Project #3
Visualizing Movie Data

Claudia Dai

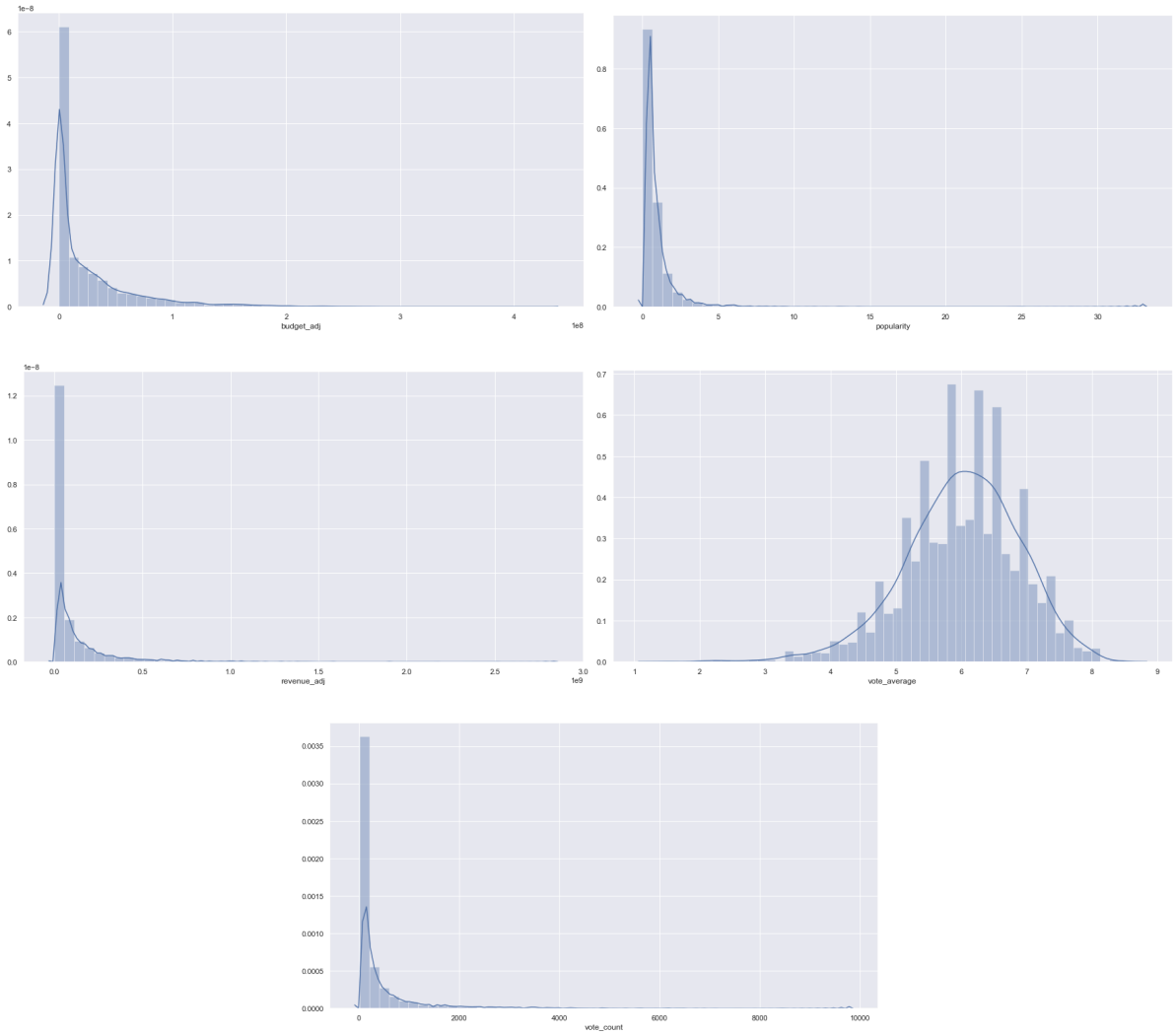
6th November, 2018

1 Data Cleanup and Attribute Selection

First, the movies data was analysed on missing values and data skewedness. The columns with the most missing values are `homepage`, `tagline`, `keywords`, and `production_companies`. Homepage is not such a big issue, but the latter three determine whether the movie is based on a novel or not and which production company produced the movie. Both of these features are needed to answer questions of this assignment. We don't want to assign these movies to a wrong genre/production company by imputing, nor do we want to assume it is "other". The same applies to `genres`. Therefore, let's drop the movies where the value of these columns are missing. We won't be using `cast`, `director`, or `Imdb_id`, so we don't need to worry about those columns.

	Missing values	**% of Total Values**
Homepage	7930	73.0
Tagline	2823	26.0
Keywords	1493	13.7
Production_companies	1030	9.5
Cast	76	0.7
Director	44	0.4
Genres	23	0.2
Imdb_id	10	0.1

Looking at some graphical distributions of the numeric column, we see that the data is very skewed for all columns, except for `vote_average`, which has a minor skew. Therefore, we should take into account **to use median** instead of mean in Tableau.



We also see that `budget_adj` and `revenue_adj` have a lot of zero values. Upon checking the first movie where budget is supposedly zero, which is Mr. Holmes, some googling reveals that the budget was in fact 10 million USD. For revenue being zero, the first movie is supposedly Wild Card, and the worldwide box office indicates 3 million USD with a production budget of 30 million USD. Damn! Someone lost money. We could dive deeper, but since this is a Tableau exercise and not a data cleaning exercise, let's assume that the budget being zero is likely to be wrong, i.e. a placeholder for a missing value, and the same for revenue being zero. Thus, we will drop the rows where budget is zero, and keep the rows where revenue is zero. We are left with 3448 rows of data, which is still quite a lot of movies for analysis. Now we need to prepare the data before feeding it into Tableau.

In order to find movies that are based on novels, a subset of data was created based on the `keywords` and `tagline` columns. With regex, the columns were searched for the appearance of "novel". In `keywords`, three phrases were found that can tell us whether or not a movie is based

on a novel, namely "based on novel", "based on graphic novel" and "inspired by novel". In `tagline`, four movies were found to be based on a novel according to the tagline phrases. This information was translated into a new column called `based_on_novel`.

For the question concerning differences between Universal Studios and Paramount Pictures, the column `production_companies` was parsed for the appearance of "Universal" and "Paramount". There were several phrases including Universal – turns out that all of those are subsidiaries of the same company NBCUniversal. There were also multiple phrases including Paramount, and it turns out that all of those are subsidiaries of the same company. Therefore, all of these phrases were considered to be part of their parent company and the `production_companies` column was tidied up such that it indicated whether the movie was produced by Universal Studios, Paramount Pictures, or others.

A column for profit adjusted was created based on the given columns `revenue_adj` and `budget_adj`.

Finally, the `genres` column needed to be cleaned up. The column held several genres in the same cell with a "|" separator. This was solved by splitting the column by the separator into multiple columns, which resulted in a wide format, and then by melting the dataset to retrieve a tall format (i.e., columnar storage).

In summary, we will use the following columns for the following purposes:

- id: join the tables
- *original_title: identify individual movies in Q4*
- production_companies: identify Universal Studios and Paramount Pictures
- release_date: to analyse the trend over time
- vote_average/popularity: to analyse the popularity
- budget_adj, revenue_adj, profit_adj: analyse financial success (or non-success)
- based_on_novel: identify movies that are based on novels and those that are not
- genre: identify the genres of the movies

2 Tableau Visualizations

The Tableau dashboard was built with Tableau's built-in colorblind palette to accommodate viewers with colorblindness.

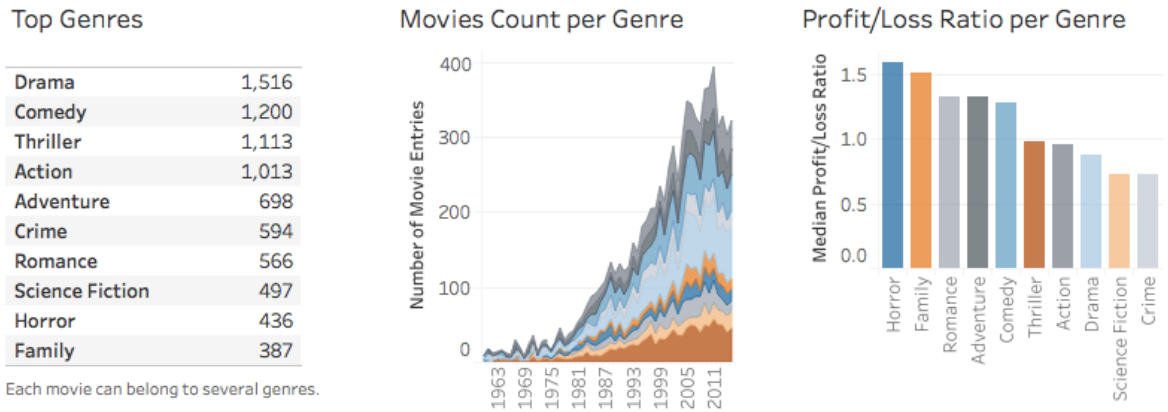
Looking at individual movies, there seems to be a few movies that don't have correct monetary values. For example, The Karate Kid, Part II shows 225 USD budget, but according to the Internet, the budget was 13 million USD. Similarly, a few other movies don't seem to have the right budget amounts declared, and some were, thus, excluded from the visualizations. However, a more thorough analysis should be done on this in the future or it should be considered to thoroughly re-scrape the data from ImdB.

A parameter for top N (3, 5 and 10) determines the top genres by count of movie entries. This parameter serves as a filter for all the dashboards, to focus on those genres that are most prominent in the movie database based on count.

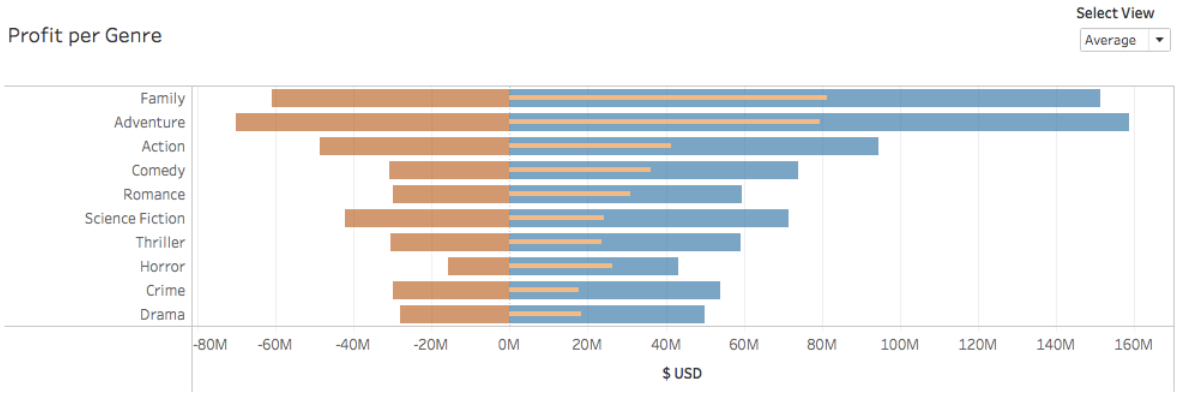
3 Questions

3.1 How have movie genres changed over time?

The amount of released movies has significantly increased over time. From around 1990s onwards there is a strong growth in movie releases.



Most of the movies released are of genre Drama, Comedy, and Thriller respectively, with Action, Adventure and Drama movies having the highest accumulated revenue and profit. Adventure, Family and Action movies bring in the highest revenue and profits on average (median). However, when looking at the profit/loss ratio per genre, Horror, Family and Romance have the highest profit-to-budget ratio.

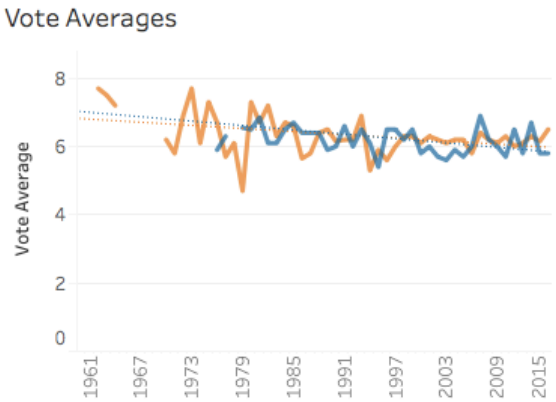


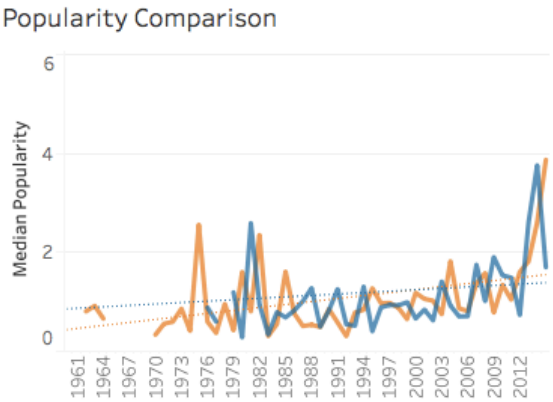
From this genre trend analysis, it becomes clear that horror seems to bring in most profit for small-budget projects on average. If you were running on a small budget, I would recommend

going for a Horror movie. If your budget does not matter much, Family or Adventure movies bring in the biggest revenues and also have quite acceptable profit-to-budget ratios.

3.2 How do the attributes differ between Universal Pictures and Paramount Pictures?

The vote averages for both Universal and Paramount are about the same on average throughout the years. However, when it comes to movie popularity, Paramount seems to have started off better with movies released in the 60s – but Universal's more recent movies have been gaining more popularity.

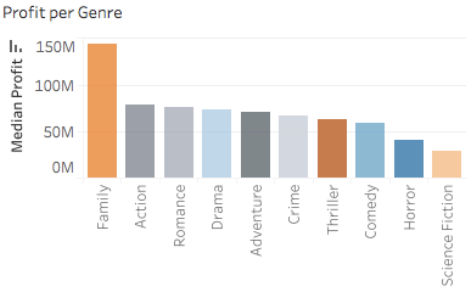


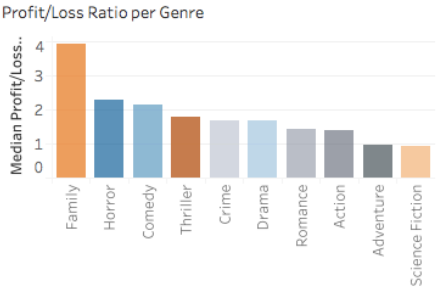


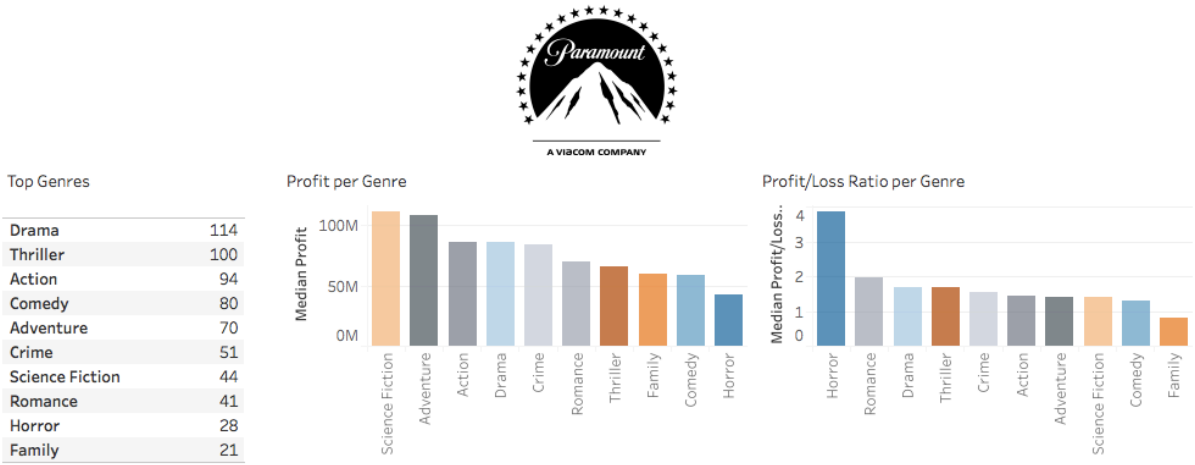
Universal released most of its movies in Comedy, Drama, Action, and Thriller respectively, while the genre Family brought in most profits on average and also has the highest profit-to-budget ratio. The genre Family seems to be a solid bet, but interestingly, Universal has released the least amount of movies in those genres when considering the top 10. Paramount released most of its movies in Drama, Thriller, Action, and Comedy – same genres but in a different count order. While Science Fiction and Adventure brought in most profits, Horror has the significantly highest profit-to-budget ratio.

Top Genres

Comedy	133
Drama	118
Action	112
Thriller	105
Adventure	76
Crime	61
Romance	59
Science Fiction	47
Horror	42
Family	28



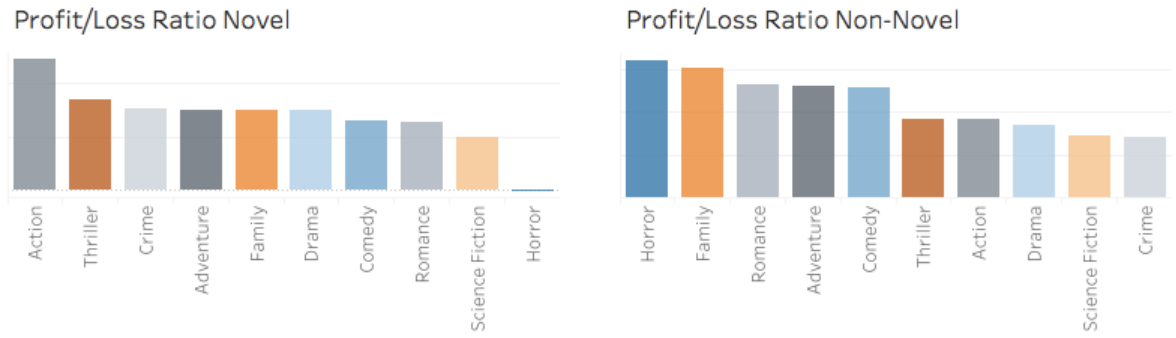




Both Universal and Paramount produce mostly Action, Comedy, Drama and Thriller movies, but Family and Horror, once again, show potential for good profit-to-budget ratios when deciding on what movie to make.

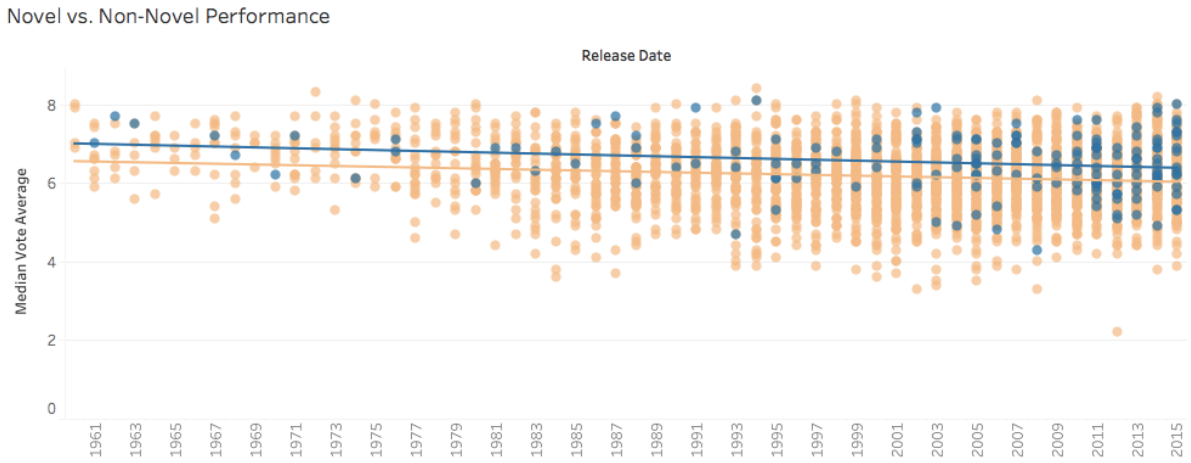
3.3 How have movies based on novels performed relative to movies not based on novels?

Novel-based movies performed well in terms of profit-to-budget ratio if they were Action movies, and very bad if they were horror movies. Interestingly, non-novels have performed best if they were Horror movies.



Generally, it is worthwhile to think about producing a movie based on a novel since throughout the years, the median vote average has remained higher than the non-novel based movie vote average. So novel-based movies are well perceived!

The production company should think about producing a novel-based movie. While non-novel based Horror movies are recommended, the production company should stay away from novel-based Horror movies, even if these generally perform well in terms of profit as described in the previous sections.



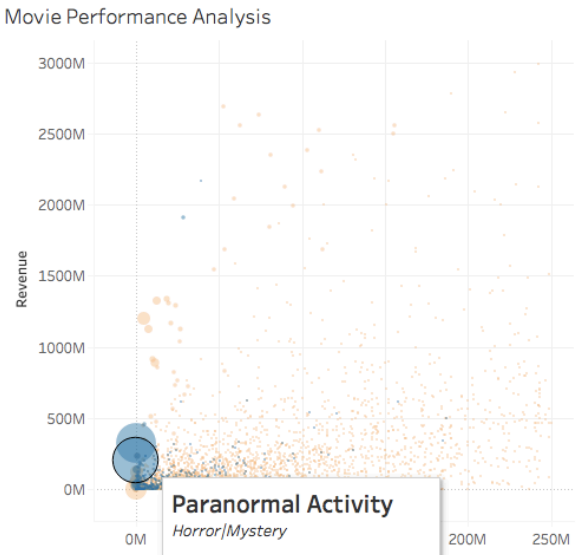
3.4 What movie should the new production company make, given only a small budget?

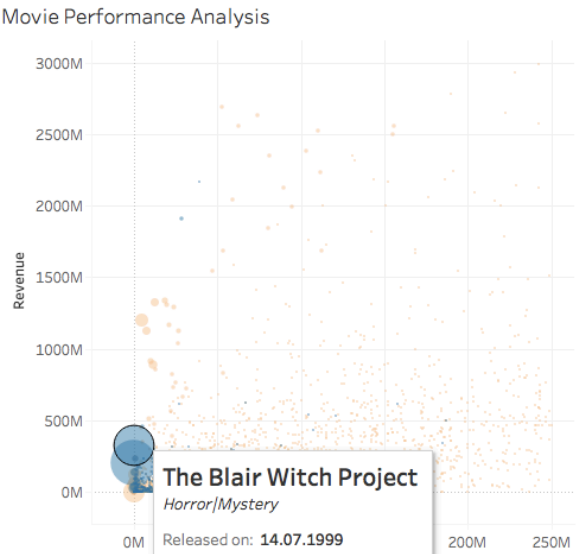
To answer this question, we are making use of the insights we gained from the first three dashboards. Even though Horror movies are not released as much as Drama, Comedy, or Thriller movies, Horror movies have the highest median profit-to-budget ratio. While Horror movies don't have a very high profit, they tend perform best on average on low budget compared to other genres. This observation can be further confirmed when looking at top competing production companies in the industry. For both Universal and Paramount, Horror didn't bring much profit, but the average profit-to-budget ratio was the highest for Paramount and second highest for Universal compared to other genres. The new production company will not have as much budget as the top competing production companies in the industry, so it makes intuitive sense for us to take a look at the Horror movie genre. However, one should be aware that while the profit-to-budget ratio is very high for Horror, this only applies to movies that are not based on novels. For movies based on novels, Horror is doing very badly when it comes to profit-to-budget ratio. It is not recommended to produce Horror based on novels.

The visualization to analyse individual movie performance leverages Tableau's power of enabling self-service analytics, allowing the users to manipulate data him or herself by filtering in order to spot opportunities and trends in the movie data. Following this logic, there are three filters to filter the data on monetary values. For example, if the user is interested in analysing which movies were made on a small budget but had high revenue, the user can make use of the budget and the revenue filter. The user can also filter the movies based on profit or loss made. The year of release date filter enables the user to look at specific year ranges. Perhaps the user is more interested in more recent movies. Finally, the genre filter allows the user to choose which genres he or she wants to have in the overall analysis.

Once again, in this visualization, when analysing the individual movies with their genre group on revenue vs. budget allocation, it becomes once again clear, that Horror movies seem to do very well in terms of revenues provided their rather small average budgets. For example, The Blair Witch Project made 324 million USD revenue on a 32k USD budget, and Paranormal Activity

made 200 million USD revenue on a 15k USD budget! Another example is The Exorcist, which made whopping 2 billion USD revenue on 40 million USD budget.





In conclusion, the production company should decide on the budget for the new movie, look at trends coming from the top production companies, and also consider novel-based movies. It is recommended for the new production company to produce a Horror movie on low budget that is based on a novel. It is also recommended to conduct further analysis on market trends, demographics of target group, and make use of predictive analytics to make a final informed decision on the new movie.