EDA is all about knowing about the data set, using graphical methods to summarise the main characteristics of the data set.

Sunkara Venkata Krishna

VS2539
N16740683

I found this dataset on kaggle, this data belongs to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls.The target variable is categorical having two values 'Yes' or 'No'.

The goal of this classification is to predict if the client is going to subscribe or not.

There are 22 independent variables and 1 dependent variable. below if the explination of each of them in more detail

1)age (numeric)

2)job : type of job is a categorical variable having ‘admin.’,‘blue-collar’, ‘entrepreneur’,‘housemaid’,‘management’,‘retired’,‘self- employed’,‘services’, ‘student’,‘technician’,‘unemployed’,‘unknown’

3)marital : marital status is a categorical variable having ‘divorced’,‘married’, ‘single’,‘unknown’; note: ‘divorced’ means divorced or widowed)

4)education : is a categorical variable having ‘basic.4y’,‘basic.6y’,‘basic.9y’, ‘high.school’,‘illiterate’,‘professional.course’,‘university.degree’,‘unknown’)

5)default: which means if the client has credit in default is a categorical variable having ‘no’,‘yes’,‘unknown’

6)housing: which means if the client has has housing loan is a categorical variable having‘no’,‘yes’,‘unknown’

7)loan:which means if the client has has personal loan is a categorical variable having ‘no’,‘yes’,‘unknown’

8)contact: contact communication type is a categorical variable having ‘cellular’, ‘telephone’

9)month: last contact month of year is a categorical variable having ‘jan’, ‘feb’, ‘mar’, ..., ‘nov’, ‘dec’

10)day_of_week: last contact day of the week is a categorical variable having ‘mon’, ‘tue’,‘wed’,‘thu’,‘fri’

11)duration: last contact duration, in seconds is a numeric varibale.

12)campaign: number of contacts performed during this campaign and for this client is a numeric variable

13)pdays: which means the number of days that passed by after the client was last contacted from a previous campaign and is a numericvariable 999 means client was not previously contacted

14)previous: which means the number of contacts performed before this campaign and for this client and is a numeric variable

15)poutcome: which means the outcome of the previous marketing campaign, it is a categorical variable having 'failure’,‘nonexistent’,‘success’

16)emp.var.rate: means the employment variation rate - quarterly indicator and is numeric.

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

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

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

20)nr.employed: number of employees - quarterly indicatorl, is numeic

21)y : has the client subscribed a term deposit is a binary catergorical variable having values as : ‘yes’,‘no’

In [101]:
```
bankdata<-read.csv('bank-additional.csv', head = T, stringsAsFactor
bankdata
```

age	job	marital	education	default	housing	loan	contact	mor
56	housemaid	married	basic.4y	no	no	no	telephone	m
57	services	married	high.school	unknown	no	no	telephone	m
37	services	married	high.school	no	yes	no	telephone	m
40	admin.	married	basic.6y	no	no	no	telephone	m
56	services	married	high.school	no	no	yes	telephone	m

45	services	married	basic.9y	unknown	no	no	telephone	m
59	admin.	married	professional.course	no	no	no	telephone	m
41	blue-collar	married	unknown	unknown	no	no	telephone	m
24	technician	single	professional.course	no	yes	no	telephone	m
25	services	single	high.school	no	yes	no	telephone	m
41	blue-collar	married	unknown	unknown	no	no	telephone	m
25	services	single	high.school	no	yes	no	telephone	m
29	blue-collar	single	high.school	no	no	yes	telephone	m
57	housemaid	divorced	basic.4y	no	yes	no	telephone	m
35	blue-collar	married	basic.6y	no	yes	no	telephone	m
54	retired	married	basic.9y	unknown	yes	yes	telephone	m
35	blue-collar	married	basic.6y	no	yes	no	telephone	m
46	blue-collar	married	basic.6y	unknown	yes	yes	telephone	m
50	blue-collar	married	basic.9y	no	yes	yes	telephone	m
39	management	single	basic.9y	unknown	no	no	telephone	m
30	unemployed	married	high.school	no	no	no	telephone	m
55	blue-collar	married	basic.4y	unknown	yes	no	telephone	m
55	retired	single	high.school	no	yes	no	telephone	m
41	technician	single	high.school	no	yes	no	telephone	m
37	admin.	married	high.school	no	yes	no	telephone	m
35	technician	married	university.degree	no	no	yes	telephone	m
59	technician	married	unknown	no	yes	no	telephone	m
39	self-employed	married	basic.9y	unknown	no	no	telephone	m
54	technician	single	university.degree	unknown	no	no	telephone	m
55	unknown	married	university.degree	unknown	unknown	unknown	telephone	m
...	...	...	...	...	...	...	...	
35	technician	divorced	basic.4y	no	no	no	cellular	r
35	technician	divorced	basic.4y	no	yes	no	cellular	r
33	admin.	married	university.degree	no	no	no	cellular	r
33	admin.	married	university.degree	no	yes	no	cellular	r
60	blue-collar	married	basic.4y	no	yes	no	cellular	r
35	technician	divorced	basic.4y	no	yes	no	cellular	r
54	admin.	married	professional.course	no	no	no	cellular	r
38	housemaid	divorced	university.degree	no	no	no	cellular	r
32	admin.	married	university.degree	no	no	no	telephone	r
32	admin.	married	university.degree	no	yes	no	cellular	r
38	entrepreneur	married	university.degree	no	no	no	cellular	r

62	services	married	high.school	no	yes	no	cellular	r
40	management	divorced	university.degree	no	yes	no	cellular	r
33	student	married	professional.course	no	yes	no	telephone	r
31	admin.	single	university.degree	no	yes	no	cellular	r
62	retired	married	university.degree	no	yes	no	cellular	r
62	retired	married	university.degree	no	yes	no	cellular	r
34	student	single	unknown	no	yes	no	cellular	r
38	housemaid	divorced	high.school	no	yes	yes	cellular	r
57	retired	married	professional.course	no	yes	no	cellular	r
62	retired	married	university.degree	no	no	no	cellular	r
64	retired	divorced	professional.course	no	yes	no	cellular	r
36	admin.	married	university.degree	no	no	no	cellular	r
37	admin.	married	university.degree	no	yes	no	cellular	r
29	unemployed	single	basic.4y	no	yes	no	cellular	r
73	retired	married	professional.course	no	yes	no	cellular	r
46	blue-collar	married	professional.course	no	no	no	cellular	r
56	retired	married	university.degree	no	yes	no	cellular	r
44	technician	married	professional.course	no	no	no	cellular	r
74	retired	married	professional.course	no	yes	no	cellular	r

In [102]:
```
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 ...
```

In [42]:
```
library(dplyr)
str(bankdata)
```

```
Attaching package: ‘dplyr’

The following objects are masked from ‘package:data.table’:

    between, first, last

The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union


'data.frame':   41188 obs. of  21 variables:
 $ age           : int  56 57 37 40 56 45 59 41 24 25 ...
 $ job           : chr  "housemaid" "services" "services" "admin."
...
 $ marital       : chr  "married" "married" "married" "married" ..
.
 $ education     : chr  "basic.4y" "high.school" "high.school" "ba
sic.6y" ...
 $ default       : chr  "no" "unknown" "no" "no" ...
 $ housing       : chr  "no" "no" "yes" "no" ...
 $ loan          : chr  "no" "no" "no" "no" ...
 $ contact       : chr  "telephone" "telephone" "telephone" "telep
hone" ...
 $ month         : chr  "may" "may" "may" "may" ...
 $ day_of_week   : chr  "mon" "mon" "mon" "mon" ...
 $ 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      : chr  "nonexistent" "nonexistent" "nonexistent"
"nonexistent" ...
 $ 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             : chr  "no" "no" "no" "no" ...
```

In [43]:
```
colSums(is.na(bankdata))
```

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

While looking at the data I found that the data contains some unknown values.

In [44]:
```
colSums(bankdata=="unknown")
```

age	0
job	330
marital	80
education	1731
default	8597
housing	990
loan	990
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

In [45]:
```
colSums(bankdata=='')
```

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

In [46]:
```
sum(colSums(bankdata=="unknown"))
```

12718

Feature Analysis

The dataset Is having 20 features out of which 12 features are categorical variables these are job, martial, education, default, housing, loan, contact, month, day_of_week, poutcome. And the remaining 8 are numerical features like age, duration, campaign , days, previous, emp_var_rate, cons_price_idx, cons_conf_idx, euribor3m and nr_emplo yed.

The feature age is having minimum at 17 and the average values at 40 and the maximum age is 98 year. The median of the age data is about 38 years, by which we can conclude that the age is centred around middle aged people. We can also see that the first quartile and the third quartile are around 32 years and 47 years respectively which also confirms our data is more centric toward the middle half of the data.

The job feature explains about the type of job people are working in and is distributed in 7 columns with the following distribution 10422 as admins , 9254 as blue-collar, 6743 as technician, 3969 as Services,2924 as management,1720 as retired and 6156 as other. ####Similarly with the martial feature It is divided into 4 classes with 4612 as divorced, 24928 as married , 11568 as single and 80 as unknown out of the given classes the married class is having maximum number of rows.

The education feature is divided into 7 classes out of these seven classes the maximum customers are having a university degree and next highest number of customers are from the high school.

The feature default is to know if the customer has a credit in default or not.This feature is classified into 3 classes out of which the “no” category dominates the most of the data having 32588 customers.

The housing feature is evenly distributed between yes and no classes and the third class “unknown” Is the least with only 990 customers data. The feature loan is divided into three classes no, yes and unknown. Out of these three classes the “no” is having the highest number of customers data with 33950 rows.

The feature contact is divided into two classes “cellular” and “telephone”. The feature month is divided into 12 classes out of which January is having the highest number of rows. In the similar senses the feature day_of_week is also evenly distributed across all the seven classes.

The feature duration is numerical variable with the mean of duration is around 258.3 seconds . The median of this data is 180.0 seconds and the maximum values is 4918.0 seconds. Based on the mean and median values the data is having outliers because the 1st quartile and 3rd quartile and mean the data is most centric towards the values around 180.0 seconds.

The feature campaign is a numerical variable with minimum of 1 and the maximum of 56 . Based on the mean and median, this data is centred around 2.

The feature pdays is numerical but can be treated as categorical because this feature has only two values 0 and 999. This can be treated as categorical data.

The feature previous is numerical but this can also be treated as categorical because this has numerical values between 0 to 7. ####The feature poutcome is divided into 3 classes out of which nonexistent class is dominating.

The features emp_var_rate, cons_price_idx, cons_conf_idx and euribor3m are numerical variables.

The feature nr_employed is a numerical data with mean value of 5167 and median value of 519 .

Final target variable y is categorical variable with 2 classes yes and no. the target variable is dominated with 36548 customers data of No.

In [47]:
```
summary(bankdata)
```

```
      age            job              marital           education
 Min.   :17.00   Length:41188       Length:41188       Length:4118
8
 1st Qu.:32.00   Class :character   Class :character   Class :char
acter
 Median :38.00   Mode  :character   Mode  :character   Mode  :char
acter
 Mean   :40.02
 3rd Qu.:47.00
 Max.   :98.00
```

```
   default            housing              loan              contac
t
 Length:41188       Length:41188       Length:41188       Length:4
1188
 Class :character   Class :character   Class :character   Class :c
haracter
 Mode  :character   Mode  :character   Mode  :character   Mode  :c
haracter



    month           day_of_week           duration         campaig
n
 Length:41188       Length:41188       Min.   :   0.0   Min.   : 1
.000
 Class :character   Class :character   1st Qu.: 102.0   1st Qu.: 1
.000
 Mode  :character   Mode  :character   Median : 180.0   Median : 2
.000
                                       Mean   : 258.3   Mean   : 2
.568
                                       3rd Qu.: 319.0   3rd Qu.: 3
.000
                                       Max.   :4918.0   Max.   :56
.000
     pdays           previous       poutcome           emp.var.rate
 Min.   :  0.0   Min.   :0.000   Length:41188       Min.   :-3.400
00
 1st Qu.:999.0   1st Qu.:0.000   Class :character   1st Qu.:-1.800
00
 Median :999.0   Median :0.000   Mode  :character   Median : 1.100
00
 Mean   :962.5   Mean   :0.173                      Mean   : 0.081
89
 3rd Qu.:999.0   3rd Qu.:0.000                      3rd Qu.: 1.400
00
 Max.   :999.0   Max.   :7.000                      Max.   : 1.400
00
 cons.price.idx  cons.conf.idx     euribor3m      nr.employed
 Min.   :92.20   Min.   :-50.8   Min.   :0.634   Min.   :4964
 1st Qu.:93.08   1st Qu.:-42.7   1st Qu.:1.344   1st Qu.:5099
 Median :93.75   Median :-41.8   Median :4.857   Median :5191
 Mean   :93.58   Mean   :-40.5   Mean   :3.621   Mean   :5167
 3rd Qu.:93.99   3rd Qu.:-36.4   3rd Qu.:4.961   3rd Qu.:5228
 Max.   :94.77   Max.   :-26.9   Max.   :5.045   Max.   :5228
      y
 Length:41188
 Class :character
 Mode  :character
```

Categorical analysis:

As we have seen we have 12 independent variable which are categorical, since they are categorical we are calculating the prop.table, which will give us the value the table entity as a fraction of the while table. I have perfomed the same for all the 12 independent variables.

In [48]:
```
sort(round(prop.table(table(bankdata$job))*100, 2), decreasing = T
```

```
       admin.   blue-collar    technician      services    managem
ent
        25.30         22.47         16.37          9.64          7
.10
      retired  entrepreneur self-employed     housemaid    unemplo
yed
         4.18          3.54          3.45          2.57          2
.46
      student       unknown
         2.12          0.80
```

In [49]:
```
sort(round(prop.table(table(bankdata$education))*100,2),decreasing
```

```
  university.degree         high.school            basic.9y profes
sional.course
              29.54               23.10               14.68
12.73
           basic.4y            basic.6y             unknown
illiterate
              10.14                5.56                4.20
0.04
```

In [50]:
```
sort(round(prop.table(table(bankdata$marital))*100,2),decreasing =
```

```
 married   single divorced  unknown
   60.52    28.09    11.20     0.19
```

In [51]:
```
sort(round(prop.table(table(bankdata$default))*100,2),decreasing =
```

```
     no unknown     yes
  79.12   20.87    0.01
```

In [52]:
```
sort(round(prop.table(table(bankdata$housing))*100,2),decreasing =
```

```
  yes    no unknown
52.38 45.21    2.40
```

In [53]:
```
sort(round(prop.table(table(bankdata$loan))*100,2),decreasing = T)
```

```
   no   yes unknown
82.43 15.17    2.40
```

In [54]:
```
sort(round(prop.table(table(bankdata$contact))*100,2),decreasing =
```

```
cellular telephone
   63.47     36.53
```

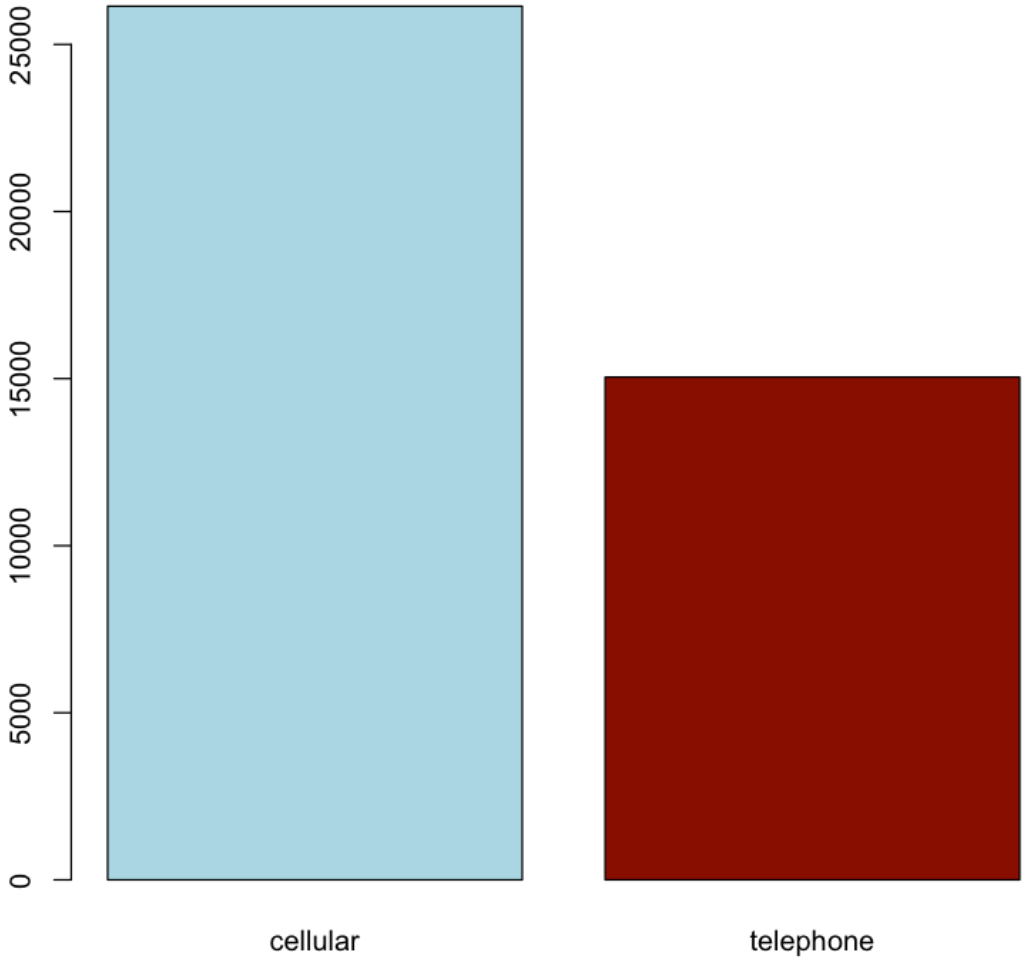
In [55]:
```
sort(round(prop.table(table(bankdata$month))*100,2),decreasing = T)
```

```
  may   jul   aug   jun   nov   apr   oct   sep   mar   dec
33.43 17.42 15.00 12.91  9.96  6.39  1.74  1.38  1.33  0.44
```

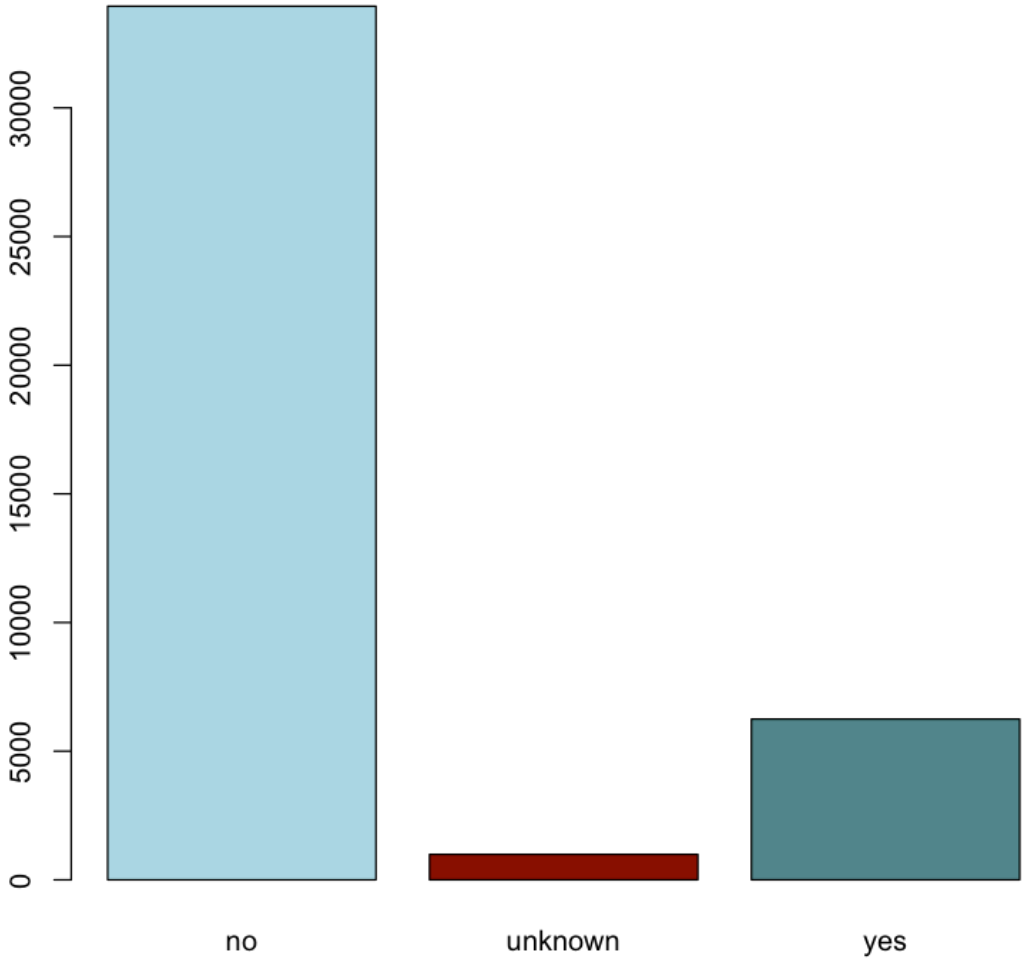
In [56]:
```
sort(round(prop.table(table(bankdata$day_of_week))*100,2),decreasin
```

```
  thu   mon   wed   tue   fri
20.94 20.67 19.75 19.64 19.00
```

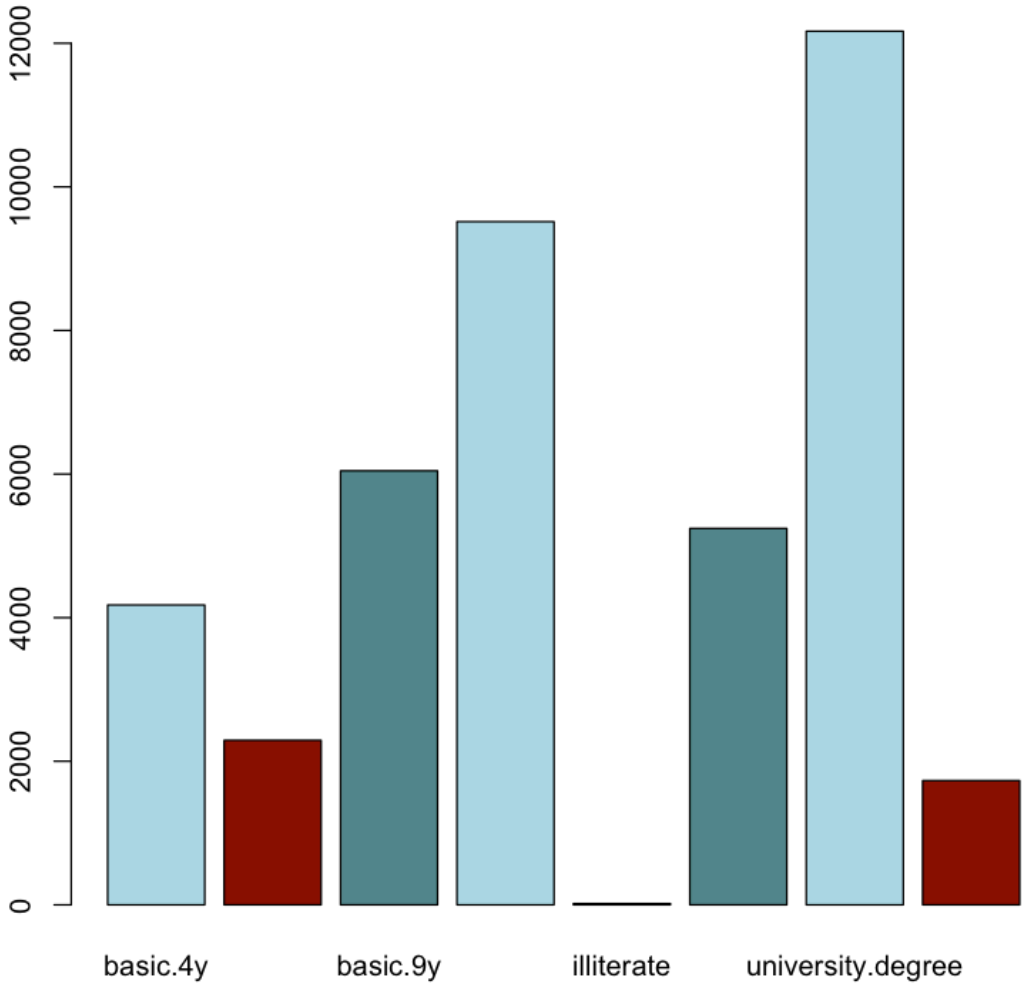
In [57]:
```
barplot(table(bankdata$contact),col=c("lightblue","darkred"))
```



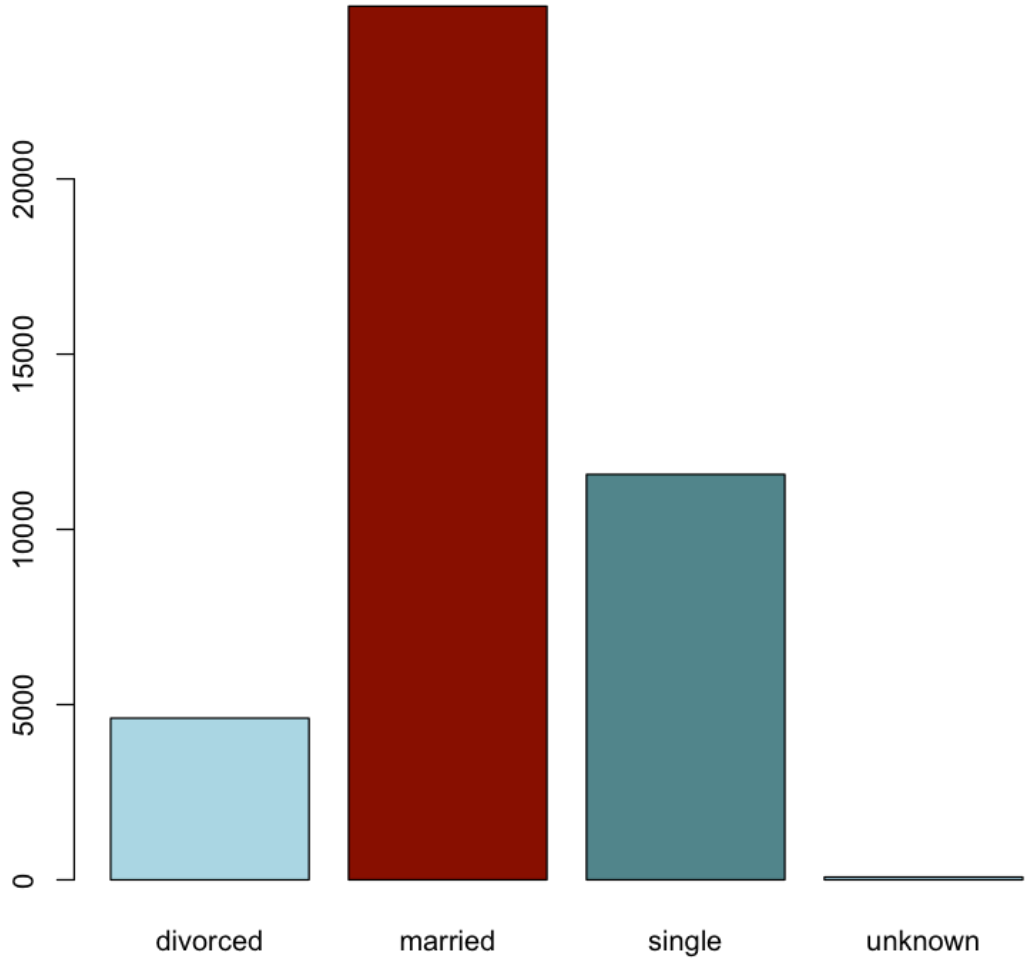
In [58]:
```
barplot(table(bankdata$loan),col=c("lightblue","darkred","cadetblue
```



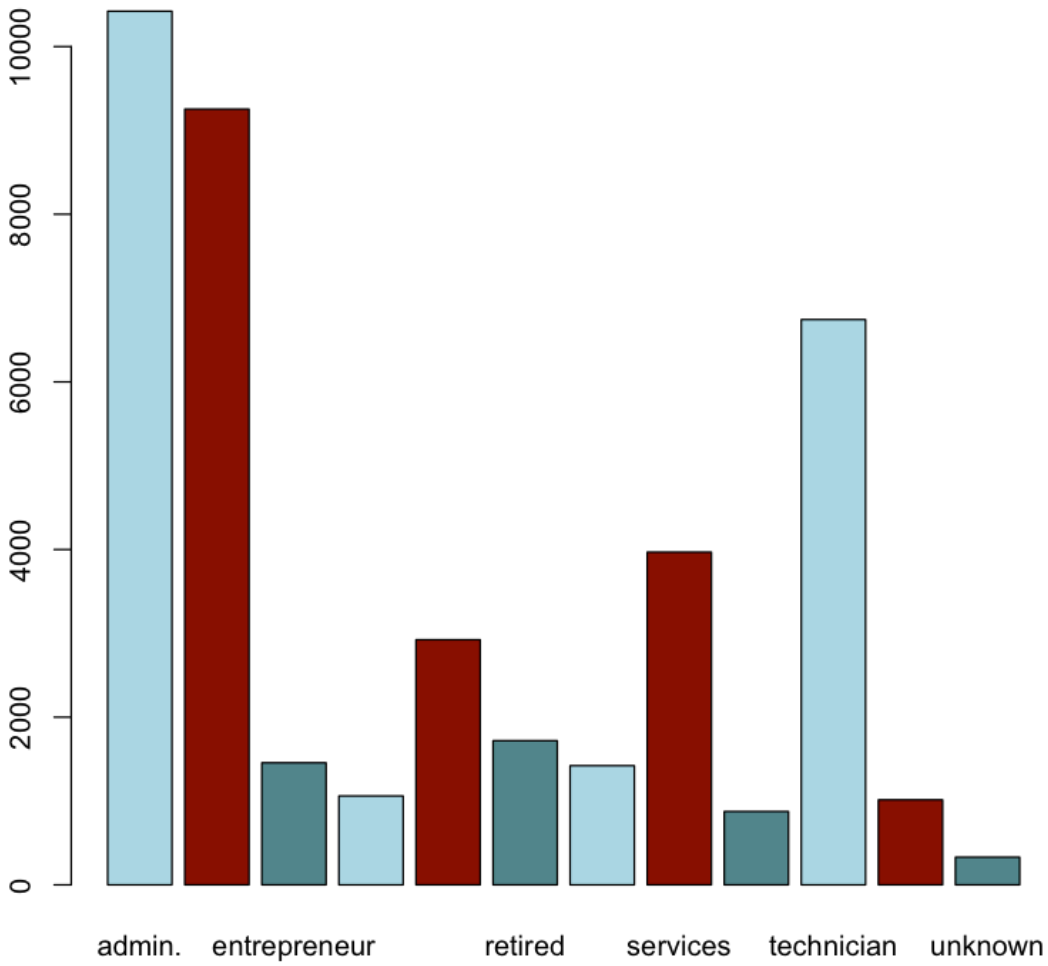
In [59]:
```
barplot(table(bankdata$education),col=c("lightblue","darkred","cade
```



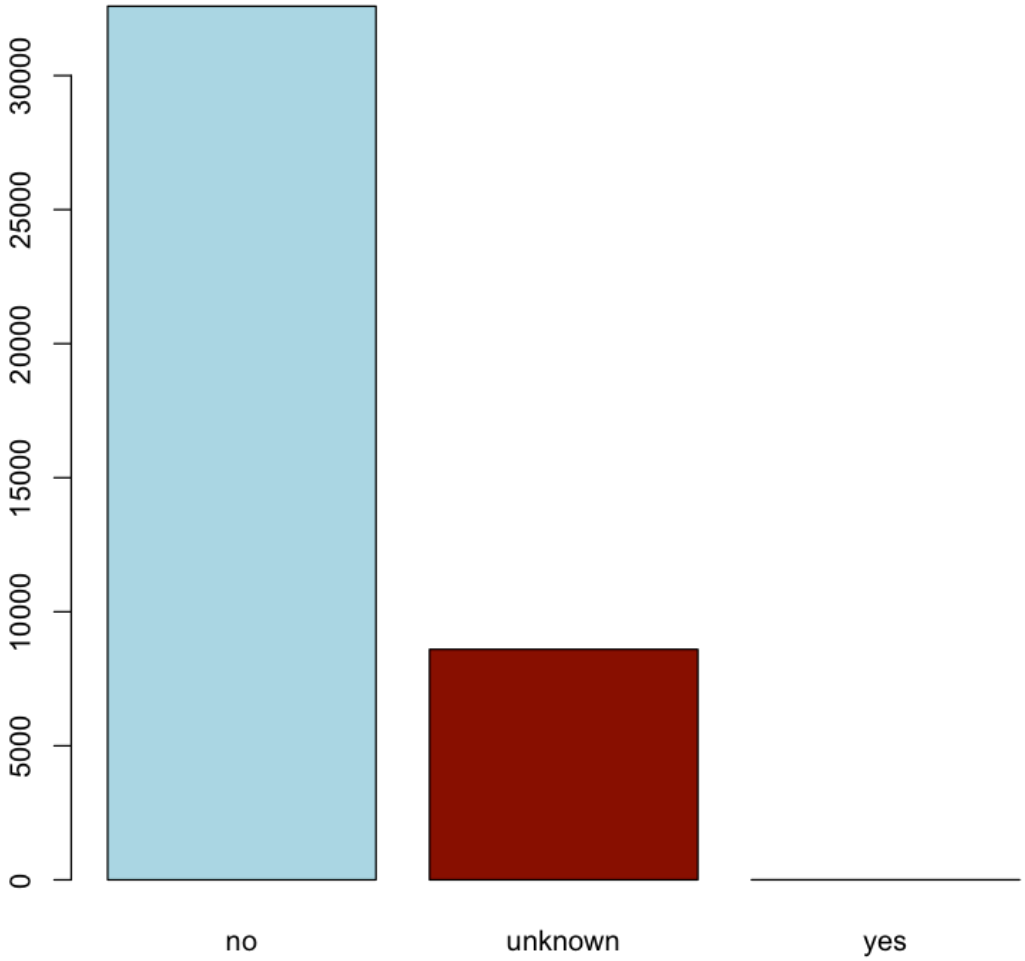
In [60]:
```
barplot(table(bankdata$marital),col=c("lightblue","darkred","cadetb
```



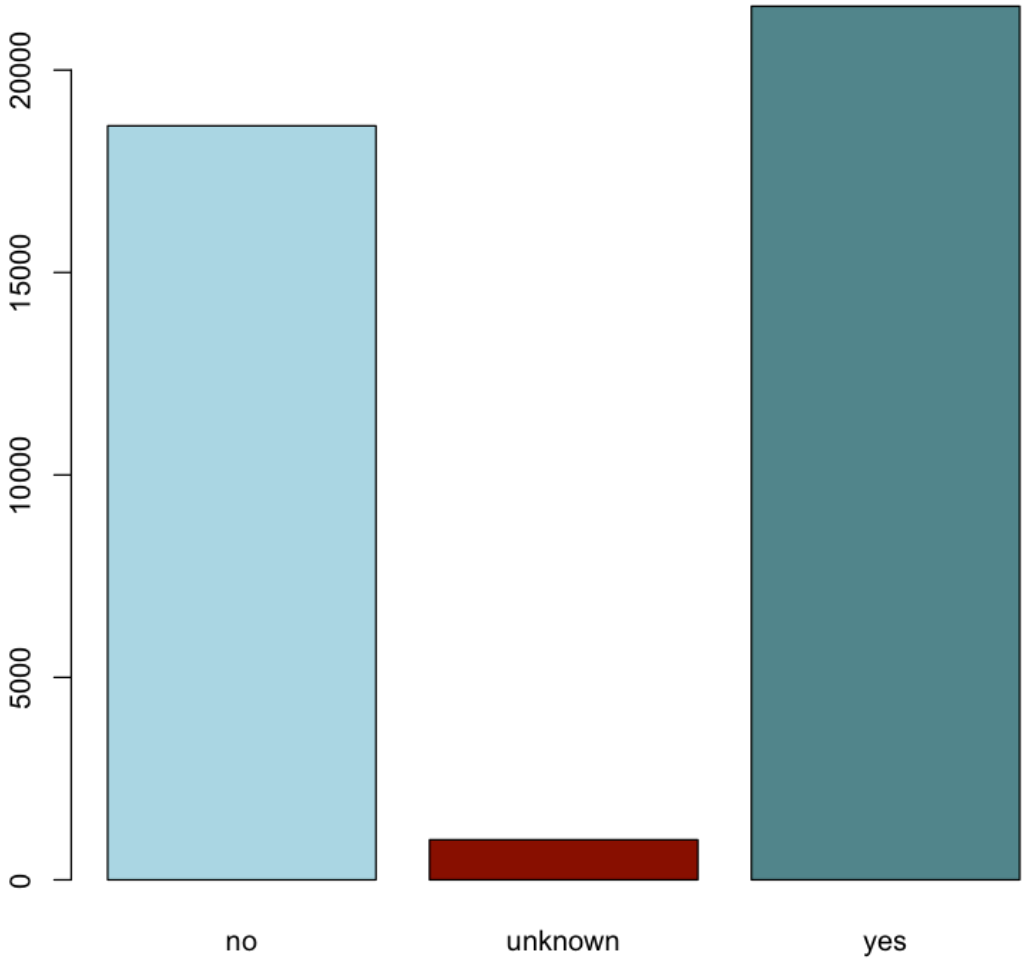
In [61]:
```
barplot(table(bankdata$job),col=c("lightblue","darkred","cadetblue4
```



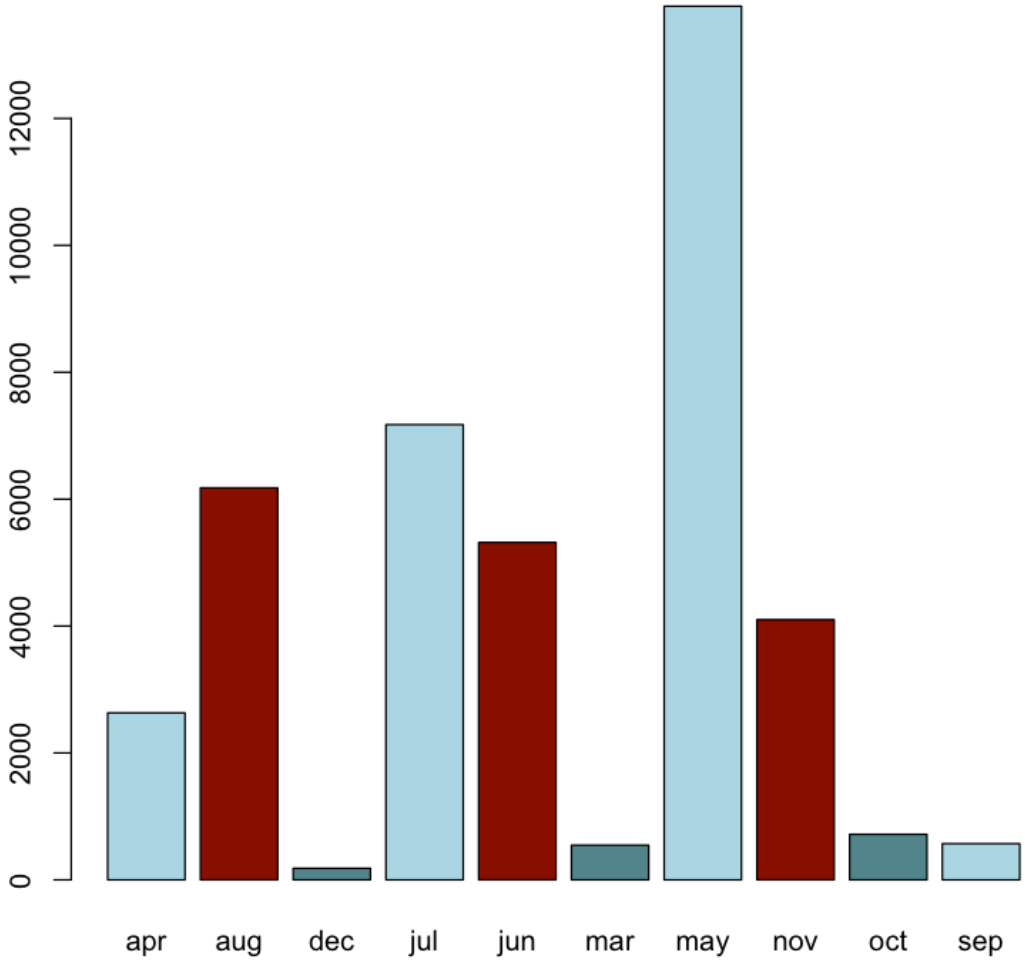
In [62]:
```
barplot(table(bankdata$default),col=c("lightblue","darkred","cadetb
```



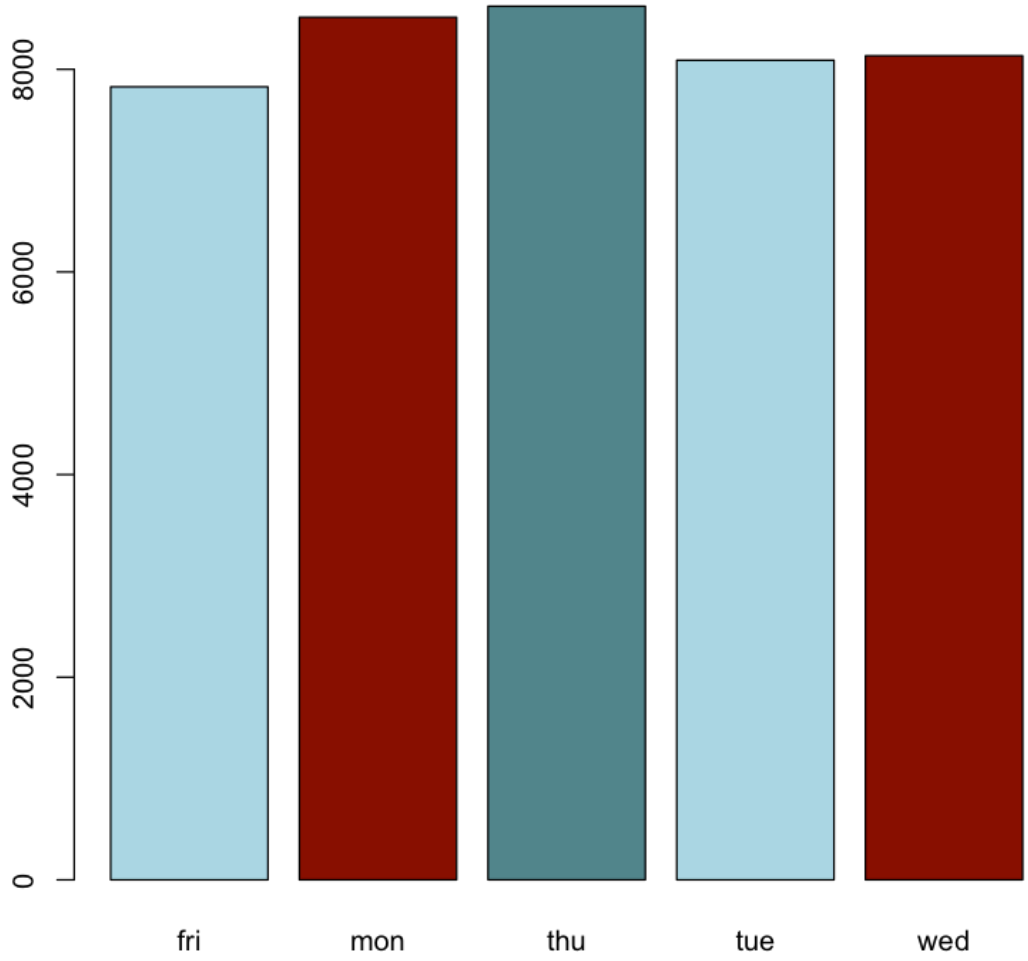
In [63]:
```
barplot(table(bankdata$housing),col=c("lightblue","darkred","cadetb
```



In [64]:
```
barplot(table(bankdata$month),col=c("lightblue","darkred","cadetblu
```



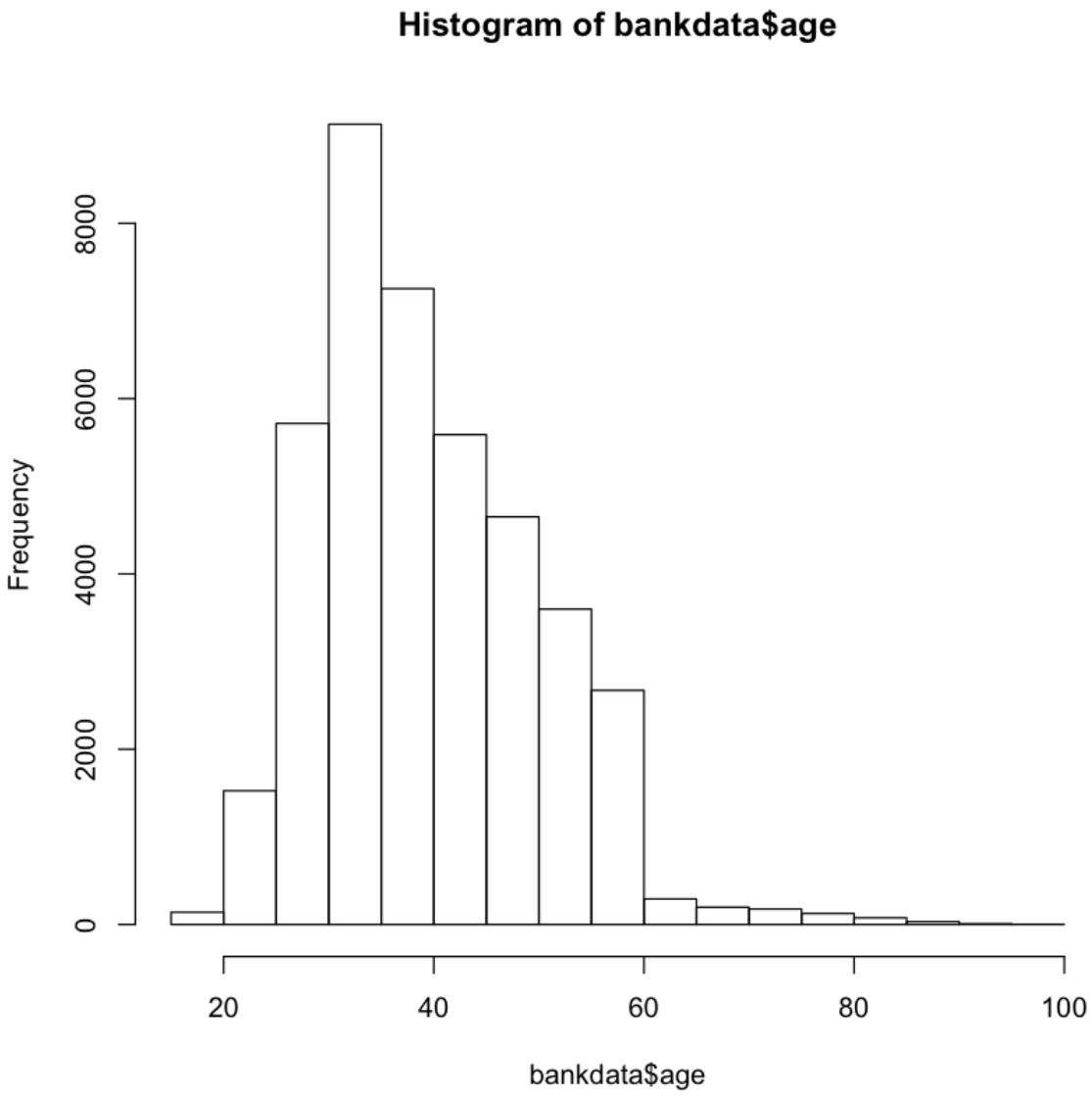
In [65]:
```
barplot(table(bankdata$day_of_week),col=c("lightblue","darkred","ca
```



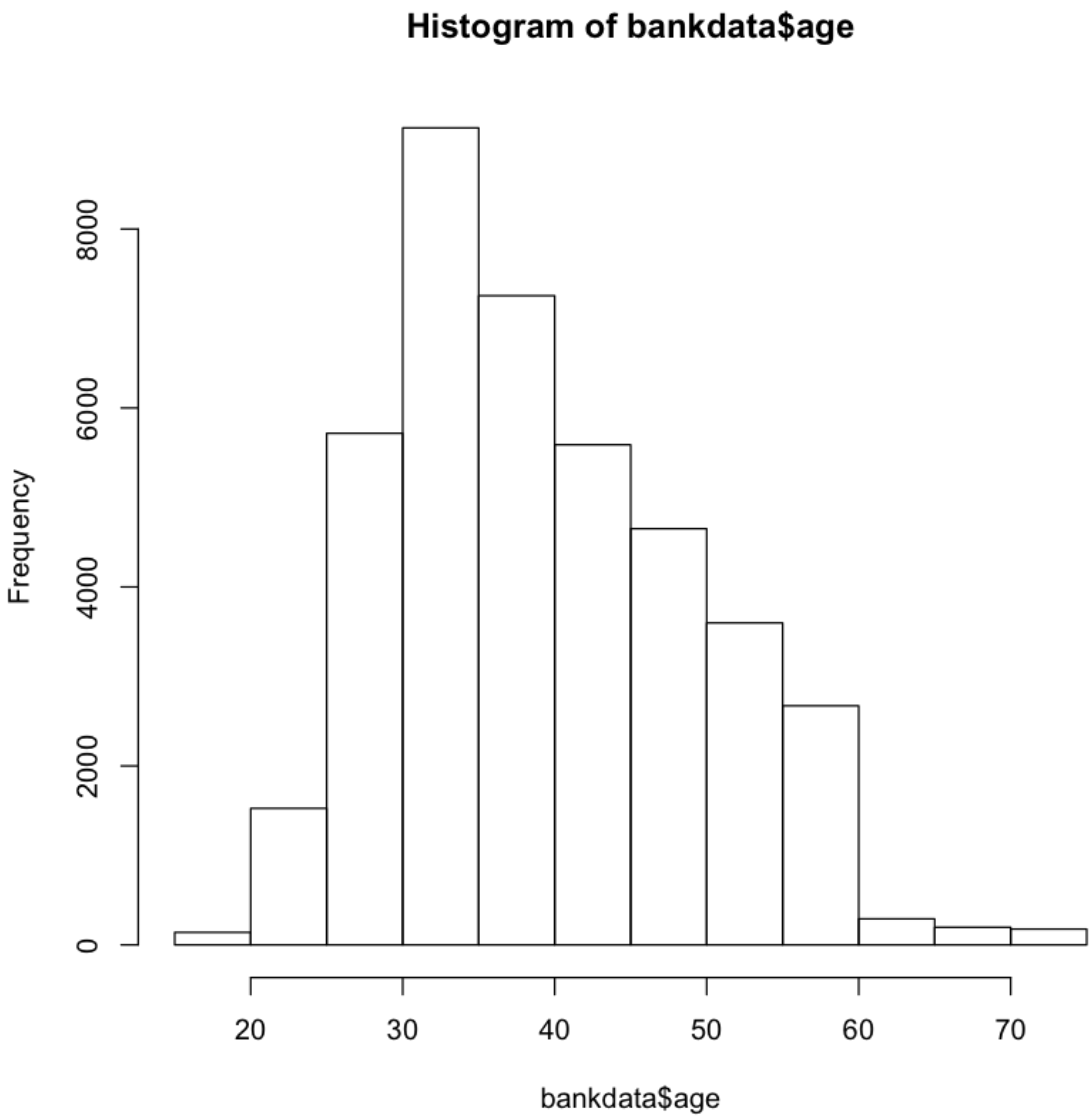
Let us see the analysis of Non- categorical features:

Here as we know age,nr_employed,duration,campaign,pdays,previous,emp_var_rate,cons_price_idx,cons_c are non caegorical features, lets us understand more about them by plotting them using a histogram. We can understand a lot about the distribution of the data using a histogram or in that case any pictorial representation

In [66]:
```
hist(bankdata$age)
```

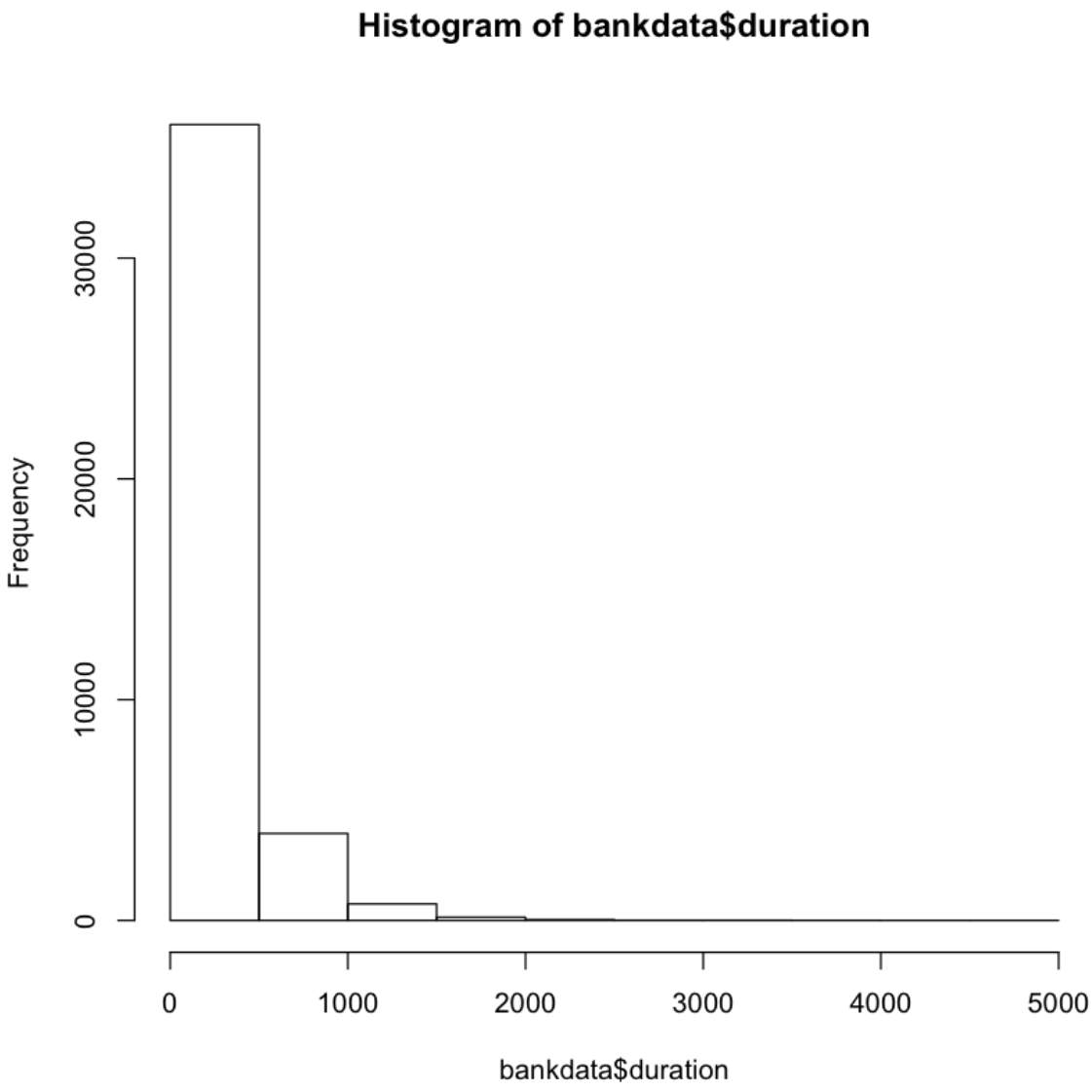


In [67]:
```
#lets us remove the unnecessary data points i.e the points which do
library(magrittr)
library(dplyr)
bankdata <- bankdata %>% filter(age<=75)
hist(bankdata$age)
```

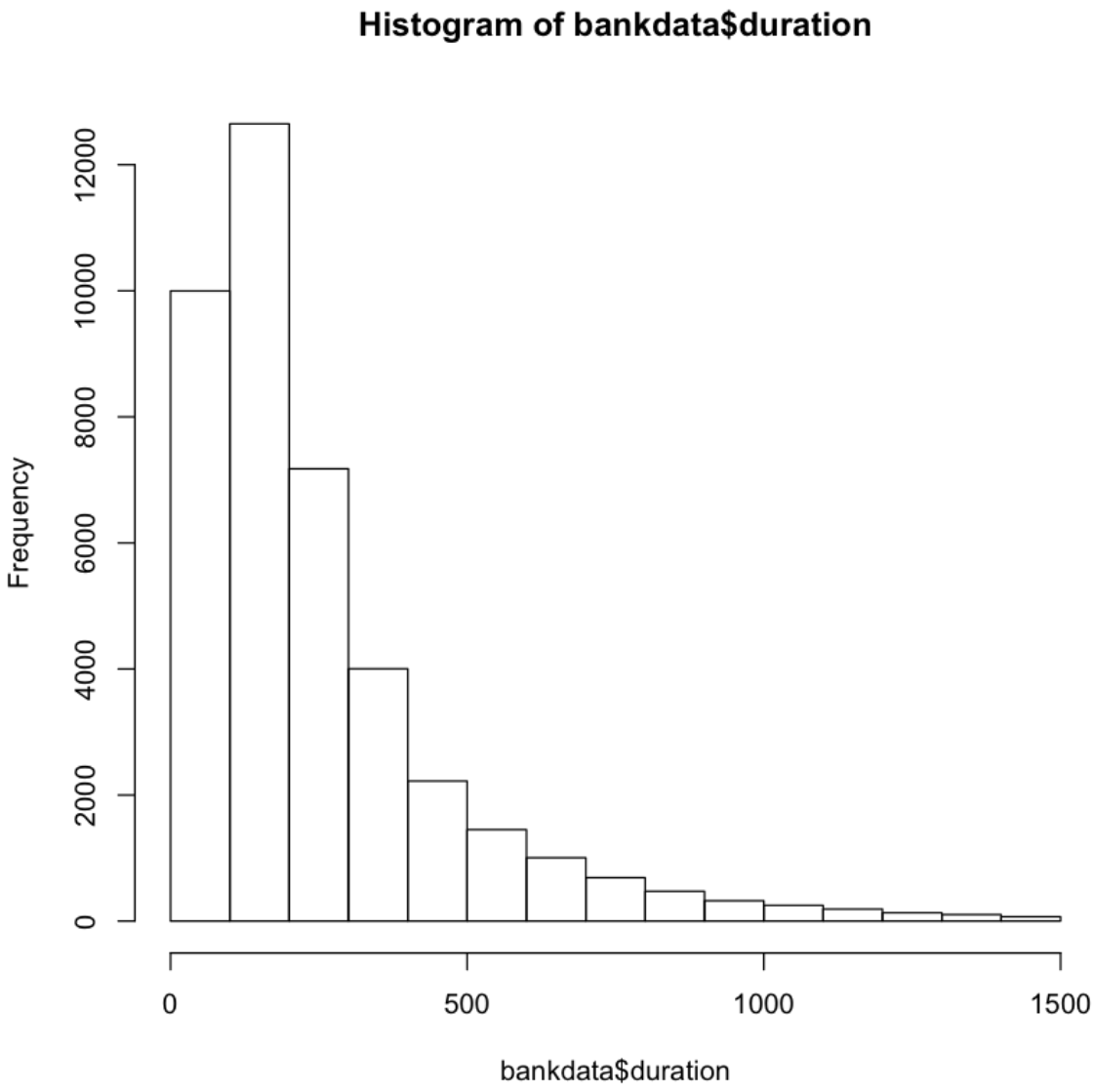


As we can see above, the age is distribued such a way that the majority of the data or people are between the age of 30 to 50(from the histogram) as we have seen above the mean of age is 39 and median is 38 which is very close to each other which means that the data is not highly skewed.

In [16]:
```
#As we have seen from the histogram below, the duration grater than
hist(bankdata$duration)
bankdata <- bankdata %>% filter(duration <= 1500)
```

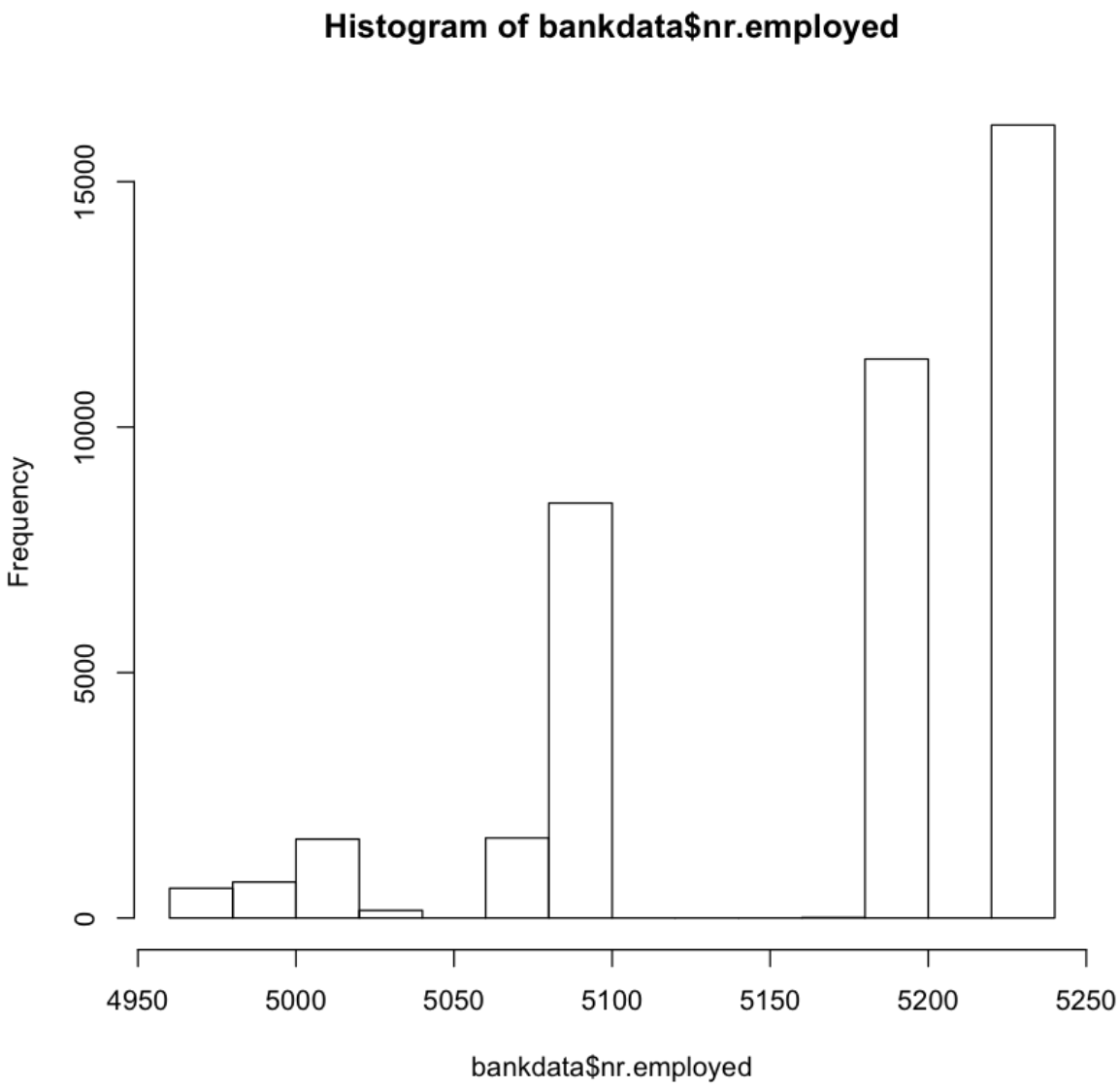


In [17]:



Here we can see that the values above 1500 have very less significance and we can consider tham as outliers and remove the points.

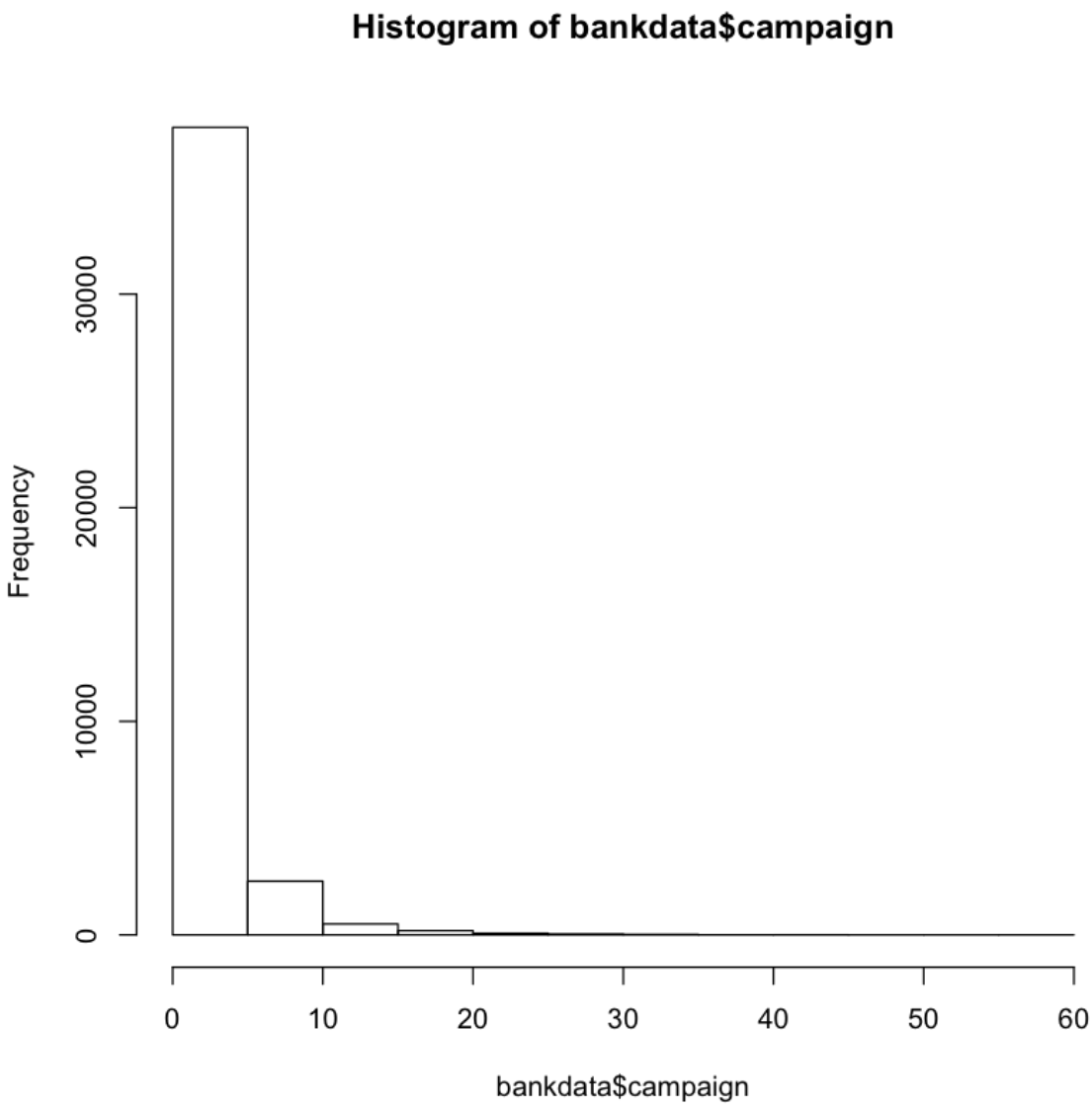
In [18]:
```
hist(bankdata$nr.employed)
```



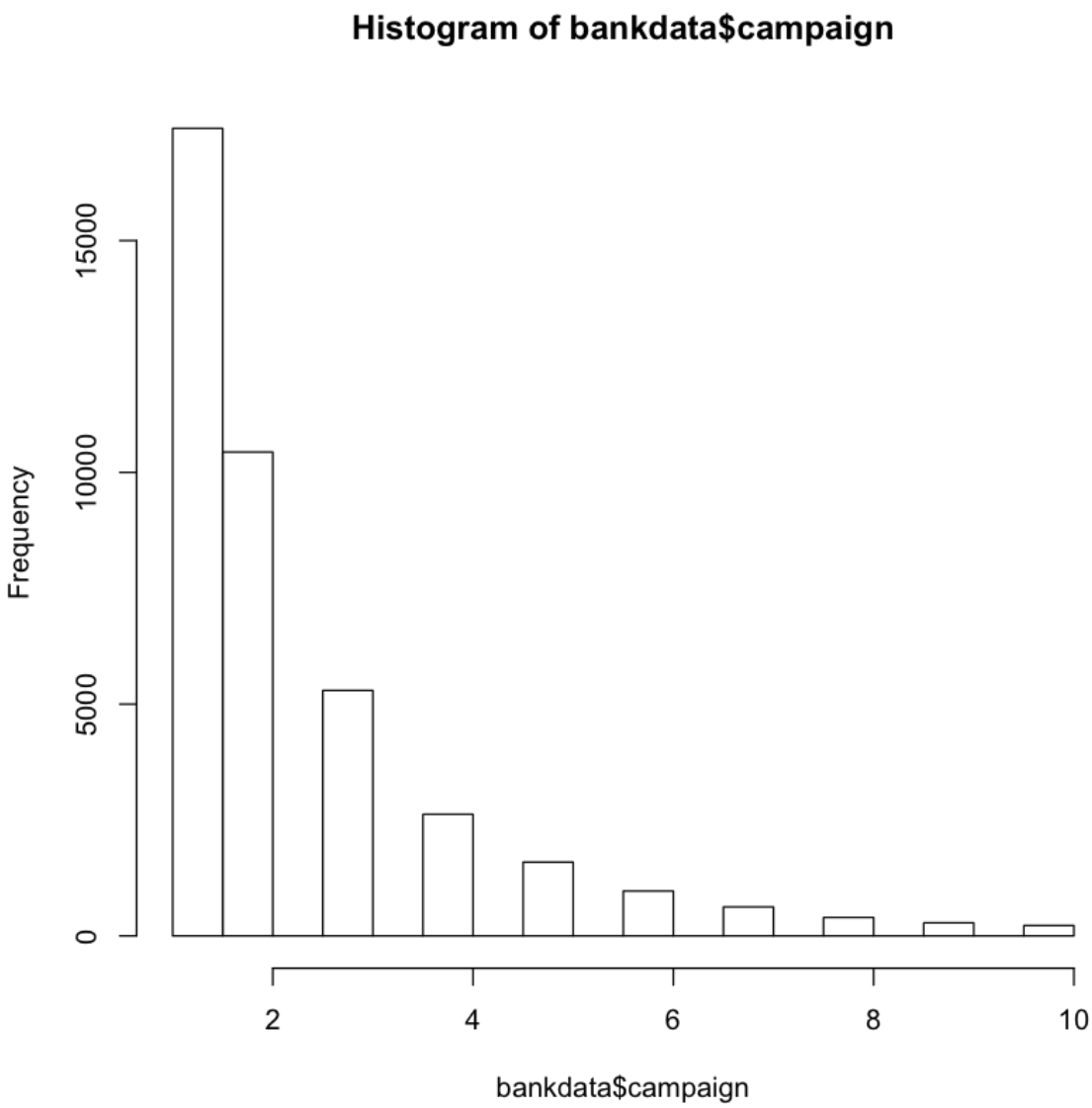
In [ ]:
```

```

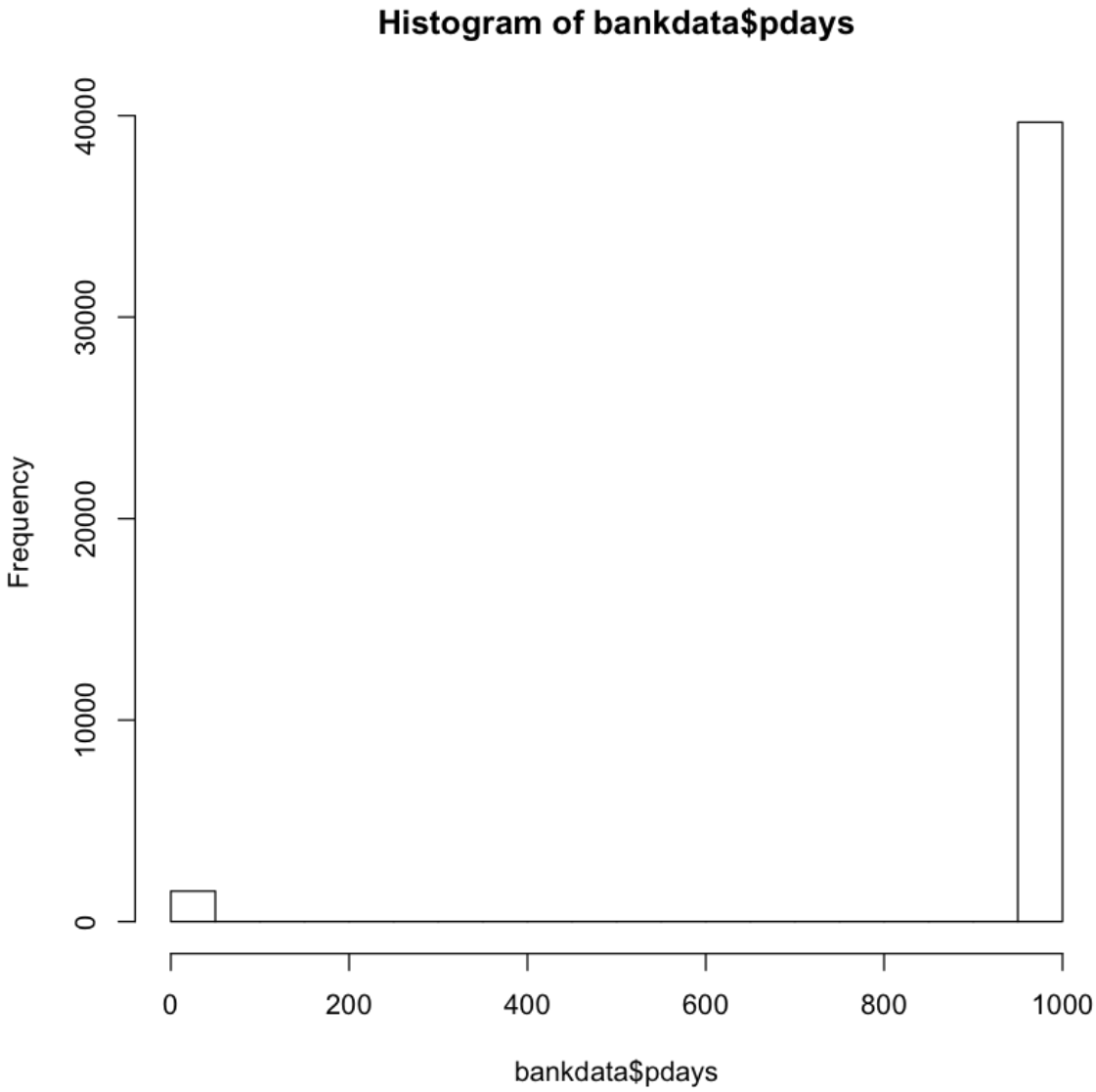
In [93]:
```
hist(bankdata$campaign)
```



In [19]:
```
# as we can see from the histogram above, the data of campaign abov
bankdata <- bankdata %>% filter(campaign <= 10)
hist(bankdata$campaign)
```

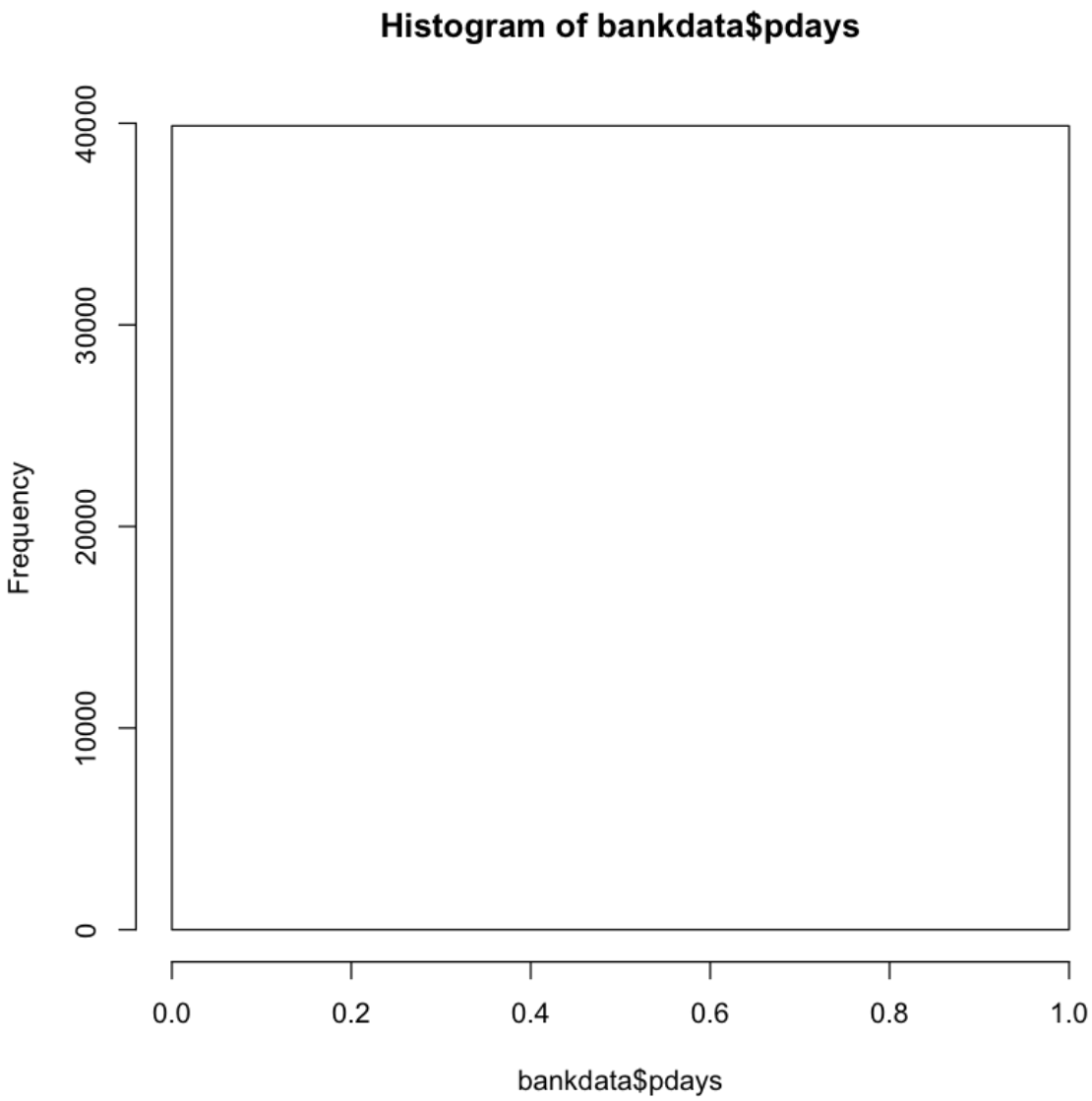


In [94]:
```
hist(bankdata$pdays)
```



In [23]:
```
#Lets also remove the unnecessary data points here.
bankdata<-bankdata %>% mutate(pdays=if_else(pdays==999,0,1))
hist(bankdata$pdays)
table(bankdata$pdays)
```

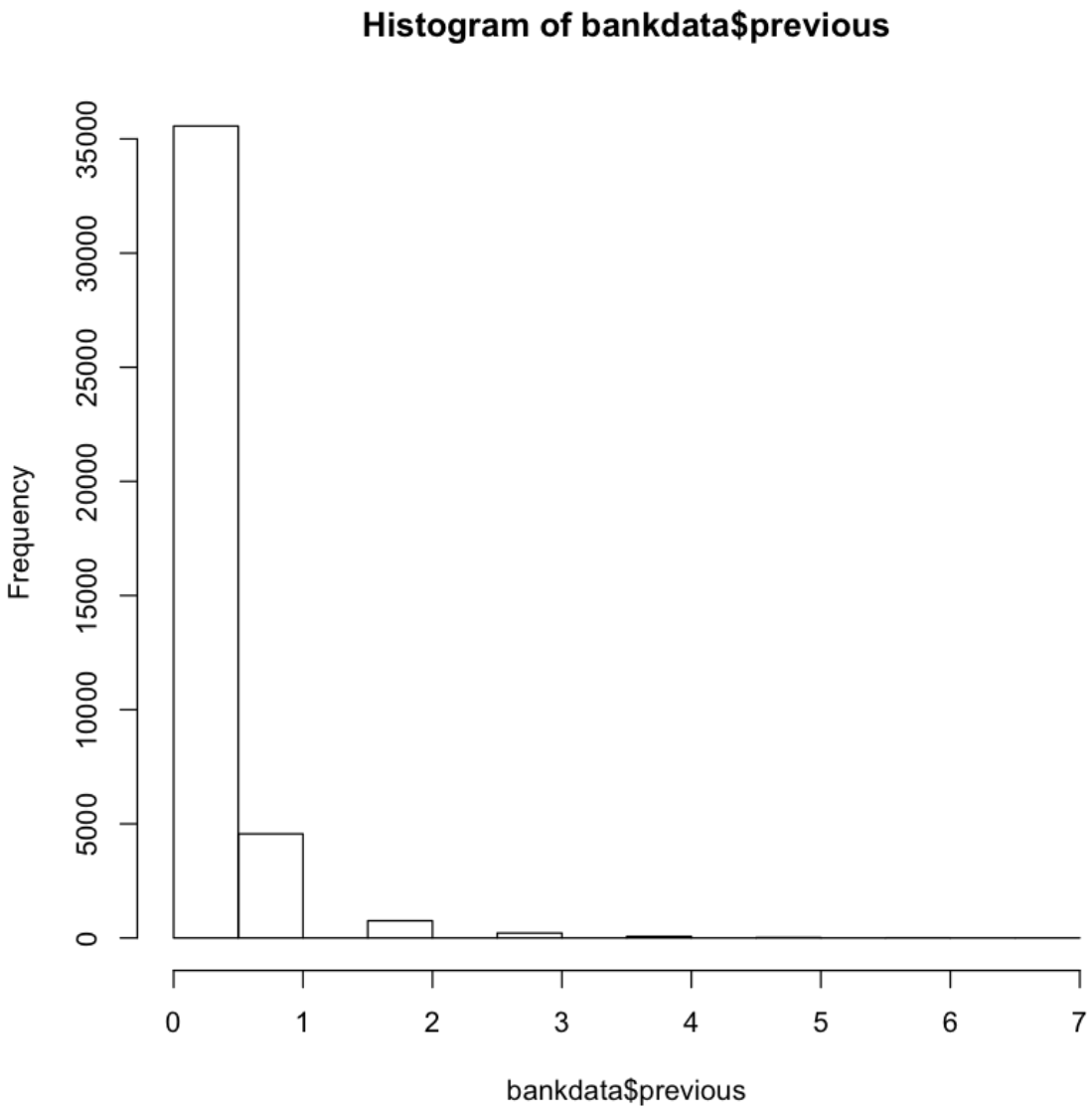
```
    1
39871
```



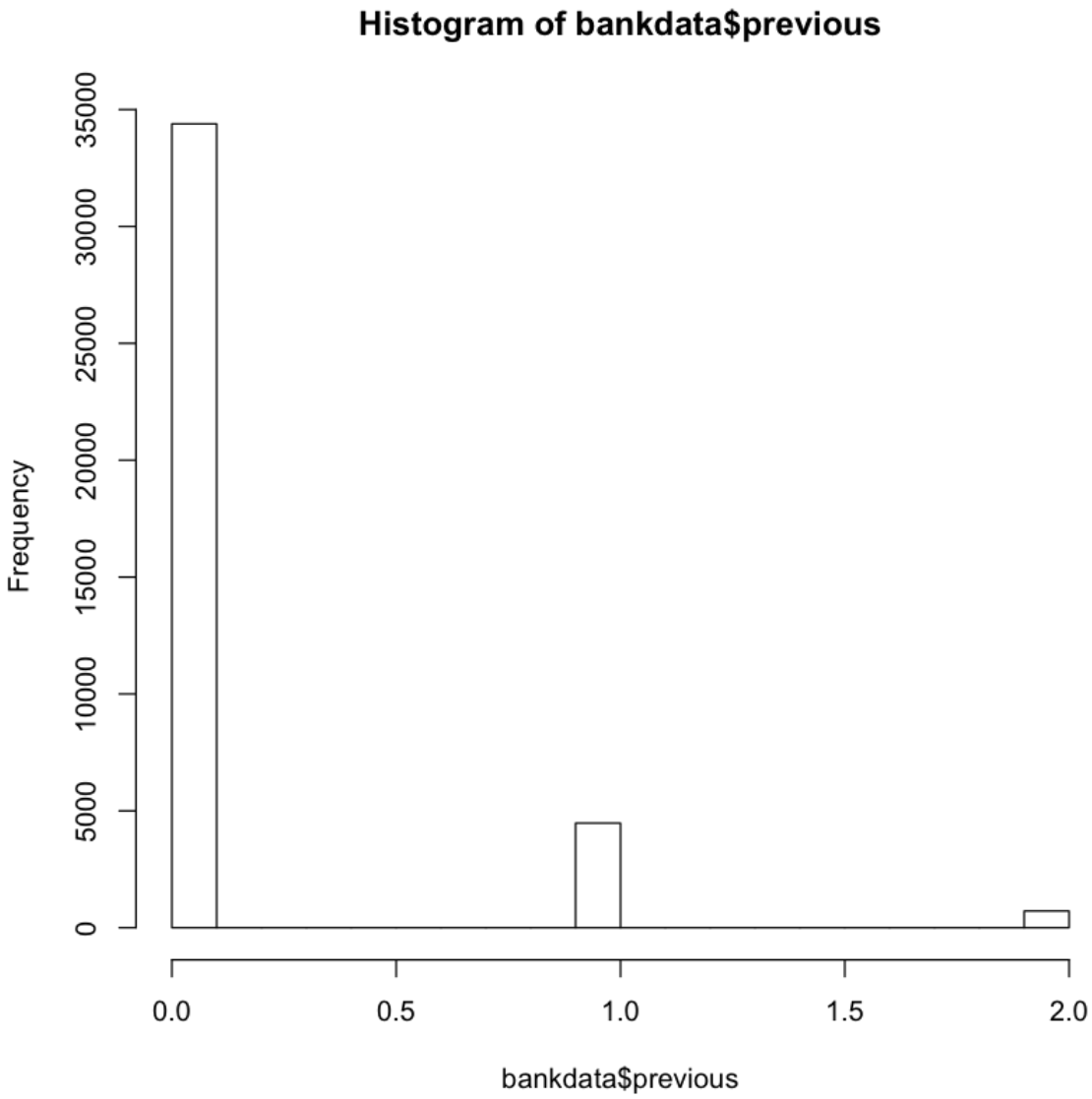
As we can see from the above histogra, comparing them seeing the dataset we can undestand that before contatcing the person in the dataset it is initially represented as 999 and after contacting it is 1, we need to fix this.

so we only kept the values of 999, 0,1 and removed if there are any other values

In [96]:
```
hist(bankdata$previous)
```

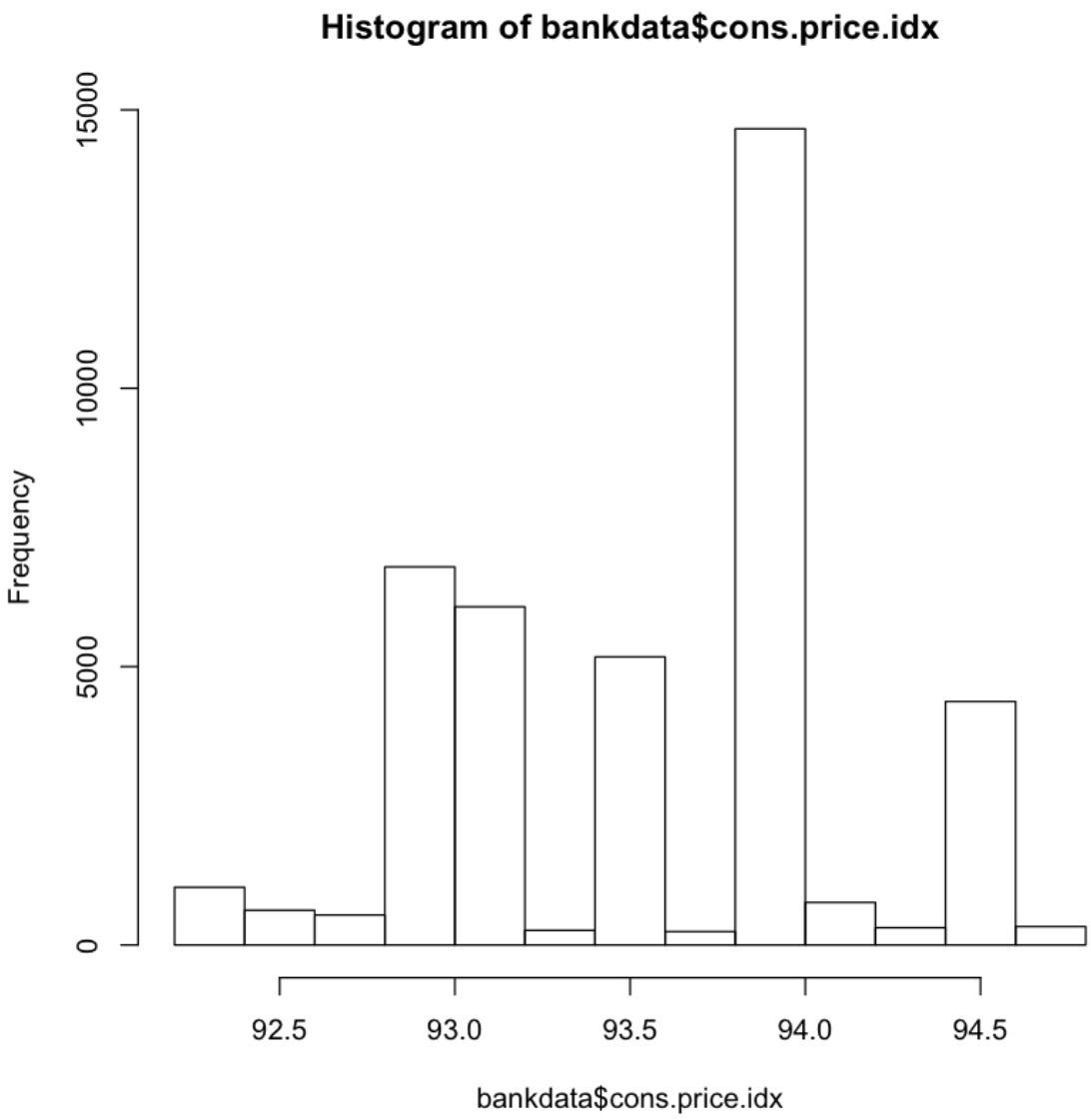


In [24]:
```
# as we can see, we can remove all the points below 2
bankdata <- bankdata %>% filter(previous <=2)
hist(bankdata$previous)
```



Here also as we can see, we can discard the rows with 3,4 and other values as they dont add any value to our dataset.

In [99]:
```
hist(bankdata$cons.price.idx)
```

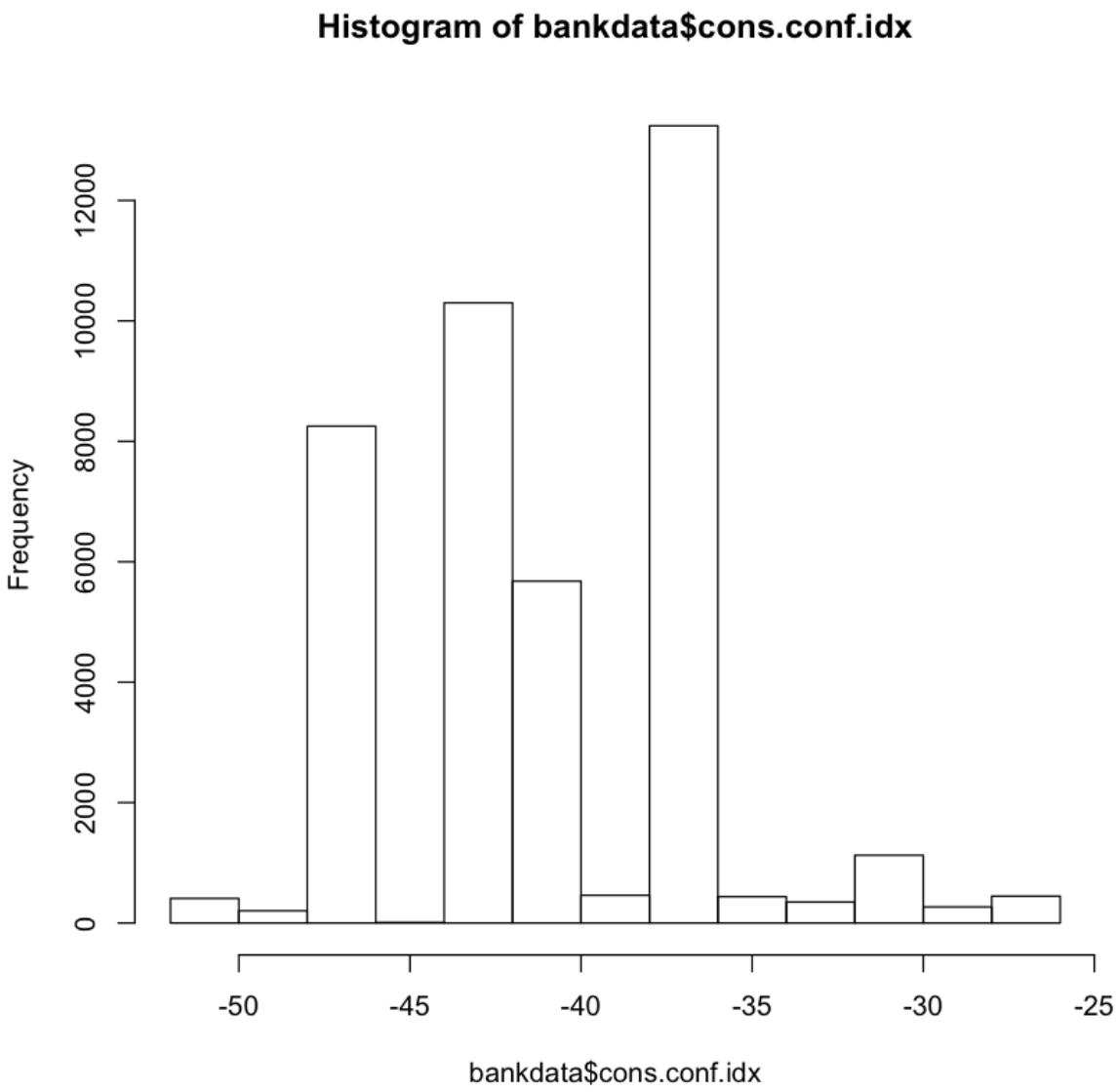


In [119]:
```
hist(bankdata$emp_var_rate)
```

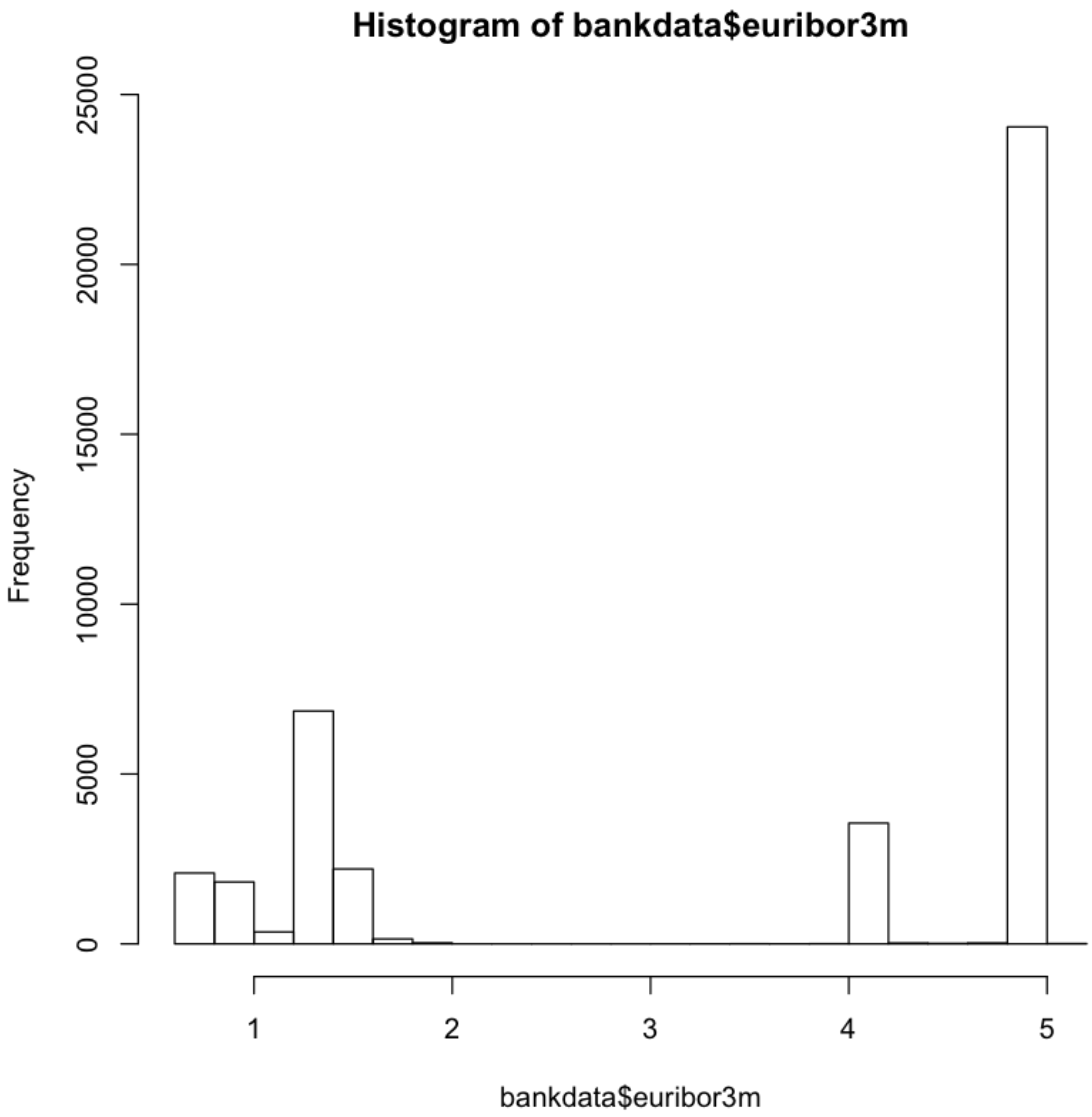
```
Error in hist.default(bankdata$emp_var_rate): 'x' must be numeric
Traceback:

1. hist(bankdata$emp_var_rate)
2. hist.default(bankdata$emp_var_rate)
3. stop("'x' must be numeric")
```

In [100]:
```
hist(bankdata$cons.conf.idx)
```



In [102]:
```
hist(bankdata$euribor3m)
```



In [120]:
```
# we still need to purify our dataset as there are other independen
# we need to fix it by remove those data points.
bankdata = bankdata %>% filter(job != "unknown")
bankdata = bankdata %>% filter(default != "unknown")
bankdata = bankdata %>% filter(housing != "unknown")
bankdata = bankdata %>% filter(loan != "unknown")
bankdata = bankdata %>% filter(marital != "unknown")
bankdata = bankdata %>% filter(education != "unknown")
colSums(bankdata=="unknown")
colSums(bankdata=='')
```

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

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

Analysing each of the individual features(numerical) with the target vaiable

Basically crosstable and using this analysis will give the analysis of each of the categorical varibale. for example we have poutcome variable,where the categorical values are failure, nonexistent, success. 'failure' accounts for 86% of no values in the target variable and 14% accounts for yes.

Same way 'nonexistent' accounts for 91% of no values and 8.7% of yes values. categorical value success accounts for 35% of no values in target. so this analysis. this also tell us which features are more valuable or are giving more information to the target varibale.

In [122]:
```
library(gmodels)
CrossTable(bankdata$job,bankdata$y, prop.c = F, prop.t = F, prop.ch
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|-------------------------|


Total Observations in Table:  30488


               | bankdata$y
 bankdata$job |        no |       yes | Row Total |
--------------|-----------|-----------|-----------|
       admin. |      7521 |      1216 |      8737 |
              |     0.861 |     0.139 |     0.287 |
--------------|-----------|-----------|-----------|
  blue-collar |      5223 |       452 |      5675 |
              |     0.920 |     0.080 |     0.186 |
--------------|-----------|-----------|-----------|
 entrepreneur |       988 |       101 |      1089 |
              |     0.907 |     0.093 |     0.036 |
--------------|-----------|-----------|-----------|
    housemaid |       603 |        87 |       690 |
              |     0.874 |     0.126 |     0.023 |
--------------|-----------|-----------|-----------|
   management |      2025 |       286 |      2311 |
              |     0.876 |     0.124 |     0.076 |
--------------|-----------|-----------|-----------|
      retired |       859 |       357 |      1216 |
              |     0.706 |     0.294 |     0.040 |
--------------|-----------|-----------|-----------|
self-employed |       960 |       132 |      1092 |
              |     0.879 |     0.121 |     0.036 |
--------------|-----------|-----------|-----------|
     services |      2599 |       258 |      2857 |
              |     0.910 |     0.090 |     0.094 |
--------------|-----------|-----------|-----------|
      student |       407 |       203 |       610 |
              |     0.667 |     0.333 |     0.020 |
--------------|-----------|-----------|-----------|
   technician |      4832 |       641 |      5473 |
              |     0.883 |     0.117 |     0.180 |
--------------|-----------|-----------|-----------|
   unemployed |       612 |       126 |       738 |
              |     0.829 |     0.171 |     0.024 |
--------------|-----------|-----------|-----------|
 Column Total |     26629 |      3859 |     30488 |
--------------|-----------|-----------|-----------|
```

In [123]:
```
CrossTable(bankdata$housing, bankdata$y, prop.c = F, prop.t = F, pr
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|-------------------------|


Total Observations in Table:  30488


                 | bankdata$y
bankdata$housing |        no |       yes | Row Total |
-----------------|-----------|-----------|-----------|
              no |     12250 |      1717 |     13967 |
                 |     0.877 |     0.123 |     0.458 |
-----------------|-----------|-----------|-----------|
             yes |     14379 |      2142 |     16521 |
                 |     0.870 |     0.130 |     0.542 |
-----------------|-----------|-----------|-----------|
    Column Total |     26629 |      3859 |     30488 |
-----------------|-----------|-----------|-----------|
```

In [124]:
```
 CrossTable(bankdata$contact, bankdata$y, prop.c = F, prop.t = F, p
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|-------------------------|


Total Observations in Table:  30488


                 | bankdata$y
bankdata$contact |        no |       yes | Row Total |
-----------------|-----------|-----------|-----------|
        cellular |     17170 |      3273 |     20443 |
                 |     0.840 |     0.160 |     0.671 |
-----------------|-----------|-----------|-----------|
       telephone |      9459 |       586 |     10045 |
                 |     0.942 |     0.058 |     0.329 |
-----------------|-----------|-----------|-----------|
    Column Total |     26629 |      3859 |     30488 |
-----------------|-----------|-----------|-----------|
```

In [125]:
```
CrossTable(bankdata$month, bankdata$y, prop.c = F, prop.t = F, pro
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|-------------------------|


Total Observations in Table:  30488


                | bankdata$y
bankdata$month |        no |       yes | Row Total |
---------------|-----------|-----------|-----------|
           apr |      1647 |       468 |      2115 |
               |     0.779 |     0.221 |     0.069 |
---------------|-----------|-----------|-----------|
           aug |      4140 |       533 |      4673 |
               |     0.886 |     0.114 |     0.153 |
---------------|-----------|-----------|-----------|
           dec |        83 |        74 |       157 |
               |     0.529 |     0.471 |     0.005 |
---------------|-----------|-----------|-----------|
           jul |      4569 |       512 |      5081 |
               |     0.899 |     0.101 |     0.167 |
---------------|-----------|-----------|-----------|
           jun |      3162 |       452 |      3614 |
               |     0.875 |     0.125 |     0.119 |
---------------|-----------|-----------|-----------|
           mar |       236 |       246 |       482 |
               |     0.490 |     0.510 |     0.016 |
---------------|-----------|-----------|-----------|
           may |      9033 |       700 |      9733 |
               |     0.928 |     0.072 |     0.319 |
---------------|-----------|-----------|-----------|
           nov |      3131 |       365 |      3496 |
               |     0.896 |     0.104 |     0.115 |
---------------|-----------|-----------|-----------|
           oct |       355 |       287 |       642 |
               |     0.553 |     0.447 |     0.021 |
---------------|-----------|-----------|-----------|
           sep |       273 |       222 |       495 |
               |     0.552 |     0.448 |     0.016 |
---------------|-----------|-----------|-----------|
  Column Total |     26629 |      3859 |     30488 |
---------------|-----------|-----------|-----------|
```

In [126]:
```
CrossTable(bankdata$day_of_week, bankdata$y, prop.c = F, prop.t =
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|-------------------------|


Total Observations in Table:  30488


                     | bankdata$y
bankdata$day_of_week |        no |       yes | Row Total |
---------------------|-----------|-----------|-----------|
                 fri |      5058 |       676 |      5734 |
                     |     0.882 |     0.118 |     0.188 |
---------------------|-----------|-----------|-----------|
                 mon |      5573 |       706 |      6279 |
                     |     0.888 |     0.112 |     0.206 |
---------------------|-----------|-----------|-----------|
                 thu |      5516 |       879 |      6395 |
                     |     0.863 |     0.137 |     0.210 |
---------------------|-----------|-----------|-----------|
                 tue |      5166 |       789 |      5955 |
                     |     0.868 |     0.132 |     0.195 |
---------------------|-----------|-----------|-----------|
                 wed |      5316 |       809 |      6125 |
                     |     0.868 |     0.132 |     0.201 |
---------------------|-----------|-----------|-----------|
        Column Total |     26629 |      3859 |     30488 |
---------------------|-----------|-----------|-----------|
```

In [127]:
```
CrossTable(bankdata$poutcome, bankdata$y, prop.c = F, prop.t = F,
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|-------------------------|


Total Observations in Table:  30488


                   | bankdata$y
bankdata$poutcome |        no |       yes | Row Total |
------------------|-----------|-----------|-----------|
          failure |      2953 |       508 |      3461 |
                  |     0.853 |     0.147 |     0.114 |
------------------|-----------|-----------|-----------|
      nonexistent |     23264 |      2572 |     25836 |
                  |     0.900 |     0.100 |     0.847 |
------------------|-----------|-----------|-----------|
          success |       412 |       779 |      1191 |
                  |     0.346 |     0.654 |     0.039 |
------------------|-----------|-----------|-----------|
     Column Total |     26629 |      3859 |     30488 |
------------------|-----------|-----------|-----------|
```

CHI Squared test: To find if two categorical features are dependent on eachother, though chisquared test we can see if one two varaibles are dependent on eachother then value of one variable will change the probalbility distribution of another.

If the value of Xsquare is less than p value then the varaibles are independent, else dependent.

In [109]:
```
chisq.test(bankdata$housing,bankdata$y)
```

```
        Pearson's Chi-squared test

data:  bankdata$housing and bankdata$y
X-squared = 5.6845, df = 2, p-value = 0.05829
```

In [110]:
```
chisq.test(bankdata$job,bankdata$y)
```

```
        Pearson's Chi-squared test

data:  bankdata$job and bankdata$y
X-squared = 961.24, df = 11, p-value < 2.2e-16
```

In [111]:
```
chisq.test(bankdata$poutcome,bankdata$y)
```

```
        Pearson's Chi-squared test

data:  bankdata$poutcome and bankdata$y
X-squared = 4230.5, df = 2, p-value < 2.2e-16
```

In [112]:
```
chisq.test(bankdata$campaign,bankdata$y)
```

```
Warning message in chisq.test(bankdata$campaign, bankdata$y):
“Chi-squared approximation may be incorrect”

        Pearson's Chi-squared test

data:  bankdata$campaign and bankdata$y
X-squared = 218.86, df = 41, p-value < 2.2e-16
```

In [113]:
```
chisq.test(bankdata$day_of_week,bankdata$y)
```

```
        Pearson's Chi-squared test

data:  bankdata$day_of_week and bankdata$y
X-squared = 26.145, df = 4, p-value = 2.958e-05
```

In [114]:
```
chisq.test(bankdata$month,bankdata$y)
```

```
        Pearson's Chi-squared test

data:  bankdata$month and bankdata$y
X-squared = 3101.1, df = 9, p-value < 2.2e-16
```

In [115]:
```
chisq.test(bankdata$contact,bankdata$y)
```

```
        Pearson's Chi-squared test with Yates' continuity correcti
on

data:  bankdata$contact and bankdata$y
X-squared = 862.32, df = 1, p-value < 2.2e-16
```

In [116]:
```
chisq.test(bankdata$default,bankdata$y)
```

```
Warning message in chisq.test(bankdata$default, bankdata$y):
“Chi-squared approximation may be incorrect”


        Pearson's Chi-squared test

data:  bankdata$default and bankdata$y
X-squared = 406.58, df = 2, p-value < 2.2e-16
```

In [117]:
```
chisq.test(bankdata$loan,bankdata$y)
```

```
        Pearson's Chi-squared test

data:  bankdata$loan and bankdata$y
X-squared = 1.094, df = 2, p-value = 0.5787
```

In [118]:
```
chisq.test(bankdata$previous,bankdata$y)
```

```
Warning message in chisq.test(bankdata$previous, bankdata$y):
“Chi-squared approximation may be incorrect”


        Pearson's Chi-squared test

data:  bankdata$previous and bankdata$y
X-squared = 2299.4, df = 7, p-value < 2.2e-16
```

In [ ]:
```

```

Calculating Variable Importance:

Here we are finiding the variable importance, this gives us a numerical values of how much each of the features which impacts the highest to the target variable. month and duration, emp.var.rate have the highest importance, housing and marital have the lowest importance.

In [103]:
```
library(data.table)
library(mltools)
library(party)
```

In [104]:
```
cf1 <- cforest( y~ . , data= bankdata, control=cforest_unbiased(mtr
```

```
         Random Forest using Conditional Inference Trees

Number of trees:  50

Response:  y
Inputs:  age, job, marital, education, default, housing, loan, con
tact, month, day_of_week, duration, campaign, pdays, previous, pou
tcome, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.
employed
Number of observations:  41188
```

In [107]:
```
a<-varimp(cf1)
a
sort(a,decreasing = TRUE)
```

age	0.00019924787227024
job	0.00026918255591476
marital	-0.000240153064590618
education	0.000102922741967409
default	0.000188691693606915
housing	-0.00012535462162697
loan	6.729563897869e-05
contact	0.00182621890875503
month	0.0179362670713202
day_of_week	0.0014145279408854
duration	0.0383545556508544
campaign	0.000489542785511646
pdays	0.00291350531107739
previous	0.00104374216533615
poutcome	0.00422115194299664
emp.var.rate	0.015566404961404
cons.price.idx	0.0076690637989048
cons.conf.idx	0.00382265619845616
euribor3m	0.0104176288183678
nr.employed	0.0147931648743155

duration	0.0383545556508544
month	0.0179362670713202
emp.var.rate	0.015566404961404
nr.employed	0.0147931648743155
euribor3m	0.0104176288183678
cons.price.idx	0.0076690637989048
poutcome	0.00422115194299664
cons.conf.idx	0.00382265619845616
pdays	0.00291350531107739
contact	0.00182621890875503
day_of_week	0.0014145279408854
previous	0.00104374216533615
campaign	0.000489542785511646
job	0.00026918255591476
age	0.00019924787227024
default	0.000188691693606915
education	0.000102922741967409
loan	6.729563897869e-05
housing	-0.00012535462162697
marital	-0.000240153064590618

In [ ]:

In [ ]: