Nzula Priscilla Malombe

Part Time

5th November 2023

Instructor Name: Samwel Jane

Outline

- **Business Problem**
- **Data Understanding**
- **Data cleaning and Preparation**
- **Data Exploration and Analysis**
- **Methods**
- **Data Visualization Analysis**

Business Problem

- 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 they don’t know anything about creating movies.
- Explore what types of films are currently doing the best at the box office.
- Translate the findings into actionable insights that the head of Microsoft's new movie studio can use to help decide what type of films to create

Data Understanding

- The data is contained in a zipped Data file which comprises of various movies datasets. We will therefore work with three data files;
- ***1.imdb.title.basics***
- ***2.imdb.title.ratings***
- ***3.bom.movie_gross***
- The 'title.basics.csv file contains 146144 rows and 6 columns which represent the basic information about the movies. Like the genre, title and the start year.
- The title.ratings.csv file contains 73856 rows and 3 columns which show the average rating of the movies and the number of votes casted.
- The movie_gross.csv file contains 3387 rows and 5 columns which basically shows the income generated by various movies. It shows income made both domestically and foreign.
- We will focus on the following features;
- averagerating, numvotes, domestic_gross, genres and foreign_gross to get the type of films that are doing the best.

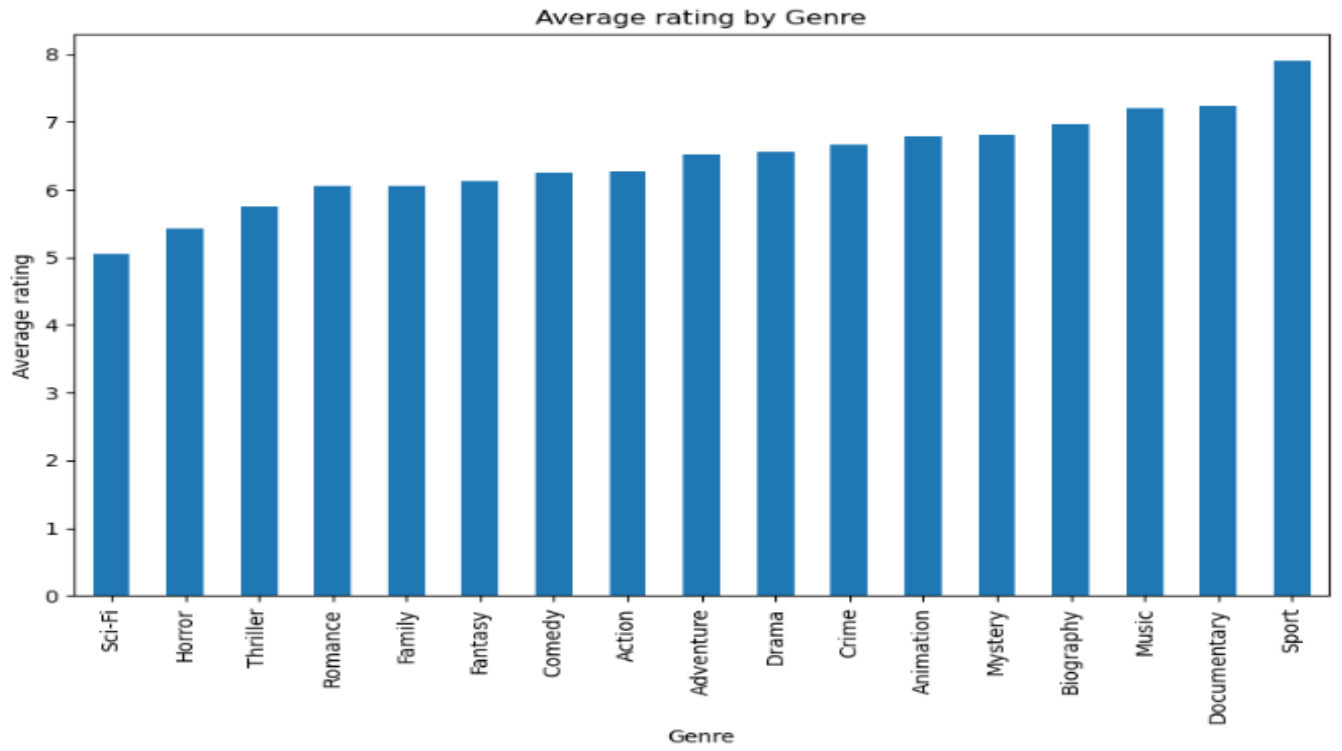
Data Cleaning and Preparation

- We handle the missing values in movies and basics data files.
- The columns domestic_gross and foreign_gross are important for analysis and therefore can not be dropped.
- Fill in the missing values with zero using .replace function.
- Also combine the two columns to obtain total_gross_income column.
- We use .drop function to do away with unnecessary columns.
- Merge the three data frames to obtain one and rename it df_rating_basics_movies for easy analysis.
- Also split the column genres using .split function.

Data Exploration and Analysis

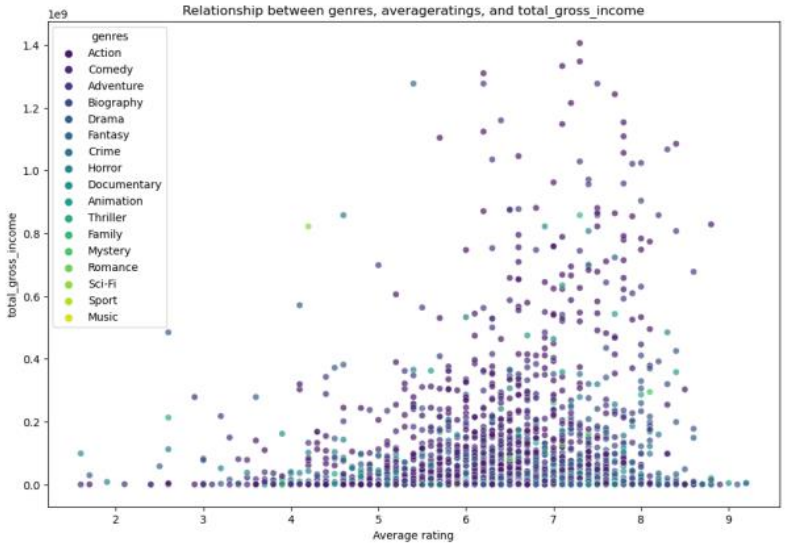
- Use .groupby, .sort and aggregate functions to analyse the data.
- Explore the types of genres with their corresponding total income, average rating and the number of votes.
- The most popular genre being the one with the highest number of votes as per the rating data recorded.
- The best genres being the ones with the highest average rating.
- The most profitable being the one that has the highest total gross income generated.

Methods
Average rating vs Genre



As per the bar above the Sport and Adventure genres have the highest average ratings while Sci-Fi genres has the lowest average rating.

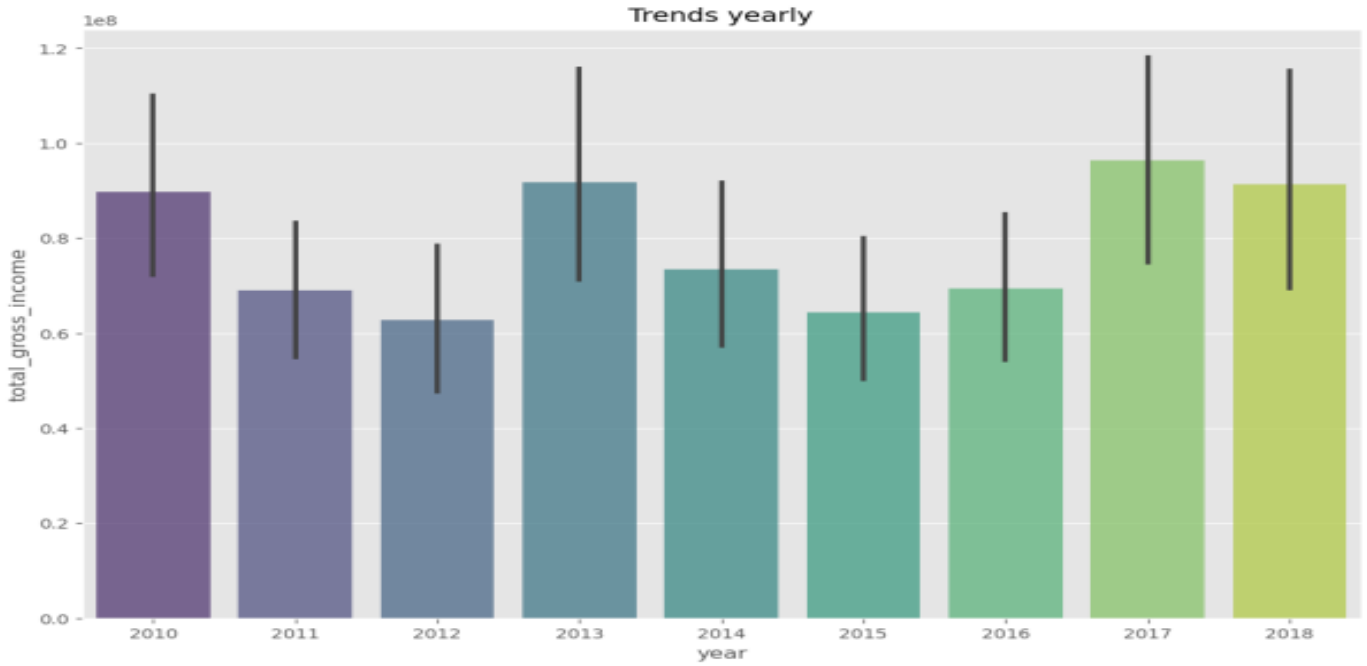
Average rating vs Total gross income



From the scatterplot above,the adventure, documentary ,drama, comedy and action genres have the highest rating.

While action and adventure having the highest total gross income.

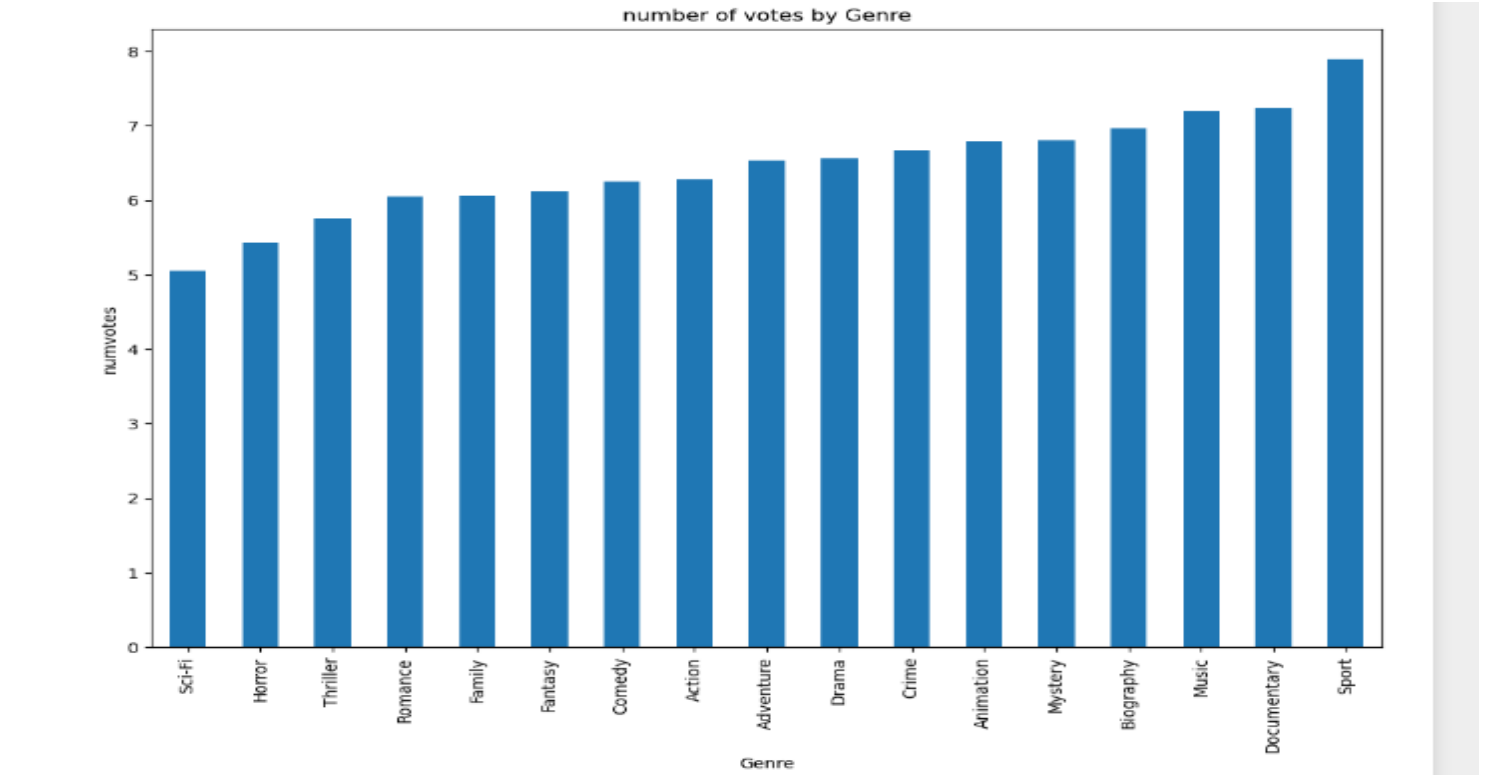
Trends yearly



From the above boxplot we note that the total gross income fluctuates with time.

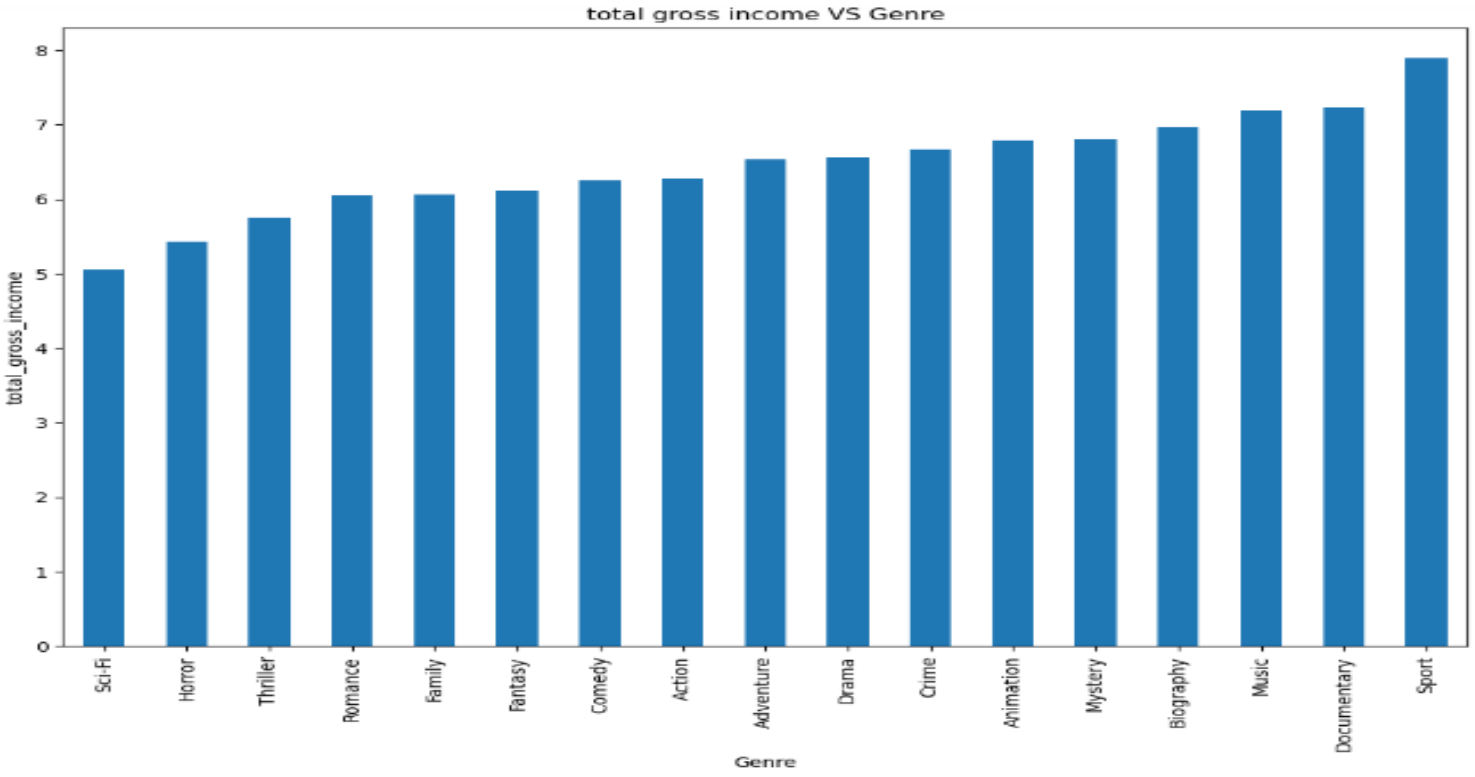
One genre could be do well in a certain year and also perform poorly in another year.

Numvotes vs Genre



From the above bar mystery, action, adventure, crime and biography genres have the highes number of votes.

Total gross income vs Genre



From the above bar chart, mystery, action, adventure, crime and biography have the highest total gross income.

Data Visualization Analysis

- We visualize the films data to help in strategic analysis by making the data more understable and easy to discover various trends.
- Using matplotlib.pyplot as plt, seaborn as sns and %matplotlib inline libraries. We plot bar graphs and scatterplots that show the trends of various genres.
- The bar graphs shows the relationship between genres, average rating and number of votes.
- The scatterplot show the relationship between genres, total gross income and average rating.
- The bar and scatterplots help in communicating insights to the Microsoft stakeholders to help them venture in the right type of films.

Findings

- From the bars and scatterplot above , we observe that ;
- The Sport and Adventure genres have the highest average ratings while Sci-Fi genres has the lowest average rating.
- Most people seem love the Adventure and Sport genres .
- The adventure, documentary ,drama, comedy and action genres have the highest rating.
- While action, adventure, animation, biography and comedy having the highest total gross income. To mean they are more profitable to venture in.
- The mystery, action, adventure, crime and biography genres have the highest number of votes.
- With the family and sport having no votes at all.

Conclusion

- From the analysis above, the Microsoft's new studio can create the Adventure, Action, biography , crime and mystery genres . They are the first five genres whose total gross income is very high. And by creating these genres Microsoft is can make good profits both internationally and locally.
- Also, the mystery ,action ,adventure, crime and biography genres have the highest votes. Meaning they are also preferable by the targeted clientele.
- For additional genres the stakeholders can include drama and comedy genres. They have a positive good rating.