ENTERING THE FILM INDUSTRY: A DATA-DRIVEN STRATEGY FOR BOX OFFICE SUCCESS

THIS WORK WAS FACILITATED BY GROUP 7:

Lily Chepngetich

Tiffany Eva

Faith Koech

Munene Gitonga

Timothy Munene

Thomas Amuti

HIGHLIGHTS

- Project Overview
- Business Understanding
  - Business Goals
- Data Understanding
- Data Preparation, Analysis and Visualization
- Results
- Recommendations
- Suggestions for further analysis

PROJECT OVERVIEW

- This project:
  - Analysed box office performance and movie attributes.
  - Discovered patterns linked to commercially successful films.
    - This was done using Exploratory and Inferential Data Analysis.

BUSINESS UNDERSTANDING AND GOAL

- Major companies are investing heavily in the film and video industry.
- Our company plans to enter this industry by launching a movie studio.
  - But they lack film production strategies and experience.
- We are therefore tasked with:
  - Determining which movie genres are the most commercially and financially viable.

BUSINESS QUESTIONS

Main business question:
What movie genres are the most commercially successful and financially viable?

Specific Questions:
1. Which genres are commonly produced?
2. Which genres received the highest audience ratings?
3. Which genres attracted the largest audiences/ viewership?
4. What was the production costs of the 5 most common movie genres?
5. What was the income generated from the 5 most common movie genres?

REGRESSION ANALYSIS QUESTION

Sub-analysis question:

1. What is the relationship between production costs and gross income?

DATA UNDERSTANDING

We used the following data sources:

1. IMDB – Provided in SQL database format, containing detailed movie metadata, ratings, and cast/crew information.

2. Box Office Mojo – CSV/TSV files with box office performance data, including domestic and international gross.

3. The Numbers – CSV/TSV files offering financial data such as production budgets, box office performance, and revenue trends.

LOADING NECESSARY LIBRARIES

- We used several libraries to analyse the dataset:

1. Pandas - to read, edit and save our python files.
2. Matplotlib and seaborn - for visualizations.
3. Statsmodels and scipy.stats – for inferential statistics and regression models.

DATA PREPARATION

- Prior to data analysis we:
  1. Reviewed all 3 datasets(IMDB, The Numbers, Box Office).

  2. Selected the needed columns.
     - Movie genres, average rating, number of votes and ordering from IMDB.
     - Domestic gross, foreign gross from the Box Office dataset.
     - Production budget, Domestic gross and Worldwide gross earning from The Numbers.

  3. Sorted out missing data.
     - 1 entry missing in the IMDB dataset.
     - 3387 entries missing in the Box Office dataset.

DATA PREPARATION

- We also:

4) Identified and addressed any duplicate records.
   - There were no duplicates in the datasets.

4) Ensured correct column data type assignments.

5) Prepared clean and reliable datasets for analysis.

DATA ANALYSIS AND VISUALIZATION

- We combined:

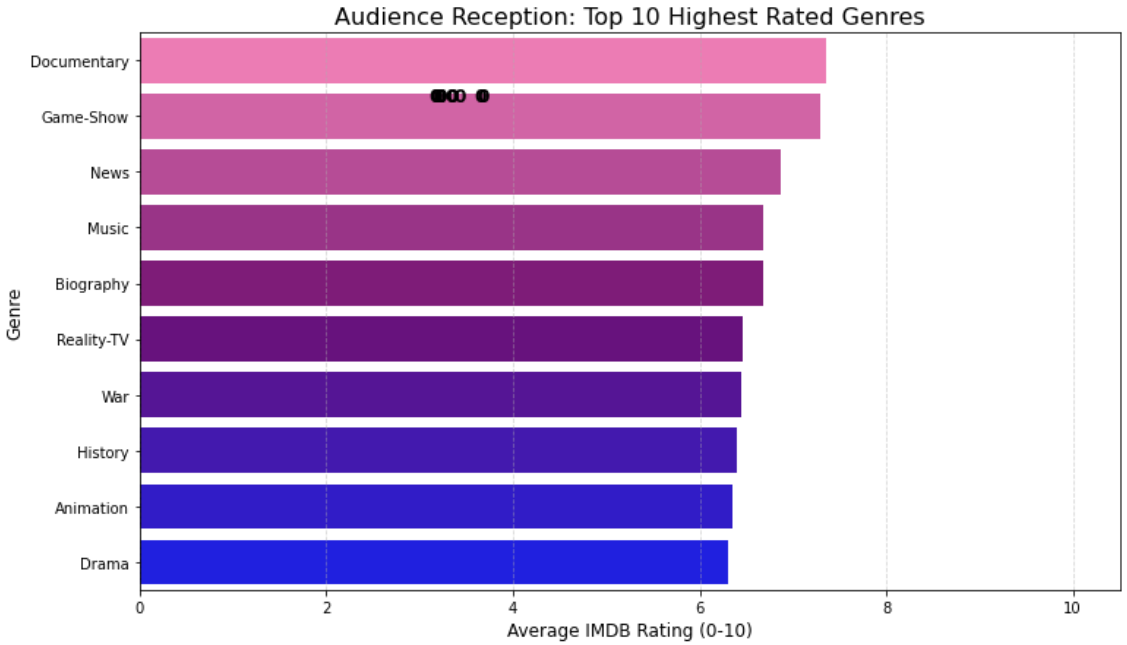
    1. Descriptive data analysis - to determine measures of central tendency.
    2. Inferential statistics - to analyse differences in genres on variables such as rating.
    3. Regression analysis - to assess relationship between production costs and income generated.

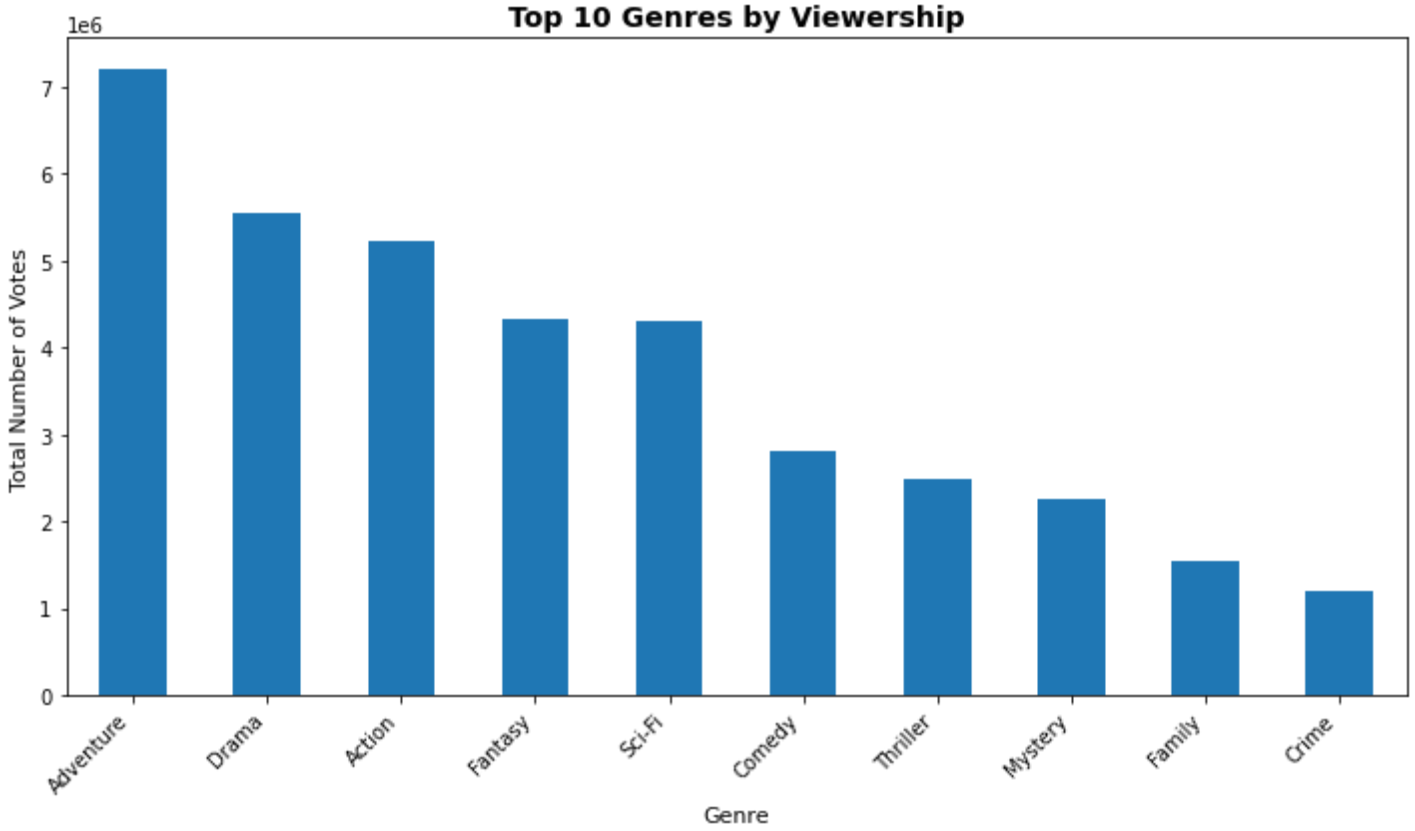
- Findings were presented mostly in bar graphs.

RESULTS

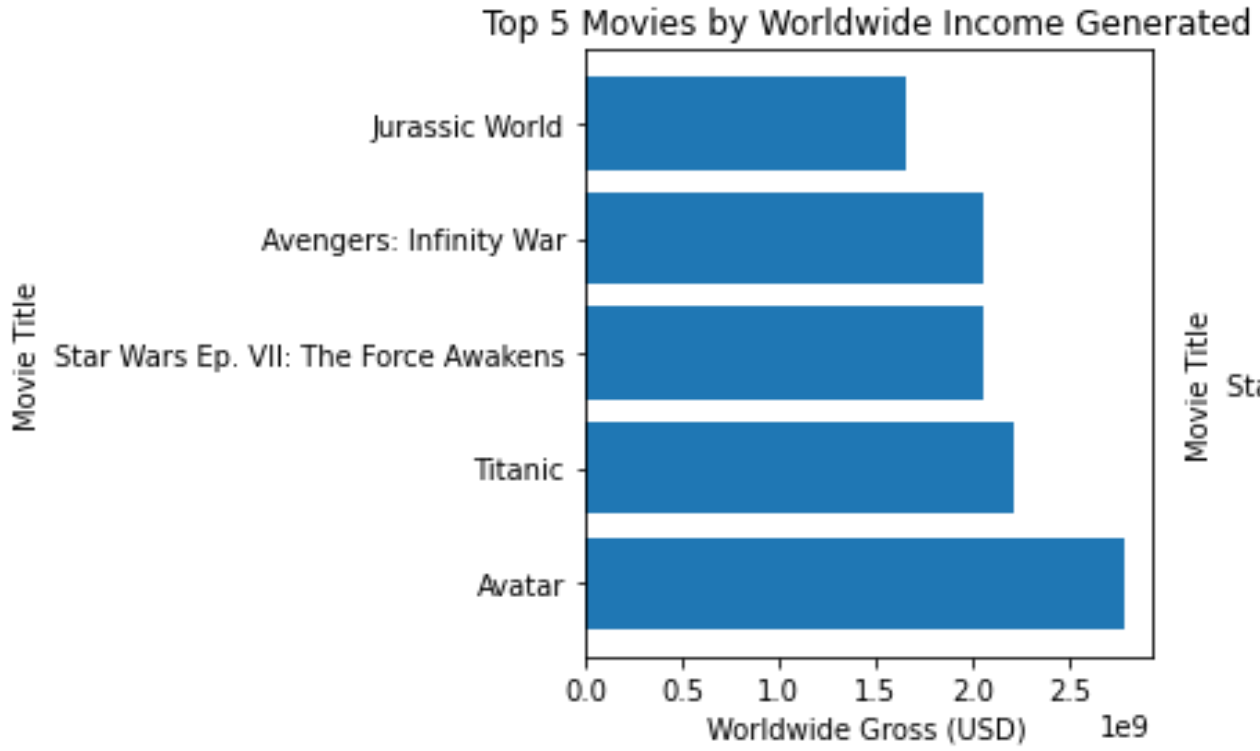
- ✓Drama, Comedy and Documentary were most commonly produced.
- ✓Biography, Music and Documentary received the highest total votes.
- ✓Documentary, Gameshow and News received the highest audience ratings.
- ✓Action, Adventure, and Science Fiction had the highest income generated.
- ✓On average, every dollar spent on production yields about $3.14 in income generated.

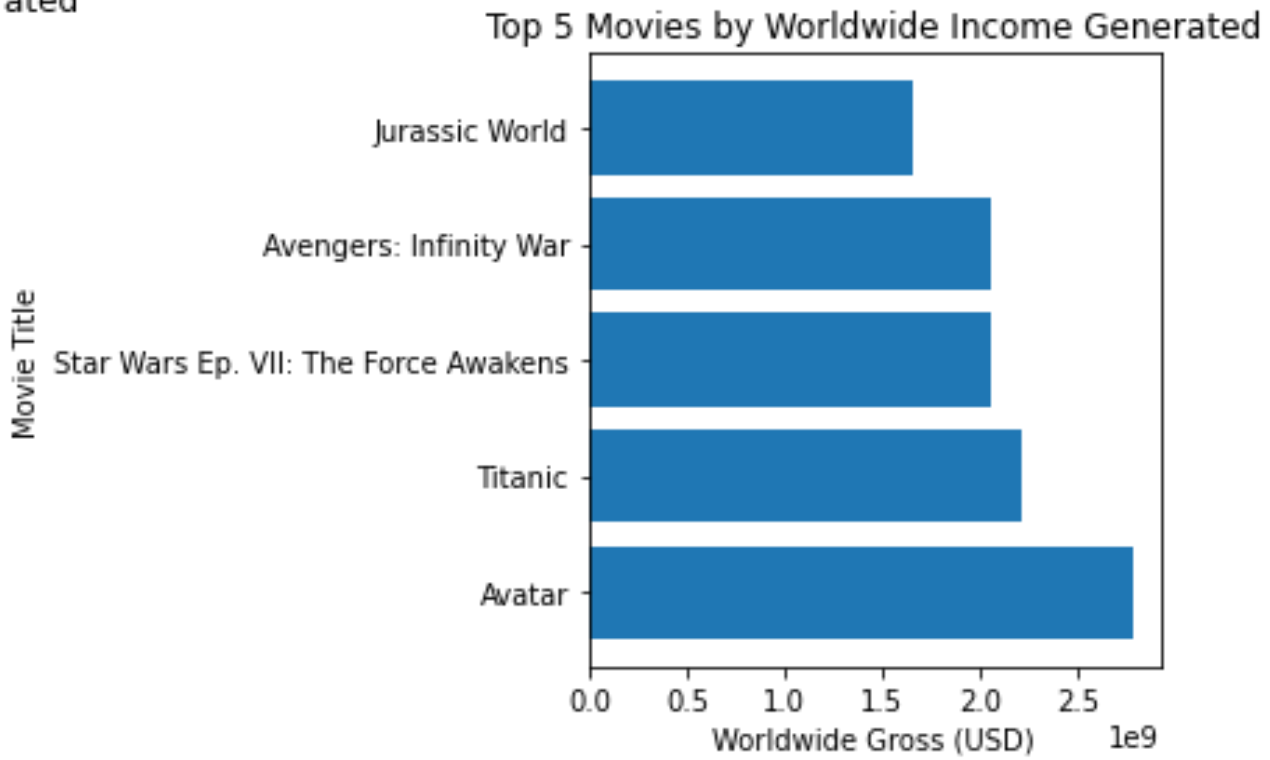
THE MOST VIEWED AND RATED MOVIE GENRES



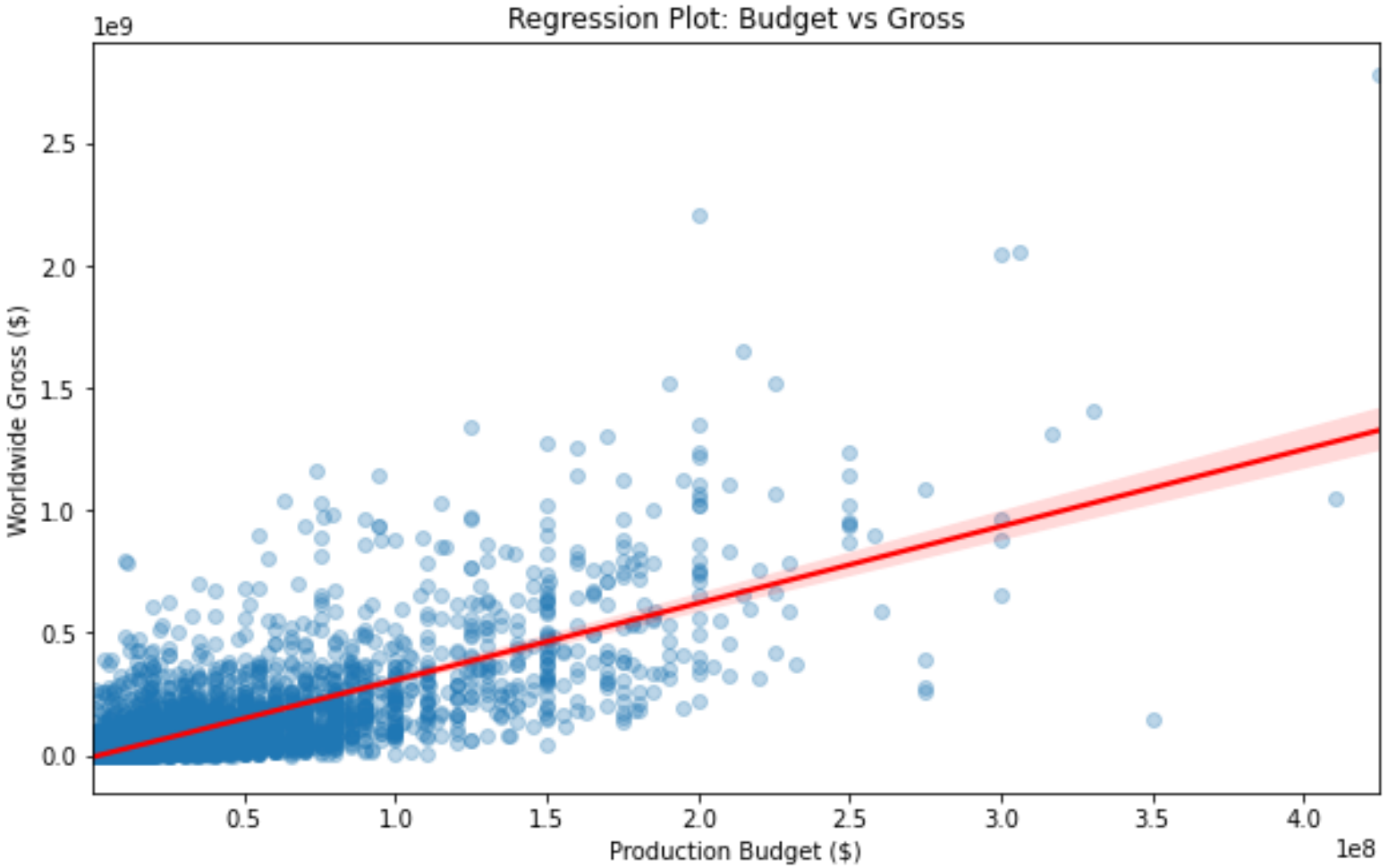


THE HIGHEST INCOME GENERATING MOVIES FROM THE NUMBERS AND BOX-OFFICE DF RESPECTIVELY





REGRESSION ANALYSIS BETWEEN PRODUCTION COSTS AND INCOME GENERATION



BUSINESS RECOMMENDATIONS

- *Recalling our business goal*:
  - Which movie genres are the most commercially successful and financially viable?
  - ✓ We recommend documentaries since they are commonly produced, had one of the highest total votes and audience ratings.

LIMITATIONS

1. The rotten tomatoes movie source - rating column had both numerical and non-numerical grade ratings that made data analysis difficult.
2. The tmdb movie source - movie genres were coded under the ID column, and there was no references to use, to help us know which movie ID belonged to which movie genre.
3. We could not use IMDB for regression analysis as it didn't have the columns or data on production budgets or gross incomes for the movie titles. Similarly, Box office mojo lacked the production cost column and therefore was not used for regression analysis.
4. Genre classifications, during assessment of the production costs versus income generated per movie genres, were checked manually using IMDb https://www.imdb.com/list/ls564012879/ since there was no unique identifier for the numbers data frame that connects genre and budget data in our data frames.

NEXT STEPS

- We suggest that for future analysis:
  - Improvement of Genre Classification be done to address genre inconsistencies across datasets by creating a standardized genre mapping.
  - We also suggest Temporal Trends and Lifecycle Analysis of genre performance over time to identify emerging trends and declining popularity.


Thank you!
Any Questions?