DATA SCIENCE FINAL PROJECT: DATA JOURNALISM WITH ANIME DATASETS

Kıvanç Adıgüzel
191180003

Ümmü Nur Gülmez
191180762

Yasin Alihan Demir
191180753

Venonat

ABSTRACT

In this paper, we try to explain what did we do for the final project. Data journalism involves using data as a primary source to uncover and tell stories. It involves analyzing and visualizing data in order to communicate findings and insights to a wider audience. Anime datasets can be a rich source of information for data journalism, as they often contain a large amount of data on characters, plotlines, and other aspects of anime series and films. By analyzing these datasets, data journalists can uncover trends and patterns within the anime industry, and use this information to tell compelling stories about the genre and its fans. Data journalism with anime datasets can also help to shed light on issues such as representation and diversity within the industry, and can be a powerful tool for bringing attention to these important topics.

Keywords

Anime, watchers, right target, users, audience, data journalism.

1. INTRODUCTION

Data journalism is the practice of using data as a means of finding, analyzing, and presenting information in a way that tells a story or informs the public. In recent years, there has been a growing interest in using data to understand and report on various aspects of society, including politics, economics, and culture. One increasingly popular area of data journalism is the use of anime datasets.

Anime, a term used to refer to animated productions from Japan, has gained a significant following around the world. There is a wealth of data available about anime, including information about the characters, storylines, and production details of various shows and movies. By analyzing this data, journalists can gain insights into the trends and patterns that shape the anime industry, and use this information to inform their reporting and storytelling.

In this paper, we will explore the potential of using anime datasets in data journalism. We will discuss the types of data that are available, the tools and techniques that can be used to analyze this data, and the ways in which this information can be used to report on and tell stories about the world of anime. We will also examine some of the ethical considerations that arise when using data in journalism, and suggest best practices for ensuring that data is used responsibly and transparently.

2. METHODOLOGY

The purpose of this study is to analyze anime datasets more effectively and finding answers for these questions:

• How should we use the anime datasets more efficient?

• How can we offer a better user experience?

• How can we show data journalism over it?

• How should we focus on the right target for users?

• How can we turn audience's data useable?

• Also, we trying to find an answer to what should we don't have to do?

Firstly, we collect and organize data. We found and downloaded datasets from online sources and Kaggle. Then we clean and preprocess the data. This step involves checking the data for duplicated or not and formatting it in a way that is suitable for analysis. We start to explore and analyze the data. This step involves using statistical and visualization techniques to understand patterns and trends in the data. This step also includes creating graphs, plots, and tables to visualize the data and using statistical tests to determine the significance of any relationships. Our main goal is offering a better user experience. Because of that we train our datasets. While training we use collaborative filtering and content-based filtering. Collaborative filtering is a technique that can filter out items that a user might like on the basis of reactions by similar users. It works by searching a large group of people and finding a smaller set of users with tastes similar to a particular user. We use cosine similarity which is a metric used to measure how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together. The smaller the angle, higher the cosine similarity. Content-based filtering is a method of recommendation in which

the characteristics of an item (in this case, an anime series) are used to make recommendations to a user. In content-based filtering, a profile is created for each user based on their preferences or history of interactions with items, and recommendations are made based on the similarity between the profile of the user and the characteristics of the items. To apply content-based filtering to an anime dataset, we needed to collect data on the characteristics of different anime series. To make recommendations to a particular user, we compare the profile of the user to the characteristics of different anime series and recommend the ones that are most similar to the user's profile.One challenge with content-based filtering is that it relies on having a good representation of the characteristics of the items. This may be difficult to achieve for a large and diverse dataset like an anime dataset, as it can be difficult to capture all of the relevant characteristics of an anime series in a simple and concise manner. In these cases, a hybrid approach that combines content-based filtering with collaborative filtering (which takes into account the preferences of other users) may be more effective.


```
In [45]:   1   give_recommendation("Death Note")
```

Recommendations for Death Note viewers :

Out[45]:

No	Anime Name	Rating
1	Hachimitsu to Clover Specials	7.850000
2	Trapp Ikka Monogatari	7.750000
3	Major S1	8.420000
4	Hakkenden: Touhou Hakken Ibun	7.570000
5	Nazotokine	4.860000
6	ef: A Tale of Melodies.	8.180000
7	Saki Achiga-hen: Episode of Side-A Specials	7.630000
8	One Piece: Oounabara ni Hirake! Dekkai Dekkai Chichi no Yume!	7.430000
9	Kizumonogatari II: Nekketsu-hen	8.730000
10	Go! Princess Precure Movie: Go! Go!! Gouka 3-bondate!!!	6.890000

Figure 1. Recommendation for Death Note Viewers


```
In [44]:   1   give_recommendation("Naruto")
```

Recommendations for Naruto viewers :

Out[44]:

No	Anime Name	Rating
1	To LOVE-Ru Darkness OVA	7.820000
2	Hanbun no Tsuki ga Noboru Sora	7.690000
3	Mai-HiME	7.590000
4	Doraemon Movie 28: Nobita to Midori no Kyojin Den	7.540000
5	Rurouni Kenshin Special	7.510000
6	Pikmin Short Movies	7.270000
7	Deadman Wonderland OVA	7.120000
8	Marimo no Hana: Saikyou Butouha Shougakusei Densetsu	5.120000
9	Shinpi no Hou	5.370000
10	Toaru Majutsu no Index: Endymion no Kiseki	7.710000

Figure 2. Recommendation for Naruto Viewers

3. DATASETS

- profiles.csv
- anime.csv
- anime2.csv
- ONE PIECE.csv
- rating.csv
- animes.csv
- reviews.csv
- profiles.csv

4. EXPERIMENTS
4.1. General Anime Analysis

Collaborative filtering is a method of making recommendations based on the preferences of a group of users. It works by predicting the rating or preference a user would give to an item, based on the ratings and preferences of similar users. Content-based filtering is a method of making recommendations based on the characteristics of an item. It works by analyzing the attributes of the items that a user has liked in the past, and using that information to recommend similar items. Both collaborative filtering and content-based filtering can be used to improve the user experience with an anime dataset. Collaborative filtering can help identify anime that are popular among a user's peers, while content-based filtering can help identify anime that are similar to ones that a user has previously enjoyed. It is possible to combine collaborative filtering and content-based filtering in a hybrid approach, which can provide more accurate recommendations by considering both the preferences of similar users and the characteristics of the anime. By using two methods we can offer different anime for viewers of a specific anime.

We also tried to offer a truer perspective to anime by analyzing datasets. First, we clean datasets. The img_url and link columns were removed because it contained meaningless data. Then we also removed missing values from other columns. After cleaning datasets, we started to analyze. We found correlation between Ratings & Popularity and Ranked & Popularity.
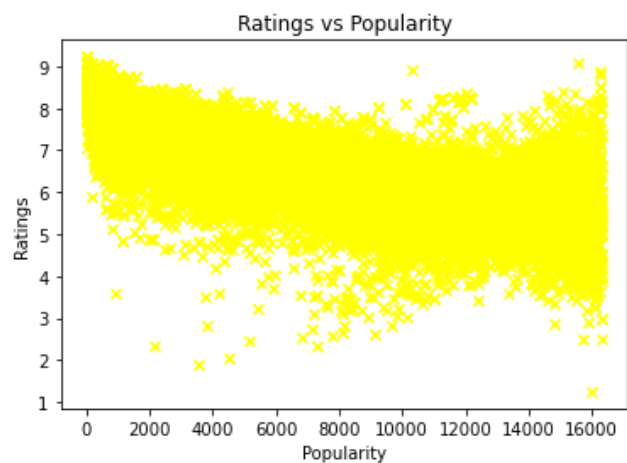

Figure 1. Ratings vs Popularity

We analyzed sources of anime to find their inspiration sources. According to our datasets the majority of anime were created with original storylines.
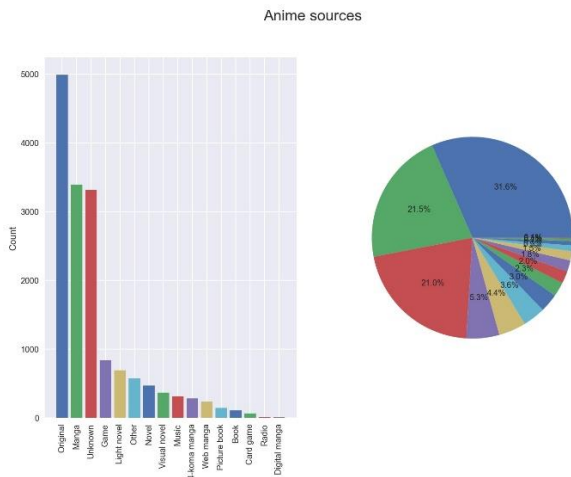

Figure 2. Shows Anime Sources

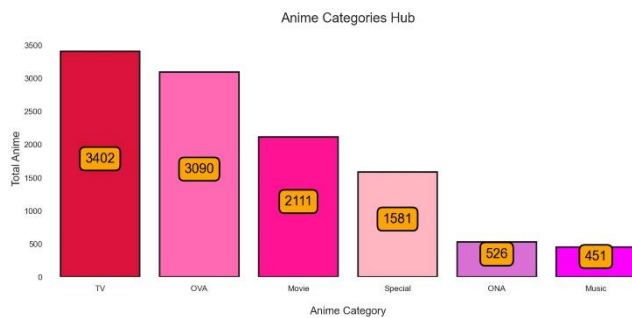Anime are made for different platforms so they have dissimilar formats. Most of them were made for TV.

Figure 3. Shows Anime Categories

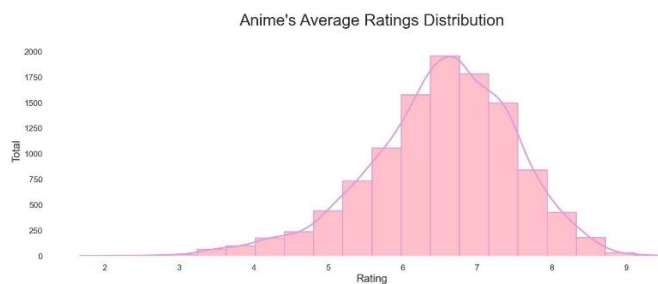We analyzed anime ratings to observe how much they were watched and to obtain an average value. We realized that average rating of anime are around between 6-7.

Figure 4. Shows Anime's Average Ratings

Anime has certain fan bases. We compared communities to find the anime with the largest fan base:
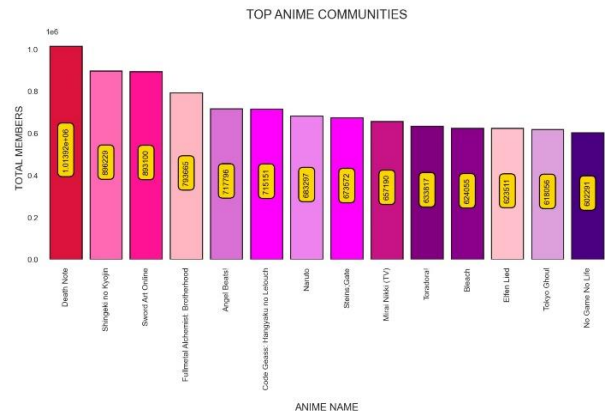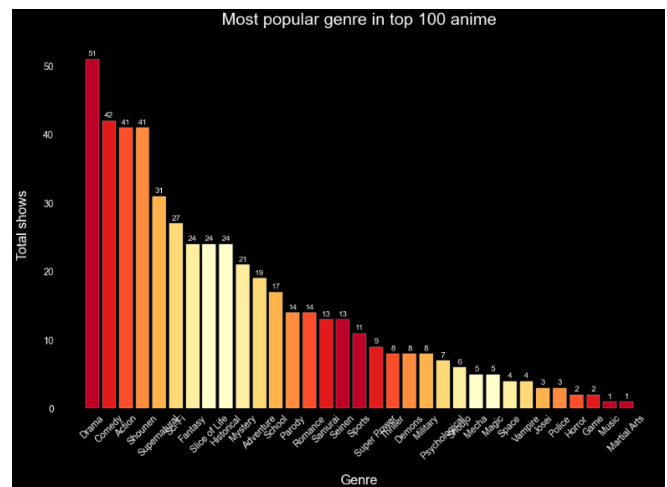

Figure 5. Shows Top Anime Communities

We found drama as a most popular genre:

Figure 6. Shows Most Popular Genres

We examined the number of anime that have been released since 1920 to observe whether interest in anime has increased or decreased, and we saw that there has been a significant increase:

Figure 7. Shows Anime Released by Year Since 1920

Also we found correlation between age & ratings:
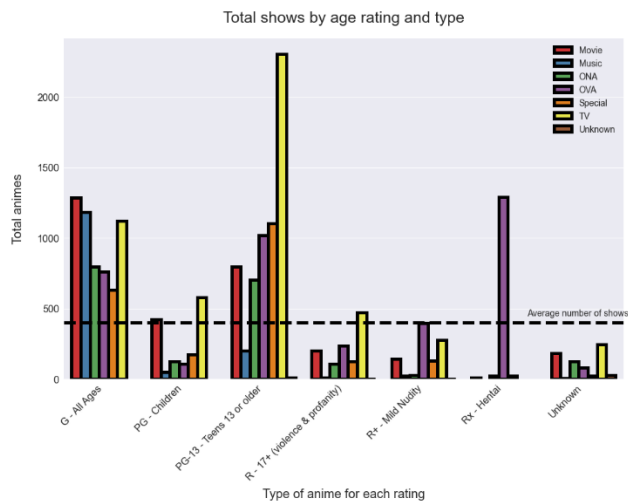

Figure 8. Shows Total Shows by Age Rating and Type

4.2. One Piece Anime Analysis

Finally, we aimed to obtain more specific results by analyzing the One Piece anime in particular. One Piece had the most episodes released in 2014 and the least number of episodes in 1999.
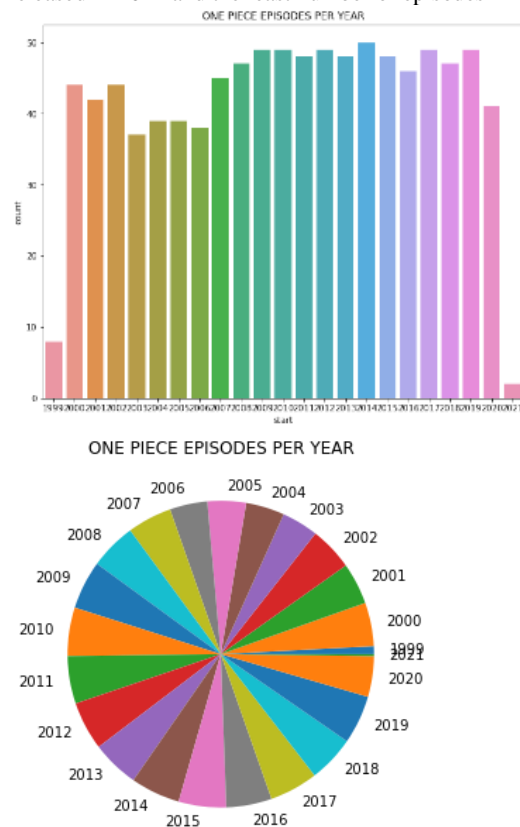

Figure 1-2. Shows One Piece Episodes Per Year

After analyzing the ratings for One Piece, we observed that it was most watched in 2015 and least watched in 2012:
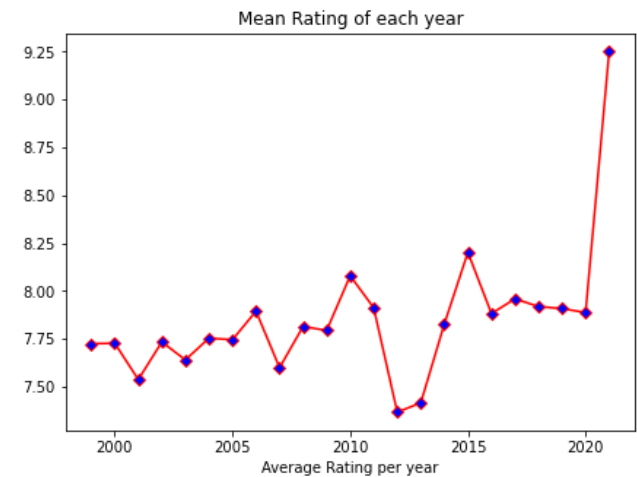

Figure 3. Shows One Piece Mean Rating of Each Year

Finally, we reached the conclusion that the most watched episode of the One Piece anime is "Kanashiki Kettou: Luffy tai Sanji (Kouhen)." The least watched episode is "Going into Action! The Implacable New Admiral of the Fleet." The difference between the most and least watched episodes was least in 1999 and highest in 2017.
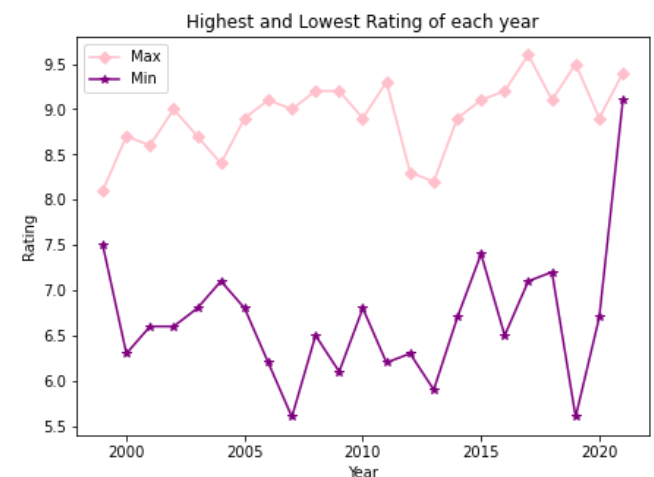

Figure 4. Shows One Piece's Highest & Lowest Rating of Each Year

5. RESULTS

Our analysis of the anime dataset revealed several interesting trends and patterns in the industry.

First, we found that certain genres of anime, such as drama and comedy, are consistently more popular than others, as measured by average ratings and viewership. This suggests that these genres have a broad appeal among anime fans.

We also discovered that the popularity of specific anime series tends to vary significantly over time. For example, we found that some series that were initially well-received by critics and audiences declined in popularity after the first season, while others gained in popularity as the series progressed.

Furthermore, we observed that interest in anime has significantly increased over the years. Overall, our analysis highlights the complexity and diversity of the anime industry, and suggests that there are many factors that influence the success of an anime series.

6. CONCLUSIONS

In conclusion, collaborative filtering and content-based filtering are effective methods for improving the user experience with an anime dataset. By analyzing the ratings and preferences of similar users, collaborative filtering can help identify popular anime among a user's peers. On the other hand, content-based filtering can help recommend anime that are similar to ones that a user has previously enjoyed based on the characteristics of the anime. By using these two methods, it is possible to offer a variety of anime recommendations to viewers of a specific anime. Through our analysis of anime datasets, we were able to gain a deeper understanding of the anime industry. By cleaning and organizing the data, we were able to identify trends and patterns in the ratings, popularity, and genre of anime. We also found that the majority of anime are created with original storylines and are primarily made for TV. Additionally, we observed a significant increase in the number of anime released since 1920, indicating a growing interest in the medium. Overall, our analysis has provided valuable insights into the world of anime and has allowed us to offer more personalized and accurate recommendations to users.

7. REFERENCES

[1] https://www.kaggle.com/datasets/a9ece97f83e99ab5955ddf7ab9c3f3a9047ba5cdbb08189b11e7243630d969d8

[2] https://data.world/datasets/anime

[3] https://www.kaggle.com/datasets/marlesson/myanimelist-