DSCI 510 Final Project Proposal

Project Topic:

In today's highly developed film and television industry, over ten thousand films of varying genres and lengths are released worldwide each year. Beyond enjoying the immersive experiences these films provide, we can't help but wonder: Do others share our opinions on a particular movie? Is there a specific genre that consistently garners the most widespread audience approval? What factors contribute to a film's overall positive or negative reception? Therefore, in this project, we've chosen to utilize data from IMDb, the world's largest online film database, to conduct a detailed exploration and analysis of this topic.

Team Members:

- Ester Hu (jhu05525@usc.edu)
- Qi Shen (qshen840@usc.edu)

1. What problem are you trying to solve?

Movie success is often discussed through subjective opinions, yet it is unclear which film characteristics most strongly correlate with higher audience ratings. This project aims to identify patterns in movie quality by examining how factors such as genre, runtime, release year, and vote count influence IMDb ratings. The main goal is to answer: What movie attributes consistently align with higher viewer ratings, and what trends can be observed across genres, time and release year?

2. How will you collect data and from where?

1. Scrape the IMDb Top 250 webpage
   - URL: https://www.imdb.com/chart/top
   - We will extract each film's title and IMDb ID from the static HTML list page.
2. Retrieve structured metadata using the OMDb or TMDB public API
   - Using each movie's IMDb ID, we will collect: Genre(s); IMDb rating; Vote count; Runtime;Release year;Additional metadata as available.

3. What analysis will you do?

RQ1 Genre & Rating: Do certain film genres consistently achieve higher IMDb ratings than others, and how do average ratings differ across major genres within the IMDb Top 250?

RQ2 Runtime & Rating: Is there a meaningful relationship between a film's runtime and its IMDb rating? Specifically, do longer films tend to receive higher ratings, or is there no consistent pattern across the Top 250?

RQ3 Release Year Trends: How have IMDb Top 250 movie ratings changed over time? Do older films, recent films, or films from certain decades show systematically higher or lower ratings?

RQ4 Popularity (Vote Count) vs Rating: To what extent does movie popularity, measured by vote count, correlate with audience ratings? What distinct patterns emerge among:

(1) highly rated films with very large vote counts (mainstream classics),

(2) highly rated films with relatively low vote counts (niche but critically acclaimed films), and

(3) highly popular films with only moderate scores?

4. What visualizations will you create?

(1) Bar Chart - Average IMDb Rating by Genre

This chart will compare the mean ratings of major genres in the IMDb Top 250 to identify which categories consistently receive higher scores.

(2) Scatter Plot - Runtime vs Rating

A scatter plot will reveal whether longer films tend to earn higher ratings, and whether runtime shows any noticeable positive or negative trend with rating.

(3) Line Chart - Rating Trends by Release Year or Decade
This chart will show how average ratings change over time, helping us observe decade-level or year-level patterns such as “classic films vs modern films.”
(4) Bubble Scatter Plot - Vote Count vs Rating
A bubble plot will display the relationship between vote count (popularity) and rating, where bubble size represents vote volume. This helps us compare:

- highly rated mainstream films (large bubbles + high ratings),
- niche films (small bubbles + high ratings), and
- popular but moderately rated films.