**Projet : Analyse d'une base de données de films**

**Objectif :**

L'objectif de ce projet est de lire un fichier CSV contenant des informations sur des films (titre, genre, année de sortie, score IMDb, etc.), puis d'analyser ces données pour obtenir des informations sur les films les mieux notés, les genres les plus populaires, et les tendances de production par année.

**Outils utilisés :**

* Python
* Pandas (pour la manipulation des données)
* Matplotlib et Seaborn (pour la visualisation)
* Jupyter Notebook (ou un IDE comme VS Code)

**Missions :**

**Mission 1 : Chargement et exploration des données**

**Objectif :** Charger et explorer le fichier CSV pour comprendre sa structure et les données qu'il contient.

1. **Chargement des données** :
   * Utiliser Pandas pour charger le fichier CSV contenant les informations sur les films.

python

Copier le code

import pandas as pd

# Charger les données à partir du fichier CSV

df = pd.read\_csv('movies.csv')

# Afficher les premières lignes du fichier

print(df.head())

1. **Exploration des données** :
   * Examiner les colonnes disponibles, les types de données et vérifier la présence de valeurs manquantes.

python

Copier le code

# Afficher les informations sur le DataFrame

print(df.info())

# Résumer les statistiques descriptives

print(df.describe())

# Vérifier les valeurs manquantes

print(df.isnull().sum())

**Mission 2 : Nettoyage des données**

**Objectif :** Nettoyer le dataset pour enlever les erreurs et préparer les données pour l'analyse.

1. **Gestion des valeurs manquantes** :
   * Gérer les valeurs manquantes (par exemple, en les supprimant ou en les remplaçant par des valeurs appropriées).

python

Copier le code

# Supprimer les lignes avec des valeurs manquantes dans les colonnes essentielles (par ex. titre, score IMDb)

df = df.dropna(subset=['titre', 'imdb\_score'])

1. **Conversion des types de données** :
   * S'assurer que les colonnes comme l'année de sortie sont bien au bon format (int).

python

Copier le code

# Convertir l'année de sortie en entier

df['année\_sortie'] = pd.to\_numeric(df['année\_sortie'], errors='coerce')

**Mission 3 : Analyse exploratoire des données**

**Objectif :** Explorer les données pour identifier des insights simples.

1. **Top 10 des films les mieux notés** :
   * Identifier les 10 films avec le score IMDb le plus élevé.

python

Copier le code

# Top 10 des films les mieux notés

top\_10\_films = df.nlargest(10, 'imdb\_score')

print(top\_10\_films[['titre', 'imdb\_score']])

1. **Distribution des films par genre** :
   * Compter combien de films appartiennent à chaque genre.

python

Copier le code

# Compter le nombre de films par genre

films\_par\_genre = df['genre'].value\_counts()

print(films\_par\_genre)

1. **Tendances de production par année** :
   * Analyser combien de films ont été produits chaque année.

python

Copier le code

# Films produits par année

films\_par\_année = df.groupby('année\_sortie')['titre'].count()

print(films\_par\_année)

**Mission 4 : Visualisation des données**

**Objectif :** Créer des graphiques pour illustrer les insights trouvés.

1. **Bar chart des genres les plus populaires** :
   * Visualiser le nombre de films par genre dans un graphique en barres.

python

Copier le code

import matplotlib.pyplot as plt

import seaborn as sns

# Bar chart des genres les plus populaires

plt.figure(figsize=(10,6))

sns.barplot(x=films\_par\_genre.index, y=films\_par\_genre.values)

plt.xticks(rotation=45)

plt.title('Nombre de films par genre')

plt.show()

1. **Évolution du nombre de films par année** :
   * Visualiser la tendance de production de films au fil des années avec un graphique linéaire.

python

Copier le code

# Graphique de la production de films par année

plt.figure(figsize=(10,6))

films\_par\_année.plot(kind='line')

plt.title('Nombre de films produits par année')

plt.xlabel('Année')

plt.ylabel('Nombre de films')

plt.show()

1. **Top 10 des films les mieux notés** :
   * Afficher les scores IMDb des 10 meilleurs films dans un graphique en barres.

python

Copier le code

# Bar chart des top 10 films les mieux notés

plt.figure(figsize=(10,6))

sns.barplot(x='imdb\_score', y='titre', data=top\_10\_films)

plt.title('Top 10 des films avec les meilleurs scores IMDb')

plt.show()

**Mission 5 : Rapport final**

**Objectif :** Synthétiser les résultats dans un rapport simple.

1. **Résumé des découvertes** :
   * Présenter les principaux insights :
     + Le genre de film le plus populaire.
     + Le top 10 des films les mieux notés.
     + L’évolution du nombre de films produits par année.
2. **Prochaines étapes** :
   * Proposer des pistes pour approfondir l'analyse (par exemple, analyser la corrélation entre les genres et les scores IMDb, ou l'influence des réalisateurs sur les notes).

**Résumé du projet :**

Le projet consiste à analyser une base de données de films en Python, à utiliser des techniques de manipulation de données avec **Pandas**, à visualiser des informations avec **Matplotlib** et **Seaborn**, et à tirer des insights sur les films les mieux notés, les genres les plus populaires et les tendances de production.