# CS661 Project: Analysis of TMDB Movie Dataset

Swarnim Verma Om Bhartiya Saaumitra Raaj Harshita Awasthi
Aryamann Srivastava Kshitiz Tyagi Aryan Deo Tejas Shrivastava

**Abstract**

Movies play a vital role in reflecting society, and influencing global audiences. This project presents a comprehensive analysis of The Movie Database dataset to uncover patterns and insights within the film industry. Using rich metadata—including genres, ratings, popularity, budgets, revenues, and cast—we conduct exploratory data analysis and visualizations. Our aim is to identify cinematic trends, key success factors, and the industry's evolution. Leveraging Python and its visualization libraries, we build an interactive dashboard that highlights correlations between production elements and commercial performance, offering actionable insights for filmmakers, distributors, and analysts.

## Introduction

In the digital age, the global film industry has transformed into a data-rich domain where every release contributes to a vast and growing pool of structured information. From box office figures to audience reviews, the movie ecosystem offers a compelling opportunity to study cultural shifts, audience behavior, and the commercial dynamics of storytelling. The TMDB (The Movie Database) platform curates an extensive and publicly accessible collection of metadata on films across languages and countries, making it an ideal foundation for visual data exploration.

Our primary objectives in this analysis are to:

- Understand how genre preferences have shifted across time and geographies.
- Explore correlations between factors such as popularity, vote average, runtime, and revenue.
- Apply data cleaning, transformation, and exploratory analysis techniques to extract meaningful insights from the dataset.

## Dataset Description

The dataset used is a secondary source from Kaggle, enriched via the TMDB API. It offers comprehensive details on global film releases, including genres, audience feedback, production data, and financial metrics. Comprising a primary CSV file with auxiliary JSON fields (e.g., genres, countries), it has been preprocessed for analysis. The project combines static CSV data with dynamic API data to provide a complete view of movie attributes..

| Column Name | Description |
|---|---|
| id | Unique movie identifier |
| title | Movie title |
| release_date | Release date of the movie |
| runtime | Movie duration in minutes |
| budget | Production budget in USD |
| revenue | Total revenue generated in USD |
| popularity | Popularity score assigned by TMDB |
| vote_average | Average user rating (scale of 0–10) |
| vote_count | Total number of user votes |
| genres | List of genres (in JSON format) |
| production_countries | Countries involved in production |
| spoken_languages | Languages spoken in the movie |
| keywords | Descriptive keywords or themes |
| status | Movie status (e.g., Released, Post Production) |
| original_language | Original language of the film (ISO code) |

## Tasks and Methodology

To meet our objectives, we structured the analysis into a series of technical tasks, combining data wrangling, exploration, and interactive visualization.

### Data Cleaning and Integration

- Merge CSV data with TMDB API responses.
- Flatten JSON fields like genres and keywords.
- Remove or flag entries with zero or missing revenue, budget, or runtime.
- Normalize and preprocess columns for comparison.

### Exploratory Data Analysis (EDA)

- Summarize distributions of key variables (revenue, popularity, rating).
- Identify correlations among numerical variables.
- Understand genre frequency and rating distributions over time.

### Visual Analytics

- Use libraries like Plotly, Seaborn, and Matplotlib.
- Create interactive charts with hover and filter functionality.
- Combine plots into a unified dashboard using Dash.

### Description of Analytical Tasks

1. **Genre Trends Over Time**: Streamgraphs to illustrate how genre popularity changes across decades, highlighting cultural shifts in cinema.

2. **Revenue vs Popularity vs Rating**: A 3D bubble chart with axes for average vote, popularity, and bubble size for revenue — identifying what makes a commercially and critically successful film.
3. **Runtime Analysis**: Scatter and line plots to examine how runtime relates to rating and revenue, revealing optimal durations for engagement and success.
4. **Keyword Themes**: Bubble charts show the frequency of keywords, and Sankey diagrams visualize relationships between themes, genres, and audience feedback.
5. **Geographical Production**: Choropleth maps and bar charts compare movie output and ratings by production country, offering insights into global cinema hotspots.
6. **Clustering and Dimensionality Reduction**: PCA and t-SNE reduce multi-feature datasets into 2D clusters to group similar movies and uncover hidden structures in genre, length, and popularity.
7. **Interactive Dashboard Development**: A web-based dashboard built with Plotly Dash allows users to explore and filter data visually in real-time, providing an engaging tool for discovery.

### Dashboard and Software Stack

We will build an interactive dashboard using:

- **Plotly, Dash**: For interactive graphs and app deployment.
- **Pandas, NumPy**: Data cleaning and transformation.
- **Seaborn, Matplotlib**: Static visualizations.
- **t-SNE, PCA (sklearn)**: Clustering and dimensionality reduction.

## Preliminary Experiments

We conducted preliminary visualizations to understand the dataset:

- A bar chart showing most popular genres across the dataset.
- Heatmaps showing correlation between vote_average, popularity, and revenue.

More visualisations will follow after further feature engineering and cleaning.

## Division of Work

| **Task** | **Members Responsible** |
|---|---|
| Keyword Bubble, Frontend | Swarnim, Om, Tejas, Aryan |
| Data Cleaning, Feature Engineering | Saaumitra, Kshitiz, Om, Harshita |
| Clustering, Runtime Trends | Aryamann, Tejas, Saaumitra, Swarnim |
| Revenue and Popularity Metrics | Harshita, Aryan, Kshitiz, Om |

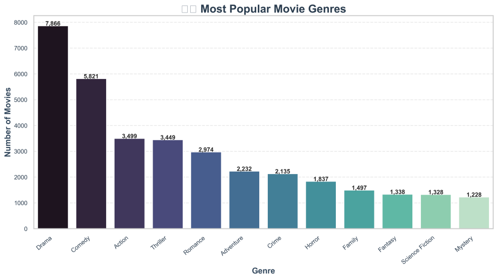

Figure 1: Most Popular Genre

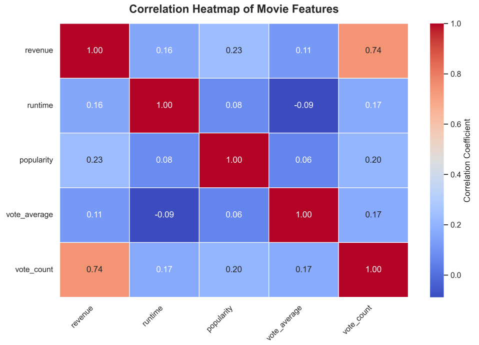

Figure 2: Feature Heatmap

# Conclusion

We aim to leverage data visualisation to understand the hidden dynamics of the movie industry. With the TMDB dataset, we will explore audience preferences, production trends, and profitability. Our interactive dashboard will aid creators and analysts in exploring cinematic data intuitively.

# References

- **TMDB API Documentation**
- **Plotly Dash Documentation**
- **TMDB Kaggle Dataset (2024) 1M movies**