SI 618 Data Manipulation & Analysis- Project Report

Title: Contributing Factors in Movie Profitability

Authors: Nuri Bae (nuri), Lionel Robert (lionelr)

April 11, 2023

Motivation

Our motivation was simple. Lionel has always been interested in movies and in particular, has always been interested in movie box office reports. It seems news about how much a movie makes is about as common as an advertisement of the movie itself, especially in believing a movie's quality is dictated by the amount of money it makes. At the same time, another common subject surrounding a movie when it releases is what the movie critics make of it. Movie critics have been around for a long time critiquing and praising movies from their unique perspective as critics to the pleasure or dismay of both audiences and filmmakers, with some praising their perspective while others believing there is little use for critics. Many movies that have been financial failures typically come with bad reviews from critics and at the same time, many successful films seem to come with positive reviews or even critical acclaim.

However, the release of the recent Avatar film received lukewarm or average reviews, with reviewers claiming that it wasn't that particularly good of a film, while at the same time, quickly making its way into becoming one of the most financially successful films in history. This has left us wondering; are audiences really affected by movie critic scores? Do bad/average scores really discourage people from watching films? Is there even a real connection between movie scores and the movie's overall gross? We have decided to undertake this research project to understand whether or not there is any particular relationship between the movie itself, including its genre, and how much money it makes, or how well the critic scores are. We also aim to examine the relationship between a movie's budget and its box office revenue and/or critic score. It is often believed that higher-budget movies tend to make more money at the box office and/or receive higher scores from critics.

There were a few similar studies done regarding figuring out whether critical reviews affect revenue. A study conducted by Emerson College (2018)[1] found that positive reviews can

[1] Emerson College (2018), Study Finds Film Critics' Moderate Influence May Have Significant Revenue Impact. CISION PR Newswire, https://www.prnewswire.com/news-releases/study-finds-film-critics-moderate-influence-may-have-significant-revenue-impact-300734203.html

increase box office revenue. Chiu et al. (2022)[2] examined the relationship between critical reviews, user reviews, and box office revenues. The results showed that critical reviews not only directly affect box office revenues but also indirectly lead to greater box office revenues by increasing active postings and user reviews volume.

Data Sources

Our chosen datasets were the TMDB 5000 Movie Dataset (5.7 MB) available on Kaggle, as well as the Rotten Tomatoes Top Movie Ratings and Technical dataset (1.79 MB) also available on Kaggle. Both of them carry different sorts of movie data we needed for this project to work. Both of them are CSV files that could easily be translated into Pandas DataFrames. The TMDB 5000 Movie Dataset features the movie title, the revenue, the budget, and the language the movie came out in, along with the movie genre among other information that may be useful to us. The Rotten Tomatoes dataset also featured the title, the date of release, the critic score, and the people (audience) score, along with an alternative genre, runtime, and a few other extra details about the movie itself. The question was figuring out whether there is some relationship between the variables in the dataset and the movie's box office. The TMDB 5000 Movie Dataset has a lot of information on the budget for movies which is important for our analysis, and this can only be accomplished by merging the two datasets because the Rotten Tomatoes dataset doesn't have budget information.

The datasets could both be found in Kaggle; the Rotten Tomatoes one being here: https://www.kaggle.com/datasets/thedevastator/rotten-tomatoes-top-movies-ratings-and-technical and the TMDB 5000 Movie Dataset being found here https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata?select=tmdb_5000_movies.csv.

Data Manipulation Methods

We merged the TMDB and Rotten Tomatoes datasets using an inner join based on the movie title. We used inner join because we needed all of the common data for each movie title for this analysis to work. This process allowed us to create a comprehensive dataset that

[2] Ya-Ling Chiu, Jiangze Du, Yide Sun, Jying-Nan Wang (2022), Do Critical Reviews Affect Box Office Revenues Through Community Engagement and User Reviews? Frontiers in Psychology, Volume 13, https://www.frontiersin.org/articles/10.3389/fpsyg.2022.900360/full

included all the relevant variables from both sources. Before merging, we performed data cleaning by converting the runtime of the Rotten Tomatoes dataset from its original format into minutes for consistency. We also removed rows containing NaN values, except for box office data. We dropped duplicate rows in both datasets after noticing quite a few movies appeared twice or more times than they should. For the box office data, we used an API call to TMDB to fill in missing values. The API URL link is https://api.themoviedb.org/3/search/movie?api_key={api_key}&query={title}&year={year} wherein you replace the api_key, title, and year with the title of the film, the year it came out, and the API key in the notebook. My API key was 2764382d496b28895007a2252ea60177.

We did face some significant challenges while handling the box office data. Initially, we wanted to use the 'revenue' column from the TMDB dataset, but this proved to be a problem considering how incomplete the data was. The box office revenue was an important part of our analysis, but about a quarter of the data was missing. We tried their official API, but it was no help, so we decided to use the box office column from the Rotten Tomatoes dataset instead since it was far more complete. This caused an additional challenge in converting object values (e.g. "303k") to their numerical counterparts. To correct this problem, we created a custom function to parse and convert the object values into their numeric equivalents. All of this allowed for a clear dataset that enabled us to explore the relationship between box office revenue and many other import movie-related variables effectively.

The code for the conversion of revenue to float can be found here:

```
def revenue_to_float(revenue_str):
    """this converts the revenue column from string to float

    Args:
        revenue_str (object): the revenue column as a string

    Returns:
        float64: the revenue column but as a float

    Note: This particular code was optimized (not written) by ChatGPT-3.
    """
    if not isinstance(revenue_str, str):
        return None

    # Extract the numerical value and the suffix (K or M)
    match = re.match(r'\$([\d.,]+)([KkMm]?)', revenue_str)
```

```
        if not match:
            return revenue_str

        value, suffix = match.groups()
        value = float(value.replace(',', ''))

        if suffix.lower() == 'k':
            value *= 1e3  # Convert to thousands
        elif suffix.lower() == 'm':
            value *= 1e6  # Convert to millions

        return value
```

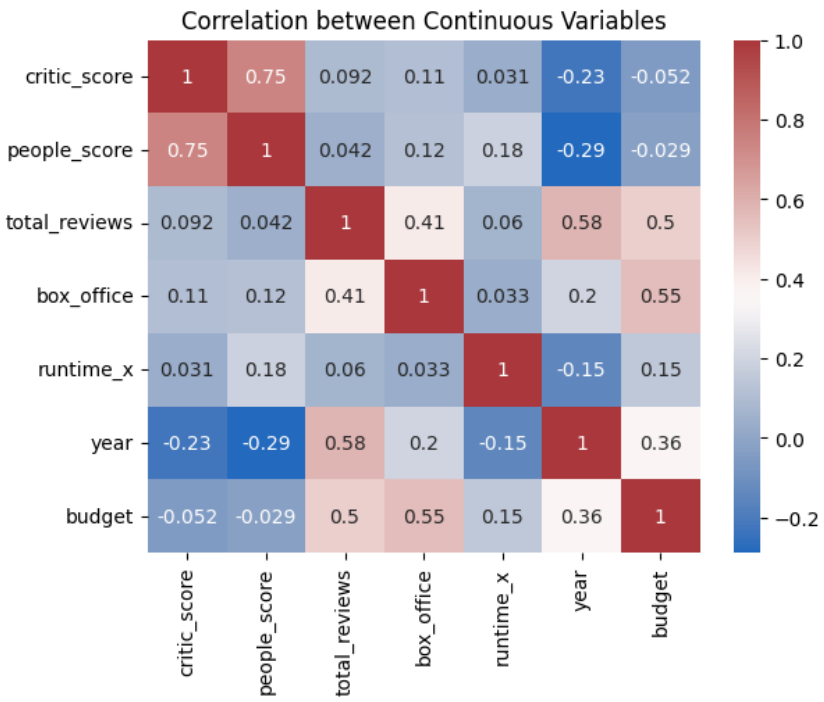
We have also converted and processed several columns in our dataset into a more usable format for our analysis. For example, we grouped languages into two categories: English and non-English. Out of the 340 movies in our dataset, 313 were in English while the remaining 27 movies were spread across 12 different languages. Given the small number of movies in languages other than English, we grouped them into a single category labeled "non-English." Another example is the box office revenue; to create a heatmap between revenue and genre, we needed to cut the continuous variable, revenue, into bins. We decided to use percentiles to define these bins, which allowed us to group the movies into ten roughly equally sized bins.

Analysis and Visualization

Relationship between revenue and other continuous variables

Correlation analysis

A heatmap was created to see the correlation between the variables of our interests (i.e., critic score, audience score, number of reviews, box office revenue, run time, and the year the movie was released, and budget). The highest correlation coefficient (0.75) was found between critic_score and people_score. From a box office revenue perspective, the budget has the strongest correlation (0.55) with box_office than other variables of our interest such as critic_score (0.11), people_score (0.12), total_reviews (0.41), runtime (0.033), and year (0.2).



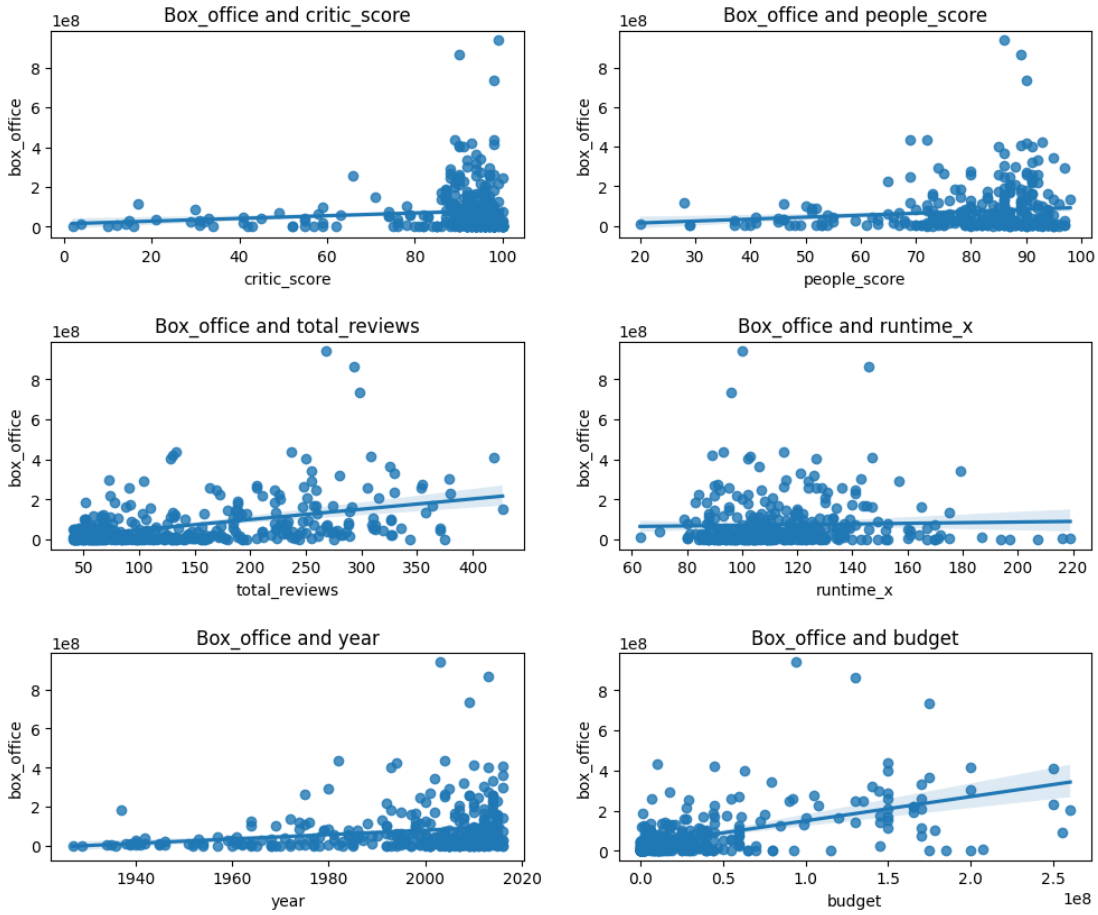
OLS regression analysis

A linear regression model was fitted to the data, with critic_score as the independent variable and revenue as the dependent variable. A small R-squared value (0.011) indicates that the model doesn't accurately capture the true relationship between the dependent and independent variables; only 1.1% of the variability in box_office can be explained by critic_score. Moreover, the p-value of 0.0505 suggests that the relationship between critic_score and box_office may not be statistically significant at the 0.05 level. The histogram and QQ plot in the notebook shows that the critic_score is skewed to the left while the box_office is skewed to the right. Given the skewness of the data, we have considered using log transformation of the revenue, but that didn't solve the problem; it actually decreased the correlation between revenue and other variables, so we decided to use the original revenue, not the log-transformed one.

Budget is a better predictor for box office revenue as it has a larger R-squared value (0.307) and a p-value (8.56e-29) that is much smaller than our alpha of 0.05, which means that 30.7% of the variability in revenue can be explained by budget and there is a statistically significant relationship between those two variables.

We also considered other variables as independent variables in our linear model. When using total_reviews as an independent variable, the model explained the variation of revenue better (R-squared = 0.168), and there is a statistically significant relationship between

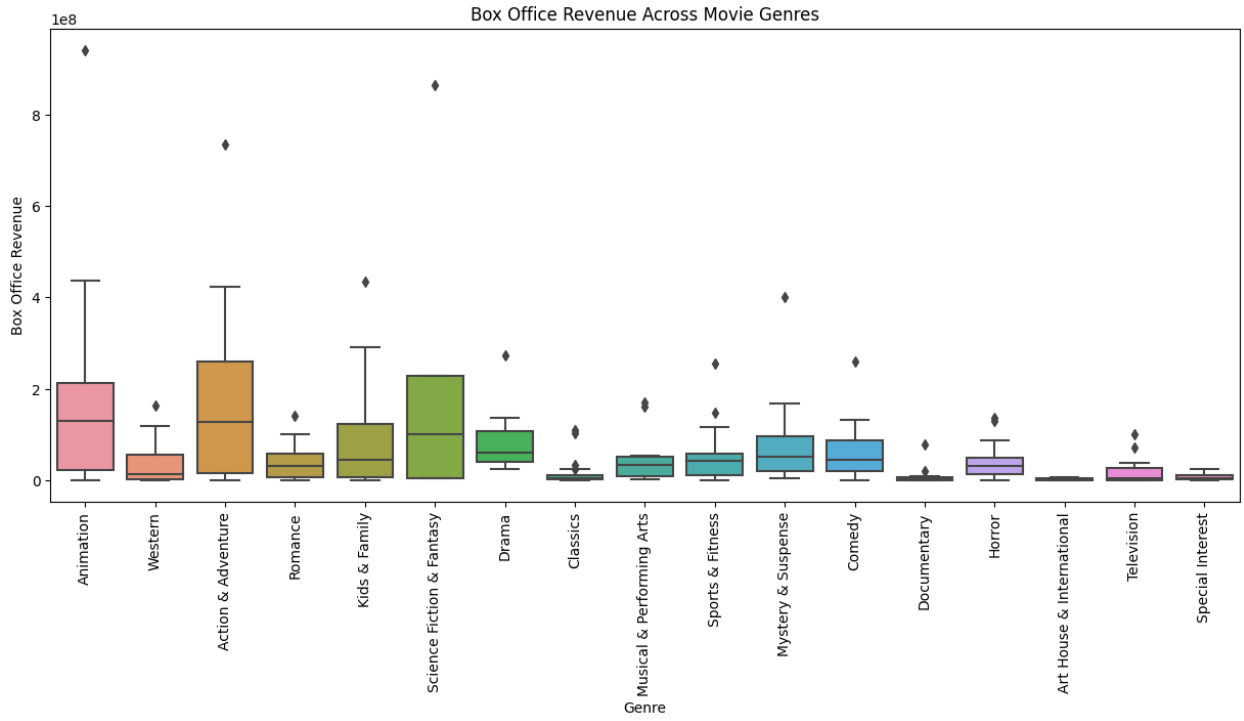
total_reviews and box_office at the 0.05 level. If using multiple regression with independent variables of budget and total_reviews, the R-squared increased to 0.338. Adding people_score to the independent variables increased the R-squared to 0.360. In both cases, p-values were much smaller than 0.05. The regression plots also show the same results. Among the variables of our interest, the budget showed the strongest relationship with box_office.



Relationship between revenue and discrete variables

Genres and box office revenue: ANOVA test

Do movies of some genres make more (or less) money than other genres? First, we created a boxplot. From the boxplot, it looks like there are some differences in revenue between different genres. But, is the difference statistically significant?

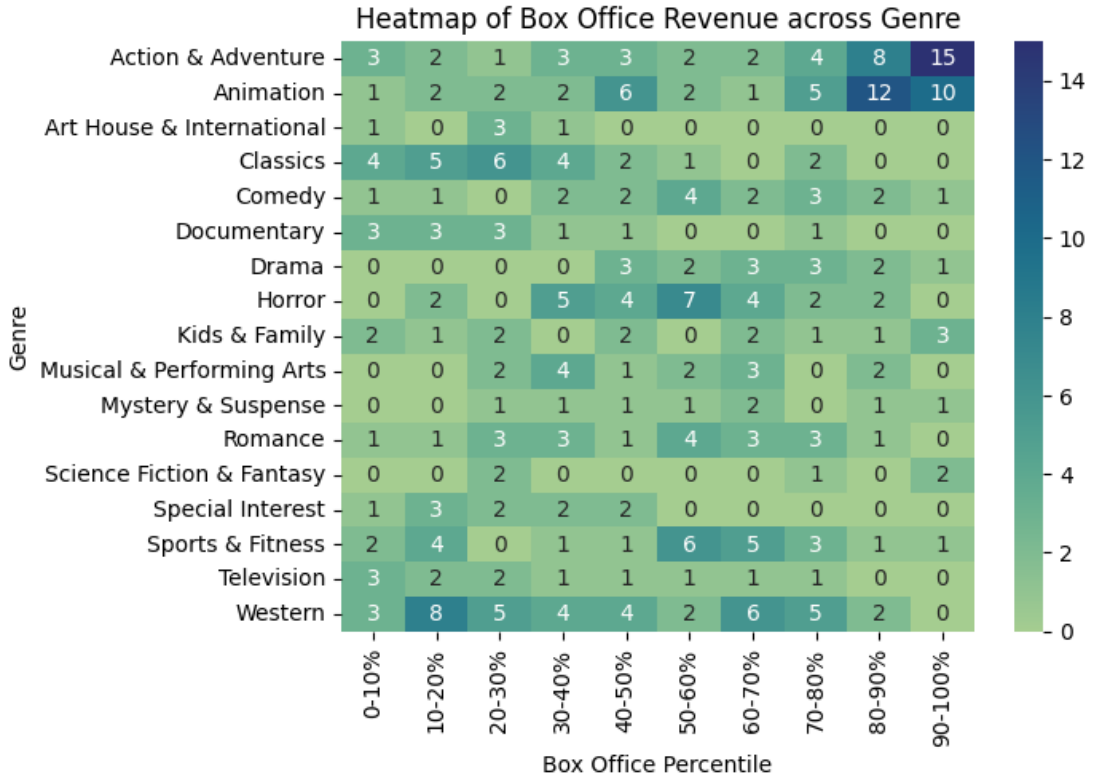


To create a heatmap between the box office revenue and movie genres, revenue was cut into 10 bins according to their percentiles. This enabled the movies to be roughly equally distributed to each bin, thus preventing any single bin from having too much influence on the heatmap. Cutting a continuous variable into discrete bins is not necessarily the best way to visualize the variable because it may result in a loss of information and could potentially affect the accuracy of the analysis. However, we found that the heatmap can complement the boxplot as it can provide a visual representation of the density of data points within each bin. For example, while it seems that the revenues for Science Fiction & Fantasy spread out pretty evenly across the range in the boxplot, we can find that there are 2 Science Fiction & Fantasy movies between the 20-30 percentile of the total revenue, then only 1 movie between 70-80 percentile, and finally 2 movies between 90-100 percentile.

The code to perform this is here:

```
# cut into 10 bins
revenue_cat = pd.cut(
   movies.box_office,
   bins=[0, 619800.0, 2441236.0, 6270000.0, 14371152, 28835451.0,
         45660000.0, 73950000, 117700322, 217040000.0, 940335536.0],
   labels=['0-10%', '10-20%', '20-30%', '30-40%', '40-50%',
```

```
            '50-60%', '60-70%', '70-80%', '80-90%', '90-100%']
)
# create a new column with revenue categories
movies.insert(1, 'box_office_bin', revenue_cat)
# create heatmap
heatmap2 = sns.heatmap(pd.crosstab(
   movies.type, movies.box_office_bin), cmap='crest', annot=True)
heatmap2.set_xlabel("Box Office Percentile")
heatmap2.set_ylabel("Genre")
heatmap2.set_title("Heatmap of Box Office Revenue across Genre")
```



The boxplot and heatmap between box office revenue and genre seem to generally suggest that there is a difference in gross between genres. To have a more accurate answer to this question, we conducted an ANOVA test. The p-value of 2.080638094903269e-12 is smaller than 0.05. Thus, we reject the null hypothesis in favor of the alternative hypothesis and conclude that there is a statistically significant difference in revenue across different genres. Now we know there is a significant difference in revenue between genres, but we don't know which genres have the difference. So, we performed Tukey's HSD test. We have 17 genres, and it's too long to

list all the groups that have statistically significant differences in their box office revenues. Generally speaking, Action & Adventure, Animation, Science Fiction & Fantasy have significant differences with many other genres while Art House & International, Classics, Comedy, Documentary, Drama, Horror, Kids & Family, Musical & Performing Arts, Mystery & Suspense, Romance, Special Interest do not have significant differences with many other genres.

Language and movie scores/box office revenue: t-test

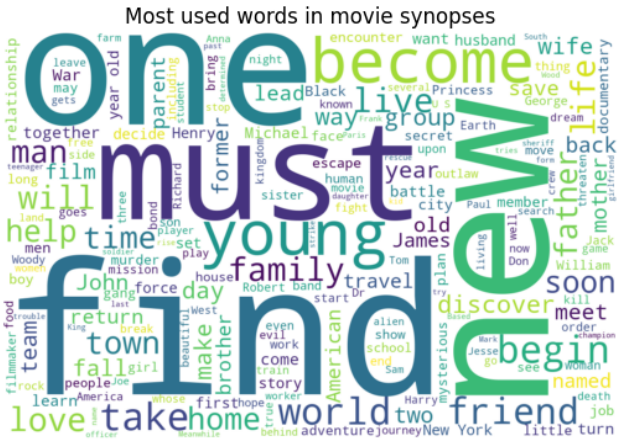
As discussed earlier in the data manipulation section, we have 13 languages where the majority of the movies are in English. Hence, we grouped languages into two categories: English and non-English. We used three variables, namely people_score, critic_score, and box_office for the T-test to compare the means between movies in English and movies in non-English languages. Our results showed that for all three variables, the p-values were smaller than 0.05. Specifically, the p-values were 0.037 for people_score, 0.014 for critic_score, and 0.013 for box_office. These results suggest that there is a significant difference between the mean people_scores, critic_scores, and revenues of movies in English and non-English languages. However, some data requirements for a T-test are violated;

- Normal distribution: People_score and critic_score are skewed to the left while box_office is skewed to the right. However, a violation of normality may still yield accurate p-values in moderate to large samples. Our dataset has 340 samples and we think it would be reasonably considered of moderate size.
- A balanced design: It is ideal to have the same number of subjects in each group. In our case, movies in English have a lot more samples than movies in non-English.

Additional visualization

Wordcloud from movie synopses

We created a word cloud to visualize which words are most frequently occurring in movie synopses. From the synopses of the 340 movies in our dataset, some examples of the most frequently used words are: “find,” “new,” “one,” “must,” “young, “become,” “family,” “begin,” “love,” “take,” “world,” “friend,” “live,” “life,” “father, “man,” “help,” “time,” “soon.”



Conclusion

Overall, our research shows a weak correlation between critic score and box office revenue, with a correlation coefficient of 0.11. This indicates that the relationship between these two variables is not strong. OLS regression analysis also indicates that the critic score is not a good predictor for box office revenue, and there is no statistically significant relationship between the two variables. This is in contrast to the findings from Emerson College (2018) and Chiu et al. (2022) have found critic scores positively affect box office revenue. There could be several reasons for this discrepancy. One possibility is the differences in the samples used in the studies; our sample consists of 340 movies that were released between 1927 and 2016 whereas Chiu et al. (2022) used only 100 movies between 2010 and 2015. Data sources are also different; we used data from Rotten Tomatoes and TMDB (The Movie Database) while Chiu et al. (2022) used data from IMDB (Internet Movie Database) and Box Office Mojo. Emerson College's (2018) research used Rotten Tomatoes, but their sample consists of 1,100 movies and they grouped reviews into two categories: negative reviews and positive reviews. Instead of dividing reviews into two groups, we used the critic scores ranging from 0 to 100.

By using the critic scores, we were able to examine the correlation between the level of critical acclaim and box office revenue. Our findings of a weak correlation between the two variables suggest that although critic scores may have some influence on box office revenue, there are likely other factors that have an impact on a movie's financial success. It turned out that budget has a much stronger correlation with box office revenue than critic scores. The number of reviews also has a significant relationship with revenue. Our ANOVA and T-test results suggest that the revenue has a relationship with movie genre and language. This implies important insight into the complex nature of the relationship between various factors and movie profitability.

Statement of Work

We, Nuri and Lionel, as a team came up with research questions that we aimed to answer through data analysis using our proposed movie datasets. After identifying our research questions, Nuri took the lead in matching each research question with adequate and proper analysis techniques, ensuring that we employed the appropriate methods to answer each question effectively. Lionel took the lead in data manipulation, ensuring that our datasets were cleaned and prepared for analysis. Nuri also contributed to data manipulation, providing valuable ideas and support in ensuring that our data was ready for analysis. Following data manipulation, we both engaged in data analysis and visualization. We worked collaboratively to analyze the data and provide insights that helped us answer our research questions. Our analysis also included the visualization techniques such as charts and graphs to present our findings in a clear and concise manner.

We believe that our collaboration was efficient, effective, and successful. We successfully defined roles and responsibilities, set clear expectations, communicated regularly, and were respectful and open-minded to each other's different backgrounds. This allowed us to work effectively as a team and achieve our objectives. One shortcoming of our collaboration was the use of collaboration tools. We attempted to use git for easier file sharing and version control, but somehow it did not work well for us, and we could not use it. For future collaborations, using Git would be even more efficient.

Overall, we were able to successfully answer our research questions and present our findings through data analysis and visualization. The collaborative effort between Nuri and Lionel allowed us to efficiently and effectively conduct the necessary data manipulation and analysis to answer our research questions.