Exploratory Data Analysis: Movie Lens

Sam Cook

The Movie Lens data set is a data set consisting of movie reviews. The data includes the movies, their, reviews, as well as demographic information of the reviewers like age, occupation, and zip code.

We can see what the most popular movies by looking at how many reviews there are for the movie. These movies may not be the best movies, but they have the most reviews and it could be inferred that they have been viewed the most.

```
Star Wars (1977)                    583
Contact (1997)                      509
Fargo (1996)                        508
Return of the Jedi (1983)           507
Liar Liar (1997)                    485
English Patient, The (1996)         481
Scream (1996)                       478
Toy Story (1995)                    452
Air Force One (1997)                431
Independence Day (ID4) (1996)       429
```

This is not surprising, as these films were box office hits. We can also evaluate the most popular movies by age group. If we look at reviews for people under age twenty, the top five reviewed are

```
Scream (1996)                                       65
Contact (1997)                                      50
Liar Liar (1997)                                    48
Star Wars (1977)                                    47
Return of the Jedi (1983)                           44
```

These are all the same movies, just in a slightly different order. If we look at the most reviewed movies from adults over age 60, our results are

```
English Patient, The (1996)     17
Fargo (1996)                    12
Full Monty, The (1997)          10
Titanic (1997)                  10
Godfather, The (1972)            9
```

An interesting analysis for the future would involved the genres of popular movies by age groups. We can see the age demographics of the reviews in the histogram to the right. Most users are in the age range from approximately 20 to approximately 40.

User age distribution