Simon Fraser University

CMPT 353, Computational Data Science

Safonov Oleh, 301575136

Summer 2024

Demographic Insights and Predictive Modeling of Anime Ratings

This project aims to explore a Kaggle dataset that contains MyAnimeList.com data in order to provide a showcase of website's user demographics, anime industry trends and to build a practical model that could predict useful information for anime studios and their publishers.

The dataset contains three tables and additional dataset:

1. AnimeList.csv (14,478 total entries)
2. UserList.csv (302,675 total entries)
3. UserAnimeList.csv (80,076,112 total entries)
4. World cities database (44,383 total entries)

AnimeList.csv contains information about different anime shows, like its identifier, title and its variations, source from which its adapted, genres, studio etc. UserList.csv contains information provided by users like their username, id, amount of shows watched, amount of episodes watched, gender, location, birth date etc. UserAnimeList.csv used to store users and shows they rated as well as the amount of episodes from every show users watched.

Since our project aims to explore user demographics, a certain amount of preprocessing needs to be done before the dataset is ready to use. First of all, we start with the AnimeList table where we keep only TV series and exclude special episodes, full-length movies and musical clips because we are interested in multiple-episode shows. Then, we process genres, adaptation source and studios into one-hot encoded columns for later use. Secondly, we process the UserList where we filter out users missing important information, calculate user age at the moment of dataset creation, and use the World cities database to map and extract user country. And finally, in UserAnimeList we filter out entries containing removed users and shows to avoid errors and inconsistencies.

After the initial data collection, it is time to discuss the project structure. The project consists of four Jupyter Notebook files each responsible for a certain part of the project:

1. CleanUp.ipynb – must be run before other notebooks to get the data into correct format
2. DataExploration.ipynb – focuses on plotting demographic information and industry trends
3. HypothesisTesting.ipynb – used to test out multiple hypotheses about population and show rating
4. RatingRegression.ipynb – engineering features and using previously found information to build a model for anime rating prediction

Jupyter notebooks were chosen for convenience of not having to run the code in order to see the results.

Now, we are going to explain the methods of data analysis used in these notebooks. The first analysis of the Data Exploration notebook are three histograms of user age distribution [Figure 1]. The second analysis consists of two choropleth maps that show the absolute and per capita amount of website users per country [Figure 2]. The maps show only sovereign states and their territories, with country borders being factual instead of legal. The third analysis of the Data Exploration is another three histograms: mean user ratings distribution, and number of watched episodes distribution [Figure 3]. The last analysis of this notebook is two bar plots showing top-15 most popular anime genres and top-15 highest rated genres [Figure 4].

The first test of Hypothesis Testing notebook aims to see whether its true that female users watch more Romance shows than male users. After creating the distributions, the data was heavily right-skewed and required a square root transformation for a somewhat-normal distribution [Figure 5]. We used T-test and Mann-Whitney U-test to see if these distributions had different means. The second test of this notebook addresses the question from the first notebook: are high amounts of empty accounts and accounts with zero mean score related? Once the data was extracted and placed in the contingency table [Figure 6], we used Chi-squared test to determine if there was a relationship. Third test of this notebook aims to test if there is a trend of anime ratings increasing for newer shows [Figure 7]. The test calculated a correlation coefficient between the release year and average rating by that year. The last test of the notebook examined if there was a difference in mean score based on the adaptation source material. We used pairwise Tukey test p-values and a Tukey plot [Figure 8] to determine if that was the case.

Based on the information from the previous notebooks we have established that anime genre, adaptation source and the amount of episodes in the anime show a strong relationship with its score. Additionally, we have added the release season and studio columns because we are sure that they are also impactful on the result. Given all this information, we are going to try creating a machine learning model in RatingRegression notebook that will predict potential score the anime might get based on this information. After encoding the additional features, we split the data into training and testing sets and trained the model. We chose XGboost algorithm as it is known to be robust for data science tasks and because it is a boosted ensemble algorithm with decision trees, which perform really well with categorical data like source, studio or genre. The model is then tested, cross-validated and saved for later use. The notebook also provides an example of using sample data.

After performing all tests and examinations, it is time to show notable results of these tests. The age extraction plots show that user ages are normally distributed with a mean near 25 years, age distributions by gender show similar results with female distribution containing less samples (the female population is significantly smaller). Choropleth maps show a few peculiar findings: high-population Western countries seem to be where most users come from. This makes total sense since the website supports only English, French, German and Spanish languages. Another finding in this set is that Poland has more users than any other country by a wide margin. Per capita map shows a similar trend, but with Scandinavian countries in the lead, followed by Canada, Australia and New Zealand. This also makes total sense since former British colonies and Scandinavian countries have small populations and have high percentage of English-speakers. Number of watched episodes does show only one interesting thing, the mode of episodes watched is actually zero. User ratings distribution also shows that there are abnormal peaks around 0 and 10 which are probably users trying their best to skew ratings for shows they like/dislike. The genre bar chart gave obvious results, most common genres are the most popular ones.

Hypothesis testing yielded some results too. The romance anime preference test gave a very small p-value which proves that the means are different, the negative t-statistic implies that female users prefer romance anime more, although the magnitude of this statistic is small, so the difference is minimal. The empty accounts hypothesis test gave a small p-value and thus we conclude that there is indeed a connection between accounts being empty and used for setting multiple zero ratings, this was probably done to lower some anime ratings. The correlation coefficient between release year and anime rating shows a moderate positive correlation that implies that ratings slowly grow as we move to newer shows. Tukey's pairwise test and plot have concluded that adaptation source does not affect ratings most of the time, but for some special sources the difference is notable.

The XGboost model that we trained has a 72% accuracy and 74% validation score which allows for some approximate predictions, but is not enough to be confident in future shows' success or failure.

Now we are going to draw conclusions based on the results. First of all, MyAnimeList developers need to implement a much more strict registration as almost half of users were missing crucial information or were

straight up bot accounts. Secondly, MyAnimeList would significantly increase their user count if they also adapted their website to Asian and African languages. The website also could benefit from a new rating system that would prohibit users from rating shows they haven't watched a single episode in. Secondly, for anime studio management it would be beneficial to try producing shows that could culturally appeal to users from Eastern Europe/Scandinavia. The studios should not try determining their target audience by gender, as it was shown to have little effect on anything examined. We can also suggest anime studios to adapt more Light novels because their adaptations seem to be very liked by the audience. (Interesting enough, there is a big trend for Light Novel adaptation in 2024). We would not advise using the produced model to predict ratings without additional data or training.

As always, the project couldn't go flawlessly without any complications or shortcomings. Processing user country was a complicated task that took too much time and we had to compromise. Distributions for T-test were not normal enough so we had to double-check with a non-parametrized test. Although performance isn't that bad, we wish we could restart the project with pyspark to speed up dataframe operations. Another shortcoming comes from our model training, and that is the need for additional data to make the accuracy high enough. The dataset lost the majority of entries after filtering, user counts used were around 22% of the original user counts, similar thing happened to other tables. We also wish to have used more unique plots, but the data was fitting histograms too well most of the time. If we had more time, we could have also analyzed full-length animated movies and compared the to multi-episode tv shows.

Visualizations

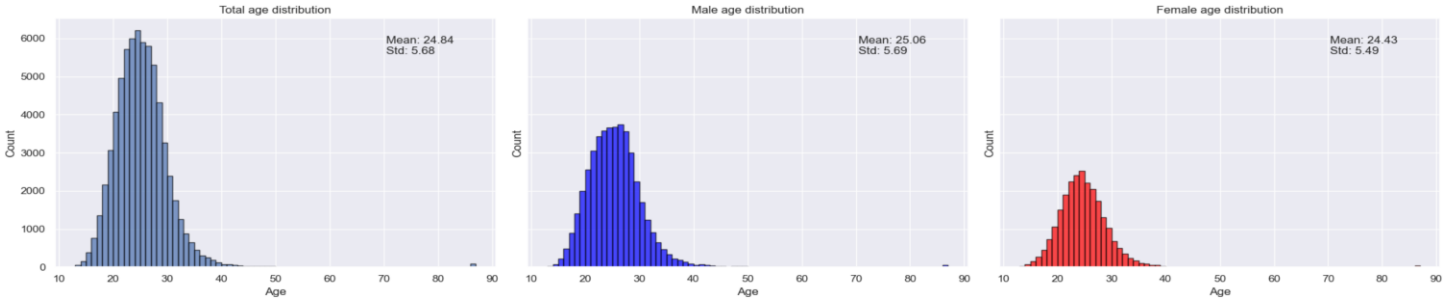

Figure 1 – General age distribution and gender age distributions

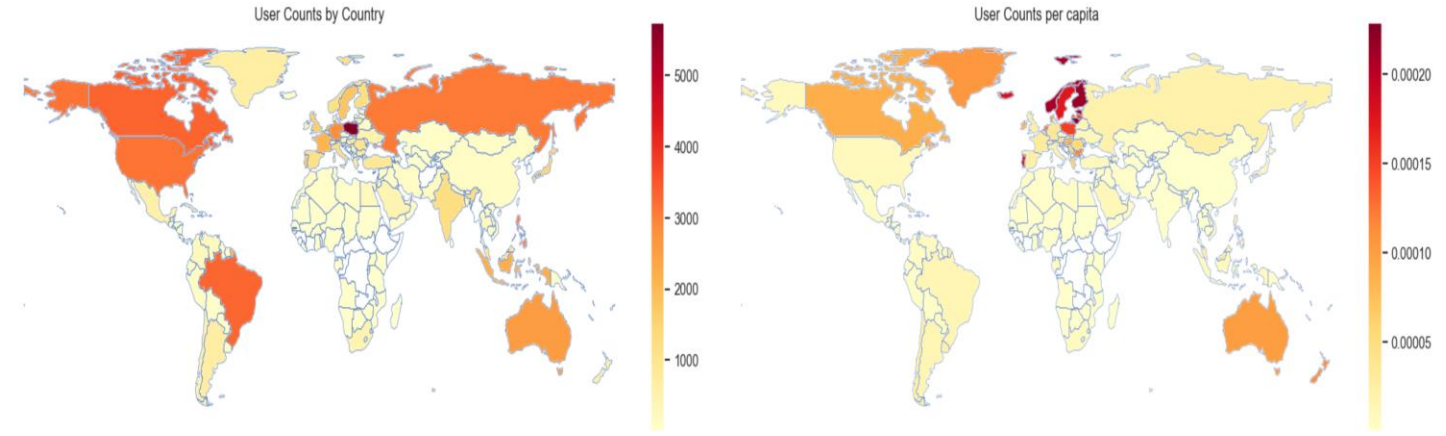

Figure 2 – Choropleth maps of user counts

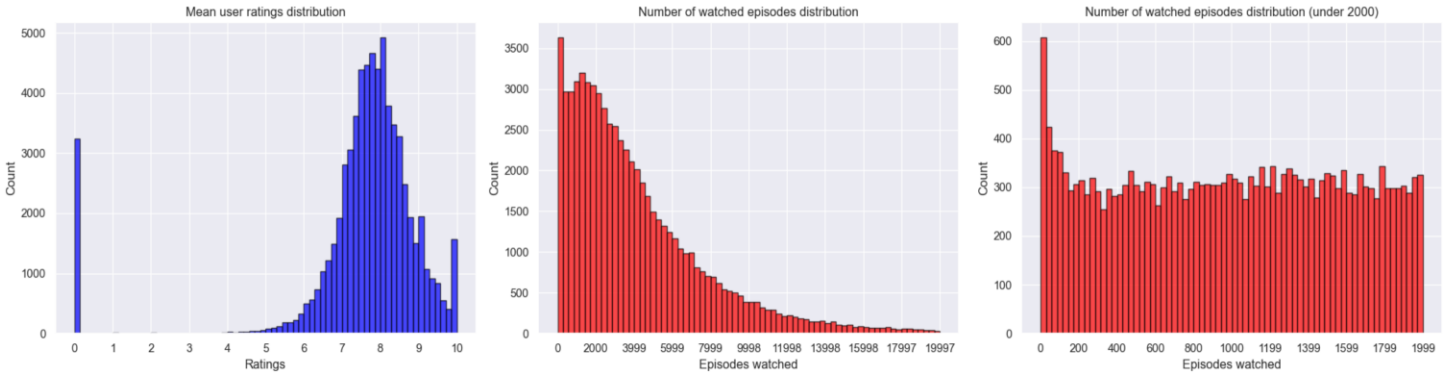

Figure 3 – User ratings and episodes watched distributions

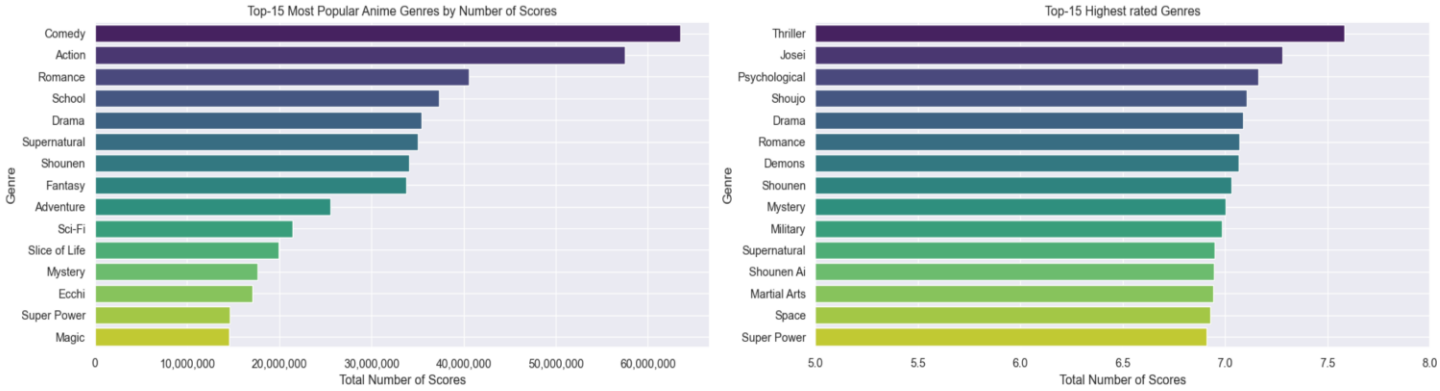

Figure 4 – Most popular and highest-rated genres

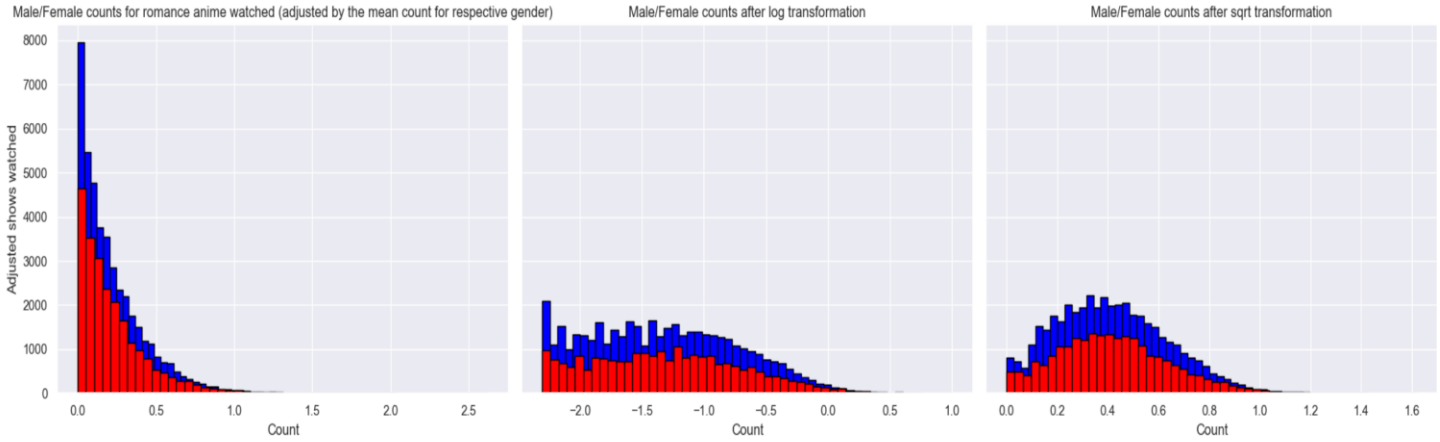

Figure 5 – Romance anime views adjusted by average count

	Zero Episodes	Non-Zero Episodes
Zero Mean Score	1704	1530
Other	65050	65050

Figure 6 – Zero views contingency table

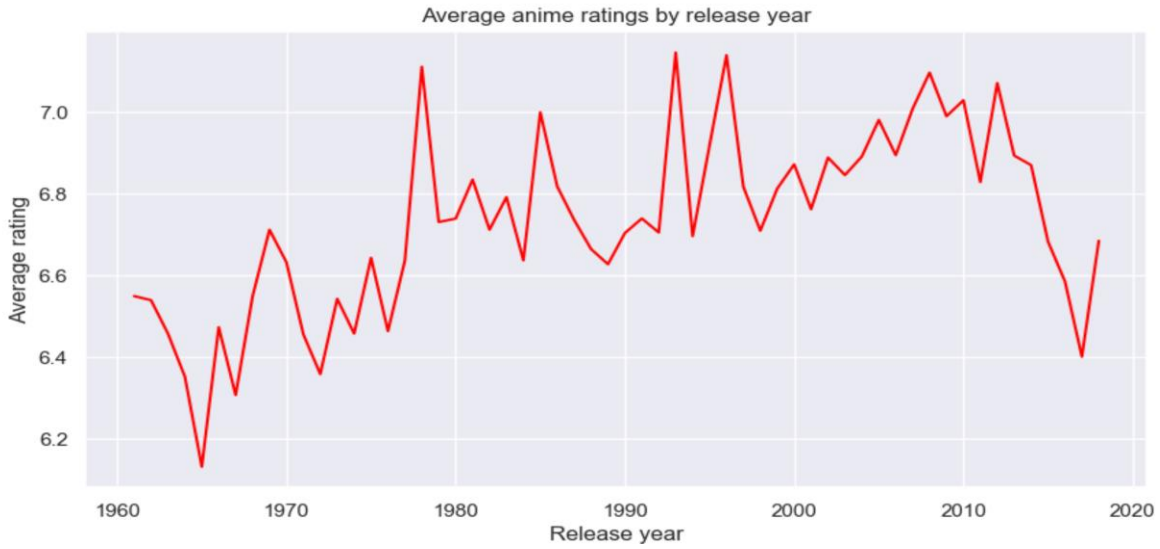

Figure 7 – Average anime ratings by year

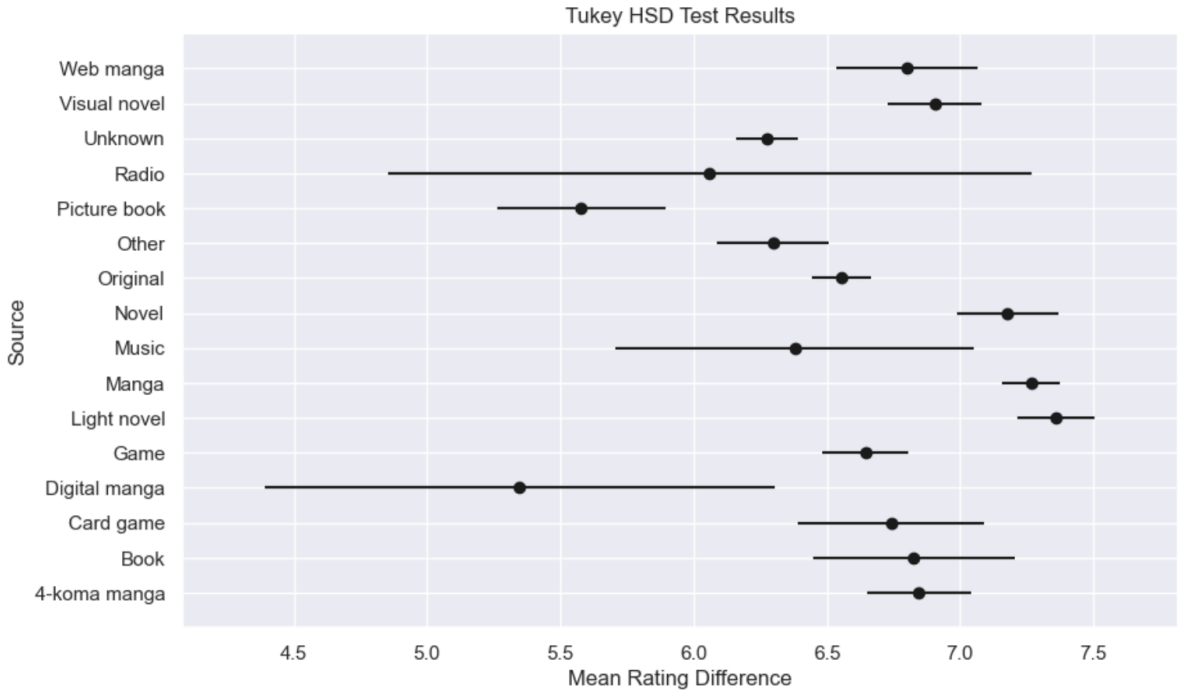

Figure 8 – Tukey's test bar plot

Acknowledgements

1. World Cities database. DOI: 10.34740/kaggle/dsv/7903661
2. MyAnimeList Dataset. DOI: 10.34740/kaggle/dsv/45582
3. World Cities Database. Link: https://www.naturalearthdata.com/

Project Experience Summary

Oleh: In this project I worked on data extraction and pre-processing using Pandas, regular expressions and dictionary mappings by extracting and one-hot encoding seasons and studios as well as filtering out invalid user accounts and mapping their reported location to a possible country and then dropped the filtered users out of the lists to make sure that all of the entries could be used in later analysis. I created distributions and country data by extracting and working with Pandas Dataframes. While researching for a possible solution for the country mapping problem, I tested out 3 different city/country datasets by mapping values and examining derived dataframes in order to find the most suitable one that allowed for correct mapping of cities and their areas to countries. I adjusted the country maps to contain only sovereign states by filtering out anything that did not fit the description to make sure that smaller unrecognized nations don't skew the per capita plot. I plotted multiple graphs using MatPlotLib, Seaborn and geopandas/geoplot to build histograms, choropleth maps, bar-charts, line-plots and contingency tables to allow visual human analysis resulting in many of previously mentioned findings. I practiced forming hypotheses and testing them with Student's T-test, Mann-Whitney U-test, Chi-squared test, Correlation coefficient as well as getting data into semi-normal shape by transforming in order to produce test results that we drew the conclusions from. I analyzed the poor performance of linear regression model, researched possible models and figured out the best set of features by previous analyses in order to build an XGboost model with a much better performance that could provide ratings. I analyzed the results of all the mentioned tests/plots to derive conclusions and suggestions to anime studios and MyAnimeList developers. I created and managed the Git repository, gitignore, readme files and package requirements to ensure that users can run the code if needed.

Xing Yu: In this project I worked on cleaning, analyzing and drawing conclusion from the MyAnimeList Dataset. Using pandas, I cleaned up and one hot encoded a dataset comprising of categorical data for use in a regression model. I used ANOVA and turkeyhsd to analyze the difference in scores based off the initial source material. Also tested the using of linear regression in predicting scores. In I found that linear does not yield good results for predicting users score. However, manga and light novel adaptation yield a higher score than other source material. This could be used by studios to decide what to adapt next for the highest user score