Un petit coup de polish - Nettoyage de fichiers Excel avec R

Thomas Vroylandt*

Résumé

Les données qui nourrissent nos analyses sont souvent issues de fichiers Excel (xls ou xlsx), envoyées par d'autres services, disponibles en *open data* ou simplement issues d'outils de collecte de données. La première étape de toute bonne analyse est alors de parvenir à importer ces données correctement. Pourtant ces fichiers Excel ne sont pas tout le temps pensés pour être importés par un ordinateur et un logiciel statistique. Ils peuvent être truffés de formatages - texte en gras, cellules colorées, indentations, cellules fusionnées, pour ne donner que quelques exemples - qui complexifie leur import. Bien qu'il s'agisse d'une étape classique et cruciale, il est courant de buter sur un problème de ce type. Cette présentation se propose d'aborder plusieurs niveaux pour importer ces fichiers et les transformer en un format *tidy* (Wickham (2014)), facilement exploitable par la suite, à l'aide de packages comme {readxl} (Wickham and Bryan (2023)), {tidyxl} (Garmonsway (2023a)) ou encore {unpivotr} (Garmonsway (2023b)). Elle s'appuie sur des jeux de données réels disponibles en accès libre.
Mots-clefs : Excel - Data Cleaning - Nettoyage de données

Développement

Cette présentation s'articulera principalement autour de jeux de données réels et disponibles en ligne plus ou moins complexes à importer et nettoyer, issus d'administrations françaises productrices de données (Insee, Urssaf CN, Drees, etc.).

Elle s'attache à proposer des solutions pratiques, nourries par l'expérience et de complexité croissante pour quelques cas courants compliquant l'import des données :

- tableaux mal positionnés sur la feuille
- onglets contenant chacun une année/région/etc.
- formatages avancés (cellules fusionnées, indentations)
- information sous une autre forme que le texte (gras, couleur)
- fichiers différents d'une année à l'autre
- plusieurs tableaux à la suite dans un même fichier

Références

Garmonsway, Duncan. 2023a. *Tidyxl: Read Untidy Excel Files.* https://CRAN.R-project.org/package=tidyxl.

———. 2023b. *Unpivotr: Unpivot Complex and Irregular Data Layouts.* https://CRAN.R-project.org/package=unpivotr.

Wickham, Hadley. 2014. "Tidy Data." *The Journal of Statistical Software* 59.

Wickham, Hadley, and Jennifer Bryan. 2023. *Readxl: Read Excel Files.* https://CRAN.R-project.org/package=readxl.

*Kantiles, thomas@kantiles.com