

|  |
| --- |
| **Master of IT in Business**  **ISSS616 Applied Statistical Analysis with R** |
| **Project Proposal: Gotta Watch ‘Em All** |
| **Team Members:**  Geraldine Pang,  Kasi Gopikanna,  Kevin Nathaniel,  Benjamin Tan,  Tan Kah Kyee  **Date:**  28th November 2018 |

**Contents**

[Introduction 2](#_Toc531194977)

[Overall Concept 2](#_Toc531194978)

[Data Sources 3](#_Toc531194979)

[Data Preparation 3](#_Toc531194980)

[Specific Statistical Techniques to be used 4](#_Toc531194981)

[Design of User Interface 5](#_Toc531194982)

[Exploration of Results 6](#_Toc531194983)

[Conclusion 8](#_Toc531194984)

[Appendix 9](#_Toc531194985)

# Introduction

Anime, which is the Japanese term for animation, refers to all forms of animated media produced from or associated with Japan, and differs from western cartoons, focuses more on character development, distinct artwork, engaging storylines with unique themes and genres. Japanese fans who are obsessed with anime or manga (comics) are often termed as “otaku”. However, this fascination with anime is not just limited to Japan. This popular artform has since propagated beyond Japan and has become a worldwide phenomenon with Western adaptations of popular Japanese anime series like “Ghost in the Shell”, “Astro Boy”, “Naruto”, “Bleach and “Death Note”.

The purpose of this data analysis project is to examine the popular Japanese anime subculture, identify the anime user demographics and derive interesting insights into this unique and quirky subculture.

# Overall Concept

The Kaggle dataset was obtained from a contributor who had crawled the data from [MyAnimeList.net](https://myanimelist.net/). The dataset is a representative sample of members of the anime community who watch anime via the website and in turn leaves feedback and comments on these anime episode/series.

User information captured by the website is tagged to unique user ids, allowing us to analyse the information for demographic trends and preferences, among other things. The information includes user-specific details, such as users’ gender, location, anime ratings, and birth dates, and anime-specific information such as anime genre, number of episodes and production studio.

Using the dataset, this project has developed a user interface to serve two main users, namely:

1. Assist a casual anime fan using descriptive analysis to determine what popular anime series to follow and whether there are other like-minded fans in their home country whom they can form a fan-base with; and
2. Assist anime producers using descriptive analysis to know who their target audience are by performing viewers’ segmentation of existing anime fanbase. This is followed by using inferential analysis via regression modelling to determine the independent variables which have significant influence on the performance measures that the producers will be keen to look at (i.e. new anime series popularity and potential viewership count). Armed with this information, this will enable producers to finetune their anime offerings based on what the masses want and to develop the next big anime hit.

# Data Sources

Three files from Kaggle’s myanimelist dataset, which have been already been prepared by the data contributor, will be used. These files have filtered out users with missing data in demographic fields, and outliers with erroneously large episodes viewed, and last login dates. Anime listings without studio or sources, or those that have not aired are not used as well.

1. **anime\_cleaned.csv** contains 6,668 records of unique anime series, listing the title in various languages, genre, number of episodes, episode length, production studio, licensor, producer, popularity, rating, score, airing date, episodes, and anime adaptation source.
2. **user\_cleaned.csv** contains information about 108,711 unique anime users on the myanimelist.net website, namely username, registration date, last online date, birth date, gender, location as well as their individual statistics on anime “consumption”. Some of these values are input from the users themselves, rather than automatically generated.
3. **animalist\_cleaned.csv** contains lists of anime viewed by all users with their username, anime ID, score, status and timestamp when the record last updated. There are 31,284,030 unique anime records, of which 57% contain user ratings information.

# 

# Data Preparation

Further data preparation is required prior to analysis. Only columns of interest are retained, and entries with incomplete/missing values are removed.

For **anime\_cleaned.csv file**, there are a total of 33 column inputs. Certain input columns (e.g. anime title in Japanese, image URL) are not required for our analysis, hence these columns are removed. Table 1 (see Appendix) summarises the variables from the Animelist.csv file which were accepted or rejected. A total of 23 columns remain in the anime\_cleaned.csv file after the unnecessary columns are removed.

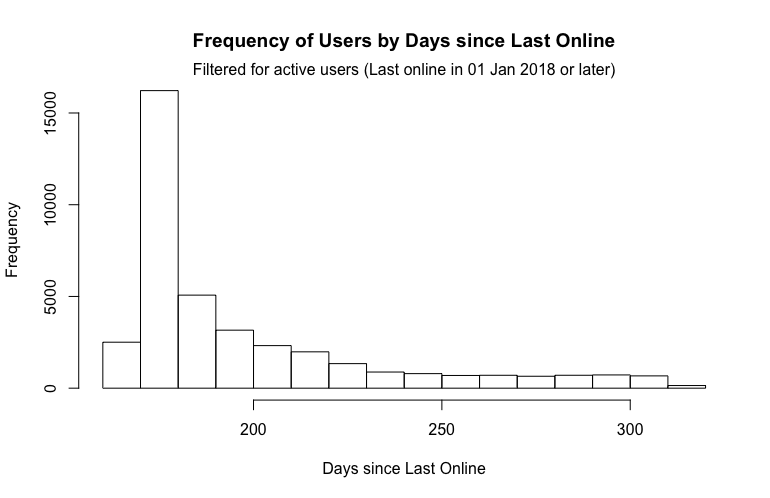
Data preparation was also done for the **user\_cleaned.csv file**, as described in Table 2 in Appendix. A total of 15 columns remain in the user\_cleaned.csv file after the unnecessary columns are removed. Similarly, this was done for **animelist\_cleaned.csv file**, as described in Table 3 in Appendix. A total of 8 columns remain.

These three files are combined. Age column is also derived using the birth date with the As.date and Age\_calc functions within R. The data dictionary for these columns can be found in Table 4 of Appendix.

The location variable was further cleaned, as many entries were not entered properly. Using R code, entries were searched against a reference list of countries/cities and a new column of countries was created. For users without any matches, manual searching of the entries was done to ensure that all of the users had a legitimate country assigned to them.

Users without valid entries, even after the code and manual search, or those which were outright gibberish were thus removed from the analysis.

Users data was then filtered for users active in 2018 only. The histogram below shows the distribution of the filtered users by their last online date.



10% of these users were sampled for the main analysis. The main dataset, **animelists\_cleaned.csv** contains variables as shown in Table 3. The final step is to join the anime\_cleaned and user\_cleaned csv files to animelists\_cleaned using anime\_id and username as identifiers. The final dataset we will use has 1,725,697 observations with 50 variables, describing each anime watched by the 10% sample of recently active users, with the user and anime information. The flowchart (Figure 1 in Appendix) summarises the entire data preparation and sampling process.

# Specific Statistical Techniques to be used

With the data set cleaned and properly sorted, we can then proceed with the visualisation of data, utilising specific statistical analysis techniques as described below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Technique | Variables | User Input | Description |
| Descriptive Analysis | | | | |
| 1 | 5-point analysis +  Boxplot | Score and Duration of Anime | Variable | Display the mean/mode/Q1-3 as a table/boxplot |
| 2 | Bar Chart + Table | Y: Continuous variables describing top Anime (favourites,members,score, duration\_min,aired from year, episodes scored\_by,rank ,popularity  X: Top N anime | Y: Category  X: Number of Anime to Display using slider | Users can choose which continuous variable of anime they want to explore and display the results accordingly. This is targeted at producers who want to derive an overview of the key factors which have an impact on the top Anime |
| 3 | Geographic distribution of Users | Age  Gender  Country | Country,  Age Range and Top N users using slider | Users can choose country, age range of users they want to find potentially near them based on the top N user |
| Inferential Analysis | | | | |
| 4 | Regression | Key Performance Measure: score, popularity, members  Independent variables:  Duration of episode, favorites, scored\_by etc | Users can choose target outcome variable and independent variables | Regression Model with Statistically Significant Variables, Scatterplot of relationship with target variable and independent variable |

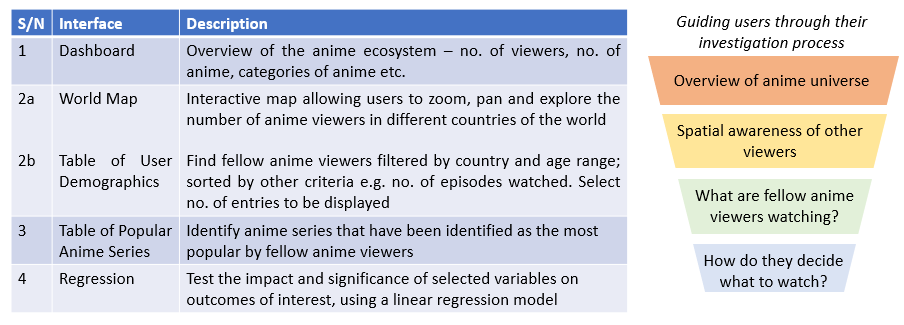
These results will then be visualised in an R Shiny application under different tabs in the user interface.

# Design of User Interface

The team has given much thought into the design of the user interface (UI), with the main objective of meeting the two main users’ needs in mind.

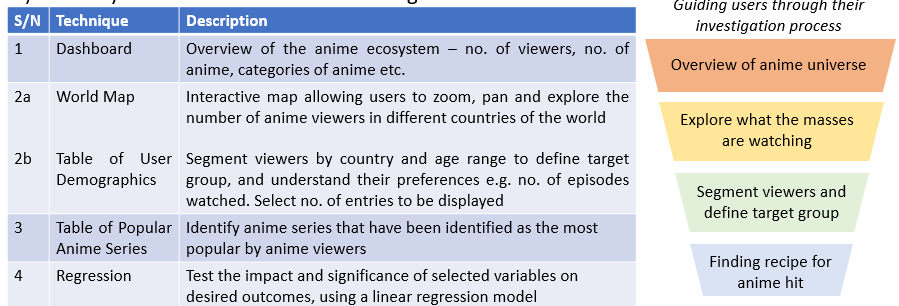
As such, the design of the UI follows a “funnel effect” whereby the users are first given an overview summary of the anime statistics before they input their selections and filter the results they require. This UI design will help them in understanding the entire dataset as well as obtaining micro level data they require.

For the case of anime fans, the UI allows for descriptive analysis to be made with the end goal of finding out what popular anime series to follow next, distribution of anime users worldwide and to assist them in finding like-minded fans in their home country to form a fan club with. Users can also search for particular users and derive key characteristics about them



*Fig 1a: User interface targeted at anime users*

For anime producers, the UI allows for descriptive analysis for preliminary data exploration to discover the current top anime series by key anime characteristics such as anime score and the target audience through viewer segmentation. This will be followed by inferential analysis to determine the possible factors that will help them develop the next big anime hit. Producers can also search for certain anime and derive key statistics.

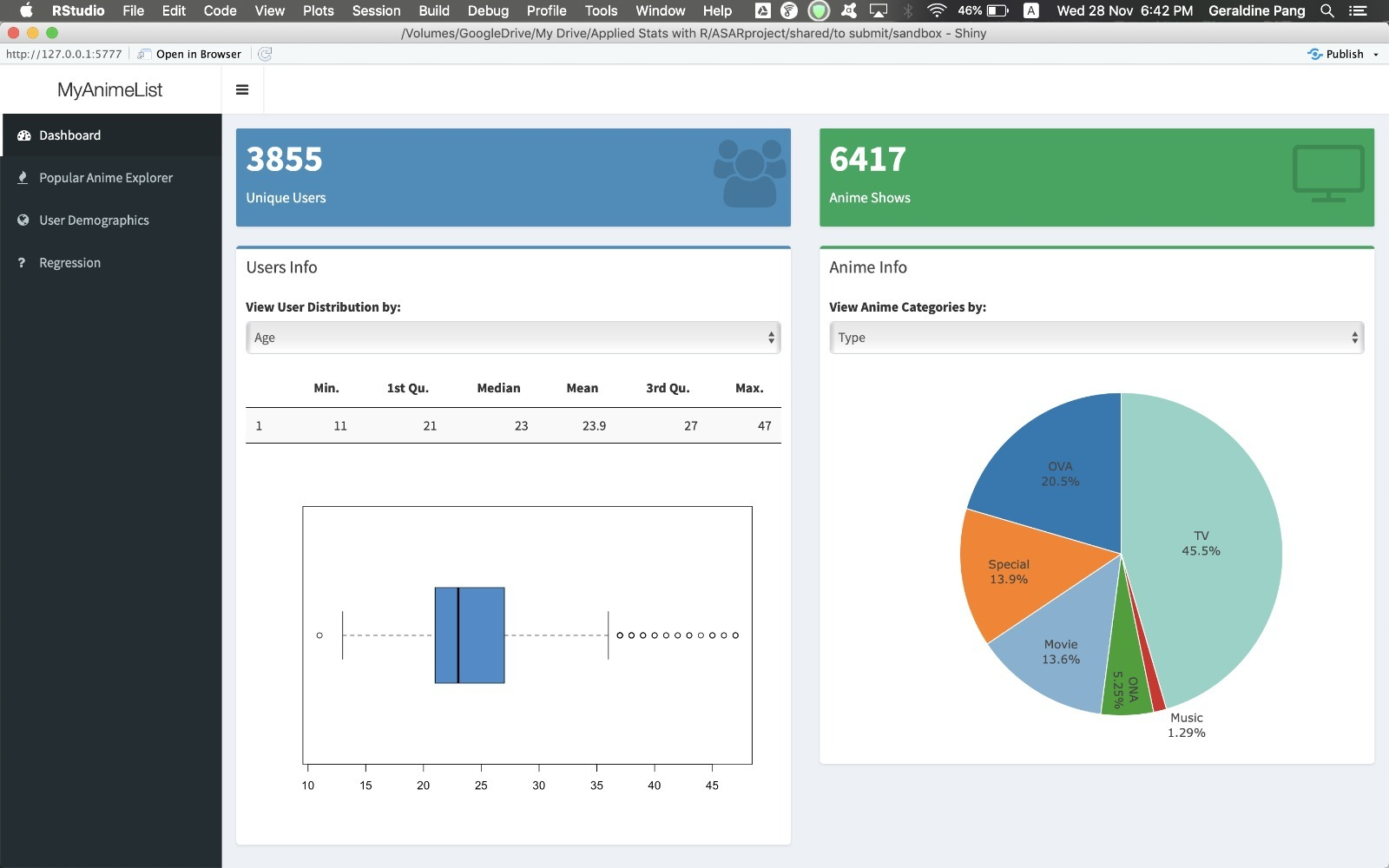


*Fig 1b: User interface targeted at anime producers*

# Exploration of Results

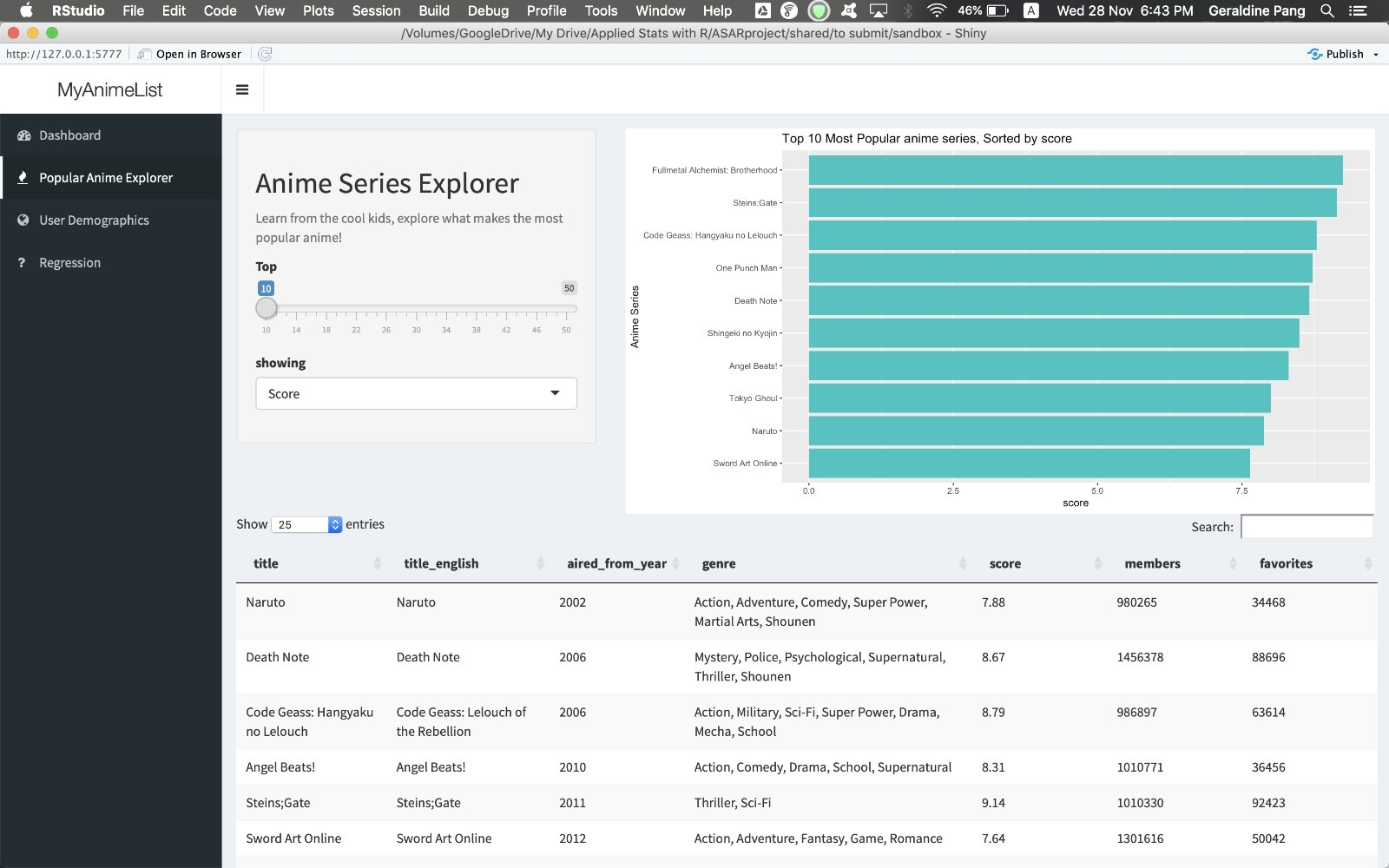
1. Dashboard

The dashboard provides an overview of the dataset being explored. It shows the number of unique anime users and anime categories. Also, we can derive box plot summary statistics of anime users based on user variables such as days last online, age which the user can choose.



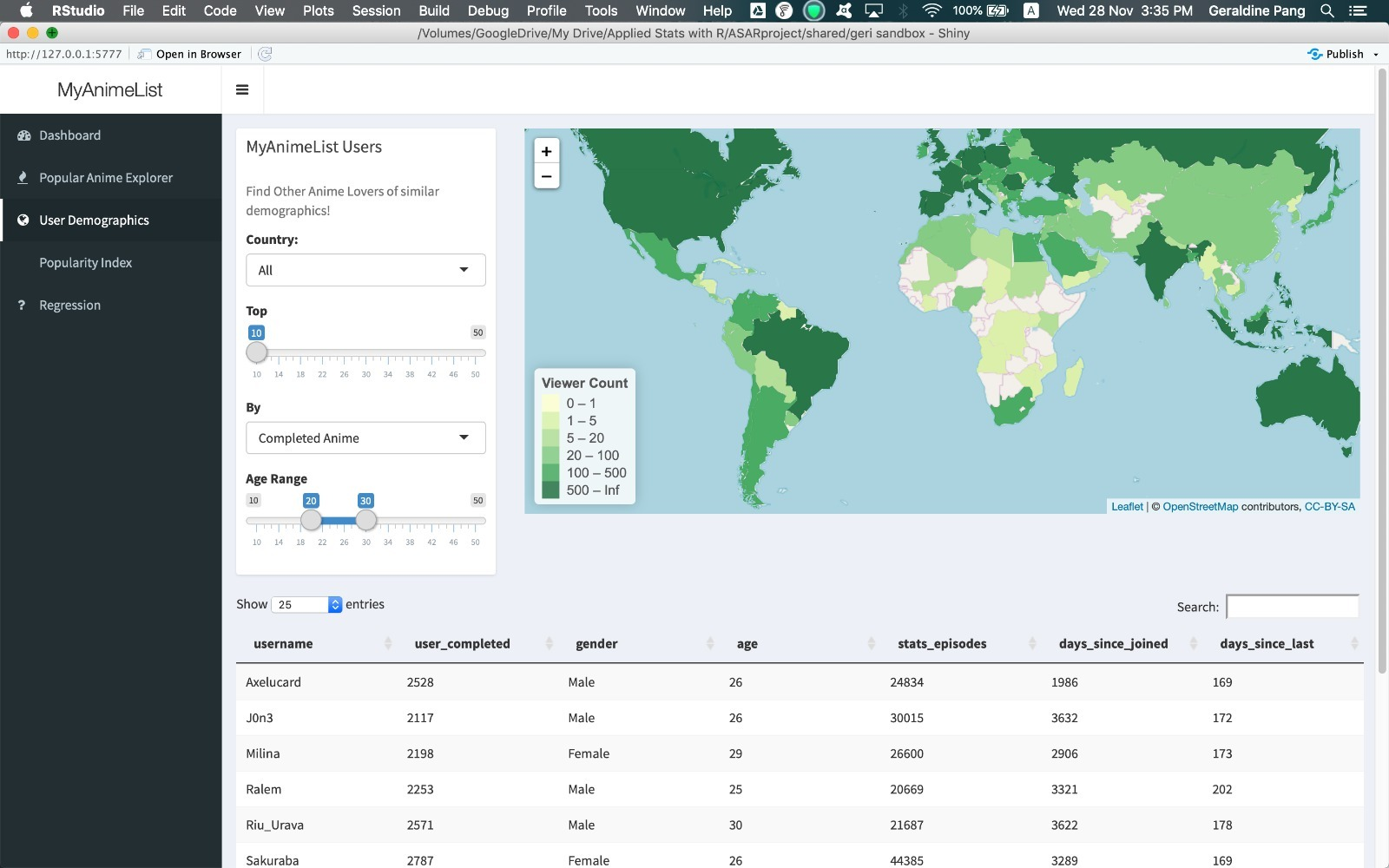
1. Popular Anime Explorer

Next, we will explore the top Anime based on Score or other key Anime variables. We can restrict the view to top 50 results. Users can also search for anime based on genre etc in the lower end of the page. This would help Anime producers significantly in identifying the success factors of a popular anime.



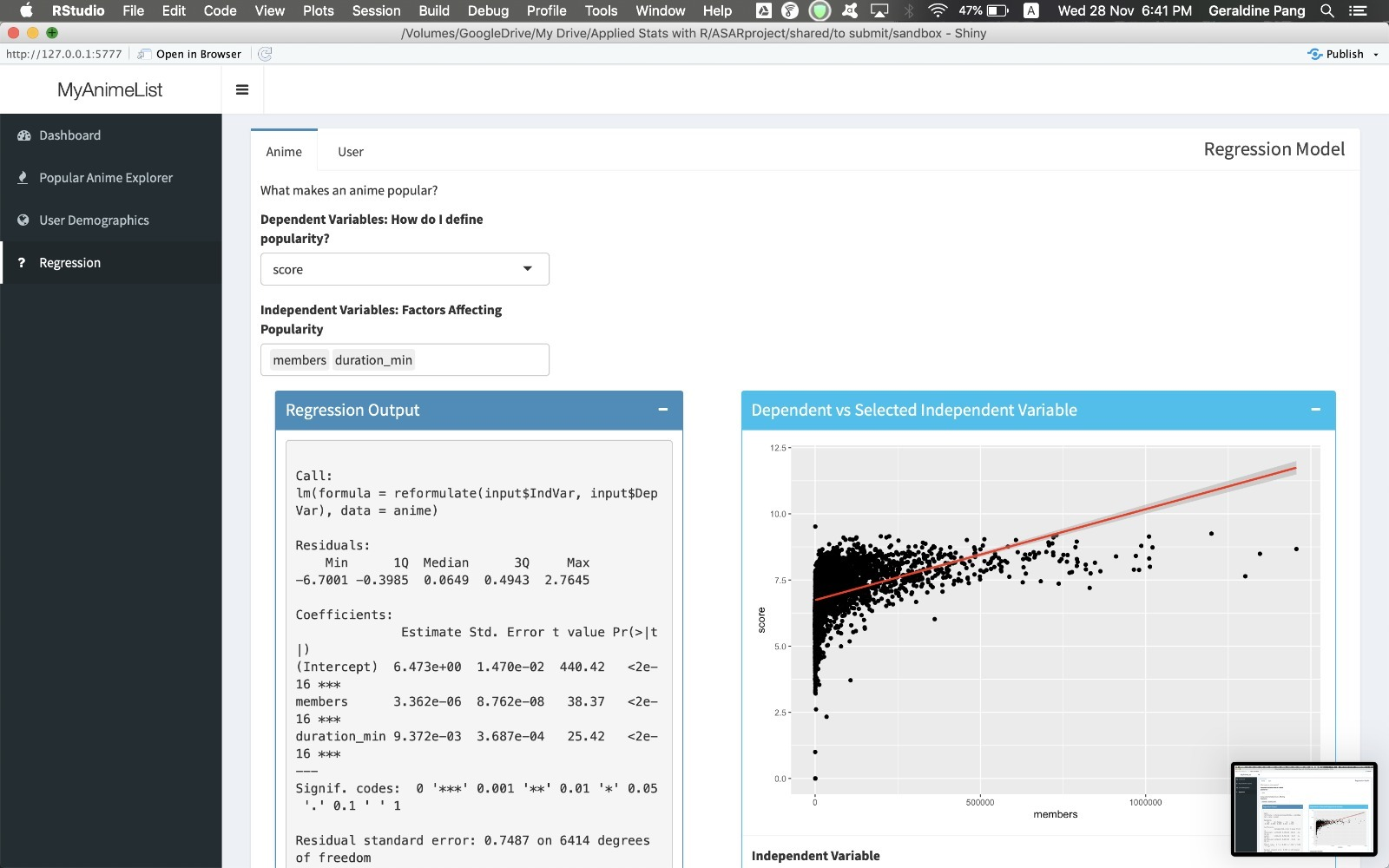
1. User Demographics

In this dashboard, users can identify fellow users based on country, age preference. This is very useful in expanding the anime community and making new friends. Also, there would be an interactive map where users can zoom, pan and explore the number of users in different countries of the world. As users hover over individual countries, the country name and number of users from that country will pop up.



1. Regression

For Regression, we decided to identify the key variables which would be significant to producers mainly and derive the interdependent relationships. We used the stepwise method learnt in our R module to predict the variables which are statistically significant on the Anime Score, Anime Popularity, Anime Members and Top Users based on Number of days spent watching Anime. Also, we can view a scatterplot to estimate the strength of the relationship between the target and explanatory variables.



We will see the regression results are generally good based on the adjusted R-squared. The key variables for Anime popularity, score, members are Rank, Scored\_by, Score, Members,

Duration\_min while the key variables for Time spent watching by Users are Days

since Joined, Stats\_Episodes, User\_Completed, Age.

# Conclusion

Our Shiny application has served its purpose to provide valuable information to the 2 main user groups, namely:

1. Anime fan – Assist them in finding popular anime series to watch and provide a listing of anime fans within their own country for meet-up. Also, could potentially establish a fan club to further propagate the popularity of anime subculture.
2. Anime producers – Assist them in producing the next big anime hit by knowing the characteristics of their targeted viewer segment and have a statistical way of determining the variables which may influence the success of their next anime series. It is hoped that production studios can tap on these during their anime ideation process to create a refreshing anime that suits their target group’s preferences.

# Appendix

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Variable** | **Data Type** | **Status** | **Remarks** |
| Anime\_ID | Identifier | Accepted |  |
| Title | Categorical | Accepted |  |
| Title\_English | Categorical | Accepted |  |
| Title\_Japanese | Categorical | Rejected | Cannot extract much information that can be deciphered |
| Title\_Synonym | Categorical | Rejected | Not much value from the information |
| Image\_URL | List | Rejected | Images are not necessary |
| Type | Categorical | Accepted |  |
| Source | Categorical | Accepted |  |
| Episodes | Numerical | Accepted |  |
| Status | Categorical | Rejected | Already captured in the Airing variable |
| Airing | Ordinal | Accepted |  |
| Aired\_String | Ordinal | Accepted |  |
| Aired | Ordinal | Accepted |  |
| Duration | Categorical | Accepted | Converted from string to numeric (e.g. “24 mins per ep” to 24.0) |
| Rating | Categorical | Accepted |  |
| Score | Continuous | Accepted |  |
| Scored\_By | Categorical | Accepted |  |
| Rank | Continuous | Accepted |  |
| Popularity | Continuous | Accepted |  |
| Members | Continuous | Accepted |  |
| Favourites | Continuous | Accepted |  |
| Background | Categorical | Rejected | Background of the anime is not important |
| Premiered | Categorical | Rejected | The data is represented in seasons and they do not really matter as this information is already captured in the Aired variable |
| Broadcast | Categorical | Accepted |  |
| Related | Categorical | Rejected | The data is not structured and not much value can be derived anyway |
| Producer | Categorical | Rejected | Too many producers and not much value can be derived from this |
| Licensor | Categorical | Rejected | The Studio variable would be a better gauge of top production statistics. Licensor does not give much information |
| Studio | Categorical | Accepted |  |
| Genre | Categorical | Accepted |  |
| Opening Theme | Categorical | Rejected | The theme songs are too varied, and no significant analysis can be derived |
| Ending Theme | Categorical | Rejected | The theme songs are too varied, and no significant analysis can be derived |
| Duration\_Min | Continuous | Accepted | Derived from Duration |
| Aired From | Date | Accepted | Derived from Aired |
| Related IDs | List | Accepted | Derived from Related |

*Table 1: Column input variables in the anime\_cleaned csv file and the decision on whether to accept or reject them for further analysis.*

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Data Type** | **Status** | **Remarks** |
| Username | Identifier | Accepted | Unique identifiers for each user (used to merge with other data table) |
| User\_id | Identifier | Rejected | Unique IDs only found in this table |
| user\_watching | Numerical | Accepted |  |
| user\_completed | Numerical | Accepted |  |
| user\_onhold | Numerical | Accepted |  |
| user\_dropped | Numerical | Accepted |  |
| user\_plantowatch | Numerical | Accepted |  |
| user\_days\_spent\_watching | Numerical | Accepted | Rounded up with Ceiling() |
| Gender | Categorical | Accepted |  |
| Location | Categorical | Accepted | Requires further cleaning to extract exact country names:   * UTF-8 characters removed * Puncutations, digits removed * Changed to lower case |
| Birth\_date | Ordinal | Accepted | Converted into age using (difference from Sys.Date) |
| Access\_rank | - | Rejected | All “NA” values |
| join\_date | Ordinal | Accepted | Derive days since joined (difference from Sys.Date) |
| last\_online | Ordinal | Accepted | Derive days since last online (difference from Sys.Date) |
| stats\_mean\_score | Continuous | Accepted |  |
| stats\_rewatched | Continuous | Accepted |  |
| stats\_episodes | Continuous | Accepted |  |

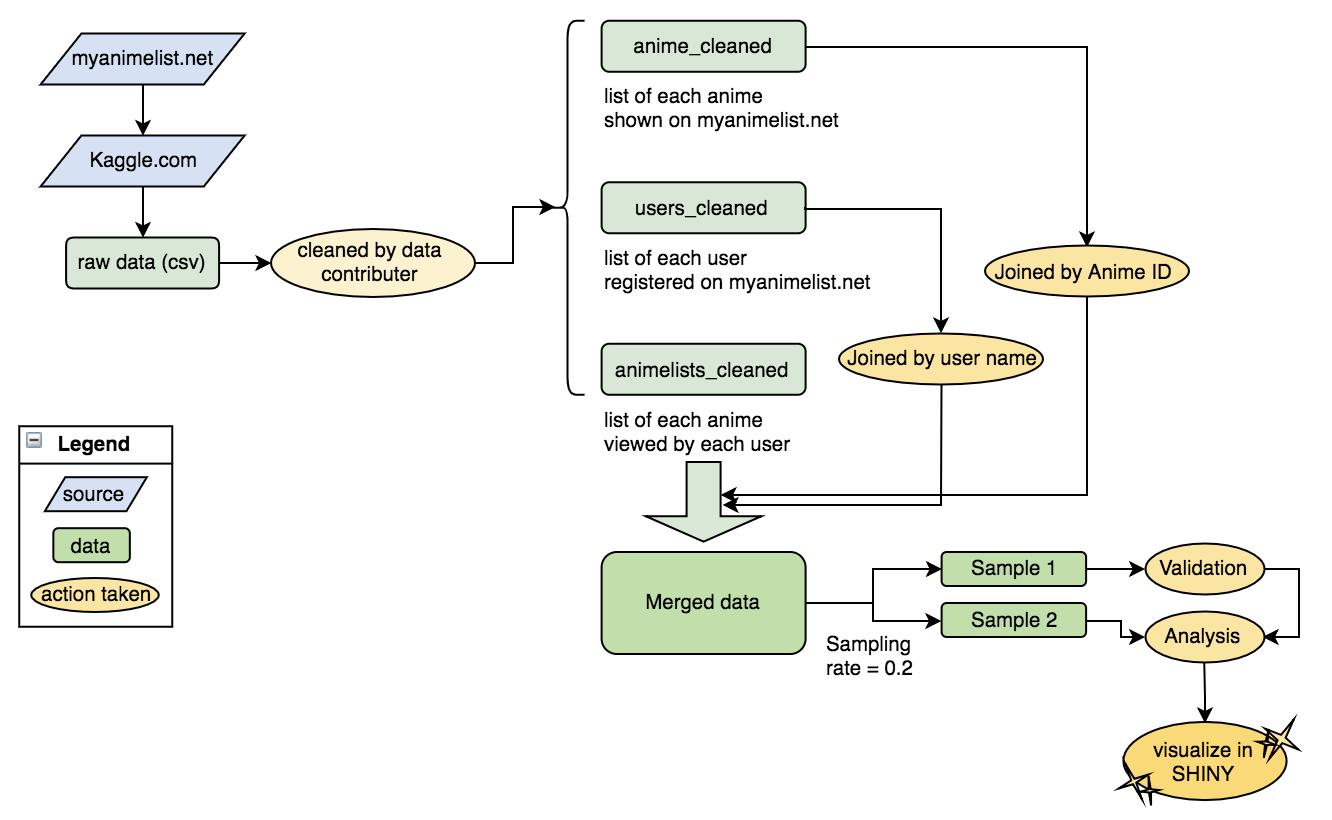
*Table 2: Column input variables in the users\_cleaned csv file and the decision whether to accept or reject them for further analysis.*

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Data Type** | **Status** | **Remarks** |
| username | Identifier | Accepted | Identifier for merging users\_cleaned table |
| Anime\_Id | Identifier | Accepted | Identifier for merging anime\_cleaned table |
| my\_watched\_episodes | Continuous | Accepted |  |
| my\_start\_date | Date | Rejected | 84% of the data is missing (coded as “0000-00-00" instead of real dates) |
| my\_finish\_date | Date | Rejected | 81% of the data is missing (coded as “0000-00-00" instead of real dates) |
| my\_score | Continuous | Accepted |  |
| My\_status | Ordinal | Accepted | Contains integer values representing:   * 1: watching * 2: completed * 3: on hold * 4: dropped * 6: plan to watch   Other values are not known (80% of the data). |
| my\_rewatching | Binary | Accepted | 20% NA |
| My\_rewatching\_ep | Continuous | Accepted | One outlier was removed. |
| My\_last\_updated | Date | Accepted |  |
| my\_tags | Categorical | Rejected | Missing in 29,289,278 rows (93% of the data) |

*Table 3: Column input variables in the animelists\_cleaned csv file and the decision whether whether to accept or reject them for further analysis.*

|  |  |
| --- | --- |
| **Column** | **Definition** |
| Anime\_ID | Identifier for Anime |
| Title | Anime Title |
| Title\_English | Anime Title Translated |
| Type | Where the Anime can be viewed (E.g. TV, Internet) |
| Source | The type of Anime (E.g. Manga, OVA- original video animation) |
| Episodes | Episodes for each specific Anime |
| Airing | Whether the anime has aired finishing or still airing |
| Aired\_String | Aired duration range (formatted) |
| Aired | Aired duration range (unformatted) |
| Duration | Duration of Anime per episode (unformatted) |
| Rating | Censor board Classifications |
| Score | Score of Anime given by Users |
| Scored\_By | Number of Users who scored the Anime |
| Rank | Average Rank of Anime gathered from Users |
| Popularity | Rank tabulated based on number of Members for each Anime |
| Members | Number of Users watching Anime |
| Favorites | Number of Users who added Anime to its Favorites |
| Broadcast | The day and timing of the Anime |
| Studio | The studio who produced the Anime |
| Genre | The genre of Anime (E.g. Action, Adventure, Comedy) |
| Duration\_Min | Duration of Anime per episode(formatted) |
| Aired\_From\_Year | Year Anime started Airing from |
| Related IDs | Related Sequels and Prequels for each Anime |
| Username | Unigue name identifier of Anime user |
| user\_watching | Number of users watching a particular Anime currently |
| user\_completed | Number of users completed watching particular Anime |
| user\_onhold | Number of users watching a particular Anime which is On Hold |
| user\_dropped | Number of users who dropped watching a particular Anime |
| user\_plantowatch | Number of users who plan to watch a particular Anime |
| user\_days\_spent\_watching | Number of users spent watching a particular Anime |
| gender | Gender of Anime user |
| location | Location of Anime user (This would be later cleaned, according to country) |
| Birth\_date | Birth date of Anime user |
| join\_date | Date when Anime user joined Animelist.net |
| last\_online | Date when Anime user is last online on Animelist.net |
| stats\_mean\_score | User average score based on usage of Animelist.net |
| stats\_rewatched | User average score based on anime rewatched by user |
| stats\_episodes | Number of Episodes watched by user |
| my\_watched\_episodes | Number of Anime watched by user |
| my\_score | Score of Anime based on User and Anime (combined) |
| My\_status | Score of user based on status |
| my\_rewatching | Number of anime rewatching |
| My\_rewatching\_ep | Number of anime episodes rewatching |
| My\_last\_updated | Time when user last updated on Animelist.net |
| Age | Age of Anime user |

*Table 4: Data dictionary of combined columns in final data file.*



*Figure 1: Overview of entire data preparation/analysis process*