DECEMBER 2022

MICROSOFT MOVIE STUDIO INITIATIVE

Movie Industry Exploratory Data Analysis.

Summary.

Microsoft wants to enter the multi-billion dollar movie-making industry, To goal was the explore what type of movies are doing best at box office.

Findings:

- **Sci-Fi** genre has the highest potential of the box office success. On the other hand, its distribution of total gross also suggests that the movies' success vary and there are risks involved. The other genres with high success potential are **Animation** and **Adventure**.
- The movies that have a **runtime between 125 minutes and 180 minutes** are highly likely to succeed.
- There is **no significant correlation** between **the average IMDB rating** and **the total gross**. However, the results showed that the movies obtained a rating 8 and above had higher mean of total gross.

Outline

BUSINESS PROBLEM

DATA

METHODS

RESULTS

CONCLUSIONS

LIMITATIONS

NEXT STEPS

Business Problem.

Microsoft sees an opportunity to branch out to the multi-billion dollar movie-making industry. They decided to open a Microsoft Movie Studio but the main challenge is that it's a new industry for Microsoft.

I defined 3 main questions to answer:

- Which genres worth investing in?
- Do the longer movies have higher box office success?
- Is there a correlation between the IMDB ratings and the worldwide total gross?

Data.

Box office business can be measure in terms of the revenue. Therefore, the **main metric** we are exploring is **the total gross**.

The **variables** explored are **genre**, **movie length** in minutes and **average user ratings**.

The time period of the data is between 2010 and 2018.

The data is coming from 3 datasets from 2 separate websites.

IMDB Title Basics:

Movie Titles
Release Years
Genres
Runtimes in Minutes

IMDB Title Ratings:

Movie Titles
Average Ratings
Number of Votes

Box Office Mojo Movie Gross:

Movie Titles
Foreign Gross
Domestic Gross
Release Year
Studio

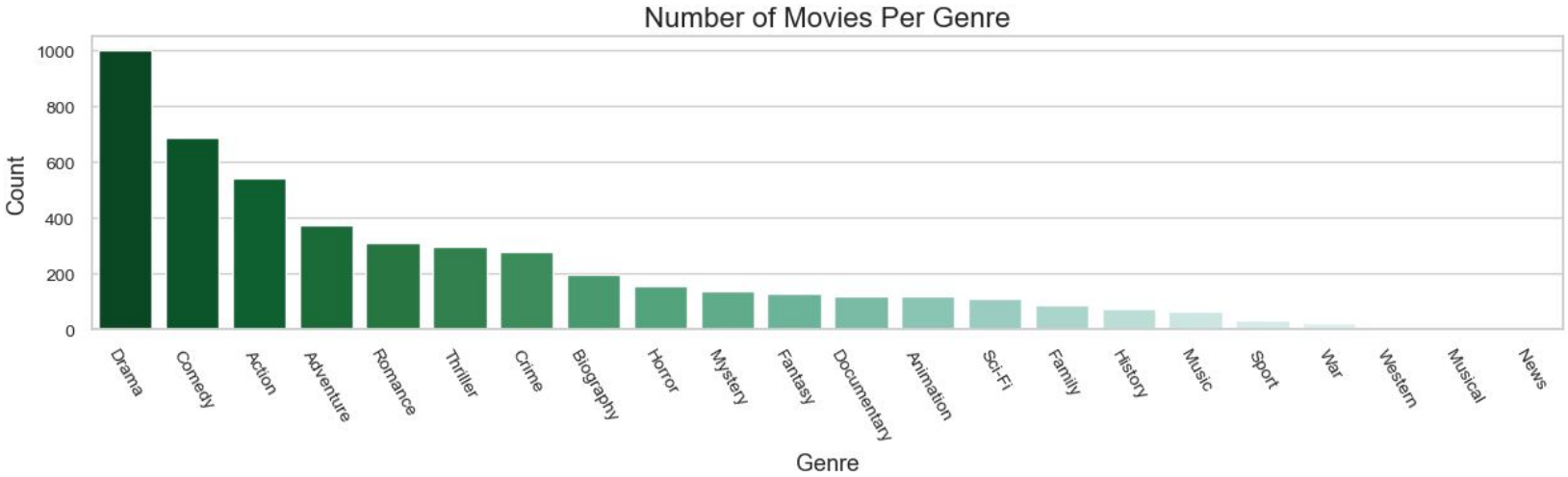
Methods.

- I removed columns and rows that were not part of the study.
- Either filled the null values or drop the rows that consist of them.
- Merged all the DataFrames together as my Master DataFrame and saved the cleaned version of this DataFrame in my repository.
- Used descriptive statistics and visualizations.

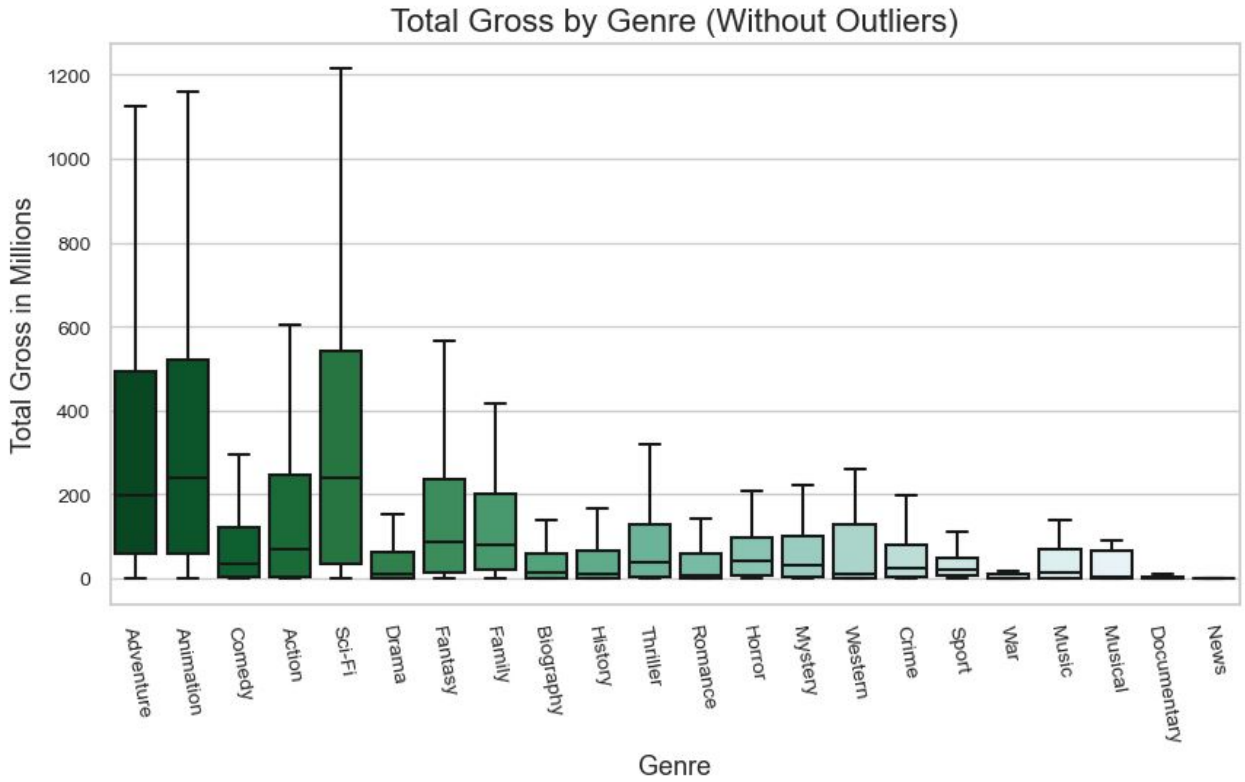
Results

Genres.

- The analysis includes 1942 movies with either single or multiple genres.
- To make it easier to analyse, I split the genres into their own rows and analysed each genre individually.
- By splitting them, the entries increased to 4773.
- The following graph shows the total number of movies per genre.



Genres.



The results of the analysis showed that the **Sci-Fi genre has the highest potential** of the box office success. On the other hand, its distribution of total gross also suggests that the movies' success vary and there are risks involved.

The other genres with high success potential are **Animation** and **Adventure**.

Movie Length.

- The analysis includes 1934 movies
- The runtimes are assigned to five categories:
    - Anything below 41 min is a short film
    - Anything between 41 min and 100 min is a short format feature-film
    - Anything between 101 min and 125 min is a medium format feature-film
    - Anything between 125 min and 180 min is a long format feature-film
    - Anything above 180 min is a super sized feature-film
- The table below shows the total number of movies per movie length.

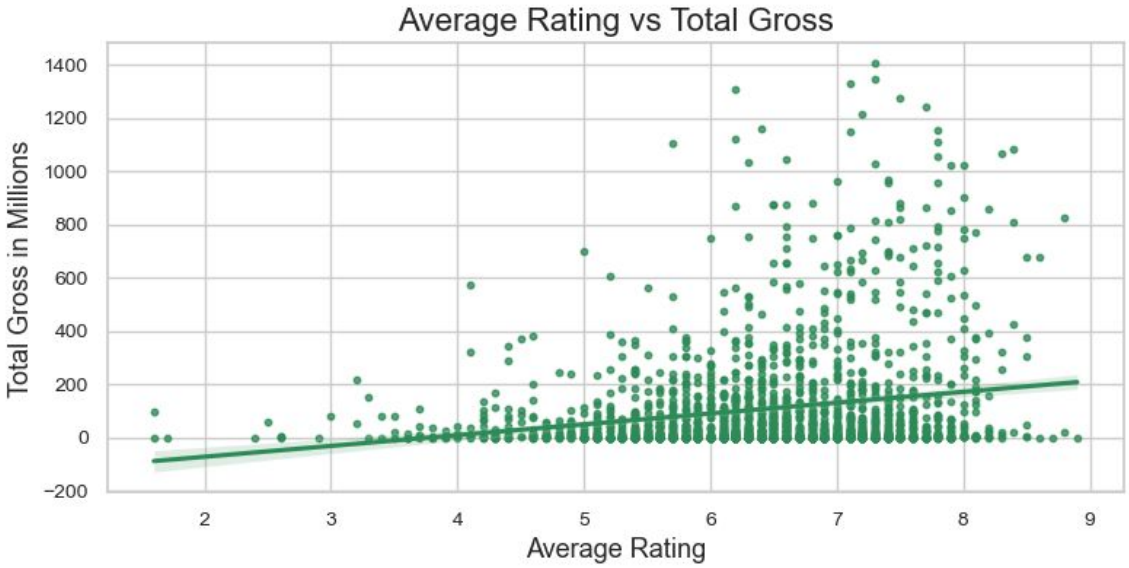
	movie_length
Medium Format Feature-Film	862
Short Format Feature-Film	661
Long Format Feature-Film	401
Short Film	7
Super-Sized Feature-Film	3

Movie Length.

The **long-format** movies that have a runtime of 125-180 minutes have higher box office success. However, when the runtime goes beyond 180 minutes, the total gross starts to decline.

Therefore, Microsoft should produce a movie in a long format but shouldn't exceed 180 minutes mark.

Average Rating.



There is a **weak positive correlation** (r=0.21) between the average rating and the total gross figure. The association between them is **negligible**.

It is interesting to see that the higher rating movies **do not necessarily** records higher gross figures.

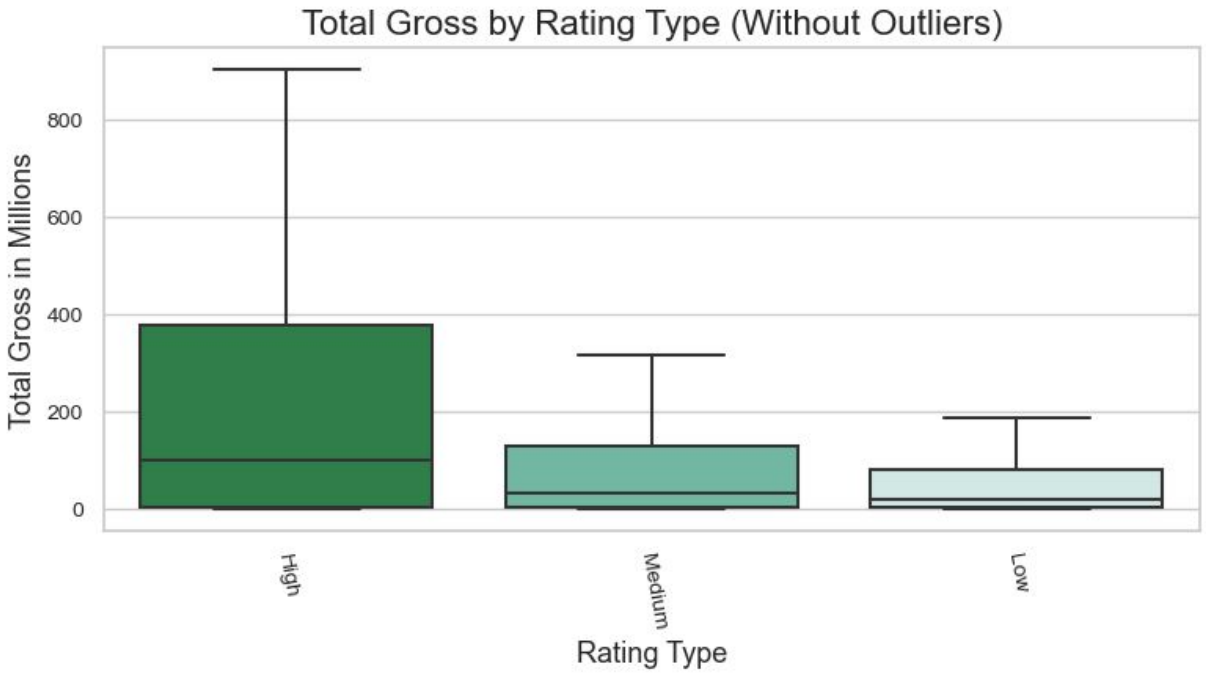
I also explored the correlation between the long format movies and the average rating. The results were very close (r=0.30) and the association was again negligible.

Average Rating.

- I also used average rating as a category variable.
- Included 1918 movies in the analysis
- Assigned one of the following categorical values to each movie based on their average rating:
    - Anything lower than or equal to 6 will be categorised as low
    - Anything greater than or equal to 8 will be categorised as high
    - Everything in between will be medium
- The table below shows the total number of movies per rating category.

	rating_type
Medium	1248
Low	591
High	79

Average Rating.



The results showed that the movies obtained a **rating 8 and above** had higher mean of total gross.

Conclusions.

My findings have shown a direction of what should be investigated even further. They provide a baseline of more in depth analysis work.

Recommendations:

- Microsoft should invest in **Sci-Fi**, **Animation**, or **Adventure** genres.
- The **long-format movies** that have a runtime of **125-180 minutes** have higher box office success. However, when the runtime goes beyond 180 minutes, the total gross starts to decline. Therefore, Microsoft should produce a movie in a long format but shouldn't exceed 180 minutes mark.
- There is **no significant correlation** between **average movie ratings** and **box office revenue**. Being said that, I observed the higher rating movies tend to perform well at the box office. The recommended action here is not to focus on the ratings purely and **use the high-rating movies as a benchmark purely.**

Limitations.

In this study, I worked with constraints.

The most important constraint was the delivery time of the analysis results. To derive some actionable insights, I had to limit the data I worked with and only selected three datasets. However, it became clear that more data was needed to explore the business problem.

Next Steps.

Moving forward, I highly recommend considering the followings:

- **The production cos**t and **the return on investment** should be part of the following studies. We know the total gross but we don't know the cost of making a high-gross movie and how profitable it could be.
- I explored the genres separately but it would provide great value if the **combination of these genres** are explored as well.
- There are **more variables** that wasn't a part of this study but could potentially give actionable insights such as **release dates, creative types,** and **production methods**.
- This study only consists of **movies released between 2010 and 2018**. It doesn't answer how the most recent movies performed and if the **global pandemic** affected the overall trends. There is value to explore these questions in a separate study before making decisions.

Thank you!

Hazal Aydin

h.aydinhazal@gmail.com

https://www.linkedin.com/in/hazalaydin/

https://github.com/hazal-aydin