



Nova

Movie Recommendation Systems

Group 6

Vedangi Sawant
Pooja Mahesh
Sneha Kalidindi
Antoni Korycki
Pranav Raj

Laying the Groundwork: Understanding the TMDB Dataset

Project Objective:

To build a personalized and scalable **movie recommendation system** using

- Demographic Filtering
- Content-Based Filtering
- Collaborative Filtering

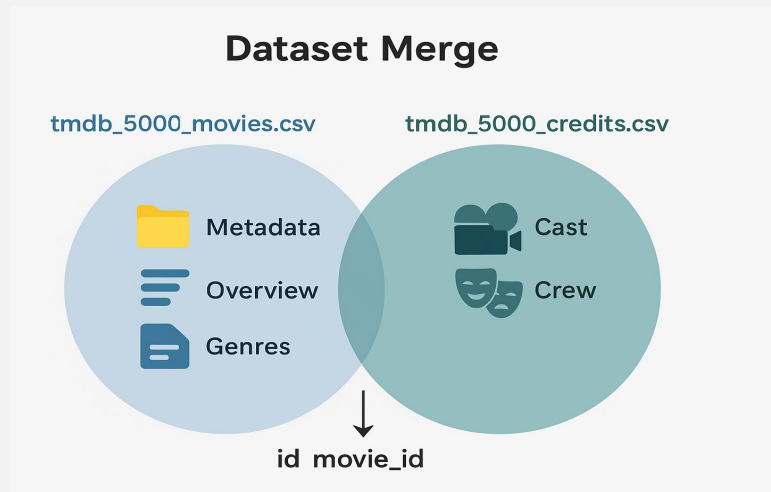
Dataset Utilized:

`tmdb_5000_movies.csv`

➤ Contains ~4,800 movies with info like **title, genres, overview, keywords, vote count**, etc.

`tmdb_5000_credits.csv`

➤ Includes **cast & crew** data such as actors, directors, and job roles.



Merge Strategy:

Join Key: `id` (from movies) and `movie_id` (from credits)

Ensures each movie has **complete contextual data** (metadata + cast/crew)

Result: A unified dataset with 20+ rich features per movie, ready for analysis

From Metadata to Meaningful Features

Features Extracted from the Merged Dataset

- **Genres** – parsed and extracted as a clean list (e.g., ['Action', 'Adventure'])
- **Overview** – cleaned movie descriptions for use in text similarity
- **Keywords** – thematic tags (e.g., ['culture clash', 'alien'])
- **Director** – extracted from crew where job == 'Director'
- **Top 3 Cast Members** – selected based on screen time/order in cast

Tools & Techniques Used

- `ast.literal_eval()` to parse JSON-like strings
- Python functions to loop, filter, and transform feature fields
- Prioritized **compact and informative** features for downstream models

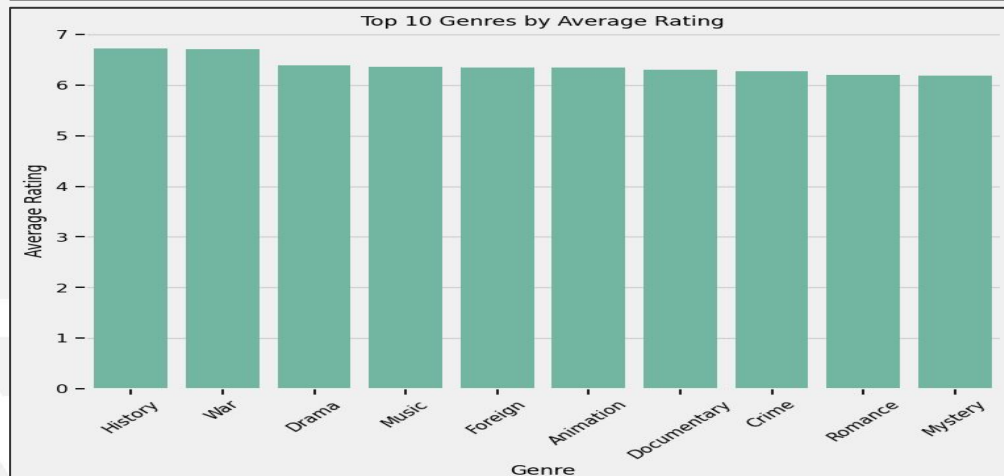
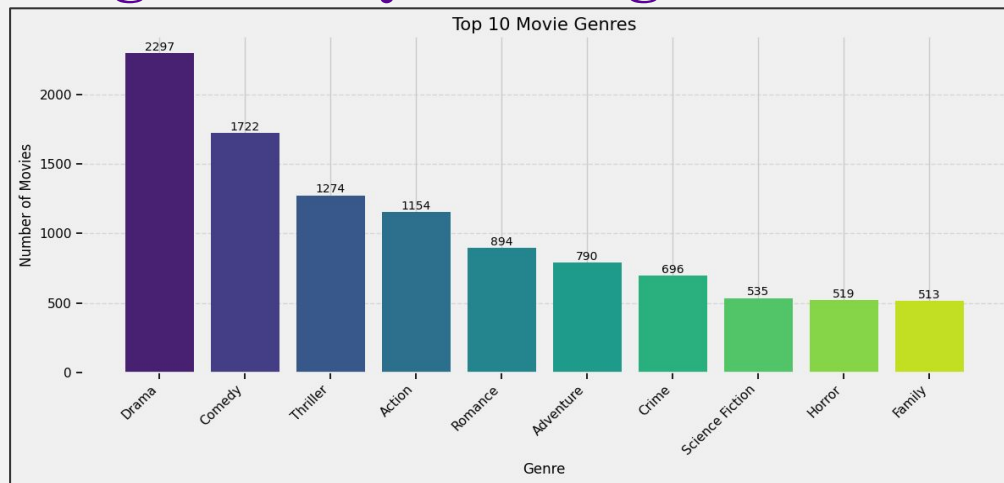
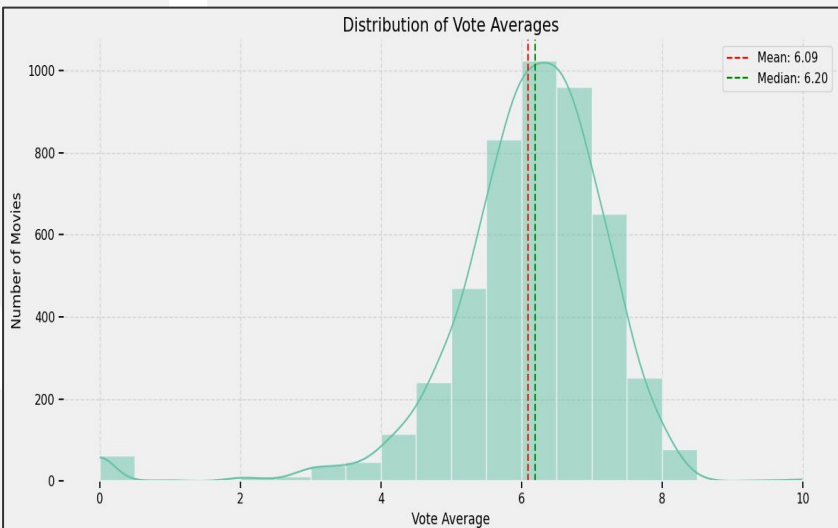
Why This Matters? These features are the **core inputs** to our content-based and hybrid models. They reflect the semantic and visual essence of each movie — enabling more personalized and accurate recommendations.

Visualization and Insights : Key Findings



Insight:

- **Most movies are rated between 6.0–6.5**, showing generally positive audience sentiment.
- **Drama and Comedy** are the most produced genres, dominating the dataset by volume.
- **History and War** films receive the highest average ratings, despite lower production counts.



Visualization and Insights : Weighted Average

A common approach to rank movies by a weighted rating formula:

$$\text{Weighted Rating (WR)} = (v/(v+m)) * R + (m/(v+m)) * C$$

Where:

R = average rating for the movie

v = number of votes for the movie

m = minimum votes required to be considered

C = mean vote across all movies

We use this for the demographic filtering which involves recommendations based on general popularity and basic info

Average vote across all movies (C) = 6.094458333333334

Minimum votes required (m) = 236.0

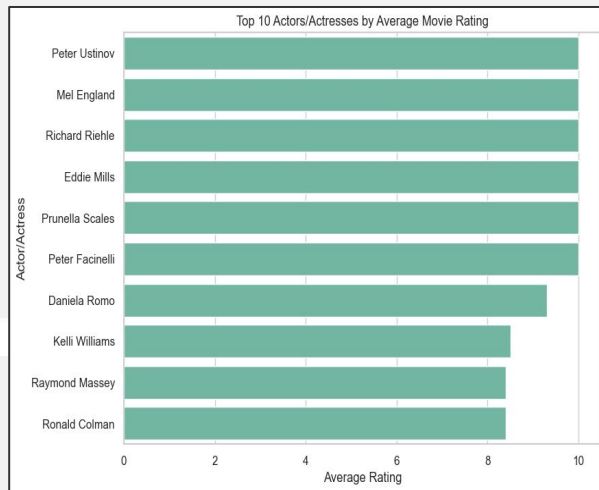
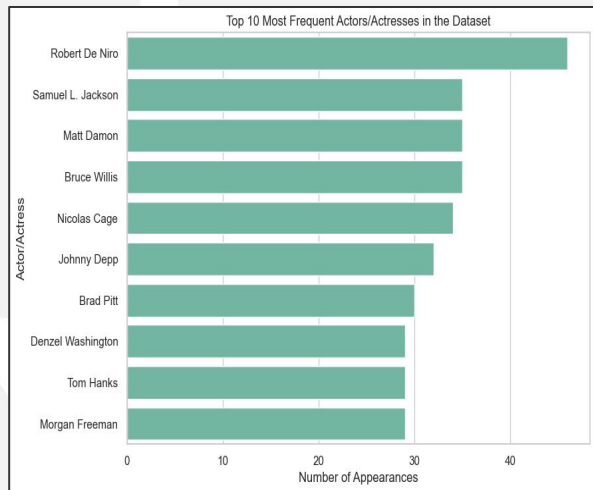
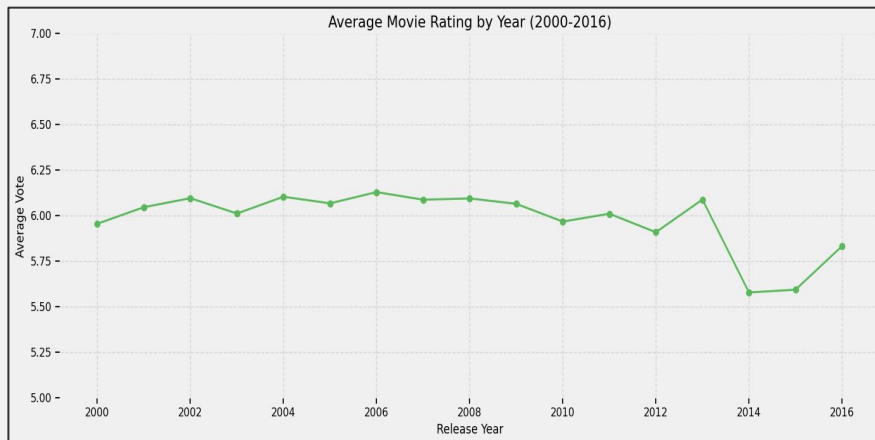
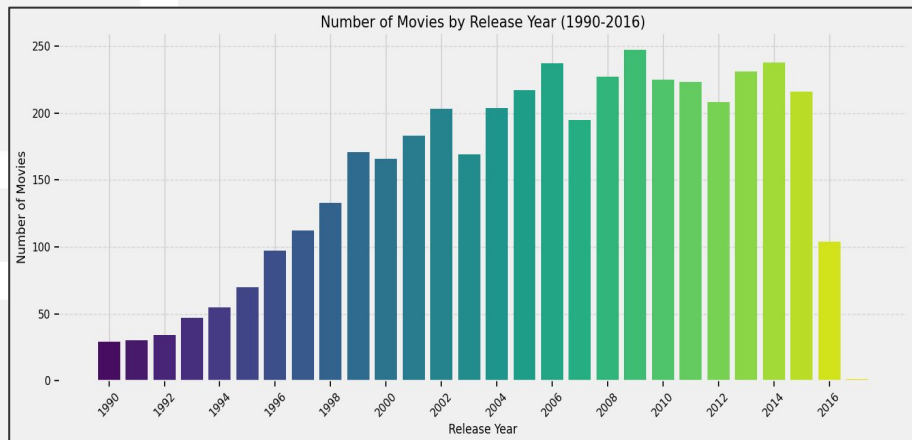
=== Top 10 Movies by Weighted Rating ===

	title	vote_count	vote_average	score
1881	The Shawshank Redemption	8205	8.5	8.432744
3337	The Godfather	5893	8.4	8.311224
662	Fight Club	9413	8.3	8.246056
3232	Pulp Fiction	8428	8.3	8.239923
1818	Schindler's List	4329	8.3	8.185979
3865	Whiplash	4254	8.3	8.184074
2294	Spirited Away	3840	8.3	8.172299
65	The Dark Knight	12002	8.2	8.159396
2731	The Godfather: Part II	3338	8.3	8.154363
809	Forrest Gump	7927	8.2	8.139127

Top 10 Directors by Average Rating (with >100 total votes)



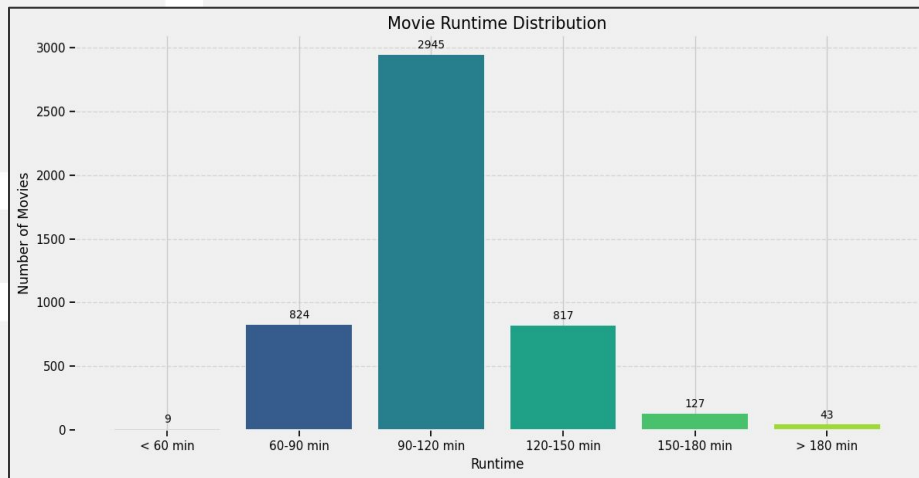
Trends in Movie Releases, Ratings, and Actor Impact






Insight:

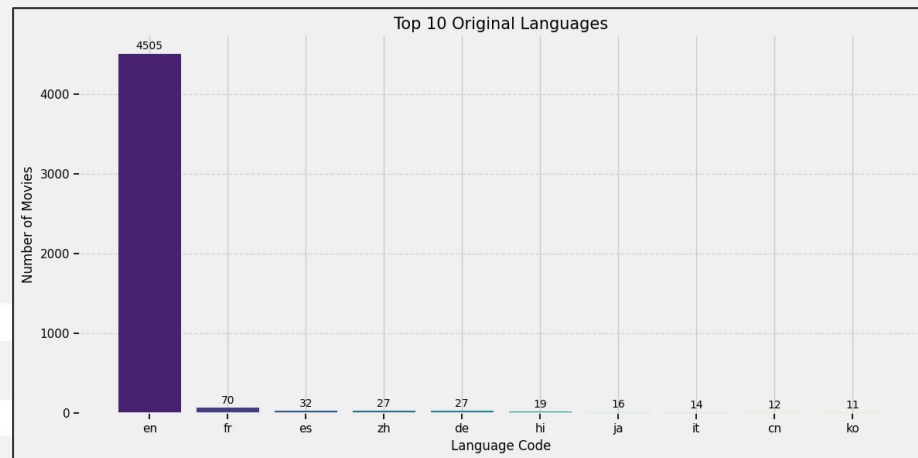
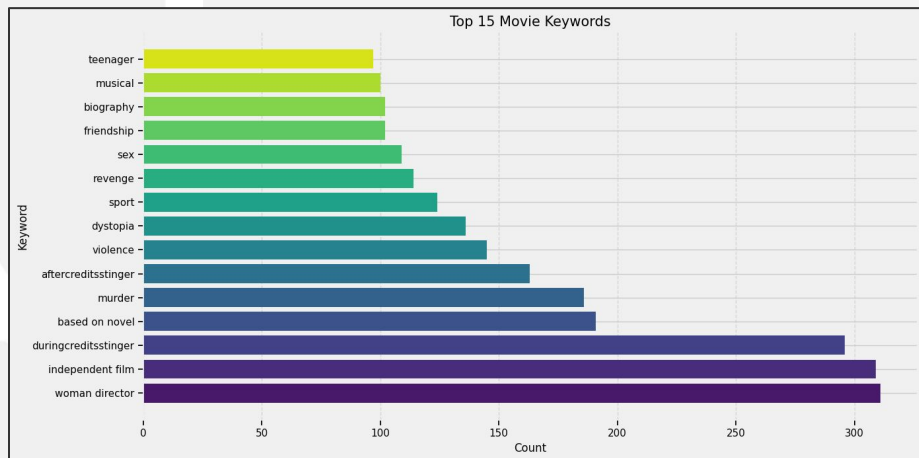
- 🎬 **Movie production peaked around 2010–2015**, showing a steady rise from the 1990s.
- ★ **Average ratings remained stable (~6.0–6.2)**, with a slight dip in 2014.
- 👤 **Robert De Niro and Samuel L. Jackson** were among the most frequent actors, while **Peter Ustinov and Mel England** starred in the highest-rated movies on average.

Patterns in Movie Runtime, Language, and Themes



Insight:

-  **Most movies are between 90–120 minutes**, aligning with standard theatrical runtime expectations.
-  **Popular keywords** highlight frequent themes like *"based on a true story"*, *"hero"*, and *"woman director"*.
-  **English dominates** as the primary production language, with very few films produced in other languages.

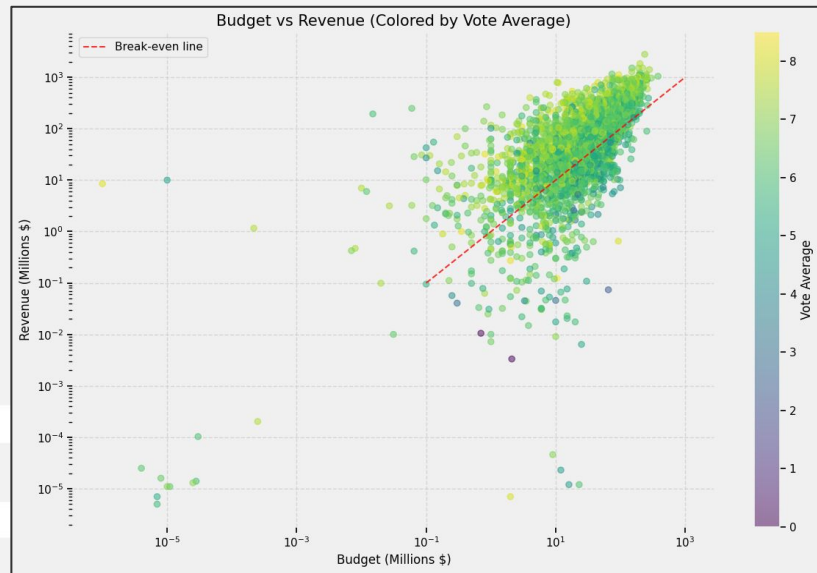
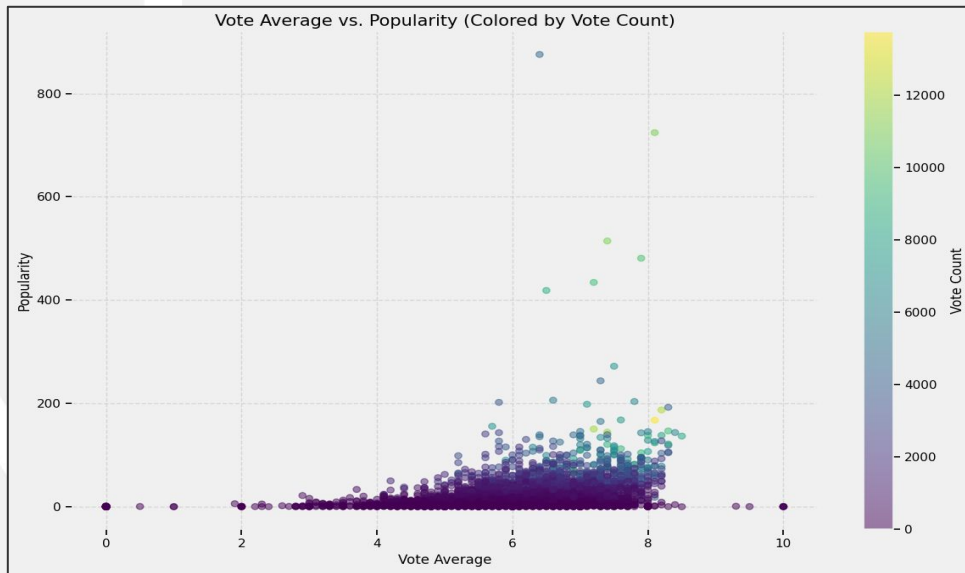
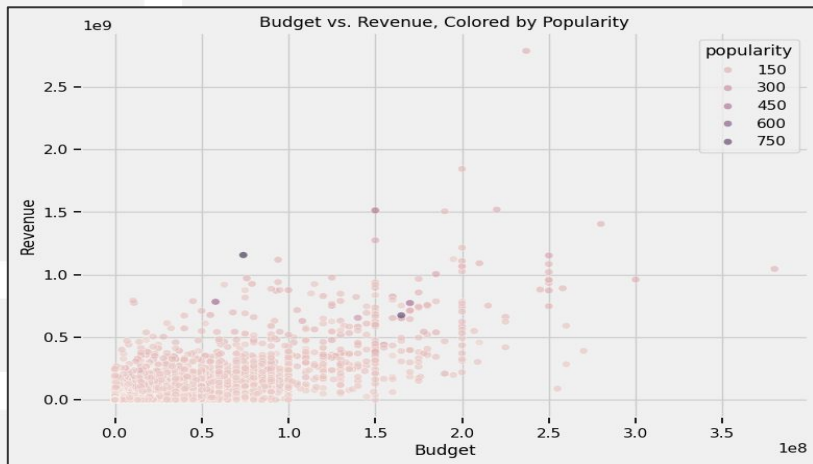


Exploring Relationships Between Budget, Ratings, and Popularity






Insight:

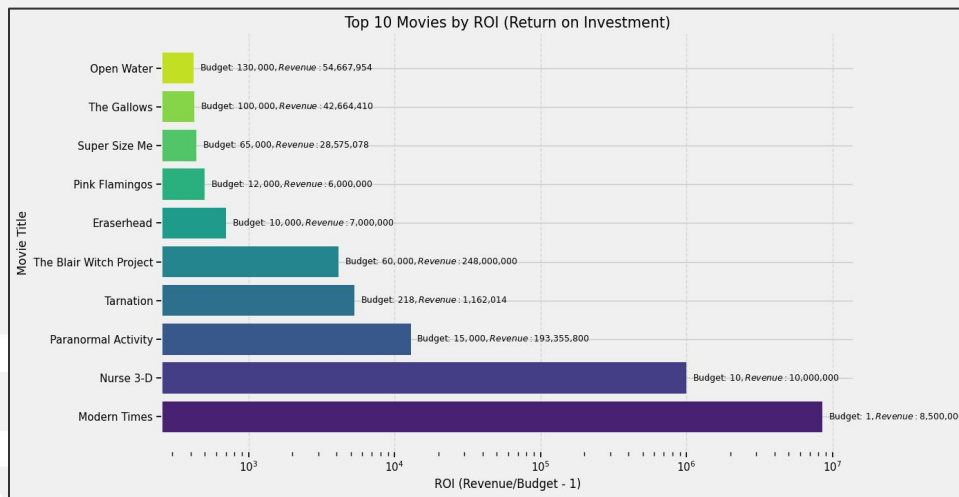
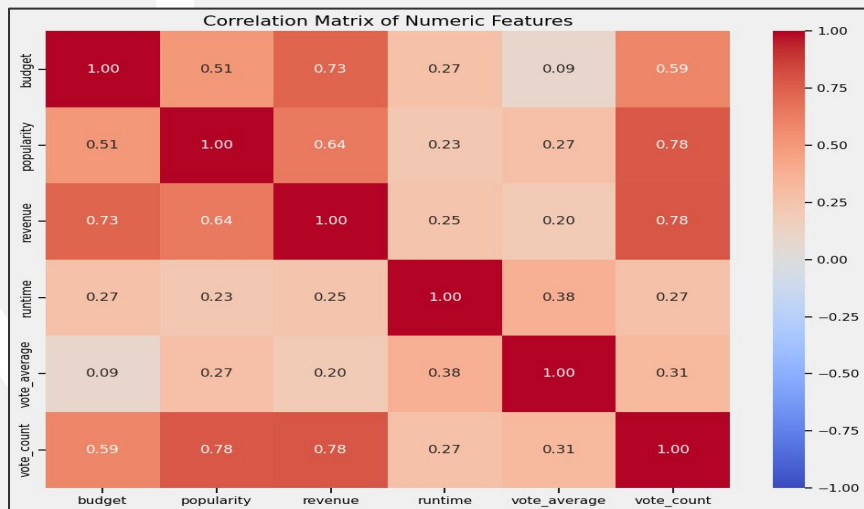
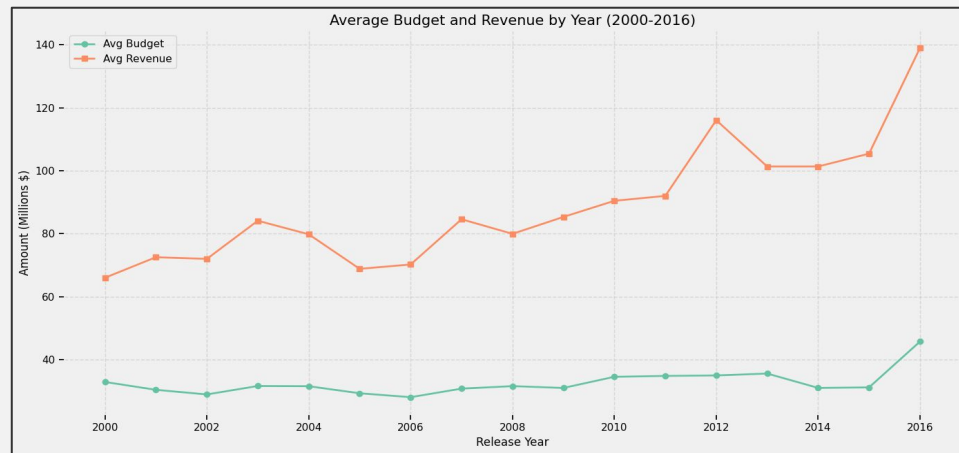
- 💰 **Higher budgets tend to yield higher revenues**, especially for popular films — but there are clear outliers with low returns.
- ★ **Movies with higher vote averages generally attract more popularity**, but a strong vote count is even more predictive.
- 📈 Most profitable films cluster **above the break-even line**, showing that **well-rated, mid-to-high budget films perform best financially**.



Financial Trends, Correlations & High-ROI Films

Insight:

-  **Budgets have increased sharply** over the years, while average revenue remained relatively flat — highlighting rising production costs.
-  **Strong correlations** exist between popularity, revenue, and vote count — but **ratings** are less correlated with commercial success.
-  Movies like **Modern Times** and **Paranormal Activity** delivered **exceptionally high ROI**, proving low-budget films can yield massive returns.



Demo of the Implementation of the various prediction Models and the output on our User Interface



Thank You