Table of Contents

Film analysis for possible Microsoft venture

1. Business understanding

Link to Jupyter Notebook: http://localhost:8890/notebooks/student.ipynb

Link to GitHub Repository: https://github.com/sophia14324/dsc-phase-1-project-v2-4

1.2. Understanding the problem

For as long as there have been movies, film producers, actors, directors and everyone else involved in bring movies to life have relied on huge scale productions to bring in the big bucks. Thus, for the last few decades, movies studios particularly Hollywood have continuously perfected the model to massively increase the cost and returns of movies. It Is no wonder that companies want to venture into the business. This project is about Microsoft wanting to hypothetically venture into the movie production business. The project is tasked to data scientists to analyze movie databases (Rotten Tomatoes, IMDB, TMDB, and Box Office Mojo) and determine film characterizes that will make them better understand the movie industry. This will be achieved by answering the data analysis questions: are movies making money? which movie studios obtain the highest gross domestically? And What is the correlation between runtime minutes and production budget?

1.3. Problem statement

The problem statement is to determine if movies are profitable so that Microsoft can start producing movies. To investigate this, our hypothesis will be:

- Movies are not profitable should not venture into the movie business (Null hypothesis)
- Movies are profitable and Microsoft should venture into the movie business (Alternate hypothesis)

As a big tech company, production of movies is not Microsoft's field. Today, there are a lot of streaming services such as Hulu, Amazon and Netflix that have taken of the industry. It is thus a highly competitive business and so Microsoft has to be at a competitive advantage to venture into film making.

2. Data Understanding

2.1. Data collection

The data was collected from the zippedData folder in the GitHub repository. It was extracted and the following movie datasets used:

- Rotten Tomatoes = rt.movie_info.tsv.gz
- The Numbers = tn.movie_budgets.csv.gz
- Box Office Mojo (BOM) datasets = bom.movie_gross.csv.gz

The data included in the datasets are: movie titles, release dates, production budget, domestic & foreign gross, studios, ratings, genres, directors, writers, runtime minutes, reviews, movie synopses among others.

2.2. Specifying the questions.

The following are the questions to analyze in order to provide recommendations to Microsoft:

1. Are movies making money?
2. Which movie studios obtain the highest gross domestically?
3. What is the correlation between runtime minutes and production budget?

3. Data Preparation

3.1. Selecting data

The data to be used in order to test the null hypothesis (movies are not profitable) are in the data frame Box Office Mojo (BOM), columns movie, production budget and domestic gross. I'll also use the same dataset (Box Office Mojo (BOM)), columns (studio and domestic gross) to answer question 2. Finally, the analysis to find out the correlation between runtime minutes and production budget will be in the datasets Rotten Tomatoes and The Numbers.

4. Data Cleaning

A variety of data cleaning with the panda's library was done on the datasets. The datasets were checked for duplicate values, missing values and outliers. No duplicates were recorded. The missing values were either kept or dropped depending on the use of the data. Other data cleaning procedures such as converting the data types to the appropriate dtypes were also done to ensure uniformity and readability. Some calculations were done i.e., subtraction between domestic gross and production budget to determine the profits movies make.

5. Modeling

This project used the library matplotlib to visualize the variables of the data analysis questions. This communicated connections between data points and structures and conclusions were able to be made from the visualizations.

6. Evaluation

6.1. Results

Question 1: Are movies making money?

Yes, movies are making money than losing money as seen by number of movies, production budget and domestic gross. To evaluate profitability on any dataset, it is imperative the have revenue and costs. In any movie production business there exists some funds/budget that the production team have to spend. For example, money for props, makeup, cars and other finished goods inventory. Revenue is the money they make, domestically or internationally. I chose to analyze the domestic gross as opposed to the international gross because Microsoft's headquarters

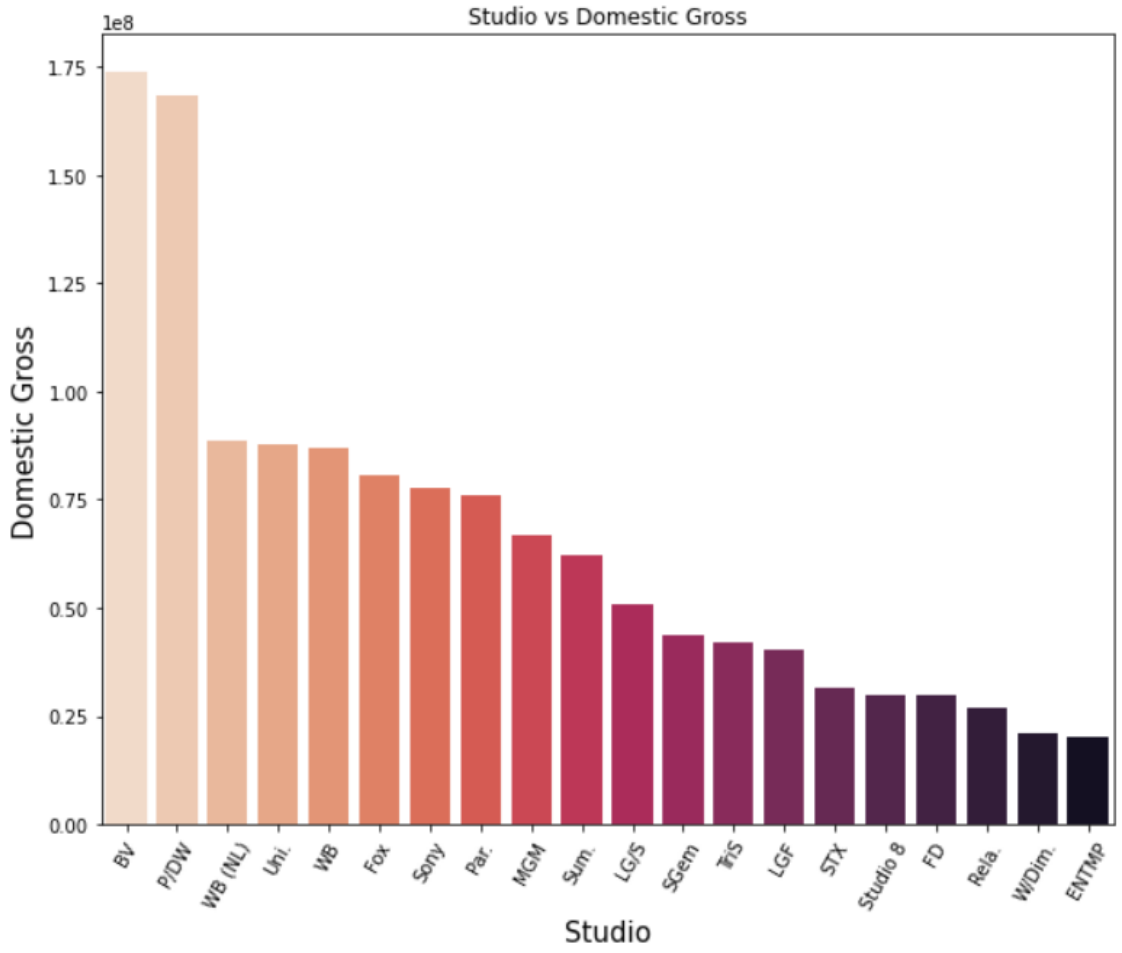
are primarily in the US and if they were to start making movies they would start in the US. The numbers for domestic would matter first.

This question was important in understanding the total amount of money that can be made making movies. It is also important for Microsoft to decide if the venture into movie making is worth to fund versus using the money on other Microsoft projects.

Question 2: Which movie studios obtain the highest gross domestically?

The highest grossing studio is BV, followed by P/DW etc. Microsoft could start observing the two studios and get inspiration if they want to start their own studio. Different studios obtain different levels of gross as observed from the figure.

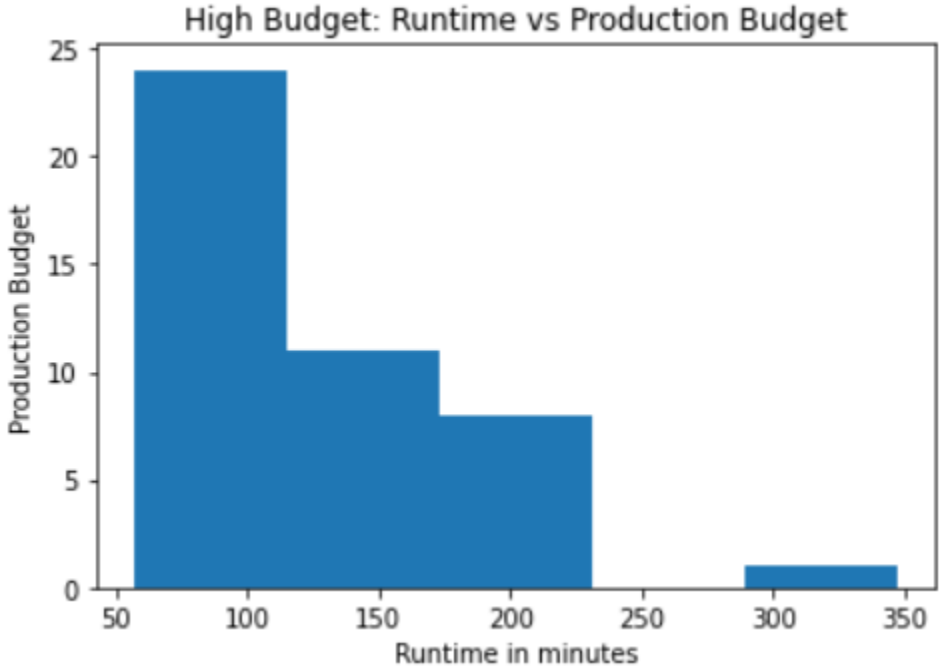
Now that it is clear the movies production is profitable, Microsoft will require to decide on the studio to use for its movies for even better investment. Microsoft is among the best technology firms. So, to continue to be the best it cannot associate itself with cheap or low-quality studios which ay damage Microsoft overall, even if movies generally make profits.



Questions 3: What is the correlation between runtime minutes and production budget?

There is a correlation between runtime minutes and production budget. Movies with few minutes have more production budgets while those with longer running times have lower production budgets.

Once Microsoft start producing movies, it will be imperative to understand some aspects in the movie industry. For example, it would be of interest for them to know how the budget estimate are made, and how these production budgets relate to the length of a movie. If a movie is long the that means there might be more scenes, meaning more materials that are to be included in the production budget. And vice versa.



7. Conclusion

The data understanding, data preparation and data cleaning allowed me analyze, model and evaluate the data on the different datasets. The key takeaways are that movies are profitable; BV and P/DW studios are most profitable compared to the rest and there is a correlation between the runtime minutes and production budget (Less runtimes = more production budget and vice versa).

8. Recommendation

Based on these findings the recommendations to Microsoft are:

1. Microsoft should venture into the movies making business
2. For optimal profits, Microsoft should emulate or have their movies produced by movie studios BV or P/DW.
3. Microsoft should keep their movies however long as long as they know the runtime minutes have a direct relationship with the production budget

9. Next Steps

The next steps I would pursue would be:

1. Genre: Analyze and determine the highest grossing genre domestically and internationally
2. Directors & Actors: Analyze and determine how the choice of directors, producers and/or actors affect the ratings and grossing of a movie.
3. American subscription streaming service: Look into ratings and reviews of different services such as Hulu, Amazon and Netflix to make recommendations to Microsoft that they would follow to increase their viewing rates and maximize profits.