Part 1 - data

jean wills

06/07/2020

## Project: Portuguese Bank Marketing Data

# Step 1: list of the packages required

# install.packages("lattice")  
# install.packages("ggplot2")  
# install.packages("mlbench")  
# install.packages("caret")  
# install.packages("plyr")  
# install.packages("GGally")  
# install.packages("reshape2")  
# install.packages("psych")  
# install.packages("dplyr")  
  
# install.packages("normalr")  
# install.packages("rJava")  
# install.packages("RWeka")  
#

# Step 2a: read the data and look at the data structure

# setwd("~/Users/jeanwills/Desktop/CKME136/")  
BM <- read.csv("/Users/jeanwills/Desktop/CKME136/bank\_full.csv", header=T, sep = ";", stringsAsFactors = T, na.strings = "NA")  
# look at the data structure   
BM\_mini <- read.csv("/Users/jeanwills/Desktop/CKME136/bank.csv", header=T, sep = ";", stringsAsFactors = T, na.strings = "NA")  
# let's check number of complete cases for no data missing at all -> no missing data!  
sum(complete.cases(BM))

## [1] 45211

# step 2b - Look at the bank data summary

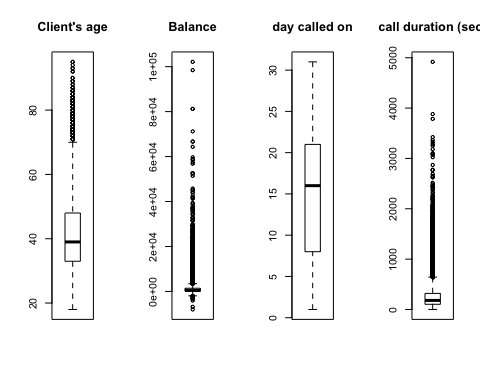
summary(BM)

## age job marital education   
## Min. :18.00 blue-collar:9732 divorced: 5207 primary : 6851   
## 1st Qu.:33.00 management :9458 married :27214 secondary:23202   
## Median :39.00 technician :7597 single :12790 tertiary :13301   
## Mean :40.94 admin. :5171 unknown : 1857   
## 3rd Qu.:48.00 services :4154   
## Max. :95.00 retired :2264   
## (Other) :6835   
## default balance housing loan contact   
## no :44396 Min. : -8019 no :20081 no :37967 cellular :29285   
## yes: 815 1st Qu.: 72 yes:25130 yes: 7244 telephone: 2906   
## Median : 448 unknown :13020   
## Mean : 1362   
## 3rd Qu.: 1428   
## Max. :102127   
##   
## day month duration campaign   
## Min. : 1.00 may :13766 Min. : 0.0 Min. : 1.000   
## 1st Qu.: 8.00 jul : 6895 1st Qu.: 103.0 1st Qu.: 1.000   
## Median :16.00 aug : 6247 Median : 180.0 Median : 2.000   
## Mean :15.81 jun : 5341 Mean : 258.2 Mean : 2.764   
## 3rd Qu.:21.00 nov : 3970 3rd Qu.: 319.0 3rd Qu.: 3.000   
## Max. :31.00 apr : 2932 Max. :4918.0 Max. :63.000   
## (Other): 6060   
## pdays previous poutcome y   
## Min. : -1.0 Min. : 0.0000 failure: 4901 no :39922   
## 1st Qu.: -1.0 1st Qu.: 0.0000 other : 1840 yes: 5289   
## Median : -1.0 Median : 0.0000 success: 1511   
## Mean : 40.2 Mean : 0.5803 unknown:36959   
## 3rd Qu.: -1.0 3rd Qu.: 0.0000   
## Max. :871.0 Max. :275.0000   
##

# we see that 7 attributes are numeric and the rest are now factors  
# we will come back to this and change the classes where required

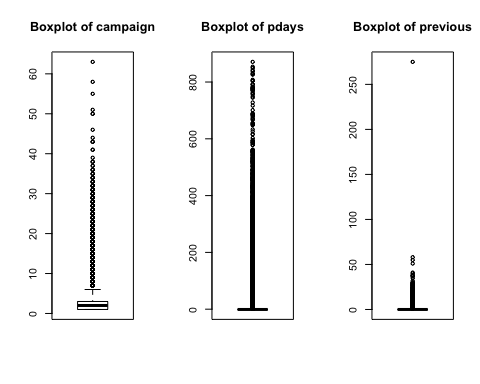
# Step 3a: plot the boxplots of the numeric data

# age has outliers above ~70  
par(mfrow=c(1,4))  
boxplot(BM$age, main = "Client's age")  
# balance has a large number of outliers above Q3  
boxplot(BM$balance, main = "Balance")  
# day has a large number of outliers above Q3  
boxplot(BM$day, main = "day called on")  
# duration has a large number of outliers above Q3  
boxplot(BM$duration, main = "call duration (sec)")



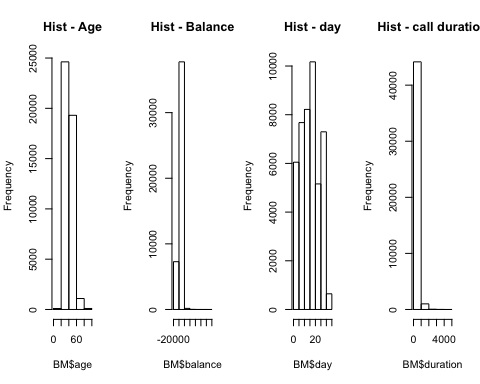
# Step 3b

par(mfrow=c(1,3))  
# camapign has a large number of outliers above the Q3  
boxplot(BM$campaign, main = "Boxplot of campaign")  
# pdays has a VERY large number of outliers above Q3  
boxplot(BM$pdays, main = "Boxplot of pdays")  
# duration has a large number of outliers above Q3  
boxplot(BM$previous, main = "Boxplot of previous")



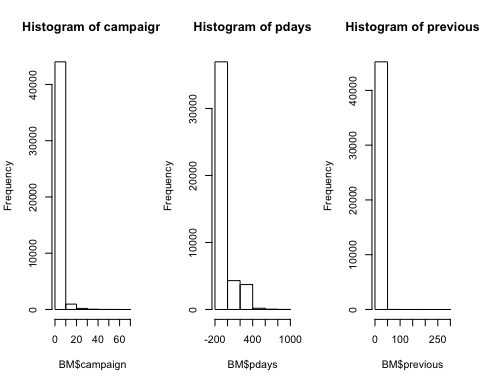
# Step 4a: plot histograms to reveal skewness / normality

par(mfrow=c(1,4))  
# age looks skewed right  
hist(BM$age, main = "Hist - Age", breaks = 5)  
# balance is skewed right  
hist(BM$balance, main = "Hist - Balance", breaks = 5)  
# day somewhat skewed right  
hist(BM$day, main = "Hist - day", breaks = 10)  
# duration is skewed right  
hist(BM$duration, main = "Hist - call duration", breaks = 5)



# Step 4b: plot histograms to reveal skewness / normality

par(mfrow=c(1,3))  
# campaign is skewed right - most data is in 1 day  
hist(BM$campaign, main = "Histogram of campaign", breaks = 5)  
# pdays skewed right  
hist(BM$pdays, main = "Histogram of pdays", breaks = 5)  
# previous skewed right  
hist(BM$previous, main = "Histogram of previous", breaks = 5)



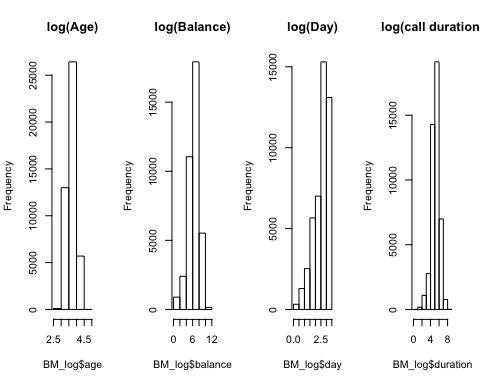
# 4c logs of the numeric data and replot the histograms…

probably won’t use this as some data is negative and <1 and doing log(data) does NOT work for them but keep for now

# day is somewhat better in the original histogram but will still change to be consistent  
# NaNs produced.....need # >1.0 - balance and duration would need adjustments  
BM\_log<-BM  
BM\_log$age<-log(BM$age)  
# balance has negative values   
BM\_log$balance<-log(BM$balance)

## Warning in log(BM$balance): NaNs produced

BM\_log$day<-log(BM$day)  
# duration has zeros  
BM\_log$duration<-log(BM$duration)  
#  
par(mfrow=c(1,4))  
hist(BM\_log$age, main = "log(Age)", breaks = 5)  
hist(BM\_log$balance, main = "log(Balance)", breaks = 5)  
hist(BM\_log$day, main = "log(Day)", breaks = 5)  
hist(BM\_log$duration, main = "log(call duration)", breaks = 10)

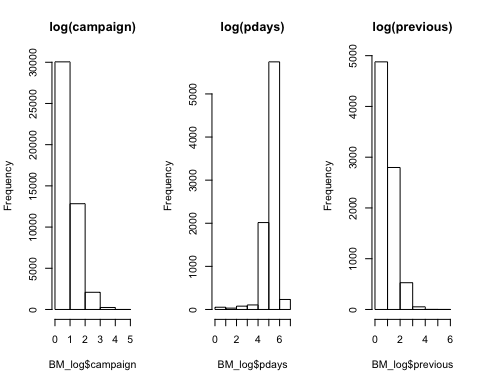


# 4d logs of the numeric data and replot the histograms…

# NaNs produced.....need # >1.0 - pdays and previous would need adjustments  
BM\_log$campaign<-log(BM$campaign)  
# pdays has -1  
BM\_log$pdays<-log(BM$pdays)

## Warning in log(BM$pdays): NaNs produced

# previous has 0 and 1  
BM\_log$previous<-log(BM$previous)  
# str(BM\_log)  
par(mfrow=c(1,3))  
hist(BM\_log$campaign, main = "log(campaign)", breaks = 5)  
hist(BM\_log$pdays, main = "log(pdays)", breaks = 5)  
hist(BM\_log$previous, main = "log(previous)", breaks = 5)

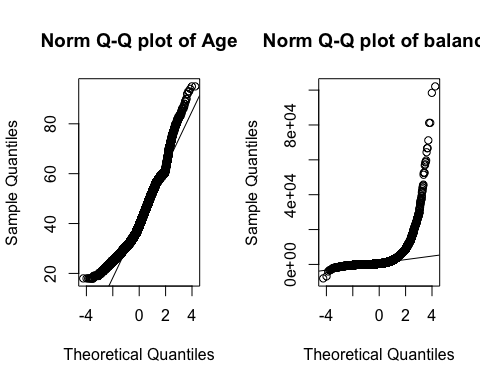


# most seem normal except campaign is still right skewed

The logs work for most of the data but not all…..so leaving data as is

# Step 5a: Q-Q plots

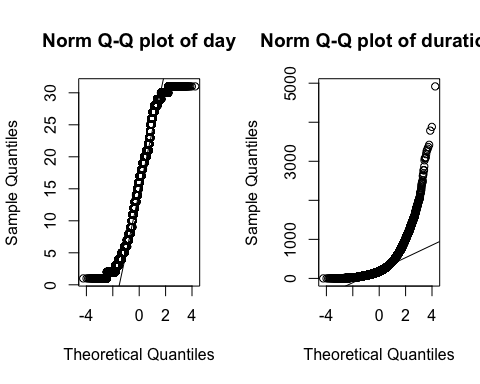
par(mfrow=c(1,2))  
qqnorm(BM$age, main = "Norm Q-Q plot of Age")  
qqline(BM$age)  
qqnorm(BM$balance, main = "Norm Q-Q plot of balance")  
qqline(BM$balance)



# both not normal

# Step 5b: Q-Q plots

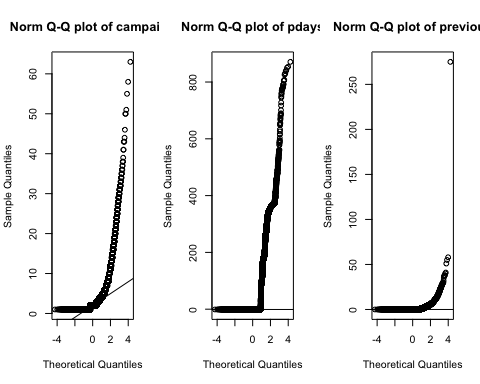
par(mfrow=c(1,2))  
qqnorm(BM$day, main = "Norm Q-Q plot of day")  
qqline(BM$day)  
qqnorm(BM$duration, main = "Norm Q-Q plot of duration")  
qqline(BM$duration)



# both not normal

# Step 5c: Q-Q plots

par(mfrow=c(1,3))  
qqnorm(BM$campaign, main = "Norm Q-Q plot of campaign")  
qqline(BM$campaign)  
qqnorm(BM$pdays, main = "Norm Q-Q plot of pdays")  
qqline(BM$pdays)  
qqnorm(BM$previous, main = "Norm Q-Q plot of previous")  
qqline(BM$previous)



# all 3 not normal

# Step 6: Shapiro Tests for Normality on numeric data

IF p<0.05 then the numeric data is not normal and significant Shapiro requires dataset size under 5,000 so using BM\_mini version # — All numeric attributes are NOT normal

shapiro.test(BM\_mini$age)

##   
## Shapiro-Wilk normality test  
##   
## data: BM\_mini$age  
## W = 0.95951, p-value < 2.2e-16

shapiro.test(BM\_mini$balance)

##   
## Shapiro-Wilk normality test  
##   
## data: BM\_mini$balance  
## W = 0.50151, p-value < 2.2e-16

shapiro.test(BM\_mini$day)

##   
## Shapiro-Wilk normality test  
##   
## data: BM\_mini$day  
## W = 0.96072, p-value < 2.2e-16

shapiro.test(BM\_mini$duration)

##   
## Shapiro-Wilk normality test  
##   
## data: BM\_mini$duration  
## W = 0.74754, p-value < 2.2e-16

shapiro.test(BM\_mini$campaign)

##   
## Shapiro-Wilk normality test  
##   
## data: BM\_mini$campaign  
## W = 0.56082, p-value < 2.2e-16

shapiro.test(BM\_mini$pdays)

##   
## Shapiro-Wilk normality test  
##   
## data: BM\_mini$pdays  
## W = 0.47041, p-value < 2.2e-16

shapiro.test(BM\_mini$previous)

##   
## Shapiro-Wilk normality test  
##   
## data: BM\_mini$previous  
## W = 0.35998, p-value < 2.2e-16

# Step 7a: test for correlations within the numeric attributes

Since we know the numeric data is not-normal, we use Spearman instead of Pearson method The correlation heat map is created Pearson method is default and p>0.05 means NOT correlated Spearman method - if p<0.05 means NOT correlated

# if p<0.05 then significant meaning correlated  
# simple example with 2 variables  
# if we do this, we need y as numeric 0/1  
# cor.test(BM$previous,BM$age, method="spearman")  
# cor.test(BM$previous,BM$age)  
# also: cor(BM$previous,BM$age)

# Step 7b: Table of correlations for all data

Correlation test can also be considered a Feature Removal method

# test ALL data for correlations  
library(lattice)  
library(ggplot2)  
BM\_num<-BM  
# num<- subset(BM\_01, select = c("age", "balance", "day", "duration", "campaign", "pdays", "previous", "y"))  
BM\_num$job<- as.numeric(BM\_num$job) #12  
BM\_num$marital<- as.numeric(BM\_num$marital) #4  
BM\_num$education<- as.numeric(BM\_num$education) #4  
BM\_num$default<- as.numeric(BM\_num$default) #2  
BM\_num$housing<- as.numeric(BM\_num$housing) #2  
BM\_num$loan<- as.numeric(BM\_num$loan) #2  
BM\_num$contact<- as.numeric(BM\_num$contact) #3  
BM\_num$month<- as.numeric(BM\_num$month) #12  
BM\_num$poutcome<- as.numeric(BM\_num$poutcome) #4  
# correlations data   
# Identify highly correlated features in caret r package  
# ensure the results are repeatable  
set.seed(12)  
library(mlbench)  
library(caret)  
# leaving out y so only 16 not 17  
corMatrix<-cor(BM\_num[, c(1:16)])  
print(corMatrix)

## age job marital education default  
## age 1.000000000 -0.0218679434 -0.403240136 -1.068066e-01 -0.017879304  
## job -0.021867943 1.0000000000 0.062045485 1.667067e-01 -0.006853085  
## marital -0.403240136 0.0620454852 1.000000000 1.085761e-01 -0.007023365  
## education -0.106806594 0.1667067239 0.108576125 1.000000e+00 -0.010717690  
## default -0.017879304 -0.0068530852 -0.007023365 -1.071769e-02 1.000000000  
## balance 0.097782739 0.0182315155 0.002121918 6.451404e-02 -0.066745057  
## housing -0.185513082 -0.1253628132 -0.016095882 -9.079024e-02 -0.006025218  
## loan -0.015655273 -0.0330039210 -0.046892524 -4.857353e-02 0.077234241  
## contact 0.026221067 -0.0820633039 -0.039201423 -1.109276e-01 0.015404140  
## day -0.009120046 0.0228555732 -0.005261364 2.267105e-02 0.009423899  
## month -0.042357405 -0.0928695791 -0.006990661 -5.730383e-02 0.011485783  
## duration -0.004648428 0.0047436409 0.011852173 1.935105e-03 -0.010021461  
## campaign 0.004760312 0.0068386259 -0.008994100 6.255137e-03 0.016821531  
## pdays -0.023758014 -0.0244550401 0.019172254 5.235498e-05 -0.029979361  
## previous 0.001288319 -0.0009106174 0.014973243 1.756963e-02 -0.018329405  
## poutcome 0.007366903 0.0110103583 -0.016850456 -1.936137e-02 0.034898194  
## balance housing loan contact day  
## age 0.097782739 -0.185513082 -0.015655273 0.02622107 -0.009120046  
## job 0.018231515 -0.125362813 -0.033003921 -0.08206330 0.022855573  
## marital 0.002121918 -0.016095882 -0.046892524 -0.03920142 -0.005261364  
## education 0.064514043 -0.090790237 -0.048573533 -0.11092757 0.022671046  
## default -0.066745057 -0.006025218 0.077234241 0.01540414 0.009423899  
## balance 1.000000000 -0.068768316 -0.084350246 -0.02727294 0.004502585  
## housing -0.068768316 1.000000000 0.041322866 0.18812289 -0.027981649  
## loan -0.084350246 0.041322866 1.000000000 -0.01087301 0.011370158  
## contact -0.027272944 0.188122888 -0.010873011 1.00000000 -0.027936231  
## day 0.004502585 -0.027981649 0.011370158 -0.02793623 1.000000000  
## month 0.019777231 0.271480739 0.022144853 0.36114488 -0.006027676  
## duration 0.021560380 0.005075449 -0.012411972 -0.02083930 -0.030206341  
## campaign -0.014578279 -0.023598707 0.009979846 0.01961438 0.162490216  
## pdays 0.003435322 0.124178400 -0.022753639 -0.24481646 -0.093044074  
## previous 0.016673637 0.037076150 -0.011043488 -0.14781140 -0.051710497  
## poutcome -0.020967337 -0.099970667 0.015457767 0.27221380 0.083459682  
## month duration campaign pdays previous  
## age -0.042357405 -0.004648428 0.004760312 -2.375801e-02 0.0012883192  
## job -0.092869579 0.004743641 0.006838626 -2.445504e-02 -0.0009106174  
## marital -0.006990661 0.011852173 -0.008994100 1.917225e-02 0.0149732426  
## education -0.057303833 0.001935105 0.006255137 5.235498e-05 0.0175696313  
## default 0.011485783 -0.010021461 0.016821531 -2.997936e-02 -0.0183294048  
## balance 0.019777231 0.021560380 -0.014578279 3.435322e-03 0.0166736367  
## housing 0.271480739 0.005075449 -0.023598707 1.241784e-01 0.0370761497  
## loan 0.022144853 -0.012411972 0.009979846 -2.275364e-02 -0.0110434883  
## contact 0.361144884 -0.020839303 0.019614376 -2.448165e-01 -0.1478113997  
## day -0.006027676 -0.030206341 0.162490216 -9.304407e-02 -0.0517104967  
## month 1.000000000 0.006313636 -0.110030865 3.306469e-02 0.0227271448  
## duration 0.006313636 1.000000000 -0.084569503 -1.564770e-03 0.0012030569  
## campaign -0.110030865 -0.084569503 1.000000000 -8.862767e-02 -0.0328552897  
## pdays 0.033064690 -0.001564770 -0.088627668 1.000000e+00 0.4548196355  
## previous 0.022727145 0.001203057 -0.032855290 4.548196e-01 1.0000000000  
## poutcome -0.033038191 0.010925350 0.101587641 -8.583616e-01 -0.4897518649  
## poutcome  
## age 0.007366903  
## job 0.011010358  
## marital -0.016850456  
## education -0.019361368  
## default 0.034898194  
## balance -0.020967337  
## housing -0.099970667  
## loan 0.015457767  
## contact 0.272213798  
## day 0.083459682  
## month -0.033038191  
## duration 0.010925350  
## campaign 0.101587641  
## pdays -0.858361643  
## previous -0.489751865  
## poutcome 1.000000000

highCorr <- findCorrelation(corMatrix, cutoff=0.5)  
print(highCorr)

## [1] 16

# results:

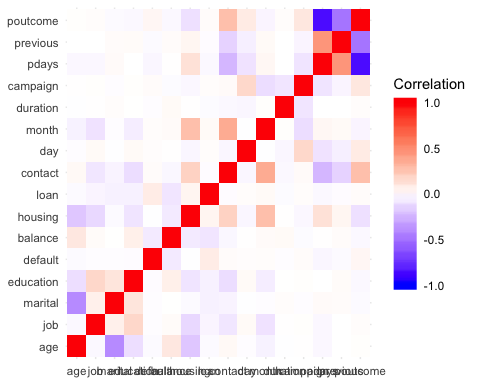
only 3 >= 0.5 corr: poutcome to pdays is -0.858 (negative) poutcome to previous is -0.49 (negative) previous to pdays is 0.455 (mild positive) Keep all for now

# Step 7c: do a correlation heat map to visualize the data

library(plyr)  
library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(ggplot2)  
library(reshape2)  
library(caret)  
# BM\_num was created above  
# leaving out y so only 16 not 17  
# this one prints extra info - ggpairs(BM\_num[, c(1,16)])  
# require(scales)  
bnk\_core<- cor(BM\_num[, c(1:16)])  
bnk\_melt<- melt(bnk\_core, varnames=c("x", "y"),value.name="Correlation")  
# summarize the correlation matrix  
highlyCorrelated <- findCorrelation(bnk\_core, cutoff=0.5)  
# print indexes of highly correlated attributes  
ggplot(bnk\_melt, aes(x=x, y=y)) +  
 geom\_tile(aes(fill=Correlation)) +   
 scale\_fill\_gradient2(low="blue", mid="white", high="red", guide=guide\_colorbar(ticks=FALSE, barheight=10),limits=c(-1,1)) +  
 theme\_minimal() +  
 labs(x=NULL, y=NULL)



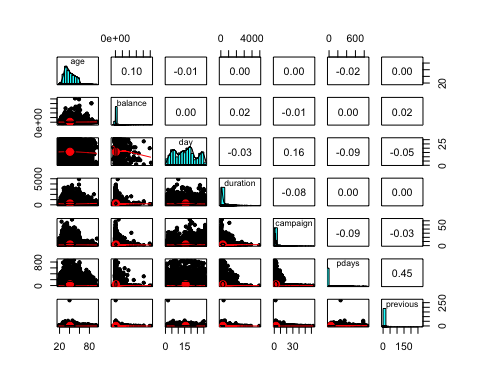
# Step 7d: scatterplot matrix of the numeric data - (this takes a bit of time)

# BM\_num created above   
library(psych)

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

# pairs works too  
# top right part shows correlations, diagonal shows histograms, and bottom left shows the  
# scatterplots, with the circles showing strength of correlation (circle~little, oval~lot)  
# took out y in the end  
pairs.panels(BM[c("age","balance","day","duration","campaign","pdays","previous")])



# Step 8a: Pearson chi-sq test for correlations of non-numeric data

This is a sample of what could be done if we did not do the heat map above Also: chisq.test(table(BMmarital))$expected and $expected shows what the results should look like if true under null hypothesis

# test for one attribute at a time against all the others for correlation  
# if p<0.05 then not significant   
chisq.test(BM$job, BM$marital)

##   
## Pearson's Chi-squared test  
##   
## data: BM$job and BM$marital  
## X-squared = 3837.6, df = 22, p-value < 2.2e-16

chisq.test(BM$job, BM$education)

##   
## Pearson's Chi-squared test  
##   
## data: BM$job and BM$education  
## X-squared = 28483, df = 33, p-value < 2.2e-16

# all p-values ~ zero - none significant

#step 8b: Pearson chi-sq test for correlations of non-numeric data # extra - test y with all categorical attributes against it

chisq.test(BM$y, BM$job)

##   
## Pearson's Chi-squared test  
##   
## data: BM$y and BM$job  
## X-squared = 836.11, df = 11, p-value < 2.2e-16

chisq.test(BM$y, BM$marital)

##   
## Pearson's Chi-squared test  
##   
## data: BM$y and BM$marital  
## X-squared = 196.5, df = 2, p-value < 2.2e-16

chisq.test(BM$y, BM$education)

##   
## Pearson's Chi-squared test  
##   
## data: BM$y and BM$education  
## X-squared = 238.92, df = 3, p-value < 2.2e-16

chisq.test(BM$y, BM$housing)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: BM$y and BM$housing  
## X-squared = 874.82, df = 1, p-value < 2.2e-16

chisq.test(BM$y, BM$loan)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: BM$y and BM$loan  
## X-squared = 209.62, df = 1, p-value < 2.2e-16

chisq.test(BM$y, BM$contact)

##   
## Pearson's Chi-squared test  
##   
## data: BM$y and BM$contact  
## X-squared = 1035.7, df = 2, p-value < 2.2e-16

chisq.test(BM$y, BM$month)

##   
## Pearson's Chi-squared test  
##   
## data: BM$y and BM$month  
## X-squared = 3061.8, df = 11, p-value < 2.2e-16

chisq.test(BM$y, BM$poutcome)

##   
## Pearson's Chi-squared test  
##   
## data: BM$y and BM$poutcome  
## X-squared = 4391.5, df = 3, p-value < 2.2e-16

# all p-values ~ zero - this is not good! we want correlation!!