Tutoriel ggplot2 - Auto-évaluation

Tarik HAKAM

22/12/2020

1. Critères d'évaluation

1. Comportement du Rmd à l'exécution
2. Qualité de la rédaction du dossier
3. Accessibilité, didactisme et pertinence du dossier
4. Qualité et lisibilité du Rmarkdown
5. Qualité des applications permettant d'illustrer le package

2. Lien vers le document commenté

En cliquant **ici**, vous trouverez le lien menant à mon GitHub hébergeant le fruit de ma réalisation.

3. Auteurs du document commenté

Le document évalué dans le cadre de ce rendu a été produit par moi-même, étudiant en Master 2 Data Management à Paris School of Business.

4. Synthèse du document

A travers ma réalisation, vous retrouverez une brève introduction de la librairie **ggplot2**, puis différents types de plots réalisés à l'aide du dataset `iris`.

Vous retrouverez une guidance et des explications afin de réaliser vous-même des :

- plots simple plot sans ggplot2
- plots simple plot avec ggplot2
- histogrammes
- fonctions de densité
- fonctions de densité en fonction des espèces
- boite à moustache
- split de fonctions de densité
- diagrammes en batons

5. Extrait commenté des parties de code

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***.

```
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.

```
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
```

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")
```

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")
```

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")
```

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")
```

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")
```

6. Evaluation du travail suivant les 5 critères précités

1. Comportement du Rmd à l'exécution

L'exécution du code Rmd se réalise parfaitement.

2. Qualité de la rédaction du dossier

La rédaction de ce document me semble de bonne qualité.

Elle suit cette logique de se vouloir être un tutoriel accessible autour des fonctions du package ggplot2 de manière efficace, suivant un fil conducteur.

3. Accessibilité, didactisme et pertinence du dossier

La lecture de ce dossier me semble également aisée et accessible.

Les descriptions sont plutôt explicites et suffisamment illustrées.

Je pense pouvoir faire adhérer le lecteur à ma production. Le document a vocation à transmettre une connaissance et j'espère que la finalité didactique sera profitable aux lecteurs.

4. Qualité et lisibilité du RMarkdown

Le RMarkdown me semnle bien écrit, lisible et aéré.

je pense qu'il s'agit d'un code bien réalisé.

5. Qualité des applications permettant d'illustrer le package

Les applications me sont simples, de qualités et maitrisées.

7. Conclusion

Selon moi, il s'agit globalement d'un tutoriel profitable.

Je pense avoir fourni un dossier recherché, documenté et accessible.

Ce document apporte une bonne approche pour découvrir ce package.

Vous retrouvez ce document sur mon **GitHub**.