Data Analysis on Anime Recommendation System using MyAnimeList dataset

Pavan Sundar Reddy Guthikonda  
 Master’s in Data Science  
 University of Missouri-Kansas City, Missouri, USA  
 pgq66@umkc.edu

Padma Sai Paladugu  
 Master’s in Computer Science  
 University of Missouri-Kansas City, Missouri, USA  
 pp9hk@umkc.edu

Chandu Bodepudi  
 Master’s in Computer Science  
 University of Missouri-Kansas City, Missouri, USA  
cbthz@umkc.edu

ABSTRACT

The COVID-19 pandemic brought about a transformative shift in daily life, with lockdowns and social distancing measures prompting a surge in demand for online entertainment. Amid this surge, anime, a dynamic and diverse form of animated entertainment originating in Japan, became a captivating source of solace and diversion. Its captivating storytelling and artistic expression, free from language barriers, resonated with a global audience. As viewers turned to online streaming platforms, the challenge of offering personalized recommendations for anime content emerged. Anime enthusiasts sought fresh and engaging content tailored to their unique preferences. This research endeavors to address this challenge by harnessing big data concepts. Our research begins by ingesting and storing data in the resilient Hadoop Distributed File System (HDFS), ensuring scalability and fault tolerance. Subsequently, PySpark is employed to process the data, encompassing vital tasks such as data cleaning, transformation, and format standardization, including addressing missing values, eliminating duplicates, and establishing data uniformity. To create a comprehensive and engaging user experience, Tableau is utilized for data visualization, enabling the development of dynamic dashboards and visualizations. These tools not only showcase recommendation results but also provide insights into user engagement metrics. At the core of our research lies the implementation of collaborative filtering algorithms, making use of PySpark's machine-learning libraries to extract valuable insights from the processed data. By considering a user's viewing history, ratings, and preferences, we aim to provide personalized anime suggestions that transcend language barriers, enriching the streaming experience. In an era characterized by evolving media consumption habits and a growing desire for cross-cultural content, our research seeks to bridge the gap between entertainment and technology. It aims to provide a tailored streaming experience that empowers users to explore, enjoy, and connect with anime in a personalized and meaningful way.

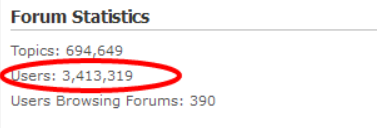
KEYWORDS

COVID-19 pandemic, online entertainment, anime, personalized recommendations, big data, Hadoop Distributed File System, PySpark, data processing, data visualization, collaborative filtering, user engagement, and cross-cultural content.

Pavan Sundar Reddy Guthikonda, Padma Sai Paladugu, Chandu Bodepudi. 2023. Data Analysis on Anime Recommendation System using MyAnimeList dataset.

**INTRODUCTION**

The world of anime, a dynamic and diverse form of animated entertainment originating in Japan, has captivated a global audience with its captivating storytelling and artistic expression. As the popularity of anime continues to soar, one platform has emerged as a central hub for enthusiasts and fans alike - MyAnimeList (MAL). MyAnimeList, often abbreviated as MAL, stands as a testament to the worldwide fervor for anime. It serves as an online platform and community-driven website dedicated to anime and manga enthusiasts. With a vast and growing user base, MAL has become an indispensable resource for anime fans. As of 2023, the MyAnimeList community boasts an impressive count of 3,413,319 registered users.



**Figure**-**1**: MyAnimeList Users Count

This thriving digital ecosystem offers users a plethora of features, from tracking their anime and manga consumption to rating and reviewing titles, engaging in forums, and discovering new favorites. One significant aspect of MAL's role in the anime community is its mobile application, which brings the world of anime closer to users' fingertips. The MyAnimeList application has seamlessly integrated into the lives of anime enthusiasts, allowing them to carry their passion in their pockets. Users can access their watchlists, discover new series, rate and review titles, and participate in discussions, all while on the go. This introduction sets the stage for a closer examination of MyAnimeList, its vast user base, and the significance of its application in enhancing the anime experience for users worldwide. In this digital age, where accessibility and connectivity are paramount, MAL has become an essential companion for anime aficionados, shaping their viewing experiences and fostering a sense of community among fans.

**RELATED WORK**

Recommendations on Netflix: Diving Deeper into the Matter (Malinowski, J., & Zimányi, E., 2019). This study explores the intricacies of recommendations on the Netflix platform, delving into the underlying mechanisms that contribute to personalized content suggestions. By presenting their findings at the 10th ACM Conference on Recommender Systems, Malinowski and Zimányi contribute valuable insights into the evolving landscape of recommender systems, specifically focusing on the widely popular streaming service, Netflix.

Matrix Factorization Techniques for Recommender Systems (Koren, Y., Bell, R., & Volinsky, C., 2009). Koren, Bell, and Volinsky delve into matrix factorization techniques as applied to recommender systems. This seminal work, published in the journal Computer, emphasizes the use of matrix factorization to enhance the accuracy of recommendations. The authors explore the collaborative filtering approach, shedding light on its effectiveness and providing a foundational understanding of the techniques employed in building robust recommender systems.

Collaborative Filtering Recommender Systems (Ekstrand, M. D., Riedl, J. T., & Konstan, J. A., 2011). In this comprehensive review published in Foundations and Trends in Human-Computer Interaction, Ekstrand, Riedl, and Konstan delve into collaborative filtering recommender systems. The authors provide an in-depth analysis of collaborative filtering techniques, offering a historical perspective and summarizing the foundational principles that guide the development of such systems. This work serves as a key reference for researchers and practitioners in the field of recommender systems.

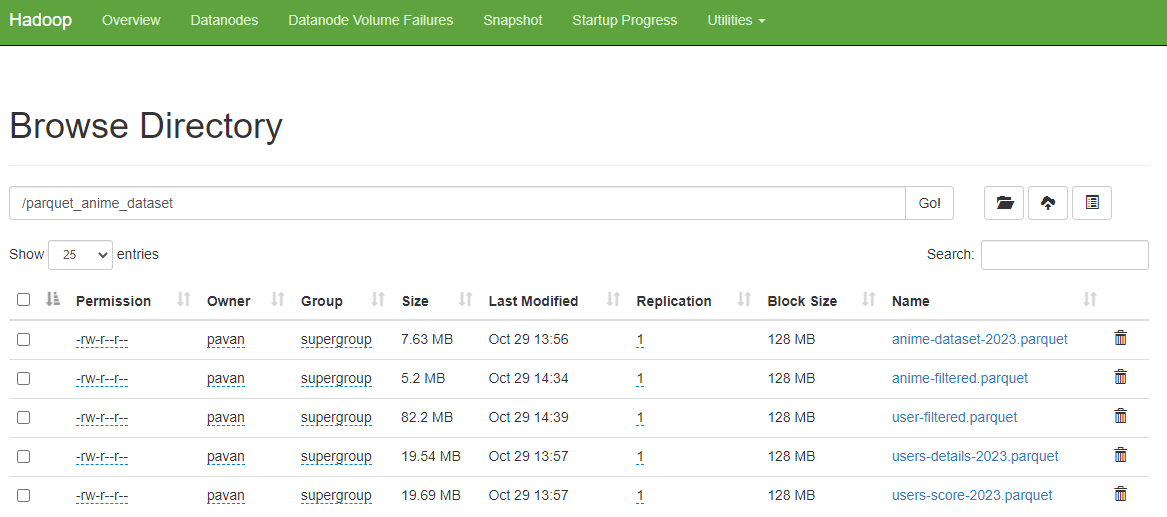
A Survey of Link Recommendation Systems in the Context of Social Networking Sites (Cano, E., & Scherp, A., 2018). Cano and Scherp contribute to the understanding of link recommendation systems, particularly within the context of social networking sites. Their survey, published in ACM Computing Surveys, reviews the landscape of link recommendation systems, providing insights into the challenges and strategies employed in suggesting relevant links within social networks. This work is valuable for researchers exploring recommendation systems beyond content and entertainment platforms.

Supervised Machine Learning: A Review of Classification Techniques (Kotsiantis, S., Zaharakis, I., & Pintelas, P., 2007). Kotsiantis, Zaharakis, and Pintelas offer a comprehensive review of supervised machine learning techniques, with a specific focus on classification. Published in Informatica, this review provides an overview of classification techniques, their applications, and the evolving landscape of supervised machine learning. The work is foundational for understanding the broader machine-learning context and its relevance to recommendation systems.

**PROPOSED TECHNIQUES**

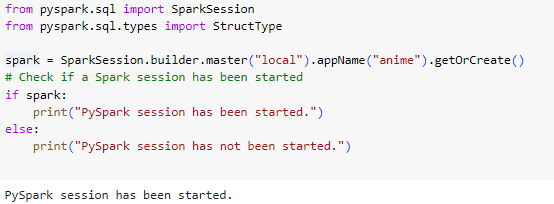
The proposed technique for data analysis of an anime recommendation system using the MyAnimeList dataset comprises four main steps:

1. **Data Ingestion and Storage**: The first step involves collecting data from the MyAnimeList dataset and storing it in a scalable and fault-tolerant storage system, such as Hadoop Distributed File System (HDFS). This ensures that the dataset is readily accessible for analysis. Convert CSV to Parquet and store it in HDFS



**Figure-2:** HDFS data ingestion

2. **Data Processing with PySpark**: PySpark is utilized to clean, transform, and prepare the dataset. Tasks include addressing missing values, removing duplicates, and standardizing data formats. PySpark's distributed computing capabilities enable efficient data processing.



**Figure**-3: PySpark Setup

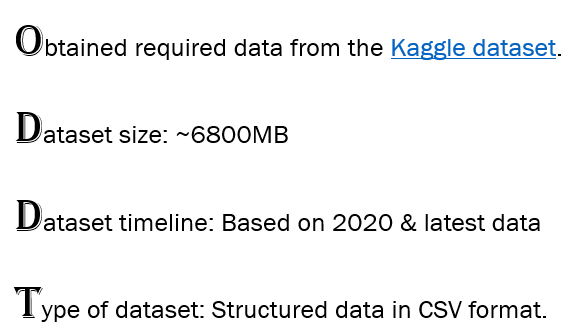
3. **Data Visualization Using Matplotlib**: The analyzed data is visualized using Tableau, creating interactive dashboards and visualizations. These tools present recommendation results and user engagement metrics in an accessible and user-friendly manner.

4. **Collaborative Filtering Algorithms**: The core of the technique involves implementing collaborative filtering algorithms using PySpark's machine-learning libraries. These algorithms consider a user's viewing history, ratings, and preferences to generate personalized anime suggestions, enhancing the overall streaming experience.

This integrated approach leverages data processing, visualization, and collaborative filtering to develop a robust recommendation system that offers personalized anime suggestions, improving the user experience on platforms like MyAnimeList.

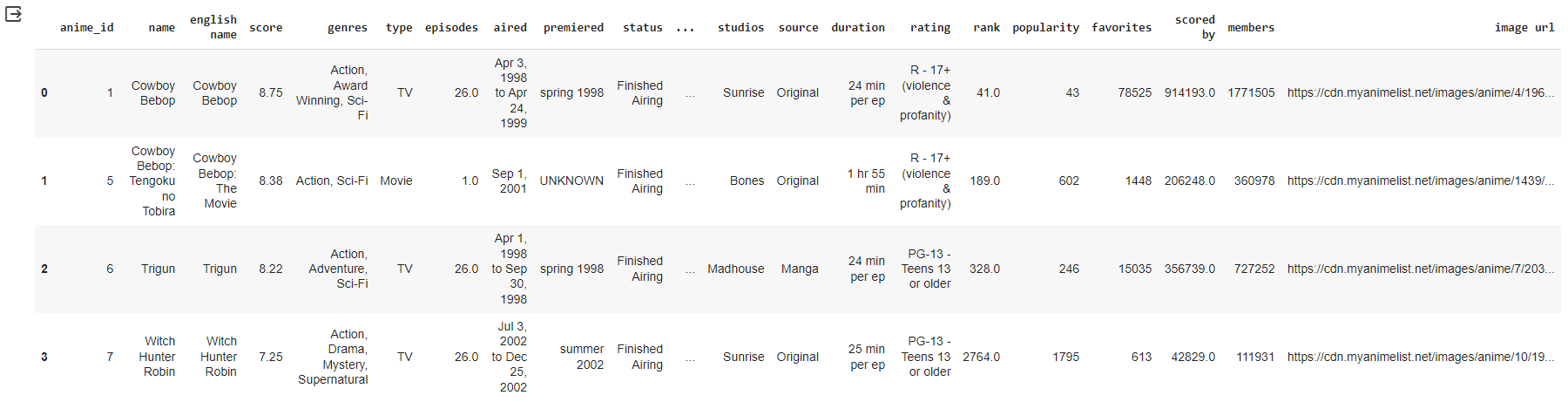
**DATA AND METHODOLOGY**

The dataset was obtained from Kaggle, a popular platform for data science and machine learning datasets. The dataset size is approximately 6800MB, indicating a substantial amount of data for analysis. The dataset encompasses information based on the year 2020 and includes the latest available data. This timeline allows us to capture recent trends and developments in the anime domain. The dataset is structured and is provided in CSV (Comma-Separated Values) format. Structured data implies that the information is organized in a tabular format with well-defined columns and rows.



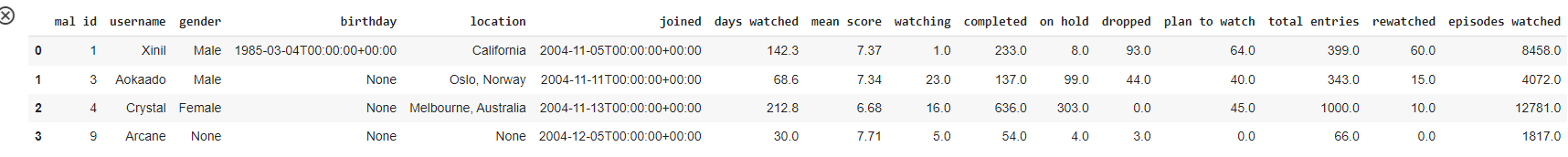
**Figure-2:** Data information

anime-dataset-2023.csv dataset contains the columns listed as 'anime\_id', 'name', 'english name', 'score', 'genres', 'type', 'episodes', 'aired', 'premiered', 'status', 'producers', 'licensors', 'studios', 'source', 'duration', 'rating', 'rank', 'popularity', 'favorites', 'scored by', 'members' and 'image url'.



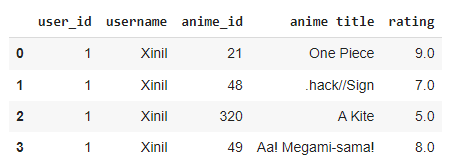
**Figure-3:** anime-dataset-2023.csv

users-details-2023.csv dataset contains the columns listed as 'mal id', 'username', 'gender', 'birthday', 'location', 'joined', 'days watched', 'mean score', 'watching', 'completed', 'on hold', 'dropped', 'plan to watch', 'total entries', 'rewatched' and 'episodes watched'.



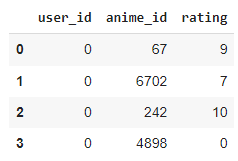
**Figure-4**: users-details-2023.csv

users-score-2023.csv dataset contains the columns listed as 'user\_id', 'username', 'anime\_id', 'anime title' and 'rating'.



**Figure-5**: users-score-2023.csv

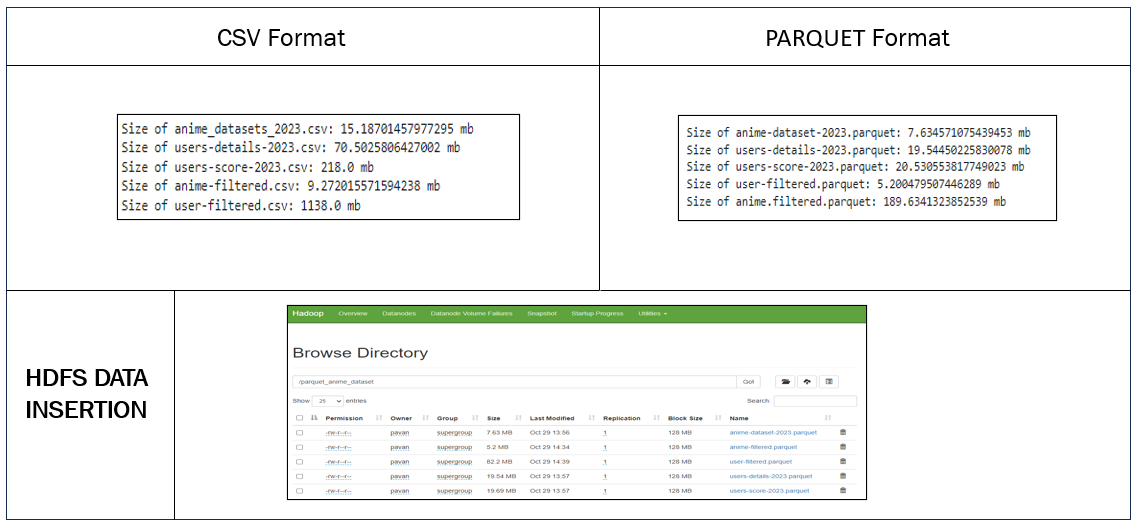
users-filtered.csv dataset contains the columns listed as 'user\_id', 'username', 'anime\_id', 'anime title' and 'rating'.



**Figure-6**: user-filtered.csv

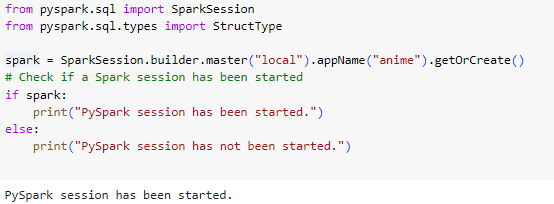
**WORK DONE**

To optimize storage and enhance data retrieval efficiency, the CSV dataset was converted to the Parquet format. Parquet offers benefits such as improved compression, faster query performance, and better compatibility with data processing tools, making it an ideal choice for large datasets and analytical workloads. All the datasets have been stored in HDFS.



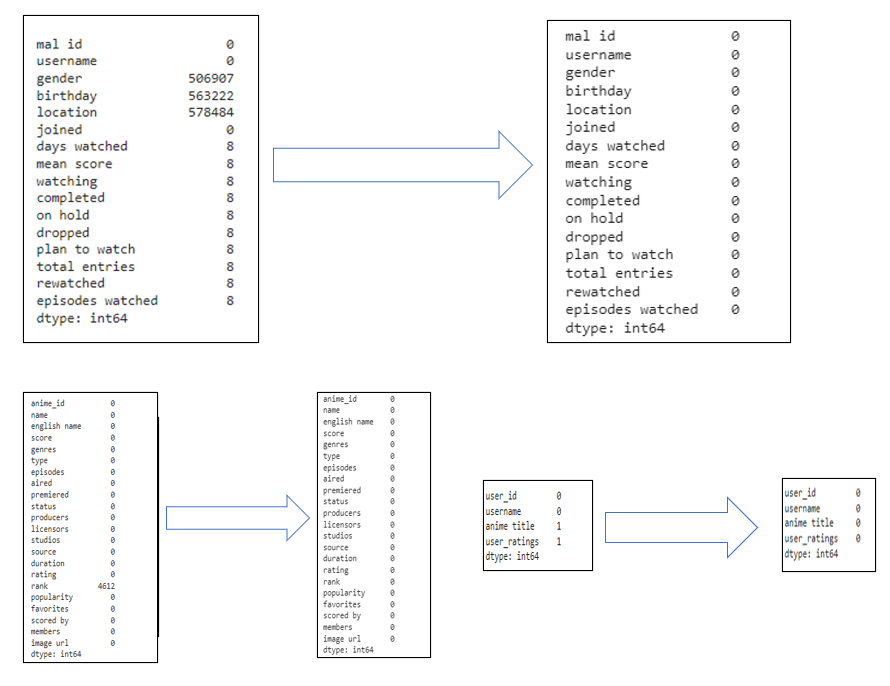
**Figure-7:** CSV to Parquet format. HDFS data insertion

PySpark is utilized to clean, transform, and prepare the dataset. Tasks include addressing missing values, removing duplicates, and standardizing data formats. PySpark's distributed computing capabilities enable efficient data processing.



**Figure-8**: PySpark Setup

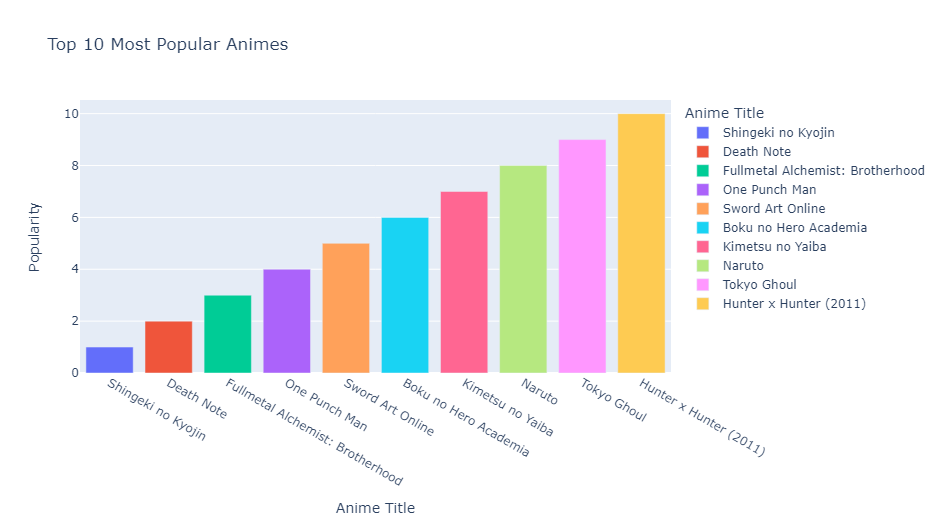
In all the datasets, for columns that contain categorical data (e.g., 'gender' in users-details-2023.csv), the fillna() method is used. This method fills missing values with a specified value or a calculated value. In the context of categorical data, it might involve filling missing values with the most frequent category (mode) or a specific category that makes sense in the given context. For columns that contain numerical data (e.g., 'score' in anime-dataset-2023.csv), the mean() method is used. This involves filling missing numerical values with the mean (average) value of the available data. It is a common strategy when dealing with numerical features to maintain the overall statistical characteristics of the dataset.



**Figure-9**: Before & after comparison of datasets

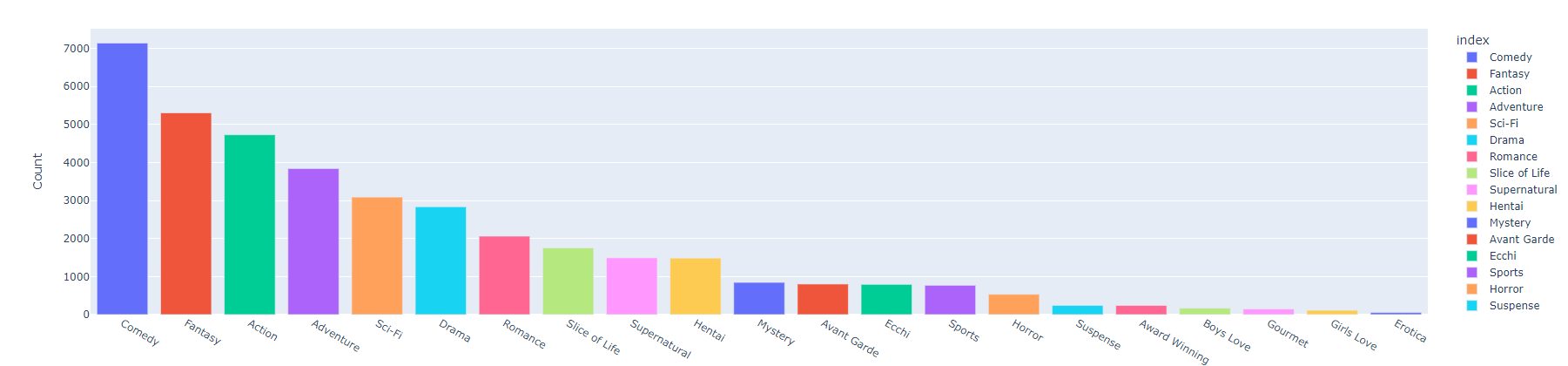
The analyzed data is visualized using Tableau, creating interactive dashboards and visualizations. These tools present recommendation results and user engagement metrics in an accessible and user-friendly manner.

Below figure-10 is a visually appealing bar chart that illustrates the distribution of anime titles across different types, making it easy to see which types have a higher or lower count.



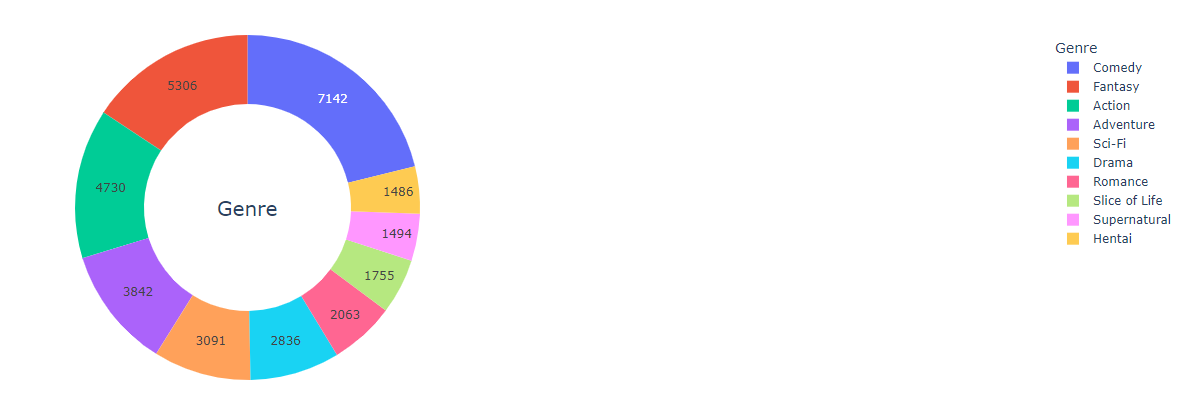
**Figure-10:** Top 10 Most Popular Animes

Below figure-11 is a representation of the distribution of anime titles across different genres. It helps in understanding which genres are more prevalent in the dataset.



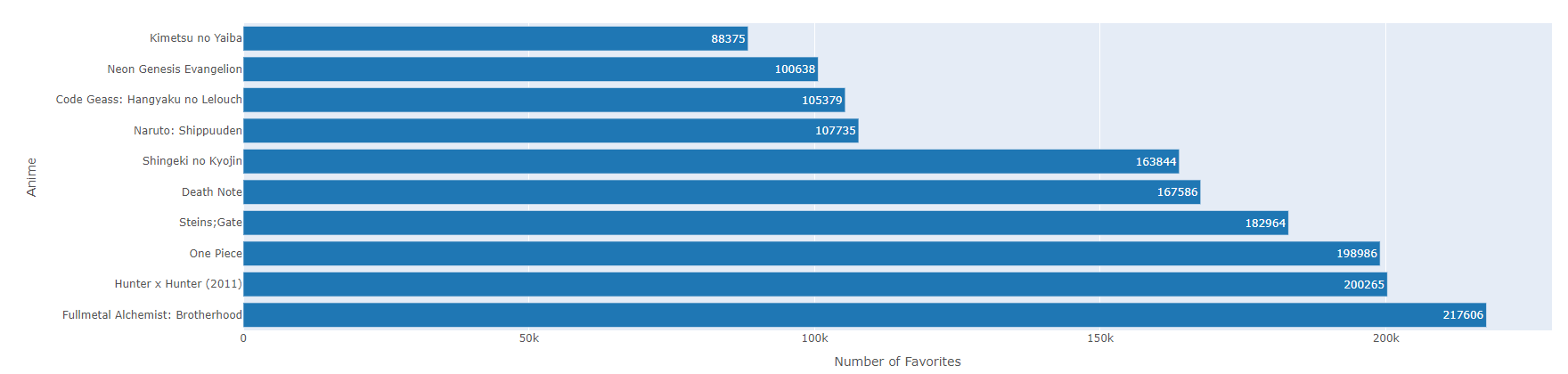
**Figure-11:** Count of Anime Titles by Genre

Below pie chart shows the distribution of anime genres, allowing you to quickly grasp which genres are more prevalent in the dataset and their relative proportions.



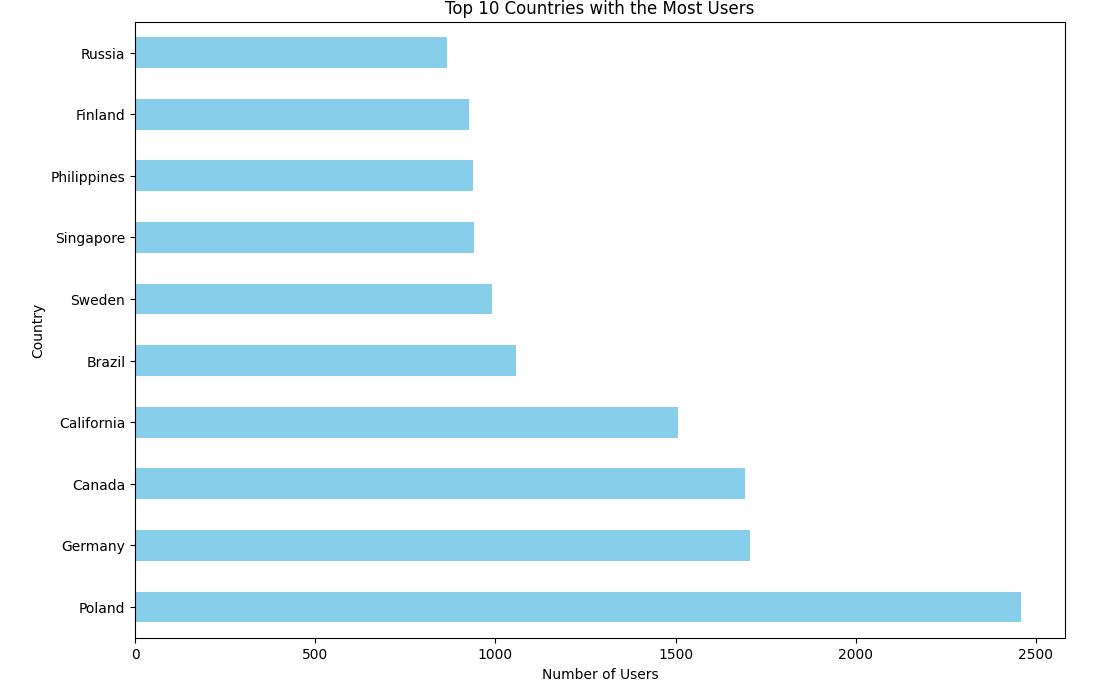
**Figure-12:** Distribution of Anime Genres

Below chart provides a quick visual summary of the most favorited anime in the dataset, allowing viewers to identify the top favorites and compare their popularity.



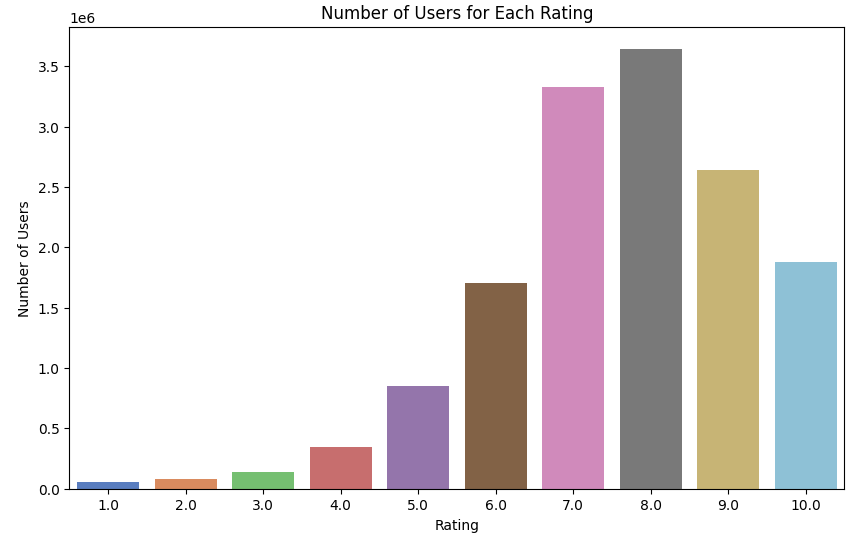
**Figure-13:** Top 10 Most Favorited Anime

Below figure visualizes the top 10 countries with the most users



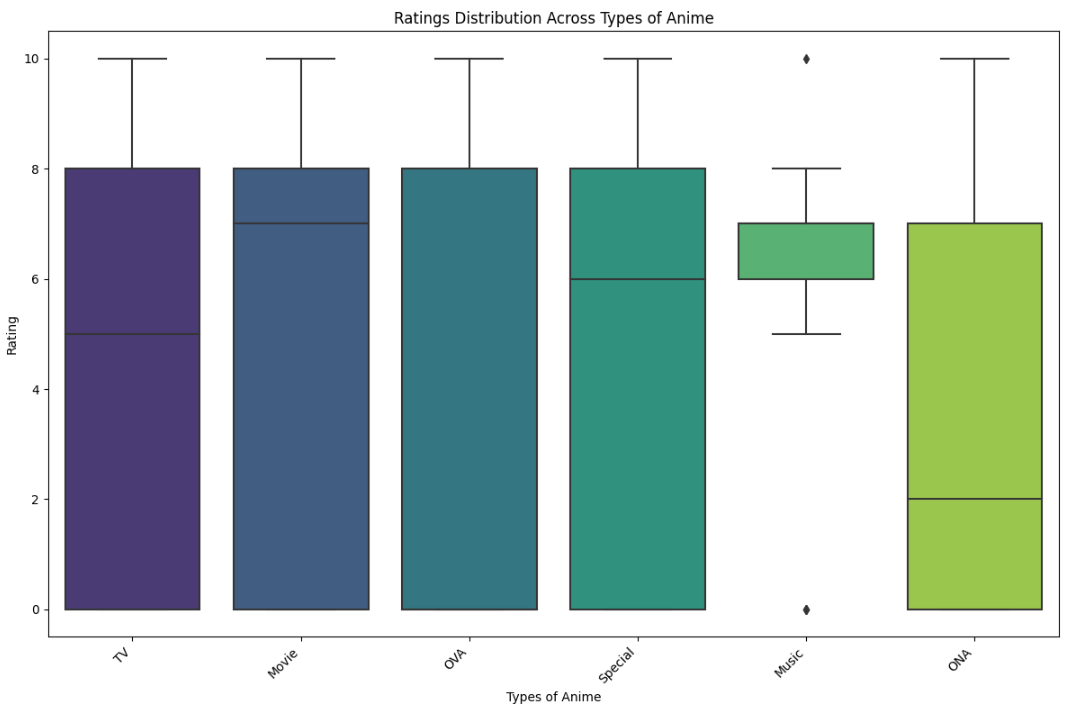
**Figure-14: T**op 10 countries with the most users

This visualization provides insights into how many users gave each rating, offering a distribution of user preferences in terms of anime ratings.

****

**Figure-15:** Number of Users for Each Rating

This visualization allows for a comparative analysis of the distribution of ratings among different types of anime, offering insights into how the audience perceives and rates anime based on their types.

****

**Figure-16:** Ratings Distribution Across Types of Anime

**ANALYSIS & RESULTS**

AUTHOR CONTRIBUTIONS:

Author 1 was primarily responsible for data ingestion and storage. They managed the process of collecting the MyAnimeList dataset and converting it from CSV to Parquet format. This author ensured that the data was efficiently stored in the Hadoop Distributed File System (HDFS) for subsequent analysis. Additionally, Author 1 played a key role in coordinating tasks and ensuring smooth communication within the team. They facilitated collaboration and acted as a liaison between other team members. Prepared well-structured presentation to demonstrate.

Author 2 took charge of data visualization using Matplotlib, and Pyplot. They created preliminary visualizations, including interactive dashboards and charts, to present insights from the dataset. These visualizations offered initial perspectives on anime genre distributions, user activity metrics, and dataset characteristics. Author 2 also contributed to data quality control by identifying and addressing issues in the dataset, ensuring that the visualization data accurately represented the underlying dataset.

Author 3 is primarily responsible for the upcoming steps in the project. They will lead the implementation of machine learning (ML) algorithms using PySpark's machine-learning libraries to develop predictive models. This author will also take the lead in the development of the anime recommendation system. Author 3 is engaged in ongoing research and analysis to inform the ML algorithms and recommendation system. They are responsible for staying up to date with the latest research and best practices in the field.

Collectively, the team members work collaboratively to ensure the success of the project, with each member contributing their expertise in their assigned areas. This division of tasks enables the project to progress efficiently while leveraging each team member's strengths and skills.

REFERENCES

1. Malinowski, J., & Zimányi, E. (2019). Recommendations on Netflix: Diving deeper into the matter. In Proceedings of the 10th ACM Conference on Recommender Systems (pp. 405-409).
2. Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8), 30-37.
3. Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2011). Collaborative filtering recommender systems. Foundations and Trends® in Human-Computer Interaction, 4(2), 81-173.
4. Cano, E., & Scherp, A. (2018). A survey of link recommendation systems in the context of social networking sites. ACM Computing Surveys (CSUR), 51(6), 119.
5. Kotsiantis, S., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. Informatica, 31(3), 249-268.Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00