L'exploration de données avec R

Congrès annuel de l'ACL

Guilherme D. Garcia

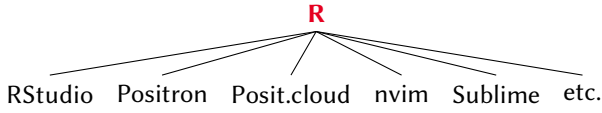
gdgarcia.ca

Juin 2025

Un langage pour l'analyse de données

- Les options les plus populaires : R, Python, Julia ; SPSS, Stata, SAS, MatLab
- R a son propre EDI (éditeur), mais RStudio/Positron est une option supérieure

☞ Plus de 20 000 extensions/bibliothèques disponibles (mai 2025)



Notre objectif aujourd'hui

Les étapes d'une analyse

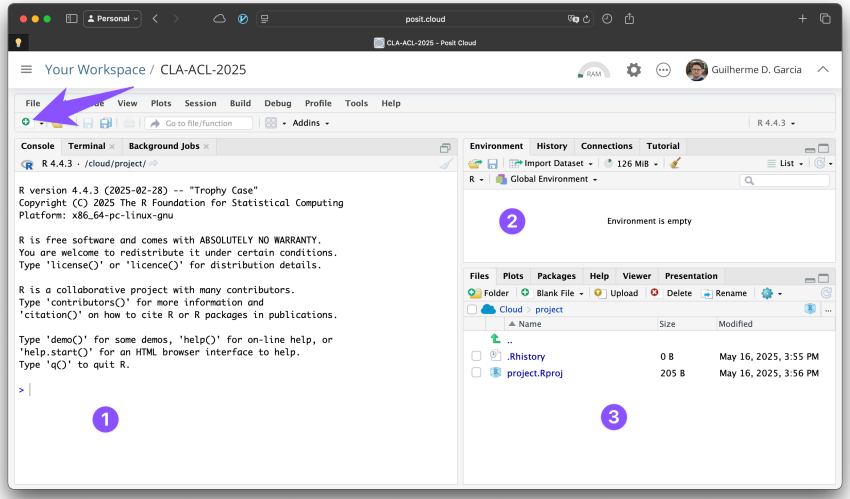
☞ Les étapes suivantes sont typiques dans une étude quantitative (quel que soit l'outil utilisé)

1. Importer nos données
2. Nettoyer/organiser les variables pertinentes
3. Explorer les patrons d'intérêt (y compris la visualisation de données)
4. Analyser les données à partir des tests, modèles, etc.

☞ Aujourd'hui, on explore R à partir des étapes 1–3

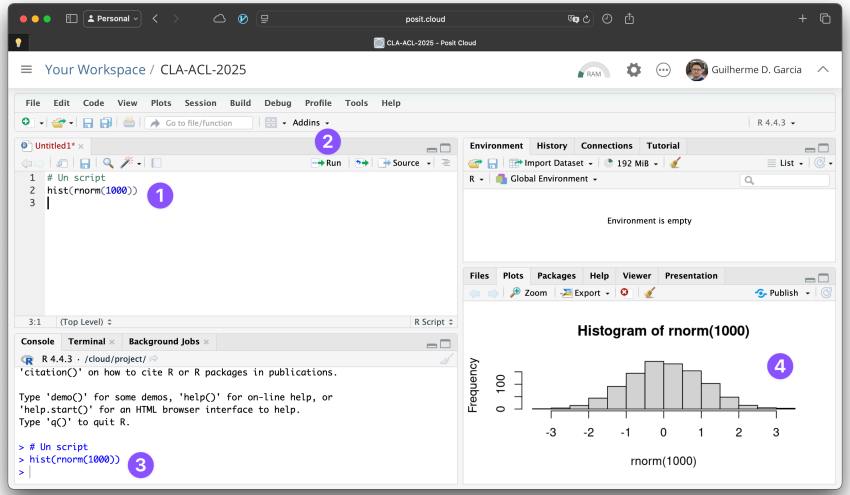
Notre outil

`posit.cloud` : EDI en ligne et gratuit (25 heures par mois)



Notre outil

`posit.cloud` : EDI en ligne et gratuit (25 heures par mois)



Un script

Notre interface avec R

- Si vous êtes familiarisé avec Excel ou SPSS, un script sera **complètement différent**

☞ Un script nous permet de documenter/reproduire une analyse de façon complète et détaillée

```r
# Ceci est un commentaire
  library(tidyverse) # Une bibliothèque

  d <- c(1, 4, 5) # La création d'une variable (un vecteur)
```

- Aujourd'hui on travaillera directement dans un script `.R` pour mieux comprendre le langage

- **Des questions avant de commencer ?**
- Allez sur `posit.cloud` et créez un compte si vous voulez reproduire notre analyse !

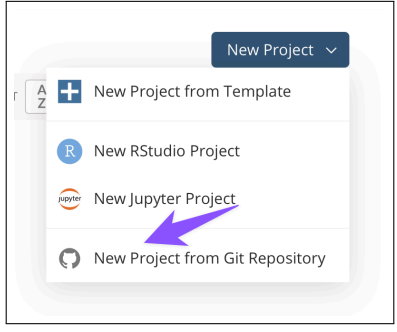
Suggestions de lecture

- Meilleure option pour R en général : Wickham *et al.* (2023)
- En linguistique : Winter (2019) ; Garcia (2021) ; Sonderegger (2023)

Comment commencer

Après avoir créé un compte sur `posit.cloud`

- Cliquez sur « New project » → « New Project from Git Repository »
- Entrez cette adresse : `https://github.com/guilhermegarcia/CLA_data_exp_R.git`



Références I

Guilherme D Garcia : *Data visualization and analysis in second language research*. Routledge, New York NY, 2021.

Morgan Sonderegger : *Regression modeling for linguistic data*. MIT Press, Cambridge, MA, 2023.

Hadley Wickham, Mine Çetinkaya Rundel et Garrett Grolemund : *R for data science*. O'Reilly Media, Inc., Sebastopol, CA, 2023. Available at `https://r4ds.had.co.nz/index.html`.

Bodo Winter : *Statistics for linguists : an introduction using R*. Routledge, New York, 2019.