Task -3-

Web Scraping for Moviemeter and IMDB

Prepared by:

- *Nada Amgad*
- *Ahmed Sayed*

Table of Contents

Required Libraries and Dependencies:

- pandas
- matplotlib
- re
- scrapy
- CrawlerProcess → from scrapy.crawler
- Selector → from scrapy.selector

Data Analysis Findings:

Movies with highest IMDB and Moviemeter

First 5 movies with highest moviemeter:

- The Shawshank Redemption
- The Godfather
- The Godfather: Part II
- The Green Mile
- Pulp Fiction

First 5 movies with highest IMDB:

- The Shawshank Redemption
- The Godfather
- The Lord Of The Rings: The Return Of The King
- 12 Angry Men6
- Schindler's List

Steps:

1. Sort the input column of the dataframe in descending order
2. Get the 1st "n" entries needed

Movies with lowest IMDB and Moviemeter

First 5 movies with lowest moviemeter:

- T Dial M for Murder
- Batman Begins
- The Great Dictator
- The Help
- Room

First 5 movies with lowest IMDB:

- The Grapes Of Wrath
- Fargo
- La Battaglia Di Algeri
- The Third man
- Prisoners

Steps:

Same steps as the previous points except that we pass **reverse=True** as an argument

1. Sort the input column of the dataframe in descending order
2. Get the 1st "n" entries needed

Most Found Genre

After making Some Analysis The most found Genre through the whole dataset is: **"Drama"**

Steps:

1. A list of all possible genres is created
2. Count the occurrence of every genre through the whole dataset
3. Append each tuple containing the genre and its count to a list
4. Get the most common genre from the list of tuples

Number of Movies released in each decade

Decade
1940 → 8
1950 → 10
1960 → 9
1970 → 10
1980 → 16
1990 → 32
2000 → 25
2010 → 22
2020 → 2

Steps:

1. Dividing the date column by 10
2. Converting it to integer
3. Multiplying it by 10 to get the accurate decade
4. Returning a dataframe grouped by decades and getting the count of each decade

Visualizations:

Bar chart for distribution of ratings Visualization

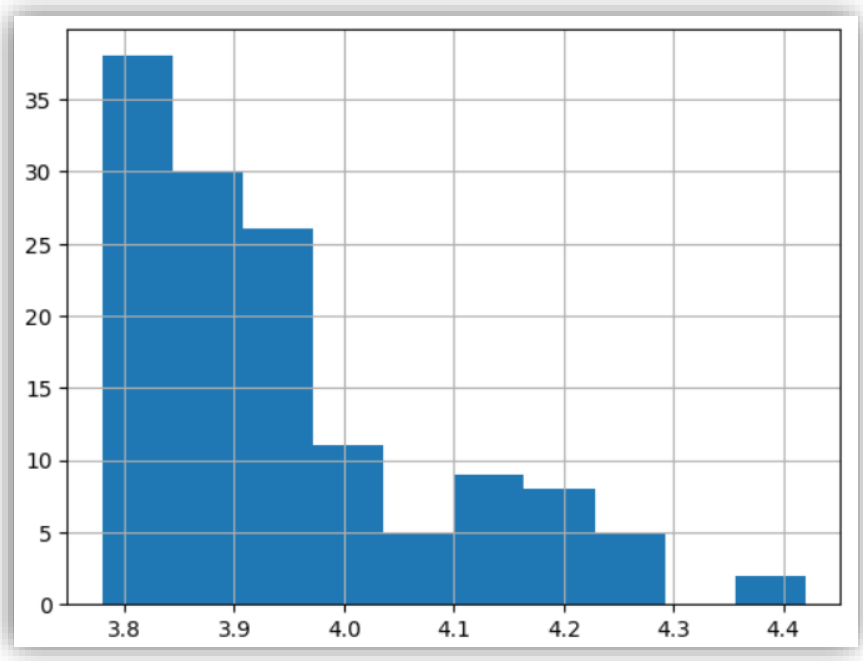

Figure 1: Rating Distribution

Bar chart for distribution of movie genres Visualization

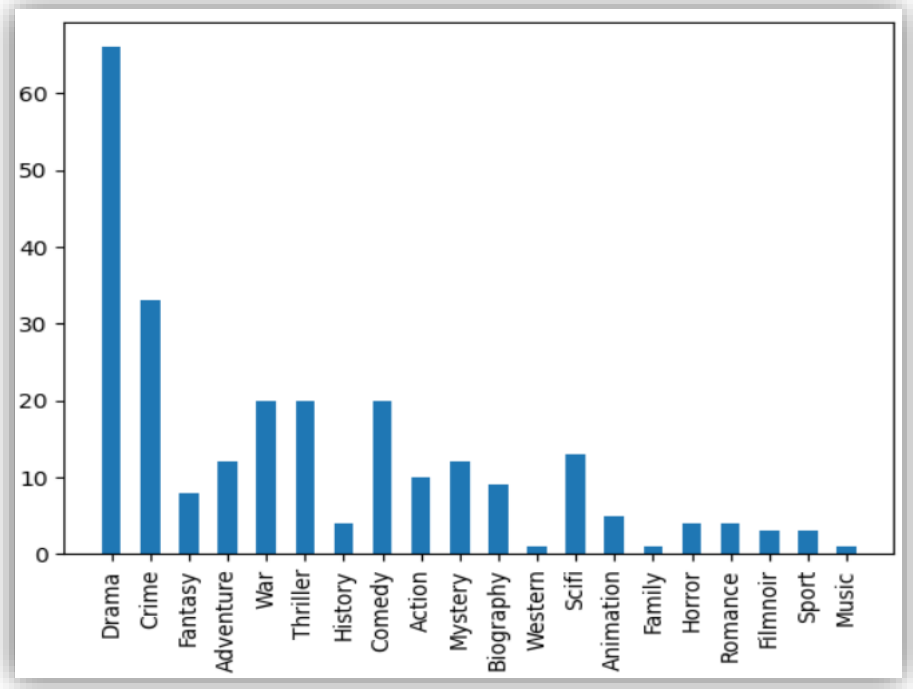

Figure 2: Movie Genres Distribution

Data Cleaning:

Case Consistent

The whole combined data set became case consistent → Lower Case.

Remove Padding

Removed left and right paddings if any.

Dropping Duplicates

Drop any row duplicates if found.

Finding Outliers Statistically

Checking for outliers but there were no outliers and it's not a good choice to remove anything in our case.

Cleaning Rating Column

Replace the commas in the ratings with decimal point to be easier in analysis.

Cleaning Duration Column

Extracting only the duration and converting it to integers