DSC_PT08P2_Phase_2_Project_Grp_1_Final

Project Overview

Business Problem

Our company recognizes the growing trend among major industry players in producing original video content and intends to establish a new movie studio to capitalize on this opportunity. However, we currently lack the necessary expertise and insights into filmmaking. To effectively guide the studio's development, we have been assigned the task of investigating the types of films that are currently excelling at the box office. The challenge lies in not only identifying these successful film trends but also converting this information into practical recommendations that the studio's leadership can use to make informed decisions about future film productions. Failure to do so may result in misguided investments and an inability to compete in the market.

Business Objectives

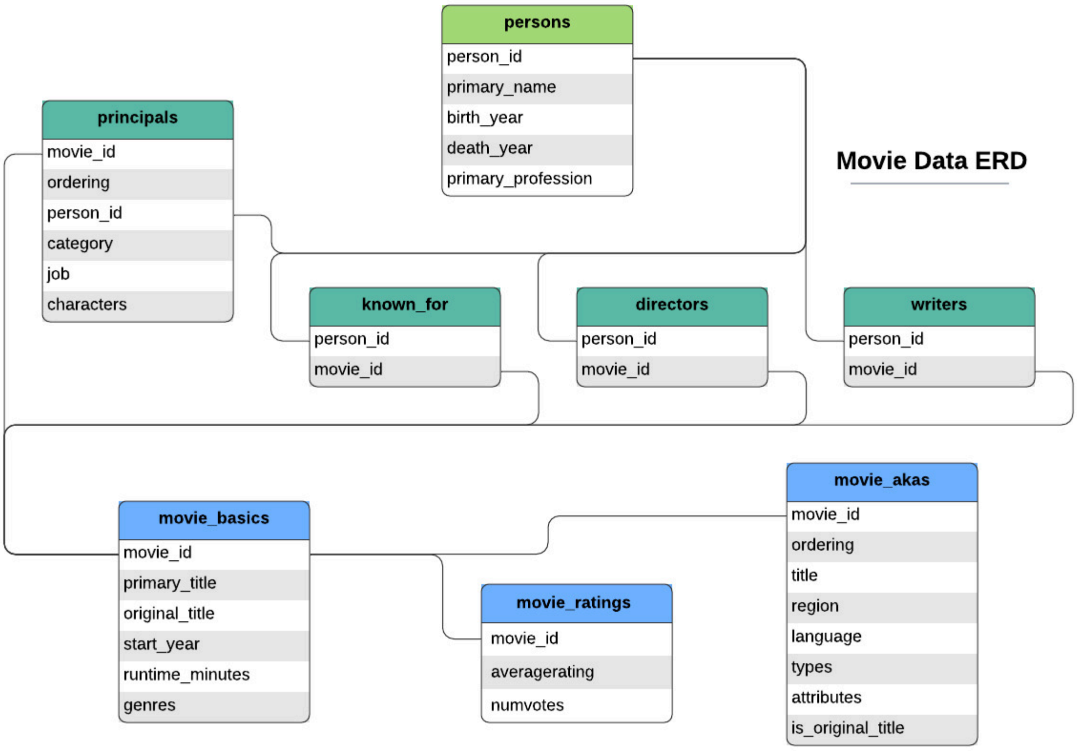
- What are the trends in movie release dates and what impact do they have on revenue?
- Is there a relationship between production budget and profitability and ROI of a movie? i.e Does a higher production budget automatically result to a higher profitability and vice versa.
- Determine which film genres are currently performing best at the box office.
- Look at trends in audience preferences and box office performance over recent years.
- Understand Audience Demographics by identifying the demographics of audiences for top-performing genres.
- Evaluate production costs associated with different genres.
- Understand the relationship between production cost and revenue streams of movies
- Evaluate the if movie rating affect production costs.
- Provide actionable insights for the new movie studio.

Sources of Data

In the folder `zippedData` are movie datasets from:

- Box Office Mojo (https://www.boxofficemojo.com/)
- IMDB (https://www.imdb.com/)
- Rotten Tomatoes (https://www.rottentomatoes.com/)
- TheMovieDB (https://www.themoviedb.org/)
- The Numbers (https://www.the-numbers.com/)

Because it was collected from various locations, the different files have different formats. Some are compressed CSV (comma-separated values) or TSV (tab-separated values) files that can be opened using spreadsheet software or `pd.read_csv` , while the data from IMDB is located in a SQLite database.



Note that the above diagram shows ONLY the IMDB data. You will need to look carefully at the features to figure out how the IMDB data relates to the other provided data files.

Data

The data used for this project include:

- movie_gross ()
- movie_budget ()
- movie_info ()
- im.db ()

Data Understanding

The Libraries

The files are built-in and can only be accessed once they are imported into the library. The libraries used in the project are listed below:

import pandas as pd import numpy as np import sqlite3 import seaborn as sns import matplotlib.pyplot as plt import matplotlib_inline import matplotlib.image as mpimg import statsmodels.api as sm import zipfile import os from scipy import stats from scipy.stats import pearsonr, ttest_ind from statsmodels.stats.diagnostic import linear_rainbow sns.set_style()

Loading of the data

The data was loaded in two ways: from CSV files and from a database. The CSV files were located in the unzipped directory, where ./Data/dsc-phase-2-project-v3-main.zip was extracted. Within this directory, the zippedData folder contained various datasets.

Additionally, data was loaded directly from the database.

- movie_budget = pd.read_csv("unzipped/dsc-phase-2-project-v3-main/zippedData/tn.movie_budgets.csv.gz")
- movie_gross = pd.read_csv("unzipped/dsc-phase-2-project-v3-main/zippedData/bom.movie_gross.csv.gz")
- movie_info = pd.read_csv("unzipped/dsc-phase-2-project-v3-main/zippedData/rt.movie_info.tsv.gz", sep='\t', encoding='ISO-8859-1')

Description of data

- Movie_info
  - All the ID column details are 1560 with none missing, datatypes are integer(1) and object(11)
  - The number of rows and columns are 1560 , 12 in number.
  - The mean of the unique data in movie_info was 1007.30.
  - The standard deviation of the ID column in movie_info was 579.16. This shows how data deviates from the mean.
  - The minimum value of ID is 1 with the maximum being 2000
- Movie_budget
  - The data has 5782 rows with none with missing values.
  - The data has 5 columns with text type of data while the ID column has integer data type.
  - ID column in Movie_budget has 5782 unique values counted.
  - The mean of the unique data in movie_budget was 50.37.
  - The standard deviation of the ID column in movie_budget was 28.82. This shows how data deviates from the mean.
  - The minimum value of ID is 1 with the maximum being 100.
- Movie_gross
  - The data has 3387 rows with title column having no missing values. Title is the unique identifier in the dataset.
  - The data has 3 columns(title,studio,foreign_gross) with text type of data, the domestic_gross and year columns have floats and integer data types, respectively.
  - The standard deviation of domestic gross for the movies was about 67 million dollars, showing a large spread
  - The 75th percentile of the domestic gross was 2.79 million dollars with the median at 1.2 million dollars. This shows that the 75% of the movies have a domestic gross less than 2.79 million dollars. 25% of the movies have a domestic gross of less than 1.2 million dollars.
  - The minimum domestic gross value for the movies was 100 dollars , the maximum domestic gross value stood at 936.7 million dollars.

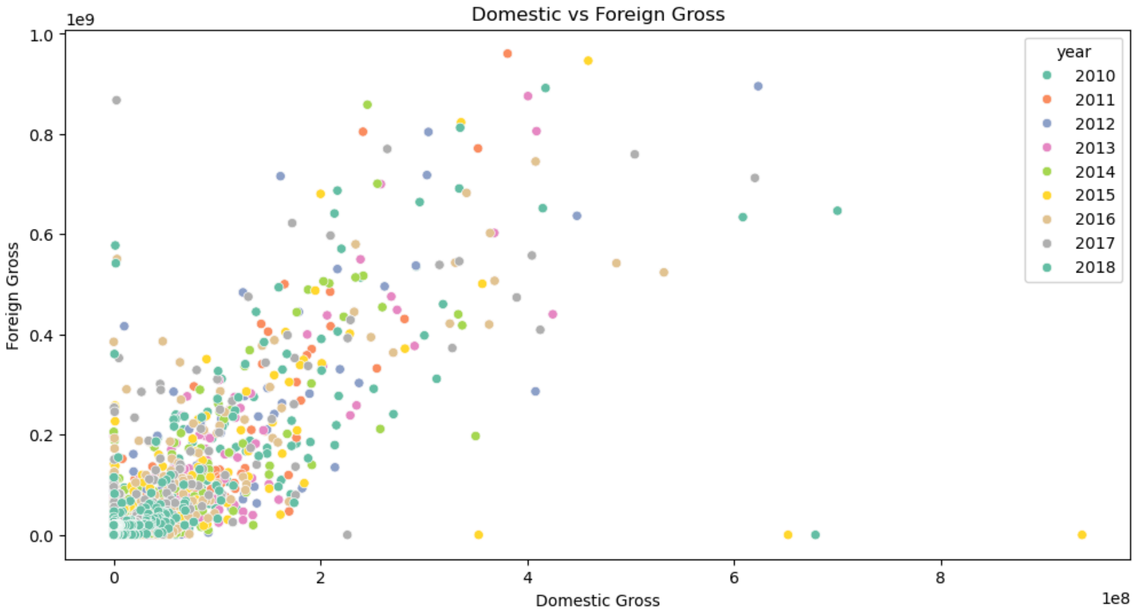
Cleaning of the data

- Movie_info
  - Finding the sum of the missing values, calculating its percentage and dropping columns with missing values above 50%.
  - Studio, box_office and currency columns have the highest number of missing values, all above 50% of the values.
  - The categorical columns the missing values are filled using the mode

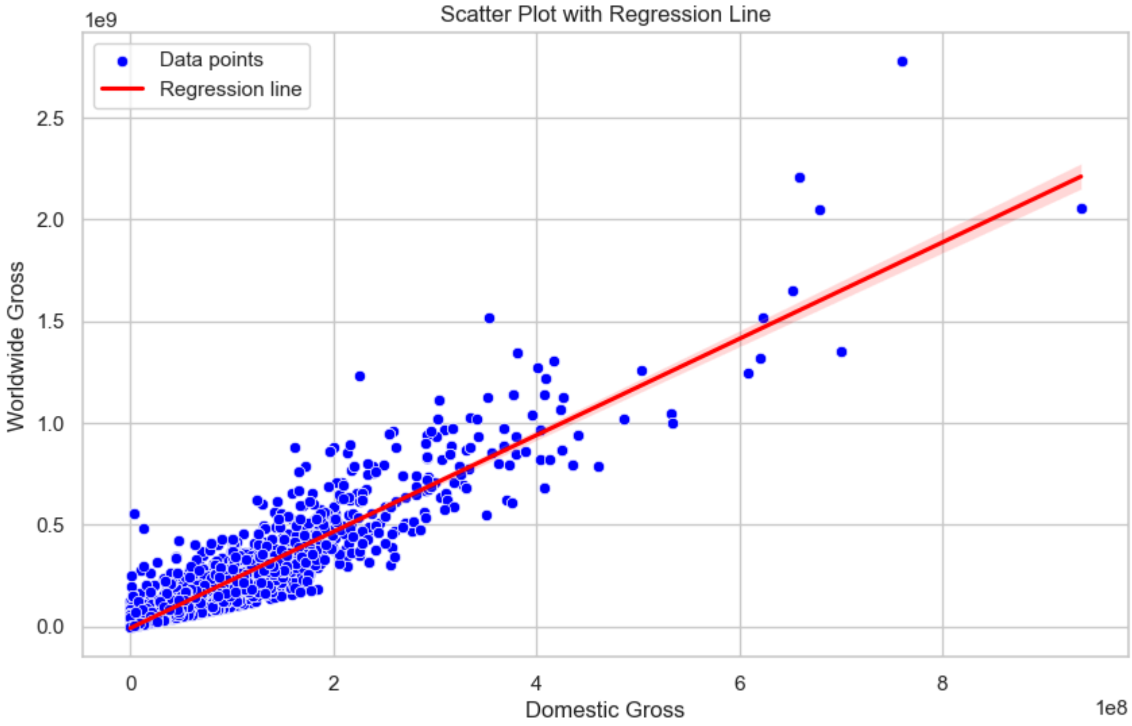
  - Splitting of the genre list to refer to specific genres, dropping the genre column and renaming of the genre list column to genre
  - Convert of the dates to Date Time Format
- Movie_budget
  - The release date being converted to DateTime Format
  - The removal of the dollar sign from the numerical columns
- Movie_gross
  - The commas from the 'foreign_gross' column were removed and the column converted to numeric
  - Convert to foreign_gross to numeric
  - The rows with missing values in 'studio' and 'domestic_gross' columns were removed
  - Replacing the Missing Values in 'foreign_gross' with the median
- IMDB Database
  - The extraction of the column names from the movie_basics table
  - Joining the Movie_basics table and the Movie_ratings table using Movie id which is the primary key.
  - Convert to a dataframe
  - Checking for the missing values, tre runtime and genres column had missing values which were dropped.
  - Removed the columns with duplicate names.
- A merge between the clean Movie_info dataset with Movie_budget dataset

Visualization(s)

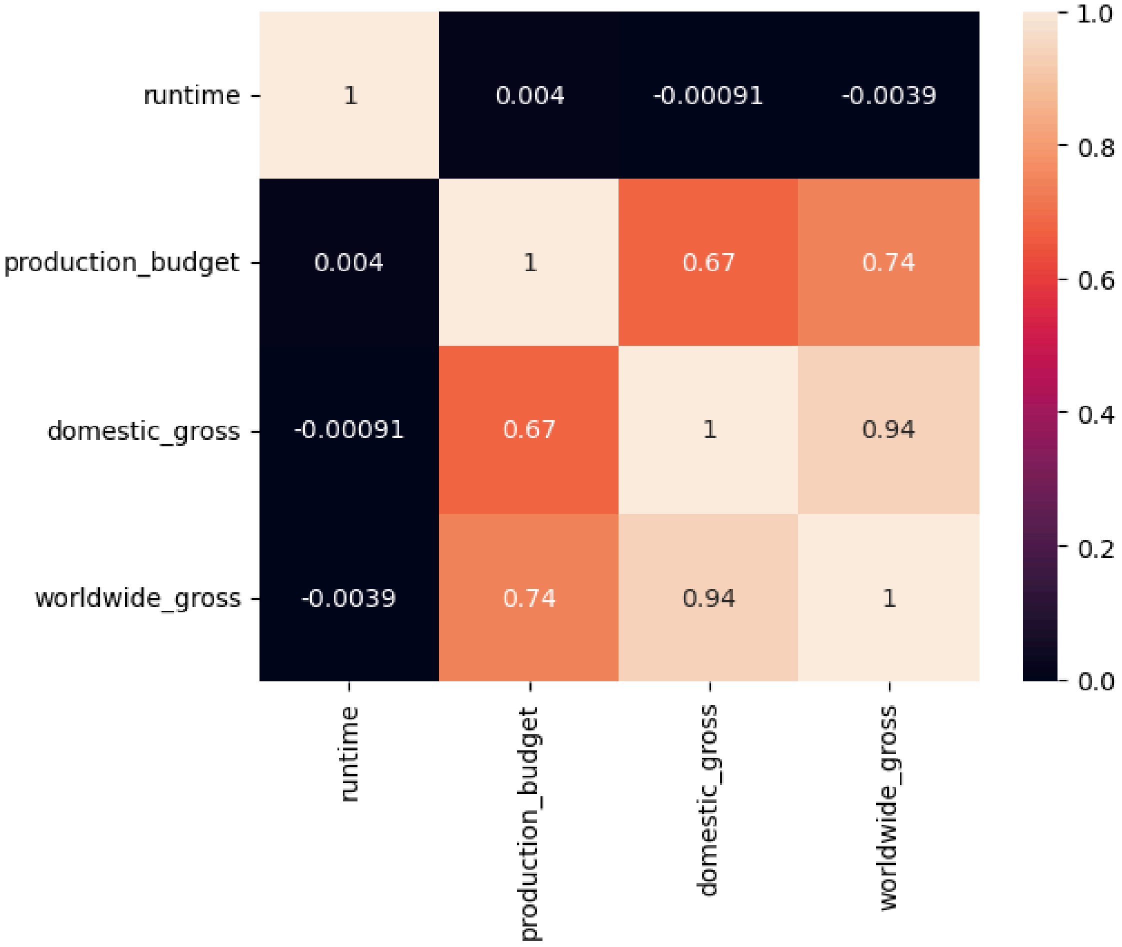
1. Relationship between Domestic Gross and Foreign Gross



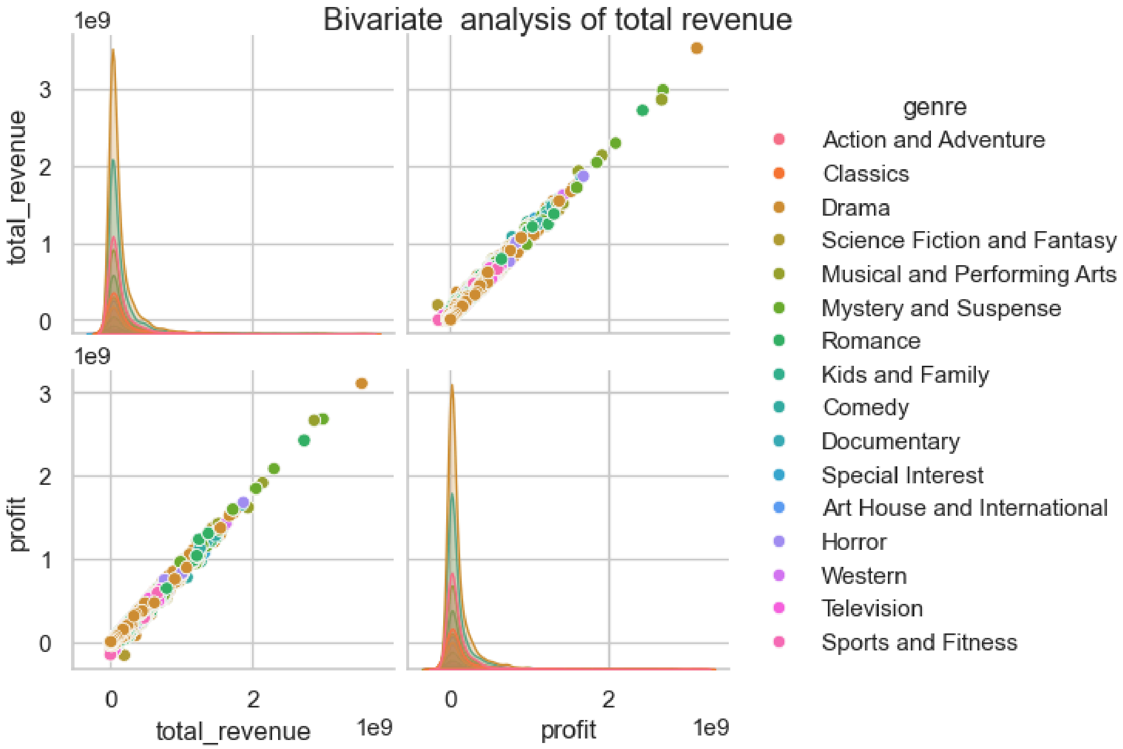
2. Scatter Plot with Regression Line of Domestic Gross and Worldwide Gross



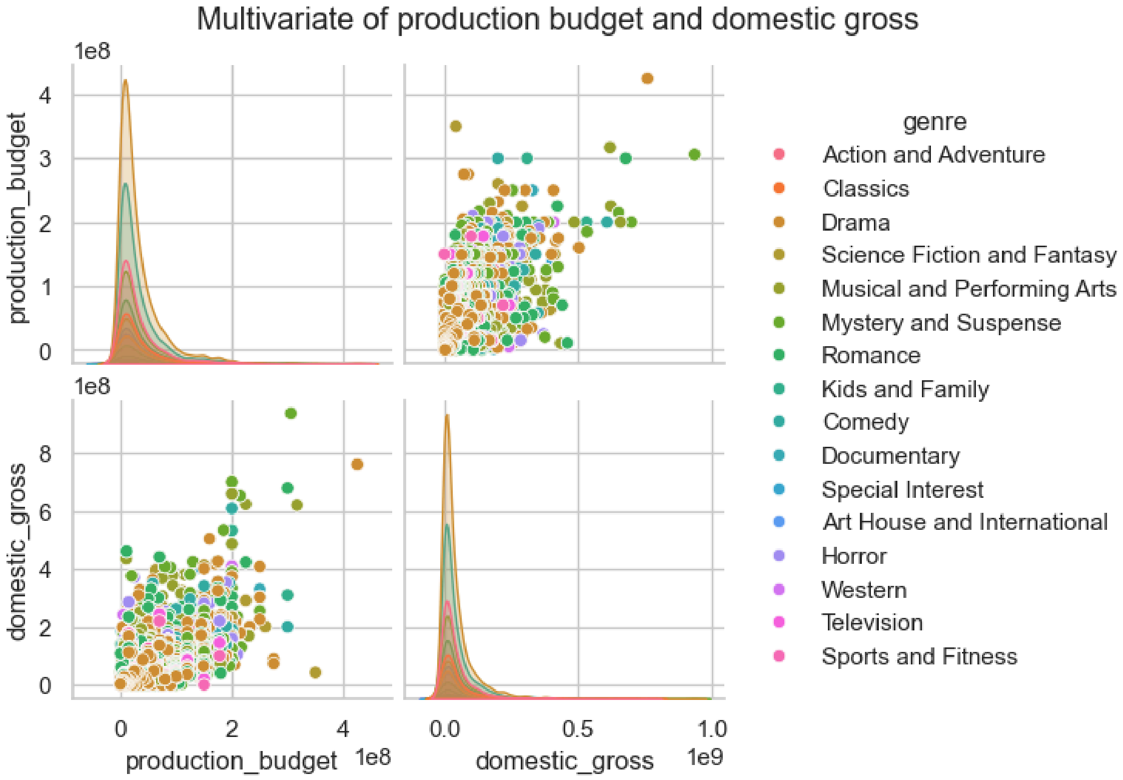
3. Heatmap Showing Correlation of the Numerical variables



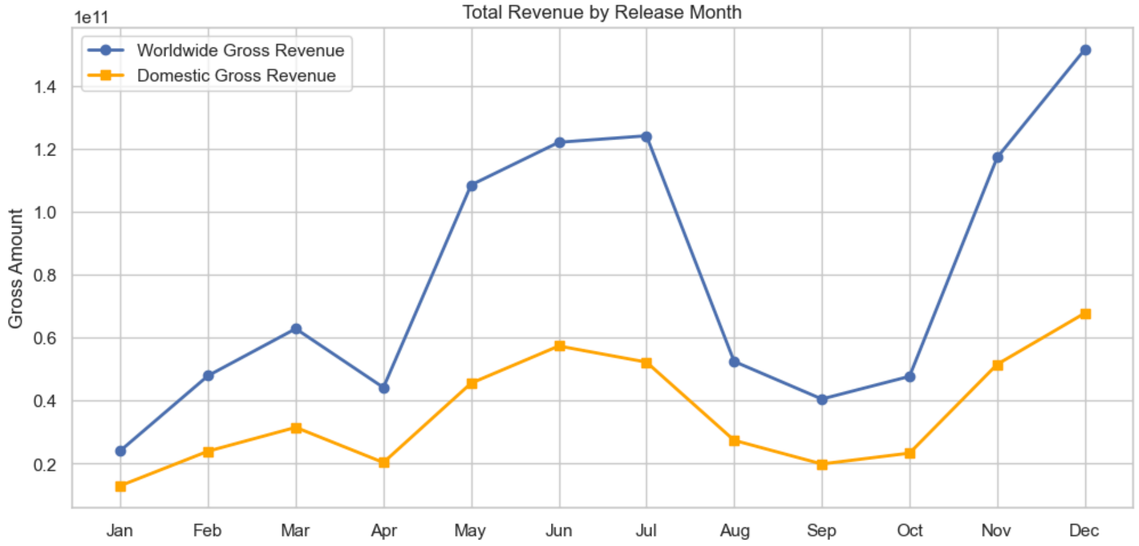
4. Bivariate Analysis of Total Revenue



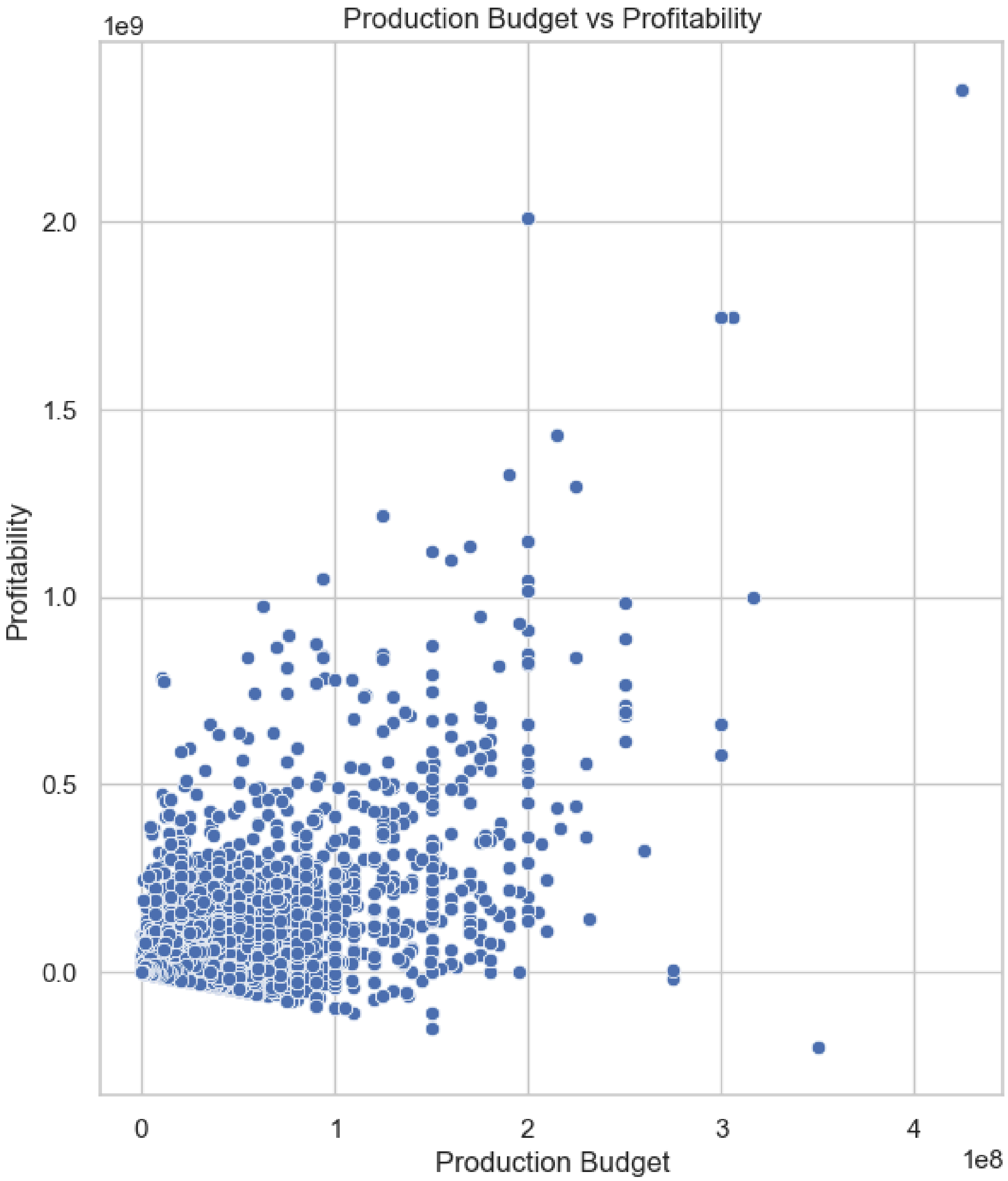
5. Multivariate Analysis of Production budget and Domestic Gross



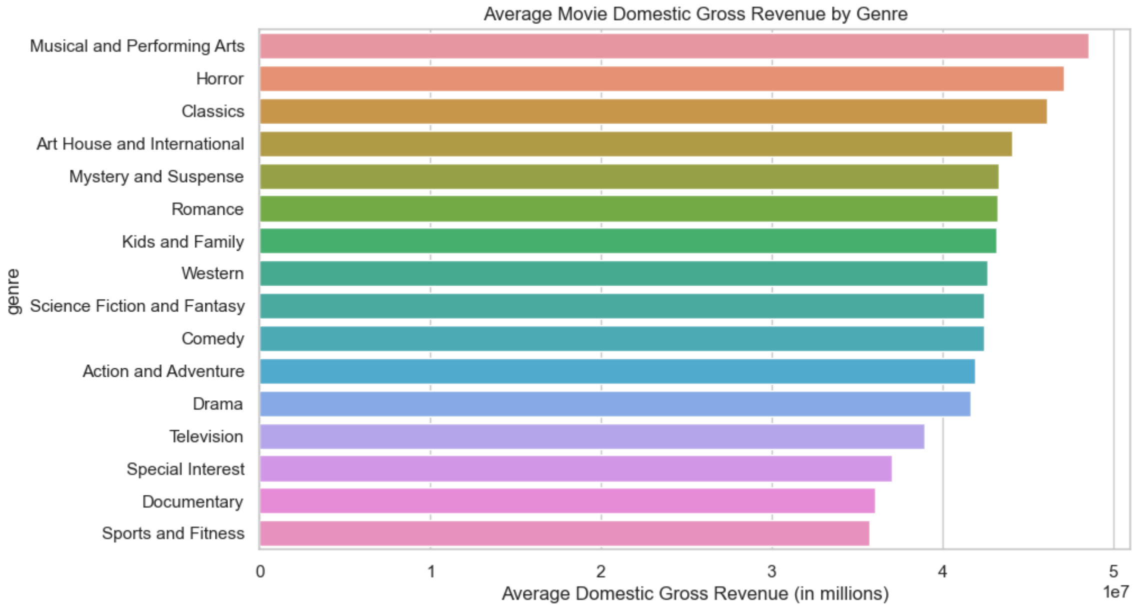
6. Analysis of Total Revenue by Movie Release Month



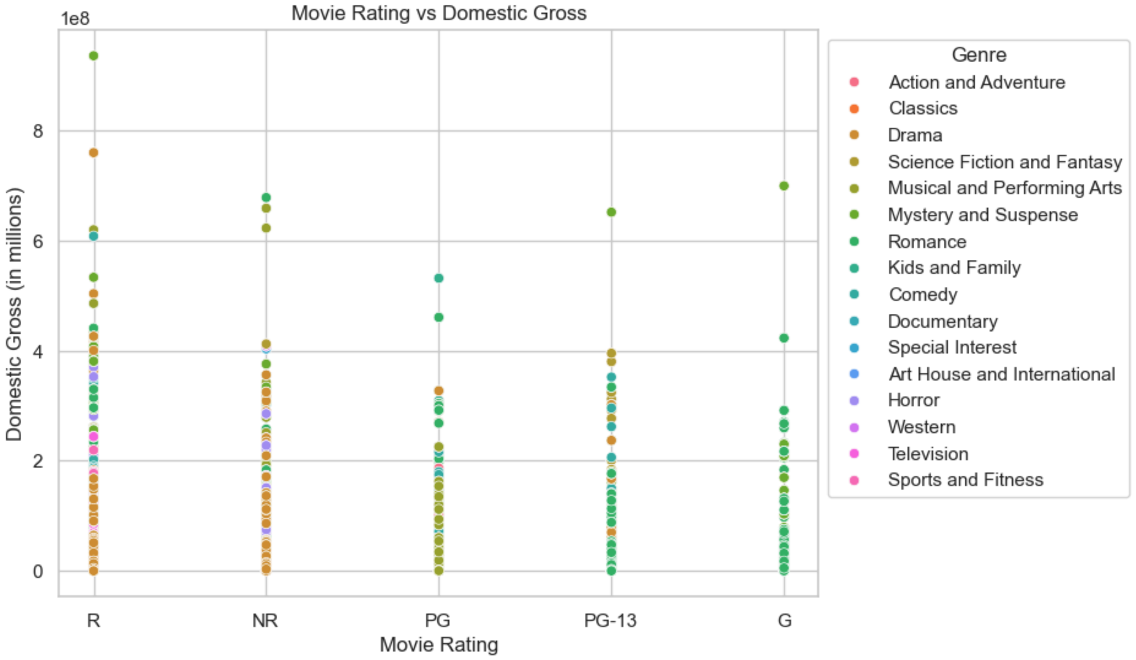
7. Analysis of Production Budget vs Profitability



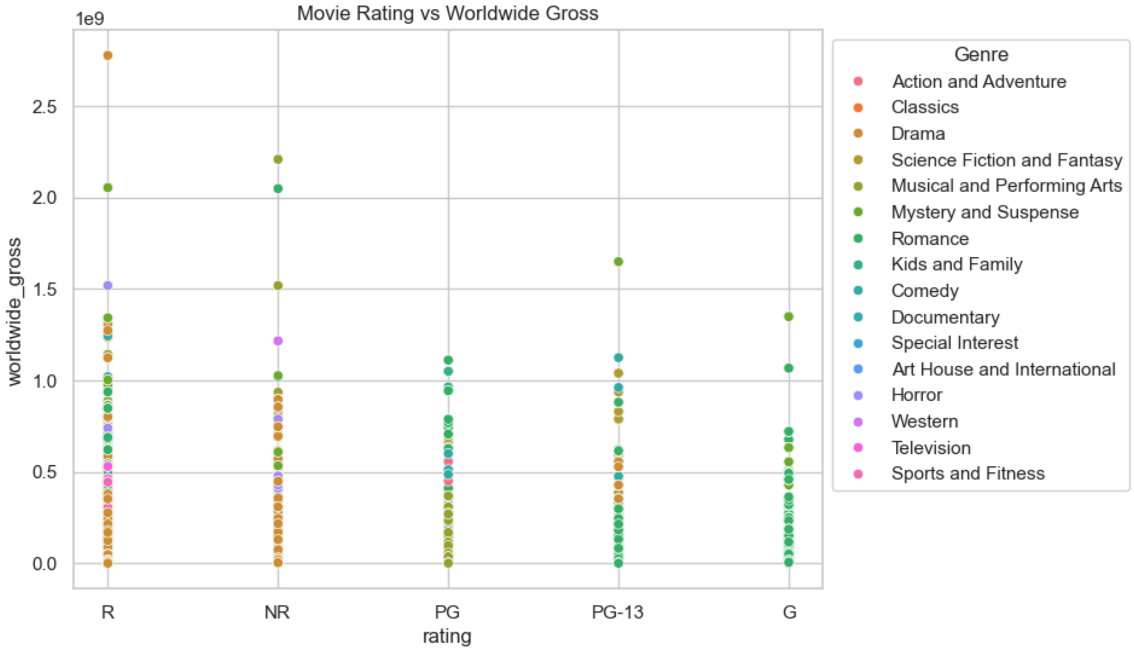
8. Average Movie Domestic Revenue by Genre



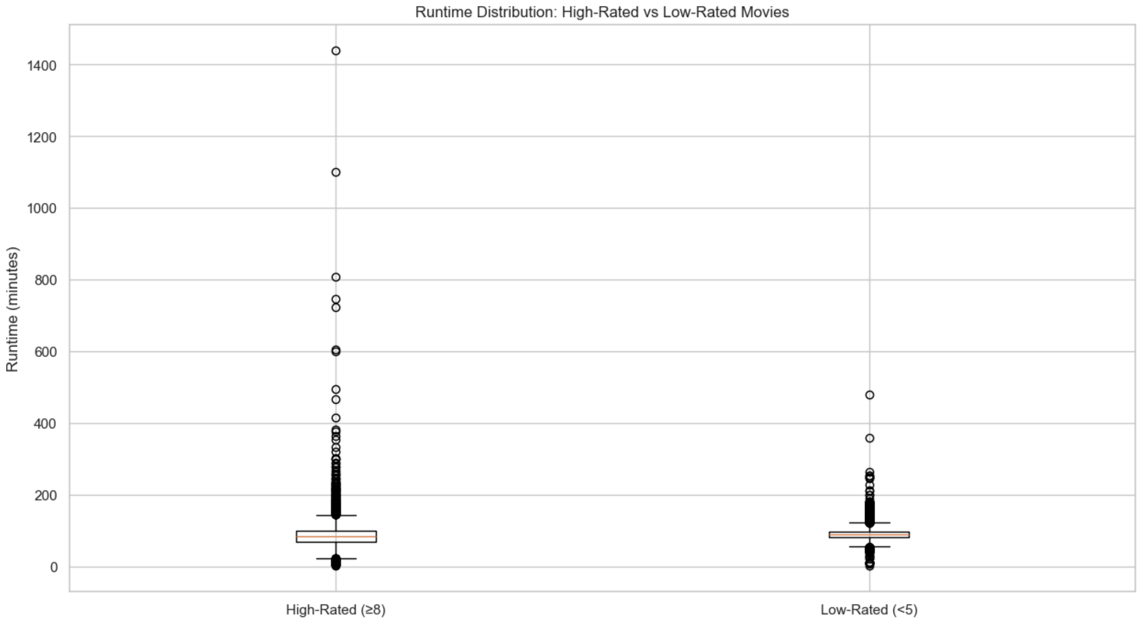
9. Analysis of Movie Rating vs Domestic Gross



10. Analysis of Movie Rating vs Worldwide Gross



11. Analysis of Movie Runtime and Movie Ratings



Conclusion(s)

1. Correlation Analysis

- Correlation between the foreign gross and domestic gross had a strong positive. This implies that an increase in domestic gross could also be reflected with an increase in foreign gross.

- Domestic gross and the year have a weak correlation, showing a weak relationship. Equally foreign gross and the year have a weak relationship. For predictive modelling, the positive correlation between the domestic gross and foreign gross will be an important relationship to consider.

- Production budget has a positive correlation with both domestic gross and worldwide gross, with the later being stronger at 0.7392. This relationship implies that production budgets of movies have a relatively strong and positive relationship with both domestic and gross earning.

- Runtime has a weak relationship with production budget, domestic and worldwide gross. The correlation between runtime and production budget was 0.0040. Equally, the correlation between runtime with both domestic and worldwide gross was negative at -0.000906 and -0.003864, depicting a very weak relationship hence the revenue values as a function of runtime cannot be modeled using a linear model.

- Similarly, runtime as a function of movie production budget cannot be accurately modeled using a linear model because the p-value of 4.13e-09 obtained is significantly smaller than the usual significance level of 0.05.

- There is a positive linear relationship between domestic gross and foreign gross as well as domestic gross and worldwide gross. Hence a linear model is suitable to model domestic gross, foreign gross and worldwide gross revenues for the movies.

- The higher production budgets generally correlate with higher total revenue, most movies fall within a lower budget and revenue range.
- There is a moderate positive correlation between production budget and profitability, but a very weak negative correlation between production budget and ROI.

2. Genre Revenue Trends in Domestic and Worldwide Markets

- Musical and Performing Arts consistently outperforms other genres in both domestic and worldwide gross revenue.

- Horror and Science Fiction and Fantasy also show strong performance in both domestic and worldwide markets.

- Classics and Documentary have a higher average domestic gross compared to their worldwide performance.

- Certain genres, such as Drama (orange) and Science Fiction and Fantasy (green), are more likely to achieve higher total revenue and profit, as these points appear more frequently among the higher values

3. Production Budget and Genre Analysis

- Genres such as Action and Adventure and Science Fiction and Fantasy are associated with both high production budgets

- In contrast, genres like Documentary and Special Interest tend to have smaller budgets and lower domestic revenue

- Most movies have relatively low production budgets and domestic grosses, with only a few movies achieving very high values in either category.

4. Insights on Movie Ratings

- High-rated movies(with ratings >=8) tend to be slightly shorter on average, but they're actually more likely to be over 2 hours long with a percentage of 9.9% compared to 7.9% for low rated movies (with ratings < 5)
- The standard deviation of High rated movies of 39.04 is significant. The minimum movie runtime was 4 minutes long with the maximum at 1440. The 1440 minutes suggests existence of outliers in the dataset since the average average runtime for high rated movies was 88.62 minutes.
- The standard deviation of low rated movies of 19.04 is significant, implying big variations in the run time. The minimum movie runtime was 4 minutes long with the maximum at 480. The 480 minutes suggests existence of outliers in the dataset. The average runtime for low rated movies was 92.71 minutes.

Recommendation(s)

- We are recommending to the company to prioritize releasing more movies during the months of June, July and December. December is the most profitable month and therefore aiming at this timeframe could maximize on box office revenue.
- We recommend to the company to utilize lower budget or experimental films in months

```
In [ ]:
```