ANALYSIS BY ETHEN DCOSTA

IMDB - MOVIE ANALYSIS

USING PYTHON (PANDAS, NUMPY, SEABORN, MATPLOTLIB)

HELLO!

- I am **Ethen D'Costa**, a detail-oriented and analytical fresher with a strong foundation in data analysis, visualization, and problem-solving.
- I am proficient in tools like Python, SQL, Excel, Power BI, and Tableau, with hands-on experience in creating dashboards and performing exploratory data analysis (EDA).
- I excel at identifying trends, drawing insights, and presenting data-driven recommendations.
- I am eager to apply my technical and analytical skills to real-world business challenges and contribute to organizational success.

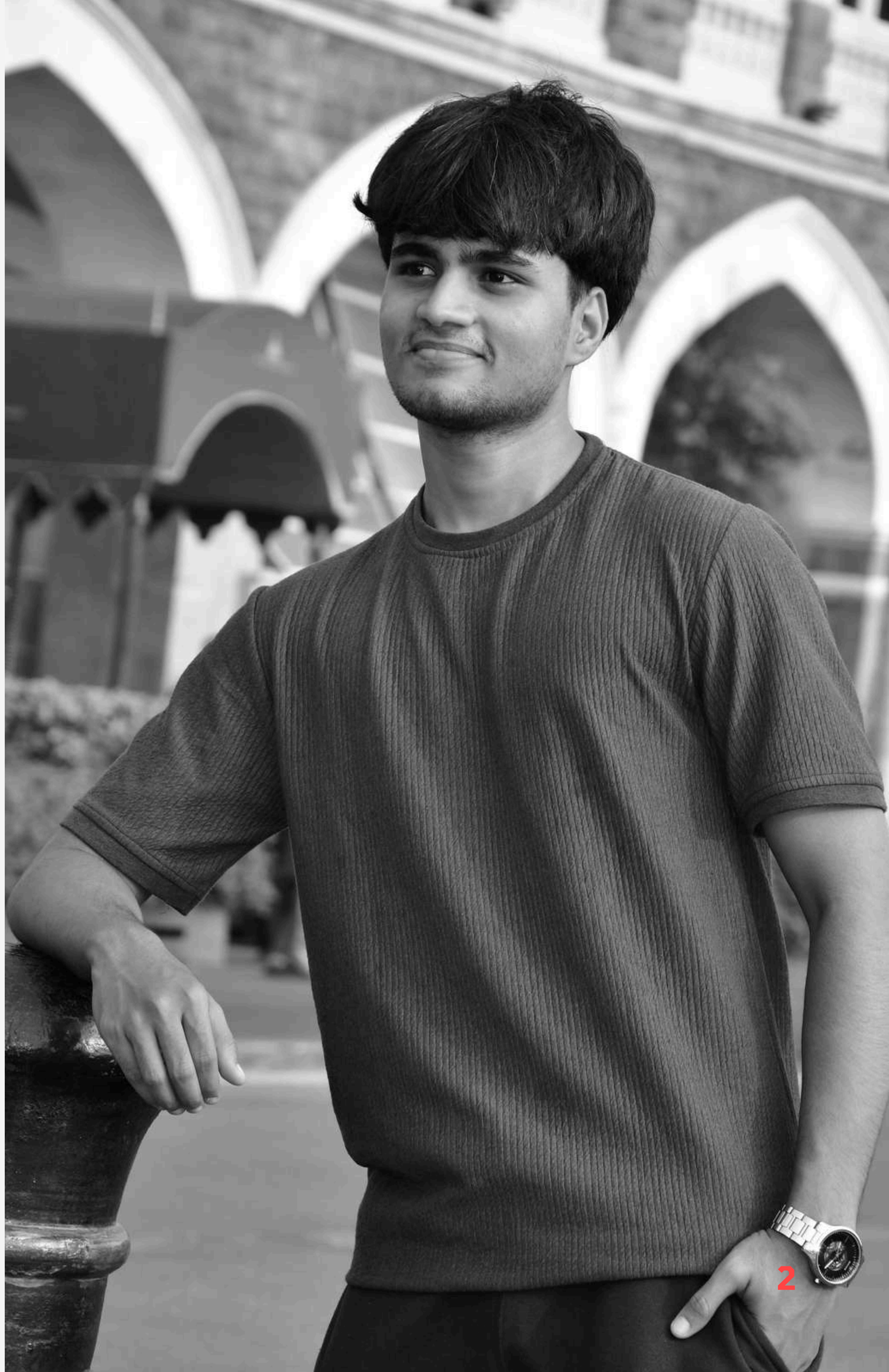


Table of Contents

OBJECTIVE

- As a data analyst intern at IMDB, you have been tasked with exploring and analyzing the IMDB Movies dataset.

- Your goal is to answer specific business questions, gain insights into movie trends, and deliver actionable recommendations.

- Using Python and libraries such as Pandas, NumPy, Seaborn, and Matplotlib, perform analysis to help IMDB better understand genre popularity, score trends, and factors influencing movie success.

PROJECT SETUP AND DATA LOADING

01

What libraries are required for this project, and why are they useful in data analysis?

1. Pandas
Pandas is a powerful library for data manipulation and analysis.

2. NumPy
NumPy is used for numerical computations and handling large arrays and matrices.

3. Matplotlib
Matplotlib is a low-level library for creating static, interactive, and animated visualizations.

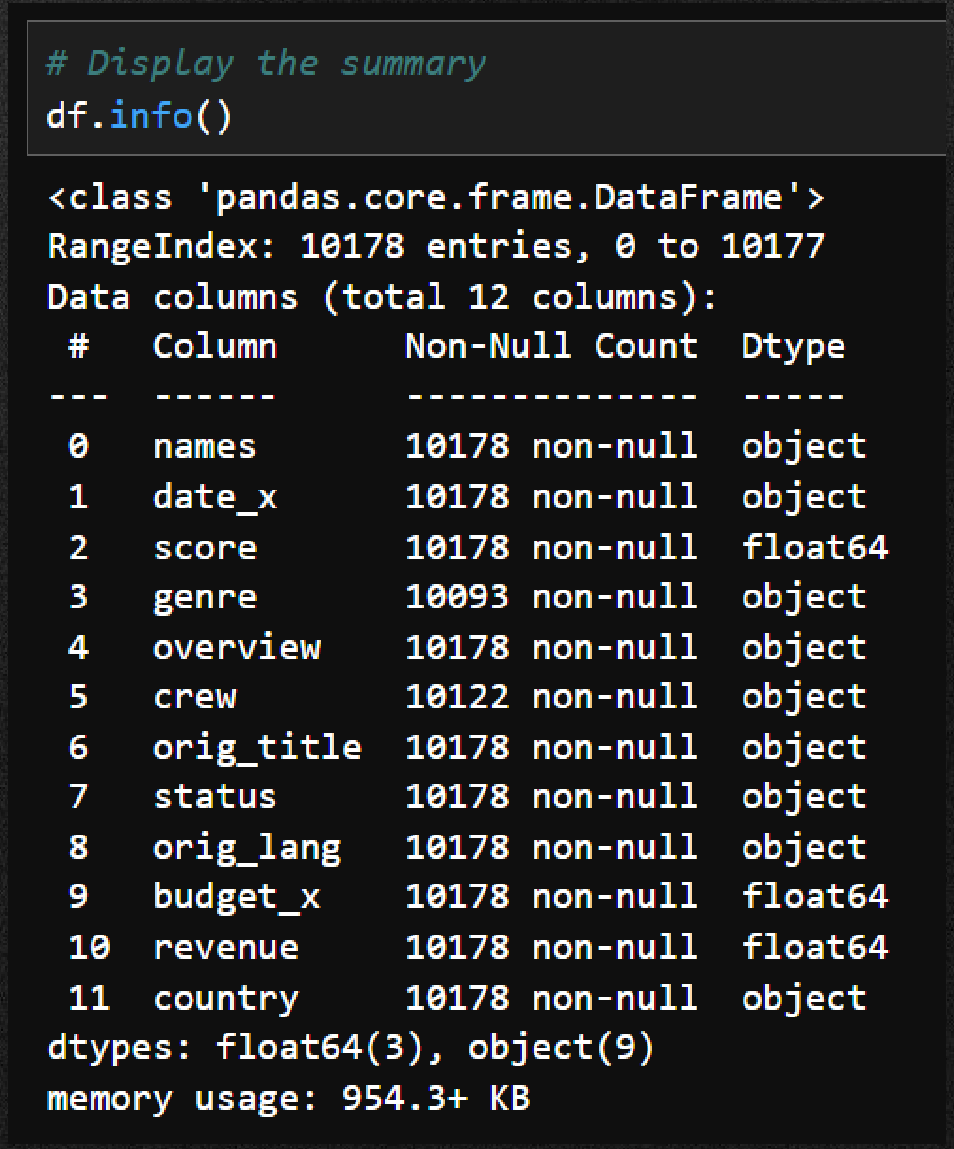
4. Seaborn
Seaborn is a high-level library for statistical data visualization built on top of Matplotlib.

02

Load the dataset. What is the shape of the dataset? What does each row and column represent?

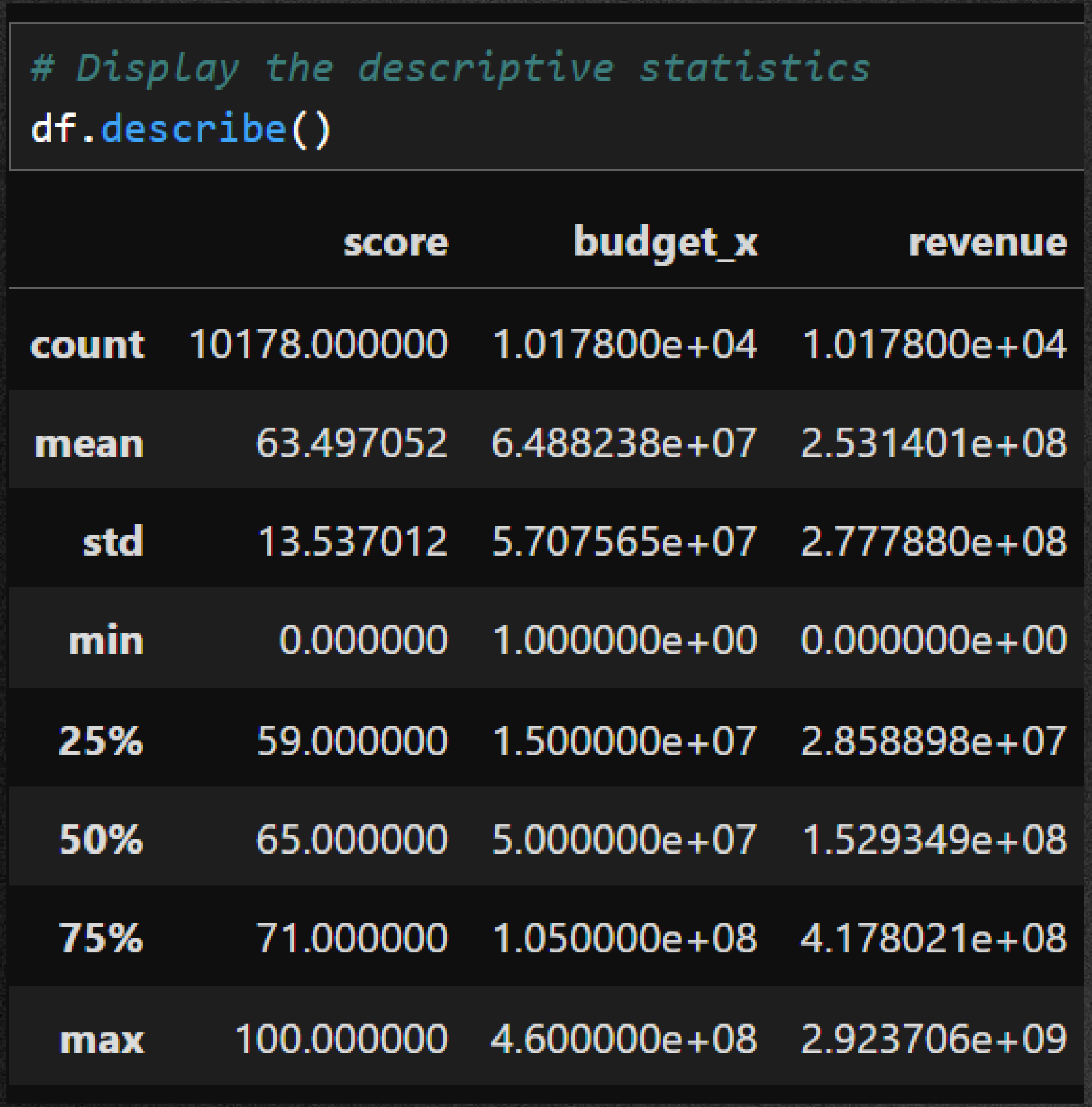
- The shape of the dataset is **(10178, 12).**
- Each row in the dataset corresponds to a movie and its components.
- Each column in the dataset corresponds to the various factors contributing to the movie such as the name of the movie, the date when the movie was released, the score, genre, overview, crew, original title, status, original language in which the movie was created, budget, revenue and the country where it was created.

DATA OVERVIEW AND BASIC EXPLORATION



```python
# Display the summary
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10178 entries, 0 to 10177
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   names       10178 non-null  object
 1   date_x      10178 non-null  object
 2   score       10178 non-null  float64
 3   genre       10093 non-null  object
 4   overview    10178 non-null  object
 5   crew        10122 non-null  object
 6   orig_title  10178 non-null  object
 7   status      10178 non-null  object
 8   orig_lang   10178 non-null  object
 9   budget_x    10178 non-null  float64
 10  revenue     10178 non-null  float64
 11  country     10178 non-null  object
dtypes: float64(3), object(9)
memory usage: 954.3+ KB
```



```python
# Display the descriptive statistics
df.describe()
```

	score	budget_x	revenue
count	10178.000000	1.017800e+04	1.017800e+04
mean	63.497052	6.488238e+07	2.531401e+08
std	13.537012	5.707565e+07	2.777880e+08
min	0.000000	1.000000e+00	0.000000e+00
25%	59.000000	1.500000e+07	2.858898e+07
50%	65.000000	5.000000e+07	1.529349e+08
75%	71.000000	1.050000e+08	4.178021e+08
max	100.000000	4.600000e+08	2.923706e+09

1. Score Column:

- Mean: 63.50
- Median (50%): 65.00
- Std (Standard Deviation): 13.54
- Range: 0 to 100

The mean and median are close, suggesting that the score distribution is approximately symmetric, though slightly left-skewed since the mean is less than the median. The standard deviation indicates some variability, but most values likely cluster around the mean, given the limited range (0–100).

2. Budget_x Column:

- Mean: $64.88M
- Median (50%): $50M
- Std: $57.08M
- Range: $1 to $460M

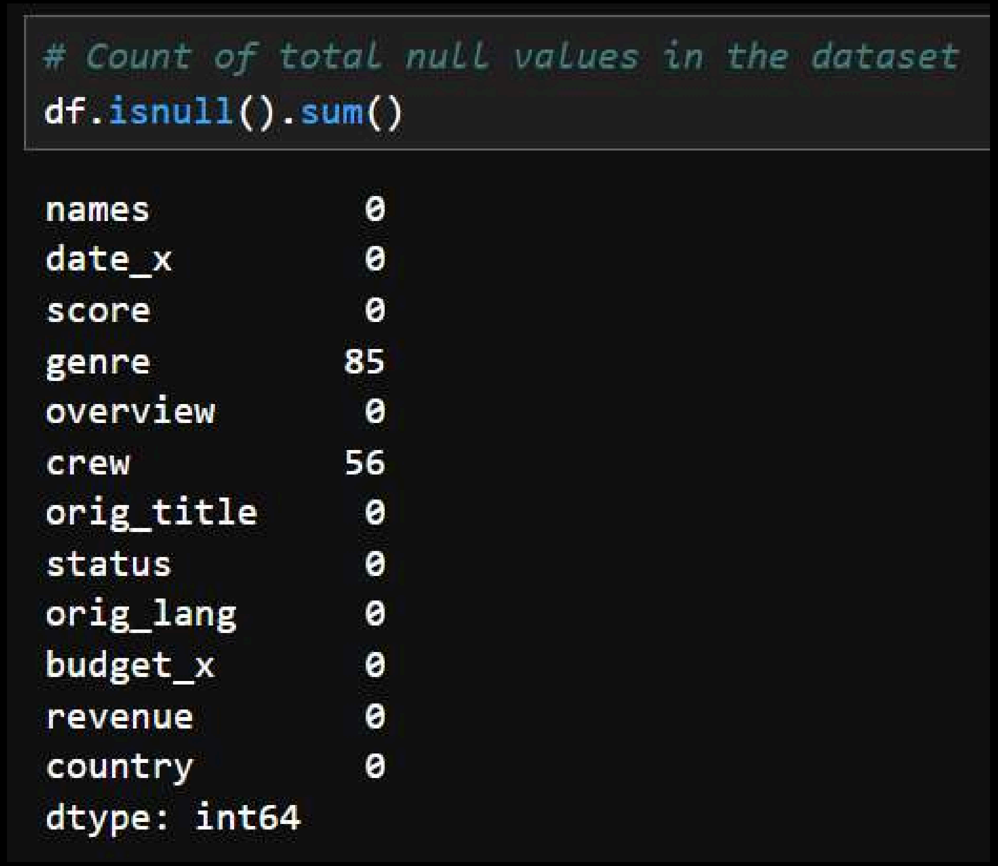
The mean is higher than the median, suggesting a right-skewed distribution (a few very high-budget projects drive the mean up). The high standard deviation relative to the mean further supports the presence of extreme outliers in budgets.

3. Revenue Column:

- Mean: $253.14M
- Median (50%): $152.93M
- Std: $277.79M
- Range: $0 to $2.92B

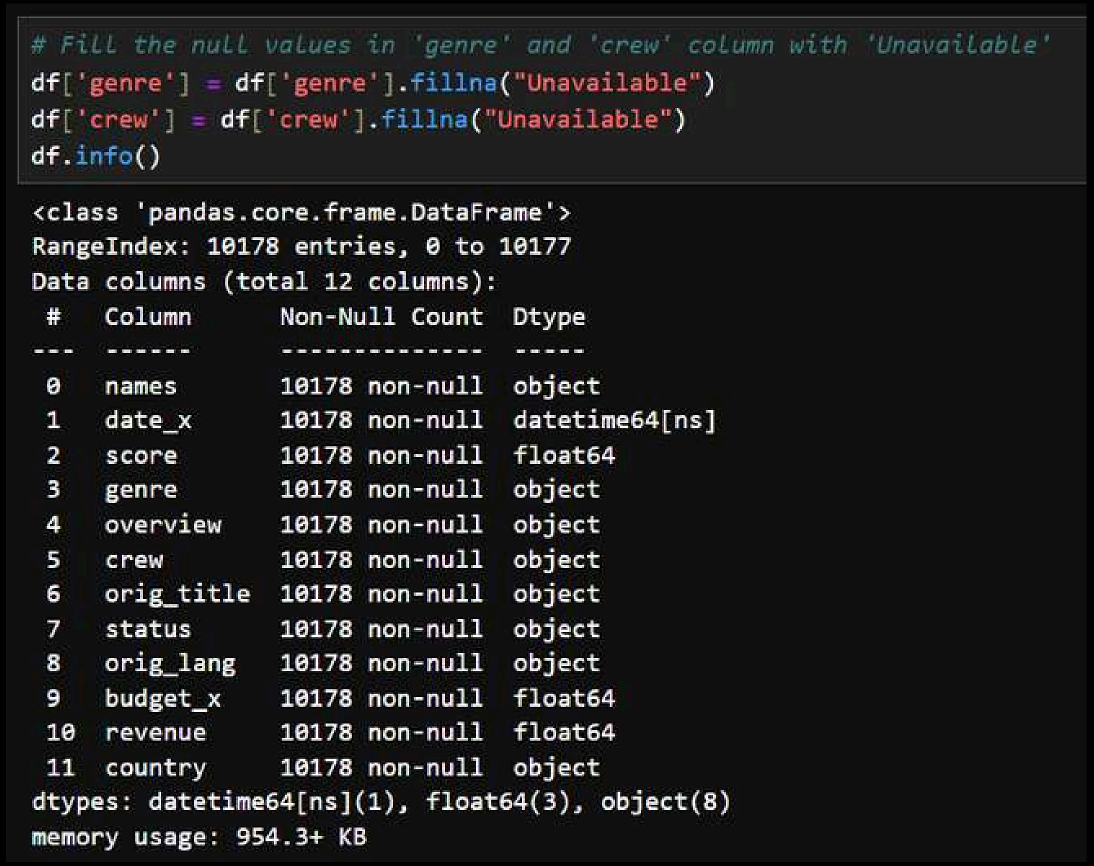
Similar to budget_x, the revenue column is also right-skewed due to the mean being much higher than the median. The very large range and high standard deviation indicate significant variability in revenue, likely driven by a few blockbuster films with exceptionally high earnings.

DATA CLEANING



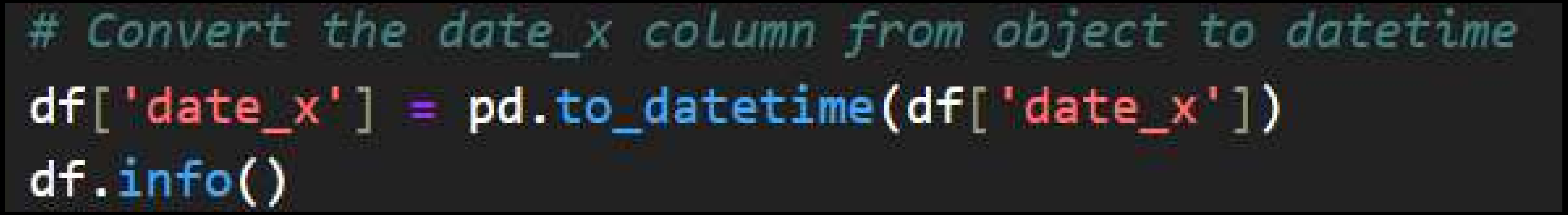
```
# Count of total null values in the dataset
df.isnull().sum()
```

```
names          0
date_x         0
score          0
genre         85
overview       0
crew          56
orig_title     0
status         0
orig_lang      0
budget_x       0
revenue        0
country        0
dtype: int64
```



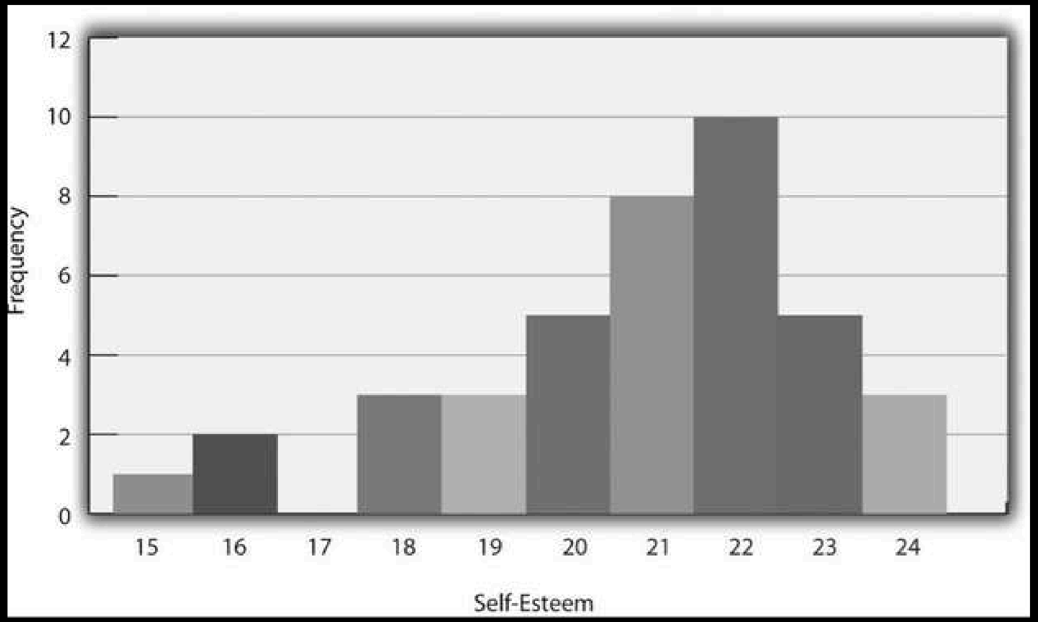
```
# Fill the null values in 'genre' and 'crew' column with 'Unavailable'
df['genre'] = df['genre'].fillna("Unavailable")
df['crew'] = df['crew'].fillna("Unavailable")
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10178 entries, 0 to 10177
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   names       10178 non-null  object
 1   date_x      10178 non-null  datetime64[ns]
 2   score       10178 non-null  float64
 3   genre       10178 non-null  object
 4   overview    10178 non-null  object
 5   crew        10178 non-null  object
 6   orig_title  10178 non-null  object
 7   status      10178 non-null  object
 8   orig_lang   10178 non-null  object
 9   budget_x    10178 non-null  float64
 10  revenue     10178 non-null  float64
 11  country     10178 non-null  object
dtypes: datetime64[ns](1), float64(3), object(8)
memory usage: 954.3+ KB
```

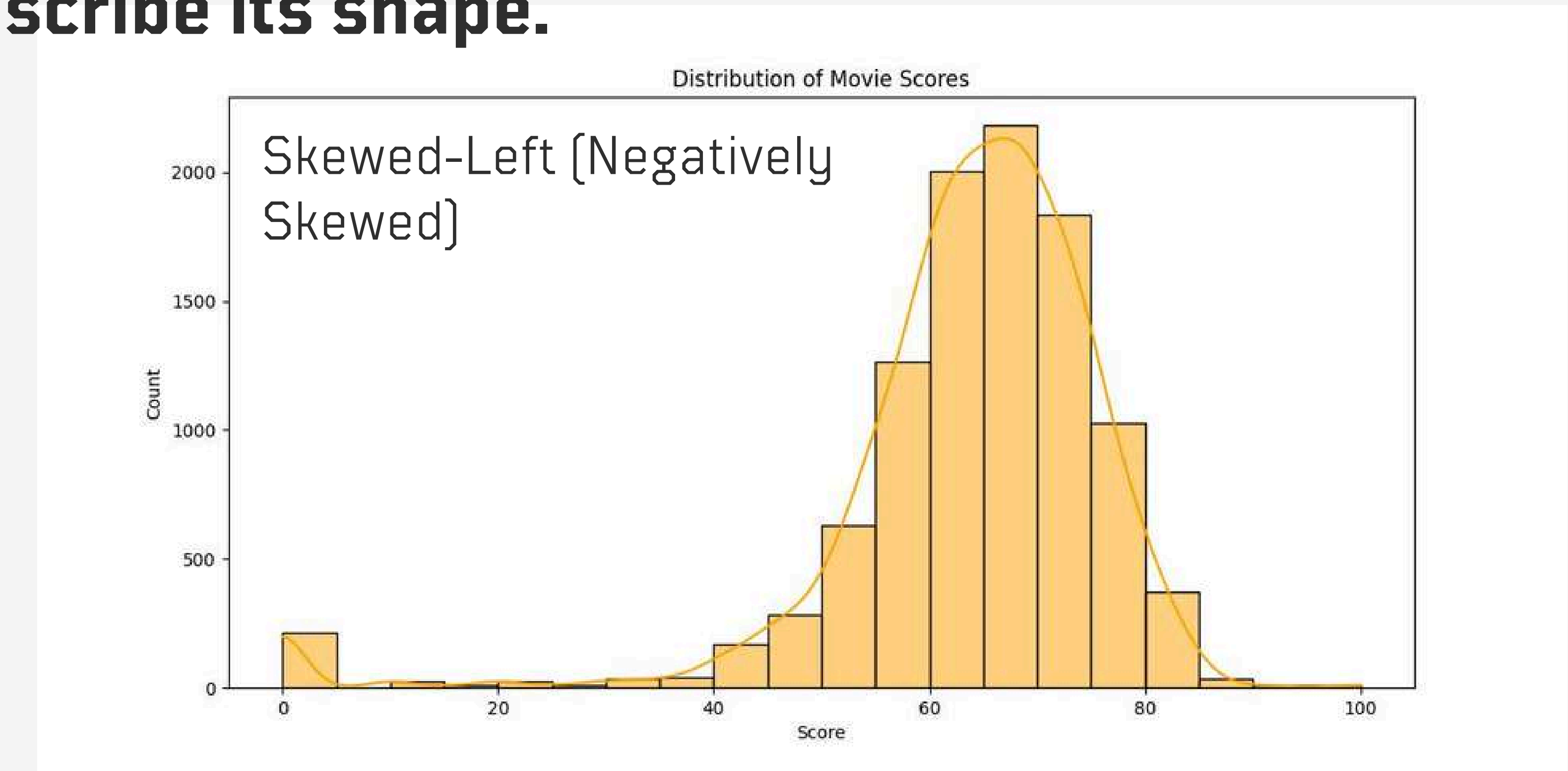


```
# Convert the date_x column from object to datetime
df['date_x'] = pd.to_datetime(df['date_x'])
df.info()
```

UNIVARIATE ANALYSIS

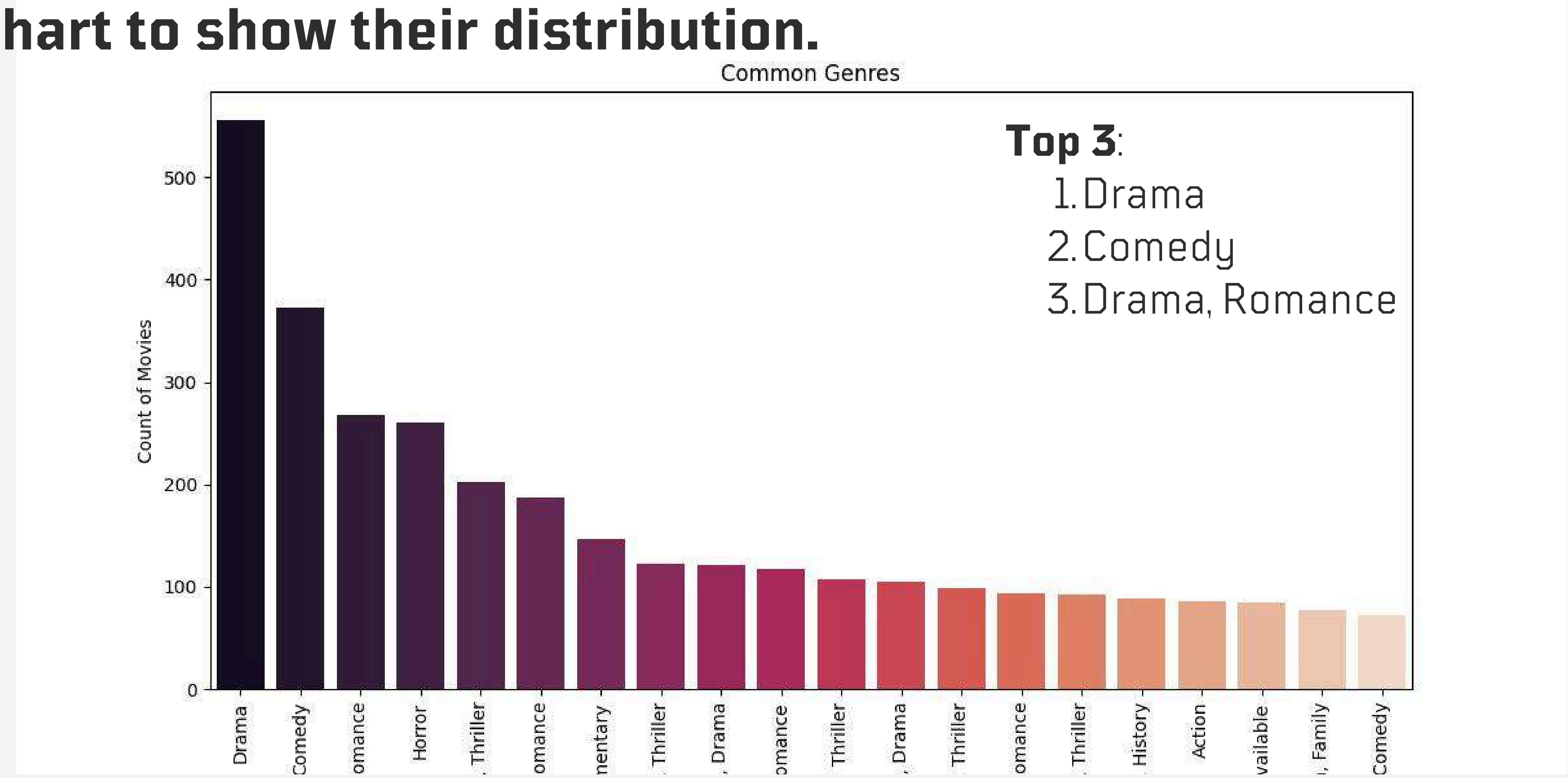


What is the distribution of movie scores? Plot a histogram and describe its shape.



Skewed-Left (Negatively Skewed)

What are the most common genres in the dataset? Use a bar chart to show their distribution.

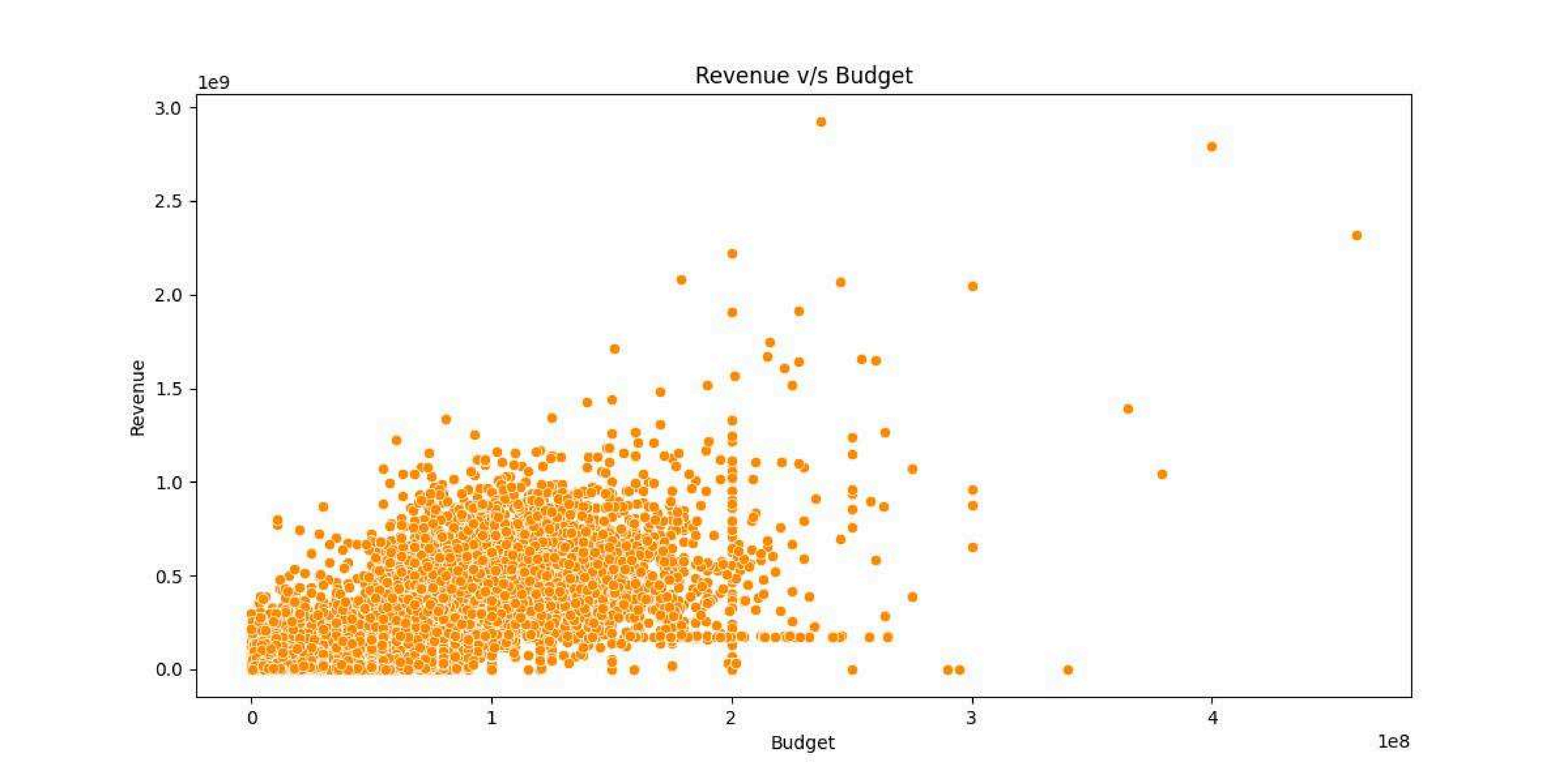


Top 3:

1. Drama
2. Comedy
3. Drama, Romance

BIVARIATE ANALYSIS

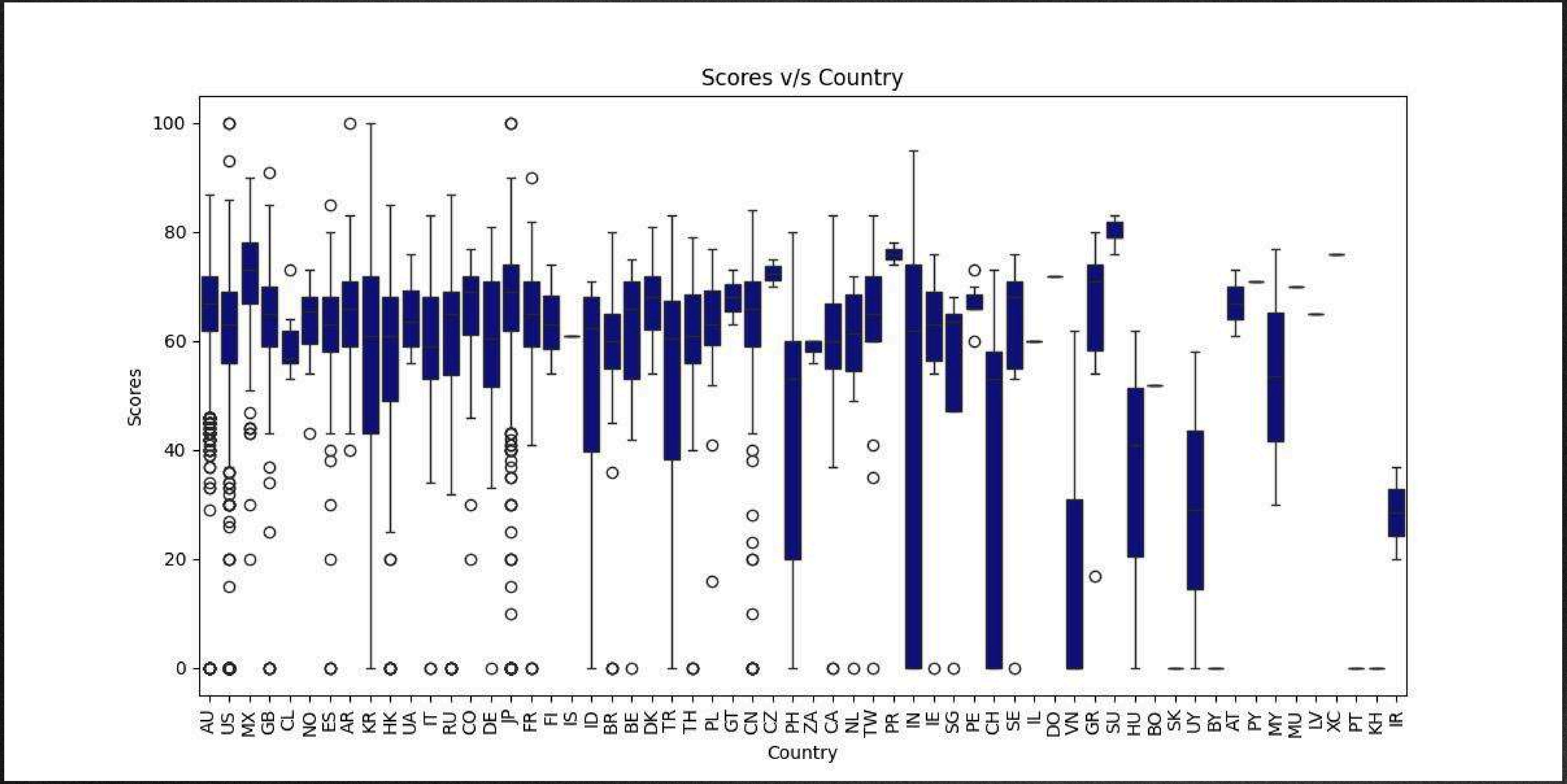
Is there a relationship between a movie's budget and its revenue? Plot a scatter plot and describe any observed trend.



There appears to be a general positive trend, where movies with higher budgets tend to generate higher revenues. However, the correlation is not perfect, as there is significant spread.

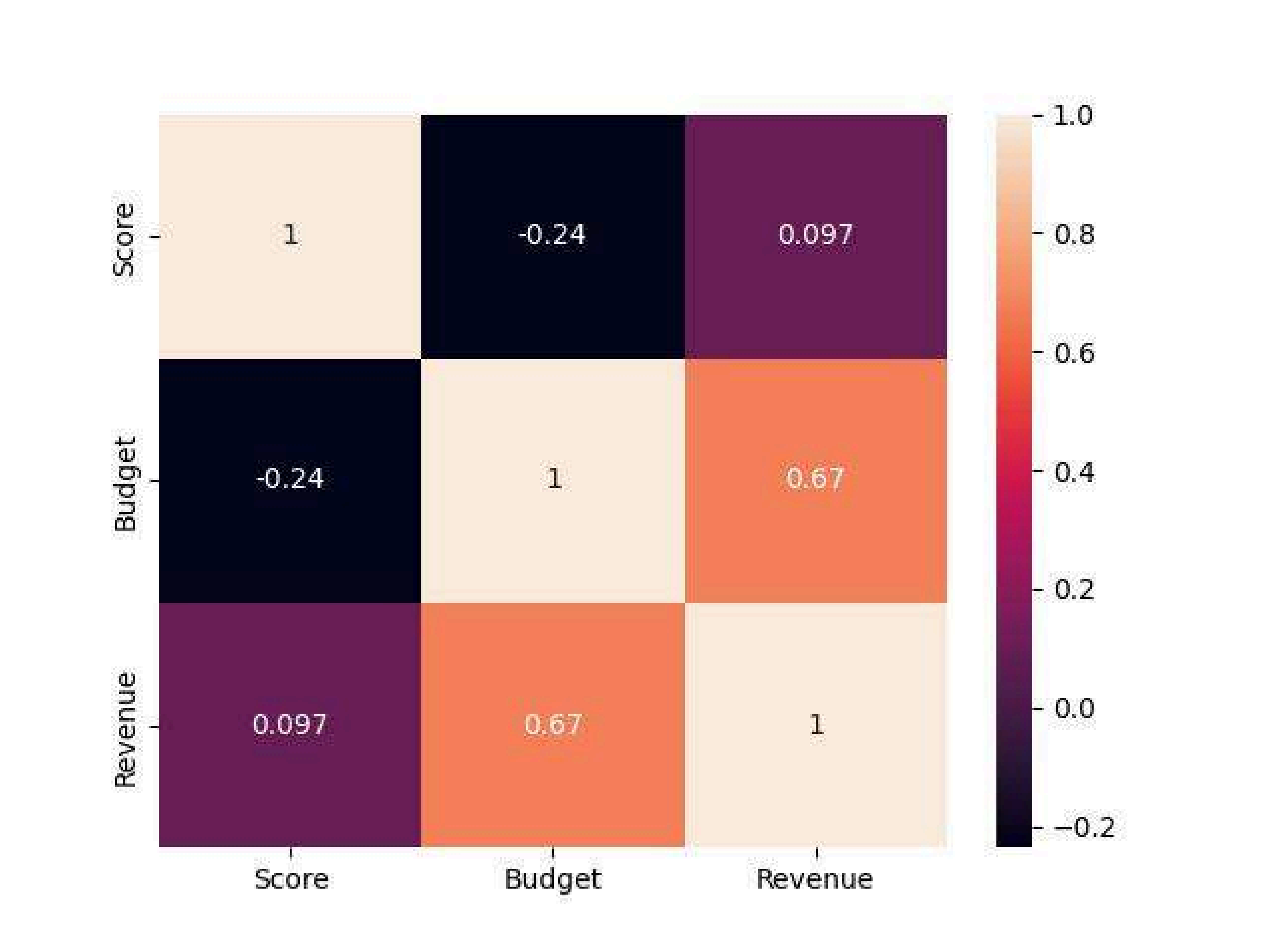
BIVARIATE ANALYSIS

How do scores vary by country? Use a boxplot to visualize the differences in scores across country.



BIVARIATE ANALYSIS

Is there a correlation between the score a movie received against its budget and revenue? Create a heatmap and calculate the correlation coefficient. What can you conclude?



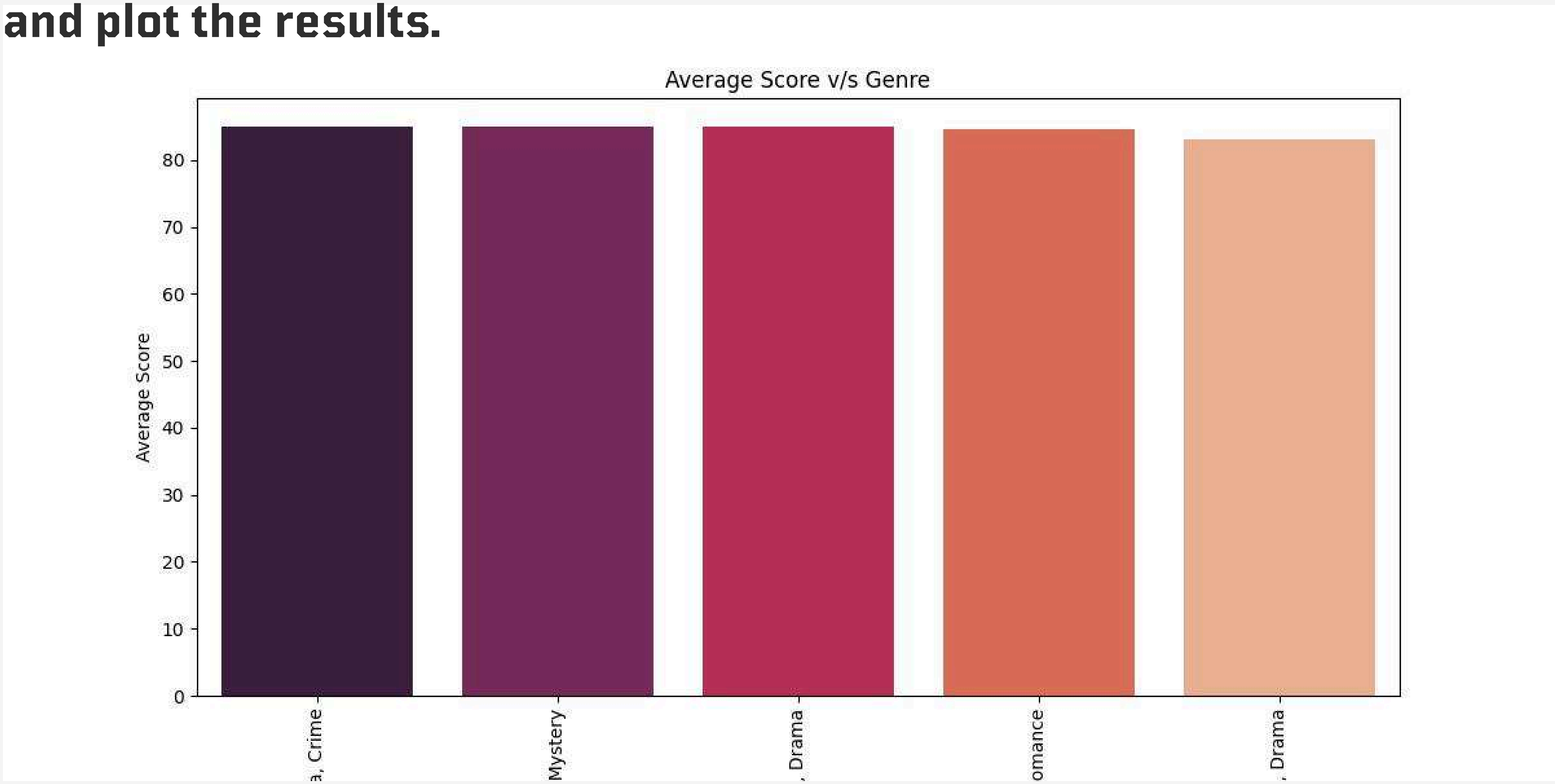
Score vs. Budget:
A weak negative correlation (-0.24) suggests that higher budgets may slightly correspond to lower scores, though the relationship is not strong.

Score vs. Revenue:
A very weak positive correlation (0.097) indicates almost no relationship between scores and revenue.

Budget vs. Revenue:
A moderate positive correlation (0.67) shows that higher budgets are often associated with higher revenues.

GENRE-SPECIFIC ANALYSIS

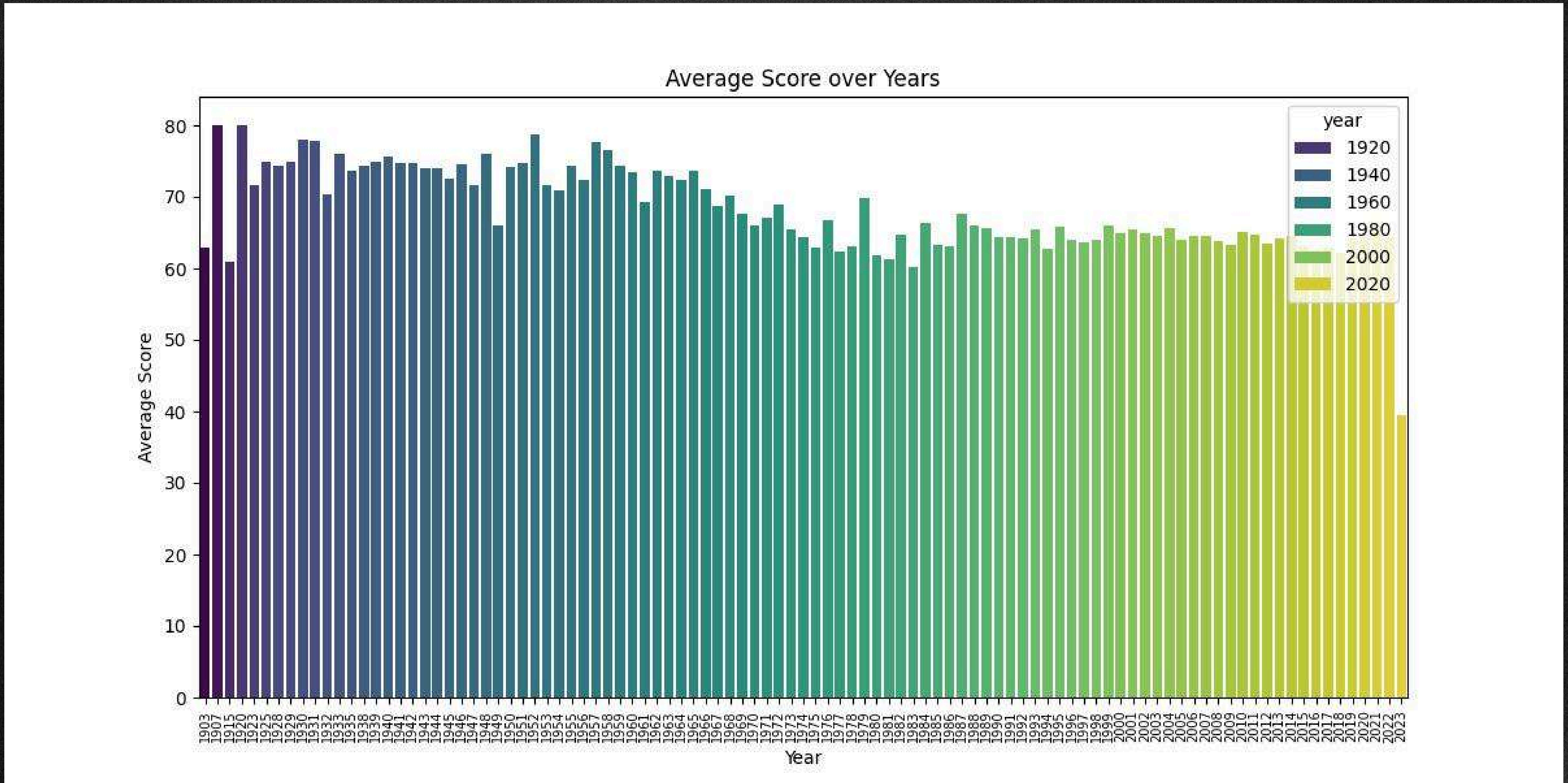
Which genre has the highest average score? Calculate the average score for each genre and plot the results.



Fantasy, Drama and Crime is the genre with the highest average score.

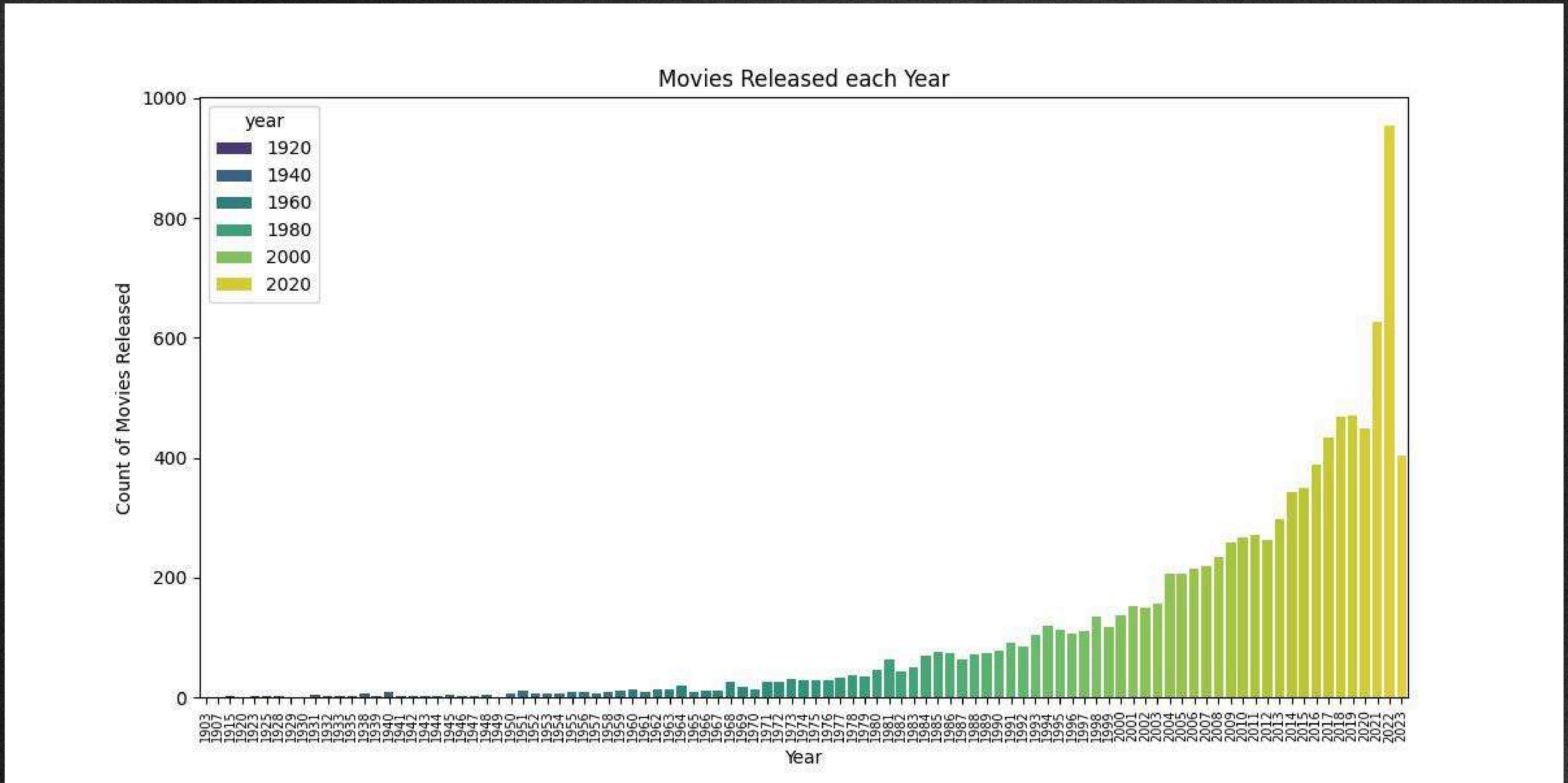
YEAR AND TREND ANALYSIS

How has the average score changed over the years? Plot the average score for each year.



YEAR AND TREND ANALYSIS

Which years had the highest and lowest number of movie releases? Plot the number of movies released each year.



<u>Year with the highest number of movie releases</u>

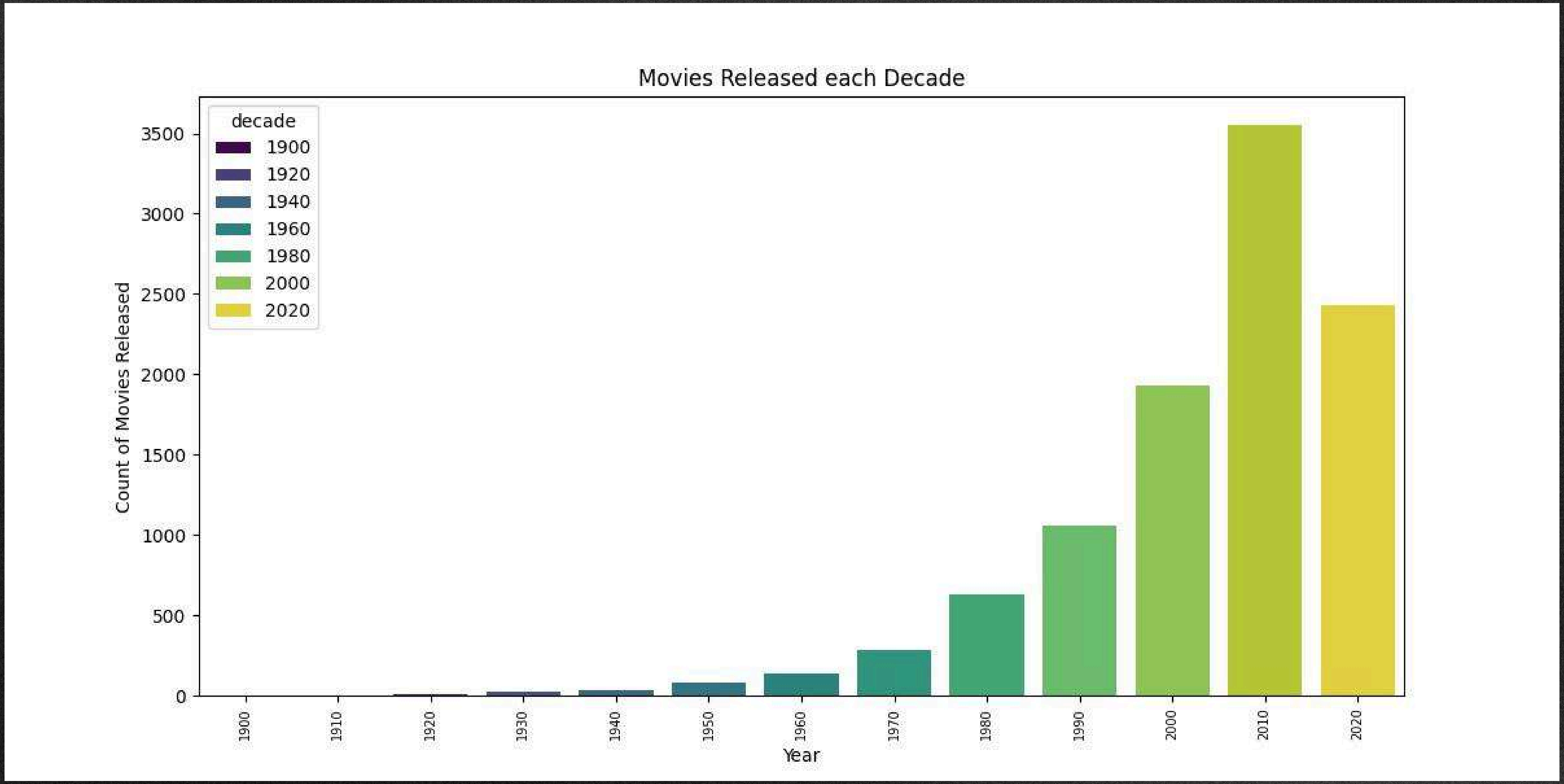
2022

<u>Year with the lowest number of movie releases</u>

1903

YEAR AND TREND ANALYSIS

Plot the number of movies released each decade.



2010 decade saw the most number of movie releases (3553).

INSIGHTS AND SUMMARY

1. Based on your analysis, what are three major insights you learned about movie trends, popular genres, or movie scores?

- Genre Popularity Over Time: Certain genres, such as action and adventure, have seen a consistent rise in popularity over the years, likely driven by advancements in special effects and global box office appeal. In contrast, genres like westerns or musicals have experienced a decline, possibly due to changes in audience preferences and cultural trends.
- Impact of Budget on Movie Scores: High-budget movies often perform better in terms of audience and critic ratings, as they can invest in better visual effects, renowned directors, and top-tier actors. However, there are exceptions, with some low-budget films (e.g., independent dramas or thrillers) achieving critical acclaim due to strong storytelling and innovative filmmaking.
- Seasonal Release Trends and Scores: Movies released during summer or the holiday season tend to have higher box office earnings and audience ratings. These periods are strategically chosen for blockbuster films that cater to family and mass audiences.

2. What additional questions could be explored with this dataset, or what other data would be helpful to gain a deeper understanding?

- How does a movie's revenue affect its score? Do more commercially successful movies tend to have better ratings?
- Is there a significant relationship between high IMDB scores and box office success?
- Are there differences in scores or revenue between movies produced in different countries?
- Which genres achieve the highest revenue-to-budget ratio, and what factors contribute to their success?

CONTACT ME

Email

ethendcosta5@gmail.com

LinkedIn

www.linkedin.com/in/ethendcosta

GitHub

https://github.com/EthenDcosta5

THANK YOU!