 **Télécharger le dataset MovieLens :**

* Accède à MovieLens Dataset et choisis une version adaptée à ton projet (le fichier 100k est parfait pour débuter).
* Télécharge les données et stocke-les dans ton dossier /data.

 **Analyse exploratoire des données :**

* Charge le dataset avec Pandas :

python

import pandas as pd

ratings = pd.read\_csv("data/ml-100k/u.data", sep="\t", names=["user\_id", "item\_id", "rating", "timestamp"])

movies = pd.read\_csv("data/ml-100k/u.item", sep="|", encoding="latin-1", names=["item\_id", "title", "release\_date", "video\_release\_date", "imdb\_url"] + [f"genre\_{i}" for i in range(19)])

* Inspecte les données :

python

print(ratings.head())

print(movies.head())

print(ratings.info())

print(movies.info())

 **Nettoyage des données :**

* **Supprime les colonnes inutiles :** Retire timestamp, video\_release\_date, etc.
* **Traite les valeurs manquantes :** Vérifie si des champs contiennent des NaN et décide de les supprimer ou de les remplir.
* **Formatage des types de données :** Assure-toi que les types sont corrects (datetime, int, etc.).

 **Structuration et exportation dans une base relationnelle :**

* Configure une base de données SQLite ou PostgreSQL.
* Crée les tables nécessaires :

sql

CREATE TABLE ratings (

user\_id INTEGER,

item\_id INTEGER,

rating INTEGER,

timestamp INTEGER

);

CREATE TABLE movies (

item\_id INTEGER,

title TEXT,

release\_date TEXT,

imdb\_url TEXT

);

* Insère les données dans la base :

python

import sqlite3

conn = sqlite3.connect("data/movielens.db")

ratings.to\_sql("ratings", conn, if\_exists="replace", index=False)

movies.to\_sql("movies", conn, if\_exists="replace", index=False)

conn.close()

 **Tests de la base de données :**

* Vérifie si les données sont bien accessibles :

python

conn = sqlite3.connect("data/movielens.db")

df = pd.read\_sql("SELECT \* FROM ratings LIMIT 5", conn)

print(df)

conn.close()

 **Résumé de l'exploration :**

* Génère un rapport court des données :
  + Nombre d’utilisateurs, films, et évaluations.
  + Distribution des évaluations (graphiques Matplotlib/Seaborn).

Exemple :

python

import matplotlib.pyplot as plt

ratings["rating"].value\_counts().plot(kind="bar")

plt.title("Distribution des évaluations")

plt.xlabel("Évaluations")

plt.ylabel("Fréquence")

plt.show()