TP 1 Pandas

Exercice 1 : Analyse des ventes

Dataset: *supermarket_sales – Sheet1*

1. Charger le dataset et afficher un résumé statistique.
2. Identifier et gérer les valeurs manquantes si elles existent.
3. Calculer le chiffre d'affaires total par ville et déterminer la ville avec les ventes les plus élevées.
4. Trouver les 3 catégories de produits (Product line) ayant le chiffre d'affaires moyen le plus élevé.
5. Calculer le pourcentage des ventes totales par genre (Male vs Female).
6. Identifier les 5 factures (Invoice ID) ayant le montant total le plus élevé.

Exercice 2 : Analyse des performances des élèves

Dataset: *StudentsPerformance.csv*

1. Charger le dataset et afficher le nombre d'élèves par genre.

2. Comparer les scores moyens selon le niveau d'éducation des parents.
3. Déterminer le nombre d'élèves ayant obtenu un score parfait (100) dans au moins une matière.
4. Trouver la corrélation entre les scores des différentes matières.

Exercice 3 :

Dataset: *telecom_churn.csv*

1. Charger le dataset et afficher la taille du dataframe.
2. Afficher les colonnes et les informations du dataframe
3. Changer le type de la colonne churn en entier
4. Afficher les caractéristiques statistiques de chaque caractéristique numérique puis des caractéristiques non numériques.
5. Afficher la distribution de churn
6. Faire le tri décroissant par Total des frais de jour
7. Quelle est la proportion d'utilisateurs qui ont churnés dans notre dataframe ?

8. Combien de temps (en moyenne) les utilisateurs qui ont churnés passent-ils au téléphone pendant la journée ?
9. Quelle est la durée maximale des appels internationaux chez les utilisateurs fidèles (Churn == 0) qui n'ont pas de forfait international ?
10. Afficher les colonnes de state à area code avec l'indexation par nom
11. Afficher la dernière ligne du DataFrame
12. Afficher les 5 premières lignes avec les éléments de state terminant par 'V'
13. Revenez à la colonne international plan et remplacez 'No' par False et 'Yes' par True

Exercice 4:

Dataset : hubble_data.csv

1. Chargez le dataset **hubble_data.csv**
2. Renommez les colonnes distance et recession_velocity en **dist** et **rec_vel** respectivement.
3. Chargez le dataset hubble_data_no_headers.csv en attribuant des noms aux colonnes **dist et rec_vel.**
4. Affichez les informations sur ces données
5. Sélectionnez la colonne dist en utilisant la fonction **tail()**
6. Affichez les 5 premières lignes de la colonne **dist** de deux manières différentes
7. Calculez l'énergie en utilisant la formule suivante $E = K \cdot dist + 0.5 \cdot v^2$ sachant que $K = 100$ et v représente la colonne recession_velocity.
8. Ajoutez la colonne dist2 qui contient la colonne $dist^2$
9. Supprimez la colonne dist2
10. Supprimez la deuxième ligne du dataset puis vérifiez si elle a été bien supprimer (attention paramêtre inplace).
11. Modifiez la colonne dist pour qu'elle représente l'index du dataset
12. Affichez le type puis calculez la moyenne et la médiane de la colonne énergie.
13. Afficher les caractéristiques statistiques de la colonne énergie
14. Calculez et affichez la matrice covariance des données
15. Affichez la valeur 194.3740 de la matrice de deux manières différentes
16. Mettre à jour les valeurs de la colonne E = 1620 pour dist =2.0
17. Mettre à jour les valeurs de la colonne E = 1800 pour dist =2.0 en utilisant .at