Data wrangling, généralement les données n’arrivent pas dans un seul fichier, elles sont cachés il faut aller les chercher. Le concept de data wrangling consiste à retravailler ces données que l’on reçoit sous différent formats pour qu’on puisse en faire qqch et les mettre sous un format que l’on peut travailler.

Data cleaning à cette étape toutes les données sont bien rangés bien cadrés, on peut donc bien travailler avec ces données. On retravaille les données pour n’avoir que de int et des float pour pouvoir faire du data analysis/machine learning/data science. On va essayer de minimiser tout ce qui est valeur manquantes et erreurs

CSV : Command Seperated Values

Les données dirty (avant Data wrangling) sont composées de :

- date

- texte

- nombres

Les données clean (il n’y a que des nombres, float ou int) sont composées de :

- numérique

- catégorique

- ordinal

Catégorique :

ONE-HOT-ENCODING : consiste à prendre les valeurs d’une colonne de dataset dirty pour les transposer en différentes colonnes. Exemple : des vêtements qui sont stocké dans un dataset avec leur couleur. En clean on veut seulement des nombres, on transpose donc les différentes couleurs de tous les vêtements en colonnes différentes (1colonne bleu, 1 colonne rouge, 1 colonne verte etc...) et on remplira ces colonnes par False/0 (si le vêtement n’appartient pas à la catégorie de couleur) ou par True/1 (si le vêtement appartient à la catégorie).

Ordinal :

Exemple : Pauvre<Moyen<Riche -> 0<1<2