Assignment Report

The Influence of Anime

Rahul Kuruppassery Johny

Rohit Shinde

6th December, 2022

Declaration on Plagiarism

This form must be filled in and completed by the student submitting an assignment

Name/s:	Rahul Kuruppassery Johny Rohit Shinde
Student Number/s:	22262108 22260213
Programme:	MSc in Computing FT
Module Code:	CA682
Assignment Title:	Data Visualisation
Submission Date:	6th December 2022
Module Coordinator:	Dr Suzanne Little

I/We declare that this material, which I/we now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I/We understand that plagiarism, collusion, and copying are grave and serious offenses in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I/We have read and understood the Assignment Regulations. I/We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the sources cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.

I/We have read and understood the referencing guidelines found at
http://www.dcu.ie/info/regulations/plagiarism.shtml, https://www4.dcu.ie/students/az/plagiarism
and/or recommended in the assignment guidelines

Name: Rahul Kuruppassery Johny Date: 06-Dec-2022

Name: Rohit Subhash Shinde Date: 06-Dec-2022

Abstract

Anime is Japanese animation that is both hand-drawn and computer-generated. Outside of Japan and in English, anime specifically refers to Japanese animation. Animes have become a major source of digital entertainment, with stories frequently set in beautiful or terrifying magical worlds. As with many works of fiction, one can escape to another world, identify with charismatic people, learn about a different culture, or simply enjoy the art itself. They have increased in popularity tremendously over the years, particularly among Western audiences. As a result, human inventiveness, globalization, and commercialism have produced a rich byproduct [1]. In this context, we hope to provide meaningful visualisations of the raw numbers gathered from the pieces of art that we admire. We highlight all animes registered on MyAnimeList.net from 1917 to 2018, and their diversity through genre classification, Furthermore, we prioritize an engaging, elegant, and entertaining user experience so that our visualisations are not only instructive but also visually appealing. The visualisations showed us various different things like which genres are popular among the different genders, the top-rated animes, and the distribution of MyAnimeList.net users around the world.

Datasets

Anime content is consumed in many different ways and forms, for our visualisation, we found this dataset on kaggle [2] that contains 300 thousand users, 14 thousand anime metadata, and 80 million. ratings from MyAnimeList.net. The dataset consists of three CSV files, anime*cleaned.csv, animelists*cleaned.csv, and *users_cleaned.csv*, that contains:

- animecleaned.csv contains the list of anime, with title, title synonyms, genre, studio, licensor, producer, duration, rating, score, airing date, episodes, source (manga, light novel, etc.), and many other important data about individual anime providing sufficient information about trends in time about important aspects of anime.
- *animelists*cleaned.csv contains information about users who watch anime, namely username, registration date (join_date), last online date, birth date, gender, location, and lots of aggregated values from their anime lists.
- *users_cleaned.csv* contains anime lists of all users. Per each record, here is username, anime ID, score, status etc.

Due to the amount of information that animelistscleaned.csv contains, the size of the file is a little over 2.2GB. Meanwhile, the other two files still contain different important information that we will be using for our Visualisations. With this, we can say our dataset contains volume and variety.

Data Exploration, Processing, Cleaning and/or Integration

What did you need to do to prepare the dataset(s) to create your graph/chart?

After collecting the data and exploring it a bit, we first decided on what data and columns would be relevant to us to tell the story of anime.

The three files, *anime_cleaned.csv, animelists_cleaned.csv,* and *users_cleaned.csv*, had 108711 rows & 17 columns, 31284030 rows & 11 columns, and 6668 rows & 33 columns respectively. We decided to drop the columns that we don't need from all three tables and then combine them together for our visualisation purpose.

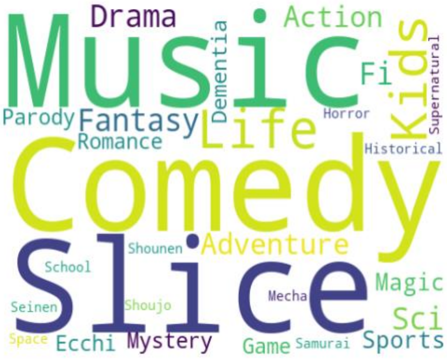
- Dropped unnecessary columns from all three tables.
- Removed all null values from all three tables.
- *animelists_cleaned* contains columns like username, anime_id which are also present in *users_cleaned* and anime*_cleaned respectively.*
- *So, we used the merge function in pandas and merged the three tables into one.*

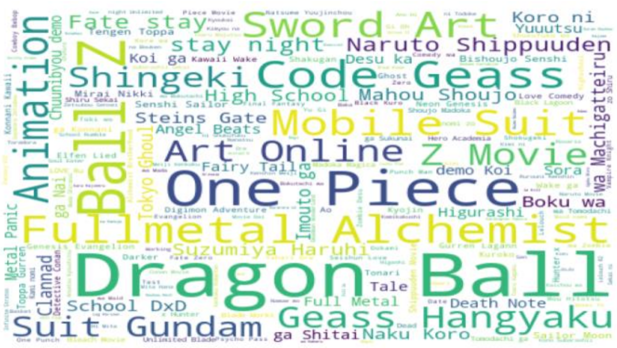
Data Exploration

The visualisations in this section are not the final visualisations but just the ones that we created to help us understand our data a bit better.

Anime Genre & Title word cloud

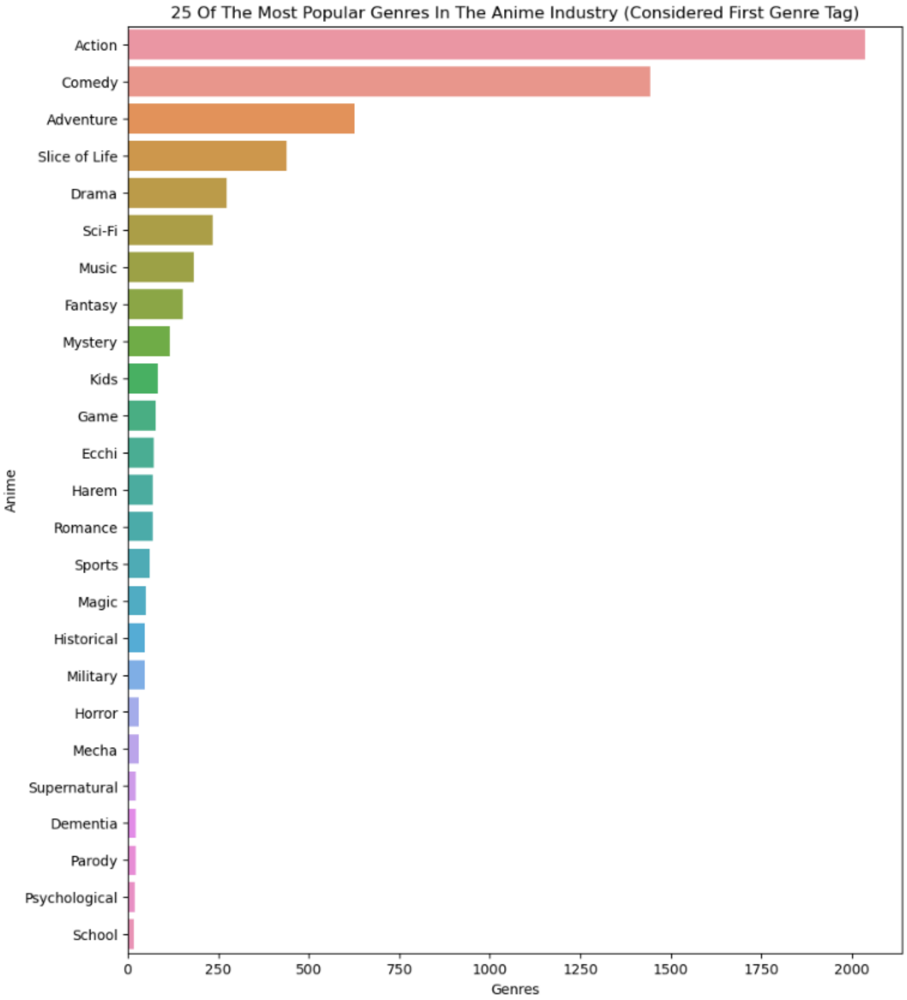
There are more than 6000 unique anime in our dataset and after cleaning we had the data for almost 20 million users who watch anime on myanimelist.net. We decided to use world cloud because it is quick and informative. The word cloud gave us a sense of what are the most watched anime on myanimelist and this gave us an idea to plot the bubble chart for the most rated anime. We have also plotted a bubble chart for the different genres. From this we can see Music, comedy, etc is most popular but here we have considered all different multi labels of genres and not just the main genre.





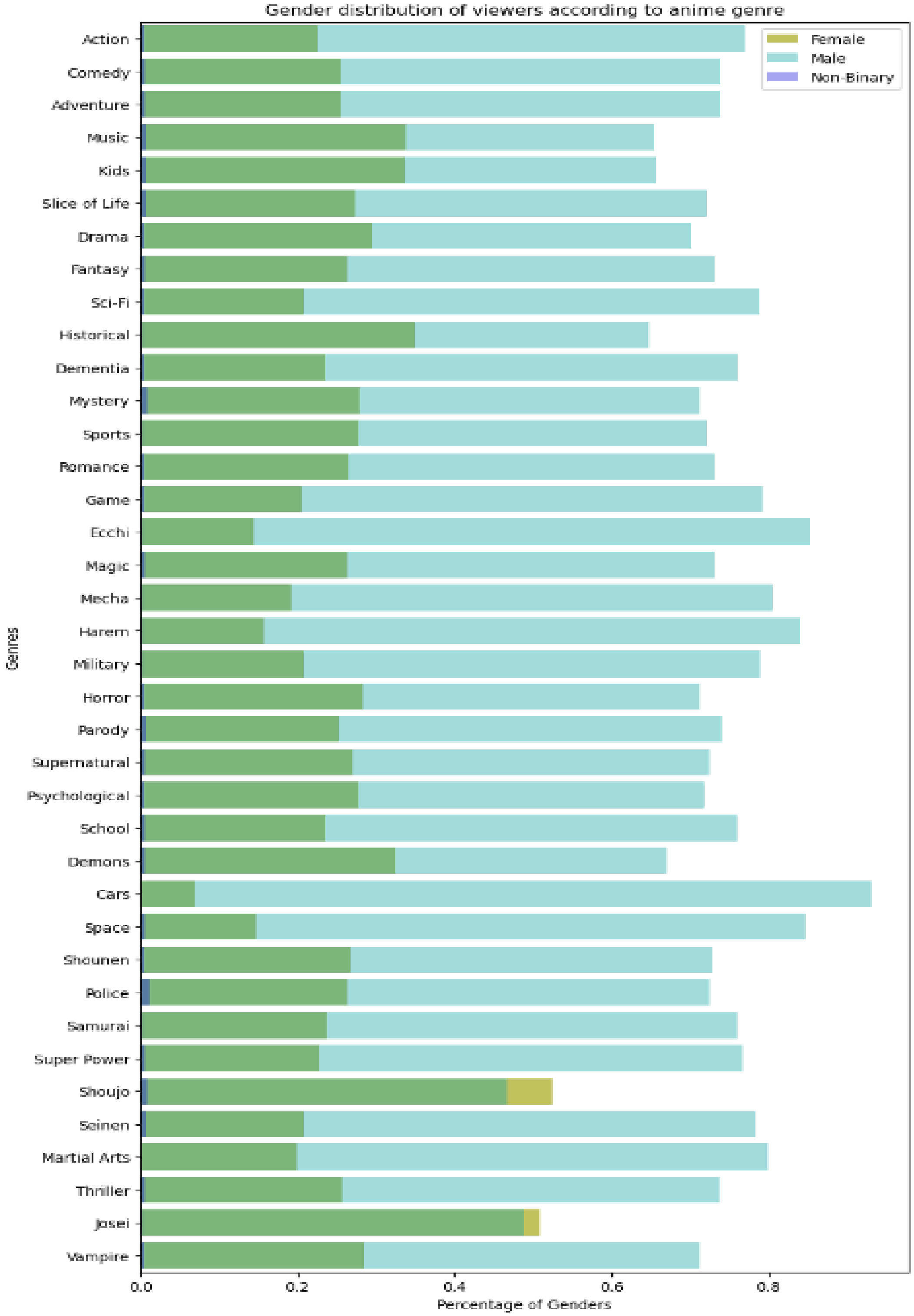
Top 25 anime genres:

Anime has several different genres as evident from the word cloud and we wanted to find out which are the most popular ones. Like any normal movie or book or story, each anime has multiple genre tags, therefore we are only considering the first genre associated with each anime as its main genre. We have identified all the different genres in our dataset and then simply counted the occurrences of these genres in the first label for each anime (or row) and since we have around 38 different genres, we've decided to just visualize the top 25 anime genres using a sorted bar graph.



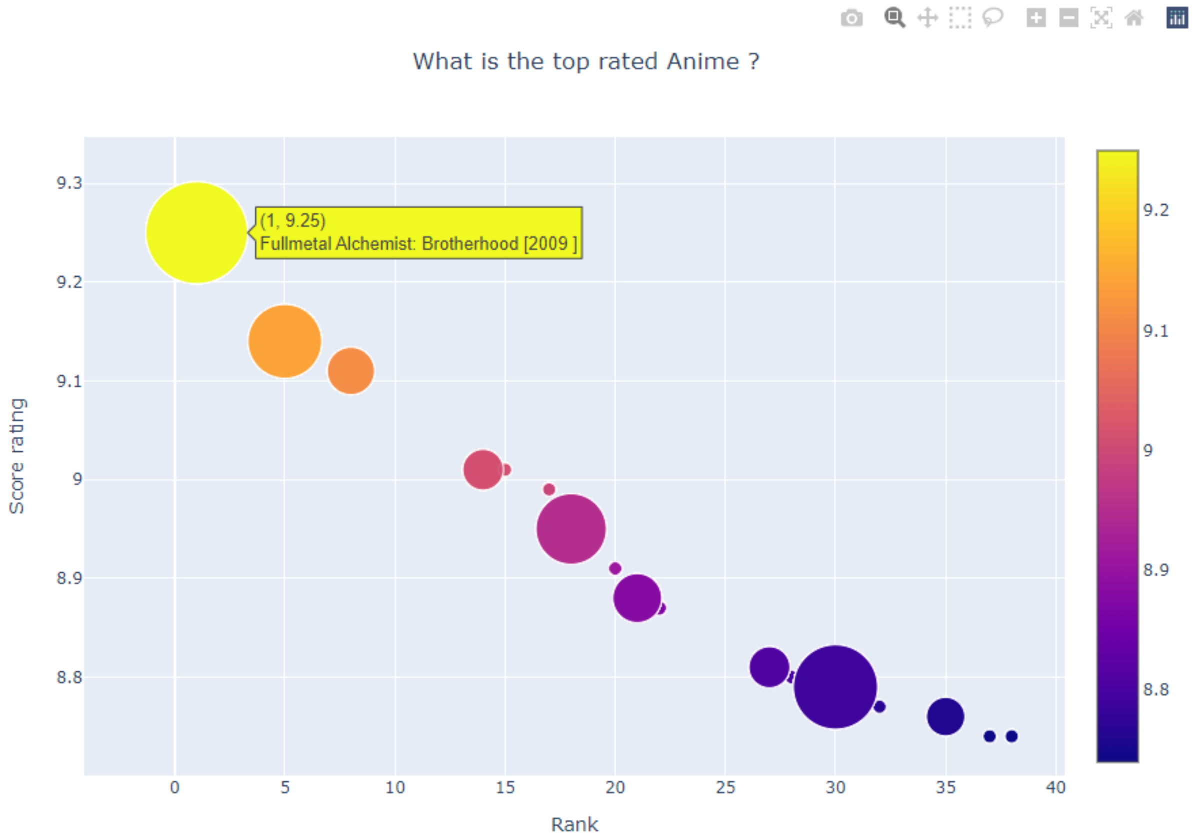
Visualisation

For the **first** visualisation, we wanted to compare the spread of our anime viewers based on their gender in different genres.



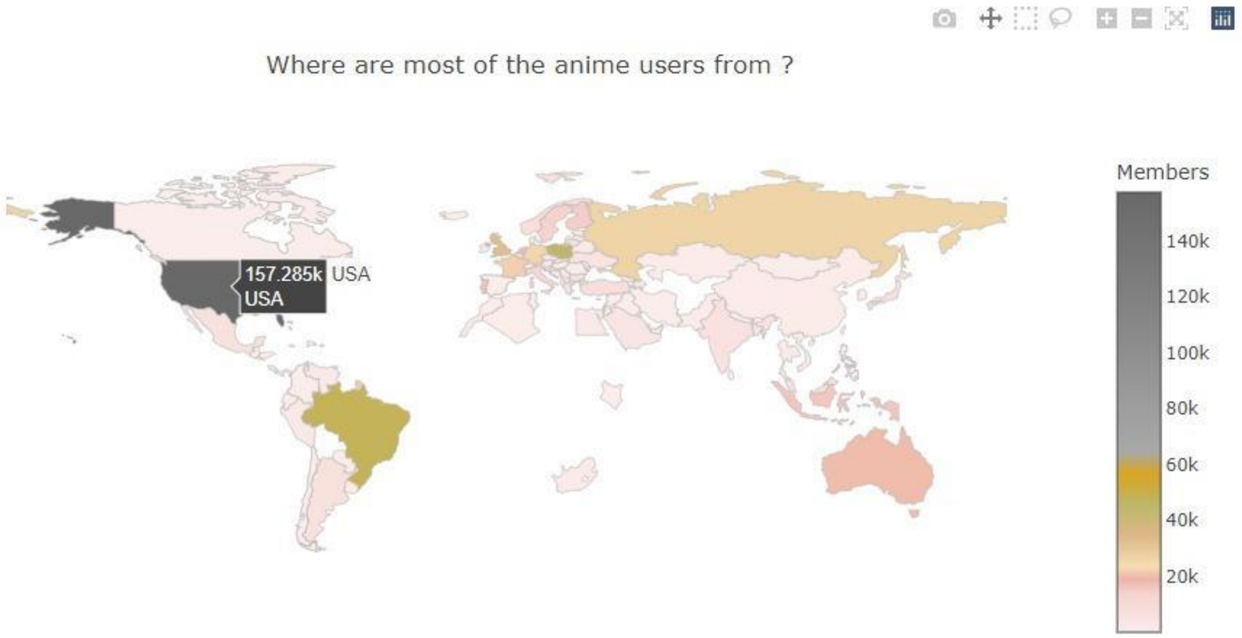
For this purpose, we went with a stacked bar chart. Stacked charts are an incredibly effective tool to compare total values across categories and it was perfect to compare the gender distribution in each of our genres. We can see from the chart that male viewers are top in most of the genres. Female viewers lead in two genres 'Shoujo' and 'Josei'. Non-Binary make up a small percentage in most of the genres but this is mainly because there were very few users who identified as Non-Binary in the first place. For color selection we have referred to your document on loop titled 'How to use color' and Stephen Few's Practical Rules for Using Color in Charts [4][5]. For this stacked bar chart, we went with simple natural colors, but we also didn't just stick to one color with different shades because this is a stacked chart and we need to be able to easily identify the different compositions of genders in each genre.

For the **second** visualisation, we're following up with our Anime title wordcloud. We are going to find out the most rated anime that have been rated by more than 100k users. Since our merged dataset was ordered and since it had a large number of rows that were causing memory issues when we were trying to process it in jupyter notebook, we have decided to take 1 million random rows by using the sample function. Then we take the ratings of those 1 million users and we are going to normalize the ratings using the formula:



For the **final** visualisation, we wanted to show how the users are distributed across the world. For this purpose, we took use of the location column in our dataset. Now, we couldn't just directly use it because from the looks of it, the location column was a user input field and thus there were loads of multiple representations and even fake entries like ‘The Moon’, ‘Behind you’, ‘World’. We obviously didn’t want to plot those values, so we first took all the locations of the users in one list and we sorted out the locations by taking a list of all the countries in the world and their popular cities and updated our list. This also covers cases like let's say the user has his location as Texas or TX, we’ve set that value to USA.

We followed this tutorial to plot our map: [Geographical plotting of maps with Plotly]



Tools and libraries used

We started off by exploring the data briefly on Microsoft Excel and also some initial cleaning with it. Both of us have backgrounds as software developers and thus we found it more comfortable to use python with Jupyter notebook for the rest of the cleaning, processing, and visualisation. We have used libraries like seaborn, matplotlib, and plotly for our main visualisations.

Conclusion

Westernization is a topic that is often spoken about in several parts of the world. Hollywood, pop culture, TV shows, and several other contents have heavily influenced the modern societies of different parts of the world. From our visualisations, we’re able to see the impact of an Eastern country’s cultural influence on the world with Anime. Anime content is only growing in popularity and the impact of the culture is growing all around the world. The final visualisation clearly shows that the USA has the largest number of anime viewers on myanimelist.net across the world. Being anime fans ourselves, we enjoyed going through the large dataset and cleaning it. We worked on the data gathering, cleaning, and processing together and we split our work for the various graphs. But since this was interesting and a bit new to us, we both ended up working together equally on all the different visualisations. Some of the problems we faced while working on this visualisation project were mainly with cleaning the data. The data was quite large since it contained almost 80 million ratings, we could barely even load it on our computers. There were also many missing data and several misrepresented data which we had to clean. Since it wasn't loading properly on excel we were kind of forced to use python to clean some of it but the various tutorials we followed helped us and now we're glad we cleaned it using python and not manually using excel. For the first stacked bar chart, it would have been better if we had some more data on non-binary users. For the second visualisation, we initially tried it with all the data in our merged dataset but we often ran into memory issues, so we took a sample of 1 million rows. We are sure python can handle much more data but we need to investigate a bit more on how to do it efficiently. For the remaining visualisations, we feel we have properly used the relevant data by processing the information from our dataset. We have tried to make sure that we follow the basic guidelines when it comes to color selection. While doing this assignment, we understood how many different options and ways are there for visualizing, and at one point we were even confused about what colors to use. We were only able to decide after going through the notes on the loop. We have also tried to use a colorblind-friendly palette [6] for our visualisations.

References

[1] lements, Jonathan. *Anime: A history*. Bloomsbury Publishing, 2017.

[2] Matěj Račinský, "MyAnimeList Dataset." Kaggle, 2018, doi: 10.34740/KAGGLE/DSV/45582.

[3] Srivastav,Ashish,"Normalization Formula.", WallStreetMojo,https://www.wallstreetmojo.com/normalization-formula/

[4] Little,Suzanne,How to use colour

[5] Few,Stephen,2008, Practical Rules for Using Color in Charts

[6] Nichols, David,"Coloring for Colorblindness",David Math Logic