**Project Documentation: IMDb Movie and TV Show Data Extraction and Visualization**

**Prepared by:** Muhammad Maroof Tahir

**1. Project Overview**

**Project Objective**

The primary objective of this project is to:

* **Extract** relevant movie and TV show data from IMDb.
* **Organize** the data in a structured format using pandas and CSV files.
* **Visualize** the insights through Matplotlib and Power BI.
* **Develop** a user-friendly dashboard to showcase findings interactively.

**Target Website**

* **Website**: IMDb (Internet Movie Database)
* **Data to be Extracted**:
  + Titles (Movies & TV Shows)
  + Directors and Cast
  + Release Dates
  + IMDb Ratings
  + Genres
  + User Reviews
  + Box Office/Budget (for movies)
  + Runtime, Languages, and Filming Locations
  + Trailer Links and Images

**Technologies Used**

* **Python**: Programming language for data extraction
* **pandas**: Data manipulation and CSV management
* **Matplotlib**: Data visualization
* **Power BI**: For creating an interactive dashboard
* **BeautifulSoup**: Web scraping tool to parse HTML

**2. Planning and Website Structure Analysis**

**IMDb Website Exploration**

The IMDb website is divided into several sections:

* **Top 250 Movies and TV Shows**: Popular lists including rankings, titles, ratings, and release years.
* **Genres and Crew Information**: Found on individual title pages.
* **Box Office** (for movies): Contains information on budget and box office earnings.

**Key Data Points for Scraping**

* **Movie/TV Show Titles**: Names of movies or TV shows.
* **Release Year**: The year a movie or TV show was released.
* **IMDb Rating**: Audience ratings available on IMDb.
* **Genres**: The type or category of the movie/TV show.
* **Director and Cast Details**: Names of the director and main actors.
* **Runtime**: Duration of the movie or TV show.
* **Box Office Data**: Budgets and earnings for movies.
* **Plot Summary**: A brief description of the story.

**3. Data Extraction**

**Web Scraping Approach**

* **Tools Used**: BeautifulSoup was used for scraping, as it provides flexibility in selecting specific HTML elements.
* **Steps**:
  1. **Send Requests**: Python requests library is used to access IMDb pages.
  2. **Parse HTML**: BeautifulSoup parses HTML to locate specific data elements like titles, genres, and ratings.
  3. **Data Structuring**: The scraped data is organized into pandas DataFrames for ease of use.

**Sample Python Code for Scraping**

python

Copy code

import requests

from bs4 import BeautifulSoup

import pandas as pd

# Example of scraping IMDb Top 250 movies

url = "https://www.imdb.com/chart/top"

response = requests.get(url)

soup = BeautifulSoup(response.content, 'html.parser')

# Extracting movie titles and ratings

titles = [title.text for title in soup.select('td.titleColumn a')]

ratings = [rating.text for rating in soup.select('td.imdbRating strong')]

# Create a pandas DataFrame

df = pd.DataFrame({

'Title': titles,

'Rating': ratings

})

# Save the data to a CSV file

df.to\_csv('imdb\_top\_250.csv', index=False)

**4. Data Organization and Modeling**

**Data Format**

* **Storage**: The extracted data is stored in CSV files for future analysis.
* **Data Manipulation**: pandas is used to clean, format, and normalize the data.

**Key Data Fields**

* **Movies Table**: Contains fields like movie\_id, title, release\_year, rating, and box\_office.
* **Genres Table**: Includes genre\_id, movie\_id, and genre\_name.
* **Cast Table**: Links movies to cast members with actor\_id, movie\_id, and role.

**5. Data Analysis and Visualization**

**Power BI Integration**

* **Data Loading**: The CSV files containing the scraped data are imported into Power BI.
* **Charts & Visualizations**:
  + **Bar Charts**: Display top-rated movies/TV shows based on IMDb ratings.
  + **Line Charts**: Track box office performance over time.
  + **Heatmaps**: Show the correlation between genres and ratings.

**Sample Visualizations:**

* **Top 10 Rated Movies by Genre**
* **Box Office Revenue over Time**
* **Average IMDb Ratings Across Genres**

**6. Final Steps and Dashboard Deployment**

**Data Cleaning and Optimization**

* The data is cleaned and formatted to ensure consistency (e.g., handling missing ratings or duplicate entries).
* Ensure all numeric fields (like ratings and box office earnings) are normalized for analysis.

**Dashboard Publishing**

* Once the visualizations are finalized, the Power BI dashboard is published for online access, making it shareable and interactive.

**7. Conclusion**

This project successfully extracts data from IMDb, organizes it into a structured format, and visualizes it using both Python and Power BI. The dashboard provides key insights into the data, helping users analyze movie and TV show trends, ratings, and box office performance efficiently.