Northeastern University
College of Engineering

Movie Recommendation System

Min Gong
Wei Chieh Wang

(857-350-5361)
(857-930-3520)

gong.mi@husky.neu.edu
wang.weichi@husky.neu.edu

Contents

1. Problem Setting

A recommendation system that provides suggestions to users which based on their own preference. The main goal of this project is to build a recommendation engine that recommends movies to users. We will be developing an Item Based Collaborative Filter.

2. Project Definition

1. Sort out the raw data.

2. Determine the top 10 movies in each decade

3. Determine the top 10 movies in each genre.

4. Build a model for new users.

5. Build a model for old users.

3. Data Description

We have two files for this project.

The file **movies_1.csv** contains information about moiveId, title, year (the years it was released), genres(animation, adventure, action…) for 10327 movies.

The file **ratings.csv** contains information about userID, movieId (the movie that was rated by the user), ratings(the user give to the movie) for 668 users.

4. Data sources

We download the data from this website: Data flair blog.

https://data-flair.training/blogs/data-science-r-movie-recommendation/

It is a website that include tutorial and data mining project.

5. Data Exploration

First, we want to make rankings of best movies in each decade and best movies in each genre. This could do a help for us to understand raw data and be a recourse for us to recommend specific movies to new users in our following model.

Ranking criteria:

1. For annually best, we set a threshold 20%: only the movies which receive over 20% ratings from users can be ranked, for genre's best, we set it as 5%;

2. For candidates over the threshold, we take average rating scores as the measures to rank them.

Following are some figures for Best movies in decades and genres

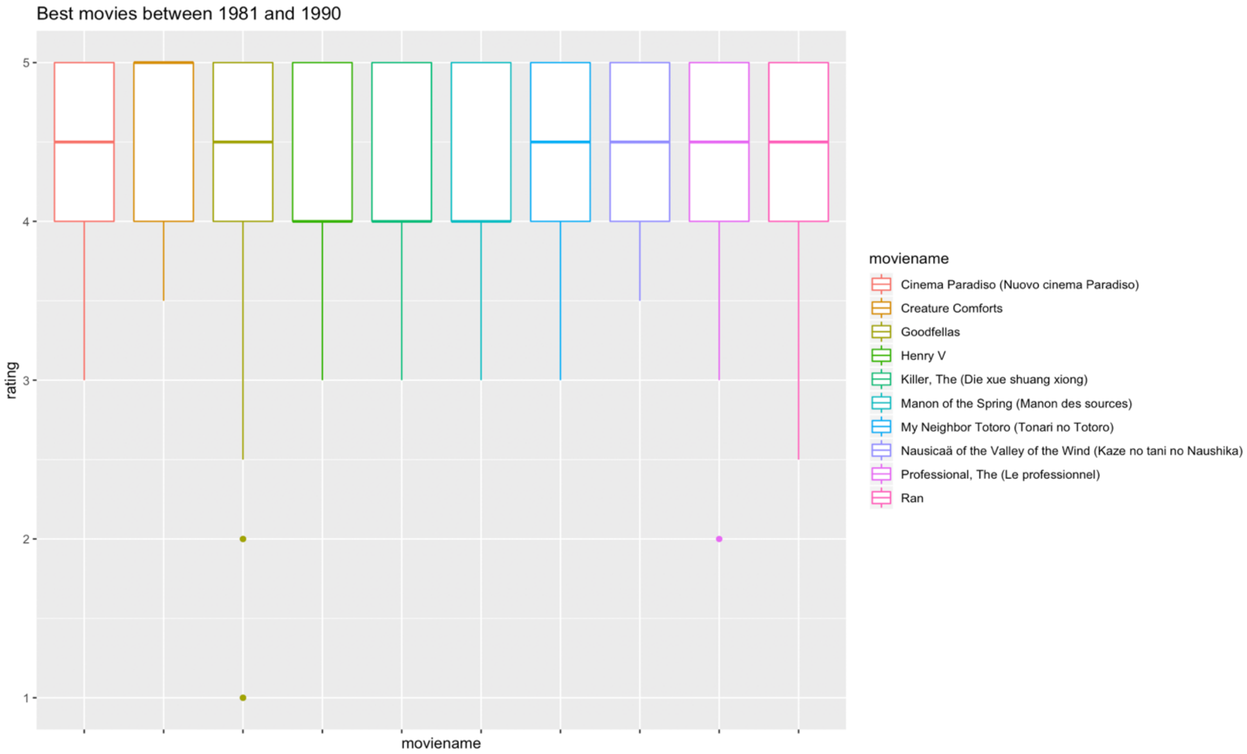


Figure 1. Best Movies in 80's

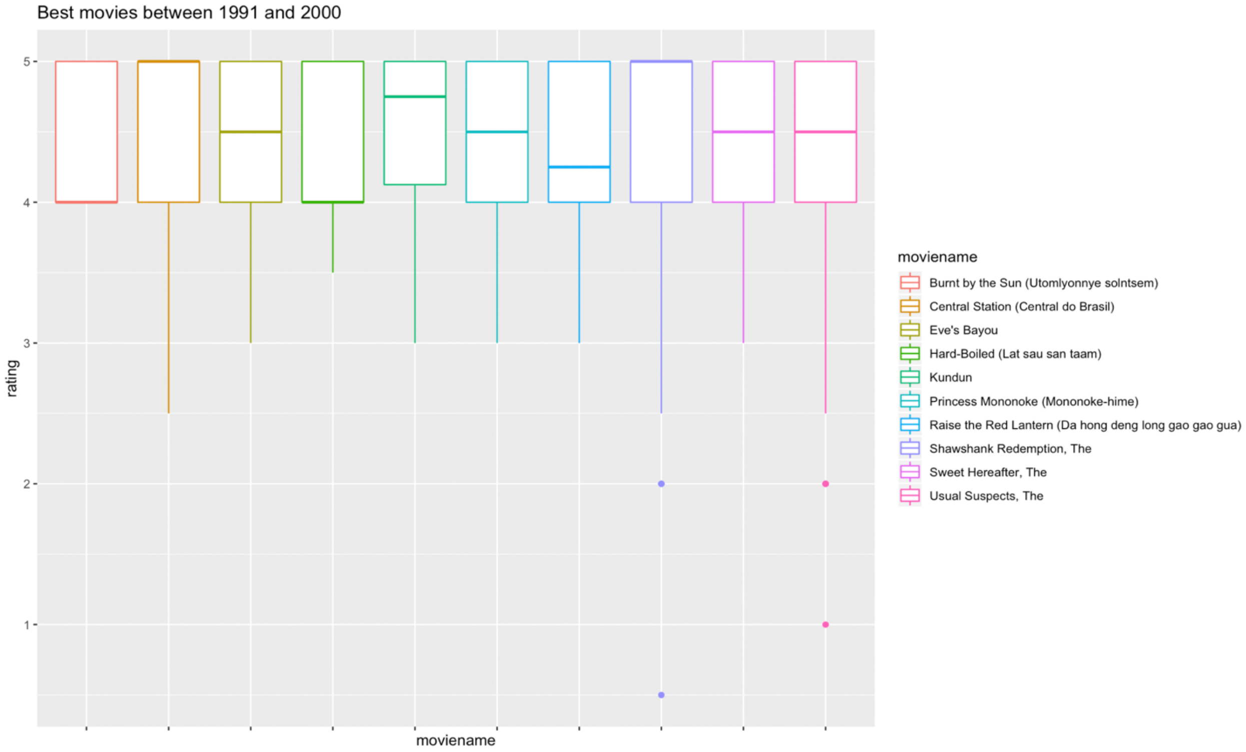


Figure 2. Best movies in 90's

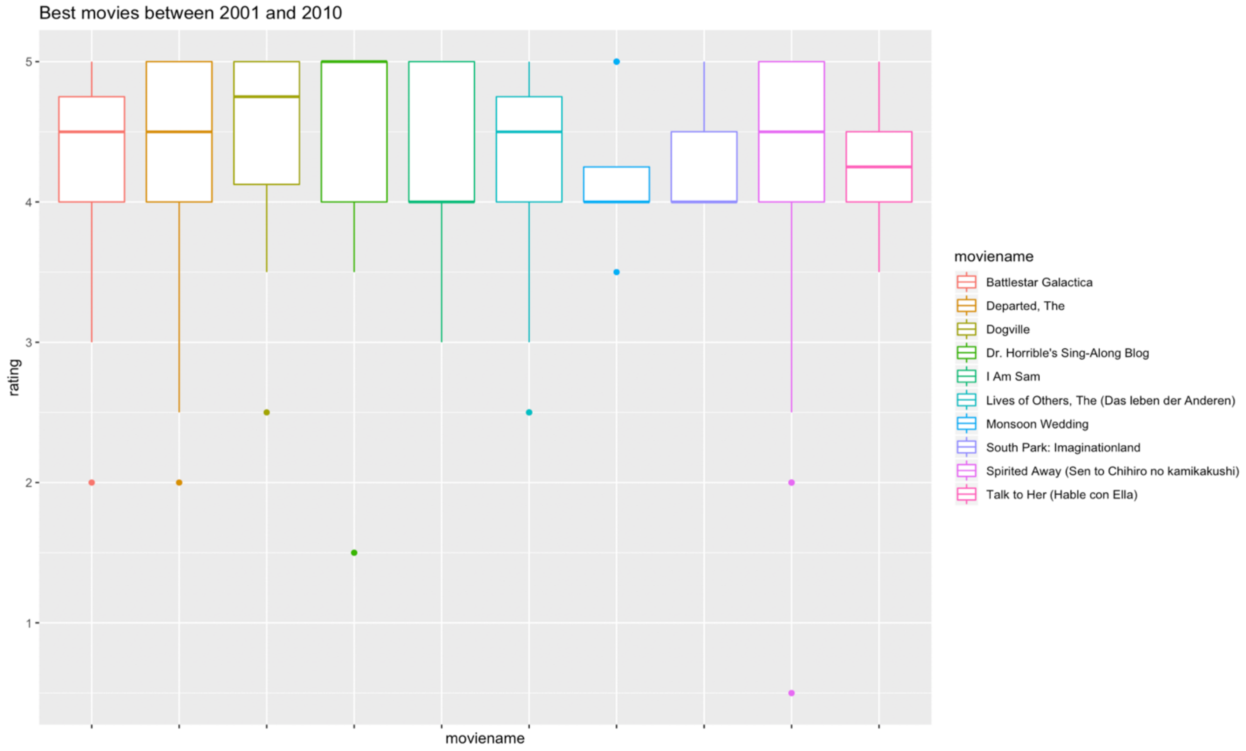


Figure 3. Best movies in 90's

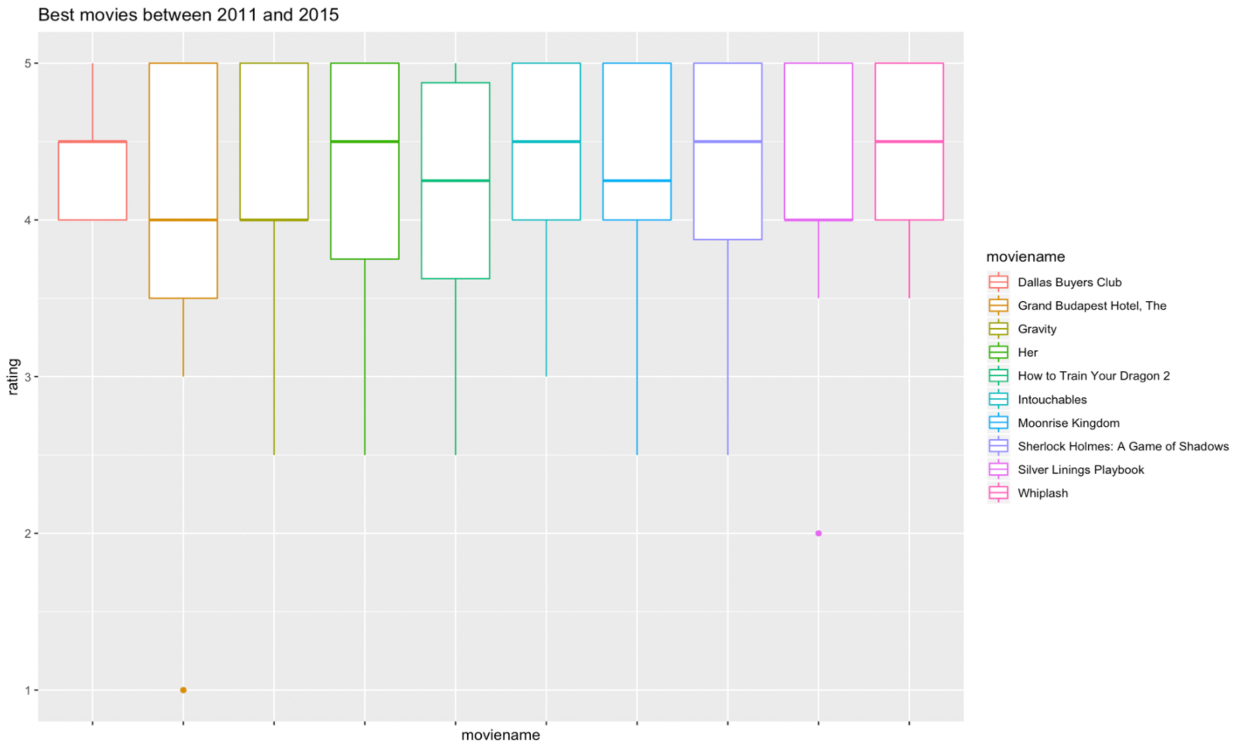


Figure 4. Best movies in 2011-2015

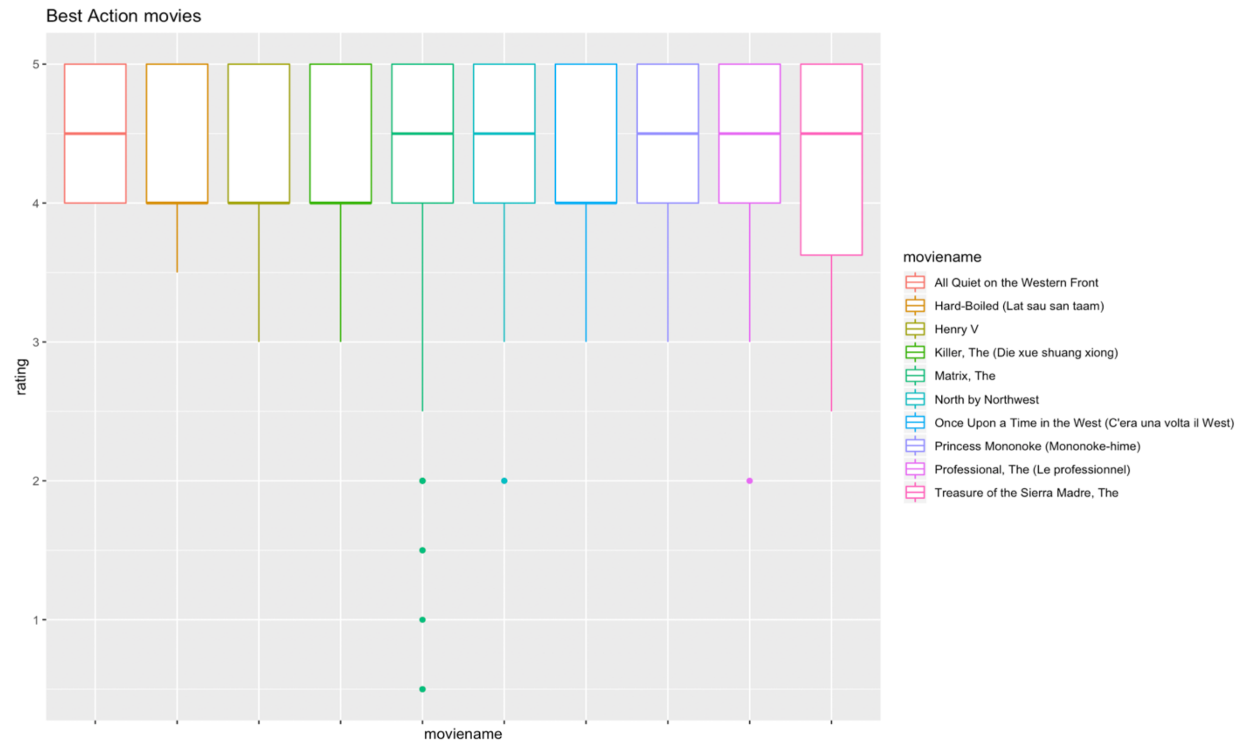


Figure 5. Top 10 in Action movies

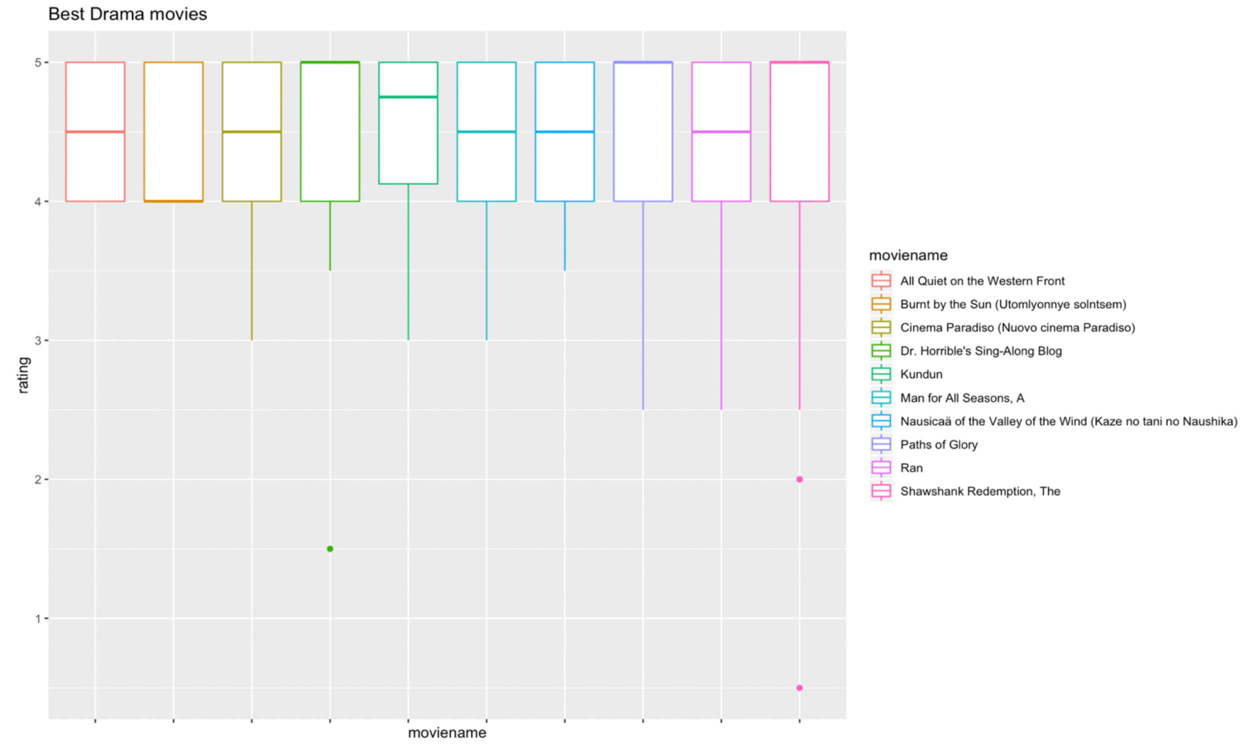


Figure 6. Top 10 in Drama movies

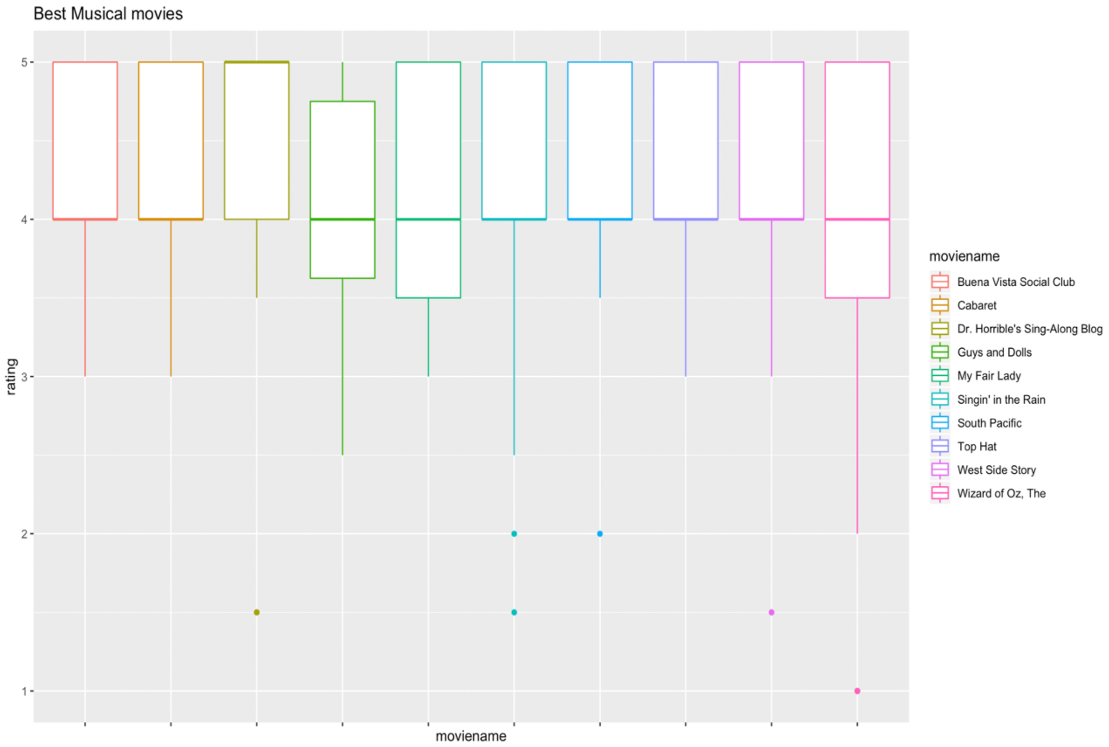


Figure 7. Top 10 in Musical movies

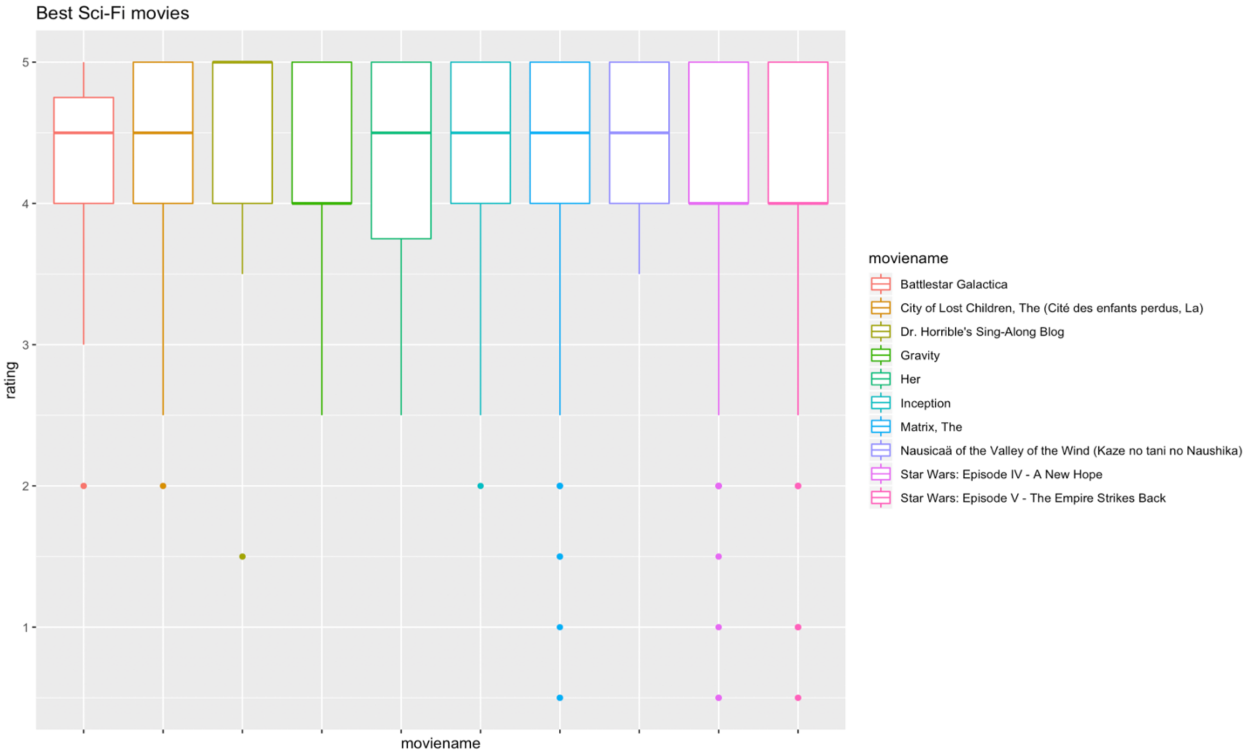


Figure 8. Top 10 in Si-Fi movies

We can take above information as reference recommendations for new users who don't have any rating history.

6. Data Mining Task

Data preprocess:

Data reduction: We delete a predictor “timestamp” due to it’s useless.

Data transformation: For the data ratings, we use 3 variables for each genre to denote a user's watching history: number of movies watched, mean and mode of the ratings, mode of years of the movies.

We separate our customers to new and old.

For the new users, which have never watched movies on our platform before, we use association rules(item-based) to deal with the data.

For the old users, which have watched movies on our platform for more than 5 movies, we use collaborative filtering (user-based) to deal with the data.

7. Data mining model

Recommendation system for new users

We will ask new users to input his/her preferred genres and years (we offer options).

After user input the genres and year, he/she will obtain an output, 5 recommended movies. (Priority: genres > years).

For example, the user input years for 2010, genres for Action and we will recommend him the top 5 actions movies in 2010; if there is not good action movies in 2010, we will give him/her an action movie 2009 or 2011 rather than some other type of movies in 2010.

Checking the intersections of genres and years selected, return top 5 movies with highest ratings, if no enough movies in the intersection, relax years to +- 1 year; if still no, +- 2 years.

Recommendation system for old users

We use 3 variables for each genre to denote a user's watching history: number of movies watched, mean and mode of the ratings, mode of years of the movies.

For example, in genre Action movie.

1. First, we count the number of action movies that the user watched.
2. Second, we count the mean or the mode of the ratings that he gave action movies.
3. Third, we count the mode of years of the action movies that he watched.
4. After obtaining the information above, we need to do the normalization because the range of our variable is different.

We repeat these steps for all genres and we will get a data frame which has 55 columns, that we call it user_info_norm.

We take this data frame to compute distance (opposite of similarity, we use RMSE to denote it) of each pair of users.

First, we plan to find top 3 similar users according the distance between one specific user.

Second, we want to find all movies the user hasn't watched but his/her neighbors have and, we will recommend him/her these movies.

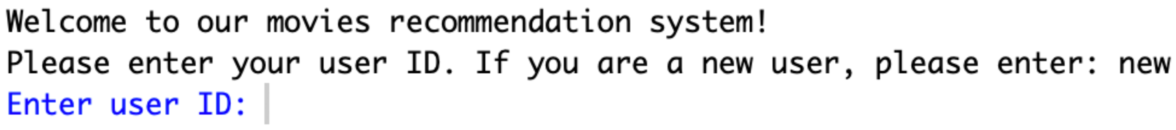
8. Performance Evaluation

Because our project is unsupervised learning, unfortunately, it does not like supervised learning model which has validation data set to evaluate the model. The only one way that we can test our model is inviting our classmates/friends to use our recommendation system and then we can know whether our model work or not.

9. Result of Project

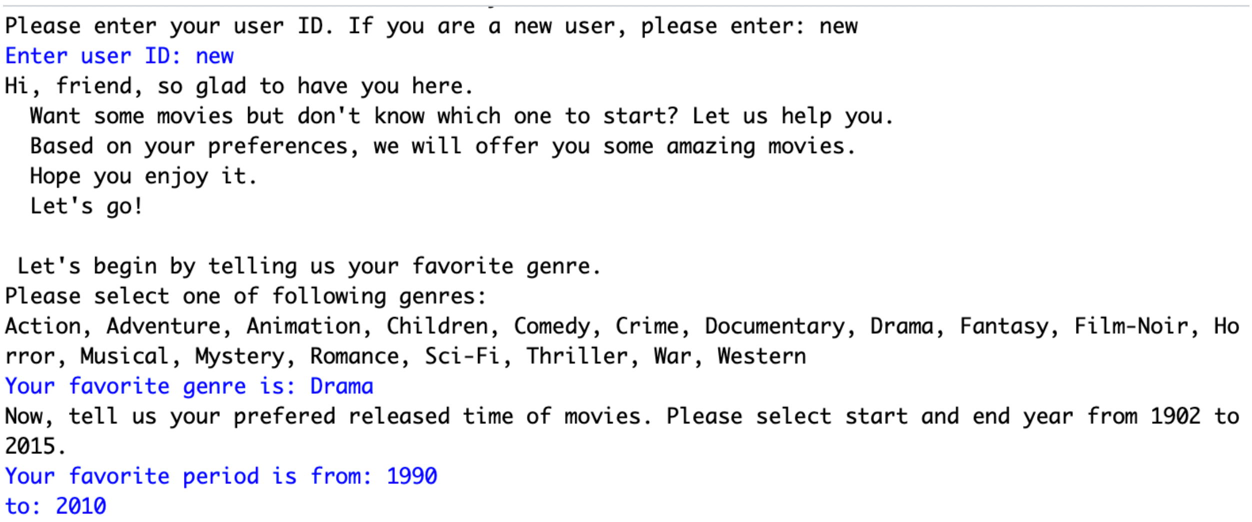
We made two models for new and old users successfully last.

For new users, they input genre and year and then they will get recommended movies.



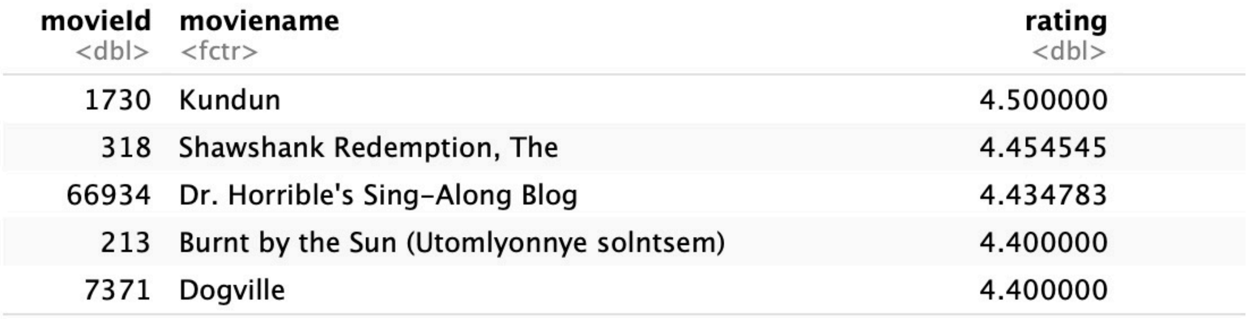
```
Welcome to our movies recommendation system!
Please enter your user ID. If you are a new user, please enter: new
Enter user ID:
```

Figure 9. movie recommended system process for new user



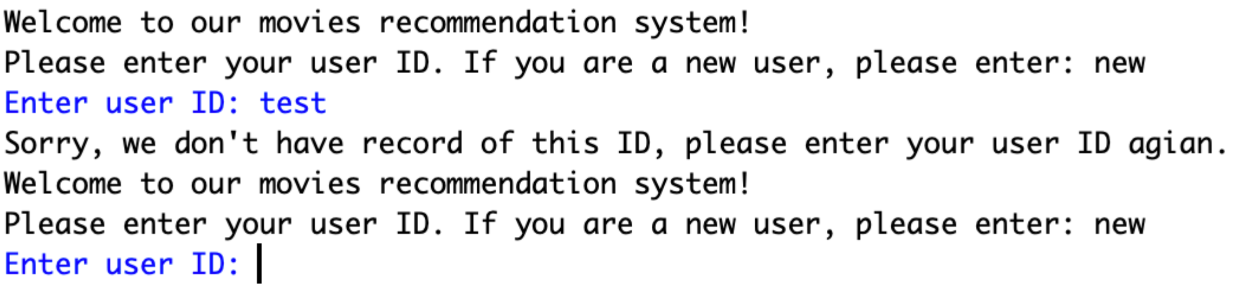
```
Please enter your user ID. If you are a new user, please enter: new
Enter user ID: new
Hi, friend, so glad to have you here.
  Want some movies but don't know which one to start? Let us help you.
  Based on your preferences, we will offer you some amazing movies.
  Hope you enjoy it.
  Let's go!

 Let's begin by telling us your favorite genre.
Please select one of following genres:
Action, Adventure, Animation, Children, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Ho
rror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western
Your favorite genre is: Drama
Now, tell us your prefered released time of movies. Please select start and end year from 1902 to
2015.
Your favorite period is from: 1990
to: 2010
```

Figure 10. movie recommended system process for new user



movieId <dbl>	moviename <fctr>	rating <dbl>
1730	Kundun	4.500000
318	Shawshank Redemption, The	4.454545
66934	Dr. Horrible's Sing-Along Blog	4.434783
213	Burnt by the Sun (Utomlyonnye solntsem)	4.400000
7371	Dogville	4.400000

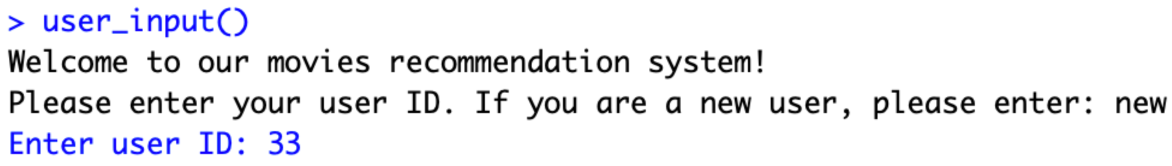
Figure 11. movie recommended system result for new user



```
Welcome to our movies recommendation system!
Please enter your user ID. If you are a new user, please enter: new
Enter user ID: test
Sorry, we don't have record of this ID, please enter your user ID agian.
Welcome to our movies recommendation system!
Please enter your user ID. If you are a new user, please enter: new
Enter user ID: |
```

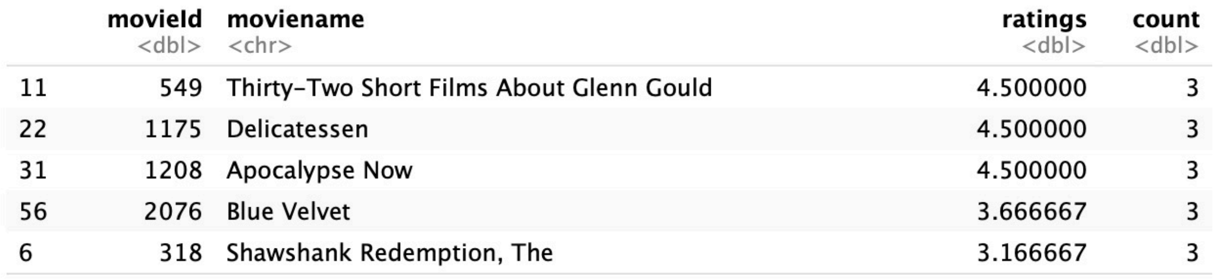
Figure12. movie recommended system process for non-new user and non-old user.

For old users, after they input their UserID, they will get recommended movies.



```
> user_input()
Welcome to our movies recommendation system!
Please enter your user ID. If you are a new user, please enter: new
Enter user ID: 33
```

Figure 13. movie recommended system process for old user



	movieId	moviename	ratings	count
	<dbl>	<chr>	<dbl>	<dbl>
11	549	Thirty-Two Short Films About Glenn Gould	4.500000	3
22	1175	Delicatessen	4.500000	3
31	1208	Apocalypse Now	4.500000	3
56	2076	Blue Velvet	3.666667	3
6	318	Shawshank Redemption, The	3.166667	3

Figure 14. movie recommended system result for old user

10. Impact of Project Outcomes

Recommendation Systems are very popular type of machine learning applications that are used in lots of fields like Youtube, Netflix, Google, Facebook. Although our project outcomes may just be kind of prototype or may not be so perfect, the algorithms in the world are an improvement over the traditional classification algorithms as they can take many classes of input and provide similarity ranking based algorithms to provide the user with accurate results. These recommendation systems have evolved over time and have incorporated many advanced machine learning techniques to provide the users with the content that they want.