

MOVIE ANALYSIS DATA REPORT
(Analyzing movie Profitability using SQL and Python)

Team Members

Dean Mutie

Winnie Njoroge

Michelle Mwende

Laban Leploote

Alice Wangui

Date: 5th November 2025

Movie Analysis

1.0 Introduction

Business Background / Overview

By building its own film studio, the company plans to enter the film production industry. The management team recognises the potential for substantial returns in producing original video content, given the rapid growth of streaming services and the increasing profitability of popular films. But the corporation has never made a movie before, and it doesn't know which movie genres do best at the box office.

The corporation needs data-driven insights into movie performance trends, including the relationships among production budgets, genres, ratings, and revenue, to make well-informed investment decisions. By comprehending these components, the business will be able to better organise its production schedule, lower financial risks, and increase box office revenue.

The business must first comprehend the strategic patterns that propel contemporary movie achievement in order to properly enter such a competitive market. The film industry has become more and more data-driven, with big studios mainly depending on analytics to identify the kind of films that draw sizable audiences, earn high ratings, and produce substantial global profits. These days, a movie's commercial success is greatly influenced by elements like audience preferences, genre appeals, production costs, and company prestige. New studios face greater financial risk and a significant degree of uncertainty if these dynamics are not well understood. In order to position itself competitively and create films with the greatest potential for profitability, the corporation must examine past trends and performance metrics.

Proposed Solution

In order to identify important variables affecting box office performance, this research suggests a thorough data analysis of current film datasets. Production budgets, profits, audience ratings, vote counts, genres, and studio performance will all be examined in the analysis. To determine which aspects of films result in more profitability, visualisations and insights will be employed. The results will serve as a basis for future modelling and strategic decision-making regarding the kinds of films to produce and invest in, even if no predictive model is currently in use.

A set of practical insights that will direct the company's new film studio in choosing lucrative genres, establishing reasonable production budgets, and coordinating creative endeavours with consumer preferences is the anticipated result of this investigation. The results will reduce the financial risks involved in making film investment decisions and aid in the creation of a data-driven production strategy.

1.1 Problem Statement

The corporation intends to open a new film studio, but it lacks data-driven insights into what makes films popular at the box office and adequate industry experience. The studio runs the danger of making expensive investments in films that might not draw viewers or yield sufficient profits if it does not comprehend the critical elements that affect profitability and audience engagement, such as production budgets, genres, and ratings. This ambiguity emphasises the necessity of a methodical investigation of the data currently available in the film business in order to guide strategic budgetary and production choices.

1.2 Objectives

- To determine the relationship between the production budget and the profitability
- To determine the impact of ratings and vote counts on the gross revenue
- To explore trends in movie performance over the years
- To examine the performance of existing studios
- To assess the combined effect of vote count, rating, and budget on worldwide revenue

2.0 Data Understanding

The purpose of this step is to familiarize ourselves with the datasets, assess their structure, and identify potential quality issues that may affect later analysis or modeling.

2.1 Data Sources and Collection

The data used in this project was sourced from publicly available raw data accessed through the GitHub repository: (<https://github.com/learn-co-curriculum/dsc-phase-2-project-v3.git>.)

The datasets were in different formats, some as CSV(comma-separated values), others as TSV(tab-separated values), and one as a SQLite database. These datasets were systematically filtered into a comprehensive database named **movies_Data.db**. The **movies_Data.db** comprises four core tables: **bom_movie_gross**, **movie_basics**, **movie_ratings**, and **tn_movie_budgets**.

2.2 Dataset Overview

The database comprises four primary tables: **bom_movie_gross**, **movie_basics**, **movie_ratings**, and **tn_movie_budgets**. The summary of each table's structure is as follows

- **bom_movie_gross table**

Contains box office revenue data.

It has 3387 records and 5 attributes: title, studio, domestic_gross, foreign_gross, and year. The table is largely complete, although there are 5 missing entries in the studio column, 28 missing entries in the domestic_gross column, and 1350 missing entries in the foreign_gross column. The foreign_gross column exhibits significant incompleteness and is stored as an object rather than a numeric type.

- **movie_basics table**

Provides fundamental metadata about movies.

It has 146144 records and 6 attributes: movie_id, primary_title, original_title, start_year, runtime_minutes, and genres. There are 21 missing values in the original_title column, 31739 missing values in the runtime_minutes column and 5408 missing values in the genres column.

- **movie_ratings table**

Contains audience and critic rating data derived from IMDb. It has 73856 records and 3 attributes: movie_id, average rating, and numvotes. The table is relatively clean with no missing values.

- **tn_movie_budgets table**

Includes detailed financial data.

It has 5782 records and 6 attributes: id, release_date, movie, production_budget, domestic_gross, and worldwide_gross. Although the table is complete with no missing values, the columns production_budget, domestic_gross, and worldwide_gross are stored as an object type and contain non-numeric symbols such as dollar signs and commas. The values must be converted into a numerical format for further analysis.

2.3 Data Quality Checks

To ensure reliability and consistency, several data quality checks were conducted.

Missing Values

Missing records in critical fields such as original_title and runtime_minutes were dropped. Categorical columns like studio and genres were imputed with the placeholder 'Unknown'. Numerical columns such as domestic_gross and foreign_gross were imputed using the median to maintain distribution integrity.

Data Type and Format Corrections

production_budget, domestic_gross, and worldwide_gross were cleaned by removing commas and dollar signs and converted to numeric for accurate computation.

3.0 Data Preparation

The goal of data preparation was to clean, transform, and structure the dataset in a way that allows for meaningful analysis and ensures the data is suitable for modelling.

3.1 Data Selection and Integration

After looking over the available datasets, we determined which pertinent columns were required for our study. In order to produce a comprehensive dataset, we concatenated datasets that contained related information based on important identifiers (such as movie ID). Irrelevant or redundant columns were dropped to streamline the analysis. We thoroughly examined every table to comprehend its structure, data types, and important columns prior to combining. This made it easier for us to decide which columns to utilise as merging keys:

Check Table information .info()

We were able to comprehend the data types, identify important columns, and identify any possible problems, including missing values or irrelevant columns, by examining the table information (.info()) and previewing the column names.

Inspect columns to find out which keys will be used during the merge

Each dataset's first rows and column names were examined using the table_columns function. This assisted us in determining which columns, like movie_id, primary_title, and start_year, could function as merge keys for merging the databases. We were able to accurately integrate the datasets and prevent repetition because we understood the column names and structure.

This method allowed for precise and significant merges in later stages by ensuring that we completely comprehended the structure of our datasets before moving further.

Handling Missing Values, Duplicates, and Errors

Missing values were identified in key columns. Depending on the type of data, they were handled by either:

- Imputation (e.g., using the median for numerical columns or a placeholder like 'Unknown' for categorical data).
- Dropping rows if the missing data was too extensive or non-critical.

There were no duplicates in our dataset. Data errors, such as impossible values or out-of-range numbers, were corrected.

According to the dataset's analysis of missing values, the **runtime_minutes** column had **31,739** missing entries, **genres** had **5,408** missing values, **domestic_gross** had **28** missing entries, and **foreign_gross** had **1,350** missing values.

Verifying Data Types

Production_budget, domestic_gross, worldwide_gross, and foreign_gross were among the numerical columns in the datasets that were kept as strings with dollar signs and commas. The formatting characters were eliminated, and the data were changed to numeric (float) types in order to get these columns ready for examination. This guarantees the accurate analysis and utilisation of all financial data in computations like profit, budget comparisons, and visualisations.

3.2 Feature Engineering

Additional features were created to enrich the dataset and support deeper insights.

Merging

In this section, we merged the tables and formed a table named “**final_merged**” using the SQL Queries.

We merged movie_basics and movie_ratings using the primary key movie_id.

The movie_id primary key was used to integrate the movie_basics and movie_ratings databases. The merged_movie_ratings dataframe, which was created by this merging, compiles all of the crucial data required for analysis. In order to provide a comprehensive and uniform dataset for future research, every movie entry now contains both its basic characteristics (such as title, year, and genre) and the rating metrics that go along with them.

We merged the movie_gross table with the merged_movie_ratings using an inner join.

We combined the movie_gross table with the previously combined merged_movie_ratings dataset in order to compile all pertinent movie data. To make sure that only films that were present in both datasets were included, we employed an inner join on the movie title and release year. In this step, the dataset with basic movie facts and ratings was supplemented with financial data (domestic and foreign gross).

we finally merged movie_budget to the merged_rg table above using an inner join

We used an inner join to combine the movie_budget table with the merged_rg dataset in order to add more production-related financial data to the dataset. The movie title (primary_title in merged_rg and movie in movie_budget) underwent the merge. To prevent duplication, just the relevant movie_budget columns—production_budget, domestic_gross, and worldwide_gross were included.

Adding New Columns

More computed columns were added to the dataset to improve it and facilitate extra analysis. A new field called total_gross was added to the movie_budget and movie_gross databases. This column shows the total revenue from both domestic and foreign markets, making it simpler to compare the overall performance of the film and to undertake further studies including total revenue.

4.0 Data Analysis and Visualization

After cleaning and preparing the datasets, we focus on Exploratory Data Analysis and Visualization to uncover meaningful patterns and relationships within the movie_Data.db.

The analysis explores how factors such as budget, revenue, ratings, genres, and studios influence the overall movie performance.

Using visual exploration and statistical summaries, the data is examined through univariate, bivariate, and multivariate analyses to identify key trends, correlations, and insights that guide further modeling and interpretation.

4.1 Univariate Analysis

Univariate Analysis was used to understand the distribution and behavior of individual variables relevant to the movie performance.

Distribution of prevalent genres

According to the dataset, a few genres dominated film production between 2010 and 2018. The most common genres were drama(26.8%), documentary(23.8%), and comedy(13.0%), which accounted for more than 60% of all films produced during this time period.

To visualize this distribution, Figure 1 represents a pie chart illustrating the proportion of the 15 genres, highlighting the significant dominance of Drama, Documentary, and Comedy film production trends.

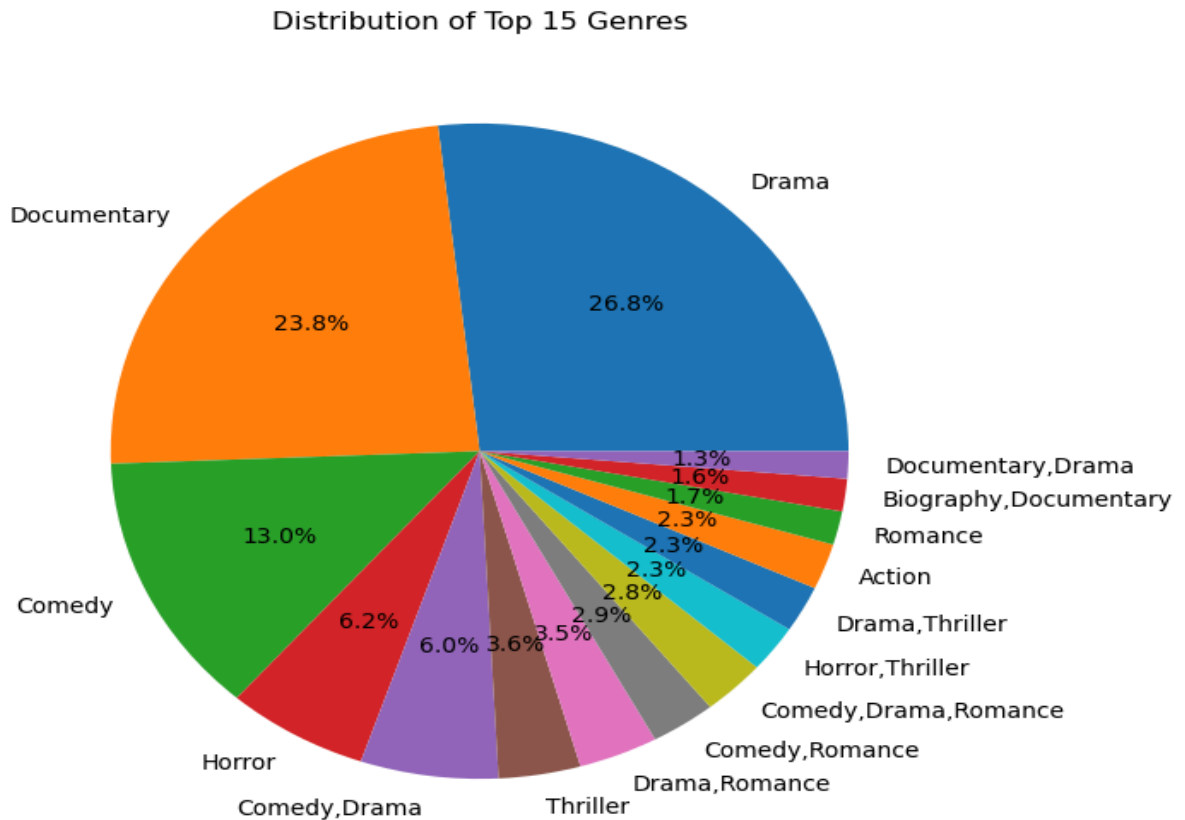


Figure 1: Distribution of prevalent genres

4.2 Bivariate Analysis

Examining the top 10 best-performing studios

According to our chart, Disney (BV) tops the film business in overall gross revenue, closely followed by Fox, Universal(Uni), and Warner Bros. (WB). When compared to other studios in the dataset, these studios consistently produce noticeably larger box office returns.

Additionally, there is a discernible drop in revenue from BV to LG/S, suggesting that the top few studios control a sizable portion of the industry. This trend implies that while smaller studios contribute far less to overall box office performance, major studios profit from higher production budgets, stronger distribution networks, and well-established franchises.

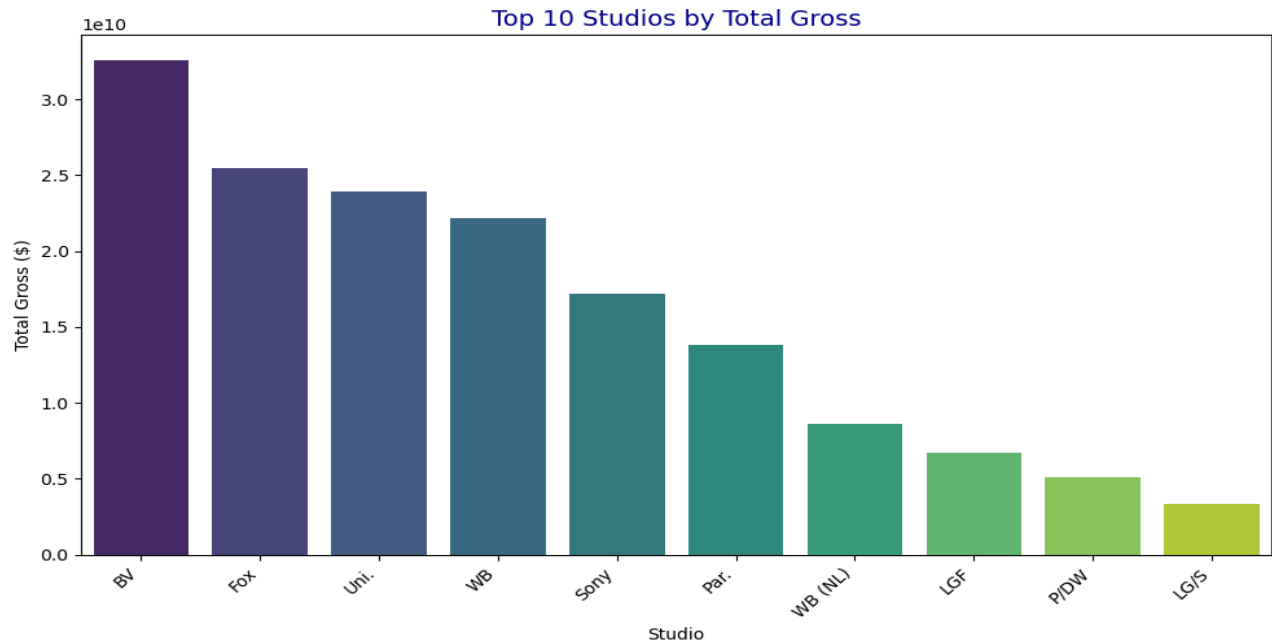


Figure 2: Examining the top 10 best-performing studios

Examining the Top 10 Highest-Grossing Genres.

According to the graph, Adventure-Animation-Comedy and Action-Adventure-Sci-Fi both make significantly more money at the box office than any other genre. The remaining genres, like Action-Adventure-Fantasy, Action-Adventure-Comedy, and Action-Adventure-Animation, make much less money after these top two. In general, the most popular genres are those that combine inventive aspects like science fiction, fantasy, animation, or humour with action or adventure.

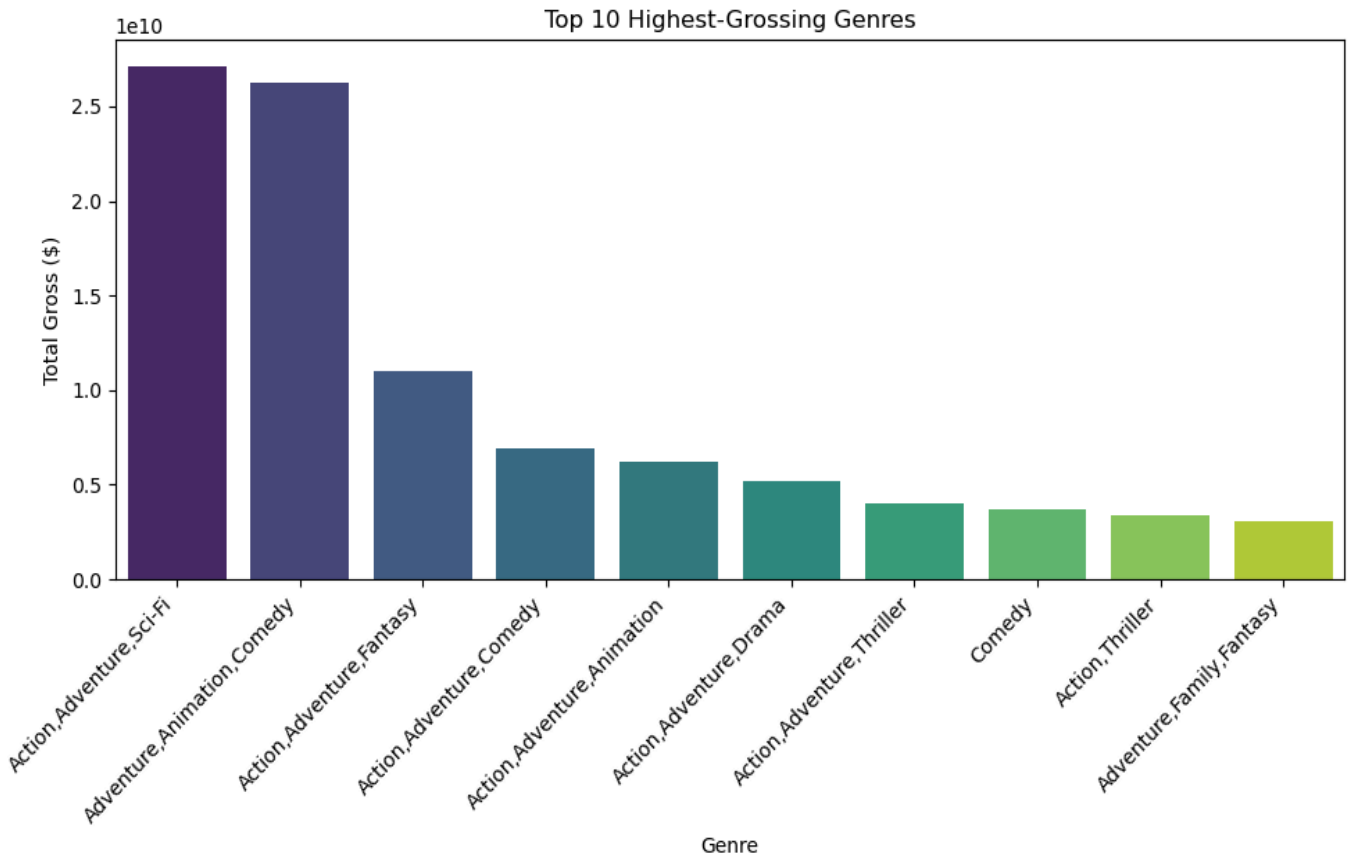


Figure 3: Examining the Top 10 Highest-Grossing Genres.

4.3 Multivariate Analysis

Evaluating the effect of Average rating, Vote count, and Production budget on worldwide revenue

Assessing the Impact of Average Rating, Vote Count, and Production Budget on Worldwide Revenue

As shown in Figure 4, there is a positive multivariate relationship between Average Rating, Vote Count, Production Budget, and Worldwide Revenue. The visualization shows that films with better audience ratings and engagement (shown by larger bubbles reflecting vote counts) tend to have higher global revenues

Furthermore, films with bigger production budgets (represented by lighter-colored bubbles) typically report higher earnings, demonstrating the importance of financial investment in

commercial performance. This pattern implies that film success is determined by the combined impact of critical reception, audience participation, and budget size, all of which contribute to higher global box office returns

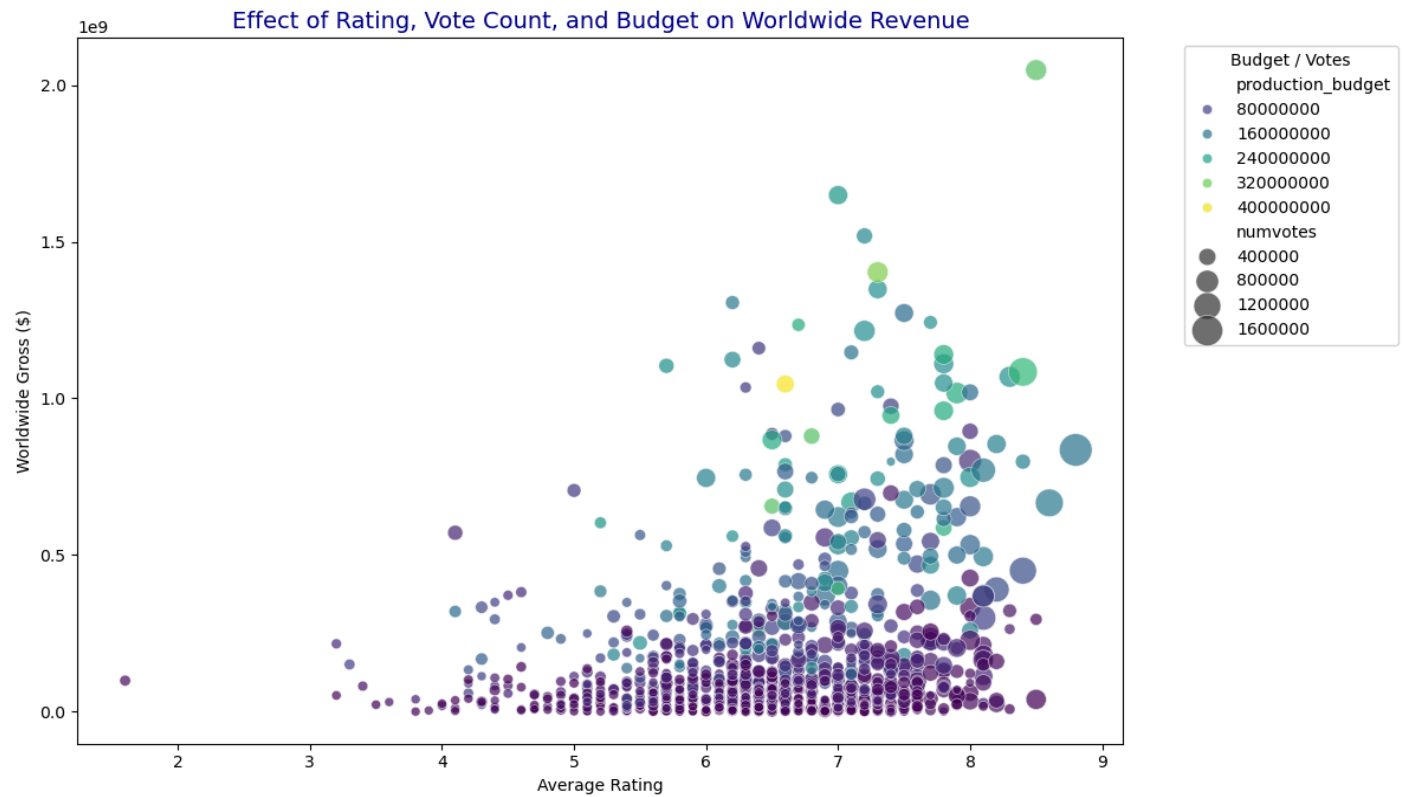


Figure 4: Evaluating the effect of Average rating, Vote count, and Production budget on worldwide revenue

5.0 Summary(Conclusion)

In conclusion, the analysis provides valuable insights into the key drivers that determine success in the film industry.

- **Relationship between Production Budget and Profitability**

The scatter plot revealed a positive correlation between production budget and profit, indicating that higher investment generally leads to higher profitability. However, the relationship was not strictly linear—some low-budget films achieved substantial profits, suggesting that strategic resource allocation and effective marketing can enhance returns even for smaller productions.

- **Impact of Ratings and Vote Counts on Gross Revenue**

Ratings and vote counts both show a weak but positive relationship with worldwide gross revenue. Movies with average ratings between 6 and 8 and higher vote counts tend to perform better financially. This suggests that audience engagement and perception (reflected through ratings and votes) influence box office outcomes, though other factors such as marketing, release timing, and franchise strength also play significant roles.

- **Trends in Movie Performance Over the Years**

Over time, there is an upward trend in overall movie revenue, reflecting growth in global audiences and expanded market reach. Recent years have seen a concentration of high-grossing films tied to established franchises and cinematic universes, indicating the growing dominance of blockbuster-driven strategies in the industry.

- **Performance of Existing Studio**

The analysis shows that the film industry is heavily concentrated among a few major studios — primarily Disney, Fox, Universal, and Warner Bros. Disney leads significantly, benefiting from its strong brand identity, diversified franchises (e.g., Marvel, Pixar), and large marketing budgets. Smaller studios contribute less to global earnings but may succeed through niche markets or independent storytelling

6.0 Recommendations

To remain competitive and profitable, the studio needs to make data-driven strategic decisions. The following proposals are centered on budget optimization, increasing audience engagement, investing in high-performing genres, diversifying the production process, and leveraging data to guide future investments.

- **Optimize Budget Allocation**

The studio should invest strategically in production budgets, ensuring a balance between scale and cost-efficiency. While high budgets can yield greater profits, careful financial planning and market research are essential to avoid overspending on underperforming projects.

- **Enhance Audience Engagement and Ratings**

Focusing on storytelling quality, originality, and audience satisfaction can improve ratings and boost long-term revenue. Encouraging early audience engagement through digital campaigns and leveraging social media reviews can amplify visibility and increase vote counts.

- **Capitalize on Genre and Franchise Trends**

The studio should continue to invest in high-performing genres such as Action, Adventure, Sci-Fi, and Animation, which have proven global appeal. Developing sequels or shared universes can sustain audience loyalty and generate consistent revenue streams.

- **Diversify Studio Strategies**

The studio should also consider targeting niche audiences, streaming platforms, or lower-budget productions with high creative value to remain competitive. Collaboration with larger distributors could also enhance exposure and profitability.

- **Leverage Data for Future Forecasting**

Regular performance tracking across key metrics like budget, ratings, votes, and studio output will help predict future trends and optimize investment decisions. Data-driven approaches can

improve profitability forecasting and reduce financial risk.