Data 621 Hw1

Vivian Kong

6/5/2018

```
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(plyr)
```

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

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

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

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
library(stringr)
library(imputeTS)
library(DataExplorer)
library(cbanalysis)
library(gvlma)
require(knitr)
```

```
## Loading required package: knitr
```

```
require(lubridate)
```

```
## Loading required package: lubridate
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:plyr':
##
##     here
## The following object is masked from 'package:base':
##
##     date
```

```
moneyball <- read.csv("https://raw.githubusercontent.com/xkong100/data-621/master/Hw1/moneyball-trainin
kable(head(moneyball))
```

INDEX	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATT
1	39	1445	194	39	
2	70	1339	219	22	
3	86	1377	232	35	
4	70	1387	209	38	
5	82	1297	186	27	
6	75	1279	200	36	

```
nrow(moneyball)
```

```
## [1] 2276
```

```
ncol(moneyball)
```

```
## [1] 17
```

```
cleanNames <- function(df) {
    name_list <- names(df)
    name_list <- gsub("TEAM_", "", name_list)
    names(df) <- name_list
    df
}

moneyball <- cleanNames(moneyball)
kable(head(moneyball))
```

INDEX	TARGET_WINS	BATTING_H	BATTING_2B	BATTING_3B	BATTING_HR	BATTING_BB	BA
1	39	1445	194	39	13	143	
2	70	1339	219	22	190	685	
3	86	1377	232	35	137	602	
4	70	1387	209	38	96	451	
5	82	1297	186	27	102	472	
6	75	1279	200	36	92	443	
There ar	e 2276 rows an	d 16 columns	.				

```
summary(moneyball)
```

```
##      INDEX         TARGET_WINS       BATTING_H      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  
##                                                                  
##    BATTING_3B       BATTING_HR       BATTING_BB      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     
##    BASERUN_SB      BASERUN_CS     BATTING_HBP      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                   
##   PITCHING_HR     PITCHING_BB      PITCHING_SO        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                       
##   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    
```

```
attach(moneyball)
sd(TARGET_WINS)
```

```
## [1] 15.75215
```

```
sd(BATTING_H)
```

```
## [1] 144.5912
```

```
sd(BATTING_2B)
```

```
## [1] 46.80141
```

```
sd(BATTING_3B)
```

```
## [1] 27.93856
```

```
sd(BATTING_HR)
```

```
## [1] 60.54687
```

```
sd(BATTING_BB)
```

```
## [1] 122.6709
```

```
sd(BATTING_SO,na.rm = TRUE)
```

```
## [1] 248.5264
```

```
sd(BASERUN_SB,na.rm = TRUE)
```

```
## [1] 87.79117
```

```
sd(BASERUN_CS,na.rm = TRUE)
```

```
## [1] 22.95634
```

```
sd(BATTING_HBP,na.rm = TRUE)
```

```
## [1] 12.96712
```

```
sd(PITCHING_H)
```

```
## [1] 1406.843
```

```
sd(PITCHING_HR)
```

```
## [1] 61.29875
```

```
sd(PITCHING_BB)
```

```
## [1] 166.3574
```

```
sd(PITCHING_SO,na.rm = TRUE)
```

```
## [1] 553.085
```

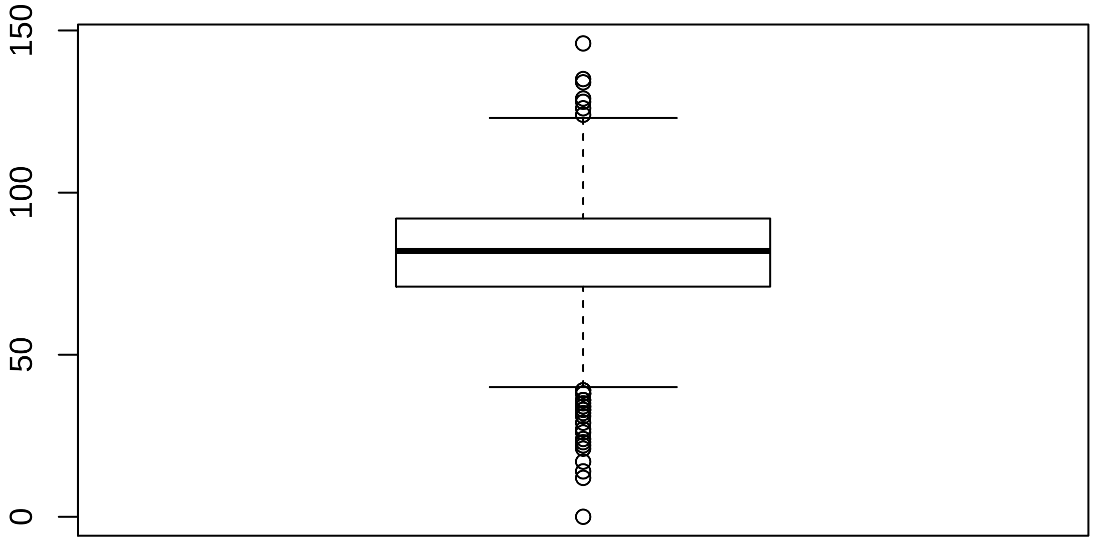
```
sd(FIELDING_E)
```

```
## [1] 227.771
```

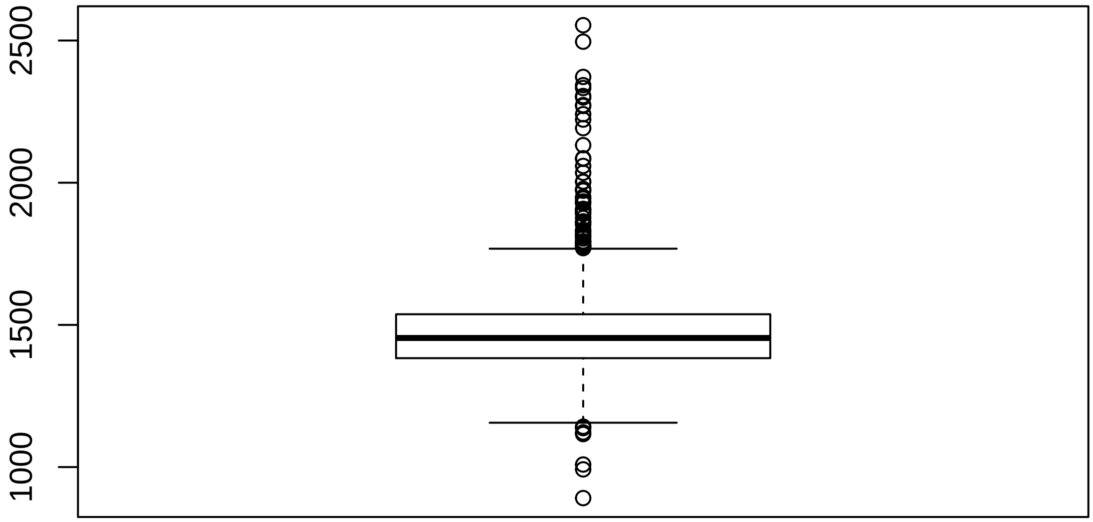
```
sd(FIELDING_DP, na.rm =TRUE)
```

```
## [1] 26.22639
```

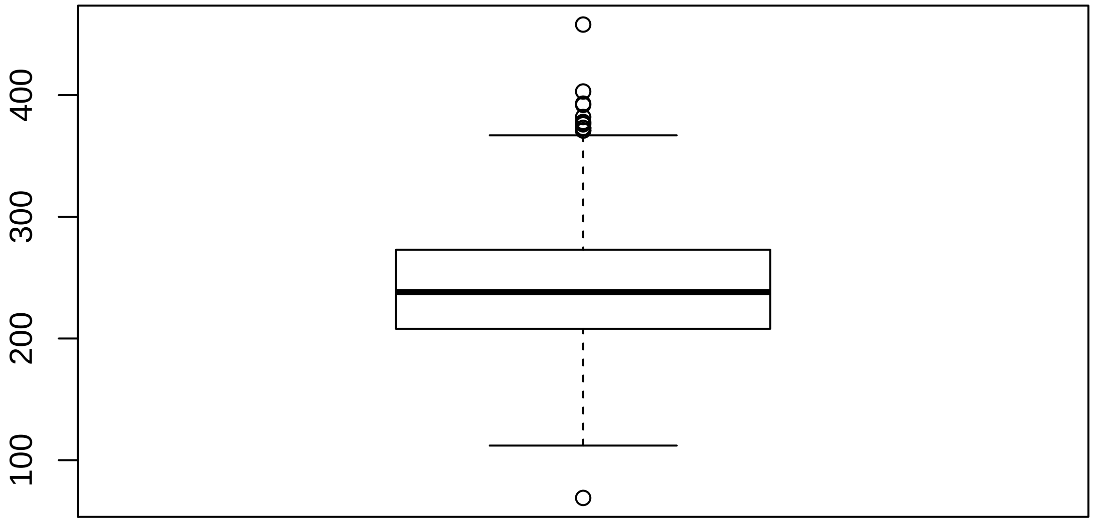
```
boxplot(TARGET_WINS)
```



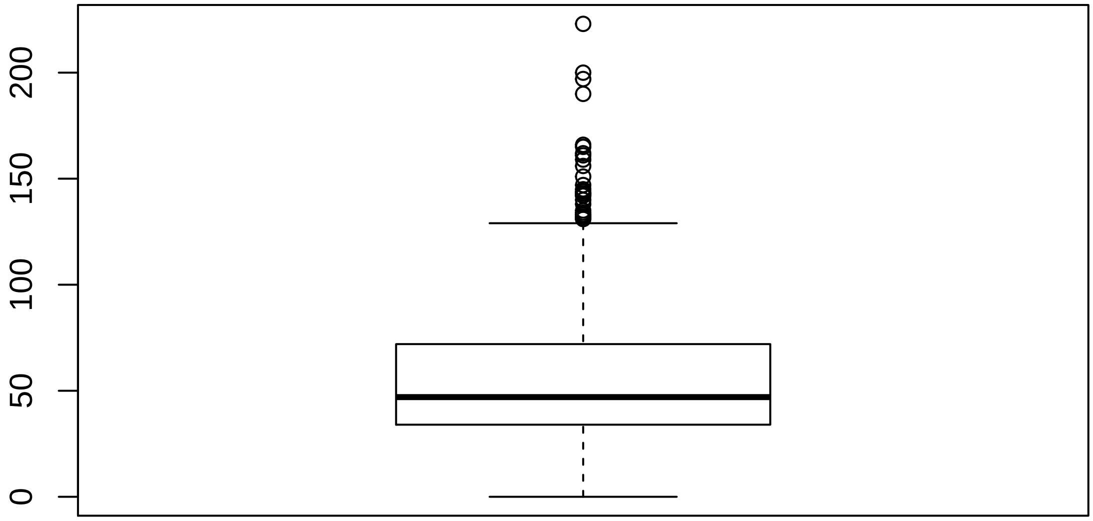
```
boxplot(BATTING_H)
```



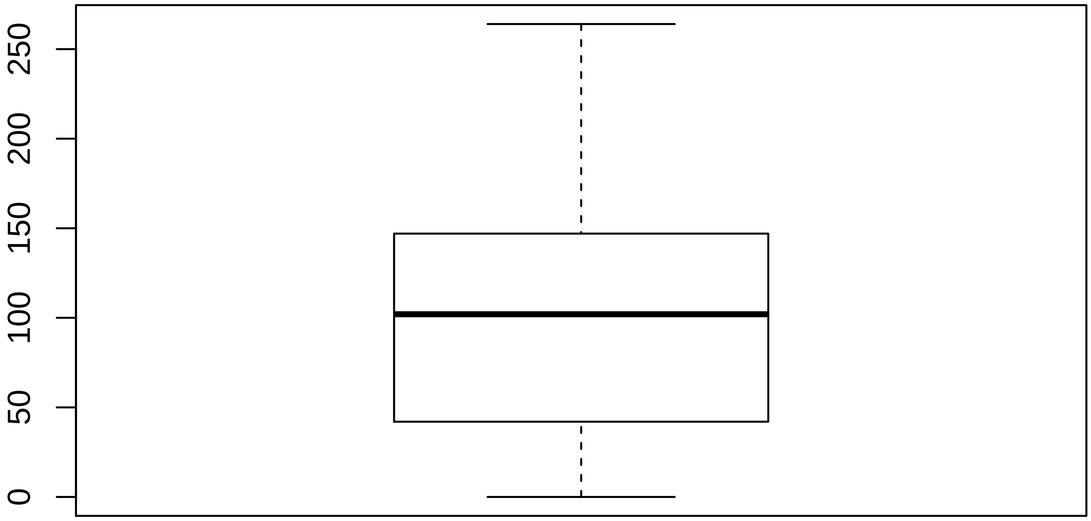
```r
boxplot(BATTING_2B)
```



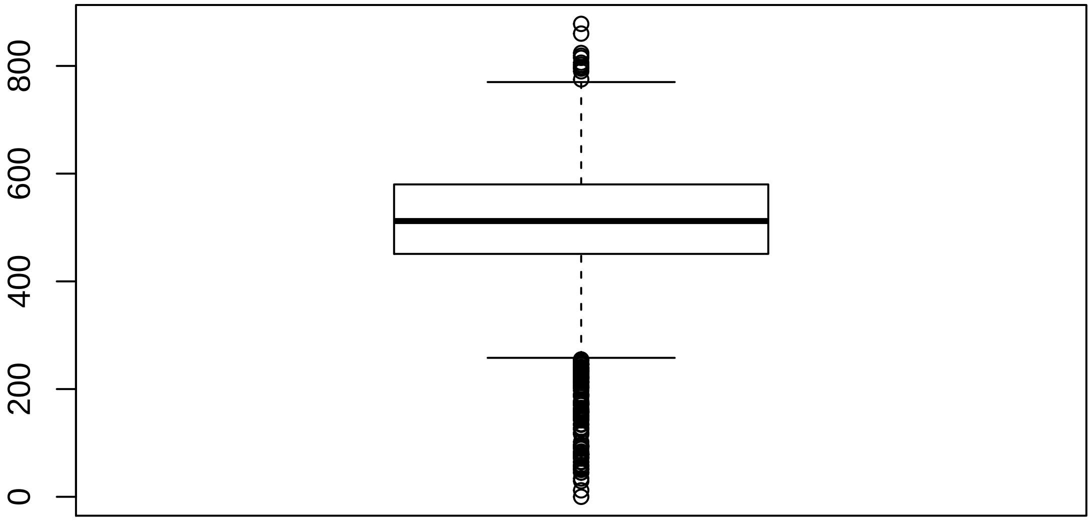
```r
boxplot(BATTING_3B)
```



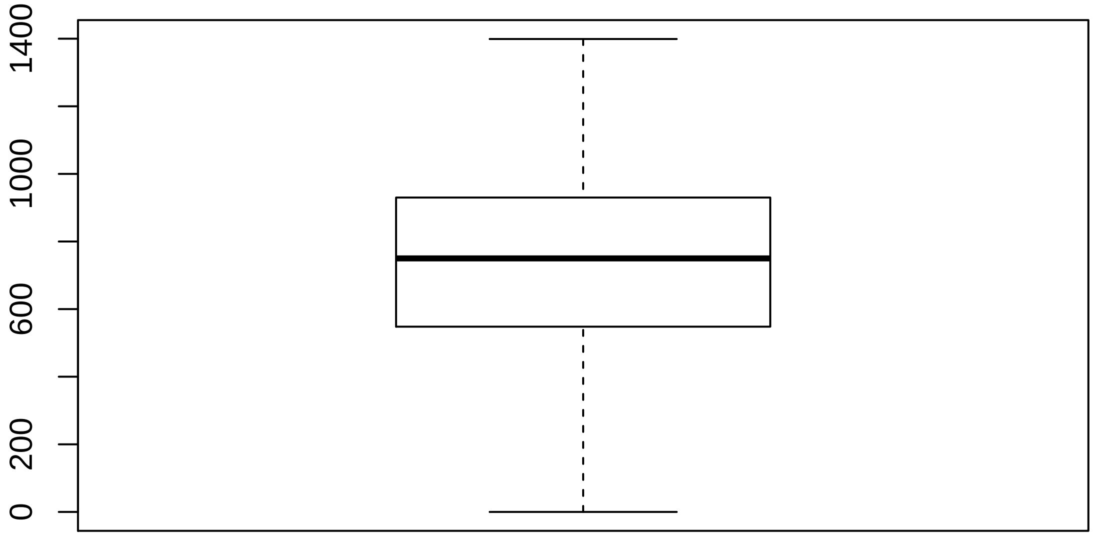
```
boxplot(BATTING_HR)
```



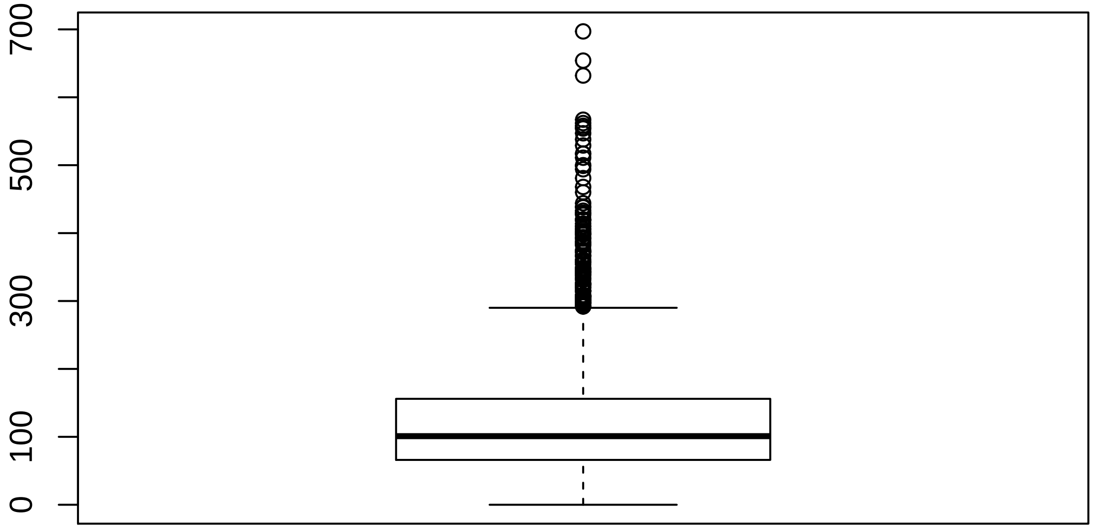
```
boxplot(BATTING_BB)
```



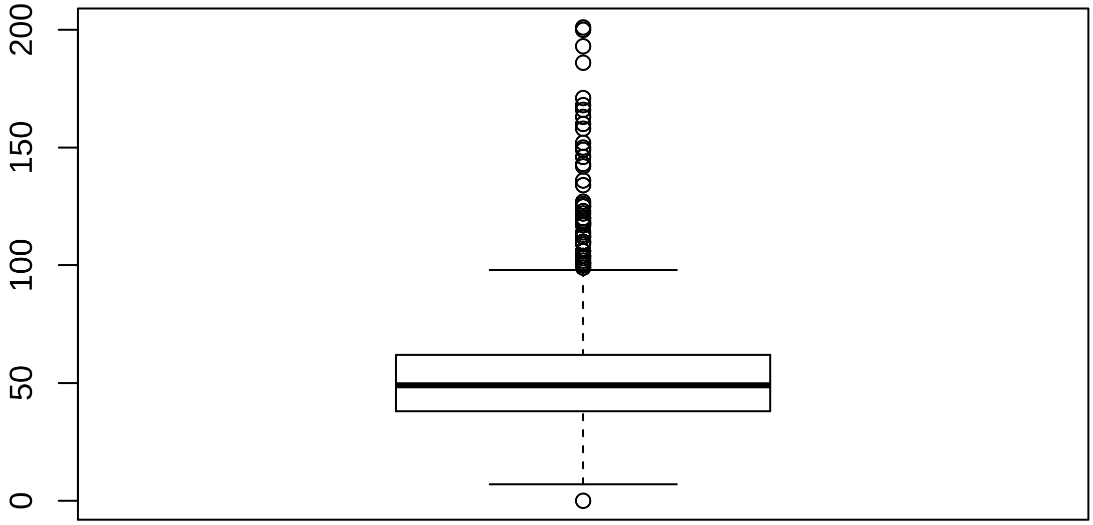
```
boxplot(BATTING_SO,na.rm = TRUE)
```



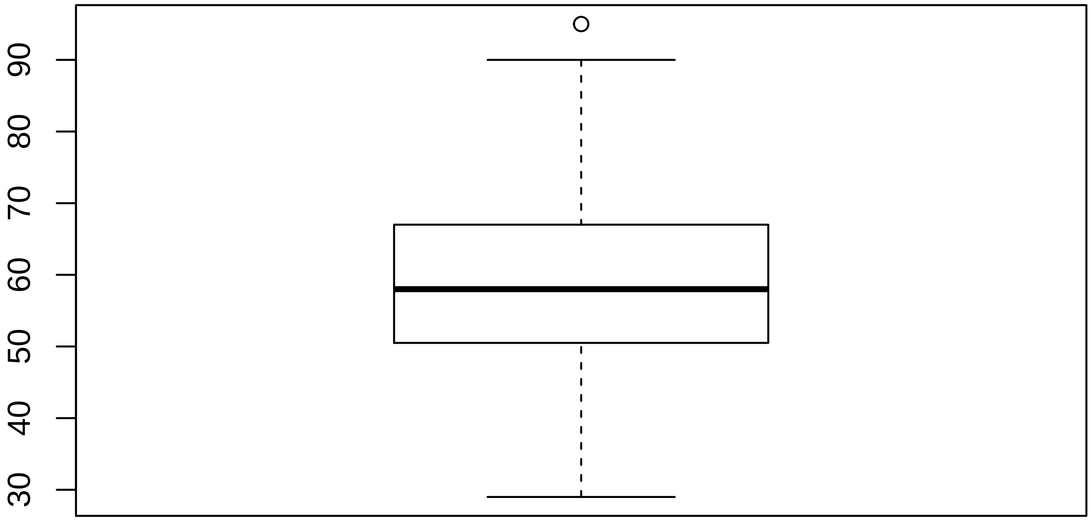
```
boxplot(BASERUN_SB,na.rm = TRUE)
```



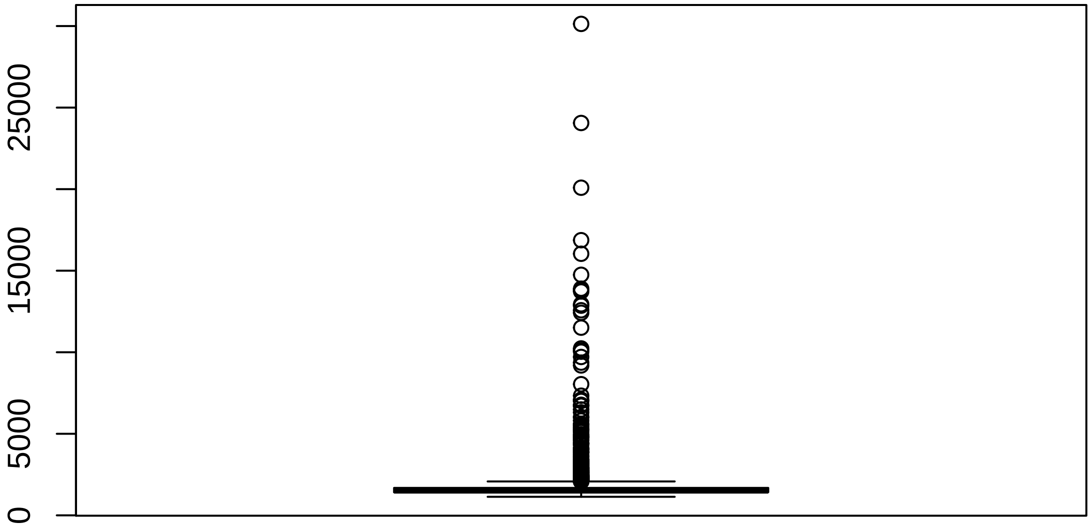
```
boxplot(BASERUN_CS,na.rm = TRUE)
```



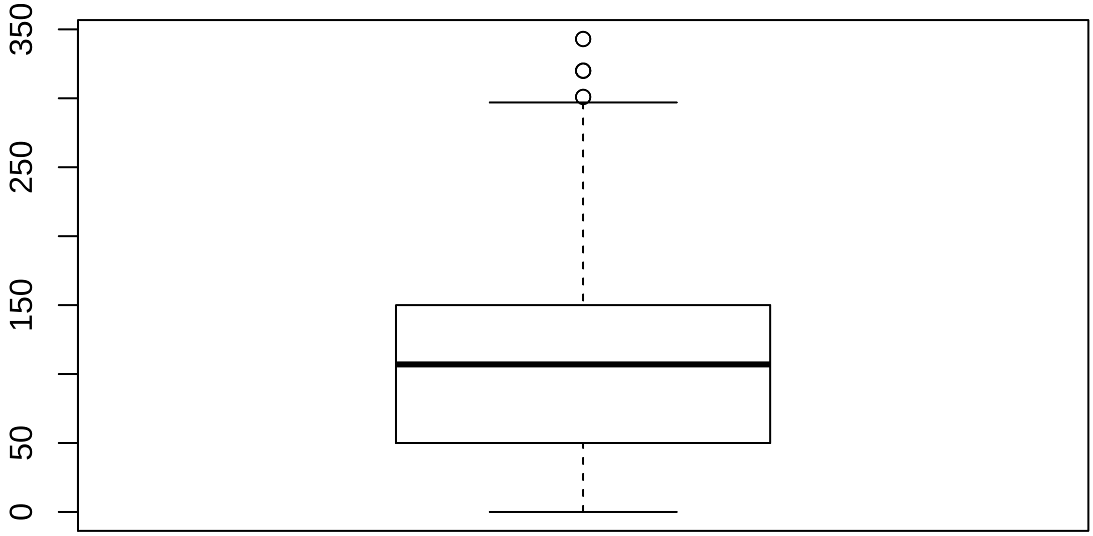
```
boxplot(BATTING_HBP,na.rm = TRUE)
```



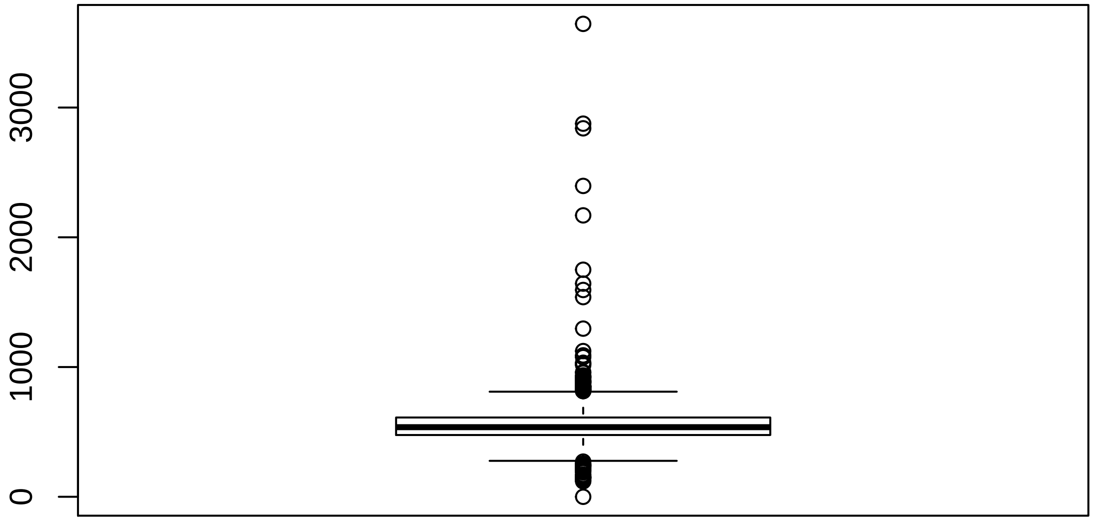
```
boxplot(PITCHING_H)
```



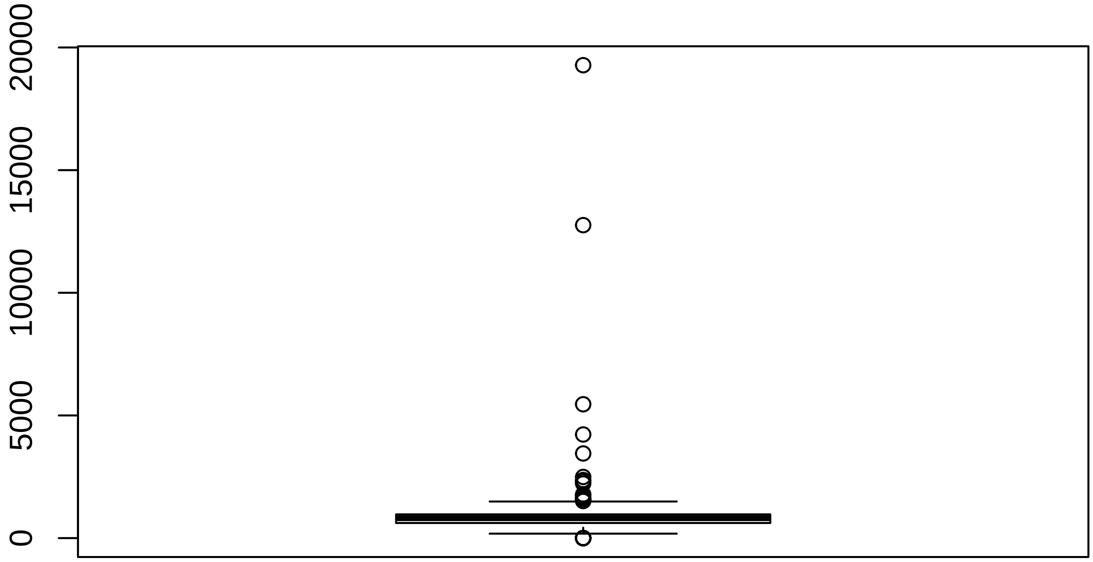
```
boxplot(PITCHING_HR)
```



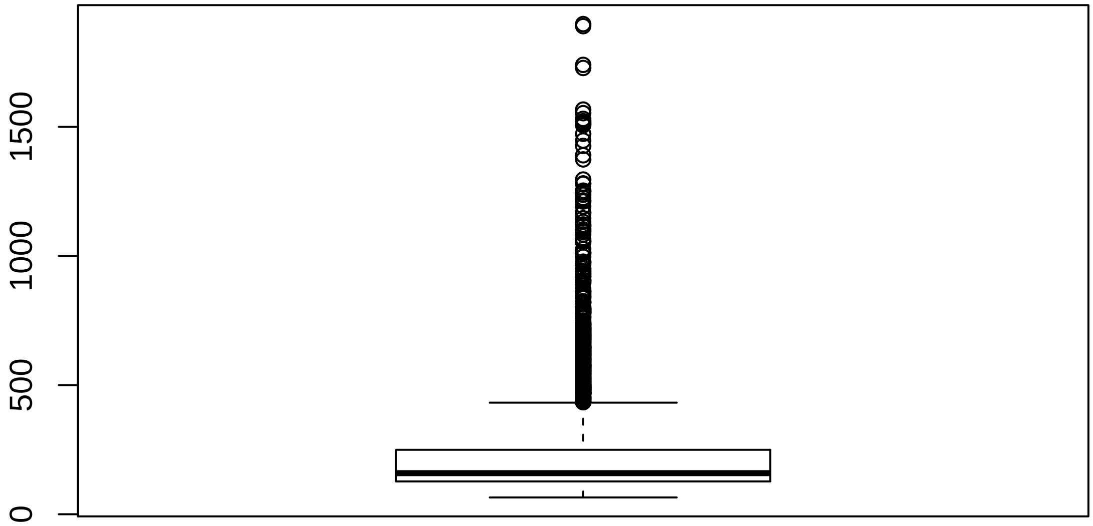
```
boxplot(PITCHING_BB)
```



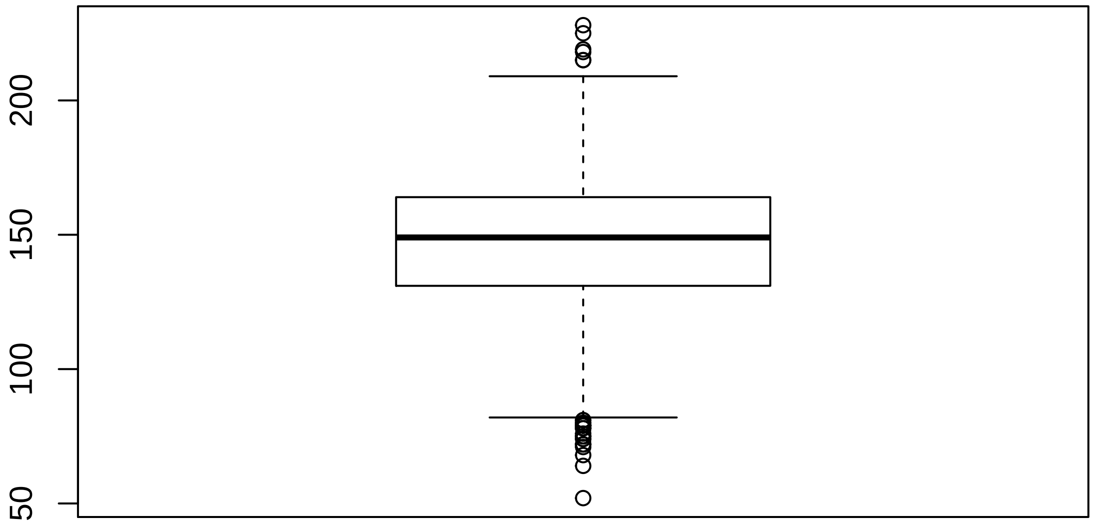
```
boxplot(PITCHING_SO,na.rm = TRUE)
```



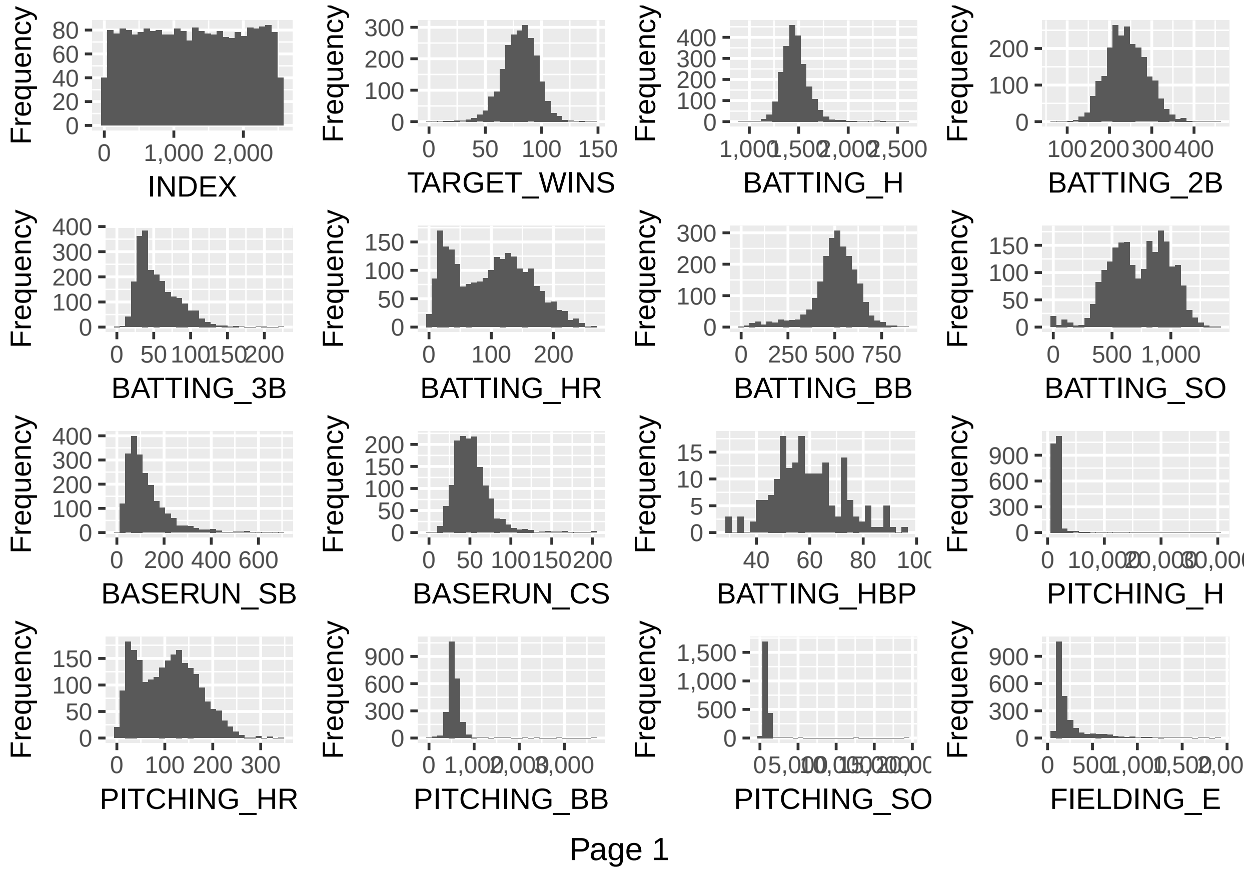
```
boxplot(FIELDING_E)
```

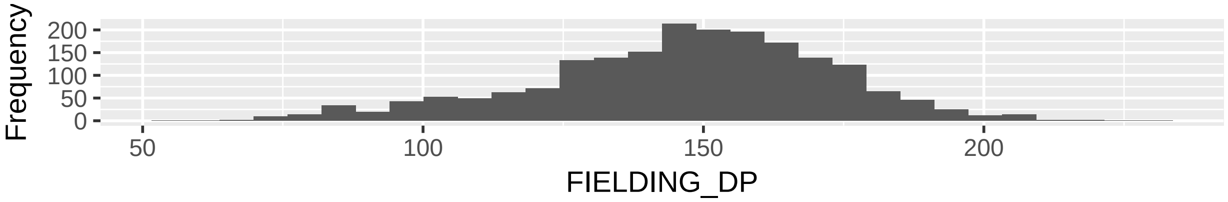


```
boxplot(FIELDING_DP, na.rm =TRUE)
```



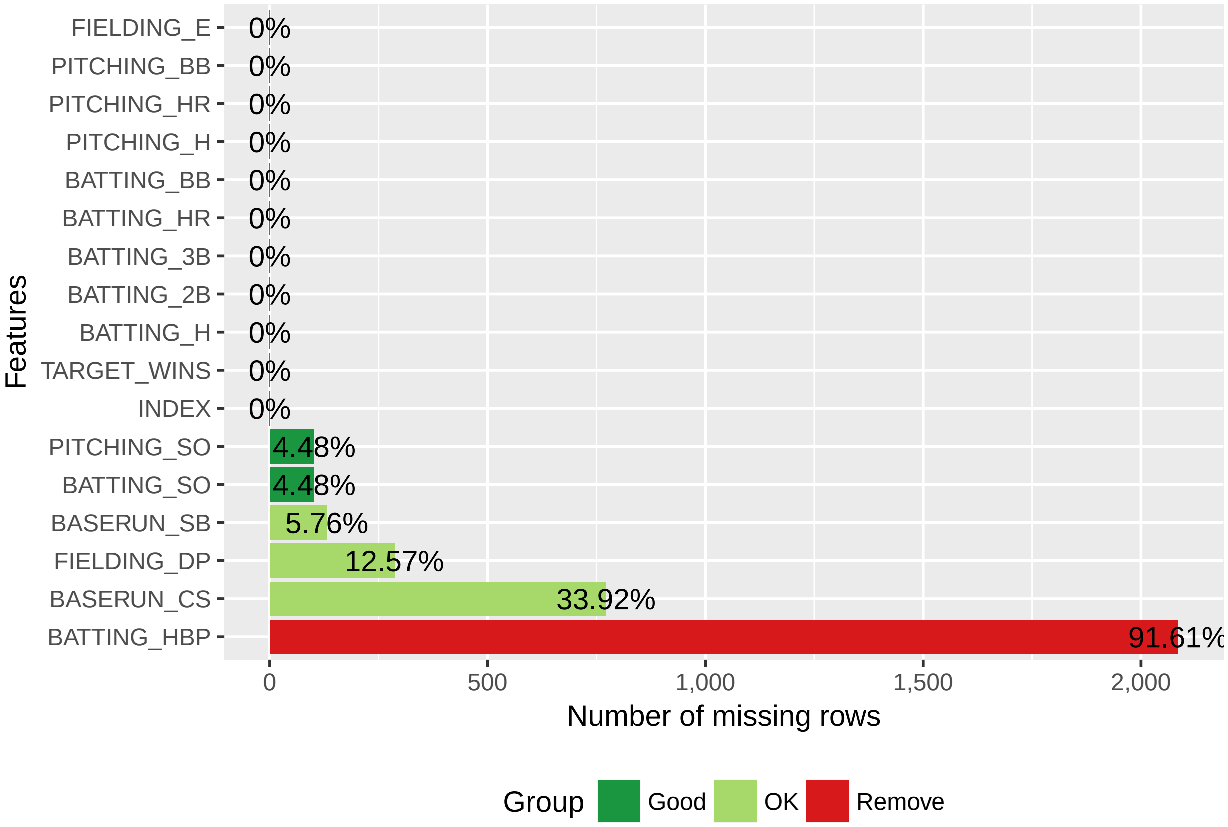
```
plot_histogram(moneyball)
```



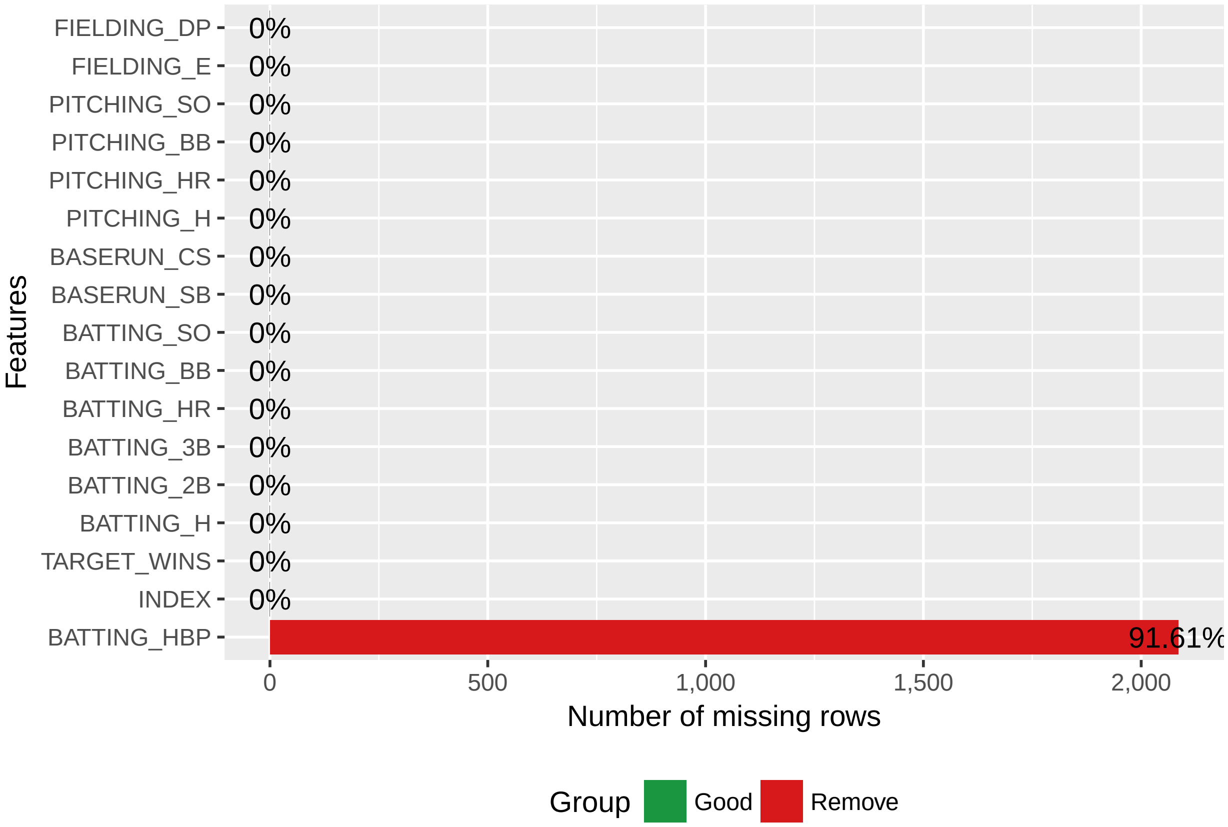


Page 2

```
plot_missing(moneyball)
```



```
moneyball[is.na(moneyball$BATTING_SO),"BATTING_SO"] <- median(na.omit(moneyball$BATTING_SO))
moneyball[is.na(moneyball$PITCHING_SO),"PITCHING_SO"] <- median(na.omit(moneyball$PITCHING_SO))
moneyball[is.na(moneyball$BASERUN_SB),"BASERUN_SB"] <- median(na.omit(moneyball$BASERUN_SB))
moneyball[is.na(moneyball$FIELDING_DP),"FIELDING_DP"] <- median(na.omit(moneyball$FIELDING_DP))
moneyball[is.na(moneyball$BASERUN_CS),"BASERUN_CS"] <- median(na.omit(moneyball$BASERUN_CS))
plot_missing(moneyball)
```



```
# By observing the data, I see that TEAM_BATTING_H is the sum of 1B, 2B, 3B, HR, to find 1B and analyze
attach(moneyball)
```

```
## The following objects are masked from moneyball (pos = 3):
##
##     BASERUN_CS, BASERUN_SB, BATTING_2B, BATTING_3B, BATTING_BB,
##     BATTING_H, BATTING_HBP, BATTING_HR, BATTING_SO, FIELDING_DP,
##     FIELDING_E, INDEX, PITCHING_BB, PITCHING_H, PITCHING_HR,
##     PITCHING_SO, TARGET_WINS
```

```
moneyball<- moneyball %>% mutate(BATTING_1B=BATTING_H-BATTING_2B-BATTING_3B,Total_batting=1*BATTING_1B+
kable(head(moneyball))
```

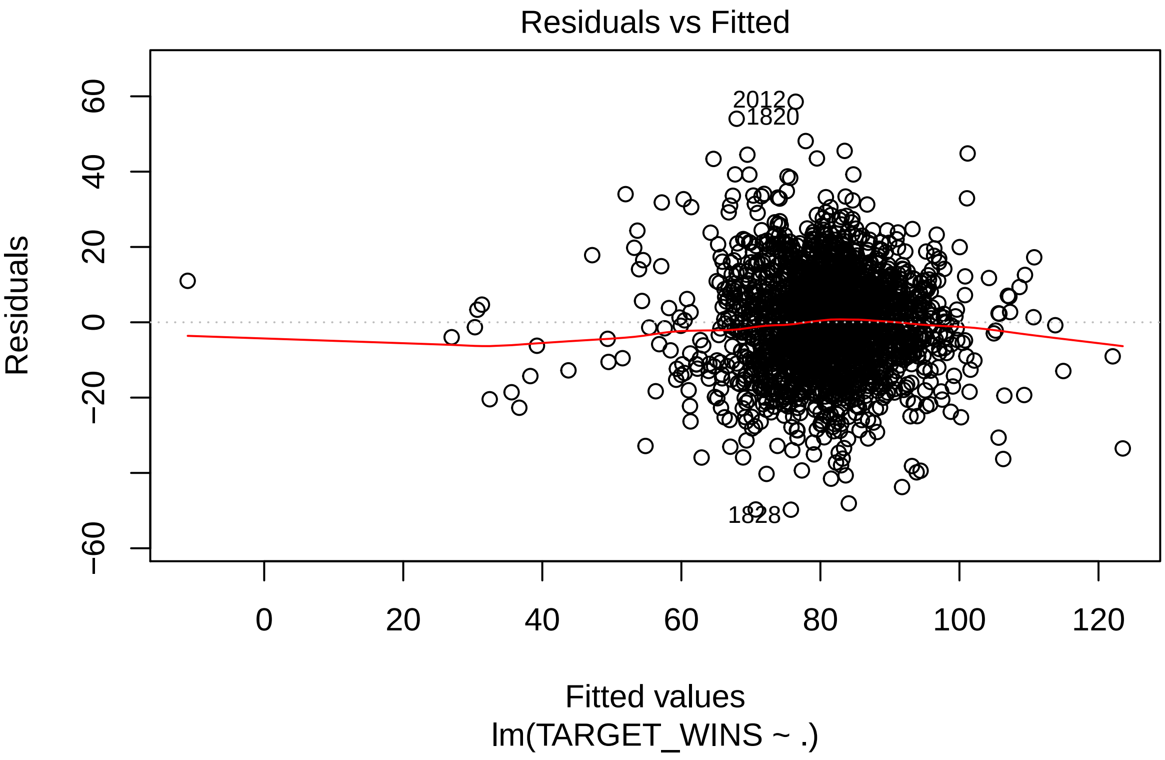
TARGET_WINS	BATTING_2B	BATTING_3B	BATTING_HR	BATTING_BB	BATTING_SO	BASERUN_
39	194	39	13	143	842	
70	219	22	190	685	1075	
86	232	35	137	602	917	
70	209	38	96	451	922	
82	186	27	102	472	920	
75	200	36	92	443	973	

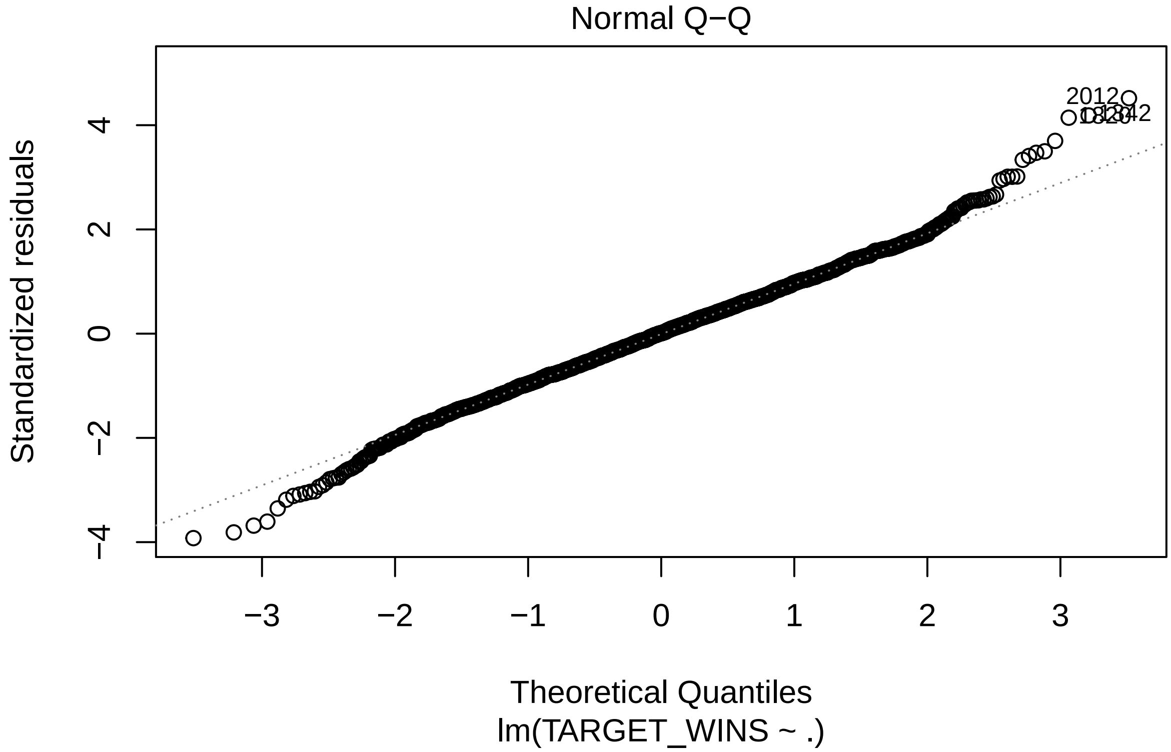
```
cor(moneyball)
```

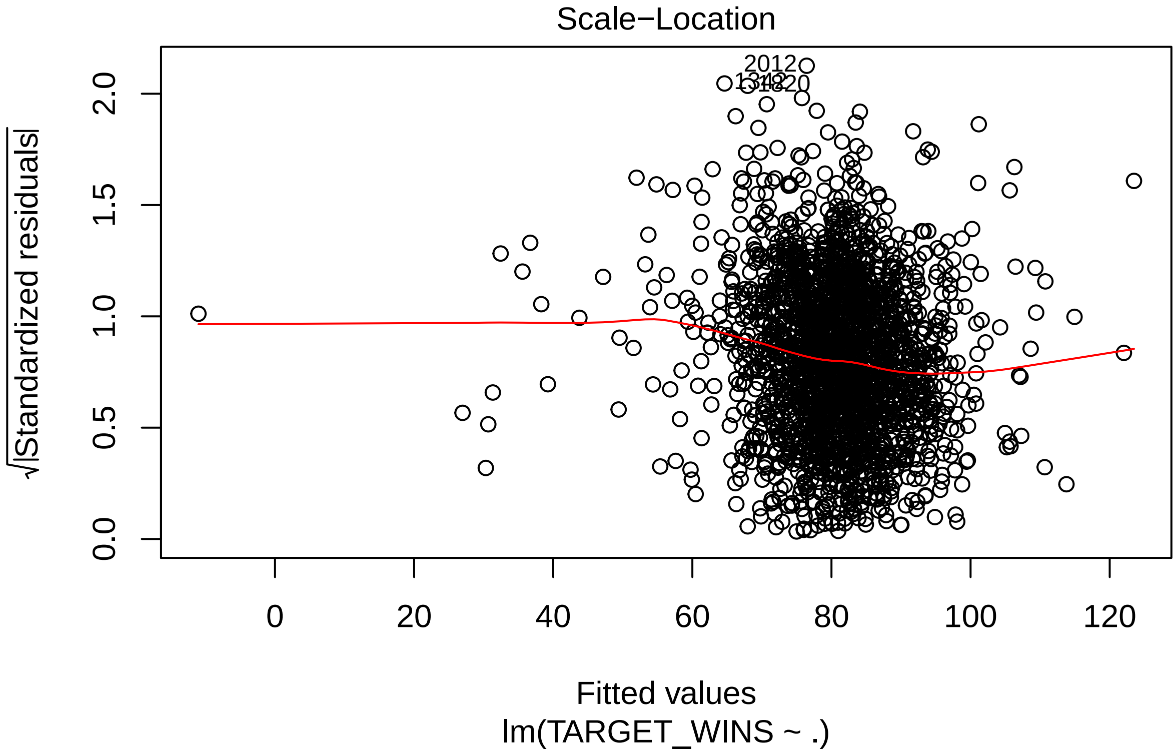
```
##              TARGET_WINS  BATTING_2B   BATTING_3B BATTING_HR  BATTING_BB
## TARGET_WINS   1.00000000  0.28910365  0.142608411  0.1761532  0.23255986
## BATTING_2B    0.28910365  1.00000000 -0.107305824  0.4353973  0.25572610
## BATTING_3B    0.14260841 -0.10730582  1.000000000 -0.6355669 -0.28723584
## BATTING_HR     0.17615320  0.43539729 -0.635566946  1.0000000  0.51373481
## BATTING_BB     0.23255986  0.25572610 -0.287235841  0.5137348  1.00000000
## BATTING_SO    -0.03058135  0.15173438 -0.655709613  0.6930076  0.37148892
## BASERUN_SB     0.12361087 -0.18340432  0.485740156 -0.4068891 -0.04268402
## BASERUN_CS     0.01595982 -0.04584955  0.136181182 -0.2254587 -0.04581766
## PITCHING_H    -0.10993705  0.02369219  0.194879411 -0.2501455 -0.44977762
## PITCHING_HR    0.18901373  0.45455082 -0.567836679  0.9693714  0.45955207
## PITCHING_BB    0.12417454  0.17805420 -0.002224148  0.1369276  0.48936126
## PITCHING_SO   -0.07579967  0.06213042 -0.254238104  0.1774182 -0.02017989
## FIELDING_E    -0.17648476 -0.23515099  0.509778447 -0.5873391 -0.65597081
## FIELDING_DP   -0.03008630  0.25696798 -0.227771884  0.3916524  0.32963974
## BATTING_1B     0.34579395  0.33580405  0.347822719 -0.0318712 -0.12886347
## Total_batting  0.39892151  0.75439415 -0.136638042  0.7493183  0.36408258
##                BATTING_SO  BASERUN_SB  BASERUN_CS  PITCHING_H PITCHING_HR
## TARGET_WINS   -0.03058135  0.12361087  0.01595982 -0.10993705  0.18901373
## BATTING_2B     0.15173438 -0.18340432 -0.04584955  0.02369219  0.45455082
## BATTING_3B    -0.65570961  0.48574016  0.13618118  0.19487941 -0.56783668
## BATTING_HR     0.69300765 -0.40688907 -0.22545867 -0.25014548  0.96937140
## BATTING_BB     0.37148892 -0.04268402 -0.04581766 -0.44977762  0.45955207
## BATTING_SO     1.00000000 -0.21178758 -0.10250193 -0.37571553  0.63286033
## BASERUN_SB    -0.21178758  1.00000000  0.23324171  0.03957227 -0.38005624
## BASERUN_CS    -0.10250193  0.23324171  1.00000000 -0.05259183 -0.22818525
## PITCHING_H    -0.37571553  0.03957227 -0.05259183  1.00000000 -0.14161276
## PITCHING_HR    0.63286033 -0.38005624 -0.22818525 -0.14161276  1.00000000
## PITCHING_BB    0.03498809  0.12928969 -0.04722893  0.32067616  0.22193750
## PITCHING_SO    0.41618159 -0.06424741 -0.05653800  0.26693587  0.19691491
## FIELDING_E    -0.58259305  0.32615276 -0.02917821  0.66775901 -0.49314447
## FIELDING_DP    0.11089804 -0.27023400 -0.10200214 -0.04464784  0.38959550
## BATTING_1B    -0.48464372  0.09474682 -0.01375594  0.33253091  0.04579447
## Total_batting  0.24141254 -0.21340675 -0.16245457 -0.01596413  0.77829405
##                PITCHING_BB  PITCHING_SO  FIELDING_E  FIELDING_DP
## TARGET_WINS    0.124174536 -0.075799674 -0.17648476 -0.030086302
## BATTING_2B     0.178054204  0.062130422 -0.23515099  0.256967975
## BATTING_3B    -0.002224148 -0.254238104  0.50977845 -0.227771884
## BATTING_HR     0.136927564  0.177418187 -0.58733910  0.391652434
## BATTING_BB     0.489361263 -0.020179893 -0.65597081  0.329639737
## BATTING_SO     0.034988093  0.416181592 -0.58259305  0.110898035
## BASERUN_SB     0.129289686 -0.064247407  0.32615276 -0.270234003
## BASERUN_CS    -0.047228927 -0.056538002 -0.02917821 -0.102002137
## PITCHING_H     0.320676162  0.266935871  0.66775901 -0.044647837
## PITCHING_HR    0.221937505  0.196914911 -0.49314447  0.389595503
## PITCHING_BB    1.000000000  0.482172000 -0.02283756  0.192348657
## PITCHING_SO    0.482172000  1.000000000 -0.02332278  0.009552324
## FIELDING_E    -0.022837561 -0.023322782  1.00000000 -0.227394807
## FIELDING_DP    0.192348657  0.009552324 -0.22739481  1.000000000
## BATTING_1B     0.047792487 -0.279280625  0.31333793  0.110655544
## Total_batting  0.182240384 -0.012524535 -0.28654467  0.371830404
##                BATTING_1B Total_batting
## TARGET_WINS    0.34579395    0.39892151
## BATTING_2B     0.33580405    0.75439415
## BATTING_3B     0.34782272   -0.13663804
## BATTING_HR    -0.03187120    0.74931833
## BATTING_BB    -0.12886347    0.36408258
## BATTING_SO    -0.48464372    0.24141254
## BASERUN_SB     0.09474682   -0.21340675
## BASERUN_CS    -0.01375594   -0.16245457
## PITCHING_H     0.33253091   -0.01596413
## PITCHING_HR    0.04579447    0.77829405
## PITCHING_BB    0.04779249    0.18224038
## PITCHING_SO   -0.27928062   -0.01252454
## FIELDING_E     0.31333793   -0.28654467
## FIELDING_DP    0.11065554    0.37183040
## BATTING_1B     1.00000000    0.54607257
## Total_batting  0.54607257    1.00000000
```

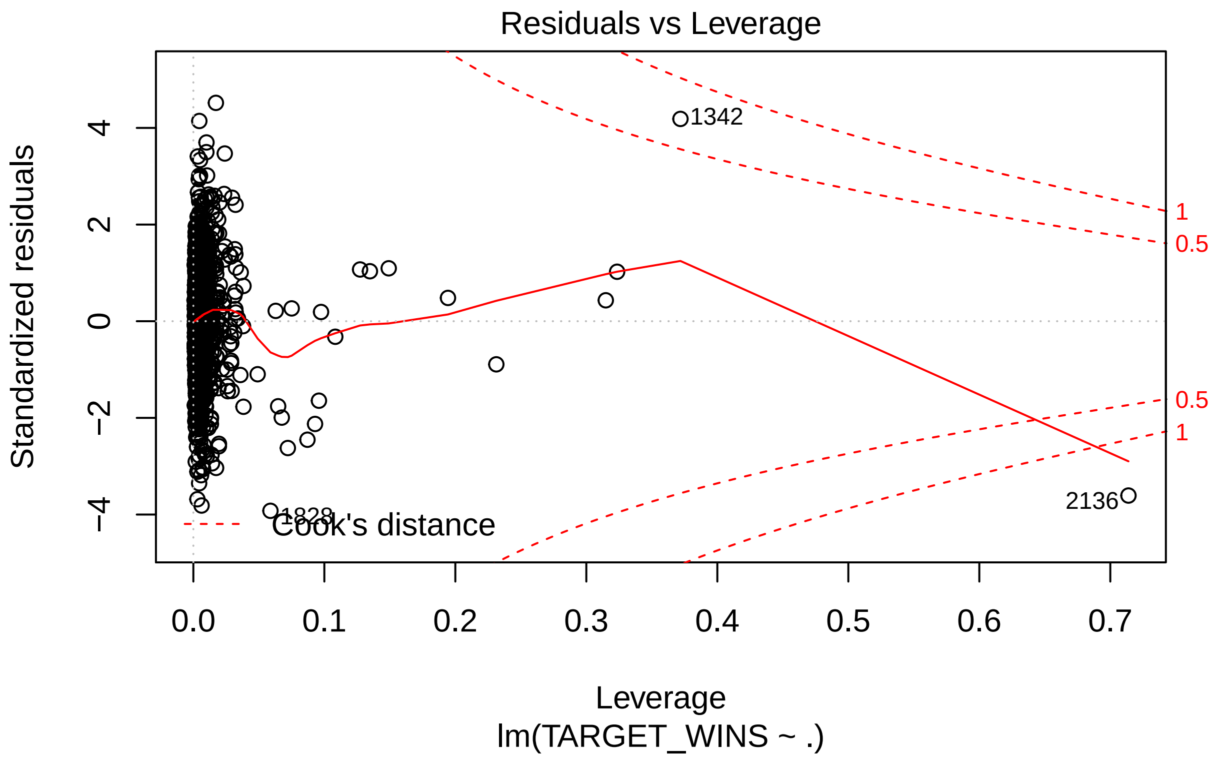
Model 1: Backwards Selection

```
m1 <-lm(TARGET_WINS~.,data=moneyball)
plot(m1)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(TARGET_WINS ~ .)
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
2012
1820
1342


Scale−Location
√|Standardized residuals|
Fitted values
lm(TARGET_WINS ~ .)
2.0
1.5
1.0
0.5
0.0
0
20
40
60
80
100
120
2012
1342
1820



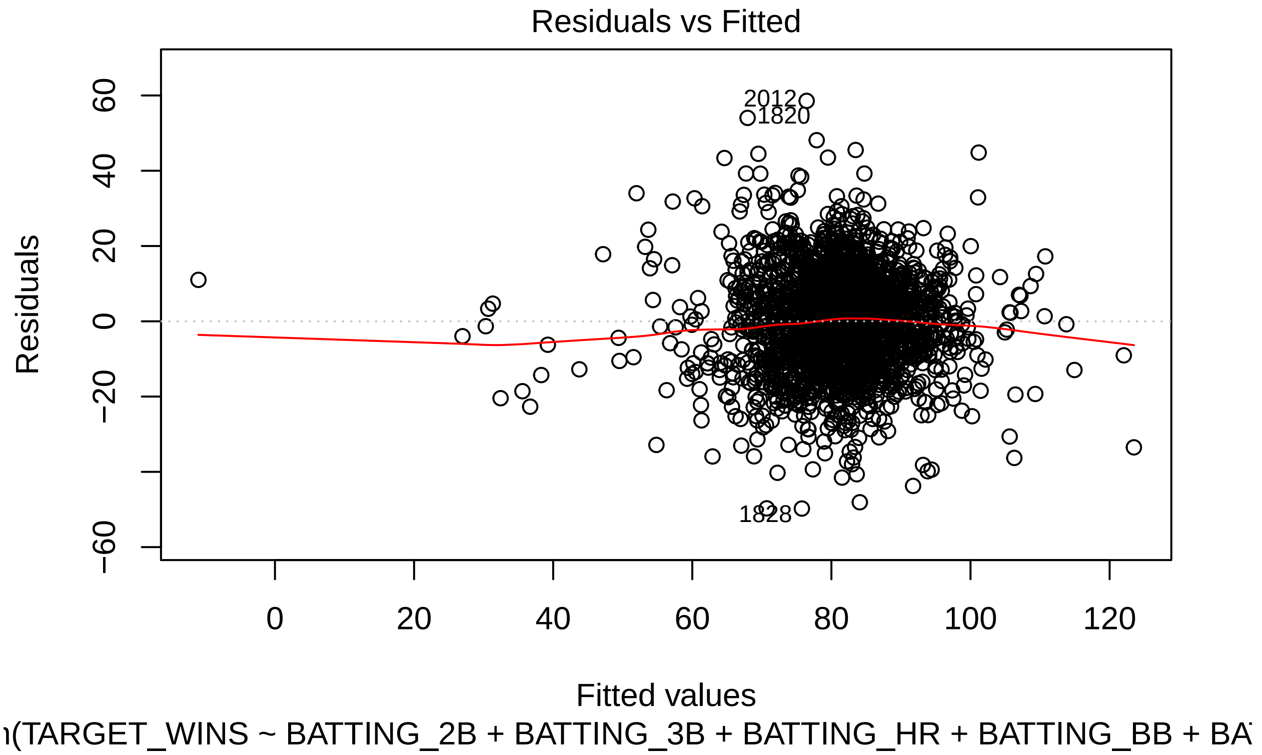
```r
summary(m1)
```

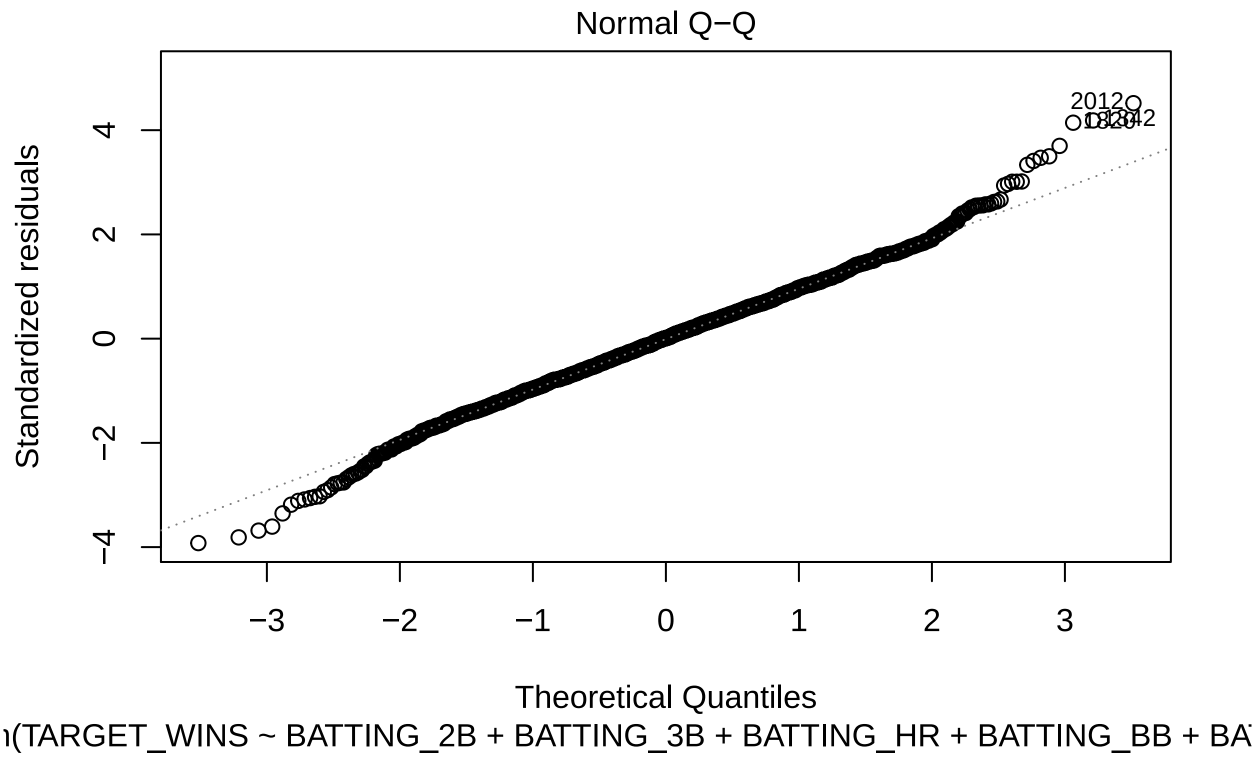
```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = moneyball)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.753  -8.626   0.120   8.395  58.561
##
## Coefficients: (1 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   23.6421579  5.3902272   4.386 1.21e-05 ***
## BATTING_2B     0.0279578  0.0073363   3.811 0.000142 ***
## BATTING_3B     0.1133940  0.0159335   7.117 1.48e-12 ***
## BATTING_HR     0.0527325  0.0274915   1.918 0.055219 .
## BATTING_BB     0.0104483  0.0058377   1.790 0.073621 .
## BATTING_SO    -0.0084323  0.0025461  -3.312 0.000941 ***
## BASERUN_SB     0.0254236  0.0043565   5.836 6.12e-09 ***
## BASERUN_CS    -0.0110027  0.0157842  -0.697 0.485829
## PITCHING_H    -0.0008456  0.0003674  -2.302 0.021444 *
## PITCHING_HR    0.0129626  0.0243894   0.531 0.595135
## PITCHING_BB    0.0007798  0.0041571   0.188 0.851231
## PITCHING_SO    0.0028156  0.0009219   3.054 0.002284 **
## FIELDING_E    -0.0195325  0.0024609  -7.937 3.23e-15 ***
## FIELDING_DP   -0.1217801  0.0129421  -9.410  < 2e-16 ***
## BATTING_1B     0.0489152  0.0036949  13.239  < 2e-16 ***
## Total_batting         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

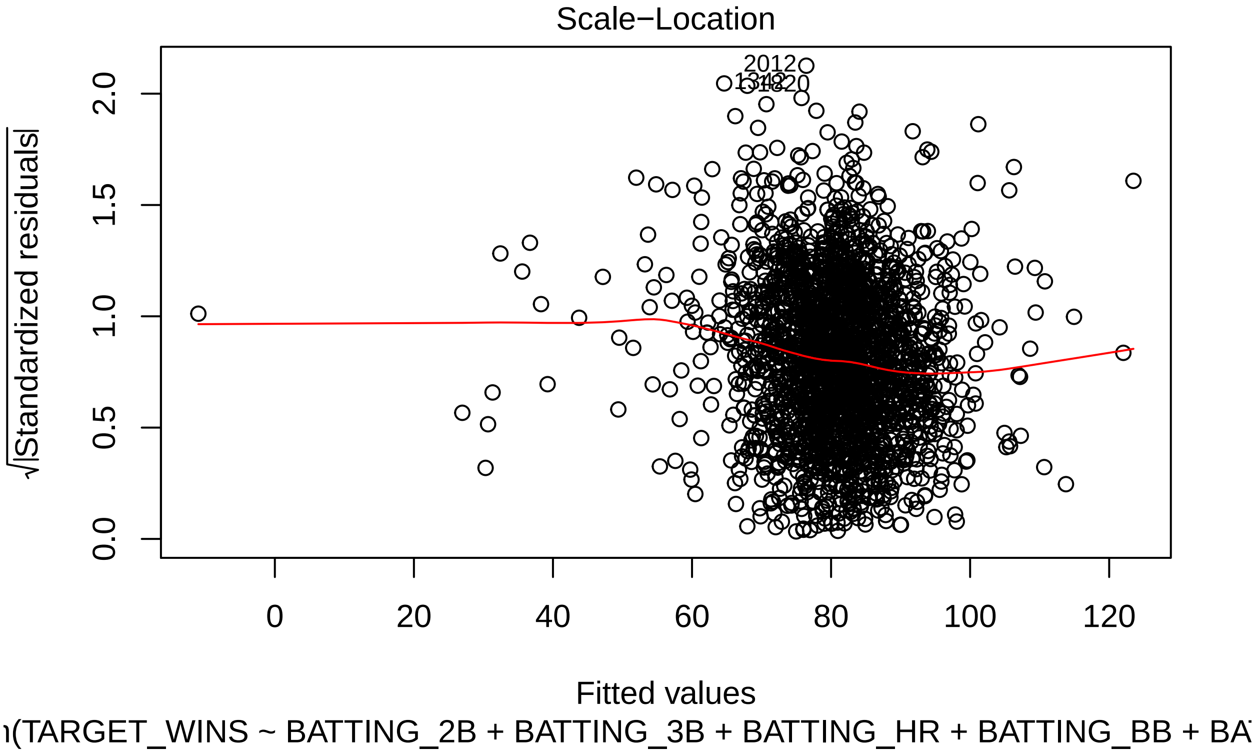
```
##
## Residual standard error: 13.07 on 2261 degrees of freedom
## Multiple R-squared:  0.3154, Adjusted R-squared:  0.3111
## F-statistic:  74.4 on 14 and 2261 DF,  p-value: < 2.2e-16
```

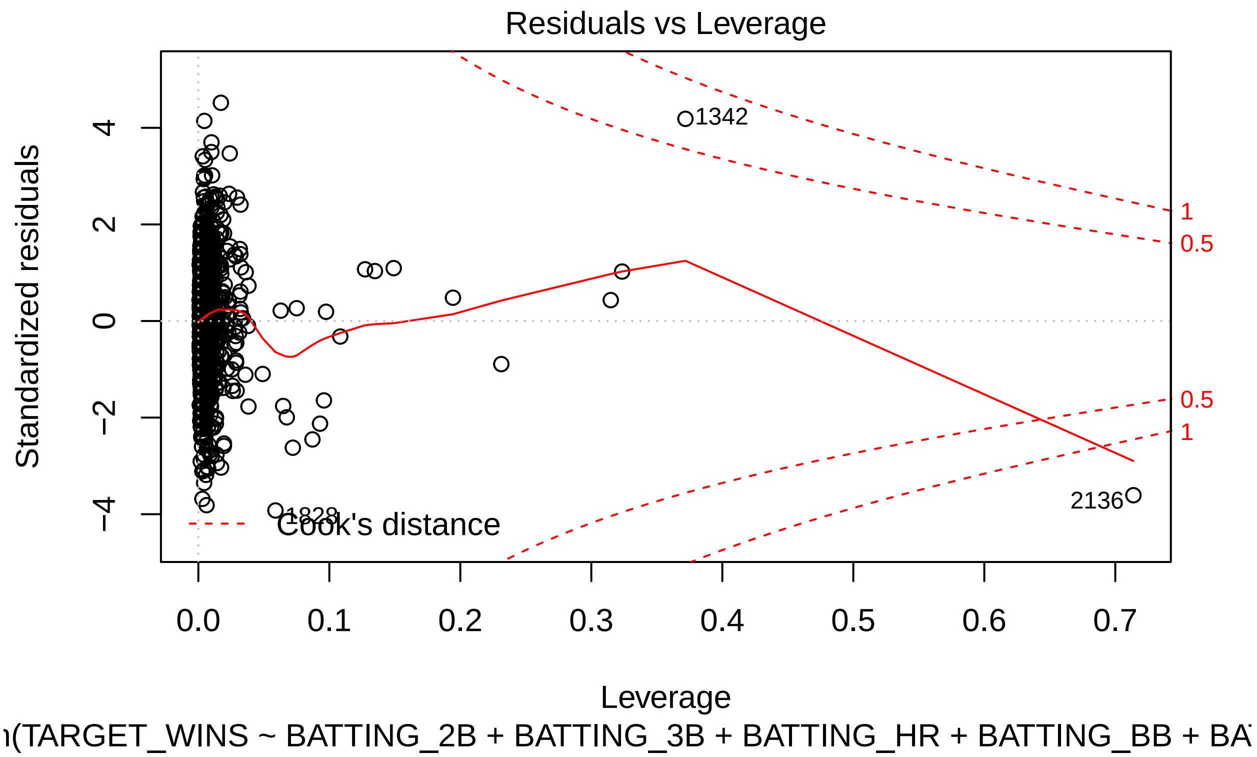
Model 2 Get rid of "Total_Batting" for now

```
m2 <- update(m1,~.-Total_batting)
plot(m2)
```




Normal Q−Q
Standardized residuals
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
2012
1820
1342
Theoretical Quantiles
ı(TARGET_WINS ~ BATTING_2B + BATTING_3B + BATTING_HR + BATTING_BB + BA


Scale−Location
√|Standardized residuals|
2.0
1.5
1.0
0.5
0.0
0
20
40
60
80
100
120
2012
1342
1820
Fitted values
ı(TARGET_WINS ~ BATTING_2B + BATTING_3B + BATTING_HR + BATTING_BB + BA

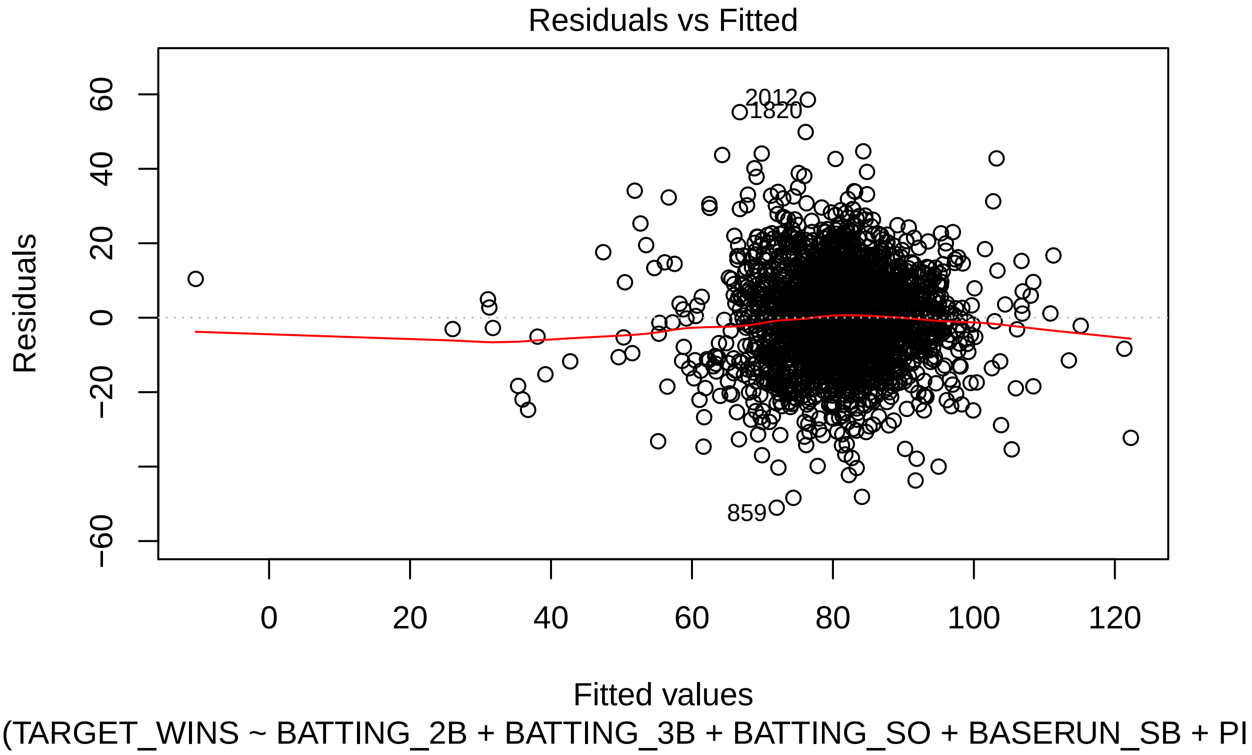


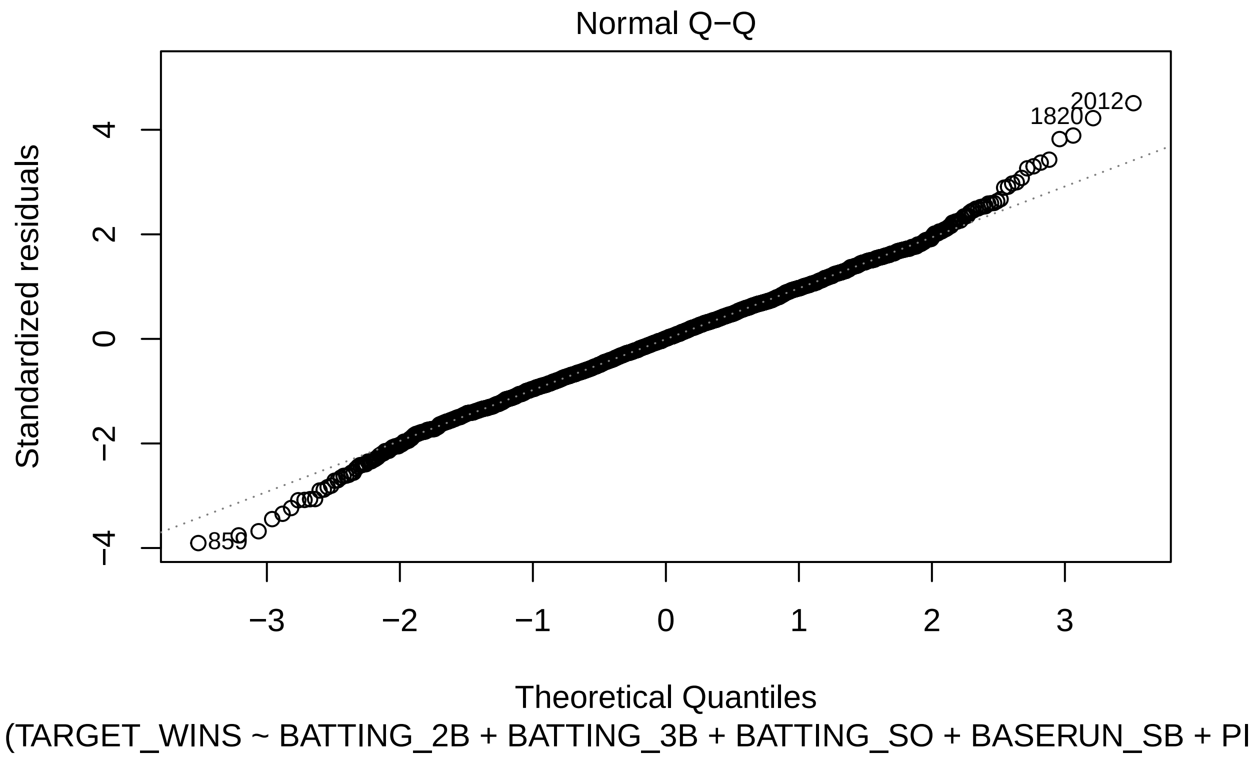
```r
summary(m2)
```

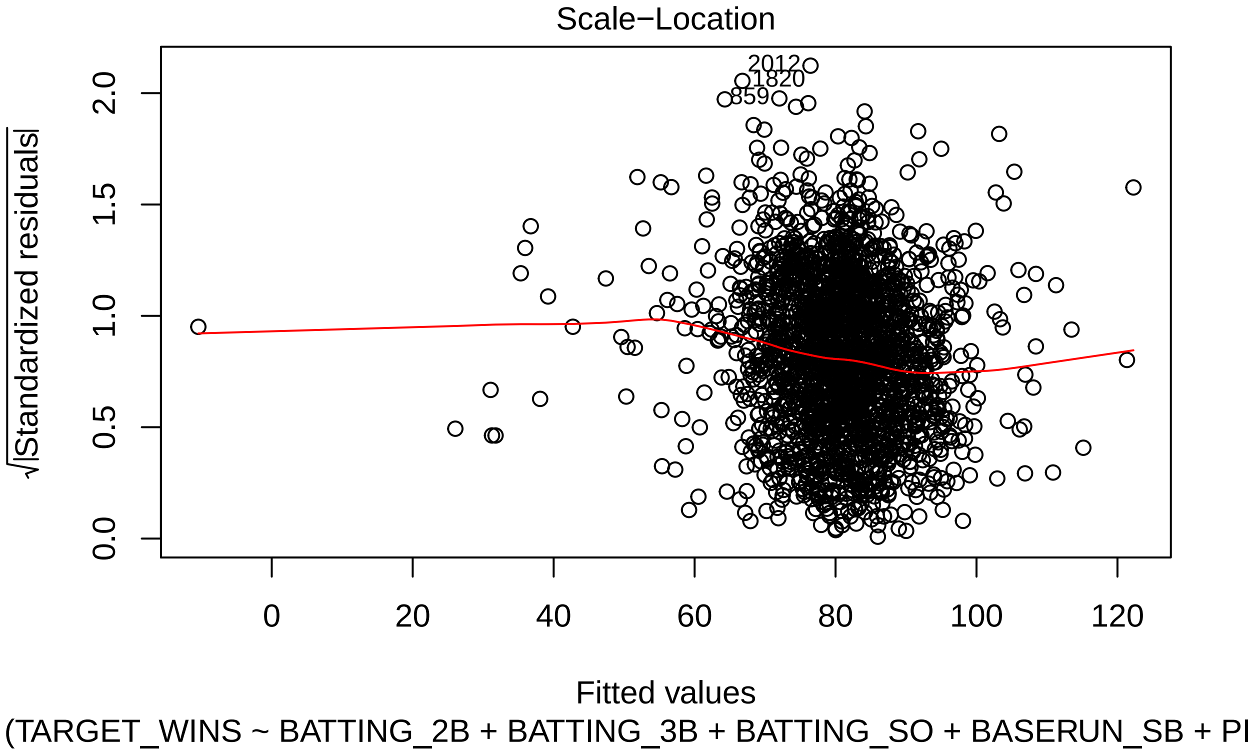
```
##
## Call:
## lm(formula = TARGET_WINS ~ BATTING_2B + BATTING_3B + BATTING_HR +
##     BATTING_BB + BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H +
##     PITCHING_HR + PITCHING_BB + PITCHING_SO + FIELDING_E + FIELDING_DP +
##     BATTING_1B, data = moneyball)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.753  -8.626   0.120   8.395  58.561
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.6421579  5.3902272   4.386 1.21e-05 ***
## BATTING_2B   0.0279578  0.0073363   3.811 0.000142 ***
## BATTING_3B   0.1133940  0.0159335   7.117 1.48e-12 ***
## BATTING_HR   0.0527325  0.0274915   1.918 0.055219 .
## BATTING_BB   0.0104483  0.0058377   1.790 0.073621 .
## BATTING_SO  -0.0084323  0.0025461  -3.312 0.000941 ***
## BASERUN_SB   0.0254236  0.0043565   5.836 6.12e-09 ***
## BASERUN_CS  -0.0110027  0.0157842  -0.697 0.485829
## PITCHING_H  -0.0008456  0.0003674  -2.302 0.021444 *
## PITCHING_HR  0.0129626  0.0243894   0.531 0.595135
## PITCHING_BB  0.0007798  0.0041571   0.188 0.851231
## PITCHING_SO  0.0028156  0.0009219   3.054 0.002284 **
## FIELDING_E  -0.0195325  0.0024609  -7.937 3.23e-15 ***
## FIELDING_DP -0.1217801  0.0129421  -9.410  < 2e-16 ***
## BATTING_1B   0.0489152  0.0036949  13.239  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.07 on 2261 degrees of freedom
## Multiple R-squared:  0.3154, Adjusted R-squared:  0.3111
## F-statistic:  74.4 on 14 and 2261 DF,  p-value: < 2.2e-16
```

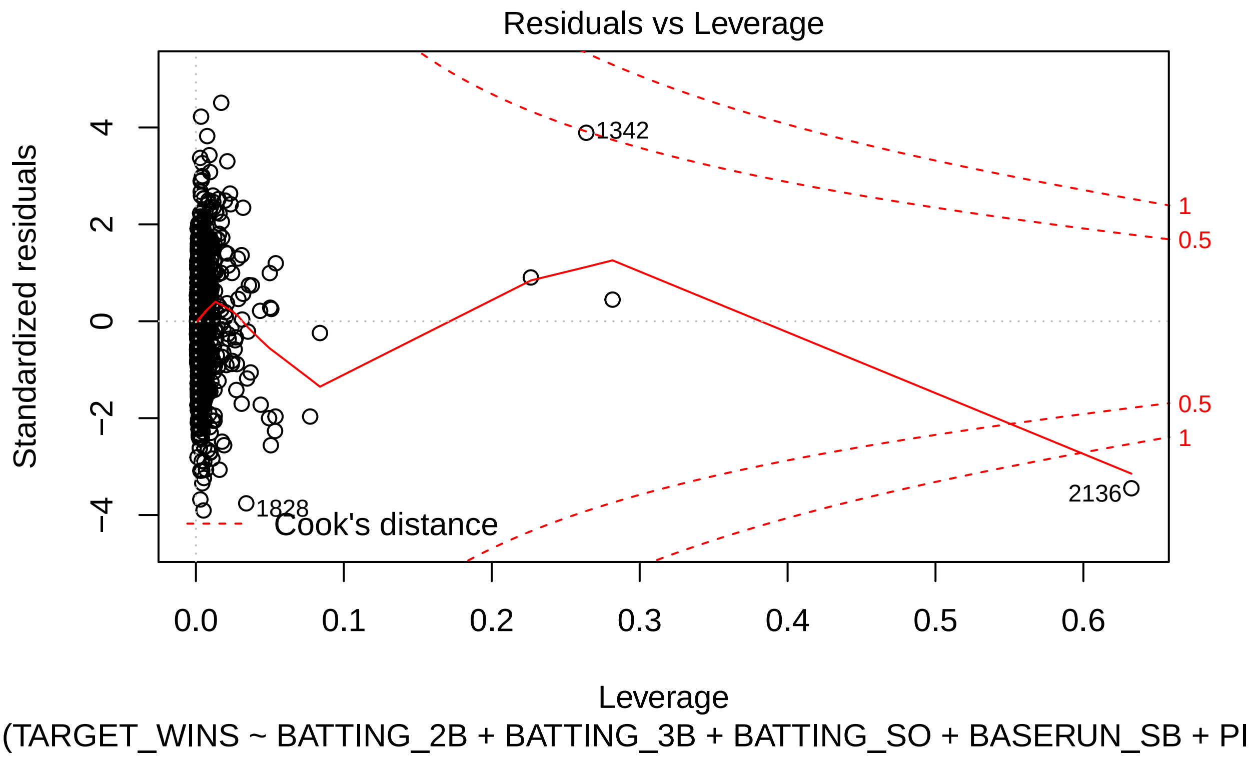
Model 3: get rid of variables are not statistically significant.

```
m3<-update(m1,~.-BATTING_HR-BATTING_BB-BASERUN_CS-PITCHING_HR-PITCHING_BB)
plot(m3)
```




Normal Q−Q
Standardized residuals
4
2
0
−2
−4
2012
1820
859
−3
−2
−1
0
1
2
3
Theoretical Quantiles
(TARGET_WINS ~ BATTING_2B + BATTING_3B + BATTING_SO + BASERUN_SB + PI


Scale−Location
√|Standardized residuals|
2.0
1.5
1.0
0.5
0.0
2012
1820
859
0
20
40
60
80
100
120
Fitted values
(TARGET_WINS ~ BATTING_2B + BATTING_3B + BATTING_SO + BASERUN_SB + PI

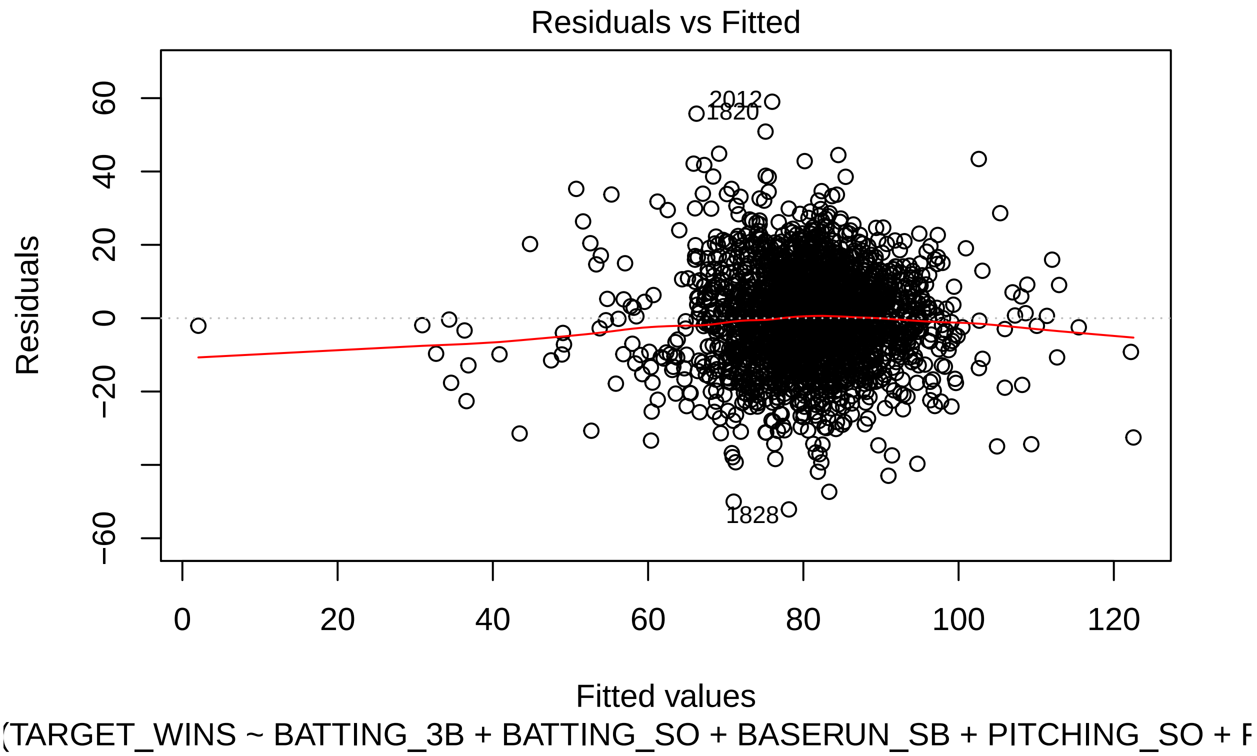


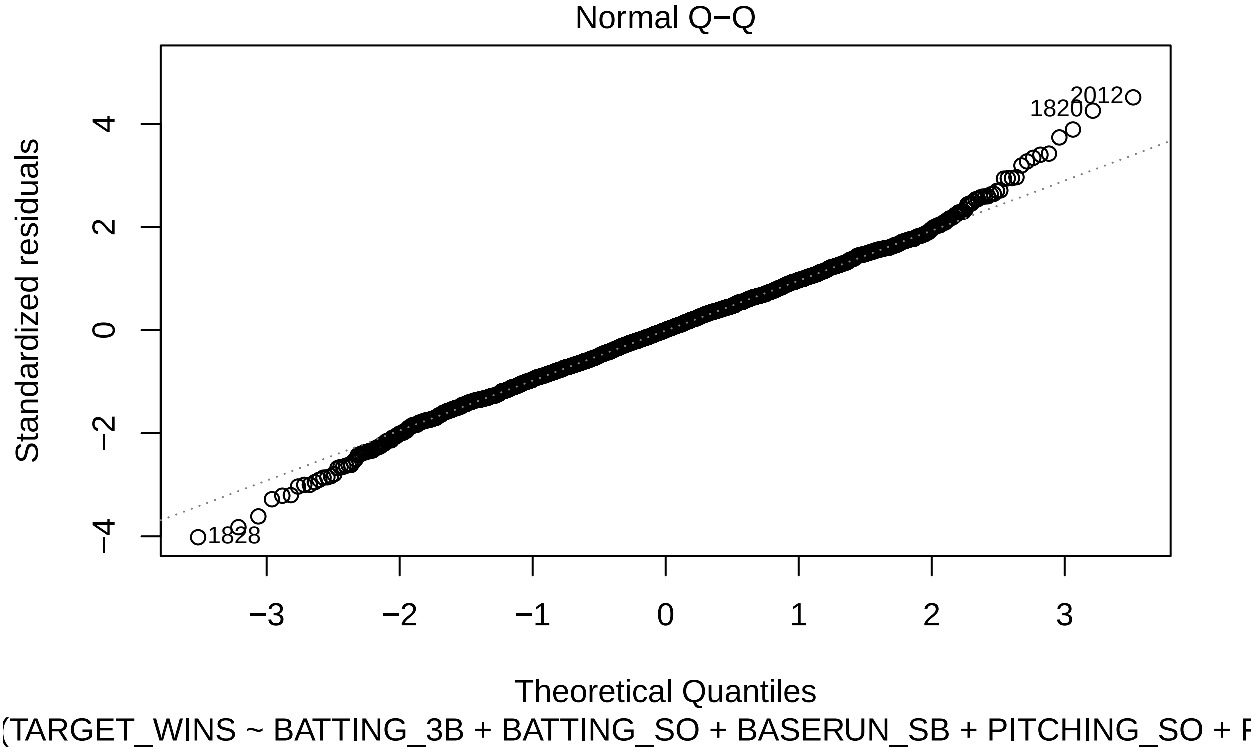
```r
summary(m3)
```

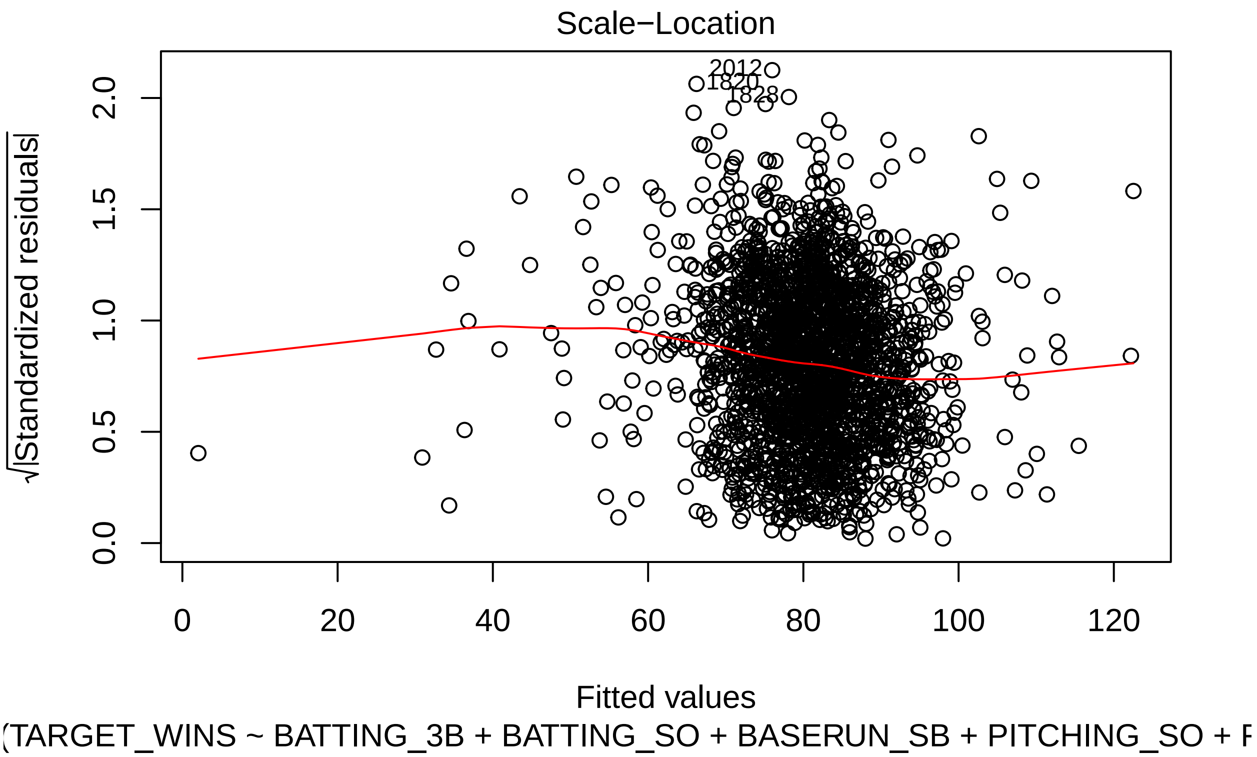
```
##
## Call:
## lm(formula = TARGET_WINS ~ BATTING_2B + BATTING_3B + BATTING_SO +
##     BASERUN_SB + PITCHING_H + PITCHING_SO + FIELDING_E + FIELDING_DP +
##     BATTING_1B + Total_batting, data = moneyball)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.028  -8.645   0.079   8.538  58.554
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    28.5417286  4.9219496   5.799 7.61e-09 ***
## BATTING_2B     -0.0094319  0.0094685  -0.996 0.319289
## BATTING_3B      0.0605133  0.0156578   3.865 0.000114 ***
## BATTING_SO     -0.0099177  0.0024244  -4.091 4.45e-05 ***
## BASERUN_SB      0.0298779  0.0039613   7.542 6.64e-14 ***
## PITCHING_H     -0.0008022  0.0003216  -2.494 0.012699 *
## PITCHING_SO     0.0029552  0.0006735   4.388 1.20e-05 ***
## FIELDING_E     -0.0226345  0.0021496 -10.530  < 2e-16 ***
## FIELDING_DP    -0.1125269  0.0126728  -8.879  < 2e-16 ***
## BATTING_1B      0.0284034  0.0050326   5.644 1.87e-08 ***
## Total_batting   0.0191872  0.0023145   8.290  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.1 on 2265 degrees of freedom
## Multiple R-squared:  0.3115, Adjusted R-squared:  0.3085
## F-statistic: 102.5 on 10 and 2265 DF,  p-value: < 2.2e-16
```

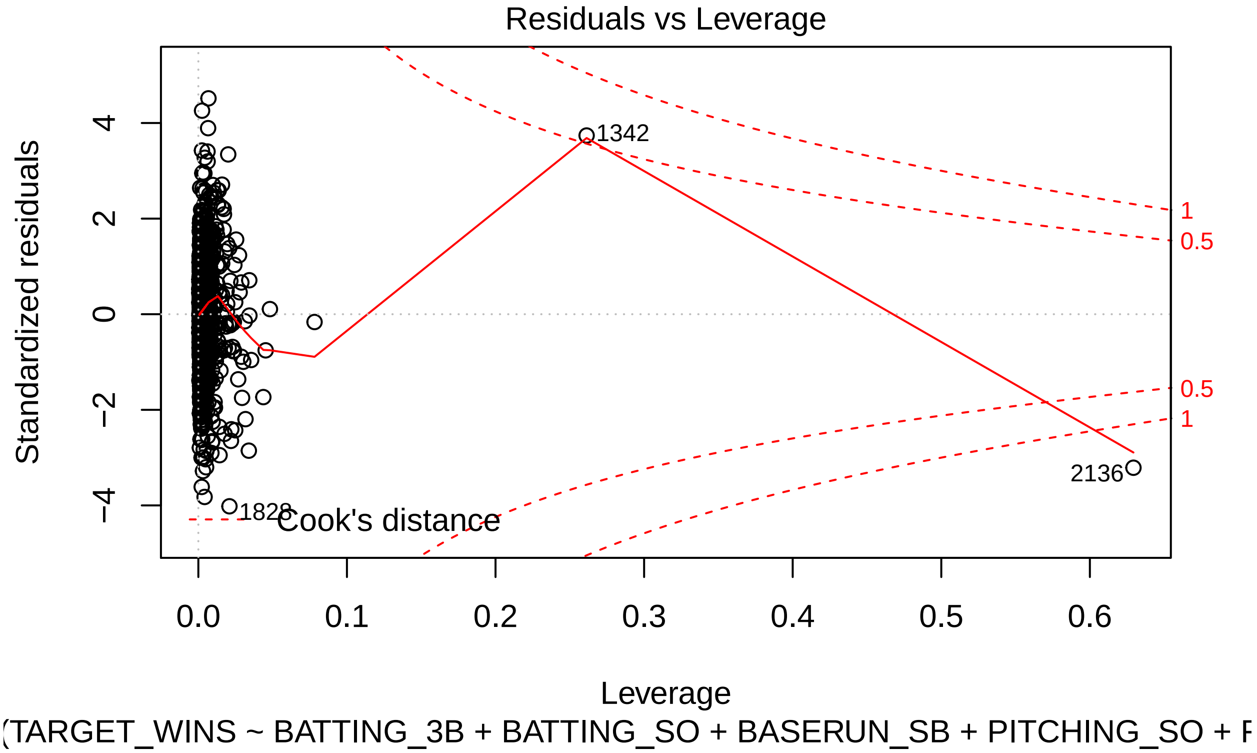
Model4: Keep getting rid of the insignificant variable.

```
m4 <-update(m3,~.-BATTING_2B-PITCHING_H)
plot(m4)
```









```r
summary(m4)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ BATTING_3B + BATTING_SO + BASERUN_SB +
##     PITCHING_SO + FIELDING_E + FIELDING_DP + BATTING_1B + Total_batting,
##     data = moneyball)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.140  -8.690   0.037   8.443  59.016
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   27.996811   4.798551   5.834 6.17e-09 ***
## BATTING_3B     0.069968   0.015282   4.578 4.94e-06 ***
## BATTING_SO    -0.007831   0.002273  -3.446 0.000580 ***
## BASERUN_SB     0.030903   0.003936   7.851 6.33e-15 ***
## PITCHING_SO    0.002079   0.000591   3.517 0.000444 ***
## FIELDING_E    -0.025823   0.001712 -15.081  < 2e-16 ***
## FIELDING_DP   -0.111019   0.012628  -8.792  < 2e-16 ***
## BATTING_1B     0.028995   0.004849   5.979 2.60e-09 ***
## Total_batting  0.017044   0.001734   9.830  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.11 on 2267 degrees of freedom
## Multiple R-squared:  0.3093, Adjusted R-squared:  0.3068
## F-statistic: 126.9 on 8 and 2267 DF,  p-value: < 2.2e-16
```

Evaluation by using our model

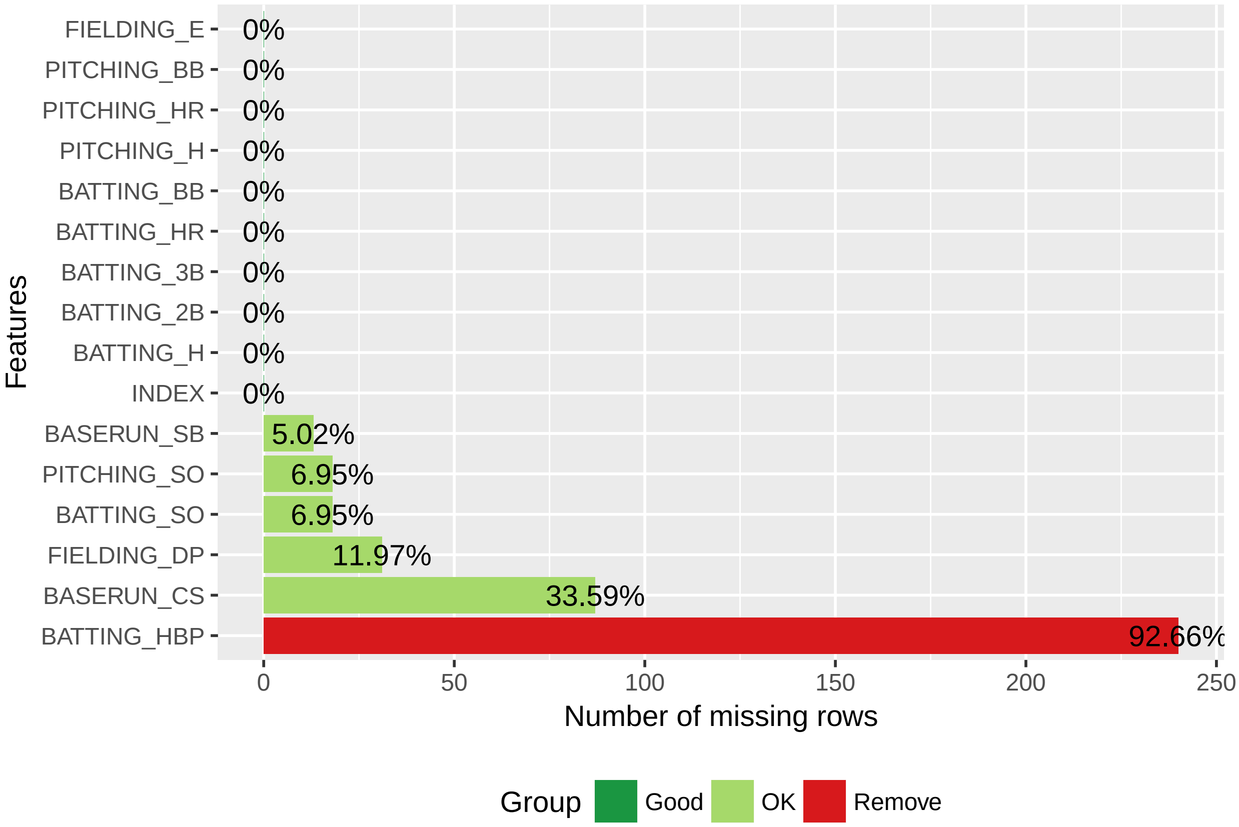
```
evaluation <- read.csv("https://raw.githubusercontent.com/xkong100/data-621/master/Hw1/moneyball-evalua
kable(head(evaluation))
```

INDEX	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_
9	1209	170	33	83	
10	1221	151	29	88	
14	1395	183	29	93	
47	1539	309	29	159	
60	1445	203	68	5	
63	1431	236	53	10	

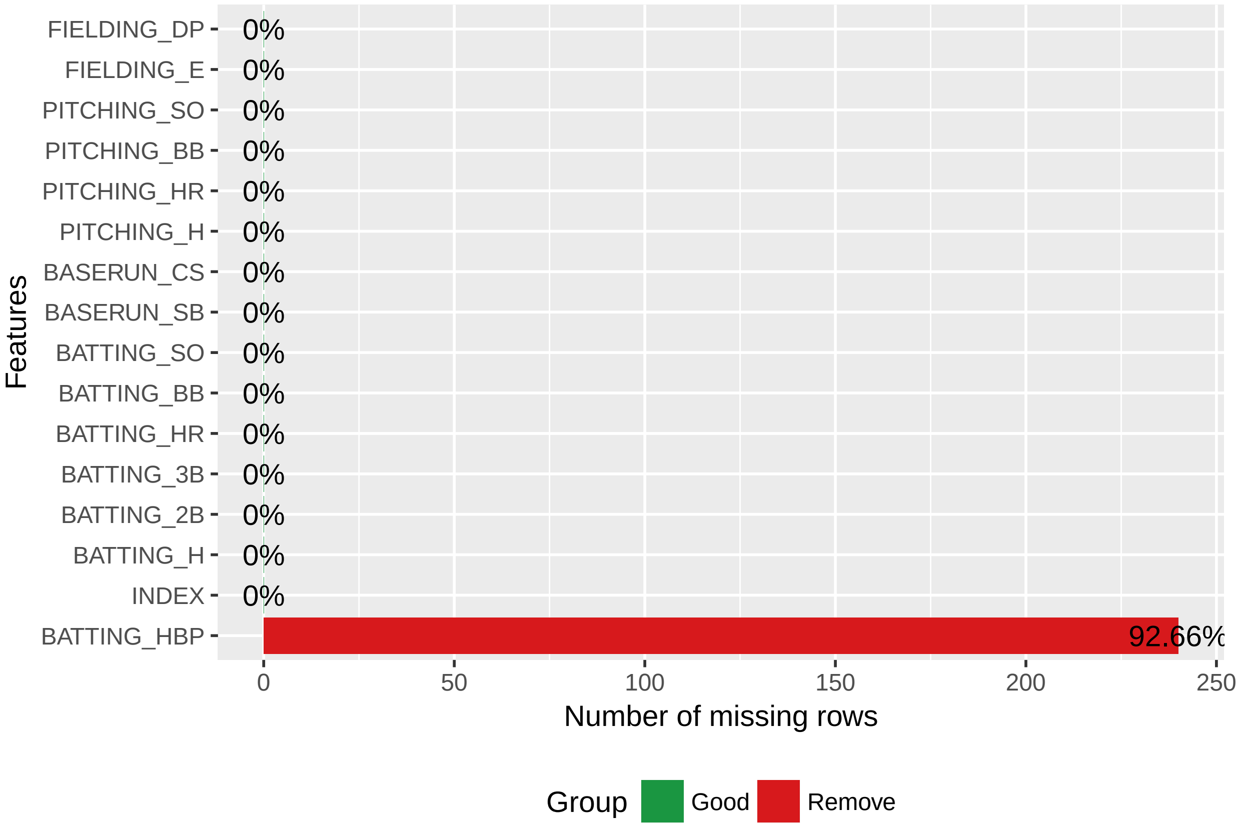
```
evaluation <- cleanNames(evaluation)
kable(head(evaluation))
```

INDEX	BATTING_H	BATTING_2B	BATTING_3B	BATTING_HR	BATTING_BB	BATTING_SO	BASE
9	1209	170	33	83	447	1080	
10	1221	151	29	88	516	929	
14	1395	183	29	93	509	816	
47	1539	309	29	159	486	914	
60	1445	203	68	5	95	416	
63	1431	236	53	10	215	377	

```
plot_missing(evaluation)
```



```
evaluation[is.na(evaluation$BATTING_SO),"BATTING_SO"] <- median(na.omit(evaluation$BATTING_SO))
evaluation[is.na(evaluation$PITCHING_SO),"PITCHING_SO"] <- median(na.omit(evaluation$PITCHING_SO))
evaluation[is.na(evaluation$BASERUN_SB),"BASERUN_SB"] <- median(na.omit(evaluation$BASERUN_SB))
evaluation[is.na(evaluation$FIELDING_DP),"FIELDING_DP"] <- median(na.omit(evaluation$FIELDING_DP))
evaluation[is.na(evaluation$BASERUN_CS),"BASERUN_CS"] <- median(na.omit(evaluation$BASERUN_CS))
plot_missing(evaluation)
```



```
attach(evaluation)
```

```
## The following objects are masked from moneyball (pos = 3):
##
##     BASERUN_CS, BASERUN_SB, BATTING_2B, BATTING_3B, BATTING_BB,
##     BATTING_H, BATTING_HBP, BATTING_HR, BATTING_SO, FIELDING_DP,
##     FIELDING_E, INDEX, PITCHING_BB, PITCHING_H, PITCHING_HR,
##     PITCHING_SO
```

```
## The following objects are masked from moneyball (pos = 4):
##
##     BASERUN_CS, BASERUN_SB, BATTING_2B, BATTING_3B, BATTING_BB,
##     BATTING_H, BATTING_HBP, BATTING_HR, BATTING_SO, FIELDING_DP,
##     FIELDING_E, INDEX, PITCHING_BB, PITCHING_H, PITCHING_HR,
##     PITCHING_SO
```

```
evaluation<- evaluation %>% mutate(BATTING_1B=BATTING_H-BATTING_2B-BATTING_3B,Total_batting=1*BATTING_1
kable(head(evaluation))
```

BATTING_2B	BATTING_3B	BATTING_HR	BATTING_BB	BATTING_SO	BASERUN_SB	BASERUN_C
170	33	83	447	1080	62	50
151	29	88	516	929	54	39
183	29	93	509	816	59	47
309	29	159	486	914	148	57
203	68	5	95	416	92	49
236	53	10	215	377	92	49

```
pred <- predict(m1, evaluation,type='response')
```

```
## Warning in predict.lm(m1, evaluation, type = "response"): prediction from a
## rank-deficient fit may be misleading
```

```
final <- data.frame(cbind(pred, moneyball$TARGET_WINS))
```

```
## Warning in cbind(pred, moneyball$TARGET_WINS): number of rows of result is
## not a multiple of vector length (arg 1)
```

```
kable(head(final))
```

pred	V2
64.02285	39
65.73235	70
75.27890	86
85.74341	70
66.37855	82
69.83817	75