1. Name of the project and team members

Title: Analysis of Key Factors for Movie Success Using TMDb Data

Yuchen Wu, email: ywu55711@usc.edu, USC ID: 9788517696, GitHub: ywu55711
Yang Zhao, email: yzhao657@usc.edu, USC ID: 7748154612, GitHub: yzhao657-lyz

2. What problem are you trying to solve?

Understanding what makes a movie financially successful or critically well-received is an important question for both the film industry and audiences. This project uses data from the TMDb API to identify data-driven patterns that explain how budget, revenue, popularity, ratings, and genre characteristics influence a film's success. The project focuses on the following research questions:

Budget vs. Revenue:

What is the relationship between a movie's production budget and its box office revenue? Is a higher budget a reliable predictor of financial success?

Popularity, Ratings, and Success:

How are a movie's popularity (TMDb's proprietary metric), its average user rating, and its financial revenue interrelated?

Genre Analysis:

How do different movie genres differ in their typical budget, revenue, and audience rating? Which genres tend to perform best financially or critically?

The Impact of Time:

Is there a correlation between a movie's release period (e.g., summer, holiday season) and its rating or popularity?

By focusing on these four questions, the project tries to provide an objective and comprehensive analysis of the measurable factors that contribute to both commercial success and audience reception in the movie industry.

3. How will you collect data and from where?

We will use the Python requests library to make HTTP GET requests to the TMDb API.

First, we will obtain movie titles. The specific data fields to be collected include title, release date, genres, production budget, revenue, popularity score, average user rating. The raw data will be stored in its raw JSON format and then cleaned and structured for analysis.

4. What analysis will you do and what visualizations will you create?

Data cleaning: Using Pandas to get dates, extract release year, convert budget/revenue to numeric values, movie genre, remove missing values, and produce a clean dataset for analysis.

Analysis: We will analyze the correlation between budget and revenue. Relationship among popularity, rating, and revenue. Additionally, comparisons of average budget, revenue, and rating by genre-level. And regression or trend analysis relating film release time to rating or popularity.

Visualization: We plan to use scatter plots for Budget vs. Revenue(Q1) and Popularity vs. Revenue(Q2). Bar chart for Average budget/revenue/rating per genre(Q2, Q3). Line plot for Average rating by release time(Q4). We also consider using a heatmap to find the Correlation between numeric variables.