Part 1 - data

jean wills

06/07/2020

## Project: Portuguese Bank Marketing Data

# Step 1: read the data and look at the data structure

# setwd("~/Users/jeanwills/Desktop/CKME136/")  
# BM <- read.csv("/Users/jeanwills/Desktop/CKME136/1\_data/bank\_full.csv", header=T, sep = ";", stringsAsFactors = T, na.strings = "NA")  
# look at the data structure   
# since we are using the smaller file, we will do the statistics on the smaller file  
BM\_mini <- read.csv("/Users/jeanwills/Desktop/CKME136/1\_Data/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!  
BM<- BM\_mini  
sum(complete.cases(BM))

## [1] 4521

# step 2 - Look at the bank data summary

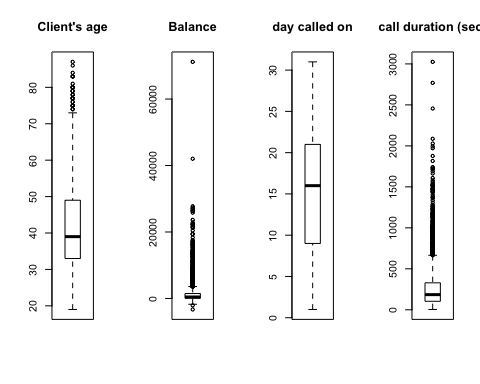
summary(BM)

## age job marital education default   
## Min. :19.00 management :969 divorced: 528 primary : 678 no :4445   
## 1st Qu.:33.00 blue-collar:946 married :2797 secondary:2306 yes: 76   
## Median :39.00 technician :768 single :1196 tertiary :1350   
## Mean :41.17 admin. :478 unknown : 187   
## 3rd Qu.:49.00 services :417   
## Max. :87.00 retired :230   
## (Other) :713   
## balance housing loan contact day   
## Min. :-3313 no :1962 no :3830 cellular :2896 Min. : 1.00   
## 1st Qu.: 69 yes:2559 yes: 691 telephone: 301 1st Qu.: 9.00   
## Median : 444 unknown :1324 Median :16.00   
## Mean : 1423 Mean :15.92   
## 3rd Qu.: 1480 3rd Qu.:21.00   
## Max. :71188 Max. :31.00   
##   
## month duration campaign pdays   
## may :1398 Min. : 4 Min. : 1.000 Min. : -1.00   
## jul : 706 1st Qu.: 104 1st Qu.: 1.000 1st Qu.: -1.00   
## aug : 633 Median : 185 Median : 2.000 Median : -1.00   
## jun : 531 Mean : 264 Mean : 2.794 Mean : 39.77   
## nov : 389 3rd Qu.: 329 3rd Qu.: 3.000 3rd Qu.: -1.00   
## apr : 293 Max. :3025 Max. :50.000 Max. :871.00   
## (Other): 571   
## previous poutcome y   
## Min. : 0.0000 failure: 490 no :4000   
## 1st Qu.: 0.0000 other : 197 yes: 521   
## Median : 0.0000 success: 129   
## Mean : 0.5426 unknown:3705   
## 3rd Qu.: 0.0000   
## Max. :25.0000   
##

# 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

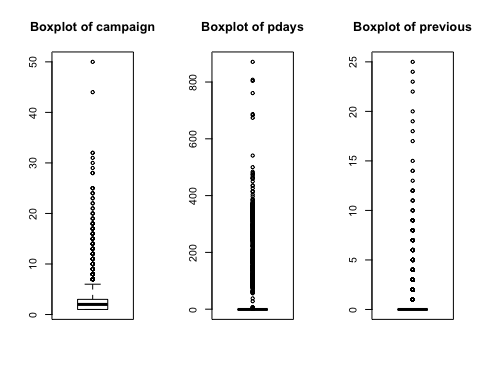
# Step 3a: plot the boxplots of the numeric data

# 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)")



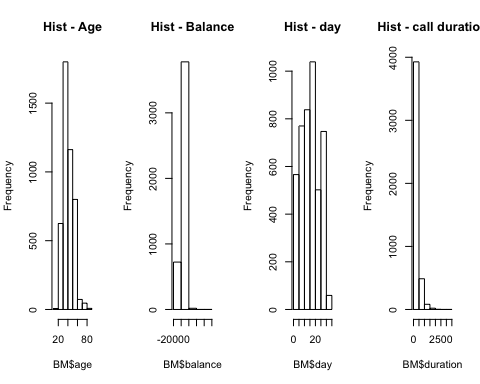
# Step 3b

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")



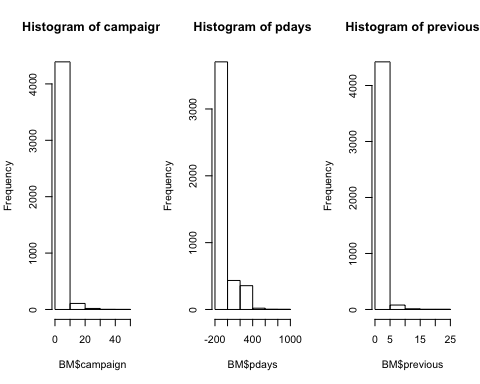
# Step 4a: plot histograms to reveal skewness / normality

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)



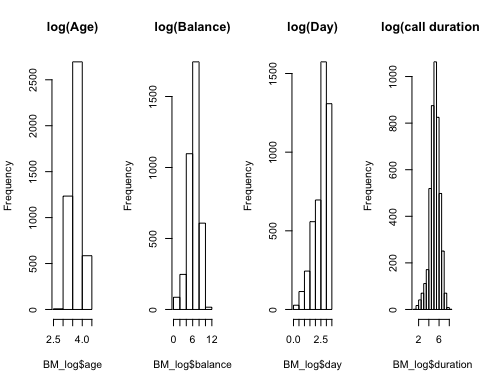
# 4c logs of the numeric data and replot the histograms…

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

# 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)

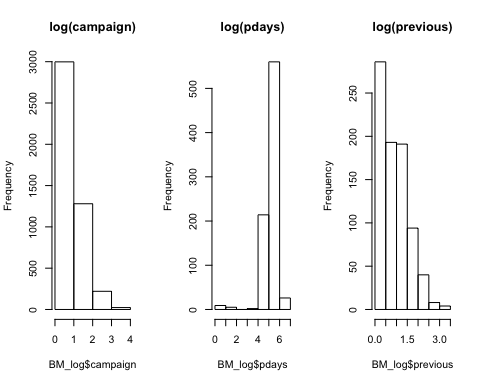


# 4d logs of the numeric data and replot the histograms…

# 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)

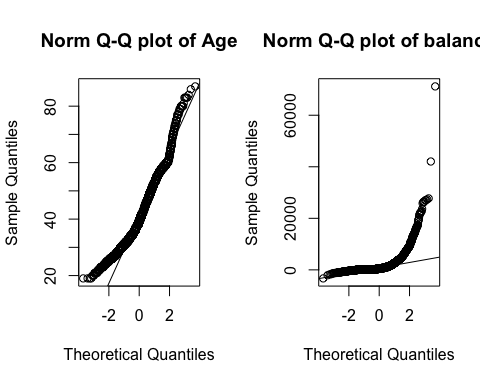


# most seem normal except campaign is still right skewed

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

# Step 5a: Q-Q plots

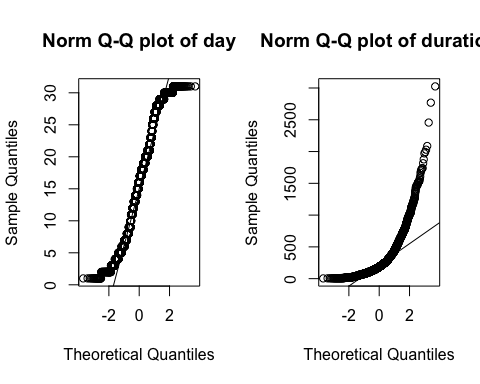
par(mfrow=c(1,2))  
qqnorm(BM$age, main = "Norm Q-Q plot of Age")  
qqline(BM$age)  
qqnorm(BM$balance, main = "Norm Q-Q plot of balance")  
qqline(BM$balance)



# both not normal

# Step 5b: Q-Q plots

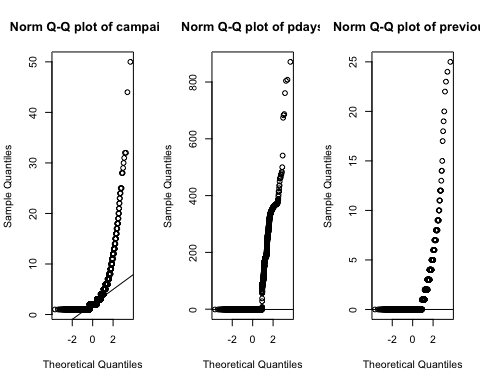
par(mfrow=c(1,2))  
qqnorm(BM$day, main = "Norm Q-Q plot of day")  
qqline(BM$day)  
qqnorm(BM$duration, main = "Norm Q-Q plot of duration")  
qqline(BM$duration)



# 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.  
# 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")

## Warning in cor.test.default(BM$previous, BM$age, method = "spearman"): Cannot  
## compute exact p-value with ties

##   
## Spearman's rank correlation rho  
##   
## data: BM$previous and BM$age  
## S = 1.5391e+10, p-value = 0.9664  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.0006265035

cor.test(BM$previous,BM$age)

##   
## Pearson's product-moment correlation  
##   
## data: BM$previous and BM$age  
## t = -0.23602, df = 4519, p-value = 0.8134  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.03265847 0.02564260  
## sample estimates:  
## cor   
## -0.003510917

# 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 -0.021499961 -0.381485029 -0.121612622 -0.017884654  
## job -0.021499961 1.000000000 0.069390257 0.170160220 0.008323818  
## marital -0.381485029 0.069390257 1.000000000 0.102713866 -0.020744766  
## education -0.121612622 0.170160220 0.102713866 1.000000000 -0.010533920  
## default -0.017884654 0.008323818 -0.020744766 -0.010533920 1.000000000  
## balance 0.083820142 0.009796670 0.024971105 0.057724782 -0.070885956  
## housing -0.193887639 -0.128352529 -0.029851263 -0.087069953 0.006880645  
## loan -0.011249693 -0.040245325 -0.045210303 -0.054085876 0.063993945  
## contact 0.015161382 -0.074067813 -0.068235607 -0.110553557 0.008447985  
## day -0.017852632 0.012865366 0.008794166 0.014925521 -0.013260779  
## month -0.040714424 -0.096612538 -0.035855384 -0.050086264 0.014296700  
## duration -0.002366889 -0.006739013 0.006618952 -0.014878249 -0.011614902  
## campaign -0.005147905 -0.002739255 0.005914831 -0.001723268 -0.012347686  
## pdays -0.008893530 -0.022759577 0.017050091 0.012076680 -0.026316613  
## previous -0.003510917 0.005029474 0.038027745 0.023983386 -0.026656428  
## poutcome -0.009320096 0.013049385 -0.027715900 -0.032134814 0.039031567  
## balance housing loan contact day  
## age 0.083820142 -0.193887639 -0.011249693 0.015161382 -0.017852632  
## job 0.009796670 -0.128352529 -0.040245325 -0.074067813 0.012865366  
## marital 0.024971105 -0.029851263 -0.045210303 -0.068235607 0.008794166  
## education 0.057724782 -0.087069953 -0.054085876 -0.110553557 0.014925521  
## default -0.070885956 0.006880645 0.063993945 0.008447985 -0.013260779  
## balance 1.000000000 -0.050227069 -0.071349288 -0.009664946 -0.008677052  
## housing -0.050227069 1.000000000 0.018450768 0.196454093 -0.031291012  
## loan -0.071349288 0.018450768 1.000000000 -0.007318501 -0.004879370  
## contact -0.009664946 0.196454093 -0.007318501 1.000000000 -0.033807273  
## day -0.008677052 -0.031291012 -0.004879370 -0.033807273 1.000000000  
## month 0.023112568 0.266629529 0.016329353 0.370077452 -0.014794567  
## duration -0.015949918 0.015740454 -0.004997142 -0.011380060 -0.024629306  
## campaign -0.009976166 -0.003573574 0.017119988 0.012278171 0.160706069  
## pdays 0.009436676 0.116893273 -0.031086406 -0.243222943 -0.094351520  
## previous 0.026196357 0.038620583 -0.022114829 -0.187232031 -0.059114394  
## poutcome -0.029268438 -0.093092510 0.027027669 0.267172511 0.073713700  
## month duration campaign pdays previous  
## age -0.0407144239 -0.0023668895 -0.005147905 -0.008893530 -0.003510917  
## job -0.0966125378 -0.0067390133 -0.002739255 -0.022759577 0.005029474  
## marital -0.0358553838 0.0066189516 0.005914831 0.017050091 0.038027745  
## education -0.0500862641 -0.0148782488 -0.001723268 0.012076680 0.023983386  
## default 0.0142967002 -0.0116149018 -0.012347686 -0.026316613 -0.026656428  
## balance 0.0231125680 -0.0159499184 -0.009976166 0.009436676 0.026196357  
## housing 0.2666295288 0.0157404536 -0.003573574 0.116893273 0.038620583  
## loan 0.0163293525 -0.0049971418 0.017119988 -0.031086406 -0.022114829  
## contact 0.3700774517 -0.0113800601 0.012278171 -0.243222943 -0.187232031  
## day -0.0147945673 -0.0246293065 0.160706069 -0.094351520 -0.059114394  
## month 1.0000000000 0.0008513212 -0.108914520 0.033291702 0.046899360  
## duration 0.0008513212 1.0000000000 -0.068382000 0.010380242 0.018080317  
## campaign -0.1089145196 -0.0683819999 1.000000000 -0.093136818 -0.067832630  
## pdays 0.0332917025 0.0103802422 -0.093136818 1.000000000 0.577561827  
## previous 0.0468993605 0.0180803172 -0.067832630 0.577561827 1.000000000  
## poutcome -0.0301894767 0.0004778985 0.110702753 -0.859245131 -0.636371930  
## poutcome  
## age -0.0093200956  
## job 0.0130493854  
## marital -0.0277159001  
## education -0.0321348144  
## default 0.0390315665  
## balance -0.0292684384  
## housing -0.0930925099  
## loan 0.0270276689  
## contact 0.2671725110  
## day 0.0737136998  
## month -0.0301894767  
## duration 0.0004778985  
## campaign 0.1107027535  
## pdays -0.8592451306  
## previous -0.6363719297  
## poutcome 1.0000000000

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

## [1] 16 14

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

## to do this, create all numeric data first (copied from Part 2):

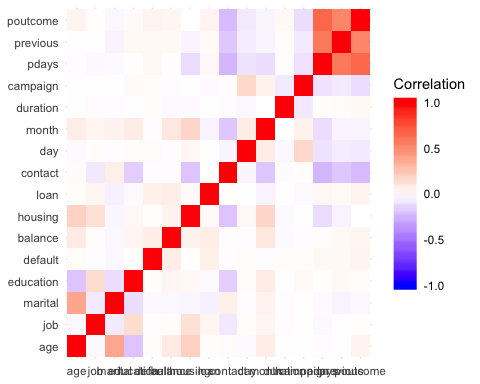
BM\_n <- BM  
BM\_n$job<- as.numeric(BM\_n$job) #12  
# marital: 1-single, 2-married, 3-divorced  
BM\_n$marital<- ifelse(BM\_n$marital == c("single"), 1,   
 ifelse(BM\_n$marital== c("married"), 2, 3))  
# education: 0:unknown, 1: primary, 2:secondary, 3:divorced  
BM\_n$education<- ifelse(BM\_n$education == c("unknown"), 0,   
 ifelse(BM\_n$education == c("primary"), 1,   
 ifelse(BM\_n$education == c("secondary"), 2, 3)))  
# default, housing, loan: if yes then 0 else 1  
# BM\_num$housing<- as.numeric(BM\_num$housing) #2  
BM\_n$default<- ifelse(BM\_n$default == c("yes"), 0, 1) #2  
BM\_n$housing<- ifelse(BM\_n$housing == c("yes"), 0, 1) #2  
BM\_n$loan<- ifelse(BM\_n$loan == c("yes"), 0, 1) #2  
BM\_n$contact<- as.numeric(BM\_n$contact) #3  
# month: jan:1, feb:2.....dec:12  
BM\_n$month<- ifelse(BM\_n$month == "jan", 1,   
 ifelse(BM\_n$month == "feb", 2,   
 ifelse(BM\_n$month == "mar", 3,  
 ifelse(BM\_n$month == "apr", 4,   
 ifelse(BM\_n$month == "may", 5,   
 ifelse(BM\_n$month == "jun", 6,  
 ifelse(BM\_n$month == "jul", 7,  
 ifelse(BM\_n$month == "aug", 8,  
 ifelse(BM\_n$month == "sep", 9,  
 ifelse(BM\_n$month == "oct", 10,  
 ifelse(BM\_n$month == "nov", 11, 12)))))))))))  
# poutcome: 0:unknown,other, 1:failure, 2: success  
BM\_n$poutcome<- ifelse(BM\_n$poutcome == c("failure"), 1, ifelse(BM\_n$poutcome== c("success"), 2, 0))   
# result: BM\_num with only numeric data (NOT scaled)  
# extra step for BM\_mini  
BM\_num<-BM\_n  
rm(BM\_n)

# 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)



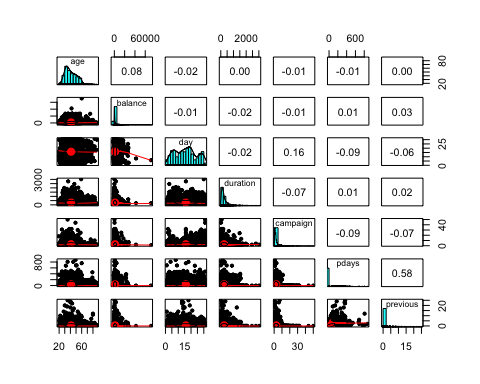
# Step 7d: scatterplot matrix of the numeric data - (this takes a bit of time)

# 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(BM$job, BM$marital))$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)

## Warning in chisq.test(BM$job, BM$marital): Chi-squared approximation may be  
## incorrect

##   
## Pearson's Chi-squared test  
##   
## data: BM$job and BM$marital  
## X-squared = 373.18, df = 22, p-value < 2.2e-16

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

## Warning in chisq.test(BM$job, BM$education): Chi-squared approximation may be  
## incorrect

##   
## Pearson's Chi-squared test  
##   
## data: BM$job and BM$education  
## X-squared = 2840, df = 33, p-value < 2.2e-16

# all p-values ~ zero - none significant

# Step 8b: Pearson chi-sq test for correlations of categorical data to the target variable

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

## Warning in chisq.test(BM$y, BM$job): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: BM$y and BM$job  
## X-squared = 68.988, df = 11, p-value = 1.901e-10

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

##   
## Pearson's Chi-squared test  
##   
## data: BM$y and BM$marital  
## X-squared = 19.03, df = 2, p-value = 7.374e-05

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

##   
## Pearson's Chi-squared test  
##   
## data: BM$y and BM$education  
## X-squared = 15.237, df = 3, p-value = 0.001625

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

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: BM$y and BM$housing  
## X-squared = 48.885, df = 1, p-value = 2.715e-12

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

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: BM$y and BM$loan  
## X-squared = 21.872, df = 1, p-value = 2.915e-06

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

##   
## Pearson's Chi-squared test  
##   
## data: BM$y and BM$contact  
## X-squared = 87.87, df = 2, p-value < 2.2e-16

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

## Warning in chisq.test(BM$y, BM$month): Chi-squared approximation may be  
## incorrect

##   
## Pearson's Chi-squared test  
##   
## data: BM$y and BM$month  
## X-squared = 250.5, 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 = 386.88, df = 3, p-value < 2.2e-16

# all p-values ~ zero - this is not good! we want correlation!!  
# this means that it may be difficult to get meaningful answers!