W203 Lab 02 Research Proposal - Team 02

Divya Menghani, Hamsini Sankaran, Israel Ayode, Sivakumar Thiyagarajan

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. Below are the research question, the X and 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)*

The variables that influences the X concept "Budget" are the Movie length (runtime), No. of Cast and Crew (derived from 'credits' column) and Genre (genres). Likewise, the variables that influences the Y concept "Revenue" are the 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) and the Release Date. The Movie investors are the key actors who can influence the both the X and the Y concept.

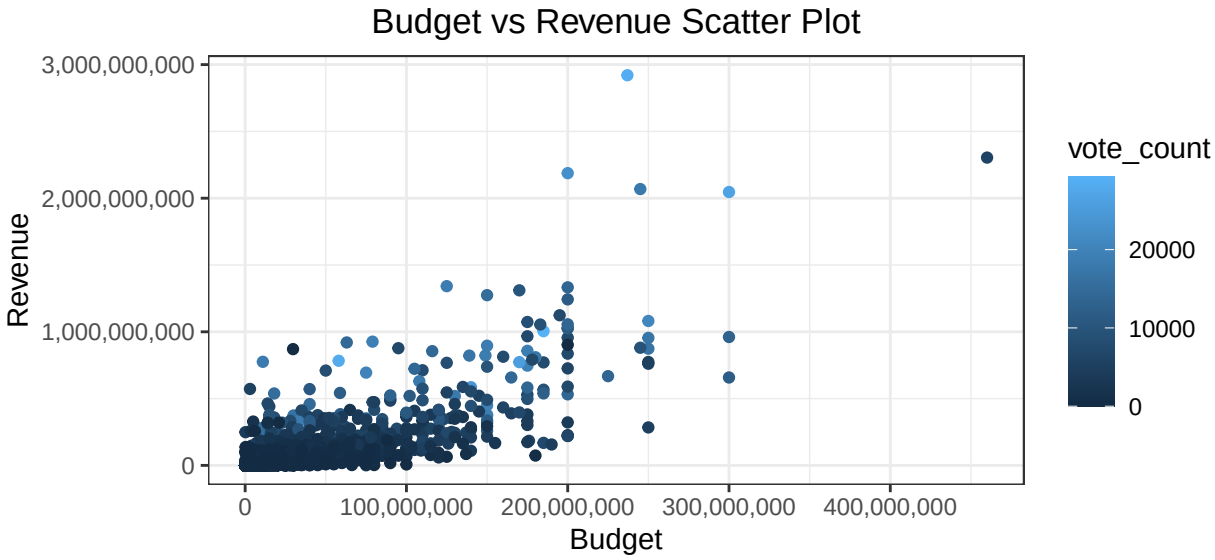
The dataset is collected from kaggle. The fields like the movie details,keywords and credits are collected from the TMDB open API.The dataset comprises of metadata for more than 700,000 movies listed in the TMDB Dataset. It is relevant and provides an opportunity to explore and analyze the relationship 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.We operatuonalize 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 budget 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).

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


Budget vs Revenue Scatter Plot
Revenue
3,000,000,000
2,000,000,000
1,000,000,000
0
0
100,000,000
200,000,000
300,000,000
400,000,000
Budget
vote_count
20000
10000
0