1. Project Title:

"A Combined Ranking from IMDb and Box Office Mojo"

Team Members:
Zhengshu Zhang (USC ID: [4664738582], Email: [zhengshu@usc.edu], GitHub: [forestzs], Url:https://github.com/forestzs
Zhuoran Deng (USC ID: [5873397495], Email: [dengzhuo@usc.edu])

2. Problem Statement

Streaming platforms and movie fans often rely on different signals to decide what to watch, such as the movie's evaluation score and sales.
However, these multiple dimensions do not always agree. And in this project, we ask:To what extent do movie quality, user popularity, and box office success align, and how can we combine them into a single, reproducible ranking?We will build a small movie ranking system that integrates three different ranking signals from public web sources, and analyze where the rankings agree or disagree.

3. Data Sources and Collection Plan

I.IMDb Top 250 Movies – used as a proxy for long-term quality

II.IMDb Most Popular Movies (MOVIEmeter) – used as a proxy for current popularity

III.Box Office Mojo Weekend Box Office (US) – used as a proxy for box office performance

We are planning to use the Python crawler to obtain the data we need from these websites. The crawler will be scheduled to run at defined intervals to capture updates in popularity indexes like MOVIEmeter, ensuring temporal consistency in our dataset. After collection, all raw data will be stored in structured CSV formats and cleaned to remove duplicates or empty values, normalize movie titles across platforms. This standard process methods will allow us team to integrate long-term quality indicators, real-time popularity signals, and financial performance metrics into a unified data suitable for statistical analysis and modeling.

4. Planned Analysis and Visualizations

All three variables will be scaled to the 0–1 range using min–max normalization to make them directly comparable. To construct a single composite measure of overall movie performance, I will apply Principal Component Analysis (PCA) to the normalized variables and use the first principal component (PC1) as a data-driven performance index.For visualization, I will use:A correlation matrix heatmap to show pairwise relationships among rating, box office, popularity, and PC1.PCA plots (e.g., scree plot and a scatter plot of movies along PC1) to visualize how films are positioned on the combined performance dimension.