Crédit photographique: Laura Kapfe, Unsplash

R ET RSTUDIO POUR L'ANALYSE DE DONNÉES TEXTUELLES

Pascal Brissette

Automne 2023

Pourquoi analyser des textes avec un langage de programmation?

1. Un langage de programmation offre beaucoup plus de ***liberté*** qu'un logiciel même très bien fait.
2. Un logiciel, surtout s'il est très puissant (exemple: Excel), utilise beaucoup de ***ressources*** pour simplement fonctionner.
3. Avec la programmation, vous conservez les ***traces*** de chaque opération de chacune des parties de votre analyse. Cela est utile pour vous et pour ceux qui voudraient ***reproduire*** votre expérience.
4. La programmation contribue à formaliser une analyse.
5. La programmation est un appel à la ***créativité*** et à la ***collaboration***.
6. Les ordinateurs et les LLM ne lisent pas des textes comme l'être humain, mais ils sont excellents pour repérer des régularités, des motifs, des corrélations, des cas aberrants. Leur mode de lecture est complémentaire à celui de l'être humain.

Crédit photographique: Janko Ferlic, Unsplash

Qu'est-ce que R?

- R est un langage de programmation complet et un logiciel gratuit distribué sous licence GNU (utilisation, partage, modification autorisés);
- créé en 1993 par Ross Ihaka et Robert Gentleman à l'Université d'Auckland (NZ), inspiré du langage S-PLUS.
- développé et distribué par des statisticiens pour l'analyse de données (intègre de nombreuses fonctions statistiques);
- la première version officielle du langage R a été publiée en février 2000.

Crédit photographique: Markus Spiske, Unsplash

Pourquoi R en particulier?

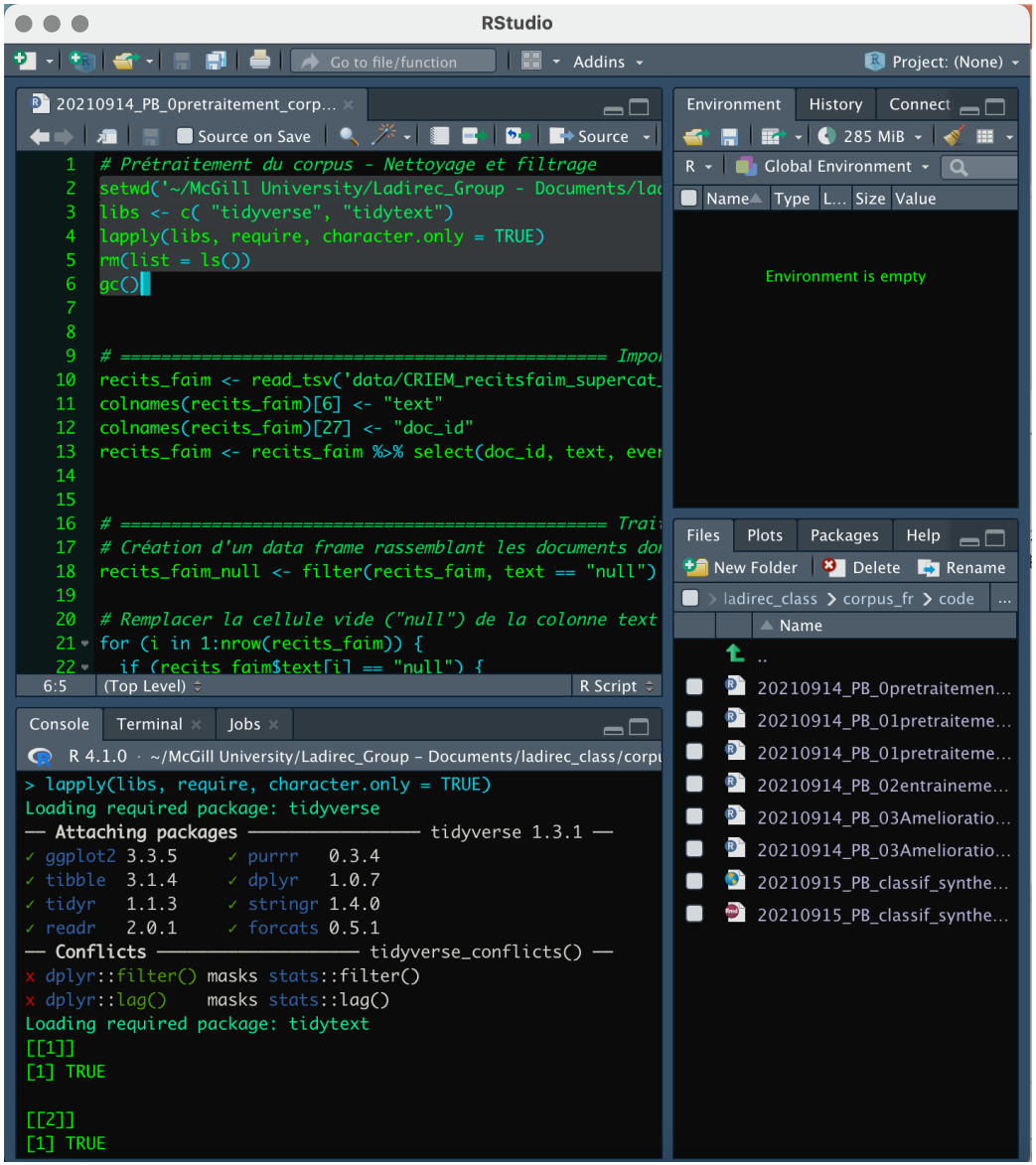
ATOUTS

- R est ***gratuit***, son code est ***ouvert*** et son développement est soutenu par une très large ***communauté scientifique*** (abondante documentation en anglais et en français);
- langage interprété (non compilé): ***exécution facile*** ;
- il comporte plus de ***25 000 extensions*** et peut faire beaucoup plus que des calculs statistiques;
- RStudio (Posit), son environnement de développement, est ***convivial***, ***attrayant*** et ***bien intégré***;
- il peut produire des ***graphiques*** et documents variés (Word, Power Point, pdf, etc.) et de très ***grande qualité***.

LIMITES

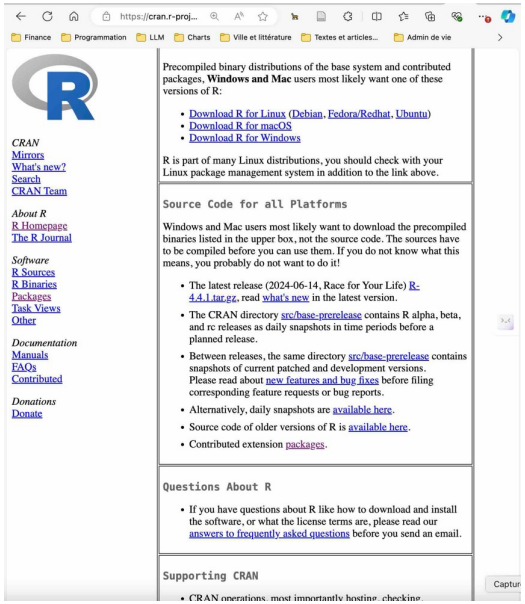
- Il faut consacrer un certain nombre d'heures à l'apprentissage de la syntaxe;
- la plupart de ses opérations se font « en mémoire », ce qui veut dire que l'exécution de certaines tâches peuvent être impossibles ou plus lentes qu'avec d'autres langages (Julia, Go, C#, C++).

RSTUDIO


RStudio
Go to file/function
Addins
Project: (None)
20210914_PB_0pretraitement_corp...
Source on Save
Source
Prétraitement du corpus - Nettoyage et filtrage
setwd('~/McGill University/Ladirec_Group - Documents/lad
libs <- c("tidyverse", "tidytext")
lapply(libs, require, character.only = TRUE)
rm(list = ls())
gc()
# ========== Impor
recits_faim <- read_tsv('data/CRIEM_recitsfaim_supercat_
colnames(recits_faim)[6] <- "text"
colnames(recits_faim)[27] <- "doc_id"
recits_faim <- recits_faim %>% select(doc_id, text, ever
# ========== Trai
# Création d'un data frame rassemblant les documents do
recits_faim_null <- filter(recits_faim, text == "null")
# Remplacer la cellule vide ("null") de la colonne text
for (i in 1:nrow(recits_faim)) {
if (recits_faim$text[i] == "null") {
6:5
(Top Level)
R Script
Environment
History
Connect
285 MiB
R
Global Environment
Name
Type
L...
Size
Value
Environment is empty
Files
Plots
Packages
Help
New Folder
Delete
Rename
ladirec_class
corpus_fr
code
Name
..
20210914_PB_0pretraitemen...
20210914_PB_01pretraiteme...
20210914_PB_01pretraiteme...
20210914_PB_02entraineme...
20210914_PB_03Amelioratio...
20210914_PB_03Amelioratio...
20210915_PB_classif_synthe...
20210915_PB_classif_synthe...
Console
Terminal
Jobs
R 4.1.0 · ~/McGill University/Ladirec_Group - Documents/ladirec_class/corpu
> lapply(libs, require, character.only = TRUE)
Loading required package: tidyverse
── Attaching packages ──── tidyverse 1.3.1 ──
✓ ggplot2 3.3.5 ✓ purrr 0.3.4
✓ tibble 3.1.4 ✓ dplyr 1.0.7
✓ tidyr 1.1.3 ✓ stringr 1.4.0
✓ readr 2.0.1 ✓ forcats 0.5.1
── Conflicts ──── tidyverse_conflicts() ──
x dplyr::filter() masks stats::filter()
x dplyr::lag() masks stats::lag()
Loading required package: tidytext
[[1]]
[1] TRUE
[[2]]
[1] TRUE

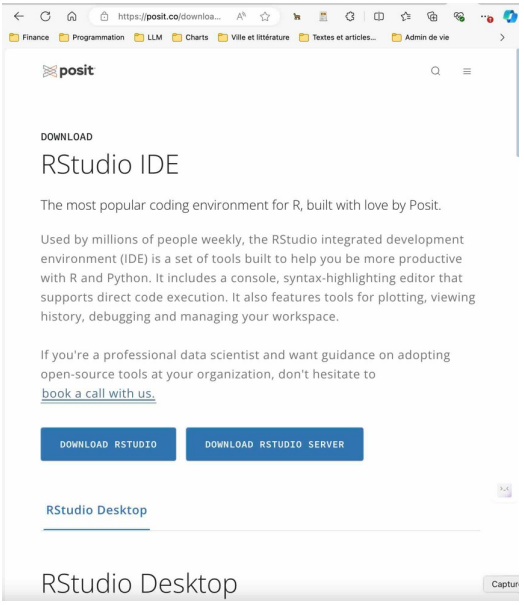
Installer RStudio sur son ordinateur

TÉLÉCHARGER ET INSTALLER LA DERNIÈRE VERSION DU LANGAGE R



https://cran.r-project.org/

TÉLÉCHARGER ET INSTALLER LA DERNIÈRE VERSION DE RSTUDIO

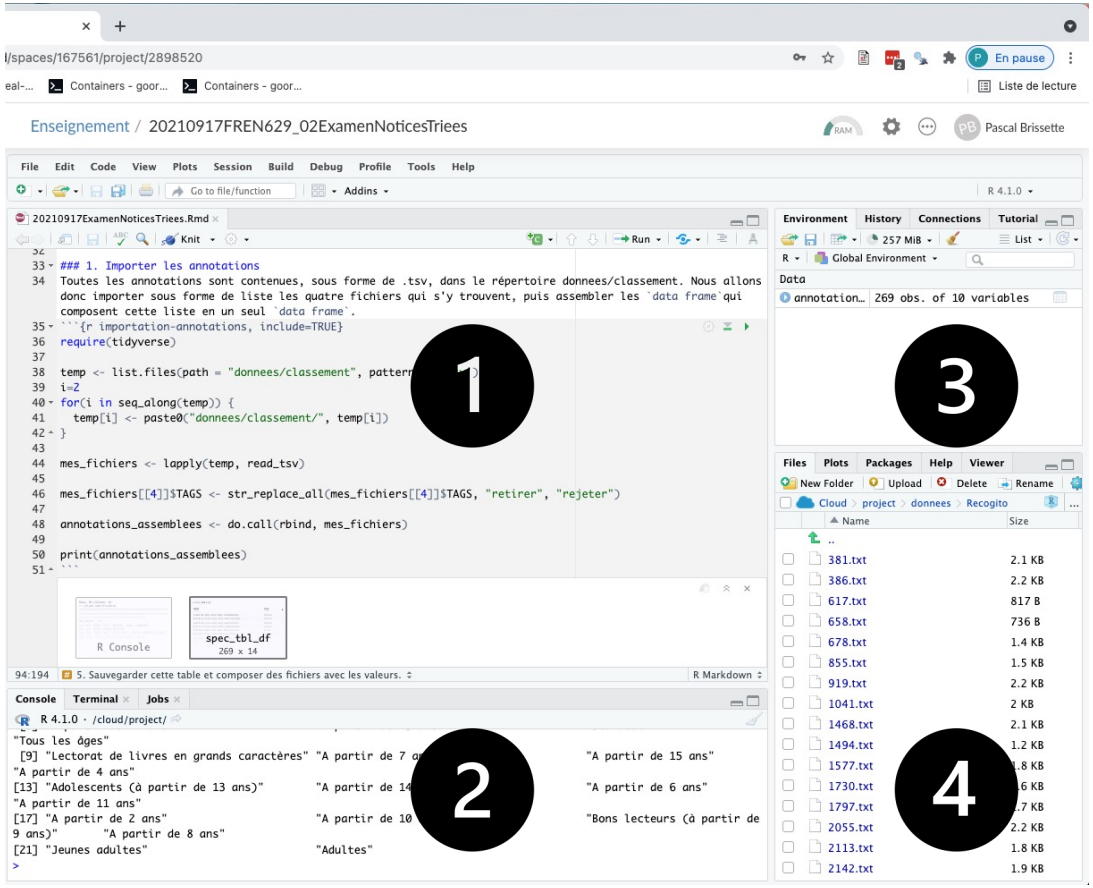


https://posit.co/downloads/

Environnement de développement RStudio

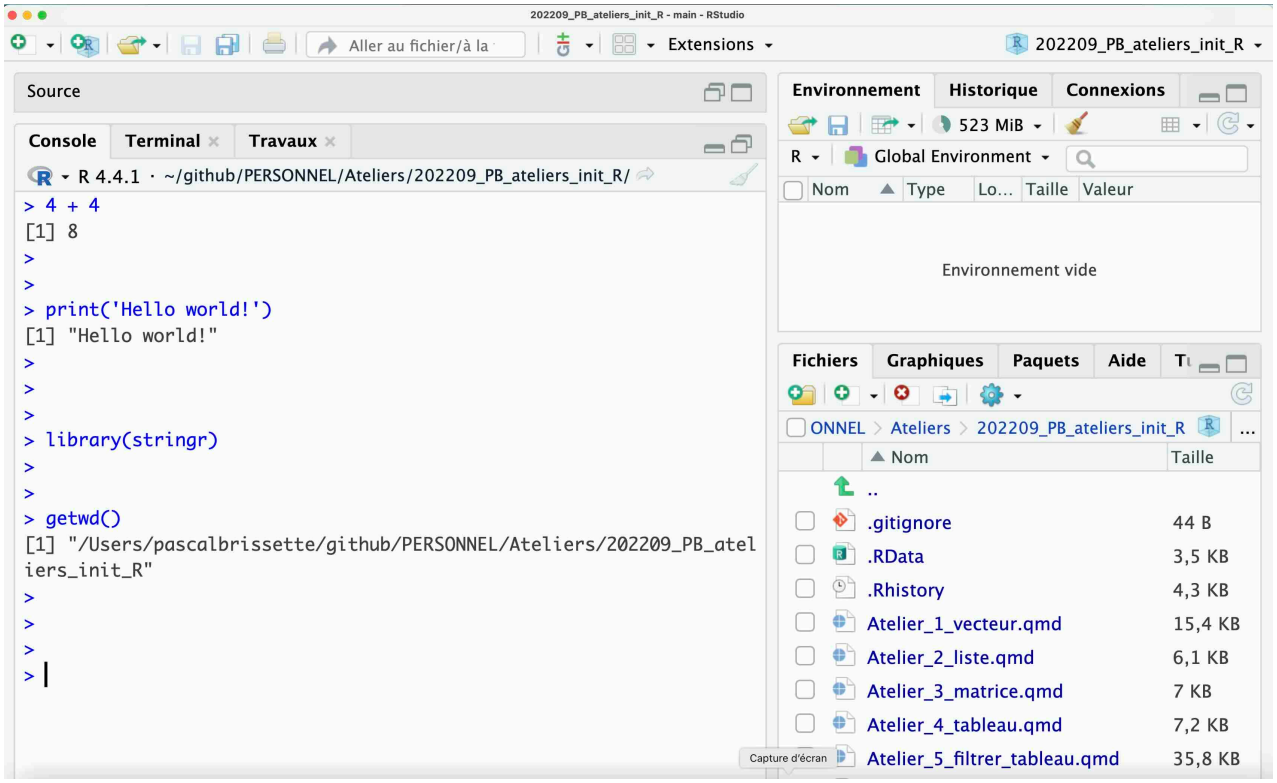
Par défaut

1. Éditeur de script
2. Console (exécution des commandes)
3. Objets en mémoire
4. Dossiers/graphiques/librairies/aide



À vous de jouer!

- Ouvrez RStudio
- Dans L'espace de la console, écrivez 4 + 4 et enfoncez la touche « Return ».
- Voilà! Vous avez fait vos premiers pas dans R!
- Ci-contre, la commande getwd() retourne le chemin du répertoire de travail (*working directory*)

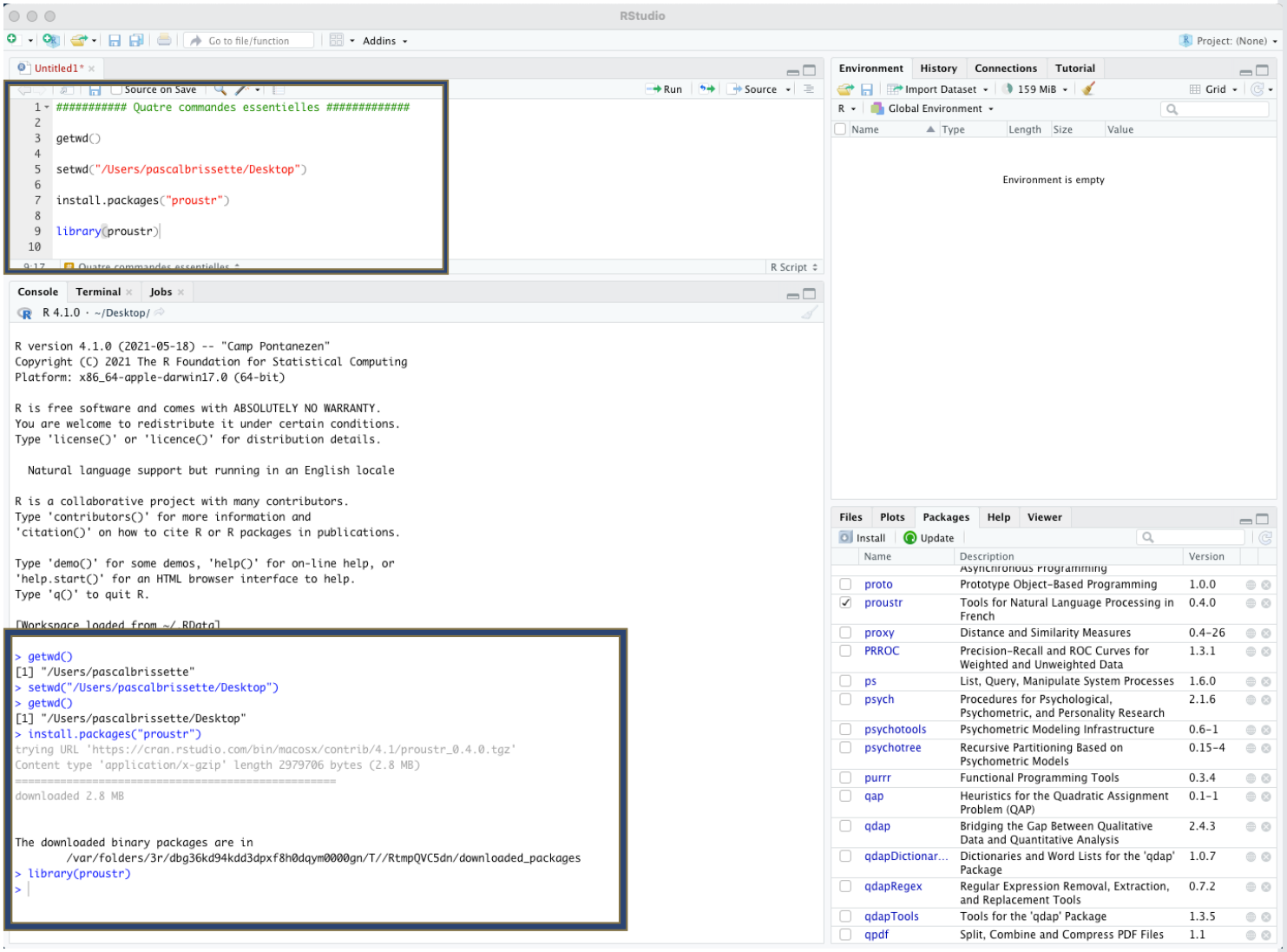


Quatre commandes essentielles

getwd( )

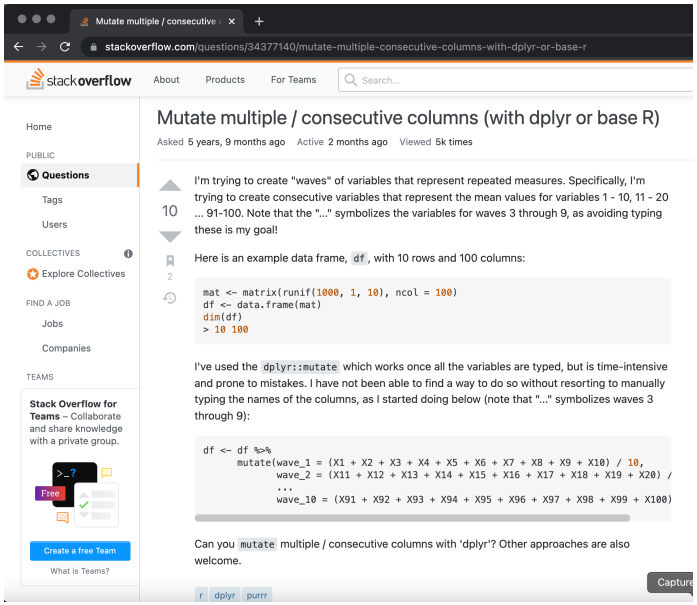
install.packages( )

library( )

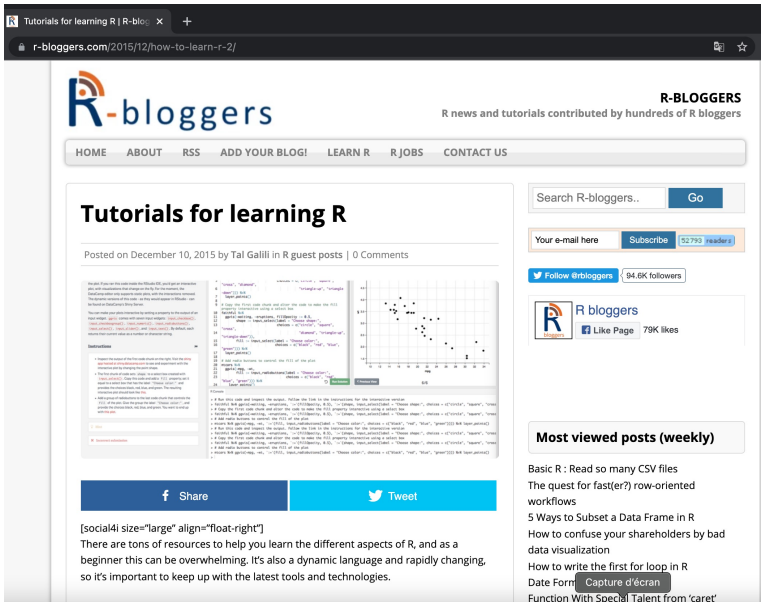


Vous avez des questions? Consultez la communauté R!

STACK OVERFLOW



R-BLOGGERS



Ressources

CRAN Project - https://cran.r-project.org/

RStudio https://www.rstudio.com/

RStudio Cloud https://rstudio.cloud/

Stack Overflow https://stackoverflow.com/questions/tagged/r

R-Bloggers https://www.r-bloggers.com/