Jason Oliai

Econ 494

Dr. L

R Code

########################

###NBA 2016-2017 DATA###

########################

#EPLORATORY ANALYSIS#

library(readxl)

NBA <- read\_excel("R/NBA 2016 - 2017.xlsx")

View(NBA)

head(NBA)

tail(NBA)

summary(NBA)

hist(NBA\_2016\_2017$TWITTER\_FOLLOWER\_COUNT\_MILLIONS)

#CLEANING DATA#

NBA\_1 <- subset(NBA, ACTIVE\_TWITTER\_LAST\_YEAR >= 1) #Cleaned data to only select NBA players from the data who used twitter#

install.packages("dplyr") #EASY FOR DATA MANIPULATION

library(dplyr)

NBA\_2 <- select(NBA\_1, -TWITTER\_HANDLE, -WIKIPEDIA\_HANDLE,) #My data had many variables that were unimportant for understanding how on court play leads to twitter activity and influence

NBA\_3 <- select(NBA\_2, -PIE\_RANK, -TS\_PCT\_RANK, -PLAYER\_ID, -TEAM\_ID, -NET\_RATING,

-OREB\_PCT, -DREB\_PCT, -TM\_TOV\_PCT, -EFG\_PCT, -TS\_PCT, -PACE,

-PIE, -GP\_RANK, -W\_RANK, -L\_RANK, -W\_PCT\_RANK, -MIN\_RANK, -NET\_RATING\_RANK,

-AST\_PCT\_RANK, -AST\_TO\_RANK, -OREB\_PCT\_RANK, -DREB\_PCT\_RANK, -REB\_PCT\_RANK,

-TM\_TOV\_PCT\_RANK, -EFG\_PCT\_RANK, -PACE\_RANK, -PIE\_RANK, -FGM\_PG\_RANK, -FGA\_PG\_RANK,

-FG\_PCT\_RANK, -CFID,)

View(NBA\_3) #went from 63 to 30 variables

dim(NBA\_3)

(dim(NBA)[1]-dim(NBA\_3)[1])/dim(NBA)[1] #7% of the data lost after 3 steps of cleaning

mean(NBA\_3$PTS) #15.21

NBA\_4 <- subset(NBA\_3, PTS >= 12) #wanted to only see activity in players

#who average more than 12 ppg, indicating that they are above the mean and that they are top players

View(NBA\_4)

install.packages('ggplot2')

library(ggplot2)

ggplot(NBA\_3, aes(x=PTS, y=TWITTER\_FOLLOWER\_COUNT\_MILLIONS)) + geom\_point()+geom\_smooth(method=lm)

ggplot(NBA\_4, aes(x=PTS, y=twitter\_count)) + geom\_point()+geom\_smooth(method=lm)

NBA\_5 <- select(NBA\_4, -USG\_PCT, -OFF\_RATING, -DEF\_RATING, -AST\_RATIO\_RANK, -AST\_RATIO)

#taking out repetitive variables that could give the same information as others

View(NBA\_5)

(dim(NBA)[1]-dim(NBA\_5)[1])/dim(NBA)[1] #39% of the original data lost after step 5

install.packages("writexl")

#Importing R data into excel

library(writexl)

write\_xlsx(NBA\_5,"C:\\Users\\Jason Oliai\\Documents\\R\\TIDYNBA.xlsx")

#HISTOGRAMS game stats

Win<- NBA\_5$W\_PCT

hist(Win)

Offensive <- NBA\_5$OFF\_RATING\_RANK

hist(Offensive)

FGA<-NBA\_5$FGA

hist(FGA)

FGM<-NBA\_5$FGM

hist(FGM)

PPG<-NBA\_5$PTS

hist(PPG)

#Histograms non-game stats

AGE <- NBA\_5$AGE

hist(AGE)

salary<- NBA\_5$SALARY\_MILLIONS

hist(salary)

twitter\_count <- NBA\_5$TWITTER\_FOLLOWER\_COUNT\_MILLIONS

hist(twitter\_count)

#Age Compared to Twitter Following

plot(NBA\_5$AGE, NBA\_5$TWITTER\_FOLLOWER\_COUNT\_MILLIONS)

#Twitter Following Compared to Salary

plot(NBA\_3$TWITTER\_FOLLOWER\_COUNT\_MILLIONS, NBA\_3$SALARY\_MILLIONS)

#Points Per Game Back Court

barplot(c(31.60, 29.1, 18.1, 23.9, 28.9),

main="PPG Averages Back Court", xlab="Back Court Players", ylab="PPG", names.arg=c("Russel Westrbrook", "James Harden", "Chris Paul", "Jimmy Butler", "Isaiah Thomas"))

#Points Per Game Front Court

barplot(c(28, 25.1, 30.96, 26.4, 20.2),

main="PPG Averages Front Court", xlab="Front Court Players", ylab="PPG", names.arg=c("Anhtony Davis", "Kevin Durant", "LeBron James", "DeMarcus Cousins", "Joel Embiid"))

#GGPlot2 Box Plots OFFENSIVE RATING

install.packages('ggplot2')

library(ggplot2)

data(NBA\_5)

ggplot(NBA\_5, aes(Offensive)) + geom\_boxplot()

#Regression AGE & TWITTER\_COUNT

ggplot(NBA\_5, aes(x=AGE, y=twitter\_count)) + geom\_point(size=2, shape=23)

ggplot(NBA\_5, aes(x=AGE, y=twitter\_count)) + geom\_point()+geom\_smooth(method=lm)

#Regression Offensive Rank & FGM

ggplot(NBA\_5, aes(x=Offensive, y=FGM)) + geom\_point(size=2, shape=23)

ggplot(NBA\_5, aes(x=Offensive, y=FGM)) + geom\_point()+geom\_smooth(method=lm)

#GGPlot2 Box Plots Wins

ggplot(NBA\_5, aes(W)) + geom\_boxplot()

#GGPlot2 Box Plots MINUTES PER GAME

ggplot(NBA\_5, aes(MIN)) + geom\_boxplot()