Exploratory data analysis

Montserrat Guillen

2017

Contents

Introduction

This is a very short introduction to the exploration of data using RStudio and Rmarkdown.

This document sequentially applies a set of Data Science techniques to gain insights from the Direct Marketing campaign of a Portuguese Banking Institution.

First we need to read the data from the file "bank-additional-full.csv"

```
#setwd("..")
bank<-read.table(file="bank-additional-full.csv",header=T,sep=";")
```

The dataset contains information on `41188` clients and `21`variables.

*Input variables:

*# 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 - housing: has housing loan? (binary: “yes”,“no”)

7 - loan: has personal loan? (binary: “yes”,“no”)

*# related with the last contact of the current campaign:

8 - contact: contact communication type (categorical: “unknown”, “telephone”,“cellular”)

9 - month: last contact month of year (categorical: “jan”, “feb”, “mar”, . . . , “nov”, “dec”)

10 - day_of_week: last contact day of the month (numeric)

11 - duration: last contact duration, in seconds (numeric)

other attributes:

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - 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)

14 - previous: number of contacts performed before this campaign and for this client (numeric)

15 - poutcome: outcome of the previous marketing campaign (categorical: “unknown”,“other”,“failure”,“success”)

*# social and economic context attributes

16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)

17 - cons.price.idx: consumer price index - monthly indicator (numeric)

18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

19 - euribor3m: euribor 3 month rate - daily indicator (numeric)

20 - nr.employed: number of employees - quarterly indicator (numeric)

*Output variable (desired target):

21 - **y - has the client subscribed a term deposit?** (binary: “yes”=1,“no”=0)

Names of the variables

We print de names of the variables

```
colnames(bank)
```

```
##  [1] "age"            "job"            "marital"        "education"
##  [5] "default"        "housing"        "loan"           "contact"
##  [9] "month"          "day_of_week"    "duration"       "campaign"
## [13] "pdays"          "previous"       "poutcome"       "emp.var.rate"
## [17] "cons.price.idx" "cons.conf.idx"  "euribor3m"      "nr.employed"
## [21] "y"
```

We will use function **attach** so that we can call variable just by their name instead of bank$name

```
attach(bank)
```

An overview of the Data Analysis

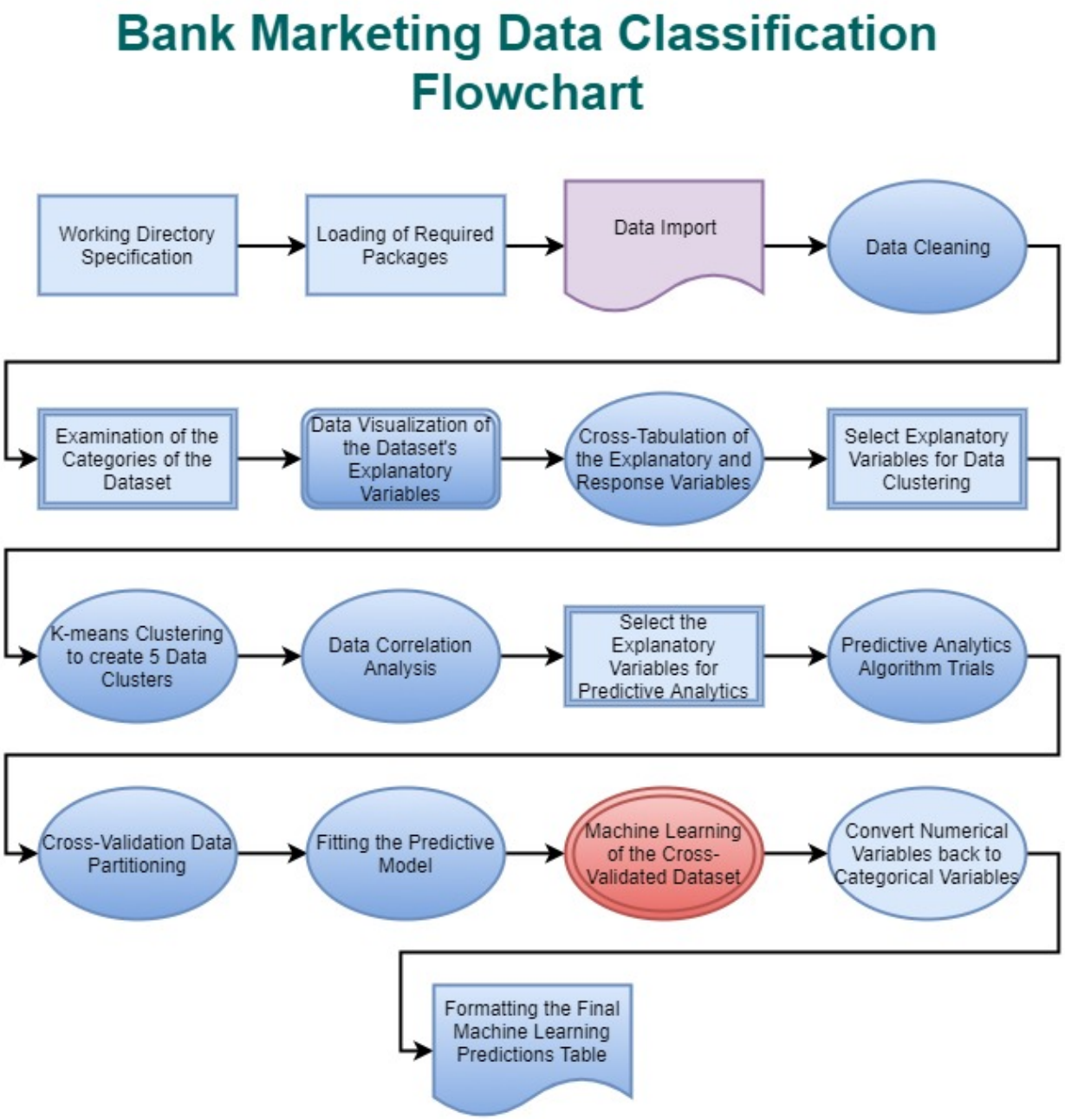


image:

Required packages

The function, "install.packages()", downloads and installs R programming language packages from CRAN-like repositories or from local files.

```
# Required Packages
# install.packages("ggplot2") # plotting
# install.packages("knitr") # report formatting
# install.packages("cluster") # kmeans clustering
```

```r
# install.packages("HSAUR") # silhouette plotting
# install.packages("fpc") # numbers cluster plot
# install.packages("lattice") # cluster plotting
# install.packages("rpart") # Decision Tress data classification
# install.packages("kernlab") # Support Vector Machines machine learning
# install.packages("randomForest") # Random Forest machine learning

library(ggplot2)
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
```

```r
#library(knitr)
#library(cluster)
#library(HSAUR)
#library(fpc)
#library(lattice)
#library(rpart)
#library(kernlab)
#library(randomForest)
```

Session information

This is information on the R version used in this example

```r
sessionInfo()
```

```
## R version 3.4.3 (2017-11-30)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 15063)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=Spanish_Spain.1252  LC_CTYPE=Spanish_Spain.1252
## [3] LC_MONETARY=Spanish_Spain.1252 LC_NUMERIC=C
## [5] LC_TIME=Spanish_Spain.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] dplyr_0.7.4   ggplot2_2.2.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.14     bindr_0.1        knitr_1.17       magrittr_1.5
```

```
##  [5] munsell_0.4.3    colorspace_1.3-2 R6_2.2.2         rlang_0.1.4
##  [9] stringr_1.2.0    plyr_1.8.4       tools_3.4.3      grid_3.4.3
## [13] gtable_0.2.0     htmltools_0.3.6  yaml_2.1.15      lazyeval_0.2.1
## [17] rprojroot_1.2    digest_0.6.12    assertthat_0.2.0 tibble_1.3.4
## [21] bindrcpp_0.2     glue_1.2.0       evaluate_0.10.1  rmarkdown_1.8
## [25] stringi_1.1.6    compiler_3.4.3   scales_0.5.0     backports_1.1.1
## [29] pkgconfig_2.0.1
```

Sample of records

```
head(bank)
```

```
##   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
```

Data visualization

Data Summary of the Bank Additional Dataset

```
summary(bank)
```

```
##       age                 job           marital
##  Min.   :17.00   admin.     :10422   divorced: 4612
##  1st Qu.:32.00   blue-collar: 9254   married :24928
##  Median :38.00   technician : 6743   single  :11568
##  Mean   :40.02   services   : 3969   unknown :   80
##  3rd Qu.:47.00   management : 2924
##  Max.   :98.00   retired    : 1720
##                  (Other)    : 6156
##                education        default         housing
##  university.degree  :12168   no     :32588   no     :18622
##  high.school        : 9515   unknown: 8597   unknown:  990
```

```
##  basic.9y           : 6045   yes    :    3   yes    :21576
##  professional.course: 5243
##  basic.4y           : 4176
##  basic.6y           : 2292
##  (Other)            : 1749
##       loan            contact          month       day_of_week
##  no     :33950   cellular :26144   may    :13769   fri:7827
##  unknown:  990   telephone:15044   jul    : 7174   mon:8514
##  yes    : 6248                     aug    : 6178   thu:8623
##                                    jun    : 5318   tue:8090
##                                    nov    : 4101   wed:8134
##                                    apr    : 2632
##                                    (Other): 2016
##     duration         campaign          pdays          previous
##  Min.   :   0.0   Min.   : 1.000   Min.   :  0.0   Min.   :0.000
##  1st Qu.: 102.0   1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.000
##  Median : 180.0   Median : 2.000   Median :999.0   Median :0.000
##  Mean   : 258.3   Mean   : 2.568   Mean   :962.5   Mean   :0.173
##  3rd Qu.: 319.0   3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.000
##  Max.   :4918.0   Max.   :56.000   Max.   :999.0   Max.   :7.000
##
##         poutcome      emp.var.rate      cons.price.idx  cons.conf.idx
##  failure    : 4252   Min.   :-3.40000   Min.   :92.20   Min.   :-50.8
##  nonexistent:35563   1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.7
##  success    : 1373   Median : 1.10000   Median :93.75   Median :-41.8
##                      Mean   : 0.08189   Mean   :93.58   Mean   :-40.5
##                      3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.4
##                      Max.   : 1.40000   Max.   :94.77   Max.   :-26.9
##
##    euribor3m      nr.employed     y
##  Min.   :0.634   Min.   :4964   no :36548
##  1st Qu.:1.344   1st Qu.:5099   yes: 4640
##  Median :4.857   Median :5191
##  Mean   :3.621   Mean   :5167
##  3rd Qu.:4.961   3rd Qu.:5228
##  Max.   :5.045   Max.   :5228
##
```

Specific statistical measures

```
sapply(bank[c("age", "duration")], median, 1)
```

```
##      age duration
##       38      180
```

Plots: Pie chart

```
table(y)
```

```
## y
##    no   yes
## 36548  4640
```

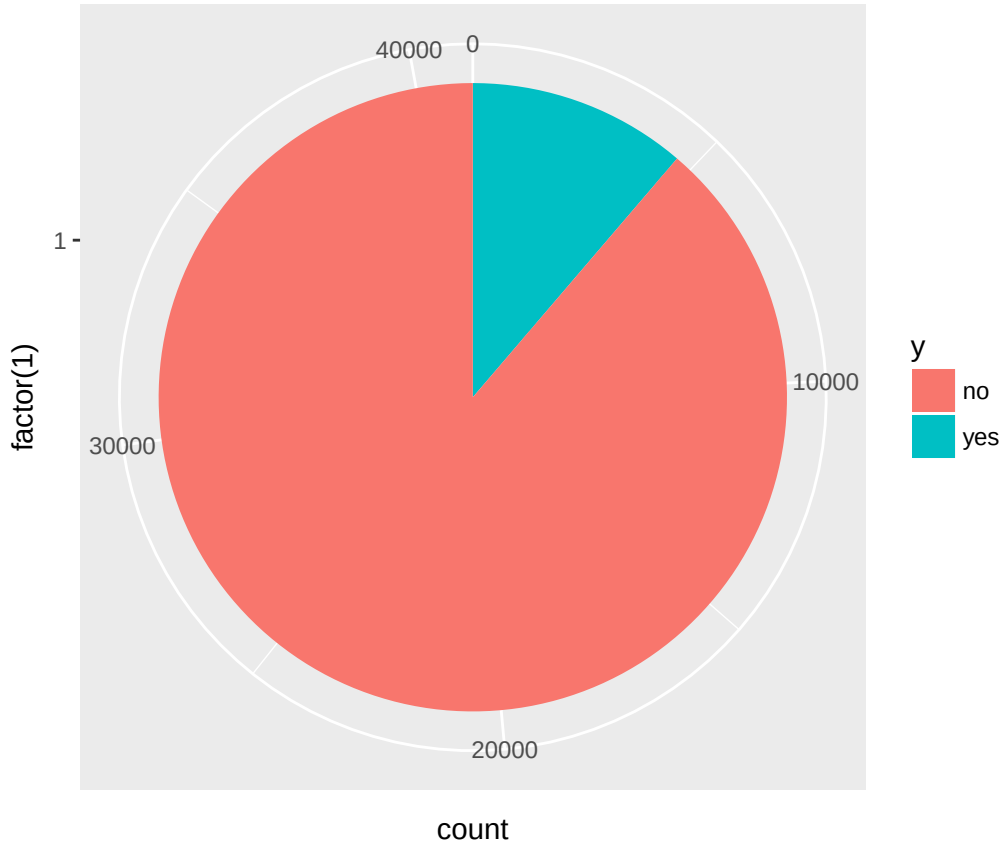
```r
prop.table(table(y))
```

```
## y
##        no       yes
## 0.8873458 0.1126542
```

```r
round(prop.table(table(y))*100, 2)
```

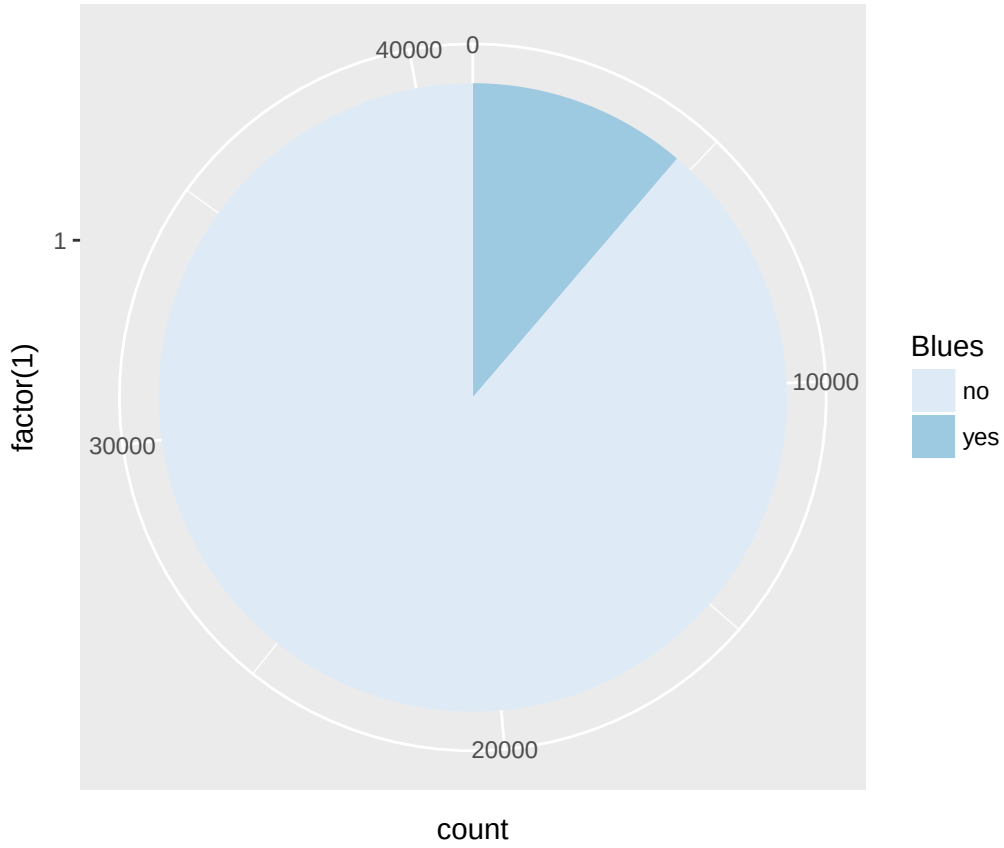
```
## y
##    no   yes
## 88.73 11.27
```

```r
ggplot(bank, aes(x=factor(1), fill=y))+
  geom_bar(width = 1)+
  coord_polar("y")
```



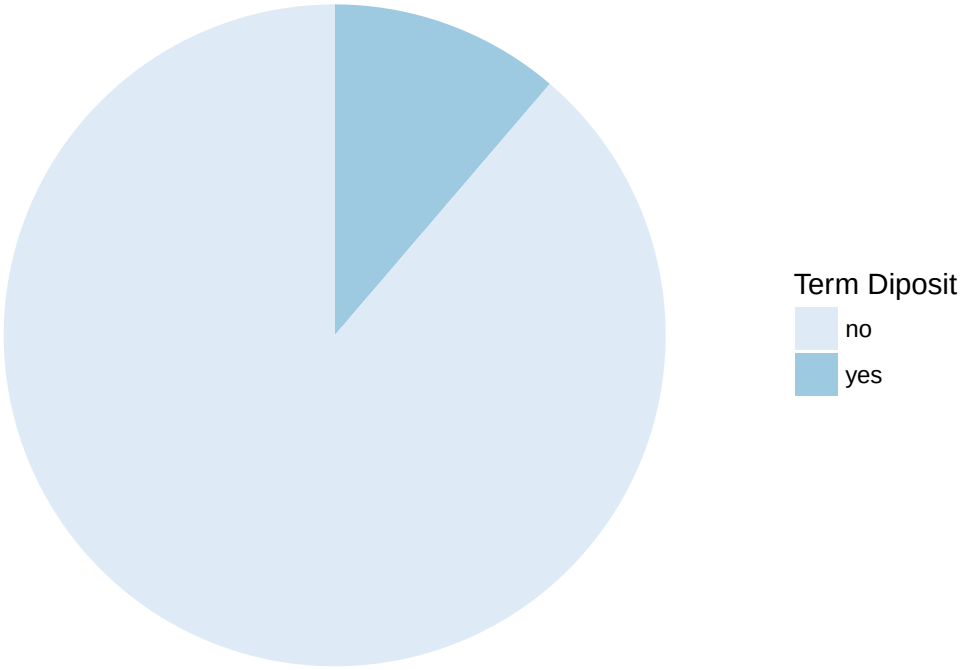
We now use another palette and labeling.

```r
ggplot(bank, aes(x=factor(1), fill=y))+
  geom_bar(width = 1)+
  coord_polar("y")+ scale_fill_brewer("Blues")
```



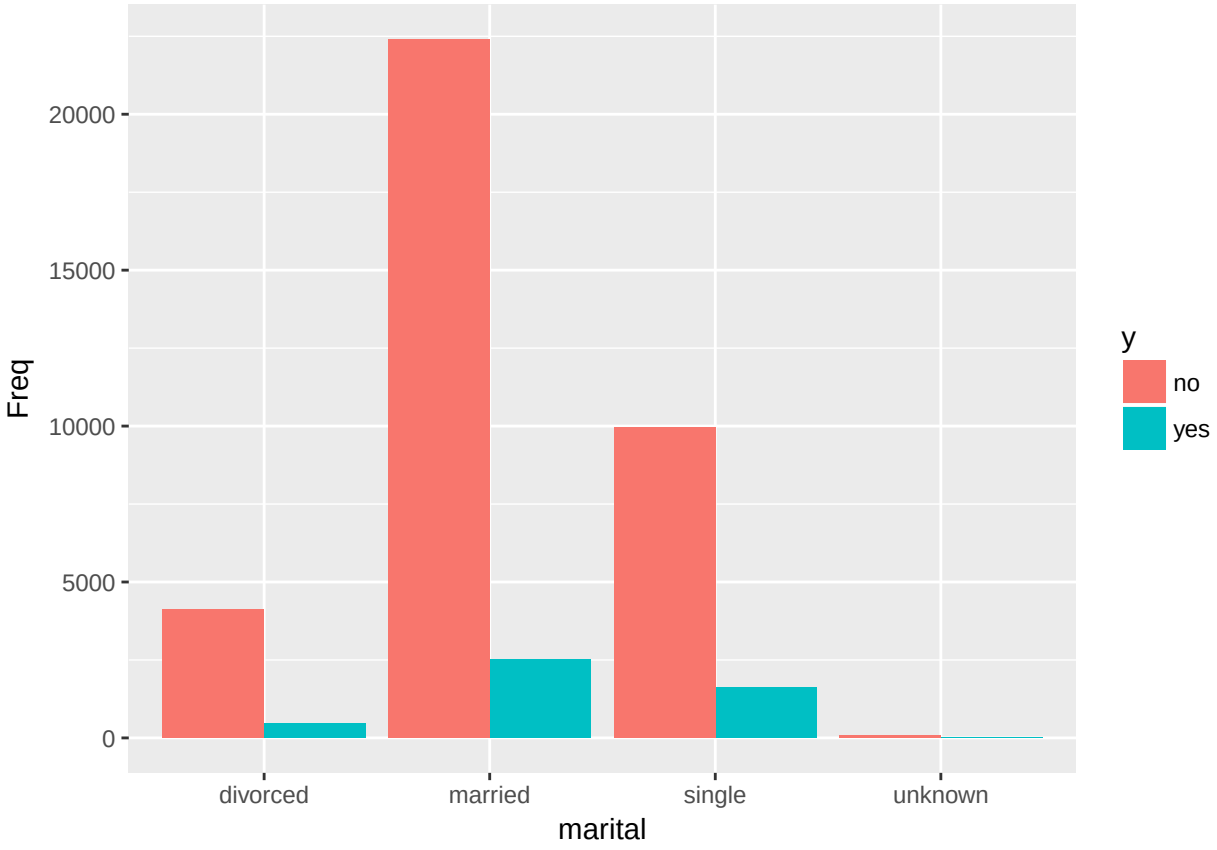

```
blank_theme <- theme_minimal()+
  theme(
  axis.title.x = element_blank(),
  axis.title.y = element_blank(),
  panel.border = element_blank(),
  panel.grid=element_blank(),
  axis.ticks = element_blank(),
  plot.title=element_text(size=14, face="bold")
  )

ggplot(bank, aes(x=factor(1), fill=y))+
  geom_bar(width = 1)+
  coord_polar("y")+ scale_fill_brewer("Term Diposit")+ blank_theme +
  theme(axis.text.x=element_blank())+
  theme(axis.text.y=element_blank())
```



Grouping bars: Frequency

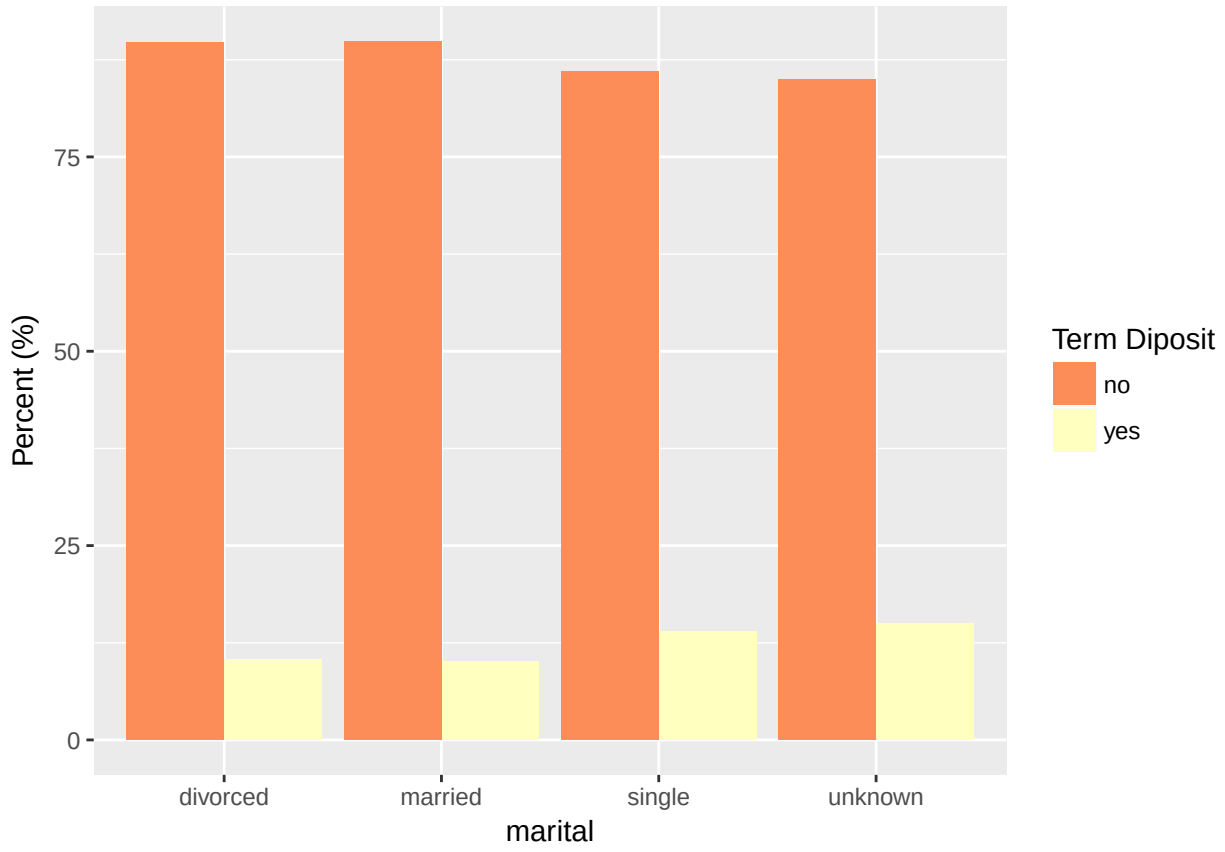
```
t=data.frame(table(y, marital))
ggplot(t, aes(x=marital, y=Freq, fill=y)) +
  geom_bar(position='dodge', stat='identity')
```



Grouping bars: Percent

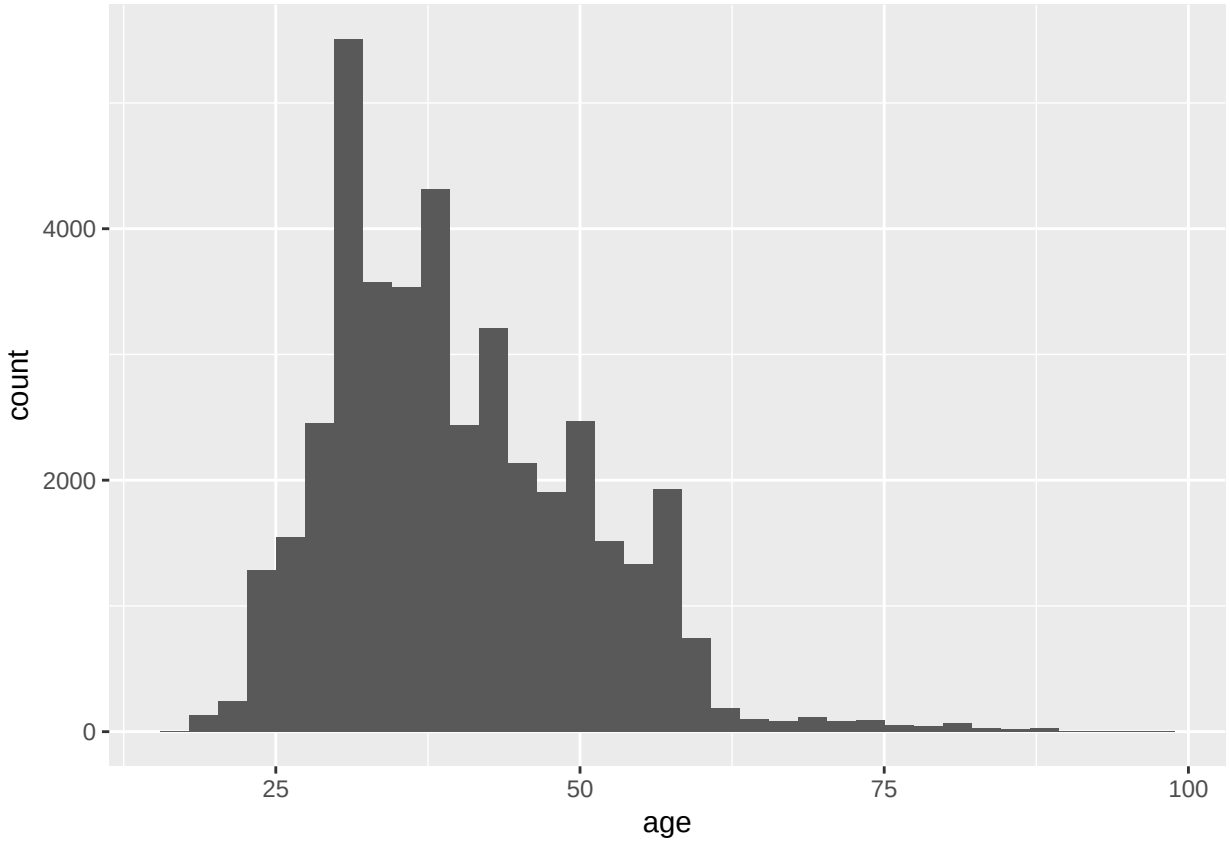
You can try a number of different palettes "Greens", "Set1", "Set2", etc. . .

```
t=data.frame(prop.table(table(y ,marital), 2))
ggplot(t, aes(x=marital, y=Freq*100, fill=y)) +
  geom_bar(position='dodge', stat='identity')+
  ylab("Percent (%)")+ scale_fill_brewer("Term Diposit", palette="Spectral")
```

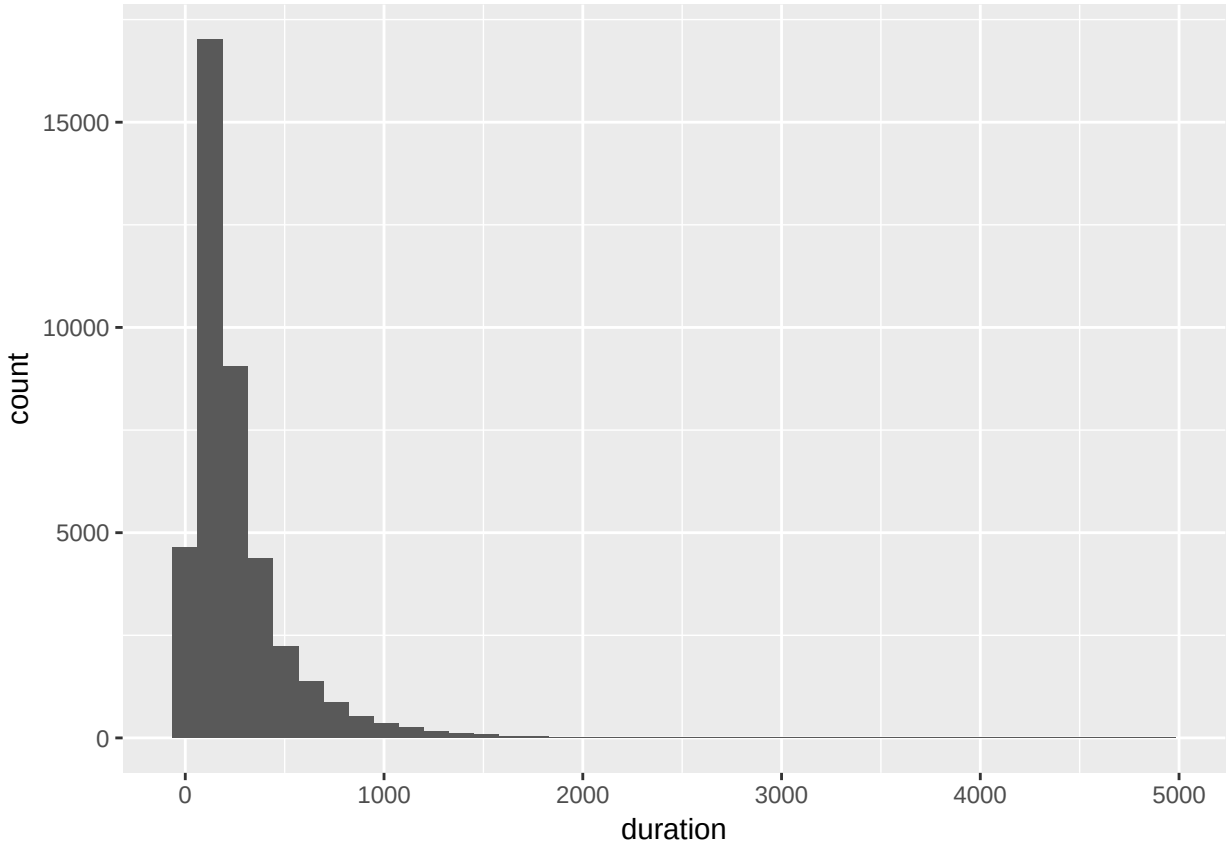


Histograms of age and duration

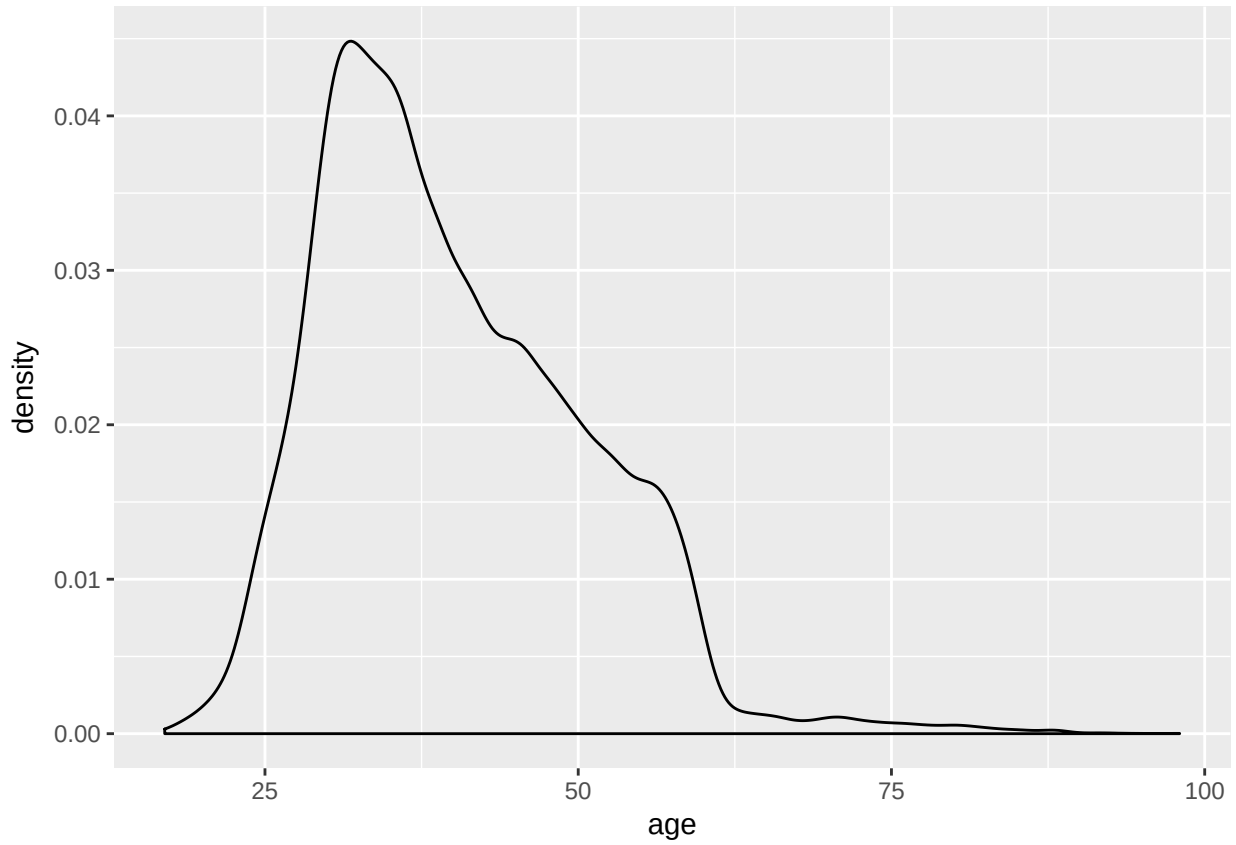
```
ggplot(data=bank, aes(age)) + geom_histogram(bins=35)
```



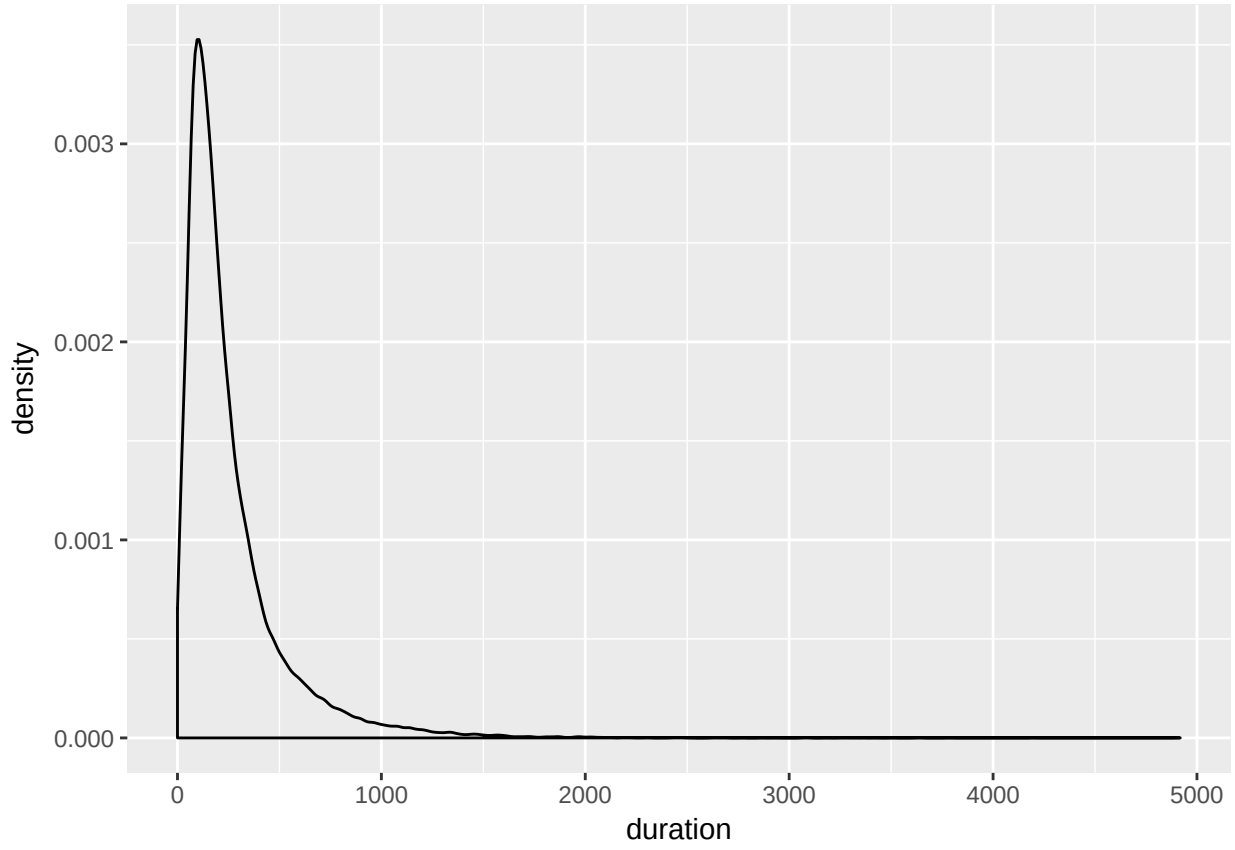

```
ggplot(data=bank, aes(duration)) + geom_histogram(bins=40)
```




```
ggplot(data=bank, aes(age)) + geom_density(alpha=1)
```




```
ggplot(data=bank, aes(duration)) + geom_density()
```



Plot of both

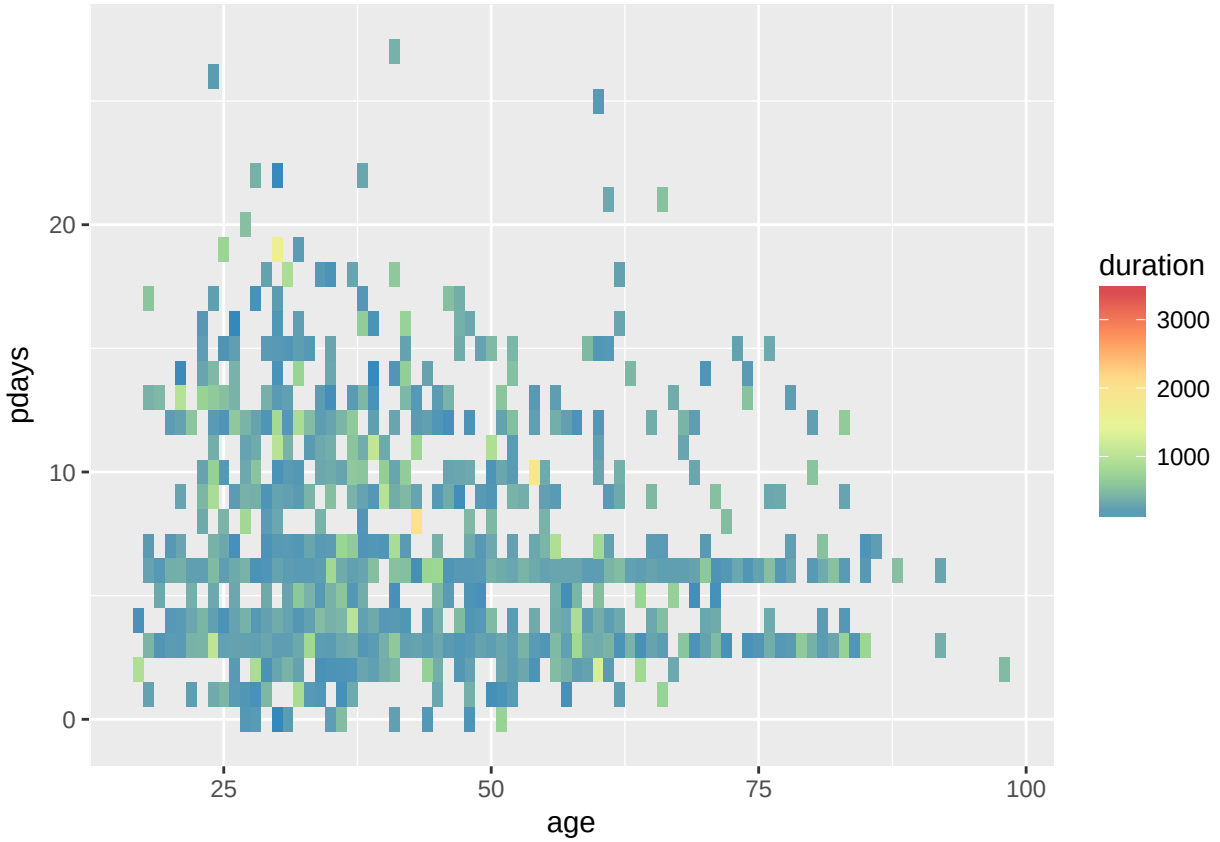
Palettes:

Diverging BrBG, PiYG, PRGn, PuOr, RdBu, RdGy, RdYlBu, RdYlGn, Spectral

Qualitative Accent, Dark2, Paired, Pastel1, Pastel2, Set1, Set2, Set3

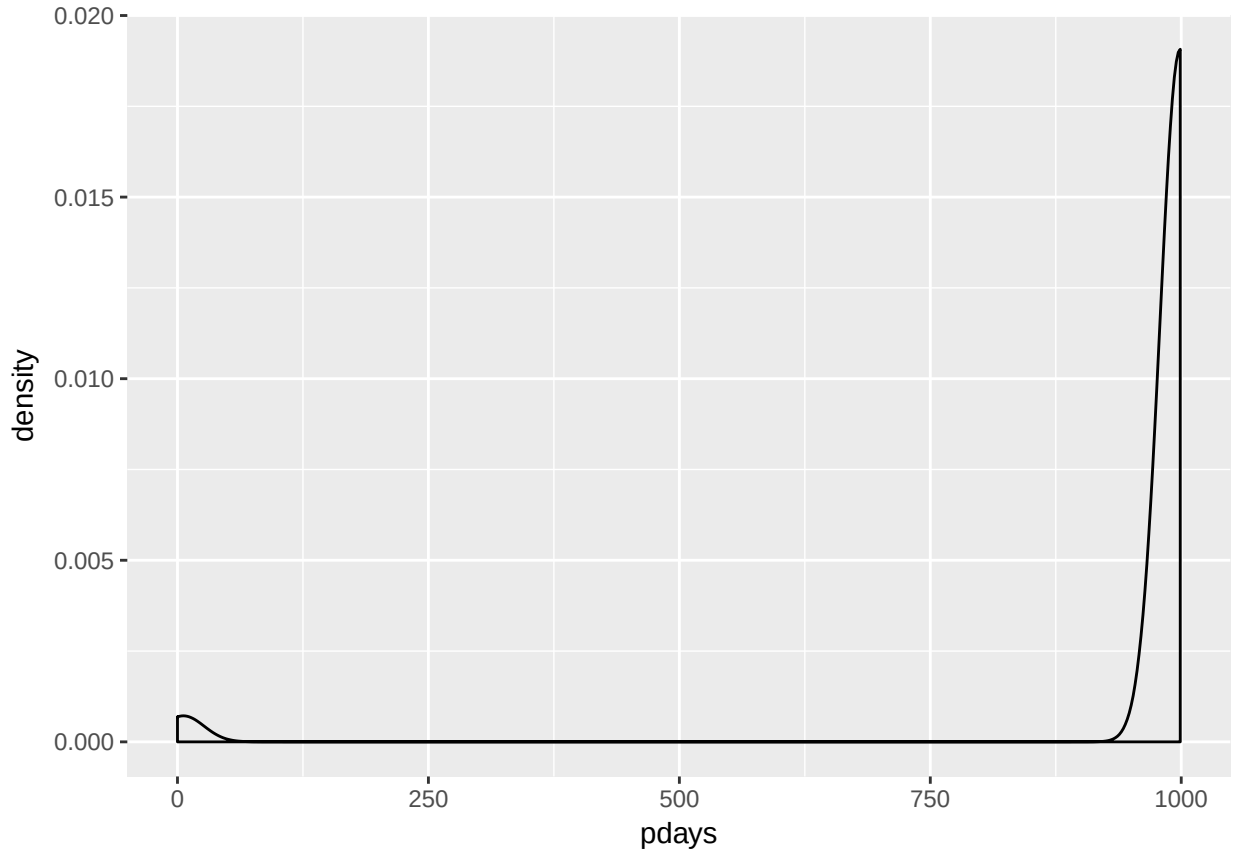
Sequential Blues, BuGn, BuPu, GnBu, Greens, Greys, Oranges, OrRd, PuBu, PuBuGn, PuRd, Purples, RdPu, Reds, YlGn, YlGnBu, YlOrBr, YlOrRd

```
 t2=subset(bank, pdays<999)
ggplot(t2) +
 geom_tile(aes(age, pdays,fill = duration))+
  scale_fill_distiller(palette = "Spectral")
```

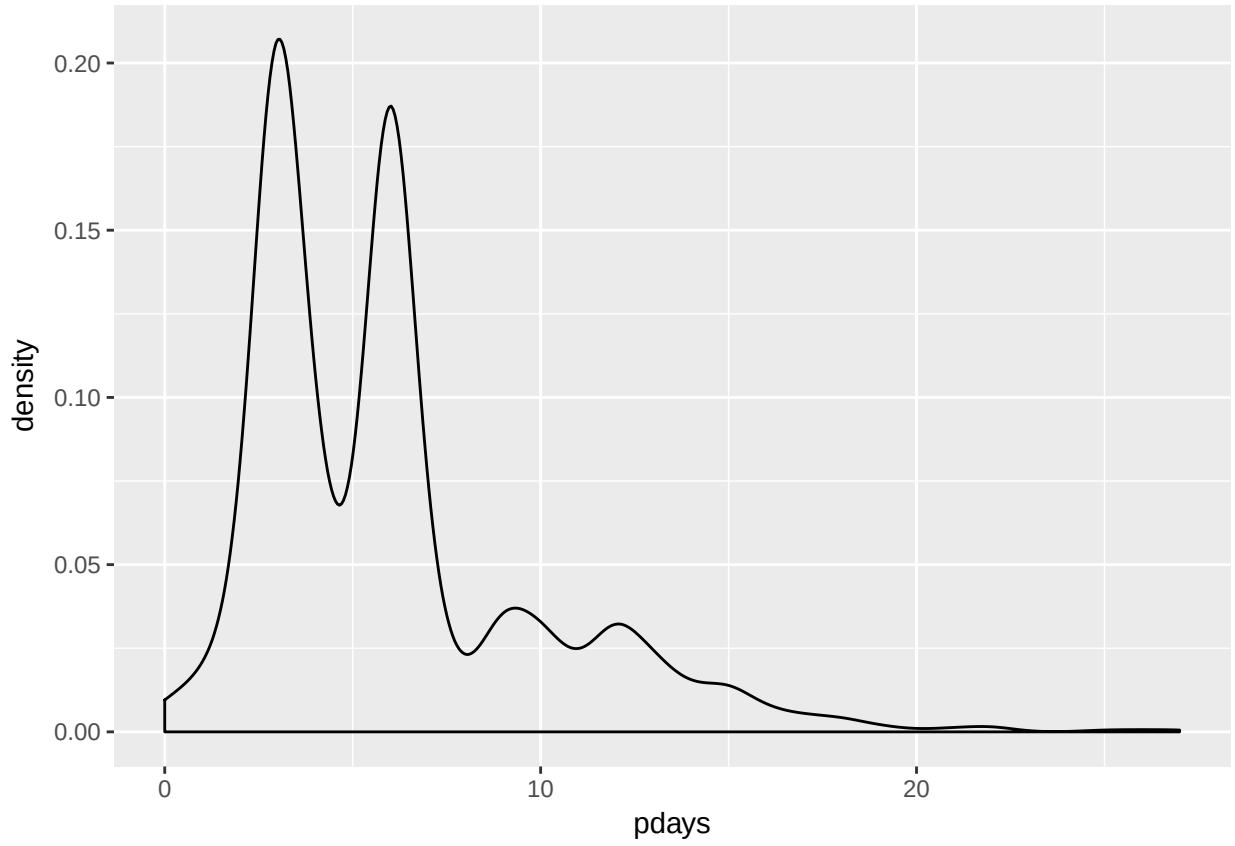


We have a problem with the variable **pdays**, because when the customer was not contacted, the value is 999.

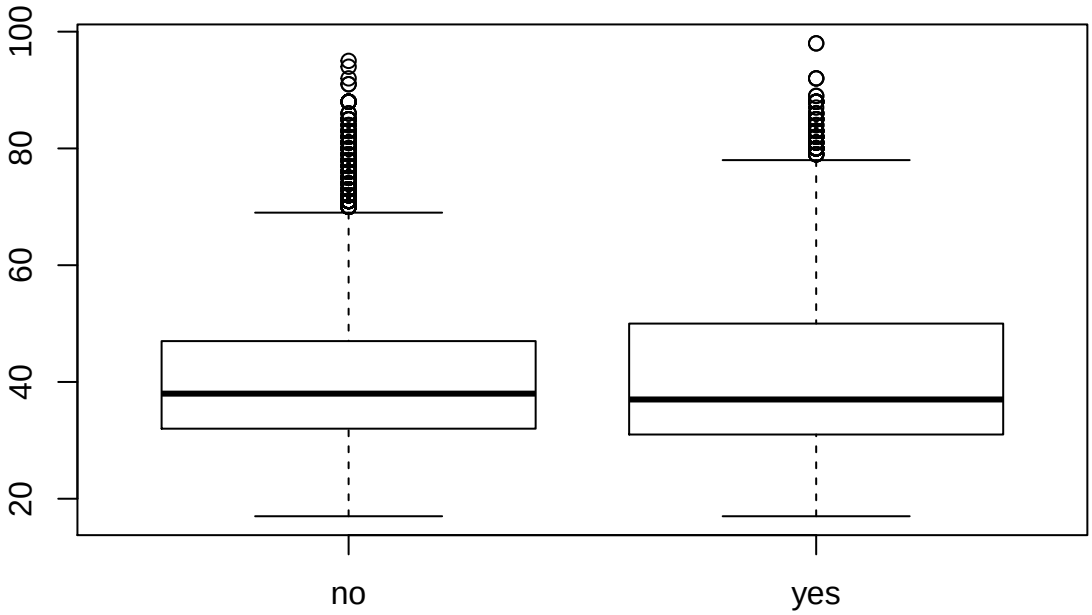
```
ggplot(data=bank, aes(pdays))+ geom_density()+scale_fill_brewer()
```



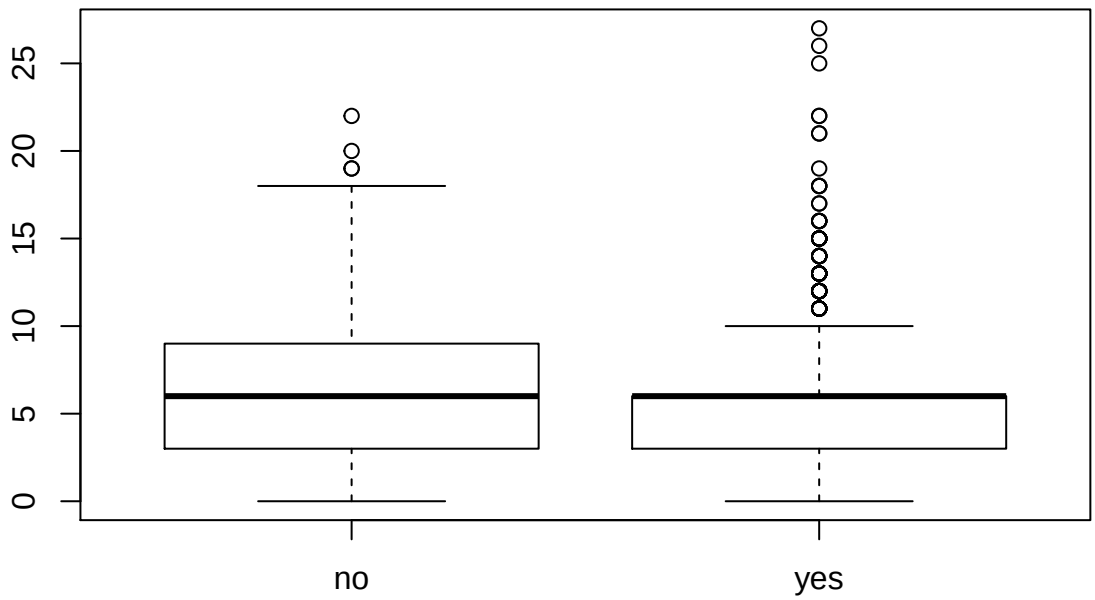

```
ggplot(data=t2, aes(pdays))+ geom_density()+scale_fill_brewer()
```




```
plot(y, age)
```




```
plot(t2$y, t2$pdays)
```



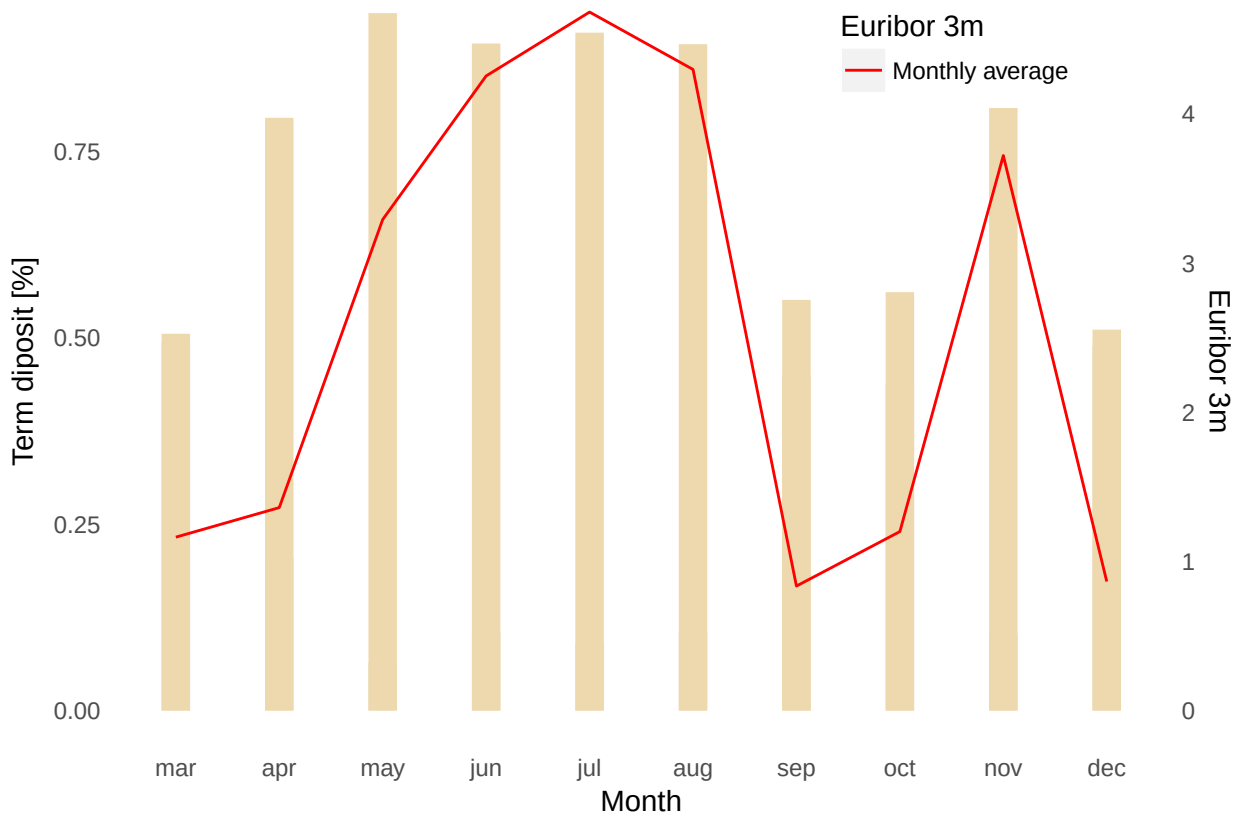
Evolution over time

```
t3=data.frame(prop.table(table(y, month),2))
t3$month_order=factor(as.character(t3$month), levels = c("mar", "apr", "may", "jun","jul", "aug", "sep"

t4 = group_by(bank, month) %>% summarise(Euribor=mean(euribor3m)) %>% ungroup()
t4$month_order=factor(as.character(t4$month), levels = c("mar", "apr", "may", "jun","jul", "aug", "sep"

ggplot(t3, aes(x=month_order, y=Freq)) +
  theme(plot.background = element_blank(),
              # panel.grid.minor = element_blank(),
              #  panel.grid.major = element_blank(),
                panel.border = element_blank(),
                panel.background = element_blank(),
                axis.ticks = element_blank(), # axis.title = element_blank()
        )+
    geom_linerange(t3, mapping=aes(x=month_order, ymin=0, ymax=Freq), colour = "wheat2", alpha=1, size=5
   geom_line(t4, mapping=aes(x=month_order, y=Euribor/5, group=1, colour= "Monthly average ")) +
scale_y_continuous(sec.axis = sec_axis(~.*5, name = "Euribor 3m")) +
scale_colour_manual(values = c("red")) +
 labs(y = "Term diposit [%]",
                x = " Month ",
                colour = "Euribor 3m") +
```

```
theme(legend.position = c(0.8, 0.9))
```



Reference

More information on graphics with R (ggplot2)

http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html