Nico and Jacob Project Report

IMDB Highest Grossing and Highest Rated Films of All Time

1. Introduction

IMDB is one of the most popular movie critic platforms and news sources in the world. Individuals often reference and utilize IMDB in order to make informed decisions on new movie releases and old ones. IMDB themselves ranks movies in a variety of lists, but does ranking have anything to do with the amount of money a film makes? Without taking a deep dive one might think that a higher rated movie would make more money, more people want to watch better movies, right? We wanted to find out how strong a correlation there is between the amount of money a film makes and how highly it is ranked in the top 1,000 movies of all time. We also wanted to find out if a director appeared more than once on both lists. Is it possible some directors appear more often than others and who appears the most? We also wanted to find out what the average grossing amount was for each audience rating. Do films that are rated differently gross more or less on average?

In this project we plan on using IMDB's lists to better understand the relationship between moving rankings and box office moneymakers. We will be scraping two separate lists and putting all the acquired data into two separate data frames prior to answering our questions.

2. Data

In this project we utilized two sources of data. The first one, IMDb's Top 1000 list of highest rated movies of all time and the second one, IMDb's Top 1000 Highest-Grossing Movies of All Time.

2a. *IMDb Top 1000*

The first website we scraped to collect the necessary data was the IMDb's Top 1000 list. We utilized Selenium and Pandas, both Python libraries with unique tools, inside of a Jupyter Notebook to execute the code required. From this site we scraped the movie's rank, the movie's name, the release date, and the audience rating for said movie. The name of the web scraping script is called *00_IMDbT1_scraper.ipynb* and it is located within the python notebooks folder. The data that was scraped was initially saved to a data frame within the scraping file. It was then saved to a comma separated values file called *IMDbT1_raw.csv*, located within the data folder.

This data file did require data cleaning, all be it a very small amount. During the scraping process, we realized that the rank was attached to the movie name within the html page. This was separated and cleaned post scrape. We also had to adjust the columns types as they were never set and we needed them to make sense logically in order to do our analysis. We set rank to an integer, movie name to a string, release date to an integer, and audience rating to a string. We did all of this within the same scraping file, *00_IMDbT1_scraper.ipynb* before being saved to its final form csv file.

2b. IMDb Top 1000 Highest-Grossing Movies of All Time

The second website we scraped to collect the other half of our necessary data was IMDb's Highest Grossing list of 1000 movies. Like the first site we scraped, we also used Selenium and Pandas libraries to have the right tools within Python to make this happen. From this site and respective list, we scraped the rank, movie name, amount grossing, and director(s). This web scraping script is named *00_IMDbHG_scraper.ipynb* and it is located within the python notebooks folder. The data was scraped into a data frame within the scraping file first then ultimately saved to a comma separated values file called *IMDbHG_raw.csv*, it can be found in the data folder alongside the other csv file.

This data file required slightly more cleaning before we were able to utilize it. Like the first data file we separated the rank that was attached to the movie name and adjusted column types to be integers or text for respective columns. We also had to remove the dollar sign in the grossing amount column to properly set it as an integer column type and get rid of the unnecessary symbol as it would play no part in our analysis. This cleaning was done in the same scraping file named *00_IMDbHG_scraper.csv* before it was saved into a final csv file.

3. **Setting Up Data Frames for Questions**

Some of the questions we were looking to answer required the combination of both data frames, as expected. We started this by analyzing how and where our combination would occur. Both data frames contain a movie's ranks and a movie's name. We decided to merge the data frame here and take the IMDb Top 1000 Highest Rated list rankings and movie names and combine them with the respective gross earnings and director(s) if they appeared in the Top 1000 Highest Grossing list. We found this narrowed our data set down to 208 movies. So now we can tell that there were 208 movies that appeared in both lists, and we had a dataset that we can work with in order to answer two of our three questions. We named this data frame final_df as it is the final merging we needed to do in order to answer the first two questions we had. This was done in *01_Creates_CSV_For_02_and_03.ipynb* prior to creating each question solution script. That final data frame we used to answer our first two questions was saved to a csv file called *merged_df.csv*, and it is saved in the data folder. A table of this dataset is below in the Merged Data Frame Dictionary for Table 1.

Table 1: Merged Data Frame

Column	Type	Source	Description
rank	Text	IMDbT1	Rank of the movie 1 through 1000
movie_name	Text	IMDbT1	Name of the movie
gross_earnings	Numeric	IMDbHG	Gross earnings of the movie
director(s)	Text	IMDbHG	Director(s) for the movie

3a

Our third question was a bit more unique in terms of what data it required. While it did not need data from both data sets, and it only required data from IMDb's Top 1000 Highest Grossing list the data was still altered from what was originally scraped. We will touch on what we did to achieve this unique data frame in the question and analysis section. Below is the dictionary for Table 2 which contains all the data needed to answer our third question. The set up for this dictionary was done in the *03_Question3_answers.ipynb* file prior to finding the answer to our question inside of the python notebooks folder.

Table 2: Grouped / Average Data Frame

Column	Type	Source	Description
audience_rating	Text	IMDbHG	Audience rating for said movie: PG-13, R, G, etc.
average_gross_earnings	Numeric	IMDbHG	Average gross earning for said audience rating

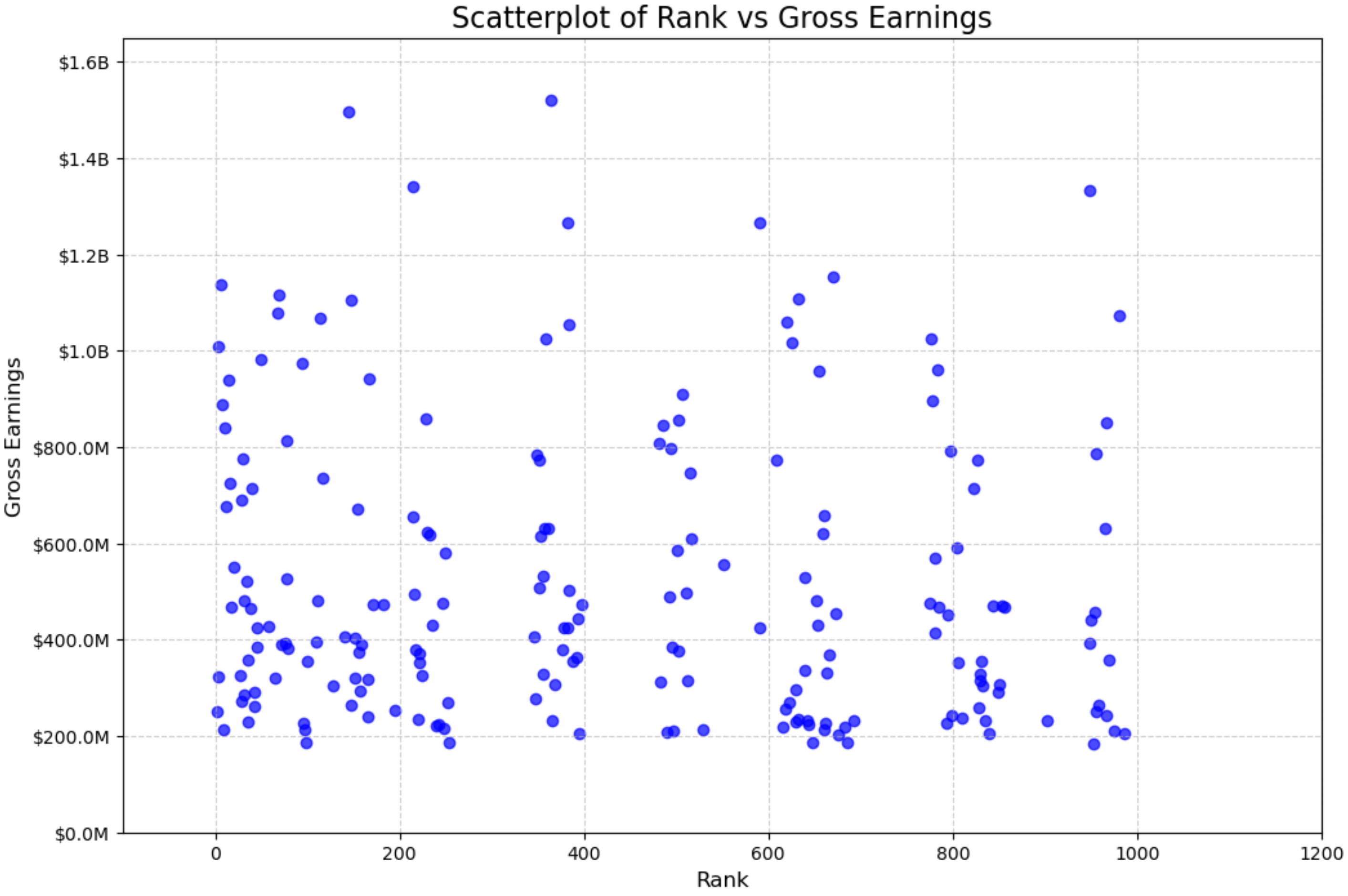
4. **Questions / Analysis**

4a. Does movie ranking / rating have anything to do with how much a movie makes?

This is the first question we wanted to answer. We utilized the *merged_df.csv* file from Table 1 shown above to construct a scatter plot graph as well as correlation metrics to get a better understanding of this question and answer. For this situation we only had to utilize the rank and gross earnings columns. The correlation coefficient between rank and gross earnings came out to -0.1196. This shows that a movie's quality rating and how much it sells in the box office has no correlation at all. This is not what we expected. By nature, we thought that on average the higher quality movies would be selling more in the box office but based on the data in this instance that appears to not be true.

To visualize this, we created a scatterplot that is found on the next page under Scatterplot Question 1. This scatterplot helps reinforce our finders of the non-existent correlation. It also gave us some more unique insights. For example, very few movies that are both highly rated and highly sold break into one and a half billion dollars sold. With this scatterplot we were also able to see that the majority of movies regardless of rank 0 through 1000 appear to be between 200 and 600 million dollars sold. With this our first question was not only answered but we were also given insight a little bit deeper into our data. The correlation as well as the visual was done within *02_Question1_answer.ipynb* inside of the python notebooks folder.

Visual 1: Scatterplot of Rank vs Gross Earnings

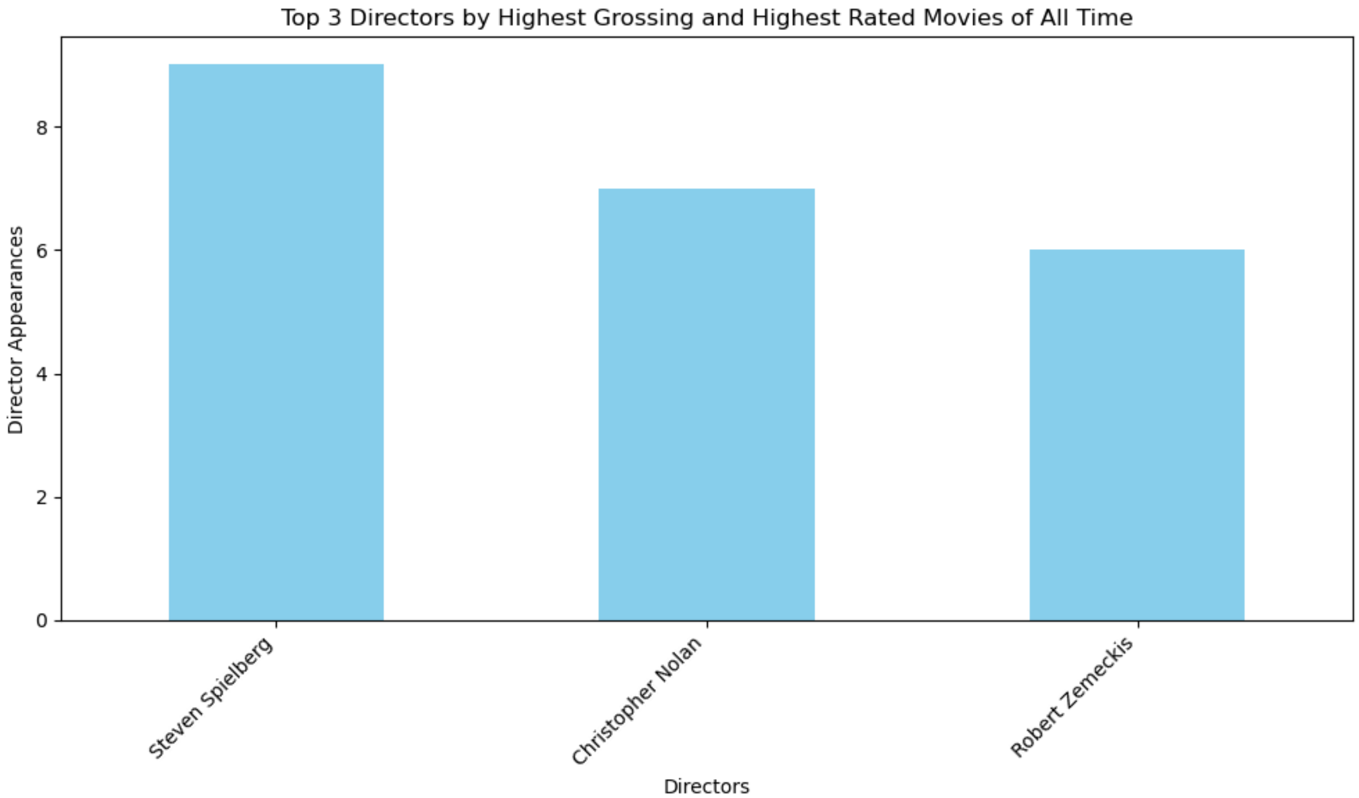


4b. Does a director appear more than once who has highly rated movies and highly sold movies?

This is the second question we wanted to answer. We were intrigued to know if certain directors produce high quality films as well as highly sold films more often than others, and if they exist what their names are. To find the answer to this question we utilized the merged data frame from Table 1 which was inside of *merged_df.csv*. Like our first question, the merged data frame from that file contained all the data we needed to find this answer. In this case we needed to find the counts for each instance in the director(s) column. We also wanted to find the average number of appearances in both lists for the directors to gain some insight into whether directors create box office hits and critic hits often. The average came out to 1.54 meaning on average a director appears 1.54 times in both lists. The reality of this scenario is that most directors likely appear once, and this average is inflated by directors like Christopher Nolan or Steven Spielberg who appear much more often. This is what we suspected since major films are often directed by big name directors who have a history of producing great films.

We found that Steven Spielberg, Christopher Nolan, and Robert Zemeckis appeared the most out of any other director in both lists. To visualize the top three directors who appear the most to get a visual representation our question we placed the data into a bar graph in descending order. We displayed the top three directors with the most appearances in both lists. This visualization can be found below in Visual 2. The average we found along with the visualization can be found in *03_Question2_answers.ipynb* within the python notebooks folder.

Visual 2: Top 3 Directors by Highest Grossing and Highest Rated Movies of All Time



4c. What was the average grossing amount by audience rating? What audience rating makes the most money?

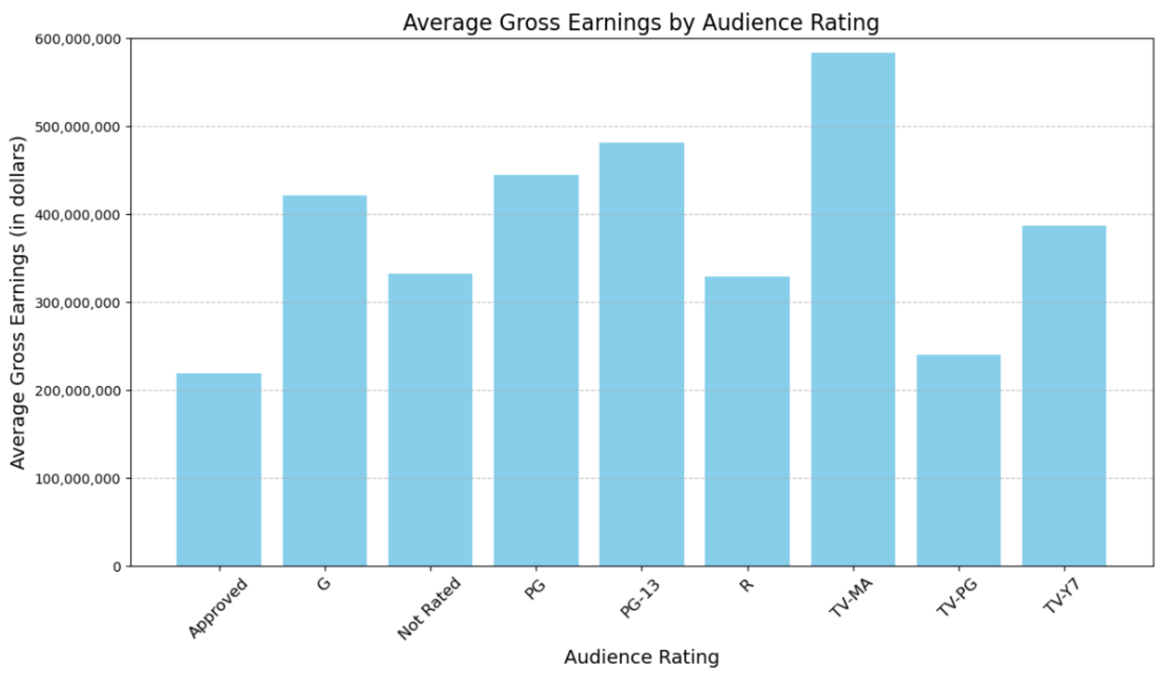
The last question we wanted to answer dealt with the average grossing amount by the audience rating. We were interested in finding this because we wanted to see if there is a correlation between the audience rating and how much money a movie makes. To answer this question, we leveraged the gross_earnings and audience_rating columns from the *IMDbHG_raw.csv file*. No merging was required because the necessary data was available in a clean and structured format so we did not have to use the created *merged_df.csv* like we did in question one and two.

Our goal was to calculate the average gross earnings for each unique audience rating category. We grouped the data by the audience_rating column and calculated the mean gross earnings for each group.

The findings revealed that the audience rating TV-MA had the highest average gross earnings, followed by PG-13 and PG. On the other hand, ratings such as Approved and TV-PG had the lowest average gross earnings. This suggests that movies with a more mature rating tend to perform better financially, likely due to a bigger audience appeal and more flexibility on the content directors can make in a movie.

To visualize the results, we plotted the average gross earnings for each audience rating using a bar graph. This graph highlights the differences in average earnings across the audience ratings and allows for easy interpretation of the data. This visualization can be found below in Visual 3. The averages we found can be found in *04_Question3_answers.ipynb* within the python notebooks folder.

Visual 3: Average Gross Earnings by Audience Rating



5. Conclusion

In this project, we analyzed the relationship between movie rankings, box office gross earnings, and audience ratings using IMDb’s Top 1000 Highest Rated Movies and Top 1000 Highest Grossing Movies lists. In summary, we investigated three key questions and found the following results.

1. *Does movie ranking / rating have anything to do with how much a movie makes?*

   The correlation coefficient between a movie’s ranking and its gross earnings was -0.1196, showing no significant correlation. This contradicts our initial expectations of thinking that higher-rated movies would earn more money. The scatterplot further shows that most movies gross between 200 and 600 million dollars, with few surpassing 1.5 billion dollars.

2. *Does a director appear more than once who has highly rated movies and highly sold movies?*

   On average, directors appeared 1.54 times in both lists, with the average being inflated by directors like Christopher Nolan, Steven Spielberg, and Robert Zemeckis. This supports the idea that successful directors tend to create both highly rated and high-earning films. The bar graph provided visualizes the top three directors with the most appearances.

3. *What was the average grossing amount by audience rating? What audience rating makes the most money?*

   Movies with a TV-MA rating had the highest average gross earnings, followed by PG-13 and PG, meanwhile Approved and TV-PG had the lowest. This suggests that mature audience ratings tend to perform better in terms of earnings. A bar graph provided illustrates the differences.

This project has several limitations, such as, the data was restricted to the Top 1000 movies from IMDb, which may not fully represent the entire movie industry. Additionally, the gross earnings were not adjusted for inflation which could affect the comparisons with the newer and older movies.

Future work on the project could involve expanding the dataset to show movies beyond IMDb’s Top 1000 lists. It would also be interesting to analyze the question of whether certain genres or release years can influence the amount of money a movie makes.