CSC 255 - Dealing with Data
Spencer Rudnick, Nobuki Harata, Michael Paquette
2-2-17

# Stories on Movie Data

## Motivation

Our original goal was to build an application where a user could input his or her demographic information and rely on a predictive model that would take that information and recommend movies to the user. The idea was to compare the user's information with ratings made by other users. Because time was not on our side we decided to flip the project on its head. We asked the question “what makes a movie popular?” The purpose was to look at what kind of movie would be popular with a given target audience. Our updated goal was to display interesting and informative stories about the popularity of films through illustrations.

For the scope of this project we focused primarily on genre preferences tied to the users’ Sex variable. We would have liked to perform in-depth analysis and create visualizations comparing and contrasting preferences based on Occupation, Location, and Age, however relying exclusively on Sex (a binary variable for the purposes of this study) freed up our time to do more in-depth analysis of how genres relate to one another. Much of the code we used to compare and contrast preferences based on a user’s Sex could be repurposed to find connections between other demographics data and movie preferences.

## Data Collection

We began exploring movies using a dataset from MovieLens that contained 20 million ratings from 2013. We converted the files to csv files which were usable in R and began inspecting the data. However, we found out that this dataset did not provide demographic data. Our initial goal was to make a movie recommender based on demographic data and previously liked movies, so we decided to use a different dataset which included demographic data. We settled on the 1M dataset made in 2003,  which was also from MovieLens.

This 1M dataset came in three different .dat files: Users, Movies, and Ratings, which contained one million ratings. When computing this dataset, it was much faster than the larger 20M dataset. The group agreed that the tradeoff was worth losing the large amount of data. The 2003 dataset provided occupation, zip code, age, and sex data for each user.

## Data Cleaning

The first step we took to clean the data was to convert the .dat files to .csv files as mentioned above. We used a basic text editor to replace a set of double colons with commas and saved them as a .csv file. Next, we made the decision to remove any movie with less than 23 ratings and the ratings associated with those movies. This was because having a low number of ratings for those movies made the ratings less reliable and more susceptible to chance.

The Ratings table was then merged into the Users table. Each rating was stored in a list of complex numbers associated with the user who made the rating. Storing the MovieID and Rating in a complex number allowed us to keep them grouped. The main reason we combined the two tables was to allow us to more easily connect ratings with demographic information. Another reason we did this was to match a 10,000 row limit on tables set by Heroku on their free web server that we planned to use. Heroku was later dropped due to not having enough memory.

The next step that we took was to process data into new tables to give us a cleaner picture of the patterns in the dataset. We created a new table, Genres, to analyze the popularity of each genre. This table contained a list of each genre, average male ratings, average female ratings, the difference in the two average ratings, and an average score. This score was created to account for both the rating count and the average rating. This was calculated by: score=multiplier*adjustedRating. The multiplier was the log(number of ratings) with a base of the first quartile of number of ratings. The adjusted rating was the averageRating - 1. This gave us a fairly even balance of ratings count with average ratings. We decided to do this because neither number of ratings nor average ratings alone wouldn't give us the best representation of the popularity of the movie.

## Results

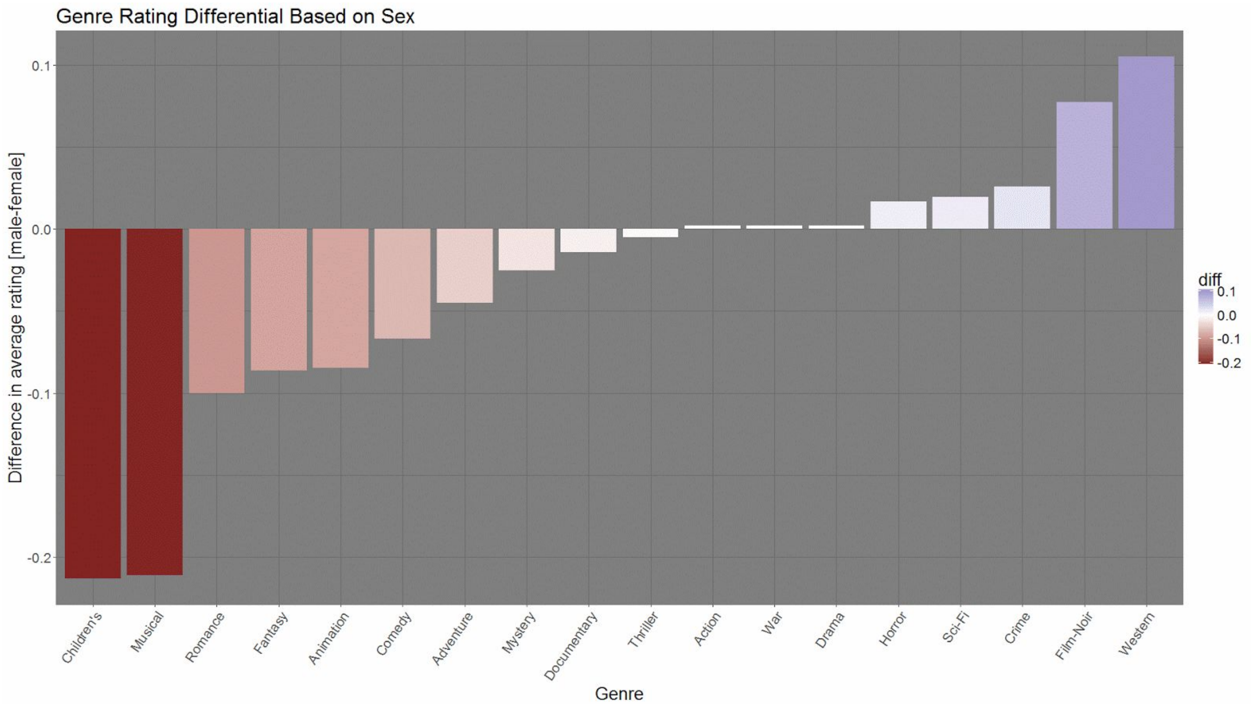

Figure 1

The first interesting find that we had was the relation between genre preference and sex. Figure 1 illustrates the difference in rating between males and females for each of the genres. We decided to look into this because we saw that there was almost no difference in the overall average rating between sexes, but we expected a difference, so we took a deeper look into which genres they prefered. The difference between male and female average rating for each genre was plotted on the y axis. This difference is the same difference calculated in the genres table. A negative difference indicates that the female average rating was higher than the male average rating. The largest difference that prefer females is the children's genre which had a difference of -0.213, while it was western movies for males with a difference of 0.105.

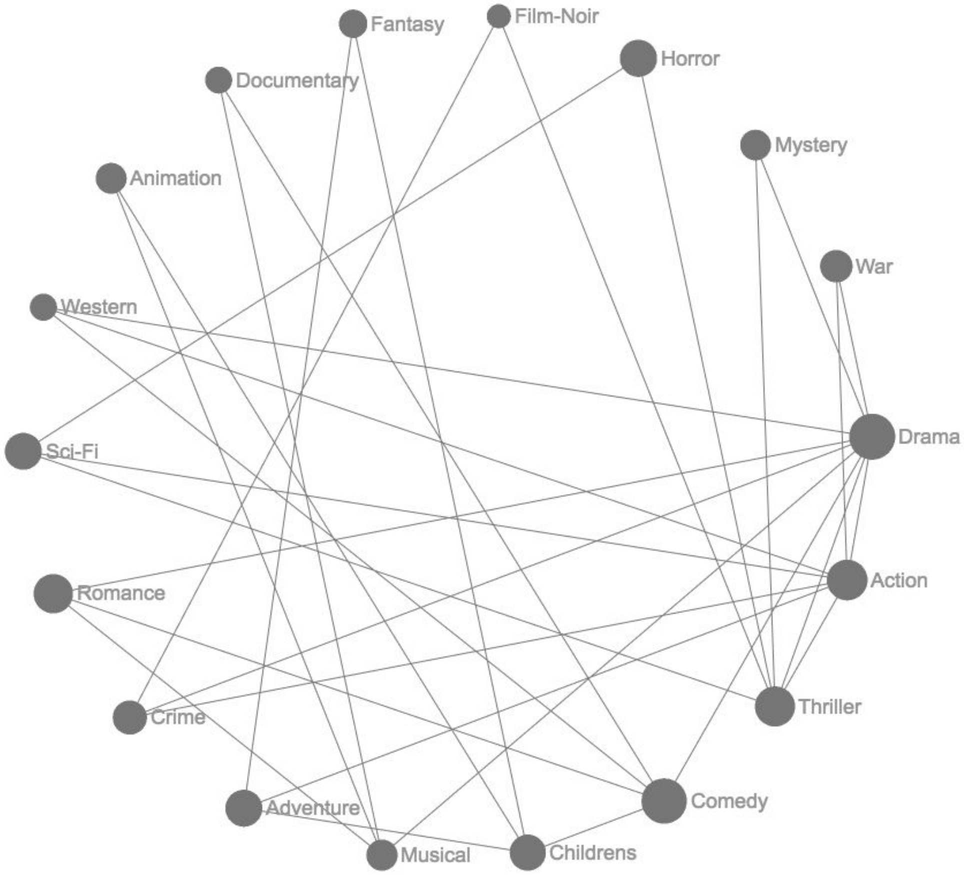

Figure 2

The second visualization we made that we found interesting was figure 2. In this figure, we plotted each genre in a circular positioning, ordered in terms of most- to least-connected, with each genre represented by a node of various sizes. The size of the node denotes the number of movies in that category. However, we took the log of the number of movies because the difference between the most popular genres and least popular genres were so large that we weren't able to see the smaller circles while keeping the larger ones a reasonable size. Each line connecting the nodes represents a strong connection between the two genres. A strong connection was determined if it met at least one of the two criteria we set. If the two genres had more than 83 connections (50th percentile of connections) it was a strong connection. If a genre had no connections greater than the median, we selected that genre's strongest two connections to include in the graph.

## Analysis

Figure one showed a very predictable pattern for the preferences in male and female ratings. As stereotypically assumed, females prefered children's, musicals,  romance, and fantasy movies. Males on the other hand prefered movies such as western, film-noir, crime, sci-fi, and horror. Surprisingly, thriller, action, war, and drama movies were rated evenly between sexes.

The children's and musical movies had the greatest difference with over -0.2, both of which were prefered by females. The difference in children's movies could come from the fact that stereotypically, females historically tend to spend more time with children and have more exposure to those movies. Furthermore,  females may even rate movies based on how much their own children like the movies, while males may tend to base their rating on their own interests. We don't think there's any particularly interesting reasoning for the preference in musicals for females. Stereotypically, females are thought to like musicals, and many musical movies are geared towards the female audience.

The connectivity graph shown in Figure two shows some interesting things. Because the nodes are ordered around the circle by number of connections, all the genres that have many connections are on grouped together. Drama has the most connections at 9 genre connections above the median. The number of movies associated with drama is 2,761, which is 70% of all the movies.  Drama invokes a lot of emotions and plays on situations that many people can relate to. It is essential to many story lines. Even if a movie is sci-fi, without a drama aspect it may be uninteresting.

## Conclusion

It is clear that there is a connection between demographic data and the preferences in the movies of the users. As we saw previously, the gender and the genre preference was shown with predictable results. One of our limitations was that we did not have the resources to hire a statistician into our programmer heavy team. We would like to expand our workers to include statisticians when moving forward in the future. They would be able to provide valuable insight on how to analyze the data to find deeper, more reliable connections between a person's demographic information and their movie preferences.

## Future Work

If we had more time to continue the project, we would look deeper and wider into the demographic data, including variables such as age, occupation, and location. We would also take into account more variables related to movies such as the year of release, movie length, and any awards the movie received.

With more detailed and in-depth statistical analysis, we would develop a predictive model which would power the app we originally planned to create, a movie recommending app that relies on users' demographic data and previously liked movies to suggest movies they may also like.

We would also like to analyze the difference between the least popular movies and the most popular movies. We would slice off the top and bottom quartiles of movies based on their score. We would then analyze these sets of movies, exploring similarities and differences among and between sets.