Otim135 /
group3-phase2-project

Code Issues Pull requests Actions Projects Wiki Security Insights Setting

End of Phase 2 Project for Group 3, Part Time Data Science Class

☆ 0 stars 0 forks 1 watching Branches Activity
Tags

Public repository

1 Branch 0 Tags Go to file t Go to file + Add file Code ···

otim-obote Modified some codes and erased redundant data 7f98571 · now

data	updates	yesterday
ipynbs	Finalized on the Final Notebook	21 minutes ago
.gitattributes	Added large zip file using Git LFS	4 days ago
.gitignore	Add im.db to .gitignore	3 days ago
Index.ipynb	Modified some codes and erased re...	now
Merged_index.ipynb	Renamed the index.ipynb to Merge...	3 hours ago
README.md	Update README.md	6 minutes ago
genre_financials_insights.csv	created a new csv file called genre_f...	13 hours ago

README

MOVIES INDUSTRY ANALYSIS

1.0 Overview

Our company sees all the big companies creating original content and wants to join the fun. A decision has been made to create a new movie studio, but the company doesn't know anything about creating movies. Our team has been tasked with exploring what types of films are currently doing the best at the box office. We must then translate those findings into actionable insights that the head of the company's new movie studio can use to help decide what type of films to create.

2.0 Business Understanding

The movie-making industry is a complex, multi-billion-dollar global market that includes the development, production, distribution, and exhibition of films. This industry operates at the intersection of art, entertainment, and commerce, primarily producing films that attract audiences and generate substantial revenue.

2.1. Industry Context

The movie-making industry includes major film studios, production companies, and a vast network of support services such as talent agencies, production equipment suppliers, and post-production services. It is also heavily influenced by technological advancements, evolving customer preferences, and global marketing dynamics.

Revenue in the movie industry is derived from multiple sources as below:

- Box Office Sales – the primary revenue from ticket sales in theatres
- Streaming Platforms – services like Netflix
- Home Entertainment – includes digital purchases

2.2. Business Objective

The project analyzes select movie industry data, including audience preferences (IM.Db ratings) and financial performance (e.g. budgets and revenue) to make informed decisions that maximize profitability and audience satisfaction. This will also allow the company to derive actionable insights for the movie studio business.

Analyzing the identified data sources helps us determine which genres, storylines, and themes resonate most with audiences. This information can guide the studios in selecting projects more likely to succeed critically and commercially.

By studying ratings and reviews, producers can tailor content to specific age groups, cultural backgrounds, and other demographics that show high interest in certain types of movies.

We will see how analyzing financial data helps studios understand how budget size correlates with box office revenue. These insights can assist in deciding whether a high-budget blockbuster or a smaller film is a better investment for a particular market.

Our analysis will focus on the following objectives:

1. Identify Popular Genres: Analyze which genres tend to have higher ratings, more viewer engagement, high audience ratings, and are most profitable.
2. Analyze Characteristics of High-Rated Movies: Examine factors such as runtime, year of release, and genre combinations to see if they correlate with higher ratings.
3. Investigate Trends Over Time: Look at how preferences in ratings, movie length, and genres have evolved over the years, highlighting trends that may be valuable for the new studio to consider.
4. Correlate financial performance with popularity and ratings.

5. Offer genre and studio strategies for maximum ROI.

3.0 Data Sources and Understanding

We work with the following datasets:

- IMDb Database: Contains movie ratings, genres, and key details.
- TMDb Dataset: Includes popularity metrics, audience ratings, and genre encodings.
- Budget Dataset: Provides production budgets, domestic and worldwide revenue.

Key Questions for Data Understanding:

- What is the distribution of genres in the dataset?
- Are there missing values in critical columns (e.g., ratings, genres, runtime)?
- How are ratings distributed across movies?
- What are the relationships between tables that can help us analyze contributor impact (e.g., directors, writers)?

4.0 Data Preparation

Steps:

1. Backing up of the data to avoid accidental loss during the data cleaning.
2. Understanding the data by reviewing the structure, and checking for data types, and constraints in the database. We also look at formatting and key relationships between tables.
3. We identify and remove duplicate records to reduce redundancy using methods such as DISTINCT in SQL or functions in data processing tools like Pandas (drop_duplicates()).
4. We then identify missing data points (NULL or NaN values) and decide on how to handle them: remove, replace with default values, or use statistical methods (e.g., mean, median, mode).
5. We structural errors by identifying issues with naming inconsistencies, typos, or formatting issues (e.g., Sci-Fi vs.SciFI). Therefore, standardizing naming conventions and formats.
6. Removing columns or rows that are not relevant to our analysis
7. Fix data type mismatches by ensuring each column has the correct data type (e.g., dates stored as DateTime objects, numbers stored as int or float). Then convert types as necessary for consistency.
8. Standardize formatting; capitalization, and number precision to uniformity (e.g., changing all entries to lowercase or uppercase).
9. Identifying outliers using statistical methods and deciding on how to handle them: correct, sap, or remove them based on context.

10. Check referential integrity by verifying that the relationship between tables is maintained (e.g., foreign keys pointing to existing primary keys).

5.0 Modelling/ Analysis

The analysis includes several statistical and visualization techniques to uncover insights:

Univariate Analysis: Examining the distribution of single variables, such as production budgets and gross revenue, through histograms and summary statistics.

Multivariate Analysis: Combining multiple factors (budget, genre, year) to analyze financial outcomes using heatmaps, scatter plots, and regression models.

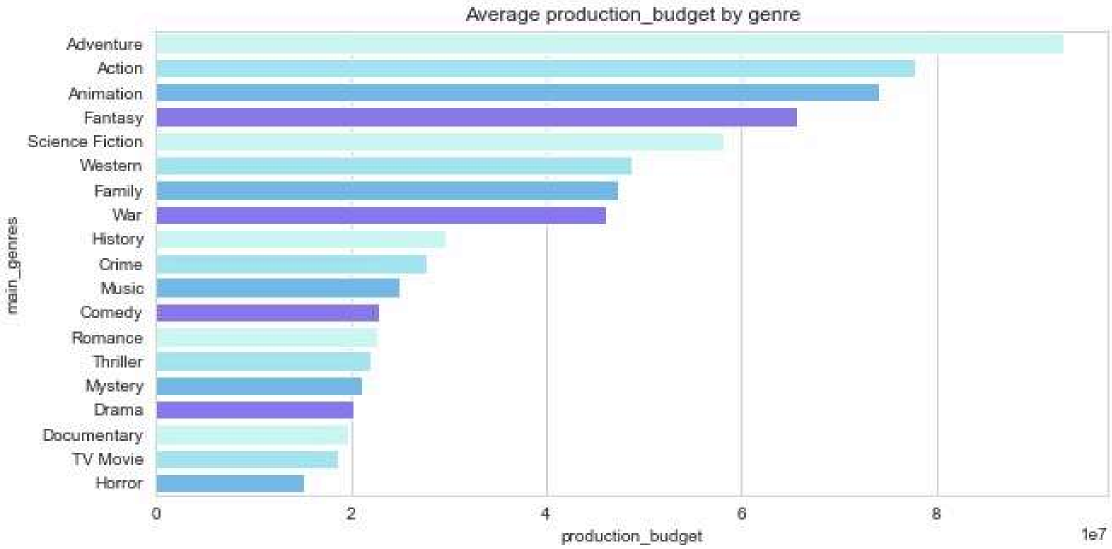
Correlation Analysis: Exploring relationships between financial variables to understand associations (e.g., production budget vs. worldwide gross).

Regression Analysis: Modeling the relationship between production budgets and worldwide gross to predict potential revenue.

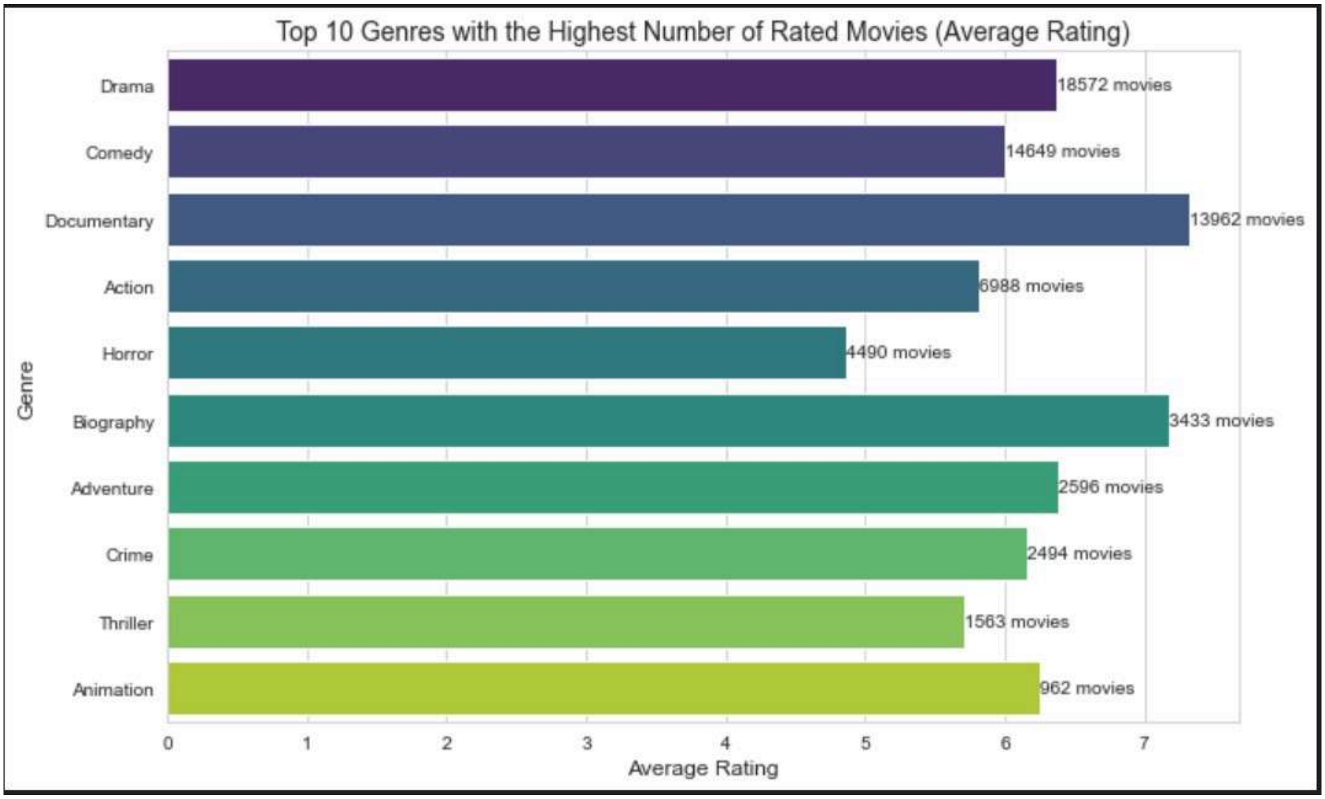
5.1. Visualizations

The following visualizations are included in this analysis:

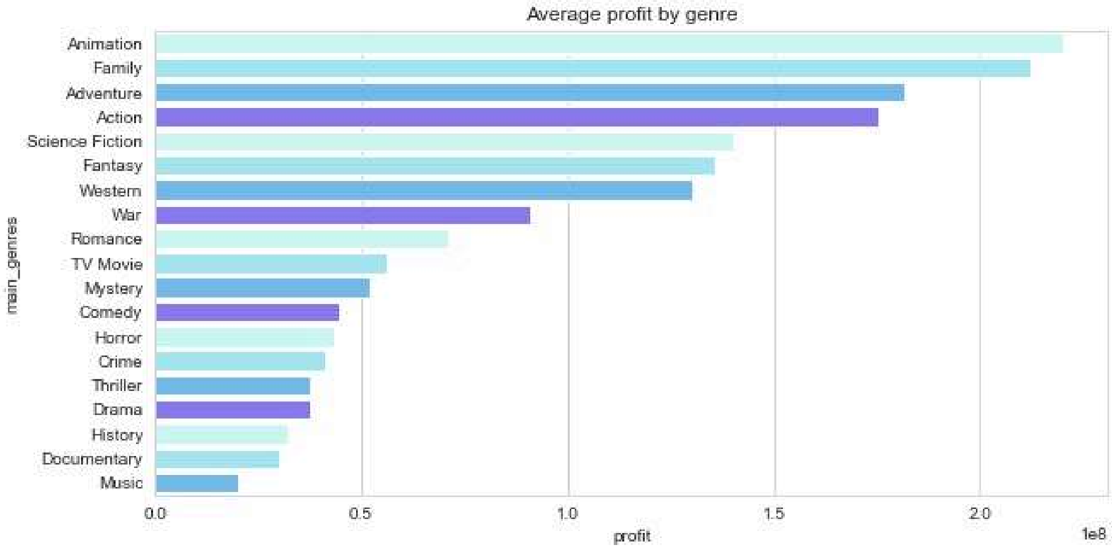
Showing movie count by genre to highlight popularity and distribution.



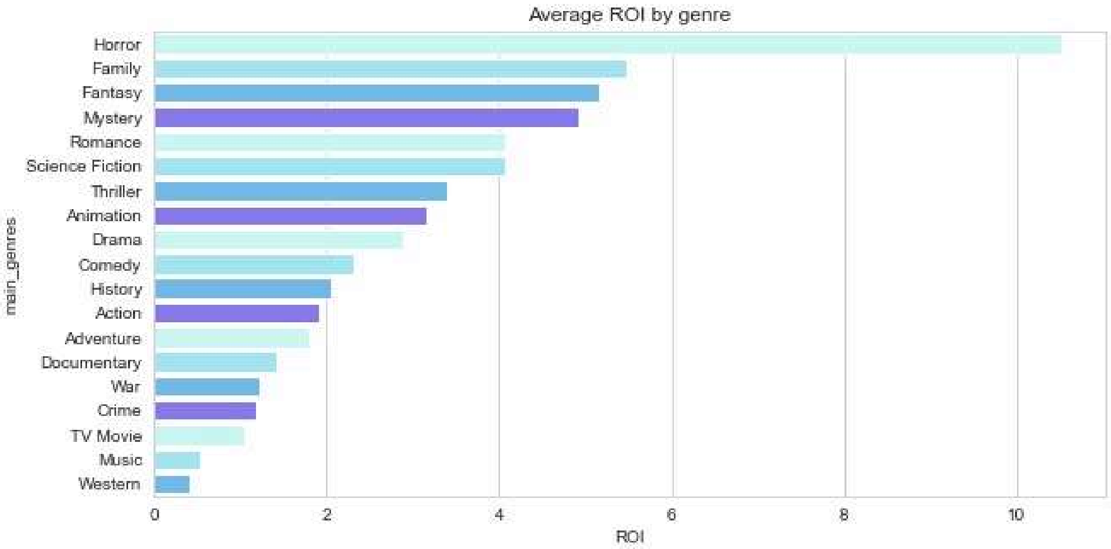
Illustrating movie trends over time.



Highlighting the average profit across genres.



Visualizing the contribution of various genres to overall revenue.



Key Findings

Some insights uncovered during the analysis:

1. Genres like Action and Adventure are often associated with higher budgets and revenue.
2. Production budget has a positive correlation with worldwide gross, though other factors also play a significant role in profitability.
3. Movies with high rating tend to dominate in revenue, but not always in ROI, indicating varied profitability.

Pre-requisites:

Python libraries: pandas, matplotlib, seaborn, statsmodels

Tableau Public for interactive visualizati

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 9

Languages

- ● **Jupyter Notebook** 100.0%