ANALYSIS OF BEST PERFORMING MOVIES CURRENTLY

Author:

Monicah Iwagit Okodoi

Client:

Microsoft

Project Overview

Seeing all big companies creating original video content, Microsoft wants to also join and have decided to create a new movie studio, but do not have the knowledge about virtual video creation. I have been assigned the task by Microsoft to figure out what are the measures that they are going to take for them to venture in this field. I was provided with several data files for the task, to analyze and give the head of Microsoft's new movie studio recommendations based on my findings to succeed in the field of movie creation.

Business Problem

Microsoft as a company wants to start on creating original video content but do not have enough knowledge about movie creation to move forward with their plan.

Objectives

Microsoft has the following objectives:

- Finding which genres of the movies perform well in the dataset to receive the most public attention.
- Determining the best time to release a movie.
- Which director is associated with the most popular genre?

Using several data frames read from the Box Office it helped in discovering patterns and relationships in the data in order to make better business decisions.

Data mining will aid in spotting movie trends depending on various attributes, develop smarter methods for movie creation and accurately predict the movie performance.

METHOD:

CRISP DM

I will be following the CRISP DM process for this task

The **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (*CRISP-DM*) is a process model that serves as the base for a data science process. It has six sequential phases:

1. Business understanding – What does the business need?
2. Data understanding – What data do we have / need? Is it clean?
3. Data preparation – How do we organize the data for modeling?
4. Modeling – What modeling techniques should we apply?
5. Evaluation – Which model best meets the business objectives?
6. Deployment – How do stakeholders access the results?

Data and Analysis Overview

In this analysis, I will perform an analysis on large data sets containing different types of movies. The data includes many different types of information about each movie, ranging from the release date, the director, the studio, average rating, rating, gross domestic and foreign and many other information obtained from different movie sites, we see this when reading the separate data files.

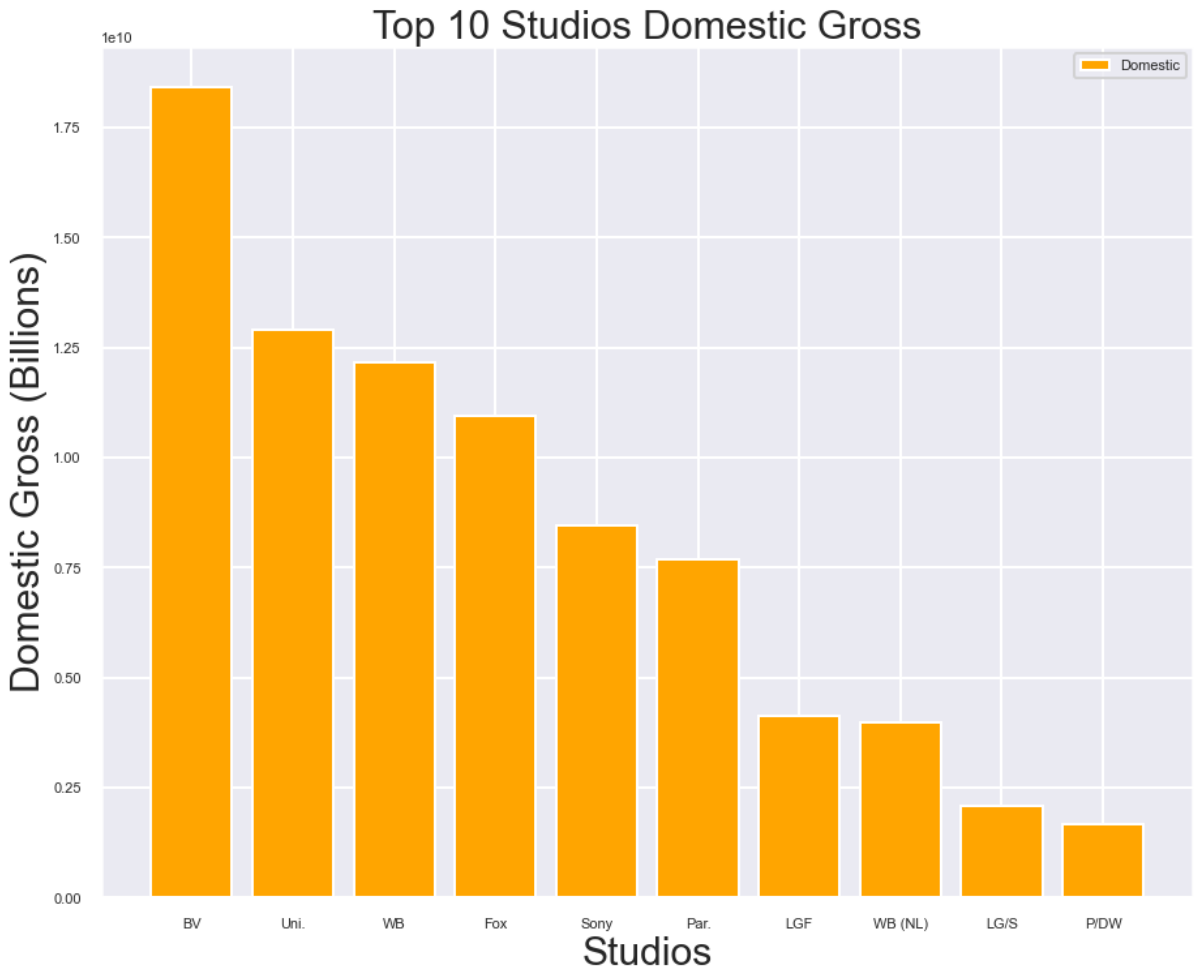
I utilized three different data sources for my analysis in order to have the most comprehensive view of the current movie performance.

- The Box Office Mojo Data: which was provided as a zipped data in csv format, containing 5 columns and 3387 movies in total. The data set was obtained from the Box Office website which ranges from 2010-2018. From the mojo data we see that most movies were filmed in IFC studio.
- Rotten Tomatoes Data: The data obtained was in a csv format with 1560 rows and 12 columns. From the data we see that the most produced genre from value counts is Drama followed by comedy.
- IMDB Data: this is a sql data and I preferred working with movie basics and movie ratings table so as to compare movies performance using average rating and genre.

I will begin my analysis by performing some descriptive analysis on each data set. Through this I will be able to obtain trends in the data pertaining to what needs to be known for a movie to be successful. This analysis will mainly be done through examination of graphs of particular attributes.

Results

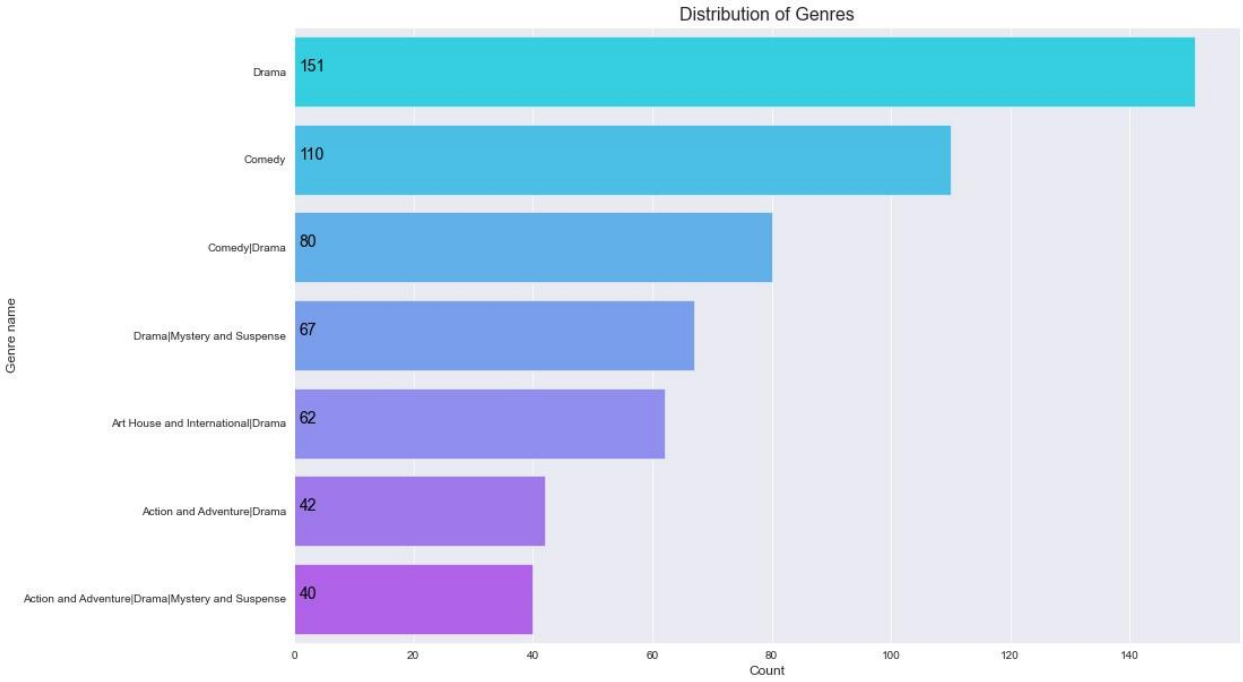
According to BOM data, on determining top 10 studios with their respective domestic gross:



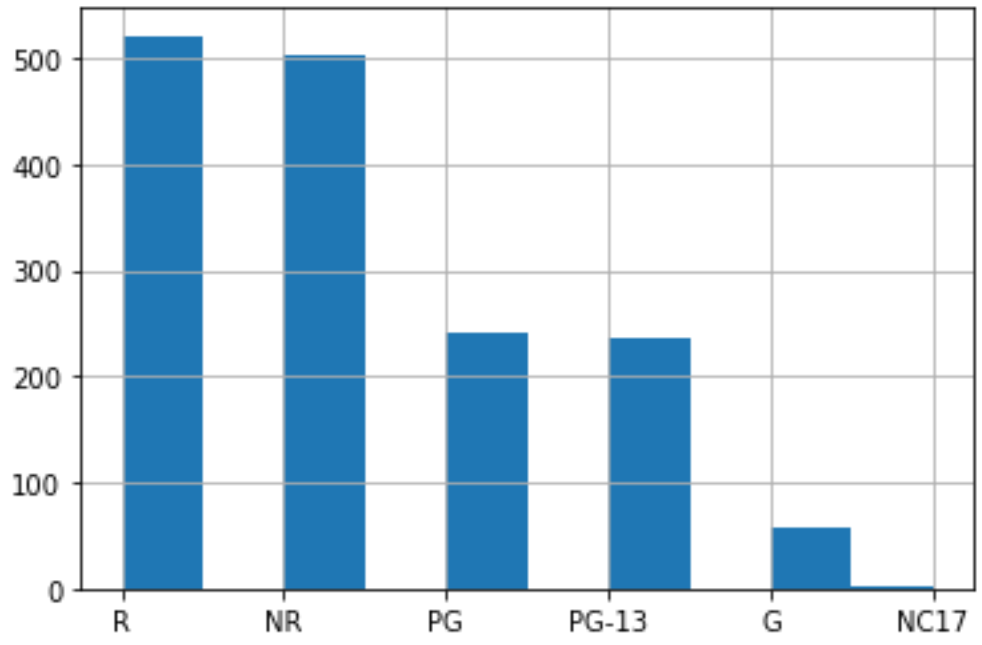
The domestic gross depends on the choice of studio.

Using the Rotten Tomatoes Data, I was able to obtain the count of each genre. Drama has the most counts followed by comedy and the least was Comedy| Horror| Mystery and Suspense.

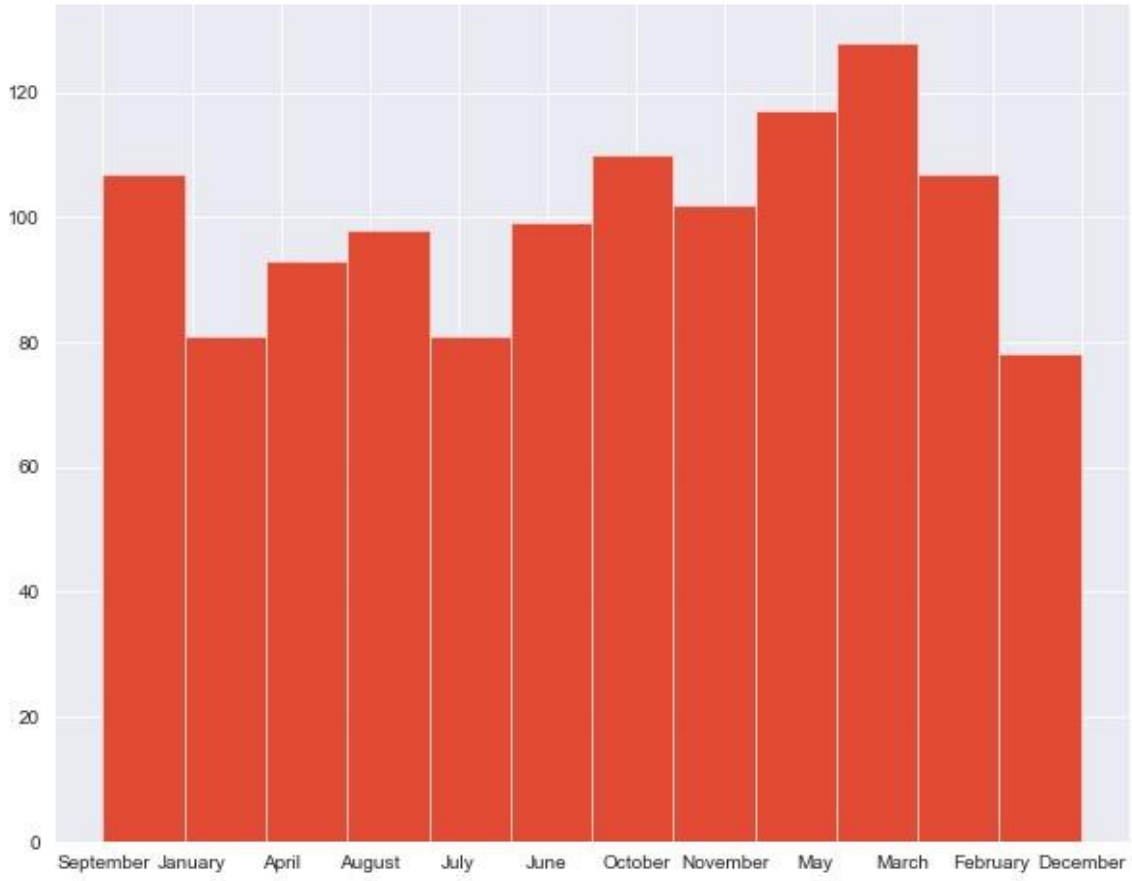
The diagram below illustrates the distribution of genres using the rotten tomatoes data:



Most movies were rated as R(Restricted)



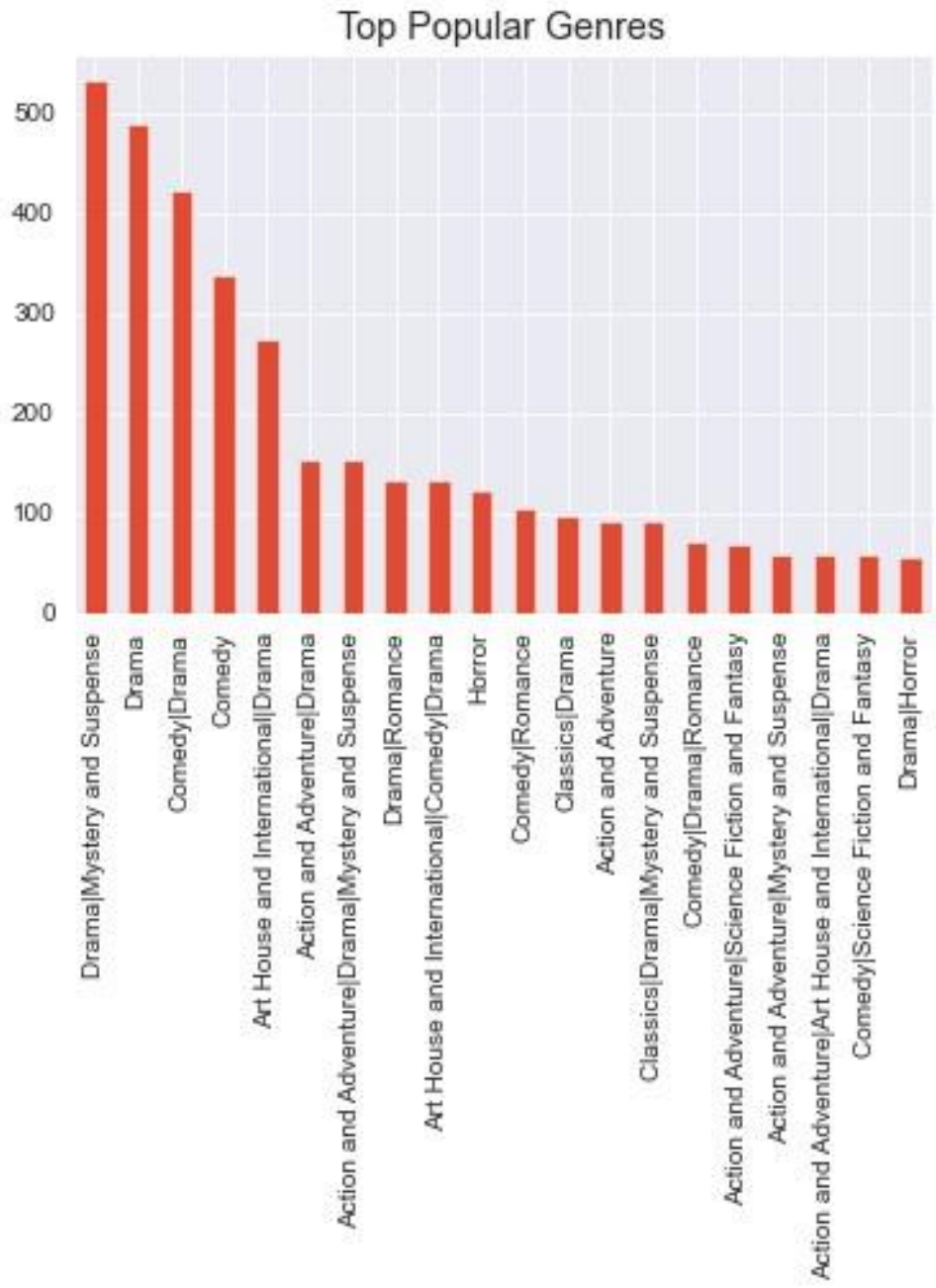
March, May, October, February are September the top 5 months in which movies are released. We can see this through the graph I plotted below:



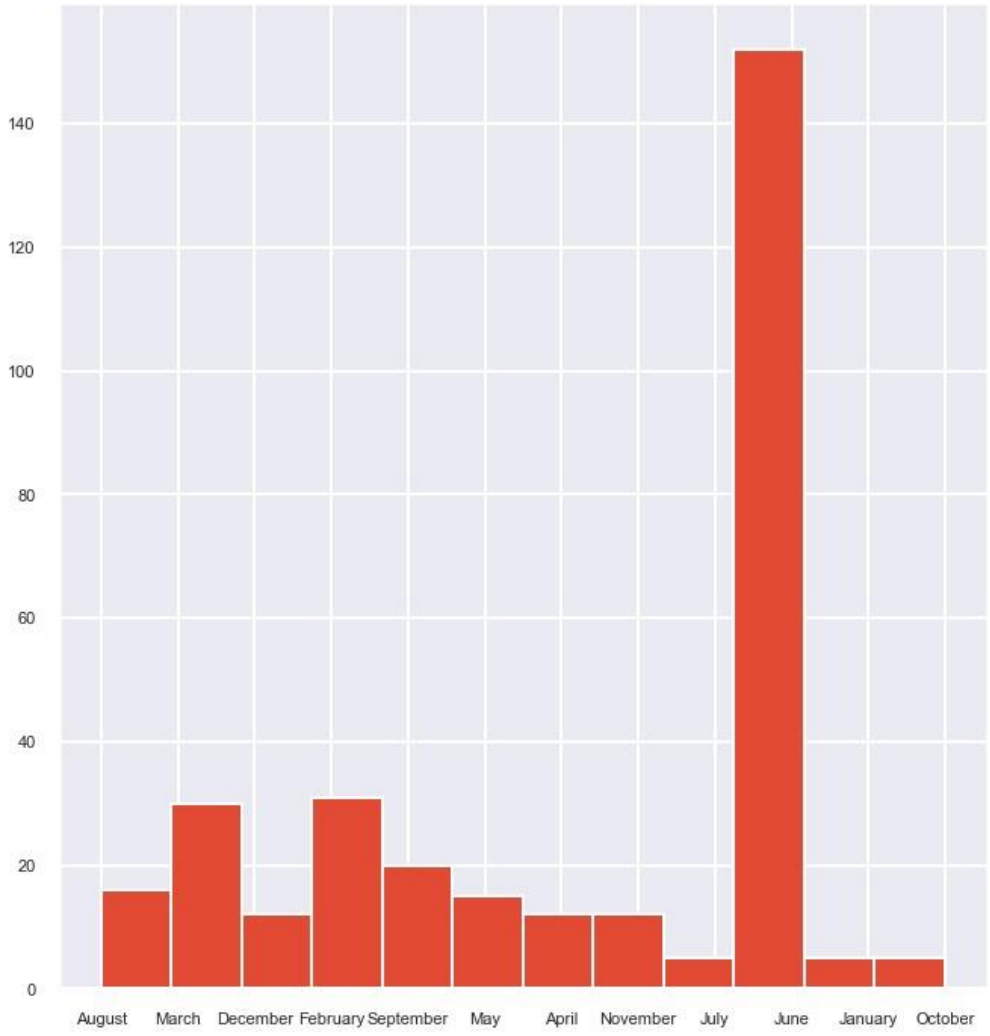
On merging two data sets, the BOM and The Rotten Tomatoes;

The top popular genre is Drama| Mystery and Suspense followed by Drama and Comedy| Mystery and Suspense| Romance being the least popular.

Graphical representation:

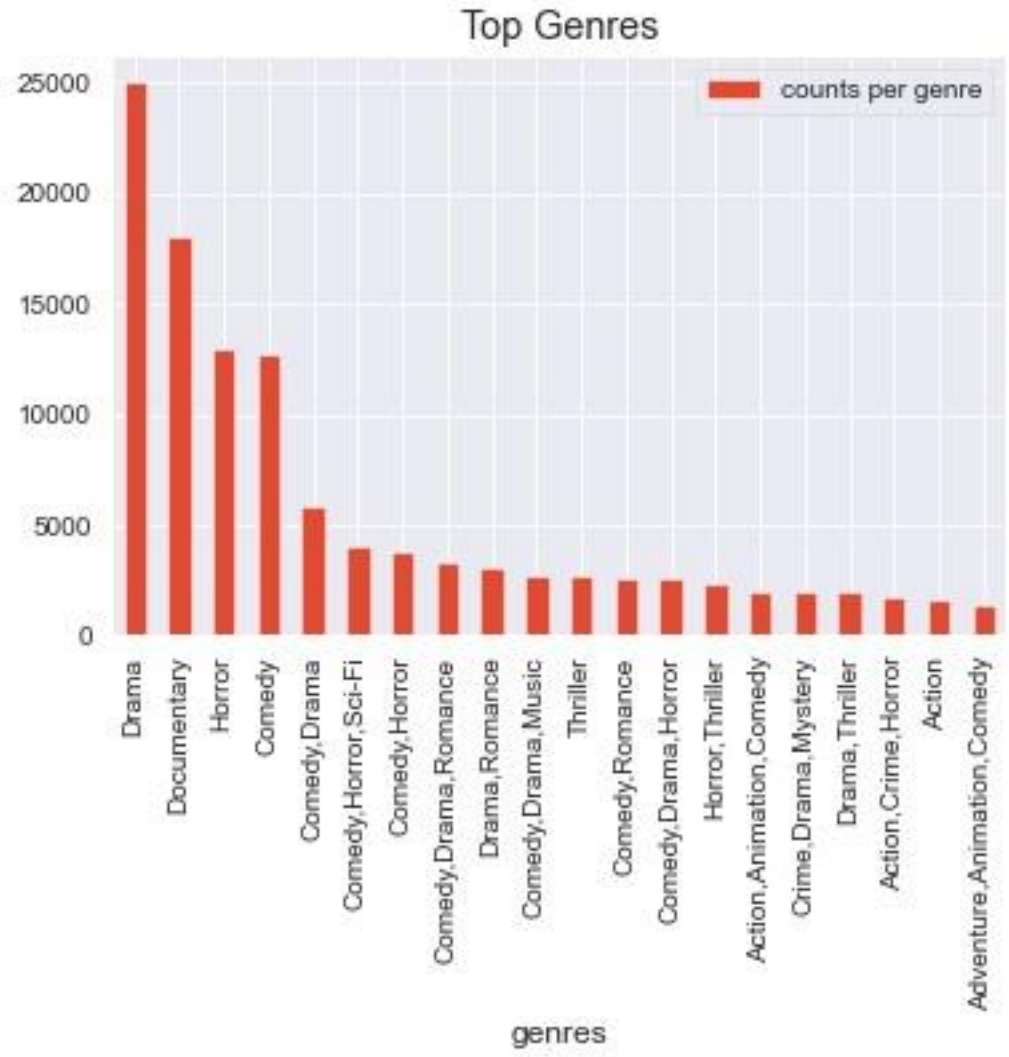


I went ahead and created a data frame for Drama| Mystery and Suspense genre so as to analyze it further.



We see that most Drama| Mystery and Suspense movies are released in the month of June as compared to any other time of the year.

For the IMDB data: I joined the movie basics, movie rating and persons table using movie id. For this data set the genre with the most value count is Drama as illustrated



Conclusion

1. Microsoft should produce either a Drama, Comedy, Mystery, Fantasy and Suspense movie and release it in either February, June or September.
2. From the BOM, for Microsoft to have a good return on the domestic gross they should go for the top most ranked studios.
3. In movie production, Drama genre to be precise, for it to have a good average rating and receive most public attention Microsoft should consider having the best director in production.