MICROSOFT MOVIE STUDIO PROJECT

Jupyter Notebook: [Python Notebook]
GitHub Repository: [GitHub Repo]

BUSINESS UNDERSTANDING

- The role assumed here is that of a Data Scientist working for the American multinational technology corporation, Microsoft. The task is to explore and analyze different movie datasets to understand what types of films are currently doing the best at the box office. These findings should thereafter be translated into actionable insights that the head of Microsoft's new movie studio can use to help decide what type of films to create.

- According to this Holywood Reporter article, we see that the metric used to determine the performance of a movie at the box office may vary. Therefore, it is important to clarify that within this analysis, the metric that shall be used to determine the success of a movie will be based on the return on investment (RoI) of the highest-grossing films. RoI is an important performance measure used by businesses to evaluate the profitability of an investment or compare the efficiency of several different investments.

- To further understand the types of movies that are currently performing the best at the box office, this analysis will look into the impact the following features have on the ROI:
    - Runtime

- Production Budget
- Gross Revenue
- Release Date
- Genre
- Directors
- Writers
- Rating

PROBLEM STATEMENT

- Identify the films that are currently performing the best at the box office

DATA UNDERSTANDING

- This analysis uses datasets obtained from two renowned movie websites:

    - The Numbers
    - IMDB

- From the first dataset which is in .csv format, we have the names, release dates, and monetary information of films that have been released and those are are yet to be released. With RoI being the target variable, the monetary data (production budget and worldwide gross) columns are the main reason this dataset was selected. Furthermore, the movie title and release date columns make this dataset an all-round good starting point for the data preparation stage.

- From the second dataset which is in .db format, we have a database with 8 tables containing different types of non-monetary information about films such as their directors, writers, and genres, the ratings, and more. This information will be used to understand the characteristics of the films that are currently performing the best at the box office.

DATA PREPARATION

SELECTING DATA

Because we are analyzing using the monetary data as a measure of success, all the columns from the first dataset are going to be used except the ‘id’ and ‘domestic_gross’ columns. These are:

- release_date
- movie
- production_budget
- worldwide_gross

To obtain the features that we shall be analyzing RoI against, the second dataset would be used. This is because the second dataset contains the metadata about the films. We will select tables:

- movie_basics
- movie_ratings
- directors
- writers

DATA CLEANING

The process used to analyze the data in this project is as follows:

- Narrow the first dataset down to the movies that have been released within the past ten years using the 'release_date' column. This is because we are looking for films that are currently performing the best.

- Sort the DataFrame by the worldwide gross column in descending order and slice the first 200 rows. We shall be analyzing the RoI based on the highest-grossing films. Furthermore, narrowing the DataFrame to the top 200 records will ensure that the

films selected have a gross that is around $300 million or more. This is a good worldwide gross for a film.

- Create a new column in the DataFrame to store the RoI value that will be calculated by dividing the worldwide gross by the production budget and multiplying the result by 100 to represent the percentage.

- Add more one-to-one features that will be used to classify the data. These features that we are looking to analyze are:
    - Runtime
    - Genres
    - Rating

- These features can be obtained by joining the movie_basics table by using the 'primary_title' column and the movie_ratings table using the newly joined 'movie_id' column.

- Clean the data by dropping the rows that have been joined to the wrong movie from the IMDb datasets and those containing missing values.

    - By evaluating the year values from the 'release_date' column and the 'start_year' column that was joined from the movie_basics table, we can determine the records that have discrepancies and drop them.

    - Drop the films that contain duplicate movie titles. Though multiple films can share a title, it is not a good idea to include them in the analysis as there is no way of verifying whether or not the information is correct.

    - The last step would be to drop the columns that do not contain variables that we are interested in. These are: 'id', 'domestic_gross', 'primary_title', 'start_year', 'release_year', and 'numvotes'.

After performing this analysis we ended up with a DataFrame containing the following columns:

- 'movie_id' - The unique identifier for each movie from the IMDB dataset.
- 'release_date' - The release date of the film.
- 'movie' - The title of the film.

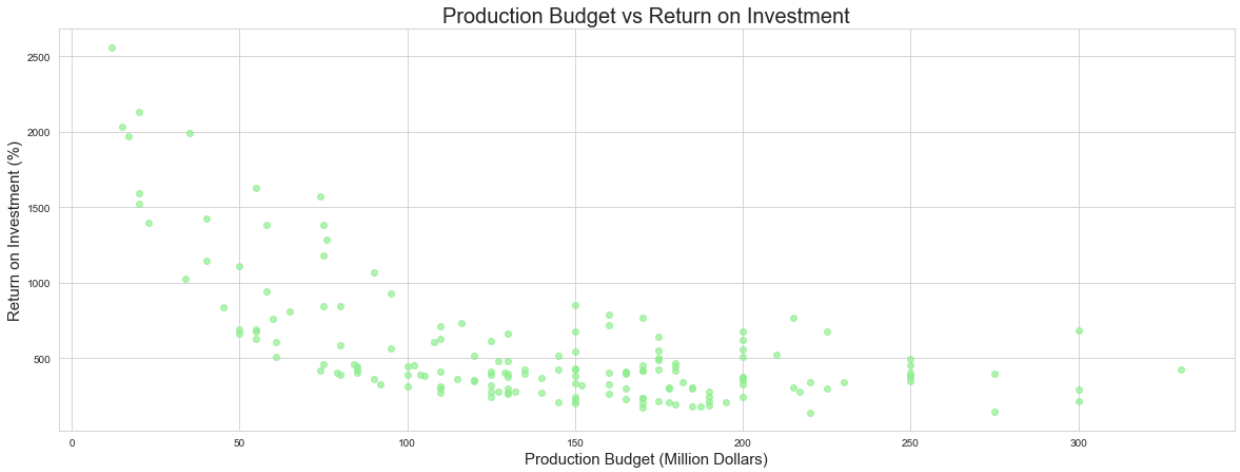
- 'production_budget' - The production budget of the film.
- 'worldwide_gross' - The worldwide gross of the film.
- 'RoI' - The return on investment of the film.
- 'runtime_minutes' - The length of the movie in minutes.
- 'genres' - The genres of the film.
- 'averagerating' - The average rating of the film.

The approach chosen to prepare the data is appropriate as it enables the analysis of the target variable (RoI) against the non-monetary features making it easier to get insights. The other features such as 'directors', and 'writers' will be implemented in the analysis phase in separate DataFrames since they contain one-to-many relationships. That would have made the analysis more difficult.

DATA ANALYSIS

- For the analysis phase we chose our primary target variable, RoI, and analyzed the rest of the features against it:
    - Runtime - The goal was to identify whether the length of a film has any impact on the return on investment. After plotting the scatter plot of the runtime against the RoI, we see that runtime has no direct impact on the return on investment of a film. This coupled with the Pearson correlation coefficient of -0.1 which is closer to 0, indicates that the runtime does not have any impact on the return on investment of a film.

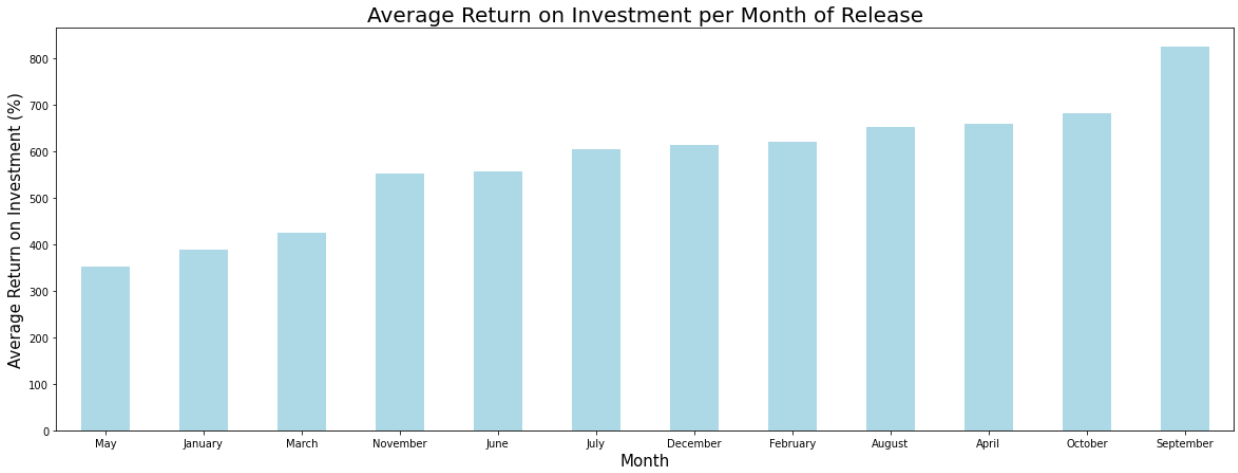
- Production Budget - This analysis aimed to identify whether spending more money on the production of a film results in a higher return on investment. From the scatter plot used, we see that there is a negative correlation between the production budget and the RoI. However, this relationship is not linear. From 0 to 100 million dollars, the correlation is negative. However, from 100 to 300 million dollars, there is no distinct correlation between the return on investment and the production budget. Looking at the Pearson correlation coefficient, we see that the correlation coefficient (-0.6) suggests a moderately negative relationship between the return on investment and the production budget which confirms what we see on the scatter plot. Therefore, the more money spent on the production of a film, the return on investment generally less.

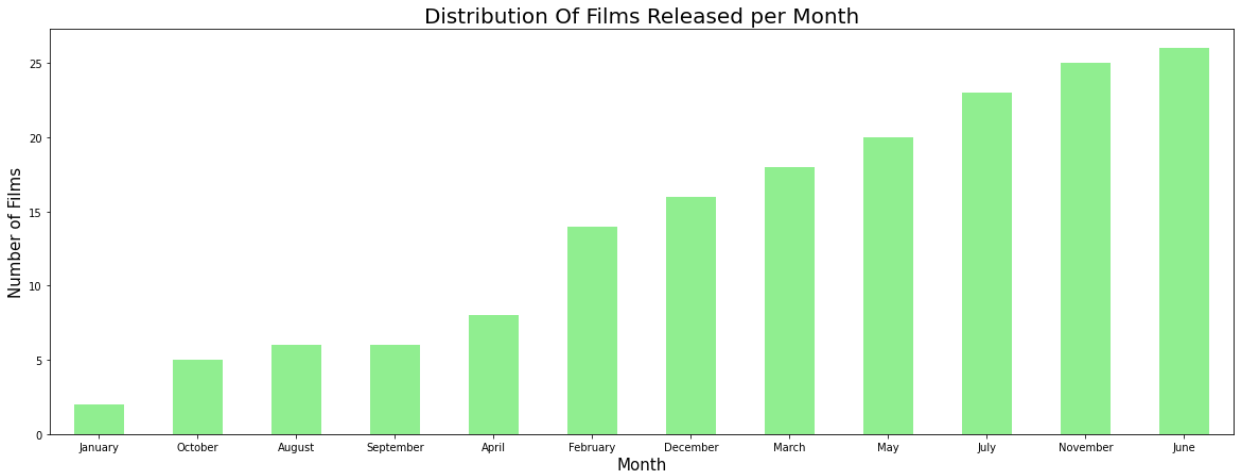


- Worldwide Gross - This analysis aimed to identify whether there is a relationship between films that gross the highest and the return on investment. From the scatter plot used, there is no distinct relationship between the worldwide gross and the RoI. Looking at the Pearson correlation coefficient, we see that the correlation coefficient (0.15) is closer to 0, therefore, suggesting that the worldwide gross of a film has no impact on its RoI.

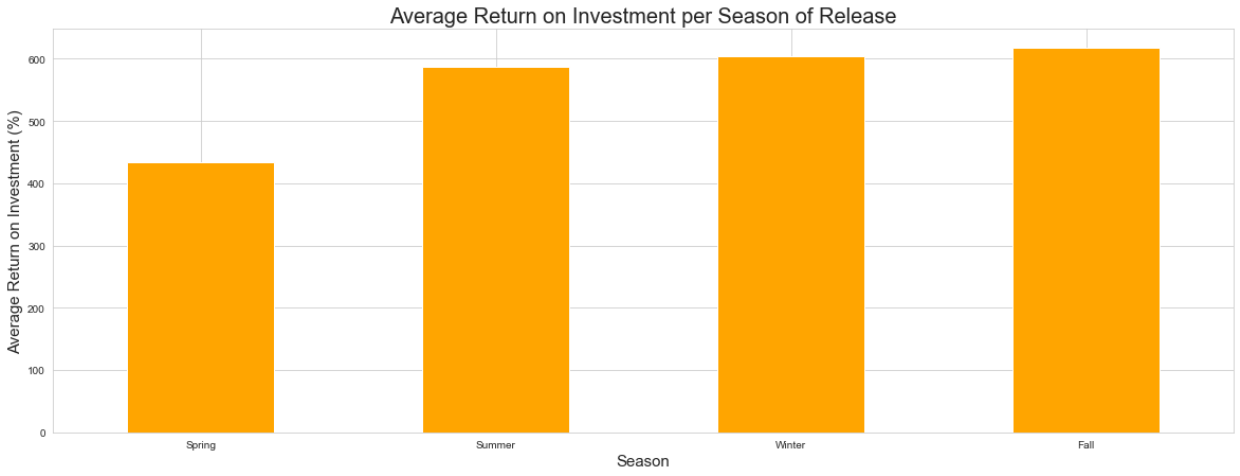
- Release Date - This analysis aimed to identify the impact that time of release has on the return on investment of a film. This analysis was split into parts: Month and Season.

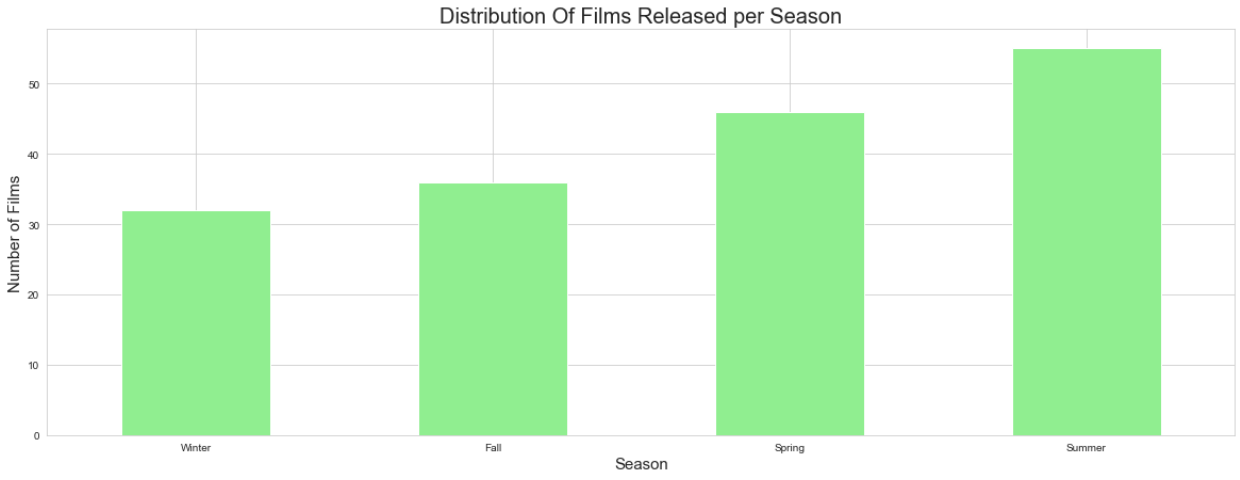
    - The analysis by month showed that films released in September averagely yield a higher RoI, however, after looking at the film distribution, we see that the distribution was not even and may perhaps be the reason why September, which had fewer films than most, averaged a higher RoI.





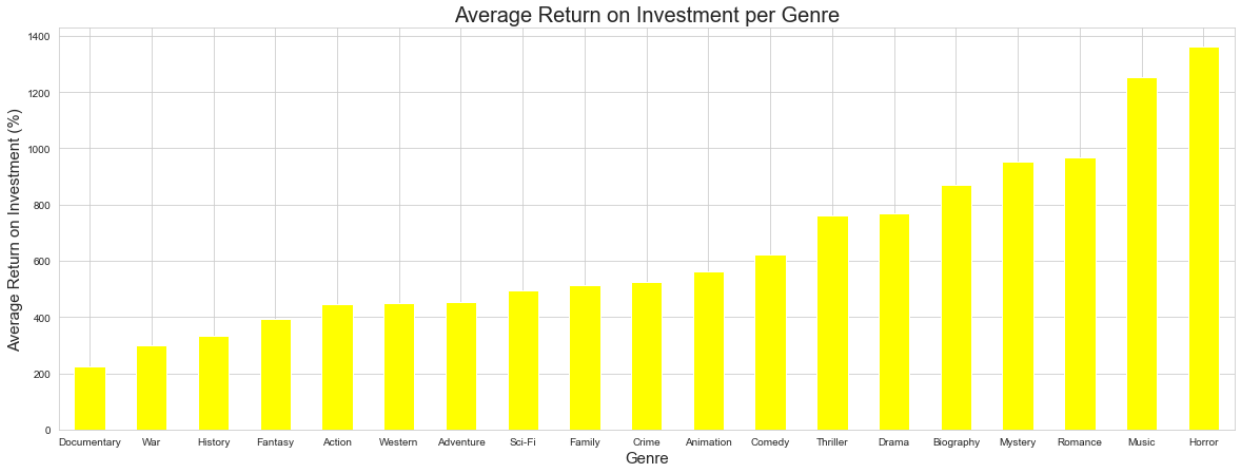
- The analysis per season shows that the highest return on investment is achieved in the Fall season, however, since the Fall season had the lowest number of films, we see that it isn't an accurate season to consider. When we look at the Summer season though, we see that it had a comparable RoI to the Fall season despite having the most films released. This, therefore, suggests that films released in the Summer are the best performing.

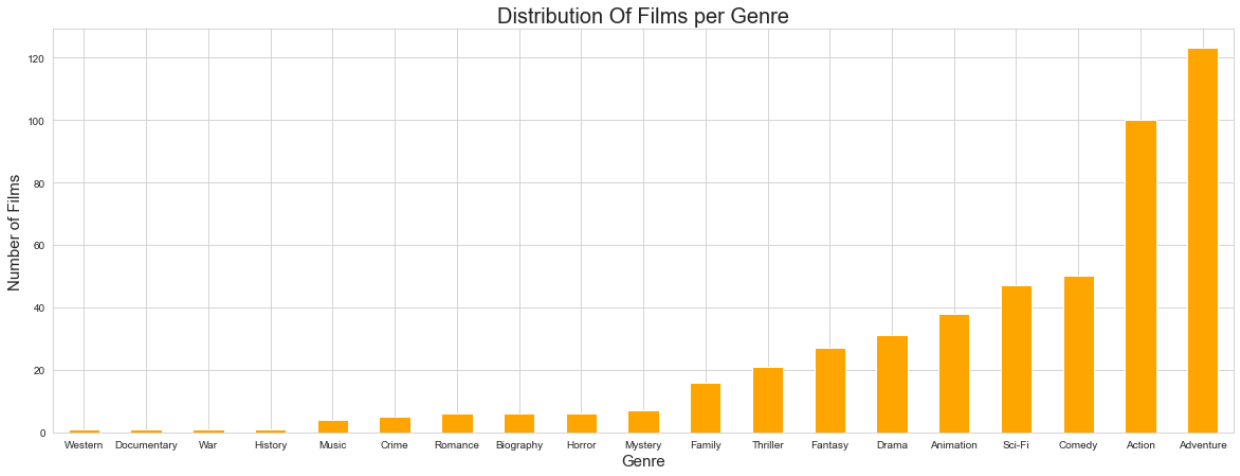




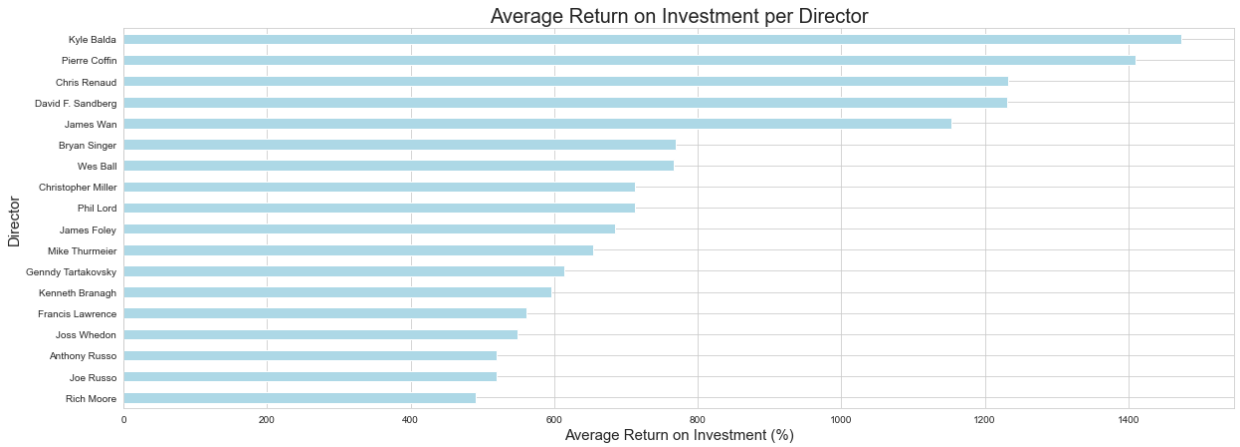
- Genres - The aim of this analysis was to identify the genre that produces the highest RoI. From the barplot used, we see that the genre that produces the highest RoI is Horror followed by Music. However, when we look at the

number of films per genre, we can see that the distribution of the genre of films in the dataset is not even. Therefore, we need to take into account that the reason that the 'Horror' and 'Music' genre have a higher average return on investment may be because there are fewer films classified under them. It is however important to note that of the Top 200 Highest Grossing Films that were originally selected, the majority were Adventure and Action. This suggests that Adventure and Action films averagely make the most money. As a result, this makes it difficult to derive any definite conclusion as to what genre yields the highest average RoI.

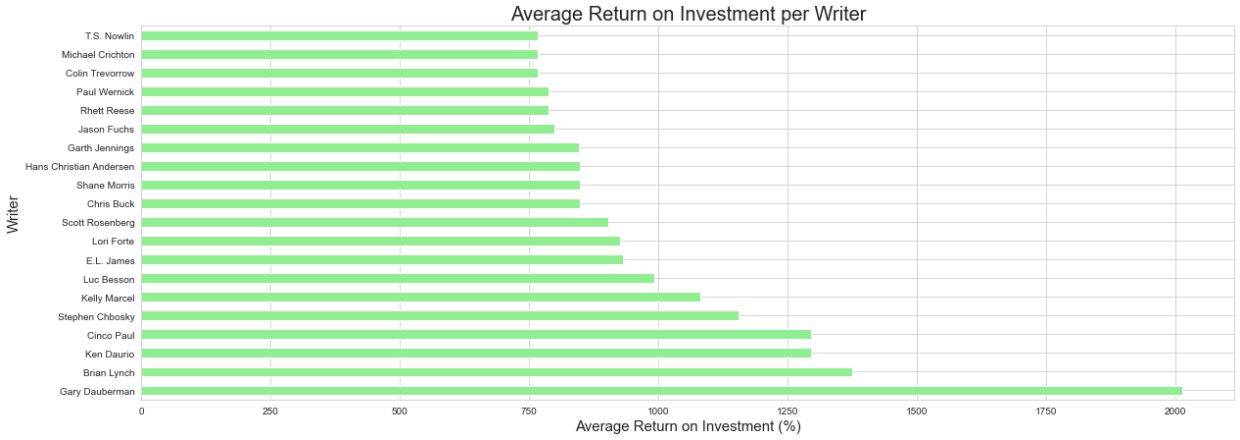




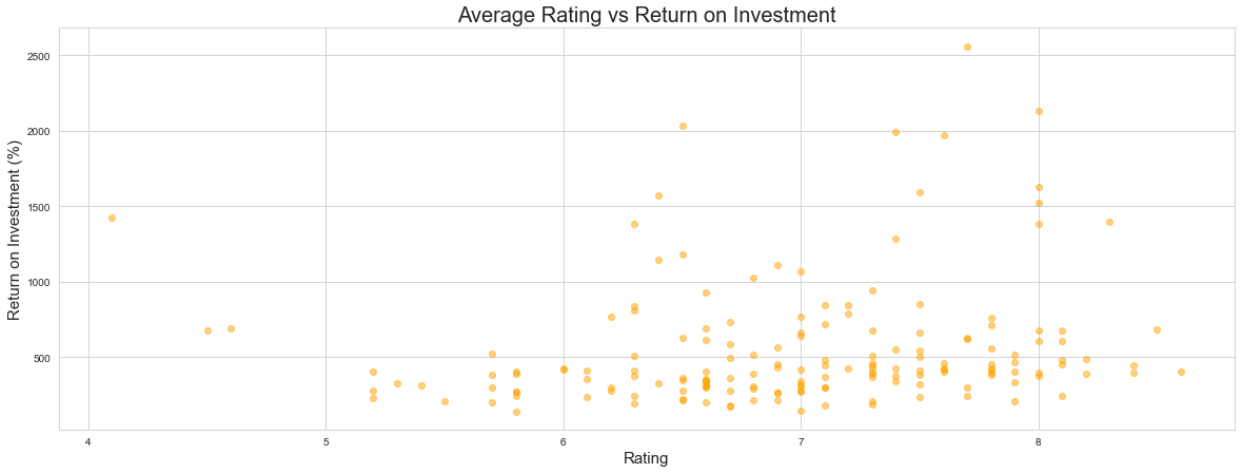
- Directors - This analysis aimed to identify the directors with the highest average RoI for the films that they have directed. We only focused on the directors who had multiple films within the prepared dataset. From the plot, we can see that films directed by Kyle Balda, Pierre Coffin, Chris Rennaud, David F. Sandberg, and James Wan produced the highest return on investment.



- Writers - The aim of this analysis was to identify the writers with the highest average RoI for the films that they have written. We only focused on the writers who have written multiple films within the prepared dataset. From our plot, we can see that films written by Gary Dauberman produced the highest return on investment.



- Rating - This final analysis aimed to identify the impact that return on investment has on the rating of a film. The scatter plot makes it difficult to see any distinct relationship between the rating and the RoI. However, looking at the Pearson correlation coefficient (0.15), we see that it is closer to 0, therefore, confirming that there is indeed no relationship between the rating of a film and the RoI



CONCLUSION

- This analysis leads to the following recommendations for the types of films that are the best performing at the box office:
    - The length of a film (Runtime) of a film has no impact on its box office performance.
    - The Production Budget of a film has a moderately negative correlation with its return on investment.
    - The Worldwide Gross of a film has no impact on its return on investment.
    - Movies released in the Summer are more likely to yield a higher return on investment. Though September (Fall) had the highest average return on investment, the number of films released was significantly less than in other months. Therefore, it would not have been an accurate measure to determine the optimal time to release a movie.
    - 'Horror' and 'Music' genres are more likely to have a higher return on investment. It is however important to note that 'Action' and 'Adventure' films are the top most grossing.
    - Films directed by Kyle Balda, Pierre Coffin, Chris Rennaud, David F. Sandberg, and James Wan produce the highest return on investment.
    - Films written by Gary Dauberman produce the highest return on investment.
    - The success of a film is not determined by the film’s rating

- This analysis may not fully solve the business problem because there are other factors that affect the performance of a film such as data that was not available during this analysis (for example, the amount of money that is spent on the film's marketing), as well as other unpredictable circumstances going on in the world (pandemics, economic downturns, war, etc).
- In order to further improve this analysis, we would need to look at more financial data such as the amount of money that is spent on marketing, the social impact of the films as well as the main cast of the films.