Description of data

The data we're using in our project come from MyAnimeList.net, a site on which people can upload their lists of watched anime (animated, typically Japanese, TV shows and movies). MyAnimeList (henceforth referred to as MAL) is similar to iMDB: users can rate anime (1-10), leave reviews, and interact with other users in forums. In our analysis, we use a Kaggle dataset composed of data scraped from MAL regarding users and the anime they've watched.

Descriptions of datasets

The Kaggle dataset is composed of several files, some of which are "cleaned" versions of others. A table of files and descriptions is as follows:

file	description
`AnimeList.csv`	raw list of anime on MyAnimeList
`UserAnimeList.csv`	collection of the individual anime that each user on the site has watched
`UserList.csv`	list of users on the MyAnimeList site
`anime_filtered.csv`	filtered version of `AnimeList.csv`
`anime_cleaned.csv`	cleaned version of `anime_filtered.csv`
`animelists_filtered.csv`	filtered version of `UserAnimeList.csv`
`animelists_cleaned.csv`	cleaned version of `animelists_filtered.csv`
`users_filtered.csv`	filtered version of `UserList.csv`
`users_cleaned.csv`	cleaned version of `users_filtered.csv`

The filtered versions of the files contain only those users with birth date, location, and gender filled. The cleaned versions of the filtered files have outliers removed (e.g., users with excessively large watchtimes), and nonsensical data removed (such as users with episode counts that exceed the episode count of a given anime). Of these files, we're going to be using the cleaned ones. The missing data that was removed in the cleaning process is not necessary in order to ensure that our analysis is done with sufficient rows; since we started with 80 million users, the paring-down of the usercount left us with enough data.

Location data cleaning

Since we want to analyze user location as one of our features, we needed to parse the locations given to us by users. Unfortunately, MAL has no standardized way for users to input their locations; instead, users simply type in whatever they want. This led to our having to discard a large quantity of data due to poorly-defined locations.

Furthermore, we decided (for the sake of simplicity) to focus our analysis solely on the users whose location text implied that they were in the United States, with the goal of analyzing location on a state-by-state basis. To do this, we made several dictionaries of locations, and used these dictionaries to parse the location strings given by the users. We created a string-parsing function using these dictionaries and then applied it to the relevant dataframe columns. Once this was done, we were able to isolate the rows which had location data that was convertable to a US state, and put that data into a separate dataframe. Finally, we converted said dataframe back into a CSV to get the final version of the `users_cleaned.csv` dataset.