**HO CHI MINH UNIVERSITY OF SCIENCE**

**FACULTY OF INFORMATION TECHNOLOGY**

**FINAL PRESENTATION**

**Subject: Intro to Data Science Class:** 22KHDL1  
**Group:** 5

**Students:**

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

**TP. Hồ Chí Minh, tháng 12, năm 2024**

Contents

[**1.** **Project Plan** 4](#_Toc185019990)

[**2.** **Data Collection** 5](#_Toc185019991)

[**a.** **Rotten Tomatoes** 5](#_Toc185019992)

[**b.** **MetaCritic** 7](#_Toc185019993)

[**c.** **The Numbers** 8](#_Toc185019994)

[**d.** **Investopedia** 10](#_Toc185019995)

[**3.** **Data Preprocessing** 12](#_Toc185019996)

[**a.** **Handling missing data** 12](#_Toc185019997)

[- **Numerical features** 12](#_Toc185019998)

[- **Categorical features** 13](#_Toc185019999)

[**b.** **Data Standardization** 15](#_Toc185020000)

[- **Numerical features** 15](#_Toc185020001)

[- **Categorical features** 15](#_Toc185020002)

[**c.** **Outliers Survey** 15](#_Toc185020003)

[- **Numerical features** 15](#_Toc185020004)

[- **Categorical features** 18](#_Toc185020005)

[**d.** **Synthesis in schools of using multiple sources** 18](#_Toc185020006)

[**e.** **Data Consistency** 19](#_Toc185020007)

[- **Data Type** 19](#_Toc185020008)

[- **Data Duplicate** 19](#_Toc185020009)

[- **Meaning of each column** 19](#_Toc185020010)

[- **Meaning of each row** 20](#_Toc185020011)

[**4.** **Data Preprocessing + Exploration** 20](#_Toc185020012)

[**a.** **Question 1** 20](#_Toc185020013)

[- **Content** 20](#_Toc185020014)

[- **Result (Based on chart)** 21](#_Toc185020015)

[- **Comment and conclusion** 22](#_Toc185020016)

[**b.** **Question 2** 23](#_Toc185020017)

[- **Content** 23](#_Toc185020018)

[- **Result (Based on chart)** 23](#_Toc185020019)

[- **Comment and conclusion** 24](#_Toc185020020)

[**c.** **Question 3** 24](#_Toc185020021)

[- **Content** 24](#_Toc185020022)

[- **Result (Based on chart)** 25](#_Toc185020023)

[- **Comment and conclusion** 28](#_Toc185020024)

[**d.** **Question 4** 29](#_Toc185020025)

[- **Content** 29](#_Toc185020026)

[- **Result (Based on chart)** 29](#_Toc185020027)

[- **Comment and conclusion** 30](#_Toc185020028)

[**e.** **Question 5** 31](#_Toc185020029)

[- **Content** 31](#_Toc185020030)

[- **Result (Based on chart)** 32](#_Toc185020031)

[- **Comment and conclusion** 32](#_Toc185020032)

[**5.** **Data Modelling** 33](#_Toc185020033)

[**a.** **Which models are used?** 33](#_Toc185020034)

[**b.** **Features Selection** 34](#_Toc185020035)

[- **For numerical features** 34](#_Toc185020036)

[- **For categorical columns** 34](#_Toc185020037)

[**c.** **Data Preprocessing** 35](#_Toc185020038)

[- **For numerical features** 35](#_Toc185020039)

[- **For categorical features** 36](#_Toc185020040)

[**e.** **Model Evaluation** 36](#_Toc185020041)

[- **Introduction to metrics used** 36](#_Toc185020042)

[- **Comparison** 37](#_Toc185020043)

[**f.** **Results and which is the best model** 39](#_Toc185020044)

[- **Results** 39](#_Toc185020045)

[- **Conclusion** 40](#_Toc185020046)

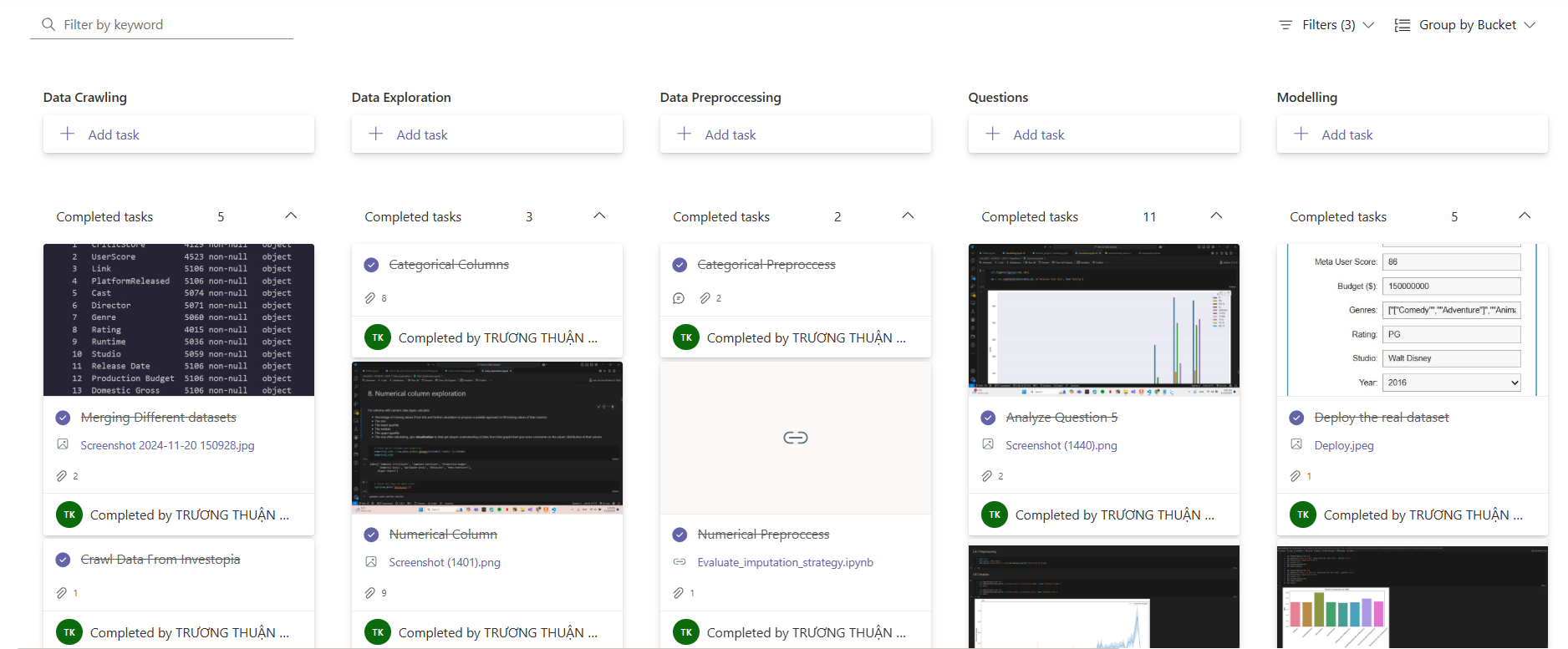
[**g.** **Output of the model and meaning** 40](#_Toc185020047)

[- **Output** 40](#_Toc185020048)

[- **Meaning** 40](#_Toc185020049)

# **Project Plan**

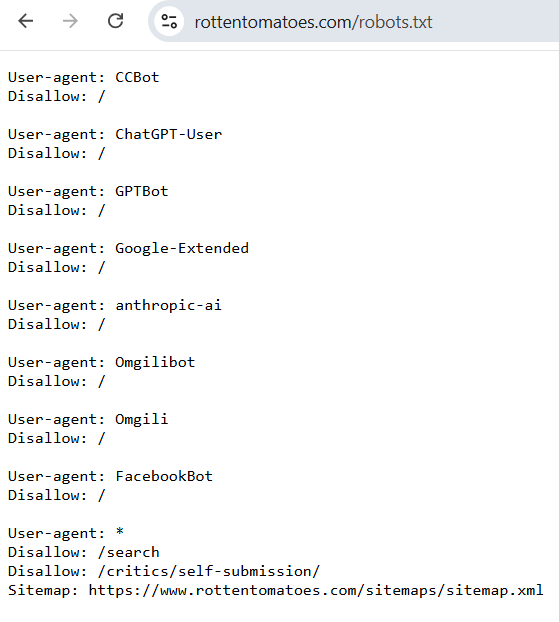
* **GitHub**: https://github.com/AshuraOtsuki/Intro2DS---HCMUS---2024
* **Microsoft Planner:** https://planner.cloud.microsoft/webui/plan/CPAR6teeCE-pZBMR6o\_p-sgAEG6v?tid=40127cd4-45f3-49a3-b05d-315a43a9f033

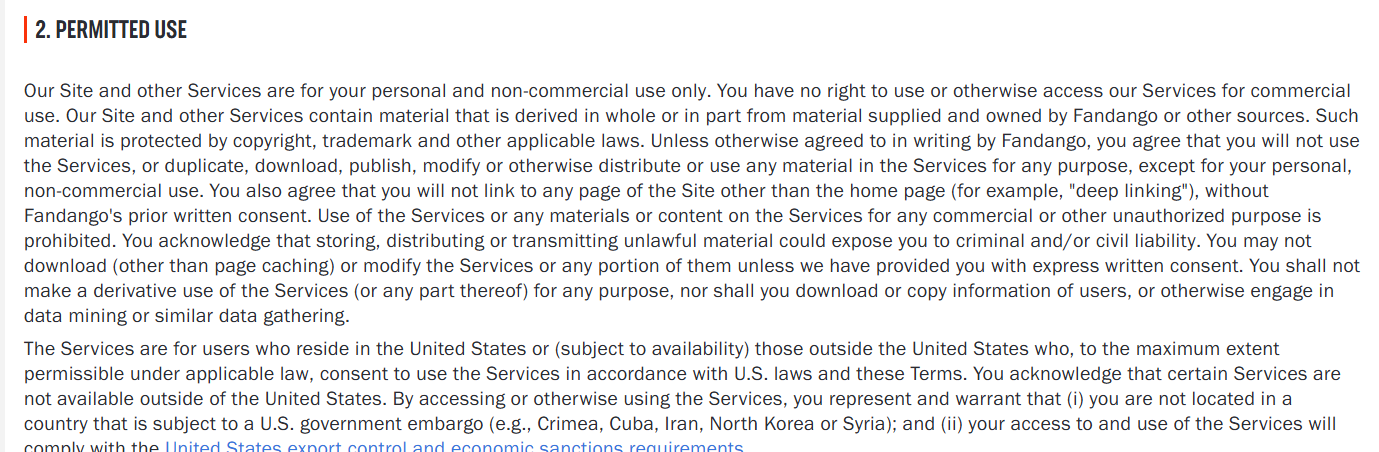


# **Data Collection**

## **Rotten Tomatoes**

* **Source:** <https://www.rottentomatoes.com>
* **Legality:**
* **Link**: https://www.rottentomatoes.com/robots.txt
* Specific bots mentioned (Facebook, GPT, and OMG) are **not allowed to crawl anything**.
* Others can crawl the site **except for the disallowed paths** like /search and /critics/self-submission/.
* As long as we are not using data for commercial, we can safely crawl data from Rotten Tomatoes





* **Reliability**:
* As the world’s most trusted and recognized source of movie and TV reviews, Rotten Tomatoes and the Tomatometer score have served as the most reliable home of entertainment recommendations for over 25 years

## **MetaCritic**

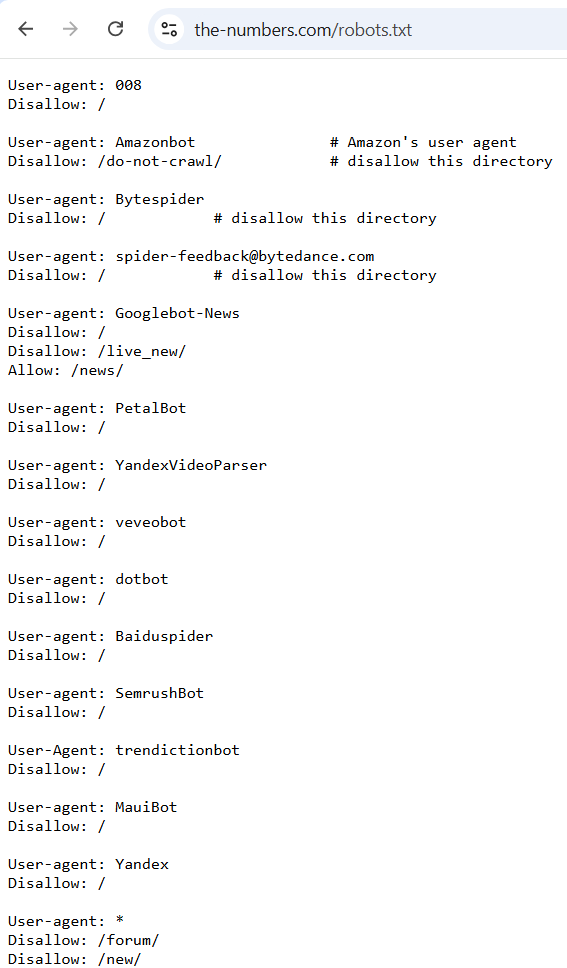
* **Source:** <https://www.metacritic.com>
* **Legality:**
* **Link:** https://www.metacritic.com/robots.txt
* Specific bots (e.g., AhrefsBot, GPTBot, SemrushBot, etc.) **are not allowed to crawl any part of the website.**
* Others are allowed to crawl most of the site, except for specific restricted paths **(e.g., /search, /login, /user, etc.).**



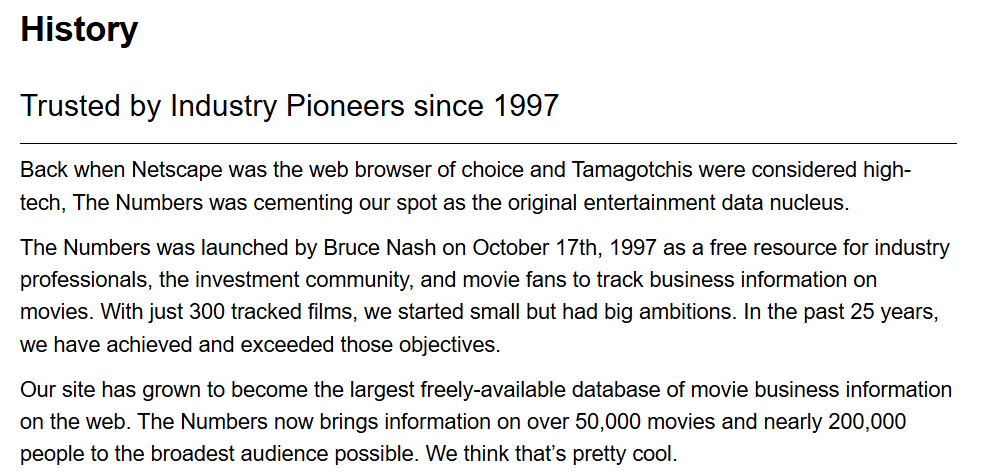
* **Reliability:**
* Metacritic is **reliable** for summarizing critical consensus, especially for movies, TV shows, and video games. However, to get the most value:
* Focus on the **critic reviews** for professional insights.
* Take **user scores** with a grain of salt, especially for controversial titles.

## **The Numbers**

* **Source:** <https://www.the-numbers.com>
* **Legality:**
* **Link:** https://www.the-numbers.com/robots.txt
* **Complete Blocking:** Specific bots (e.g., 008, Bytespider, PetalBot, Yandex) are not allowed to crawl any part of the site.
* **Partial Crawling:** Bots like Googlebot-News, BingBot, Applebot, and others have restricted access to certain directories and must follow crawl delays.
* **General Bots and Normal Users:** Most bots and users can crawl unrestricted sections of the site while avoiding the disallowed directories.

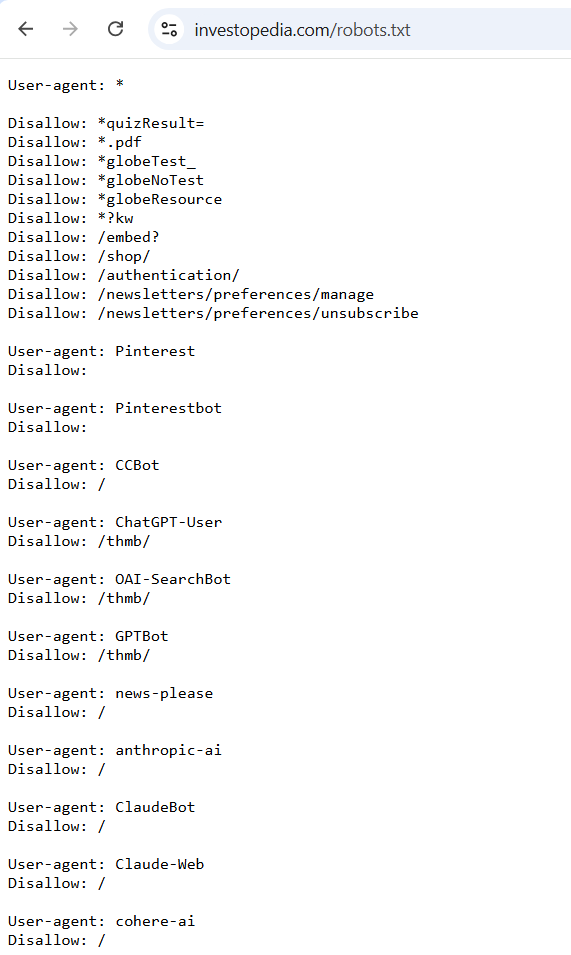
****

* **Reliability:**
* **The Numbers** is reliable for:
* Historical box office data.
* Basic production budgets and profitability metrics.
* Broad industry trends and financial analysis.

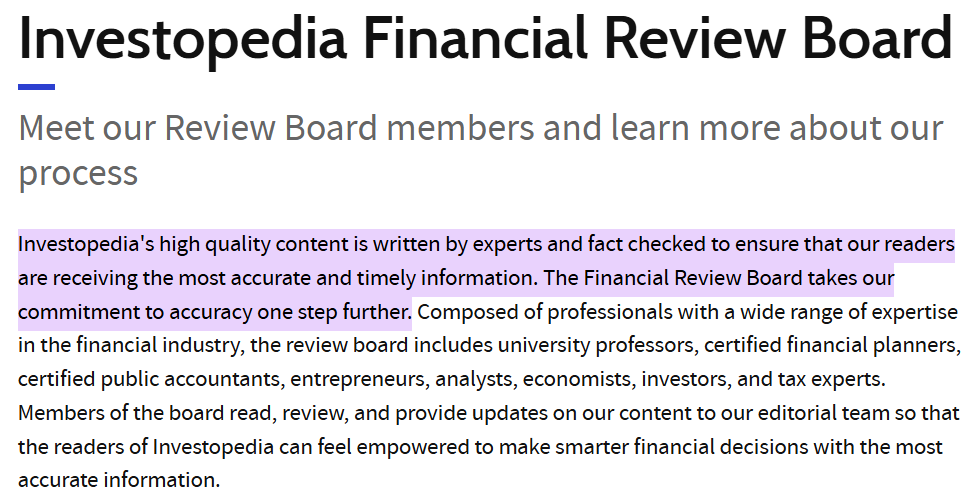
****

## **Investopedia**

* **Source:** [**https://www.investopedia.com**](https://www.investopedia.com/robots.txt)
* **Legality:**
* **Link:** [**https://www.investopedia.com/robots.txt**](https://www.investopedia.com/robots.txt)
* **Complete Blocking:** Specific bots (e.g., 008, Bytespider, PetalBot, Yandex) are not allowed to crawl any part of the site.
* **Partial Crawling:** Bots like Googlebot-News, BingBot, Applebot, and others have restricted access to certain directories and must follow crawl delays.
* **General Bots:** Most bots can crawl unrestricted sections of the site while avoiding the disallowed directories.



* Reliability:
* **Investopedia** is highly reliable for:
* Learning financial concepts.
* Building foundational knowledge of investing, personal finance, and economics.
* Using tools for financial planning and decision-making.

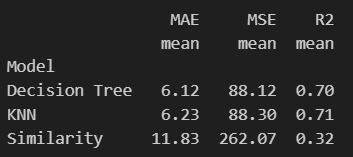


# **Data Preprocessing**

## **Handling missing data**

### **Numerical features**

* Due to complexity of other categorical features such as: Directors, Cast, Studios, we only use available numerical features to build models to fill missing data.
* **Selected features:** 'Tomatoes CriticScore', 'Tomatoes UserScore', 'Metascore', 'Meta UserScore'.
* First, we test with 3 approaches: Similarity Based Technique, KNN Imputer and Decision Tree Regressor.
* **Similarity Based Technique**: Missing values in the ratings matrix are filled by leveraging similarity scores, which are calculated based on the inverse of the mean absolute difference between users and critics' ratings. This ensures that missing values are imputed based on the preferences of users with similar rating patterns.
* **KNN Imputer:** The KNNImputer works by finding the k-nearest neighbors (based on a specified distance metric) for the data points with missing values. It then imputes the missing values using the mean or median (depending on the specified strategy) of the neighboring data points (If all features are null, we will fill mean of each column for that row).
* **Decision Tree Regressor:** Decision trees handle missing data by either ignoring instances with missing values, imputing them using statistical measures, or creating separate branches. During prediction, the tree follows the training strategy, applying imputation or navigating a dedicated branch for instances with missing data.
* **Evaluation:**



KNN and Decision Tree are performing similarly in terms of MAE, MSE, and R². Both models are significantly better than the Similarity model.

However, **KNN would likely be the better choice** because:

* Movie ratings are typically continuous, with values influenced by the preferences of users/critics (neighbors). KNN works well for this kind of problem because it relies on the similarity between users/critics, which is particularly useful in collaborative filtering tasks like movie rating prediction.
* KNN can predict the missing ratings by considering the ratings of similar users or movies, which is a natural fit for this type of task.

### **Categorical features**

* **Genres, Directors and Cast**: Since the unique values of these columns are too many, we will adapt similarity-based technique.
* **In terms of Genres:**
* Takes a movie row and finds n\_similar (default=5) movies based on: Cast match (if available), Director match (if available)
* If not enough similar movies are found:
  + Expands search to include all movies with same Director
  + Removes duplicates from the combined results
* Falls back to random sampling if no attributes match
* **In terms of Directors:**
* Takes a movie row and finds n\_similar (default=5) movies based on: Cast intersection (at least one matching actor), Exact Genre match
* If not enough similar movies are found:
  + Expands search to include all movies in the same Genre.
  + Removes duplicates from the combined results.
* **In terms of Cast:**
* Takes a movie row and finds n\_similar (default=5) movies based on: Director match (if available) andGenre match (if available).
* If not enough similar movies are found:
  + Expands search to include all movies with same Director.
  + Removes duplicates from the combined results
* Falls back to random sampling if no attributes match.
* **Studio**: Using hybrid approach combining **similarity-based approach** and **Random Forest Classifier:**
* **Similarity-Based Method:**
  + This approach tries to find movies that are similar to the one with the missing Studio value based on shared attributes (Cast, Director, Genre, and Rating).
  + For each movie with a missing Studio, we find other movies in the dataset with matching attributes. The most frequent Studio among these similar movies is used to fill in the missing value.
* **Machine Learning Model (Random Forest):**
  + When no suitable similar movies can be found, we fall back on using a trained machine learning model to predict the Studio.
  + The model is trained on known records where the Studio is not missing. We use the Cast, Director, Genre, and Rating as features for predicting the Studio of a movie.
* **Rating**: Using the same hybrid approach as Studio:
* First fill the missing data with similarity-based approach
* If the similarity-based approach failed, it falls back on **Random Forest**
* **Release Date:** Based on the popularity of genres, we will fill the similar month and year for movies that have the similar genres:
* To determine which years are most popular for each genre, we group the dataset by `Genre` and `Release Year` and count how many movies were released each year for each genre.
* For each movie with a missing release year, we attempt to predict the year based on the genre’s popularity trend. If the genre has a trend for multiple years, the year with the highest movie count is chosen as the predicted year.
* If a genre doesn’t have sufficient data or trends (e.g., a genre with no available release year data), we use the median year of all movies as a fallback prediction.
* **Platform Released, Link and Runtime**: We will remove as they are irrelevant for our data exploration and runtime.

## **Data Standardization**

### **Numerical features**

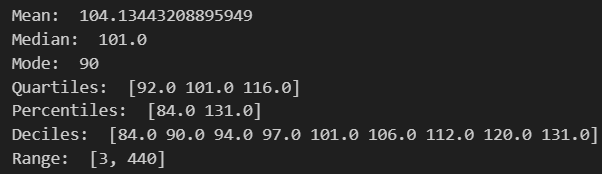
* At first, numerical features are all object, therefore we transformed it into float type since only float can contain nan.

### **Categorical features**

* **Casts, Directors and Genres:** Because these features have so many unique values, we will split it and put it into a list.
* **Ratings**: We converted it into category of Pandas.
* **Runtime**: Converted into a dictionary {‘hours’, ‘minutes’, seconds’}
* **Release Date:** Converted into datetime type standard of Pandas ‘%Y-%M-%D’.

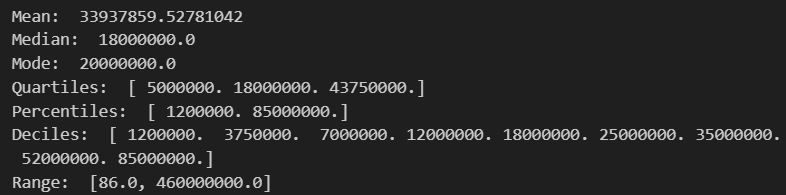
## **Outliers Survey**

### **Numerical features**

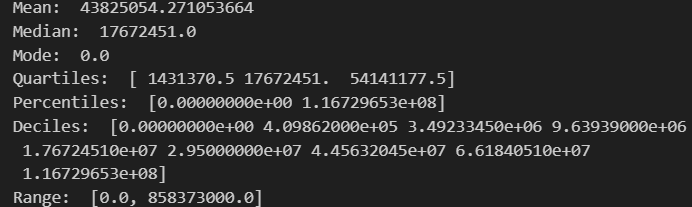
* **Runtime:**
* **Lowerbound** = Q1 - 1.5\*IQR = 56
* **Upperbound** = Q3 +1.5\*IQR = 152
* **Range** = [3,440]

**=> There are outliers: Values less than 56.0 (e.g., 3).**

**Values greater than 152.0 (e.g., 440).**

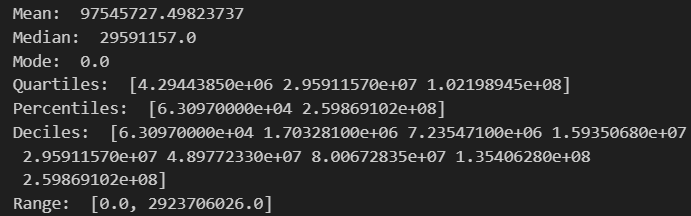
* **Production Budget:**
* **Lowerbound** = −52125000 **(negative, so no lower bound here)**
* **Upperbound** =102875000
* **Range** = [86, 460000000]

**=> There are outliers: Any value greater than 102875000 is an outlier.**

* **Domestic Gross:**
* **Lowerbound** = −77624340 **(negative, so no lower bound here)**
* **Upperbound** = 133197381
* **Range** = [0, 858373000]

**=> There are significant outliers in feature. These are likely the very large values (e.g 858373000) that exceed the IQR upper bound of 133197381.**

* **Worldwide Gross:**

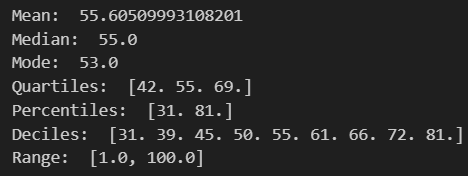
****

* **Lowerbound** = −140563923.25 (negative, so no lower bound here)
* **Upperbound** = 249830411.25
* **Range** = [0, 2923706026]

**=> The feature contains significant outliers. These outliers are very high values (e.g., 2923706026) that fall well outside the IQR upper bound of**

**249830411.25. The dataset has a long tail with many low values (including 0.0) and a few extremely high values skewing the mean upward.**

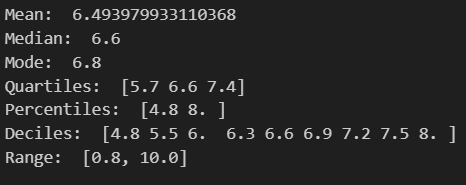
* **Metascore:**

****

* **Lowerbound** = 1.5
* **Upperbound** =109.5
* **Range** = [1, 100]

**=> This feature appears to have no significant outliers. It is symmetrically distributed around the mean, with values spread evenly between the minimum and maximum**

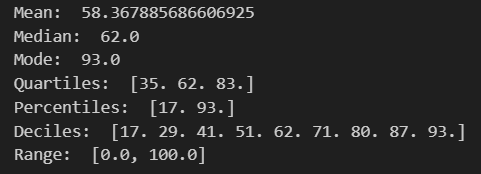
* **Meta Userscore:**

****

* **Lowerbound** = 3.15
* **Upperbound** = 9.95
* **Range** = [0.8, 10]

**=> The feature otherwise appears to be symmetrically distributed with a central tendency around 6.5, as evidenced by the alignment of the mean, median, and mode.**

* **Tomatoes CriticScore:**

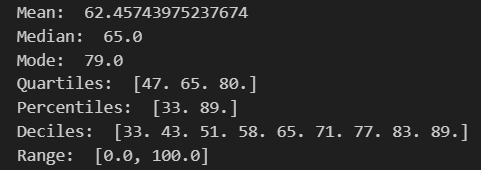
****

* **Lowerbound** = -37 (negative, so no lower bound here)
* **Upperbound** = 155
* **Range** = [0, 100]

**=> No outliers detected based on the IQR method.**

**The feature appears to have a slight left skew, as indicated by the mean being slightly lower than the median. This skewness is minimal and does not indicate extreme asymmetry.**

* **Tomatoes UserScore:**

****

* **Lowerbound** = -2.5 (negative, so no lower bound here)
* **Upperbound** = 129.5
* **Range** = [0, 100]

**=> The feature appears to be well within expected boundaries, and there are no extreme outliers.**

**The slight left skew (mean < median) suggests that the feature might have a few lower values that are pulling the mean down, but no significant outliers are present.**

### **Categorical features**

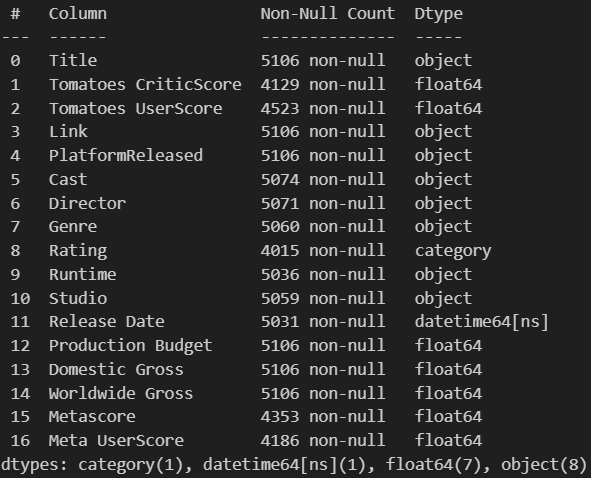
* There are no outliers method for these features.

## **Synthesis in schools of using multiple sources**

* First, we will take data from The Numbers as a base data, since we have planned to model data to predict on economic features
* Second, we will take the name of each movie, normalize it and put it into the template API of Metacritic and Rotten Tomatoes:
* Metacritic: [https://www.metacritic.com/movie/{movie](https://www.metacritic.com/movie/%7Bmovie)\_name}
* Rotten Tomatoes: [https://www.rottentomatoes.com/m/{movie\_name](https://www.rottentomatoes.com/m/%7Bmovie_name)}
* After successfully collecting all necessary schools of datasets, the scores varying between 2 sources of each movie are not too wide.
* Finally, we complete the dataset with 3 different sources: Rotten Tomatoes, Metacritic and The Numbers.

## **Data Consistency**

### **Data Type**

****

* All types of features are converted into the right type of their standard types.

### **Data Duplicate**

* There are no duplicated rows in the dataset

### **Meaning of each column**

|  |  |
| --- | --- |
| **Column** | **Meaning** |
| Title | The film title |
| CriticScore and UserScore | Percentage-based scores from critics and users, respectively |
| Link | URLs to movie pages |
| PlatformReleased | Indicates release platforms, such as Cinema. |
| Cast and Director | Names of the main cast members and director |
| Genre, Rating, and Runtime | Film genre, content rating (e.g., PG, R), and runtime |
| Studio | The studio responsible for production or distribution |
| Release Date | Specific release date |
| Production Budget, Domestic Gross, and Worldwide Gross | Financial details in terms of budget and revenue |
| Metascore and Userscore | Average scores given by critics and users on Metacritic website |

### **Meaning of each row**

* Each line of the dataset is a record of a movie, and all records include various features of the movie like Name, Genres, Cast, Budget, Revenue, Runtime, etc

# **Data Preprocessing + Exploration**

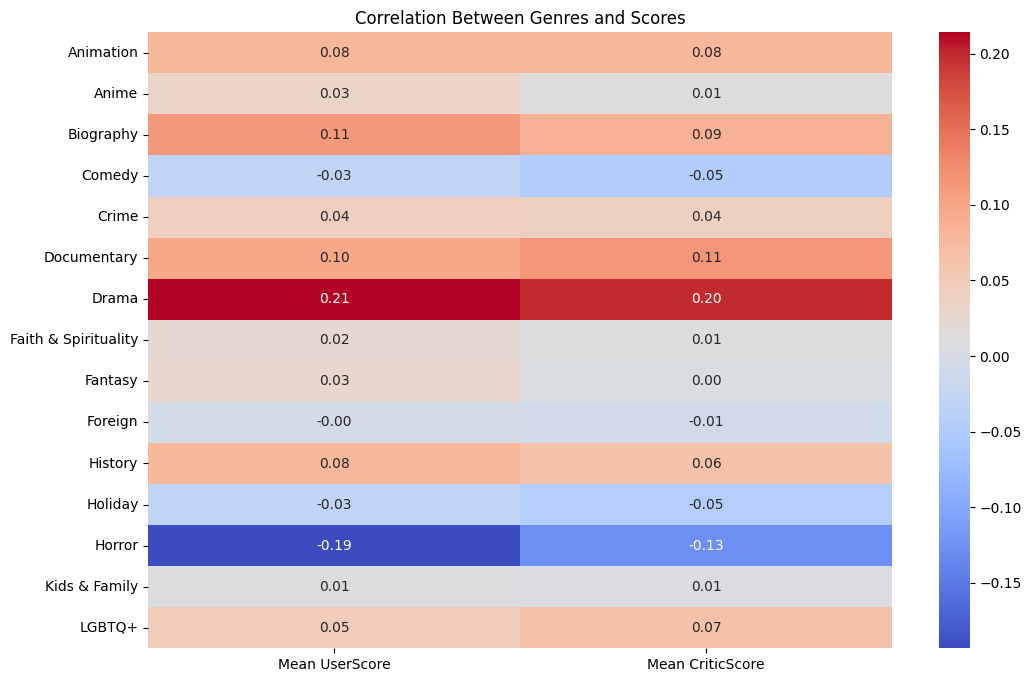
From the dataset that we collected, we proposed the problem of exploring and analyzing movie data to uncover insights into factors affecting movie performance, including ratings, revenue, and critical reception.

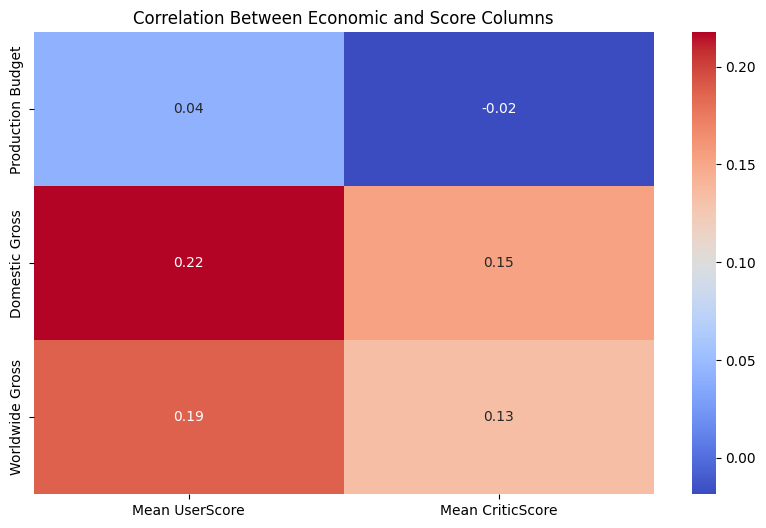
## **Question 1**

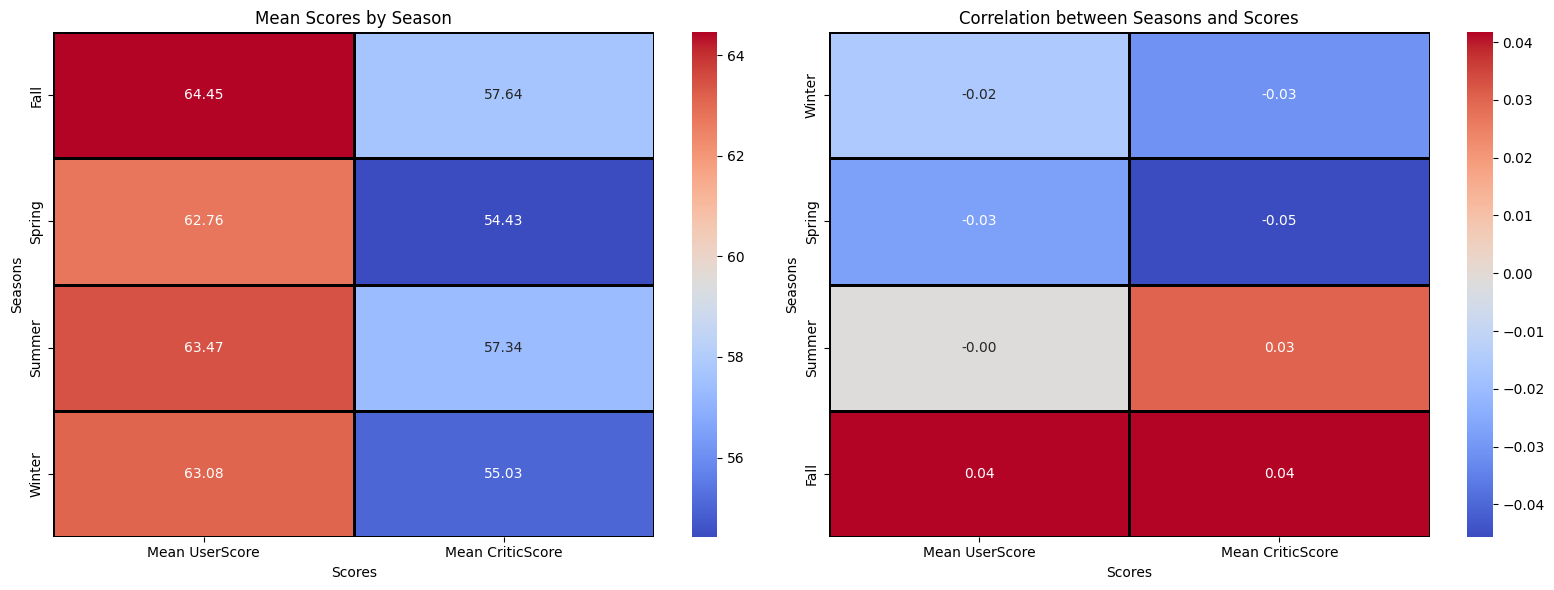
### **Content**

* What are the primary factors influencing critic and user scores?

### **Result (Based on chart)**







### **Comment and conclusion**

* Revenue and Genre are key factors: Higher-grossing movies and specific genres like drama receive better scores.
* Horror films generally under-perform in both audience and critic ratings.
* Older movies have a nostalgic or critical edge in scoring.
* Seasonal releases, especially in Fall, tend to fare better critically and commercially.

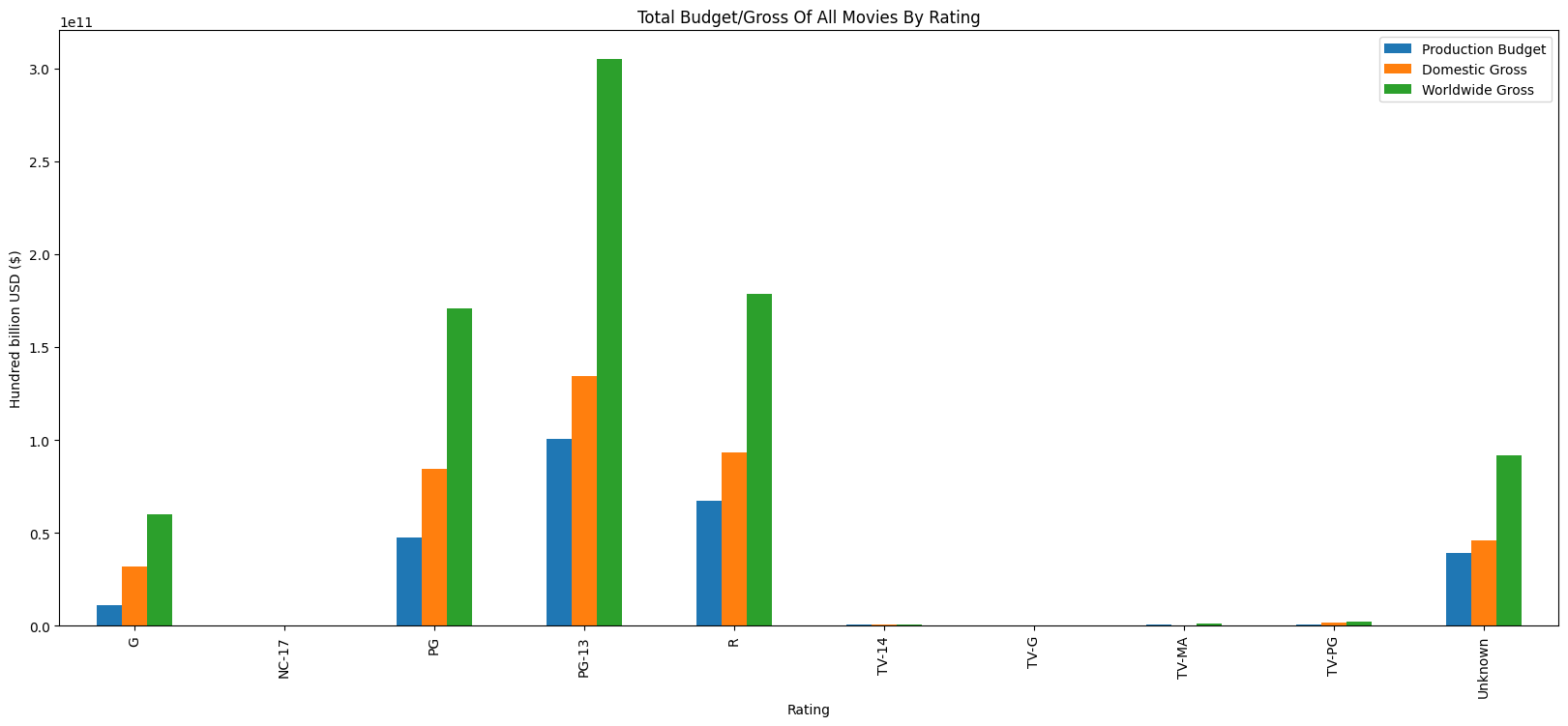
**=> Strategic release timing and effective marketing are crucial for improving reception.**

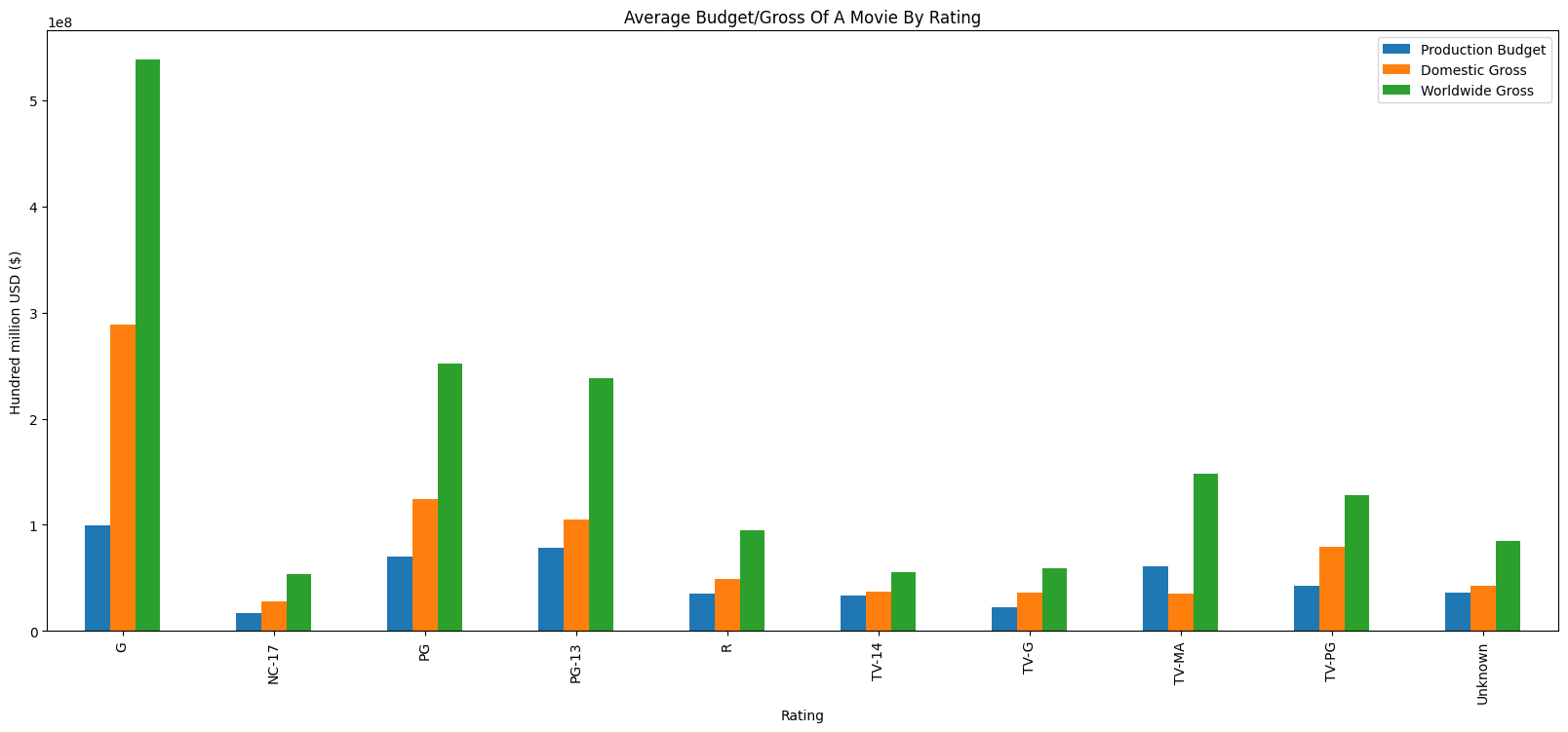
## **Question 2**

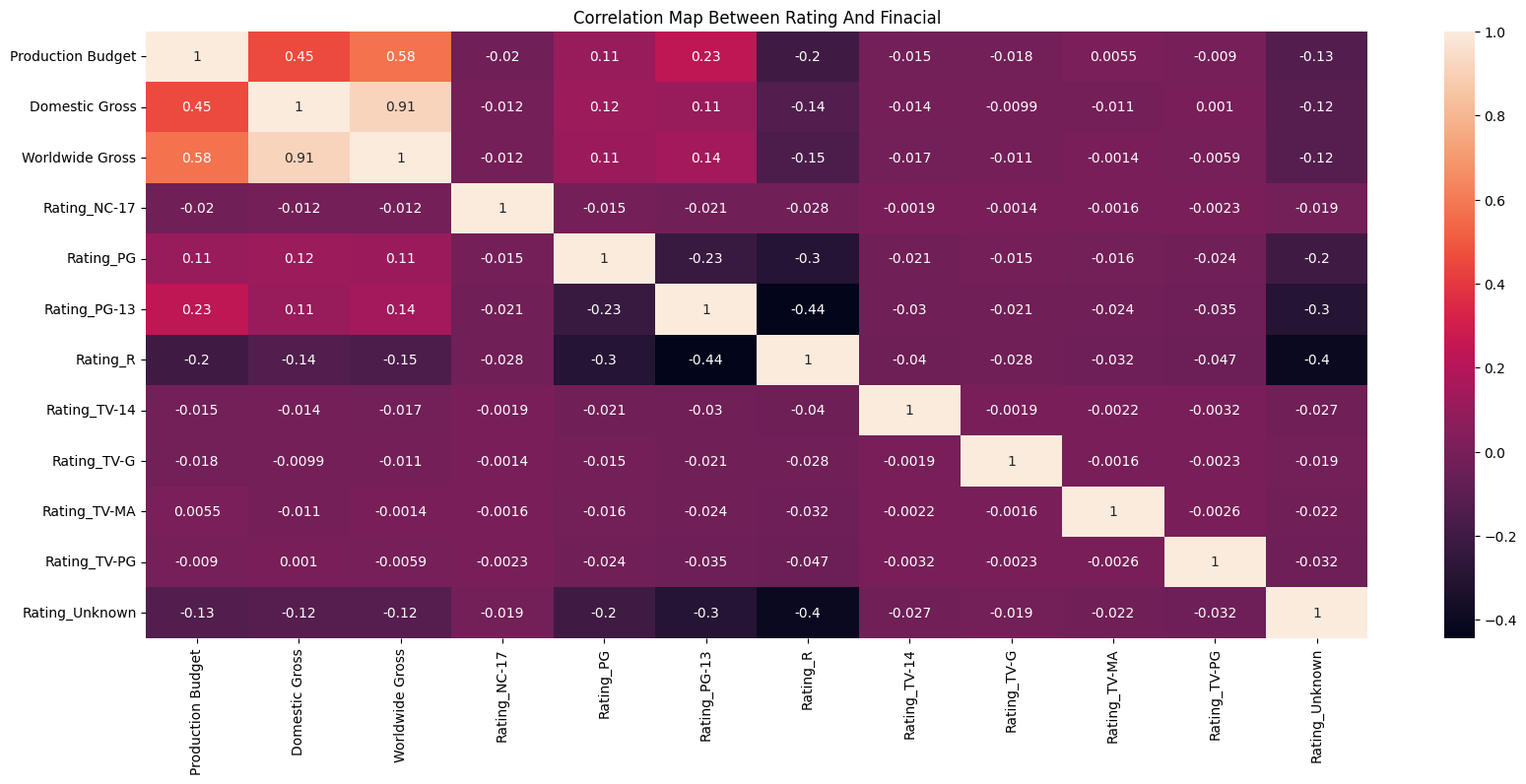
### **Content**

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

### **Result (Based on chart)**







### **Comment and conclusion**

* PG-13 rated movies tend to have the highest Production Budgets and box office Grosses, benefiting from broad audience appeal, extensive marketing, and a balance of accessibility and thrilling content.
* R and G rated movies can perform very well but also require a high Production Budget.
* On the other hand, G rated films, with their universal appeal to all age groups, often have the best return on investment, benefiting from lower production costs, strong family and global appeal, and extensive merchandising opportunities.

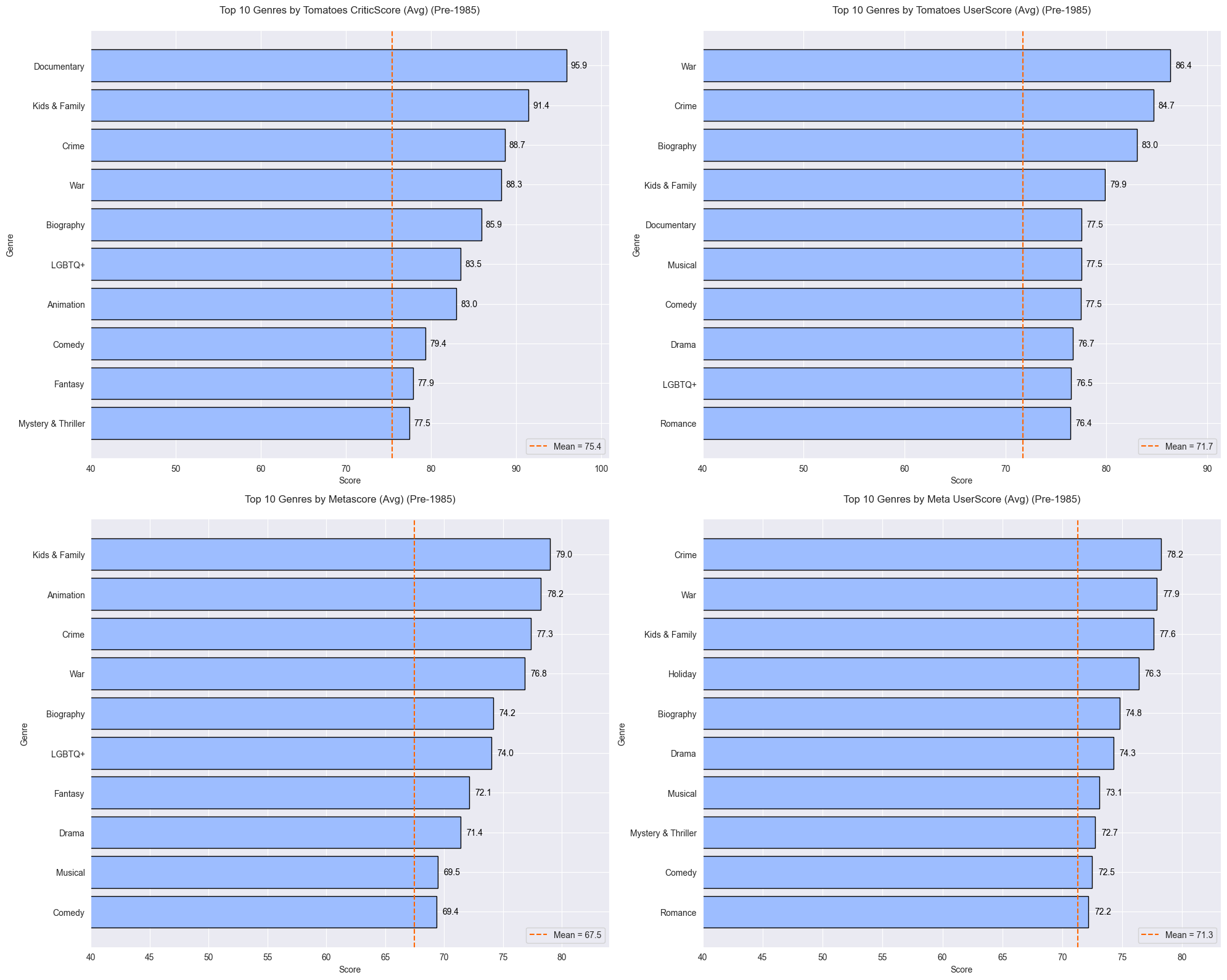
**=> Overall, PG-13 movies dominate in financial success, but G and PG films also prove to be financially viable, especially when factoring in long-term revenue from licensing and merchandise.**

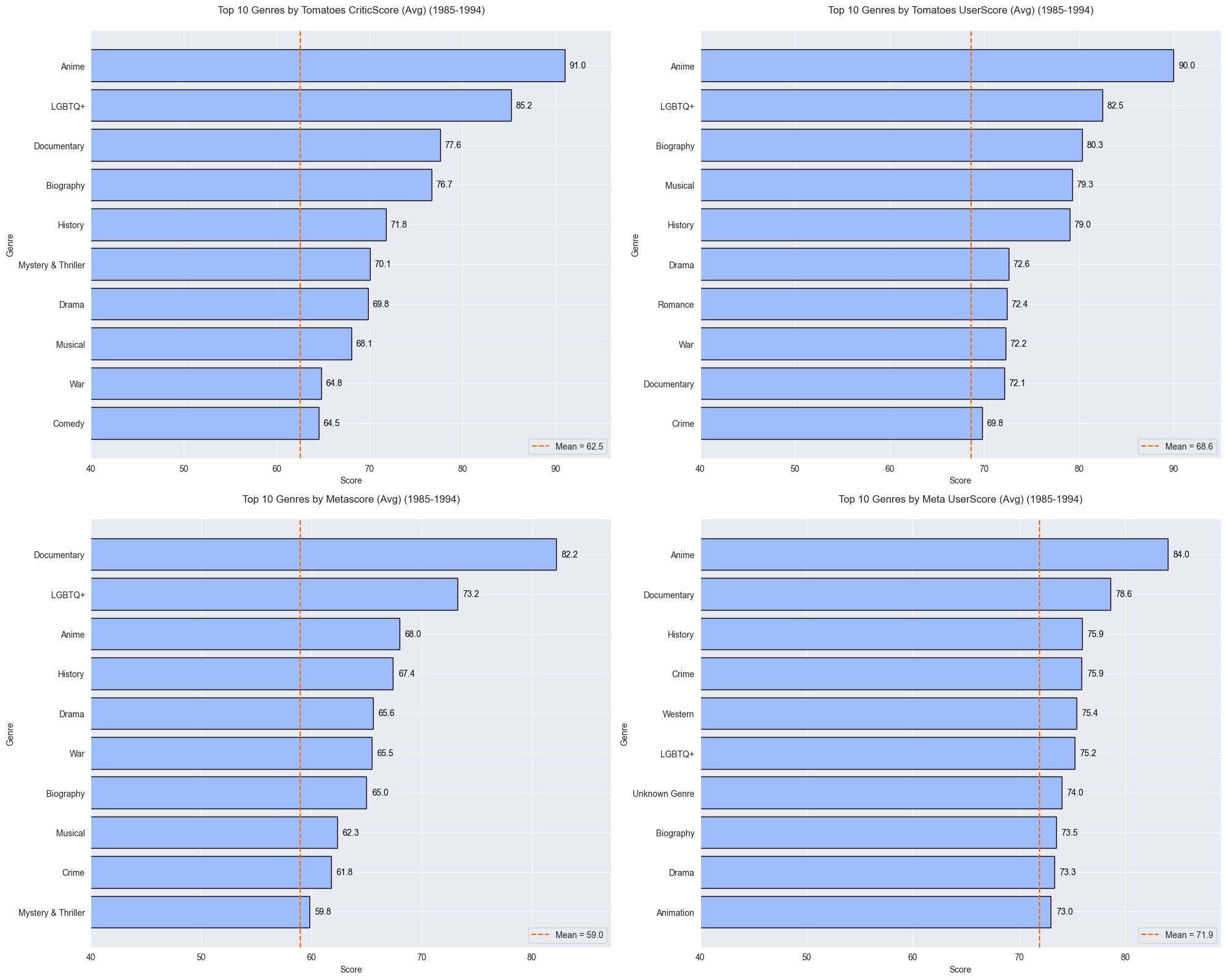
## **Question 3**

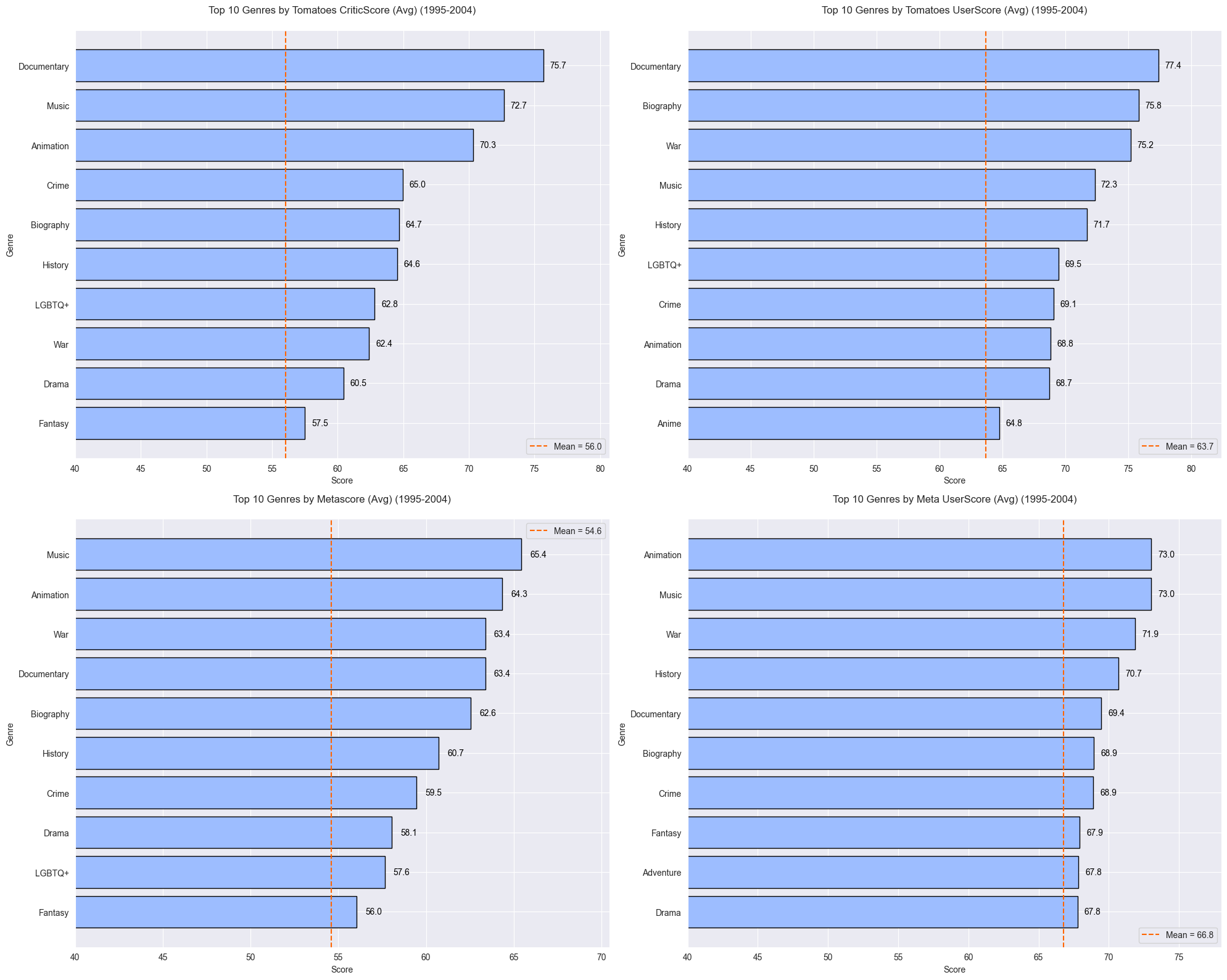
### **Content**

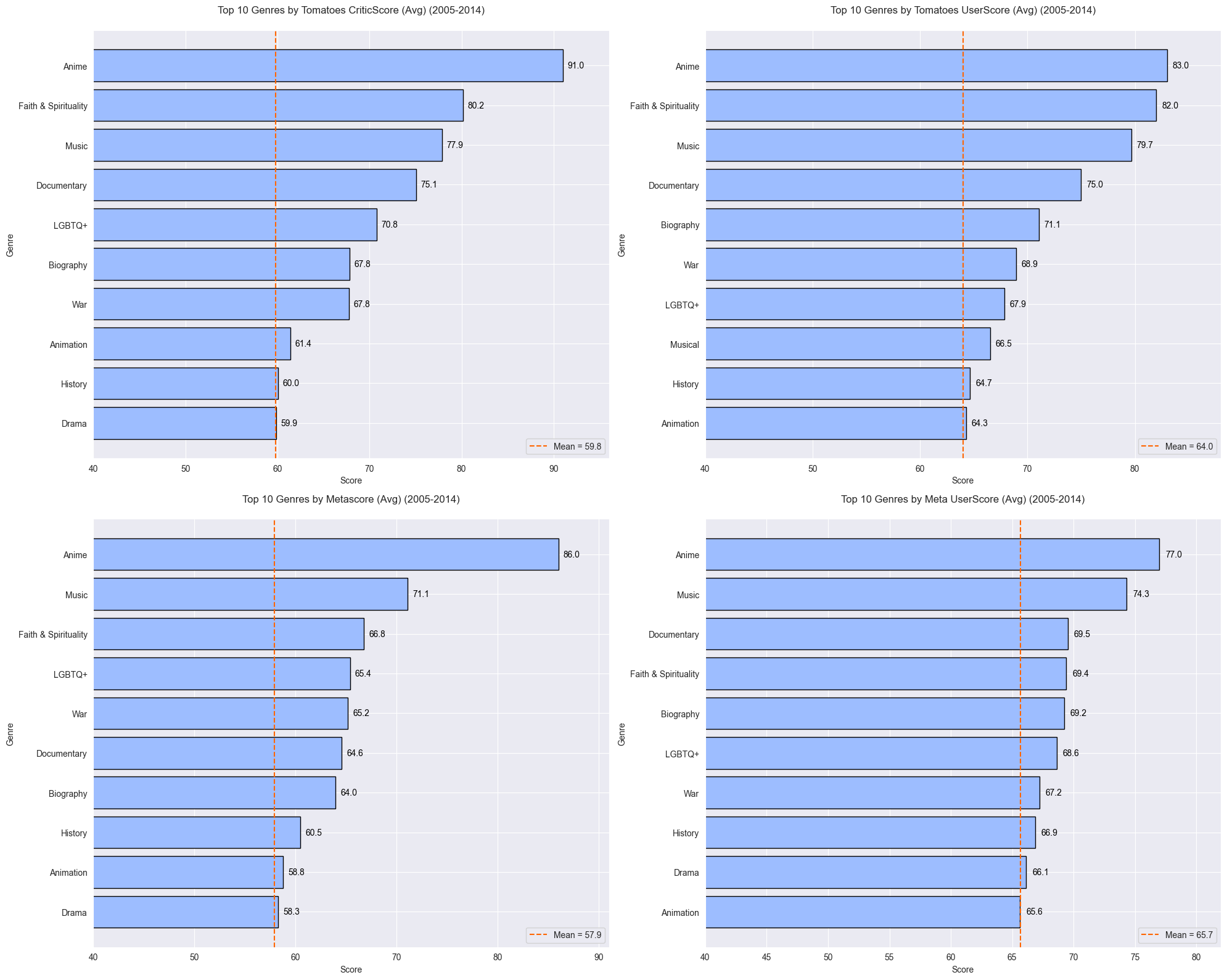
* What genres tend to perform better in terms of critic and user scores?

### **Result (Based on chart)**









### **Comment and conclusion**

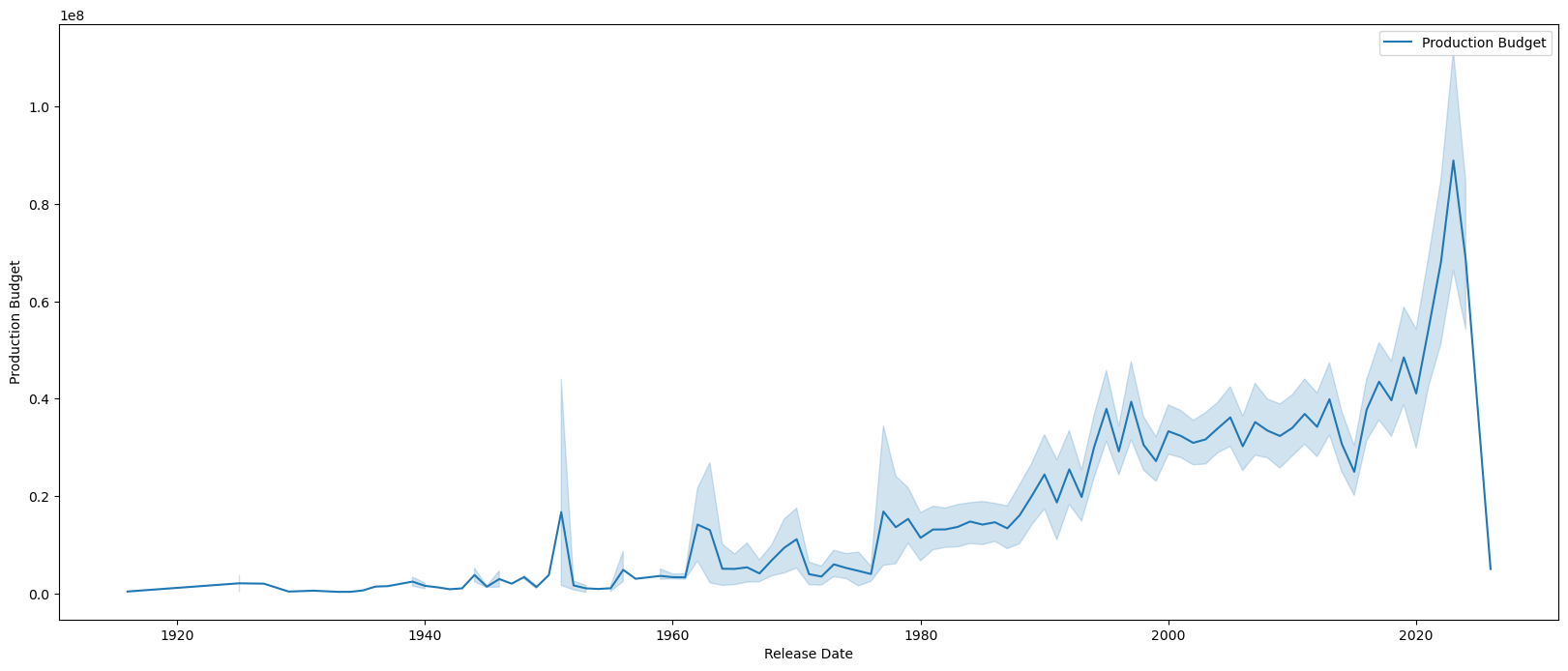
* **Pre-1985:** This era marks a time when Documentary, Kids & Family, War, and Crime movies thrived and were highly acclaimed.
* 1985 - 1994: The Anime genre achieved its dominance for the first time, while the strong growth of the LGBTQ+ genre began reshaping societal perspectives on sexual orientation.
* **1995 - 2004:** The Documentary genre maintained its steady prominence, and was accompanied by the rapid rise of emerging genres like Music and Animation.
* **2005 - 2014:** Anime reclaimed its dominance after a decade of losing the number one position, alongside the remarkable rise of emerging genres like Music and Faith & Spirituality.
* **2015 - 2024:** Anime maintained its dominance, accompanied by the sustainable development of the Music genre. Additionally, critics expressed a strong preference for the LGBTQ+ genre, while users showed a clear fondness for the Faith & Spirituality genre.

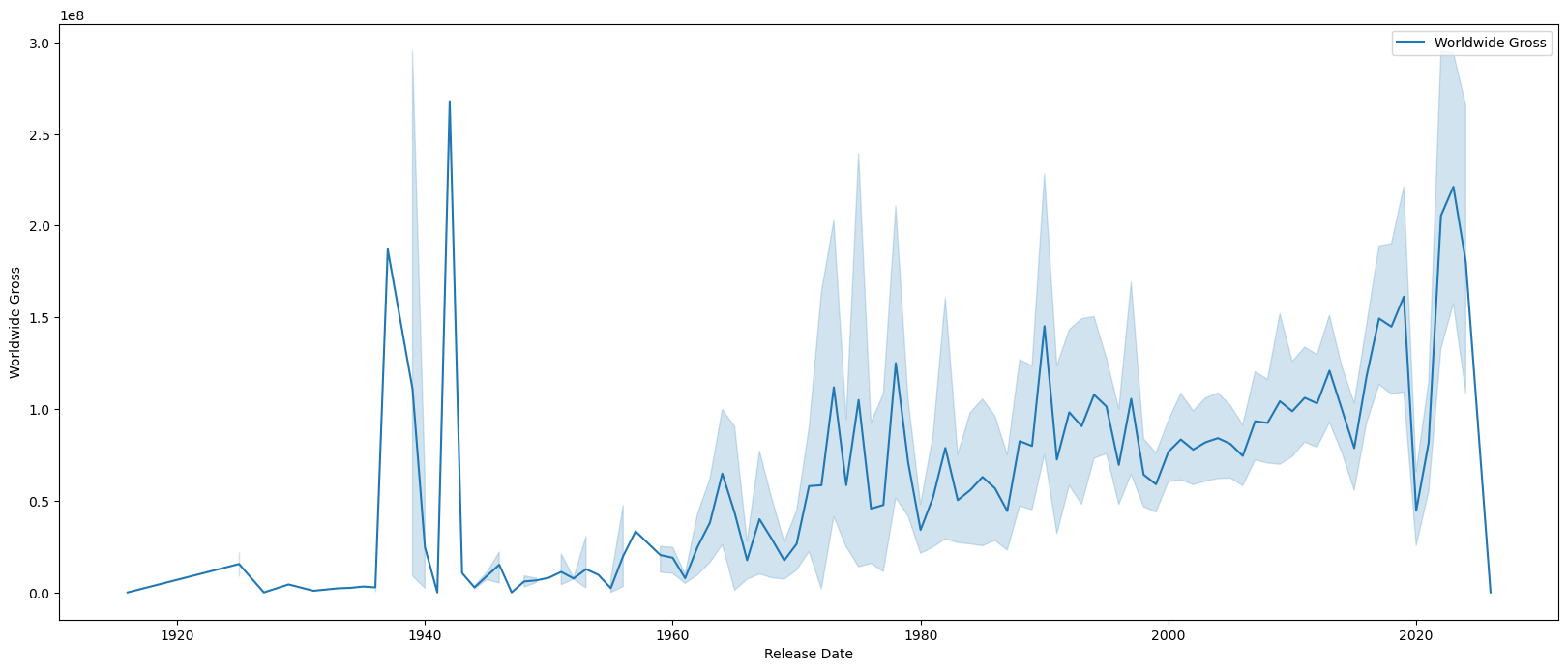
## **Question 4**

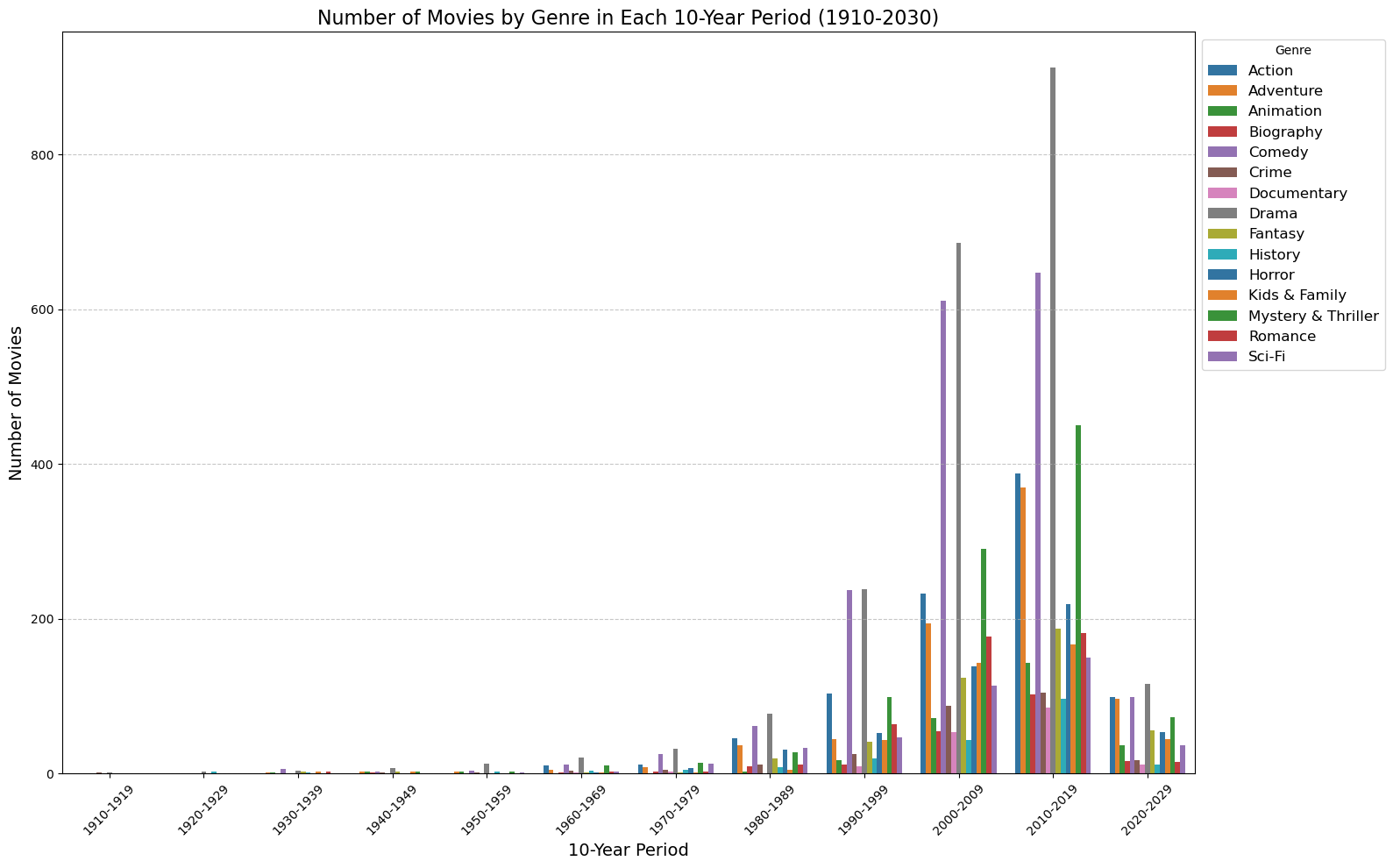
### **Content**

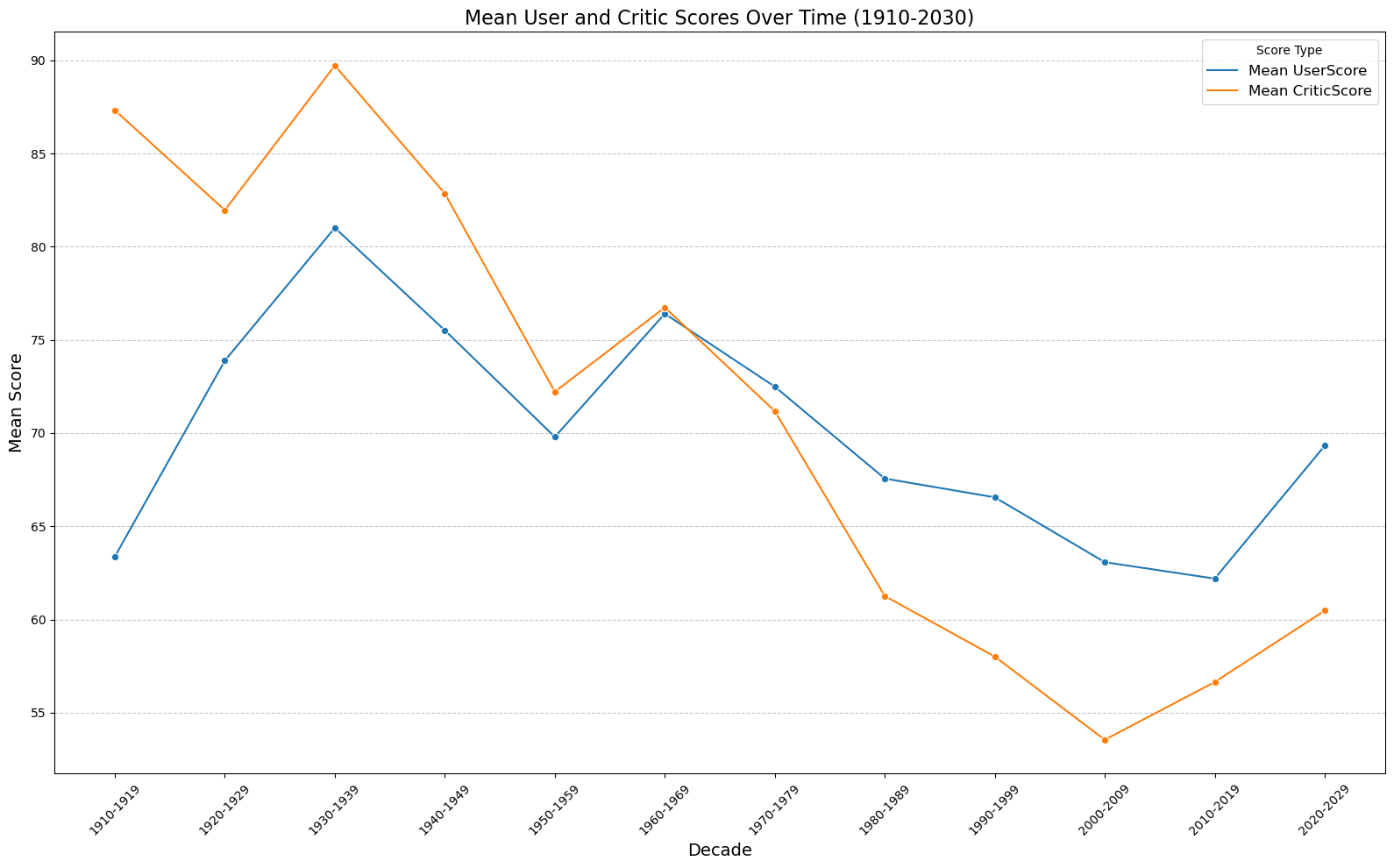
* Are there trends in movie production and performance over time?

### **Result (Based on chart)**









### **Comment and conclusion**

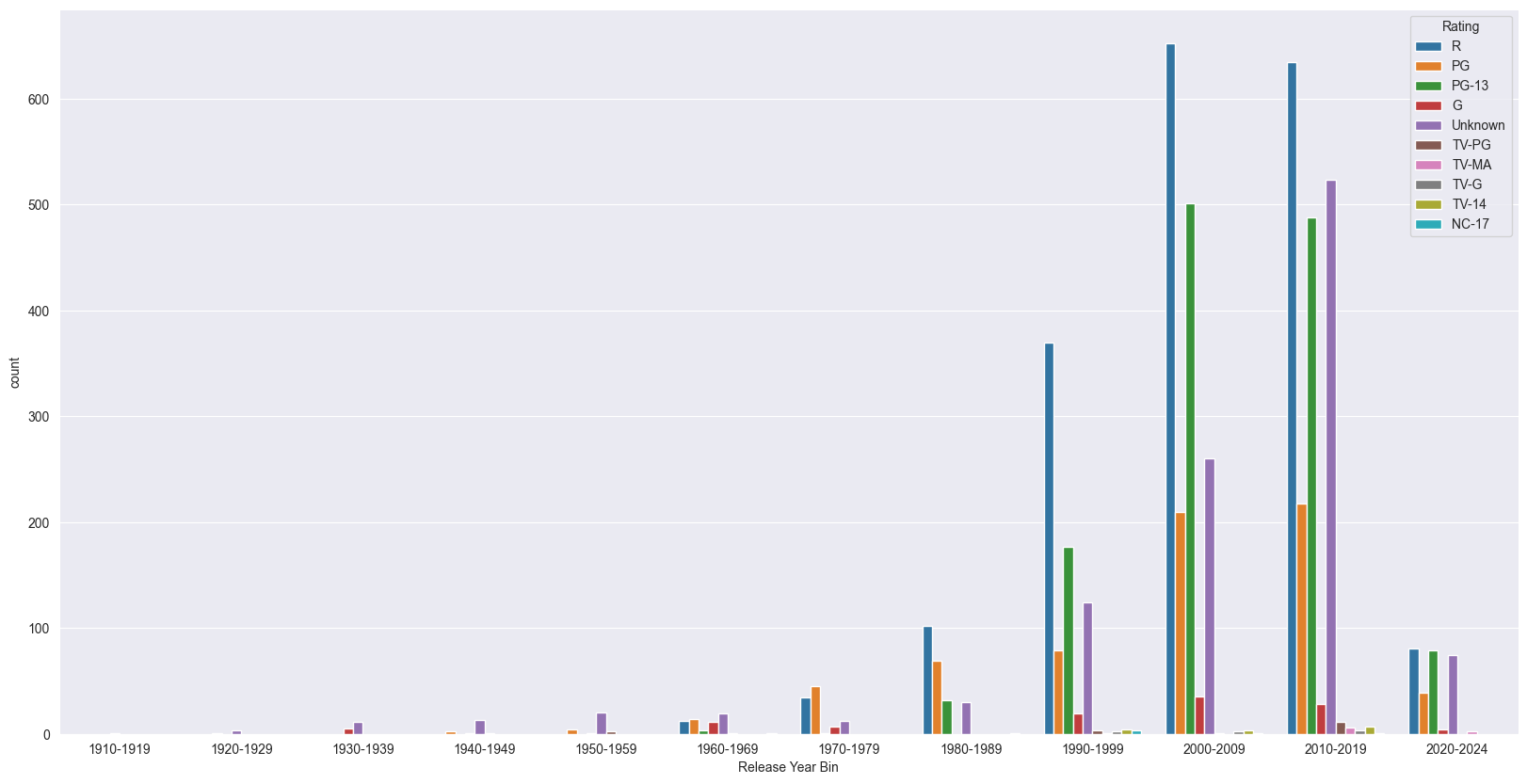
* **Financial aspect insights**:
* Higher production budgets are strongly correlated with increased worldwide gross, indicating that greater investment often leads to higher revenue, likely due to improved quality, marketing, and distribution.
* Post-1990, advancements in technology and the rise of blockbuster franchises have further driven growth in both budgets and revenue.
* **Genre-Specific insights:**
* Action and Adventure remain top genres, reflecting audience demand for high-stakes, visually dynamic stories, while Comedy consistently caters to the desire for light-hearted content.
* Animation and Family saw rapid growth in the late 1990s and 2000s, highlighting their appeal to family audiences.
* Sci-Fi and Fantasy experienced significant post-2000 growth, driven by technological advancements and cultural trends.
* Documentaries, though produced in smaller numbers, gained steady interest in the 2000s, reflecting a growing appetite for real stories and social issues.
* **Critic and audience ratings trend:**
* The gap between User and Critic scores has widened over time, especially during the 1980s and 1990s, diverging from the early alignment seen in the 1950s.
* Recent decades show an upward trend in User Scores, suggesting increased audience satisfaction, possibly influenced by cultural shifts, nostalgia, or evolving filmmaking styles.
* Differences in evaluation criteria between critics and audiences, along with the impact of streaming platforms in the 2000s, have further shaped this critic-audience divide.

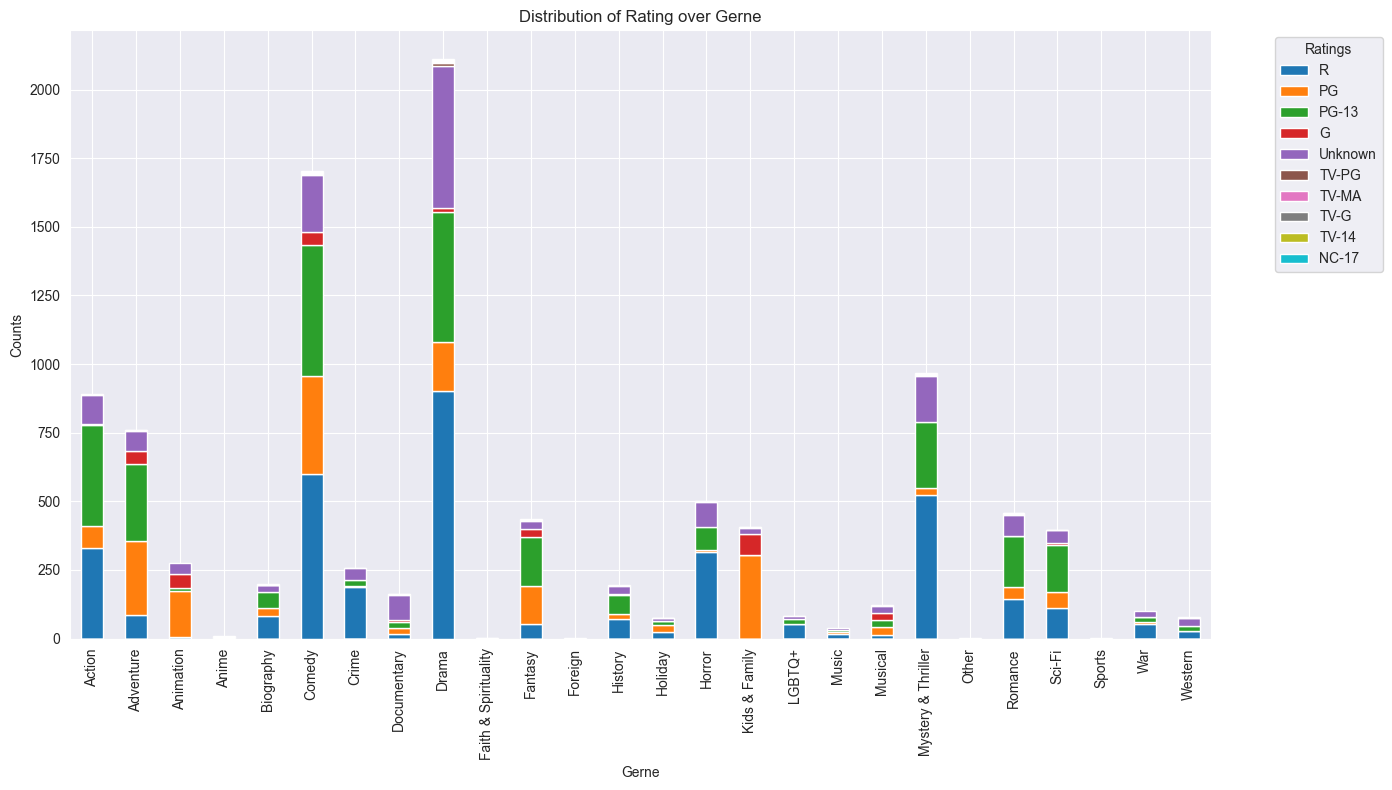
## **Question 5**

### **Content**

* What is the trend of rating movies over time?

### **Result (Based on chart)**





### **Comment and conclusion**

* The trend in movie ratings has evolved in response to changing societal norms, audience preferences, and the rise of new genres.
* As film became a more powerful medium for storytelling, the introduction of the Motion Picture Association of America's (MPAA) rating system in 1968 helped provide clearer guidelines for age-appropriate content.
* In the following decades, the popularity of genres like action, drama, and horror led to a rise in R-rated films, particularly as filmmakers sought more freedom to explore mature themes. Meanwhile, PG and PG-13 ratings became more common with the growing emphasis on family-friendly and teen-targeted films.
* To this day, while the number of R-rated films remains significant, there has been a continued rise in PG-13 and PG blockbusters, especially in the superhero and fantasy genres, which strike a balance between mature storytelling and wider audience appeal.

# **Data Modelling**

## **Which models are used?**

Here are the list of models used for this project:

|  |  |
| --- | --- |
| **Model** | **Description** |
| Decision Tree | A decision supports recursive partitioning structure that uses a tree-like model of decisions and their possible consequences, including [chance](https://en.wikipedia.org/wiki/Probability) event outcomes, resource costs, and [utility](https://en.wikipedia.org/wiki/Utility). |
| Random Forest | An ensemble learning method for classification, regression and other tasks that works by creating a multitude of decision trees during training. |
| Gradient Boosting | A very powerful machine learning algorithm, often wins many Kaggle competitions. |
| XGBoots | Extreme Gradient Boosting, based on Gradient Boosting but with huge improvement |

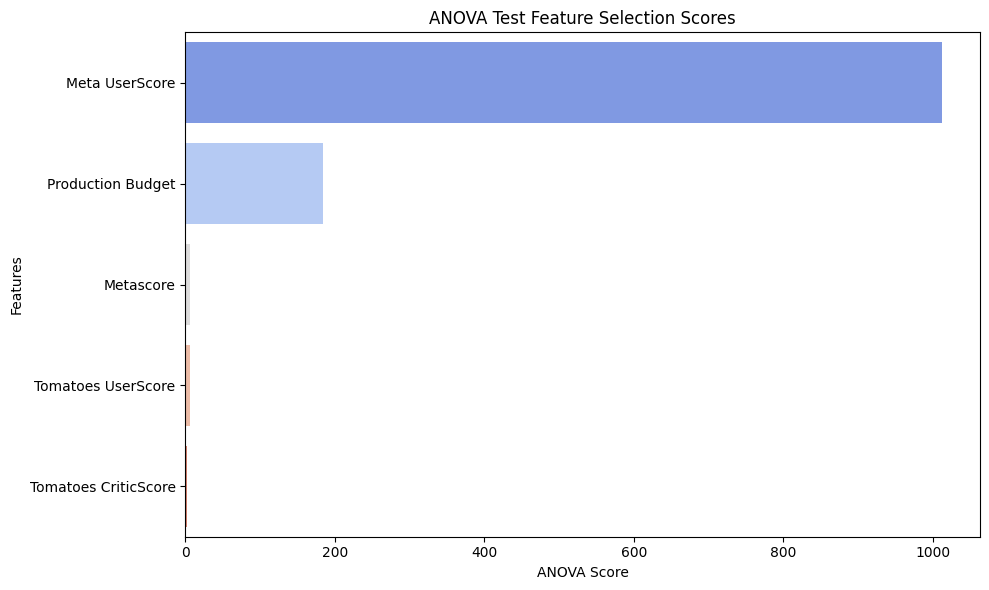
We also use GridSearchCV, a tool from the scikit-learn library used for hyperparameter tuning in machine learning. It essentially automates the process of finding the optimal combination of hyperparameters for a given machine learning model.

## **Features Selection**

### **For numerical features**

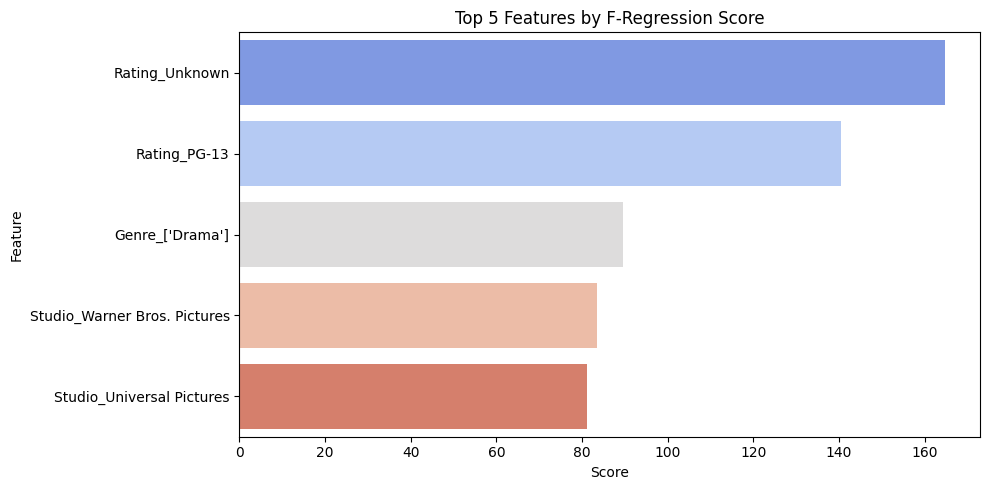
* Anova Test for numerical feature selection: The ANOVA (Analysis of Variance) test is typically used for feature selection in classification problems where the target variable is categorical. However, ANOVA concepts can still be applied in regression problems, albeit in a slightly different context.

We concluded that these features (**Meta Userscore, Production Budget, Metascore, Tomatoes UserScore and Tomatoes CriticScore**) will be all included into our models.

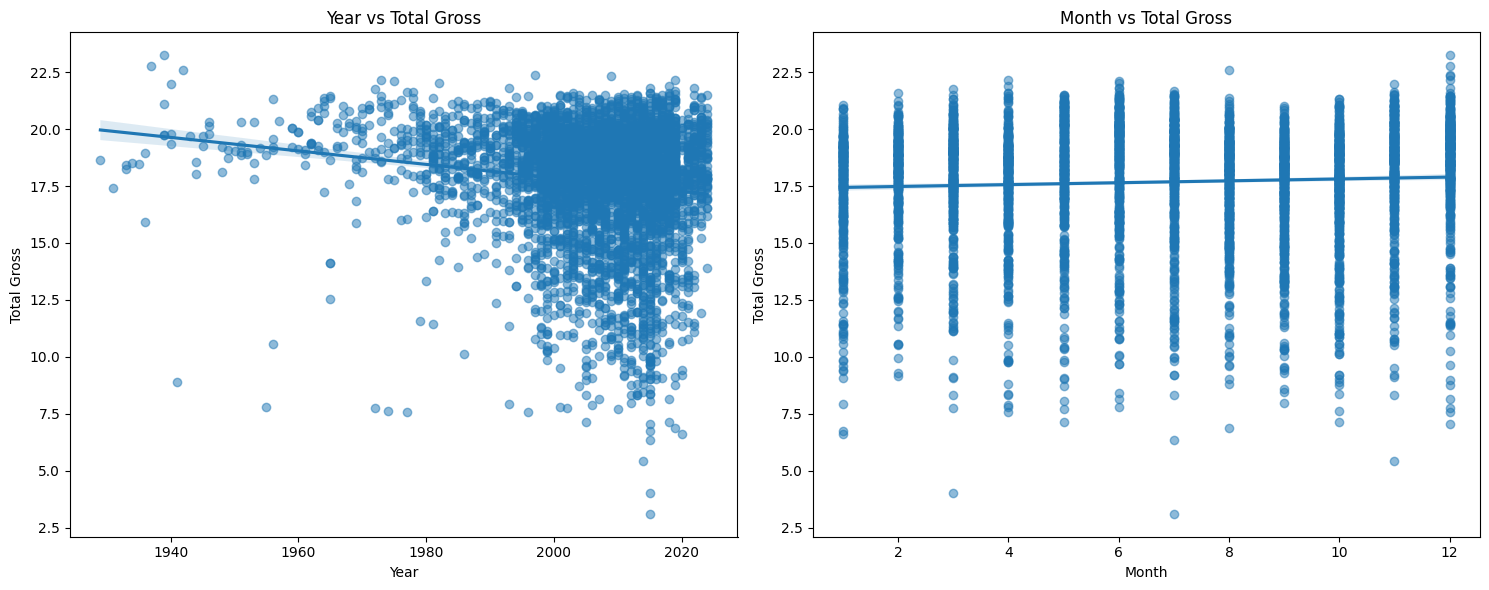


### **For categorical columns**

* As can be seen from the plot, some members of **Genre**, **Rating ,**and **Studio** show up in the top 5 F-Regression Score, therefore we will need to include these columns for modeling.



* **Time data:** Although the correlation between Year and Month and Total Gross is weak, we still keep it because both features are relevant for the movie industry. But we need to modify it into bin to reduce potential noise while preserving temporal information.



## **Data Preprocessing**

Preprocessing is a critical step in preparing the dataset for machine learning models. It ensures that data is clean, consistent, and in a format suitable for analysis.

### **For numerical features**

* + Apply logarithm transformation for money values: Applying logarithm for **Production Budget** and **Worldwide Gross** plays an important role in training the Machine learning model. It helps reduce the scale, and skewness and stabilize the variance of the money data.

### **For categorical features**

* **Genre** is reformatted from a string representation of a list to an actual list of genres. Then, one-hot encoding is applied to each unique genre
* **Rating** is one-hot encoded using OneHotEncoder.
* **Studio** is one-hot encoded for the top n\_studio studios (Which movie produced by top studio (has the top highest number of movies produced) is labeled 1 \), with other studios grouped into 'Studio Other'.
* **Director** and **Cast** are one-hot encoded as the same for **Studio**, although we don't use this feature, we still leave there if needed in future.
* **Year** and **Month** are binned into 10 years and 3 months respectively, to catch the trend of movies.

1. **Train/test/valid split.**

* Train set size: 80%.
* Test set size: 20%.
* In this project we don’t split the validation set, instead we will use cross\_val\_score on train set to evaluate the model.

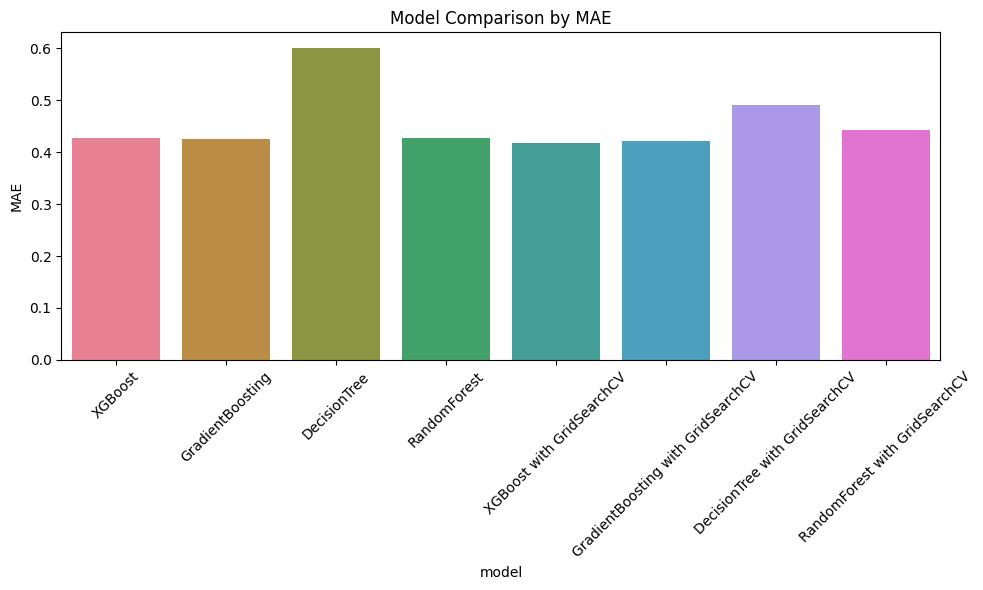
## **Model Evaluation**

### **Introduction to metrics used**

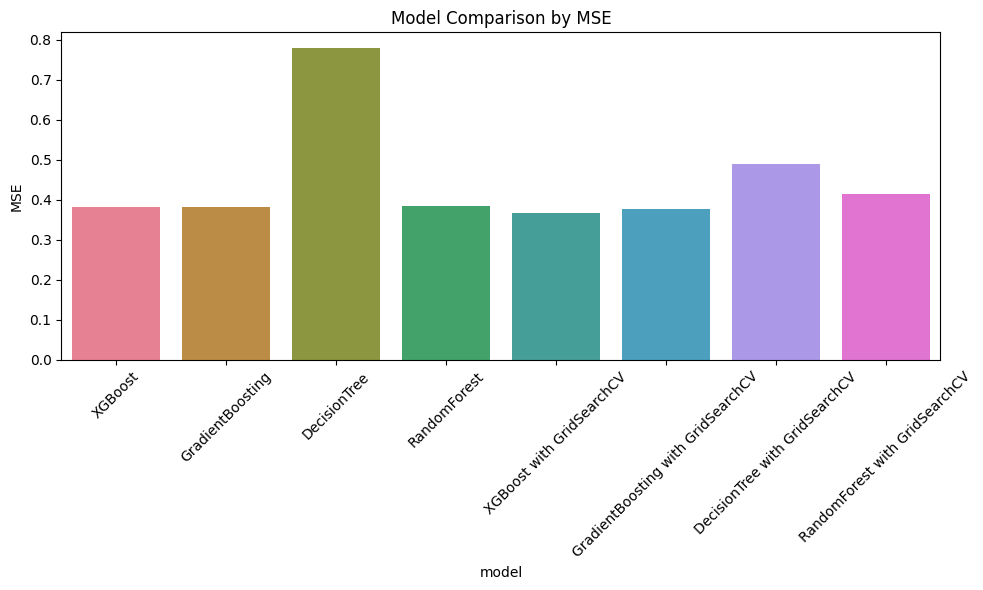
|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Description** | **Advantage** | **Disadvantage** |
| **Mean Absolute Error (MAE)** | Measures the average magnitude of the errors in a set of predictions, without considering their direction. | - Easy to interpret as it represents the average error in the same units as the target variable.  - Less sensitive to outliers compared to MSE. | - Not differentiable at zero, which can make it difficult to use with some optimization algorithms.  - Treats all errors equally, which might not be desirable in some contexts. |
| **Mean Squared Error (MSE)** | Measures the average of the squares of the errors. It is more sensitive to outliers than MAE. | - Differentiable, making it suitable for gradient-based optimization algorithms.  - Penalizes larger errors more heavily, which can be useful if larger errors are particularly undesirable. | - More sensitive to outliers because it squares the errors, which can disproportionately affect the metric.  - The units of MSE are the square of the target variable's units, which can make it less interpretable. |
| **R-squared (R2)** | Represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. | - Represents the proportion of variance explained by the model, which is easy to interpret.  - Useful for comparing the goodness-of-fit of different models. | - Can be misleading for non-linear models.  - High R2 does not necessarily mean a good model; it can be high due to overfitting.  - Can be negative if the model performs worse than a horizontal line (mean of the target variable). |

### **Comparison**

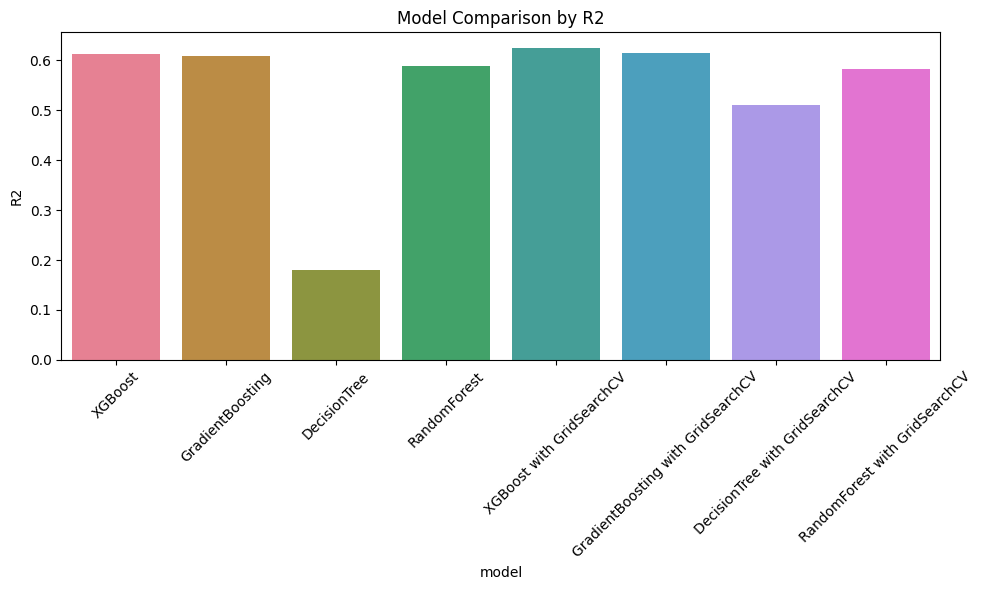
* **Model comparison by MAE :**

****

* **Model comparison by MSE:**



* **Model comparison by R2 Score:**



* **Key insights:**
* DecisionTree performs poorly across all metrics, standing out as a weak performer.
* Ensemble models (XGBoost, GradientBoosting, RandomForest) are the top choices, with slight variations between tuned and untuned versions.
* GridSearchCV versions of XGBoost, GradientBoosting, and RandomForest do not show a dramatic improvement over their default counterparts. However, their performance is consistent.
* Consistency: Ensemble models outperform across all metrics, demonstrating their strength in handling complex patterns.

## **Results and which is the best model**

### **Results**

* Best Performers: XGBoost and GradientBoosting, both tuned and untuned, are the top models due to their tight clustering and alignment with diagonal.
* Weakest Performer: DecisionTree (both tuned and untuned) shows significant scatter and poor alignment, making it unsuitable for this dataset.
* Impact of Tuning: GridSearchCV provides marginal improvements for ensemble models (GradientBoosting and RandomForest) but has minimal effect on XGBoost.

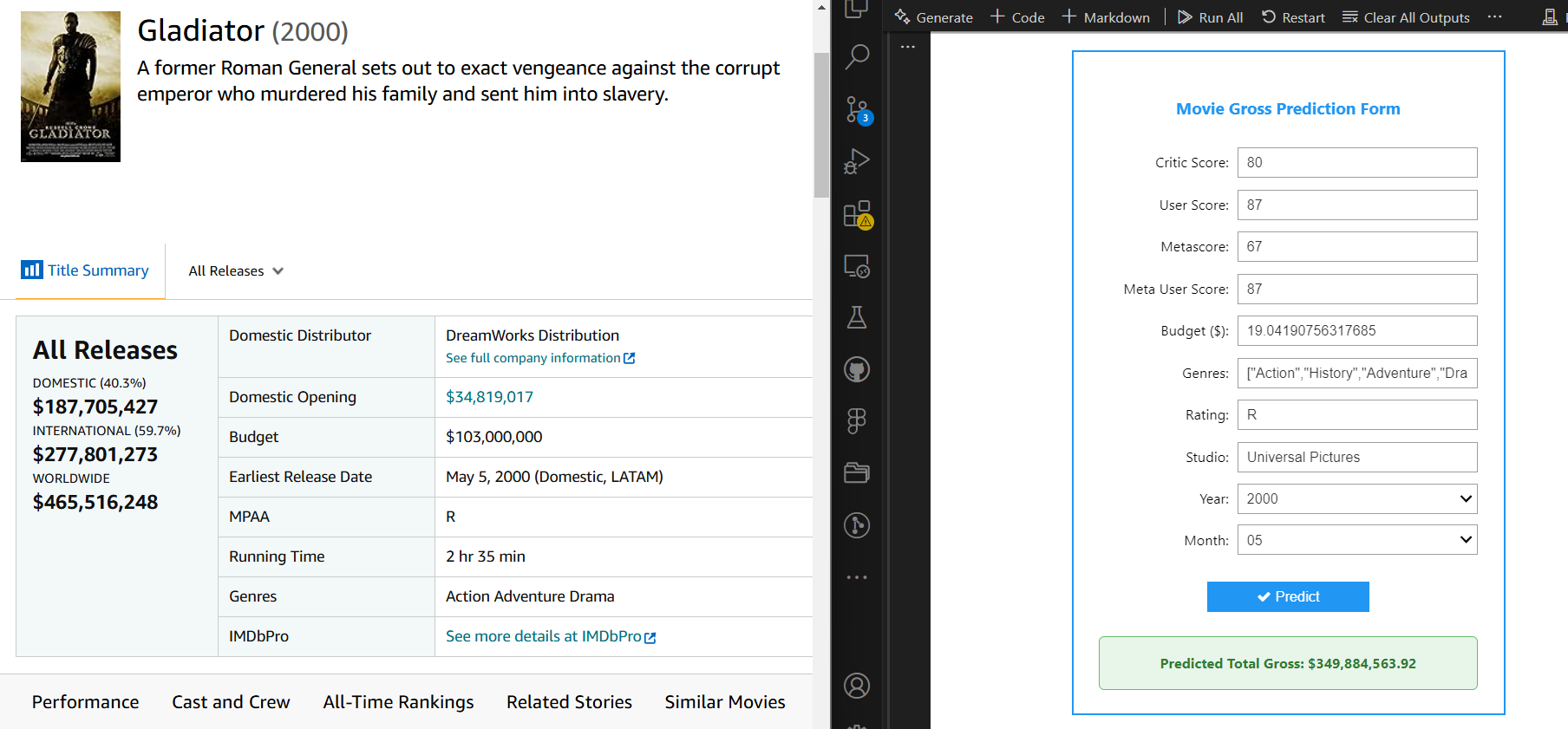
### **Conclusion**

* In comparing GradientBoosting and XGBoost, the performance differences in prediction accuracy are minimal.
* The key distinction lies in training efficiency. XGBoost is almost twice as fast as GradientBoosting, making it significantly more suitable for rapid iterations, larger datasets, or time-sensitive applications.
* Given its comparable prediction accuracy, faster training time, and suitability for deployment in production environments, XGBoost is the preferred choice, especially when efficiency and scalability are critical.

**=> XGBoost is the best model.**

## **Output of the model and meaning**

### **Output**



### **Meaning**

* With the output of the models, as well as the issue in the modeling process, is the predicted Total Gross of a movie. However, since the output is in natural logarithm form, it needs to be exponentiated to obtain the actual monetary values.
* In the example above, we try predicting the gross of Gladiator movie, the gap is about 120 million dollars, this can be because of other factors like social-related factors such as trending and unexpected blockbusters.