Data621-hw1

Joseph Elikishvili

February 26, 2018

1. Introduction

In this paper we will explore the set containing 2200 records. Each record represents a professional baseball team from the years 1871 to 2006 inclusive. Each record has the performance of the team for the given year, with all of the statistics adjusted to match the performance of a 162 game season. We will build several linear regression models on the training data to predict number of wins for the team.

2. Data Exploration

First we will load the data and review the dimensions of the dataset

```
## [1] 2276   17
```

So we have a total of 2276 records and 17 variables.

Lets Review the data

```
##  [1] "INDEX"            "TARGET_WINS"      "TEAM_BATTING_H"
##  [4] "TEAM_BATTING_2B"  "TEAM_BATTING_3B"  "TEAM_BATTING_HR"
##  [7] "TEAM_BATTING_BB"  "TEAM_BATTING_SO"  "TEAM_BASERUN_SB"
## [10] "TEAM_BASERUN_CS"  "TEAM_BATTING_HBP" "TEAM_PITCHING_H"
## [13] "TEAM_PITCHING_HR" "TEAM_PITCHING_BB" "TEAM_PITCHING_SO"
## [16] "TEAM_FIELDING_E"  "TEAM_FIELDING_DP"
```

It appears we have Index which is just the index of the data and can be removed and we have our target variable Target_Wins and a total of 15 predictor variables we will be working with.

Next we will scan for missing values as those can affect the process

```
##            INDEX      TARGET_WINS   TEAM_BATTING_H  TEAM_BATTING_2B
##                0                0                0                0
##  TEAM_BATTING_3B  TEAM_BATTING_HR  TEAM_BATTING_BB  TEAM_BATTING_SO
##                0                0                0              102
##  TEAM_BASERUN_SB  TEAM_BASERUN_CS TEAM_BATTING_HBP  TEAM_PITCHING_H
##              131              772             2085                0
## TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E
##                0                0              102                0
## TEAM_FIELDING_DP
##              286
```

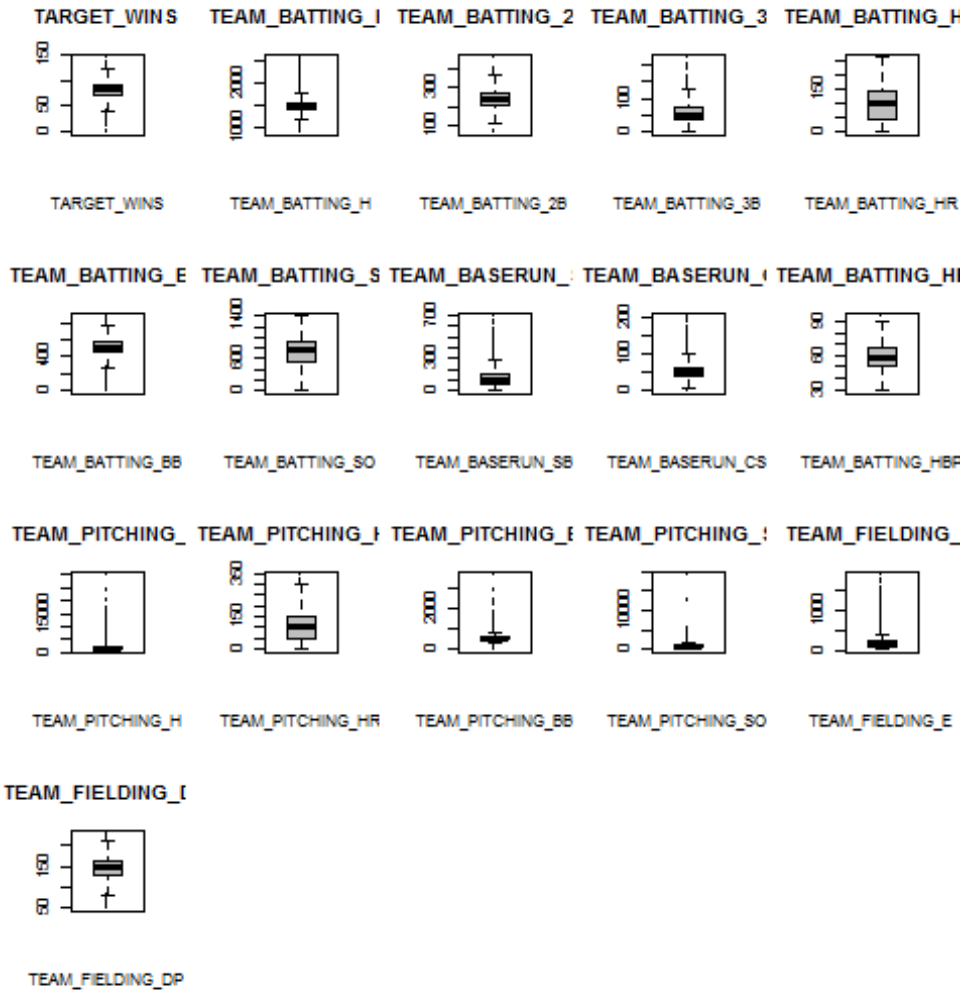
We can see that we have a reasonable number of missing values of the most of the predictors with the exception of TEAM_BATTING_HBP (2085) and TEAM_BASERUN_CS (772). These are significant numbers of missing data points considering we have a total 2276 records. One option is to remove the 2 columns, but we will attempt to impute the missing data and see if it can still be useful to us before removing the columns.

Next we will review the predictors given to us to better understand the data we are dealing with. The table gives us a basic overview of the data

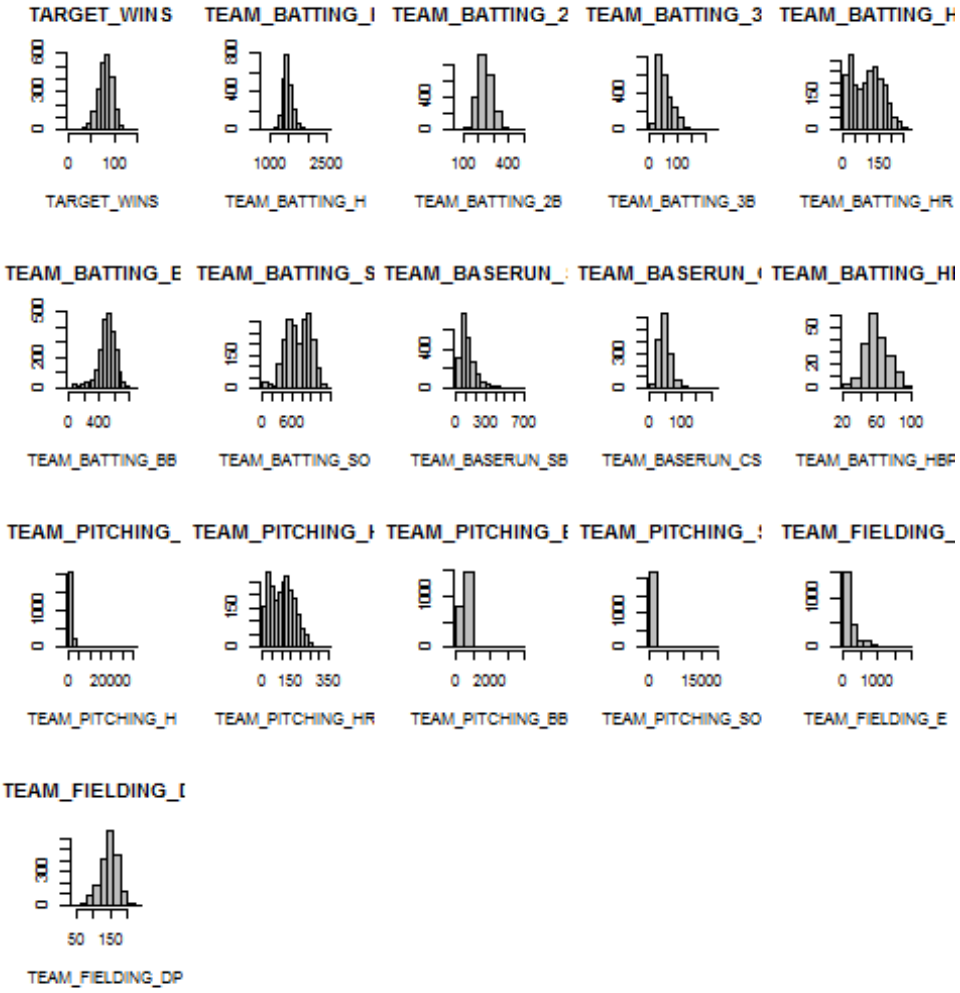
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
TARGET_WINS	1	2276	80.79	15.75	82.0	81.31	14.83	0	146	146	-0.40	1.03	0.33
TEAM_BATTING_H	2	2276	1469.27	144.59	1454.0	1459.04	114.16	891	2554	1663	1.57	7.28	3.03
TEAM_BATTING_2B	3	2276	241.25	46.80	238.0	240.40	47.44	69	458	389	0.22	0.01	0.98
TEAM_BATTING_3B	4	2276	55.25	27.94	47.0	52.18	23.72	0	223	223	1.11	1.50	0.59
TEAM_BATTING_HR	5	2276	99.61	60.55	102.0	97.39	78.58	0	264	264	0.19	-0.96	1.27
TEAM_BATTING_BB	6	2276	501.56	122.67	512.0	512.18	94.89	0	878	878	-1.03	2.18	2.57
TEAM_BATTING_SO	7	2174	735.61	248.53	750.0	742.31	284.66	0	1399	1399	-0.30	-0.32	5.33
TEAM_BASERUN_SB	8	2145	124.76	87.79	101.0	110.81	60.79	0	697	697	1.97	5.49	1.90
TEAM_BASERUN_CS	9	1504	52.80	22.96	49.0	50.36	17.79	0	201	201	1.98	7.62	0.59
TEAM_BATTING_HBP	10	191	59.36	12.97	58.0	58.86	11.86	29	95	66	0.32	-0.11	0.94
TEAM_PITCHING_H	11	2276	1779.21	1406.84	1518.0	1555.90	174.95	1137	30132	28995	10.33	141.84	29.49
TEAM_PITCHING_HR	12	2276	105.70	61.30	107.0	103.16	74.13	0	343	343	0.29	-0.60	1.28
TEAM_PITCHING_BB	13	2276	553.01	166.36	536.5	542.62	98.59	0	3645	3645	6.74	96.97	3.49
TEAM_PITC	1	21	817.	553.	813.	796.	257	0	1927	1927	22.	671.	11.

HING_SO	4	74	73	09	5	93	.23		8	8	17	19	86
TEAM_FIEL	1	22	246.	227.	159.	193.	62.	65	18	18	2.9	10.9	4.7
DING_E	5	76	48	77	0	44	27		98	33	9	7	7
TEAM_FIEL	1	19	146.	26.2	149.	147.	23.	52	22	17	-	0.18	0.5
DING_DP	6	90	39	3	0	58	72		8	6	0.3		9
											9		

Next we will create a boxplot of each of the predictors to visualize the variability of the data



And finally we will review the histograms of the predictors to see the distribution and the skew.



Next we will impute the missing data. Even though we have 2 predictors that have a lot of missing values, we will use them nevertheless and review the results. When we get to model selection, we will try a model with the predictors that have been heavily imputed and compare to the model without them.

For imputation we will use missForest package which will choose the best suited method and will impute the missing data.

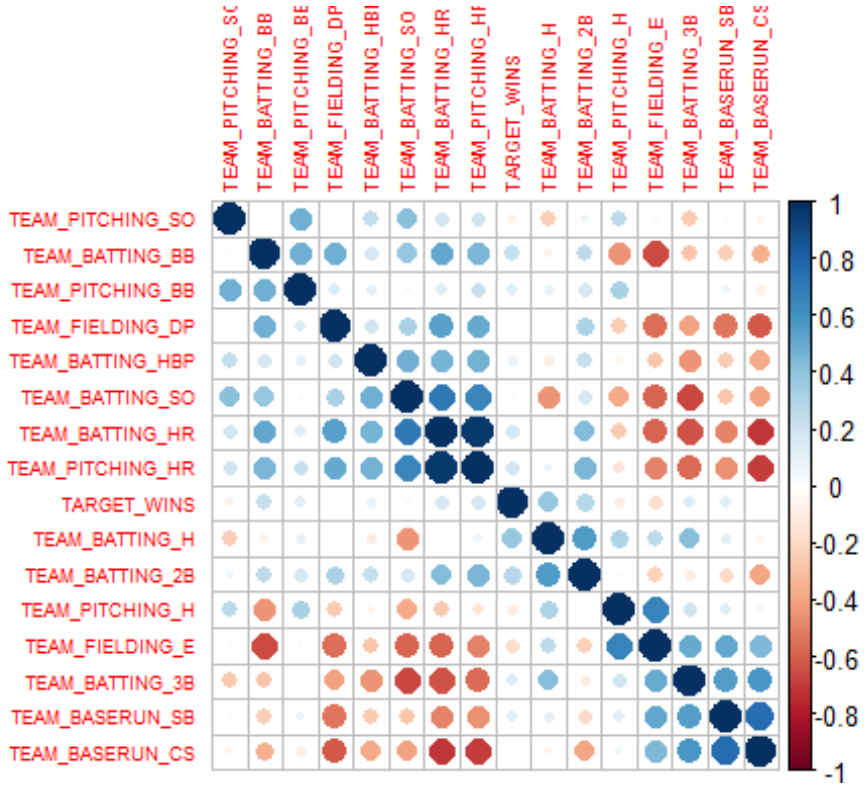
```
##   missForest iteration 1 in progress...done!
##   missForest iteration 2 in progress...done!
##   missForest iteration 3 in progress...done!
##   missForest iteration 4 in progress...done!
##   missForest iteration 5 in progress...done!
```

Lets review the imputed data and scan for missing values.

```
##       TARGET_WINS    TEAM_BATTING_H  TEAM_BATTING_2B  TEAM_BATTING_3B
##                 0                 0                0                0
##  TEAM_BATTING_HR  TEAM_BATTING_BB  TEAM_BATTING_SO  TEAM_BASERUN_SB
##                 0                 0                0                0
##  TEAM_BASERUN_CS TEAM_BATTING_HBP  TEAM_PITCHING_H TEAM_PITCHING_HR
##                 0                 0                0                0
## TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E TEAM_FIELDING_DP
##                 0                 0                0                0
```

We no longer have any missing values and we are clear to proceed further.

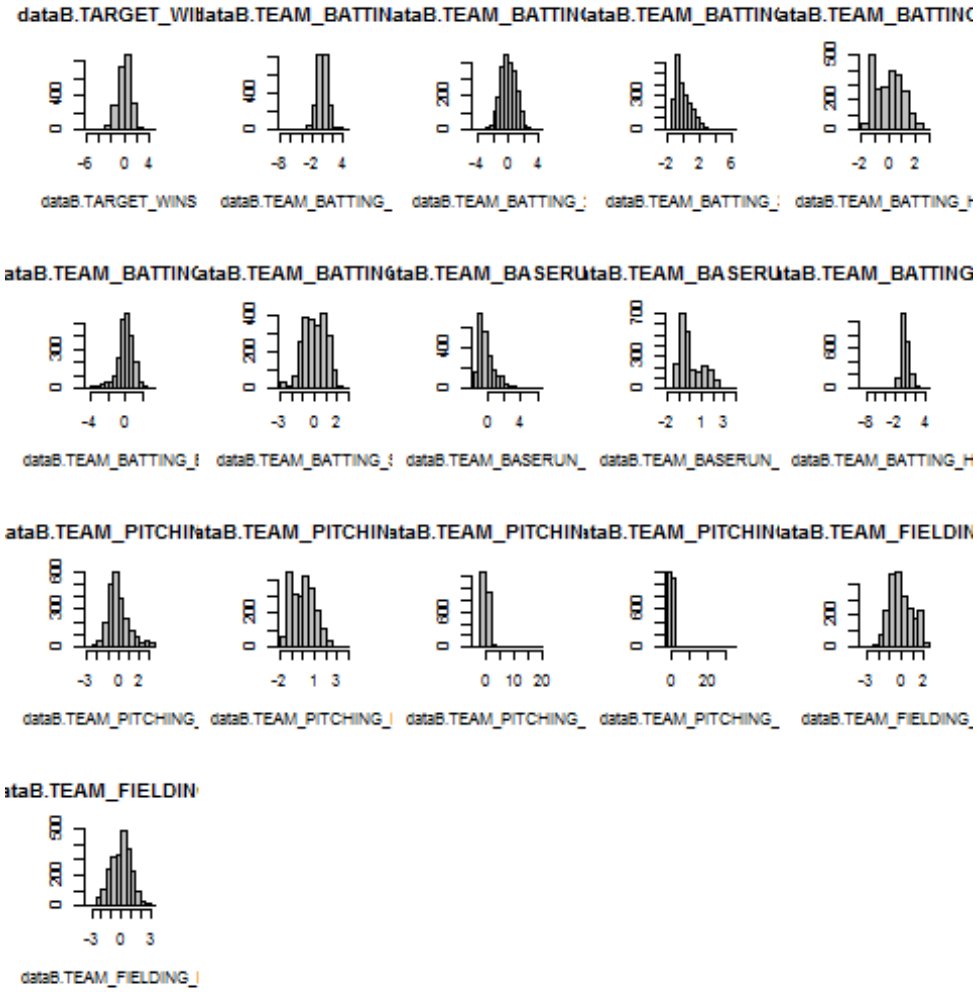
Next we will create the correlation plot to visually identify correlated predictors.



3. Data Preparation

Next we will use the Boxcos transformation to normalize the data. Once normalized, this will allow us to better work with the data. We will use Boxcox transformation to automatically select the appropriate transformation algorytm for our data

Next we will visualize data to see how well it was normalized after boxcox transformation and compare the results to the histograms before we applied the transformation.



We can see that boxcox transformation has had a very positive result on the dataset and data is now normalized for the most part.

4. Build Model

First we will create a basic model consisting of all imputed data

```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = datanorm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9030 -0.5053  0.0024  0.5198  3.5847
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.506e-11  1.675e-02   0.000 1.000000
## TEAM_BATTING_H    4.286e-01  3.619e-02  11.844  < 2e-16 ***
## TEAM_BATTING_2B  -4.260e-02  2.707e-02  -1.574 0.115630
## TEAM_BATTING_3B   1.558e-01  3.024e-02   5.153 2.79e-07 ***
## TEAM_BATTING_HR   2.936e-02  1.072e-01   0.274 0.784229
## TEAM_BATTING_BB   2.484e-01  3.699e-02   6.717 2.34e-11 ***
## TEAM_BATTING_SO  -4.207e-01  4.219e-02  -9.972  < 2e-16 ***
## TEAM_BASERUN_SB   2.277e-01  3.086e-02   7.377 2.27e-13 ***
## TEAM_BASERUN_CS   1.626e-01  3.579e-02   4.544 5.81e-06 ***
## TEAM_BATTING_HBP  1.410e-01  2.017e-02   6.991 3.58e-12 ***
## TEAM_PITCHING_H  -1.833e-01  4.006e-02  -4.576 4.99e-06 ***
## TEAM_PITCHING_HR  2.435e-01  9.746e-02   2.499 0.012527 *
## TEAM_PITCHING_BB -1.382e-01  3.646e-02  -3.789 0.000155 ***
## TEAM_PITCHING_SO  1.738e-01  3.081e-02   5.642 1.89e-08 ***
## TEAM_FIELDING_E  -4.631e-01  3.796e-02 -12.198  < 2e-16 ***
## TEAM_FIELDING_DP -1.854e-01  2.341e-02  -7.918 3.75e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.799 on 2260 degrees of freedom
## Multiple R-squared:  0.3659, Adjusted R-squared:  0.3616
## F-statistic: 86.92 on 15 and 2260 DF,  p-value: < 2.2e-16
```

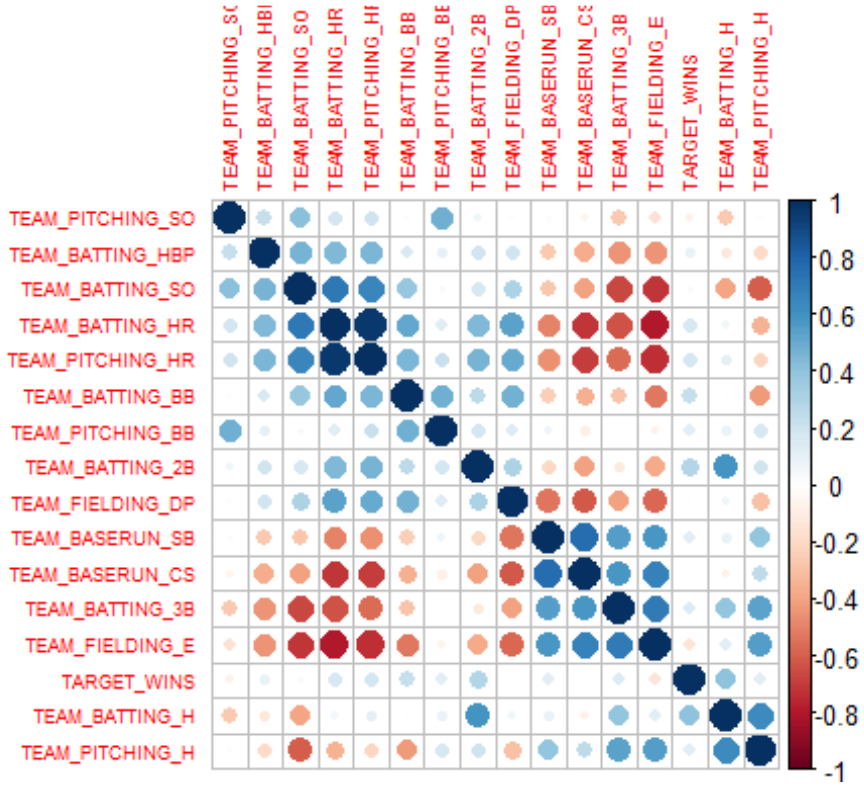
Next we will remove the heavily imputed predictors and compare the results

```
##
## Call:
## lm(formula = TARGET_WINS ~ . - TEAM_BASERUN_CS - TEAM_BATTING_HBP,
##     data = datanorm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9171 -0.5086  0.0080  0.5265  3.6180
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)       1.539e-11  1.698e-02   0.000 1.000000
## TEAM_BATTING_H    4.219e-01  3.667e-02  11.505  < 2e-16 ***
## TEAM_BATTING_2B  -4.407e-02  2.731e-02  -1.613 0.106791
## TEAM_BATTING_3B   1.480e-01  3.021e-02   4.899 1.03e-06 ***
## TEAM_BATTING_HR  -1.058e-01  1.072e-01  -0.987 0.323599
## TEAM_BATTING_BB   2.566e-01  3.735e-02   6.870 8.28e-12 ***
## TEAM_BATTING_SO  -3.783e-01  4.244e-02  -8.914  < 2e-16 ***
## TEAM_BASERUN_SB   2.978e-01  2.593e-02  11.483  < 2e-16 ***
## TEAM_PITCHING_H  -1.886e-01  4.037e-02  -4.671 3.17e-06 ***
## TEAM_PITCHING_HR  3.267e-01  9.813e-02   3.329 0.000884 ***
## TEAM_PITCHING_BB -1.493e-01  3.679e-02  -4.059 5.10e-05 ***
## TEAM_PITCHING_SO  1.878e-01  3.109e-02   6.038 1.81e-09 ***
## TEAM_FIELDING_E  -4.760e-01  3.824e-02 -12.448  < 2e-16 ***
## TEAM_FIELDING_DP -2.162e-01  2.327e-02  -9.292  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8102 on 2262 degrees of freedom
## Multiple R-squared:  0.3474, Adjusted R-squared:  0.3436
## F-statistic: 92.62 on 13 and 2262 DF,  p-value: < 2.2e-16
```

So it appears that imputing the 2 columns did not hurt our model in fact the extra data slightly helps the model so we will leave those in.

For the next model we will try to eliminate some of the highly correlated predictors, so we will run correlation matrix again on the normalized data



We will remove TEAM_BATTING_HR and TEAM_FIELDING_E as they are very highly correlated to other predictors

```
##
## Call:
## lm(formula = TARGET_WINS ~ . - TEAM_BATTING_HR - TEAM_FIELDING_E,
##     data = datanorm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7743 -0.5276  0.0200  0.5552  3.4505
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.393e-11  1.728e-02   0.000 1.000000
## TEAM_BATTING_H    4.793e-01  3.622e-02  13.233  < 2e-16 ***
## TEAM_BATTING_2B   1.761e-02  2.747e-02   0.641 0.521537
## TEAM_BATTING_3B   6.668e-02  2.956e-02   2.256 0.024173 *
## TEAM_BATTING_BB   2.910e-01  3.512e-02   8.285  < 2e-16 ***
## TEAM_BATTING_SO  -3.219e-01  4.273e-02  -7.534 7.09e-14 ***
## TEAM_BASERUN_SB   1.826e-01  3.150e-02   5.796 7.76e-09 ***
## TEAM_BASERUN_CS   1.307e-01  3.648e-02   3.582 0.000348 ***
## TEAM_BATTING_HBP  1.620e-01  2.063e-02   7.855 6.15e-15 ***
## TEAM_PITCHING_H  -3.055e-01  3.851e-02  -7.934 3.31e-15 ***
## TEAM_PITCHING_HR  3.493e-01  3.674e-02   9.505  < 2e-16 ***
## TEAM_PITCHING_BB -1.544e-01  3.371e-02  -4.579 4.94e-06 ***
## TEAM_PITCHING_SO  1.732e-01  3.016e-02   5.745 1.05e-08 ***
## TEAM_FIELDING_DP -1.487e-01  2.395e-02  -6.210 6.30e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8246 on 2262 degrees of freedom
## Multiple R-squared:  0.324,  Adjusted R-squared:  0.3201
## F-statistic: 83.39 on 13 and 2262 DF,  p-value: < 2.2e-16
```

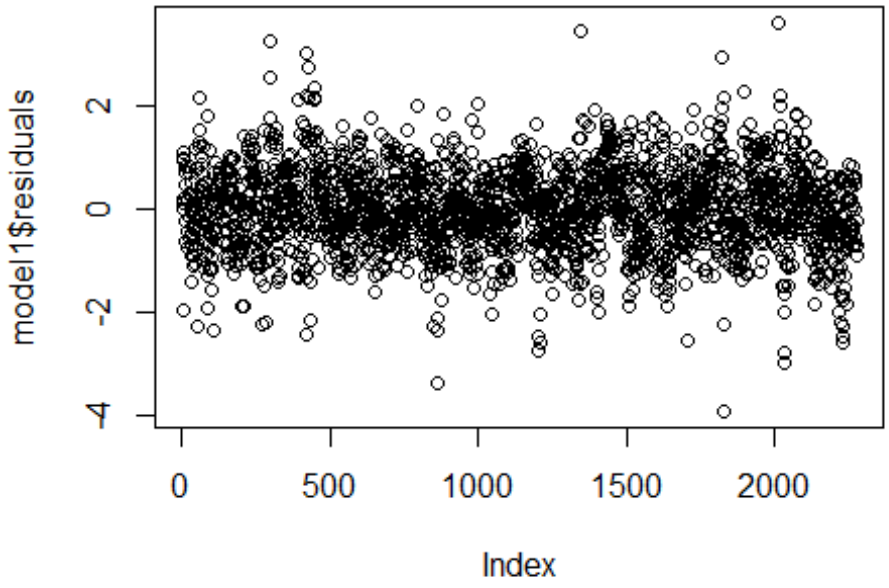
It appears that removing highly correlated predictors also does not help our model.

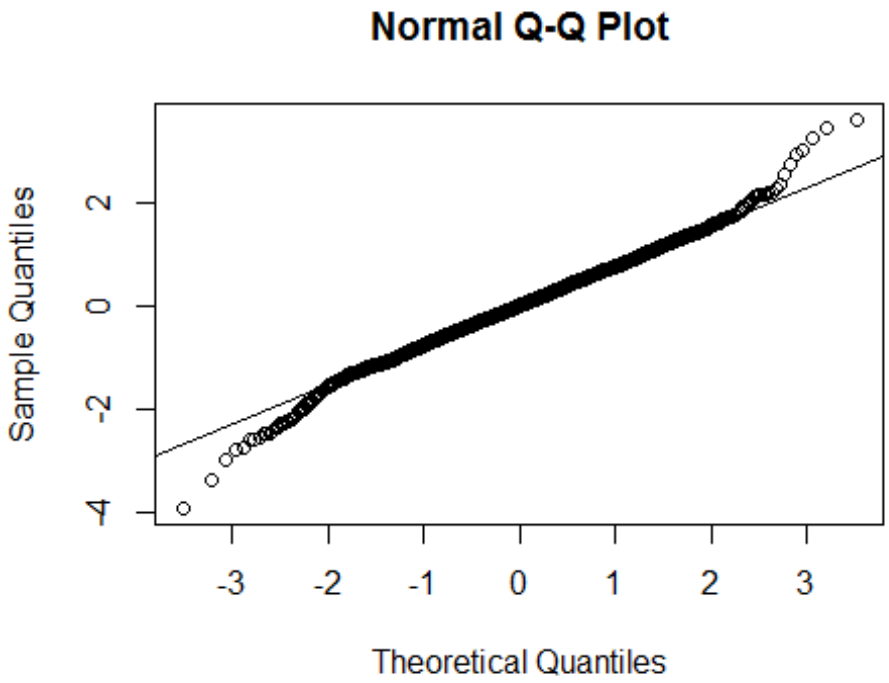
After comparing RMSE, R-squared and F-statistics of the 3 models, we will go ahead and use the model1 it is the simplest model and uses all the predictors, but since we did not find any significant benefit with other models, We will use the simplest model containing all the predictors.

5. Select Model

We decided to select the final model as it is the model in its purest form containing all the predictors, we tried to omit imputed data, drop some of the highly correlated predictors and removing some of the less relevant predictors, but did not see much improvement, therefore we decided to stay with the base model.

Next we will review residual plot and qqplot of the model. We want to make sure the residuals are random and the variance is constant.





```
##   missForest iteration 1 in progress...done!
##   missForest iteration 2 in progress...done!
##   missForest iteration 3 in progress...done!
##   missForest iteration 4 in progress...done!
```

References

Complete code is located at:
https://github.com/jelikish/Cuny1/blob/master/Spring2018/621/hw1