DATA 621 Howewok 1

9/24/2021

Prepared by Critical Thinking Group 3 - Dominika Markowska-Desvallons and Shana Green

Introduction

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.

Objective

The objective is to build a multiple linear regression model on the training data to predict the number of wins for the given team. We can only use the variables provided (or variables that we will derive from the variables provided)

Data Exploration

Dataset

The moneyball training set contains 17 columns - including the target variable "TARGET_WINS" - and 2276 rows, covering baseball team performance statistics from the years 1871 to 2006 inclusive. The data has been adjusted to match the performance of a typical 162 game season. The data-set was entirely numerical and contained no categorical variables. There was also focus on all the variables to see which if any have missing data.

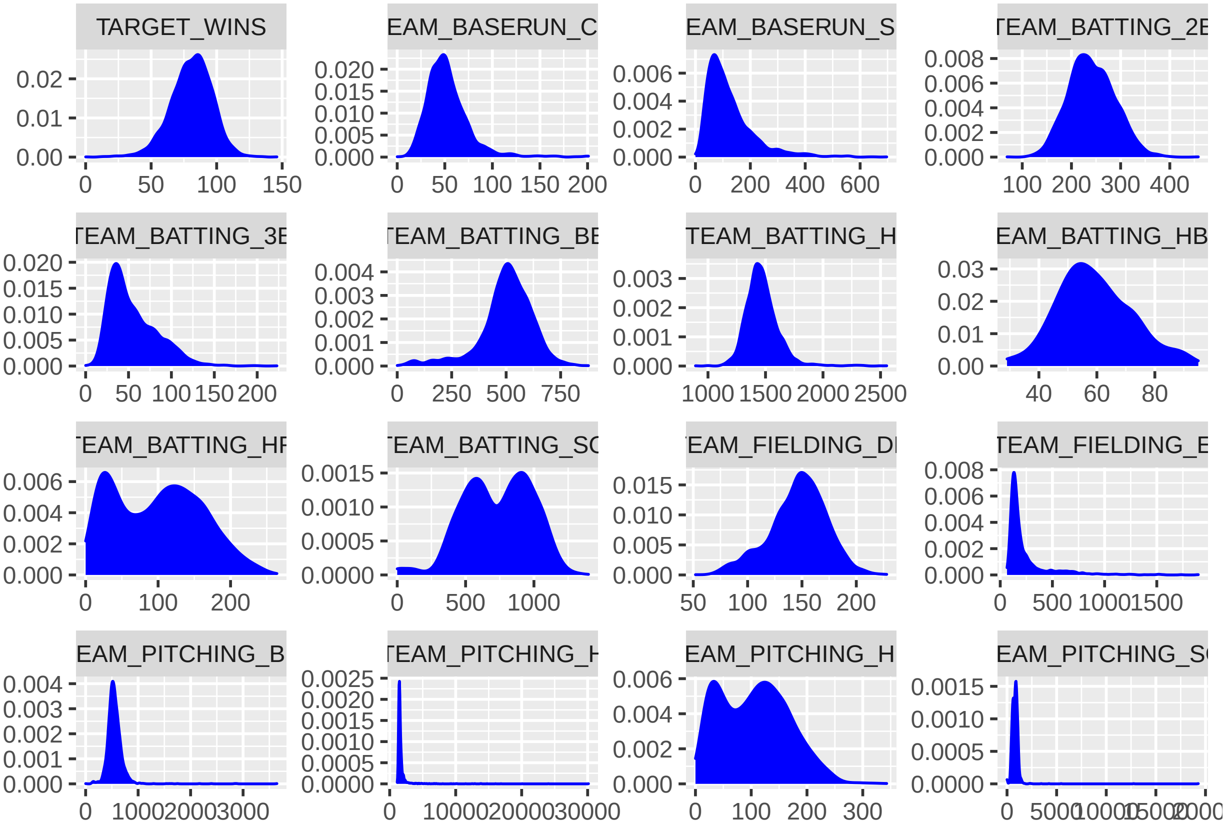
```
## Rows: 2,276
## Columns: 17
## $ INDEX            <int> 1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 15, 16, 17, 18, 1~
## $ TARGET_WINS      <int> 39, 70, 86, 70, 82, 75, 80, 85, 86, 76, 78, 68, 72, 7~
## $ TEAM_BATTING_H   <int> 1445, 1339, 1377, 1387, 1297, 1279, 1244, 1273, 1391,~
## $ TEAM_BATTING_2B  <int> 194, 219, 232, 209, 186, 200, 179, 171, 197, 213, 179~
## $ TEAM_BATTING_3B  <int> 39, 22, 35, 38, 27, 36, 54, 37, 40, 18, 27, 31, 41, 2~
## $ TEAM_BATTING_HR  <int> 13, 190, 137, 96, 102, 92, 122, 115, 114, 96, 82, 95,~
## $ TEAM_BATTING_BB  <int> 143, 685, 602, 451, 472, 443, 525, 456, 447, 441, 374~
## $ TEAM_BATTING_SO  <int> 842, 1075, 917, 922, 920, 973, 1062, 1027, 922, 827, ~
## $ TEAM_BASERUN_SB  <int> NA, 37, 46, 43, 49, 107, 80, 40, 69, 72, 60, 119, 221~
## $ TEAM_BASERUN_CS  <int> NA, 28, 27, 30, 39, 59, 54, 36, 27, 34, 39, 79, 109, ~
## $ TEAM_BATTING_HBP <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ TEAM_PITCHING_H  <int> 9364, 1347, 1377, 1396, 1297, 1279, 1244, 1281, 1391,~
## $ TEAM_PITCHING_HR <int> 84, 191, 137, 97, 102, 92, 122, 116, 114, 96, 86, 95,~
## $ TEAM_PITCHING_BB <int> 927, 689, 602, 454, 472, 443, 525, 459, 447, 441, 391~
## $ TEAM_PITCHING_SO <int> 5456, 1082, 917, 928, 920, 973, 1062, 1033, 922, 827,~
## $ TEAM_FIELDING_E  <int> 1011, 193, 175, 164, 138, 123, 136, 112, 127, 131, 11~
## $ TEAM_FIELDING_DP <int> NA, 155, 153, 156, 168, 149, 186, 136, 169, 159, 141,~
```

Getting min, max, median, mean, 1st quarter, 3rd quater.

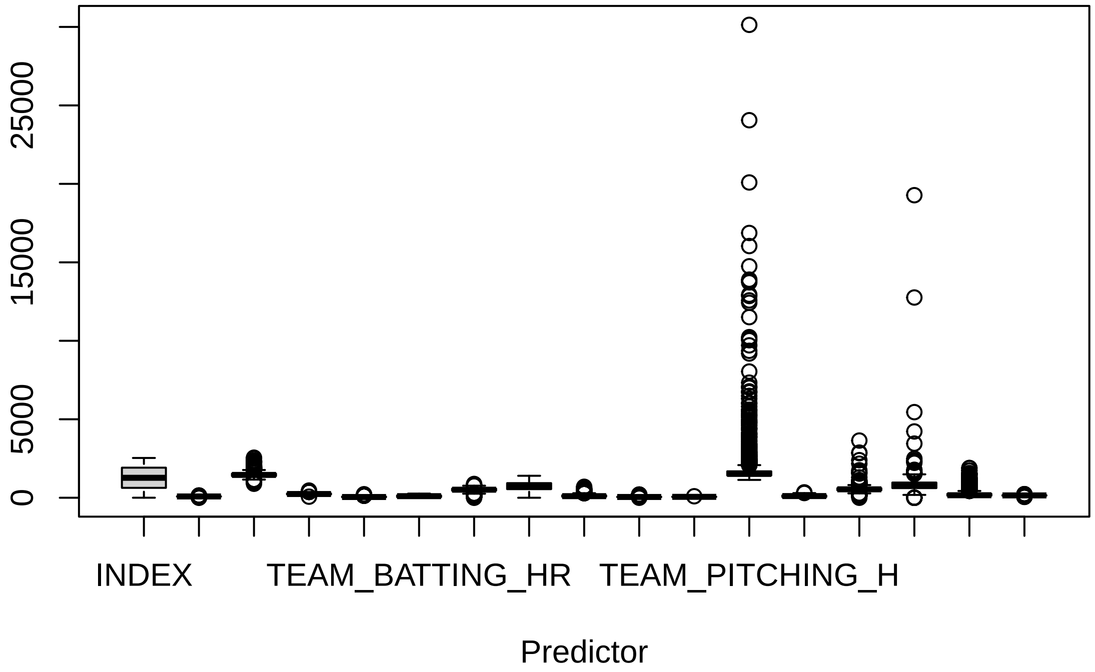
```
##      INDEX         TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B
##  Min.   :   1.0   Min.   :  0.00   Min.   : 891   Min.   : 69.0  
##  1st Qu.: 630.8   1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0  
##  Median :1270.5   Median : 82.00   Median :1454   Median :238.0  
##  Mean   :1268.5  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

Summarizing there are many variables that appeared with unusually extreme values such as TEAM_PITCHING_H and 30132.0. We need to look much closer at the data and analyze the extreme values and get more information regarding that. In histograms below, the data shows multiple graphs with right skews while only a few have left-skew.

```
## Warning: Removed 3478 rows containing non-finite values (stat_density).
```

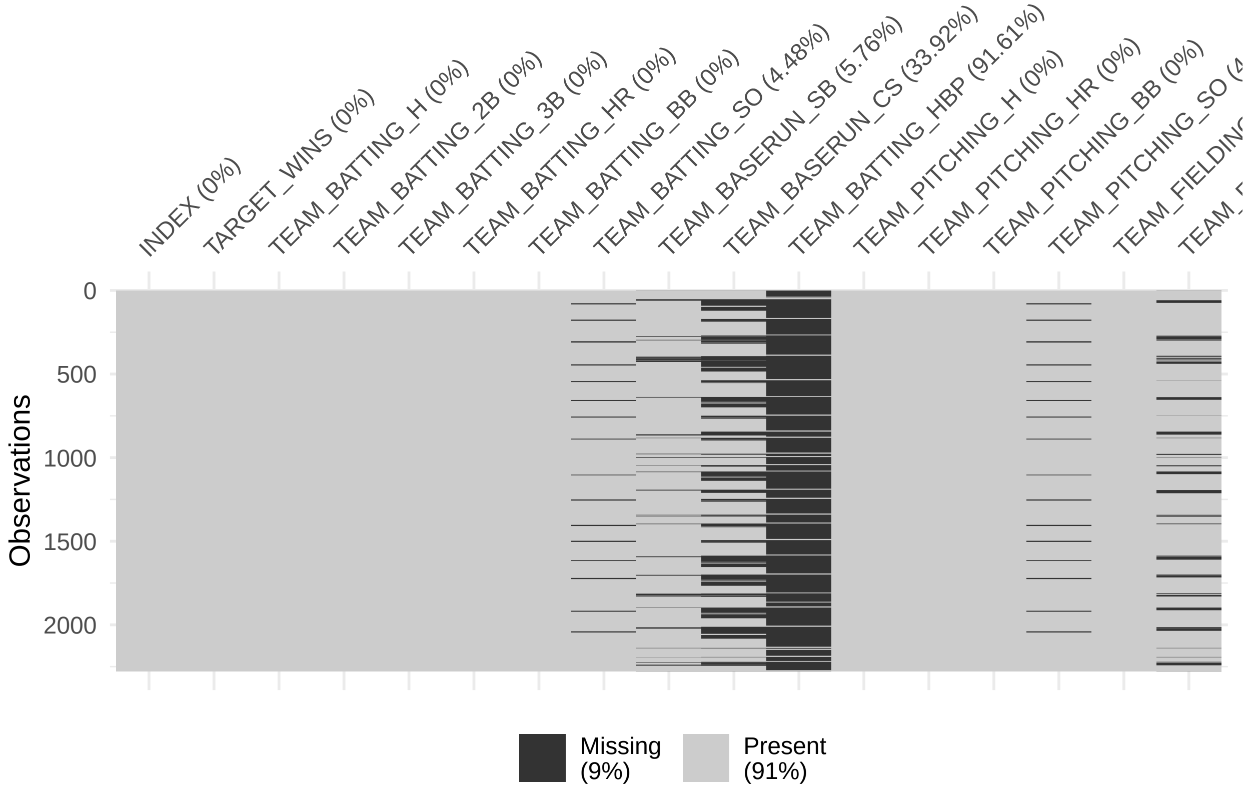


The boxplots show the spread of data within the dataset, and show various outliers. As seen in the graph below, TEAM_PITCHING_H seems to have the highest spread with the most outliers.



Check missing data

```
vis_miss(training)
```

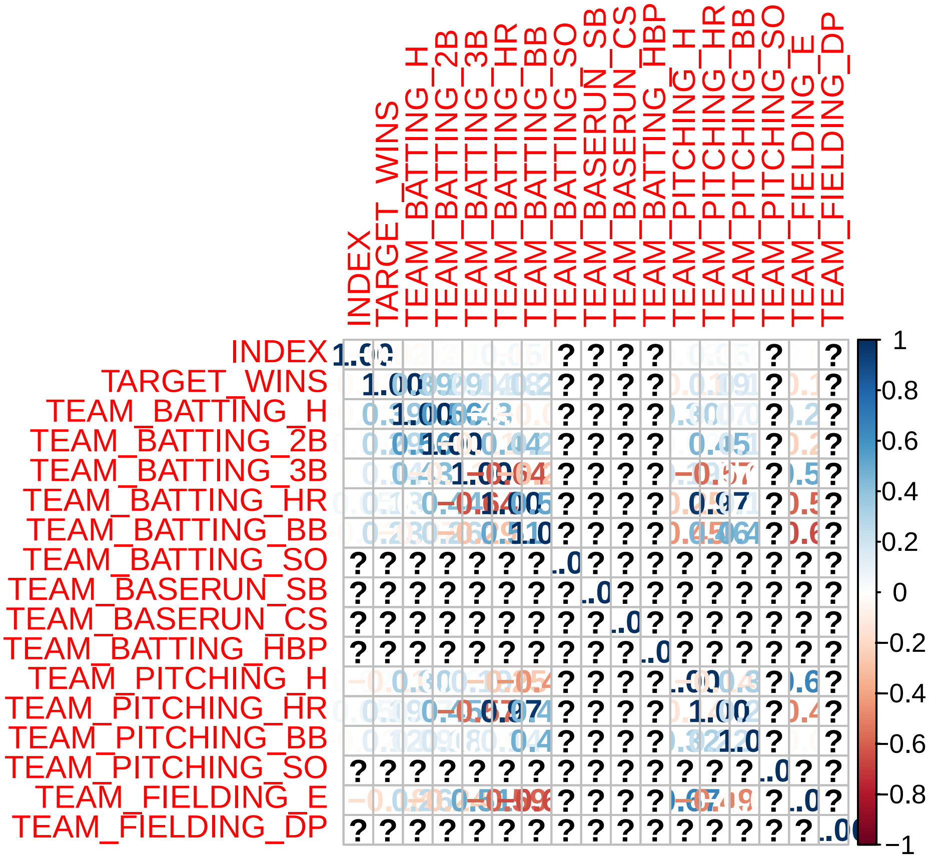


```
mb_cor <- cor(training)
round(mb_cor, 3)
```

```
##                   INDEX TARGET_WINS TEAM_BATTING_H TEAM_BATTING_2B
## INDEX             1.000      -0.021         -0.018           0.011
## TARGET_WINS      -0.021       1.000          0.389           0.289
## TEAM_BATTING_H   -0.018       0.389          1.000           0.563
## TEAM_BATTING_2B   0.011       0.289          0.563           1.000
## TEAM_BATTING_3B  -0.006       0.143          0.428          -0.107
## TEAM_BATTING_HR   0.051       0.176         -0.007           0.435
## TEAM_BATTING_BB  -0.027       0.233         -0.072           0.256
## TEAM_BATTING_SO      NA          NA             NA              NA
## TEAM_BASERUN_SB      NA          NA             NA              NA
## TEAM_BASERUN_CS      NA          NA             NA              NA
## TEAM_BATTING_HBP     NA          NA             NA              NA
## TEAM_PITCHING_H   0.017      -0.110          0.303           0.024
## TEAM_PITCHING_HR  0.051       0.189          0.073           0.455
## TEAM_PITCHING_BB -0.015       0.124          0.094           0.178
## TEAM_PITCHING_SO     NA          NA             NA              NA
## TEAM_FIELDING_E  -0.009      -0.176          0.265          -0.235
## TEAM_FIELDING_DP     NA          NA             NA              NA
##                  TEAM_BATTING_3B TEAM_BATTING_HR TEAM_BATTING_BB
## INDEX                     -0.006           0.051          -0.027
## TARGET_WINS                0.143           0.176           0.233
## TEAM_BATTING_H             0.428          -0.007          -0.072
## TEAM_BATTING_2B           -0.107           0.435           0.256
## TEAM_BATTING_3B            1.000          -0.636          -0.287
## TEAM_BATTING_HR           -0.636           1.000           0.514
## TEAM_BATTING_BB           -0.287           0.514           1.000
## TEAM_BATTING_SO               NA              NA              NA
## TEAM_BASERUN_SB               NA              NA              NA
## TEAM_BASERUN_CS               NA              NA              NA
## TEAM_BATTING_HBP              NA              NA              NA
## TEAM_PITCHING_H            0.195          -0.250          -0.450
## TEAM_PITCHING_HR          -0.568           0.969           0.460
## TEAM_PITCHING_BB          -0.002           0.137           0.489
## TEAM_PITCHING_SO              NA              NA              NA
## TEAM_FIELDING_E            0.510          -0.587          -0.656
## TEAM_FIELDING_DP              NA              NA              NA
##                  TEAM_BATTING_SO TEAM_BASERUN_SB TEAM_BASERUN_CS
## INDEX                         NA              NA              NA
## TARGET_WINS                   NA              NA              NA
## TEAM_BATTING_H                NA              NA              NA
## TEAM_BATTING_2B               NA              NA              NA
## TEAM_BATTING_3B               NA              NA              NA
## TEAM_BATTING_HR               NA              NA              NA
## TEAM_BATTING_BB               NA              NA              NA
## TEAM_BATTING_SO                1              NA              NA
## TEAM_BASERUN_SB               NA               1              NA
## TEAM_BASERUN_CS               NA              NA               1
## TEAM_BATTING_HBP              NA              NA              NA
## TEAM_PITCHING_H               NA              NA              NA
## TEAM_PITCHING_HR              NA              NA              NA
## TEAM_PITCHING_BB              NA              NA              NA
## TEAM_PITCHING_SO              NA              NA              NA
## TEAM_FIELDING_E               NA              NA              NA
## TEAM_FIELDING_DP              NA              NA              NA
##                  TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR
## INDEX                          NA           0.017            0.051
## TARGET_WINS                    NA          -0.110            0.189
## TEAM_BATTING_H                 NA           0.303            0.073
## TEAM_BATTING_2B                NA           0.024            0.455
## TEAM_BATTING_3B                NA           0.195           -0.568
## TEAM_BATTING_HR                NA          -0.250            0.969
## TEAM_BATTING_BB                NA          -0.450            0.460
## TEAM_BATTING_SO                NA              NA               NA
## TEAM_BASERUN_SB                NA              NA               NA
## TEAM_BASERUN_CS                NA              NA               NA
## TEAM_BATTING_HBP                1              NA               NA
## TEAM_PITCHING_H                NA           1.000           -0.142
## TEAM_PITCHING_HR               NA          -0.142            1.000
## TEAM_PITCHING_BB               NA           0.321            0.222
## TEAM_PITCHING_SO               NA              NA               NA
## TEAM_FIELDING_E                NA           0.668           -0.493
## TEAM_FIELDING_DP               NA              NA               NA
##                  TEAM_PITCHING_BB TEAM_PITCHING_SO TEAM_FIELDING_E
## INDEX                      -0.015               NA          -0.009
## TARGET_WINS                 0.124               NA          -0.176
## TEAM_BATTING_H              0.094               NA           0.265
## TEAM_BATTING_2B             0.178               NA          -0.235
## TEAM_BATTING_3B            -0.002               NA           0.510
## TEAM_BATTING_HR             0.137               NA          -0.587
## TEAM_BATTING_BB             0.489               NA          -0.656
## TEAM_BATTING_SO                NA               NA              NA
## TEAM_BASERUN_SB                NA               NA              NA
## TEAM_BASERUN_CS                NA               NA              NA
## TEAM_BATTING_HBP               NA               NA              NA
## TEAM_PITCHING_H             0.321               NA           0.668
## TEAM_PITCHING_HR            0.222               NA          -0.493
## TEAM_PITCHING_BB            1.000               NA          -0.023
## TEAM_PITCHING_SO               NA                1              NA
## TEAM_FIELDING_E            -0.023               NA           1.000
## TEAM_FIELDING_DP               NA               NA              NA
##                  TEAM_FIELDING_DP
## INDEX                          NA
## TARGET_WINS                    NA
## TEAM_BATTING_H                 NA
## TEAM_BATTING_2B                NA
## TEAM_BATTING_3B                NA
## TEAM_BATTING_HR                NA
## TEAM_BATTING_BB                NA
## TEAM_BATTING_SO                NA
## TEAM_BASERUN_SB                NA
## TEAM_BASERUN_CS                NA
## TEAM_BATTING_HBP               NA
## TEAM_PITCHING_H                NA
## TEAM_PITCHING_HR               NA
## TEAM_PITCHING_BB               NA
## TEAM_PITCHING_SO               NA
## TEAM_FIELDING_E                NA
## TEAM_FIELDING_DP                1
```

check correlation

```
M<-cor(training)
corrplot(M, method="number")
```



DATA PREPARATION