

# [M1MINT] PROJET Machine Learning Exploration non supervisée avec l'algorithmedes k-means & Prédiction avec CART

Mamadou DIOUF

14 novembre 2020

## INTRODUCTION

Les **coyotes**, Canis latrans, les lynx roux(**bobcats**), Lynx rufus, et les renards gris(**gray foxes**), Urocyon cinereoargenteus sont tous des mésoprédateurs mammifères communs dans la Californie côtière et se trouvent en sympatrie dans une grande partie de l'Amérique du Nord. Les excréments produits par ces trois animaux sont assez similaires, mais ont historiquement été largement différenciées par la morphologie.

Pour vérifier, l'efficacité de la morphologie classification des excréments en espèces en construisant des modèles prédictifs pour l'identification des espèces avec un ensemble d'excréments bien décrits et vérifiés par l'ADN.**Rachel E. B. Reid** a compilé une base de données de traits morphologiques, biogéochimiques et contextuels pour un ensemble de 122 ADN vérifiés des excréments de bobcat, de coyote et de gray fox: data(scat) dans la librairie caret.

```
library(tree)
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(rpart)

data("scat")
summary(scat)

##      Species        Month       Year      Site     Location
##  bobcat   :57  November :17   Min.   :2011  ANNU:92    edge   :38
##  coyote   :28   January  :16  1st Qu.:2011  YOLA:18   middle  :47
##  gray_fox:25  April    :14 Median  :2012           off_edge:25
##                  September:14 Mean   :2012
##                  June     :13  3rd Qu.:2012
##                  October  :12 Max.   :2013
##                  (Other)  :24
```

```

##      Age        Number       Length       Diameter
## Min. :1.000   Min. :1.000   Min. : 2.500   Min. : 7.80
## 1st Qu.:3.000 1st Qu.:2.000 1st Qu.: 6.500 1st Qu.:16.07
## Median :3.000 Median :2.000 Median : 9.000 Median :18.05
## Mean   :3.345 Mean  :2.618 Mean  : 9.298 Mean  :18.56
## 3rd Qu.:5.000 3rd Qu.:3.000 3rd Qu.:11.500 3rd Qu.:21.32
## Max.  :5.000  Max. :7.000  Max. :20.500  Max. :30.00
##                               NA's :6
##
##      Taper        TI        Mass       d13C
## Min. : 2.30   Min. :0.230   Min. : 0.94   Min. :-29.85
## 1st Qu.:17.30 1st Qu.:0.990 1st Qu.: 5.66 1st Qu.:-28.08
## Median :25.80 Median :1.430 Median : 9.75 Median :-27.47
## Mean   :27.43 Mean  :1.602 Mean  :12.46 Mean  :-26.86
## 3rd Qu.:37.40 3rd Qu.:1.890 3rd Qu.:17.61 3rd Qu.:-26.45
## Max.  :91.50  Max. :8.680 Max. :53.70 Max. :-19.67
## NA's  :17     NA's :17    NA's :1     NA's :2
##
##      d15N        CN        ropey      segmented
## Min. : 1.840  Min. : 4.500  Min. :0.0000  Min. :0.0000
## 1st Qu.: 5.620 1st Qu.: 6.200 1st Qu.:0.0000 1st Qu.:0.0000
## Median : 6.885 Median : 7.250 Median :1.0000 Median :1.0000
## Mean   : 7.436 Mean  : 8.399 Mean  :0.5636 Mean  :0.5636
## 3rd Qu.: 8.305 3rd Qu.: 8.650 3rd Qu.:1.0000 3rd Qu.:1.0000
## Max.  :18.000  Max. :23.600 Max. :1.0000 Max. :1.0000
## NA's  : 2      NA's : 2
##
##      flat        scrape
## Min. :0.00000  Min. :0.00000
## 1st Qu.:0.00000 1st Qu.:0.00000
## Median :0.00000 Median :0.00000
## Mean   :0.05455 Mean  :0.04545
## 3rd Qu.:0.00000 3rd Qu.:0.00000
## Max. :1.00000  Max. :1.00000
##

```

```
str(scat)
```

```

## 'data.frame': 110 obs. of 19 variables:
## $ Species : Factor w/ 3 levels "bobcat","coyote",...: 2 2 1 2 2 2 1 1 1 ...
## $ Month   : Factor w/ 9 levels "April","August",...: 4 4 4 4 4 4 4 4 4 ...
## $ Year    : int 2012 2012 2012 2012 2012 2012 2012 2012 2012 ...
## $ Site    : Factor w/ 2 levels "ANNU","YOLA": 2 2 2 2 2 2 1 1 1 ...
## $ Location: Factor w/ 3 levels "edge","middle",...: 1 1 2 2 1 1 3 3 3 2 ...
## $ Age     : int 5 3 3 5 5 5 1 3 5 5 ...
## $ Number  : int 2 2 2 2 4 3 5 7 2 1 ...
## $ Length  : num 9.5 14.9 8.5 8.9 6.5 5.5 11.20.5 ...
## $ Diameter: num 25.7 25.4 18.8 18.1 20.7 21.2 15.7 21.9 17.5 18 ...
## $ Taper   : num 41.9 37.1 16.5 24.7 20.1 28.5 8.2 19.3 29.1 21.4 ...
## $ TI      : num 1.63 1.46 0.88 1.36 0.97 1.34 0.52 0.88 1.66 1.19 ...
## $ Mass    : num 15.9 17.6 8.4 7.4 25.4 ...
## $ d13C   : num -26.9 -29.6 -28.7 -20.1 -23.2 ...
## $ d15N   : num 6.94 9.87 8.52 5.79 7.01 8.28 4.2 3.89 7.34 6.06 ...

```

```

## $ CN      : num  8.5 11.3 8.1 11.5 10.6 9 5.4 5.6 5.8 7.7 ...
## $ ropey   : int  0 0 1 1 0 1 1 0 0 1 ...
## $ segmented: int  0 0 1 0 1 0 1 1 1 1 ...
## $ flat    : int  0 0 0 0 0 0 0 0 0 ...
## $ scrape  : int  0 0 1 0 0 0 1 0 0 0 ...

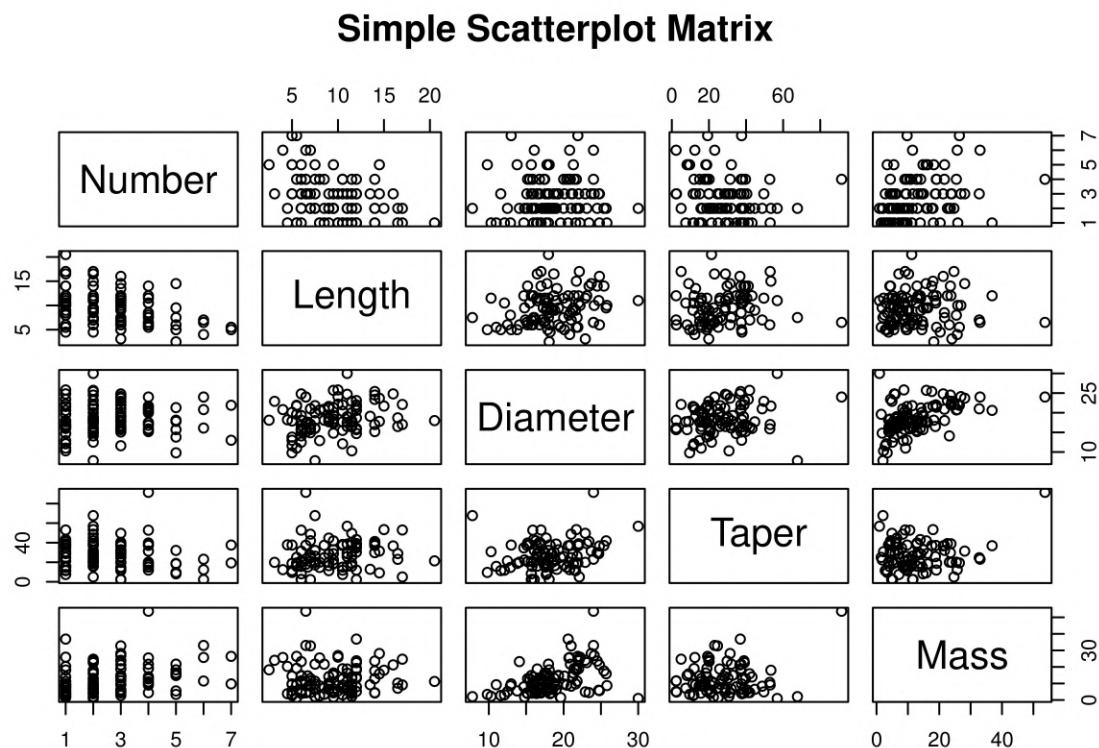
```

variables	unités	Description
Scat diameter	mm	mesure au point le plus large au 10e de millimètre près
Scat length	cm	longueur de la pièce la plus longue à 0,5 cm près
Taper length	mm	longueur du cône le plus long, le long de l'axe du scat
Degree of taper	sans unité	rapport de la longueur du cône au diamètre de l'éclat
Number of pieces	integer	nombre d'excréments séparés
Scat mass	grams	poids sec total après lyophilisation et cuisson
Segmented?	NA	le scat montre-t-il une segmentation? 1 = oui, 0 = non
Ropey?	NA	est-ce que les excréments semblent cordés / tordus / tissés? 1 = oui, 0 = non
Flat?	NA	est-ce que le scat est une flaqué d'eau plate qui manque d'autres traits morphologiques? 1 = oui, 0 = non
Location	3 point scale	variable catégorielle décrivant l'emplacement des excréments sur le sentier / route - milieu, bord ou hors bord
Scrape?	NA	y a-t-il une éraflure près de la crasse? 1 = oui, 0 = non
d15N	sans unité	delta-N-15 est une mesure du rapport entre les deux isotopes stables de l'azote , $^{15}N : ^{14}N$
d13C	sans unité	delta-C-treize est une signature isotopique, une mesure du rapport entre les isotopes stables $^{13}C : ^{12}C$
TI	sans unité	titane
C:N Ratio	sans unité	rapport des atomes de carbone aux atomes d'azote dans les excréments, qui est une approximation du degré de carnivore de l'animal

Nous avons comme variables morphologiques Number, Length, Diameter, Taper, Mass, ropey, segmented, flat et scrape, biogéochimiques d15N, d13C, CN et contextuels Month, Year, Site et Location.

Exploration non supervisée avec l'algorithme des k-means

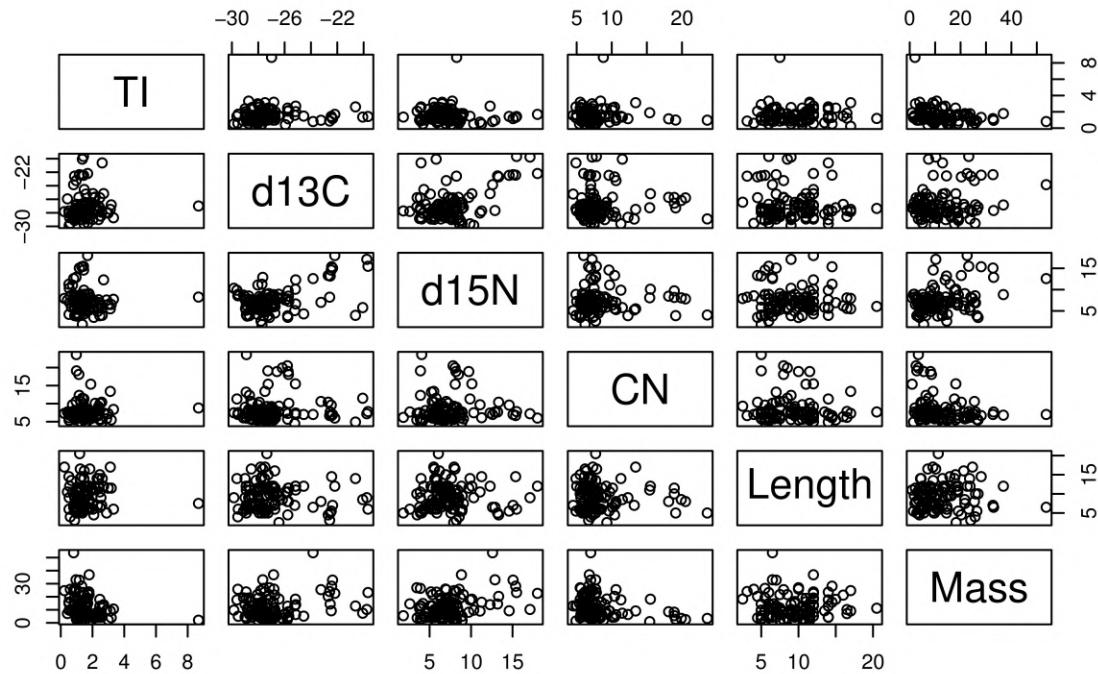
```
pairs(~Number+Length+Diameter+Taper+Mass,main="Simple Scatterplot Matrix",data=scat)
```



```
##
```

```
pairs(~TI+d13C+d15N+CN+Length+Mass,main="Simple Scatterplot Matrix",data=scat)
```

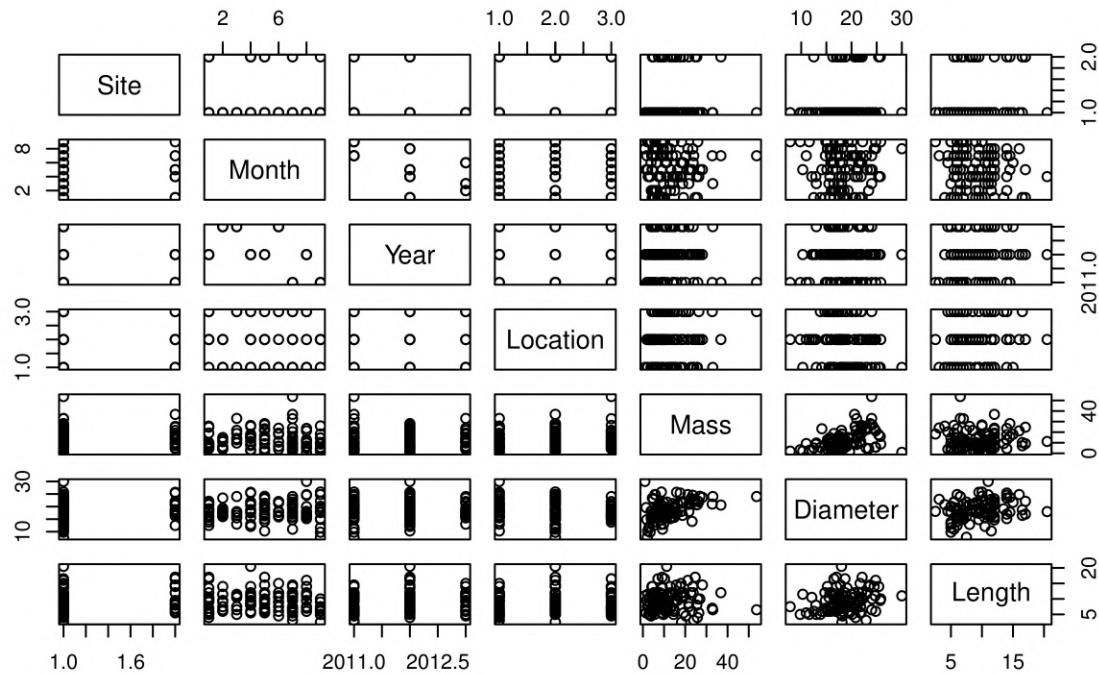
## Simple Scatterplot Matrix



Nous constatons des correlations fortes des variables. Surtout entre les variables biogéochimiques.

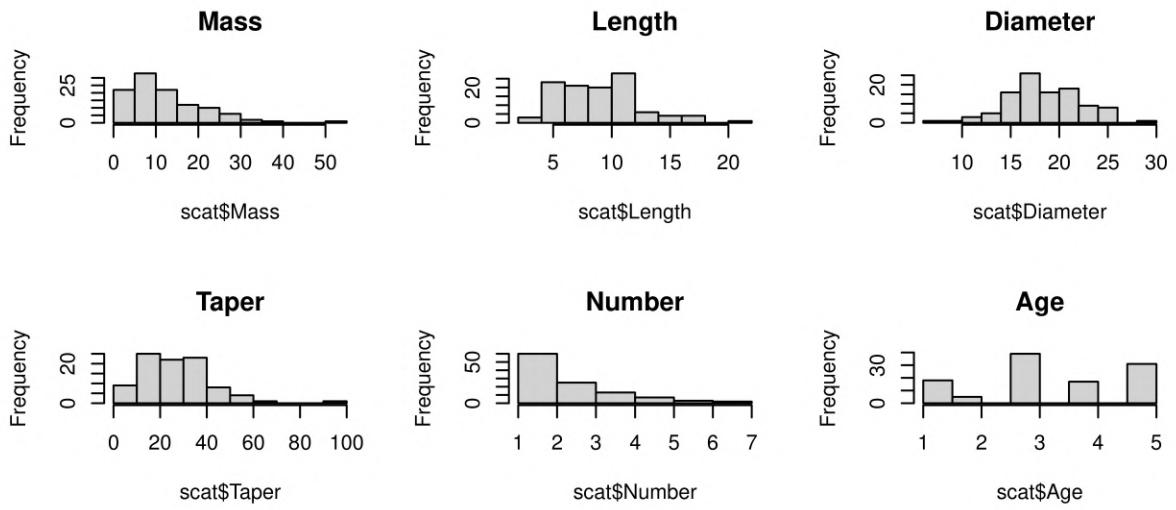
```
pairs(~Site+Month+Year+Location+Mass+Diameter+Length, main="Simple Scatterplot Matrix", data=scat)
```

## Simple Scatterplot Matrix



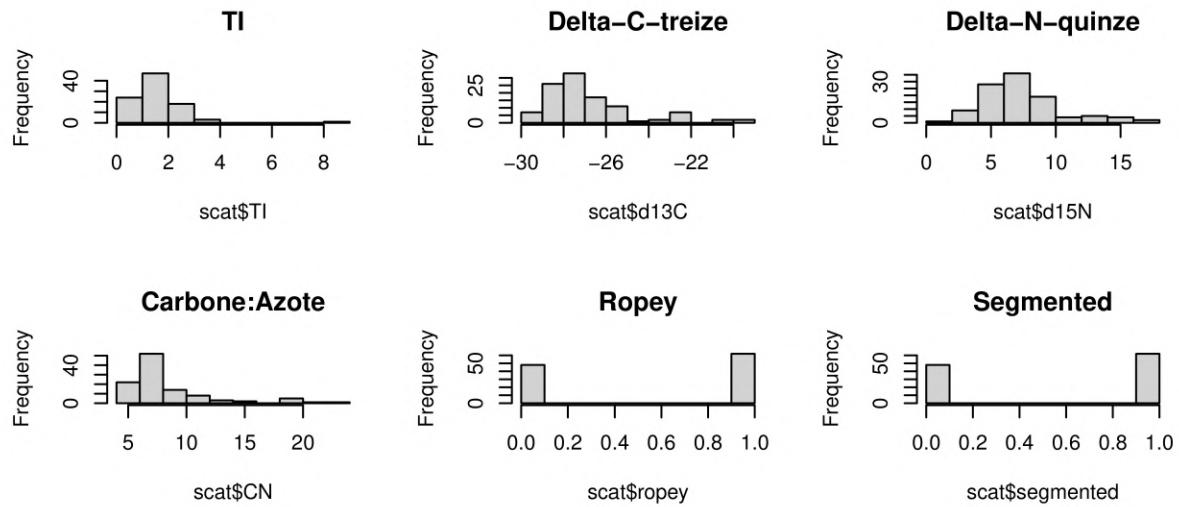
```
##
```

```
par(mfrow=c(3,3))
hist(scat$Mass, main = "Mass")
hist(scat$Length, main = "Length")
hist(scat$Diameter, main = "Diameter")
hist(scat$Taper, main = "Taper")
hist(scat$Number, main = "Number")
hist(scat$Age, main = "Age")
```

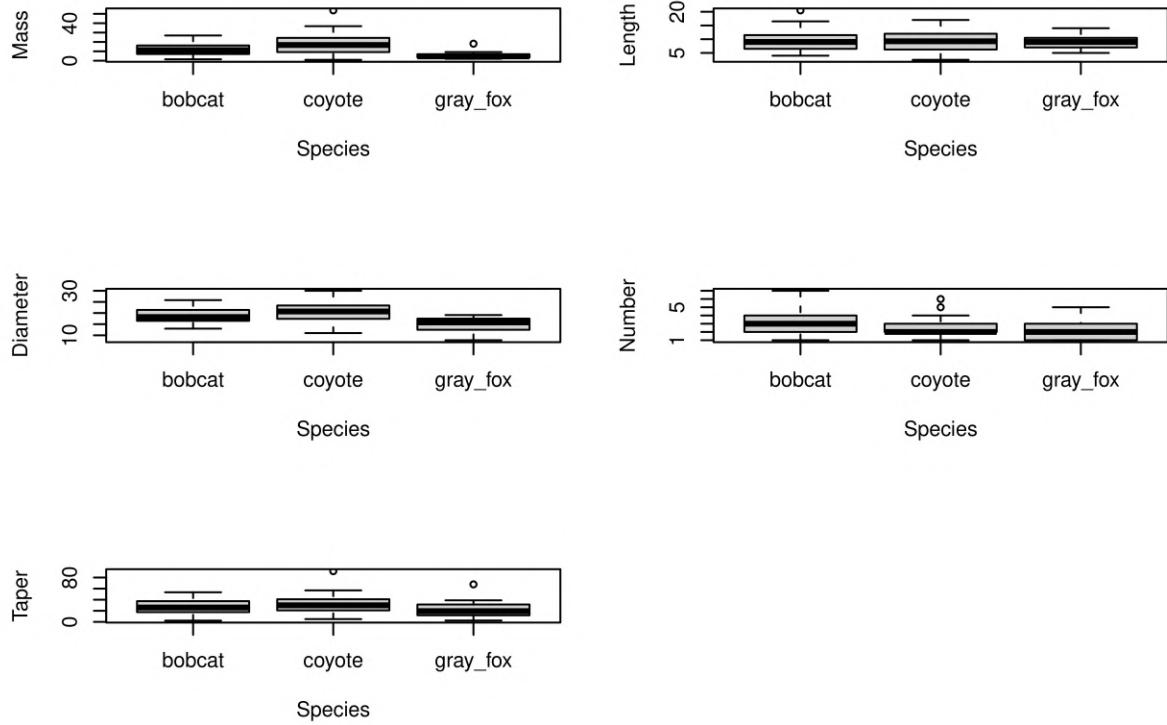


##

```
par(mfrow=c(3,3))
hist(scat$TI, main = "TI")
hist(scat$d13C, main = "Delta-C-treize")
hist(scat$d15N, main = "Delta-N-quinze")
hist(scat$CN, main = "Carbone:Azote")
hist(scat$ropey, main = "Ropey")
hist(scat$segmented, main = "Segmented")
```

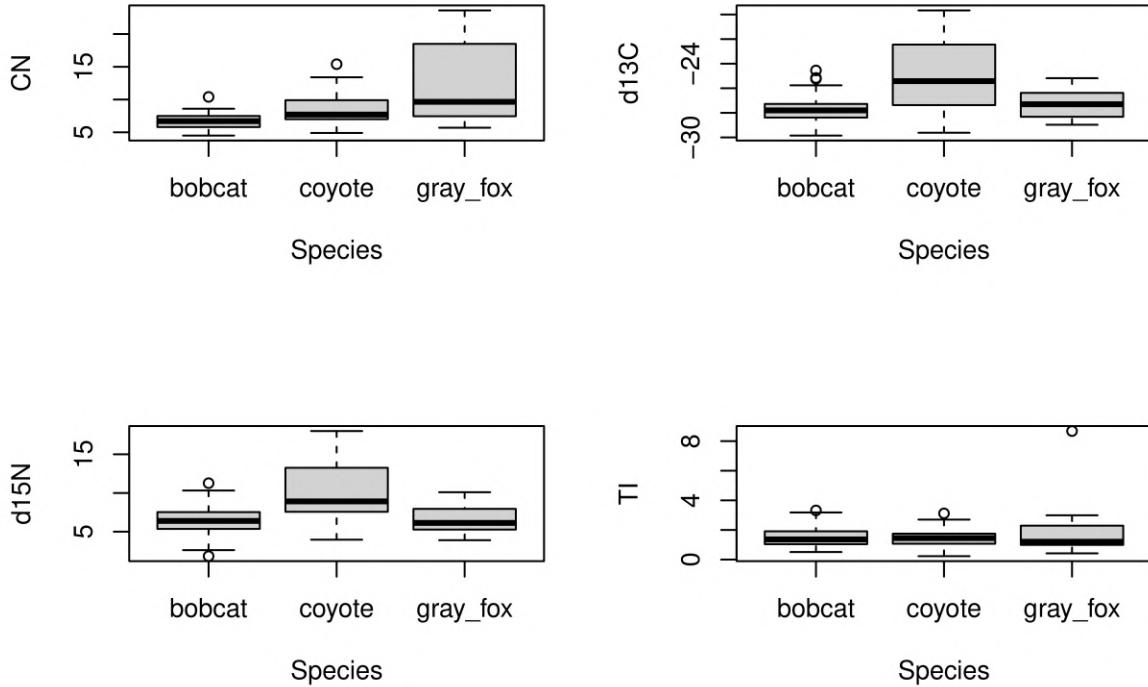


```
par(mfrow=c(3,2))
boxplot(Mass~Species,data=scat)
boxplot(Length~Species,data=scat)
boxplot(Diameter~Species,data=scat)
boxplot(Number~Species,data=scat)
boxplot(Taper~Species,data=scat)
```



Les médianes de Mass et Diameter de grayfox sont plus bas que celles des autres.  
 Les médianes de longueur sont sensiblement égales pour toutes les espèces.  
 Les bobcat ont une médiane plus grande par rapport aux coyotes et gray\_fox.

```
par(mfrow=c(2,2))
boxplot(CN~Species,data=scat)
boxplot(d13C~Species,data=scat)
boxplot(d15N~Species,data=scat)
boxplot(TI~Species,data=scat)
```

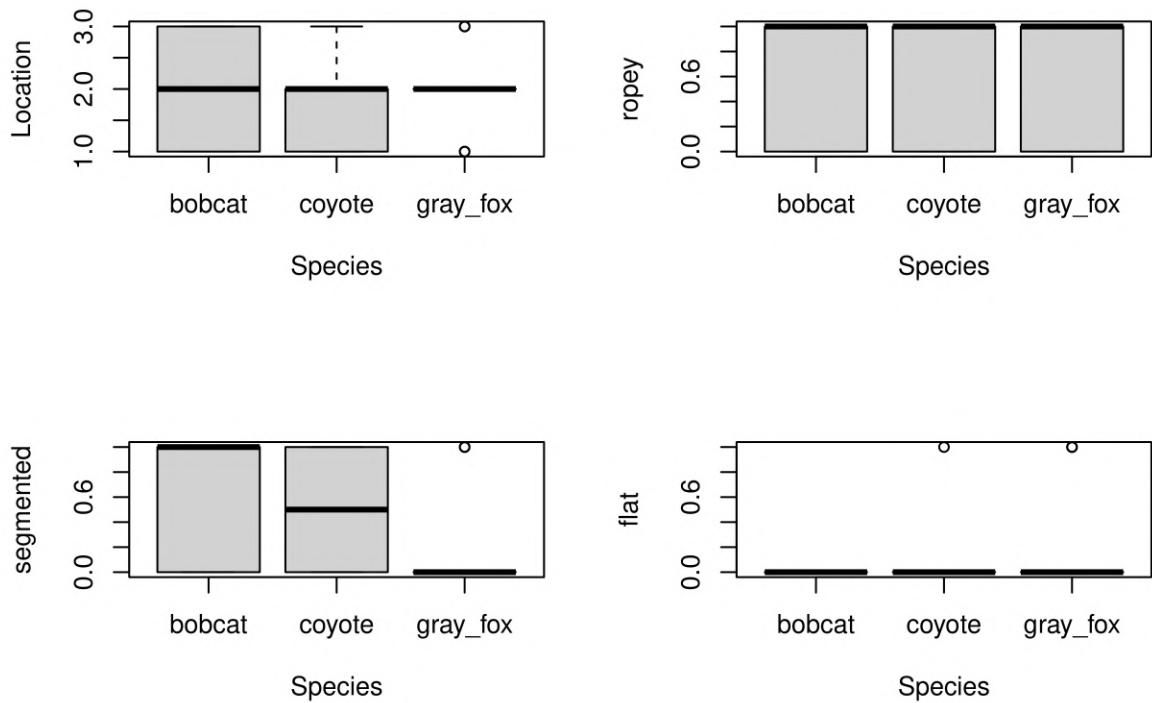


Les médianes de d15N et d13C pour coyote supérieure à celle des autres qui ont des médianes très proches.  
La médiane de CN pour les gray\_fox est supérieure à celle des autres.

##

```
te=scat
te$Location=as.numeric(scat$Location)
te$Site=as.numeric(scat$Site)
par(mfrow=c(2,2))

boxplot(Location~Species,data=te)
boxplot(ropey~Species,data=scat)
boxplot(segmented~Species,data=scat)
boxplot(flat~Species,data=scat)
```

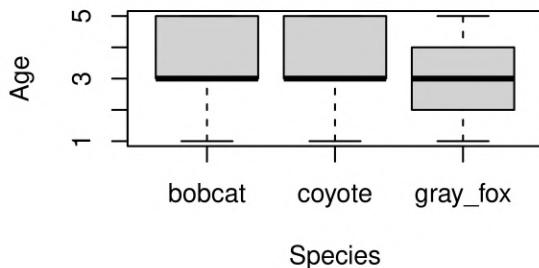
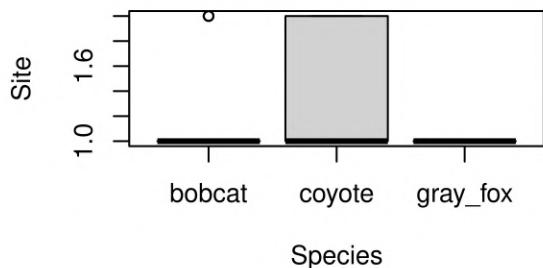
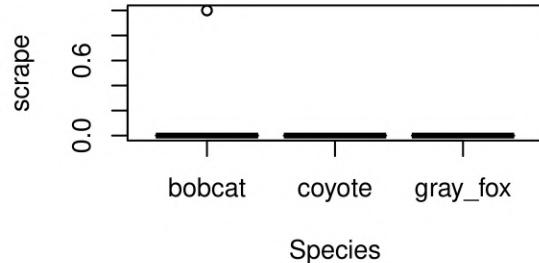
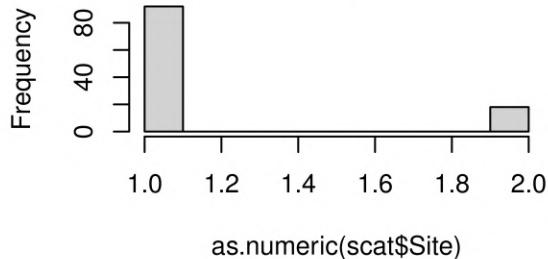


La médiane d'excréments segmentés est très élevée pour bobcat par rapport aux autres tandis que celle de gray\_fox est nulle.

##

```
par(mfrow=c(2,2))
hist(as.numeric(scat$Site))
boxplot(scrape~Species,data=scat)
boxplot(Site~Species,data=te)
boxplot(Age~Species,data=te)
```

### Histogram of as.numeric(scat\$Site)



```
rowna=sum(rowSums(is.na(scat))!=0)
colna=sum(colSums(is.na(scat))!=0)
print(paste('Le nombre de lignes contenant de NA est: ',rowna),quote = FALSE)
```

```
## [1] Le nombre de lignes contenant de NA est: 19
```

```
print(paste('Le nombre de colonnes contenant de NA est: ',colna),quote = FALSE)
```

```
## [1] Le nombre de colonnes contenant de NA est: 7
```

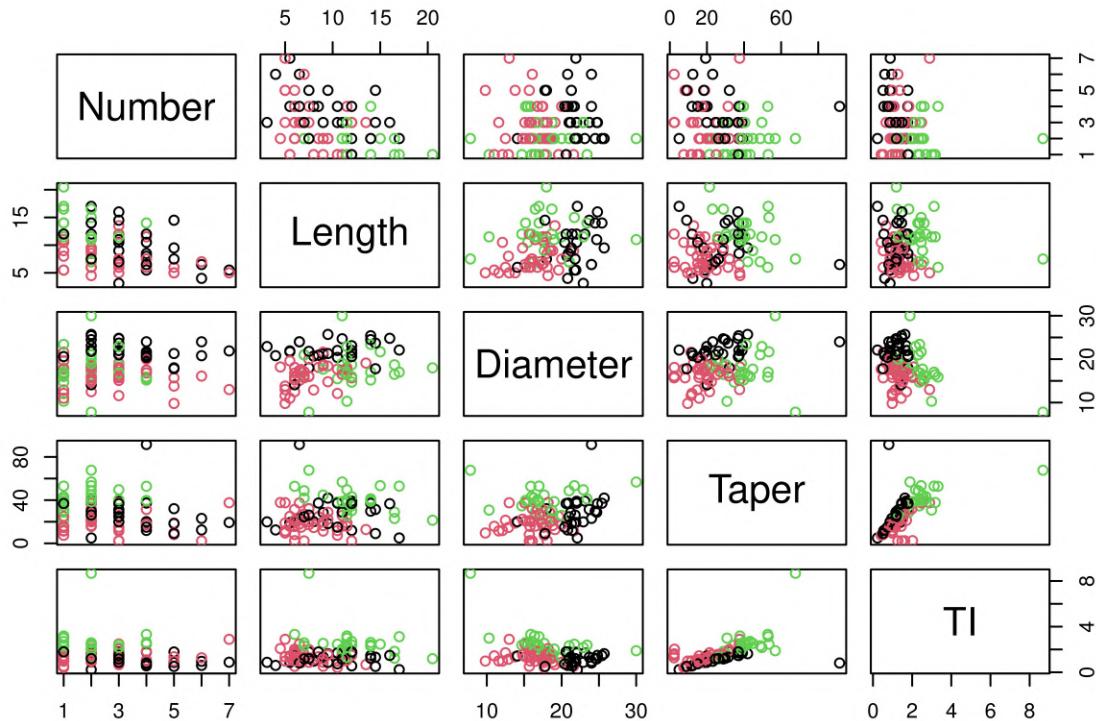
Comme nous avons des données manquantes, alors pour la méthode des kmeans nous pouvons soit supprimer les individus avec des données manquantes, soit remplacer le NA par la moyenne de la colonne où elle se trouve.

Cependant la méthode des kmeans est utilisée que pour variable quantitative

### kmeans avec données manquantes supprimées

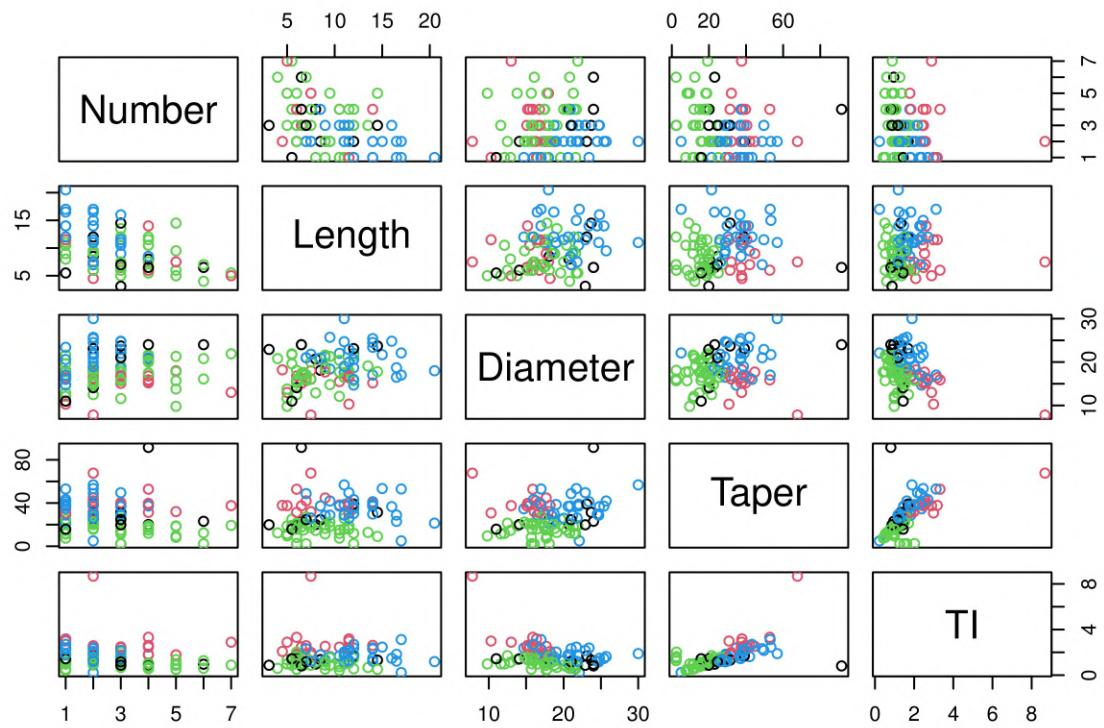
```
scat.nona=data.frame(na.omit(scat)) #supprime les NAs
```

```
scat.scale = scale(scat.nona[,c(3,6:15)]) #centrage des valeurs quantitatives
km=kmeans(scat.scale,3)
plot(scat.nona[,7:11], col = km$cluster) #visualisation sur les variables 7 à 11
points(km$centers, col = 1:3, pch = 8)
```



Nous constatons trois groupes: l'un d'eux est inclus dans l'union des deux autres qui une intersection petite.

```
km=kmeans(scat.scale,4)
plot(scat.nona[,7:11], col = km$cluster) #visualisation sur les variables 7 à 11
points(km$centers, col = 1:4, pch = 8)
```



```

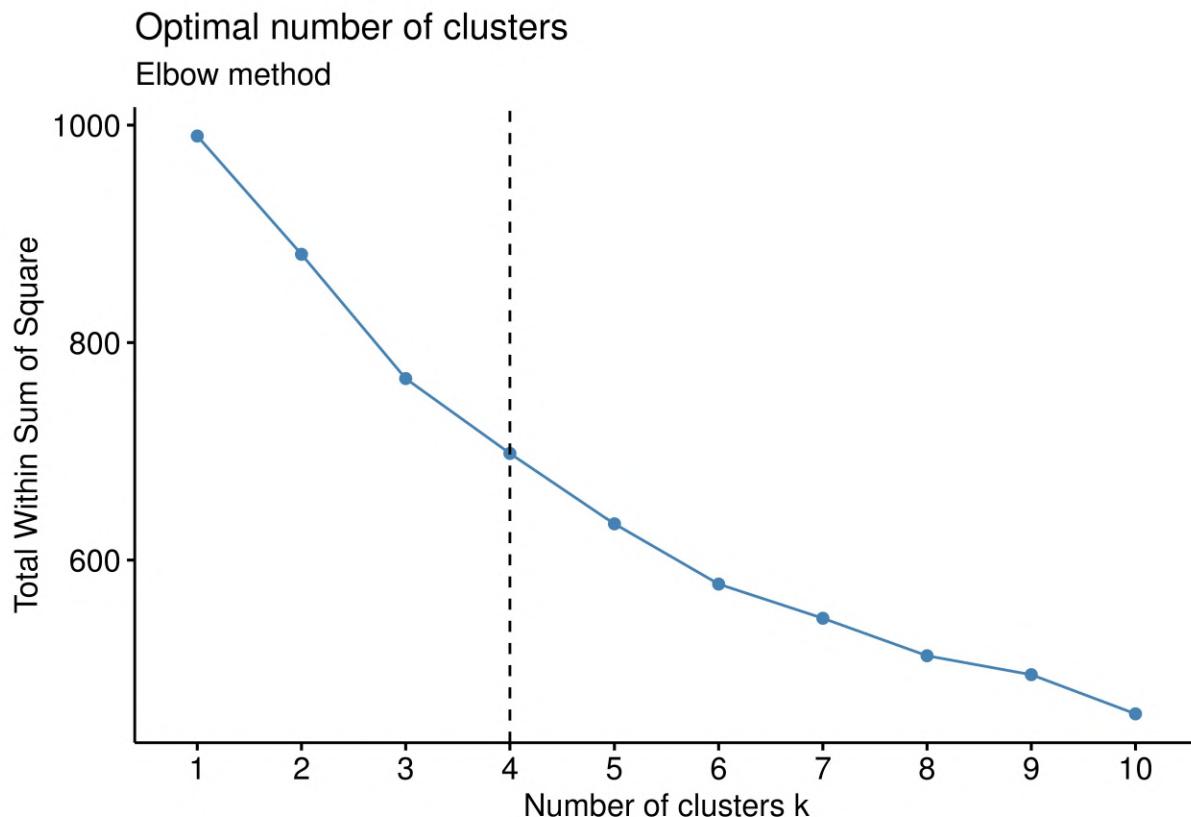
library(factoextra)

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

library(NbClust)

fviz_nbclust(scat.scale, kmeans, method = "wss") +
  geom_vline(xintercept = 4, linetype = 2) +
  labs(subtitle = "Elbow method")

```



La valeur optimale de  $k$  pour est 4 pour des  $k$  entre 0 et 10.

kmeans avec données manquantes remplacer par la moyenne

```

scat.moy = data.frame(scat)
for (j in c(3,6:15)) #variable quantitatives
{colm=mean(scat.nona[,j])
 for ( i in 1:110)
 {
  if (is.na(scat.moy[i,j]))
  {
   scat.moy[i,j]=colm
  }
 }
}

summary(scat.moy)

##      Species      Month       Year      Site      Location
##  bobcat :57  November :17   Min.   :2011  ANNU:92    edge   :38
##  coyote  :28   January  :16  1st Qu.:2011  YOLA:18   middle  :47
##  gray_fox:25   April    :14 Median  :2012           off_edge:25

```

```

## September:14 Mean :2012
## June :13 3rd Qu.:2012
## October :12 Max. :2013
## (Other) :24
##      Age       Number       Length       Diameter
## Min. :1.000 Min. :1.000 Min. : 2.500 Min. : 7.80
## 1st Qu.:3.000 1st Qu.:2.000 1st Qu.: 6.500 1st Qu.:16.15
## Median :3.000 Median :2.000 Median : 9.000 Median :18.15
## Mean :3.345 Mean :2.618 Mean : 9.298 Mean :18.55
## 3rd Qu.:5.000 3rd Qu.:3.000 3rd Qu.:11.500 3rd Qu.:21.15
## Max. :5.000 Max. :7.000 Max. :20.500 Max. :30.00
##
##      Taper       TI       Mass       d13C
## Min. : 2.30 Min. :0.230 Min. : 0.940 Min. : -29.85
## 1st Qu.:18.82 1st Qu.:1.095 1st Qu.: 5.728 1st Qu.: -28.06
## Median :27.47 Median :1.593 Median : 9.980 Median : -27.45
## Mean :27.44 Mean :1.602 Mean :12.456 Mean : -26.86
## 3rd Qu.:36.85 3rd Qu.:1.780 3rd Qu.:17.290 3rd Qu.: -26.48
## Max. :91.50 Max. :8.680 Max. :53.700 Max. : -19.67
##
##      d15N       CN       ropey       segmented
## Min. : 1.840 Min. : 4.500 Min. :0.0000 Min. :0.0000
## 1st Qu.: 5.670 1st Qu.: 6.225 1st Qu.:0.0000 1st Qu.:0.0000
## Median : 6.925 Median : 7.300 Median :1.0000 Median :1.0000
## Mean : 7.434 Mean : 8.388 Mean :0.5636 Mean :0.5636
## 3rd Qu.: 8.275 3rd Qu.: 8.575 3rd Qu.:1.0000 3rd Qu.:1.0000
## Max. :18.000 Max. :23.600 Max. :1.0000 Max. :1.0000
##
##      flat       scrape
## Min. :0.00000 Min. :0.00000
## 1st Qu.:0.00000 1st Qu.:0.00000
## Median :0.00000 Median :0.00000
## Mean : 0.05455 Mean : 0.04545
## 3rd Qu.:0.00000 3rd Qu.:0.00000
## Max. :1.00000 Max. :1.00000
##

```

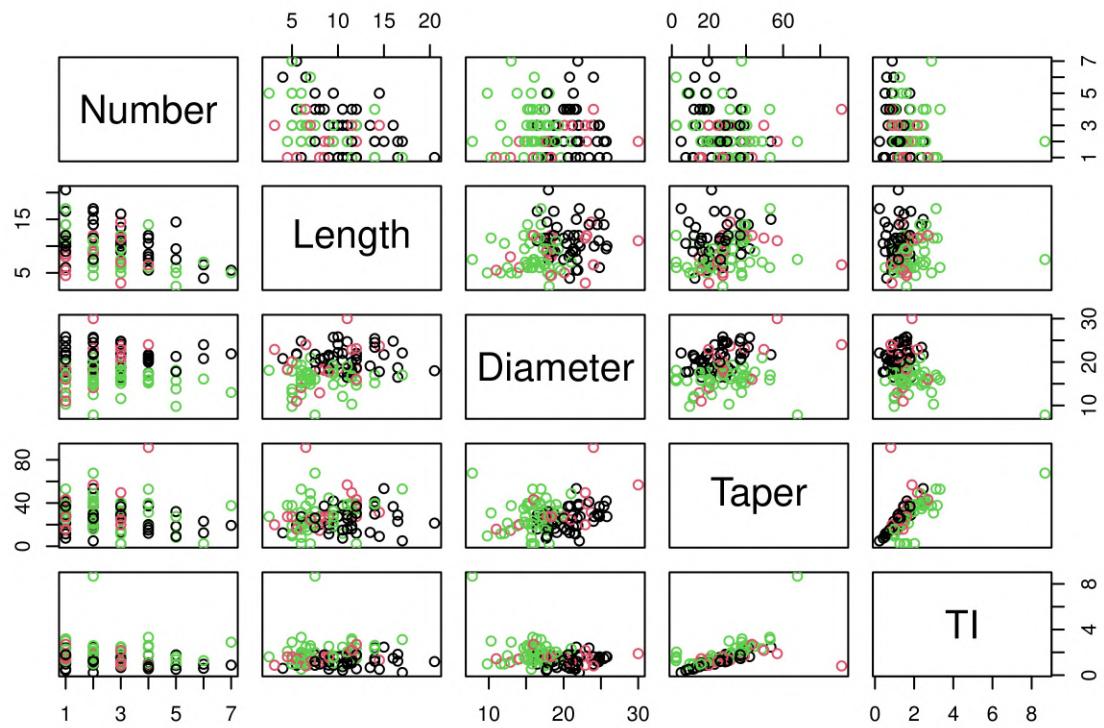
Nous voyons bien que les Nas sont remplacées.

```
##
```

```

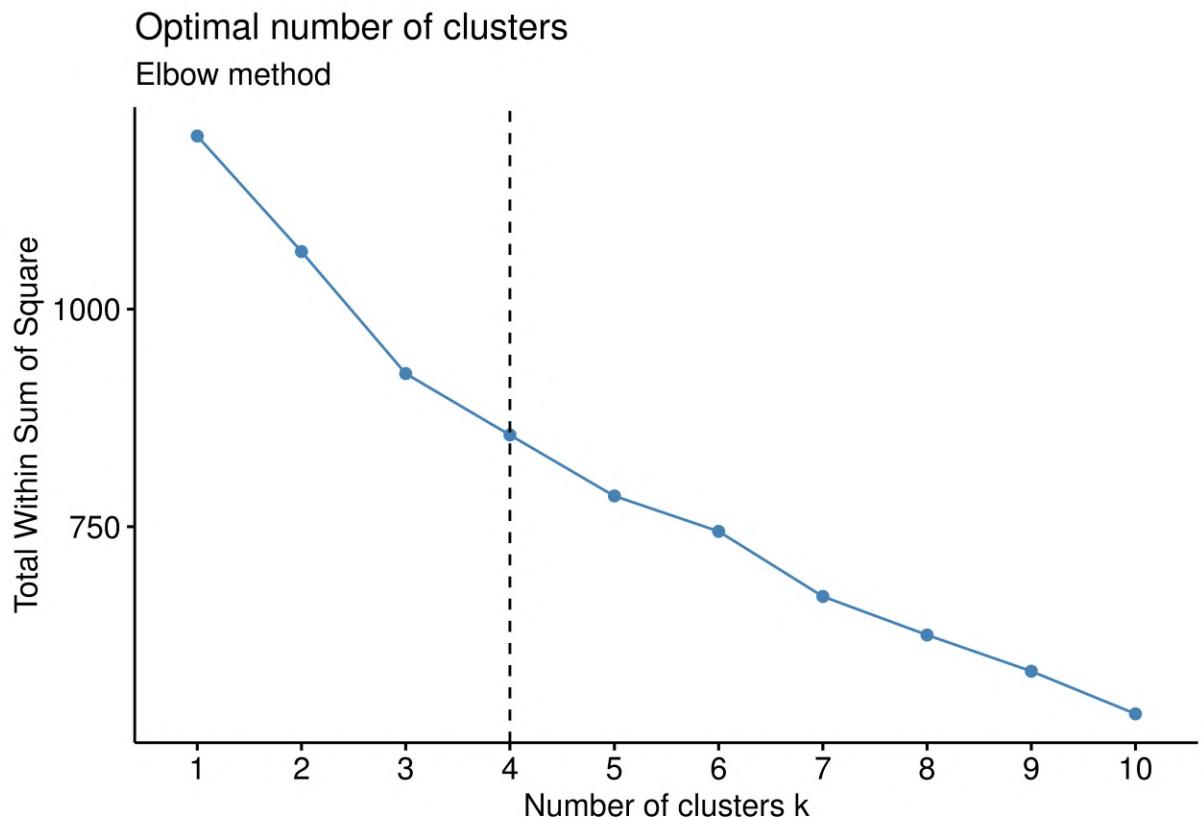
scat.scale1 = scale(scat.moy[,c(3,6:15)]) #centrage des valeurs quantitatives
km3=kmeans(scat.scale1,3)
plot(scat.moy[,7:11], col = km3$cluster) #visualisation sur les variables 7 à 11
points(km3$centers, col = 1:3, pch = 8)

```



```
##
```

```
fviz_nbclust(scat.scale1, kmeans, method = "wss") +
  geom_vline(xintercept = 4, linetype = 2) +
  labs(subtitle = "Elbow method")
```



Nous ne constatons pas de différence visible et nous avons le même optimal  $k = 4$ .

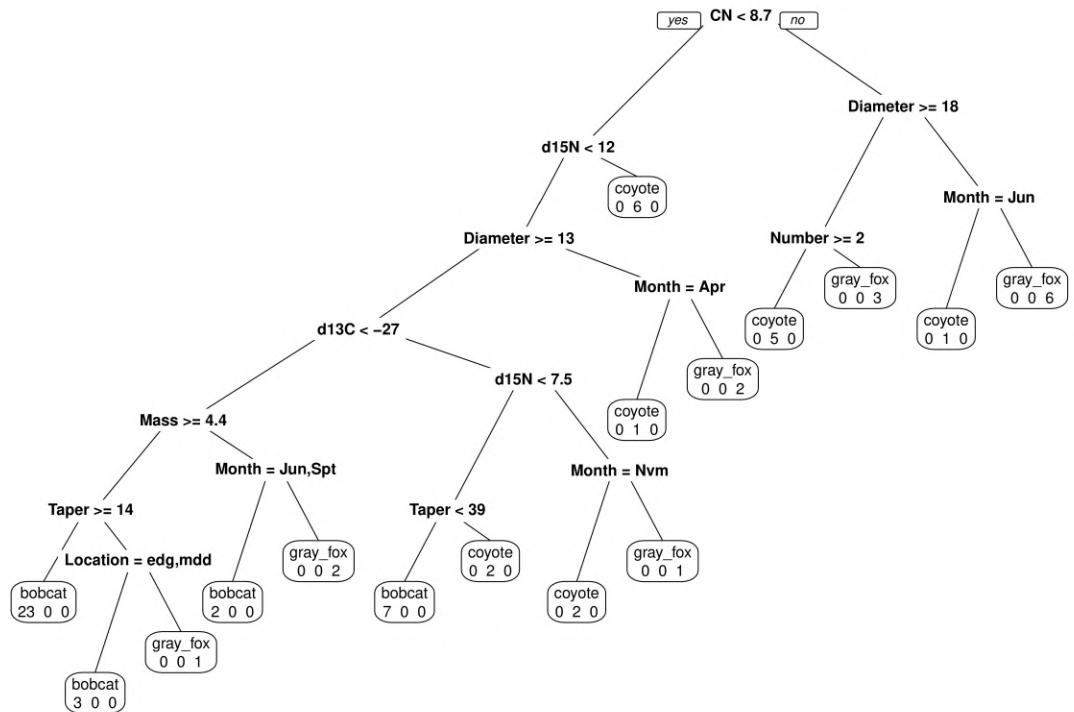
### Prédiction avec CART (Arbre de décision)

```
library(rpart.plot)

p <- createDataPartition(y=scat$Species,p=60/100,list=FALSE) # 60% entrainement
train <- scat[p,]    #Jeu d'entraînement
test <- scat[-p,]   #Jeu de test

model <- rpart(Species~. , data = train, control = rpart.control(minsplit = 1,cp=0))

prp(model, extra = 1, cex = 0.5)
```



```

pred <- predict(model,test,type = "class")

confM<-confusionMatrix(pred,test$Species)
confM

## Confusion Matrix and Statistics
##
##             Reference
## Prediction bobcat coyote gray_fox
##   bobcat      18      2      3
##   coyote       2      7      2
##   gray_fox     2      2      5
##
## Overall Statistics
##
##                 Accuracy : 0.6977
##                 95% CI : (0.5387, 0.8282)
##   No Information Rate : 0.5116
##   P-Value [Acc > NIR] : 0.01036
##
##                 Kappa : 0.5062
##
##   Mcnemar's Test P-Value : 0.97759
##
## Statistics by Class:
##
```

```

##          Class: bobcat Class: coyote Class: gray_fox
## Sensitivity      0.8182      0.6364      0.5000
## Specificity     0.7619      0.8750      0.8788
## Pos Pred Value   0.7826      0.6364      0.5556
## Neg Pred Value   0.8000      0.8750      0.8529
## Prevalence       0.5116      0.2558      0.2326
## Detection Rate   0.4186      0.1628      0.1163
## Detection Prevalence 0.5349      0.2558      0.2093
## Balanced Accuracy 0.7900      0.7557      0.6894

print(paste("L'erreur de prédiction est estimée à:", 1 - confM$overall[1]), quote = FALSE)

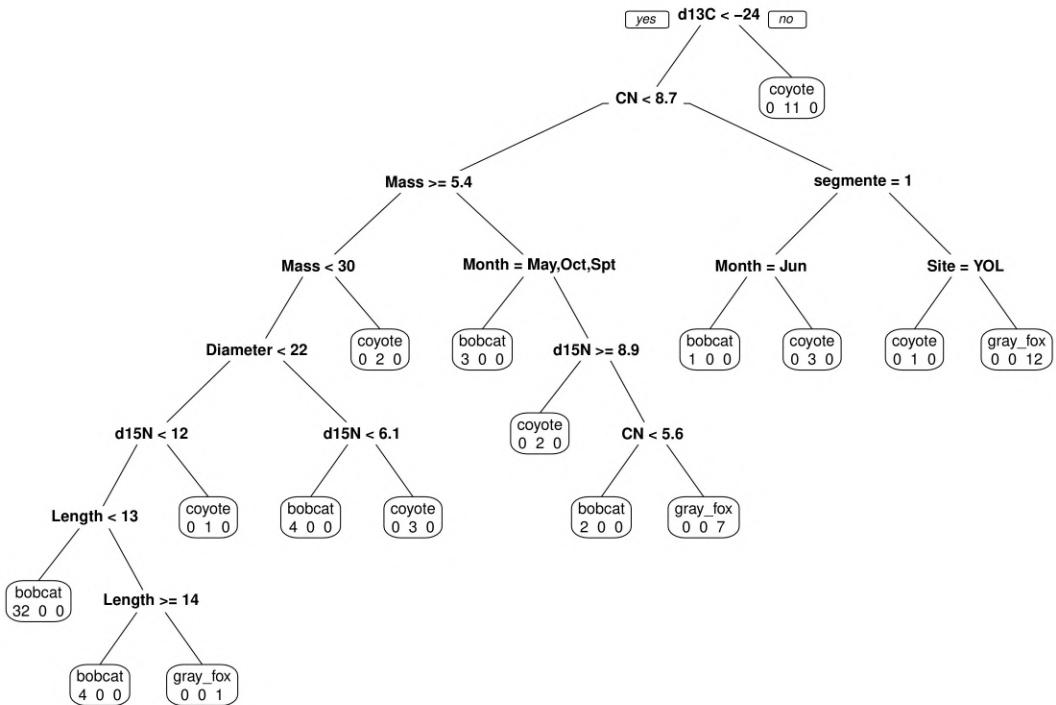
## [1] L'erreur de prédiction est estimée à: 0.302325581395349

p <- createDataPartition(y=scat$Species, p=80/100, list=FALSE) # 80% entraînement
train <- scat[p,]    #Jeu d'entraînement
test <- scat[-p,]    #Jeu de test

model <- rpart(Species~. , data = train, control = rpart.control(minsplit = 1, cp=0))

prp(model, extra = 1, cex = 0.5)

```



```

pred <- predict(model,test,type = "class")

confM1<-confusionMatrix(pred,test$Species)
confM1

## Confusion Matrix and Statistics
##
##             Reference
## Prediction  bobcat  coyote  gray_fox
##   bobcat      8       1       2
##   coyote      2       3       0
##   gray_fox    1       1       3
##
## Overall Statistics
##
##                 Accuracy : 0.66667
##                 95% CI : (0.4303, 0.8541)
##     No Information Rate : 0.5238
##     P-Value [Acc > NIR] : 0.1371
##
##                 Kappa : 0.4556
##
## McNemar's Test P-Value : 0.6444
##
## Statistics by Class:
##
```

```

##          Class: bobcat Class: coyote Class: gray_fox
## Sensitivity      0.7273      0.6000      0.6000
## Specificity     0.7000      0.8750      0.8750
## Pos Pred Value   0.7273      0.6000      0.6000
## Neg Pred Value   0.7000      0.8750      0.8750
## Prevalence       0.5238      0.2381      0.2381
## Detection Rate   0.3810      0.1429      0.1429
## Detection Prevalence 0.5238      0.2381      0.2381
## Balanced Accuracy 0.7136      0.7375      0.7375

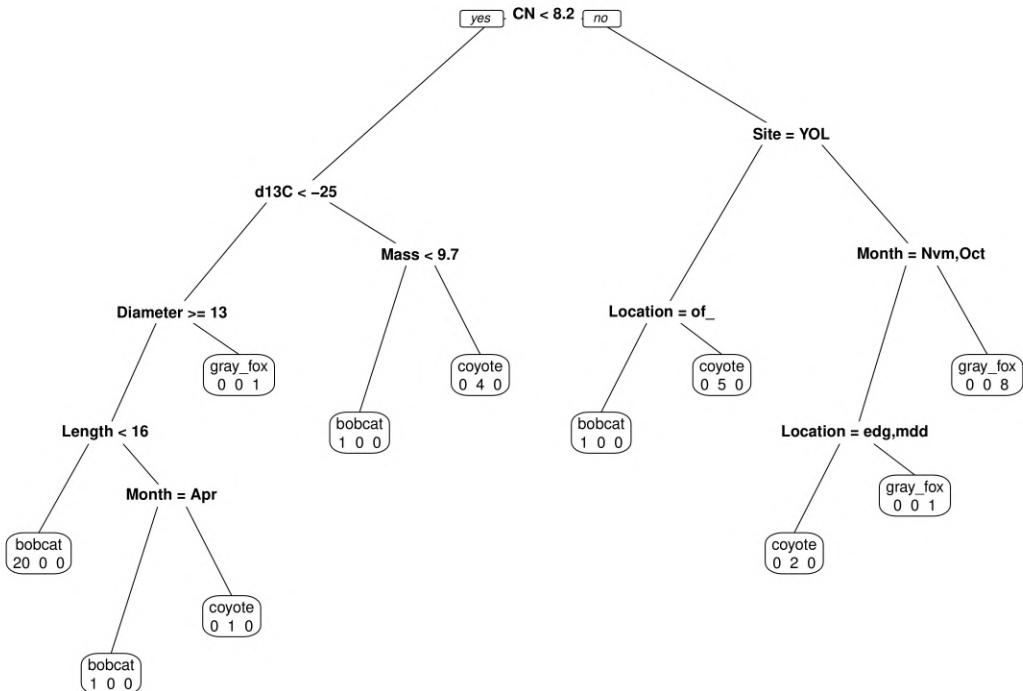
print(paste("L'erreur de prédiction pour 20% test est estimée de à:", 1 - confM1$overall[1]), quote = FALSE)
## [1] L'erreur de prédiction pour 20% test est estimée de à: 0.3333333333333333

p <- createDataPartition(y=scat$Species, p=40/100, list=FALSE) # 40% entraînement
train <- scat[p,]    #Jeu d'entraînement
test <- scat[-p,]    #Jeu de test

model <- rpart(Species~., data = train, control = rpart.control(minsplit = 1, cp=0))

prp(model, extra = 1, cex = 0.5)

```



```

pred <- predict(model,test,type = "class")

confM2<-confusionMatrix(pred,test$Species)
confM2

## Confusion Matrix and Statistics
##
##             Reference
## Prediction bobcat coyote gray_fox
##     bobcat      30      4      7
##     coyote       1      9      1
##     gray_fox     3      3      7
##
## Overall Statistics
##
##                 Accuracy : 0.7077
##                 95% CI : (0.5817, 0.814)
##     No Information Rate : 0.5231
##     P-Value [Acc > NIR] : 0.001877
##
##                 Kappa : 0.498
##
## McNemar's Test P-Value : 0.221385
##
## Statistics by Class:
##
```

```

##          Class: bobcat Class: coyote Class: gray_fox
## Sensitivity      0.8824      0.5625      0.4667
## Specificity      0.6452      0.9592      0.8800
## Pos Pred Value    0.7317      0.8182      0.5385
## Neg Pred Value    0.8333      0.8704      0.8462
## Prevalence        0.5231      0.2462      0.2308
## Detection Rate    0.4615      0.1385      0.1077
## Detection Prevalence 0.6308      0.1692      0.2000
## Balanced Accuracy 0.7638      0.7608      0.6733

```

```

for (pr in seq(0.01,1,0.15))
{ p <- createDataPartition(y=scat$Species,p=pr,list=FALSE)
  train <- scat[p,]  #Jeu d'entraînement
  test <- scat[-p,]  #Jeu de test

  model <- rpart(Species~. , data = train, control = rpart.control(minsplit = 1,cp=0))
  pred <- predict(model,test,type = "class")

  confM3<-confusionMatrix(pred,test$Species)
  print(paste("L'erreur de prédiction est estimée à:",  1 - confM3$overall[1],"avec", pr*100,"% comme train"))
}

## [1] L'erreur de prédiction est estimée à: 0.728971962616823 avec 1 % comme train
## [1] L'erreur de prédiction est estimée à: 0.461538461538462 avec 16 % comme train
## [1] L'erreur de prédiction est estimée à: 0.4 avec 31 % comme train
## [1] L'erreur de prédiction est estimée à: 0.293103448275862 avec 46 % comme train
## [1] L'erreur de prédiction est estimée à: 0.390243902439024 avec 61 % comme train
## [1] L'erreur de prédiction est estimée à: 0.28 avec 76 % comme train
## [1] L'erreur de prédiction est estimée à: 0.4444444444444444 avec 91 % comme train

```

Nous pouvons conclure que la précision n'est pas du pourcentage du jeu d'entraînement.