# Import the files

Data1<-read.csv("HW1DATA1.csv")
Data2<-read.csv("HW1DATA2.csv")

ls()

# Using SQL, extract only the needed variables

library(sqldf)
install.packages("sqldf")

str(Data1)

Data1a<-sqldf('select TRADES, AGE, RBAL, MATCHKEY from Data1')

str(Data1a)
Data2a<-sqldf('select DELQID, CRELIM, MATCHKEY from Data2')

#merge the two datasets

?merge

#note that Data1 has fewer values...

Data3<-merge(Data1a, Data2a, by="MATCHKEY", all.x="TRUE")

dim(Data3)

summary(Data3)

#Prove that there are no duplicate MATCHKEYS in the file

sqldf('select count(distinct MATCHKEY) as count from Data3')

#Report the structure and overall summary statistics

dim(Data3)

str(Data3)

attach(Data3)

summary(Data3)

#note that since MATCHKEY is an identifier and not really a variable, this should be ommitted from the summary
Nomatch <- Data3[c(2:6)]

summary(Nomatch)

# Create the following new variables:
# If the DELQID = 0 then RISK = NEW, if the DELQID = 1 or 2 then RISK = GOOD, if the DELQID = 3 + then RISK = BAD.

Data3$RISK[Data3$DELQID == 0] <- "NEW"
Data3$RISK[Data3$DELQID == 1] <- "GOOD"
Data3$RISK[Data3$DELQID == 2] <- "GOOD"
Data3$RISK[Data3$DELQID > 2] <- "BAD"

head(Data3)

Data3 <- Data3[, -8]

table(Data3$RISK)

# If the CRELIM <= 2000 then LIMIT = 1, if the CRELIM 2000 <= 5000 then LIMIT = 2, if the CRELIM > 5000 then LIMIT = 3.

Data3$limit <- ifelse(Data3$CRELIM <= 2000, 1, ifelse(Data3$CRELIM <= 4000, 2, 3))

table(Data3$limit)

# If a value for any variable is greater than 4 standard deviations above the mean, it should be imputed with the median.
summary(Data3)

#from the summary, you will find the means for AGE, TRADES, RBAL and CRELIM...the others do not make sense.
#determine the standard deviation for each of these

attach(Data3)

sd(AGE)
sd(TRADES)
sd(RBAL)

#from this information, you will see that age is not impacted because 4x the std is 60. The average age is 47.
#There are no observations over 127. The other three are impacted.

Data3$TRADES<-ifelse(Data3$TRADES >70,17,Data3$TRADES)
Data3$RBAL<-ifelse(Data3$RBAL >44549,7617,Data3$RBAL)

summary(Data3)
table(Data3$RISK)

#Using SQL, report the count, average, median, min and max for AGE, RBAL and TRADES by RISK. Do the same by LIMIT.
#Note that I have provided a subset of what is required...but I will assume that you can follow this...

Risktrades<-sqldf('select Risk, count(Risk) N, avg(TRADES) AVG_TRADES from Data3 group by Risk')
Risktrades
Riskcrelim<-sqlf('select Risk, count(Risk) N, avg(CRELIM) AVG_CRELIM from Data3 group by Risk')
Riskcrelim

Riskage<-sqlf('select Risk, count(Risk) N, avg(AGE) AVG_AGE from Data3 group by Risk')
Riskage

Riskrbal<-sqlf('select Risk, count(Risk) N, avg(RBAL) AVG_RBAL from Data3 group by Risk')
Riskrbal