Bank_Marketing_DataAnalysis.R

nandinisahni

Fri Feb 15 16:49:26 2019

```
#Import Data
bank <- read.csv("~/Documents/Study/Semester2/
Multivariate/bank-additional/bank-additional-full.csv",
sep=";")
bank_marketing <- bank

#View(bank_marketing)
head(bank_marketing)
##   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
```

```
str(bank_marketing)
## '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 ...
```

```
#Boxplots to check for any outliers
boxplot(bank_marketing$age, main="Age Box
plot",yaxt="n", xlab="Age", horizontal=TRUE)
```
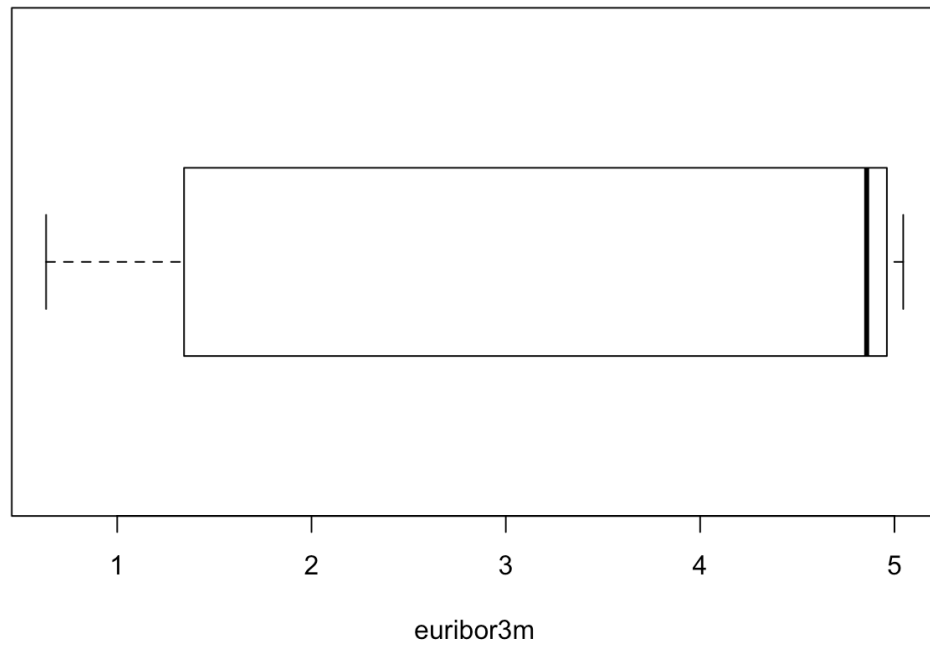
Age Box plot

Age
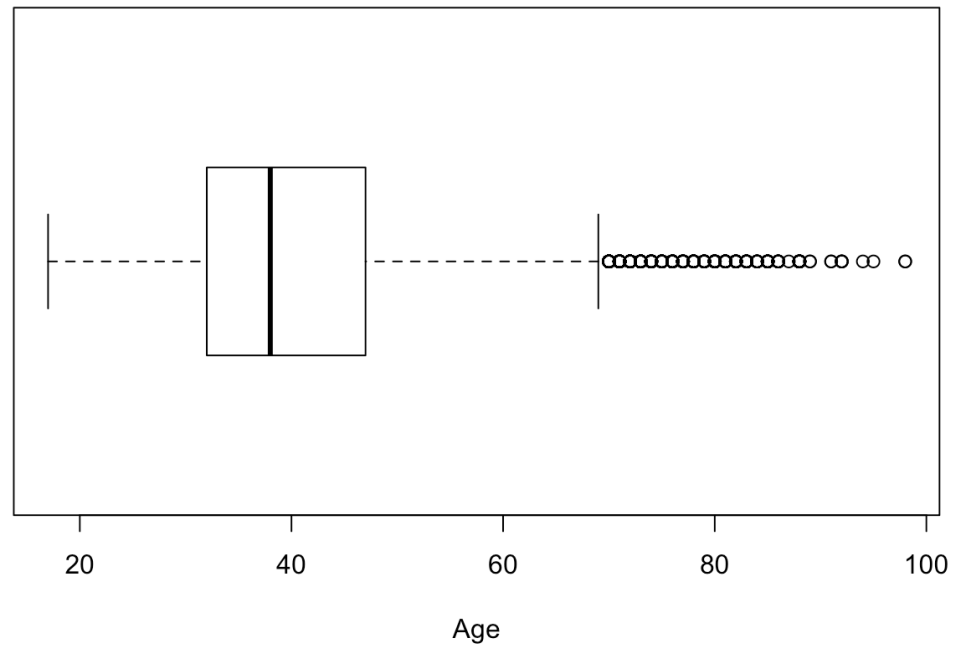
```
boxplot(bank_marketing$euribor3m, main="Euribor3m Box
plot",yaxt="n", xlab="euribor3m", horizontal=TRUE)
```
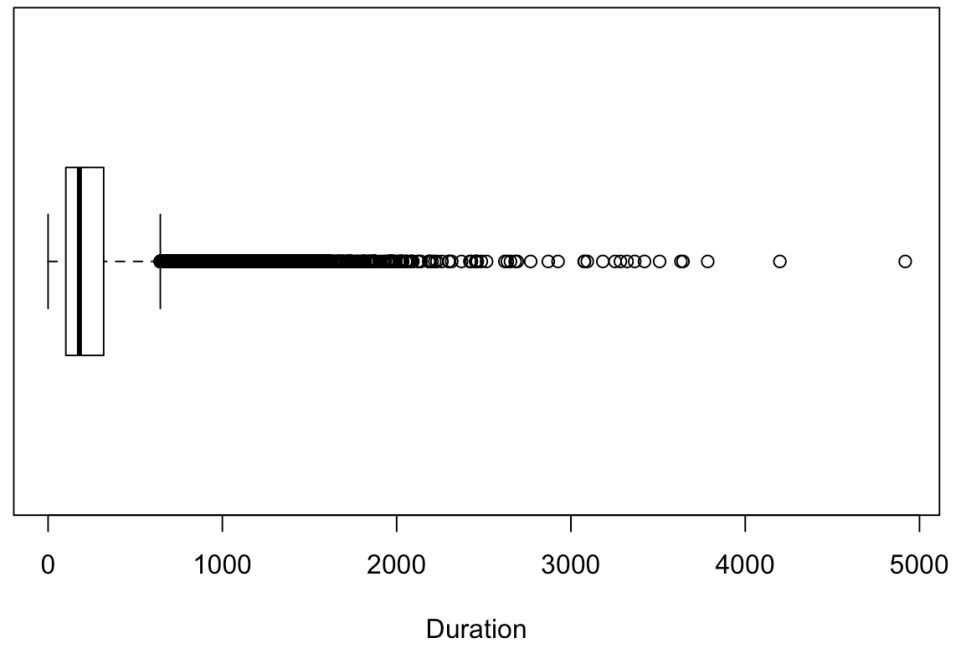
Euribor3m Box plot

euribor3m

```
boxplot(bank_marketing$duration, main="Duration Box
plot",yaxt="n", xlab="Duration", horizontal=TRUE)

#Plotting Histograms and bargraphs for different
coulmns using ggplot
library(ggplot2)
```
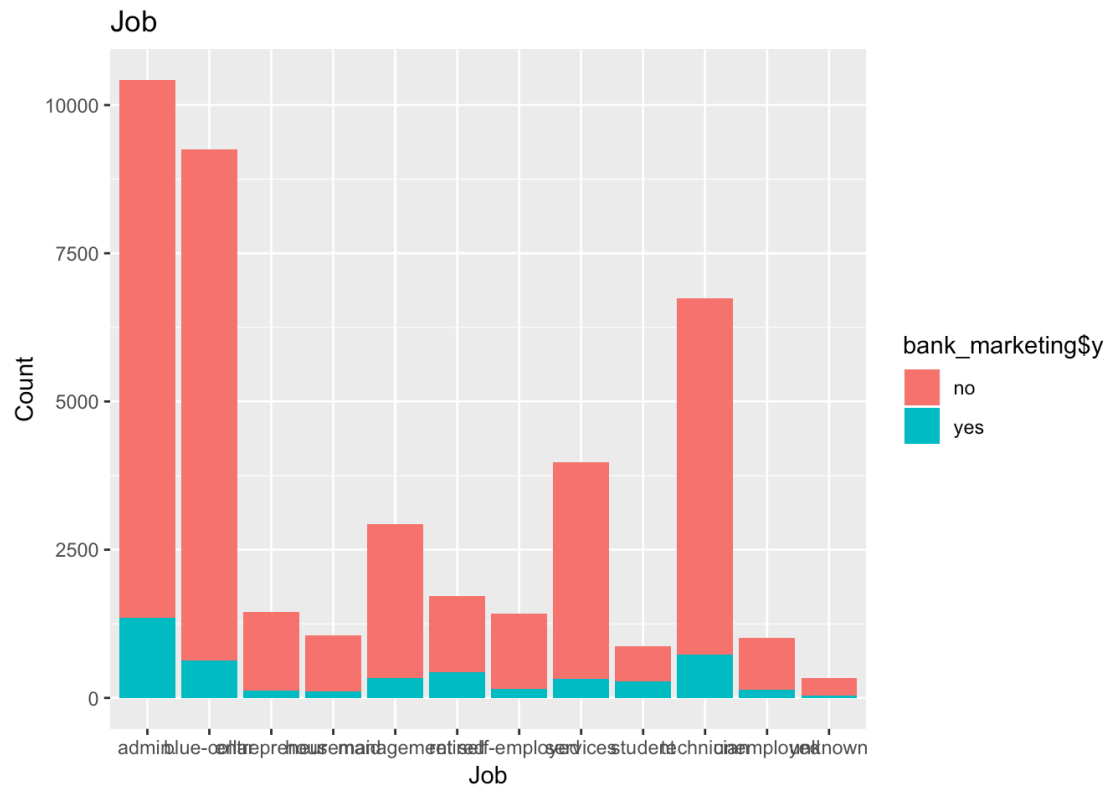
Duration Box plot

Duration

```
ggplot(bank_marketing,aes(x=bank_marketing$age,fill=ban
k_marketing$y)) + geom_histogram(binwidth=1) +
  labs(y= "Count", x="Age", title = "Age")
```

Age
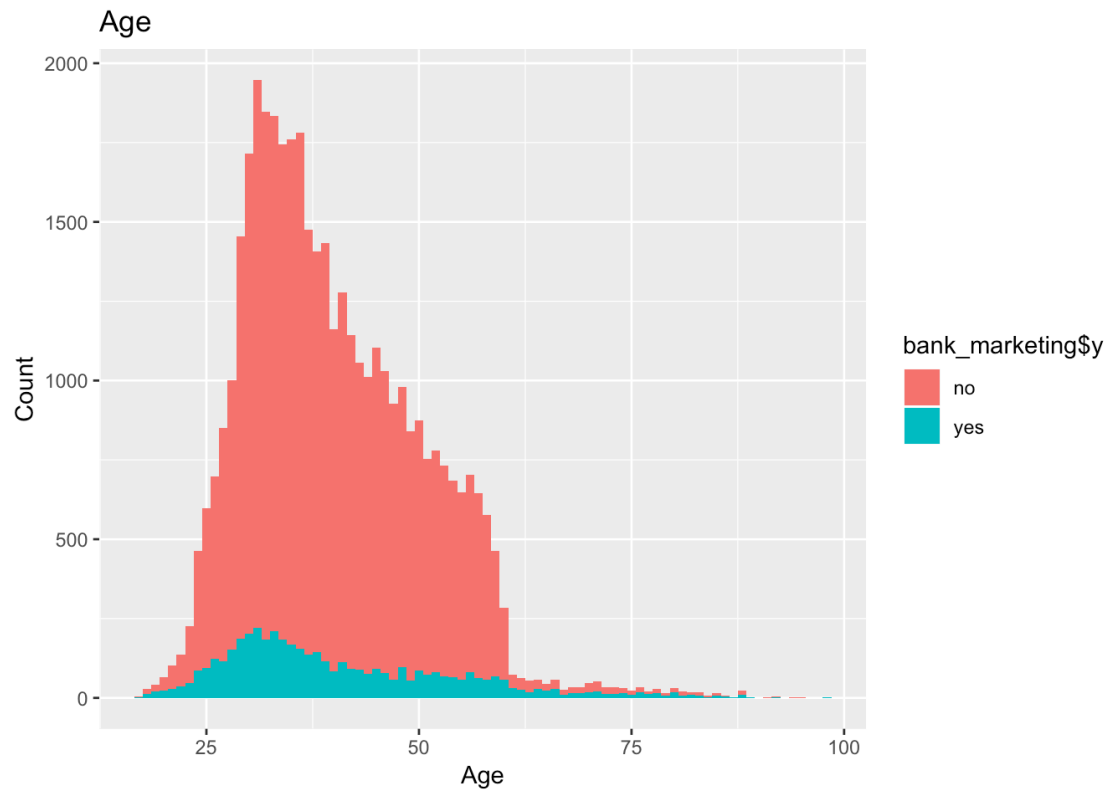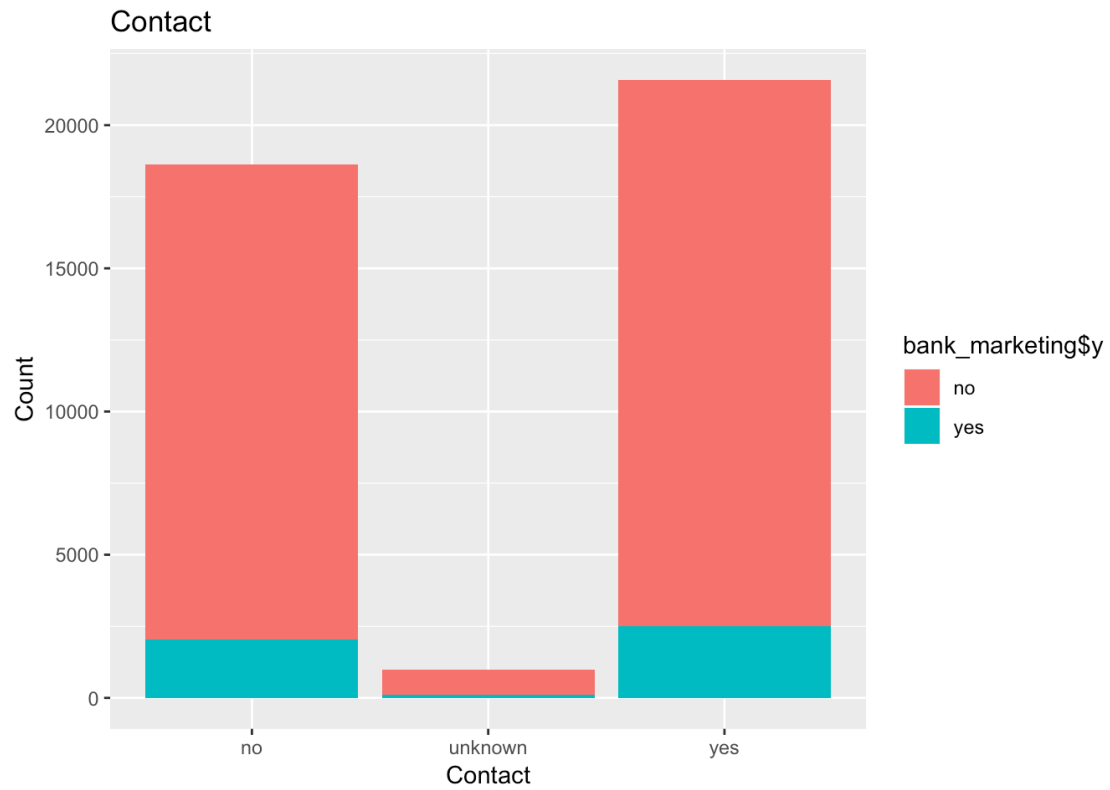


```
ggplot(bank_marketing,
aes(x=bank_marketing$job,fill=bank_marketing$y)) +
geom_bar()+
  labs(y= "Count", x="Job", title = "Job")
```

```
ggplot(bank_marketing,
aes(x=bank_marketing$marital,fill=bank_marketing$y)) +
geom_bar() +
  labs(y= "Count", x="Marital", title = "Marital")
```

Marital
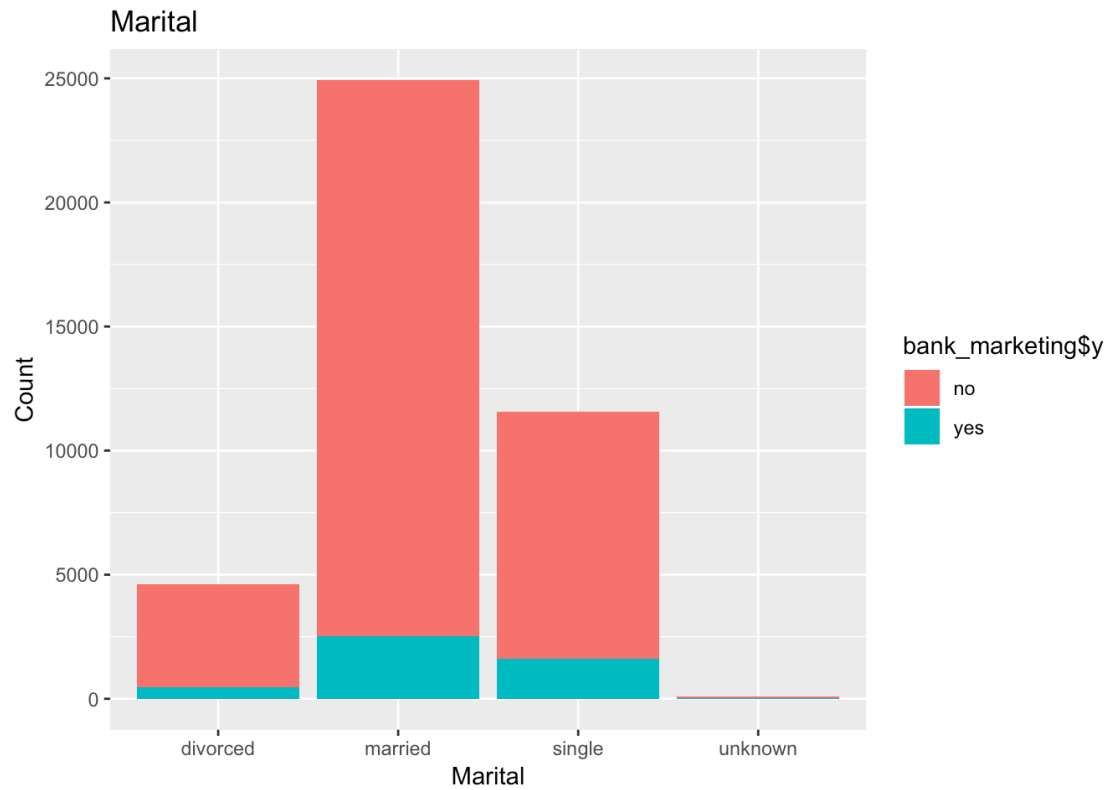

```r
ggplot(bank_marketing,
aes(x=bank_marketing$education,fill=bank_marketing$y))
+ geom_bar()+
  labs(y= "Count", x="Education", title = "Education")
```
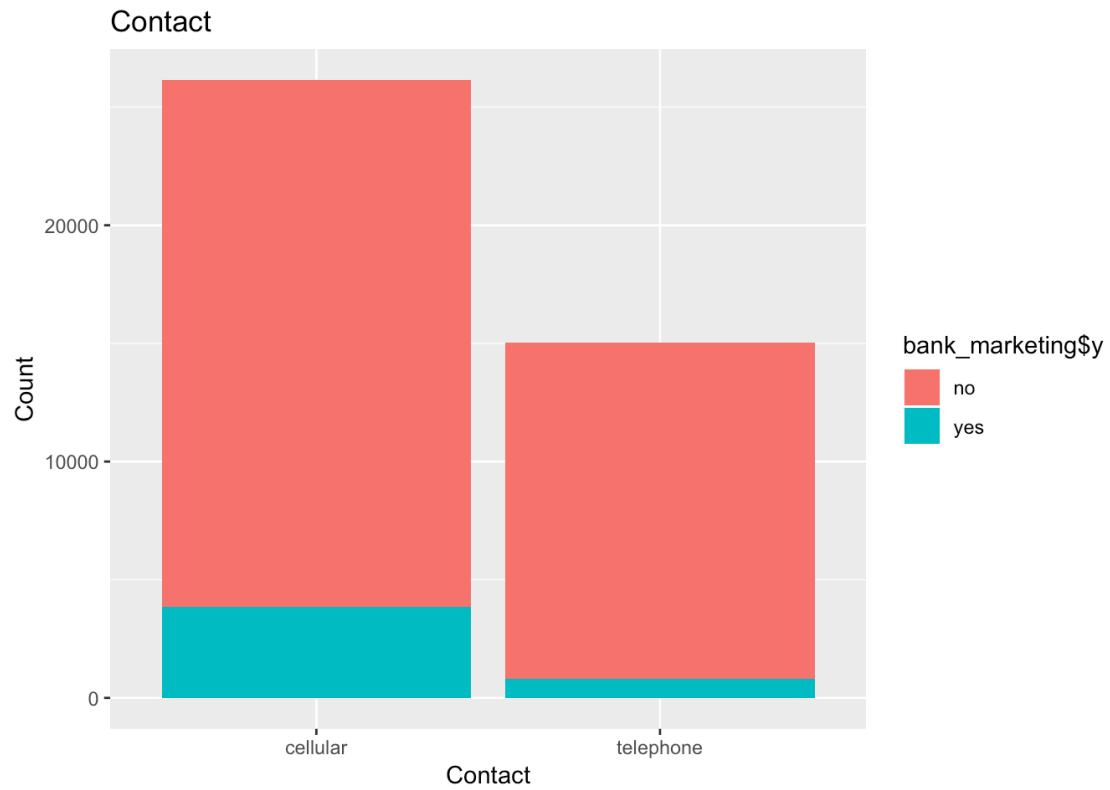
Education
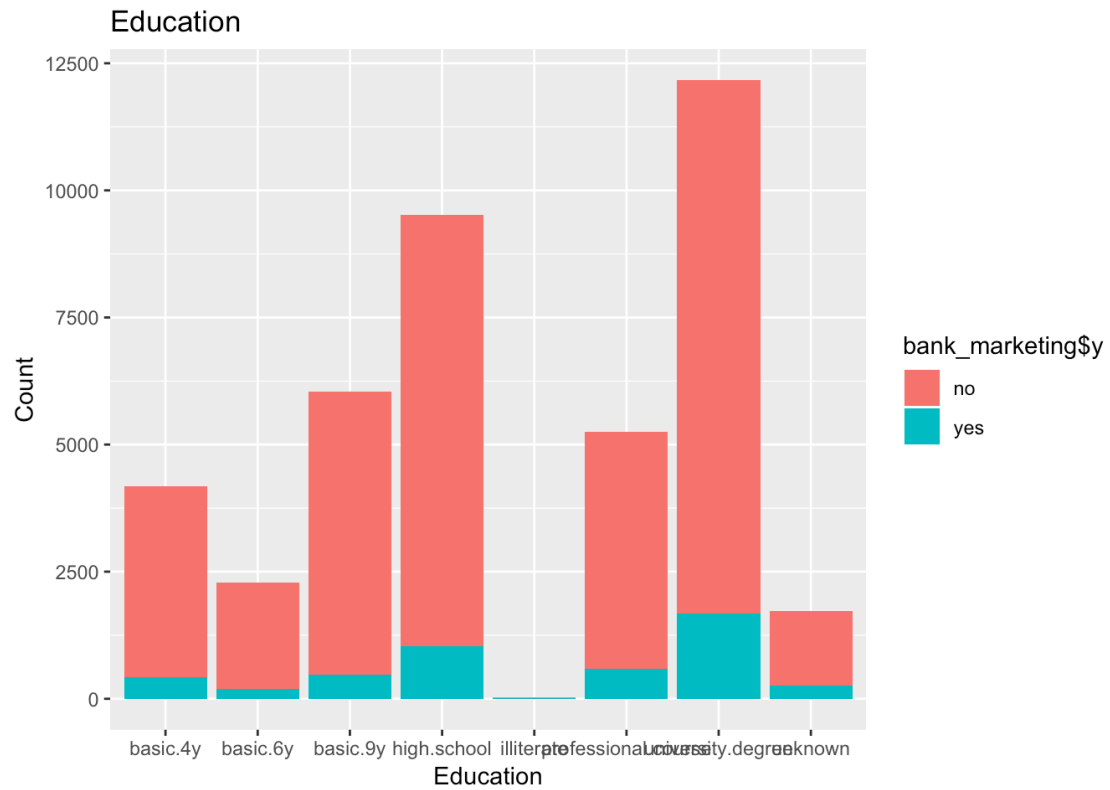

```
ggplot(bank_marketing,
aes(x=bank_marketing$default,fill=bank_marketing$y)) +
geom_bar()+
  labs(y= "Count", x="Contact", title = "Contact")
```
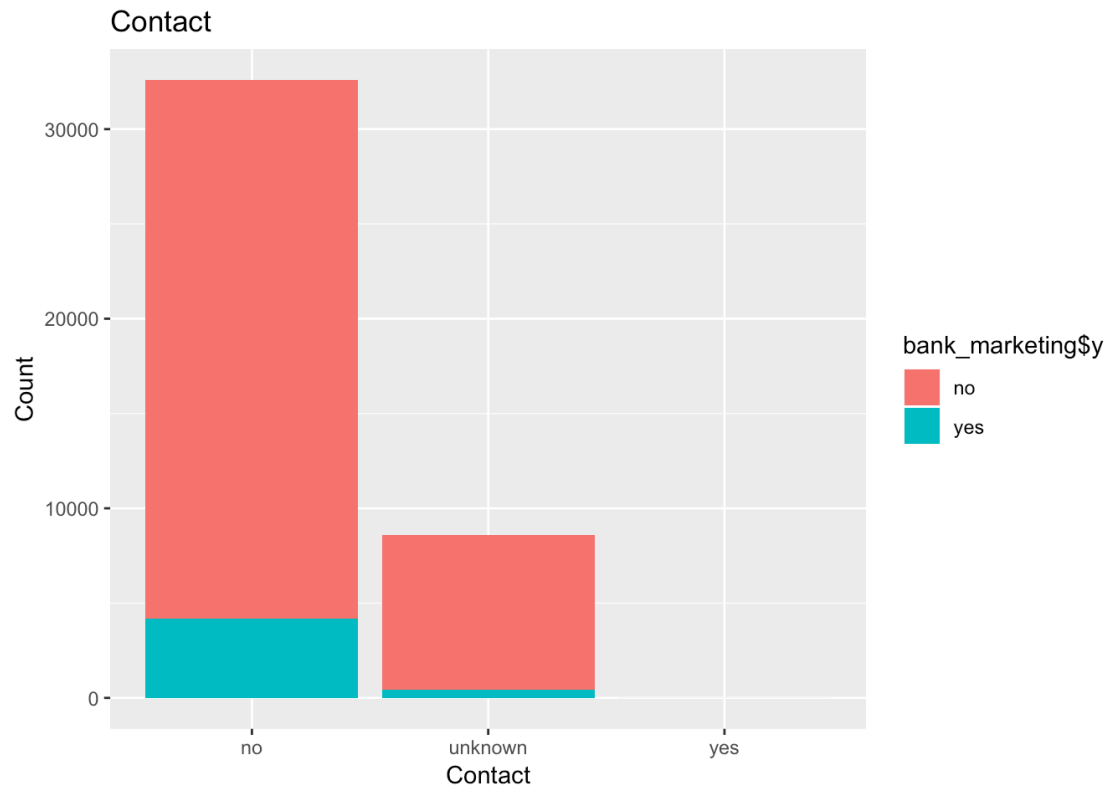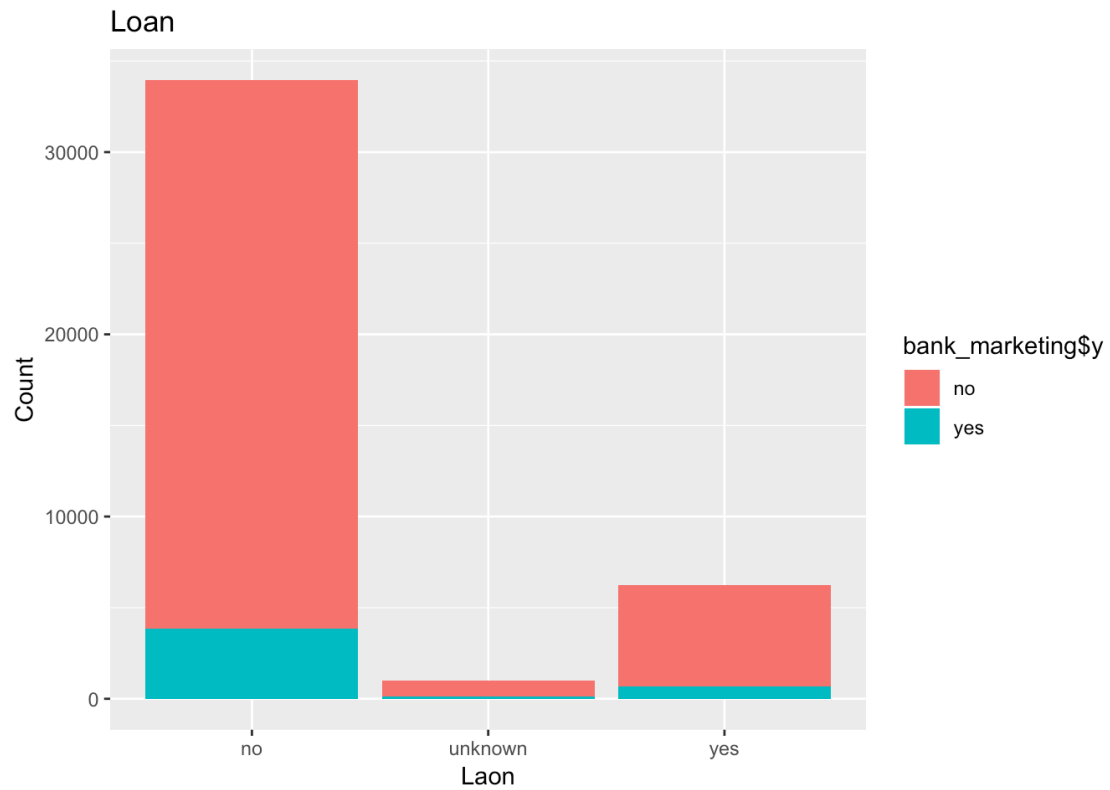
Contact


```
ggplot(bank_marketing,
aes(x=bank_marketing$housing,fill=bank_marketing$y)) +
geom_bar()+
  labs(y= "Count", x="Contact", title = "Contact")
```
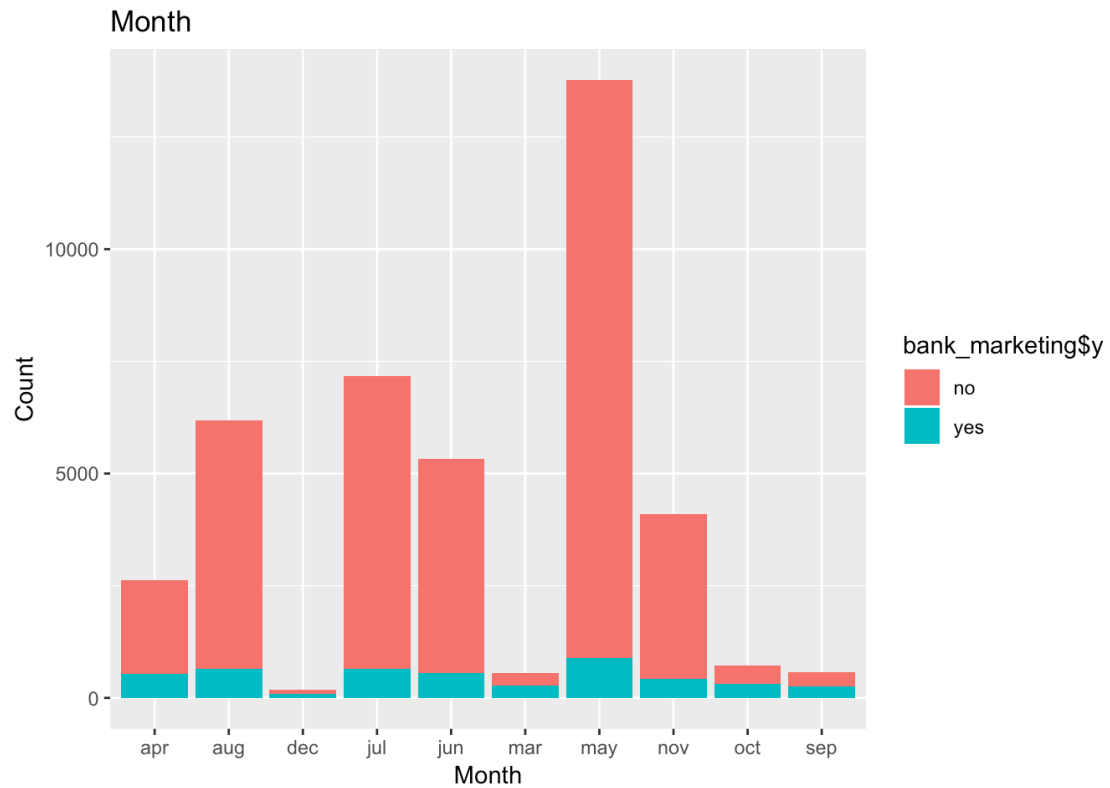
Contact


```
ggplot(bank_marketing,
aes(x=bank_marketing$contact,fill=bank_marketing$y)) +
geom_bar()+
  labs(y= "Count", x="Contact", title = "Contact")
```
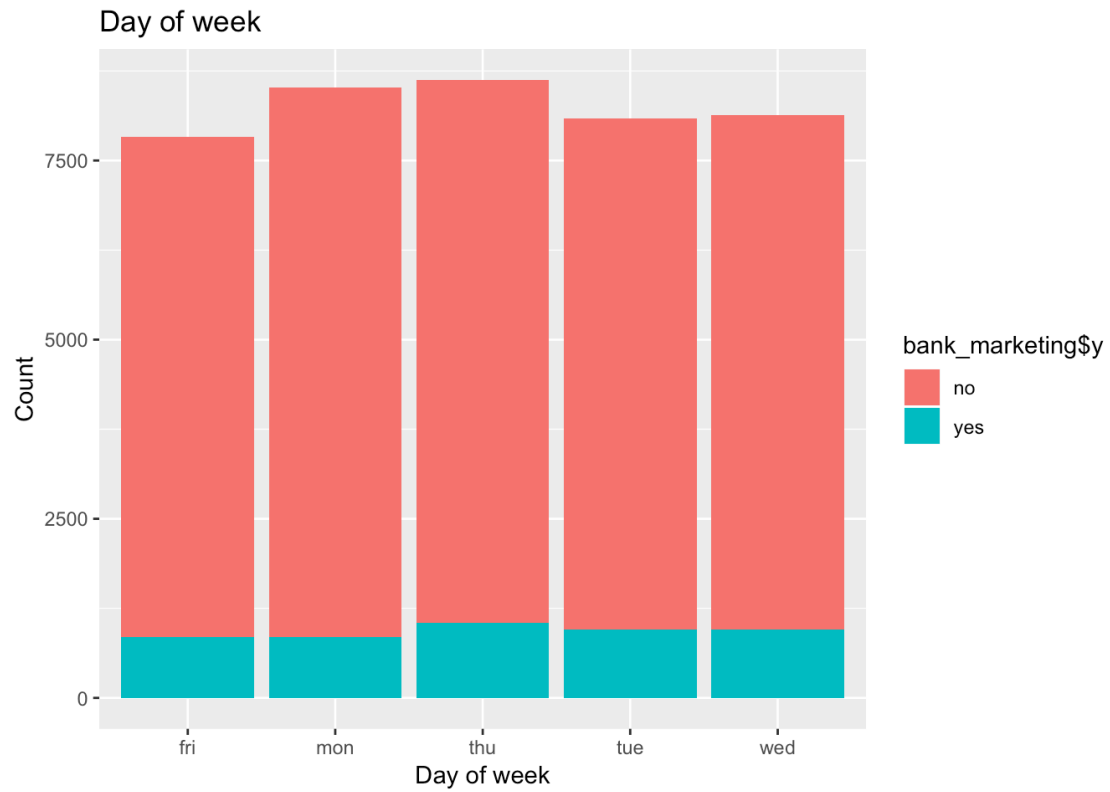
Contact


```
ggplot(bank_marketing,
aes(x=bank_marketing$loan,fill=bank_marketing$y)) +
geom_bar()+
  labs(y= "Count", x="Laon", title = "Loan")
```

Loan


```
ggplot(bank_marketing,
aes(x=bank_marketing$month,fill=bank_marketing$y)) +
geom_bar()+
  labs(y= "Count", x="Month", title = "Month")
```

Month


```
ggplot(bank_marketing,
aes(x=bank_marketing$day_of_week,fill=bank_marketing$y)
) + geom_bar()+
  labs(y= "Count", x="Day of week", title = "Day of
week")
```

Day of week


```
ggplot(bank_marketing,
aes(x=bank_marketing$poutcome,fill=bank_marketing$y)) +
geom_bar() +
  labs(y= "Count", x="poutcome", title = "Poutcome")
```

Poutcome
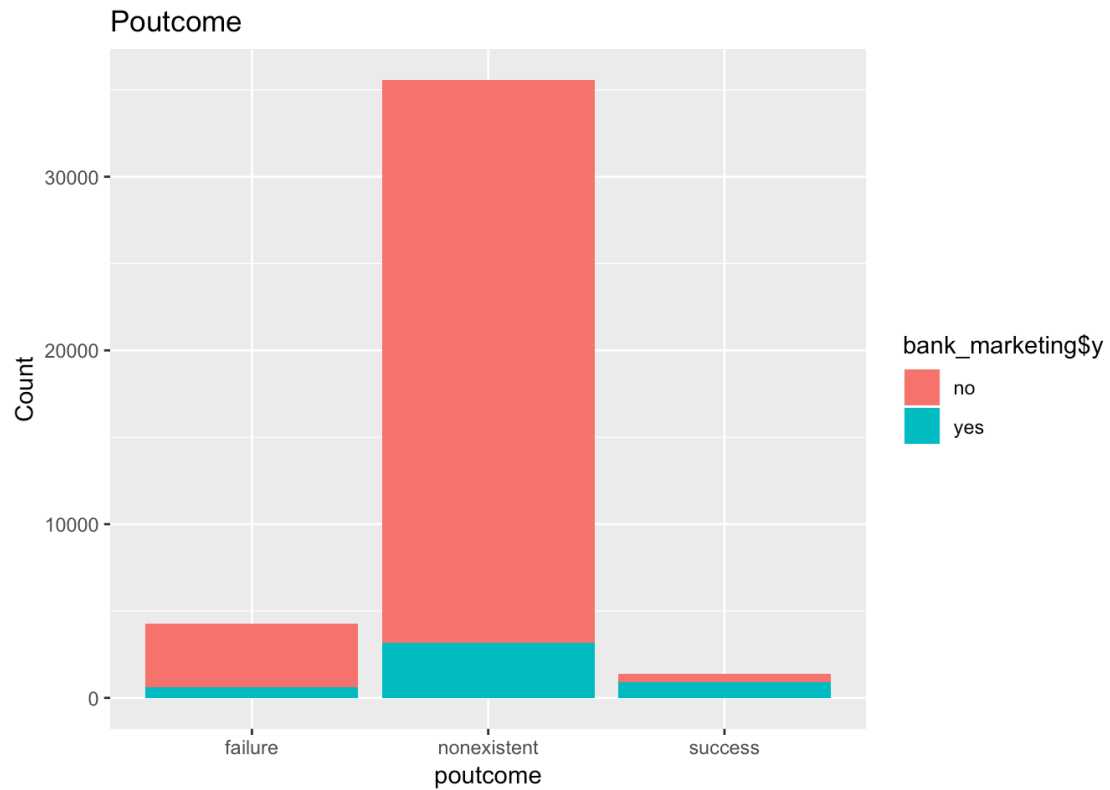

```
#Diagonal boxplot
library(SciViews)
bank_marketing_int=bank_marketing[c(10:14)]
pairs(bank_marketing_int,
labels=c("Duration","Campaign","pdays","previous"),font
.labels=2)
```
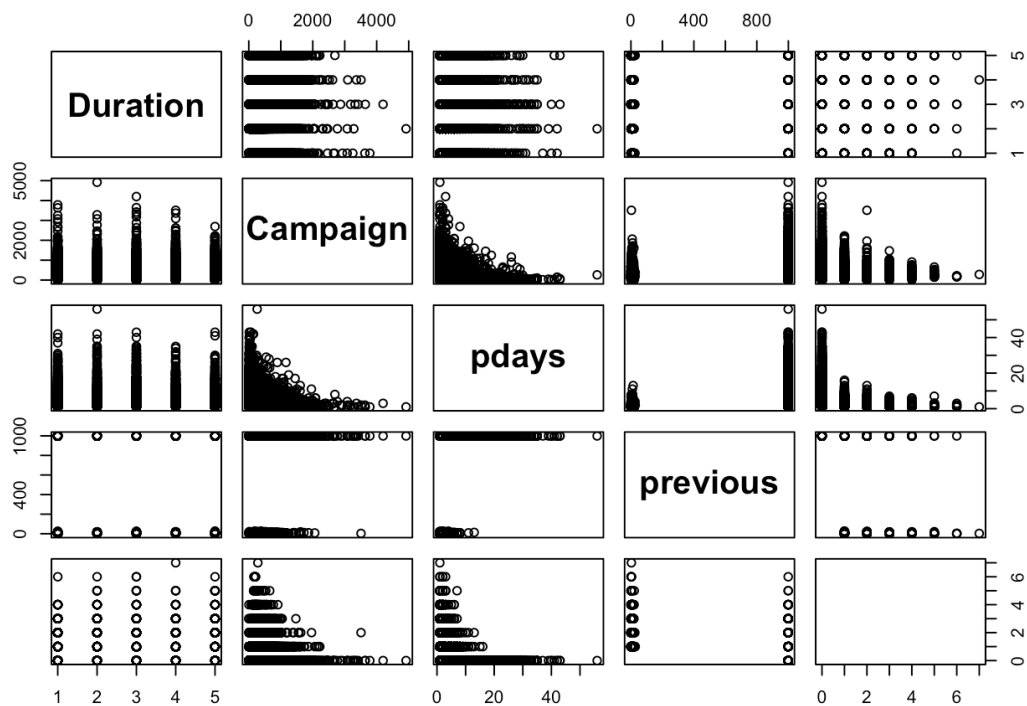
```
#3d Scatterplot plotted between age, duration and
campaign to analyse the relationship between the three
library(scatterplot3d)
s3d <-
scatterplot3d(bank_marketing$age,bank_marketing$duratio
n,bank_marketing$campaign,pch=c(1,16)
[as.numeric(bank$y)],xlab="Age", ylab="Duration",
angle=45,zlab="Campaign",
lty.hide=2,type="h",y.margin.add=0.1,font.axis=2,font.l
ab=2)
```
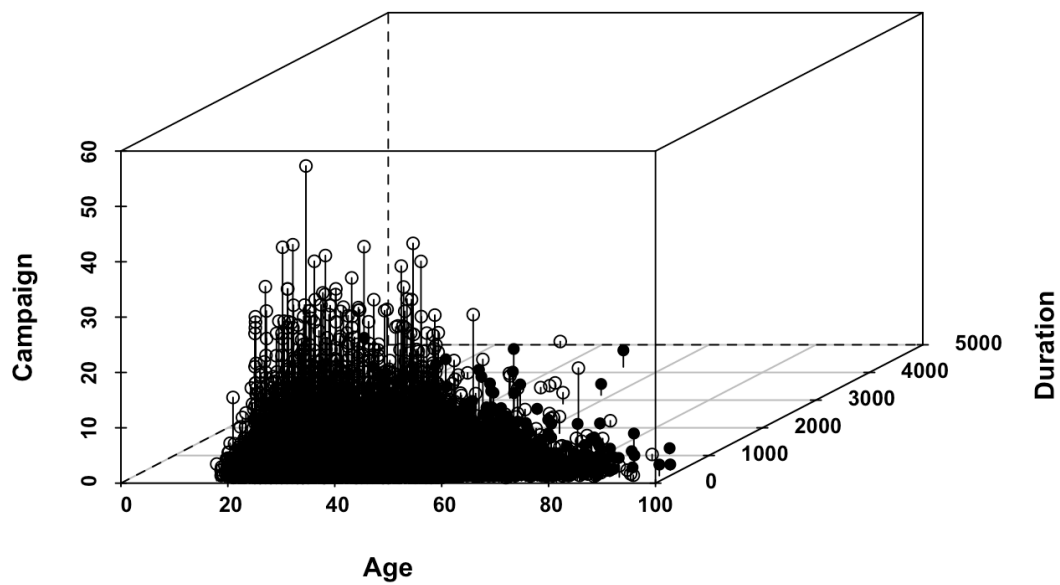
```
library(MVA)
## Loading required package: HSAUR2
## Loading required package: tools
#Chiplot plotted to analyse the relation between age
and duration
mlab = "Age of the Customer"
plab = "Duration"
with(bank_marketing, plot(age, duration, xlab = mlab,
ylab = plab, cex.lab = 0.9))
```
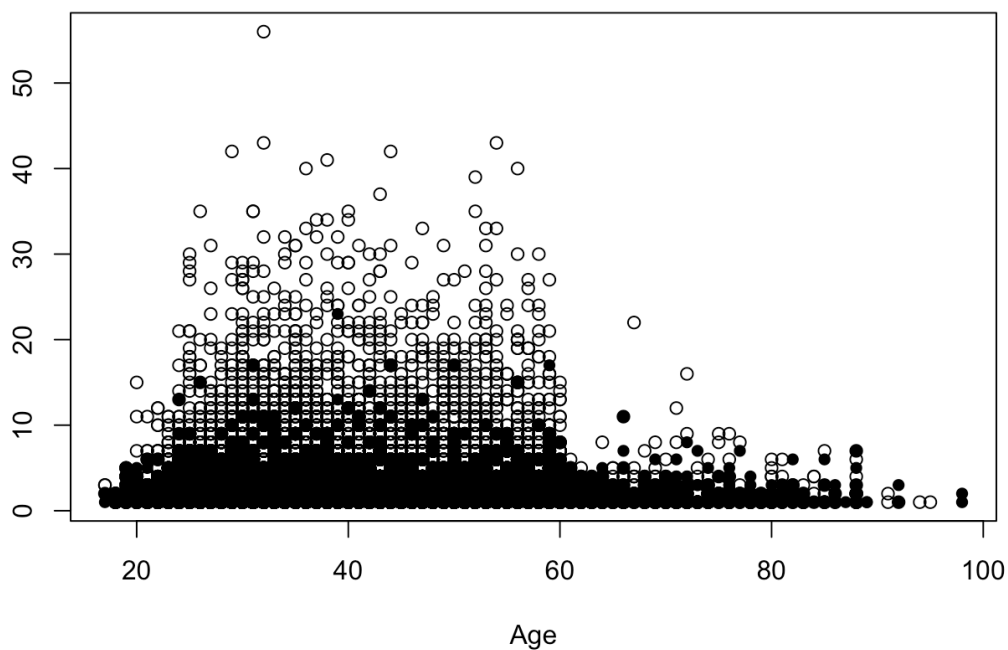
Age of the Customer

```
with(bank_marketing, chiplot(age, duration))
```


λ
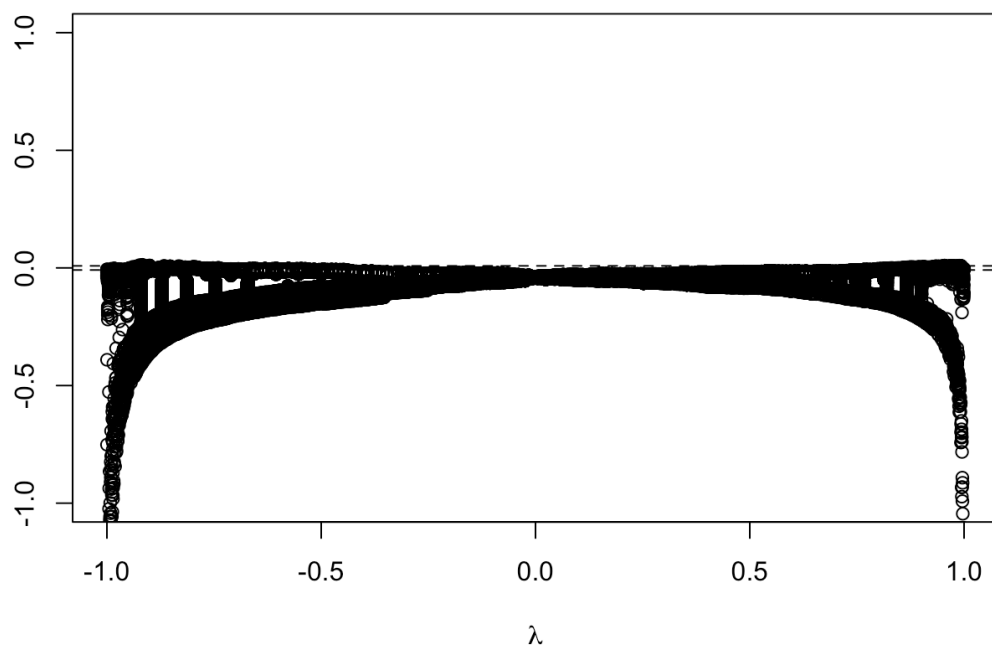
```
#Scatterplots
attach(bank_marketing)
plot(age, duration, pch=c(1,16)
[as.numeric(y)],xlab="Age",ylab="Duration")
```
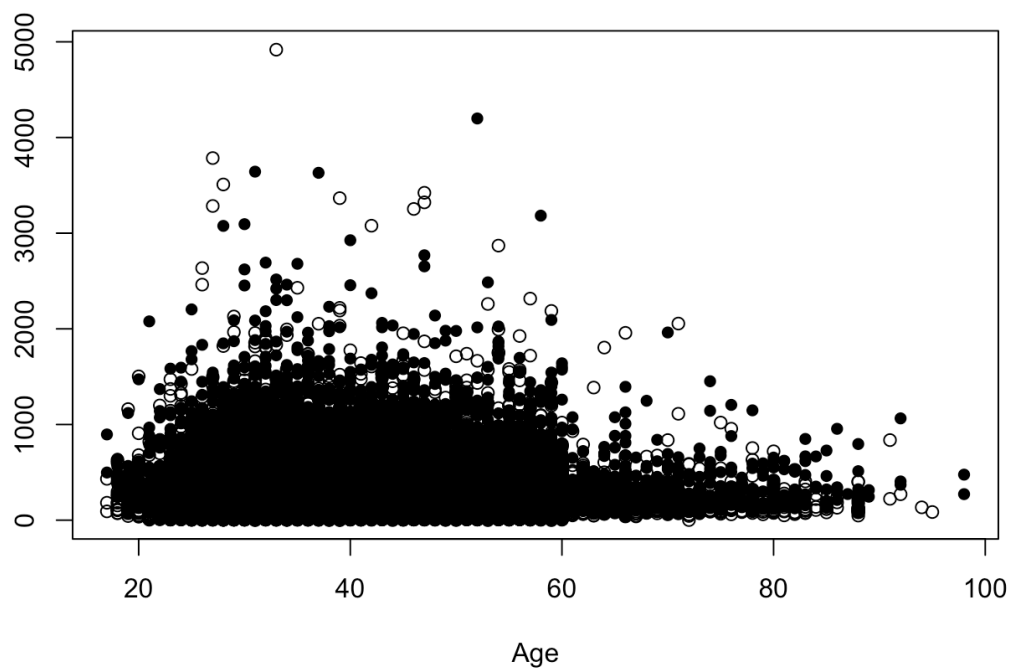


```
plot(age, campaign, pch=c(1,16)[as.numeric(y)],
xlab="Age", ylab="Campaign")
```

```
detach(bank_marketing)

#pairs(bank_marketing, panel = function (x, y, ...)
{points(x, y, ...);abline(lm(y ~ x), col = "grey")},
pch = ".", cex = 1.5)

#ScatterplotMatrix
library(car)
## Loading required package: carData
scatterplotMatrix(~age+duration+campaign+pdays+previous
| bank_marketing$y, data=bank_marketing,
var.labels=c("age","Duration","Campaign","pdays","previ
ous"),cex.labels=0.7,
diagonal="boxplot",smooth=FALSE,reg.line=FALSE,pch=c(1,
16),col=rep("black",2), legend.plot=FALSE)
```
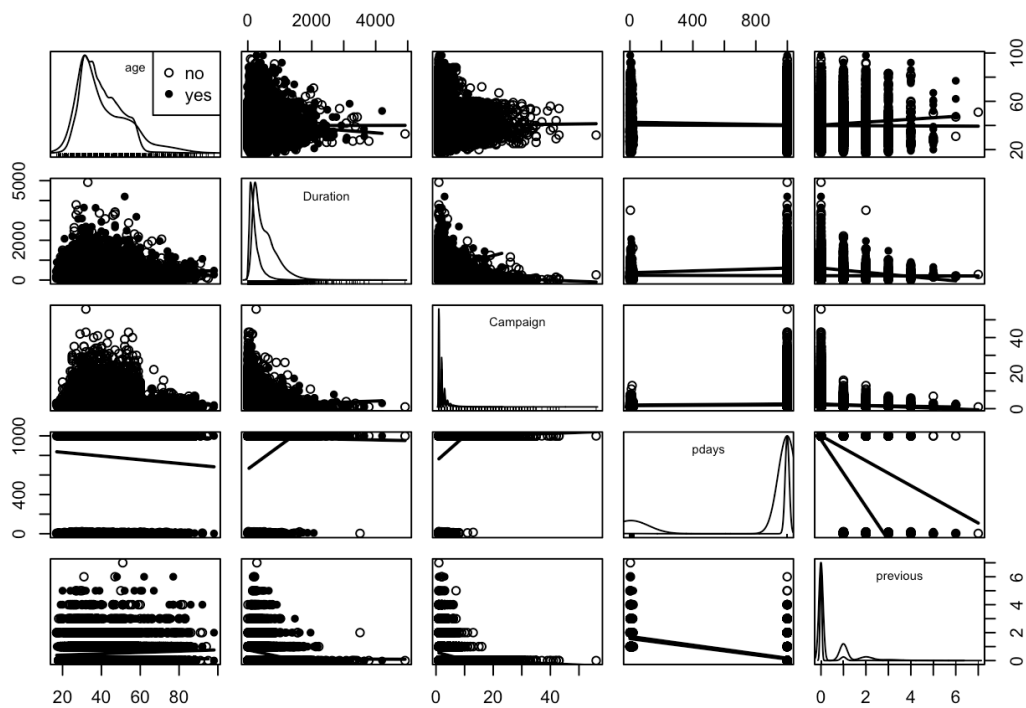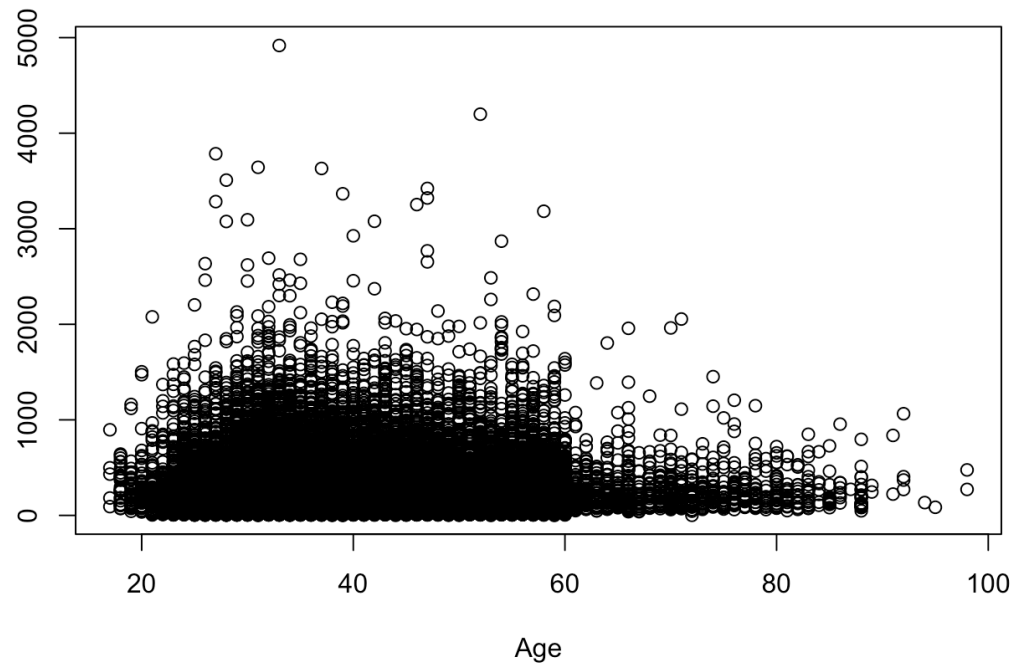
```
#bvplot
bank_marketing_age_dur=data.frame(bank_marketing$age,
bank_marketing$duration)
bvbox(bank_marketing_age_dur, mtitle = "", xlab = mlab,
ylab = plab)
```
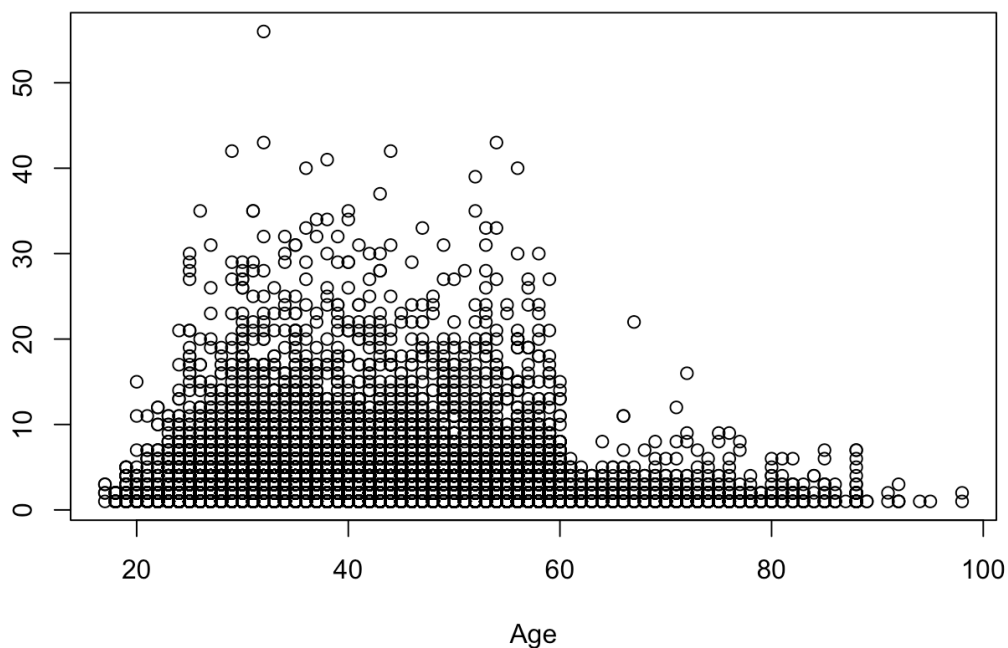
Age of the Customer

```r
y_int=ifelse(bank_marketing$y=='no', 0, 1)

plot(bank_marketing$age, bank_marketing$duration,
pch=c(1,16)[y_int],xlab="Age",ylab="Duration")
```

```
plot(bank_marketing$age, bank_marketing$campaign,
pch=c(1,16)[y_int],xlab="Age", ylab="Campaign")
```

```
#Instead of using splom using psych library it includes
splom , and give better correlation for factor features
library(psych)
## Warning: package 'psych' was built under R version
3.5.2
##
## Attaching package: 'psych'
## The following object is masked from 'package:car':
##
##      logit
## The following objects are masked from
'package:ggplot2':
##
##      %+%, alpha
pairs.panels(bank[,c(1:8,21)])
```
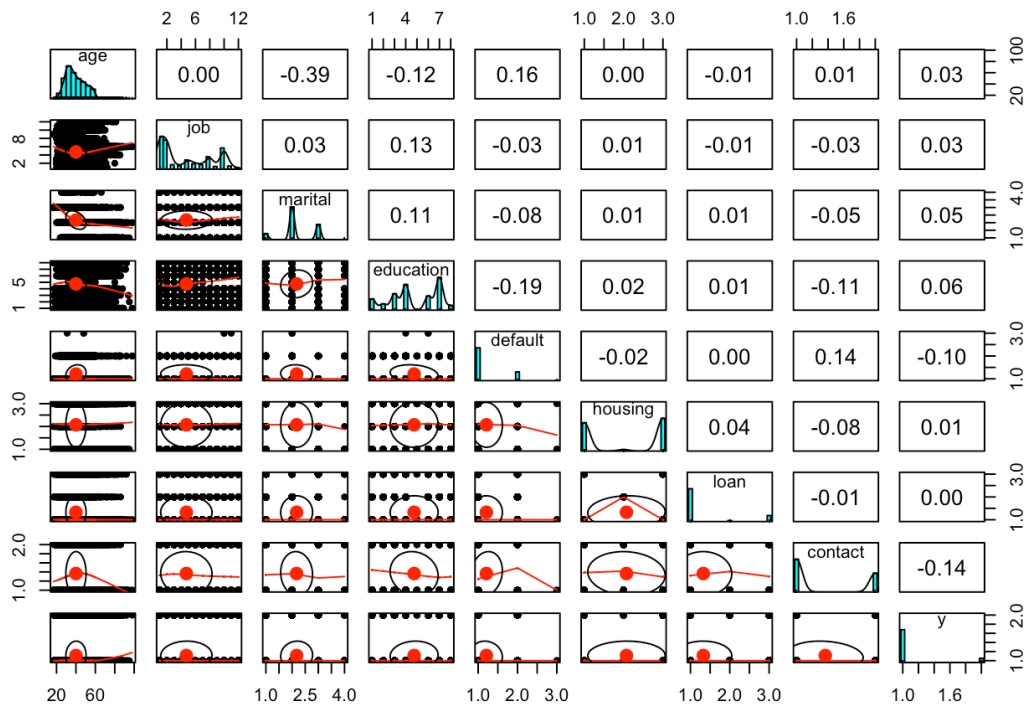
```
pairs.panels(bank[,c(9:21)])
```
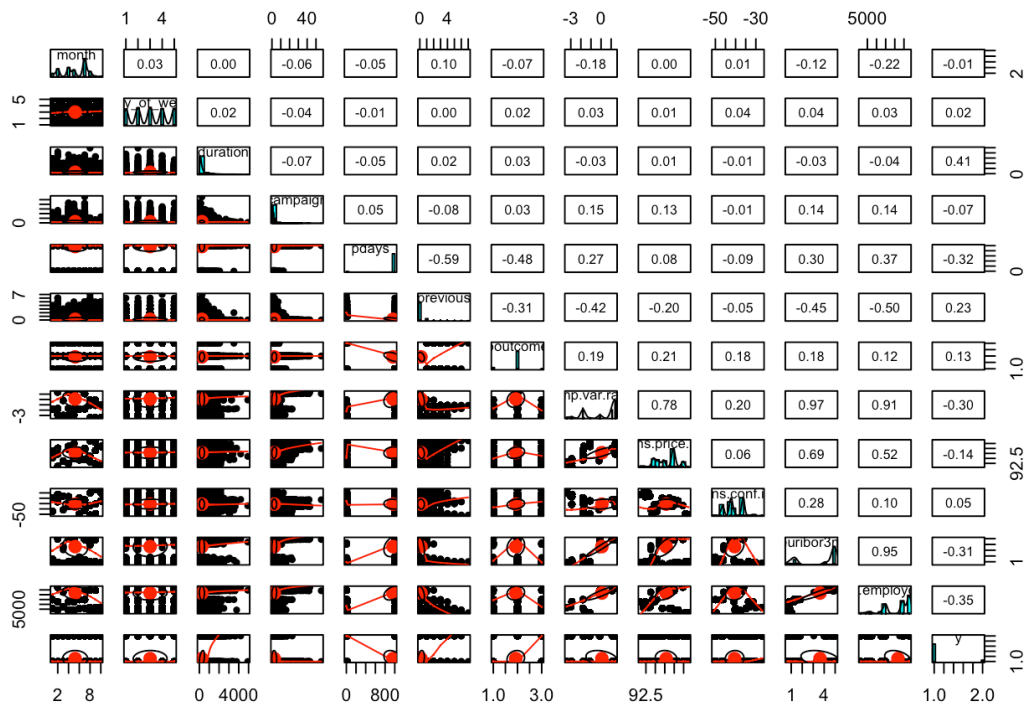
```
library(ggplot2)
qplot(bank_marketing$pdays,bank_marketing$duration,data
=bank_marketing,colour=y,size=3)
```