PROJET MACHINE LEARNING

Saâd Qriouet n° étudiant : 20171683 M1 MINT

11/14/2020

1. Vos données

Chargement des librairies et du jeu de données

```
library(MASS)
library(FactoMineR)
library(ggplot2)
library(ggpubr)
library(qcc)
```

```
## Package 'qcc' version 2.7
```

```
## Type 'citation("qcc")' for citing this R package in publications.
```

```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(rpart)
library(rpart.plot)
library(lattice)
library(caret)
library(e1071)

data(crabs)
```

2. Exploration non supervisée avec l'analyse en composantes principales

A. Presentation du jeu de données :

```
dim(crabs)
```

```
## [1] 200   8
```

```
names(crabs)
```

```
## [1] "sp"    "sex"   "index" "FL"    "RW"    "CL"    "CW"    "BD"
```

```
summary(crabs)
```

```
##  sp      sex         index            FL              RW              CL
##  B:100   F:100   Min.   : 1.0   Min.   : 7.20   Min.   : 6.50   Min.   :14.70
##  O:100   M:100   1st Qu.:13.0   1st Qu.:12.90   1st Qu.:11.00   1st Qu.:27.27
##                  Median :25.5   Median :15.55   Median :12.80   Median :32.10
##                  Mean   :25.5   Mean   :15.58   Mean   :12.74   Mean   :32.11
##                  3rd Qu.:38.0   3rd Qu.:18.05   3rd Qu.:14.30   3rd Qu.:37.23
##                  Max.   :50.0   Max.   :23.10   Max.   :20.20   Max.   :47.60
##        CW              BD
##  Min.   :17.10   Min.   : 6.10
##  1st Qu.:31.50   1st Qu.:11.40
##  Median :36.80   Median :13.90
##  Mean   :36.41   Mean   :14.03
##  3rd Qu.:42.00   3rd Qu.:16.60
##  Max.   :54.60   Max.   :21.60
```

Le jeu de données contient 5 mesures morphologiques sur 200 crabes de deux couleurs et sexes différents, de l'espèce Leptograpsus variegatus collectées à Fremantle en Australie occidentale.

Nombre de variables : 8

La variable index est la variable identifiante, on ne va pas y preter attention. Hormis la variable index, il y a deux typologies concernant les variables : quantitatives et qualitatives.

Les variables quantitatives sont : FL, RW, CL, CW et BD donc 5.

Les variables qualitatives sont : sp et sex donc 2.

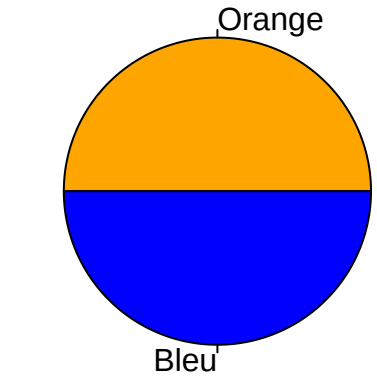
B. Brève description du jeu de données :

a) Variable sp :

```
effectifs <- table(crabs$sp)
frequences <- effectifs/sum(effectifs)
tableau <- cbind(effectifs,frequences)
tableau
```

```
##   effectifs frequences
## B       100        0.5
## O       100        0.5
```

```
pie(table(crabs$sp),labels=c("Orange","Bleu"), col=c("orange","blue"), main="Répartition des espèces", 
```



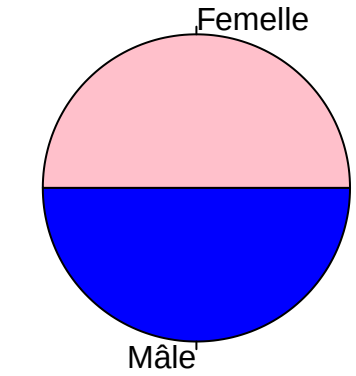
Il y a autant de crabes bleus que de crabes oranges.

b) Variable sex :

```
effectifs <- table(crabs$sex)
frequences <- effectifs/sum(effectifs)
tableau <- cbind(effectifs,frequences)
tableau
```

```
##   effectifs frequences
## F       100        0.5
## M       100        0.5
```

```
pie(table(crabs$sex), labels=c("Femelle","Mâle"), col=c("pink","blue"), main="Répartition des sexes", c
```



Il y a autant de crabes mâles que de crabes femelles.

c) Variable FL :

```
# Calcul de moyenne et variance :
mean(crabs$FL)
```

```
## [1] 15.583
```

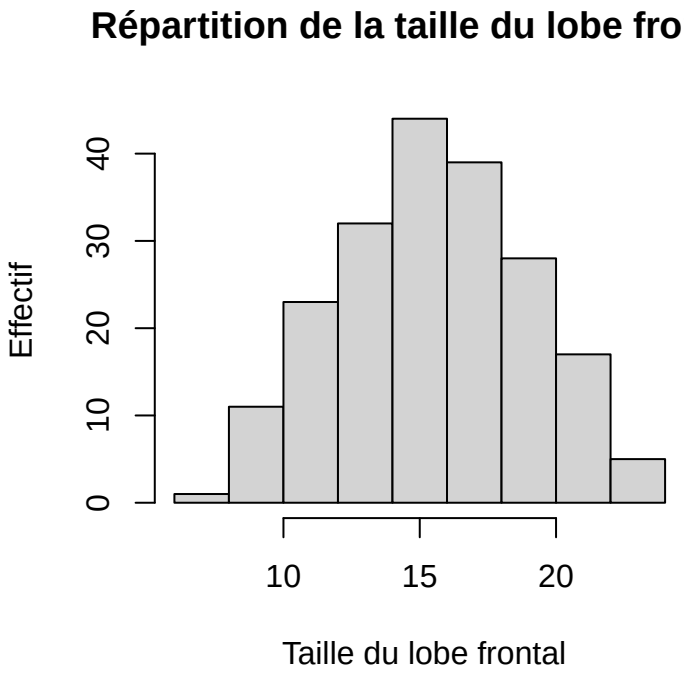
```
var(crabs$FL)
```

```
## [1] 12.2173
```

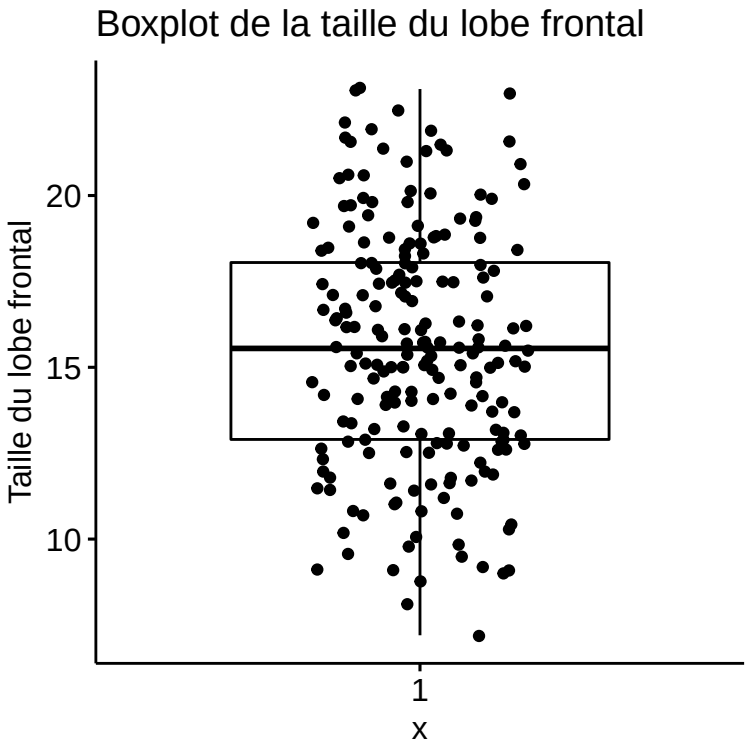
```
# Construction du tableau d'effectifs,fréquences et de fréquences cumulées
reshist <- hist(crabs$FL,plot=FALSE,right=FALSE)
effectifs <- reshist$counts
frequences <- effectifs/sum(effectifs)
frequencescum <- cumsum(frequences)
tableau <- cbind(effectifs,frequences,frequencescum)
tableau
```

```
##       effectifs frequences frequencescum
##  [1,]         1      0.005         0.005
##  [2,]        11      0.055         0.060
##  [3,]        21      0.105         0.165
##  [4,]        31      0.155         0.320
##  [5,]        47      0.235         0.555
##  [6,]        35      0.175         0.730
##  [7,]        31      0.155         0.885
##  [8,]        18      0.090         0.975
##  [9,]         5      0.025         1.000
```

```
# Histogramme
hist(crabs$FL, main = "Répartition de la taille du lobe frontal", xlab = " Taille du lobe frontal", ylab
```




```
# Box-plot
ggboxplot(crabs$FL, main = "Boxplot de la taille du lobe frontal ", ylab = "Taille du lobe frontal", ad
```



La moyenne de la taille du lobe frontal est de 15.583, avec des valeurs allant de 1 à 50 (variance très élevée). On observe quelques points abérrants sur le boxplot.

d) Variable RW :

```
# Calcul de moyenne et variance :
mean(crabs$RW)
```

```
## [1] 12.7385
```

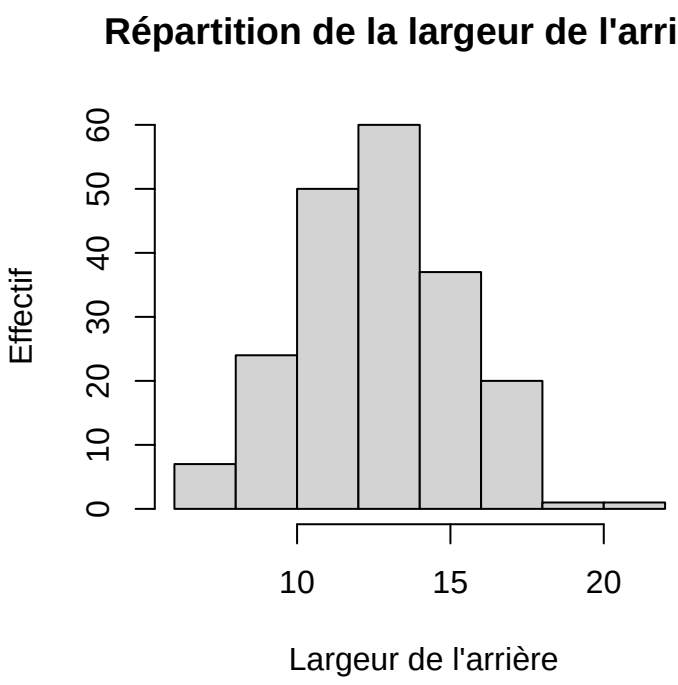
```
var(crabs$RW)
```

```
## [1] 6.622078
```

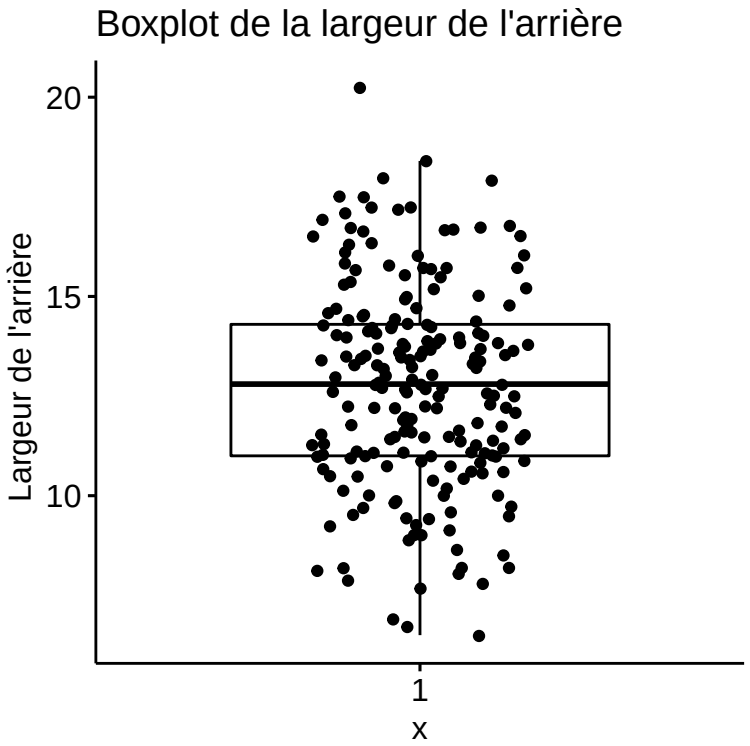
```
# Construction du tableau d'effectifs,fréquences et de fréquences cumulées
reshist <- hist(crabs$RW,plot=FALSE,right=FALSE)
effectifs <- reshist$counts
frequences <- effectifs/sum(effectifs)
frequencescum <- cumsum(frequences)
tableau <- cbind(effectifs,frequences,frequencescum)
tableau
```

```
##      effectifs frequences frequencescum
## [1,]         6      0.030         0.030
## [2,]        22      0.110         0.140
## [3,]        52      0.260         0.400
## [4,]        57      0.285         0.685
## [5,]        39      0.195         0.880
## [6,]        21      0.105         0.985
## [7,]         2      0.010         0.995
## [8,]         1      0.005         1.000
```

```
# Histogramme
hist(crabs$RW, main = "Répartition de la largeur de l'arrière", xlab = " Largeur de l'arrière", ylab =
```




```
# Box-plot
ggboxplot(crabs$RW, main = "Boxplot de la largeur de l'arrière ", ylab = "Largeur de l'arrière ", add =
```



La moyenne de la largeur de l'arrière est de 12.7385, avec des valeurs allant de 7.2 à 23.1 (variance élevée). On observe quelques points abérrants sur le boxplot.

e) Variable CL :

```
# Calcul de moyenne et variance :
mean(crabs$CL)
```

```
## [1] 32.1055
```

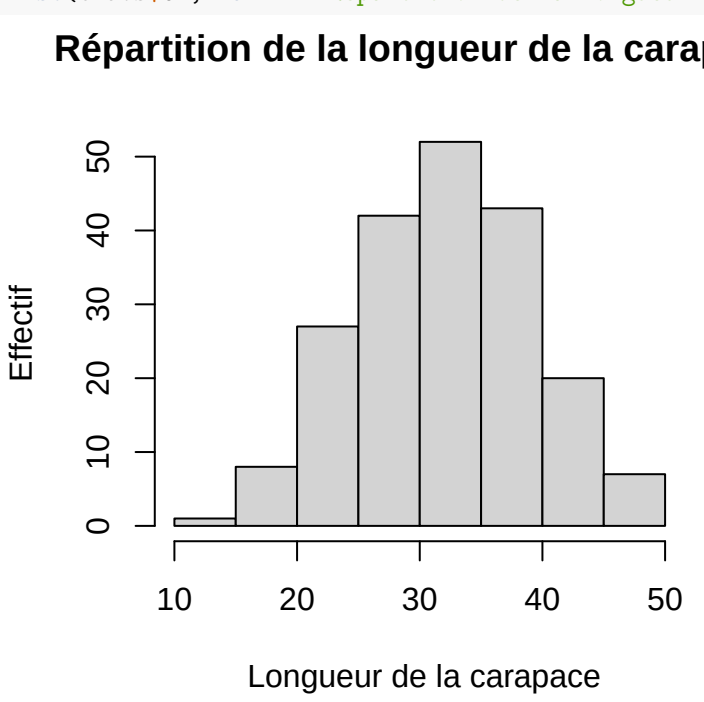
```
var(crabs$CL)
```

```
## [1] 50.67992
```

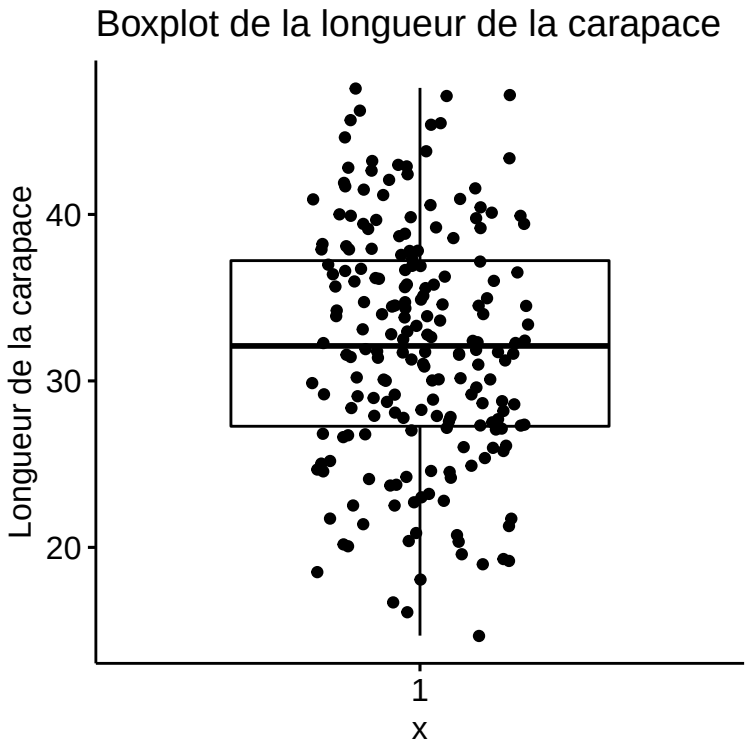
```
# Construction du tableau d'effectifs,fréquences et de fréquences cumulées
reshist <- hist(crabs$CL,plot=FALSE,right=FALSE)
effectifs <- reshist$counts
frequences <- effectifs/sum(effectifs)
frequencescum <- cumsum(frequences)
tableau <- cbind(effectifs,frequences,frequencescum)
tableau
```

```
##      effectifs frequences frequencescum
## [1,]         1      0.005         0.005
## [2,]         8      0.040         0.045
## [3,]        26      0.130         0.175
## [4,]        41      0.205         0.380
## [5,]        53      0.265         0.645
## [6,]        43      0.215         0.860
## [7,]        21      0.105         0.965
## [8,]         7      0.035         1.000
```

```
# Histogramme
hist(crabs$CL, main = "Répartition de la longueur de la carapace", xlab = " Longueur de la carapace", yl
```




```
# Box-plot
ggboxplot(crabs$CL, main = "Boxplot de la longueur de la carapace ", ylab = "Longueur de la carapace", a
```



La moyenne de la longueur de la carapace est de 32.1055, avec des valeurs allant de 14.7 à 47.6 (variance très élevée).

On observe quelques points abérrants sur le boxplot.

f) Variable CW :

```
# Calcul de moyenne et variance :
mean(crabs$CW)
```

```
## [1] 36.4145
```

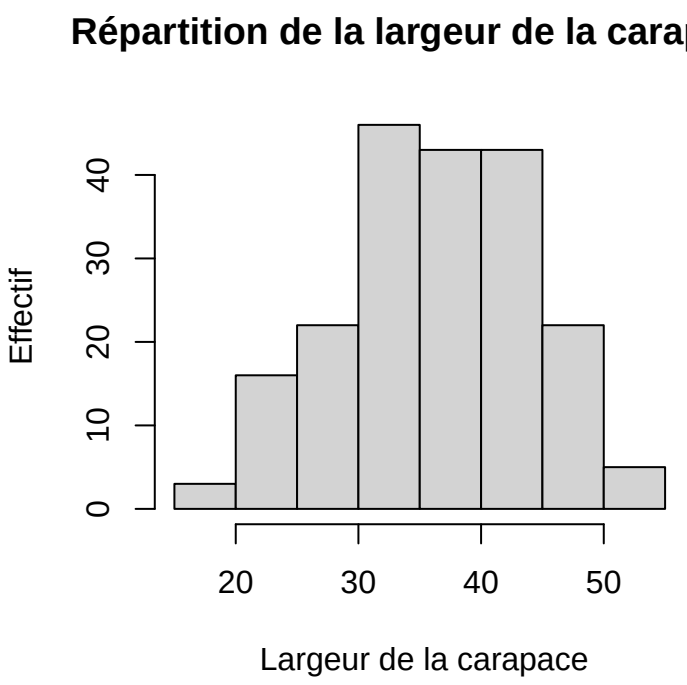
```
var(crabs$CW)
```

```
## [1] 61.96768
```

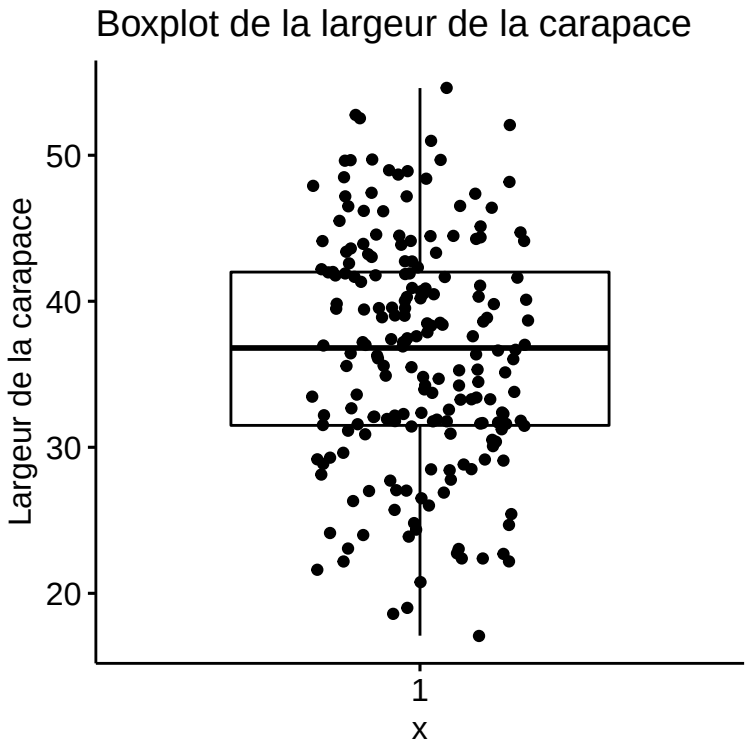
```
# Construction du tableau d'effectifs,fréquences et de fréquences cumulées
reshist <- hist(crabs$CW,plot=FALSE,right=FALSE)
effectifs <- reshist$counts
frequences <- effectifs/sum(effectifs)
frequencescum <- cumsum(frequences)
tableau <- cbind(effectifs,frequences,frequencescum)
tableau
```

```
##      effectifs frequences frequencescum
## [1,]         3      0.015         0.015
## [2,]        16      0.080         0.095
## [3,]        22      0.110         0.205
## [4,]        46      0.230         0.435
## [5,]        42      0.210         0.645
## [6,]        44      0.220         0.865
## [7,]        22      0.110         0.975
## [8,]         5      0.025         1.000
```

```
# Histogramme
hist(crabs$CW, main = "Répartition de la largeur de la carapace", xlab = " Largeur de la carapace", ylab
```




```
# Box-plot
ggboxplot(crabs$CW, main = "Boxplot de la largeur de la carapace", ylab = "Largeur de la carapace", add
```



La moyenne de la largeur de la carapace est de 36.4145, avec des valeurs allant de 17.1 à 54.6 (variance très élevée).

On observe quelques points abérrants sur le boxplot.

g) Variable BD :

```
# Calcul de moyenne et variance :
mean(crabs$BD)
```

```
## [1] 14.0305
```

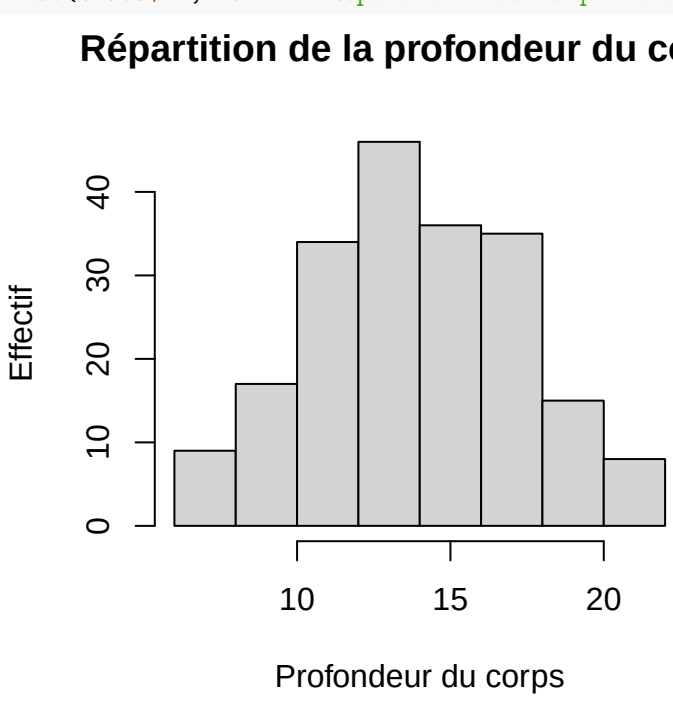
```
var(crabs$BD)
```

```
## [1] 11.72907
```

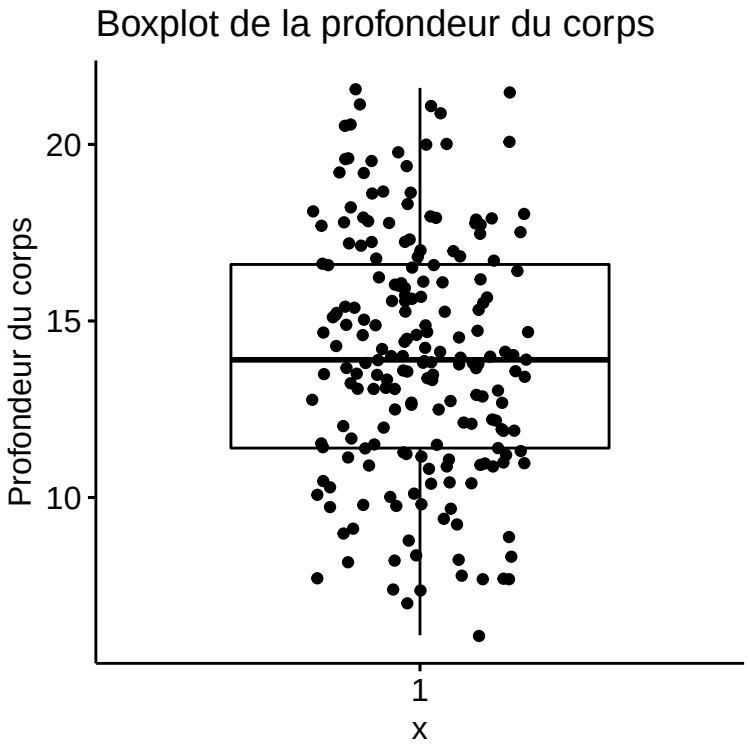
```
# Construction du tableau d'effectifs,fréquences et de fréquences cumulées
reshist <- hist(crabs$BD,plot=FALSE,right=FALSE)
effectifs <- reshist$counts
frequences <- effectifs/sum(effectifs)
frequencescum <- cumsum(frequences)
tableau <- cbind(effectifs,frequences,frequencescum)
tableau
```

```
##      effectifs frequences frequencescum
## [1,]         9      0.045         0.045
## [2,]        16      0.080         0.125
## [3,]        33      0.165         0.290
## [4,]        43      0.215         0.505
## [5,]        39      0.195         0.700
## [6,]        35      0.175         0.875
## [7,]        15      0.075         0.950
## [8,]        10      0.050         1.000
```

```
# Histogramme
hist(crabs$BD, main = "Répartition de la profondeur du corps ", xlab = "Profondeur du corps", ylab = "E
```




```
# Box-plot
ggboxplot(crabs$BD, main = "Boxplot de la profondeur du corps", ylab = "Profondeur du corps", add = "ji
```



La moyenne de la largeur de la carapace est de 14.0305, avec des valeurs allant de 6.1 à 21.6 (variance élevée).

On observe quelques points abérrants sur le boxplot.

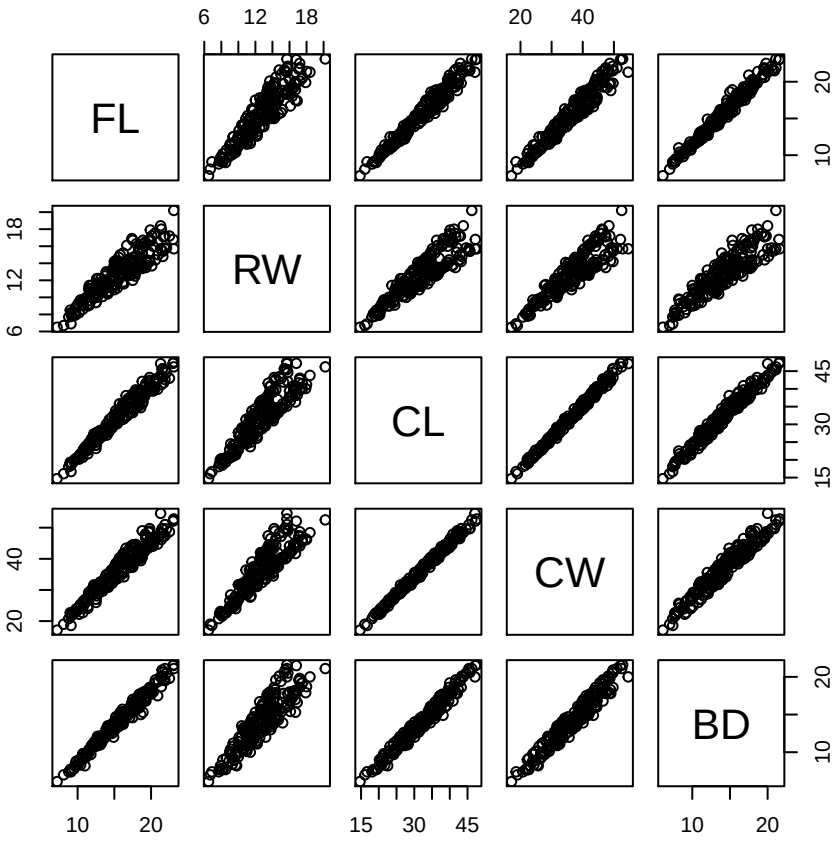
C. Utilisation de l'ACP :

Afin de continuer notre analyse exploratoire, nous allons réaliser une ACP :

a) Scatterplot avant l'ACP :

Avant d'appliquer la méthode de l'ACP, nous allons créer un jeu de données contenant uniquement les variables quantitatives et y réaliser un scatterplot pour observer les liens entre les variables :

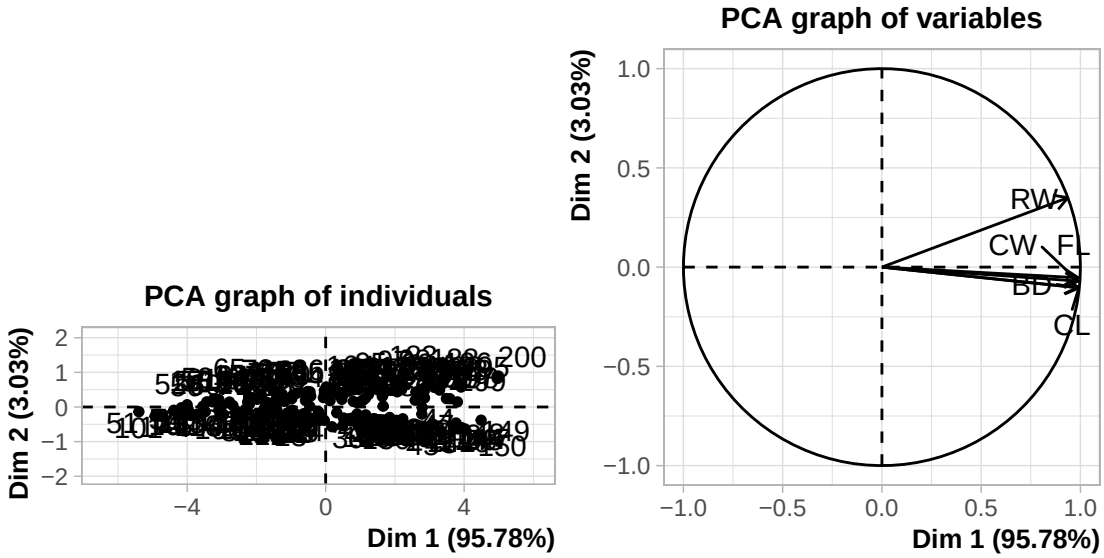
```
crabsquant <- crabs[,4:8]
pairs(crabsquant)
```



b) Première ACP :

Nous allons réaliser une ACP (basique) avec ce nouveau jeu de données :

```
res_pca <- PCA(crabsquant)
```




```
summary(res_pca)
```

```
##
## Call:
## PCA(X = crabsquant)
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5
## Variance               4.789   0.152   0.047   0.011   0.002
## % of var.             95.777   3.034   0.933   0.223   0.034
## Cumulative % of var.  95.777  98.810  99.743  99.966 100.000
##
## Individuals (the 10 first)
##        Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr
## 1   |  4.937 | -4.928  2.535  0.996 | -0.268  0.238  0.003 | -0.122  0.160
## 2   |  4.387 | -4.386  2.009  0.999 | -0.094  0.029  0.000 | -0.039  0.017
## 3   |  4.133 | -4.129  1.780  0.998 | -0.169  0.094  0.002 |  0.034  0.012
## 4   |  3.892 | -3.884  1.575  0.996 | -0.246  0.199  0.004 |  0.015  0.002
## 5   |  3.841 | -3.834  1.535  0.996 | -0.224  0.166  0.003 | -0.015  0.002
## 6   |  2.962 | -2.953  0.910  0.994 | -0.220  0.160  0.006 |  0.038  0.016
## 7   |  2.680 | -2.678  0.749  0.999 |  0.039  0.005  0.000 |  0.082  0.072
## 8   |  2.575 | -2.548  0.678  0.979 | -0.363  0.435  0.020 |  0.063  0.042
## 9   |  2.593 | -2.585  0.698  0.994 | -0.117  0.045  0.002 |  0.062  0.042
## 10  |  2.213 | -2.206  0.508  0.994 |  0.079  0.021  0.001 |  0.157  0.264
##       cos2
## 1    0.001 |
## 2    0.000 |
## 3    0.000 |
## 4    0.000 |
## 5    0.000 |
```

```
## 6   0.000 |
## 7   0.001 |
## 8   0.001 |
## 9   0.001 |
## 10  0.005 |
##
## Variables
##       Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
## FL |  0.989 20.434  0.979 | -0.054  1.893  0.003 | -0.115 28.172  0.013 |
## RW |  0.937 18.325  0.878 |  0.350 80.664  0.122 |  0.003  0.014  0.000 |
## CL |  0.992 20.538  0.984 | -0.104  7.195  0.011 |  0.067  9.590  0.004 |
## CW |  0.987 20.350  0.975 | -0.070  3.261  0.005 |  0.141 42.585  0.020 |
## BD |  0.987 20.352  0.975 | -0.103  6.987  0.011 | -0.096 19.639  0.009 |
```

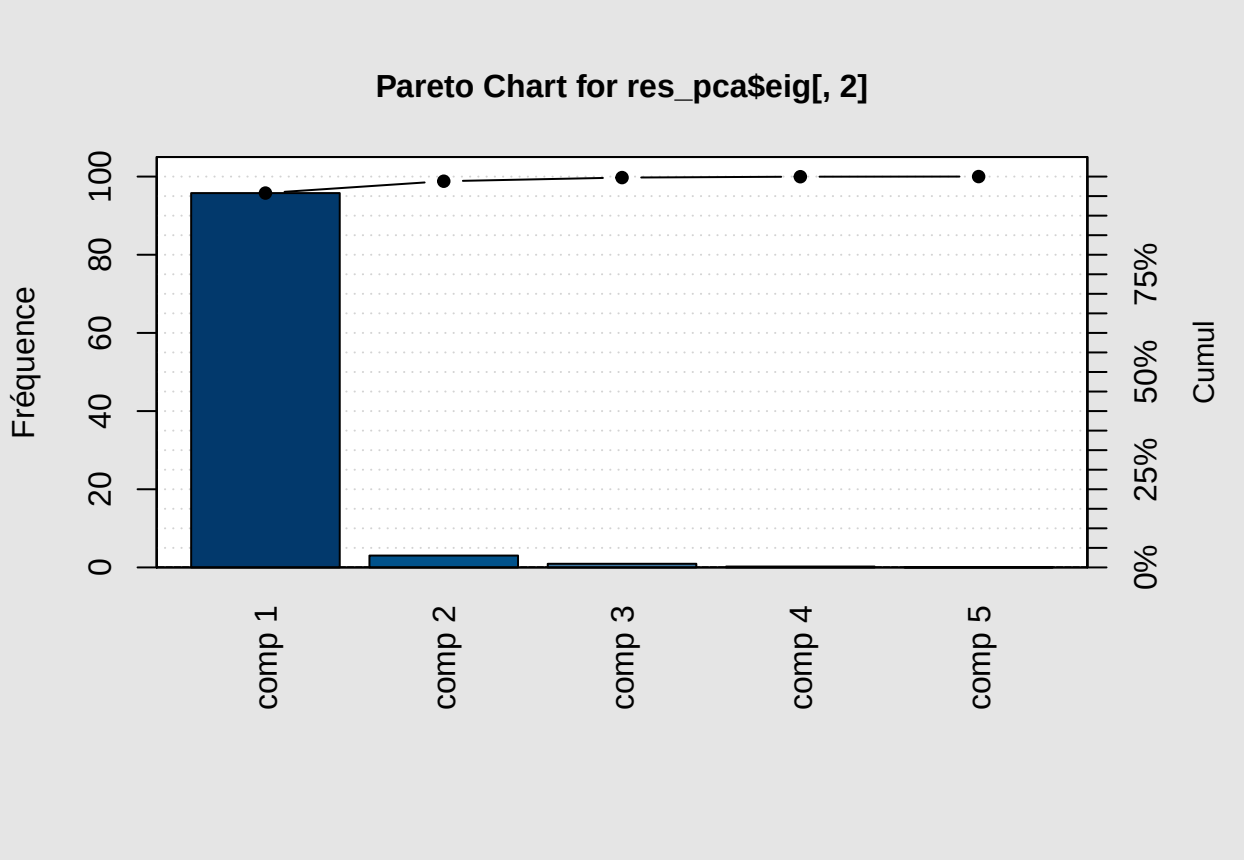
Tout d'abord, nous voyons que le graphique des individus est illisible en partie a cause des libéllés des individus se superposent.

Ensuite, à propos du cercle des corrélations, nous n'avons pas d'informations sur les flèches (quelle flèche correspond à quelle variable ?) et on observe aussi des chevauchements entre les variables, ce qui rend le graphique difficilement lisible.

Nous allons choisir les dimensions et contrôler la pertinence de l'ACP avec le tableau des variances expliquées et un diagramme de Pareto.

Analyse du tableau des variances expliquées et affichage du diagramme de Pareto :

```
pareto.chart(res_pca$eig[,2], cumperc = seq(0, 100, by = 5), ylab = "Fréquence", ylab2 = "Cumul")
```




```
##
## Pareto chart analysis for res_pca$eig[, 2]
##             Frequency    Cum.Freq.   Percentage Cum.Percent.
##   comp 1  95.77669569  95.77669569  95.77669569  95.77669569
##   comp 2   3.03370413  98.81039982   3.03370413  98.81039982
##   comp 3   0.93265948  99.74305930   0.93265948  99.74305930
##   comp 4   0.22270714  99.96576645   0.22270714  99.96576645
##   comp 5   0.03423355 100.00000000   0.03423355 100.00000000
```

On observe que les 2 premières composantes couvrent plus de 98% (dont RW avec 95% de variance expliquée) de la variabilité, on peut donc réduire le nombre de dimensions des données à seulement 2 au lieu de 5

c) Nouvelle ACP avec les modifications :

Pour différencier les individus en fonction de leur espece et leur sexe, nous allons réintegrer les variables "sex" et "sp" en tant que variables supplémentaires (à l'aide de la fonction "cbind()"), ainsi nous aurons 4 couleurs et symboles pour différencier les types de crabes que nous détaillerons plus tard.

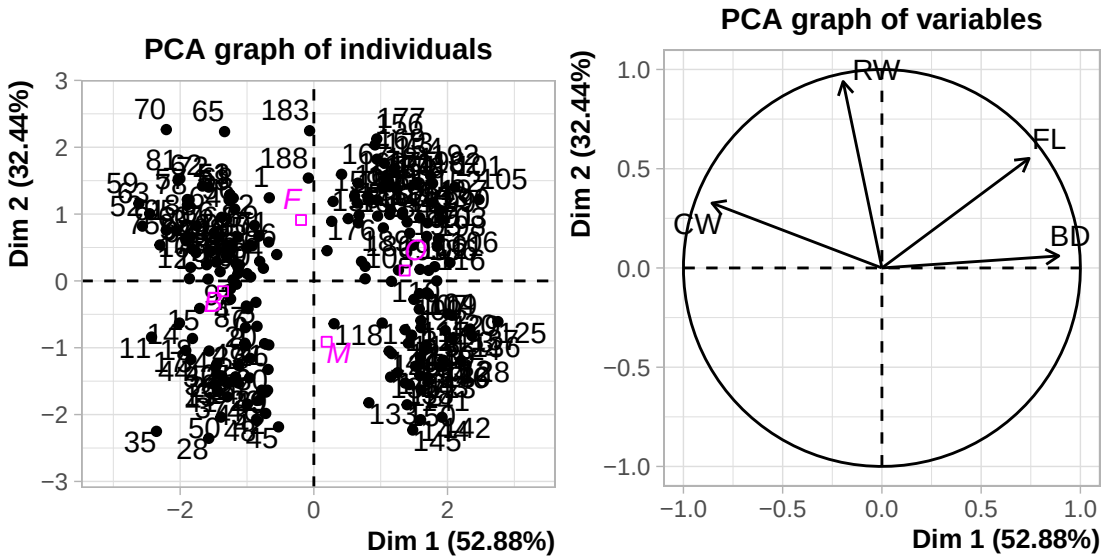
Normalisation des variables quantitatives sur le jeu de données en fonction de la variable CL :

```
crabsquant2 <- crabsquant[,c(1,2,4,5)]/crabsquant$CL
```

Réintégration des variables qualitatives :

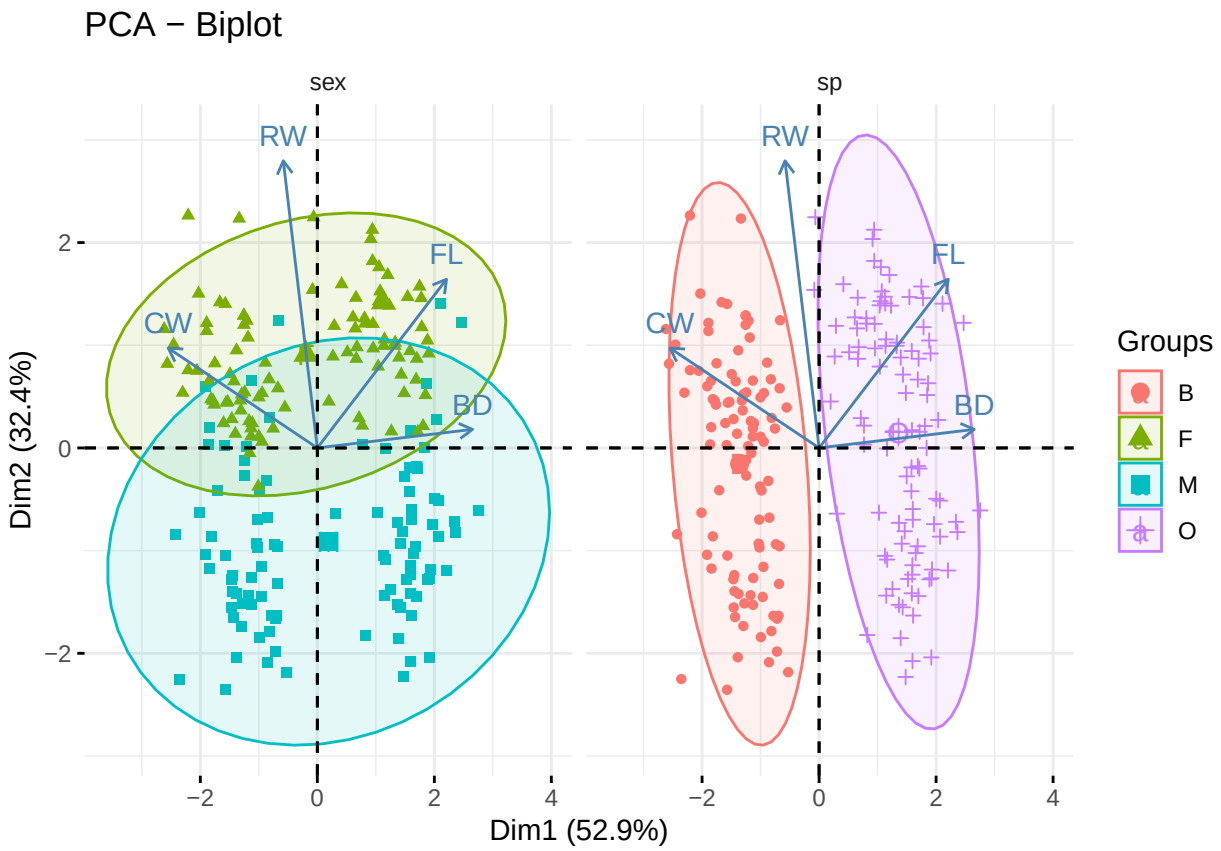
```
crabs2 <- cbind(crabs[1:2],crabsquant2)
```

```
res.pca2 <- PCA(crabs2,quali.sup = 1:2)
```



Affichage du cercle de corrélation sur le graphique des individus :

```
fviz_pca_biplot(res.pca2, habillage = 1:2, label = "var", addEllipses=TRUE, ellipse.level=0.95)
```



Ainsi avec les modifications réalisées, nous avons le nouvel ACP en fonction des variables quantitatives normalisées ci-dessus avec le code couleur suivant :

- Vert triangle : crabe femelle
- Vert carré : crabe male
- Bleu plus : crabe orange
- Rouge cercle : crabe bleu

Nous remarquons que cette nouvelle ACP est bien meilleure que la première, nous avons une meilleure précision et clarté concernant la représentation des graphiques. Nous avons réussis a séparer les différents types de crabes comme le montre le biplot.

Concernant le cercle de corrélations :

L'axe des abscisses représente la dimension 1, et différencie les crabes selon la variable sp. En effet, les crabes oranges sont ceux qui ont une valeur élevée en dimension 1 (à droite de la droite verticale), contrairement aux crabes bleus qui une faible valeur dans cette dimension (à gauche de la droite verticale). Ainsi, les crabes oranges ont un rapport entre les variables BD et CL élevé, et un rapport CW/CL faible.

L'axe des ordonnées représente la dimension 2, et différencie les crabes selon leur sexe. En effet, les crabes femelle sont celles qui ont une valeur élevée en dimension 2 (en haut de la droite horizontale), contrairement aux crabes males (meme si c'est moins flagrant). Plus le rapport RW/CL est élevé, plus le crabe sera susceptible d'être une femelle.

Nous avons donc réalisé une ACP permettant de prédire le type d'un crabe en fonction de certaines de ces mesures morphologiques.

3. Prédition avec CART (Arbre de décision)

Ici, nous devons prédire la variable sp qui est une variable qualitative. Nous allons donc procéder à une analyse discriminante à l'aide d'un arbre de classification.

A. Arbre de classification :

Transformation de la variable sp :

```
crabs$sp <- as.factor(crabs$sp)
```

Nous allons exclure la variable index car c'est la variable identifiante, elle ne nous apporte aucune information pertinente, nous allons donc construire un jeu de données sans index :

```
crabs3 <- crabs[,-3]
```

Dans cette partie, nous allons subdiviser le jeu de données en deux echantillons : un d'apprentissage (70% du jeu de base) que l'on va utiliser pour construire l'arbre de décision, et un de test (30%) pour réaliser des prédictions et évaluer les performances du modèle.

Création des echantillons :

```
set.seed(100)
trainIndex <- createDataPartition(crabs3$sp, p=0.7, list=FALSE, times = 1)
print(length(trainIndex)) # taille de l'échantillon d'apprentissage
```

```
## [1] 140
```

```
print(head(trainIndex,10)) # 10 premiers individus de l'apprentissage
```

```
##       Resample1
##  [1,]         1
##  [2,]         2
##  [3,]         3
##  [4,]         4
##  [5,]         5
##  [6,]         7
##  [7,]         9
##  [8,]        10
##  [9,]        12
## [10,]        14
```

```
crabsTrain <- crabs3[trainIndex,]
crabsTest <- crabs3[-trainIndex,]
```

Distribution des espèces dans les echantillons :

```
print(table(crabsTrain$sp))
```

```
##
##  B  O
## 70 70
```

```
print(table(crabsTest$sp))
```

```
##
##  B  O
## 30 30
```

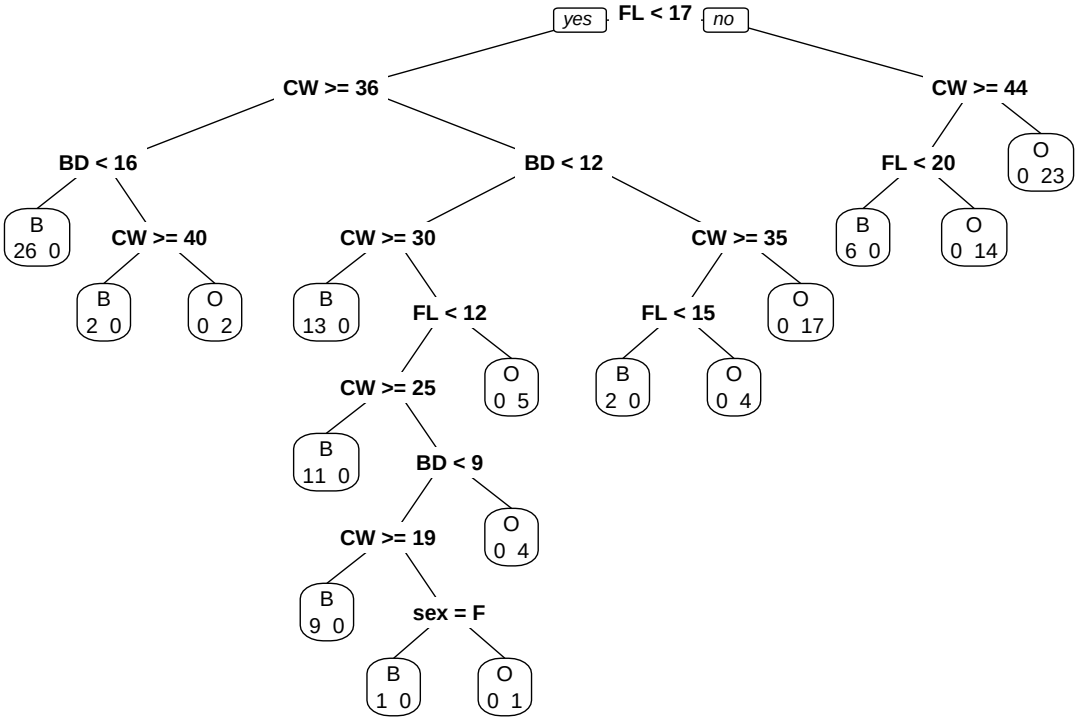
Construction de l'arbre de décision complet

```
modtree <- rpart(sp~., data = crabsTrain, minsplit=2, cp=0) # feuilles contenants au moins 2 observatio
```

Cet arbre complet à été construit de sorte a ce qu'il y ait au moins 2 observations dans chaque feuille, et sans contrainte sur la qualité du découpage.

Affichage de l'arbre complet :

```
prp(modtree, extra = 1, cex = 0.7)
```



Chaque nœud représente une question, la réponse non étant toujours dans la branche droite de l'arbre. Ce dernier comporte 16 feuilles et est assez volumineux.

Chaque feuille est étiquetée par la décision associée (ici B ou O) et par l'effectif classe par espèce des crabes affectés à la feuille. Par exemple, la feuille la plus à gauche classe les crabes en B, avec 26 crabes de cette espèce et aucun de l'espèce O.

Nous allons determiner la compléxité qui minimise l'erreur estimée.

Calcul de la compléxité optimale :

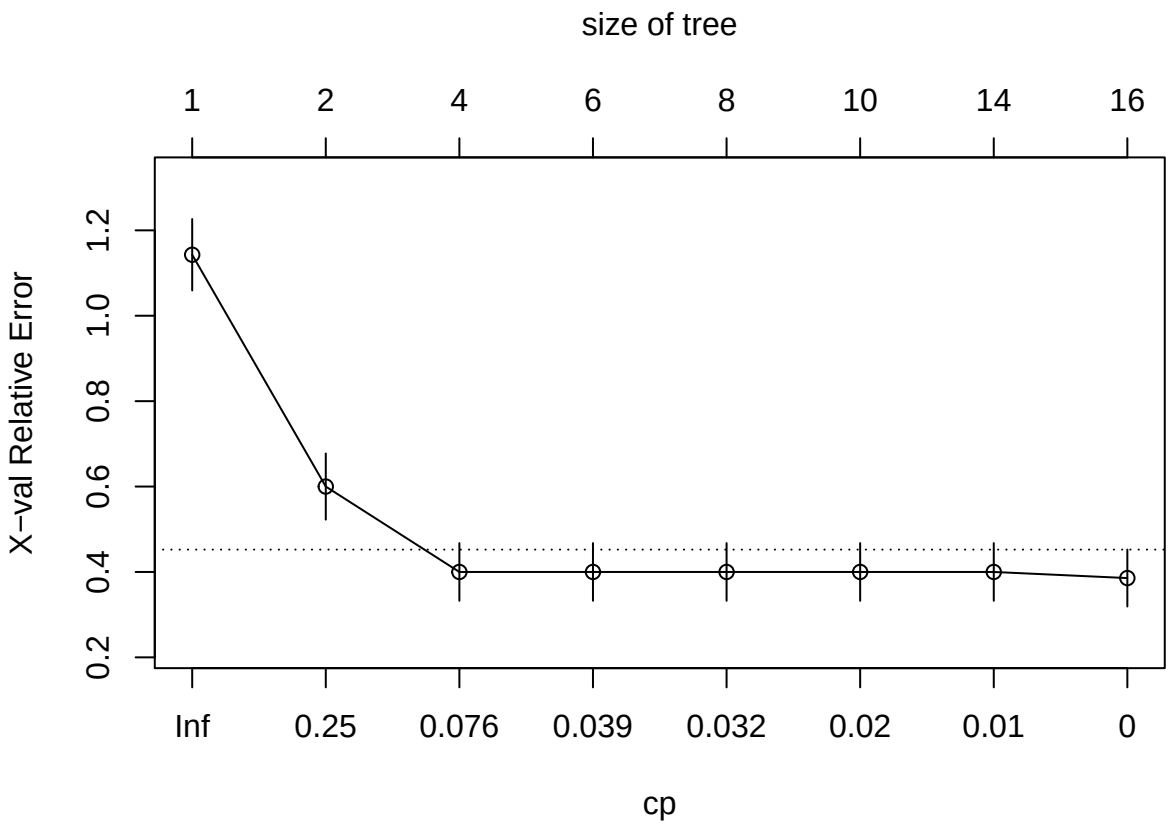
Pour choisir le bon niveau de simplification (ou encore le bon nombre de feuilles), nous allons procèder par validation croisée.

Affichage du tableau de CP :

```
printcp(modtree)
```

```
##
## Classification tree:
## rpart(formula = sp ~ ., data = crabsTrain, minsplit = 2, cp = 0)
##
## Variables actually used in tree construction:
## [1] BD  CW  FL  sex
##
## Root node error: 70/140 = 0.5
##
## n= 140
##
##          CP nsplit rel error  xerror     xstd
## 1 0.4428571      0  1.000000 1.14286 0.083649
## 2 0.1357143      1  0.557143 0.60000 0.077460
## 3 0.0428571      3  0.285714 0.40000 0.067612
## 4 0.0357143      5  0.200000 0.40000 0.067612
## 5 0.0285714      7  0.128571 0.40000 0.067612
## 6 0.0142857      9  0.071429 0.40000 0.067612
## 7 0.0071429     13  0.014286 0.40000 0.067612
## 8 0.0000000     15  0.000000 0.38571 0.066690
```

```
plotcp(modtree)
```



Comme prévu, les performances vont d'abord s'améliorer quand on va augmenter le nombre de feuilles puis vont se dégrader à cause du sur-apprentissage.

La valeur optimale de cp (la complexité qui minimise l'erreur estimée) pour l'élagage est généralement la valeur la plus à gauche en dessous de la droite horizontale (ici, ça semble être 0.076).

B. Arbre simplifié :

Afin de determiner la compléxité optimale, nous aurions pu nous contenter de determiner graphiquement à l'aide du graphique précédent, et de mettre directement la valeur trouvée sur cp (ie : mettre cp = 0.076). Nous allons privilégier une automatisation du processus de la manière suivante :

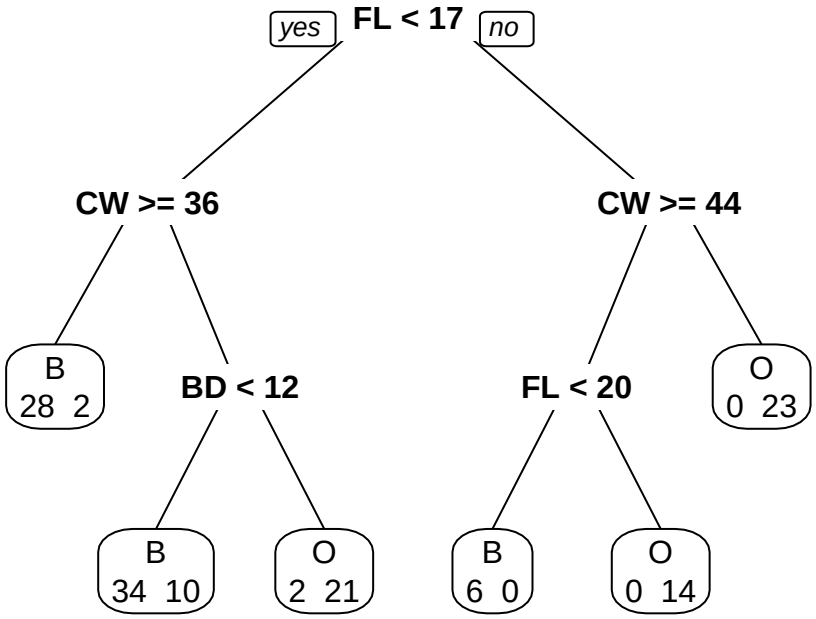
```
cp <- modtree$cptable[,1]
xerror <- modtree$cptable[,4]
xstd <- modtree$cptable[,5]
line <- min(xerror)+xstd[which.min(xerror)]
```

Création de l'arbre simplifié :

```
indexTree <- which(modtree$cptable[,4] < line)[1]
modtree2 <-prune(modtree, cp = sqrt(cp[indexTree]*cp[indexTree+1]))
```

Affichage du nouvel arbre :

```
prp(modtree2, extra = 1)
```



Nous avons donc l'arbre simplifié, il comporte 6 feuilles et est beaucoup plus facile a lire que l'arbre complet.

C. Performances de l'arbre :

Afin d'évaluer les performances de l'arbre, nous allons tester l'arbre sur le jeu de données de test et voir à quel point le modèle est performant.

Prédiction et distribution des espèces prédites sur l'échantillon test :

```
pred <- predict(modtree2, newdata = crabsTest, type = "class")
print(table(pred))
```

```
## pred
##  B  O
## 34 26
```

Le modèle à prédit 34 crabes d'espèce Bleu et 26 Oranges, regardons si la prédiction est bonne ou non.

Affichage de la matrice de confusion et différents indicateurs d'évaluation :

```
mat <- confusionMatrix(data = pred, reference = crabsTest$sp, positive = "O")
print(mat)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  B  O
##          B 28  6
##          O  2 24
##
##                Accuracy : 0.8667
##                  95% CI : (0.7541, 0.9406)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 2.603e-09
##
##                   Kappa : 0.7333
##
##  Mcnemar's Test P-Value : 0.2888
##
##             Sensitivity : 0.8000
##             Specificity : 0.9333
##          Pos Pred Value : 0.9231
##          Neg Pred Value : 0.8235
##              Prevalence : 0.5000
##          Detection Rate : 0.4000
##    Detection Prevalence : 0.4333
##       Balanced Accuracy : 0.8667
##
##        'Positive' Class : O
##
```

On constate que la qualité de prédiction dépend de l'espèce du crabe.

Le taux de prédiction correctes total est est d'environ 87 %.

En effet, sur les 34 crabes Oranges , le taux de prévisions correctes est d'environ 82 %, et sur les 26 crabes Bleus, le taux de prévisions correctes est d'environ 92 %.

Le modèle est tres efficace pour prédire les crabes Bleus, et est assez bon pour prédire les crabes Oranges.

On conclut que le modèle est efficace et réalise de bonnes prédictions sur le jeu d'apprentissage. Pour avoir de meilleurs résultats, on pourrait entrainer le modèle et le tester sur de plus gros jeux de données.