R code

```
#libraries
library(mice)
library(knitr)
library(VIM)
library(readr)
library(readr)
library(leaps)
library(MASS)
#importing data and setting working direct.
setwd("~/Desktop/NW MSDS/MSDS411/Unit 1/Moneyball/")
moneyball=read.csv("moneyball.csv",header=T)
#Exploratory Data Analysis
str(moneyball)
head(moneyball)
summary(moneyball)
correlations <- cor(moneyball)
round(correlations,2)
par(mfrow=c(1,2))
hist(moneyball$TARGET_WINS, col = "#A71930", xlab = "TARGET_WINS", main = "Histogram of
Wins")
boxplot(moneyball$TARGET_WINS, col = "#A71930", main = "Boxplot of Wins")
par(mfrow = c(1,1))
##
#Batting
par(mfrow=c(2,2))
hist(moneyball$TEAM_BATTING_H, col = "#A71930", xlab = "Team_Batting_H", main =
"Histogram of Hits")
hist(moneyball$TEAM_BATTING_2B, col = "#09ADAD", xlab = "Doubles", main = "Histogram of
Doubles")
boxplot(moneyball$TEAM_BATTING_H, col = "#A71930", main = "Boxplot of Hits")
boxplot(moneyball$TEAM_BATTING_2B, col = "#09ADAD", main = "Boxplot of Doubles")
par(mfrow=c(1,1))
##
par(mfrow=c(2,2))
hist(moneyball$TEAM_BATTING_3B, col = "#A71930", xlab = "Triples", main = "Histogram of
Triples")
hist(moneyball$TEAM_BATTING_HR, col = "#DBCEAC", xlab = "Home Runs", main = "Histogram
of Home Runs")
boxplot(moneyball$TEAM_BATTING_3B, col = "#A71930", main = "Boxplot of Triples")
boxplot(moneyball$TEAM_BATTING_HR, col = "#DBCEAC", main = "Boxplot of Home Runs")
par(mfrow=c(1,1))
##
```

```
par(mfrow=c(2,3))
hist(moneyball$TEAM_BATTING_BB, col = "#A71930", xlab = "Walks", main = "Histogram of
Walks")
hist(moneyball$TEAM_BATTING_SO, col = "#09ADAD", xlab = "Strikeouts", main = "Histogram
of Strikeouts")
hist(moneyball$TEAM_BATTING_HBP, col = "#DBCEAC", xlab = "Hit By Pitches", main =
"Histogram of HBP")
boxplot(moneyball$TEAM_BATTING_BB, col = "#A71930", main = "Boxplot of Walks")
boxplot(moneyball$TEAM_BATTING_SO, col = "#09ADAD", main = "Boxplot of Strikeouts")
boxplot(moneyball$TEAM_BATTING_HBP, col = "#DBCEAC", main = "Boxplot of HBP")
par(mfrow=c(1,1))
##
par(mfrow=c(2,2))
hist(moneyball$TEAM_BASERUN_SB, col = "#A71930", xlab = "Stolen Bases", main = "Histogram
of Steals")
hist(moneyball$TEAM_BASERUN_CS, col = "#DBCEAC", xlab = "Caught Stealing", main =
"Histogram of CS")
boxplot(moneyball$TEAM_BASERUN_SB, col = "#A71930", main = "Boxplot of Steals")
boxplot(moneyball$TEAM_BASERUN_CS, col = "#DBCEAC", main = "Boxplot of CS")
par(mfrow=c(1,1))
##
#pitching
par(mfrow=c(2,2))
hist(moneyball$TEAM_PITCHING_H, col = "#A71930", xlab = "Hits Against", main = "Histogram
of Hits Against")
hist(moneyball$TEAM_PITCHING_HR, col = "#09ADAD", xlab = "Home Runs Against", main =
"Histograms of HR Against")
boxplot(moneyball$TEAM_PITCHING_H, col = "#A71930", main = "Boxplot of Hits Against")
boxplot(moneyball$TEAM_PITCHING_HR, col = "#09ADAD", main = "Boxplot of HR Against")
par(mfrow=c(1,1))
##
par(mfrow=c(2,2))
hist(moneyball$TEAM_PITCHING_BB, col = "#A71930", xlab = "Walks Allowed", main =
"Histogram of Walks Allowed")
hist(moneyball$TEAM_PITCHING_SO, col = "#DBCEAC", xlab = "Strikeouts", main = "Histograms
of Strikeouts")
boxplot(moneyball$TEAM_PITCHING_BB, col = "#A71930", main = "Boxplot of Walks Allowed")
boxplot(moneyball$TEAM_PITCHING_SO, col = "#DBCEAC", main = "Boxplot of Strikeouts")
par(mfrow=c(1,1))
##
#fielding
par(mfrow=c(2,2))
hist(moneyball$TEAM_FIELDING_DP, col = "#A71930", xlab = "Double Plays", main =
"Histogram of Double Plays")
```

```
hist(moneyball$TEAM_FIELDING_E, col = "#09ADAD", xlab = "Errors Committed", main =
"Histogram of Errors Committed")
boxplot(moneyball$TEAM_FIELDING_DP, col = "#A71930", main = "Boxplot of Double Plays")
boxplot(moneyball$TEAM_FIELDING_E, col = "#09ADAD", main = "Boxplot of Errors
Committed")
par(mfrow=c(1,1))
##
#scatter matrix

panel.cor <- function(x, y, digits=2, prefix="", cex.cor, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits=digits)[1]
  txt <- paste(prefix, txt, sep="")
  if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r)
}

# Batting Stats and Wins
pairs(moneyball[2:8], lower.panel=panel.smooth, upper.panel = panel.cor)

#Baserunning  Stats and Wins
pairs(~ moneyball$TARGET_WINS + moneyball$TEAM_BASERUN_CS +
moneyball$TEAM_BASERUN_SB, lower.panel = panel.smooth)

#Pitcher Stats and Wins
pairs(~ moneyball$TARGET_WINS + moneyball$TEAM_PITCHING_BB +
moneyball$TEAM_PITCHING_H +
      moneyball$TEAM_PITCHING_HR + moneyball$TEAM_PITCHING_SO, lower.panel =
panel.smooth)

#mice package
head(moneyball)
summary(moneyball)
md.pattern(moneyball)
m <- md.pairs(moneyball);m
pbox(moneyball,pos=1,int=FALSE,cex=0.7)
imp <- mice(moneyball)
imp
head(complete(imp))
moneyballfilled <- complete(imp)
#capOutlier <- function(x){
```

```
  #qnt <- quantile(x, probs=c(.25, .75), na.rm = T)
  #caps <- quantile(x, probs=c(.01, .99), na.rm = T)
  #H <- 1.5 * IQR(x, na.rm = T)
  #x[x < (qnt[1] - H)] <- caps[1]
  #x[x > (qnt[2] + H)] <- caps[2]
  #return(x)
#}
#moneyballfilled$TARGET_WINS=capOutlier(moneyballfilled$TARGET_WINS)
#moneyballfilled$TEAM_BATTING_H=capOutlier(moneyballfilled$TEAM_BATTING_H)
#moneyballfilled$TEAM_BATTING_2B=capOutlier(moneyballfilled$TEAM_BATTING_2B)
#moneyballfilled$TEAM_BATTING_3B=capOutlier(moneyballfilled$TEAM_BATTING_3B)
#moneyballfilled$TEAM_BATTING_HR=capOutlier(moneyballfilled$TEAM_BATTING_HR)
#moneyballfilled$TEAM_BATTING_BB=capOutlier(moneyballfilled$TEAM_BATTING_BB)
#moneyballfilled$TEAM_BATTING_SO=capOutlier(moneyballfilled$TEAM_BATTING_SO)
#moneyballfilled$TEAM_BASERUN_SB=capOutlier(moneyballfilled$TEAM_BASERUN_SB)
#moneyballfilled$TEAM_BATTING_HBP=capOutlier(moneyballfilled$TEAM_BATTING_HBP)
#moneyballfilled$TEAM_PITCHING_H=capOutlier(moneyballfilled$TEAM_PITCHING_H)
#moneyballfilled$TEAM_PITCHING_HR=capOutlier(moneyballfilled$TEAM_PITCHING_HR)
#moneyballfilled$TEAM_PITCHING_BB=capOutlier(moneyballfilled$TEAM_PITCHING_BB)
#moneyballfilled$TEAM_PITCHING_SO=capOutlier(moneyballfilled$TEAM_PITCHING_SO)

write.csv(moneyballfilled, file = "moneyballcleanedtrain.csv")
summary(moneyballfilled)
## model creation
mse <- function(sm)
  mean(sm$residuals^2)
# model 1
model1 <- lm(TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_BB + TEAM_BASERUN_SB
+ TEAM_PITCHING_HR +
               TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_PITCHING_H +
TEAM_PITCHING_BB, data = moneyballfilled)
summary(model1)
vif(model1)

# model 2 (stepwise approach)
stepwisemodel <- lm(formula = TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B +
TEAM_BATTING_HR +
               TEAM_BATTING_H +
               TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
TEAM_BASERUN_CS + TEAM_PITCHING_HR +
               TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDING_E +
TEAM_FIELDING_DP, data = moneyballfilled)
model2 <- stepAIC(stepwisemodel, direction = "both")
summary(model2)
```

```
vif(model2)

# model 3 (subset models)

subsets <- regsubsets(TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B +
TEAM_BATTING_HR +
                TEAM_BATTING_H +
                TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
TEAM_BASERUN_CS + TEAM_PITCHING_HR +
                TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDING_E +
TEAM_FIELDING_DP,  data = moneyballfilled, nvmax = 12, nbest = 1)
subsets
summary(subsets)
plot(subsets, scale="adjr2")

subset1 <- lm(TARGET_WINS ~
           TEAM_BATTING_H, data = moneyballfilled)
summary(subset1)

subset2 <- lm(TARGET_WINS ~
            TEAM_BATTING_H + TEAM_FIELDING_E, data = moneyballfilled)
summary(subset2)

subset3 <- lm(TARGET_WINS ~
            TEAM_BATTING_H + TEAM_FIELDING_E + TEAM_BASERUN_SB, data = moneyballfilled)
summary(subset3)

subset4 <- lm(TARGET_WINS ~
            TEAM_BATTING_H + TEAM_FIELDING_E + TEAM_BASERUN_SB + TEAM_FIELDING_DP,
data = moneyballfilled)
summary(subset4)

subset5 <- lm(TARGET_WINS ~
            TEAM_BATTING_H + TEAM_BATTING_HR + TEAM_BATTING_SO + TEAM_FIELDING_E +
TEAM_BASERUN_SB, data = moneyballfilled)
summary(subset5)

subset6 <- lm(TARGET_WINS ~ TEAM_FIELDING_DP + TEAM_FIELDING_E + TEAM_BATTING_SO
+ TEAM_BASERUN_SB + TEAM_BATTING_HR + TEAM_BATTING_H,
            data = moneyballfilled)
summary(subset6)

subset7 <- lm(TARGET_WINS ~ TEAM_FIELDING_DP + TEAM_FIELDING_E + TEAM_BATTING_SO
+ TEAM_BASERUN_SB + TEAM_BATTING_HR + TEAM_BATTING_H + TEAM_PITCHING_BB,
```

```
          data = moneyballfilled)
summary(subset7)

subset8 <- lm(TARGET_WINS ~ TEAM_FIELDING_DP + TEAM_FIELDING_E + TEAM_BATTING_SO
+ TEAM_BASERUN_SB + TEAM_BATTING_HR + TEAM_BATTING_H + TEAM_PITCHING_BB +
            TEAM_BATTING_2B, data = moneyballfilled)
summary(subset8)

subset9 <- lm(TARGET_WINS ~ TEAM_FIELDING_DP + TEAM_FIELDING_E + TEAM_BATTING_SO
+ TEAM_BASERUN_SB + TEAM_BATTING_HR + TEAM_BATTING_H + TEAM_PITCHING_BB +
            TEAM_BATTING_3B + TEAM_BASERUN_CS , data = moneyballfilled)
summary(subset9)

subset10 <- lm(TARGET_WINS ~ TEAM_FIELDING_DP + TEAM_FIELDING_E +
TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_BATTING_HR + TEAM_BATTING_H +
TEAM_PITCHING_BB +
            TEAM_BATTING_3B + TEAM_BATTING_2B + TEAM_BASERUN_CS , data =
moneyballfilled)
summary(subset10)

subset11 <- lm(TARGET_WINS ~ TEAM_FIELDING_DP + TEAM_FIELDING_E +
TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_BATTING_HR + TEAM_BATTING_H +
TEAM_PITCHING_BB +
             TEAM_BATTING_3B + TEAM_BATTING_2B + TEAM_BASERUN_CS +
TEAM_PITCHING_SO , data = moneyballfilled)
summary(subset11)
vif(subset11)

subset12 <- lm(TARGET_WINS ~ TEAM_FIELDING_DP + TEAM_FIELDING_E +
TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_BATTING_HR + TEAM_BATTING_H +
TEAM_PITCHING_BB +
             TEAM_BATTING_3B + TEAM_BATTING_2B + TEAM_BASERUN_CS +
TEAM_PITCHING_SO + TEAM_BATTING_BB , data = moneyballfilled)
summary(subset12)
vif(subset12)

##########
#metrics
AIC(model1)
AIC(model2)
AIC(subset1)
AIC(subset2)
AIC(subset3)
AIC(subset4)
```

```
AIC(subset5)
AIC(subset6)
AIC(subset7)
AIC(subset8)
AIC(subset9)
AIC(subset10)
AIC(subset11)
AIC(subset12)
mse(model1)
mse(model2)
mse(subset1)
mse(subset2)
mse(subset3)
mse(subset4)
mse(subset5)
mse(subset6)
mse(subset7)
mse(subset8)
mse(subset9)
mse(subset10)
mse(subset11)
mse(subset12)
####
#fixing Test DATA
moneyball_test=read.csv("moneyball_test.csv",header=T)
md.pattern(moneyball_test)
m2 <- md.pairs(moneyball_test);m
pbox(moneyball_test,pos=1,int=FALSE,cex=0.7)
imp2 <- mice(moneyball_test)
imp2
head(complete(imp2))
moneyball_test_filled <- complete(imp2)
pbox(moneyball_test_filled,int=FALSE,cex=0.7)
summary(moneyball_test_filled)
#Scoring
moneyball_test_filled$P_TARGET_WINS <-
  30.9909146
-  0.0980972 * moneyball_test_filled$TEAM_FIELDING_DP
-  0.0351613 * moneyball_test_filled$TEAM_FIELDING_E
-  0.0163923 * moneyball_test_filled$TEAM_BATTING_SO
+  0.0457459 * moneyball_test_filled$TEAM_BASERUN_SB +
+  0.0861166 * moneyball_test_filled$TEAM_BATTING_HR +
+  0.0447497 * moneyball_test_filled$TEAM_BATTING_H +
+  0.0017051 * moneyball_test_filled$TEAM_PITCHING_BB +
```

```
+  0.0266074 * moneyball_test_filled$TEAM_BATTING_3B -
-  0.0176317 * moneyball_test_filled$TEAM_BATTING_2B -
-  0.0107410 * moneyball_test_filled$TEAM_BASERUN_CS +
+  0.0029857 * moneyball_test_filled$TEAM_PITCHING_SO +
+  0.0049060 * moneyball_test_filled$TEAM_BATTING_BB

#subset for file submission
prediction <- moneyball_test_filled[c("INDEX","P_TARGET_WINS")]
prediction$P_TARGET_WINS[(prediction$P_TARGET_WINS < 40)] = 40
prediction$P_TARGET_WINS[(prediction$P_TARGET_WINS > 115)] = 115
## writen csv for submission
write.csv(prediction, file = "logan_strouse_predictions.csv")
```