Group 5 Final Project: IMDb Data Visualization and Analysis

Group Members: Yang Aaron, Saez Dylan, Sveinsson Hrólfur, Miao Rui

Section 1: Introduction and Background: Investigating the movie industry: What contributes to success, and who is our target audience?

IMDb is a popular source for movie, TV, and celebrity content. Essentially, IMDb is an extensive database containing information including, but not limited to, production background, plot, ratings, and movie summaries. According to IMDb's website, the database stores more than 6 million movies and their respective information. Amazon has owned this database since 1998. In this project, we take a look at a .csv file of IMDb information and look to explore the data set to identify key indicators for movie success, such as release months, duration, budget, average movie ratings, and genre. We define *movie success* as US gross income. The team's interest in movies and data visualization and analysis is the primary motivating factor for picking this data set. The visualizations created from this project are intended for audiences that share the team's interest in movies and data.

Section 2: Objectives and Goals

Firstly, we learned how to find a good dataset. Secondly, we learned how to ask a question before visualizing the data. Thirdly, we learned how to visualize data so people can understand it clearly, easily see patterns, and get some insight. Fourthly, we learned how to build a dashboard to illustrate our project. Finally, our project goal is to make USA movies successful. The final goal is to determine the critical indicators for movie success through data visualization and analysis. US movie production companies and their executives are our primary target audience. Based on our recommendation, these stakeholders can use our analysis to improve their movie production process and decision-making, resulting in higher US gross income.

Section 3: Dataset

The data was downloaded from a Kaggle competition. Specifically, the dataset titled "IMDb movies.csv". The initial dataset has 85,855 rows and 23 columns. Since the dataset has around 85,000 rows and several columns containing null values, subsetting the dataset was needed to organize the data. After collaboration, the team decided to pick these columns: title, year, date_published, genre, duration, country,

language, actors, description, avg_vote, budget, usa_gross_income, and worldwide_gross_income. After eliminating the extra columns, the team subsetted the data, keeping only rows that had data in every column, which left the team with a final data set of 8,099 rows and 13 columns. There are no locations in the final dataset, however, 'country' existed in the initial dataset.

Section 4: Questions

The questions we will ask of the data are posed as recommendations that our target audience may take into consideration when producing/planning movies. The variable of interest that we will analyze (dependent variable) is US Gross Income. Therefore, in order for our target audience to maximize US Gross Income, we will provide them with insights from the dataset.

Question 1: Are the number of movies produced increasing over time, and if so, are the average ratings also increasing over time? Why or why not?

From figure 1 below, we can see a slow, increasing trend in movies produced from 1945 to 2006 as can be seen from the upward trend for the line chart. The number of movies produced was 326 in 2006 but was only 1 in 1945. These are small numbers because the data has been subsetted. However, when thinking about the year 1945, we can add context to these numbers by hypothesizing that demand for movies was low during the world war periods between 1914 and 1945. When the economy started to turn around, and people could start living normal lives again, we speculate that those two factors promoted movie production growth.

There is a slow, decreasing trend in average rating over time, as indicated by the green to the red color difference in figure 1. This also holds when looking at the original data with around 85,000 observations. Why is that? Maybe people feel that older movies are more genuine, original, and creative and that they were produced with more passion than newer movies. The rush to produce nowadays and other factors such as originality could be affecting the audience's feelings towards new movies.

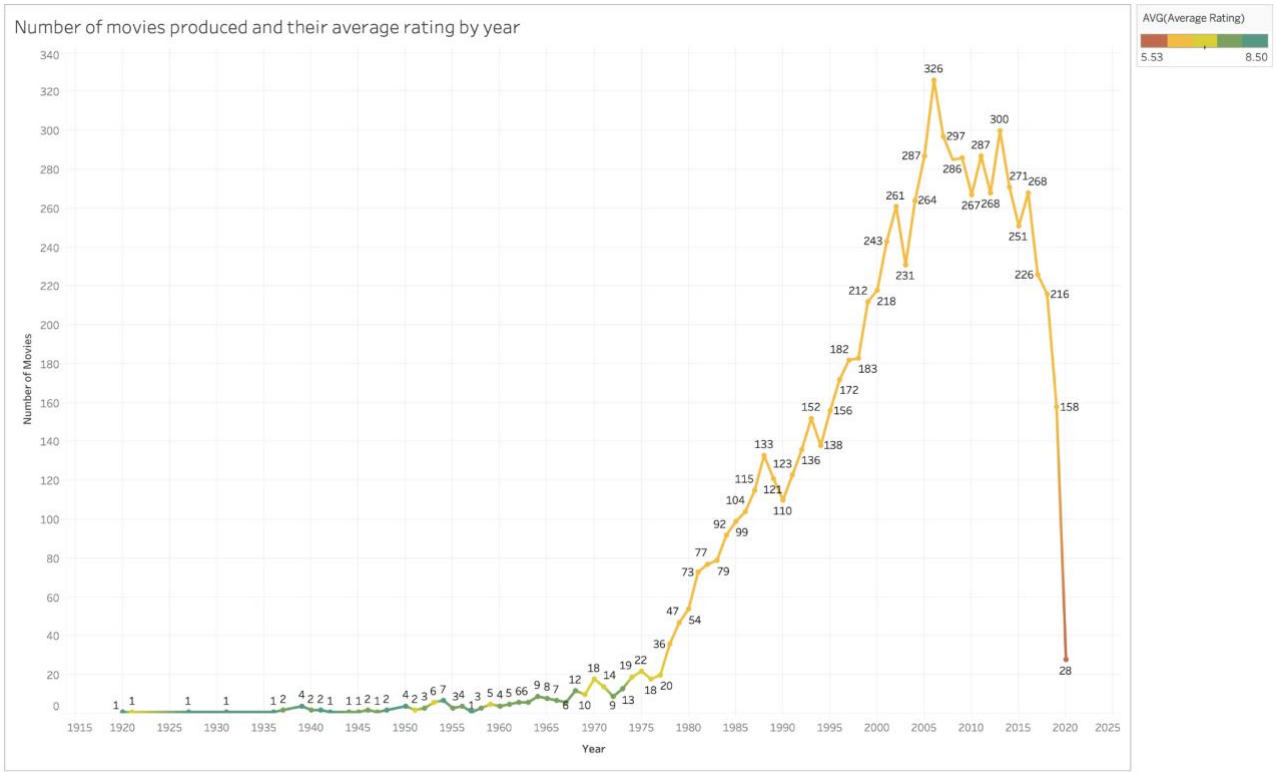

Figure 1: Number of movies produced and their average rating by year.

Question 2:

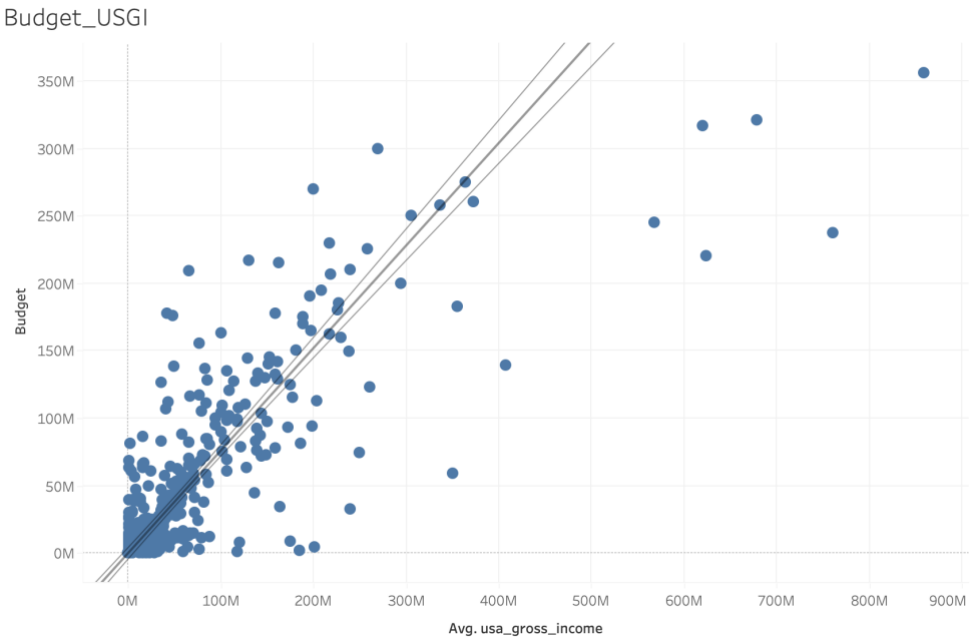

Figure 2: Correlation between Budget and Average US Gross Income

a) What is the relationship between US Gross Income and Budget?

Based on the data, as the budget for movies increase, average US gross income increases as well. Therefore, a positive relationship may be concluded to exist between these two variables. In this graph, the aim is to show how the target audience may want to consider their budgets for movies. If they typically have a wide range of budgets, maybe they would like to save their resources for higher-budget films to maximize their US gross income. This is important when it comes to financial planning and budgeting.

b) What is the relationship between US Gross Income and Duration?

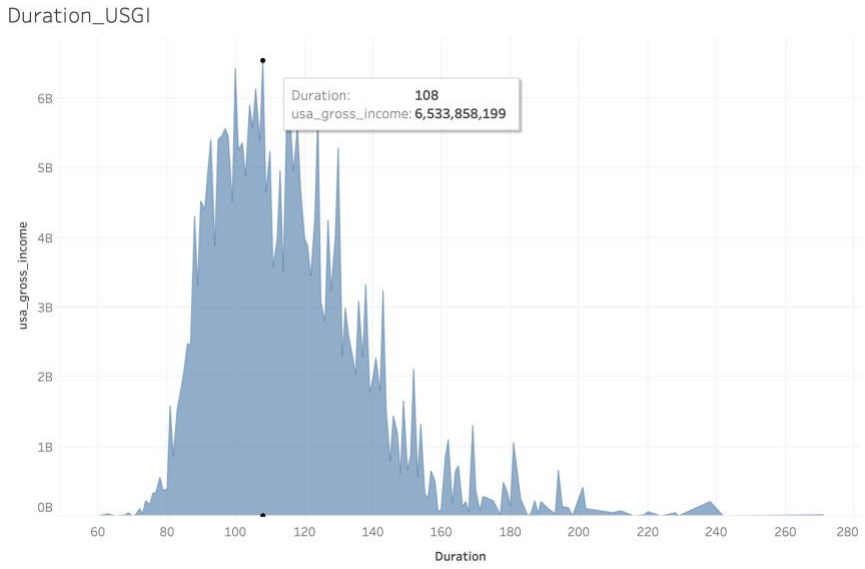

Figure 3: Duration (minutes) and US Gross Income

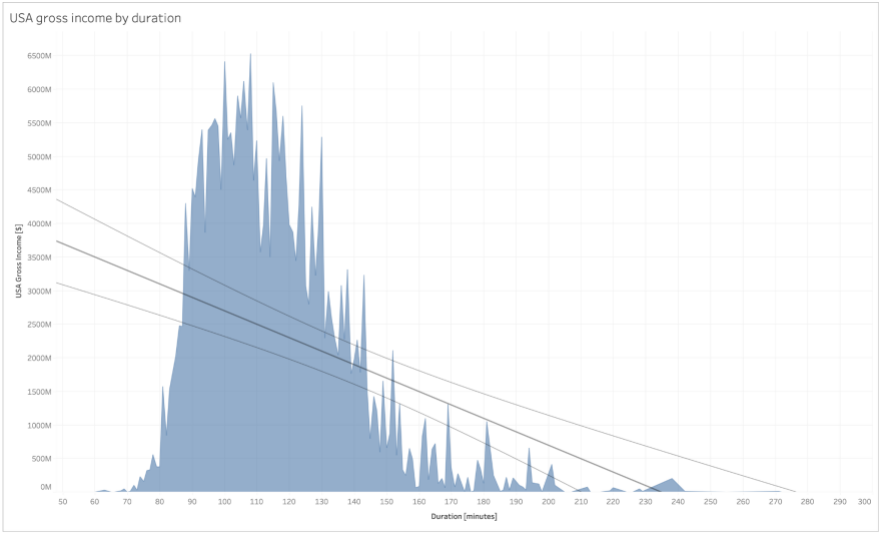

Figure 4: Duration (minutes) and US Gross Income with Trend Linei

Based on the data, movies with a duration of 108 minutes have the greatest sum of US gross income out of all other durations listed in the dataset. The second graph including a trend line indicates: as duration increases past 108 minutes, budget begins to fall; prior to 108 minutes, budget increases. This graph aims to reveal to the target audience a time range they may want to focus on when planning a movie. This is important because it may give the production company/executives insight into the attention span of audiences and when to cut off a movie's duration.

Question 3: What combination of genres has the highest average US Gross Income in the U.S.?

After counting the combinations from hundreds of different permutations, we found out the genre combination that has the highest average gross income is Family and Sci-Fiction. A typical example of this kind of movie is E.T. - L'extra-terrestre, a very classic movie which was published in 1982. This combination of genres satisfies the majority of peoples' preferences, science fiction and the family members can have fun together.

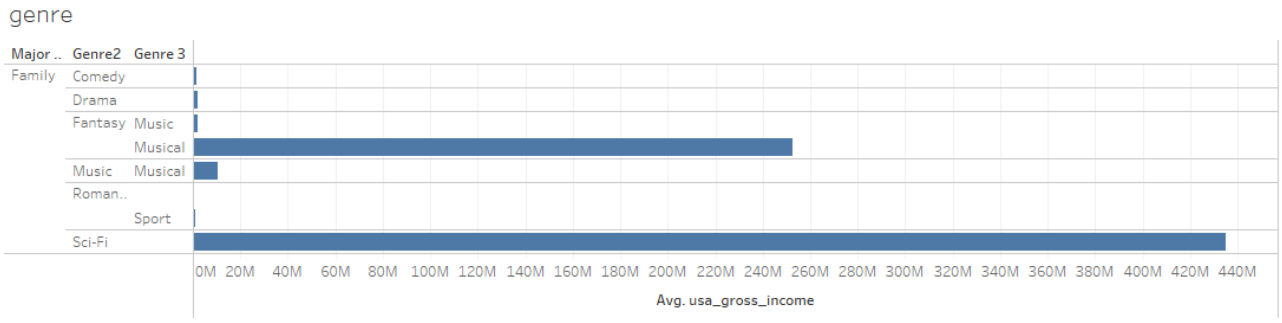

Figure 5: The genre combination that has the highest average gross income.

Question 4: When is the most profitable time to release a movie?

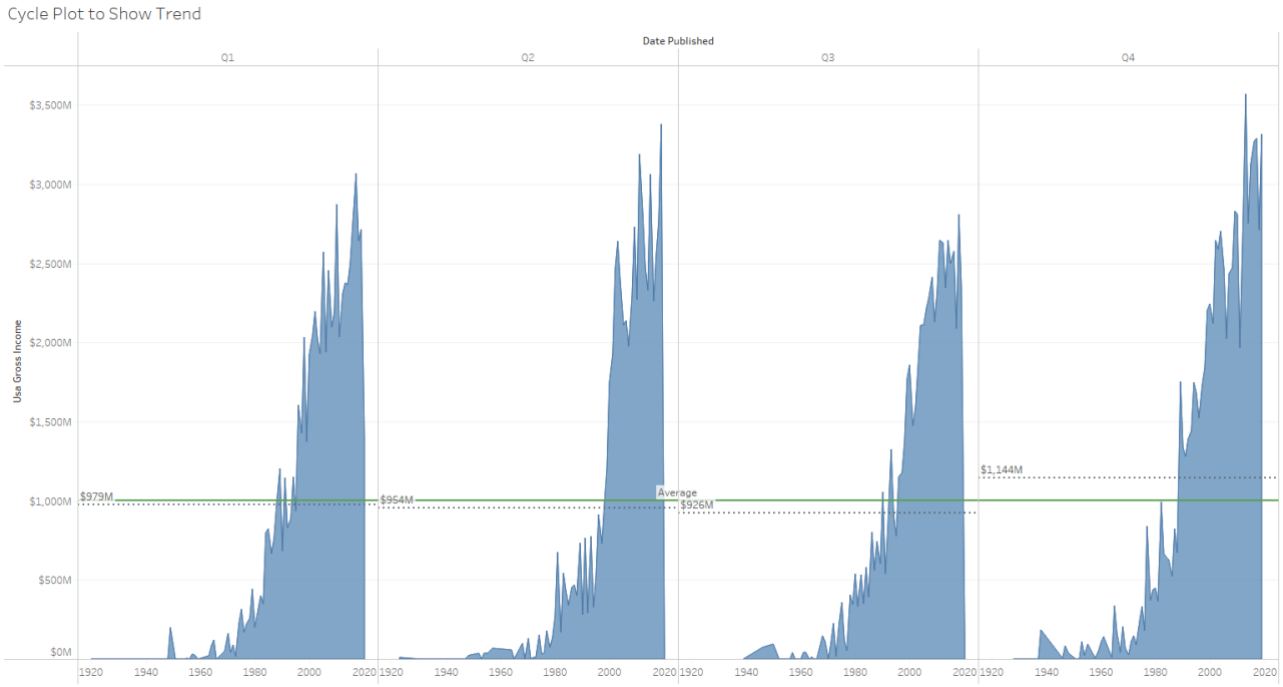

Figure 6: Cycle plot above, we see the highest average quarter of US gross income is the 4th quarter.

If we used a line chart to show a trend, we could see the trend across a year. But all the additional insights would be missed. Although the line chart made sense, it would be hard to see details and let movie company executives decide. Thus, we chose to use a cycle plot to show the trend. We could see each quarter and whether it increased or decreased. Also, we could compare each quarter to see which had the most and the least US gross income.

As you can see from the above cycle plot, we calculated average US gross income for each quarter and average US gross income from 1920 to 2020. The cycle plot displays that the highest average US gross income is $1144 million in the 4th quarter. Second is $979 million in the 1st quarter. Third is $954 million in the 2nd quarter. Fourth is $926 million in the 3rd quarter. Also, we draw the green line for average US gross income. The cycle plot discovers only the 4th quarter has above average US gross income. Thus, we know that the most profitable time for the movie market is the 4th quarter of a year. Therefore, we recommend that movie company executives pay attention to the 4th quarter movie market.

Question 5: What is the relationship between US Gross Income and Average Rating?

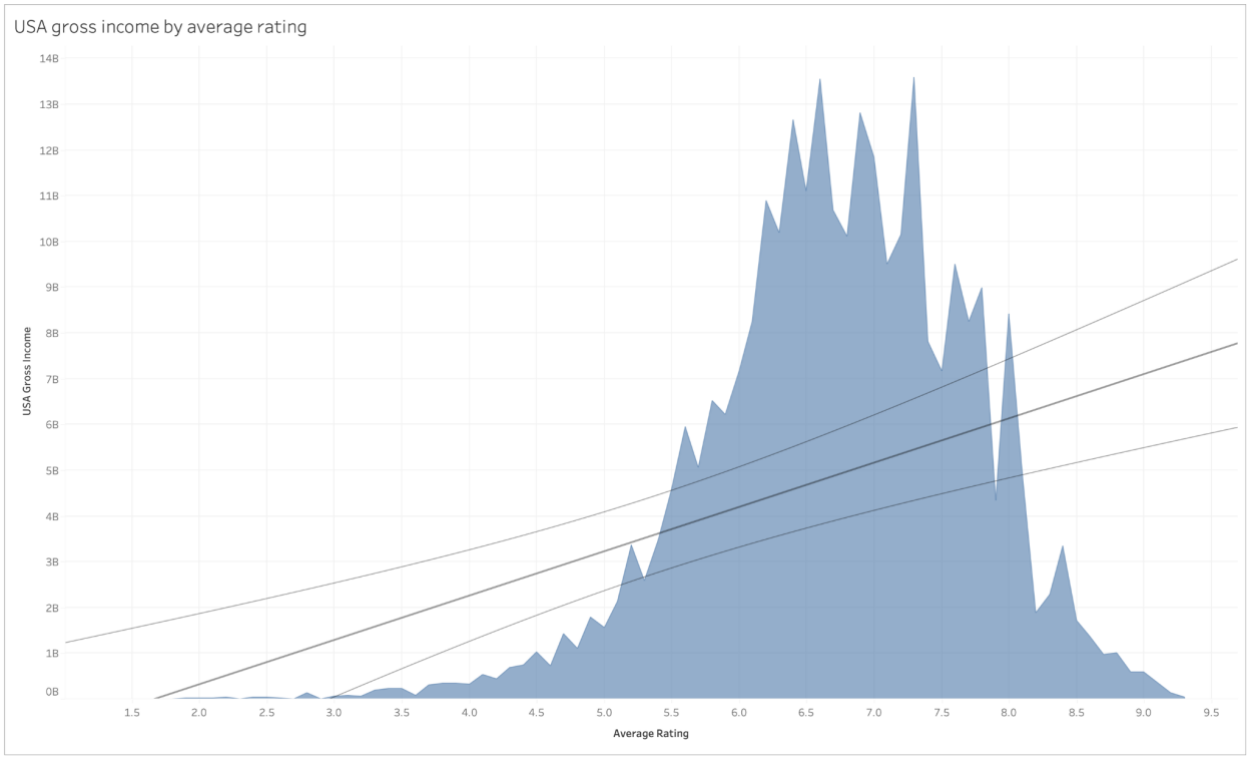

Figure 7: Average Rating and US Gross Income

From figure 7 above, we can see that as the average rating increases, the sum of US gross income increases. Movie producers should focus on producing high-quality original content that fans appreciate.

References

Boulenger, T. (2016). An analysis of movies by ratings.

https://rpubs.com/ashtom/movierating

Kaggle. IMDb movies extensive dataset. [Dataset].

https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset?select=IMDb+movies.csv

Mccoy, S. (2021). Murach's Python for Data Analysis. Tableau & Python integration.