CUNY The City University of New York | School of Professional Studies

DATA 621 – Business Analytics and Data Mining

Homework #1 Assignment Requirements

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

Deliverables:

- A write-up submitted in PDF format. Your write-up should have four sections. Each one is described below. You may assume you are addressing me as a fellow data scientist, so do not need to shy away from technical details.

- Assigned predictions (the number of wins for the team) for the evaluation data set.

- Include your R statistical programming code in an Appendix.

Write Up:

1. DATA EXPLORATION (25 Points)

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.

a. Mean / Standard Deviation / Median
b. Bar Chart or Box Plot of the data
c. Is the data correlated to the target variable (or to other variables?)
d. Are any of the variables missing and need to be imputed "fixed"?

2. DATA PREPARATION (25 Points)

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations.

a. Fix missing values (maybe with a Mean or Median value)
b. Create flags to suggest if a variable was missing
c. Transform data by putting it into buckets
d. Mathematical transforms such as log or square root (or use Box-Cox)
e. Combine variables (such as ratios or adding or multiplying) to create new variables

3. BUILD MODELS (25 Points)

Using the training data set, build at least three different multiple linear regression models, using different variables (or the same variables with different transformations). Since we have not yet covered automated variable selection methods, you should select the variables manually (unless you previously learned Forward or Stepwise selection, etc.). Since you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done.

Discuss the coefficients in the models, do they make sense? For example, if a team hits a lot of Home Runs, it would be reasonably expected that such a team would win more games. However, if the coefficient is negative (suggesting that the team would lose more games), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

4. SELECT MODELS (25 Points)

Decide on the criteria for selecting the best multiple linear regression model. Will you select a model with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your model.

For the multiple linear regression model, will you use a metric such as Adjusted R^2, RMSE, etc.? Be sure to explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output. Using the training data set, evaluate the multiple linear regression model based on (a) mean squared error, (b) R^2, (c) F-statistic, and (d) residual plots. Make predictions using the evaluation data set.

DATA 621 – Business Analytics and Data Mining

Homework 1

Ramnivas Singh

2022-03-06

1.0 Summary

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

Deliverables:

1. A write-up submitted in PDF format. Your write-up should have four sections. Each one is described below. You may assume you are addressing me as a fellow data scientist, so do not need to shy away from technical details.
2. Assigned predictions (the number of wins for the team) for the evaluation data set.
3. Include your R statistical programming code in an Appendix.

Write-up sections :

1. Data Exploration
2. Data Preparation
3. Build Models
4. Select Models

2.0 Data Exploration

First of all, load the data and analyze to get some insights like summary, how the data got distributed and correlation between variables and understand the data by using stats, plots and summary. The objective of this is analysis is to develop a better understanding of the data to include its shape, central tendencies, completeness (missing data) and its correlation to our response variable Target_Wins.

```
##   TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_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
##
##  TEAM_BATTING_HR  TEAM_BATTING_BB TEAM_BATTING_SO  TEAM_BASERUN_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
##  TEAM_BASERUN_CS TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_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
##  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.:  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
```

```
##      n
## 1 2276
```

```
##  [1] "TARGET_WINS"      "TEAM_BATTING_H"   "TEAM_BATTING_2B"  "TEAM_BATTING_3B"
##  [5] "TEAM_BATTING_HR"  "TEAM_BATTING_BB"  "TEAM_BATTING_SO"  "TEAM_BASERUN_SB"
##  [9] "TEAM_BASERUN_CS"  "TEAM_BATTING_HBP" "TEAM_PITCHING_H"  "TEAM_PITCHING_HR"
## [13] "TEAM_PITCHING_BB" "TEAM_PITCHING_SO" "TEAM_FIELDING_E"  "TEAM_FIELDING_DP"
```

View rows and columns, variable types

Glimpse of the data shows that all variables are numeric, no ctegorical variable is present here. We do lots of NA for few predcitors in the data set. In our further analysis we will try to identify :

- Structure of the each predictors
- How Many NA and Zero , is it significant to remove them or replace them with some predicted value.
- Statistical summary of the data

```
## Rows: 2,276
## Columns: 16
## $ 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,~
```

Sample 6 rows with sample 7 columns

```
##   TARGET_WINS TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B TEAM_BATTING_HR
## 1          39           1445             194              39              13
## 2          70           1339             219              22             190
## 3          86           1377             232              35             137
## 4          70           1387             209              38              96
## 5          82           1297             186              27             102
## 6          75           1279             200              36              92
##   TEAM_BATTING_BB TEAM_BATTING_SO TEAM_BASERUN_SB TEAM_BASERUN_CS
## 1             143             842              NA              NA
## 2             685            1075              37              28
## 3             602             917              46              27
## 4             451             922              43              30
## 5             472             920              49              39
## 6             443             973             107              59
##   TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB
## 1               NA           9364              84             927
## 2               NA           1347             191             689
## 3               NA           1377             137             602
## 4               NA           1396              97             454
## 5               NA           1297             102             472
## 6               NA           1279              92             443
##   TEAM_PITCHING_SO TEAM_FIELDING_E TEAM_FIELDING_DP
## 1             5456            1011              NA
## 2             1082             193             155
## 3              917             175             153
## 4              928             164             156
## 5              920             138             168
## 6              973             123             149
```

Show entire dataset of training data

Show 10 **entries** **Search:**

	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB	TEAM_BATTING_SO	TEAM_BASEF
1	39	1445	194	39	13	143	842	
2	70	1339	219	22	190	685	1075	
3	86	1377	232	35	137	602	917	
4	70	1387	209	38	96	451	922	
5	82	1297	186	27	102	472	920	
6	75	1279	200	36	92	443	973	
7	80	1244	179	54	122	525	1062	
8	85	1273	171	37	115	456	1027	
9	86	1391	197	40	114	447	922	
10	76	1271	213	18	96	441	827	

Showing 1 to 10 of 2,276 entries

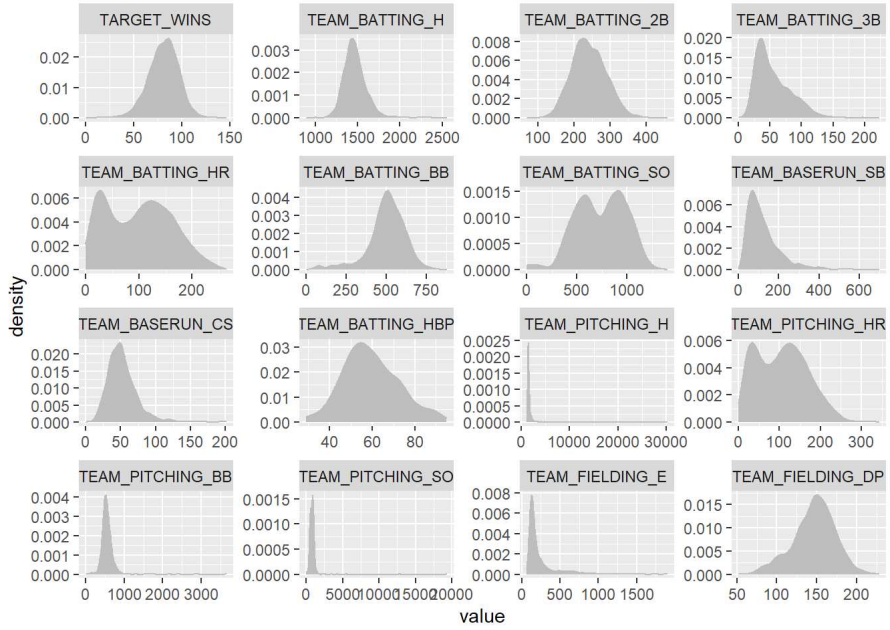
Previous 1 2 3 4 5 … 228 Next

Here are some key points from data exploration:

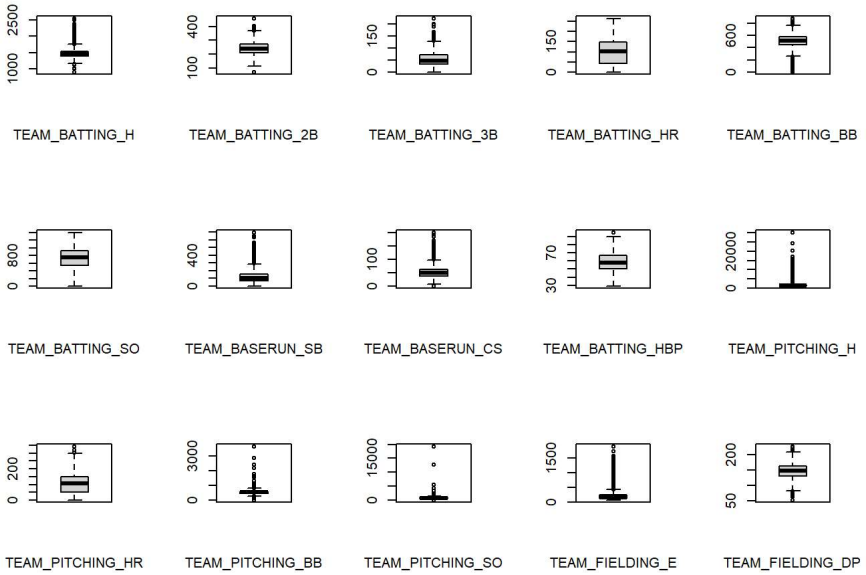
- here are multiple variables with missing (NA) values and TEAM-BATTING_HBP has the highest NAs.
- The data is generally complete, however, six variables have missing data.
- The lowest complete rate is for the variable Hit By Pitch, with a rate of only 8%.
- The data set includes 2276 rows, 16 columns with all variables are numeric
- The response variable appears to be normally or near-normally distributed.

Additional Data Exploration

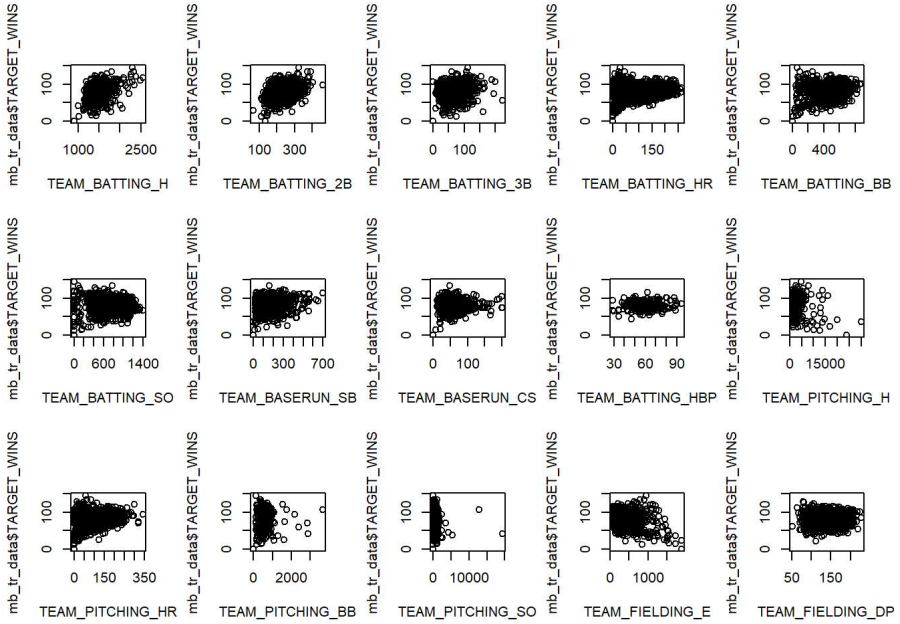
Skewness in the data :



The majority of the explanatory variables appear to be normal or near-normal. There are however, several variable that have bi-modal distributions (Batting_HR, SO, Pitching_HR, Batting_SO) and others that are right-skewed (Fielding, Pitching_BB, Pitching_H)



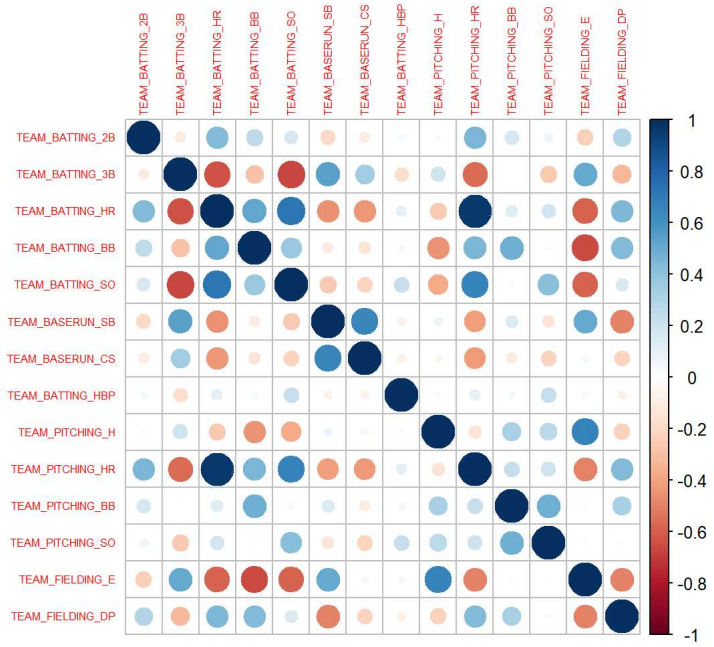
Response Variable & Correlations

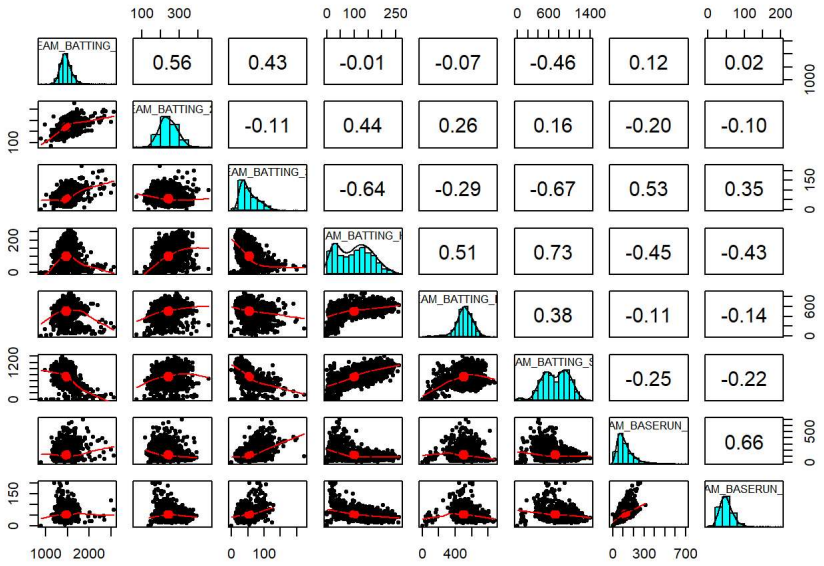


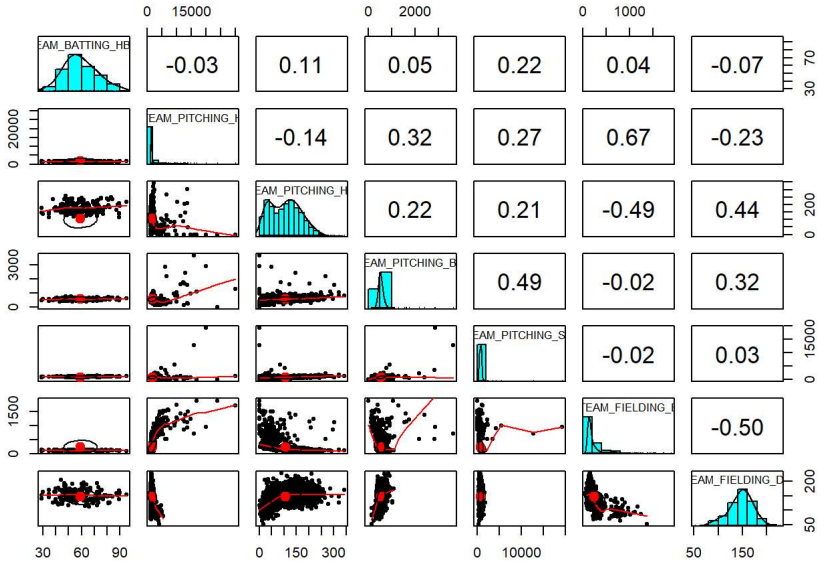
```
##  [1] "TEAM_BATTING_H"   "TEAM_BATTING_2B"  "TEAM_BATTING_3B"  "TEAM_BATTING_HR"
##  [5] "TEAM_BATTING_BB"  "TEAM_BATTING_SO"  "TEAM_BASERUN_SB"  "TEAM_BASERUN_CS"
##  [9] "TEAM_BATTING_HBP" "TEAM_PITCHING_H"  "TEAM_PITCHING_HR" "TEAM_PITCHING_BB"
## [13] "TEAM_PITCHING_SO" "TEAM_FIELDING_E"  "TEAM_FIELDING_DP"
```

```
##                  TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B TEAM_BATTING_HR
## TEAM_BATTING_H       1.00000000      0.56177286      0.21391883      0.39627593
## TEAM_BATTING_2B      0.56177286      1.00000000      0.04203441      0.25099045
## TEAM_BATTING_3B      0.21391883      0.04203441      1.00000000     -0.21879927
## TEAM_BATTING_HR      0.39627593      0.25099045     -0.21879927      1.00000000
## TEAM_BATTING_BB      0.19735234      0.19749256     -0.20584392      0.45638161
## TEAM_BATTING_SO     -0.34174328     -0.06415123     -0.19291841      0.21045444
## TEAM_BASERUN_SB      0.07167495     -0.18768279      0.16946086     -0.19021893
## TEAM_BASERUN_CS     -0.09377545     -0.20413884      0.23213978     -0.27579838
## TEAM_BATTING_HBP    -0.02911218      0.04608475     -0.17424715      0.10618116
## TEAM_PITCHING_H      0.99919269      0.56045355      0.21250322      0.39549390
## TEAM_PITCHING_HR     0.39495630      0.24999875     -0.21973263      0.99993259
## TEAM_PITCHING_BB     0.19529071      0.19592157     -0.20675383      0.45542468
## TEAM_PITCHING_SO    -0.34445001     -0.06616615     -0.19386654      0.20829574
## TEAM_FIELDING_E     -0.25381638     -0.19427027     -0.06513145      0.01567397
## TEAM_FIELDING_DP     0.01776946     -0.02488808      0.13314758     -0.06182222
##                  TEAM_BATTING_BB TEAM_BATTING_SO TEAM_BASERUN_SB
## TEAM_BATTING_H        0.19735234     -0.34174328      0.07167495
## TEAM_BATTING_2B       0.19749256     -0.06415123     -0.18768279
## TEAM_BATTING_3B      -0.20584392     -0.19291841      0.16946086
## TEAM_BATTING_HR       0.45638161      0.21045444     -0.19021893
## TEAM_BATTING_BB       1.00000000      0.21833871     -0.08806123
## TEAM_BATTING_SO       0.21833871      1.00000000     -0.07475974
## TEAM_BASERUN_SB      -0.08806123     -0.07475974      1.00000000
## TEAM_BASERUN_CS      -0.20878051     -0.05613035      0.62473781
## TEAM_BATTING_HBP      0.04746007      0.22094219     -0.06400498
## TEAM_PITCHING_H       0.19848687     -0.34145321      0.07395373
## TEAM_PITCHING_HR      0.45659283      0.21111617     -0.18948057
## TEAM_PITCHING_BB      0.99988140      0.21895783     -0.08741902
## TEAM_PITCHING_SO      0.21793253      0.99976835     -0.07351325
## TEAM_FIELDING_E      -0.07847126      0.30814540      0.04292341
## TEAM_FIELDING_DP     -0.07929078     -0.12319072     -0.13023054
##                  TEAM_BASERUN_CS TEAM_BATTING_HBP TEAM_PITCHING_H
## TEAM_BATTING_H      -0.093775445      -0.02911218      0.99919269
## TEAM_BATTING_2B     -0.204138837       0.04608475      0.56045355
## TEAM_BATTING_3B      0.232139777      -0.17424715      0.21250322
## TEAM_BATTING_HR     -0.275798375       0.10618116      0.39549390
## TEAM_BATTING_BB     -0.208780510       0.04746007      0.19848687
## TEAM_BATTING_SO     -0.056130355       0.22094219     -0.34145321
## TEAM_BASERUN_SB      0.624737808      -0.06400498      0.07395373
## TEAM_BASERUN_CS      1.000000000      -0.07051390     -0.09297789
## TEAM_BATTING_HBP    -0.070513896       1.00000000     -0.02769699
## TEAM_PITCHING_H     -0.092977893      -0.02769699      1.00000000
## TEAM_PITCHING_HR    -0.275471495       0.10675878      0.39463199
## TEAM_PITCHING_BB    -0.208470154       0.04785137      0.19703302
## TEAM_PITCHING_SO    -0.055308336       0.22157375     -0.34330646
## TEAM_FIELDING_E      0.207701189       0.04178971     -0.25073028
## TEAM_FIELDING_DP    -0.006764233      -0.07120824      0.01416807
##                  TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO
## TEAM_BATTING_H         0.39495630       0.19529071      -0.34445001
## TEAM_BATTING_2B        0.24999875       0.19592157      -0.06616615
## TEAM_BATTING_3B       -0.21973263      -0.20675383      -0.19386654
## TEAM_BATTING_HR        0.99993259       0.45542468       0.20829574
## TEAM_BATTING_BB        0.45659283       0.99988140       0.21793253
## TEAM_BATTING_SO        0.21111617       0.21895783       0.99976835
## TEAM_BASERUN_SB       -0.18948057      -0.08741902      -0.07351325
## TEAM_BASERUN_CS       -0.27547150      -0.20847015      -0.05530834
## TEAM_BATTING_HBP       0.10675878       0.04785137       0.22157375
## TEAM_PITCHING_H        0.39463199       0.19703302      -0.34330646
```

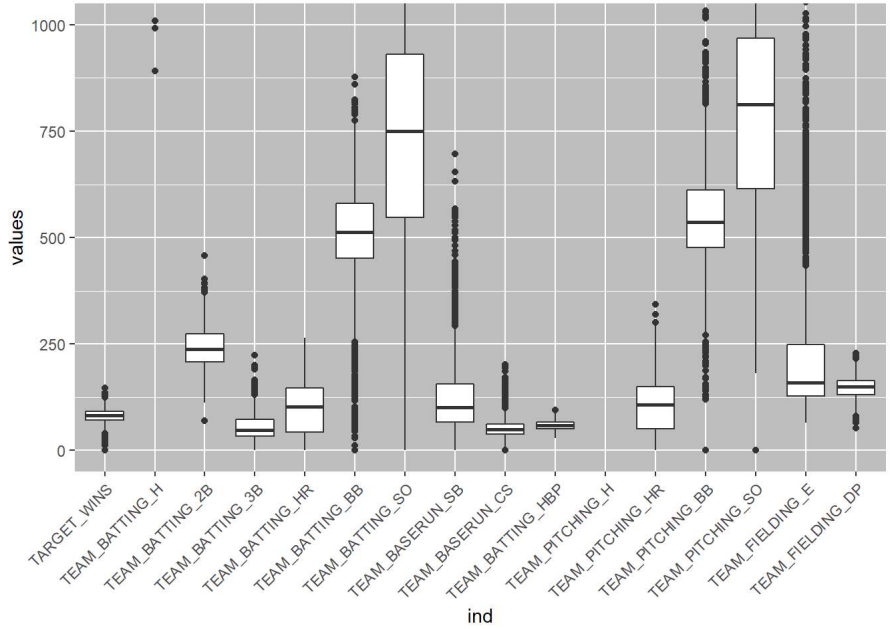
```
## TEAM_PITCHING_HR       1.00000000       0.45580983       0.20920115
## TEAM_PITCHING_BB       0.45580983       1.00000000       0.21887700
## TEAM_PITCHING_SO       0.20920115       0.21887700       1.00000000
## TEAM_FIELDING_E        0.01689330      -0.07692315       0.31008407
## TEAM_FIELDING_DP      -0.06292475      -0.08040645      -0.12492321
##                  TEAM_FIELDING_E TEAM_FIELDING_DP
## TEAM_BATTING_H       -0.25381638      0.017769456
## TEAM_BATTING_2B      -0.19427027     -0.024888081
## TEAM_BATTING_3B      -0.06513145      0.133147578
## TEAM_BATTING_HR       0.01567397     -0.061822219
## TEAM_BATTING_BB      -0.07847126     -0.079290775
## TEAM_BATTING_SO       0.30814540     -0.123190715
## TEAM_BASERUN_SB       0.04292341     -0.130230537
## TEAM_BASERUN_CS       0.20770119     -0.006764233
## TEAM_BATTING_HBP      0.04178971     -0.071208241
## TEAM_PITCHING_H      -0.25073028      0.014168073
## TEAM_PITCHING_HR      0.01689330     -0.062924751
## TEAM_PITCHING_BB     -0.07692315     -0.080406452
## TEAM_PITCHING_SO      0.31008407     -0.124923213
## TEAM_FIELDING_E       1.00000000      0.040205814
## TEAM_FIELDING_DP      0.04020581      1.000000000
```




0.56
0.43
-0.01
-0.07
-0.46
0.12
0.02
-0.11
0.44
0.26
0.16
-0.20
-0.10
-0.64
-0.29
-0.67
0.53
0.35
0.51
0.73
-0.45
-0.43
0.38
-0.11
-0.14
-0.25
-0.22
0.66


-0.03
0.11
0.05
0.22
0.04
-0.07
-0.14
0.32
0.27
0.67
-0.23
0.22
0.21
-0.49
0.44
0.49
-0.02
0.32
-0.02
0.03
-0.50

Outliers



Missing, NA and Zero

We are trying to see how many `NA` is present in the dataset.

variable	n	percent
TEAM_BATTING_HBP	2085	92%
TEAM_BASERUN_CS	772	34%
TEAM_FIELDING_DP	286	13%
TEAM_BASERUN_SB	131	5.8%
TEAM_BATTING_SO	102	4.5%
TEAM_PITCHING_SO	102	4.5%

variable	n	percent
TEAM_BATTING_SO	20	0.9%
TEAM_PITCHING_SO	20	0.9%
TEAM_BATTING_HR	15	0.7%
TEAM_PITCHING_HR	15	0.7%
TEAM_BASERUN_SB	2	0.1%
TEAM_BATTING_3B	2	0.1%

variable	n	percent
TARGET_WINS	1	0%
TEAM_BASERUN_CS	1	0%
TEAM_BATTING_BB	1	0%
TEAM_PITCHING_BB	1	0%

As can be inferred from above, there are very few zero values exists.

3.0 Data Preparation

Data preparation will include addressing missing data, outliers and feature engineering or creating new variables.

Outliers - The box plot for Wins should some very low values (even zero). According to major league baseball, the lowest number of wins recorded by a team was 20 by the Cleveland Spiders in 1899. Therefore, I will remove all rows from the data set with wins less than 20. The highest number of wins was 116, earned by the Seattle Mariners in 2001. I will also adjust the data set accordingly.

Missing Data - EDA identified variables with missing data. Given strategy I must address the missing data for two variables: Hit By Pitch and Caught Stealing. I will utilize historical major league baseball averages of these two variable as my replacement data.

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

```
##  [1] "TEAM_BATTING_H"   "TEAM_BATTING_2B"  "TEAM_BATTING_3B"  "TEAM_BATTING_HR"
##  [5] "TEAM_BATTING_BB"  "TEAM_BATTING_SO"  "TEAM_BASERUN_SB"  "TEAM_BASERUN_CS"
##  [9] "TEAM_BATTING_HBP" "TEAM_PITCHING_H"  "TEAM_PITCHING_HR" "TEAM_PITCHING_BB"
## [13] "TEAM_PITCHING_SO" "TEAM_FIELDING_E"  "TEAM_FIELDING_DP"
```

```
##  [1] "TEAM_BATTING_H"   "TEAM_BATTING_2B"  "TEAM_BATTING_3B"  "TEAM_BATTING_HR"
##  [5] "TEAM_BATTING_BB"  "TEAM_BATTING_SO"  "TEAM_BASERUN_SB"  "TEAM_BASERUN_CS"
##  [9] "TEAM_BATTING_HBP" "TEAM_PITCHING_HR" "TEAM_PITCHING_BB" "TEAM_PITCHING_SO"
## [13] "TEAM_FIELDING_E"  "TEAM_FIELDING_DP"
```

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

```
##  [1] "TEAM_BATTING_H"   "TEAM_BATTING_2B"  "TEAM_BATTING_3B"  "TEAM_BATTING_HR"
##  [5] "TEAM_BATTING_BB"  "TEAM_BATTING_SO"  "TEAM_BASERUN_SB"  "TEAM_BASERUN_CS"
##  [9] "TEAM_BATTING_HBP" "TEAM_PITCHING_HR" "TEAM_PITCHING_SO" "TEAM_FIELDING_E"
## [13] "TEAM_FIELDING_DP"
```

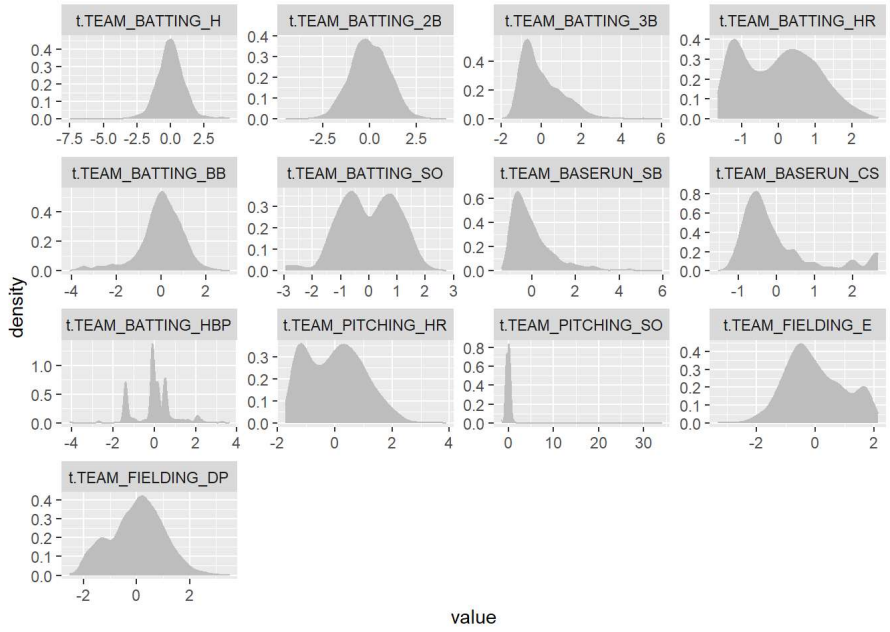
Imputing the NAs using Mice(pmm - predictive mean matching)

```
##
##  iter imp variable
##    1   1  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    1   2  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    1   3  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    1   4  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    1   5  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    2   1  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    2   2  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    2   3  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    2   4  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    2   5  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    3   1  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    3   2  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    3   3  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    3   4  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    3   5  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    4   1  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    4   2  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    4   3  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    4   4  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    4   5  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    5   1  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    5   2  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    5   3  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    5   4  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##    5   5  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_PITCHING_SO  TEAM_FIELDING_DP
```

```
##  TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B  TEAM_BATTING_HR
##  Min.   : 891   Min.   : 69.0   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:1383   1st Qu.:208.0   1st Qu.: 34.00   1st Qu.: 42.00
##  Median :1454   Median :238.0   Median : 47.00   Median :102.00
##  Mean   :1469   Mean   :241.2   Mean   : 55.25   Mean   : 99.61
##  3rd Qu.:1537   3rd Qu.:273.0   3rd Qu.: 72.00   3rd Qu.:147.00
##  Max.   :2554   Max.   :458.0   Max.   :223.00   Max.   :264.00
##  TEAM_BATTING_BB TEAM_BATTING_SO  TEAM_BASERUN_SB TEAM_BASERUN_CS
##  Min.   :  0.0   Min.   :   0.0   Min.   :  0     Min.   :  0.00
##  1st Qu.:451.0   1st Qu.: 542.8   1st Qu.: 67     1st Qu.: 42.00
##  Median :512.0   Median : 735.0   Median :105     Median : 56.00
##  Mean   :501.6   Mean   : 728.0   Mean   :132     Mean   : 73.34
##  3rd Qu.:580.0   3rd Qu.: 925.0   3rd Qu.:166     3rd Qu.: 85.25
##  Max.   :878.0   Max.   :1399.0   Max.   :697     Max.   :201.00
##  TEAM_BATTING_HBP TEAM_PITCHING_HR TEAM_PITCHING_SO  TEAM_FIELDING_E
##  Min.   :29.00    Min.   :  0.0    Min.   :    0.0   Min.   :  65.0
##  1st Qu.:51.00    1st Qu.: 50.0    1st Qu.:  612.0   1st Qu.: 127.0
##  Median :51.00    Median :107.0    Median :  805.0   Median : 159.0
##  Mean   :52.44    Mean   :105.7    Mean   :  811.8   Mean   : 246.5
##  3rd Qu.:56.00    3rd Qu.:150.0    3rd Qu.:  961.0   3rd Qu.: 249.2
##  Max.   :95.00    Max.   :343.0    Max.   :19278.0   Max.   :1898.0
##  TEAM_FIELDING_DP
##  Min.   : 52.0
##  1st Qu.:124.0
##  Median :145.0
##  Mean   :141.3
##  3rd Qu.:162.0
##  Max.   :228.0
```

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

```
##  t.TEAM_BATTING_H    t.TEAM_BATTING_2B  t.TEAM_BATTING_3B t.TEAM_BATTING_HR
##  Min.   :-7.537074   Min.   :-4.48108   Min.   :-1.9776   Min.   :-1.64521
##  1st Qu.:-0.573089   1st Qu.:-0.68949   1st Qu.:-0.7606   1st Qu.:-0.95153
##  Median :-0.003988   Median :-0.03019   Median :-0.2953   Median : 0.03944
##  Mean   : 0.000000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000
##  3rd Qu.: 0.586908   3rd Qu.: 0.69827   3rd Qu.: 0.5995   3rd Qu.: 0.78267
##  Max.   : 4.390097   Max.   : 4.05391   Max.   : 6.0042   Max.   : 2.71505
##  t.TEAM_BATTING_BB  t.TEAM_BATTING_SO  t.TEAM_BASERUN_SB t.TEAM_BASERUN_CS
##  Min.   :-4.08866   Min.   :-2.95076   Min.   :-1.3897   Min.   :-1.5330
##  1st Qu.:-0.41215   1st Qu.:-0.75100   1st Qu.:-0.6841   1st Qu.:-0.6551
##  Median : 0.08511   Median : 0.02819   Median :-0.2838   Median :-0.3625
##  Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.63944   3rd Qu.: 0.79826   3rd Qu.: 0.3586   3rd Qu.: 0.2488
##  Max.   : 3.06871   Max.   : 2.71938   Max.   : 5.9511   Max.   : 2.6682
##  t.TEAM_BATTING_HBP t.TEAM_PITCHING_HR t.TEAM_PITCHING_SO t.TEAM_FIELDING_E
##  Min.   :-4.11371   Min.   :-1.72432   Min.   :-1.49540   Min.   :-3.3092
##  1st Qu.:-0.07987   1st Qu.:-0.90864   1st Qu.:-0.36802   1st Qu.:-0.7163
##  Median :-0.07987   Median : 0.02123   Median :-0.01249   Median :-0.1424
##  Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.52490   3rd Qu.: 0.72271   3rd Qu.: 0.27488   3rd Qu.: 0.7096
##  Max.   : 3.64121   Max.   : 3.87123   Max.   :34.01707   Max.   : 2.1432
##  t.TEAM_FIELDING_DP
##  Min.   :-2.53783
##  1st Qu.:-0.64889
##  Median : 0.06651
##  Mean   : 0.00000
##  3rd Qu.: 0.69350
##  Max.   : 3.50065
```



4.0 Build Models

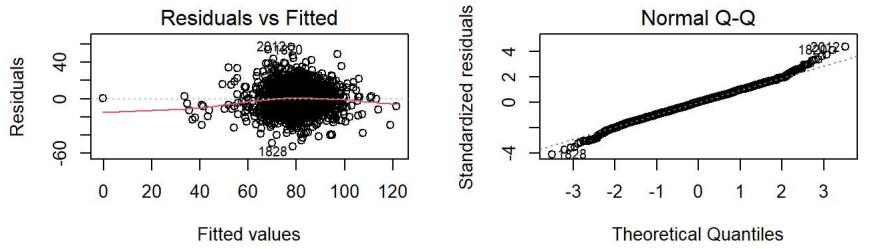
Lets utilize the training data set to create the various models. Use selected variable to build several models to predict wins. The variables selected reflect my strategy of using variables that are related to runs scored and/or runs allowed. Next, in subsequent sections will select the best model and apply the test data set to that model.

Model 1 - All variables included

```
## [1] "TEAM_BATTING_SO"  "TEAM_BASERUN_SB"  "TEAM_BASERUN_CS"  "TEAM_PITCHING_SO"
## [5] "TEAM_FIELDING_DP"
```

```
##   TARGET_WINS     TEAM_BATTING_BB TEAM_BATTING_SO  TEAM_BASERUN_SB
##  Min.   :  0.00   Min.   :  0.0   Min.   :   0.0   Min.   :  0.0
##  1st Qu.: 71.00   1st Qu.:451.0   1st Qu.: 556.8   1st Qu.: 67.0
##  Median : 82.00   Median :512.0   Median : 735.6   Median :106.0
##  Mean   : 80.79   Mean   :501.6   Mean   : 735.6   Mean   :124.8
##  3rd Qu.: 92.00   3rd Qu.:580.0   3rd Qu.: 925.0   3rd Qu.:151.0
##  Max.   :146.00   Max.   :878.0   Max.   :1399.0   Max.   :697.0
##  TEAM_BASERUN_CS  TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB
##  Min.   :  0.00   Min.   : 1137   Min.   :  0.0    Min.   :   0.0
##  1st Qu.: 44.00   1st Qu.: 1419   1st Qu.: 50.0    1st Qu.: 476.0
##  Median : 52.80   Median : 1518   Median :107.0    Median : 536.5
##  Mean   : 52.80   Mean   : 1779   Mean   :105.7    Mean   : 553.0
##  3rd Qu.: 54.25   3rd Qu.: 1682   3rd Qu.:150.0    3rd Qu.: 611.0
##  Max.   :201.00   Max.   :30132   Max.   :343.0    Max.   :3645.0
##  TEAM_PITCHING_SO  TEAM_FIELDING_E  TEAM_FIELDING_DP TEAM_TOTAL_BASES
##  Min.   :    0.0   Min.   :  65.0   Min.   : 52.0    Min.   :1026
##  1st Qu.:  626.0   1st Qu.: 127.0   1st Qu.:134.0    1st Qu.:1947
##  Median :  817.7   Median : 159.0   Median :146.4    Median :2126
##  Mean   :  817.7   Mean   : 246.5   Mean   :146.4    Mean   :2120
##  3rd Qu.:  957.0   3rd Qu.: 249.2   3rd Qu.:161.2    3rd Qu.:2285
##  Max.   :19278.0   Max.   :1898.0   Max.   :228.0    Max.   :3290
```

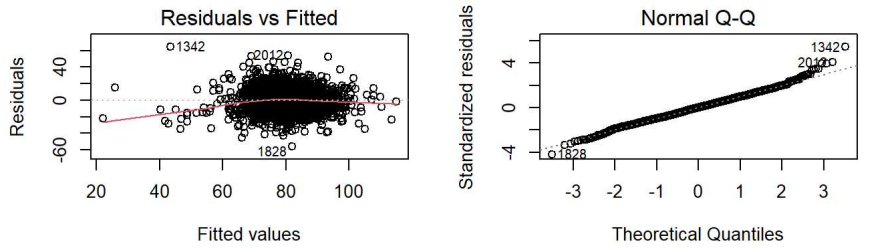
```
##  TEAM_BATTING_BB  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS
##         5.480737         4.541755         1.785279         1.198380
##  TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO
##         3.362423         5.925567         4.922824         3.037166
##  TEAM_FIELDING_E TEAM_FIELDING_DP TEAM_TOTAL_BASES
##         4.120002         1.351790         2.865975
```

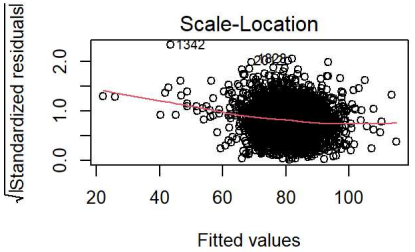


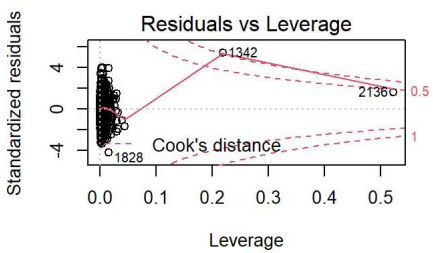
Model 2 - Excludes variables based on possible Multicollinearity

Below shows the summary, vif and diagnostics plot when TEAM_BATTING_SO, TEAM_PITCHING_BB, TEAM_PITCHING_H, TEAM_PITCHING_HR variables are excluded.

```
##  TEAM_BATTING_BB  TEAM_BASERUN_SB  TEAM_BASERUN_CS TEAM_PITCHING_SO
##         2.166066         1.483321         1.143406         1.013311
##  TEAM_FIELDING_E TEAM_FIELDING_DP TEAM_TOTAL_BASES
##         2.221597         1.320254         1.214327
```



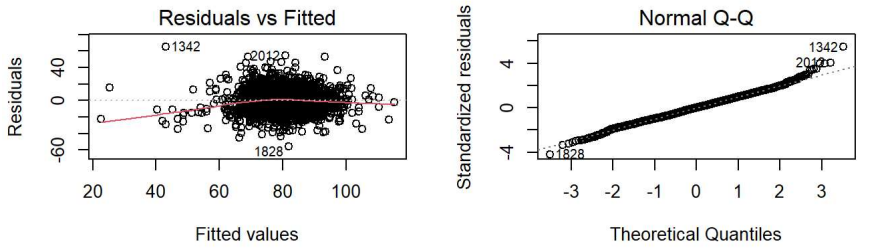


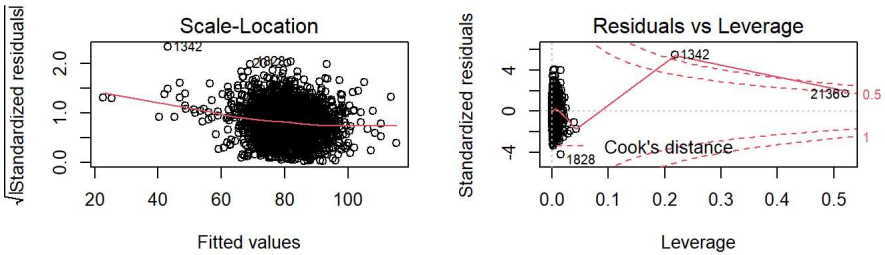


Model 3 - Excludes variable based on insignificant P-value

Below shows the summary, vif and diagnostics plot when TEAM_BASERUN_CS variable is excluded.

```
##  TEAM_BATTING_BB  TEAM_BASERUN_SB TEAM_PITCHING_SO  TEAM_FIELDING_E
##         2.126959         1.357521         1.008308         2.154007
## TEAM_FIELDING_DP TEAM_TOTAL_BASES
##         1.320254         1.200559
```





5.0 Select Models & Predictions

```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = tr_prep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.473  -8.628   0.271   8.340  56.943
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      41.207911   4.622688   8.914  < 2e-16 ***
## TEAM_BATTING_BB   0.004656   0.005256   0.886   0.3758
## TEAM_BATTING_SO  -0.014422   0.002417  -5.968 2.79e-09 ***
## TEAM_BASERUN_SB   0.034669   0.004318   8.029 1.57e-15 ***
## TEAM_BASERUN_CS  -0.012185   0.016159  -0.754   0.4509
## TEAM_PITCHING_H  -0.000823   0.000359  -2.293   0.0220 *
## TEAM_PITCHING_HR -0.014456   0.010938  -1.322   0.1864
## TEAM_PITCHING_BB  0.003042   0.003674   0.828   0.4078
## TEAM_PITCHING_SO  0.001942   0.000888   2.187   0.0288 *
## TEAM_FIELDING_E  -0.018998   0.002455  -7.740 1.48e-14 ***
## TEAM_FIELDING_DP -0.124068   0.013059  -9.501  < 2e-16 ***
## TEAM_TOTAL_BASES  0.031484   0.001797  17.521  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.14 on 2264 degrees of freedom
## Multiple R-squared:  0.3078, Adjusted R-squared:  0.3044
## F-statistic: 91.53 on 11 and 2264 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ . - TEAM_BATTING_SO - TEAM_PITCHING_BB -
##     TEAM_PITCHING_H - TEAM_PITCHING_HR, data = tr_prep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.956  -9.284  -0.038   8.686  64.587
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      32.4154807  3.2790113   9.886  < 2e-16 ***
## TEAM_BATTING_BB   0.0079896  0.0033765   2.366   0.0181 *
## TEAM_BASERUN_SB   0.0372378  0.0040218   9.259  < 2e-16 ***
## TEAM_BASERUN_CS   0.0156849  0.0161282   0.973   0.3309
## TEAM_PITCHING_SO -0.0011024  0.0005241  -2.103   0.0355 *
## TEAM_FIELDING_E  -0.0115523  0.0018417  -6.273 4.24e-10 ***
## TEAM_FIELDING_DP -0.1218095  0.0131868  -9.237  < 2e-16 ***
## TEAM_TOTAL_BASES  0.0285276  0.0011951  23.870  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.42 on 2268 degrees of freedom
## Multiple R-squared:  0.276,  Adjusted R-squared:  0.2738
## F-statistic: 123.5 on 7 and 2268 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ . - TEAM_BATTING_SO - TEAM_PITCHING_BB -
##     TEAM_PITCHING_H - TEAM_PITCHING_HR - TEAM_BASERUN_CS, data = tr_prep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.893  -9.256  -0.047   8.660  64.986
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      33.6918542  3.0048608  11.212  < 2e-16 ***
## TEAM_BATTING_BB   0.0075483  0.0033459   2.256   0.0242 *
## TEAM_BASERUN_SB   0.0383768  0.0038474   9.975  < 2e-16 ***
## TEAM_PITCHING_SO -0.0011382  0.0005228  -2.177   0.0296 *
## TEAM_FIELDING_E  -0.0118647  0.0018134  -6.543 7.44e-11 ***
## TEAM_FIELDING_DP -0.1218116  0.0131867  -9.237  < 2e-16 ***
## TEAM_TOTAL_BASES  0.0284039  0.0011883  23.902  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.42 on 2269 degrees of freedom
## Multiple R-squared:  0.2757, Adjusted R-squared:  0.2738
## F-statistic:   144 on 6 and 2269 DF,  p-value: < 2.2e-16
```

Based on the 3 models, there is no significant difference in R2, Adjusted R2 and RMSE even when i did the treatment for multi-collinearity. The 3rd one will be selected although the R-squared value is not the highest because possible multicollinearity is addressed and all included variables appear to contribute significantly to the model. The TARGET_WINS of the evaluation data set will be predicted using this Model 3. I decided to use model3 for the predictions considering its more parsimonious model.

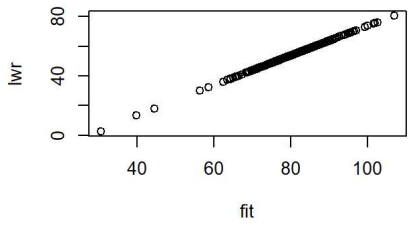
Predictions:

We had to modify our predictions a bit because our final model a) predicted wins > 260 for one observation and b) -783 wins for another. This is clearly poor performance and it may be important to find better options for our model. For now, we simply modify these outlier observations so those maxs and mins are replaced with the maxes and mins of our final training set. For the evaluation dataset also we will be doing all the pre-processing steps. Removing the variables:

```
##
##  iter imp variable
##    1   1  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    1   2  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    1   3  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    1   4  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    1   5  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    2   1  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    2   2  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    2   3  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    2   4  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    2   5  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    3   1  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    3   2  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    3   3  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    3   4  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    3   5  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    4   1  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    4   2  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    4   3  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    4   4  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    4   5  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    5   1  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    5   2  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    5   3  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    5   4  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
##    5   5  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_BATTING_HBP  TEAM_FIELDING_DP
```

```
## [1] "TEAM_BATTING_SO"  "TEAM_BASERUN_SB"  "TEAM_BASERUN_CS"  "TEAM_PITCHING_SO"
## [5] "TEAM_FIELDING_DP"
```

```
##       fit              lwr              upr
##  Min.   : 30.63   Min.   : 2.583   Min.   : 58.67
##  1st Qu.: 75.68   1st Qu.:49.332   1st Qu.:102.02
##  Median : 81.31   Median :54.976   Median :107.65
##  Mean   : 80.46   Mean   :54.091   Mean   :106.82
##  3rd Qu.: 86.07   3rd Qu.:59.706   3rd Qu.:112.44
##  Max.   :106.95   Max.   :80.524   Max.   :133.37
```



6.0 References

Bibliography

Diez, D.M., Barr, C.D., & Cetinkaya-Rundel, M. (2015). OpenIntro Statistics, Third Edition. Open Source. Print

Faraway, J. J. (2015). Extending linear models with R, Second Edition. Boca Raton, FL: Chapman & Hall/CRC. Print

Fox, John (2016). Applied Regression Analysis and Generalized Linear Models, Third Edition. Los Angeles, CA: Sage. Print.

7.0 Resource Links

http://www.baseball-almanac.com/ (http://www.baseball-almanac.com/)

http://tangotiger.net/wiki_archive/Base_Runs.html (http://tangotiger.net/wiki_archive/Base_Runs.html)

https://www.kaggle.com/junkal/selecting-the-best-regression-model (https://www.kaggle.com/junkal/selecting-the-best-regression-model)

https://www.listendata.com/2018/03/regression-analysis.html (https://www.listendata.com/2018/03/regression-analysis.html)

```
---
title: "DATA 621 - Business Analytics and Data Mining"
subtitle: "Homework 1"
author: "Ramnivas Singh"
date: "`r Sys.Date()`"
output:
  pdf_document:
    toc: yes
    toc_depth: '5'
  html_document:
    theme: default
    highlight: espresso
    toc: yes
    toc_depth: 5
    toc_float:
      collapsed: yes
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)
```

1.0 Summary

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:
{width=100%,height=80%}

Deliverables:
1. A write-up submitted in PDF format. Your write-up should have four
sections. Each one is described
below. You may assume you are addressing me as a fellow data scientist, so do
not need to shy away
from technical details.
2. Assigned predictions (the number of wins for the team) for the evaluation
data set.
3. Include your R statistical programming code in an Appendix.

Write-up sections :
1. Data Exploration
2. Data Preparation
3. Build Models
4. Select Models
```

\clearpage

```{r warning=FALSE, message=FALSE}
library(knitr)
library(corrgram)
library(mice)
library(caret)
library(e1071)
library(tidyr)
library(dplyr)
library(ggplot2)
library(psych)
library(reshape)
library(stringr)
library(DT)
library(data.table)
library(kableExtra)
library(corrplot)
library(DMwR2)
library(ggcorrplot)
library(car)
```

2.0 Data Exploration

First of all, load the data and analyze to get some insights like summary, how the data got distributed and correlation between variables and understand the data by using stats, plots and summary. The objective of this is analysis is to develop a better understanding of the data to include its shape, central tendencies, completeness (missing data) and its correlation to our response variable Target_Wins.

```{r}

mb_tr_data <- read.csv("https://raw.githubusercontent.com/rnivas2028/MSDS/
Data621/HW1/moneyball-training-data.csv")
mb_tr_data <- mb_tr_data %>%select(-INDEX)
mb_eval_data <- read.csv("https://raw.githubusercontent.com/rnivas2028/MSDS/
Data621/HW1/moneyball-evaluation-data.csv")
mb_eval_data <- mb_eval_data %>% select(-INDEX)
```

```{r}
summary(mb_tr_data)
```

```{r}
count(mb_tr_data)
```

```{r}
names(mb_tr_data)
```

View rows and columns, variable types

Glimpse of the data shows that all variables are numeric, no ctegorical variable is present here. We do lots of NA for few predcitors in the data set. In our further analysis we will try to identify :

+ Structure of the each predictors
+ How Many NA and Zero , is it significant to remove them or replace them with some predicted value.
+ Statistical summary of the data

```{r, warning=FALSE, message=FALSE}
glimpse(mb_tr_data)
```

Sample 6 rows with sample 7 columns
```{r, warning=FALSE, message=FALSE}
head(mb_tr_data)
```

Show entire dataset of training data
```{r, warning=FALSE, message=FALSE}
DT::datatable(mb_tr_data, options = list(pagelength=5))
```

Here are some key points from data exploration:

* here are multiple variables with missing (NA) values and TEAM-BATTING_HBP has the highest NAs.
* The data is generally complete, however, six variables have missing data.
* The lowest complete rate is for the variable Hit By Pitch, with a rate of only 8%.
* The data set includes 2276 rows, 16 columns with all variables are numeric
* The response variable appears to be normally or near-normally distributed.

Additional Data Exploration
Skewness in the data :
```{r}
mb_tr_data1 = melt(mb_tr_data)
ggplot(mb_tr_data1, aes(x= value)) +
    geom_density(fill = "grey", color="grey") +
  facet_wrap(~variable, scales ="free", ncol = 4)
```

The majority of the explanatory variables appear to be normal or near-normal. There are however, several variable that have bi-modal distributions (Batting_HR, SO, Pitching_HR, Batting_SO) and others that are right-skewed (Fielding, Pitching_BB, Pitching_H)

```{r}
```

```
par(mfrow=c(3,5))
x <- c(2:16)
for (val in x) {
  boxplot(mb_tr_data[,val], xlab=names(mb_tr_data[val]))
}
```

Response Variable & Correlations
```{r}
par(mfrow=c(3,5))
for (val in x) {
plot(mb_tr_data[,val],mb_tr_data$TARGET_WINS, xlab=names(mb_tr_data[val]))
}
```

```{r}
mb_tr_data2 <- mb_tr_data[,-1 ]
names(mb_tr_data2)
cor(drop_na(mb_tr_data2))
```

```{r, echo=FALSE, warning=FALSE, message=FALSE}
mat<-as.matrix(cor(mb_tr_data2[-1],use="pairwise.complete.obs"))
corrplot(mat,tl.cex=.5)
```

```{r}
pairs.panels(mb_tr_data2[1:8])
pairs.panels(mb_tr_data2[9:15])
```

Outliers
```{r}
ggplot(stack(mb_tr_data), 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'))
```

Missing, NA and Zero

We are trying to see how many `NA` is present in the dataset.
```{r, warning=FALSE, message=FALSE}
mb_tr_data %>%
  gather(variable, value) %>%
  filter(is.na(value)) %>%
  group_by(variable) %>%
  tally() %>%
  mutate(percent = n / nrow(mb_tr_data) * 100) %>%
```

```
  mutate(percent = paste0(round(percent, ifelse(percent < 10, 1, 0)), "%"))
%>%
  arrange(desc(n)) %>%
  kable() %>%
  kable_styling()
```

```{r, warning=FALSE, message=FALSE}
mb_tr_data %>%
  gather(variable, value) %>%
  filter(value == 0) %>%
  group_by(variable) %>%
  tally() %>%
  mutate(percent = n / nrow(mb_tr_data) * 100) %>%
  mutate(percent = paste0(round(percent, ifelse(percent < 10, 1, 0)), "%"))
%>%
  arrange(desc(n)) %>%
  kable() %>%
  kable_styling()
```
As can be inferred from above, there are very few zero values exists.

\clearpage

3.0 Data Preparation

Data preparation will include addressing missing data, outliers and feature engineering or creating new variables.

Outliers - The box plot for Wins should some very low values (even zero). According to major league baseball, the lowest number of wins recorded by a team was 20 by the Cleveland Spiders in 1899. Therefore, I will remove all rows from the data set with wins less than 20. The highest number of wins was 116, earned by the Seattle Mariners in 2001. I will also adjust the data set accordingly.

Missing Data - EDA identified variables with missing data. Given strategy I must address the missing data for two variables: Hit By Pitch and Caught Stealing. I will utilize historical major league baseball averages of these two variable as my replacement data.

The variable TEAM_BATTING_HBP is having mostly missing values so the variable will be removed completely.
```{r}
mb_tr_data_f <- mb_tr_data[,-1 ]
names(mb_tr_data_f)

mb_tr_data_f <- mb_tr_data_f[,-10 ]
names(mb_tr_data_f )
```

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

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

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

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

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

```{r}
t = preProcess(imputed_mb_tr_data_Data,
                   c("BoxCox", "center", "scale"))
mb_tr_data_final = data.frame(
      t = predict(t, imputed_mb_tr_data_Data))
summary(mb_tr_data_final)
```

```{r}
mb_tr_data_final1 = melt(mb_tr_data_final)
ggplot(mb_tr_data_final1, aes(x= value)) +
    geom_density(fill = "grey", color="grey")+
  facet_wrap(~variable, scales = 'free')
```

\clearpage

4.0 Build Models

Lets utilize the training data set to create the various models. Use selected variable to build several models to predict wins. The variables selected reflect my strategy of using variables that are related to runs scored and/or runs allowed. Next, in subsequent sections will select the best model and apply the test data set to that model.

Model 1 - All variables included

```{r}
tr_prep <- mb_tr_data %>%
    mutate(TEAM_TOTAL_BASES =
           TEAM_BATTING_H + TEAM_BATTING_2B
           + (2 * TEAM_BATTING_3B) + (3 * TEAM_BATTING_HR))

#remove variable
tr_prep = select(tr_prep,
                 -TEAM_BATTING_H,
                 -TEAM_BATTING_2B,
                 -TEAM_BATTING_3B,
```

```
                -TEAM_BATTING_HR,
                -TEAM_BATTING_HBP)
# Impute missing data
colnames(tr_prep)[colSums(is.na(tr_prep)) > 0]

#impute

tr_prep = tr_prep %>%
  mutate(TEAM_BASERUN_CS =
           ifelse(is.na(TEAM_BASERUN_CS),
                  mean(TEAM_BASERUN_CS, na.rm=TRUE), TEAM_BASERUN_CS)) %>%

  mutate(TEAM_BASERUN_SB =
           ifelse(is.na(TEAM_BASERUN_SB),
                  mean(TEAM_BASERUN_SB, na.rm=TRUE), TEAM_BASERUN_SB)) %>%

  mutate(TEAM_PITCHING_SO =
           ifelse(is.na(TEAM_PITCHING_SO),
                  mean(TEAM_PITCHING_SO, na.rm=TRUE), TEAM_PITCHING_SO)) %>%

  mutate(TEAM_BATTING_SO =
           ifelse(is.na(TEAM_BATTING_SO),
                  mean(TEAM_BATTING_SO, na.rm=TRUE), TEAM_BATTING_SO)) %>%

  mutate(TEAM_FIELDING_DP =
           ifelse(is.na(TEAM_FIELDING_DP),
                  mean(TEAM_FIELDING_DP, na.rm=TRUE), TEAM_FIELDING_DP))

summary(tr_prep)
```

```{r}
model1 <- lm(TARGET_WINS ~., data = tr_prep)
```

```{r}
vif(model1)
par(mfrow=c(2,2))
plot(model1)
```

Model 2 - Excludes variables based on possible Multicollinearity
Below shows the summary, vif and diagnostics plot when TEAM_BATTING_SO, TEAM_PITCHING_BB, TEAM_PITCHING_H, TEAM_PITCHING_HR variables are excluded.

```{r}
model2 <- lm(TARGET_WINS ~ .
             - TEAM_BATTING_SO
             - TEAM_PITCHING_BB
             - TEAM_PITCHING_H
             - TEAM_PITCHING_HR, data = tr_prep)
```

```{r}
```

```
vif(model2)
par(mfrow=c(2,2))
plot(model2)
```

Model 3 - Excludes variable based on insignificant P-value
Below shows the summary, vif and diagnostics plot when TEAM_BASERUN_CS variable is excluded.

```{r}
### Model 3 - Excludes Insignificant variables
model3 <- lm(TARGET_WINS ~ .
             - TEAM_BATTING_SO
             - TEAM_PITCHING_BB
             - TEAM_PITCHING_H
             - TEAM_PITCHING_HR
             - TEAM_BASERUN_CS, data = tr_prep)

```

```{r}
vif(model3)
par(mfrow=c(2,2))
plot(model3)
```

\clearpage

5.0 Select Models & Predictions

```{r}
summary(model1)
summary(model2)
summary(model3)
```

Based on the 3 models, there is no significant difference in R2, Adjusted R2 and RMSE even when i did the treatment for multi-collinearity. The 3rd one will be selected although the R-squared value is not the highest because possible multicollinearity is addressed and all included variables appear to contribute significantly to the model. The TARGET_WINS of the evaluation data set will be predicted using this Model 3. I decided to use model3 for the predictions considering its more parsimonious model.

Predictions:
We had to modify our predictions a bit because our final model a) predicted wins > 260 for one observation and b) -783 wins for another. This is clearly poor performance and it may be important to find better options for our model. For now, we simply modify these outlier observations so those maxs and mins are replaced with the maxes and mins of our final training set. For the evaluation dataset also we will be doing all the pre-processing steps. Removing the variables:

```{r}
mb_eval_data_f <- mb_eval_data[,-1 ]
mb_eval_data_f <- mb_eval_data_f[,-10 ]
mb_eval_data_f <- mb_eval_data_f[,-11 ]
imputed_mb_eval_data_Data <- mice(mb_eval_data_f, m=5, maxit = 5, method =
'pmm')
imputed_mb_eval_data_Data <- complete(imputed_mb_eval_data_Data)
t = preProcess(imputed_mb_eval_data_Data,
                   c("BoxCox", "center", "scale"))
mb_eval_data_final = data.frame(
      t = predict(t, imputed_mb_eval_data_Data))

eval_prep <- mb_eval_data %>%
    mutate(TEAM_TOTAL_BASES =
           TEAM_BATTING_H + TEAM_BATTING_2B
           + (2 * TEAM_BATTING_3B) + (3 * TEAM_BATTING_HR))

#remove variable
eval_prep = select(eval_prep,
                 -TEAM_BATTING_H,
                 -TEAM_BATTING_2B,
                 -TEAM_BATTING_3B,
                 -TEAM_BATTING_HR,
                 -TEAM_BATTING_HBP)
# Impute missing data
colnames(eval_prep)[colSums(is.na(eval_prep)) > 0]

#impute

eval_prep = eval_prep %>%
mutate(TEAM_BASERUN_CS =
           ifelse(is.na(TEAM_BASERUN_CS),
                  mean(TEAM_BASERUN_CS, na.rm=TRUE), TEAM_BASERUN_CS)) %>%

mutate(TEAM_BASERUN_SB =
         ifelse(is.na(TEAM_BASERUN_SB),
                mean(TEAM_BASERUN_SB, na.rm=TRUE), TEAM_BASERUN_SB)) %>%

mutate(TEAM_PITCHING_SO =
         ifelse(is.na(TEAM_PITCHING_SO),
                mean(TEAM_PITCHING_SO, na.rm=TRUE), TEAM_PITCHING_SO)) %>%

mutate(TEAM_BATTING_SO =
         ifelse(is.na(TEAM_BATTING_SO),
                mean(TEAM_BATTING_SO, na.rm=TRUE), TEAM_BATTING_SO)) %>%

mutate(TEAM_FIELDING_DP =
         ifelse(is.na(TEAM_FIELDING_DP),
                  mean(TEAM_FIELDING_DP, na.rm=TRUE), TEAM_FIELDING_DP))
```

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

```

```{r}
summary(eval_data)
```
```{r}
par(mfrow=c(2,2))
plot(eval_data)
```

\clearpage

6.0 References
Bibliography
Diez, D.M., Barr, C.D., & Cetinkaya-Rundel, M. (2015). OpenIntro Statistics, Third Edition. Open Source. Print

Faraway, J. J. (2015). Extending linear models with R, Second Edition. Boca Raton, FL: Chapman & Hall/CRC. Print

Fox, John (2016). Applied Regression Analysis and Generalized Linear Models, Third Edition. Los Angeles, CA: Sage. Print.

7.0 Resource Links
http://www.baseball-almanac.com/

http://tangotiger.net/wiki_archive/Base_Runs.html

https://www.kaggle.com/junkal/selecting-the-best-regression-model

https://www.listendata.com/2018/03/regression-analysis.html
```