NON-TECHNICAL PRESENTATION, PHASE 1 PROJECT

OVERVIEW

Microsoft, an influential American multinational technology company, is committed to its core mission of empowering individuals and organizations worldwide to achieve more through relentless innovation. The company's diverse operations span various sectors, including the ubiquitous Windows operating system, the renowned Microsoft Office productivity suite, and a robust presence in the cloud computing realm through Microsoft Azure.

In an era of intense competition, exemplified by tech titans like Amazon venturing into movie studios and streaming services, Microsoft seeks to enhance its competitive edge by investmenting in a movie studio. Leveraging its extensive expertise in data analytics and cloud computing, Microsoft has the potential to amass valuable user data, encompassing preferences and viewing habits. This data could be harnessed for invaluable insights, personalized content recommendations, and precision-targeted advertising.

Furthermore, the allure of streaming services transcends borders, offering a global audience. Microsoft, recognizing this expansive opportunity, may strategically seize the moment to extend its global footprint and amplify brand recognition on a worldwide scale.

BUSINESS UNDERSTANDING

Microsoft has made a strategic decision to venture into the movie production industry. However the company lacks the expertise on film production among them audience preferences and industry dynamics. With this pressing need, microsoft needs to gain contextual awareness of the movie industry landscape and levarage data analysis to inform this venture. The primary objective of this project is to identify and analyze the key factors and characteristics that drive box office success and therefore inform on the type of movies to create. High-performing films consistently generate significant box office revenue, making it essential to explore the elements that influence and determine box office earnings.

DATA UNDERSTANDING

Initial data is from 4 sources,

- Box Office Mojo
  Contains data in csv format. The columns are 5 representing the title(object/ string),studio(an object/string), domestic gross(float) , foreign gross(object) and year(int) of 3386 movies.
  There are 38 null values in the domestic gross column 1350 in the foreign gross and 2 in the studio column. Rows with null values are dropped.
  Change the foreign gross to a float data type.
  The dataset is now clean for analysis
- IMDb.
  A sqlite3 database with 8 columns of principals, persons, known_for, directors, writers, movie_basics, movie_ratings and movie_akas.
- The MovieDB
  Contains data in CSV format. The data has 9 columns and 2518 rows. The columns are genre_ids (object/string), id (int), original_language(object/string), original_title(object),

popularity(float), release_date (object/string), title (object/string), vote_average (float) and vote_count(int).
There are duplicates which are dropped.
The genre_ids has a significant number empty lists records, drop the entire column.
Drop the original title, vote average and vote count columns since they have no use in the analysis.
Rename the release date to year, extracting only the year from the initial release date

- The Numbers
  Contains data in csv format. The data has 6 columns and 5782 rows. The columns are id (int), release_date (string/object), movie (object/string), production_budget (object/string), domestic_gross (object/string), worlwide_gross (object/string).
  The dataset has no values but contains placeholders in the domestic gross and worldwide gross columns. Drop the rows with the placeholder value ($0).
  Change the worldwide and domestic gross columns to a float datatype.
  Rename the movie, worldwifde gross and movie columns to title and foreign gross
  Change the release date to year, while deriving the year from the initial release date to repopulate the column
  The dataset is now ready for analysis

Feature Engineering

Add total revenue column as the sum of domestic gross and foreign gross to the BOM and The Numbers Dataset.
Merge The Numbers DataFrame and IMDb data to create a dataframe which will be used for further analysis e.g Grouping data according to genres
Merging The movie dataset and The numbers dataset to create a dataframe which will be used for further analysis.

DATA ANALYSIS

Univariate analysis

Total Revenue:

Measure of dispersion:

- The mean is 142,399,275
- The median is 54,200,060
- The mode is 16,000,000

The total revenue is a positively skewed distribution. The mean of the data is greater than the median. The mean is pulled towards the right tail while the median is the middle value of the data. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.

Production Budget

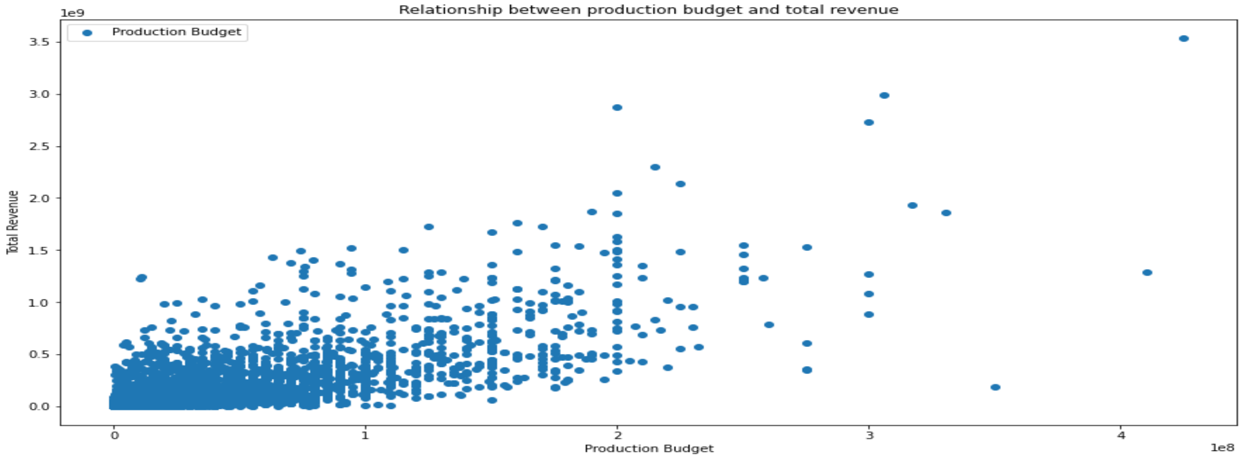
Measure of central tendency:

- The mean is 142,399,275
- The median is 54,200,060
- The mode is 16,000,000

The total revenue is a positively skewed distribution. The mean of the data is greater than the median. The mean is pulled towards the right tail while the median is the middle value of the data. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left.

Bivariate analysis

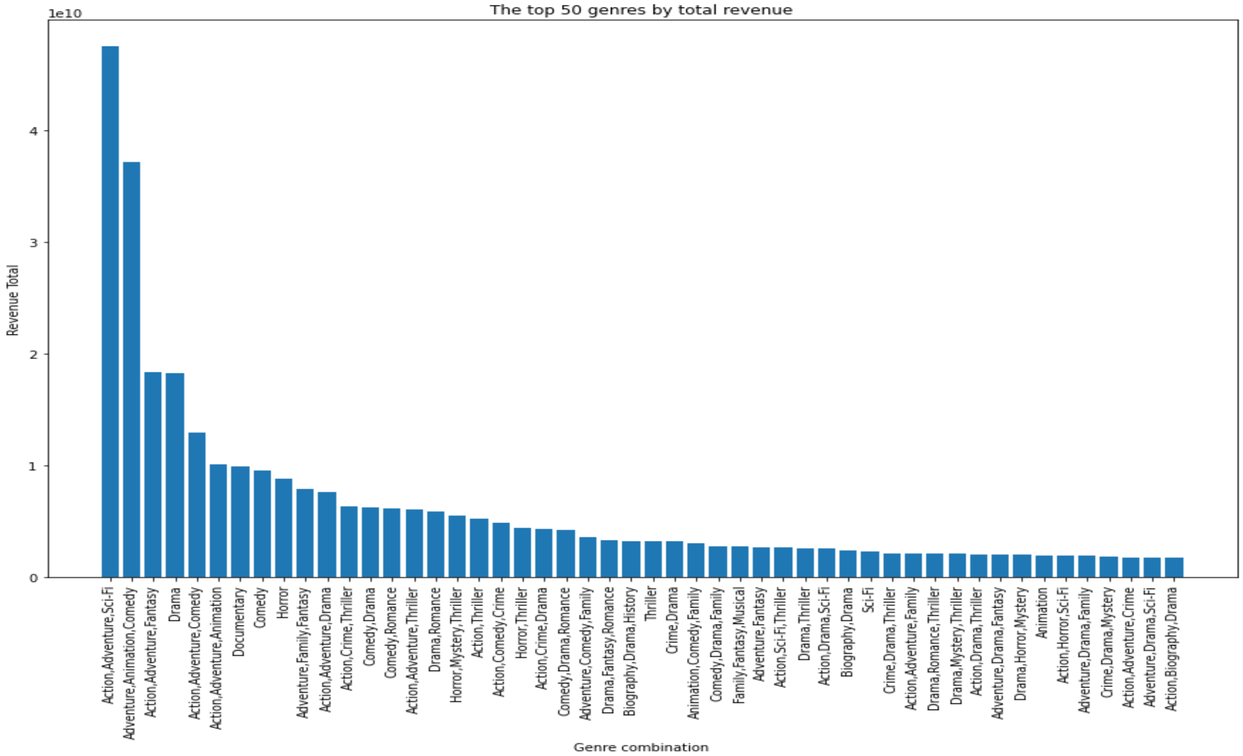
Production budget and total revenue:



There is a strong positive correlation between height and weight. As the production budget increases, total revenue tends to increase. However it is not a perfect relationship.x

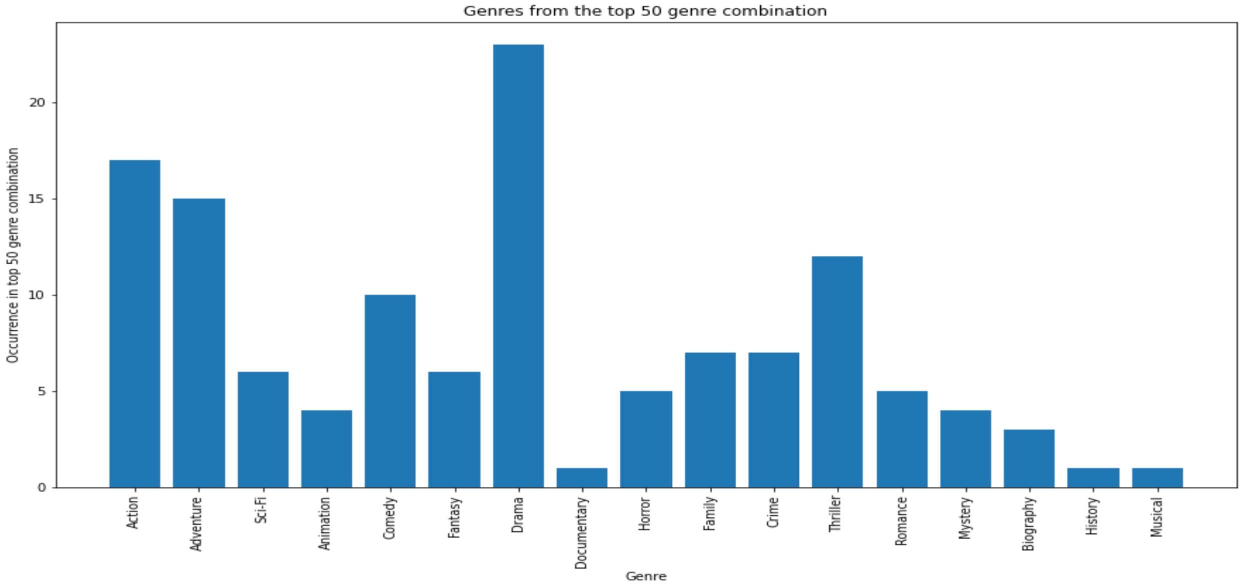
Grouping movies with genres

The top 50 genre categories by total revenue are:



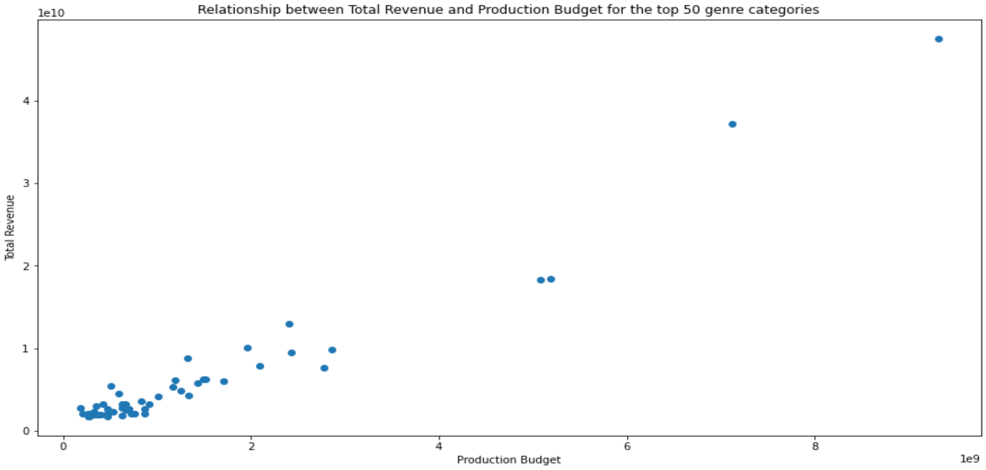
These are genres categories of the movies. They are made up of different genres. These genres may occur more than once in different genre categories.

The occurrence of these genres can be showed as:



Drama genre occurs most, 23 times in the top 50 genre categories. Documentary, History and Musical genres occur only once in the top 50 genre categories.

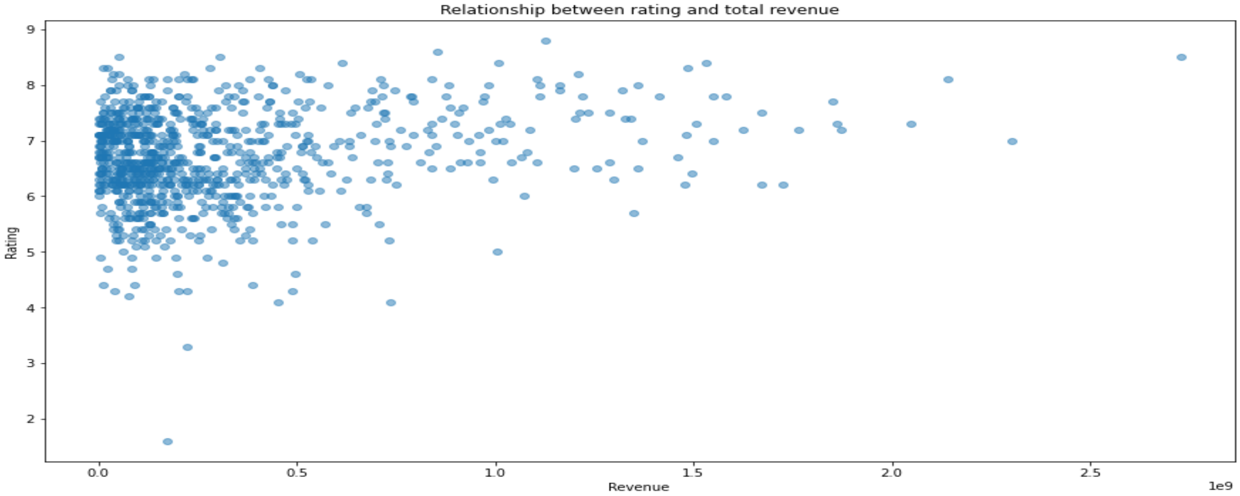
What is the relationship between the total revenue and production budget for the top 50 genre categories?



There is a high positive between the production budget and the total revenue. The relationship here is stronger than when movies are generalised. As the production budget increases, the total revenue increases.

<u>Relationship between IMDb rating and total revenue</u>

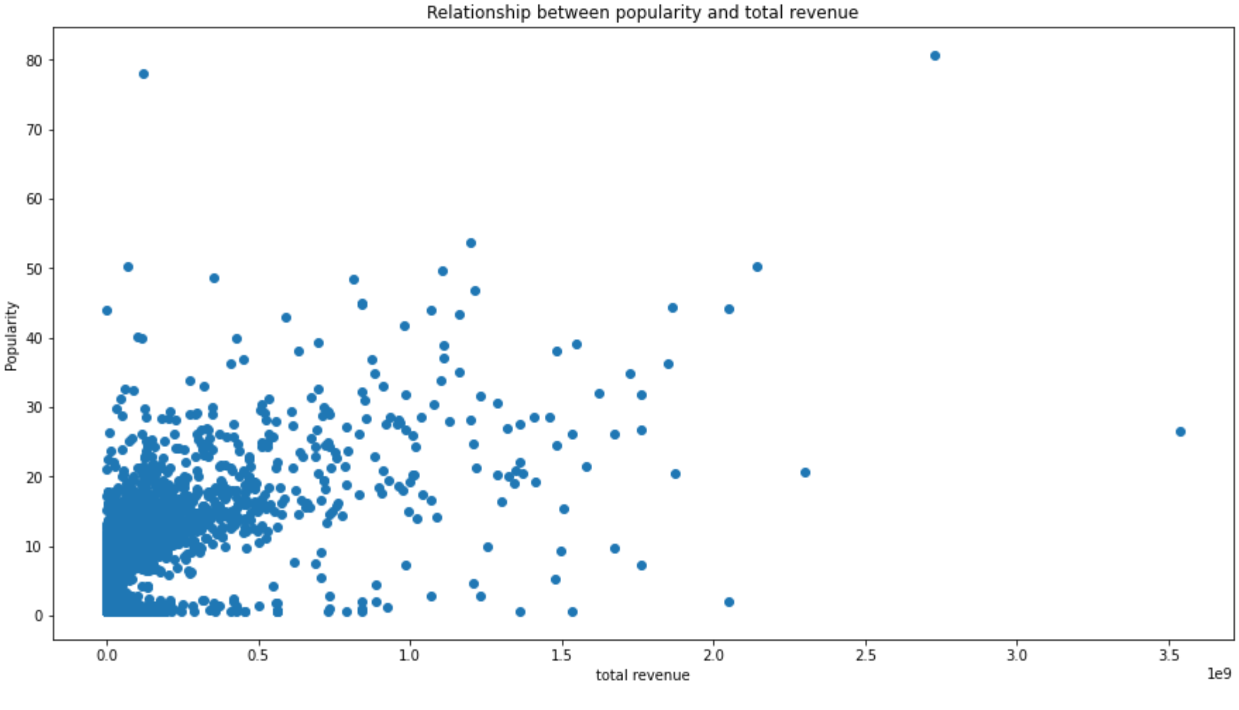
We can explore the potential relationship between movie ratings and total revenue in the IMDb dataset by creating a scatter plot, allowing us to visually analyze any correlation between these two variables.



There is a very weak positive relationship between the IMDb rating and total revenue of the movies. Highly rated movies do not tend to have high revenue.

Relationship between popularity and total revenue

The Movie DB dataset has a column popularity. We could use this column to explore if there is any relationship between the popularity of the movies and the total revenue by each movie



There is a moderate positive relationship between the popularity of a movie and the total revenues. As the popularity increases, the revenue tends to increase even though moderately.

Recomendations

Production Budget

Budget Allocation Strategy - Allocate budgets strategically based on the observed positive correlation. Consider allocating resources to high-budget productions that have the potential generate substantial revenue. However be selective and base your decision on other factors such as genre analysis as the correlation is not perfect.

Genres

Genre Prioritization - Focus resources and efforts on genres that have consistently performed well in terms of total revenue, the genres in the top 50 genre categories. Direct more resources, both financial and creative to projects within these high revenue genres.

Budget optimization Strategy - Given the strong correlation between budget and revenue in the top genres, consider allocating budgets strategically to projects with these genres to maximise returns.

Risk Management - A near perfect relationship indicates a lower level of risk in budget allocation for these genres. It suggests well funded project in the top genres are likely to yield more positive returns.

Popularity

Popularity driven marketing - Invest in effective marketing strategies that enhance the popularity of movies. Engage with the target audience through strategies like promotions, social media and interactive campaigns to boost anticipation and interest.

Ratings

Quality vs Commercial Appeal - While high IMdb ratings are important for prestige and reputation, recognize that ratings do not correlate with revenue. Diversify your movie portfolio to include high rated projects as well as commercially driven projects.

NEXT STEPS

Cast - Further analysis could be carried out to analyse if there is any relationship between the cast of a movie and its revenue. Do the A tier directors and actors have any correlation to the total revenue of a film?

Genres Trend - What is the genre trend over the years? What are the genres that are trendung now? What are the emerging genres?