Mini Project—1

IMDB 2024 Data Scraping and Visualizations

Introduction

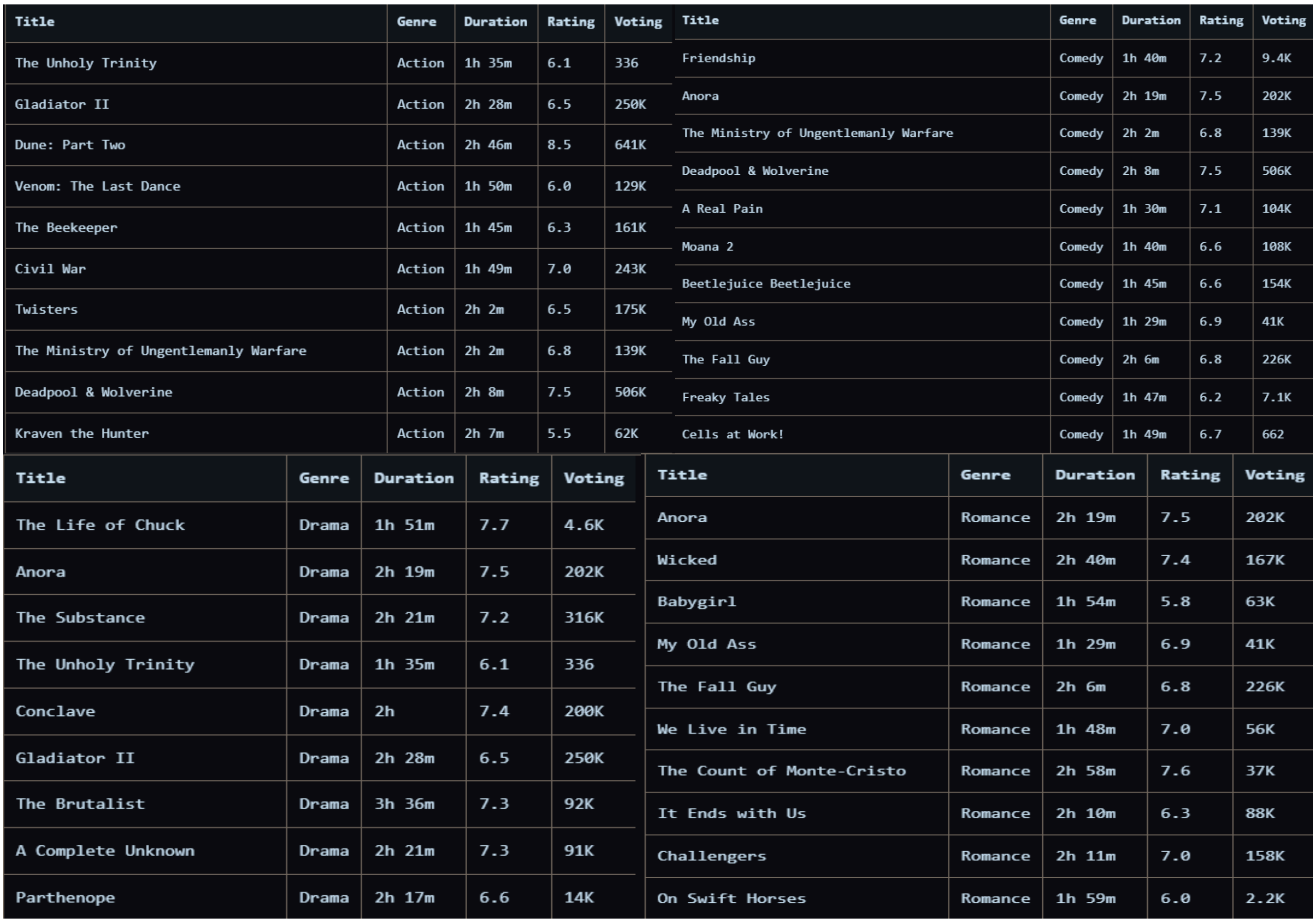
This project focuses on extracting and analyzing movie data from IMDb for the year 2024. The task involves scraping data such as movie names, genres, ratings, voting counts, and durations from IMDb's 2024 movie list using Selenium. The data will then be organized genre-wise, saved as individual CSV files, and combined into a single dataset stored in an SQL database. Finally, the project will provide interactive visualizations and filtering functionality using Streamlit to answer key questions and allow users to customize their exploration of the dataset.

Approach

DATA EXTRACTION USING SELENIUM

During the IMDb data extraction process, we initially relied on XPath expressions to locate and extract elements such as movie titles, genres, durations, ratings, and vote counts. However, we encountered several challenges with this approach. IMDb's HTML structure is highly dynamic and deeply nested, making XPath expressions long, brittle, and difficult to maintain. Even minor changes in the page layout would break the XPaths, resulting in scraping failures. Furthermore, the inconsistency in the availability of certain fields across different movie entries led to frequent exceptions, which added to the complexity of error handling. Due to these limitations, we explored CSS selectors as an alternative. CSS selectors proved to be more concise, readable, and resilient to structural changes on the webpage. They also allowed for easier handling of optional or missing elements without interrupting the scraping process. This shift significantly improved the robustness and maintainability of the scraping logic, making CSS selectors the preferred choice for the majority of the extraction tasks.

Beyond element extraction, we also faced difficulties in implementing pagination and scroll-based loading. IMDb's interface loads new movie entries either through paginated links or dynamic content loading triggered by scrolling. Detecting the end of pagination and interacting with dynamically rendered buttons required careful handling, including scroll automation and retry logic. In some cases, the "Next" button would not appear until a scroll event occurred, or it would be replaced by a JavaScript-rendered element. To overcome this, we used a combination of scrolling scripts, presence checks, and fallback mechanisms to ensure full traversal across all pages. This comprehensive approach to both element extraction and navigation significantly improved the reliability and completeness of the scraped dataset. Finally,addressing the challenges the details of the movies are extracted based on genre and made into separate csv files and kept under a single folder named genre_sep_files as action.csv,romance.csv,drama.csv,thriller.csv and horror.csv respectively.



Title	Genre	Duration	Rating	Voting
The Unholy Trinity	Action	1h 35m	6.1	336
Gladiator II	Action	2h 28m	6.5	250K
Dune: Part Two	Action	2h 46m	8.5	641K
Venom: The Last Dance	Action	1h 50m	6.0	129K
The Beekeeper	Action	1h 45m	6.3	161K
Civil War	Action	1h 49m	7.0	243K
Twisters	Action	2h 2m	6.5	175K
The Ministry of Ungentlemanly Warfare	Action	2h 2m	6.8	139K
Deadpool & Wolverine	Action	2h 8m	7.5	506K
Kraven the Hunter	Action	2h 7m	5.5	62K

Title	Genre	Duration	Rating	Voting
Friendship	Comedy	1h 40m	7.2	9.4K
Anora	Comedy	2h 19m	7.5	202K
The Ministry of Ungentlemanly Warfare	Comedy	2h 2m	6.8	139K
Deadpool & Wolverine	Comedy	2h 8m	7.5	506K
A Real Pain	Comedy	1h 30m	7.1	104K
Moana 2	Comedy	1h 40m	6.6	108K
Beetlejuice Beetlejuice	Comedy	1h 45m	6.6	154K
My Old Ass	Comedy	1h 29m	6.9	41K
The Fall Guy	Comedy	2h 6m	6.8	226K
Freaky Tales	Comedy	1h 47m	6.2	7.1K
Cells at Work!	Comedy	1h 49m	6.7	662

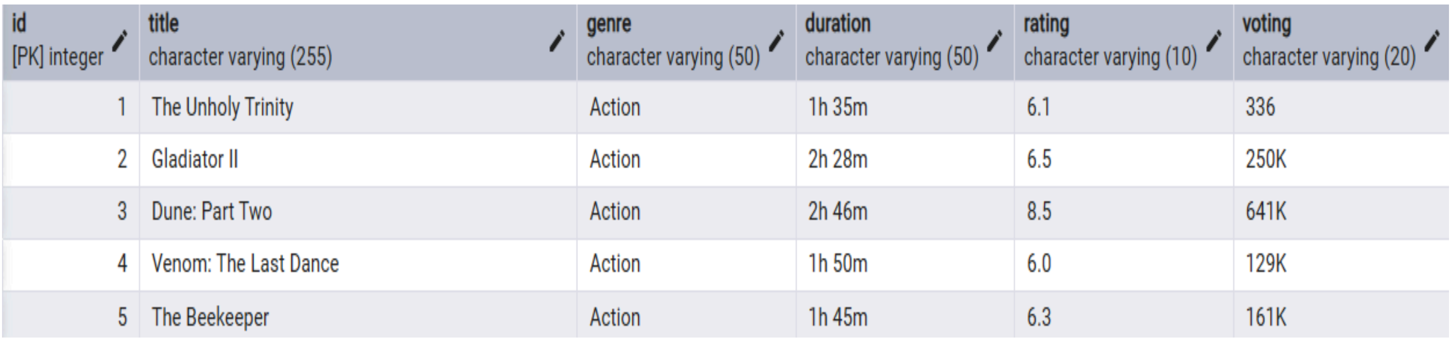
Title	Genre	Duration	Rating	Voting
The Life of Chuck	Drama	1h 51m	7.7	4.6K
Anora	Drama	2h 19m	7.5	202K
The Substance	Drama	2h 21m	7.2	316K
The Unholy Trinity	Drama	1h 35m	6.1	336
Conclave	Drama	2h	7.4	200K
Gladiator II	Drama	2h 28m	6.5	250K
The Brutalist	Drama	3h 36m	7.3	92K
A Complete Unknown	Drama	2h 21m	7.3	91K
Parthenope	Drama	2h 17m	6.6	14K

Title	Genre	Duration	Rating	Voting
Anora	Romance	2h 19m	7.5	202K
Wicked	Romance	2h 40m	7.4	167K
Babygirl	Romance	1h 54m	5.8	63K
My Old Ass	Romance	1h 29m	6.9	41K
The Fall Guy	Romance	2h 6m	6.8	226K
We Live in Time	Romance	1h 48m	7.0	56K
The Count of Monte-Cristo	Romance	2h 58m	7.6	37K
It Ends with Us	Romance	2h 10m	6.3	88K
Challengers	Romance	2h 11m	7.0	158K
On Swift Horses	Romance	1h 59m	6.0	2.2K

Fig:Sample of csv files of movies based on genre

For facilitating query and visualization process the extracted csv files are merged into a single csv file called movies_2024.csv.

For this project I have used PostgreSQL and created a database called movies containing the data from movies_2024.csv.



id [PK] integer	title character varying (255)	genre character varying (50)	duration character varying (50)	rating character varying (10)	voting character varying (20)
1	The Unholy Trinity	Action	1h 35m	6.1	336
2	Gladiator II	Action	2h 28m	6.5	250K
3	Dune: Part Two	Action	2h 46m	8.5	641K
4	Venom: The Last Dance	Action	1h 50m	6.0	129K
5	The Beekeeper	Action	1h 45m	6.3	161K

Fig:Sample schema structure from movies DB in PostgreSQL database

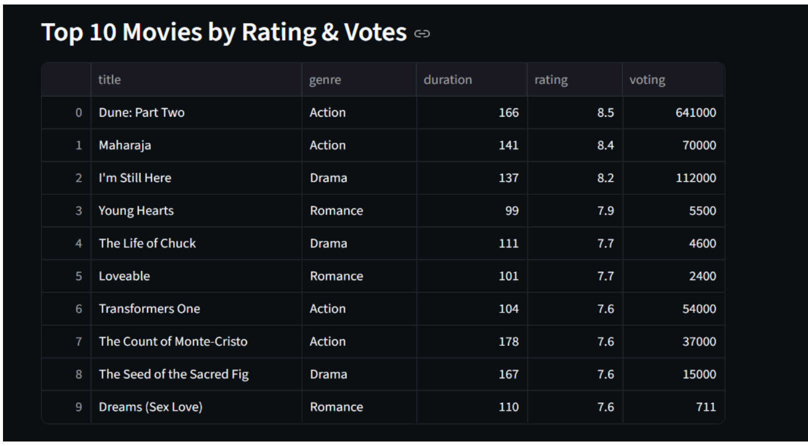
DATA ANALYSIS,VISUALIZATION AND FILTRATION

This Streamlit application provides an interactive dashboard for analyzing and exploring IMDb movies released in 2024. The app connects to a PostgreSQL database and retrieves movie data including title, genre, duration, rating, and voting count. It processes the raw data by converting durations (e.g., “1h 45m”) into total minutes and normalizing voting counts (e.g., “5.6K” to 5600). The application features two main sections: "Movie Trends & Analysis - 2024" and "Find Your Movie".

In the "Movie Trends & Analysis - 2024" section, the app presents a series of visualizations and data tables. These include the top 10 movies by rating and votes, genre distribution, average duration and voting trends by genre, and a histogram of rating distributions. It also identifies the top-rated movie per genre, most popular genres by voting, and highlights the shortest and longest movies based on runtime. Additionally, a heatmap displays average ratings across genres, and a scatter plot visualizes the correlation between ratings and voting counts.

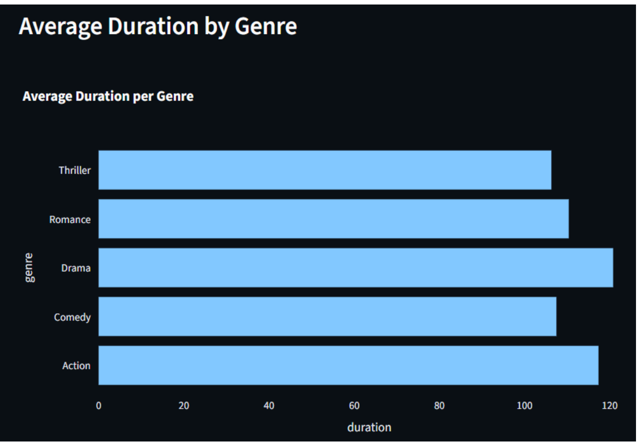
The "Find Your Movie" section allows users to filter movies based on genre, rating range, vote count, duration, and keyword-based title search. Filtered results are displayed in a table, and further visual analysis is provided using bar charts, pie charts, and scatter plots similar to the trends section but specific to user-selected criteria. The app dynamically updates insights such as genre-wise statistics and extreme duration movies within the filtered dataset.

Overall, this dashboard offers a user-friendly, data-driven exploration of recent movie trends and helps users discover high-rated and popular movies based on their preferences. It combines efficient data transformation, visualization with Plotly, and interactive filtering through Streamlit’s UI components.

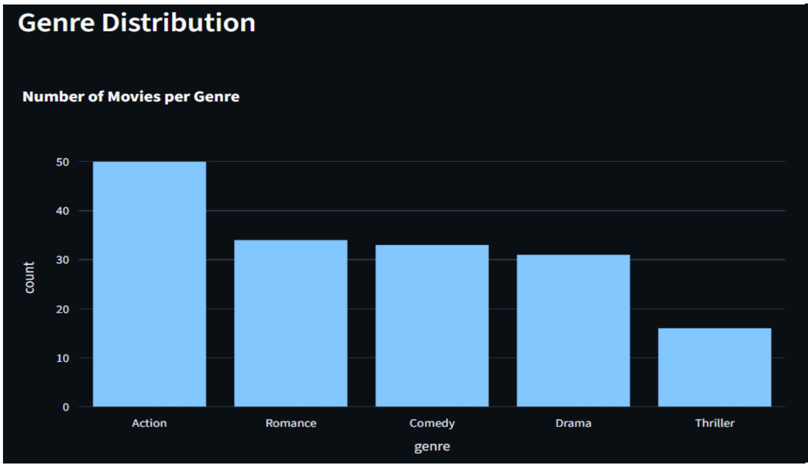


Top 10 Movies by Rating & Votes

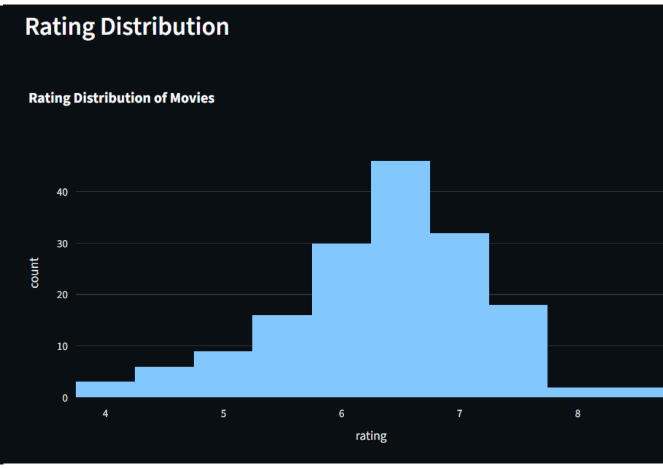
	title	genre	duration	rating	voting
0	Dune: Part Two	Action	166	8.5	641000
1	Maharaja	Action	141	8.4	70000
2	I'm Still Here	Drama	137	8.2	112000
3	Young Hearts	Romance	99	7.9	5500
4	The Life of Chuck	Drama	111	7.7	4600
5	Loveable	Romance	101	7.7	2400
6	Transformers One	Action	104	7.6	54000
7	The Count of Monte-Cristo	Action	178	7.6	37000
8	The Seed of the Sacred Fig	Drama	167	7.6	15000
9	Dreams (Sex Love)	Romance	110	7.6	711



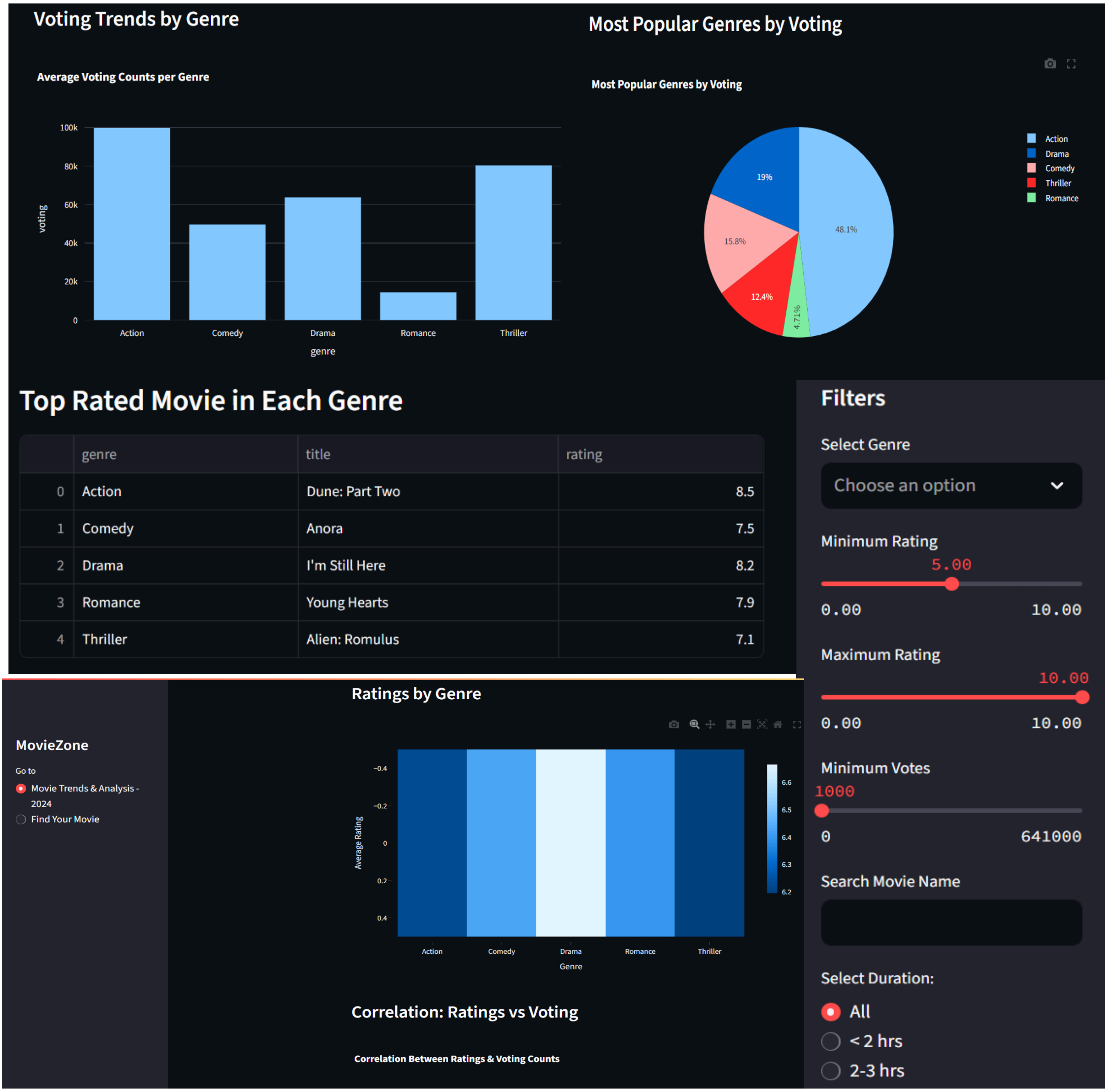
Average Duration by Genre



Genre Distribution



Rating Distribution



Top Rated Movie in Each Genre

	genre	title	rating
0	Action	Dune: Part Two	8.5
1	Comedy	Anora	7.5
2	Drama	I'm Still Here	8.2
3	Romance	Young Hearts	7.9
4	Thriller	Alien: Romulus	7.1

Filters

Select Genre: Choose an option

Minimum Rating: 5.00 (0.00 – 10.00)

Maximum Rating: 10.00 (0.00 – 10.00)

Minimum Votes: 1000 (0 – 641000)

Search Movie Name

Select Duration:
● All
○ < 2 hrs
○ 2-3 hrs

MovieZone

Go to
● Movie Trends & Analysis - 2024
○ Find Your Movie

Fig:WalkThrough of overall application