CSE46801 - Information Visualization
Assignment 2: Anime Data Analysis

Student name: Nguyen Minh Duc
Student ID: 20202026

1 Teaser of the application

Please refer to Figure 1 for an overview of my system. Or alternatively, click this link to have first-hand experience.

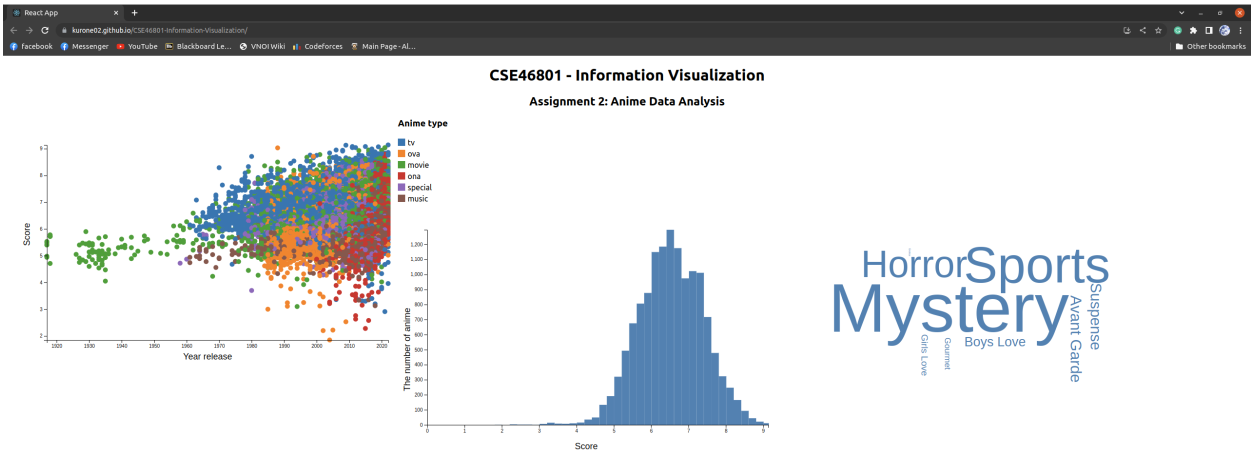

Figure 1: Teaser of the application

2 Introduction to data

The anime industry has been growing rapidly in recent years. More and more people have started watching this particular style of animation from Japan. Understanding the preferences of anime fans would benefit studios in production and marketing. Moreover, as the industry is growing, more and more anime series and movies are produced creating a large database, which could be utilized to understand the most popular genres, themes, and characters, helping the studios to create content that is likely to resonate with their target audience. Lastly, analyzing anime data can also aid in the study of Japanese culture and its influence on popular media worldwide. By analyzing trends and patterns in anime consumption and production, one could gain a better understanding of the cultural and social factors that shape media consumption and its impact on society.

In this assignment, I will use “MyAnimeList Anime and Manga Datasets” from HERNÀNDEZ [1] containing 24043 entries of different animes. The data is crawled from “MyAnimeList” [2], a well-known database that stores new anime series, movies, and manga (Japanese-style comics) every season. The data was last updated on 25th July 2022.

3 Data Preprocessing

The data set contains 24043 rows and 39 columns of different attributes for each anime, namely, ID, Title, Type, Scores, Release year, Genres, etc. The data set contains a lot of `NaN` values and most of them are unstructured. After consideration, I decided to keep some important attributes that will be useful for analysis: Title, Type, Scores, Start year, and Genres. All of the `NaN` values will be converted to 0 in the `scores` attribute, and remove every entry that has a `NaN` type, start year, and genres, reducing the number of anime in the data set down to approximately 13000 entries.

For this section, I read the data with `d3.csv`, and filter out all of the invalid data points.

```
d3.csv("/anime.csv")
    .then(data => {
      const filtered_data = data.filter(item => item.type && item.start_year &&
          item.score && item.genres);
      for(let i = 0; i < filtered_data.length; i++) {
        filtered_data[i].start_year = parseInt(filtered_data[i].start_year);
        filtered_data[i].score = parseFloat(filtered_data[i].score);
        filtered_data[i].genres = JSON.parse(filtered_data[i].genres.replace(/’/g,
            ’"’));
      }
      setFullData(filtered_data);
      setData(filtered_data);
      setDataScatterDistr(filtered_data);
    });
```

Listing 1: Filtering data

4 Application features

In this application, I propose three visualization features that can help users analyze the patterns in the anime audience’s perception of different types and genres of different animes. The features are as follows

- A scatter plot showing the relationship between the release year and the score of every anime in the data set.
- A bar chart showing the distribution of the chosen animes.
- A word cloud visualization in which the genre with the highest average score will be the largest.

4.1 Scatter plot

In this visualization view, I extracted two attributes from every data point, namely release years and scores to see if recently made anime will have better ratings than those made in the past. Please refer to Figure 2 for a preview. Users can choose different types of anime to analyze their performance over the years by clicking on the legends on the right of the scatter plot. For example, if the user clicks on the movie type, the visualization will be updated similarly to Figure 3.

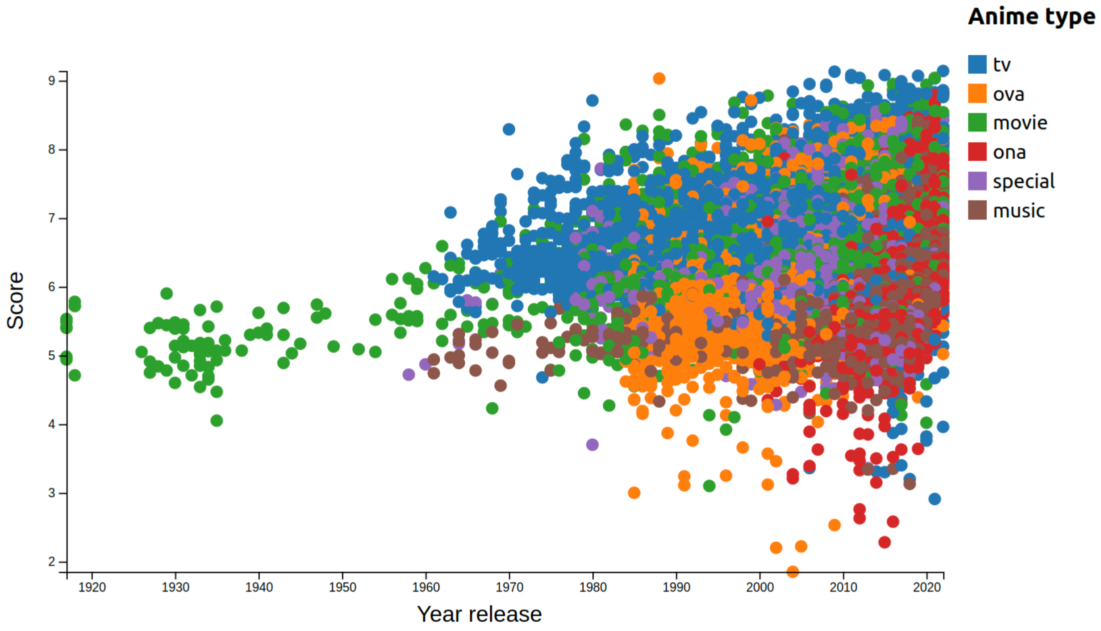

Figure 2: First view: scatter plot

4.2 Bar chart

In the second visualization view, I aggregate all of the chosen data points into 50 bins of scores ranging from 0 to 10. Users can hover on each bar to see more details including the precise number of anime with the respective score.

4.3 Word cloud

In this final visualization view, I used the word cloud idiom in visualization to display the most successful genre in the chosen anime, i.e., the genre with the highest average score. The preview is in Figure 5.

4.4 Interactions

In my visualization system, there is a bidirectional link between the first and the second view, i.e., the scatter plot and the bar chart. The other link is an unidirectional link from the scatter plot to the word cloud. Users can “brush” a region in scatter to choose some of the data points and both the bar chart and the word cloud will be updated accordingly. Additionally, when the user clicks on a bar in the score distribution, the system will filter out all of the data points not having a score lying within the chosen interval. Please refer to Figure 6 for a preview.

5 Design rationale

Since the data set I chose has tens of thousands of data points, a scatter plot is an easy choice thanks to its scalability. Also, to distinguish between two different anime types, I use the qualitative color scheme so that users can have a better time recognizing the patterns. As for the bar chart, we also need to summarize the data into a distribution, hence, a bar chart is a perfect solution. Lastly, the choice of the word cloud is hard to justify to obtain the best visualization for the best performing genre, but since the bar chart has already been utilized for another purpose, the word cloud is the next thing that can effectively visualize words.

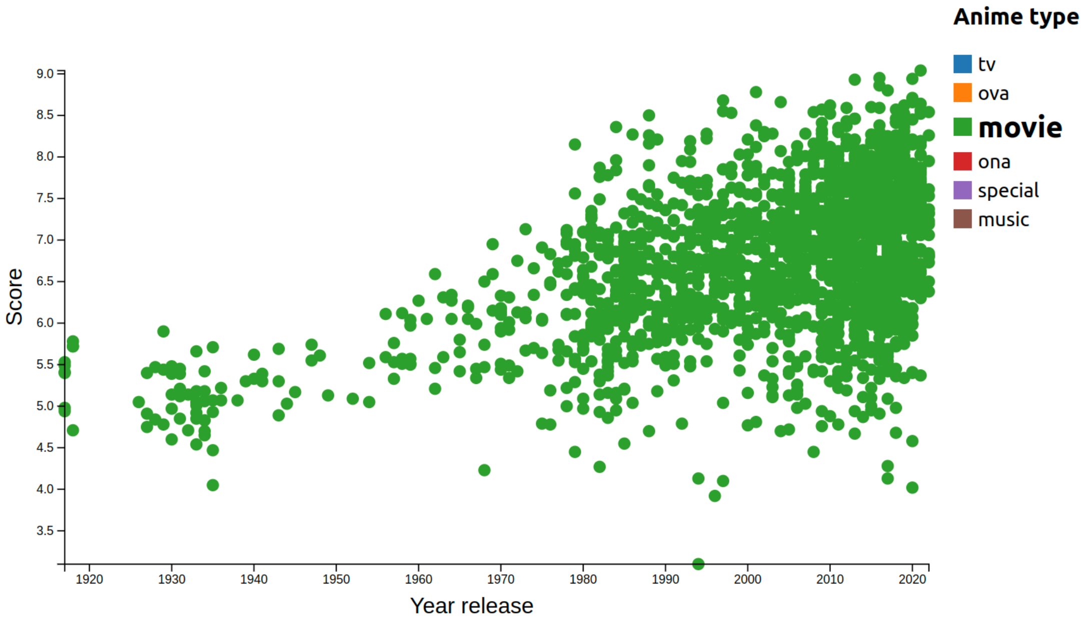

Figure 3: Example when user clicks on movie

6 Usage scenarios

Users can use my system to analyze the anime industry trends, hence, have better decision-making in doing something in the market.

7 Observation

As one can observe from my system, most of the time, the recently made animes, on average, slightly increase in terms of the rating score compared to the past ones. Also, the score distribution forms a normal distribution, having quite a high average.

8 Running explanation

I already deployed the application at https://kurone02.github.io/CSE46801-Information-Visualization/. If one wants to build the system from scratch, follow the following steps

- Install `node` and `npm`
- Obtain the source code in the zip file, unzip it
- Go into the code's directory
- Run `npm install` to install all dependencies
- Run `npm start` to start a development server

I would prefer users to test my system with the provided link.

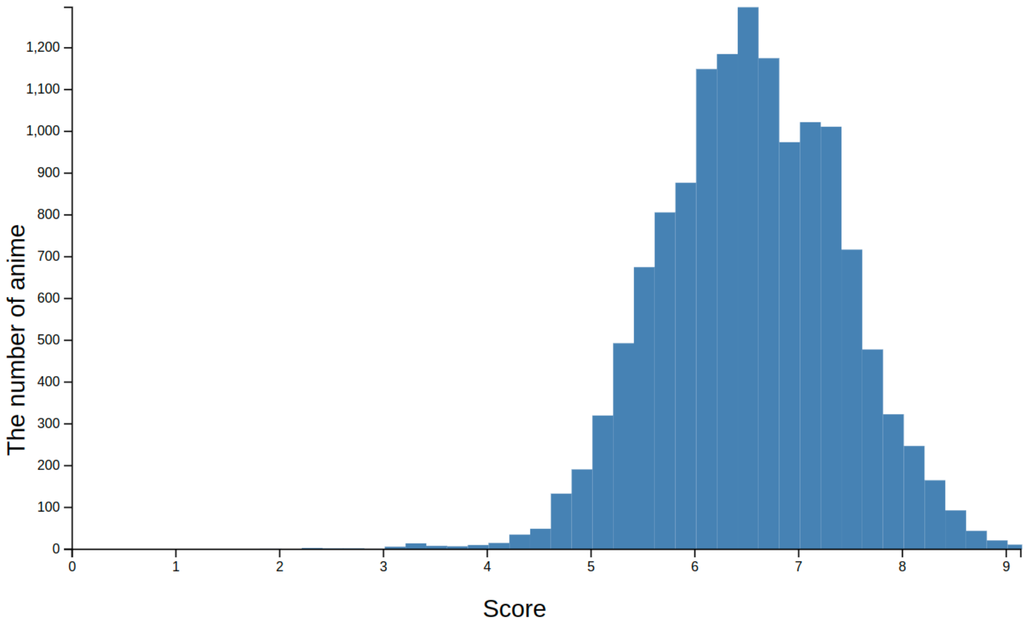

Figure 4: Second view: bar chart

References

[1] https://www.kaggle.com/datasets/andreuvallhernndez/myanimelist?select=anime.csv

[2] https://myanimelist.net/

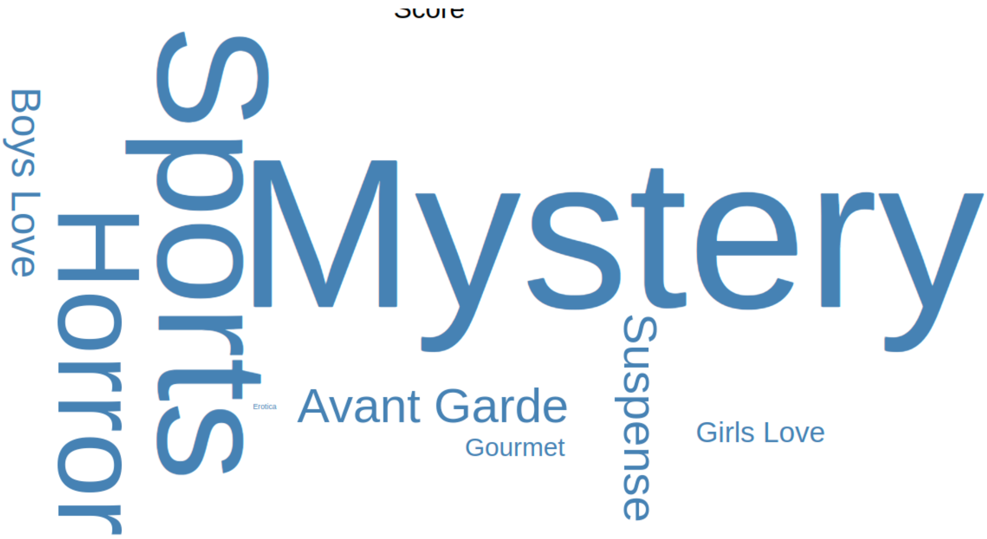

Figure 5: Third view: Word Cloud

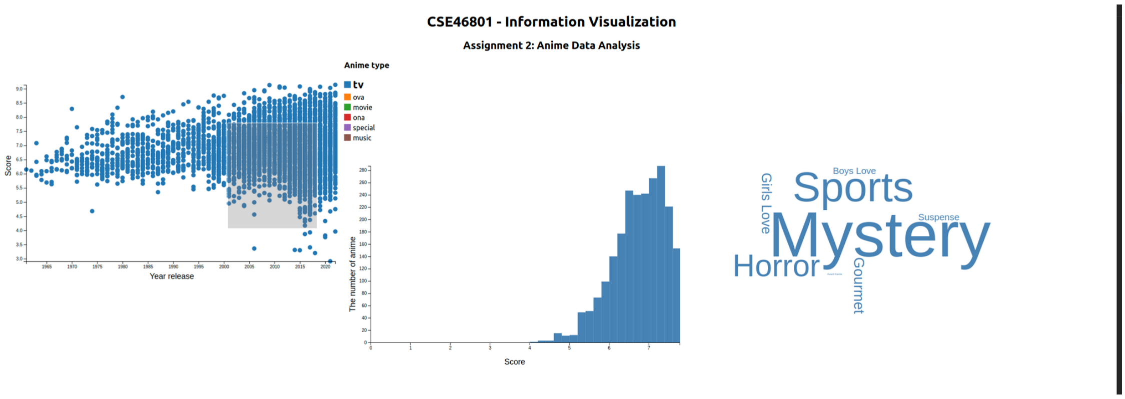

Figure 6: View changes