Project 1 Module 1

Introduction

The goal of this project is to gain a deeper understanding of the film industry and helping Microsoft open a film studio. The problem is they are a technology company and need to do detailed analysis in order to create a presentation that explores the larger industry and what films are doing best at the box office. The data includes box office mojo, IMDB, rotten tomatoes and themoviedb.org.

Designing the Experiment

Before beginning the analysis, I established the need to understand the macro metrics of the film industry. I began by looking at the distribution of total gross, profits, domestic, foreign gross and other financial metrics. These key metrics determine success/failure in the industry and require a careful analysis in order to unpack and find what truly drives success for a movie.

The project begins by understanding the distribution of the data. This informs the relationship between observations (such as profit) and can be used for plotting relationships between variables in the dataset. Some visualizations I used here are histograms and distribution plots. More are listed below.

Clearly, all profit, production and total gross metrics were skewed to the left indicating a density of films performing at very average numbers with huge breakaway successes being far and in between but having a huge impact on the overall revenue of top performing studios.

Data cleaning and Processing

Tn_movie_budget:

The data types for key numerical metrics were all loaded as objects. In order to remove any symbols or commas and convert to a $ sign, I used a function I found on r/learnpython that achieves this.

```
valid = '1234567890.'

def clean(data):

    #returns function that clears out nonvalid character

    return int(''.join(filter(lambda char: char in valid, data)))
```

I then assigned the new, int type column to their previous column titles.

I added the columns of Gross profit and profit margin to enable deeper analysis going forward.

Bom_movie_gross:

I had to change the foreign_gross column to a float from an object.

Exploratory Data Analysis

Value Counting Functions

.max() – returns highest value

.min() – returns lowest value

. mean() – returns mean

.value_counts – returns counts of unique values

.groupby() – Splits data into groups based on a specific criteria dependent on available axes. Maps labels to group names and returns a list.

Df.nlargest(n = 100) – returns the largest 100 values of a list

Data Visualization and EDA

sns.regplot()

sns.pairplot()

sns.jointplot()

sns.boxplot()

sns.distplot()

plt.plot()

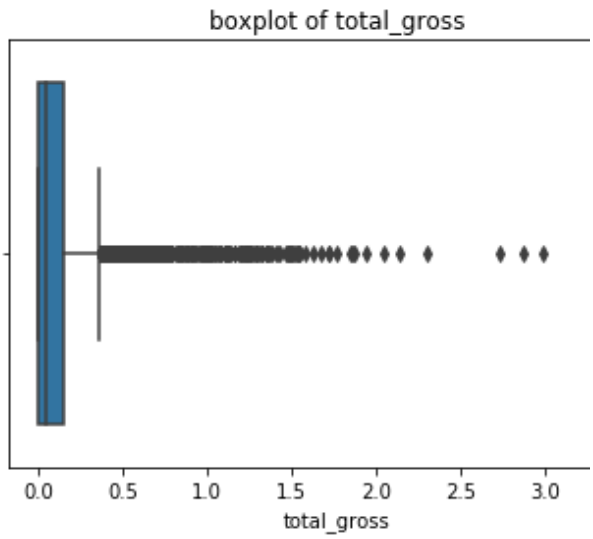
plt.hist()

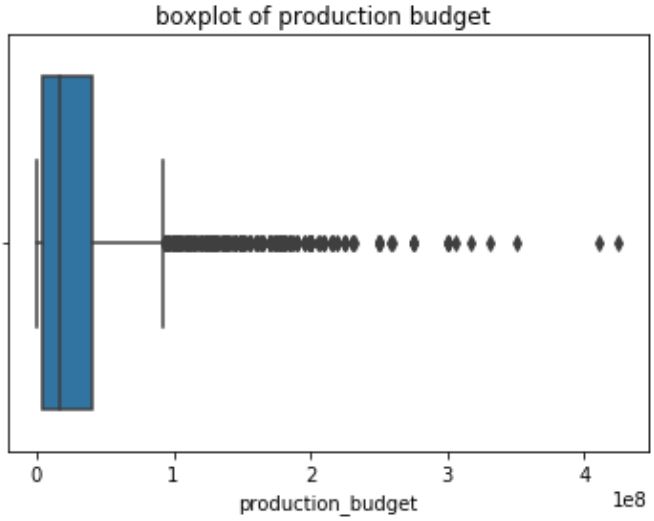
plt.scatter()

Observations

Question 1: What is the distribution of profits and revenue across the film industry and why does it matter? What are their relationships?

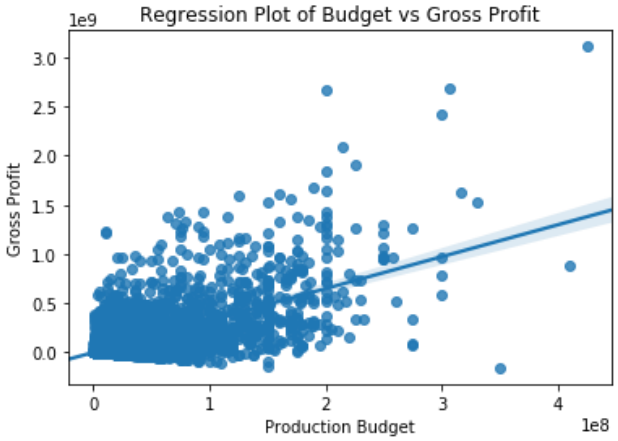
The distribution of the data indicates density around the mean which skews left. This shows that most movies perform to an average rate, do not lose money but are not huge grossing blockbusters either. I used this to inform my process of digging into the top 100 movies further to figure out what makes a film successful.

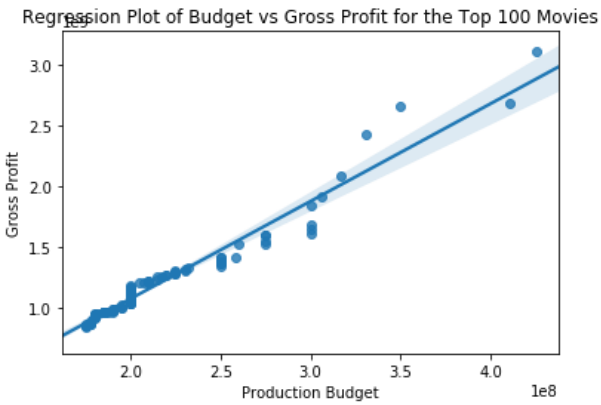




Question 2: Is there a correlation between the top 100 grossing films and their budgets?

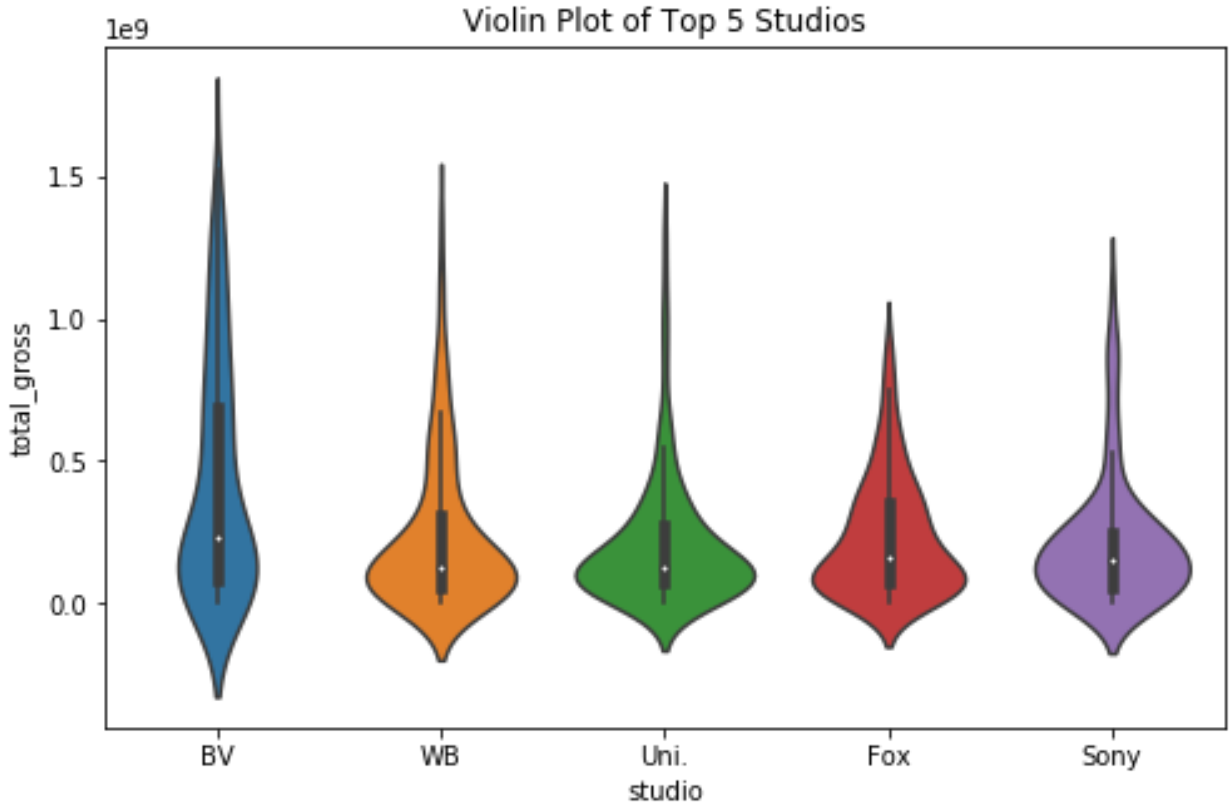
Determining and objective, provable link between budget and gross is not established in this analysis. More factors then budget certainly drive total gross and would require further analysis. However, studios indicate that they are moving towards high budget blockbusters for a big percentage of their films and there is a solid case to be made for high production budgets in the coming years. There is a clear relationship between the production budget of the top 100 films indicating that if you are looking to make a big movie, it is advisable to spend more to earn more. However, this is not a guaranteed success with the top production budgets and top earners being break away successes. My recommendation would be to set a minimal threshold for all films produced to ensure a quality of production competitive with success cases.





Question 3 : What are the most successful studios and their distribution of revenue and profit?

Studios show similar spreads with big winners and lots of diverse outcomes in terms of profits and losses. The top 5 highest performing studios all had a break way success as shown in the violin plot. If you are looking to break into the top 5 studios, a breakaway success that tops box office is necessary. The average top 5 studio produces just under 16 films per year indicating a high output and high gross per output. Disney has stated that they produced 12 films in 2019(a date range outside this dataset and analysis) and are continuing to consolidate this into less movies with the goal of each being a blockbuster.

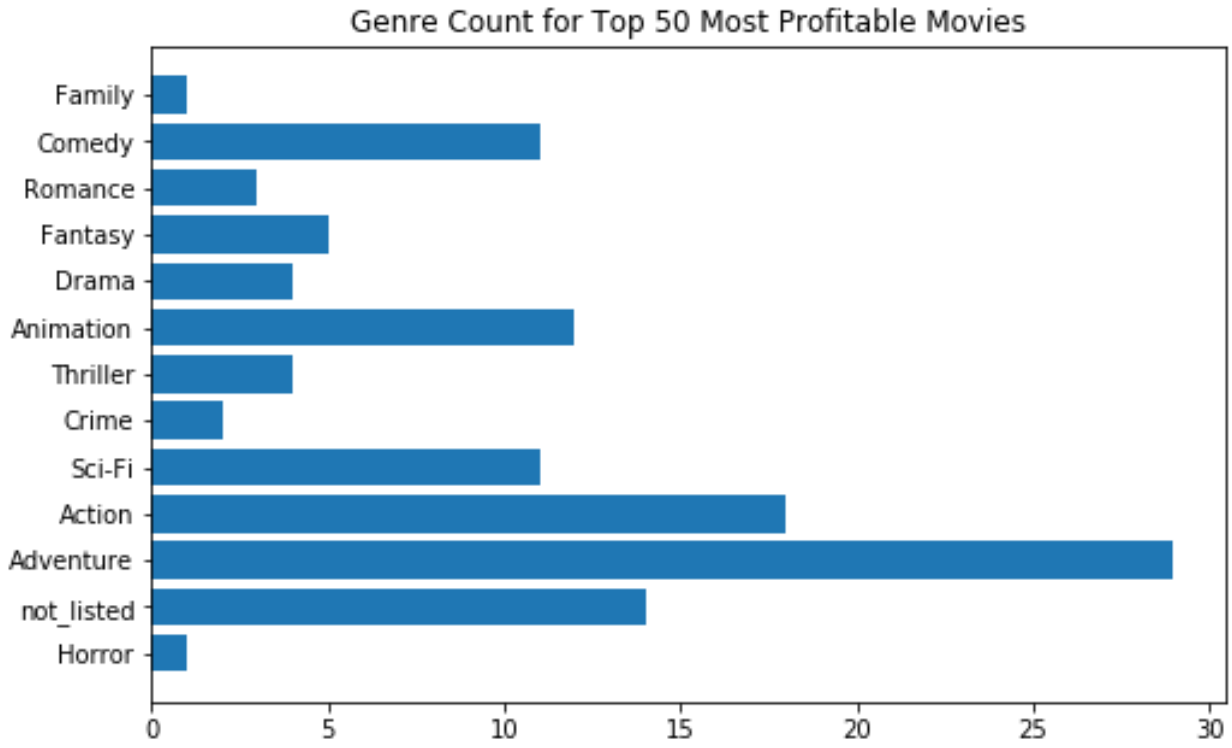


Question 4: What are the top 50 highest grossing movies and genres?

The final concrete insight of this analysis is the huge profitability of the action and adventure genres as indicated in the histogram above. Though multiple genres show strong levels of

profitability, action and adventure top the list. The 3rd highest is 'not listed' indicating further data enrichment is needed to label these films correctly for a more accurate analysis.

Overall, the film industry is not low risk with a high chance of failure even when a large budget is allocated. By looking at certain categories such as action & adventure as well as further analysis into breakaway successes would be a great start to a next phase of analysis. There is a difference in the highest margins and the total grossing movies with the films being in the horror categories. This is an insight that can inform budgeting for movies depending on genre and maximizing low or medium budget movies.



Question 5: How can we maximize our budgets?

Given the disparity between highest margins films and total grossing films, there is definitely a deeper analysis into profit margins. Developing a low cost, high return strategy is a useful endeavor for a studio that is newly launched and looking for quick wins to establish itself. If Microsoft chooses to go low budget, high margins, the category of horror is looking particularly promising.

Total Gross	Highest Margins
Avatar	Deep Throat
Star Wars Ep. VII: The Force Awakens	Paranormal Activity
Titanic	The Blair Witch Project

Conclusion

In conclusion, there are 4 main recommendations from this analysis.

1. The industry has a huge spread of success and failure with most films being clustered around a very average performance in box office. Production budget does initially appeared to be connected to gross profit.
2. The top 5 highest performing studios all had a breakaway success as shown in the violin plot. If you are looking to enter into the top 5 studios, a breakaway success that tops box office is necessary. The average studio in top 5 made 15.97 films per year
3. Adventure and Action genres are the highest performing genres.
4. The highest profit margins are in the adult and horror categories. If you are looking to maximize a small budget, these genres (especially horror!) are worth looking into depending on the target audience. The top 50 profit margins vary significantly from the top 50 total grossing movies.

Next Steps

Time series analysis to analyze the best release dates

Investigate the teams behind the highest grossing movies to potentially recruit them

Analyze ROI per genre to determine what genres have the most longevity

Assess the values of reviews and votes in a film's success.