



*"Have you ever felt lost in  
the sea of endless movie  
choices, unsure of what  
to watch next?"*







# *ENHANCED CONTENT BASED MOVIE RECOMMENDATION SYSTEM*

TEAM E: LOHITHA PULIJALA  
PRAGNA PULIPATI  
SUKRUTHI YAKKALA  
KAGGLE ID: [LOHITHAP123@]





## *Project Overview:*

- *Development of an advanced content-based movie recommender system focusing on movie keywords; plot overviews, and explicit user preferences to enhance the quality and personal relevance of movie recommendations.*
- *Motivation: The integration of user preferences seeks to bridge the gap between generic content-based recommendations and user-centric suggestions, aiming to boost user engagement and satisfaction.*
- *Research Question:*
- *Can incorporating user preferences into content-based recommendations improve the relevancy and user satisfaction of the recommended movies*





- *Existing Approaches:*• Classic content-based and collaborative filtering methodologies, hybrid models leveraging both user and item metadata.
- *Gap Identified:*• A noticeable lack of personalization and user preference consideration in pure content-based models.
- *Our Contribution:*• Our model seeks to intertwine user preferences with content metadata to render more personalized and user-relevant recommendations





## *DATASET INTRODUCTION:*

*Dataset: TMDB Movie Metadata*  
*Source: TMDB*  
*Movie Metadata on Kaggle*

*([https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata?select=tmdb\\_5000\\_movies.csv](https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata?select=tmdb_5000_movies.csv))*



*Number of Columns: 24*

*•size of dataset:(4809, 24)*

*•Column Names: Budget (int64) | Genres (string) |  
Homepage (string) | ID (int64) | Keywords (string) |  
Original Language (string) | Original Title (string) |  
Overview (string) | Popularity (float64) | Production  
Companies (string) | Production Countries (string) |  
Release Date (string) | Revenue (int64) | Runtime  
(float64) | Spoken Languages (string) | Status (string)  
| Tagline (string) | Title (string) | Vote Average  
(float64) | Vote Count (int64) | UserId (int64) | MovieId  
(int64) | Userrating (float64) | Director (string)*



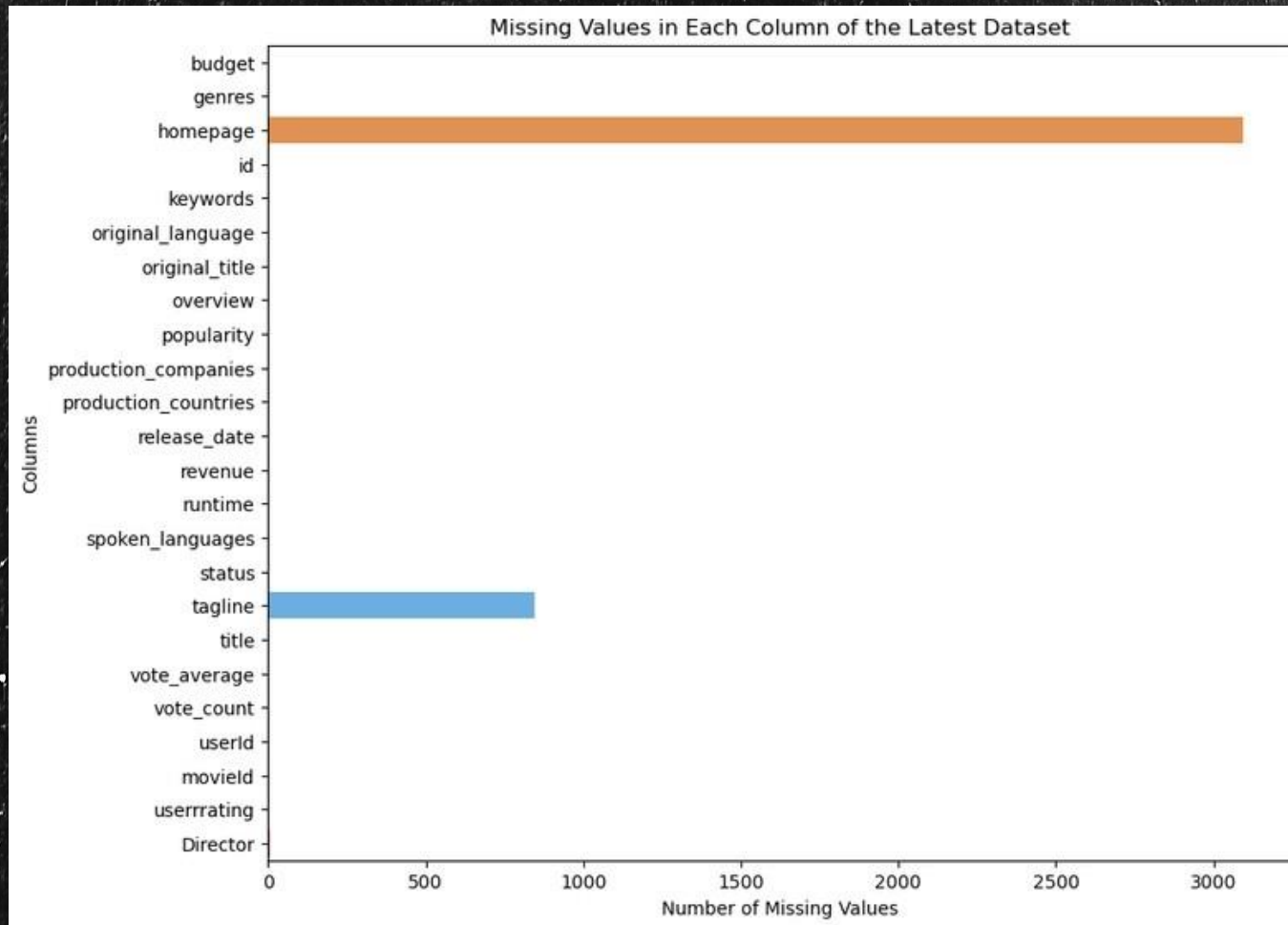
## LITERATURE REVIEW:



- *Kaggle.com: These are some references used in implementing idea*
- *<https://www.kaggle.com/code/ashutosh39/movie-recommendation-using-text-mining>*
- *<https://www.kaggle.com/code/imsakshimittal/movie-recommender-system>*
- *<https://www.kaggle.com/code/yassermessahli/movies-recommender-system-content-based>*
- *Based on these studies, following issues were identified*
- *How to suggest movies when there is no consumer data statistics?*
- *What class of movie attributes can be adopted for the recommender system?*
- *How to compute the likeness between two movies?*
- *How to validate the results we get?*

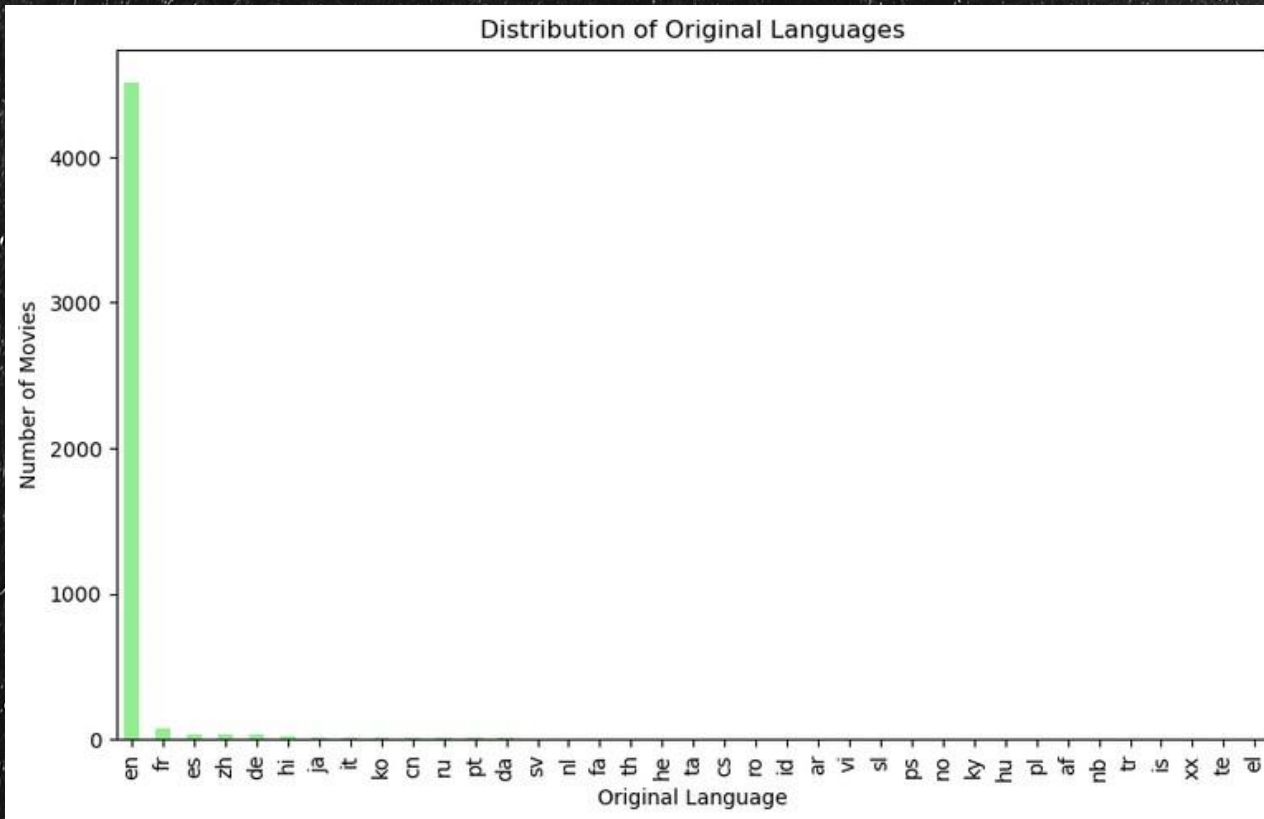


# EXPLORE DATA ANALYSIS



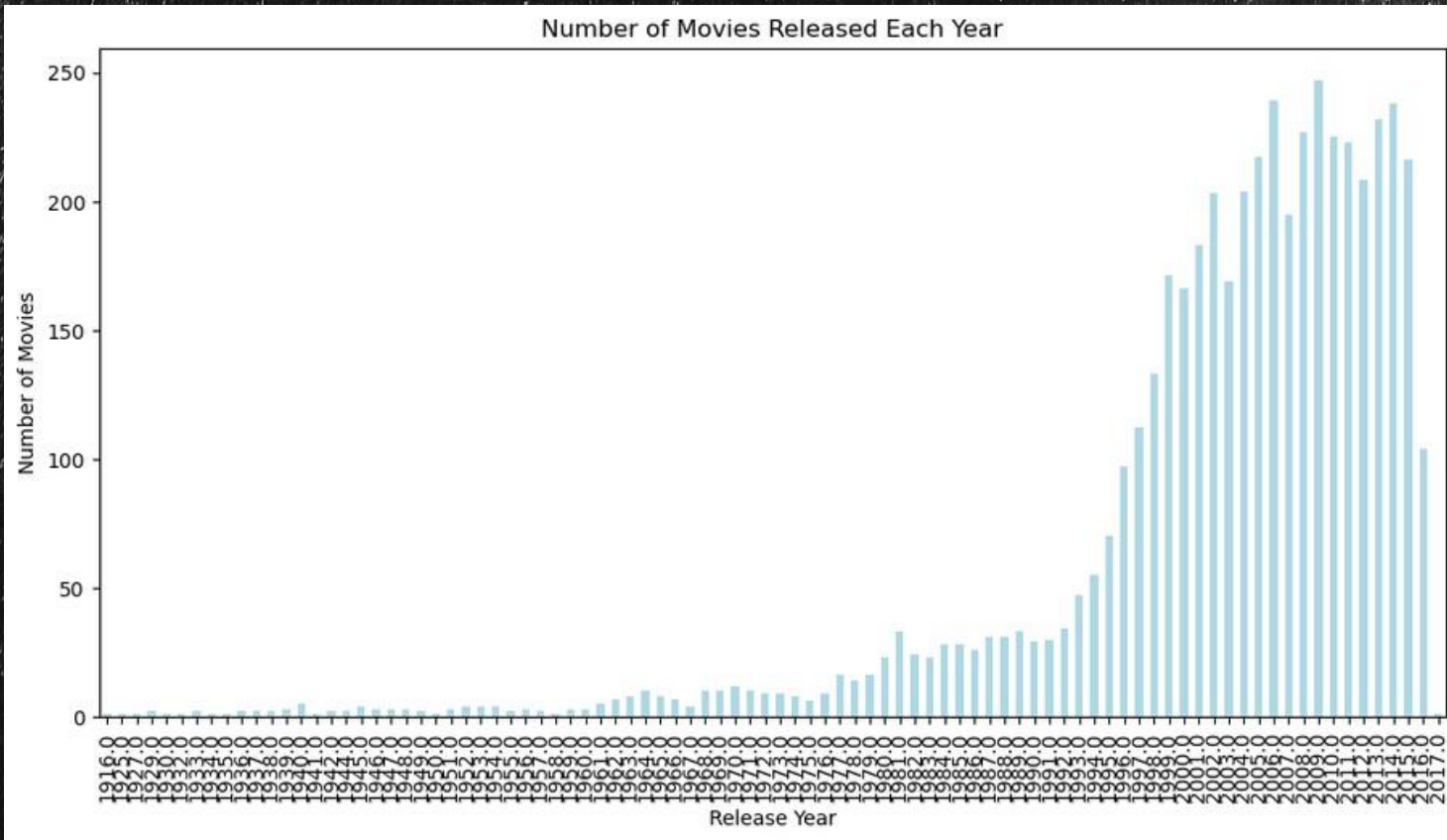


# EXPLORE DATA ANALYSIS



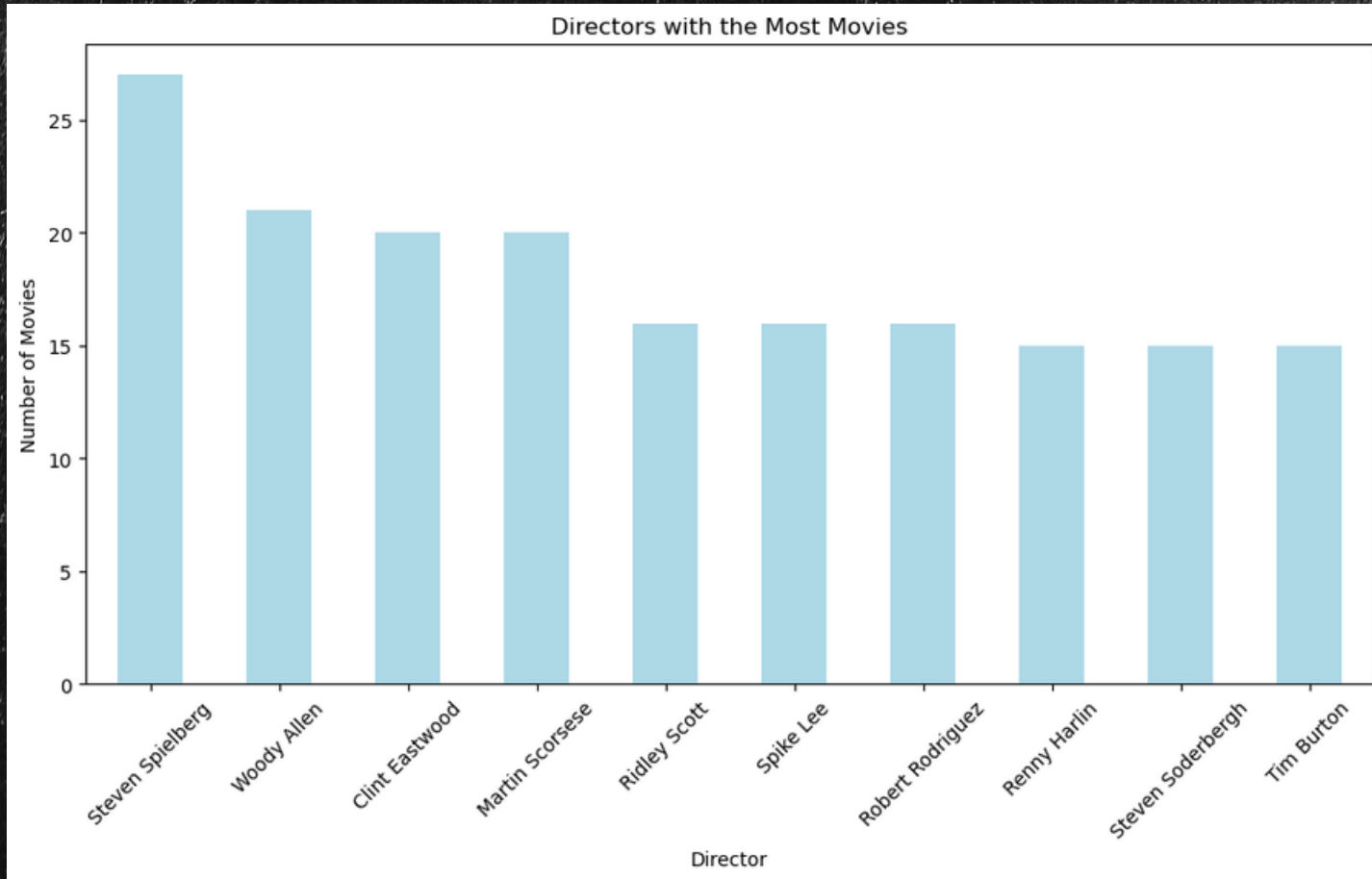


# EXPLORE DATA ANALYSIS



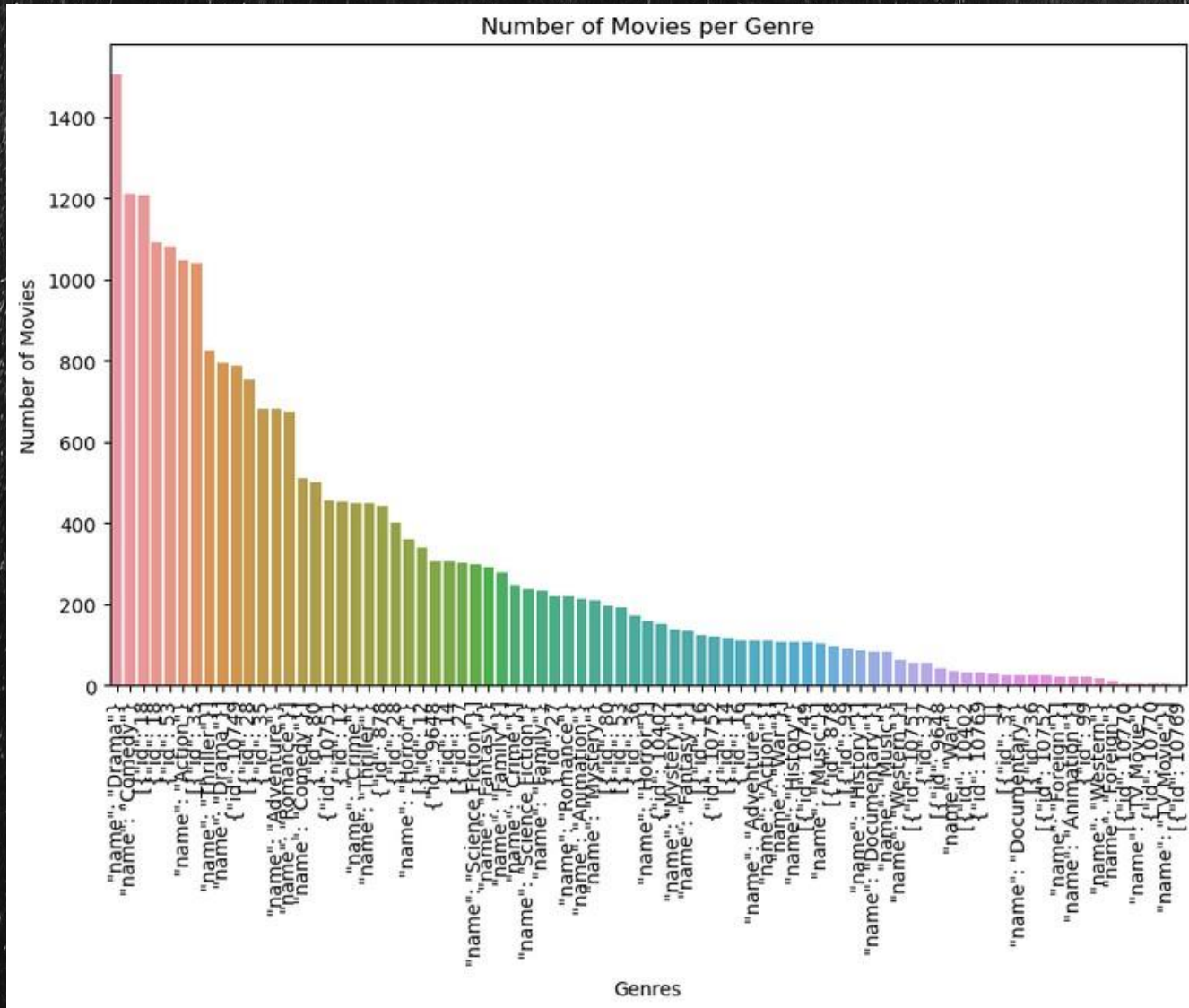


# EXPLORE DATA ANALYSIS





# MULTIVARIANT ANALYSIS

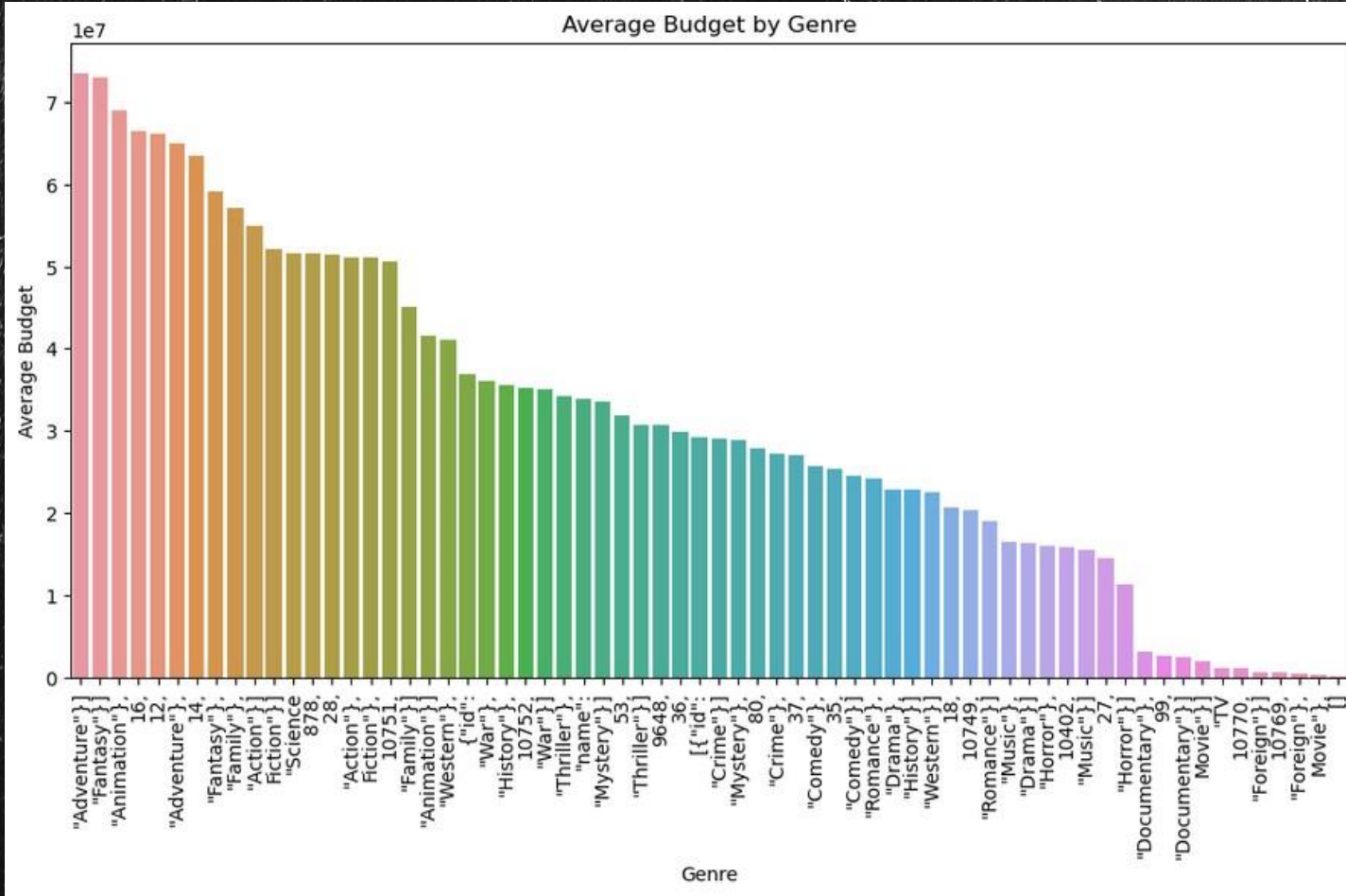


Total number of genres: 81

Average number of movies per genre: 301.01

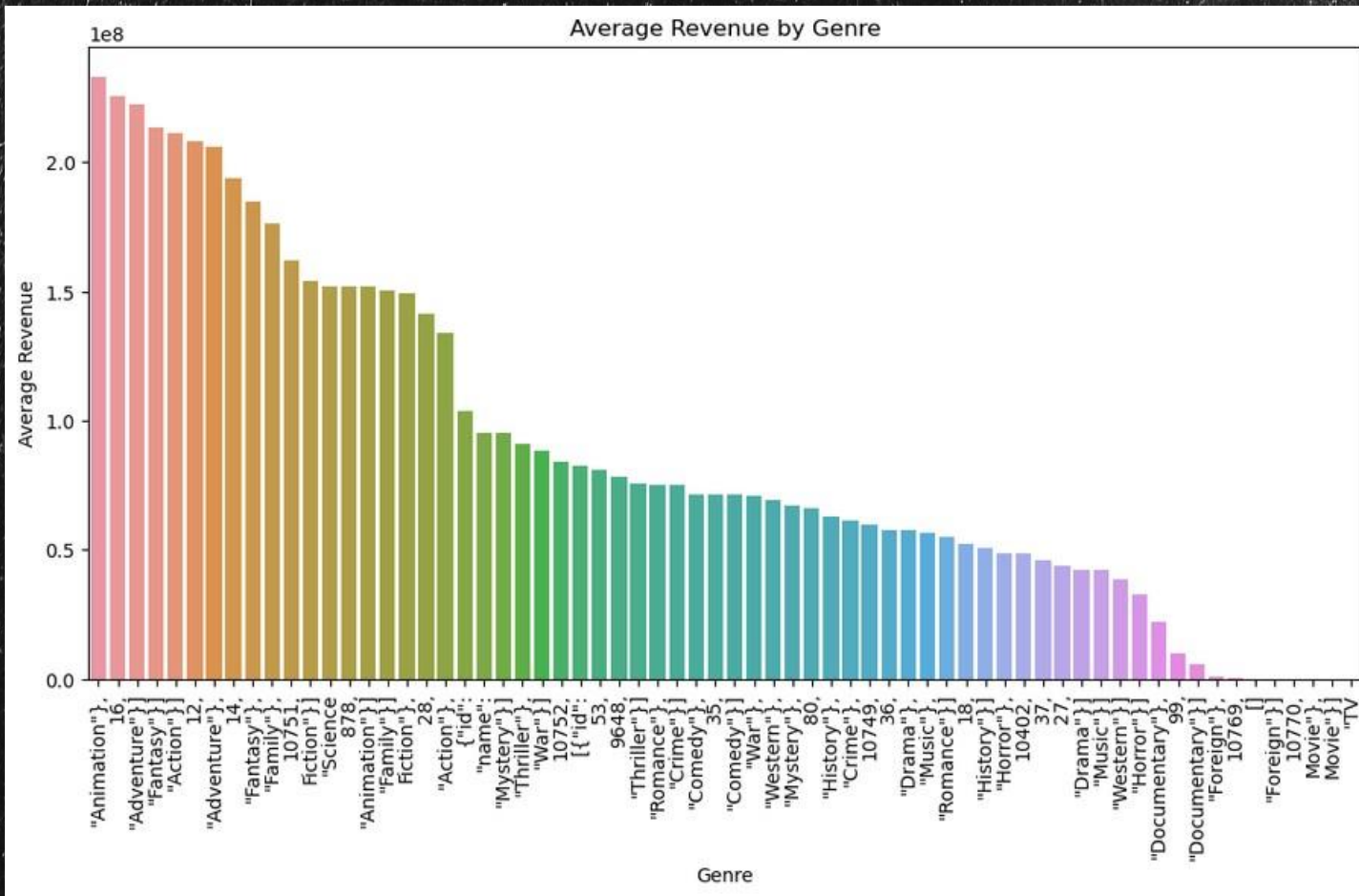


# MULTIVARIANT ANALYSIS





# MULTIVARIANT ANALYSIS





# MULTIVARIATE ANALYSIS

- Data trends show that genres with larger investments often yield higher revenues, supporting the utility of genre prediction in financial forecasting.
- Certain genres demonstrate a strong return on investment, with 'Family' and 'Animation' being prime examples, highlighting the value of genre prediction in resource allocation.
- Genres such as 'Western' and 'History' tend to have lower revenue relative to their budget, emphasizing the role of genre prediction in mitigating financial risk.
- The consistent high performance of genres like 'Animation' and 'Adventure' in revenue generation indicates market preference, which can be leveraged through genre prediction.
- Genres that manage to attract significant viewership with moderate budgets, such as 'Thriller' and 'Comedy', underscore the importance of genre prediction for effective marketing strategies.
- The necessity for larger budgets in certain genres due to production complexities underlines the need for genre prediction in production strategy formulation.
- The high revenues of certain genres reflect their current popularity, which could be predicted to capitalize on market trends and audience preferences.

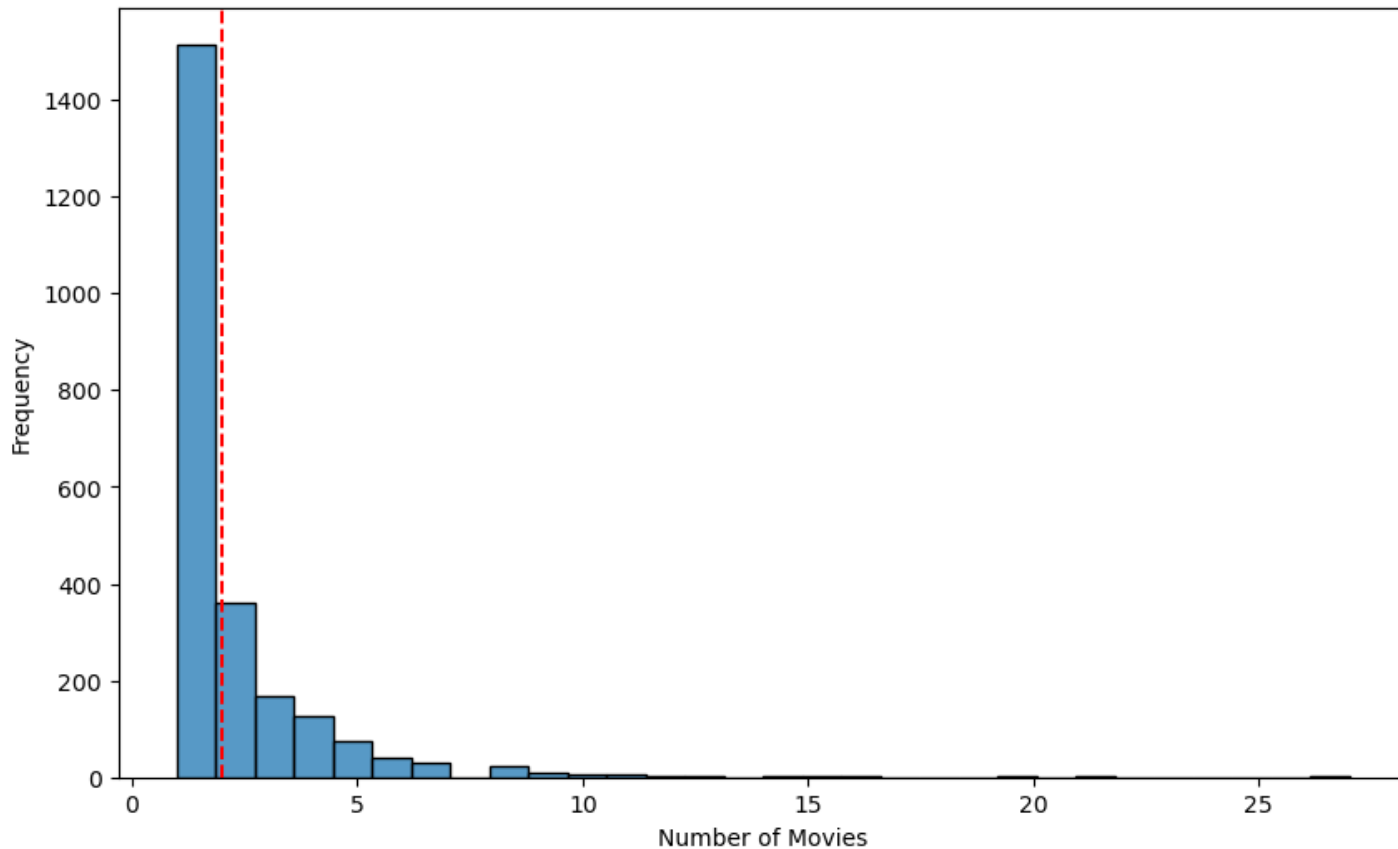




# MULTIVARIANT ANALYSIS



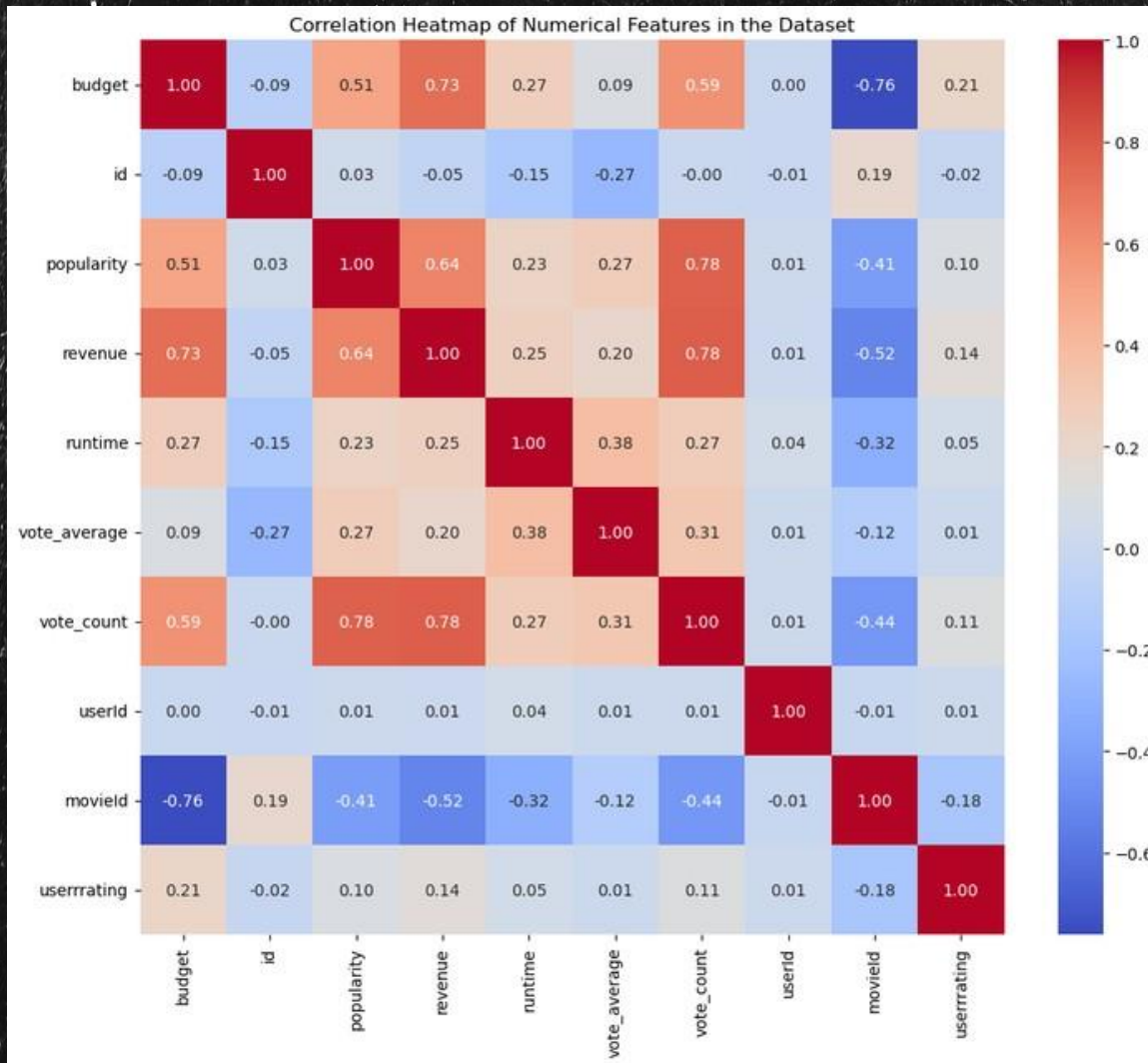
Distribution of Movies per Director



Total Unique Directors: 2385  
Average Number of Movies per Director:  
2.0138364779874216



# MULTIVARIANT ANALYSIS




- High budget genres tend to yield higher revenues, affirming the financial stakes in accurate genre classification.
- Popular genres draw more ratings, emphasizing genre's role in audience engagement metrics.



# MODEL PREDICTION



 Movie Analytics System

Enter Movie Title:

Avatar

Number of Recommendations:

15

Submit

Show Metrics

Show Genre Metrics Plot

Show Director Accuracy

## Genre Prediction:

**Features:** The features for genre prediction are the combined textual data from the 'genres' and 'keywords' columns of your dataset. These features are combined into a single column named 'combined\_features', which is then transformed using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization.

**Target Variable:** The target variable for genre prediction through RandomForest and Support Vector Machine is the genres of the movies. This is obtained from the 'genres' column and transformed into a binary matrix using MultiLabelBinarizer.



## MODEL PREDICTION:

Director Prediction:

**Features:** The features for director prediction are the same 'combined\_features' used in genre prediction. These are also transformed using TF-IDF vectorization and then standardized using StandardScaler.

**Target Variable:** The target variable for director prediction is not explicitly defined in your code, as K-Means clustering is an unsupervised learning algorithm

However, the clustering aims to group movies into clusters (in this case, two clusters) based on the similarity of their features. We are using these clusters to infer whether a director is among the top 1000 or bottom 1000 directors, based on the frequency of their movies in the dataset.

We got **accuracy of 0.27** for director prediction since it has high number of unique directors of **2385**.





# MOEL PREDICTION



**Inverse Document Frequency (IDF):** This measures the importance of the term across a set of documents. The IDF of a specific term is calculated as:

$$\text{IDF}(t, D) = \log \left( \frac{\text{Total number of documents } D}{\text{Number of documents with term } t \text{ in them}} \right)$$

TF-IDF is used to transform the textual data in the 'combined\_features' column (which combines 'genres' and 'keywords') into a numerical form.

This numerical representation captures the importance of words (terms) in each movie's description relative to the entire dataset, making it suitable for use in machine learning models.



## MODEL PREDICTION

### 2. Cosine Similarity:

Cosine similarity is a metric used to measure how similar two documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The formula for cosine similarity between two vectors  $A$  and  $B$  is:

$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

- After transforming the 'combined\_features' into TF-IDF vectors, cosine similarity is used to find similarities between movies based on these features.
- Specifically, it is used to calculate the similarity between the TF-IDF vectors of different movies.
- Movies with higher cosine similarity scores are considered more similar in terms of their genres and keywords.





# RESULTS



Classification Metrics				
Random Forest Metrics:				
	precision	recall	f1-score	support
Action	1.00	0.99	0.99	248
Adventure	0.98	0.97	0.98	167
Animation	1.00	0.82	0.90	45
Comedy	1.00	0.99	0.99	337
Crime	0.99	0.79	0.88	151
Documentary	0.00	0.00	0.00	15
Drama	1.00	0.98	0.99	445
Family	0.90	0.62	0.73	97
Fantasy	1.00	0.52	0.68	99
Fiction	0.99	1.00	1.00	114
Foreign	0.00	0.00	0.00	7
History	1.00	0.83	0.91	35
Horror	0.98	0.83	0.90	108
Movie	0.00	0.00	0.00	4
Music	1.00	0.10	0.18	30
Mystery	1.00	0.77	0.87	66
Romance	0.97	0.92	0.94	176
Science	0.99	1.00	1.00	114
TV	0.00	0.00	0.00	4
Thriller	1.00	1.00	1.00	252
War	0.86	0.23	0.36	26
Western	0.00	0.00	0.00	16
micro avg	0.99	0.89	0.94	2556
macro avg	0.76	0.61	0.65	2556
weighted avg	0.97	0.89	0.92	2556
samples avg	0.96	0.88	0.91	2556



# REESULTS

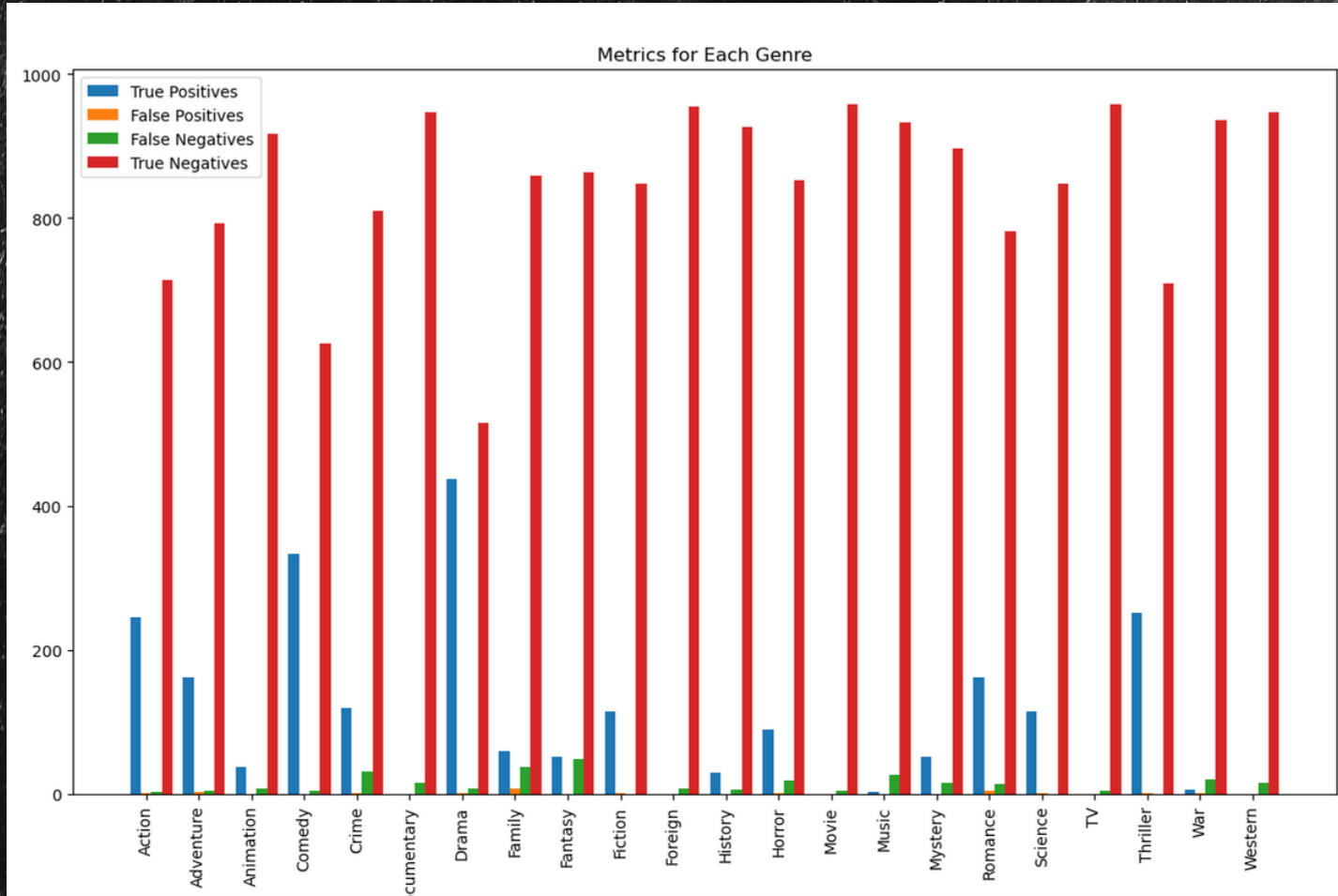


## SVM Metrics:

	precision	recall	f1-score	support
Action	0.99	1.00	1.00	248
Adventure	0.98	1.00	0.99	167
Animation	0.96	0.98	0.97	45
Comedy	1.00	1.00	1.00	337
Crime	0.96	0.99	0.98	151
Documentary	1.00	0.07	0.12	15
Drama	1.00	1.00	1.00	445
Family	0.86	0.92	0.89	97
Fantasy	1.00	0.98	0.99	99
Fiction	1.00	0.99	1.00	114
Foreign	0.00	0.00	0.00	7
History	0.97	0.97	0.97	35
Horror	0.98	0.99	0.99	108
Movie	0.00	0.00	0.00	4
Music	0.86	0.83	0.85	30
Mystery	1.00	0.98	0.99	66
Romance	0.97	0.99	0.98	176
Science	1.00	0.99	1.00	114
TV	0.00	0.00	0.00	4
Thriller	1.00	1.00	1.00	252
War	0.81	0.96	0.88	26
Western	0.00	0.00	0.00	16
micro avg	0.98	0.97	0.98	2556
macro avg	0.79	0.76	0.75	2556
weighted avg	0.97	0.97	0.97	2556
samples avg	0.96	0.96	0.96	2556



# RESULTS





# GITHUB PAGE SNAPSHOT

GITHUB Repository Link:

<https://github.com/DS606-Lohitha-Pragna-Sukruthi?tab=repositories>



DS606-Lohitha-Pragna-Sukruthi / DS6060TEAME

Tutor Application

Code Issues Pull requests Actions Projects 1 Wiki Security Insights Settings

DS6060TEAME Public

main 1 Branch 0 Tags

Go to file Add file Code

DS606-Lohitha-Pragna-Sukruthi	Update README.md	3611443 · 2 minutes ago	17 Commits
DS606 (1).ipynb	Add files via upload	3 months ago	
DS6060_TEAM_E.ipynb	Add files via upload	last month	
DS606WEEKCODE (1).ipynb	Add files via upload	2 months ago	
DS606_TeamE_LOHITHAPULIJALA_PRAGNAPUL...	Add files via upload	4 minutes ago	
README.md	Update README.md	2 minutes ago	
main1textcode	Create main1textcode	3 months ago	
tmdb_5000_movies_with_director_UPDATED.csv	Add files via upload	last month	
tmdb_5000_movies_with_director_UPDATED3.csv	Add files via upload	last month	

README

About

ENHANCED CONTENT BASED MOVIE RECOMMENDATION SYSTEM

Readme Activity 0 stars 1 watching 1 fork

Releases

No releases published  
[Create a new release](#)

Packages

No packages published  
[Publish your first package](#)

Languages



# REFERENCES



- F. Maxwell Harper and Joseph A. Konstan. 2015. *TheMovieLens Datasets: History and Context*  
• <https://ieeexplore.ieee.org/document/9767172>
- Jannach, D., et al. (2010). *Recommender Systems: An Introduction*.
- Lops, P., et al. (2011). *Content-based Recommender Systems: State of the Art and Trends*.
- Mikolov, T., et al. (2013). *Distributed Representations of Words and Phrases and their Compositionality*.
- TMDb Movie Metadata on Kaggle  
• <https://www.kaggle.com/code/ashutosh39/movie-recommendation-using-text-mining>  
• <https://www.kaggle.com/code/imsakshimittal/movie-recommender-system>  
• <https://www.kaggle.com/code/yassermessahli/movie-s-recommender-system-content-based>



## ***FUTURE WORK:***

- *Future Approaches:*

*To enhance model performance, we can explore various Machine Learning methodologies, including Sentiment Analysis and other Natural Language Processing techniques*





