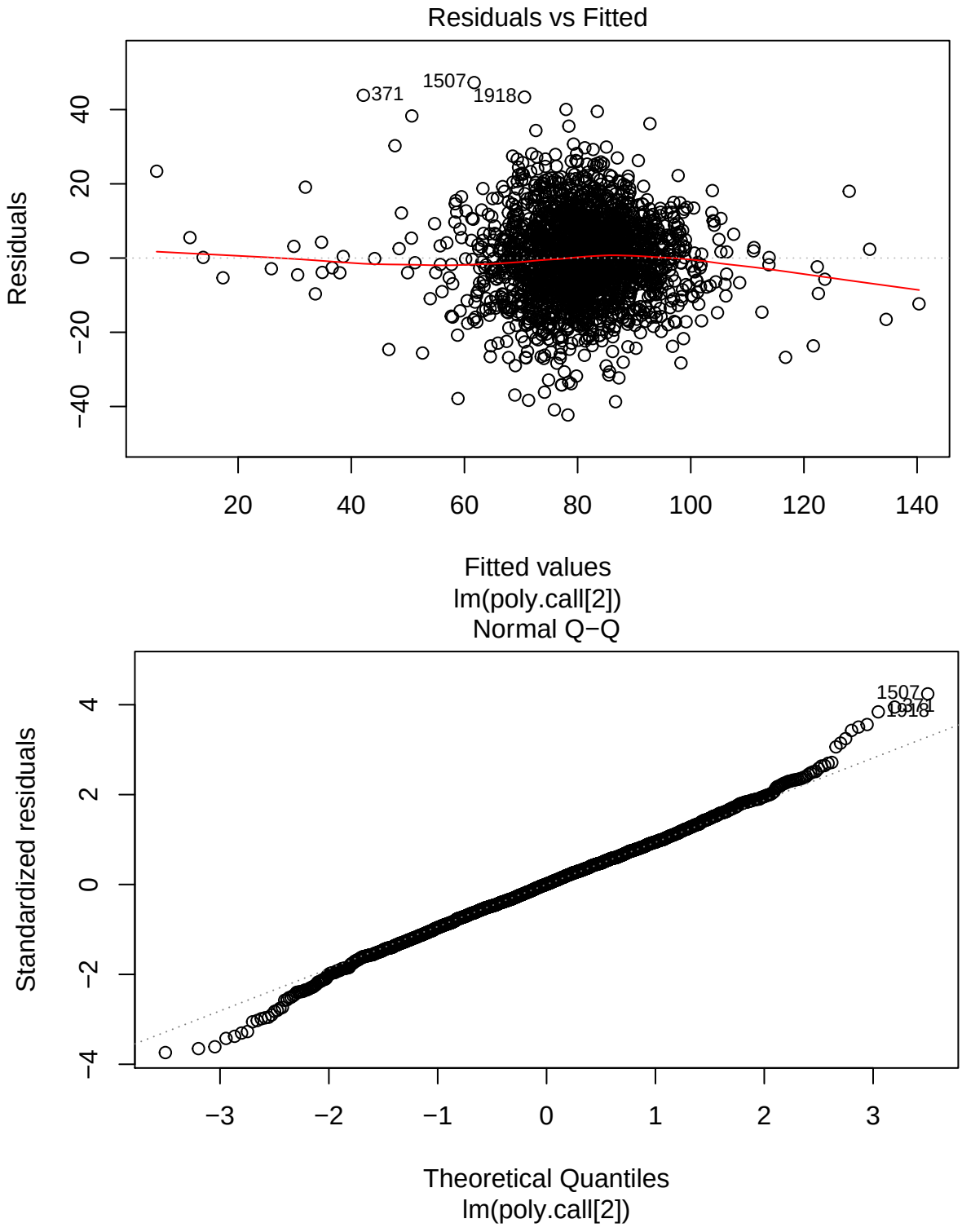
Final Model Review

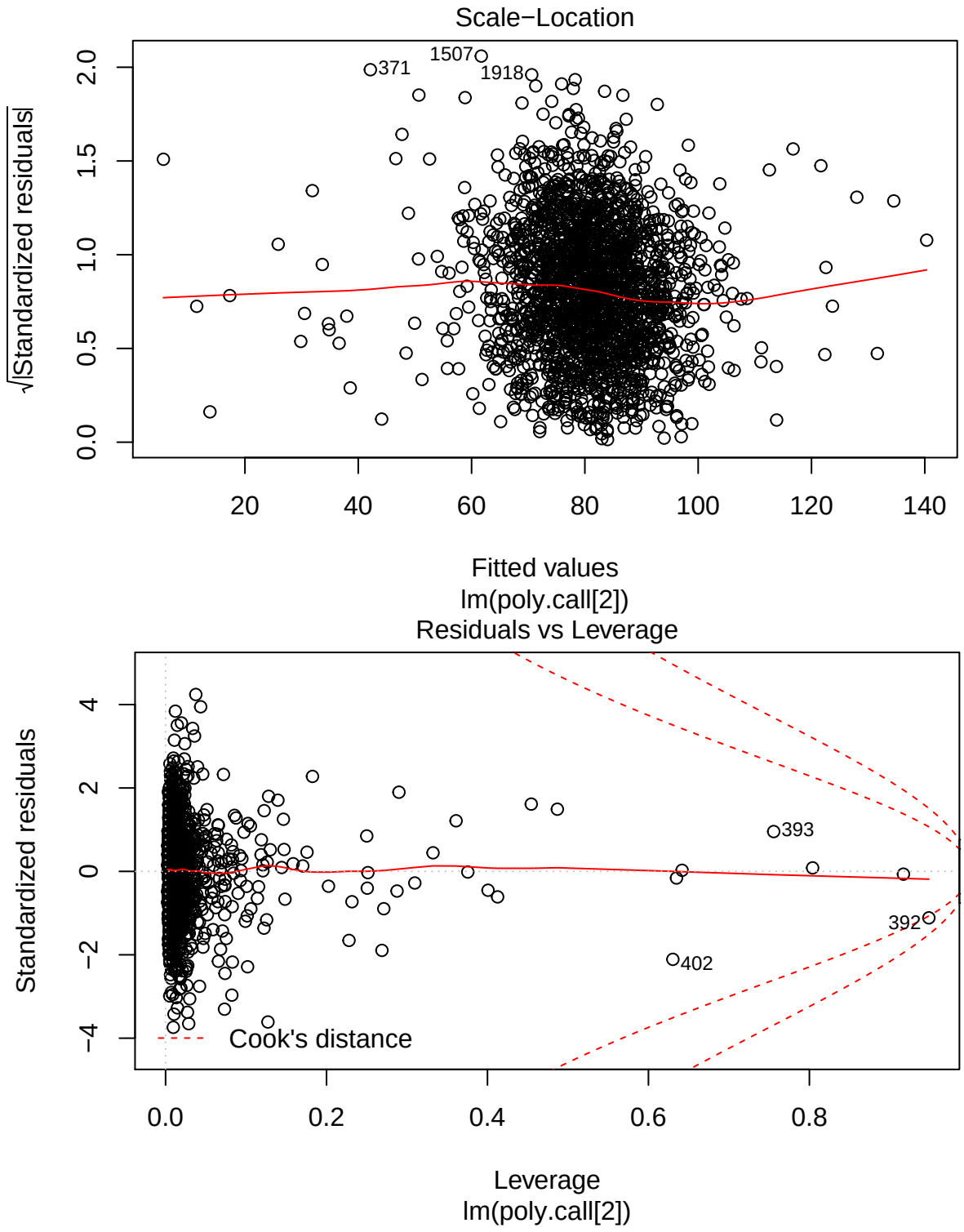
First let's review all the expected diagnostics of our final model.

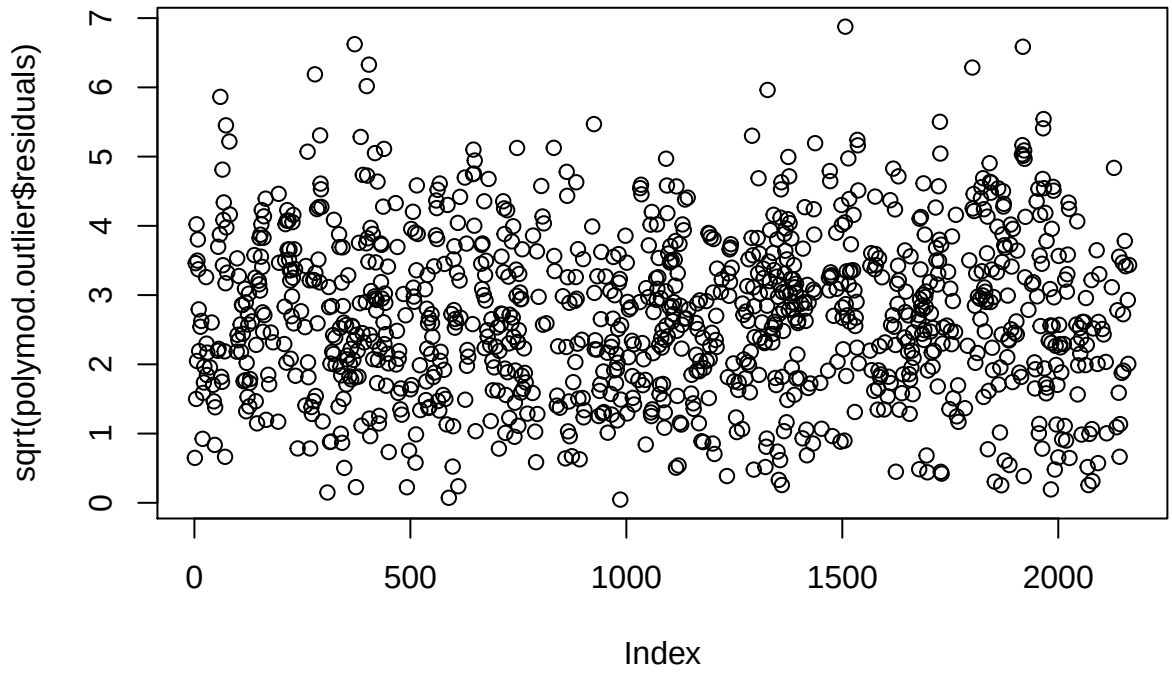
Our final model is:

```
## (TARGET_WINS ~ BATTING_BB + BATTING_SO + BASERUN_CS + PITCHING_HR +
##     PITCHING_BB + PITCHING_SO + FIELDING_E + FIELDING_DP + I(BATTING_2B^2) +
##     I(BATTING_3B^2) + I(BATTING_BB^2) + I(BATTING_SO^2) + I(BASERUN_SB^2) +
##     I(BASERUN_CS^2) + I(PITCHING_H^2) + I(PITCHING_BB^2) + I(PITCHING_SO^2) +
##     I(FIELDING_E^2) + I(FIELDING_DP^2) + I(BATTING_1B^2) + I(BASERUN_StealPer^2) +
##     I(BATTING_2B^3) + I(BATTING_3B^3) + I(BATTING_HR^3) + I(BATTING_SO^3) +
##     I(BASERUN_CS^3) + I(PITCHING_H^3) + I(PITCHING_BB^3) + I(FIELDING_E^3) +
##     I(FIELDING_DP^3) + I(BATTING_2B^4) + I(BATTING_3B^4) + I(BATTING_HR^4) +
##     I(BATTING_BB^4) + I(BATTING_SO^4) + I(BASERUN_CS^4) + I(PITCHING_H^4) +
##     I(PITCHING_BB^4) + I(FIELDING_E^4) + I(FIELDING_DP^4) + I(BATTING_1B^4))()
```

Let's review the diagnostic plots and a plot of the residuals.


Residuals vs Fitted
371
1507
1918
Residuals
40
20
0
−20
−40
20
40
60
80
100
120
140
Fitted values
lm(poly.call[2])
Normal Q−Q
1507
Standardized residuals
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(poly.call[2])

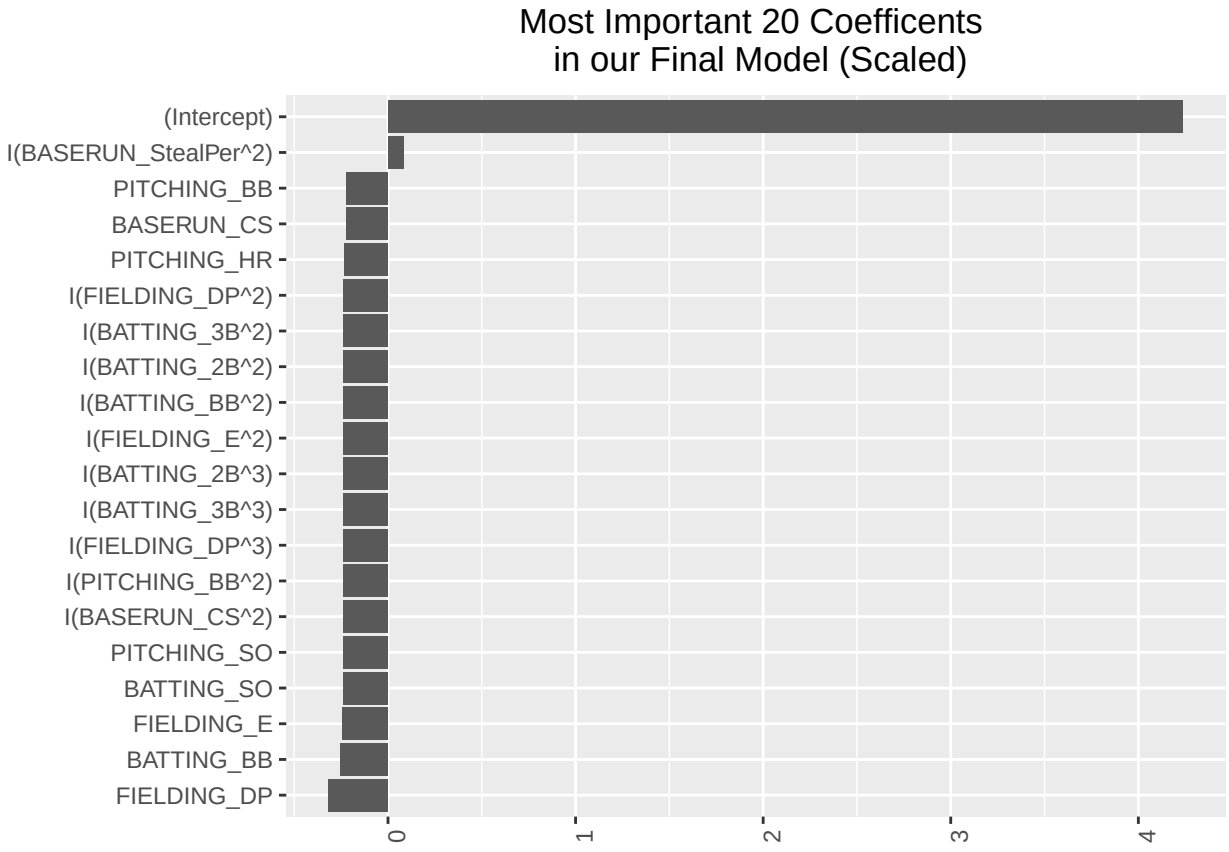

Scale−Location
√|Standardized residuals|
371
1507
1918
0.0
0.5
1.0
1.5
2.0
20
40
60
80
100
120
140
Fitted values
lm(poly.call[2])
Residuals vs Leverage
Standardized residuals
393
392
402
Cook's distance
1
0.5
0.5
1
−4
−2
0
2
4
0.0
0.2
0.4
0.6
0.8
Leverage
lm(poly.call[2])



Everything looks to be in scope here, since we removed individual observations that were skewing our diagnostics back in section 3.

Plot the top Coefficients of our model

What's interesting here and may point to a poor model is that essentially our intercept coefficient gives each observation 185.1 wins, and then most other coefficients subtract from there. For visual ease, the coefficients below have been scaled.

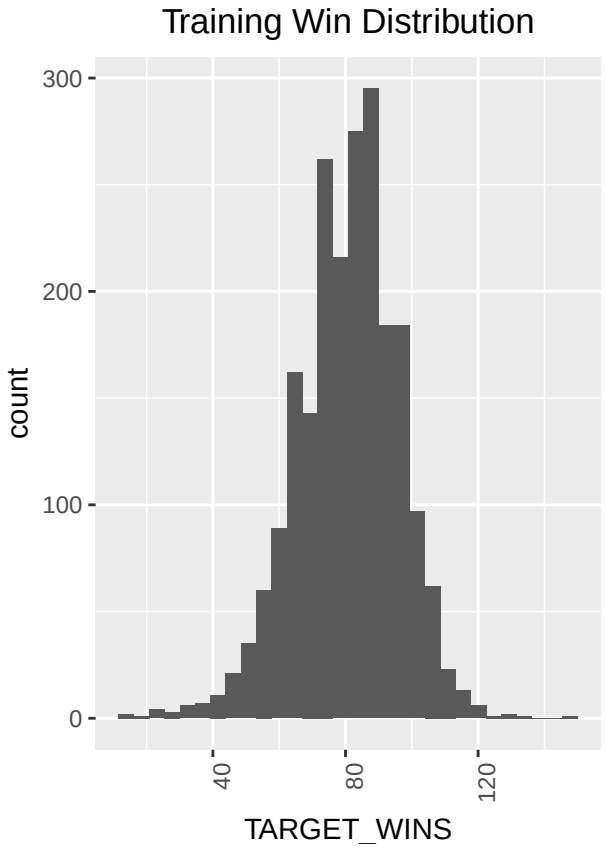


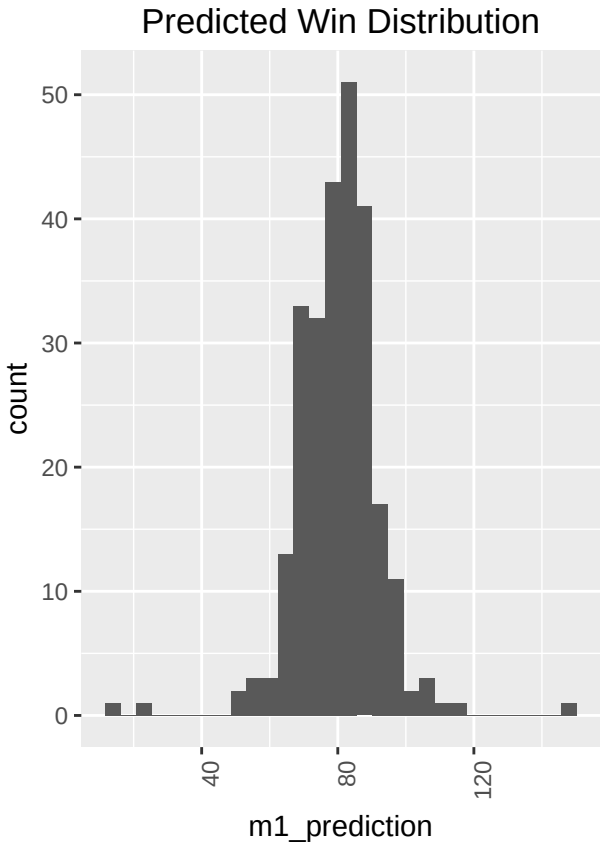
Predictions

We had to modify our predictions a bit because our final model a) predicted wins > 260 for one observation and b) -783 wins for another. This is clearly poor performance and it may be important to find better options for our model.

For now, we simply modify these outlier observations so those maxs and mins are replaced with the maxes and mins of our final training set.

Compare predicted to original distribution





Appendix

- For full output code visit: LINK
- For predicted values over test set visit: LINK