Les DataFrames de Pandas

Exercice 1 : Consommation des véhicules

- *Lire les noms des véhicules du fichier Véhicules.csv, affichez ensuite le dataFrame.*
- *Transformez la feature 'Unnamed: 0' en index, et supprimez ensuite cette colonne. Donnez le nom 'Véhicules' à l'index.*
- *Quel est la taille du dataFrame. De quel type sont les différentes features.*
- *Combien y-a-t-il de valeurs manquantes.*
- *Faire une analyse statistique des données (describe). Comment peut-on analyser les données obtenues, pensez-vous qu'il y ait des outliers.*
- *Affichez les véhicules par ordre de prix croissant.*
- *Combien y-a-t-il de véhicules avec un prix supérieur à 1e+5 euros.*
- *Utiliser maintenant la fonction cut de pandas pour découper les prix des véhicules en 20 bacs. Comment se répartissent ces bacs en fonction du nombre de valeurs dans chaque bac. Comment interpréter le fait qu'il y ait une seule voiture dans le bac 19 et presque toutes les autres dans les premiers bacs.*
- *A partir du boolean indexing récupérer dans un dataFrame les véhicules ayant un prix inférieur à 1e+5. Combien de véhicules ont été supprimés.*
- *Faire une nouvelle analyse statistique des données que constatez-vous en particulier sur le prix moyen des véhicules.*
- *Affichez le niveau de corrélation entre variables. Pensez-vous qu'il y ait un lien entre le prix et la consommation. Quelle est la feature qui semble le plus impacter la consommation. Quelles sont les features les moins inter-dépendantes.*
- *Importez maintenant la bibliothèque pyplot de matplotlib et afficher via la fonction scatter les prix en fonction des consommations. Pensez-vous que la dépendance soit linéaire ou quadratique.*

Exercice 2 : Émission de gaz à effet de serre

- *Lire les données du fichier Gaz.csv. Vous y ajouterez une option de lecture dtype={'INSEE' : str } afin d'éviter un varning. En effet le code INSEE mélange des valeurs entières pour presque toutes les communes sauf la*

corse qui intègre les caractères A et B pour les deux départements et est donc de type string.

- *Faites en sorte que l'index soit sur le code INSEE.*
- *Quel est la taille du dataFrame. De quel type sont les différentes features. Affichez également en utilisant la fonction value_counts le nombre de chaque type.*
- *Combien y-a-t-il de valeurs manquantes dans les différentes features.*
- *Faire une analyse statistique des données (describe). Comment peut-on analyser les données obtenues, pensez-vous qu'il y ait des outliers.*
- *Affichez quelles sont les communes qui n'ont pas de données sur les déchets.*
- *Quels sont les départements qui produisent le plus d'émission de co2 dues aux déchets (celles qui ont plus de 2e+5).*
- *Même question pour l'agriculture. Vous afficherez ensuite le nombre de valeurs manquantes sur l'agriculture dans les départements. Quels sont les deux départements qui ont le moins d'émission de co2 due à l'agriculture.*
- *Sélectionnez les samples qui ont une émission de co2 supérieure a 3e+5 sur la feature 'Autres transports international'. Que constatez-vous.*
- *Sélectionnez les samples qui ont une émission de co2 supérieure a 1e+5 sur la feature 'Routier'. Vous comptabiliserez la répartition par départements. Quels sont les deux départements qui semblent émettre le plus de co2.*