## Project 1 – Movie Data Analysis

Jesse Heaton, Jonathan Koo, Khyati Thakore

Written Analysis:

How does a movie’s rating affect its performance in the top 1000 grossing movies? With the graph of each movie’s rating, we can tell a couple things about the ratings of a movie. For example, critic ratings are usually lower than watcher ratings by a couple integers. Also we can notice that almost all movie ratings are closer together and that there are no outliers in our ratings with a very high watcher rating and a very low critic rating or vice versa since we have an r-value close to 1. Other than these observable details/patterns, the rating seems to have no real effect on how well a movie does.

You may be asking, does the studio that created a movie matter for how it will perform? Yes to a certain extent. This is because a large majority of the movies in the top 1000 grossing movies, 4 studios (Disney, WB, Universal, Paramount) have over half of the movies in the top 1000. However When comparing the studios to their movies ratings, there is almost no discernible difference between the ratings since they seem to average out quite nicely at around a 7 with the more ratings that each studio has.

Looking at the box office sales categorized by release date, we decided to break this down by month released and year released. We observed that total box office sales steadily increased in the years measured, but also decided to look at the total movies in the dataset released in each year and the average box office sales per year. The total movies released per year also generally increased year by year, with the bar graph looking similar to the total box office sales per year. Looking at the average box office sales, however, there still appears to be a general upward trend during the years measured, with a few outlier years in between.

The analysis broken down by the month shows the total box office sales are greater in the summer months (May, June, July) and holiday months (November and December). We also looked at average box office sales per month of release, and observed generally the same pattern, with April also having a high average for box office sales. This makes sense since those are months when people are likely to have more free time to see movies.

As shown in the graphs for rating, the total box office sales for PG-13 movies was significantly higher than the rest. We also wanted to look at the average box office sales per category to account for simply more movies in one category. Comparing the average box office sales, PG-13 was still the highest average grossing category, followed closely by G and PG rated movies. This makes sense, as those three categories are more age inclusive and can cater to a wider audience.

We also wanted to examine the effect of runtime on box office sales. With a correlation coefficient of 0.26, there is a weak positive correlation between them. Meaning, that the runtime of a movie has little effect on the success of a movie. The graph also shows that a movie should run for at least 70 minutes if it wants to break 100 million in the box office. Another thing studios may want to keep in mind is that movies that run longer than 200 minutes see less sales than normally predicted.

Next, we checked how the genre could indicate box office sales. We examined the genres of successful movies; with Star Wars, Avatar, Black Panther, and 2 Avengers movies in the top 5 highest box office sales, it’s no wonder that action and adventure movies have some of the highest bars. The other bar, Animation, covers many family-related content which also seems to prove successful. The least successful theatrical releases are documentaries.

Limitations of our analysis:

We began looking at the data from a specific dataset from Kaggle, so we are limited to the list of 16,000 movies from the list. From there, we only looked at the top 1,000 movies from that dataset, per the API limit for 1,000 calls/day. Further, not all the 1,000 movies that we tried to search for were found in the OMDb database. Our original dataset also only included movies up through 2019, not accounting for any trends from the past three years. Our analysis has some limitations as well, such as using the box office sales and ratings as our primary metrics of "success" of the movies. There are also other factors which we could have analyzed here, such as the influence of top actors on the box office success, the directors/producers of the movies, awards - which would have required additional data that we were not able to find or more complex analysis.

If we did not have the above limitations and wanted to do some further analysis, we would first increase the number of movies we include in the data to get more accurate conclusions. We could look at other metrics for success, such as awards won. We could have also analyzed the influence of the actors, directors, and producers on the success of the movies. If we were able to find proper data relating to the budget of each film, we could compare that to the box office sales to observe any trends and calculate net earnings for each movie. Since our original dataset did not include global data or streaming data, we could find relevant data and do the same analysis for global movies and streaming movies and viewing. We could also try to find more current data and observe trends from the past three years, when the pandemic changed both the industry and the way consumers watch movies.