New Movie Studio Investment

MICROSOFT COMPANY

Done by JOSELINE A.

1.0 Overview and the Business understanding.

- Welcome team to the presentation of the analysis of the movie studio data sets.
- The purposes of this analysis is to be able to help advice Microsoft company the important factors to consider while investing in a new Movie Studio business.
- These are the two data sets used in the analysis.
- **Box Office Mojo Data:**
  - Overview of variables such as movie titles, gross revenues, etc.
- **IMDb Data:**
  - Overview of variables in movie basics (title, genres, runtime) and movie ratings (average ratings, number of votes).
- More insights on these two datasets will be addressed more on the presentation.

1.1 The Report is sub-divided following section.

Data cleaning and Quality

- Identify and handle missing or inconsistent data in key columns (e.g., domestic gross, foreign gross, genres, runtime minutes, average rating, nom votes).Standardize data formats, especially for numeric and categorical variables.

Genre Analysis:

- Analyze the performance of different genres based on average ratings, runtime, and gross revenue.

Revenue Analysis:

- Calculate and compare domestic and foreign gross revenues.
- Identify the highest-grossing movies and studios.

Correlation Analysis:

- Investigate the relationships between numeric variables such as runtime minutes, average rating, and Num votes.
- Identify any significant correlations that could inform strategic decisions.

Studio Performance:

- Compare the performance of different studios based on their movies' gross revenue, ratings, and number of votes.
- Identify the most successful studios over the years.

Recommendations and the next steps to follow.

Continuation….

- These are:
  - 1. Loading necessary libraries these are Numpy
  - 2. Loading Dataset from a CSV file or from a Table.
  - 3. Summarization of Data to understand Dataset (Descriptive Statistics)
  - 4. Visualization of Data to understand Dataset (Plots, Graphs etc.)
  - 5. Data pre-processing and Data transformation.

1.2 Goals

- Our goal is to:
  - Determine the types of films that are currently successful at the box office and translate those findings into concrete business recommendations.
  - The genres that are doing well both foreign and domestically.
  - Determining which movies studio has been doing well over the years both domestically an in foreign countries.
  - Recommendations of the best genres to invest in that are doing wonderfully well in the industry.
  - Graphical representation that supports the analysis
  - Use easy to read codes.

2.0 Data Collection and Preparation

- **2.1 Data sources.**
- The two data sets used in this report analysis is are
  - Box Office Mojo (bom.movie_gross.csv
  - IMDb (im.db)

- 2.2 Data extraction:
  - The csv file which stands for the comma separated values using pandas. Pandas is one of the data analysis libraries which are used on entire data analysis.
  - The formula is (“pd.read_csv(‘bom.movie_gross’)

The IMDb file was imported using SQL. The two tables that was important for this analysis was the movie basics and movie ratings.

These were imported using the sqlite3 library a connection created to the databases using the querying languages.

2.1 Table presentation of the two datasets

- Importing the bom.movie_gross.csv This will provide the first 5 columns and rows
- `df = pd.read_csv('bom.movie_gross.csv')`
- `df`

title	studio	domestic_gross	foreign_gross	year	
0	Toy Story 3	BV	415000000.0	652000000	2010
1	Alice in Wonderland (2010)	BV	334200000.0	691300000	2010
2	Harry Potter and the Deathly Hallows Part 1	WB	296000000.0	664300000	2010
3	Inception	WB	292600000.0	535700000	2010
4	Shrek Forever After	P/DW	238700000.0	513900000	2010

The import tables of the im.db database tables

```
# Loading the movie_basics table
This tables gives a preview of the movie basics tables imported from im.db database
movie_basics_df = pd.read_sql_query("SELECT * FROM movie_basics", conn)
movie_basics_df
```

	movie_id	primary_title	original_title	start_year	runtime_minutes	genres
0	tt0063540	Sunghursh	Sunghursh	2013	175.0	Action,Crime,Drama
1	tt0066787	One Day Before the Rainy Season	Ashad Ka Ek Din	2019	114.0	Biography,Drama
2	tt0069049	The Other Side of the Wind	The Other Side of the Wind	2018	122.0	Drama
3	tt0069204	Sabse Bada Sukh	Sabse Bada Sukh	2018	NaN	Comedy,Drama
4	tt0100275	The Wandering Soap Opera	La Telenovela Errante	2017	80.0	Comedy,Drama,Fantasy

```
This will provide the first 5 columns and rows
# Load the movie_basics table
movie_ratings_df = pd.read_sql_query("SELECT * FROM movie_ratings", conn)
movie_ratings_df
```

movie_id	averagerating	numvotes	
0	tt10356526	8.3	31
1	tt10384606	8.9	559
2	tt1042974	6.4	20
3	tt1043726	4.2	50352
4	tt1060240	6.5	21

3.0 Data cleaning

- In this section, I handled the missing data sets, looked for duplicates within the datasets and finally any formatting inconsistencies.
- Merge datasets based on common identifiers (e.g., movie IDs) in the im.db database.
- Further analysis is now done with the cleaned dataset.

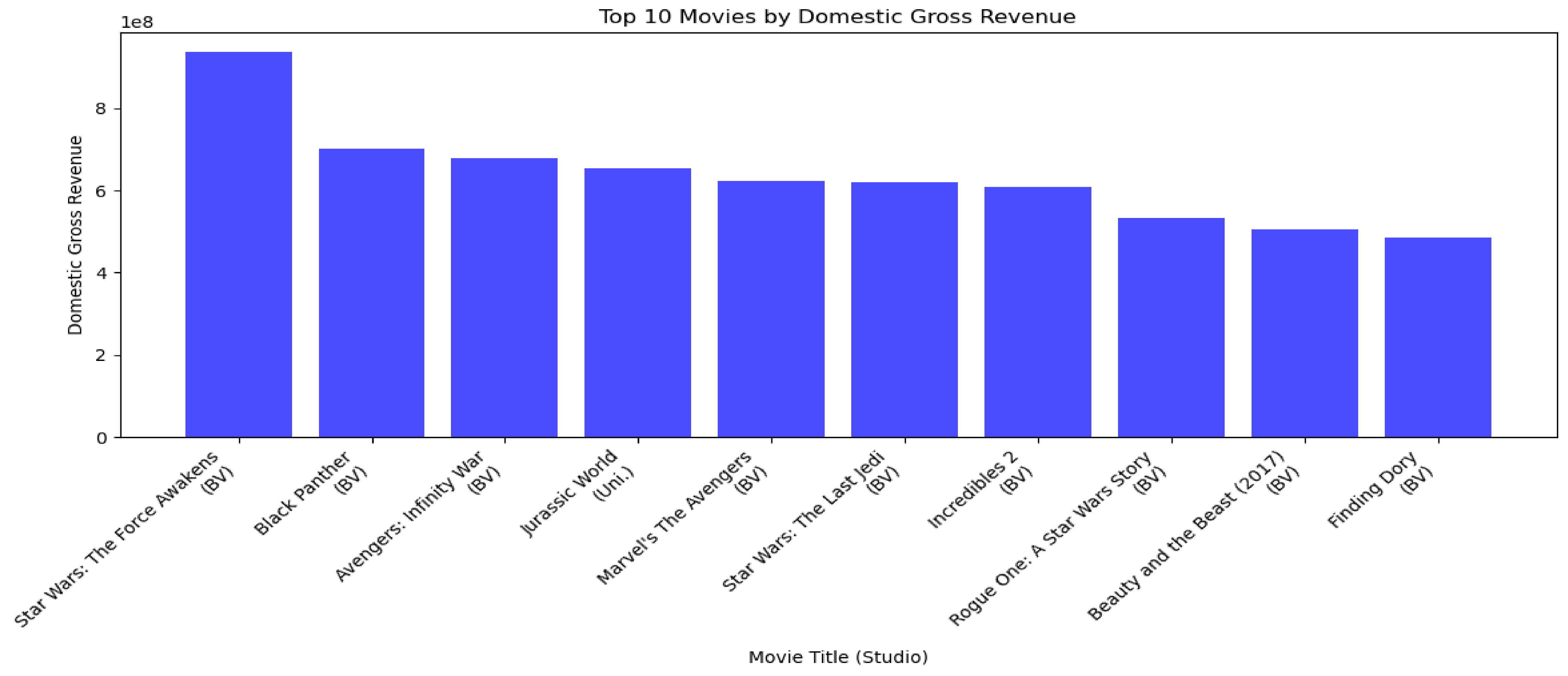
4.0 Approach Analysis

- The preferred methods used on the Analysis are:
  - **Descriptive statistics,** includes the mean, median mode and the interquartile ranges of the columns in the two data sets. An example is the mean of the domestic gross and foreign gross column.
  - **Correlation Analysis**: Investigate correlations between ratings, genres, and box office success.
  - The correlation analysis helps determine the variables which are directly related and hence a better decision making on which genre to invest in and expect a profit or an excellence in the project.

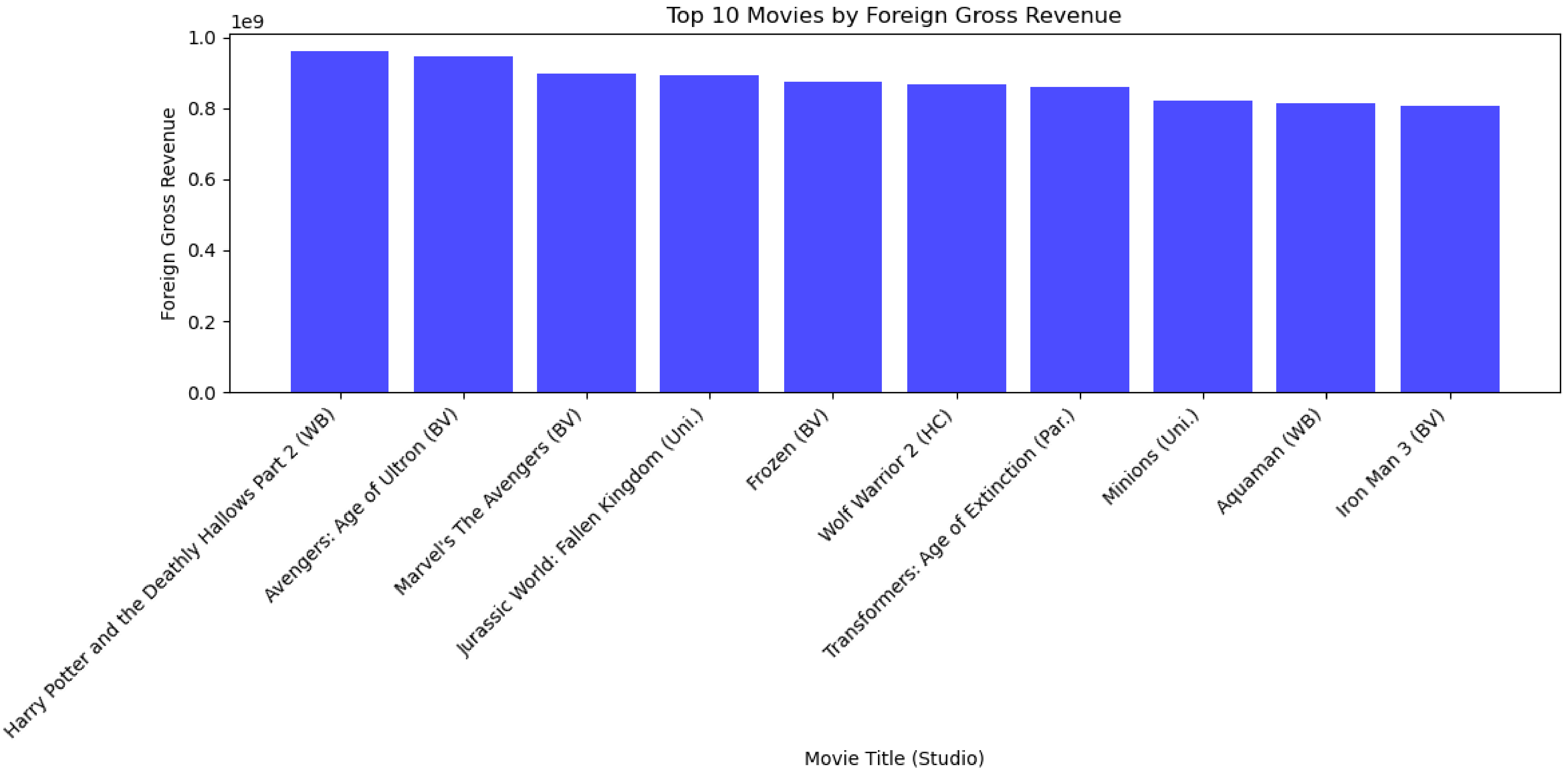
5.0 VISUALIZATION

Bar plot for the top 10 most grossing movie domestically

The following graphical representation are from the bom.movie_gross.csv



The total of top 10 foreign gross



Brief explanations on the first two graphs

- From the two graphical representations, it is quite clear that the BV movie studio is doing exceptionally well as both from the above analysis are grossing well both domestically and in foreign regions.

- This means that the majority of the audience fans of the Action, fiction, Comedy and superhero genre.

- This is the best section to invest in and open a new movie studio company on as it already has a large audience dominated by BV studios.

Most Grossing Movie Both Domestically and Foreign and the studio

	Movie_id	title	studio	domestic_gross	foreign_gross	total_gross	year
	727	Marvel's The Avengers	BV	623400000.0	895500000.0	1.518900e+09	2012

5.1 Finding the studio with the most number of movies

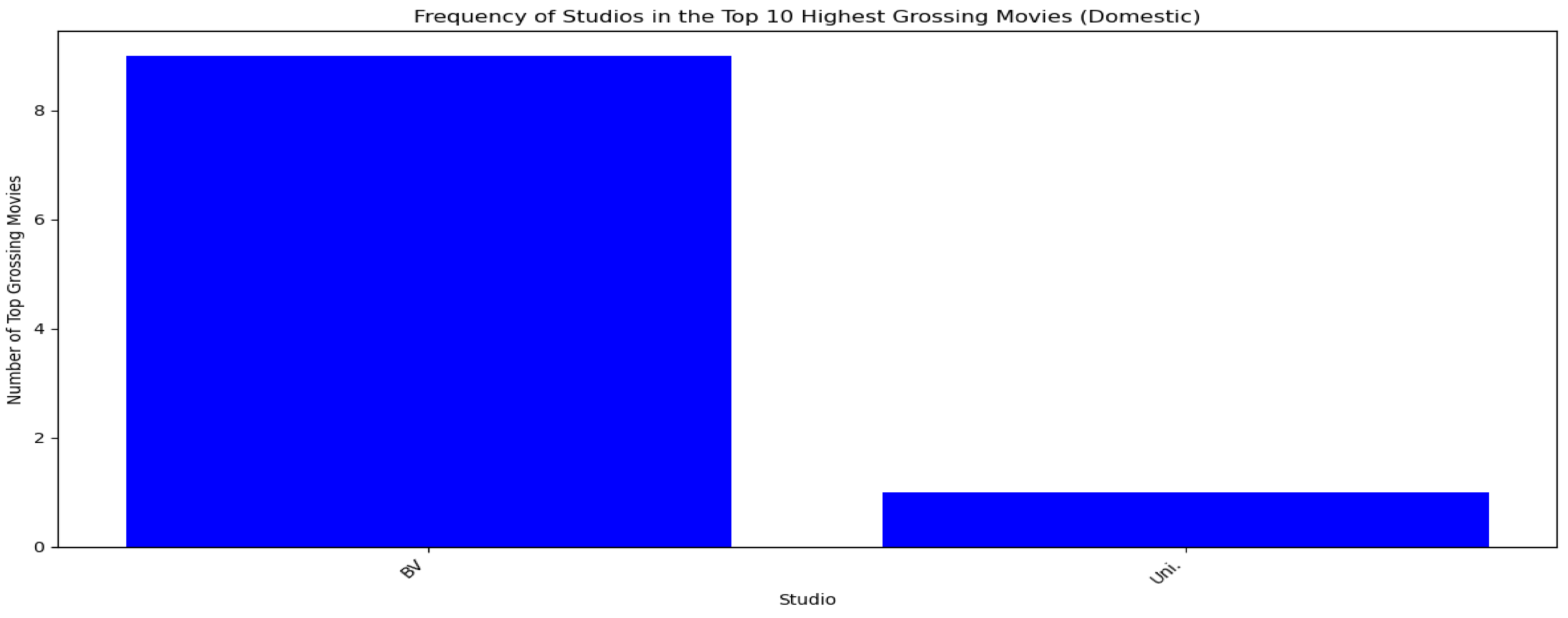
Studio with the most number of movies: IFC:166 movies

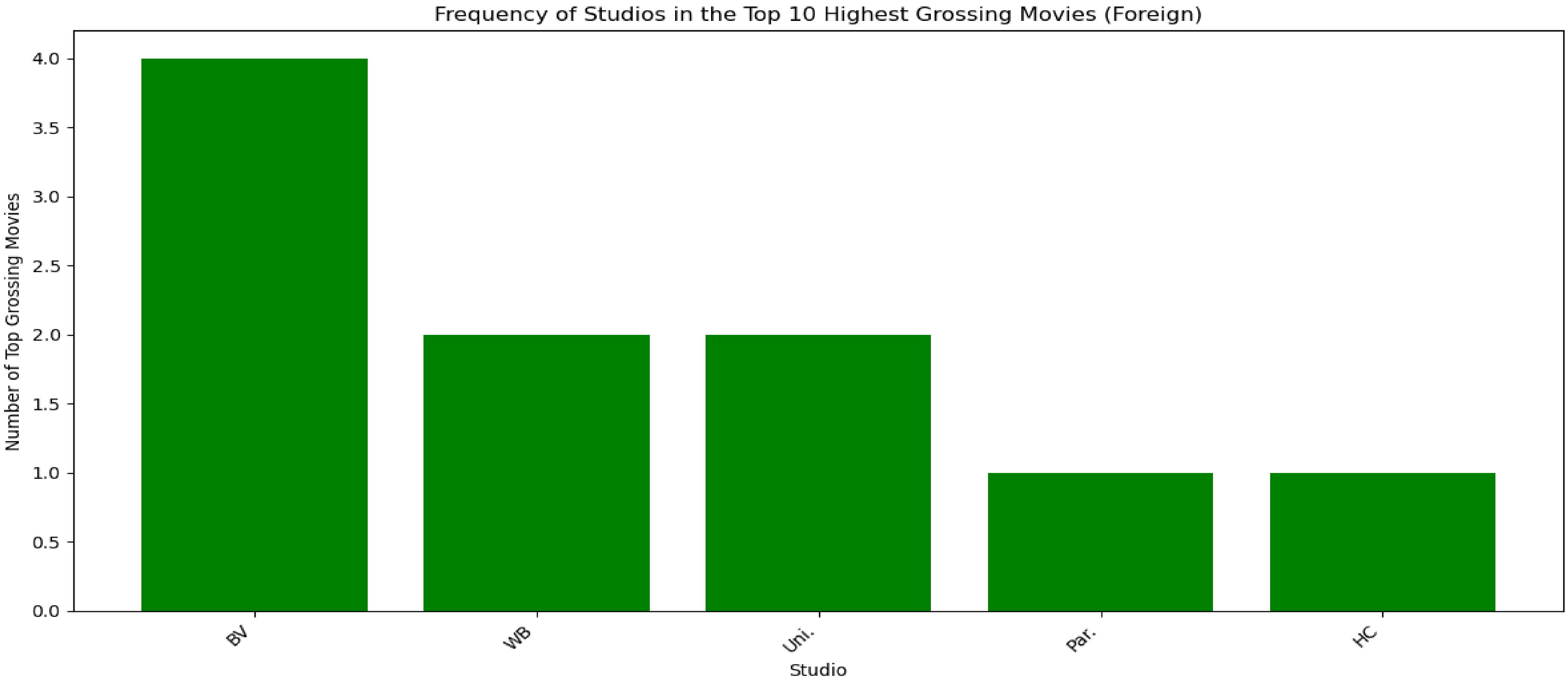
This company has the highest number of movies but does not appear in top get gross income both domestically and foreign regions.166

- Studio with the mos

5.2 HISTOGRAM ANALYSIS

- 1.1 The studio with highest domestic gross and foreign gross





The graphs above show the frequency of studios in the top 10 highest grossing movies for both domestic and foreign markets. From the analysis:

Warner Bros. leads in both categories, having the highest number of top-grossing movies in both domestic and foreign markets with a count of 5 movies each.

This insight can guide Microsoft to consider collaborating with or benchmarking against Warner Bros. for their new movie studio.

5.3. The SQL im.db database visualization

In the SQL table, I worked with these two tables movie basics and movie ratings

- The movie ratings table has no missing entries.
- -This means that no cleaning needed.

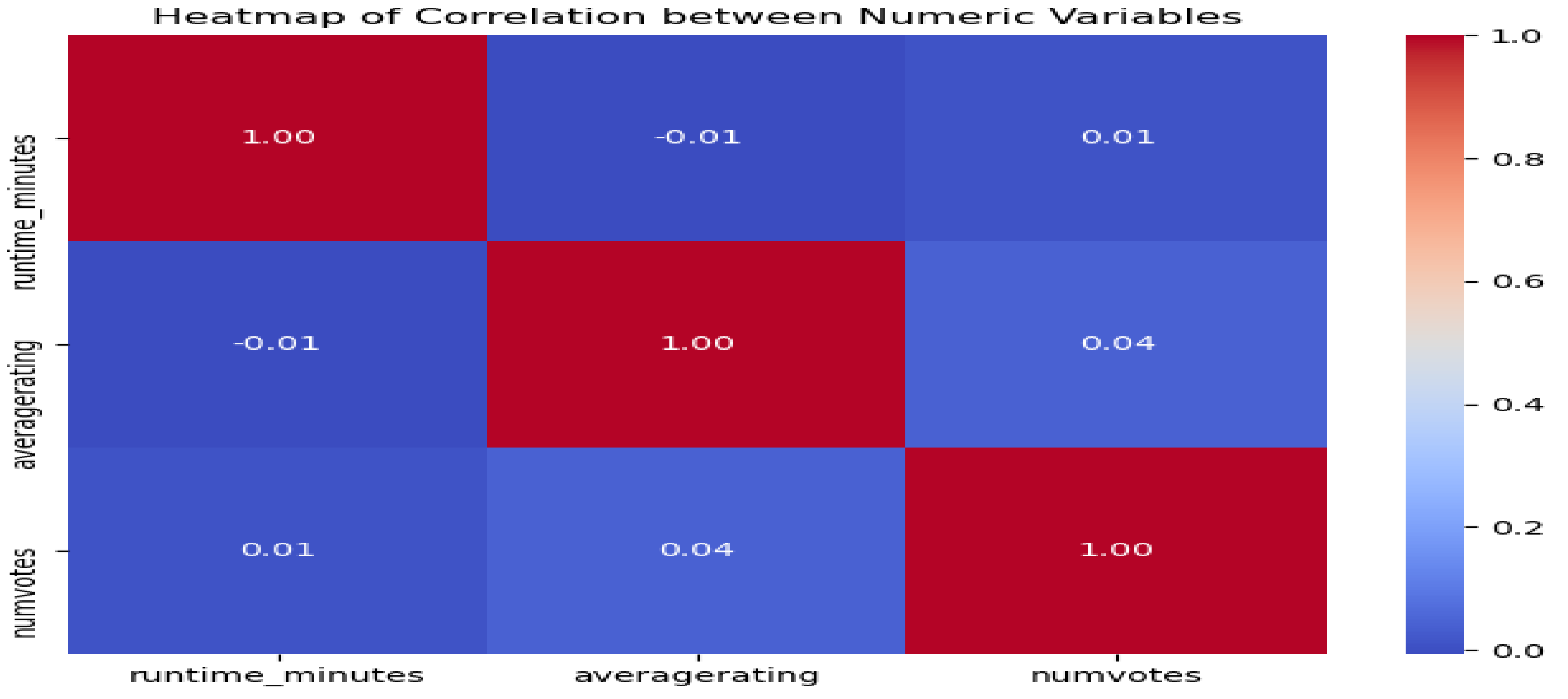
- I cleaned the movie_basics table with the missing column using the formula below.
- # Calculate the percentage of missing values for each column
- missing_percentage = movie_basics_df.isnull().mean() * 100

5.3.1 Merging of the two tables on the im.db database

- Merging the two tables using JOIN query and filling in th missing entries for each column
- This will help in efficient analysis, cleansing of the data and combining data.
- For example, from the data base, the movie id column in the two tables, movie basics and movie rating had to be merged and the two columns from the movie ratings, (averagerating and numvotes) being only two columns merged into the tabel movie basics.
- The im.db data base analysis visualization falls below

Heatmap correlation.

Visualizing the correlation between numeric variables such as runtime, average rating, and number of votes. This will help understand how these variables are related to each other.



Discussion

- From the table, Correlation coefficients range from -1 to 1, where:
  - 1 indicates a perfect positive correlation.
  - -1 indicates a perfect negative correlation.
  - 0 indicates no correlation.

More insights

- **1.1Correlation between runtime minutes and average rating:**
  - Correlation Coefficient: -0.01
  - Interpretation: There is an almost negligible negative correlation between the runtime of a movie and its average rating.

- This suggests that the runtime of a movie does not significantly affect its average rating.
- **Correlation between runtime minutes and Num votes:**
  - Correlation Coefficient: 0.01
  - Interpretation: There is an almost negligible positive correlation between the runtime of a movie and the number of votes it receives.
  - This indicates that the runtime of a movie does not significantly affect the number of votes it gets.

Conti….

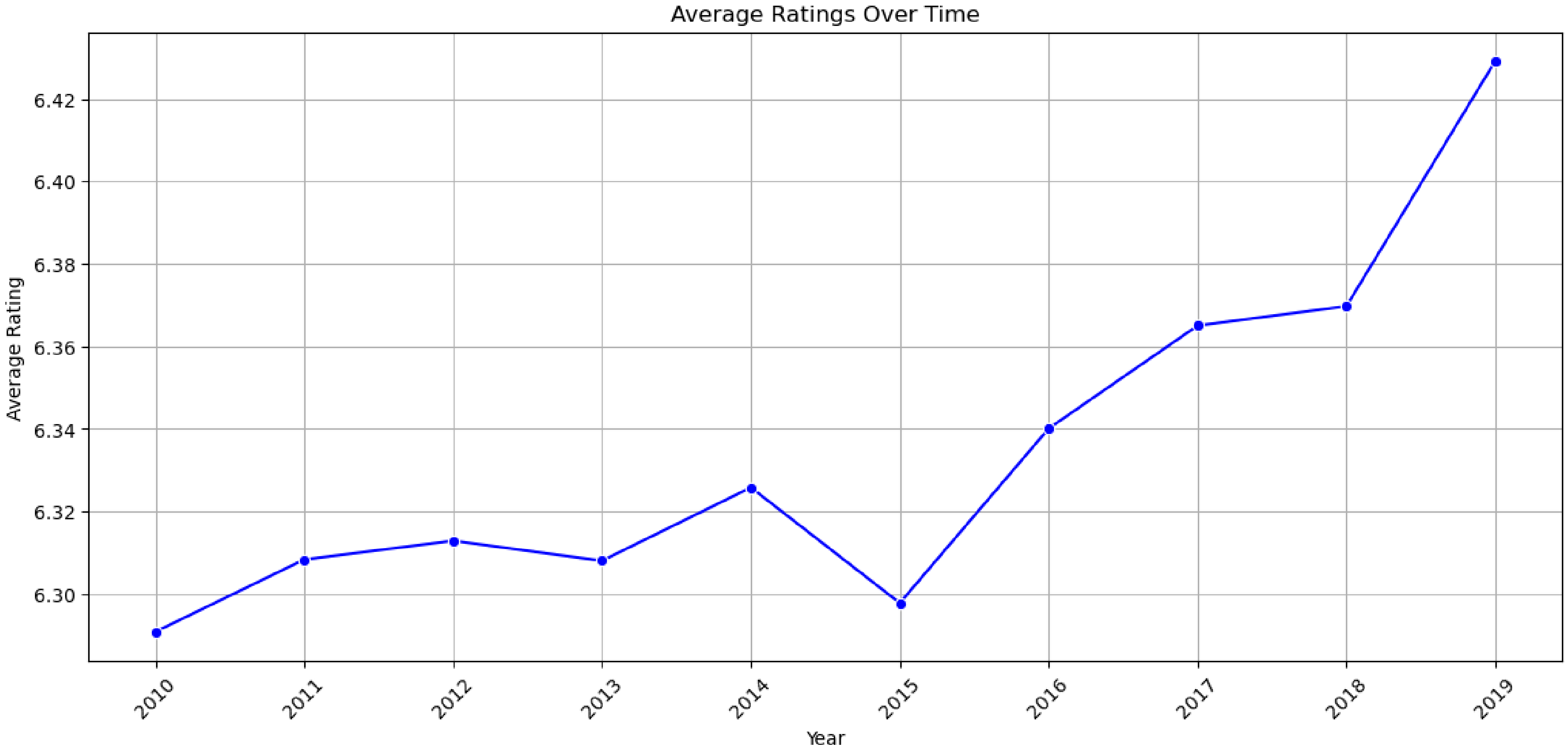
Correlation between averagerating and numvotes:

- Correlation Coefficient: 0.04
- Interpretation: There is a very weak positive correlation between the average rating of a movie and the number of votes it receives. This implies that movies with higher average ratings tend to receive slightly more votes, but the relationship is very weak.

- Overall, the correlation analysis suggests that runtime_minutes, averagerating, and numvotes are largely independent of each other in the given dataset.

Line Plot of Average Ratings Over Time

Plot the average ratings of movies over time to see if there are any trends or patterns in how ratings have changed over the years



Line graph explanation

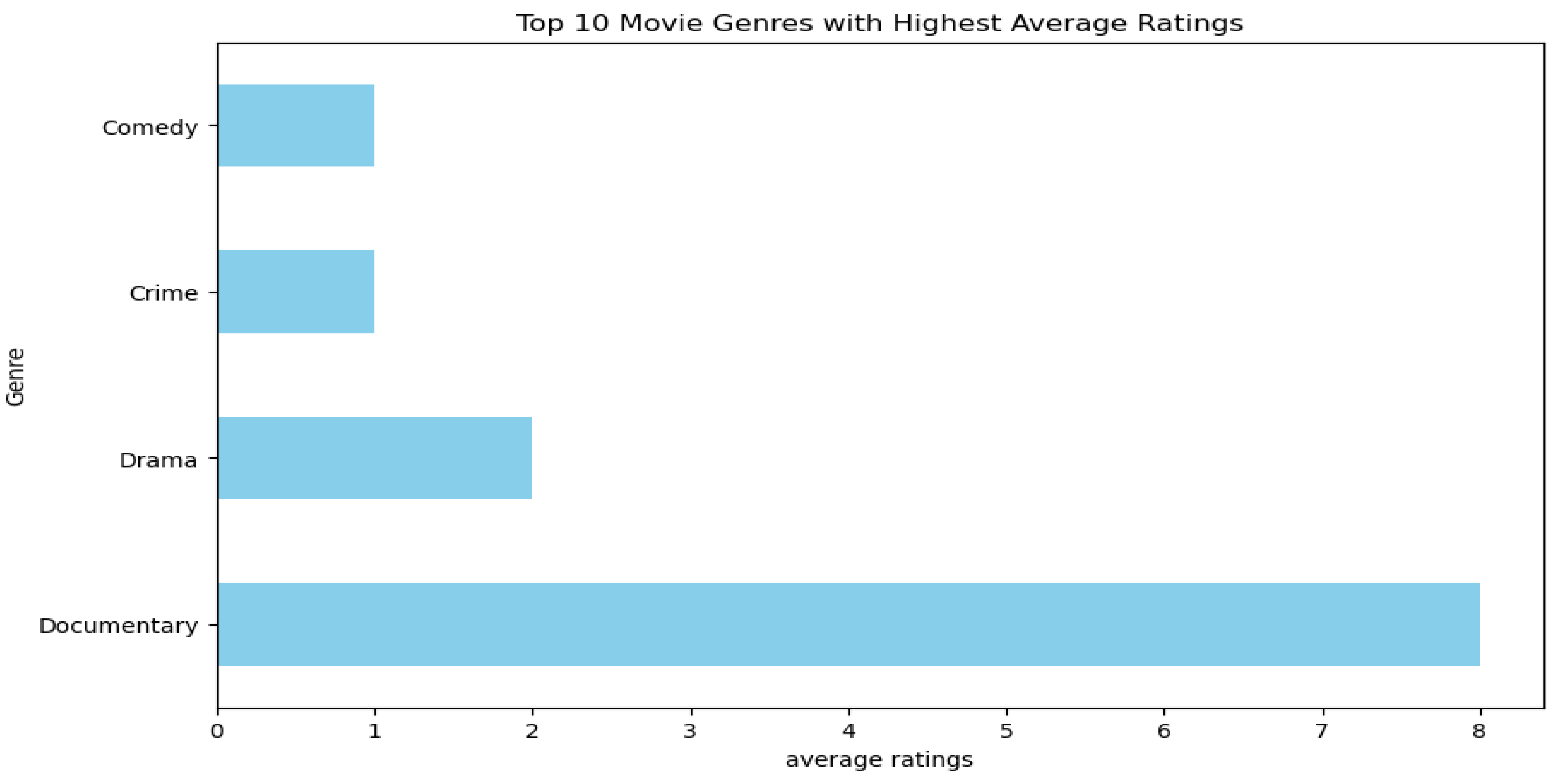
- It is evidently clear that the average number of ratings has been increasing over the years steadily.

- The year 2019 had the highest averaged number of votes.

- This might have been directly influenced by the increased streaming platforms and heavy social media promotions.

-

Graphical representation of the top 10 genres with the highest rating



Graphical representation of the top 10 genres with the highest rating

- As per the graphical analysis, it is clear that the Documentary genre is leading in the number of averaged ratings.
- The remaining three are, drama, crime and comedy with a little it lower averaged numbers.
- This is clear indication to Microsoft team that we can venture into other genre such as the documentary as it has the highest averaged ratings.

6.0 Analysis Findings

- From the data analysis and with the help of graphical representation
  - The best performing genres in the movie industry is the Action, Sci-fi, documentaries, superhero genres.

- **History and Documentary Genres**:
  - Positioned closely together with moderate ratings (~6.65) and runtimes (~70 minutes).
  - Shows that these genres maintain a balance and are moderately appreciated by audiences.
- It is quite clear that the BV movie studio is doing exceptionally well as both from the above analysis are grossing well both domestically and in foreign regions.

- This means that the majority of the audience fans of the Action, fiction, Comedy and superhero genre.

- This is the best section to invest in and open a new movie studio company on as it already has a large audience dominated by BV studios.

Recommendations for Microsoft's New Movie Studio:

- **Heavy capitalize on the high domestic and foreign income from the top movies like the high performing studios such as BV studios.**
  - The genre of such movies has a large number of audience and that is why it is doing well both locally and internationally.
- Capitalize on Moderate Genres (Documentary, Biography, Music, Sport):
  - **These genres balance runtime and ratings well. They can be a safe bet for producing content that is likely to be well-received by a wide audience.**
- **Experiment with Shorts:**
  - Although they have low ratings, shorts have the advantage of being low-risk due to their shorter length. Experimenting with high-quality short films might capture a niche market

Next steps…

- **Data Collection:**
- Gather additional data on factors like marketing spend, social media engagement, audience demographics, and critical review
- **Data Enrichment:**
- Integrate external datasets such as box office performance, awards won, and streaming platform availability.

Appreciation

- **Thank You**
- Thank you for your attention! I’m happy to answer any questions you may have.
- Feel free to reach out to me via LinkedIn for further discussions or collaborations.
- **Your Name**

- LinkedIn: https://www.linkedin.com/in/joseline-odhiambo-094a9122b/