APPENDIX: R PROGRAM FILE FOR ASSIGNMENT 2

#A1 data prep code taken from Portal

## Data prep ##

nfl.raw <- read.csv("NFLdraft.csv", head= T, strip.white= T , stringsAsFactors = F)

# head(nfl.raw)

# tail(nfl.raw)

# str(nfl.raw)

nfl.raw$Pos[nfl.raw$Pos == "LS"] <- "C"

nfl <- within(nfl.raw, {

Pos <- factor(Pos)

PosGroup <- factor(ifelse(Pos %in% c("C", "DE", "DT", "OG", "OT", "TE"), "Linemen",

ifelse(Pos %in% c("CB", "WR", "FS"), "Small Backs", "Big Backs")))

ProTeam <- factor(matrix(unlist(strsplit(Drafted, " / ")), ncol= 4, byrow = T)[,1])

Round <- factor(matrix(unlist(strsplit(Drafted, " / ")), ncol= 4, byrow = T)[,2])

Overall <- factor(matrix(unlist(strsplit(Drafted, " / ")), ncol= 4, byrow = T)[,3])

Year <- factor(matrix(unlist(strsplit(Drafted, " / ")), ncol= 4, byrow = T)[,4])

Overall <- as.numeric(gsub("[^0-9]", "", Overall))

HtFt <- as.numeric(sub("-.\*", "", Ht))

HtIn <- as.numeric(sub("\*.-", "", Ht))

Ht <- 12\*HtFt + HtIn

})

#nfl <- subset(nfl, select= -c(Link, Drafted, Year, HtFt, HtIn)) # Remove unused columns

head(nfl) # Check

str(nfl)

View(nfl)

#### Part A #####

#1.

par(family = "Courier New")

yd40 <- (6.0 - nfl$Yd40)

par(mfrow=c(1,1))

plot(yd40, nfl$Vertical)

#2.

fit <- lm(nfl$Vertical ~ yd40)

#3.

abline(coef(fit), col="red")

#4.

par(mfrow=c(1,2))

plot(fit, 1); plot(fit, 2)

#5.

#Transforming X

b <- exp(yd40)

exponential <- lm(nfl$Vertical ~ b)

par(mfrow=c(1,1))

plot(b, nfl$Vertical)

abline(coef(exponential), col="red")

par(mfrow=c(1,2))

plot(exponential,1); plot(exponential, 2)

#6.

c <- (yd40)^2

squared <- lm(nfl$Vertical ~ c)

par(mfrow=c(1,1))

plot(c, nfl$Vertical)

abline(coef(squared), col="red")

par(mfrow=c(1,2))

plot(squared, 1); plot(squared, 2)

par(mfrow=c(1,1))

#7.

summary(fit)

anova(fit)

summary(exponential)

anova(exponential)

summary(squared)

anova(squared)

#8.

abs\_dffits <- data.frame(abs(dffits(squared)))

abs\_dffits > 0.3

#dffits row numbers 26,75,205

nfl$Name[26]

nfl$Name[75]

nfl$Name[209]

lev\_value <- 2.5\* mean(influence.measures(squared)$infmat[,6])

lev <- data.frame(influence.measures(squared)$infmat[,6])

lev > lev\_value

#leverage row number 11

nfl$Name[11]

#plotting the squared model

#high leverage point

plot(c, nfl$Vertical)

abline(coef(squared), col="red")

points(c[11], nfl$Vertical[11], col="red")

#high DFFIT points

points(c[26], nfl$Vertical[26], col="blue")

points(c[75], nfl$Vertical[75], col="blue")

points(c[209], nfl$Vertical[209], col="blue")

#### Part B ####

#1.

receiving <- read.csv("Receiving.csv", head= T, strip.white= T , stringsAsFactors = F)

rushing <- read.csv("Rushing.csv", head= T, strip.white= T , stringsAsFactors = F)

View(receiving)

View(rushing)

install.packages("dplyr")

library(dplyr)

#2

rec2 <- select(receiving, Player, G:TD)

rush2 <- select(rushing, Player, G:X20.)

colnames(rec2)

colnames(rec2) <- c("Name", "REC\_G", "REC\_REC", "REC\_YDS", "REC\_YDSG", "REC\_AVG", "REC\_LNG", "REC\_X20.", "REC\_X40.", "REC\_TD")

colnames(rush2)

colnames(rush2) <- c("Name", "RUN\_G", "RUN\_ATT", "RUN\_ATTG", "RUN\_YDS", "RUN\_YDSG", "RUN\_AVG", "RUN\_TD", "RUN\_LNG", "RUN\_X20.")

rec.df <- data.frame(rec2)

rush.df <- data.frame(rush2)

install.packages("plyr")

require(plyr)

#3.

new <- join(rush.df, rec.df, by="Name", type="full")

new2 <- join(nfl, new, by="Name", type="left")

#4.

yardDash <- (6 - new2$Yd40)

fit2 <- lm(new2$REC\_YDS ~ yardDash)

summary(fit2)

#6.

par(mfrow=c(1,2))

plot(yardDash, new2$REC\_YDS); plot(fit2, 1)

#7.

fit3 <- lm(new2$RUN\_ATTG ~ new2$Overall)

summary(fit3)

#8.

par(mfrow=c(1,1))

plot(new2$Overall, new2$RUN\_ATTG)