The dataset consists of three files:

1. Users.csv

To recap, it contains information about users who watch anime, namely username, registration date (join\_date), last online date, birth date, gender, location, and lots of aggregated values from their anime lists.

Users birth\_date, last\_online and join\_date columns were converted to datetime and users with invalid dates for the aforementioned columns were dropped. The last online column data was updated using the user’s most recent myanimelist’s my\_last\_updated value because it was discovered that the last\_online value lagged behind a user’s most recent animelist update time.

Two users ('warrose','DeliberatorEVA') had an unusually high number of my\_watched\_episodes; these users were deleted from the dataset. Specifically, ***warrose*** had the highest number of animes with unsually high no of rewatches per episode (65535 rewatches), way much higher than the number of total episodes; he was an outlier that skewed distributions of calculations such as the no of days spent watching anime. **DeliberatorEVA** also had a sizeable number of animes with a unusually high no of rewatches per episode; in addition, he had not scored any of the anime in his list.

1. Animelists.csv

To recap, it contains anime lists of all users. Per each record, here is username, anime ID, score, status and timestamp when was this record last updated.

While exploring the animelists.csv data, I discovered unusual my\_watched\_episodes data for a number of users. There were users who had rewatched an anime 65535 times. After doing some reading on the myanimelists forum, I discovered myanimelist allows users to input the watched episode amount for ongoing anime with no official airing end date. Also in late October, early September of 2014 a moderator's account got compromised and the episode count of a lot of entries changed to 9001 episodes. If people added an entry as "completed" to their list, the episode count was automatically filled with the number of episodes (9001 in this case).’Currently Airing’ anime with an episode count of 0 (equivalent of ‘unknown’ status on myanimelist) were identified and rows with a value higher than 2444 for the watched episode count were deleted from the dataset. The benchmark was 2444 because the longest running anime series Sazae-san has roughly that many aired episodes. Next, rows with my\_watched\_episodes > (total episodes) \* (number of times user has rewatched the anime) were repleaced with total\_episodes.

1. Animes.csv

To recap, it contains a list of anime, with title, title synonyms, genre, studio, licensor, producer, duration, rating, score, airing date, episodes, source (manga, light novel etc.) and many other important data about individual anime.

Animes with no genre info were deleted and (aired from year) info was extracted from aired string and aired column.