# Projet Final d'Analyse de Données

Code

**Lucas Rodriguez Pereira - Anass Lahrech - Taha Bouziane**

# Analyse du jeu de données: Haberman's Survival

## Introduction

Nous trations le jeu de données intitulé : Haberman's Survival Data

Le jeu de données suivant contient des cas récoltés d'une étude menée entre 1958 et 1970 à l'hôpital Billings de l'université de Chicago sur les patiens qui ont survécu à une opération chirurgicale pour le cancer du sein.

Le nombre d'observations fournies dans ce jeu de données est de 306 avec 4 variables.

Le jeu de données ne contient de valeurs manquantes.

Ce jeu de données nous permettra de faire une étude afin de prédire la classe, c'est-à-dire si les femmes qui ont subi une opérration pour le cancer du sein survivront au moins 5 ans après l'opération.

Ci-dessous un exemple des observations du jeu de données :

Hide

```
setwd('./')
haberman <- read.table('haberman.data',sep=',',col.names = c('age','yearOfOperation',
 'nodes', "survivalStatus"))
head(haberman)
```

| | age<br><int> | yearOfOperation<br><int> | nodes<br><int> | survivalStatus<br><int> |
|---|---|---|---|---|
| 1 | 30 | 64 | 1 | 1 |
| 2 | 30 | 62 | 3 | 1 |
| 3 | 30 | 65 | 0 | 1 |
| 4 | 31 | 59 | 2 | 1 |
| 5 | 31 | 65 | 4 | 1 |
| 6 | 33 | 58 | 10 | 1 |

6 rows

## Les variables

Les variables dont on dispose sont :
1. l'âge du patient à l'âge de l'opération (numérique)
2. l'année de l'opération du patient (année 1900) (numérique)

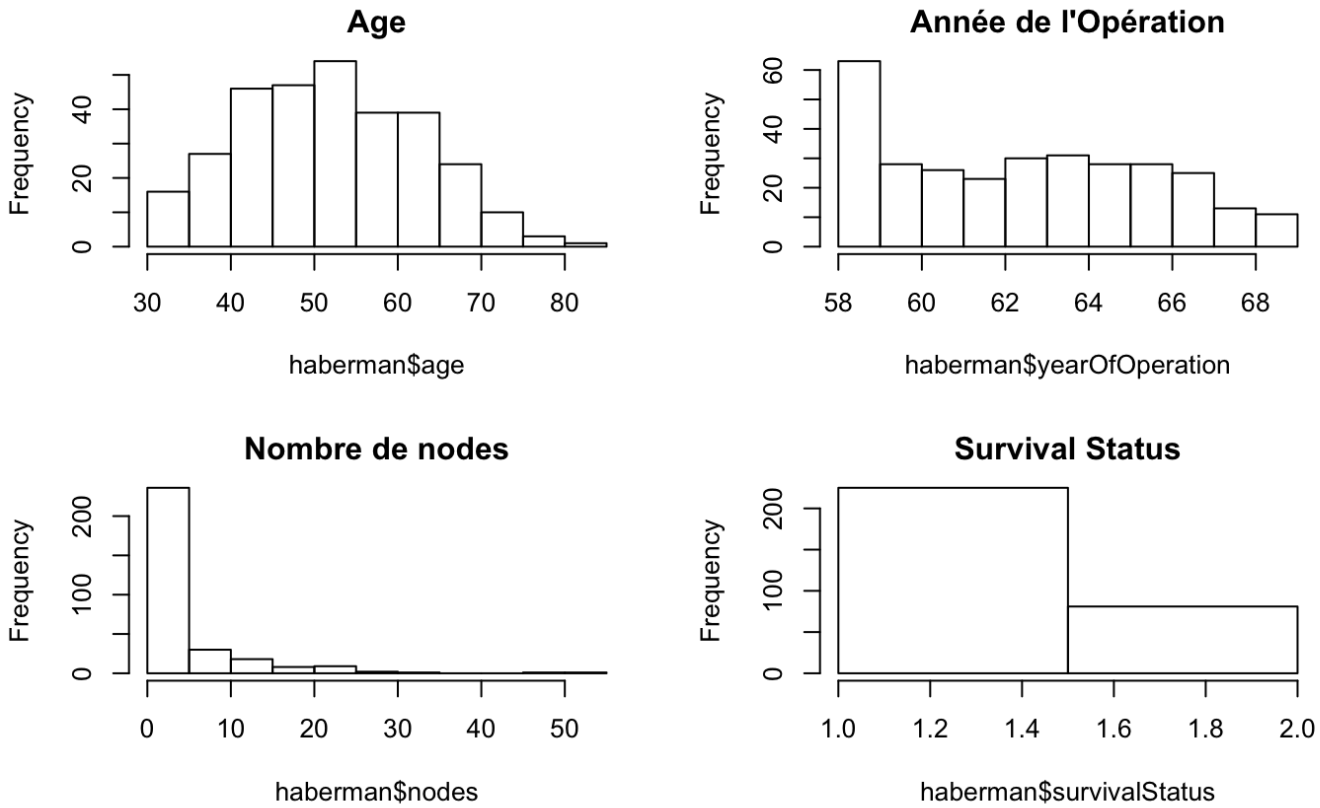
3. Nombre de nœuds auxiliaires positifs détectés (numérique)
4. Statut de survie, 1 pour un patient qui a survécu 5 ans ou plus après l'opération, 2 pour un patient qui n'a pas survécu plus de 5 ans. (binaire)

Hide

```
par(mfrow=c(2,2))
hist(haberman$age, main = "Age")
hist(haberman$yearOfOperation, main = "Année de l'Opération")
```

Hide

```
hist(haberman$nodes, main = "Nombre de nodes")
hist(haberman$survivalStatus,nclass=2, main = "Survival Status")
```

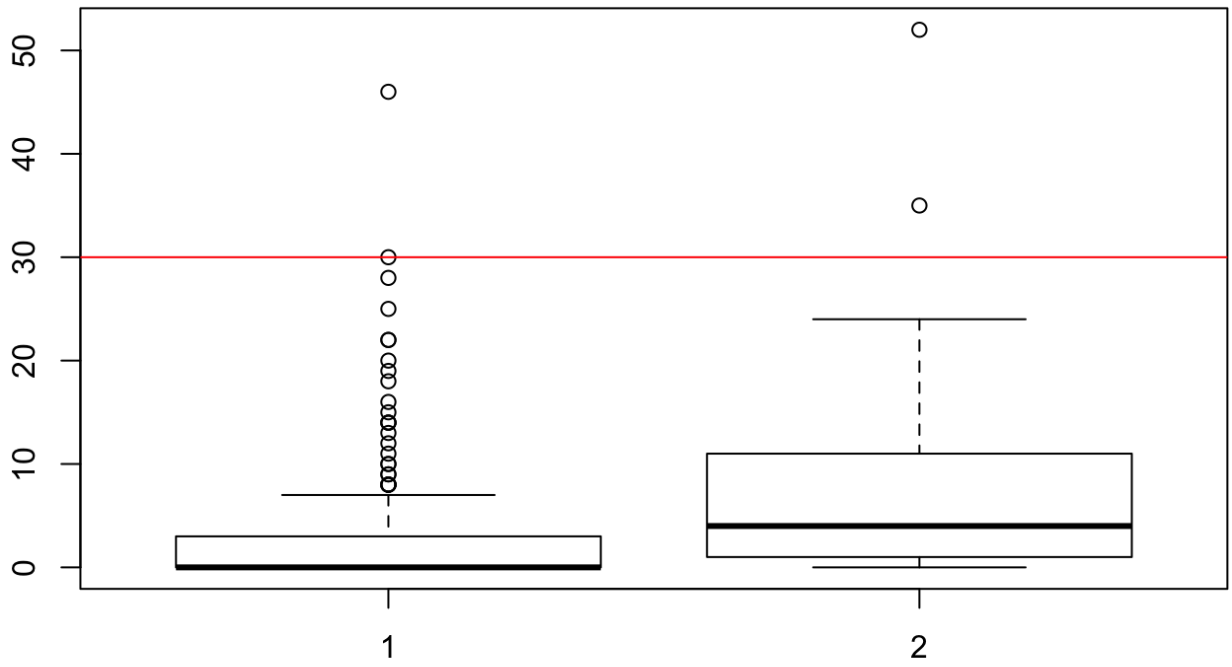



# Suppression des données aberrantes

D'après les précédents histogrammes, on se demande si le jeu de donnée ne contient pas de données aberrantes, notamment dans le nombre de nodes. Sans la connaissance ou l'aide d'un spécialiste, c'est impossible d'être sûr si ces mesures sont bien erronées. Pour ce faire, il faut analyser d'un peu plus près cette variable.

Hide

```
boxplot(haberman$nodes~haberman$survivalStatus)
abline(h=30,col="red")
```




La ligne rouge est ce qu'on a consideré la frontière (f=30) au-dessus de laquelle les valeurs sont des points extrêmes très éloignés du reste. Donc, on a décidé de traiter ces données considerées aberrantes en les supprimmant, car elles ne représentent que trois observations.

Hide

```
haberman<-haberman[haberman$nodes<=30,]
```

# Statistiques univariées

La moyenne de l'âge est en général 52.46, et cela ne change pas beaucoup par rapport à la valeur du **survivalStatus**. La moyenne de l'année de l'opération ne varie pas non plus. En revanche, le nombre de nodes varient. Sa moyenne passe de 2.598, pour ce qui ont survécu au moins 5 ans après l'opération, à 6,544 pour ce qui n'ont pas survécu. On se demande donc, quelles sont les corrélations entre les variables ?

## All data

Hide

```
#all data
summary(haberman)
```

```
      age        yearOfOperation     nodes        survivalStatus
 Min.   :30.00   Min.   :58.00   Min.   : 0.000   Min.   :1.000
 1st Qu.:44.00   1st Qu.:60.00   1st Qu.: 0.000   1st Qu.:1.000
 Median :52.00   Median :63.00   Median : 1.000   Median :1.000
 Mean   :52.46   Mean   :62.86   Mean   : 3.627   Mean   :1.261
 3rd Qu.:61.00   3rd Qu.:65.50   3rd Qu.: 4.000   3rd Qu.:2.000
 Max.   :83.00   Max.   :69.00   Max.   :30.000   Max.   :2.000
```

## Filtré par les personnes qui sont mortes

Hide

```
#Survived
summary(haberman[haberman$survivalStatus == 1,])
```

```
      age         yearOfOperation     nodes         survivalStatus
 Min.   :30.00   Min.   :58.00   Min.   : 0.000   Min.   :1
 1st Qu.:43.00   1st Qu.:60.00   1st Qu.: 0.000   1st Qu.:1
 Median :52.00   Median :63.00   Median : 0.000   Median :1
 Mean   :52.01   Mean   :62.84   Mean   : 2.598   Mean   :1
 3rd Qu.:60.00   3rd Qu.:66.00   3rd Qu.: 3.000   3rd Qu.:1
 Max.   :77.00   Max.   :69.00   Max.   :30.000   Max.   :1
```

### Filtré par les personnes qui ont survécu

Hide

```
#Died
summary(haberman[haberman$survivalStatus == 2,])
```

```
      age         yearOfOperation     nodes         survivalStatus
 Min.   :34.00   Min.   :58.0    Min.   : 0.000   Min.   :2
 1st Qu.:46.00   1st Qu.:59.5    1st Qu.: 1.000   1st Qu.:2
 Median :53.00   Median :63.0    Median : 4.000   Median :2
 Mean   :53.75   Mean   :62.9    Mean   : 6.544   Mean   :2
 3rd Qu.:61.00   3rd Qu.:65.0    3rd Qu.:11.000   3rd Qu.:2
 Max.   :83.00   Max.   :69.0    Max.   :24.000   Max.   :2
```

# Corrélation entre les variables

La matrice de corrélation confime les hypothèses qu'on a créé à partir des statistiques univarées. Il y a une corrélation faible (presque nulle) entre les variables. La plus expressive mais toujours pas très importante est la corrélation entre le nombre de nodes et le **survivalStatus**. On l'analyse premièrement.

Ci-desouss la matrice de corrélation.

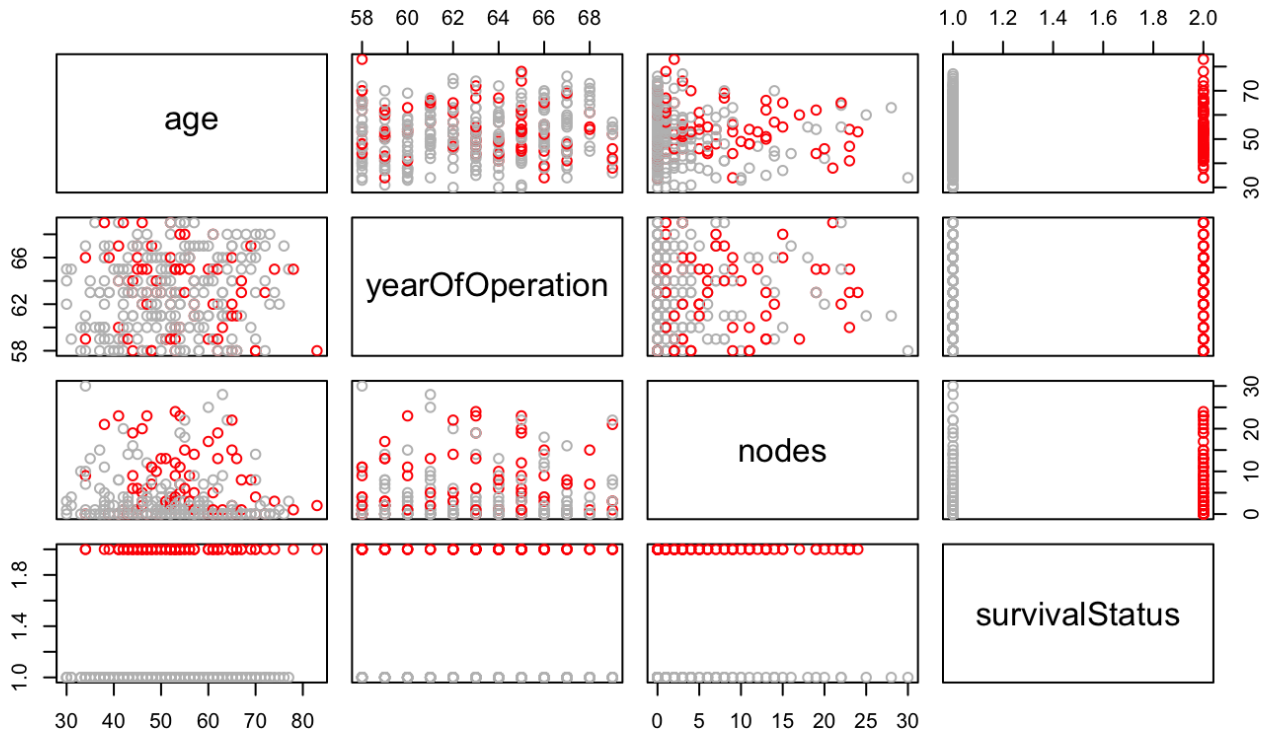
Hide

```
#no strong linear correlation
cor(haberman, method = "spearman")
```

```
                      age yearOfOperation       nodes survivalStatus
age              1.00000000     0.086687417 -0.09949479    0.058380007
yearOfOperation  0.08668742     1.000000000 -0.03403204    0.004919198
nodes           -0.09949479    -0.034032041  1.00000000    0.319068735
survivalStatus   0.05838001     0.004919198  0.31906873    1.000000000
```

Hide

```
colors<-rep("black",nrow(haberman))
colors[haberman$survivalStatus == 1] <- "grey"
colors[haberman$survivalStatus == 2] <- "red"
pairs(haberman, col=colors)
```




# Relation entre nodes et statut de survie

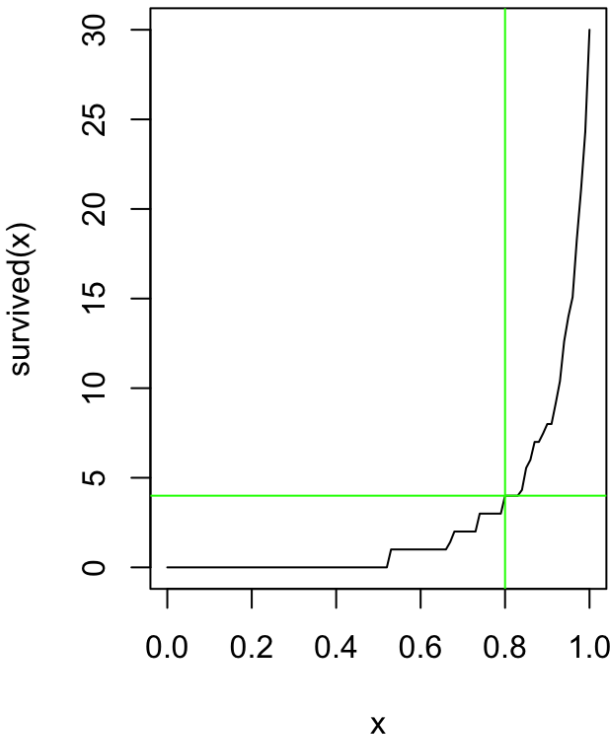
Ici on conclue que le nombre de nodes est important. Pour les personnes qui ont survécu au moins 5 après l'opération, leurs nombre de noeuds était inférieur à 4 nodes. Tandis que les personnes qui n'ont pas survécu plus de 5 ans, leurs nombre de noeuds approchait en moyenne 13 nodes. Seulement 20% de ceux qui ont survécus avaient plus de 4 noeuds. Par contre, 50% de ceux que n'ont pas survécu avaient plus de 4 noeuds.

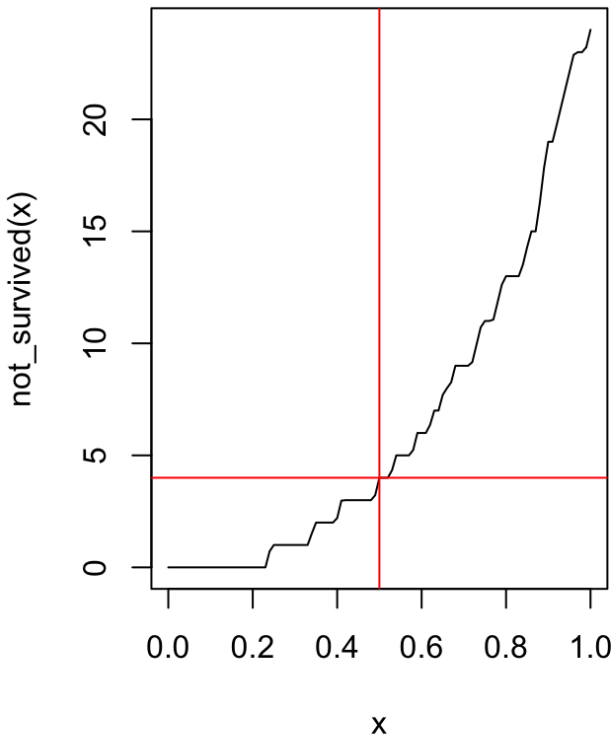
Hide

```
layout(matrix(c(1,2),ncol=2))
survived<-function(x) {
  return(quantile(haberman$nodes[haberman$survivalStatus == 1], probs=x))
}
not_survived<-function(x) {
  return(quantile(haberman$nodes[haberman$survivalStatus == 2], probs=x))
}
curve(survived,from=0,to=1)
abline(h=4,v=0.8, col="green")
```

Hide

```
curve(not_survived,from=0,to=1)
abline(h=4,v=0.5, col="red")
```



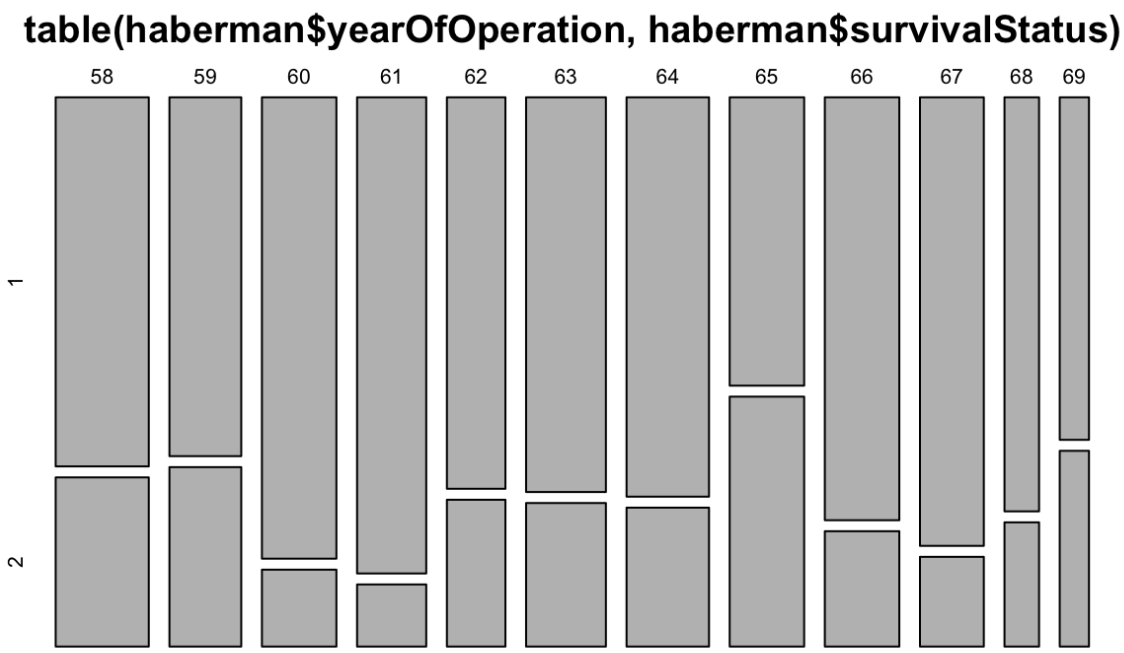




# Relation entre l'âge et le statut de survie

Entre l'âge et le statut de survie, on ne distingue pas de corrélation. Le graphe ci-dessous nous donne une image plus visuel des gens qui ont survécu plus de 5 ans et ceux qui ont survécu moins de 5 ans par rapport à l'âge.

Hide

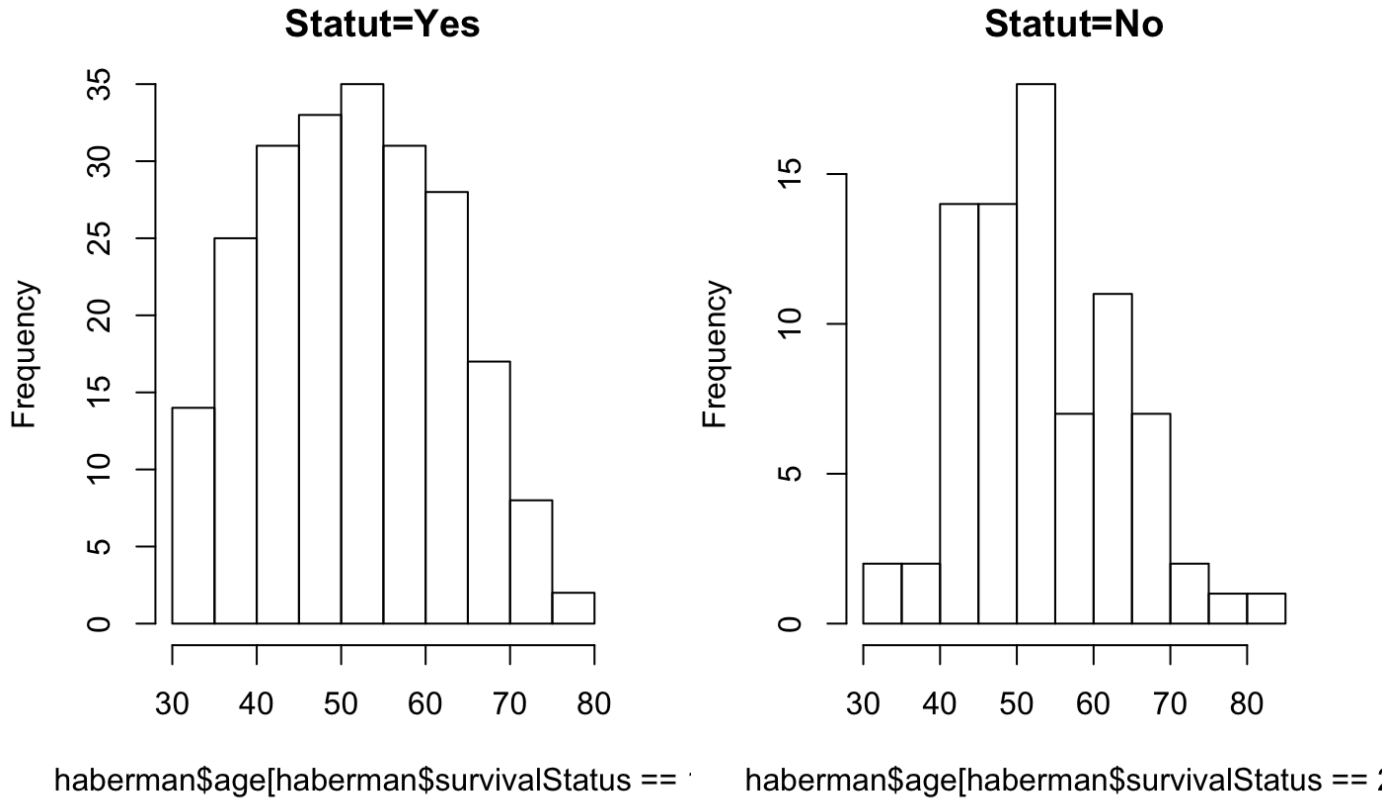
```
plot(table(haberman$yearOfOperation,haberman$survivalStatus))
```




On déduit aussi grâce aux graphes suivant qu'il n'y a pas de relation entre l'âge et le statut de survie, la matrice de corrélation faite précédément confirme cette conclusion. Les deux variables (Statut=1 et Statut=2) suivent ici une loi normale.

Hide

```
layout(matrix(c(1,2),ncol=2))
hist(haberman$age[haberman$survivalStatus==1],main = "Statut=Yes")
hist(haberman$age[haberman$survivalStatus==2],main = "Statut=No")
```



##Régression linéaire simple

En faisant une régression linéaire entre le statut de survie et le nombre de nodes.
On observe que la qualité du modèle est très médiocre, car il explique que 8% de la variabilité des status de survie.

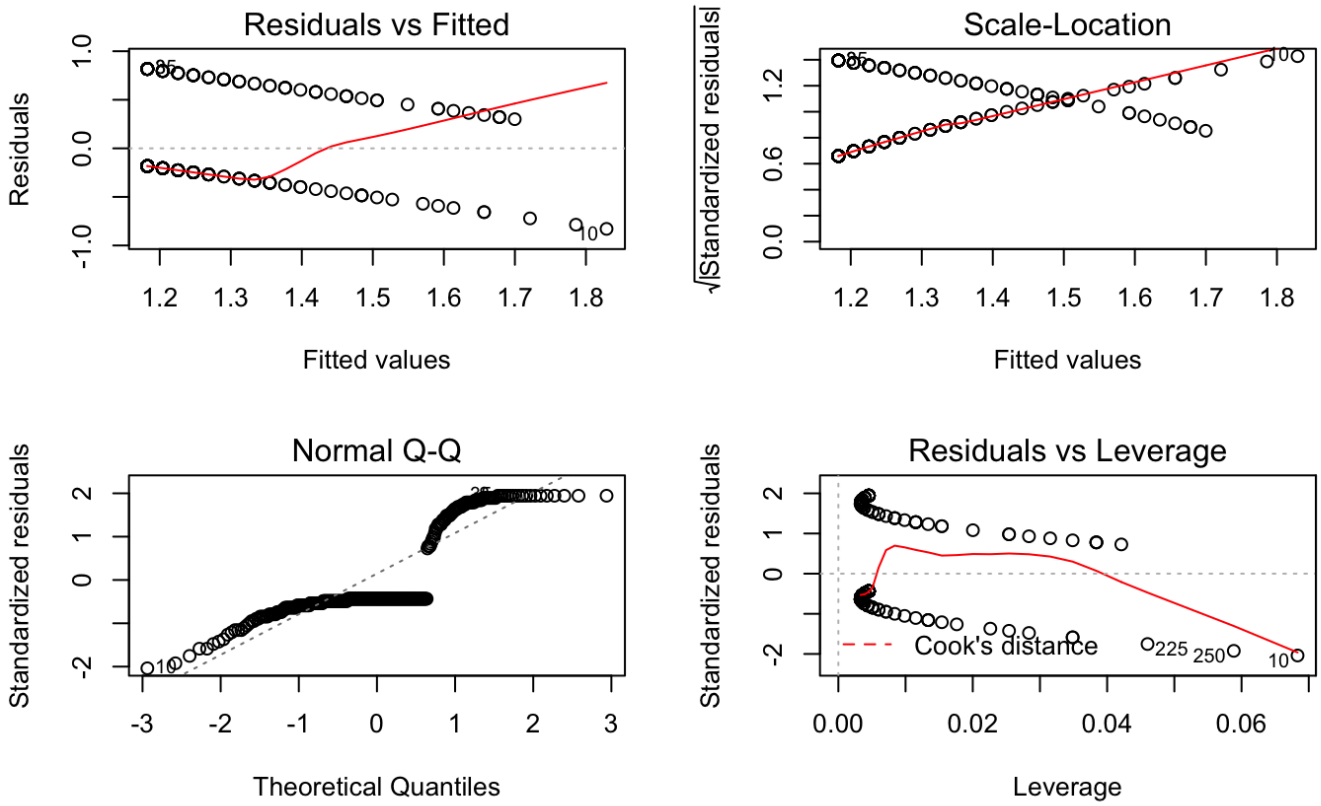
Hide

```
model=lm(haberman$survivalStatus~haberman$nodes)
summary(model)
```

```

Call:
lm(formula = haberman$survivalStatus ~ haberman$nodes)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8290 -0.2041 -0.1826  0.3218  0.8174

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.182567   0.028360  41.699  < 2e-16 ***
haberman$nodes 0.021549   0.004074   5.289 2.37e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4213 on 301 degrees of freedom
Multiple R-squared:  0.08503,   Adjusted R-squared:  0.08199
F-statistic: 27.97 on 1 and 301 DF,  p-value: 2.37e-07
```

Hide

```
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(model)
```




# PCA

On distingue grâce aux graphes ci-dessous, qu'auncune variable n'est corrélée avec les composantes principales. On ne peut pas donc conclure.

Hide

```
install.packages('FactoMineR', dependencies = TRUE)
```

```
trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.5/FactoMineR_1.4
1.tgz'
Content type 'application/x-gzip' length 3625207 bytes (3.5 MB)
==================================================
downloaded 3.5 MB
```

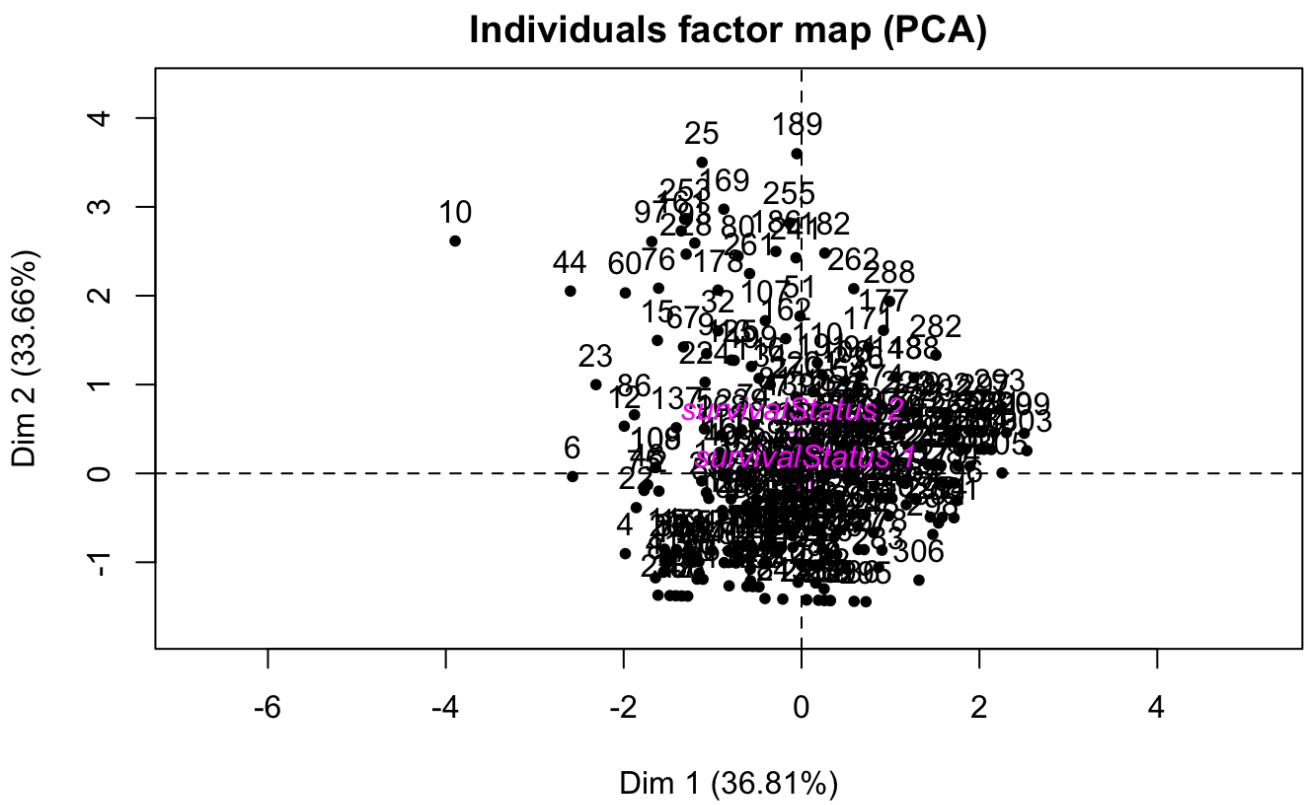
```

The downloaded binary packages are in
    /var/folders/2g/x43p5shj75321bbfljtm_w2w0000gn/T//RtmprbKFLQ/downloaded_packages
```

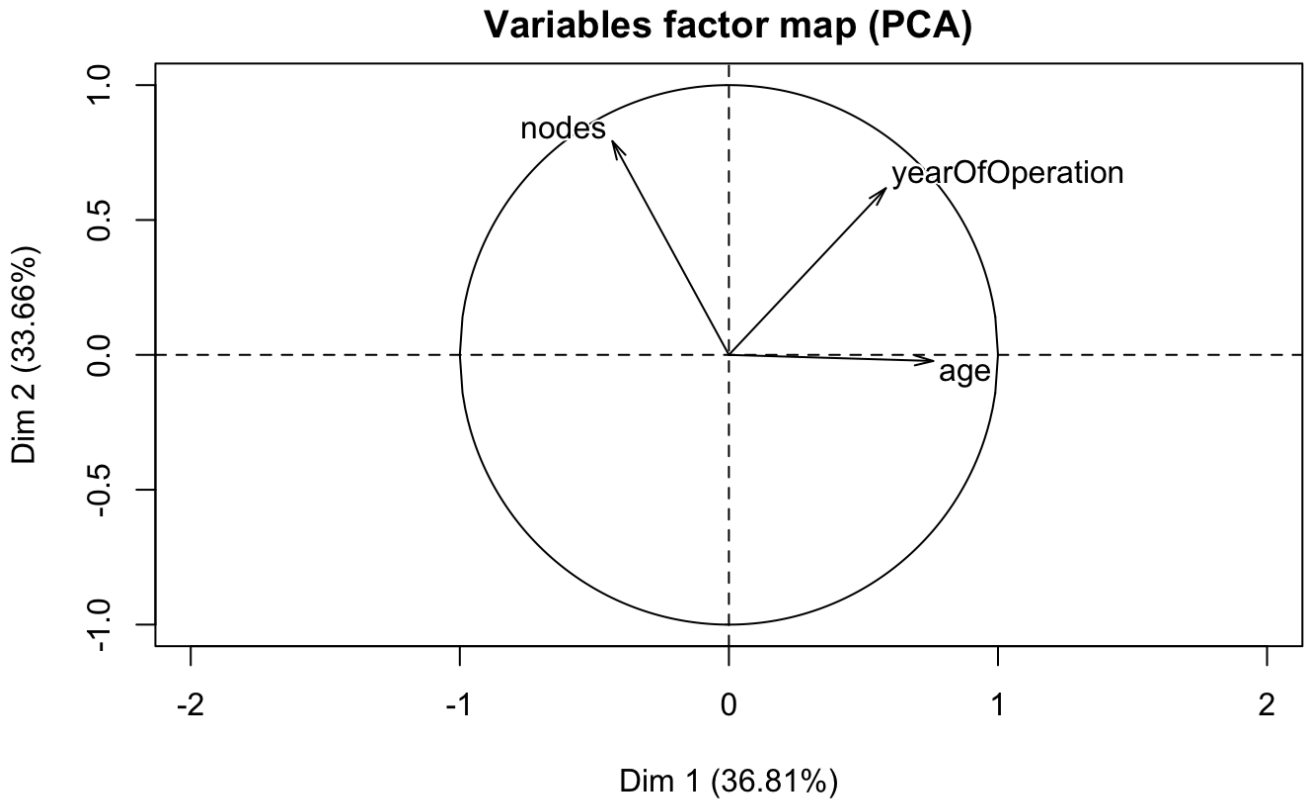
Hide

```
library('FactoMineR')
```

Hide

```
result<-PCA(haberman, scale.unit = TRUE, quali.sup = c(4))
```




#Conclusion

Est ce que l'âge des patiens au moment de l'opération influence le statut de survie ?

Suite à l'analyse de ce jeu de donnée, la réponse est donc non vu l'absence de corrélation.

Est-ce que le nombre de nodes auxiliaires positifs influence de le statut de survie ?

La réponse à cette question est oui, vu que dans ce jeu de données la seule variable qui corrèle positivement avec le statut de survie est le nombre de lymph nodes.

Nous avons maintenant une confirmation de notre intuition. À l'avenir, nous souhaiterons avoir l'avis d'un expert pour savoir si grâce à son expertise et les résultats obtenus, nous pouvons prédire le statut de survie.