Mini Project 1. IMDb 2024 Movie Analysis Dashboard

Step 1: Scraping_Final

Need to import time, pandas, selenium, webdrivers,action,key

1. IMDb URL for movies released in 2024 - driver = webdriver.Chrome()
2. url=https://www.imdb.com/search/title/?title_type=feature&release_date=2024-01-01,2024-12-31
3. To find only action and animation and animation and crime and history one by one manually.
4. Open the URL
5. Allow the page to load (you can adjust the sleep time if necessary)
6. Define a function to click the "Load More" button
7. Keep clicking "Load More" until it's no longer available
8. Initialize lists to store the scraped data
9. movie_items – Store data
10. Create a DataFrame using the extracted data
11. Save the DataFrame to a CSV file (optional)

Step 2: Merge CSV

Need to import and Pandas.

1. Path to genre CSVs created in Script 1
   csv_folder = r"C:\Users\mugil\Project_imdb\Final Scrapping"
2. Read all CSV files in the folder
3. Merge them all into one DataFrame
4. Optional: Remove duplicates based on 'Title' and 'Genre' to keep multi-genre movies
5. Save merged dataset

After Completion - print("Merged dataset saved as 'merged5_imdb_2024.csv'")

Step 3: Data_Cleaning

Need to Import pandas, numpy, re

1. Load the dataset
2. Analysis the dataset by using – head, describe, info, df.isnull().sum()
3. --- Fill NaN values ---
4. Fill with a specific value (e.g., 0 for votes, mean/median for rating/duration)
5. Analysis the dataset by using – head, describe, info, df.isnull().sum()
6. --- Step 2: Convert 'Duration' to total minutes ---using function
7. ---Drop rows with missing data in essential columns ---
8. --- Fix data types ---

df.to_csv('new_cleaned_merged5_imdb_2024.csv', index=False)

Mini Project 1. IMDb 2024 Movie Analysis Dashboard

Step 4: Data Import to MYSQL DB

!pip install sqlalchemy pymysql pandas

Import pandas and create engine to connect SQL DB – (from sqlalchemy import create_engine)

1. Load the cleaned merged CSV
2. MySQL connection details
3. Create SQLAlchemy engine

engine = create_engine(f"mysql+pymysql://{username}:{password}@{host}:{port}/{database}")

4. Upload the DataFrame to the new table

Print the data - Data uploaded successfully to table 'imdb_2024_movies' in datab
ase 'project1_imdb'

To view Datas in DB

#select count(*) as sn from project1_imdb.imdb_2024_movies

#select genre from project1_imdb.imdb_2024_movies group by genre

Step 5: Creating Intractive Dashboards using Streamlit Using (MY SQL Work Bench)

import streamlit, pandas, sqlalchemy, matplotlib.pyplot, seaborn,plotly.express, time

1. pip install streamlit pandas matplotlib seaborn plotly mysql-connector-python sqlalchemy
2. Function to load MySQL data using SQLAlchemy engine
3. Function to load MySQL data using SQLAlchemy engine

3.1.1 Create an SQLAlchemy engine

3.1.2 Query to fetch data from the database

3.1.3 Use the engine to load data into a pandas DataFrame

df = load_mysql_data()

4. Check if data is loaded correctly
5. FILTERS and Sibebards
6. Apply filters using
7. VISUALIZATIONS

7.1. Top 10 Movies by Rating and Voting Count

7.2. Genre Distribution

7.3. Average Duration by Genre

7.4. Voting Trends by Genre

7.5. Rating Distribution

7.6. Genre-Based Rating Leaders

7.7. Most Popular Genres by Voting

7.8. Duration Extremes

7.9. Ratings by Genre (Heatmap)

7.10. Correlation: Ratings vs Votes

(OR) Step 5.1: Creating Intractive Dashboards using Streamlit Using (MY CSV)

import streamlit, pandas, matplotlib.pyplot, seaborn,plotly.express, time

1. pip install streamlit pandas matplotlib seaborn plotly mysql-connector-python sqlalchemy

2. Read the CSV an Load to dataframe

3. FILTERS and Sibebards

4. Apply filters using

5. VISUALIZATIONS

5.1. Top 10 Movies by Rating and Voting Count

5.2. Genre Distribution

5.3 Average Duration by Genre

5.4. Voting Trends by Genre

5.5. Rating Distribution

5.6. Genre-Based Rating Leaders

5.7. Most Popular Genres by Voting

5.8. Duration Extremes

5.9. Ratings by Genre (Heatmap)

5.10. Correlation: Ratings vs Votes

To get Ouput – Save the code in VS Code and run the following scripts

PS C:\Users\mugil> cd project_imdb\final_scrapping

PS C:\Users\mugil\project_imdb\final_scrapping> streamlit run 5_imdb_dashboard_mysql.py

You can now view your Streamlit app in your browser.

Local URL: http://localhost:8501

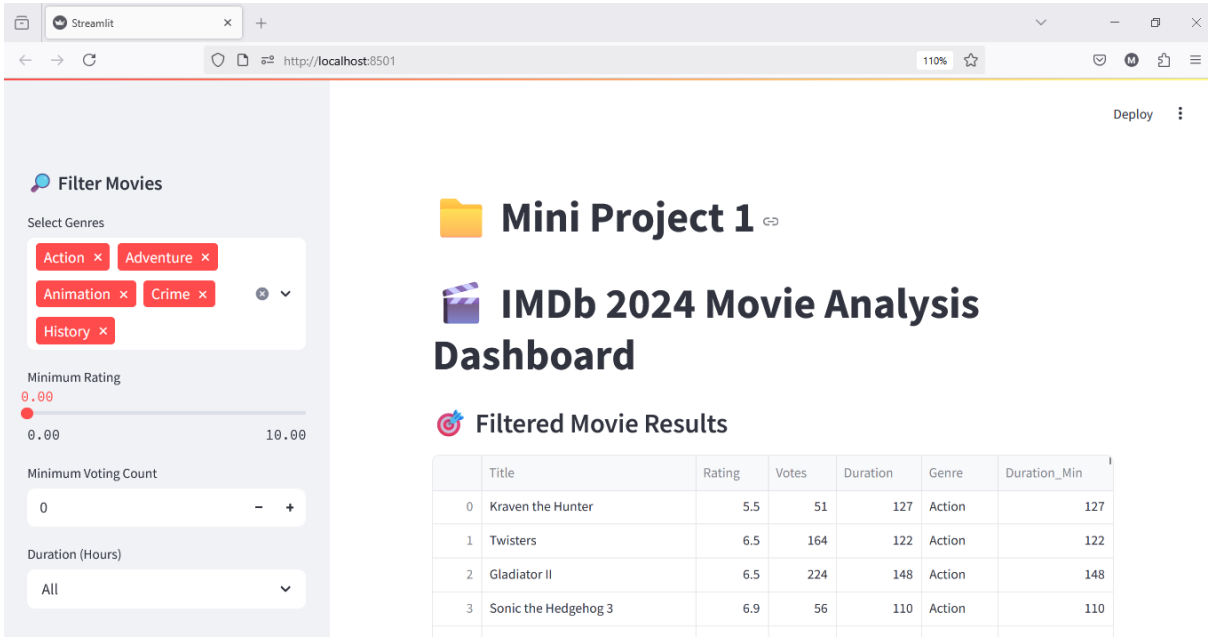
Network URL: http://192.168.0.153:8501

Output will be displayed in Brower contains interactive dashboard.

Screenshots are attached.

Mini Project 1. IMDb 2024 Movie Analysis Dashboard

Reference 1:



Reference 2: