CUNY DATA 621 HW1: Moneyball

Group 2: Elina Azrilyan, Charls Joseph, Mary Anna Kivenson, Sunny Mehta, Vinayak Patel

March 01, 2020

Introduction

In this project we will examine a dataset of baseball team data from 1871 to 2006. Our goal is to build a multivariate regression model capable of predicting Wins out of sample. We will explore and prepare the data, build and select models, and discuss commentary and conclusions along the way.

Data Exploration

Read Data

Here, we read the dataset and shorten the feature names for better readibility in visualizations.

```
##   WINS bt_H bt_2B bt_3B bt_HR bt_BB bt_SO br_SB br_CS bt_HBP ph_H ph_HR
## 1   39 1445   194    39    13   143   842    NA    NA     NA 9364    84
## 2   70 1339   219    22   190   685  1075    37    28     NA 1347   191
## 3   86 1377   232    35   137   602   917    46    27     NA 1377   137
## 4   70 1387   209    38    96   451   922    43    30     NA 1396    97
## 5   82 1297   186    27   102   472   920    49    39     NA 1297   102
## 6   75 1279   200    36    92   443   973   107    59     NA 1279    92
##   ph_BB ph_SO fd_E fd_DP
## 1   927  5456 1011    NA
## 2   689  1082  193   155
## 3   602   917  175   153
## 4   454   928  164   156
## 5   472   920  138   168
## 6   443   973  123   149
```

Summary

First, we take a look at a summary of the data. A few things of interest are revealed:

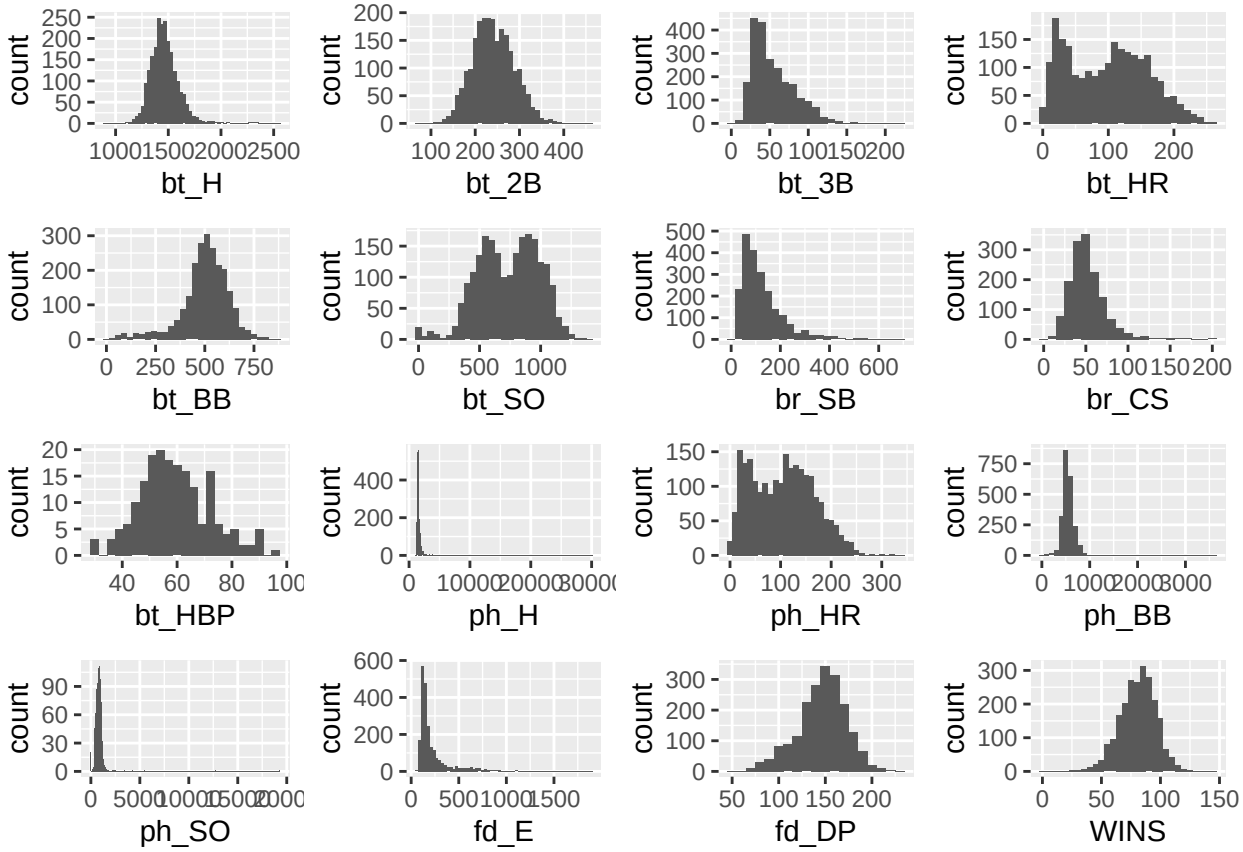
- bt_SO, br_SB, br_CS, bt_HBP, ph_SO, and fd_DP have missing values
- The max values of ph_H, ph_BB, ph_SO, and fd_E seem abnormally high

```
##       WINS             bt_H          bt_2B           bt_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  
##                                                                  
##      bt_HR            bt_BB           bt_SO            br_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    
##      br_CS           bt_HBP           ph_H           ph_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                                   
##      ph_BB            ph_SO             fd_E            fd_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    
```

Histogram

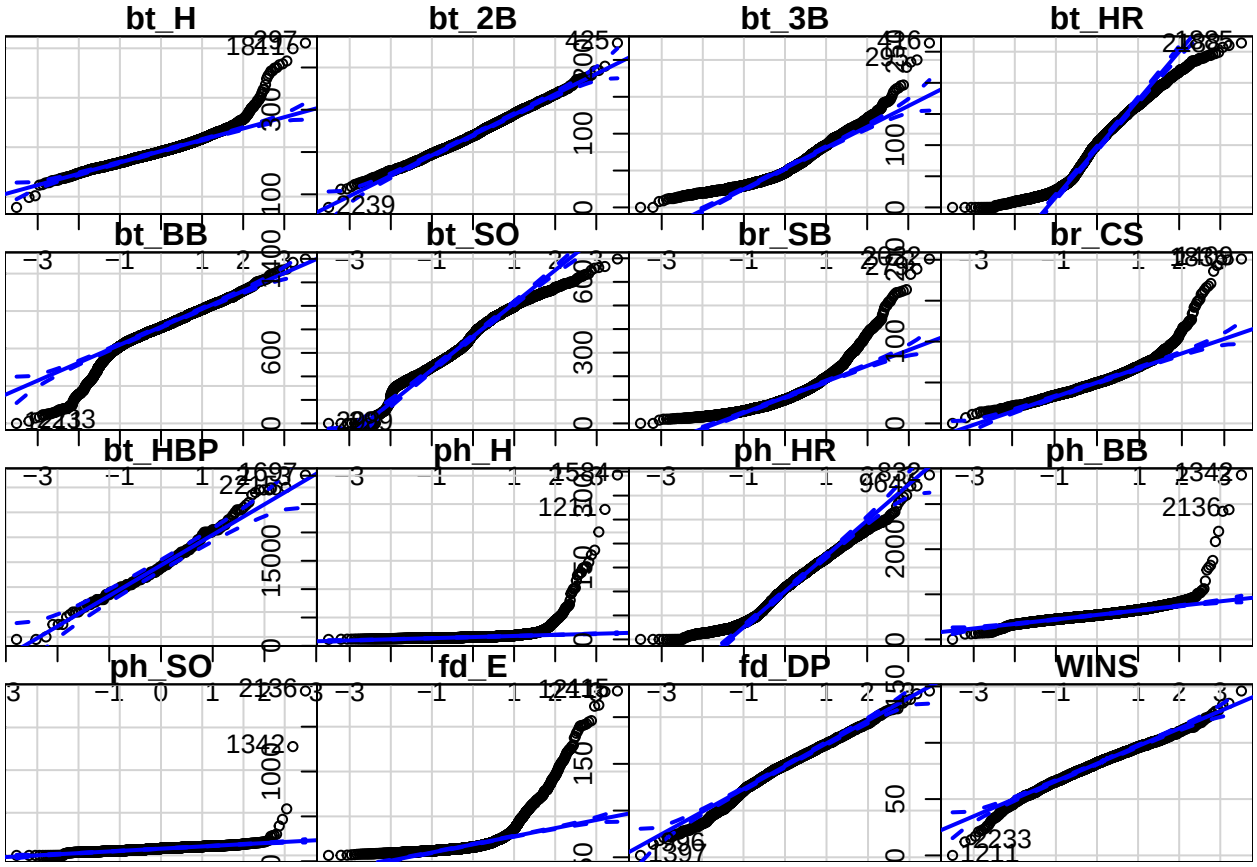
Next, we create histograms of each of the features and target variable.

- bt_H, bt_2B, bt_BB, br_CS, bt_HBP, fd_DP, WINS all have normal distributions
- ph_H, ph_BB, ph_SO, and fd_E are highly right-skewed



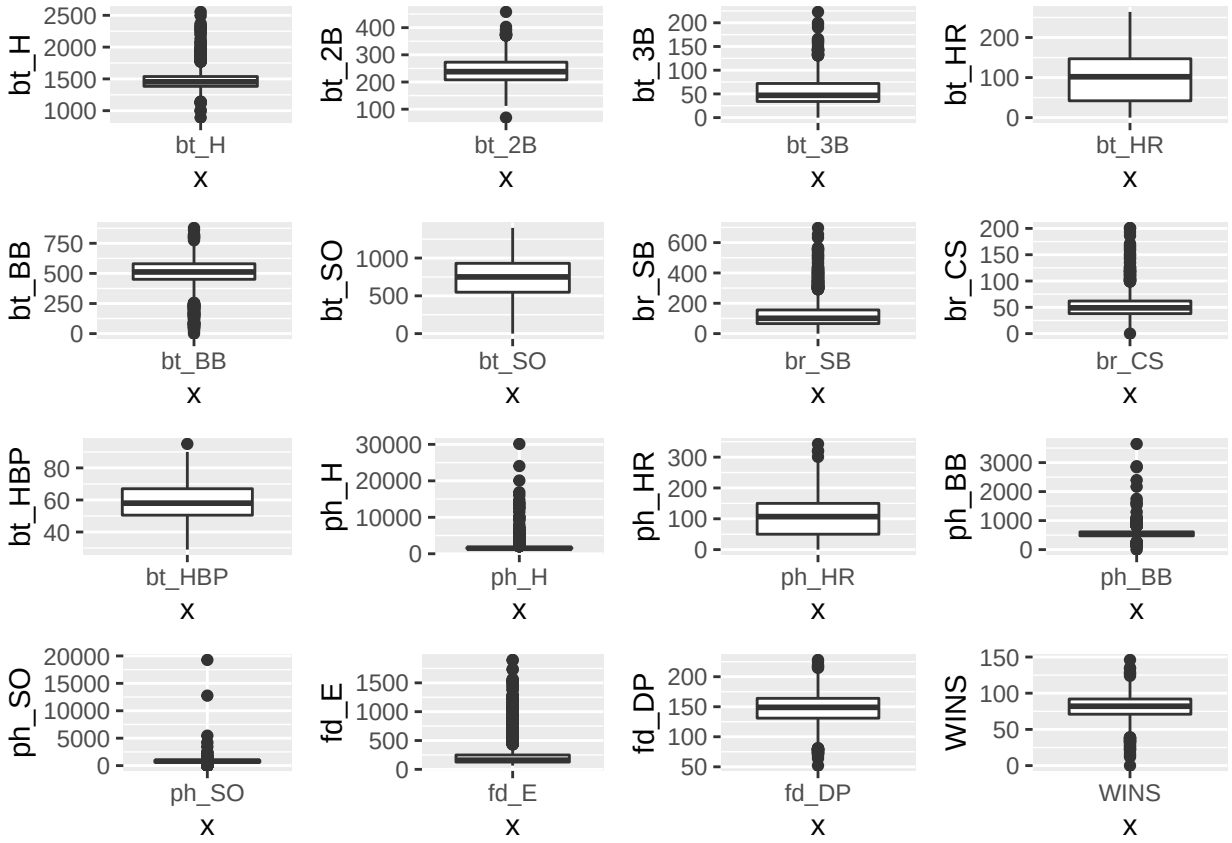
QQ Plots

- Most of the features are not lined up with the theoretical QQ plot, however this will be addressed by the models we build.



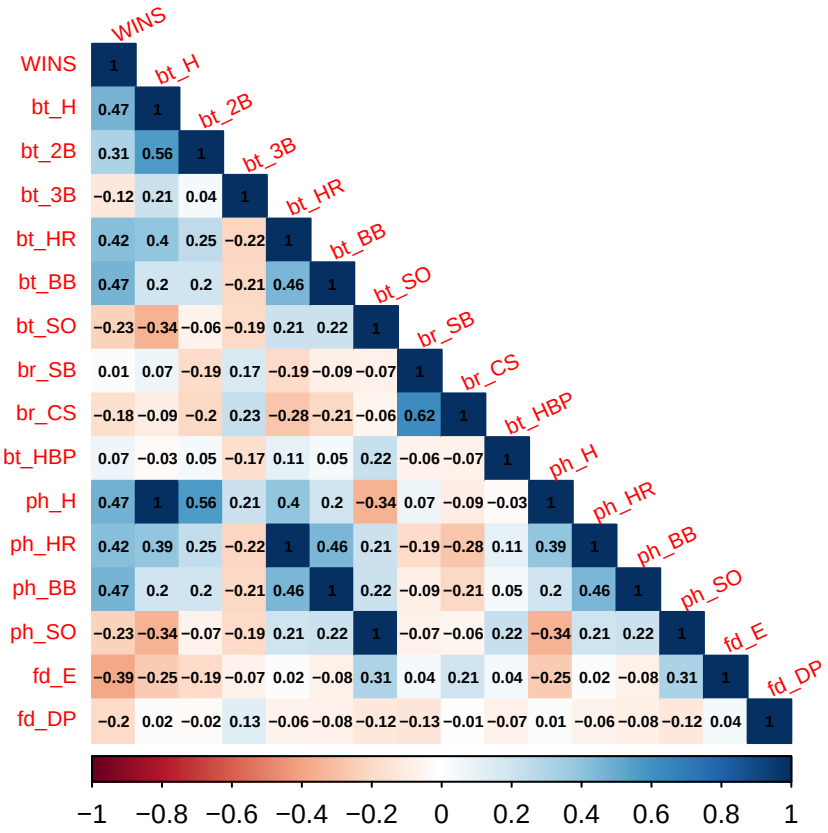
Boxplot

- Most of the boxplots shown below reflect a long right tail with many outliers.



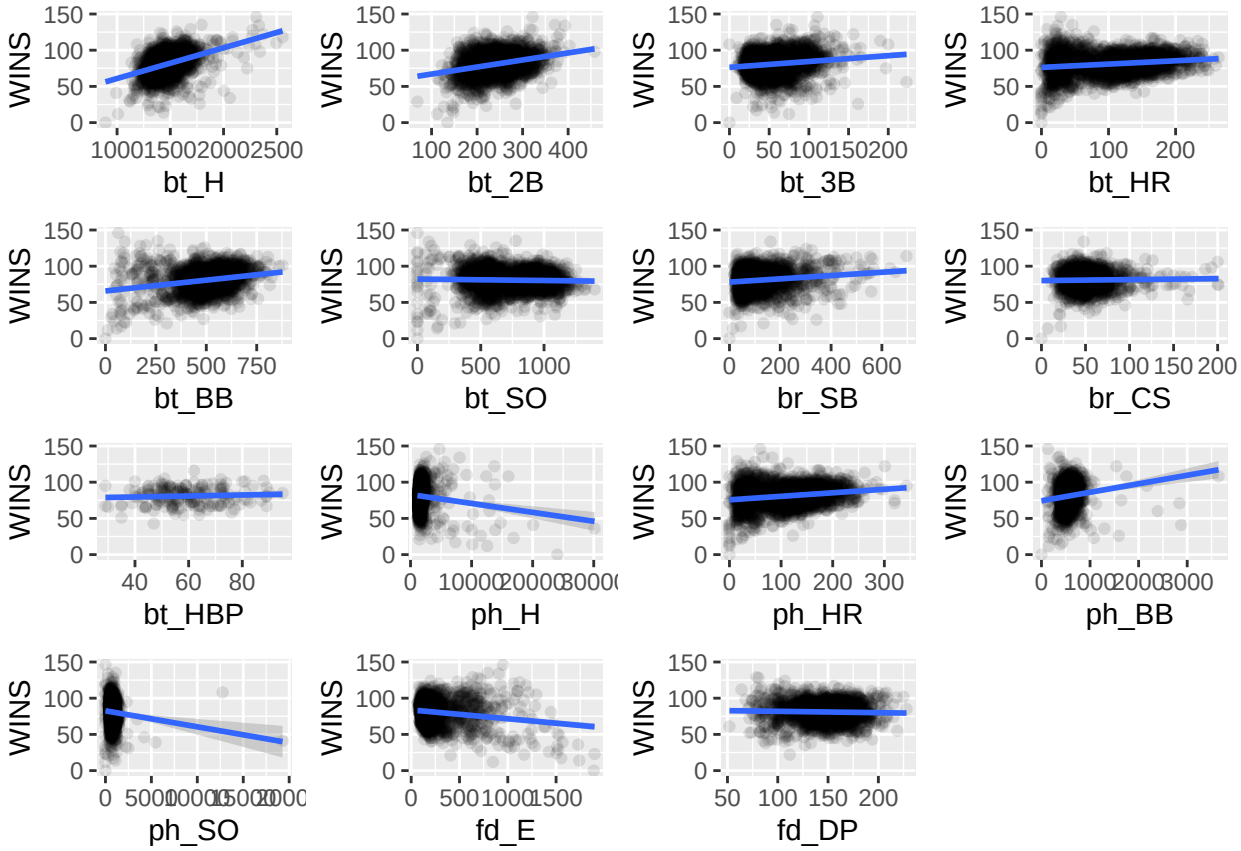
Correlation Plot

- There is a strong positive correlation between ph_H and bt_H
- There is a strong positive correlation between ph_HR and bt_HR
- There is a strong positive correlation between ph_BB and bt_BB
- There is a strong positive correlation between ph_SO and bt_SO
- There seems to be a weak correlation between bt_HBP/br_SB and Wins



Scatter Plots

Here, we see a scatter plot of each of the feature variables with the target variable.



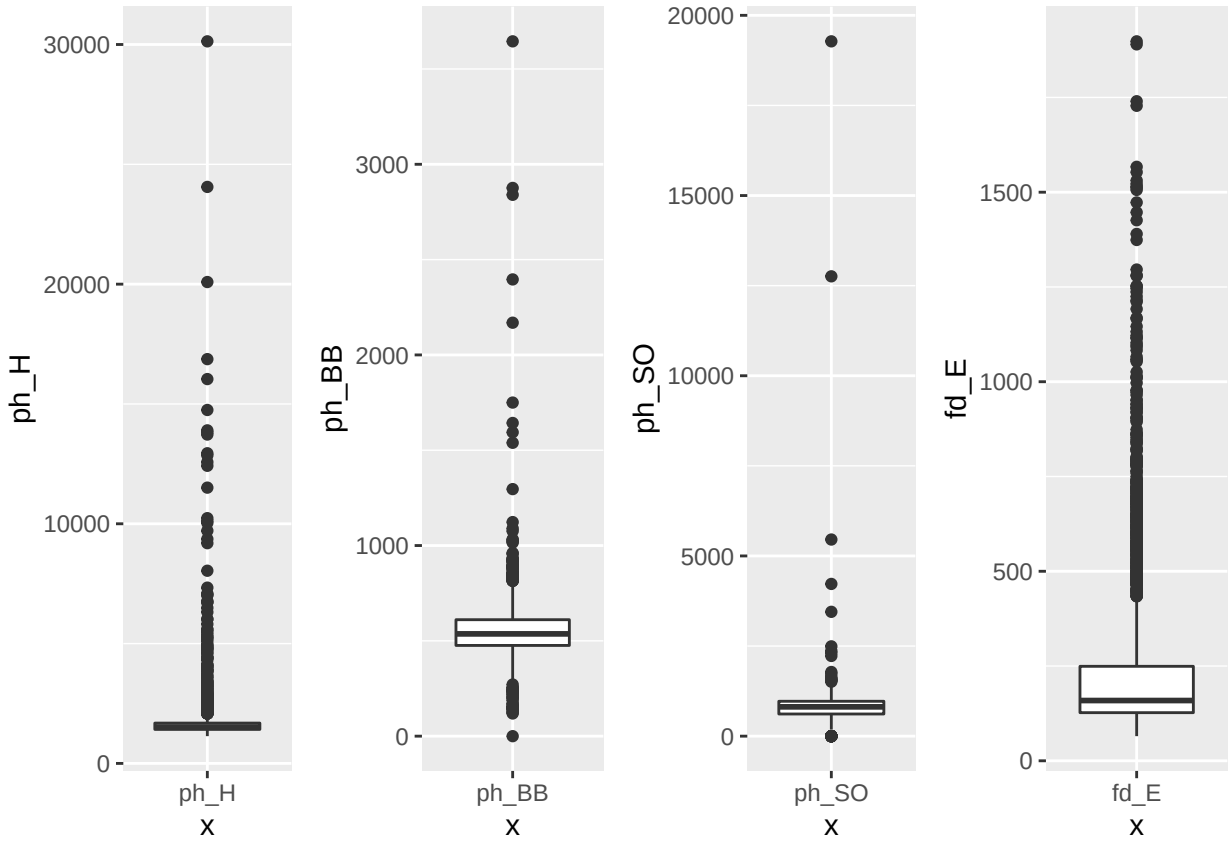
Data Preparation

Outliers

Extreme Values

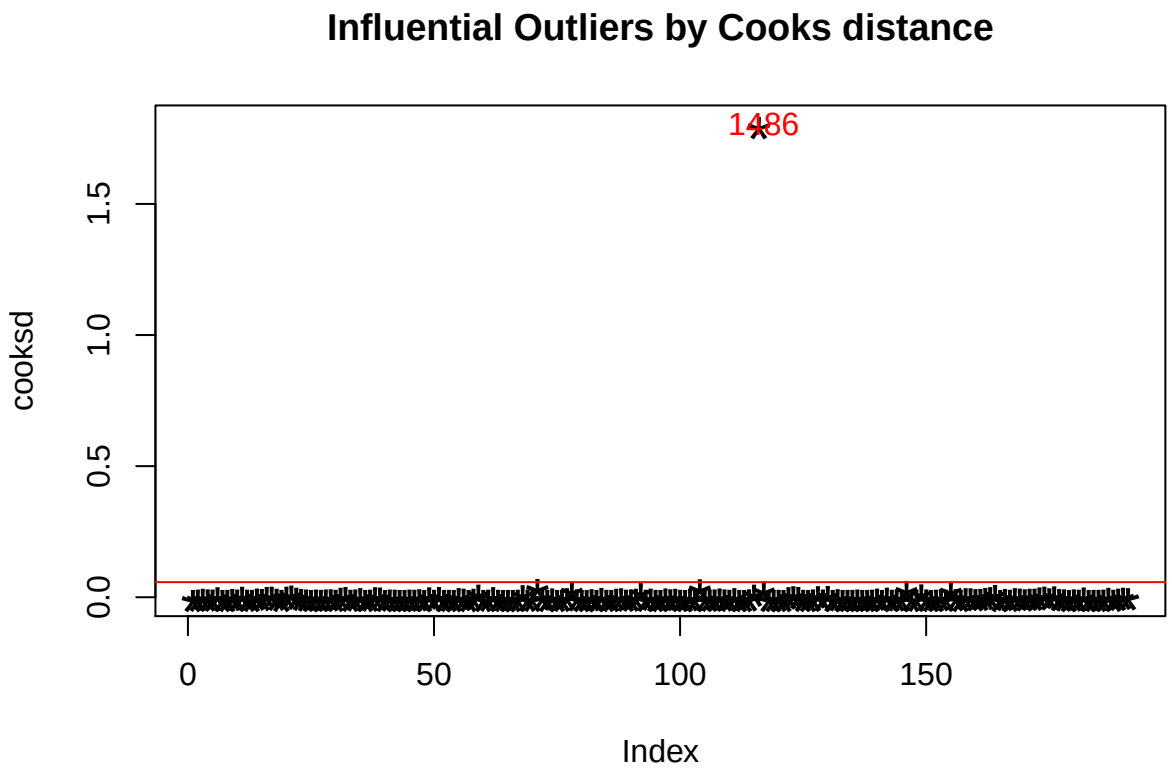
While exploring the data, we noticed that the max values of ph_H, ph_BB, ph_SO, and fd_E seem abnormally high.

We see that the record for most hits in a season by team (ph_H) was set at 1,724 in 1921. However, we also know that the datapoints were normalized for 162 games in a season. To take a moderate approach, we will remove the some of the most egggregious outliers that are seen in these variables.



Cooks Distance

We will also remove influential outliers using Cooks distance.



We remove the influential outliers.

Fill Missing Values

The following features have missing values.

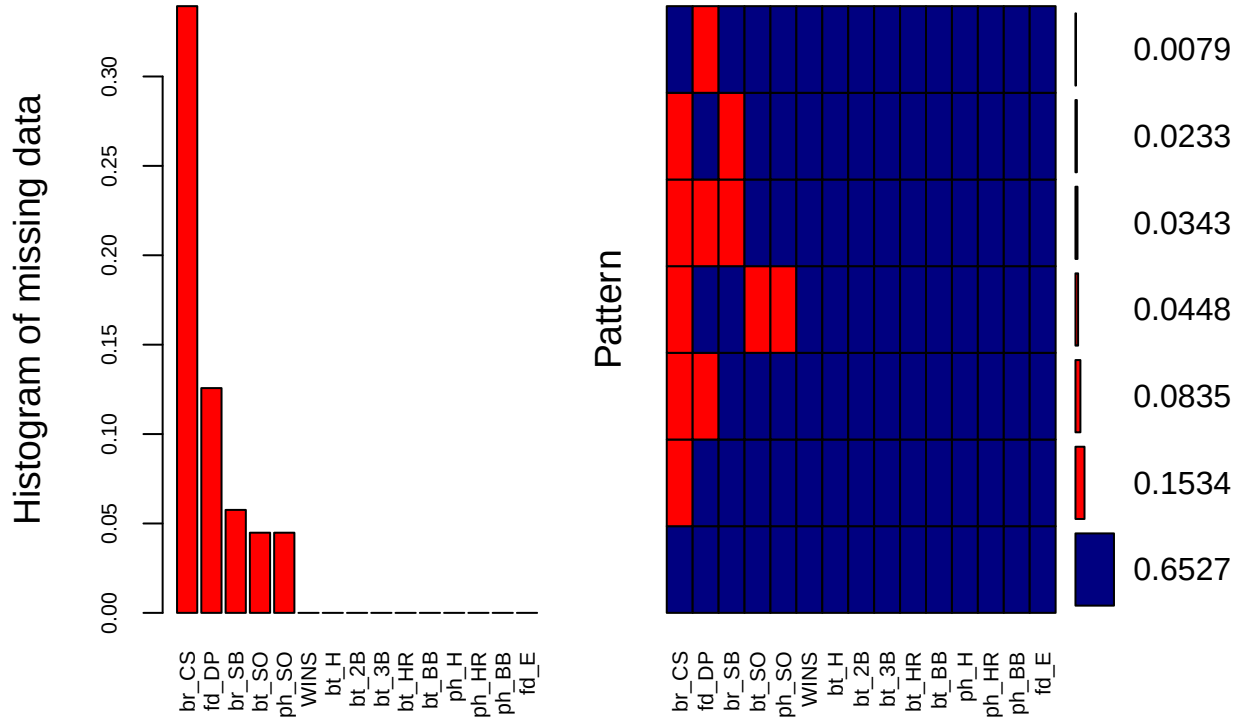
- bt_SO - Strikeouts by batters
- br_SB - Stolen bases
- br_CS - Caught stealing
- bt_HBP - Batters hit by pitch (get a free base)

- ph_SO - Strikeouts by pitchers
- fd_DP - Double Plays

Since most values in bt_HBP are missing (90%), we will drop this feature.

Multivariate Imputation by Chained Equations (mice)

We will use Multivariable Imputation by Chained Equations (mice) to fill the missing variables.




```
##
##  Variables sorted by number of missings:
```

```
##  Variable      Count
##     br_CS 0.33934066
##     fd_DP 0.12571429
##     br_SB 0.05758242
##     bt_SO 0.04483516
##     ph_SO 0.04483516
##      WINS 0.00000000
##      bt_H 0.00000000
##     bt_2B 0.00000000
##     bt_3B 0.00000000
##     bt_HR 0.00000000
##     bt_BB 0.00000000
##      ph_H 0.00000000
##     ph_HR 0.00000000
##     ph_BB 0.00000000
##      fd_E 0.00000000
```

Address Correlated Features

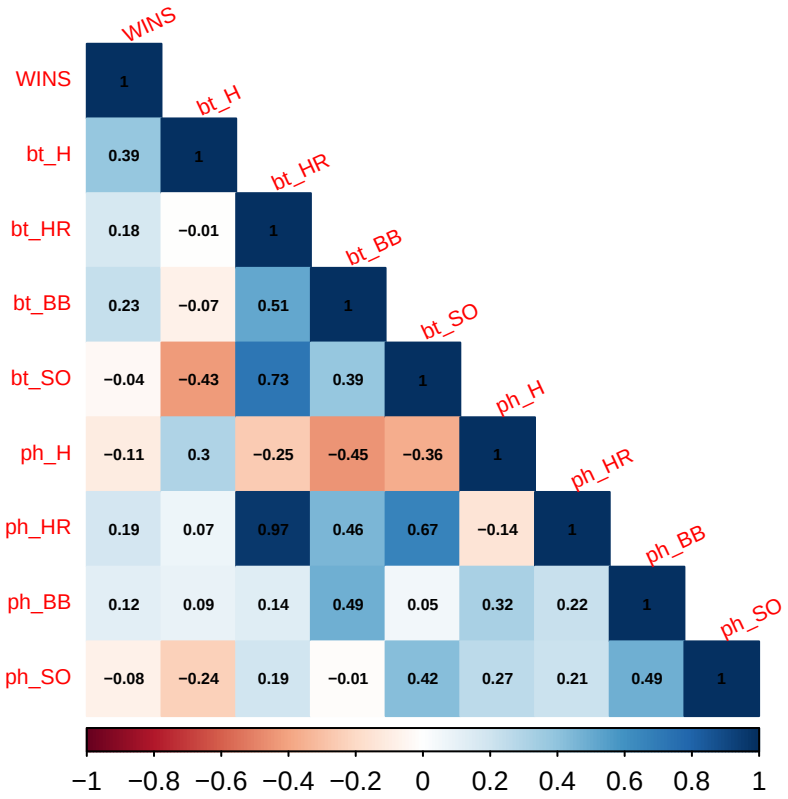
While exploring the data, we noticed several features had strong positive linear relationships.

Let's run a Variance Inflation Factor test to detect multicollinearity. Features with a VIF score > 10 will be reviewed.

```
##      bt_H     bt_2B     bt_3B     bt_HR     bt_BB     bt_SO     br_SB
##  3.816108  2.471669  2.949210 36.587699  6.786885  5.272760  3.747356
##     br_CS      ph_H     ph_HR     ph_BB     ph_SO      fd_E     fd_DP
##  4.007571  4.180244 29.607551  6.301651  3.372521  5.375395  1.990500
```

Let's make another correlation plot with only these features.

- bt_SO (strikeouts by batters) and bt_H (base hits by batters) have a strong positive correlation
- bt_H (base hits by batters) and bt_BB (walks by batters) have a strong positive correlation
- ph_BB (walks allowed) and bt_BB (walks by batters) have a strong negative correlation
- ph_SO (strikeouts by pitchers) and bt_SO (strikeouts by batters) have a moderate negative correlation
- ph_HR (homeruns allowed) and bt_HR (homeruns by batters) have a strong negative correlation
- ph_SO (strikeouts by pitchers) and ph_BB (walks allowed) have a moderate negative correation



To fix this, we can remove some correlated features and combine others.

- Remove bt_HR. It has an extremely strong correlation with ph_HR.
- Remove bt_SO. It has an extremely strong correlation with ph_SO.
- Replace bt_H (total base hits by batters) with BT_1B = bt_H - BT_2B - BT_3B - BT_HR (1B base hits)
- Replace ph_BB and bt_BB as a ratio of walks by batters to walks allowed

These adjustments result in less multicollinearity.

```
##    bt_2B    bt_3B    br_SB    br_CS     ph_H    ph_HR    ph_SO     fd_E
## 1.546254 2.304995 3.530792 3.973626 3.671210 2.324498 1.895449 7.103387
```

```
##    fd_DP    bt_1B       BB
## 1.940269 2.692229 5.736130
```

Create Output

```
##   TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B TEAM_BATTING_HR
## 1           1209             170              33              83
## 2           1221             151              29              88
## 3           1395             183              29              93
## 4           1539             309              29             159
## 5           1445             203              68               5
## 6           1431             236              53              10
##   TEAM_BATTING_BB TEAM_BATTING_SO TEAM_BASERUN_SB TEAM_BASERUN_CS
## 1             447            1080              62              50
## 2             516             929              54              39
## 3             509             816              59              47
## 4             486             914             148              57
## 5              95             416              NA              NA
## 6             215             377              NA              NA
##   TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB
## 1               NA            1209               83              447
## 2               NA            1221               88              516
## 3               NA            1395               93              509
## 4               42            1539              159              486
## 5               NA            3902               14              257
## 6               NA            2793               20              420
##   TEAM_PITCHING_SO TEAM_FIELDING_E TEAM_FIELDING_DP
## 1             1080              140             156
## 2              929              135             164
## 3              816              156             153
## 4              914              124             154
## 5             1123              616             130
## 6              736              572             105
```

Conclusions from Data Examination

Thoroughly understanding our data before building models is a crucial step. Utilizing domain knowledge and common sense can go a long way in aiding the prediction process. In calling upon baseball subject matter considerations here, we can make a few comments.

In general, the Pitching data looks strange, relative to the Batting data. A team's pitching and batting should be largely independent. And in theory we know pitchers giving up hits should be detrimental to winning, while batters getting hits should be beneficial. We observe the latter in the data, but not the former.

Also, the scale of some of the pitching variables are wildly different than we'd expect. It appears as though some of the pitching data, particularly Hits, has been drastically altered such that even after data cleaning, imputation, removal, and transformation, it may still not be of much use in a predictive model.

On the Batter and Baserunning side, the Stolen Bases, Caught Stealing, and Hit By Pitch variables are missing a lot of values. This may be detrimental to their utility in the model fitting process.

Build Models

Linear Model 1.

We will begin with all independent variables and use the back elimination method to eliminate the non-significant ones.

```
##
## Call:
## lm(formula = WINS ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.796  -8.302   0.294   8.189  60.610
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.981e+01  7.430e+00   9.396  < 2e-16 ***
## bt_H         4.211e-02  3.667e-03  11.482  < 2e-16 ***
## bt_2B       -2.299e-02  8.864e-03  -2.593  0.00956 **
## bt_3B        5.437e-02  1.646e-02   3.304  0.00097 ***
## bt_HR        1.694e-01  3.079e-02   5.501 4.19e-08 ***
## bt_BB        2.592e-02  6.044e-03   4.288 1.88e-05 ***
## bt_SO       -1.309e-02  2.487e-03  -5.262 1.56e-07 ***
## br_SB        4.711e-02  4.926e-03   9.563  < 2e-16 ***
## br_CS        1.761e-02  1.002e-02   1.757  0.07914 .
## ph_H         1.237e-03  4.387e-04   2.820  0.00485 **
## ph_HR       -8.339e-02  2.826e-02  -2.951  0.00320 **
```

```
## ph_BB       -1.087e-02  4.251e-03  -2.557  0.01062 *
## ph_SO        1.157e-03  8.905e-04   1.299  0.19410
## fd_E        -5.690e-02  3.413e-03 -16.675  < 2e-16 ***
## fd_DP       -1.021e-01  1.253e-02  -8.151 5.91e-16 ***
## bt_1B              NA         NA      NA       NA
## BB          -3.971e+01  5.805e+00  -6.841 1.01e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.48 on 2258 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3692, Adjusted R-squared:  0.365
## F-statistic: 88.09 on 15 and 2258 DF,  p-value: < 2.2e-16
```

We will start by eliminating the variables with high p-values and lowest significance from the model

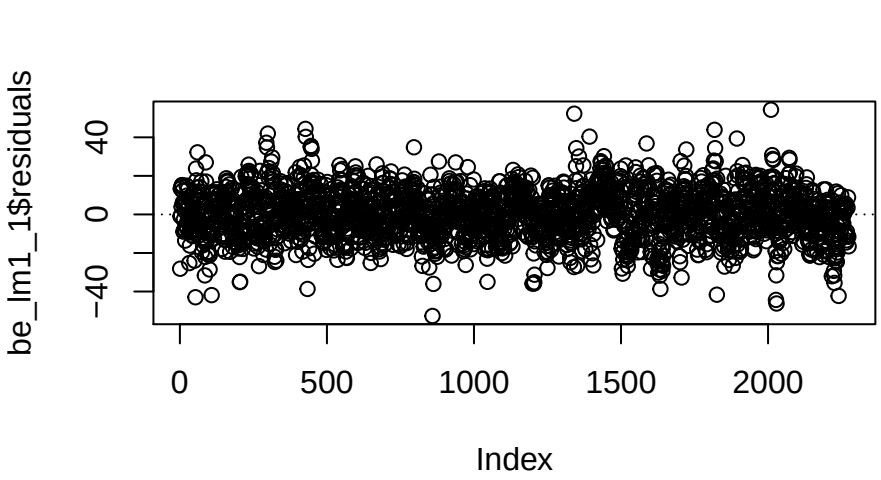
Let's take a look at the resulting model:

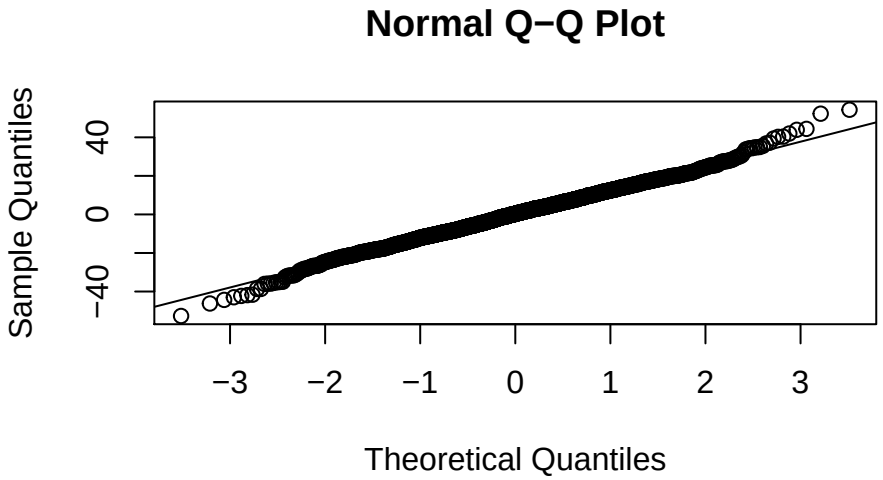
```
##
## Call:
## lm(formula = WINS ~ bt_H + bt_SO + br_SB + ph_HR + fd_E + fd_DP +
##     BB, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.679  -8.588   0.313   8.416  54.309
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  53.885683   6.074116   8.871  < 2e-16 ***
## bt_H          0.041010   0.002515  16.308  < 2e-16 ***
## bt_SO        -0.014202   0.002081  -6.826 1.12e-11 ***
## br_SB         0.054109   0.003802  14.231  < 2e-16 ***
## ph_HR         0.065946   0.007995   8.248 2.70e-16 ***
## fd_E         -0.047803   0.002895 -16.512  < 2e-16 ***
## fd_DP        -0.093645   0.011954  -7.834 7.22e-15 ***
## BB          -13.493559   3.846392  -3.508  0.00046 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 12.67 on 2266 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3475, Adjusted R-squared:  0.3455
## F-statistic: 172.4 on 7 and 2266 DF,  p-value: < 2.2e-16
```

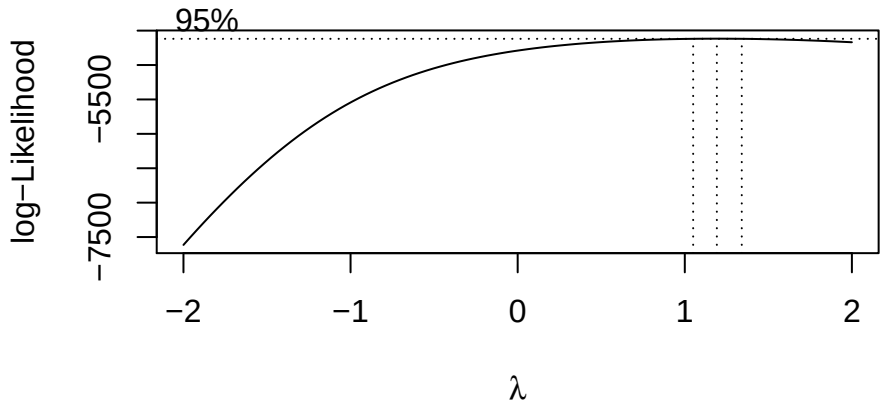
We are seeing high significance indicators and p-values of 0 across all 10 remaining variables, however our R squared value is rather low - 36.

The next step is to check residuals plot and QQ plot to check the validity of our model.





Both of these plots show that the model is a reasonable model. There is no pattern evident in the residuals and normality assumptions is close enough, even though there are some outliers.



We are going to use Box-Cox transformation to determine if a transformation is required.

Lambda is close to 1, so no transformation is needed.

Linear Model 2.

This Linear Model will be built using the variables we believe would have the highest corelation with WINs.

THe following variables will be used: - Base Hits by batters (1B,2B,3B,HR) - Walks by batters - Stolen bases - Strikeouts by batters - Errors - Strikeouts by pitchers - Double Plays - Hits allowed

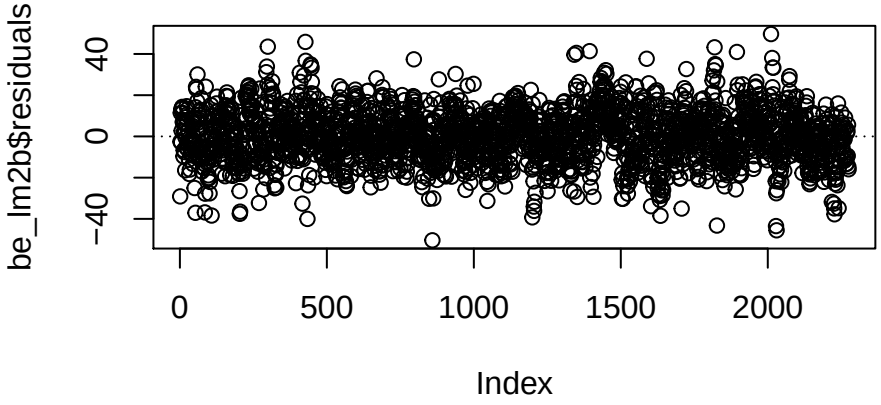
```
##
## Call:
## lm(formula = WINS ~ bt_H + bt_BB + br_SB + bt_SO + fd_E + ph_SO +
##     fd_DP + ph_H, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.280  -8.532   0.195   8.472  51.026
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

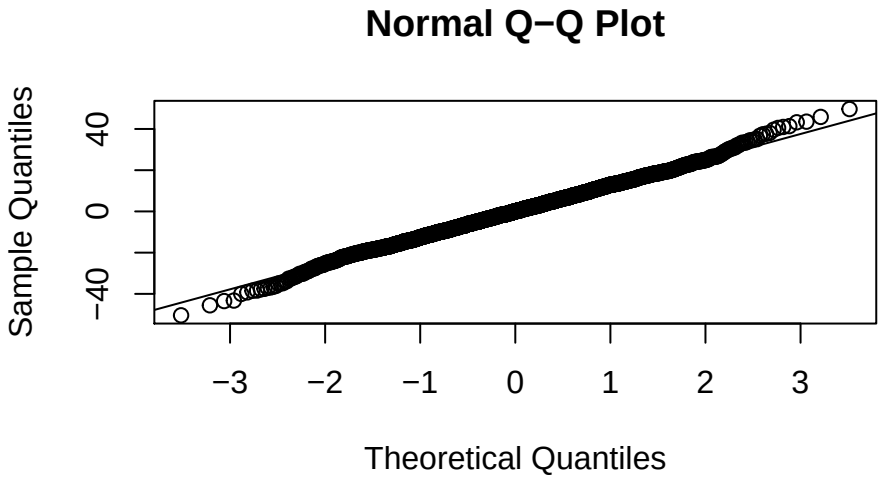
```
## (Intercept) 15.3478849  4.0521441   3.788 0.000156 ***
## bt_H         0.0503162  0.0022029  22.841  < 2e-16 ***
## bt_BB        0.0149574  0.0029804   5.019 5.61e-07 ***
## br_SB        0.0461432  0.0040190  11.481  < 2e-16 ***
## bt_SO       -0.0031153  0.0016300  -1.911 0.056111 .
## fd_E        -0.0404161  0.0027066 -14.933  < 2e-16 ***
## ph_SO        0.0003777  0.0006765   0.558 0.576744
## fd_DP       -0.0887691  0.0124373  -7.137 1.28e-12 ***
## ph_H         0.0012386  0.0003394   3.650 0.000268 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.84 on 2266 degrees of freedom
## Multiple R-squared:  0.3378, Adjusted R-squared:  0.3355
## F-statistic: 144.5 on 8 and 2266 DF,  p-value: < 2.2e-16
```

Let's remove the two variables with low significance:

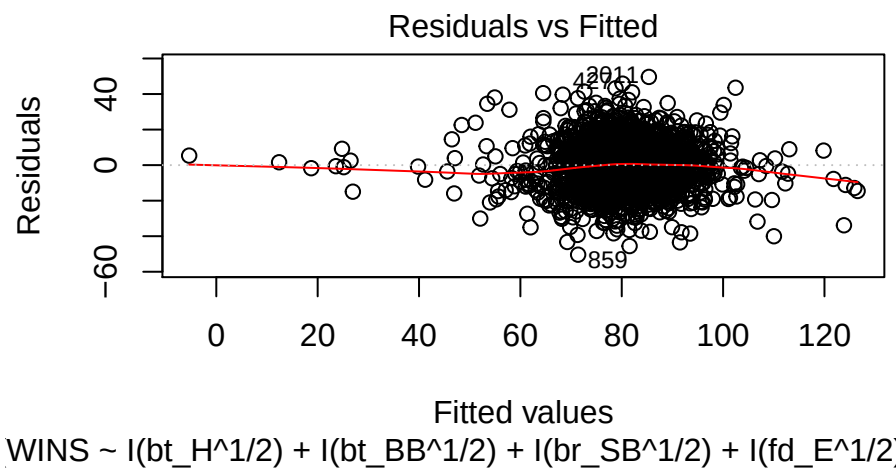
```
##
## Call:
## lm(formula = WINS ~ I(bt_H^1/2) + I(bt_BB^1/2) + I(br_SB^1/2) +
##     I(fd_E^1/2) + I(fd_DP^1/2) + I(ph_H^1/2), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.397  -8.582   0.048   8.357  49.640
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.5848707  3.4321787   3.375  0.00075 ***
## I(bt_H^1/2)   0.1031101  0.0039836  25.884  < 2e-16 ***
## I(bt_BB^1/2)  0.0292814  0.0059503   4.921 9.23e-07 ***
## I(br_SB^1/2)  0.0917733  0.0075812  12.105  < 2e-16 ***
## I(fd_E^1/2)  -0.0779180  0.0049657 -15.691  < 2e-16 ***
## I(fd_DP^1/2) -0.1809329  0.0246963  -7.326 3.27e-13 ***
## I(ph_H^1/2)   0.0025417  0.0005909   4.301 1.77e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

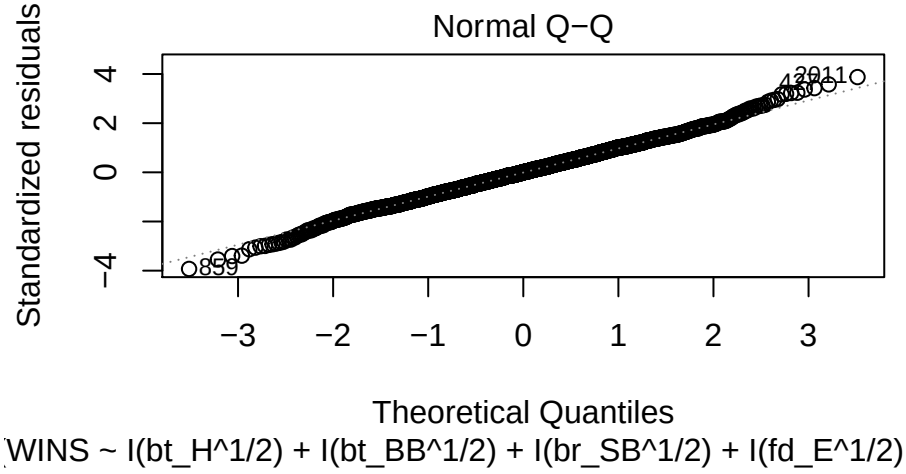
```
## Residual standard error: 12.85 on 2268 degrees of freedom
## Multiple R-squared:  0.3367, Adjusted R-squared:  0.335
## F-statistic: 191.9 on 6 and 2268 DF,  p-value: < 2.2e-16
```

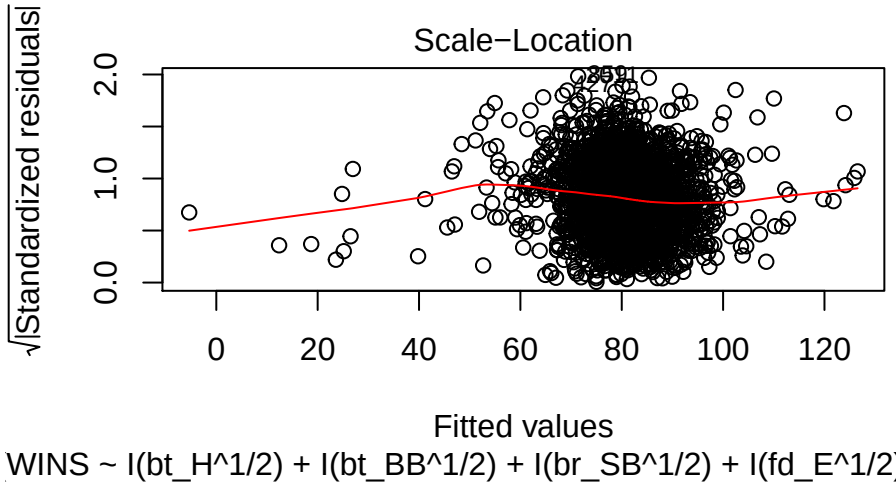


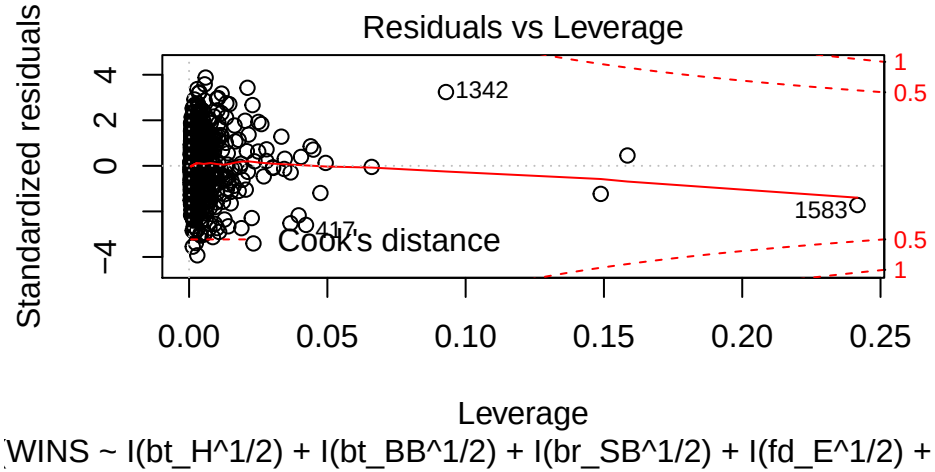



```
##   bt_H bt_2B bt_3B bt_HR bt_BB bt_SO br_SB br_CS bt_HBP ph_H ph_HR ph_BB
## 1 1209   170     33    83   447  1080    62    50     NA 1209    83   447
## 2 1221   151     29    88   516   929    54    39     NA 1221    88   516
## 3 1395   183     29    93   509   816    59    47     NA 1395    93   509
## 4 1539   309     29   159   486   914   148    57     42 1539   159   486
## 5 1445   203     68     5    95   416    NA    NA     NA 3902    14   257
## 6 1431   236     53    10   215   377    NA    NA     NA 2793    20   420
##   ph_SO fd_E fd_DP
## 1  1080  140   156
## 2   929  135   164
## 3   816  156   153
## 4   914  124   154
## 5  1123  616   130
## 6   736  572   105
```


Residuals vs Fitted
Residuals
Fitted values
lm(WINS ~ I(bt_H^1/2) + I(bt_BB^1/2) + I(br_SB^1/2) + I(fd_E^1/2) + I(
2011
859


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(WINS ~ I(bt_H^1/2) + I(bt_BB^1/2) + I(br_SB^1/2) + I(fd_E^1/2) + I(
2011
859


Scale−Location
√|Standardized residuals|
Fitted values
lm(WINS ~ I(bt_H^1/2) + I(bt_BB^1/2) + I(br_SB^1/2) + I(fd_E^1/2) + I(

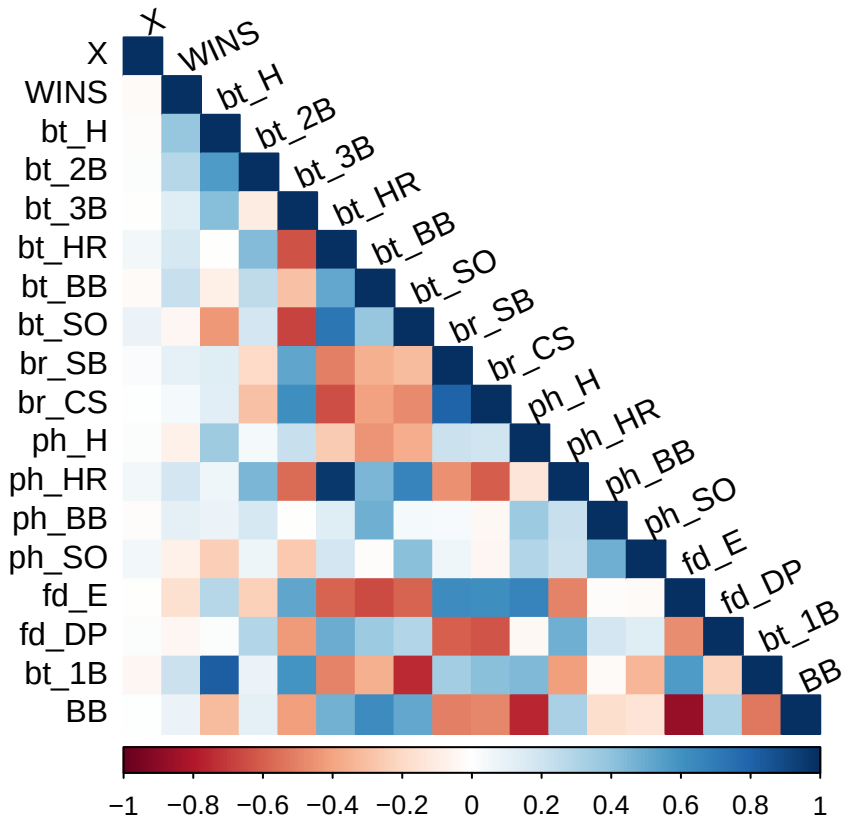

Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
lm(WINS ~ I(bt_H^1/2) + I(bt_BB^1/2) + I(br_SB^1/2) + I(fd_E^1/2) + I(
1342
1583
417

Further Model Building and Model Selection

Load the dataset

Load the data set that was curated after the preliminary explanatory analysis.

Plotted a correlation matrix on the original data set



Found that there is a data point with 0 value which needs to be corrected to a non-zero val for BOX-COX transformation later

```
##         X WINS bt_H bt_2B bt_3B bt_HR bt_BB bt_SO br_SB br_CS  ph_H ph_HR
## 1211 1211    0  891    135     0     0     0     0     0     0 24057     0
```

```
##      ph_BB ph_SO fd_E fd_DP bt_1B BB
## 1211     0     0 1890   219   756 NA
```

Remove the index column that got added in the preliminary step. Also lift each datapoints by 1 to fix the zero data point.

Split the data into training and test data set(80% training and 20% testing)

Assumption of Ordinary Least square regression

Before building and trying out different linear regression models, we will review the assumptions for the OLS algorithm to make it perform well.

1. Residual Error should be normally distributed
2. Absence of hetroschdasticity
3. Absence of Colinearity

We will check these assumption/factors while reviewing the results of each model.

Full Model

Fitting a full model with all remaining independent variables(“bt_H” “bt_2B” “bt_3B” “bt_HR” “bt_BB” “bt_SO” “br_SB” “br_CS” “ph_H” “ph_HR” “ph_BB” “ph_SO” “fd_E” “fd_DP” “bt_1B” “BB”) and the response variable WINS

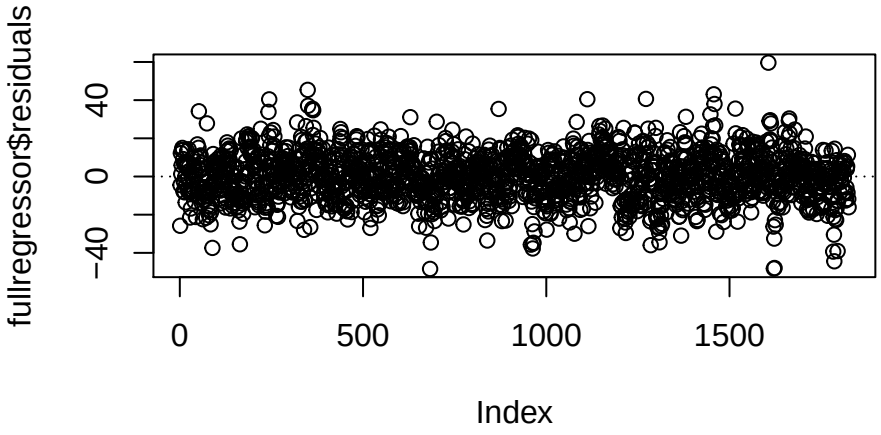
```
##  [1] "WINS"  "bt_H"  "bt_2B" "bt_3B" "bt_HR" "bt_BB" "bt_SO" "br_SB"
##  [9] "br_CS" "ph_H"  "ph_HR" "ph_BB" "ph_SO" "fd_E"  "fd_DP" "bt_1B"
## [17] "BB"
```

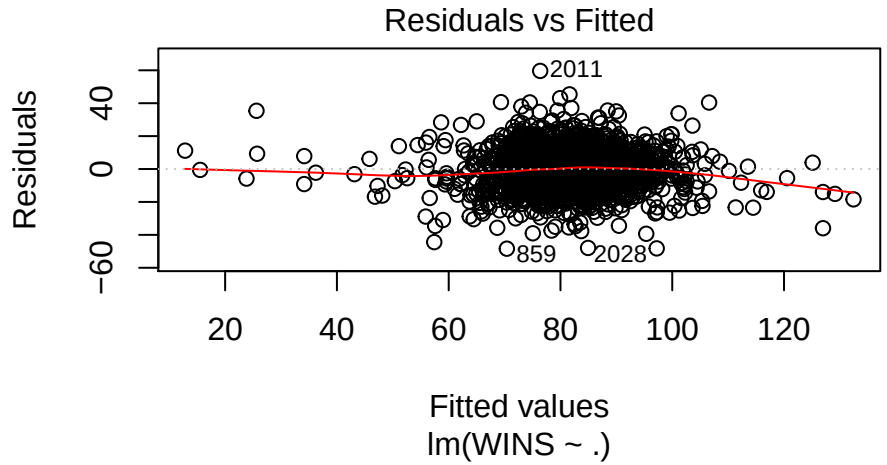
create a dataframe for holding the regression metrics.

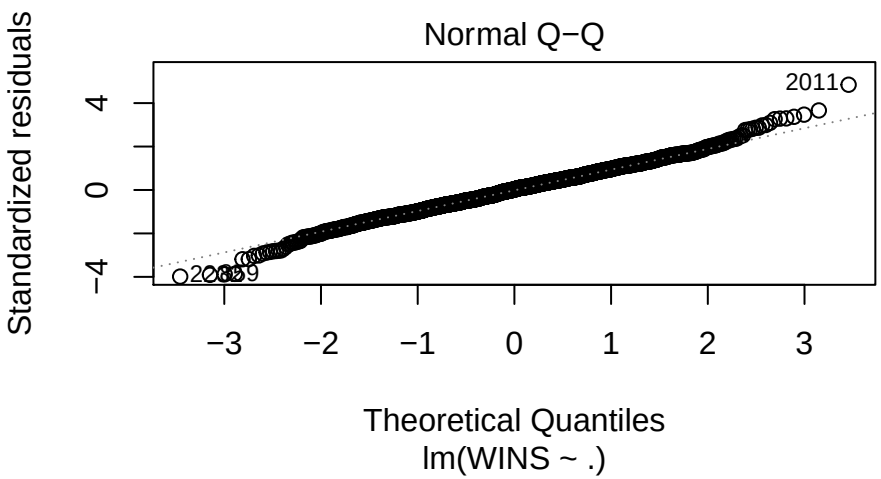
Full model Stats

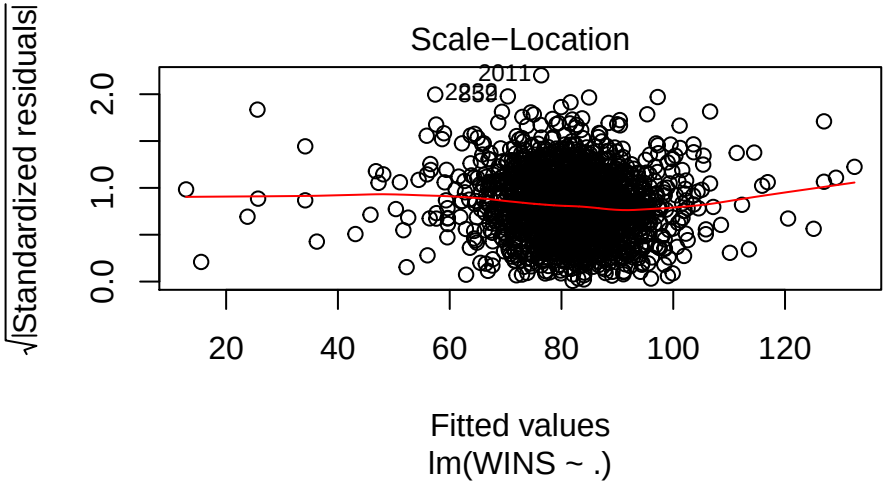
```
##
## Call:
## lm(formula = WINS ~ ., data = training_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

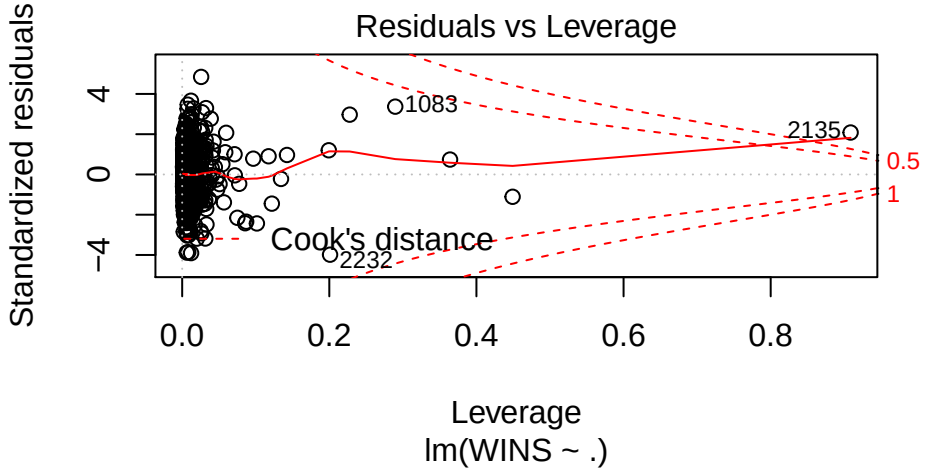
```
## -48.405  -8.181   0.329   7.759  59.627
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.115e+02  1.330e+01   8.383  < 2e-16 ***
## bt_H         4.365e-02  4.012e-03  10.882  < 2e-16 ***
## bt_2B       -2.487e-02  9.817e-03  -2.533  0.01139 *
## bt_3B        6.384e-02  1.844e-02   3.461  0.00055 ***
## bt_HR        1.558e-01  3.360e-02   4.638 3.78e-06 ***
## bt_BB        4.178e-02  6.877e-03   6.076 1.50e-09 ***
## bt_SO       -1.176e-02  2.726e-03  -4.316 1.68e-05 ***
## br_SB        4.624e-02  5.487e-03   8.426  < 2e-16 ***
## br_CS        1.364e-02  1.122e-02   1.216  0.22433
## ph_H         1.479e-03  4.608e-04   3.210  0.00135 **
## ph_HR       -7.223e-02  3.069e-02  -2.354  0.01869 *
## ph_BB       -2.114e-02  4.834e-03  -4.373 1.30e-05 ***
## ph_SO        8.737e-04  9.125e-04   0.958  0.33841
## fd_E        -5.399e-02  3.684e-03 -14.655  < 2e-16 ***
## fd_DP       -1.150e-01  1.381e-02  -8.326  < 2e-16 ***
## bt_1B               NA         NA      NA       NA
## BB          -4.238e+01  6.263e+00  -6.766 1.78e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.47 on 1809 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3918, Adjusted R-squared:  0.3868
## F-statistic: 77.69 on 15 and 1809 DF,  p-value: < 2.2e-16
```


fullregressor$residuals
Index


Residuals vs Fitted
Residuals
2011
859
2028
Fitted values
lm(WINS ~ .)


Normal Q−Q
Standardized residuals
2011
Theoretical Quantiles
lm(WINS ~ .)


Scale−Location
√|Standardized residuals|
2011
Fitted values
lm(WINS ~ .)



Test evaluation Metrics and prediction results

We see the two independent variables has p-value > 0.05. we will remove this independent variable from the model and try its performance.

Since the R-square and RMSE of the model is not that great and it shows a possible underfitting problem. We will try out some transformations like backward elimination, square, logarithmic and BOX-COX transformations and review the results.

```
## Warning in predict.lm(fullregressor, newdata = test_set): prediction from a
## rank-deficient fit may be misleading
```

```
##    predictions actual
## 5     66.72792     83
## 10    66.58775     77
## 21    74.98691     71
## 29    75.99609     82
## 46    87.60120     86
## 48    91.52018     99
```

```
##   rmse_test rmse_train r2_train r2_test adj_r2_train
## 1   12.8927    12.4108   0.3918   0.241       0.3868
```

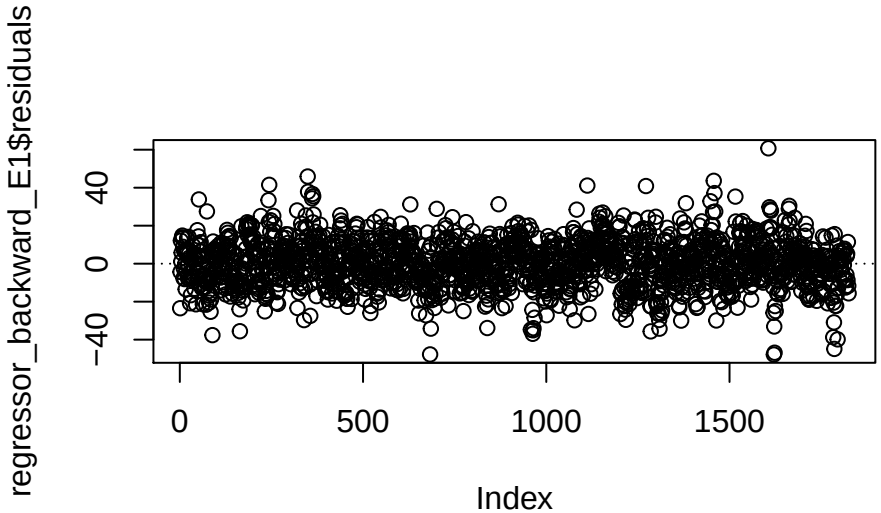
Backward Elimination

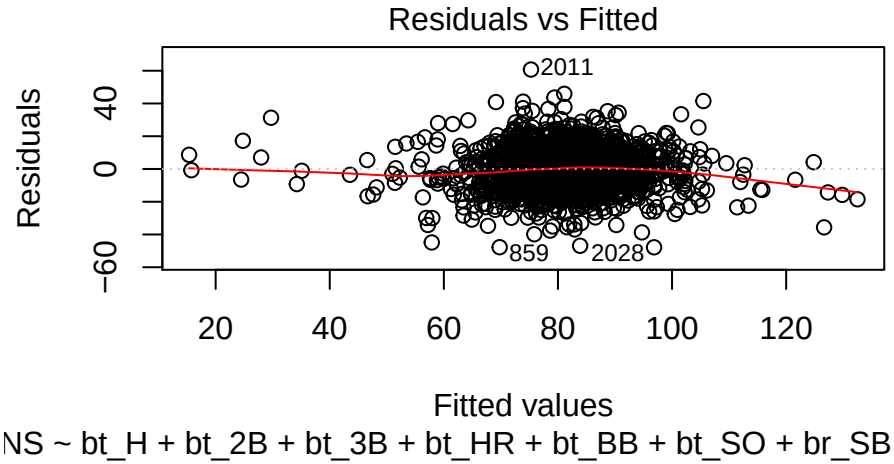
Through backward elimination process, some independent variables(ph_SO,br_CS, bt_1B) that has pvalue more than the significance level of 0.05 were removed from the full model

Backward elimination Model Stats

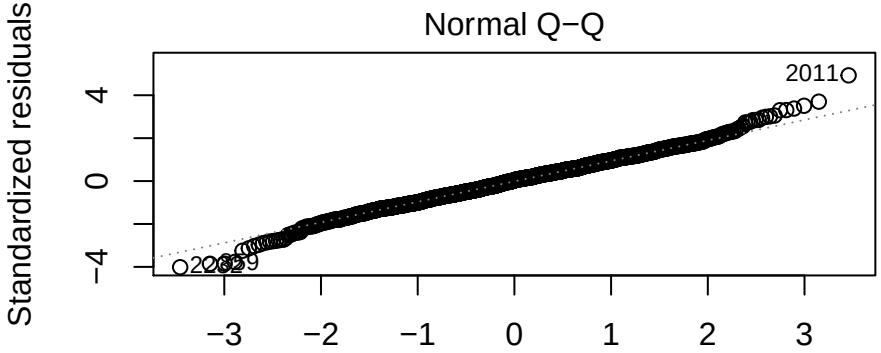
```
##
## Call:
## lm(formula = WINS ~ bt_H + bt_2B + bt_3B + bt_HR + bt_BB + bt_SO +
##     br_SB + ph_H + ph_HR + ph_BB + fd_E + fd_DP + BB, data = training_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.845  -8.187   0.450   7.839  60.727
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.139e+02  1.319e+01   8.637  < 2e-16 ***
## bt_H         4.346e-02  3.973e-03  10.937  < 2e-16 ***
## bt_2B       -2.517e-02  9.732e-03  -2.587 0.009765 **
## bt_3B        6.687e-02  1.833e-02   3.648 0.000272 ***
## bt_HR        1.623e-01  3.267e-02   4.967 7.42e-07 ***
## bt_BB        3.854e-02  6.228e-03   6.188 7.51e-10 ***
## bt_SO       -1.088e-02  2.511e-03  -4.335 1.54e-05 ***
## br_SB        5.046e-02  4.592e-03  10.990  < 2e-16 ***
## ph_H         1.441e-03  4.589e-04   3.139 0.001722 **
## ph_HR       -8.049e-02  2.969e-02  -2.711 0.006778 **
## ph_BB       -1.847e-02  4.188e-03  -4.411 1.09e-05 ***
## fd_E        -5.440e-02  3.631e-03 -14.982  < 2e-16 ***
## fd_DP       -1.171e-01  1.344e-02  -8.710  < 2e-16 ***
## BB          -4.282e+01  6.236e+00  -6.866 9.05e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.47 on 1811 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.391,  Adjusted R-squared:  0.3866
```

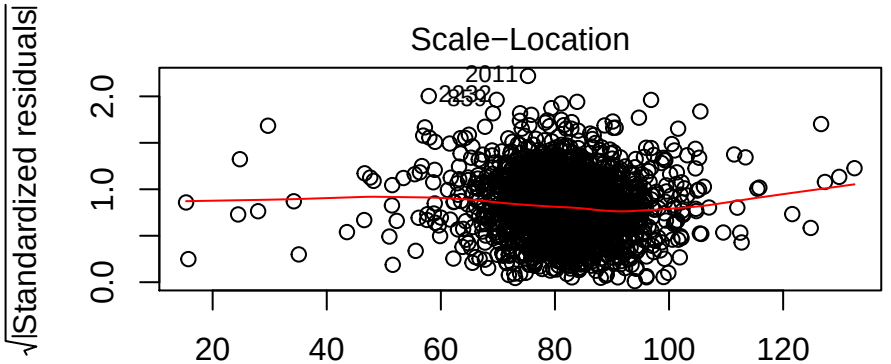
```
## F-statistic: 89.42 on 13 and 1811 DF,  p-value: < 2.2e-16
```



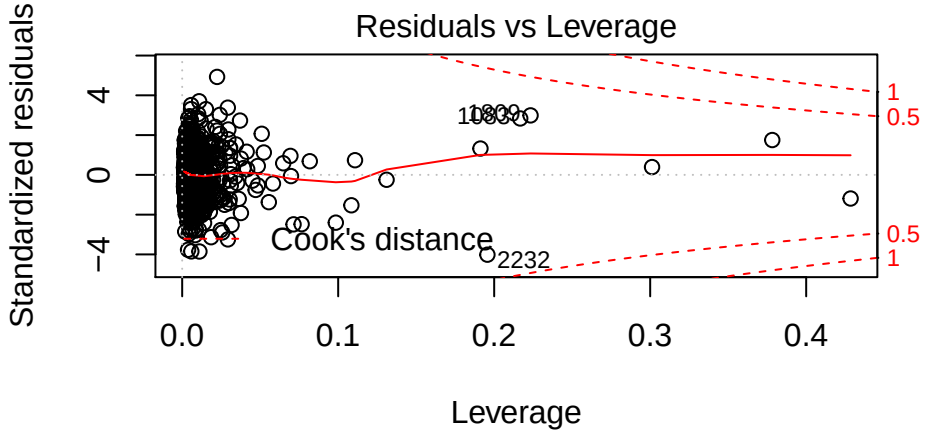


NS ~ bt_H + bt_2B + bt_3B + bt_HR + bt_BB + bt_SO + br_SB + ph





NS ~ bt_H + bt_2B + bt_3B + bt_HR + bt_BB + bt_SO + br_SB + ph NS ~ bt_H + bt_2B + bt_3B + bt_HR + bt_BB + bt_SO + br_SB + ph



NS ~ bt_H + bt_2B + bt_3B + bt_HR + bt_BB + bt_SO + br_SB + ph

Test evaluation Metrics and prediction results

```
##    predictions actual
## 5     66.80905     83
## 10    66.85463     77
## 21    74.79397     71
## 29    76.03111     82
## 46    87.70131     86
## 48    91.57863     99
```

```
##   rmse_test rmse_train r2_train r2_test adj_r2_train
## 1   12.9301    12.4195    0.391  0.2381       0.3866
```

Backward Elimination + removal of colinear Variables

Performing a VIF Test on all variables to remove some independent variables which are colinear(VIF has more than 5)

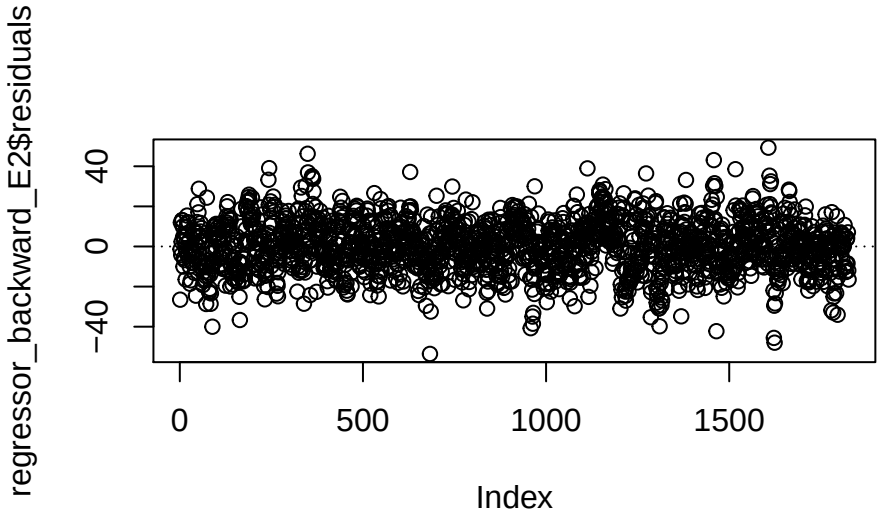
```
##      bt_H     bt_2B     bt_3B     bt_HR     bt_BB     bt_SO     br_SB
```

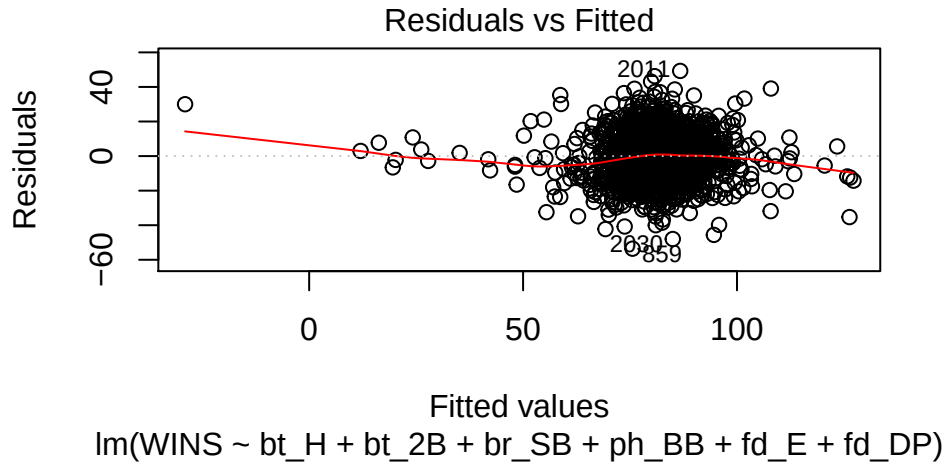
```
##  4.014112  2.464797  3.033585 46.653155  5.981965  4.528813  2.695233
##      ph_H     ph_HR     ph_BB      fd_E     fd_DP        BB
##  4.896792 39.716223  4.674980  8.358823  1.898830 10.757738
```

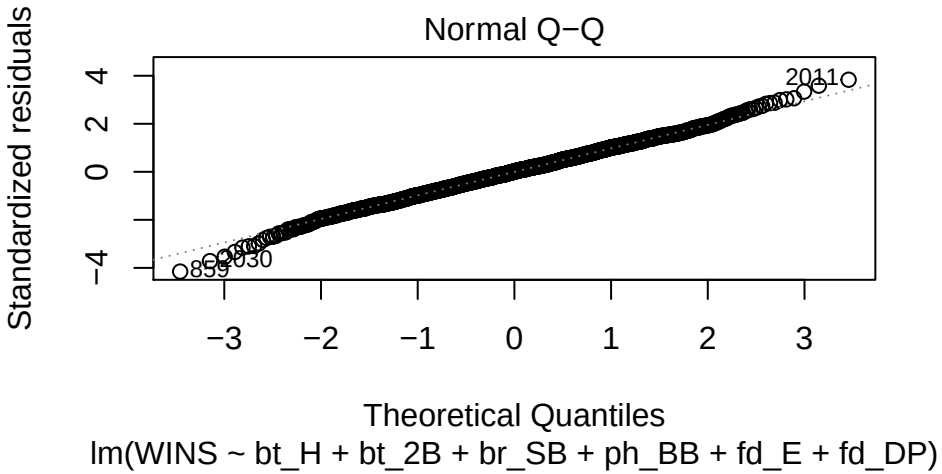
remove Colinear variables bt_HR + ph_HR + BB + bt_BB. Also remove variables with p-value > 0.05 (bt_3B, bt_SO, ph_H)

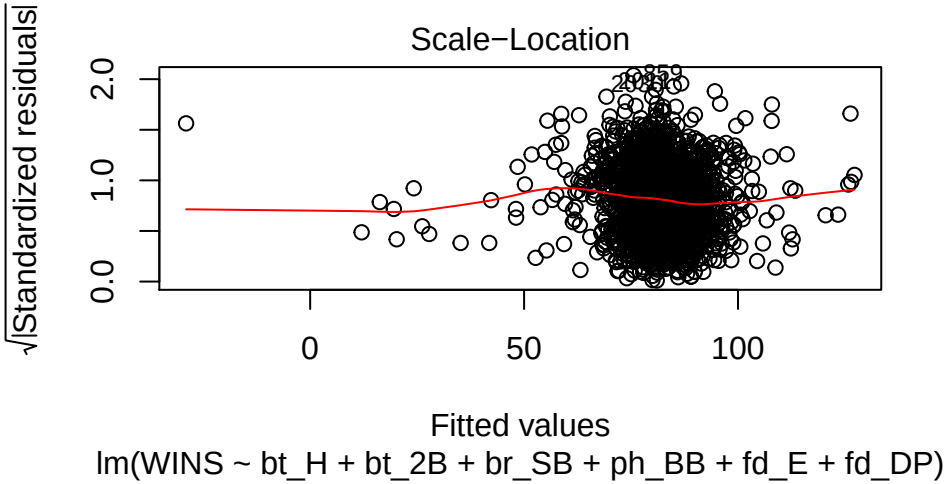
Backward elimination Model Stats(with removal of colinear variables)

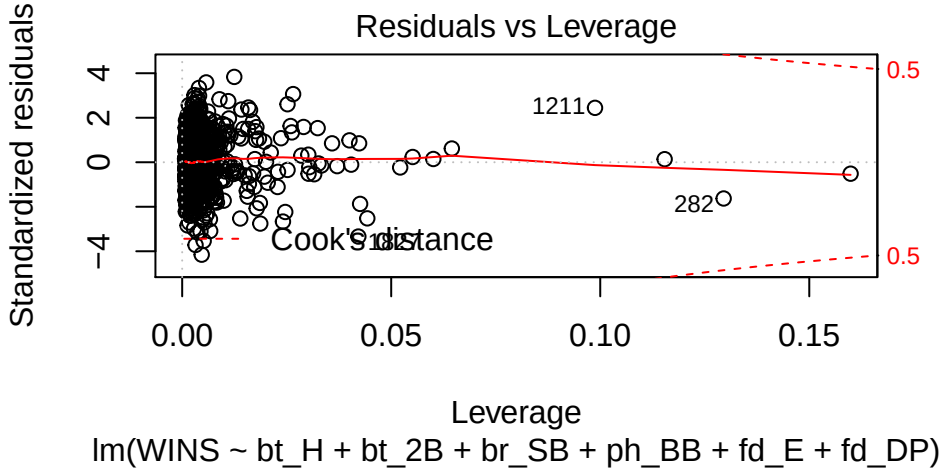
```
##
## Call:
## lm(formula = WINS ~ bt_H + bt_2B + br_SB + ph_BB + fd_E + fd_DP,
##     data = training_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.579  -8.577   0.122   8.482  49.278
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.207764   3.593651   3.397 0.000696 ***
## bt_H         0.059449   0.002877  20.662  < 2e-16 ***
## bt_2B       -0.021329   0.009053  -2.356 0.018579 *
## br_SB        0.040128   0.004117   9.747  < 2e-16 ***
## ph_BB        0.006392   0.001998   3.200 0.001399 **
## fd_E        -0.038440   0.001792 -21.450  < 2e-16 ***
## fd_DP       -0.084982   0.013246  -6.416 1.78e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.93 on 1819 degrees of freedom
## Multiple R-squared:  0.3511, Adjusted R-squared:  0.349
## F-statistic: 164.1 on 6 and 1819 DF,  p-value: < 2.2e-16
```


regressor_backward_E2$residuals
−40
0
40
0
500
1000
1500
Index


Residuals vs Fitted
Residuals
−60
0
40
2011
2030
859
0
50
100
Fitted values
lm(WINS ~ bt_H + bt_2B + br_SB + ph_BB + fd_E + fd_DP)


Normal Q−Q
Standardized residuals
−4
0
2
4
2011
859
2030
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(WINS ~ bt_H + bt_2B + br_SB + ph_BB + fd_E + fd_DP)


Scale−Location
√|Standardized residuals|
0.0
1.0
2.0
0
50
100
Fitted values
lm(WINS ~ bt_H + bt_2B + br_SB + ph_BB + fd_E + fd_DP)




```
##    predictions actual
## 5      70.70862     83
## 10     70.34577     77
## 21     77.07268     71
## 29     79.39163     82
## 46     84.77783     86
## 48     88.10494     99
```

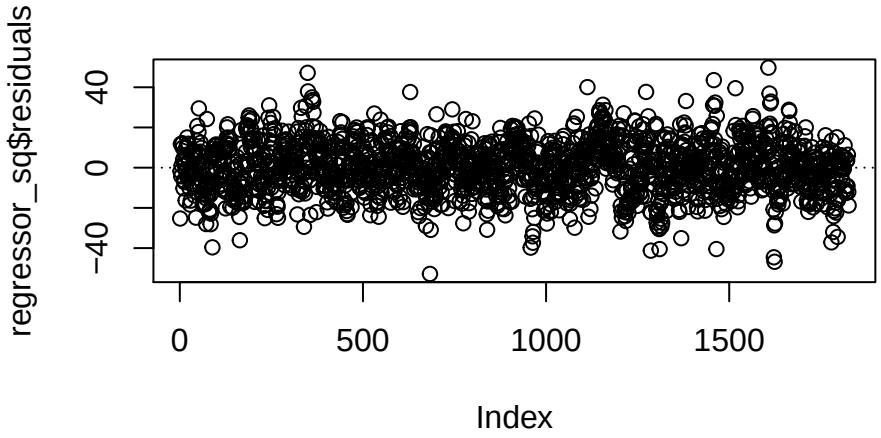
```
##   rmse_test rmse_train r2_train r2_test adj_r2_train
## 1   12.7212    12.9056   0.3511  0.2413        0.349
```

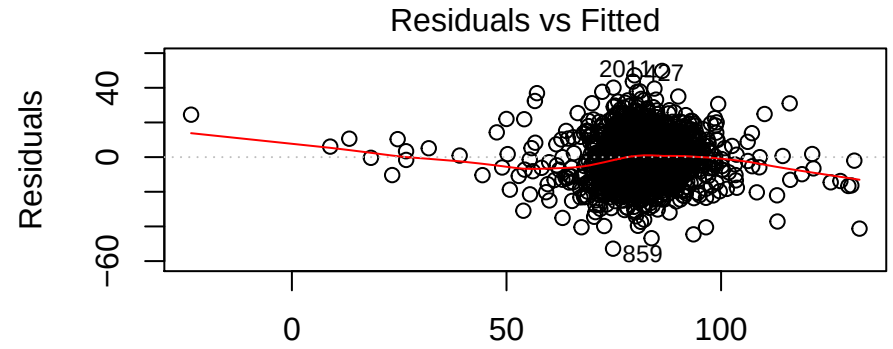
Square transformation Model

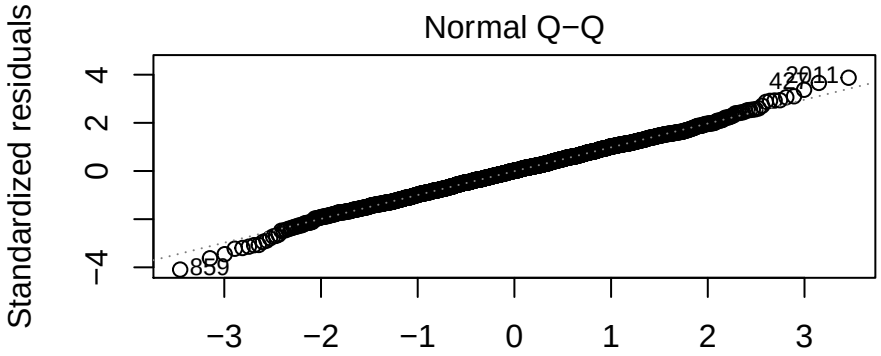
Square transformation Model Stats

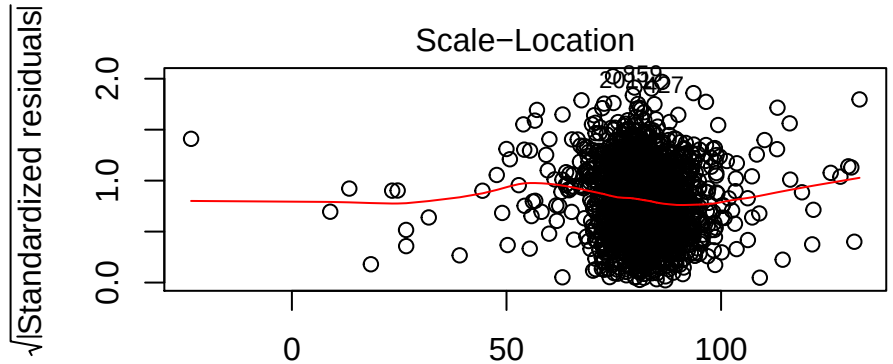
```
##
## Call:
## lm(formula = WINS ~ I(bt_H^2) + I(bt_2B^2) + (br_SB^2) + (ph_BB^2) +
```

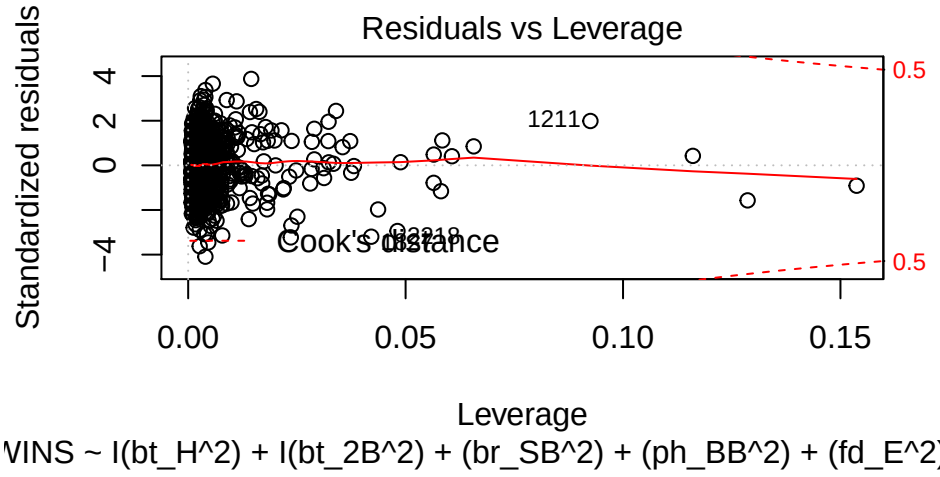
```
##     (fd_E^2) + (fd_DP^2), data = training_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.831  -8.729   0.132   8.529  49.738
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.641e+01  2.486e+00  22.694  < 2e-16 ***
## I(bt_H^2)    1.858e-05  8.849e-07  21.002  < 2e-16 ***
## I(bt_2B^2)  -3.519e-05  1.734e-05  -2.029  0.04256 *
## br_SB        4.152e-02  4.116e-03  10.088  < 2e-16 ***
## ph_BB        5.967e-03  1.998e-03   2.987  0.00286 **
## fd_E        -3.987e-02  1.795e-03 -22.219  < 2e-16 ***
## fd_DP       -8.490e-02  1.322e-02  -6.424 1.69e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.93 on 1819 degrees of freedom
## Multiple R-squared:  0.3508, Adjusted R-squared:  0.3487
## F-statistic: 163.8 on 6 and 1819 DF,  p-value: < 2.2e-16
```


regressor_sq$residuals
−40
0
40
0
500
1000
1500
Index


Residuals vs Fitted
Residuals
−60
0
40
0
50
100
Fitted values
MINS ~ I(bt_H^2) + I(bt_2B^2) + (br_SB^2) + (ph_BB^2) + (fd_E^2)


Normal Q−Q
Standardized residuals
−4
0
2
4
−3
−2
−1
0
1
2
3
Theoretical Quantiles
MINS ~ I(bt_H^2) + I(bt_2B^2) + (br_SB^2) + (ph_BB^2) + (fd_E^2)


Scale−Location
√|Standardized residuals|
0.0
1.0
2.0
0
50
100
Fitted values
MINS ~ I(bt_H^2) + I(bt_2B^2) + (br_SB^2) + (ph_BB^2) + (fd_E^2)



MINS ~ I(bt_H^2) + I(bt_2B^2) + (br_SB^2) + (ph_BB^2) + (fd_E^2) -

Test evaluation Metrics and prediction results

RMSE(test) has improved, but R-square is reduced slightly.

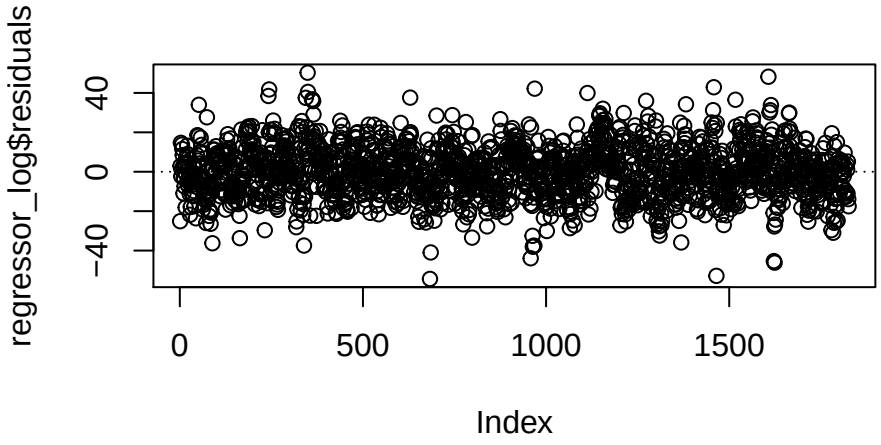
```
##    predictions actual
## 5     71.49189     83
## 10    71.68233     77
## 21    76.78273     71
## 29    79.73048     82
## 46    84.73991     86
## 48    88.01546     99
```

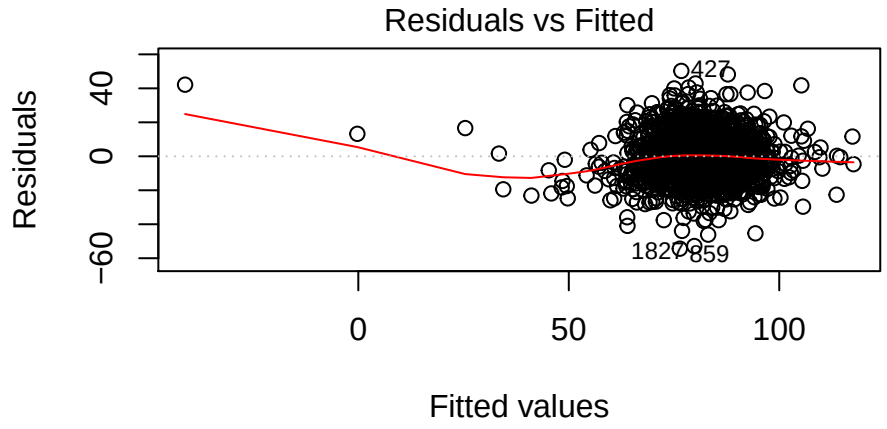
```
##   rmse_test rmse_train r2_train r2_test adj_r2_train
## 1   12.6774    12.9088   0.3508  0.2459       0.3487
```

Logarithmic transformation

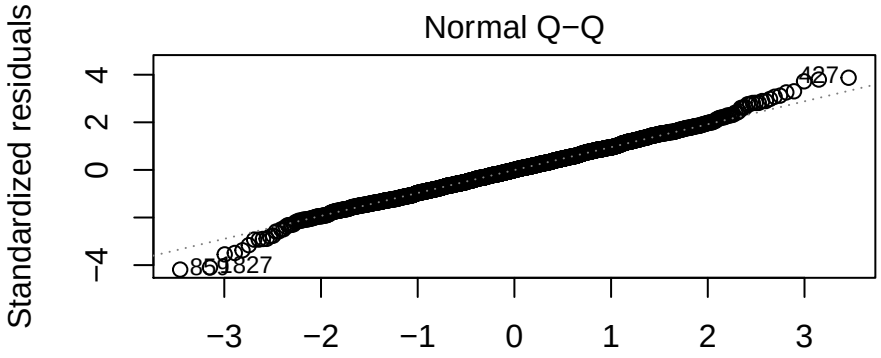
Logarithmic transformation Stats

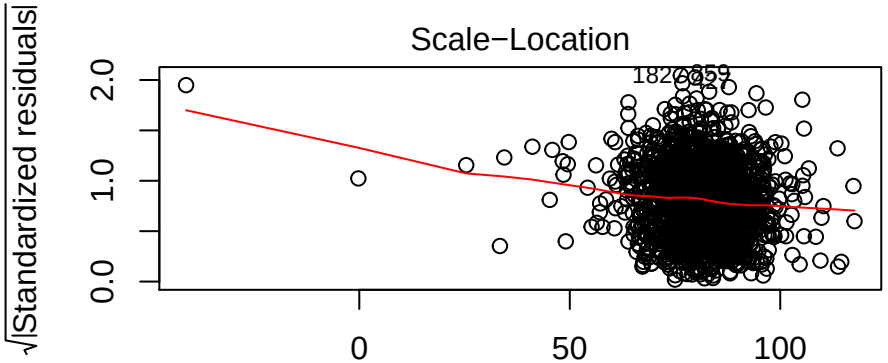
```
##
## Call:
## lm(formula = WINS ~ log1p(bt_H) + log1p(bt_2B) + log1p(br_SB) +
##     log1p(ph_BB) + log1p(fd_E) + log1p(fd_DP), data = training_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.379  -8.525  -0.049   8.348  50.265
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -453.2839    26.5293 -17.086  < 2e-16 ***
## log1p(bt_H)    97.0853     4.7161  20.586  < 2e-16 ***
## log1p(bt_2B)   -9.6544     2.3418  -4.123 3.91e-05 ***
## log1p(br_SB)    4.8338     0.5874   8.229 3.56e-16 ***
## log1p(ph_BB)    3.5732     1.1583   3.085  0.00207 **
## log1p(fd_E)   -14.5104     0.7717 -18.804  < 2e-16 ***
## log1p(fd_DP)  -17.8524     1.8683  -9.555  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.02 on 1819 degrees of freedom
## Multiple R-squared:  0.3419, Adjusted R-squared:  0.3397
## F-statistic: 157.5 on 6 and 1819 DF,  p-value: < 2.2e-16
```



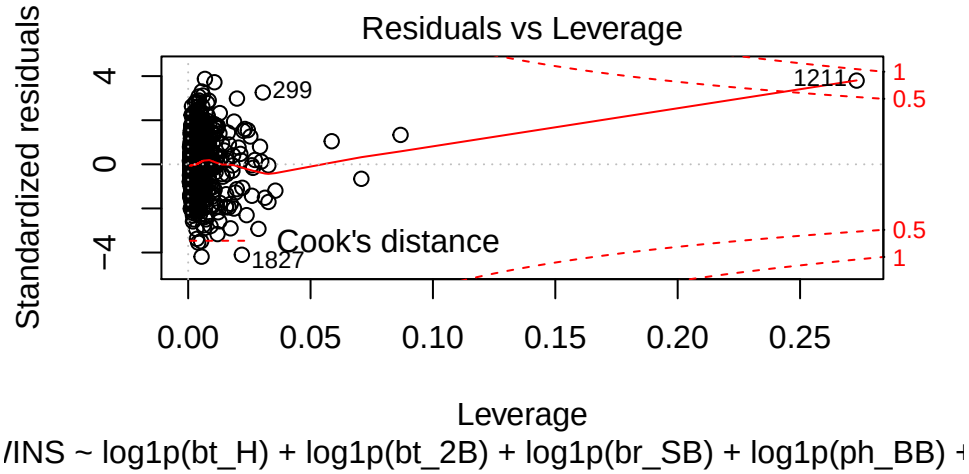


/INS ~ log1p(bt_H) + log1p(bt_2B) + log1p(br_SB) + log1p(ph_BB)





/INS ~ log1p(bt_H) + log1p(bt_2B) + log1p(br_SB) + log1p(ph_BB) /INS ~ log1p(bt_H) + log1p(bt_2B) + log1p(br_SB) + log1p(ph_BB)



/INS ~ log1p(bt_H) + log1p(bt_2B) + log1p(br_SB) + log1p(ph_BB) +

Test evaluation Metrics and prediction results

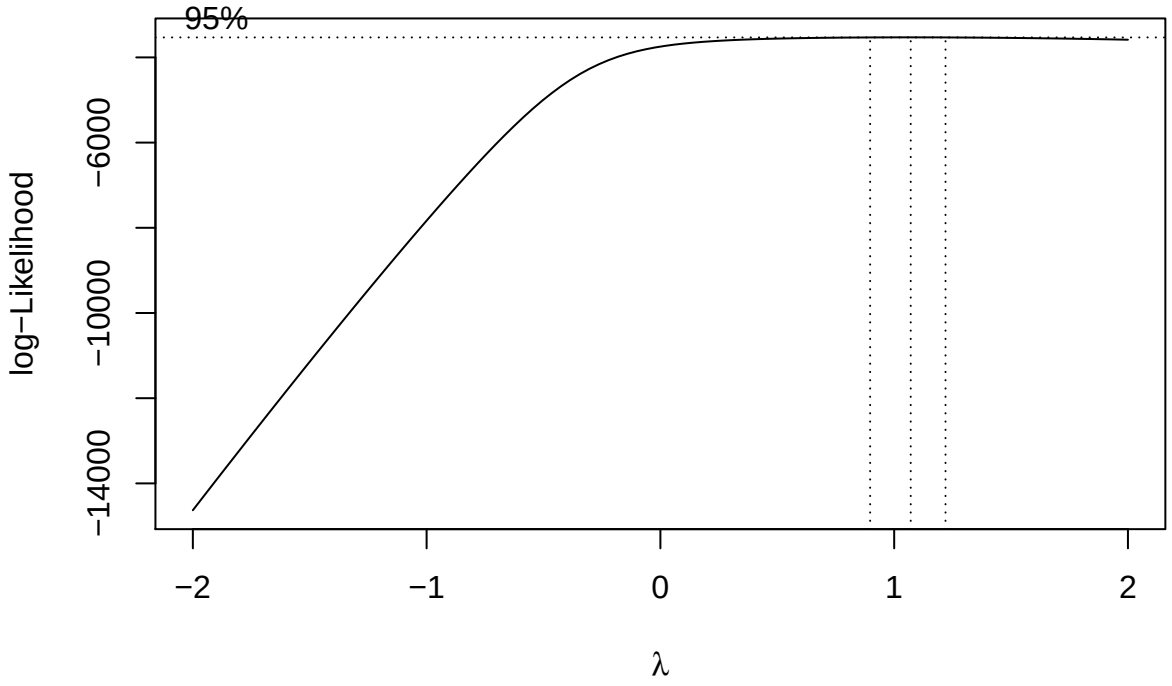
Residual Error plot developed a slight cure and OLS assumptions are not met.

```
##    predictions actual
## 5     69.82943     83
## 10    69.84392     77
## 21    74.99316     71
## 29    82.81690     82
## 46    88.47094     86
## 48    94.84237     99
```

```
##   rmse_test rmse_train r2_train r2_test adj_r2_train
## 1   13.0509    12.9974   0.3419  0.2117       0.3397
```

Box Cox transformation

Trying out a Box-Cox transformation. Used the best model so far which is backward elimination model to apply Box-Cox transformation. Lambda comes close to 1. So it doesnt make any difference and there is no need to apply the Box-Cox tranformation.



Cross Validation

Performing a cross validation algorithm if it make some improvement.

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -53.579  -8.577   0.122   8.482  49.278
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.207764   3.593651   3.397 0.000696 ***
## bt_H         0.059449   0.002877  20.662  < 2e-16 ***
## bt_2B       -0.021329   0.009053  -2.356 0.018579 *
## br_SB        0.040128   0.004117   9.747  < 2e-16 ***
## ph_BB        0.006392   0.001998   3.200 0.001399 **
## fd_E        -0.038440   0.001792 -21.450  < 2e-16 ***
## fd_DP       -0.084982   0.013246  -6.416 1.78e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.93 on 1819 degrees of freedom
## Multiple R-squared:  0.3511, Adjusted R-squared:  0.349
## F-statistic: 164.1 on 6 and 1819 DF,  p-value: < 2.2e-16
```

```
##    predictions actual
## 5     70.70862     83
## 10    70.34577     77
## 21    77.07268     71
## 29    79.39163     82
## 46    84.77783     86
## 48    88.10494     99
```

```
##   rmse_test rmse_train r2_train r2_test adj_r2_train
## 1   12.7212        NaN   0.3511  0.2413        0.349
```

regressor	Rsquare(Train-set)	Adjusted-RSquare(Training-set)	RMSE(Train-set)	RMSE(Test-set)	R-Square(Test)
Full-Model	0.3918	0.3868	12.4108	12.8927	0.241
Backward elimination-1	0.391	0.3866	12.4195	12.9301	0.2381
Backward elimination-2	0.3511	0.349	12.9056	12.7212	0.2413
Square Transformation	0.3508	0.3487	12.9088	12.6774	0.2459
Log Transformation	0.3419	0.3397	12.9974	13.0509	0.2117
Cross Validation	0.3511	0.349	NaN	12.7212	0.2413

Summary of findings

1. Model built using Backward Elimination-2 and Square transformation looks to be the best among all models considering comparetively low RMSE(test) and comparetively good R-square values.
2. Less R^2 and high RMSE shows an underfitting problem. The dataset doesnt follow a linear relationship with the response variable. None of the transformation helped improving the metrics.
3. Although the model exhibits an underfitting problem, it slightly met the ordinary least square assumptions.

a. Residuals doesnt have high variance.
b. Residual QQ plots gives a slight straight line.

4. Residual Error plot developed a slight cure and OLS assumptions are not met.
5. Metrics shows RMSE(Train) and RMSE(TEST) is almost same and dont have much differences. But the R^2 has some difference for all the models. We will see some overfitting solution and check the model gets improved further in next step.

Does overfitting exist?

RMSE(train) and RMSE(test) doesn't indicate there is a problem of overfitting, but R^2 has some difference. We want to see if the model gets improved using some of the underfitting solutions.

Ridge Regression

Lets try out the ridge regression with cross validation.

```
## 7 x 1 sparse Matrix of class "dgCMatrix"
##                           1
## (Intercept) 16.789041438
## bt_H         0.053359100
## bt_2B       -0.007399118
## br_SB        0.035809230
## ph_BB        0.006233650
## fd_E        -0.034062969
## fd_DP       -0.080906289
```

R^2(test) and RMSE doesn't get improved

```
##    predictions actual
## 5     70.99658     83
## 10    71.00643     77
## 21    77.27990     71
## 29    79.50153     82
## 46    84.89866     86
## 48    88.04531     99
```

```
##       RMSE   Rsquare
## 1 12.71836 0.2403934
```

Lasso

Lets try out the Lasso regression with cross validation.

```
## 7 x 1 sparse Matrix of class "dgCMatrix"
##                         1
## (Intercept) 12.536563485
## bt_H         0.058912750
## bt_2B       -0.019708708
## br_SB        0.039811044
## ph_BB        0.006247125
## fd_E        -0.038077537
## fd_DP       -0.084283559
```

R^2 and RMSE doesn't get improved

```
##       RMSE   Rsquare
## 1 12.72148 0.2410859
```

Elastic net

Lets try out the Elastic net regression with cross validation.

```
## 7 x 1 sparse Matrix of class "dgCMatrix"
##                         1
```

```
## (Intercept) 12.567573359
## bt_H         0.058888868
## bt_2B       -0.019710177
## br_SB        0.039783371
## ph_BB        0.006269394
## fd_E        -0.038058422
## fd_DP       -0.084345388
```

R^2 and RMSE doesn't get improved

```
##    predictions actual
## 5     70.72962     83
## 10    70.41252     77
## 21    77.10584     71
## 29    79.41875     82
## 46    84.80441     86
## 48    88.09797     99
```

```
##       RMSE   Rsquare
## 1 12.72105 0.2411239
```

Overfitting solution didnt take any effect, so there is no improvement to the best model we identified above.

Prediction of evaluation data set

Conclusions on Model Building and Selection

When building a regression model, it may seem easy to throw as many variables into the mix as it takes to get a high R^2. But that will not usually lead to the best and most predictive model out of sample. Often a model with a lower R^2 may in fact be superior if it is more sensible, violates fewer assumptions, has more desirable residual plots, and was built using strong domain knowledge.

This baseball data was an example of a situation in which, though the absolute R^2 values may not have been the highest, the strongest models were ones in which all assumptions were examined and out of sample predictions were tested.

Evaluations

Below we have shown the actual predictions for two of our above models.

Evaluating using Square transformation model

INDEX	WINS
9	68
10	69
14	77
47	88
60	65
63	69
74	83
83	79
98	73
120	76
123	74
135	79
138	76
140	79
151	83
153	77
171	75
184	79
193	74
213	90
217	83
226	86
230	79
241	72
291	84
294	88
300	48
348	77
350	85
357	80
367	88
368	84
372	82
382	82
388	79

INDEX	WINS
396	81
398	76
403	86
407	88
410	89
412	81
414	90
436	17
440	101
476	91
479	92
481	102
501	77
503	72
506	80
519	78
522	88
550	77
554	74
566	78
578	81
596	91
599	78
605	67
607	81
614	90
644	77
692	87
699	82
700	83
716	106
721	75
722	84
729	80
731	89
746	86
763	75

INDEX	WINS
774	75
776	82
788	86
789	85
792	82
811	83
835	76
837	81
861	85
862	89
863	96
871	78
879	81
887	81
892	81
904	84
909	88
925	88
940	85
951	91
976	74
981	88
983	89
984	85
989	88
995	104
1000	88
1001	87
1007	81
1016	74
1027	84
1033	80
1070	79
1081	64
1084	58
1098	79
1150	84

INDEX	WINS
1160	62
1169	82
1172	86
1174	95
1176	92
1178	83
1184	82
1193	87
1196	81
1199	78
1207	78
1218	92
1223	72
1226	71
1227	69
1229	72
1241	89
1244	92
1246	77
1248	90
1249	92
1253	88
1261	81
1305	81
1314	86
1323	85
1328	70
1353	75
1363	79
1371	92
1372	83
1389	70
1393	75
1421	92
1431	75
1437	75
1442	72

INDEX	WINS
1450	77
1463	81
1464	81
1470	82
1471	80
1484	85
1495	27
1507	73
1514	80
1526	75
1549	86
1552	66
1556	91
1564	76
1585	98
1586	103
1590	90
1591	98
1592	91
1603	84
1612	80
1634	79
1645	75
1647	81
1673	91
1674	90
1687	80
1688	95
1700	82
1708	75
1713	77
1717	72
1721	74
1730	79
1737	88
1748	85
1749	86

INDEX	WINS
1763	81
1768	91
1778	98
1780	91
1782	52
1784	65
1794	123
1803	73
1804	84
1819	78
1832	80
1833	84
1844	70
1847	78
1854	80
1855	77
1857	84
1864	76
1865	81
1869	76
1880	91
1881	83
1882	82
1894	82
1896	80
1916	81
1918	73
1921	104
1926	95
1938	80
1979	67
1982	68
1987	84
1997	80
2004	94
2011	80
2015	80

INDEX	WINS
2022	81
2025	79
2027	82
2031	77
2036	93
2066	76
2073	83
2087	80
2092	82
2125	70
2148	82
2162	93
2191	76
2203	85
2218	79
2221	77
2225	86
2232	77
2267	86
2291	74
2299	88
2317	86
2318	86
2353	85
2403	68
2411	89
2415	82
2424	83
2441	74
2464	84
2465	80
2472	55
2481	94
2487	42
2500	72
2501	76
2520	85

INDEX	WINS
2521	84
2525	77

Evaluating using Backward elimination model

INDEX	WINS
9	67
10	67
14	77
47	88
60	66
63	69
74	83
83	79
98	72
120	76
123	74
135	80
138	77
140	80
151	83
153	77
171	74
184	79
193	74
213	90
217	83
226	86
230	79
241	72
291	84
294	89
300	51
348	77
350	86

INDEX	WINS
357	80
367	88
368	85
372	82
382	82
388	79
396	81
398	75
403	86
407	88
410	89
412	82
414	91
436	16
440	102
476	92
479	93
481	103
501	76
503	71
506	80
519	79
522	89
550	77
554	73
566	78
578	81
596	91
599	77
605	65
607	81
614	90
644	77
692	88
699	82
700	83
716	106

INDEX	WINS
721	76
722	85
729	80
731	89
746	86
763	75
774	75
776	83
788	86
789	86
792	82
811	83
835	75
837	81
861	85
862	89
863	96
871	79
879	82
887	81
892	81
904	85
909	88
925	89
940	85
951	92
976	74
981	89
983	89
984	85
989	88
995	103
1000	89
1001	88
1007	82
1016	74
1027	85

INDEX	WINS
1033	80
1070	79
1081	65
1084	57
1098	79
1150	84
1160	61
1169	82
1172	86
1174	96
1176	92
1178	83
1184	82
1193	88
1196	81
1199	78
1207	78
1218	93
1223	70
1226	70
1227	66
1229	70
1241	89
1244	92
1246	77
1248	91
1249	92
1253	88
1261	82
1305	81
1314	87
1323	85
1328	71
1353	74
1363	79
1371	91
1372	83

INDEX	WINS
1389	69
1393	74
1421	92
1431	75
1437	74
1442	71
1450	76
1463	81
1464	82
1470	82
1471	80
1484	85
1495	31
1507	73
1514	80
1526	74
1549	86
1552	67
1556	92
1564	76
1585	98
1586	103
1590	91
1591	98
1592	91
1603	85
1612	80
1634	79
1645	75
1647	81
1673	92
1674	90
1687	80
1688	95
1700	82
1708	74
1713	77

INDEX	WINS
1717	72
1721	74
1730	79
1737	87
1748	85
1749	86
1763	81
1768	89
1778	98
1780	91
1782	53
1784	65
1794	120
1803	73
1804	84
1819	78
1832	81
1833	85
1844	70
1847	78
1854	80
1855	77
1857	84
1864	76
1865	82
1869	75
1880	92
1881	83
1882	82
1894	82
1896	80
1916	81
1918	73
1921	104
1926	96
1938	81
1979	66

INDEX	WINS
1982	68
1987	85
1997	81
2004	94
2011	80
2015	79
2022	81
2025	79
2027	82
2031	76
2036	89
2066	75
2073	83
2087	80
2092	82
2125	71
2148	81
2162	93
2191	76
2203	85
2218	78
2221	76
2225	86
2232	77
2267	87
2291	74
2299	89
2317	86
2318	87
2353	85
2403	65
2411	89
2415	83
2424	83
2441	73
2464	84
2465	80

INDEX	WINS
2472	56
2481	95
2487	34
2500	71
2501	76
2520	85
2521	84
2525	77