MSDS6372\_project2

Balaji Avvaru

3/11/2020

## 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

library(gmodels)

## Warning: package 'gmodels' was built under R version 3.6.3

library(dplyr)

##   
## 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(ggplot2)  
library(tidyr)  
library(SDMTools)

## Warning: package 'SDMTools' was built under R version 3.6.2

library(readr)  
library(digest)  
library(ISLR)

## Warning: package 'ISLR' was built under R version 3.6.2

library(car)

## Warning: package 'car' was built under R version 3.6.2

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

library(leaps)

## Warning: package 'leaps' was built under R version 3.6.2

library( Matrix)

##   
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

library(foreach)

## Warning: package 'foreach' was built under R version 3.6.2

library(glmnet)

## Warning: package 'glmnet' was built under R version 3.6.2

## Loaded glmnet 3.0-2

library(VIM)

## Warning: package 'VIM' was built under R version 3.6.2

## Loading required package: colorspace

## Loading required package: grid

## Loading required package: data.table

## Warning: package 'data.table' was built under R version 3.6.2

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

## VIM is ready to use.   
## Since version 4.0.0 the GUI is in its own package VIMGUI.  
##   
## Please use the package to use the new (and old) GUI.

## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues

##   
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':  
##   
## sleep

library(mice)

## Warning: package 'mice' was built under R version 3.6.2

## Loading required package: lattice

## Registered S3 methods overwritten by 'lme4':  
## method from  
## cooks.distance.influence.merMod car   
## influence.merMod car   
## dfbeta.influence.merMod car   
## dfbetas.influence.merMod car

##   
## Attaching package: 'mice'

## The following object is masked from 'package:tidyr':  
##   
## complete

## The following objects are masked from 'package:base':  
##   
## cbind, rbind

library(corrgram)

## Warning: package 'corrgram' was built under R version 3.6.2

## Registered S3 method overwritten by 'seriation':  
## method from   
## reorder.hclust gclus

##   
## Attaching package: 'corrgram'

## The following object is masked from 'package:lattice':  
##   
## panel.fill

library(car)  
library(tidyverse)

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

## v tibble 2.1.3 v stringr 1.4.0  
## v purrr 0.3.3 v forcats 0.4.0

## -- Conflicts --------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x purrr::accumulate() masks foreach::accumulate()  
## x data.table::between() masks dplyr::between()  
## x mice::complete() masks tidyr::complete()  
## x Matrix::expand() masks tidyr::expand()  
## x dplyr::filter() masks stats::filter()  
## x data.table::first() masks dplyr::first()  
## x dplyr::lag() masks stats::lag()  
## x data.table::last() masks dplyr::last()  
## x Matrix::pack() masks tidyr::pack()  
## x car::recode() masks dplyr::recode()  
## x purrr::some() masks car::some()  
## x purrr::transpose() masks data.table::transpose()  
## x Matrix::unpack() masks tidyr::unpack()  
## x purrr::when() masks foreach::when()

library(limma)  
library(gridExtra)

## Warning: package 'gridExtra' was built under R version 3.6.2

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

## Load the data

setwd("C:/Users/BelajiAvvaru/Desktop/Docs/SMU/MSDS 6372/Project 2/")  
#bankdata<- read.csv(file.choose())  
bankdata<-read.csv("bank-additional-full.csv",header=T, sep=";")  
head(bankdata)

## age job marital education default housing loan contact month  
## 1 56 housemaid married basic.4y no no no telephone may  
## 2 57 services married high.school unknown no no telephone may  
## 3 37 services married high.school no yes no telephone may  
## 4 40 admin. married basic.6y no no no telephone may  
## 5 56 services married high.school no no yes telephone may  
## 6 45 services married basic.9y unknown no no telephone may  
## day\_of\_week duration campaign pdays previous poutcome emp.var.rate  
## 1 mon 261 1 999 0 nonexistent 1.1  
## 2 mon 149 1 999 0 nonexistent 1.1  
## 3 mon 226 1 999 0 nonexistent 1.1  
## 4 mon 151 1 999 0 nonexistent 1.1  
## 5 mon 307 1 999 0 nonexistent 1.1  
## 6 mon 198 1 999 0 nonexistent 1.1  
## cons.price.idx cons.conf.idx euribor3m nr.employed y  
## 1 93.994 -36.4 4.857 5191 no  
## 2 93.994 -36.4 4.857 5191 no  
## 3 93.994 -36.4 4.857 5191 no  
## 4 93.994 -36.4 4.857 5191 no  
## 5 93.994 -36.4 4.857 5191 no  
## 6 93.994 -36.4 4.857 5191 no

## number of rows, number of columns and structure of bank data

nrow(bankdata)

## [1] 41188

ncol(bankdata)

## [1] 21

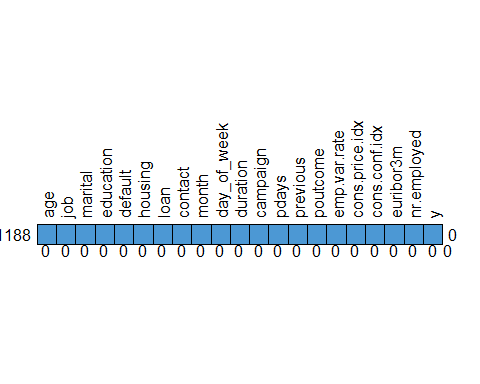
str(bankdata)

## 'data.frame': 41188 obs. of 21 variables:  
## $ age : int 56 57 37 40 56 45 59 41 24 25 ...  
## $ job : Factor w/ 12 levels "admin.","blue-collar",..: 4 8 8 1 8 8 1 2 10 8 ...  
## $ marital : Factor w/ 4 levels "divorced","married",..: 2 2 2 2 2 2 2 2 3 3 ...  
## $ education : Factor w/ 8 levels "basic.4y","basic.6y",..: 1 4 4 2 4 3 6 8 6 4 ...  
## $ default : Factor w/ 3 levels "no","unknown",..: 1 2 1 1 1 2 1 2 1 1 ...  
## $ housing : Factor w/ 3 levels "no","unknown",..: 1 1 3 1 1 1 1 1 3 3 ...  
## $ loan : Factor w/ 3 levels "no","unknown",..: 1 1 1 1 3 1 1 1 1 1 ...  
## $ contact : Factor w/ 2 levels "cellular","telephone": 2 2 2 2 2 2 2 2 2 2 ...  
## $ month : Factor w/ 10 levels "apr","aug","dec",..: 7 7 7 7 7 7 7 7 7 7 ...  
## $ day\_of\_week : Factor w/ 5 levels "fri","mon","thu",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ duration : int 261 149 226 151 307 198 139 217 380 50 ...  
## $ campaign : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ pdays : int 999 999 999 999 999 999 999 999 999 999 ...  
## $ previous : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ poutcome : Factor w/ 3 levels "failure","nonexistent",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ emp.var.rate : num 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...  
## $ cons.price.idx: num 94 94 94 94 94 ...  
## $ cons.conf.idx : num -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 ...  
## $ euribor3m : num 4.86 4.86 4.86 4.86 4.86 ...  
## $ nr.employed : num 5191 5191 5191 5191 5191 ...  
## $ y : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...

# missing data

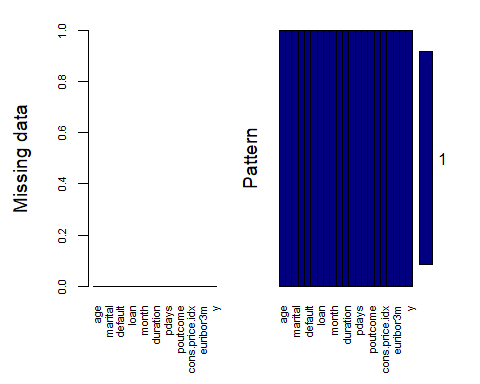
#Display missing-data patterns  
md.pattern(bankdata, plot=TRUE, rotate.names = TRUE)

## /\ /\  
## { `---' }  
## { O O }  
## ==> V <== No need for mice. This data set is completely observed.  
## \ \|/ /  
## `-----'



## age job marital education default housing loan contact month day\_of\_week  
## 41188 1 1 1 1 1 1 1 1 1 1  
## 0 0 0 0 0 0 0 0 0 0  
## duration campaign pdays previous poutcome emp.var.rate cons.price.idx  
## 41188 1 1 1 1 1 1 1  
## 0 0 0 0 0 0 0  
## cons.conf.idx euribor3m nr.employed y   
## 41188 1 1 1 1 0  
## 0 0 0 0 0

#Display missing-data in a bar-plot  
mice\_plot <- aggr(bankdata, col=c('navyblue','yellow'),  
 numbers=TRUE, sortVars=TRUE,  
 labels=names(bankdata), cex.axis=.7,  
 gap=3, ylab=c("Missing data","Pattern"))



##   
## Variables sorted by number of missings:   
## Variable Count  
## age 0  
## job 0  
## marital 0  
## education 0  
## default 0  
## housing 0  
## loan 0  
## contact 0  
## month 0  
## day\_of\_week 0  
## duration 0  
## campaign 0  
## pdays 0  
## previous 0  
## poutcome 0  
## emp.var.rate 0  
## cons.price.idx 0  
## cons.conf.idx 0  
## euribor3m 0  
## nr.employed 0  
## y 0

## Summary of response variable

summary(bankdata$y)

## no yes   
## 36548 4640

## EDA on continuous variables

##“age” “duration” “campaign” “pdays” “previous” “emp.var.rate” “cons.price.idx” ## “cons.conf.idx” “euribor3m” “nr.employed”

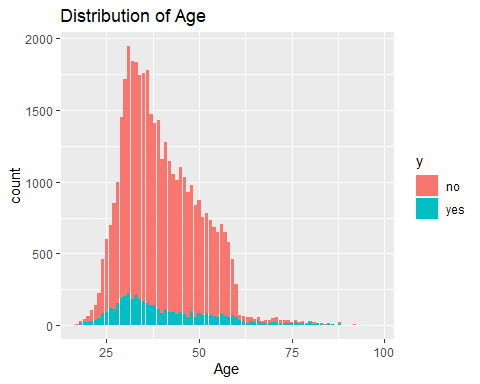
names(bankdata)[sapply(bankdata, class) == "integer" | sapply(bankdata, class) == "numeric"]

## [1] "age" "duration" "campaign" "pdays"   
## [5] "previous" "emp.var.rate" "cons.price.idx" "cons.conf.idx"   
## [9] "euribor3m" "nr.employed"

## Summary on age variable  
summary(bankdata$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 17.00 32.00 38.00 40.02 47.00 98.00

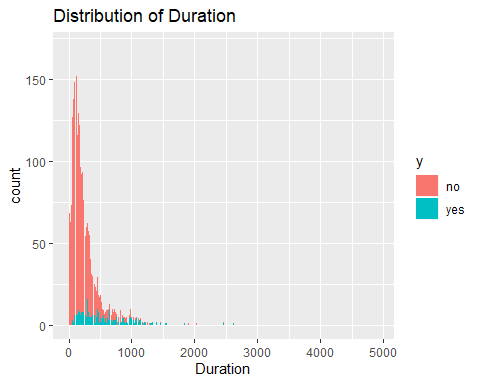
bankdata %>% ggplot(aes(x = age, fill = y)) + geom\_bar() + ggtitle("Distribution of Age") + xlab("Age")



## Summary on duration variable  
  
summary(bankdata$duration)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 102.0 180.0 258.3 319.0 4918.0

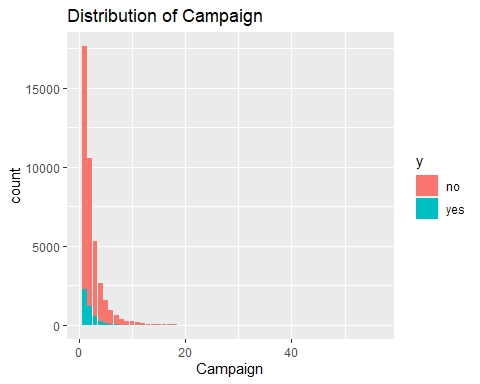
bankdata %>% ggplot(aes(x = duration, fill = y)) + geom\_bar() + ggtitle("Distribution of Duration") + xlab("Duration")



## Summary on campaign variable  
  
summary(bankdata$campaign)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 1.000 2.000 2.568 3.000 56.000

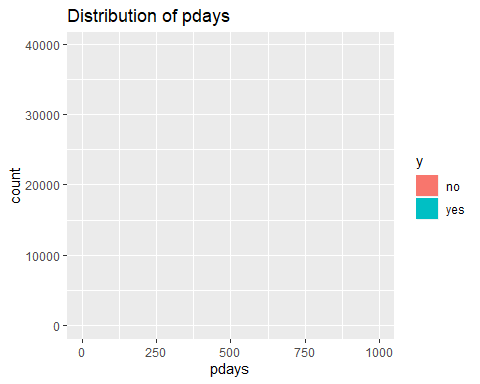
bankdata %>% ggplot(aes(x = campaign, fill = y)) + geom\_bar() + ggtitle("Distribution of Campaign") + xlab("Campaign")



## Summary on pdays variable  
  
summary(bankdata$pdays)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 999.0 999.0 962.5 999.0 999.0

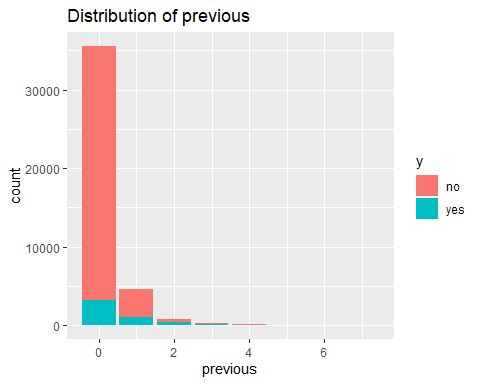
bankdata %>% ggplot(aes(x = pdays, fill = y)) + geom\_bar() + ggtitle("Distribution of pdays") + xlab("pdays")



## Summary on previous variable  
  
summary(bankdata$previous)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 0.000 0.173 0.000 7.000

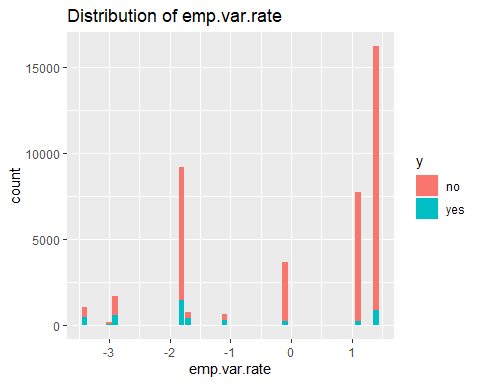
bankdata %>% ggplot(aes(x = previous, fill = y)) + geom\_bar() + ggtitle("Distribution of previous") + xlab("previous")



## Summary on emp.var.rate variable  
  
summary(bankdata$emp.var.rate)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -3.40000 -1.80000 1.10000 0.08189 1.40000 1.40000

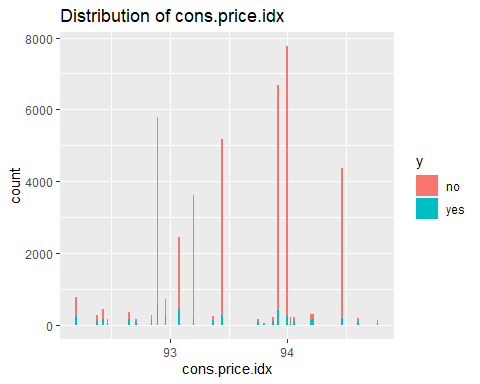
bankdata %>% ggplot(aes(x = emp.var.rate, fill = y)) + geom\_bar() + ggtitle("Distribution of emp.var.rate") + xlab("emp.var.rate")



## Summary on cons.price.idx variable  
  
summary(bankdata$cons.price.idx)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 92.20 93.08 93.75 93.58 93.99 94.77

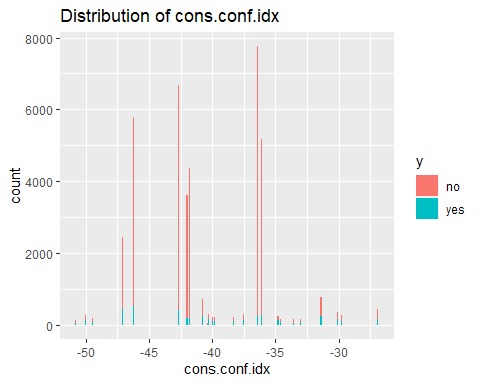
bankdata %>% ggplot(aes(x = cons.price.idx, fill = y)) + geom\_bar() + ggtitle("Distribution of cons.price.idx") + xlab("cons.price.idx")



## Summary on cons.conf.idx variable  
  
summary(bankdata$cons.conf.idx)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -50.8 -42.7 -41.8 -40.5 -36.4 -26.9

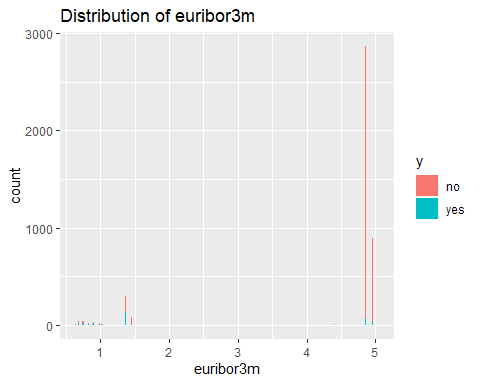
bankdata %>% ggplot(aes(x = cons.conf.idx, fill = y)) + geom\_bar() + ggtitle("Distribution of cons.conf.idx") + xlab("cons.conf.idx")



## Summary on euribor3m variable  
  
summary(bankdata$euribor3m)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.634 1.344 4.857 3.621 4.961 5.045

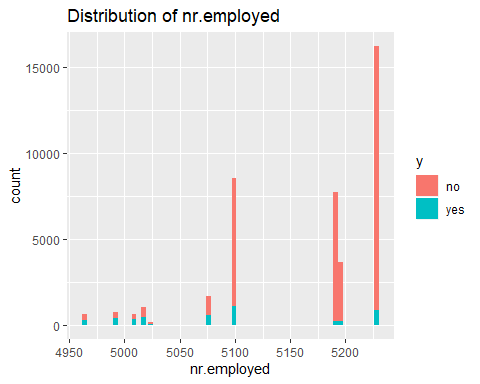
bankdata %>% ggplot(aes(x = euribor3m, fill = y)) + geom\_bar() + ggtitle("Distribution of euribor3m") + xlab("euribor3m")



## Summary on nr.employed variable  
  
summary(bankdata$nr.employed)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4964 5099 5191 5167 5228 5228

bankdata %>% ggplot(aes(x = nr.employed, fill = y)) + geom\_bar() + ggtitle("Distribution of nr.employed") + xlab("nr.employed")



## EDA on categorical variables

## “job” “marital” “education” “default” “housing” “loan” “contact” “month”

## “day\_of\_week” “poutcome”

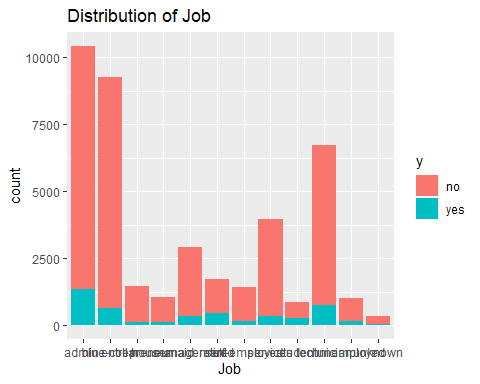
names(bankdata)[sapply(bankdata, class) == "factor"]

## [1] "job" "marital" "education" "default" "housing"   
## [6] "loan" "contact" "month" "day\_of\_week" "poutcome"   
## [11] "y"

## Summary on Job variable  
summary(bankdata$job)

## admin. blue-collar entrepreneur housemaid management   
## 10422 9254 1456 1060 2924   
## retired self-employed services student technician   
## 1720 1421 3969 875 6743   
## unemployed unknown   
## 1014 330

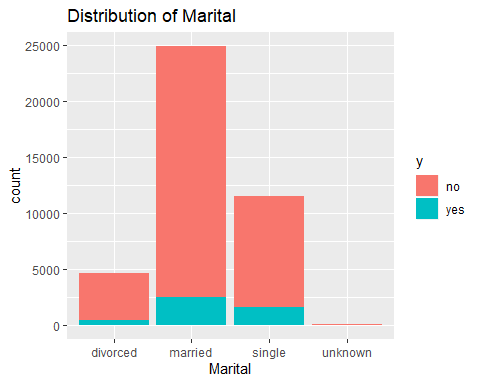
bankdata %>% ggplot(aes(x = job, fill = y)) + geom\_bar() + ggtitle("Distribution of Job") + xlab("Job")



## Summary on marital variable  
  
summary(bankdata$marital)

## divorced married single unknown   
## 4612 24928 11568 80

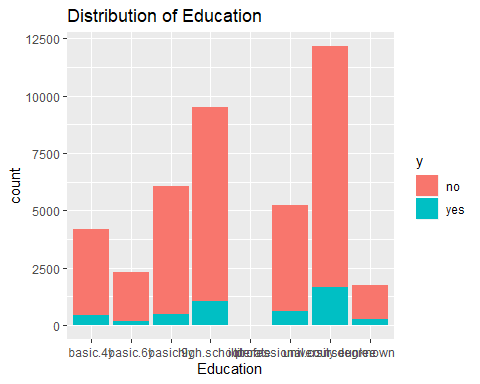
bankdata %>% ggplot(aes(x = marital, fill = y)) + geom\_bar() + ggtitle("Distribution of Marital") + xlab("Marital")



## Summary on education variable  
  
summary(bankdata$education)

## basic.4y basic.6y basic.9y high.school   
## 4176 2292 6045 9515   
## illiterate professional.course university.degree unknown   
## 18 5243 12168 1731

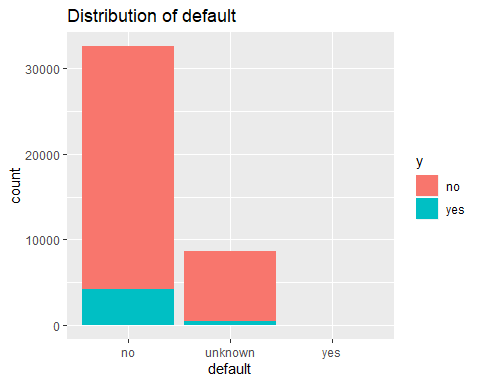
bankdata %>% ggplot(aes(x = education, fill = y)) + geom\_bar() + ggtitle("Distribution of Education") + xlab("Education")



## Summary on default variable  
  
summary(bankdata$default)

## no unknown yes   
## 32588 8597 3

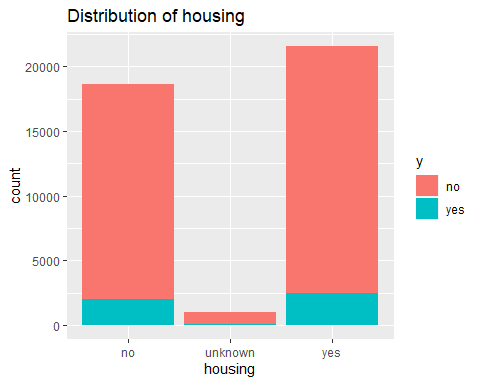
bankdata %>% ggplot(aes(x = default, fill = y)) + geom\_bar() + ggtitle("Distribution of default") + xlab("default")



## Summary on housing variable  
  
summary(bankdata$housing)

## no unknown yes   
## 18622 990 21576

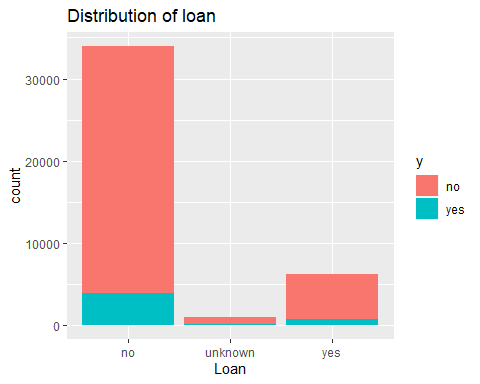
bankdata %>% ggplot(aes(x = housing, fill = y)) + geom\_bar() + ggtitle("Distribution of housing") + xlab("housing")



## Summary on Loan variable  
  
summary(bankdata$loan)

## no unknown yes   
## 33950 990 6248

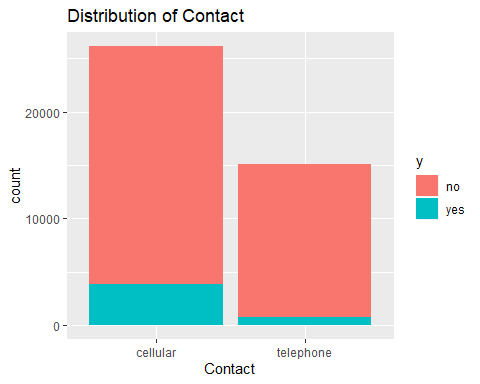
bankdata %>% ggplot(aes(x = loan, fill = y)) + geom\_bar() + ggtitle("Distribution of loan") + xlab("Loan")



## Summary on Contact variable  
  
summary(bankdata$contact)

## cellular telephone   
## 26144 15044

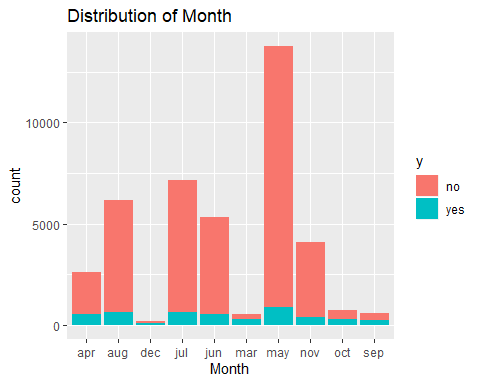
bankdata %>% ggplot(aes(x = contact, fill = y)) + geom\_bar() + ggtitle("Distribution of Contact") + xlab("Contact")



## Summary on Month variable  
  
summary(bankdata$month)

## apr aug dec jul jun mar may nov oct sep   
## 2632 6178 182 7174 5318 546 13769 4101 718 570

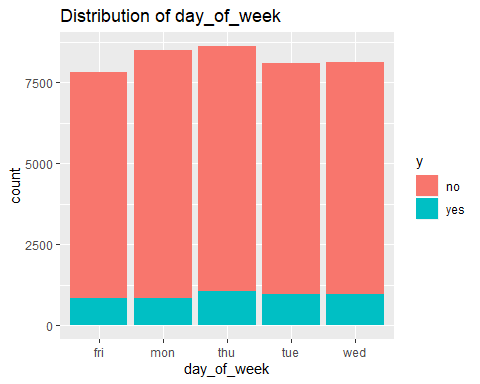
bankdata %>% ggplot(aes(x = month, fill = y)) + geom\_bar() + ggtitle("Distribution of Month") + xlab("Month")



## Summary on day\_of\_week variable  
  
summary(bankdata$day\_of\_week)

## fri mon thu tue wed   
## 7827 8514 8623 8090 8134

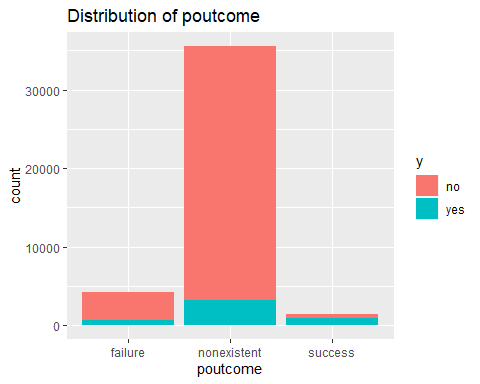
bankdata %>% ggplot(aes(x = day\_of\_week, fill = y)) + geom\_bar() + ggtitle("Distribution of day\_of\_week") + xlab("day\_of\_week")



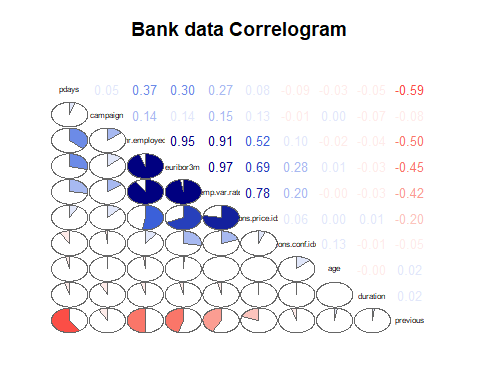
## Summary on poutcome variable  
  
summary(bankdata$poutcome)

## failure nonexistent success   
## 4252 35563 1373

bankdata %>% ggplot(aes(x = poutcome, fill = y)) + geom\_bar() + ggtitle("Distribution of poutcome") + xlab("poutcome")



#correlation plot for all continuous variables  
corrgram(bankdata, order=TRUE,  
 upper.panel=panel.cor, lower.panel=panel.pie, main="Bank data Correlogram")



## euribor3m and nr.employed are highly correlated (0.95)

## emp.var.rate and euribor3m are highly correlated (0.97)

## emp.var.rate and nr.employed are highly correlated (0.91)