Logiciels spécialisés: R

Manipulation et présentation des données avec R

Paolo Crosetto - 24 heures – Septembre 2021

Dans ce cours, on va utiliser le logiciel R, et en particulier son 'dialect' le *tidyverse* pour manipuler, visualiser, et analyser des données, et pour produire automatiquement des rapports statistiques dynamiques et des sites web avec les résultats de l'analyse. L'accent est mis sur l'analyse de jeux de données existants, et utilise R plutôt comme un outil de statistique appliquée qu'un outil de programmation, simulation et analyse économétrique.

Pré-requis:
aucun, mais une quelque familiarité avec la statistique et les données est bienvenue.

Pré-requis techniques:

- les étudiants sont priés de venir en cours avec leurs PC portables
- (si pas possible, on va pouvoir utiliser les PC des salles informatiques)
- installation de R (gratuit, ici: https://pbil.univ-lyon1.fr/CRAN/)
- installation de Rstudio (gratuit, ici: https://www.rstudio.com/products/rstudio/download/#download)
- installation du package *tidyverse*:
  - ouvrez Rstudio
  - assurez-vous que votre PC soit connecté à internet
  - dans la console (en bas à droite) tapez `install.packages("tidyverse")`
  - allez boire un café (cela prend quelques minutes)
- installation de GIT (gratuit, ici: https://git-scm.com/downloads)
- création d'un compte sur github (gratuit, ici: https://github.com/)

Structure du cours:

1. **prise en main du logiciel:** R, Rstudio, pourquoi R et non pas un autre logiciel, ressources (gratuites) en ligne pour apprendre (livres, sites, twitter, blogs, wikis…).
2. **workflow** : travailler de façon efficace, seuls ou à plusieurs ; documenter son code ; utiliser des *repository* ouverts pour gérer le code et l'interaction avec des coauteurs ou le public. Usage de base de GIT et de github.
3. **plotting :** dire la vérité et mentir avec les données, bad and good plots; ggplot, the grammar of graphics; #tidytuesday (package de référence : *ggplot2*).
4. **manipuler les données :** sélectionner, filtrer, transformer, nettoyer, reshape, merge, création de variables, données 'propres' et 'ordonnées', tidy data (package de référence : *dplyr* et *tidyr*).
5. **analyser les données**: statistiques descriptives, analyse par groupe, modèles simples de régression ; appliquer une analyse de façon récursive, pour chaque groupe ; comparer les résultas (package de référence : *dplyr, purrr* et *broom*).
6. **web crawling** : extraire des données de sites web (package de référence : *rvest, RCrawler*)
7. **créer des rapports de stat**: rapports dynamiques avec Rmarkdown et knitr; mise à jour des rapports (package de référence *: knitr*) ; publication des rapports sur github.

Contrôle final:
Les étudiants auront deux semaines pour produire un rapport statistique sur des données qu'ils auront choisi (une liste de jeux de données va être fournite à l'avance). Le rapport sera rédigé en Rmarkdown et publié sur la page github de chauqe étudiant. Le développement se fera également sur github et les codes sources seront ouverts et accessibles. Exemple d'un produit final possible: https://jtanwk.github.io/us-solar/#how_have_solar_panel_costs_changed