

(Re)découverte de R... en 1h45

Ecole de Bioinformatique AVIESAN-IFB – Roscoff – Novembre 2023

Charlotte Berthelier charlotte.berthelier@sb-roscoff.fr

Elise Jacquemet elise.jacquemet@pasteur.fr

Slides d'Hugo Varet – hugo.varet@pasteur.fr

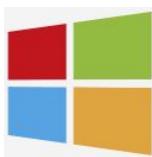


R en quelques mots

Langage de programmation qui permet de :

1. manipuler des données : importer, transformer, exporter
2. faire des analyses statistiques plus ou moins complexes : description, exploration, modélisation...
3. créer des (jolies) figures

Disponible sur [RCRAN](#)



Historique :

- 1993 : début du projet R
- 2000 : sortie de R 1.0.0
- 2023 : R 4.3.2

Avantages et inconvénients

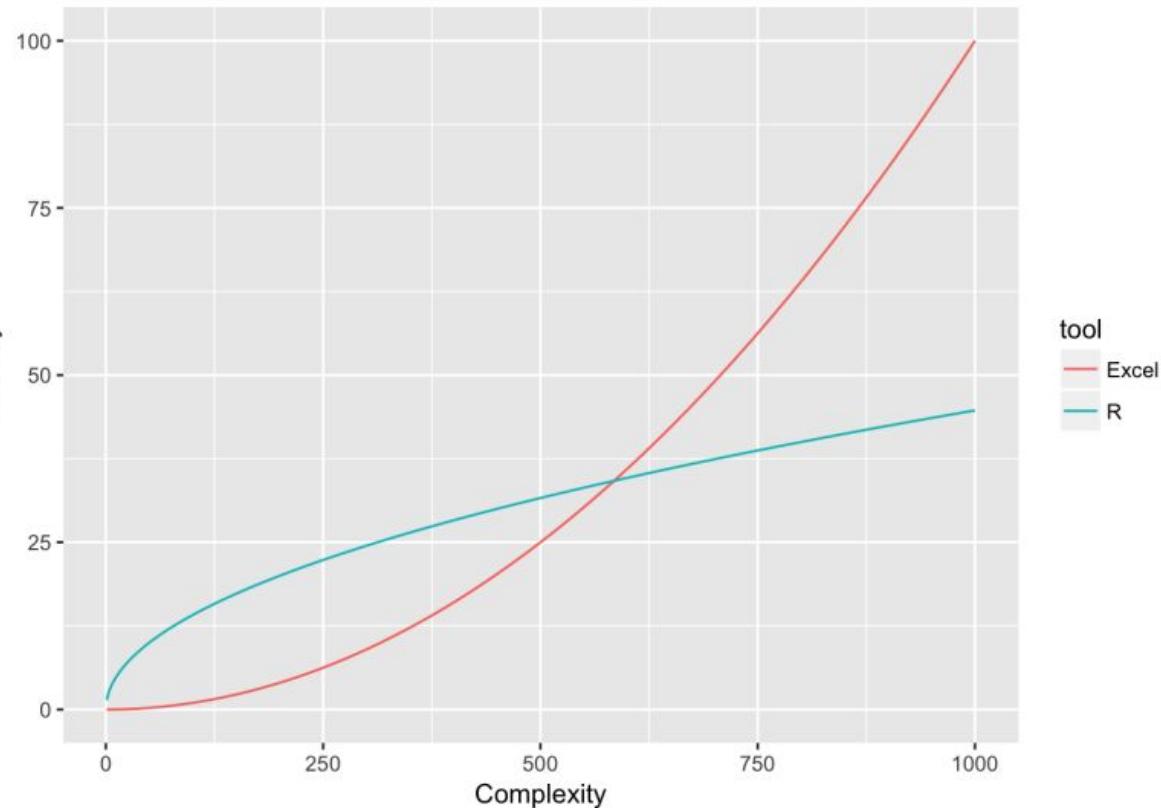
Avantages :

- Souplesse d'utilisation pour réaliser des analyses statistiques
- Libre et gratuit, même s'il existe maintenant des versions payantes de RStudio (shiny et/ou server)
- Reproductibilité des analyses en écrivant/sauvegardant les commandes R dans des scripts (Rmd)
- Large communauté d'utilisateurs/aide en ligne
- Grand nombre de packages spécifiques

Inconvénients :

R vs Excel

Difficulty vs. Complexity



Covid : le Royaume-Uni passe à côté de milliers de cas à cause... d'un fichier Excel arrivé à saturation

Les autorités sanitaires britanniques ont reconnu que près de 16.000 cas de coronavirus en Angleterre sont passés sous le radar au cours de la semaine écoulée à cause d'un problème dans le chargement des données.

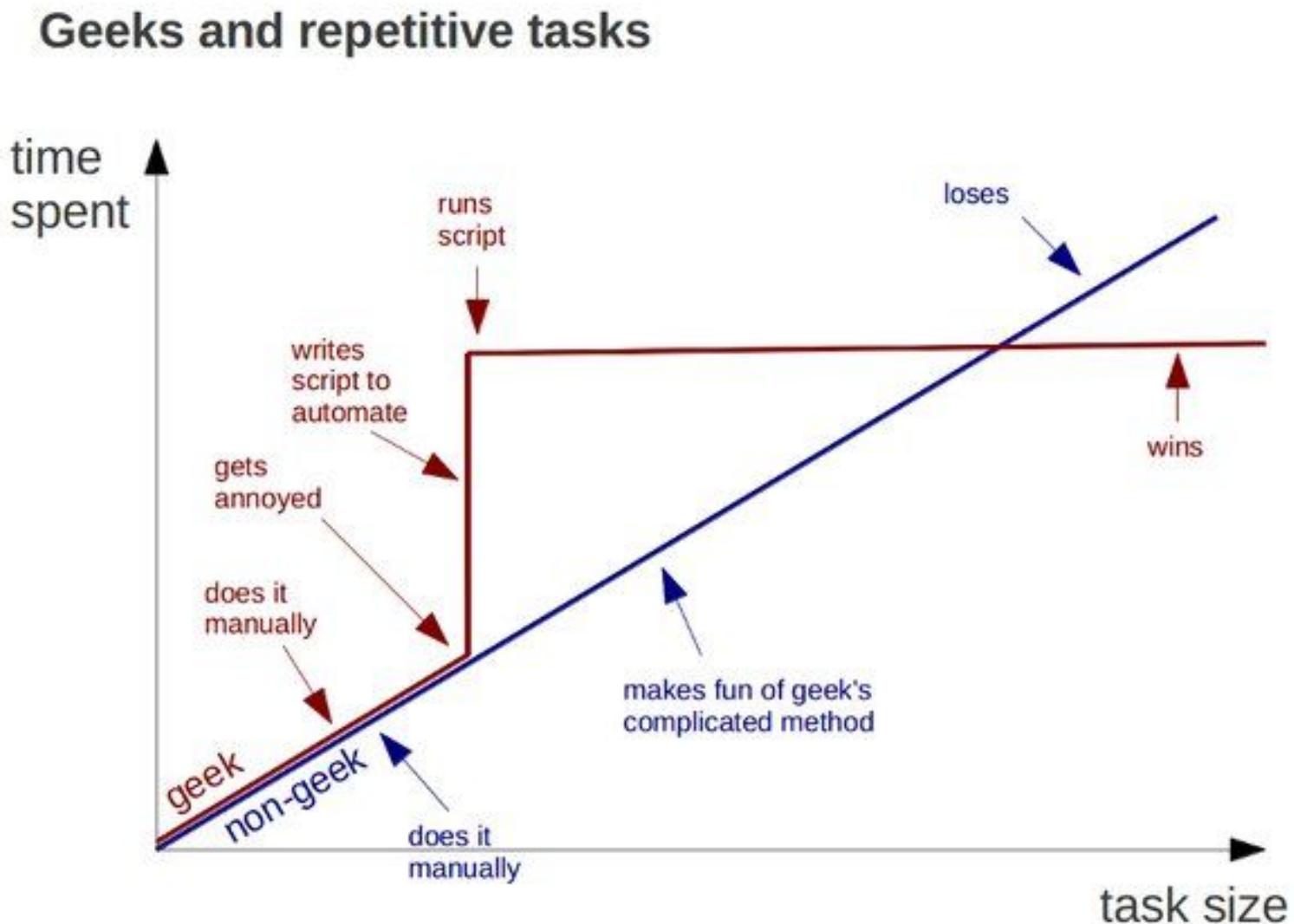
Lire plus tard Europe Partager Commenter



[Alexandre Counis, Les Echos, 5 oct. 2020](#)

Source: R-bloggers

Geeks and repetitive tasks



R sait tout faire

Lire un tableau de données	read.table()
Fusionner deux tableaux	merge()
Filtrer des lignes	data [data\$x > 10]
Sélectionner des colonnes	data [, c("x", "y")]
Rechercher une chaîne de caractères	grep()
Calculer une moyenne	mean()
Réaliser une ACP	prcomp()
Additionner deux matrices	mat1 + mat2
Exporter un tableau de données	write.table()
Calculer une variance	var()
Régression linéaire	lm()
Tracer une courbe	plot()
Tester une hypothèse	t.test()
Dessiner un histogramme	hist()
Convertir des données	as.matrix()

Modes d'utilisation (liste non exhaustive)



Localemement via le terminal



Localemement via RStudio (utilisation classique)



Sur un serveur via le terminal et une connexion ssh



Sur un serveur via un navigateur web pour accéder à RStudio server

Ouverture ou connexion à RStudio

2 alternatives :

1. Vous connecter via **Jupyter lab de l'IFB**

<https://jupyterhub.cluster.france-bioinformatique.fr>

puis cliquer sur l'icône RStudio



2. Vous connecter au **serveur Web RStudio de l'IFB**

<https://rstudio.cluster.france-bioinformatique.fr>

puis vous identifier

Sign in to RStudio

Username:

Password:

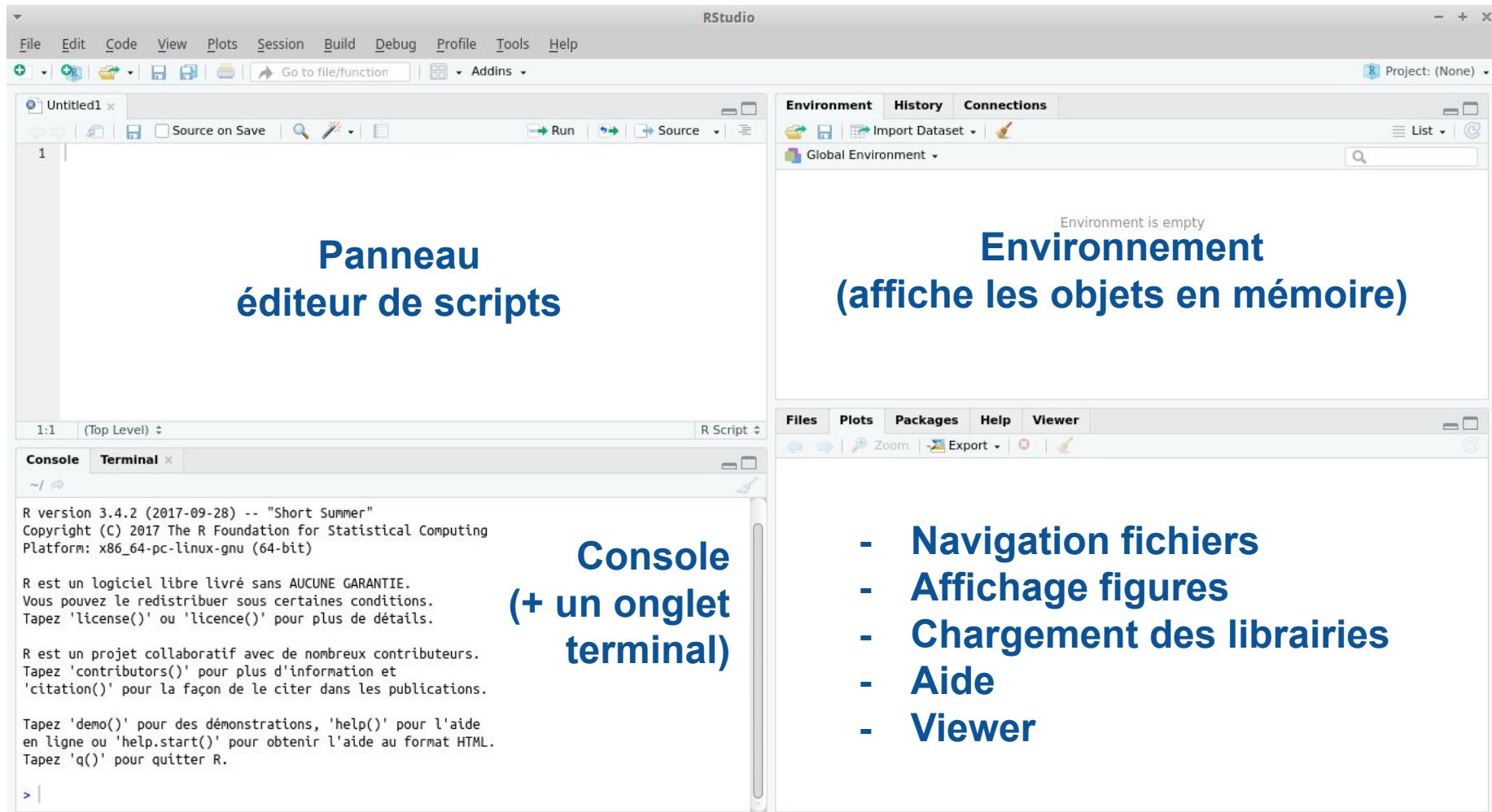
Stay signed in

Sign In



RStudio

- Disponible depuis 2011
- Logiciel facilitant l'utilisation de R via 4 panneaux
- Chaque panneau présente plusieurs onglets (fonctionnalités complémentaires)



R sait tout faire : il compte

Tapez les commandes suivantes dans le panneau Console de RStudio

2 + 3

4 * 5

6 / 4

1:10

8:-9

1,2

1.2

Notion de variable/objet

```
a <- 2  
print(a)  
a
```

Créer une variable nommée a et lui assigner une valeur
Afficher la valeur de la variable a
Même résultat: si on écrit le nom de variable, R l'imprime

Environment		History	Connections
			Import Dataset ▾
	Global Environment ▾		
Values			
a	2		
AplusB	5		
b	3		

```
b <- 3  
a_plus_b <- a + b  
print(a_plus_b)
```

Assigner une valeur à une seconde variable
Effectuer un calcul avec 2 variables
Afficher le contenu de la variable a_plus_b

```
a <- 7  
print(a_plus_b)
```

Changer la valeur de a
Note: le contenu de a_plus_b n'est pas modifié

Environment		History	Connections
			Import Dataset ▾
	Global Environment ▾		
Values			
a	7		
AplusB	5		
b	3		

```
a_plus_b <- a + b  
print(a_plus_b)
```

On recalcule a_plus_b
La nouvelle valeur tient compte de la modification de a

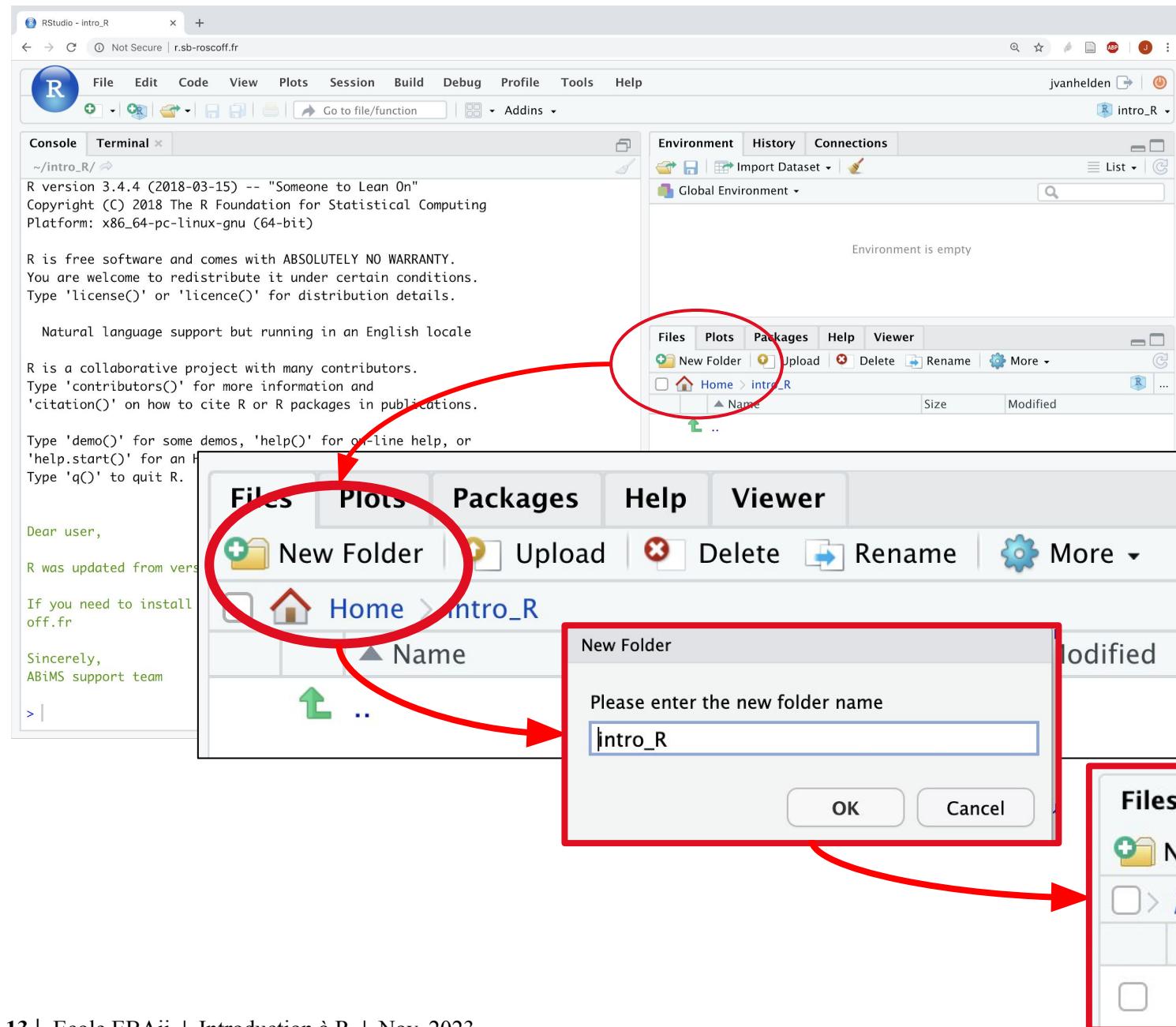
```
vec1 <- c(1,10)  
vec2 <- 1:10  
vec2 + a  
vec3 <- c("riri", "fifi", "loulou")  
vec2 / 2  
vec3 / 2
```

Créer un vecteur
Créer un vecteur contenant une séquence d'entiers de 1 à 10
Somme d'un vecteur et d'un nombre
Vecteur de chaînes de caractères
Diviser un vecteur de nombres par un nombre
Diviser des chaînes de caractères par un nombre

Noms de variables interdits: TRUE, FALSE, T, F, c, t, pi, data, LETTERS, letters, ...

Cas pratique : manipulation de matrices de données

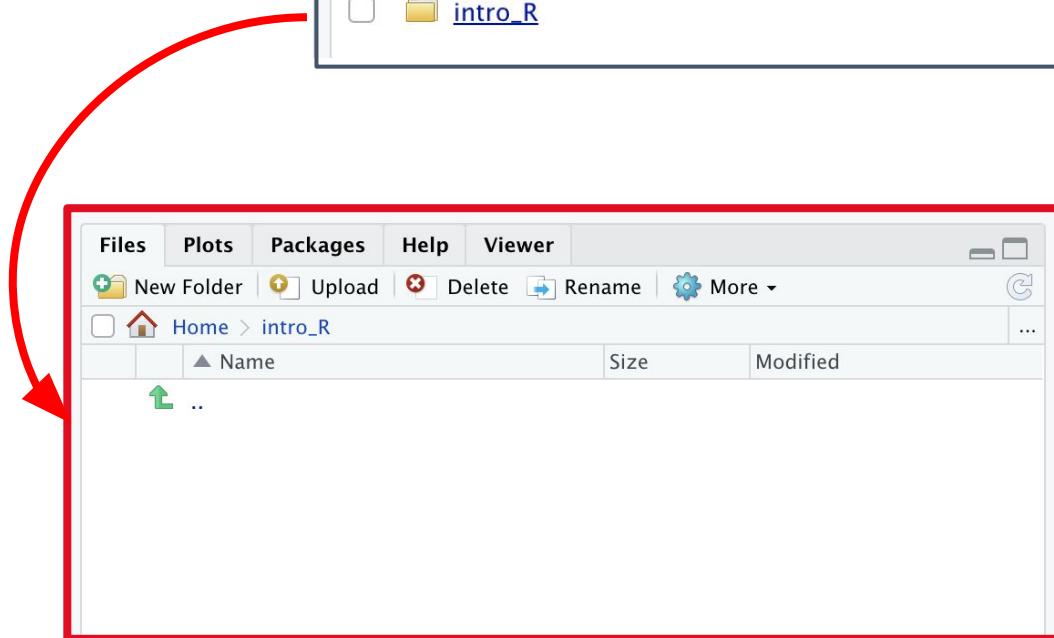
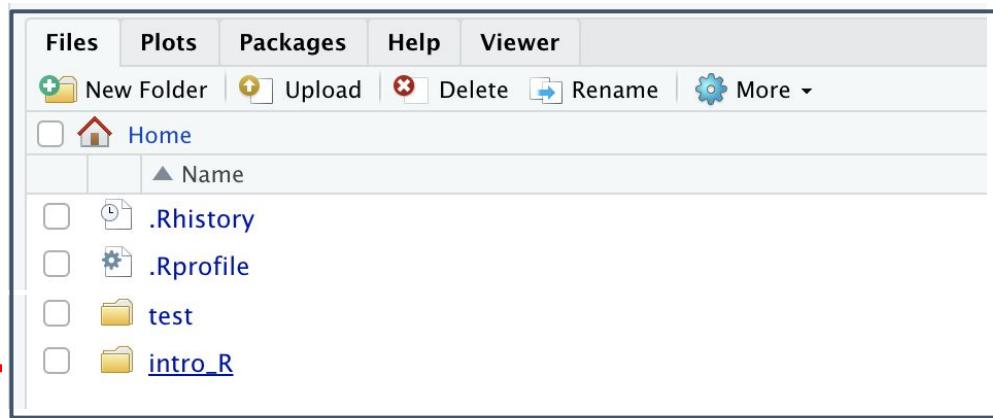
Création d'un dossier `intro_R` pour vos résultats de ce TP



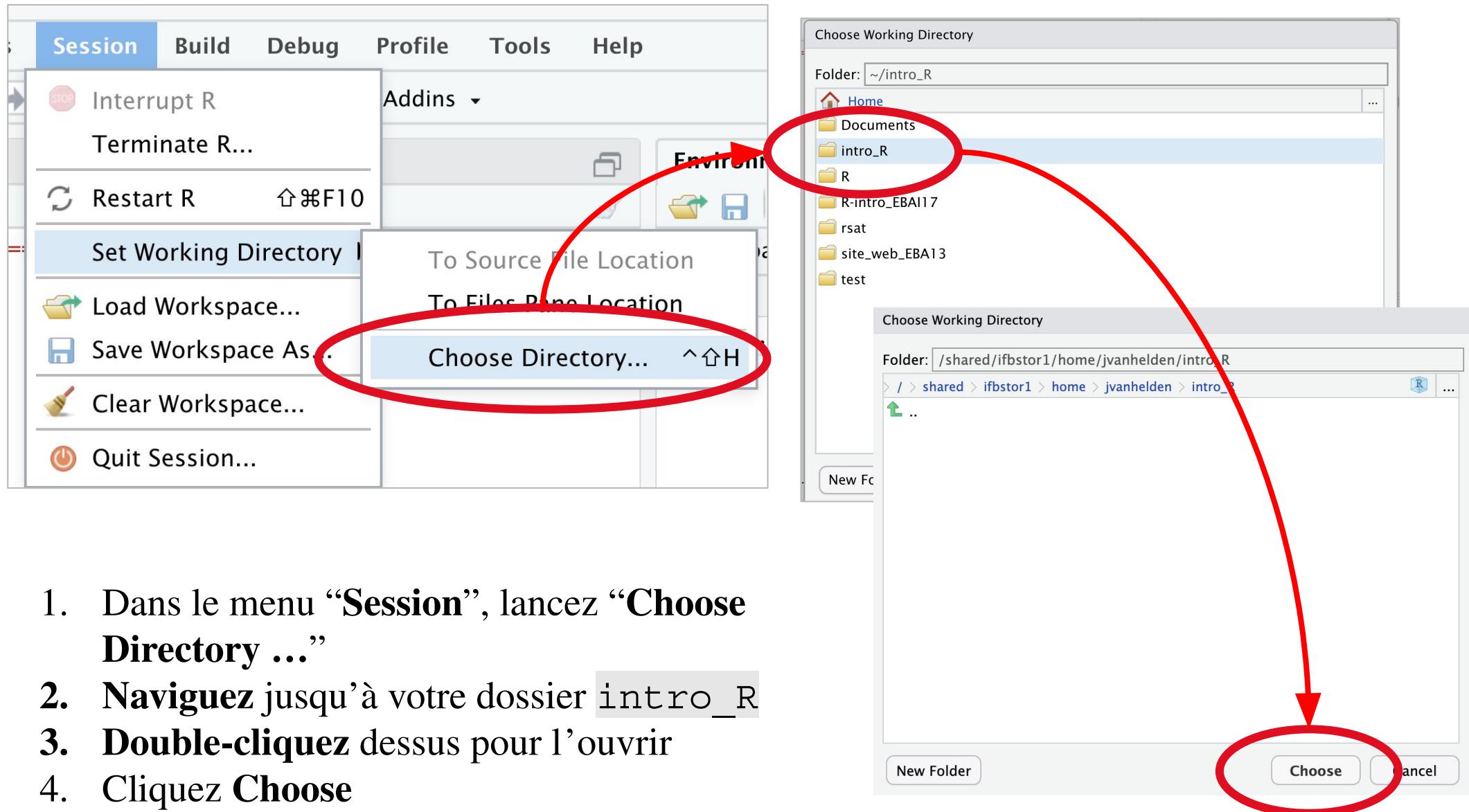
Déplacement dans le dossier “intro_R”

Double-cliquez sur le dossier “intro_R”, pour vous y déplacer.

Puisque vous venez de créer le dossier il est vide (image du bas).



Définissez votre dossier espace de travail (working directory)



Téléchargez un fichier depuis RStudio

A partir d'un navigateur Web, téléchargez et enregistrez **sur votre ordi** les fichiers de données :

- fruits.tsv / .csv / .xlsx : tableau de données de fruits disponible en 3 différents formats

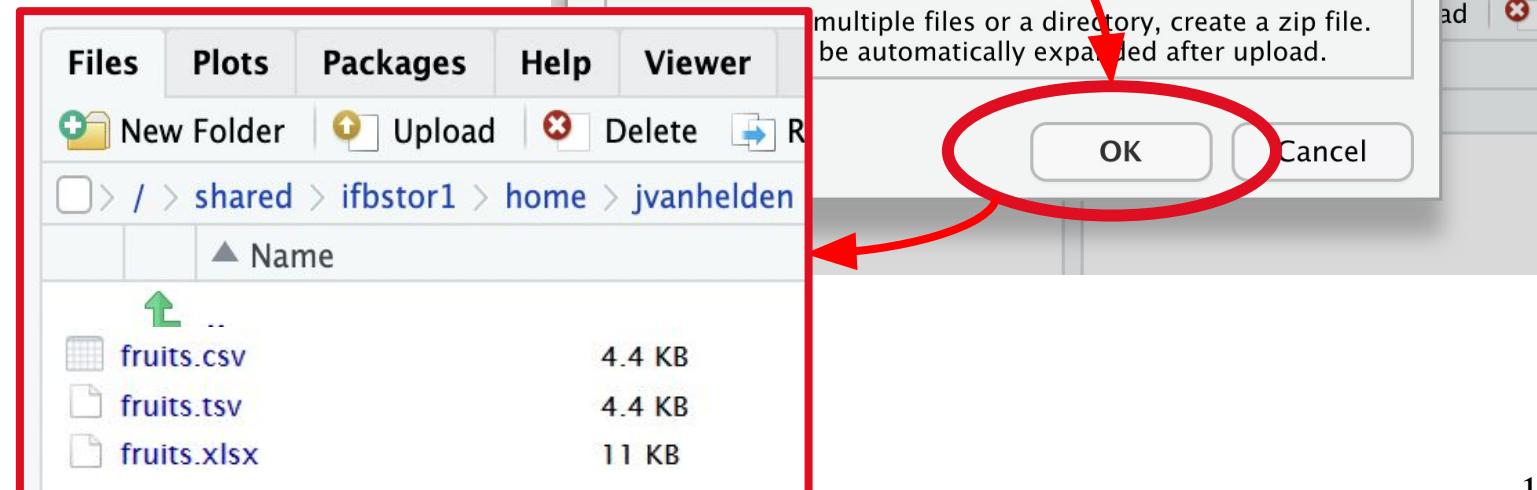
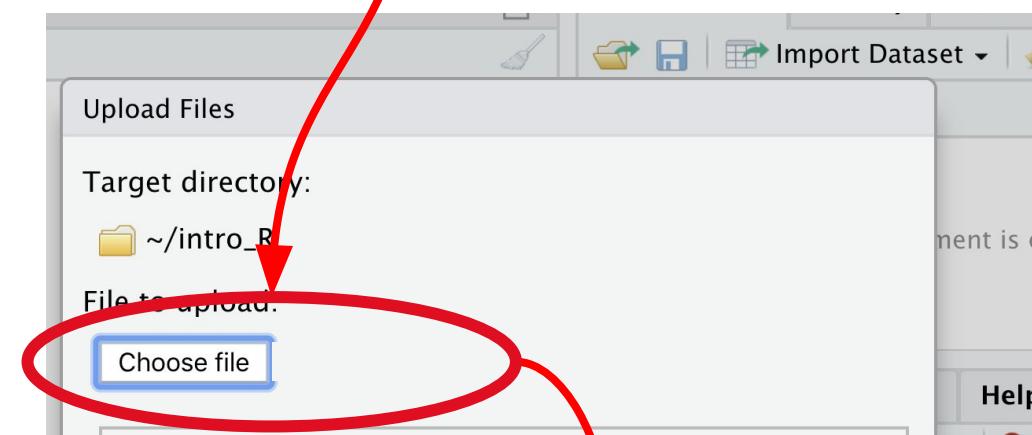
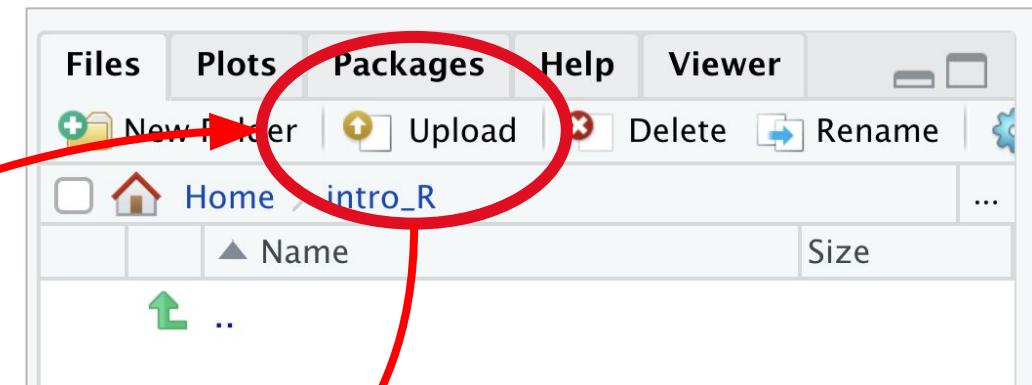
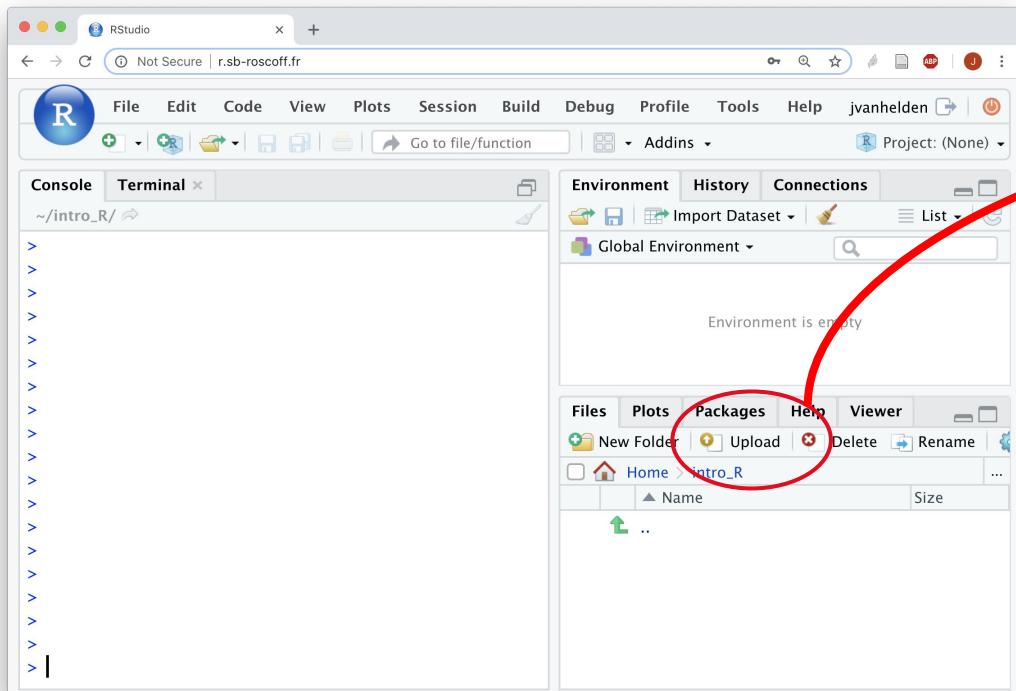
Fichiers disponibles sur le Moodle :

<https://moodle.france-bioinformatique.fr/course/view.php?id=22>

Attention: veillez à sauvegarder les fichiers

- sous leur nom original,
- avec les extensions respectives (certains navigateurs omettent l'extension, ce qui poserait problème pour la suite du TP)

Téléversement (“upload”) des données

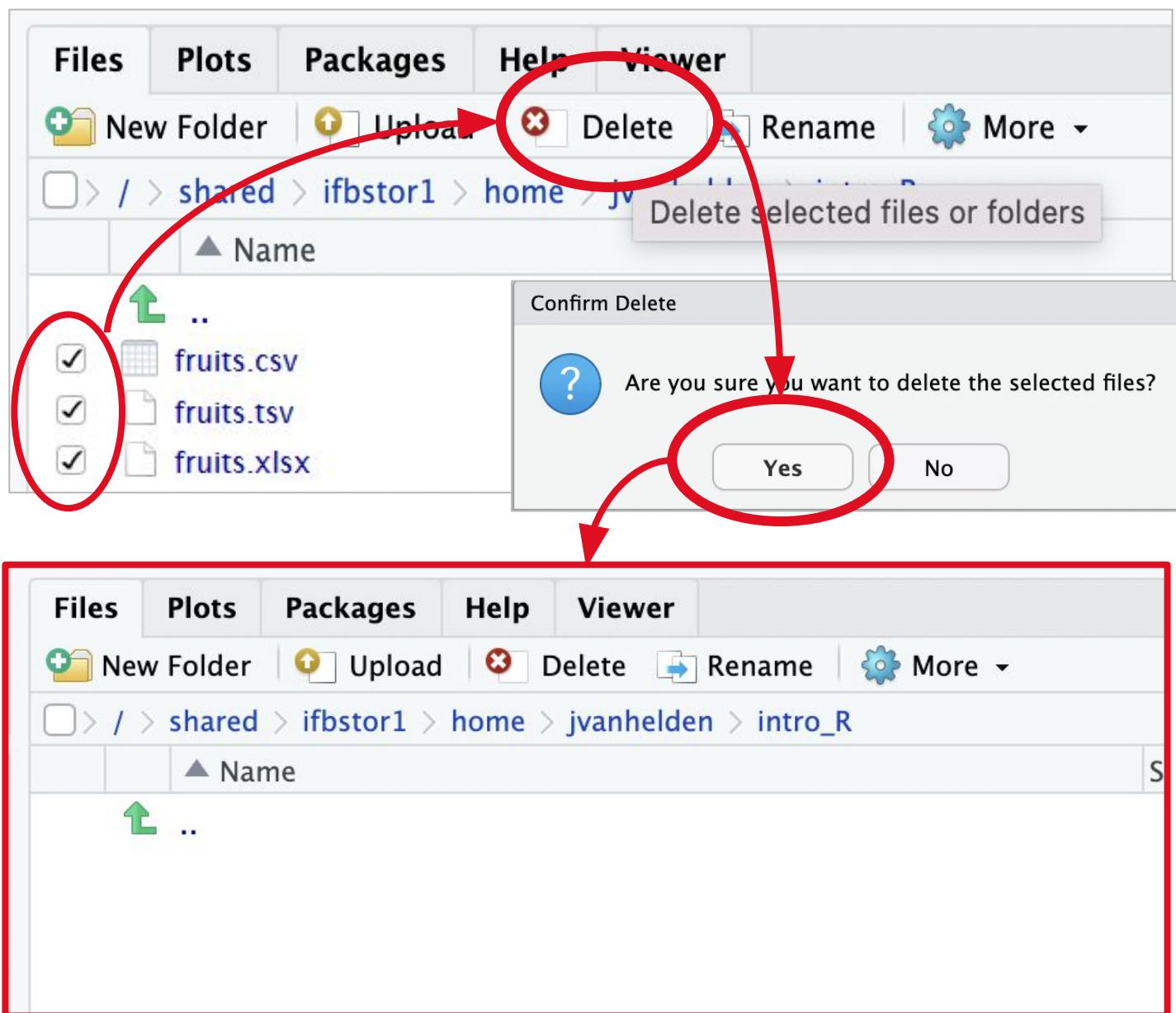


Au moyen du bouton “Upload”, téléversez les fichiers depuis votre ordinateur vers votre compte sur le serveur.

On efface tout et on recommence

1. Sélectionnez les deux fichiers
2. Effacez-les sans pitié

(nous allons vous montrer deux autres façons de les téléverser)



The “R geek” way (V2, directement depuis Rstudio)



Revenir à la maison (note: R interprète le caractère “~” comme le “HOME” de Linux; et cela marche aussi pour Windows!)
`setwd("~/")`

Définir une variable qui indique le chemin du dossier de travail (working directory).

`my_work_dir <- "~/intro_R"`

S'il n'existe pas encore, créer le dossier de travail. (Commande Unix équivalente: "mkdir -p ~/intro_R")

`dir.create(my_work_dir, recursive = TRUE, showWarnings = FALSE)`

Où suis-je ? (Commande Unix équivalente: "pwd")

`getwd()`

Aller dans ce dossier de travail (Commande Unix équivalente: "cd ~/intro_R")

`setwd(my_work_dir)`

Et maintenant, où suis-je ?

`getwd()`

Qu'y a-t-il par ici ? (Commande Unix équivalente: "ls")

`list.files()`

`dir() ## Un autre nom pour la même commande`

Télécharger un fichier : the “geek” way (V2)

Nous avons montré ci-dessus comment télécharger des fichiers en utilisant l’interface graphique de RStudio. Alternativement, on peut télécharger des fichiers au moyen de la commande R **download.file**.

Les commandes suivantes permettent de télécharger les fichiers utilisés pour les exercices.

```
download.file(url =
  "https://github.com/IFB-ElixirFr/EBAII/blob/master/2023/ebaiin1/intro_R/data_fruits/fruits.csv",
  destfile = "fruits.csv")
```

```
download.file(url =
  "https://github.com/IFB-ElixirFr/EBAII/blob/master/2023/ebaiin1/intro_R/data_fruits/fruits.tsv",
  destfile = "fruits.tsv")
```

```
download.file(url =
  "https://github.com/IFB-ElixirFr/EBAII/blob/master/2023/ebaiin1/intro_R/data_fruits/fruits.xlsx",
  destfile = "fruits.xlsx")
```

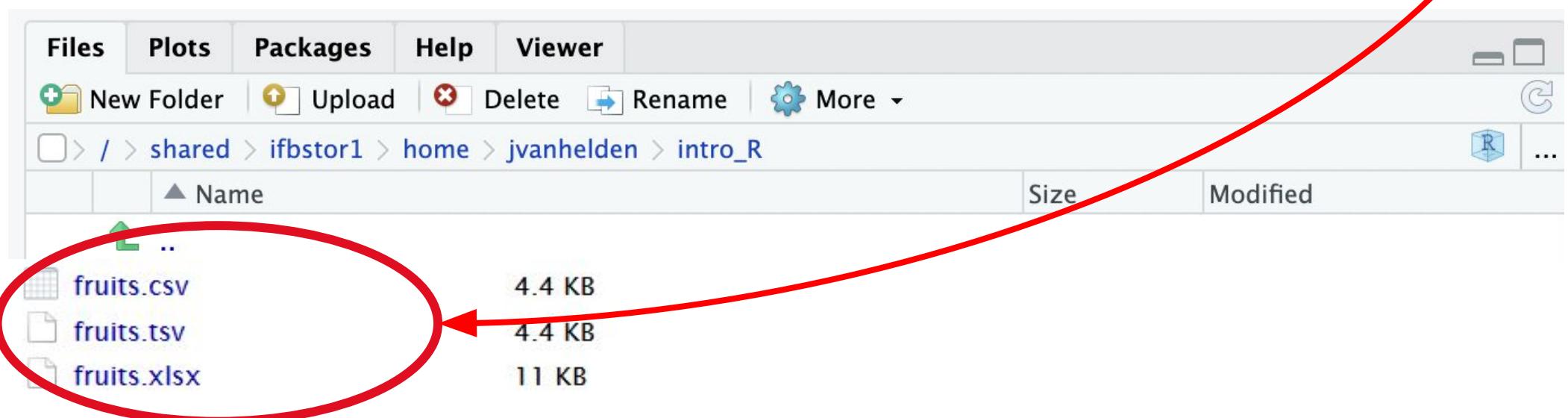
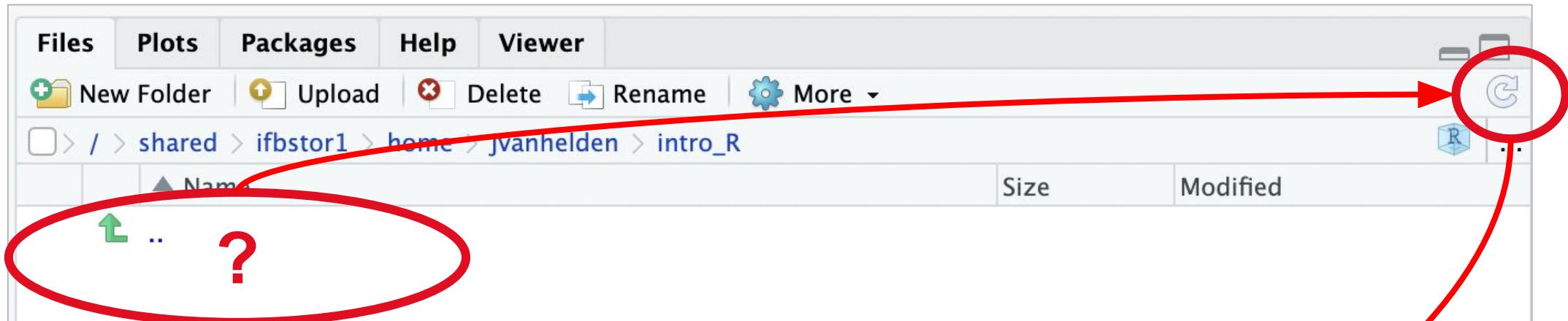
Note : équivalent de la commande “wget” sous Unix.

Qu'y a-t-il par ici ? (Commande Unix équivalente: "ls") **list.files()**

Actualisation du dossier

Dans certains cas, il faut actualiser le contenu du dossier pour pouvoir voir le nouveau sous-dossier.

Vérifiez ensuite si `intro_R` apparaît bien dans le contenu de votre dossier principal.



Chargement des données (dans la mémoire de R)

Charger le contenu du fichier "expression.txt" dans une variable nommée "exprs".

```
fruits <- read.table(file = "fruits.tsv", header = TRUE, sep = "\t")
```

Accéder à l'aide d'une fonction

```
help(read.table)
```

Notation alternative

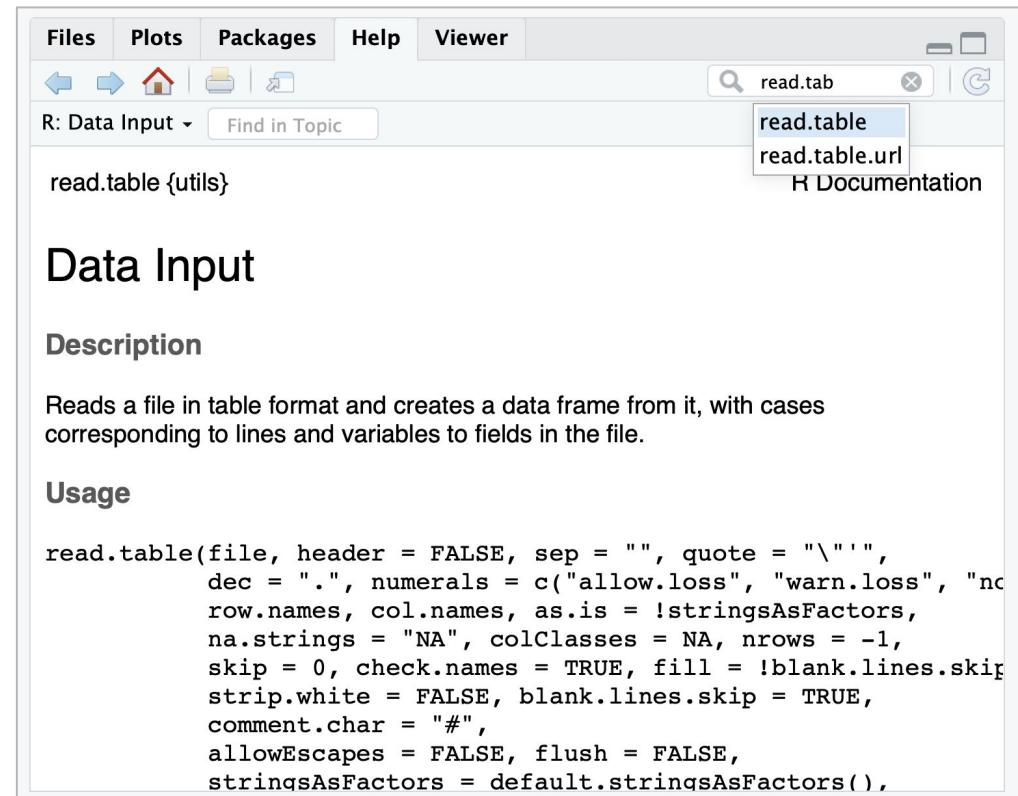
```
?read.table
```

Approche pratique :

1. demande à Google

“Comment lire une table en R ?”

2. adapte l'exemple



Recherche interactive sous RStudio

- Sélectionner l'onglet “Help” du panneau inférieur droit.
- Taper “read.table” dans la boîte de recherche.

Affichage de l'objet “fruits”

Imprimer toutes les valeurs.

`print(fruits)`

Affichage des premières lignes de l'objet
`head(fruits)`

Affichage des dernières lignes de l'objet
`tail(fruits)`

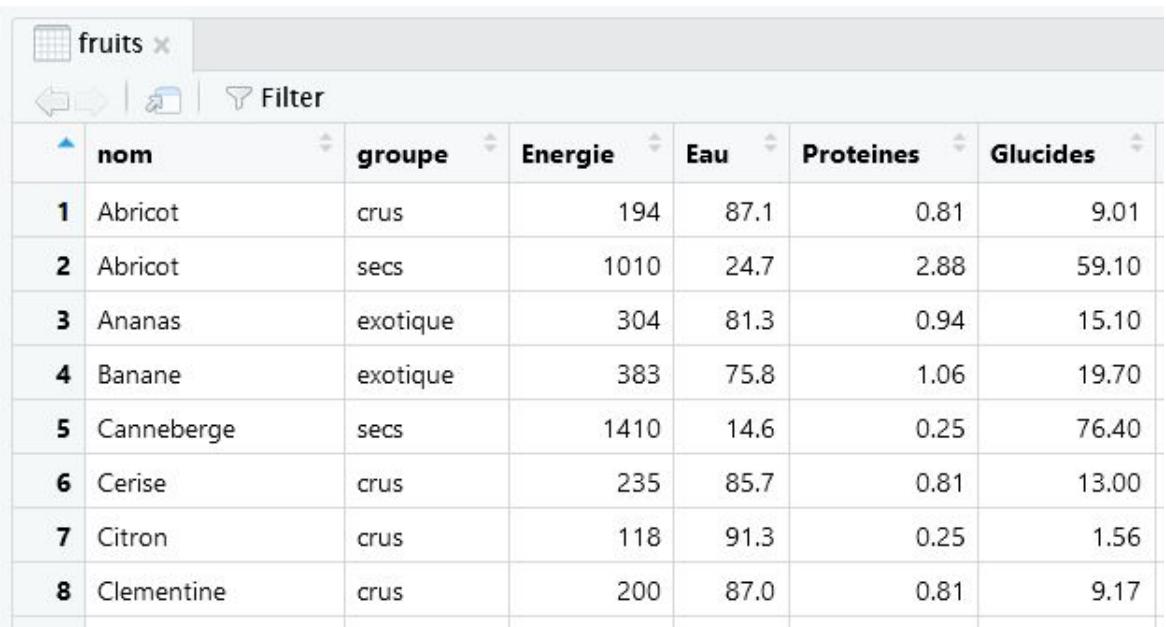
Un peu plus de lignes

`head(fruits, n = 20)`

Explorer le tableau dans un panneau de visualisation

`View(fruits)`

Note: vous pouvez cliquer sur une en-tête de colonne pour trier les données



	nom	groupe	Energie	Eau	Proteines	Glucides
1	Abricot	crus	194	87.1	0.81	9.01
2	Abricot	secs	1010	24.7	2.88	59.10
3	Ananas	exotique	304	81.3	0.94	15.10
4	Banane	exotique	383	75.8	1.06	19.70
5	Canneberge	secs	1410	14.6	0.25	76.40
6	Cerise	crus	235	85.7	0.81	13.00
7	Citron	crus	118	91.3	0.25	1.56
8	Clementine	crus	200	87.0	0.81	9.17

`> head(fruits, n = 10)`

	nom	groupe	Energie	Eau	Proteines
1	Abricot	crus	194	87.1	0.81
2	Abricot	secs	1010	24.7	2.88
3	Ananas	exotique	304	81.3	0.94
4	Banane	exotique	383	75.8	1.06
5	Canneberge	secs	1410	14.6	0.25
6	Cerise	crus	235	85.7	0.81
7	Citron	crus	118	91.3	0.25
8	Clementine	crus	200	87.0	0.81
9	CompoteMultiFruits	compote	279	82.9	0.25
10	CompotePomme	compote	432	72.9	0.23

Caractéristiques d'un tableau de données

Dimensions :

`ncol(fruits)`

`nrow(fruits)`

`dim(fruits)`

Nombre de colonnes

Nombre de lignes

Dimensions

Noms des colonnes et des lignes :

`colnames(fruits)`

Noms des colonnes

`names(fruits)`

idem

`rownames(fruits)`

noms des lignes

Résumé rapide des données par colonne :

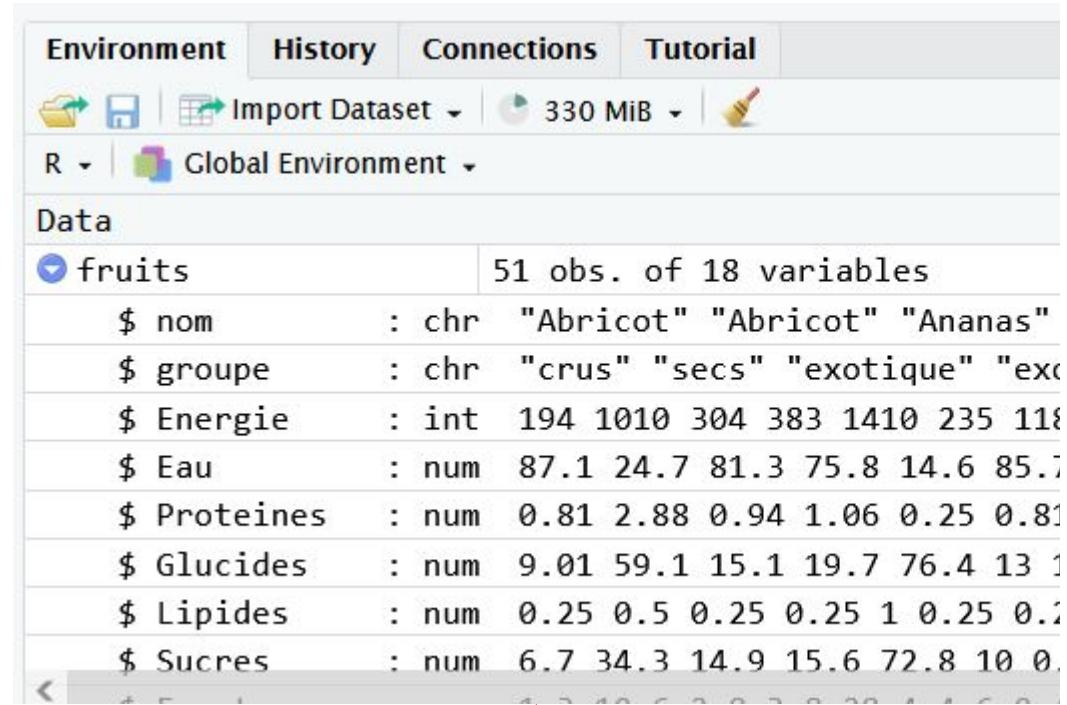
`summary(fruits)`

Statistiques par colonne

`str(fruits)`

Structure de la variable

-> Même résultats que dans le panneau “Environment”



fruits		51 obs. of 18 variables
\$ nom	:	chr "Abricot" "Abricot" "Ananas"
\$ groupe	:	chr "crus" "secs" "exotique" "exc"
\$ Energie	:	int 194 1010 304 383 1410 235 118
\$ Eau	:	num 87.1 24.7 81.3 75.8 14.6 85.7
\$ Proteines	:	num 0.81 2.88 0.94 1.06 0.25 0.81
\$ Glucides	:	num 9.01 59.1 15.1 19.7 76.4 13 1
\$ Lipides	:	num 0.25 0.5 0.25 0.25 1 0.25 0.2
\$ Sucres	:	num 6.7 34.3 14.9 15.6 72.8 10 0.

Affichage de colonnes d'un tableau

Afficher les noms des colonnes

colnames(fruits)

Valeurs stockées dans la colonne nommée "nom"

fruits\$nom

Notation alternative

fruits[, "nom"]

Sélection de plusieurs colonnes.

fruits[, c("nom", "groupe")]

Sélection de colonnes par leur indice

fruits[, 2]

fruits[, c(3, 2)]

fruits[, c(1:5)]

Sélection ou suppression de colonnes d'un tableau -> subset()

R: Subsetting data.tables ▾

Find in Topic

subset.data.table {data.table}

R Documentation

Subsetting data.tables

Description

Returns subsets of a data.table.

Usage

```
## S3 method for class 'data.table'  
subset(x, subset, select, ...)
```

Arguments

x data.table to subset.

subset logical expression indicating elements or rows to keep

select expression indicating columns to select from data.table

... further arguments to be passed to or from other methods

Sélection ou suppression de colonnes d'un tableau

Sélectionner la première colonne de fruits :

fruitsbis <- subset(fruits, select = nom)

fruitsbis <- subset(fruits, select = 1)

Supprimer une ou plusieurs colonnes de fruits avec leur noms :

fruitsbis <- subset(fruits, select = -nom)

fruitsbis <- subset(fruits, select = -c(nom, Glucides, Sucres))

Supprimer la première colonne de fruits avec son index :

fruitsbis <- subset(fruits, select = -1)

Supprimer plusieurs colonnes de fruits avec leur index :

fruitsbis <- subset(fruits, select = -c(2, 3))

Figures en R

Chargement des données (poids de poussins suivant 4 régimes différents au cours du temps):

```
data("ChickWeight")
```

```
summary(ChickWeight)
```

Sélection des données pour T=21 :

```
dta <- ChickWeight[ChickWeight$Time == 21, ]
```

Affichage du design :

```
table(dta$Diet)
```

Visualisation des données avec **boxplot()** :

```
boxplot(weight ~ Diet, data=dta)
```

```
stripchart(weight ~ Diet, data=dta, add=TRUE, vertical=TRUE)
```

Visualisation avec **ggplot()** :

```
ggplot(data=dta, mapping=aes(x=Diet, y=weight, fill=Diet)) +
```

```
  geom_boxplot() +
```

```
  geom_jitter(width = 0.1, height = 0)
```

Take home messages

- Tout est faisable avec R
- **Définir et comprendre l'opération mathématique/statistique** avant de chercher la fonction R correspondante
- R est un langage :
 - plusieurs types et structures de données (out of scope)
 - énormément de commandes à découvrir (out of scope)
 - Google est votre ami
- Une infinité de :
 - ressources en ligne
 - tutoriels pour des analyses spécifiques (e.g. DESeq2 pour le RNA-Seq)
- Bonnes pratiques : <https://style.tidyverse.org/syntax.html>

Serveur RStudio

<https://rstudio.cluster.france-bioinformatique.fr/>



Jupyter lab (inclut un serveur RStudio et plein d'autres choses)

<http://jupyterhub.cluster.france-bioinformatique.fr/>



Une question ? Un besoin ? Un problème ? **Contactez la communauté IFB**

<https://community.france-bioinformatique.fr/>



Ressources

The collage includes:

- Base R Cheat Sheet**: A summary of basic R functions like mean, str, and install.packages.
- Advanced R Cheat Sheet**: A detailed guide on environments, search paths, and function environments.
- RStudio IDE :: CHEAT SHEET**: A comprehensive guide to RStudio's interface, including sections on documents and apps, write code, R support, and pro features.
- R Markdown :: CHEAT SHEET**: A guide to R Markdown, covering what it is, how to use it in RStudio, and its workflow.
- R Studio**: The official RStudio logo.

R

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

RStudio

<https://rstudio.com/>



<https://www.r-bloggers.com/>



<https://thinkr.fr/>

Rstudio Cheatsheets (un tas de thèmes):

<https://rstudio.com/resources/cheatsheets/>