Notes_for_Moneyball

Exploratory analysis of money-ball.csv

These are some notes on the data

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#load data
setwd('C:/Program Files/GitHub/business_analytics')
dat<-read.csv('./data/moneyball-training-data.csv',header = TRUE)%>%select(-INDEX)
dat<-dat%>%select(-TEAM_BATTING_HBP)
```

This data is probably not real data.

I noticed while looking at *TEAM_BATTING_SO* that there were 64 non NA data points that were below the all time record for lowest team strikeouts, as well as many SO totals being zero, which is impossible.

fewest strikeouts all time 1921 Cincinnati Reds : 308

https://www.baseball-almanac.com/recbooks/rb_strike2.shtml

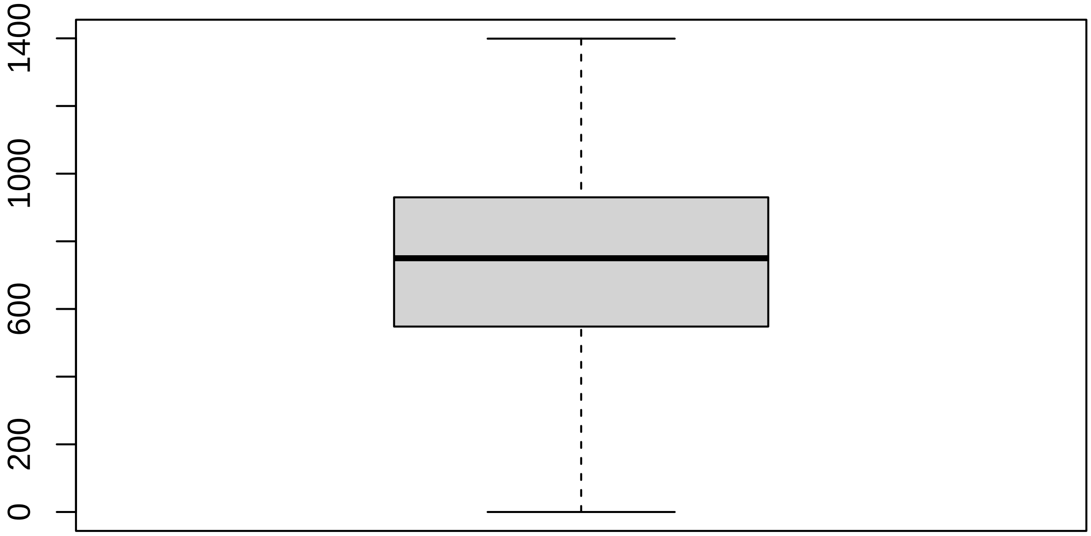
```
dat<-dat%>%filter(!is.na(TEAM_BATTING_SO))

a<-dat%>%filter(TEAM_BATTING_SO<308)%>%count()
b<-dat%>%filter(TEAM_BATTING_SO>1595)%>%count()
data.frame('SO below all time'=as.integer(a),'SO above all time'=as.integer(b))
```

```
##   SO.below.all.time SO.above.all.time
## 1                64                 0
```

Note that they are not counted as outliers

```
out_vals<-boxplot(dat$TEAM_BATTING_SO)$out
```



```
out_vals
```

```
## numeric(0)
```

In this particular column, I wonder if they arent counted as outliers due to the bimodal nature of the distribution.

Having sliced up TEAM_BATTING_SO a couple of ways, I do not think that it will correlate with TEAM_WINS even if we can disentangle the factors causing the bimodality, but this should raise our suspicions of other features.

NET_SB has higher correlation than TEAM_BASERUN_SB + TEAM_BASERUN_CS

not a major difference, but might be worth putting in.

```
dat_sb<-dat%>%mutate(TEAM_NET_SB=TEAM_BASERUN_SB-TEAM_BASERUN_CS)
dat_sb<-dat_sb%>%filter(!is.na(TEAM_BASERUN_CS & TEAM_BASERUN_SB))
sb<-round(cor(dat_sb$TARGET_WINS,dat_sb$TEAM_BASERUN_SB),3)
cs<-round(cor(dat_sb$TARGET_WINS,dat_sb$TEAM_BASERUN_CS),3)
sb_cs<-sb+cs
data.frame('NET_SB'=round(cor(dat_sb$TARGET_WINS,dat_sb$TEAM_NET_SB),3),'SB+CS'=sb_cs,'SB'=sb,'CS'=cs)
```

```
##   NET_SB SB.CS    SB    CS
## 1  0.185 0.176 0.154 0.022
```

#Total bases with linear weights to replace TEAM_BATTING_H, TEAM_BATTING_2B, TEAM_BATTING_3B…etc

I don't know if this is kosher, but if I use general linear weights developed by `baseball prospectus` and add all the hitting columns into a total bases column, I get a nice strong correlation. (I can get a similar correlation with simpler weights (2B=2,3B=3,4B=4))

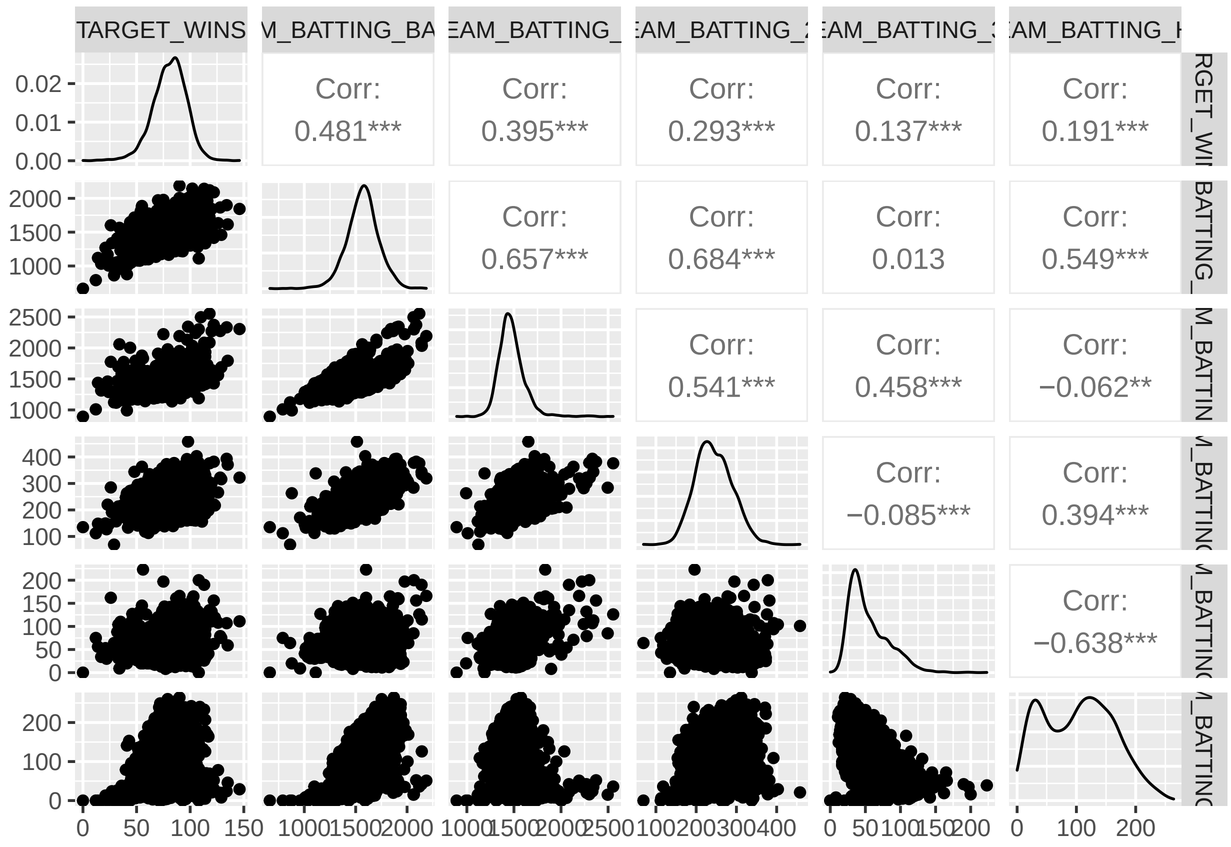
This might save us big on the Rˆ2 by using less features.

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
dat_tb<-dat%>%mutate(TEAM_BATTING_BASES=.7*(TEAM_BATTING_H-TEAM_BATTING_2B-TEAM_BATTING_3B-TEAM_BATTING_

dat_tb<-dat_tb%>%select(TARGET_WINS,TEAM_BATTING_BASES,TEAM_BATTING_H,TEAM_BATTING_2B,TEAM_BATTING_3B,T
ggpairs(dat_tb)
```



LM with these adjustments

```
dat_final<-dat%>%mutate(TEAM_NET_SB=TEAM_BASERUN_SB-TEAM_BASERUN_CS)
dat_final<-dat_final%>%mutate(TEAM_BATTING_BASES=.7*(TEAM_BATTING_H-TEAM_BATTING_2B-TEAM_BATTING_3B-TEAM
#dat_final<-dat_final%>%drop_na()
lm_1<-lm(TARGET_WINS~TEAM_BATTING_BASES+TEAM_NET_SB,data=dat_final)
summary(lm_1)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_BASES + TEAM_NET_SB,
##     data = dat_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.528  -8.054   0.252   7.863  35.526
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -5.044189   3.540020  -1.425    0.154
## TEAM_BATTING_BASES  0.052301   0.002217  23.588  < 2e-16 ***
## TEAM_NET_SB         0.055771   0.008542   6.529 9.02e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 11.31 on 1501 degrees of freedom
##   (670 observations deleted due to missingness)
## Multiple R-squared:  0.2955, Adjusted R-squared:  0.2945
## F-statistic: 314.8 on 2 and 1501 DF,  p-value: < 2.2e-16
```