IMDB Data Mining Project

Richard Joerger
Rochester Institute of Technology
raj2348@g.rit.edu

Yifei Sun
Rochester Institute of Technology
ys8800@g.rit.edu

Ajeeta Khatri
Rochester Institute of Technology
ak6038@g.rit.edu

Zhuo Liu
Rochester Institute of Technology
zl9901@g.rit.edu

1. OVERVIEW

The purpose of this paper is to explain our data set and to begin investigating that data set through the lens of data mining. We've decided to continue working with the IMDb data set as we have experience working with it which should reduce the amount of time it takes for us to get "up to speed" when it comes to understanding the irregularities of the data. An example of this time reduction is that we didn't have to spend time figuring out how the data is interconnected, instead we could spend our time focusing on how to prepare our data for analysis.

2. DATA SET

It's possible to summarize our data set in one word: entertainment. The data hosted and shared by IMDb spans movies, short films, TV episodes, TV specials, TV movies, videos, and video games. It's not just titles and release dates, there is also birth and death information about all of the people who worked on these projects to bring them together and make the entertainment product people were enjoying.

So starting in the most broad sense, in the title.basic data, an entertainment product has an original title, a primary title, a type - which is one of the aforementioned types -, a start and end year, a release region, if the title is for adults, a run time in minutes, and a list of genres. There is also a unique identifier for each title called tconst which is the mechanism that is used to link together the various segments of data. In a separate file we have information about the various localized versions of an entertainment product. This table is called title.akas, it has titleId - which is just a different name for tconst as these values map 1 to 1 -, the ordering of a product within the entire data set, the title of the movie as well as the region for that title, the language of the title, and if it is the original version. It's important to note that these two tables, when merged together, provide all of the information about an entertainment product for every where in the world and maintains the connection between the original and all of the localized versions. Ratings are something that entertainment could do without but regardless, everything has someone out there who is an opinion of it. The rating table is quite simple. It has a tconst for the entertainment product, the average rating, and the number of votes

An important aspect of any entertainment product are the people who made it because without them, there wouldn't be anything to entertain us. The IMDb data set also has this kind of information. For each person they have a unique constant - nmconst -, a primary name, a birth and death year, a list of the primary professions, and what titles they are known for which is a list of tconst values. This information can be used to link those who work on movies to the movies they're best known for, which provides us another relationship to analyze. The intermediate table which allow us to link these two in a far better fashion is the principals table. This table has a tcosnt value, the ordering of the person in the credits, the persons's name, their category in that film, their job name if they fall within a subset of the category. An example of this would be the director category may contain someone who is the director of photography. They fall within the director category but they're job pertained to managing and "directing" the cameraman to match the director's vision. The last column in the principals section is the character name, if it is applicable. A director doesn't have a character name but an actor will have a real name and the character they're playing.

The very last table is the episode table. This relates two tconst values. The first is the individual episode tconst, the second is the tconst for the season that episode belongs to. This table also contains data about what season number the episode is in as well as what the episode number itself is.

3. DATA SET CONJECTURES

One the things that happens as you begin to understand a data set is you start connecting the dots between the different aspects of the data. For example, since time has gone on more people have probably engaged with newer entertainment products which means that more people will have a greater variety of opinions and therefore those entertainment products will have a score that is far closer to the middle. Additionally, we believe there will be a correlation between average rating and run time. Looking at 1 we can see the distribution that there appears to be a bi-modal distribution around 8 and 5.

Then looking at figure 2 we see that there really is no obvious correlation between run time and the average rating.

Looking at 3 we can see that as time goes on, the amount of ratings increases as time goes on and that the spread of ratings for release year vs. average rating.

4. DATA MINING IDEAS

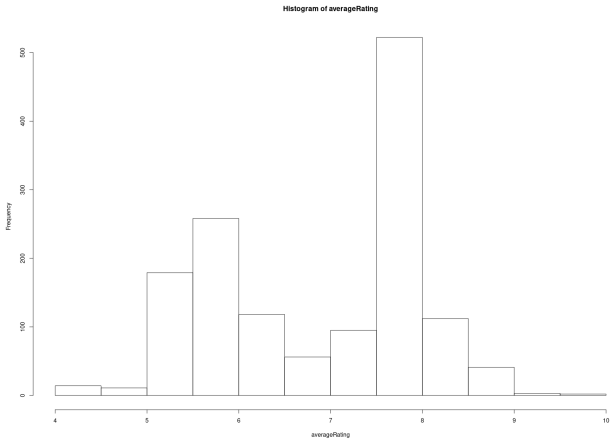

Figure 1: Histogram of Average Rating

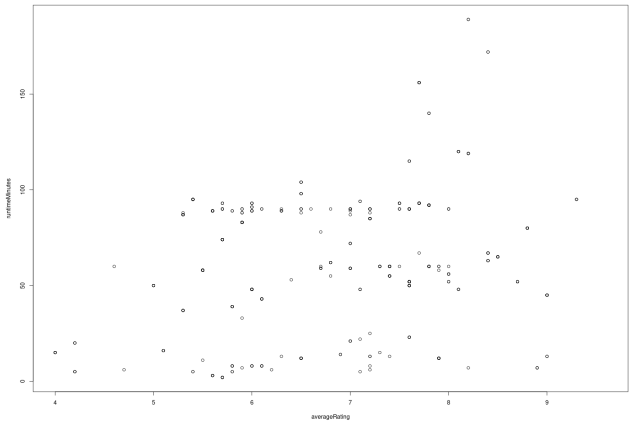

Figure 2: Average rating vs running time

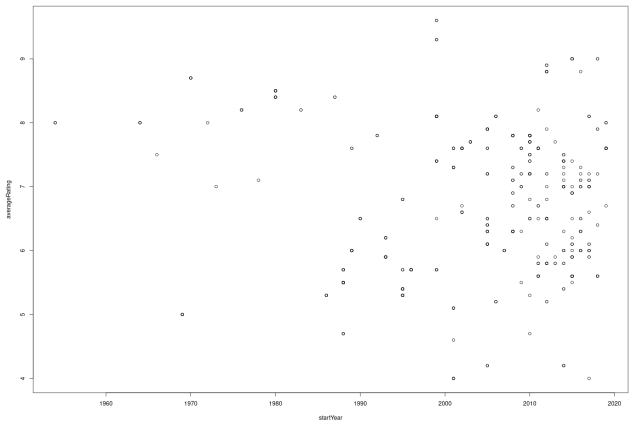

Figure 3: Release Year mapped against Average Rating

We have been unable to narrow down to one particular data mining idea so we instead propose two with the hopes that you can help guide us into which is a more interesting approach. Our first idea is being able to predict the rating of a movie based on the actors, directors, and producers of the film. The second idea would be to determine if a movie is likely to be localized to a different region than it was original made for by looking at the actors, producers, directors, and rating of the film. We also figure that for either of these proposed data mining approaches, rather than looking at the names of these people, which is a hard feature to quantify on, instead we could use the number of entertainment products which they have been a part of as the data feature. Another possibility for figuring out features would be to quantify how many highly rated entertainment products each has been a part of as that would likely be a good indication if the piece of entertainment would end up being localized. We believe either these approaches would be considered as data mining approaches because they're

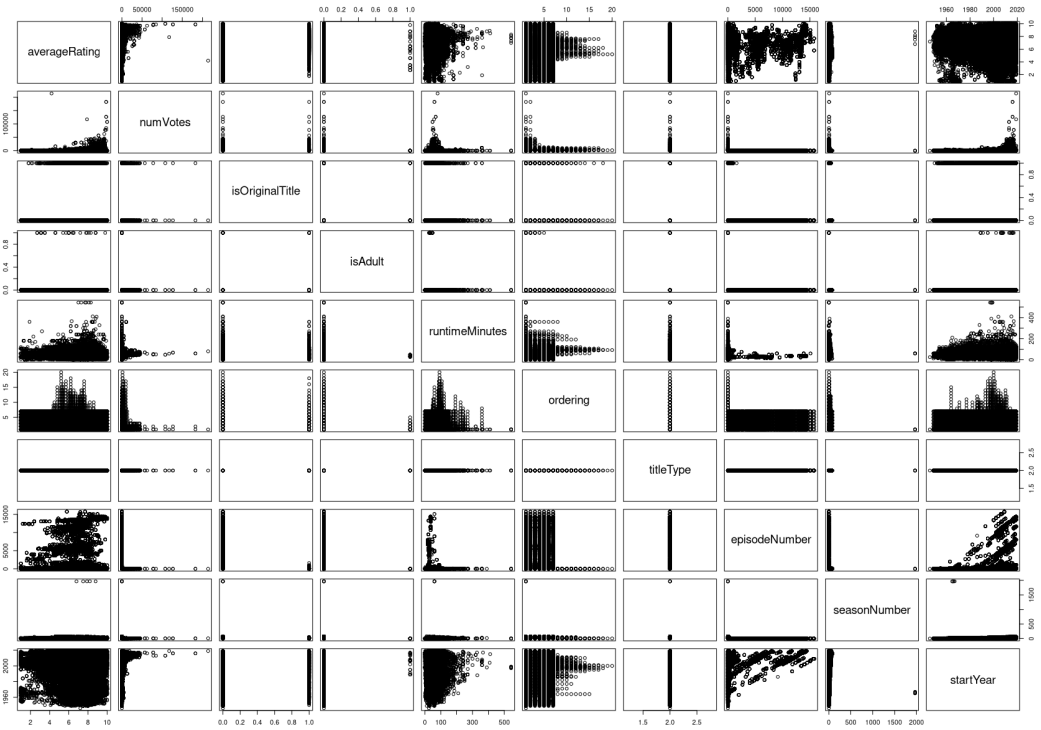

Figure 4: Pairs of comparison

```
   tconst             titleType  primaryTitle        originalTitle          isAdult
Length:1411        Min.   :2    Length:1411        Length:1411        Min.   :0.00000
Class :character   1st Qu.:2    Class :character   Class :character   1st Qu.:0.00000
Mode  :character   Median :2    Mode  :character   Mode  :character   Median :0.00000
                   Mean   :2                                         Mean   :0.02126
                   3rd Qu.:2                                         3rd Qu.:0.00000
                   Max.   :2                                         Max.   :1.00000

 startYear       endYear      runtimeMinutes       genres             ordering
Min.   :1954   Min.   : NA    Min.   :  2.00   Length:1411        Min.   : 1.000
1st Qu.:1996   1st Qu.: NA    1st Qu.: 52.00   Class :character   1st Qu.: 1.000
Median :2008   Median : NA    Median : 60.00   Mode  :character   Median : 2.000
Mean   :2004   Mean   :NaN    Mean   : 65.92                      Mean   : 2.896
3rd Qu.:2012   3rd Qu.: NA    3rd Qu.: 90.00                      3rd Qu.: 4.000
Max.   :2019   Max.   : NA    Max.   :189.00                      Max.   :11.000
               NA's   :1411   NA's   :216

   title              region            language           types
Length:1411        Length:1411        Length:1411        Length:1411
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character

 attributes        isOriginalTitle  averageRating     numVotes       parentTconst
Length:1411        Min.   :0.0000   Min.   :4.000   Min.   :   5.0   Length:1411
Class :character   1st Qu.:0.0000   1st Qu.:5.900   1st Qu.:  10.0   Class :character
Mode  :character   Median :0.0000   Median :7.400   Median :  46.0   Mode  :character
                   Mean   :0.1538   Mean   :6.904   Mean   : 321.7
                   3rd Qu.:0.0000   3rd Qu.:7.700   3rd Qu.: 272.0
                   Max.   :1.0000   Max.   :9.600   Max.   :2223.0

seasonNumber     episodeNumber       nconst           primaryName
Min.   : 1.000   Min.   : 0.000   Length:1411        Length:1411
1st Qu.: 1.000   1st Qu.: 1.000   Class :character   Class :character
Median : 1.000   Median : 3.000   Mode  :character   Mode  :character
Mean   : 2.053   Mean   : 6.395
3rd Qu.: 1.000   3rd Qu.: 5.000
Max.   :52.000   Max.   :65.000
NA's   :725      NA's   :725

 birthYear       deathYear     primaryProfession
Min.   :1931   Min.   :1970   Length:1411
1st Qu.:1937   1st Qu.:1970   Class :character
Median :1937   Median :1996   Mode  :character
Mean   :1942   Mean   :1987
3rd Qu.:1937   3rd Qu.:1996
Max.   :1974   Max.   :1996
NA's   :1402   NA's   :1402
```

Figure 5: Data Summary