Movie Data Analysis

Table of contents

01 Members of Group

Members of group

22127148 **Dương Nhật Huy**	**22127224** **Trương Thuận Kiệt**
22127257 **Phạm Minh Mẫn**	**22127492** **Hồ Đăng Phúc**

02
Project Overview

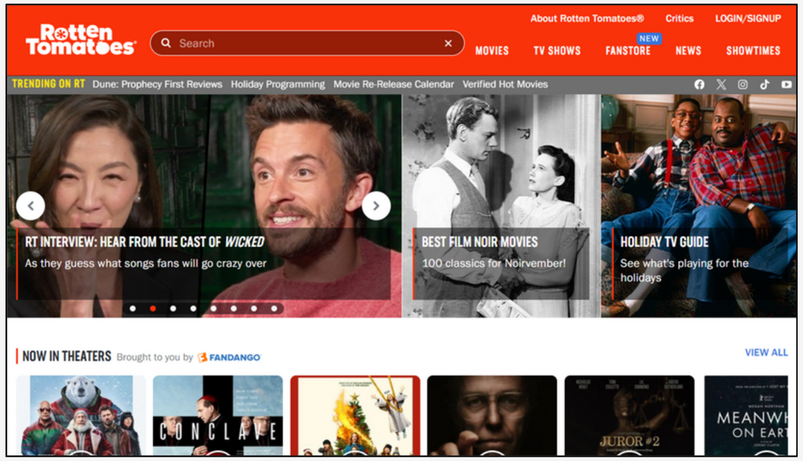
Reasons for Project

- To explore and analyze movie data to uncover insights into factors affecting movie performance, including ratings, revenue, and critical reception.
- To address missing data challenges and demonstrate effective preprocessing and modeling techniques.
- To enhance understanding of data cleaning, visualization, and modeling in a real-world dataset.

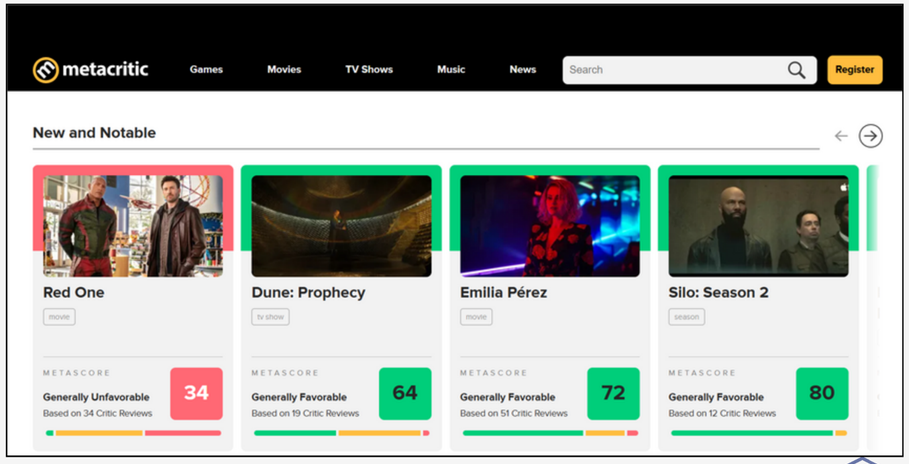
Data Sources

Data Rating Web

01
Rotten Tomatoes



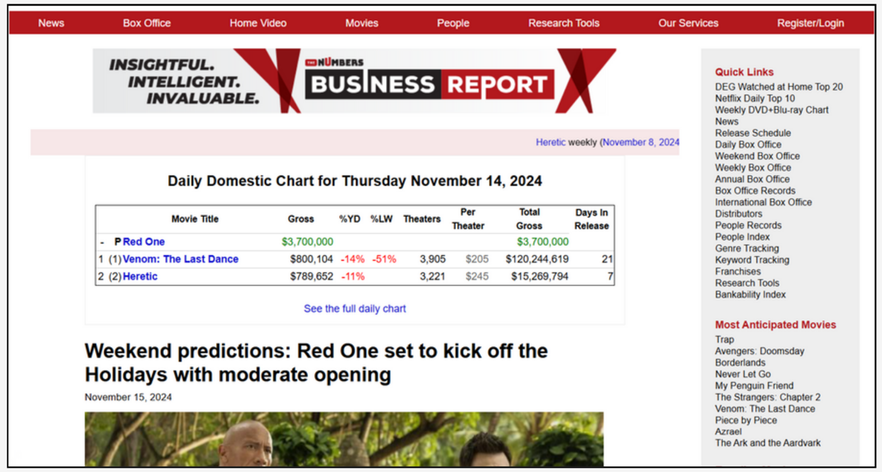
02
Meta Critic



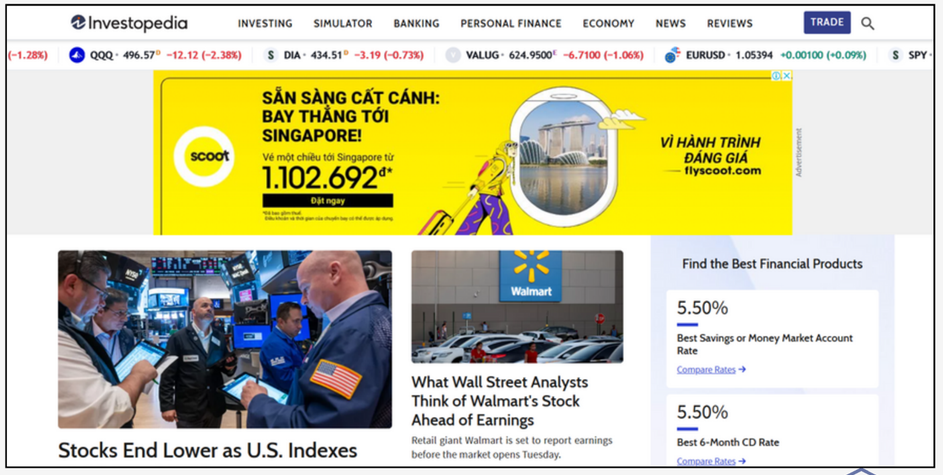
Data Sources

Budget/ Revenue and Inflation rate

01
The Numbers



02
Investopedia



03 Data overview

Data Overview

Data information

Size

Data type

The dataset contains approximately 5,106 movies with 17 features.

```
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Title                 5106 non-null   object
 1   Tomatoes CriticScore  4129 non-null   float64
 2   Tomatoes UserScore    4523 non-null   float64
 3   Link                  5106 non-null   object
 4   PlatformReleased      5106 non-null   object
 5   Cast                  5074 non-null   object
 6   Director              5071 non-null   object
 7   Genre                 5060 non-null   object
 8   Rating                4015 non-null   category
 9   Runtime               5036 non-null   object
 10  Studio                5059 non-null   object
 11  Release Date          5031 non-null   datetime64[ns]
 12  Production Budget     5106 non-null   float64
 13  Domestic Gross        5106 non-null   float64
 14  Worldwide Gross       5106 non-null   float64
 15  Metascore             4353 non-null   float64
 16  Meta UserScore        4186 non-null   float64
dtypes: category(1), datetime64[ns](1), float64(7), object(8)
```

The dataset fully contains the general features of each movie

Key features

Title

Movie name

Critic/User Scores

Ratings from critics and users

Financials

Production budgets, domestic and worldwide gross revenue

Categoricals

Genre, platform, studio, cast, director

Time Information

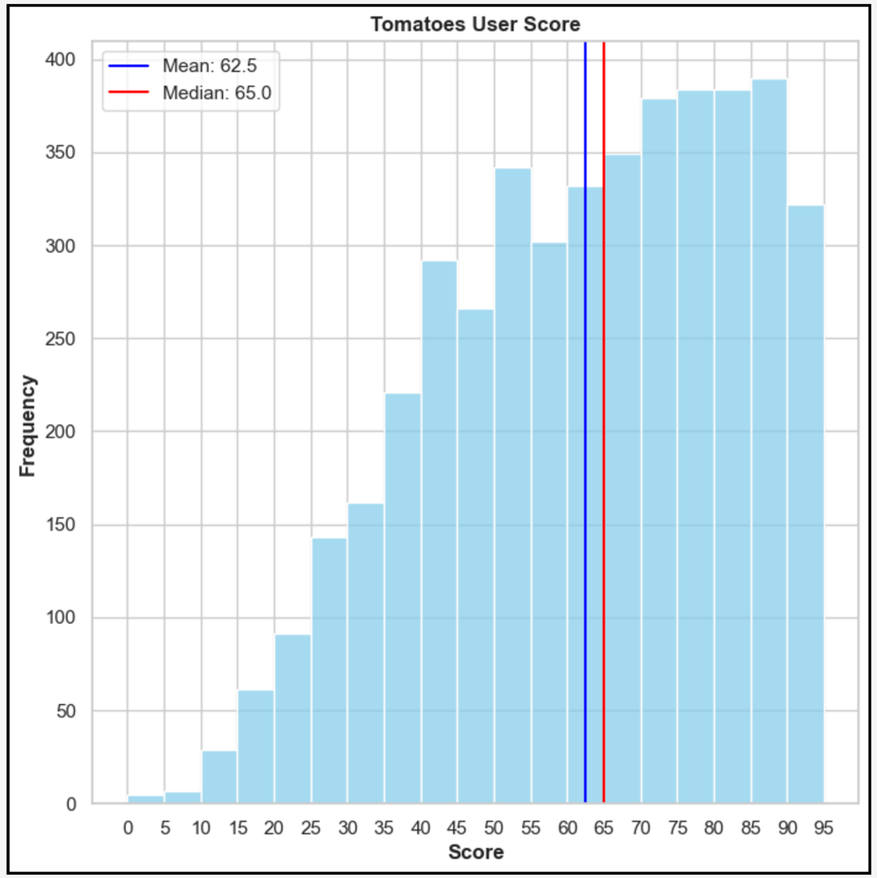
Release date, runtime

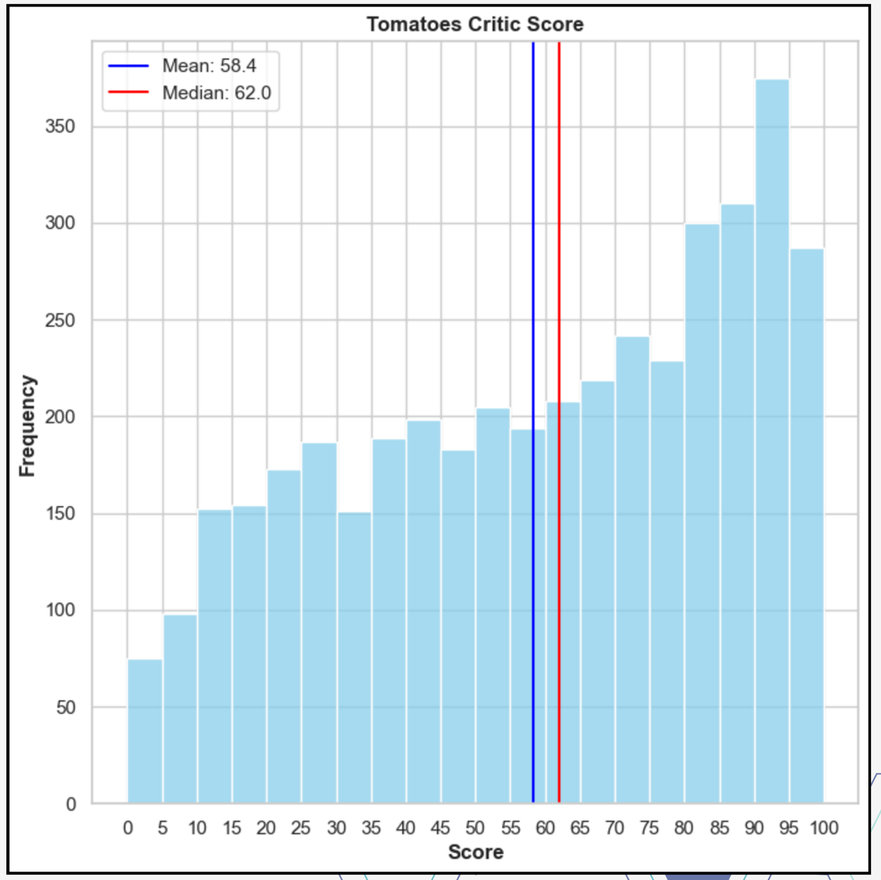
Numerical Columns

Rotten Tomatoes Scores

All the scores are out of 100

- Tomatoes UserScore: Scores from users/ viewers from Rotten Tomatoes
- Tomatoes CriticScore: Scores from critics from Rotten Tomatoes



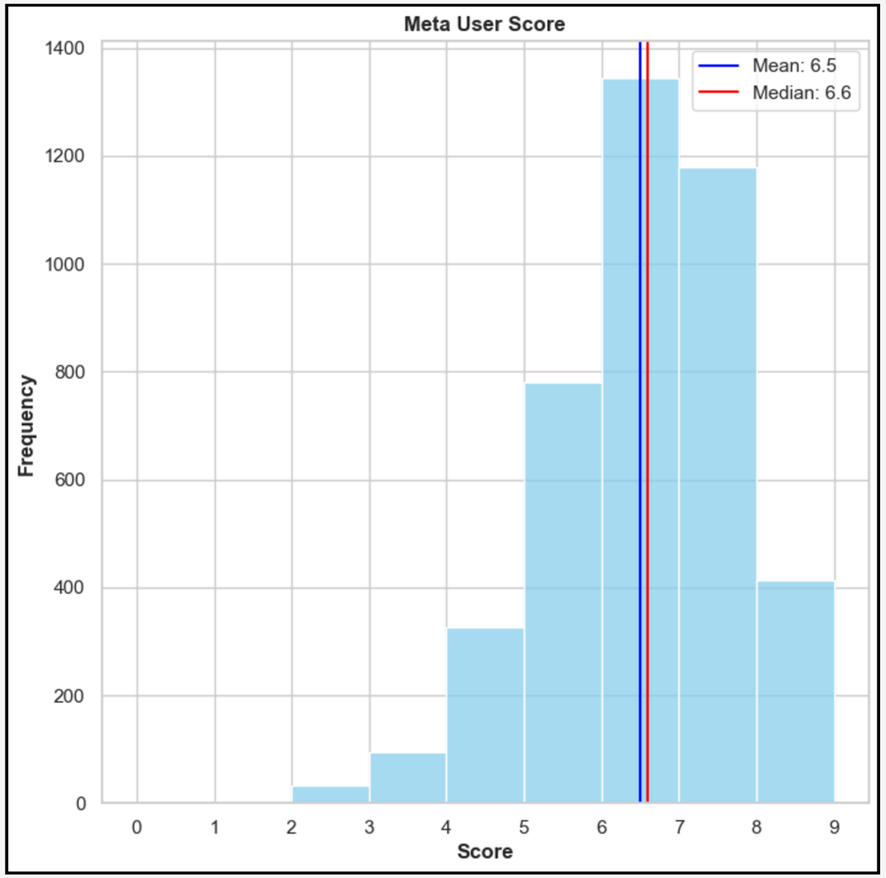


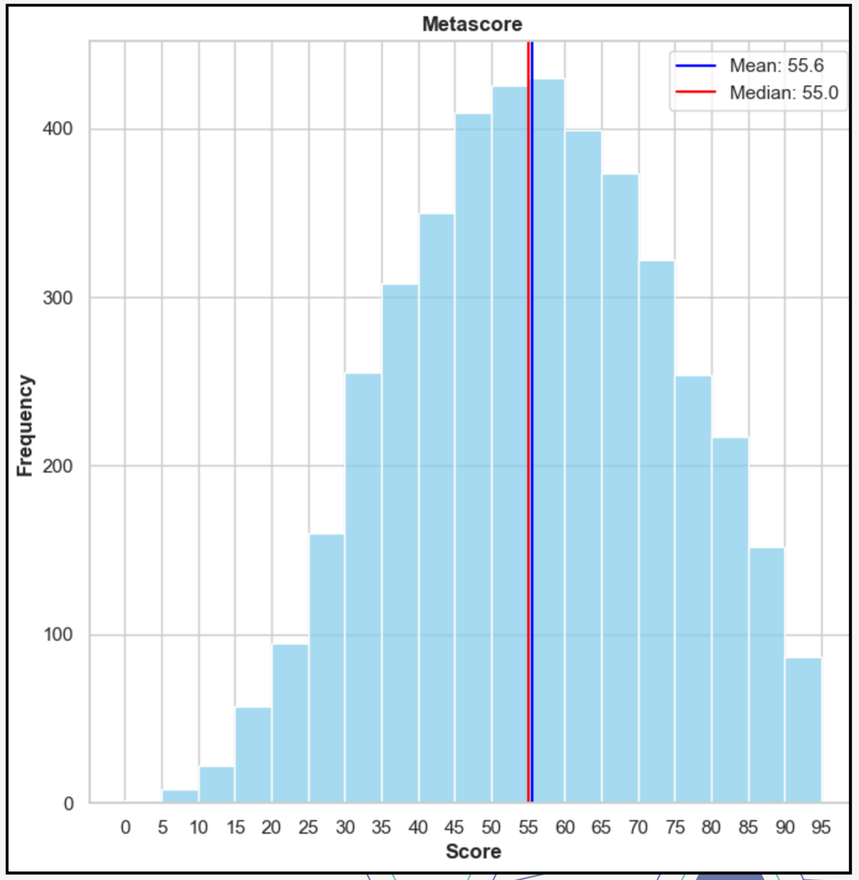
Numerical Columns

MetaCritic Scores

MetaScore is out of 100, except for Meta User Score is out of 10

- Meta User Score: Scores from users/ viewers from MetaCritic
- MetaScore: Scores from critics from MetaCritic



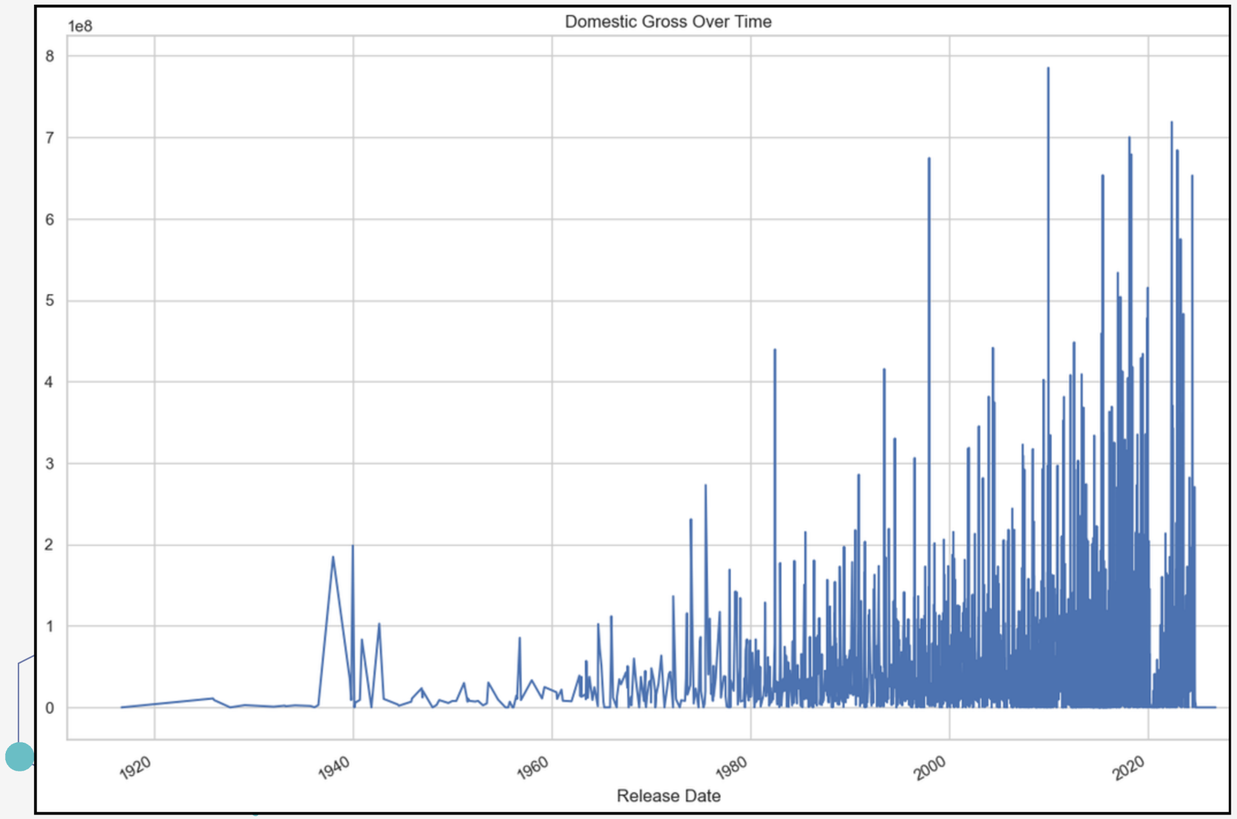


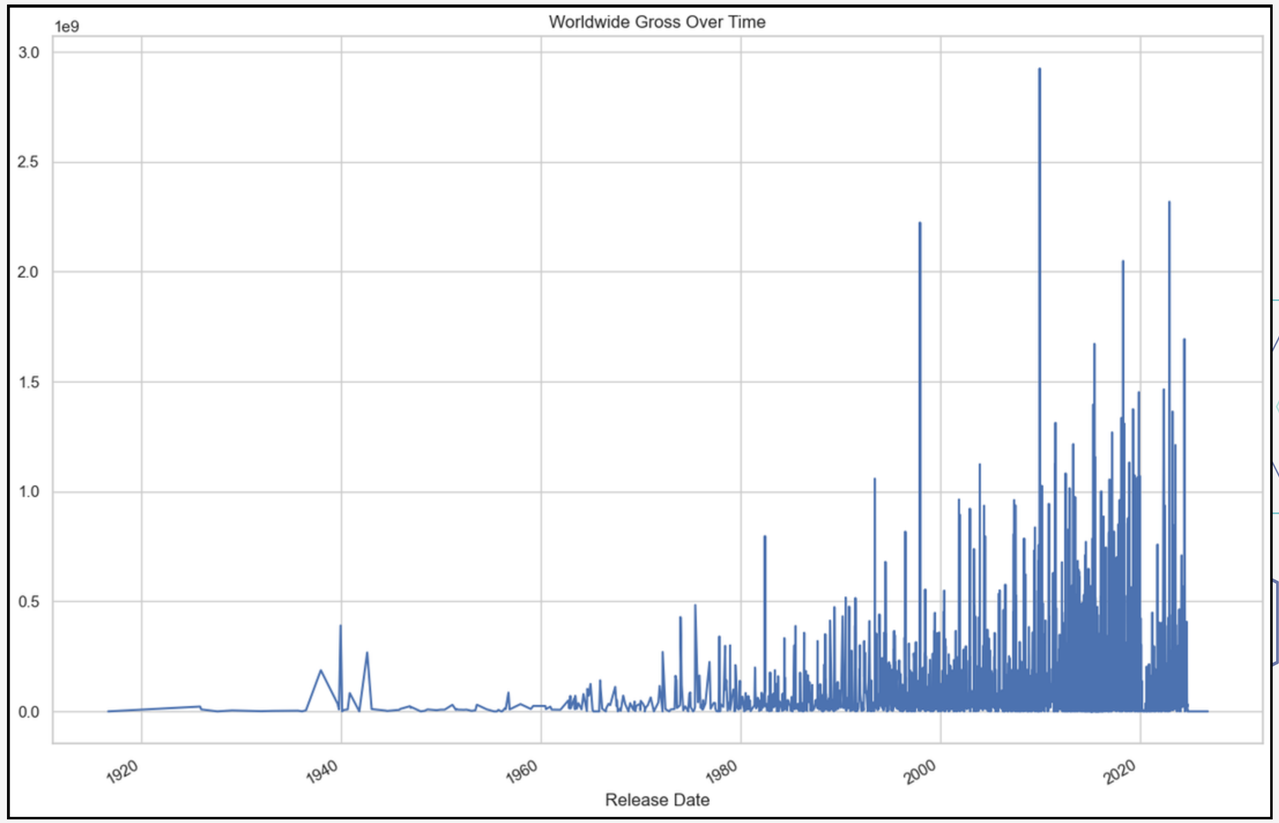
Numerical Columns

Economic Features

All the factors are calculated in billions

- Domestic/Worldwide Gross: Revenues from national/international market
- Production Budget: The expenditure spent to produce movies



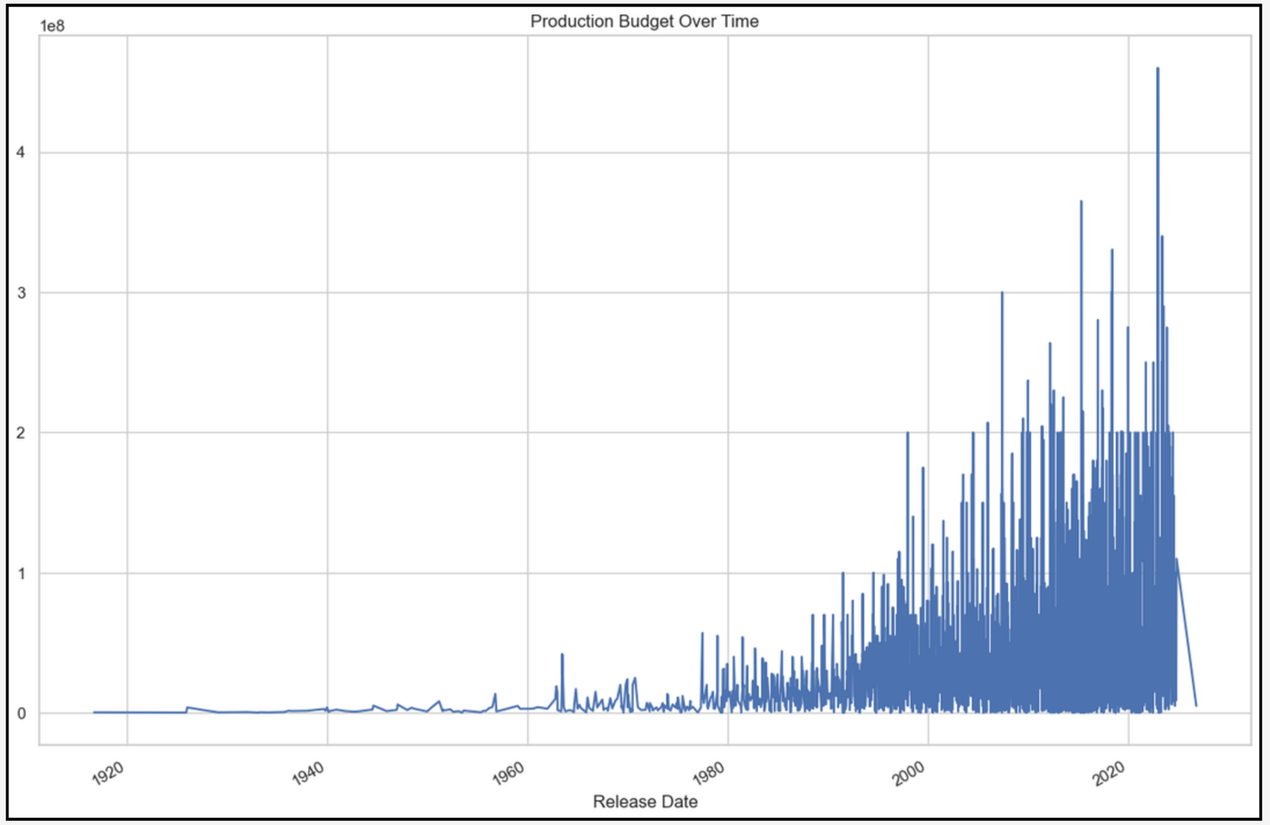


Numerical Columns

Economic Features

All the factors are calculated in billions

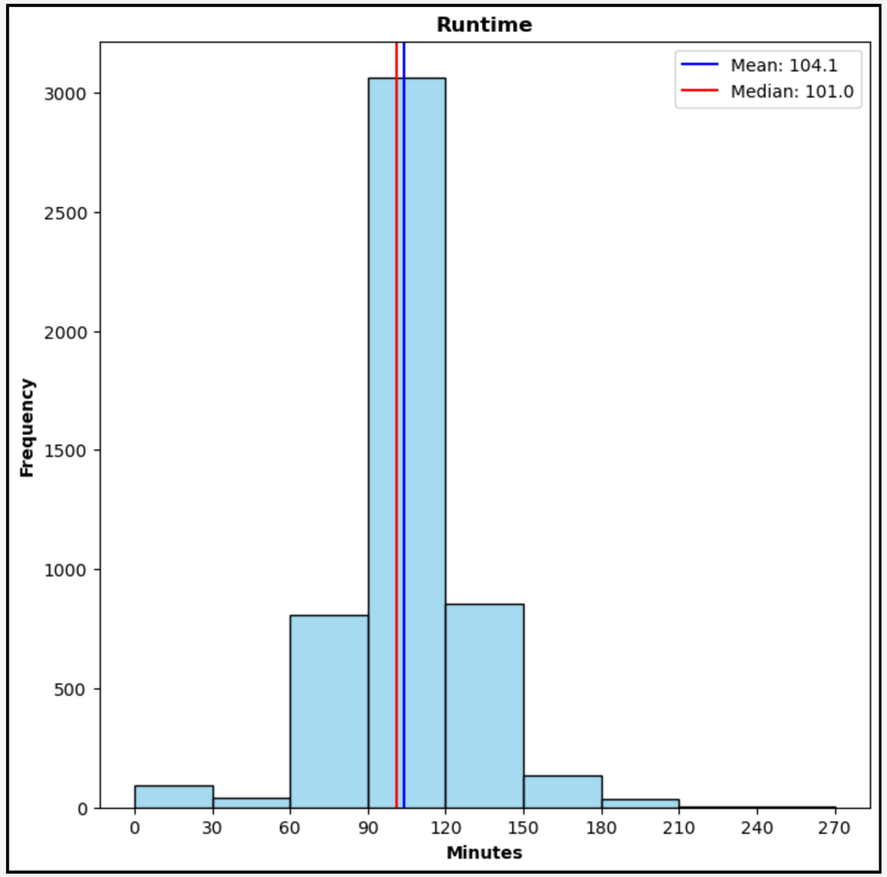
- Domestic/Worldwide Gross: Revenues from national/international market
- Production Budget: The expenditure spent to produce movies



Numerical Columns

Runtime

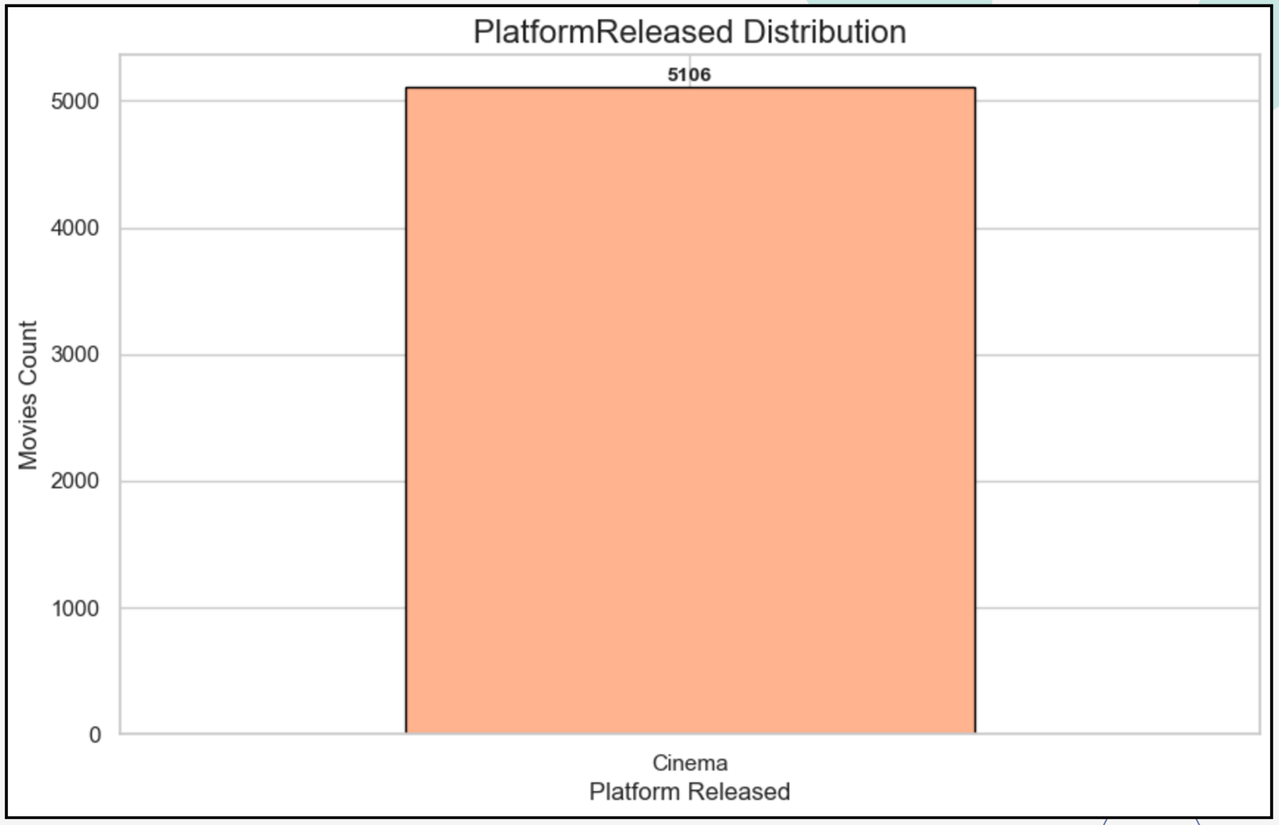
Time duration of each movie



Categorical Columns

Platform Released

Places that movies are released like Netflix, Cinema, etc

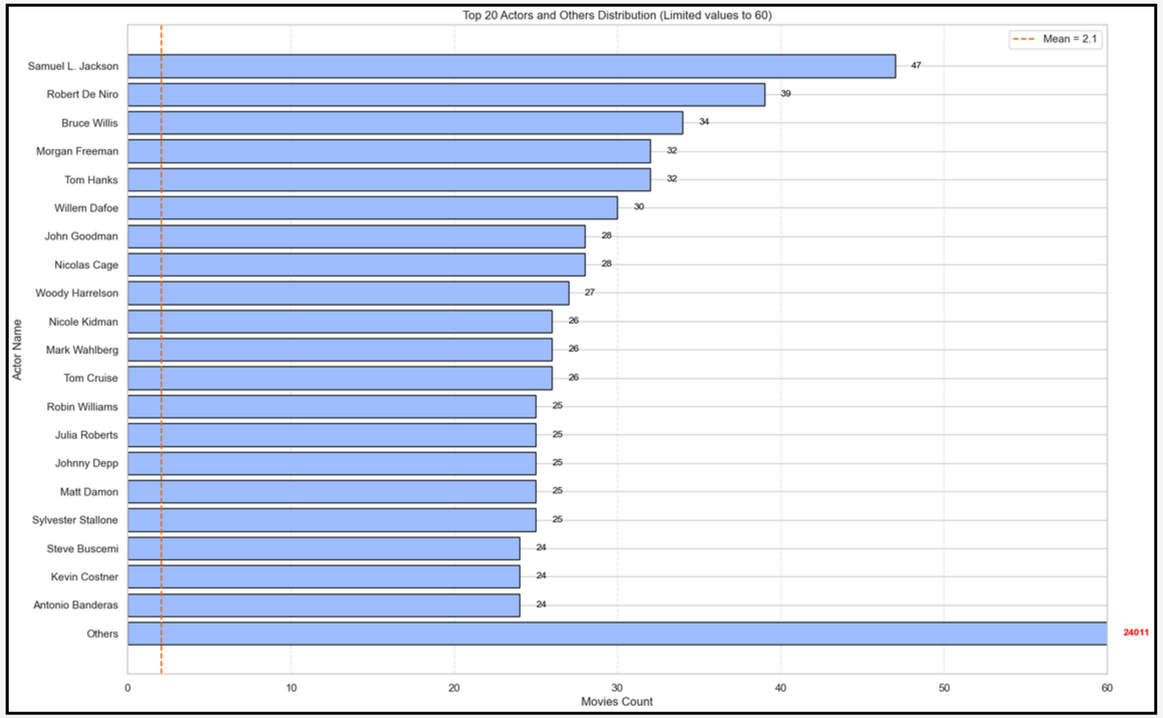


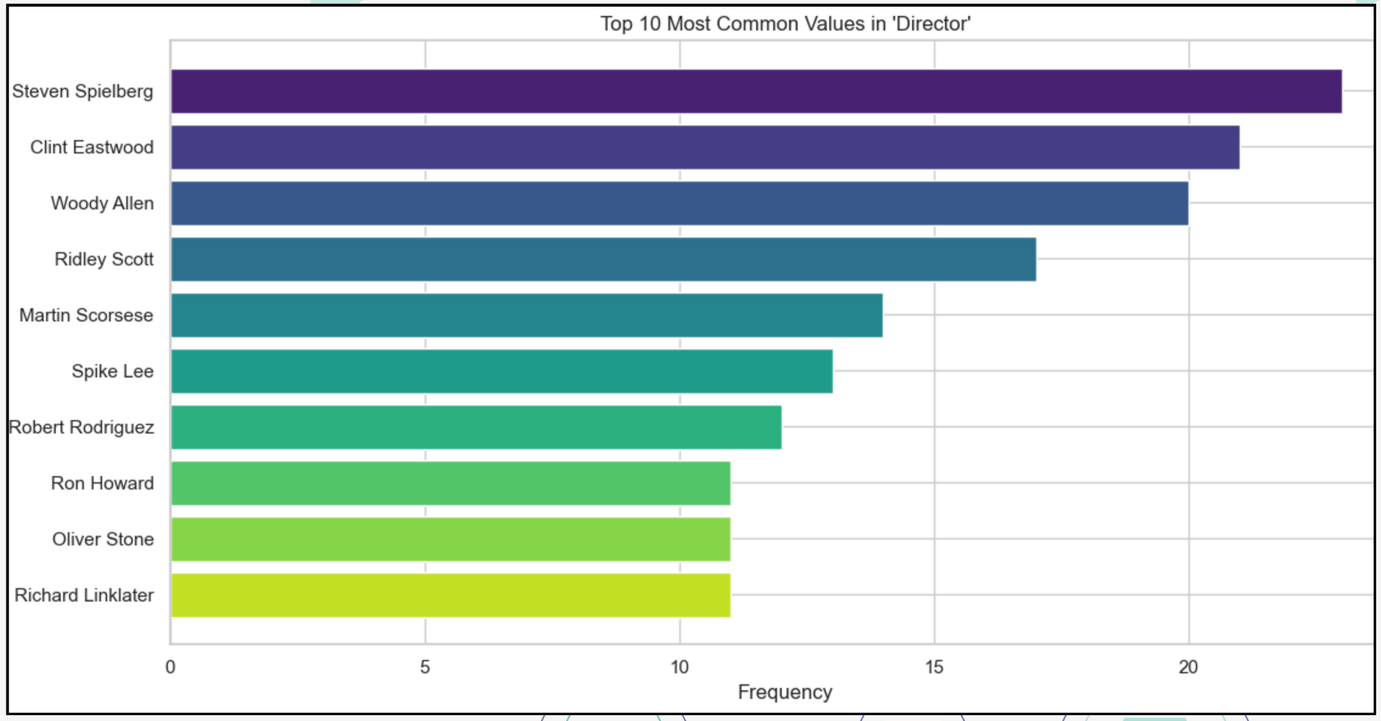
Categorical Columns

Casts, Directors and Studios

All factors that contribute to creating a movie

- Casts: Actors/Actresses that played in a movie
- Directors: Who have control over movie production
- Studios: Companies that are responsible for releasing movies



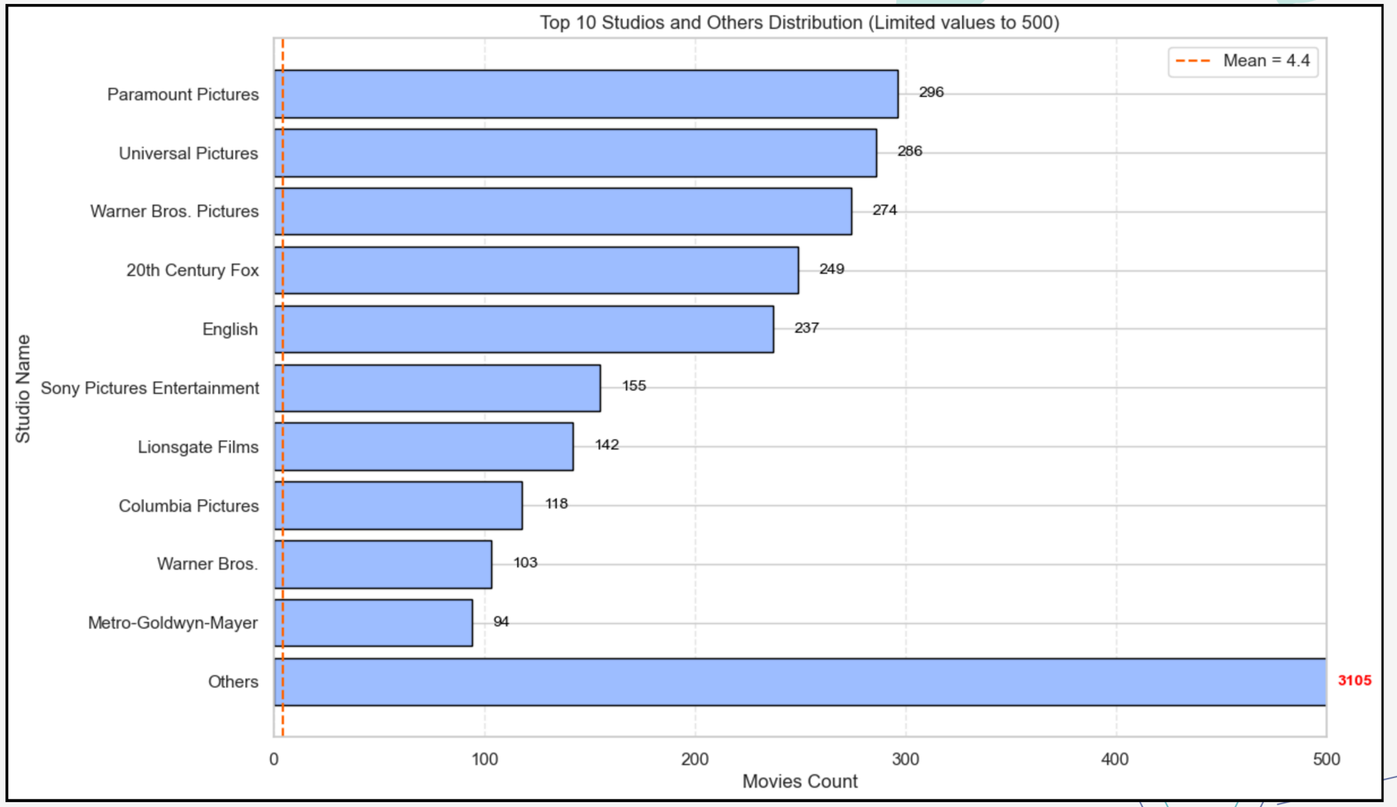


Categorical Columns

Casts, Directors and Studios

All factors that contribute to creating a movie

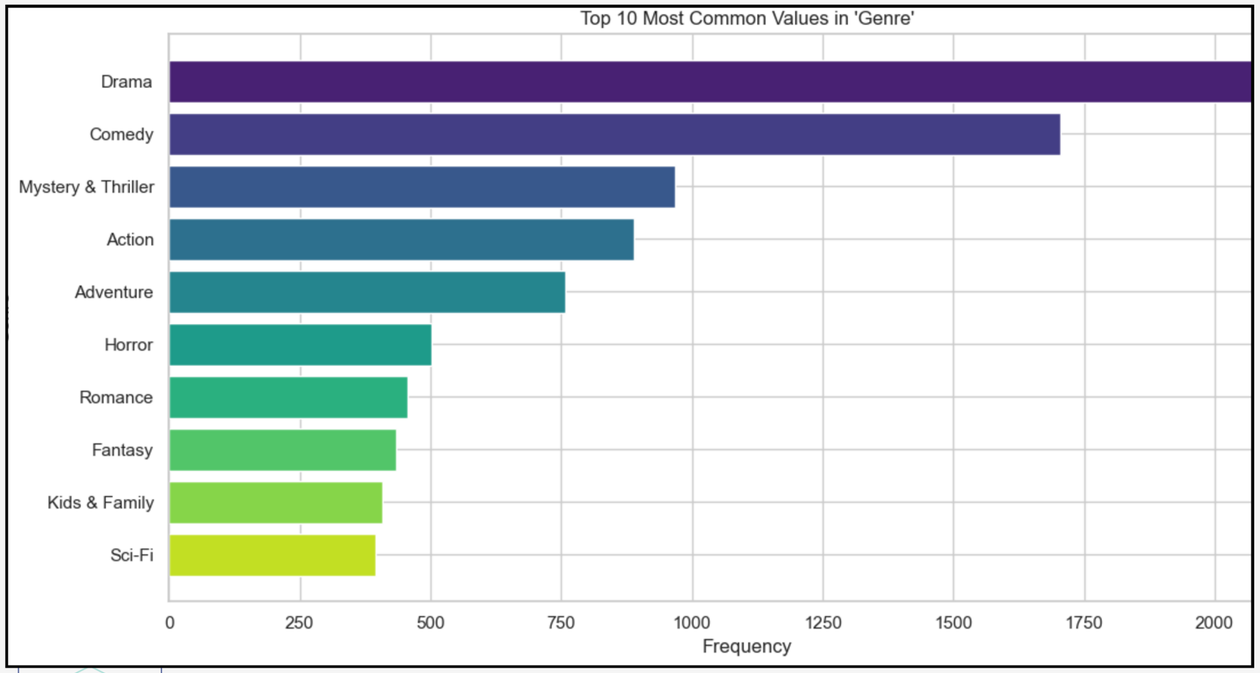
- Casts: Actors/Actresses that played in a movie
- Directors: Who have control over movie production
- Studios: Companies that are responsible for releasing movies

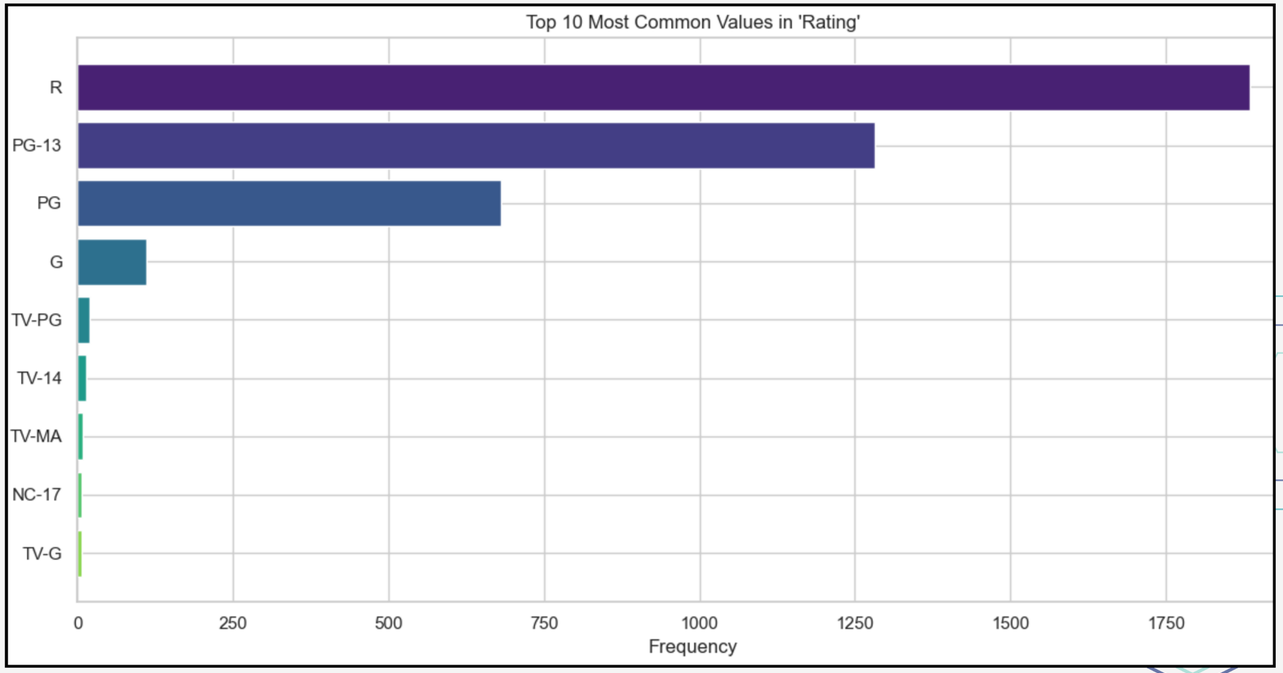


Categorical Columns

Genres and Ratings

- Genres: A category of artistic composition that creates a movie
- Ratings: Classifies films based on their suitability for audiences

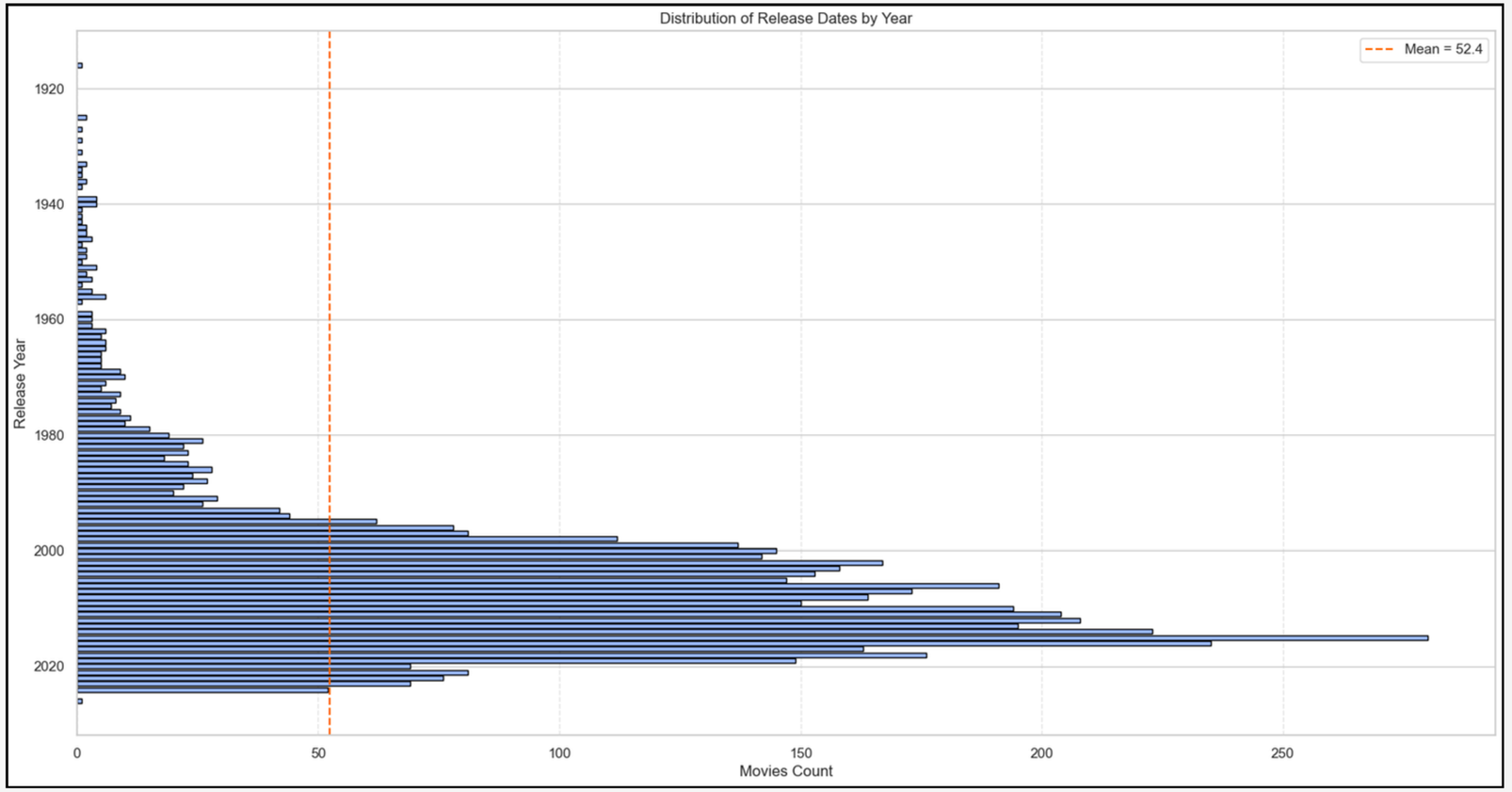




Categorical Columns

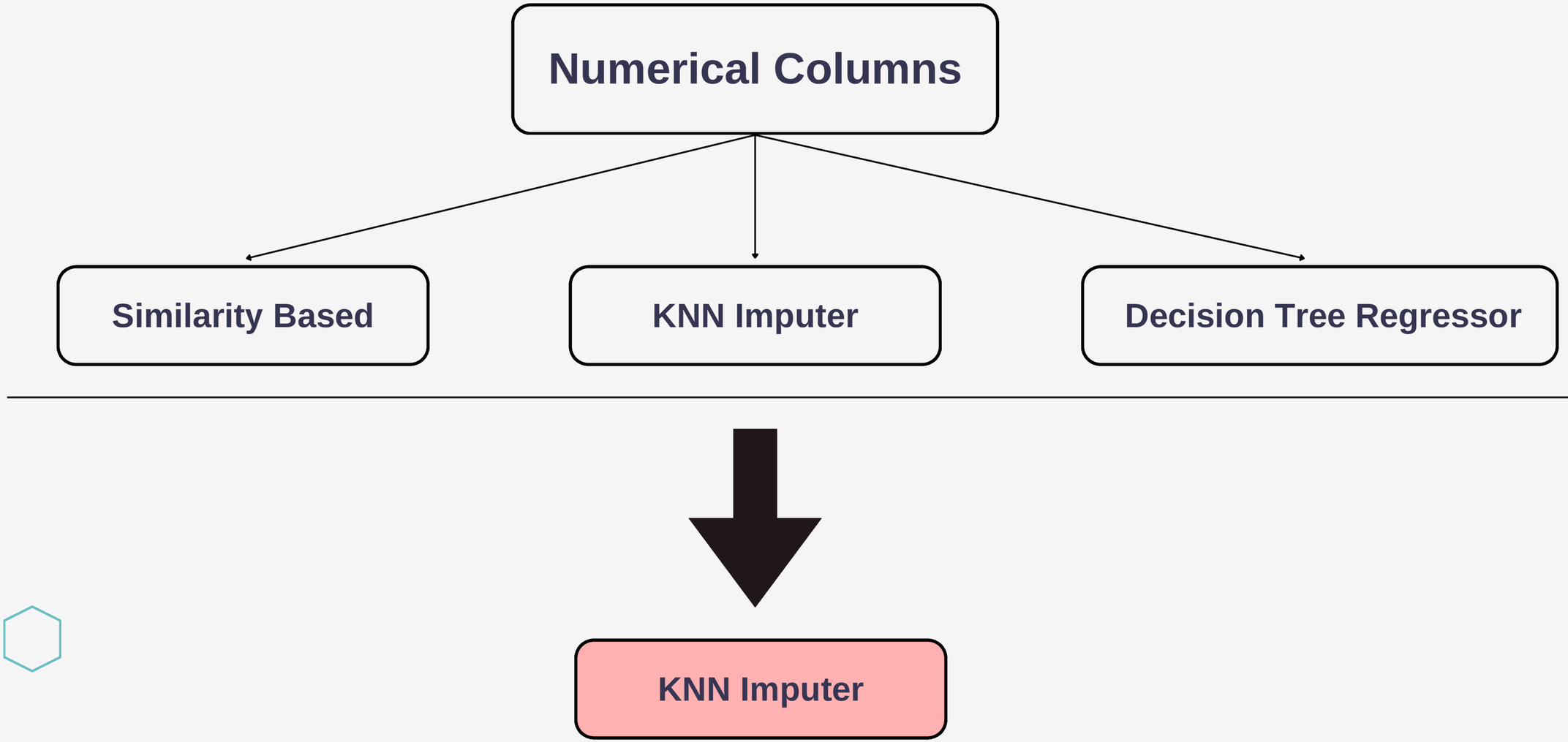
Release Dates

Date that a movie is released

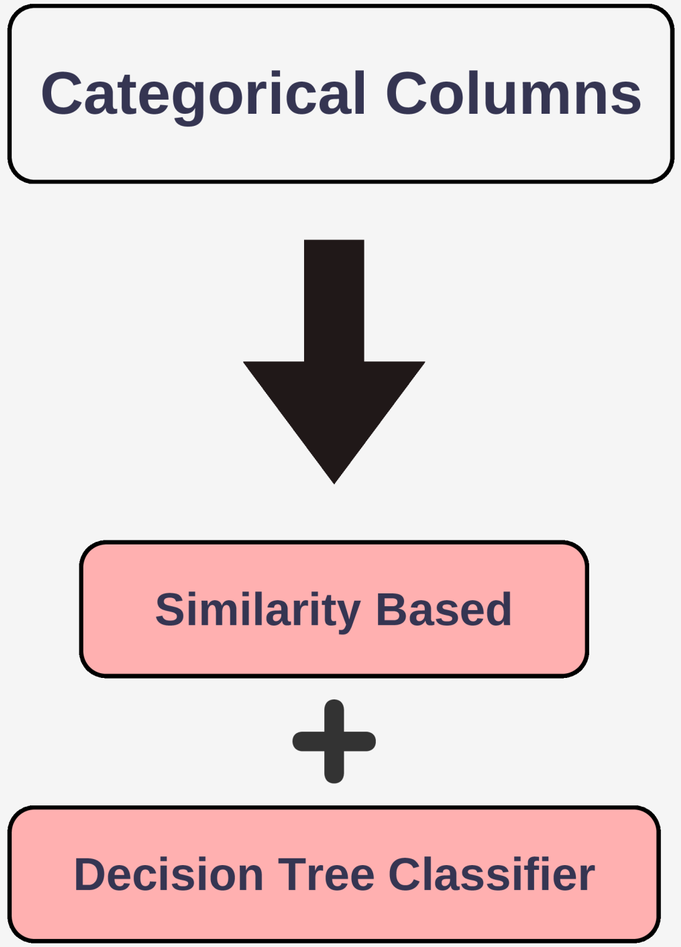


04
Data Preprocessing

Handle Missing Values



Handle Missing Values

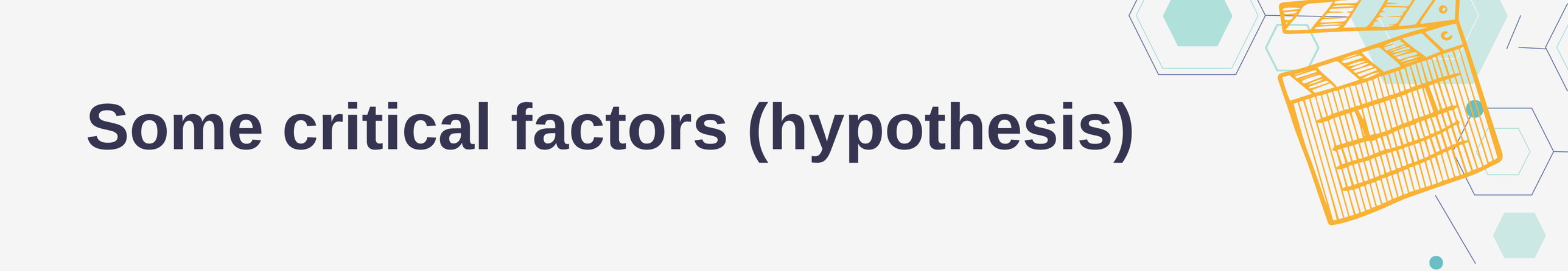


05
Questions

Question 1:

What are the primary factors influencing critic and user scores?

Some critical factors (hypothesis)



- Genre
- Production Budget
- Release Time
- Marketing & Promotion

Preprocessing

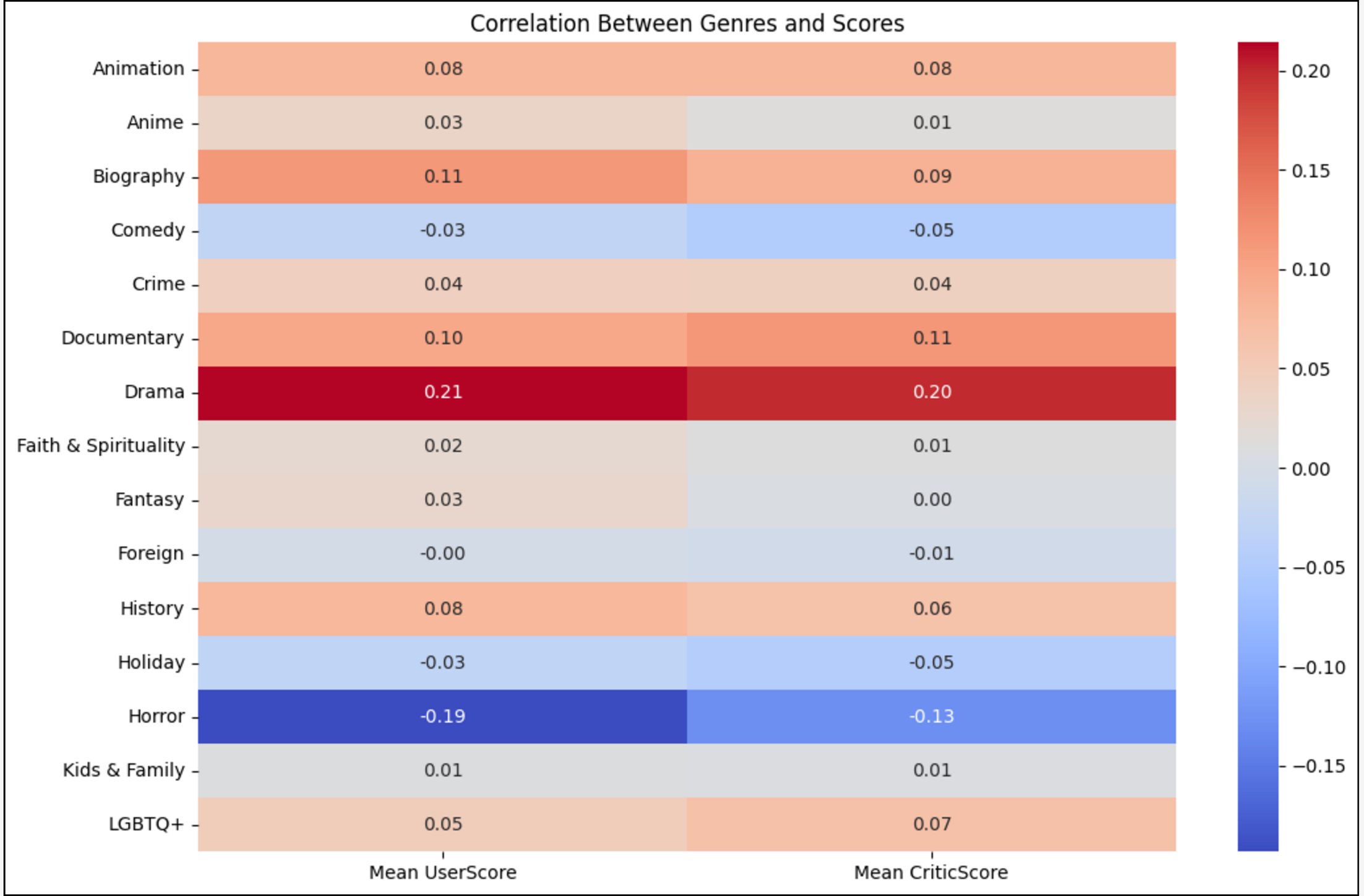
One-hot encode the Genre column

Split release date into year, month and day

Calculate the mean score for each movie

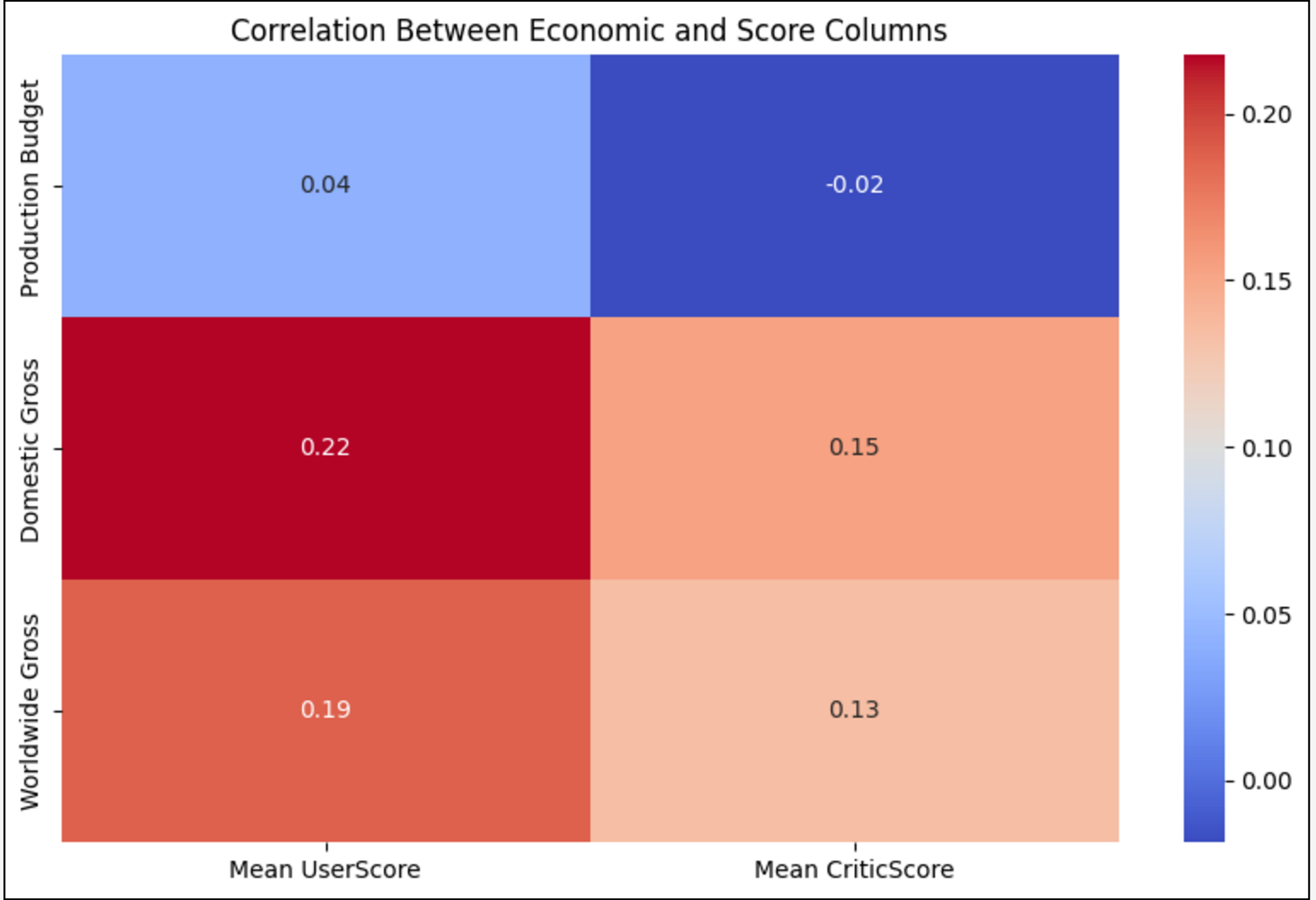
One-hot encode the Rating column

Drop columns that are not needed


Correlation Between Genres and Scores
Animation
Anime
Biography
Comedy
Crime
Documentary
Drama
Faith & Spirituality
Fantasy
Foreign
History
Holiday
Horror
Kids & Family
LGBTQ+
0.08
0.08
0.03
0.01
0.11
0.09
-0.03
-0.05
0.04
0.04
0.10
0.11
0.21
0.20
0.02
0.01
0.03
0.00
-0.00
-0.01
0.08
0.06
-0.03
-0.05
-0.19
-0.13
0.01
0.01
0.05
0.07
Mean UserScore
Mean CriticScore
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15

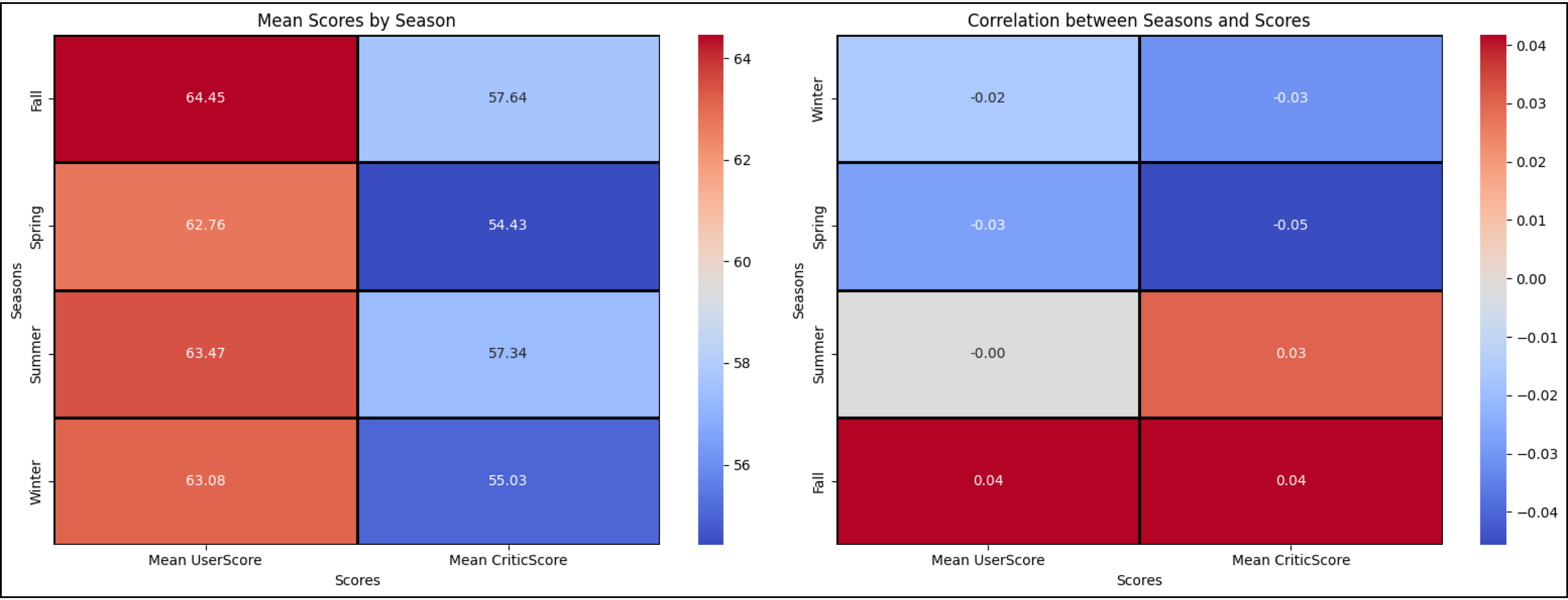
Key Observations

- **Strong Positive Correlations**:
  - Drama
  - Documentary
- **Weak or Negative Correlations**:
  - Horror
  - Comedy
- **Mixed Correlations:**
  - Animation
  - Biography


Correlation Between Economic and Score Columns
Production Budget
Domestic Gross
Worldwide Gross
0.04
-0.02
0.22
0.15
0.19
0.13
Mean UserScore
Mean CriticScore
0.20
0.15
0.10
0.05
0.00

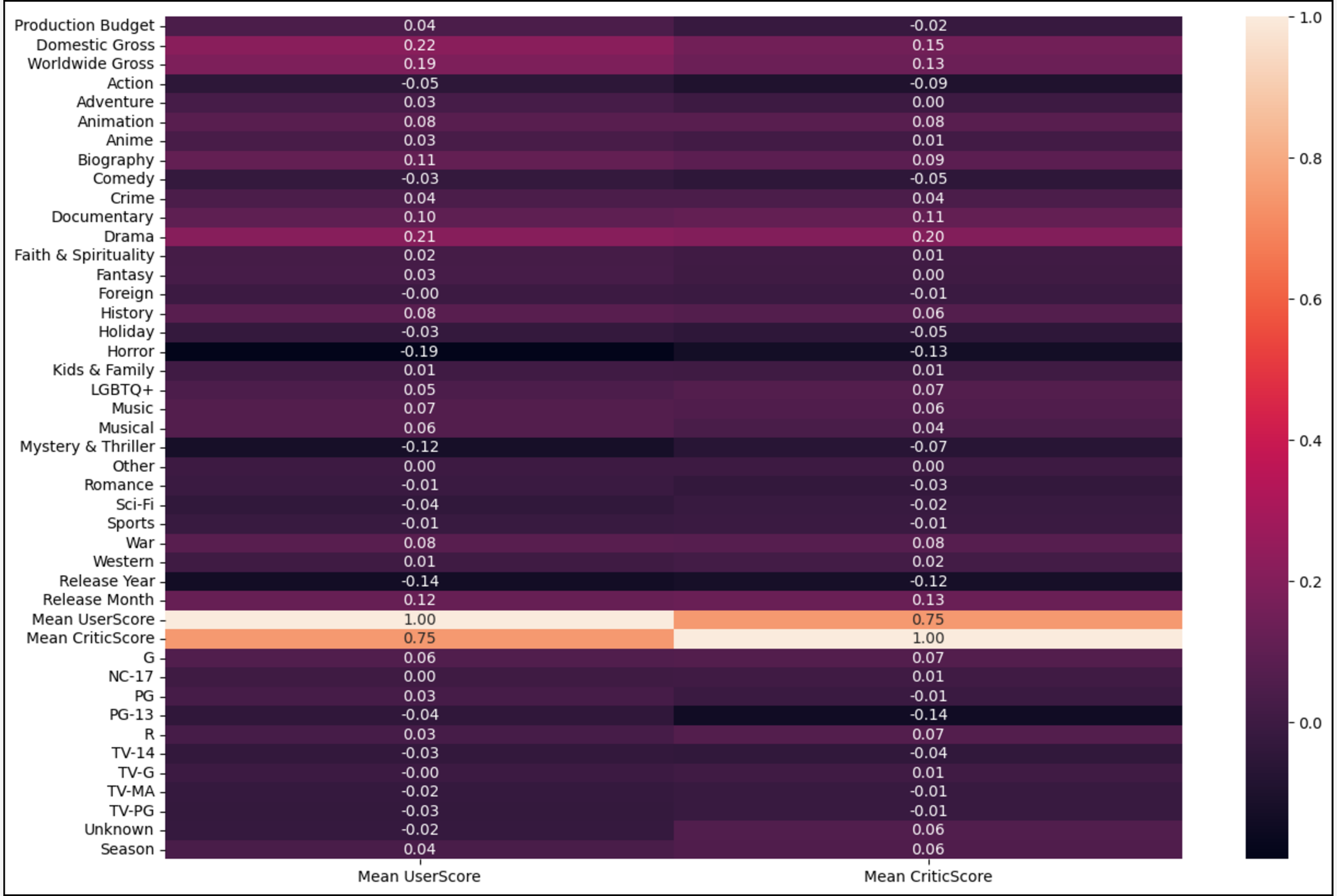
Key Observations

- **Summary**:
  - **Production Budget**: Weak correlation with ratings.
  - **Domestic and Worldwide Gross**: Strong positive correlation with ratings
- **Inferences**:
  - **Box Office and Acclaim**: Commercial success often aligns with critical and audience approval.
  - **Budget vs. Quality**: A large budget alone is not a reliable predictor of a film's quality or success.


Mean Scores by Season
Fall
Spring
Summer
Winter
Seasons
64.45
57.64
62.76
54.43
63.47
57.34
63.08
55.03
Mean UserScore
Mean CriticScore
Scores
64
62
60
58
56
Correlation between Seasons and Scores
Winter
Spring
Summer
Fall
Seasons
-0.02
-0.03
-0.03
-0.05
-0.00
0.03
0.04
0.04
Mean UserScore
Mean CriticScore
Scores
0.04
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.04

Key Observations

- **Mean Scores by Season:**
  - **Winter**: Generally sees higher scores from both users and critics.
  - **Spring**: Lower scores compared to other seasons.
  - **Summer and Fall**: Moderate level of scores.
- **Correlation Between Seasons and Scores**
  - **Weak Correlations:** The correlations are generally weak.
  - **Slight Variations:** There are subtle variations in the correlations for user and critic scores across seasons.


Production Budget 0.04 -0.02
Domestic Gross 0.22 0.15
Worldwide Gross 0.19 0.13
Action -0.05 -0.09
Adventure 0.03 0.00
Animation 0.08 0.08
Anime 0.03 0.01
Biography 0.11 0.09
Comedy -0.03 -0.05
Crime 0.04 0.04
Documentary 0.10 0.11
Drama 0.21 0.20
Faith & Spirituality 0.02 0.01
Fantasy 0.03 0.00
Foreign -0.00 -0.01
History 0.08 0.06
Holiday -0.03 -0.05
Horror -0.19 -0.13
Kids & Family 0.01 0.01
LGBTQ+ 0.05 0.07
Music 0.07 0.06
Musical 0.06 0.04
Mystery & Thriller -0.12 -0.07
Other 0.00 0.00
Romance -0.01 -0.03
Sci-Fi -0.04 -0.02
Sports -0.01 -0.01
War 0.08 0.08
Western 0.01 0.02
Release Year -0.14 -0.12
Release Month 0.12 0.13
Mean UserScore 1.00 0.75
Mean CriticScore 0.75 1.00
G 0.06 0.07
NC-17 0.00 0.01
PG 0.03 -0.01
PG-13 -0.04 -0.14
R 0.03 0.07
TV-14 -0.03 -0.04
TV-G -0.00 0.01
TV-MA -0.02 -0.01
TV-PG -0.03 -0.01
Unknown -0.02 0.06
Season 0.04 0.06
Mean UserScore
Mean CriticScore
1.0
0.8
0.6
0.4
0.2
0.0

Key Observations

- **Summary**:
  - **Consistent Perception**: Movies that perform well with critics tend to be appreciated by users too.
  - **Mutual Influences**: High critic scores may influence user expectations, or user enthusiasm might validate critic assessments.

Conclusion

- Domestic and Worldwide Gross
- Seasonality
- Genre Influences
- Ratings
- Production Budget
- Marketing Indicators

- Strategic release timing
- Effective marketing

Question 2:

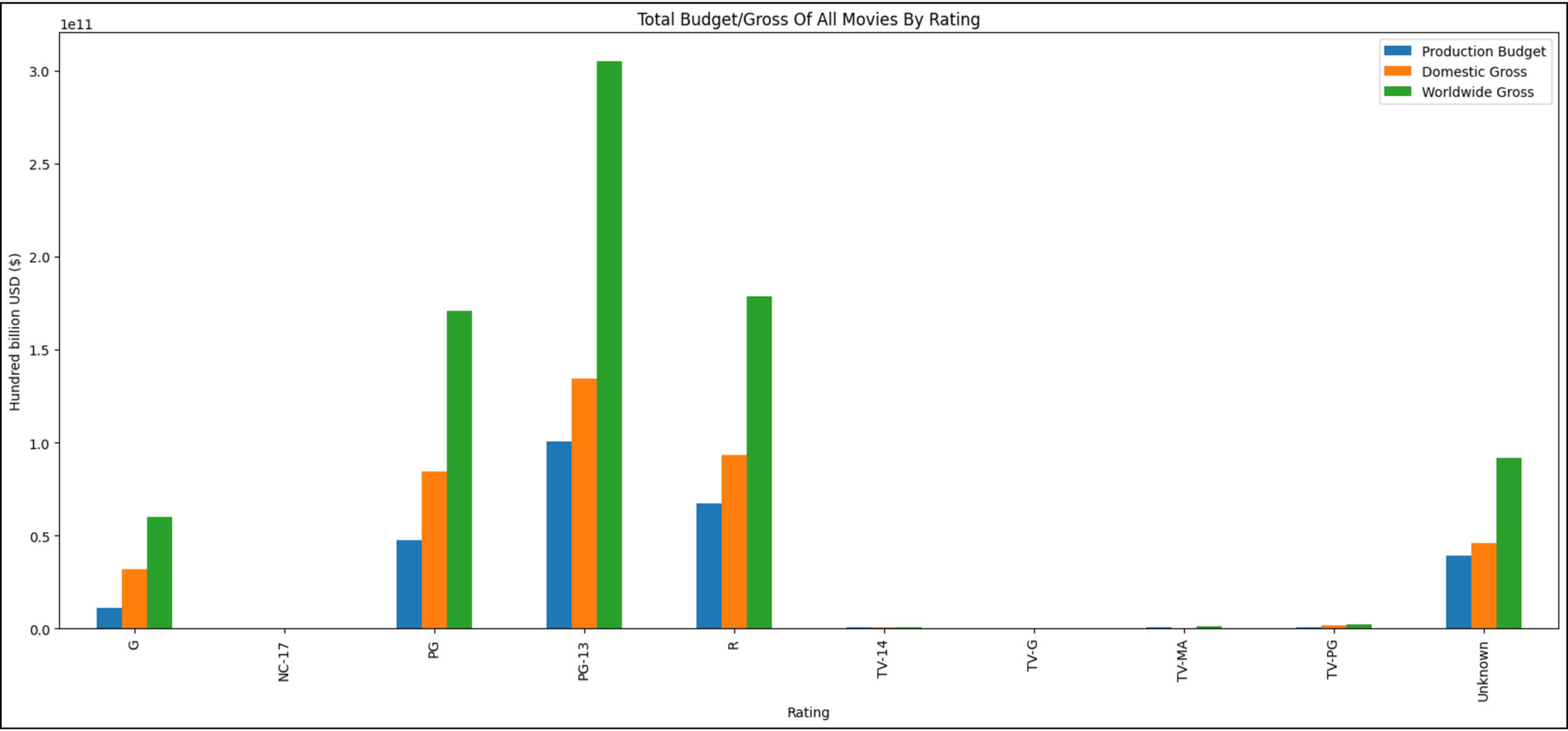
How do financial metrics like budget and gross revenue correlate with ratings?

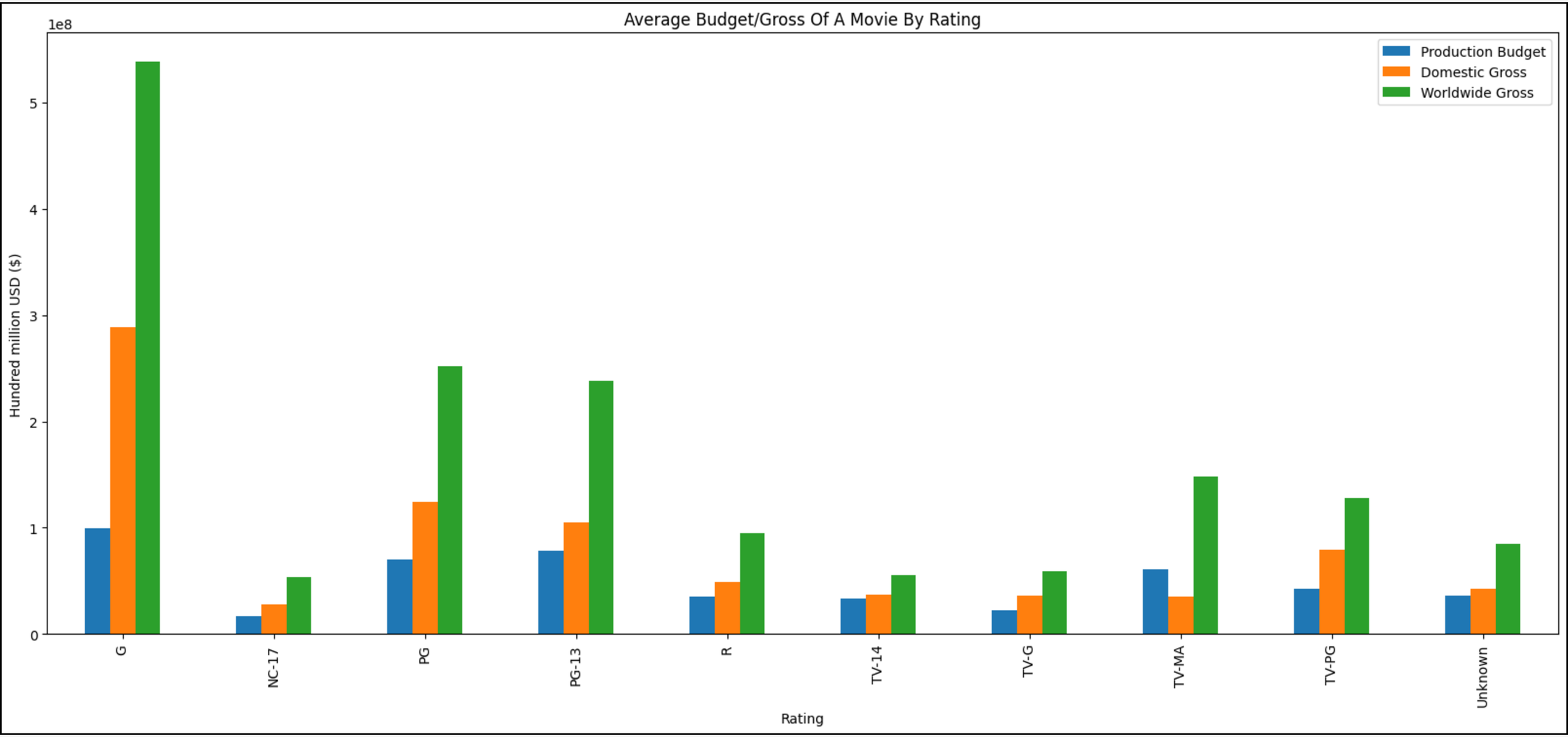
Preprocessing

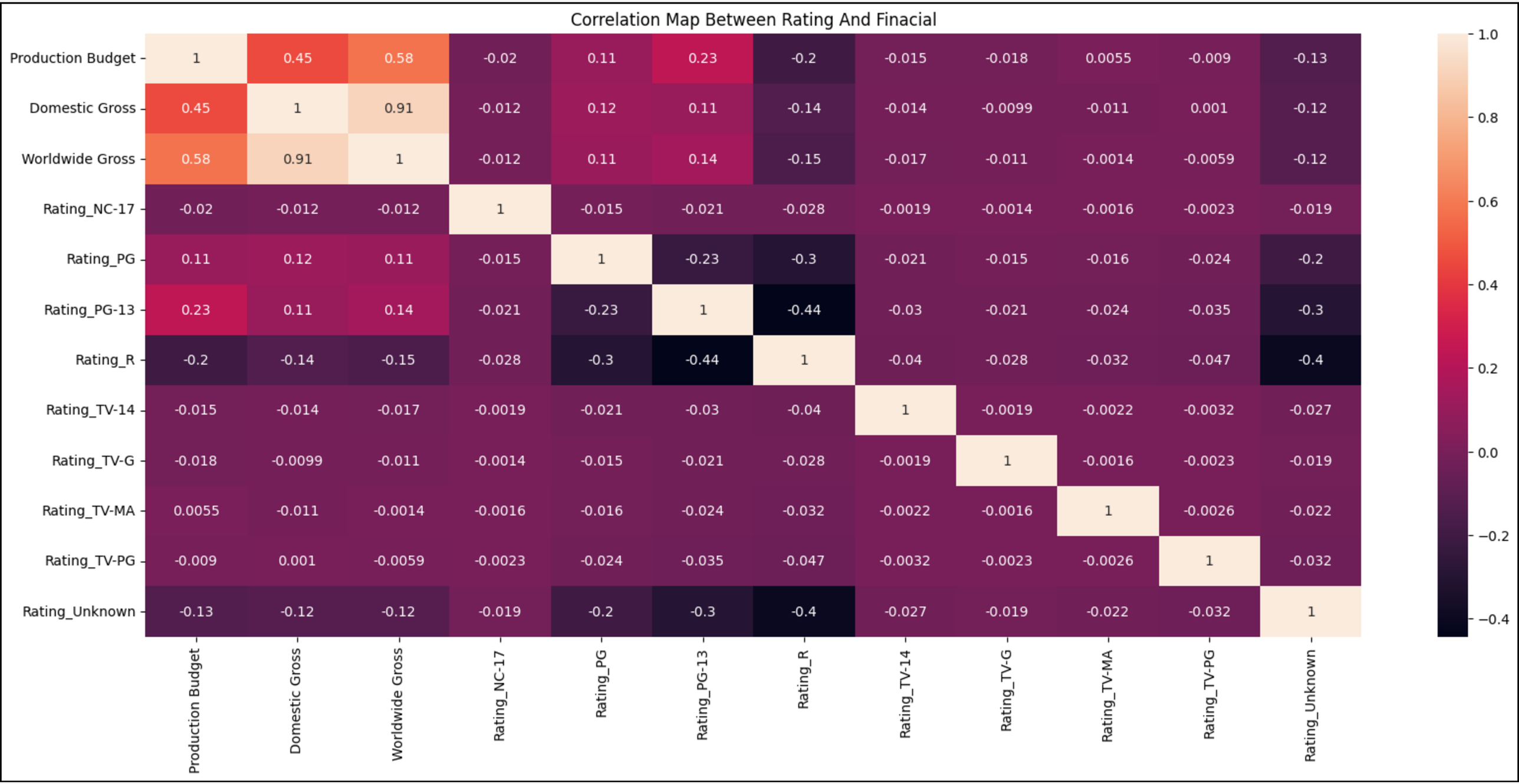
Remove unnecessary columns

Get release year and unit for each year

Apply inflation


Total Budget/Gross Of All Movies By Rating
1e11
Hundred billion USD ($)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Production Budget
Domestic Gross
Worldwide Gross
G
NC-17
PG
PG-13
R
TV-14
TV-G
TV-MA
TV-PG
Unknown
Rating


Average Budget/Gross Of A Movie By Rating
1e8
Production Budget
Domestic Gross
Worldwide Gross
Hundred million USD ($)
0
1
2
3
4
5
G
NC-17
PG
PG-13
R
TV-14
TV-G
TV-MA
TV-PG
Unknown
Rating


Correlation Map Between Rating And Finacial
Production Budget 1 0.45 0.58 -0.02 0.11 0.23 -0.2 -0.015 -0.018 0.0055 -0.009 -0.13
Domestic Gross 0.45 1 0.91 -0.012 0.12 0.11 -0.14 -0.014 -0.0099 -0.011 0.001 -0.12
Worldwide Gross 0.58 0.91 1 -0.012 0.11 0.14 -0.15 -0.017 -0.011 -0.0014 -0.0059 -0.12
Rating_NC-17 -0.02 -0.012 -0.012 1 -0.015 -0.021 -0.028 -0.0019 -0.0014 -0.0016 -0.0023 -0.019
Rating_PG 0.11 0.12 0.11 -0.015 1 -0.23 -0.3 -0.021 -0.015 -0.016 -0.024 -0.2
Rating_PG-13 0.23 0.11 0.14 -0.021 -0.23 1 -0.44 -0.03 -0.021 -0.024 -0.035 -0.3
Rating_R -0.2 -0.14 -0.15 -0.028 -0.3 -0.44 1 -0.04 -0.028 -0.032 -0.047 -0.4
Rating_TV-14 -0.015 -0.014 -0.017 -0.0019 -0.021 -0.03 -0.04 1 -0.0019 -0.0022 -0.0032 -0.027
Rating_TV-G -0.018 -0.0099 -0.011 -0.0014 -0.015 -0.021 -0.028 -0.0019 1 -0.0016 -0.0023 -0.019
Rating_TV-MA 0.0055 -0.011 -0.0014 -0.0016 -0.016 -0.024 -0.032 -0.0022 -0.0016 1 -0.0026 -0.022
Rating_TV-PG -0.009 0.001 -0.0059 -0.0023 -0.024 -0.035 -0.047 -0.0032 -0.0023 -0.0026 1 -0.032
Rating_Unknown -0.13 -0.12 -0.12 -0.019 -0.2 -0.3 -0.4 -0.027 -0.019 -0.022 -0.032 1
Production Budget
Domestic Gross
Worldwide Gross
Rating_NC-17
Rating_PG
Rating_PG-13
Rating_R
Rating_TV-14
Rating_TV-G
Rating_TV-MA
Rating_TV-PG
Rating_Unknown
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Key Observations

- **PG-13 Movies:**
  - Highest budget and gross.
  - Moderate budget-gross correlation.
  - Broad appeal and global success.
- **R & PG Movies:**
  - Strong budget and gross.
  - R: Mature content, limited reach.
  - PG: Kids-focused, less broad appeal.
- **G-Rated Movies:**
  - Best gross-to-budget ratio.
  - Low costs, high global appeal.
  - Dominated by Disney and Pixar.

Conclusions

PG-13

Highest Production Budgets and box office Grosses

R and PG

Perform well but high Production Budget

G

Best return on investment

Question 3:

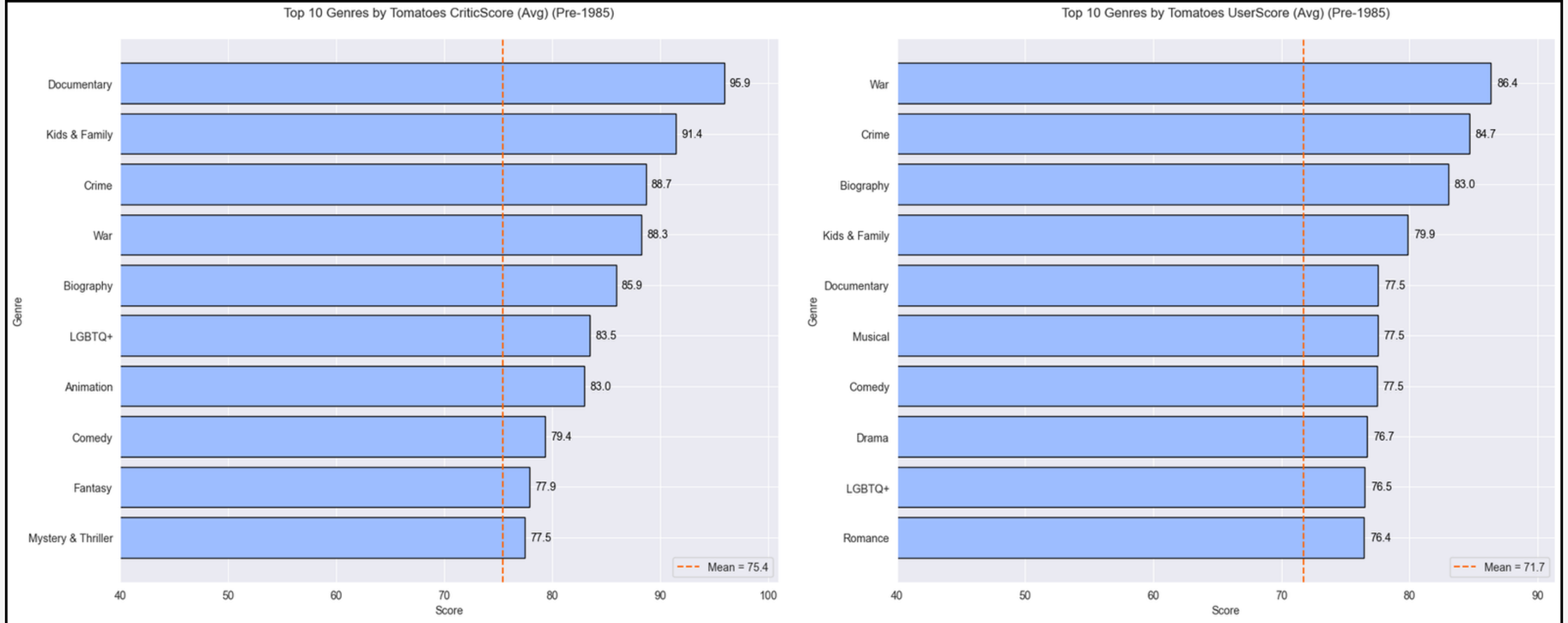
What genres tend to perform better in terms of critic and user scores?

Preprocessing

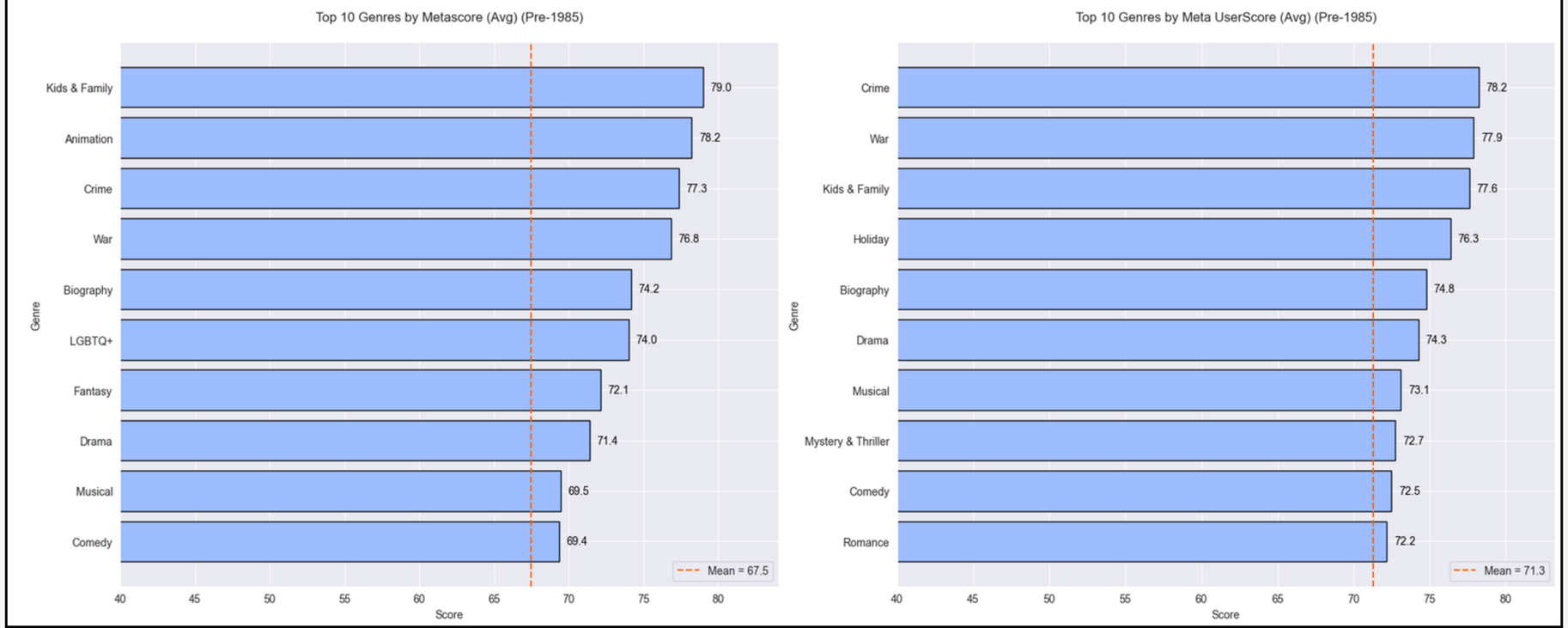
Seperate rows with multiple genres

Divide data into different "time period" bins

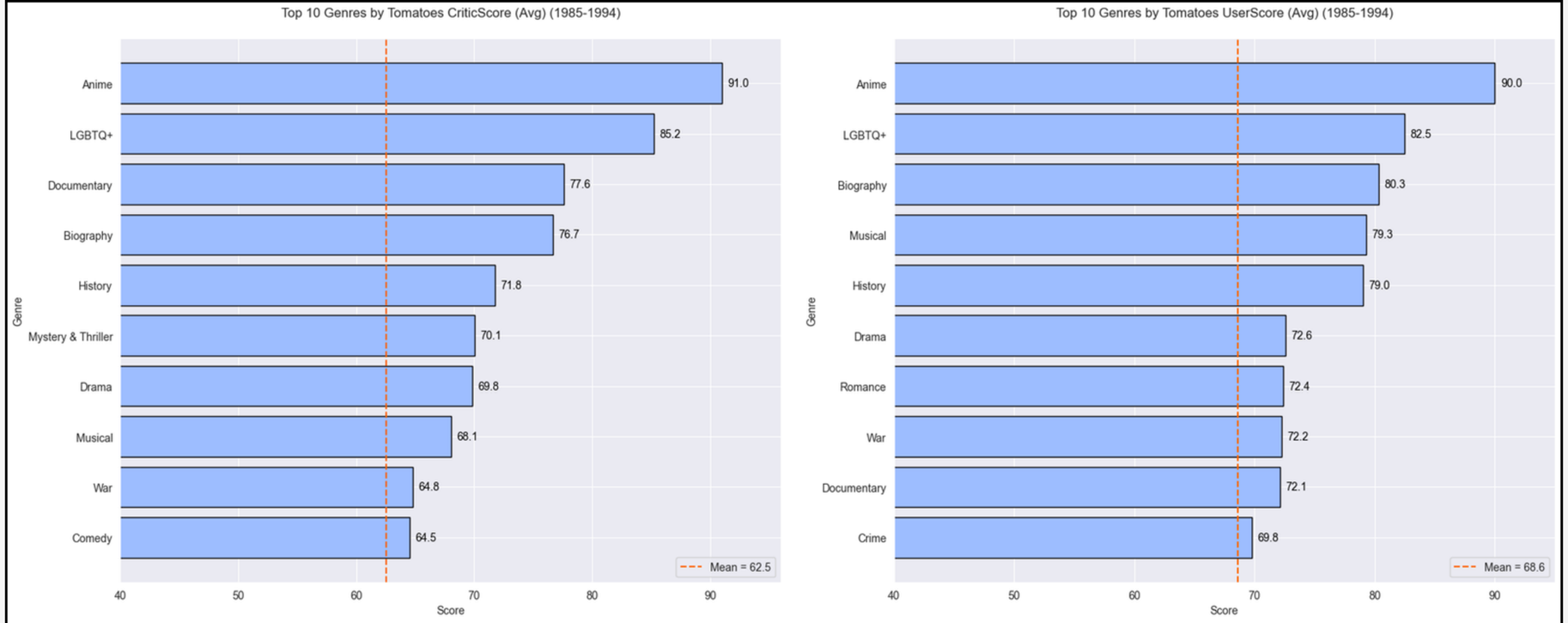
Calculate mean score by genres


Top 10 Genres by Tomatoes CriticScore (Avg) (Pre-1985)
Documentary 95.9
Kids & Family 91.4
Crime 88.7
War 88.3
Biography 85.9
LGBTQ+ 83.5
Animation 83.0
Comedy 79.4
Fantasy 77.9
Mystery & Thriller 77.5
Mean = 75.4
Genre
Score
40 50 60 70 80 90 100

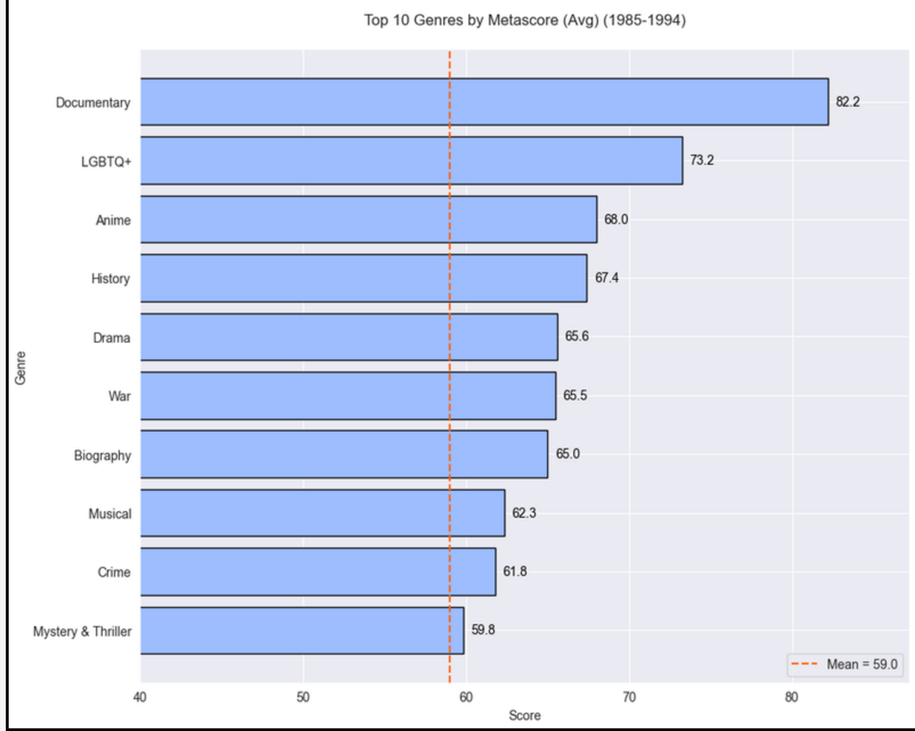
Top 10 Genres by Tomatoes UserScore (Avg) (Pre-1985)
War 86.4
Crime 84.7
Biography 83.0
Kids & Family 79.9
Documentary 77.5
Musical 77.5
Comedy 77.5
Drama 76.7
LGBTQ+ 76.5
Romance 76.4
Mean = 71.7
Genre
Score
40 50 60 70 80 90

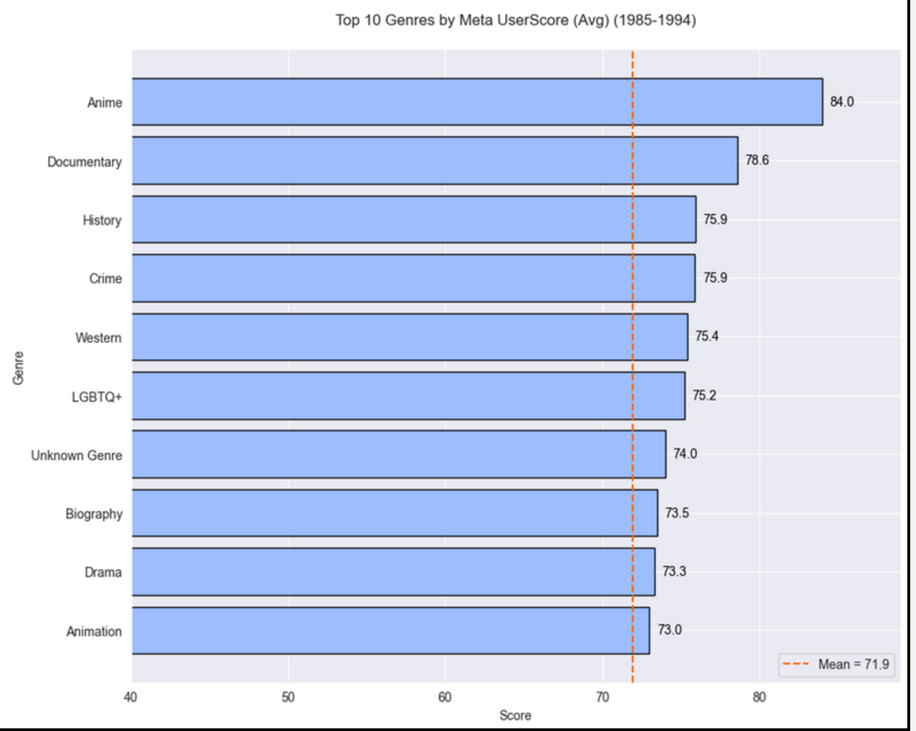

Top 10 Genres by Metascore (Avg) (Pre-1985)
Kids & Family 79.0
Animation 78.2
Crime 77.3
War 76.8
Biography 74.2
LGBTQ+ 74.0
Fantasy 72.1
Drama 71.4
Musical 69.5
Comedy 69.4
Mean = 67.5
Genre
Score
40 45 50 55 60 65 70 75 80

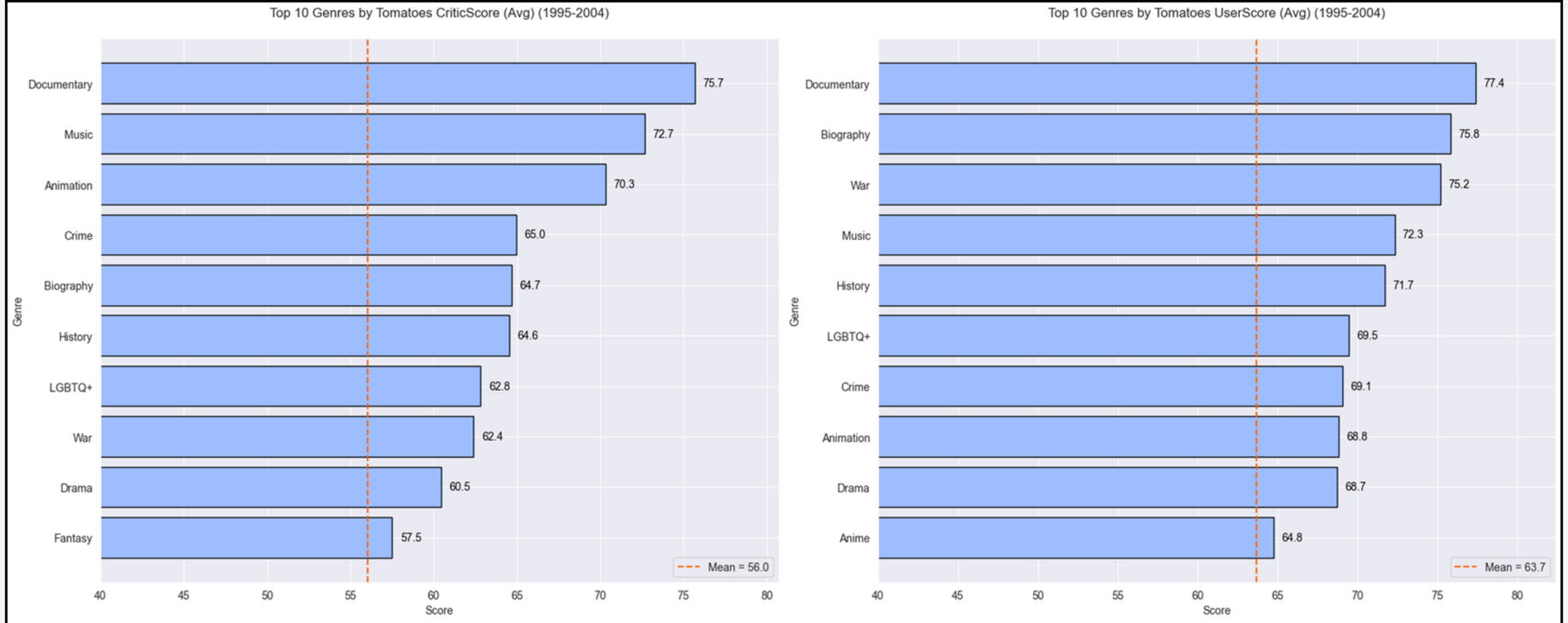
Top 10 Genres by Meta UserScore (Avg) (Pre-1985)
Crime 78.2
War 77.9
Kids & Family 77.6
Holiday 76.3
Biography 74.8
Drama 74.3
Musical 73.1
Mystery & Thriller 72.7
Comedy 72.5
Romance 72.2
Mean = 71.3
Genre
Score
40 45 50 55 60 65 70 75 80

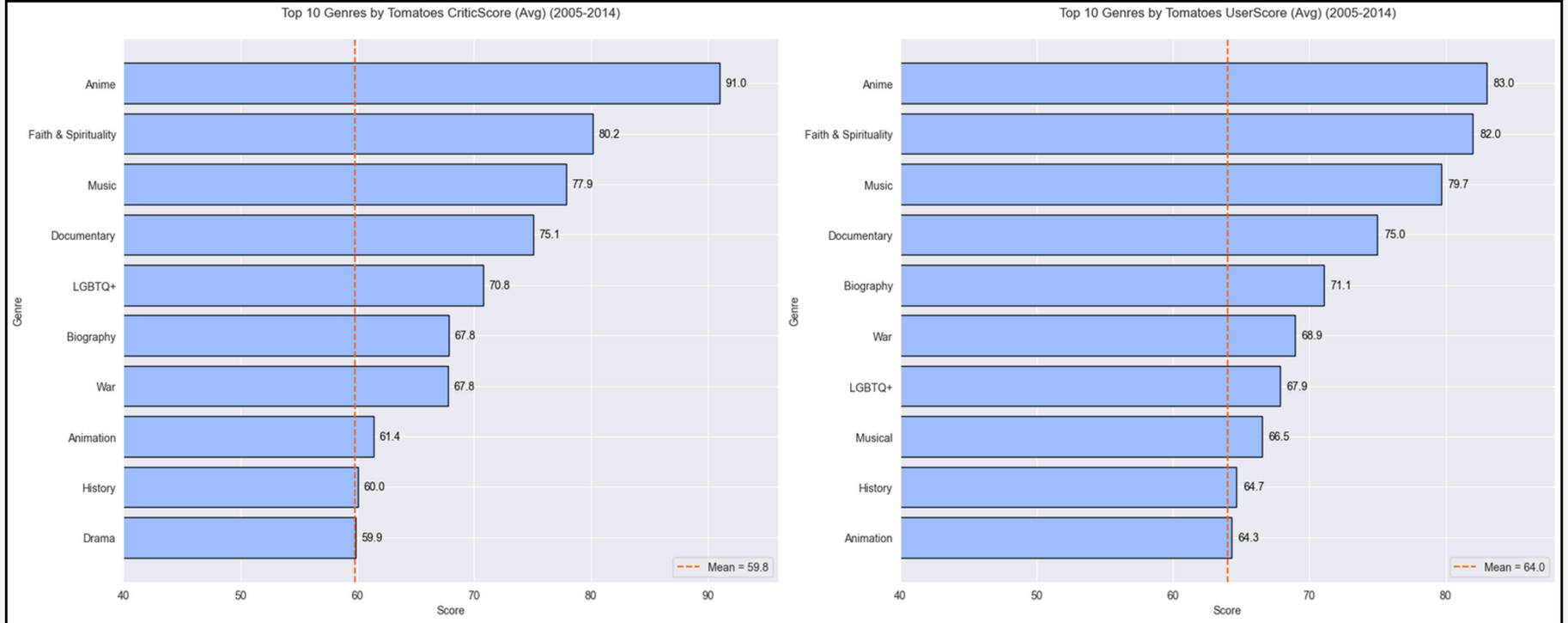

Top 10 Genres by Tomatoes CriticScore (Avg) (1985-1994)
Anime 91.0
LGBTQ+ 85.2
Documentary 77.6
Biography 76.7
History 71.8
Mystery & Thriller 70.1
Drama 69.8
Musical 68.1
War 64.8
Comedy 64.5
Mean = 62.5
Genre
Score
40 50 60 70 80 90

Top 10 Genres by Tomatoes UserScore (Avg) (1985-1994)
Anime 90.0
LGBTQ+ 82.5
Biography 80.3
Musical 79.3
History 79.0
Drama 72.6
Romance 72.4
War 72.2
Documentary 72.1
Crime 69.8
Mean = 68.6
Genre
Score
40 50 60 70 80 90

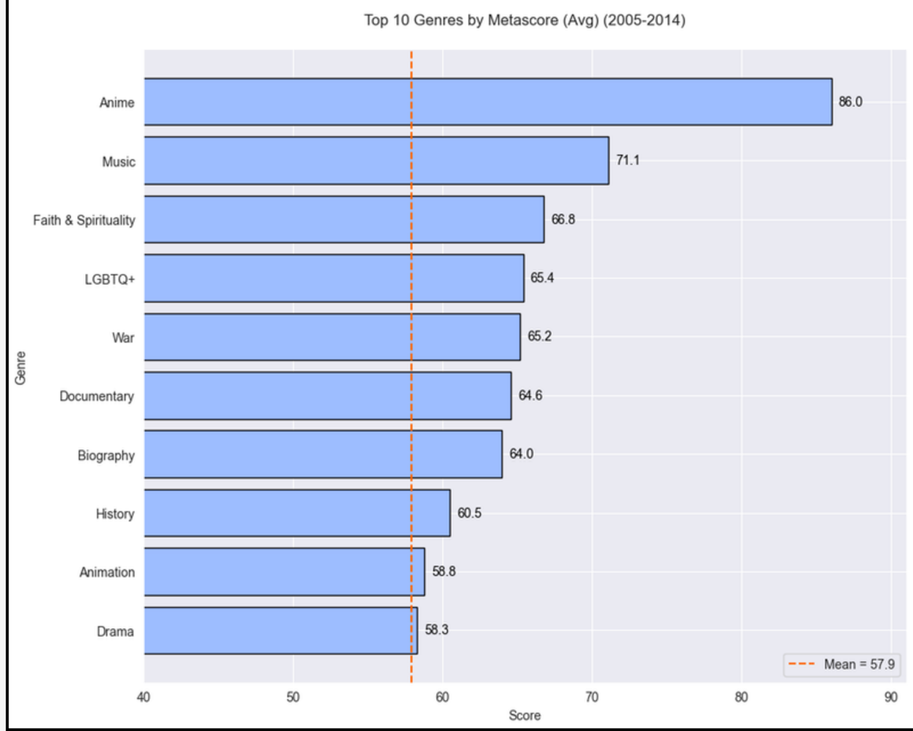

Top 10 Genres by Metascore (Avg) (1985-1994)
Documentary 82.2
LGBTQ+ 73.2
Anime 68.0
History 67.4
Drama 65.6
War 65.5
Biography 65.0
Musical 62.3
Crime 61.8
Mystery & Thriller 59.8
Mean = 59.0
Genre
Score
40 50 60 70 80

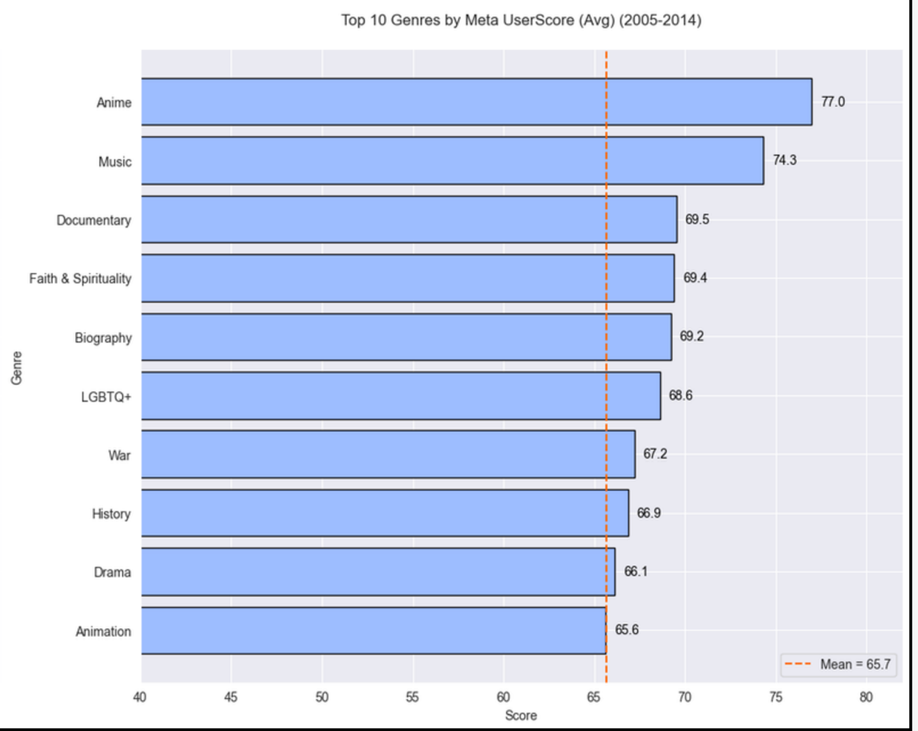

Top 10 Genres by Meta UserScore (Avg) (1985-1994)
Anime 84.0
Documentary 78.6
History 75.9
Crime 75.9
Western 75.4
LGBTQ+ 75.2
Unknown Genre 74.0
Biography 73.5
Drama 73.3
Animation 73.0
Mean = 71.9
Genre
Score
40 50 60 70 80

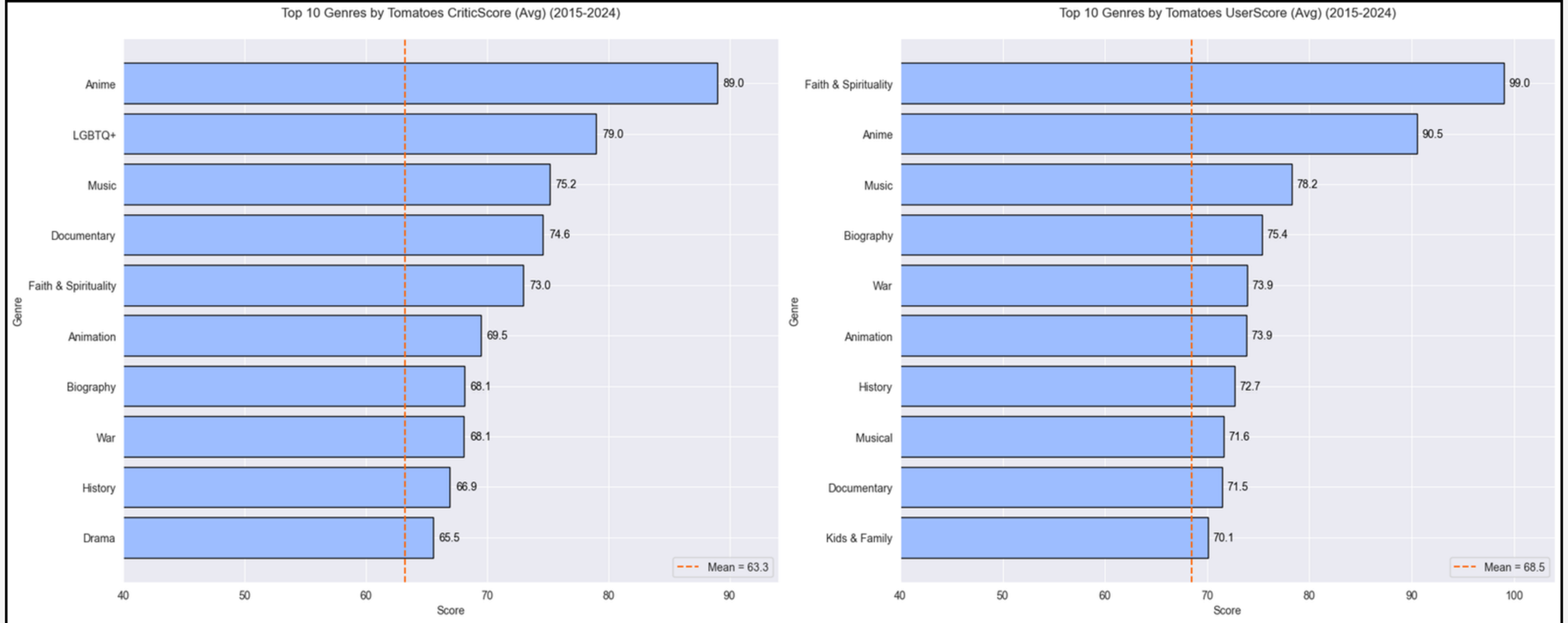

Top 10 Genres by Tomatoes CriticScore (Avg) (1995-2004)
Documentary 75.7
Music 72.7
Animation 70.3
Crime 65.0
Biography 64.7
History 64.6
LGBTQ+ 62.8
War 62.4
Drama 60.5
Fantasy 57.5
Mean = 56.0
Genre
Score
Top 10 Genres by Tomatoes UserScore (Avg) (1995-2004)
Documentary 77.4
Biography 75.8
War 75.2
Music 72.3
History 71.7
LGBTQ+ 69.5
Crime 69.1
Animation 68.8
Drama 68.7
Anime 64.8
Mean = 63.7
Genre
Score
Top 10 Genres by Metascore (Avg) (1995-2004)
Mean = 54.6
Music 65.4
Animation 64.3
War 63.4
Documentary 63.4
Biography 62.6
History 60.7
Crime 59.5
Drama 58.1
LGBTQ+ 57.6
Fantasy 56.0
Genre
Score
Top 10 Genres by Meta UserScore (Avg) (1995-2004)
Animation 73.0
Music 73.0
War 71.9
History 70.7
Documentary 69.4
Biography 68.9
Crime 68.9
Fantasy 67.9
Adventure 67.8
Drama 67.8
Mean = 66.8
Genre
Score


Top 10 Genres by Tomatoes CriticScore (Avg) (2005-2014)
Anime 91.0
Faith & Spirituality 80.2
Music 77.9
Documentary 75.1
LGBTQ+ 70.8
Biography 67.8
War 67.8
Animation 61.4
History 60.0
Drama 59.9
Mean = 59.8
Genre
Score
40 50 60 70 80 90

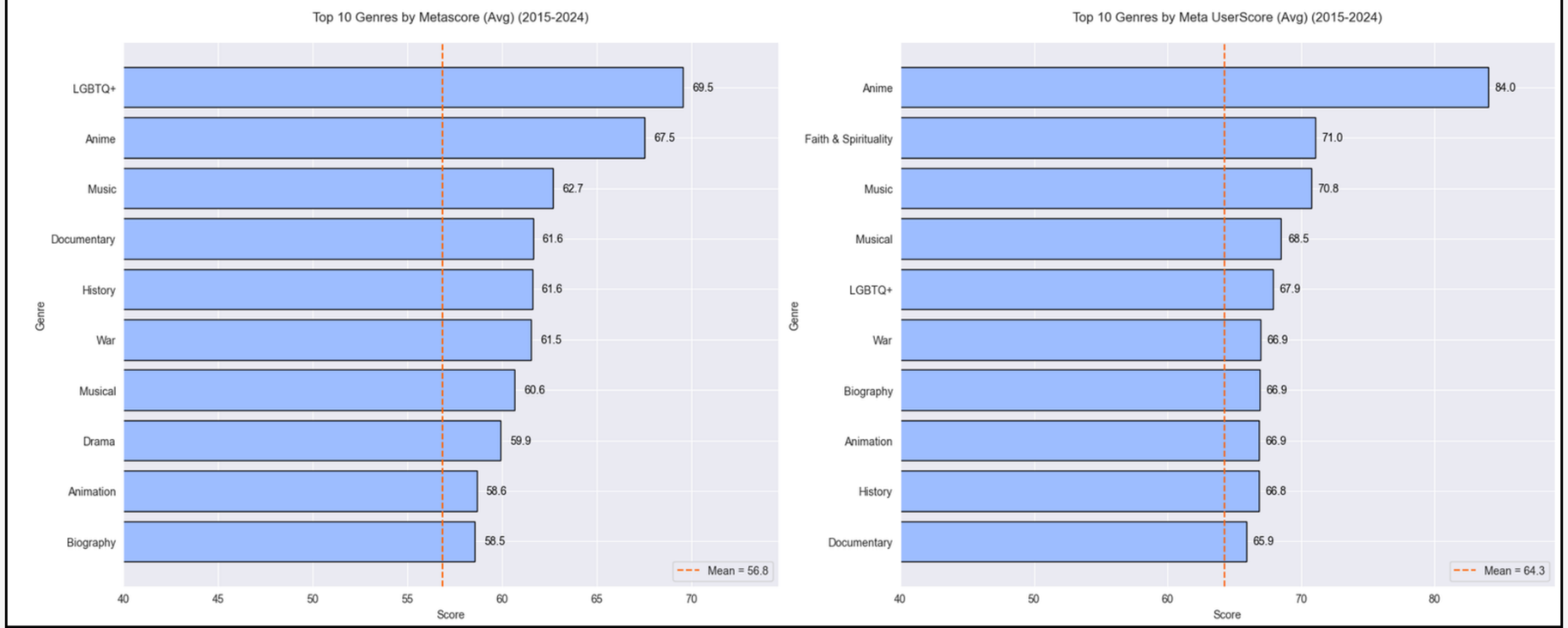
Top 10 Genres by Tomatoes UserScore (Avg) (2005-2014)
Anime 83.0
Faith & Spirituality 82.0
Music 79.7
Documentary 75.0
Biography 71.1
War 68.9
LGBTQ+ 67.9
Musical 66.5
History 64.7
Animation 64.3
Mean = 64.0
Genre
Score
40 50 60 70 80


Top 10 Genres by Metascore (Avg) (2005-2014)
Anime 86.0
Music 71.1
Faith & Spirituality 66.8
LGBTQ+ 65.4
War 65.2
Documentary 64.6
Biography 64.0
History 60.5
Animation 58.8
Drama 58.3
Mean = 57.9
Genre
Score
40 50 60 70 80 90


Top 10 Genres by Meta UserScore (Avg) (2005-2014)
Anime 77.0
Music 74.3
Documentary 69.5
Faith & Spirituality 69.4
Biography 69.2
LGBTQ+ 68.6
War 67.2
History 66.9
Drama 66.1
Animation 65.6
Mean = 65.7
Genre
Score
40 45 50 55 60 65 70 75 80


Top 10 Genres by Tomatoes CriticScore (Avg) (2015-2024)
Anime 89.0
LGBTQ+ 79.0
Music 75.2
Documentary 74.6
Faith & Spirituality 73.0
Animation 69.5
Biography 68.1
War 68.1
History 66.9
Drama 65.5
Mean = 63.3
Genre
Score
40 50 60 70 80 90

Top 10 Genres by Tomatoes UserScore (Avg) (2015-2024)
Faith & Spirituality 99.0
Anime 90.5
Music 78.2
Biography 75.4
War 73.9
Animation 73.9
History 72.7
Musical 71.6
Documentary 71.5
Kids & Family 70.1
Mean = 68.5
Genre
Score
40 50 60 70 80 90 100


Top 10 Genres by Metascore (Avg) (2015-2024)
LGBTQ+ 69.5
Anime 67.5
Music 62.7
Documentary 61.6
History 61.6
War 61.5
Musical 60.6
Drama 59.9
Animation 58.6
Biography 58.5
Mean = 56.8
Genre
Score
40 45 50 55 60 65 70

Top 10 Genres by Meta UserScore (Avg) (2015-2024)
Anime 84.0
Faith & Spirituality 71.0
Music 70.8
Musical 68.5
LGBTQ+ 67.9
War 66.9
Biography 66.9
Animation 66.9
History 66.8
Documentary 65.9
Mean = 64.3
Genre
Score
40 50 60 70 80

Conclusions

Pre-1985

Documentary, Kids & Family, War, and Crime

1985 - 1994

Anime & growth of LGBTQ+

1995 - 2004

Documentary Music and Animation

2005 - 2014

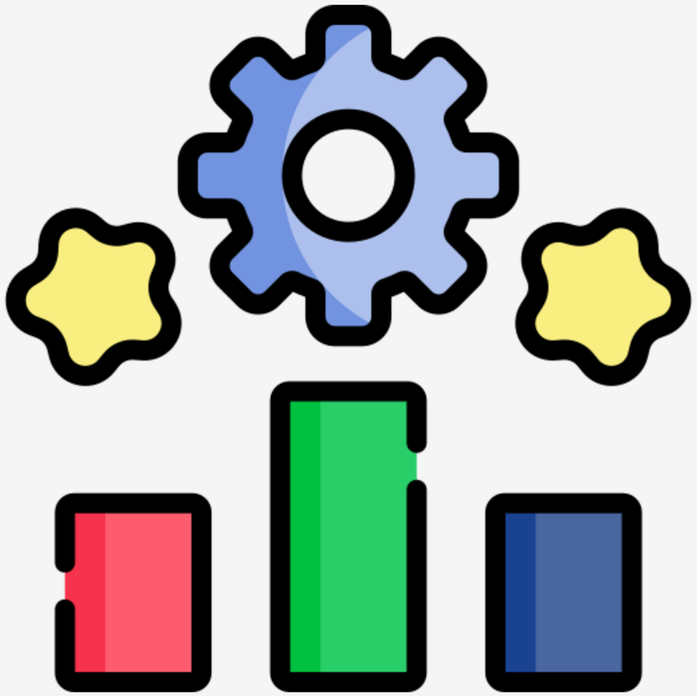
Anime Dominance

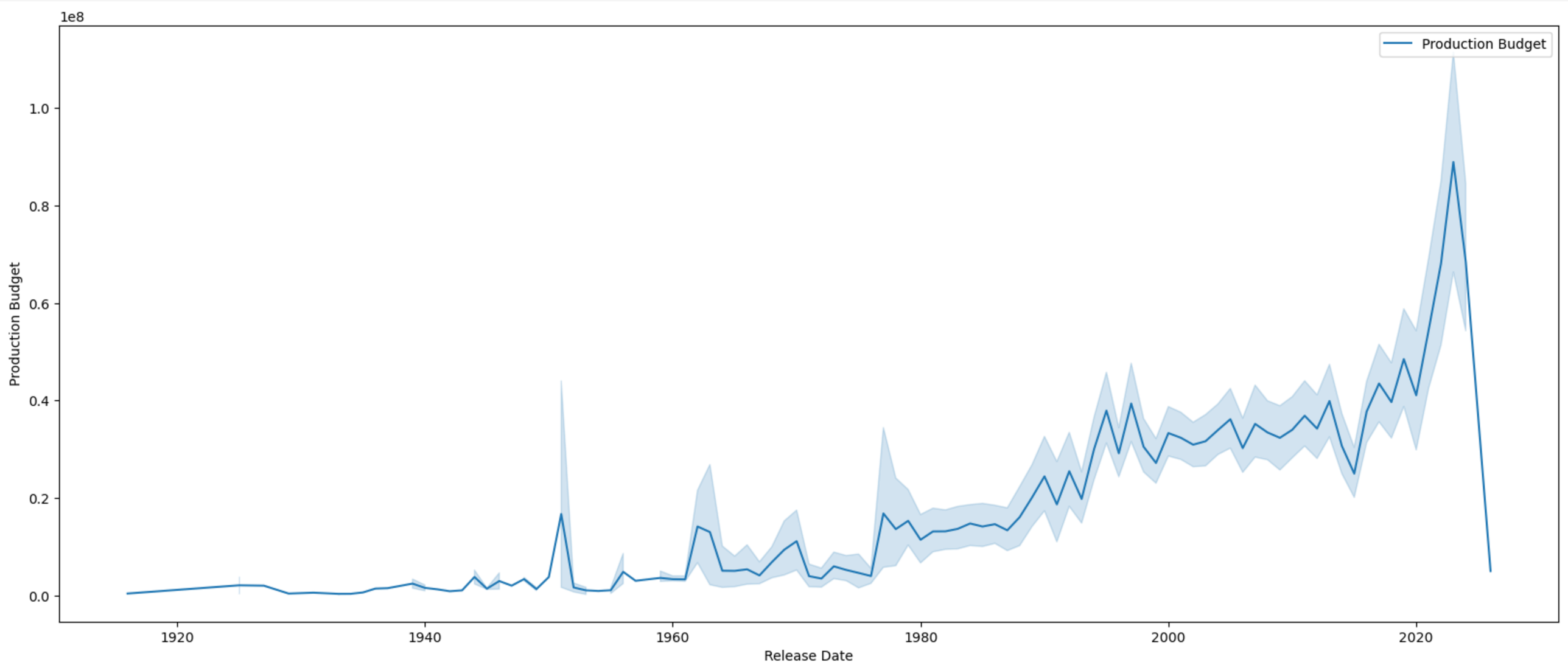
2015 - 2024

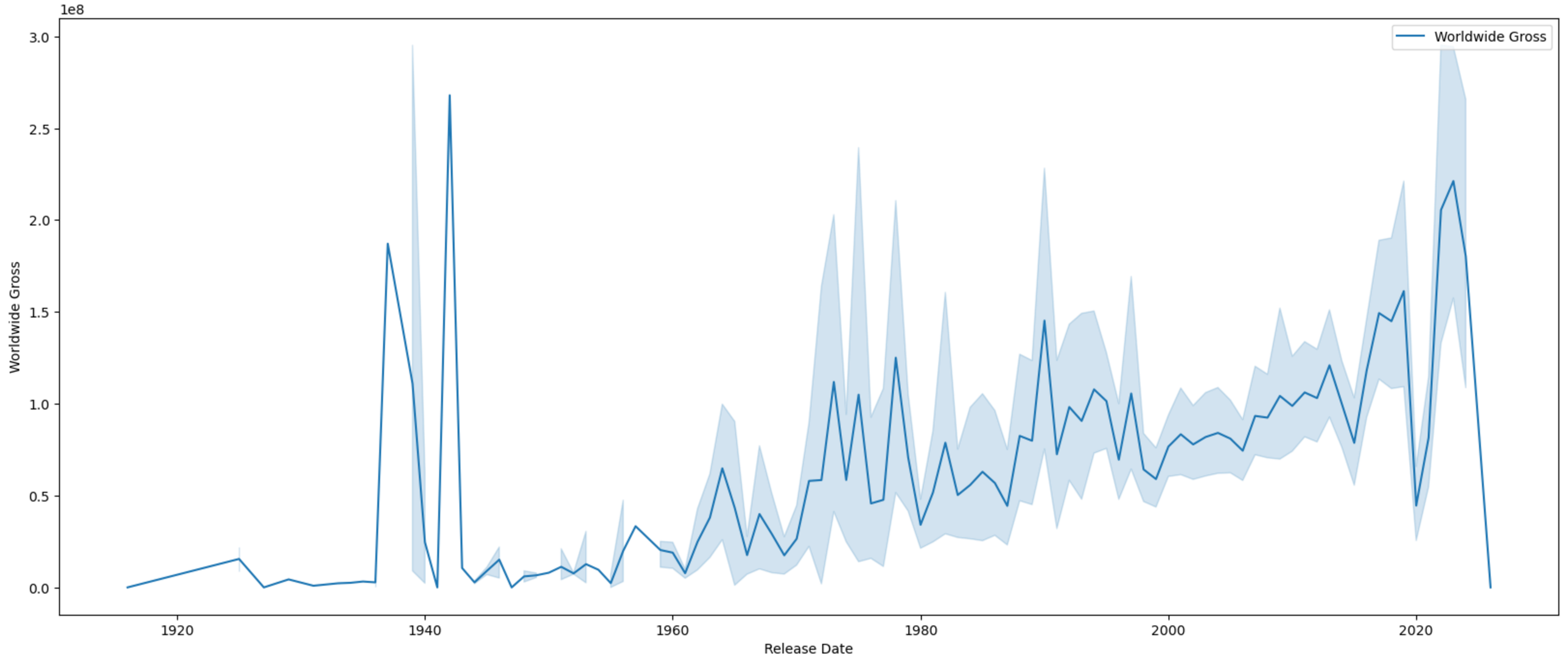
Anime still dominating, fondness in Faith & Spirituality

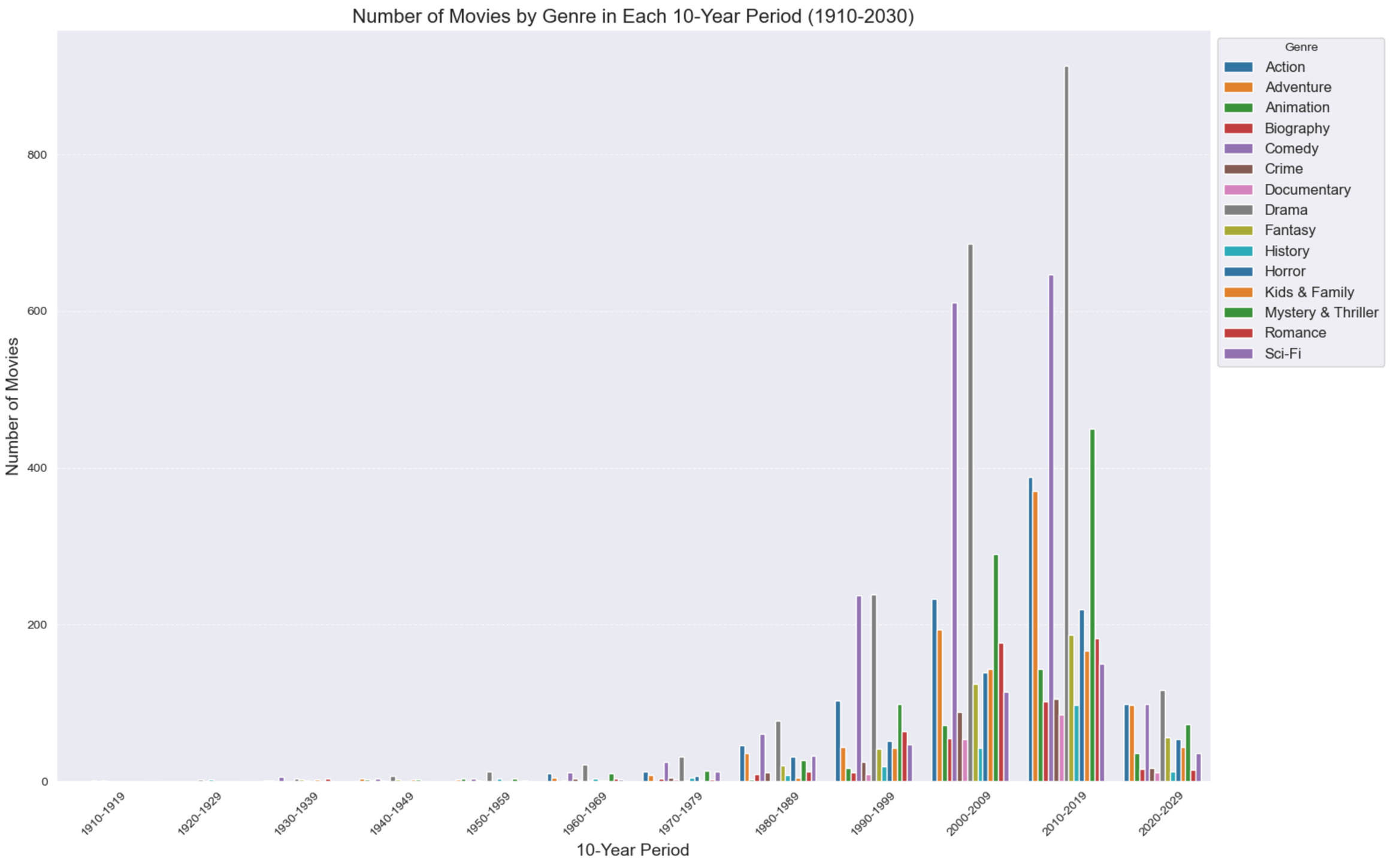
Question 4:

Are there trends in movie production and performance over time?




1e8
Production Budget
1.0
0.8
0.6
0.4
0.2
0.0
Production Budget
1920
1940
1960
1980
2000
2020
Release Date


1e8
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Worldwide Gross
1920
1940
1960
1980
2000
2020
Release Date
Worldwide Gross


Number of Movies by Genre in Each 10-Year Period (1910-2030)
Number of Movies
0
200
400
600
800
1910-1919
1920-1929
1930-1939
1940-1949
1950-1959
1960-1969
1970-1979
1980-1989
1990-1999
2000-2009
2010-2019
2020-2029
10-Year Period
Genre
Action
Adventure
Animation
Biography
Comedy
Crime
Documentary
Drama
Fantasy
History
Horror
Kids & Family
Mystery & Thriller
Romance
Sci-Fi


Mean User and Critic Scores Over Time (1910-2030)
Score Type
Mean UserScore
Mean CriticScore
Mean Score
90
85
80
75
70
65
60
55
1910-1919
1920-1929
1930-1939
1940-1949
1950-1959
1960-1969
1970-1979
1980-1989
1990-1999
2000-2009
2010-2019
2020-2029
Decade

Conclusion

Financial aspect

- Higher budgets = higher revenue.
- Tech and franchises boosted growth post-1990.
- COVID-19 caused a sharp 2020 decline.

Genre trends

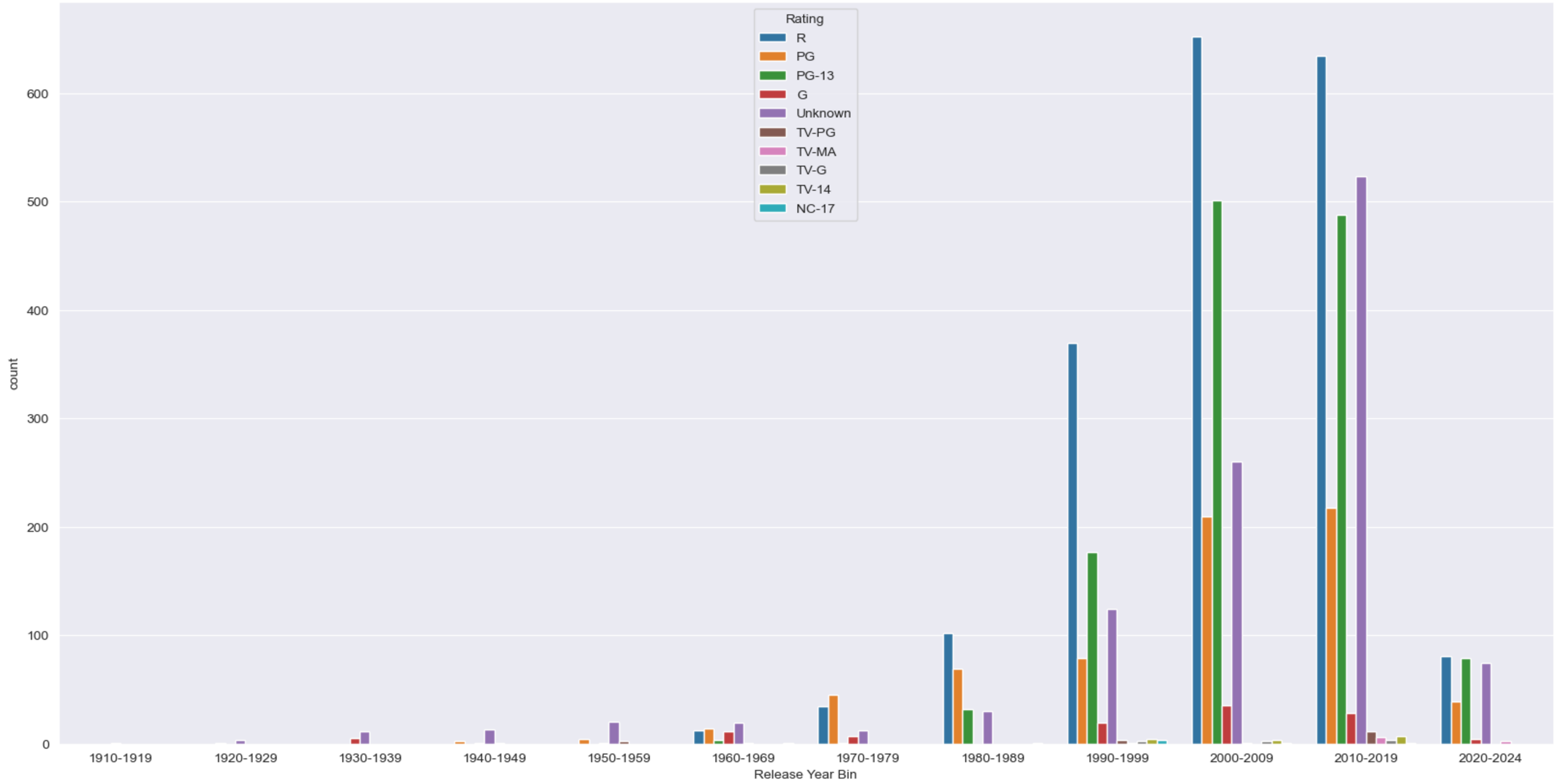
- Action/Adventure: Always popular.
- Comedy: Steady demand.
- Animation/Family: Growth since 1990s.

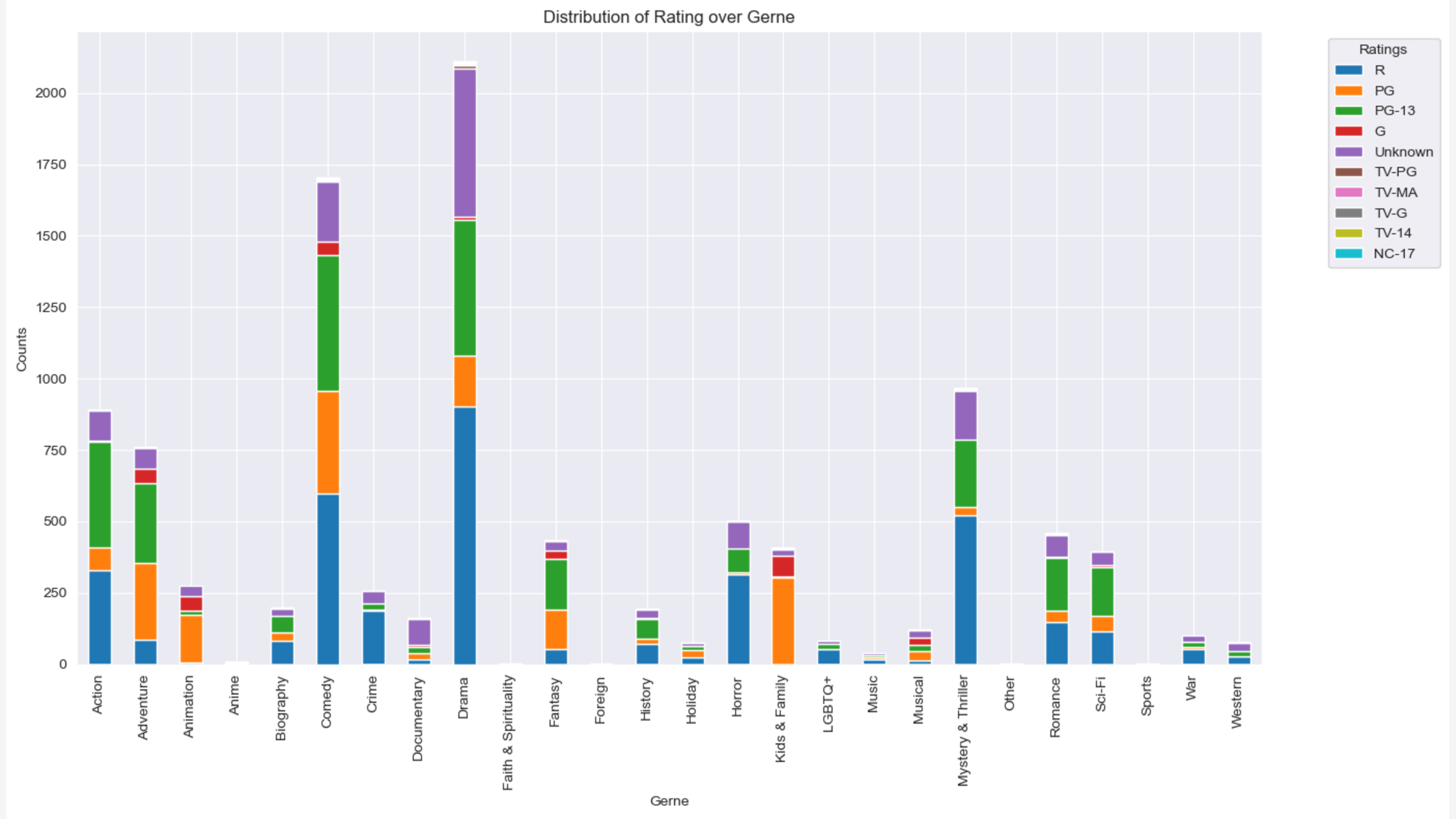
Critic and audience ratings

- User-Critic gap widens since 1980s.
- User Scores recovering recently.
- Streaming reshapes audience ratings.

Question 5:

What is the trend of rating movies over time?


Rating
R
PG
PG-13
G
Unknown
TV-PG
TV-MA
TV-G
TV-14
NC-17
count
0
100
200
300
400
500
600
1910-1919
1920-1929
1930-1939
1940-1949
1950-1959
1960-1969
1970-1979
1980-1989
1990-1999
2000-2009
2010-2019
2020-2024
Release Year Bin


Distribution of Rating over Gerne
Counts
Gerne
Ratings
R
PG
PG-13
G
Unknown
TV-PG
TV-MA
TV-G
TV-14
NC-17
0
250
500
750
1000
1250
1500
1750
2000
Action
Adventure
Animation
Anime
Biography
Comedy
Crime
Documentary
Drama
Faith & Spirituality
Fantasy
Foreign
History
Holiday
Horror
Kids & Family
LGBTQ+
Music
Musical
Mystery & Thriller
Other
Romance
Sci-Fi
Sports
War
Western

Conclusion

Reflects changing norms and audience preferences.

MPAA ratings (1968) set age-appropriate guidelines.

Rise of R-rated films for mature themes in action, drama, and horror.

PG and PG-13 grew with family-friendly and teen-focused films.

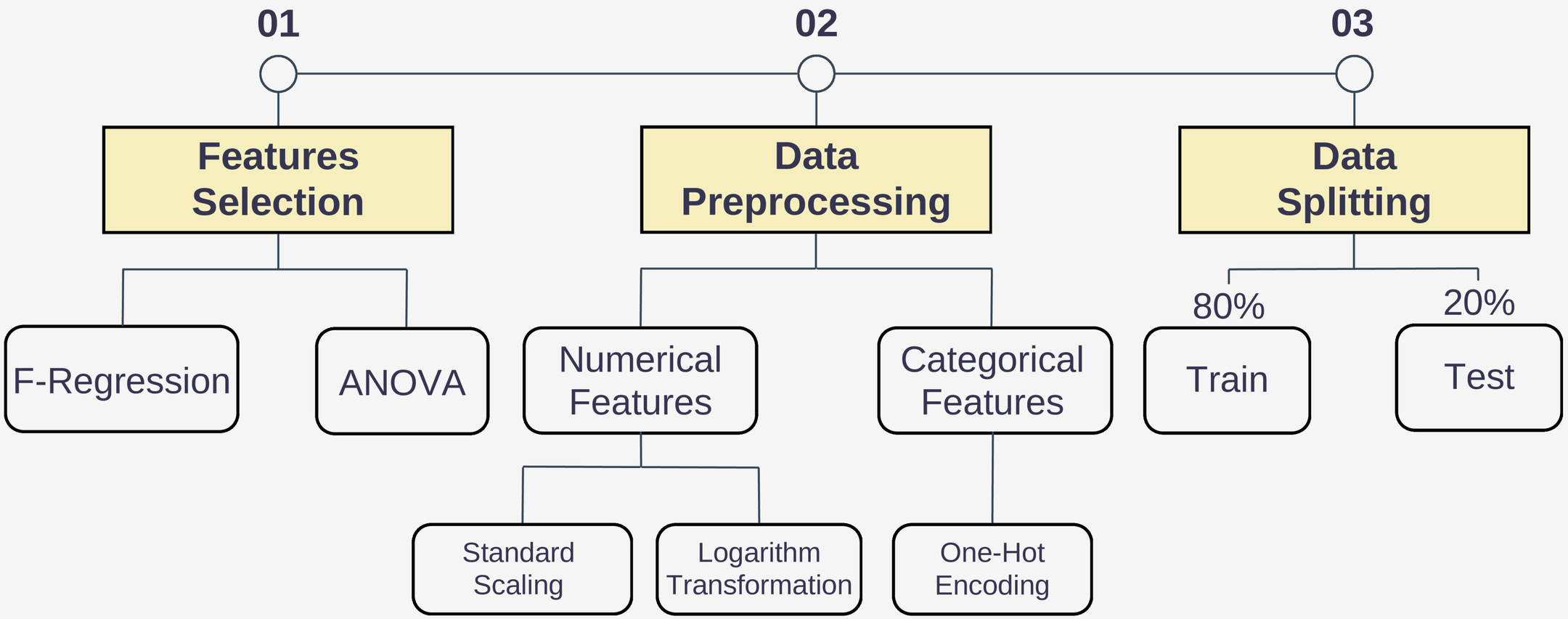
PG-13 dominates modern blockbusters, balancing mature themes with broad appeal.

06
Modelling

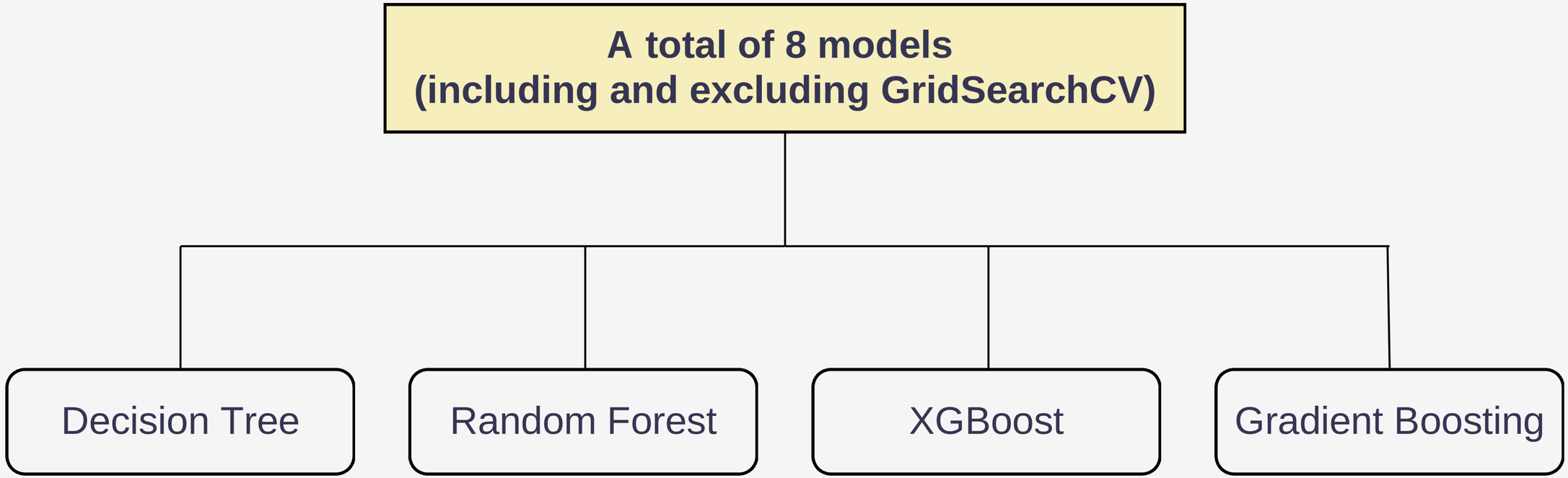
Main Reason

Predict the gross of movies

Data Preparation



Models Training



Models Testing

Model Name	MAE	MAE (GridSearchCV)	Mean R-Squared	Mean R-Squared (GridSearchCV)
XGBoost	0.425	0.417	0.599	0.607
Gradient Boosting	0.429	0.423	0.596	0.6
Decision Tree	0.575	0.493	0.177	0.474
Random Forest	0.425	0.441	0.567	0.573

Models Evaluation

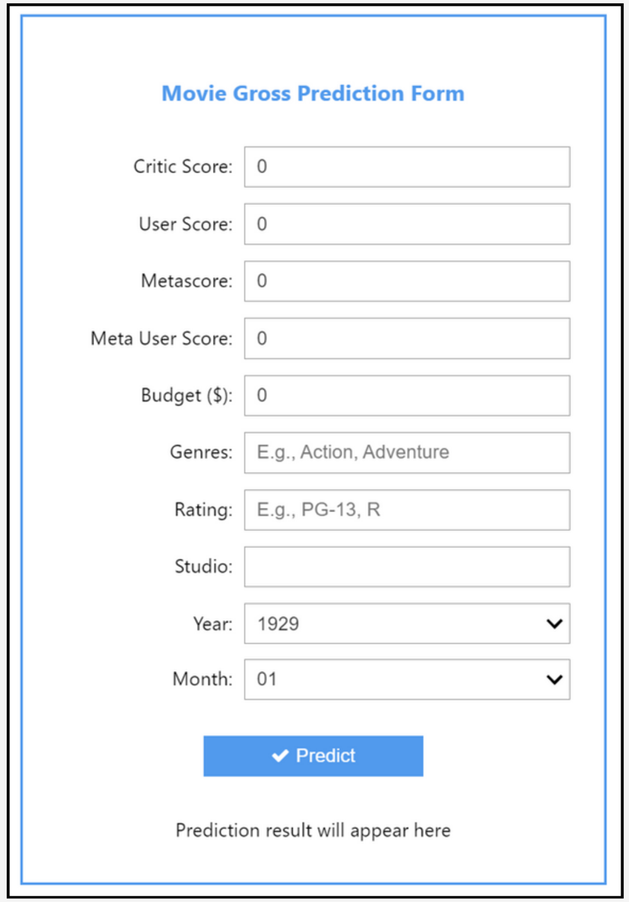
Gradient Boosting vs. XGBoost (with GridSearchCV)

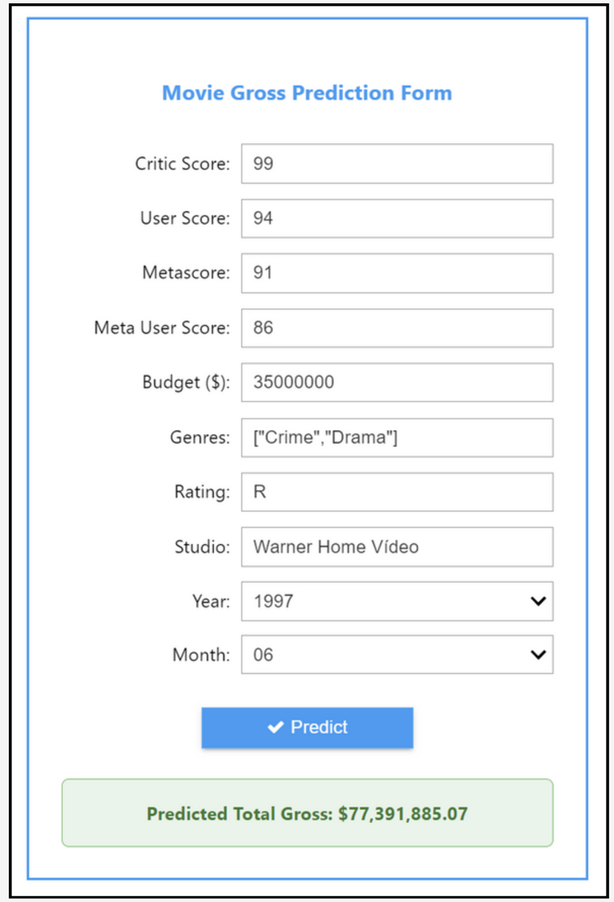
Model Name	MAE	Mean R-Squared	Traning Time
Gradient Boosting	0.423	0.6	35s
XGBoost	0.417	0.607	18s
Difference (%)	+ 1 %	- 1%	+ 95%

Models Evaluation

XGBoost with GridSearchCV is the best model

Prediction


Movie Gross Prediction Form
Critic Score: 0
User Score: 0
Metascore: 0
Meta User Score: 0
Budget ($): 0
Genres: E.g., Action, Adventure
Rating: E.g., PG-13, R
Studio:
Year: 1929
Month: 01
Predict
Prediction result will appear here


Movie Gross Prediction Form
Critic Score: 99
User Score: 94
Metascore: 91
Meta User Score: 86
Budget ($): 35000000
Genres: ["Crime","Drama"]
Rating: R
Studio: Warner Home Video
Year: 1997
Month: 06
Predict
Predicted Total Gross: $77,391,885.07

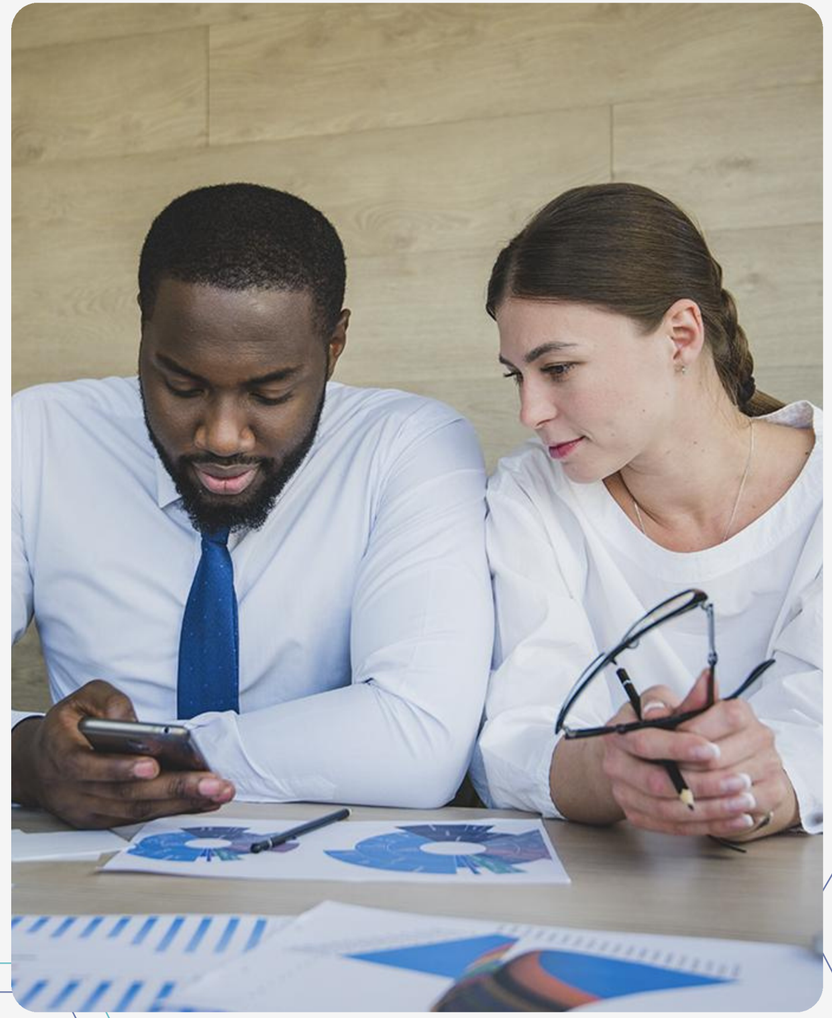
Reflection

Challenges

- Overcoming anti-scraping measures (CAPTCHA, dynamic content).
- Handling complex data structures with multiple values in columns like Cast, Director, Genre.
- Resolving notebook conflicts on GitHub and distributing team tasks.
- Handling missing values, heavy one-hot encoding, and categorical inconsistencies.

Lessons Learned

- Proficient in data crawling (Selenium, Scrapy) and imputation techniques (KNN, Decision Tree).
- Gained insights into data science workflow, hyperparameter tuning, and feature impacts (genres, ratings, etc.).
- Learned to create interactive form interfaces in Jupyter Notebook for smoother workflows.
- Improved collaboration, real-world problem-solving, and data cleaning skills.



Thanks for listening!