


 **Bryson1798** / **dsc-phase-1-project-v7** Public View license 0 stars  0 forks Star Watch[Code](#) [Issues](#) [Pull requests](#) [Actions](#) [Projects](#) [Wiki](#) [Security](#) [Insights](#) [Settings](#) master ▾

...



Bryson1798 ...

30 minutes ago

[View code](#) README.md

Box Office Analysis: Identifying Best Performing Movie Genres for Microsoft's New Studio

Overview

This project analyzes the datasets from 3 movie websites namely, [Box Office Mojo](#), [TheMovieDB](#) and [The Numbers](#). The 3 datasets are merged into one pandas DataFrame to enable a more in-depth analysis and better findings with regards to the top/best performing movie genres in the Box Office. Microsoft can use the findings from this analysis to help decide what type of films to create so as to stay at par in the movie industry.

Business Understanding

Based on the business problem, which is, Microsoft wants to get in on the fun of creating movies/original video content but they have no knowledge of creating movies, I have formulated 5 business questions whereby I will use my dataset to extract meaningful findings which can be translated into actionable insights for the Head of Microsoft's new movie studio to help him/her decide on what types of films to create. These business questions are:

1. What are the top 3 best performing genres of movies at the box office?
2. What is the relationship between production budget and the success of a movie both domestically and worldwide?
3. What is the competitive landscape of the movie industry in terms of market share?
4. How does the release time of a movie contribute to its success?

Data Understanding

The data sources for this analysis are 3 websites namely:

- [Box Office Mojo](#)
- [TheMovieDB](#)
- [The Numbers](#)

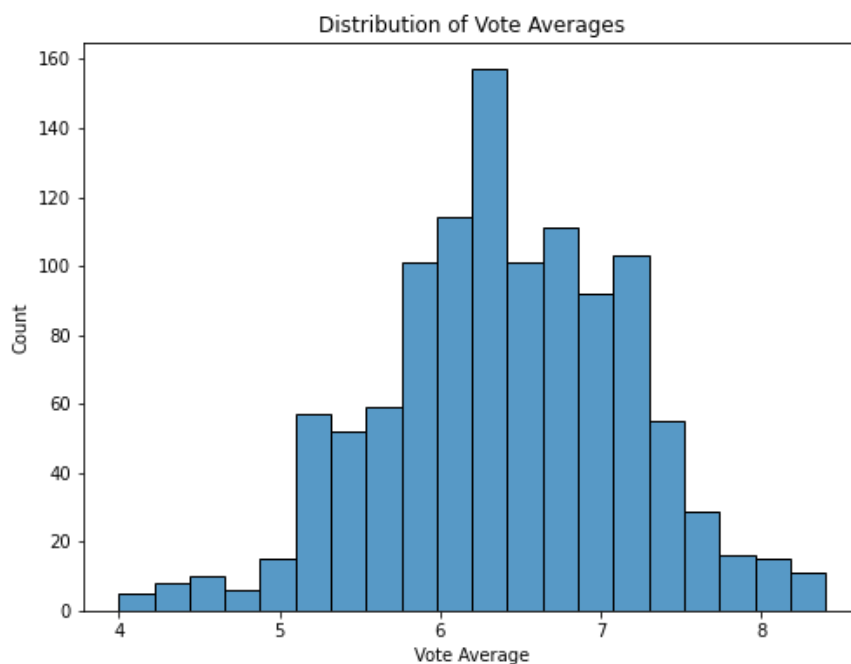
I therefore have 3 separate CSV data files:

- `bom.movie_gross.csv.gz` : each record represents a movie title, with attributes of that movie (eg. `domestic_gross`).
- `tmdb.movies.csv.gz` : each record represents a movie title as well, with attributes such as `release_date` .
- `tn.movie_budgets.csv.gz` : each record represents a movie title as well, with attributes such as `production_budget` .

Also, note that the data may not reflect the most-up-to-date trends and performances in the movie industry since its scope is upto 2020.

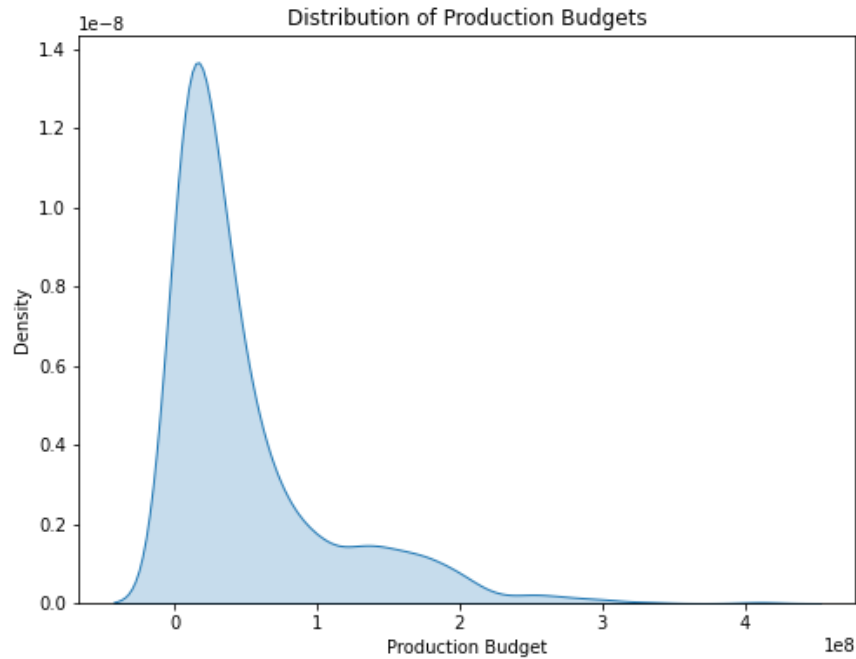
Univariate Analysis

i) A histogram to show distribution of vote averages



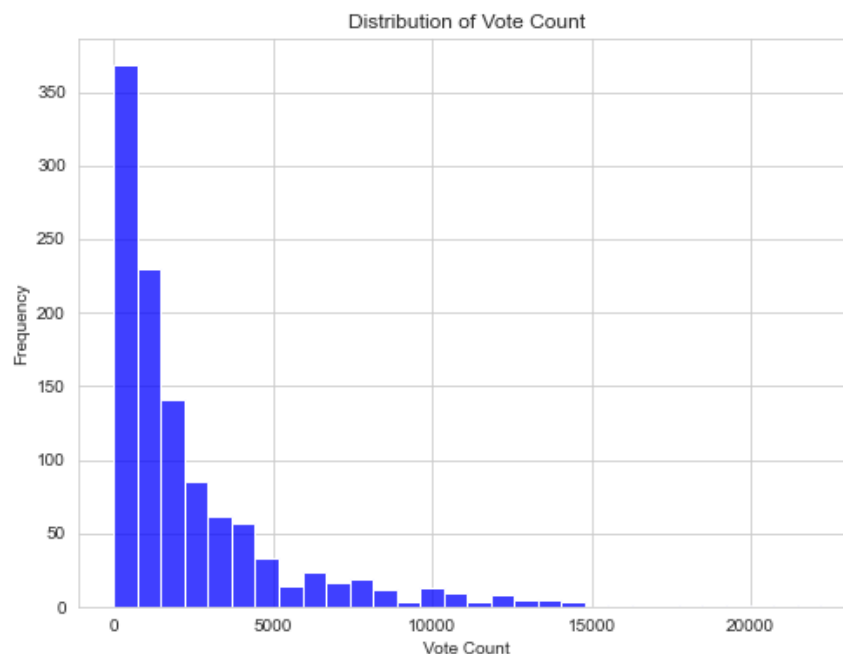
From the histogram above, I can observe that a majority of movies received a vote average of between 6 and 6.5.

ii) A kernel density plot to show the distribution of production budgets



From the above kernel density, the peak denotes the value of the production budget that is most frequently observed among the movies. I can observe that the peak of the distribution is skewed to the right, this suggests that a majority of movies have lower production budgets, with fewer movies having higher production budgets.

iii) A histogram to show the distribution of vote count

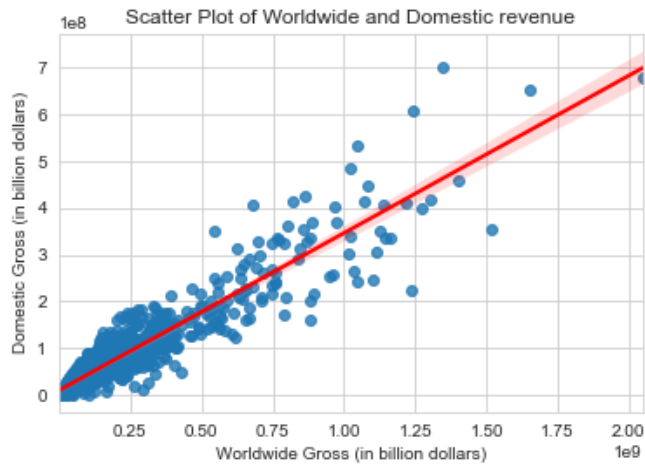


From the above histogram, I can observe that a majority of the movies have low vote counts since the peak of the histogram is skewed to the right.

Bivariate Analysis

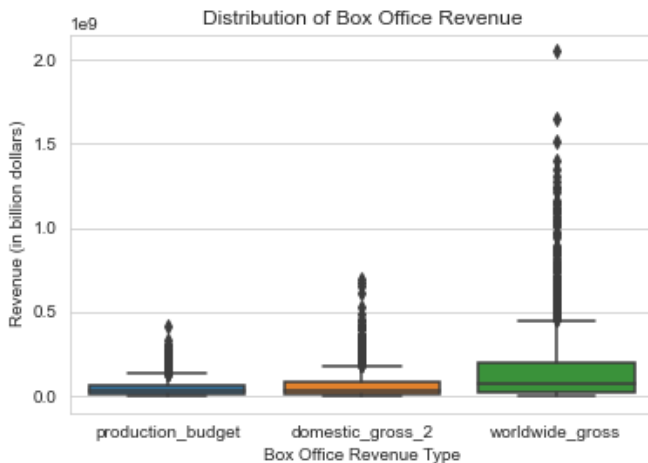
Here I will be generating plots to describe the relationships between different features/variables.

i) A scatter plot showing the relationship between `worldwide_gross` and `domestic_gross`



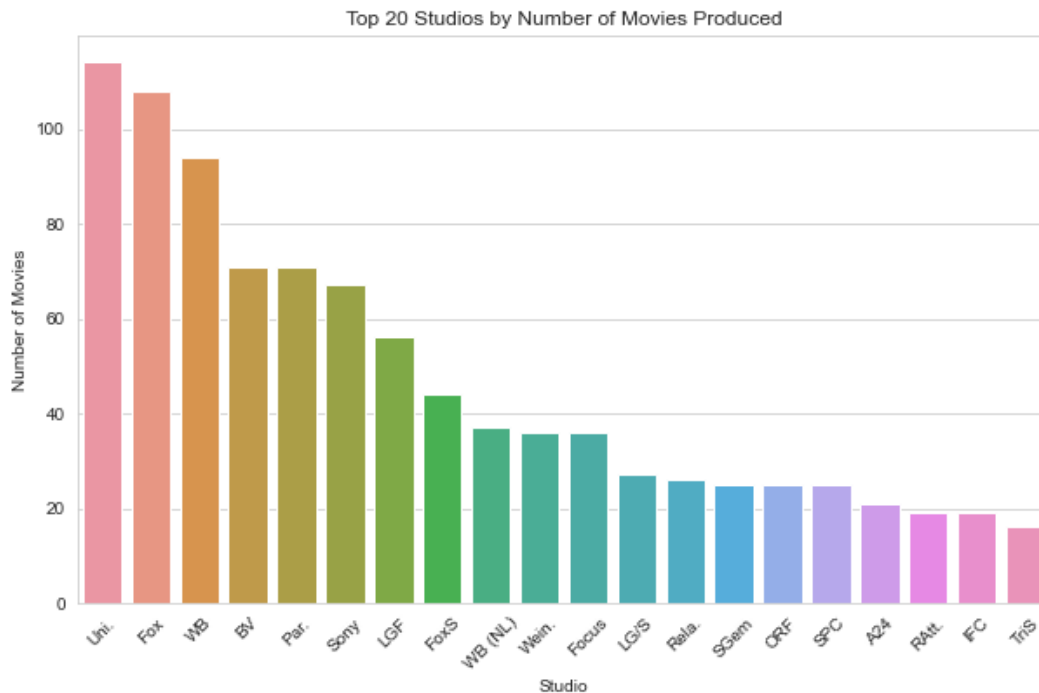
Using the line of best fit from the above scatter plot, I can observe and conclude that there is a strong positive correlation between `worldwide_gross` and `domestic_gross_2` variables. This means that as the domestic revenue from a movie increases, its worldwide revenue increases too.

ii) A box plot showing the distribution of `production_budget`, `domestic_gross_2` and `worldwide_gross` revenues.



The boxplot above shows the distribution of `production_budget`, `domestic_gross_2` and `worldwide_gross` variables. There's quite a vast number of outliers in all the three variables. But given that this is a Movies dataset, I will keep the outliers, since they are representative of the real world data whereby we can have some Box Office movies performing exceptionally well in the market.

iii) A bar plot showing the Top 20 Studios by the number of movies produced



The bar plot above shows the distribution of number of movies produced by studios. From the bar plot, it is clear that Uni. Studio takes the lead in movie production followed by Fox and WB. But it is important to note that producing more movies doesn't necessarily mean that it is the best performing studio. Other factors such as return on investment must be considered too.

Performing Aggregations to answer Business Question 1

The business question:

- What are the top 3 best performing types of movies at the box office?

Here I use the `groupby()` built-in method and group the dataset by the `genres` column to answer the question.

Also of importance to note is that some movies are a mixture of different genres. So it begs the question, to either treat each genre individually or as a whole? The answer to this question depends on the analysis being performed. If one is analyzing the popularity of each individual genre separately, then they would treat each genre individually. However, if one is analyzing the popularity of movies with a specific combination of genres, then they would treat the list of genres as a whole.

That said, I will be treating the list of genres as a whole, since a movie can be of one genre or also a combination of different genres.

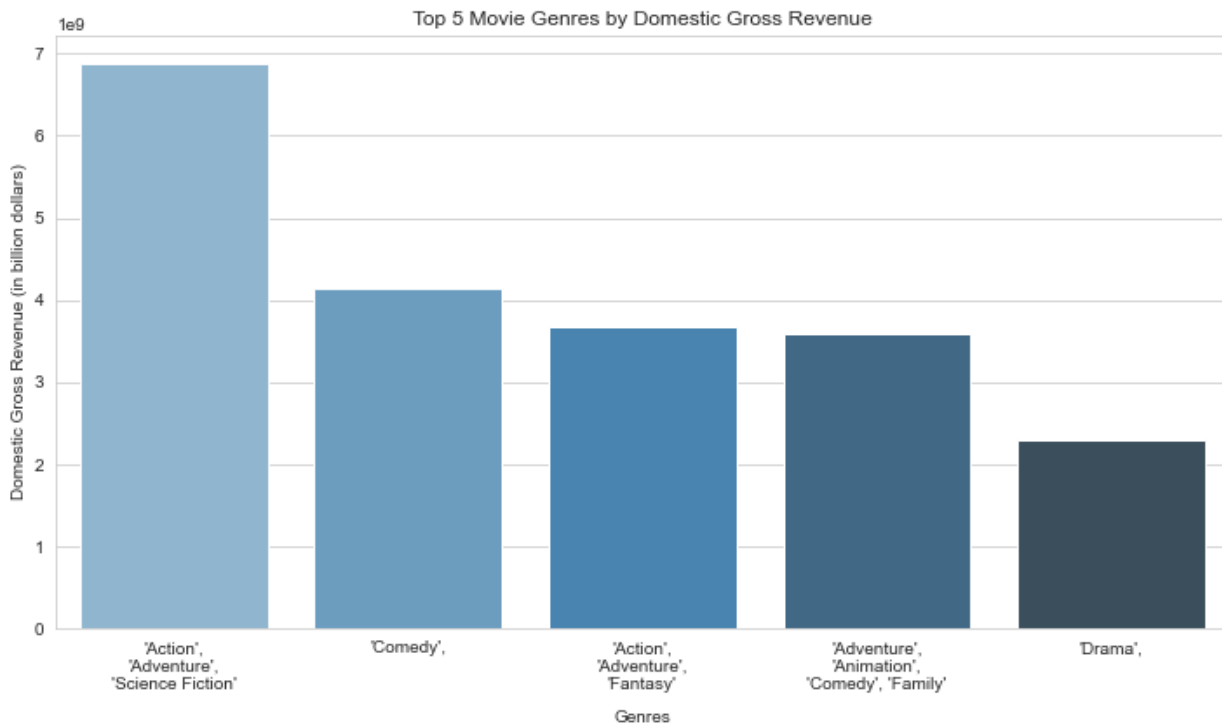
To gain a better understanding of the current trends in the movie industry, it may be useful to explore multiple perspectives and consider how they relate to one another. I will compare the top genres across different groupings and look for patterns or similarities that may provide insight into overarching trends.

It's also important to keep in mind that the top genres by one metric may not necessarily be the same as the top genres by another metric. For example, a genre may be highly profitable but not very popular among audiences, or it may receive high ratings but not generate a lot of revenue. Therefore, it's important to consider multiple metrics when analyzing the data to get a more comprehensive understanding of the current trends in the movie industry.

To come up with a final top 3, I will consider the top genres across multiple metrics and determine which genres are consistently ranking high across the board. For instance, I will create plots that show the rankings for each genre by `domestic_gross_2`, `worldwide_gross`, `popularity`, `vote_count`, and `vote_average`, and then compare the results.

In addition, you can get the definitions of the various `genre_ids` values at [TheMovieDB genre_ids definitions](#)

i) Group the DataFrame by `genres` and sum `domestic_gross_2`

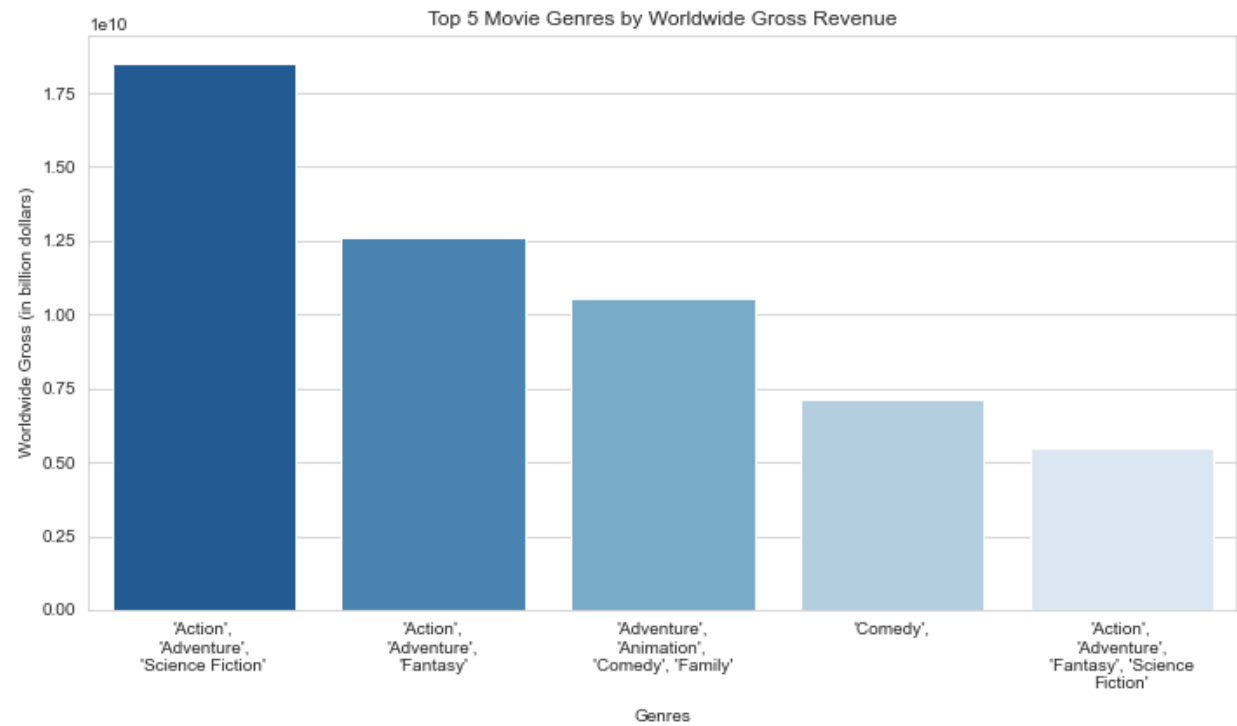


Using the definitions of the various `genre_ids` values at [TheMovieDB genre_ids definitions](#);

From the above results, the top 5 `genres` sorted by `domestic_gross_2` are:

1. Action, Adventure, Science Fiction
2. Comedy
3. Action, Adventure, Fantasy
4. Adventure, Animation, Comedy, Family
5. Drama

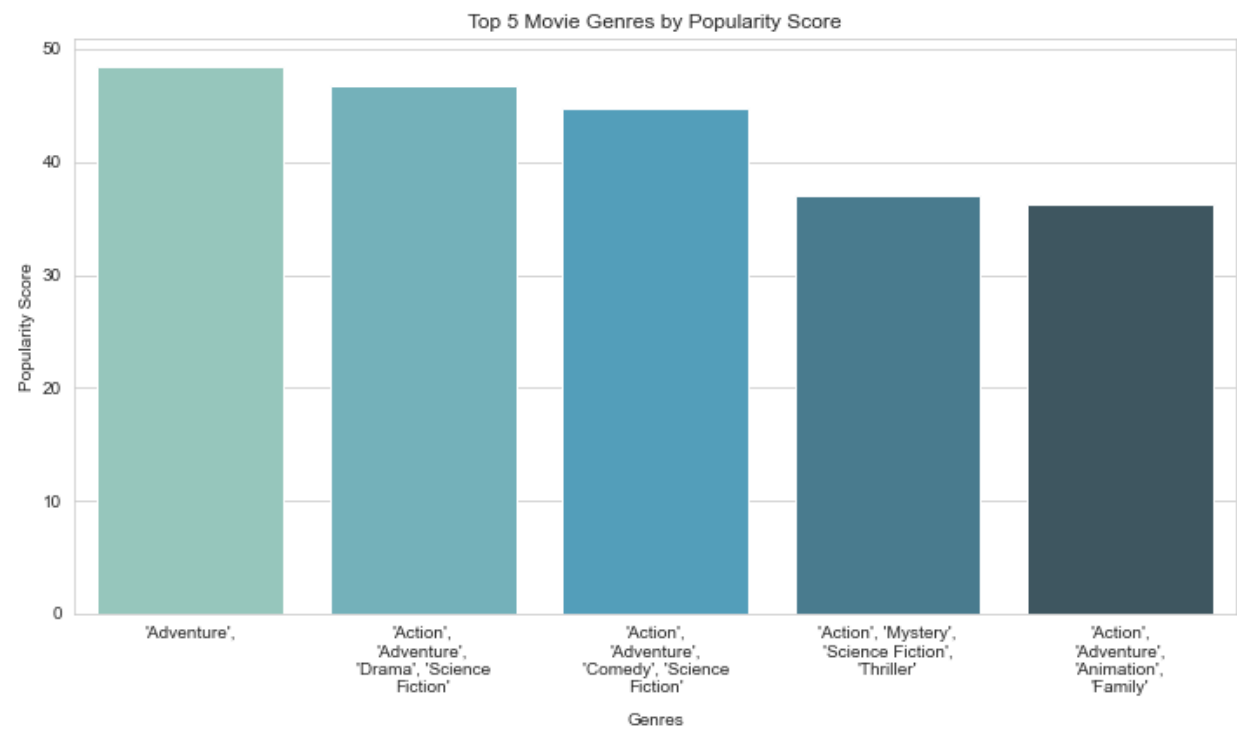
ii) Group the DataFrame by `genres` and sum `worldwide_gross`



From the above results, the top 5 genres sorted by worldwide_gross are:

- 1. Action, Adventure, Science Fiction
- 2. Action, Adventure, Fantasy
- 3. Adventure, Animation, Comedy, Family
- 4. Comedy
- 5. Action, Adventure, Fantasy, Science Fiction

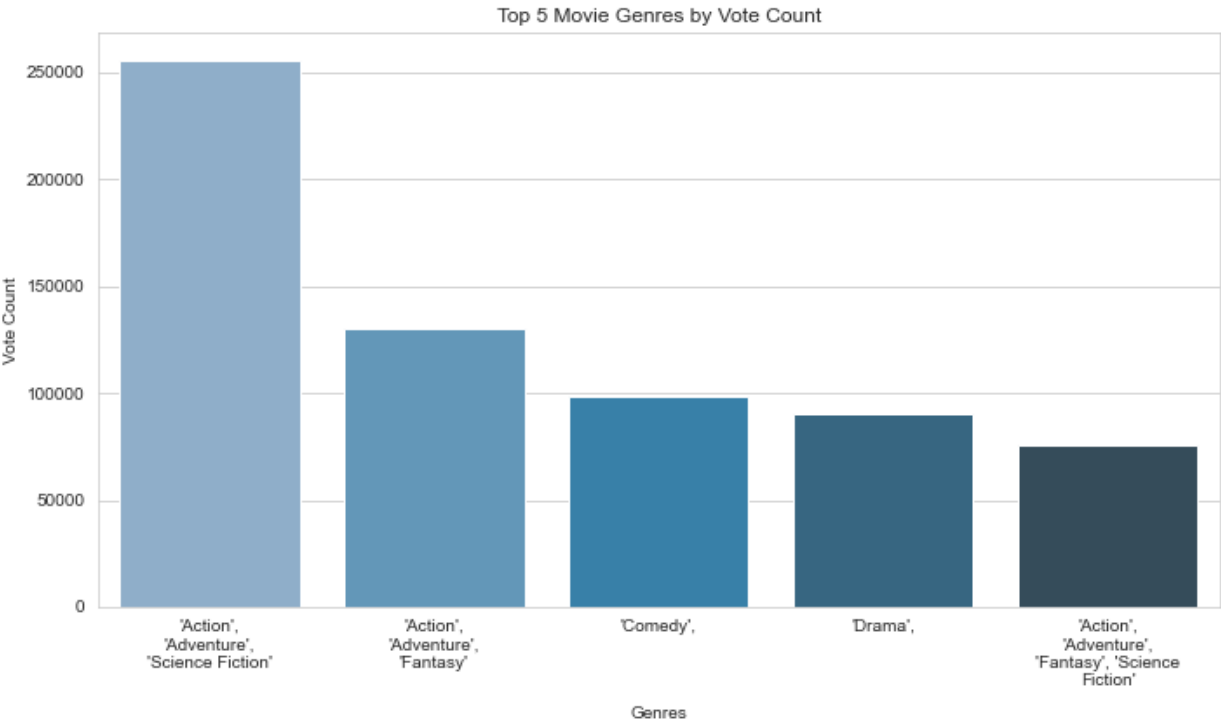
iii) Group the DataFrame by genres and mean of popularity



From the above results, the top 5 genres sorted by popularity are:

- 1. Adventure
- 2. Action, Adventure, Drama, Science Fiction
- 3. Action, Adventure, Comedy, Science Fiction
- 4. Action, Mystery, Science Fiction, Thriller
- 5. Action, Adventure, Animation, Family

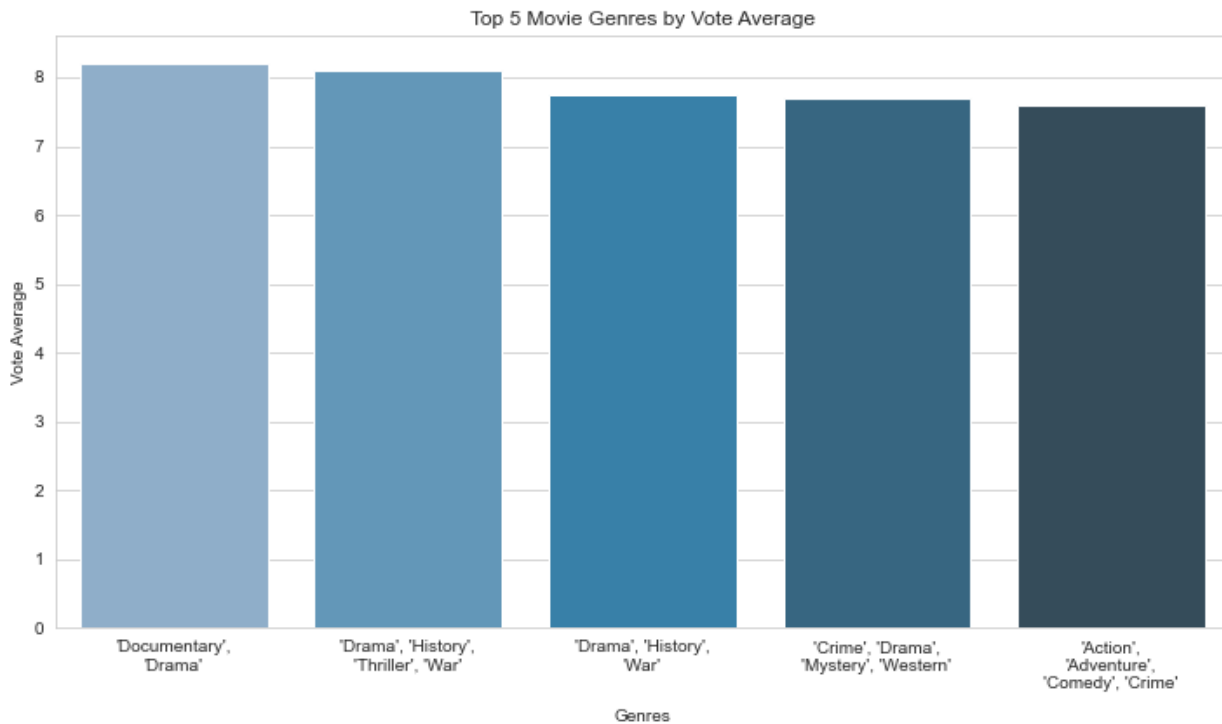
iv) Group the DataFrame by genres and sum vote_count



From the above results, the top 5 genres sorted by vote_count are:

- 1. Action, Adventure, Science Fiction
- 2. Action, Adventure, Fantasy
- 3. Comedy
- 4. Drama
- 5. Action, Adventure, Fantasy, Science Fiction

v) Group the DataFrame by genres and mean of vote_average



From the above results, the top 5 genres sorted by `vote_average` are:

1. Documentary, Drama
2. Drama, History, Thriller, War
3. Drama, History, War
4. Crime, Drama, Mystery, Western
5. Action, Adventure, Comedy, Crime

Therefore, from these 5 sets of analyses I can conclude that the top 3 best performing types of movies are:

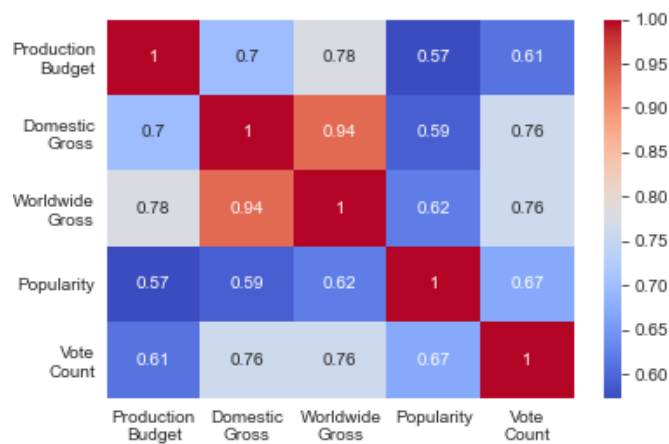
1. Action, Adventure, Science Fiction
2. Action, Adventure, Fantasy
3. Comedy

Creating a correlation matrix to answer Business Question 2

The business question:

- What is the relationship between production budget and the success of a movie both domestically and worldwide?

Below I create a correlation matrix for the features which are indicators of a successful movie in the Box Office



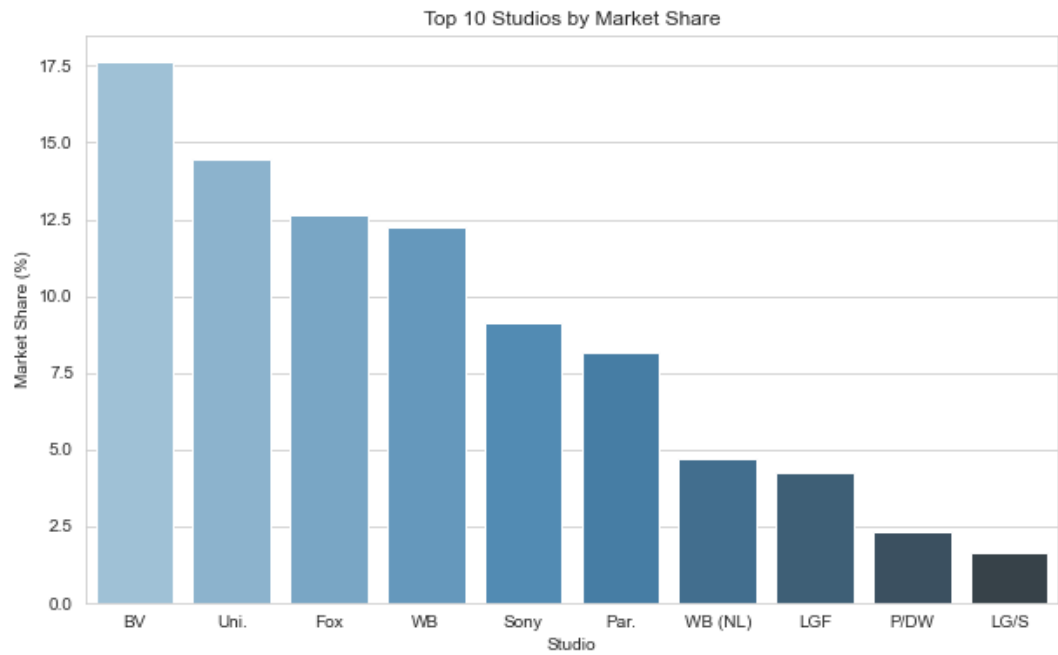
- The correlation coefficient between `production_budget` and `worldwide_gross` is **0.78**; and between `production_budget` and `domestic_gross_2` is **0.7**. Both figures indicate a strong positive correlation between the variables. This means that as `production_budget` increases, `worldwide_gross` and `domestic_gross_2` tend to increase as well. Therefore, the production budget allocated to a movie production is a good indicator of its success or failure in the Box Office in terms of revenue.
- Also, another important correlation coefficient to note is between `domestic_gross_2` and `worldwide_gross` which is **0.94**. This indicates a strong positive correlation between the two variables. This means that a movie that is successful locally is highly likely to be successful internationally/worldwide as the relationship between the two variables is relatively strong.

Performing Aggregations and Engineering a new feature `market_share` to answer Business Question 3

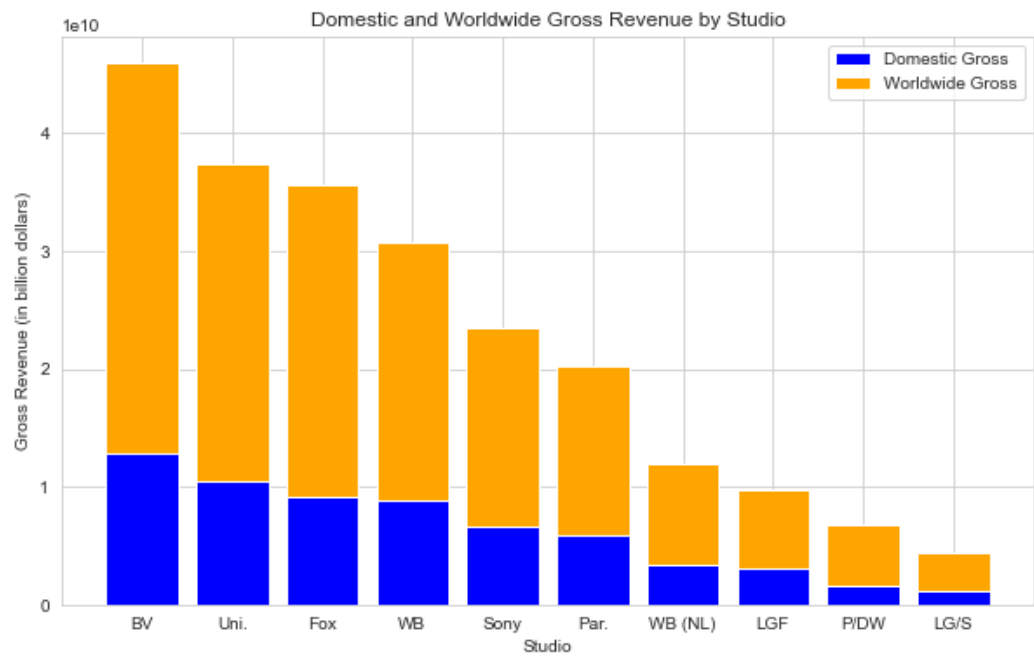
The business question:

- What is the competitive landscape of the movie industry in terms of market share?

I find this to be an important aspect because as Microsoft dives into the movie industry, it's imperative they understand who they are going to be competing against in terms of capturing the audience's attention.



The `market_share` is calculated as the sum of a studio's domestic revenue divided by the total domestic revenue of all the studios multiplied by 100%. Based on the above analysis, BV Studio has the highest domestic and worldwide gross revenue, the highest total vote count, and the highest mean popularity. BV Studio has a market share of 17.6%, followed by Uni. with 14.4% and Fox with 12.7%. This suggests that BV Studio is the clear leader in many metrics. This provides insight into the competitive landscape of the movie industry based on the chosen metrics.



The stacked bar plot shows the domestic and worldwide gross revenue of the top 10 movie studios, broken down by region, i.e. Domestic revenue and Worldwide revenue. Each bar represents a studio, and is divided into two sections: blue for domestic gross revenue, and orange for worldwide gross revenue.

The height of each bar represents the total gross revenue for that studio, and the width of each section represents the proportion of that revenue coming from the domestic or worldwide market. For example, the tallest bar represents BV Studio, and we can see that the majority of its revenue comes from the international market.

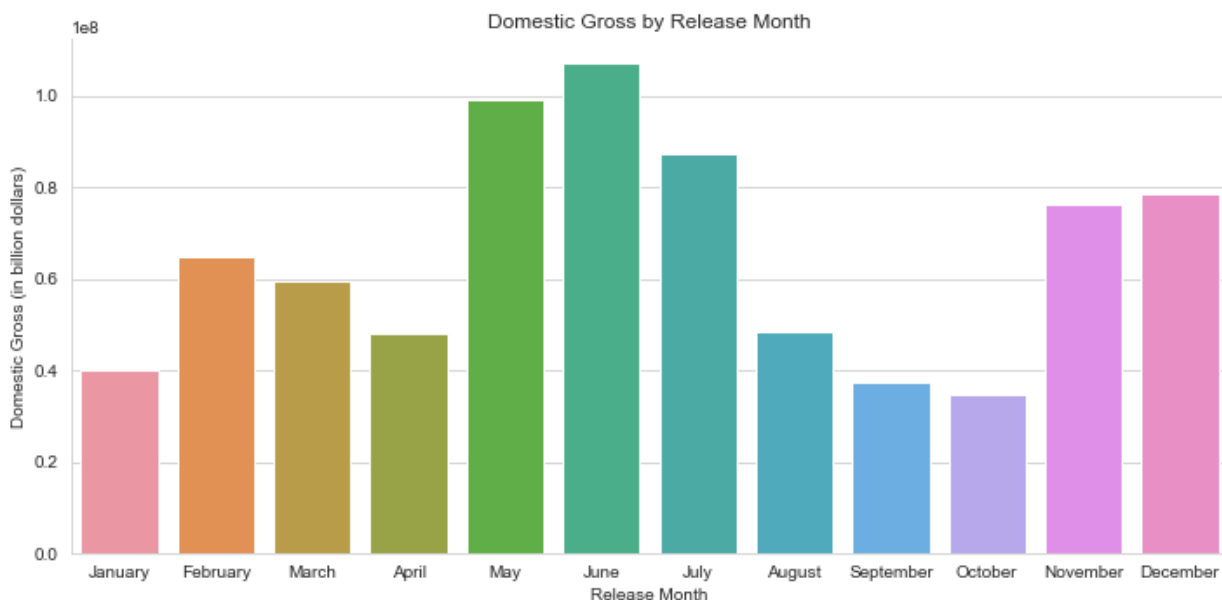
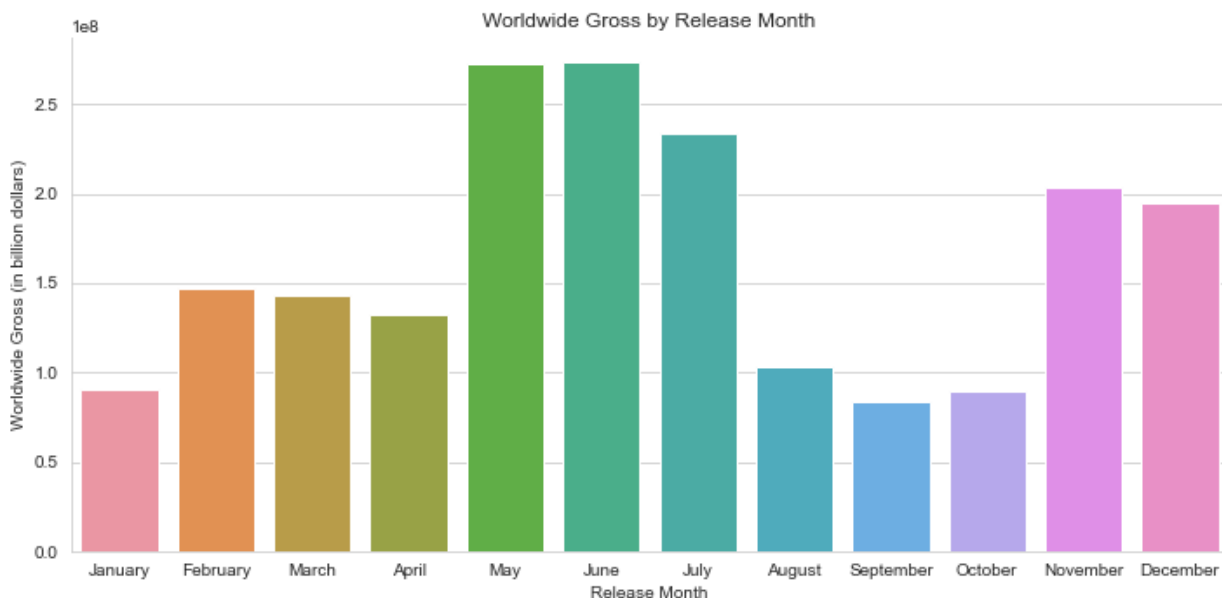
Overall, this plot helps us visualize the revenue breakdown of the top movie studios, and can provide insights into the relative importance of different regions for these studios. This is also an indicator to Microsoft's new movie studio, in that they can anticipate more of their revenue generated from the international market, as compared to the domestic market.

Performing Aggregations to answer Business Question 4

The business question:

- How does the release time of a movie contribute to its success?

Here I will be using the `worldwide_gross` and `domestic_gross_2` variables as my units of measurement to depict the revenues generated in the different months and hence determine when it is suitable to release a movie, thereby increasing its chances of success.



The above plots show the average worldwide gross and domestic gross by release month for the movies in the dataset. The x-axis shows the months of the year, and the y-axis shows the average gross in billion dollars. Each bar represents the average gross for a particular month. The "Worldwide Gross by Release Month" plot shows that the months of May, June and July have the highest average worldwide gross, while the months of September and October have the lowest. The "Domestic Gross by Release Month" plot show a similar trend, with May, June and July having the highest average domestic gross, and September and October having the lowest. Overall, these plots suggest that releasing a movie in May, June or July may lead to higher gross revenue, both domestically and worldwide, while releasing a movie in September or October may result in lower gross revenue.

Conclusion

This analysis leads to four **recommendations** that will enable Microsoft get into the movie industry with a resounding success for the movies that will be produced/created.

1. Based on the findings of the top 3 best performing types of movies in the Box Office, Microsoft can consider producing movies around the genre combinations of:

- **Action, Adventure & Science Fiction**
- **Action, Adventure, Fantasy**
- **Comedy**

Also, they can play around the genres creatively and come up with something a bit unique, for example, a combination of **Action, Adventure, Comedy** or even **Action, Science Fiction, Fantasy** to see the response and reaction from the movie lovers.

2. Based on the findings of strong positive correlation between production budget and domestic gross, and production budget and worldwide gross, for the genres stated above in the first recommendation, the Head of Microsoft's new movie studio can liaise with the finance department and ensure that sufficient budgetary allocation is made to film production. This would enable the several aspects involved in film production to be taken care of sufficiently, for instance;

- **Production equipment:** Getting the latest equipment and editing tools is key to producing high quality video content.
- **Visuals and Sound:** The visual elements of a movie are critical in creating an immersive and engaging experience for viewers. This includes everything from the cinematography and special effects to the costumes and set design. A good soundtrack can help set the tone of a film and enhance the emotional impact of key scenes.
- **Marketing:** Finally, filmmakers will need to consider the marketing and distribution of the new movie. They will need to think about how they will promote the movie and ensure that it is being distributed in a way that will reach their intended audience. Effective marketing will certainly increase the movie's popularity, which in turn may mean success for the new movie.

Therefore, the investment in a movie's production really influences its success in the Box Office.

3. Based on the findings of how competitive the movie industry is in terms of market share, Microsoft will need to differentiate itself in order to stand out, for example;
- Microsoft's new movie studio could focus on producing high-quality movies that are unique and have compelling storylines, in line with the best performing genres specified in the first recommendation.

- Partnering with well-known and respected directors and actors.
 - Leveraging innovative marketing and distribution strategies to reach wider audiences.
4. Based on the findings of the best months to release a movie being May, June and July, the Microsoft new movie studio should consider releasing movies around this time. There could be various factors that contribute to the high revenues in May, June and July. One possibility is that these months fall within the summer blockbuster season, which typically runs from May to August, where studios release highly anticipated movies that are expected to perform well at the box office, as well as audience availability. Therefore, if Microsoft takes advantage of this period, the movies released are likely to yield higher gross revenues.

Next Steps

Further analyses could yield additional insights to further improve decision-making for the movie genres to produce in the new Microsoft studio:

- A further analysis into the directors/actors in the film industry. This analysis could provide insights on whom to hire during production of movies. Involving highly rated/successful directors & actors in production increases the probability of success for a movie.
- A further analysis into the reviews made by the public by conducting sentiment analysis on social media and other online platforms to gauge the public's reaction to movie trailers, posters, and other promotional materials. This analysis can help Microsoft studio identify potential issues with marketing campaigns or aspects of the movie that may not be well-received by the audience and hence make adjustments to its marketing and production strategies in real-time, potentially increasing the chances of success for its movies.

Releases

No releases published

[Create a new release](#)

Packages

No packages published

[Publish your first package](#)

Languages

● Jupyter Notebook 100.0%