Untitled

#Title: "MSDS 6372 Group Project 2: Bank Project - Predicting if a customer will subscribe to a term deposit."  
#Group Members: Nicole Norelli,Nneanna Okpara,Sowmya Mani  
#Date: March 25 2021  
  
#Introduction: This Project is about the Bank Market Analysis to predict if a customer will subscribe to a term deposit  
  
#The data set used for this analysis consists of 2 Data sets:  
  
#Bank Full.csv and Bank.csv  
  
#Bank client data:  
#1 - age (numeric)  
#2 - job : type of job (categorical:"admin.","unknown","unemployed","management","housemaid","entrepreneur","student",  
# "blue-collar","self-employed","retired","technician","services")   
#3 - marital : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)  
#4 - education (categorical: "unknown","secondary","primary","tertiary")  
#5 - default: has credit in default? (binary: "yes","no")  
#6 - balance: average yearly balance, in euros (numeric)   
#7 - housing: has housing loan? (binary: "yes","no")  
#8 - loan: has personal loan? (binary: "yes","no")  
# related with the last contact of the current campaign:  
#9 - contact: contact communication type (categorical: "unknown","telephone","cellular")   
#10 - day: last contact day of the month (numeric)  
#11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")  
#12 - duration: last contact duration, in seconds (numeric)  
# other attributes:  
#13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)  
#14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)  
#15 - previous: number of contacts performed before this campaign and for this client (numeric)  
#16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")  
  
#Output variable (desired target):  
#17 - y - has the client subscribed a term deposit? (binary: "yes","no")

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

#Libraries loaded for the ANalysis  
library(XML)

## Warning: package 'XML' was built under R version 4.0.3

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.0.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(RCurl)

## Warning: package 'RCurl' was built under R version 4.0.3

library(httr)

## Warning: package 'httr' was built under R version 4.0.3

library(jsonlite)

## Warning: package 'jsonlite' was built under R version 4.0.3

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.0.3

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.0.4 v stringr 1.4.0  
## v tidyr 1.1.2 v forcats 0.5.0  
## v readr 1.4.0

## Warning: package 'ggplot2' was built under R version 4.0.3

## Warning: package 'tibble' was built under R version 4.0.3

## Warning: package 'tidyr' was built under R version 4.0.3

## Warning: package 'readr' was built under R version 4.0.3

## Warning: package 'purrr' was built under R version 4.0.3

## Warning: package 'stringr' was built under R version 4.0.3

## Warning: package 'forcats' was built under R version 4.0.3

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x tidyr::complete() masks RCurl::complete()  
## x dplyr::filter() masks stats::filter()  
## x purrr::flatten() masks jsonlite::flatten()  
## x dplyr::lag() masks stats::lag()

library(naniar)

## Warning: package 'naniar' was built under R version 4.0.3

library(GGally)

## Warning: package 'GGally' was built under R version 4.0.3

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

library(ggplot2)  
library(class)  
library(caret)

## Warning: package 'caret' was built under R version 4.0.3

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

## The following object is masked from 'package:httr':  
##   
## progress

library(knnp)

## Warning: package 'knnp' was built under R version 4.0.3

##   
## Attaching package: 'knnp'

## The following object is masked from 'package:class':  
##   
## knn

library(e1071)

## Warning: package 'e1071' was built under R version 4.0.3

library(ggplot2)  
library(maps)

## Warning: package 'maps' was built under R version 4.0.3

##   
## Attaching package: 'maps'

## The following object is masked from 'package:purrr':  
##   
## map

library(dplyr)  
library(mapproj)

## Warning: package 'mapproj' was built under R version 4.0.3

library(ggplot2)  
library(dplyr)  
library(ggcorrplot)

## Warning: package 'ggcorrplot' was built under R version 4.0.3

#Import the Bank Full Data  
Bank\_Full<-read.csv('C:/Sowmya/SMU/03\_Applied Stats/Group Project 2/bank-full.csv' ,sep=";")  
  
#Quick Peek at the SUmmary data of the available dataset  
summary(Bank\_Full)

## age job marital education   
## Min. :18.00 Length:45211 Length:45211 Length:45211   
## 1st Qu.:33.00 Class :character Class :character Class :character   
## Median :39.00 Mode :character Mode :character Mode :character   
## Mean :40.94   
## 3rd Qu.:48.00   
## Max. :95.00   
## default balance housing loan   
## Length:45211 Min. : -8019 Length:45211 Length:45211   
## Class :character 1st Qu.: 72 Class :character Class :character   
## Mode :character Median : 448 Mode :character Mode :character   
## Mean : 1362   
## 3rd Qu.: 1428   
## Max. :102127   
## contact day month duration   
## Length:45211 Min. : 1.00 Length:45211 Min. : 0.0   
## Class :character 1st Qu.: 8.00 Class :character 1st Qu.: 103.0   
## Mode :character Median :16.00 Mode :character Median : 180.0   
## Mean :15.81 Mean : 258.2   
## 3rd Qu.:21.00 3rd Qu.: 319.0   
## Max. :31.00 Max. :4918.0   
## campaign pdays previous poutcome   
## Min. : 1.000 Min. : -1.0 Min. : 0.0000 Length:45211   
## 1st Qu.: 1.000 1st Qu.: -1.0 1st Qu.: 0.0000 Class :character   
## Median : 2.000 Median : -1.0 Median : 0.0000 Mode :character   
## Mean : 2.764 Mean : 40.2 Mean : 0.5803   
## 3rd Qu.: 3.000 3rd Qu.: -1.0 3rd Qu.: 0.0000   
## Max. :63.000 Max. :871.0 Max. :275.0000   
## y   
## Length:45211   
## Class :character   
## Mode :character   
##   
##   
##

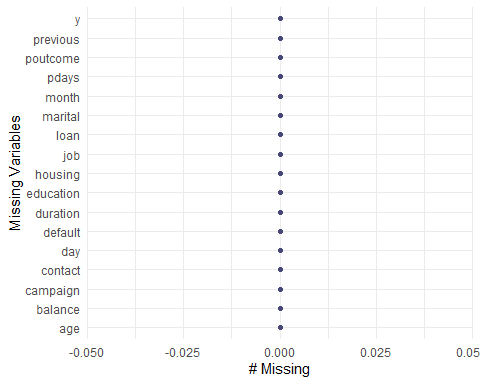
str(Bank\_Full)

## 'data.frame': 45211 obs. of 17 variables:  
## $ age : int 58 44 33 47 33 35 28 42 58 43 ...  
## $ job : chr "management" "technician" "entrepreneur" "blue-collar" ...  
## $ marital : chr "married" "single" "married" "married" ...  
## $ education: chr "tertiary" "secondary" "secondary" "unknown" ...  
## $ default : chr "no" "no" "no" "no" ...  
## $ balance : int 2143 29 2 1506 1 231 447 2 121 593 ...  
## $ housing : chr "yes" "yes" "yes" "yes" ...  
## $ loan : chr "no" "no" "yes" "no" ...  
## $ contact : chr "unknown" "unknown" "unknown" "unknown" ...  
## $ day : int 5 5 5 5 5 5 5 5 5 5 ...  
## $ month : chr "may" "may" "may" "may" ...  
## $ duration : int 261 151 76 92 198 139 217 380 50 55 ...  
## $ campaign : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ pdays : int -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...  
## $ previous : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ poutcome : chr "unknown" "unknown" "unknown" "unknown" ...  
## $ y : chr "no" "no" "no" "no" ...

#Checking for Missing Data  
sapply(Bank\_Full,function(x) sum(is.na(x)))

## age job marital education default balance housing loan   
## 0 0 0 0 0 0 0 0   
## contact day month duration campaign pdays previous poutcome   
## 0 0 0 0 0 0 0 0   
## y   
## 0

gg\_miss\_var(Bank\_Full)+xlab("Missing Variables")



#The Bank dataset has 45,211 observations with 17 variables providing more information on the Bank Clients.There is no missing data in the data set.  
  
#Test Data set  
Bank\_Test<-read.csv('C:/Sowmya/SMU/03\_Applied Stats/Group Project 2/bank.csv' ,sep=";")  
  
#Quick Peek at the SUmmary data of the available dataset  
summary(Bank\_Test)

## age job marital education   
## Min. :19.00 Length:4521 Length:4521 Length:4521   
## 1st Qu.:33.00 Class :character Class :character Class :character   
## Median :39.00 Mode :character Mode :character Mode :character   
## Mean :41.17   
## 3rd Qu.:49.00   
## Max. :87.00   
## default balance housing loan   
## Length:4521 Min. :-3313 Length:4521 Length:4521   
## Class :character 1st Qu.: 69 Class :character Class :character   
## Mode :character Median : 444 Mode :character Mode :character   
## Mean : 1423   
## 3rd Qu.: 1480   
## Max. :71188   
## contact day month duration   
## Length:4521 Min. : 1.00 Length:4521 Min. : 4   
## Class :character 1st Qu.: 9.00 Class :character 1st Qu.: 104   
## Mode :character Median :16.00 Mode :character Median : 185   
## Mean :15.92 Mean : 264   
## 3rd Qu.:21.00 3rd Qu.: 329   
## Max. :31.00 Max. :3025   
## campaign pdays previous poutcome   
## Min. : 1.000 Min. : -1.00 Min. : 0.0000 Length:4521   
## 1st Qu.: 1.000 1st Qu.: -1.00 1st Qu.: 0.0000 Class :character   
## Median : 2.000 Median : -1.00 Median : 0.0000 Mode :character   
## Mean : 2.794 Mean : 39.77 Mean : 0.5426   
## 3rd Qu.: 3.000 3rd Qu.: -1.00 3rd Qu.: 0.0000   
## Max. :50.000 Max. :871.00 Max. :25.0000   
## y   
## Length:4521   
## Class :character   
## Mode :character   
##   
##   
##

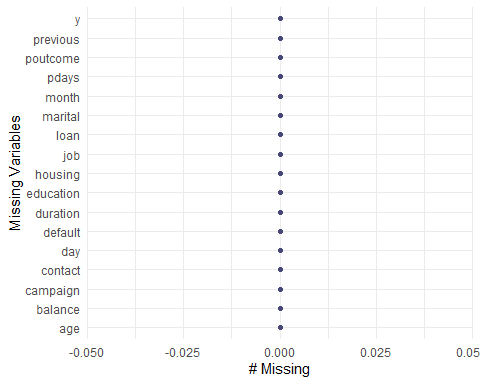
str(Bank\_Test)

## 'data.frame': 4521 obs. of 17 variables:  
## $ age : int 30 33 35 30 59 35 36 39 41 43 ...  
## $ job : chr "unemployed" "services" "management" "management" ...  
## $ marital : chr "married" "married" "single" "married" ...  
## $ education: chr "primary" "secondary" "tertiary" "tertiary" ...  
## $ default : chr "no" "no" "no" "no" ...  
## $ balance : int 1787 4789 1350 1476 0 747 307 147 221 -88 ...  
## $ housing : chr "no" "yes" "yes" "yes" ...  
## $ loan : chr "no" "yes" "no" "yes" ...  
## $ contact : chr "cellular" "cellular" "cellular" "unknown" ...  
## $ day : int 19 11 16 3 5 23 14 6 14 17 ...  
## $ month : chr "oct" "may" "apr" "jun" ...  
## $ duration : int 79 220 185 199 226 141 341 151 57 313 ...  
## $ campaign : int 1 1 1 4 1 2 1 2 2 1 ...  
## $ pdays : int -1 339 330 -1 -1 176 330 -1 -1 147 ...  
## $ previous : int 0 4 1 0 0 3 2 0 0 2 ...  
## $ poutcome : chr "unknown" "failure" "failure" "unknown" ...  
## $ y : chr "no" "no" "no" "no" ...

#Checking for Missing Data  
sapply(Bank\_Test,function(x) sum(is.na(x)))

## age job marital education default balance housing loan   
## 0 0 0 0 0 0 0 0   
## contact day month duration campaign pdays previous poutcome   
## 0 0 0 0 0 0 0 0   
## y   
## 0

gg\_miss\_var(Bank\_Test)+xlab("Missing Variables")



#The Test data set has 4,521 observations with 17 variables with no missing variables

## Including Plots

You can also embed plots, for example:

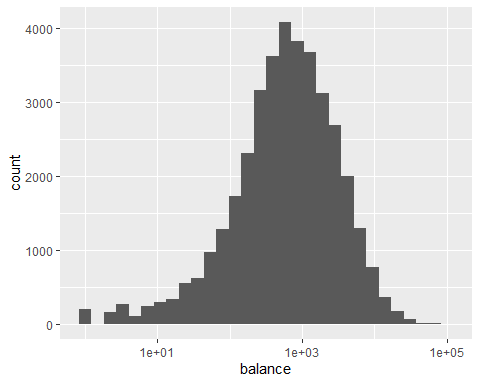
#sapply(Bank\_Full, function(x) sum(x %in% common\_na\_strings)) # missing values using other than NA  
#sapply(Bank\_Full, function(x) sum(x %in% common\_na\_numbers)) # missing values using other than NA  
#sapply(Bank\_Full, function(x) sum(is.na(x)))  
#sapply(Bank\_Full, function(x) sum(x %in% 'unknown'))  
  
#Analysing the data with plots  
Bank\_Full%>%ggplot(aes(x=balance))+geom\_histogram()+ scale\_x\_log10()

## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous x-axis

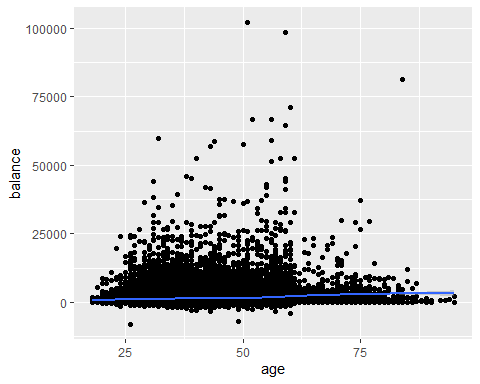
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 7280 rows containing non-finite values (stat\_bin).



#log of balance shows the data is normally distributed  
  
#Analysing the age and balance  
Bank\_Full%>%ggplot(aes(x=age,y=balance))+geom\_point()+geom\_smooth()

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

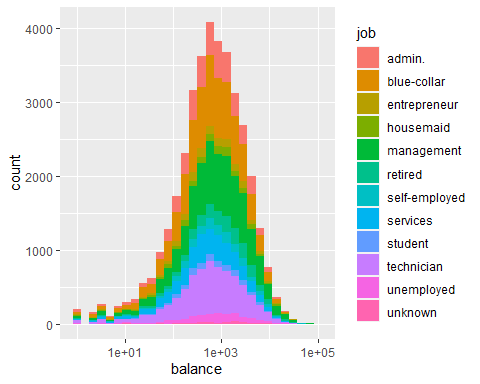


Bank\_Full%>%ggplot(aes(x=balance,fill=job))+geom\_histogram()+ scale\_x\_log10()

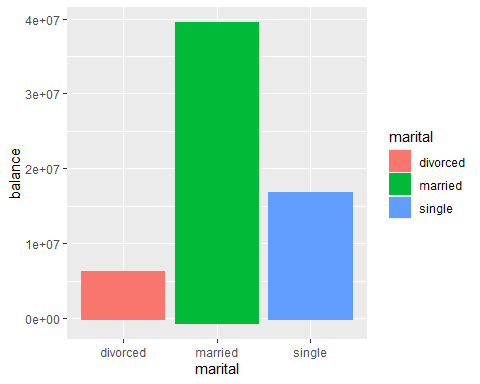
## Warning in self$trans$transform(x): NaNs produced  
  
## Warning in self$trans$transform(x): Transformation introduced infinite values in  
## continuous x-axis

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

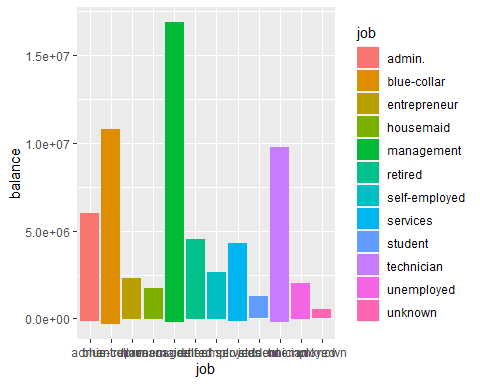
## Warning: Removed 7280 rows containing non-finite values (stat\_bin).



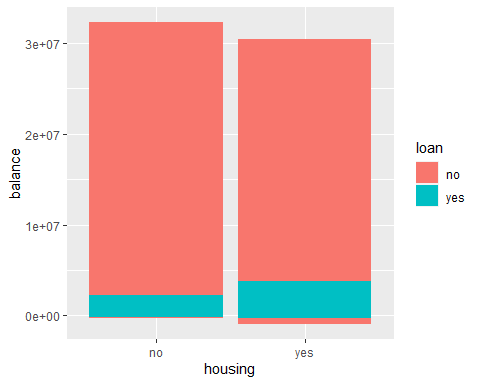
Bank\_Full%>%ggplot(aes(x=marital,y=balance,fill=marital))+geom\_col()



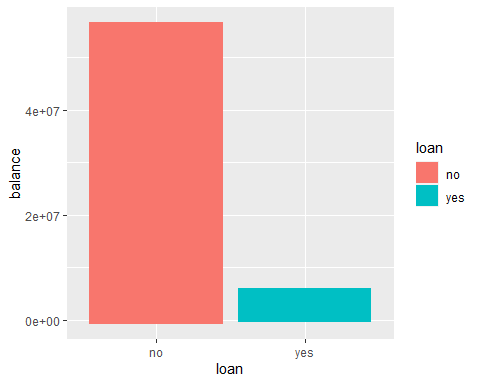
Bank\_Full%>%ggplot(aes(x=job,y=balance,fill=job))+geom\_col()



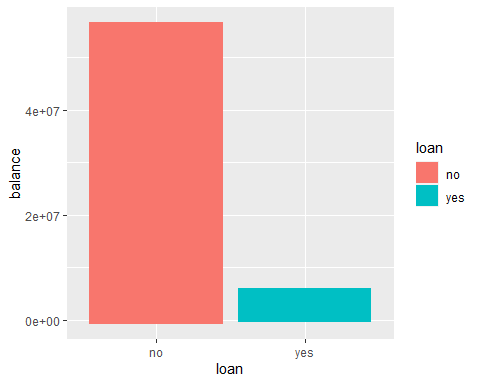
Bank\_Full%>%ggplot(aes(x=housing,y=balance,fill=loan))+geom\_col()



Bank\_Full%>%ggplot(aes(x=loan,y=balance,fill=loan))+geom\_col()



Bank\_Full%>%ggplot(aes(x=loan,y=balance,fill=loan))+geom\_col()

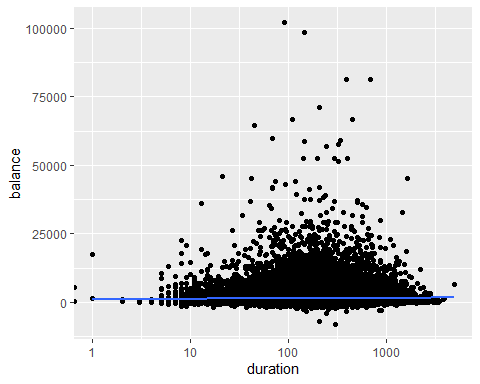


Bank\_Full%>%ggplot(aes(x=duration,y=balance))+geom\_point()+ scale\_x\_log10()+geom\_smooth()

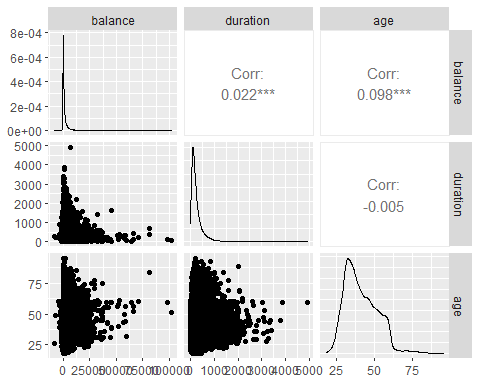
## Warning: Transformation introduced infinite values in continuous x-axis  
  
## Warning: Transformation introduced infinite values in continuous x-axis

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

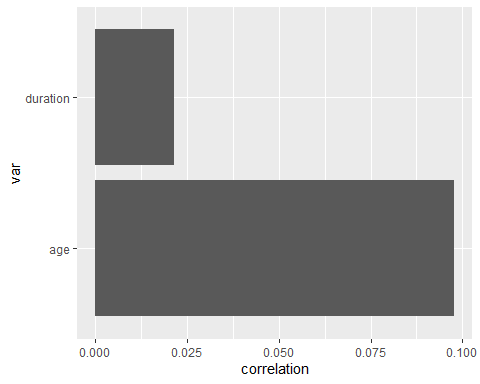
## Warning: Removed 3 rows containing non-finite values (stat\_smooth).



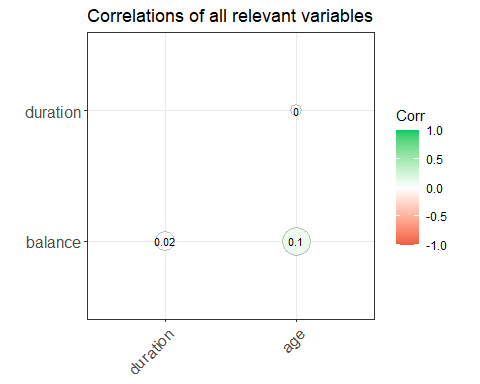
Bank\_gg<-Bank\_Full%>%select(balance, duration,age )  
  
ggpairs(Bank\_gg)



cor.xy <- cor(Bank\_Full %>% dplyr::select(balance, duration,age ), use = "complete.obs")  
LE.cor <- data.frame(var = rownames(cor.xy)[-1], correlation = cor.xy[-1, 1])  
LE.cor %>% ggplot(aes(x = var, y = correlation)) + geom\_col() + coord\_flip()



corr <- Bank\_Full %>% dplyr::select(balance, duration,age)   
corr <- round(cor(corr), 2)  
ggcorrplot(corr, type = "lower",  
 lab = TRUE, lab\_size = 3, method = "circle",  
 colors = c("tomato2", "white", "springgreen3"),  
 title = "Correlations of all relevant variables",  
 ggtheme = theme\_bw())



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.