ANIME ODYSSEY: EXPLORING GENRES, CLUSTERING & TRENDS

AKSHAT PORWAL (AKSHATP), PRATHAMESH JOSHI (PRATHUJ) & SETTASIT CHITPHENTOM (SETTASIT)

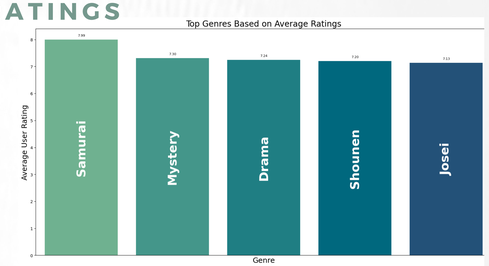
INTRODUCTION

The anime industry has experienced rapid growth in recent years, with an ever-expanding library of anime titles and a growing user base on platforms like MyAnimeList. However, with this growth comes the challenge of providing users with personalized recommendations, enhancing user satisfaction, and understanding the genre associations.

RESEARCH QUESTIONS

1. Which anime genres are most highly rated by users based on average scores?
2. How does the hierarchical clustering of anime genres and user grouping into genre clusters provide insights into diverse user preferences?
3. How does the predictive performance of the recommendation system in rating predictions, and what are the effective strategies to recommend unwatched anime based on user reviews?
4. What is the temporal trend in user preferences concerning the release timing of anime, and to what extent do genres influence user ratings of anime?
5. How do word clouds, derived from anime synopses within specific genres, illuminate the unique themes characterizing each genre?

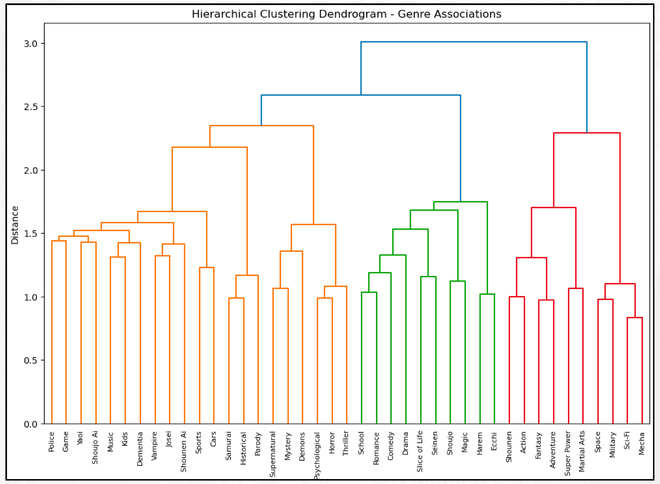
TOP GENRES BASED ON AVERAGE RATINGS

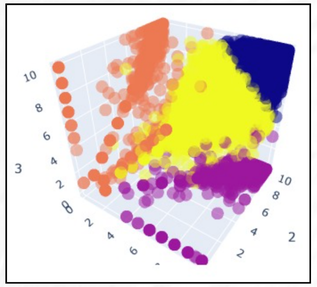


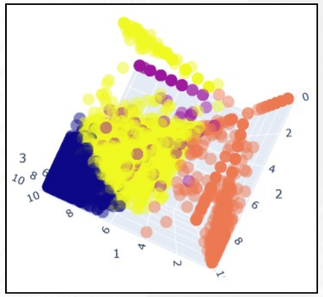
EXPLORING THE JACCARD SIMILARITIES BASED ON GENRES

Anime: Fullmetal Alchemist: Brotherhood	
One Piece: Heart of Gold	0.75
One Piece Movie 14: Stampede	0.66
Akame ga Kill!	0.62

ANALYSING USER CLUSTERS



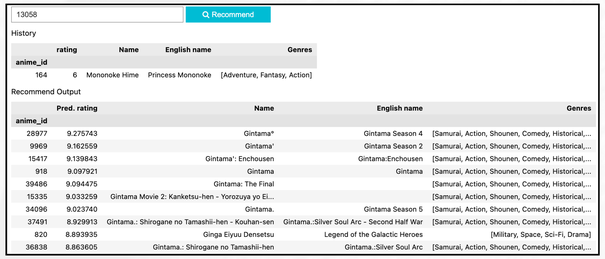




USER RECOMMENDATION

Prediction performance is evaluated via 3-fold cross-validation on reviews from ~317,000 users (13.4 million reviews). After training on the complete dataset, the model can predict ratings for unwatched anime, facilitating personalized recommendations based on anticipated scores.

Recommendation Model	Mean Absolute Error (MAE)
Singular Value Decomposition (SVD)	0.88
K-Nearest Neighbors	0.95



13058 | Recommend

History

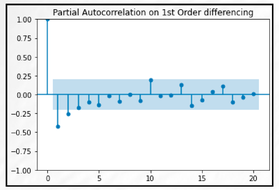
anime_id	rating	Name	English name	Genres
164	6	Mononoke Hime	Princess Mononoke	[Adventure, Fantasy, Action]

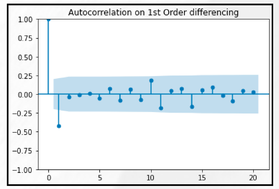
Recommend Output

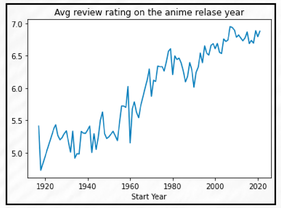
anime_id	Pred. rating	Name	English name	Genres
28977	9.275743	Gintama°	Gintama Season 4	[Samurai, Action, Shounen, Comedy, Historical,...
9969	9.162559	Gintama'	Gintama Season 2	[Samurai, Action, Shounen, Comedy, Historical,...
15417	9.139843	Gintama°: Enchousen	Gintama:Enchousen	[Samurai, Action, Shounen, Comedy, Historical,...
918	9.097921	Gintama	Gintama	[Samurai, Action, Shounen, Comedy, Historical,...
39486	9.094475	Gintama: The Final		[Samurai, Action, Shounen, Comedy, Historical,...
15335	9.033259	Gintama Movie 2: Kanketsu-hen - Yorozuya yo Ei...		[Samurai, Action, Shounen, Comedy, Historical,...
34096	9.023740	Gintama.	Gintama Season 5	[Samurai, Action, Shounen, Comedy, Historical,...
37491	8.929913	Gintama.: Shirogane no Tamashii-hen - Kouhan-sen	Gintama.:Silver Soul Arc - Second Half War	[Samurai, Action, Shounen, Comedy, Historical,...
820	8.893935	Ginga Eiyuu Densetsu	Legend of the Galactic Heroes	[Military, Space, Sci-Fi, Drama]
36838	8.863605	Gintama.: Shirogane no Tamashii-hen	Gintama.:Silver Soul Arc	[Samurai, Action, Shounen, Comedy, Historical,...

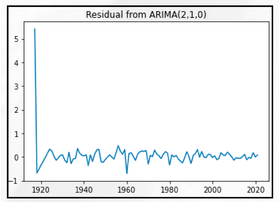
USER TEMPORAL PREFERENCE

Differencing of the average ratings on release year results in a weakly stationary time series. Investigation of partial autocorrelation in the differenced data indicates that an ARIMA(2, 1, 0) model is well-suited for capturing the underlying patterns in the pre-differenced dataset, showcasing its effectiveness in modeling the temporal dynamics of the data.

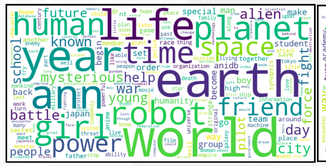




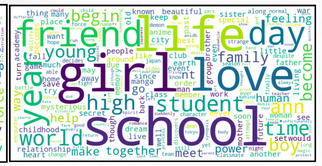




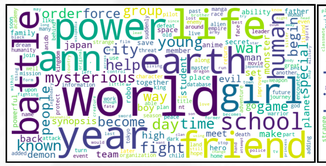
GENRE- SPECIFIC WORD CLOUDS



Genre: Sci-fi



Genre: Romance



Genre: Action

Genre: Comedy

FINDINGS

- User can be classified into 4 groups based on how users rate the anime.
- There are multiple approaches in recommending anime to users; Using similarity of genres, or recommendation system which based on peer reviews.
- The average rating based on release year of anime fits ARIMA(2,1,0) can be used to predict average rating of future anime.
- Word Clouds suggest genres associated with certain keywords. Yet, there are keywords which universally associated with anime itself.

Anime Odyssey: Exploring Genres, Clustering & Trends

Akshat Porwal (akshatp), Prathamesh Joshi (prathuj), and Settasit Chitphentom (settasit)

Abstract

The anime industry has undergone substantial growth in recent times, marked by an extensive array of anime titles and an escalating user base on platforms like MyAnimeList. This growth presents a challenge in delivering personalized recommendations to users, augmenting user satisfaction, and unraveling complex genre associations. In this study, a multifaceted methodology is employed to address these challenges. It involves analyzing Jaccard similarities among anime based on genres, conducting hierarchical clustering to understand genre structures, employing the k-means method for user segmentation, and placing emphasis on the development of a User Recommendation System and exploration of User Temporal Preferences. Additionally, a word cloud analysis is conducted on anime synopses to extract key thematic elements.

1 Introduction

The anime industry's rapid expansion has led to an immense collection of anime titles and an ever-increasing audience base on platforms like MyAnimeList. While this growth is promising, it introduces significant challenges, including the need for personalized recommendations, elevated user satisfaction, and a deeper comprehension of genre connections. This study endeavors to tackle these challenges by employing a comprehensive methodology. To unravel the intricate web of anime genres, Jaccard's similarities are scrutinized. Further insights into genre structures are derived through hierarchical clustering analysis. User segmentation is then achieved using the k-means method, focusing particularly on the development of an innovative User Recommendation System. Moreover, the study delves into User Temporal Preferences to understand the evolving patterns of user behavior over time. Additionally, an exploration of anime synopses through a word cloud analysis sheds light on prevalent thematic elements. This multifaceted approach aims to enrich the understanding of user preferences, genre dynamics, and thematic content prevalent in the diverse realm of anime.

2 Objectives

The primary objectives of this project are to:

1. **Identify Top-Rated Genres:** Explore and determine the anime genres that receive the highest average ratings from users, providing insights into popular and well-received genres within the anime community.

2. **Analyze User Preferences:** Investigate how hierarchical clustering of anime genres, coupled with user grouping into genre clusters, contributes to a nuanced understanding of diverse user preferences and viewing patterns.

3. **Evaluate Recommendation System:** Assess the predictive performance of the recommendation system in generating accurate rating predictions. Explore effective strategies for recommending unwatched anime based on user reviews.

4. **Examine Temporal Trends:** Investigate temporal trends in user preferences, specifically focusing on the release timing of anime. Analyze the extent to which genres influence user ratings over time.

5. **Illuminate Genre-Specific Themes:** Utilize word clouds generated from anime synopses to highlight and interpret unique themes characterizing each genre, enhancing our understanding of the narrative elements associated with distinct anime categories.

These objectives collectively aim to provide comprehensive insights into user preferences, system performance, and thematic elements within the anime landscape, fostering a deeper understanding of the dynamics within the anime community.

3 The Dataset

The project utilizes the MyAnimeList Database 2020, sourced from Kaggle. This extensive dataset encompasses information on 17,562 anime titles and preferences from 325,772 different users. The dataset comprises:

- **animelist.csv:** Contains user-specific anime lists with scores, watching statuses, and episode details.
- **rating_complete.csv:** A subset of animelist.csv focused on completely watched anime and user ratings.
- **anime.csv:** Provides general information about each anime, including genres, studios, and ratings.
- **HTML files:** Available for scraping detailed information like reviews, recommendations, and stats.

Data collection spanned from February 26th to March 20th, yielding a diverse and comprehensive dataset for conducting analyses on user preferences, recommendation strategies, temporal trends, and thematic exploration across anime genres.

4 Methodology

Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) unfolded in two phases, beginning with a meticulous examination of anime title genres and metadata. To address the challenge posed by multiple genres per

anime, the dataframe underwent an explosion, enabling an insightful analysis of genre popularity. The subsequent calculation of average genre popularity highlighted the top 5 genres, offering a comprehensive snapshot of genre dynamics.

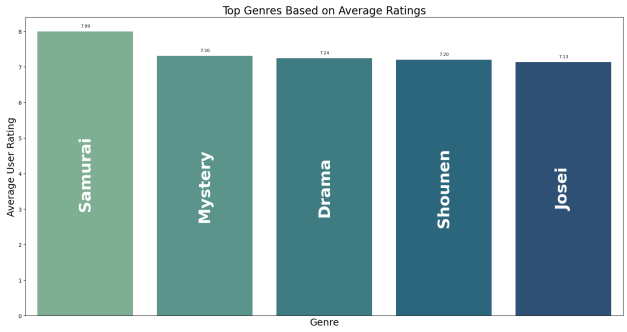


Figure 1: Top 5 Genres on Average Rating

Transitioning to the realm of user reviews, the EDA unearthed intriguing patterns. Despite the predominance of titles with fewer than 2500 reviews, the non-uniform distribution revealed that select anime, particularly renowned ones, attracted significantly more reviews. An exploration of average ratings by release year per genre unveiled positive correlations, indicating enduring viewer appreciation over time.

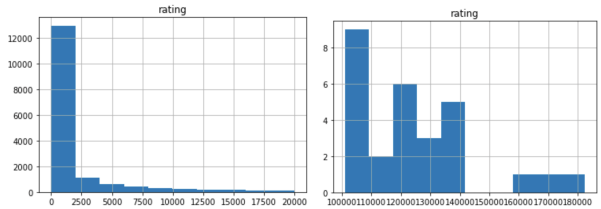


Figure 2: Exploring Average Ratings

However, exceptions in genres labeled as "unknown" and certain minor genres introduced nuances to this overarching trend. This multifaceted EDA lays the foundation for comprehensive insights into genre dynamics and user engagement in the anime landscape.

4.1 Exploring Jaccard Similarities Based on Genres

In our exploration of anime data, we delved into uncovering similarities between different anime by utilizing the Jaccard similarity measure based on their genre classifications. The initial step involved curating a comprehensive dataset encompassing anime details and diverse genre combinations. Preprocessing efforts focused on structuring genre information into sets conducive to subsequent set operations.

Once genres were appropriately formatted as sets for each anime entry, Jaccard similarities were systematically computed. Leveraging the Jaccard index formula, which considers intersection and union of genre sets, we calculated similarity scores for anime pairs. This iterative process involved comparing genre compositions and quantifying similarity based on shared and distinct genre classifications.

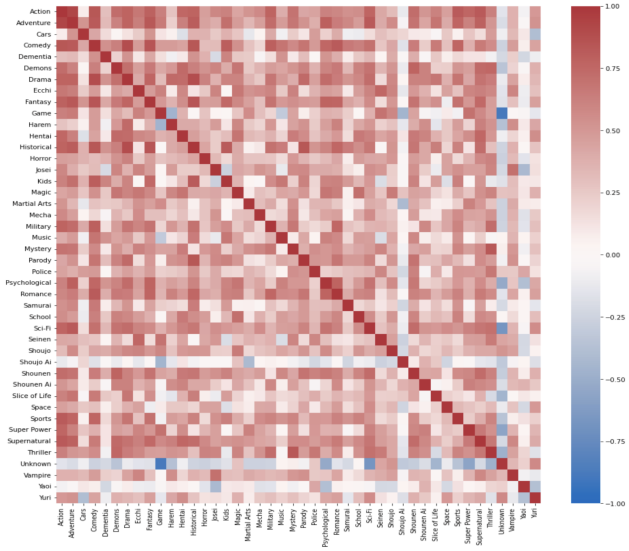


Figure 3: Genre-wise Average Rating Correlation Heatmap Over Release Years

The resulting output provides a matrix or dictionary of Jaccard similarity scores, offering insights into genre-based resemblances among anime. This exploration lays the foundation for understanding anime relationships based on genre affinities, potentially informing recommendations and enhancing user experiences.

4.2 Analyzing User Clusters

Creating a one-hot encoded matrix for movie genres, especially with multiple genre associations, involves systematically transforming categorical data into binary format. Parsing the 'Genres' column and extracting unique genre labels initiates the process. Expanding the DataFrame with new columns representing each genre, the binary indicators are iteratively updated based on movie genre tags.

Hierarchical clustering using Ward linkage is an agglomerative technique, merging the closest clusters to form a hierarchical structure. Ward linkage, known for minimizing variance during merging, is suited for applications prioritizing variance minimization.

Users were segmented by K-means clustering based on their movie genre preferences. Utilizing a user-genre matrix, the K-means algorithm grouped users with similar genre inclinations. This approach identified distinct user segments, enabling targeted content recommendations or marketing strategies aligned with each genre-based cluster's preferences.

4.3 User Recommendation System

The SVD model showcases commendable prediction performance. The evaluation metrics from the 3-fold cross-validation is used to measure reliability and effectiveness of the model in predicting user preferences for anime. Additionally, the comparison with K-Nearest Neighbors provides valuable insights into the relative strengths of different collaborative filtering techniques

To illustrate the practical utility of the SVD model, a user-friendly demo has been developed. Users can input their unique identifier, and the demo will present their viewing history alongside personalized recommendations.

Figure 4: Recommendation system demo with user ratings and recommended items

4.4 User Temporal Preference

The relationship between genres and rating when grouping the titles by released year. This release year is determined by the date the title first aired. The correlation between genres is high as one title has more than one genre.

By focusing on the average user ratings for each anime title which is grouped by its released year. The average rating for newer titles seems to have higher ratings than older counterparts. Furthermore, the first-order differencing of the data results in a weakly stationary time series with which the Autoregressive model can be applied.

4.5 Genre-Specific Word Clouds

To generate a word cloud from anime synopses, the process involves several sequential stages. We began by collecting the synopses data. Upon acquisition, the synopses underwent essential preprocessing steps, including text standardization (lowercasing), punctuation removal, tokenization into individual words, and exclusion of common stopwords using libraries like NLTK (Natural Language

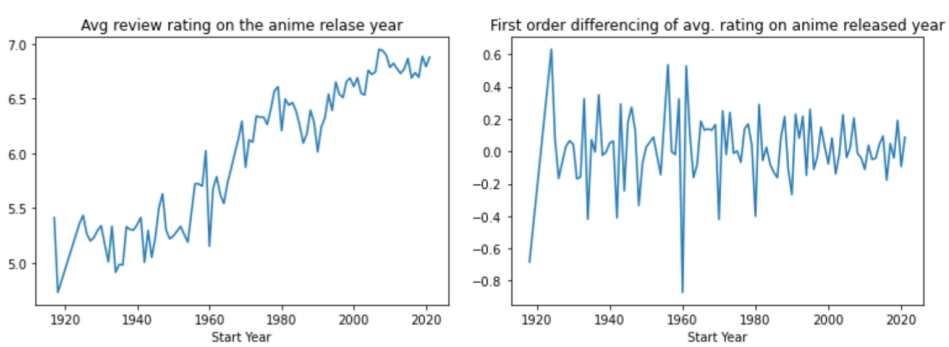


Figure 5: (left) Average rating of anime by releasing year (right) first order differencing of average rating of anime by releasing year

Toolkit). Further normalization techniques like lemmatization or stemming might be applied. After the preparation phase, the cleaned text is ready for word frequency computation, tallying the occurrence of each word within the processed synopses. Subsequently, using Python's 'wordcloud' library, we generated a visual representation of the word cloud. Adjusting parameters such as image size, background color, and font style resulted in a word cloud offering an illustrative depiction of prevalent terms and themes extracted from anime synopses—a concise yet informative snapshot of the underlying content.

5 Results and Conclusions

In delving into the expansive realm of anime, our study has yielded insightful results that cast light on fundamental aspects such as user preferences, genre associations, and temporal trends. Our exploration of user preferences brought to the forefront top-rated genres, including Samurai, Mystery, Drama, Shounen, and Josei. Utilizing Jaccard similarity measures, we unraveled nuanced relationships among anime genres, offering a nuanced perspective on the diverse landscape of genre-based resemblances. Taking a deeper dive, we applied Jaccard similarities to discern connections between anime genres. For instance, in the case of Fullmetal Alchemist: Brotherhood, we identified its top three anime counterparts based on Jaccard similarity, shedding further light on the intricate web of genre relationships within the anime domain.

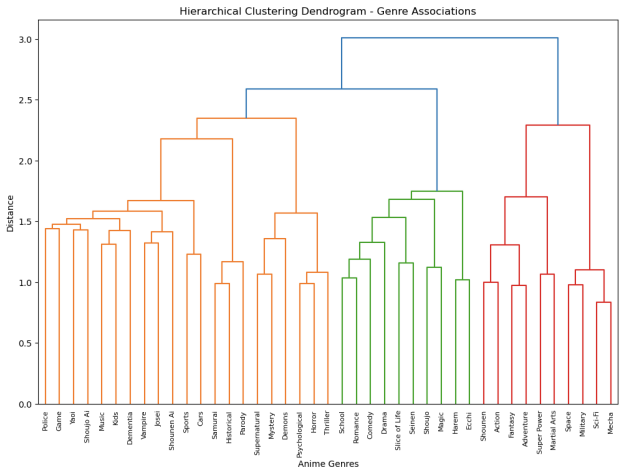


Figure 6: Hierarchial Clustering of Genres

Our user segmentation analysis categorizes viewers into distinct segments based on their interaction with anime genres. Each segment is represented by a specific color, aiding in visual interpretation. The first segment, identified by the color orange, is associated with Genre Clusters 2 and 3 in the plane, indicating a preference for movies within these clusters while rarely engaging with those in Cluster 1. The magenta segment aligns with Genre Clusters 1 and 2, showcasing a tendency to view movies in Cluster 3 less frequently. The yellow segment occupies the middle part of the three axes, symbolizing users who provide average ratings across all three genre clusters. Finally, the blue segment, positioned at the maximum end of the axes, signifies users who consistently assign higher ratings to movies in all three clusters. This segmentation approach, cou-

pled with visual representation, provides a comprehensive understanding of user preferences and behaviors within the anime genre landscape.

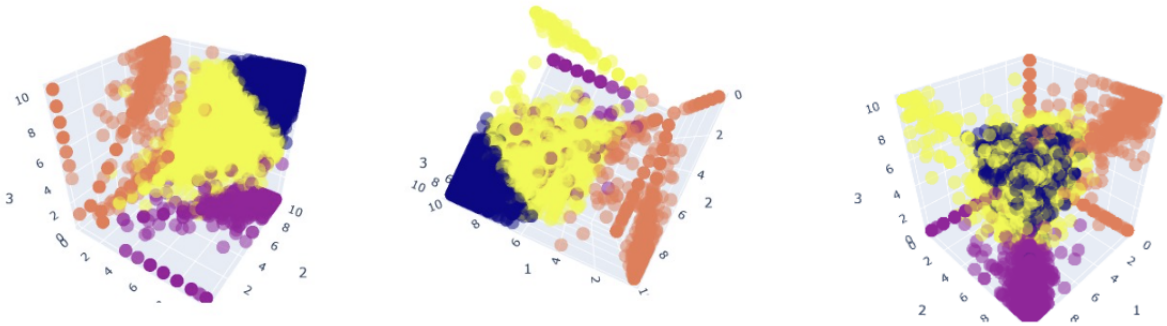


Figure 7: User Clusters

Additionally, the analysis uncovered an interesting trend by establishing a correlation between average anime ratings and their release years, revealing a fitting model in ARIMA(2,1,0) for predicting the average rating of forthcoming anime titles. This predictive model serves as a valuable tool in anticipating the potential success of future anime releases.

The SVD and KNN models are able to reliably predict user ratings with average mean absolute error of 0.87 and 0.95 respectively.

```
    Evaluating RMSE, MAE of algorithm SVD on 3 split(s).

                  Fold 1  Fold 2  Fold 3  Mean    Std
RMSE (testset)    1.1765  1.1749  1.1763  1.1759  0.0007
MAE (testset)     0.8790  0.8777  0.8789  0.8785  0.0006
Fit time          12.93   12.49   12.95   12.79   0.21
Test time         6.14    6.15    6.26    6.18    0.05
{'test_rmse': array([1.1765492 , 1.17493892, 1.17632392]),
'test_mae': array([0.87901695, 0.87771758, 0.87887208]),
'fit_time': (12.930427074432373, 12.493419647216797, 12.945121049880981),
'test_time': (6.144198894500732, 6.145112037658691, 6.2577879428863525)}

    Evaluating RMSE, MAE of algorithm KNNBasic on 3 split(s).

                  Fold 1  Fold 2  Fold 3  Mean    Std
RMSE (testset)    1.2941  1.2918  1.2924  1.2928  0.0010
MAE (testset)     0.9587  0.9581  0.9584  0.9584  0.0002
Fit time          190.73  171.00  173.67  178.47  8.74
Test time         598.17  549.93  575.58  574.56  19.71
{'test_rmse': array([1.29408973, 1.29182792, 1.29243483]),
'test_mae': array([0.95870405, 0.95810879, 0.9584189 ]),
'fit_time': (190.73062705993652, 170.99515986442566, 173.67348313331604),
'test_time': (598.1731100082397, 549.9259850978851, 575.5787467956543)}
```

As for user temporal analysis, the series autocorrelation and partial autocorrelation suggests that it would fit using AR(2) model which is ARIMA(2,1,0) for the pre-differencing data. As seen in the model diagnosis, the P-value for coefficients are lower than 0.05.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                   97
Model:                 ARIMA(2, 1, 0)   Log Likelihood                  19.945
Date:                Thu, 07 Dec 2023   AIC                            -33.890
Time:                        15:20:23   BIC                            -26.197
Sample:                             0   HQIC                           -30.780
                                 - 97
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5404      0.105     -5.164      0.000      -0.746      -0.335
ar.L2         -0.2422      0.123     -1.970      0.049      -0.483      -0.001
sigma2         0.0385      0.005      7.494      0.000       0.028       0.049
===================================================================================
Ljung-Box (L1) (Q):                   0.10   Jarque-Bera (JB):                23.19
Prob(Q):                              0.75   Prob(JB):                         0.00
Heteroskedasticity (H):               0.28   Skew:                            -0.82
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.77
===================================================================================
```

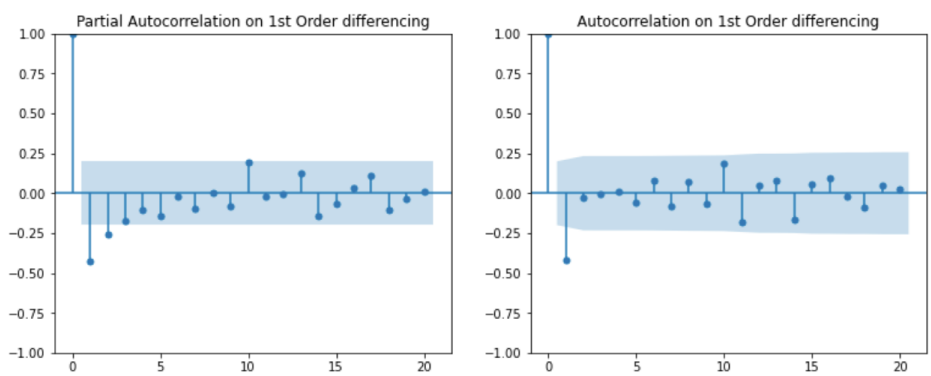


Figure 8: PCF and ACF plot of 1st order differencing of avgerage rating by release year

In our final analysis, we explore word clouds generated from the synopses of four distinct genres, revealing prevalent themes such as school, world, girl, and life. This insightful examination of genre-specific word clouds offers a concise glimpse into the recurring motifs and narrative elements characterizing each genre within the anime landscape.

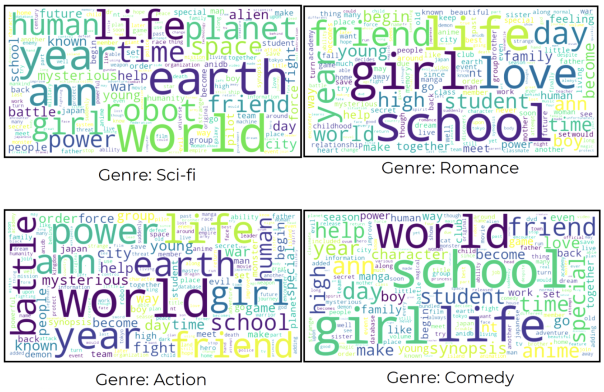


Figure 9: Word Clouds

6 Future Work

In future endeavors, enhancing the user segmentation methodology to encompass more intricate behavioral patterns and preferences holds the potential to unveil deeper insights into nuanced user groups within the anime community. Additionally, refining recommendation systems through the integration of hybrid models, merging genre similarities with collaborative filtering or content-based approaches, could amplify the accuracy and personalization of anime suggestions.

Furthermore, delving into advanced time series models beyond ARIMA, such as seasonal models or machine learning-based forecasting techniques, might enhance the precision of predicting average ratings for future anime releases based on their release years. Additionally, a more profound exploration of semantic analysis in word clouds, employing natural language processing (NLP) techniques, could reveal intricate genre-specific themes and sentiment analysis within anime synopses.

Moreover, incorporating sentiment analysis on user reviews and utilizing user sentiment as a feature in recommendation systems could provide a more nuanced understanding of user preferences. Finally, the exploration of collaborative filtering techniques on user review data, leveraging matrix factorization methods or deep learning architectures, has the potential to advance the accuracy of personalized anime recommendations based on peer reviews and user behaviors.

References

[1] Girsang, A. S., Al Faruq, B., Herlianto, H. R., Simbolon, S. (2020). Collaborative Recommendation System in Users of Anime Films. Journal of Physics: Conference Series, 1566(1), 012057. DOI:10.1088/1742-6596/1566/1/012057

[2] Lee, J. H., Shim, Y., Jett, J. (2015). Analyzing User Requests for Anime Recommendations. JCDL '15: Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries, 269–270. DOI:10.1145/2756406.2756969e

[3] Agarwal, K. (2022, February 17). Analysis of Anime Ratings Dataset: https://medium.com/@krishnax16/analysis-of-anime-ratings-dataset-54cbde90f23d

[4] How Does Collaborative Filtering Work in Recommender Systems. Turing Knowledge Base. https://www.turing.com/kb/collaborative-filtering-in-recommender-system

[5] Hug, N. (2015). Surprise' Python Library documentation. Welcome to Surprise' documentation! - Surprise 1 documentation. https://surprise.readthedocs.io/en/stable/