Tutoriel ggplot2

Tarik HAKAM

10/12/2020

Introduction

La librairie **ggplot2** est une librarie R du package *tidyverse*, développée selon les principes développés par Leland Wilkinson dans son ouvrage *The Grammar of Graphics*.

C'est une librairie de visualisation de données développée par Hadley Wickham.

ggplo2 permet donc la construction de graphiques complexes de manière efficace avec une syntaxe cohérente et puissante.

ggplot2 part du principe que les données relatives à un graphique sont stockées dans un data frame (tableau de données).

Pour commencer, nous allons charger le dataset "iris".

C'est à partir de ces données que nous réaliserons nos tests et que nous créerons les différents types de plots.

```
data("iris")
```

Visualisation du dataset "iris"

Par la commande précédente, vous avez chargé le dataset sur votre environnement.

Nous allons maintenant afficher le 6 premières lignes du data frame afin de prendre connaissance des différentes variables et valeurs disponibles.

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Affichage du nom des colonnes

La commande ***attr*** permet d'afficher le nom des colonnes de votre data frame.

```r
attr(iris,"names")
```

```
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
```

Affichage de statistiques à partir des données

La fonction **summary()** permet d'obtenir la description statistique d'une variable ou d'une table de données.

Pour une variable donnée, la fonction renvoie 5 valeurs : le minimum (Min.), le premier quartile (1st Qu.), la médiane (Median), la moyenne (Mean), le troisième quartile (3rd Qu.) et le maximum (Max).

```r
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

Chargement de la library ggplot2

Au préalable, nous aurons éxécuter la fonction ***install.packages("ggplot2")*** qui permet de télécharger le package **ggplot2** et de l'installer sur notre machine.

Cette commande n'est à exécuter qu'une fois.

Puis exécuter la fonction **library("ggplot2")** qui permet ensuite de charger le package et de rendre les fonctionnalités de celui-ci disponibles (cette fonction est à exécuter à chaque fois que l'on ouvre RStudio).

```r
library("ggplot2")
```

Simple plot sans ggplot2

Pour le premier plot, nous allons le générer sans utiliser **ggplot2** de façon à pouvoir comparer les différentes possibilité de langage et les différences esthétiques.

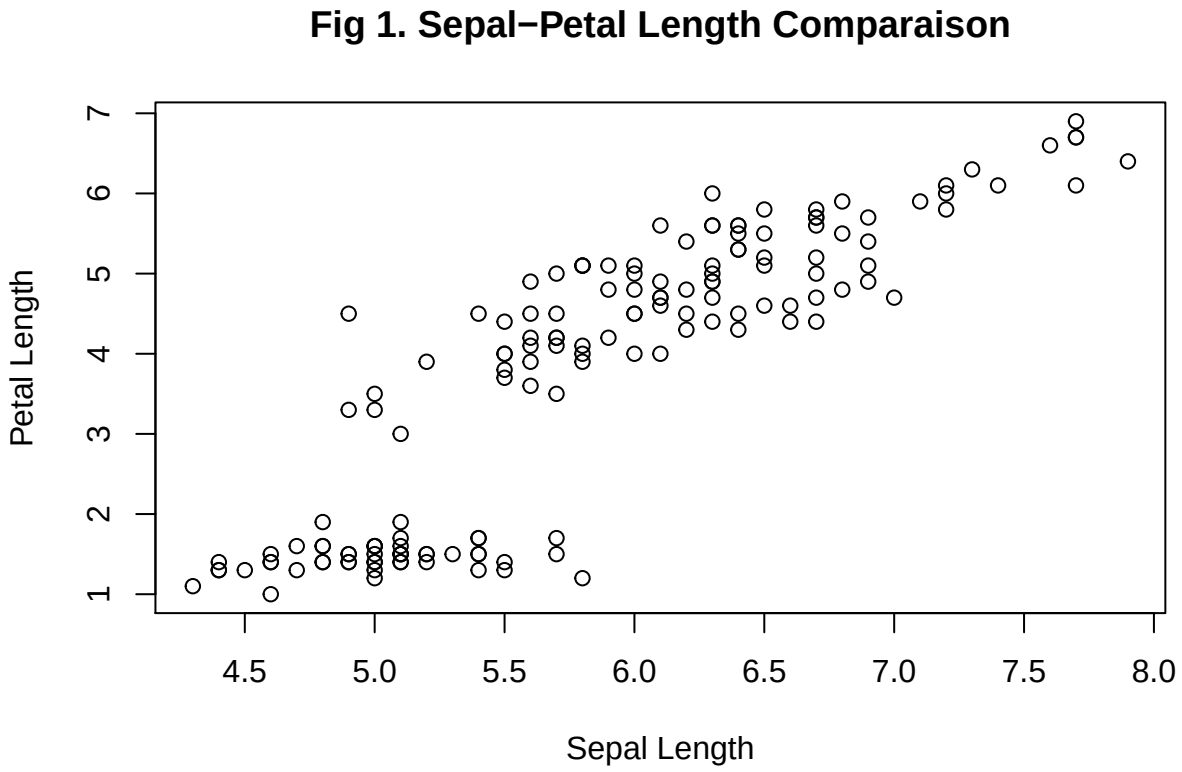
En abscisse, nous choissons d'afficher la longueur des sépales.

Et en ordonnée, nous afficherons la longueur des pétales.

Par défaut, le plot est légendé à l'aide des intitulés des colonnes sélectionnées.

A l'aide des commandes *xlab*, *ylab* et *main*, nous allons respectivement renommer les 2 axes *Sepal Length* et *Petal Lenght* et attribuer le titre suivant au graphique ***Sepal-Petal Length Comparaison***.

```r
plot(x=iris$Sepal.Length, y=iris$Petal.Length,
     xlab="Sepal Length", ylab="Petal Length",  main="Fig 1. Sepal-Petal Length Comparaison")
```



Simple plot avec ggplot2

Pour le 2ème plot, nous allons le gérérer un graphique de nuages de points à l'aide de la librairie **ggplot2**, tout reprenant les mêmes dénominations d'axes et de titre.

Pour se faire, nous allons y ajouter les caractéristiques suivantes :

data = iris : pour extraire les données du dataset "iris"

aes(x = ..., y = ...) : pour définir les valeurs attribuer aux axes x et y

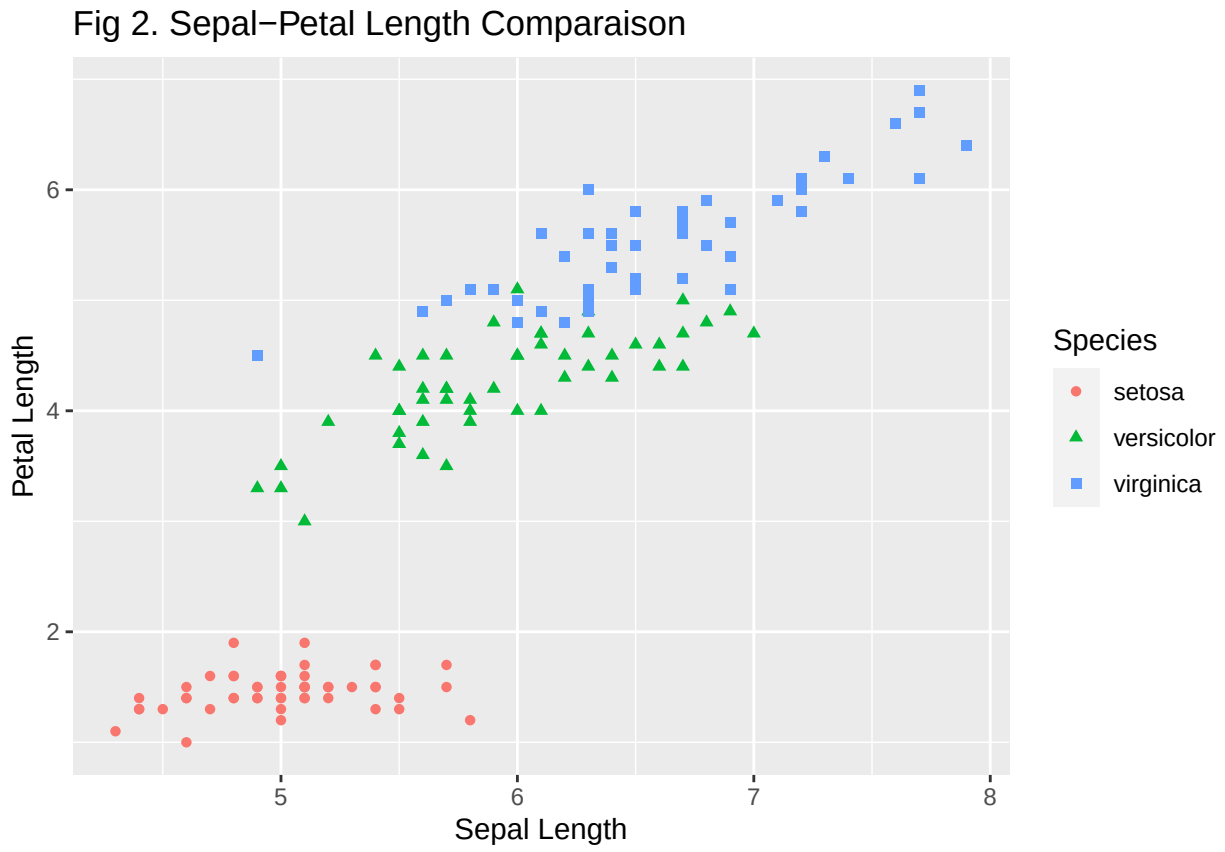
geom_point : pour générer le graphique à points

aes(color=Species, : pour attribuer des couleurs en fonction des espèces

shape=Species) : pour attribuer une forme aux points en fonction des espèces

ggtitle pour attribuer un titre.

```r
scatter <- ggplot(data=iris, aes(x = Sepal.Length, y = Petal.Length))
scatter + geom_point(aes(color=Species, shape=Species)) +
  xlab("Sepal Length") +  ylab("Petal Length") +
  ggtitle("Fig 2. Sepal-Petal Length Comparaison")
```



Histogramme

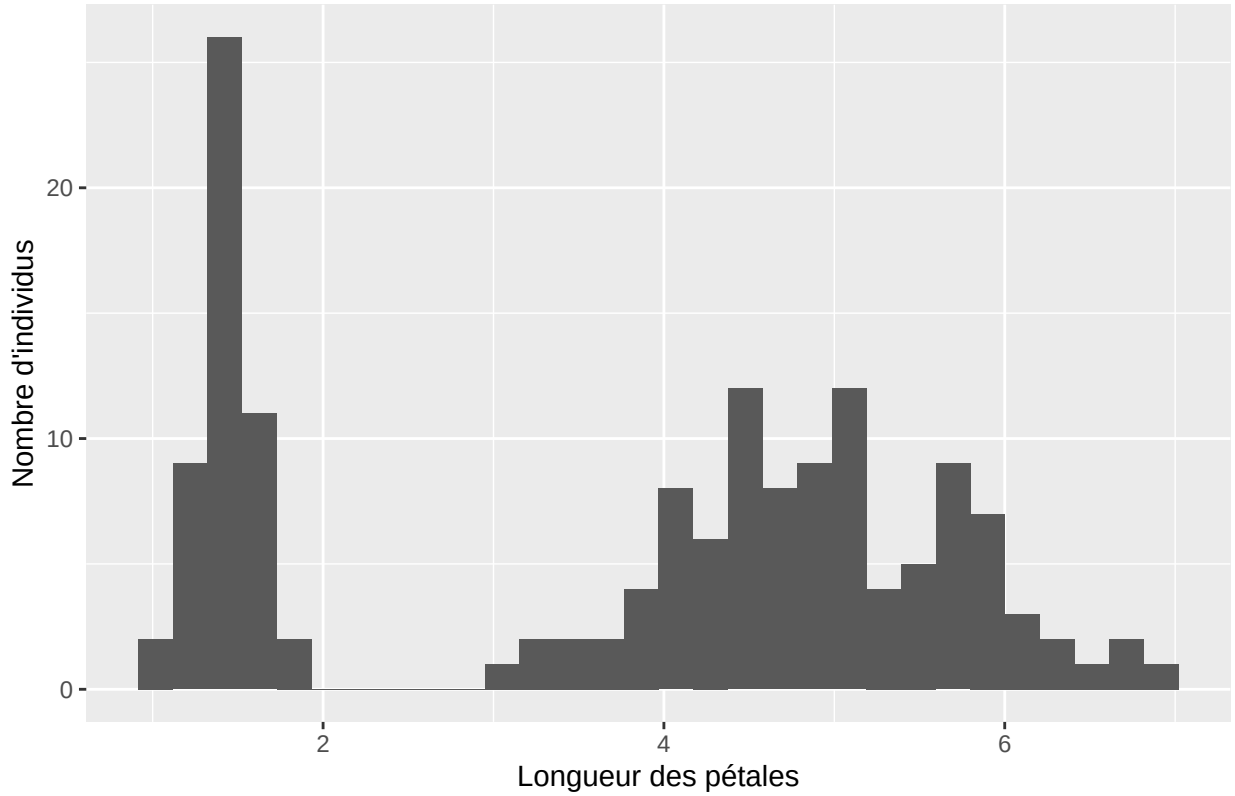
Pour le 3ème plot, sur le même modèle, nous allons générer un histogramme qui a pour objectif d'afficher le nombre d'invidus en fonction de la longueur des pétales.

Vous n'aurez donc qu'une seule variable exprimée en x, la longueur des pétales.

L'ordonnée y, vous affichera le nombre d'individus par population.

```
hist_p <- ggplot(iris, aes(Petal.Length)) + geom_histogram() +
  xlab("Longueur des pétales") +  ylab("Nombre d'individus") +
  ggtitle("Fig 3. Nombre d'individus en fonction de la longueur des pétales")
hist_p
```

Fig 3. Nombre d'individus en fonction de la longueur des pétales

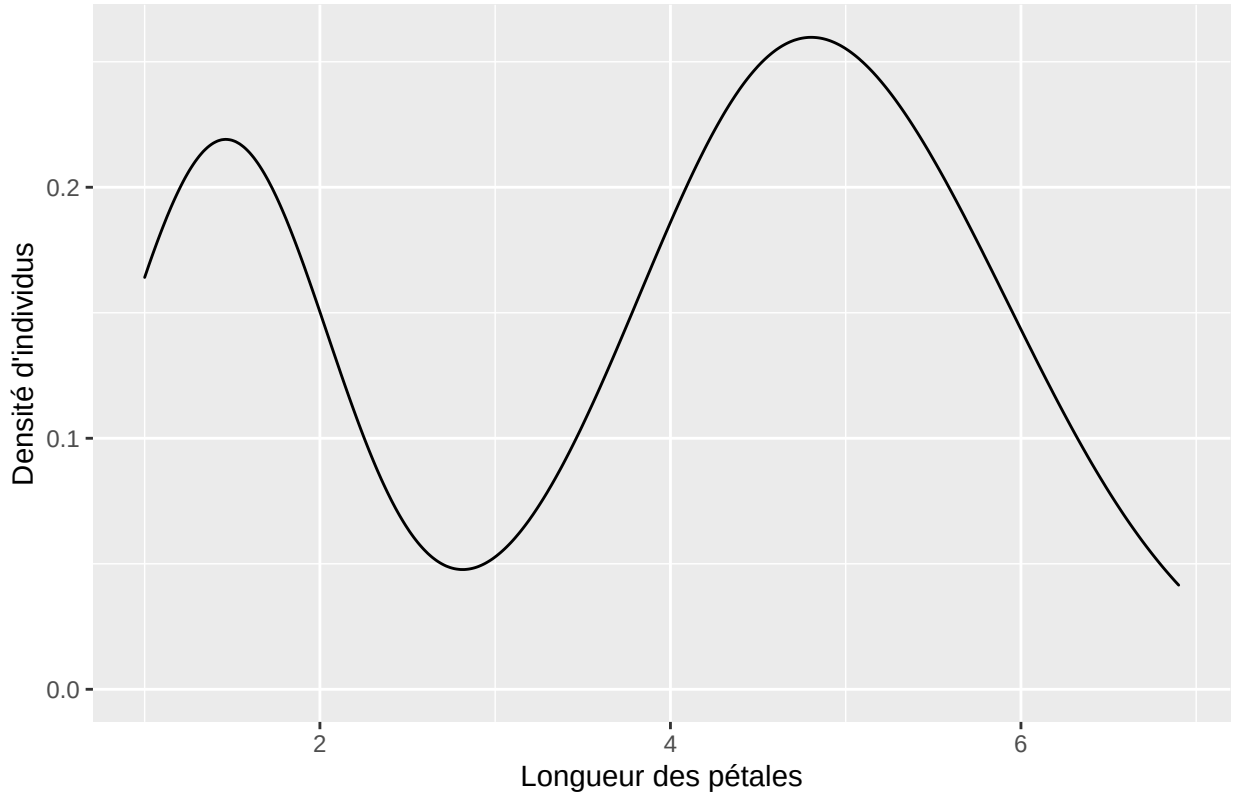


Fonction de densité

Pour le 4ème plot, nous allons générer la fonction de densité de la longueur des pétales de l'ensemble des espèces.

```
ggplot(iris, aes(Petal.Length)) + geom_density() +
  xlab("Longueur des pétales") +  ylab("Densité d'individus") +
  ggtitle("Fig 4. Fonction de densité de la longueur des pétales")
```

Fig 4. Fonction de densité de la longueur des pétales

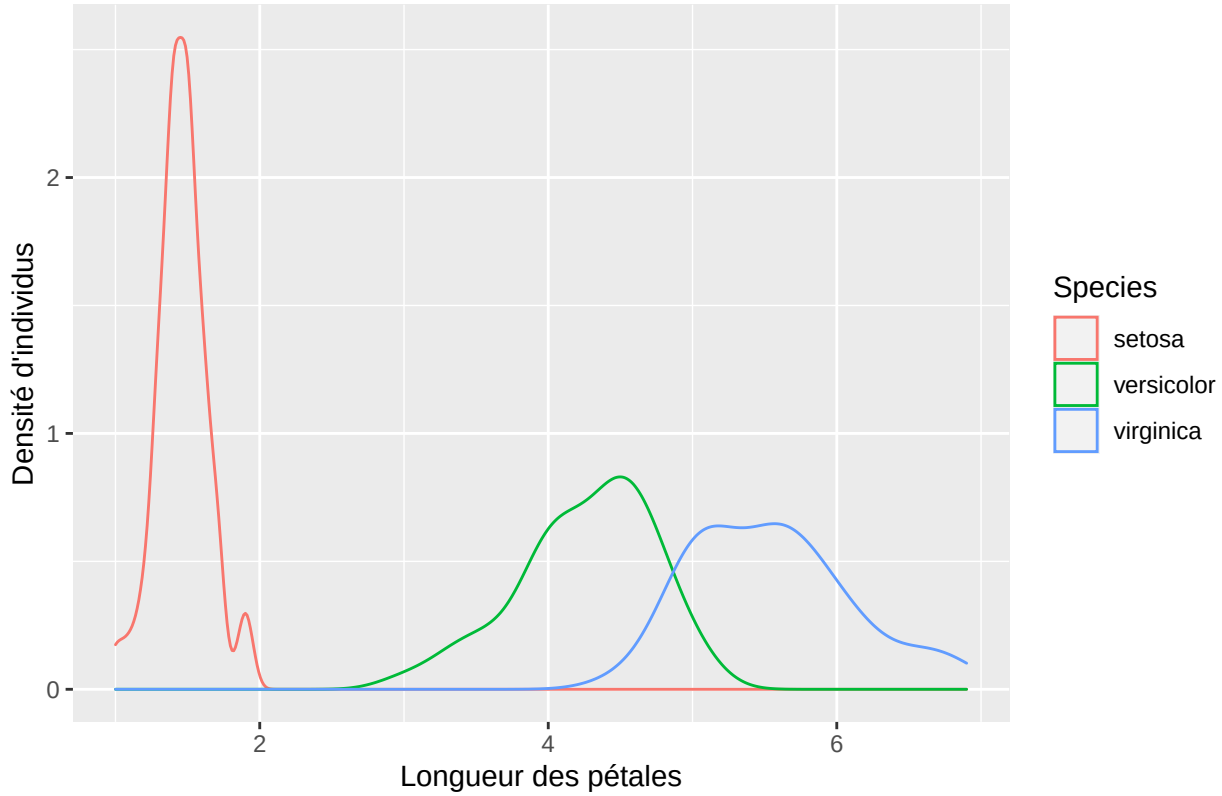


Fonction de densité en fonction des espèces

Pour le 5ème plot, nous allons générer la fonction de densité de la longueur des pétales par espèces dans le même repère.

```
ggplot(iris, aes(Petal.Length, color = Species)) + geom_density() +
  xlab("Longueur des pétales") +  ylab("Densité d'individus") +
  ggtitle("Fig 5. Fonction de densité de la longueur des pétales par espèce")
```

Fig 5. Fonction de densité de la longueur des pétales par espèce



Boite à moustache

Pour le 6ème plot, nous allons générer une boite à moustache de la longueur des pétales par espèces.

Une boîte à moustache ou encore appelé box-plot est un graphique tout simple qui permet de résumer une variable de manière simple et visuel, d'identifier les valeurs extrêmes et de comprendre la répartition des observations.

Ci-dessous, vous trouverez les détails sur le sens de lecture de ce genre graphique afin de l'utiliser simplement : - La valeur centrale du graphique est la médiane (il existe autant de valeur supérieures qu'inférieures à cette valeur dans l'échantillon).

Les bords du rectangle sont les quartiles :

- Pour le bord inférieur, un quart des observations ont des valeurs plus petites et trois quart ont des valeurs plus grandes.
- Le bord supérieur suit le même raisonnement.
- Les extrémités des moustaches sont calculées en utilisant 1.5 fois l'espace interquartile (la distance entre le 1er et le 3ème quartile).

On peut remarquer que 50% des observations se trouvent à l'intérieur de la boîte.

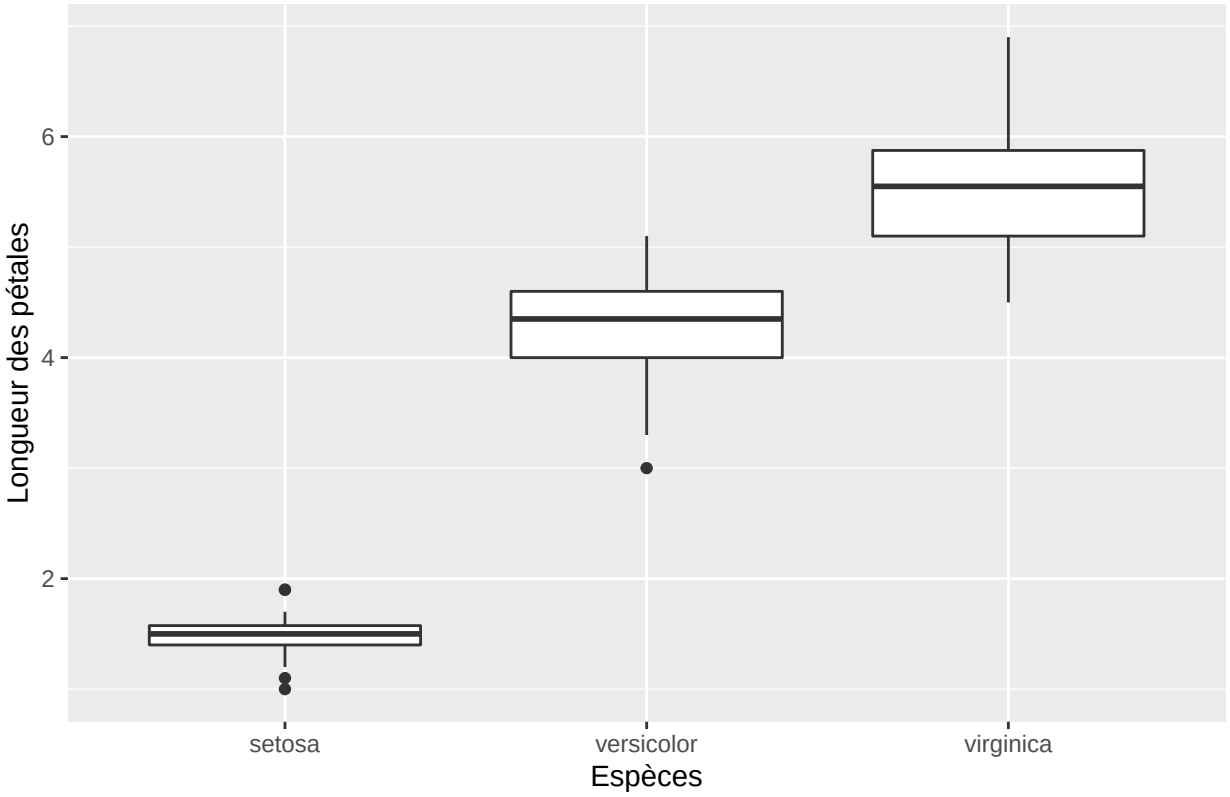
Les valeurs à l'extérieur des moustaches sont représentées par des points.

On ne peut pas dire que si une observation est à l'extérieur des moustaches alors elles est une valeur aberrante.

Par contre, cela indique qu'il faut étudier plus en détail cette observation.

```
ggplot(iris, aes(x = Species, y = Petal.Length)) + geom_boxplot() +
  xlab("Espèces") +  ylab("Longueur des pétales") +
  ggtitle("Fig 6. Boite à moustache de la longueur des pétales par espèce")
```

Fig 6. Boite à moustache de la longueur des pétales par espèce



Split des fonctions de densité

Pour le 7ème plot, nous allons générer les fonctions de densité de la longueur des pétales par espèces dans 3 repères distincts.

```
ggplot(iris, aes(x = Petal.Length, color = Species)) + geom_density() + facet_wrap(~ Species) +
  xlab("Longueur des pétales pour chaque espèce") +  ylab("Densité d'individus") +
  ggtitle("Fig 7. Fonction de densité de la longueur des pétales par espèce")
```

Fig 7. Fonction de densité de la longueur des pétales par espèce

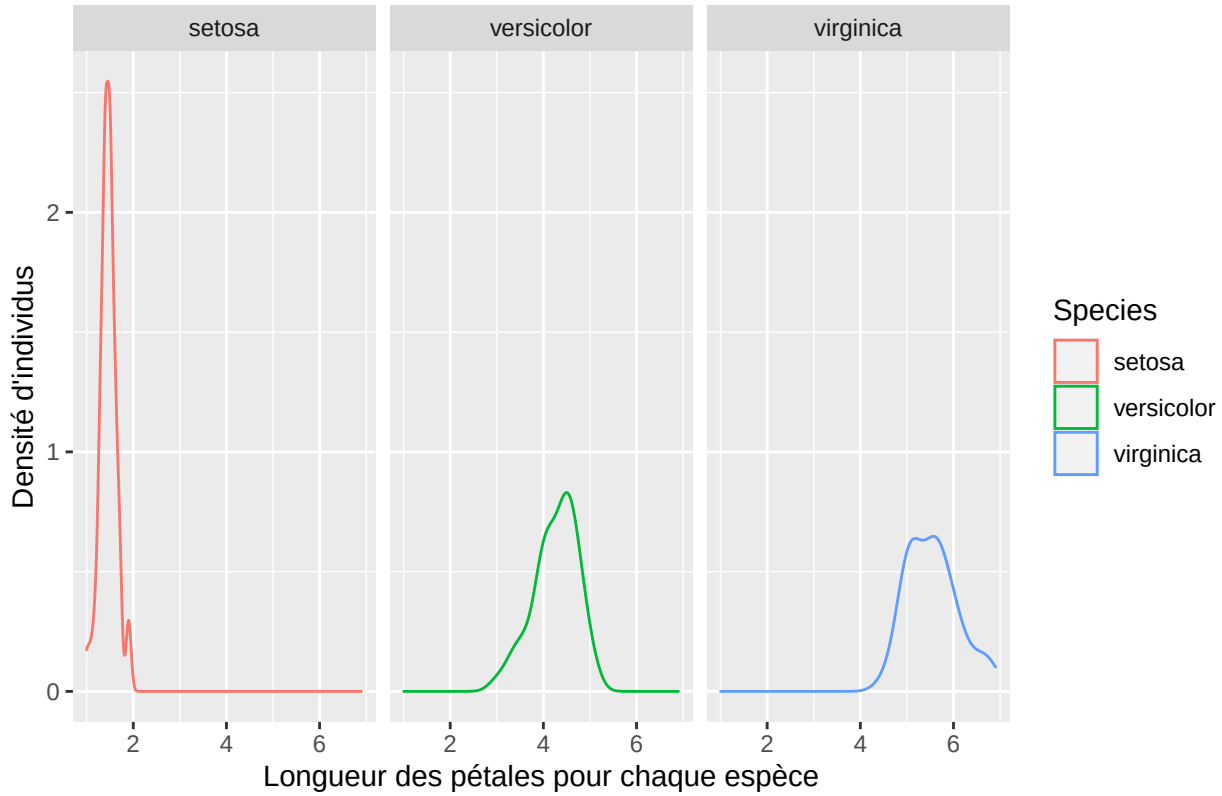


Diagramme en batons

Pour le 8ème et dernier plot, nous allons générer un diagramme en batons de la variable espèce.

```
b <- ggplot(iris, aes(x=Species))
b + geom_bar(aes(fill = Species)) +
  xlab("Espèces") +  ylab("Nombre d'individus") +
  ggtitle("Fig 8. Nombre d'individus en fonction des espèces")
```

Fig 8. Nombre d'individus en fonction des espèces