R Notebook

Contents

1. Data Exploration

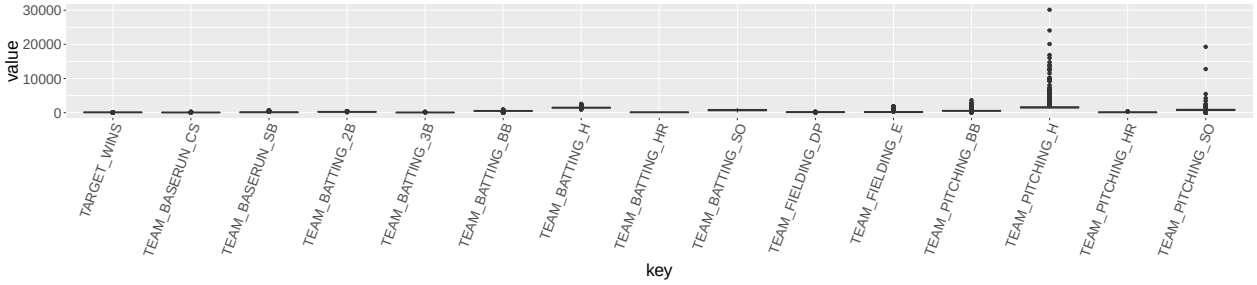
Describe the size and the variables in the moneyball training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas.

Summary Statistics

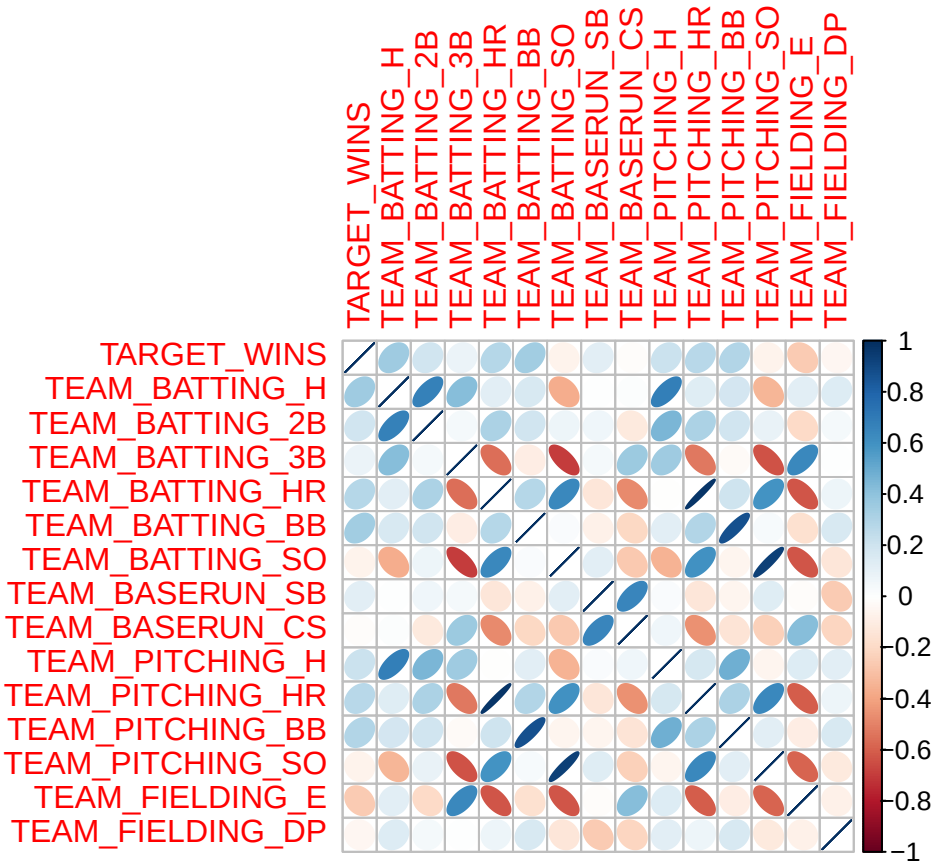
```
##      INDEX         TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B
##  Min.   :   1.0   Min.   :  0.00   Min.   : 891   Min.   : 69.0
##  1st Qu.: 630.8   1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0
##  Median :1270.5   Median : 82.00   Median :1454   Median :238.0
##  Mean   :1268.5   Mean   : 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
```

Box Plot



##Correlation Check




```
##                 TARGET_WINS TEAM_BATTING_H TEAM_BATTING_2B
## TARGET_WINS       1.00000000    0.359102657      0.19377063
## TEAM_BATTING_H    0.35910266    1.000000000      0.67597673
## TEAM_BATTING_2B   0.19377063    0.675976730      1.00000000
## TEAM_BATTING_3B   0.08273171    0.424243387      0.04159537
## TEAM_BATTING_HR   0.28346449    0.121533598      0.31393818
## TEAM_BATTING_BB   0.34829184    0.169265952      0.19769347
## TEAM_BATTING_SO  -0.06289963   -0.365563214      0.07894699
## TEAM_BASERUN_SB   0.12035129    0.006327368      0.06540808
## TEAM_BASERUN_CS  -0.01119995    0.014056697     -0.11760390
## TEAM_PITCHING_H   0.21610375    0.687898063      0.45457406
## TEAM_PITCHING_HR  0.27872264    0.136743031      0.31787719
## TEAM_PITCHING_BB  0.29493756    0.188208391      0.19232998
## TEAM_PITCHING_SO -0.06673654   -0.330034165      0.09116245
## TEAM_FIELDING_E  -0.25450741    0.121279902     -0.19416232
## TEAM_FIELDING_DP -0.04949709    0.148619532      0.04710052
##                 TEAM_BATTING_3B TEAM_BATTING_HR TEAM_BATTING_BB
## TARGET_WINS          0.08273171     0.283464494      0.34829184
## TEAM_BATTING_H       0.42424339     0.121533598      0.16926595
## TEAM_BATTING_2B      0.04159537     0.313938180      0.19769347
## TEAM_BATTING_3B      1.00000000    -0.554984363     -0.09728038
## TEAM_BATTING_HR     -0.55498436     1.000000000      0.28987751
## TEAM_BATTING_BB     -0.09728038     0.289877505      1.00000000
## TEAM_BATTING_SO     -0.69034724     0.640159102      0.02384741
## TEAM_BASERUN_SB      0.04553554    -0.138348758     -0.07810816
## TEAM_BASERUN_CS      0.36934182    -0.470715823     -0.20628472
```

```
## TEAM_PITCHING_H       0.35643177    -0.002382248      0.12825248
## TEAM_PITCHING_HR     -0.52736563     0.971652181      0.29599870
## TEAM_PITCHING_BB     -0.02376740     0.200030244      0.87467545
## TEAM_PITCHING_SO     -0.63929275     0.597391699      0.03414065
## TEAM_FIELDING_E       0.64109245    -0.621227175     -0.16115860
## TEAM_FIELDING_DP      0.00887003     0.070219016      0.16343997
##                  TEAM_BATTING_SO TEAM_BASERUN_SB TEAM_BASERUN_CS
## TARGET_WINS          -0.06289963     0.120351295     -0.01119995
## TEAM_BATTING_H       -0.36556321     0.006327368      0.01405670
## TEAM_BATTING_2B       0.07894699     0.065408077     -0.11760390
## TEAM_BATTING_3B      -0.69034724     0.045535545      0.36934182
## TEAM_BATTING_HR       0.64015910    -0.138348758     -0.47071582
## TEAM_BATTING_BB       0.02384741    -0.078108165     -0.20628472
## TEAM_BATTING_SO       1.00000000     0.126866866     -0.26423776
## TEAM_BASERUN_SB       0.12686687     1.000000000      0.65233884
## TEAM_BASERUN_CS      -0.26423776     0.652338836      1.00000000
## TEAM_PITCHING_H      -0.34335935     0.021181111      0.06909025
## TEAM_PITCHING_HR      0.60500635    -0.130403756     -0.45085011
## TEAM_PITCHING_BB     -0.05535063    -0.059437263     -0.14133608
## TEAM_PITCHING_SO      0.93221165     0.134325496     -0.23168177
## TEAM_FIELDING_E      -0.62668120    -0.011939775      0.42695850
## TEAM_FIELDING_DP     -0.13647576    -0.255985407     -0.21424801
##                  TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB
## TARGET_WINS          0.216103749       0.27872264       0.29493756
## TEAM_BATTING_H       0.687898063       0.13674303       0.18820839
## TEAM_BATTING_2B      0.454574060       0.31787719       0.19232998
## TEAM_BATTING_3B      0.356431774      -0.52736563      -0.02376740
## TEAM_BATTING_HR     -0.002382248       0.97165218       0.20003024
## TEAM_BATTING_BB      0.128252478       0.29599870       0.87467545
## TEAM_BATTING_SO     -0.343359353       0.60500635      -0.05535063
## TEAM_BASERUN_SB      0.021181111      -0.13040376      -0.05943726
## TEAM_BASERUN_CS      0.069090246      -0.45085011      -0.14133608
## TEAM_PITCHING_H      1.000000000       0.17271371       0.48745563
## TEAM_PITCHING_HR     0.172713710       1.00000000       0.31601752
## TEAM_PITCHING_BB     0.487455634       0.31601752       1.00000000
## TEAM_PITCHING_SO    -0.056621268       0.64681064       0.12892998
## TEAM_FIELDING_E      0.144892622      -0.60020742      -0.09659142
## TEAM_FIELDING_DP     0.122916134       0.07928388       0.16263604
##                  TEAM_PITCHING_SO TEAM_FIELDING_E TEAM_FIELDING_DP
## TARGET_WINS           -0.06673654     -0.25450741      -0.04949709
## TEAM_BATTING_H        -0.33003416      0.12127990       0.14861953
## TEAM_BATTING_2B        0.09116245     -0.19416232       0.04710052
## TEAM_BATTING_3B       -0.63929275      0.64109245       0.00887003
## TEAM_BATTING_HR        0.59739170     -0.62122718       0.07021902
## TEAM_BATTING_BB        0.03414065     -0.16115860       0.16343997
## TEAM_BATTING_SO        0.93221165     -0.62668120      -0.13647576
## TEAM_BASERUN_SB        0.13432550     -0.01193978      -0.25598541
## TEAM_BASERUN_CS       -0.23168177      0.42695850      -0.21424801
## TEAM_PITCHING_H       -0.05662127      0.14489262       0.12291613
## TEAM_PITCHING_HR       0.64681064     -0.60020742       0.07928388
## TEAM_PITCHING_BB       0.12892998     -0.09659142       0.16263604
## TEAM_PITCHING_SO       1.00000000     -0.58691832      -0.11885440
## TEAM_FIELDING_E       -0.58691832      1.00000000      -0.07870923
## TEAM_FIELDING_DP      -0.11885440     -0.07870923       1.00000000
```

2. Data Preparation

Dealing with missing values

In order to deal with missing values, I will drop hit by pitch(TEAM_BATTING_HBP), Since it has too many missing values. I will then run a function that replaces all missing values with the meadian of the column it is in

```
##      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.: 556.8
##  Median : 47.00   Median :102.00   Median :512.0   Median : 735.6
##  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.: 925.0
##  Max.   :223.00   Max.   :264.00   Max.   :878.0   Max.   :1399.0
##  TEAM_BASERUN_SB TEAM_BASERUN_CS  TEAM_PITCHING_H TEAM_PITCHING_HR
##  Min.   :  0.0   Min.   :  0.00   Min.   : 1137   Min.   :  0.0
##  1st Qu.: 67.0   1st Qu.: 44.00   1st Qu.: 1419   1st Qu.: 50.0
##  Median :106.0   Median : 52.80   Median : 1518   Median :107.0
##  Mean   :124.8   Mean   : 52.80   Mean   : 1779   Mean   :105.7
##  3rd Qu.:151.0   3rd Qu.: 54.25   3rd Qu.: 1682   3rd Qu.:150.0
##  Max.   :697.0   Max.   :201.00   Max.   :30132   Max.   :343.0
##  TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E  TEAM_FIELDING_DP
##  Min.   :   0.0   Min.   :    0.0   Min.   :  65.0   Min.   : 52.0
##  1st Qu.: 476.0   1st Qu.:  626.0   1st Qu.: 127.0   1st Qu.:134.0
##  Median : 536.5   Median :  817.7   Median : 159.0   Median :146.4
##  Mean   : 553.0   Mean   :  817.7   Mean   : 246.5   Mean   :146.4
##  3rd Qu.: 611.0   3rd Qu.:  957.0   3rd Qu.: 249.2   3rd Qu.:161.2
##  Max.   :3645.0   Max.   :19278.0   Max.   :1898.0   Max.   :228.0
```

Massage the dataset:

The dataset is missing a variable for singles, so I will add a variable called, TEAM_BATTING_1B, this variable will be hits - HR's - 2B - 3B. Also to eal with the fact that TEAM_PITCHING_H, and TEAM_PITCHING_SO are very spread out, I will take the square root of theese values.

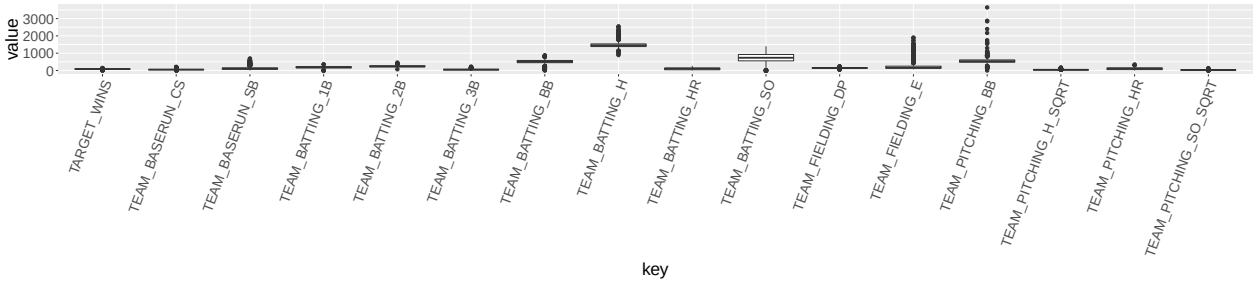
Exploration cleaning data

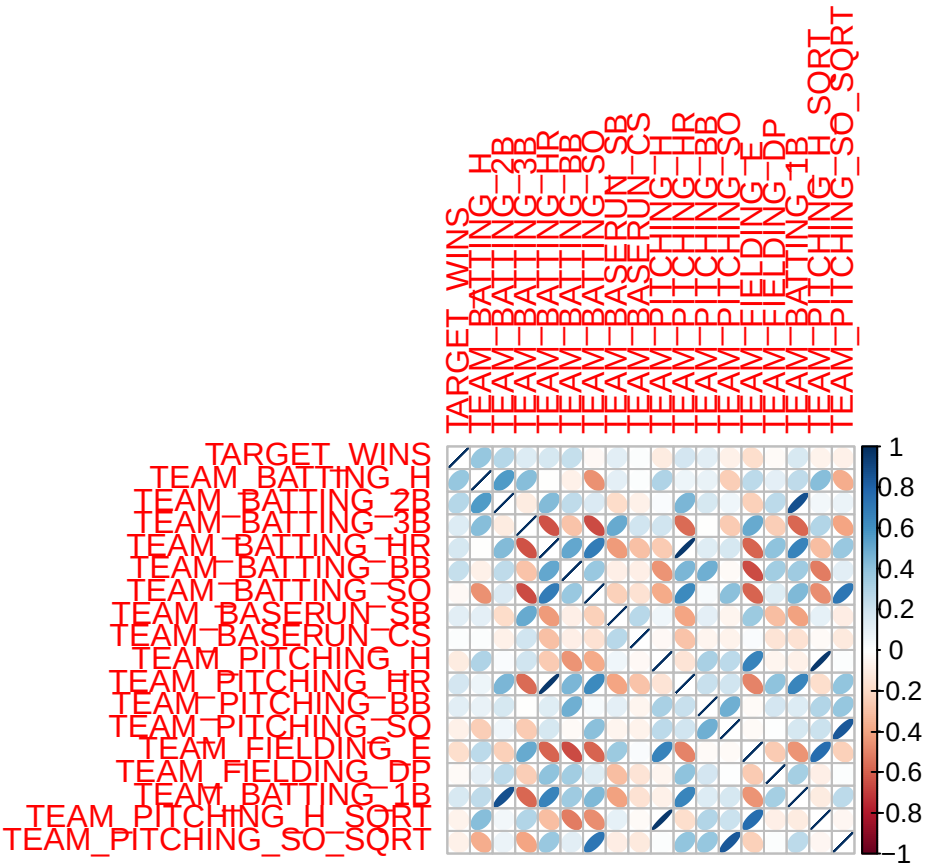
Summary Stats:

```
##      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.: 556.8
##  Median : 47.00   Median :102.00   Median :512.0   Median : 735.6
##  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.: 925.0
##  Max.   :223.00   Max.   :264.00   Max.   :878.0   Max.   :1399.0
##  TEAM_BASERUN_SB TEAM_BASERUN_CS  TEAM_PITCHING_H TEAM_PITCHING_HR
##  Min.   :  0.0   Min.   :  0.00   Min.   : 1137   Min.   :  0.0
##  1st Qu.: 67.0   1st Qu.: 44.00   1st Qu.: 1419   1st Qu.: 50.0
##  Median :106.0   Median : 52.80   Median : 1518   Median :107.0
##  Mean   :124.8   Mean   : 52.80   Mean   : 1779   Mean   :105.7
##  3rd Qu.:151.0   3rd Qu.: 54.25   3rd Qu.: 1682   3rd Qu.:150.0
##  Max.   :697.0   Max.   :201.00   Max.   :30132   Max.   :343.0
##  TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E  TEAM_FIELDING_DP
##  Min.   :   0.0   Min.   :    0.0   Min.   :  65.0   Min.   : 52.0
##  1st Qu.: 476.0   1st Qu.:  626.0   1st Qu.: 127.0   1st Qu.:134.0
##  Median : 536.5   Median :  817.7   Median : 159.0   Median :146.4
##  Mean   : 553.0   Mean   :  817.7   Mean   : 246.5   Mean   :146.4
##  3rd Qu.: 611.0   3rd Qu.:  957.0   3rd Qu.: 249.2   3rd Qu.:161.2
##  Max.   :3645.0   Max.   :19278.0   Max.   :1898.0   Max.   :228.0
##  TEAM_BATTING_1B TEAM_PITCHING_H_SQRT TEAM_PITCHING_SO_SQRT
##  Min.   :-27     Min.   : 33.72       Min.   :  0.00
##  1st Qu.:149     1st Qu.: 37.67       1st Qu.: 25.02
##  Median :186     Median : 38.96       Median : 28.60
##  Mean   :186     Mean   : 41.12       Mean   : 27.98
##  3rd Qu.:226     3rd Qu.: 41.02       3rd Qu.: 30.94
##  Max.   :357     Max.   :173.59       Max.   :138.85
```

BoxPlot





###Correlation

```
##                          TARGET_WINS TEAM_BATTING_H TEAM_BATTING_2B
## TARGET_WINS              1.00000000    0.388767521      0.28910365
## TEAM_BATTING_H           0.38876752    1.000000000      0.56284968
## TEAM_BATTING_2B          0.28910365    0.562849678      1.00000000
## TEAM_BATTING_3B          0.14260841    0.427696575     -0.10730582
## TEAM_BATTING_HR          0.17615320   -0.006544685      0.43539729
## TEAM_BATTING_BB          0.23255986   -0.072464013      0.25572610
## TEAM_BATTING_SO         -0.03067847   -0.450618504      0.15494194
## TEAM_BASERUN_SB          0.12297192    0.114034614     -0.18982749
## TEAM_BASERUN_CS          0.01556444    0.011640619     -0.07393756
## TEAM_PITCHING_H         -0.10993705    0.302693709      0.02369219
## TEAM_PITCHING_HR         0.18901373    0.072853119      0.45455082
## TEAM_PITCHING_BB         0.12417454    0.094193027      0.17805420
## TEAM_PITCHING_SO        -0.07578725   -0.245447770      0.06170843
## TEAM_FIELDING_E         -0.17648476    0.264902478     -0.23515099
## TEAM_FIELDING_DP        -0.02884126    0.115487966      0.26277009
## TEAM_BATTING_1B          0.16741117    0.252408467      0.87333024
## TEAM_PITCHING_H_SQRT    -0.06416499    0.421815721      0.05291298
## TEAM_PITCHING_SO_SQRT   -0.07066711   -0.367848237      0.08324953
##                          TEAM_BATTING_3B TEAM_BATTING_HR TEAM_BATTING_BB
## TARGET_WINS                  0.142608411     0.176153200      0.23255986
## TEAM_BATTING_H               0.427696575    -0.006544685     -0.07246401
## TEAM_BATTING_2B             -0.107305824     0.435397293      0.25572610
## TEAM_BATTING_3B              1.000000000    -0.635566946     -0.28723584
## TEAM_BATTING_HR             -0.635566946     1.000000000      0.51373481
## TEAM_BATTING_BB             -0.287235841     0.513734810      1.00000000
```

```
## TEAM_BATTING_SO          -0.657375248     0.696558278      0.37315562
## TEAM_BASERUN_SB           0.501029711    -0.428348107     -0.08187366
## TEAM_BASERUN_CS           0.194662155    -0.290739192     -0.08462458
## TEAM_PITCHING_H           0.194879411    -0.250145481     -0.44977762
## TEAM_PITCHING_HR         -0.567836679     0.969371396      0.45955207
## TEAM_PITCHING_BB         -0.002224148     0.136927564      0.48936126
## TEAM_PITCHING_SO         -0.254024989     0.176956412     -0.02039633
## TEAM_FIELDING_E           0.509778447    -0.587339098     -0.65597081
## TEAM_FIELDING_DP         -0.245749000     0.406149199      0.34049087
## TEAM_BATTING_1B          -0.578029383     0.668755334      0.35062185
## TEAM_PITCHING_H_SQRT      0.293605232    -0.307866149     -0.51797920
## TEAM_PITCHING_SO_SQRT    -0.398811640     0.378130225      0.12772493
##                       TEAM_BATTING_SO TEAM_BASERUN_SB TEAM_BASERUN_CS
## TARGET_WINS               -0.03067847      0.12297192      0.01556444
## TEAM_BATTING_H            -0.45061850      0.11403461      0.01164062
## TEAM_BATTING_2B            0.15494194     -0.18982749     -0.07393756
## TEAM_BATTING_3B           -0.65737525      0.50102971      0.19466215
## TEAM_BATTING_HR            0.69655828     -0.42834811     -0.29073919
## TEAM_BATTING_BB            0.37315562     -0.08187366     -0.08462458
## TEAM_BATTING_SO            1.00000000     -0.23156655     -0.15661491
## TEAM_BASERUN_SB           -0.23156655      1.00000000      0.27872176
## TEAM_BASERUN_CS           -0.15661491      0.27872176      1.00000000
## TEAM_PITCHING_H           -0.37513495      0.06084269     -0.03690036
## TEAM_PITCHING_HR           0.63656192     -0.39800787     -0.28952252
## TEAM_PITCHING_BB           0.03651072      0.11859723     -0.05352832
## TEAM_PITCHING_SO           0.41623330     -0.05454760     -0.06862178
## TEAM_FIELDING_E           -0.58328727      0.36863178      0.02356448
## TEAM_FIELDING_DP           0.13100129     -0.30249971     -0.14025285
## TEAM_BATTING_1B            0.44925536     -0.40128516     -0.15606082
## TEAM_PITCHING_H_SQRT      -0.46952886      0.10944674     -0.02865742
## TEAM_PITCHING_SO_SQRT      0.72554922     -0.09652256     -0.11314495
##                       TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB
## TARGET_WINS               -0.10993705       0.18901373      0.124174536
## TEAM_BATTING_H             0.30269371       0.07285312      0.094193027
## TEAM_BATTING_2B            0.02369219       0.45455082      0.178054204
## TEAM_BATTING_3B            0.19487941      -0.56783668     -0.002224148
## TEAM_BATTING_HR           -0.25014548       0.96937140      0.136927564
## TEAM_BATTING_BB           -0.44977762       0.45955207      0.489361263
## TEAM_BATTING_SO           -0.37513495       0.63656192      0.036510721
## TEAM_BASERUN_SB            0.06084269      -0.39800787      0.118597230
## TEAM_BASERUN_CS           -0.03690036      -0.28952252     -0.053528320
## TEAM_PITCHING_H            1.00000000      -0.14161276      0.320676162
## TEAM_PITCHING_HR          -0.14161276       1.00000000      0.221937505
## TEAM_PITCHING_BB           0.32067616       0.22193750      1.000000000
## TEAM_PITCHING_SO           0.26685582       0.19643263      0.481971904
## TEAM_FIELDING_E            0.66775901      -0.49314447     -0.022837561
## TEAM_FIELDING_DP          -0.05828433       0.40149241      0.187772148
## TEAM_BATTING_1B           -0.07603717       0.65129072      0.147228606
## TEAM_PITCHING_H_SQRT       0.96785568      -0.17628549      0.290779233
## TEAM_PITCHING_SO_SQRT      0.01464440       0.39528544      0.382184010
##                       TEAM_PITCHING_SO TEAM_FIELDING_E TEAM_FIELDING_DP
## TARGET_WINS               -0.075787253     -0.17648476     -0.028841263
## TEAM_BATTING_H            -0.245447770      0.26490248      0.115487966
## TEAM_BATTING_2B            0.061708426     -0.23515099      0.262770086
```

```
## TEAM_BATTING_3B           -0.254024989      0.50977845     -0.245749000
## TEAM_BATTING_HR            0.176956412     -0.58733910      0.406149199
## TEAM_BATTING_BB           -0.020396332     -0.65597081      0.340490875
## TEAM_BATTING_SO            0.416233300     -0.58328727      0.131001287
## TEAM_BASERUN_SB           -0.054547601      0.36863178     -0.302499715
## TEAM_BASERUN_CS           -0.068621784      0.02356448     -0.140252846
## TEAM_PITCHING_H            0.266855819      0.66775901     -0.058284329
## TEAM_PITCHING_HR           0.196432629     -0.49314447      0.401492411
## TEAM_PITCHING_BB           0.481971904     -0.02283756      0.187772148
## TEAM_PITCHING_SO           1.000000000     -0.02323692      0.009324802
## TEAM_FIELDING_E           -0.023236920      1.00000000     -0.252971612
## TEAM_FIELDING_DP           0.009324802     -0.25297161      1.000000000
## TEAM_BATTING_1B            0.175108969     -0.44277116      0.336076436
## TEAM_PITCHING_H_SQRT       0.210733378      0.76370436     -0.083174583
## TEAM_PITCHING_SO_SQRT      0.849873463     -0.23382460      0.014445280
##                      TEAM_BATTING_1B TEAM_PITCHING_H_SQRT
## TARGET_WINS               0.16741117          -0.06416499
## TEAM_BATTING_H            0.25240847           0.42181572
## TEAM_BATTING_2B           0.87333024           0.05291298
## TEAM_BATTING_3B           -0.57802938          0.29360523
## TEAM_BATTING_HR           0.66875533          -0.30786615
## TEAM_BATTING_BB           0.35062185          -0.51797920
## TEAM_BATTING_SO           0.44925536          -0.46952886
## TEAM_BASERUN_SB           -0.40128516          0.10944674
## TEAM_BASERUN_CS           -0.15606082          -0.02865742
## TEAM_PITCHING_H           -0.07603717          0.96785568
## TEAM_PITCHING_HR          0.65129072          -0.17628549
## TEAM_PITCHING_BB          0.14722861           0.29077923
## TEAM_PITCHING_SO          0.17510897           0.21073338
## TEAM_FIELDING_E           -0.44277116          0.76370436
## TEAM_FIELDING_DP          0.33607644          -0.08317458
## TEAM_BATTING_1B           1.00000000          -0.10042553
## TEAM_PITCHING_H_SQRT      -0.10042553          1.00000000
## TEAM_PITCHING_SO_SQRT     0.26372825          -0.04800748
##                      TEAM_PITCHING_SO_SQRT
## TARGET_WINS                    -0.07066711
## TEAM_BATTING_H                 -0.36784824
## TEAM_BATTING_2B                 0.08324953
## TEAM_BATTING_3B                -0.39881164
## TEAM_BATTING_HR                 0.37813022
## TEAM_BATTING_BB                 0.12772493
## TEAM_BATTING_SO                 0.72554922
## TEAM_BASERUN_SB                -0.09652256
## TEAM_BASERUN_CS                -0.11314495
## TEAM_PITCHING_H                 0.01464440
## TEAM_PITCHING_HR                0.39528544
## TEAM_PITCHING_BB                0.38218401
## TEAM_PITCHING_SO                0.84987346
## TEAM_FIELDING_E                -0.23382460
## TEAM_FIELDING_DP                0.01444528
## TEAM_BATTING_1B                 0.26372825
## TEAM_PITCHING_H_SQRT           -0.04800748
## TEAM_PITCHING_SO_SQRT           1.00000000
```

There does not seem to be much co-linearity in this data.

3. Build Models

In this section I will create 3 models. Looking at the correlation, the most positively correlated to wins is total hits, the most negative correlation is home runs againts. I will create 3 models one with each variable sperately and one with both variables combined.

Model 1

Take Team batting total hits as this has the most positive correlation to target wins.

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.768  -8.757   0.856   9.762  46.016
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    18.562326   3.107523   5.973 2.69e-09 ***
## TEAM_BATTING_H  0.042353   0.002105  20.122  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.52 on 2274 degrees of freedom
## Multiple R-squared:  0.1511, Adjusted R-squared:  0.1508
## F-statistic: 404.9 on 1 and 2274 DF,  p-value: < 2.2e-16
```

This model says that every team gets 18 wins and when you add a hit that adds .04 wins

Model 2

For model two, I am going to look at how TEAM_PITCHING_HR effects target wins. This is a measure of how bad a teams pitching is. This variable also had the most negative correlation to target wins.

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_PITCHING_HR, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.657  -9.956   0.636  10.055  67.477
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      75.656920   0.646540 117.018   <2e-16 ***
## TEAM_PITCHING_HR  0.048572   0.005292   9.179   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 15.47 on 2274 degrees of freedom
## Multiple R-squared:  0.03573,    Adjusted R-squared:  0.0353
## F-statistic: 84.25 on 1 and 2274 DF,  p-value: < 2.2e-16
```

This model says that if each team starts with 75 wins, for each Homerun that is hit the team would win .04 more games. To me this does not make much sense.

Model 3

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_PITCHING_HR,
##     data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.575  -9.120   0.702   9.578  49.406
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      16.058261   3.074565   5.223 1.92e-07 ***
## TEAM_BATTING_H    0.041071   0.002078  19.761  < 2e-16 ***
## TEAM_PITCHING_HR  0.041514   0.004903   8.468  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.3 on 2273 degrees of freedom
## Multiple R-squared:  0.1771, Adjusted R-squared:  0.1764
## F-statistic: 244.6 on 2 and 2273 DF,  p-value: < 2.2e-16
```

In this regession for each hit or each homerun that is scored against a team it adds .04 wins from a baseline of 16 wins.

4. Select Models

Since model 3 has the higest R^2, I am going to select hat model. The main thing with this model is becasue R^2 is low the model may not be very acurate, however since it takes into account the 2 most influential offensive and defensive variables should be an ok model.

Make Predictions

To run the predictions, I read in the evaluation dataset run it through the model and get the output. The first 20 rows of my predictions are:

```
##    Predicted_Wins TEAM_BATTING_H TEAM_PITCHING_HR
## 1        69.15897           1209              83
## 2        69.85939           1221              88
## 3        77.21335           1395              93
## 4        85.86750           1539             159
## 5        75.98733           1445              14
## 6        75.66142           1431              20
```

```
## 7        76.45062          1430              40
## 8        74.56090          1385              39
## 9        68.80474          1259              25
## 10       76.00857          1397              62
## 11       76.86708          1427              53
## 12       84.68263          1496             173
## 13       82.51603          1420             196
## 14       83.70223          1460             185
## 15       79.86314          1411             141
## 16       81.30594          1434             153
## 17       74.80740          1297             132
## 18       82.42150          1446             168
## 19       69.21235          1276              18
## 20       91.60154          1715             123
```

5. Appendix (R Code)

Imports

Data Exploration

1. Summary Statistics

```
training <- read.csv( "moneyball-training-data.csv")

summary(training)
```

Box Plot

```
training_tidy <- gather(data =training,  key=key,value=value,-INDEX  )
training_tidy <-  subset(training_tidy, key != "TEAM_BATTING_HBP")
ggplot(data=training_tidy) + geom_boxplot(aes(x=key, y=value)) +
  theme(text = element_text(size=20),
        axis.text.x = element_text(angle=70, hjust=1))
```

Correlation Check

```
training_corr_plot <- training[ , !(names(training) %in% c("TEAM_BATTING_HBP", "INDEX" ))]
training_corr_plot <- training_corr_plot[complete.cases(training_corr_plot), ]
M <- cor(training_corr_plot)
corrplot(M, method = "ellipse")
```

```
M
```

2.Data Preparation

Dealing with missing values

In order to deal with missing values, I will drop hit by pitch(TEAM_BATTING_HBP), Since it has too many missing values. I will then run a function that replaces all missing values with the meadian of the column it is in

```
training <- training[ , !(names(training) %in% c("TEAM_BATTING_HBP" ))]


for(i in 1:ncol(training)){
  training[is.na(training[,i]), i] <- mean(training[,i], na.rm = TRUE)
}

summary(training)
```

Massage the dataset:

The dataset is missing a variable for singles, so I will add a variable called, TEAM_BATTING_1B, this variable will be hits - HR's - 2B - 3B. Also to eal with the fact that TEAM_PITCHING_H, and TEAM_PITCHING_SO are very spread out, I will take the square root of theese values.

```
training$TEAM_BATTING_1B <- training$TEAM_BATTING_H - (training$TEAM_BATTING_H+training$TEAM_BATTING_3B
```

```
training$TEAM_PITCHING_H_SQRT <- sqrt(training$TEAM_PITCHING_H)
training$TEAM_PITCHING_SO_SQRT <- sqrt(training$TEAM_PITCHING_SO)
```

Exploration cleaning data

Summary Stats:

```
summary(training)
```

BoxPlot

```
training_tidy <- gather(data =training,  key=key,value=value,-INDEX  )
training_tidy <-  subset(training_tidy, key != "TEAM_BATTING_HBP")
training_tidy <-  subset(training_tidy, key != "TEAM_PITCHING_H")
training_tidy <-  subset(training_tidy, key != "TEAM_PITCHING_SO")
ggplot(data=training_tidy) + geom_boxplot(aes(x=key, y=value)) +
  theme(text = element_text(size=20),
        axis.text.x = element_text(angle=70, hjust=1))
```

Correlation

```
training_corr_plot <- training[ , !(names(training) %in% c("TEAM_BATTING_HBP", "INDEX" ))]
training_corr_plot <- training_corr_plot[complete.cases(training_corr_plot), ]
```

```
M <- cor(training_corr_plot)
corrplot(M, method = "ellipse")
```

```
M
```

3. Build Models

Model 1

```
model1 <- lm(TARGET_WINS~TEAM_BATTING_H, data = training )
summary(model1)
```

Model 2

```
model2 <- lm(TARGET_WINS~TEAM_PITCHING_HR, data = training )
summary(model2)
```

Model 3

```
model3 <- lm(TARGET_WINS~TEAM_BATTING_H+TEAM_PITCHING_HR, data = training )
summary(model3)
```

4. Select Models

Make Predictions

```
testing <- read.csv("moneyball-evaluation-data.csv")
testing$Predicted_Wins <- predict (model3, testing[,c("TEAM_BATTING_H", "TEAM_PITCHING_HR" )] )

head(testing[,c("Predicted_Wins", "TEAM_BATTING_H", "TEAM_PITCHING_HR" )],20)
```