Minor Project Report — IMDb Movie Ratings Analyzer

Submitted by: Virendra Mahajan

Course: Data Science Minor Project

Tool Stack: Python, Requests, BeautifulSoup, Pandas, Matplotlib

---

1) Introduction

In this project, I developed a data analysis system that scrapes movie information from IMDb (Internet Movie Database) to study trends in ratings, genres, and directors. The project's goal was to transform unstructured web data into a clean, structured dataset and perform Exploratory Data Analysis (EDA) to discover meaningful insights.

IMDb is one of the most trusted sources for movie ratings, providing valuable information about audience preferences, popular genres, and director performance. Through web scraping and visualization, this project demonstrates how Data Science techniques can extract, process, and interpret real-world online data.

---

2) Objective

The main objectives of this project are:

To collect movie metadata and ratings from IMDb using Python and BeautifulSoup.

To clean and organize the data into a structured format suitable for analysis.

To perform Exploratory Data Analysis (EDA) to identify rating trends, genre popularity, and top directors.

To visualize results through graphs and statistical summaries.

Goal: To build a simple data-driven insight system for IMDb movie ratings using real web data.

---

3) Methodology

Tools and Libraries Used

Python 3.12

Requests → For sending HTTP requests and fetching IMDb pages.

BeautifulSoup (bs4) → For parsing and extracting HTML data.

Pandas & NumPy → For data manipulation and cleaning.

Matplotlib → For data visualization (histograms, boxplots, scatter plots).

Dataset

Data was collected from IMDb title pages using 15 movie IDs (e.g., tt0111161, tt0468569).

Total titles scraped: 15

Key fields collected:

Title, Year, Rating, Votes, Genres, Runtime, Certificate, Directors

Steps Performed

Scraped IMDb movie data via Python script (scrape_imdb.py).

Cleaned and stored data in CSV format (imdb_movies.csv).

Conducted analysis and generated visualizations using eda_plots.py.

Exported key statistics and top results to summary.json and CSV files.

---

4) Implementation

Commands Executed

python src/scrape_imdb.py --ids data/raw/title_ids_sample.txt --out data/processed/imdb_movies.csv

python src/eda_plots.py --input data/processed/imdb_movies.csv

File Description

scrape_imdb.py:

Fetches each IMDb page by ID

Extracts fields like title, rating, votes, genres, and directors

Saves all results into data/processed/imdb_movies.csv

Creates reports/summary.json (mean rating, median, top genres)

eda_plots.py:

Loads the processed CSV

Plots rating distributions, ratings by genre and decade, and votes vs ratings

Generates tables: Top 10 Movies by Votes and Top Directors by Average Rating

Saves all plots to src/figures/ and summary tables to reports/

---

5) Results and Observations

Dataset Summary

Metric	Value
Total Movies Scraped	15
Mean Rating	8.9
Median Rating	8.8

Top Genres (from summary.json)

Drama

Action

Adventure

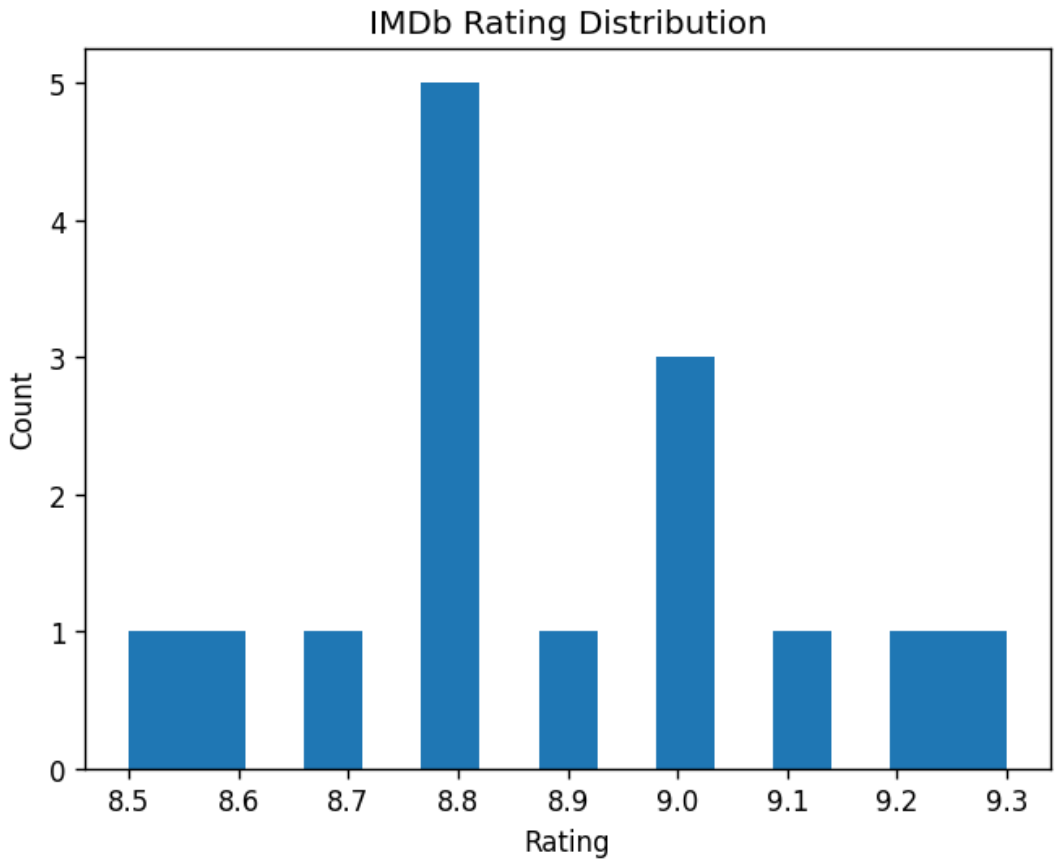
Crime

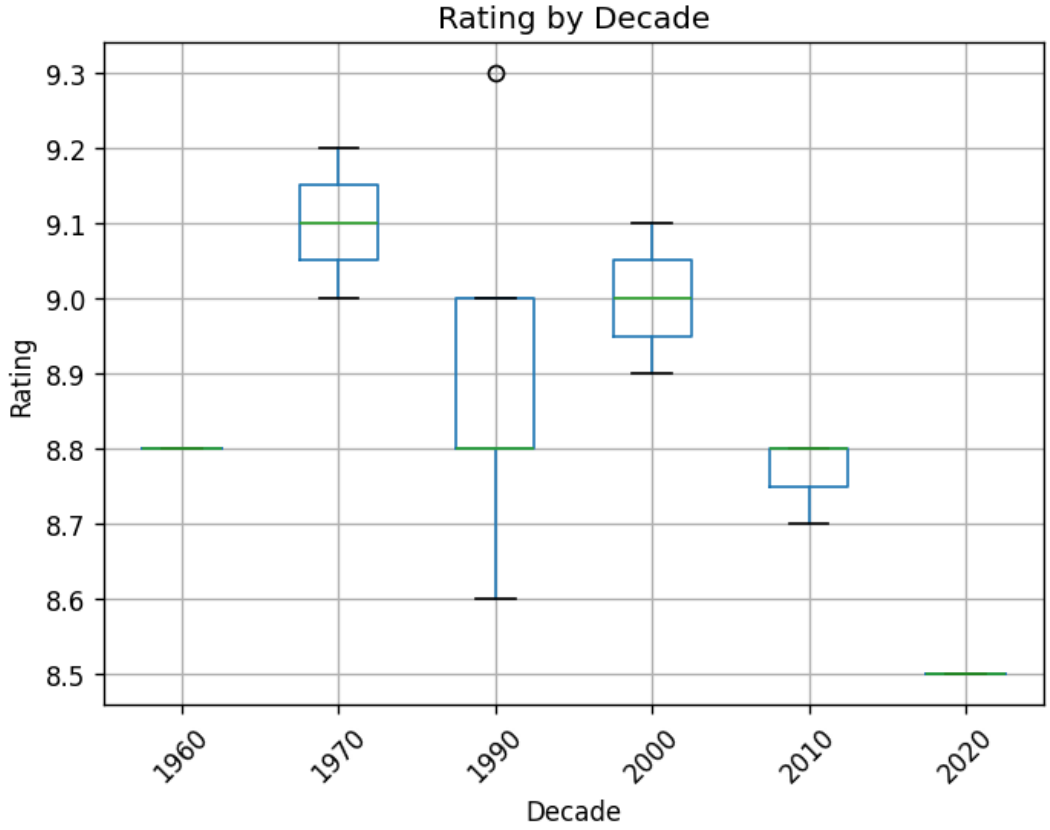
Sci-Fi

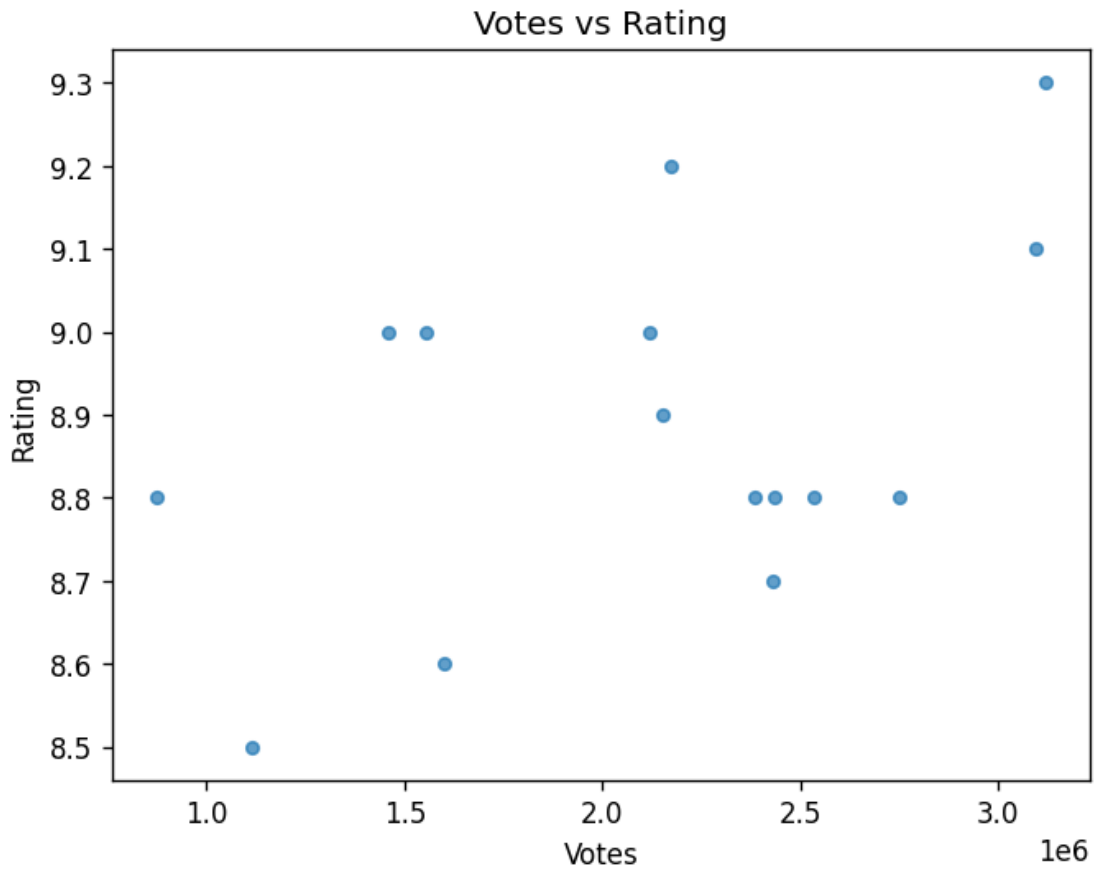
Top 10 Movies by Votes

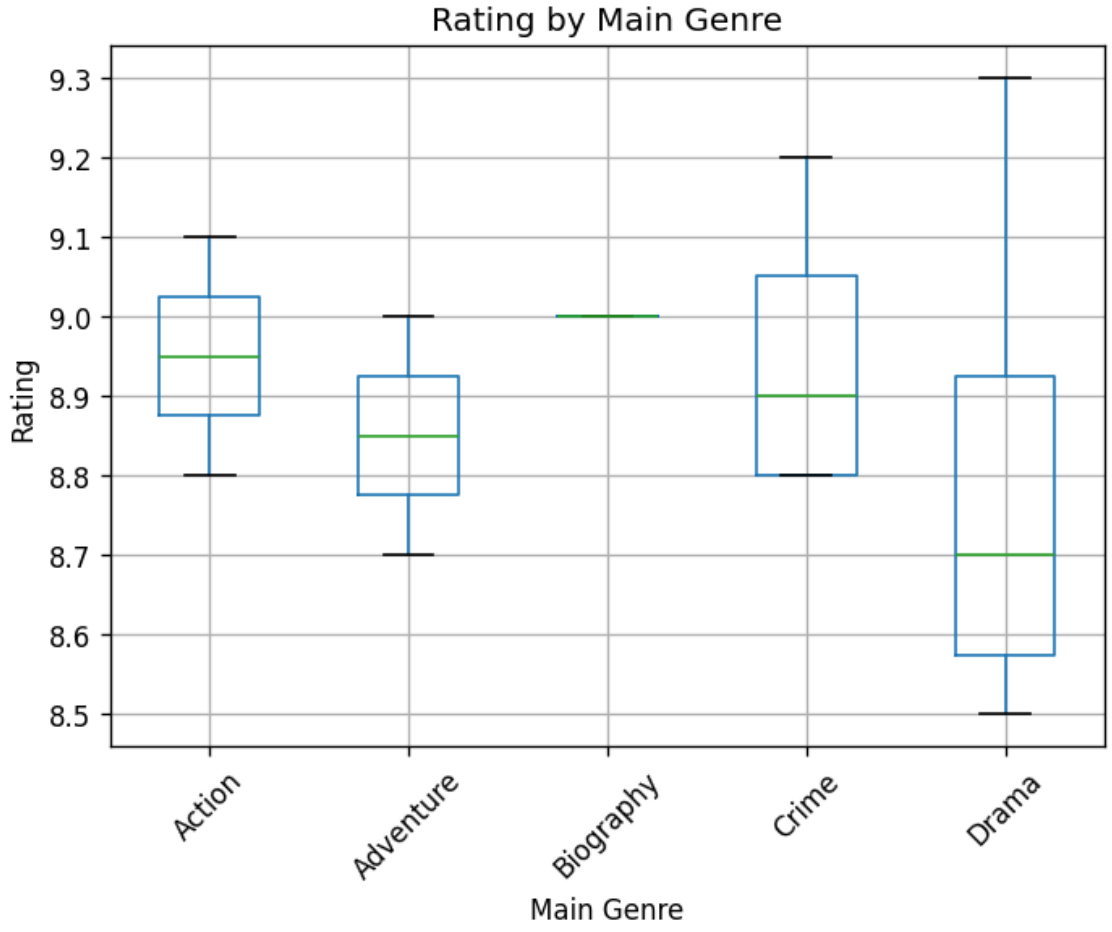
Rank	Movie	Year	Rating	Votes
1	The Shawshank Redemption	1994	9.3	3,118,227
2	The Dark Knight	2008	9.1	3,094,110
3	Inception	2010	8.8	2,748,918
4	Fight Club	1999	8.8	2,533,103
5	Forrest Gump	1994	8.8	2,435,468
6	Interstellar	2014	8.7	2,428,231
7	Pulp Fiction	1994	8.8	2,382,192
8	The Godfather	1972	9.2	2,174,427
9	The Lord of the Rings: The Fellowship of the Ring	2002	8.9	2,153,608
10	The Return of the King	2004	9.0	2,117,985

Visual Findings


IMDb Rating Distribution
Count
Rating
0
1
2
3
4
5
8.5
8.6
8.7
8.8
8.9
9.0
9.1
9.2
9.3


Rating by Decade
Rating
Decade
9.3
9.2
9.1
9.0
8.9
8.8
8.7
8.6
8.5
1960
1970
1990
2000
2010
2020


Votes vs Rating
Rating
9.3
9.2
9.1
9.0
8.9
8.8
8.7
8.6
8.5
1.0
1.5
2.0
2.5
3.0
Votes
1e6


Rating by Main Genre
Rating
9.3
9.2
9.1
9.0
8.9
8.8
8.7
8.6
8.5
Action
Adventure
Biography
Crime
Drama
Main Genre

Rating Distribution: Most ratings lie between 8.5–9.3.

Genre Boxplot: Drama and Action movies have consistently high ratings.

Votes vs Rating Scatter: More popular movies (higher votes) tend to have higher ratings.

Rating by Decade: 1990s and 2000s show strong consistency in high-rated titles.

Director Insights

Top Directors (min 2 films):

Director	No. of Films	Avg Rating
Christopher Nolan	3	8.9
Peter Jackson	2	8.95

---

6) Conclusion

This project successfully demonstrated how Data Science can be applied to extract and analyze online data from IMDb.
Using Python's scraping and visualization tools, I built an end-to-end mini analytics workflow that identifies rating patterns, top genres, and influential directors.

Key Learnings:

Gained hands-on experience in web scraping and HTML parsing.

Learned how to clean, transform, and visualize data effectively.

Developed insight into audience trends in movies.

Future Enhancements:

Scrape more movies (Top 250 list).

Integrate OMDb API for safer, faster data access.

Add NLP analysis for user reviews (sentiment analysis).

---

7) References

IMDb Official Site: https://www.imdb.com

Python Documentation: https://docs.python.org/3/

BeautifulSoup Docs: https://www.crummy.com/software/BeautifulSoup/

Pandas Documentation: https://pandas.pydata.org/docs/

📎 Attachments

data/processed/imdb_movies.csv – Final dataset

reports/summary.json – Summary statistics

reports/top10_by_votes.csv – Top movies

reports/top_directors_avg_rating.csv – Director rankings

src/figures/ – All generated plots