#assignment 6.1

a. Create a visual for representing missing values in the dataset.

library(ggplot2)

library(data.table)

temp <- tempfile()

download.file('http://archive.ics.uci.edu/ml/machine-learning-databases/00222/bank.zip',temp)

data <- read.table(unz(temp, 'bank-full.csv'), sep = ';', stringsAsFactors = FALSE)

data<- data.table(data)

setnames(data, c('age', 'job', 'marital', 'education', 'default', 'balance', 'housing', 'loan', 'contact',

'day\_of\_week', 'month', 'duration', 'campaign', 'pdays', 'previous', 'poutcome', 'y'))

for(i in c('age', 'job','loan','campaign')) set(data, sample(1:45212, 1000), i, NA)

for(i in c('default', 'housing', 'day\_of\_week', 'previous')) set(data, sample(1:45212, 500),i, NA)

set(data,sample(1:45212, 10000),'education', NA)

data[,.(Names = names(data), Is\_Missing = colSums(is.na(data)))]

# data[,':='(age = as.numeric(age), balance = as.numeric(balance), day\_of\_week = as.numeric(day\_of\_week), duration =as.numeric(duration))]

# data[,':='(campaign = as.numeric(campaign), pdays = as.numeric(pdays), previous = as.numeric(previous))]

plot\_Missing <- function(data\_in, title = NULL){

temp\_df <- as.data.frame(ifelse(is.na(data\_in), 0, 1))

temp\_df <- temp\_df[,order(colSums(temp\_df))]

data\_temp <- expand.grid(list(x = 1:nrow(temp\_df), y = colnames(temp\_df)))

data\_temp$m <- as.vector(as.matrix(temp\_df))

data\_temp <- data.frame(x = unlist(data\_temp$x), y = unlist(data\_temp$y), m = unlist(data\_temp$m))

ggplot(data\_temp) + geom\_tile(aes(x=x, y=y, fill=factor(m))) + scale\_fill\_manual(values=c("white", "black"), name="Missing\n(0=Yes, 1=No)") + theme\_light() + ylab("") + xlab("") + ggtitle(title)

}

plot\_Missing(data[,(colSums(is.na(data)) > 0), with = FALSE], title = "Missing Values")

