1. Name, Team Members

Name: Movie Rating Prediction and Influential Factors Analysis Based on TMDB Data

Team Members:
Team Member 1: Rui Wang

- USC email: rwang671@usc.edu
- USC ID: 8179828591

Team Member 2: I - Chen Li

- USC email: ichenli@usc.edu
- USC ID: 2685490589

2. Short Description

This project uses the TMDB API to create a movie data analysis workflow to acquire movie data. It can automatically retrieve movie metadata from multiple pages, such as title, release year, rating, genre, director, and cast, primarily focusing on popular films. The project aims to collect movie information in a structured manner, writing it into a CSV file for subsequent work. Analysis can be performed from perspectives such as movie popularity, rating, genre, and the number of films the actor/director has appeared in. This is beneficial for researching movie trends, recommending movies, and understanding the film market.

3. Data

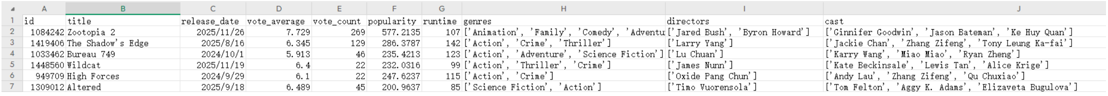
（1）Data Sources

Source Type	Description
TMDB Movie API	Used to get raw movie metadata such as rating, vote_count, release_date, cast, directors, popularity, runtime, etc.

（2）Data Columns

Column Name	Description
id	Unique identifier for each movie (Primary Key).
title	Official title of the movie.
release_date	Date when the movie was released (format: YYYY-MM-DD).
vote_average	Average user rating scored on a 0–10 scale.
vote_count	Number of individual user votes contributing to the rating.
popularity	Popularity index generated by TMDB based on user activity and interactions.
runtime	Duration of the movie in minutes.
genres	List of genres assigned to the movie (e.g., Action, Fantasy, Comedy).
directors	List of directors associated with the movie.
cast	Leading cast members (typically top-billed actors).

（3）Data Samples



	A	B	C	D	E	F	G	H	I	J
1	id	title	release_date	vote_average	vote_count	popularity	runtime	genres	directors	cast
2	1084242	Zootopia 2	2025/11/26	7.729	269	577.2135	107	['Animation', 'Family', 'Comedy', 'Adventu	['Jared Bush', 'Byron Howard']	['Ginnifer Goodwin', 'Jason Bateman', 'Ke Huy Quan']
3	1419406	The Shadow's Edge	2025/8/16	6.345	129	286.3787	142	['Action', 'Crime', 'Thriller']	['Larry Yang']	['Jackie Chan', 'Zhang Zifeng', 'Tony Leung Ka-fai']
4	1033462	Bureau 749	2024/10/1	5.913	46	235.4213	123	['Action', 'Adventure', 'Science Fiction']	['Lu Chuan']	['Karry Wang', 'Miao Miao', 'Ryan Zheng']
5	1448560	Wildcat	2025/11/19	6.4	22	232.0316	99	['Action', 'Thriller', 'Crime']	['James Nunn']	['Kate Beckinsale', 'Lewis Tan', 'Alice Krige']
6	949709	High Forces	2024/9/29	6.1	22	247.6237	115	['Action', 'Crime']	['Oxide Pang Chun']	['Andy Lau', 'Zhang Zifeng', 'Qu Chuxiao']
7	1309012	Altered	2025/9/18	6.489	45	200.9637	85	['Science Fiction', 'Action']	['Timo Vuorensola']	['Tom Felton', 'Aggy K. Adams', 'Elizaveta Bugulova']

The dataset consists of 101 rows and 10 columns, where each row represents an individual movie record retrieved from the TMDB Movie API, and each column corresponds to a specific movie attribute such as ratings, popularity, runtime, genres, and cast information.

4. Data Cleaning, Analysis & Visualization

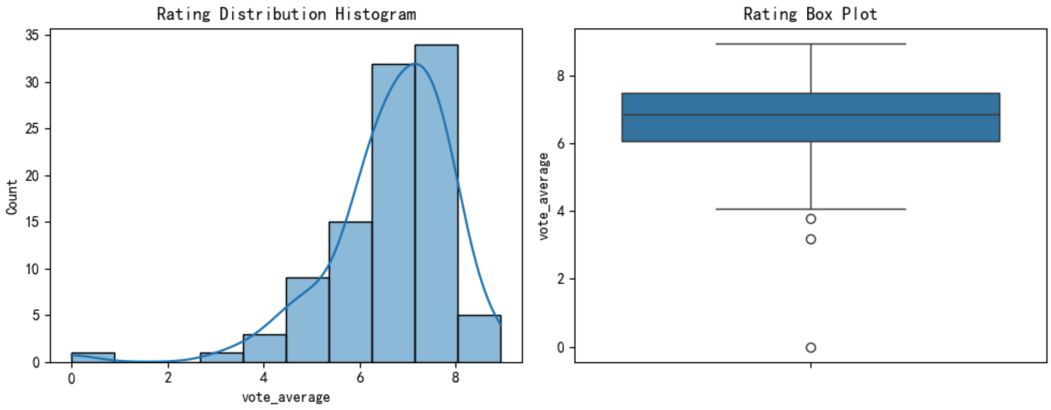
4.1 Data Cleaning and Preparation

Before conducting any analysis, the movie dataset obtained from the TMDB API was cleaned and organized to ensure consistency and usability. The raw data included attributes such as release date, rating, popularity, genres, directors, and cast members; however, several fields were stored in inconsistent formats.

Specifically, the *genres*, *directors*, and *cast* fields were originally represented as strings rather than iterable lists. These fields were converted into list-based structures to support statistical analysis and frequency counting. Additionally, the *release_date* field was transformed into a standardized datetime format to avoid potential errors in time-based operations.

After formatting, the dataset was examined for missing values. Records lacking core numerical attributes—such as *popularity*, *vote_count*, and *runtime*—were removed to prevent bias and instability during the analysis and potential model-training stages. As a result, a clean and well-structured dataset was obtained, suitable for further exploration and visualization.

4.2 Exploratory Data Analysis

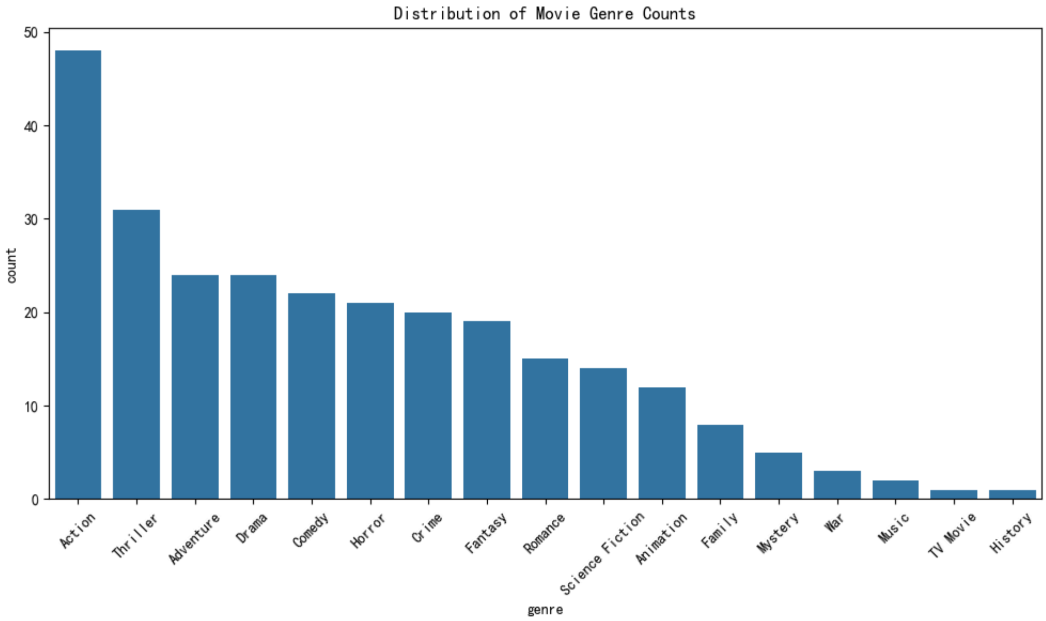


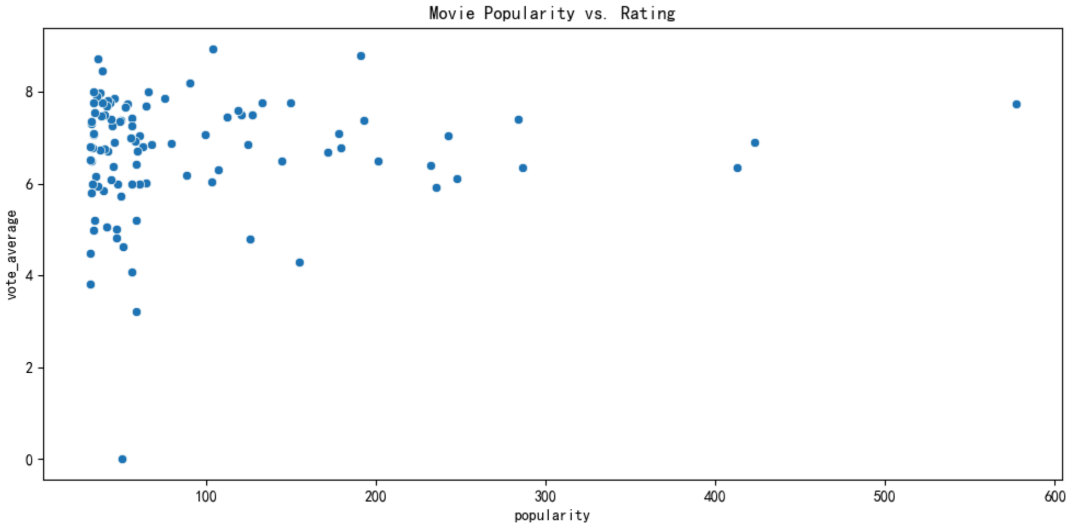
Once the data was cleaned, exploratory data analysis was performed to understand overall patterns and distributions.

The analysis began with the distribution of movie ratings. Histograms and box plots indicated that most ratings were concentrated between 6 and 8, suggesting that the majority of movies received generally positive but not extreme reviews. This implies that highly polarized ratings were relatively rare in the dataset.

Next, the frequency of different movie genres was analyzed. A bar chart was used to visualize the number of movies within each genre. The results showed that Action, Thriller, and Science Fiction were the most common genres, indicating higher production volume and stronger commercial maturity. However, high production frequency did not necessarily correspond to higher audience ratings, a finding further examined in later analysis.

4.3 Visualization and Relationship Analysis





To investigate factors influencing movie ratings, correlation analysis was conducted between ratings and numerical features such as popularity, vote count, and runtime. Scatter plots were used to visualize the relationship between movie popularity and ratings.

The visualizations revealed a positive relationship between popularity and ratings: movies with higher popularity generally tended to have higher ratings. However, the wide dispersion of data points suggested that popularity alone does not guarantee critical acclaim, indicating that other factors also play a role in determining audience evaluations.

5. Changes from Original Proposal

During the project, some parts of the research plan were modified. Initially, I wanted more data to allow for more feature variables and more detailed analysis. However, during data collection, I discovered some expected fields were missing, had different formats, or were inaccessible. Continuing with the previous methods would reduce the accuracy and readability of the results. Therefore, to ensure the quality of the analysis, the data scope was appropriately narrowed, time permitting. Some valuable and meaningful indicators were retained, and more time was spent on data cleaning to ensure data quality. Additionally, my initial plan included making more complex predictions and building models. However, due to the high cost of training and parameter tuning, data analysis and visualization methods were used to make the data clearer and easier to understand, thus making the entire research process more complete.

6. Mention of Future Work

Even though the data collection of this project is done well, as well as the data cleaning and visualisation of the data. But I think there is still a need for improvement and expansion. In the future, if there is more time or more computing resources, I would like to try the following things: The first thing is to increase the scale of the data, such as increasing the number of years of data, or adding more film platforms. this way, it will be more thorough and representative. Second, this article is also a kind of visualization, analysis and analysis, which can be used in the future with some machine learning model, or deep learning model like the box office prediction, rating prediction, or similar movie recommendation to do some intelligent analysis. Moreover, we can do more textual analysis like the text sentiment analysis based on films' reviews, we can analyze the relationship of audience's sentiments and rating, and thus reach more explanatory conclusions. And it can also become a dynamic data visualization system, allowing people to choose the different types of films, different film years, and different film directors for dynamic display, making this research result more practical.