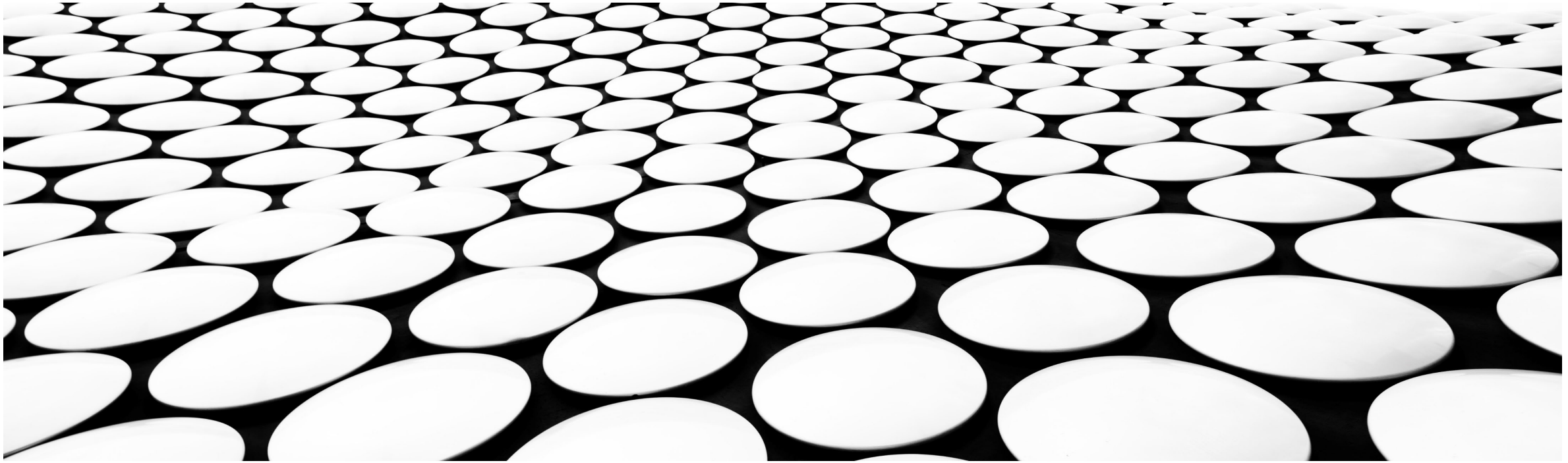
LIGHTS, CAMERA, DATABASE: EXPLORING THE WORLD OF MOVIES

BY: MARIACHARLOTE MBIYU

COURSE : DATA SCIENCE PART TIME 2023

DATE : 17.03.2023



PROJECT BACKGROUND

- A trend of large corporations developing original video content has emerged.
- Microsoft aim to follow suit by establishing their own movie studio.

Challenge :

- Microsoft lack experience in movie-creation and require insights into the types of films that are currently successful at the box office.

Expectation:

- As a independent consultant, your tasked to conduct research on the most profitable movie genres
- Provide actionable insights to the head of Microsoft new movie studio.
- The findings will aid in the decision-making process for the types of films to create.

PROJECT PROCESS OVERVIEW

1. Data source & **Data understanding**
2. **Business problem understanding**
3. Initial data insights

Data Exploration Analysis

1. Merging of the 3 data sets
2. Statistical Analysis
3, Visualization of data

Data cleaning and preparation carried out for three data sets individually.

1. **Summary of results**
2. **Recommendations**

STEP 1: DATA & BUSINESS UNDERSTANDING

Data source & Data understanding

Data was acquired from different international movie databases to support with the project.

Through initial insights, 3 databases were utilized to come to the conclusions of this study

1. rt.movie_info.tsv.gz : Each record represents standard movie information.
2. rt.reviews.tsv.gz: Each record contains reviews and ratings for the movies.
3. tn.movie_budgets.csv.gz : The file contains the production budget and gross sales per movie.

Business problem understanding

The primary objective of this study was to address the following inquiries:

1. Which movie genres are the most popular and what is their average runtime?
2. What is the relationship between production budget and revenue generated?
3. How do movie ratings and reviews influence revenue?
4. Is their a any correlation between production budget and ratings

STEP 2: DATA CLEANING

The below processes were achieved in the data cleaning process:

- Dropping of Unnecessary Columns and rows based on different criteria's.
- Filling data in empty cells.
- Creation of new columns i.e. Return on investment(ROI).
- Converting data into suitable data types i.e. dates, numeric.
- Removing of characters that don't need to be in the columns (commas, currency signs etc.)

STEP 3 : DATA EXPLORATION ANALYSIS

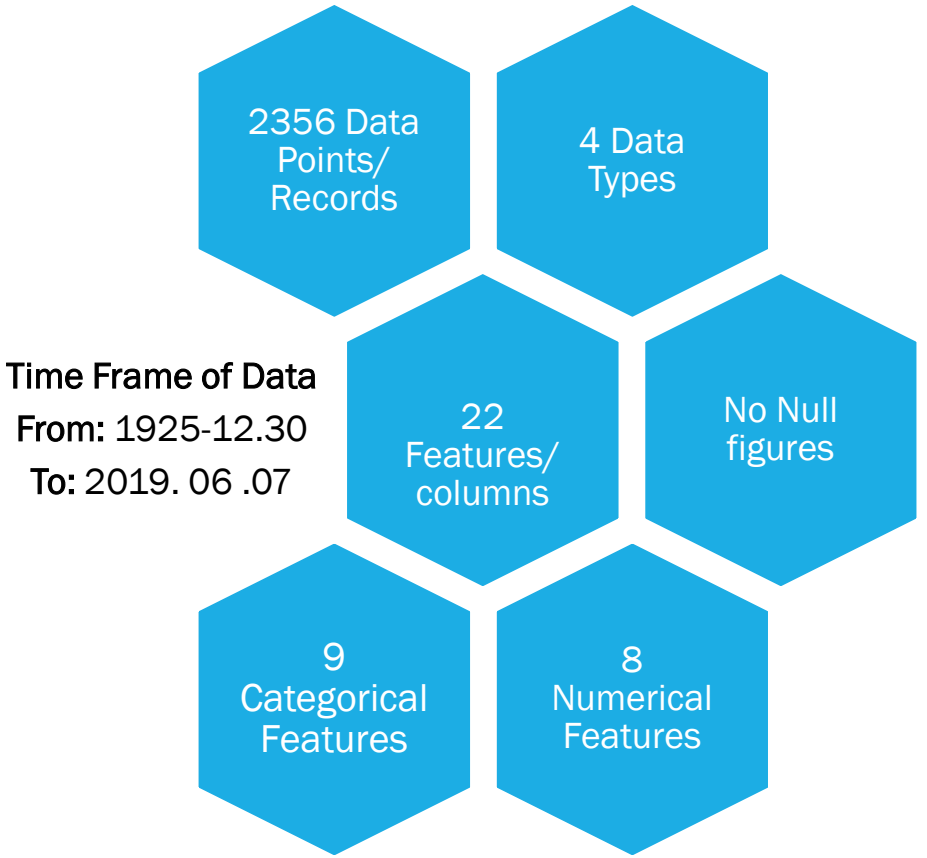
Merging

In this step , a merged Data Frame was used for Data Exploration

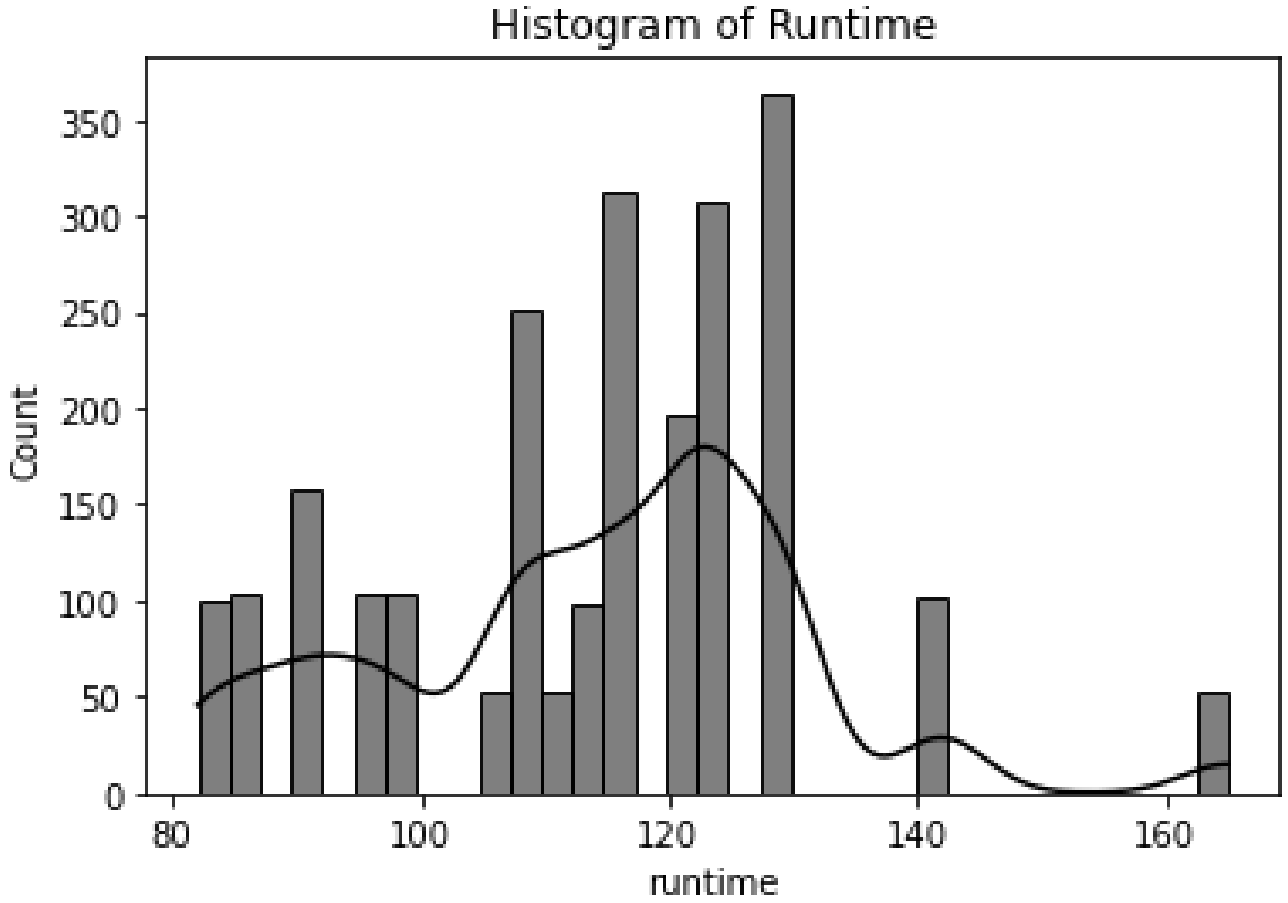
Features of the data utilized are beside :

Additionally :

- 4 features are date time objects
- 1 feature is a Boolean (True/False)

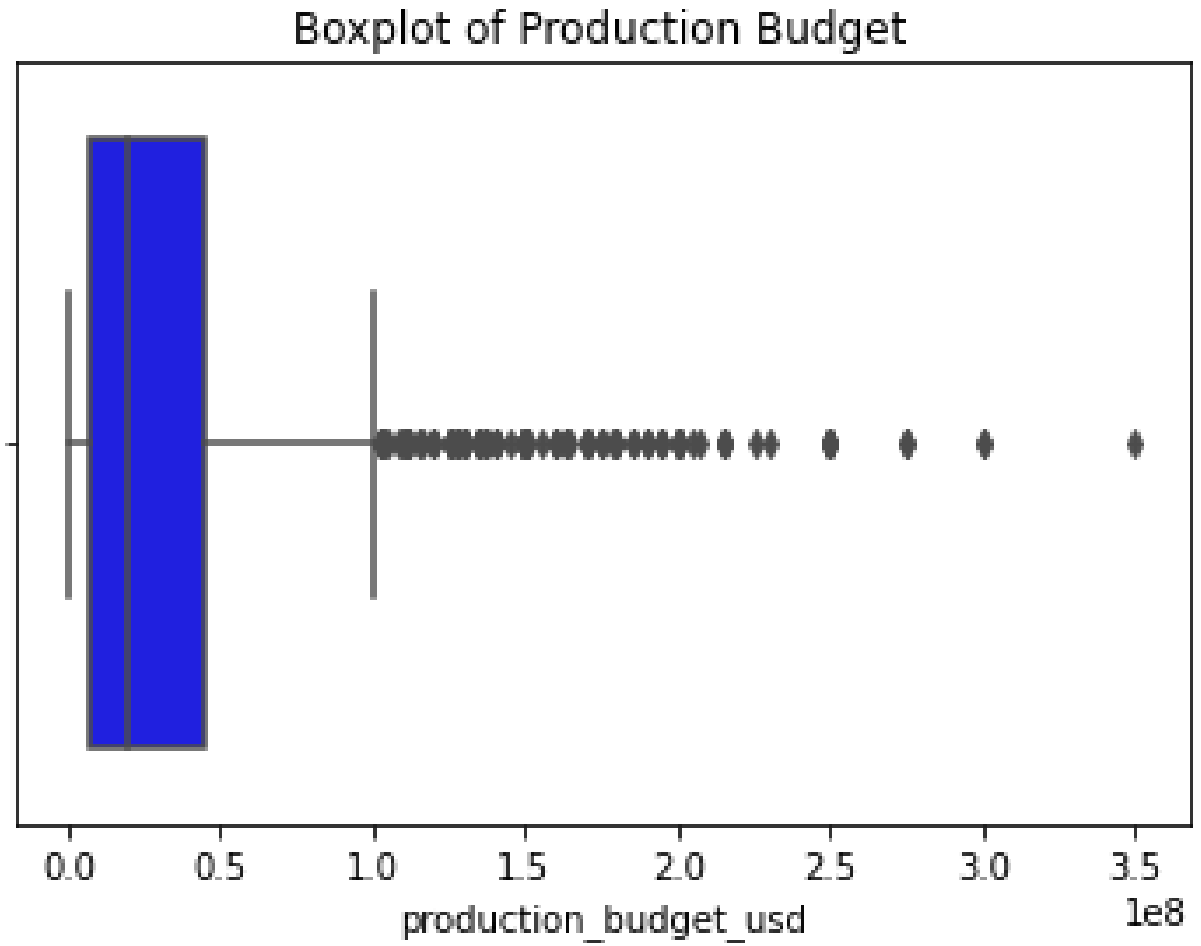


STEP 3: DATA EXPLORATION ANALYSIS SUMMARY STATISTICS - RUNTIME

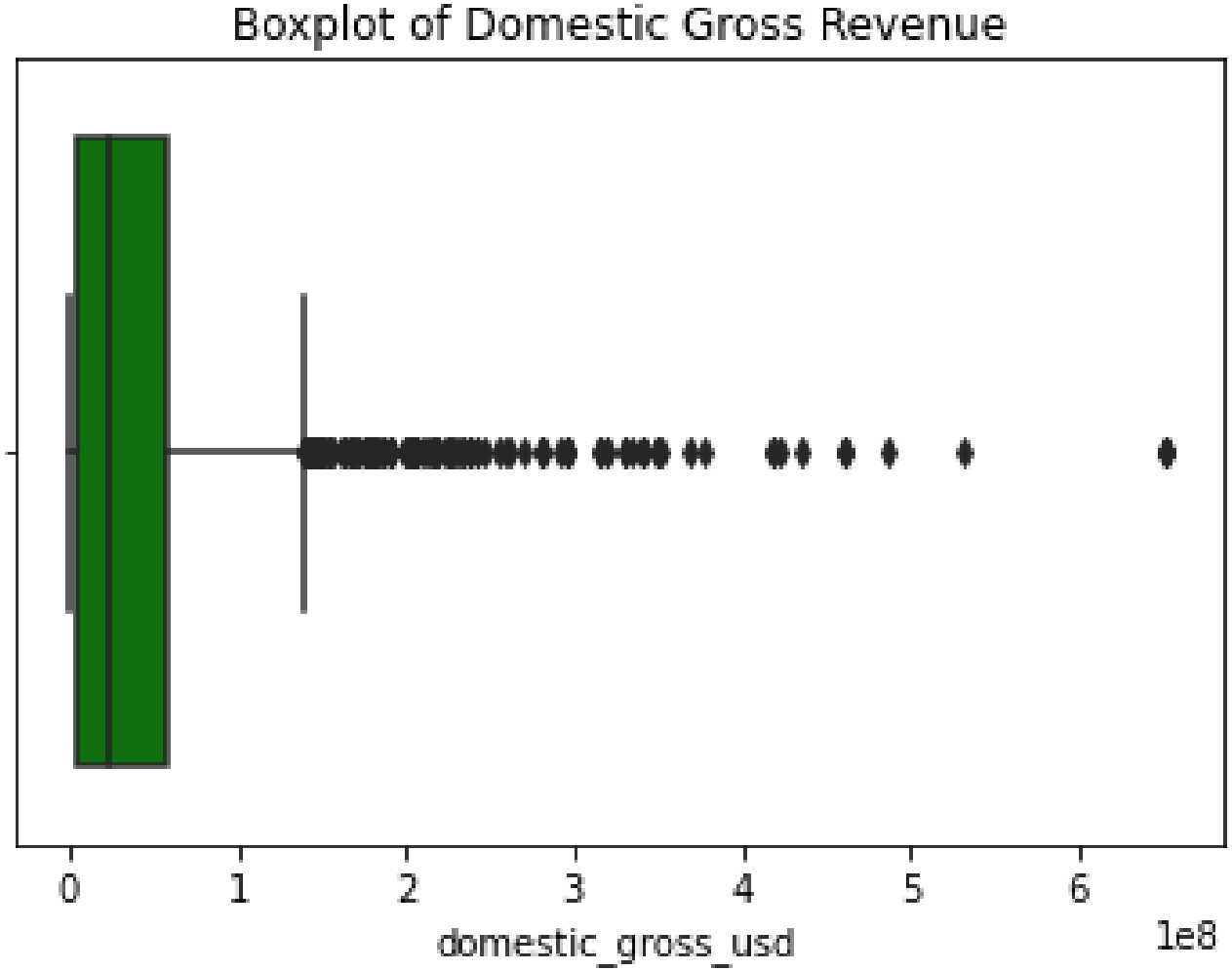


- The Average run time for movies was captured at 114 minutes

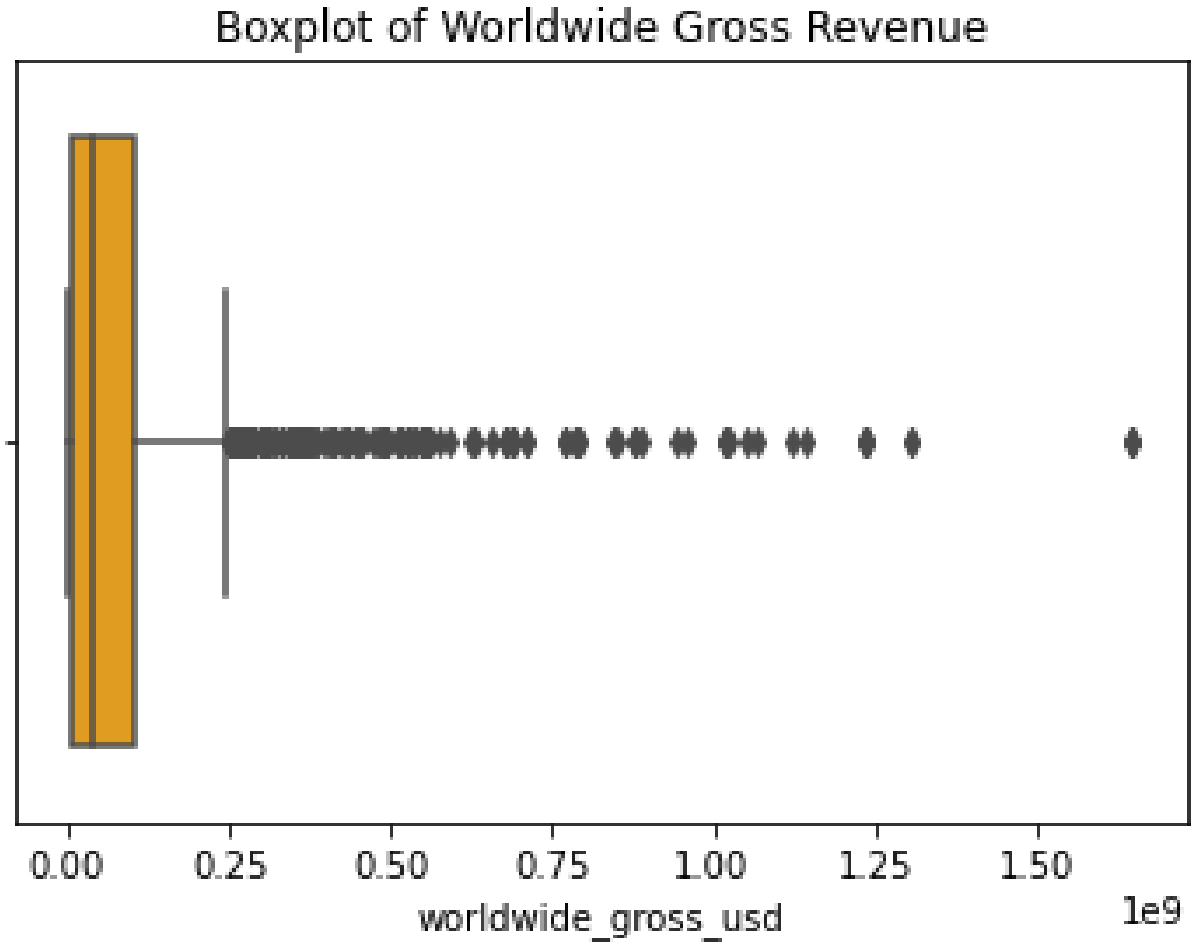
STEP 3: DATA EXPLORATION ANALYSIS
SUMMARY STATISTICS – PRODUCTION BUDGET



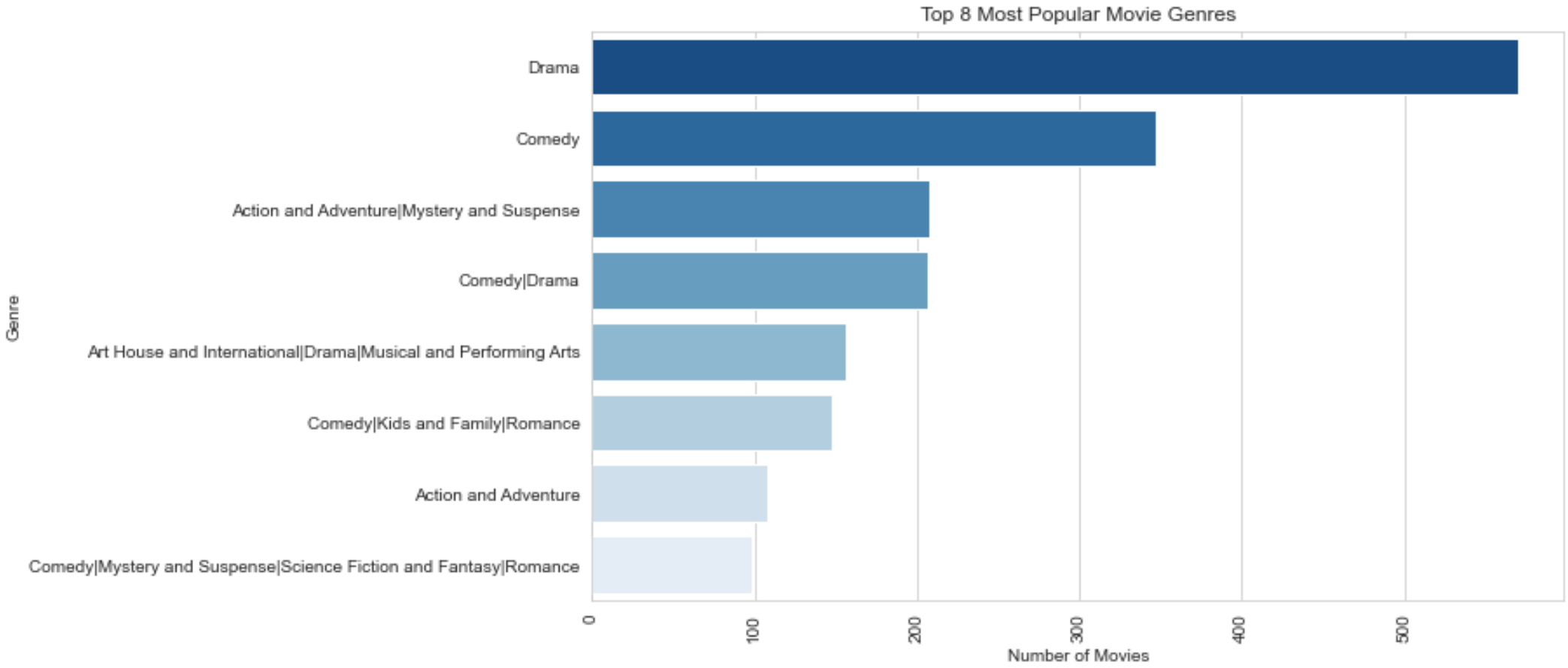
STEP 3: DATA EXPLORATION ANALYSIS
SUMMARY STATISTICS –DOMESTIC GROSS REVENUE



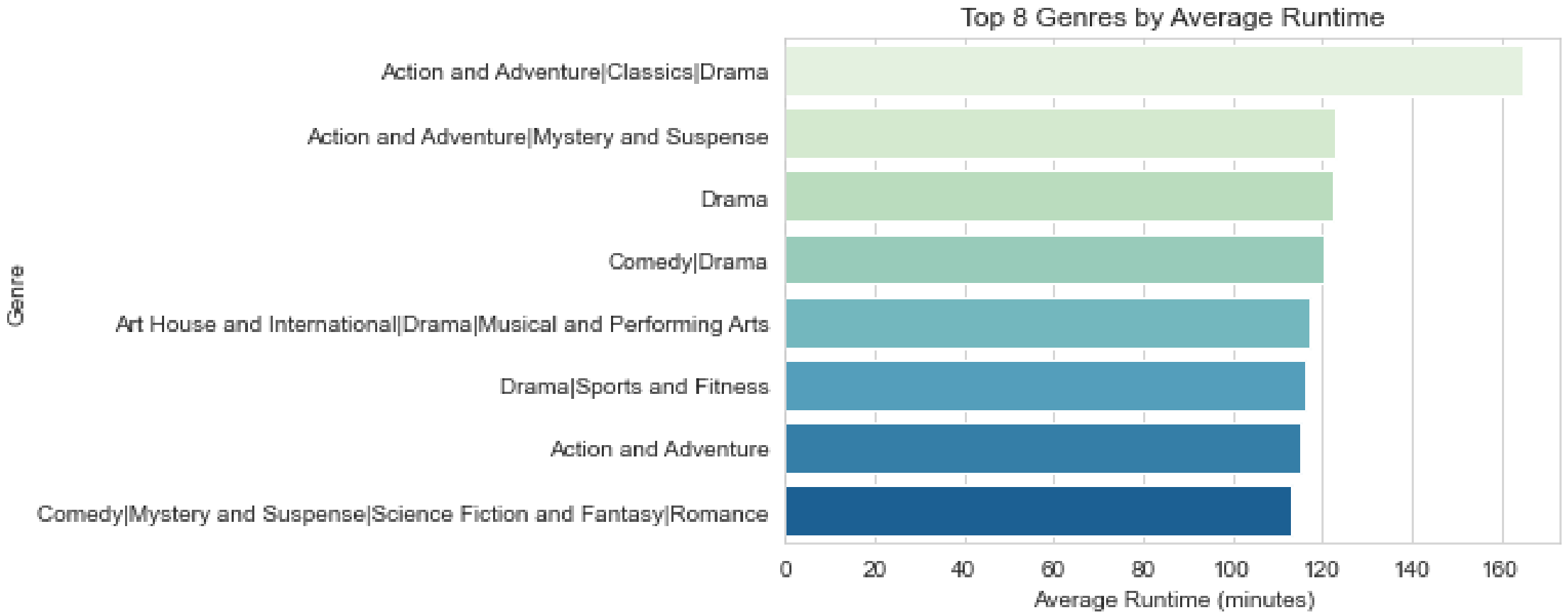
STEP 3: DATA EXPLORATION ANALYSIS
SUMMARY STATISTICS –WORLDWIDE GROSS REVENUE



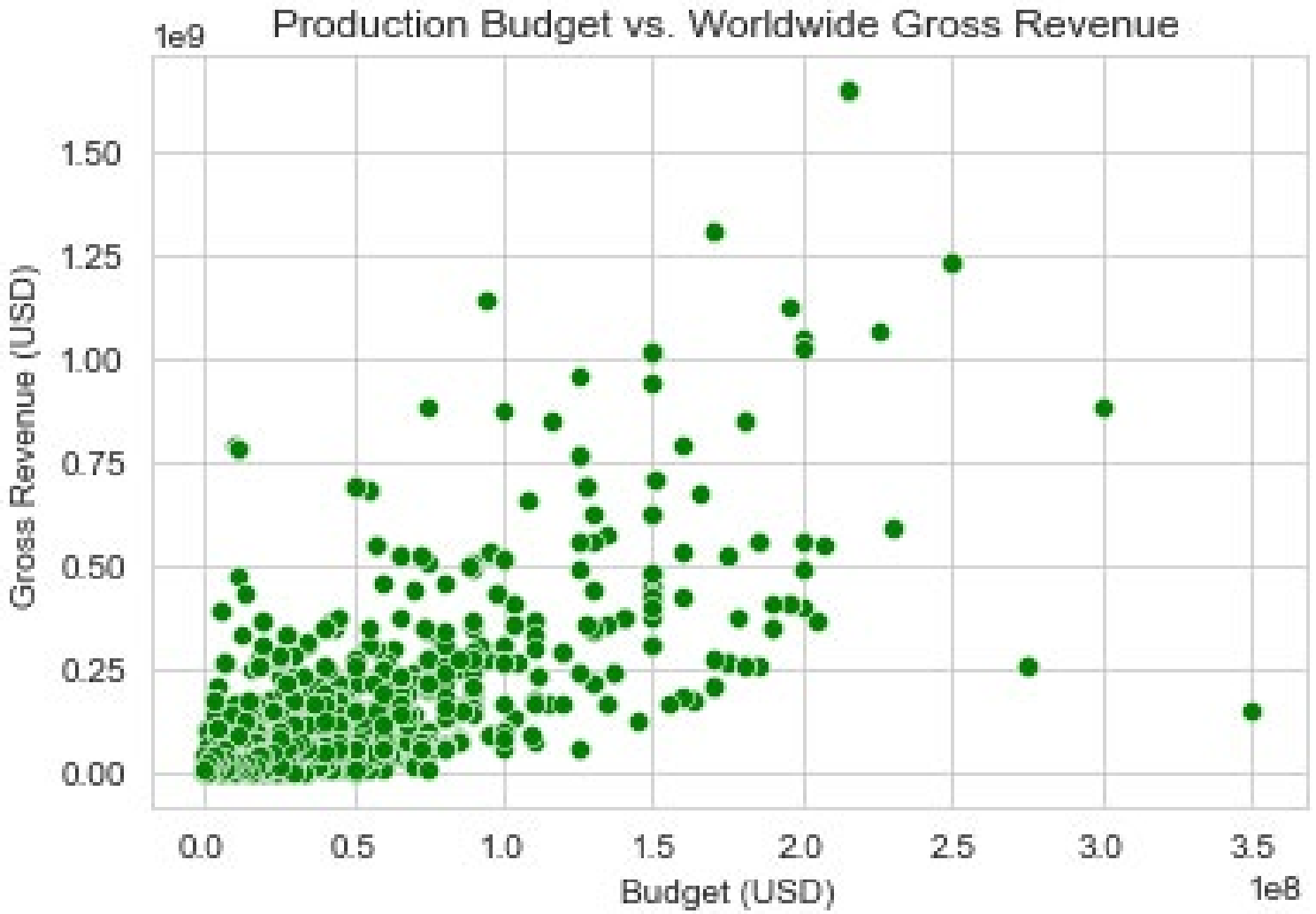
8 MOST POPULAR MOVIE GENRES



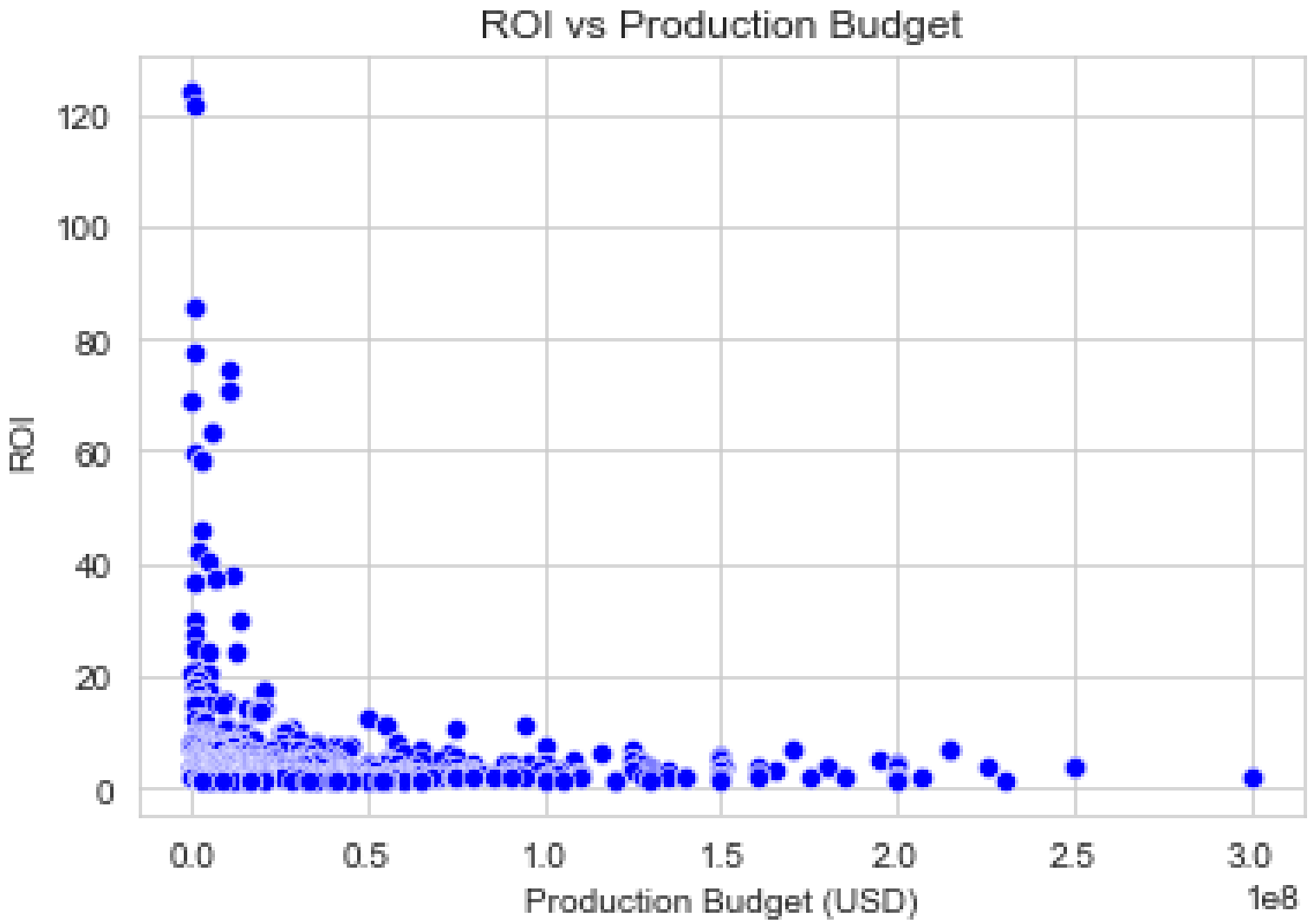
TOP 8 GENRES BY AVERAGE RUNTIME



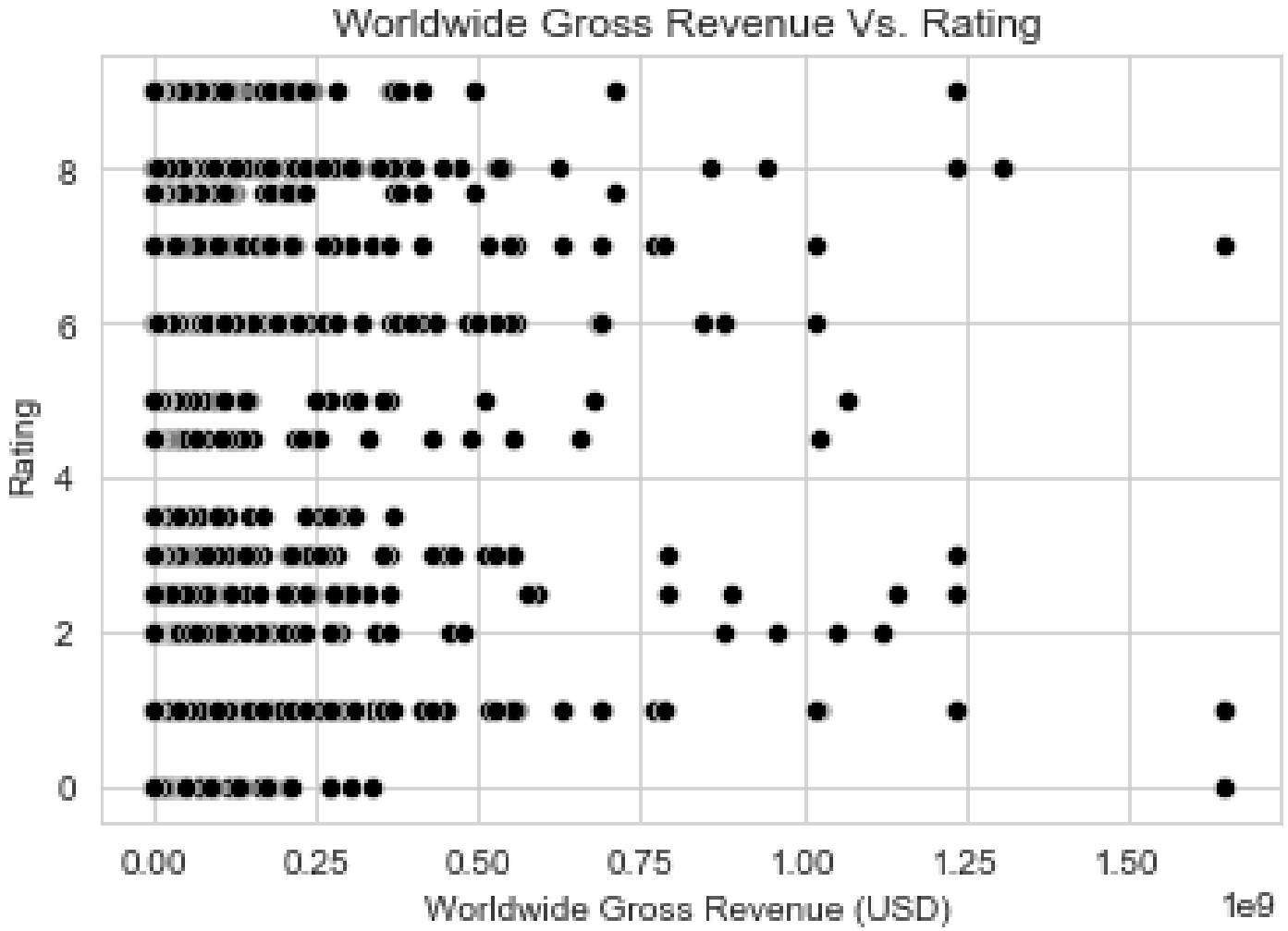
PRODUCTION BUDGET VS. WORLDWIDE GROSS REVENUE



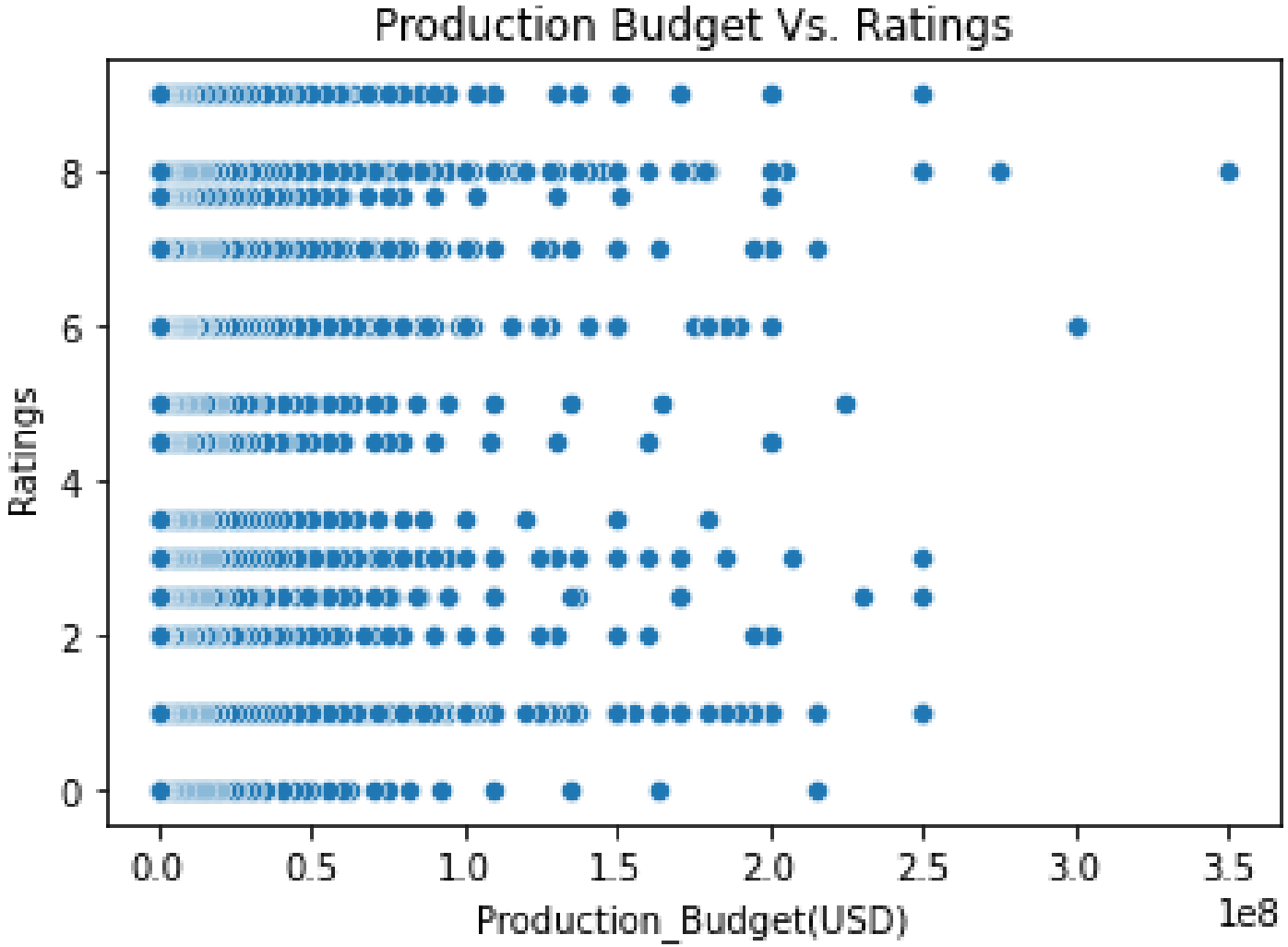
ROI VS PRODUCTION BUDGET



WORLDWIDE GROSS REVENUE VS RATING



PRODUCTION BUDGET VS RATINGS



- A slight tendency was noted on movies with higher production budgets to have slightly higher ratings, but the relationship is not very strong.

RECOMMENDATIONS

Recommendations to consider:

Genre and Runtime

- Invest in drama and comedy movies as they are the most popular genres.
- To Keep in mind the appropriate length of each movie . The average mean is 114 minutes

Budgets & Ratings

- Although the average production budget spent by movie makers in this dataset was 34.6 Million, consider that the majority of movies spent between USD 7 million and USD 45 million.
- This means a lower budget can always be utilized.
- Microsoft should not solely rely on higher production budgets as a guarantee for higher ratings or revenues.

Revenue & Ratings

- Microsoft should keep in mind that there is a weak negative correlation between worldwide gross revenue and movie ratings. This means that having high ratings does not necessarily guarantee high revenues.
- Thus should not rely heavily on ratings.

CONCLUSION

Finally, Microsoft should keep in mind that while the majority of movies in the dataset generated between USD 7 Million and USD 103 Million in worldwide revenue, the highest worldwide gross revenue registered was USD 1.65 Billion.

Therefore, it is important to remain open to the possibility of high revenue generation, and not limit investment opportunities based on past revenue trends.

THANK YOU

FURTHER QUERIES CAN BE REDIRECTED TO MY EMAIL :

MARICHARLIYU@GMAIL.COM