n <- 500

set.seed(75080)

z <- rnorm(n)

w <- rnorm(n)

x <- 5\*z + 50

y <- -100\*z+ 1100 + 50\*w

y <- 10\*round(y/10)

y <- ifelse(y<200,200,y)

y <- ifelse(y>1600,1600,y)

dt1 <- data.table('id'=1:500,'sat'=y,'income'=x,'group'=rep(1,n))

z <- rnorm(n)

w <- rnorm(n)

x <- 5\*z + 80

y <- -80\*z+ 1200 + 50\*w

y <- 10\*round(y/10)

y <- ifelse(y<200,200,y)

y <- ifelse(y>1600,1600,y)

dt2 <- data.table('id'=501:1000,'sat'=y,'income'=x,'group'=rep(2,n))

z <- rnorm(n)

w <- rnorm(n)

x <- 5\*z + 30

y <- -120\*z+ 1000 + 50\*w

y <- 10\*round(y/10)

y <- ifelse(y<200,200,y)

y <- ifelse(y>1600,1600,y)

dt3 <- data.table('id'=1001:1500,'sat'=y,'income'=x,'group'=rep(3,n))

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

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

ggplot(dtable,aes(x=income,y=sat,color=as.factor(group)))+geom\_point()

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

#Q1.1

#The data generating process is done using three groups.

#The groups in SAT are lower,middle and high income groups

#Within every group,SAT score is negatively related to income level.

#but in Between groups, the effect might be different, which is not concluded from

#the data generating process.

#Q1.2

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

dtable$id <- 1:1500

lm1 <- lm(sat~income,data=dtable)

lm2 <- lm(sat~income+group-1,data=dtable)

lm3 <- lm(sat~income,data=dtable[group==1])

lm4 <- lm(sat~income,data=dtable[group==2])

lm5 <- lm(sat~income,data=dtable[group==3])

summary(lm1) # In pooled OLS the sat and income are negatively correlated

summary(lm2) # Only the coefficient is negatively corellated all else are highly possitive

summary(lm3)

summary(lm4)

summary(lm5)

#All three groups are separately negatively correlated with sat

#Q1.3

ctree(sat~income+group,data=dtable)

ctree(sat~income,data=dtable)

ctree(sat~group,data=dtable)

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

#This plot adds no additional value to our analysis and looks vague to draw conclusion

#Q1.4

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

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

#The splitting looks more defined and linear using a glmtree model and so it is the bestto use

#Q1.5

kmeansplot <- function(wss){

wss <- data.table('Centers'=1:NROW(wss),'wss'=wss)

ggplot(wss,aes(x=Centers,y=wss)) + geom\_line() +

scale\_y\_continuous('Within group sum of squares') +

scale\_x\_discrete(limits=1:10) +

ggtitle('k-Means Elbow Plot') +

theme(aspect.ratio=1,plot.margin=grid::unit(c(0,0,0,0), "mm"))

}

wss <- rep(NA,10)

for(i in 1:10)

wss[i] <- kmeans(dtable%>%select(sat,income),i,nstart=10)$tot.withinss

kmeansplot(wss) #There are two elbows at 2 and 3 since we dont want to miss data I choose 3

kmeans(dtable%>%select(sat,income),2)$centers

kmeans(dtable%>%select(sat,income),3)$centers

#Q1.6

dtable$kmean\_3 <- as.factor(kmeans(dtable%>%select(sat,income),3)$cluster)

table(dtable$kmean\_3,dtable$group)

dtable$kmean\_2 <- as.factor(kmeans(dtable%>%select(sat,income),2)$cluster)

table(dtable$kmean\_2,dtable$group)

dtable$hclust <- as.factor(cutree(hclust(dtable%>%select(sat,income)%>%dist),3))

table(dtable$hclust,dtable$group)

#(256+343+267)/1500=57.8%

#(401+202+134)/1500=49.13%

#This model is does not seem better to identify the K means

#Q1.7

lm1 <- lm(sat~income,data=dtable)

summary(lm1)

lm2 <- lm(sat~income+kmean\_3-1,data=dtable)

summary(lm2)

lm3 <- lm(sat~income+kmean\_2-1,data=dtable)

summary(lm3)

lm4 <- lm(sat~income+hclust-1,data=dtable)

summary(lm4)

#The negative relations have all turned into possitive correlations in this modeling

#Q1.8

wss <- rep(NA,10)

for(i in 1:10)

wss[i] <- kmeans(dtable%>%select(income),i,nstart=10)$tot.withinss

kmeansplot(wss)

#The distinct elbow is now 2

kmeans(dtable%>%select(income),2)$centers

dtable$kmean\_21 <- as.factor(kmeans(dtable%>%select(income),2)$cluster)

table(dtable$kmean\_21,dtable$group)

kmeans(dtable%>%select(income),3)$centers

dtable$kmean\_31 <- as.factor(kmeans(dtable%>%select(income),3)$cluster)

table(dtable$kmean\_31,dtable$group)

summary(lm1 <- lm(sat~income+kmean\_21-1,data=dtable))

summary(lm2 <- lm(sat~income+kmean\_31-1,data=dtable))

#(489+500+484)/1500=98.2% The identification rate looks quiet good now

#The relationship is now positively correlated

#Q1.9

wss <- rep(NA,10)

for(i in 1:10)

wss[i] <- kmeans(dtable%>%select(sat,income)%>%scale,i,nstart=10)$tot.withinss

kmeansplot(wss)

kmeans(dtable%>%select(income,sat)%>%scale,3)$centers

dtable$kmean\_311 <- as.factor(kmeans(dtable%>%select(income,sat)%>%scale,3,nstart=10)$cluster)

table(dtable$kmean\_311,dtable$group) #(371+500+330)/1500=80% The identification rate has dropped but is

#still a good number

summary(lm1 <- lm(sat~income+kmean\_311-1,data=dtable))

#The coefficients look greatly positive