Phase one project analysis

EDGAR KIPRONO

DS-FT-07

MICROSOFT MOVIE ANALYSIS

INTRODUCTION-OVERVIEW

- This project is designed to address the challenge of Microsoft's entry into the dynamic and competitive movie industry.
- As technology and entertainment converge, Microsoft recognizes the strategic importance of establishing a foothold in the cinematic landscape. The real-world problem at hand involves navigating the complexities of the movie industry, understanding audience preferences, and optimizing content creation and distribution.
- Stakeholders in this endeavor include Microsoft executives, content creators, distributors, and end-users.
- By leveraging data analytics, market insights, and cutting-edge technology, this project aims to empower Microsoft to make informed decisions, tailor content to audience demands, and strategically position itself in the movie industry.

BUSINESS UNDERSTANDING

- This project aims to generate insights and recommendations to the Microsoft company though Exploratory data analysis(EDA) in the current film industry giving insights on venturing into this industry.

- This project tends to analyze the trends and patterns associated with the current success rate of films and movies in box office examining factors such as budget(cost) of movie production, top-rated movies and more.

BUSINESS UNDERSTANDING

This project tends toward the conversion of practical recommendations and insights to make more realistic and informed decision about the type of film to venture into, if it worth it, and to the right target audience. When implemented, this will be of great importance to the current stakeholders as they get to know what they are in getting to and perhaps make informed decisions like,budget allocation or even risks that might accrue

DATA UNDERSTANDING

- This data analysis of this project envelops diverse data from box office which draws its data from different movie sites such as IMDb, rotten tomatoes, the movie Db the numbers and many more.
- The dataset includes but to mention a few genre budget, movie gross, release dates, reviews, original title and popularity. These datasets will be crucial in identifying trends common in most successful movies including what the target audience preferences.

DATA UNDERSTANDING

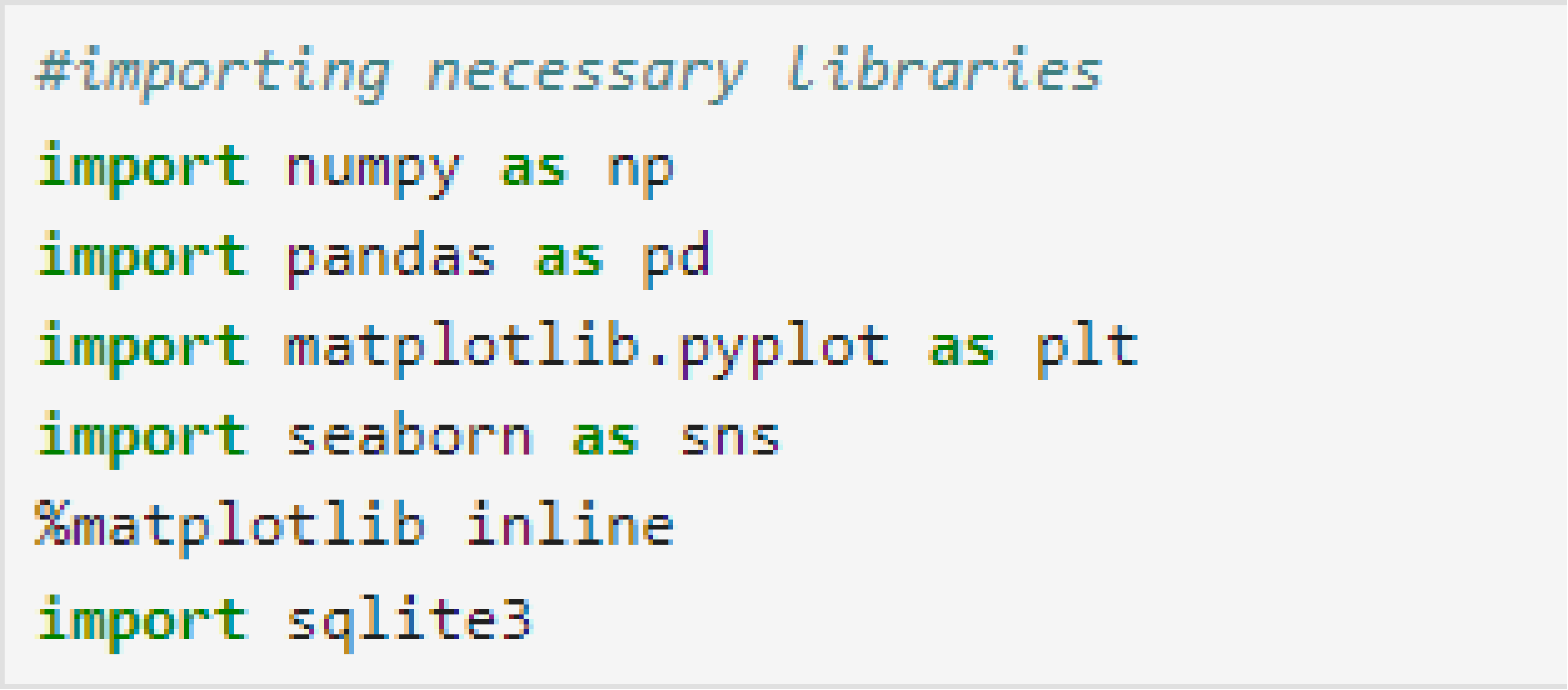
- The budget and gross is important because they measure the success rate in terms of finances and investments. Inclusion of original movie language may offer a good background and insight into regional preferences and success in the international market.
- Last but not least review will be crucial in determining the success rate of a movie in the long run. Descriptive statistics reveals the range and distribution of certain key features such as the budget cost and returns per region

DATA UNDERSTANDING

- The budget and gross will also be crucial and will be a determinant factor in realizing the success of this venture.
- However, this project also takes into inclusion limitations that come with venturing into this industry.
- This might include quality and incompleteness of the data being received from box office which does not predict future trends of the industry.

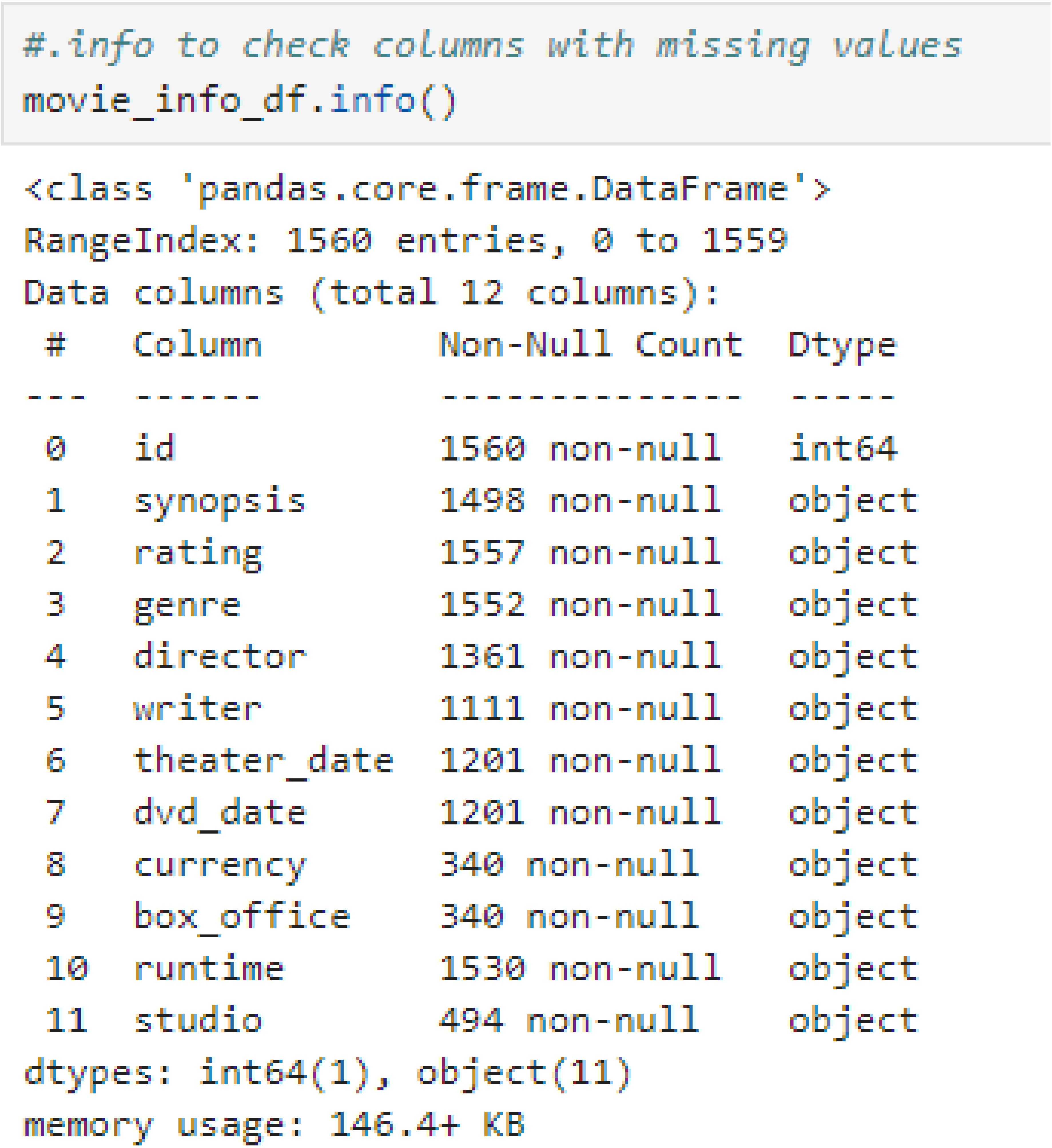
DATA UNDERSTANDING

- **Data from box office might be current and to make more informed appropriate decisions it requires one to have historical data to make more future decisions.**
- **Data privacy concerns might prove to be a limitation since peoples are more concerned with the ever changing technologies and might be worried about how their data is being utilized since we are using some of their reviews and ratings.**
- **The company should also guarantee compliance with regulations concerning controlling and gathering of information**



```
#importing necessary libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import sqlite3
```

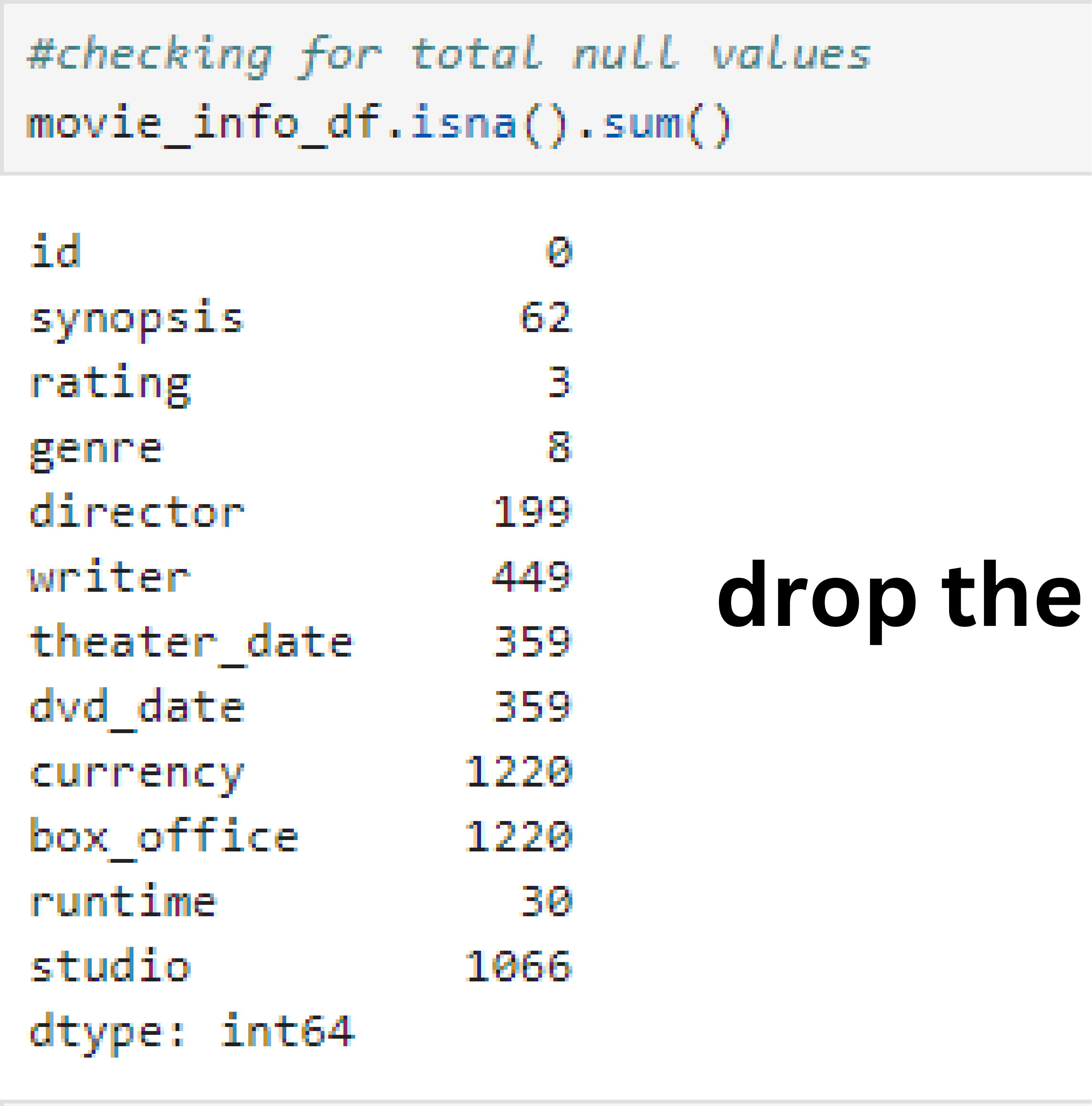
include necessary libraries use to analyze and visualize the data.Libraries such as pandas, numpy,matplotlib and seaborn. For the IMDb database we connected to the database through sqlite3.



```python
#.info to check columns with missing values
movie_info_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1560 entries, 0 to 1559
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   id            1560 non-null   int64 
 1   synopsis      1498 non-null   object
 2   rating        1557 non-null   object
 3   genre         1552 non-null   object
 4   director      1361 non-null   object
 5   writer        1111 non-null   object
 6   theater_date  1201 non-null   object
 7   dvd_date      1201 non-null   object
 8   currency      340 non-null    object
 9   box_office    340 non-null    object
 10  runtime       1530 non-null   object
 11  studio        494 non-null    object
dtypes: int64(1), object(11)
memory usage: 146.4+ KB
```

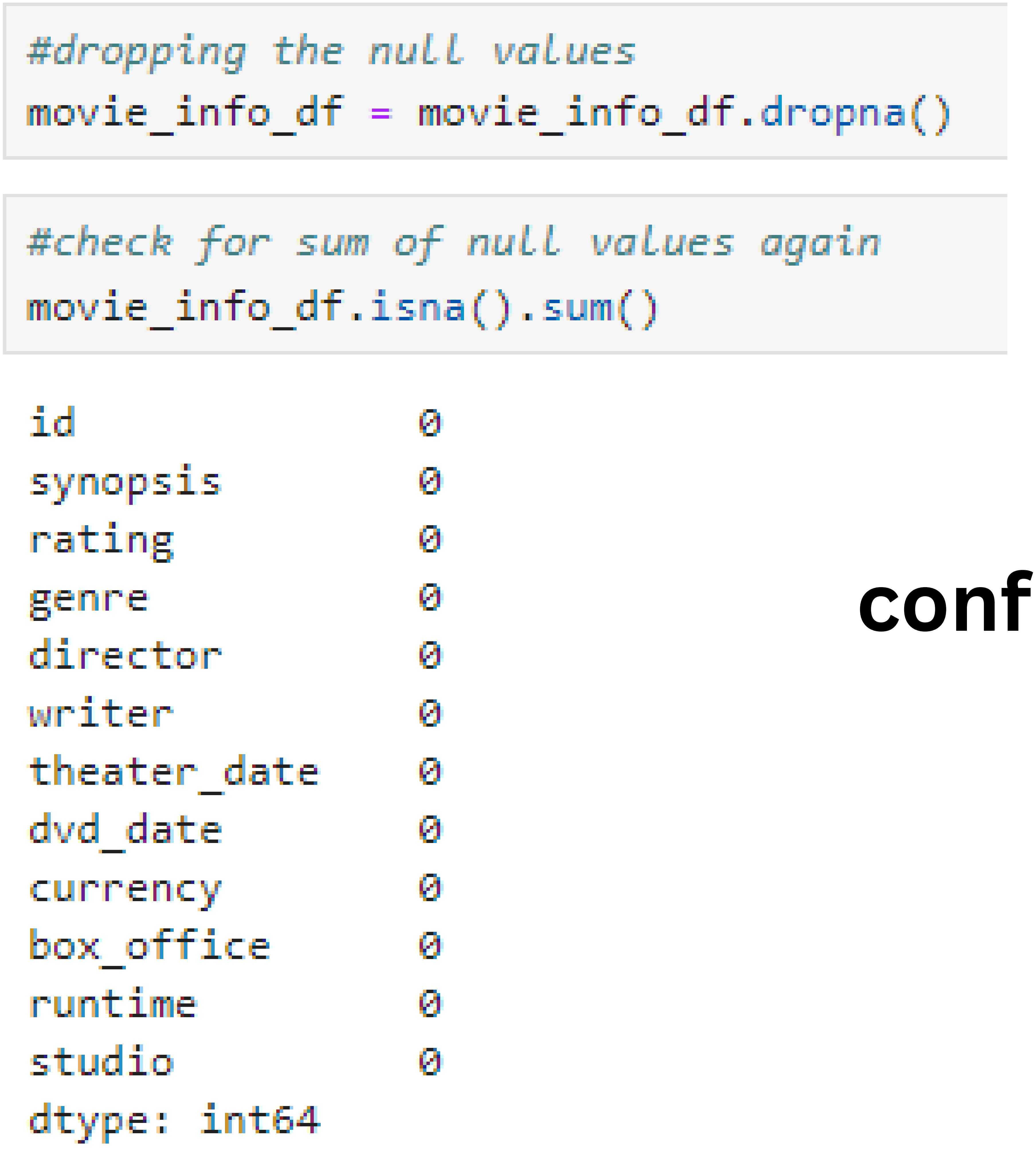
check for the completeness of data in the columns before analysis of the data begins.



```python
#checking for total null values
movie_info_df.isna().sum()
```

```
id                 0
synopsis          62
rating             3
genre              8
director         199
writer           449
theater_date     359
dvd_date         359
currency        1220
box_office      1220
runtime           30
studio          1066
dtype: int64
```

drop the null values from the rows



```
#dropping the null values
movie_info_df = movie_info_df.dropna()
```

```
#check for sum of null values again
movie_info_df.isna().sum()
```

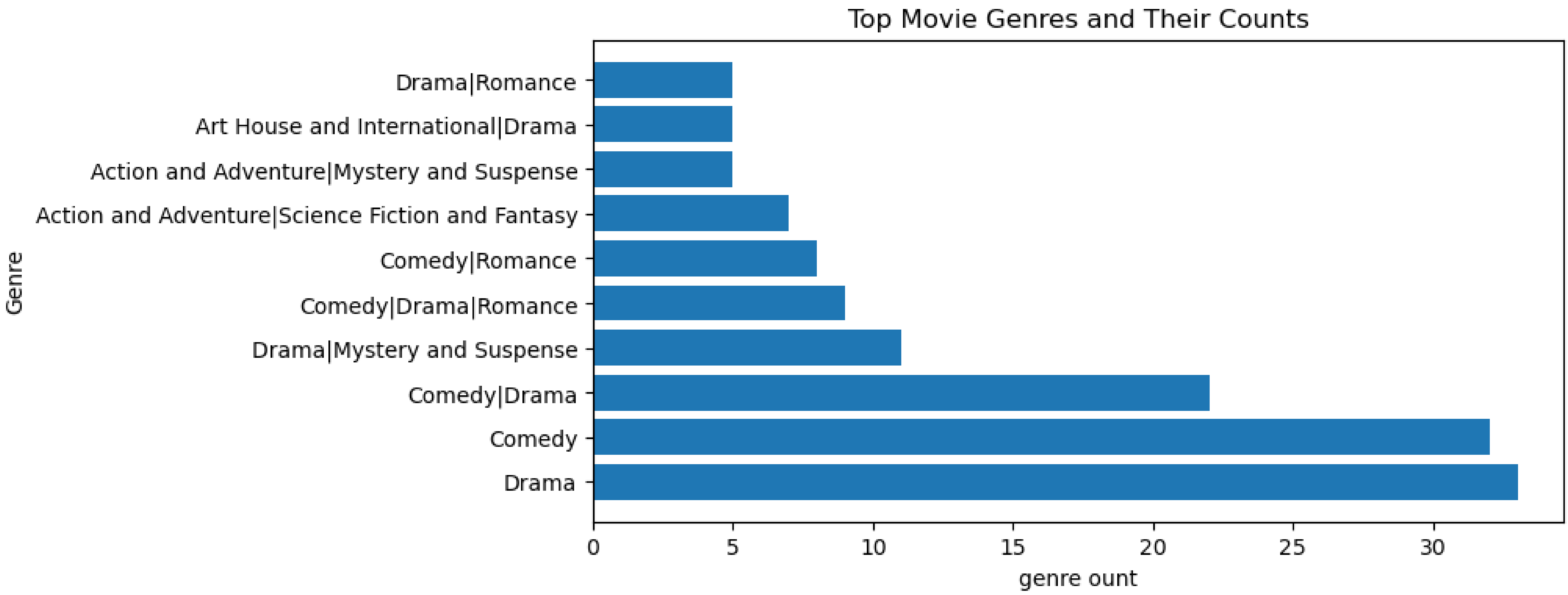
```
id              0
synopsis        0
rating          0
genre           0
director        0
writer          0
theater_date    0
dvd_date        0
currency        0
box_office      0
runtime         0
studio          0
dtype: int64
```

confirming it there is still any null values

DATA ANALYSIS

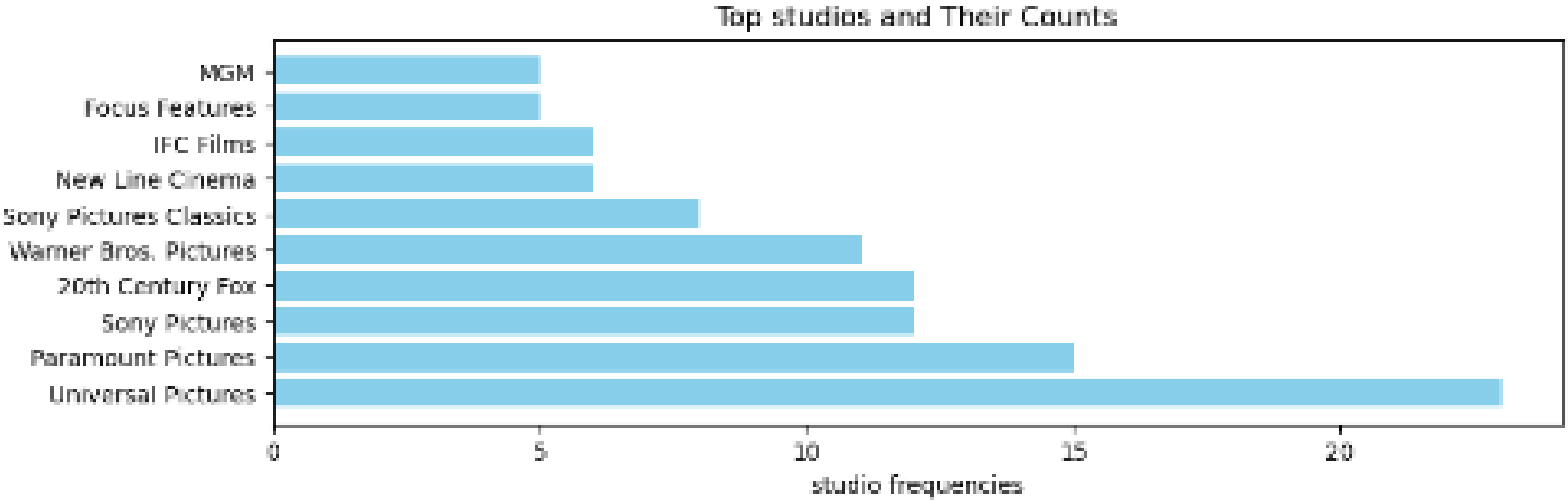
&

VISUALIZATION



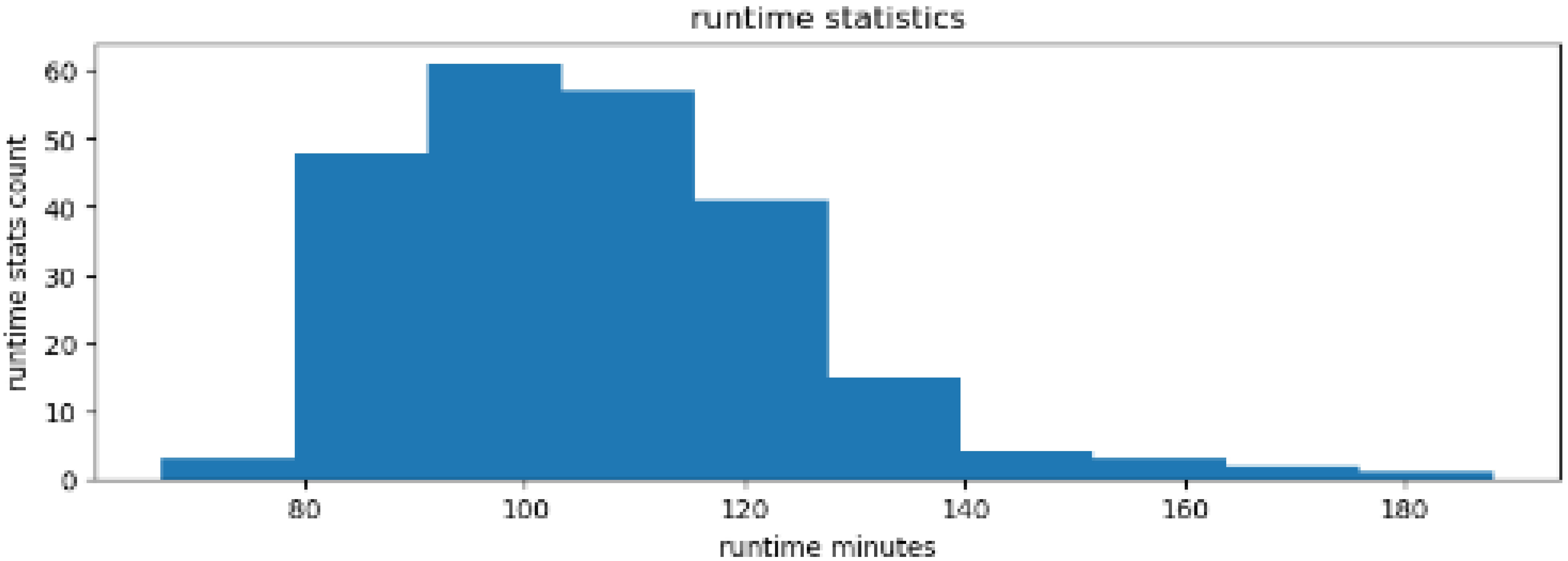
This graph shows the top movie genres and their counts

This project suggests that microsoft should be more invested in the top 5 genres



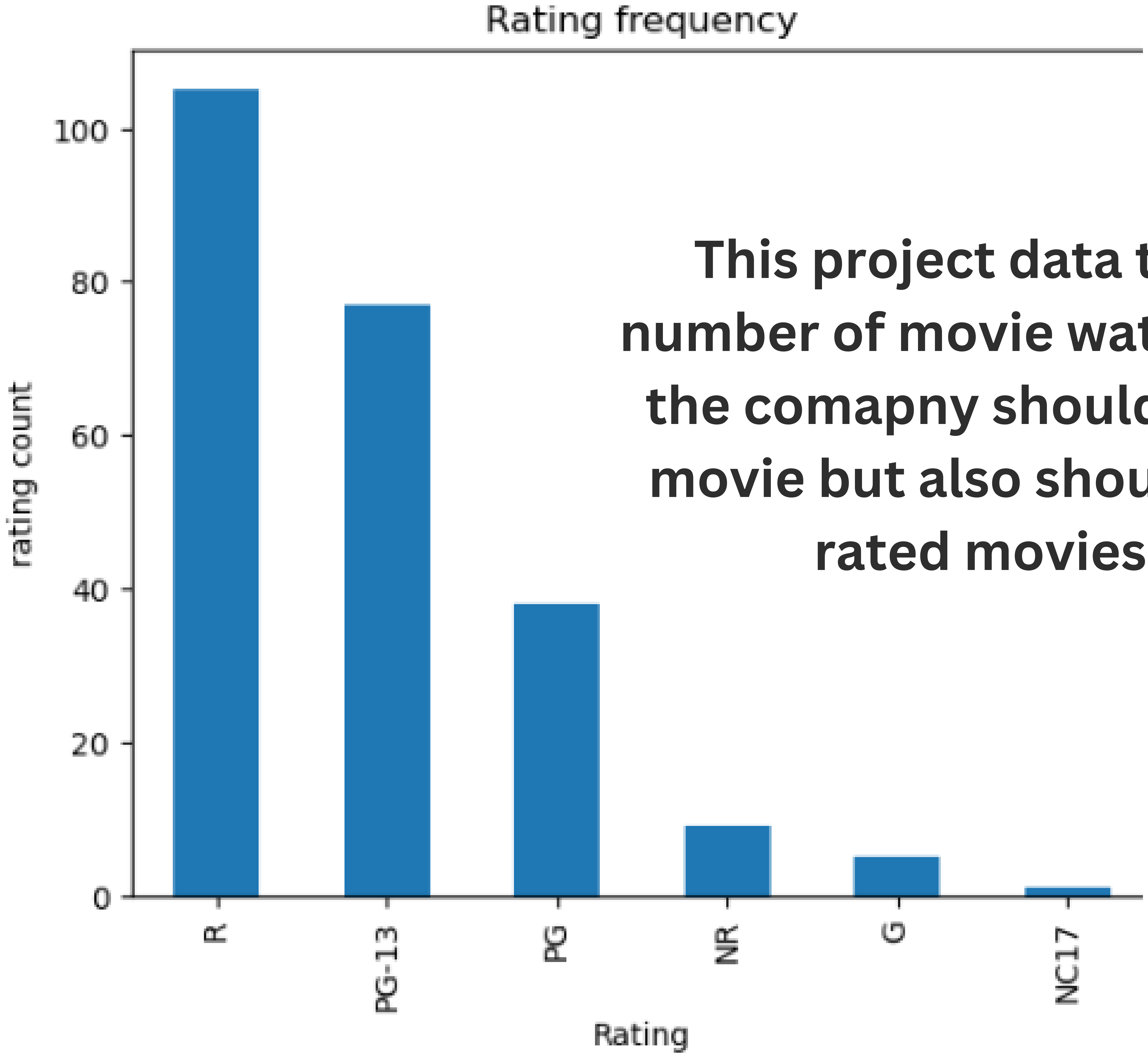
This graph shows the top studios and their counts

This project suggests that microsoft would have more competition from these studios and should try and find out success factors of these studios

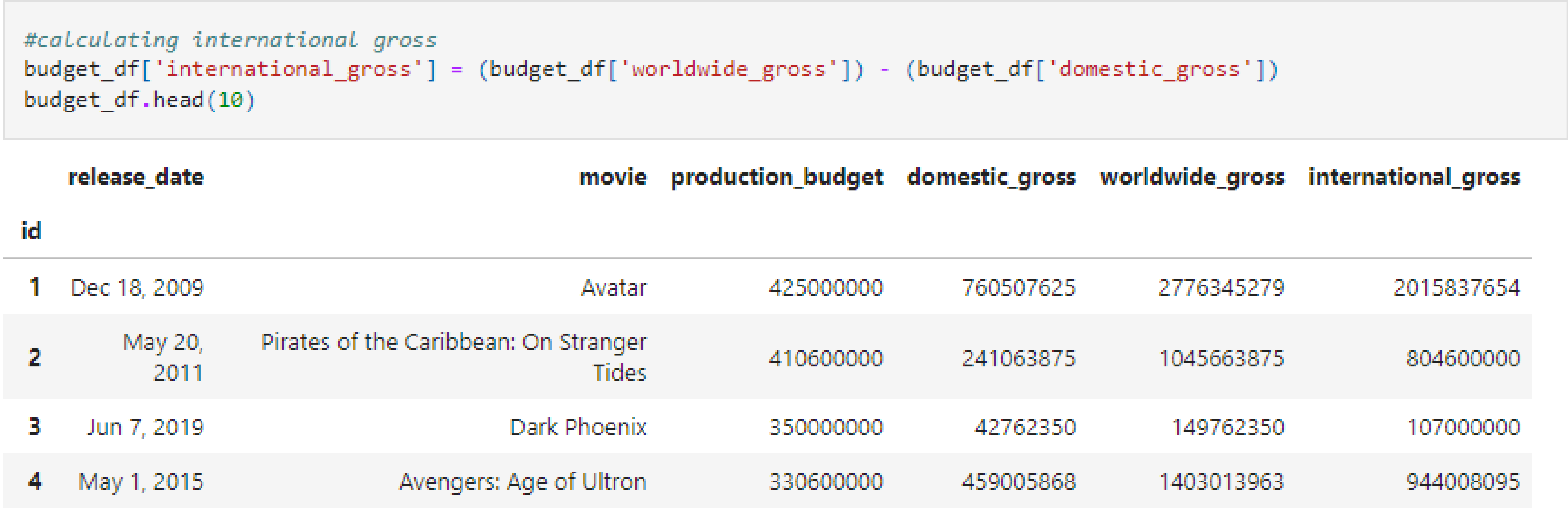


This dataframe shows the runtime stats like the mean median and standard deviation. The maximum runtime minutes is 188 minutes while the minimum runtime is 67 minutes, The average runtime minutes is 106 minutes

- **25th Percentile (Q1): 93 minutes**
- **This means that 25% of the movies in the dataset have a runtime of 93 minutes or les s. 50th Percentile (Q2 or Median): 105 minues**
- **This is the middle value of the dataset when it is ordered. It separates the lower 50% of runtimes from the upper This means 50% of the movies have a runtime of 105 minutes or less, and the other 50% have a runtime of 105 minutes or more. 75th Percentile (Q3): 117 minutes**
- **This means that 75% of the movies in the dataset have a runtime of 117 minutes**
- **Judging by this data, movies produced by microsoft should range between 90 mins and 120 mins or less.**



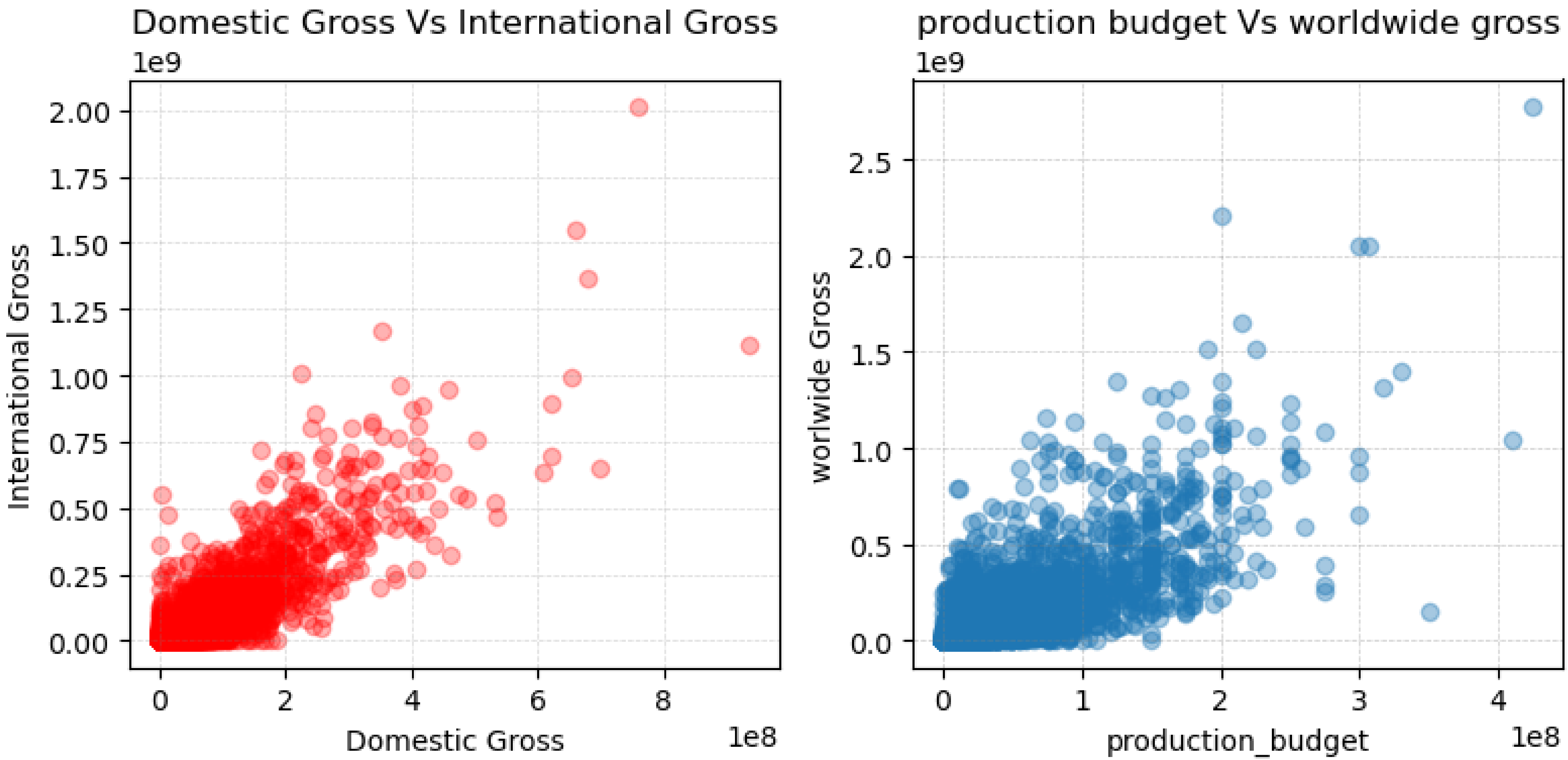
This project data then suggests that a good number of movie watchers are mostly audlts and the comapny should invest more on Restricted movie but also shoul be able to invest in PG-13 rated movies to suit all audiences



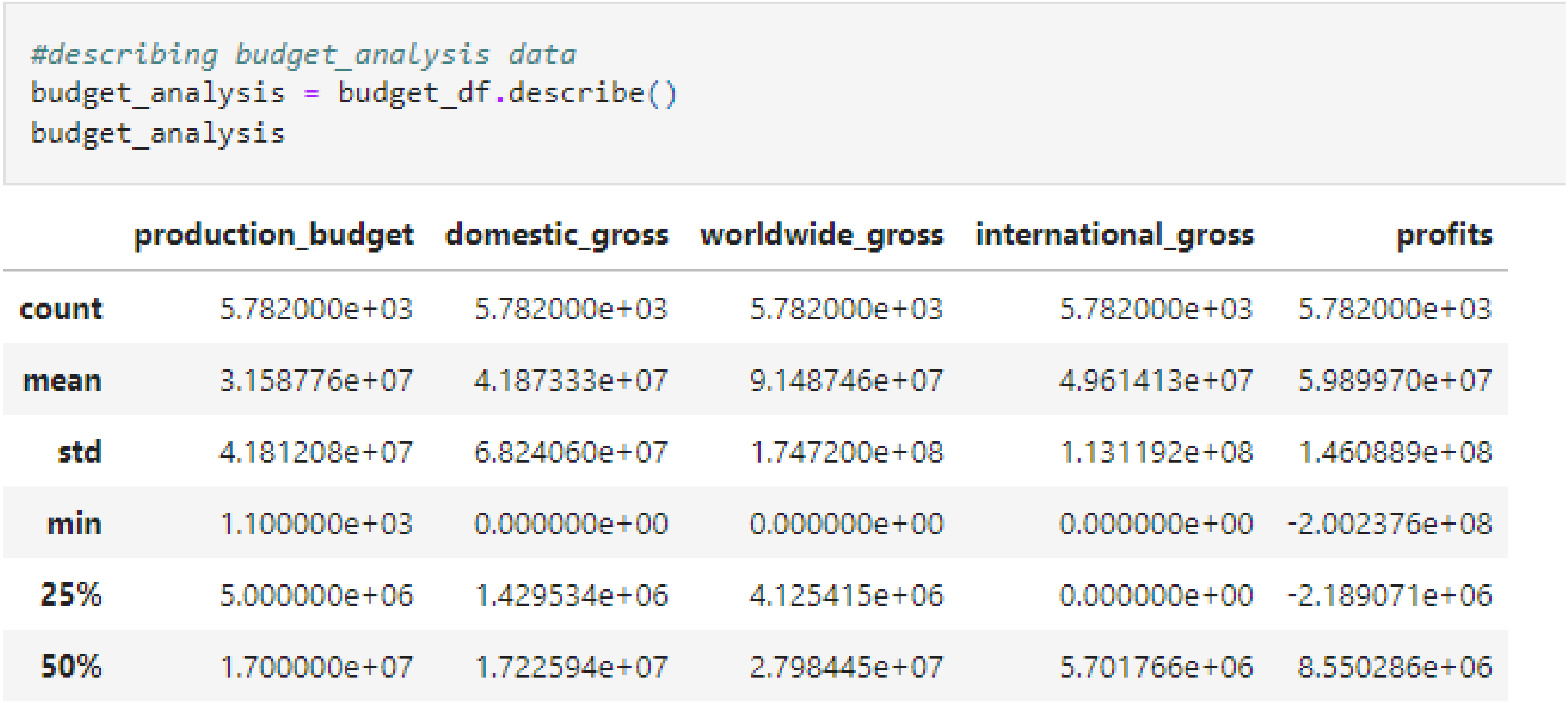
```
#calculating international gross
budget_df['international_gross'] = (budget_df['worldwide_gross']) - (budget_df['domestic_gross'])
budget_df.head(10)
```

	release_date	movie	production_budget	domestic_gross	worldwide_gross	international_gross
id						
1	Dec 18, 2009	Avatar	425000000	760507625	2776345279	2015837654
2	May 20, 2011	Pirates of the Caribbean: On Stranger Tides	410600000	241063875	1045663875	804600000
3	Jun 7, 2019	Dark Phoenix	350000000	42762350	149762350	107000000
4	May 1, 2015	Avengers: Age of Ultron	330600000	459005868	1403013963	944008095

First we get to find the international gross which is worldwide gross – domestic gross to know how much returns to movies get from the international audience.



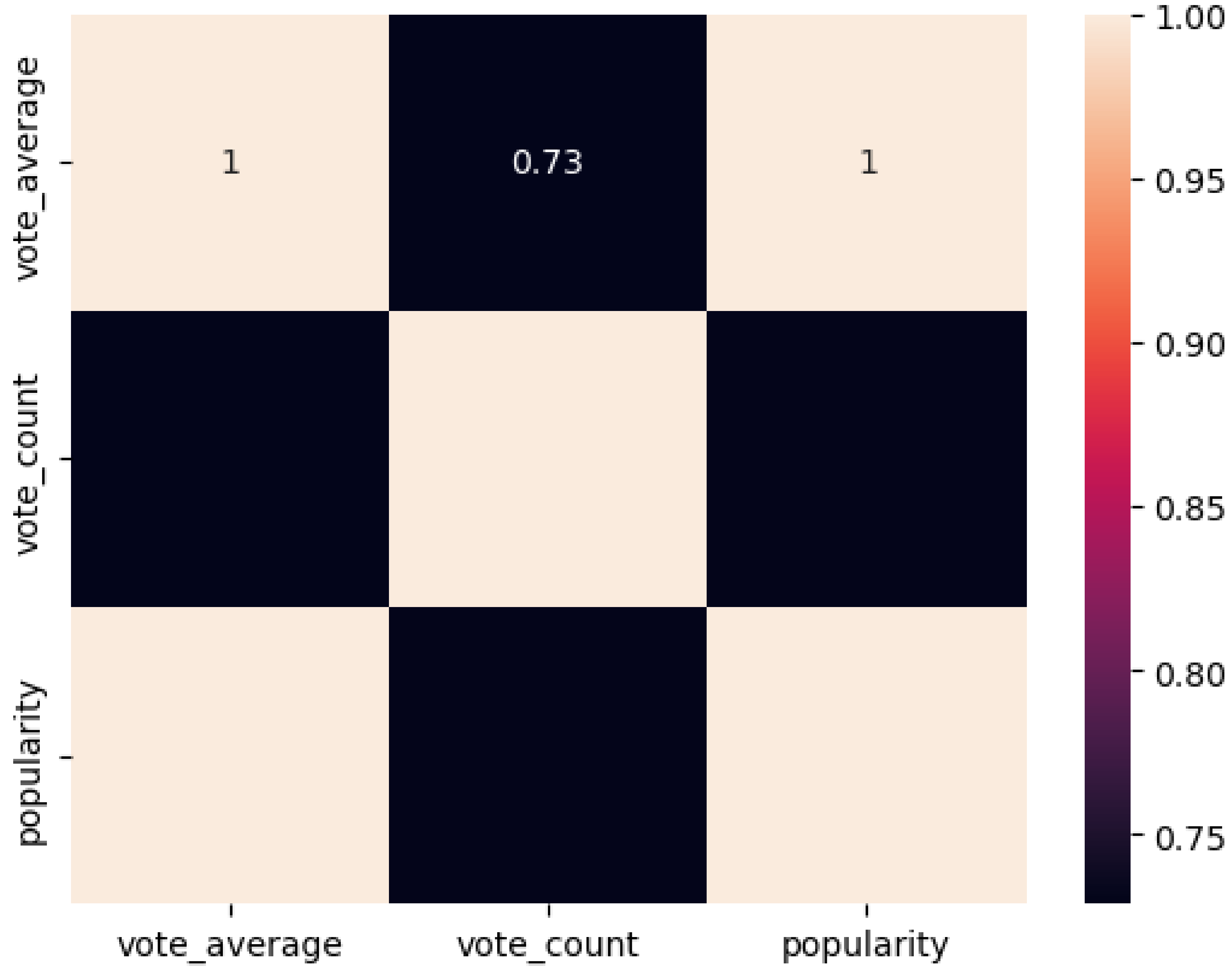
scatter graph shows the relation between domestic gross vs international gross and production budget against worldwide gross. This plot shows that an increase in domestic gross will influence an increase in international gross and production budget influences worlwide gross



```
#describing budget_analysis data
budget_analysis = budget_df.describe()
budget_analysis
```

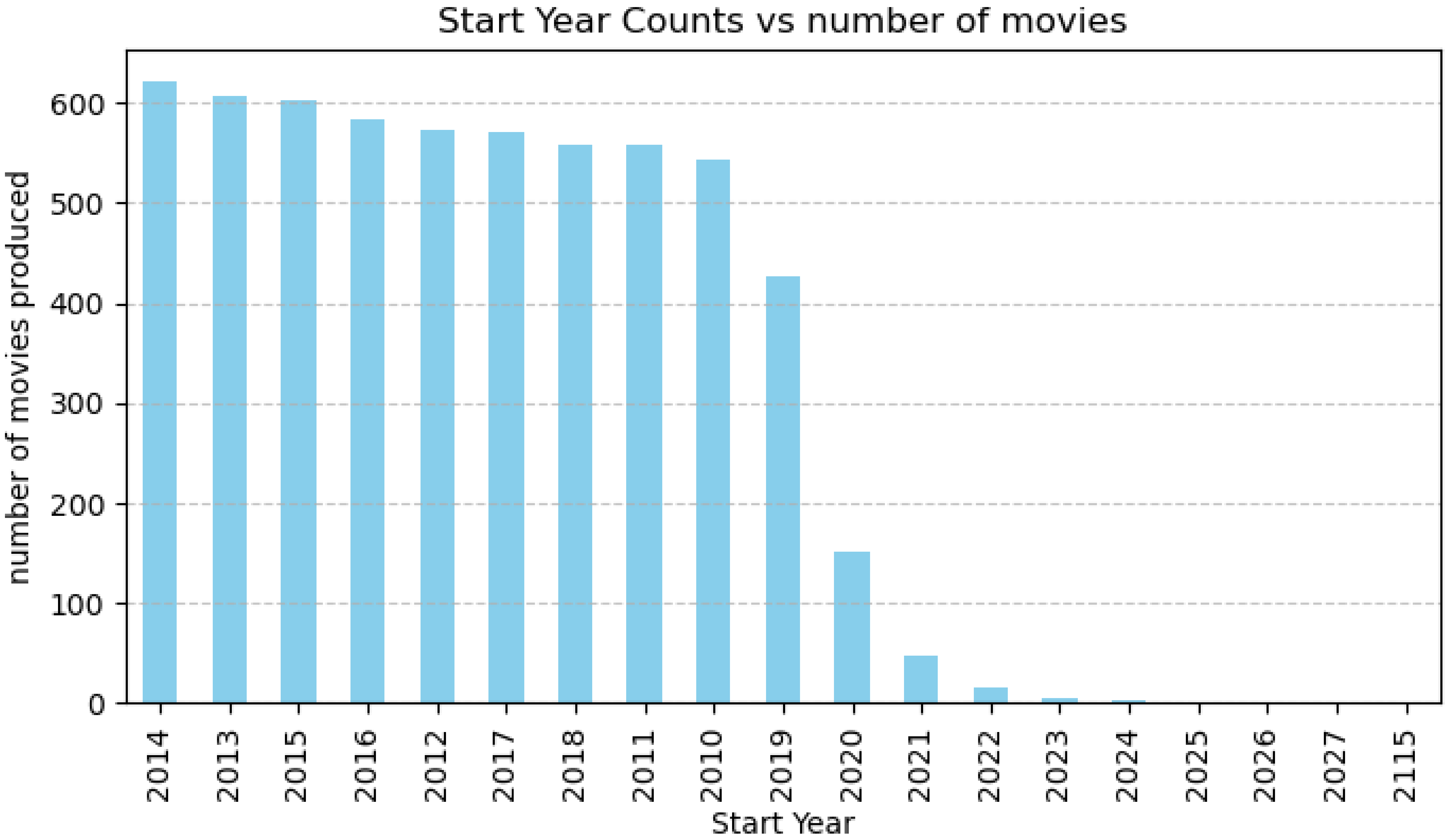
	production_budget	domestic_gross	worldwide_gross	international_gross	profits
count	5.782000e+03	5.782000e+03	5.782000e+03	5.782000e+03	5.782000e+03
mean	3.158776e+07	4.187333e+07	9.148746e+07	4.961413e+07	5.989970e+07
std	4.181208e+07	6.824060e+07	1.747200e+08	1.131192e+08	1.460889e+08
min	1.100000e+03	0.000000e+00	0.000000e+00	0.000000e+00	-2.002376e+08
25%	5.000000e+06	1.429534e+06	4.125415e+06	0.000000e+00	-2.189071e+06
50%	1.700000e+07	1.722594e+07	2.798445e+07	5.701766e+06	8.550286e+06

From the table we can see that the estimated production for starting a movie would be figures close to $31,587,760 and profits would be close to $59,899,700 but does not those would be the actual profit figures but most movies make profits of up to $8,550,286 and more



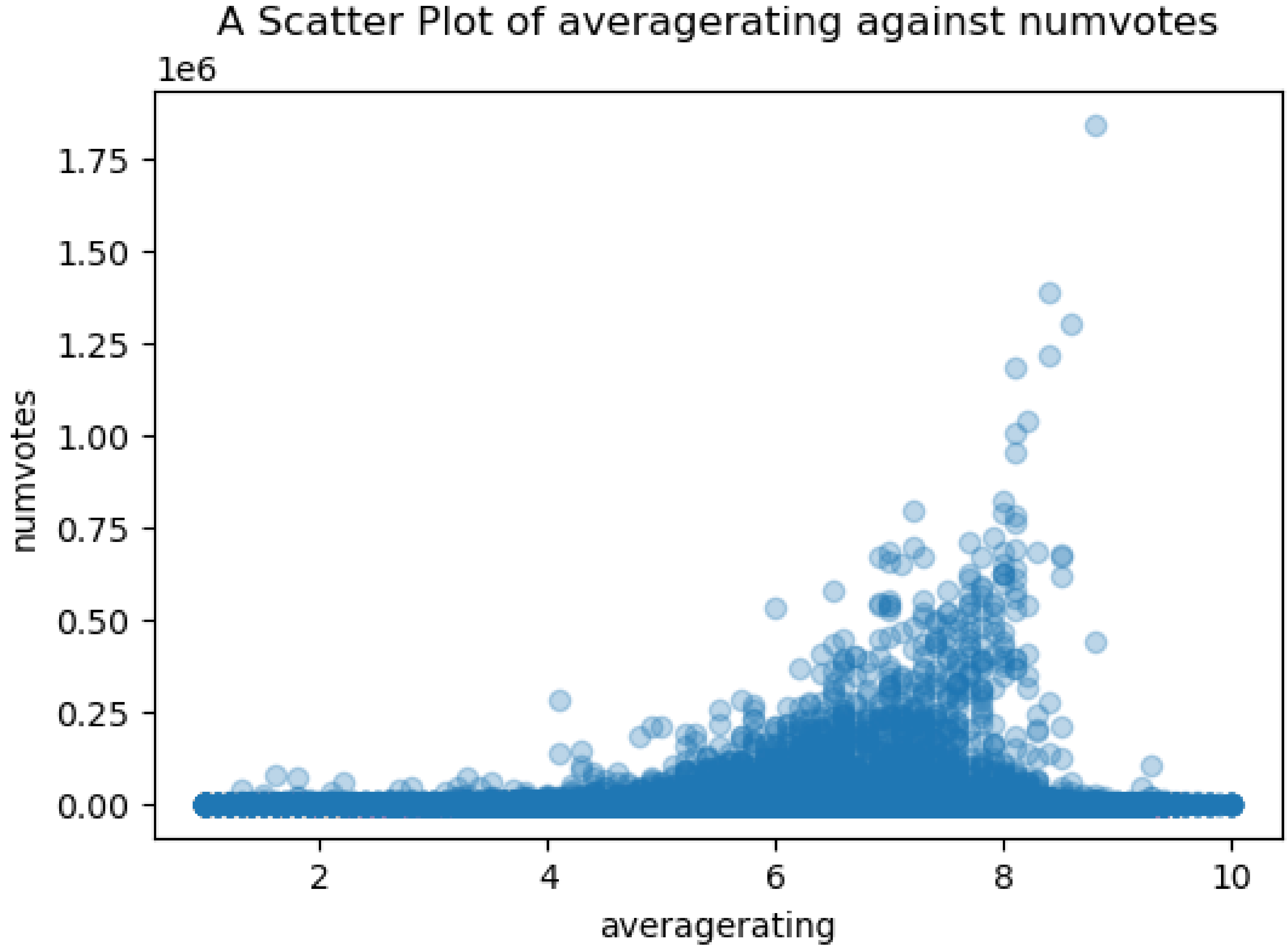
In terms of correlation of popularity vs vote average with a correlation coefficient of 1 indicates a perfect positive correlation. This means that as popularity increases, vote average in movies also increases.

It indicates a strong positive relationship between popularity and vote average. Vote count vs vote average shows a correlation coefficient of 0.73 which indicates a strong positive correlation and suggests that as the vote count increases the vote average tends to increase but it is not a perfect correlation



The graph above shows start year vs the number of movies that started on that year

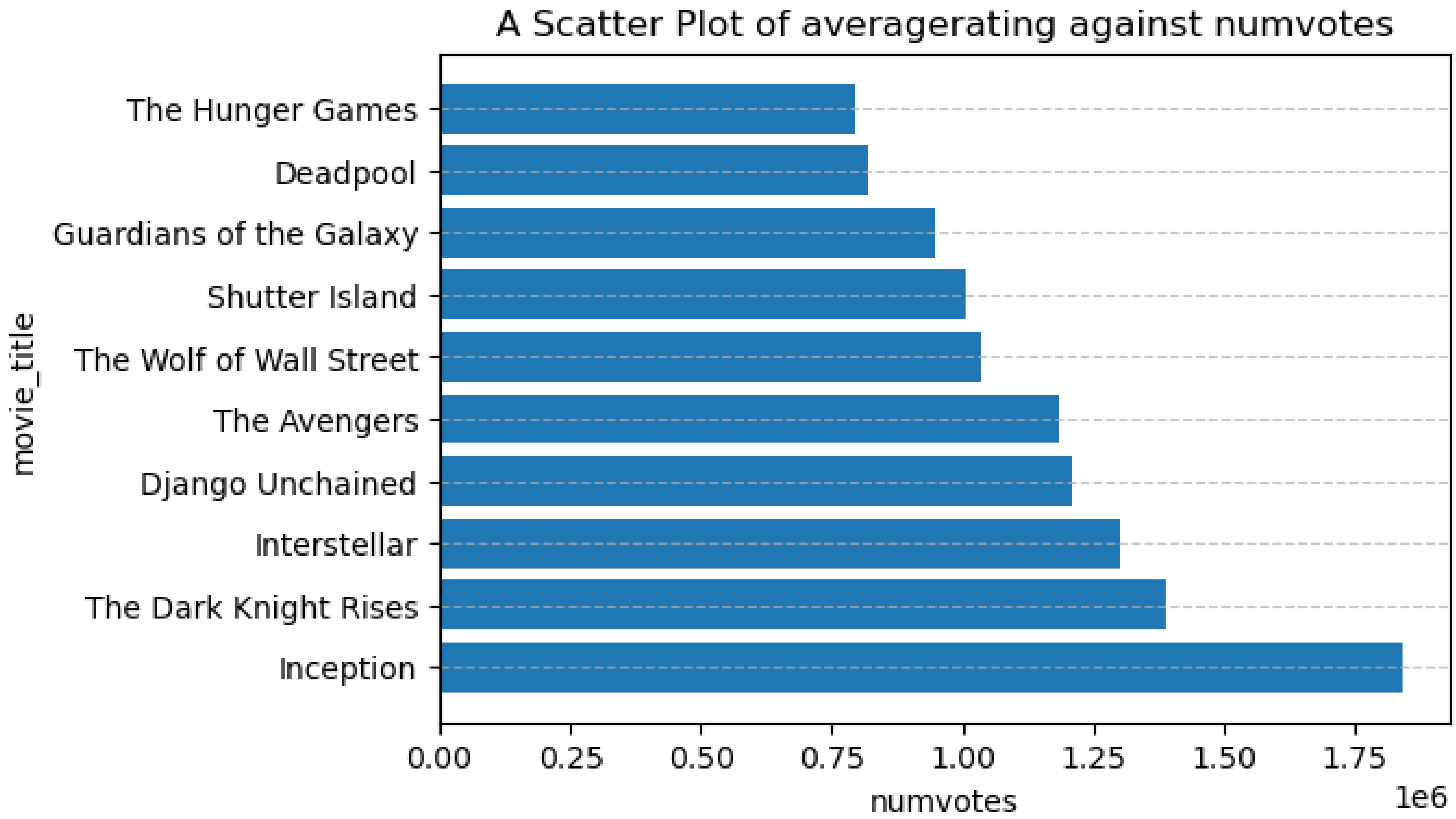
From the database the project mainly used 3 tables, movies basics movie akas and movie ratings. The first table we used was movie basics to know what number of movies were produced in a certain year. From the result we notice that 2014 had the highest number of movies produced followed by 2013. The project results displayed then shows a gradual decrease of movies produced with 2023 having the least from the data we had.



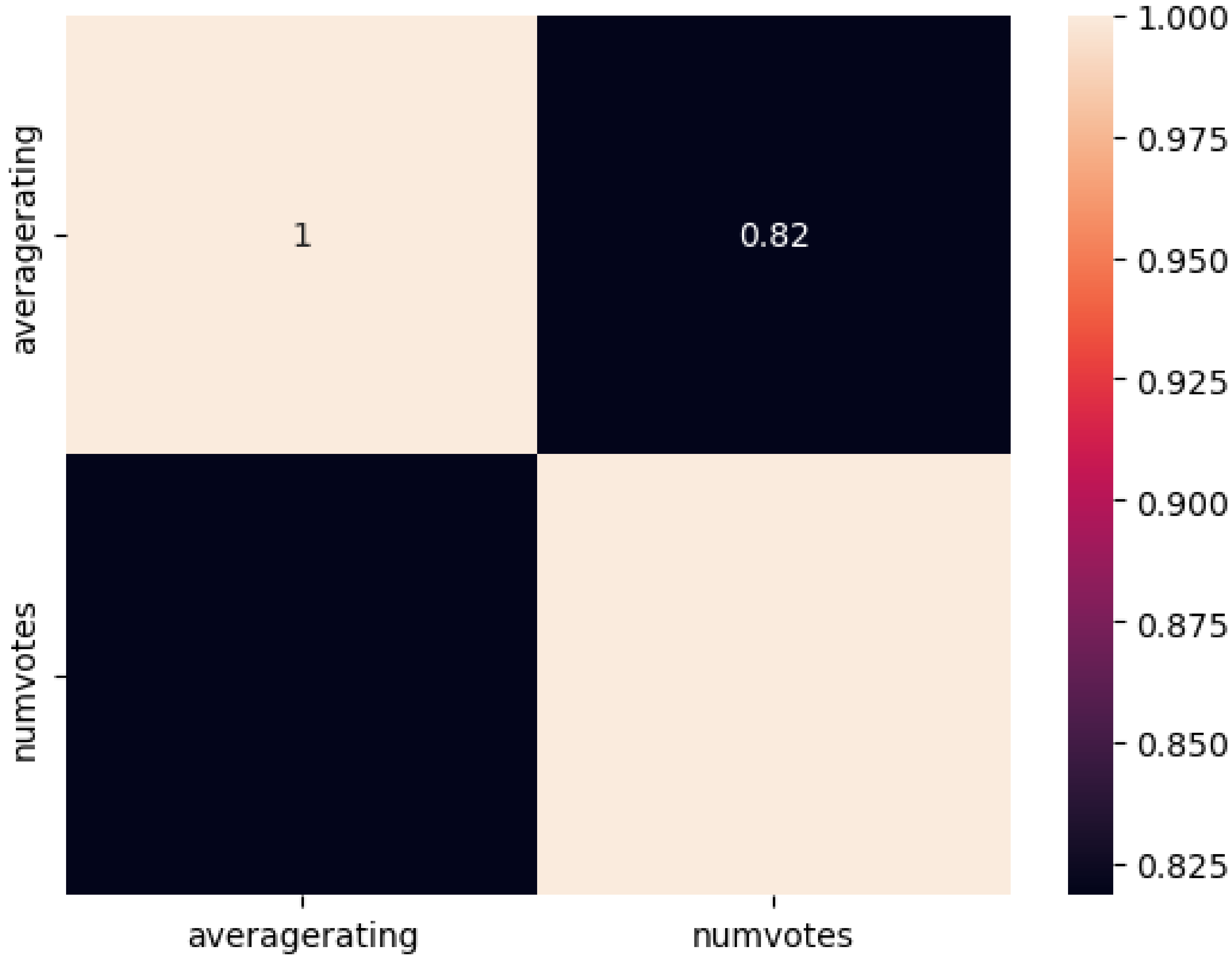
average rating and number of votes

From the movie ratings table the average rating and numvotes columns displayed in the scatter plot, we can see that alot of movies are mostly between 6 and 8 with a large number of movies being rated 7

Inception movie was top movie with a high number of counts and an average rating of 8.8 while the dark knight rises had an average rating vote of 8.4. Microsoft should analyze these type of movie content to know what type of content fascinates the audience
From the above correlation map shows the relationship between average rating and number of vote counts



this shows that as the num of vote counts increase it might influence the rating since inception had a rating of 8.8



From the graph number of votes against average rating the correlation coefficient was 0.82. This is a positive correlation coefficient suggests that as average rating increases there is a tendency for number of votes to increase.

This suggests a fairly consistent relationship between the two variables. When you observe a movie with higher rating there is a likelihood that it has received a higher number of votes

- . From the value counts conducted the top most writer was Woody Allen with a total of 4 movies written followed by John Hughes with 3 and Jim Jarsmuch with 3 too . This project then goes ahead to suggest the three writers as a good potential start to write movies for Microsoft.
- The project also suggest director Spielberg should be included in the list if the company was to outsource movie directors for the first project
- In the reviews file the project tries to find what ratings do most movies to be able to know the average rating Microsoft movies get. Most movies had a rating of 2/4 and ¾ this means Microsoft should anticipate ratings between 2 and 3 if they produce average movies. In terms of publishers, Reelings reviews was top followed by Patrick Nabbaro

Conclusion

- From this analysis the project concludes that the company should not limit the production budget if the company want to register higher profits. Production budget influences the profits made from movies
- Marketing of movies should be done worldwide more since a higher number of votes by the audience means the higher rating if the quality of the movie is high
- The company should try and collaborate with top writers and directors because of their experience in the movie industry
- More datasets should be analyzed before Microsoft joins the movie industry to ensure a good work is done on the movies.

THANK YOU

Name: Edgar Kiprono

Github:https://github.com/ed-gar-k

LinkedIn:www.linkedin.com/in/edgar-kiprono