IMDb 2024 Movie Dashboard

Author
Dinesh Kumar D

Project Overview

This project is an end-to-end movie analysis pipeline using IMDb data scraped in 2024.
It includes data scraping, cleaning, transformation, exploratory data analysis (EDA),
SQL integration, and a fully interactive Streamlit dashboard.

Project Structure

- data/csv/: Raw and cleaned CSV files (per genre and final merged data)

- notebooks/: EDA and data visualization notebook

- scripts/: Python scripts for scraping, merging, cleaning, and database loading

- screenshots/: Visual proof of dashboard features

- app.py: The Streamlit dashboard app

- requirements.txt: Python dependencies

- README.md: Project documentation

- project_report.pdf: This generated report

Project Goals

- Scrape IMDb data across multiple genres such as action, comedy, drama, romance, and horror.

- Clean, merge, and transform the data into a usable structure for analysis and dashboarding.

- Perform exploratory data analysis on various metrics like IMDb ratings, number of votes, and movie durations.

- Store the processed data in both CSV and MySQL databases.

- Build an interactive dashboard using Streamlit to visualize and filter movie trends.

Key Features

IMDb Scraper:

- Uses requests, selenium, and BeautifulSoup to scrape movie metadata.

Data Cleaning:

- Handles missing values, formats data types, and categorizes durations.

Exploratory Data Analysis:

- Histograms, bar charts, heatmaps, scatter plots, and box plots for insights.

Streamlit Dashboard:

- Real-time filters, top-10 ratings, genre charts, duration patterns, voting insights.

MySQL Integration:

- Loads cleaned data into MySQL and includes SQL scripts for analysis.

Technologies Used

- Programming Language: Python

- Libraries: pandas, numpy, matplotlib, seaborn, requests, selenium, BeautifulSoup

- Dashboard: Streamlit

- Database: MySQL with mysql-connector-python

- Tools: Jupyter Notebook, Markdown

How to Run the Project

1. **Clone the repository:**

```
git clone https://github.com/yourusername/imdb_2024_scraper_dashboard.git
```

2. **Install dependencies:**

```
pip install -r requirements.txt
```

3. **Run scraper:**

```
python scripts/imdb_scraper.py (change genres manually)
```

4. **Merge CSVs:**

```
python scripts/merge_csv_files.py
```

5. **Clean data:**

```
python scripts/movie_data_cleaner.py
```

6. **EDA:**

Run imdb_eda.ipynb and export eda_cleaned.csv

7. **Load to MySQL:**

```
python scripts/database/load_and_test_mysql.py
```

8. **SQL Analysis:**

Execute scripts/database/sql/all_queries.sql

9. **Launch Dashboard:**

```
streamlit run app.py
```

Deliverables

- Cleaned genre-wise and merged datasets (CSV)

- MySQL database with `movies_2024` table

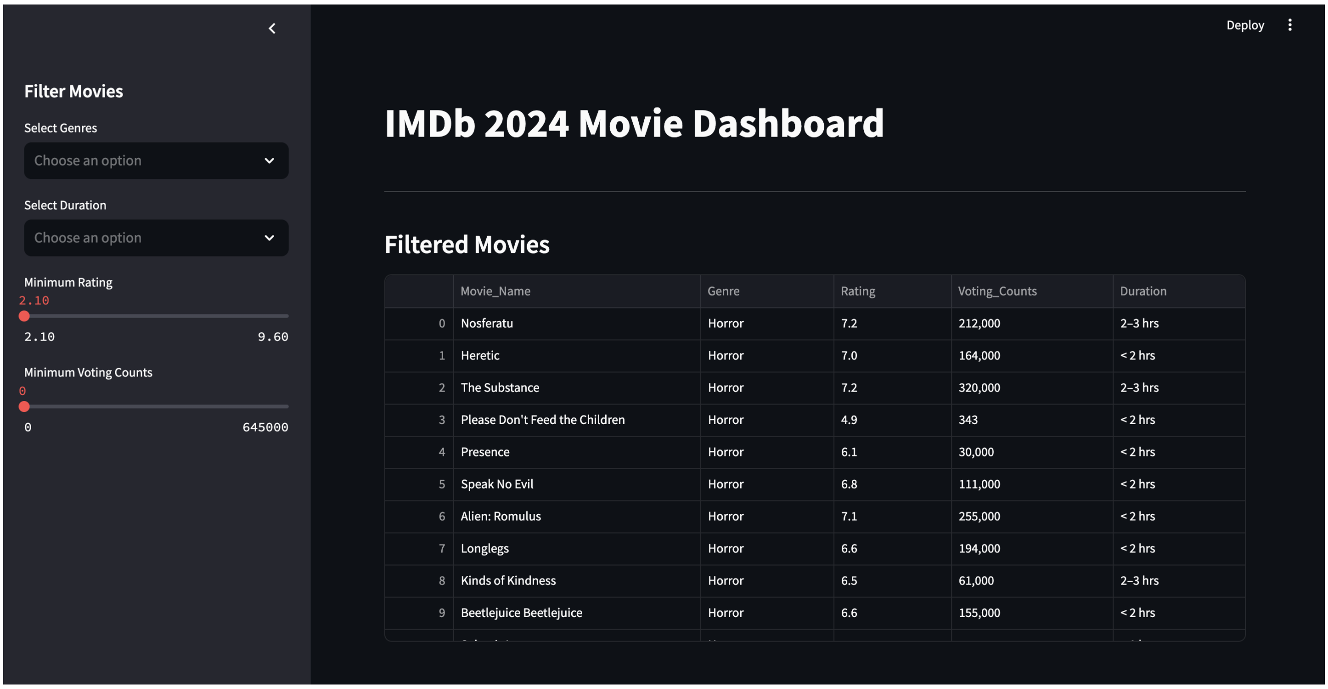
- EDA visualizations and summaries

- Streamlit dashboard with interactivity

- Professional PDF project report

- Complete deployment-ready codebase

Dashboard Screenshots



Filter Movies

Select Genres: Choose an option

Select Duration: Choose an option

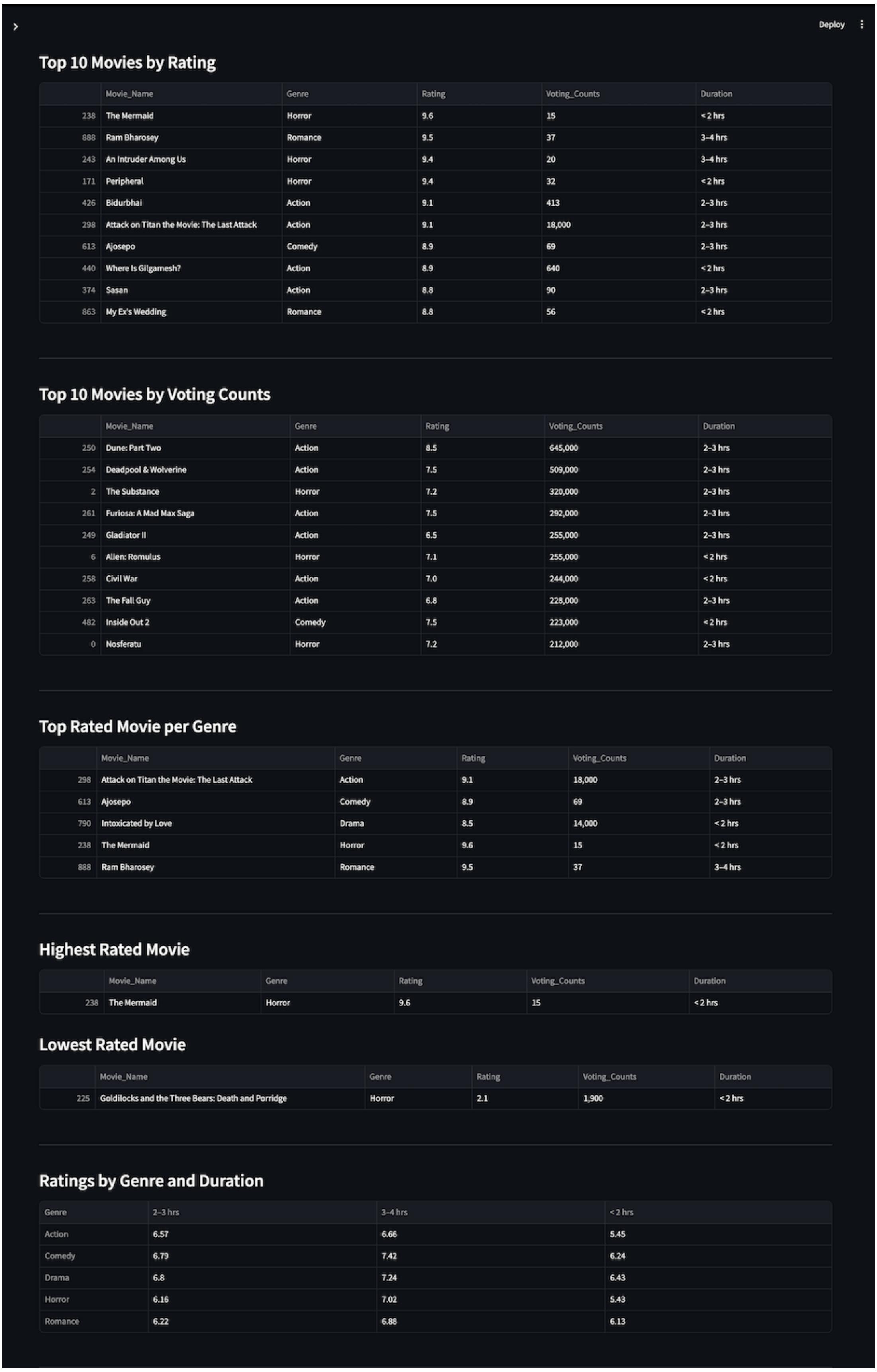
Minimum Rating: 2.10 (2.10 – 9.60)

Minimum Voting Counts: 0 (0 – 645000)

IMDb 2024 Movie Dashboard

Filtered Movies

	Movie_Name	Genre	Rating	Voting_Counts	Duration
0	Nosferatu	Horror	7.2	212,000	2–3 hrs
1	Heretic	Horror	7.0	164,000	< 2 hrs
2	The Substance	Horror	7.2	320,000	2–3 hrs
3	Please Don't Feed the Children	Horror	4.9	343	< 2 hrs
4	Presence	Horror	6.1	30,000	< 2 hrs
5	Speak No Evil	Horror	6.8	111,000	< 2 hrs
6	Alien: Romulus	Horror	7.1	255,000	< 2 hrs
7	Longlegs	Horror	6.6	194,000	< 2 hrs
8	Kinds of Kindness	Horror	6.5	61,000	2–3 hrs
9	Beetlejuice Beetlejuice	Horror	6.6	155,000	< 2 hrs



Top 10 Movies by Rating

	Movie_Name	Genre	Rating	Voting_Counts	Duration
238	The Mermaid	Horror	9.6	15	< 2 hrs
888	Ram Bharosey	Romance	9.5	37	3–4 hrs
243	An Intruder Among Us	Horror	9.4	20	3–4 hrs
171	Peripheral	Horror	9.4	32	< 2 hrs
426	Bidurbhai	Action	9.1	413	2–3 hrs
298	Attack on Titan the Movie: The Last Attack	Action	9.1	18,000	2–3 hrs
613	Ajosepo	Comedy	8.9	69	2–3 hrs
440	Where Is Gilgamesh?	Action	8.9	640	< 2 hrs
374	Sasan	Action	8.8	90	2–3 hrs
863	My Ex's Wedding	Romance	8.8	56	< 2 hrs

Top 10 Movies by Voting Counts

	Movie_Name	Genre	Rating	Voting_Counts	Duration
250	Dune: Part Two	Action	8.5	645,000	2–3 hrs
254	Deadpool & Wolverine	Action	7.5	509,000	2–3 hrs
2	The Substance	Horror	7.2	320,000	2–3 hrs
261	Furiosa: A Mad Max Saga	Action	7.5	292,000	2–3 hrs
249	Gladiator II	Action	6.5	255,000	2–3 hrs
6	Alien: Romulus	Horror	7.1	255,000	< 2 hrs
258	Civil War	Action	7.0	244,000	< 2 hrs
263	The Fall Guy	Action	6.8	228,000	2–3 hrs
482	Inside Out 2	Comedy	7.5	223,000	< 2 hrs
0	Nosferatu	Horror	7.2	212,000	2–3 hrs

Top Rated Movie per Genre

	Movie_Name	Genre	Rating	Voting_Counts	Duration
298	Attack on Titan the Movie: The Last Attack	Action	9.1	18,000	2–3 hrs
613	Ajosepo	Comedy	8.9	69	2–3 hrs
790	Intoxicated by Love	Drama	8.5	14,000	< 2 hrs
238	The Mermaid	Horror	9.6	15	< 2 hrs
888	Ram Bharosey	Romance	9.5	37	3–4 hrs

Highest Rated Movie

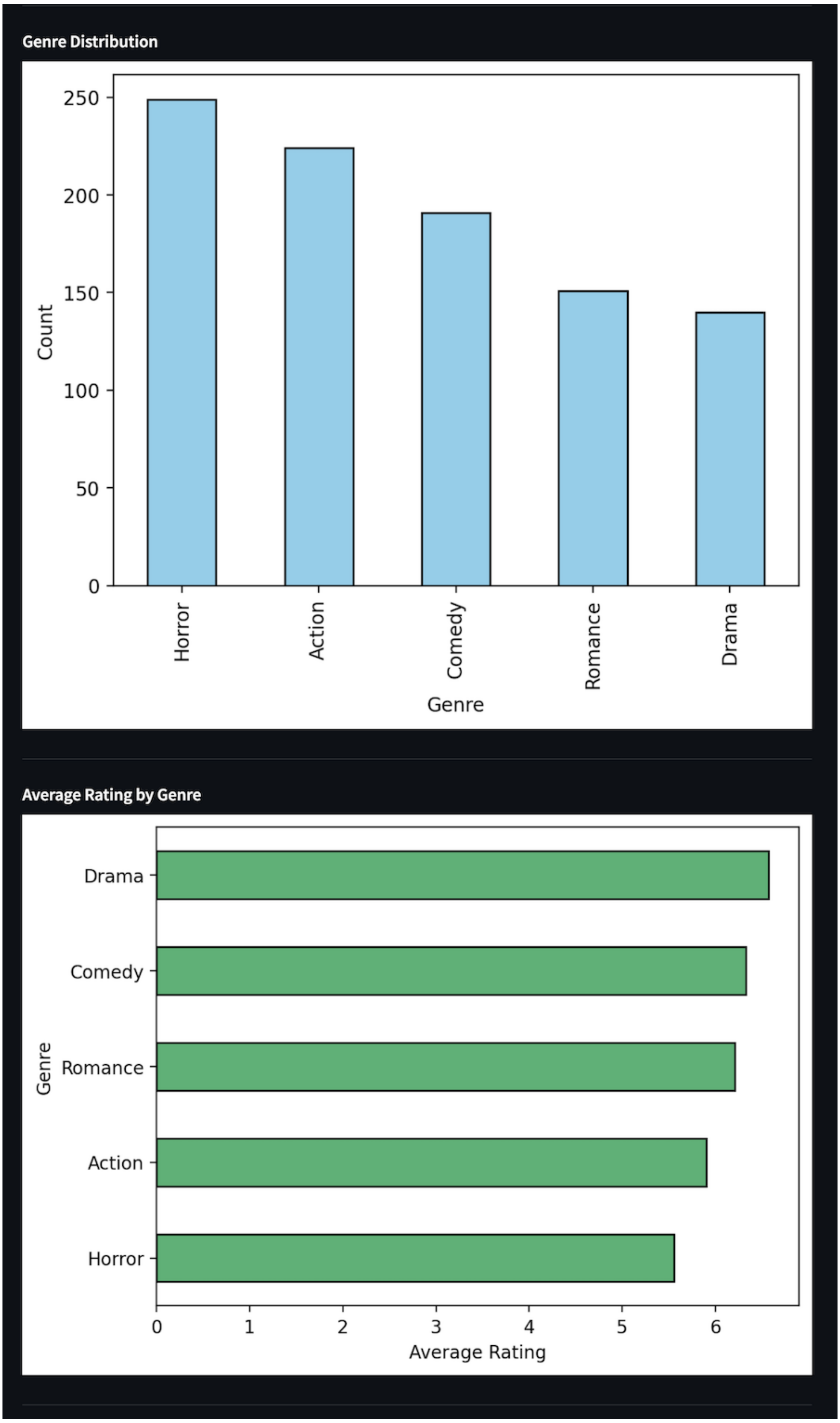
	Movie_Name	Genre	Rating	Voting_Counts	Duration
238	The Mermaid	Horror	9.6	15	< 2 hrs

Lowest Rated Movie

	Movie_Name	Genre	Rating	Voting_Counts	Duration
225	Goldilocks and the Three Bears: Death and Porridge	Horror	2.1	1,900	< 2 hrs

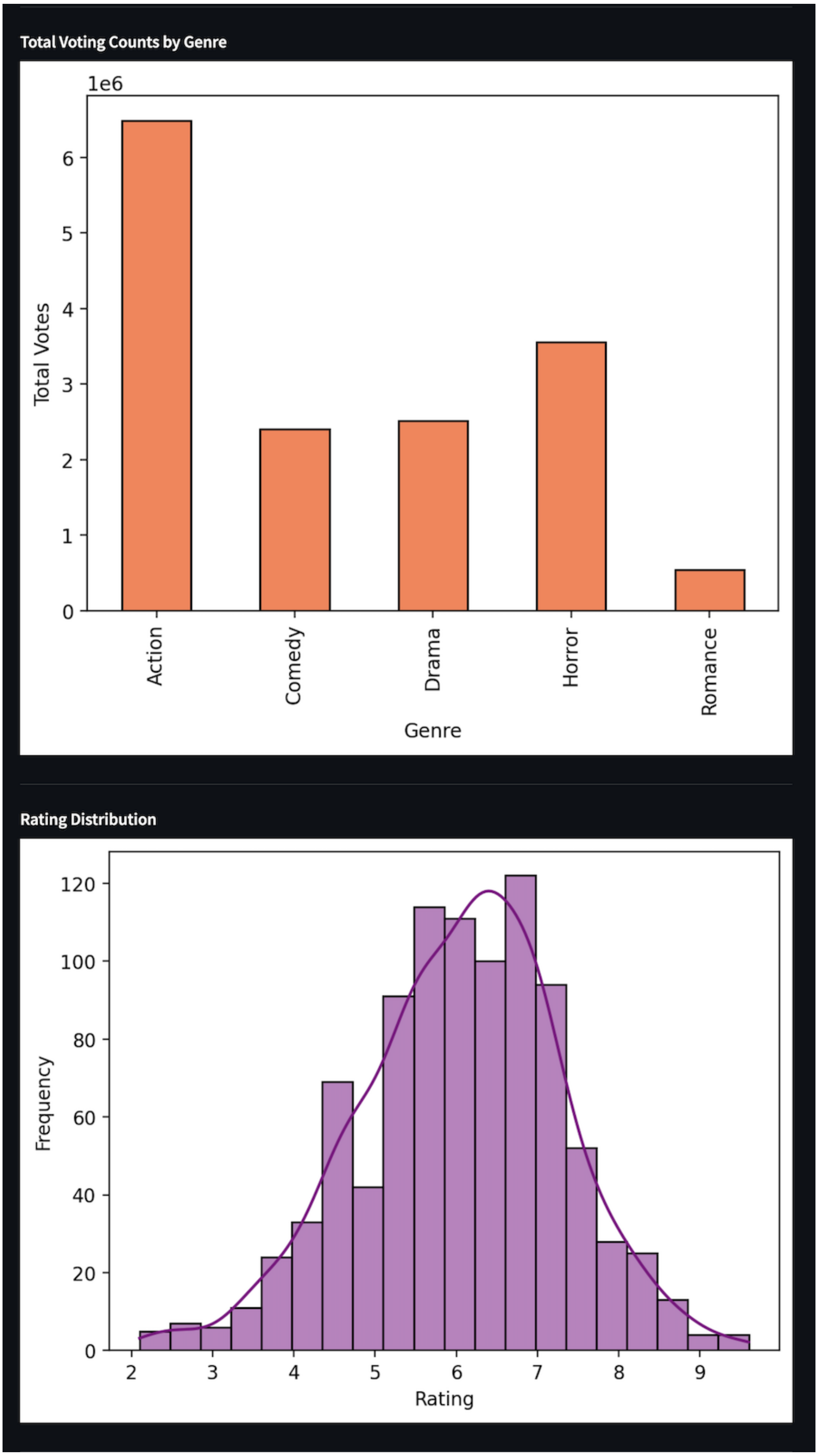
Ratings by Genre and Duration

Genre	2–3 hrs	3–4 hrs	< 2 hrs
Action	6.57	6.66	5.45
Comedy	6.79	7.42	6.24
Drama	6.8	7.24	6.43
Horror	6.16	7.02	5.43
Romance	6.22	6.88	6.13



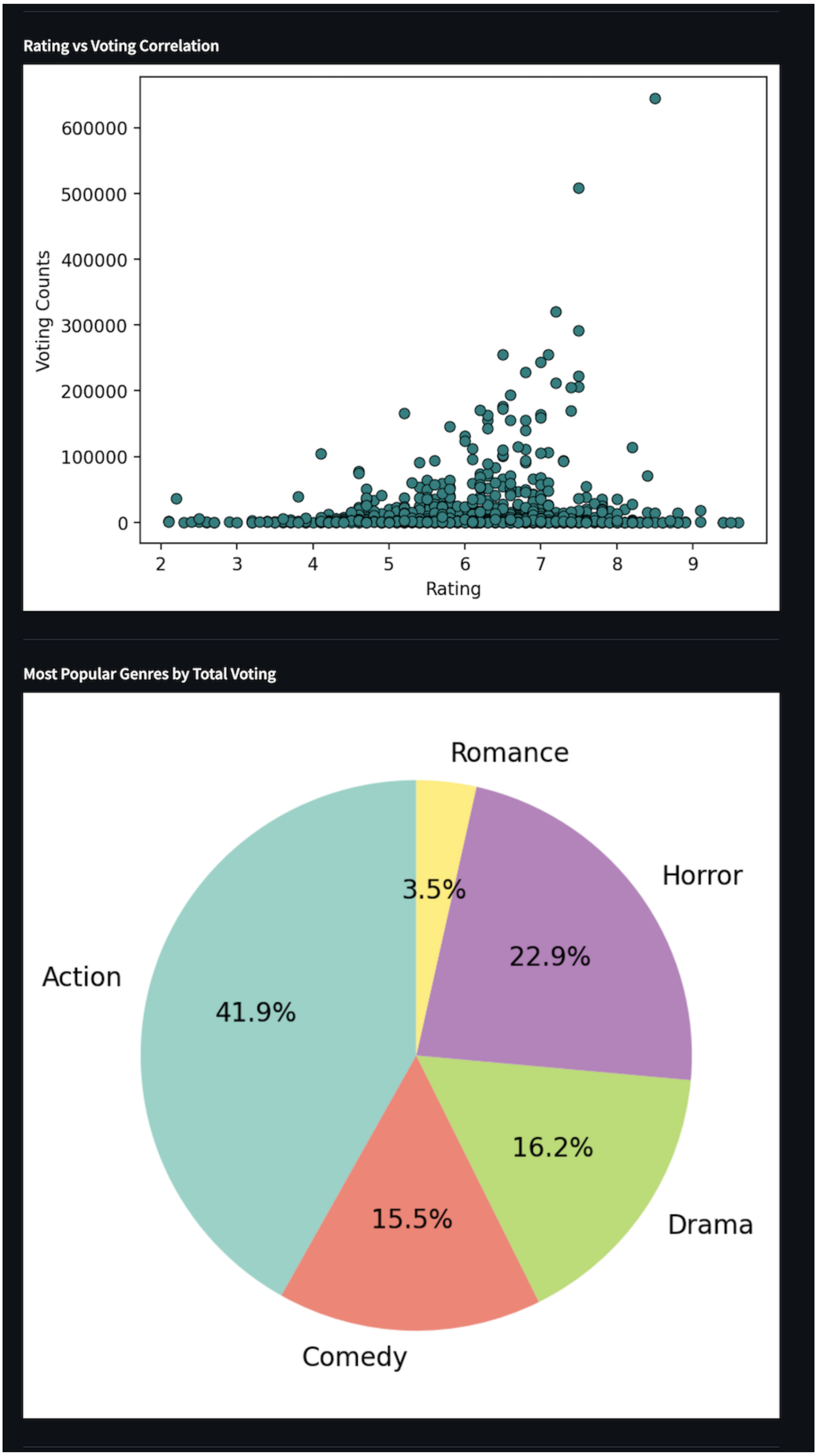
Genre Distribution

Average Rating by Genre



Total Voting Counts by Genre

Rating Distribution



Rating vs Voting Correlation

Most Popular Genres by Total Voting