R Notebook

#Load library.  
require('caret')

## Loading required package: caret

## Loading required package: lattice

## Loading required package: ggplot2

#Set project path so you don't have to type full path every time (this is the scripts folder).  
setwd('./')  
  
#Load data.  
subjData <- readRDS("../subjectData/n1396\_T1\_subjData.rds")  
  
#Pull out only variables of interest for splitting and checking demographics.  
dataToSplit <- subjData[,c('bblid','scanid','sex','race2','ageAtScan1','mood\_4factor','psychosis\_4factor','externalizing\_4factor','phobias\_4factor','overall\_psychopathology\_4factor')]  
  
#Remove any NA's from the variable you want to split on.  
dataNoNA <- subset(dataToSplit,is.na(dataToSplit$overall\_psychopathology\_4factor)==FALSE)  
  
#Count number of subjects before and after removing NAs.  
nrow(dataToSplit)

## [1] 1396

nrow(dataNoNA)

## [1] 1395

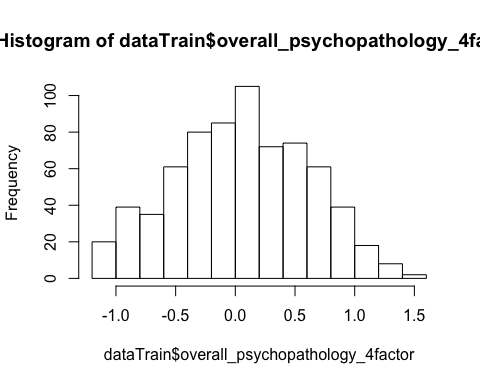
#Set a random number as the seed.  
set.seed(1234)  
  
##Split into the train and test data sets using caret.  
#p=the percentage of data that goes to training (e.g., 50%)  
#list=FALSE (gives a matrix with rows instead of list)   
#times=the number of partitions to create (number of training sets)  
trainIndex <- createDataPartition(dataNoNA$overall\_psychopathology\_4factor, p=0.5, list=F, times=1)  
  
#Pull the variables into the new train and test matrices.  
dataTrain <- dataNoNA[trainIndex,]  
dataTest <- dataNoNA[-trainIndex,]  
  
#Count number of subjects in the train and test sets.  
nrow(dataTrain)

## [1] 699

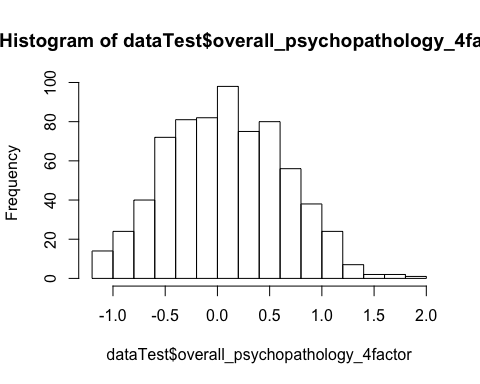
nrow(dataTest)

## [1] 696

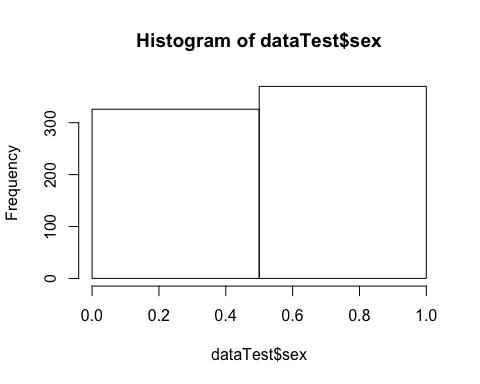
#Save the train and test samples with the demographic variables.  
#write.csv(dataTrain,'../subjectData/n699\_T1\_train.csv',row.names=FALSE, quote=FALSE)  
#write.csv(dataTest,'../subjectData/n696\_T1\_test.csv',row.names=FALSE, quote=FALSE)  
  
#Save the bblids and scanids only for NMF.  
IDs <- c("bblid", "scanid")  
bblidsScanids\_train <- dataTrain[IDs]  
bblidsScanids\_test <- dataTest[IDs]  
  
#Remove header.  
names(bblidsScanids\_train) <- NULL  
names(bblidsScanids\_test) <- NULL  
  
#Save lists.  
#write.csv(bblidsScanids\_train, file="../subjectData/n1396\_T1\_train\_bblids\_scanids.csv", row.names=FALSE)  
#write.csv(bblidsScanids\_test, file="../subjectData/n1396\_T1\_test\_bblids\_scanids.csv", row.names=FALSE)  
  
##Make histograms to see that train and test are relatively similar.  
#Overall Psych  
hist(dataTrain$overall\_psychopathology\_4factor,breaks=14)



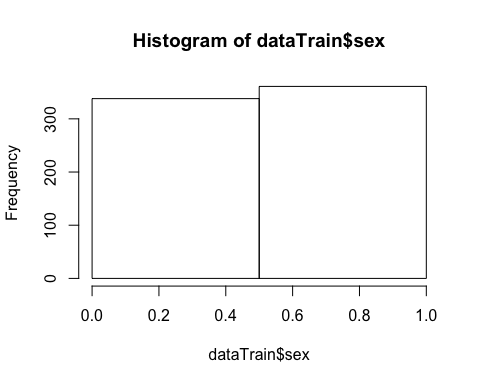
hist(dataTest$overall\_psychopathology\_4factor,breaks=14)



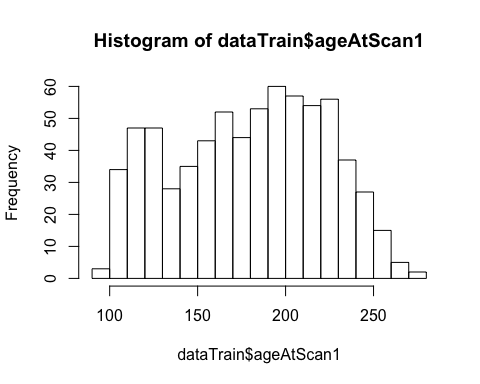
#Sex  
hist(dataTest$sex,breaks=2)



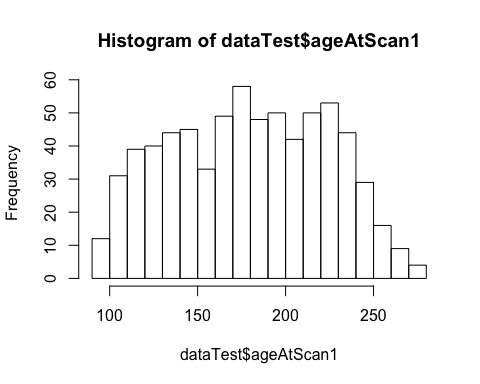
hist(dataTrain$sex,breaks=2)



#Age  
hist(dataTrain$ageAtScan1,breaks=14)



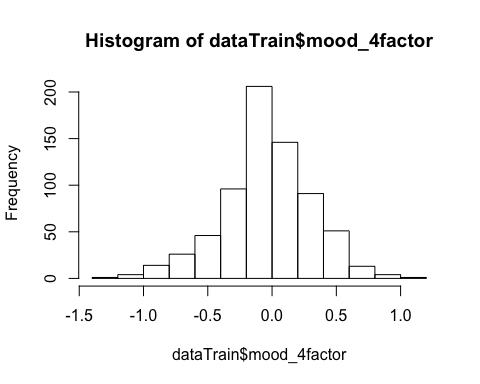
hist(dataTest$ageAtScan1,breaks=14)



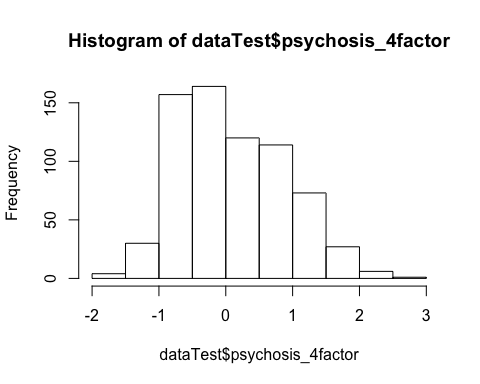
#Mood  
hist(dataTest$mood\_4factor,breaks=14)



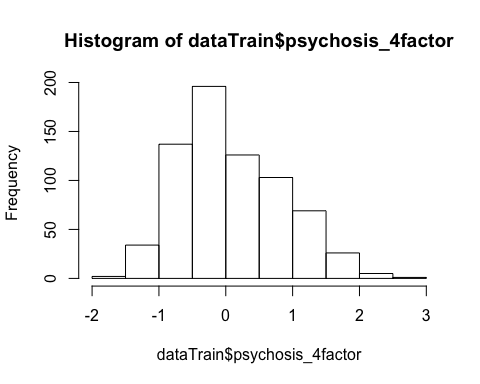
hist(dataTrain$mood\_4factor,breaks=14)



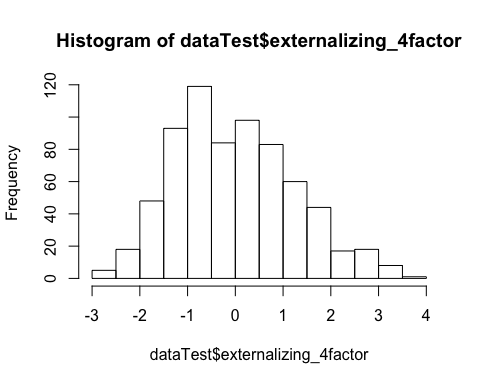
#Psychosis  
hist(dataTest$psychosis\_4factor,breaks=14)



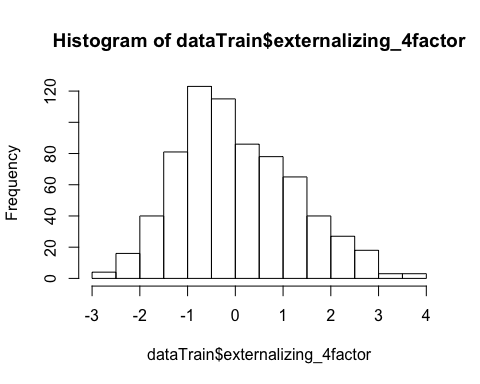
hist(dataTrain$psychosis\_4factor,breaks=14)



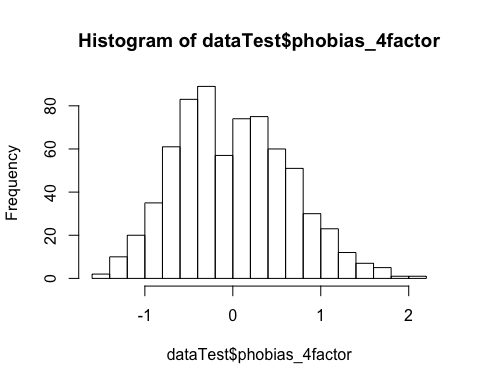
#Externalizing  
hist(dataTest$externalizing\_4factor,breaks=14)



hist(dataTrain$externalizing\_4factor,breaks=14)



#Phobias  
hist(dataTest$phobias\_4factor,breaks=14)



hist(dataTrain$phobias\_4factor,breaks=14)

