TP23 – Nettoyage de données

Correction des erreurs dans les données

Ressource : titanic.csv, Height_shoe_size.xlsx, ABRUZZO-Schools.csv

1. Ouvrir le fichier titanic.csv dans OpenRefine
2. Convertir le texte en nombre dans les colonnes adéquates et créer un projet
3. Valeur texte invalide
    a. Dans la colonne ‘sexe’ nous devrions trouver homme ou femme
    b. sexe -> Facet -> Text Facet
    c. Exemple : la ligne avec '-' indique un nom féminin, donc il peut être modifié en femme
    d. Réinitialiser la facette 'sexe' pour mettre à jour les données éditées
4. Valeur numérique invalide
    a. age -> Facet -> Numeric Facet
    b. Exemple : les âges vont de 0 à 500, ce qui est incorrect pour les âges en 1914
    c. Modifiez l'âge incorrect et réinitialisez la facette 'age'.
5. Valeur aberrante
    a. Exemple : Ouvrez le fichier Height_shoe_size.xlsx dans Google spreadsheet
    b. Utilisation de l’histogramme pour repérer les valeurs invalides
    c. Utilisez un ‘scatter plot’ pour voir la corrélation entre la hauteur et la taille de la chaussure et voir s'il existe des valeurs non valides.
    d. Sélectionnez tout -> Insérer -> Graphique
6. Doublants. Évaluation des écoles italiennes à risque en cas de tremblement de terre
    a. Ouvrez ABRUZZO-Schools.csv avec OpenRefine
    b. Convertir le texte en nombre dans les colonnes adéquates et créer un projet
    Note : Remarquez qu’une même école apparaît parfois plusieurs fois avec des risques sismiques différents (différents bâtiments). Supprimer les doublons des données dépend de l'histoire que vous voulez raconter.
    c. Exemple : Informer les lecteurs du montant maximal de risque
        i. Enlever les écoles avec moins de risques
        ii. Concaténer le nom de l'école avec la ville :
        School name -> Edit cells -> Transform-> Expression -> value + "," + cells ["City"].value
    d. Enlever les doublons
        i. School name -> Edit cells -> blank down
        ii. School name -> Facet -> text Facet
        iii. clicker en bas sur 'blank' -> ‘Invert’ (en haut)