Analyzing Genre Discrepancies and Crafting Movie Recommendation system

***Abstract*— This report delves into the intricate patterns among movie ratings, audience preferences, and clustering techniques to inform decision-making in the movie industry. Analyzing rating discrepancies among different genres, we uncover subtle patterns of critic and audience's preferences. Leveraging the clustering models, we explore the effectiveness of recommending models for movies based on the input of genre and content rating. While identifying potential biases in the clustering process, our findings suggest opportunities for personalized content delivery. Insights gained from this study paves the way for movie studios and streaming platforms to better understand audience preferences, fostering a more engaging and personalized cinematic experience.**

I. INTRODUCTION

The movie industry is one of the strongest branches of the media, reaching billions of viewers worldwide [1]. While some viewers like to get surprised by the performance of a movie by firsthand watching them, many like to get opinions before committing their time and energy into binging these movies [2]. Rotten Tomatoes is arguably the most popular website where people can find reviews and ratings for nearly any movie [3]. While most movie and series watchers pass a quick judgment without even seeing them based on the ratings from this platform [4], some viewers choose to form their own opinions for the movies and then assess whether to agree or disagree with the ratings. Rotten Tomatoes provides a platform for audiences to share their views and rate movies, while preserving the Tomatometer Rating exclusively for critics.

Exploring critic ratings and audience preferences is crucial in this diverse world of cinema [5]. Critics and audiences might consider different aspects when evaluating movies, and while there are some shared criteria, individual preferences play a significant role in shaping their opinions. For example, some genres resonate strongly with audiences, while others earn praise from critics without gaining similar audience opinions. This study aims to study the movies having large discrepancies between audience and critics ratings. Down the line, the motive is to cluster similar movies based on various features, providing a foundation for personalized movie recommendations.

II. ANALYTICAL QUESTIONS

The research tends to unravel the answers to 4 analytical questions, them being:

a) Are there movies present with large discrepancies between critic and audience ratings?

This could give rise to some interesting perceptions of critic and audience preferences of movies.

b) What are the particular genres that critics and audiences don't share similar views on?

Exploring genres where critics and audiences diverge on their opinions can be an interesting study as it can shed light on patterns on why certain genres appeal differently to them. It can also have practical implications for filmmakers seeking to navigate their target audience.

c) Is it effective to cluster similar movies based on various features?

This question leads to the analysis of movies assigned to the same cluster to ascertain the extent of their similarities.

d) Do clusters formed based on various movie features result in accurate movie recommendations?

Exploring the accuracy of movie recommendations within clusters formed based on diverse movie features introduces an understanding of clustering method's abitlity to provide accurate movie recommendations.

Each question we ask sets the stage for the next, thus helping us discover important aspects of the dataset. By addressing these questions one by one, we unravel the complexities of how people perceive and categorize movies, offering valuable insights into the world of cinema.

III. DATA (MATERIALS)

A. Key Characteristics And Why They Are Suitable For Answering The Analytical Questions

1) Genres and Ratings: The dataset includes comprehensive information on movie genres, allowing for a detailed exploration of how different genres are perceived by both critics and audiences. Ratings, such as the tomatometer and audience ratings, provide a quantitative measure of movie reception.

2) Runtime: The 'runtime' feature captures the duration of each movie, contributing to the understanding of how movie length may influence audience and critic opinions.

3) Critic and Audience Status: Features like 'critic_status' and 'audience_status' provide additional insights into the reception of movies. These status indicators can offer a categorical perspective on whether a movie is well-received by critics, audiences, or both.

4) Rating Discrepancy Metric: The calculated 'rating_discrepancy' metric, derived from tomatometer and audience ratings, serves as an important role in assessing the variance between critic and audience perspectives, while gaining the insights about the extent of discrepancies between them.

5) Temporal Dynamics: The temporal aspect of the dataset, represented by features like 'original_release_date,' enables the investigation of trends and patterns in movie releases over time. This temporal dimension is instrumental in understanding evolving preferences and industry dynamics.

These features, excluding temporal variables, are integral to the clustering process, forming the basis for the movie recommendation system. The clustering mechanism chooses genre, ratings, and other characteristics to identify similar

movies, thereby enhancing the precision of movie recommendations based on user preferences.

B. Assumptions

Assuming that the audience and critic ratings are not affected by unobserved factors beyond the dataset, such as marketing strategies or external events, which in reality could influence the ratings. Also, assuming that the provided tomatometer and audience ratings accurately reflect the opinions of critics and audiences, respectively.

IV. ANALYSIS

A) Data Preparation

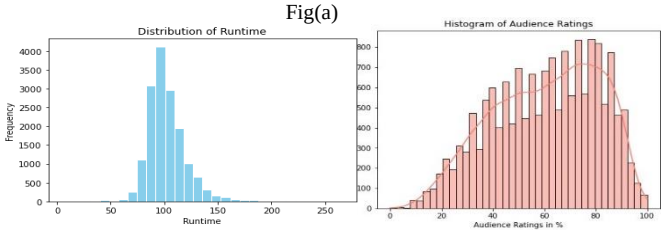
i. Dealing With the Duplicates:

After merging datasets for the purpose of comprehensive movie analysis, initial duplicates were removed using '.drop_duplicates'. However, a significant number of duplicate entries remained, prompting us to take a closer look at them based on movie titles. Nearly two-thirds of the dataset contained duplicate entries with identical titles and other features but distinct content reviews. To retain sequel information, duplicates were removed based on movie title and runtime, as it is highly unusual for sequels to have the same runtime (in minutes) even though they could have the same title. This strategy ensured the removal of non-sequential duplicates.

ii. Handling Missing Values:

Careful evaluation of missing values was done for the important features. For genre and tomatometer columns with low counts of missing values, corresponding rows were removed. However, for critical features like runtime and audience ratings, descriptive statistics and distribution analysis was performed which guided imputation decisions. Mean imputation for runtime, justified by a nearly normal distribution, preserved overall integrity. In audience ratings, a slightly skewed distribution favored median imputation, balancing central tendency preservation with outlier considerations. These approaches aimed to maintain data authenticity.

Fig(a)

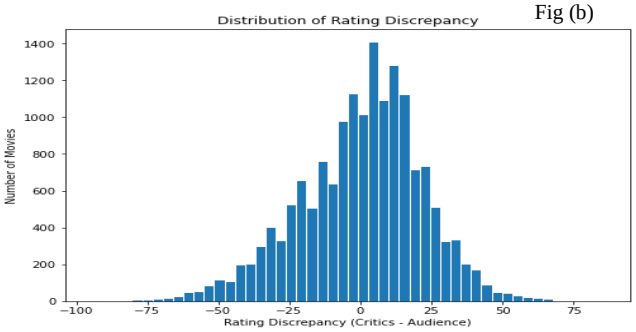


B) Data Derivation

i. Creation of Rating Discrepancy column:

The creation of the "rating_discrepancy" and "absolute_rating_discrepancy" columns is crucial for addressing the question about movies with large discrepancies between critic and audience ratings. These columns give a quantitative measure of the difference between these ratings, providing a basis for analyzing such discrepancies. The figure (b) displays the distribution of the 'rating_discrepancy' column, thus answering our first analytical question by providing evidence that movies with significant differences in opinions between critics and audiences are present in the dataset.

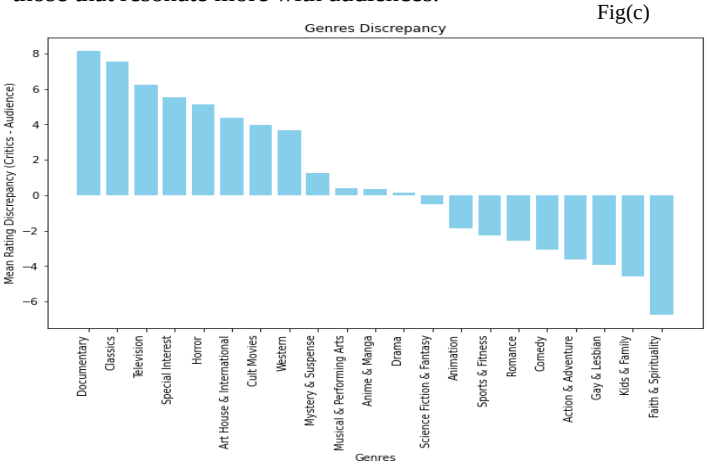
Fig (b)



ii. Genre-wise Mean Rating Analysis:

The exploration of genre column involves the process of its segregation, where each movie, often assigned with multiple genres, is effectively categorized. Employing the split function enabled a more detailed examination of individual genres. Following this, mean tomatometer and audience ratings were calculated for each genre resulting in the calculation of mean rating_discrepancies for individual genres Negative discrepancy values indicate a general preference for genres by audiences, while higher positive values suggest that critics tend to favor these genres more. The visual representation of the differences of these mean ratings further enhances the understanding of genres where average critics and audiences diverge in their views. This step provides the insights to answer one of the analytical questions as it reveals the genres that get more favour from critics, as opposed to those that resonate more with audiences.

Fig(c)



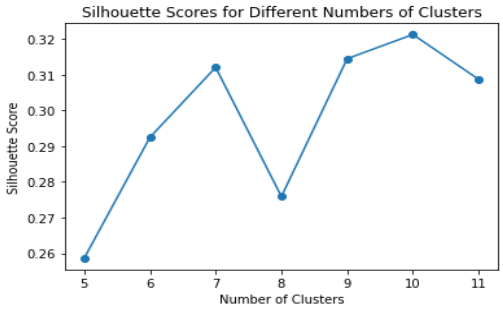
iii. Encoding and scaling of the important features:

In preparing the dataset for modelling, the emphasis was placed on the conversion of categorical variables into numerical ones. The genre column, with its multiple genre assignments, required careful encoding. Using one-hot encoding, each movie's genres were appropriately represented with 1s and 0s. The same encoding approach was applied to the content rating column. For tomatometer status, audience status, and runtime bucketized (additional column to provide an overview of movie runtime), ordinal encoding was employed. Additionally, normalization of significant columns for clustering was ensured by scaling them using MinMax Scaling technique, placing all features within the standardized range of 0 to 1.

C) Construction of Models

i. Performing K-Means Clustering:

Following the encoding and scaling of essential features for analysis, a combined dataframe with all the core features was created. To determine the optimal number of clusters, silhouette score method was used where the highest silhouette score decides the optimal number of clusters. Fig(d) shows the results of the silhouette score. Subsequently, leveraging the optimal cluster count, k-means clustering was executed based on diverse features, including genres, critics' ratings and status, audience rating and status, runtime, and rating discrepancies. Despite setting a specific random seed, the clusters formed exhibited slight variations with each run of the code. This could be due to the cluster behaviour and their sensitivity to initial conditions.



Optimal Number of Clusters: 10

Fig (d)

ii. Creating basic model for movie recommendations based on the clusters:

After forming clusters through k-means clustering, a basic model for movie recommendations was developed based on these clusters. While the dataset lacked extensive information on audience and critic preferences, the goal was to assess the practicality of the clustering approach. The model is designed to recommend movies based on simple inputs like genre and content rating. The system identifies the cluster with the highest frequency of the input feature, excluding the exact input, and recommends movies from that cluster assuming that it contains similar movies. This recommendation process prioritizes critic ratings, presenting results in decreasing order of the tomatometer rating. This model creation aims to showcase the applicability of the clusters formed on the basis of a number of features in offering personalized movie suggestions, emphasizing the technique's effectiveness in a real-world scenario.

D) Validation of Results

i. Analysing the Characteristics in individual Cluster:

The visuals of some of the clusters are shown in fig (e) produced to get the mean characteristics in each cluster to check the effectiveness of the clustering process.

In cluster 2, movies predominantly belong to the 'Gay & Lesbian' genre, with many rated as NR. Although the mean audience status is high, both critic and audience ratings are minimal in this cluster. Cluster 3, on the other hand, encompasses movies of diverse genres like 'Documentary', 'Faith & Spirituality', and 'Musical & Performing Arts', receiving high ratings from both critics and audiences. The majority of films in this cluster are also rated as NR. From this example, it is clear that relying on multiple features in the clustering process adds variety and complexity to the clusters,

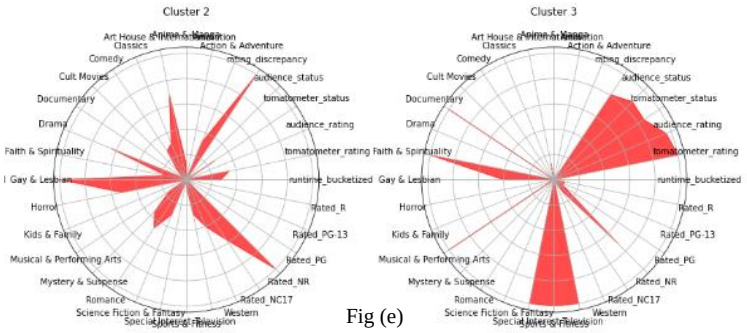


Fig (e)

thus answering the second analytical question. It becomes tricky to figure out what kinds of movies are in a cluster just by looking at radar plots. The figure below shows how the clusters are spread out based on various criteria, emphasizing the intricate nature of the groups. It underscores the importance of a detailed analysis from various angles to truly grasp what makes up each cluster.

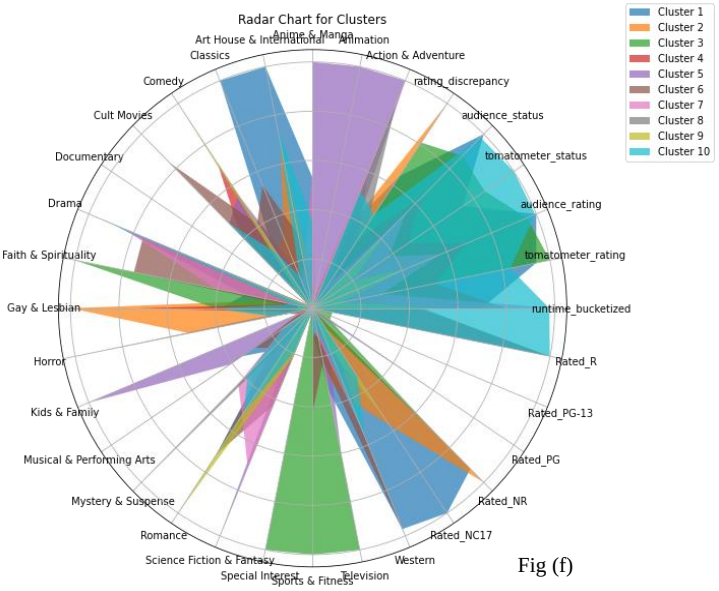


Fig (f)

V. FINDINGS, REFLECTIONS AND FUTURE WORK

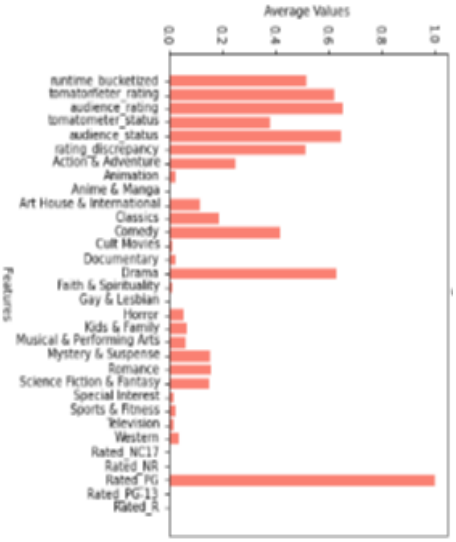
Addressing the analytical question about the particular genres that have high discrepancies between the audiences and the critics, the rating discrepancies were seen for each individual genres (fig(c)). Genres like 'Documentary' and 'Classics' saw critics more inclined, while 'Faith & Spirituality' and 'Kids & Family' resonated more with audiences. Meanwhile, genres such as 'Drama,' 'Musical & Performing Arts,' and 'Science Fiction & Fantasy' found a harmonious balance, getting assigned with comparable ratings from both critics and audiences. This finding offers a straightforward glimpse into the diverse ways critics and audiences perceive various genres.

To address the final analytical question on the effectiveness of clusters in providing accurate movie recommendations, an in-depth analysis of the model's output was conducted. Specifically, the recommendations generated by the movie recommendation system were tested using inputs like genre or content rating. Illustrating the model's output, a snippet of the table resulting from inputting the genre 'Drama' is provided.

movie_title	genres	cluster
HULK VS.	Action & Adventure, Animation, Kids &...	6
My Life as a Zucchini	Animation, Art House & International,...	6
Babe	Action & Adventure, Drama, Kids & Family, Sci...	6
Batman Beyond: Return of the Joker	Action & Adventure, Animation, Kids &...	6
20,000 Leagues Under The Sea	Action & Adventure, Drama, Kids & Family	6
Gianni e le donne (The Salt of Life)	Art House & International, Comedy, Drama	6
My Side of the Mountain	Action & Adventure, Art House & Inter...	6
Black Beauty	Action & Adventure, Drama, Kids & Family	6
Kim	Action & Adventure, Classics, Drama, Kids...	6
Batman: Assault on Arkham	Action & Adventure, Animation, Kids &...	6

Fig (g)

The model, in its process, identifies the cluster with the highest frequency of the input genre, which is cluster 6 in the case of 'Drama' genre, excludes the exact input, and suggests movies similar to those within that cluster. It can be seen from the bar chart that cluster 6 indeed has a substantial number of movies assigned to the 'Drama' genre, corroborating the model's choice. Emphasizing the assumption that viewers



Cluster 6
Fig (h)

seek movies akin to the 'Drama' genre rather than solely that genre, the recommendations originate from a cluster with analogous movies but encompassing different genres. After analysing the output closely, all the recommended movies, although of different genres than the input, share some thematic similarities with the 'Drama' genre. However, it's worth noting that some movies in the recommendations like 'My Life as a Zucchini' are exceptions. These movies have low significant correlation with the movies of the 'Drama' genre, suggesting that variations in the clusters may lead to occasional deviations in movie recommendations.

Reflecting on the clusters and the movie distributions within the clusters, the pie chart shown in fig (i) illustrates the unequal spread of movies across different clusters in the dataset. This imbalance poses challenges for the analysis, potentially leading to biased insights and non-accurate recommendations. Clusters with a disproportionately large number of movies could be dominating our model, reducing the variety of suggested films. Consequently, the system could exhibit a preference for the characteristics prevalent in over-represented clusters, skewing the accuracy of movie recommendations. To enhance the robustness and fairness of the model, addressing this imbalance would be essential to ensure a more equal representation of movies across all clusters and making a more accurate and inclusive recommendation system.

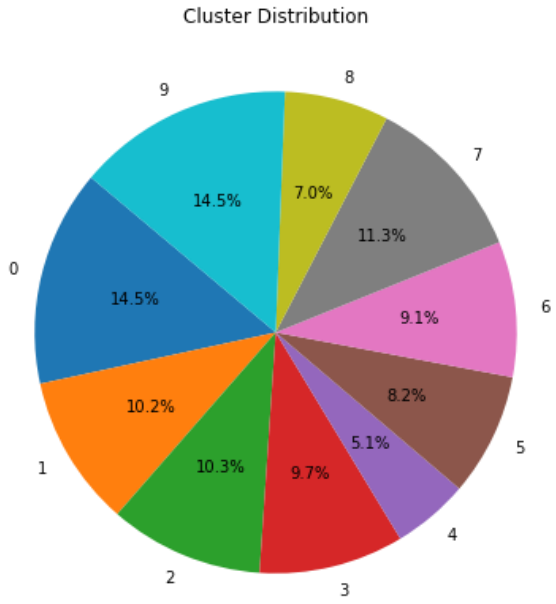


Fig (i)

Moreover, the datasets used for this reasearch lacked actual user preferences, thus limiting the model's efficiency in crafting personalized movie recommendations. Future improvements could be achieved by collecting honest user feedback on the recommendation system. This valuable input could help to enhance the model, making it more aligned with users' tastes for accurate and valuable movie suggestions.

In conclusion, the insights derived from our analysis can significantly impact decision-making within the entertainment industry, particularly for movie studios and streaming platforms. By understanding the preferences and discrepancies between critics and audiences across different genres, the people working in the entertainment industry can tailor their content creation and marketing strategies more effectively.

REFERENCES

[1] "Courtesy Walt Disney Studios/Pixar Animation Studios." Available: https://www.motionpictures.org/wp-content/uploads/2018/04/MPAA-THEME-Report-2017_Final.pdf

[2] M. Nishijima, M. Rodrigues, and T. L. D. Souza, "Is Rotten Tomatoes killing the movie industry? A regression discontinuity approach," Working Papers, Department of Economics 2021_12, University of São Paulo (FEA-USP), 2021.

[3] Rotten Tomatoes, "Rotten Tomatoes: About," *Rotten Tomatoes*, 2019. https://www.rottentomatoes.com/about

[4] X. Song, "Analyzing movie scores on IMDB and Rotten Tomatoes." https://rstudio-pubs-static.s3.amazonaws.com/336722_2193716117584b63a2a6ebb837217d85.html

[5] Alaji, A. (2023). "Investigate the Effect of Rotten Tomatoes and IMDb's Rating and Critic Reviews on Movies Publicity." International Journal of Electronics Communication and Computer Science, 14(2), 323. doi: 10.9756/INT-JECSE/V14I2.323.

Title	Word Count
Abstract	102
Introduction	234
Analytical Questions	217
Data (Materials)	275
Analysis	995
Findings, Reflections & Future Work	547