bocekout / Phase2-Project

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

View license

0 stars 1 fork 1 watching Branches Activity
Tags

Public repository

3 Branches Tags Go to file t Go to file Add file Code

bocekout Merge pull request #12 from bocekout/Ricky dfc780f · now

Presentation	whoops updating presentation file	3 hours ago
images	Almost final commit. One more thin...	4 hours ago
notebooks	Almost final commit. One more thin...	4 hours ago
.gitignore	fixed noteboook	2 days ago
CONTRIBUTING.md	add files and readme	last year
Data.zip	final commit i hope	3 hours ago
FINAL.ipynb	Almost final commit. One more thin...	4 hours ago
LICENSE.md	add files and readme	last year
README.md	fix links in readme	now

README License

A Guide to Box Office Success

Authored by Elif Surucu & Ricky Bocek

Instructions Before Proceding

All datasets used in this analysis are available in the Data.zip file - just unzip in place to use our project. The unzipped folder is already listed in the .gitignore to avoid errors with trying to upload the IMDb sqlite database file to GitHub, as it is too large when not compressed.

Project Overview

This analysis of movie data, sourced from Kaggle, The Numbers, and IMDb, investigates the financial success of movies by analyzing the risk-reward relationship between production budget and profitability, adjusted to the yearly Consumer Price Index. This analysis leverages historical movie data to identify patterns and key features, such as genres, directors, and number of principals, that correlate with higher profitability. By applying statistical techniques like ANOVA and linear regression, the project uncovers which genres and budget categories lead to the most successful movies. Additionally, the project offers recommendations for future movie production, focusing on maximizing ROI based on these identified success factors.

Business Understanding

The movie industry operates in a highly competitive environment where production companies aim to maximize profitability while minimizing financial risk. However, predicting the financial success of a movie is challenging due to various factors like genre, budget, cast, and market trends. This project seeks to address this uncertainty by identifying the key features that drive higher returns on investment (ROI) for movies. The goal is to provide movie studios and producers with actionable insights to optimize their budgets and make informed decisions about which genres, directors, and budget levels are more likely to result in profitable outcomes. By analyzing historical data, the project can help studios focus their resources on the most promising projects, thereby improving financial performance in an unpredictable market.

Data Understanding

Project data sources:

- IMDb
- The Numbers
- Kaggle - The Ultimate Film Statistics Dataset
- US Bureau of Labor Statistics

This project combines movie budget and earnings sourced from Kaggle and The Numbers with an IMDb dataset of production elements to create a feature-rich list of over 5700 movies with financial metrics, using historical Consumer Price Index (CPI) tables from the US Bureau of Labor Statistics to adjust dollar values for inflation. Key data fields include genres, production budgets, domestic and worldwide gross earnings, and director. The project focuses on maximizing ROI to evaluate the financial success of movies, while accounting for risk as determined by production budget.

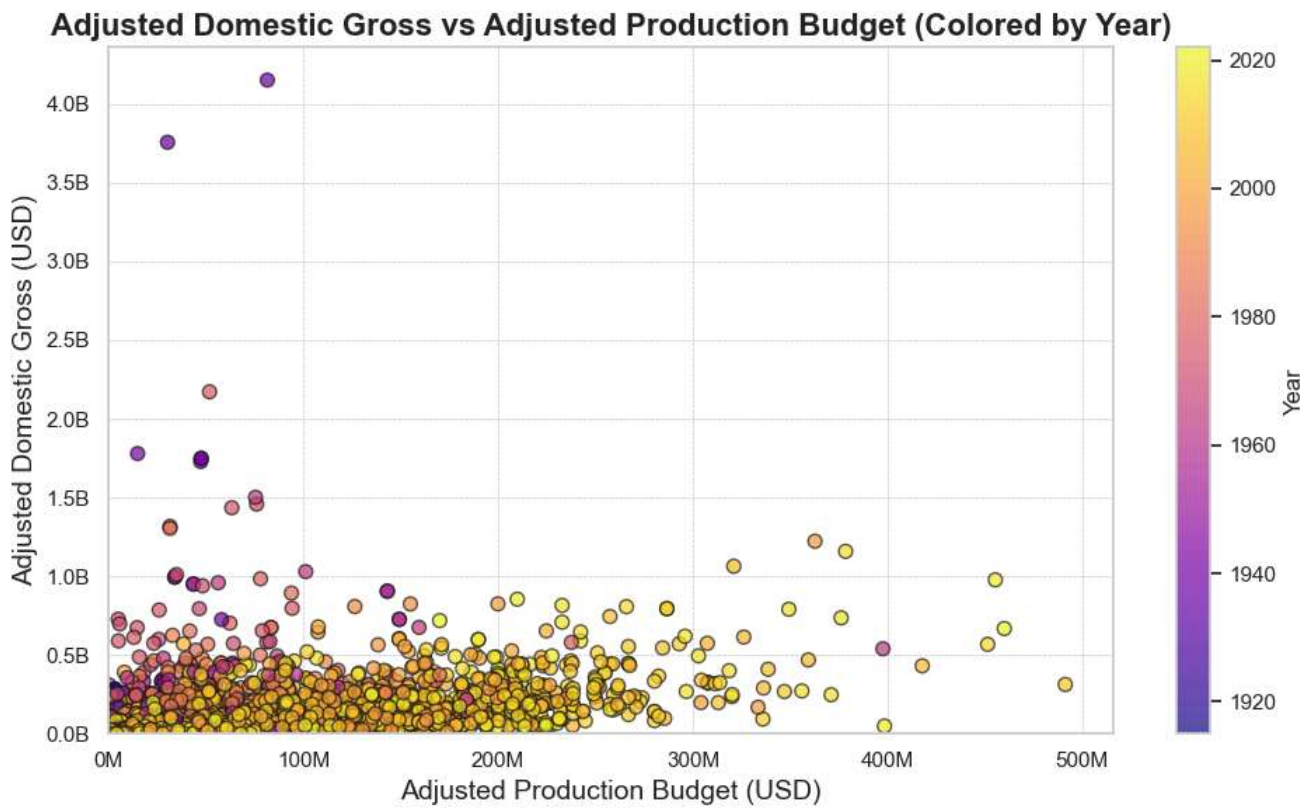
When evaluating the following recommendations, it is crucial to keep in mind that our data can *only* account for projects that actually survive to release and distribution. This analysis can help inform what projects to pursue, but none of this - aside, possibly, from director selection - is relevant to the task of assuring that a film actually comes to fruition.

Data Preparation

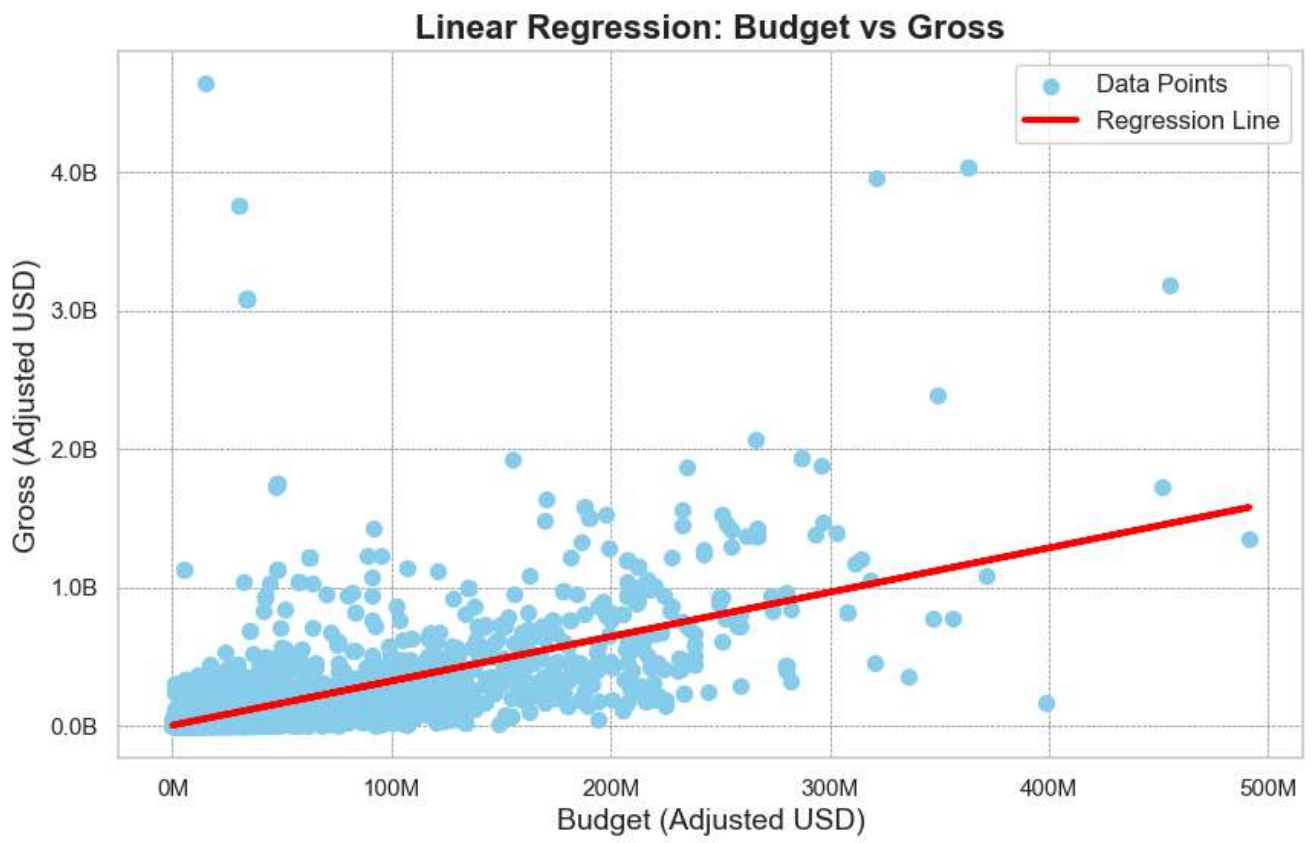
To ensure accurate analysis, we addressed missing values, removed duplicates, and ensured data consistency across all relevant fields. We merged datasets (movies, directors, and others) based on movie_id and person_id to consolidate essential information like gross earnings, genres, and director. We derived inflation-adjusted budget and revenue values to account for dollar value change over time and calculated ROI to allow comparison of success between vastly different budget levels. We split multi-genre entries to analyze each genre's impact separately on metrics like ROI. We filtered directors based on relevant criteria to ensure all recommendations remained actionable and cohesive.

Analysis and Recommendations

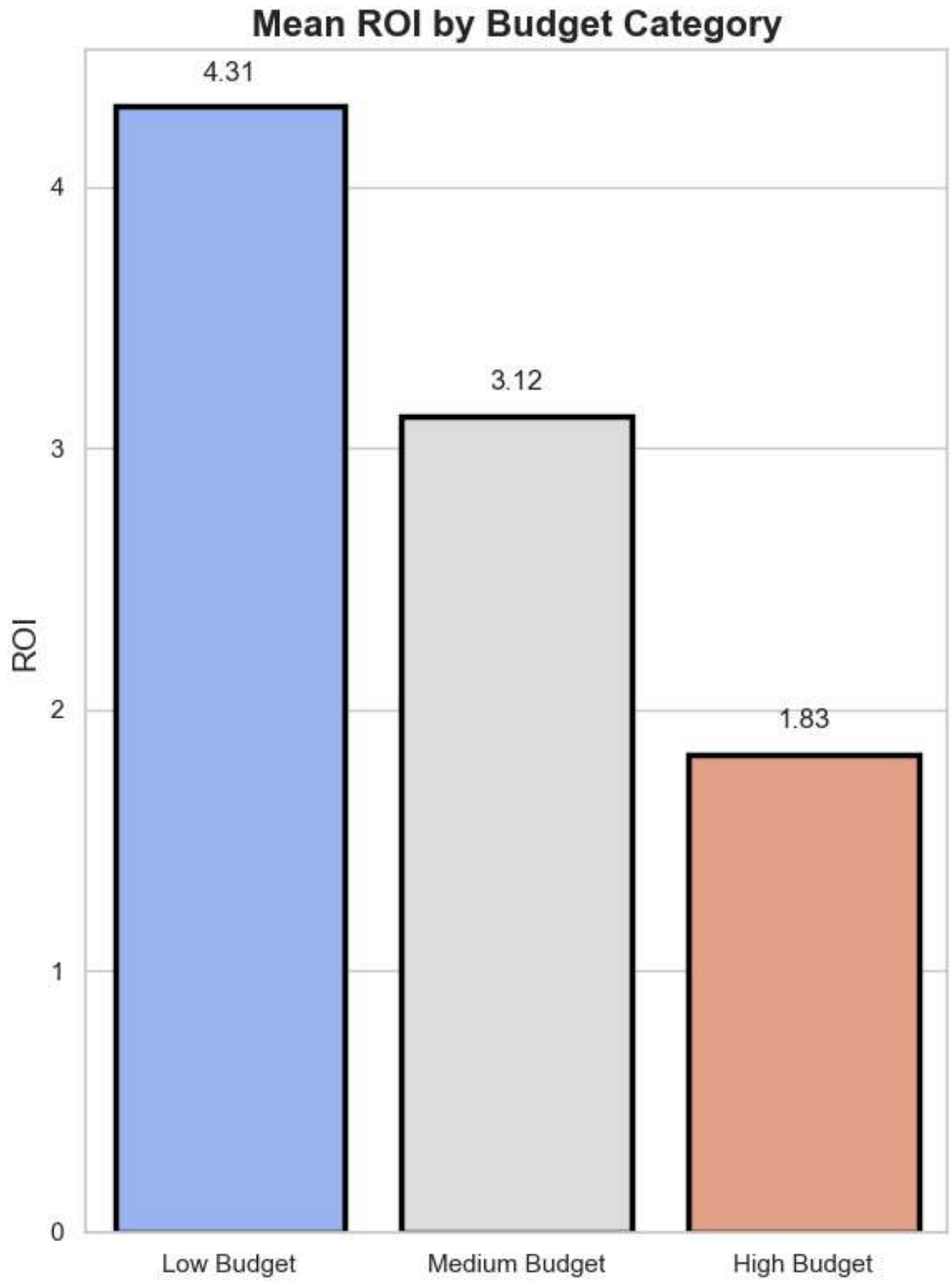
We started by comparing adjusted gross revenue to adjusted budget then added linear regression to look for direct correlation. We expanded on the analysis by comparing ROI by budget category and then by two measures of acclaim. Then perfomed ANOVA tests looking for significant differences in ROI by genre or number of principals and characterized the differences we found with a post-hoc Tukey HSD test.



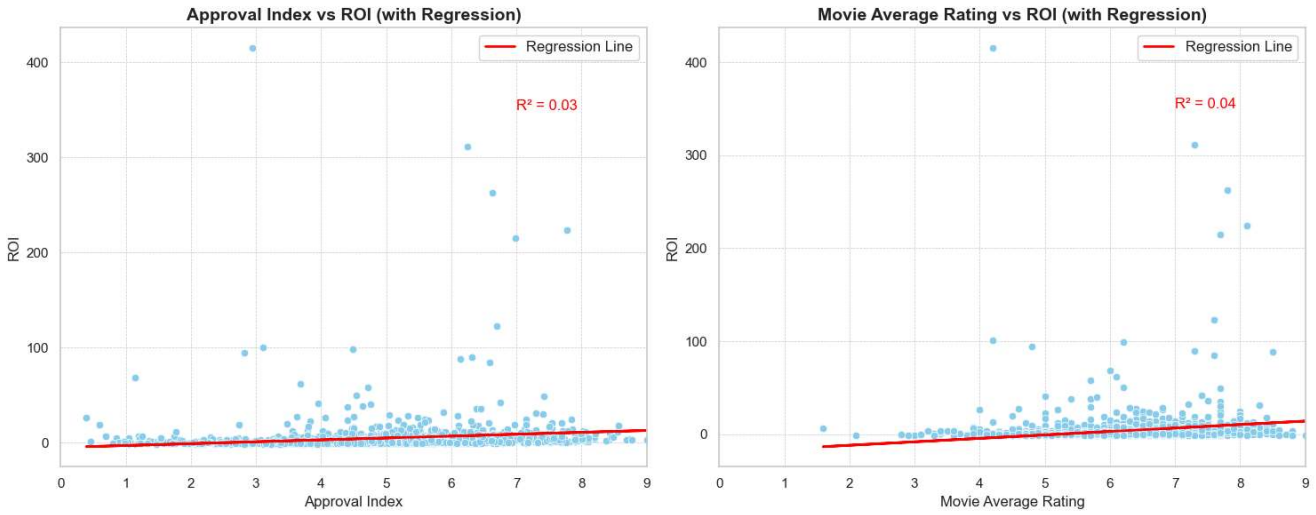
This exploratory visualization of the data emphasizes the significant variance in gross earnings and, consequently, in return on investment (ROI). The color scale indicates that more recent films (yellow and light orange) tend to have higher production budgets than older films (purple), reflecting expansion and capitalization of the industry. Given the high variance in gross revenue and the fact that our top revenue outliers are not the same as our top budget outliers, a logical question followed: "IS there a relationship between budget and gross revenue?"



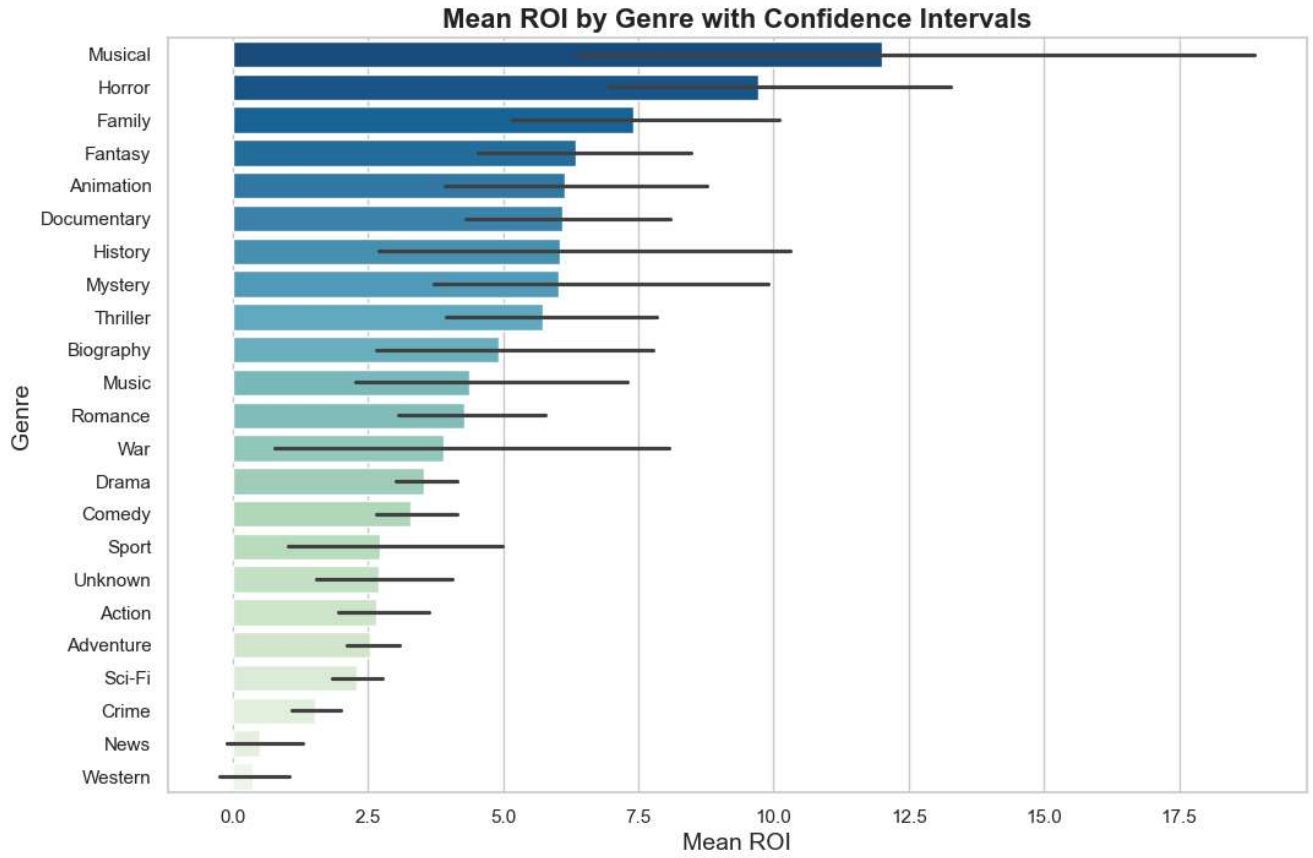
From the regression, we observe that for each (adjusted) budget dollar spent, there is an average expected revenue of $3.19 (in 2022 dollar value), with budget driving as much as 30% of the variance (an R-squared of ~0.3).



However, when analyzing budget categories, we notice that high-budget films have a diminished ROI compared to low-to-medium budget films, so the relationship is apparently not entirely linear. Increasing the budget within reasonable limits is an effective strategy for maximizing revenue, but due to diminishing average ROI and the previously-stated limitations in the data, managing risk exposure is still key - it's more worthwhile to be comfortably able to make *and release* several low-budget films than to over-extend for a higher budget film.



Both the approval index and movie average rating have very weak positive correlations with ROI, with R-squared values of 2.6% for Approval and 4% for Ratings. The relationships are small but present, as indicated by the positive slopes of both regression lines. However, the high dispersion of the data (especially in the ROI outliers) suggests other factors may have a stronger influence on ROI. In general it is recommended to ignore ratings as an avenue to ROI.



ANOVA test across genres allowed us to determine that there WAS a significant difference in mean ROI between genres. Following up with a post-hoc Tukey HSD test allowed us to describe which categories performed significantly better or worse than others. This bar plot clearly mean ROI differences between various movie genres, with error bars to assess the overall condience interval. This helps pinpoint which genres tend to perform better or worse in terms of ROI. For example, our data shows a good degree of confidence in the mean ROI for horror films, our second-highest performing category, but some high-performing genres like musicals are not as well represented in the data, so they exhibit wider confidence

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 6

Languages

- **Jupyter Notebook** 100.0%