

TD2 – Partie A – Exploration de données

Exploration de données

1. Importez le fichier « Produits.xlsx » dans un dataframe intitulé « produits »

```
import pandas as pd
produits = pd.read_excel("Produits.xlsx")
```

2. Affichez les 7 premières lignes du dataframe « produits »

```
produits.head(7)
```

	Nom	Categorie	Origine	Ville	Prix	Stock	Valeur
0	Alice Mutton	Viandes	CEE	Lyon	195.0	0	0.0
1	Aniseed Syrup	Condiments	Exterieur	Paris	50.0	25	1250.0
2	Boston Crab Meat	Poissons et fruits de mer	Exterieur	Paris	92.0	30	2760.0
3	Camembert Pierrot	Produits laitiers	Exterieur	Paris	170.0	0	0.0
4	Carnarvon Tigers	Poissons et fruits de mer	Exterieur	Lyon	312.5	0	0.0
5	Chai	Boissons	CEE	Lyon	90.0	10	900.0
6	Chang	Boissons	Exterieur	Lyon	95.0	25	2375.0

3. Affichez les 5 dernières lignes : tail()
4. Affichez un aperçu des caractéristiques des variables quantitatives.

```
produits.describe()
```

Des résultats obtenus, quelle est la moyenne de la variable « Prix » ? L'écart type de la variable « Stock » ?

5. Affichez un résumé de toutes les variables.

```
produits.dtypes
```

Quelles variables sont numériques, et lesquelles sont catégorielles ?

6. Affichez les noms, catégories, origines et prix des produits dont la catégorie est « Boissons »

```
produits.loc[produits["Categorie"]=="Boissons",["Nom","Categorie", "Origine", "Prix"]]
```

7. Affichez les noms, catégories, origines et prix des produits dont la catégorie est « Boissons » et le prix supérieur à 100

8. Affichez les noms, catégories, origines et prix des produits dont la catégorie est « Boissons » et le prix supérieur à 100 et l'origine est « CEE »
9. Affichez les noms, catégories, origines et prix des produits dont la catégorie est « Boissons » et l'origine « CEE » ; ou bien la catégorie est « Condiments ».
10. Affichez les noms, catégories, origines et prix des produits dont le prix est compris entre 70 et 100.
11. Listez les 15 produits les moins chers
12. Affichez la moyenne des prix de tous les produits par catégorie

```
produits.groupby("Categorie").agg({"Prix": "mean"})
```

13. Affichez le minimum et le maximum des prix de tous les produits par catégorie
14. Affichez la moyenne des prix des produits de type « Boissons » par origine
15. Affichez le nombre de produits par catégorie.

16. Importez le fichier « Census.xlsx » dans un dataframe et affichez les 5 premières lignes (préciser la feuille « Adult » lors de l'importation du fichier).

Ce fichier comporte un ensemble de descriptions qui permettent de prédire si une personne a un salaire supérieur à un certain seuil.

17. Combien y'a-t-il de variables et d'observations ?

```
census.shape
```

18. Quelle est la distribution des valeurs de la colonne « sex » ?

```
census["sex"].value_counts()
```

19. Quelle est la distribution des valeurs de la colonne « class » ?

20. Croiser les variables « classe » et « sex ».

```
pd.crosstab(census["sex"], census["classe"])
```

Quelle est la proportion des « more » dans l'échantillon global ? Parmi les hommes ? Parmi les femmes ?

Est-ce que ce résultat nous permet de conclure que le niveau de revenu est différent selon que l'on est un homme ou une femme ?