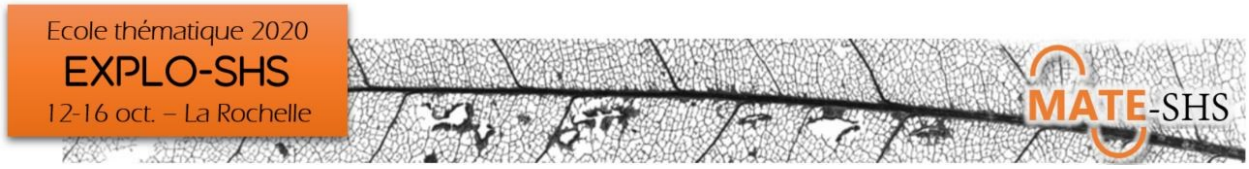

Comment utiliser le nettoyage des données pour explorer, rendre compte d'un potentiel scientifique : le langage R

Maël Theulière (en visio), Hélène Mathian

Atelier, mercredi 14 octobre 9h30-12h15

Présentation générale

Nous souhaiterions aborder au cours de cet atelier un certain nombre de « procédures » qui permettent le nettoyage des données , mais qui nécessairement les interrogent et nous conduit à faire des choix de recodage, à construire des modèles de lecture des enregistrements.

En particulier les premières explorations s'attachent à explorer la cohérence des données. Ces tests peuvent être de simples tris à plat par exemple, ou être plus complexes.

Dans un premier temps on s'attachera à illustrer :

- Le recodage de données manquantes
- Le repérage d' « outliers »
- Le nettoyage et le recodage de chaînes de caractères

Dans une deuxième partie on abordera des tests de cohérences sur des dimensions plus spécifiques que sont le temps et l'espace.

Environnement informatique

R et Rstudio

Type de données traitées - droit d'accès

Données territoriales, données open data , données réseaux sociaux

Niveau requis

Être à l'aise avec la manipulation de données en général et avec le package dplyr sous R.

Objectifs

Illustrer quelques pratiques de "nettoyage" de données statistiques

Formule pédagogique

Fourniture des données préparées et des codes. Identification des étapes et discussion autour des choix de la mise en œuvre.

Liens

R (https://www.r-project.org/) , Rstudio (https://rstudio.com/products/rstudio/download/) la liste des packages nécessaires au TP sera fournie en arrivant.