MOVIE/FILM SUGGESTION PROJECT

1. BUSINESS UNDERSTANDING

Link to GitHub Repository: [GitHub Repo]

1.1. UNDERSTANDING THE PROBLEM

The cinema of the United States, often generally referred to as Hollywood, has had a profound effect on cinema across the world since the early 20th century.

Film entertainment is big business in the United States and it was estimated that the film entertainment business generated $35.3B in revenue in 2019 . The United States is among the biggest film industries in the world in general but also in terms of tickets sold per year, ranking behind China and India.

Microsoft sees all the big companies creating original video content, and they want to get in on the fun. They have decided to create a new movie studio, but the problem is they don't know anything about creating movies. They have hired Flatiron Data Science team to help them better understand the movie industry. Our team is charged with doing data analysis and creating a presentation that explores what type of films are currently doing the best at the box office. We will then translate those findings into actionable insights that the CEO can use when deciding what type of films they should be creating.

In this report, we will investigate factors associated with commercially successful movies. This report gives a comprehensive evaluation on factors influencing the box office success of a movie such as production budget, domestic gross, international gross, worldwide gross, genre etc

1.2. PROBLEM STATEMENT

Microsoft sees all the big companies creating original video content, and they want to get in on the fun. They have decided to create a new movie studio, but the problem is they don’t know anything about creating movies. They have hired Flatiron Data Science team to help them better understand the movie industry. Our team is charged with doing data analysis and creating a presentation that explores what type of films are currently doing the best at the box office. We will then translate those findings into actionable insights that the CEO can use when deciding what type of films they should be creating.

The Problem statement is to analyze these datasets in order to help this company create movie films that will thrive in the market and generate income.With the help of our analysis questions; we will be able to enlighten microsoft on a direction/path to follow while starting their film studio.

2. DATA UNDERSTANDING

2.1. DATA COLLECTION

This project contains dataset from various credible sources:Box office,IMDB,
welll use 2 datasets the info and budget dataset
We will be using ratings,genre, director column in the first dataset while in the 2nd one we will use all columns except the release date columns our data selection: we ll use 2 datasets the info and budget dataset
We will be using ratings,genre, director column in the first dataset while in the 2nd one we will use all columns except the release date column.

2.2. DATA DESCRIPTION

This project contains 2 dataframes: rt.movie_info.tsv and tn.movie_budgets.csv.

Rt.movie_info.tsv:
The dataset contains 1560 rows and 12 columns, and the entries are all object data types except the column id which is an integer. The columns that we will be investigating are the ratings, genre director.
Ratings: These are basically measures of films based on suitability, for example adult and children categories. Our data has been categorized into 6 ratings:

- **R(Restricted):** This rating is for films specifically designed to be viewed by adults and therefore may be unsuitable for children under 17.
- **G(General guidance)**: This program is designed to be appropriate for all ages. This rating indicates a film contains nothing that would offend parents for viewing by children
- **PG(Parental Guidance):**This program is designed to be appropriate for all ages. This rating indicates a film contains nothing that would offend parents for viewing by children
- **NC-17 (Clearly Adult)**:This rating is applied to films the MPAA believes most parents will consider inappropriate for children 17 and under. It indicates only that adult content is more intense than in an R rated movie.
- **PG-13(Parents Strongly Cautioned)**:Some material may not be suited for children under age 13. May contain violence, nudity, sensuality, language, adult activities or other elements beyond a PG rating, but doesn’t reach the restricted R category.
- **NR(Not rated)**: this means the films is not rated

Genre:is a stylistic or thematic category for motion pictures based on similarities either in the narrative elements, aesthetic approach, or the emotional response to the film
We have drama,action,thriller, comedy and sometimes a combination of the genres
Write and director: there are the people who write and direct those films respectively.

Tn.movie_budgets.csv.
The dataset contains 1582 rows and 6 columns, and the entries are all sring data types
This dataset contains:
Movie: The movie

Release_date: The date the movie was released to the user viewing
Production_budget: This are all the expenses incurred during the production stages of the movie,
Domestic_gross: This means the revenue generated from a movie at their domestic country/area
Worldwide gross: means all the revenue that was generated around the wold for a particular movie.
Net income: We subtracted the production budget from the domestic gross,

2.3 TARGET POPULATION

The target population Microsoft company, its stakeholders, CEOs, managers and their esteemed customers

2.4. DESCRIBING THE QUESTION

1 What kind of movie contents in terms of genre perform the best?

2 What film rating has the highest count?

3 Does budgetary allocations affect the gross income?

4 Does movie director affect the rating/ views of a movie film?

2.4.3. EXPERIMENTAL DESIGN

1. Loading Datasets and Preparing the Data.
2. Data Cleaning to deal with Anomalies and Outliers.
3. Exploratory Data Analysis (Univariate and Bivariate Analysis).

4. Hypothesis Testing to Implement the Solution.
5. Conclusions and Recommendation.

3. DATA PREPARATION

3.1**. SELECTING DATA**

In our data selection: we ll use 2 datasets the info and budget datasetDATA

CLEANING

- In both df we checked for duplicates, missing data and the data types:
- In the info dataset we replaced the null values with missing, then dropped the columns we were not going to use,
- In the budget dataset we cleaned the $ in the columns, dropped the columns we were not using, changed the object data_types to floats and integers and added another column
- The data was also found to be consistent there being no duplicated data.

We will be using ratings,genre, director column in the first dataset while in the 2nd one we will use all columns except the release date column

3.2.

4. DATA ANALYSIS

4.1. EXPLORATORY DATA ANALYSIS

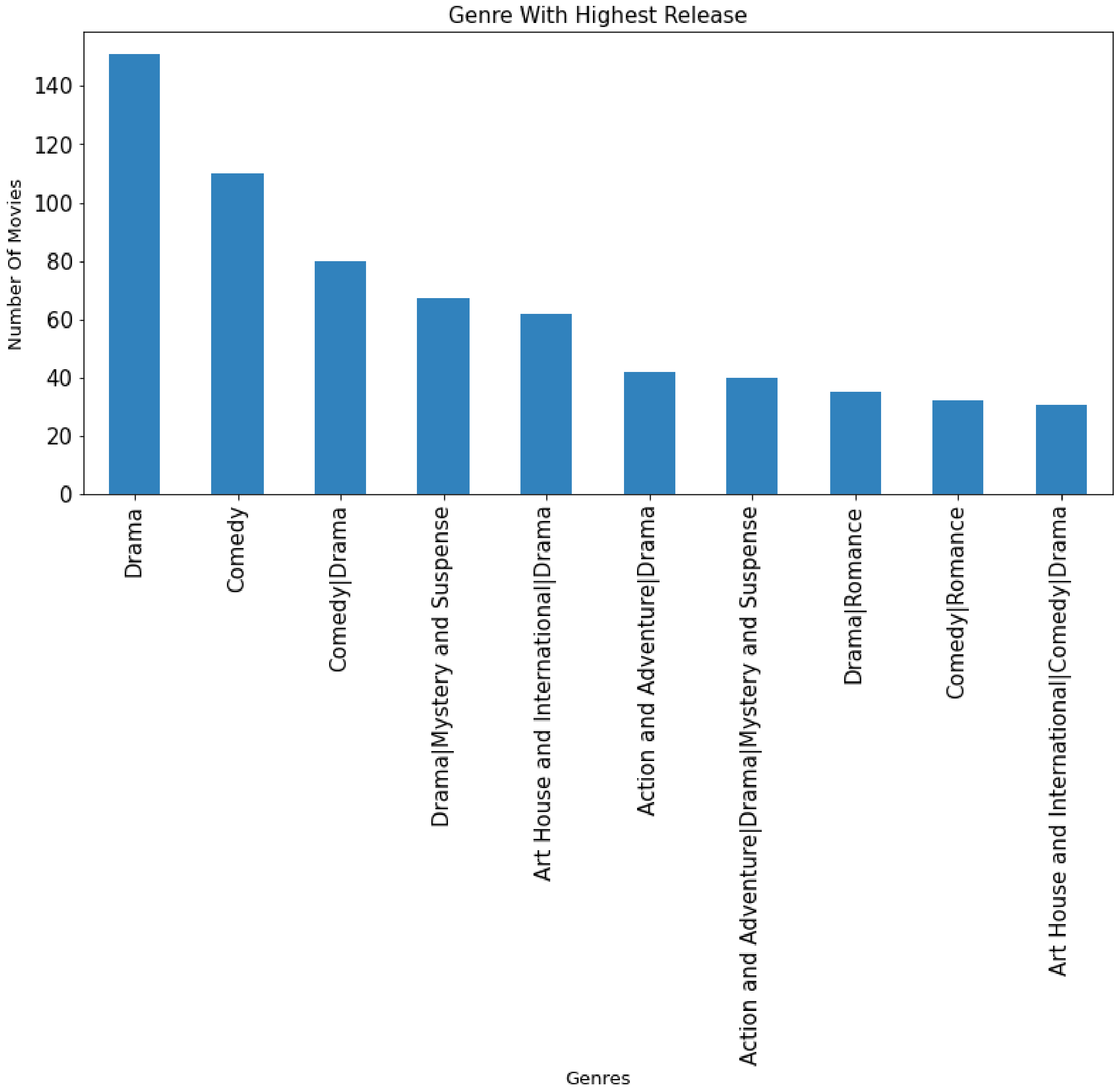
After discussing the structure of the data and cleaning the data, we have to visualize our data so that it can help us answer our questions.

DESCRIBING THE QUESTION

The first 3 questions we used the first dataset:

1 What kind of movie contents in terms of genre perform the best?

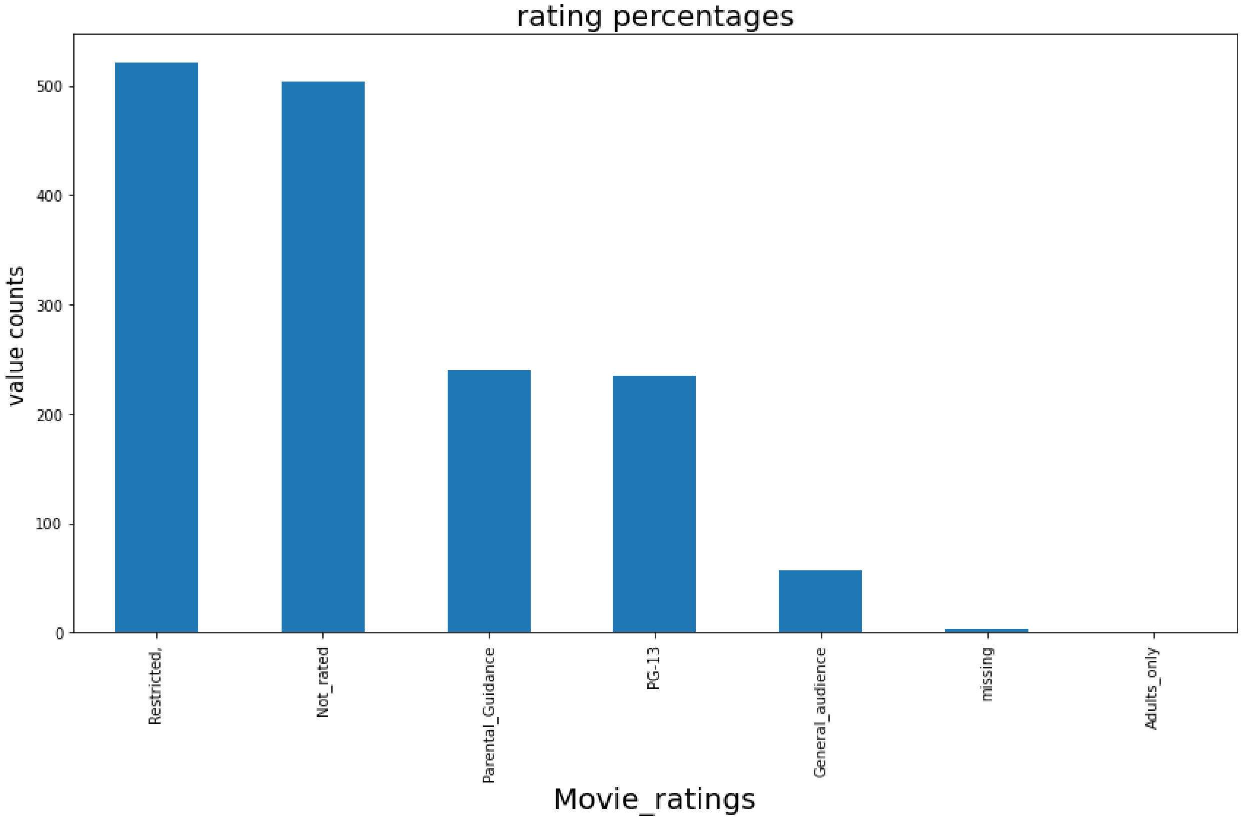
We did a value counts of the movie genres and found out that the most released movies genre was Drama, with 151 counts followed byComedy with 110 then a combination of both comedy and drama with 80.as shown below:



We see that people like drama as drama is the most prevalent, followed by Comedy, Comedy/drama.

2 What film rating has the highest count?

We did a value count and found out that the most watched/released film rating is the restricted movies, Adult films are more prevalent films, which means more adult films are being consumed and produced: Restricted 521, However; we can see that Restricted for adults only and Not rated have small difference,with a percentage of 33.34%,32.2% that's around 1.14%.



3 Does a movie director affect the rating/ views of a movie film?

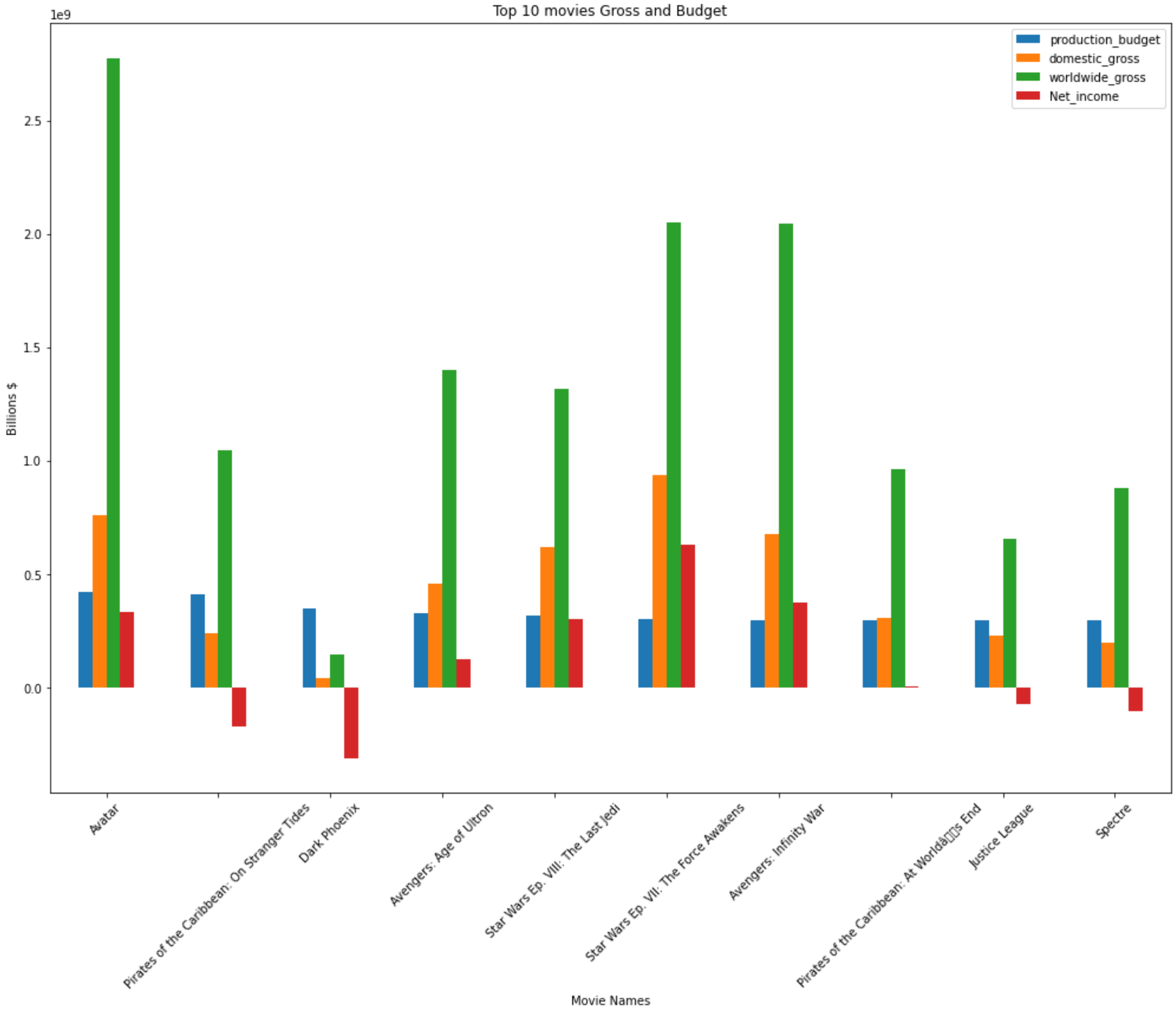
In our director column, we encountered a lot of missing values around 199 missing directors of the films.

Steven Spielberg has directed a ton of films , followed by clint Eastwood and william friedkin with a total of 4 films

4 Does budgetary allocations affect the gross income?

In this question, we used the budget dataframe, we did a value count of movies with the highest production budget, and found out that Avatar was leading with a production budget of $425000000 followed by Pirates of the Caribbean and Dark Phoenix,

We plotted a graph to see the relationship between production budget, domestic gross, worldwide gross and net income as shown below



5. CONCLUSION

Drama is the most popular genre, followed by action, comedy and drama/comedy.

'Avatar', 'Star Wars' and 'Titanic' are the most profitable movies.

Revenue is directly connected to the budget.

Movies with higher budgets have shown a corresponding increase in the revenues.

6. RECOMMENDATION

The test suggests that the best kind of film to venture into is drama and comedy,
most studios have invested in comedy, The ratings to consider is R rating
WE should invest more in budget in order to get high returns