First delivery - ADEI

Alex Rubio i Josep Bernad

6 de marÃ§ de 2019

Contents

1 Presentation

1.1 R Markdowns document

This is an R Markdown document. We are showing some examples of EDA. 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. Use * to provide emphasis such as *italics* and **bold**.

Create lists: Unordered * and + or ordered 1. 2.

1. Item 1
2. Item 2
    - Item 2a
    - Item 2b

2 Bank client data

2.1 Description

Input variables:

1. age (numeric)
2. job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3. marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4. education (categorical:'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unkno
5. default: has credit in default? (categorical: 'no','yes','unknown')
6. housing: has housing loan? (categorical: 'no','yes','unknown')
7. loan: has personal loan? (categorical: 'no','yes','unknown')# related with the last contact of the current campaign:
8. contact: contact communication type (categorical:'cellular','telephone')
9. month: last contact month of year (categorical: 'jan', 'feb', 'mar',..., 'nov', 'dec')
10. day_of_week: last contact day of the week (categorical:'mon','tue','wed','thu','fri')
11. duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
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; 999 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: 'failure','nonexistent','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)
21. y - has the client subscribed a term deposit? (binary: 'yes','no')

3 Loading packages

4 Loading data

```
#rm(list=ls())
# Load Required Packages: to be increased over the course

#setwd("C:/Users/lmontero/Dropbox/DOCENCIA/FIB-ADEI/PRACTICA/BankMarketing")
#setwd("D:/DOCENCIA/FIB-ADEI/PRACTICA/BankMarketing")

# Josep
#setwd("/Users/SigmundFreud/Developer/r-studio/laboratory-adei/data-directory")
#file_path <- "Users/SigmundFreud/Developer/r-studio/laboratory-adei/data-directory"

# Alex
setwd("D:/Google Drive/Uni/ADEI/data-directory")
file_path <- "D:/Google Drive/Uni/ADEI/data-directory"

#load(paste0("D:/Google Drive/Uni/ADEI/data-directory","5000_samples.RData"))
load(path.expand("D:/Google Drive/Uni/ADEI/data-directory/5000_samples.RData"))
summary(df)
```

```
##       age                 job           marital
##  Min.   :17.00   admin.     :1288   divorced: 546
##  1st Qu.:32.00   blue-collar:1156   married :3029
##  Median :38.00   technician : 831   single  :1416
##  Mean   :39.97   services   : 471   unknown :   9
##  3rd Qu.:47.00   management : 345
##  Max.   :92.00   retired    : 187
##                  (Other)    : 722
##                education       default        housing          loan
##  university.degree  :1431   no     :3939   no     :2226   no     :4138
##  high.school        :1169   unknown:1061   unknown: 112   unknown: 112
##  basic.9y           : 758   yes    :   0   yes    :2662   yes    : 750
##  professional.course: 668
##  basic.4y           : 493
##  basic.6y           : 272
##  (Other)            : 209
##       contact          month      day_of_week    duration
##  cellular :3182   may    :1679   fri: 948    Min.   :   4.0
##  telephone:1818   jul    : 907   mon:1017    1st Qu.: 104.0
##                   aug    : 699   thu:1031    Median : 181.0
##                   jun    : 660   tue:1005    Mean   : 263.7
##                   nov    : 502   wed: 999    3rd Qu.: 328.0
##                   apr    : 323               Max.   :3078.0
##                   (Other): 230
##     campaign          pdays          previous             poutcome
##  Min.   : 1.000   Min.   :  0.0   Min.   :0.0000   failure    : 493
##  1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.0000   nonexistent:4315
##  Median : 2.000   Median :999.0   Median :0.0000   success    : 192
##  Mean   : 2.647   Mean   :957.9   Mean   :0.1772
##  3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.0000
##  Max.   :42.000   Max.   :999.0   Max.   :5.0000
##
##   emp.var.rate     cons.price.idx  cons.conf.idx      euribor3m
##  Min.   :-3.4000   Min.   :92.20   Min.   :-50.80   Min.   :0.634
##  1st Qu.:-1.8000   1st Qu.:93.08   1st Qu.:-42.70   1st Qu.:1.344
##  Median : 1.1000   Median :93.88   Median :-41.80   Median :4.857
##  Mean   : 0.1029   Mean   :93.58   Mean   :-40.59   Mean   :3.641
##  3rd Qu.: 1.4000   3rd Qu.:93.99   3rd Qu.:-36.40   3rd Qu.:4.961
##  Max.   : 1.4000   Max.   :94.77   Max.   :-26.90   Max.   :5.045
##
##   nr.employed     y
##  Min.   :4964   no :4416
##  1st Qu.:5099   yes: 584
##  Median :5191
##  Mean   :5168
##  3rd Qu.:5228
##  Max.   :5228
##
```

5 Univariate Descriptive Analysis

Creem factors per cada variable posant abans NA a aquells valos erronis o faltants.

5.1 Transform missing and wrong data to NA's

```
#Default
sel<-which(df$default=="unknown");length(sel)
```

```
## [1] 1061
```

```
df$default[sel] <- NA
df$default <- factor(df$default)
summary(df$default)
```

```
##   no NA's
## 3939 1061
```

```
#marital
sel<-which(df$marital=="unknown");length(sel)
```

```
## [1] 9
```

```
df$marital[sel] <- NA
df$marital <- factor(df$marital)
summary(df$marital)
```

```
## divorced  married   single     NA's
##      546     3029     1416        9
```

```
#Housing
sel<-which(df$housing=="unknown");length(sel)
```

```
## [1] 112
```

```
df$housing[sel] <- NA
df$housing <- factor(df$housing)
summary(df$housing)
```

```
##   no  yes NA's
## 2226 2662  112
```

```
#Loan
sel<-which(df$loan=="unknown");length(sel)
```

```
## [1] 112
```

```
df$loan[sel] <- NA
df$loan <- factor(df$loan)
summary(df$loan)
```

```
##   no  yes NA's
## 4138  750  112
```

```
#Job
sel<-which(df$job=="unknown");length(sel)
```

```
## [1] 43
```

```
df$job[sel] <- NA
df$job <- factor(df$job)
summary(df$job)
```

```
##        admin.   blue-collar  entrepreneur     housemaid    management
##          1288          1156           181           132           345
```

```
##       retired self-employed      services       student    technician
##           187           152           471           100           831
##    unemployed          NA's
##           114            43
```

```
#Education
sel<-which(df$education=="unknown");length(sel)
```

```
## [1] 207
```

```
df$education[sel] <- NA
df$education <- factor(df$education)
summary(df$education)
```

```
##            basic.4y            basic.6y            basic.9y
##                 493                 272                 758
##         high.school          illiterate professional.course
##                1169                   2                 668
##   university.degree                NA's
##                1431                 207
```

```
#Pdays
sel<-which(df$pdays==999);length(sel)
```

```
## [1] 4793
```

```
df$pdays[sel] <- NA
df$pdays <- factor(df$pdays)
summary(df$pdays)
```

```
##    0    1    2    3    4    5    6    7    8    9   10   11   12   13   15
##    1    5   12   62   17    5   48   13    5    9    7    2    4    8    3
##   16   17   18 NA's
##    1    4    1 4793
```

5.2 Create new factors corresponding to qualitative concepts.

5.2.1 Month

```
#Modify factor levels label
df$f.month <- factor(df$month, labels=paste("Month", sep="-", levels(df$month)))
table(df$f.month)
```

```
##
## Month-apr Month-aug Month-dec Month-jul Month-jun Month-mar Month-may
##       323       699        19       907       660        66      1679
## Month-nov Month-oct Month-sep
##       502        79        66
```

```
# Define new factor categories: 1-Spring | 2-Summer | 3-Resta
df$f.season <- 3

# 1 level - spring
sel<-which(df$f.month %in% c("Month-mar","Month-apr","Month-may"))
df$f.season[sel] <-1

# 2 level - Summer
sel<-which(df$f.month %in% c("Month-jun","Month-jul","Month-aug"))
```

```
df$f.season[sel] <-2

table(df$f.season);summary(df$f.season)
```

```
##
##    1    2    3
## 2068 2266  666
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    1.00    2.00    1.72    2.00    3.00
```

```
df$f.season<-factor(df$f.season,levels=1:3,labels=c("Spring","Summer","Aut-Win"))
summary(df$f.season)
```

```
##  Spring  Summer Aut-Win
##    2068    2266     666
```

5.2.2 Job

```
#Modify factor levels label
df$f.job <- factor(df$job, labels=paste("Job", sep="-", levels(df$job)))

table(df$f.job)
```

```
##
##        Job-admin.   Job-blue-collar  Job-entrepreneur     Job-housemaid
##              1288              1156               181               132
##    Job-management       Job-retired Job-self-employed      Job-services
##               345               187               152               471
##       Job-student    Job-technician    Job-unemployed
##               100               831               114
```

```
# Define new factor categories: 1-selfemployed | 2-worker | 3-other
df$f.jobsituation<-3

# 1 level - self-employed
sel<-which(df$f.job %in% c("Job-entrepreneur","Job-housemaid","Job-self-employed"))
df$f.jobsituation[sel] <- 1

# 2 level - worker
sel<-which(df$f.job %in% c("Job-admin","Job-blue-collar","Job-management","Job-services","Job-technician
df$f.jobsituation[sel] <- 2

table(df$f.jobsituation);summary(df$f.jobsituation)
```

```
##
##    1    2    3
##  465 2803 1732
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.253   3.000   3.000
```

```
df$f.jobsituation<-factor(df$f.jobsituation,levels=1:3,labels=c("Self-employed","Worker","Other"))
summary(df$f.jobsituation)
```

```
## Self-employed        Worker         Other
##           465          2803          1732
```

5.2.3 Pdays

```
table(df$pdays)
```

```
##
##  0  1  2  3  4  5  6  7  8  9 10 11 12 13 15 16 17 18
##  1  5 12 62 17  5 48 13  5  9  7  2  4  8  3  1  4  1
```

```
# Define new factor categories: 1-contacted | 2-not contacted
df$f.prev_contacted<-2

# 1 level - contacted
sel<-which(df$pdays %in% c(1:20))
df$f.prev_contacted[sel] <- 1

# 2 level - not contacted
sel<-which(df$pdays %in% c(21:1000))
df$f.prev_contacted[sel] <- 2

table(df$f.prev_contacted);summary(df$f.prev_contacted)
```

```
##
##    1    2
##  206 4794
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   1.959   2.000   2.000
```

```
df$f.prev_contacted<-factor(df$f.prev_contacted,levels=1:2,labels=c("Contacted","No-contacted"))
summary(df$pdays)
```

```
##    0    1    2    3    4    5    6    7    8    9   10   11   12   13   15
##    1    5   12   62   17    5   48   13    5    9    7    2    4    8    3
##   16   17   18 NA's
##    1    4    1 4793
```

5.2.4 Education

```
#Modify factor levels label
df$education <- factor(df$education, labels=paste("Edu", sep="-", levels(df$education)))

table(df$education)
```

```
##
##            Edu-basic.4y            Edu-basic.6y            Edu-basic.9y
##                     493                     272                     758
##         Edu-high.school         Edu-illiterate Edu-professional.course
##                    1169                       2                     668
##   Edu-university.degree
##                    1431
```

```
# Define new factor categories: 1-mandatory | 2-nonmandatory | 3-other
df$f.education<-3

# 1 level - mandatory
sel<-which(df$education %in% c("Edu-basic.4y","Edu-basic.6y", "Edu-basic.9y", "Edu-high.school"))
df$f.education[sel] <- 1
```

```
# 2 level - nonmandatory
sel<-which(df$education %in% c("Edu-professional.course","Edu-university.degree"))
df$f.education[sel] <- 2

table(df$f.education);summary(df$f.education)
```

```
##
##    1    2    3
## 2692 2099  209
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.503   2.000   3.000
```

```
df$f.education<-factor(df$f.education,levels=1:3,labels=c("Mandatory","Non-Mandatory","Other"))
summary(df$f.education)
```

```
##     Mandatory Non-Mandatory         Other
##          2692          2099           209
```

5.2.5 Extra Factorization

```
#Housing

df$f.housing<-factor(df$housing,labels=paste("f",sep=".",levels(df$housing)))
table(df$f.housing);summary(df$f.housing);
```

```
##
##  f.no f.yes
##  2226  2662
```

```
##  f.no f.yes  NA's
##  2226  2662   112
```

```
#Marital

df$f.marital<-factor(df$marital,labels=paste("f",sep=".",levels(df$marital)))
table(df$f.marital);summary(df$f.marital);
```

```
##
## f.divorced  f.married   f.single
##        546       3029       1416
```

```
## f.divorced  f.married   f.single       NA's
##        546       3029       1416          9
```

```
#Default
df$f.default<-factor(df$default, labels=paste("f",sep=".",levels(df$default)))
table(df$f.default);summary(df$f.default)
```

```
##
## f.no
## 3939
```

```
## f.no NA's
## 3939 1061
```

```
#Loan
df$f.loan<-factor(df$loan,labels=paste("f",sep=".",levels(df$loan)))
```

```r
table(df$f.loan);summary(df$f.loan)
```

```
##
##  f.no f.yes
##  4138   750
```

```
##  f.no f.yes  NA's
##  4138   750   112
```

```r
#Contact
df$f.contact<-factor(df$contact,labels=paste("f",sep=".",levels(df$contact)))
table(df$f.contact);summary(df$f.contact)
```

```
##
##  f.cellular f.telephone
##        3182        1818
```

```
##  f.cellular f.telephone
##        3182        1818
```

```r
#Day of Week
df$f.day<-factor(df$day_of_week,labels=paste("f.day",sep=".",levels(df$day)))
table(df$f.day);summary(df$f.day)
```

```
##
## f.day.fri f.day.mon f.day.thu f.day.tue f.day.wed
##       948      1017      1031      1005       999
```

```
## f.day.fri f.day.mon f.day.thu f.day.tue f.day.wed
##       948      1017      1031      1005       999
```

5.3 Create new factors corresponding to quantitative concepts.

5.3.1 Age discreatization

```r
summary(df$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   32.00   38.00   39.97   47.00   92.00
```

```r
qulist<-quantile(df$age,seq(0,1,0.25),na.rm=TRUE)

varaux<-factor(cut(df$age,breaks=qulist,include.lowest=T))
table(varaux)
```

```
## varaux
## [17,32] (32,38] (38,47] (47,92]
##    1353    1248    1202    1197
```

```r
tapply(df$age,varaux,median)
```

```
## [17,32] (32,38] (38,47] (47,92]
##      29      35      43      53
```

```r
varaux<-factor(cut(df$age,breaks=c(17,30,40,50,95),include.lowest=T))
table(varaux)
```

```
## varaux
## [17,30] (30,40] (40,50] (50,95]
##     887    2003    1252     858
```

```
tapply(df$age,varaux,median)
```

```
## [17,30] (30,40] (40,50] (50,95]
##      28      35      45      55
```

```
df$f.age<-factor(cut(df$age,breaks=c(17,30,40,50,95),include.lowest=T))

summary(df$f.age)
```

```
## [17,30] (30,40] (40,50] (50,95]
##     887    2003    1252     858
```

```
levels(df$f.age)<-paste0("f.age-",levels(df$f.age))
```

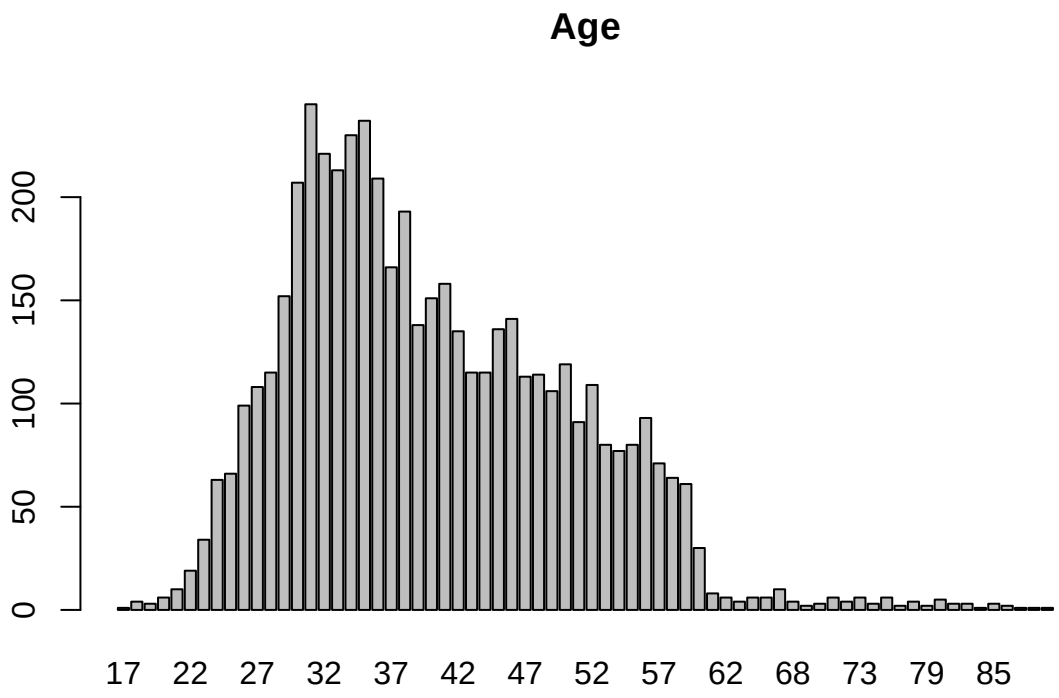
6 Exploratory Data Analysis

6.1 Age

```
summary(df$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   32.00   38.00   39.97   47.00   92.00
```

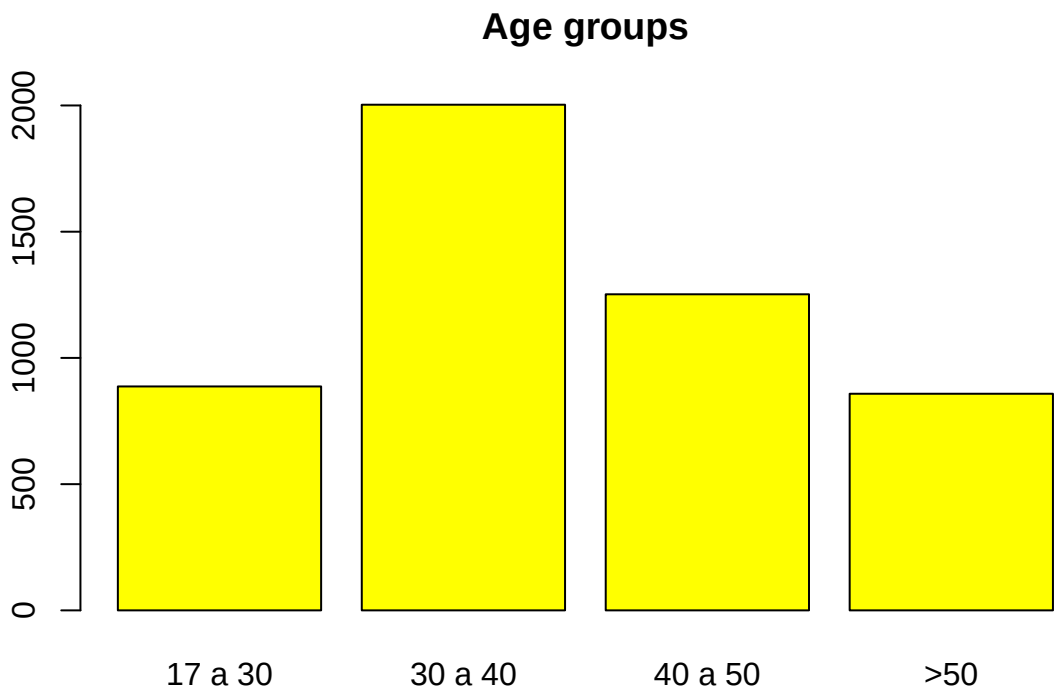
```
barplot(table(df$age), main= "Age")
```




```
summary(df$f.age)
```

```
## f.age-[17,30] f.age-(30,40] f.age-(40,50] f.age-(50,95]
##           887          2003          1252           858
```

```
barplot(table(df$f.age), main="Age groups",names.arg=c("17 a 30","30 a 40","40 a 50",">50"),col="yellow
```

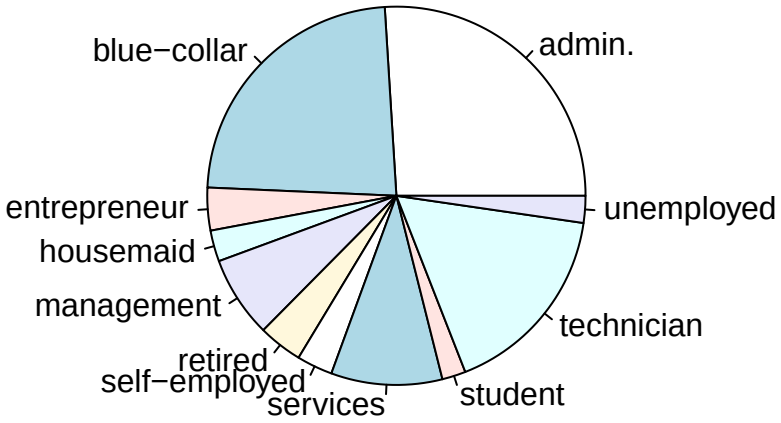


6.2 Job

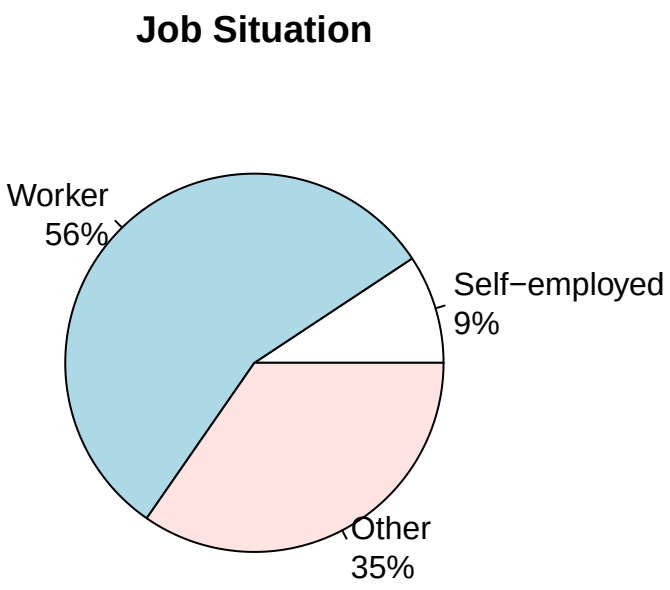
```
table(df$job)
```

```
##
##        admin.   blue-collar  entrepreneur     housemaid    management
##          1288          1156           181           132           345
##       retired self-employed      services       student    technician
##           187           152           471           100           831
##    unemployed
##           114
```

```
pie(table(df$job), main= "Job")
```




```
aux <- table(df$f.jobsituation)
pct <- round(aux/sum(aux)*100)
lbls <- paste(names(aux), "\n", pct, sep="")
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(aux,labels = lbls,
   main="Job Situation")
```

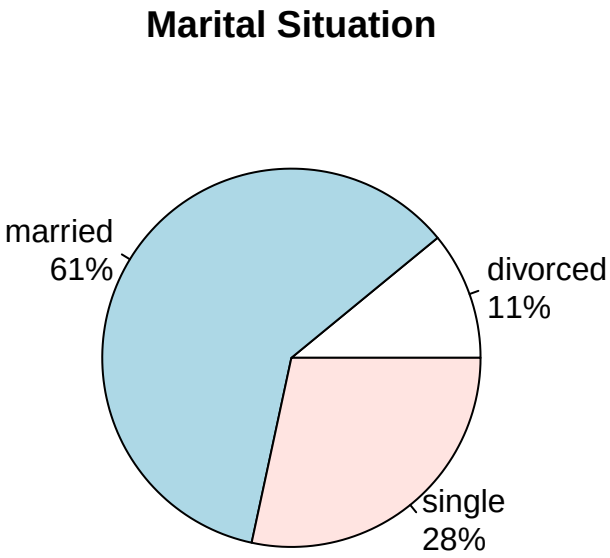


##Marital

```
table(df$marital)
```

```
##
## divorced  married   single
##      546     3029     1416
```

```
aux <- table(df$marital)
pct <- round(aux/sum(aux)*100)
lbls <- paste(names(aux), "\n", pct, sep="")
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(aux,labels = lbls,
   main="Marital Situation")
```

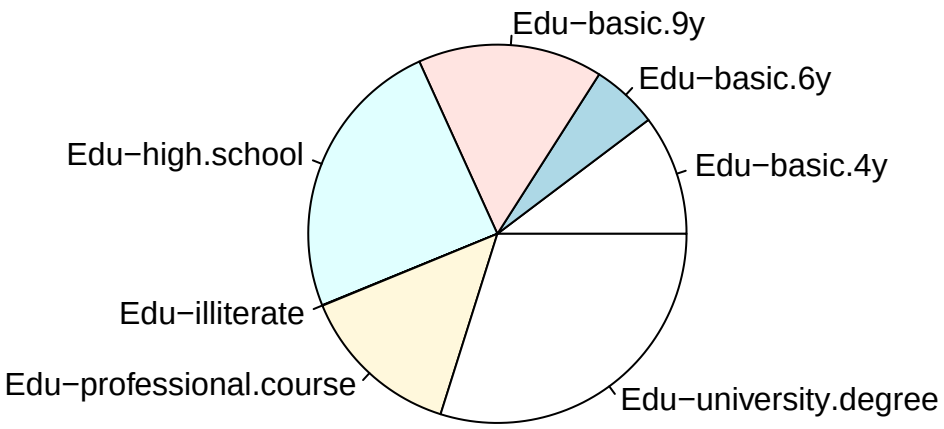


##Education

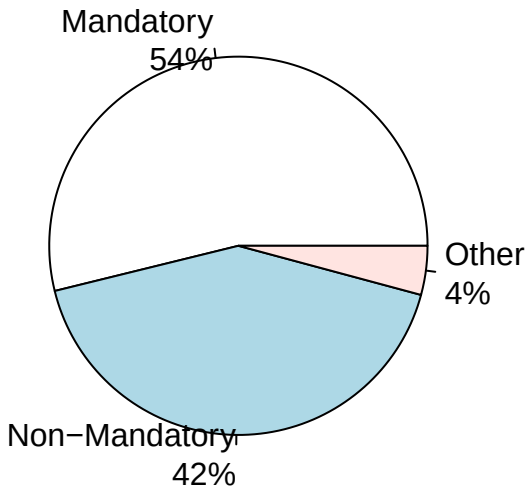
```
table(df$education)
```

```
##
##            Edu-basic.4y            Edu-basic.6y            Edu-basic.9y
##                     493                     272                     758
##         Edu-high.school         Edu-illiterate Edu-professional.course
##                    1169                       2                     668
##   Edu-university.degree
##                    1431
```

```
pie(table(df$education), main= "Education")
```




```
aux <- table(df$f.education)
pct <- round(aux/sum(aux)*100)
lbls <- paste(names(aux), "\n", pct, sep="")
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(aux,labels = lbls, main="Education Level")
```



##Default-Housing-Loan

```
table(df$default)
```

```
##
##   no
## 3939
```

```
table(df$housing)
```

```
##
##   no  yes
## 2226 2662
```

```
table(df$loan)
```

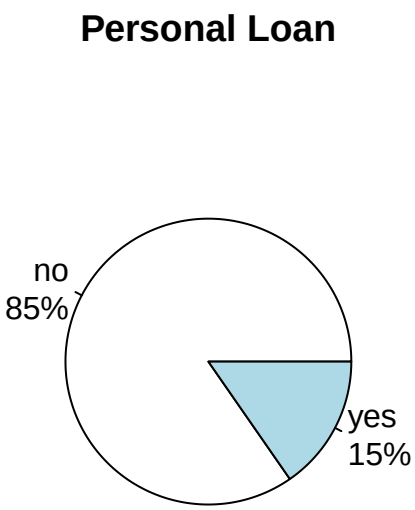
```
##
##   no  yes
## 4138  750
```

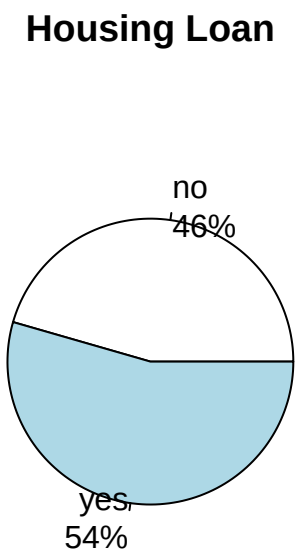
```
attach(mtcars)
```

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

```
par(mfrow=c(1,2))
aux <- table(df$loan)
pct <- round(aux/sum(aux)*100)
lbls <- paste(names(aux), "\n", pct, sep="")
```

```
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(aux,labels = lbls, main="Personal Loan")
aux <- table(df$housing)
pct <- round(aux/sum(aux)*100)
lbls <- paste(names(aux), "\n", pct, sep="")
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(aux,labels = lbls, main="Housing Loan")
```



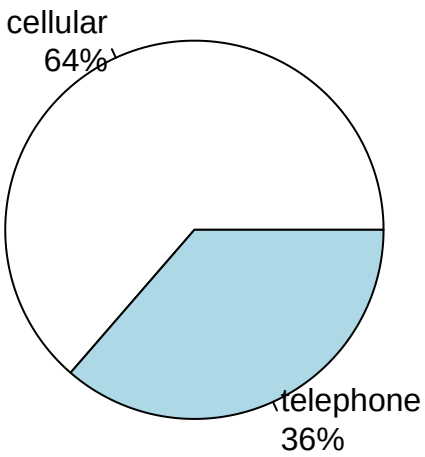


##Contact Device

```
table(df$contact)
```

```
##
##  cellular telephone
##      3182      1818
```

```
aux <- table(df$contact)
pct <- round(aux/sum(aux)*100)
lbls <- paste(names(aux), "\n", pct, sep="")
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(aux,labels = lbls, main="Contact Device")
```



##Date

```
table(df$month)
```

```
##
##  apr  aug  dec  jul  jun  mar  may  nov  oct  sep
##  323  699   19  907  660   66 1679  502   79   66
```

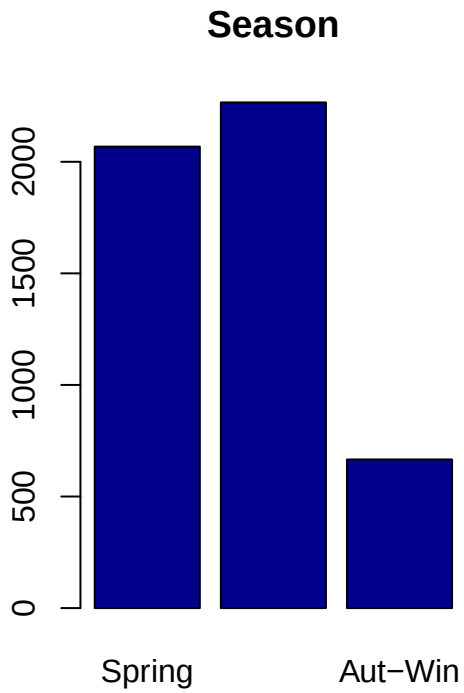
```
table(df$f.season)
```

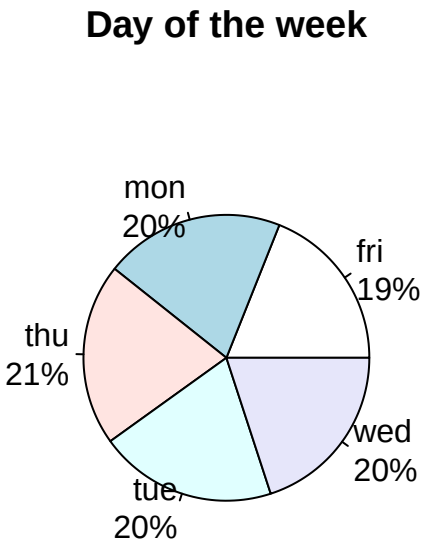
```
##
##  Spring  Summer Aut-Win
##    2068    2266     666
```

```
table(df$day_of_week)
```

```
##
##  fri  mon  thu  tue  wed
##  948 1017 1031 1005  999
```

```
par(mfrow=c(1,2))
barplot(table(df$f.season), main= "Season", col="darkblue")
aux <- table(df$day_of_week)
pct <- round(aux/sum(aux)*100)
lbls <- paste(names(aux), "\n", pct, sep="")
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(aux,labels = lbls,
   main="Day of the week")
```


Season
0
500
1000
1500
2000
Spring
Aut−Win


Day of the week
mon
20%
fri
19%
thu
21%
wed
20%
tue
20%