DSCI 510 Final Project Proposal

Fall 2025

Project Name:

Predicting Box Office Success: An Analysis of Movie Performance Factor

Team member (solo)
Chaoyang (Sunny) Yin
USC ID: 7522481537
USC email: sunnyyin@usc.edu
Github: ChaoyangYin

Problem to be solved
The film industry invests billions each year, yet many movies fail to generate profit. This project aims to investigate which factors, such as genre, bugdet, production, release time, and especially opinions, can actually predict box office revenue. An interesting yet important focus is the public opinion factor; Professional critics and audience can often disagree on the value of movies. This particular aspect may have influence on movie's box office results in different ways. By analysing the critics ratrng and audience scores, alongside with core attributes of movie, this project will try to identify which factors affect box office success the most, and uncover how opinion dynamics contribute to commercial outcomes.

Data Collecion
This project will collect form two publicly available APIs.
TMDB API will provide data on the core attributes of the movies across multiple years using the */discover/movie* endpoint. TMDB also provides each film's IMDb ID, which allows for comparision and merging outside the dataset.

The OMDb API, which returns Rotten Tomatoes critic scores, Rotten Tomatoes audience scores, and Metascore values, will be the primary data source for the opinion dynamics study. This API can be queried by IMDb ID for merging with the TMDB results. For each movie obtained from TMDB, the corresponding OMDb entry will be requested and parsed.

Data Analysis:
The data acquired from API will first be cleaned by parsing dates, trimming genres and filtering outliers. Additional derived attributes can be created, such as caculating crtitic_audience_gap, by incorporating multiple attributes.
Cross-variable analysis will then be performed to evaluate how factors like budget, genre, runtime, critic scores, and audience ratings relate to revenue and ROI. This includes correlational analysisand trend studies across years and genres.
Visualizations will be created using Matplotlib to show important patterns, including scatter plots, bar charts, and line charts that highlight trends, correlations, and differences across genres and rating sources. These plots will support the analysis by clearly showing how various factors relate to revenue, ROI, and rating differences.