



Module 4 - Visualisation des données

Exercices - Solutions

Exercice 1 :

1. Le téléchargement des données de `algea` est fait grâce à la ligne 2 du Listing 1. Ceci doit être précédé de l'installation du package `DMwR2` de R.
2. `algea` comprend 200 observations dans plusieurs rivières européennes décrites par 18 variables. Parmi ces variables se retrouvent 3 variables qualitatives : saison (`season`), taille (`size`) et vitesse (`speed`) qui décrivent l'échantillon d'eau. Les 15 autres variables sont quantitatives; elles mesurent des concentrations chimiques : `mxPH`, `mnO2`, `Cl`, `N03`, `NH4`, `oP04`, `P04`, `Chla`, `a1`, `a2`, `a3`, `a4`, `a5`, `a6`, `a7`
3. L'affichage du contenu des dix premières lignes de la base de données se fait grâce à la ligne 4 de code du Listing 1.
4. La distribution de la variable `a6` est donnée par l'histogramme de la figure 1 (a) (ligne 6 de code du Listing 1). Nous pouvons obtenir une meilleure apparence en augmentant le nombre de bins tel que illustré dans la figure 1 (b) (ligne 7 du code du Listing 1).
5. La variable `size` est une variable qualitative qui prend trois valeurs `large`, `medium`, `small`. Sa distribution est donnée par l'histogramme de la figure 2 (a) (ligne 8 du code du Listing 1).
6. Les statistiques descriptives de la variable `N03` selon la taille de la rivière `size` sont obtenues grâce au diagramme en boîte de la figure 3 (ligne 9 du code du Listing 1).

```

1  install.packages("DMwR2")
2  data(algae, package="DMwR2")
3  summary(algae)
4  head(algae, 10)
5  hist(algae$a6, main="(a) Distribution de a6", xlab="Frequence
   de a6", ylab="Occurrences")
6  hist(algae$a6, main="(b) Distribution de a6", xlab="Frequence
   de a6", ylab="Occurrences", col="blue",
7  breaks = seq(min(algae$a6), max(algae$a6), length.out = 21))
8  barplot(table(algae$size), main = "Distribution de la taille
   (size)")
9  boxplot(algae$N03~algae$size, names=c("large", "medium", "small
   "), col=c("blue", "pink", "grey"), main="Repartition de N03
   selon la taille (size)", cex.main = 1.5, cex.sub=1.5,
   cex.axis = 1, cex.lab= 1.5, ylim=c(0.001,12))

```

Listing 1 – Pseudo code de l'Exercice 1

Exercice 2 :

1. La base de données `mtcars` contient un ensemble de 32 voitures et les mesures de 11 variables liées à l'aspect et leurs performances : `mpg` (Milles/gallon), `cyl` (nombre de cylindre), `disp` (déplacement), `hp` (puissance brute), `drat` (Rapport de pont arrière), `wt` (poids (1000 lbs)), `qsec` (temps pour 1/4 de miles), `vs` (Moteur (0 = V-shaped, 1 = droit)), `am` (Transmission (0 = automatique, 1 = manuelle)), `gears` (nombre de gears), `carb` (nombre de carburateurs).
2. La Figure 4 représente la distribution du nombre de cylindre (`cylinder`) selon le nombre de gears (`gears`). Nous remarquons que le nombre d'automobile muni de 6 cylindres est très important lorsque le nombre de gears est égal à 3. Alors qu'il n'existe pas de voiture à 6 cylindre et 4 gears.

```

1  library(datasets)
2  data(mtcars)
3  summary(mtcars)
4  Numbers<-table(mtcars$cyl, mtcars$gear)
5  barplot(Numbers, main='Automobile cylinder number grouped by
   number of gears',
6  col=c('red', 'orange', 'blue'), legend=rownames(Numbers), xlab='
   Number of Gears',
7  ylab='count')

```

Listing 2 – Pseudo code de l'Exercice 1

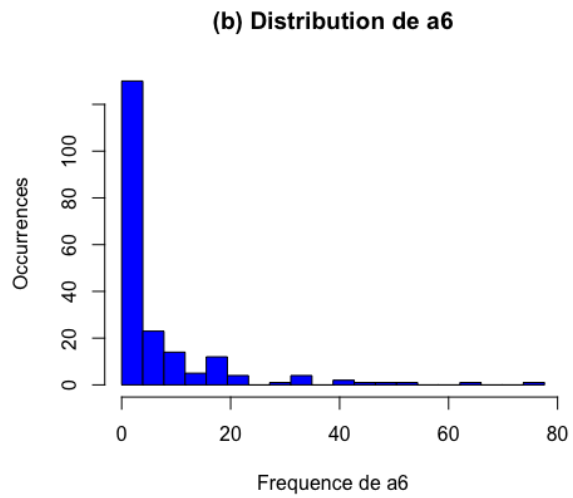
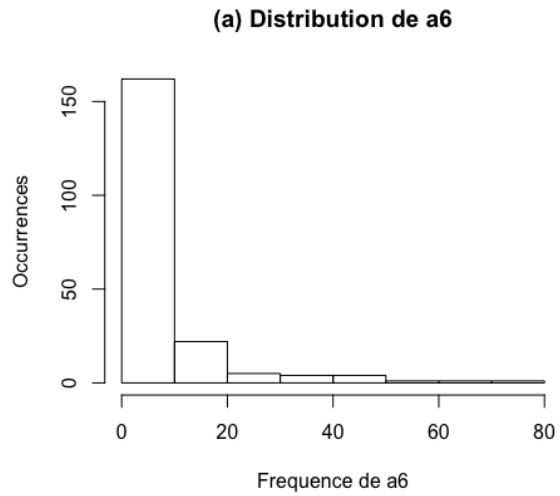


FIGURE 1 – Histogramme de la variable a6

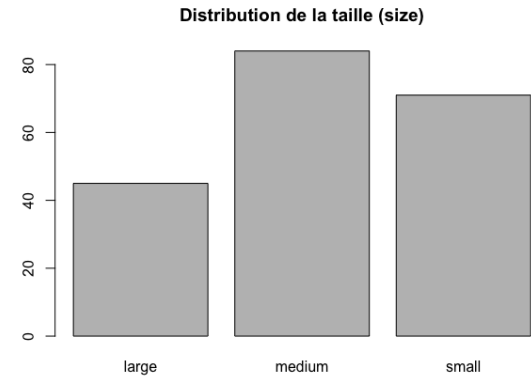


FIGURE 2 – Histogramme de la variable size

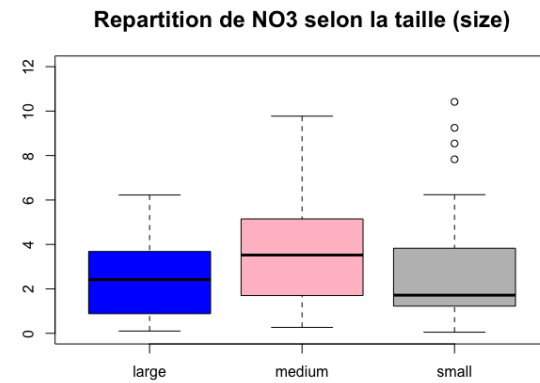


FIGURE 3 – Diagramme en boîte de NO3 selon la taille (size)

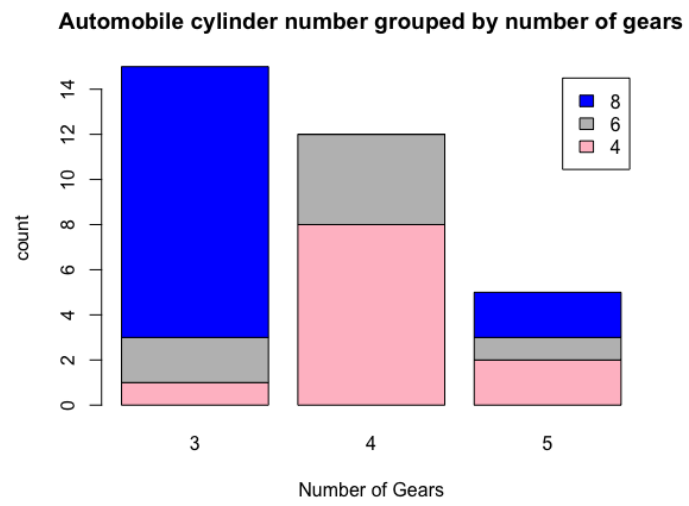


FIGURE 4 – Histogramme de la variable a6