williambguo /
ds-p2-project

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

0 stars 0 forks 1 watching Branches Tags Activity

Public repository

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

williambguo Update README.md 2321ea2 · 1 minute ago

code	Delete redundant notebook	15 hours ago
data	Giga update	18 hours ago
images	Small update	15 hours ago
.gitignore	Giga update	18 hours ago
README.md	Update README.md	1 minute ago
presentation.pdf	Rename slide deck	1 hour ago

README

Analysis of Movie Data

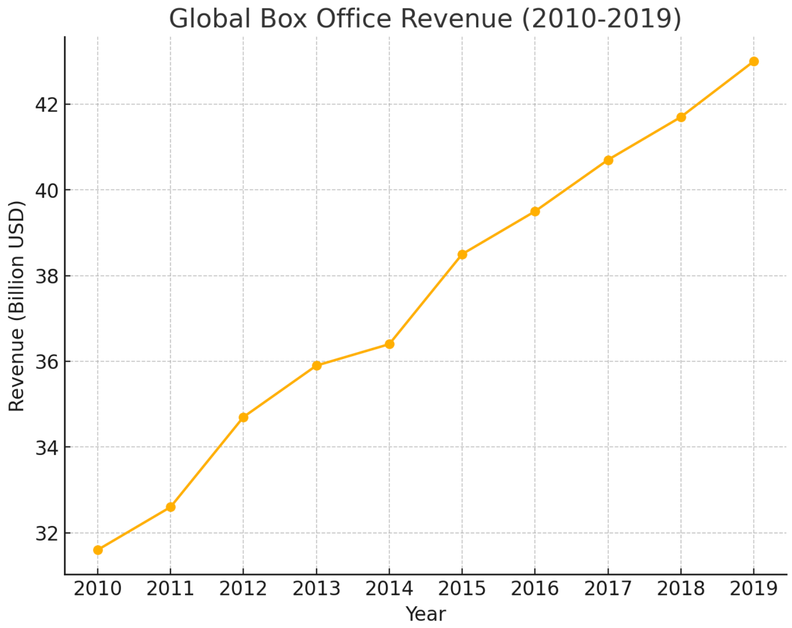
Overview

The purpose of this project is to use multiple datasets containing movie information to identify what movie-related factors lead to success as quantified by return on investment (ROI). The analysis aims to provide insights that could contribute to a film studio or production company's decision-making regarding the types of movies to produce. Ordinary Least Squares (OLS) regression is used with ROI as the dependent variable to estimate how effective selected independent variables such as film genre or film runtime are at predicting movie success. Due to the nature of the datasets I elected to use this analysis is set in 2019.

Business Understanding

Having seen how big companies especially in tech (e.g. Amazon, Apple, and Netflix) have started investing heavily in original video content production, our company is now looking for opportunities in the film industry. I have been tasked with exploring what types of films are currently doing the best at the box office.

The film industry has seen steady growth over this decade and is projected healthy growth going forward thanks to the key growth drivers such as the rise of streaming platforms and growing entertainment demands in international markets like China.



There are several market trends of note currently which are all undergirded by the globalization of content. The aforementioned rise of streaming platforms and direct-to-consumer models of original video content is one. Streaming is already a dominant force in film consumption for many consumers. In addition, the companies behind streamers are investing heavily into their own production projects. Amazon and Netflix have already begun this process. Disney and Apple - who have just launched their own streaming platforms - are sure to follow.

The dominance of franchises has been starkly clear over this decade with Marvel's Marvel Cinematic Universe (MCU) leading the way. While franchises have always performed well at the box office (see Star Wars, Jurrasic Park, Harry Potter), the MCU and Disney in particular have asserted their dominance in the 2010s. In 2019, out of the top 10 movies in terms of domestic gross box office, only the 10th ranked movie was an original idea with the rest being either sequels or part of a large franchise.

Closely linked to franchise dominance is the shift toward tentpole productions has also happened. Movie studios are increasingly focusing on large-budget tentpole productions, especially established IP and cinematic unverses. The success of those projects allows them to bankroll other projects under the same IP or cinematic unverse umbrella. Smaller mid-budget films are thus being squeezed out of theaters and are ending up on streaming platforms or at smaller distributors.

Key Objective

The main objective of this analysis is to identify the factors that positively affect a movie's ROI.

Data Understanding and Analysis

Source of Data

The data used for this analysis was acquired from two sources:

- Movie budget and box office data from The Numbers
- All other movie data such as release date, genres, and runtime from IMDb

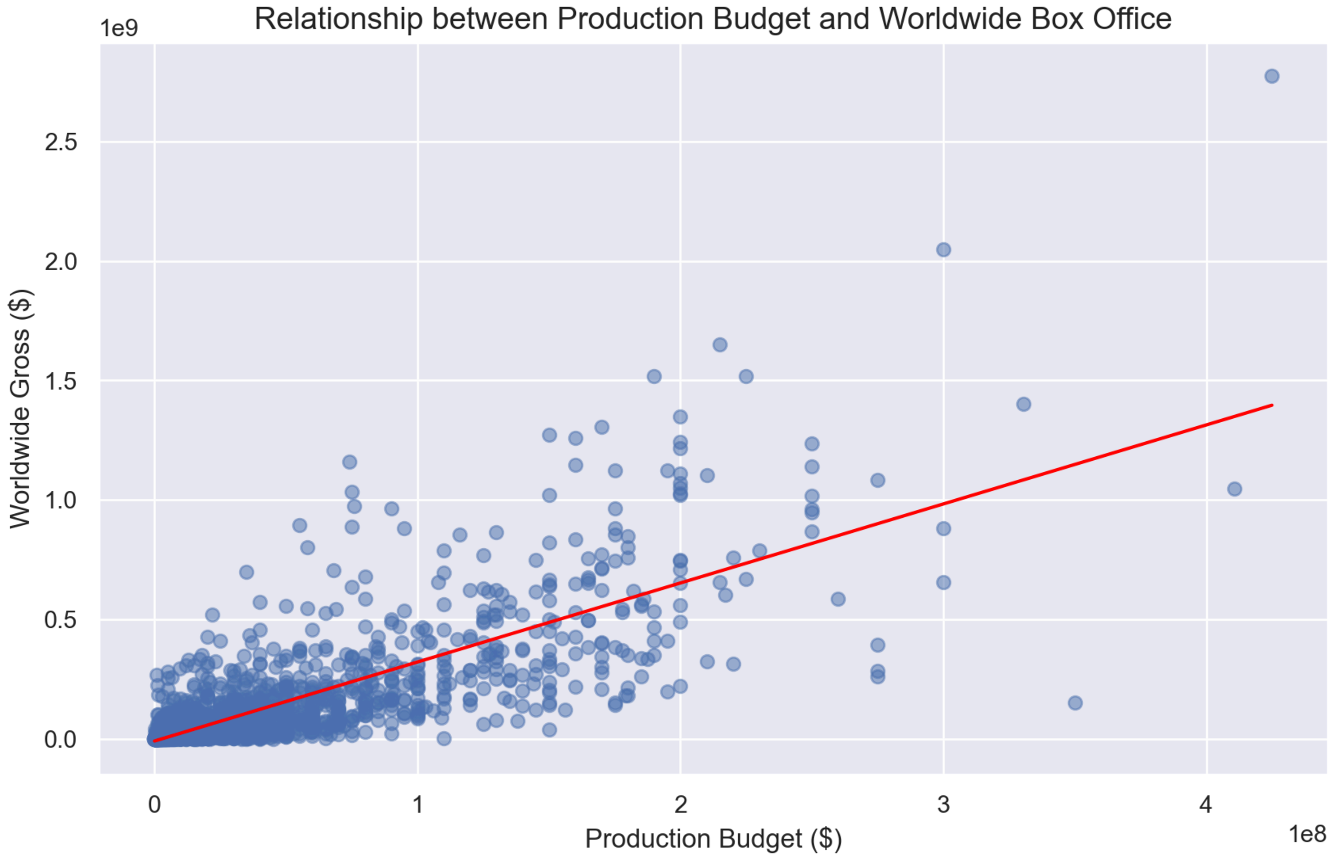
Data Cleaning

The datasets were faily clean and well-organized already so after some initial exploration and basic data cleaning I merged the table from The Numbers (TN) with one of the IMDb tables. The IMDb database file came with eight tables of which I only utilized 'movie_basics'. I merged the tables on the only common key available which was movie title.

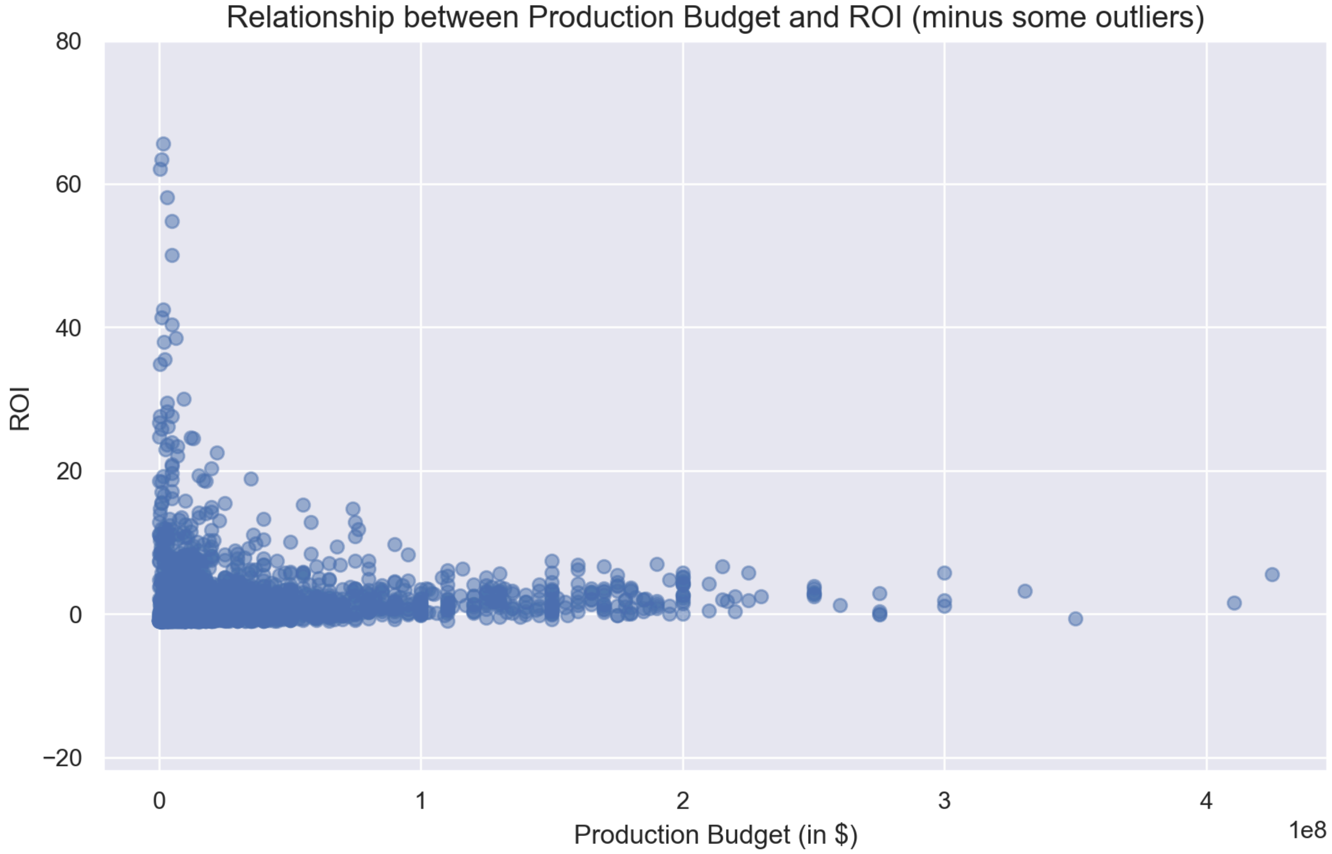
```
df_merged = pd.merge(df_tn, df_movie_basics, how='inner', on='title')
```

Visualization of Data

A basic scatter of production budget against box office revenues confirms an intuition that says the more money you invest in producing a movie the more revenue that movie will generate.



Plotting that budget against ROI, however, reveals a relationship that is not very obvious.



Feature Engineering

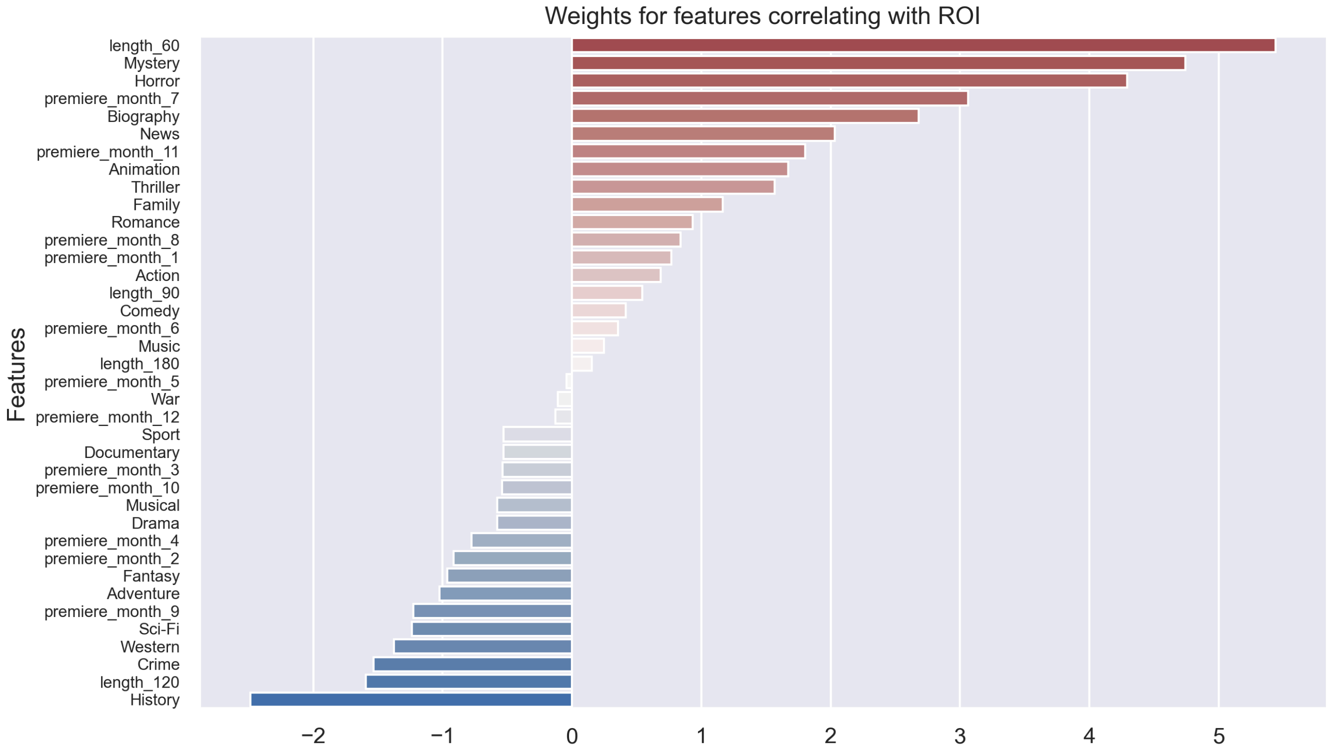
Next I did some basic feature engineering to obtain the independent variables needed for my regression analysis. Since the independent variables are all categorical I converted to a quantitative form by using dummy variabless. Below is an example with the genre feature and the first five rows of the newly created genre columns:

```
more_dummies = tn_imdb['genres'].str.get_dummies(sep=',')
```

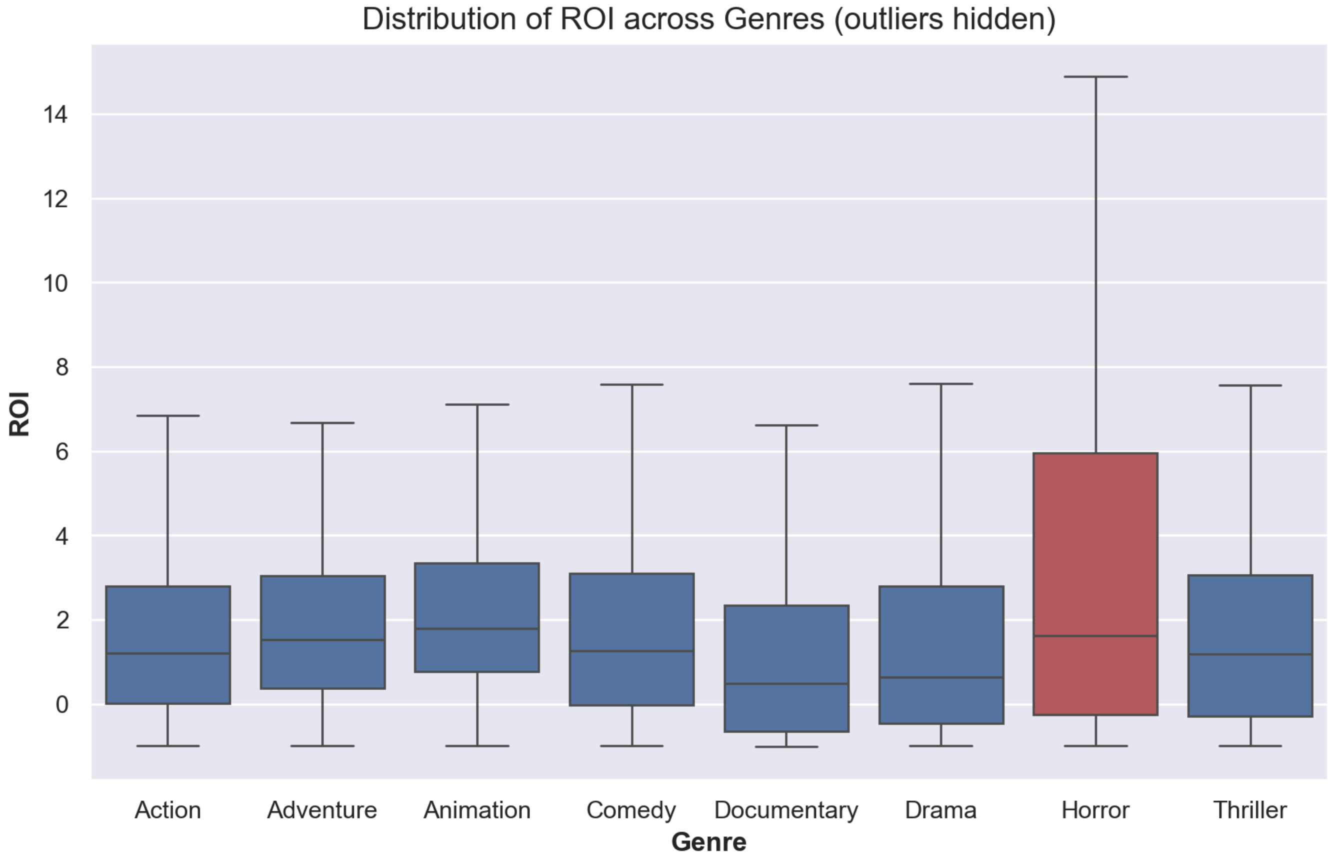
	Action	Adventure	Animation	Biography	Comedy	Crime	Documentary	Drama	Family	Fantasy	...
0	0	0	0	0	0	0	0	0	0	0	...
1	1	1	0	0	0	0	0	0	0	1	...
2	1	1	0	0	0	0	0	0	0	0	...
3	1	1	0	0	0	0	0	0	0	0	...
4	1	1	0	0	0	0	0	0	0	0	...

Results

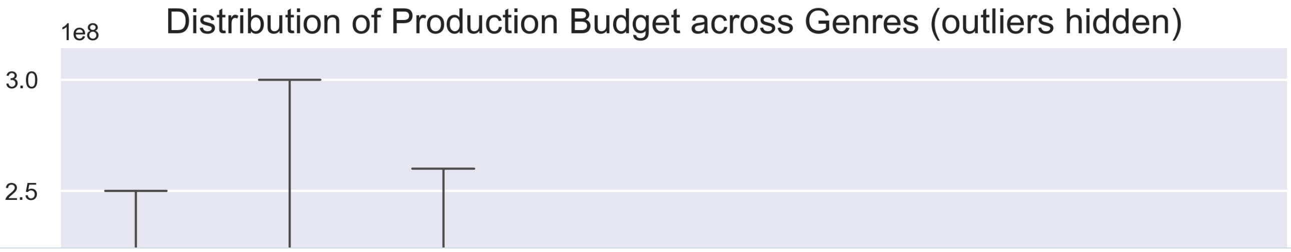
With all of the categorical variables converted into quantitative form I was able to perform my OLS regression with ROI as the dependent variable. Below is a bar plot of the regression coefficients sorted in descending order.



While the R-squared for this model is only 0.043, the results are not insignificant if statistically insignificant. The Horror and Mystery genres are especially interesting when coupling the regression findings with a plot of ROI distributions over genres. In the graph below I picked out the seven biggest movie genres and plotted the ROIs in those genres in a box plot.



The box plot provides support for the genre coefficient weights in the regression model. Horror movies, which are typically coupled with the Thriller tag as well, perform well when the success metric is ROI. A plot of production budget distributions over the same genres helps explain why.



Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- Jupyter Notebook 100.0%