Projet R n°1

UE Programmation R, Master 1 Mathématiques et Applications spécialité Ingenierie Mathématique pour les Sciences du Vivant

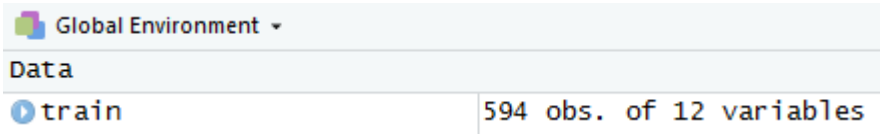
Ouhssaine Nadia

UFR Math-Info, Université Paris Descartes, 26/10/2018

Le naufrage du Titanic

Description des données

Question 1 : Charger les données du data frame *train*



Question 2 : Explorer la structure des données

Nombres d'observations et de variables / Noms des variables :

```
str(train)
```

```
## 'data.frame':    594 obs. of  12 variables:
##  $ PassengerId: int  707 706 566 244 825 754 751 649 463 438 ...
##  $ Survived   : int  1 0 0 0 0 0 1 0 0 1 ...
##  $ Pclass     : int  2 2 3 3 3 3 2 3 1 2 ...
##  $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 439 561 205 505 635 420 859 872 285 69
##  $ Sex        : Factor w/ 2 levels "female","male": 1 2 2 2 2 2 1 2 2 1 ...
##  $ Age        : num  45 39 24 22 2 23 4 NA 47 24 ...
##  $ SibSp      : int  0 0 2 0 4 0 1 0 0 2 ...
##  $ Parch      : int  0 0 0 0 1 0 1 0 0 3 ...
##  $ Ticket     : Factor w/ 681 levels "110152","110413",..: 109 171 520 659 250 352 235 621 7 238 ...
##  $ Fare       : num  13.5 26 24.15 7.12 39.69 ...
##  $ Cabin      : Factor w/ 147 levels "A10","A14","A16",..: NA NA NA NA NA NA NA NA 134 NA ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 3 3 3 3 3 3 3 3 3 3 ...
```

Il y a donc 594 observations et 12 variables. Les noms des variables sont : "PassengerId", "Survived", "Pclass", "Name", "Sex", "Age", "SibSp", "Parch", "Ticket", "Fare", "Cabin", "Embarked".

Variables qualitatives et quantitatives :

- Variables qualitatives :

```
qualit=names(train)[sapply(train,class)=='factor']
qualit
```

```
## [1] "Name"     "Sex"      "Ticket"   "Cabin"    "Embarked"
```

- Variables quantitatives :

```
quant=names(train)[sapply(train,class)=='integer']
quant
```

```
## [1] "PassengerId" "Survived"    "Pclass"      "SibSp"       "Parch"
```

Après réflexion autour de ces résultats, je constate une confusion entre les deux types de variables car certaines variables comme "PassengerId", "Survived" ou encore "Pclass" sont considérées comme des variables quantitatives alors qu'elles ne le sont pas étant donné que "PassengerId" représente un numéro de passager, "Survived" représente la survie ou la non survie du passager et "Pclass" représente la classe du passager. Cette erreur semble venir de l'encodage de ces variables qualitatives qui peuvent être parfois numérique comme par exemple les numéros de classe ou bien l'indiçage binaire des survivants (1= survivant et 0= non survivant). La machine ne comprend pas qu'il s'agit de catégories et considère cela comme des quantités.

Par ailleurs, les variables "Age" et "Fare" n'ont pas été classé, considérées seulement comme étant numérique.

Si l'on prend en considération toutes ces remarques, les variables qualitatives serait donc : "Name", "Sex", "Ticket", "Cabin", "Embarked", "PassengerId", "Survived" et "Pclass". Et les variable quantitatives : "Age", "SibSp" (= nombre de frères/soeurs/conjoints), "Parch" (= nombre de parents/enfants) et "Fare" (= tarif).

Nous allons donc modifier la *class* des variables concernées par l'erreur à l'aide du code suivant :

```
train$Age<-as.integer(train$Age)
train$Fare<-as.integer(train$Fare)
train$Survived <-as.factor(train$Survived)
train$Pclass <-as.factor(train$Pclass)
train$PassengerId <-as.factor(train$PassengerId)
```

Remarque : Ce changement de class est nécessaire pour la suite du projet, notamment lorsque l'on nous demandera de décrire les variables ou encore de comparer les variables entre elles (nous en détaillerons davantage la nécessité dans les questions concernées).

On peut maintenant réutiliser les codes précédents qui nous permettaient d'afficher les variables qualitatives et quantitatives :

- Variables qualitatives :

```
qualit=names(train)[sapply(train,class)=='factor']
qualit
```

```
## [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
## [6] "Ticket"      "Cabin"       "Embarked"
```

- Variables quantitatives :

```
quant=names(train)[sapply(train,class)=='integer']
quant
```

```
## [1] "Age"   "SibSp" "Parch" "Fare"
```

Nombre de valeurs manquantes :

```
sum(is.na(train))
```

```
## [1] 585
```

Il y a donc 585 valeurs manquantes.

Variable ayant le plus de valeurs manquantes :

```
valmanq=1
for(i in 1:11){
  if(sum(is.na(train[,valmanq]))<sum(is.na(train[,i+1]))){valmanq = i+1}}
#valmanq nous donne la position de la variable ayant le plus de valeurs manquantes
names(train[valmanq]) #ceci nous donne le nom de la varibale ayant le plus de valeurs manquantes
```

```
## [1] "Cabin"
```

Il s'agit donc de la variable "Cabin".

Question 3 : Description des variables “Survived”, “Sex”, “Plcass” et “Age”

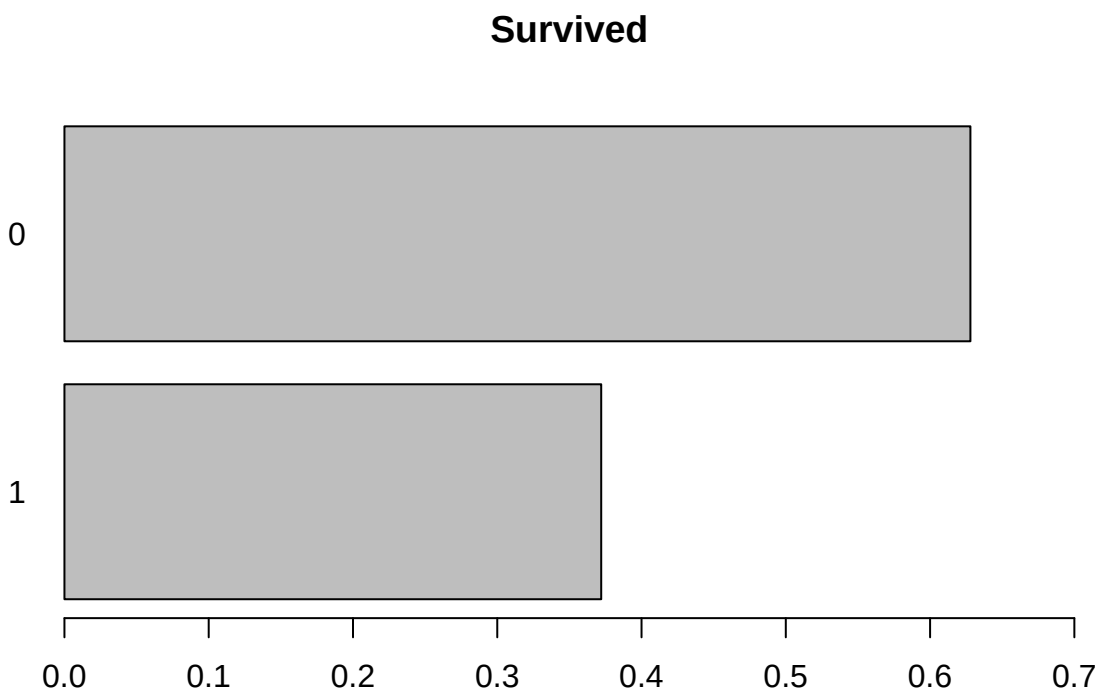
Description de Survived : *Variable qualitative*

Diagramme bâton et table des comptages :

```
require(knitr)
```

```
## Loading required package: knitr
```

```
par(mar=c(5,4,4,2)+0.1)
barplot( sort(prop.table(summary(train$Survived))), horiz=T , main="Survived",xlim=c(0,0.7),las=1)
```

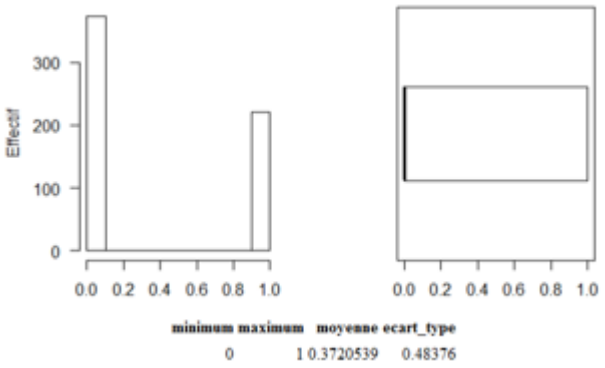



```
kable(as.data.frame(sort(summary(train$Survived),decreasing=T))
      ,col.names='Effectifs', caption="Survived")
```

Table 1: Survived

	Effectifs
0	373
1	221

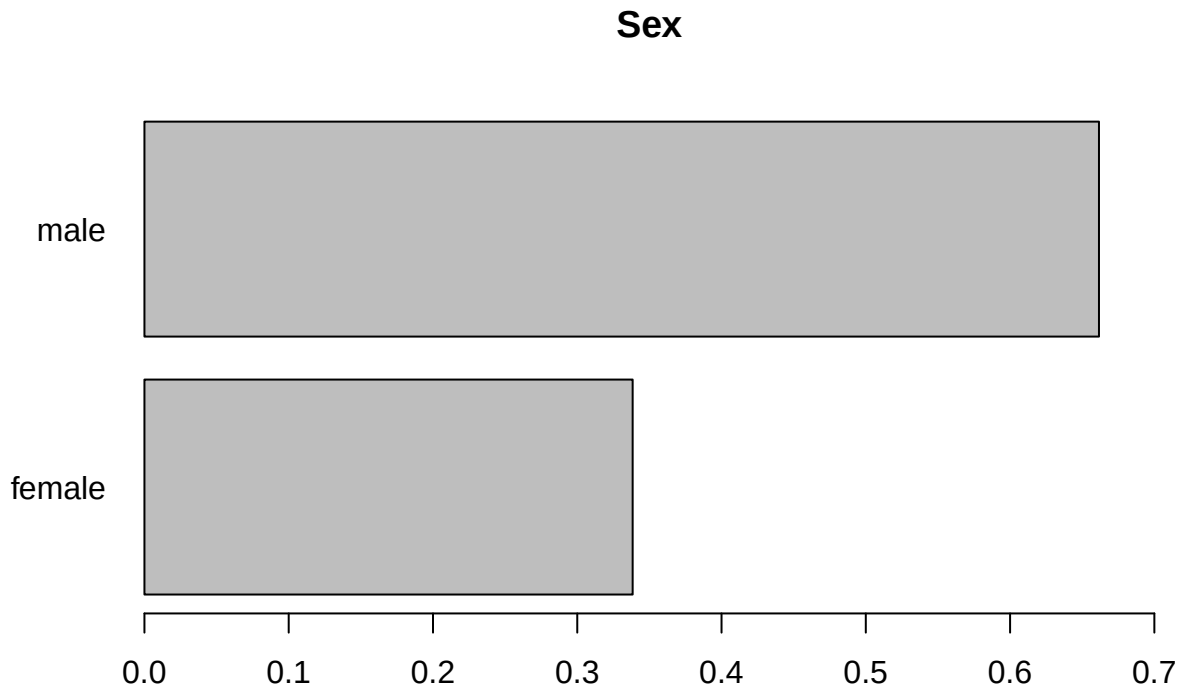
Remarque : Sans le changement de class fait à la question précédente nous aurions considéré Survived comme une variable quantitative et aurions donc fait un histogramme, un boxplot et calculer moyenne, min, max et écart-type. Ceci nous auraient donné une répresentation maladroite et peu concluante des résultats. Voici un petit aperçu de ce que l'on aurait eu :



Description de Sex : ***Variable qualitative***

Diagramme bâton et table des comptages :

```
par(mar=c(5,4,4,2)+0.1)
barplot( sort(prop.table(summary(train$Sex))), horiz=T,xlim=c(0,0.7), main="Sex", las=1)
```



```
kable(as.data.frame(sort(summary(train$Sex),decreasing=T))
      ,col.names='Effectifs', caption="Sex")
```

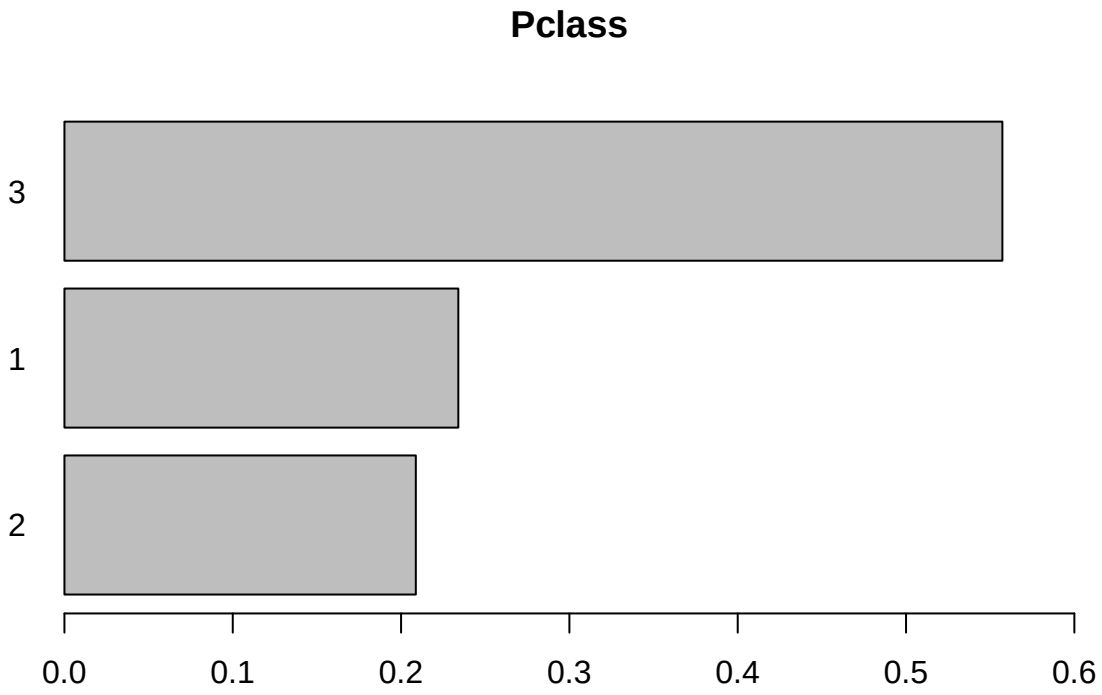
Table 2: Sex

	Effectifs
male	393
female	201

Description de Pclass : *Variable qualitative*

Diagramme bâton et table des comptages :

```
par(mar=c(5,4,4,2)+0.1)
barplot( sort(prop.table(summary(train$Pclass))), horiz=T,xlim=c(0,0.6), main="Pclass",las=1)
```

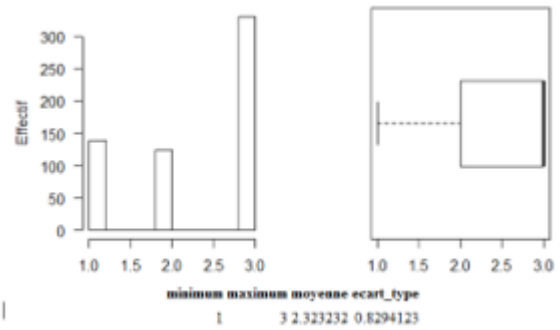


```
kable(as.data.frame(sort(summary(train$Pclass),decreasing=T))
      ,col.names='Effectifs', caption="Pclass")
```

Table 3: Pclass

	Effectifs
3	331
1	139
2	124

Remarque : Sans le changement de class fait à la question précédente, nous aurions considéré Pclass comme une variable quantitative et aurions donc fait un histogramme, un boxplot et calculer moyenne, min, max et écart-type. Ceci nous auraient donné une répresentation maladroite et peu concluante des résultats. Voici un petit aperçu de ce que l'on aurait eu :



Description de Age : ***Variable quantitative***

```
etud_Age=data.frame(minimum=min(train$Age,na.rm=TRUE)
                    ,maximum=max(train$Age,na.rm=TRUE)
                    ,moyenne=mean(train$Age,na.rm=TRUE)
                    ,ecart_type=sd(train$Age,na.rm=TRUE)
                    )
kable(etud_Age)
```

minimum	maximum	moyenne	ecart_type
0	71	29.55814	14.36676

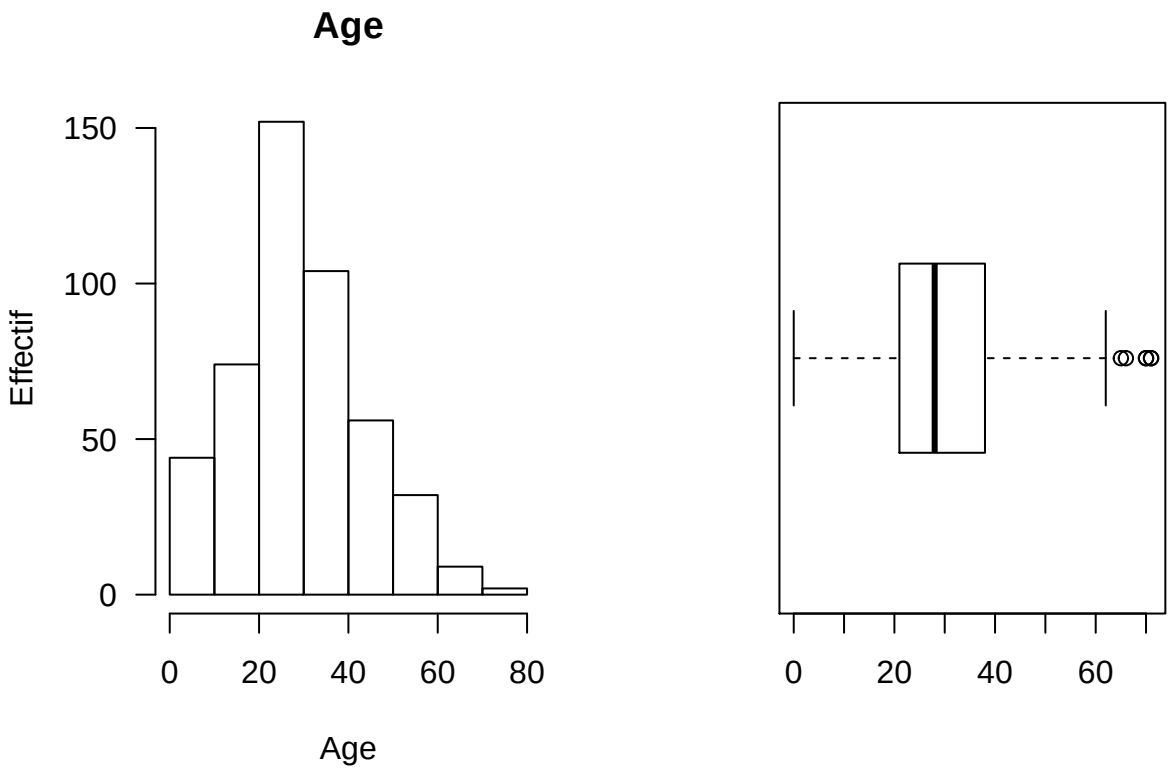
Histogramme et Boxplot (accompagné de la fonction générique Summary) :

```
par(mfrow=c(1,2))
hist(train$Age,main="Age", xlab="Age", ylab="Effectif",las=1)
summary(train$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
```

```
##    0.00   21.00   28.00   29.56   38.00   71.00     121
```

```
boxplot(train$Age,horizontal=T)
```



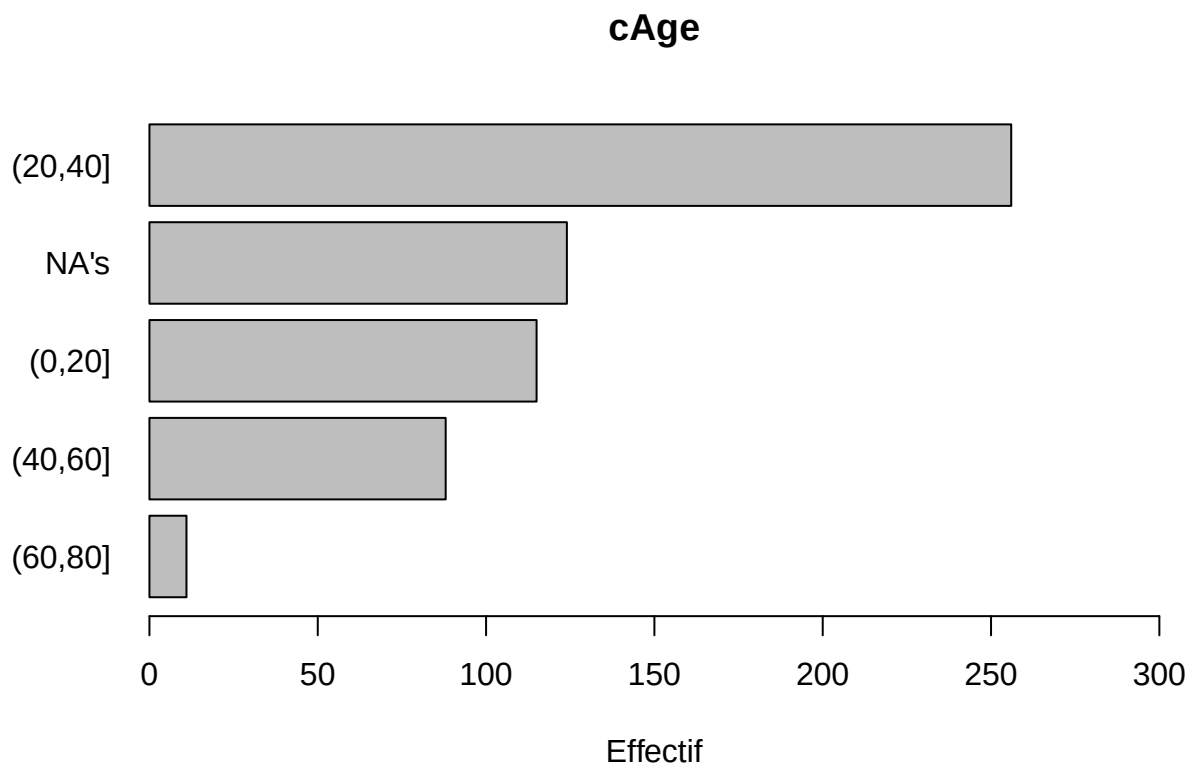
Question 4 : Construction et description de la variable cAge

Construction :

```
train$cAge<-factor(cut(train$Age, breaks = 20*(0:4)))
```

Description :

```
barplot(sort(summary(train$cAge)),main="cAge",
xlab="Effectif",horiz=TRUE,las=1,xlim=c(0,300))
```




```
kable(as.list.data.frame(sort(summary(train$cAge,decreasing=T)))
      ,col.names='Effectifs', caption="cAge")
```

Table 5: cAge

	Effectifs
(60,80]	11
(40,60]	88
(0,20]	115
NA's	124
(20,40]	256

Lien entre les variables

Question 5 : Décrire les liens entre les variables

Entre Sex et Survived :

Table de contingences :

```
etud_Sx_S <- table(train$Sex,train$Survived)
etud_Sx_S
```

```
##
##            0   1
##   female  49 152
##   male   324  69
```

Fréquences totales :

```
prop.table(etud_Sx_S)
```

```
##
##                    0          1
##   female 0.08249158 0.25589226
##   male   0.54545455 0.11616162
```

Fréquences marginales :

```
prop.table(etud_Sx_S, margin=1)
```

```
##
##                   0         1
##   female 0.2437811 0.7562189
##   male   0.8244275 0.1755725
```

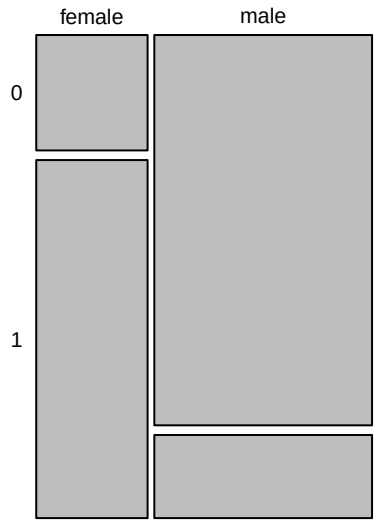
```
prop.table(etud_Sx_S, margin=2)
```

```
##
##                   0         1
##   female 0.1313673 0.6877828
##   male   0.8686327 0.3122172
```

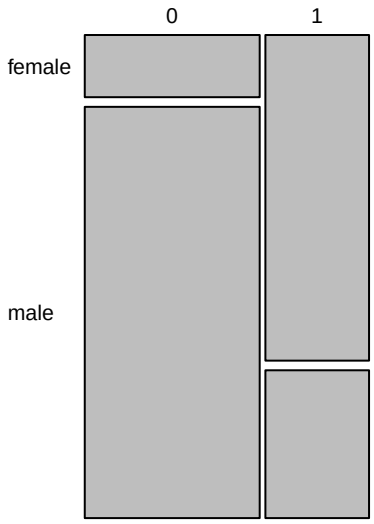
Mosaïc plot :

```
par(mfrow=c(1,2))
mosaicplot(table(train$Sex,train$Survived),main="Survived en fonction de Sex",las=1)
mosaicplot(table(train$Survived,train$Sex),main="Sex en fonction de Survived",las=1)
```

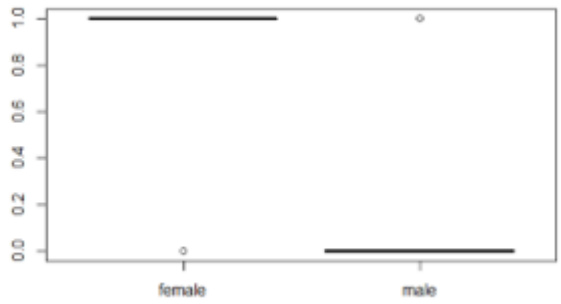
Survived en fonction de Sex



Sex en fonction de Survived



Remarque : Sans le changement de class fait à la question 2, nous aurions considéré Survived comme une variable quantitative et aurions comparé une variable qualitative à une variable quantitative. Ceci nous auraient donné une répresentation maladroite et peu concluante des résultats. Voici l'aperçu de ce que l'on aurait eu :



Entre Pclass et Survived :

Table de contingences :

```
etud_P_S <- table(train$Pclass,train$Survived)
etud_P_S
```

```
##
##       0   1
##   1  48  91
##   2  68  56
##   3 257  74
```

Fréquences totales :

```
prop.table(etud_P_S)
```

```
##
##              0          1
##   1 0.08080808 0.15319865
##   2 0.11447811 0.09427609
##   3 0.43265993 0.12457912
```

Fréquences marginales :

```
prop.table(etud_P_S, margin=1)
```

```
##
##             0         1
##   1 0.3453237 0.6546763
##   2 0.5483871 0.4516129
##   3 0.7764350 0.2235650
```

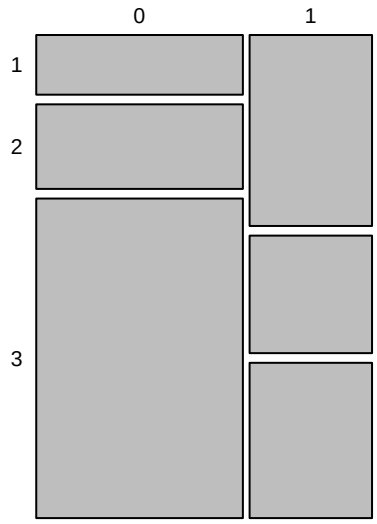
```
prop.table(etud_P_S, margin=2)
```

```
##
##             0         1
##   1 0.1286863 0.4117647
##   2 0.1823056 0.2533937
##   3 0.6890080 0.3348416
```

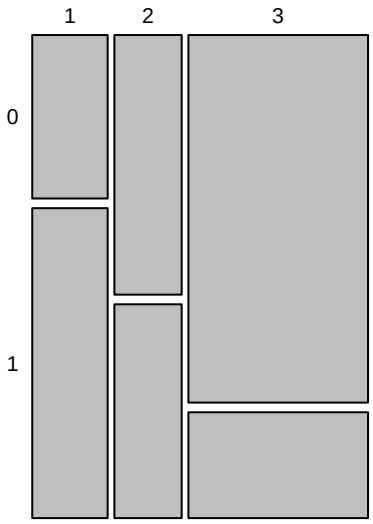
Mosaïc plot :

```
par(mfrow=c(1,2))
mosaicplot(table(train$Survived,train$Pclass),main="Pclass en fonction de Survived",las=1)
mosaicplot(table(train$Pclass,train$Survived),main="Survived en fonction de Pclass",las=1)
```

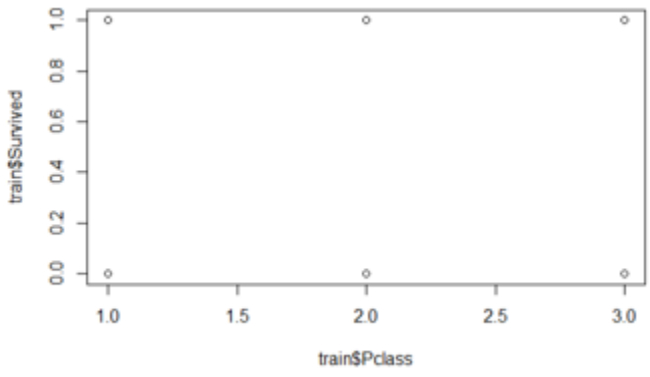
Pclass en fonction de Survived



Survived en fonction de Pclass



Remarque : Sans le changement de class fait à la question 2, nous aurions considéré Survived comme une variable quantitative et aurions comparé deux variables quantitatives. Ceci nous auraient donné une répresentation maladroite et peu concluante des résultats. Voici l'aperçu de ce que l'on aurait eu :



Entre Age et Survived :

```
tapply(train$Age, train$Survived, mean, na.rm=TRUE)
```

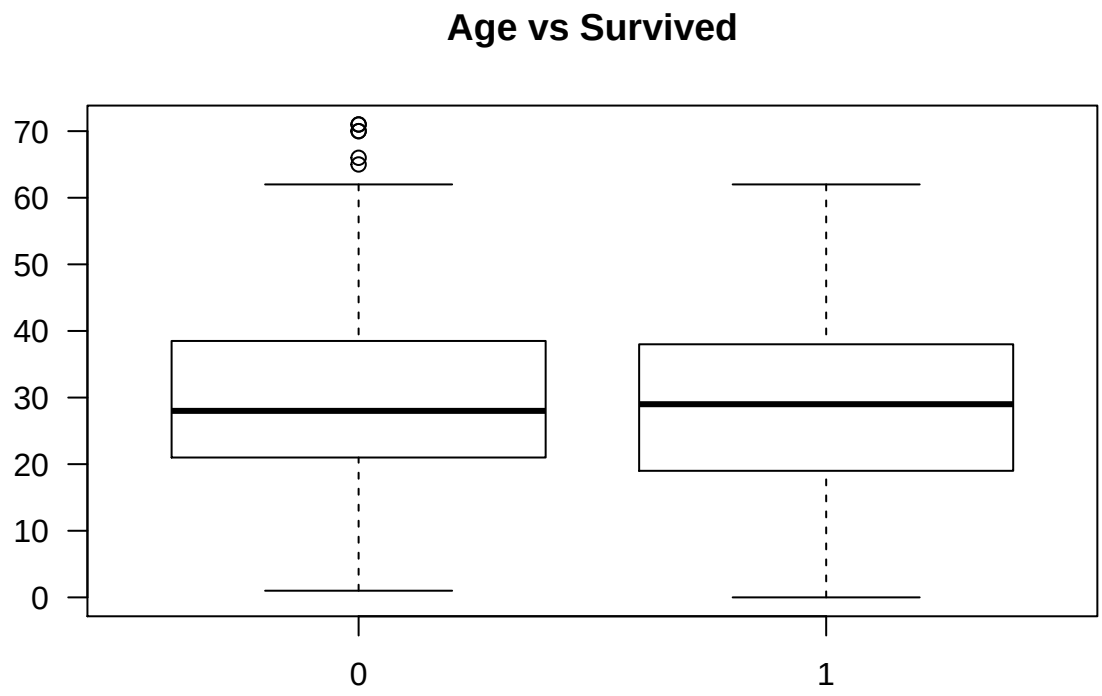
```
##        0        1
## 30.33333 28.35135
```

```
tapply(train$Age, train$Survived,summary, na.rm=TRUE)
```

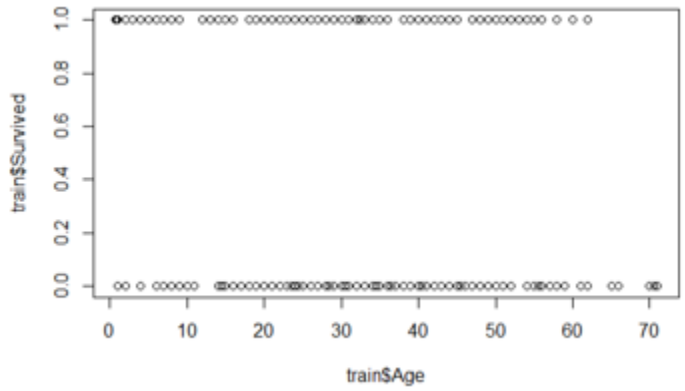
```
## $`0`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   21.00   28.00   30.33   38.25   71.00      85
##
## $`1`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   19.00   29.00   28.35   38.00   62.00      36
```

Boxplot :

```
boxplot(train$Age ~ train$Survived, data = train, las=1, main="Age vs Survived")
```



Remarque : Sans le changement de class fait à la question 2, nous aurions considéré Survived comme une variable quantitative et aurions aurions comparé deux variables quantitatives. Ceci nous auraient donné une répresentation maladroite et peu concluante des résultats. Voici l'aperçu de ce que l'on aurait eu :



Entre cAge et Survived :

Table de contingences :

```
etud_cA_S <- table(train$cAge,train$Survived)
etud_cA_S
```

```
##
##               0   1
##    (0,20]    65  50
##    (20,40] 161  95
##    (40,60]  52  36
##    (60,80]  10   1
```

Fréquences totales :

```
prop.table(etud_cA_S)
```

```
##
##                       0          1
##    (0,20]  0.13829787 0.10638298
##    (20,40] 0.34255319 0.20212766
##    (40,60] 0.11063830 0.07659574
##    (60,80] 0.02127660 0.00212766
```

Fréquence marginales :

```
prop.table(etud_cA_S, margin=1)
```

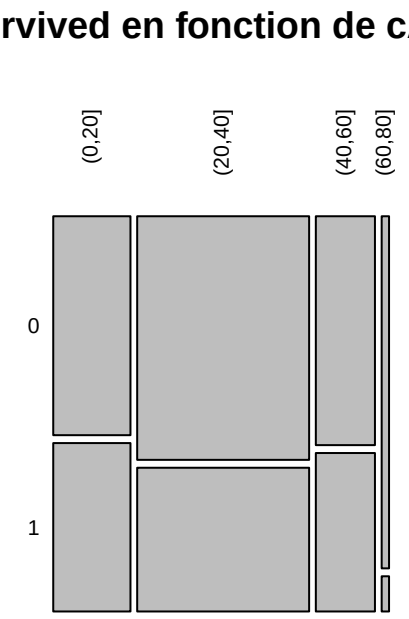
```
##
##                       0          1
##    (0,20]  0.56521739 0.43478261
##    (20,40] 0.62890625 0.37109375
##    (40,60] 0.59090909 0.40909091
##    (60,80] 0.90909091 0.09090909
```

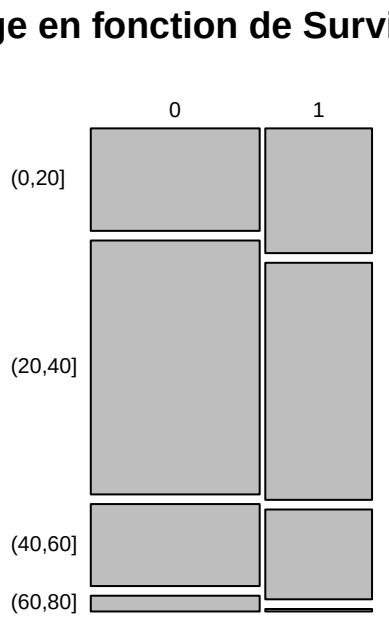
```r
prop.table(etud_cA_S, margin=2)
```

```
##
##                     0           1
##   (0,20]  0.225694444 0.274725275
##   (20,40] 0.559027778 0.521978022
##   (40,60] 0.180555556 0.197802198
##   (60,80] 0.034722222 0.005494505
```

Mosaïc plot :

```r
par(mfrow=c(1,2))
mosaicplot(table(train$cAge,train$Survived),main="Survived en fonction de cAge",las=2)
mosaicplot(table(train$Survived,train$cAge),main="cAge en fonction de Survived",las=1)
```





Question 6 : Commentaires et hypothèses

- Commentaires et hypothèse sur le lien entre Sex et Survived :

D'apèrs la table de contingences, il y a plus de survivants femme (152) que de survivants homme (69). Cet écart est d'autant plus frappant par le fait qu'il y a plus d'hommes que de femmes (201 femmes, 393 hommes). D'après les fréquences observées, on relève particulièrement que 76% des femmes ont survécu alors que 82% des hommes ont péri. Par ailleurs, 69% des survivants sont des femmes et 87% des non survivants sont des hommes. Le mosaïc plot nous illustre parfaitement ces derniers résultats. Ainsi, on peut facilement penser que le sexe influ sur la survie des passagers. Les femmes avaient, en effet, plus de chance de survivre qu'un homme.

- Commentaires et hypothèse sur le lien entre Pclass et Survived :

D'apèrs la table de contingences, on constate que la 1ère classe est la seule classe ayant plus de survivants que de non survivants (1=91 ; 0=49). La 2ème classe a très globalement autant de survivants que de non survivants (1=56 ; 0=68) et enfin la 3ème classe est la classe présentant le plus grand écart avec nettement plus de non survivants (1=74 ; 0=257). D'après les fréquences observées, on relève particulièrement que 41% des survivants sont de la 1ère classe tandis que 69% des non survivants sont de la 3ème classe. Ainsi, on peut facilement penser que la classe des passagers influ sur leur survie. Les passagers de 1ère classe ont,en effet, beaucoup plus de chance de survivre tandis que les passagers de la 3ème classe ont beaucoup moins de chance de survivre, la 2ème classe faisant intermédiaire au deux autres.

- Commentaires et hypothèse sur le lien entre cAge et Survived :

D'après la table de contingences, on constate que la tranche d'âge (20;40] est celle ayant le plus de survivants, suivi de la tranche d'âges (0;20], puis de (40;60] et enfin de (60;80]. C'est exactement le même classement pour les tranches d'âges ayant le plus de non survivant. A ce stade, on pourrait penser qu'il n'y a pas d'âge favorable à la survie car ce classement correspond également au classement par effectif de chaque tranche d'âge (du plus grand effectif au plus faible). Les fréquences observées nous confirme cette stabilité de classement entre les tranches d'âges. Cependant en regardant le mosaïc plot, on constate que pour la tranche d'âge (60;80] la différence survivant/non survivant se dessine davantage. Au vue du faible effectif de cette dernière tranche d'âge, on peut tout de même conclure que globalement, il n'y a pas d'âge favorable à la survie.

Prédiction de la survie

Question 7 : Estimation de probabilité de survie conditionnelle

- $\mathbb{P}(S = 1|Sx = \text{female})$:

```
sum(train$Sex=='female' & train$Survived==1)/sum(train$Sex=='female')
```

```
## [1] 0.7562189
```

Ainsi, $\mathbb{P}(S = 1|Sx = \text{female}) = 0.7562189$

- $\mathbb{P}(S = 1|Sx = \text{male})$:

```
sum(train$Sex=='male' & train$Survived==1)/sum(train$Sex=='male')
```

```
## [1] 0.1755725
```

Ainsi, $\mathbb{P}(S = 1|Sx = \text{male}) = 0.1755725$

- $\mathbb{P}(S = 1|P = 1)$:

```
sum(train$Pclass==1 & train$Survived==1)/sum(train$Pclass==1)
```

```
## [1] 0.6546763
```

Ainsi, $\mathbb{P}(S = 1|P = 1) = 0.6546763$

- $\mathbb{P}(S = 1|P = 2)$:

```
sum(train$Pclass==2 & train$Survived==1)/sum(train$Pclass==2)
```

```
## [1] 0.4516129
```

Ainsi, $\mathbb{P}(S = 1|P = 2) = 0.4516129$

- $\mathbb{P}(S = 1|P = 3)$:

```
sum(train$Pclass==3 & train$Survived==1)/sum(train$Pclass==3)
```

```
## [1] 0.223565
```

Ainsi, $\mathbb{P}(S = 1|P = 3) = 0.223565$

- $\mathbb{P}(S = 1|cAge = (0, 20])$:

```
sum(is.na(train$cAge)==FALSE & train$cAge=="(0,20]"
& train$Survived==1)/sum(is.na(train$cAge)==FALSE & train$cAge=="(0,20]")
```

```
## [1] 0.4347826
```

Ainsi, $\mathbb{P}(S = 1|cAge = (0, 20]) = 0.4347826$

- $\mathbb{P}(S = 1|cAge = (20, 40])$:

```
sum(is.na(train$cAge)==FALSE & train$cAge=="(20,40]"
& train$Survived==1)/sum(is.na(train$cAge)==FALSE & train$cAge=="(20,40]")
```

```
## [1] 0.3710938
```

Ainsi, $\mathbb{P}(S = 1|cAge = (20, 40]) = 0.3710938$

- $\mathbb{P}(S = 1|cAge = (40, 60])$:

```
sum(is.na(train$cAge)==FALSE & train$cAge=="(40,60]"
& train$Survived==1)/sum(is.na(train$cAge)==FALSE & train$cAge=="(40,60]")
```

```
## [1] 0.4090909
```

Ainsi, $\mathbb{P}(S = 1|cAge = (40, 60]) = 0.4090909$

- $\mathbb{P}(S = 1|cAge = (60, 80])$:

```
sum(is.na(train$cAge)==FALSE & train$cAge=="(60,80]"
& train$Survived==1)/sum(is.na(train$cAge)==FALSE & train$cAge=="(60,80]")
```

```
## [1] 0.09090909
```

Ainsi, $\mathbb{P}(S = 1|cAge = (60, 80]) = 0.09090909$

Question 8 : Construction des tables de probabilité conditionnelle

```
#P(S|P):
S_P <- prop.table(table(train$Pclass, train$Survived), margin=2)
rownames(S_P) <- c('1','2', '3')
colnames(S_P) <- c('0','1')
S_P
```

```
##
##            0         1
##   1 0.1286863 0.4117647
##   2 0.1823056 0.2533937
##   3 0.6890080 0.3348416
```

```
#P(S|Sx):
S_Sx <- prop.table(table(train$Sex, train$Survived), margin=2)
rownames(S_Sx) <- c('female','male')
colnames(S_Sx) <- c('0','1')
S_Sx
```

```
##
##                  0         1
##   female 0.1313673 0.6877828
##   male   0.8686327 0.3122172
```

```r
#P(S|cA):
S_cA <- prop.table(table(train$cAge, train$Survived), margin=2)
rownames(S_cA) <- c('(0,20]','(20,40]', '(40,60]','(60,80]')
colnames(S_cA) <- c('0','1')
S_cA
```

```
##
##                     0           1
##   (0,20]  0.225694444 0.274725275
##   (20,40] 0.559027778 0.521978022
##   (40,60] 0.180555556 0.197802198
##   (60,80] 0.034722222 0.005494505
```

```r
#P(S = 1) et P(S = 0)
S <- prop.table(table(train$Survived))
names(S)<-c('0','1')
S
```

```
##         0         1
## 0.6279461 0.3720539
```

Question 9: Fonction de probabilité de survie des passagers avec le classificateur de Bayes

```r
prob_prediction<-function(Sex, Pclass, cAge){
P1<-(S_Sx[Sex,'1']*S_P[Pclass,'1']*S_cA[cAge,'1']*S['1'])
P2<-(S_Sx[Sex,'0']*S_P[Pclass,'0']*S_cA[cAge,'0']*S['0'])
P<-P1/(P1+P2)
P
}

#A titre d'exemple :
prob_prediction("female","1","(20,40]")
```

```
##         1
## 0.9026094
```

Evaluation de la performance du classificateur

Question 10 : Charger les données du data frame *test*

```
test                    66 obs. of 4 variables
  Survived: int 0 1 0 0 0 1 0 0 1 0 ...
  Pclass : int 1 1 1 1 1 2 3 1 1 1 ...
  Sex : Factor w/ 2 levels "female","male": 2 2 2 2 2 1 2 2 2 2 ...
  cAge : Factor w/ 4 levels "(0,20]","(20,40]",..: 3 2 1 4 3 2 2 3 2 3 ...
```

Question 11 : Application de la fonction de probabilité de survie pour chaque passager de *test*

```
#Conversion des variables Sex, Pclass et cAge
test$Sex<-as.character(test$Sex)
test$Pclass<-as.character(test$Pclass)
test$cAge<-as.character(test$cAge)
#Table de la probabilité de survie pour les 66 passagers
kable(prob_prediction(test$Sex,test$Pclass,test$cAge),col.names='Probabilité de survie')
```

Probabilité de survie
0.4274326
0.3888529
0.4533924
0.0973349
0.4274326
0.8010303
0.0881202
0.4274326
0.3888529
0.4274326
0.9157815
0.3888529
0.3888529
0.3888529
0.3888529
0.6109966
0.9235596
0.9157815
0.9235596
0.9026094
0.9026094
0.8010303
0.9235596
0.3888529
0.4274326

Probabilité de survie
0.2648740
0.9026094
0.9026094
0.3888529
0.5846499
0.0881202
0.4274326
0.0973349
0.8010303
0.3888529
0.4274326
0.9026094
0.9235596
0.4274326
0.9157815
0.9026094
0.4533924
0.9157815
0.9026094
0.9026094
0.0973349
0.0973349
0.4274326
0.3888529
0.3888529
0.9235596
0.3888529
0.9026094
0.4274326
0.8010303
0.0973349
0.4533924
0.9026094
0.9235596
0.3888529
0.4274326
0.9157815
0.4533924
0.9235596
0.9235596
0.3888529

Question 12 : Prédiction de la survie avec la règle du *Maximum a Posteriori Probability*

```
test$MAP<-character(length=66)
for(i in 1:66){
survivant_pred<-(prob_prediction(test$Sex[i],test$Pclass[i],test$cAge[i])>=0.5)
if(survivant_pred==TRUE){
  test$MAP[i]='Survivant'
```

```
}else{test$MAP[i]='Non survivant'}
}
```

Nous allons afficher quelques valeurs de cette nouvelle variable à titre d'exemple.

Survived	Pclass	Sex	cAge	MAP
0	1	male	(40,60]	Non survivant
1	1	male	(20,40]	Non survivant
0	1	male	(0,20]	Non survivant
0	1	male	(60,80]	Non survivant
0	1	male	(40,60]	Non survivant
1	2	female	(20,40]	Survivant
0	3	male	(20,40]	Non survivant
0	1	male	(40,60]	Non survivant
1	1	male	(20,40]	Non survivant
0	1	male	(40,60]	Non survivant
1	1	female	(40,60]	Survivant
1	1	male	(20,40]	Non survivant

On constate que dans certains cas la prédiction de survie du passager ne correspond pas au réel statut de survie de ce passager.

Question 13 : Comparaison vecteur de prédiction de survie et vecteur du vrai statut de survie

```
#variable de comparaison entre vecteur de préduction et vecteur du vrai statut
(etud_survived_map <- table(test$Survived,test$MAP))
```

```
##
##     Non survivant Survivant
##   0            22         2
##   1            15        27
```

```
#Table de contingence
prop.table(etud_survived_map)
```

```
##
##     Non survivant  Survivant
##   0    0.33333333 0.03030303
##   1    0.22727273 0.40909091
```

Sur les 66 passagers 49 passagers ont été bien classé (cf somme de la diagonale de la 1ère table de contingence), soit une proportion d'environ 74% (cf somme de la diagonale de la 2ème table de contingence).