MBA-6005 TRAVAIL INDIVIDUEL 02

Zouheyr Ayas

Nov 15, 2020

Contents

PLAN DE TRAVAIL

PRATIQUE - PARTIE 01

Chargement de données

```
library(car)
```

```
## Loading required package: carData
```

```r
data(Salaries)
head(Salaries,10)
```

```
##         rank discipline yrs.since.phd yrs.service    sex salary
## 1       Prof          B            19          18   Male 139750
## 2       Prof          B            20          16   Male 173200
## 3   AsstProf          B             4           3   Male  79750
## 4       Prof          B            45          39   Male 115000
## 5       Prof          B            40          41   Male 141500
## 6  AssocProf          B             6           6   Male  97000
## 7       Prof          B            30          23   Male 175000
## 8       Prof          B            45          45   Male 147765
## 9       Prof          B            21          20   Male 119250
## 10      Prof          B            18          18 Female 129000
```

Exploration de la structure de données

```r
str(Salaries)
```

```
## 'data.frame':    397 obs. of  6 variables:
##  $ rank         : Factor w/ 3 levels "AsstProf","AssocProf",..: 3 3 1 3 3 2 3 3 3 3 ...
##  $ discipline   : Factor w/ 2 levels "A","B": 2 2 2 2 2 2 2 2 2 2 ...
##  $ yrs.since.phd: int  19 20 4 45 40 6 30 45 21 18 ...
##  $ yrs.service  : int  18 16 3 39 41 6 23 45 20 18 ...
##  $ sex          : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 1 ...
##  $ salary       : int  139750 173200 79750 115000 141500 97000 175000 147765 119250 129000 ...
```

QUESTION 01: Nombre d'hommes et de femmes par grade

```r
table(Salaries$rank,Salaries$sex)
```

```
##
##             Female Male
##   AsstProf      11   56
##   AssocProf     10   54
##   Prof          18  248
```

Proportions d'hommes et de femmes par grade:

```r
prop.table(table(Salaries$rank,Salaries$sex))
```

Proportion globale:

```
##
##                 Female       Male
##   AsstProf  0.02770781 0.14105793
##   AssocProf 0.02518892 0.13602015
##   Prof      0.04534005 0.62468514
```

```r
prop.table(table(Salaries$rank,Salaries$sex),1)
```

Proportion par ligne:

```
##
##                 Female       Male
##   AsstProf  0.16417910 0.83582090
##   AssocProf 0.15625000 0.84375000
##   Prof      0.06766917 0.93233083
```

```r
prop.table(table(Salaries$rank,Salaries$sex),2)
```

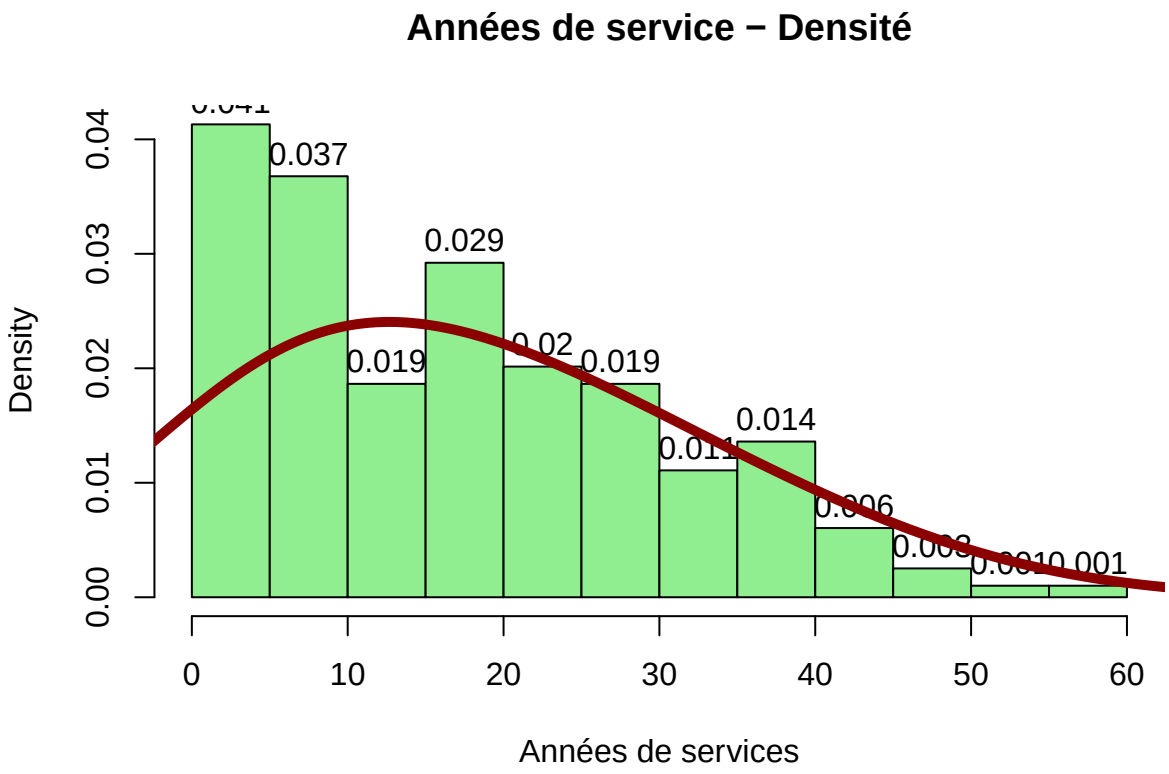
Proportion par colonne:

```
##
##                Female      Male
##   AsstProf  0.2820513 0.1564246
##   AssocProf 0.2564103 0.1508380
##   Prof      0.4615385 0.6927374
```

QUESTION 02 : Histogramme pour les années de service

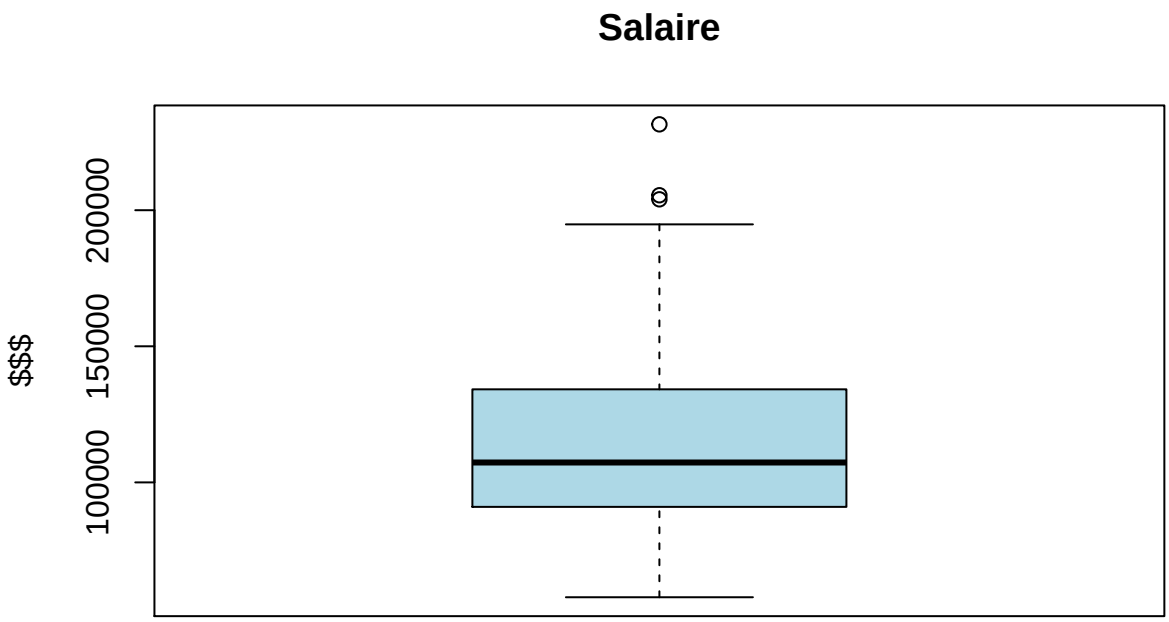
```r
hist(Salaries$yrs.service, freq = FALSE,
     main="Années de service - Densité",col="lightgreen",
     xlab = "Années de services",probability = TRUE,
     labels = TRUE)

#Dessiner la ligne de densité
lines(density(Salaries$yrs.service, bw=10),type="l",col="darkred",lwd=5)
```



QUESTION 03 : Dessiner un box plot pour le salaire

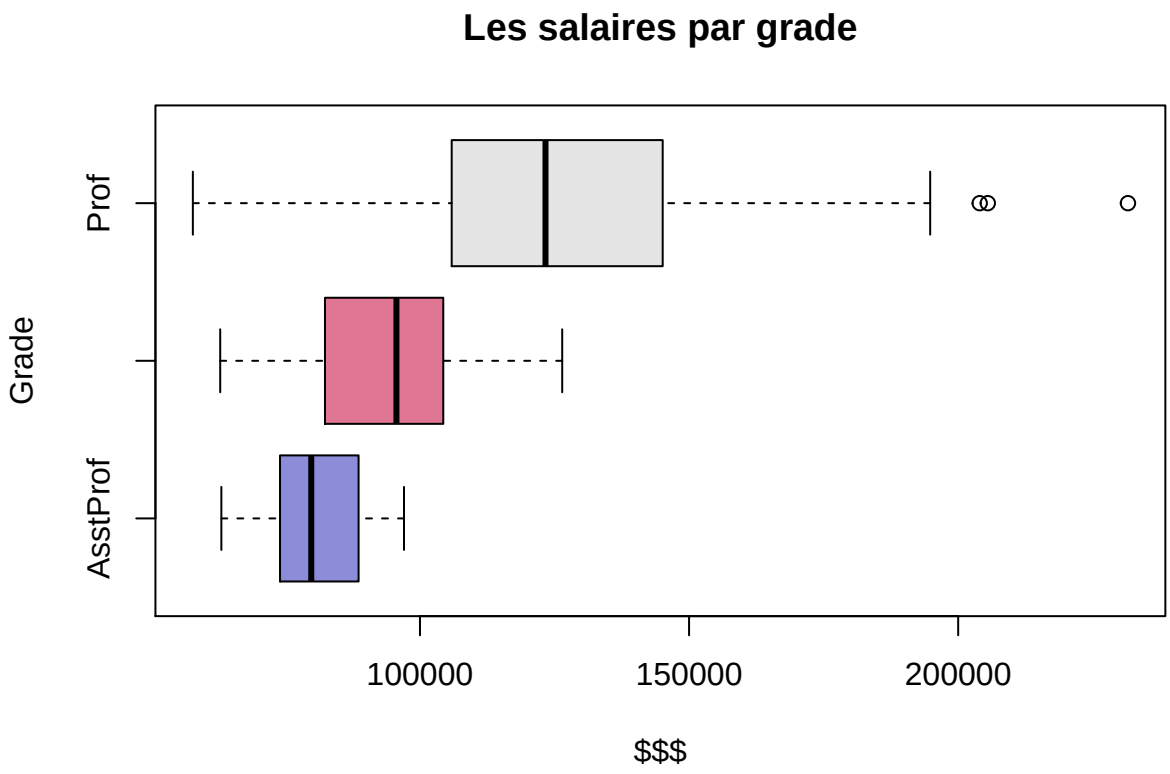
```
boxplot(Salaries$salary, col="lightblue",main="Salaire",ylab="$$$")
```



QUESTION 04: Dessiner un box plot pour le salaire par grade

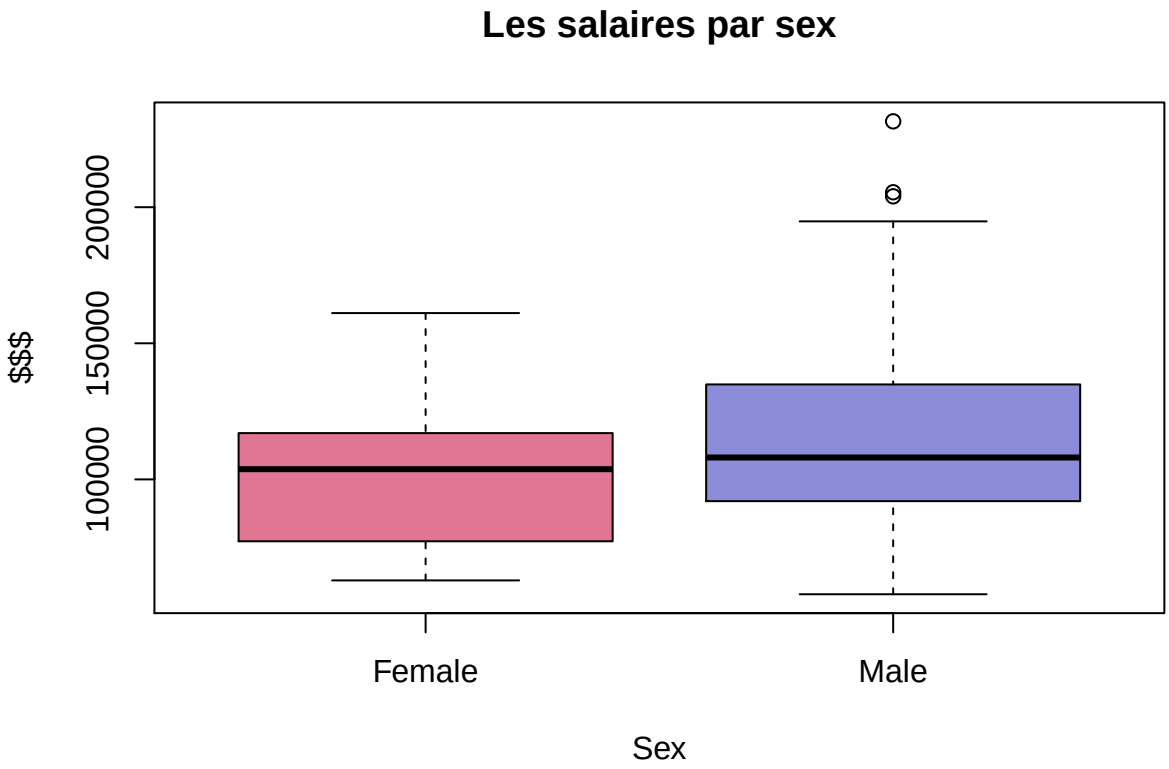
```r
#Spécification de colors
colors <- ifelse(levels(Salaries$rank)=="AsstProf" , rgb(0.1,0.1,0.7,0.5) ,
              ifelse(levels(Salaries$rank)=="AssocProf", rgb(0.8,0.1,0.3,0.6),
              "grey90"))


#Lancement de plot
boxplot(Salaries$salary  ~  (Salaries$rank), col=colors,
        main="Les salaires par grade",xlab = "$$$", ylab = "Grade",
        horizontal = TRUE)
```



Plot de salaires par Sex:

```
#spécification des colors
colors <- ifelse(levels(Salaries$sex)=="Male" , rgb(0.1,0.1,0.7,0.5) ,
              ifelse(levels(Salaries$sex)=="Female", rgb(0.8,0.1,0.3,0.6),
                     "grey90" ) )
#Plotting
boxplot(Salaries$salary  ~  (Salaries$sex), col=colors,
        main="Les salaires par sex",xlab = "Sex", ylab = "$$$")
```



EXPREMENTATIONS:

3.2.RELATION ENTRE LES VARIABLES

3.2.1.Chargement de données de CRM

```
cust.df<-read.csv("http://goo.gl/PmPkaG")
str(cust.df)
```

```
## 'data.frame':    1000 obs. of  12 variables:
##  $ cust.id          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ age              : num  22.9 28 35.9 30.5 38.7 ...
##  $ credit.score     : num  631 749 733 830 734 ...
##  $ email            : chr  "yes" "yes" "yes" "yes" ...
##  $ distance.to.store: num  2.58 48.18 1.29 5.25 25.04 ...
##  $ online.visits    : int  20 121 39 1 35 1 1 48 0 14 ...
##  $ online.trans     : int  3 39 14 0 11 1 1 13 0 6 ...
##  $ online.spend     : num  58.4 756.9 250.3 0 204.7 ...
##  $ store.trans      : int  4 0 0 2 0 0 2 4 0 3 ...
##  $ store.spend      : num  140.3 0 0 95.9 0 ...
##  $ sat.service      : int  3 3 NA 4 1 NA 3 2 4 3 ...
##  $ sat.selection    : int  3 3 NA 2 1 NA 3 3 2 2 ...
```

3.2.2.Converting data to factors

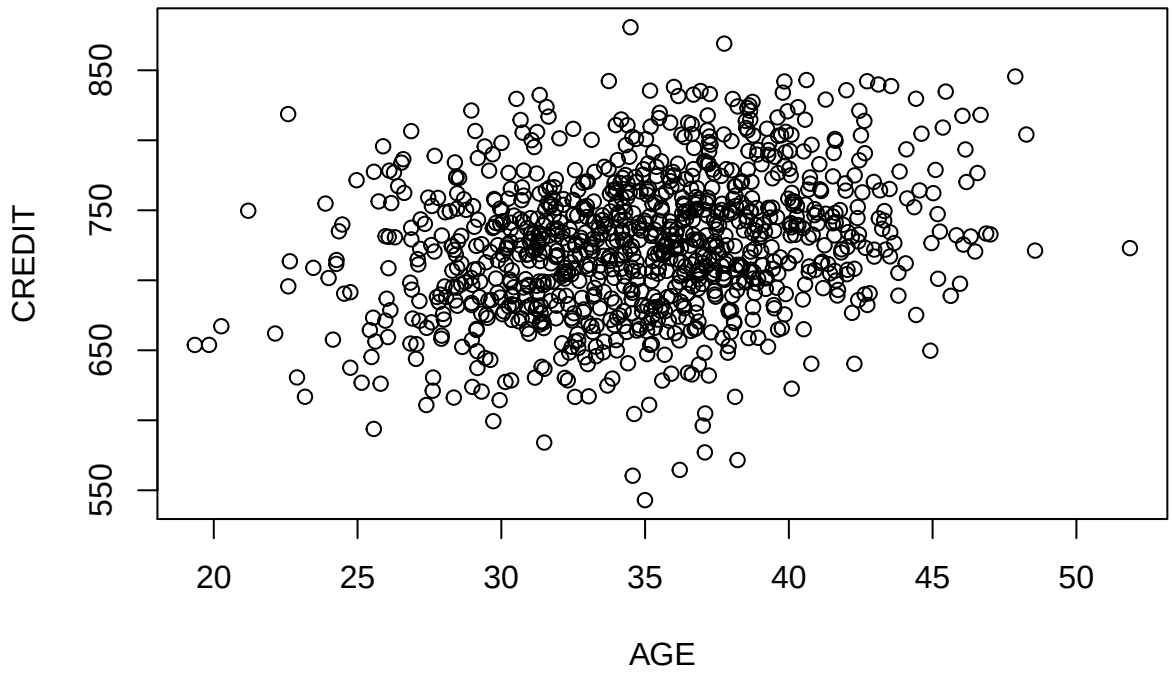
```
str(cust.df$cust.id)
```

```
##  int [1:1000] 1 2 3 4 5 6 7 8 9 10 ...
```

```
cust.df$cust.id<-factor(cust.df$cust.id)
str(cust.df$cust.id)
```

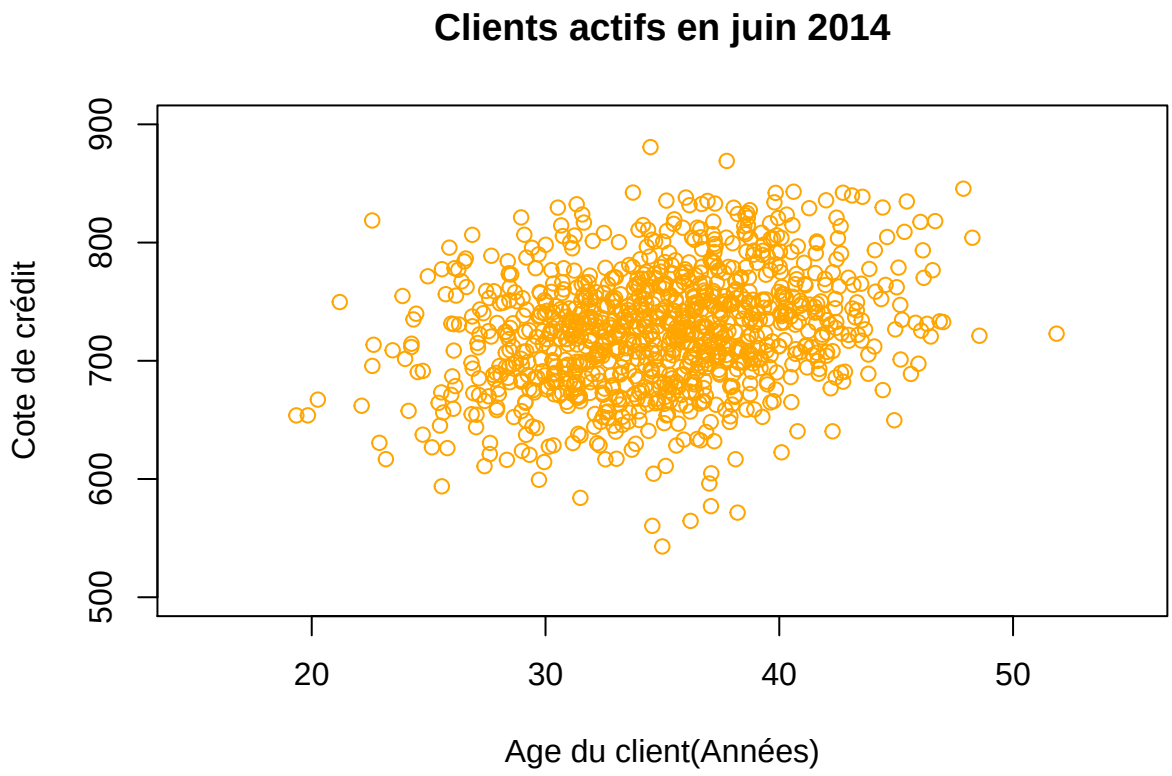
```
##  Factor w/ 1000 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
```

```
plot(x=cust.df$age,y=cust.df$credit.score,xlab = "AGE", ylab = "CREDIT")
```



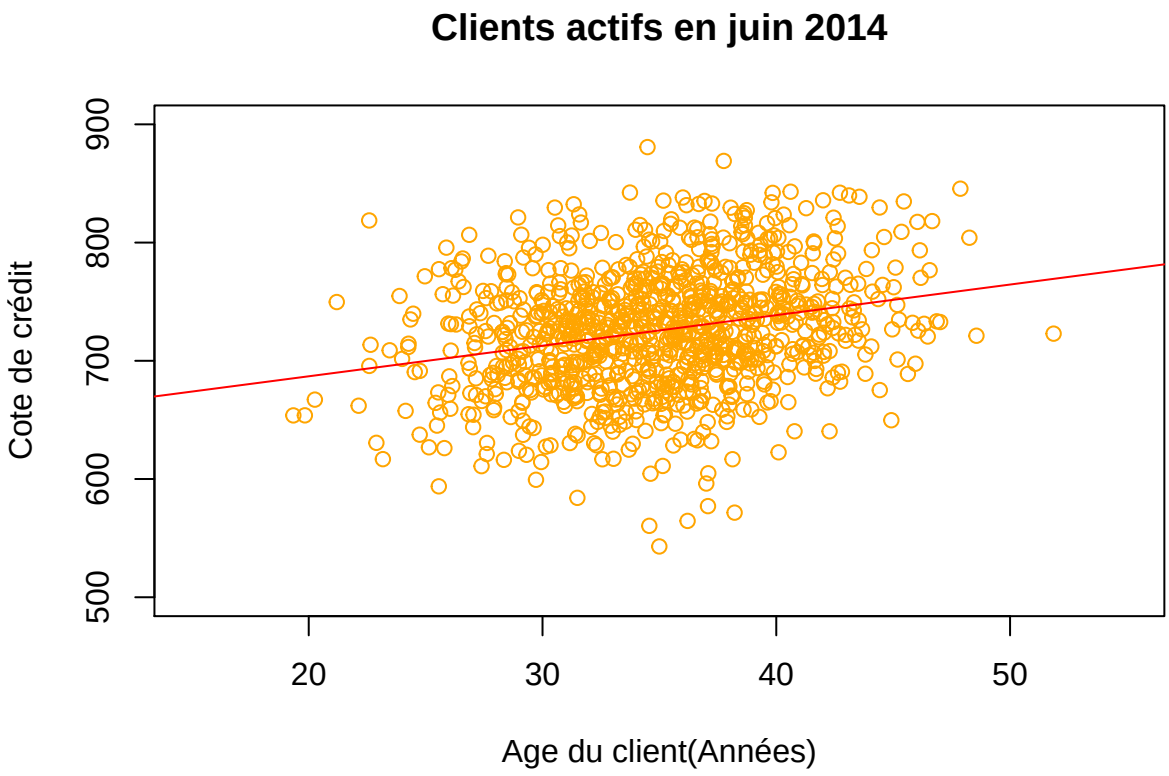
3.2.3. Add color, labels, and adjust axis limits

```
plot(cust.df$age, cust.df$credit.score,col="orange",xlim=c(15,55),
     ylim = c(500,900),main="Clients actifs en juin 2014",
     xlab = "Age du client(Années)",ylab = "Cote de crédit")
```



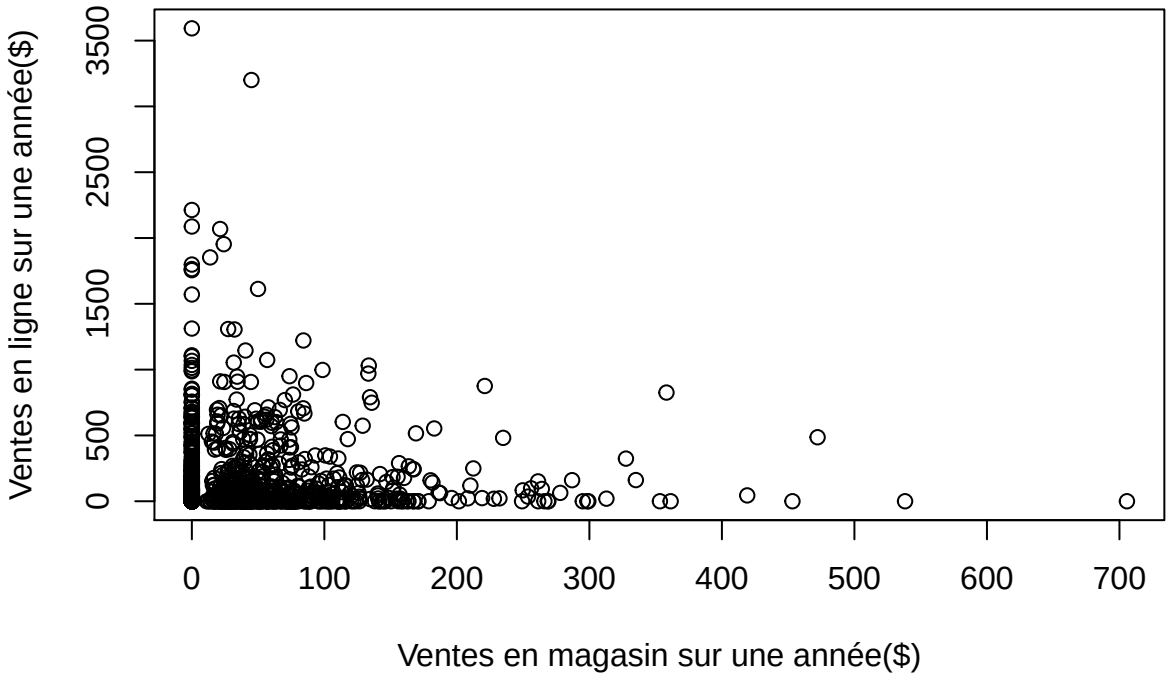
3.2.4. Regression

```
plot(cust.df$age, cust.df$credit.score,col="orange",xlim=c(15,55),
     ylim = c(500,900),main="Clients actifs en juin 2014",
     xlab = "Age du client(Années)",ylab = "Cote de crédit")
abline(lm(cust.df$credit.score ~ cust.df$age),col="red")
```



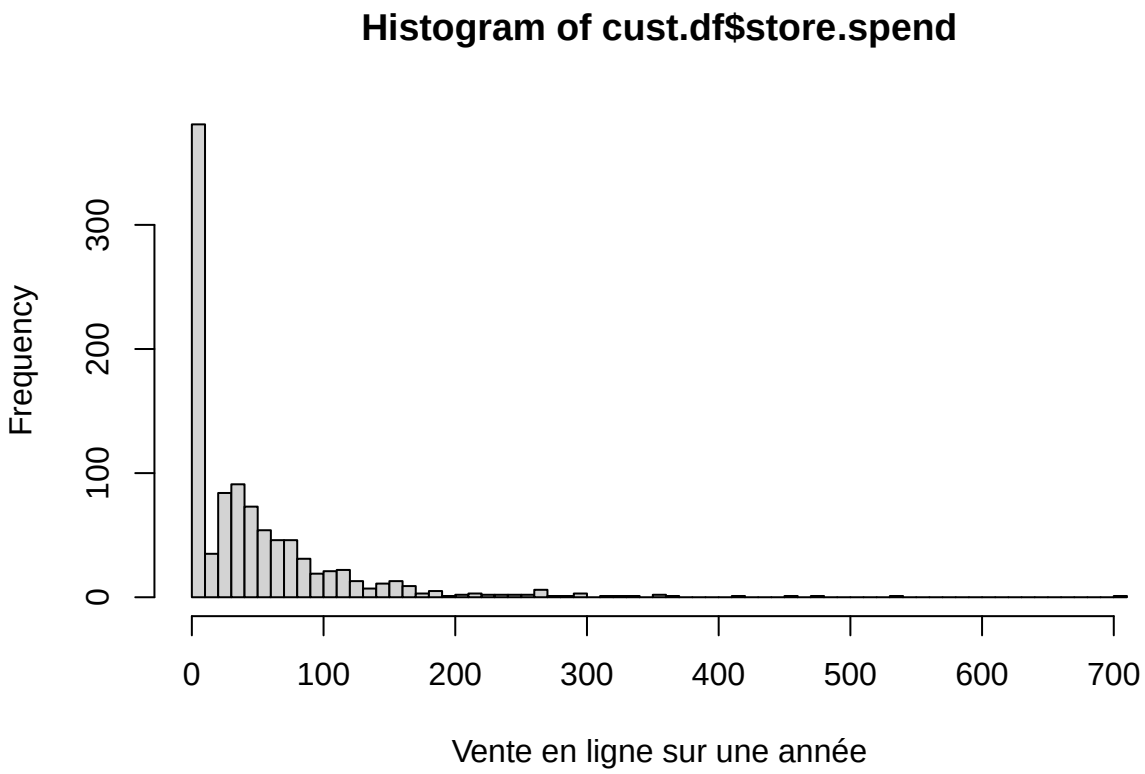
3.2.5. Ventes en ligne vs ventes en magasin

```
plot(cust.df$store.spend,cust.df$online.spend,
     xlab = "Ventes en magasin sur une année($)",
     ylab = "Ventes en ligne sur une année($)")
```



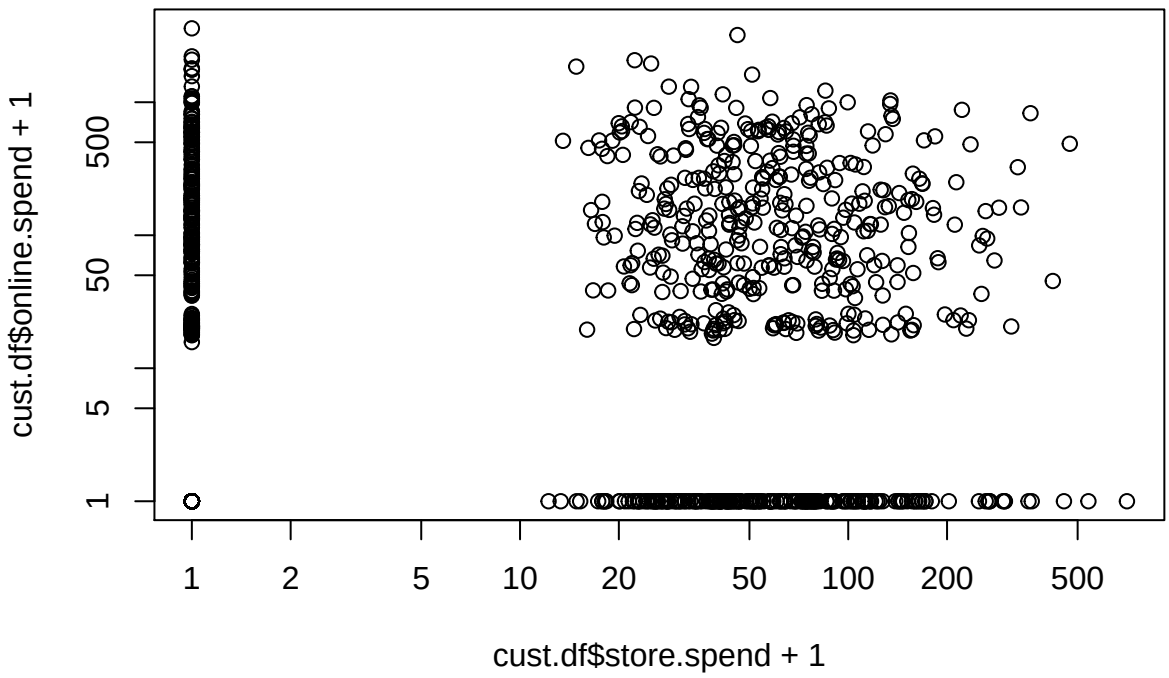
3.2.6.Histogramme des depenses en magasin:

```
hist(cust.df$store.spend,
     breaks=(0:ceiling(max(cust.df$store.spend)/10))*10,
     xlab = "Vente en ligne sur une année")
```



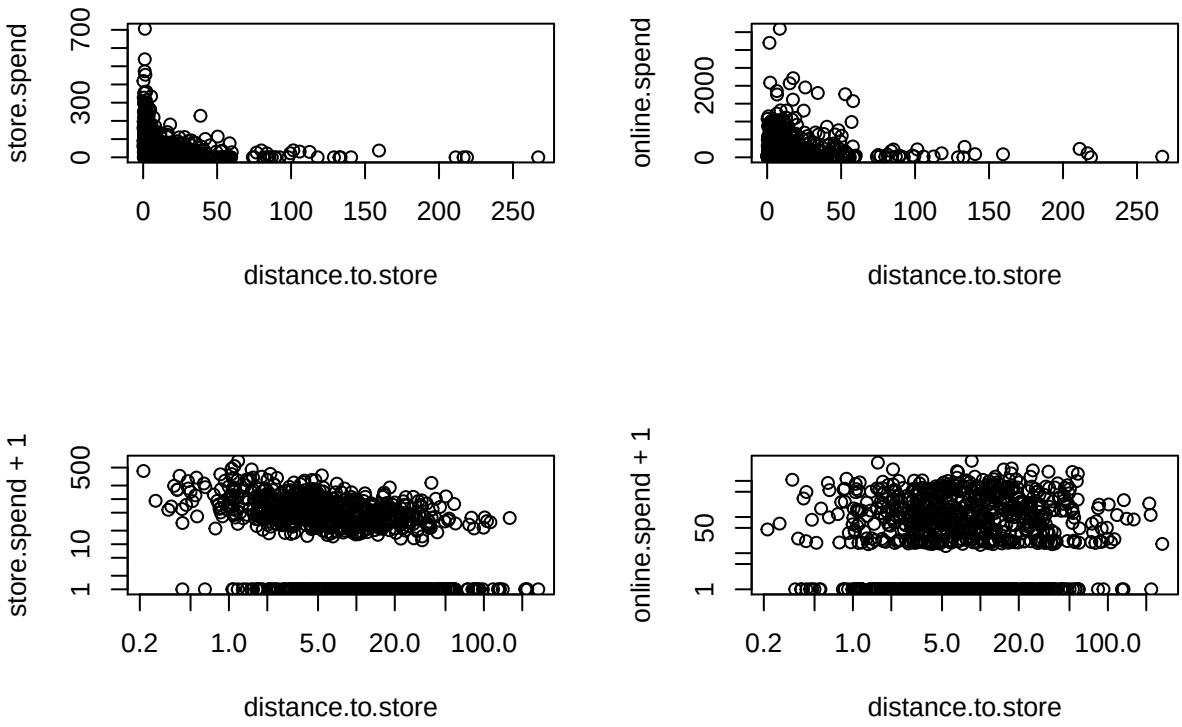
3.2.7.Utilisation de la fonction logarithmique

```
plot(cust.df$store.spend+1,cust.df$online.spend+1,log = "xy")
```



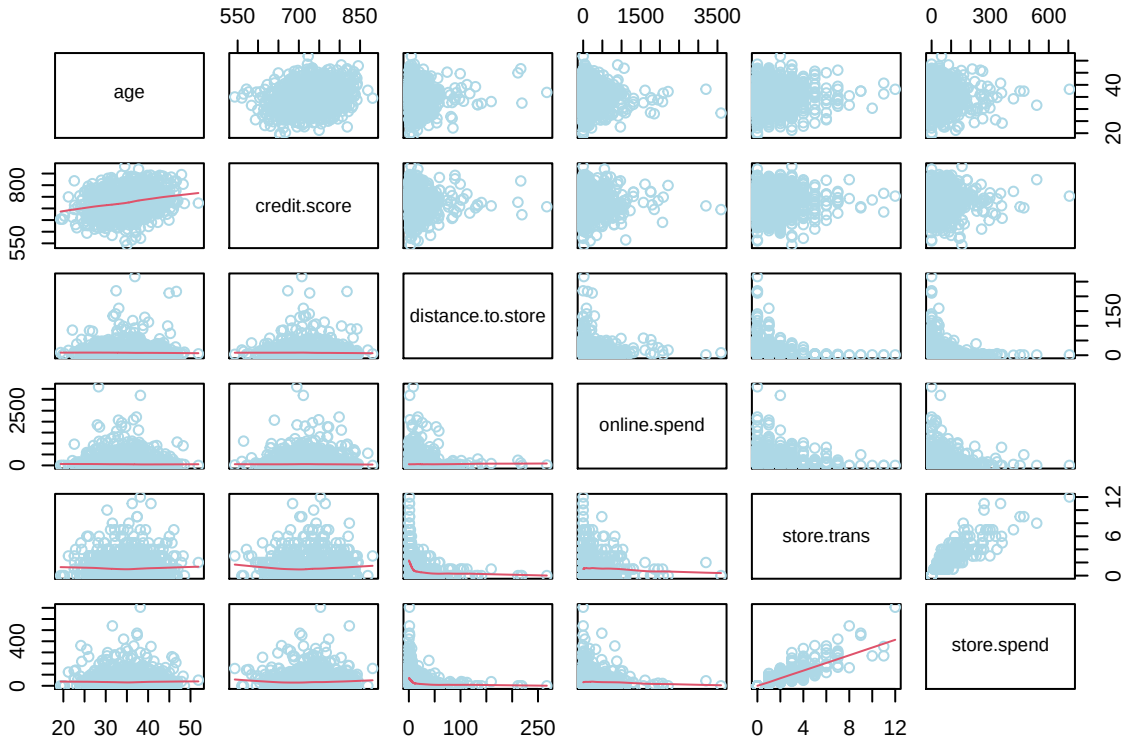
3.2.8.Multi-panel plot

```
par(mfrow=c(2, 2))
with(cust.df, plot(distance.to.store, store.spend))
with(cust.df, plot(distance.to.store, online.spend))
with(cust.df, plot(distance.to.store, store.spend+1, log="xy"))
with(cust.df, plot(distance.to.store, online.spend+1, log="xy"))
```



3.2.9.ScatterPlot Matrix

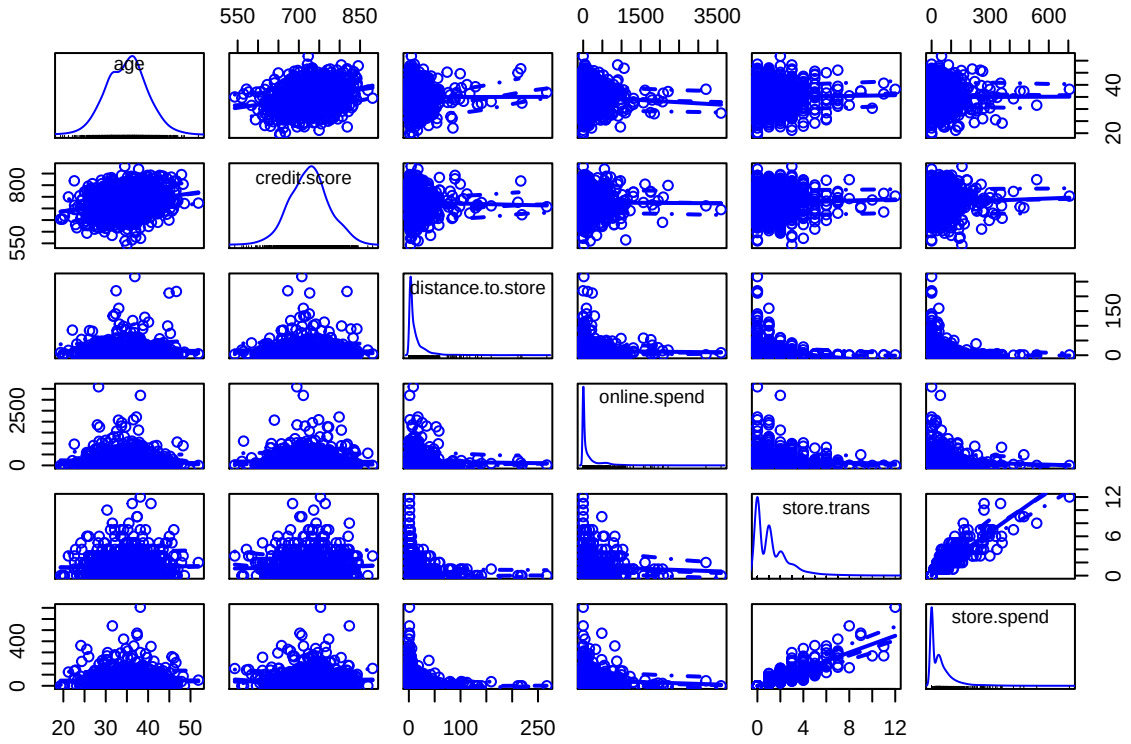
```
pairs(formula = ~ age + credit.score + distance.to.store +
online.spend + store.trans + store.spend, data=cust.df,lower.panel=panel.smooth,col="lightblue")
```



3.2.10.ScatterPlot Matrix with car data

```
library(car)
scatterplotMatrix(formula = ~ age + credit.score +
distance.to.store + online.spend + store.trans +
store.spend, data=cust.df, diagonal="histogram")
```

```
## Warning in applyDefaults(diagonal, defaults = list(method =
## "adaptiveDensity"), : unnamed diag arguments, will be ignored
```



3.2.11. Coefficient de correlation de Pearson

```
cor(cust.df$age,cust.df$credit.score)
```

```
## [1] 0.2545045
```

3.2.12. Matrice de correlation

```
cor(cust.df[, c(2, 3, 5:12)])
```

```
##                          age credit.score distance.to.store online.visits
## age                1.000000000  0.254504457        0.00198741   -0.06138107
## credit.score       0.254504457  1.000000000       -0.02326418   -0.01081827
## distance.to.store  0.001987410 -0.023264183        1.00000000   -0.01460036
## online.visits     -0.061381070 -0.010818272       -0.01460036    1.00000000
## online.trans      -0.063019935 -0.005018400       -0.01955166    0.98732805
## online.spend      -0.060685729 -0.006079881       -0.02040533    0.98240684
## store.trans        0.024229708  0.040424158       -0.27673229   -0.03666932
## store.spend        0.003841953  0.042298123       -0.24149487   -0.05068554
## sat.service                NA           NA                NA            NA
## sat.selection              NA           NA                NA            NA
```

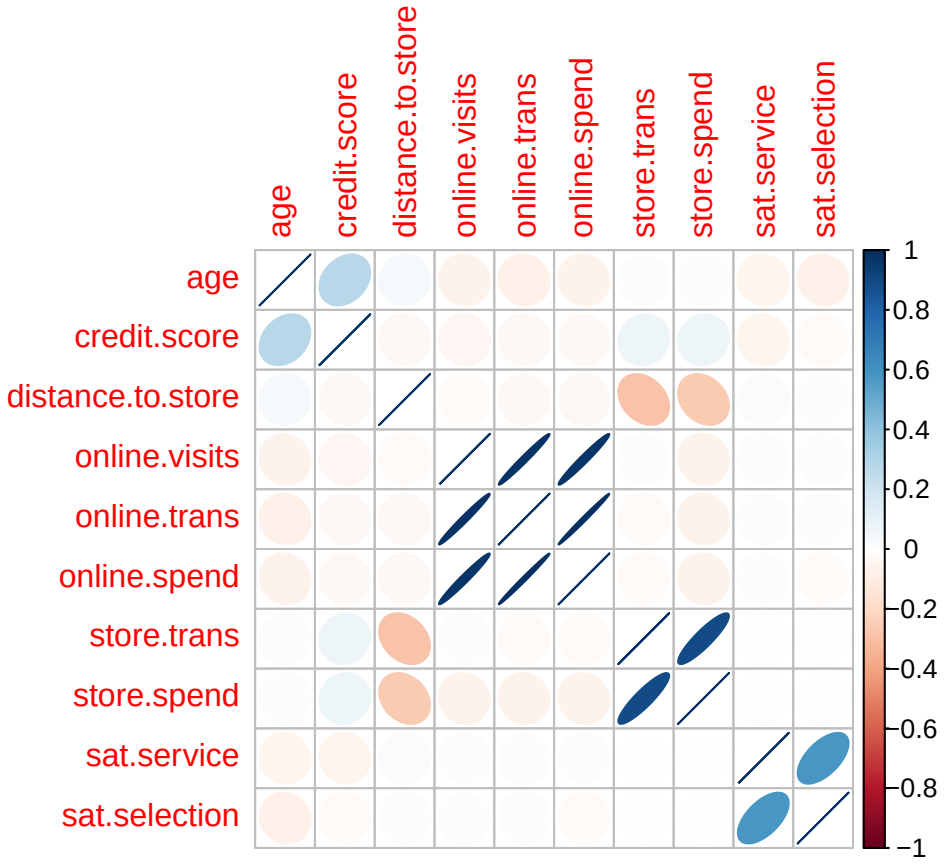
```
##                   online.trans online.spend store.trans  store.spend
## age                -0.06301994 -0.060685729  0.02422971  0.003841953
## credit.score       -0.00501840 -0.006079881  0.04042416  0.042298123
## distance.to.store  -0.01955166 -0.020405326 -0.27673229 -0.241494870
## online.visits       0.98732805  0.982406842 -0.03666932 -0.050685537
## online.trans        1.00000000  0.993346657 -0.04024588 -0.052244650
## online.spend        0.99334666  1.000000000 -0.04089133 -0.051690053
## store.trans        -0.04024588 -0.040891332  1.00000000  0.892756851
## store.spend        -0.05224465 -0.051690053  0.89275685  1.000000000
## sat.service                 NA           NA          NA           NA
## sat.selection               NA           NA          NA           NA
##                   sat.service sat.selection
## age                        NA            NA
## credit.score               NA            NA
## distance.to.store          NA            NA
## online.visits              NA            NA
## online.trans               NA            NA
## online.spend               NA            NA
## store.trans                NA            NA
## store.spend                NA            NA
## sat.service                 1            NA
## sat.selection              NA             1
```

3.2.13. Visualiser la matrice de corrélation

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
corrplot(corr=cor(cust.df[ , c(2, 3, 5:12)],
use="complete.obs"), method ="ellipse")
```



3.2.14. Transformation de données

```
cor(cust.df$distance.to.store, cust.df$store.spend)
```

```
## [1] -0.2414949
```

```
cor(1/cust.df$distance.to.store, cust.df$store.spend)
```

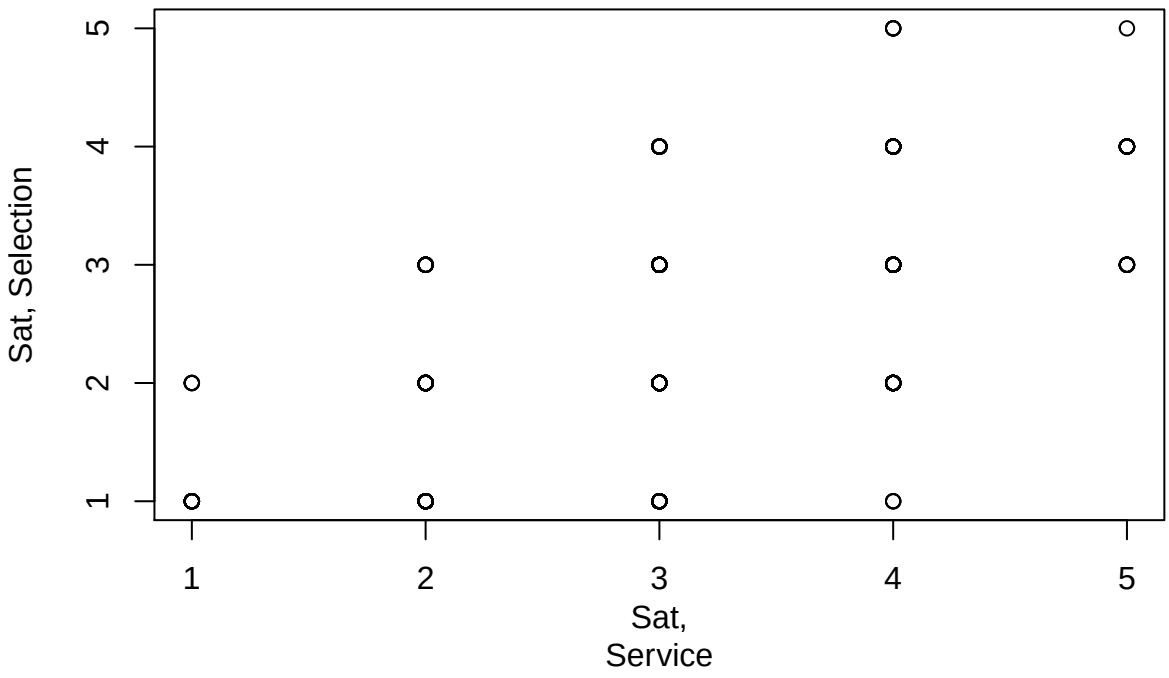
```
## [1] 0.4329997
```

```
cor(1/sqrt(cust.df$distance.to.store), cust.df$store.spend)
```

```
## [1] 0.4843334
```

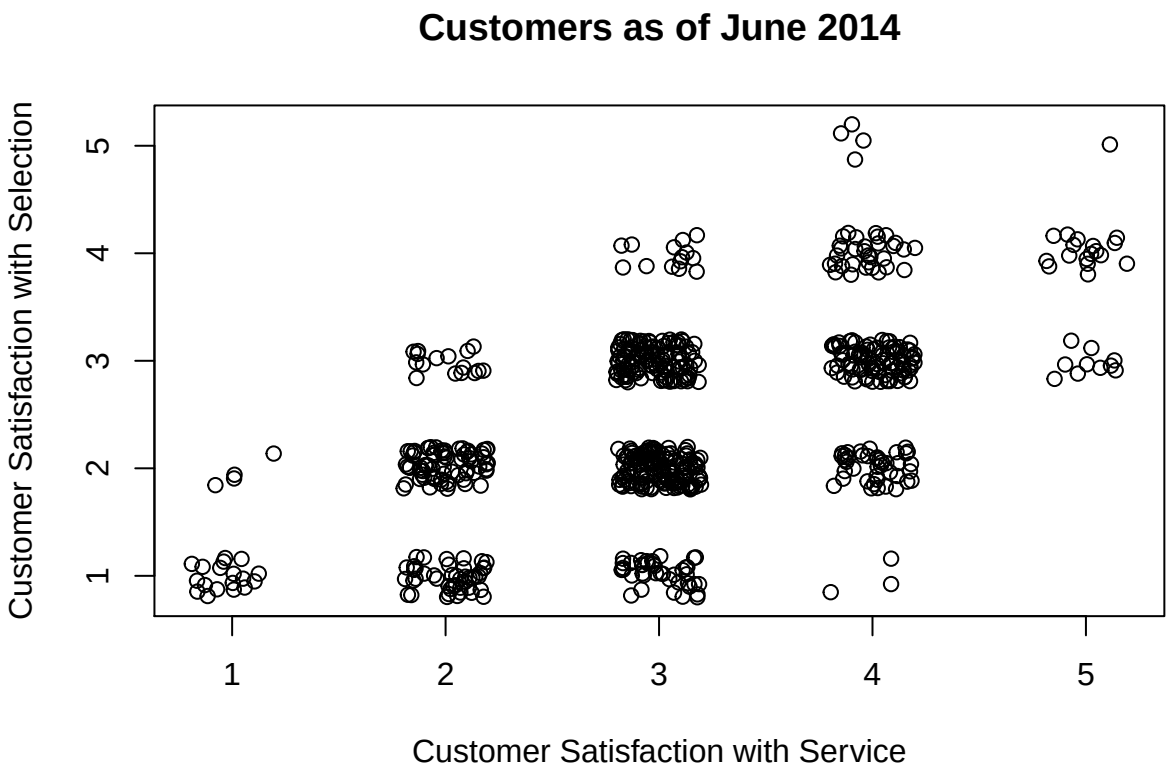
3.2.15.Exploration des associations dans des réponses d'enquête

```
plot (cust.df$sat.service, cust.df$sat.selection, xlab="Sat,
Service", ylab="Sat, Selection")
```



3.2.16.Exploration des associations dans des réponses d'enquête - suite

```
plot(jitter(cust.df$sat.service), jitter(cust.df$sat.selection),
xlab="Customer Satisfaction with Service",
ylab="Customer Satisfaction with Selection",
main="Customers as of June 2014")
```



3.2.17.Exploration des associations dans des réponses d'enquête - suite

```
resp <- !is.na(cust.df$sat.service)
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
polychoric(cbind(cust.df$sat.service[resp],
cust.df$sat.selection[resp]))
```

```
## Call: polychoric(x = cbind(cust.df$sat.service[resp], cust.df$sat.selection[resp]))
## Polychoric correlations
##    C1   C2
## R1 1.00
## R2 0.67 1.00
##
##  with tau of
```

```
##          1     2    3   4
## [1,] -1.83 -0.72 0.54 1.7
## [2,] -0.99  0.12 1.26 2.4
```

PRATIQUE - PARTIE 02

Chargement de données

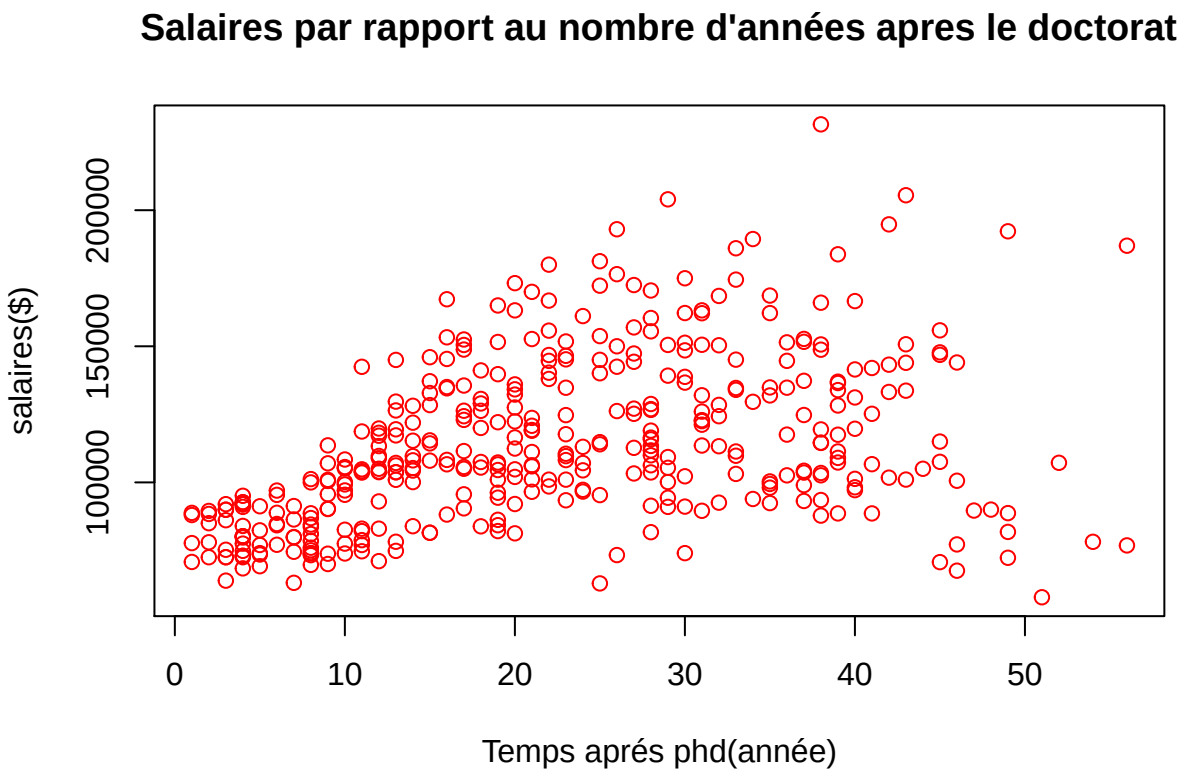
```
head(Salaries,10)
```

```
##         rank discipline yrs.since.phd yrs.service    sex salary
## 1       Prof          B            19          18   Male 139750
## 2       Prof          B            20          16   Male 173200
## 3   AsstProf          B             4           3   Male  79750
## 4       Prof          B            45          39   Male 115000
## 5       Prof          B            40          41   Male 141500
## 6  AssocProf          B             6           6   Male  97000
## 7       Prof          B            30          23   Male 175000
## 8       Prof          B            45          45   Male 147765
## 9       Prof          B            21          20   Male 119250
## 10      Prof          B            18          18 Female 129000
```

QUESTION 01 : Les salaires par rapport aux années écoulées depuis le doctorat

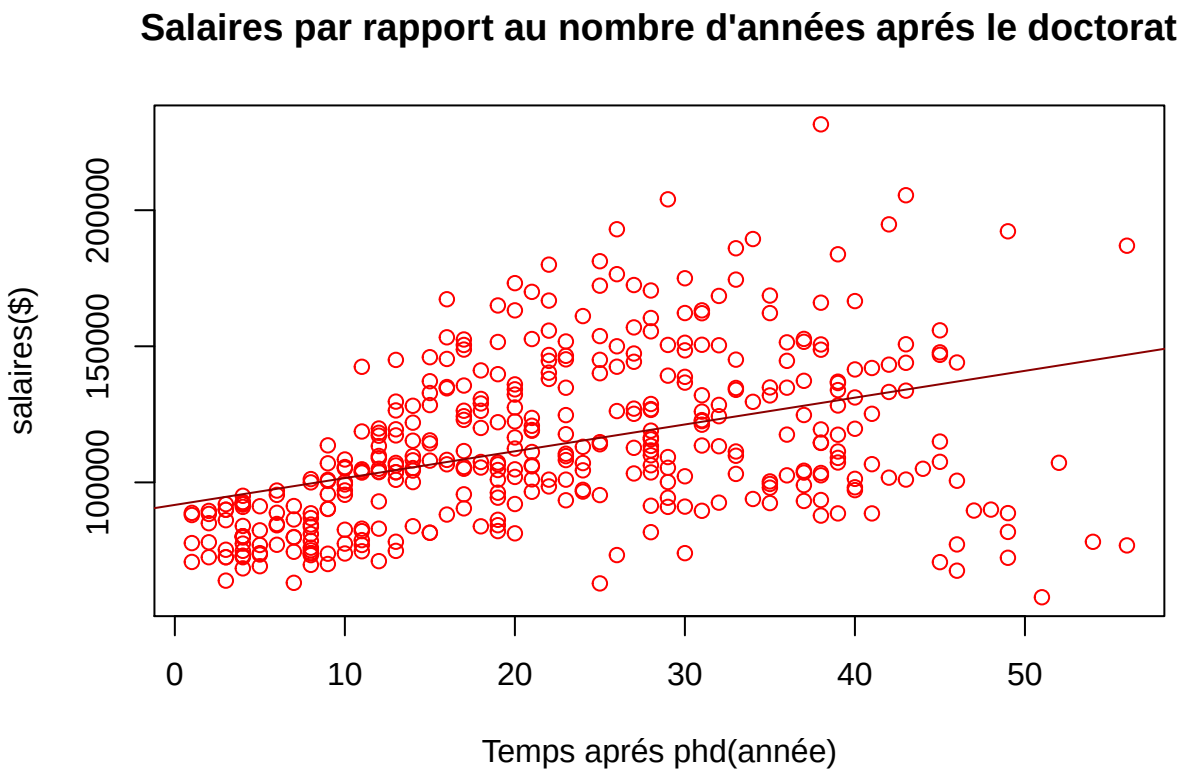
```
logSalary <- log(Salaries$salary)

plot(Salaries$yrs.since.phd,Salaries$salary,
     main = "Salaires par rapport au nombre d'années apres le doctorat",
     xlab = "Temps aprés phd(année)",
     ylab = "salaires($)",
     col="red")
```



Dessinons une droite sur le nuage de points

```
plot(Salaries$yrs.since.phd,Salaries$salary,
     main = "Salaires par rapport au nombre d'années aprés le doctorat",
     xlab = "Temps aprés phd(année)",
     ylab = "salaires($)",
     col="red")
model=lm(Salaries$salary ~ Salaries$yrs.since.phd)
abline(model,col="darkred")
```



QUESTION 02 : Corrélation entre salaire et nombre d'années aprés le doctorat

Il parait qu'il y a une linéarité entre les deux variables, Salaires et nombre d'années après le doctorat, calculons le coefficient de corrélation de Pearson pour bien évaluer le degré de dépendance

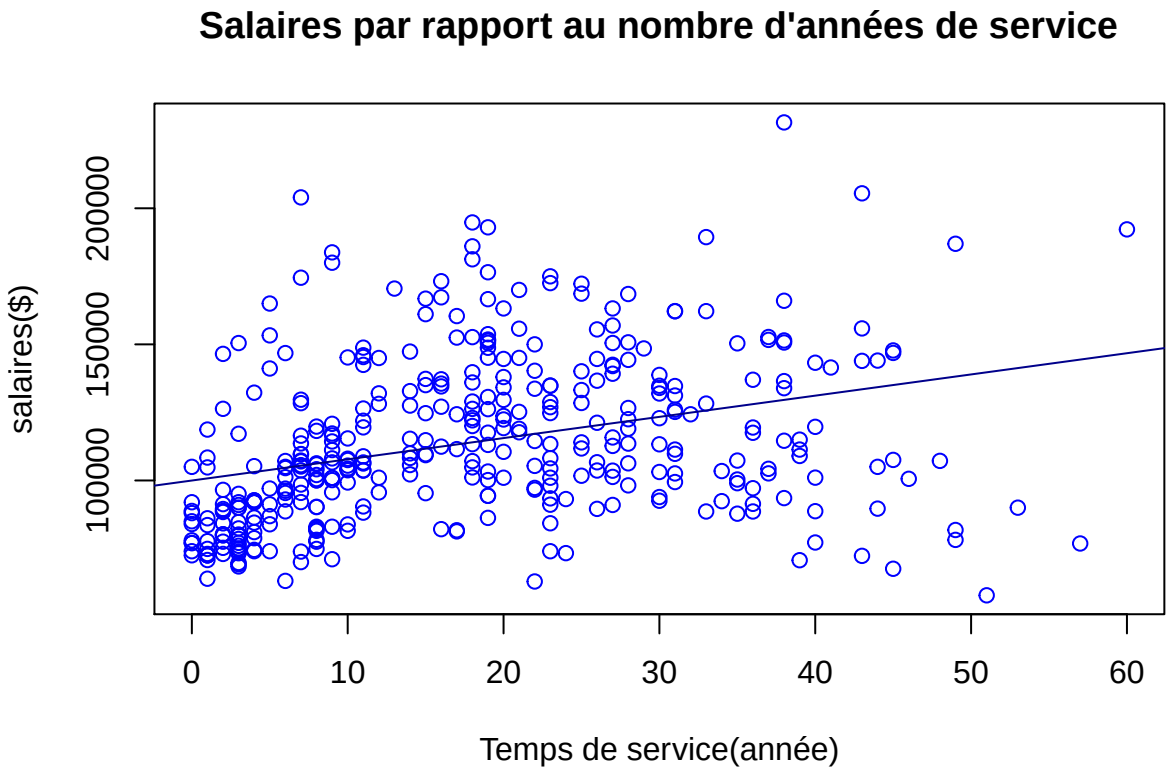
```
cor.test(Salaries$salary , Salaries$yrs.since.phd)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Salaries$salary and Salaries$yrs.since.phd
## t = 9.1775, df = 395, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3346160 0.4971402
## sample estimates:
##       cor
## 0.4192311
```

Coefficient de Pearson = 0.42 , la dépendance est considérée **modérée**

Corrélation entre la variable salaire et nombre d'années de service

```
plot(Salaries$yrs.service,Salaries$salary,
     main = "Salaires par rapport au nombre d'années de service",
     xlab = "Temps de service(année)",
     ylab = "salaires($)",
     col="blue")
model=lm(Salaries$salary ~ Salaries$yrs.service)
abline(model,col="darkblue")
```



Coefficient de corrélation

```
cor.test(Salaries$salary , Salaries$yrs.service)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Salaries$salary and Salaries$yrs.service
## t = 7.0602, df = 395, p-value = 7.529e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2443740 0.4193506
```

```
## sample estimates:
##       cor
## 0.3347447
```

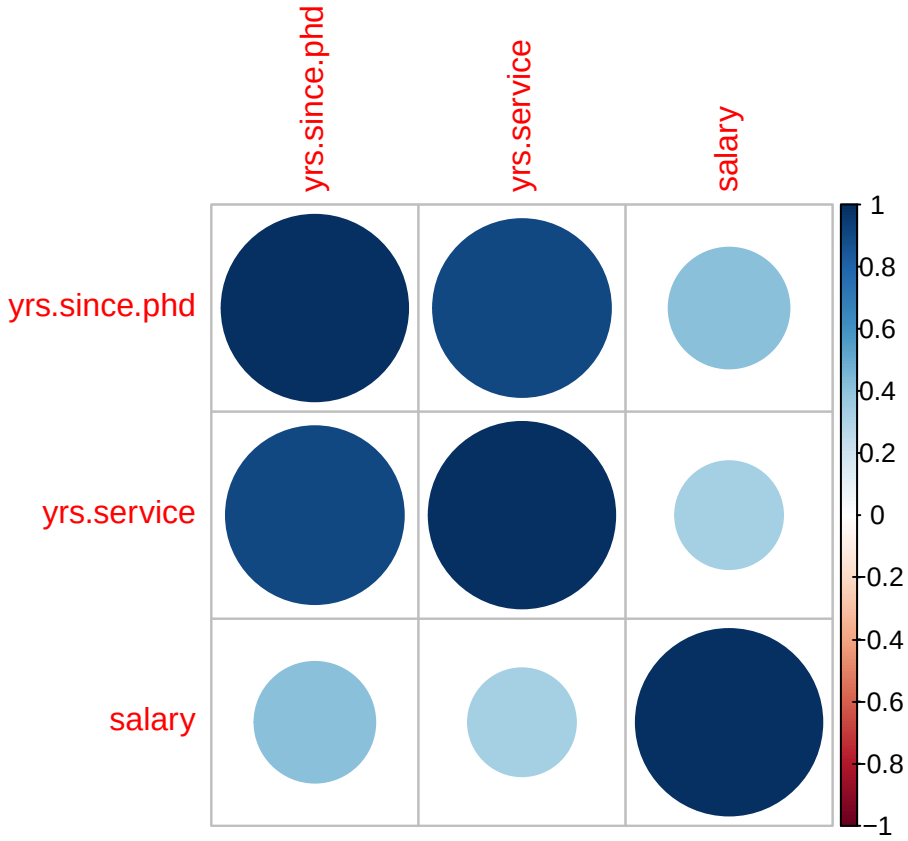
Coefficient de Pearson = 0.33 , la dépendance est considérée **faible**

P-Value

On observe que dans les deux test le P-Value est de loins inférieur à 0.05, alors les résultats sont significatives statistiquement

QUESTION 03: Visualisation de toutes les relations bivariées

```
library(corrplot)
corrplot(corr=cor(Salaries[ , c(3, 4, 6)],
use="complete.obs"), method ="circle",
sig.level =0.05,insig = "blank",)
```



On observe une corrélation relativement faible entre la variable salaire et les deux autres variables “nombre d’années de service” et "nombre d’années écoulées après le doctorat.