

Initiation à R Studio

4 juillet 2024



Layan Fessler

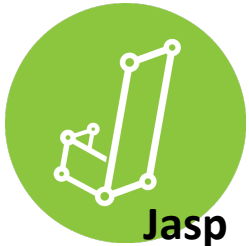
Layan.fessler@univ-grenoble-alpes.fr

Silvio Maltagliati

maltagli@usc.edu



QUELS LOGICIELS POUR L'ANALYSE DE DONNÉES ?



POUR

Facilité d'utilisation, interface graphique intuitive

Facile pour les tests statistiques courants, utilisation de **R** en arrière-plan

Tous types de statistiques, bonne interface graphique, support technique

CONTRE

Petits jeux de données, analyse statistique de **base**, **faible reproductivité**

Dépendant des **modules disponibles**, moins puissant pour les **analyses complexes**, **faible reproductivité**

Coûteux, moins **flexible** que R, scripts SPSS **moins courants** que les scripts R, **faible reproductivité**



- **Langage de programmation open source**
- Interface de **ligne de commande**
- Écrit dans **différents langages** : C, C++, FORTRAN et Java
- Système de « **packages** »
- **Nommé R** en raison des initiales de ses auteurs : Ross Ihaka et Robert Gentleman

POUR

Gratuit et open-source, extrêmement **flexible et puissant**, large **communauté** et **support**, **forte reproductibilité**

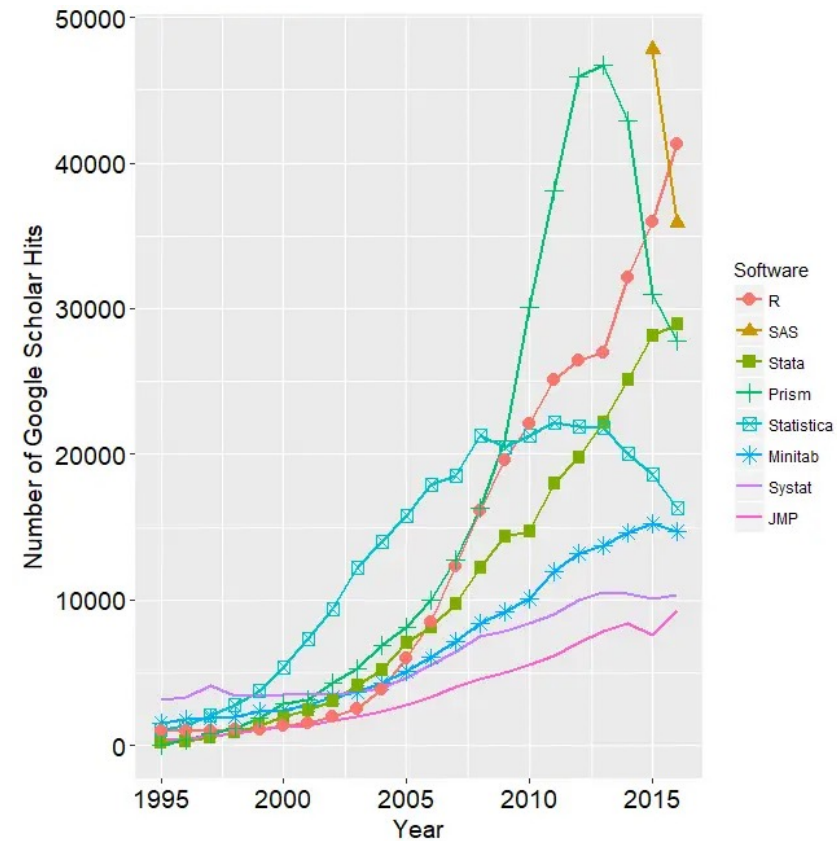
CONTRE

Courbe d'apprentissage plus élevée, nécessite une certaine familiarité avec la **programmation**

Logiciels « clique-bouton »

POURQUOI CHOISIR R ?

De plus en plus utilisé



<https://r4stats.com/articles/popularity/>

Favorise la reproductibilité et la science ouverte

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Ouvrir la science! OPEN SCIENCE OUR ACTIONS RESOURCES NEWS WHO ARE WE?

FR - EN

Acting in favour of open and shared scientific research

The French Committee for Open Science ensures the implementation of the National Open Science Policy.

→ [Discover the French Plan for Open Science](#)

→ [Read about the Committee's projects](#)

Axe 3

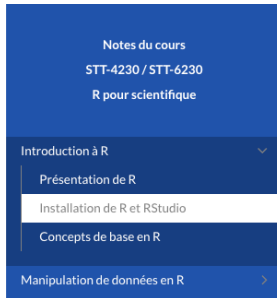
Opening up and promoting source code produced by research

Software plays a key role in scientific research, and it can be a tool, a result, and a research object. Making software source code available, with the option of modifying, reusing and disseminating them, is a major requirement to ensure the reproducibility of scientific findings and to support the creation and sharing of knowledge, in keeping with the open science ethos.

«The opening of software source codes is a major challenge for the reproducibility of scientific results.»

SHARE ON TWITTER

COMMENT INSTALLER R ET SON INTERFACE R STUDIO ?



STT-4230 > Introduction à R > Guide d'installation ou de mise à jour de R et RStudio

Guide d'installation ou de mise à jour de R et RStudio

Sophie Baillargeon, Université Laval

2021-01-15

Version PDF de cette page web : [installation_r_2021.pdf](#)

https://stt4230.rbind.io/introduction/installation_r_rstudio/



[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting Bugs
Conferences
Search
Get Involved: Mailing Lists
Get Involved: Contributing
Developer Pages
R Blog

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

News

- **R version 4.4.1 (Race for Your Life)** has been released on 2024-06-14.
- We are deeply sorry to announce that our friend and colleague Friedrich (Fritz) Leisch has died. [Read our tribute to Fritz here.](#)
- **R version 4.4.0 (Puppy Cup)** has been released on 2024-04-24.
- **R version 4.3.3 (Angel Food Cake)** (wrap-up of 4.3.x) was released on 2024-02-29.
- **Registration for userR! 2024** has opened with early bird deadline March 31 2024.

<https://www.r-project.org/>



De nombreux tutoriel sur YouTube



https://www.youtube.com/watch?v=YrEe2TLr3MI&ab_channel=TechRelatedTips



https://www.youtube.com/watch?v=i5WIMX4LK8M&ab_channel=ProgrammingKnowledge

How to Download and Install RStudio 2024

211 k vues • il y a 1 an

Tech Related Tips

This video guides about how to install rstudio in your computer. To

How to Install R and RStudio on Mac / MacOS (2024)

18 k vues • il y a 5 mois

ProgrammingKnowledge

"R Ready: How to Install R and RStudio on Mac | Quick Setup Tutorial" Welcome to

PREMIERS PAS DANS R STUDIO

- Créer un nouveau **document de** travail « R Notebook »

(File → New file → R Notebook)

- Créer un « **markdown** » pour lancer une commande de base

(Code → Insert chunk)

- Créer son premier « **vecteur** » à partir d'une commande de base

```
a <- 3 + 3
```

```
print(a) # afficher le vecteur
```

```
b <- c(1, 5, 10)
```

```
c <- c(1:10)
```

```
d <- c(« rouge », « bleu », « vert »)
```

- Convertir un vecteur en **tableau**

```
c <- as.data.frame(c)
```

PREMIERS PAS DANS R STUDIO

- Installer un **package** et charger sa bibliothèque de packages

Installer : (*Packages* → *Install* → « entrer le nom du package » → *Install*)

Méthode alternative dans un markdown : `install.packages(« nom du package »)`

Charger : Dans un markdown : `library(nom du package)`

Exemple : Installer et charger les packages suivants : « dplyr », « psych » et « readxl »



dplyr

Manipulation de données



psych

Analyses psychométriques
et psychologiques



readxl

Importation de fichier Excel
(.xls et .xlsx)

PREMIERS PAS DANS R STUDIO

■ Importer un jeu de données

Méthode 1 (Manuelle)

Import Dataset → From Excel → Sélection du fichier → Import

Méthode 2 (Automatique)

1. Localiser et Définir son « Working Directory »

Par défaut, R stocke vos travaux dans le répertoire où R est installé. Cependant, pour une meilleure organisation et pour éviter de changer sans arrêt entre les données et les travaux effectués, il est recommandé de **définir un répertoire spécifique où stocker vos scripts, vos données et vos résultats**.

Bonne pratique : localiser son working directory à chaque début de session R

```
```{r}
setwd("/Users/layanfessler/Library/CloudStorage/OneDrive-UniversitéGrenobleAlpes/ENCADREMENT/STAGE D'EXCELLENCE/2024/Atelier R")
```
```

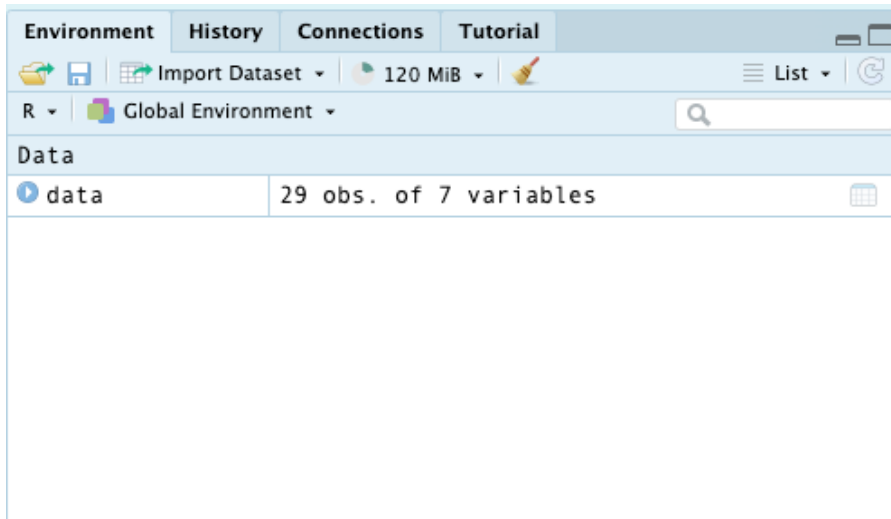
PREMIERS PAS DANS R STUDIO

■ Importer un jeu de données

Méthode 2 (Automatique)

1. Localiser et Définir son « Working Directory »
2. Importer son jeu de données (e.g., « patisseries_JO_2024.xlsx »)

```
# Localiser et définir son working directory
```{r}
working directory
setwd("/Users/layanfessler/Library/CloudStorage/OneDrive-UniversitéGrenobleAlpes/ENCADREMENT/STAGE D'EXCELLENCE/2024/Atelier R")
data <- read_excel("patisseries_JO_2024.xlsx") # Charger son jeu de données
```
```



PREMIERS PAS DANS R STUDIO

■ Inspecter son jeu de données

- Obtenir le nom des variables : `names(data)`
- Obtenir le nombre de lignes et de colonnes : `nrow(data)` ; `ncol()` ; `dim()`
- Visualiser les données d'un tableau / les premières données : `View(data)` ; `head(data)`
- Obtenir des informations sur le « type » de variables : `glimpse(data)`



1. Entiers : ``integer`` ou ``int``
2. Réels (nombres à virgule flottante) : ``numeric`` ou ``dbl`` (abréviation de "double" précision)
3. Chaînes de caractères : ``character`` ou ``chr``
4. Booléens : ``logical`` ou ``lgl``
5. Facteurs : ``factor``
6. Complexes : ``complex``

PREMIERS PAS DANS R STUDIO

- **Obtenir nos premiers indicateurs sur les variables**

- Identifier (et quantifier) des données manquantes :

Vérifier s'il y a des données manquantes

```
missing_rows <- apply(is.na(data), 1, any)
```

Nombre de données manquantes

```
sum(missing_rows)
```

Localisation des données manquantes

```
print(data[missing_rows, ])
```



Exercice

1. Y'a-t-il des **données manquantes** ?
2. Si oui, **combien**, pour quelle **variable** et quel **participant** ?

PREMIERS PAS DANS R STUDIO

- **Obtenir nos premiers indicateurs sur les variables**
 - Identifier (et quantifier) des données manquantes
 - Résumé global d'un tableau : *summary(data)*



```
##{r}
# résumé global
summary(data)
##
```

| Code_participant | Specialite_sportive | Sexe | Age | Tiramisu | Flan | Moelleux |
|------------------|---------------------|------------------|---------------|---------------|---------------|---------------|
| Min. : 1 | Length:29 | Length:29 | Min. :18.00 | Min. :1.000 | Min. :1.000 | Min. :2.000 |
| 1st Qu.: 8 | Class :character | Class :character | 1st Qu.:21.00 | 1st Qu.:4.000 | 1st Qu.:3.000 | 1st Qu.:5.000 |
| Median :15 | Mode :character | Mode :character | Median :24.00 | Median :5.000 | Median :4.000 | Median :6.000 |
| Mean :15 | | | Mean :23.03 | Mean :4.724 | Mean :4.357 | Mean :5.345 |
| 3rd Qu.:22 | | | 3rd Qu.:24.00 | 3rd Qu.:6.000 | 3rd Qu.:6.000 | 3rd Qu.:6.000 |
| Max. :29 | | | Max. :30.00 | Max. :7.000 | Max. :7.000 | Max. :7.000 |
| | | | | NA's :1 | | |

PREMIERS PAS DANS R STUDIO

- **Obtenir nos premiers indicateurs sur les variables**
- Identifier (et quantifier) des données manquantes
- Résumé global d'un tableau : *summary(data)*
- Résumé d'une seule variable **numérique** : *summary(data\$Flan)*
- **Autre fonction (très) utile** : *describe(data)*

```
## R
describe(data) # tout le jeu de données
describe(data$Flan) # Une seule variable
```



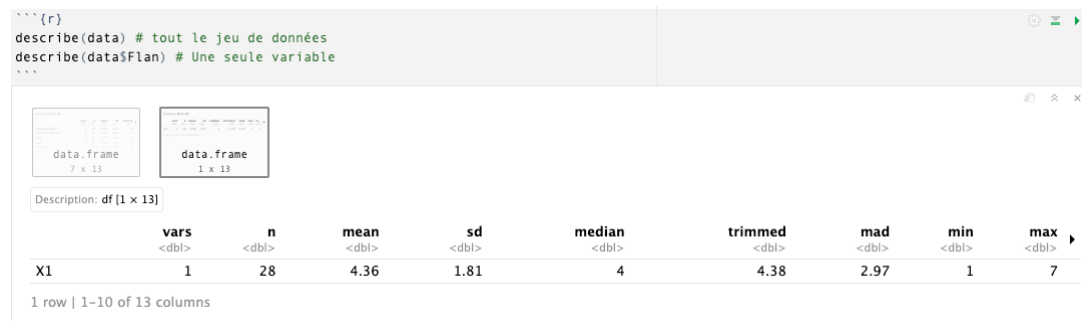
Description: df [7 x 13]

| | vars
<dbl> | n
<dbl> | mean
<dbl> | sd
<dbl> | median
<dbl> | trimmed
<dbl> | mad
<dbl> | min
<dbl> | max
<dbl> |
|----------------------|---------------|------------|---------------|-------------|-----------------|------------------|--------------|--------------|--------------|
| Code_participant | 1 | 29 | 15.00 | 8.51 | 15 | 15.00 | 10.38 | 1 | 29 |
| Specialite_sportive* | 2 | 29 | 2.00 | 0.89 | 2 | 2.00 | 1.48 | 1 | 3 |
| Sexe* | 3 | 29 | 1.55 | 0.51 | 2 | 1.56 | 0.00 | 1 | 2 |
| Age | 4 | 29 | 23.03 | 3.06 | 24 | 22.96 | 2.97 | 18 | 30 |
| Tiramisu | 5 | 29 | 4.72 | 1.85 | 5 | 4.84 | 1.48 | 1 | 7 |
| Flan | 6 | 28 | 4.36 | 1.81 | 4 | 4.38 | 2.97 | 1 | 7 |
| Moelleux | 7 | 29 | 5.34 | 1.29 | 6 | 5.44 | 1.48 | 2 | 7 |

7 rows | 1-10 of 13 columns

describe(data\$Flan)

```
## R
describe(data) # tout le jeu de données
describe(data$Flan) # Une seule variable
```



Description: df [1 x 13]

| | vars
<dbl> | n
<dbl> | mean
<dbl> | sd
<dbl> | median
<dbl> | trimmed
<dbl> | mad
<dbl> | min
<dbl> | max
<dbl> |
|----|---------------|------------|---------------|-------------|-----------------|------------------|--------------|--------------|--------------|
| X1 | 1 | 28 | 4.36 | 1.81 | 4 | 4.38 | 2.97 | 1 | 7 |

1 row | 1-10 of 13 columns

PREMIERS PAS DANS R STUDIO

- **Obtenir nos premiers indicateurs sur les variables**
 - Identifier (et quantifier) des données manquantes
 - Résumé global d'un tableau : `summary(data)`
 - Résumé d'une seule variable **numérique** : `summary(data$Flan)`
 - **Autre fonction (très) utile** : `describe(data)`
 - Résumé d'une seule variable **non numérique** : `table(data$Specialite_sportive)`



```
```{r}
table(data$Specialite_sportive) # variable non numérique
```
```

| Athletisme | Natation | Volleyball |
|------------|----------|------------|
| 11 | 7 | 11 |

PREMIERS PAS DANS R STUDIO

■ Obtenir des indicateurs par groupe

Note moyenne des flans par spécialité sportive : `describeBy(data$Flan, data$Specialite_sportive)`

OU fonction dplyr: (permet de créer un tableau)

`data %>%`

`group_by(Specialite_sportive) %>%`

`summarise(mean = mean(Flan, na.rm = TRUE), sd = sd(Flan, na.rm = TRUE))` # il y a une valeur manquante pour le groupe natation, on inclue donc la fonction **na.rm = TRUE** pour forcer à exclure les valeurs manquantes dans le calcul de la moyenne et écart-type



```
```{r}
data %>%
 group_by(Specialite_sportive) %>%
 summarise(mean = mean(Flan, na.rm = TRUE), sd = sd(Flan, na.rm = TRUE))
```
```

A tibble: 3 × 3

| Specialite_sportive
<chr> | mean
<dbl> | sd
<dbl> |
|------------------------------|---------------|-------------|
| Athlétisme | 4.818182 | 1.990888 |
| Natation | 4.166667 | 2.041241 |
| Volleyball | 4.000000 | 1.549193 |

3 rows

PREMIERS PAS DANS R STUDIO

- **Obtenir des indicateurs par groupe**

```
data %>%  
  group_by(Specialite_sportive) %>%  
  summarise(mean = mean(Flan, na.rm = TRUE), sd = sd(Flan, na.rm = TRUE))
```

- **Pour inclure plusieurs pâtisseries**

```
data %>%  
  group_by(Specialite_sportive) %>%  
  summarise(  
    mean_Flan = mean(Flan, na.rm = TRUE),  
    sd_Flan = sd(Flan, na.rm = TRUE),  
    mean_Tiramisu = mean(Tiramisu, na.rm = TRUE),  
    sd_Tiramisu = sd(Tiramisu, na.rm = TRUE),  
    mean_Moelleux = mean(Moelleux, na.rm = TRUE),  
    sd_Moelleux = sd(Moelleux, na.rm = TRUE)  
  )
```



Exercice

Les athlètes français.es pour les JO 2024 ont noté leurs pâtisseries préférées de 0 à 10. Ces notes seront utilisées par les nutritionnistes de l'équipe de France pour sélectionner les pâtisseries à réaliser pour la fin des JO, en fonction de la spécialité sportive et du sexe.

1. Quelle est la meilleure pâtisserie en fonction de la **spécialité sportive** ?
2. Quelle est la meilleure pâtisserie en fonction du **sexe** ?
3. Quelle est la meilleure pâtisserie en fonction de la **spécialité sportive** ET du **sexe** ?

PREMIERS PAS DANS R STUDIO

- **Trier des données à partir d'une variable "catégorielle"**

```
Athletisme <- subset(data, Specialite_sportive == "Athletisme")
```

- **Trier des données à partir d'une variable "numérique"**

```
scores_hauts_tiramisu <- subset(data, Tiramisu > 5)
```

- **Conserver uniquement les données non manquantes**

```
scores_hauts_tiramisu_clean <- na.omit(scores_hauts_tiramisu)
```



Exercice

1. Quelle est la meilleure pâtisserie en fonction de la **spécialité sportive** ET du **sexe** pour les athlètes **≤ 24 ans** ? # le symbole « \leq » se note « $<=$ » dans R
2. Quelle est la meilleure pâtisserie en fonction de la **spécialité sportive** ET du **sexe** pour les athlètes de **> 24 ans** ?

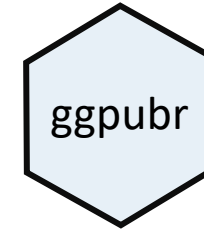
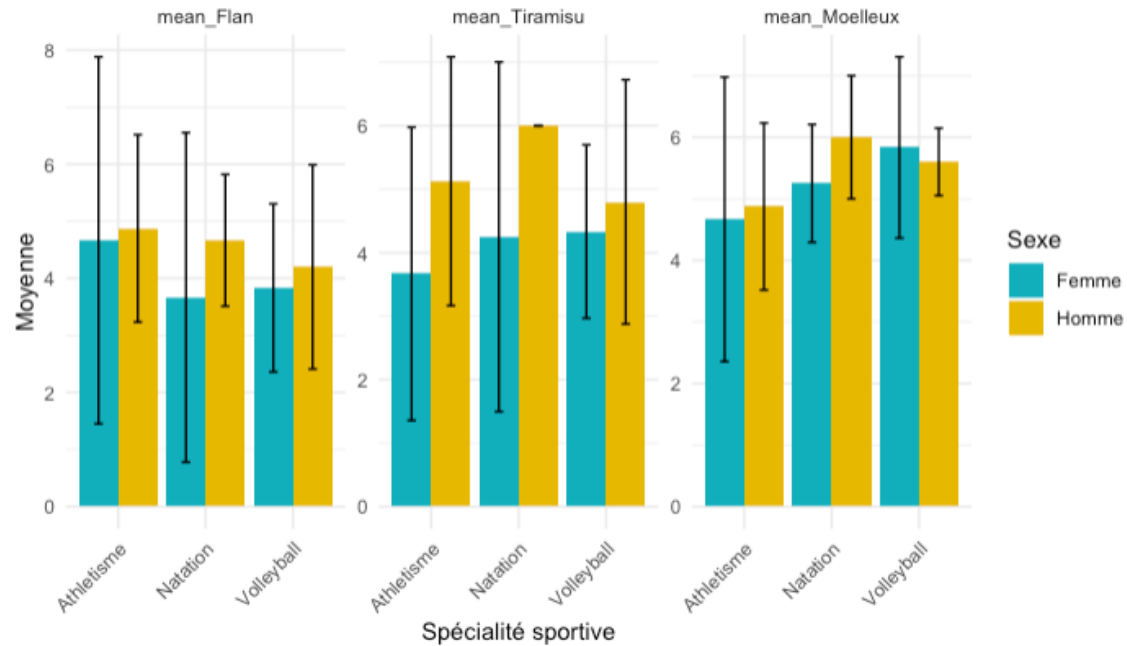
BONUS – VISUALISATION GRAPHIQUE



ggplot2

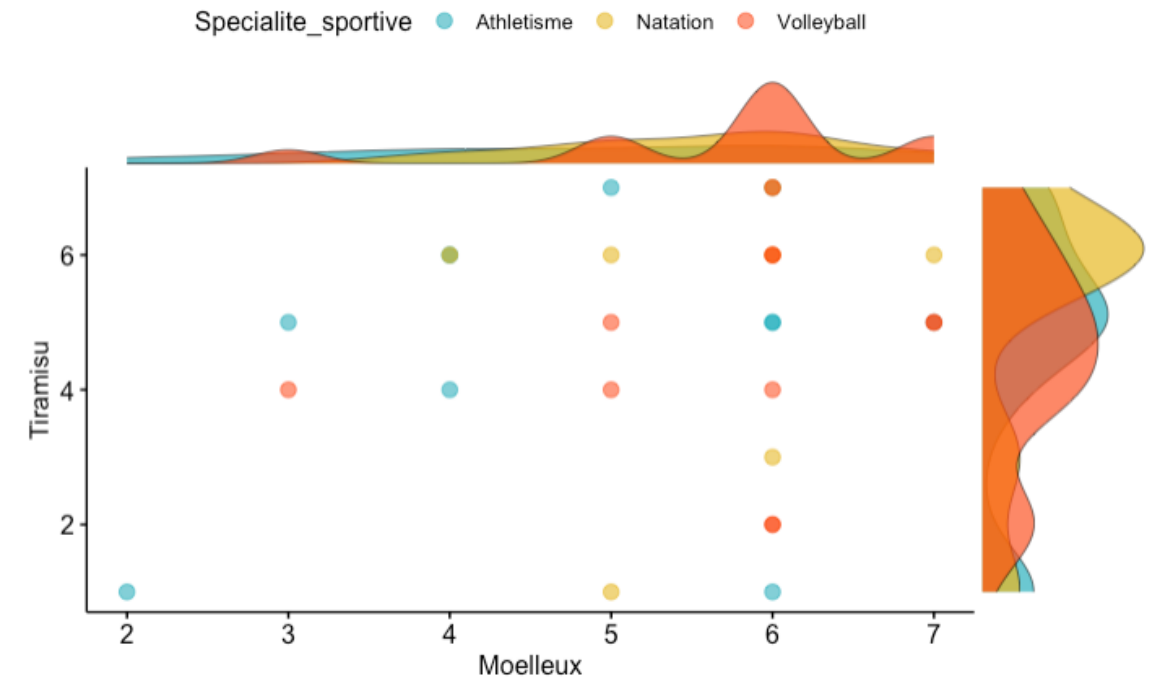
Visualisation graphique

Moyenne et écart-type des notes des pâtisseries par spécialité sportive et sexe



ggpubr

Distribution en fonction des groupes pour deux variables numériques



BONUS – STATISTIQUES INFÉRENTIELLES

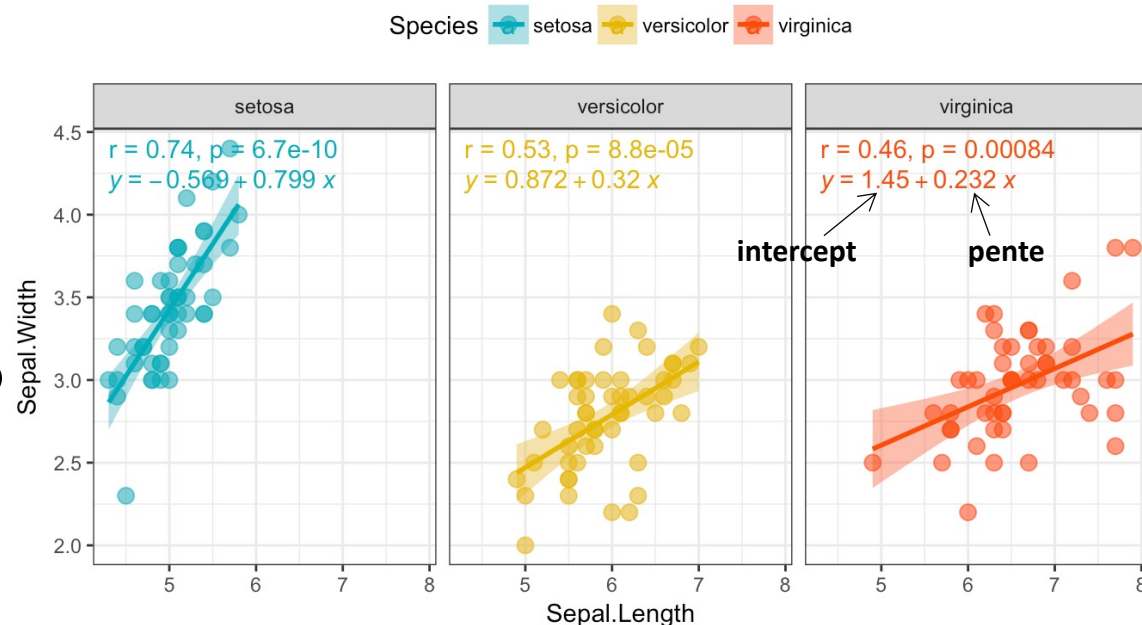


Statistiques descriptives suffisantes si on ne cherche pas à généraliser les résultats à la population générale. Sinon, nécessaire de passer par des **statistiques inférentielles**.

→ Faire des **déductions** ou des **inférences** à partir de données collectées à partir d'un échantillon, et de **généraliser** ces conclusions à une **population plus large**.

Exemple : association entre la longueur des pétales et leur largeur, en fonction de l'espèce de fleur

r = coefficient de corrélation
 p = significativité
 y = équation



Intercept (ordonnée à l'origine) : valeur de y lorsque $x = 0$

Pente : changement de y pour chaque unité de changement de x (e.g., pour chaque augmentation de 1 unité de x , y augmente de 0.232 unités)

POUR ALLER PLUS LOINS

Home > Cheat sheets > Data Science

Getting Started with R Cheat Sheet

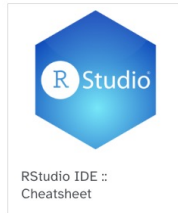
This cheat sheet will cover an overview of getting started with R. Use it as a handy, high-level reference for a quick start with R. For more detailed R Cheat Sheets, follow the highlighted cheat sheets below.

Jun 2022 · 9 min read

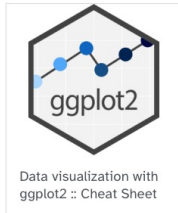
<https://www.datacamp.com/cheat-sheet/getting-started-r>

Posit Cheatsheets

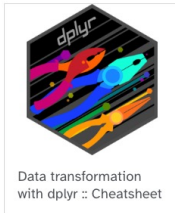
HTML versions of our popular cheatsheets. PDF versions are available to download on each cheatsheet page. There are also non-English translations available for many cheatsheets, contributed by the community.



RStudio IDE ::
Cheatsheet



Data visualization with
ggplot2 :: Cheat Sheet

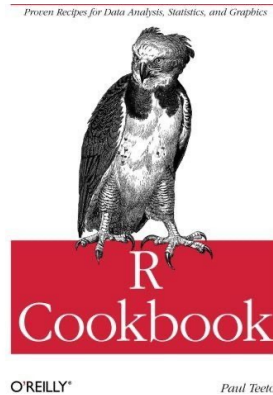


Data transformation
with dplyr :: Cheatsheet

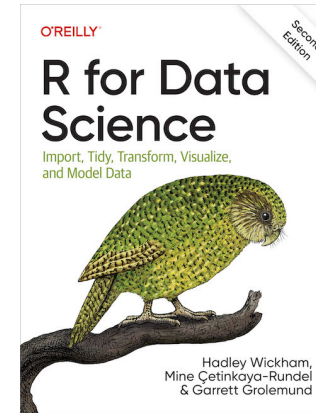
<https://rstudio.github.io/cheatsheets/>



<https://www.datacamp.com/cheat-sheet/chatgpt-cheat-sheet-data-science>



<http://www.cookbook-r.com/>



<https://r4ds.hadley.nz/>



<https://larmarange.github.io/analyse-R/>

Mathématiques et statistiques **Outils pour la recherche**

Introduction à la statistique avec R

Réf. 71007

Ce cours permet d'apprendre la statistique à l'aide du logiciel libre R. Le recours aux mathématiques est minimal. L'objectif est de savoir analyser des données en comprenant ce que l'on fait.

📅 Durée : 5 semaines ⌚ Effort : 25 heures 🔄 Rythme: ~5 heures/semaine

🌐 Langues: Français

<https://www.fun-mooc.fr/fr/cours/introduction-a-la-statistique-avec-r/>