CUNY SPS DATA 621 - CTG5 - HW1

Betsy Rosalen, Gabrielle Bartomeo, Jeremy O'Brien, Lidiia Tronina, Rose Koh

February 27, 2019

Contents

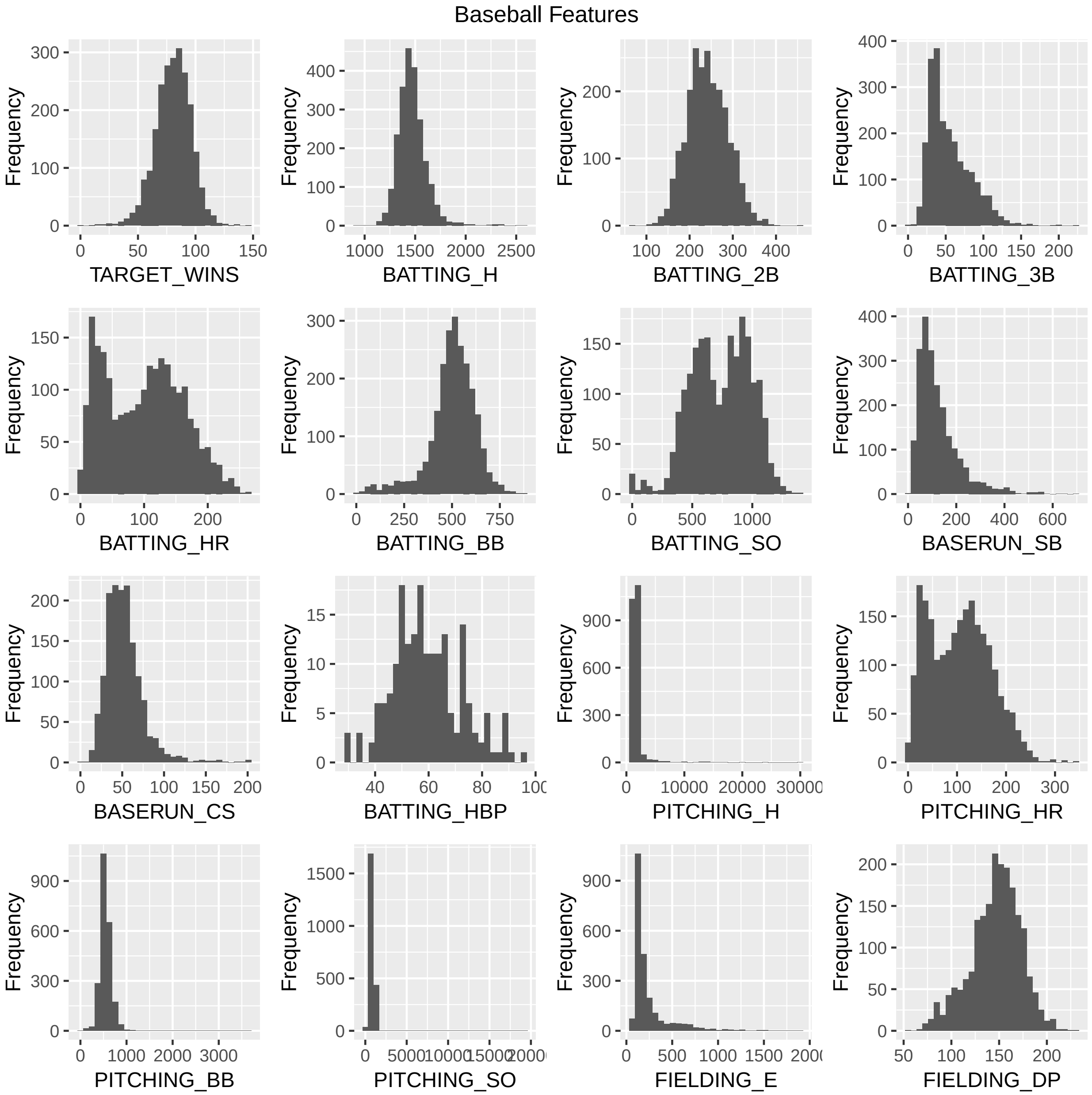
1 Data exploration

1.1 Possible writeup for Data Exploration

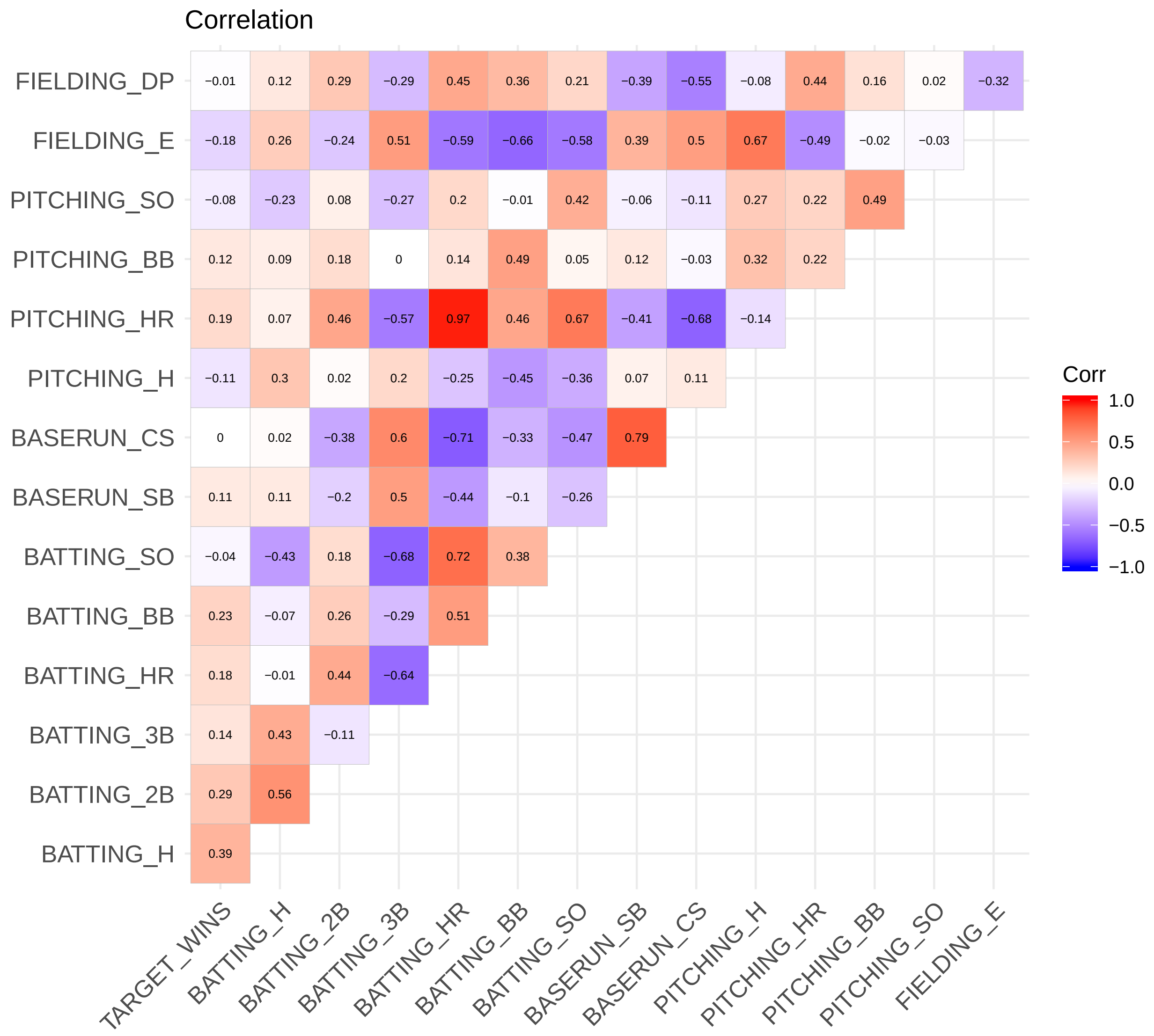
Professionals and gamblers alike are always seeking to optimize their chances of winning, whether it be sports, games, or their bets on them. Major League Baseball is a multibillion dollar industry where individual teams, players, and those who profit off of their success stand to benefit most from such optimization.

In order to determine the best way to infer whether the 162 games in a baseball team's year will result in more wins overall, data from 1871 to 2006 where each set of values represented a season for an unnamed team, totalling 2,276 records. For each team their number of wins in a given year were given with a maximum possible of 162 wins, in addition to that team's base hits, doubles, triples, homeruns, walks, and strikeouts by batters, batters hit by pitches, bases stolen by batters and the number of times they were caught stealing, the number of errors, double plays, walks, hits, and homeruns allowed, and strikeouts by pitchers.

Of all the observations gathered across these fifteen variables, 10.187% were missing information; batters hit by pitches was missing the most, with 2,085 instances of missing information. The standard deviation of the various variables hints at the intense skewing of some of the variables provided, especially the hits allowed, walks allowed and strike outs.



The theoretical effect of strikeouts by batters, batters caught stealing, errors, walks, hits, and homeruns allowed were believed to theoretically have a negative impact on the number of wins of an individual team in a given year. A closer look at the correlation between the variables painted a different picture.



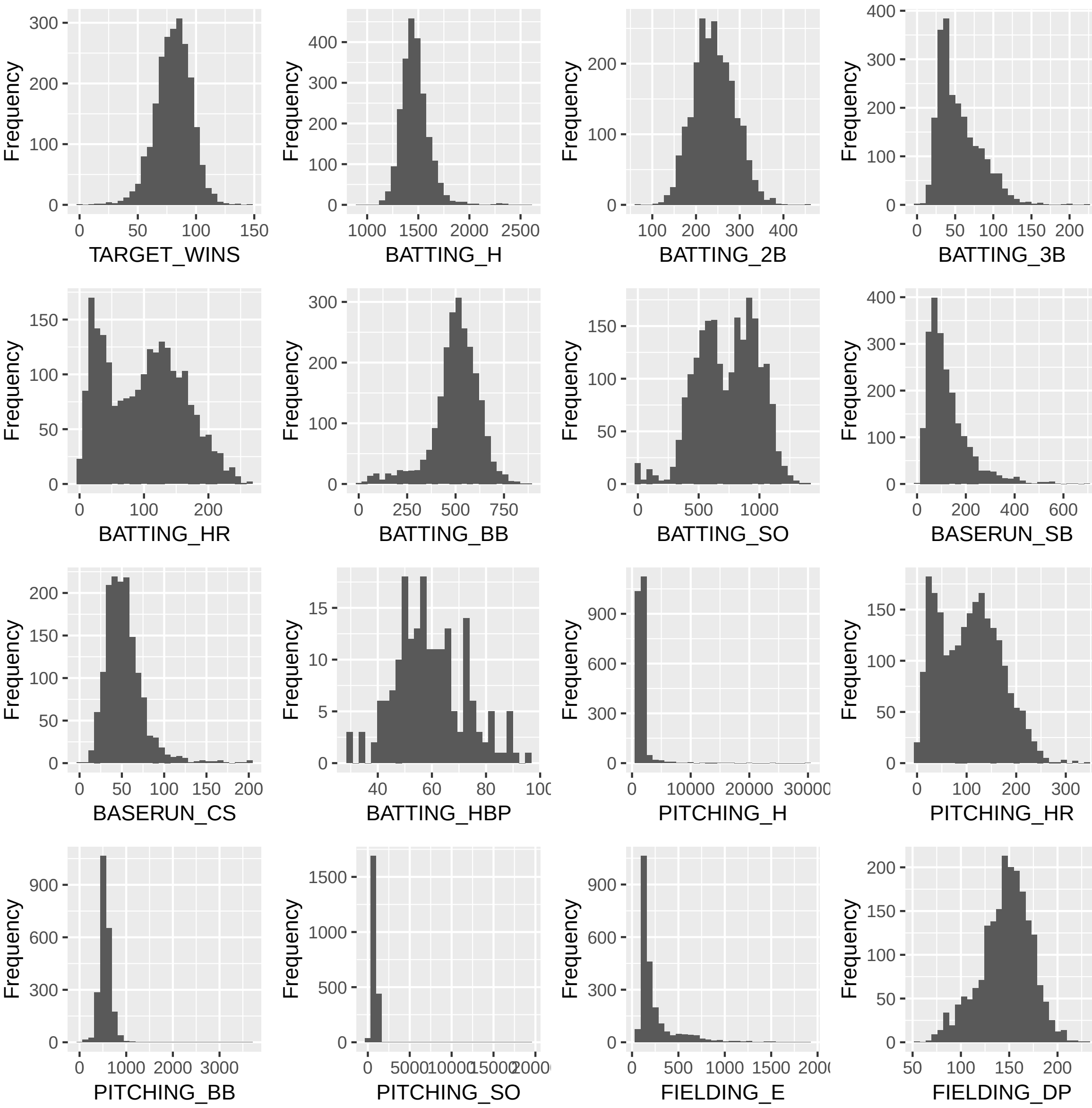
When compared to what was hypothesized, there was actually a positive impact for the number of wins for a team in a given year by walks, hits, and homeruns allowed; at the same time, variables previously thought to have a positive correlation - strikeouts by pitchers and double plays - had a negative correlation for the number of wins. The three variables with the greatest correlation to the number of wins were the hits allowed, the walks by batters, and the walks allowed. Of these, the hits allowed had a relatively low correlation with the walks by batters and the walks allowed, whereas the walks allowed and the walks by batters had a direct positive correlation with one another.

- Describe the size:

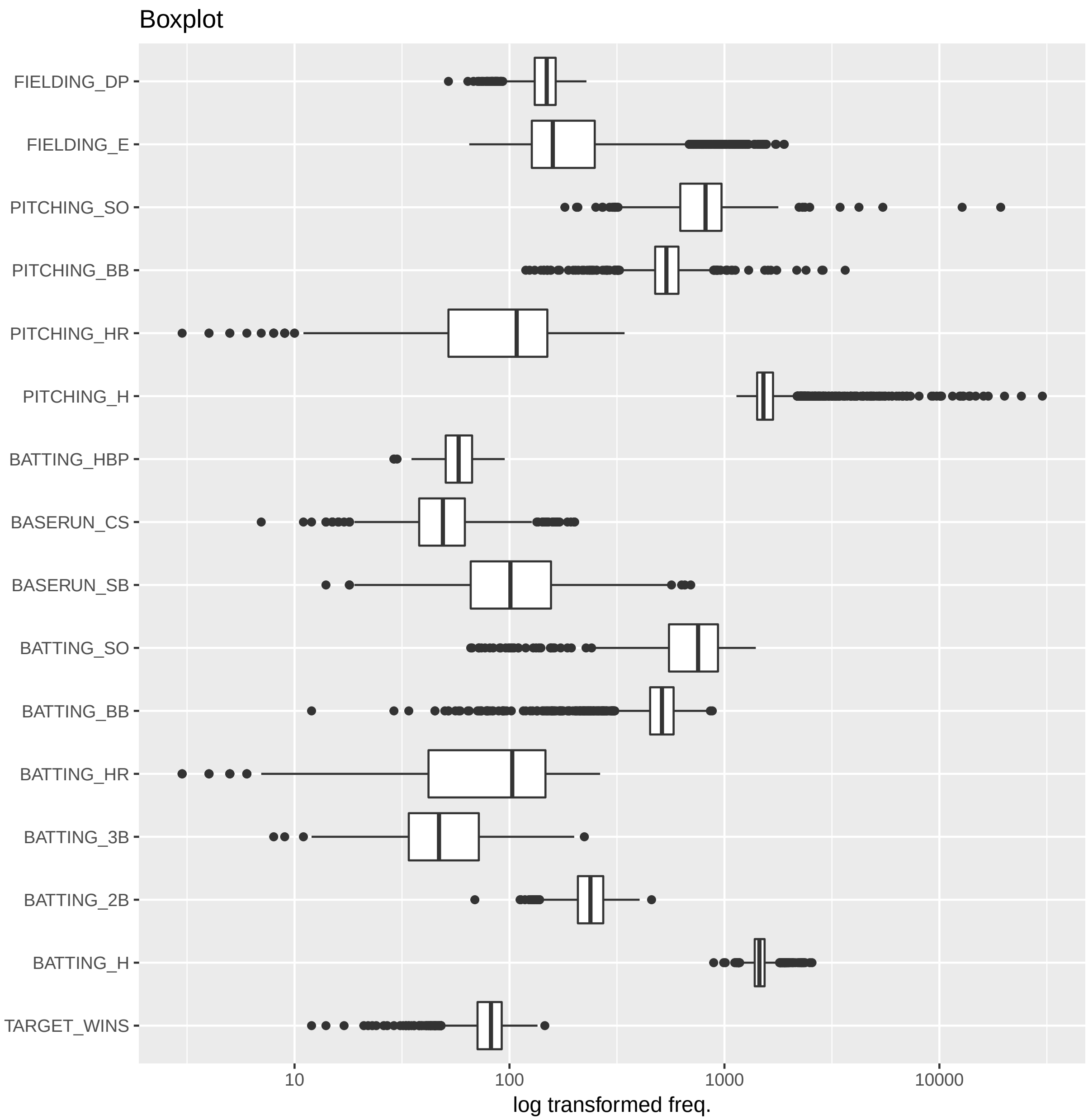
The money ball data is 144kb in size. The data contains 2,276 rows and 16 columns without the index. The variables are continuous integer. The `TARGET_WINS` is our response variable. There are 3,478 missing values out of 36,416 observations.

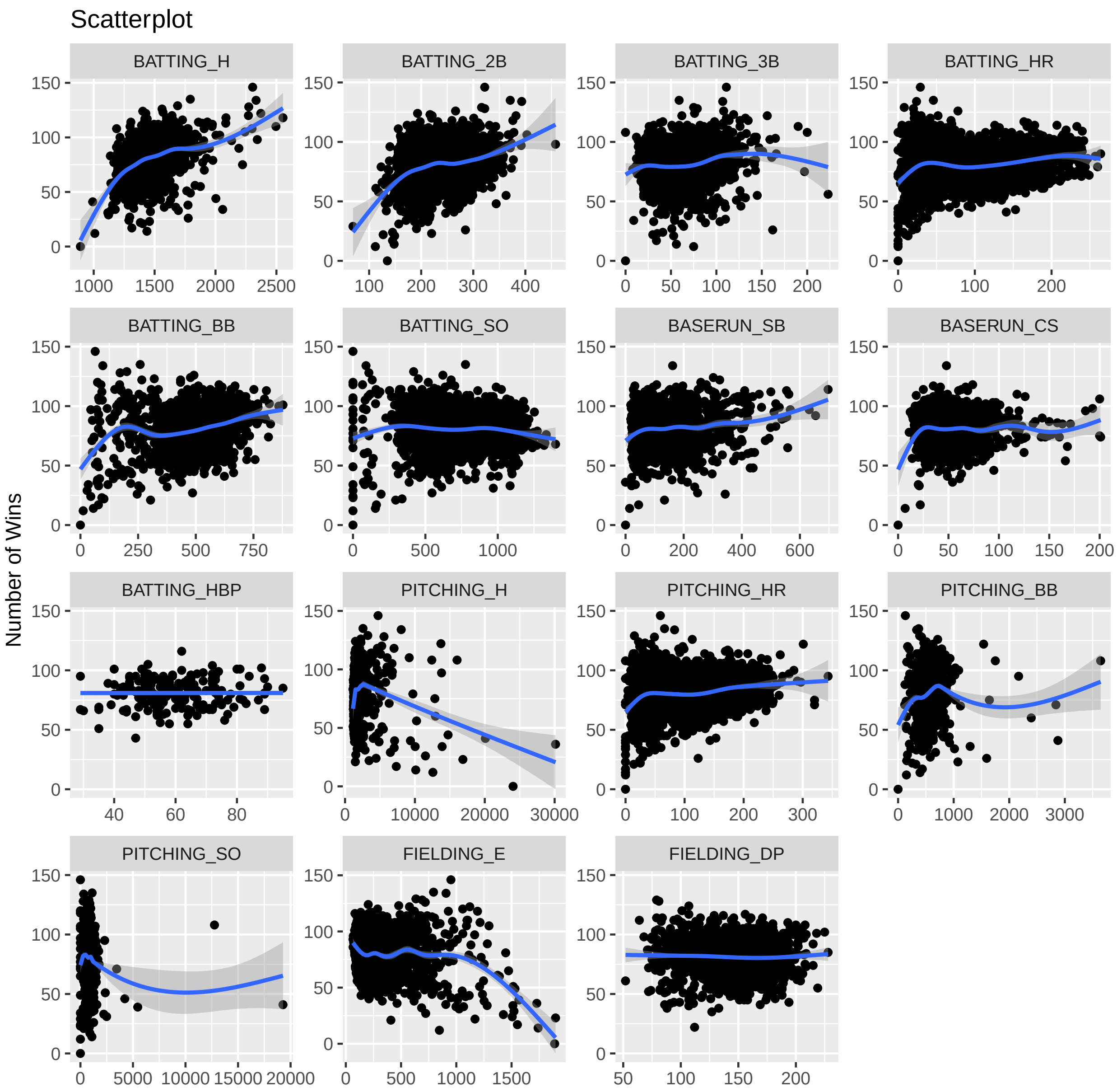
- Statistics summary

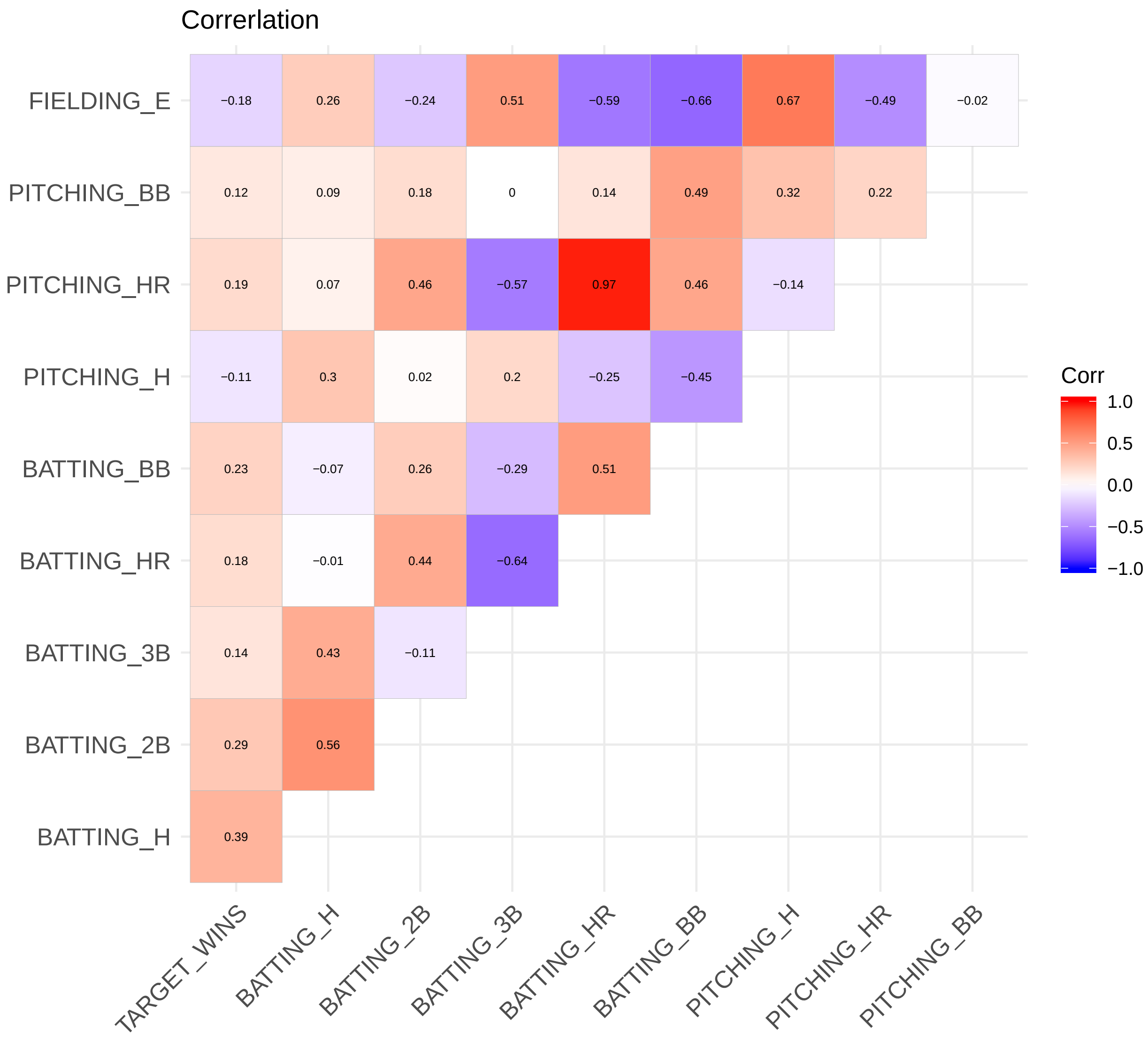
var	s	n	mean	sd med	ian	trimmed	mad mi	n m	ax ran	ge	sk
TARGET_WINS	1	2276	80.79086	15.75215	82.0	81.31229	14.8260	0	146	146	-0.39
BATTING_H	2	2276	1469.26977	144.59120	1454.0	1459.04116	114.1602	891	2554	1663	1.57
BATTING_2B	3	2276	241.24692	46.80141	238.0	240.39627	47.4432	69	458	389	0.21
BATTING_3B	4	2276	55.25000	27.93856	47.0	52.17563	23.7216	0	223	223	1.10
BATTING_HR	5	2276	99.61204	60.54687	102.0	97.38529	78.5778	0	264	264	0.18
BATTING_BB	6	2276	501.55888	122.67086	512.0	512.18331	94.8864	0	878	878	-1.02
BATTING_SO	7	2174	735.60534	248.52642	750.0	742.31322	284.6592	0	1399	1399	-0.29
BASERUN_SB	8	2145	124.76177	87.79117	101.0	110.81188	60.7866	0	697	697	1.97
BASERUN_CS	9	1504	52.80386	22.95634	49.0	50.35963	17.7912	0	201	201	1.97
BATTING_HBP	10	191	59.35602	12.96712	58.0	58.86275	11.8608	29	95	66	0.31
PITCHING_H	11	2276	1779.21046	1406.84293	1518.0	1555.89517	174.9468	1137	30132	28995	10.32
PITCHING_HR	12	2276	105.69859	61.29875	107.0	103.15697	74.1300	0	343	343	0.28
PITCHING_BB	13	2276	553.00791	166.35736	536.5	542.62459	98.5929	0	3645	3645	6.74
PITCHING_SO	14	2174	817.73045	553.08503	813.5	796.93391	257.2311	0	19278	19278	22.17
FIELDING_E	15	2276	246.48067	227.77097	159.0	193.43798	62.2692	65	1898	1833	2.99
FIELDING_DP	16	1990	146.38794	26.22639	149.0	147.57789	23.7216	52	228	176	-0.38

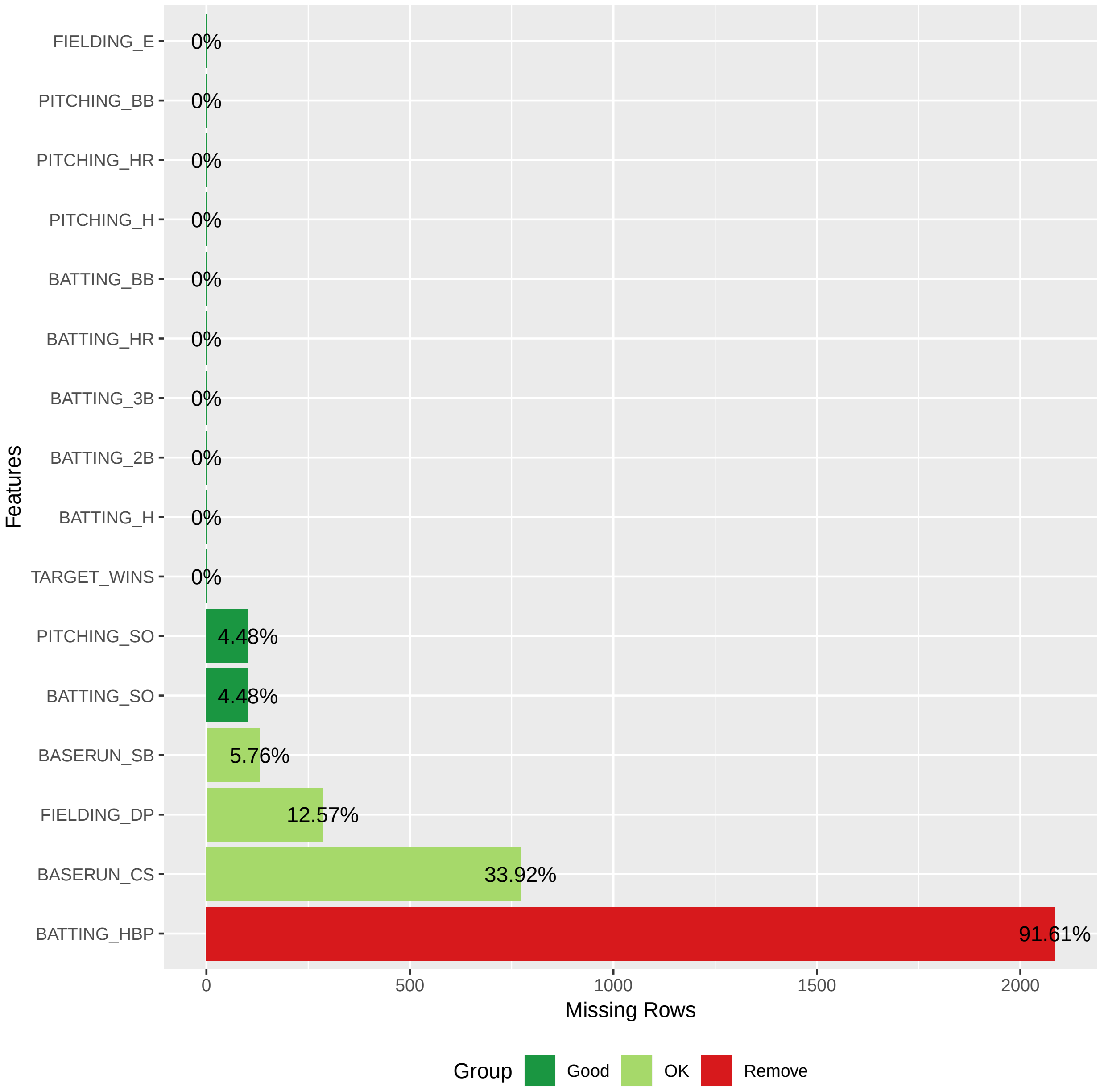


- Data visualization


Boxplot
FIELDING_DP
FIELDING_E
PITCHING_SO
PITCHING_BB
PITCHING_HR
PITCHING_H
BATTING_HBP
BASERUN_CS
BASERUN_SB
BATTING_SO
BATTING_BB
BATTING_HR
BATTING_3B
BATTING_2B
BATTING_H
TARGET_WINS
10
100
1000
10000
log transformed freq.


Scatterplot
Number of Wins
BATTING_H
BATTING_2B
BATTING_3B
BATTING_HR
BATTING_BB
BATTING_SO
BASERUN_SB
BASERUN_CS
BATTING_HBP
PITCHING_H
PITCHING_HR
PITCHING_BB
PITCHING_SO
FIELDING_E
FIELDING_DP
0
50
100
150
1000
1500
2000
2500
100
200
300
400
0
50
100
150
200
0
100
200
0
250
500
750
0
500
1000
0
200
400
600
0
50
100
150
200
40
60
80
0
10000
20000
30000
0
100
200
300
0
1000
2000
3000
0
5000
10000
15000
20000
0
500
1000
1500
50
100
150
200


Correrlation
FIELDING_E −0.18 0.26 −0.24 0.51 −0.59 −0.66 0.67 −0.49 −0.02
PITCHING_BB 0.12 0.09 0.18 0 0.14 0.49 0.32 0.22
PITCHING_HR 0.19 0.07 0.46 −0.57 0.97 0.46 −0.14
PITCHING_H −0.11 0.3 0.02 0.2 −0.25 −0.45
BATTING_BB 0.23 −0.07 0.26 −0.29 0.51
BATTING_HR 0.18 −0.01 0.44 −0.64
BATTING_3B 0.14 0.43 −0.11
BATTING_2B 0.29 0.56
BATTING_H 0.39
TARGET_WINS
BATTING_H
BATTING_2B
BATTING_3B
BATTING_HR
BATTING_BB
PITCHING_H
PITCHING_HR
PITCHING_BB
Corr
1.0
0.5
0.0
−0.5
−1.0



2 Data preparation

2.1 Possible writeup for Data Preparation

As previously mentioned, just north of 10% of the data was missing values. Missing values can lead to errors in a model, bias, and worse if left unaccounted for. Attempting to "fix" this by imputing values or guessing why the values are missing in the first place - such as concluding that the missing values are meant to be zeroes - are just as likely to help with creating a model as it is to help with creating a disaster.

One of the R packages utilized, DataExplorer, recommends removing null or missing values; it was for this reason all observations of hits by pitch were removed.<< SOURCE FOR THIS IS REQUIRED! IF NO SOURCE IS PROVIDED, CONSIDER USING "Due to the sheer volume of missing values present in the observations for hits by pitch (91.61%) it was determined the best course of action was to remove the variable altogether.” OR A VARIATION THEREOF. >> Deleting all cases with missing values, in this instance, would have shrunk the size of the dataset down to less than a tenth of its original size. For this reason, the feature itself was excluded from the dataset, rather than the cases that had no values present for it.

The other missing values - present in batting strikeouts… needs more work. x_x

2.2 Information in general, don’t use the writeups here as-is

2.3 Missing Values

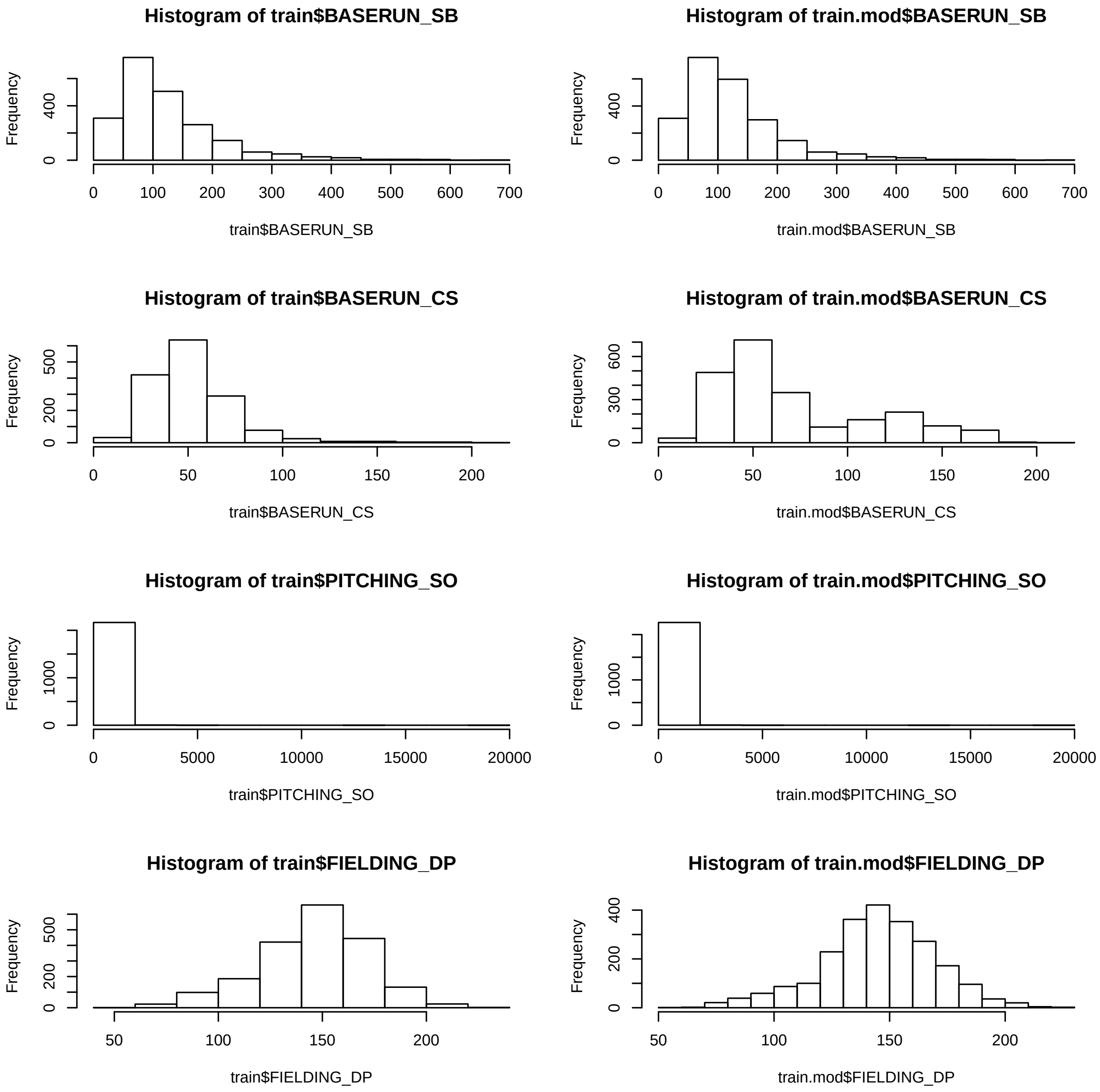
1) `Hit by pitch` missing 91.61% .

- Missing values can lead to errors and bias into a model. Fixing and imputation may help or make it worse.
- When it is just a few observations missing, modifications can be made, however, with 91.61% is a large proportion and could distort the modelling later on that it is better to ignore this column.
- The Data explorer package recommends to remove.
- From LMR: Missing Completely at Random (MCAR) The probability that a value is missing is the same for all cases. If we simply delete all cases with missing values from the analysis, we will cause no bias, although we may lose some information.
- However, there is no consensus on when to exclude missing data. Some argue that missing data more than 10% can lead to bias. Others argue that missing data patterns have greater impact than the proportion.

2) `Pitching_SO` and `Batting_SO` are missing exact same proportion 4.48% and are missing in the same observations.

2.4 NA Imputation

```
## TARGET_WINS   BATTING_H  BATTING_2B  BATTING_3B  BATTING_HR  BATTING_BB
##           0           0           0           0           0           0
##  BATTING_SO  BASERUN_SB  BASERUN_CS  PITCHING_H PITCHING_HR PITCHING_BB
##         102         131         772           0           0           0
## PITCHING_SO  FIELDING_E FIELDING_DP
##         102           0         286
```


Histogram of train$BASERUN_SB

Histogram of train.mod$BASERUN_SB

Histogram of train$BASERUN_CS

Histogram of train.mod$BASERUN_CS

Histogram of train$PITCHING_SO

Histogram of train.mod$PITCHING_SO

Histogram of train$FIELDING_DP

Histogram of train.mod$FIELDING_DP

2.5 Feature Engineering