TP21 – Nettoyage de données

Encodage des caractères

Ressource : RichestFrenchFamiliesISO-8859-1.csv, RatesSweden.csv

Exemple : RichestFrenchFamiliesISO-8859-1.csv. Tous les caractères spéciaux français sont affichés en tant que points d'interrogation.

1. Outil : Sublime Text 2
    a. File -> Re-open with Encoding -> Choose UTF-8
2. Outil : Bloc-Notes
    a. File -> Save As -> Choose UTF-8
3. Outil : Excel
    a. Data -> From Text -> Get Data -> Delimiter -> Encodeing popup menu

4. Exemple : Crowdsourcing pour enquêter sur les taux d'intérêt sur les hypothèques en Suède : RatesSweden.csv
    a. Outil : www.openRefine.org
    b. Tous les caractères spéciaux suédois sont affichés en tant que points d'interrogation.
    c. Créer un projet -> Choose File -> Caracters Encoding -> UTF-8
    d. Convertir automatiquement les valeurs en nombres : cocher 'Parse cell text into numbers'
    e. Renommer le projet et clicker 'Create Project'
    f. Remplacer un caractère
        i. En général, l'écriture est : replace (value, "ancien texte", "nouveau texte"). Par exemple: Rate -> Edit Cells -> Transform -> Expression -> replace (value, ",", "."). Ou bien: Zip Code -> Edit Cells -> Transform -> Expression -> replace (value, " ", "") (remplacer l'espace)
    g. Convertir le text en nombre :
        i. Zip Code-> Edit Cells -> common Transforms -> to numbers
    h. Analyse des données
        i. Par exemple la distribution (l'histogramme) des taux : Rate -> Facets -> Facet numérique