thomasamuti-cpu / **Movies_project_moringa**

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

This GitHub is a collaboration project with other Moringa class-mates. In it, we aim to use exploratory data analysis to generate insights for a business stakeholder.

0 stars 1 fork 0 watching Branches Activity
Tags

Public repository

main 6 Branches 0 Tags Go to file t Go to file Add file Code

AmutiMombo Edited README, Added dashboard, git_ignore, figures, notebook and pre... 549af52 · 4 minutes ago

Edited csv files for group analysis	Edited README, Added dashboard, git_ignor...	4 minutes ago
Figures	Edited README, Added dashboard, git_ignor...	4 minutes ago
Pdf documents	Edited README, Added dashboard, git_ignor...	4 minutes ago
.gitignore	Edited README, Added dashboard, git_ignor...	4 minutes ago
Movies_project_Dashboard.twb	Edited README, Added dashboard, git_ignor...	4 minutes ago
README.MD	Edited README, Added dashboard, git_ignor...	4 minutes ago
student.ipynb	Edited README, Added dashboard, git_ignor...	4 minutes ago

README

Project Overview

This project will employ exploratory data analysis to generate actionable insights that inform strategic decisions for business stakeholders.

Business Understanding

As major companies increasingly invest in original video content, our company has decided to enter the entertainment industry by launching a new movie studio. However, the organization currently lacks experience in film production and does not have a clear strategy for deciding what types of movies to create.

The primary business goal of this project is to reduce the risk associated with entering the movie industry by using data-driven insights to understand which types of films are performing best at the box office today and identify patterns related to movie genre, rating, viewership, and finanical performance (income generation and production costs).

By analyzing current box office trends and successful films, this project will provide actionable insights that can guide the leadership of the new movie studio in making informed decisions about what kinds of movies to produce. These insights will help the studio prioritize film types that are more likely to achieve commercial success, maximize return on investment, and compete effectively with established studios.

Ultimately, the findings from this analysis will support strategic decision-making around film development, allowing the company to enter the market with a clearer understanding of audience demand and industry trends.

Business Goal

To determine which movie genres are the most commercially successful and financially viable.

Business Question

What movie genres are the most commercially successful and financially viable?

Specific Questions:

a. Which movie genre is most commonly produced? b. Which genres received the highest audience ratings? Were the differences statistically significant? c. Which genres received the largest audience/ viewership? d. What was the production cost of the 5 most common movie genres? e. What income was generated by the 5 most common movie genres?

Sub-Analysis: a. What is the relationship between production costs and gross income?

Data Understanding

Data sources, description and relevant columns used

Five datasets were provided. We will focus on the following three that provide the data most relevant to the business questions above.

1. IMDB (https://www.imdb.com/) IMBD is and online database of information related to films, television, series, podcasts, video games and online streaming content. The provided data is in a SQLite database and the movie ratings table will be most relevant in answering the business questions on audience size and ratings.
2. Box office Mojo (https://www.boxofficemojo.com/) This is a subsidiary of IMDB and is an online platform that provides data on commercial performance of films i.e. ticket sales revenue. The data comes as a CSV file with columns providing the domestic gross earnings and foreign gross earnings.
3. The Numbers (https://www.the-numbers.com/) This online platform also provides a film's domestic and foreign gross revenue figures from ticket sales but also includes data on production budgets.

Data Preperation and Cleaning

The following libraries are imported for use in data preparation and cleaning:

- Pandas: to read, edit, save CSV, TSV files
- SQLite3 : to query SQLite database tables

Data Analysis

The following libraries are used for data analysis:

- Pandas
- Statsmodel and Scipy.stats for inferential statistics and regression models

Regression analysis

Data Visualisation

The following libraries are used for data analysis:

- Pandas
- Matplotlib and seaborn for visualisations

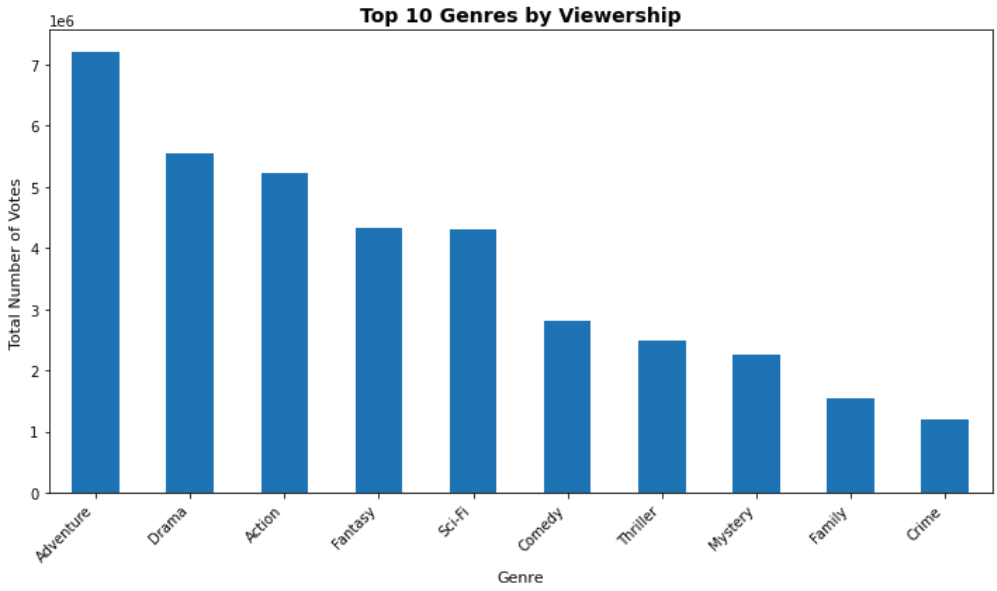
Findings from Data Analysis

From our analysis:

a. Drama, Comedy and Documentary were most commonly produced.

b. Documentary, Gameshow and News received the highest audience ratings. Biography, Music and Documetary received the highest total votes. alt text. Our ANOVA test p value was <0.05. This suggested that there is a high possibility that our null hypothesis was not true. Given that the null hypothesis was that 'Movies did not differ on ratings', our p value suggests that movie genres actually did differ on ratings. Therefore we reject the null hypothesis and accept the alternative.

c. Adventure, Drama, Action, Fantasy and Sci-Fi received the largest audience viewership.



d. Avatar, Pirates of the Carribean, Dark Pheonix, Avengers, Stars Wars had the highest production budget whereas My Date with Drew, A Plague So Pleasant, Return to the Land of Wonders, Following and the Mongol King has the lowest production budget. Being movies, comparison with movie sites had to be made to determine which genres the movies belonged to. Following comparison, analysis of production budgets from the Numbers dataset shows that the highest production costs are consistently linked to Action, Adventure, and Science Fiction films. This trend reflects the resource-heavy nature of blockbuster productions, which often involve extensive visual effects, large casts, and worldwide filming needs.

e. Avatar, Titanic, Star Wars, Avengers, Jurrasic world generated the most income from the numbers df. alt text. Marvel's Avengers, Age of Ultron, Black Panther, Harry Potter and Deathly Hallows Part 2 and Stars Wars Jedi generated the most income from the box office df. alt text. To determine the income generated from the 5 most common movie genres, a reliable identifier that links genre information fom ***IMDb*** with income data from ***The Numbers*** and ***Box Office Mojo*** movie datasets is required. Because such an identifier is not available from the datasets provided, an accurate income analysis by genre cannot be carried out without risking data quality. However genre classifications for the top 5 movies with highest income generated from ***The Numbers*** and ***Box Office Mojo*** datasets were verified manually using ***IMBd*** website, https://www.imdb.com/list/ls564012879/ . Since there was no unique identifier connecting genre and revenue data in the IMBD, Numbers and Box Office datasets, genre information serves for context rather than direct integration with the datasets. Among the ten movies with the highest income, Action, Adventure, and Science Fiction take the lead.

These genres are present in 8 out of the 10 films and make up most of the total revenue. In contrast, genres like Drama, Romance, and Fantasy show up less often but still generate substantial revenues in some cases.

Sub-Analysis:

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

Our model met the assumptions of linearity, independence of errors and equality of variances. Regression formula: gross income = 3.14(production costs) - 6.91

There is a strong, positive relationship between spending more on production and earning higher revenue. On average, every dollar spent on production yields about $3.14 in return. However, this strategy comes with higher risk; as budgets increase, the variability in returns also increases, meaning high-budget movies can result in massive profits but also significant losses. alt text.

Additional visualization (Tableau Dashboard)

text

Tableau Link:
https://public.tableau.com/app/profile/thomas.amuti/viz/Movies_project_Dashboard/Mostsuccessfulandfinanciallyviablemoviegenres?publish=yes

Business Recommendations

Based on our results:

1. Drama, Comedy and Documentary were most commonly produced
2. Biography, Music and Documentary received the highest total votes.
3. Documentary, Gameshow and News received the highest audience ratings.
4. Action, Adventure, and Science Fiction had the highest income.

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 highest audience ratings.

Limitations

1. We left out the rotten tomatoes movie source since the rating column that we would have used from it had both numerical and non_numerical grade ratings that would have made data analysis difficult.
2. We also opted to leave out the tmdb movie source since the movie genres within it 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 also not used for regression analysis.
4. Genre classifications during assessment of the production costs and 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 dataframe that connects genre and budget data in our analysis.

Suggestions for further analysis

We suggest that for future analysis:

1. Improvement of Genre Classification be done to address genre inconsistencies across datasets by creating a standardized genre mapping.

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 6

Languages

- **Jupyter Notebook** 100.0%