Module 3 - Analyse exploratoire des données

Exercices - Solutions

Exercice 1 : Lecture des données

1. Le téléchargement des données de `iris` est fait grâce à la ligne 3 du Listing 1.
2. L'ensemble des données de `iris` comprennent 3 espèces de fleurs (`setosa`, `virginica`, `versicolor`). Ces espèces sont caractérisées par 4 variables : la longueur et la largeur des sépales `Sepal.Length` et `Sepal.Width` ainsi que la longueur et la largeur des pétales `Petal.Length` et `Petal.Width` .
3. Le nombre d'échantillons dans `iris` est égal à 150. Ceci peut être obtenu par la ligne de code 7.
4. Affichage du contenu des dix premières lignes de la base de données : ligne 8 du Listing 1.

```
1 # Exercice1
2 # Telechargement des donnees
3 data(iris)
4 # Afficher un sommaire de
5 summary(iris)
6 # Nombre d'echantillons de iris
7 dim(iris)
8 # Affichage des 10 premieres lignes de l'objet "MyData"
9 head(iris, 10)
```

Listing 1 – Pseudo code de l'Exercice 1

Exercice 2 : Caractéristiques de tendances centrales et de dispersion

1. La moyenne et les trois quartiles de chaque variable sont résumés dans le Tableau 1. Ils sont calculés grâce à la commande `summary`

Variable	`Sepal.Length`	`Sepal.Width`	`Petal.Length`	`Petal.Width`
Moyenne	5.843	3.057	3.758	1.199
Q1	5.100	2.800	1.600	0.300
Q2	5.800	3.000	4.350	1.300
Q3	6.400	3.300	5.100	1.800

TABLE 1 – Moyenne et les trois quartiles de chacune des quatres variables

2. La moyenne et les trois quartiles de la variable `Sepal.Length` par espèce sont résumés dans le Tableau 2 (obtenus grace aux lignes 5 à 7 du Listing 2).

Espèce	`setosa`	`versicolor`	`virginica`
Moyenne	5.006	5.936	6.588
Q1	4.800	5.600	6.225
Q2	5.006	5.900	4.350
Q3	6.225	6.500	6.900

TABLE 2 – La moyenne et les trois quartiles de la variable `Sepal.Length`

```
1 # Exercice 2
2 # Moyenne et quartiles des variables
3 summary(iris)
4 # Moyenne de la variable Sepal.Length
5 summary(iris$Sepal.Length[iris$Species=="setosa"])
6 summary(iris$Sepal.Length[iris$Species=="versicolor"])
7 summary(iris$Sepal.Length[iris$Species=="virginica"])
```

Listing 2 – Pseudo code de l'Exercice 2

Exercice 3 : Caractéristiques de formes

1. Les coefficients d'asymétrie des quatre variables sont donnés dans le tableau 3. Ils sont obtenus grâce aux lignes de codes 4 à 7 du Listing 3.

Variable	Symétrie
Sepal.Length	0.308
Sepal.Width	0.312
Petal.Length	-0.269
Petal.Width	-0.100

TABLE 3 – Coefficients d'asymétrie

2. Les coefficients d'aplatissement des quatre variables sont donnés dans le tableau 4. Ils sont obtenus grace aux lignes de codes 8 à 11 du Listing 3.

Variable	Aplatissement
Sepal.Length	-0.605
Sepal.Width	0.138
Petal.Length	-1.416
Petal.Width	-1.358

TABLE 4 – Coefficients d'applatissement

```
1  # Exercice 3
2  # load e1071
3  library(e1071)
4  skewness(iris$Sepal.Length)
5  skewness(iris$Sepal.Width)
6  skewness(iris$Petal.Length)
7  skewness(iris$Petal.Width)
8  kurtosis(iris$Sepal.Length)
9  kurtosis(iris$Sepal.Width)
10 kurtosis(iris$Petal.Length)
11 kurtosis(iris$Petal.Width)
```

Listing 3 – Pseudo code de l'Exercice 3

3. Le coefficient d'asymétrie de la variable `Petal.Length` est égal à -0.570 pour l'espèce `versicolor`. Cette valeur étant non nulle, elle laisse prédire que la distribution de cette variable est non symétrique. De plus,cette valeur est négative, donc la moyenne est inférieure à la médiane (`Mean`=4.26 < `Median`=4.35) et l'allure de la distribution est plus décalée vers la droite. Ceci est confirmé par l'histogramme de cette variable donné par la Figure 1.

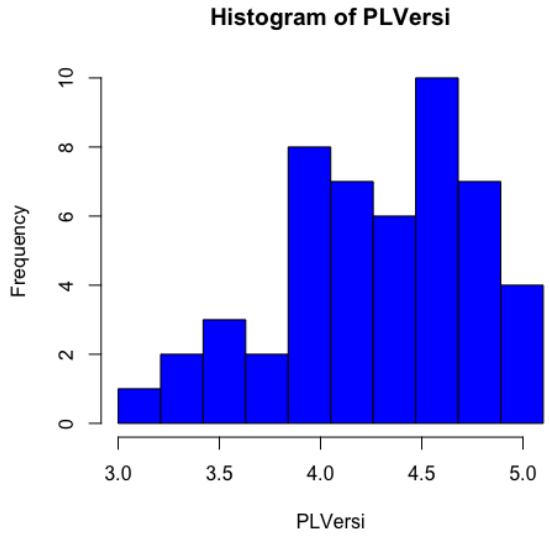

FIGURE 1 – Histogramme de la variable `Petal.Length` pour l'espèce `versicolor`

```
1 # Exercice 3
2 library(e1071)
3 PLVersi = iris$Petal.Length[iris$Species=="versicolor"]
4 skewness(PLVersi)
5 summary(PLVersi)
6 hist(PLVersi, breaks = seq(min(PLVersi), max(PLVersi),
      length.out = 11), col="blue")
```