MovieLens Recommender System

Team Neo

EBAC 4 (Mixed Group), Institute of Systems Science, NUS

[Anusuya Manickavasagam: A0163300Y,

Chinnasubbareddygari Mohan Reddy: A0163433L ,

Gello Mark Vito: A0163448Y,

Indumathi Nagarajan: A0163441M,

Muni: A0163382E, Pradeep: A0163453H]

Contents

[1 Executive Summary 2](#_Toc498291486)

[1.1 Data Selection and Preprocessing 2](#_Toc498291487)

[1.2 u.item Preprocessing (Raw Movie Data) 3](#_Toc498291488)

[1.3 u.user Preprocessing (Raw User Data) 5](#_Toc498291489)

[1.4 u.data Preprocessing (Raw Ratings Data) 5](#_Toc498291490)

[1.5 Data Partition/High Level Observation 6](#_Toc498291491)

[1.6 Model Building 7](#_Toc498291492)

[1.6.1 Apriori (Market Basket Analysis) 7](#_Toc498291493)

[1.6.2 User Demographics 10](#_Toc498291494)

[1.6.3 Genre Demographics 11](#_Toc498291495)

[1.7 Model Performance 12](#_Toc498291496)

[1.8 Movie Recommendation System Application 13](#_Toc498291497)

[1.9 Recommendations 13](#_Toc498291498)

**Abstract**

This study aims to build a movie recommender system for MovieLens. Purpose of which is to increase revenue and improve user friendliness by suggesting movies which the users are likely to watch. The proponents of this study used the association algorithm through the ratings that users made. Additionally, factors such as the users’ demography and the movies’ genre in which a certain group of audience are in were also taken into consideration.

# Executive Summary

Objectives of this study is to seek recommendations based on association of user ratings on movies. User demography and movie genre were used to aid in building the association algorithm.

After data cleaning and preparation of dataset variables, the following model is built and cross validated:

Model: Apriori (Market Basket Analysis)

Tools: R, Excel

Using the ratings dataset, the model is initially constructed by user id and movie id columns. This resulted in **941 rules** *with (support:0.2, confidence:0.8)*. Rules were filtered after taking a minimum lift of 2.1. By running the rules against the Test Dataset, model produced **82.26% accuracy**. About 4 out of 5 recommended movies will be likely watched by the user. It can be observed that the relevant rules generated by this model have strong relationship with movies in series. The prominent movies for this model are: Raiders of the Lost Ark (1981), Alien (1979), Star Wars (1977) etc.

Next, the user demographics are taken into consideration when invoking the apriori() function. This time, movies, user age, user gender, and user occupation were merged in the same dataset and is transformed into a transaction object which are then used for the apriori function(). The generated rules did not give relevant weights on age and occupation but some rules portray that males are most likely the gender of users who have watched certain movies (e.g. Star Wars is preferred by Males). Moreover, it was also repeated that movies in series are most likely be watched and therefore a good candidate when doing movie recommendation.

Lastly, we also incorporated movie’s genres. After adding the Genre into basket of items, more rules ~ 4504362 rule(s)] are created with the same support (0.2) and confidence (0.8). In order to gather more insights, we still loosening the rule parameters to 0.8 support and 0.9 confidence, on which 7074 rule(s) were created with Genre. It looks like Children & Sci.Fi Genre have enough power to draw viewers.

The “confidence” tells us that, if we take again the first record, 98% of the users who watched the Children & Sci.Fi movies, they also watched Adventure Genre too.

## Data Selection and Preprocessing

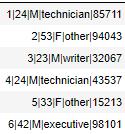
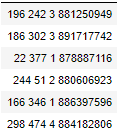
The file consisting of 100K records were used in this study. There were three datasets used:

1. u.item – contains list of movies; its id and its genre
2. u.user – contains list of users as well as the user’s demography and other characteristics
3. u.data – contains ratings of users through the combination of the above datasets

Below is the raw data from the datasets:



*Figure SEQ Figure \\* ARABIC 1 Movie: Raw*

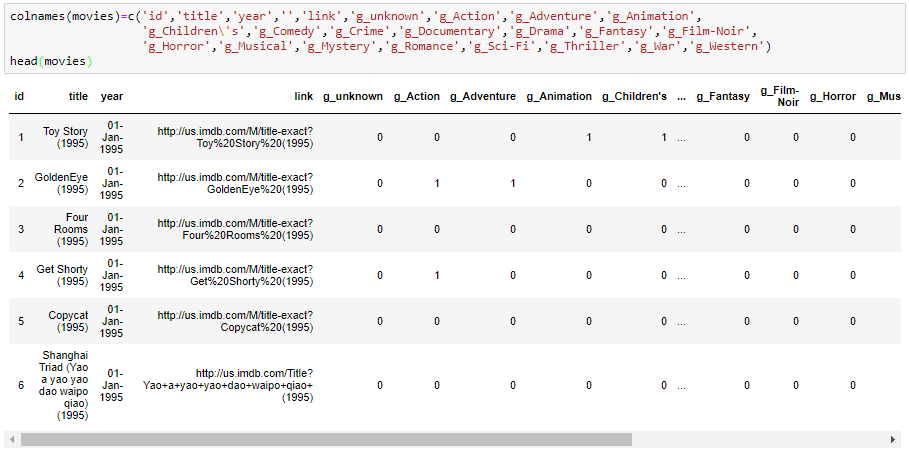


*Figure SEQ Figure \\* ARABIC 2 User: Raw*

*Figure SEQ Figure \\* ARABIC 3 Rating: Raw*

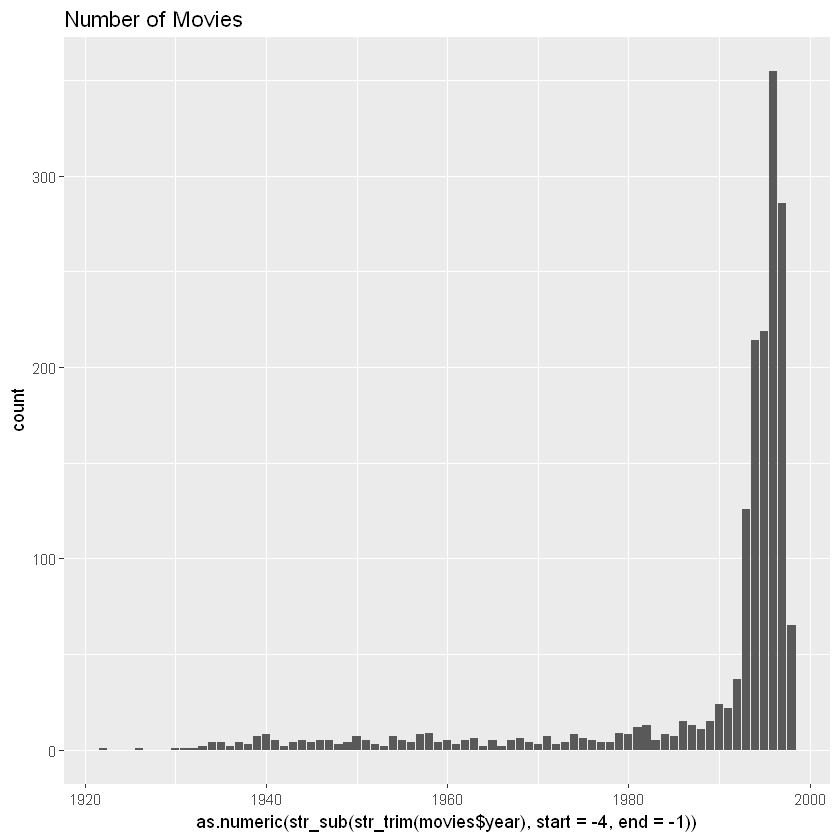
## u.item Preprocessing (Raw Movie Data)

The data is loaded and parsed using the pipeline as delimiter and assigned the columns respectively. For the genre column, the descriptions are taken from the u.genre file which consists of 19 genres.



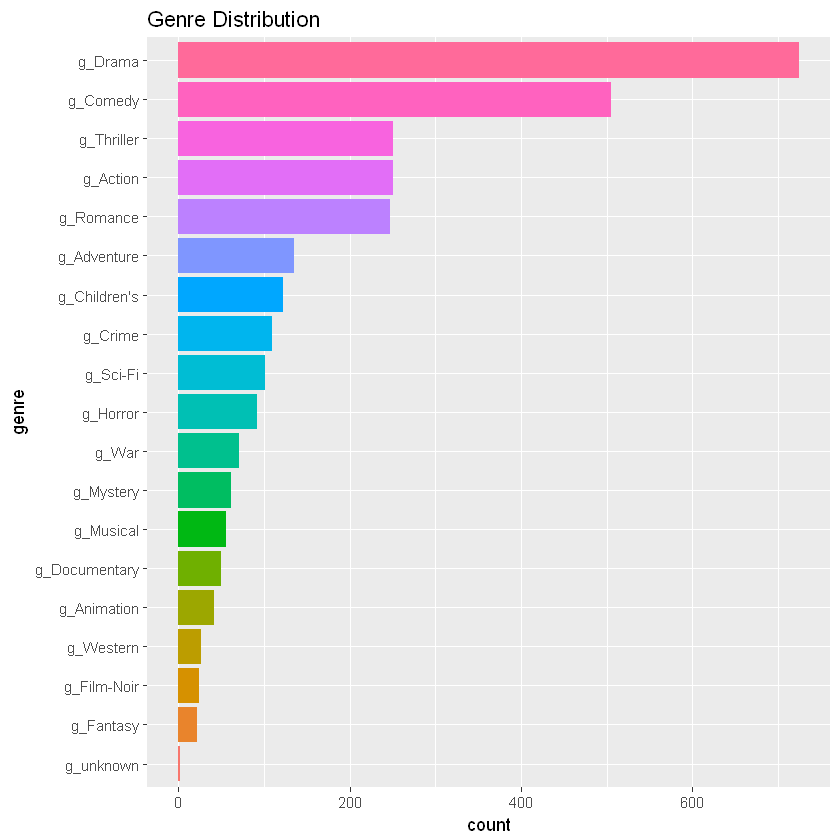
*Figure 4 Movie: Processed Data*

The plot below shows the movie count as year progresses. From the illustration below, we can see that as time pass by, more movies are being produced:



*Figure 5 Movie Count per Year*

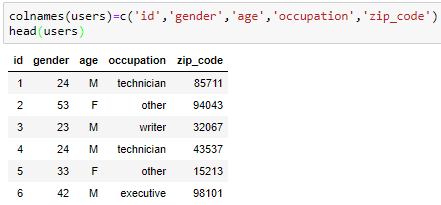
Plotting the movie genre, it can also be observed that Drama and Comedy are the most watched genres:



*Figure 6 Genre Distribution*

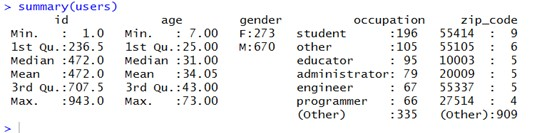
## u.user Preprocessing (Raw User Data)

The user data is loaded and parsed using tab as delimiter and assigned the columns respectively.



*Figure 7 User Data*

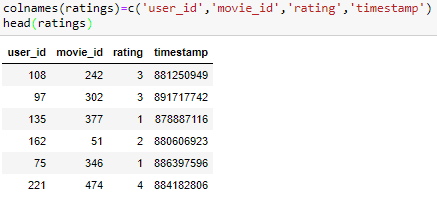
By looking at the summary, we can observer about 2/3rd of the users are Male. This ratio still looks fine and may not add unnecessary bias to our association rules. So we went ahead with this. Had it been like 90% of Male users, we might have faced data unbalancing problems.



*Figure 8 User Data Summary*

## u.data Preprocessing (Raw Ratings Data)

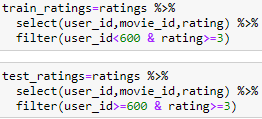
The data is loaded and parse using the pipeline as delimiter like the movies dataset. Column names are also assigned respectively:



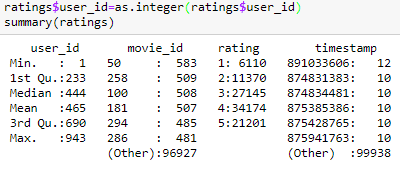
*Figure 9 Ratings Data*

## Data Partition/High Level Observation

The data was divided into training and test datasets. Using below screenshot, we can see that there were 943 users in the ratings table. We split the dataset into 70-30 proportion and used records having ratings above 3.



*Figure 10 Training and Test Datasets*



*Figure 11 Ratings Data Summary*

This also shows that the top 5 most rated movies are:

1. 50 – Star Wars (1977)
2. 258 – Contact (1997)
3. 100 – What Happened Was... (1994)
4. 181 – Low Down Dirty Shame, A (1994)
5. 294 – Ayn Rand: A Sense of Life (1997)
6. 286 – English Patient, The (1996)

All these are produced during 1994-1997.

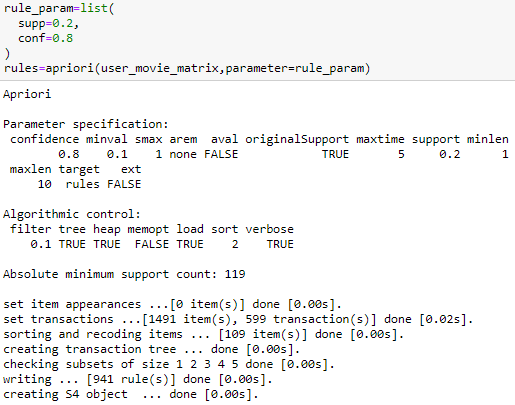
## Model Building

### Apriori (Market Basket Analysis)

Since apriori() accepts a dataset of type ‘transaction’, the training dataset has to be converted into a transaction object. This is done by saving the training dataset into a csv file and loading the data using the read.transactions() function.

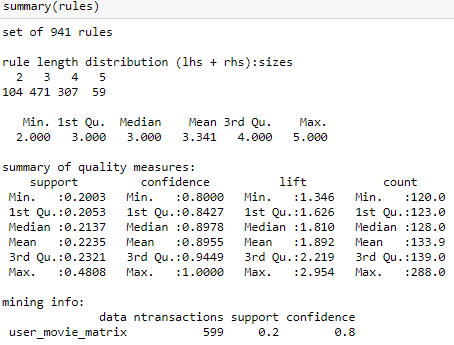
arules R package is used for this model.

We ran different combinations of support and confidence while using apriori() function. The function with supp=0.2, and conf=0.8 worked best and we have included the same below:



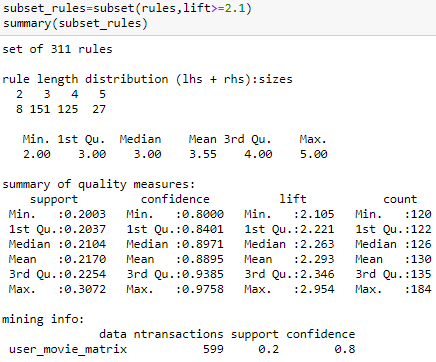
*Figure 12 apriori()*

80% confidence is used to filter low quality rules and limit the results, this gave **941 rules**.



*Figure 13 apriori() Rules Summary*

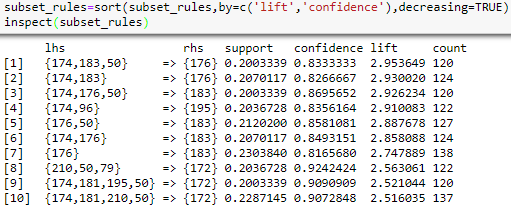
A rule with high confidence doesn’t equate to a good rule. It might be possible that a rule with 100% confidence is already applicable to all users, that is, a pair of movie is already watched by all users so there is nothing to recommend even if the movie pair rule has a 100% confidence. Considering this, we had to further filter the rules though the property ‘lift’ which signifies the independence between movies in a pair. It can be seen above that 3rd quartile of lift has a value of 2.219, so we tried filtering the rules using lift>=2.1 as baseline.



*Figure 14 Final Rules*

From 941 rules, it was further filtered to **311 relevant rules**.

Peeking at the top 10 most relevant rules sorted by lift and confidence, we can see below entries:



*Figure 15 Most Relevant Rules*

The top 10 most relevant rules are displayed below by matching with the data with the movie dataset. Here are the 10 most relevant entries from the rules:

1. Raiders of the Lost Ark (1981), Alien (1979), Star Wars (1977)

Recommendation: Aliens (1986)

1. Raiders of the Lost Ark (1981), Alien (1979)

Recommendation: Aliens (1986)

1. Raiders of the Lost Ark (1981), Aliens (1986), Star Wars (1977)

Recommendation: Alien (1979)

1. Raiders of the Lost Ark (1981), Terminator 2: Judgment Day (1991)

Recommendation: Terminator, The (1984)

1. Aliens (1986), Star Wars (1977)

Recommendation: Alien (1979)

1. Raiders of the Lost Ark (1981), Aliens (1986)

Recommendation: Alien (1979)

1. Aliens (1986)

Recommendation: Alien (1979)

1. Indiana Jones and the Last Crusade (1989), Star Wars (1977), Fugitive, The (1993)

Recommendation: Empire Strikes Back, The (1980)

1. Raiders of the Lost Ark (1981), Return of the Jedi (1983), Terminator, The (1984), Star Wars (1977)

Recommendation: Empire Strikes Back, The (1980)

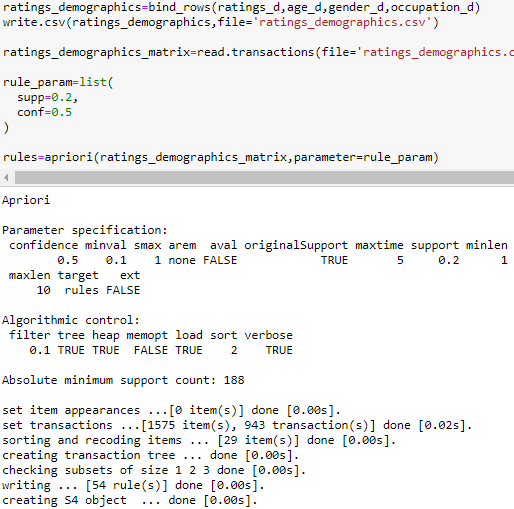
1. Raiders of the Lost Ark (1981), Return of the Jedi (1983), Indiana Jones and the Last Crusade (1989), Star Wars (1977)

Recommendation: Empire Strikes Back, The (1980)

Interestingly, a pattern here can be seen for movies with sequels like Aliens and Star Wars (which makes sense as people will probably watch the next sequel movie if they have started watching the first movie).

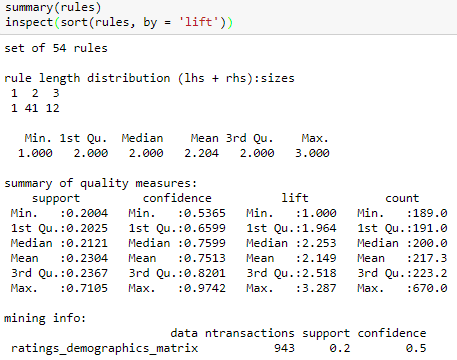
### User Demographics

We added the demographics into picture by having the movies, user age, user gender, and user occupation merged in a single data frame. By loosening the rule parameters to gather more insights, there were 54 rules as a result.



*Figure 16 Rules with Demographics*

Next, we’ll try to inspect these rules and observe the insights it tells us.



*Figure 17 Demographic Rules Summary*

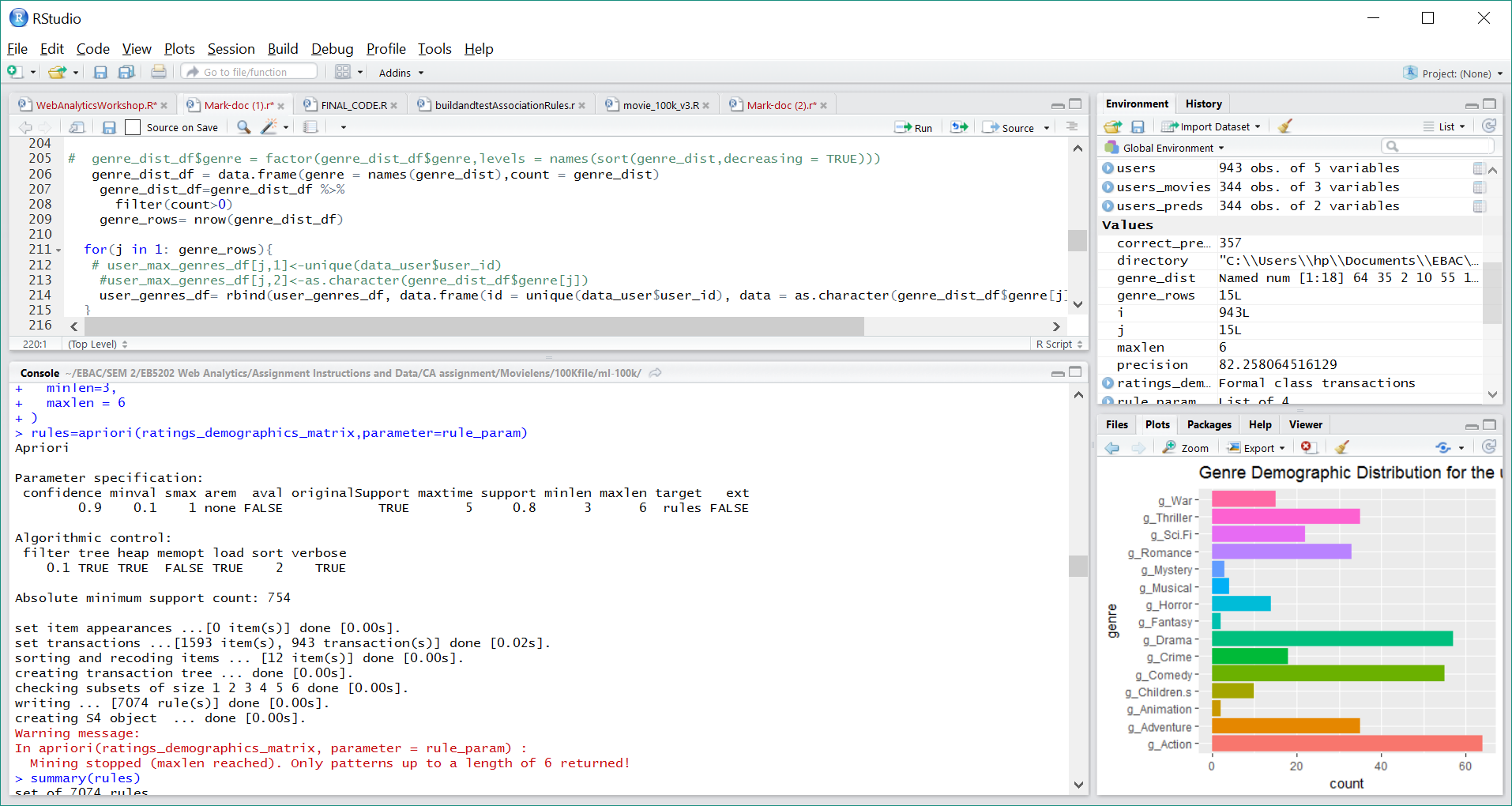
Aside from the usual user-movie recommendations, the resulting 54 rules give these insights:

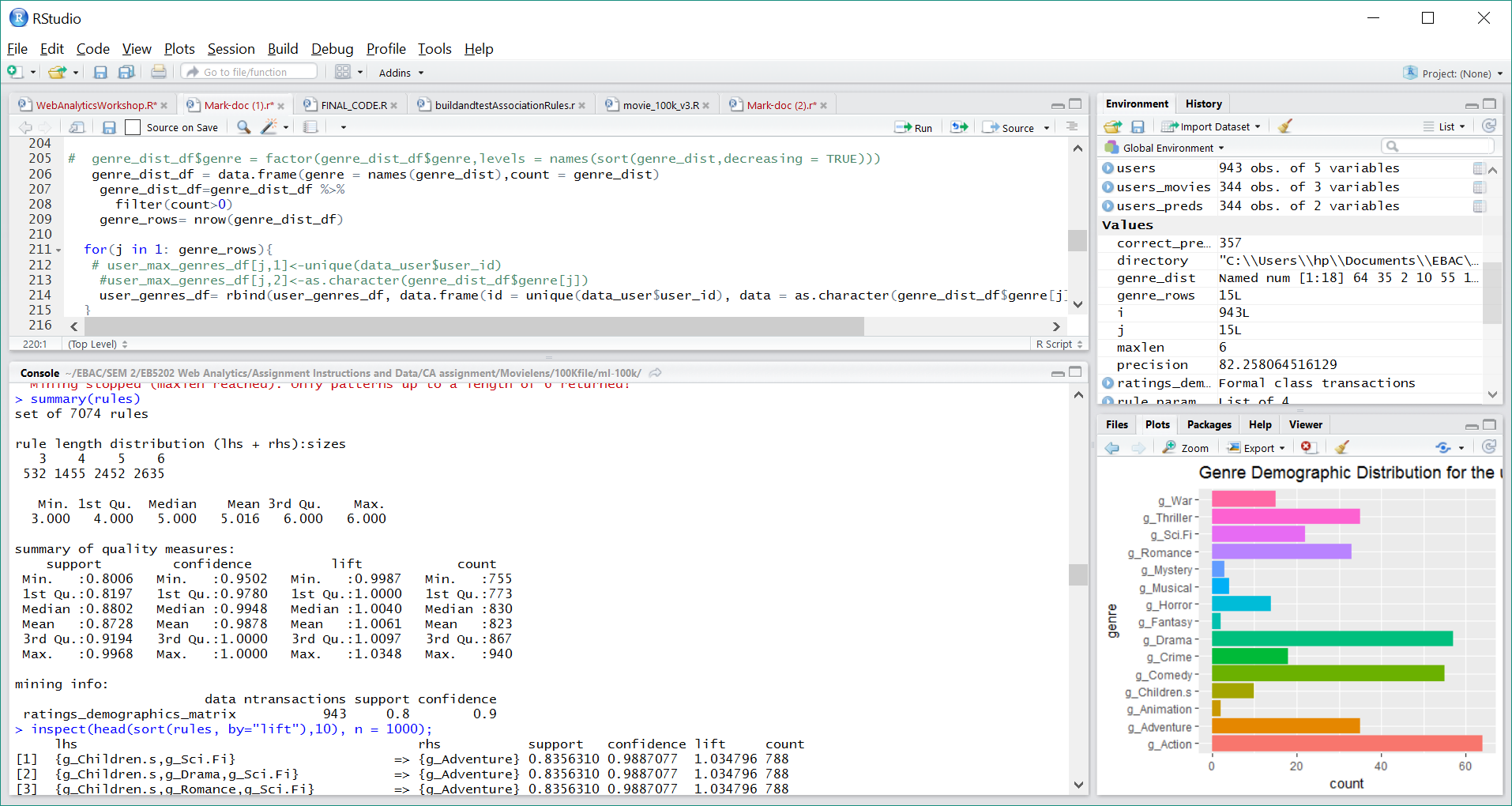
* If someone is a male and have watched *Star Wars (1977)*, the person will be likely to watch *Return of the Jedi (1983)*.
* If the user watched the following movies, it must be likely that the user is a Male:
* *Raiders of the Lost Ark (1981)*
* *Return of the Jedi (1983)*
* *Star Wars (1977)*
* *Pulp Fiction (1994)*
* *Contact (1997)*
* *Fargo (1996)*
* *Toy Story (1995)*

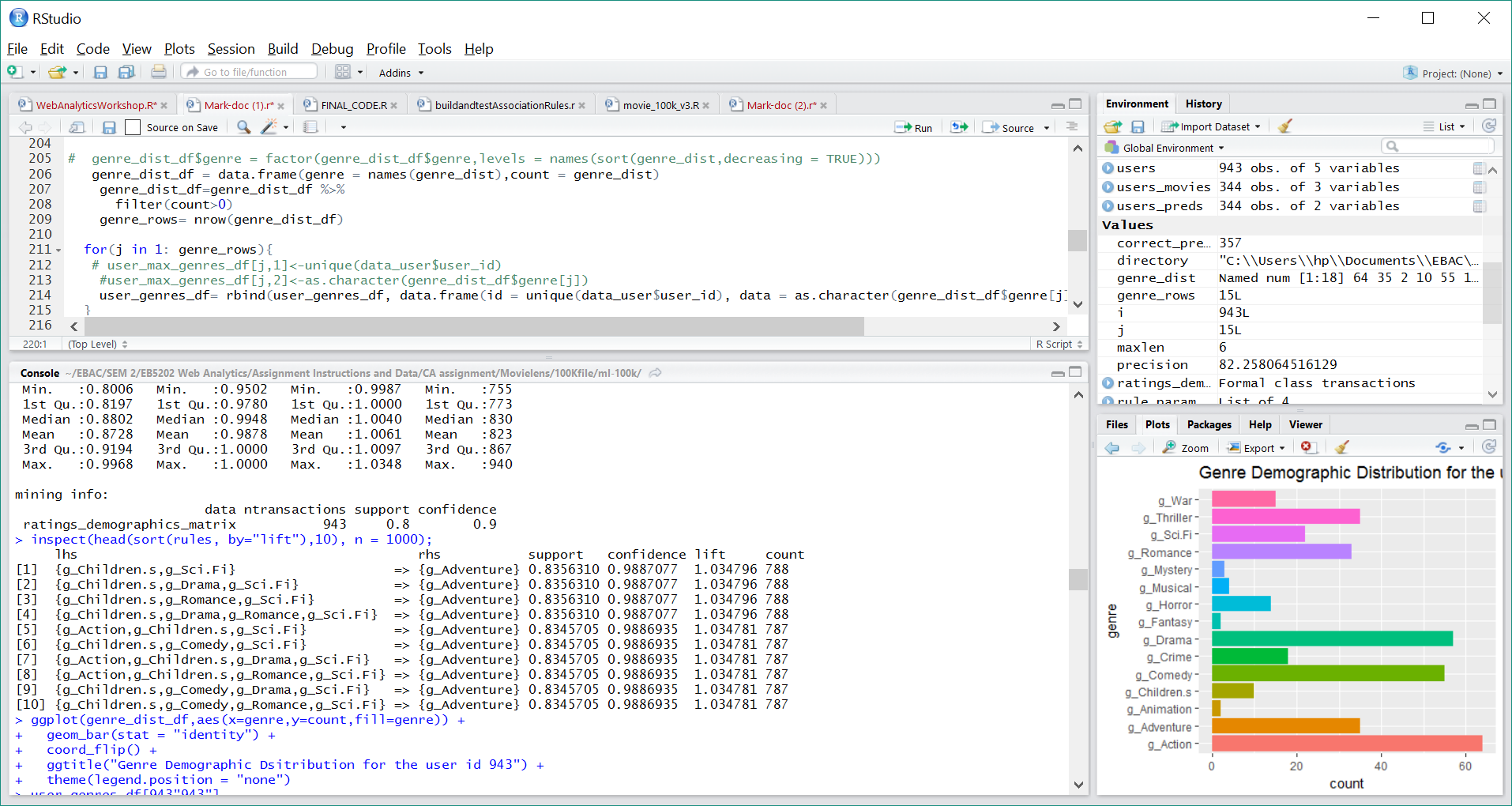
### Genre Demographics

We also incorporated movie’s genres. After adding the Genre into basket of items, more rules ~ 4504362 rule(s)] are created with the same **support (**0.2) and confidence(0.8). In order to gather more **insights,** we still loosening the rule parameters to 0.8 support and 0.9 confidence, on which 7074 rule(s) were created with Genre.

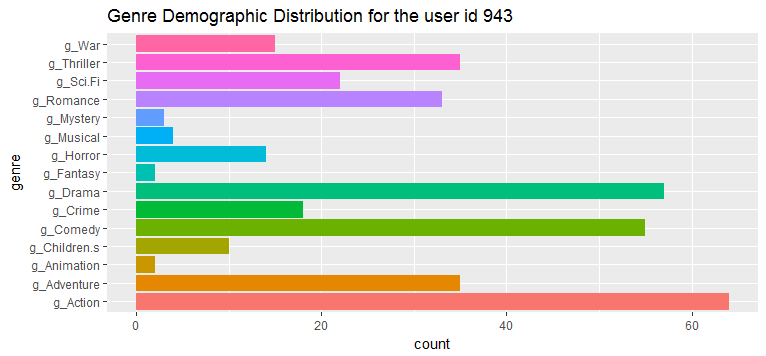
* It looks like Children & Sci.Fi Genre have enough power to draw viewers.
* The “confidence” tells us that, if we take again the first record, 98% of the users who watched the Children & Sci.Fi movies, they also watched Adventure Genre too.





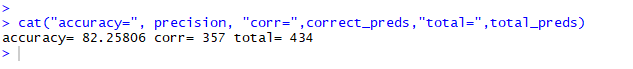


*Figure 18 Genre Demographic Rules Distribution*

 *Figure 19 Genre Demographic Distribution Plot*

## Model Performance

82.26% accuracy was achieved by the model. The figures are displayed below by validating on the test dataset:



*Figure 20 Model Precision*

Total Prediction – the recommended movies for each user are extracted and counted overall.

Correct Prediction – by comparing the actual movies watched by users in test dataset, we’ve matched all those predicted movies against the actual movies and counted them.

The model looks promising to deploy with an accuracy of predicting about 4 correct movies out of 5 recommendations. This accuracy can further be improved by incorporating the insights and findings for the rules that were generated after putting the user demographics and movie genre into picture.

## Movie Recommendation System Application

* Part one movies of sequels can be recommended. It is evident based on data that users who’ve watched these types of movies are likely to watch the succeeding movies. Some of these movies are Star Trek and Die Hard.
* If a user is a male, movies below can be recommended:
  + Raiders of the Lost Ark (1981)
  + Return of the Jedi (1983)
  + Star Wars (1977)
  + Pulp Fiction (1994)
  + Contact (1997)
  + Fargo (1996)
  + Toy Story (1995)
* Most rated movies are in the 1990’s to 2000’s. These movies can be highly recommended for especially for new users having born and grew up during these years.
* The RHS side data from the generated rules can serve as guide on what movies to recommend based on the movies the users already watched.

## Recommendations

* The study used 100K dataset, it might be worthwhile to try the 1M dataset.
* The rules with a lift higher than 1 are the one of interest. The higher value, the higher the correlation.
* To reduce the higher number of rules are being generated, may have to adjust the parameters of the algorithm, especially with support. With 100k Data set , we observed that the Rules are extremely minimized if we go for Support with 0.2 and more. Lesser than 0.2 is impacting the memory performance to generate rules.
* To run the different rules, we observed that they are CPU intensive tasks and probably using a distributed system will be helpful in running the predictions for those rules.
* Item Basket Including User Demographic, Movie Genre Demographics:

