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

## Q1 ##

The variable ‘n’ with the value 500 is the number of observations.

We then generate 500 random, normally distributed numbers.

These numbers are assigned to the variables ‘w’ and ‘z’.

They have a mean of 0 and standard deviation of 1.

The variable ‘x’ represents the income and its values are obtained from the variable ‘z’.

Variable ‘y’ represents the SAT scores. The SAT scores lie between 200 and 1600.

Variables id and group define the id and group of the variables.

The process is repeated for 3 groups and finally they are merged.

It is also observed that there is a positive relationship between income and

SAT score within each group but the relationship between income and

SAT score is a deterring one across different groups.

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

### Q2 ###

dtable <- merge(dt1 ,dt2, all=TRUE)

dtable <- merge(dtable ,dt3, all=TRUE)

dtable$group <- as.factor(dtable$group)

dtable$iid <- 1:1500

dtable$group <- as.factor(dtable$group)

summary(lm(sat~income,data=dtable))

summary(lm(sat~income+group-1,data=dtable))

summary(lm(sat~income,data=dtable[group==1]))

summary(lm(sat~income,data=dtable[group==2]))

summary(lm(sat~income,data=dtable[group==3]))

#In the pooled model, a $1000 increase in income leads to a 2.79 increase in the SAT scores.

#For the within model,

#a $1000 increase in income leads to a 20.17 point increase in SAT scores.

#In the pooled model relationship between data is present without considering the group

#In the Within group maodel the relationship considers the group number

# Varience of the group is not considered in the pooled model

# In the Other model it is considered.

ptable <- pdata.frame(dtable,index=c('group','iid'))

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

### Question 3 ###

plot(ctree(sat~income,data=dtable))

plot(ctree(sat~group,data=dtable))

plot(ctree(sat~group,data=dtable))

plot ggives mean SAT Scores with different income.

Group 1 has SAT mean score of 1106

Group 2 has SAT mean score of 1202

Group 3 has SAT mean score of 992

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

### Question 4 ###

library(partykit)

model <- glmtree(sat~income|group,data=dtable)

plot(model)

AIC(glmtree(sat~income|group,data=dtable))

15959.82

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

### Question 5 ###

kmeans.wss <- function(data,maxclu=10,seed=1,nstart=10) {

wss <- rep(NA,maxclu)

for (i in 1:maxclu) {

set.seed(seed)

model <- kmeans(data,centers=i,nstart=nstart)

wss[i] <- model$tot.withinss

}

return(wss)

}

eratio <- function(wss) {

n <- NROW(wss)

dss <- -diff(wss)

dss <- c(wss[1]/log(n),dss)

erat <- dss[1:(n-1)]/dss[2:n]

gss <- log(1+dss/wss)

grat <- gss[1:(n-1)]/gss[2:n]

return(c(which.max(erat),which.max(grat)))

}

plot.wss <- function(wss) {

plot(1:NROW(wss), wss, type="b", xlab="Number of Clusters", ylab="Aggregate Within Group SS")

}

wss <- ptable %>% kmeans.wss

plot.wss(wss)

eratio(wss) # 2 3

model <- kmeans(dtable[,.(income,sat)],2,nstart=10)

model$centers

dtable$kgroup1 <- as.factor(model$cluster)

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

### Q6 ###

table(dtable$kgroup1,dtable$group)

hclust.wss <- function(data,model=hclust(dist(data)),maxclu=10) {

wss <- rep(NA,maxclu)

for(i in 1:maxclu){

gps <- cutree(model,i)

means <- data[,lapply(.SD,mean),by=gps]

demeaned <- data-means[gps,2:(ncol(data)+1)]

wss[i] <- sum(demeaned^2)

}

return(wss)

}

wss <- hclust.wss(dtable[,.(income,sat)])

eratio(wss)

plot.wss(wss)

model <- hclust(dist(dtable[,.(income,sat)]))

dtable$hgroup <- as.factor(cutree(model,4))

table(dtable$hgroup,dtable$group)

## we have 4 groups

# we have about 268+250+202+134 correct answers out of 1500

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

## Question 7 ##

summary(lm(sat~income+kgroup1-1,data=dtable))

summary(lm(sat~income+hgroup-1,data=dtable))

## we are not able to find the within group relationships

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

## Question 8 ##

wss <- kmeans.wss(dtable[,.(income)])

eratio(wss)

plot.wss(wss)

model <- kmeans(dtable[,.(income)],3,nstart=10)

model$centers

dtable$kgroup2 <- as.factor(model$cluster)

table(dtable$kgroup2,dtable$group)

summary(lm(sat~income+kgroup2-1,data=dtable))

## estimation is highly accurate ( ) and we are able to recover the within group relationships

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

## Question 9 ##

kmeans.scale <- function(dt,k,nstart=10){

model <- kmeans(scale(dt),k,nstart=nstart)

sdmat <- ones(k)%\*%as.matrix(dt[,lapply(.SD,sd)])

mmat <- ones(k)%\*%as.matrix(dt[,lapply(.SD,mean)])

meanmat <- model$centers\*sdmat + mmat model[[10]] <- meanmat names(model)[10] <- 'means' return(model) }

wss <- kmeans.wss(dtable[,.(scale(income),scale(sat))])

wss

eratio(wss)

model <- kmeans.scale(dtable[,.(income,sat)],3)

model$centers

model$means

dtable$kgroup3 <- as.factor(model$cluster)

table(dtable$kgroup3,dtable$group)

summary(lm(sat~income+kgroup3-1,data=dtable))

## accuracy has dropped to using two variables

## kmeans is very sensitive to the input variables