1 Introduction

Exploratory Visualization has been used to analyse and represent the Anime Recommendation Dataset where new features were derived from existing features to gain new knowledge about the data.

2 About the Dataset

The dataset used for this visualisation is an **Anime Recommendations Database** [1]. This data set contains information on user preference data from 73,516 users on 12,294 anime, but only a part of the dataset has been used to visualise. The raw dataset was preprocessed to generated data required for visualisation -information about Anime "TV Shows" and "Movies" have been filtered, aggregated and rendered using the tool ***Processing.py***.

The visualisation fits into the category of **Exploratory Data Visualization** (Analyst Centric), where new data is derived from the existing once's. The dataset consists of information like, Anime Name, Genre (an anime can belong to more than one genre), type (Movie, TV show, OVA, etc.) rating, members and episodes. From this, information about **each genre** ('per-genre' data) is derived which are as follows.

- Average Rating
- Minimum and Maximum Rating.
- Anime count.
- Percentage of Anime in each Genre.
- Minimum and Maximum number of Episodes

This information has been calculated on data load. These stats are stored using a dictionary (map in python), in the case of per genre average rating: key in the genre and the value is the list of all the ratings of anime that belong to this genre and the average is calculated, which is the average rating for that genre. Similarly other derived has been calculated.

3 Exploratory Data Visualization

Exploratory Data Visualization (**EDV**) is an approach to analyzing data sets to summarize their main characteristics using visual methods. Here EDV is used to find new knowledge about how anime genre compare with each other in terms of minimum, maximum and average values for rating, episodes and anime count in each genre. To represent this derived information various methods have been used, like:

1. Pie Charts
2. Bar Graphs
3. Scatter Plots
4. Box Plots
5. Network Graphs

Apart from representing information in this format for both Anime **TV Shows** and **Movies**, a option to compare this derived information has also been provided via means of a bar chart and 2 types of pie charts (radial and area pie charts).

4 Visualisation Analysis

4.1 Data Types:

The data type for this visualization is a Table, where each Item represents an individual entity in the dataset and the attributes were as follows: Anime Name, Genre, Type, Rating, Members and Episodes.

4.2 Task(s):

The visualization of the dataset is abstracted as an action-target pair to increase the effectiveness or expressiveness of the data for the visualization.

4.2.1 Actions:

Here, the data is consumed to discover/explore the data to find new knowledge and to produce / derive new data elements to present them in user friendly format.

4.2.2 Targets:

The target of this visualization was to identify trends and extract features based on single or multiple attributes of the dataset, by understanding the correlation between genre, rating, episodes and anime count.

4.3 Visual Encoding Channels:

Various visual encoding channels were used to represent the derived data, like:

1. **Position:** In the case of scatter plot, position of the circles along the y-axis was used to represent anime count in that genre, where in case of box plot position of the box along the y-axis was used to represent average rating.
2. **Mark:** The Circles, Boxes and Bars were used to represent different types of data in the across scatter plot, box plots and bar graphs respectively.
3. **Size:** In case of pie chart: the area (size) of the arc was used to represent the percentage of anime in each genre, In Bar Graph: the height of the bar was used to depict the average rating of the genre, where in case of scatter plot, the radius of the circle was also used to represent the average rating per-genre.
4. **Colour:** Here, a unique colour was used to represent a particular genre throughout all the graphs.

5 Technical Details:

Processing.py was used to visualize the data. No additional libraries or tools were used apart from the what is available by default via processing. The code for this project can be found on my **GitHub repo**[1] under the folder data-visualization/otaku_hub

All graphs/plots created for this visualisation were written in a generic way so that it can be reused to fit any dataset of any size. Given the x and y coordinates, 2 scales (one for generic scaling and another for offset in case of outliers) and the data, as the bare minimum parameters, the visualisation functions were designed to plot the graphs dynamically. This feature has been showcased under the **Compare Genre** functionality of the visualisation, where in a user can select what genres he/she wants to compare.

6 Results:

The novelty of this visualisation lies in derived data and the new knowledge obtained through the visualisation of the said data. With this visualisation we learn that in TV Shows, the "Physiological" anime genre has the highest rating with 7.53 whereas "Music" has the lowest with a rating of 6.22. In terms of anime count, the "Comedy" genre has the highest with 1870 and "Samurai" genre with the lowest with 51. Other information like episode count per genre, and similar information for Movie type of anime can also be obtained, which otherwise would have been impossible to directly infer from the raw dataset.

The **strengths** of this visualisation lies in how it is able to obtain, visualise and compare the new knowledge obtained, which in-turn leads to information being obtained in terms of comparison data between multiple genres. It's **weakness** lies in the amount of data that has to be processed and visualised as the number of anime increases (the raw data) the data that it has to preprocess to obtain data that is to be visualised increases which results in performance issues, and as more genre become significant in the future the shear amount of data to be visualised will increased which leads to a sense of clutterness, this could be avoided by improving the visualisation techniques or the tool used.

References

[1] Kaggle.com. 2020. Anime Recommendations Database. [online] Available at: <https://www.kaggle.com/CooperUnion/anime-recommendations-database> [Accessed 17 April 2020].

[1] https://github.com/mukeshmk/data-visualization