IMDB MOVIE DATA ANALYSIS

Everybody loves watching movies, isn’t it right? It transports us into a world where anything and everything is possible. Not just that, we get to see amazing talent in terms of acting, direction, visual effects etc. But sometimes, a movie that seems to be very promising fails to do well at the box office, the opposite is also true. And that makes you wonder what is really required for a movie to be successful. Is it the actor or the Director? Is it the Genre? What about the popularity of the actors and the movie budget? How much do critic ratings affect the movie? These are the questions that we try to answer in this project.

We started by performing basic data cleaning of the dataset (Source: Kaggle) - removing redundant entries, deleting missing entries, converting categorical data to numerical, after which we tried to find a way to rank the actors and directors by integrating both the qualitative and quantitative data. We kept IMDB score as the reference and divided this score into four quantiles of 25% each. We then counted the number of movies each director had in every quantile and multiplied it with an index to get the rank. We did the same for actors but we also considered their popularity (Number of Likes on their Facebook Page). Once this was done, it was a lot easier to analyze the data. The “genres” were given as tab separated values and were also categorical. We had nine distinct genres, where we represented each of them by a unique number. The genre for a movie is calculated as follows:

We defined the categories as {'Action':1, 'Adventure':2, 'SciFi':3, 'Romance':4, 'Comedy':5, 'Horror':6, 'Drama':7, 'Thriller':8, 'Documentary':9}

So, if a movie falls under Action, Adventure and Romance, it will be represented as 124.

The main takeaways for me from this project were:

- Prioritizing the work.
- Walking through all the final details with teammates before starting the project to avoid issues when merging the final work.
- Organizing code into modules so that it is easier to modify and read.
- Learnt the basics of data analysis and front-end development.

While working with the dataset, we found several flaws, which if rectified would make the results much better. The dataset had more of English movies so we were not able to analyze if language was important to a movie’s success. So, I would start by creating my own dataset by scraping websites, making sure I had equal number of movies from each language and making sure none of the entries were missing.

Since we had data of movies from 1920 until 2016, some movies did not have their Budget in USD. To take these movies into consideration, we also need to consider the average exchange rate of the currency in the year the movie was released. We had omitted these movies in our analysis but I would like to make this conversion and include them.

Lastly, I would make the final front-end interface more interactive. Our interface was interactive but we could have made it better by adding effects for mouse-clicks and mouse-hovers in all the visualizations.

From this project, we concluded that a movie is more likely to be successful if it has a popular actor and a good director, a budget of 40 – 80 million USD and it is in a popular genre. Critic ratings do not seem to affect the movie’s success if the movie has a popular actor.

Manisha Mishra

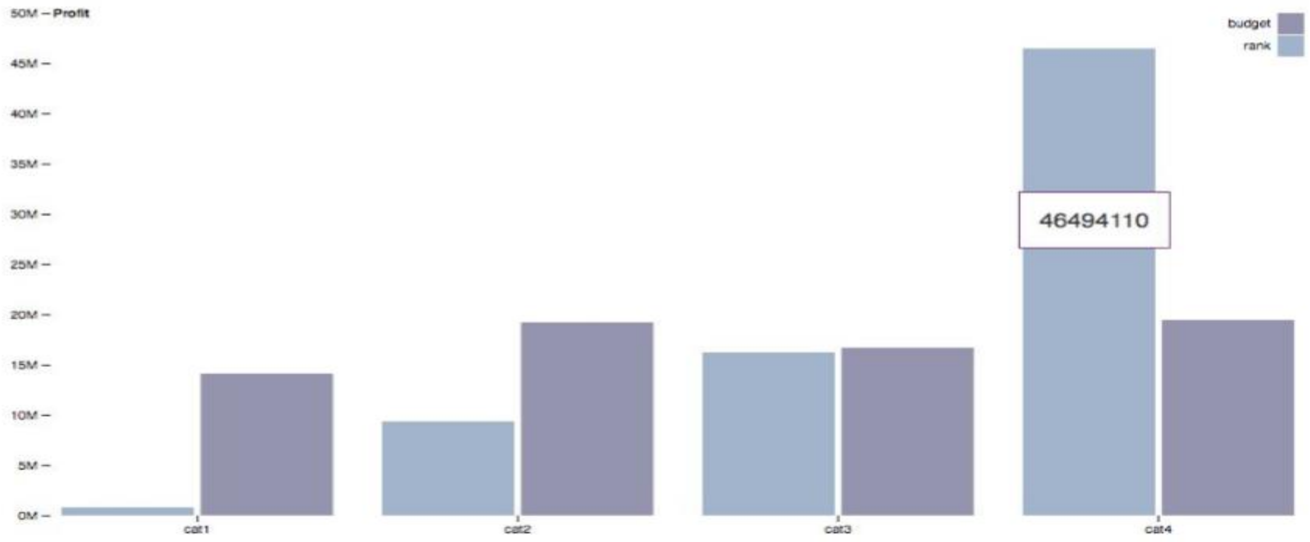

Figure 1. Quality (IMDB Score) vs Budget

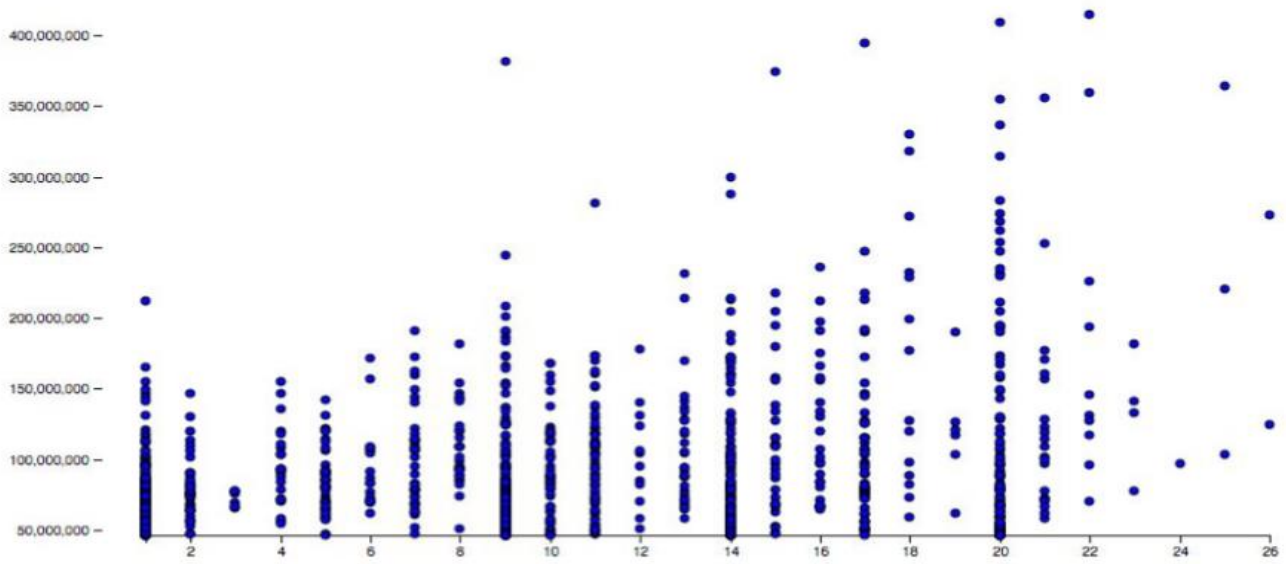

Figure 2. Director Score vs Gross

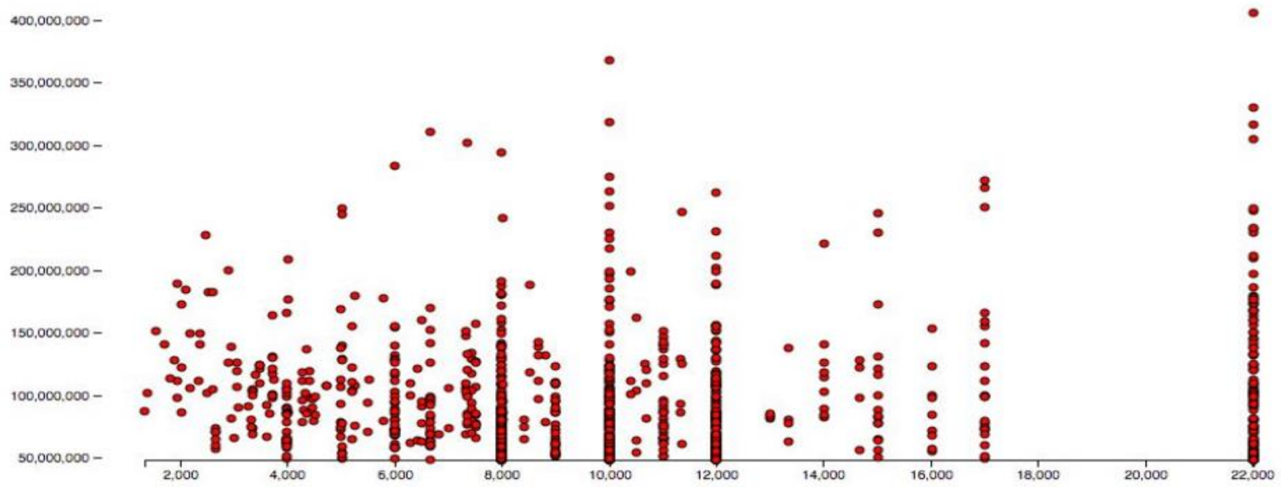

Figure 3. Actor Score vs Gross

Manisha Mishra

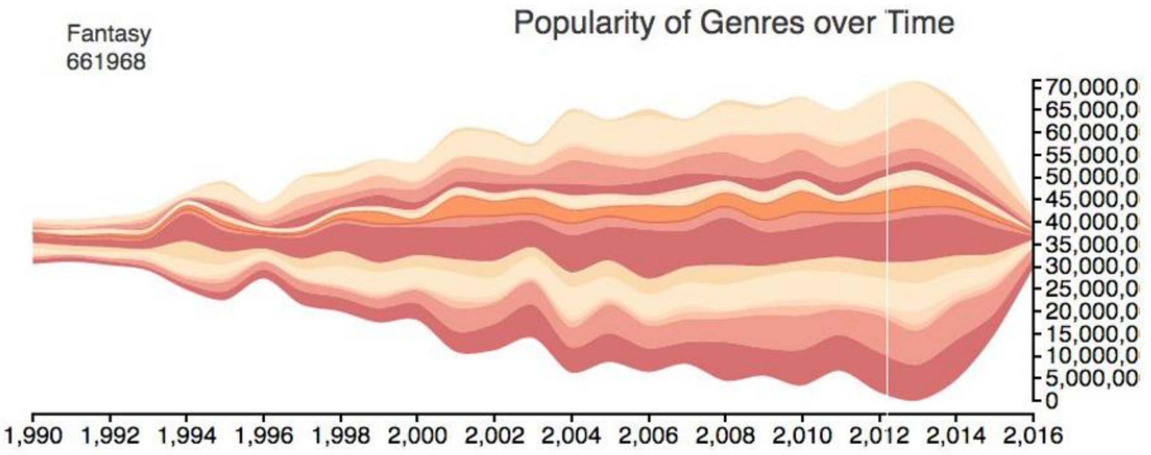

Figure 4. Popularity of Genres

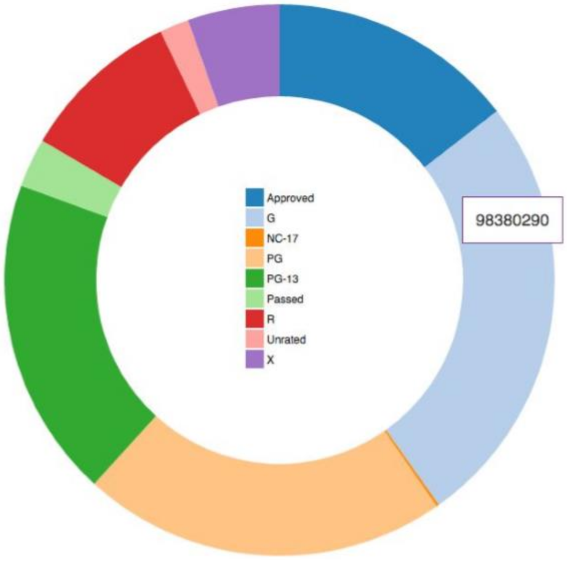

Figure 5. Ratings vs Gross

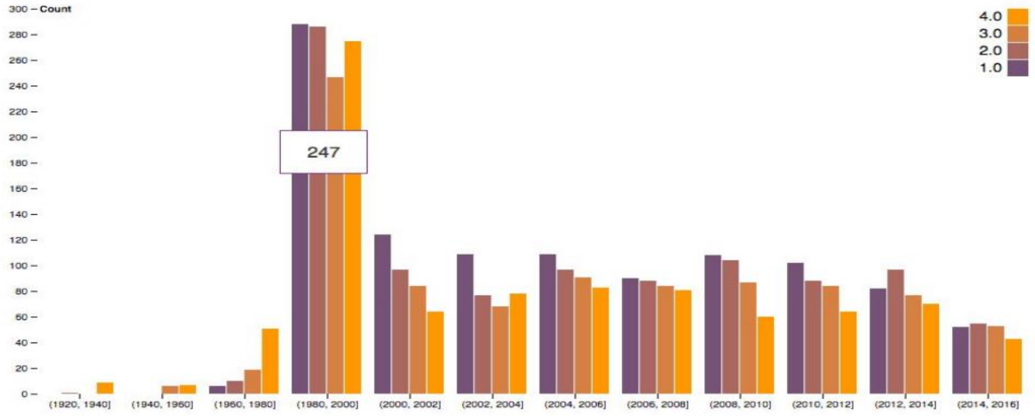

Figure 6. Quality (IMDB Score) vs Count

Manisha Mishra