MICROSOFT MOVIE PROJECT

Author: Teofilo Acholla Ligawa Gafna

1. BUSINESS UNDERSTANDING

1.1 UNDERSTANDING THE PROBLEM

Microsoft desires to venture into the movie production industry and they have set up a studio. The role that this study has assumed is that of a data scientist who is tasked with shedding light on the movie industry to determine which kind of movies perform best at the box office. The ultimate goal is to advise Microsoft on which type of movies to invest in if at all they plan to become significant players in the movie industry.

1.2 PROBLEM STATEMENT

The ultimate goal is to advise Microsoft on which type of movies to invest in if at all they plan to become significant players in the movie industry by answering the following questions:

A. What can you expect to spend on a good movie?
B. Which genre dominates among the top fifty rated movies?
C. Which genres are economical in terms of the return on investment?
D. Does the movie's length affect the film's cost?
E. When is it Lucrative to release the movie?

2. DATA UNDERSTANDING

This is where we get the data about the data. The shape in terms of the number of rows and columns, the descriptive statistics, the first five entries, and the last five entries just to get an idea of how the data should look like. We shall explore the various relationships among the variables involved in the study.

2.1 Data Collection

The data was collected from two sources. The first is the IMDb database that the study queried using SQLite in order to obtain movie_basics and movie_ratings. The movie_basics table contained data on movie id, primary title, original title, genre, release year, and runtime minutes while the movie_ratings table contained data on movie id, average ratings, and the number of votes. The study then joined the two tables using the primary key of the movie id. From The Numbers, the study had access to data on movie titles, release dates, production budgets, domestic gross revenue, and worldwide gross revenue.

2.2 Data Preparation

2.2 .1 Data Cleaning and Feature Engineering

Firstly, the study queried data from the IMDb database to obtain tables from it. The study further selected the movie_basics and movie_ratings tables, end explored the tables checking for inconsistencies such as missing values and duplicates. From movie_basics, missing values were found on the original titles, genres, and runtime minutes columns. Rows were subset on the genres and runtime columns then missing values from those columns were dropped. The original titles column was dropped. Duplicates were not found. From the movie_ratings table, inconsistencies were checked then the two tables were merged on common rows to form a new data frame with columns from both movie_basics and movie_ratings. The new data frame was feature engineered in that some column names were changed to facilitate further data preparation.

Secondly, the study loaded data from The Numbers, made a data frame out of it, and explored it for inconsistencies. Missing values and duplicates were not found, and some column names were changed to facilitate further preparation and data analysis.

Thirdly, the merger of movie_basics and movie_ratings was then joined to the data frame from The Numbers on common rows to obtain the third data frame. The study considered the possibility of having outliers but concluded that if they were removed then the analysis would be inaccurate.

The third data frame was then feature engineered to create a fourth data frame which consists populated by details of the top 30 percent of movies based on the average rating after which a subset of the third data frame was created in order to obtain data for the top 50 movies in the top 30 percent of movies based on the average rating.

The final table had 786 rows and 11 columns and was as follows:

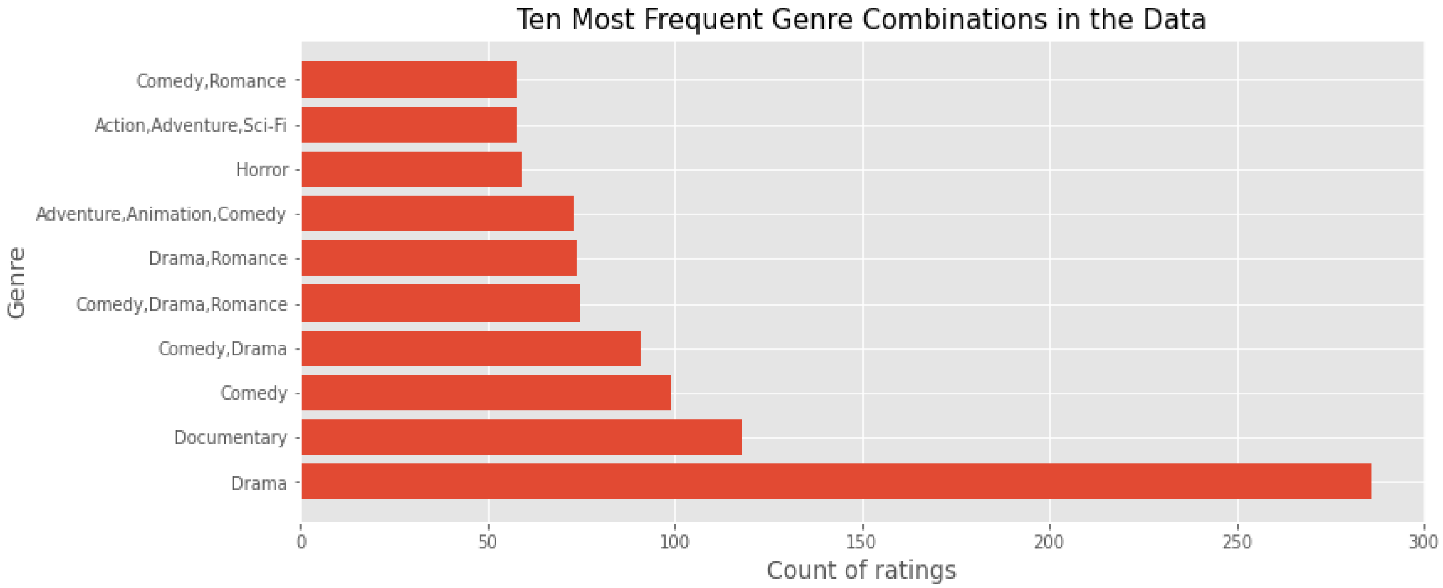
Variable	Description
title	Movie name
runtime mins	Movie length in minutes
genres	Movie genre of movie
average_rating	Mean rating of movie
release_date	Date movie was released
production_budget_usd	Production budget of the movie in USD
domestic_gross_usd	Domestic income from the movie in the US in USD
worldwide_gross_usd	Foreign income from the movie in USD
return_on_investment	Total revenue - production budget in USD
total_revenue	Foreign income + domestic income in USD
month	Month movie was released

2.3 Data Analysis Findings

The findings from the analysis were as follows:

Dominating Genre Overally

In the third data frame, the most frequent genre combinations were as follows:



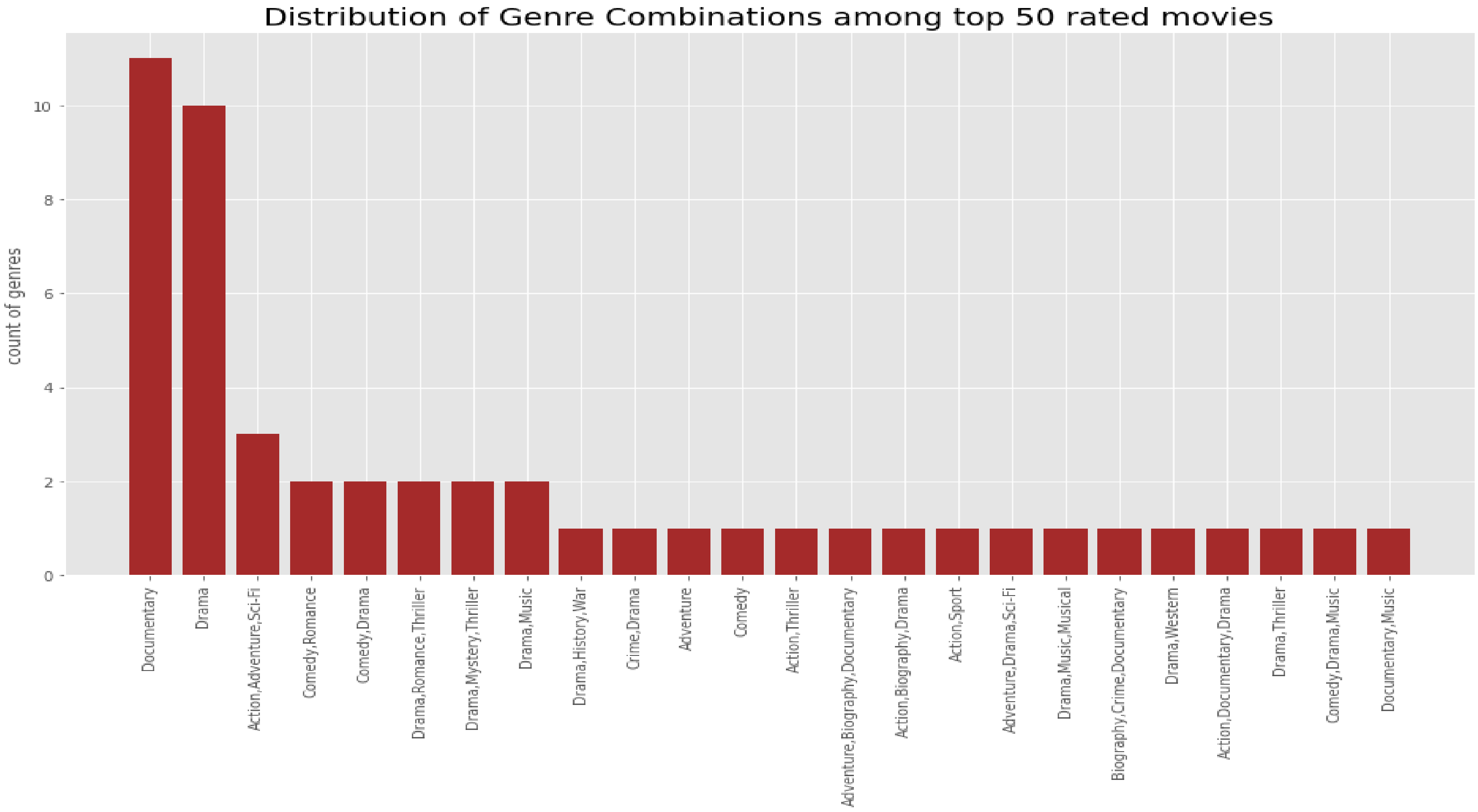
Drama was most frequent followed by Documentary and comedy.

What should Microsoft expect to spend on a good movie?

The study discovered that the mean production budget of the top 30 percent of movies based on the rating was estimated to be around 41 million US dollars and this was the amount of money that was suggested to Microsoft. This was within the interquartile range of 6.5 million US dollars to 49 million US dollars.

Which Genres dominate the top 50 rated movies?

The study discovered that Documentary and Drama dominate the top 50 charts and are followed by Action/Adventure/Sci-Fi combination as shown:



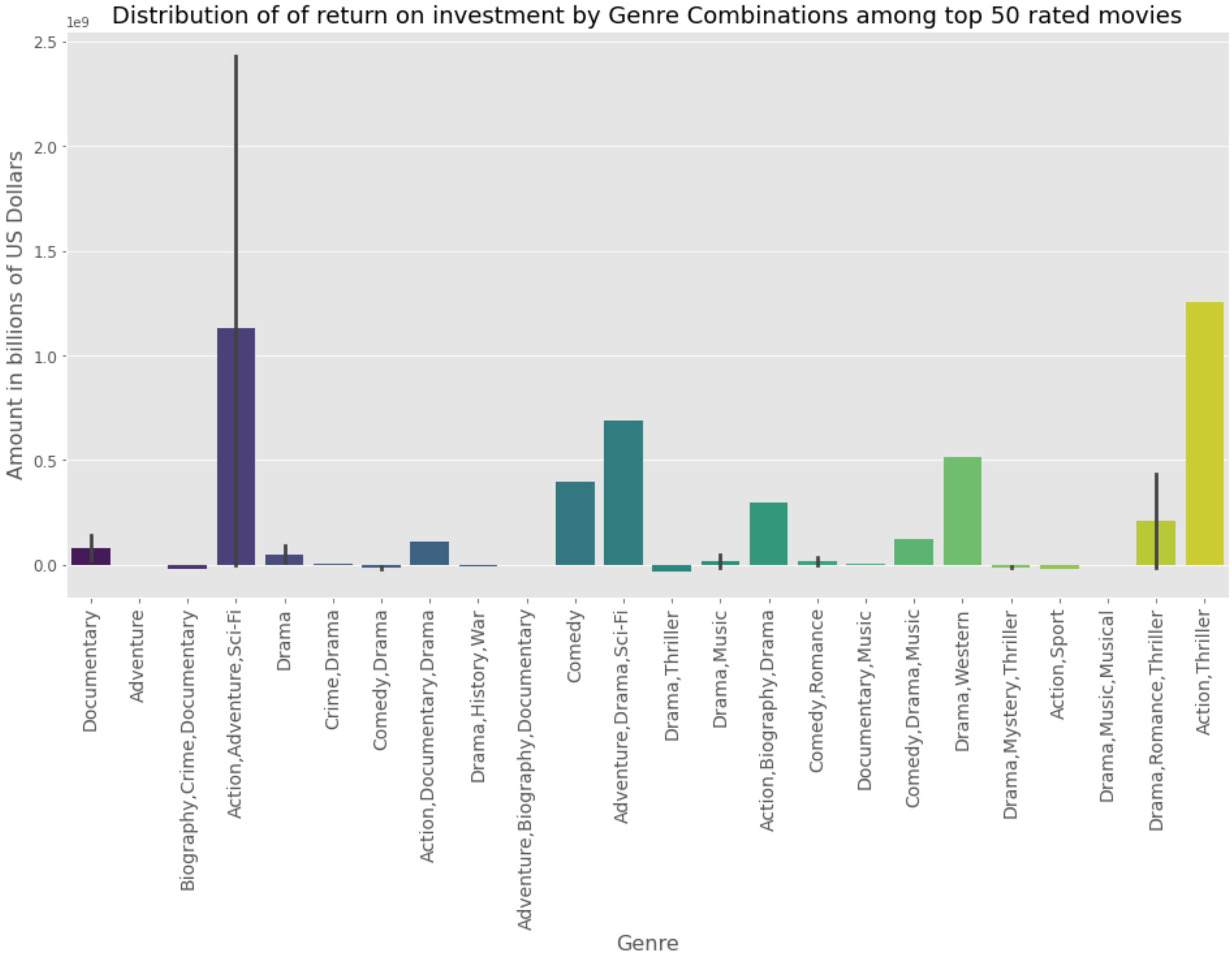
Which Genres are economical in terms of return on investment?

Among the top 50 charts, the genres that were found to have a meaningful return on investment include:

- Action/Adventure/Sci-Fi
- Action/Thriller
- Adventure/Drama/Sci-Fi

- Drama Western
- Comedy

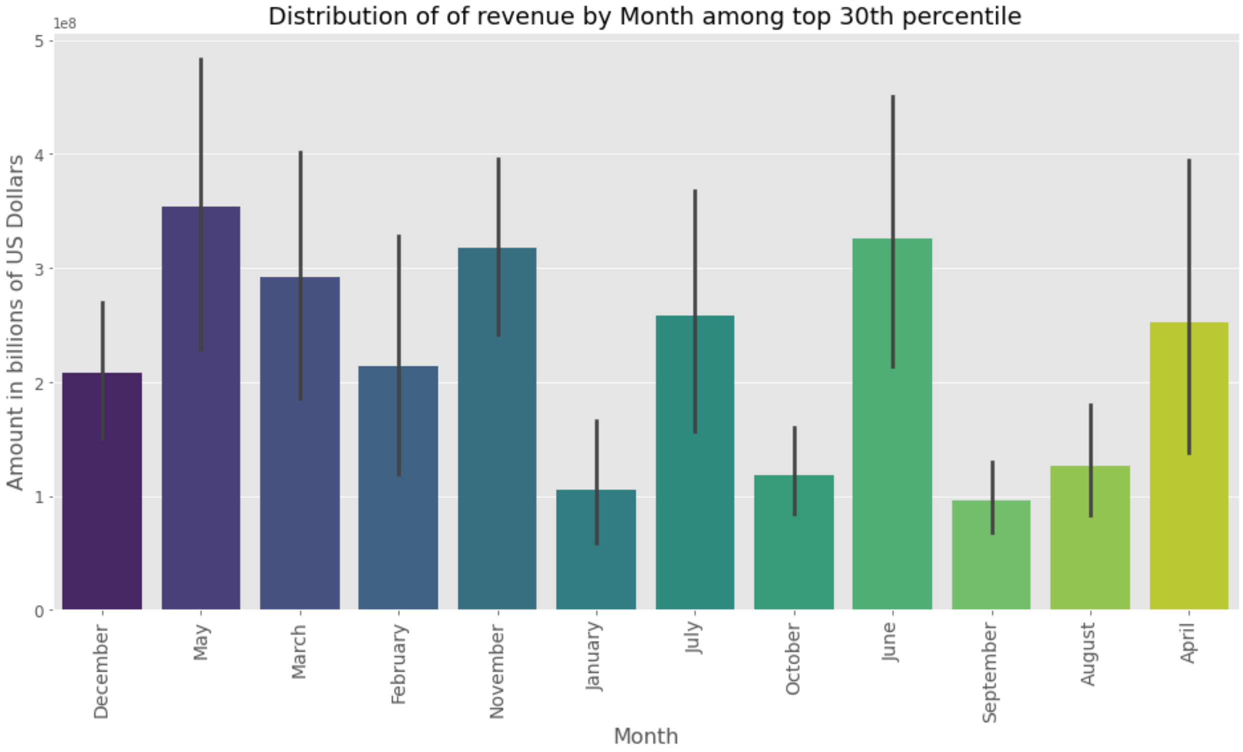
Action/Adventure/Sci-Fi lead the charts in terms of return on investment while also being the most expensive to produce.
The graph below sheds more light.



Does Length of Movie Affect Production Cost?

A correlation test was done to test whether or not there exists a relationship between the length of the movie and the production cost. The study discovered that the correlation coefficient was 0.013 and this was considered as not significant enough to have a linear relationship.

Which month Should the movie be released?



From the graphic above, May seems to be the most lucrative time to release a movie. The study discovered that there are two periods in which a studio can release movies to get a good profit. These periods include:

- The months between April and July.
- The months between November and March.

Conclusion

Microsoft's ticket to worldwide film industry success lies in producing an Action/Adventure/Sci-Fi movie on an estimated budget of around 1.1 billion US dollars released in May if they are to get a significant value for their studio investment, or produce either of or a combination of Documentary/Drama/Action and release it between April and July or between

November and March if they want to create a good reputation for themselves in the film industry

The study is convinced that Microsoft will have a successful entry into the filmmaking sector if the suggestions the study has offered are implemented. The data shows that every characteristic the study has mentioned is associated with a big global box office take, which is exactly what Microsoft will want for their initial films and beyond.

Recommendation

From the conclusion above, I would recommend that Microsoft take the following approaches:

- In the short term, the study recommends that Microsoft should focus on making a name for themselves in the short run and thus produce a Documentary/Drama/Action movie to be released between April and July or between November and March on a budget of around 41 million US dollars that will increase the probability of obtaining ratings of 8.2 and above. This will most likely cause them to gain popularity.
- In the long term, the study recommends that Microsoft should focus on getting value for their money and produce an Action/Adventure/Sci-Fi movie that is produced on a budget of around 49 million US dollars. They should do this along with producing Drama/Documentary/Action in order to rise high in the charts of movie studios.