TD 1 : Faire parler les données

Exercice 1 : Analyse de la fréquentation de parcs d'attraction

Vous disposez des données annuelles ci-dessous sur le nombre de visiteurs de deux parcs d'attractions (source `https://queue-times.com/fr/parks/9/attendances` et `https://queue-times.com/fr/parks/4/attendances`).

Tableau 1 : Fréquentation annuelle du *Parc Astérix*

Année	Nombre de visiteurs
2023	2 815 000
2022	2 632 000
2021	1 300 000
2020	1 163 000
2019	2 326 000
2018	2 174 000
2017	2 000 000
2016	1 850 000
2015	1 850 000
2014	1 800 000

Tableau 2 : Fréquentation annuelle de *Disneyland Paris*

Année	Nombre de visiteurs
2023	10 400 000
2022	9 930 000
2021	3 500 000
2020	2 620 000
2019	9 745 000
2018	9 843 000
2017	9 660 000
2016	8 400 000
2015	9 790 000
2014	9 940 000

Partie 1 : Statistiques descriptives

1. Pour éviter des calculs trop pénibles, commencez par changer l'unité utilisée pour compter les visiteurs en millions et arrondissez à un chiffre après la virgule. Par exemple s'il y a eu 1 588 433 visiteurs, nous arrondirons à 1.6 millions.

2. Calculez la moyenne, la médiane et la variance du nombre de visiteurs pour *Disneyland Paris* et pour le *Parc Astérix* sur la période 2014-2023. Que nous révèle cette analyse ?

Partie 2 : Visualisation des données

1. Réalisez un histogramme du nombre de visiteurs pour chacun des parcs (séparément) sur la période 2014-2023 en utilisant comme bacs pour le parc Astérix :

   - $[0, 1($,
   - $[1; 1.5($,
   - $[1.5; 2($,
   - $[2; 3($,
   - $[3; +\infty($

   et pour Disneyland Paris :

   - $[0; 2($,
   - $[2; 8($,
   - $[8; 9($,
   - $[9; 10($,
   - $[10; +\infty($

2. Qu'avez-vous utilisé pour l'axe vertical (axe des y) de vos histogrammes à la question précédente? Si vous avez utilisé le nombre de points de données tombant dans chaque bac, refaites les graphes en utilisant le principe de l'aire vu en cours (hauteur d'un bac = nombre de points dans le bac divisé par (nombre de points total $\times$ largeur du bac). Si vous aviez déjà utilisé le principe de l'aire, refaites l'histogramme en utilisant comme hauteur le nombre de points de données tombant dans chaque bac. Quelle version est la plus utile ?

3. Superposez la moyenne et la médiane sur vos histogrammes et commentez.

4. Tracez un graphe linéaire (*line plot*) représentant le nombre de visiteurs pour chaque parc en fonction de l'année. Quelle conclusion pouvez-vous tirer en ce qui concerne les tendances des deux parcs au fil du temps, notamment pendant les années 2020-2021 ?

5. Réalisez un graphe de dispersion (*scatter plot*) avec le nombre de visiteurs de *Parc Astérix* sur l'axe des x et le nombre de visiteurs de *EuroDisney* sur l'axe des y. Commentez.

Partie 3 : Manipulation de données tabulées

Pour pouvoir réaliser les graphiques ci-dessus automatiquement avec la bibliothèque Python `seaborn`, il est utile de structurer les données de manière appropriée.

1. Expliquez comment vous pouvez fusionner les deux tableaux de données (*Disneyland Paris* et *Parc Astérix*) en un seul tableau où chaque ligne représente une année, et où les colonnes **Parc** et **Nombre de visiteurs** indiquent respectivement le parc (EuroDisney ou Parc Astérix) et le nombre de visiteurs correspondant.

2. Pourquoi cette structure de données est-elle plus adaptée pour réaliser des visualisations avec `seaborn` par rapport à une approche par *jointure*, qui aurait deux colonnes séparées pour le nombre de visiteurs d'*EuroDisney* et de *Parc Astérix* ?

3. Une fois que vous avez le bon tableau, quel serait le code seaborn à utiliser pour générer :

   - L'histogramme des visiteurs avec les moyennes et médianes superposées ?
   - Le *line plot* des visiteurs en fonction du temps pour les deux parcs ?
   - Le graphique de corrélation entre les visiteurs des deux parcs ?