PHASE 1 PROJECT

LESLEY KAMAMO

OUTLINE

OVERVIEW

This highlights the stakeholders need in the project

BUSINESS PROBLEM

This highlights the stakeholders need in the project

DATASET

To show the datasets to be used in the analysis

METHODS

This is to show the procedures and models created from the datasets

RESULTS

To show the findings and insights based on the models created

CONCLUSION

To show the recommendations and summary of the analysis

OVERVIEW

This project requires the use Explanatory Data Analysis to generate insights

for Microsoft as our business stakeholder.

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. They require our insight to help on the same.

BUSINESS PROBLEM

This business has decided to create a New Movie Studio. To do this, they need insight into the following:

1. What genres of films are frequently viewed

2. What genres of films are produced the most
3. What genres of films tend to have higher rating

4. What films produced have higher profit return value

DATA

For this project, the dataset to be used is contained in three separate CSV files, as below:

1. `bom.movie_gross.csv`: each record represents movie title, with attributes of that movie (e.g. year). The domestic_gross and foreign_gross are represented and each movie listed has a corresponding studio of creation.

2. `title.ratings.csv`: each record represents a movie title, then the Genre and Start year columns with values on the same.

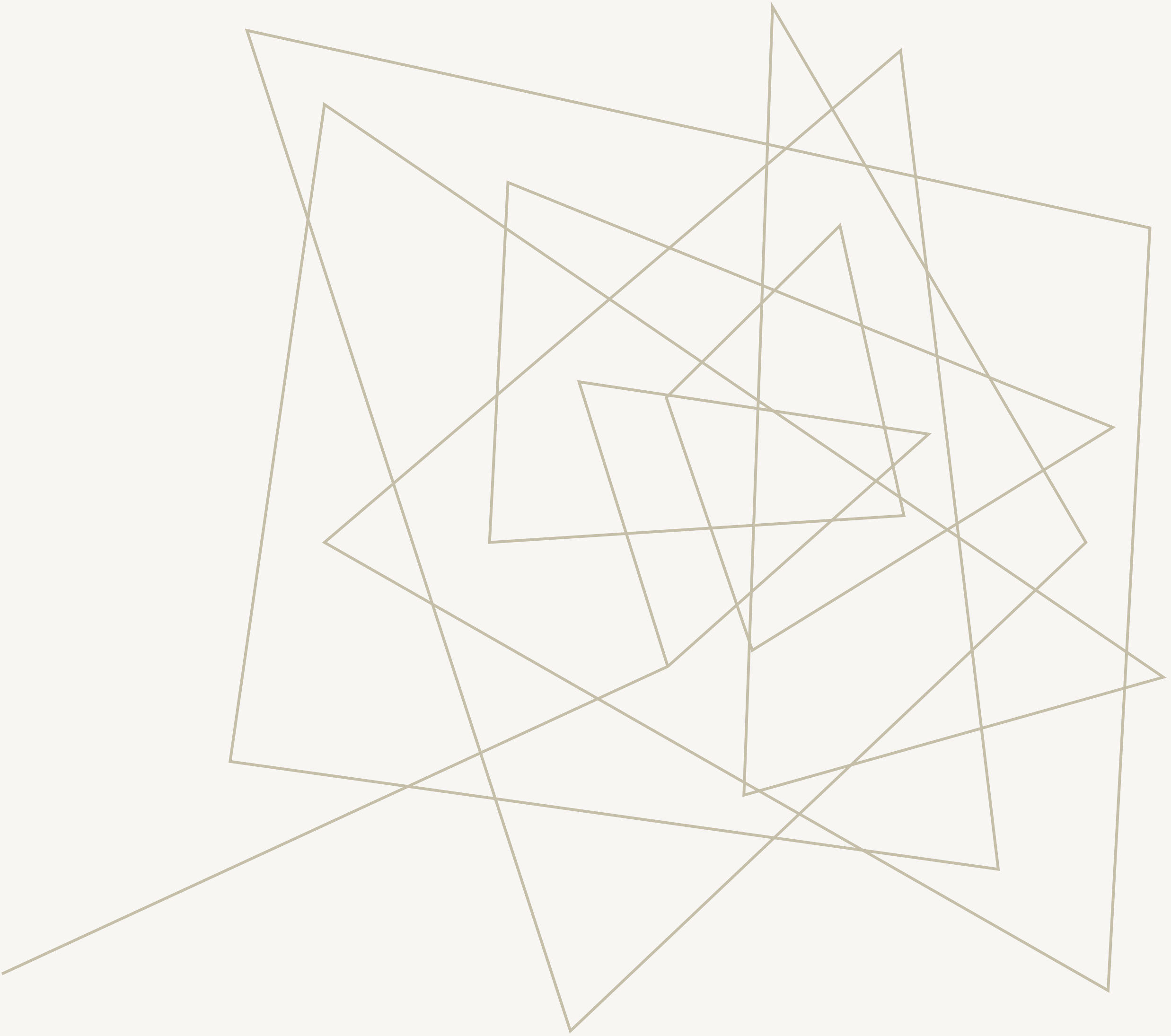
3. `title.basics.csv`: each record represents a movie, with additional columns such as the average ratings

METHODS

Data Preparation

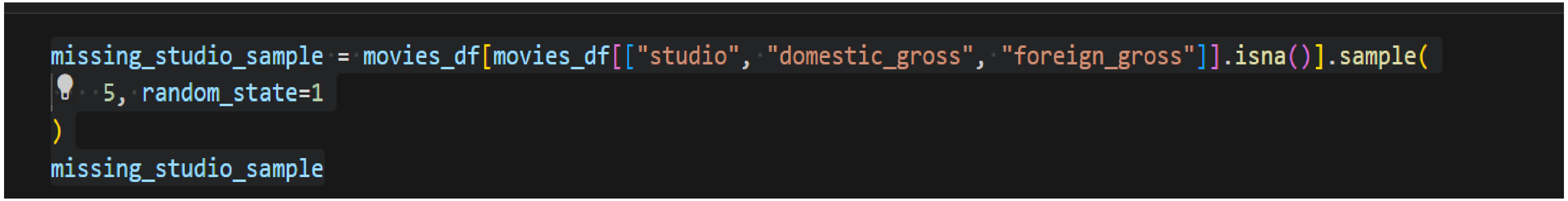
Data Cleaning

Data Modelling

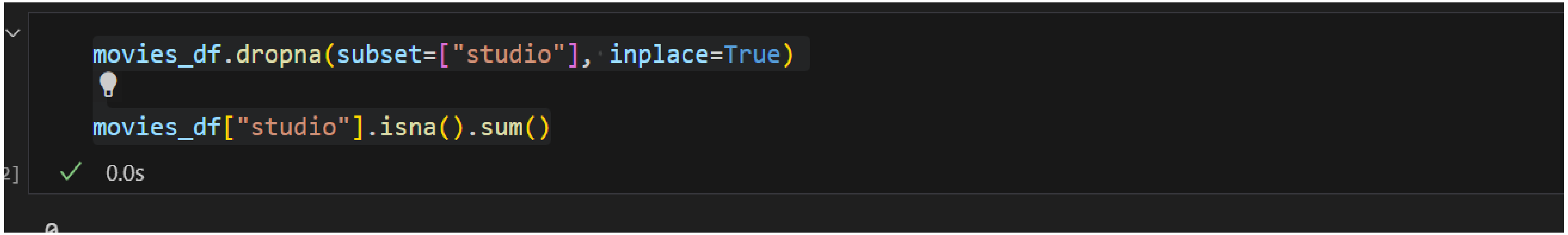


DATA PREPARATION AND CLEANING

- For this procedure, we are required to determine the rows with missing values and choose the appropriate methods to deal with such outliers, for analysis purposes. For example;



```
missing_studio_sample = movies_df[movies_df[["studio", "domestic_gross", "foreign_gross"]].isna()].sample(
    5, random_state=1
)
missing_studio_sample
```

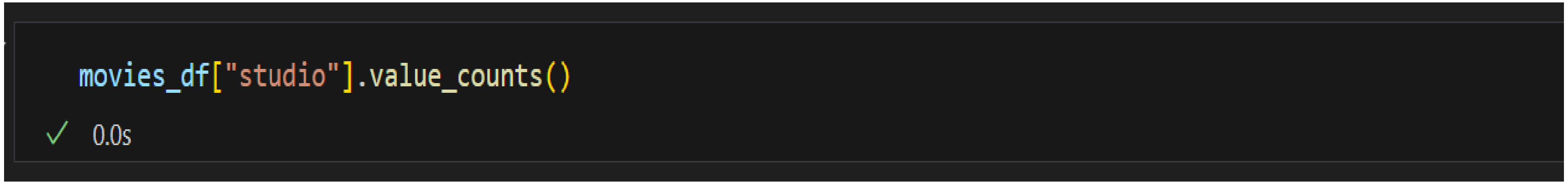


```
movies_df.dropna(subset=["studio"], inplace=True)

movies_df["studio"].isna().sum()
```

✓ 0.0s

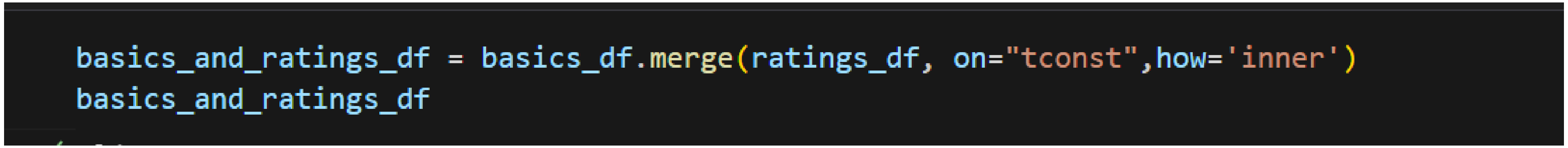
```
0
```



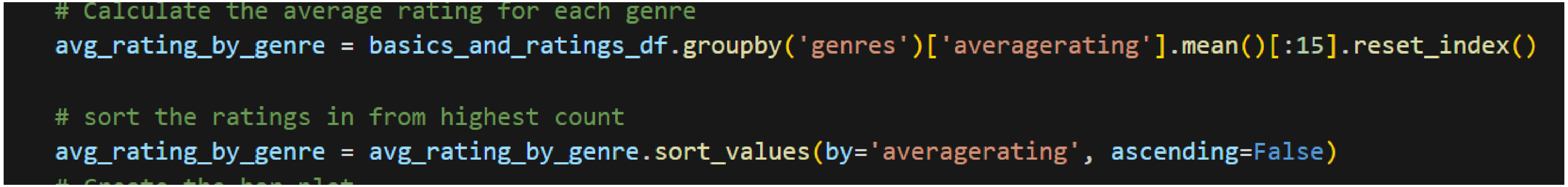
```
movies_df["studio"].value_counts()
```

✓ 0.0s

- Here we are also required to Join the data frames with common features if need be.
- The grouping of the different dataset information also comes in handy when we want to represent specific data values. For example;

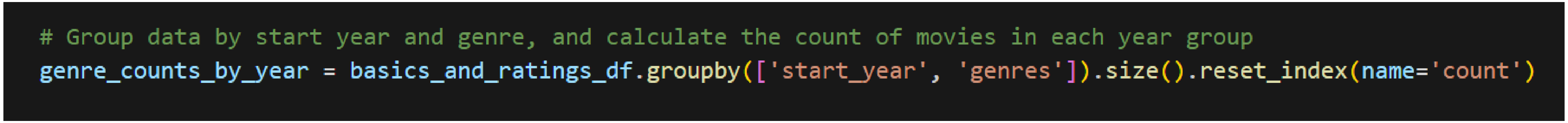


```
basics_and_ratings_df = basics_df.merge(ratings_df, on="tconst",how='inner')
basics_and_ratings_df
```

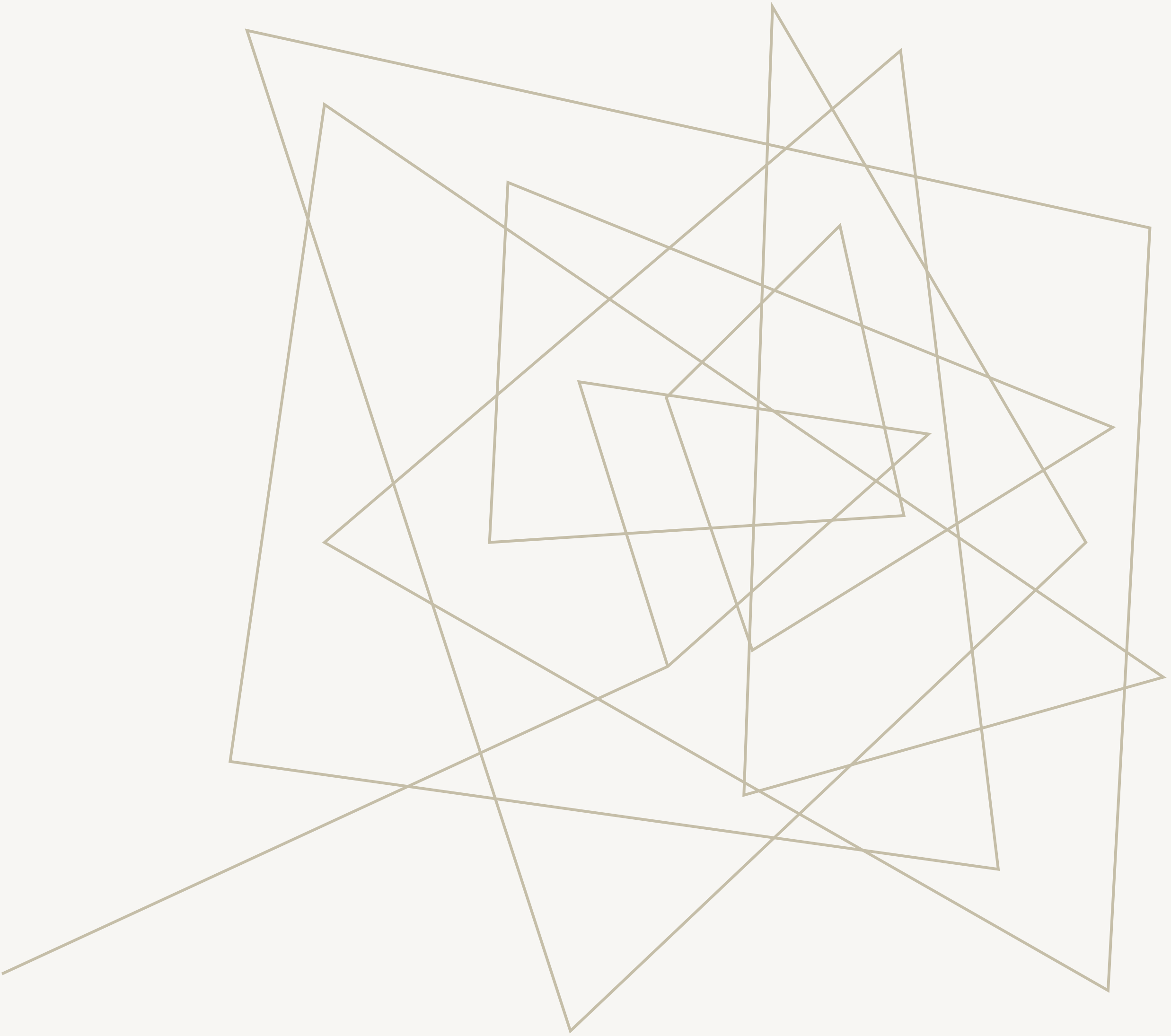


```
# Calculate the average rating for each genre
avg_rating_by_genre = basics_and_ratings_df.groupby('genres')['averagerating'].mean()[:15].reset_index()

# sort the ratings in from highest count
avg_rating_by_genre = avg_rating_by_genre.sort_values(by='averagerating', ascending=False)
```



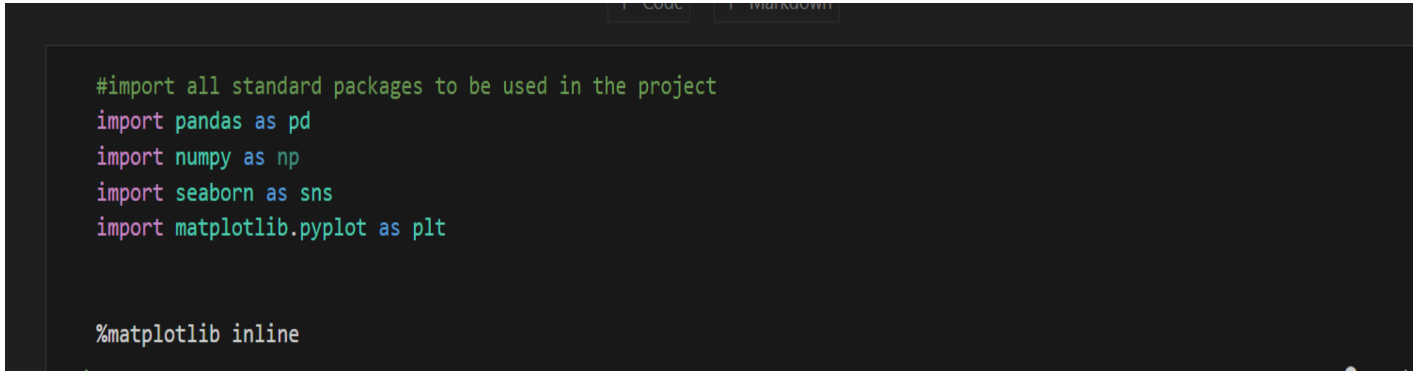
```
# Group data by start year and genre, and calculate the count of movies in each year group
genre_counts_by_year = basics_and_ratings_df.groupby(['start_year', 'genres']).size().reset_index(name='count')
```



DATA MODELLING

- For this Step, I was able to differentiate the choices in terms of the questions asked and represented data in the different models:
- Some of the plot libraries used to represent the answers or create visualization include;

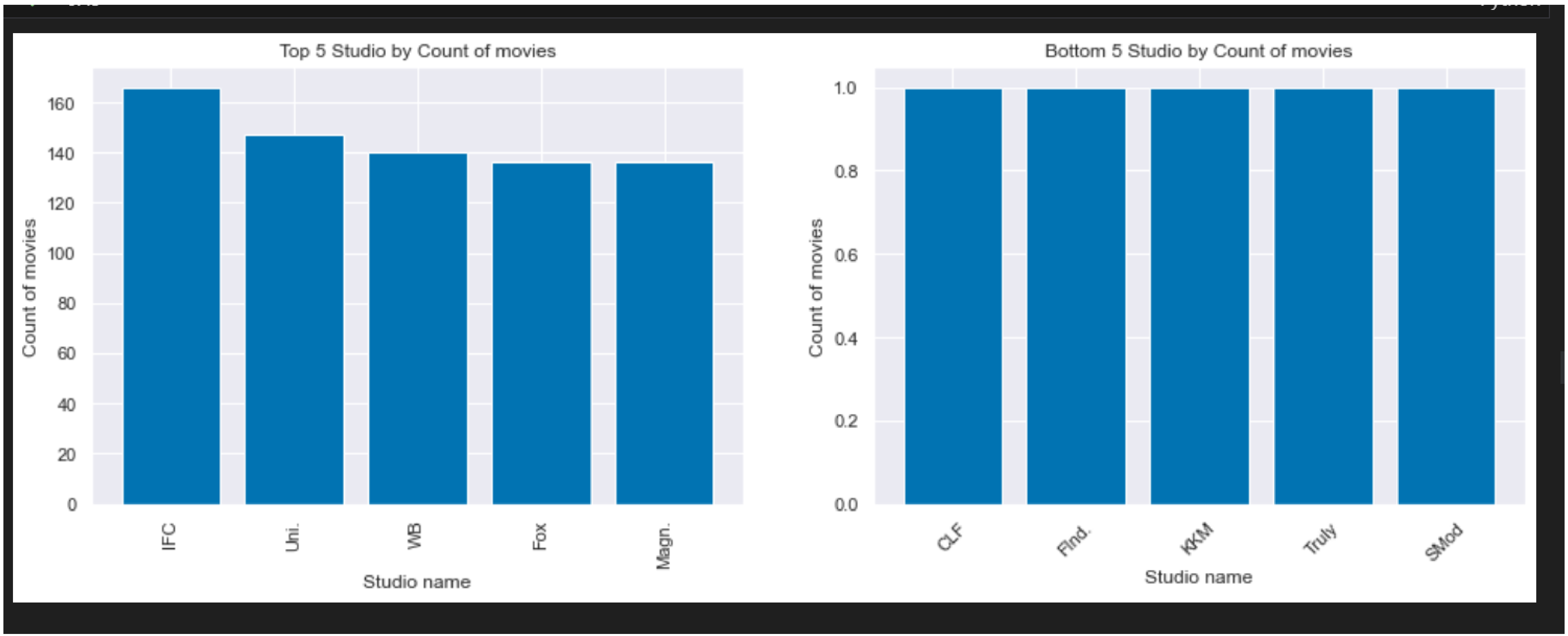
1. Matplot libraries
2. Seaborn libraries



```
#import all standard packages to be used in the project
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt


%matplotlib inline
```

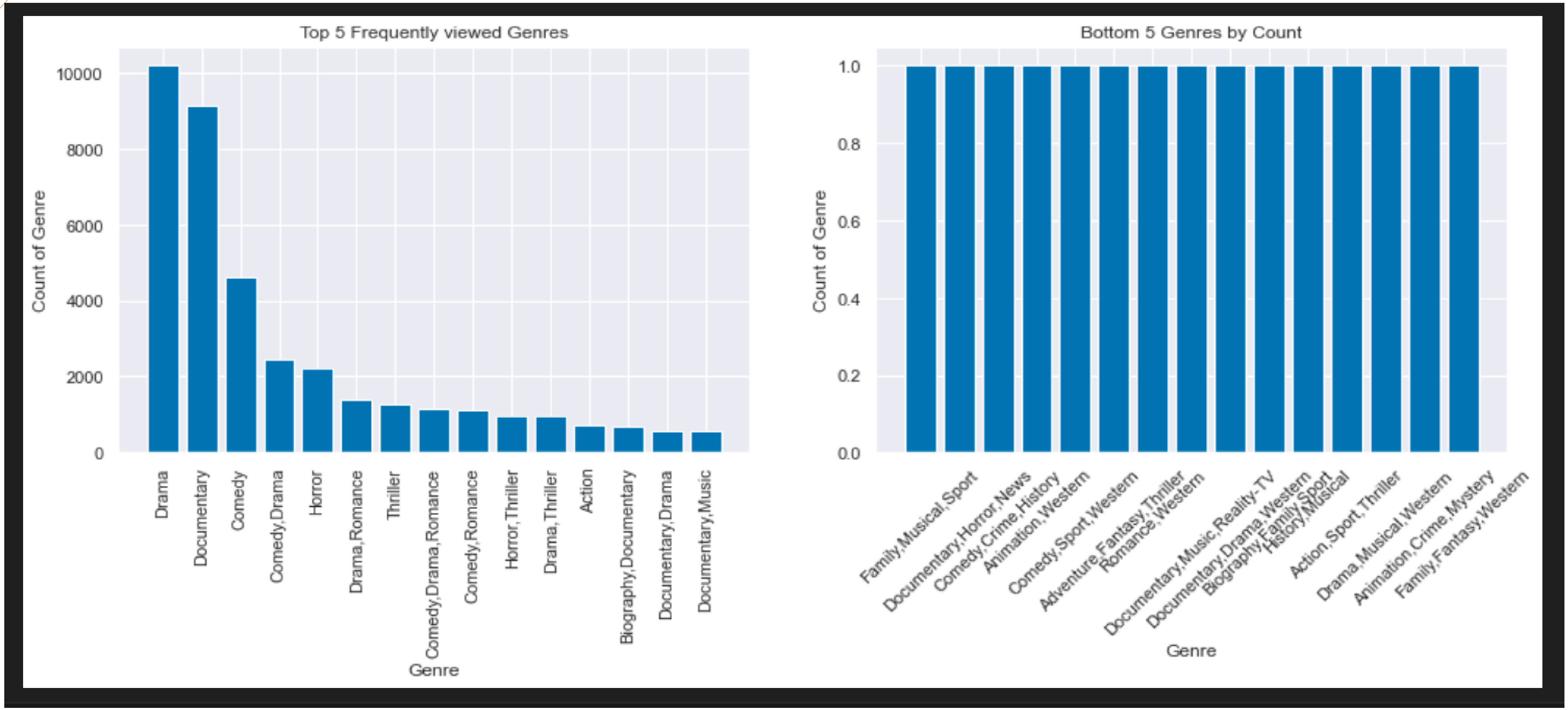
DATA MODELS



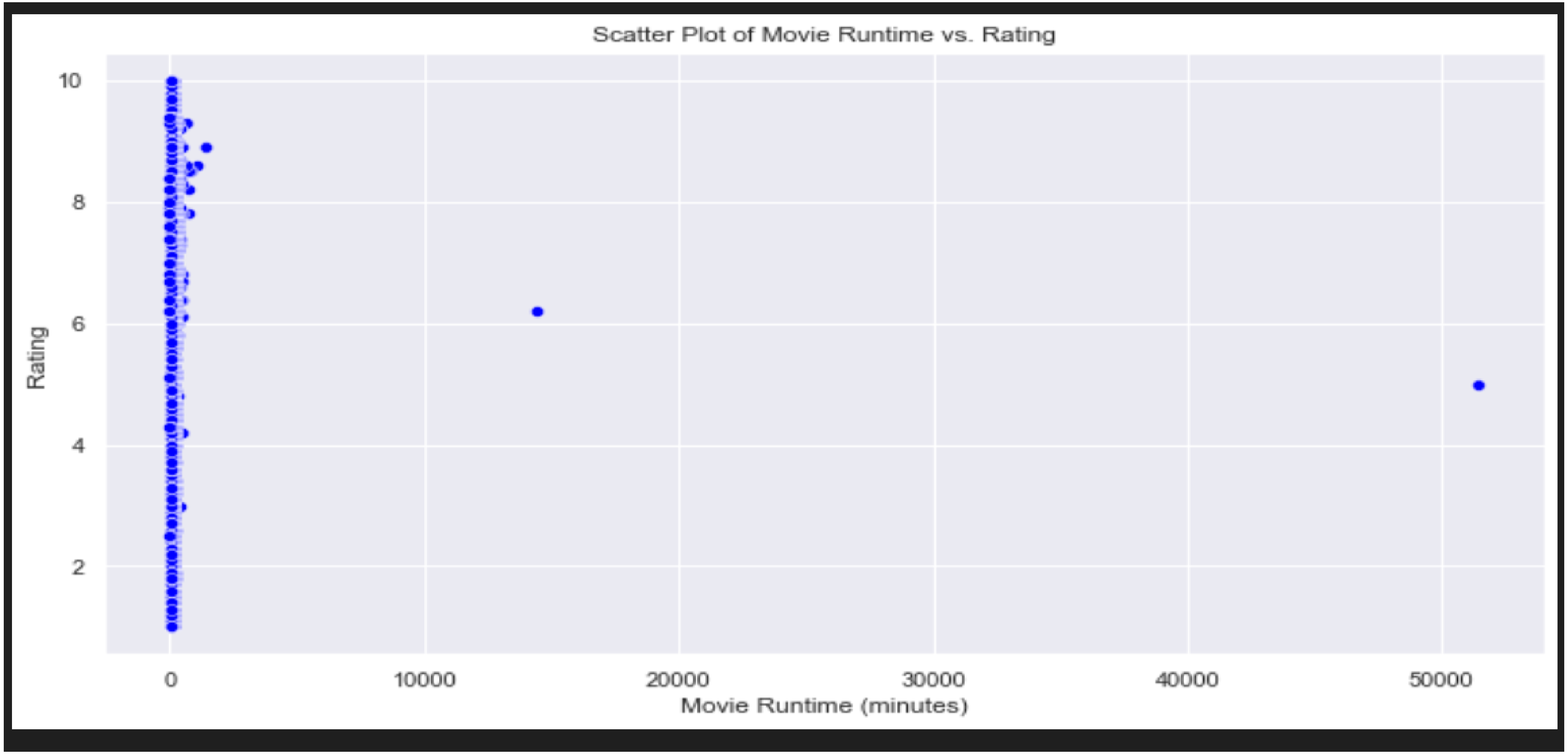
Top 5 Studio by Count of movies

Bottom 5 Studio by Count of movies

DATA MODELS

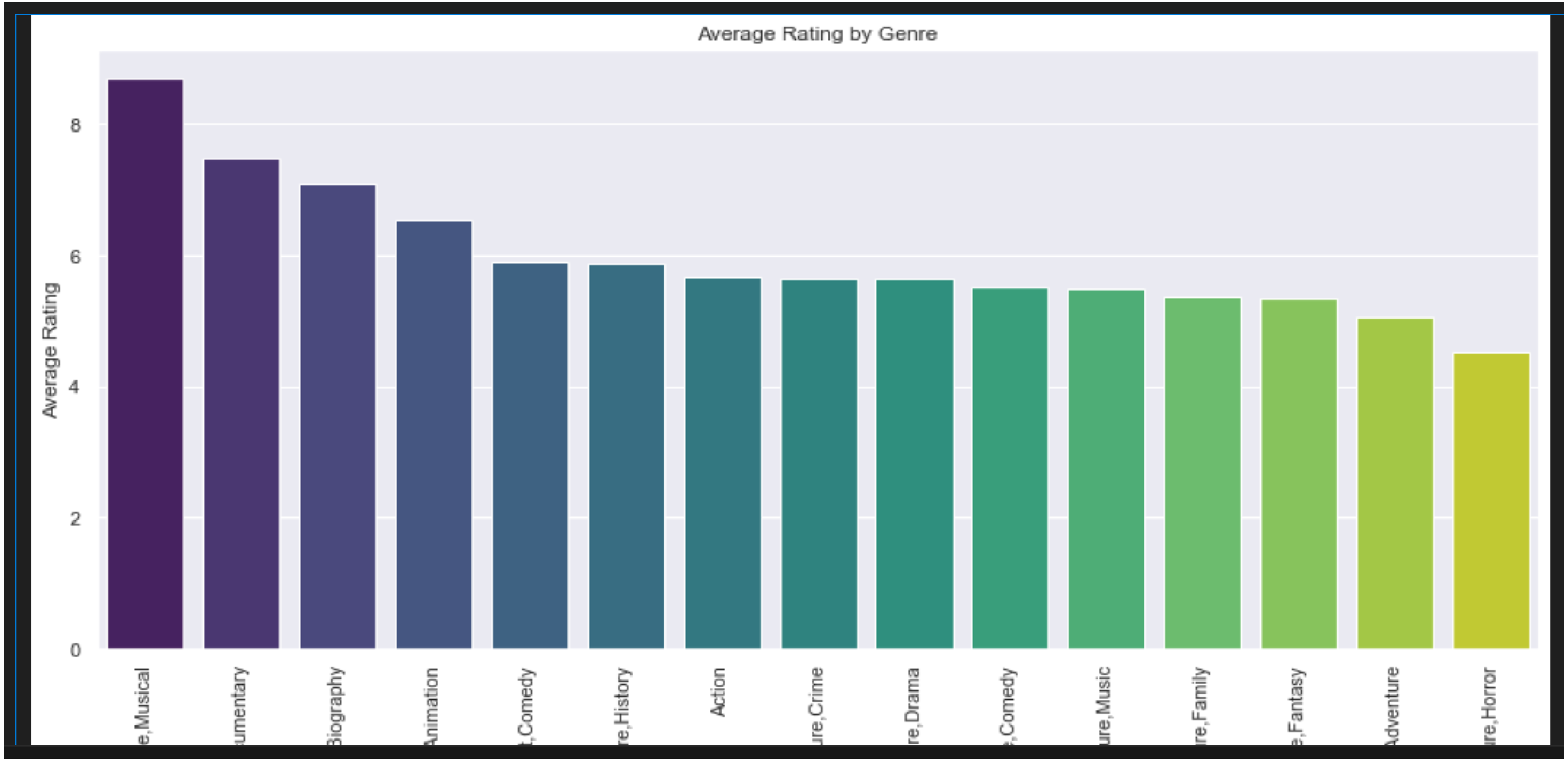

Top 5 Frequently viewed Genres
Count of Genre
0
2000
4000
6000
8000
10000
Drama
Documentary
Comedy
Comedy,Drama
Horror
Drama,Romance
Thriller
Comedy,Drama,Romance
Comedy,Romance
Horror,Thriller
Drama,Thriller
Action
Biography,Documentary
Documentary,Drama
Documentary,Music
Genre
Bottom 5 Genres by Count
Count of Genre
0.0
0.2
0.4
0.6
0.8
1.0
Family,Musical,Sport
Documentary,Horror,News
Comedy,Crime,History
Animation,Western
Comedy,Sport,Western
Adventure,Fantasy,Thriller
Romance,Western
Documentary,Music,Reality-TV
Documentary,Drama,Western
Biography,Family,Sport
History,Musical
Action,Sport,Thriller
Drama,Musical,Western
Animation,Crime,Mystery
Family,Fantasy,Western
Genre

DATA MODELS



Scatter Plot of Movie Runtime vs. Rating
Rating
10
8
6
4
2
0
10000
20000
30000
40000
50000
Movie Runtime (minutes)

DATA MODELS


Average Rating by Genre
Average Rating
0
2
4
6
8
e,Musical
umentary
Biography
Animation
,Comedy
re,History
Action
ure,Crime
re,Drama
,Comedy
ure,Music
re,Family
e,Fantasy
Adventure
ure,Horror

CONCLUSION

The analysis done for the above dataset chosen yields the following conclusion.

1. The most preferred genre / type of films are the Action-based.

2. In the recent years, the most produced type of films are the Action-based Films

3. The year that returned the most domestic gross is 2018. We are able to relate this with the genre of film produced that year - which represents the 'Action' based films.

RECOMMENDATION

From the analysis done above, I would recommend that the Stakeholders:

1. Should consider creating Action-based films

2. Should consider creating films with less runtime runtime in minutes. In doing so, this will yield to higher ratings from viewers and in turn more revenue for the stakeholder.

SUMMARY

From the analysis done above, I would recommend that the Stakeholders should consider creating Action-based films with an average runtime in minutes. In doing so, this will yield to higher ratings from viewers and in turn more revenue for the stakeholder.

THANK YOU

Lesley Kamamo

DATA SCIENCE – PART TIME

PHASE 1 PROJECT