

# Questionnaire Pratique en Analyse de Données

Formation Data Analysis - Séance 2

10 février 2025

## Instructions

- Ce questionnaire se concentre sur l'utilisation des bibliothèques Python **Pandas**, **Matplotlib** et **Seaborn** pour l'analyse de données.
- Écrivez des scripts pour chaque question et testez-les sur un dataset réel (`Titanic.csv`).
- Soyez créatifs et explorez plusieurs approches pour chaque problème.

## 1 Manipulation des Données avec Pandas

1. **Chargement des données** : Écrire un script Python pour importer un fichier CSV, afficher les **5 premières lignes** et obtenir des informations générales sur le dataset.
2. **Filtrage** : Sélectionner et afficher uniquement les lignes où la colonne **Age** est **supérieure à 30 ans**.
3. **Gestion des valeurs manquantes** : Remplacer les valeurs NaN d'une colonne numérique par sa **moyenne**.
4. **Création de nouvelles colonnes** : Ajouter une colonne **TailleFamille** qui est la somme des colonnes **SibSp** et **Parch**.
5. **Tri des données** : Trier le dataset selon l'âge en ordre croissant et afficher les **10 premières lignes**.

## 2 Statistiques Descriptives et Exploration

1. Calculer la **moyenne**, la **médiane** et l'**écart-type** d'une colonne numérique.

2. Trouver la **valeur la plus fréquente (mode)** d'une colonne catégorielle comme `Pclass` ou `Sex`.
3. Détecter les **valeurs aberrantes** d'une colonne avec la règle des **1.5 IQR** et les afficher.
4. Calculer la **corrélation** entre deux variables numériques et interpréter le résultat.
5. Générer un **tableau croisé (pivot\_table)** montrant le taux de survie (`Survived`) en fonction du sexe et de la classe (`Pclass`).

### 3 Visualisation des Données avec Matplotlib & Seaborn

1. Afficher un **histogramme** de la répartition des âges dans un dataset.
2. Générer un **boxplot** des âges selon la classe (`Pclass`) pour identifier les écarts de répartition.
3. Créer un **countplot** avec Seaborn pour afficher le nombre de passagers par classe et par sexe.
4. Tracer une **matrice de corrélation (heatmap)** avec Seaborn) pour visualiser les relations entre les variables numériques.
5. Générer un **diagramme en barres** montrant la proportion de survivants (`Survived`) en fonction du sexe.

### 4 Manipulation Avancée des Données

1. Sélectionner et afficher uniquement les colonnes `Nom`, `Âge` et `Survived`.
2. Utiliser `groupby()` pour afficher l'**âge moyen des passagers par classe**.
3. Fusionner deux datasets ayant une colonne commune `ID` et afficher les **5 premières lignes** du résultat.
4. Filtrer le dataset pour afficher uniquement les passagers de **1ère classe ayant plus de 40 ans**.
5. Créer une **nouvelle colonne Catégorie d'Âge** (Enfant, Adulte, Senior) en fonction de la variable `Âge`.

## 5 Challenge Créatif

### Défi Final

**écrivez un script pour y répondre en utilisant Pandas et Seaborn !** *Quel est l'impact de la classe et du sexe sur le taux de survie dans le Titanic ? - Montrez une analyse avec visualisation.*

Bonne exploration et amusez-vous avec les données !