Microsoft Movie Studio Analysis

Data-Driven Film Production Proposal

By Michelle Joan Wangari

Business Understanding

- Microsoft is entering a highly competitive film industry dominated by experienced studios such as Disney and Netflix. Success in this space often depends on understanding audience preferences, market trends, and historical performance of different types of films.
- Utilizing data can provide insights into the characteristics of successful movies, helping new players like Microsoft navigate decisions about content creation, marketing, and distribution. A clear understanding of these factors is essential for positioning the studio effectively in the entertainment industry.

Problem Statement

- Main-What types of movies perform best at the box office?
- Which genres are most popular with audiences?
- How does runtime, release year, and audience ratings affect a movie's success?
- What genres consistently generate the highest average box office revenue?

Objectives

- Main- Understand the type of movies perform best at the box office.
- Identifying movies that are most popular with audiences.
- Finding out how runtime, release year and audience ratings affects a movie's success.
- Genres that consistently generate the highest average box office revenue.

Data Understanding

- In data understanding the following actions were carried out:

1. Loading the data
2. Discovering the shape and structure of the dataset
3. Identifying data sources
4. Data description
5. Merging
6. Characteristics of the data

Notes

a) Shape and structure- Examining their shape and structure helps identify missing values and key columns for merging

b) Data Source- IMDB is a widely used platform for movie data, offering detailed information on titles, ratings, and genres.
Box Office Mojo provides financial performance data, specifically domestic gross revenue.
These sources are appropriate for identifying patterns in movie success, both critically and financially

c) Data Description-Columns used for data analysis is identified

d) Merging- The IMDB datasets are merged using the tconst identifier, ensuring a direct match between titles and ratings. The combined IMDB dataset is then merged with Box Office Mojo's domestic gross data using the movie title (primary title). This allows for analysis of both movie content and financial performance.

e) Characteristics of Data -The merged dataset contains information on each movie's title, genre(s), release year, runtime, audience rating, number of votes, and box office gross. This combination allows for analysis of how different features contribute to financial and audience success. Most movies have one or more genres listed, runtimes range from short films to epics, and ratings range from below 5 to above 9. Some missing or extreme values were identified and are addressed in the data cleaning section.

.

Data Cleaning

The dataset was cleaned and prepared using the following steps:

1. **Dropped missing values** in key columns: genres, runtime Minutes, and start Year.
2. **Removed duplicates** across all datasets.
3. **Handled outliers** by identifying extreme values in runtime Minutes and domestic gross.
4. Cleaned data types: **Converted** start Year to integer and formatted domestic gross as numeric.
5. **Filtered ratings** to include movies with at least 1,000 IMDB votes.
6. **Renamed columns** (e.g., title → primary Title) for consistent merging.
7. **Engineered new features**, including:

i. Decade: based on release year

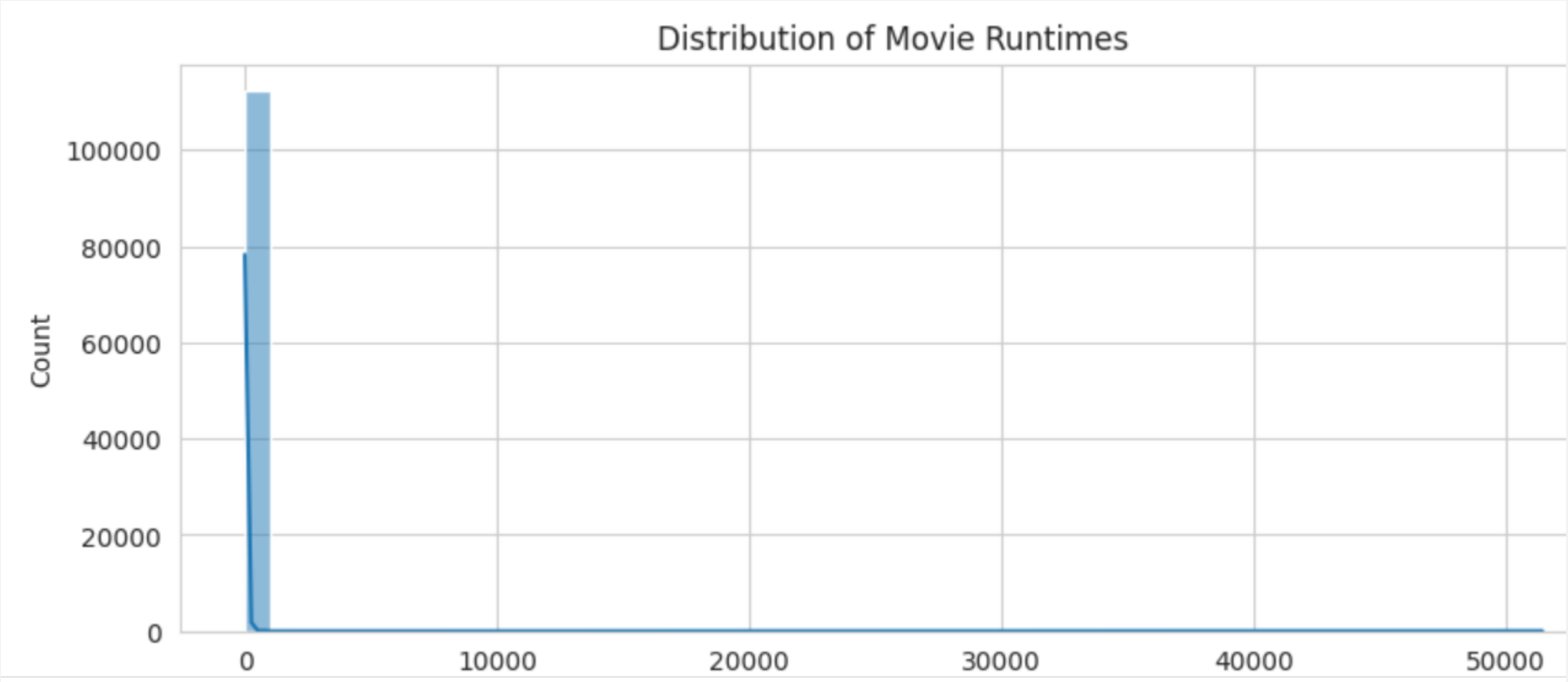
ii. Genre count: number of genres assigned to a movie

Handling outliers

This filters out:

Very short entries (e.g., short films or errors)

Very long values (likely not theatrical releases)



EDA- Exploratory Data Analysis

- Univariate analysis

I. Distribution of IMDB

II. Distribution of domestic gross

III. Distribution of movie runtimes

- Bivariate analysis

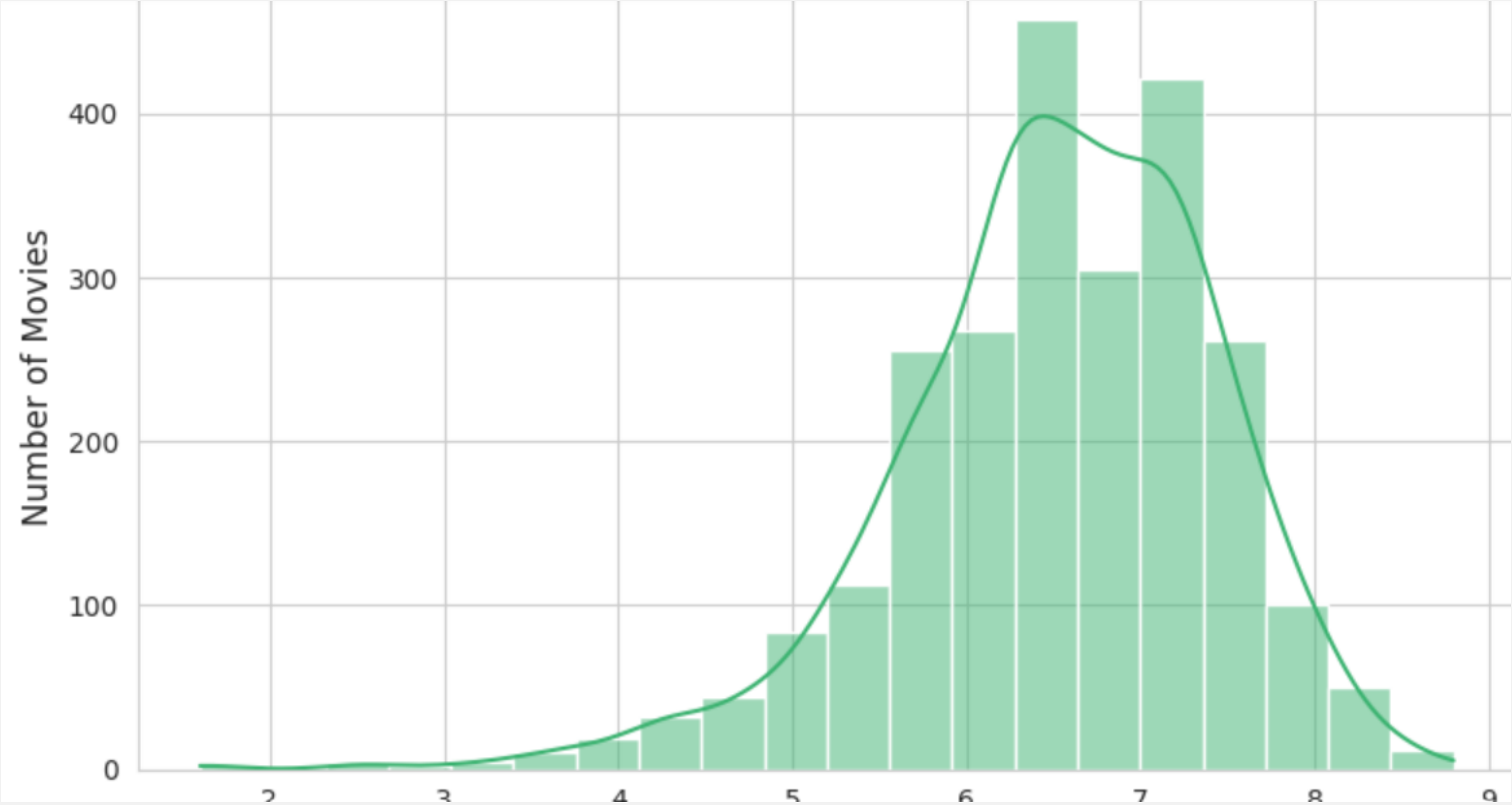
I. IMDB Rating vs. Domestic Gross

II. Genre vs. Average Gross

III. Runtime vs Domestic Gross

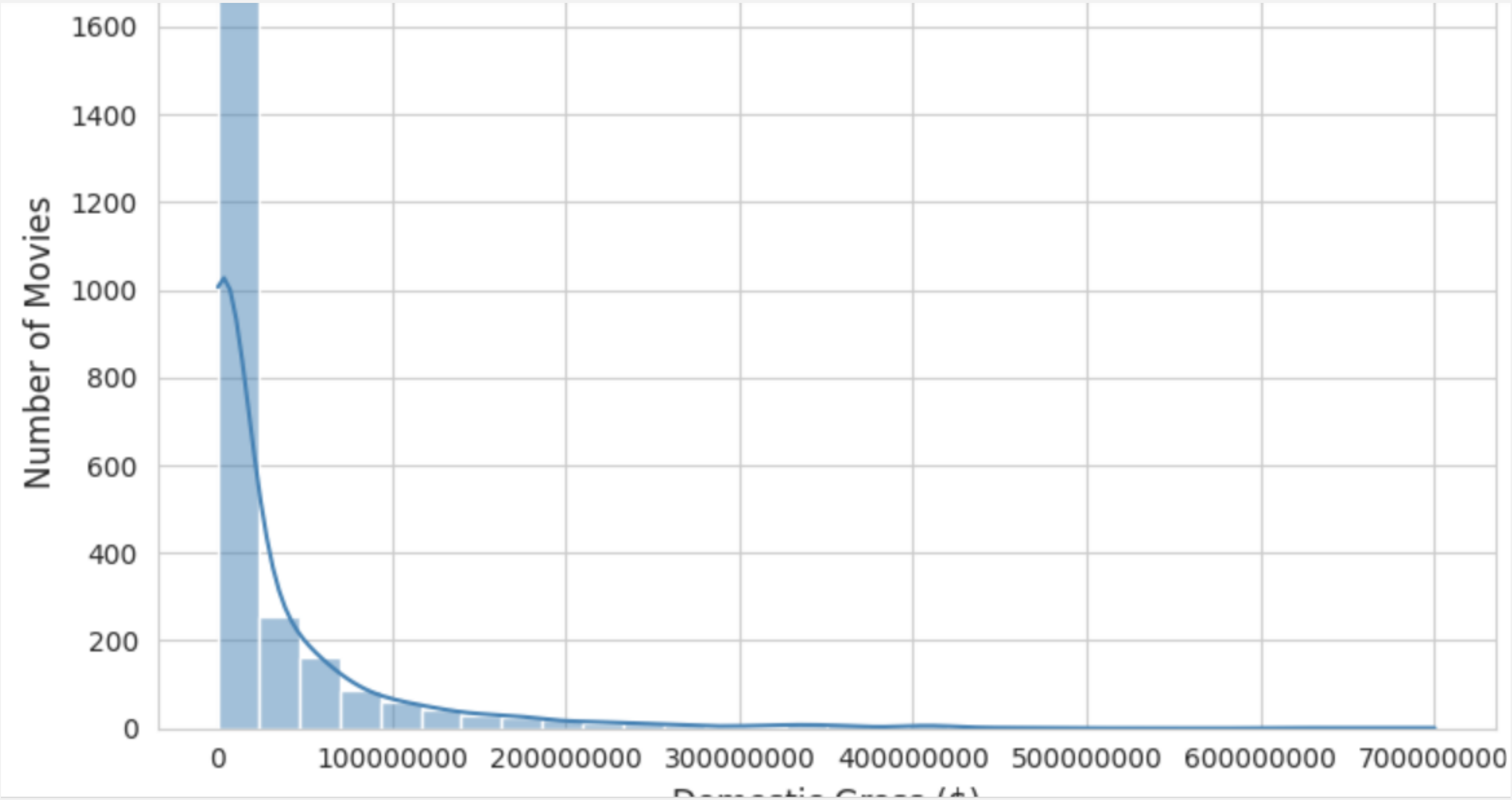
Distribution of IMDB

The majority of movies have IMDB ratings between 5.5 and 7.5. Ratings above 8 are relatively rare, suggesting that most movies are received with moderate audience approval.



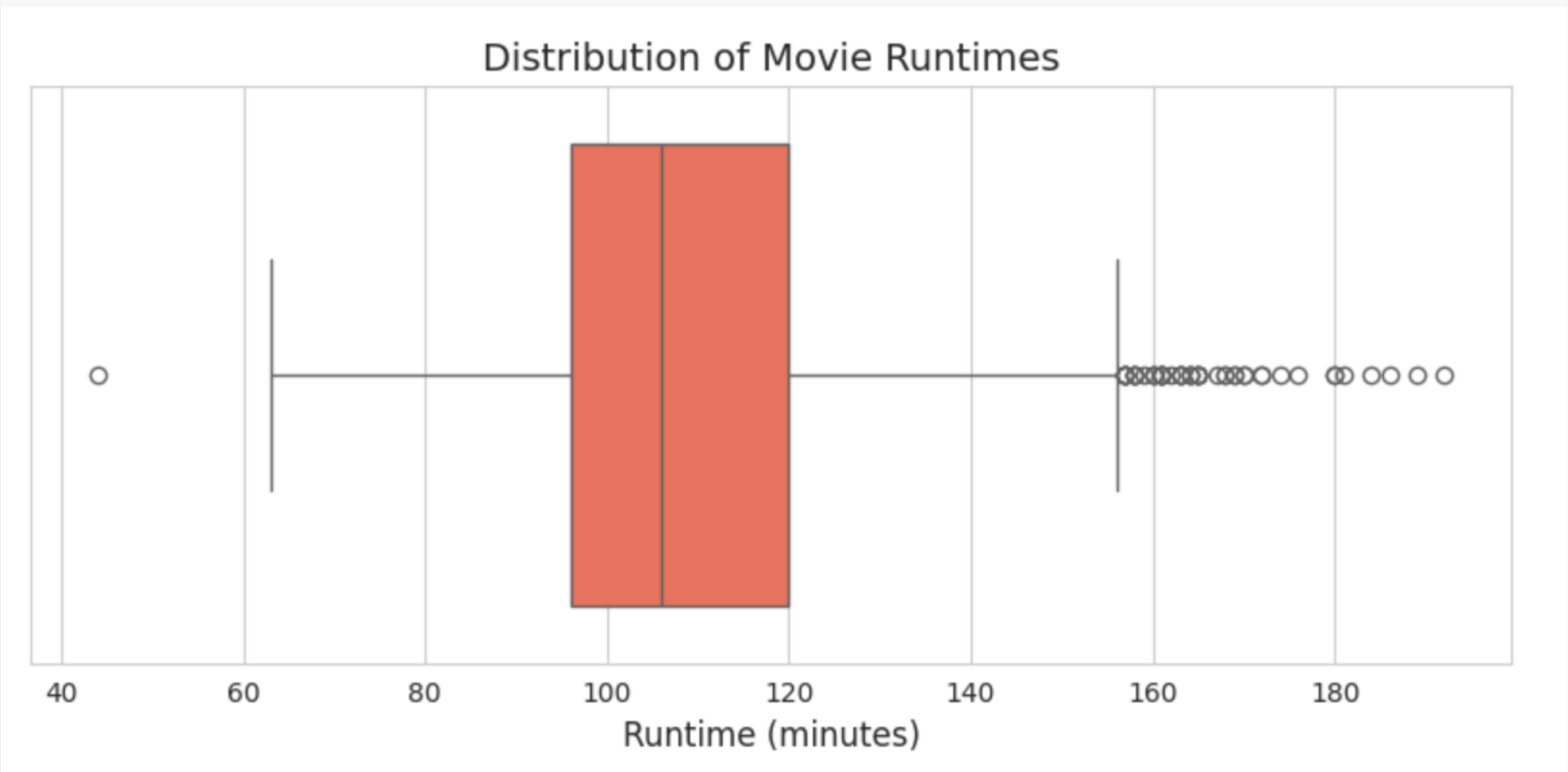
Distribution of domestic gross

Most movies earned under $100M domestically. Only a few high performers grossed significantly more, which indicates a right-skewed distribution.



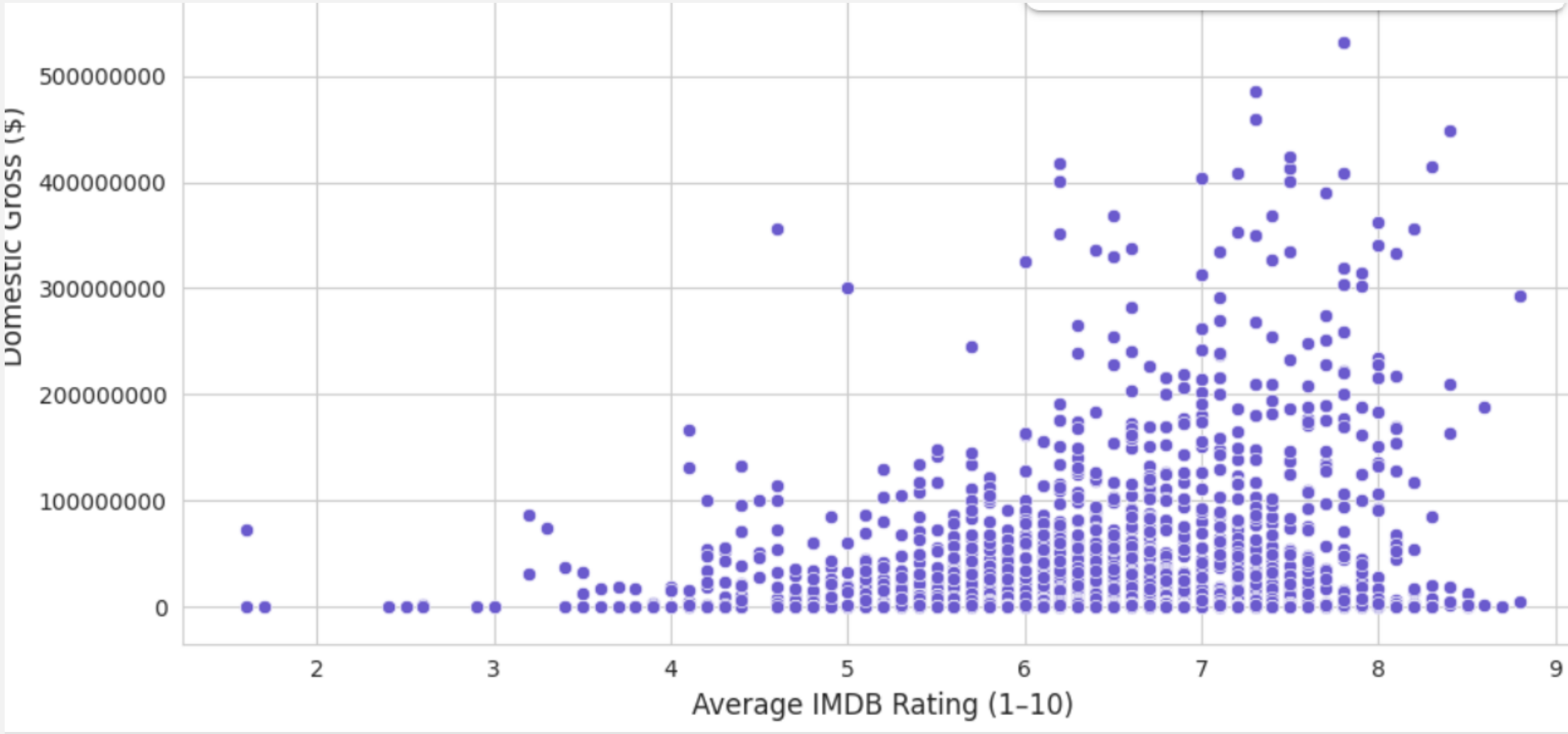
Distribution of movie runtimes

The typical runtime lies between 90–120 minutes. A few outliers exist on both ends, but most films fall within expected industry standards.



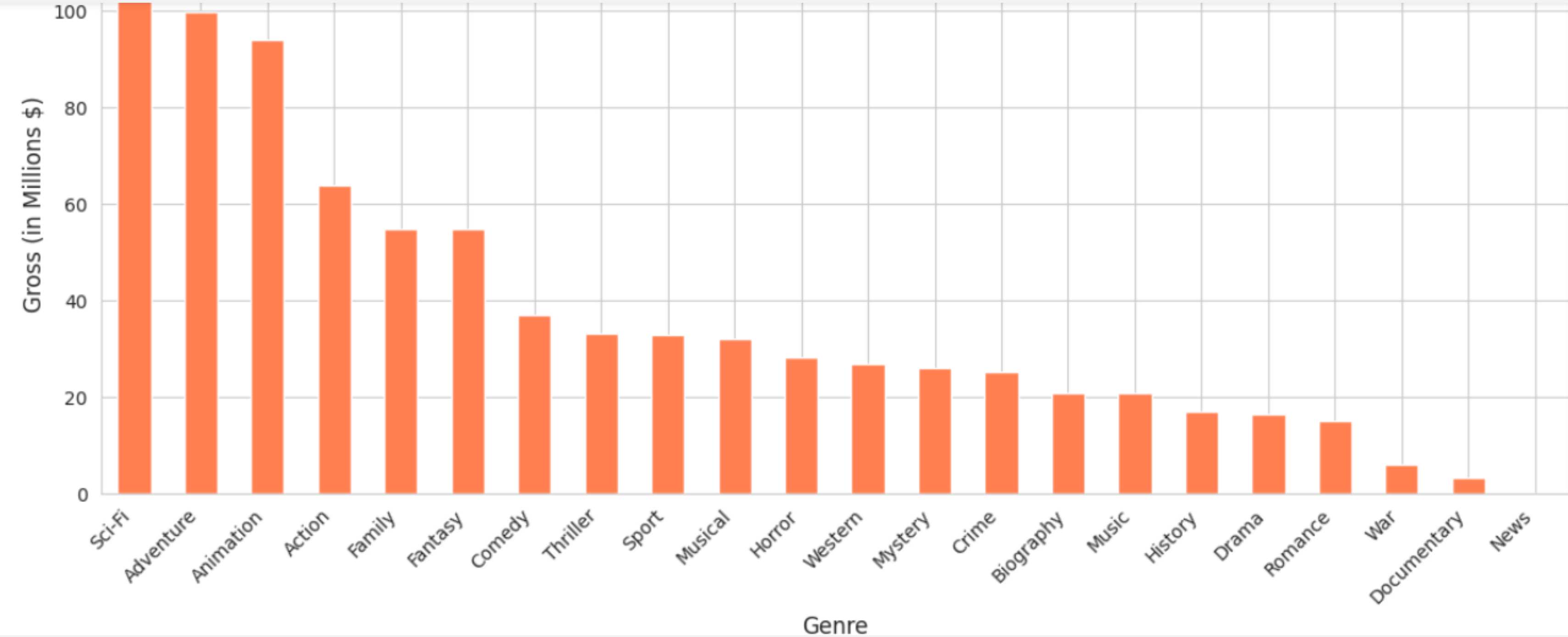
IMDB Rating vs. Domestic Gross

A slight upward trend suggests that movies with higher audience ratings tend to perform better at the box office, though the relationship is not perfectly linear.



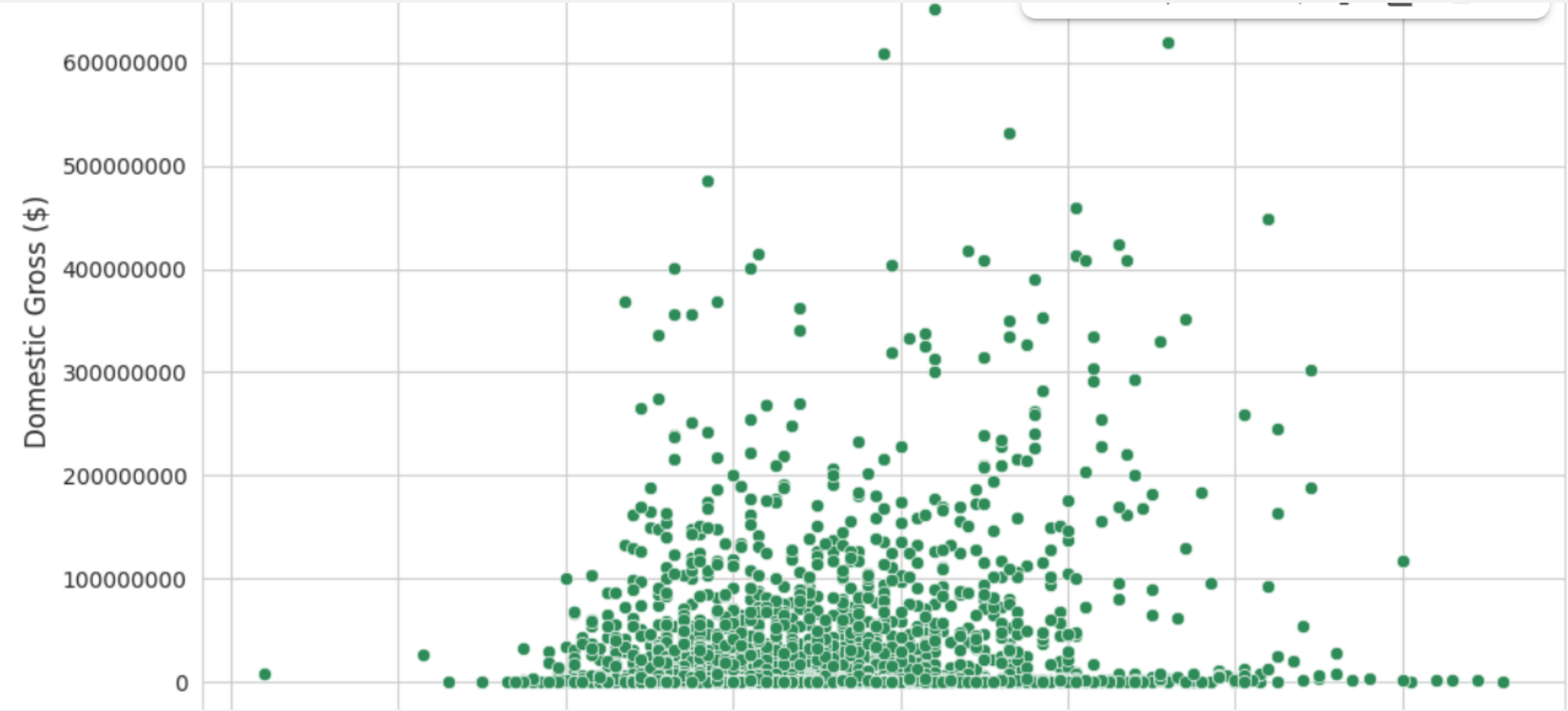
Genre vs. Average Gross

Action, Adventure, and Sci-Fi movies generate the highest average gross, suggesting they are top-performing genres.



Runtime vs Domestic Gross

Very short or very long movies tend to underperform, while those around 100–130 minutes often perform better — hinting at an ideal length for box office success.



Conclusion

- 🎬 Action, Adventure, and Sci-Fi genres consistently generate the highest average box office revenue, showing strong audience demand for high-energy, big-production films.
- 🌟 Higher IMDB ratings are loosely associated with better box office performance. Well-received movies tend to earn more, suggesting audience satisfaction contributes to financial success.
- ⏱️ Runtimes between 90 and 130 minutes are most common among successful movies. Films that are too short or excessively long are less likely to perform well.
- 🎞️ Most movies have moderate ratings (5.5–7.5), and truly high-rated titles (above 8.0) are relatively rare. These may indicate standout or niche successes.
- 💰 The majority of films gross under $100 million, showing that blockbuster-level success is rare, and reinforcing the value of identifying patterns that consistently lead to mid- or high-level returns.

Recommendations

- 🎬 Prioritize Action, Adventure, and Sci-Fi genres — these genres consistently perform the best financially and align with modern audience interests.

- 🎟️ Target PG-13-level content — these films typically appeal to a wide audience without limiting content flexibility.

- 🌟 Use early IMDB ratings or feedback to guide investment and marketing — higher audience scores are generally linked to stronger box office returns.

- ⏱️ Keep films within the 90–130-minute range — runtimes in this range match the majority of high-performing movies.

Thank you

Let’s help Microsoft create its next box office success.

Feel free to ask any questions or request clarifications.

By: Michelle Wangari Joan