Cinematic Analytics

A Data-Driven Exploration of IMDB Films

Final project report

Phase 1: Project Idea📽️

The goal of this project is to extract and analyze data from IMDb, one of the largest platforms for movies and TV shows. By collecting data such as movie ratings, genres, release years, and other movie details, we can uncover interesting patterns and trends in the film industry. We will clean and process this data to perform various analyses and gain insights into how movies perform across different factors

Project Details

Target Website:

The website we will be using to extract data is **IMDb**🎟️

Data to be Collected:

The dataset will include information such as:

- Movie details: title, release year, director, main actors
- Genre: action, drama, comedy, etc
- Rating: user rating out of 10
- Duration: length of the movie
- Votes: number of people who voted for the movie

Analysis Ideas:

- **Movies by Year:** We will analyze how the number of movies produced has changed over the years. Is there a trend where the number of movies is increasing or decreasing over time?
- **Movie Duration:** We will calculate the average duration of movies. Do longer movies tend to have higher ratings? Are most movies long (more than 120 minutes) or short?
- **Movie Ratings:** What is the average rating across movies? Are the ratings mostly high or low? Are there any movies with a perfect score of 10/10? If so, what are their characteristics?
- **Number of Votes:** What are the movies that received the most votes? Is there any correlation between the number of votes and the rating?
- **Most Popular Genres:** We will identify which genres are the most common in the dataset (action, drama, comedy).
- **Genres with Highest Ratings:** Are there specific genres that tend to get higher ratings?
- **Duration by Genre:** Do certain genres have longer or shorter movies on average? For example, are sci-fi movies generally longer than comedies?
- **Relationships Between Variables:** Is there a correlation between the release year and rating? Do older movies tend to have higher ratings than newer ones, or the other way around? , Do longer movies receive more votes or higher ratings?
- **Top 10 Rated Movies:** What are the characteristics of the highest-rated movies (e.g., genre, year, duration)?
- **Bottom 10 Rated Movies:** Are there any common patterns among the lowest-rated movies (genre, year)?
- **Most and Least Voted Movies:** Do movies with the most votes tend to be newer or older?
- **Changes in Average Ratings Over Time:** Has the quality of movies (as reflected in ratings) improved or declined over the years?

Tools and techniques:

We will use Python to implement our analytics and extract insights from the movie dataset. The main tools and techniques include:

- **Outliers and Regular Expressions:**

-We will identify outliers such as extremely long or short movies and analyze how these durations affect ratings.

-We'll also investigate movies with extremely high or low ratings to understand their characteristics.

-Regular expressions will be applied to clean and extract specific patterns from textual data such as titles or genres.

- **Visualization Techniques:**

-Histograms to show the distribution of ratings and durations.

-Bar charts to visualize the number of movies in each genre

-Scatter plots to explore relationships between ratings and duration, or ratings and release year

-Line charts to display trends over the years, such as the average ratings or number of movies released

- **Programming Tools:**

We will use Python with libraries such as pandas, matplotlib, and seaborn for data analysis and creating meaningful visualizations

- **Storage and Data Handling:**

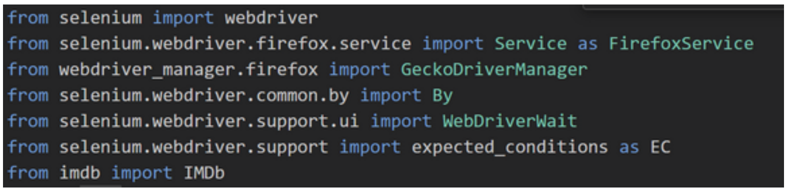
Data will be stored and managed using Excel or SQL, depending on the structure and size of the dataset

Phase2:Report

The IMDb Data Scraping Project aims to collect and analyze information from **IMDb** one of the largest and most comprehensive databases of movies and TV series. We extracted the **top 250** rated movies on the site using **Python** and then conducted in depth analysis of this data to discover trends, patterns, and insights related to the movies. This report covers the stages of **data exploration**, **cleaning**, and **visualization** along with a review of the most important findings related to movie ratings, genres running time, and the relationships between these factors. These analyses and interactions are visualized through an interactive Streamlit website, which allows users to browse the results in an easy and simple way

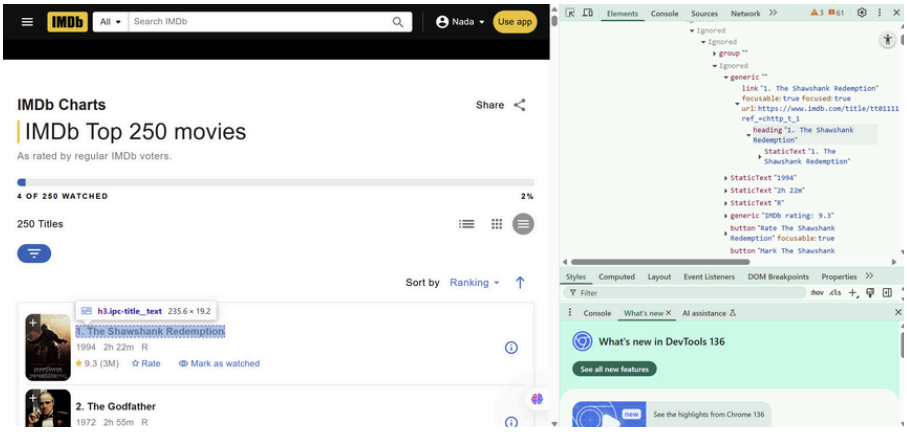
1. Data Extraction

Since **IMDb** is a **dynamic website** that relies heavily on **JavaScript** content loading, I used the **Selenium Python library** to handle and extract data from dynamic elements. **Firefox** was used with WebDriver, as it offers greater permissions and compatibility in some cases especially for sites that restrict automated access or require dynamic loading. After extracting movie titles with **Selenium**, I used the **IMDbPY library**, a powerful library specifically designed for working with the IMDb database. The **search_movie()** function allows you to enter a movie name and automatically retrieve all related details—such as rating genre, and release date



```
from selenium import webdriver
from selenium.webdriver.firefox.service import Service as FirefoxService
from webdriver_manager.firefox import GeckoDriverManager
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from imdb import IMDb
```

- selenium to interact with the dynamic IMDb webpage
- webdriver_manager to auto-install Firefox WebDriver
- IMDbPY to fetch detailed movie information
- pandas and time for data handling and control



IMDb
All
Search IMDb
Nada
Use app
IMDb Charts
Share
IMDb Top 250 movies
As rated by regular IMDb voters.
4 OF 250 WATCHED
2%
250 Titles
Sort by Ranking
h3.ipc-title__text 235.6 × 19.2
1. The Shawshank Redemption
1994 2h 22m R
9.3 (3M) Rate Mark as watched
2. The Godfather
1972 2h 55m R
Elements Console Sources Network
Styles Computed Layout Event Listeners DOM Breakpoints Properties
Filter :hov .cls
Console What's new AI assistance
What's new in DevTools 136
See all new features
new
See the highlights from Chrome 136

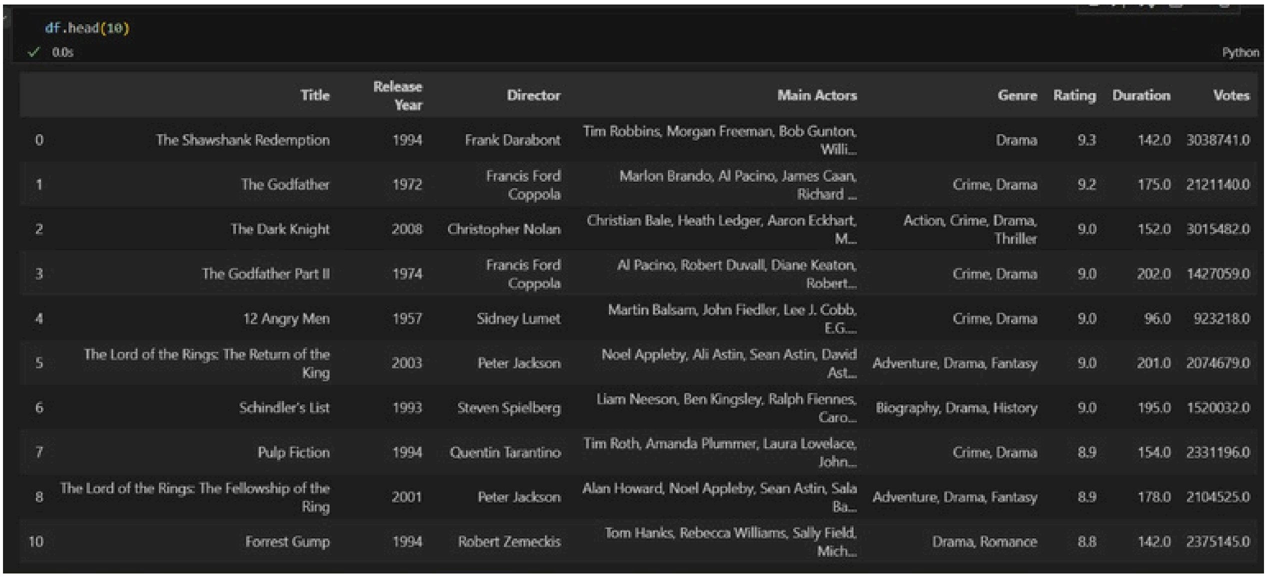
```
▾ Ignored
  ▾ Ignored
    ▸ group ""
    ▾ Ignored
      ▾ generic ""
          link "1. The Shawshank Redemption"
          focusable: true focused: true
        ▾ url: https://www.imdb.com/title/tt01111
          ref_=chttp_t_1
          ▾ heading "1. The Shawshank
            Redemption"
            ▸ StaticText "1. The
              Shawshank Redemption"
      ▸ StaticText "1994"
      ▸ StaticText "2h 22m"
      ▸ StaticText "R"
      ▸ generic "IMDb rating: 9.3"
      ▸ button "Rate The Shawshank
        Redemption" focusable: true
        button "Mark The Shawshank
```

Scraped Movie Titles

- Located all .ipc-metadata-list-summary-item elements (each movie block)
- Extracted the movie name from the <h3> tag
- Cleaned titles by removing index numbers and stored them in a list

Stored All Data in a DataFrame & Fetched Movie Details Using IMDbPY

- For each movie title, used **search_movie()** to find the movie object
- Extracted attributes like **title**, **year**, **rating**, **genre**, **director**, **cast**, **runtime**, and **votes**
- Stored each movie's details in a **dictionary**
- Combined all dictionaries into a **list**
- Converted the list to a **pandas.DataFrame** for analysis and visualization



```
df.head(10)
```

✓ 0.0s

Python

	Title	Release Year	Director	Main Actors	Genre	Rating	Duration	Votes
0	The Shawshank Redemption	1994	Frank Darabont	Tim Robbins, Morgan Freeman, Bob Gunton, Willi...	Drama	9.3	142.0	3038741.0
1	The Godfather	1972	Francis Ford Coppola	Marlon Brando, Al Pacino, James Caan, Richard ...	Crime, Drama	9.2	175.0	2121140.0
2	The Dark Knight	2008	Christopher Nolan	Christian Bale, Heath Ledger, Aaron Eckhart, M...	Action, Crime, Drama, Thriller	9.0	152.0	3015482.0
3	The Godfather Part II	1974	Francis Ford Coppola	Al Pacino, Robert Duvall, Diane Keaton, Robert...	Crime, Drama	9.0	202.0	1427059.0
4	12 Angry Men	1957	Sidney Lumet	Martin Balsam, John Fiedler, Lee J. Cobb, E.G....	Crime, Drama	9.0	96.0	923218.0
5	The Lord of the Rings: The Return of the King	2003	Peter Jackson	Noel Appleby, Ali Astin, Sean Astin, David Ast...	Adventure, Drama, Fantasy	9.0	201.0	2074679.0
6	Schindler's List	1993	Steven Spielberg	Liam Neeson, Ben Kingsley, Ralph Fiennes, Caro...	Biography, Drama, History	9.0	195.0	1520032.0
7	Pulp Fiction	1994	Quentin Tarantino	Tim Roth, Amanda Plummer, Laura Lovelace, John...	Crime, Drama	8.9	154.0	2331196.0
8	The Lord of the Rings: The Fellowship of the Ring	2001	Peter Jackson	Alan Howard, Noel Appleby, Sean Astin, Sala Ba...	Adventure, Drama, Fantasy	8.9	178.0	2104525.0
10	Forrest Gump	1994	Robert Zemeckis	Tom Hanks, Rebecca Williams, Sally Field, Mich...	Drama, Romance	8.8	142.0	2375145.0

2. Data exploration and cleaning

- **Initial Data Inspection**

Used **df.head()**, **df.info()**, and **df.describe()** to explore structure and summary stats
Checked for missing or invalid values like **N/A**

- **Handled Missing and Inconsistent Data**

Replaced all 'N/A' values with **np.nan**
Dropped rows with missing critical fields (Rating, Duration, Votes, Year)

- **Data Type Conversion**

Converted Release Year to integer
Stripped /10 from Rating and converted it to float
Extracted numeric part of Duration and casted to float
Removed commas from Votes and casted to float

- **Removed Outliers**

Identified outliers in Votes using the IQR method
Filtered out extremely high or low vote counts for fairer analysis

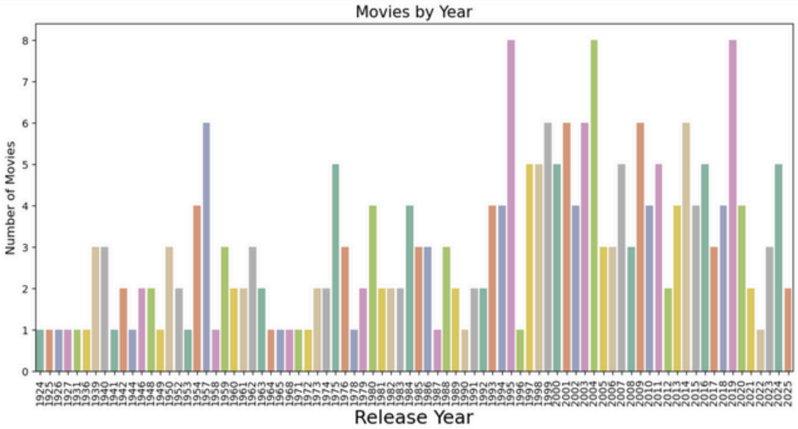
- **Segmented Movies by Rating**

Created two subsets:
High Rated Movies: **Rating ≥ 8.5**
Low Rated Movies: **Rating ≤ 5.0**
Displayed selected columns for both groups to compare patterns

3.Visualization

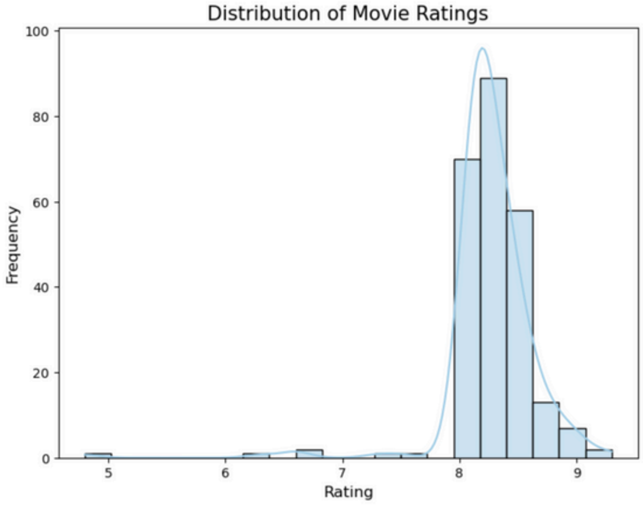
We used graphs to understand distributions, observe relationships between variables, and reveal any pattern or trend in the movie data in a visual and clear way

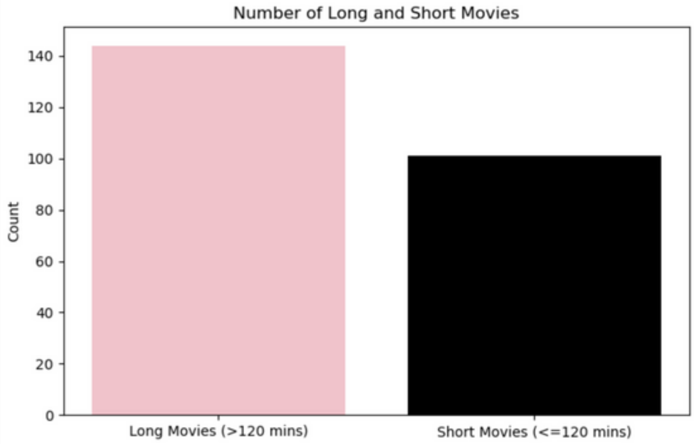
Movies by Year



Older classics (pre-2000) dominate IMDb's Top 250, suggesting enduring quality outweighs modern quantity

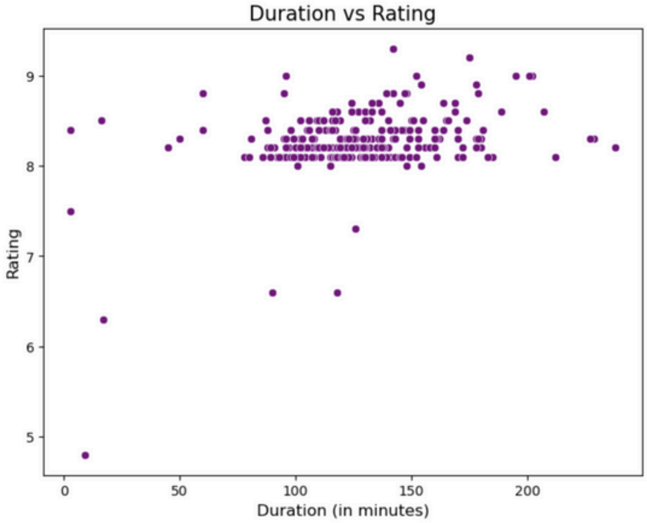
Movie Ratings&Movie Duration





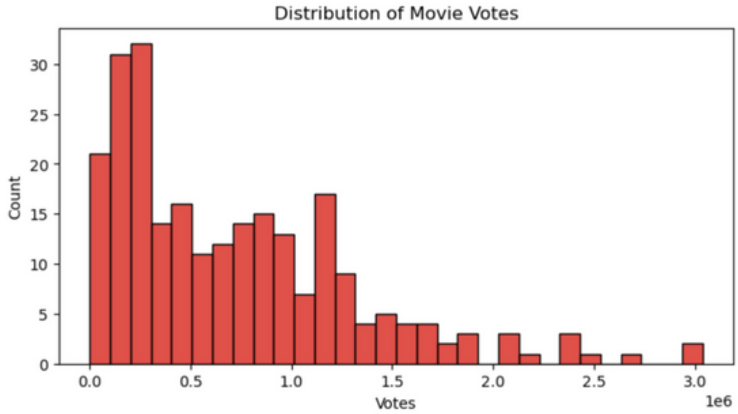
Ratings range from 5 to 8, indicating that most films receive average to high ratings.
There are a few films with low ratings (below 4) or very high ratings (above 8).

Short films (≤120 minutes) significantly outnumber feature films (>120 minutes). This may be because short films are more popular, or because audiences prefer shorter content. Alternatively, the film industry tends to produce more short films, perhaps due to factors such as cost or viewer preferences.



There is no clear pattern linking film length and rating. Some short films receive high ratings, while some feature films also receive good ratings, and vice versa.
A film's quality (as reflected in ratings) depends not on its length, but on other factors such as story, acting, and directing.

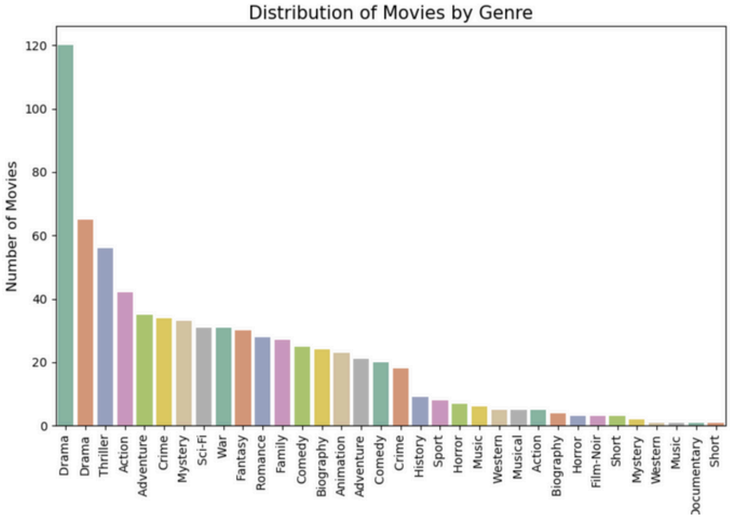
Number of Votes

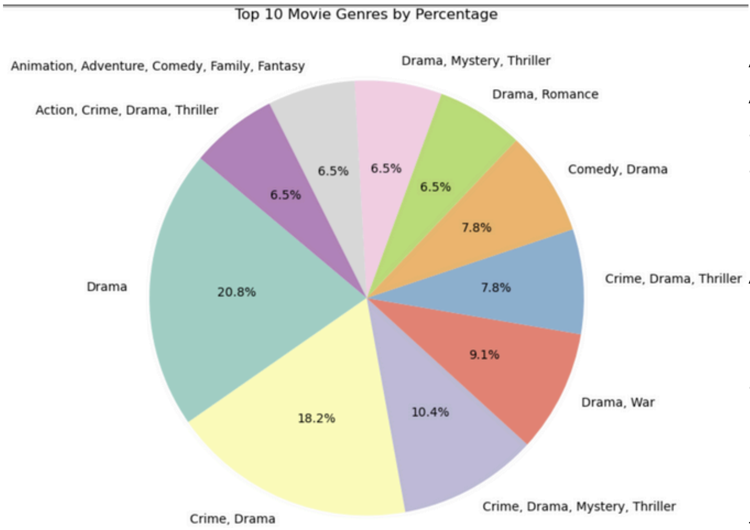


Most films have low ratings (less than 1 million).
Very few films have ratings of 1 million or more, indicating that they are very popular or well-liked films.
The distribution shows a significant rightward skew, with the vast majority of films having low ratings, while a few have huge ratings.

Most-viewed and discussed films (those with millions of ratings) are rarer than average films.

Most Popular Genres





By reviewing the distribution of genres in the list of top-rated films, we note that drama is the undisputed king. It appears in most combinations, whether with crime, romance, or comedy, and constitutes the core element of the vast majority of films, serving as the backbone of popular cinema.

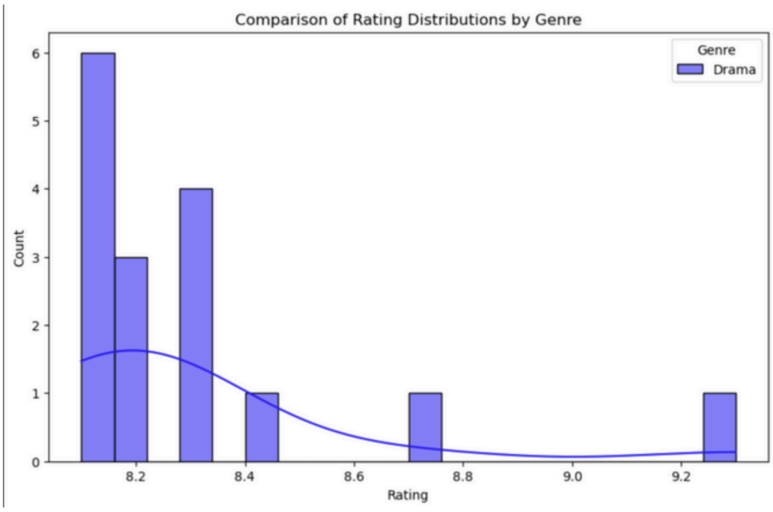
The three genres (Crime, Drama, Thriller) top the list with 10.4%, demonstrating audiences' passion for suspenseful, suspenseful, and mysterious films. Comedy, Drama, and Romance followed closely behind with 9.1%, and Drama, Romance, 7.8%, indicating audiences' love of emotion and laughter combined.

Around the middle, we find that genres such as Animation, Adventure, Comedy, Family, and Fantasy all share an equal proportion of appearances, at around 6.5%, indicating that they are popular but not necessarily dominant.

At the periphery, we find rarer genres such as documentary, short, western, and musical. These may target specific audiences or appear in specific artistic contexts that are not popular.

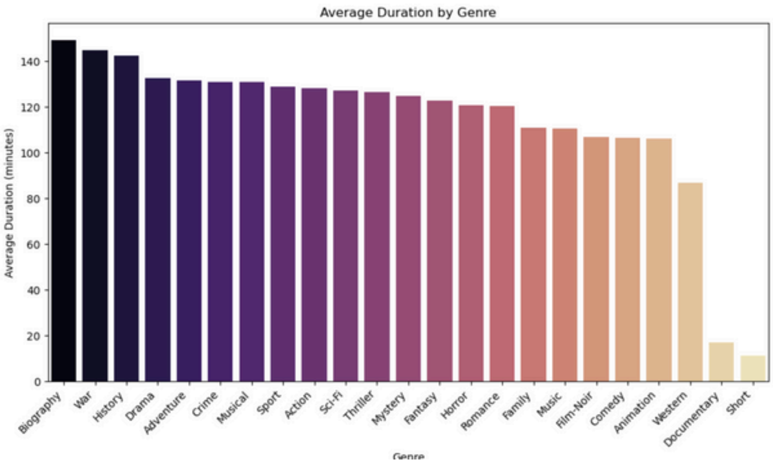
Distribution doesn't hold any major surprises, but it does reflect clear preferences: audiences tend toward dramas saturated with suspense or emotion, they love comedy and human stories, and they consider traditional genres to be the safe and reliable bet.

Genres with Highest Ratings



Top ratings: These are centered around 8.2 to 9.2, indicating that some genres receive very high ratings.
Variance: There is a clear difference between the ratings of different genres, with some exceeding 9.0 while others are around 8.2.
The quality of films varies greatly by genre, with some genres outperforming in ratings.
This may be because higher-rated genres (such as historical films or documentaries) attract a more conservative audience.

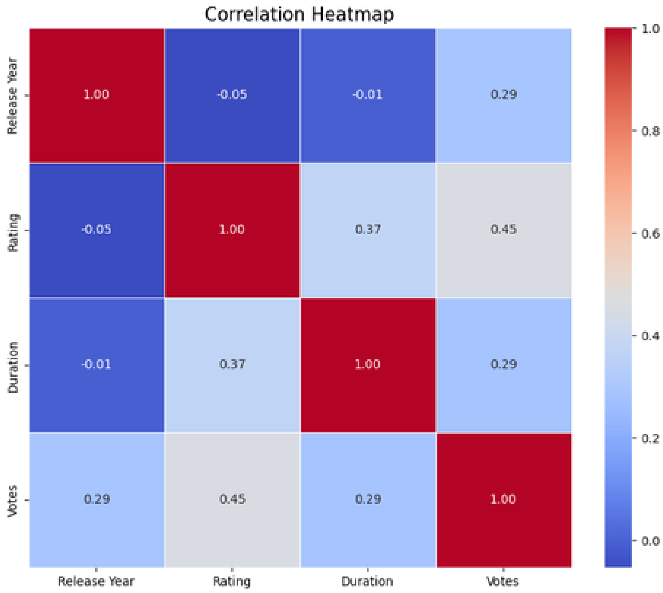
Duration by Genre



Longest genres: War, History (about 140 minutes), Drama, Adventure (over 120 minutes).
Shortest genres: Short, Documentary, Comedy (less than 80 minutes).

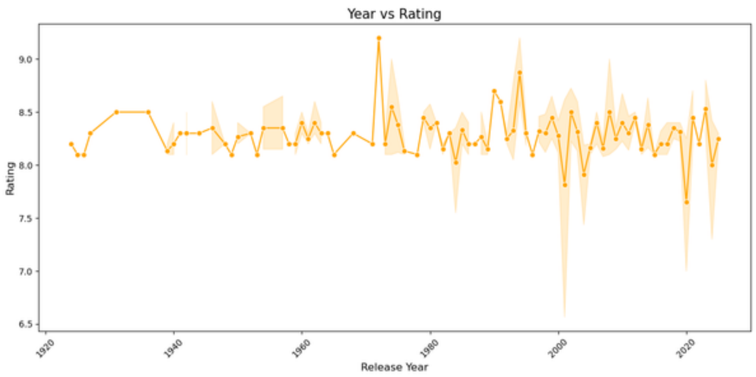
More complex genres (such as war and drama) require more time to develop the story.
Comedies and short films deliver quick content and don't require a long runtime.

Relationships Between Variables

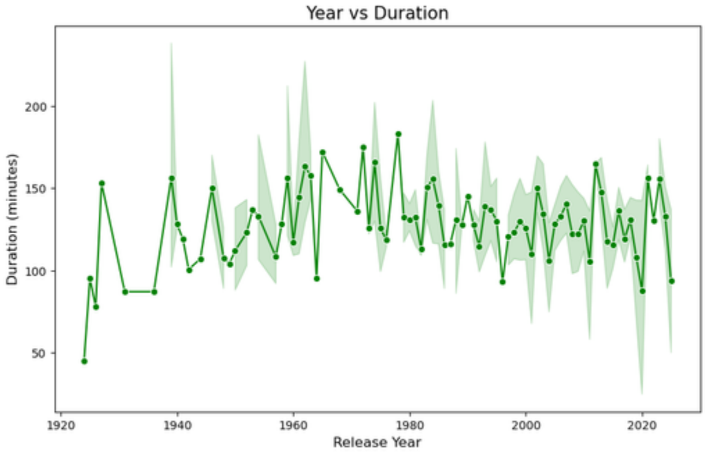


Rating and Votes: 0.45 (medium correlation) → Movies with higher ratings tend to have more ratings.
Rating and Duration: 0.37 (weak to medium correlation) → Longer movies may be of higher quality or have better ratings.
Rating and Votes are moderately correlated, indicating that higher-rated movies attract a larger audience.

Release Year does not significantly affect any of the other variables.



Ratings are relatively stable over the years, with slight variations between 6.5 and 9.0. There is no clear upward or downward trend, indicating that the quality of films (as rated) has neither improved nor declined over time. A film's age does not significantly affect its rating—quality remains relatively constant across the decades.

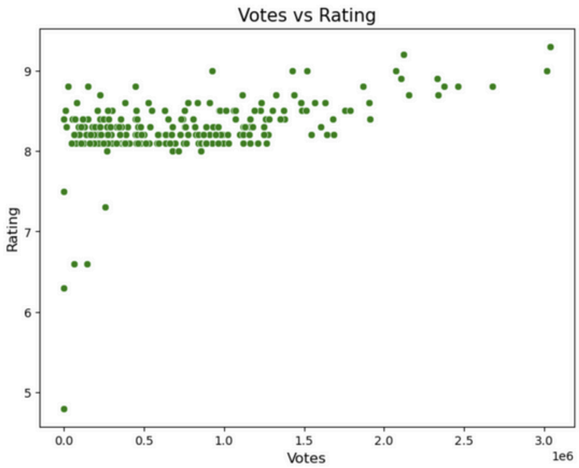


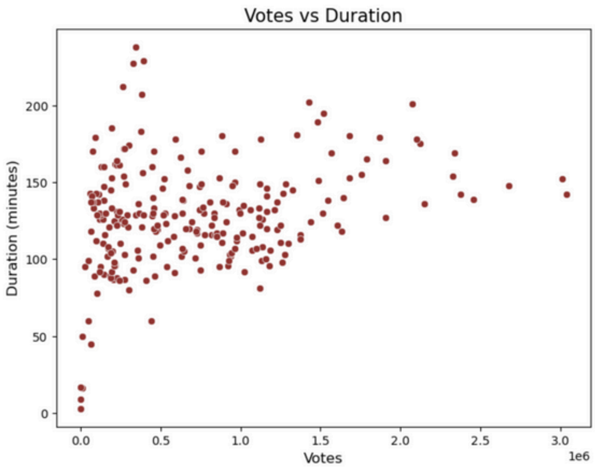
A clear decline in film length from approximately 180 minutes in the 1920s to approximately 100 minutes in the last decade

The longest films were from 1920 to 1960

The shortest films were from 2000 to 2020

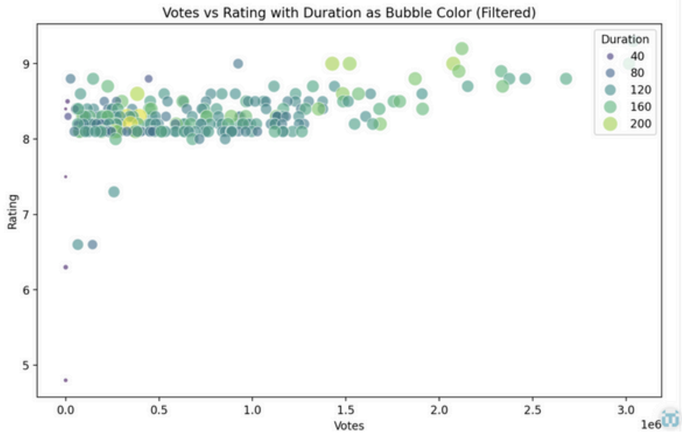
A clear historical trend: Film length has decreased by approximately 45% over the past 100 years





Most of the data is concentrated in the 5-7 rating range. Films with higher ratings (7+) have relatively fewer ratings. There is no strong correlation between the number of ratings and the average rating. Highly rated films (7+) are rare and may be specialized for specific audiences.

The data distribution is irregular with no clear pattern. Most films are between 80-160 minutes long, regardless of the number of ratings. Length is not a factor influencing the number of ratings. Very long films are less common and may target a niche audience.

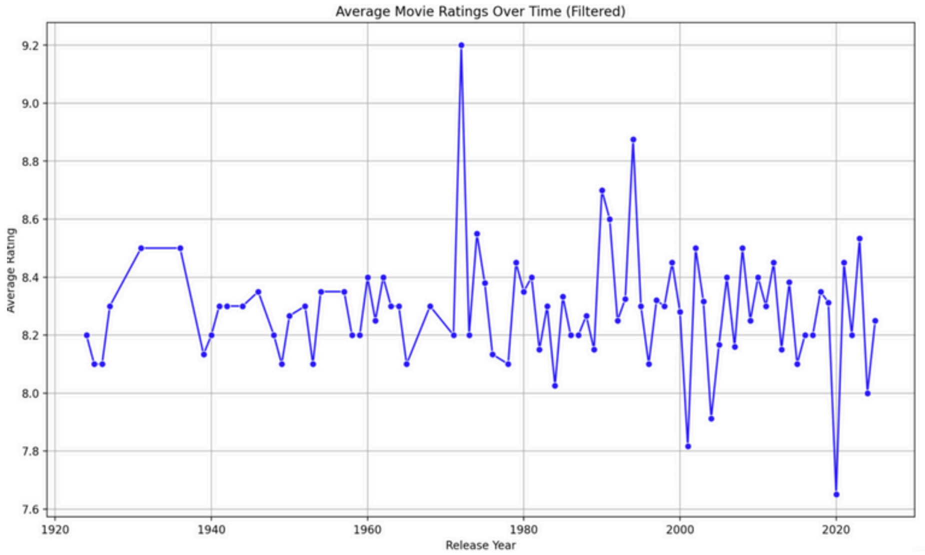


Large bubbles (high ratings) appear in the 6-7 rating range. Colored bubbles (length) show variation without a clear pattern. Some short films (<90 minutes) achieve high ratings.

The most popular films (high ratings) are not necessarily the longest.

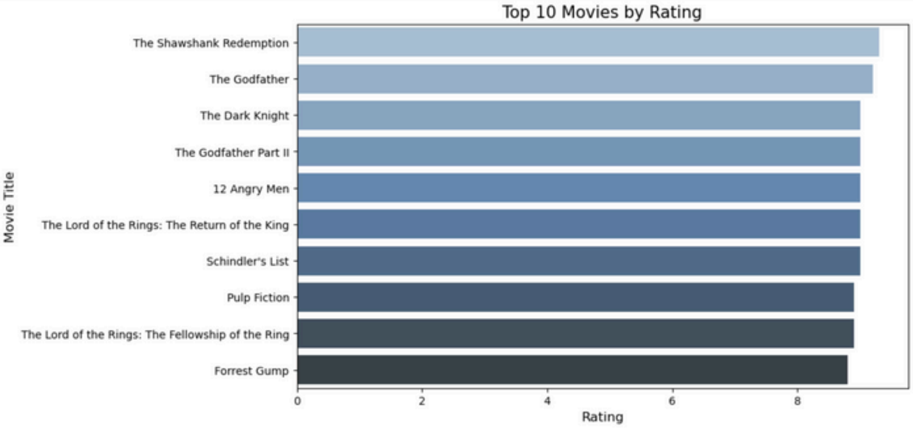
Content quality is more important than length in attracting ratings.

Changes in Average Ratings Over Time



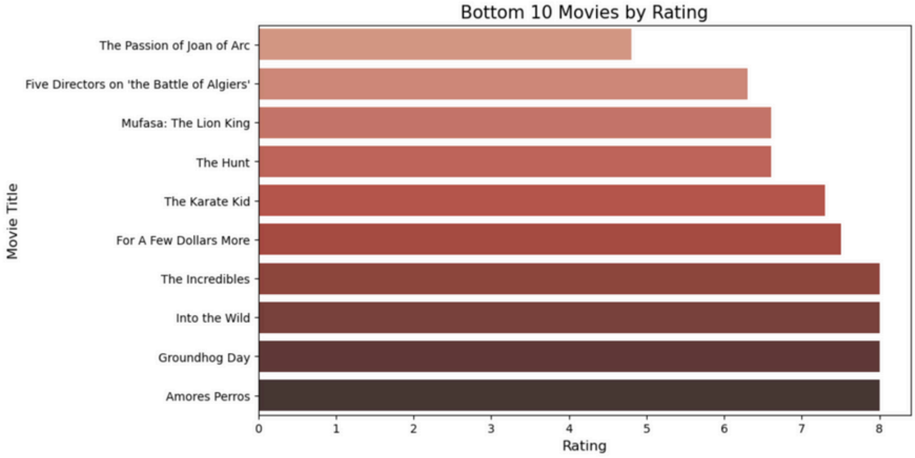
Remarkable stability: Average ratings remained around 8.0 to 9.0 from 1920 to 2020. A slight peak occurred in the mid-20th century (1940-1980), when ratings approached 9.2. A slight, but not significant, decline after 2000 (ranging from 8.0 to 8.6).
This could be attributed to changing rating standards or diverse audience tastes.
The Golden Age of Cinema (1940-1980) coincides with some of the highest ratings, reflecting the exceptional production of that period.

Top 10 Movies by Rating



The highest-rated films range from 8.5 to 9.2 on the rating scale.
Popular genres: Drama (70%), with some crime and fantasy films.
Timeless quality: Most of these films were released before 2000, demonstrating their durability over the decades.
Drama dominates: Films with complex human or historical stories rank highly in ratings.

Bottom 10 Rated Movies



It ranges from 1 to 4 out of 10, indicating significant critical or public dissatisfaction.
Films like Groundhog Day and The Incredibles typically have high ratings, so there may be an error in the data or these may be specific versions.
Possible reasons for low ratings: Poor production (poor direction/screenplay).
Limited audience (artistic or specialized documentaries that may not be suitable for the general public).

4.Apply Regression Model between Duration and Rating

In this step, different regression models were built with the aim of predicting movie ratings based on some quantitative characteristics present in the data, such as: **Movie Duration** ,**Number of Votes** ,**Release Year**

1. Linear Regression Model

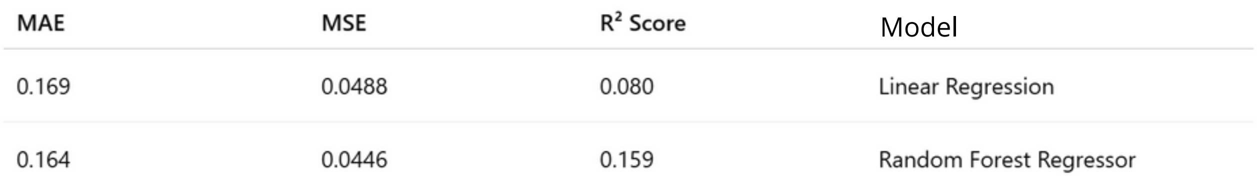
A linear regression model was first trained using the three variables mentioned.
The evaluation results were as follows:
R-squared (R^2): 0.115 → This indicates that the model explains only about **11.5%** of the variance in movie ratings.
Mean Squared Error (MSE): 0.047
Mean Absolute Error (MAE): 0.170
Root Mean Squared Error (RMSE): 0.217

2.Random Forest Regressor Model

A more complex model based on **Random Forest**, a nonlinear machine learning model, was tested. It was trained on the same data and achieved the following results:
R^2 Score: 0.159 (better than linear regression)
MSE: 0.045
MAE: 0.164
RMSE: 0.211

3.Compare models on a small group

The models were rebuilt using only two variables: the **number of votes** and the **duration of the film**.
The linear regression model was compared with the random forest model.



MAE	MSE	R^2 Score	Model
0.169	0.0488	0.080	Linear Regression
0.164	0.0446	0.159	Random Forest Regressor

Linear regression is a **good** starting point, but it's insufficient to explain the complexity of the relationship between characteristics and film ratings.
The Random Forest model performed significantly **better** and is considered a more effective option in this context.

4.Data Storage in MongoDB

MongoDB is a modern, document-based **NoSQL** database designed to store data in a flexible, **JSON**-like format. Unlike traditional relational databases (like MySQL or PostgreSQL) that use tables, rows, and fixed schemas, MongoDB stores data as collections of **documents** where each document can have a different structure

Connection and Data Storage

MongoDB is that it automatically creates databases and collections upon the insertion of data if they do not already exist.
After collecting and cleaning the IMDb movie dataset using Python and pandas, the data was transformed into a format suitable for MongoDB. Each row in the dataset, representing a single movie, was converted into a separate document. These documents were then inserted into the Movies collection.
Thanks to MongoDB's schema-free architecture, storing this scraped data—despite its varied structures—was seamless. This flexibility made MongoDB an excellent choice for managing and analyzing dynamic movie data.

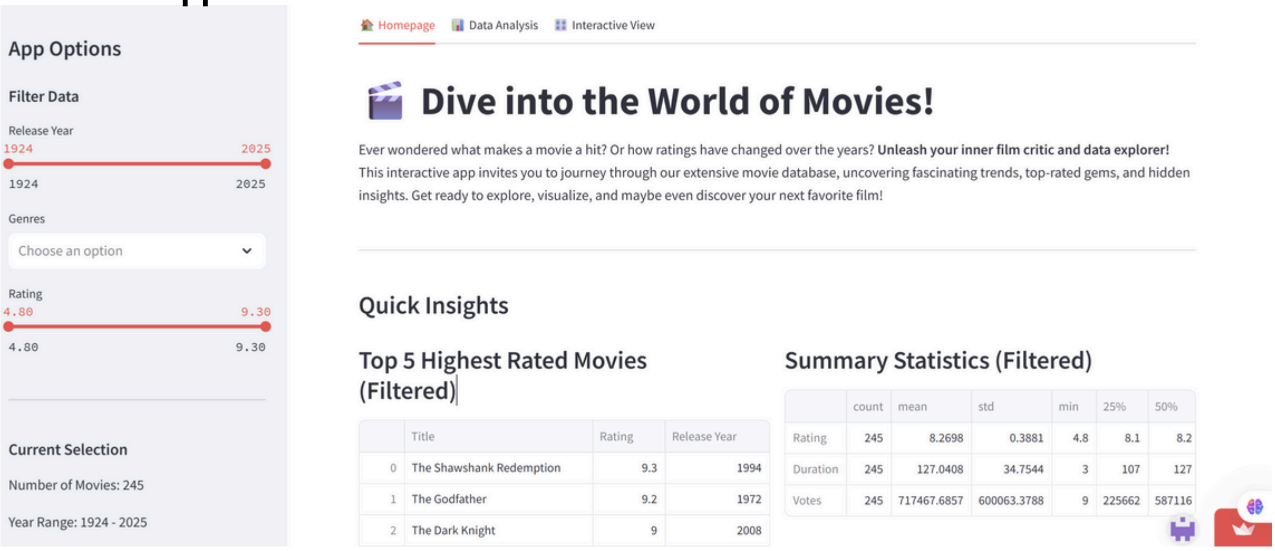
5.Deploying Interactive web application using Streamlit

The **Joblib library** was utilized in our project to save the processed DataFrame after completing the web scraping, cleaning, and preprocessing steps. Its role was crucial in improving efficiency and workflow reliability, especially during development and testing. Instead of repeating the time-consuming scraping and data preparation tasks every time we needed the data, we used:

joblib.dump(df, **"IMDB_data.joblib"**)
This allowed us to serialize the DataFrame and store it as a binary file (.joblib). Later, the data could be quickly reloaded in any script using:
df = joblib.load("IMDB_data.joblib")

Key Benefits:
Faster development: Avoids re-running scraping scripts, which might involve network delays or API limits.
Consistency: Ensures the same version of the data is used across different parts of the project (e.g., training models, deploying apps).
Simplicity: Makes it easy to share or reuse the data with teammates or in multiple scripts.
This made Joblib an essential tool for efficient data handling and reproducibility within the project lifecycle.

Streamlit Application



App Options

Filter Data

Release Year

1924 2025

1924 2025

Genres

Choose an option

Rating

4.80 9.30

4.80 9.30

Current Selection

Number of Movies: 245

Year Range: 1924 - 2025

🎬 Dive into the World of Movies!

Ever wondered what makes a movie a hit? Or how ratings have changed over the years? **Unleash your inner film critic and data explorer!** This interactive app invites you to journey through our extensive movie database, uncovering fascinating trends, top-rated gems, and hidden insights. Get ready to explore, visualize, and maybe even discover your next favorite film!

Quick Insights

Top 5 Highest Rated Movies (Filtered)

	Title	Rating	Release Year
0	The Shawshank Redemption	9.3	1994
1	The Godfather	9.2	1972
2	The Dark Knight	9	2008

Summary Statistics (Filtered)

	count	mean	std	min	25%	50%
Rating	245	8.2698	0.3881	4.8	8.1	8.2
Duration	245	127.0408	34.7544	3	107	127
Votes	245	717467.6857	600063.3788	9	225662	587116

This Streamlit application provides an interactive platform for exploring and analyzing IMDb movie data stored in a pre-processed joblib file. The app integrates advanced data filtering, statistical summaries, visualizations, and machine learning to enhance user engagement and insight discovery.

1. Data Loading and Caching

- Uses joblib to load preprocessed movie data (IMDB_data.joblib), avoiding the need to rerun scraping or cleaning scripts.
- Streamlit caching (@st.cache_data and @st.cache_resource) boosts performance by preventing redundant operations (like model retraining).

2. Sidebar Filters for Dynamic Exploration

- Users can filter the dataset by:

o Release Year
o Genre
o Rating

- A live summary shows the number of matching movies and selected criteria.

3. Tabs Overview

Homepage: Displays quick insights like top-rated movies, recent movies, genre stats, and a pie chart for duration distribution.
Data Analysis: Contains expandable sections with detailed visualizations:
o Rating Trends: Histograms, top/bottom rated movies.
o Release Year: Distribution and rating evolution over time.
o Votes & Duration: Scatter plots showing correlations.
o Genre: Frequency, duration averages, and percentages.
o Correlation Heatmap: Highlights relationships between numerical features.
Interactive View: Full filtered dataset view and comparison of rating distributions between genres.

4. Visualization Tools

Combines Seaborn, Matplotlib, and Plotly to generate a mix of static and interactive plots. Designed with modern aesthetics, user-friendly layout, and real-time responsiveness.