HW1

Aadi Kalloo

June 12, 2016

VAR_NAME	MEAN	MEDIAN
INDEX	1268.46353	1270.5
TARGET_WINS	80.79086	82.0
TEAM_BATTING_H	1469.26977	1454.0
TEAM_BATTING_2B	241.24692	238.0
TEAM_BATTING_3B	55.25000	47.0
TEAM_BATTING_HR	99.61204	102.0
TEAM_BATTING_BB	501.55888	512.0
TEAM_BATTING_SO	735.60534	750.0
TEAM_BASERUN_SB	124.76177	101.0
TEAM_BASERUN_CS	52.80386	49.0
TEAM_BATTING_HBP	59.35602	58.0
TEAM_PITCHING_H	1779.21046	1518.0
TEAM_PITCHING_HR	105.69859	107.0
TEAM_PITCHING_BB	553.00791	536.5
TEAM_PITCHING_SO	817.73045	813.5
TEAM_FIELDING_E	246.48067	159.0
TEAM_FIELDING_DP	146.38794	149.0

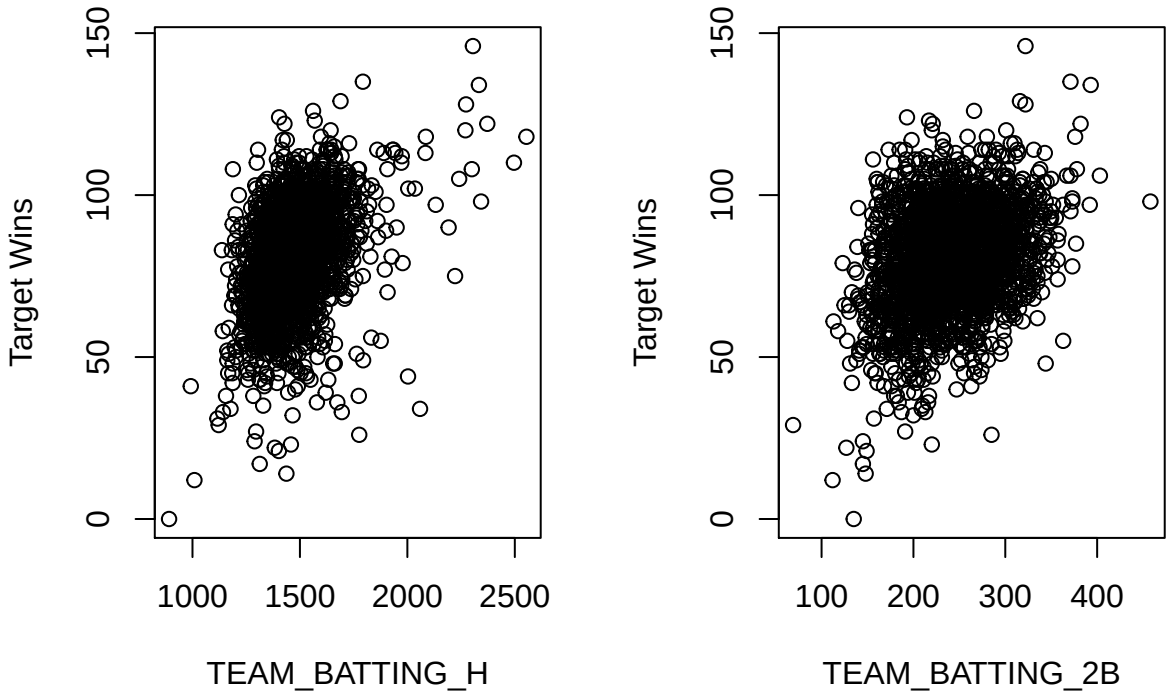
VAR_NAME	CORRELATION TO WINS (r)
TEAM_BATTING_H	0.38876752
TEAM_BATTING_2B	0.28910365
TEAM_BATTING_3B	0.14260841
TEAM_BATTING_HR	0.17615320
TEAM_BATTING_BB	0.23255986
TEAM_BATTING_SO	-0.03175071
TEAM_BASERUN_SB	0.13513892
TEAM_BASERUN_CS	0.02240407
TEAM_BATTING_HBP	0.07350424
TEAM_PITCHING_H	-0.10993705
TEAM_PITCHING_HR	0.18901373
TEAM_PITCHING_BB	0.12417454
TEAM_PITCHING_SO	-0.07843609
TEAM_FIELDING_E	-0.17648476
TEAM_FIELDING_DP	-0.03485058

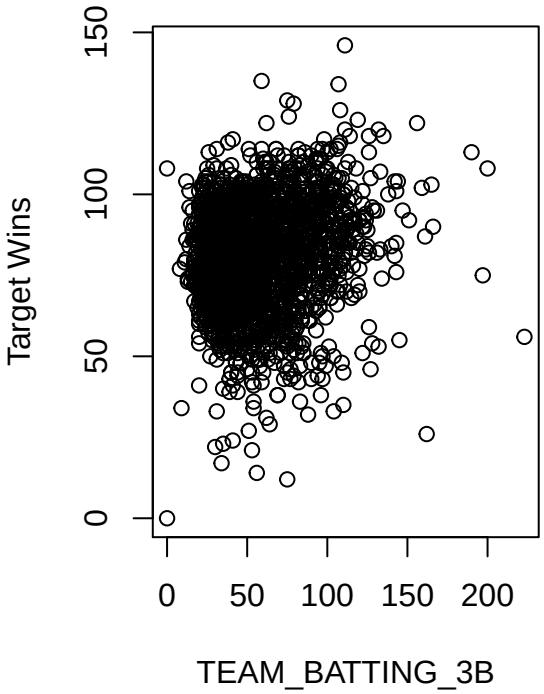
VAR_NAME	NUM_MISSING
INDEX	0
TARGET_WINS	0
TEAM_BATTING_H	0
TEAM_BATTING_2B	0
TEAM_BATTING_3B	0

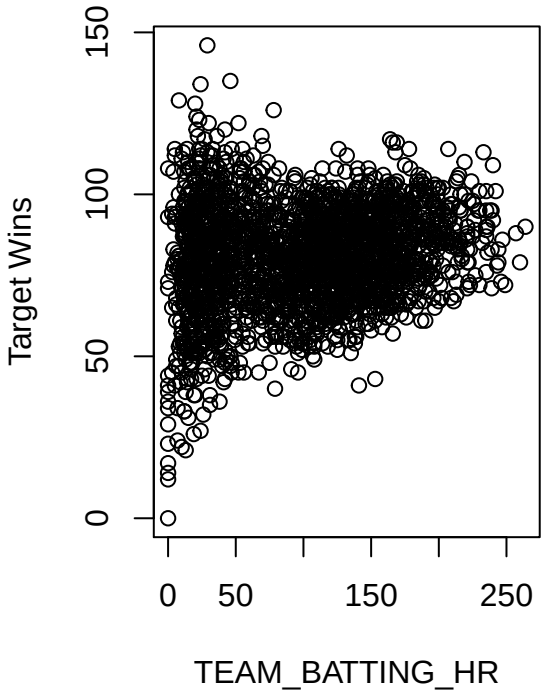
VAR_NAME	NUM_MISSING
TEAM_BATTING_HR	0
TEAM_BATTING_BB	0
TEAM_BATTING_SO	102
TEAM_BASERUN_SB	131
TEAM_BASERUN_CS	772
TEAM_BATTING_HBP	2085
TEAM_PITCHING_H	0
TEAM_PITCHING_HR	0
TEAM_PITCHING_BB	0
TEAM_PITCHING_SO	102
TEAM_FIELDING_E	0
TEAM_FIELDING_DP	286

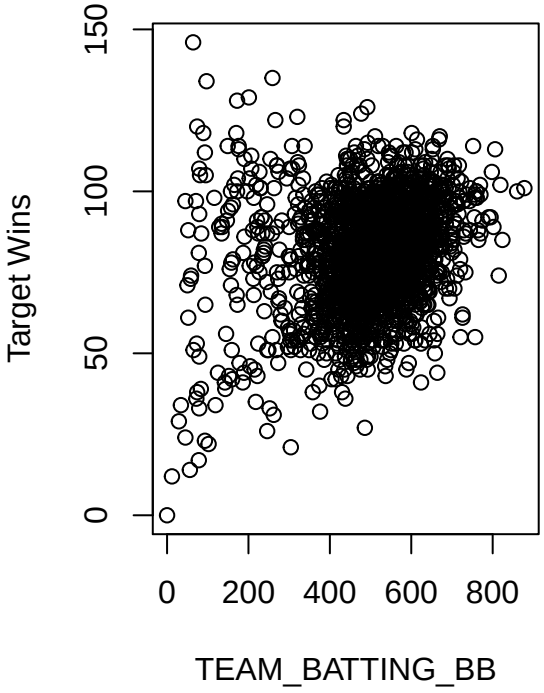
Graphs

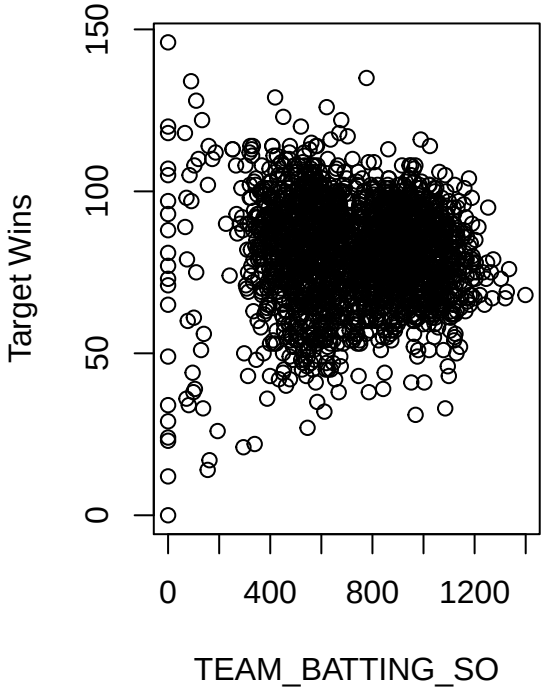
```
par(mfrow = c(1,2))
for (var_count in 3:17) {
  plot(x = trainingdata[, var_count], y = trainingdata$TARGET_WINS, xlab = names(trainingdata)[var_coun
}
```

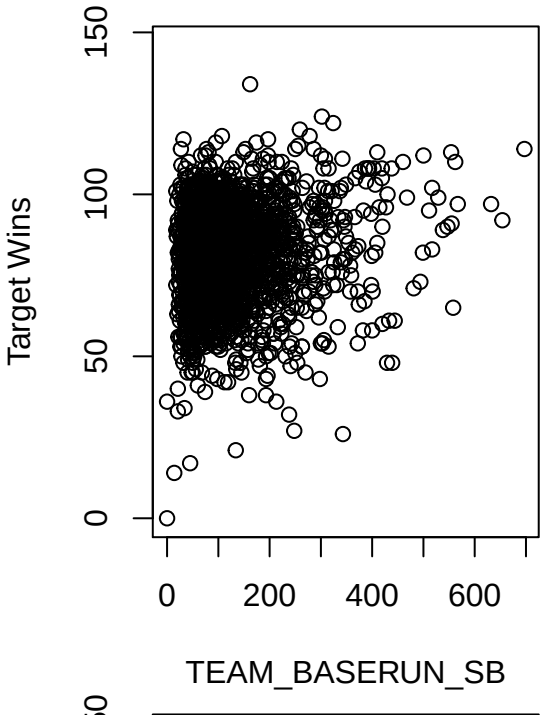


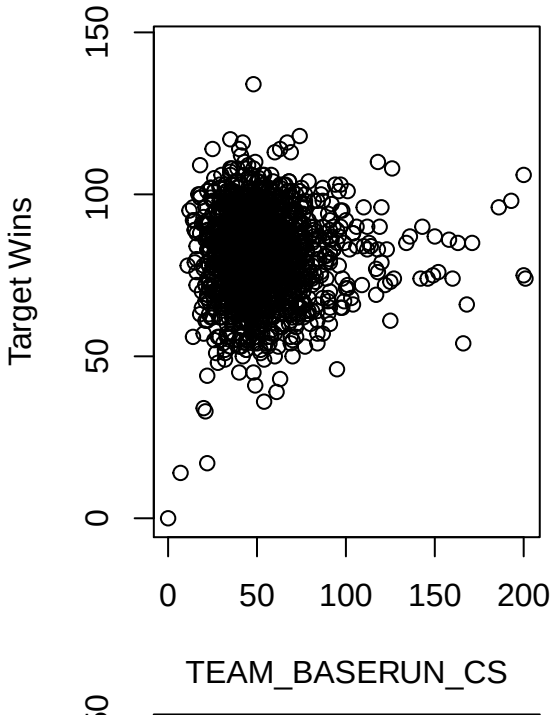

Target Wins
0
50
100
150
0
50
100
150
200
TEAM_BATTING_3B

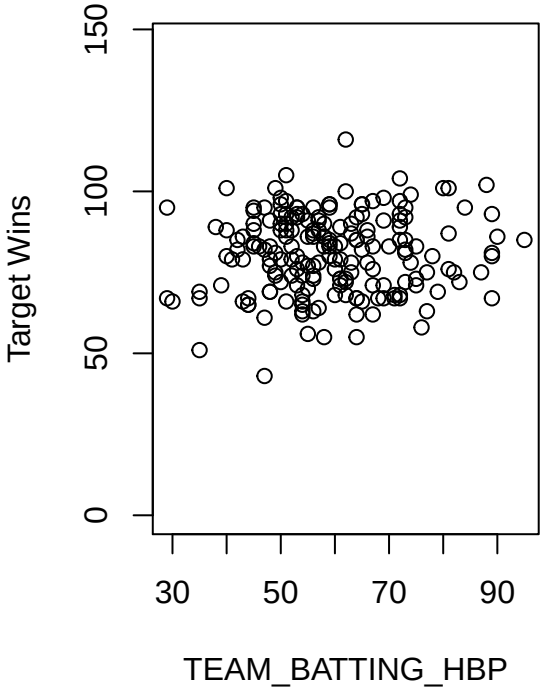

Target Wins
0
50
100
150
0
50
150
250
TEAM_BATTING_HR

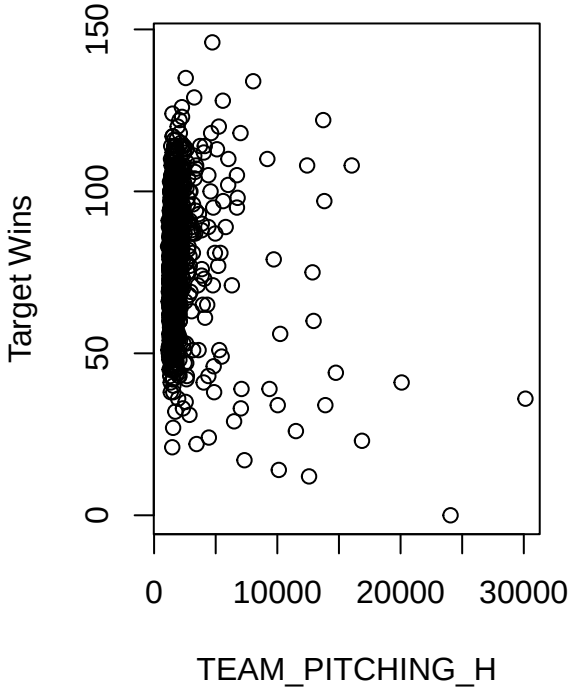

Target Wins
0
50
100
150
0
200
400
600
800
TEAM_BATTING_BB

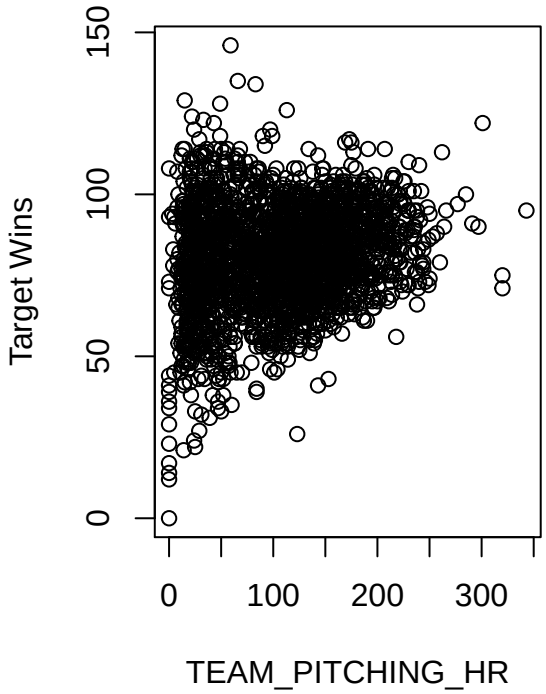

Target Wins
0
50
100
150
0
400
800
1200
TEAM_BATTING_SO

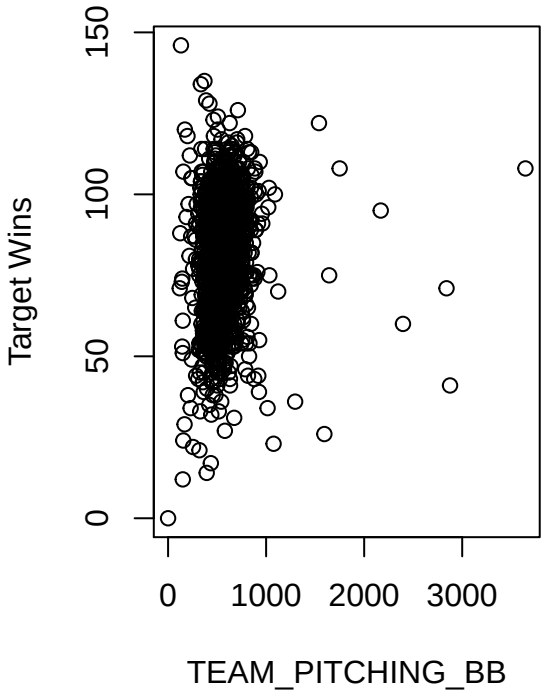

Target Wins
0
50
100
150
0
200
400
600
TEAM_BASERUN_SB

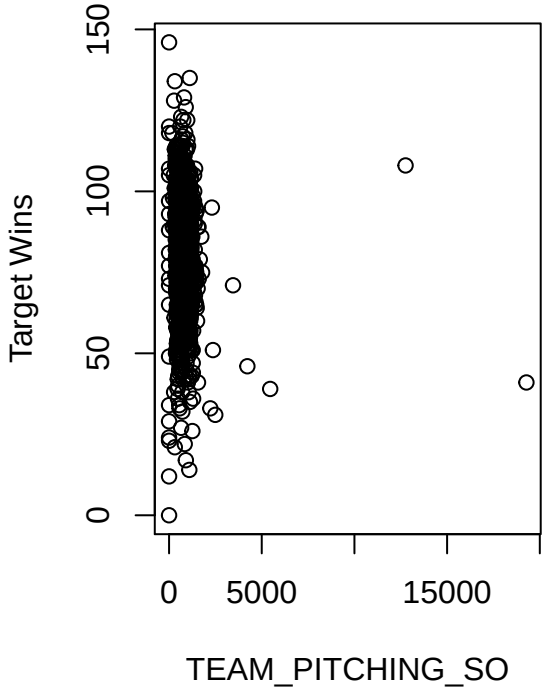

Target Wins
0
50
100
150
0
50
100
150
200
TEAM_BASERUN_CS

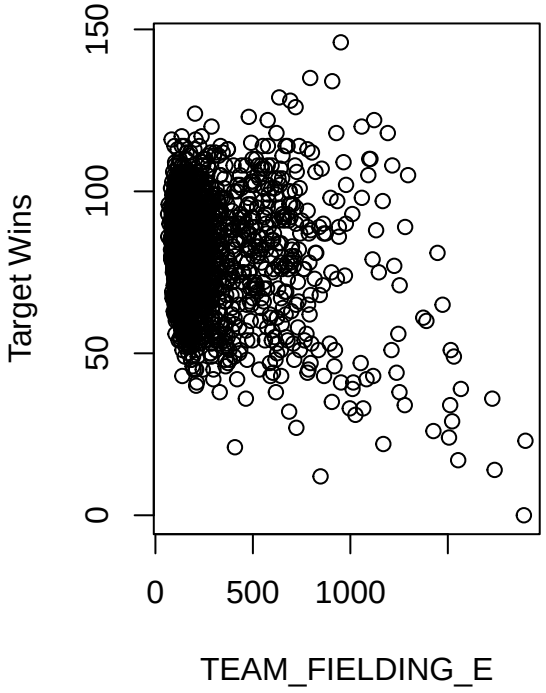

Target Wins
0
50
100
150
30
50
70
90
TEAM_BATTING_HBP

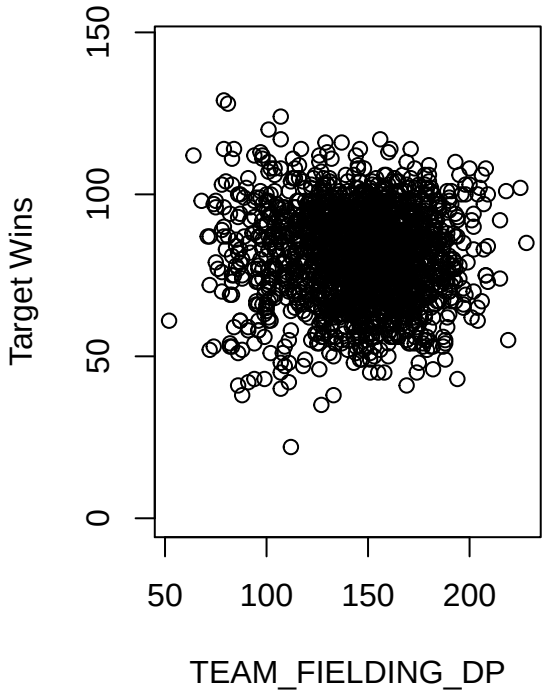

Target Wins
0
50
100
150
0
10000
20000
30000
TEAM_PITCHING_H


Target Wins
0
50
100
150
0
100
200
300
TEAM_PITCHING_HR


Target Wins
0
50
100
150
0
1000
2000
3000
TEAM_PITCHING_BB

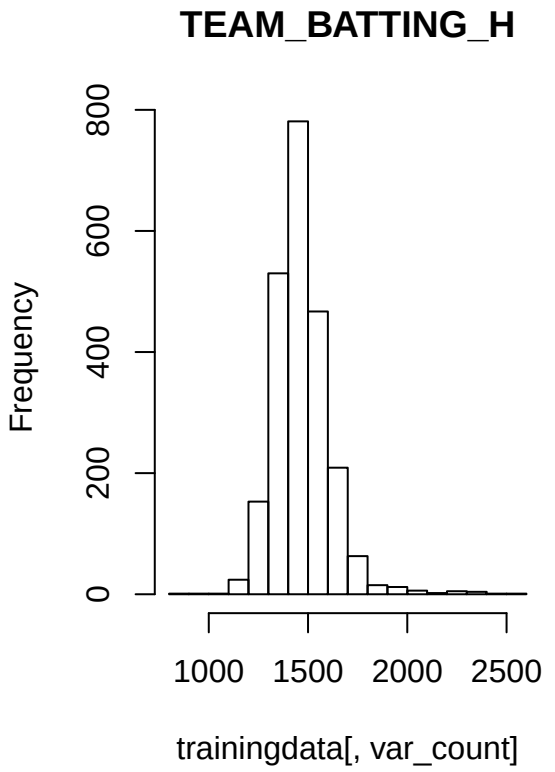

Target Wins
0
50
100
150
0
5000
15000
TEAM_PITCHING_SO

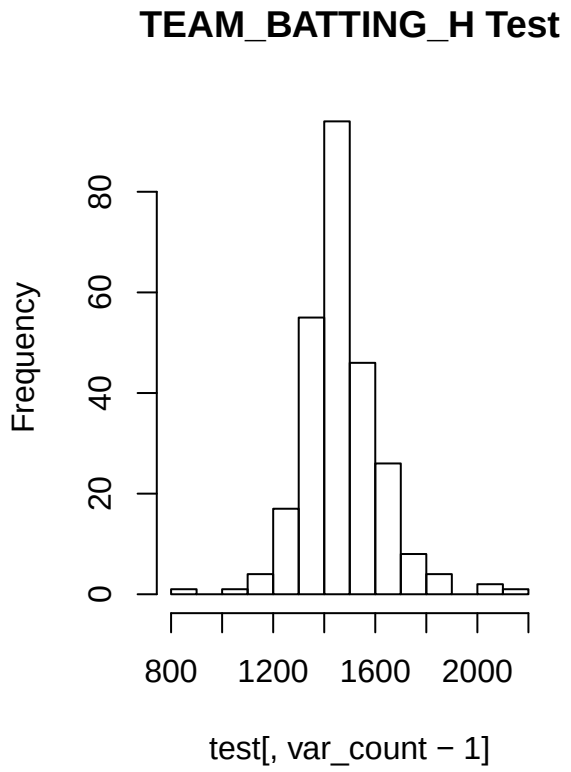

Target Wins
0
50
100
150
0
500
1000
TEAM_FIELDING_E

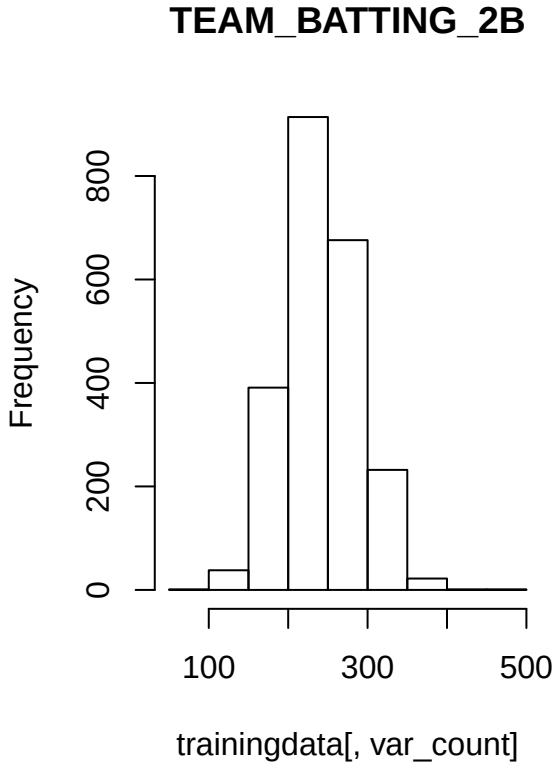


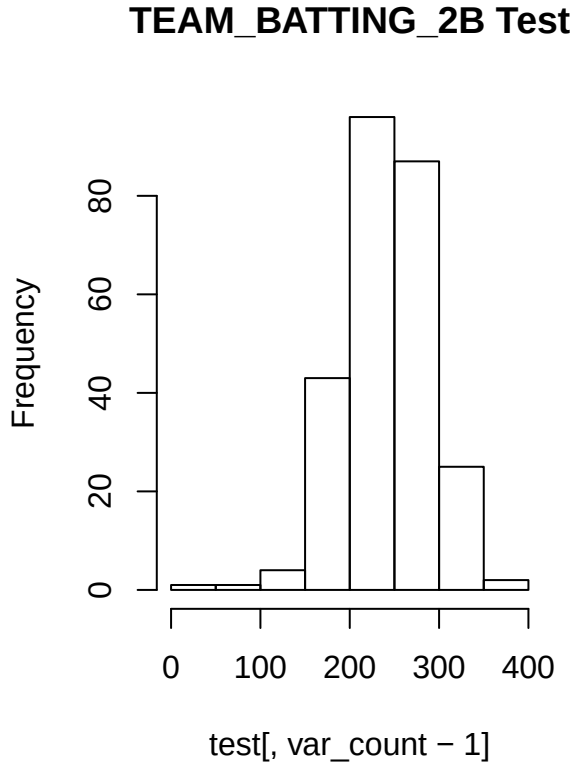

```
scale2 <- function(x) {scale(x, center = TRUE, scale = TRUE)}

training_Zscores = as.data.frame(lapply(trainingdata, scale2))
training_Zscores$INDEX = trainingdata$INDEX

test = read.csv("https://raw.githubusercontent.com/dsmilo/DATA621/master/HW1/data/moneyball-evaluation-
test_ZScores = as.data.frame(lapply(test, scale2))
test_ZScores$INDEX = test$INDEX

write.csv(training_Zscores, "trainingZ.csv")
write.csv(test_ZScores, "testZ.csv")
```

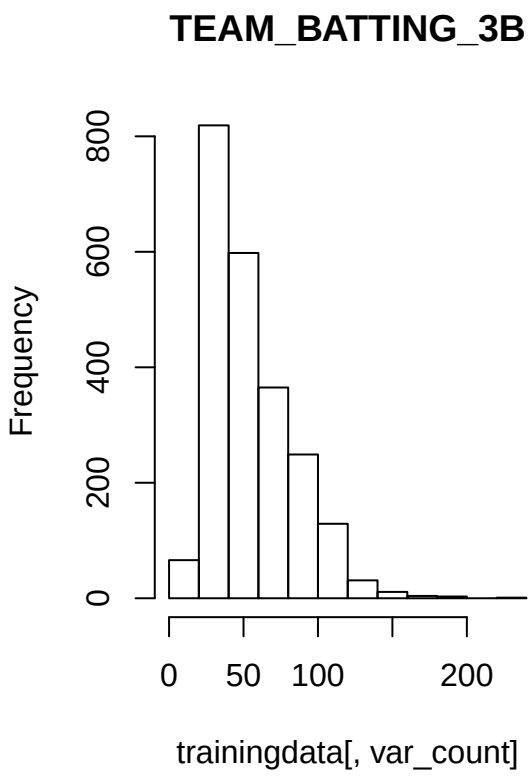
```
par(mfrow = c(1,2))
for (var_count in 3:17) {
  hist(x = trainingdata[, var_count], main = names(trainingdata)[var_count])
  hist(x = test[, var_count - 1], main = paste0(names(trainingdata)[var_count], " Test"))
}
```

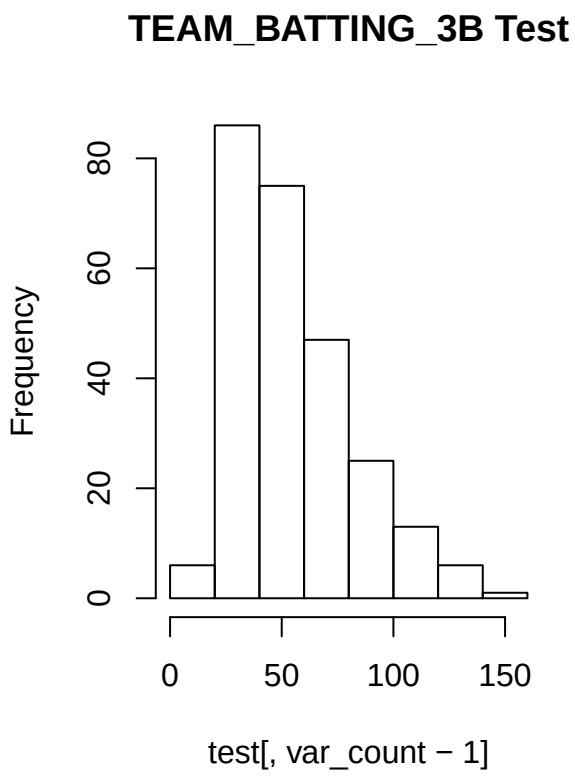

TEAM_BATTING_H
Frequency
0
200
400
600
800
1000
1500
2000
2500
trainingdata[, var_count]

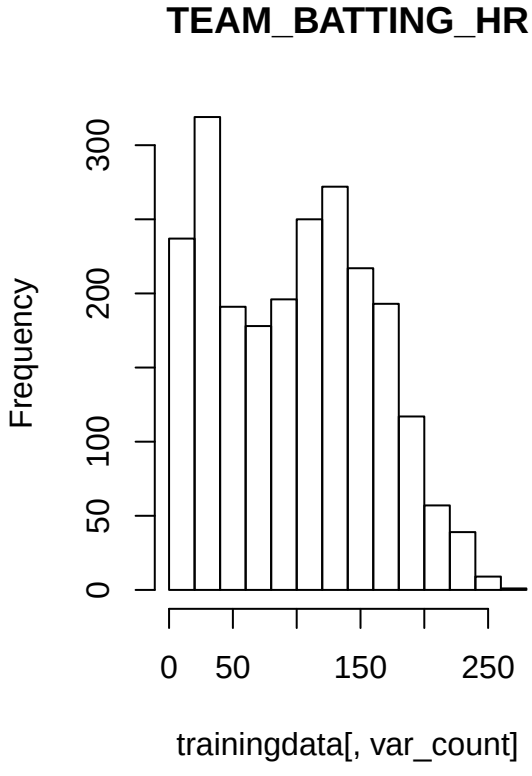

TEAM_BATTING_H Test
Frequency
0
20
40
60
80
800
1200
1600
2000
test[, var_count − 1]

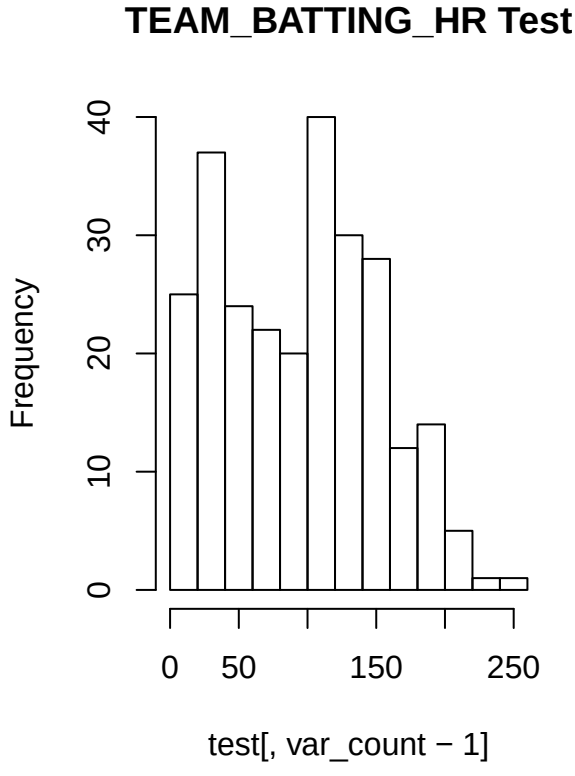

TEAM_BATTING_2B
Frequency
0
200
400
600
800
100
300
500
trainingdata[, var_count]

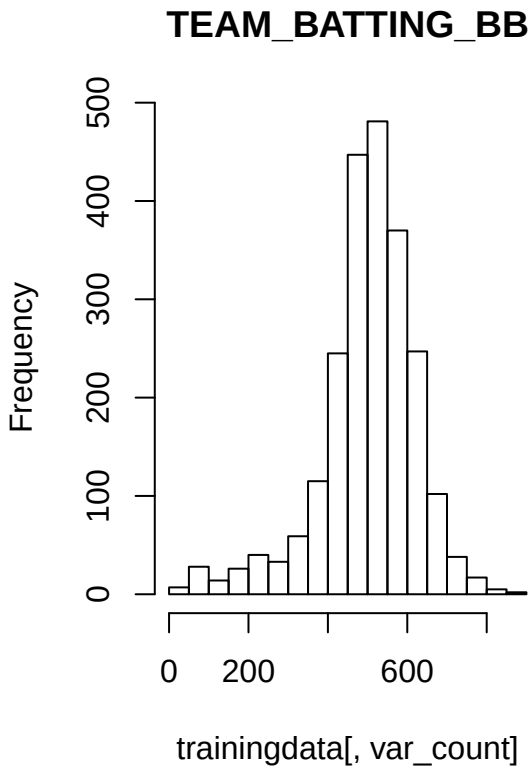

TEAM_BATTING_2B Test
Frequency
0
20
40
60
80
0
100
200
300
400
test[, var_count − 1]

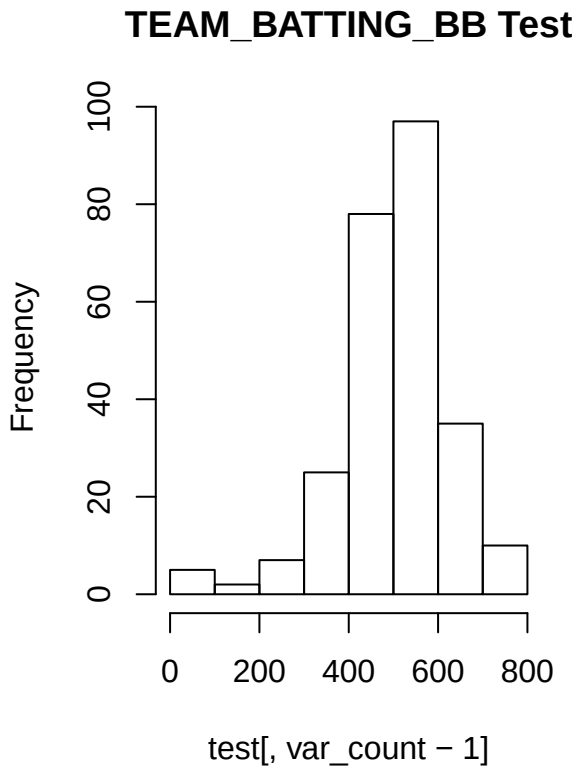

TEAM_BATTING_3B
Frequency
0
200
400
600
800
0
50
100
200
trainingdata[, var_count]

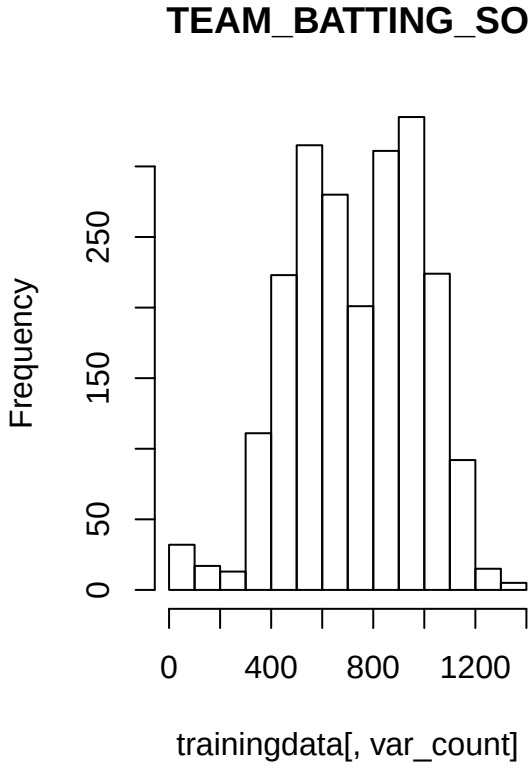

TEAM_BATTING_3B Test
Frequency
0
20
40
60
80
0
50
100
150
test[, var_count − 1]

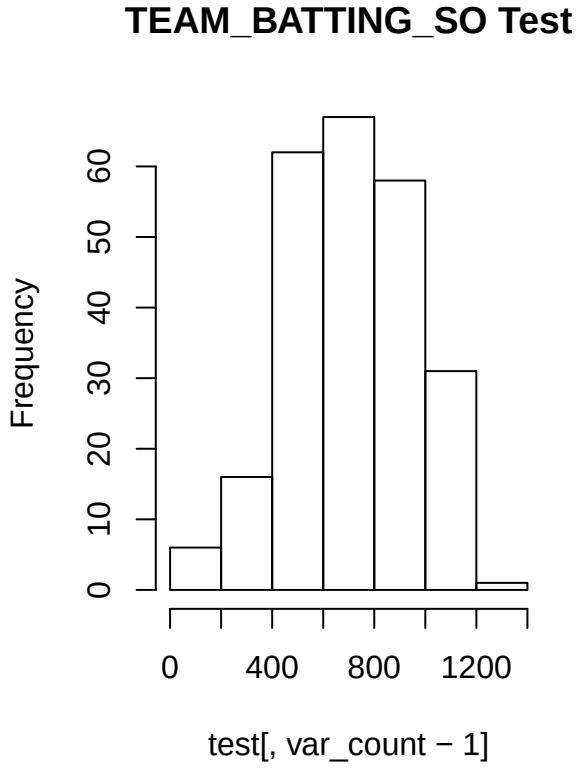

TEAM_BATTING_HR
Frequency
0
50
100
200
300
0
50
150
250
trainingdata[, var_count]

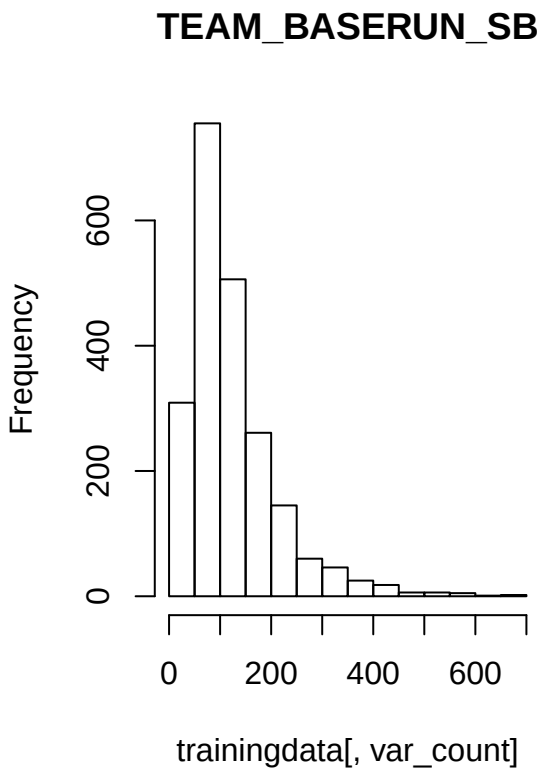

TEAM_BATTING_HR Test
Frequency
0
10
20
30
40
0
50
150
250
test[, var_count − 1]

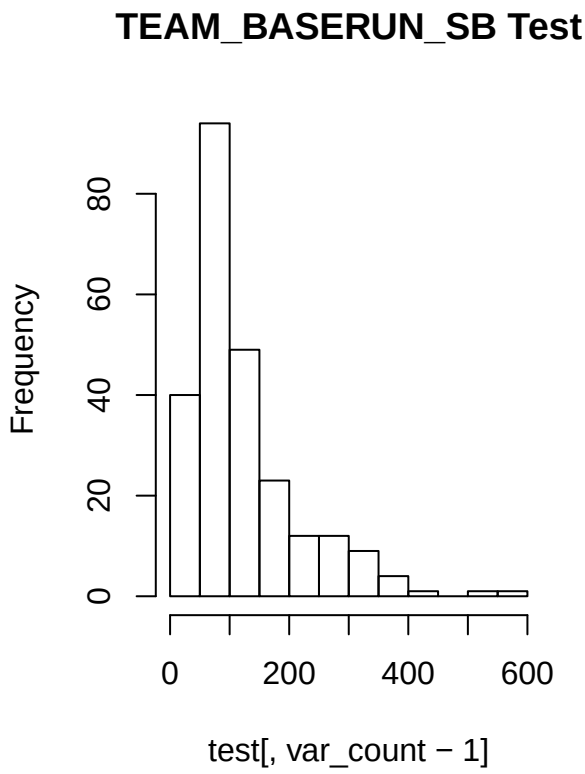

TEAM_BATTING_BB
Frequency
0
100
200
300
400
500
0
200
600
trainingdata[, var_count]

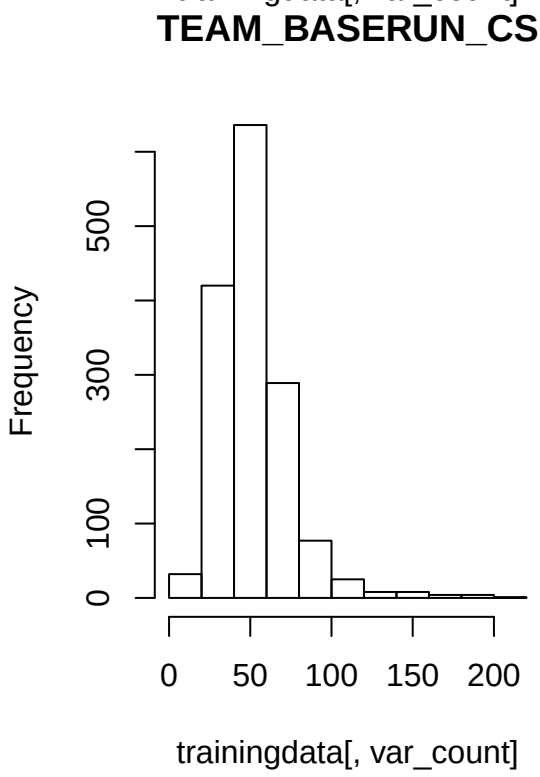

TEAM_BATTING_BB Test
Frequency
0
20
40
60
80
100
0
200
400
600
800
test[, var_count − 1]

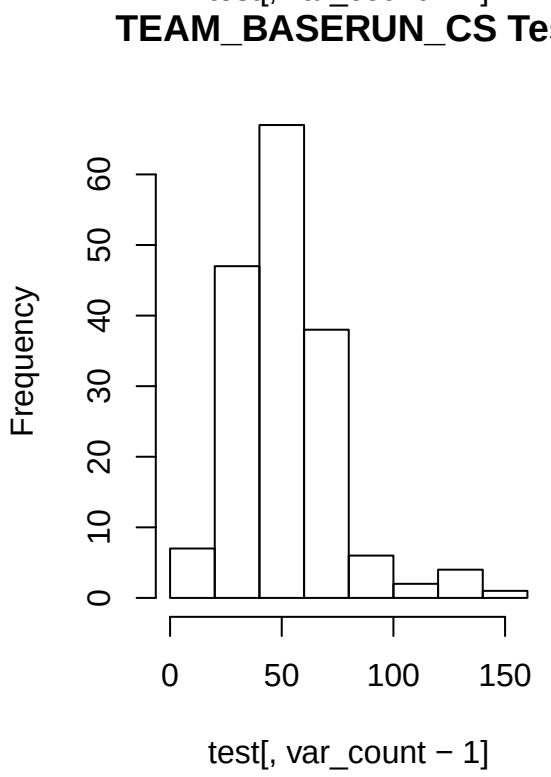

TEAM_BATTING_SO
Frequency
0
50
150
250
0
400
800
1200
trainingdata[, var_count]

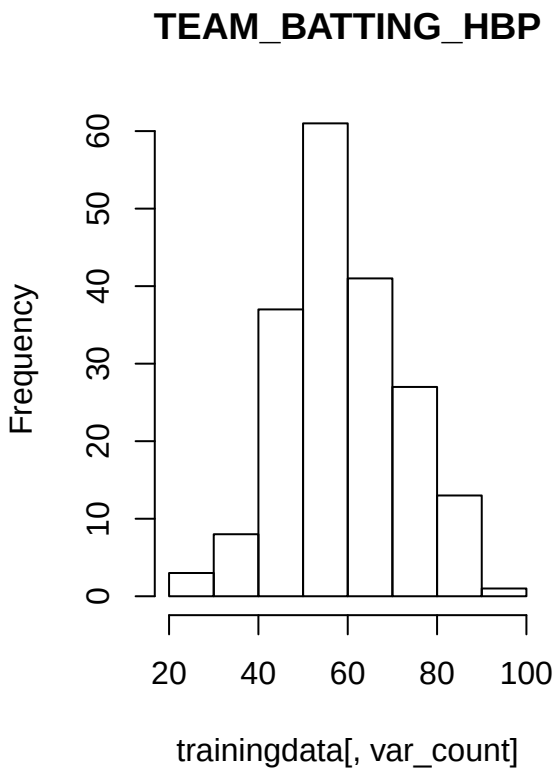

TEAM_BATTING_SO Test
Frequency
0
10
20
30
40
50
60
0
400
800
1200
test[, var_count − 1]

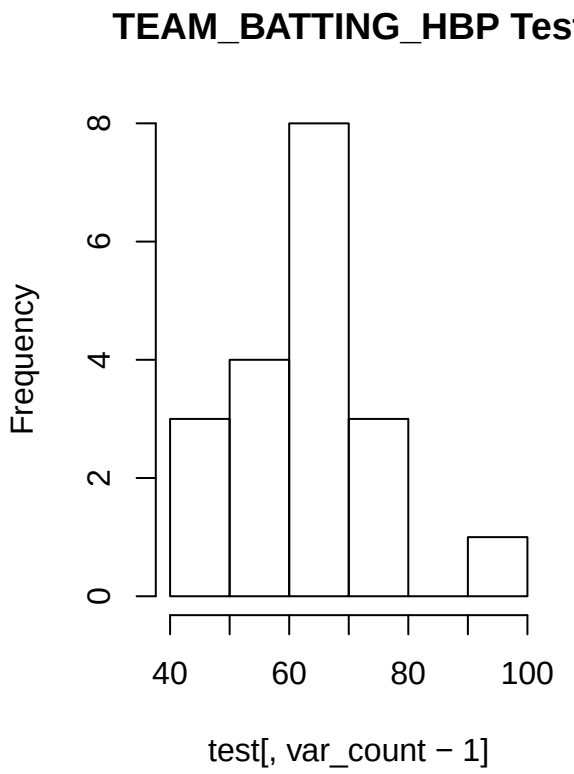

TEAM_BASERUN_SB
Frequency
0
200
400
600
0
200
400
600
trainingdata[, var_count]

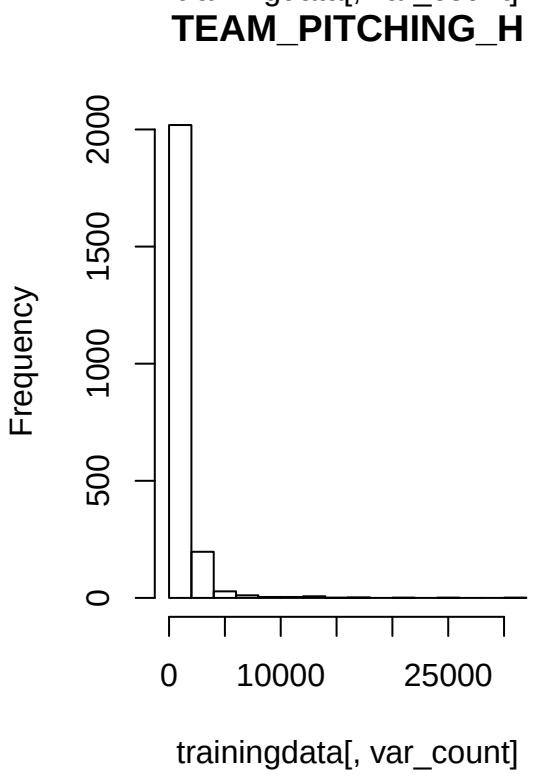

TEAM_BASERUN_SB Test
Frequency
0
20
40
60
80
0
200
400
600
test[, var_count − 1]

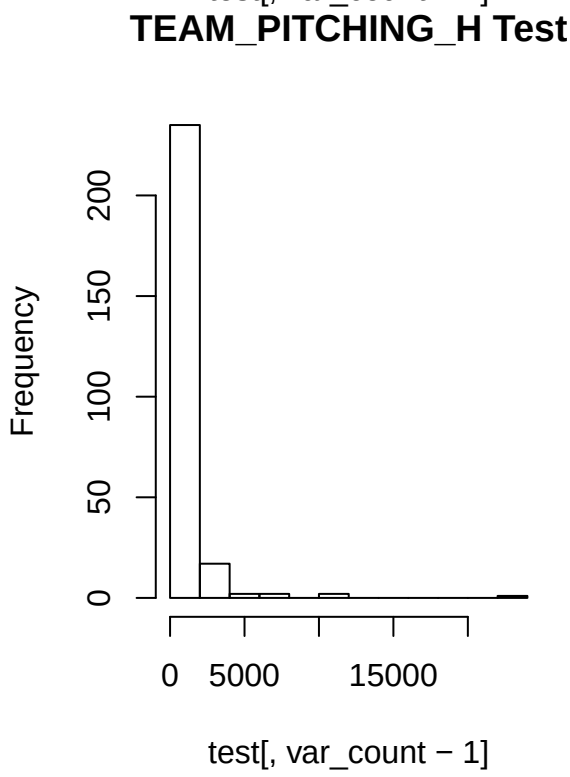

TEAM_BASERUN_CS
Frequency
0
100
300
500
0
50
100
150
200
trainingdata[, var_count]

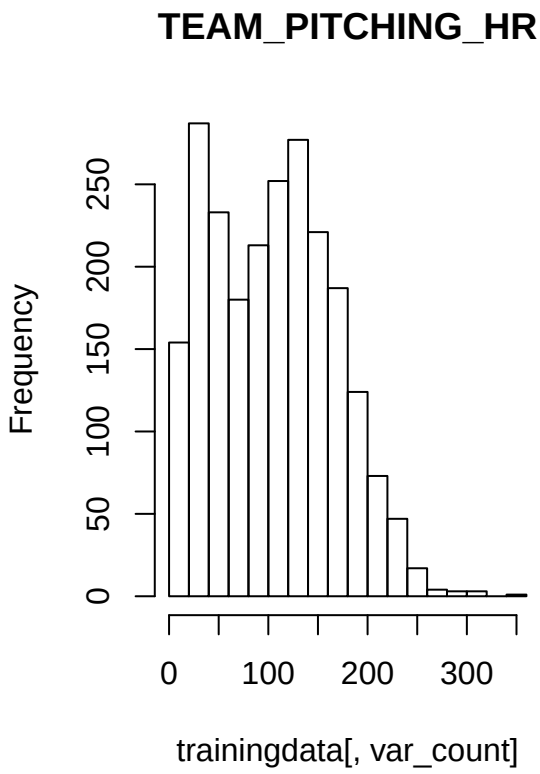

TEAM_BASERUN_CS Test
Frequency
0
10
20
30
40
50
60
0
50
100
150
test[, var_count − 1]

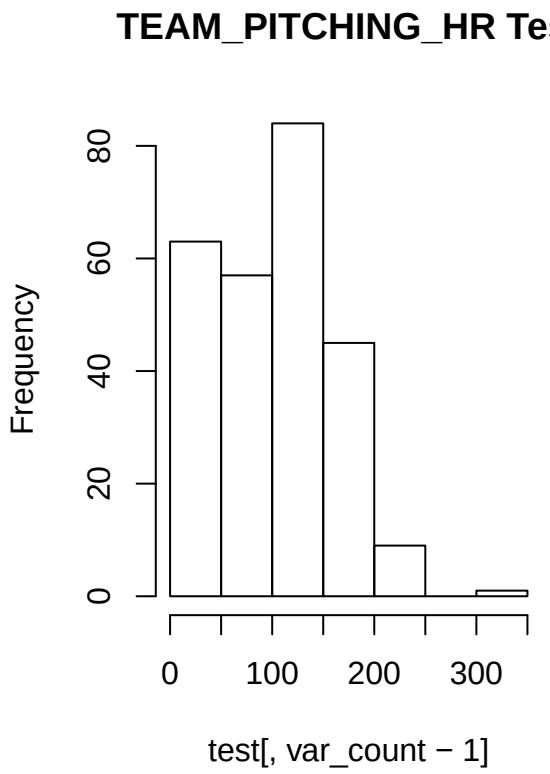

TEAM_BATTING_HBP
Frequency
0
10
20
30
40
50
60
20
40
60
80
100
trainingdata[, var_count]

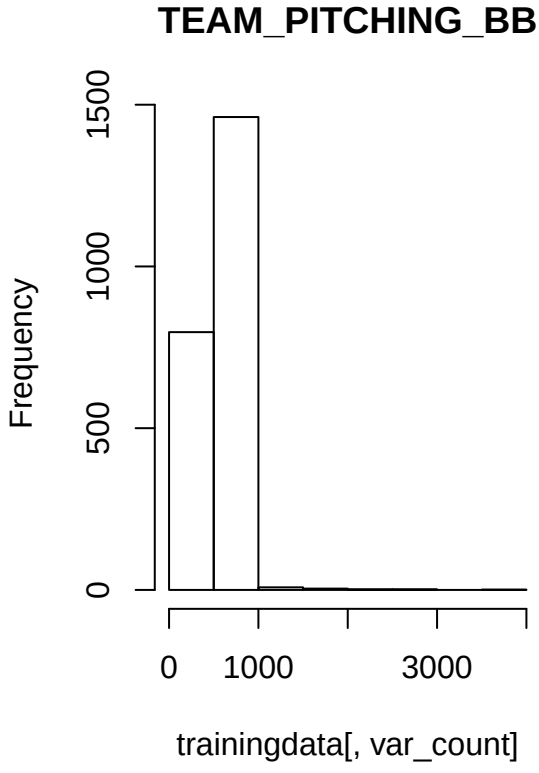

TEAM_BATTING_HBP Test
Frequency
0
2
4
6
8
40
60
80
100
test[, var_count − 1]

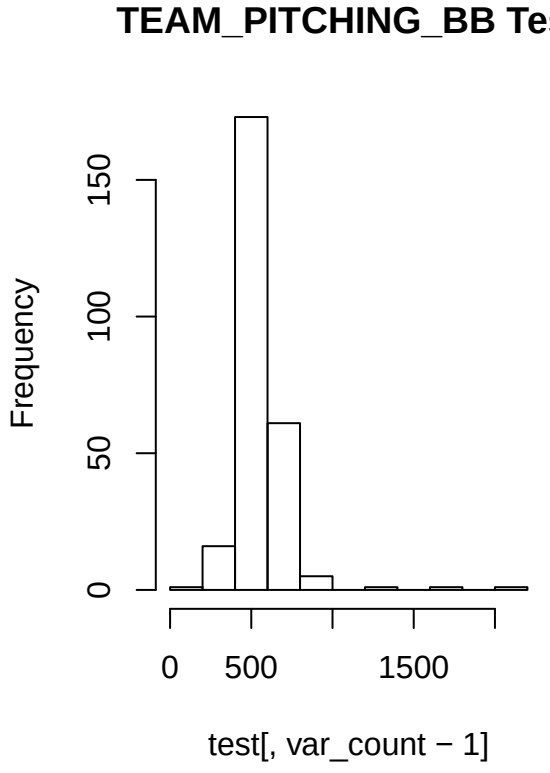

TEAM_PITCHING_H
Frequency
0
500
1000
1500
2000
0
10000
25000
trainingdata[, var_count]

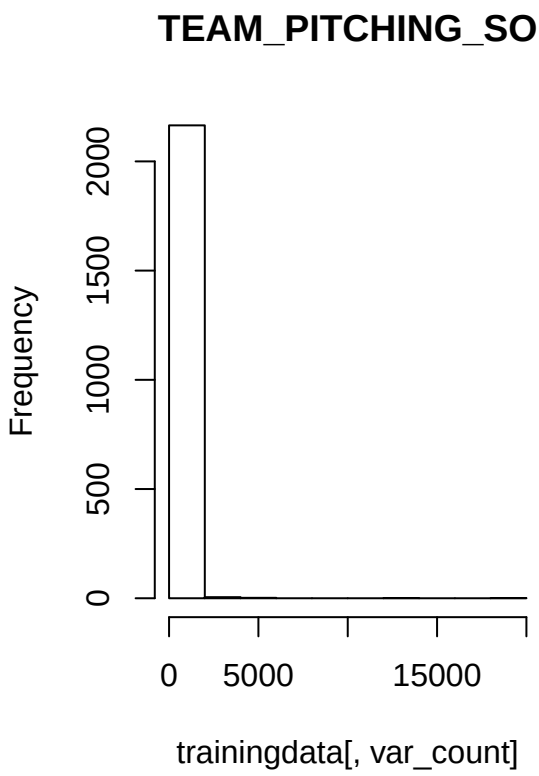

TEAM_PITCHING_H Test
Frequency
0
50
100
150
200
0
5000
15000
test[, var_count − 1]

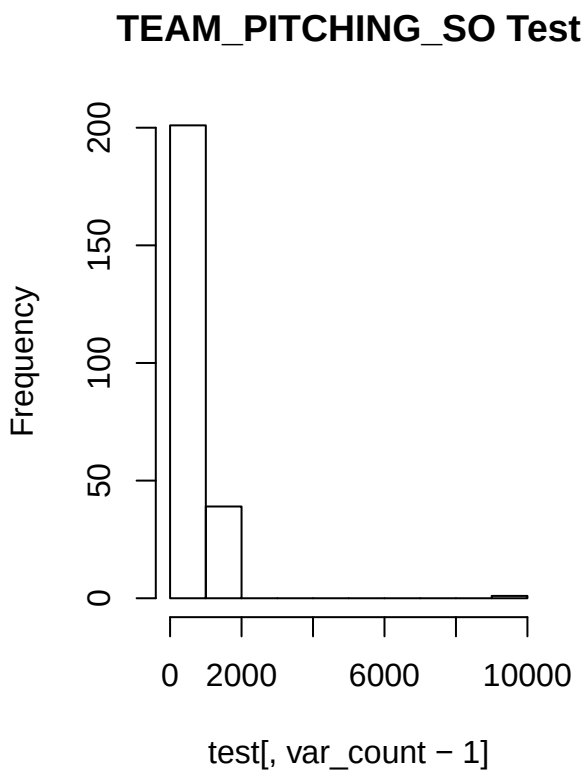

TEAM_PITCHING_HR
Frequency
0
50
100
150
200
250
0
100
200
300
trainingdata[, var_count]

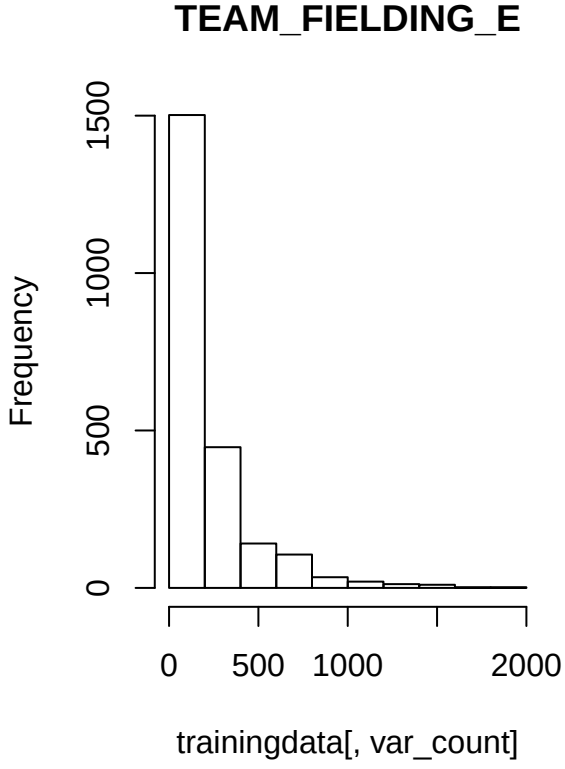

TEAM_PITCHING_HR Test
Frequency
0
20
40
60
80
0
100
200
300
test[, var_count − 1]

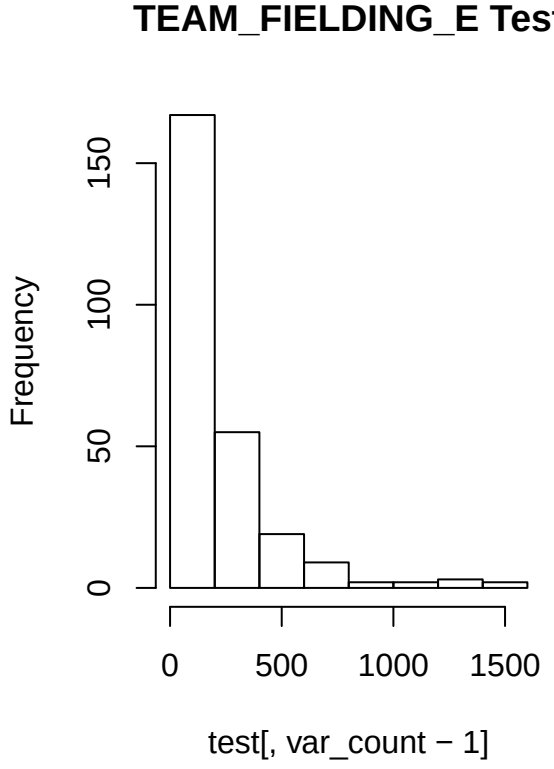

TEAM_PITCHING_BB
Frequency
0
500
1000
1500
0
1000
3000
trainingdata[, var_count]

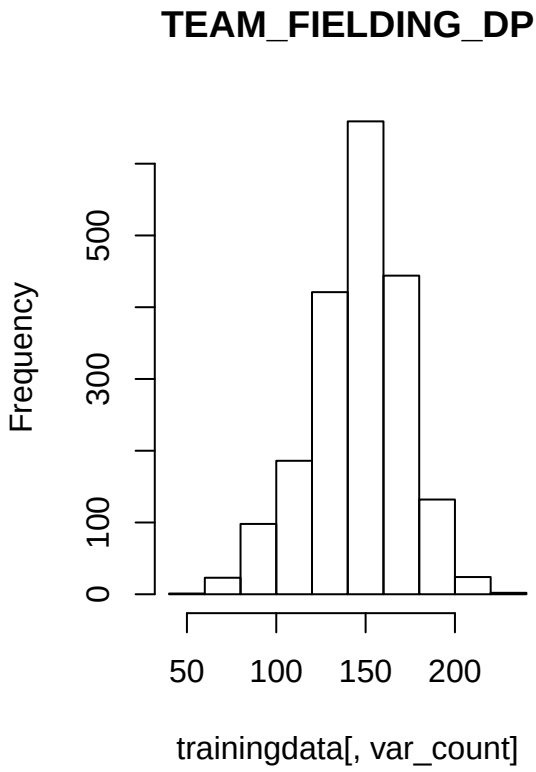

TEAM_PITCHING_BB Test
Frequency
0
50
100
150
0
500
1500
test[, var_count − 1]

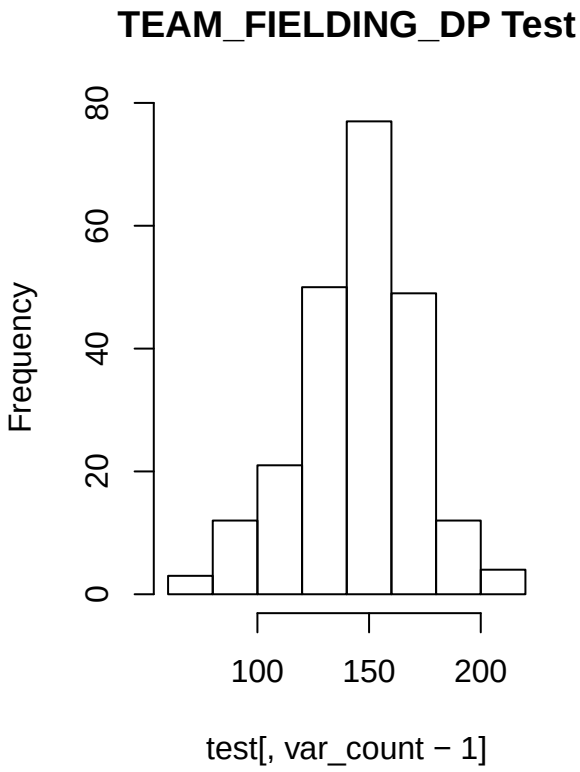

TEAM_PITCHING_SO
Frequency
0
500
1000
1500
2000
0
5000
15000
trainingdata[, var_count]


TEAM_PITCHING_SO Test
Frequency
0
50
100
150
200
0
2000
6000
10000
test[, var_count − 1]


TEAM_FIELDING_E
Frequency
0
500
1000
1500
0
500
1000
2000
trainingdata[, var_count]


TEAM_FIELDING_E Test
Frequency
0
50
100
150
0
500
1000
1500
test[, var_count − 1]


TEAM_FIELDING_DP
Frequency
0
100
300
500
50
100
150
200
trainingdata[, var_count]


TEAM_FIELDING_DP Test
Frequency
0
20
40
60
80
100
150
200
test[, var_count − 1]