

# Introduction à R et Rstudio

---

# À propos de R

---

R est un langage orienté vers le traitement et l'analyse quantitative de données, dérivé du langage S. Il est développé depuis les années 90 par un groupe de volontaires de différents pays et par une large communauté d'utilisateurs et utilisatrices. C'est un logiciel libre, publié sous [licence GNU GPL](#).

L'utilisation de R présente plusieurs avantages :

- c'est un logiciel multiplateforme, qui fonctionne aussi bien sur des systèmes Linux, Mac OS X ou Windows.
- c'est un logiciel libre, développé par ses utilisateurs et utilisatrices, diffusable et modifiable par tout un chacun.
- c'est un logiciel gratuit.
- c'est un logiciel puissant, dont les fonctionnalités de base peuvent être étendues à l'aide d'extensions développées par la communauté. Il en existe plusieurs milliers.
- c'est un logiciel avec d'excellentes capacités graphiques.

# À propos de R

Comme rien n'est parfait, on peut également trouver quelques inconvénients :

- le logiciel, la documentation de référence et les principales ressources sont en anglais. Il est toutefois parfaitement possible d'utiliser R sans spécialement maîtriser cette langue et il existe de plus en plus de ressources francophones.
- R n'est pas un logiciel au sens classique du terme, mais plutôt un langage de programmation. Il fonctionne à l'aide de scripts (des petits programmes) édités et exécutés au fur et à mesure de l'analyse.
- en tant que langage de programmation, R a la réputation d'être difficile d'accès, notamment pour ceux n'ayant jamais programmé auparavant.

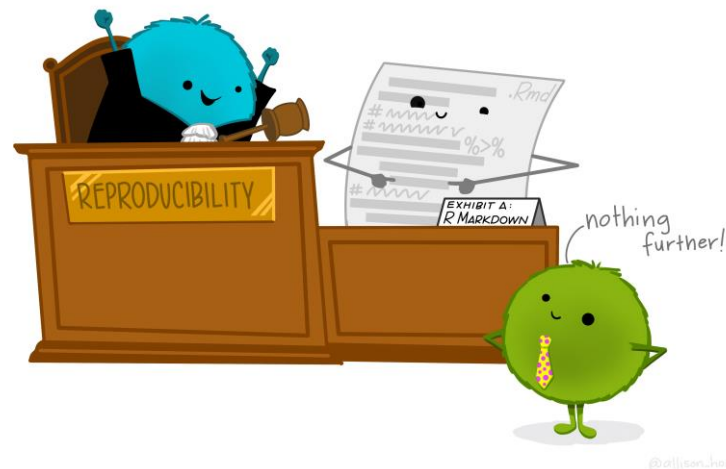




# À propos de R

Le fait de structurer ses analyses sous forme de scripts (suite d'instructions effectuant les différentes opérations d'une analyse) présente de nombreux avantages :

- le script conserve l'ensemble des étapes d'une analyse, de l'importation des données à leur analyse en passant par les manipulations et les recodages.
- on peut à tout moment revenir en arrière et corriger ou modifier ce qui a été fait.
- il est très rapide de réexécuter une suite d'opérations complexes.
- on peut très facilement mettre à jour les résultats en cas de modification des données sources.
- le script garantit, sous certaines conditions, la reproductibilité des résultats obtenus.



@allison\_hart

# À propos de RStudio

---

RStudio n'est pas à proprement parler une interface graphique pour R, il s'agit plutôt d'un *environnement de développement intégré*, qui propose des outils facilitant l'écriture de scripts et l'usage de R au quotidien.

C'est une interface bien supérieure à celles fournies par défaut lorsqu'on installe R sous Windows ou sous Mac.

RStudio est également un logiciel libre et gratuit. Une version payante existe, mais elle ne propose pas de fonctionnalités indispensables.



# Prérequis

---

Pour installer R, il suffit de se rendre sur une des pages suivantes :

- [Installer R sous Windows](#)
- [Installer R sous Mac](#)

Pour installer RStudio, rendez-vous sur [la page de téléchargement du logiciel](#) et installez la version adaptée à votre système.

---

Cette première partie est basée sur le document « Introduction au tidyverse » écrite par Julien Barnier, disponible sur <https://juba.github.io/tidyverse>

Les illustrations sont le fruit du travail de Allison Horst <https://github.com/allisonhorst/stats-illustrations>

Ces derniers sont mis à disposition selon les termes de la [Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions 4.0 International](#).



# Objets et structures de données

---



# Les objets

---

- Sur R, chaque résultat est stocké dans la mémoire vive de l'ordinateur sous forme d'objets qui ont chacun un nom.
- Un objet représente un concept, une idée. Il se matérialise par une entité qui possède sa propre identité. Dans celle-ci, l'on compte deux aspects majeurs: la structure interne et le comportement.
- L'utilisateur agit sur les objets avec des opérateurs (arithmétiques, logiques, comparaison) et des fonctions (qui sont elles mêmes des objets).

# Les objets

---

- Dans R on distingue différents types d'objets :
- caractères (strings en anglais);
- nombres (entiers ou réels);
- dates;
- valeurs logiques qui ne prennent que deux valeurs: TRUE (vrai) ou FALSE (faux);
- facteurs qui sont un format spécial dans R prévu pour les variables catégorielles.

**Ce qui se ressemble s'assemble : on ne peut pas mélanger des objets de différents types !**

# Les objets de type numérique

---

## CONTINUOUS

MEASURED DATA, can have  $\infty$  values within possible range.



I AM 3.1" TALL  
I WEIGH 34.16 grams

## DISCRETE

OBSERVATIONS CAN ONLY EXIST AT LIMITED VALUES, OFTEN COUNTS.

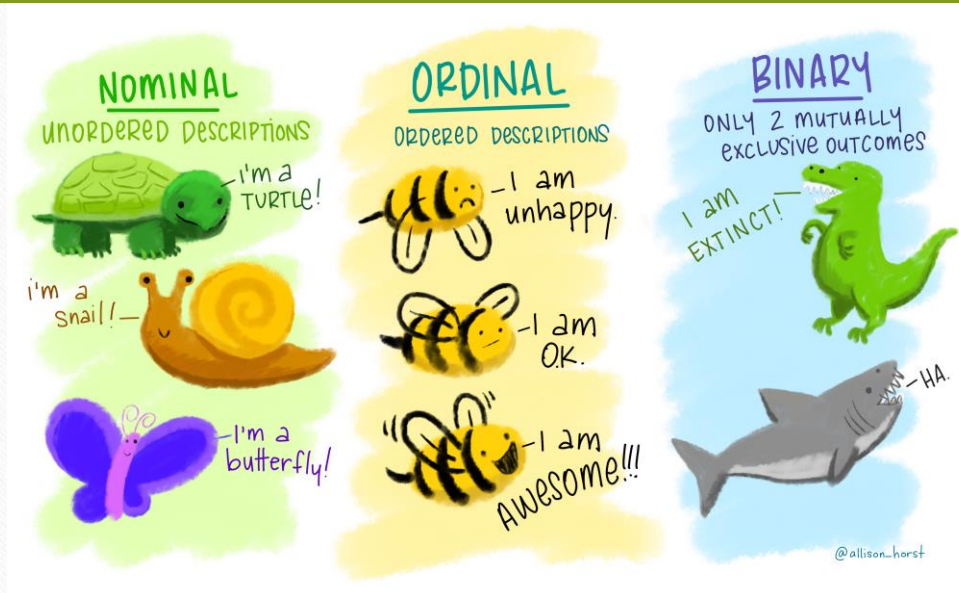


I HAVE 8 LEGS  
and  
4 SPOTS!

@allison\_korot



# Les objets de type caractère



# Les objets complexes

---

- Les vecteurs : ensemble d'éléments de même nature

Si on mélange des vecteurs de différents types, R fera une coercion selon la hiérarchie *logique < entier < réel < caractère*

- Les matrices : collection de vecteurs organisée de façon rectangulaire, ne peuvent former une matrice que des éléments de même nature.
- Les data frames : format d'organisation de données en forme rectangulaire, mais respectant la nature des données qu'elle contient, permet de combiner des variables numériques et caractères.
- Les lists : objets permettant de contenir des données non structurées de la même façon.

# Les packages

---

- Dans R, l'unité de partages de codes est le package, celui-ci contient du code, des données, de la documentation.
- Le répertoire officiel des packages R est le **Comprehensive R Archive Network**, ou CRAN
- En avril 2022 il y avait plus de 19 000 packages disponibles sur le CRAN.



- Pour installer un package depuis le CRAN, il suffit d'utiliser `install.packages("light")`. Celui-ci est ensuite activable avec `library("light")`



- Il est aussi possible de partager des packages sur Github, plateforme de partage de codes, et de les installer à l'aide du package devtools.



# À propos du *tidyverse*

---

Le *tidyverse* est une compilation de packages construits autour d'une philosophie commune et conçues pour fonctionner ensemble.

Elles facilitent l'utilisation de R dans les domaines les plus courants : manipulation des données, recodages, production de graphiques, etc.



# Manipulation de données, jointures et « tidy data »

---



- 
- La plupart du temps, les données brutes ne sont pas utilisables telles quel.
  - Il faut d'abord préparer les données à leur exploitation :
    - 1) Découverte des données
    - 2) Structuration les données pour faciliter leur exploitation
    - 3) Nettoyage des données, suppression des outliers, recodages et corrections.
    - 4) Enrichissement des données, ajout de données additionnelles
    - 5) Validation des données
    - 6) Publication

# Structuration des données : les « tidy data »

“**TIDY DATA** is a standard way of mapping the meaning of a dataset to its structure.”

—HADLEY WICKHAM

## In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement

each column a variable

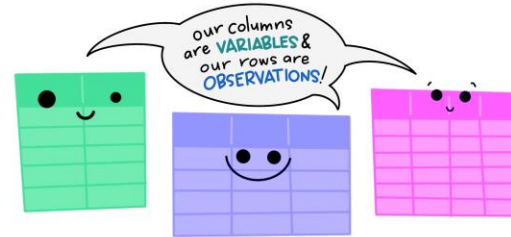
id	name	color
1	floof	gray
2	max	black
3	cat	orange
4	donut	gray
5	merlin	black
6	panda	calico

each row an observation

Wickham, H. (2014). Tidy Data. Journal of Statistical Software 59 (10). DOI: 10.18637/jss.v059.i10

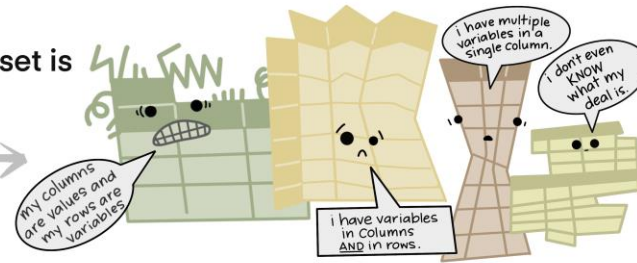
# Structuration des données : les « tidy data »

The standard structure of tidy data means that "tidy datasets are all alike..."



"...but every messy dataset is messy in its own way."

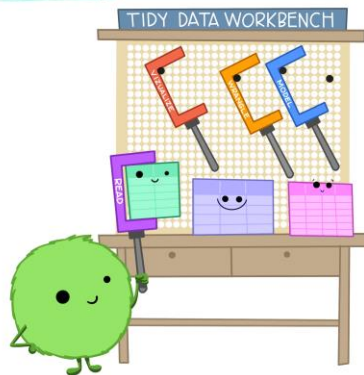
—HADLEY WICKHAM



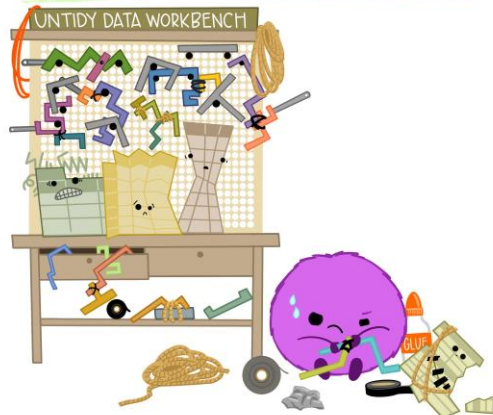


# Structuration des données : les « tidy data »

When working with tidy data,  
we can use the **same tools** in  
**similar ways** for different datasets...

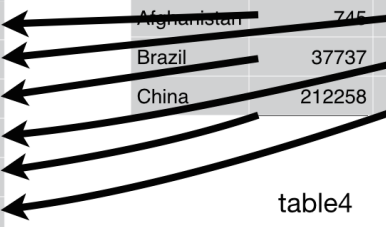


...but working with untidy data often means  
reinventing the wheel with **one-time**  
**approaches** that are **hard to iterate or reuse**.



# Structuration des données : les « tidy data »

Exemple de conversion de données en version « tidy »



country	year	cases
Afghanistan	1999	745
Afghanistan	2000	2666
Brazil	1999	37737
Brazil	2000	80488
China	1999	212258
China	2000	213766

country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

table4

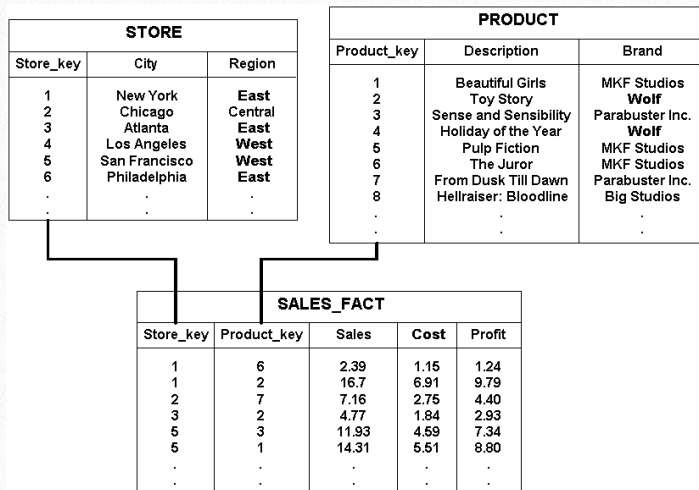
# Enrichissement des données : les jointures

---

- Il est rare de n'utiliser qu'une seule table de données pour une enquête. En général, celles-ci sont séparées en plusieurs tables en fonction de l'entité observée.
- Ces données sont appelées «données relationnelles » et sont liées les unes aux autres par une clé identifiant.

# Enrichissement des données : les jointures

- Exemple d'une base de données relationnelle





# Enrichissement des données : les jointures

- Pour récupérer des informations disponibles dans deux tables différentes, il faut les joindre par leur clé identifiant, c'est le principe de la jointure de tables.

