DATA 621 – Business Analytics and Data Mining Homework 1

14/09/2019

1. DATA EXPLORATION

There are 16 variables and 2,276 observations in the training data.

```
## Observations: 2,276
## Variables: 16
## $ TARGET_WINS      <dbl> 39, 70, 86, 70, 82, 75, 80, 85, 86, 76, 78, 68,…
## $ TEAM_BATTING_H   <dbl> 1445, 1339, 1377, 1387, 1297, 1279, 1244, 1273,…
## $ TEAM_BATTING_2B  <dbl> 194, 219, 232, 209, 186, 200, 179, 171, 197, 21…
## $ TEAM_BATTING_3B  <dbl> 39, 22, 35, 38, 27, 36, 54, 37, 40, 18, 27, 31,…
## $ TEAM_BATTING_HR  <dbl> 13, 190, 137, 96, 102, 92, 122, 115, 114, 96, 8…
## $ TEAM_BATTING_BB  <dbl> 143, 685, 602, 451, 472, 443, 525, 456, 447, 44…
## $ TEAM_BATTING_SO  <dbl> 842, 1075, 917, 922, 920, 973, 1062, 1027, 922,…
## $ TEAM_BASERUN_SB  <dbl> NA, 37, 46, 43, 49, 107, 80, 40, 69, 72, 60, 11…
## $ TEAM_BASERUN_CS  <dbl> NA, 28, 27, 30, 39, 59, 54, 36, 27, 34, 39, 79,…
## $ TEAM_BATTING_HBP <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…
## $ TEAM_PITCHING_H  <dbl> 9364, 1347, 1377, 1396, 1297, 1279, 1244, 1281,…
## $ TEAM_PITCHING_HR <dbl> 84, 191, 137, 97, 102, 92, 122, 116, 114, 96, 8…
## $ TEAM_PITCHING_BB <dbl> 927, 689, 602, 454, 472, 443, 525, 459, 447, 44…
## $ TEAM_PITCHING_SO <dbl> 5456, 1082, 917, 928, 920, 973, 1062, 1033, 922…
## $ TEAM_FIELDING_E  <dbl> 1011, 193, 175, 164, 138, 123, 136, 112, 127, 1…
## $ TEAM_FIELDING_DP <dbl> NA, 155, 153, 156, 168, 149, 186, 136, 169, 159…
```

Summary Statistics of the variables

Table continues below

TARGET_ WINS	TEAM_BAT TING_H	TEAM_BATT ING_2B	TEAM_BATT ING_3B	TEAM_BATT ING_HR	TEAM_BATT ING_BB
Min. : 0.00	Min. : 891	Min. : 69.0	Min. : 0.00	Min. : 0.00	Min. : 0.0
1st Qu.: 71.00	1st Qu.:1383	1st Qu.:208.0	1st Qu.: 34.00	1st Qu.: 42.00	1st Qu.:451.0
Median : 82.00	Median :1454	Median :238.0	Median : 47.00	Median :102.00	Median :512.0
Mean : 80.79	Mean :1469	Mean :241.2	Mean : 55.25	Mean : 99.61	Mean :501.6
3rd Qu.: 92.00	3rd Qu.:1537	3rd Qu.:273.0	3rd Qu.: 72.00	3rd Qu.:147.00	3rd Qu.:580.0
Max. :146.00	Max. :2554	Max. :458.0	Max. :223.00	Max. :264.00	Max. :878.0

NA	NA	NA	NA	NA	NA

Table continues below

TEAM_BATTING_SO	TEAM_BASERUN_SB	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H	TEAM_PITCHING_HR
Min. : 0.0	Min. : 0.0	Min. : 0.0	Min. :29.00	Min. : 1137	Min. : 0.0
1st Qu.: 548.0	1st Qu.: 66.0	1st Qu.: 38.0	1st Qu.:50.50	1st Qu.: 1419	1st Qu.: 50.0
Median : 750.0	Median :101.0	Median : 49.0	Median :58.00	Median : 1518	Median :107.0
Mean : 735.6	Mean :124.8	Mean : 52.8	Mean :59.36	Mean : 1779	Mean :105.7
3rd Qu.: 930.0	3rd Qu.:156.0	3rd Qu.: 62.0	3rd Qu.:67.00	3rd Qu.: 1682	3rd Qu.:150.0
Max. :1399.0	Max. :697.0	Max. :201.0	Max. :95.00	Max. :30132	Max. :343.0
NA’s :102	NA’s :131	NA’s :772	NA’s :2085	NA	NA

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	NA’s :102	NA	NA’s :286

Box plot of the data

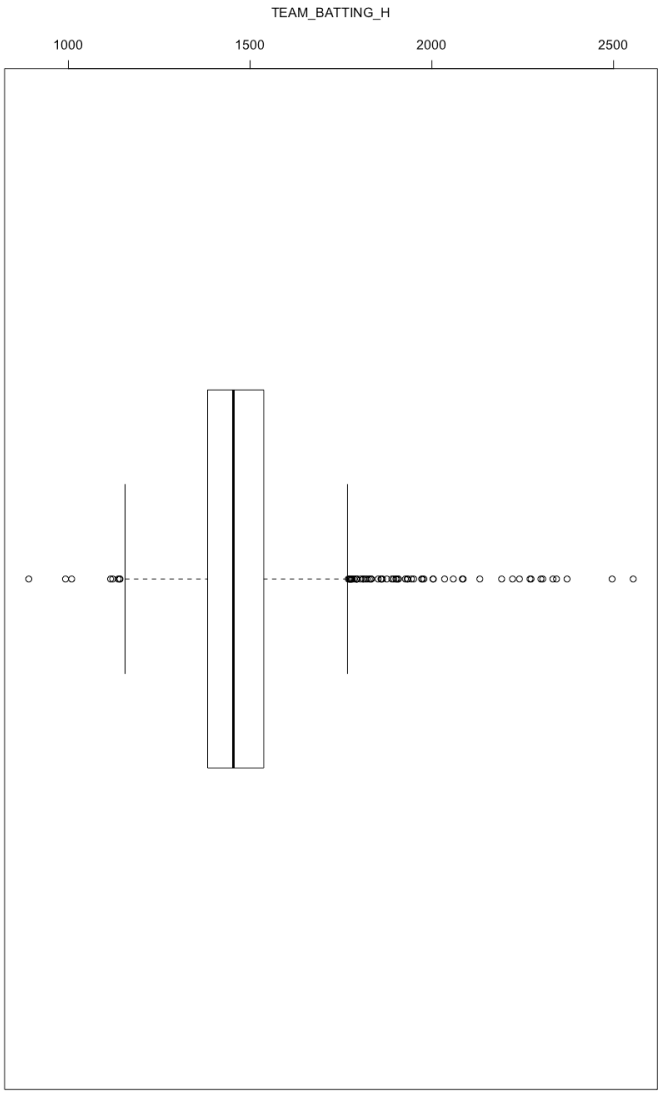
From the following box plots of the variables we can see that the following variables have outliers

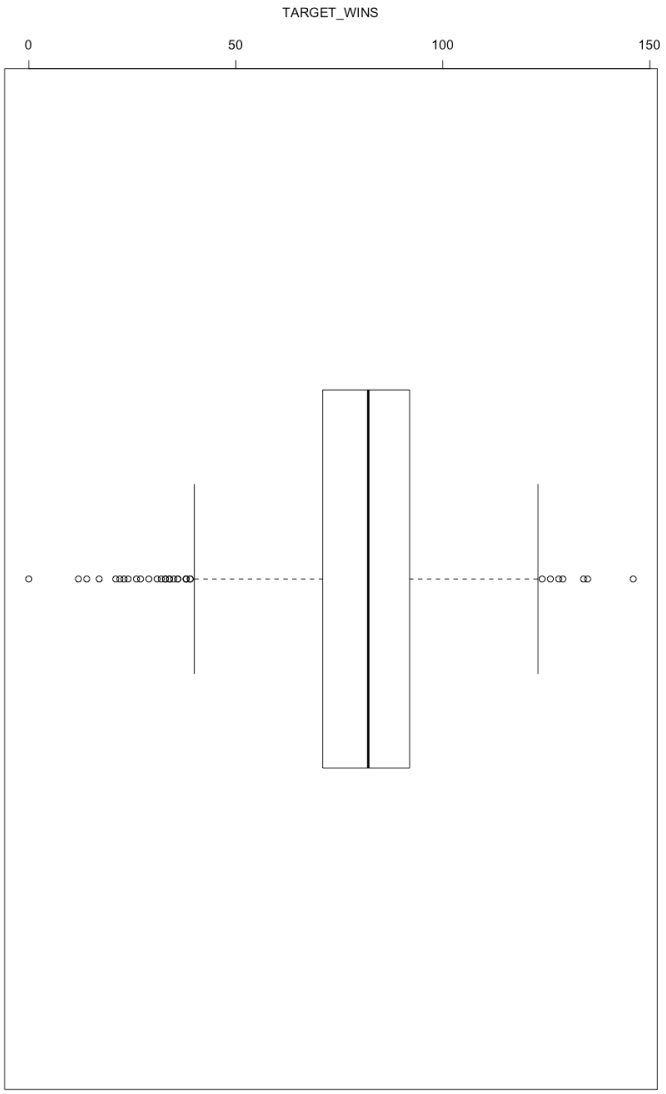
[1] "TARGET_WINS" "TEAM_BATTING_H" "TEAM_BATTING_2B" "TEAM_BATTING_3B" "TEAM_BATTING_BB"

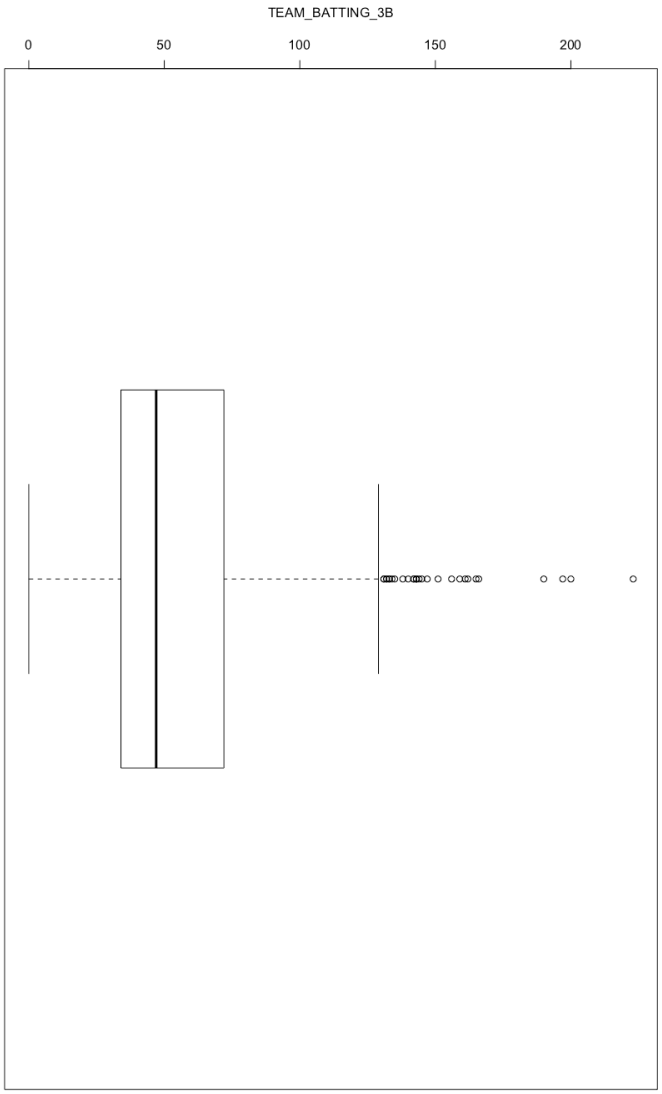
[6] "TEAM_BASERUN_SB" "TEAM_BASERUN_CS" "TEAM_BATTING_HBP" "TEAM_PITCHING_HR" "TEAM_PITCHING_BB"

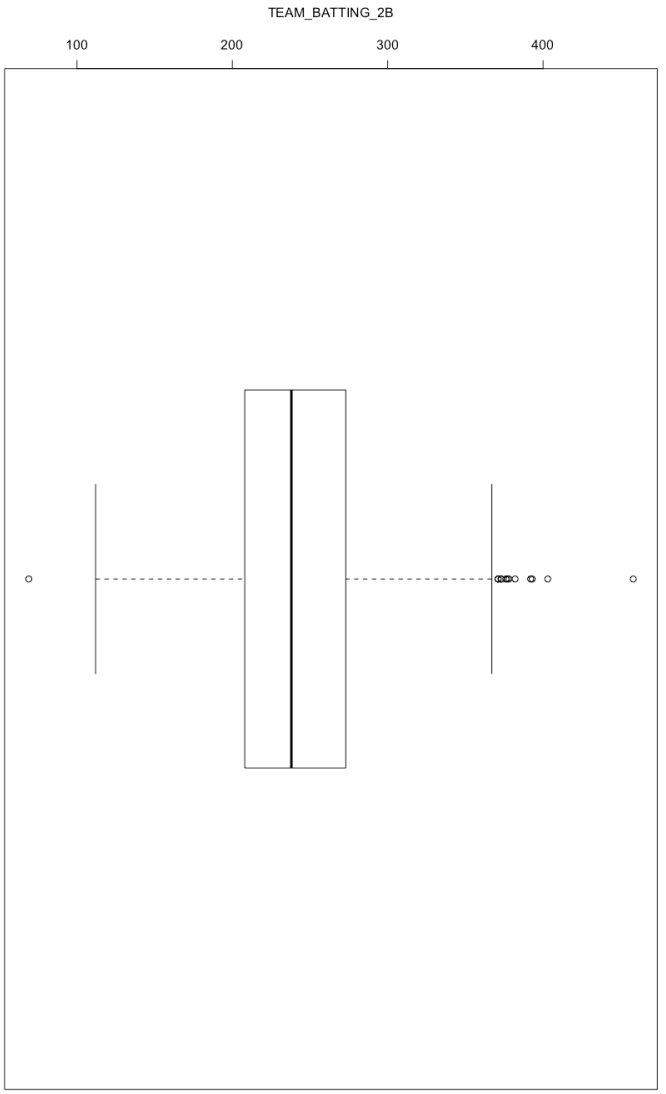
[11] "TEAM_PITCHING_SO" "TEAM_FIELDING_E" "TEAM_FIELDING_DP".

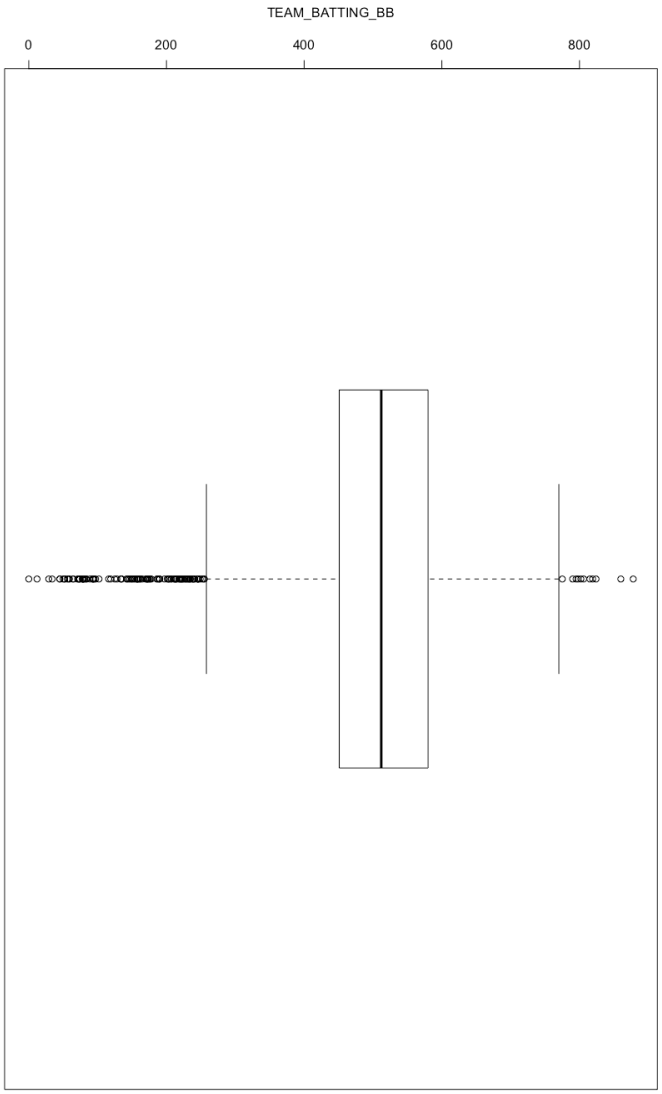
Other variables don’t have any outlier.

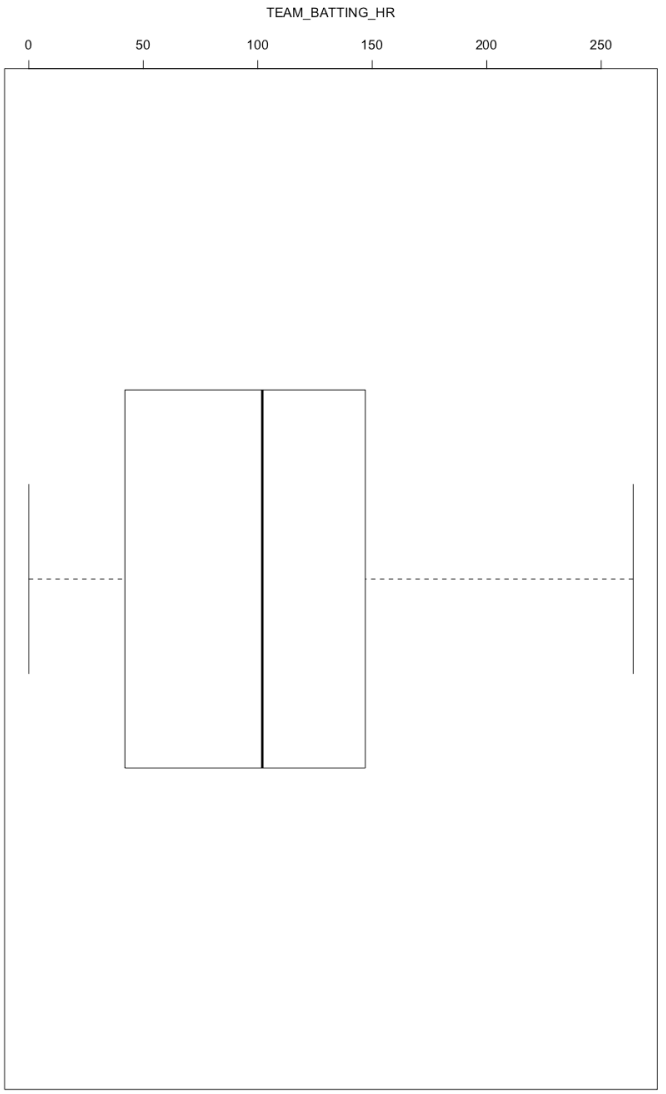

TEAM_BATTING_H
1000
1500
2000
2500

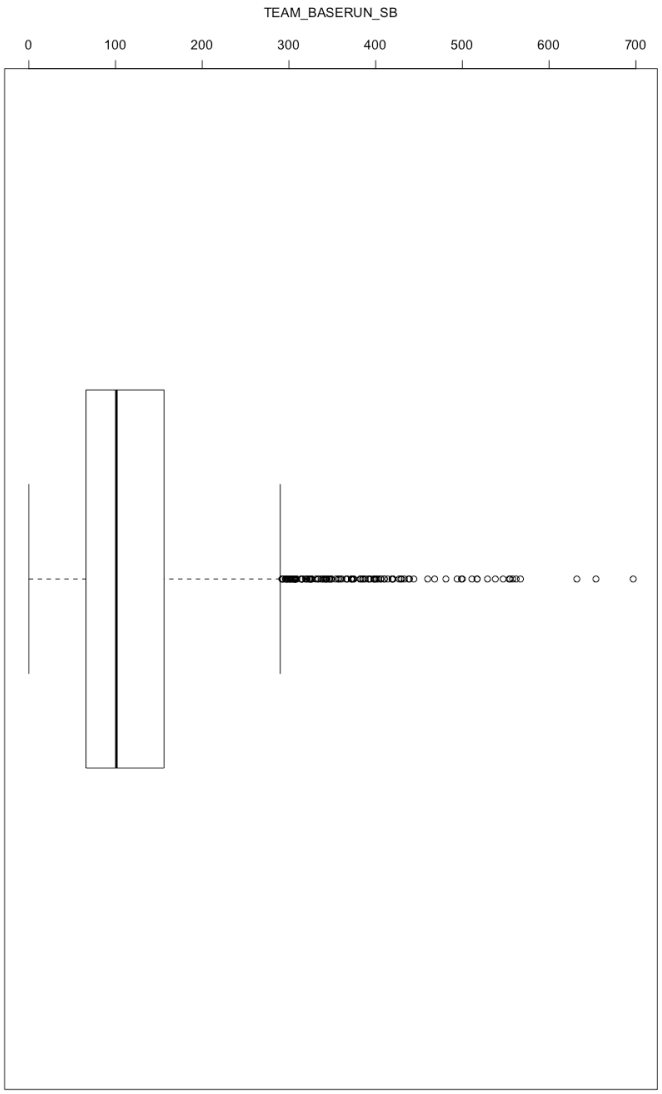

TARGET_WINS
0
50
100
150

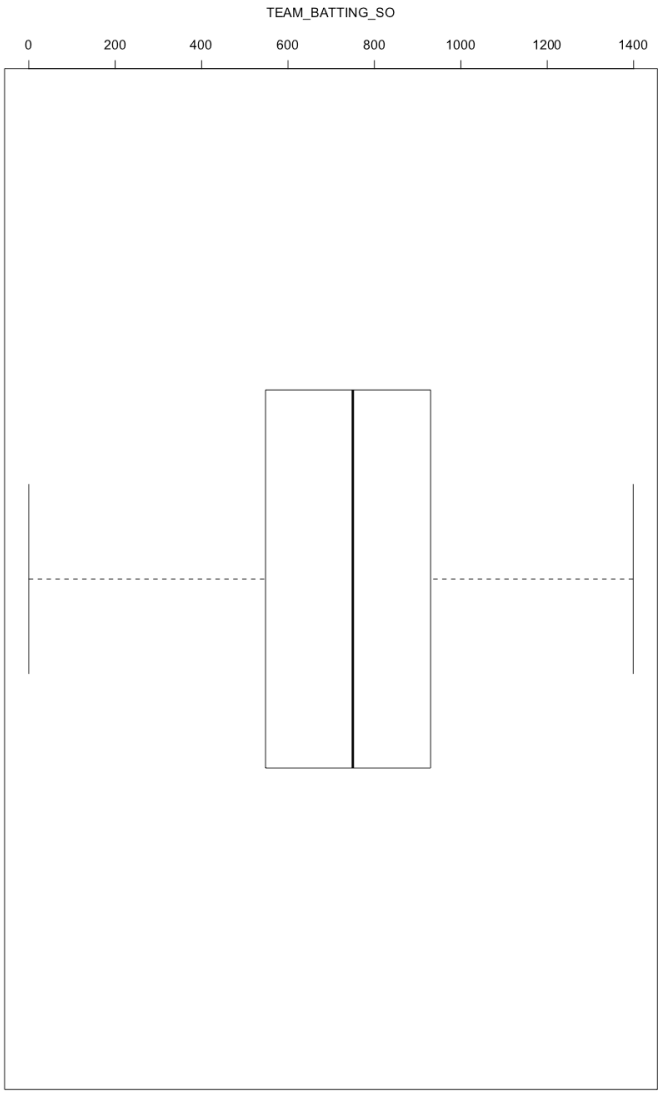

TEAM_BATTING_3B
0
50
100
150
200

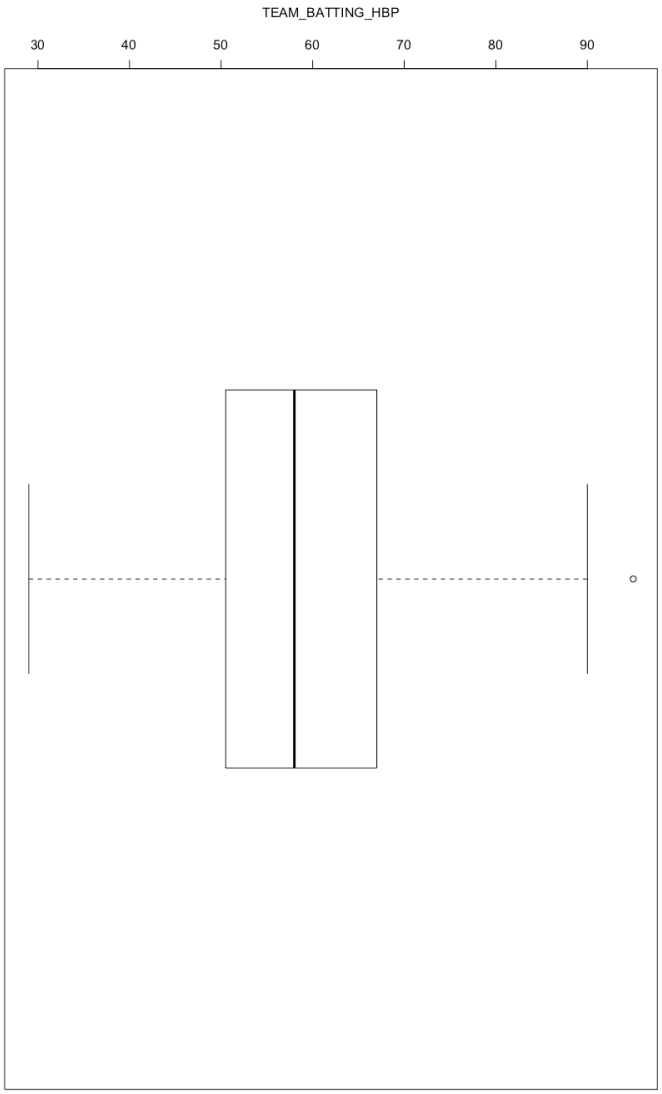

TEAM_BATTING_2B
100
200
300
400

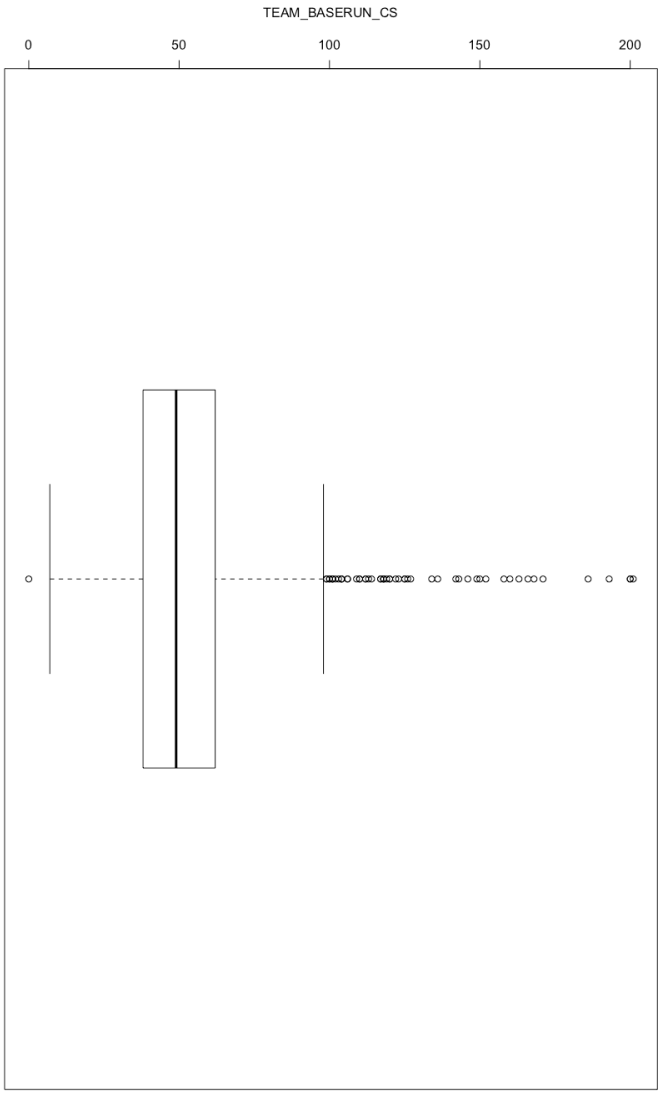

TEAM_BATTING_BB
0
200
400
600
800

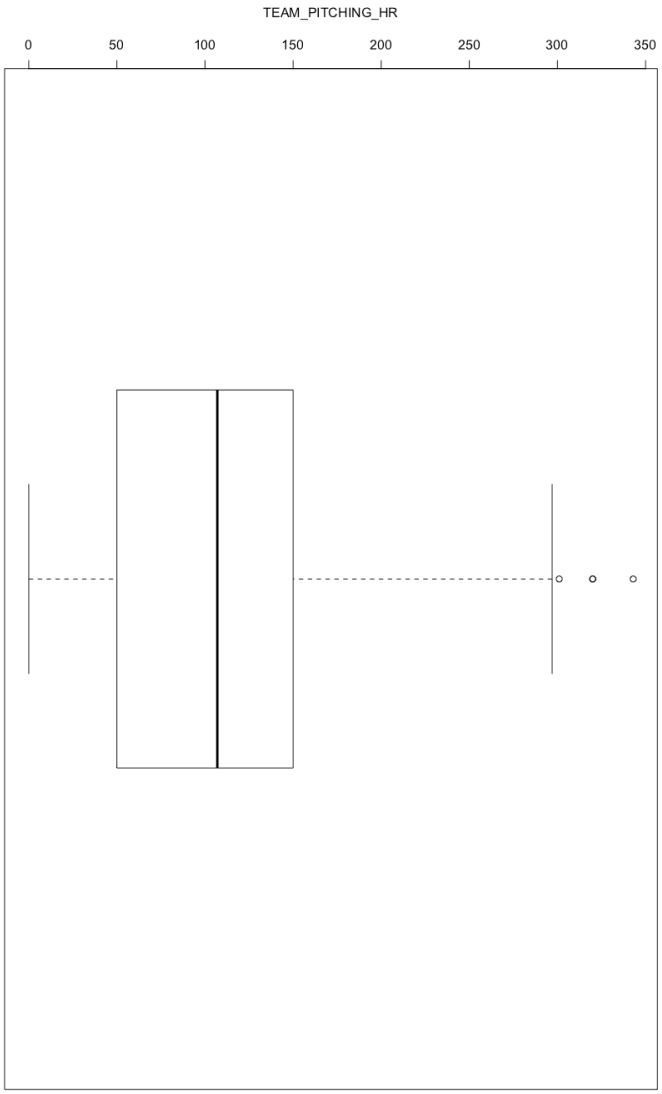

TEAM_BATTING_HR
0
50
100
150
200
250

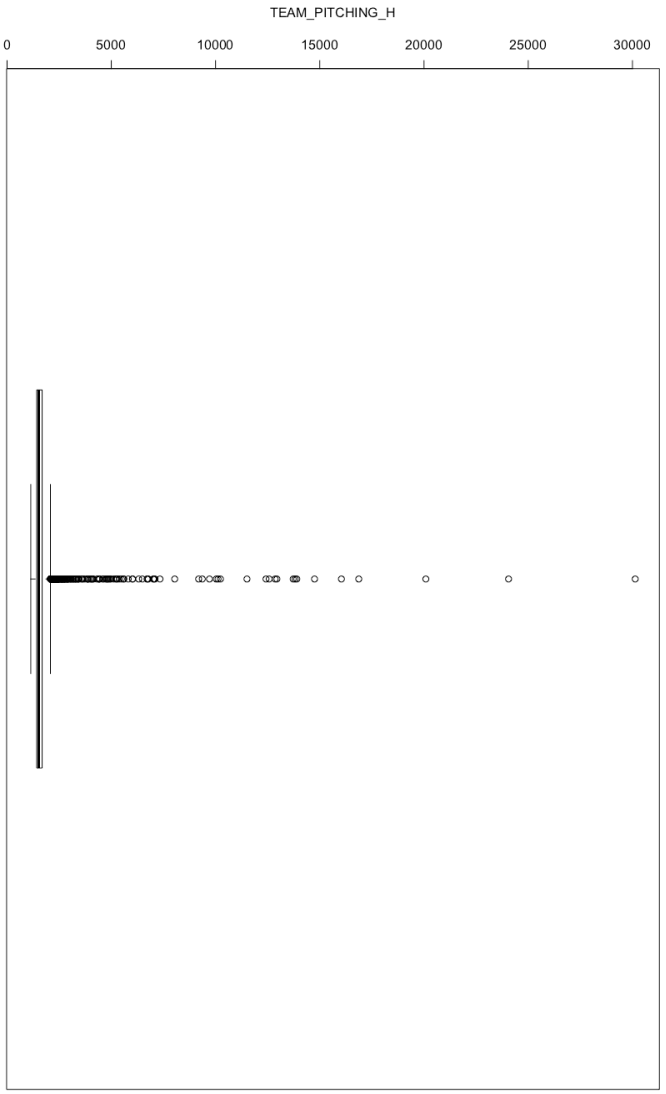

TEAM_BASERUN_SB
0
100
200
300
400
500
600
700

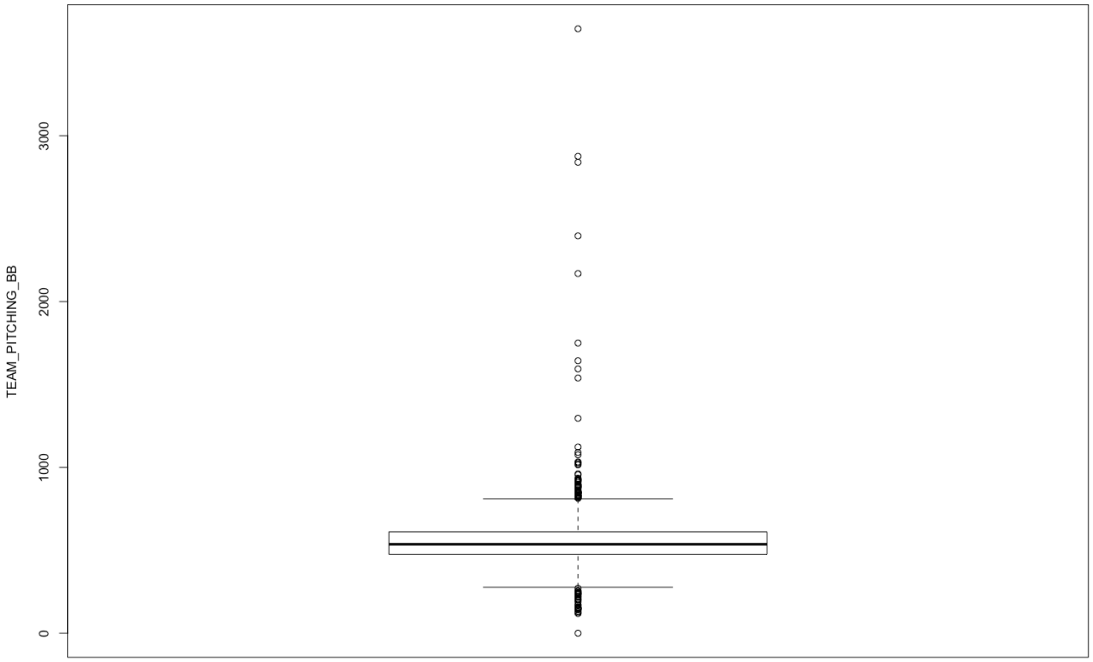

TEAM_BATTING_SO
0
200
400
600
800
1000
1200
1400

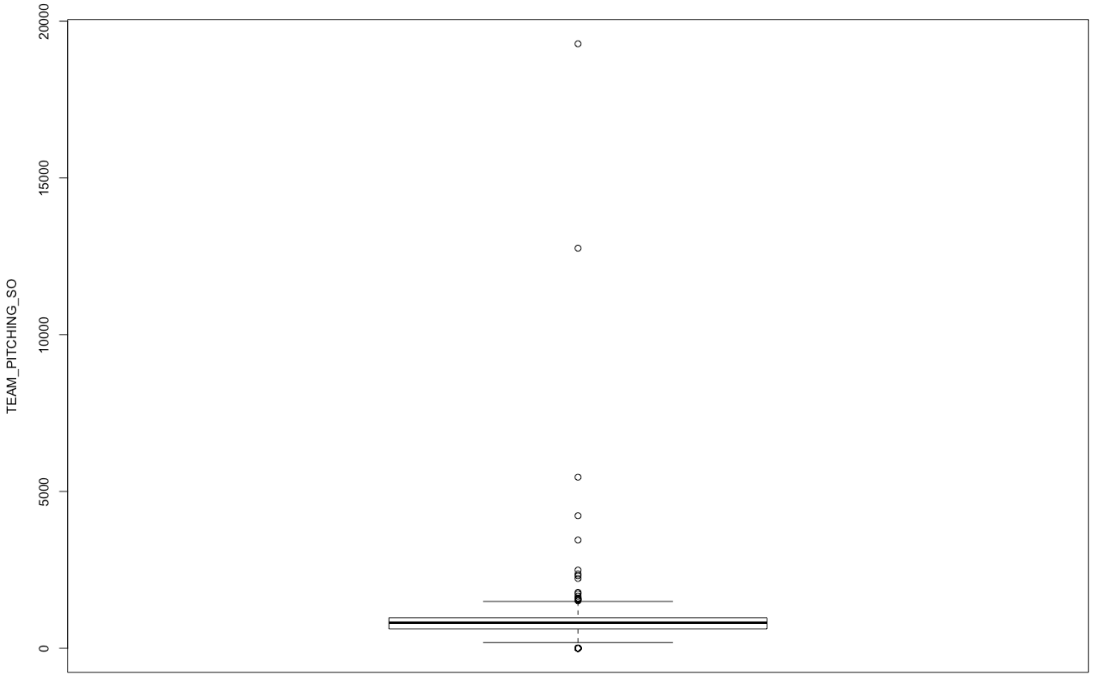

TEAM_BATTING_HBP
30
40
50
60
70
80
90

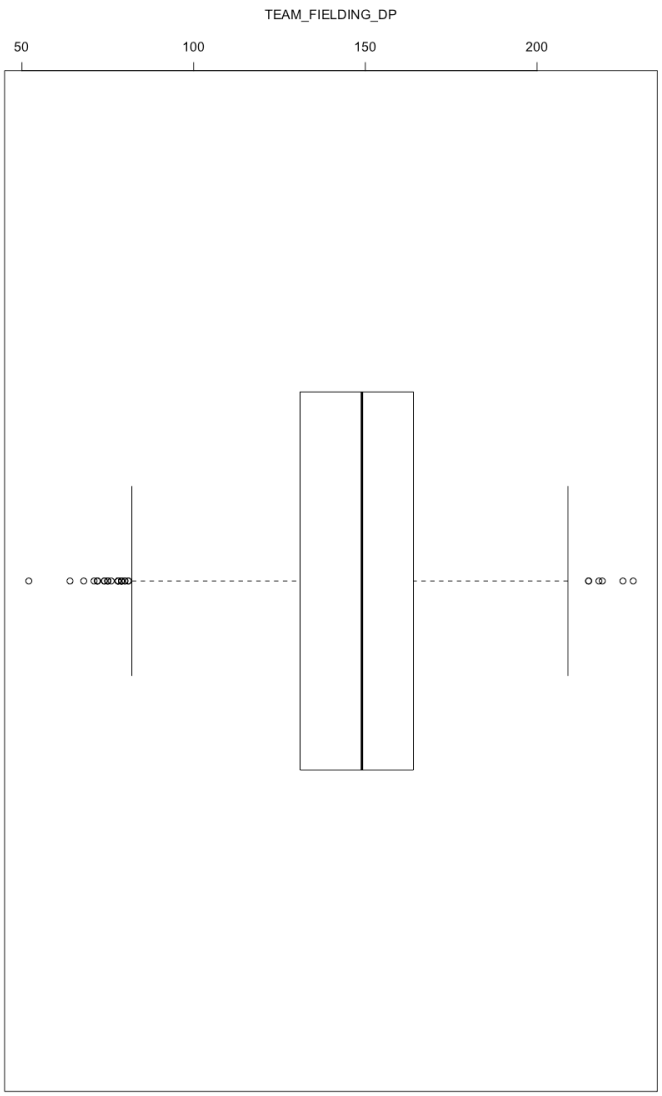

TEAM_BASERUN_CS
0
50
100
150
200

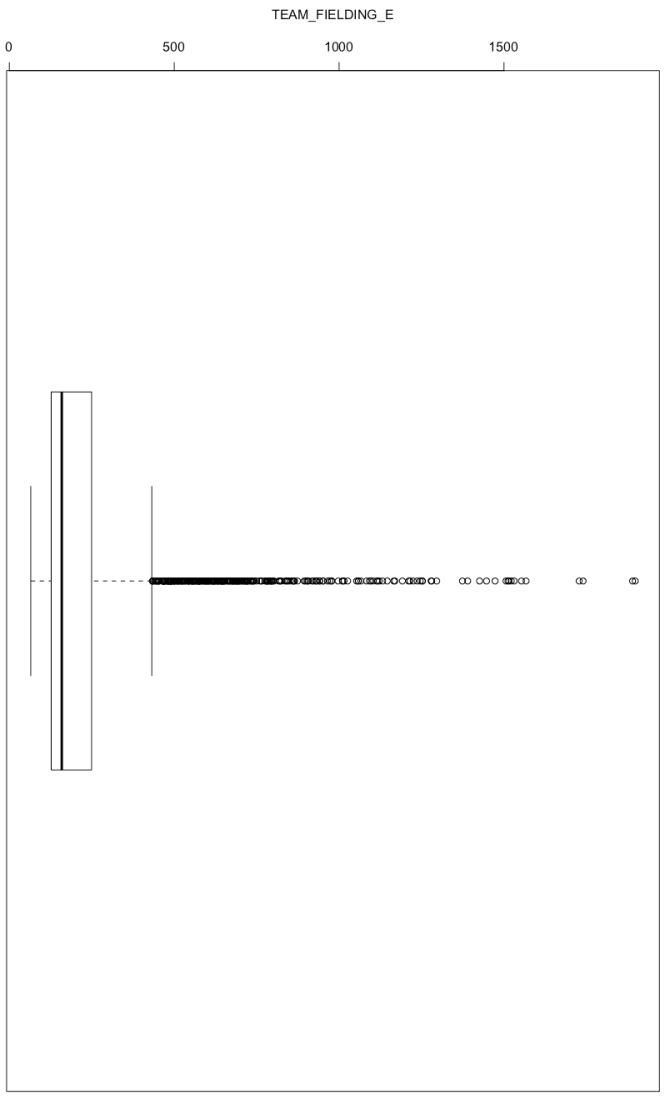

TEAM_PITCHING_HR
0
50
100
150
200
250
300
350

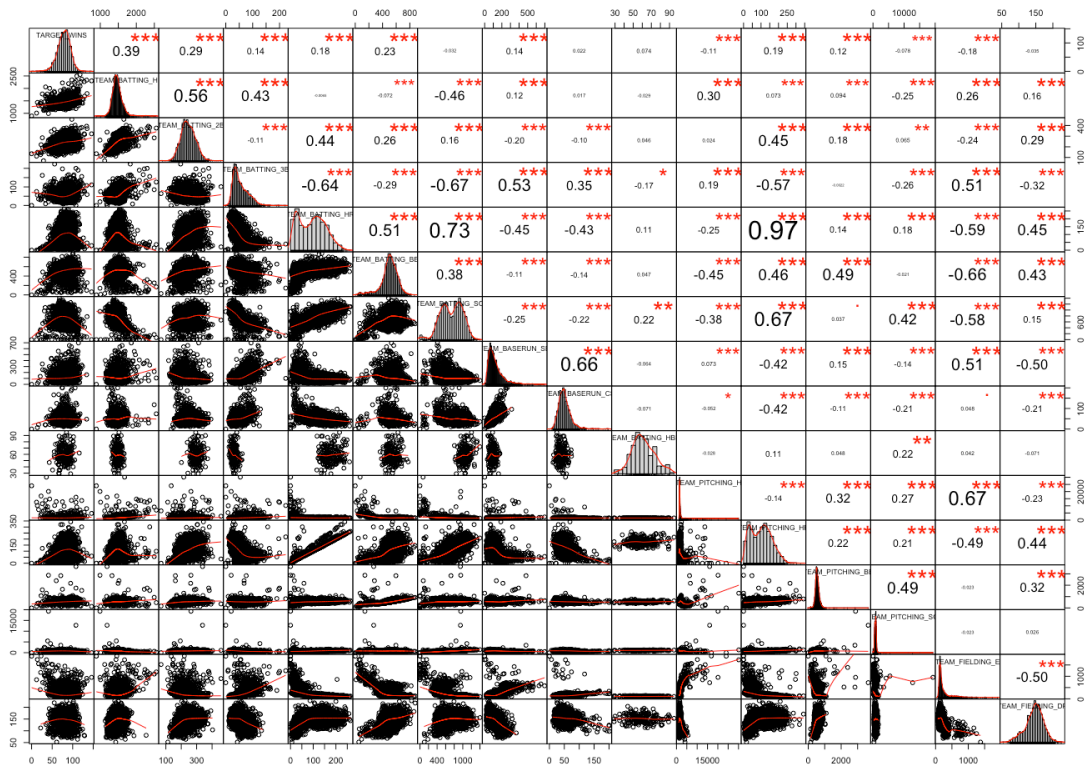

TEAM_PITCHING_H
0
5000
10000
15000
20000
25000
30000


TEAM_PITCHING_BB
0
1000
2000
3000


TEAM_PITCHING_SO
0
5000
10000
15000
20000


TEAM_FIELDING_DP
50
100
150
200


TEAM_FIELDING_E
0
500
1000
1500

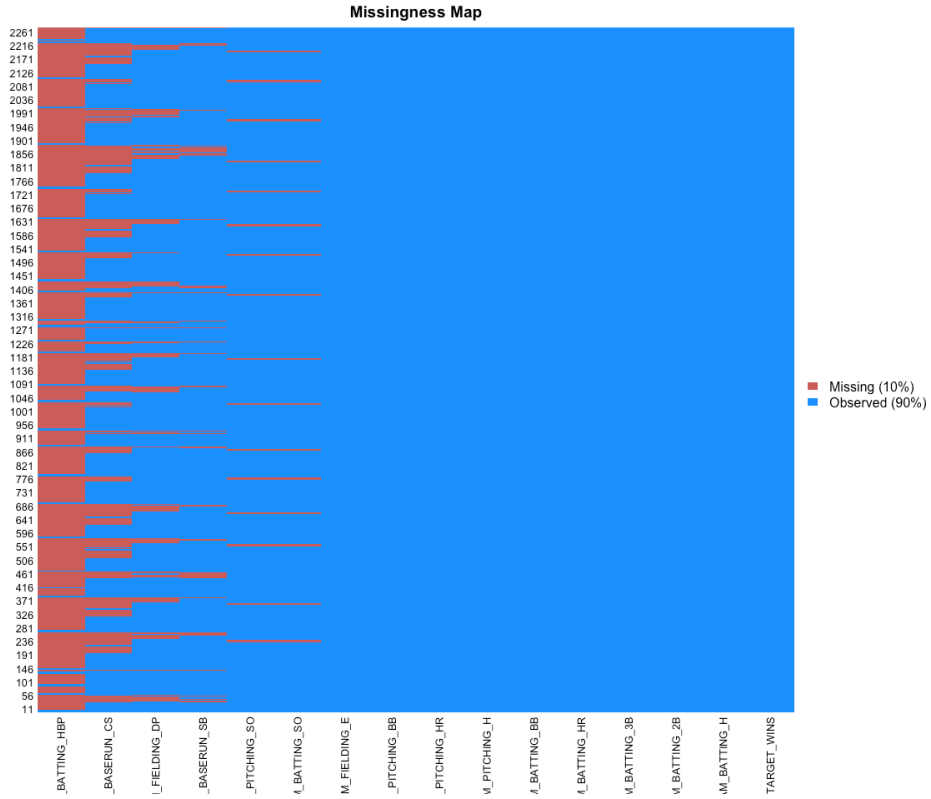


From the above plots histograms we can see that "TARGET_WINS" , "TEAM_BATTING_H" , "TEAM_BATTING_2B" ,"TEAM_BASERUN_CS" "TEAM_FIELDING_DP" are approximately normally distributed.
1,2,3,9,16

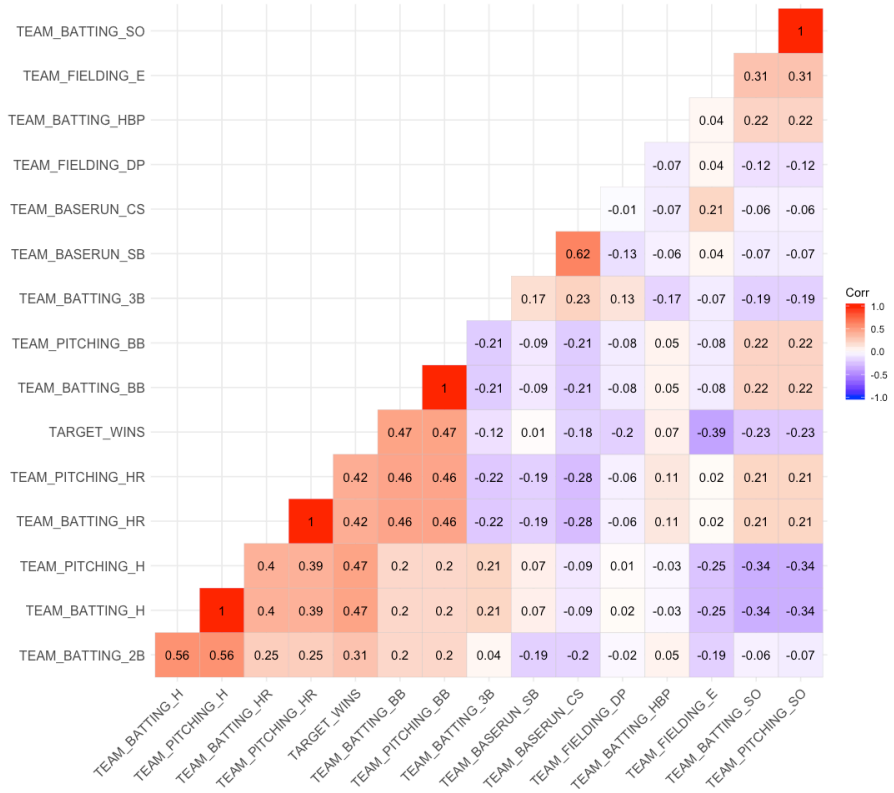
"TEAM_BATTING_3B" c"TEAM_BASERUN_SB" c"TEAM_BASERUN_CS", "TEAM_FIELDING_E"

are positively skewed. These variables have very few high values.

"TEAM_BATTING_HR" , "TEAM_BATTING_SO" , "TEAM_PITCHING_HR" has bi modal distributions. A large number of players scored two modes scores.



From the missing data plot we can see that 10% training data are missing. Most missings are in Team Batting HBP, Team BASERUN CS, Team Filding DP.



From the corrections matrix plot we can see that our Target Wins variable is highly positively correlated with Team Batting BB,Team Pitching BB and negatively correlated with Team Fielding E, Team Batting SO, Team Pitching SO.

2. DATA PREPARATION

a. Missing value imputation

```
Missing values were imputed using mice R package for both the training and
test data set.
```

b. Create flags to suggest if a variable was missing

```
##
##    0    1
## 2085  191
From the flag we can see that 191 observations had missing values.
```

c. Transform data by putting it into buckets
I've transformed TEAM_BATTING_H into 3 buckets based on 0-1200,1200-2000,2000-3000. The new variable is TEAM_BATTING_H.cat. which has 3 categories Low, Medium,High. The old variable was dropped from the data set.

d. Combine variables (such as ratios or adding or multiplying) to create new variables
stolen variable was created summing TEAM_BASERUN_CS and TEAM_BASERUN_SB. Old two variables were dropped from the data.

3. BUILD MODELS

Model 1 with all the variables

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	66.51	4.338	15.33	1.544e-50
TEAM_BATTING_2B	0.04058	0.006974	5.818	6.8e-09
TEAM_BATTING_3B	0.0825	0.01582	5.216	1.992e-07
TEAM_BATTING_HR	0.1187	0.02747	4.32	1.625e-05
TEAM_BATTING_BB	0.01267	0.005732	2.21	0.02717
TEAM_BATTING_SO	-0.0299	0.002299	-13	2.445e-37
TEAM_BATTING_HBP	0.1591	0.02919	5.451	5.54e-08
TEAM_PITCHING_H	0.001164	0.0003972	2.931	0.00341
TEAM_PITCHING_HR	0.007045	0.02432	0.2896	0.7721
TEAM_PITCHING_BB	-0.001544	0.00413	-0.3739	0.7085
TEAM_PITCHING_SO	0.001794	0.0009085	1.975	0.04841
TEAM_FIELDING_E	-0.04154	0.002752	-15.09	4.159e-49
TEAM_FIELDING_DP	-0.1157	0.01291	-8.966	6.279e-19

TEAM_BATTING_H.catMedium	9.968	2.672	3.73	0.000196
TEAM_BATTING_H.catHigh	33.9	4.349	7.794	9.828e-15
stolen	0.039	0.003167	12.31	8.961e-34

Fitting linear model: TARGET_WINS ~ .

Observations	Residual Std. Error	R^2	Adjusted R^2
2276	12.8	0.3437	0.3394

Model 1 interpretations

For 1 unit increase in TEAM_BATTING_2B holding other things constant number of wins increases by 0.04058 units.

For 1 unit increase in TEAM_BATTING_3B holding other things constant number of wins increases by 0.0825 units.

For 1 unit increase in TEAM_BATTING_HR holding other things constant number of wins increases by 0.1187 units.

For 1 unit increase in TEAM_BATTING_BB holding other things constant number of wins increases by 0.01267 units.

For 1 unit increase in TEAM_BATTING_SO holding other things constant number of wins decreases by 0.0299 units.

For 1 unit increase in TEAM_BATTING_HBP holding other things constant number of wins increases by 0.1591 units.

For 1 unit increase in TEAM_PITCHING_H holding other things constant number of wins increases by 0.001164 units.

For 1 unit increase in TEAM_PITCHING_HR holding other things constant number of wins increases by 0.007045 units.

For 1 unit increase in TEAM_PITCHING_BB holding other things constant number of wins decreases by 0.001544 units.

For 1 unit increase in TEAM_PITCHING_SO holding other things constant number of wins increases by 0.001794 units.

For 1 unit increase in TEAM_FIELDING_E holding other things constant number of wins decreases by 0.04154 units.

For 1 unit increase in TEAM_FIELDING_DP holding other things constant number of wins decreases by 0.1157 units.

For 1 unit increase in TEAM_BATTING_H.catMedium holding other things constant number of wins increases by 9.968 units.

For 1 unit increase in TEAM_BATTING_H.catHigh holding other things constant number of wins increase by 33.9 units.
For 1 unit increase in stolen holding other things constant number of wins increase by 0.039 units.

Model 2 dropping non-significant variable

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	66.73	4.238	15.75	4.316e-53
TEAM_BATTING_2B	0.04177	0.006956	6.005	2.224e-09
TEAM_BATTING_3B	0.08301	0.01567	5.298	1.287e-07
TEAM_BATTING_HR	0.1232	0.00859	14.34	1.003e-44
TEAM_BATTING_BB	0.01074	0.00316	3.397	0.0006922
TEAM_BATTING_SO	-0.02728	0.001968	-13.86	5.645e-42
TEAM_BATTING_HBP	0.1582	0.02913	5.431	6.21e-08
TEAM_PITCHING_H	0.001484	0.0003129	4.744	2.225e-06
TEAM_FIELDING_E	-0.04246	0.002718	-15.62	2.54e-52
TEAM_FIELDING_DP	-0.1157	0.01279	-9.046	3.096e-19
TEAM_BATTING_H.catMedium	8.906	2.622	3.396	0.0006944
TEAM_BATTING_H.catHigh	31.66	4.123	7.678	2.392e-14
stolen	0.03984	0.003069	12.98	3.197e-37

Fitting linear model: TARGET_WINS ~ . - TEAM_PITCHING_HR - TEAM_PITCHING_BB - TEAM_PITCHING_SO

Observations	Residual Std. Error	R^2	Adjusted R^2
2276	12.81	0.3421	0.3386

Model 2 interpreations

For 1 unit increase in TEAM_BATTING_2B holding other things constant number of wins increases by 0.04177 units.

For 1 unit increase in TEAM_BATTING_3B holding other things constant number of wins increases by 0.08301 units.

For 1 unit increase in TEAM_BATTING_HR holding other things constant number of wins increases by 0.1232 units.

For 1 unit increase in TEAM_BATTING_BB holding other things constant number of wins increases by 0.01074 units.

For 1 unit increase in TEAM_BATTING_SO holding other things constant number of wins decreases by 0.02728 units.

For 1 unit increase in TEAM_BATTING_HBP holding other things constant number of wins increases by 0.1582 units.

For 1 unit increase in TEAM_PITCHING_H holding other things constant number of wins increases by 0.001484 units.

For 1 unit increase in TEAM_FIELDING_E holding other things constant number of wins decreases by 0.04246 units.

For 1 unit increase in TEAM_FIELDING_DP holding other things constant number of wins decreses by 0.1157 units.

For 1 unit increase in TEAM_BATTING_H.catMedium holding other things constant number of wins increases by 8.906 units.

For 1 unit increase in TEAM_BATTING_H.catHigh holding other things constant number of wins increases by 31.66 units.
For 1 unit increase in stolen holding other things constant number of wins increases by 0.03984 units.

Model 3 with highly correlated variables

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	40.83	3.45	11.83	2.141e-31
TEAM_BATTING_H.catMedium	14.47	2.954	4.899	1.032e-06
TEAM_BATTING_H.catHigh	41.4	4.69	8.827	2.102e-18
TEAM_BATTING_2B	0.07106	0.00758	9.374	1.626e-20
TEAM_BATTING_HR	-0.07093	0.02402	-2.953	0.003178
TEAM_BATTING_BB	0.0215	0.003164	6.794	1.39e-11
TEAM_PITCHING_H	-0.001288	0.0002817	-4.574	5.033e-06
TEAM_PITCHING_HR	0.06504	0.02306	2.82	0.004838

Fitting linear model: TARGET_WINS ~ TEAM_BATTING_H.cat + TEAM_BATTING_2B + TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_PITCHING_H + TEAM_PITCHING_HR

Observations	Residual Std. Error	R^2	Adjusted R^2

2276 14.53 0.15 0.1489

15 Model 3 interpretations

For 1 unit increase in TEAM_BATTING_H.cat Medium holding other things constant number of wins increase by 14.47 units.

For 1 unit increase in TEAM_BATTING_H.catHigh holding other things constant number of wins increase by 41.4 units.

For 1 unit increase in TEAM_BATTING_2B holding other things constant number of wins increase by 0.07106 units.

For 1 unit increase in TEAM_BATTING_HR holding other things constant number of wins decreses by 0.07093 units.

For 1 unit increase in TEAM_BATTING_BB holding other things constant number of wins increase by 0.0215 units.

For 1 unit increase in TEAM_PITCHING_H holding other things constant number of wins decerses by 0.001288 units.

For 1 unit increase in TEAM_PITCHING_HR holding other things constant number of wins increase by 0.06504 units.

4. SELECT MODELS

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?

I've selected model 2 as the best model with slightly worse performance as it's simpler than other is more parsimonious. From the residuals plots of the 3 plots we can see that the residuals of the model 3 is more normally distributed than other models.

Discuss why you selected your model. For the multiple linear regression model, will you use a metric such as Adjusted R2, RMSE, etc.?

For multiple regression model we can use Adjusted R2,RMSE etc as here the dependent variable is continuous.

Be sure to explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output.

Here multicollinearity is eliminated by removing two highly correlated variables in the traihing and evaluation data.

Using the training data set, evaluate the multiple linear regression model based on

(a) mean squared error

model 1 has the lowest mean squared error.

```
## [1] 162.7660 163.1656 210.4413
```

(b) R2

Model 1 has highest Adjusted R2 , model 2 slightly lower and model 3 worst performer here.

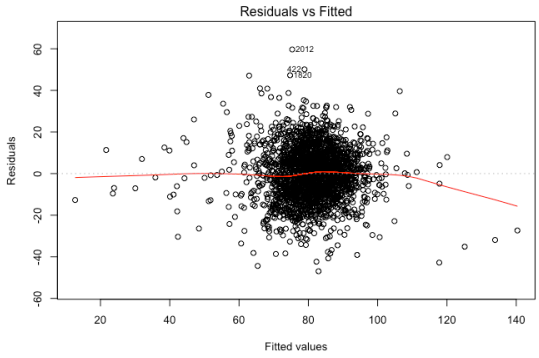
```
## [1] 0.3393862
## [1] 0.338642
## [1] 0.1489004
```

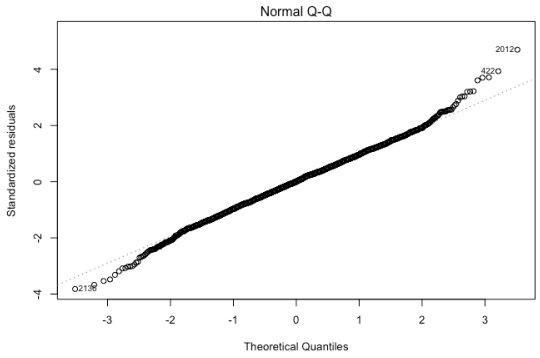
(c) F-statistic

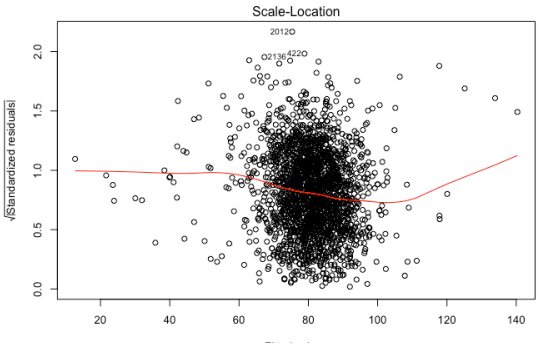
Based on f statistics model 2 is the best model

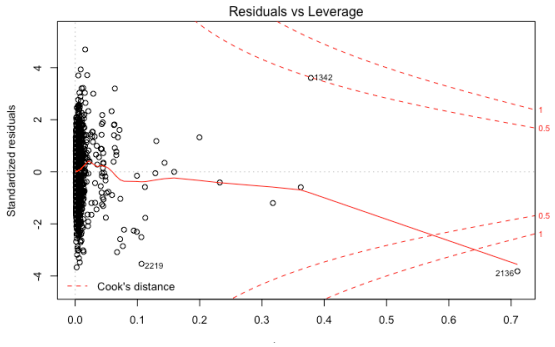
```
##      value      numdf      dendf
##   78.91779   15.00000 2260.00000
##      value      numdf      dendf
##   98.07433   12.00000 2263.00000
##      value      numdf      dendf
##   57.85895    7.00000 2268.00000
```

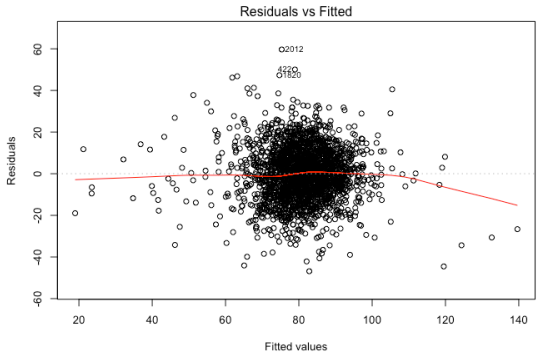
(d) residual plots.

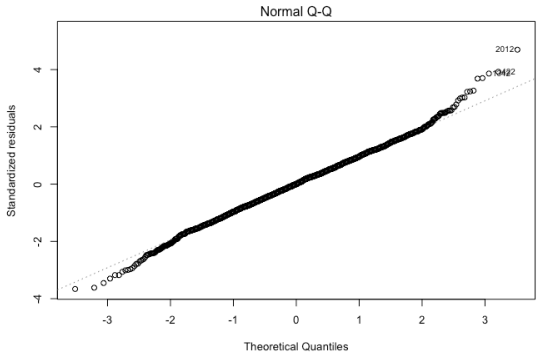

Residuals vs Fitted
Residuals
Fitted values
2012
422
1820

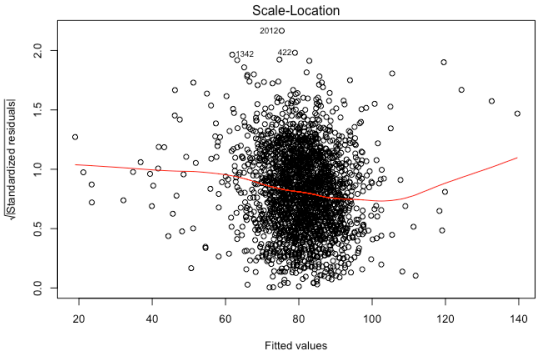

Normal Q-Q
Standardized residuals
Theoretical Quantiles
2012
422
2136

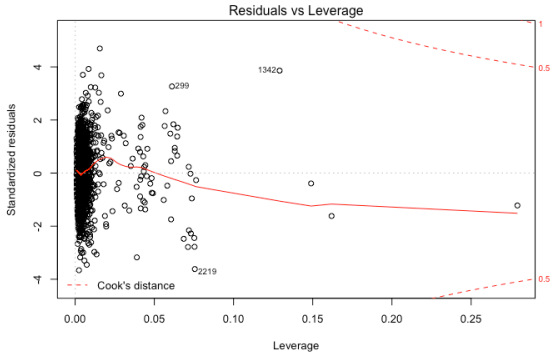

Scale-Location
√|Standardized residuals|
2012
2136
422

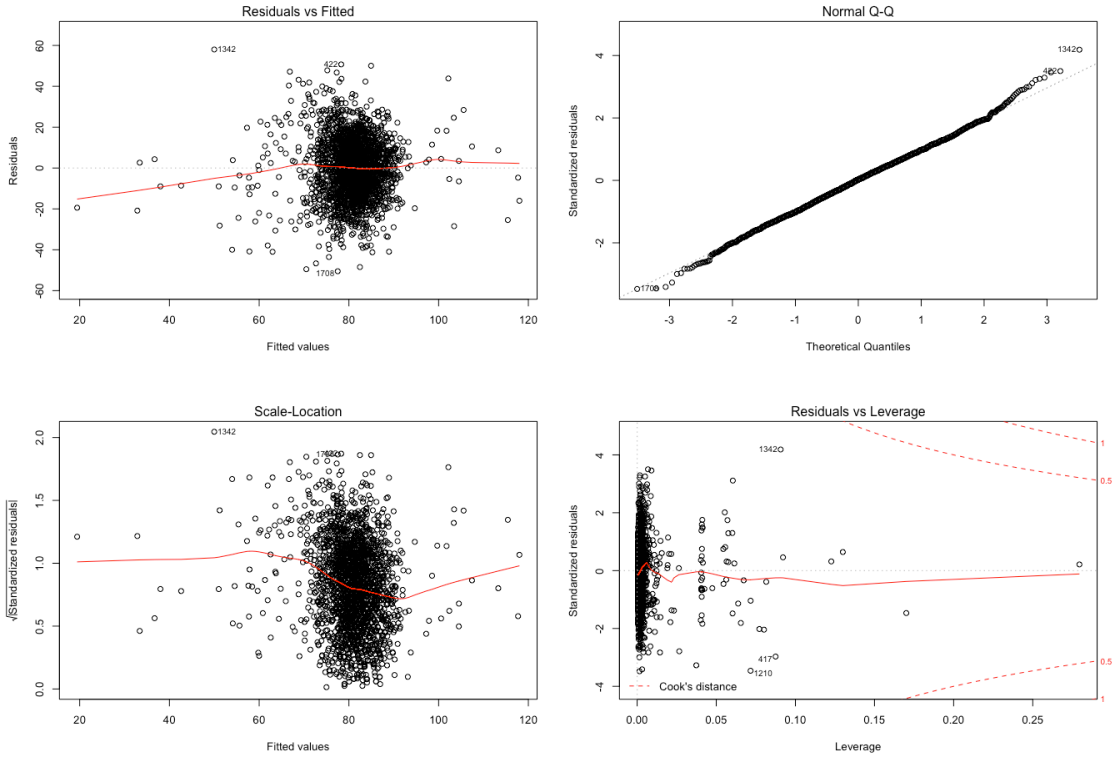

Residuals vs Leverage
Standardized residuals
1342
2219
2136
Cook's distance


Residuals vs Fitted
Residuals
Fitted values
2012
422
1820


Normal Q-Q
Standardized residuals
Theoretical Quantiles
2012


Scale-Location
√|Standardized residuals|
2012
1342
422
Fitted values


Residuals vs Leverage
Standardized residuals
1342
299
2219
Cook's distance
Leverage



Predicted values based on evaluation data.

Make predictions using the evaluation data set.

```
##   [1]  68.48767  71.38504  72.12317  84.83438  70.16620  68.73464
85.68734
##   [8]  76.00445  75.85849  78.19033  72.68052  87.01893  86.71197
88.03763
##  [15]  88.13696  83.66940  75.79230  78.40343  74.22339  95.33227
83.22888
##  [22]  82.78867  82.06625  76.80082  80.28384  88.24047  62.91863
84.15983
##  [29]  85.86068  78.14723  92.96408  85.83207  81.48234  81.11226
79.82247
##  [36]  88.57711  78.00640  91.44307  84.99535  87.58912  85.06656
89.37659
##  [43]  28.54820  98.15649  91.91464  94.94144  95.22529  86.08178
72.61789
##  [50]  92.37001  80.60373  90.31735  74.78800  77.31880  72.74980
76.98705
##  [57]  94.45525  83.10520  59.88745  80.02173  91.12336  76.47366
84.50786
##  [64]  89.88557  89.66803 102.97966  73.05131  79.37974  78.73432
88.57108
##  [71]  86.60946  75.03791  79.20755  97.31598  72.64560  76.23734
80.50222
```

```
##  [78]  83.83428  80.70453  83.93743  92.00608  89.36467  99.35351
80.38606
##  [85]  83.57266  83.57485  88.09016  82.07539  88.44194  93.52788
86.79906
##  [92]  80.78706  79.16872  93.60227  91.40149  91.19427  89.05885
98.27534
##  [99]  85.97692  89.54876  83.29764  77.68317  85.97774  82.52179
74.82078
## [106]  70.70175  57.51297  79.36434  89.55668  60.99673  87.78159
94.73241
## [113]  89.89639  88.85225  82.97740  81.25010  84.63588  84.04608
74.47662
## [120]  80.09763 101.91647  80.97909  76.04747  71.33448  72.16930
91.94506
## [127]  87.90748  83.36340 100.10939  90.71871  89.80031  84.93343
80.72216
## [134]  79.57090  86.32260  82.13408  78.75019  79.79763  87.18288
80.05348
## [141]  67.41627  80.10535  92.25060  73.38117  72.96425  75.61461
81.81643
## [148]  77.30022  76.60572  84.41243  86.27781  80.18331  42.80567
69.74522
## [155]  72.52029  70.95859  93.45498  72.35090  90.11417  82.30247
100.44653
## [162] 103.07066  95.47314 100.96040  96.90498  94.47721  80.85722
91.78653
## [169]  74.71601  83.73195  91.25980  89.58270  83.14932  95.65323
88.45345
## [176]  77.94737  80.80159  72.14018  73.38385  80.19481  92.12023
88.78278
## [183]  85.64648  88.45839  99.82228  99.32677  79.57212  61.82762
72.66872
## [190] 131.76888  79.13778  89.45419  81.33148  79.67157  79.94568
71.72765
## [197]  81.98435  89.47627  82.25859  79.79352  74.24921  73.40594
72.34378
## [204]  96.59306  82.33178  83.00158  74.15658  78.11323  85.95968
76.65071
## [211] 105.15990  89.00342  87.44518  67.67453  78.89580  81.53157
74.68869
## [218]  91.57693  75.90977  79.24671  77.60802  79.39687  82.71303
77.98005
## [225]  79.38975  81.22994  79.70342  76.84217  81.10108  72.15632
86.53929
## [232]  95.59687  80.58169  88.47550  80.14375  77.55778  75.43614
81.23768
## [239]  96.87252  75.12678  90.51053  91.18343  84.84442  79.27575
53.34976
## [246]  87.15169  82.81757  87.13769  77.96667  86.75250  80.17237
55.86816
```

```
## [253]  86.73627  19.26033  69.82397  77.22468  84.44193  88.83353
79.40591
```

- Appendix.

```
## ------------------------------------------------------------------------
library(ggcorrplot)
library(pander)
library(tidyverse)
library(PerformanceAnalytics)
library(Amelia)
library(caret)
library(mice)


## ------------------------------------------------------------------------
training <- read_csv("moneyball-training-data.csv" )
training <- training[,2:ncol(training)]
evaluation <- read_csv("moneyball-evaluation-data.csv")
colId <- evaluation$INDEX
evaluation <- evaluation[,2:ncol(evaluation)]


## ------------------------------------------------------------------------
glimpse(training)


## ------------------------------------------------------------------------
pander(summary(training), split.table=120)


## ------------------------------------------------------------------------
for(col in colnames(training)){
  boxplot(training[,col],ylab = col)
}



## ------------------------------------------------------------------------
chart.Correlation(training)


## ------------------------------------------------------------------------
```

```
missmap(training)

## ------------------------------------------------------------------------
corr <- cor(training, use="complete.ob")
ggcorrplot(corr, hc.order = TRUE, type = "lower",
   lab = TRUE)


## ------------------------------------------------------------------------
imputed_Data <- mice(training, m=1, maxit = 50, method = 'pmm', seed = 500)
complete_data <- complete(imputed_Data,1)

imputed_Data_eval <- mice(evaluation, m=1, maxit = 50, method = 'pmm', seed = 500)
complete_data_evaluation <- complete(imputed_Data_eval,1)


## ------------------------------------------------------------------------
training$flag <- 0
training$flag[rowMeans(training) > 0] <- 1
table(training$flag)


## ------------------------------------------------------------------------
# TEAM_BATTING_H 0-1200,1200-2000,2000-3000
b <- c(-Inf, 1200, 2000, Inf)
names <- c("Low", "Medium", "High")
complete_data$TEAM_BATTING_H.cat <- cut(complete_data$TEAM_BATTING_H, breaks =
b, labels = names)
complete_data_evaluation$TEAM_BATTING_H.cat <-
cut(complete_data_evaluation$TEAM_BATTING_H, breaks = b, labels = names)




## ------------------------------------------------------------------------
complete_data$stolen <- complete_data$TEAM_BASERUN_CS +
complete_data$TEAM_BASERUN_SB
complete_data_evaluation$stolen <- complete_data_evaluation$TEAM_BASERUN_CS +
complete_data_evaluation$TEAM_BASERUN_SB

## ------------------------------------------------------------------------
# drop TEAM_BATTING_H
complete_data <- complete_data %>%
  select(-TEAM_BATTING_H,-TEAM_BASERUN_CS, -TEAM_BASERUN_SB)
# drop TEAM_BATTING_H
evaluation <- evaluation %>%
```

```
  select(-TEAM_BATTING_H,-TEAM_BASERUN_CS, -TEAM_BASERUN_SB)


## ------------------------------------------------------------------------
model1 <- lm(TARGET_WINS~., data = complete_data)
pander(summary(model1))


## ------------------------------------------------------------------------
model2 <- lm(TARGET_WINS~.-TEAM_PITCHING_HR-TEAM_PITCHING_BB-
TEAM_PITCHING_SO, data = complete_data)
pander(summary(model2))


## ------------------------------------------------------------------------
model3 <- lm(TARGET_WINS~TEAM_BATTING_H.cat+TEAM_BATTING_2B
+TEAM_BATTING_HR + TEAM_BATTING_BB+
            TEAM_PITCHING_H+TEAM_PITCHING_HR, data = complete_data )
pander(summary(model3))


## ------------------------------------------------------------------------
mse <- numeric(3)
mse[[1]] <- mean((complete_data$TARGET_WINS - predict(model1))^2)
mse[[2]] <- mean((complete_data$TARGET_WINS - predict(model2))^2)
mse[[3]] <- mean((complete_data$TARGET_WINS - predict(model3))^2)
mse


## ------------------------------------------------------------------------
 summary(model1)$adj.r.squared
 summary(model2)$adj.r.squared
 summary(model3)$adj.r.squared


## ------------------------------------------------------------------------
 summary(model1)$fstatistic
 summary(model2)$fstatistic
 summary(model3)$fstatistic


## ------------------------------------------------------------------------
par(mfrow=c(2,2))
plot(model1)
par(mfrow=c(2,2))
plot(model2)
```

```
par(mfrow=c(2,2))
plot(model3)


## ------------------------------------------------------------------------
complete_data_evaluation$PredictedWins <- predict(model2, complete_data_evaluation)
complete_data_evaluation$PredictedWins

## ------------------------------------------------------------------------
df <- data.frame("Index"=colId,"Predicted wins"= complete_data_evaluation$PredictedWins)
write.csv(df,"predictions.csv")
```