Movie data analysis

Lynette Nkirote Data science Phase one project
Presentation

Business understanding

Microsoft are trying to create a new movie studio as they want to get into the industry, but they don't understand anything about creating movies.

How do we assist them to explore what types of films are currently doing the best in the box office?

we are going to use three datasets i.e.

- imdb.title.basics
- imdb.title.ratings
- bom.movie_gross

This datasets will help us to answer the following questions?

- What type of data does this datasets have?
- Does the dataset have missing values that are necessary or unnecessary
- How can we analyse this data to understand which is the top genres in the movie industry?
- Which is the top genre in most of the years?
- What is the total domestic gross used in the movie industry?
- The average rating of each top genre?
- The Frequency of the domestic gross for the top 10 studios
- And the distribution of runtime in minutes

Loading and data understanding

- Loaded the necessary libraries i.e. pandas and numpy
- Then proceeded to load the csv files i.e. movie_gross, title_basics and title _ratings
- The movie_gross dataset has 33787 rows and 4 columns where float data type is 1 and integer data type is 1 and object data type is 2
- The title_basics dataset has 146144 rows and 5 columns, where float datatype is 1,integer datatype is 1 and object data type is 4
- The title_ratings dataset has 73856 rows and 3 columns, where float datatype is 1, integer datatype is 1 and object datatype is 1

Data Cleaning

- We checked for missing values that are in each dataset to make a decision of whether they are necessary or unnecessary
- We noticed that movie_gross and title_basics dataset had missing values and we therefore decided to drop them
- title basics from 146144 rows to 112232 rows
- Movie_gross from 3387 rows to 2007 rows

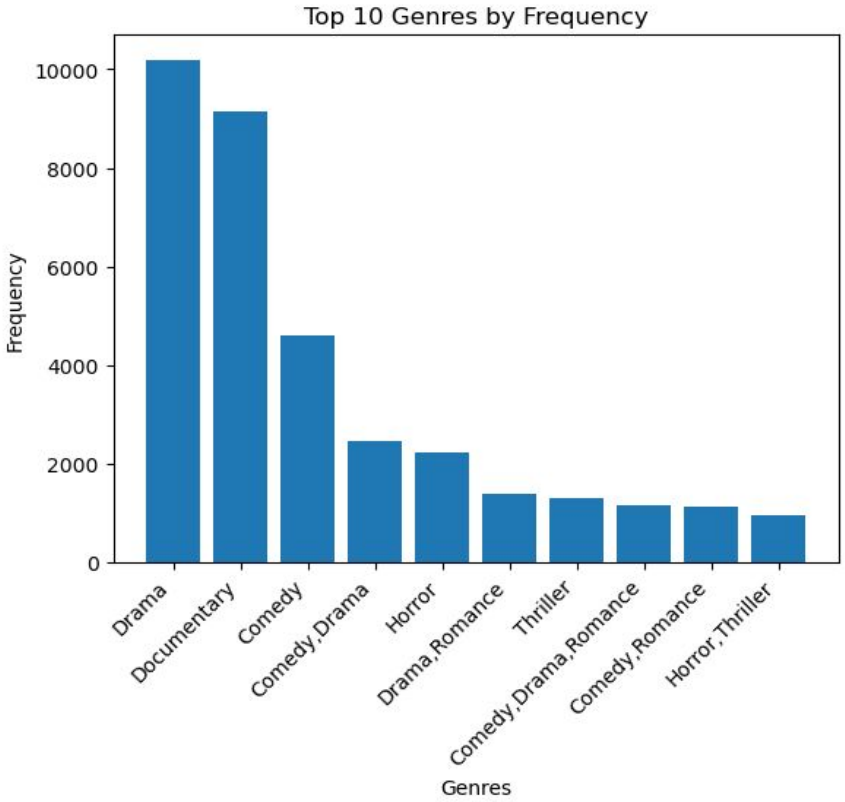
Merging Datasets

Since we are using 3 datasets it will be good to have a look and see if we can merge the datasets.

After reviewing each column in the 3 datasets, the title_basics and title_ratings dataset contained a primary key i.e. tconst

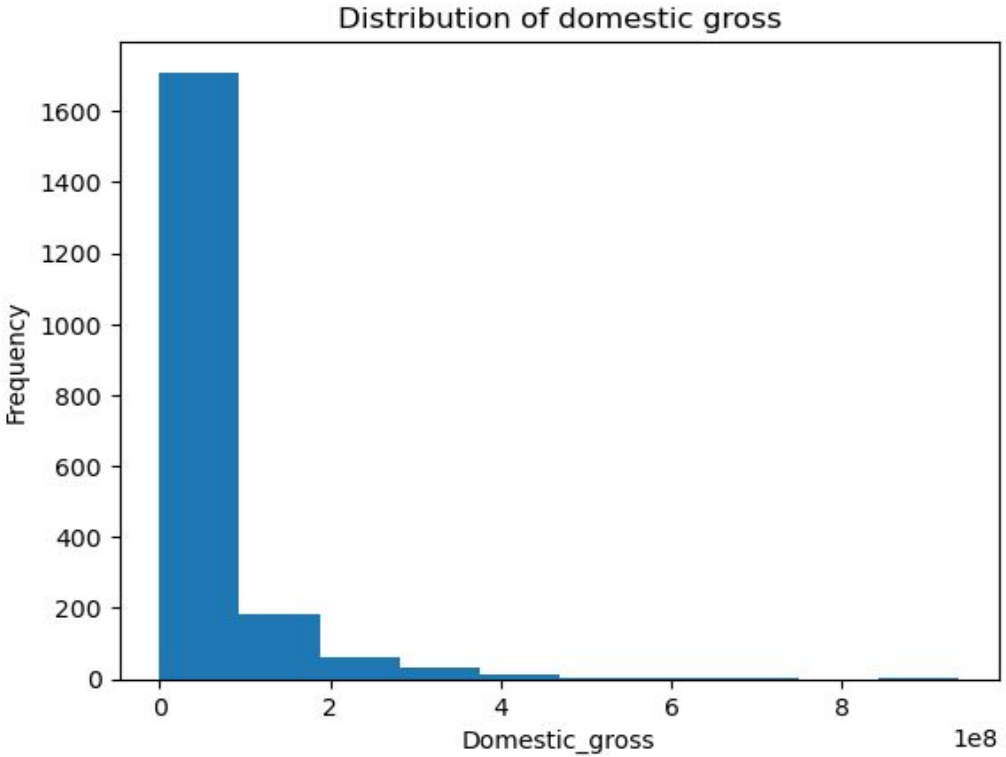
With the 2 datasets have a common column we therefore decided to merge the dataset and use it as one and named it merged_titles.

Top 10 Genres by frequency



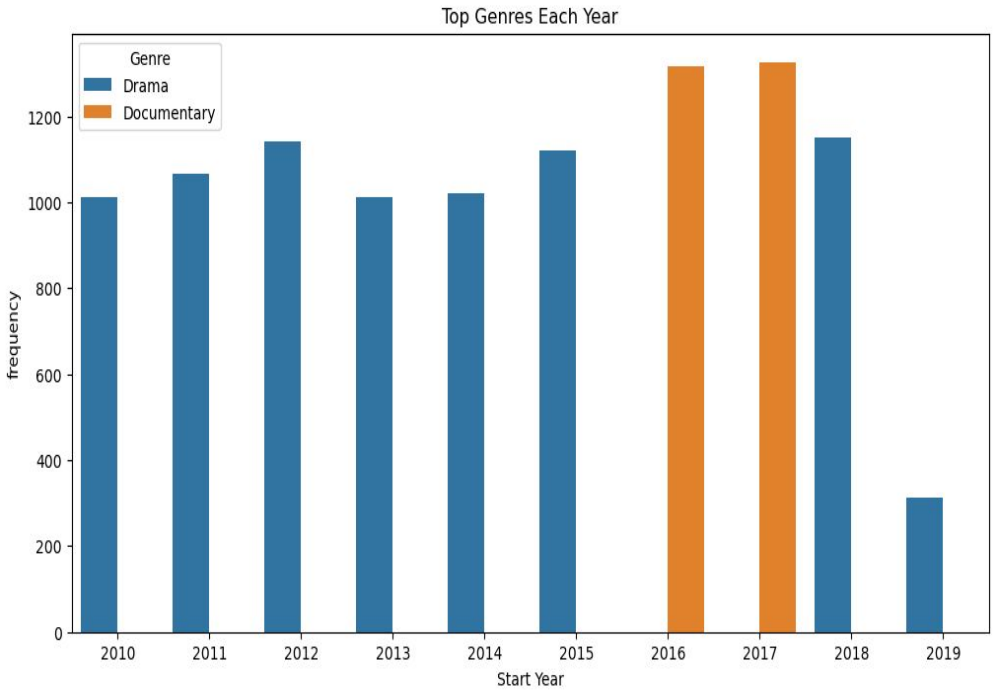
- By loading the visualisation libraries i.e. matplotlib and seaborn we will be looking at some visualisation that will help analyse the data
- The plot shown of the top 10 genres shows the frequency of the top 10 genres
- From the plot we can see that Drama is the highest.

Distribution of domestic gross



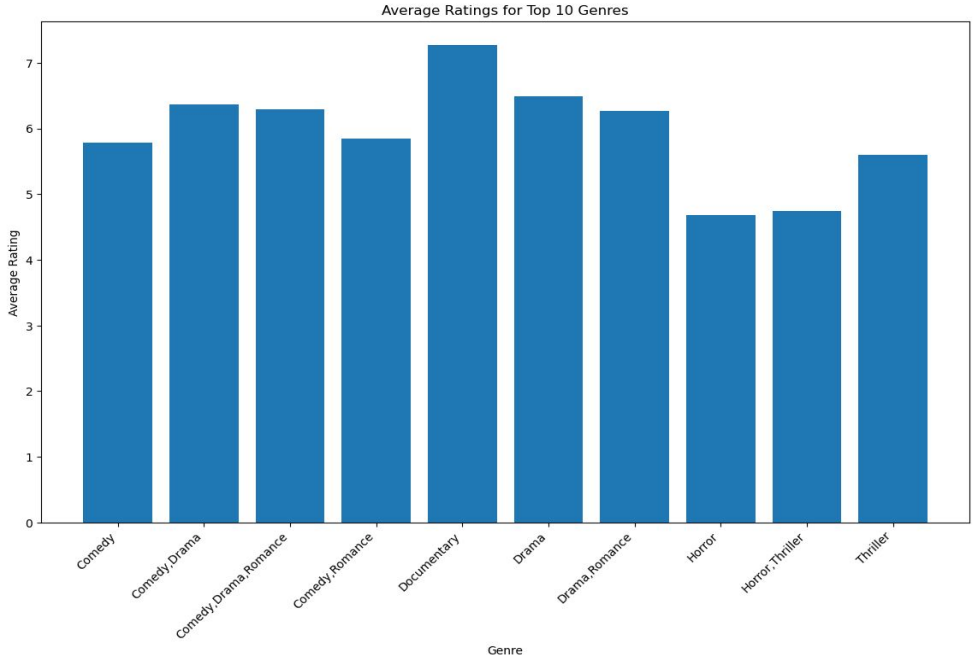
- We used the histogram plot to be able to understand the distribution of the domestic gross
- From the above plot we can say that the plot is positively skewed this means that the median is lower than the mean.

Top genres each start year



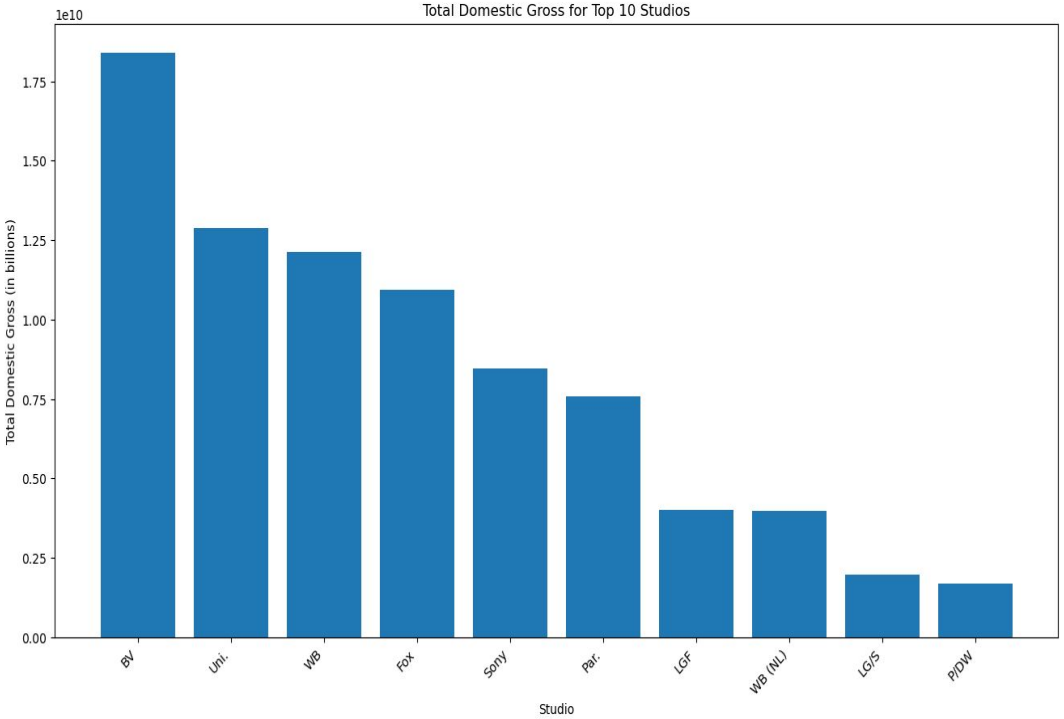
- We went on to check the top genres in each start year to understand which is top genre that is constant in all the start years
- The bar graph shows that drama is the top genre in most start years even in the latest year of 2019
- While Documentary led in 2016 & and 2017

Average rating for top 10 Genres



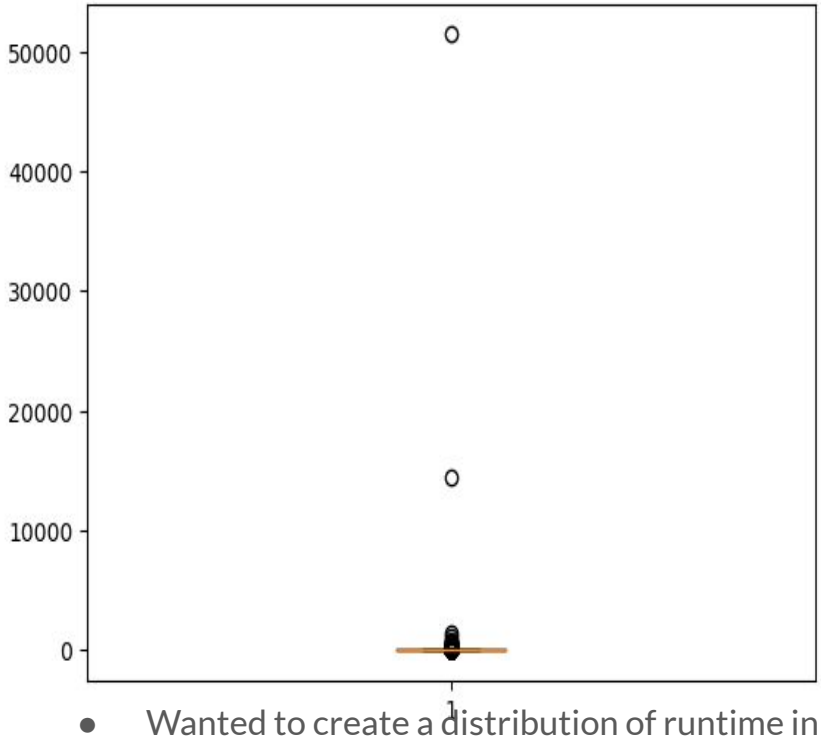
- After knowing the top 10 genres, we look at the mean average rating for the top genres to understand which one has good rating
- From the bar graph, it is clear to say that **Documentary** is the highest in rating followed by **Drama,** the least rating came from thriller movies.

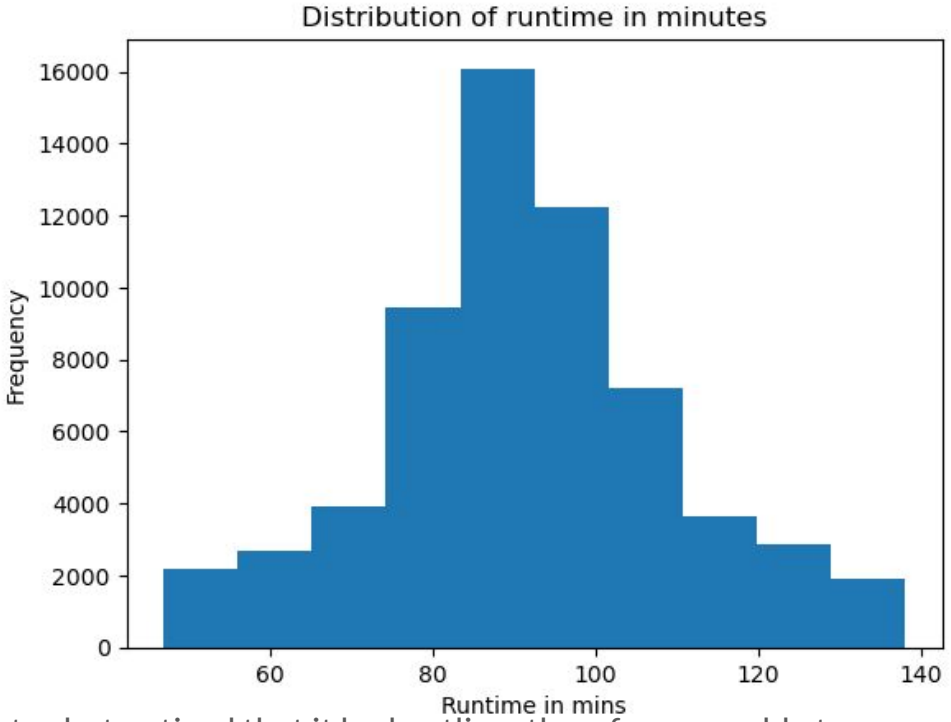
Top domestic gross amount for top 10 studios



- We want to check the top highest studio in terms of the amount of domestic gross they have made from movies.
- From the data above we can say that BV studios is the highest in the total domestic gross this means they produce good rating shows

Handling outliers and distribution of runtime in minutes





- Wanted to create a distribution of runtime in minutes but noticed that it had outliers therefore was able to identify and remove the outliers.
- After removing the outliers,from the above histogram it shows that the output is symmetrical since the mean and median are in the same range

Business conclusions

From the visualisation created this are the business recommendation that would forward to Microsoft

- Drama has a genre would be the best movie to create since we can see that it's the top most genre overly and also each of the start years, although in terms of average rating it was the 2nd highest while documentary as a genre had the highest average rating
- The domestic_gross was highest in the BV studio although the median of the domestic gross was the highest which made the data positively skewed, this should assist the microsoft team to know what is the expectation of the domestic gross
- The distribution of runtime should also help the microsoft team to also know how long the should make the move in minutes which from the analysis it shows that the distribution is symmetrical and the mean and median are in the same range
- From this findings the microsoft team can be able to have a way forward on which genre and how long the movie should take and also the amount of money they can make from the movie so as to plan on a budget.