

Statistical Analysis of Anime Series with R

Trend ANalysis and Demographics study of the otaku SUBculture

Prepared by:

Geraldine Pang, Kasi Gopikanna, Kevin Nathaniel, Benjamin Tan, Tan Kah Kyee

Applied Statistical Analysis with R

24 Oct 2018

Contents

[Introduction 2](#_Toc526409848)

[Overall Concept (Purpose and Key Idea, Target Audience) 2](#_Toc526409849)

[Data Sources (Detailed description of data sources, fields and transformed variables) 3](#_Toc526409850)

[Specific Methodology (Statistical techniques and analytics employed) 4](#_Toc526409851)

[Conclusion / Summary 5](#_Toc526409852)

# Introduction

Anime (abbreviation for the word animation) is produced in Japan and it differs from western cartoons, focusing 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 fasciation of anime isn’t limited to only Japan. This popular artform has since propagated out of Japan and has become a worldwide phenomenon with Western adaptation of popular Japanese anime series like “Ghost in the Shell”, “Astro Boy” and “Death Note”.

The purpose of this data analysis project is to examine the popular Japanese otaku culture and derive some interesting insights into this unique and quirky subculture.

# Overall Concept (Purpose and Key Idea, Target Audience)

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

Information is captured by the website and it is based on the user ids for demographics analysis and trends. The dataset information includes users’ general particulars like gender, location, anime ratings, birth dates as well as anime information such as genres, number of episodes and production studio.

Useful insights can be gathered from the demographics of anime fans as well as the ratings of anime. With the information at hand, this report can serve two purposes:

1. (Target Audience 1) for a casual anime fan to know what are the popular anime series to follow and whether there are other like-minded fans in their countries to form a fan-base, or,
2. (Target Audience 2) for anime producers to know who are their target audience and their inclination towards any particular genre of anime. This will serve as a baseline for producers to develop the next big anime series.

# Data Sources (Detailed description of data sources, fields and transformed variables)

< to decide which dataset to use cleaned or filtered? Suggestion – use the cleaned dataset and do our own data cleaning and represent it below with detailed description of what had been cleaned up>

The dataset has a total of three files:

1. AnimeList.csv contains 31 columns of data which list down the anime title, title synonyms, genre, number of episodes, episode length, production studio, licensor, producer, popularity, ranking, score, airing date, episodes, anime adaptation source and anime fans who have watched and reviewed the anime.
2. UserList.csv contains geographical information about users who watch anime, namely username, registration date, last online date, birth date, gender, location as well as their individual statistics on anime “consumption”.
3. UserAnimeList.csv contains anime lists of all users with their username, anime ID, score, status and timestamp when the record last updated.

The dataset as a whole contains:

* 302,675 unique users
* 302,573 of them with some demographic data entered (to be cleaned)
* 80,076,112 records in anime lists
* 46,358,322 records of ratings given
* 14,478 unique anime series watched/rated

<to amend according to the inputs that have been cleaned>

There is filtered version of dataset is contained in files anime\_filtered.csv, animelists\_filtered.csv and users\_filtered.csv. It consists of users who have birth date, location and gender filled. So it contains lot less animelists data. But all important characteristics like rating mean and variation, or genres in animelists is unchanged when ommiting users with some missing data, so even with filtered data we should get same information. The filtered dataset contains:

* 116,133 unique users with demographic data
* 35,802,010 records in anime lists
* 20,726,794 records of ratings given
* 14,474 unique anime series watched/rated

There is also cleaned version of the filtered dataset which consists of files anime\_cleaned.csv, animelists\_cleaned.csv and users\_cleaned.csv. This cleaned version has trucated all users with ridiculously large number of episodes in anime which obviously don't have that much episodes, watched episodes larger than number of episodes in individual anime were fixed and seen episodes and watch time were recalculated accordingly. For some users, last online was 1900 year, just weird values, so their last activity was inferred from their last animelist update timestamp. Many users incorrectly filled number of rewatched episodes. For anime where more episodes have been watched than that anime has episodes, watched episodes have been rewritten to number of episodes in that anime. Watch time and number of watched episodes have been fixed accordingly. Users too young and too old obviously were truncated too. 6 users with most episodes seen, suspiciously lots of episodes, were truncated here too. That is too few users to affect any statistics. Anime with unknown studio or unknown source were discarded too. Also anime which were not yet aired were discarded. Their ratings were removed too. Removing them did not affect much other statistics, and without studio or source they did not give much information. Mostly unknown and insignificant anime were removed that way. Airing year was calculated for all remaining anime.

# Specific Methodology (Statistical techniques and analytics employed)

Initial ideas:

1. World map displaying the users’ location to denote the concentration of anime fans
2. Histogram showing the top 10 popular anime series (based on score/rating) by year/genre/production studio
3. Histogram showing the number of anime produced per year
4. User demographics chart – age/location/gender/number of anime watched

# Conclusion / Summary