Phase 1 project

Kelvin Kipyegon Rotich

DSF-FTO6

15/09/2023

Overview

- Microsoft is a multinational technology corporation that was founded by Bill Gates and Paul Allen in 1975. It has since become one of the world's leading software and technology companies. Its early success came from its operating system, MS-DOS, which was later succeeded by Windows.
- The company's business domain spans a wide range of products and services, including operating systems for personal computers and servers, productivity software services like Microsoft Office, cloud computing with services like Azure, hardware such as the Surface line, and a variety of other applications and services.
- Microsoft's business case centers on its ability to provide essential software and technology solutions to both consumers and enterprises. With a dominant position in the operating systems market for personal computers, as well as a strong presence in office productivity software, Microsoft enjoys a significant user base. Additionally, their foray into cloud computing and services like Azure has allowed them to tap into the rapidly growing cloud market.
- Overall, Microsoft's diverse product offerings, strategic acquisitions and emphasis on cloud computing have contributed to its continuous success and status as a tech industry leader.

Business understanding

- The movie business has been around from as early as 1895.
- The industry was dominated by movie studios such as Warner Bros, 20th Century Fox and Universal.
- In recent times however, tech companies such as Netflix, Apple and Amazon have joined in on the fun and had success.
- This has enticed some interest to Microsoft and they want to start their own movie creation company but the down side to their goal is that they do not have any knowledge of the movie business.
- This analysis will help them get some insights into the business and see what they need in order for their project to succeed.

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.
- You are charged with exploring what types of films are currently doing the best at the box office.
- You must then translate those findings into actionable insights that the head of Microsoft's new movie studio can use to help decide what type of films to create.

Objectives

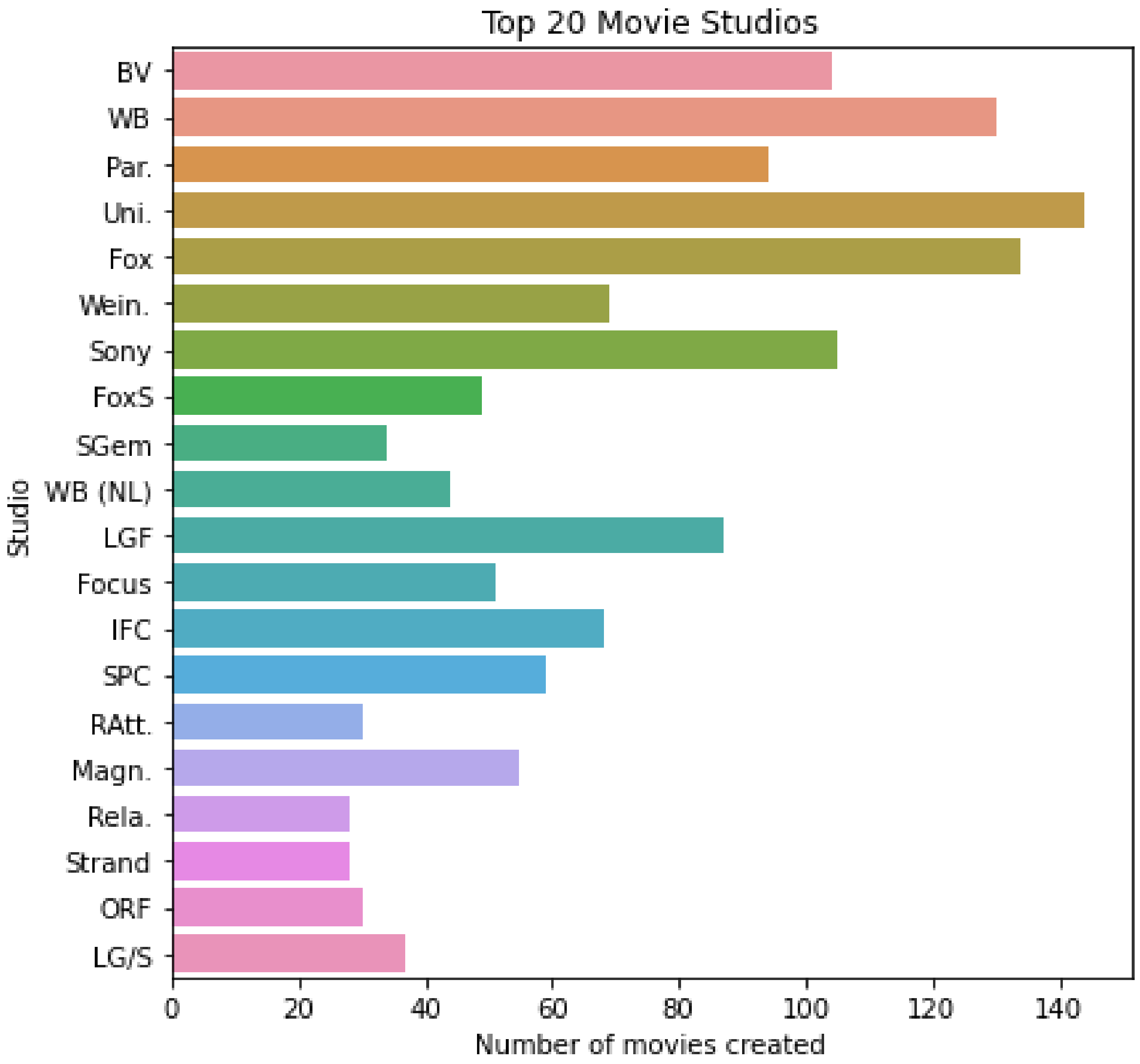
- To show how profitable the movie industry is based on the data.

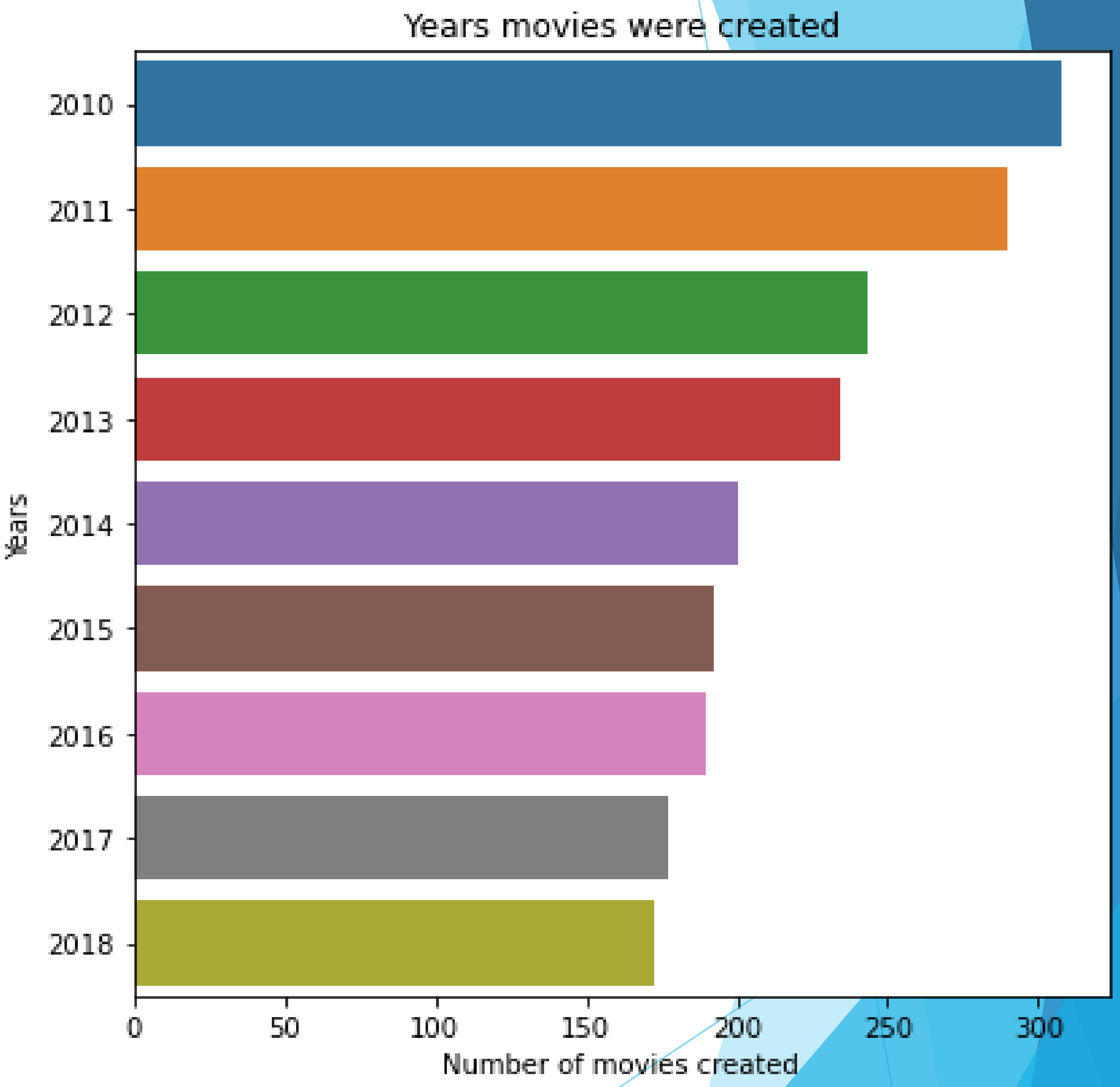
- To get some of the factors that lead to success in the movie industry based on the data.

- To look for the strong relationships between some of these factors.

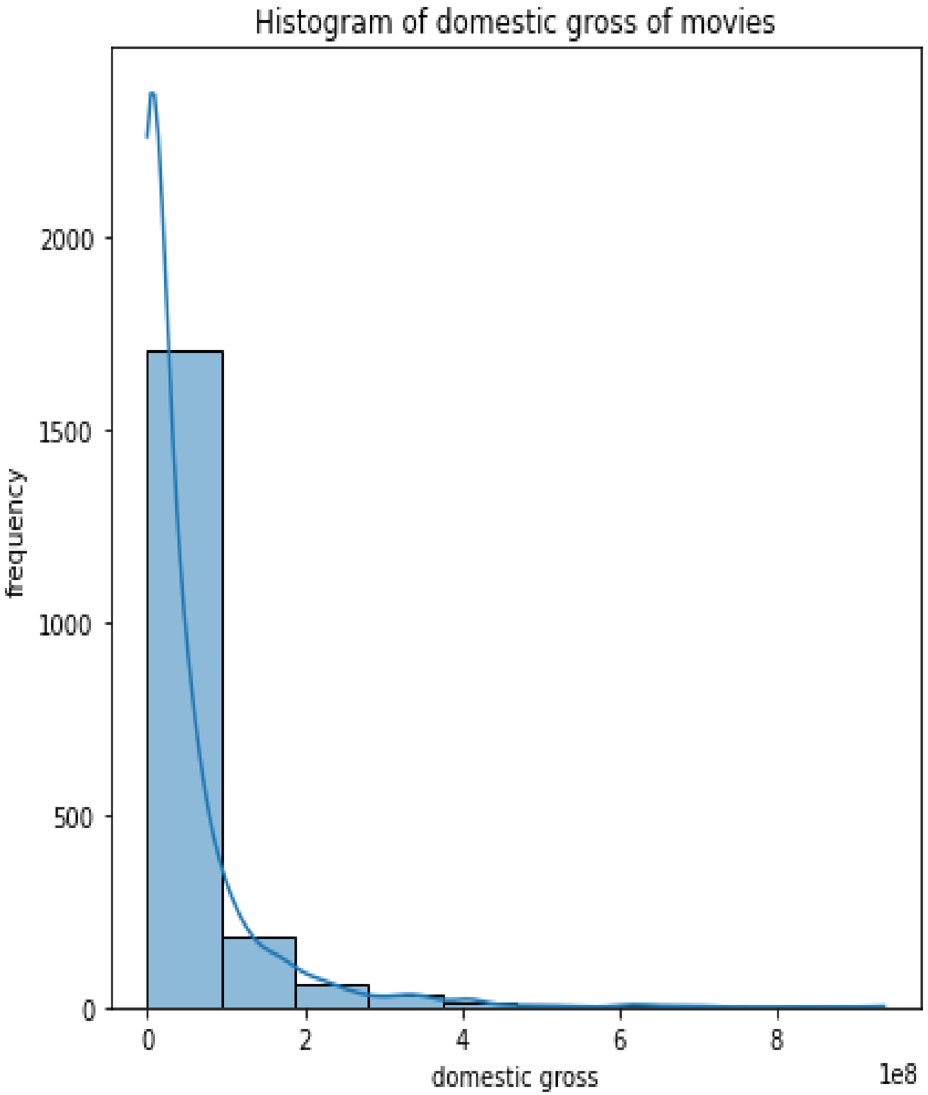
Data Analysis

- Here we analyzed one table in each of the three data frames.
- We began with the first data frame.
- First we looked at the categorical columns of the data.
- We focused on the `studio` and `year` columns.
- We found the top studios that released a lot of movies.
- We also saw the years that these movies were created and we plotted some graphs for representation.

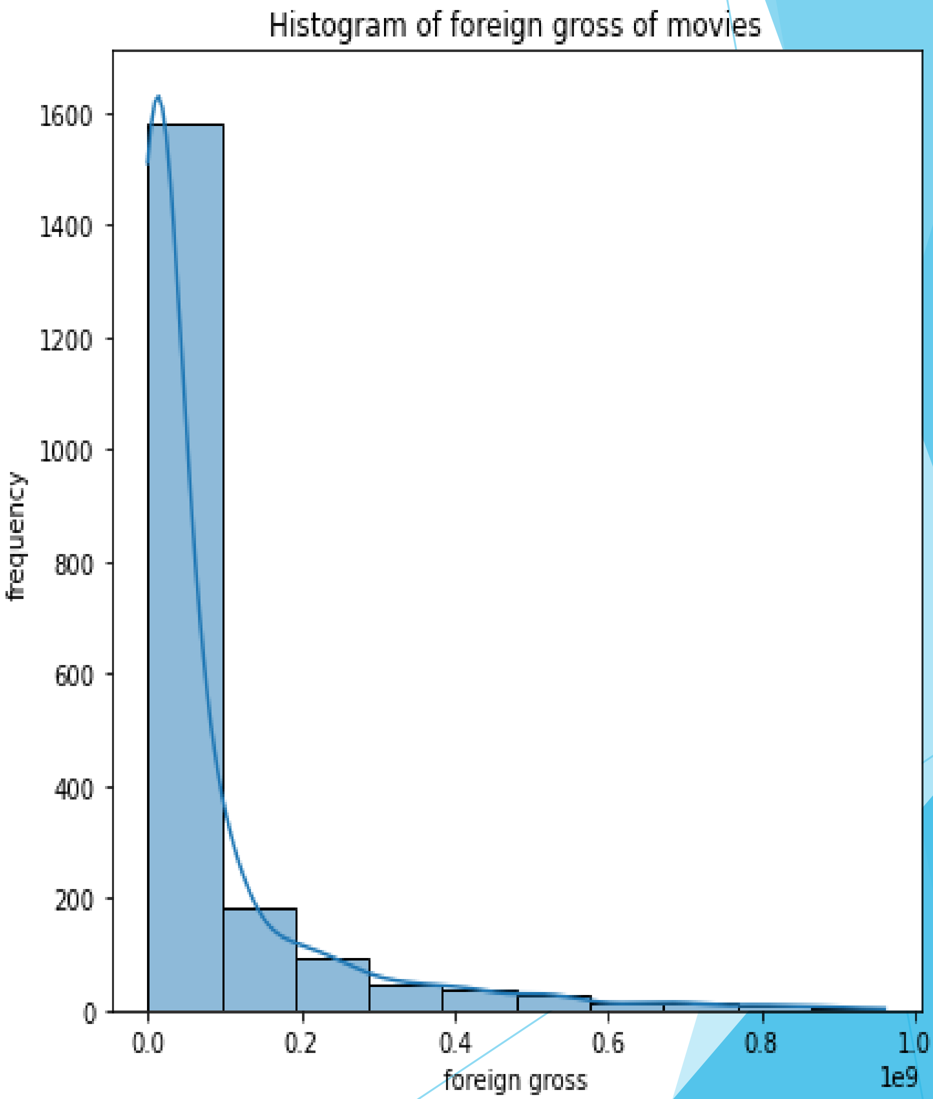

Top 20 Movie Studios
Studio
BV
WB
Par.
Uni.
Fox
Wein.
Sony
FoxS
SGem
WB (NL)
LGF
Focus
IFC
SPC
RAtt.
Magn.
Rela.
Strand
ORF
LG/S
0
20
40
60
80
100
120
140
Number of movies created


Years movies were created
Years
2010
2011
2012
2013
2014
2015
2016
2017
2018
0
50
100
150
200
250
300
Number of movies created

- From our analysis, there were 172 movie studios.
- The top 5 studios were `Uni.`, `Fox`, `WB`, `Sony` and `BV`.
- We then went to the numeric data analysis. Here, we observed `domestic_gross` and `foreign_gross`.
- We looked at the measures of central tendency and measures of dispersion.
- We also plotted some graphs for representation.

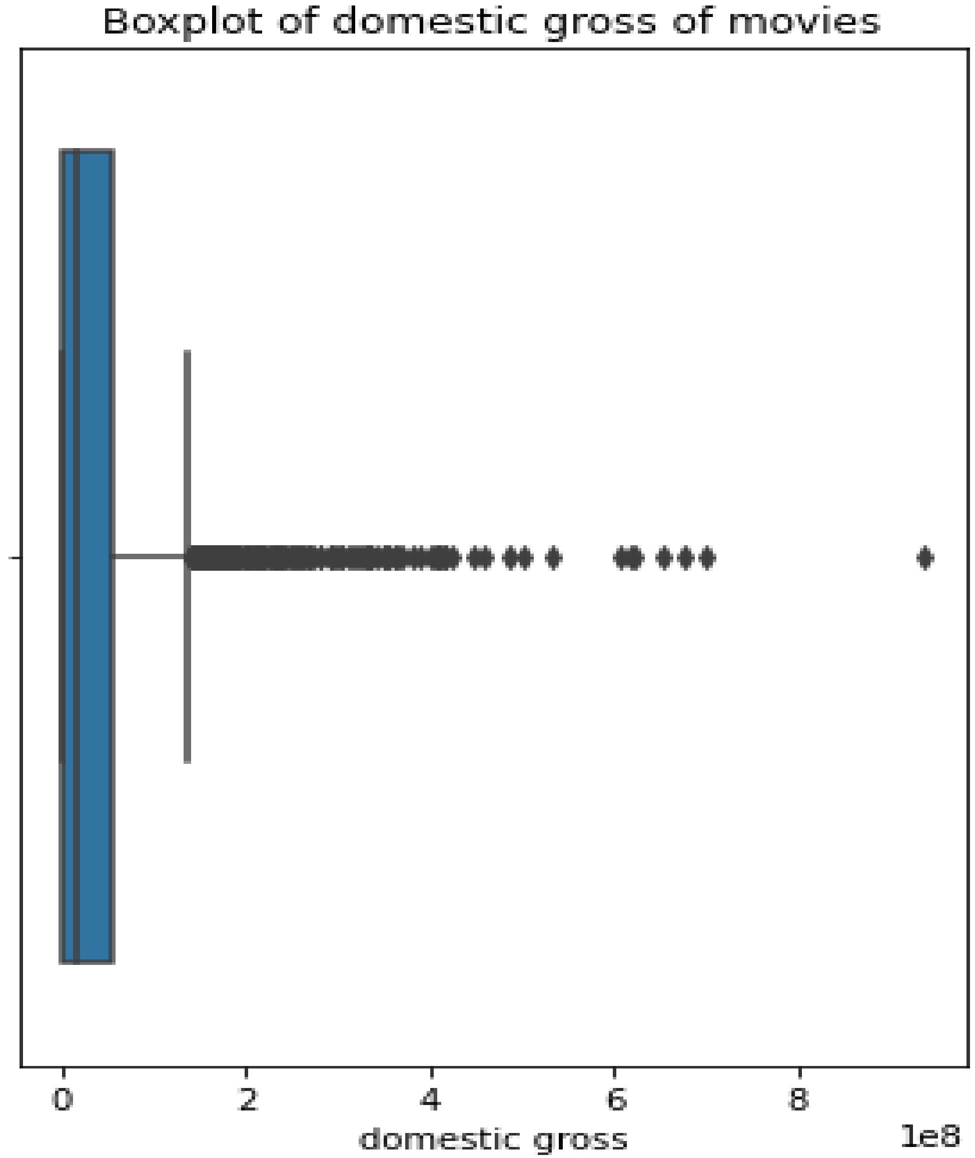


Histogram of domestic gross of movies

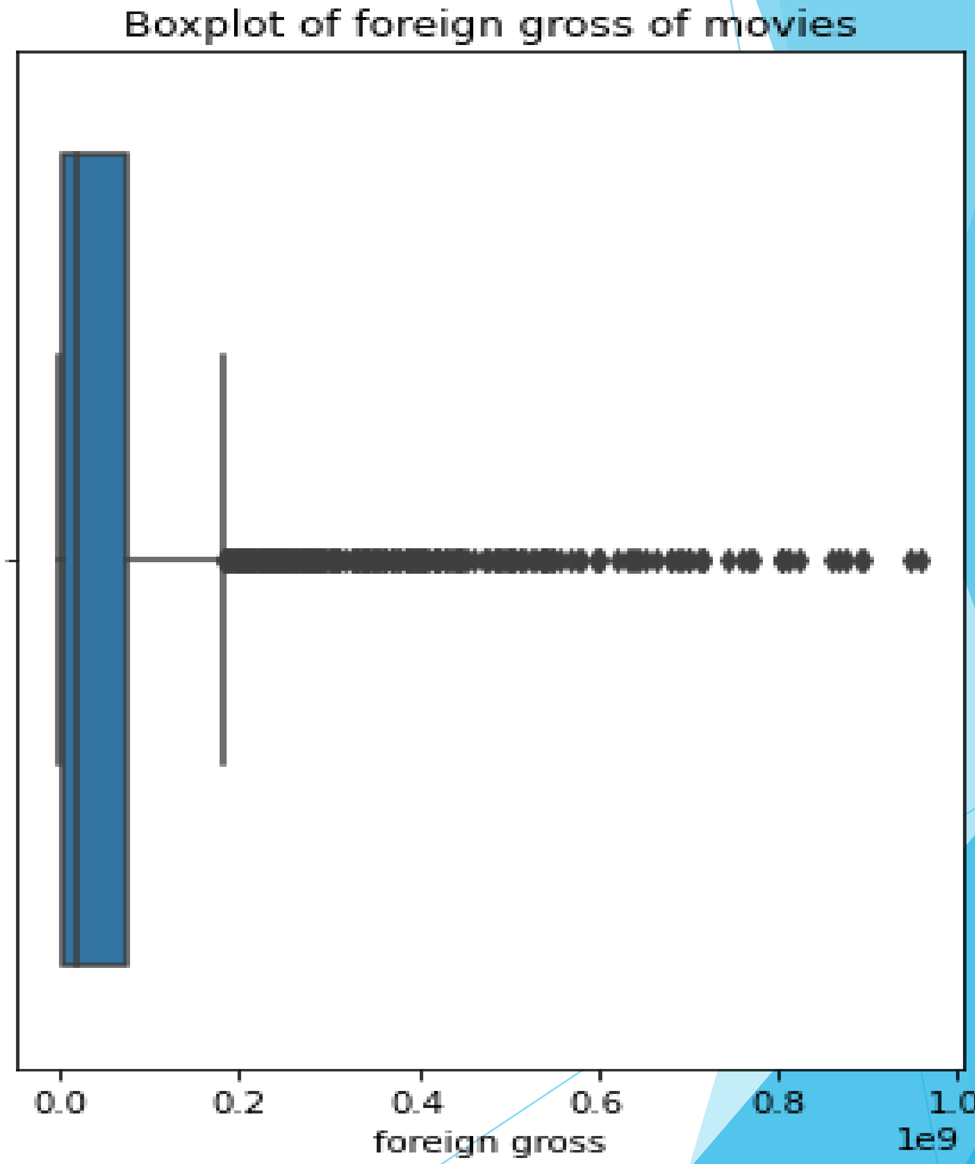


Histogram of foreign gross of movies

- The mean for domestic gross and foreign gross were $47019840 and $75790385 respectively.
- Their medians were $16700000 and $19400000 respectively.
- Based on the mean, median and the histogram, we saw that both datasets were positively skewed since the median were lower than the mean.
- This means that majority of the data was concentrated on the left side of the distribution and there were relatively few extreme values on the right side.
- However, this was not enough for the study.
- We also needed to find the measures of dispersion to see how the values were far from the mean. We also plotted boxplots for representation.

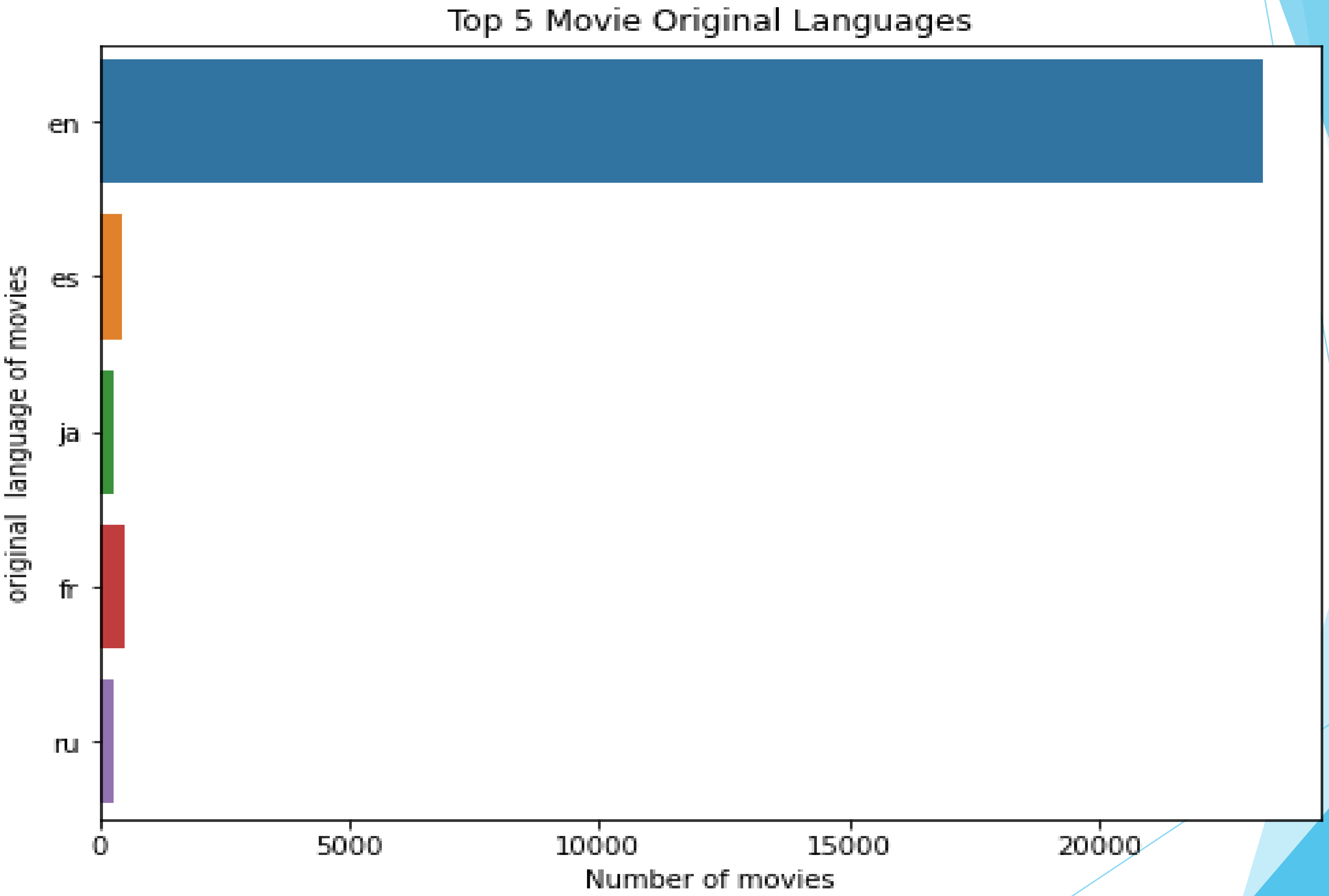


Boxplot of domestic gross of movies

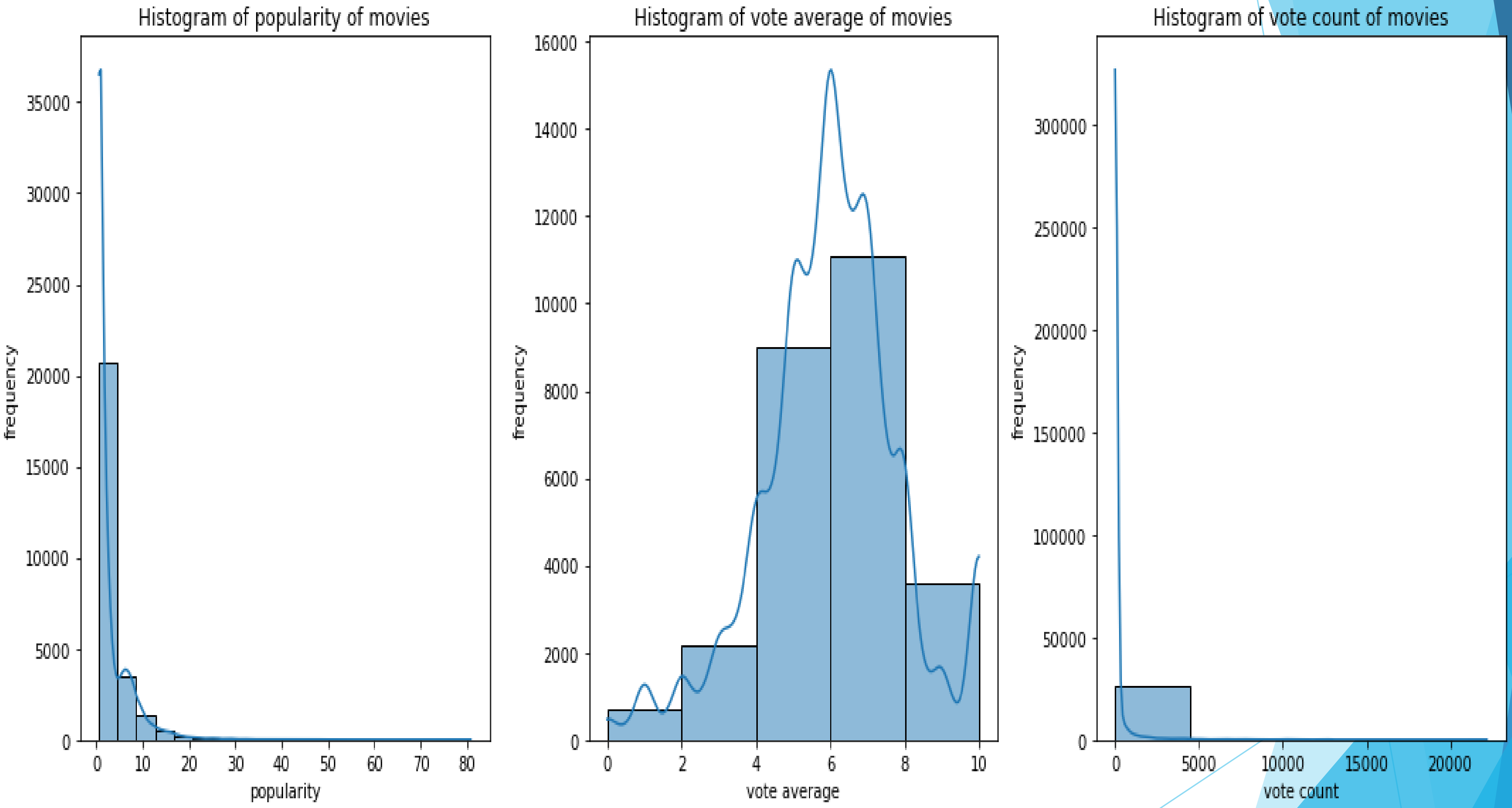


Boxplot of foreign gross of movies

- The standard deviation, lower and upper quartiles for domestic gross were $81626889, $670000 and $56050000 respectively while those of the foreign gross were $138179553, $3900000 and $75950000 respectively.
- Based on these values and the boxplots, we saw that in both data sets, most of the data was seen spread out over a wider range relative to the mean.
- This analysis fulfilled our first objective which was to show profitable the movie business is.
- We then conducted analysis on the second data frame. We focused on \`original_language\`, \`popularity\`, \`vote_average\` and \`vote_count\`.
- Our categorical column in this case was \`original_language\`. We plotted a graph for representation.


Top 5 Movie Original Languages
original language of movies
en
es
ja
fr
ru
0
5000
10000
15000
20000
Number of movies

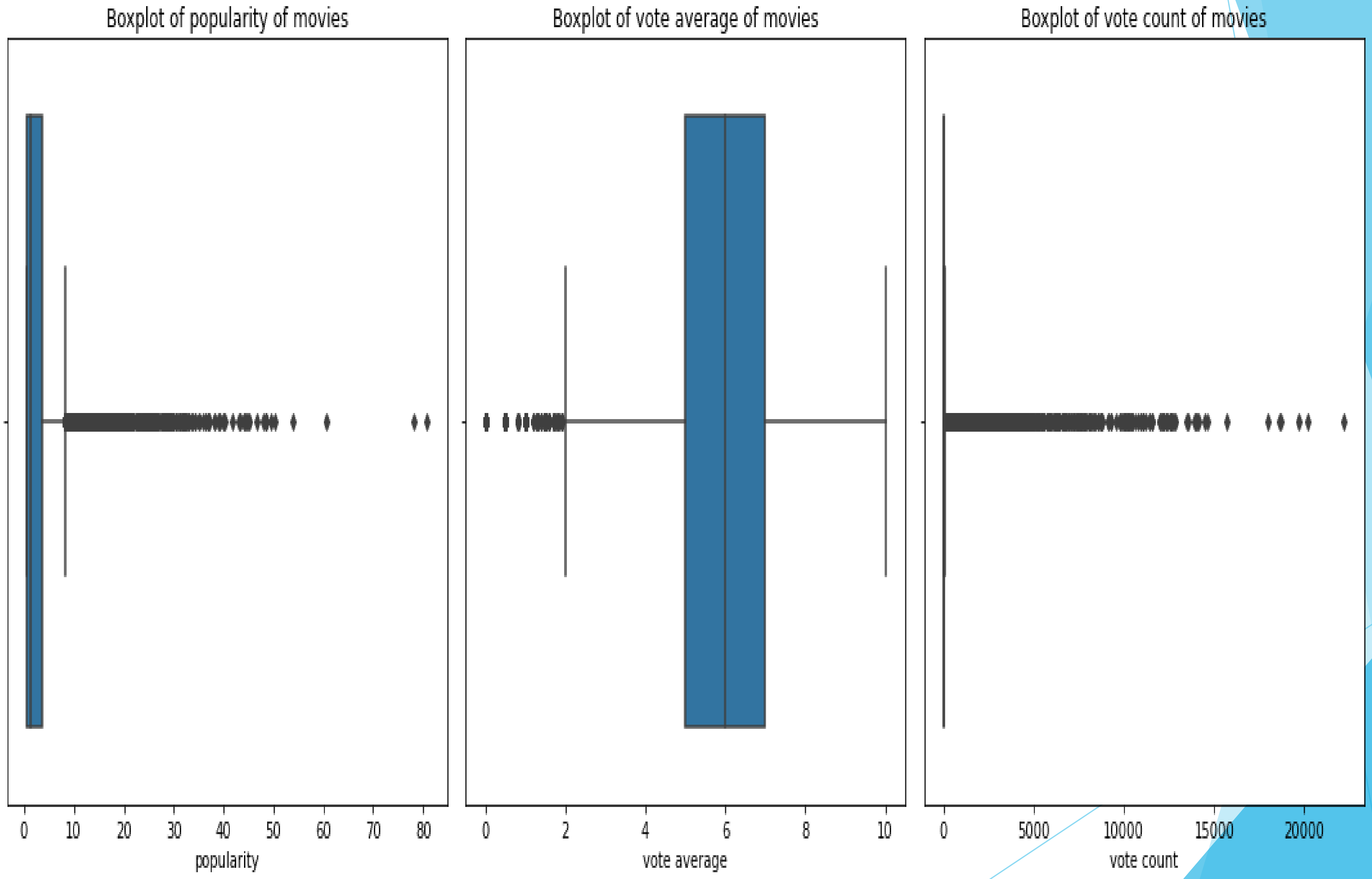
- We noticed that most of the movies had their original language as English followed by Spanish, Japanese, French and Russian.
- This means that a lot of movies consider English as the original language for their movies.
- This could be as a result of many successful movies being in English.
- However, we are not certain of this success because we don't have a way to quantify the success of the movies based on the language because we need to have a revenue column which is not in the data frame.
- With that we then went and took a look at the numeric columns.
- Here, we focused on `populartity`, `vote_average` and `vote_count`. We checked the measures of central tendency and measures of dispersion as before and plotted the graphs as before.



Histogram of popularity of movies

Histogram of vote average of movies

Histogram of vote count of movies



Boxplot of popularity of movies

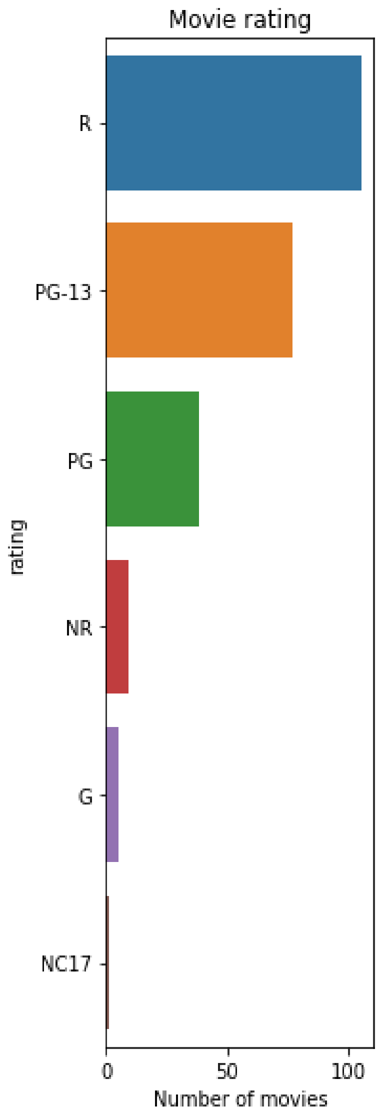
Boxplot of vote average of movies

Boxplot of vote count of movies

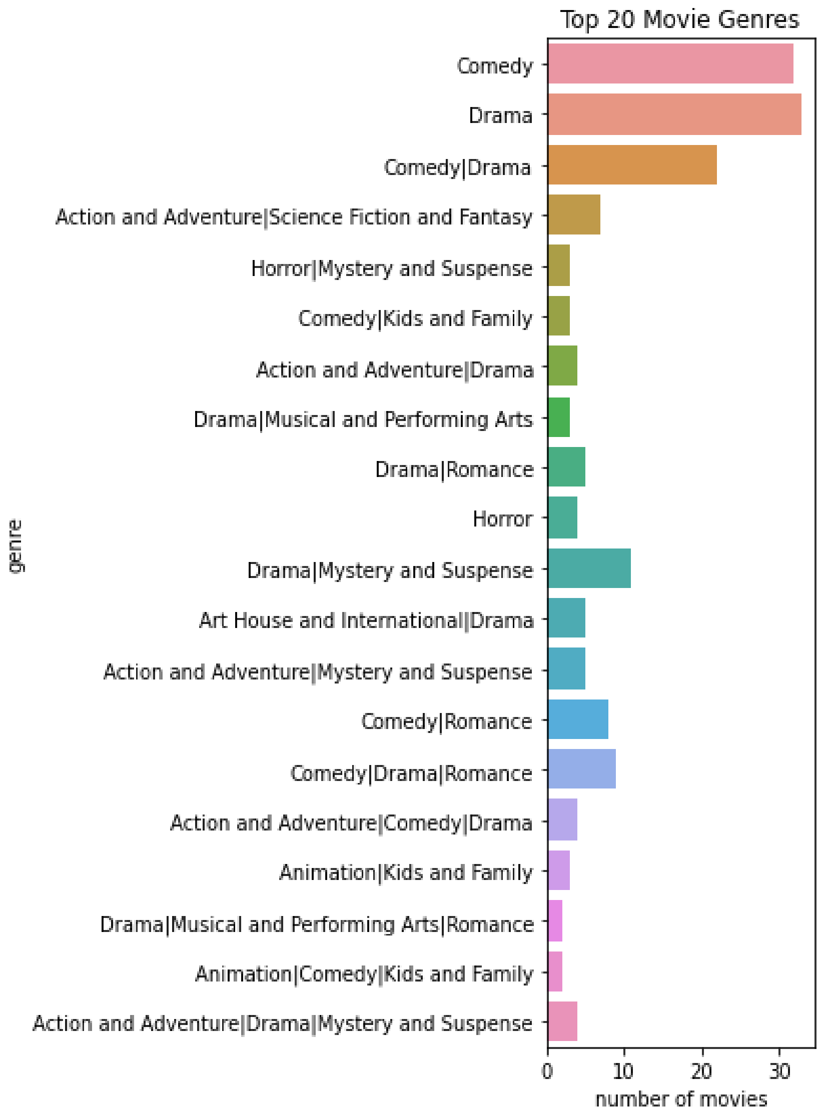
- We analyzed the popularity and vote counts columns. Their means were 3 and 194 and their medians were 1 and 5 respectively.
- Based on the mean, median and the histograms, we can see that both datasets were positively skewed since the median are lower than the mean.
- This means that majority of the data was concentrated on the left side of the distribution and there are relatively few extreme values on the right side.
- However, the vote average has a different shape.
- We can see that the histogram had a symmetrical shape. We also saw that the mean and median were equal since it was 6.
- This means that the values are spread out on both sides of the central point of the data.

- We also found the measures of dispersion to see how the values were far from the mean. We focused first on the popularity and vote count columns.
- The standard deviations of the two columns were 4 and 961 and their lower quartiles were 1 and 2 and their upper quartiles were 4 and 28 respectively. Based on their standard deviation, quartiles and boxplots, we saw that in both data sets, most of the data can be seen spread out over a wider range relative to the mean. This means that there is a high degree of variability or dispersion in the datasets.
- The vote average however has a different characteristic unlike the other two columns. Its standard deviation, lower and upper quartiles were 2, 5 and 7 respectively. We saw that the standard deviation is lower than the mean and the quartiles are close to the median. This means that the data have a relatively narrow spread and is not heavily skewed towards one extreme.

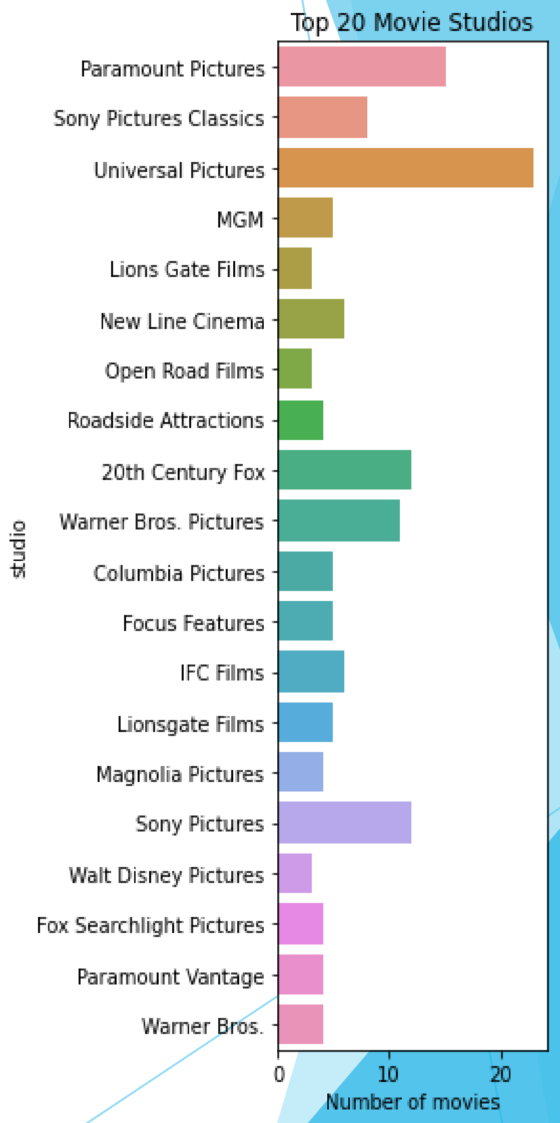
- This analysis showed that even though there were a lot of popular movies, this does not mean that the popularity of the movie is guaranteed. We needed a revenue column which would have made it easier to conduct bivariate analysis with respect to popularity. However, we did not have a revenue column.
- We then conducted analysis on the third data frame. We focused on `rating`, `genre`, `box_office` and `studio` columns.
- For the categorical analysis, we were interested in the `rating`, `genre` and `studio` columns. We also plotted the graphs for representation.



Movie rating

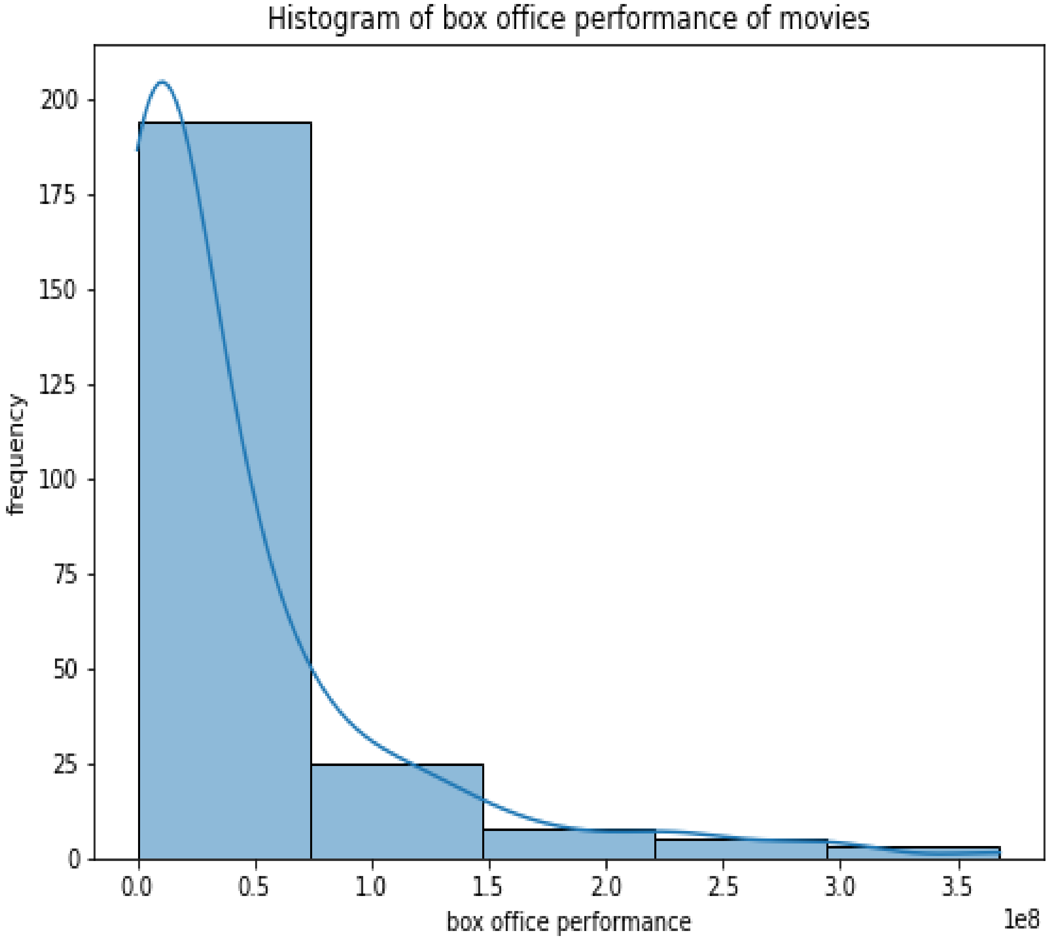


Top 20 Movie Genres

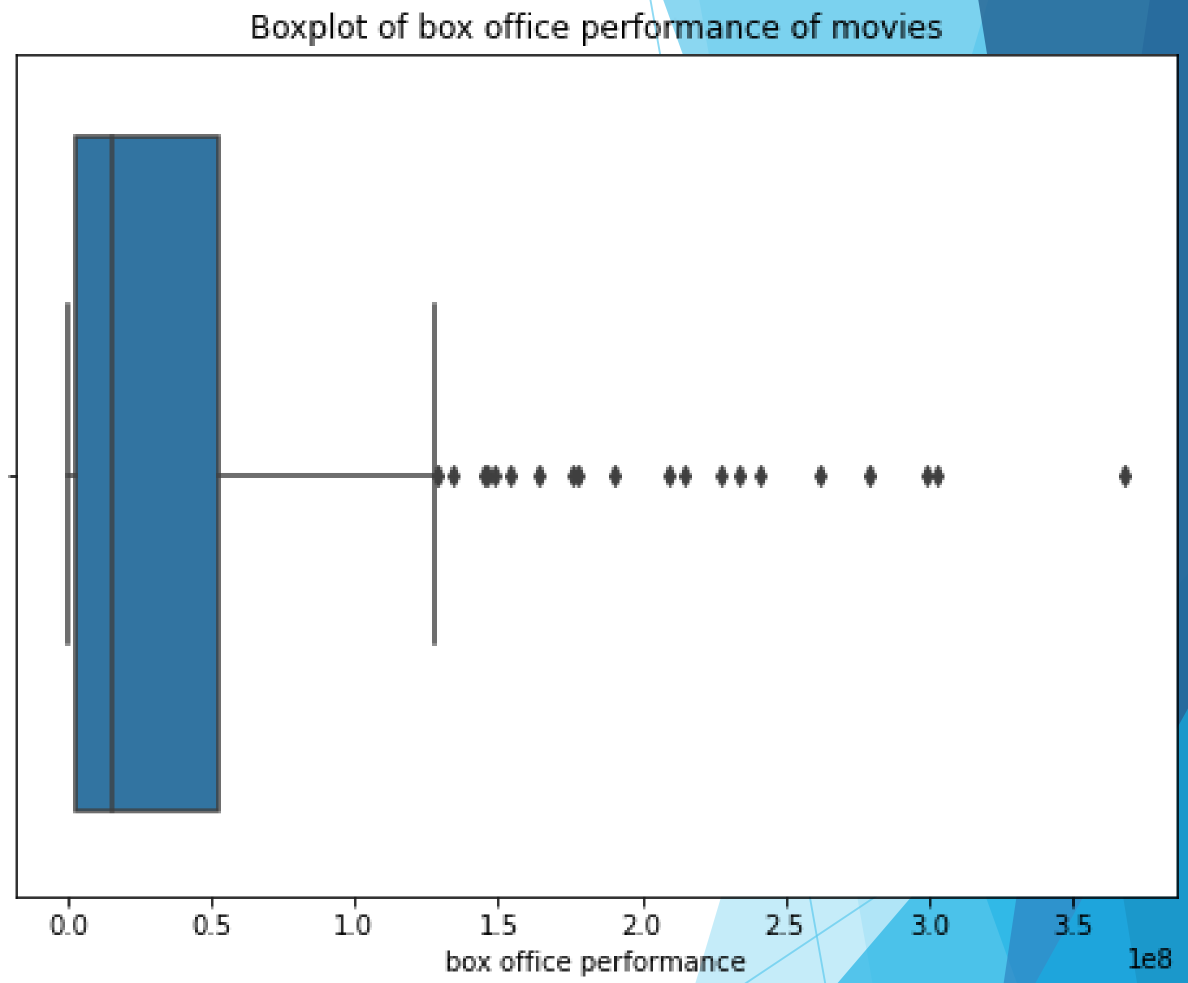


Top 20 Movie Studios

- Here saw that R-rated movies had the highest number, followed by PG-13, PG, NR, G and NC17 movies respectively.
- We also saw that the genres with a lot of movies were comedy and drama. We also noted that Universal Pictures, Paramount Pictures, Sony Pictures, 20th Century Fox and Warner Bros. Pictures were the studios with the most number of movies in the dataset.
- We then went ahead and looked at the numeric column in the dataset. For this data frame, we had only one numeric column, the `box_office` column. We then analyzed the data in the same way we did before. We also plotted graphs for representation



Histogram of box office performance of movies



Boxplot of box office performance of movies

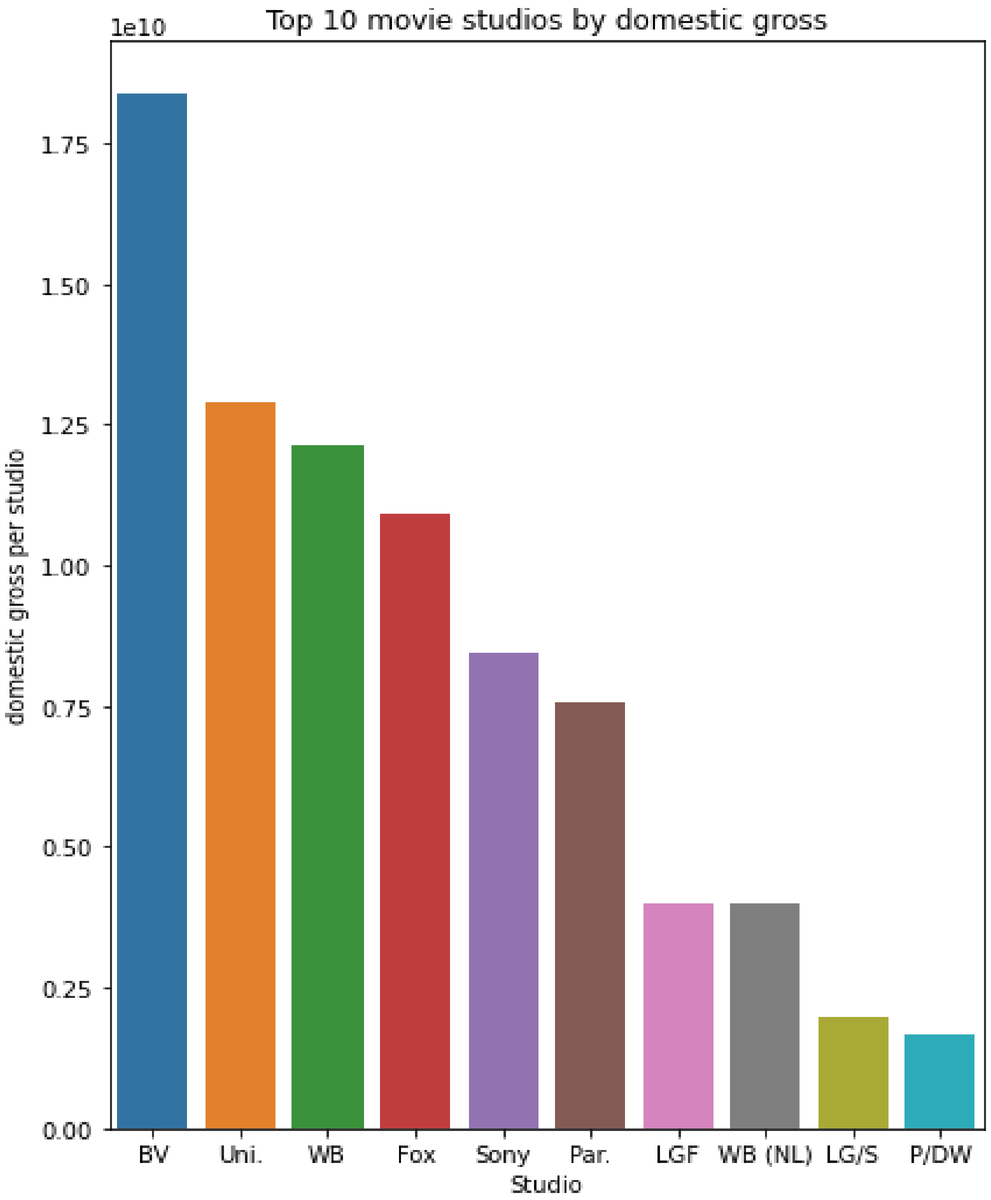
- The mean and median we found were $41958400 and $15536310 respectively.
- Based on the mean, median and the histogram, we saw that the dataset is positively skewed since the median was lower than the mean. This means that majority of the data is concentrated on the left side of the distribution and there are relatively few extreme values on the right side. However, this was not enough for the study.
- We also need to find the measures of dispersion to see how the values are far from the mean. The standard deviation, lower and upper quartiles we found were $62630156, $2302444 and $52649522.
- Based on the standard deviation, the quartiles and the boxplots, we saw that in the dataset, most of the data can be seen spread out over a wider range relative to the mean. This means that there was a high degree of variability or dispersion in the dataset.
- The univariate analysis showed us that the movie industry has earned a lot of revenue for their creation studios hence fulfilling our first objective of profitability in the movie industry.

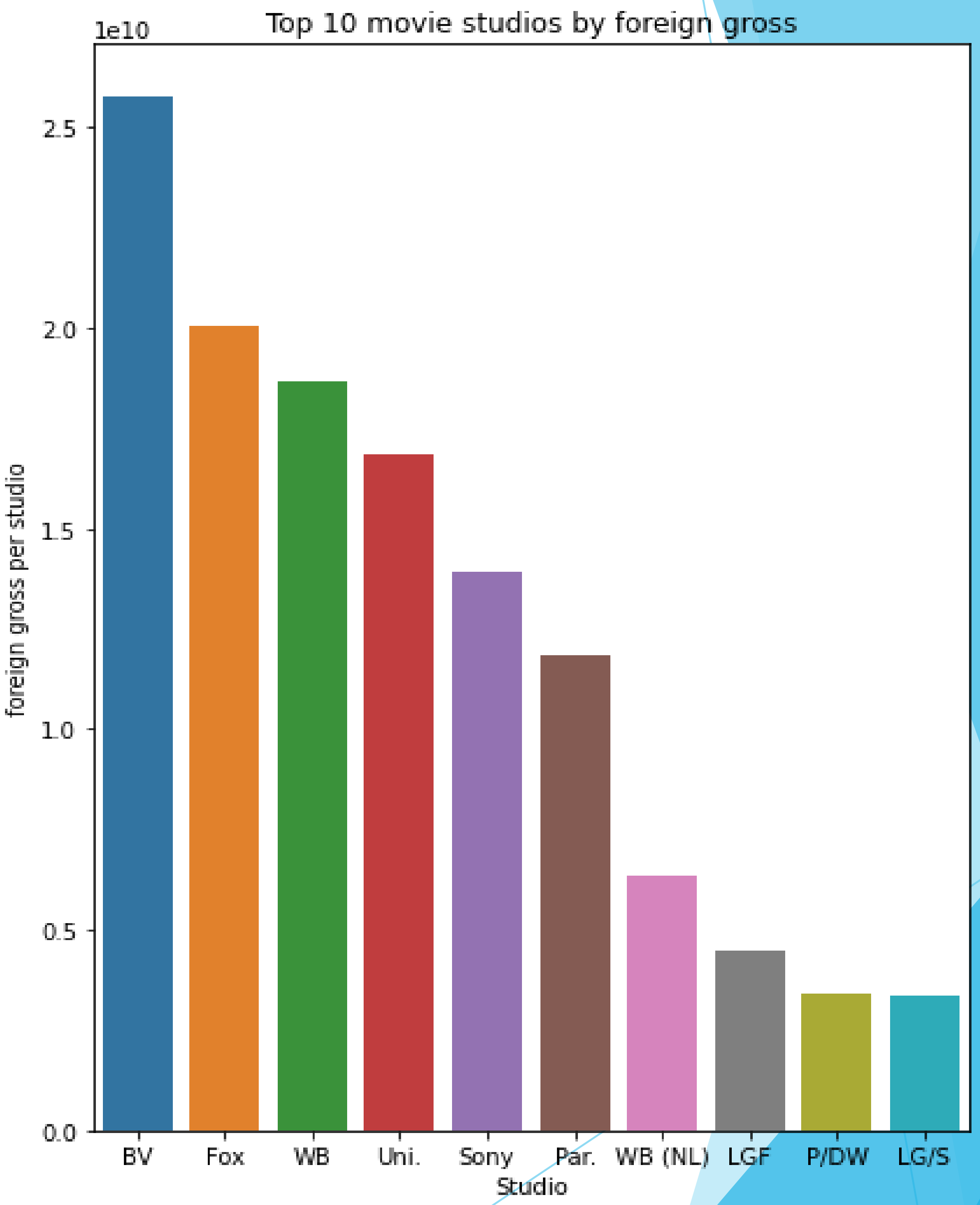
Bivariate analysis

- We began by comparing the `studio` columns with both the `domestic_gross` and the `foreign_gross` columns.

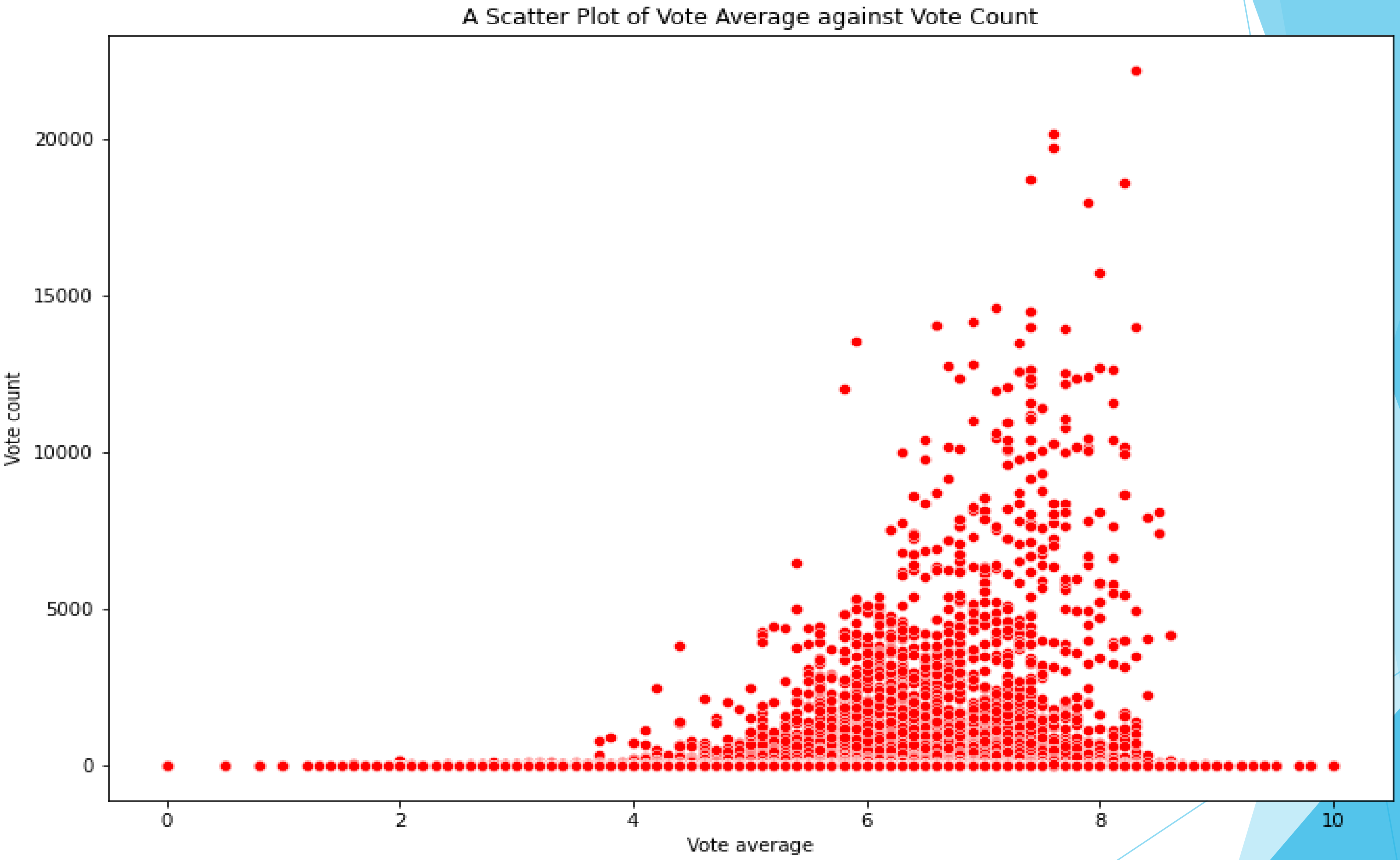
- Since we were comparing a categorical column with a numeric column, we used bar graphs for representation.

- We focused on the ten studios that earned the most revenue either domestically or globally.

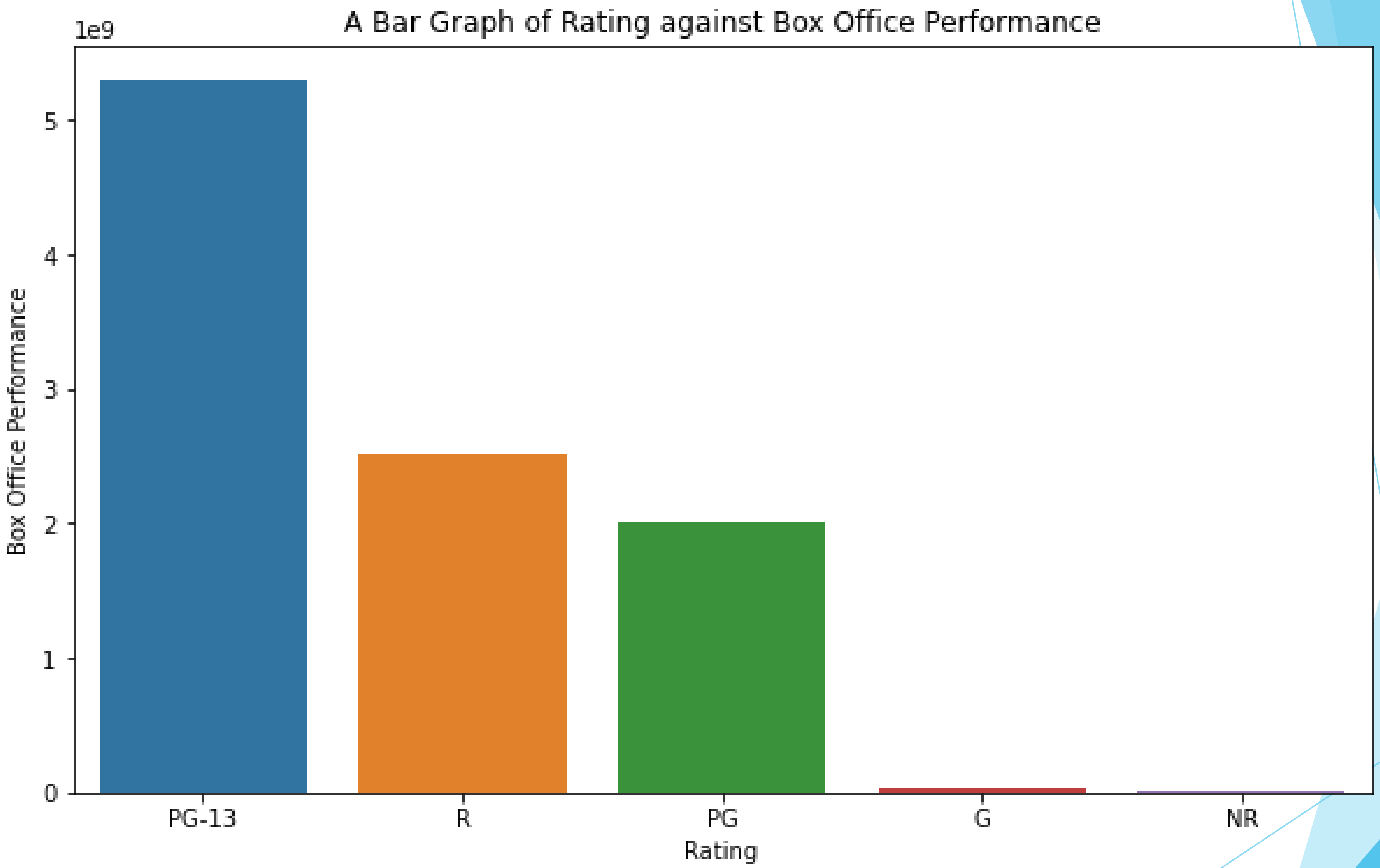

Top 10 movie studios by domestic gross
1e10
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
domestic gross per studio
BV
Uni.
WB
Fox
Sony
Par.
LGF
WB (NL)
LG/S
P/DW
Studio


Top 10 movie studios by foreign gross
1e10
2.5
2.0
1.5
1.0
0.5
0.0
foreign gross per studio
BV
Fox
WB
Uni.
Sony
Par.
WB (NL)
LGF
P/DW
LG/S
Studio

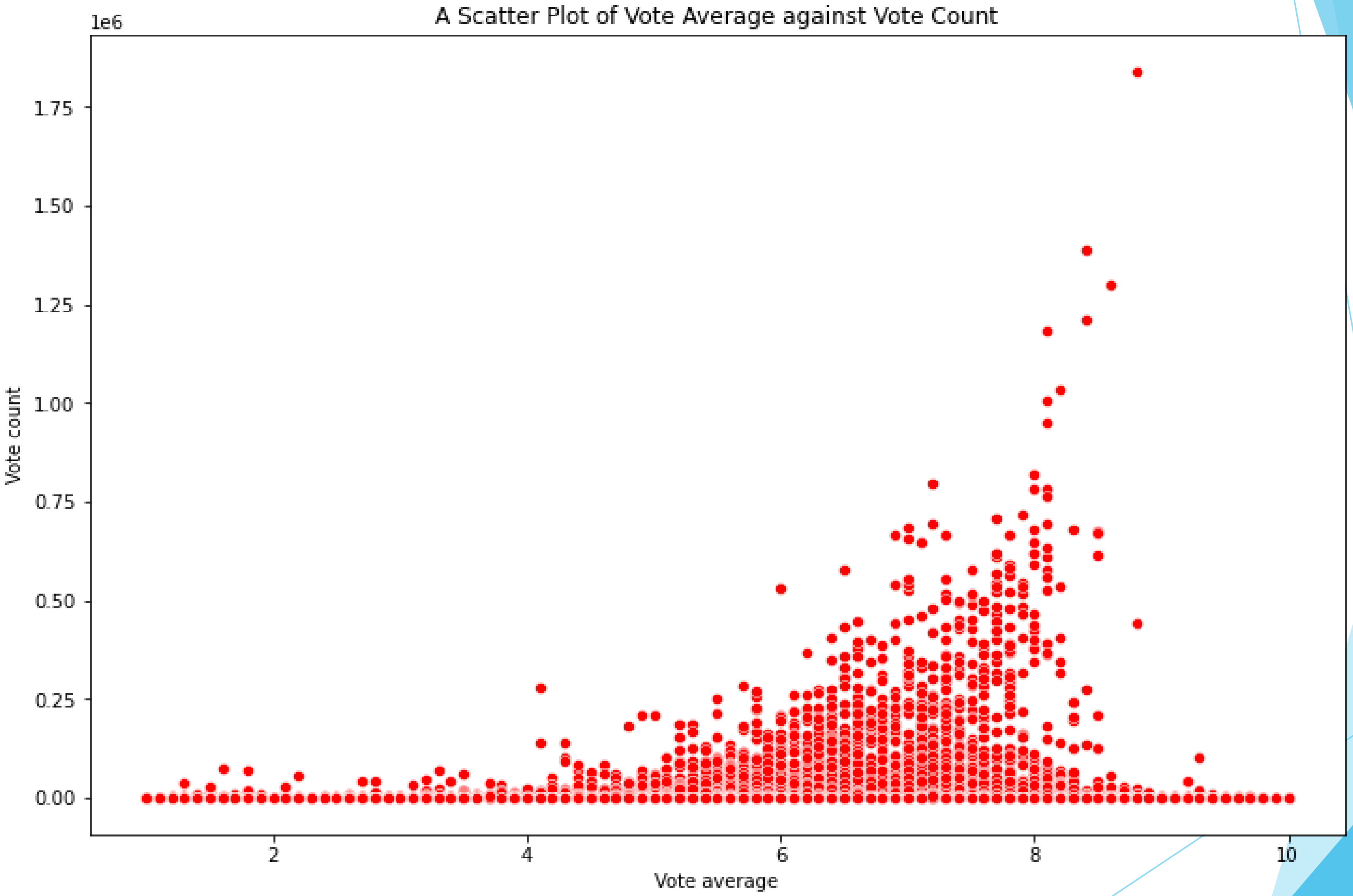
- Even though their orders look different, we saw that the top 10 studios were the same in both graphs.
- This can only mean that these studios earned the most revenue either in the United States or globally.
- This shows that the movies created by these studios are very popular in the market and are on high demand. Microsoft should therefore use them as a benchmark if they want their movie business to be successful.
- We then went ahead and analyzed the second data frame. Here, we focused on the vote average and vote count columns and created a scatter plot for representation.


A Scatter Plot of Vote Average against Vote Count
Vote count
20000
15000
10000
5000
0
Vote average
0
2
4
6
8
10

- We noticed that most of the vote counts had a vote average between 5 and 8. This showed that most of the people who voted gave from an average review to an above average review of the movies they watched.
- This means that most people were impressed by the movies they watched.
- With this we see that there is a relationship between movie review and the number of people who voted in these reviews. However, we opted to study more of this relationship in the next segment.
- We then went ahead and did analysis on the third data frame. We went with the rating and box office columns. We also plotted a bar graph for representation.


A Bar Graph of Rating against Box Office Performance
1e9
Box Office Performance
0
1
2
3
4
5
PG-13
R
PG
G
NR
Rating

- Here, we saw that the PG-13, R-rated and the PG movies had the most revenue. This means that these ratings are popular to the market.
- Therefore, movie ratings should be taken into consideration when Microsoft decides to create their movies.
- They should create a ratio of movies based on the ratings since their movies should attract the market regardless of their age.
- We then went ahead and focused on the `movie_ratings` table in the database. We used the `averagerating` and `numvotes` columns from this table for analysis. We also created a scatter plot for representation.


A Scatter Plot of Vote Average against Vote Count
1e6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Vote count
2
4
6
8
10
Vote average

- We saw that most of the vote counts had a vote average between 6 and 8. This showed that most of the people who voted gave an above average review of the movies they watched.

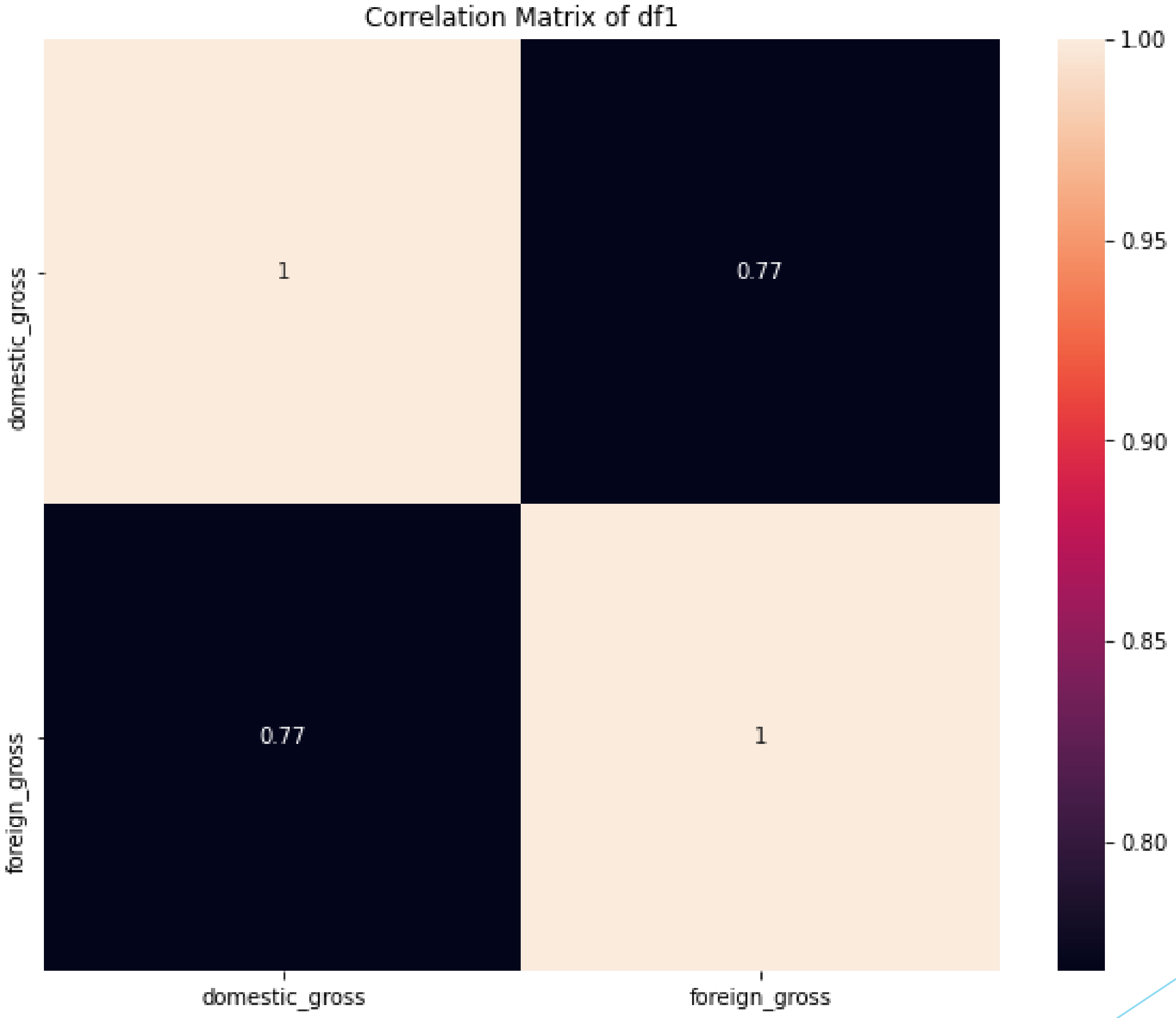
- This means that most people were impressed by the movies they watched.

- With this we see that there is a relationship between movie review and the number of people who voted in these reviews. However, we opted to study more of this relationship in the next segment.

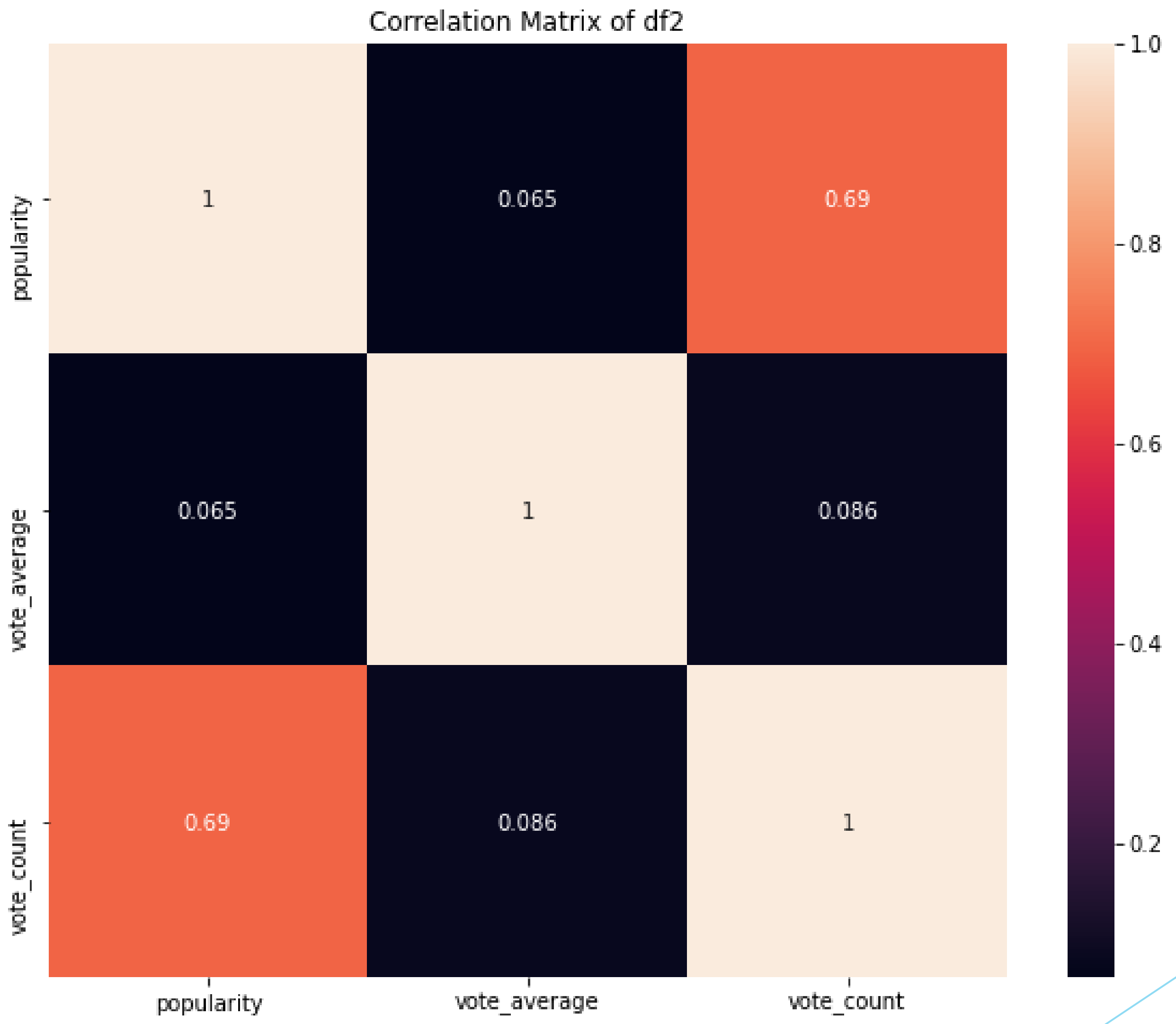
- With the bivariate analysis, we can see that studios, movie ratings, number of reviewers and review ratings are essential for movie success and these are some of the things Microsoft should be interested in before creating their movie company.

Multivariate analysis

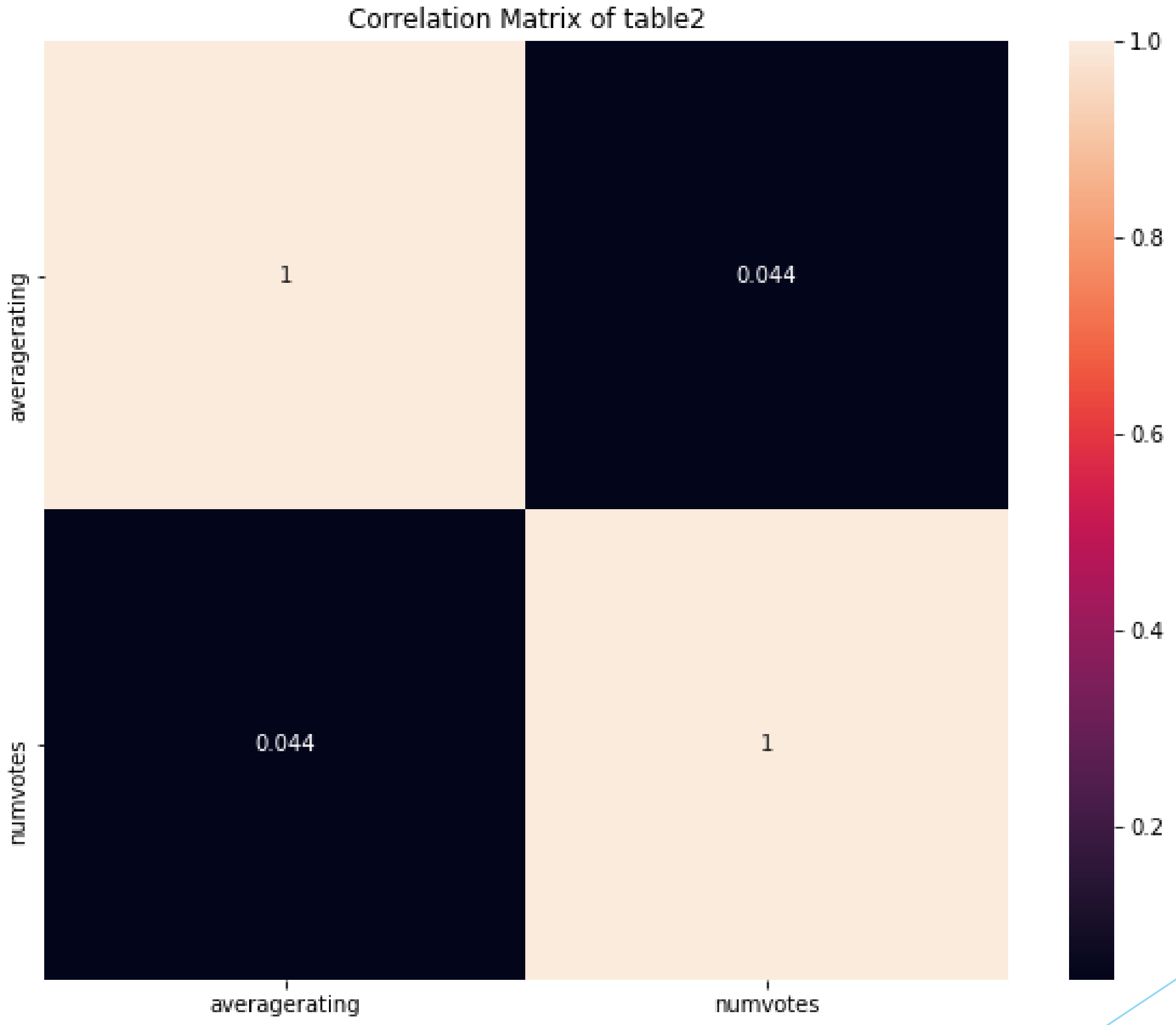
- Here, we focus on analysis of multiple columns. We created heat maps of the numerical columns to show the correlation matrices of the data. We began our analysis with the first data frame.


Correlation Matrix of df1
domestic_gross
foreign_gross
1
0.77
0.77
1
domestic_gross
foreign_gross
1.00
0.95
0.90
0.85
0.80

- Based on the correlation matrix, we saw that there was a high positive correlation between the two columns.
- This means that they have a strong linear relationship in which they tend to increase or decrease together.
- Microsoft should therefore note that the success of their movies in the United States alone would not be enough.
- They should also appeal the global market if they want to guarantee success.
- We then headed to the second data frame and plotted the heat map which showed the correlation matrix of the numerical columns.


Correlation Matrix of df2
popularity
vote_average
vote_count
1
0.065
0.69
0.065
1
0.086
0.69
0.086
1
popularity
vote_average
vote_count
1.0
0.8
0.6
0.4
0.2

- Here, we saw that the popularity and vote average columns had a low positive correlation.
- This was also seen with the vote count and vote average columns. This means that even though there may be some tendency of the variables to increase together, it is not a strong or consistent pattern.
- However, we also saw that the vote count and popularity columns had a high positive correlation.
- This means that they have a strong linear relationship in which they tend to increase or decrease together.
- For Microsoft, this means that the more the people who use their streaming service and give positive reviews to their movies, the more the popularity of the streaming service grows. Microsoft in this case should create their movies in a unique way that appeals the market.
- We then concluded the analysis by focusing on the tables in the database. In this case, we joined the two tables together and created a data frame from the tables. We then focused on the `averagerating` and `numvotes` columns for our analysis.


Correlation Matrix of table2
averagerating
numvotes
1
0.044
0.044
1
averagerating
numvotes
1.0
0.8
0.6
0.4
0.2

- We saw that `averagerating` and `numvotes` had a low positive correlation.
- This means that even though there may be some tendency of the variables to increase together, it is not a strong or consistent pattern.
- This means that even though the correlation is low, Microsoft should not ignore it.
- On the contrary, they should use it as reference when looking at the reviews of their movie company.
- The multivariate analysis has proven that there is a strong relationship between domestic gross and foreign gross.
- There is also a strong relationship between review ratings and the number of people who voted in these reviews.

Conclusion

- Based on the analysis we conducted, we can draw three conclusions.

1. The movie industry is a very profitable business
2. Some of the factors that may bring success in the movie industry are movie studios, movie ratings, number of reviewers and review ratings.
3. There is a strong relationship between domestic gross and foreign gross. There is also a strong relationship between review ratings and the number of people who voted in these reviews.