**Data Wrangling Project Final Report**

**IMDB Movie Reviews and Gross Earnings**

1. **Introduction**

With the recent pandemic becoming more stable, this is allowing movie theaters to open again. The virus caused many obstacles when it comes to the film industry. Our project could be a valuable resource to these companies when deciding when to release their productions to maximize their return. In this project, our group plans on using a dataset that includes IMDb Top 1000 movies. We can then examine the relationships between ratings, revenue, time released, along with other attributes that can provide supplemental insights to movie corporations.

In this project, we utilize data on the characteristics of movies along with their various respective worldwide gross earnings to discover trends that production companies can consider when producing their next film.

Research Questions:

* What is the distribution of movies and scores across different genres?
* Is there any relationship between IMDb ratings and worldwide gross earnings?
* Is there a relationship between the Director’s that have produced multiple films and the gross earnings for their films?
* Is there a relationship between the length of a movie and the gross earnings of that film?

1. **Data**

This project consists of two datasets. There was an existing dataset on Kaggle that contained characteristics of popular movies including specific IMDb ratings, and Box Office Mojo statistics about revenue for the films.

*2.1 IMDB Movie Reviews*

The first source of data we will use for this project is an existing set of data from *Kaggle* that consists of movies with the most popular IMDb ratings over time. The data set is a collection of 1000 observations with 12 variables that display the rating, stars, and director among other attributes that will help us solve problems and identify relationships between film ratings and in the film industry.

The format of the file from Kaggle was downloaded as a csv file. This made it easy to open within excel and eventually uploading it in RStudio for data transformation and integration. There were a few irrelevant features in the data we decided to ignore such as a link to the movie poster, certificates earned, and overview of movie. We did not include gross revenue in the current dataset since that will be something we generate from web scraping.

*2.2 Worldwide Gross Earnings*

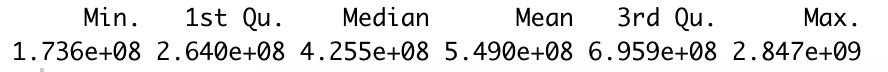
The second data source consisted of scraping data from “boxofficemojo.com”. Specifically, we used top lifetime grosses that had a list of 1000 movies. There were 5 different pages that consisted of 200 movies on each page. From this website, we had to inspect the page source of HTML code to properly scrape the data using specific URLs and loops to crawl through all the pages of movie lists.

Box Office Mojo shares statistics in 3 relevant categories. (Worldwide, Domestic, and Foreign) For the project, we chose to only include the worldwide lifetime gross earnings statistic. We also pulled the titles for each movie listed on the top lifetime grosses. It was not necessary to remove any columns since there was choice in what data is being scraped. The complete data set consisted of the titles and the corresponding worldwide lifetime gross earnings. There was 1000 observations and 2 variables.

*2.3 Summary Statistics*

We included summary statistics of Gross Earnings because it is a highlighted feature in our data (Figure 1). There is a histogram to show the distribution of gross earnings (Graph 1). We conducted a hypothesis test for normal distribution. This was done through a Shapiro test. The null hypothesis is that the gross earnings feature is normally distributed. After conducting the hypothesis test, we obtained a p-value of 2.2e-16 which is below our alpha of 0.05 so we reject our null hypothesis and conclude that the gross earnings variable is not normally distributed.

*Figure 1 Summary Stats of Gross Earnings*



*Graph 1 Gross Earnings Distribution*

Chart, histogram

Description automatically generated

*2.4 Integration of IMDb Reviews and Gross Earnings*

To successfully integrate these datasets, we only utilized the common movies between the two sets. We were able to use intersect functions within R programming to find the shared movies between our existing dataset and the one we scraped from Box Office Mojo. Overall, the final integrated data set has 193 observations and 13 variables. A data description for each variable is contained in Table 1. The code that was used to scrape, clean, and merge the datasets is included in the R script “WranglingProject.R”

*Table 1 Data Dictionary*

|  |  |  |
| --- | --- | --- |
| **Column** | **Type** | **Description** |
| Series\_Title | text | Name of the movie |
| Released\_Year | date | Year that movie was released |
| Runtime | numeric | How long the movie is in minutes |
| Genre | text | Genre of the movie |
| IMDb\_Rating | numeric | Rating of the move given on the IMDB site (1-10) |
| Meta\_score | numeric | Metacritic score (1-100) |
| Director | text | Director of the movie |
| Star1 | text | Lead Actor 1 |
| Star2 | text | Lead Actor 2 |
| Star3 | text | Lead Actor 3 |
| Star4 | text | Lead Actor 4 |
| No\_of\_Votes | numeric | Total number of reviews |
| GrossEarnings | numeric | Worldwide lifetime gross earnings |

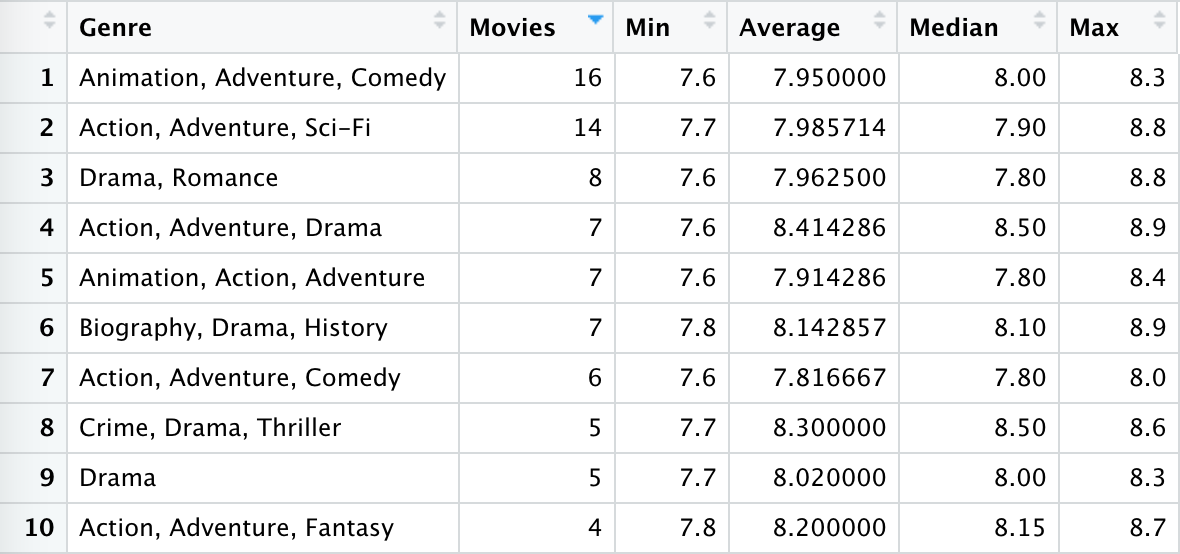
1. **Analysis**

The goal of this project is to analyze top-rated and top-grossing movies to determine whether there is any relationship between IMDb ratings and world-wide gross earnings.

*3.1 Movie Ratings and Earnings by Genre*

What is the distribution of movies and scores across different genres? We created a dplyr summary table that calculates the number of movies within each genre, and summary statistics (minimum, maximum, average, and median overall rating) for the IMDb ratings. Since there are seventy-nine different combinations of genres, we focused on the top ten to get an idea of the most common types of movies included in our data. Table 2 shows the top ten genres by movie totals.

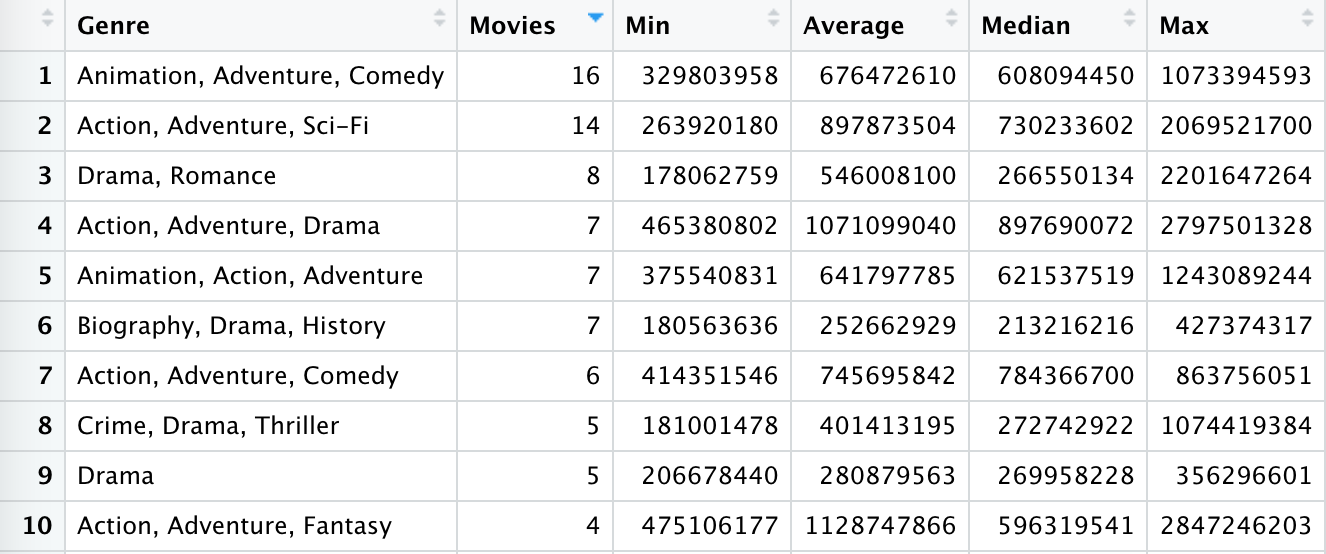
*Table 2 Genre Rating Summary (top 10 rows, 79 rows total)*



The summary table shows that most movies included in both IMDb’s highest ratings and Box Mojo’s top gross revenue earner are Animated, Adventure and Comedy films. There were sixteen movies within this genre, and while they did not have the highest average rating, it’s important to note that these are the most common genres within our collected data. This table shows that the more common genres do not include movies that have the highest average ratings, so genre has little to no impact on the ratings the movies received.

To extract more information about the genre groups, we also decided to create an identical table that displays summary statistics of worldwide earnings to see if genre plays any role in determining worldwide gross earnings. Table 3 below shows the top ten genres by movie totals.

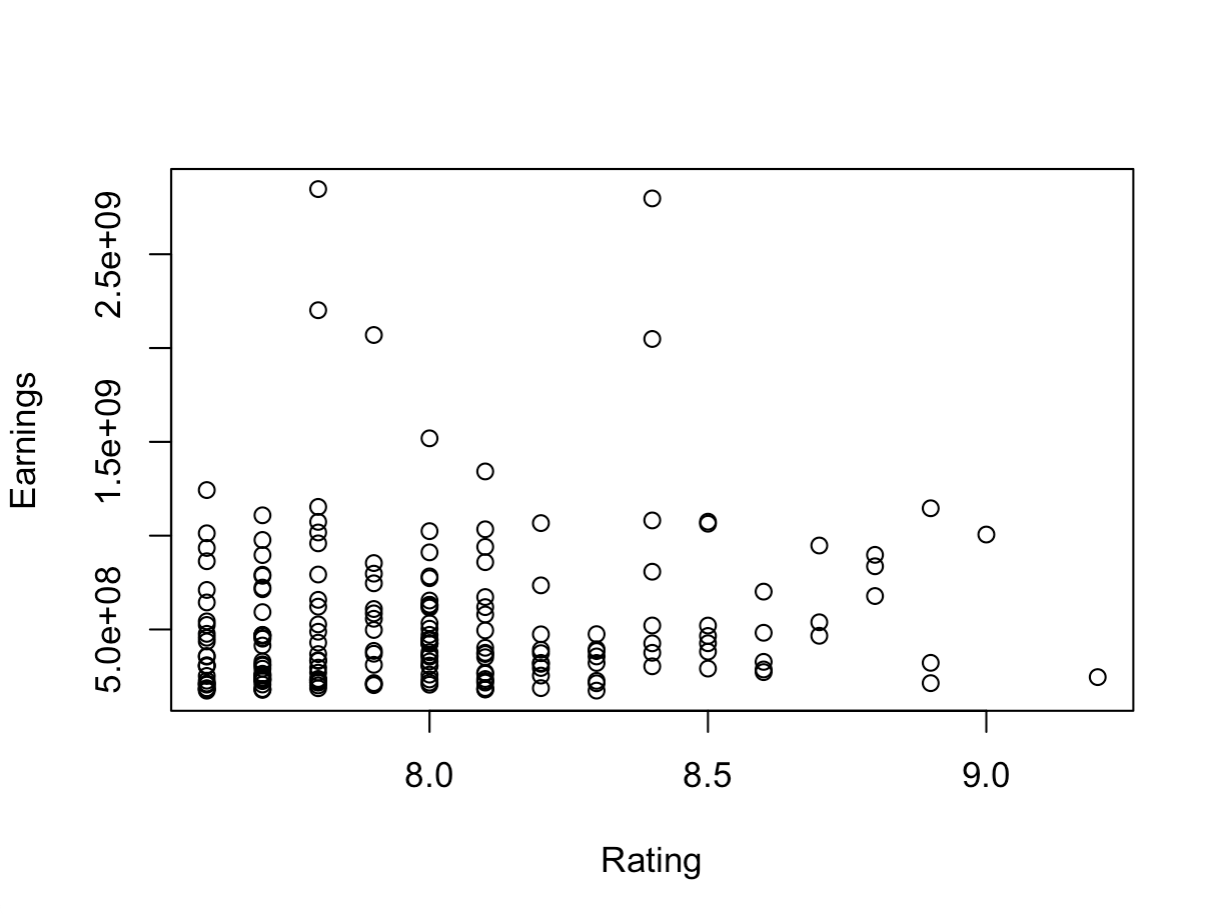
*Table 3 Genre Earnings Summary (top 10 rows, 79 rows total)*



The summary table above helps to display that even the most popular genres include movies that did not perform as well as movies that belong to other genres. For example, the top genre Animation, Adventure, and Comedy has large minimum earnings, but have an average much lower than the 2nd, 4th, and 5th ranked genres. The maximum is also much lower than some of the other genres in the top 10. This information helps to prove that the more common genres do not include movies that have the highest earnings, so genre has little impact on the gross earnings of the movies included in our data.

*3.2 Ratings vs. Earnings*

Is there any relationship between IMDb ratings and worldwide gross earnings? We created a scatter plot chart to plot ratings against gross earnings to see if any positive relationship exists. We then used the correlation function to calculate the true correlation between ratings and earnings to give a definitive measure of the relationship that exists. Graph 2 below displays the scatter plot and correlation results.

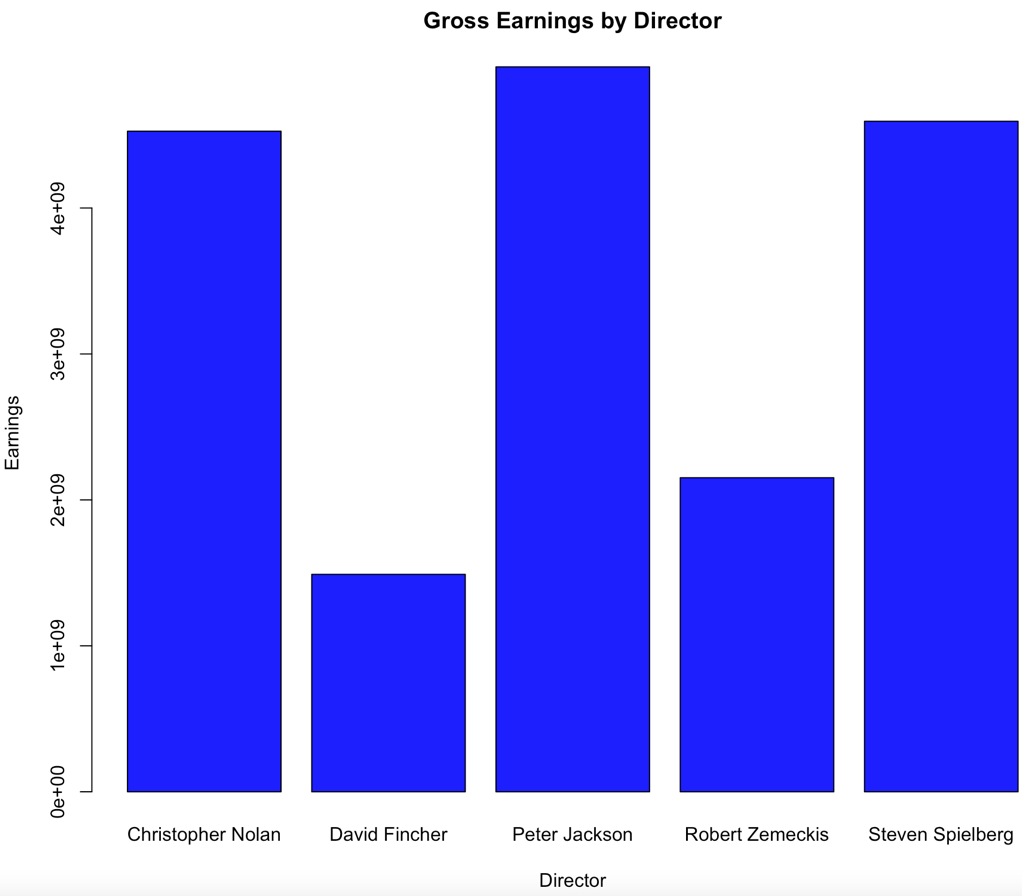
*Graph 2 Ratings and Earnings*

The scatter plot above visualizes the distribution of ratings vs. earnings, helping to see the relationships and outliers within the data. The plot does not clearly display a positive relationship between ratings and earnings, so we utilized correlation to provide evidence on whether a positive or negative relationship exists between the two variables. The correlation between ratings and gross earnings was 0.1040319, which is a weak positive relationship. This information confirms that the movies included in our data have a positive relationship that exists in which when a movie receives a higher IMDb score it is predicted that it will also generate more earnings than lower rated movies.

3.3 *Directors vs. Earnings*

Is there a relationship between the Director’s that have produced multiple films and the gross earnings for their films? Within our dataset, there are 120 different directors, some of which have directed multiple films. We analyzed the top 5 directors that have the highest frequency of films directed. The top 5 directors include: Steven Spielberg (9), Christopher Nolan (6), David Fincher (5), Peter Jackson (5), and Robert Zemeckis (5). It’s logical to assume that the directors that have produced the most films would generate the highest earnings; however, we want to interpret these results further. Therefore, we will analyze this correlation through a column chart that sums the total earnings for the five most common directors. See Graph 3.

*Graph 3 Directors vs Earnings*



Based on this column chart, we can conclude that the frequency of films directed is inconclusive of the total gross earnings. Steven Spielberg has directed the most films (9) and the second closest to him was Christopher Nolan (6). Yet, Peter Jackson, (who only directed 5 films) is the clear leader with the highest amount of gross earnings. This is not surprising as all 5 of the films that Peter Jackson directed were among the top 8 films with the highest gross earnings from the sample that we looked at (see Table 4 below).

*Table 4 Top 8 Highest Grossing Films*



3.4 *Movie Length vs. Earnings*

Is there a relationship between the length of a movie and the gross earnings of that film? We used multiple methods of evaluation and visualization to find results of how runtime is related to gross earnings. A scatterplot was generated using the qplot function within the ggplot2 library to visualize the relationship between runtime and earnings that movie grossed. (Graph 3) The correlation function was also used to put a value on this relationship. Another visualization was used to categorize the movies into short, medium, or long movie lengths. We can then use a bar plot to compare the length type to gross earnings. (Graph 4)

*Graph 4 Runtime & Gross Earnings*

Chart, scatter chart

Description automatically generated

We can pull various interpretations from the scatterplot. Majority of the movies falls within 100 – 150 minutes which holds the average runtime of 128.72. A slightly positive trend is observed from the graph but there are also not many data points with a longer movie length of 175 minutes. A correlation value between runtime and gross earnings is 0.139167 which indicates a weak positive relationship. There are a few outliers but do not affect the overall conclusions that runtime affects gross earnings.

*Graph 5 Length Type & Gross Earning Average*

Chart, bar chart

Description automatically generated

Background information behind graph 5 includes how the length types were created. Movies that range from 0-100 minutes were labeled as “Short”. Movies that are between 100-150 minutes are labeled as “Medium”. Movies that are greater than 150 minutes are labeled as “Long”. The average gross earnings were used to aggregate each length type. We interpret the graph to show that long movies have the highest average gross earning value, but this could be due to the small number of long movies within the dataset resulting in skewed averages.

1. **Conclusion**

This project incorporates data from Kaggle and Box Office Mojo. These two datasets contain various characteristics for popular movies such as ratings, movie titles, revenue, etc. We downloaded the Kaggle dataset as a CSV which allowed us to transform and analyze it in R Studio. We ignored certain features that did not contribute to our end goal. Our second dataset from boxofficemojo.com was a collection of the top 1000 movies characterized by lifetime gross revenue. We scraped our second dataset by inspecting page source code and scraping specific URLs through loops to create a readable movie list. Next, we integrated the IMDB reviews and gross earnings using the intersect function within R Studio to find the movies that were found in both datasets. Our integrated dataset consisted of 193 observations/movies and 13 total variables.

For the analysis portion, we decided to determine whether there was a relationship between IMDB movie ratings and the movie’s world-wide gross earnings. We created a dplyr summary that gave the top 10 genre ratings by number of movies. We can conclude that genre does not have a high impact on the rating of the movie from our dataset. We created a table with the same genres but categorized it by gross earnings. Again, this resulted in genre having little impact on gross earnings on a movie from our dataset.

Next, we wanted to analyze whether there was a relationship between IMDB ratings and world-wide gross earnings by creating a scatter plot that showed the correlation between rating and earnings. We concluded from this analysis that the higher a movie’s IMDB rating, the higher the ratings that movie will receive.

Following this, we analyzed if there was a relationship between directors and earnings. We found inconclusive evidence that the frequency of films directed by a single director was correlated with total gross earnings. However, on table 4 it shows that certain directors have a higher likelihood of producing higher gross earnings.

Lastly, we analyzed the relationship between movie length and earnings. We generated a scatter plot using the qplot function. Besides a small number of outliers, we concluded that the correlation between run time and gross earnings has a weak positive relationship meaning that run time does not affect gross earnings greatly. Next, we created a bar plot between different movie length type and gross earnings. Once again, we found a weak positive relationship that included a small number of outliers but found that movie length type does not highly affect gross earnings.

Some limitations that we had were not enough variables in one of our datasets that could cause skewness in our analysis. Another limitation that we encountered was that only 193 observations were found on both datasets. If we could’ve had a higher number of observations on both datasets, our analysis might have explained more. For future work, we could scrape a larger number of websites to get more detailed observations. We could also find better correlations using country, region, state, etc. This might give us a better insight into film companies into who or what to edit within their movies. We believe that the analysis we did would give many companies in the film industry a better understanding of what variables produce the highest gross earnings and movie rating.

Sources

Scraped Data:

<https://www.boxofficemojo.com/chart/top_lifetime_gross/?ref_=bo_cso_ac>

Existing Dataset:

<https://www.kaggle.com/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows>