W203 Lab 02 Research Proposal - Team 02

Ayoade Israel, Divya Menghani, Hamsini Sankaran, Sivakumar Thiyagarajan

The movie industry is a huge multi-billion dollar business, with movies created at various budgets ranging from low to high. However, one of the significant factors determining a movie's success is its revenue. In this proposal, we aim to investigate the correlation between the "Budget" and the "Revenue" of the movies. Using an exploratory dataset, we found a graphical relationship between the fields "Revenue" and "Budget" of the movies. Below are the research question, the X concept and the Y concept.

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

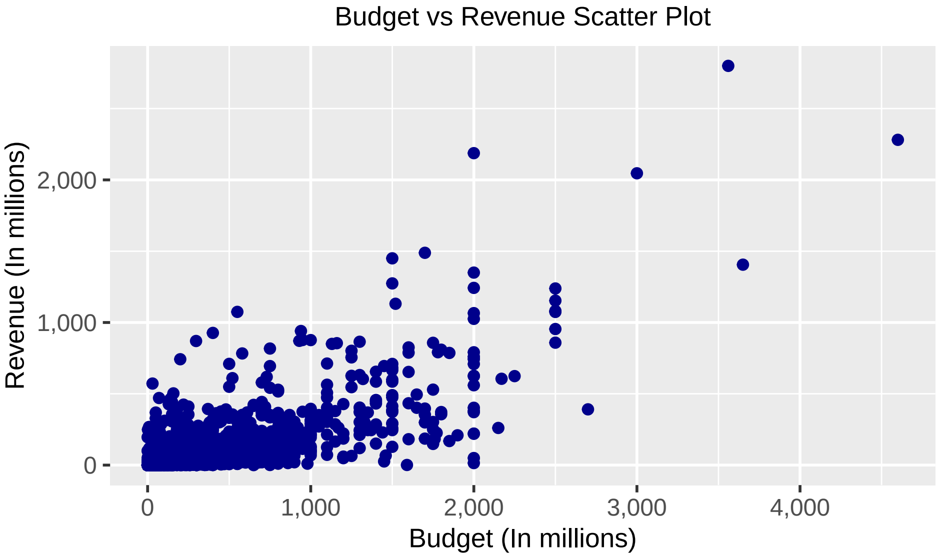
The key actor of the X concept and the intended audience of the Y concept is the movie production firms. These companies are motivated to enhance their movie revenue (Y concept), and they hold authority over the X concept, which pertains to the budget. The covariates influencing the Y concept "Revenue" are the Number of Cast and Crew (derived from the 'credits' column), Movie length (runtime), Number of production companies (derived from the column production_companies), Genre (genres), and the Release Month(derived from the field release_date).

We gathered the dataset from Kaggle. The dataset comprises metadata for more than 700,000 movies listed in the TMDB Dataset. It has 722986 rows and 20 columns. It is relevant and provides an opportunity to analyze the relationship between the budget and the revenue of the movies. The dataset contains unique information on films, with each movie being distinct from the others and identified by an id column in the dataset. Additionally, the distribution of the X and Y concept is identical, satisfying the assumptions of IID. Because each movie is not subject to multiple samplings, the dataset is regarded as cross-sectional.

We operationalize the variables Budget(X) and Revenue(Y) 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). The unit of observation of the X concept, Y concept, and the covariates are indicated in the below table.

Budget	**Revenue**	**Genre**	**Movie Length**	**Number of Cast and Crew**	**Number of production companies**	**Release Date**
$ amount	$ amount	movie category	minutes	count	count	Date

We also plotted a correlation heatmap and observed that our X concept is having strong correlation with the Y concept when compared to other covariates. The below plot shows the relationship between the X concept (Budget) and Y concept (Revenue):


Budget vs Revenue Scatter Plot
Revenue (In millions)
Budget (In millions)