Moneyball training data exploration

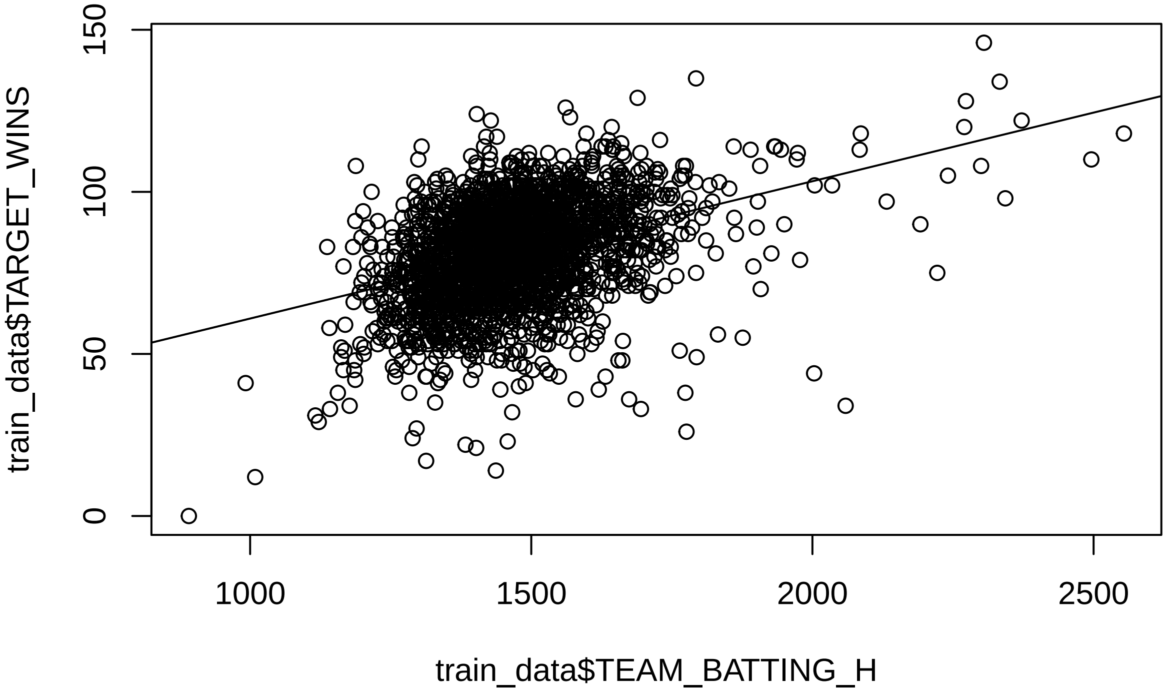
Arindam

June 11th, 2016

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     891    1383    1454    1469    1537    2554 
```

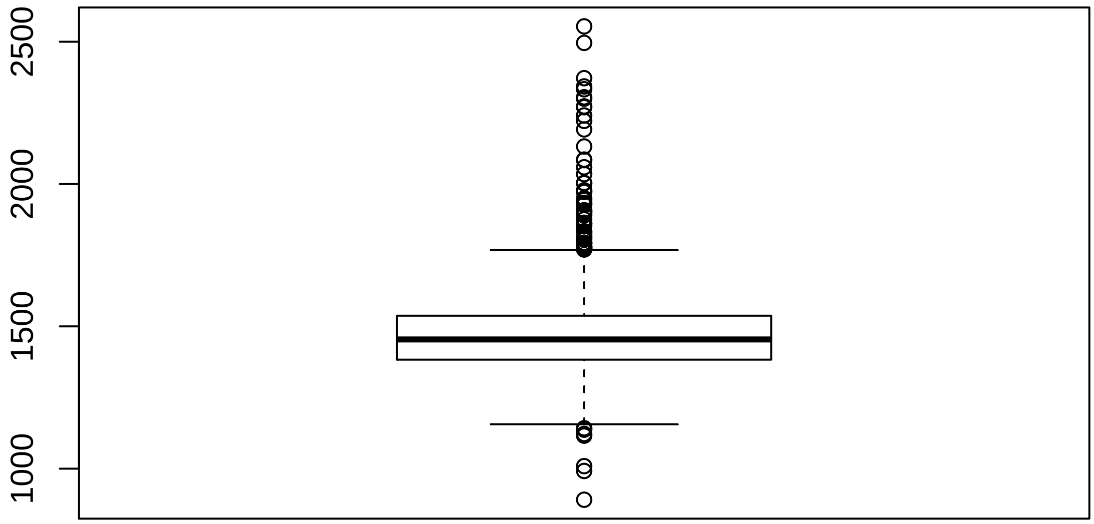
Analysis with TEAM_BATTING_H

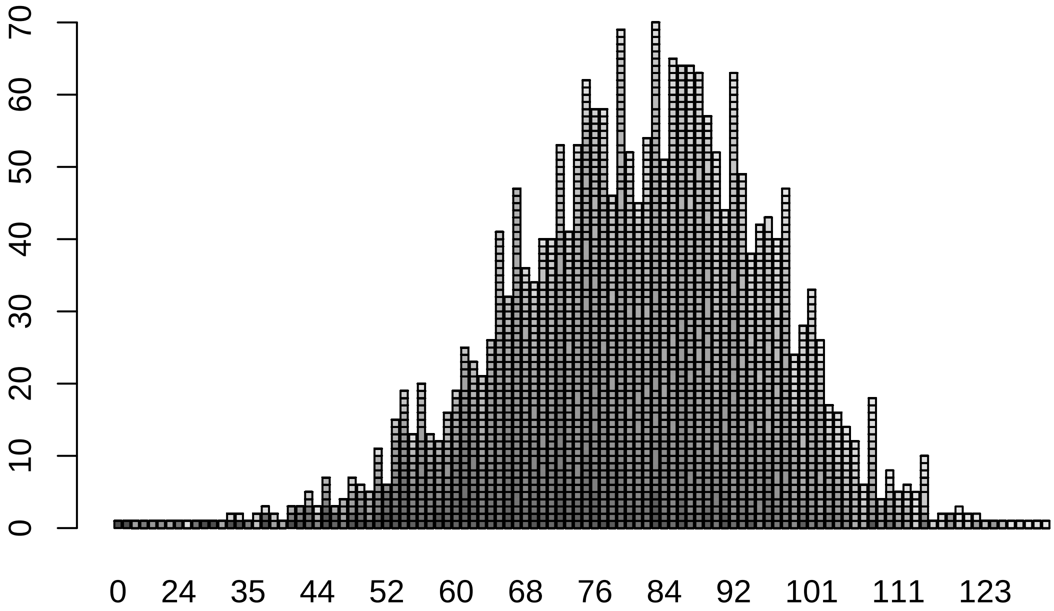
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     891    1383    1454    1469    1537    2554 
```



```
## 
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H, data = train_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -71.768  -8.757   0.856   9.762  46.016 
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    18.562326   3.107523   5.973 2.69e-09 ***
## TEAM_BATTING_H  0.042353   0.002105  20.122  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.52 on 2274 degrees of freedom
## Multiple R-squared:  0.1511, Adjusted R-squared:  0.1508
## F-statistic: 404.9 on 1 and 2274 DF,  p-value: < 2.2e-16
```

```
## [1] 0.3887675
```

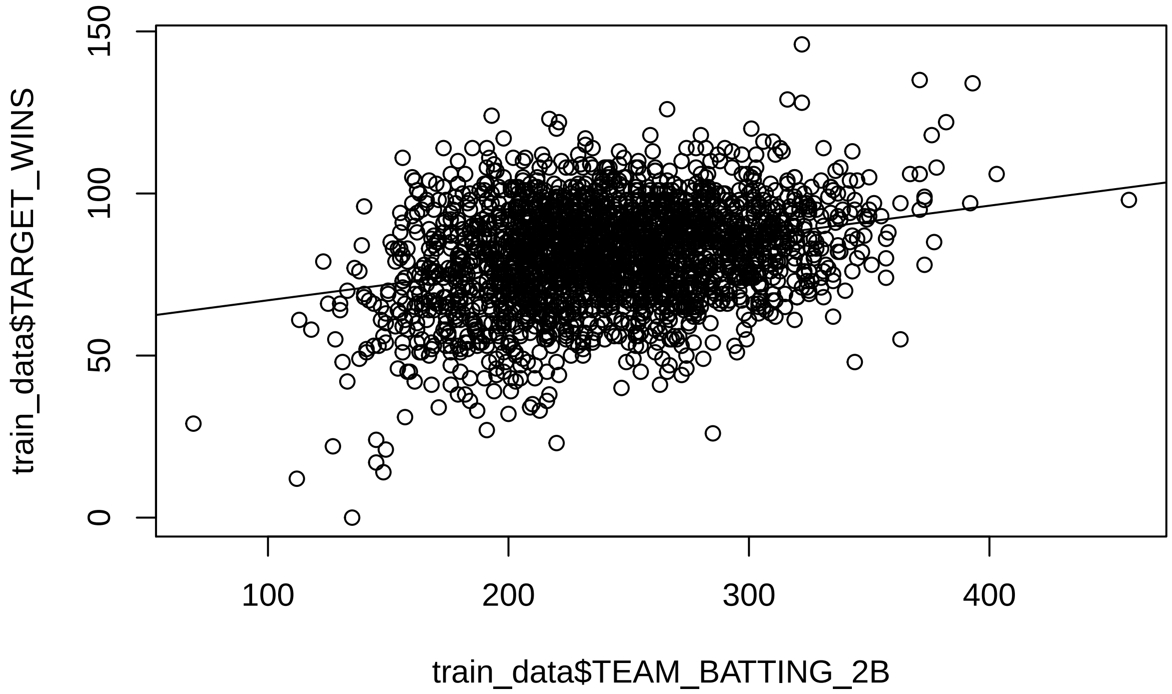




looking at the COR and regression line looks like this variable can be used as input variable to the model. Also looking at the box plot it appears outliers need to be handled for better result.

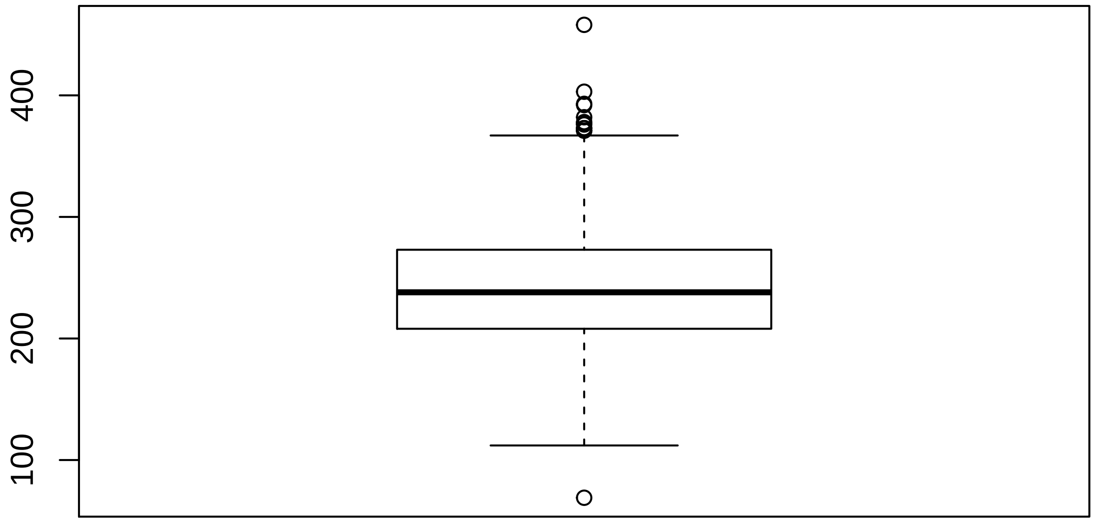
Analysis with TEAM_BATTING_2B

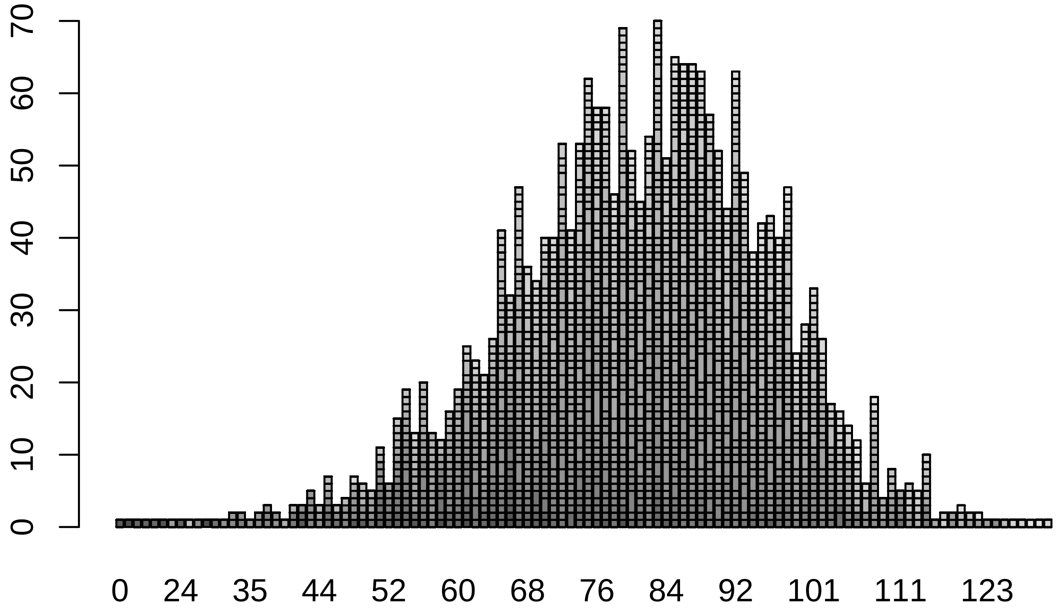
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    69.0   208.0   238.0   241.2   273.0   458.0
```



```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_2B, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -70.453  -9.572   0.636  10.135  57.351
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     57.316365   1.660403   34.52   <2e-16 ***
## TEAM_BATTING_2B  0.097305   0.006757   14.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.08 on 2274 degrees of freedom
## Multiple R-squared:  0.08358,    Adjusted R-squared:  0.08318
## F-statistic: 207.4 on 1 and 2274 DF,  p-value: < 2.2e-16
```

```
## [1] 0.2891036
```

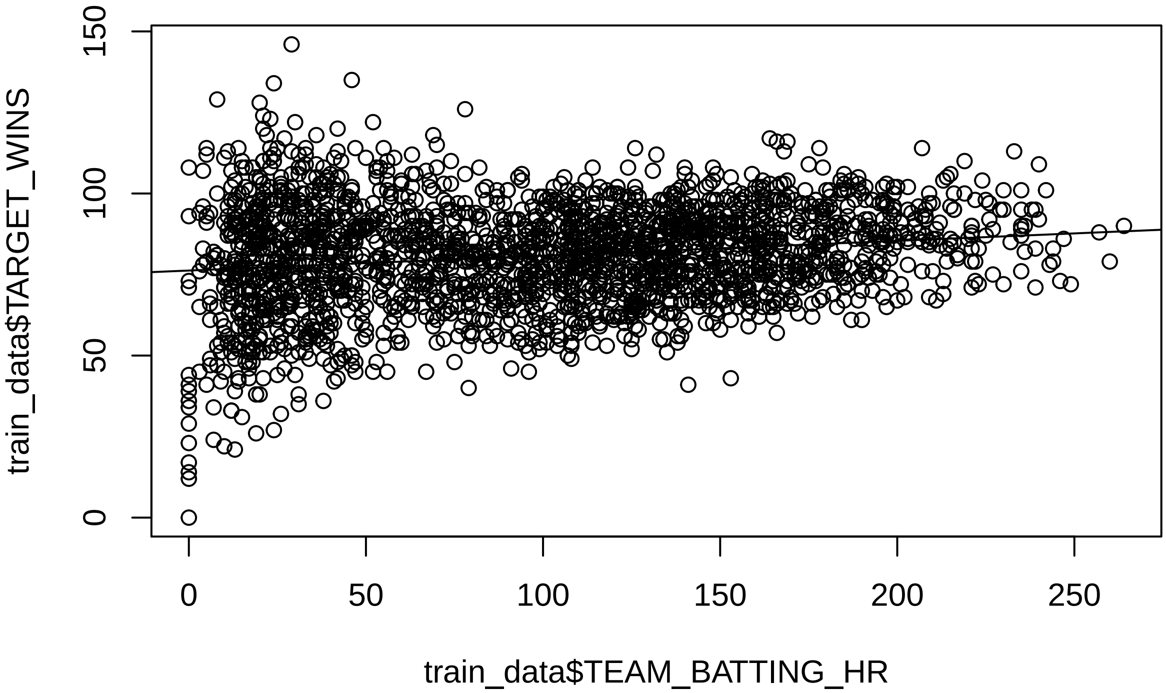

100
200
300
400



## This variable has COR coefficient 0.28 and can be used as predictor variable. This variabe has long right tail and will need some outlier handling

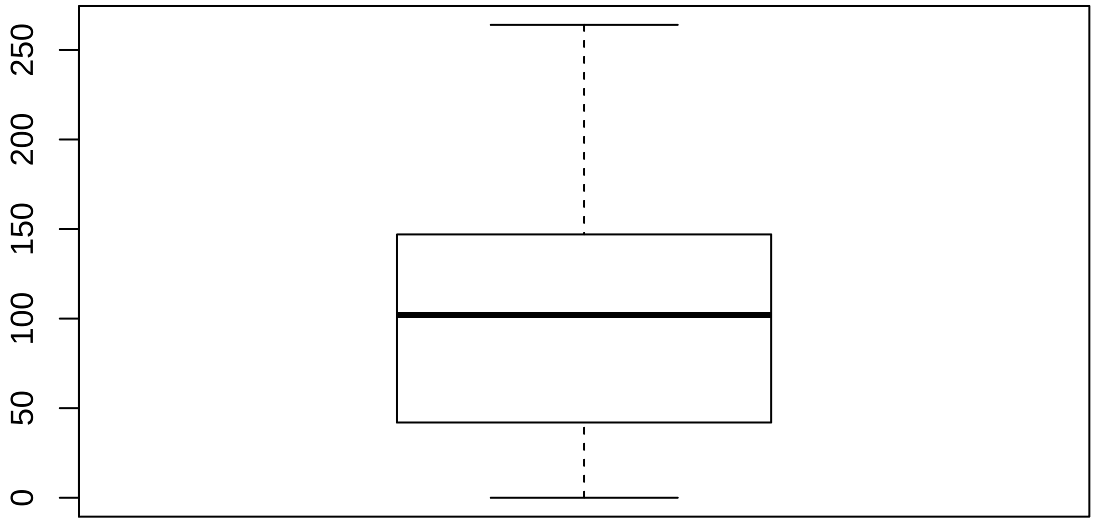
Analysis with TEAM_BATTING_HR

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   42.00  102.00   99.61  147.00  264.00 
```



```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_HR, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.226  -9.909   0.520  10.218  68.445
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     76.22576    0.62599 121.768   <2e-16 ***
## TEAM_BATTING_HR  0.04583    0.00537   8.534   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.51 on 2274 degrees of freedom
## Multiple R-squared:  0.03103,    Adjusted R-squared:  0.0306
## F-statistic: 72.82 on 1 and 2274 DF,  p-value: < 2.2e-16
```

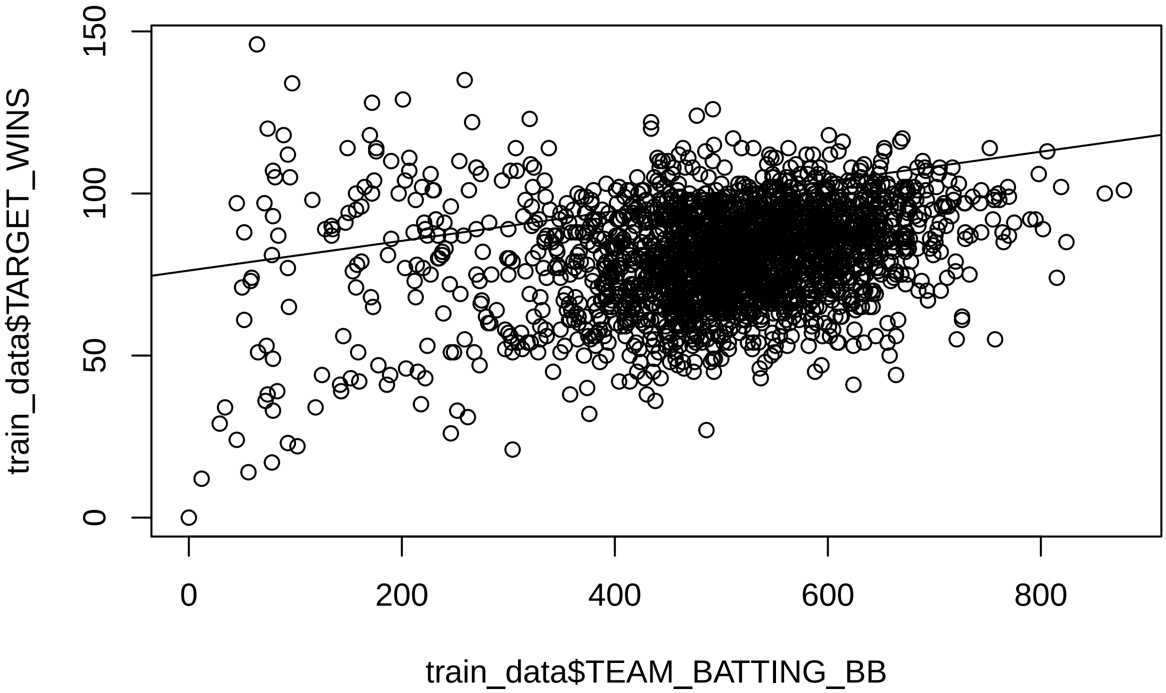
```
## [1] 0.1761532
```



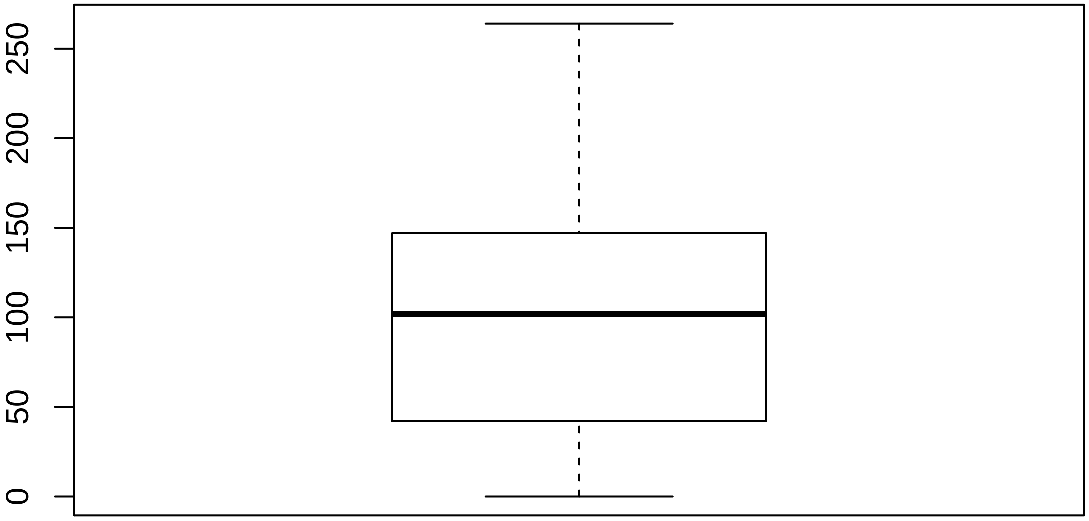
## This variable is very interesting as there are cases where team has less home run but higher wins and also there are teams who have more home runs but less wins. COR is 0.17, as this is an important factors in a baaseball match i would like to use this variable in the prediction model.

Analysis with TEAM_BATTING_BB

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0   451.0   512.0   501.6   580.0   878.0
```



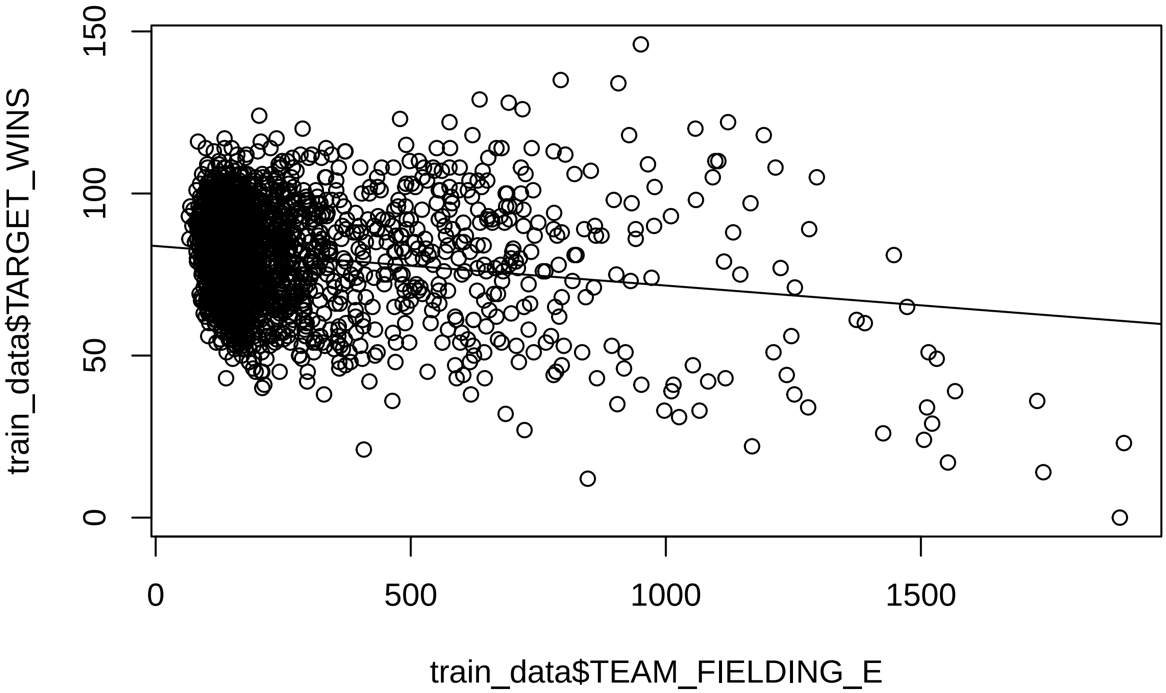
```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_HR, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.226  -9.909   0.520  10.218  68.445
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     76.22576    0.62599 121.768   <2e-16 ***
## TEAM_BATTING_HR  0.04583    0.00537   8.534   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.51 on 2274 degrees of freedom
## Multiple R-squared:  0.03103,    Adjusted R-squared:  0.0306
## F-statistic: 72.82 on 1 and 2274 DF,  p-value: < 2.2e-16

## [1] 0.1761532
```



Analysis of the chart indicates that relationship this variable can be used a predictor with COR 0.17

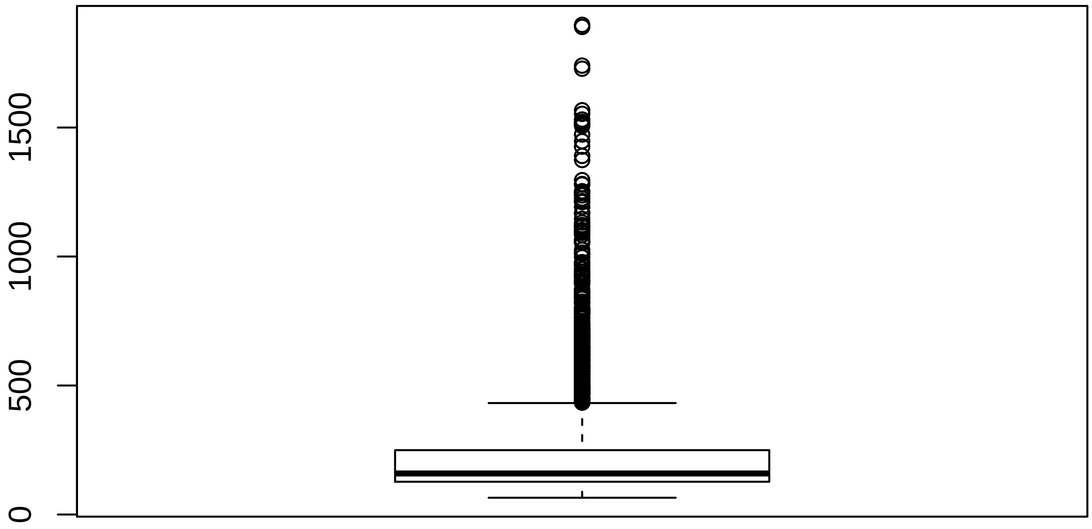
Analysis with TEAM_FIELDING_E

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    65.0   127.0   159.0   246.5   249.2  1898.0
```



```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_FIELDING_E, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.461 -10.078   0.697  10.318  73.808
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     83.799234   0.479030  174.94   <2e-16 ***
## TEAM_FIELDING_E -0.012205   0.001427   -8.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.51 on 2274 degrees of freedom
## Multiple R-squared:  0.03115,    Adjusted R-squared:  0.03072
## F-statistic:  73.1 on 1 and 2274 DF,  p-value: < 2.2e-16
```

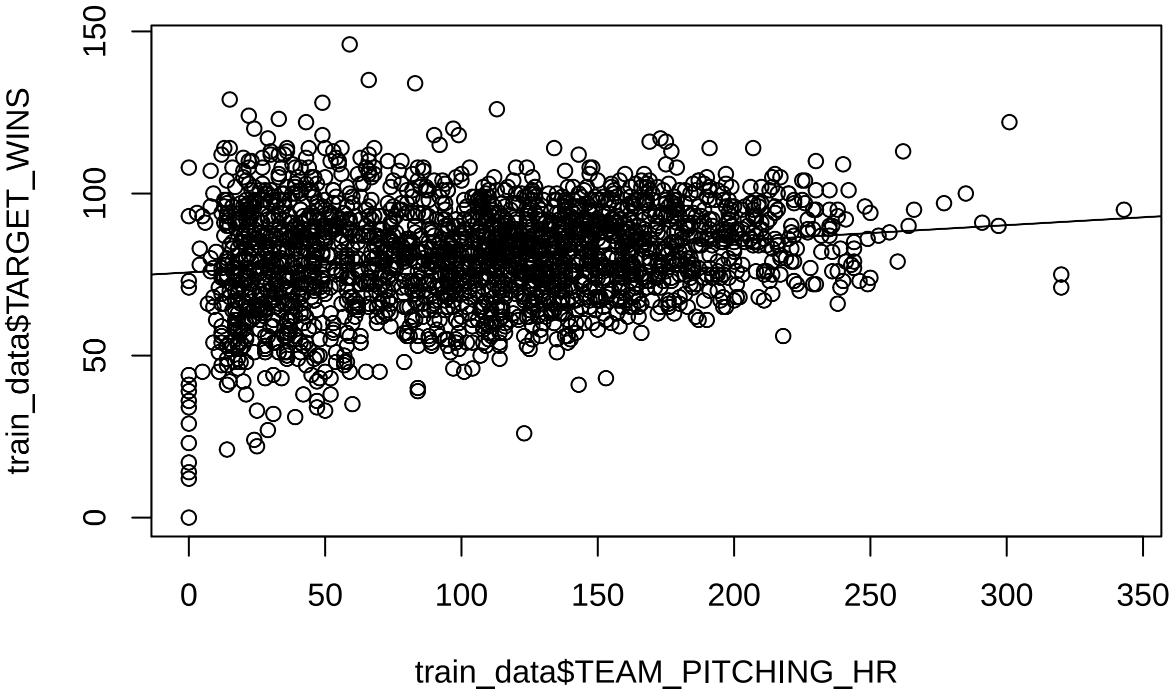
```
## [1] -0.1764848
```



## this variables clearly shows increase in error rates leads to nagative impact on win number and has COR -0.17 and can be selected for the model. Also this variable is very skewed in one side and needed outlier exclusion

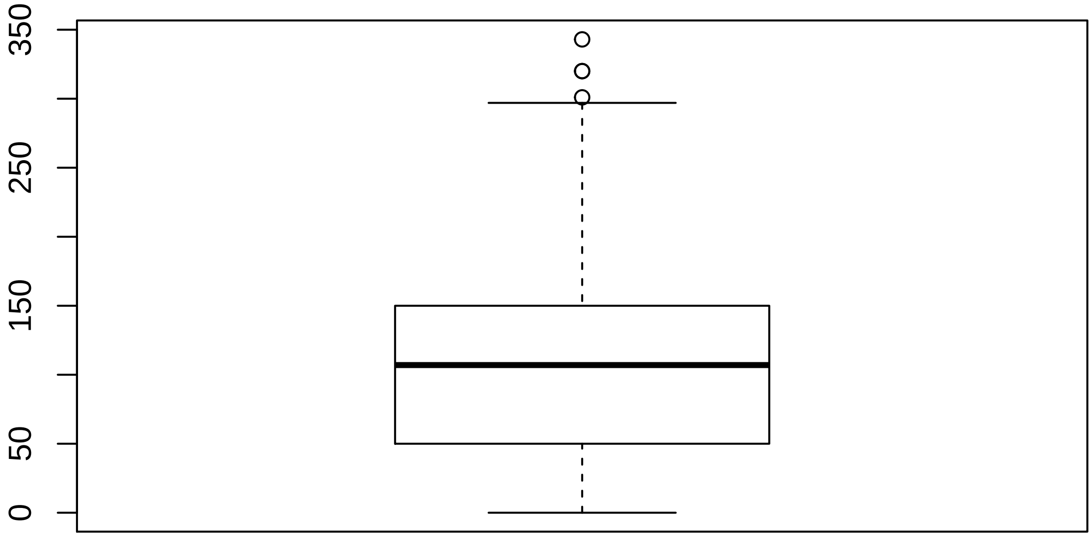
Analysis with TEAM_PITCHING_HR

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    50.0   107.0   105.7   150.0   343.0
```



```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_PITCHING_HR, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.657  -9.956   0.636  10.055  67.477
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      75.656920   0.646540 117.018   <2e-16 ***
## TEAM_PITCHING_HR  0.048572   0.005292   9.179   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.47 on 2274 degrees of freedom
## Multiple R-squared:  0.03573,    Adjusted R-squared:  0.0353
## F-statistic: 84.25 on 1 and 2274 DF,  p-value: < 2.2e-16
```

```
## [1] 0.1890137
```



Looking at chart it looks like concentration of points are towards the left upper side of the chart which implies lesser hits allowed higher win number. Also this variable has cor 0.18 and selected as predictor for the model.