

W203 Lab 02 Research proposal

Datasci 203, Section 10, Group 2, Lab 2

Divya Menghani, Hamsini Sankaran, Israel Ayode, Sivakumar Thiagarajan

The movie industry is a huge multi-billion dollar business, with movies being created at various budgets ranging from low to high. One of the major factors that determine the success of a movie is its revenue. In this proposal, we aim to investigate the relationship between the “Budget” and the “Revenue” of the movie.

1. What is the research question? Specifically, what is the X concept and what is the Y concept?

- *Research Question: How does the budget of a movie influence its revenue?*
- *X Concept: Budget - budget (metric)*
- *Y Concept: Revenue - revenue (metric)*

Other variables that influences the X concept “Budget” are:

- Movie length (runtime) - No. of Cast and Crew (derived from ‘credits’ column) - Genre (genres)

Other variables that influences the Y concept “Revenue” are:

- Movie length (runtime) - Vote count (vote_count) - Movie title length (no. of characters in movie title) - No. of Cast and Crew (derived from ‘credits’ column) - Genre (genres) - Release Date

2. Who is the actor who can change your X concept?

Movie investors are the key actors who can influence the X concept, which is the budget of the movie.

3. Who is the audience who would care about changes in the Y concept?

Movie investors are the key actors who can influence the Y concept, which is the revenue of the movie.

4. What is the data source? What variables will you use to operationalize X and Y?

The dataset is obtained for “kaggle” website. The dataset comprises of metadata for more than 700,000 movies listed in the TMDB Dataset. It is relevant and provides an opportunity to explore and analyze the relationship between the budget and the revenue of the movies. Since the dataset is unique on movies, each movie is independent from one another and uniquely identified by an id column in the dataset. The movie details, keywords and credits are collected from the TMDB open API. We operationalize the variables X and Y as indicated below:

- Operationalization of X: Budget (budget) We operationalize the budget in terms of \$ amount.
- Operationalization of Y: Revenue (revenue) We operationalize the revenue in terms of \$ amount.

We also need to consider the impact of omitted variables such as the movie’s storyline, competition from other movies released on the same date, and the film’s rating given by the Motion Picture Association (MPA).

5. What is the unit of observation? That is, does each row of the data represent a person, a review, a hotel stay, or something else?

Each row of data in the dataset is uniquely identified by a movie.

- Budget is measured by \$ amount
- Revenue is measured by \$ amount
- Genre represents movie category
- Movie Length is represented as count
- Vote count is represented as count
- Movie title length is represented as count
- No. of Cast and Crew is represented as count

