# Complementary Visualization & Profitability for Filmmaking

Deepkumar Patel
DATA-53000
Data Science Program
Lewis University
deepkumarpatel@lewisu.edu

Julian Cagnazzo
DATA-53000
Data Science Program
Lewis University
juliandcagnazzo@lewisu.edu

Morgan Heyboer
DATA-53000
Data Science Program
Lewis University
morganrheyboer@lewisu.edu

***Abstract - The goal of this project is to merge two complementary datasets related to movies and extract interesting patterns which will help generate more revenue in box office sales. This report is highly focused towards comparing revenue and other useful metrics which can be put to use as a guide for any filmmaker.***

## I. Introduction

The raw dataset included in this report comes from two different sources; TMDB [1] website and Kaggle [2], and both of these datasets are located on Kaggle, however the original source for the second dataset is still unknown. The movie database (TMDB) is a popular, user editable database for shows and movies, and most of the attributes in this dataset are categorical along with some continuous attributes. The second dataset contains mostly continuous attributes, and a few categorical attributes. To create the final movie dataset, both datasets were merged, based on the movie title. In the later section, we will see how this dataset was cleaned and transformed to generate explanatory data visualizations.

The main purpose of this report is to merge two complementary datasets and extract useful trends and patterns which will potentially help filmmakers generate more revenue for the movies they will be producing. This will not only help media production companies grow and make more profit, but it will also help filmmakers connect more with their audience by showing them the types of movies they like to see. Even though this report only focuses on movies, this type of analysis can be applied to TV shows, seasons, and even episodes. This report shows the projections of the revenue a movie might make; therefore, it should not be taken as concrete evidence. A successful movie involves many immeasurable attributes such as script, compelling storyline, editing and many more, however, most of the measurable attributes are included in this dataset.

The future sections of this report describe the data preprocessing, dataset, nine different explanatory data visualizations, along with a discussion, and a conclusion. Annexed is a list of sections with a more detailed description. Section II describes the process that is involved before any type of analysis is made on the dataset; this includes merging, transforming and removing features from the data set. Section III describes the merged dataset. Section IV includes explanatory data visualizations along with an explanation. Section V includes a discussion regarding the visualizations. Section VI is the conclusion of this report.

## II. Data Preprocessing

As mentioned earlier, this report uses two complementary datasets, therefore, in order to merge them properly it is important to clean and transform them before it can be used for final analysis. The original raw dataset of TMDB contained 4,803 instances with 20 attributes, and the second dataset contained 5,043 instances with 19 attributes. For the purpose of this report, only a few chosen columns were extracted because they made the most sense and were presentable. Data preprocessing is done using python in a Jupyter notebook, which is included with the submission of this report. This process itself is divided into three main steps.

### *A. Format column names and remove irrelevant columns*

This step includes renaming the column names to avoid any confusion between the two datasets and it also helps understand the columns we are dealing with. Since there was a goal in mind, it was easy to drop most of the columns that did not provide value to our final analysis.

### *B. Add Parsed columns*

There were a couple of columns in the raw dataset which contained JSON objects and concatenated strings, these types of columns were parsed to extract the information that was needed for the analysis. Once the column is fully parsed, each value of a particular cell is then added as a new attribute in the data frame. This step increased the number of attributes in the data frame, but it gave the opportunity to analyze the individual values that were once part of one giant column.

### *C. Merge and rearrange columns*

After applying the above steps to individual datasets, they were both merged based on the title of the movie. After merging both the datasets, all columns were rearranged for better readability.

After all the above steps, the final dataset contained 5,048 instances and 70 attributes, but it was further filtered in Tableau as per the goal of this report. In addition, new columns were generated based on the old columns. The final tableau worksheet and file is also included with the submission of this report.

## III. Data Description

After the preprocessing step, and after careful data exploration, we are left with the following attributes shown in Table I. Some of the columns were computed and derived using different columns. In addition, for each visual, all the null instances were removed, and for some columns the range was adjusted while still being able to accurately present the data. The final number of attributes that were used for this report were 13, and using this data the next section will cover the explanatory data visualizations.

**Table I - Movie Dataset**

| ***Attribute*** | ***Type*** | ***Example Value*** | ***Description*** |
|---|---|---|---|
| Genre | Nominal (string) | Comedy | Movie genre |
| Duration | Numeric (integer) | 120 | Length of the movie |
| Word count (Computed) | Numeric (integer) | 17,729 | Counts of words spoken in the movie |
| Movie Facebook Likes | Numeric (integer) | 33,000 | Facebook likes for the movie |
| IMDB Score | Numeric (float) | 7.9 | IMDB rating for the movie |
| TMDB Popularity | Numeric (float) | 150.438 | Average popularity of the movie on TMDB |

|  |  |  |  |
|---|---|---|---|
| Critic Review Count | Numeric (integer) | 723 | Critic reviews count for the movie |
| Release Date | Interval (date) | 12/10/2009 | Date movie was released |
| Profit (Computed) | Numeric (integer) | 24,139.100 | Profit made on the movie |
| Content Rating | Nominal (string) | PG-13 | Content rating of the movie |
| Actors | Nominal (string) | Johnny Depp | Lead actor of the movie |
| Revenue | Numeric (integer) | 300,000,000 | Revenue for the movie |
| Budget | Numeric (integer) | 240,000,000 | Budget for the movie |

## IV. Methods

### A. *Explanatory Data Visualization*

The following visualizations were created using Tableau v.2020.3.2. Movies with budgets under 10 million dollars were excluded from visualization in order to focus on larger movies.

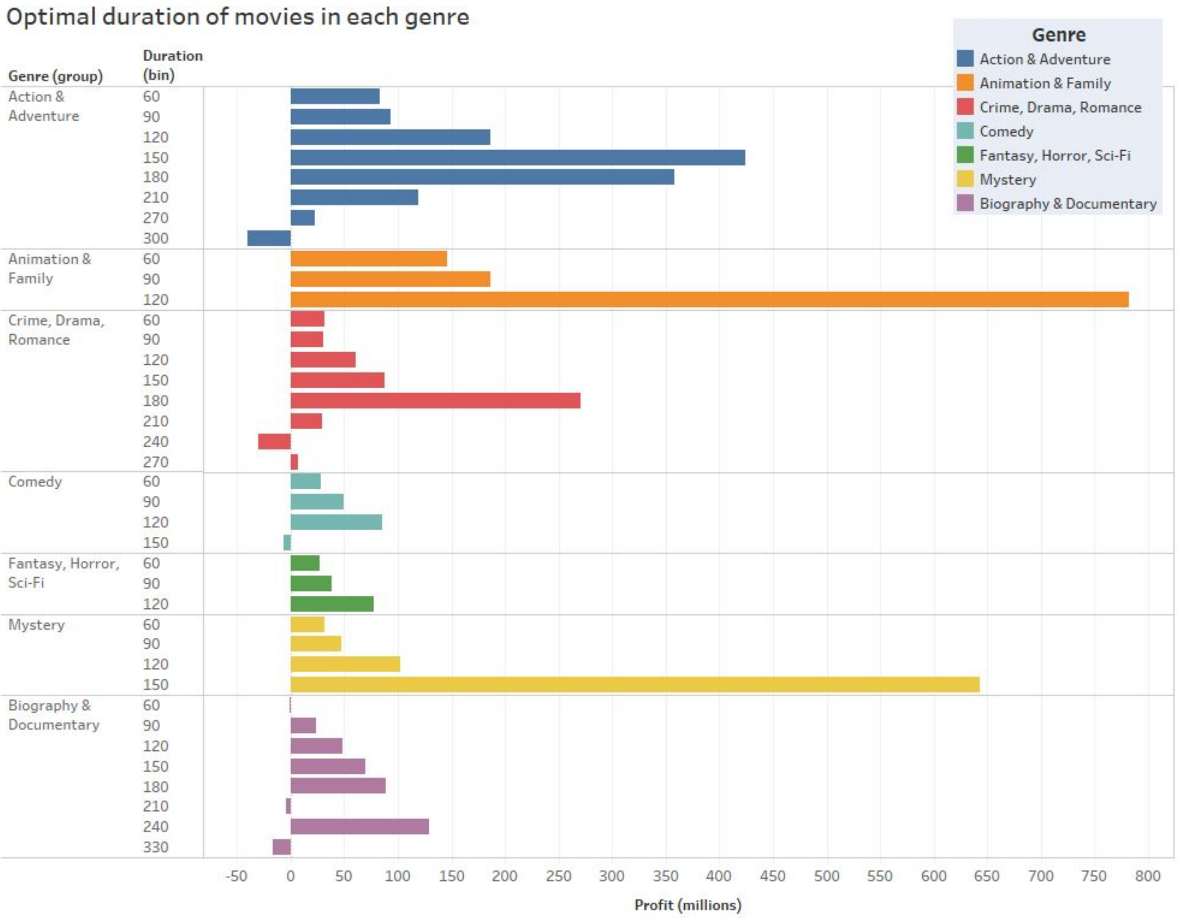

**Figure 1. Optimal duration of movies in each genre**

Each bar in the fig.1 represents the amount of profit earned on average by movies of the duration and genre specified on the y-axis. From this graph, we can see that the optimal length of a movie depends heavily on the genre. The optimal duration ranges from 120 - 240 minutes depending on the genre. Fantasy, Horror and  Sci-Fi has the shortest optimal duration at 120 minutes and Biography and Documentary has the longest optimal duration at 240 minutes.

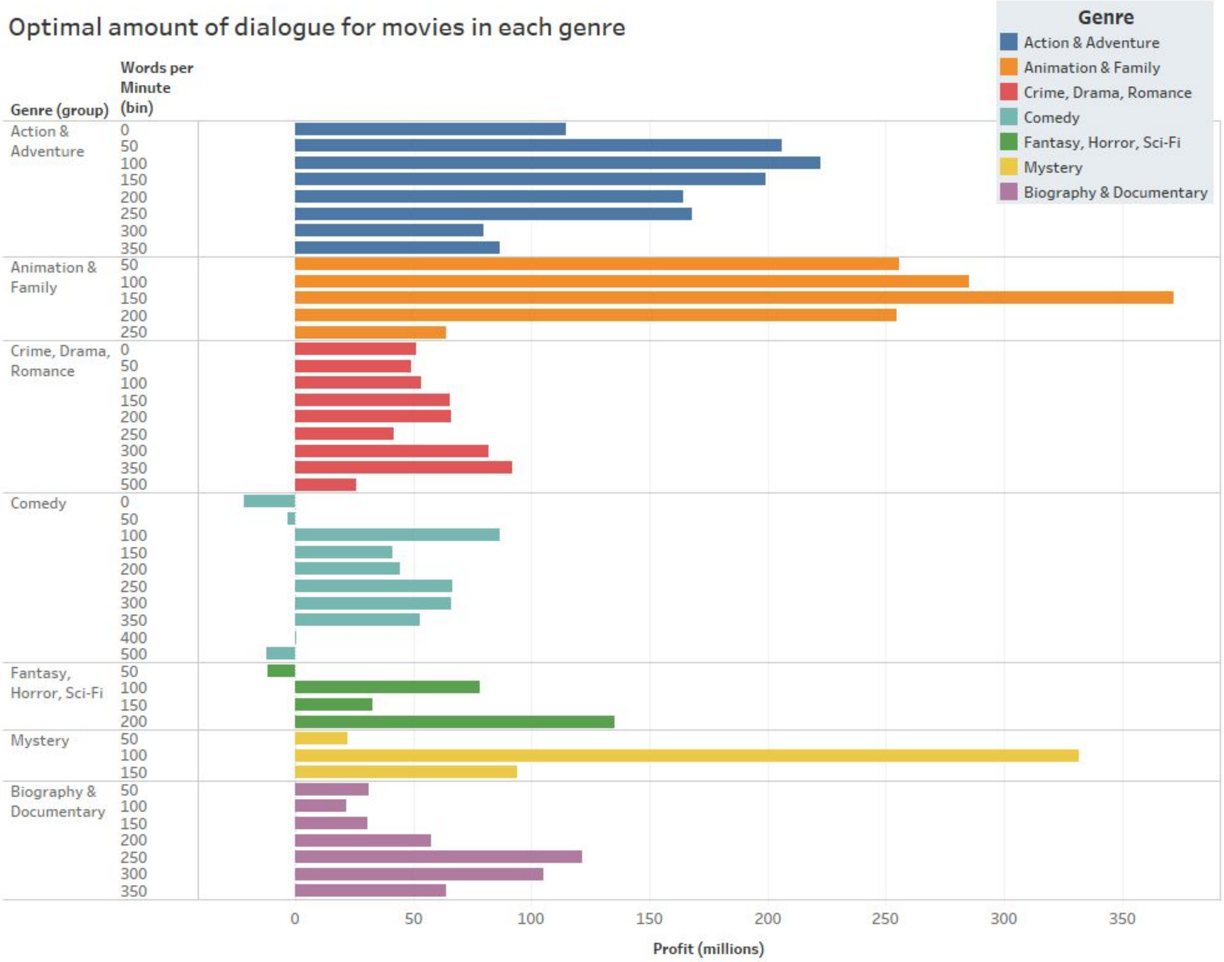

**Figure 2. Optimal amount of dialogue for movies in each genre**

Each bar in fig.2 represents the amount of profit earned on average by movies of the duration and genre specified on the y-axis. From this graph, we can see that the optimal amount of dialogue depends heavily on the genre. Optimal dialogue ranges from 100 - 350 words per minute of screentime. Crime, Drama, and Romance benefit the most from a lot of dialogue with the optimal amount being 350 words per minute. Most other genres have an optimal amount of dialogue closer to 100 words per minute of screentime.

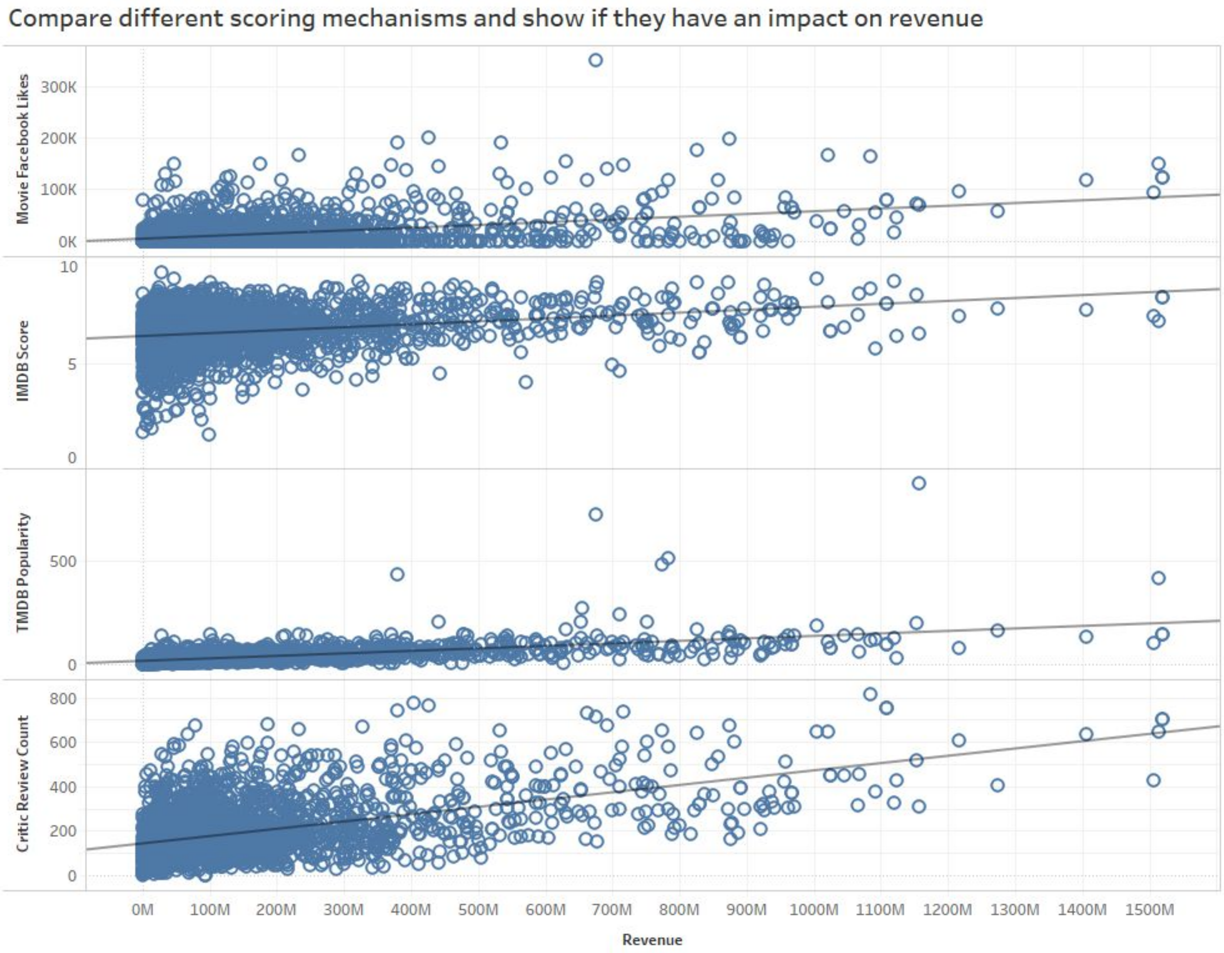

**Figure 3. Compare different scoring mechanisms and show if they have an impact on revenue.**

Each circle in fig.3 represents a movie that has been given the score specified on the y-axis and generated the revenue defined on the x-axis. From the trendline fit to this data, we can see there is not a strong correlation between these scoring metrics and the amount of revenue generated by a movie.

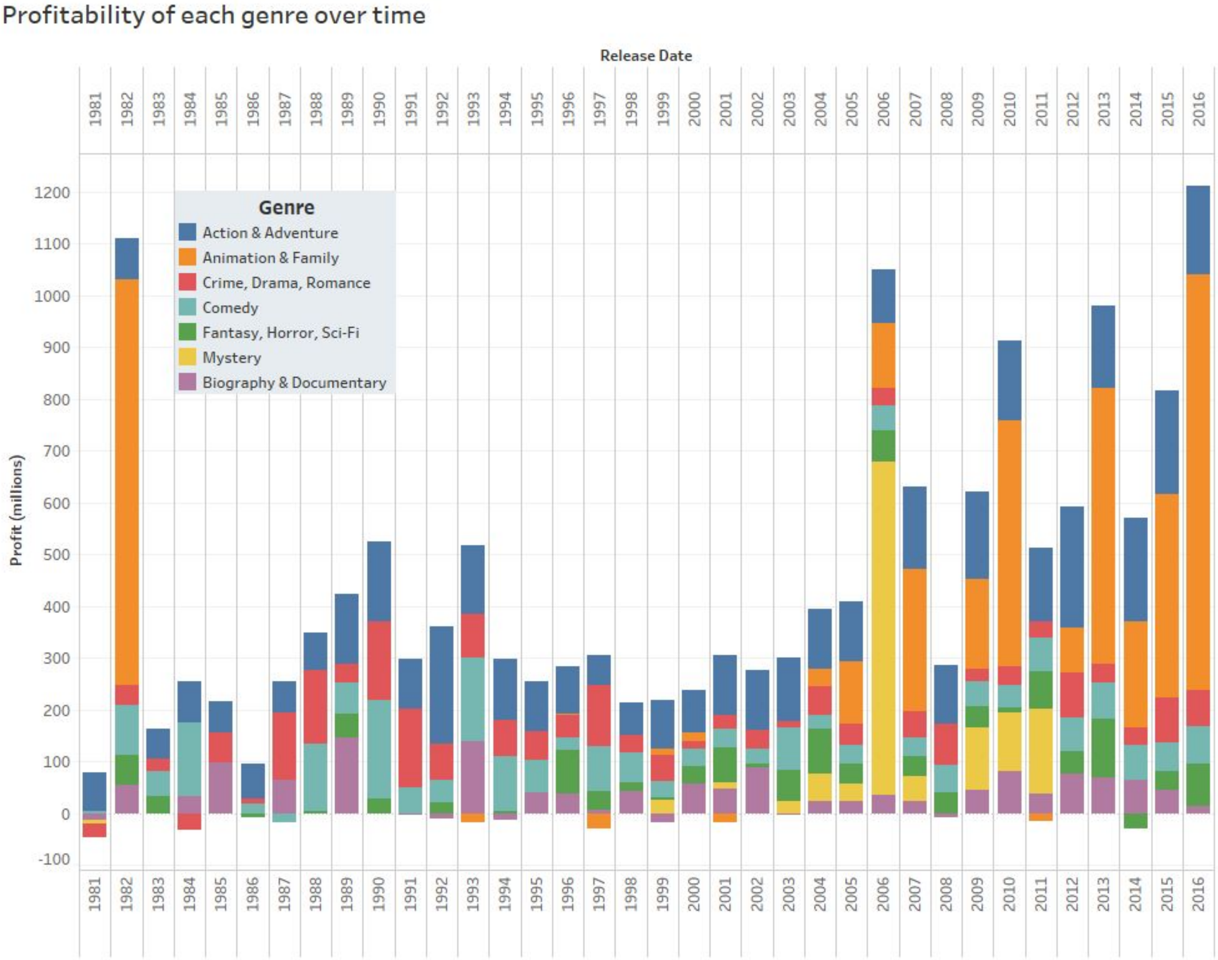

**Figure 4. Profitability of each genre over time.**

Each bar in fig.4 represents the average profit generated by all the movies in our dataset during the year specified in the x-axis. Each bar is broken down into colors showing how much of the total profit was generated by movies in each genre. From this graph, we can see that animation and family movies have seen an explosion on profitability over the last 15 years. We also see that action and adventure movies have been some of the most consistently profitable movies over the last 40 years.

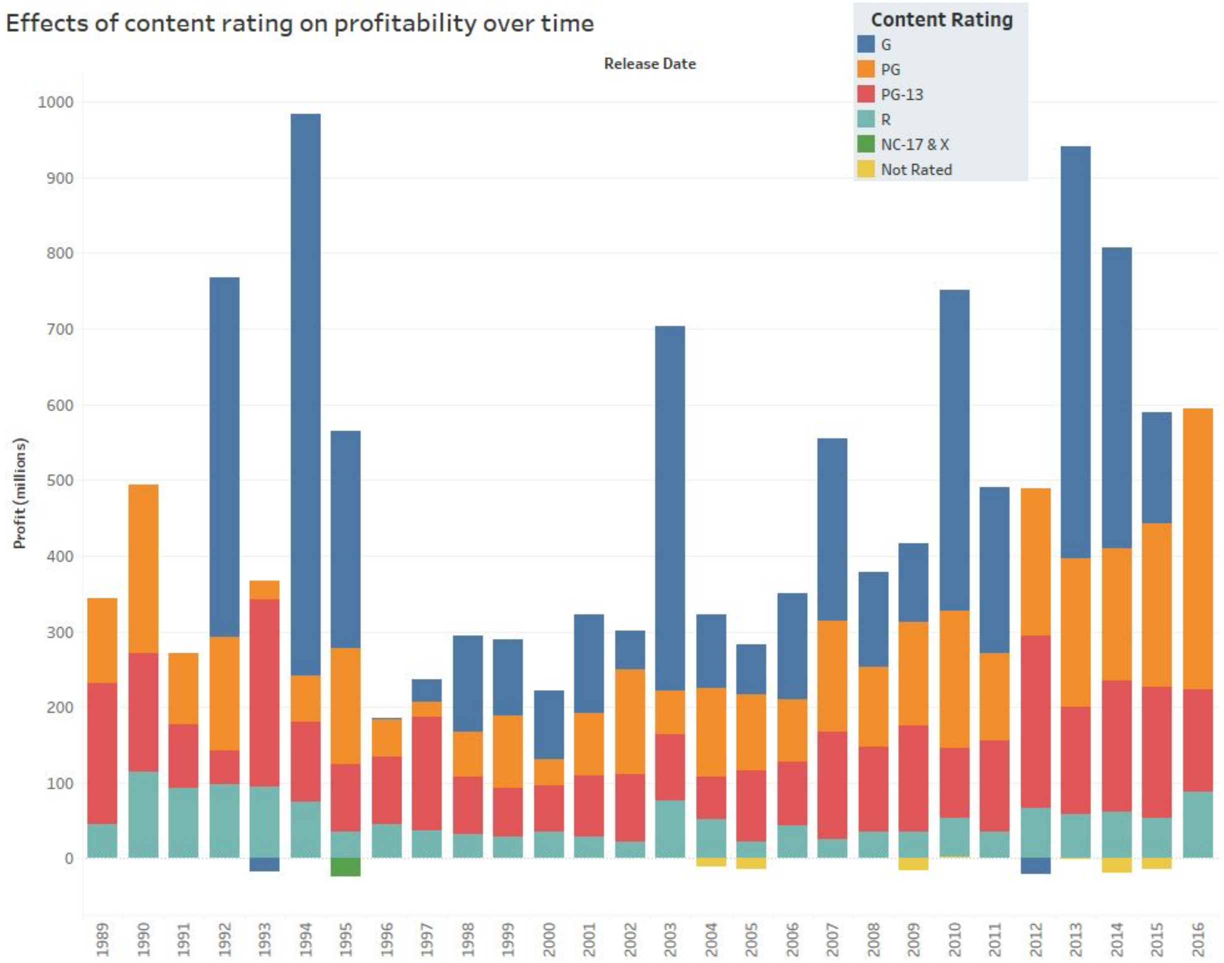

**Figure 5. Effects of content rating on profitability over time**

Each bar in fig.5 represents the average amount of profit generated by all the movies in our dataset during the year specified on the x-axis. Each bar is broken down into colors showing how much of the profit was generated by movies with each content rating in the legend. From this graph, we can see that having a stricter content rating can seriously limit the amount of profit generated from a movie. Most of the profit generated by movies over that last 30 years have had a content rating of PG-13 or lower.

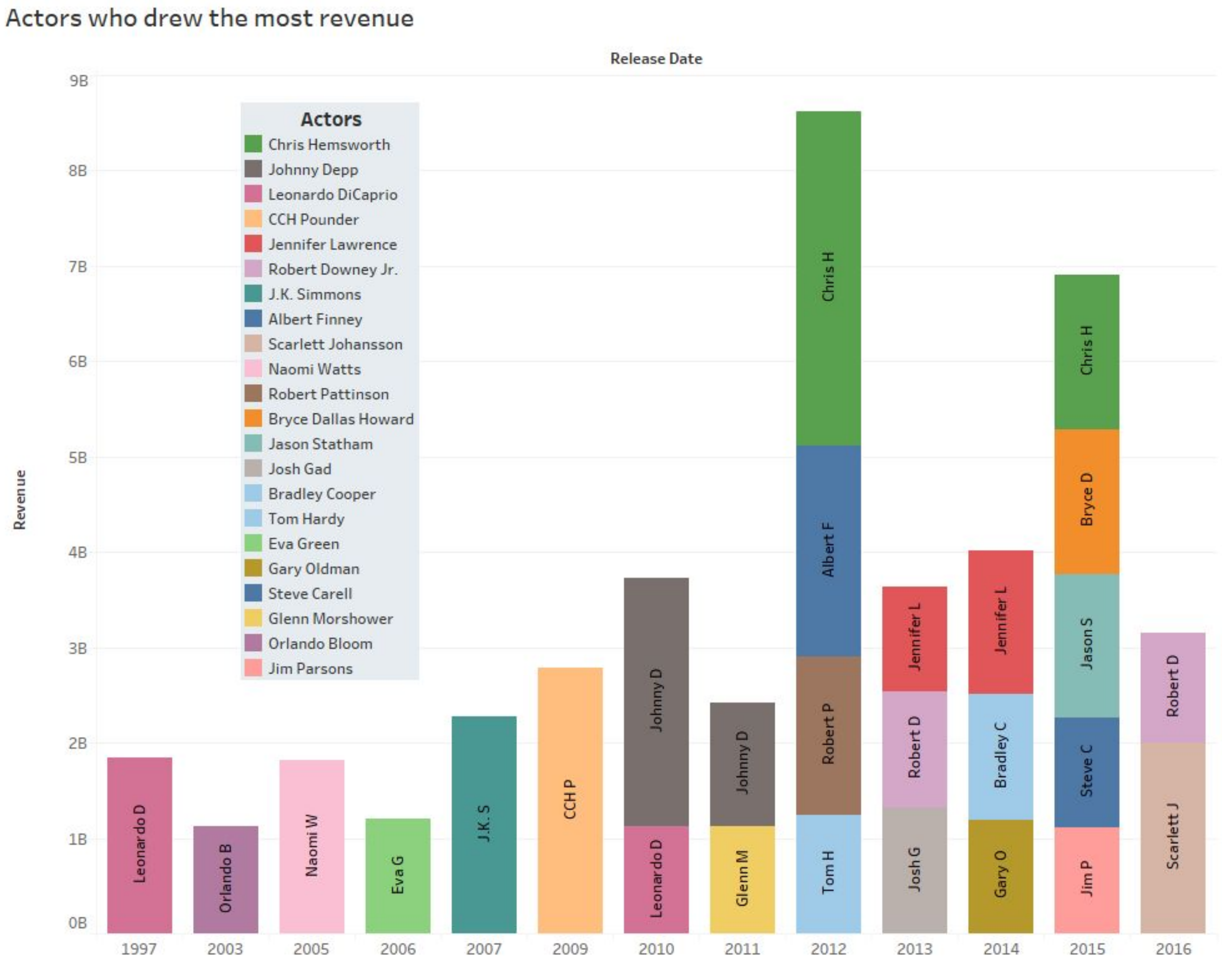

**Figure 6. Actors who drew the most revenue**

In fig.6 graph shows the revenue brought in by movies starring the most successful actors in the year specified on the x-axis. In order to be included in this graph actors must have starred in movies with a sum of revenues totalling at least 1 billion dollars in a single year.

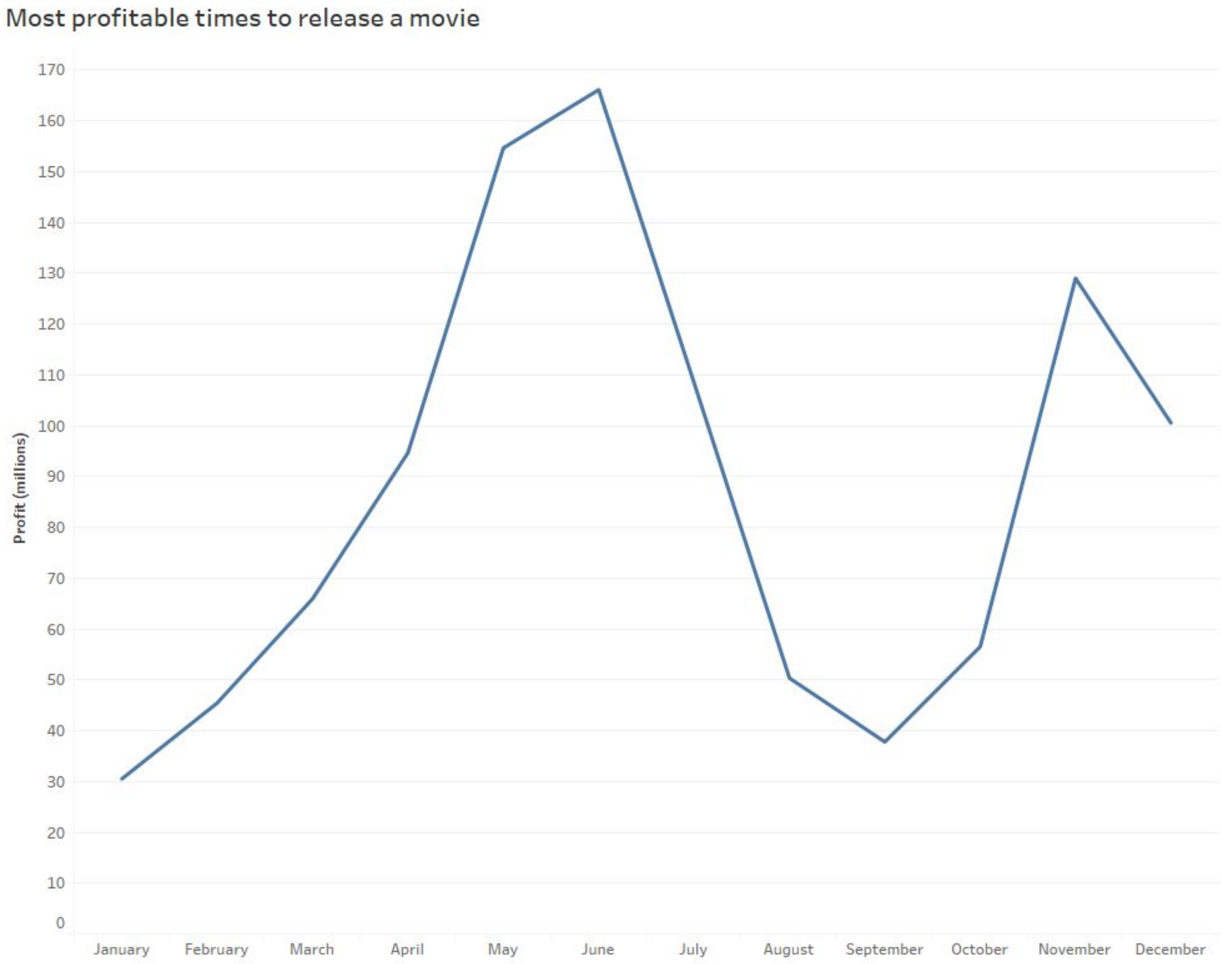

**Figure 7. Most profitable times to release a movie**

The line in fig.7 represents the average profit generated by movies released during the month specified on the x-axis. We can see from this graph that the most profitable times to release a movie are during the months of June and November.

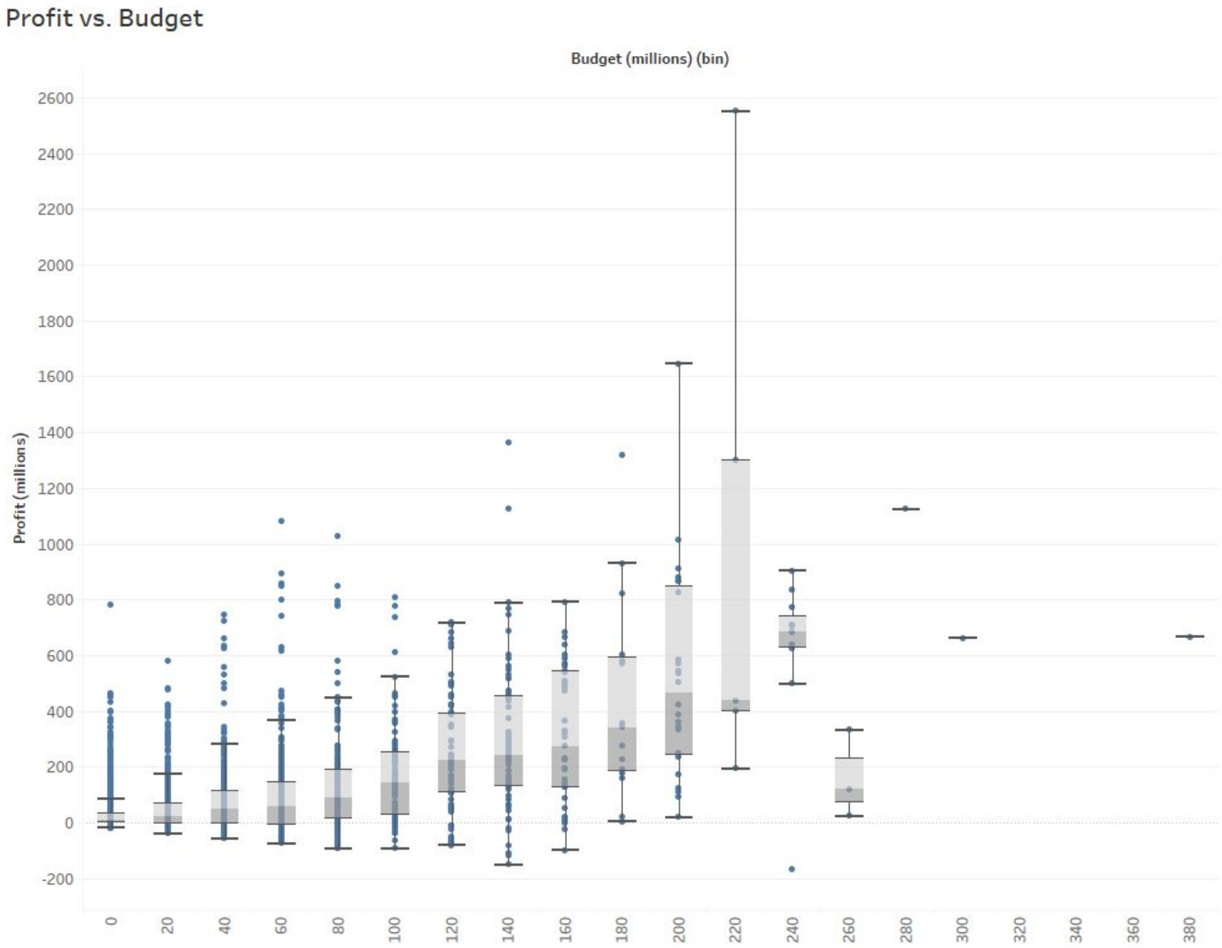

**Figure 8. Effects of budget size on profitability of movies**

Each box in fig.8 represents the profits generated by movies with the budgets specified in the x-axis. From this graph we can see that there is a positive correlation between a movie's budget and its profitability.

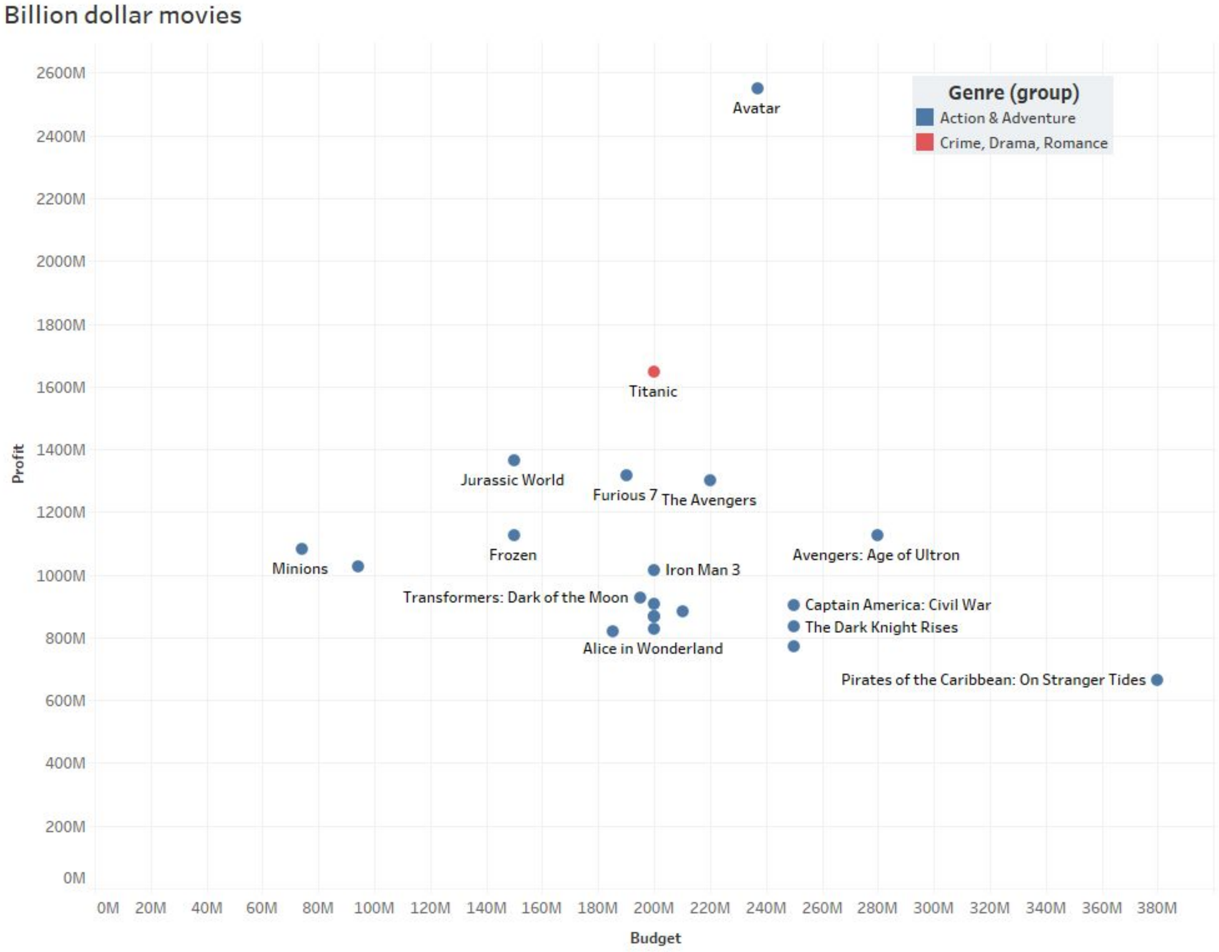

**Figure 9. Billion dollar movies**

In fig.9 graph shows the profit earned and budget of each movie in our dataset that had a revenue of over a billion dollars. From this graph we can see that the overwhelming majority of these ‘billion dollar movies’ are action and adventure movies.

## V. Discussion

### *A. Initial Impressions*

Our initial hypothesis after exploring the dataset was that there are several attributes of a movie that can impact the revenue generated: budget, genre, content rating, and release month. We explored other aspects of movie productions, like lead actor and scoring mechanisms (Facebook likes, IMDB score, etc.) and thought that they did not affect revenue as much.

We can see in Figure 1 that there is an ideal duration length that will generate the most revenue, but it differs slightly for each genre. For Action or Adventure movies, the ideal length is 150 minutes, but for Biography and Documentary movies, it is 240 minutes. We can see a similar relationship between genre and dialog in Figure 2. The ideal amount of dialog appears to vary between 100 words per minute in the Comedy genre to 250 words in the Biography and Documentary genres.

Figure 3 shows a moderately weak relationship between the scoring mechanisms retrieved from the dataset and the revenue that the movie generates. The trend lines for each of the scoring mechanisms all report an R-squared value of less than 0.4, showing that this is not a significant indicator of how much revenue a movie will generate.

Figures 4 and 7 show the effect that the release year and month have on the profit of the movie, which would directly impact the revenue. Figure 4 shows that as we move through time, the profit made by movies has increased drastically. It also breaks down the profit generated by genre, so we can see that Animation and Family genres have generated the most profit in the recent years. Figure 7 shows how the month that a movie is released affects the profit it generates as well. We can see a spike in profit around the summer months as well as around the November holiday season.

Figure 5 shows the relationship between the content rating and the profit a movie generates over time. We can see that content ratings of R and above don't make as much money as G, PG, and PG-13 rated movies, especially in recent years.

Figure 6 shows how much revenue movies made based on the lead actor over time. At an initial glance, you can see that some actors generate quite a bit of revenue (for example Chris Hemsworth) in recent years. However, there is such a variety in the lead actor of movies that it is hard to find any specific pattern.

Figure 7 shows how the budget of a movie can impact the profit it generates, while Figure 9 shows specific examples of movies that generated more than 1 billion dollars in revenue. From Figure 7, we can see that low budget movies tend to generate lower profits than high budget movies. However, that relationship doesn't seem to be as strong when the budget begins to exceed 240 million dollars. Figure 9 gives more context to this relationship by showing specific examples of movies, and where they fall when mapping their budget versus profit.

### *B. How Visualization Theories/Best Practices were Applied*

You can see color theory is used in quite a few of the visualizations. Because we were doing a lot of analysis surrounding the genre of the movie, colors were assigned to each genre in the dataset. Due to the categorical nature of the genre field, color was the most appropriate way to visualize that field. We also grouped together some of the genres (for example, Action and Adventure genres) to reduce the number of possible categories and therefore to limit the number of colors seen on the visualization at once.

Gestalt's Laws are used throughout the visualizations as well. One example is the law of proximity. In Figures 3 and 8, you can see many overlapping data points, which show that the values they represent are extremely close to each other. We can see this especially well in Figure 3. Gestalt's Law of relative size is also used regularly in the visualizations, as can be seen in the bar charts in Figures 1, 2, 4, 5, and 6.

## VI. Conclusions

### *A. Conclusions Based on Data Set*

Our initial hypothesis postulated that budget, genre, content rating, and release month had the most direct effect of a movie's revenue. When media production companies are screening movie ideas, our data shows that these four attributes should be top of mind.

A larger budget for a movie can lead to more special effects, more popular actors/actresses, as well as other critical attributes that lead to a successful movie. All of these combined can result in a very profitable movie, if balanced correctly. However, it must be cautioned that having too large of a budget can lead to a negative effect on the profit, as is seen in Figure 7. A prime example of this can be seen in Figure 9. The movie *Pirates of the Caribbean: On Stranger Tides* had a budget of $380 million, but only made a profit of $665 million. Based on the data in Figure 7, the ideal budget for any movie is between $220 million and $240 million.

The genre and content rating of the movie also have a large impact on the revenue generated by a movie. These two attributes are most likely related. Based on the data shown in Figures 4 and 5, a company's best chance at a large revenue would be to produce a PG rated movie that is family friendly and animated or a PG-13 movie in the Action or Adventure genres. These types of movies seem to produce the most revenue.

The time of year that a movie is released has a significant impact on the revenue generated by a movie. The months of April, May, and June had some of the highest profit numbers shown in Figure 7. November and December also reported high profits. The summer months most likely generated the most profit because school is usually done by that time and teenagers and young adults want something fun to do on their summer breaks. The weather is usually nice out as well so people are more inclined to want to go somewhere instead of stay at home. The months of November and December are an odd place to see an increase in profit, but it can be argued that it's for the same reason as the summer months. There is usually a school break around the holidays, leaving more leisure time for students. Working adults also typically take more time off in these months to spend time with their families and watching a movie is a great way to be together. Releasing movies in the summer or around the winter holidays appears to give the best profits.

### *B. Alternative Conclusions*

While this report focused on objective data surrounding movies (i.e. budget, release date, genre, etc.), many aspects of a movie cannot be objectified, calculated, or measured. Aspects like a storyline or character arcs are critical to a movie's success and cannot be measured.

### ***References***

[1] The Movie Database (TMDB), "TMDB 5000 Movie Dataset" Kaggle, 29 August 2016, www.kaggle.com/tmdb/tmdb-movie-metadata/metadata

[2] G. Mpoy, "Movie Metadata" Kaggle, 29 January 2019, www.kaggle.com/bobirino/movie-metadata/metadata