











Phase_Two_Project Public

 JoelKy-coder create zip  720e1ed · 22 minutes ago 

	Data	create zip	22 minutes ago
	.gitattributes	Initial commit	last week
	.gitignore	gitignore	2 hours ago
	Group one phase ...	Graphs	7 hours ago
	Linegraph2.png	Added some visuals	6 hours ago
	Phase_Two_Project...	Notebook PDF upload	6 hours ago
	README.md	Update README.md	5 hours ago
	barplott.png	Added some visuals	6 hours ago
	barploty.png	Added some visuals	6 hours ago
	presentation.pdf	Add files via upload	1 hour ago

Statistical

[Readme](#)[Activity](#)[0 stars](#)[2 watching](#)[0 forks](#)

Releases

No releases published

[Create a new release](#)

Packages

No packages published

[Publish your first package](#)

Contributors 5



Languages

● Jupyter Notebook 100.0%

Group One Phase_Two_Project

Project Overview

This project aims to explore key trends in the movie industry, including genres, themes, and audience preferences that are driving box office performance.

Project Description

This project explores revenue trends in the movie industry using statistical hypothesis testing. The primary objective is to determine whether there is a significant difference in movie revenues over movie votes, accounting for variables like genre, ratings, duration and release year. The analysis follows the CRISP-DM methodology for structured and systematic data analysis.

Hypothesis Testing in Movies Data:

Analyzing Revenue Trends Over Number of Movie Votes

Objectives

- To identify the effect of ratings across different genres.
- To determine the genres that generates highest revenue/profit.

- To identify the trends in revenue generation. Formulate hypotheses about movie revenue trends.
- Visualize results for actionable insights.

Dataset

- [IMDB](#) and [BOM](#)
- **Features:**
 - Revenue details: domestic_gross, foreign_gross
 - Movie details: title, year, runtime, genre, ratings
- **Access:** Publicly available for download from [Here](#).

Methodology

This project is structured using the CRISP-DM framework:

1. Business Understanding:

- Formulated the hypothesis: "There is no significant difference in movie revenues over number of movie votes."

2. Data Understanding:

- Performed exploratory data analysis (EDA) to identify patterns and relationships.

3. Data Preparation:

- Combined data from two different sources (SQLite DB and CSV file),
Cleaned the dataset, handled missing values, and filtered outliers.

4. Modeling and Testing:

- Applied hypothesis testing methods linear regression

5. Evaluation:

- Interpreted the results and validated the hypothesis.

Hypotheses

1. Primary Hypothesis:

- **Null Hypothesis (H_0):** There is no significant difference in revenue over number of votes.
- **Alternative Hypothesis (H_1):** There is a significant difference in revenue over number of votes.

Tools

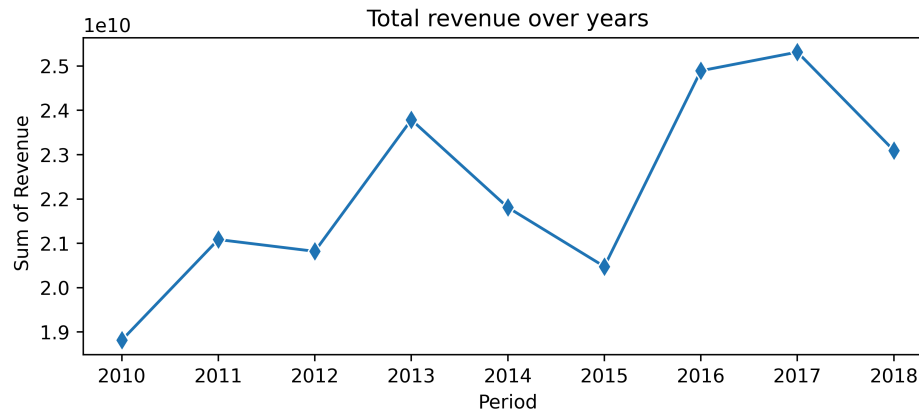
- **Languages:** Python
- **Libraries:**
 - pandas and numpy for data wrangling
 - sqlite3 for db reading
 - statsmodels for modelling
 - matplotlib and seaborn for visualization

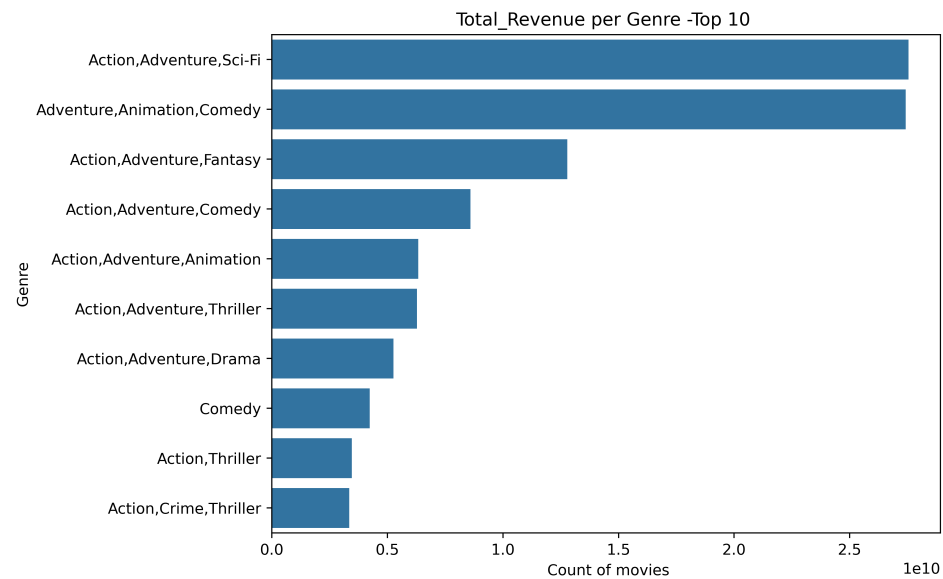
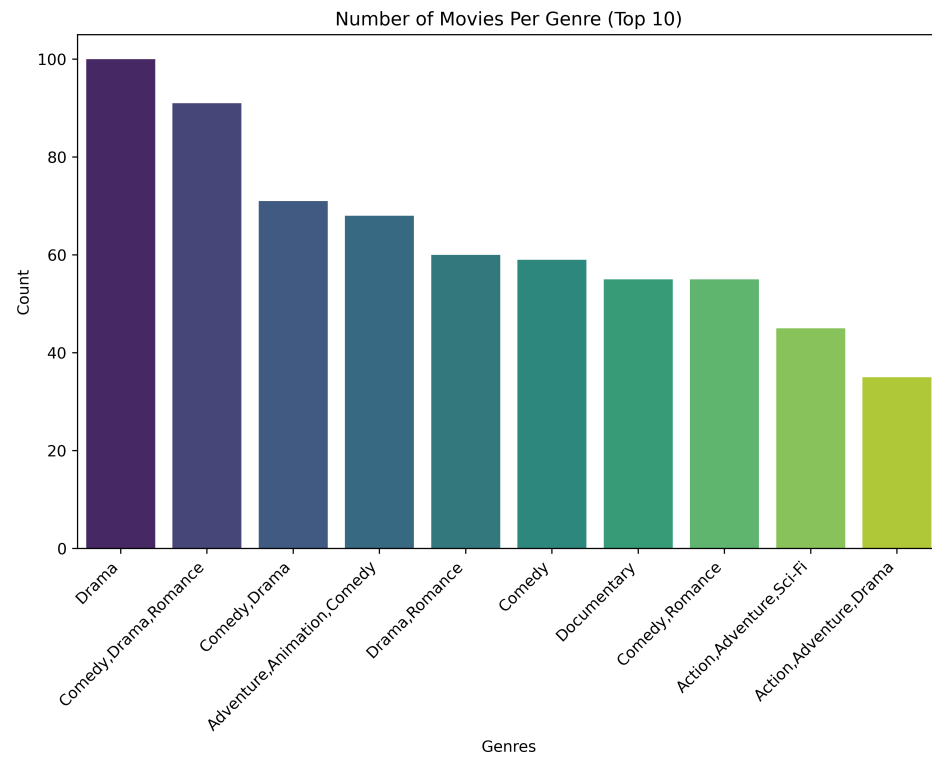
Results

- The hypothesis testing revealed a significant positive trend in movie revenues over number of movie votes ($p < 0.05$).

- Movies with significantly higher votes yielded more revenue compared to movies with lower votes.
- Though Drama movies were highest in production, the combination of Action, Adventure, Sci-Fi and Adventure, Animation, Comedy generated more revenue

Quick view of visualization





Conclusion

Revenue by Genre: Genres like Action, Adventure, and Sci-Fi consistently generate higher revenues, emphasizing the need to focus on producing films within these popular categories to capture broad audience appeal.

Audience Engagement and Revenue: Movies with higher audience engagement (reflected by the number of votes) show a significant positive correlation with total revenue. This indicates the importance of building a strong online presence and encouraging audience feedback to drive financial success.

Revenue Growth Over Time: Total movie revenue has steadily increased over the years, highlighting the movie industry's growth and its expanding global reach.

Model Limitations: While the linear regression model identified the number of votes as a significant predictor of revenue, its relatively low R-squared value (46.4%) and multicollinearity suggest the need to incorporate additional variables, such as genres, ratings, and studios, to improve the model's predictive accuracy.

Recommendations

Focus on High-Grossing Genres: Prioritize films in successful genres like Action and Adventure, which consistently perform well at the box office. Understanding audience preferences within these genres can significantly boost profitability.

Boost Audience Engagement: Since audience votes strongly correlate with revenue, the studio should actively encourage participation through targeted campaigns and fan-driven promotions, both before and after a film's release.

Track Revenue Trends: To stay competitive in the evolving movie industry, the studio should track changes in revenue patterns over time. Monitor changes in consumer behavior, regional preferences, and technology (e.g., streaming) to stay ahead of industry shifts.

Leverage Studio Branding and Partnerships: Analyze the success patterns of top-performing studios and identify opportunities for improvement or collaboration. Partnering with leading studios or well-known directors can enhance brand visibility and drive higher revenue for films.

Future Ideas

- Analyze revenue trends across specific genres to identify high-performing categories.