

Aide pour le rapport du projet

C.Maugis-Rabusseau & O. Roustant

2023-09-29

Contents

1	Consignes de rédaction	1
2	Quelques éléments de Rmarkdown	1
3	Etude du jeu de données Iris	2
3.1	Récupération des données	2
3.2	Statistiques descriptives	3
4	Reference	4

Ce document a pour but de vous donner des consignes pour la rédaction du rapport et vous donner des éléments de démarrage pour la rédaction d'un rapport en Rmarkdown.

1 Consignes de rédaction

Votre rapport doit synthétiser votre travail d'étude du jeu de données abordé durant le projet commun. Il doit comprendre :

- une organisation par sections, sous-sections, ... une introduction et une conclusion
- pour chaque méthode d'analyse considérée : expliquer son principe et l'objectif, la mettre en application, commenter les résultats
- Toute figure doit avoir une légende et doit être commentée
- Même remarque pour les tableaux de résultats
-

2 Quelques éléments de Rmarkdown

Des éléments de rédaction d'un document Rmarkdown sont donnés dans la partie 4 des tutoriels R (ici)

Vous pouvez également trouver de nombreux exemples sur le web.

Nous rappelons ici quelques points :

- Vous devez créer un documents Rmarkdown au format de sortie PDF
- Vous pouvez organiser votre document en sections, sous-sections, ... grâce à #, ##, ...
- Vous pouvez dans l'en-tête de votre document Rmarkdown
 - préciser le titre du document, les auteurs, la date, ...
 - ajouter une table des matières avec l'option `toc` dans le `output:pdf_document`
 - ajouter des macro Latex dans `header-includes`
 - ajouter une bibliographie avec `bibliography`:
 - ...
- Vous pouvez mettre du code R dans des chunks R et jouer sur les options comme
 - `echo=F` pour ne pas afficher le code dans le rapport
 - `eval=F` pour ne pas l'évaluer dans le rapport
 - `fig.height`, `fig.width`, ... pour maîtriser la taille des figures
 - `fig.cap` pour mettre une légende aux figures
 - `message=F` pour ne pas afficher les messages de R
 - ...
- Vous pouvez mettre des formules mathématiques Latex entre `$...$`. On peut aussi utiliser `\begin{equation} \end{equation}`, ...
- Vous pouvez mettre également du code python dans des chunks python et en utilisant la librairie `reticulate` (voir tutoriel)

3 Etude du jeu de données Iris

On va ici utiliser le célèbre jeu de données des Iris pour illustrer quelques points de rédaction en Rmarkdown. Vous êtes donc invités à parcourir en même temps le `.pdf` et le `.Rmd` pour comprendre les points de syntaxe.

3.1 Récupération des données

Les données Iris ont été collectées par Edgar Anderson [`@iris`]. Ce sont les mesures en centimètres des variables suivantes : longueur du sépale (`Sepal.Length`), largeur du sépale (`Sepal.Width`), longueur du pétale (`Petal.Length`) et largeur du pétale (`Petal.Width`) pour trois espèces d'iris : *Iris setosa*, *I. versicolor* et *I. virginica*. Les données sont disponibles de base sous R et on les récupère donc avec la fonction `data(iris)`. On affiche ici les premières lignes du jeu de données :

##	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

On retrouve bien que les données sont composées de 150 individus, de 4 variables quantitatives et d'une variable qualitative *Species*. Dans la suite, nous notons Y la variable *Species* et X la matrice composées des 4 autres variables

$$X = (X_{ij}), \quad i \in \{1, \dots, 150\}, \quad j \in \{1, \dots, 4\}.$$

On peut facilement repasser les données en python à l'aide de la librairie `reticulate` et les commandes suivantes :

3.2 Statistiques descriptives

Nous faisons ici quelques statistiques descriptives pour prendre en main les données.

3.2.1 La variable *Species*

Nous commençons par la variable *Species* (vecteur Y) qui est une variable qualitative. La figure suivante nous permet de contrôler que nous avons bien 50 individus par espèce.

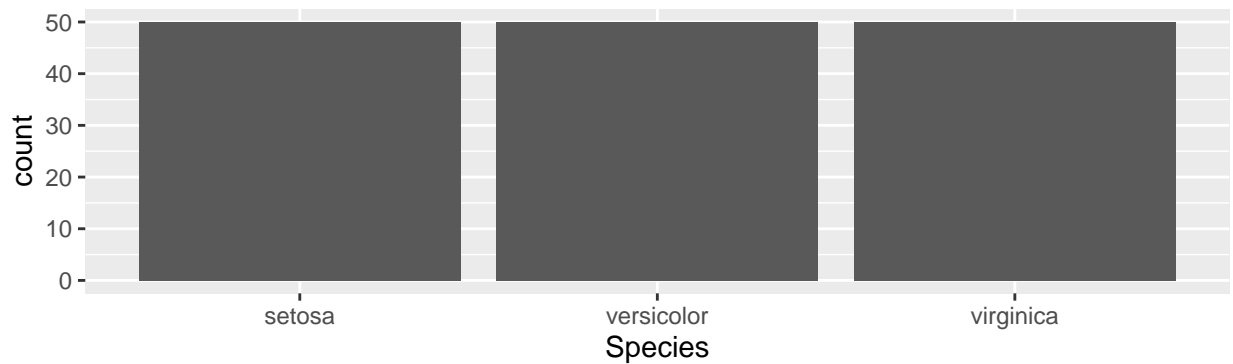


Figure 1: Barplot de la variable Species.

3.2.2 Les 4 variables quantitatives

Nous nous intéressons ici aux 4 variables quantitatives (matrice X). La figure suivante montre les corrélations entre les 4 variables. On peut remarquer que la largeur et la longueur des pétales sont fortement corrélées positivement, ce n'est pas le cas pour les sépales.

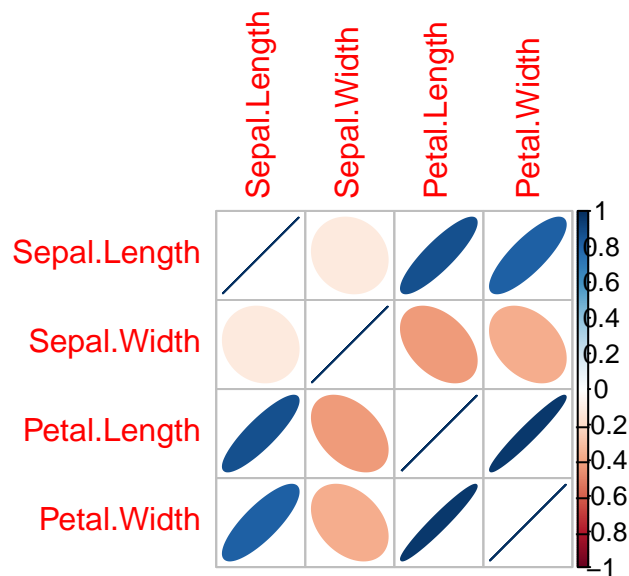
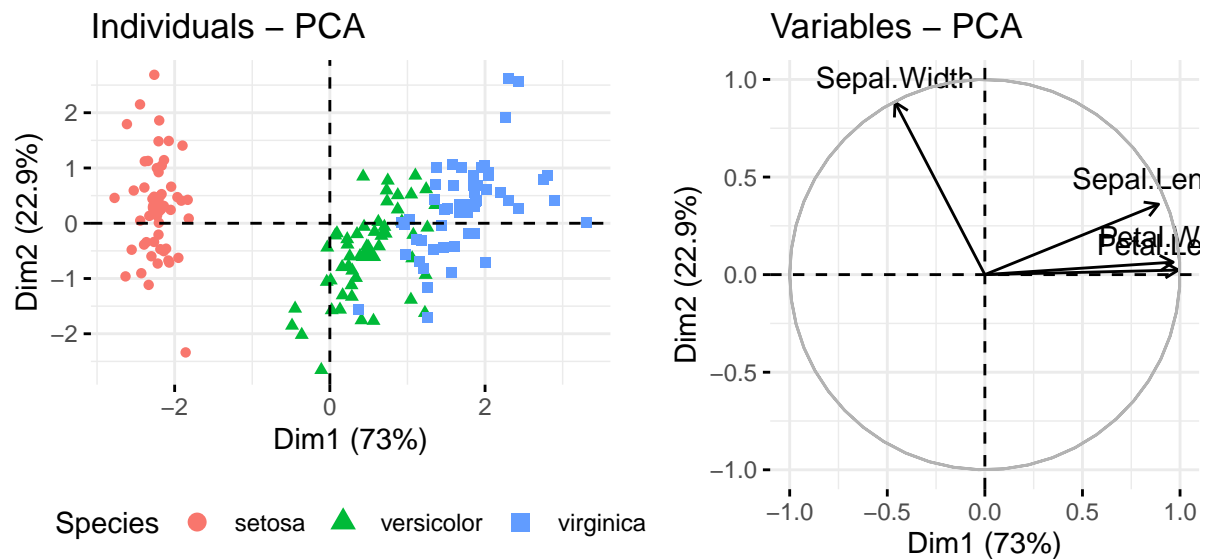


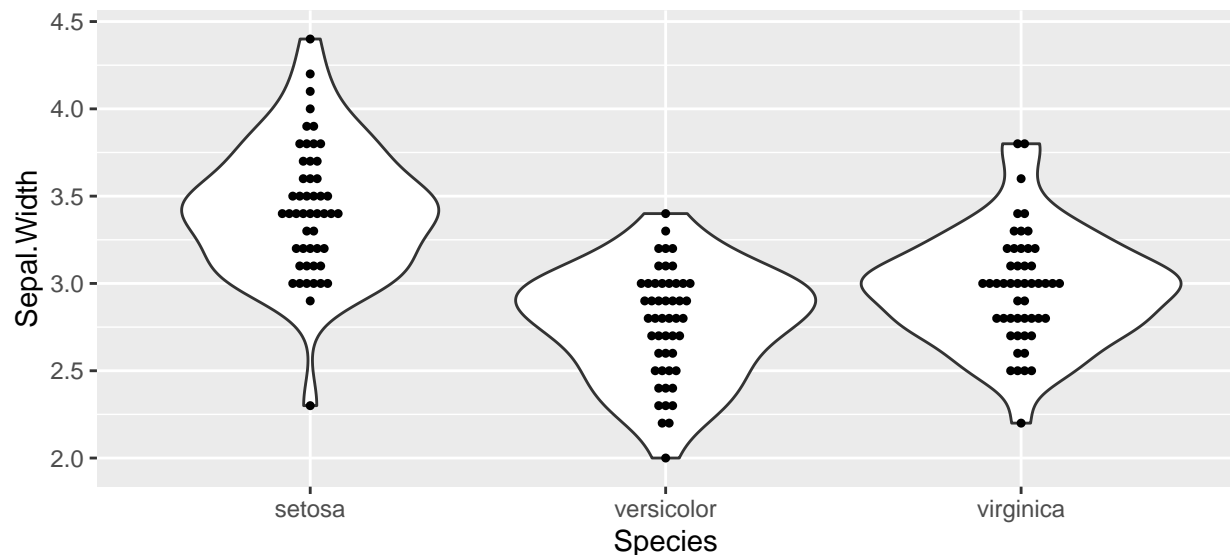
Figure 2: Matrice des corrélations entre les 4 variables quantitatives

A l'aide de la librairie **FactoMineR**, on met en place une ACP. On représente ici les individus projetés sur le premier plan factoriel, la couleur correspondant à l'espèce ainsi que les corrélations des variables quantitatives initiales avec les deux premières composantes principales.

```
respca=PCA(iris,quali.sup=5,graph=F)
g1=fviz_pca_ind(respca,label="none",habillage=5)+ theme(legend.position="bottom")
g2=fviz_pca_var(respca)
grid.arrange(g1,g2,ncol=2)
```



On peut remarquer que les *Setosa* se distinguent des deux autres espèces principalement par la largeur de leur sépales. On peut appuyer ce point à l'aide de la figure suivante



4 Reference