Overview

In this homework assignment, you will explore, analyze and model a data set containing approximately 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.

Your objective is to build a multiple linear regression model on the training data to predict the number of wins for the team. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:

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

Data Exploration:

Load the data and understand the data by using some stats and plots. The dataset consists of 17 elements, with 2276 total cases. There are multiple variables with missing (NA) values and TEAM-BATTING_HBP has the highest NAs.

Summary and descriptive statistics Descriptive statistics is used here to summarize the data to gather insights into the information contained in the dataset.

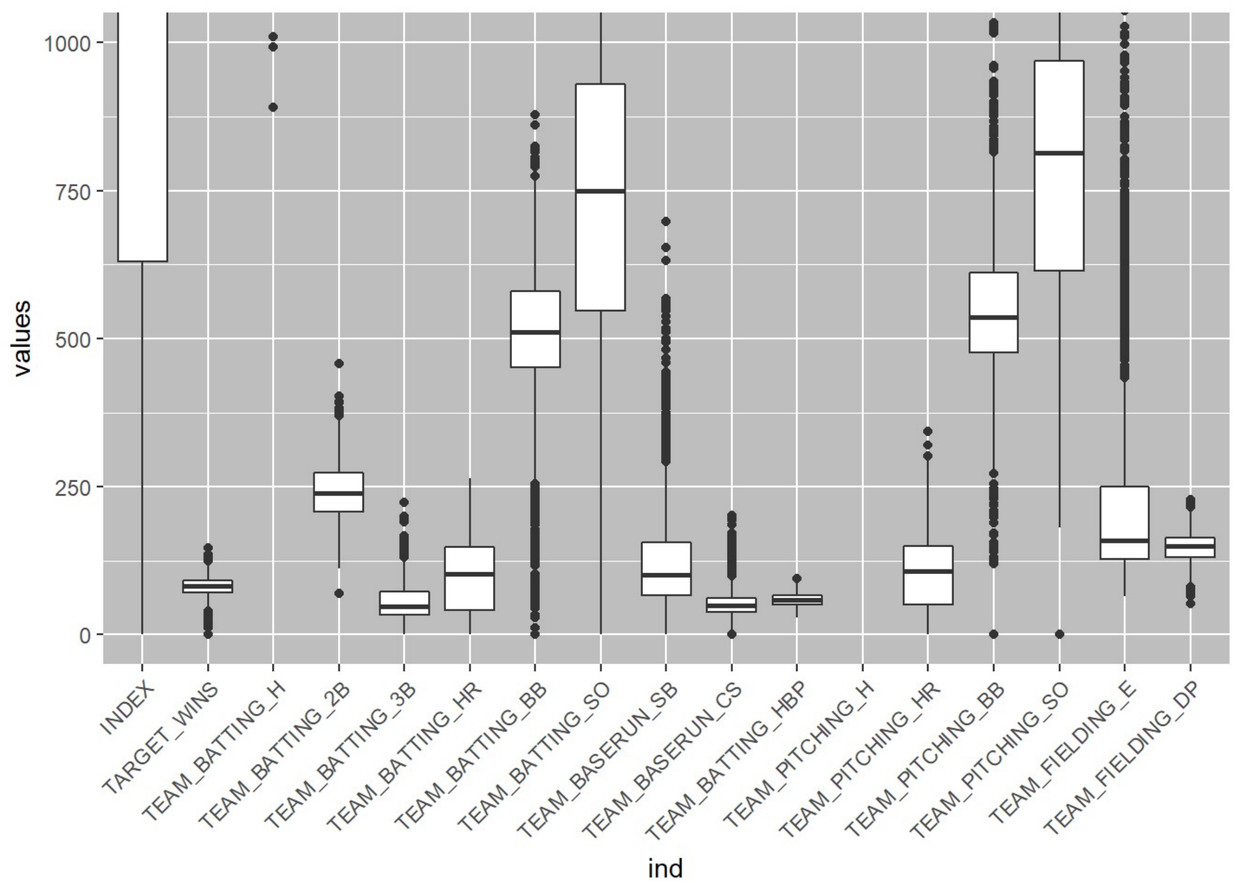
The descriptive statistics below shows the mean, mode, standard deviation, minimum and maximum of each variable in the dataset.

```
##      INDEX         TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B
##  Min.   :   1.0   Min.   :  0.00   Min.   : 891   Min.   : 69.0  
##  1st Qu.: 630.8   1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0  
##  Median :1270.5   Median : 82.00   Median :1454   Median :238.0  
##  Mean   :1268.5   Mean   : 80.79   Mean   :1469   Mean   :241.2  
##  3rd Qu.:1915.5   3rd Qu.: 92.00   3rd Qu.:1537   3rd Qu.:273.0  
##  Max.   :2535.0   Max.   :146.00   Max.   :2554   Max.   :458.0  
##                                                                  
##  TEAM_BATTING_3B  TEAM_BATTING_HR  TEAM_BATTING_BB TEAM_BATTING_SO
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.0   Min.   :   0.0 
##  1st Qu.: 34.00   1st Qu.: 42.00   1st Qu.:451.0   1st Qu.: 548.0 
##  Median : 47.00   Median :102.00   Median :512.0   Median : 750.0 
##  Mean   : 55.25   Mean   : 99.61   Mean   :501.6   Mean   : 735.6 
##  3rd Qu.: 72.00   3rd Qu.:147.00   3rd Qu.:580.0   3rd Qu.: 930.0 
##  Max.   :223.00   Max.   :264.00   Max.   :878.0   Max.   :1399.0 
##                                                    NA's   :102    
##  TEAM_BASERUN_SB TEAM_BASERUN_CS TEAM_BATTING_HBP TEAM_PITCHING_H
##  Min.   :  0.0   Min.   :  0.0   Min.   :29.00    Min.   : 1137  
##  1st Qu.: 66.0   1st Qu.: 38.0   1st Qu.:50.50    1st Qu.: 1419  
##  Median :101.0   Median : 49.0   Median :58.00    Median : 1518  
##  Mean   :124.8   Mean   : 52.8   Mean   :59.36    Mean   : 1779  
##  3rd Qu.:156.0   3rd Qu.: 62.0   3rd Qu.:67.00    3rd Qu.: 1682  
##  Max.   :697.0   Max.   :201.0   Max.   :95.00    Max.   :30132  
##  NA's   :131     NA's   :772     NA's   :2085                    
##  TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E
##  Min.   :  0.0    Min.   :   0.0   Min.   :    0.0   Min.   :  65.0 
##  1st Qu.: 50.0    1st Qu.: 476.0   1st Qu.:  615.0   1st Qu.: 127.0 
##  Median :107.0    Median : 536.5   Median :  813.5   Median : 159.0 
```

```
##  Mean   :105.7    Mean   : 553.0   Mean   :  817.7   Mean   : 246.5
##  3rd Qu.:150.0    3rd Qu.: 611.0   3rd Qu.:  968.0   3rd Qu.: 249.2
##  Max.   :343.0    Max.   :3645.0   Max.   :19278.0   Max.   :1898.0
##                                    NA's   :102
##  TEAM_FIELDING_DP
##  Min.   : 52.0
##  1st Qu.:131.0
##  Median :149.0
##  Mean   :146.4
##  3rd Qu.:164.0
##  Max.   :228.0
##  NA's   :286
```

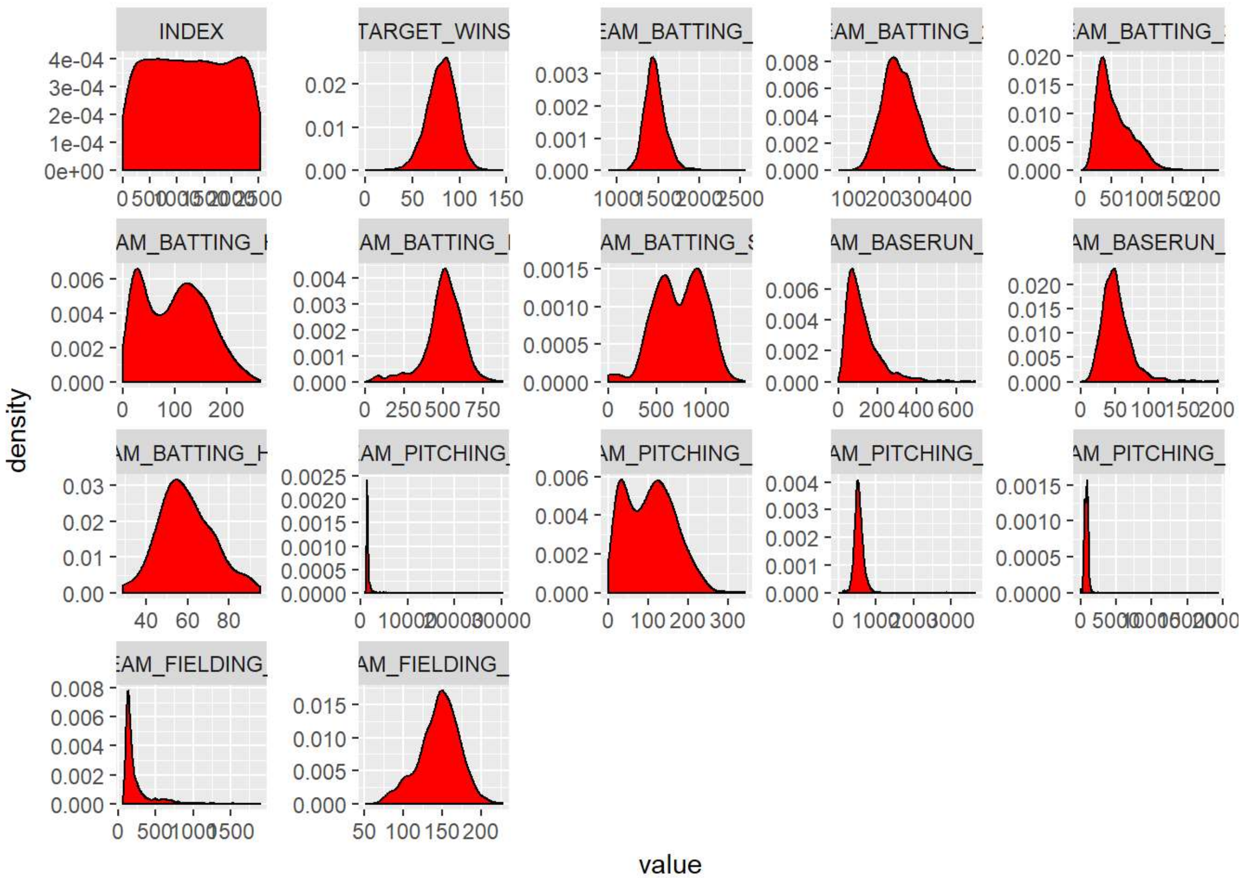
Checking for outliers:

Outlier detection is very important for the model performance. Below you can see that there are some outliers in that data.



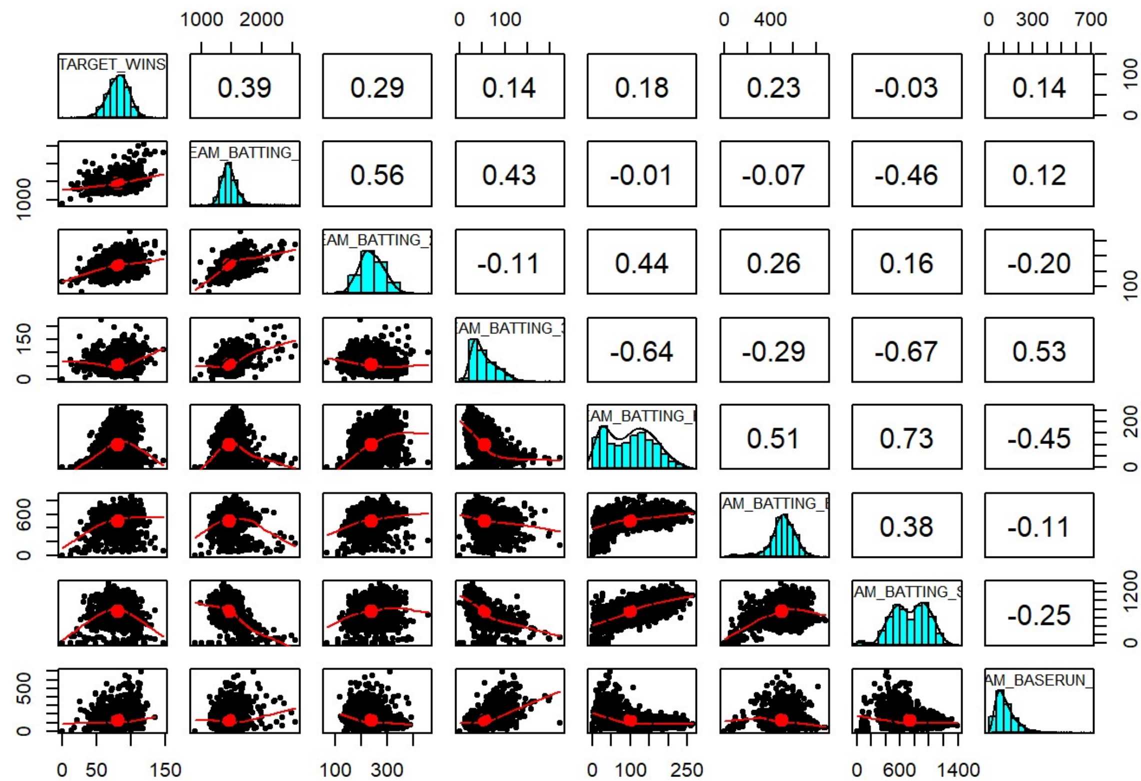
Checking for skewness in the data:

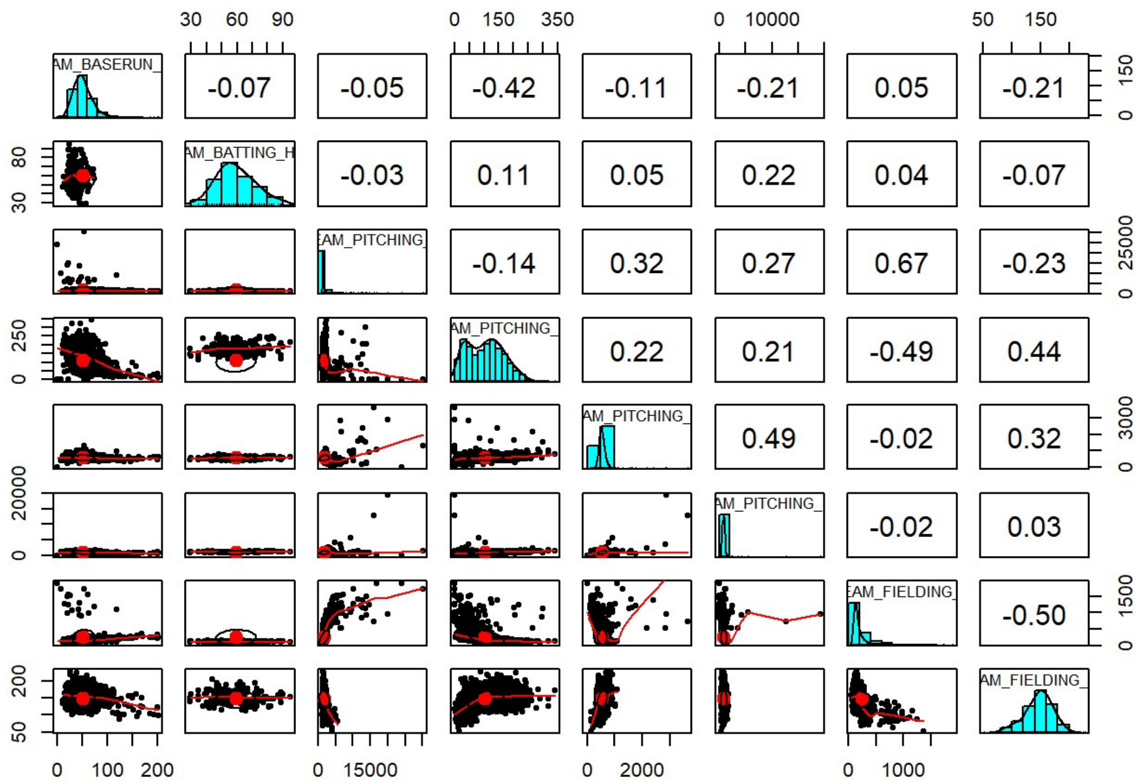
Examining skewness and outliers in the data is important prior to choosing the model. This is important because some models will require transformation of the data. As seen there are several variables that are skewed and also there are outliers.



Finding correlations:

We can see there are some positively and some negatively correlated variables. Looking at the plot, we can see that certain variables are more related than others.


TARGET_WINS
EAM_BATTING_
EAM_BATTING_2
AM_BATTING_3
AM_BATTING_I
AM_BATTING_E
AM_BATTING_S
AM_BASERUN_
0.39
0.29
0.14
0.18
0.23
-0.03
0.14
0.56
0.43
-0.01
-0.07
-0.46
0.12
-0.11
0.44
0.26
0.16
-0.20
-0.64
-0.29
-0.67
0.53
0.51
0.73
-0.45
0.38
-0.11
-0.25
1000 2000
0 100
0 400
0 300 700
0 100
1000
100
0 150
0 200
0 600
0 1200
0 500
0 50 150
100 300
0 100 250
0 600 1400



Data Preparation:

Removal of Data:

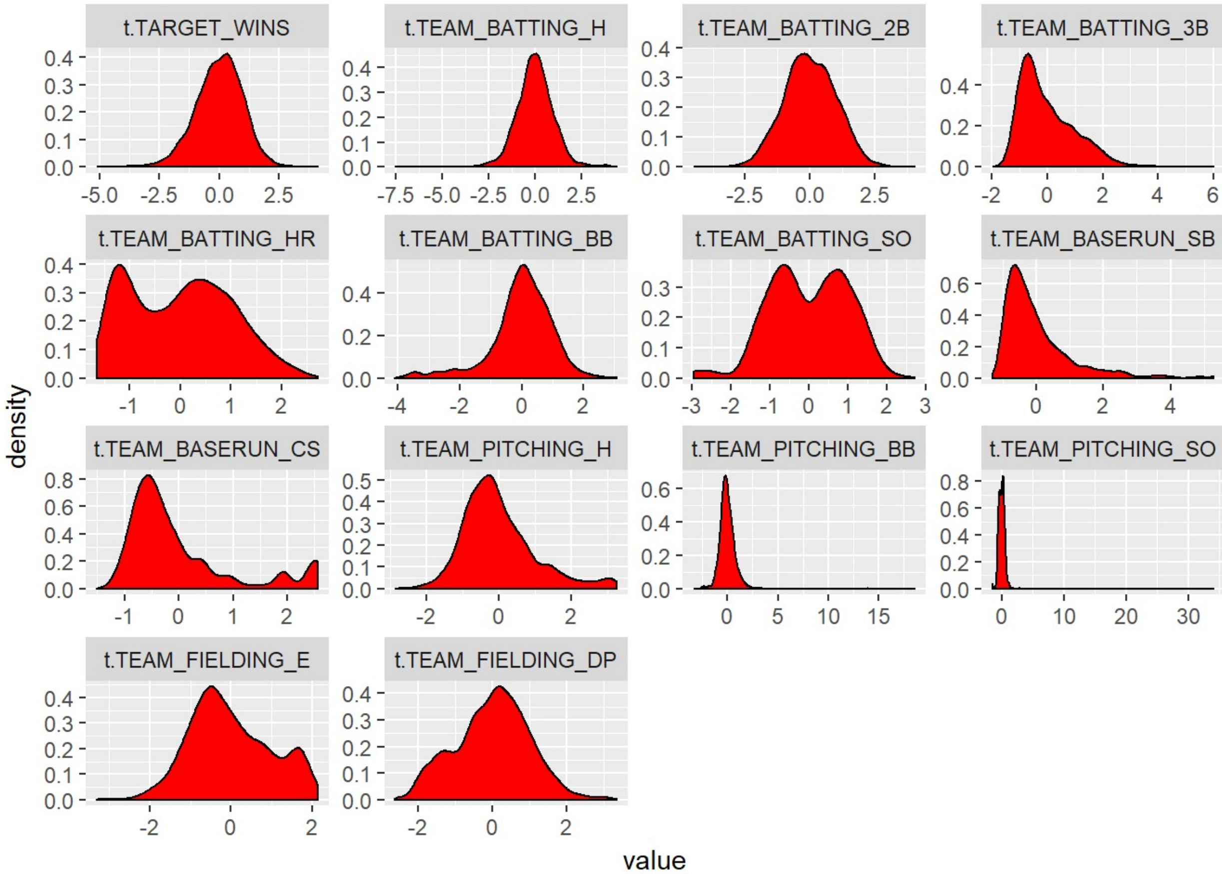
The variable TEAM_BATTING_HBP is having mostly missing values so the variable will be removed completely. TEAM_PITCHING_HR and TEAM_BATTING_HR are highly correlated, so we can remove one of them.

Imputation of Missing (NA) values:

The data will be imputed via prediction using the MICE (Multivariate Imputation) library using pmm - predictive mean matching method.

Data transformation:

Centering and scaling was used to transform individual predictors in the dataset using the caret library. Below is the plot after the data transformation.



Build Models:

Model1:

With all variables:

```
##
## Call:
## lm(formula = t.TARGET_WINS ~ ., data = mtd_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4847 -0.5019 -0.0032  0.5140  3.8244
##
## Coefficients:
```

```
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.549e-11  1.705e-02   0.000    1.000
## t.TEAM_BATTING_H     4.223e-01  3.654e-02  11.558  < 2e-16 ***
## t.TEAM_BATTING_2B   -3.720e-02  2.754e-02  -1.351    0.177
## t.TEAM_BATTING_3B    1.708e-01  2.998e-02   5.699 1.37e-08 ***
## t.TEAM_BATTING_HR    2.257e-01  3.805e-02   5.932 3.45e-09 ***
## t.TEAM_BATTING_BB    1.466e-01  3.481e-02   4.213 2.62e-05 ***
## t.TEAM_BATTING_SO   -3.549e-01  4.063e-02  -8.736  < 2e-16 ***
## t.TEAM_BASERUN_SB    2.369e-01  3.225e-02   7.345 2.87e-13 ***
## t.TEAM_BASERUN_CS    4.776e-02  3.393e-02   1.408    0.159
## t.TEAM_PITCHING_H   -1.899e-01  3.853e-02  -4.928 8.90e-07 ***
## t.TEAM_PITCHING_BB   4.180e-03  3.361e-02   0.124    0.901
## t.TEAM_PITCHING_SO   1.247e-01  2.980e-02   4.185 2.97e-05 ***
## t.TEAM_FIELDING_E   -4.872e-01  3.854e-02 -12.641  < 2e-16 ***
## t.TEAM_FIELDING_DP  -2.020e-01  2.325e-02  -8.686  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8135 on 2262 degrees of freedom
## Multiple R-squared:  0.3419, Adjusted R-squared:  0.3382
## F-statistic: 90.42 on 13 and 2262 DF,  p-value: < 2.2e-16
```

Model2:

With only the significant variables:

```
##
## Call:
## lm(formula = t.TARGET_WINS ~ t.TEAM_BATTING_H + t.TEAM_BATTING_3B +
##     t.TEAM_BATTING_HR + t.TEAM_BATTING_BB + t.TEAM_BATTING_SO +
##     t.TEAM_BASERUN_SB + t.TEAM_PITCHING_SO + t.TEAM_PITCHING_H +
##     t.TEAM_PITCHING_SO + t.TEAM_FIELDING_E + t.TEAM_FIELDING_DP,
##     data = mtd_final)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.5299 -0.4978 -0.0048  0.5167  3.7841 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         1.541e-11  1.706e-02   0.000        1    
## t.TEAM_BATTING_H    3.920e-01  3.055e-02  12.830  < 2e-16 ***
## t.TEAM_BATTING_3B   1.776e-01  2.976e-02   5.966 2.81e-09 ***
## t.TEAM_BATTING_HR   2.238e-01  3.766e-02   5.942 3.26e-09 ***
## t.TEAM_BATTING_BB   1.494e-01  2.232e-02   6.692 2.76e-11 ***
## t.TEAM_BATTING_SO  -3.653e-01  3.906e-02  -9.354  < 2e-16 ***
## t.TEAM_BASERUN_SB   2.664e-01  2.607e-02  10.218  < 2e-16 ***
## t.TEAM_PITCHING_SO  1.200e-01  2.197e-02   5.462 5.23e-08 ***
## t.TEAM_PITCHING_H  -1.910e-01  3.550e-02  -5.382 8.14e-08 ***
## t.TEAM_FIELDING_E  -4.698e-01  3.753e-02 -12.517  < 2e-16 ***
## t.TEAM_FIELDING_DP -2.071e-01  2.232e-02  -9.281  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.8137 on 2265 degrees of freedom
## Multiple R-squared:  0.3407, Adjusted R-squared:  0.3378 
## F-statistic: 117.1 on 10 and 2265 DF,  p-value: < 2.2e-16
```

Model3:

Further reducing the variables(TEAM_PITCHING_SO and TEAM_BATTING_SO are having high correlation, TEAM_BATTING_H and TEAM_PITCHING_H are also having high correlation, TEAM_BATTING_SO and TEAM_PITCHING_SO are also having high correlation):

```
## 
## Call:
## lm(formula = t.TARGET_WINS ~ t.TEAM_BATTING_H + t.TEAM_BATTING_3B + 
##     t.TEAM_BATTING_HR + t.TEAM_BATTING_BB + t.TEAM_BATTING_SO + 
```

```
##     t.TEAM_BASERUN_SB + t.TEAM_FIELDING_E + t.TEAM_FIELDING_DP,
##     data = mtd_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4615 -0.5149 -0.0021  0.5225  4.5628
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.440e-12  1.720e-02   0.000        1
## t.TEAM_BATTING_H     2.885e-01  2.431e-02  11.870  < 2e-16 ***
## t.TEAM_BATTING_3B    1.862e-01  2.986e-02   6.234 5.40e-10 ***
## t.TEAM_BATTING_HR    1.856e-01  3.741e-02   4.961 7.52e-07 ***
## t.TEAM_BATTING_BB    1.803e-01  2.113e-02   8.532  < 2e-16 ***
## t.TEAM_BATTING_SO   -2.504e-01  3.478e-02  -7.200 8.15e-13 ***
## t.TEAM_BASERUN_SB    2.244e-01  2.501e-02   8.972  < 2e-16 ***
## t.TEAM_FIELDING_E   -4.961e-01  3.645e-02 -13.610  < 2e-16 ***
## t.TEAM_FIELDING_DP  -2.120e-01  2.240e-02  -9.464  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8207 on 2267 degrees of freedom
## Multiple R-squared:  0.3289, Adjusted R-squared:  0.3265
## F-statistic: 138.9 on 8 and 2267 DF,  p-value: < 2.2e-16
```

Select models and predictions:

From the three models, I decided to use model3 for the predictions considering its more parsimonious model. There is no significant difference in R2, Adjusted R2 and RMSE even when i did the treatment for multi-collinearity.

Predictions:

For the evaluation dataset also we will be doing all the preprocessing steps that we did for the training data.

```r
eval_data <- predict(model3, newdata = med_final, interval="prediction")
eval_data
```

```
##              fit         lwr        upr
## 1   -1.20733701 -2.82022586  0.40555184
## 2   -0.98275416 -2.59426358  0.62875526
## 3   -0.59511173 -2.20593900  1.01571555
## 4    0.28690291 -1.32422638  1.89803219
## 5   -1.05350237 -2.66620271  0.55919797
## 6   -0.87989580 -2.49249885  0.73270725
## 7    0.31907656 -1.29568037  1.93383349
## 8   -0.69090577 -2.30245499  0.92064344
## 9   -0.67265861 -2.28452518  0.93920795
## 10  -0.55720513 -2.16780354  1.05339327
## 11  -0.91111908 -2.52361968  0.70138151
## 12  -0.02731931 -1.64015563  1.58551702
## 13   0.08585405 -1.52837761  1.70008570
## 14   0.01017680 -1.60273913  1.62309272
## 15   0.34475957 -1.26956438  1.95908352
## 16  -0.50040118 -2.11226411  1.11146175
## 17  -0.72579980 -2.33767302  0.88607342
## 18  -0.11818645 -1.72912563  1.49275274
## 19  -0.67811962 -2.29031657  0.93407733
## 20   0.37585547 -1.23619352  1.98790446
## 21   0.33408626 -1.27771461  1.94588712
## 22   0.20203863 -1.40969818  1.81377545
## 23   0.10075249 -1.51084353  1.71234850
## 24  -0.70905687 -2.32065453  0.90254078
## 25   0.14163136 -1.47047522  1.75373794
## 26   0.47214873 -1.14102732  2.08532478
## 27  -0.60985095 -2.23470893  1.01500702
## 28  -0.52902760 -2.13984965  1.08179445
## 29   0.31886420 -1.29379671  1.93152512
## 30  -0.54480059 -2.15751196  1.06791078
## 31   0.71240335 -0.90019930  2.32500600
```

```
## 32   0.36202491 -1.24886238  1.97291221
## 33   0.34297429 -1.26904188  1.95499046
## 34   0.20885994 -1.40542788  1.82314777
## 35   0.04053154 -1.57077462  1.65183770
## 36   0.11531639 -1.49870494  1.72933771
## 37  -0.25147218 -1.86157985  1.35863550
## 38   0.46654868 -1.14711715  2.08021451
## 39   0.06102769 -1.55044446  1.67249983
## 40   0.46685128 -1.14579270  2.07949526
## 41   0.17041526 -1.44200934  1.78283985
## 42   1.33936894 -0.27554348  2.95428137
## 43  -1.58376700 -3.21035551  0.04282151
## 44   1.64463604  0.02071209  3.26855999
## 45   0.66703450 -0.94780278  2.28187178
## 46   1.00738063 -0.60576550  2.62052676
## 47   1.04129068 -0.57156248  2.65414384
## 48  -0.44146522 -2.05260335  1.16967290
## 49  -0.81394162 -2.42521716  0.79733391
## 50  -0.12195282 -1.73268996  1.48878431
## 51  -0.34503001 -1.95604738  1.26598736
## 52   0.21352115 -1.39813082  1.82517311
## 53  -0.44335264 -2.05549204  1.16878676
## 54  -0.27644900 -1.88856859  1.33567059
## 55  -0.62328873 -2.23414084  0.98756339
## 56   0.07608083 -1.53565286  1.68781452
## 57   0.67090466 -0.94148178  2.28329111
## 58  -0.40364360 -2.01517792  1.20789073
## 59  -1.11612811 -2.72892181  0.49666559
## 60  -0.21833311 -1.82917534  1.39250912
## 61   0.42886163 -1.18219149  2.03991475
## 62   0.05674609 -1.55908886  1.67258103
## 63   0.41479668 -1.19623218  2.02582554
## 64   0.29610193 -1.31777380  1.90997766
## 65   0.37404321 -1.23990770  1.98799412
```

```
## 66   1.38823955 -0.22829455  3.00477366
## 67  -0.62628163 -2.23800178  0.98543852
## 68  -0.35210034 -1.96415217  1.25995150
## 69  -0.21650258 -1.82811921  1.39511404
## 70   0.42366475 -1.18941599  2.03674549
## 71   0.27251288 -1.34105249  1.88607826
## 72  -0.38453357 -1.99978788  1.23072075
## 73  -0.20379979 -1.81718648  1.40958689
## 74   0.53038477 -1.08447059  2.14524014
## 75  -0.27323424 -1.88627776  1.33980929
## 76  -0.25594513 -1.86897823  1.35708797
## 77   0.42404340 -1.18714322  2.03523003
## 78   0.06246996 -1.54862632  1.67356624
## 79  -0.64931484 -2.26055271  0.96192303
## 80  -0.47748091 -2.08898462  1.13402281
## 81   0.19465583 -1.41652330  1.80583496
## 82   0.32393378 -1.28733360  1.93520116
## 83   0.79924224 -0.81298714  2.41147162
## 84  -0.48456128 -2.09761624  1.12849368
## 85   0.24334017 -1.36849871  1.85517904
## 86  -0.20179059 -1.81483692  1.41125574
## 87   0.16465019 -1.44773433  1.77703470
## 88   0.31905366 -1.29120525  1.92931258
## 89   0.80537636 -0.80782495  2.41857767
## 90   0.76121791 -0.85040630  2.37284212
## 91   0.21905734 -1.39328083  1.83139550
## 92   0.73998824 -0.87844396  2.35842043
## 93  -0.50396241 -2.11489894  1.10697412
## 94   0.10608083 -1.50562012  1.71778178
## 95   0.07545302 -1.53616873  1.68707476
## 96   0.09130598 -1.52000335  1.70261531
## 97   0.61389533 -1.00067677  2.22846743
## 98   1.14177435 -0.47252381  2.75607251
## 99   0.43769586 -1.17485332  2.05024504
```

```
## 100  0.37483017 -1.23825292  1.98791326
## 101 -0.10384623 -1.71549114  1.50779868
## 102 -0.48368008 -2.09490820  1.12754804
## 103  0.25685762 -1.35350534  1.86722058
## 104  0.26396050 -1.34770049  1.87562149
## 105 -0.49624664 -2.11013904  1.11764576
## 106 -0.97767800 -2.59157434  0.63621834
## 107 -1.53346506 -3.15002578  0.08309565
## 108 -0.07288058 -1.68517753  1.53941638
## 109  0.75944824 -0.85225881  2.37115528
## 110 -1.41132133 -3.02581517  0.20317251
## 111  0.36168807 -1.24900685  1.97238298
## 112  0.41946277 -1.19171114  2.03063668
## 113  0.76497301 -0.84579115  2.37573716
## 114  0.71587158 -0.89557533  2.32731849
## 115  0.04779631 -1.56352049  1.65911312
## 116  0.04233730 -1.56870918  1.65338379
## 117  0.26826032 -1.34409942  1.88062006
## 118  0.10031777 -1.50997534  1.71061089
## 119 -0.44090675 -2.05254918  1.17073568
## 120  0.03809180 -1.57492298  1.65110658
## 121  0.95667002 -0.65620124  2.56954128
## 122 -0.68686811 -2.29912211  0.92538588
## 123 -0.74970645 -2.36182781  0.86241491
## 124 -0.96716685 -2.58276841  0.64843471
## 125 -0.83077148 -2.44292203  0.78137906
## 126  0.19113161 -1.42057693  1.80284015
## 127  0.38115804 -1.23100821  1.99332430
## 128 -0.36107579 -1.97209514  1.24994357
## 129  0.64349288 -0.96817157  2.25515732
## 130  0.43215465 -1.17979867  2.04410798
## 131  0.20787858 -1.40326095  1.81901812
## 132  0.13841392 -1.47364615  1.75047398
## 133 -0.67841669 -2.29460174  0.93776836
```

```
## 134 -0.06358876 -1.67567606  1.54849854
## 135  1.24427368 -0.37288989  2.86143724
## 136 -0.20285013 -1.81602567  1.41032541
## 137 -0.26148706 -1.87281691  1.34984280
## 138 -0.22615042 -1.83675364  1.38445281
## 139  1.10903335 -0.51067021  2.72873691
## 140 -0.06355680 -1.67467165  1.54755806
## 141 -1.23410379 -2.84748492  0.37927733
## 142 -0.46759991 -2.07967852  1.14447869
## 143  0.60449424 -1.00747938  2.21646785
## 144 -0.58138858 -2.19317964  1.03040247
## 145 -0.18651159 -1.79845018  1.42542699
## 146 -0.40560691 -2.01621618  1.20500236
## 147 -0.42587454 -2.03726230  1.18551322
## 148  0.03574422 -1.57511276  1.64660120
## 149 -0.13004157 -1.74235657  1.48227343
## 150  0.34670622 -1.26413888  1.95755132
## 151  0.11203640 -1.49988670  1.72395951
## 152  0.46835586 -1.14590010  2.08261182
## 153 -1.29665815 -2.91913748  0.32582118
## 154 -1.07028852 -2.68258000  0.54200295
## 155 -0.01689578 -1.62864489  1.59485333
## 156 -0.99145324 -2.60408026  0.62117378
## 157  0.83887791 -0.77383410  2.45158991
## 158 -0.65882024 -2.27057998  0.95293949
## 159  0.54969462 -1.06223786  2.16162710
## 160 -0.23594490 -1.84844565  1.37655584
## 161  1.21252451 -0.40310316  2.82815218
## 162  1.66385849  0.04771429  3.28000269
## 163  0.98825730 -0.62443702  2.60095161
## 164  1.38272308 -0.23334455  2.99879071
## 165  1.10221426 -0.51381035  2.71823886
## 166  0.94769704 -0.66641514  2.56180922
## 167  0.15195781 -1.46019228  1.76410789
```

```
## 168  0.16676299 -1.44593415  1.77946013
## 169 -0.72379047 -2.33614896  0.88856803
## 170 -0.02278968 -1.63477255  1.58919319
## 171  0.65862593 -0.95320888  2.27046074
## 172  0.48137419 -1.12988834  2.09263672
## 173  0.12167545 -1.48924061  1.73259151
## 174  0.79987887 -0.81204314  2.41180089
## 175  0.01618564 -1.59463534  1.62700663
## 176 -0.15021456 -1.76203348  1.46160437
## 177  0.11701228 -1.49579352  1.72981809
## 178 -0.87731581 -2.48990234  0.73527071
## 179 -0.32193016 -1.93219590  1.28833559
## 180 -0.17549973 -1.78634734  1.43534788
## 181  0.45969801 -1.15562275  2.07501877
## 182  0.32056478 -1.29212493  1.93325448
## 183  0.45859366 -1.15328172  2.07046903
## 184  0.54983277 -1.06174159  2.16140713
## 185  0.56438577 -1.05246171  2.18123324
## 186  0.85718200 -0.76316203  2.47752603
## 187  0.49595760 -1.11889770  2.11081291
## 188 -0.76756555 -2.37989888  0.84476778
## 189 -1.14495687 -2.75709871  0.46718498
## 190  1.77924965  0.16141830  3.39708101
## 191 -0.41012079 -2.02259721  1.20235563
## 192  0.05574687 -1.55508654  1.66658029
## 193 -0.58201169 -2.19300462  1.02898125
## 194 -0.46240638 -2.07356633  1.14875357
## 195 -0.40231981 -2.01479563  1.21015602
## 196 -1.08913399 -2.70185707  0.52358909
## 197 -0.43111712 -2.04183649  1.17960225
## 198  0.79688783 -0.81716105  2.41093671
## 199  0.07175139 -1.53934242  1.68284520
## 200  0.31238287 -1.29897961  1.92374536
## 201 -0.56851917 -2.18186704  1.04482870
```

```
## 202  0.15440653 -1.45739299  1.76620606
## 203 -0.07018623 -1.68401522  1.54364277
## 204  0.65236918 -0.95887172  2.26361009
## 205  0.07676805 -1.53463097  1.68816707
## 206  0.22989135 -1.38138352  1.84116623
## 207  0.11840273 -1.49366352  1.73046898
## 208  0.17306915 -1.43890297  1.78504127
## 209  0.12656131 -1.48476510  1.73788773
## 210 -0.48683492 -2.09854503  1.12487518
## 211  1.43138968 -0.18224288  3.04502224
## 212  0.30471847 -1.30709570  1.91653264
## 213  0.03602408 -1.57574882  1.64779697
## 214 -1.20659967 -2.81864144  0.40544210
## 215 -0.79412720 -2.40691091  0.81865652
## 216  0.15149928 -1.45962288  1.76262143
## 217 -0.22039601 -1.83443584  1.39364382
## 218  0.66757903 -0.94408588  2.27924393
## 219 -0.18256583 -1.79335710  1.42822545
## 220  0.09789910 -1.51300026  1.70879847
## 221 -0.36145536 -1.97295483  1.25004412
## 222 -0.59208254 -2.20464606  1.02048098
## 223 -0.07889398 -1.69005991  1.53227195
## 224 -0.31980425 -1.93375035  1.29414184
## 225 -0.02543383 -1.64906980  1.59820215
## 226 -0.19645387 -1.80713088  1.41422314
## 227 -0.14052288 -1.75150081  1.47045505
## 228 -0.18205069 -1.79425327  1.43015188
## 229  0.44199527 -1.16911381  2.05310436
## 230 -0.26949127 -1.88235594  1.34337339
## 231 -0.01188703 -1.62480513  1.60103108
## 232  0.58451562 -1.02682349  2.19585473
## 233  0.02354812 -1.58897152  1.63606776
## 234  0.25702737 -1.35544334  1.86949807
## 235 -0.20997497 -1.82070395  1.40075401
```

```
## 236 -0.35358054 -1.96412101  1.25695992
## 237 -0.30379566 -1.91685704  1.30926571
## 238  0.09498279 -1.51736930  1.70733489
## 239  0.78578279 -0.82717524  2.39874083
## 240 -0.69263314 -2.30377007  0.91850379
## 241  0.32207223 -1.28883570  1.93298017
## 242  0.75439457 -0.85799678  2.36678593
## 243  0.28414799 -1.32725259  1.89554858
## 244  0.17107585 -1.44070315  1.78285485
## 245 -1.51881090 -3.13427675  0.09665494
## 246  0.11323414 -1.49909571  1.72556400
## 247 -0.18464372 -1.79544146  1.42615403
## 248  0.18614103 -1.42514783  1.79742990
## 249 -0.35972057 -1.97087405  1.25143291
## 250  0.38619624 -1.22788325  2.00027572
## 251  0.18046579 -1.43129271  1.79222428
## 252 -0.68605576 -2.29969940  0.92758789
## 253  0.80067365 -0.81286939  2.41421669
## 254 -2.76155514 -4.38814877 -1.13496151
## 255 -0.80845155 -2.41972677  0.80282367
## 256 -0.34039494 -1.95391587  1.27312598
## 257  0.20497135 -1.40694314  1.81688584
## 258  0.06663220 -1.54441074  1.67767514
## 259 -0.33100279 -1.94295047  1.28094489
```

```
summary(eval_data)
```

```
##       fit                 lwr               upr
##  Min.   :-2.76156   Min.   :-4.3881   Min.   :-1.135
##  1st Qu.:-0.40786   1st Qu.:-2.0194   1st Qu.: 1.204
##  Median : 0.06247   Median :-1.5486   Median : 1.674
##  Mean   : 0.00000   Mean   :-1.6127   Mean   : 1.613
##  3rd Qu.: 0.37444   3rd Qu.:-1.2391   3rd Qu.: 1.988
##  Max.   : 1.77925   Max.   : 0.1614   Max.   : 3.397
```

Appendex:

---

title: "Data621 - Assignment1"

author: "Ritesh Lohiya"

date: "June 16, 2018"

output: html_document

#HW #1 Assignment - Moneyball Model

Overview In this homework assignment, you will explore, analyze and model a data set containing approximately 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.

Your objective is to build a multiple linear regression model on the training data to predict the number of wins for the team. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:

```{r}

#install.packages('caret')

#install.packages('e1071', dependencies=TRUE)

library(knitr)

library(stringr)

library(tidyr)

library(dplyr)

library(ggplot2)

library(psych)

library(reshape)
```

```
library(corrgram)

library(mice)

library(caret)

library(e1071)
```

#DATA EXPLORATION:

Load the data and understand the data by using some stats and plots.

```{r}
mtd <- read.csv("https://raw.githubusercontent.com/Riteshlohiya/Data621-Assignment-1/master/moneyball-training-data.csv")

count(mtd)

names(mtd)

summary(mtd)
```

The dataset consists of 17 elements, with 2276 total cases. There are multiple variables with missing (NA) values and TEAM-BATTING_HBP has the highest NAs.

Checking for outliers:

```{r}
ggplot(stack(mtd), aes(x = ind, y = values)) +

  geom_boxplot() +

  coord_cartesian(ylim = c(0, 1000)) +

  theme(legend.position="none") +

  theme(axis.text.x=element_text(angle=45, hjust=1)) +

  theme(panel.background = element_rect(fill = 'grey'))
```

Checking for skewness in the data

```{r}
mtd1 = melt(mtd)
ggplot(mtd1, aes(x= value)) +
  geom_density(fill='red') + facet_wrap(~variable, scales = 'free')
```

As seen there are several variables that are skewed and also there are outliers.

Finding correlations:

```{r}
mtd2 <- mtd[,-1 ]
names(mtd2)
cor(drop_na(mtd2))
```

```{r}
pairs.panels(mtd2[1:8])
pairs.panels(mtd2[9:16])
```

We can see there are some positively and some negatively correlated variables.

#DATA PREPARATION

Removing the variables:

```{r}
mtd_f <- mtd[,-1 ]
```

names(mtd_f)

```

The variable TEAM_BATTING_HBP is having mostly missing values so the variable will be removed completely.

```{r}
mtd_f <- mtd_f[,-10 ]
names(mtd_f )
```

TEAM_PITCHING_HR and TEAM_BATTING_HR are highly correlated, so we can remove one of them.

```{r}
mtd_f <- mtd_f[,-11 ]
names(mtd_f)
```

Imputing the NAs using Mice(pmm - predictive mean matching)

```{r}
imputed_mtd_Data <- mice(mtd_f, m=5, maxit = 5, method = 'pmm')
imputed_mtd_Data <- complete(imputed_mtd_Data)
summary(imputed_mtd_Data)
```

Centering and scaling was used to transform individual predictors in the dataset using the caret library.

```{r}
t = preProcess(imputed_mtd_Data,

c("BoxCox", "center", "scale"))

mtd_final = data.frame(

t = predict(t, imputed_mtd_Data))

summary(mtd_final)

```

```{r}

mtd_final1 = melt(mtd_final)

ggplot(mtd_final1, aes(x= value)) +

geom_density(fill='red') + facet_wrap(~variable, scales = 'free')

```

#BUILD MODELS:

Model1:

With all variables:

```{r}

model1 <- lm(t.TARGET_WINS ~., mtd_final)

summary(model1)

```

Model2:

With only the significant variables:

```{r}

model2 <- lm(t.TARGET_WINS ~ t.TEAM_BATTING_H + t.TEAM_BATTING_3B + t.TEAM_BATTING_HR + t.TEAM_BATTING_BB + t.TEAM_BATTING_SO +

t.TEAM_BASERUN_SB + t.TEAM_PITCHING_SO + t.TEAM_PITCHING_H +
t.TEAM_PITCHING_SO + t.TEAM_FIELDING_E + t.TEAM_FIELDING_DP, mtd_final)

summary(model2)

```

Model3:

Further reducing the variables(TEAM_PITCHING_SO and TEAM_BATTING_SO are having high correlation, TEAM_BATTING_H and TEAM_PITCHING_H are also having high correlation, TEAM_BATTING_SO and TEAM_PITCHING_SO are also having high correlation):

```{r}

model3 <- lm(t.TARGET_WINS ~ t.TEAM_BATTING_H  + t.TEAM_BATTING_3B  +
t.TEAM_BATTING_HR  + t.TEAM_BATTING_BB + t.TEAM_BATTING_SO +
t.TEAM_BASERUN_SB  + t.TEAM_FIELDING_E + t.TEAM_FIELDING_DP, mtd_final)

summary(model3)

```

#SELECT MODELS AND PREDICTION:

```{r}

summary(model1)

summary(model2)

summary(model3)

```

From the three models, I decided to use model3 for the predictions considering its more parsimonious model. There is no significant difference in R2, Adjusted R2 and RMSE even when i did the treatment for multi-collinearity.

#PREDICTION:
```

For the evaluation dataset also we will be doing all the preprocessing steps.

```{r}
med <- read.csv("https://raw.githubusercontent.com/Riteshlohiya/Data621-Assignment-1/master/moneyball-evaluation-data.csv")
```

Removing the variables:

```{r}
med_f <- med[,-1 ]
names(med_f)
```

```{r}
med_f <- med_f[,-10 ]
names(med_f )
```

```{r}
med_f <- med_f[,-11 ]
names(med_f)
```

Imputing the NAs using Mice(pmm - predictive mean matching)

```{r}
imputed_med_Data <- mice(med_f, m=5, maxit = 5, method = 'pmm')
imputed_med_Data <- complete(imputed_med_Data)
summary(imputed_med_Data)
```

Centering and scaling was used to transform individual predictors in the dataset using the caret library.

````
```{r}
t = preProcess(imputed_med_Data,
               c("BoxCox", "center", "scale"))
med_final = data.frame(
    t = predict(t, imputed_med_Data))


summary(med_final)
```
````

````
```{r}
eval_data <- predict(model3, newdata = med_final, interval="prediction")
eval_data
```
````

````
```{r}
summary(eval_data)
```
````