LP-Informatique Décisionnelle

Travaux Pratiques

TP1 : Manipulation de données avec Pandas

Objectifs :

1- Installation de l'environnement de travail.

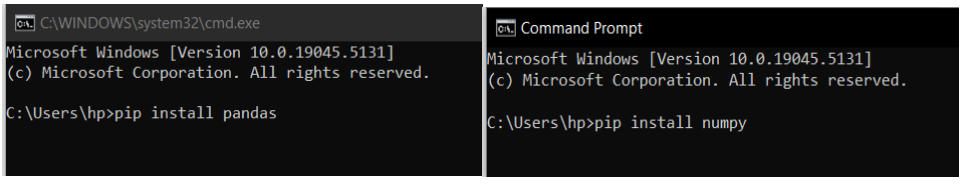
2- Introduction à NumPy et Pandas pour manipuler et analyser des données, en appliquant des techniques de nettoyage, d'exploration et de transformation de données adaptées à l'analyse décisionnelle.

- Installer la dernière version de Python via le site web officiel : https://www.python.org/downloads/.

  Cocher la case « add python.exe to PATH »

  A la fin de l'installation, cochez « Disable path lenght Limit »

- Installer les bibliothèques NumPy et Pandas avec « **pip »** dans l'invite de commande**.**



- Installer Jupyter notebook avec la commande **«pip install notebook»**

Lancer Jupyter via cmd : **jupyter notebook** et créer un notebook, le renommé **« Nom_Prenom_ID_TP1 ».**

Exercice 1 : Chargement et consultation de données

1) Créez le DataFrame ci-dessous avec Pandas et affichez les cinq premières lignes.

date	Produit	region	Ventes	Profit	satisfaction_client	ad_depense
15/01/2023	Product A	Nord	1500	300	4	200
05/02/2023	Product D	Ouest	500	80	2	100
07/02/2023	Product E	Nord	0		1	50
10/02/2023	Product F	Sud	1200	250	4	180
15/03/2023	Product G	Est	700	100	2	
20/03/2023	Product I	Ouest	1100	200	3	150
10/04/2023	Product J	Sud	0	0		0
12/04/2023	Product K	Est	900	130	4	100
17/04/2023	Product M	Nord		350	4	200

2) Quelles informations générales pouvez-vous obtenir sur ce DataFrame (colonnes, types de données, taille) ?
3) Affichez les statistiques descriptives des colonnes numériques.
4) Affichez le nombre total de lignes et de colonnes du DataFrame.
5) Vérifiez si toutes les valeurs de la colonne ***ventes*** sont positives.
6) Affichez la liste des noms de colonnes du DataFrame

Exercice 2 : Nettoyage de données

1) Identifiez les colonnes ayant des valeurs manquantes
2) Affichez le nombre de valeurs manquantes par colonne.
3) Supprimez toutes les lignes ayant des valeurs manquantes dans les colonnes ***ventes*** et ***profit*** uniquement, et afficher le dataframe complet.
4) Remplacez les valeurs manquantes dans la colonne ***client_satisfaction*** par la valeur moyenne de la colonne, et afficher les données de la colonne ***client_satisfaction.***
5) Supprimez les lignes doublons, s'il y en a.
6) Vérifiez si toutes les valeurs manquantes ont été traitées. Sinon ; remplacez les valeurs manquantes par la valeur 0.
7) Vérifiez la présence des doublons.
8) Trier les lignes du DataFrame en ordre décroissant selon la colonne Ventes.
9) Calculez la somme des ventes pour chaque région.

Exercice 3 : Manipulation avancée

1) Filtrez les lignes où la valeur dans la colonne ***Ventes*** est supérieure à 1000.
2) Ajoutez une nouvelle colonne ***profit_margin***, qui représente le ratio ***profit / ventes***.
3) Convertir la colonne ***date*** en format **datetime** et extraire les mois dans une nouvelle colonne ***Mois***.
4) Calculez et afficher le chiffre d'affaires total (***ventes***) par mois.
5) Affichez les 5 produits ayant les plus hauts profits.
6) Affichez les ventes moyennes (***ventes***) par région (***region***).
7) Créez une colonne ***high_sales*** avec la valeur ***True*** si les ventes (***ventes***) sont supérieures à 1000 et ***False*** sinon.
8) Calculez le nombre total de produits vendus pour chaque niveau de satisfaction client (***client_satisfaction***).
9) Créer une colonne ***high_profit*** avec la valeur **1** si le profit est supérieur à la moyenne, sinon **0**.