UNIVERSITE D'ABOMEY-CALAVI

INSTITUT DE MATHEMATIQUES ET DE SCIENCES PHYSIQUES

Code UE: DTS2204

INTITULÉ DE L'UE: Data Science

INTITULÉ ECUE: Principes de la data science

MASSE HORAIRE: 30H (20H+10H)

ENSEIGNANT: Dr. Olouladé B. Moctard

Exercices

Mars 2025

Exercice1

Vous disposez d'un fichier Excel nommé 'employes_dataset.xlsx' contenant 1000 enregistrements sur des employés dans une entreprise multinationale. Chaque ligne représente un employé avec différentes informations personnelles et professionnelles.

Objectifs : Explorer, analyser et nettoyer ce jeu de données à l'aide de la bibliothèque Pandas, et utiliser NumPy lorsque cela est pertinent pour les calculs statistiques.

Consignes:

1. Afficher les 10 premières lignes du fichier.
2. Afficher les noms de colonnes du fichier.
3. Compter le nombre d'hommes et de femmes.
4. Identifier les 5 pays les plus représentés.
5. Calculer le salaire moyen, médian, minimum, maximum et l'écart-type à l'aide de NumPy.
6. Donner l'âge moyen par département.
7. Trouver la ville ayant le plus grand nombre d'employés.
8. Lister les 10 employés les mieux payés.
9. Calculer le nombre d'employés par département et par sexe.
10. Générer un graphique montrant la distribution des âges.
11. Identifier les colonnes contenant des valeurs manquantes.
12. Remplacer les NaN de la colonne 'Télétravail (%)' par la moyenne de cette colonne.
13. Remplacer les NaN de la colonne 'Télétravail (%)' par une valeur estimée pertinente (par exemple, la moyenne par département ou selon l'âge).
14. Supprimer les lignes dont la colonne 'Performance (Note)' est manquante.
15. Proposez une stratégie pertinente pour gérer les valeurs manquantes dans 'Performance (Note)' : suppression, remplissage conditionnel, ou autre — justifiez votre choix.
16. Convertir la colonne 'Date d'embauche' en format datetime.
17. Créer une colonne 'Ancienneté (années)' basée sur la date d'embauche.
18. Supprimer les doublons éventuels du fichier.
19. Uniformiser les majuscules dans les colonnes 'Nom', 'Prénom', 'Ville', 'Pays'.
20. Créer une colonne 'Email valide' qui indique si l'e-mail semble valide (xxx@xx.xx').
21. Supprimer les valeurs extrêmes (outliers) dans la colonne 'Salaire' en utilisant l'écart interquartile (IQR).
22. Vérifier si la distribution des âges suit une loi normale à l'aide de NumPy.
23. Créer une colonne 'prime' : si la performance est ≥ 4 et l'ancienneté ≥ 5 ans $\rightarrow$ 1000, sinon 0 (avec NumPy).
24. Encoder la colonne 'Sexe' en valeurs numériques (0 pour Femme, 1 pour Homme)
25. Utilisez pd.cut() pour créer une nouvelle colonne 'Tranche d'âge' avec les intervalles suivants : 0-25, 26-30, 31-40.
26. Ajoutez une colonne de langages (listes) fictive à 4 lignes et utilisez 'explode()' pour séparer chaque ligne par langage.
27. Transformez votre DataFrame avec un MultiIndex sur 'Département' et 'Sexe', puis utilisez 'stack()' et 'unstack()' pour observer la structure.

28. Détectez les outliers dans la colonne ‘Salaire’ avec la méthode de l’IQR, et créez un DataFrame les contenant.
29. Encodez la colonne ‘Département’ avec ‘get_dummies()’ pour préparer les données à une analyse automatique.
30. Ajoutez une colonne de date fictive (date d’inscription) et utilisez ‘.dt.year’ et ‘.dt.month’ pour extraire les informations temporelles.
31. Créez une fonction personnalisée pour catégoriser l’âge et appliquez-la à votre DataFrame avec ‘pipe()’.
32. Utilisez Seaborn pour afficher un ‘barplot’ de l’âge moyen par département.
33. Sauvegarder le DataFrame nettoyé dans un nouveau fichier Excel nommé ‘employes_nettoyé.xlsx’.
34. Créer un histogramme de la répartition des âges.
35. Créer un graphique à barres montrant le nombre d’employés par département.
36. Créer un camembert (pie chart) montrant la répartition hommes/femmes.
37. Créer un boxplot des salaires par département.
38. Générer une heatmap de corrélation entre les colonnes numériques (âge, salaire, performance, télétravail, ancienneté).
39. Tracer une courbe de l’évolution des embauches par année.
40. Créer un graphique combiné (barres + ligne) montrant le salaire moyen et la performance moyenne par département.
41. Visualiser la distribution des salaires avec un KDE plot (seaborn).
42. Tracer une carte thermique du nombre d’employés par pays et par sexe (tableau croisé sous forme de heatmap).
43. Générer un scatter plot entre l’âge et le salaire, en coloriant par note de performance.

Exercice2

Vous disposez d’un fichier Excel nommé ‘ecommerce_transactions.xlsx’ contenant 10 000 transactions effectuées par des clients d’une plateforme e-commerce internationale. Chaque ligne représente une commande passée par un client. Votre objectif est d’explorer, nettoyer, analyser et visualiser les données pour en extraire des informations utiles.

Consignes globales :

1. Charger les données dans un DataFrame Pandas et afficher un aperçu général du jeu de données.

2. Afficher les dimensions du DataFrame et le type de chaque colonne.

3. Identifier les colonnes contenant des valeurs manquantes et proposer une stratégie adaptée pour les traiter (ex : selon la catégorie ou la méthode de paiement).

4. Supprimer les doublons éventuels.

5. Créer une colonne ‘Année-Mois’ à partir de la colonne ‘Date’ pour faciliter l’analyse temporelle.

6. Afficher les 5 pays générant le plus de chiffre d’affaires total.

7. Calculer le chiffre d’affaires total par catégorie de produits.

8. Identifier les marques les plus vendus (en quantité) dans chaque catégorie.

9. Afficher les méthodes de paiement les plus utilisées par pays.

10. Déterminer la dépense moyenne par client et afficher les 10 plus gros clients.

11. Calculer la note moyenne par catégorie de marque et par pays.

12. Identifier les commandes avec une note manquante et déterminer s’il existe un schéma selon la catégorie ou le pays.

13. Utiliser NumPy pour calculer la moyenne, la médiane, l’écart-type et les percentiles des montants totaux.

14. Créer une colonne ‘Client fidèle’ : True si le client a effectué plus de 5 commandes.

15. Créer un graphique à barres montrant le chiffre d’affaires mensuel total.

16. Créer un pie chart montrant la répartition des ventes par catégorie.

17. Créer un boxplot comparant le montant des commandes par méthode de paiement.

18. Générer un heatmap des notes moyennes des clients par pays et par catégorie.

19. Créer un scatter plot entre la quantité et le montant total pour détecter des anomalies.

20. Sauvegarder le DataFrame nettoyé dans un fichier Excel nommé ‘ecommerce_transactions_clean.xlsx’.