VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELAGAVI 590018

Report on

“MOVIE RECOMMENDATION SYSTEM”

By

RIYANCHHI AGRAWAL (1BM17CS076)
SHAKSHI PANDEY (1BM17CS092)
SHEHYAAZ KHAN NAYAZI (1BM17CS094)

Under the Guidance of

Prof. Bhoomika A P
Assistant Professor, Department of CSE
B.M.S College of Engineering
Work carried out at

Department of Computer Science and Engineering
B.M.S College of Engineering
(Autonomous College under VTU)
P.O. Box No.: 1908, Bull Temple Road, Bangalore-560 019
2020-2021

B.M.S COLLEGE OF ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Data Science using R self-study project titled **“Movie Recommendation System”** has been carried out by RIYANCHHI AGRAWAL (1BM17CS076), SHAKSHI PANDEY (1BM17CS092), SHEHYAAZ KHAN NAYAZI (1BM17CS094) during the academic year 2020-2021.

Signature of the guide
Prof. Bhoomika A P
Assistant Professor,
Department of Computer Science and Engineering
B.M.S College of Engineering, Bangalore

B.M.S COLLEGE OF ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

We, RIYANCHHI AGRAWAL (1BM17CS076), SHAKSHI PANDEY (1BM17CS092), SHEHYAAZ KHAN NAYAZI (1BM17CS094), students of 7th Semester, B.E, Department of Computer Science and Engineering, B.M.S College of Engineering, Bangalore hereby declare that, this Data Science using R self-study project work entitled "**MOVIE RECOMMENDATION SYSTEM**" has been carried out by us under the guidance of Prof. Bhoomika A P, Assistant Professor, Department of CSE, B.M.S College of Engineering, Bangalore during the academic semester August-December 2020. We also declare that to the best of our knowledge and belief, the assignment reported here is not from part of any other report by any other students.

Signature of the Candidates

RIYANCHHI AGRAWAL (1BM17CS076)

SHAKSHI PANDEY (1BM17CS092)

SHEHYAAZ KHAN NAYAZI (1BM17CS094)

INDEX

SI. No.	TOPIC	PAGE No.
1.	OBJECTIVE	
2.	INTRODUCTION	
3.	DESIGN MODULES	
4.	DETAILED DESCRIPTION OF MODULES	
5.	SCREENSHOTS	
6.	NEW LEARNINGS AND FUTURE ENHANCEMENTS	
7.	REFERENCES	

I. DESCRIPTION:

a. OBJECTIVE :

On the Internet, where the number of choices is overwhelming, there is a need to filter, prioritize and efficiently deliver relevant information in order to alleviate the problem of information overload, which has created a potential problem to many Internet users.

Recommender systems solve this problem by searching through large volumes of dynamically generated information to provide users with personalized content and services.

Here different characteristics and potentials of different prediction techniques in recommendation systems is explored in order to serve as a compass for research and practice in the field of recommendation systems.

The main goal of this data science project is to build a recommendation engine that recommends movies to users. This R project is designed to help to understand the functioning of how a recommendation system works. The recommendation is developed using Item Based Collaborative Filtering.

b. INTRODUCTION :

Analysis of data examines large and different types of data and studies their growth and development.

It involves three major stages-

- ■ Pre-processing the data
- ■ Data Analysis
- ■ Visualizing the data.
- ■ Data Recommendations

Data Pre-processing - It involves methods and techniques used to discover knowledge from the data. As data is most likely to be imperfect, inconsistent and sometimes redundant it cannot be directly used in the next step i.e. Data Mining process. Data preprocessing stage enables us to process data by adapting it to the requirements presented by each data mining algorithm.

The **Data Analysis** stage is dedicated to carrying out the actual analysis task, which typically involves one or more types of analytics. This stage can be iterative in nature, especially if the data analysis is exploratory, in which case analysis is repeated until the appropriate pattern or correlation is uncovered. The exploratory analysis approach will be explained shortly, along with confirmatory analysis.

Depending on the type of analytic result required, this stage can be as simple as querying a dataset to compute an aggregation for comparison. On the other hand, it can be as challenging as combining data mining and complex statistical analysis techniques to discover patterns and anomalies or to generate a statistical or mathematical model to depict relationships between variables.

Data Visualization- The ability to analyze massive amounts of data and find useful insights carries little value if the only ones that can interpret the results are the analysts.The Data Visualization stage is dedicated to using data visualization techniques and tools to graphically communicate the analysis results for effective interpretation by business users.

Business users need to be able to understand the results in order to obtain value from the analysis and subsequently have the ability to provide feedback.The results of completing the Data Visualization stage provide users with the ability to perform visual analysis, allowing for the discovery of answers to questions that users have not yet even formulated.

Data Recommendations- Recommendation system provides the facility to understand a person's taste and find new, desirable content for them automatically based on the pattern between their likes and rating of different items. Filtering means filtering products based on ratings and other user data. Recommendation systems use three types of filtering: collaborative, user-based and a hybrid approach. In collaborative filtering, a comparison of users' choices is done and recommendations given.

Recommendation algorithms are at the core of the movie viewer applications. They provide users with personalized suggestions to reduce the amount of time and frustration to find something great content to watch.

Collaborative filtering- is a family of algorithms where there are multiple ways to find similar users or items and multiple ways to calculate rating based on ratings of similar users. Depending on the choices you make, you end up with a type of **collaborative filtering** approach.

It is a method to predict a user's taste and find the items that a user might prefer on the basis of information collected from various other users having similar tastes or preferences. It takes into consideration the basic fact that if person X and person Y have a certain reaction for some items then they might have the same opinion for other items too.

Item Based collaborative filtering- The very first step is to build the model by finding similarity between all the item pairs. The similarity between item pairs can be found in different ways. One of the most common methods is to use cosine similarity.

Formula for Cosine Similarity:

$$\boxed{Similarity(\overline{A},\overline{B}) = \frac{\overline{A}\bullet\overline{B}}{||\overline{A}||*||\overline{B}||}}$$

Pearson's Method- The Pearson correlation coefficient, also referred to as Pearson's r, the Pearson product-moment correlation coefficient, or the bivariate correlation, is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and −1.

Formula for pearson's method:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

$x_{\{i\}}$ = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

$y_{\{i\}}$= values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

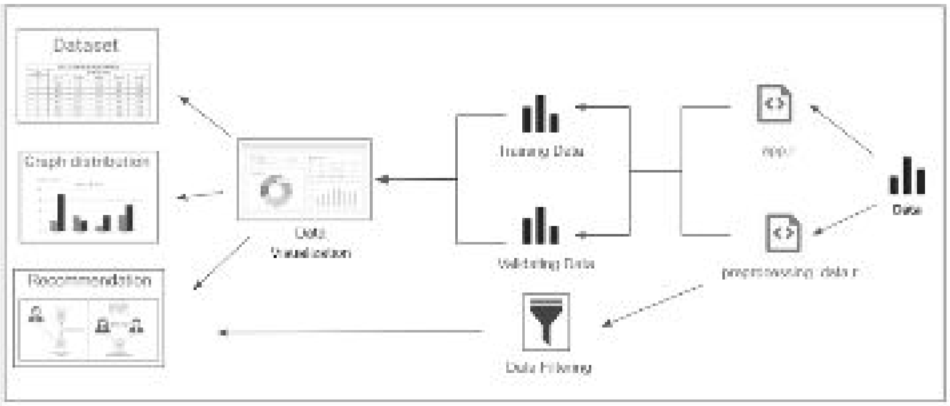
Demographic Filtering- Demographic Filtering (DF) technique uses the demographic data of a user to determine which items
may be appropriate for recommendation.

So we will be using Demographic Filtering to get the best ranking score for these apps where,

- v is the number of votes for the app;
- m is the minimum votes required to be listed in the chart;
- R is the average rating of the app; And
- C is the mean vote across the whole report

$$\text{Weighted Rating (WR)} = \left(\frac{v}{v+m} . R\right) + \left(\frac{m}{v+m} . C\right)$$

c. DESIGN MODULES :



d. DETAILED DESCRIPTION OF EACH MODULES

- Requirement and Functionality

➢ Functional Requirements-

- The system must be able to recommend movies to users based on their preferences using Collaborative Filtering.
- The system should compare different Collaborative Filtering models.
- The system should analyze the dataset and show it visually through graphs.

➢ System Requirements-

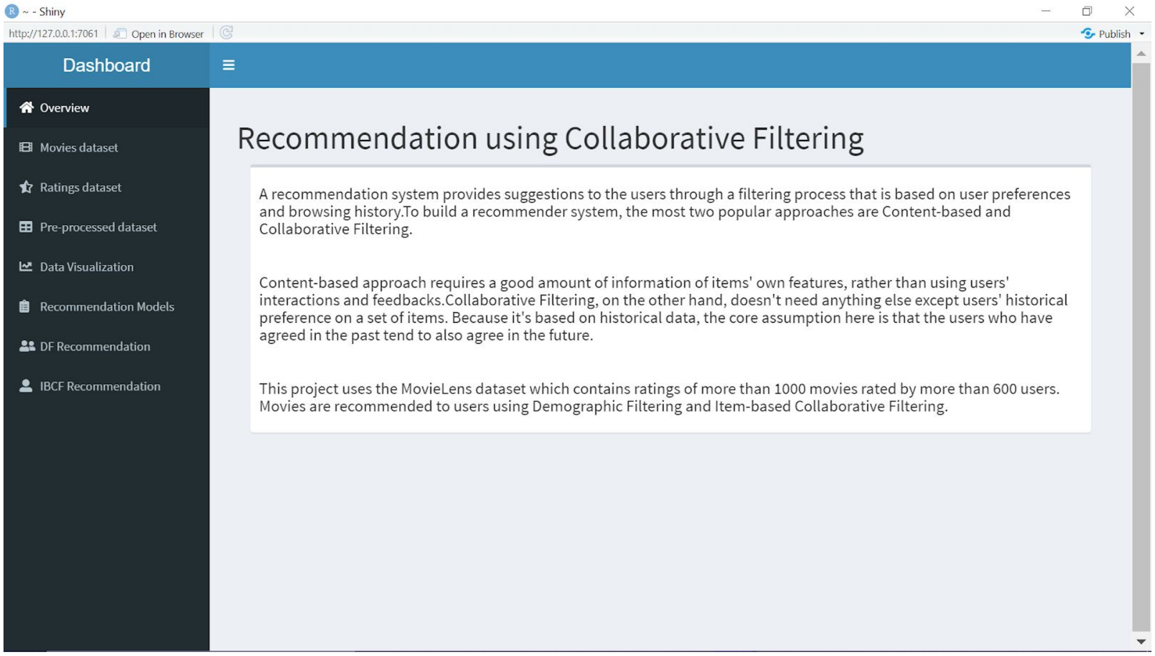
- System Specifications-
    - Windows(version 7 or above)/Linux OS
    - 500Gb-1Tb HDD/SSD
    - 4GB RAM or above
- Tools Required-
    - R Studio
    - R programming language

- Techniques / Mechanisms used

To perform the above mentioned functionalities following python libraries are used-

- Ggplot2 - It is a data visualization package for the statistical programming language R.
- Reshape2 - It is an R package written by Hadley Wickham that makes it easy to transform data between wide and long formats.
- Data.table - It is widely used for fast aggregation of large datasets, add/update/remove of columns, quicker ordered joins, and a fast file reader.
- Dplyr - dplyr is primarily a set of functions designed to enable dataframe manipulation in an intuitive, user-friendly way.
- Tidyr - Tidy data is data that's easy to work with: it's easy to munge (with dplyr), visualise and model (with R's hundreds of modelling packages).
- Shiny - Shiny is an R package that makes it easy to build interactive web apps straight from R.
- Tibble - It is a modern reimagining of the data.frame, keeping what time has proven to be effective, and throwing out what is not.
- Forcats - levels of a factor can be checked using the forcats library.
- DT - DT provides an R interface to the JavaScript library DataTables.
- Recommenderlab - Provides lab for Developing and Testing Recommender Algorithms

II SCREENSHOTS AND EXPLANATION



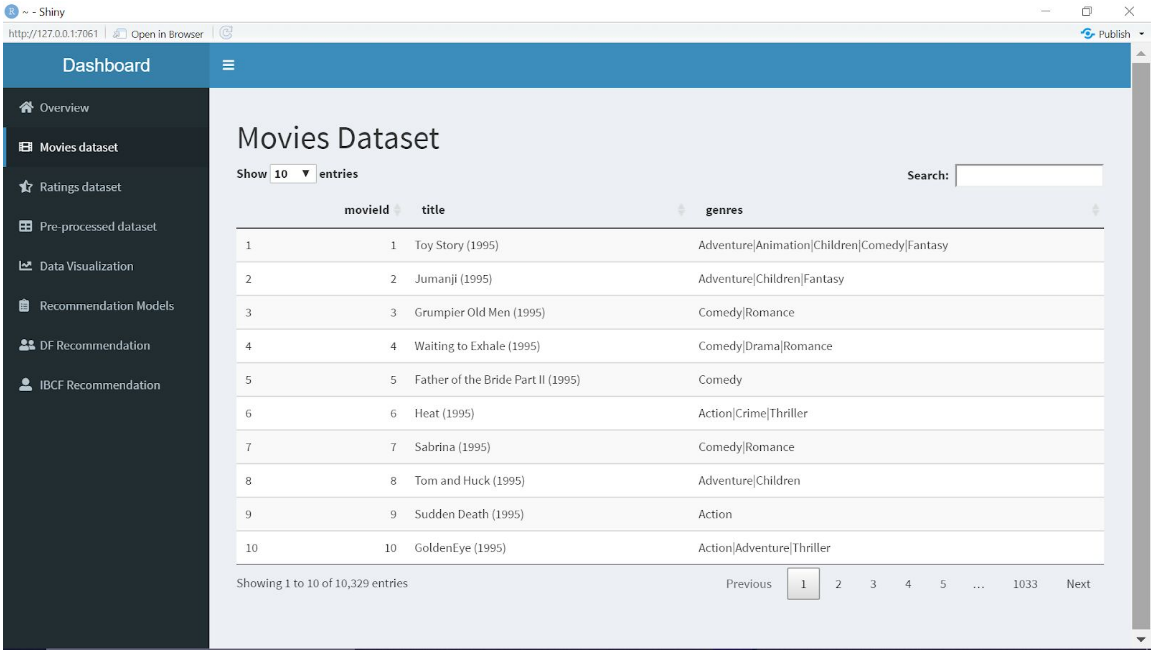
Recommendation using Collaborative Filtering

A recommendation system provides suggestions to the users through a filtering process that is based on user preferences and browsing history.To build a recommender system, the most two popular approaches are Content-based and Collaborative Filtering.

Content-based approach requires a good amount of information of items' own features, rather than using users' interactions and feedbacks.Collaborative Filtering, on the other hand, doesn't need anything else except users' historical preference on a set of items. Because it's based on historical data, the core assumption here is that the users who have agreed in the past tend to also agree in the future.

This project uses the MovieLens dataset which contains ratings of more than 1000 movies rated by more than 600 users. Movies are recommended to users using Demographic Filtering and Item-based Collaborative Filtering.

Description of Content-based and Collaborative Filtering



Movies Dataset

Show 10 entries

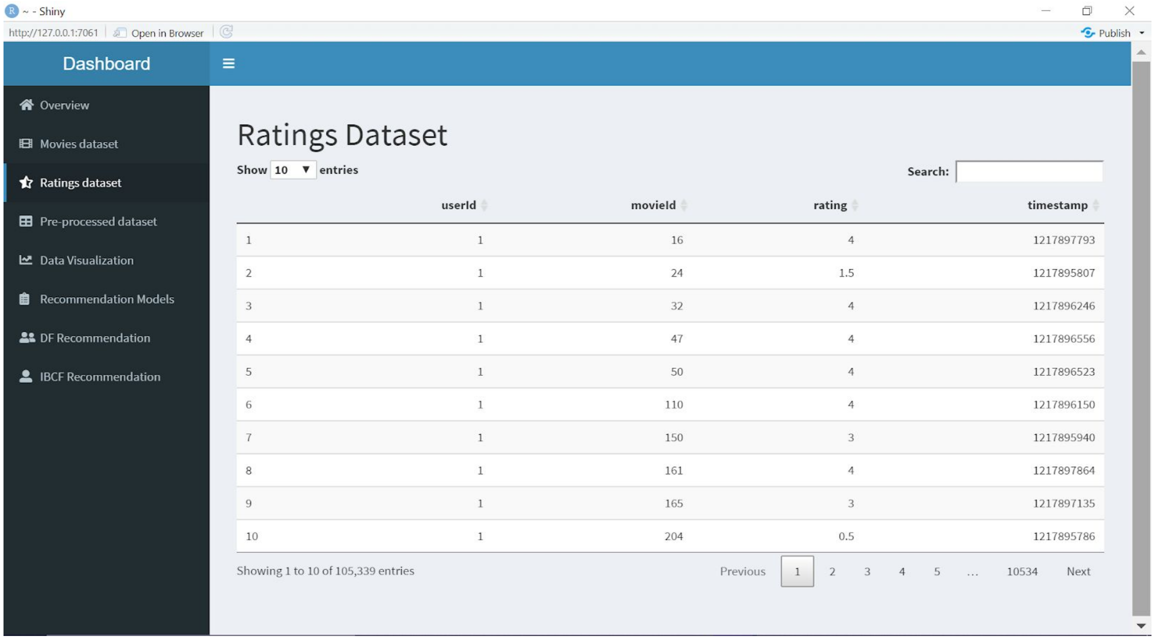
Search:

	movieId	title	genres
1	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
2	2	Jumanji (1995)	Adventure|Children|Fantasy
3	3	Grumpier Old Men (1995)	Comedy|Romance
4	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
5	5	Father of the Bride Part II (1995)	Comedy
6	6	Heat (1995)	Action|Crime|Thriller
7	7	Sabrina (1995)	Comedy|Romance
8	8	Tom and Huck (1995)	Adventure|Children
9	9	Sudden Death (1995)	Action
10	10	GoldenEye (1995)	Action|Adventure|Thriller

Showing 1 to 10 of 10,329 entries

Previous 1 2 3 4 5 … 1033 Next

Movie details in the MovieLens dataset



Ratings Dataset

Show 10 ▼ entries

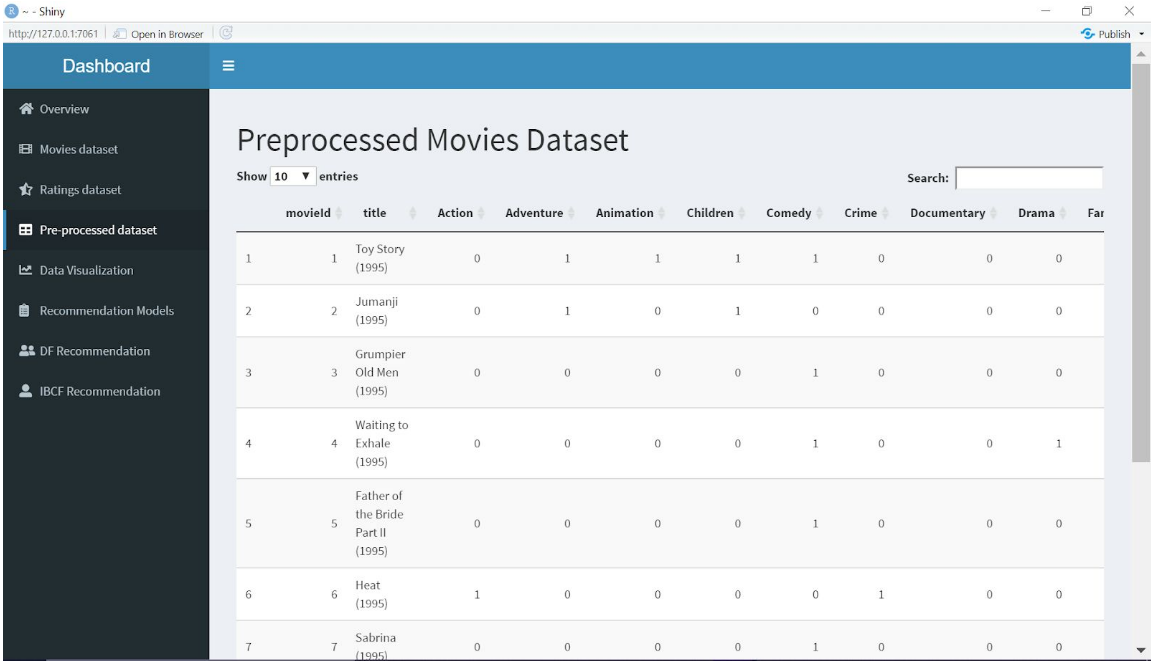
Search:

	userId	movieId	rating	timestamp
1	1	16	4	1217897793
2	1	24	1.5	1217895807
3	1	32	4	1217896246
4	1	47	4	1217896556
5	1	50	4	1217896523
6	1	110	4	1217896150
7	1	150	3	1217895940
8	1	161	4	1217897864
9	1	165	3	1217897135
10	1	204	0.5	1217895786

Showing 1 to 10 of 105,339 entries

Previous 1 2 3 4 5 … 10534 Next

User ratings in the MovieLens dataset



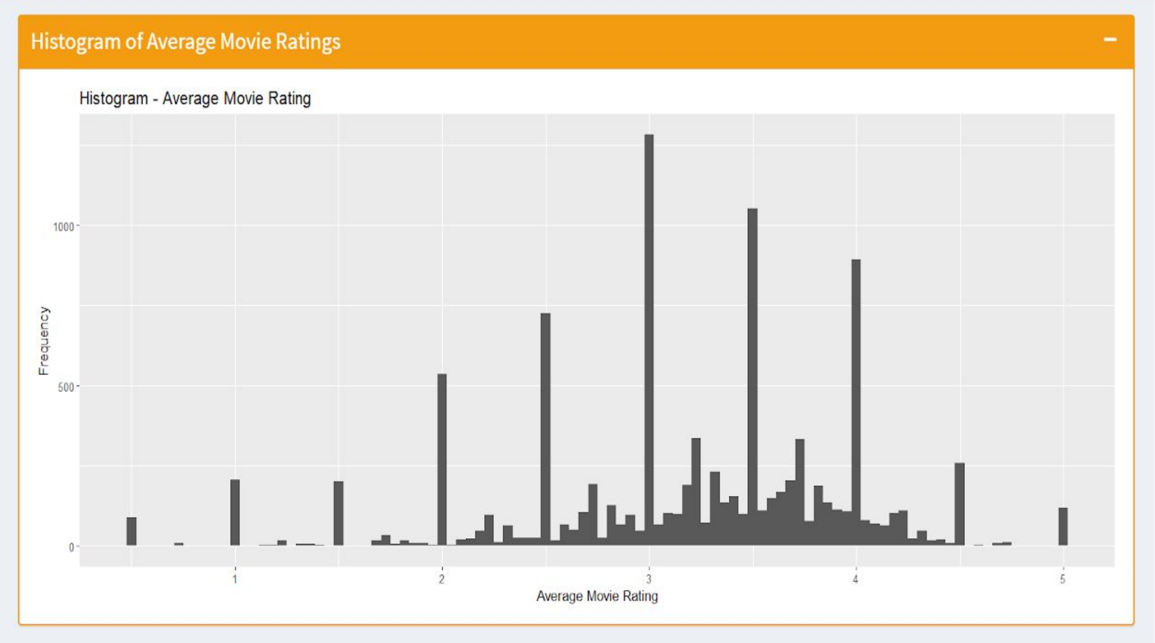
Preprocessed Movies Dataset

Show 10 ▼ entries

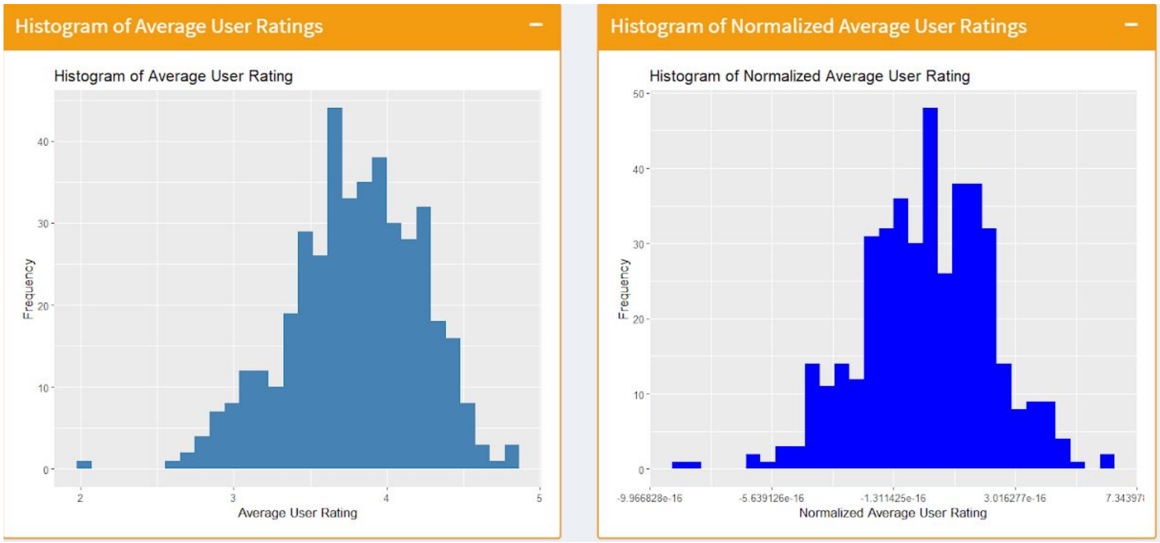
Search:

	movieId	title	Action	Adventure	Animation	Children	Comedy	Crime	Documentary	Drama	Far
1	1	Toy Story (1995)	0	1	1	1	1	0	0	0	
2	2	Jumanji (1995)	0	1	0	1	0	0	0	0	
3	3	Grumpier Old Men (1995)	0	0	0	0	1	0	0	0	
4	4	Waiting to Exhale (1995)	0	0	0	0	1	0	0	1	
5	5	Father of the Bride Part II (1995)	0	0	0	0	1	0	0	0	
6	6	Heat (1995)	1	0	0	0	0	1	0	0	
7	7	Sabrina (1995)	0	0	0	0	1	0	0	0	

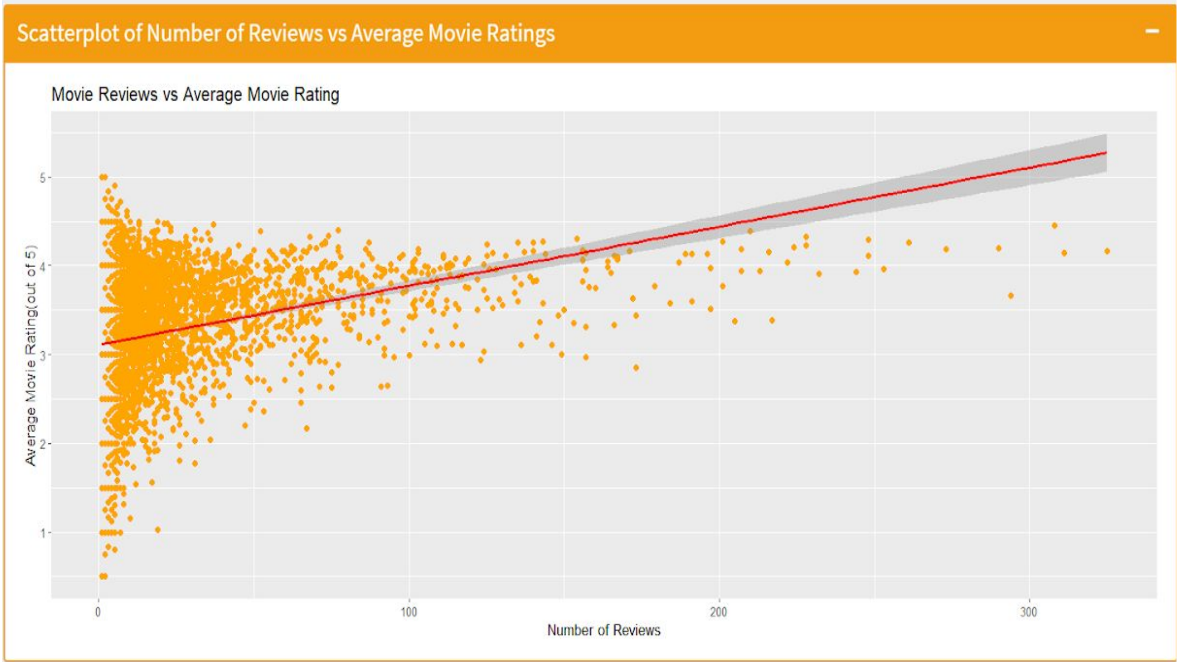
Pre-processed data with separate columns for each genre and a demographic score for every movie



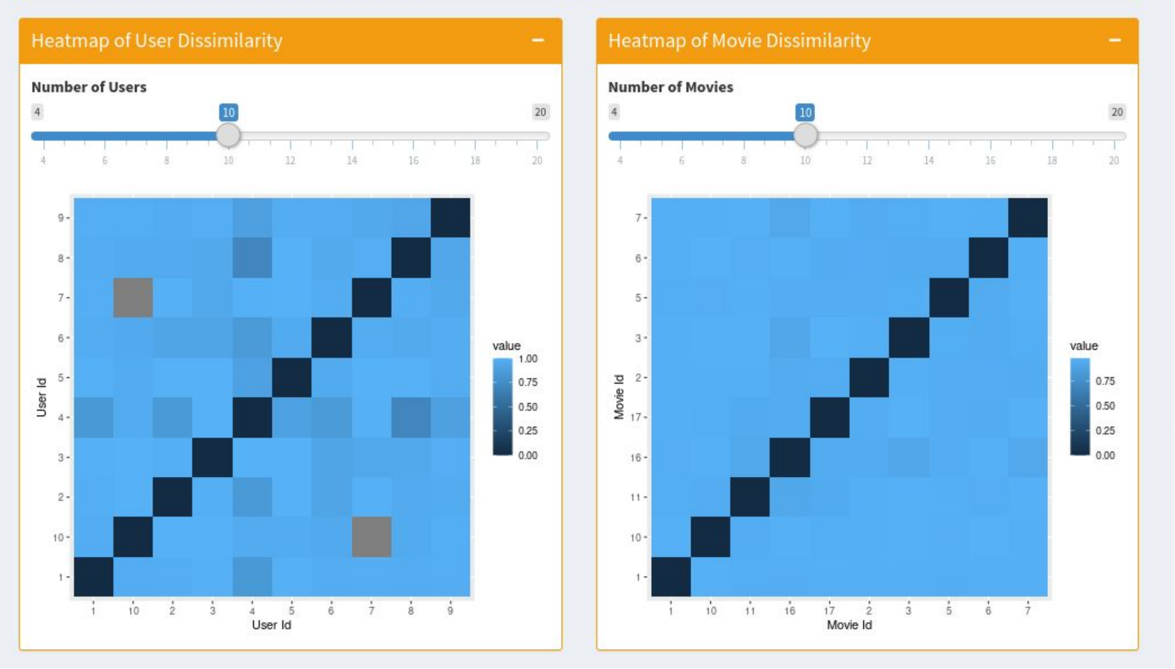
Histogram of Average Movie Rating, this shows that most movies have an average rating of 3, indicating that most movies are of good quality



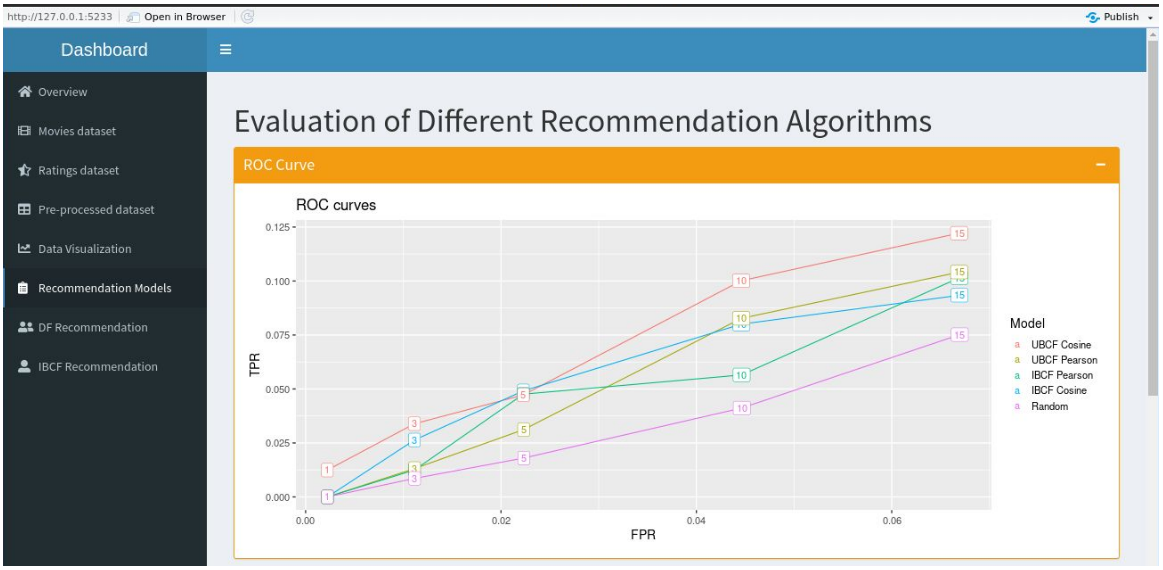
Histogram of Average User Rating and Normalised User Rating, this shows that most users have provided a rating between 3 and 4



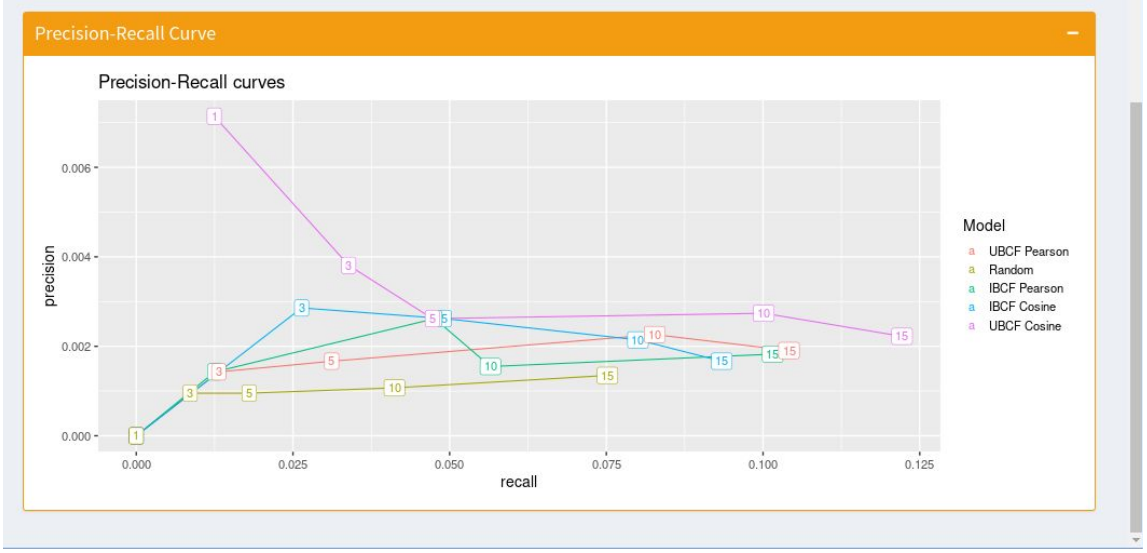
Scatterplot of Average movie rating against number of movie reviews, the red line shows the best fit line for this plot. A slight positive trend can be seen between a movie's average rating and its number of reviews.



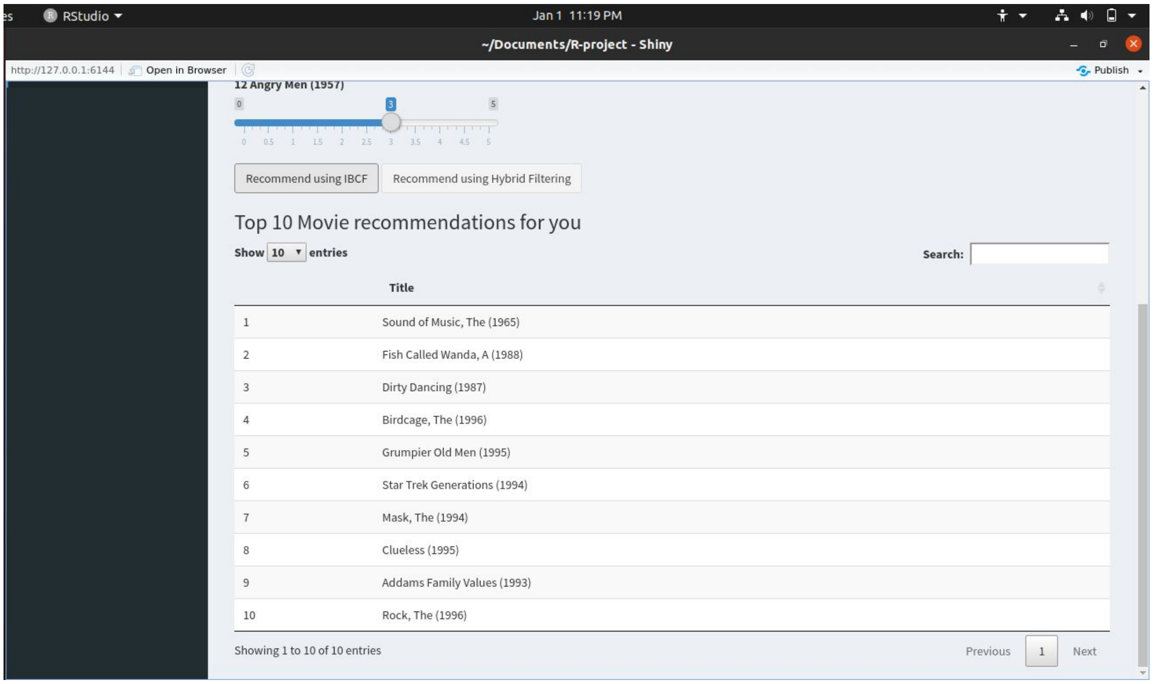
A Heatmap of User and Movie dissimilarity, this indicates the dissimilarity and similarity between different users and movies



ROC curve of different recommendation algorithms, this shows that UBCF Cosine has the largest Area under ROC curve



Precision-Recall curve of different recommendation algorithms, this shows that UBCF Pearson method has the largest Area under the PR curve



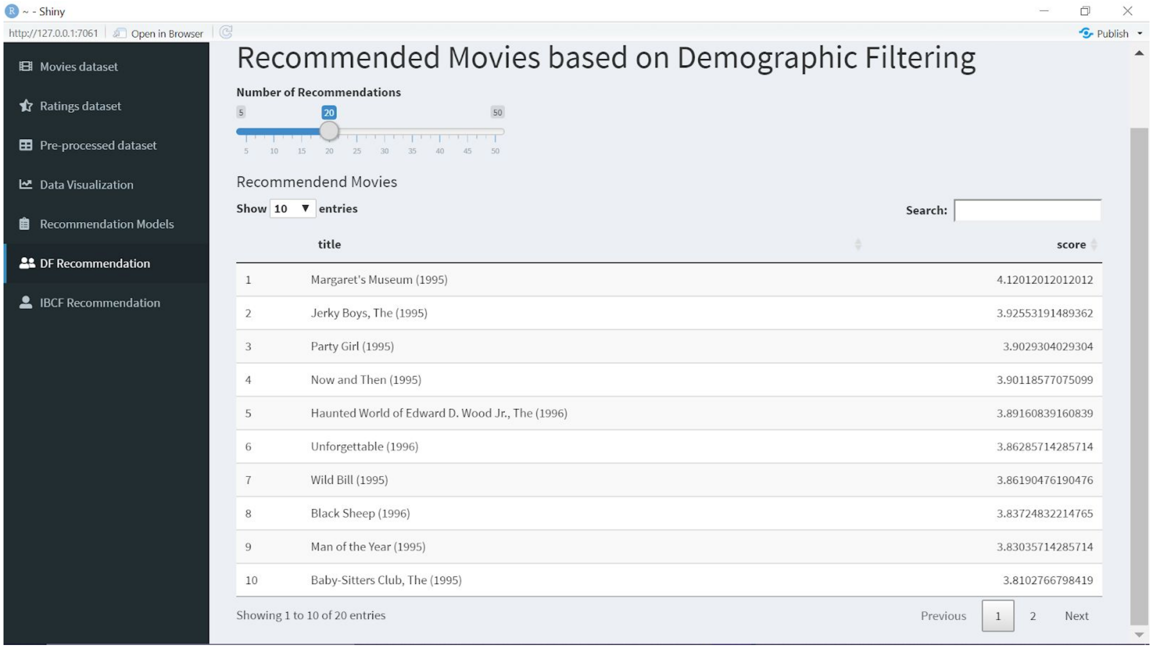
Top 10 Movie recommendations for you

Show 10 entries Search:

	Title
1	Sound of Music, The (1965)
2	Fish Called Wanda, A (1988)
3	Dirty Dancing (1987)
4	Birdcage, The (1996)
5	Grumpier Old Men (1995)
6	Star Trek Generations (1994)
7	Mask, The (1994)
8	Clueless (1995)
9	Addams Family Values (1993)
10	Rock, The (1996)

Showing 1 to 10 of 10 entries Previous 1 Next

Top 10 movie recommendations using IBCF Pearson method, the movies are recommended in descending order of a user's preferences, with the most preferred movie recommended first.



Recommended Movies based on Demographic Filtering

Number of Recommendations: 20 (5–50)

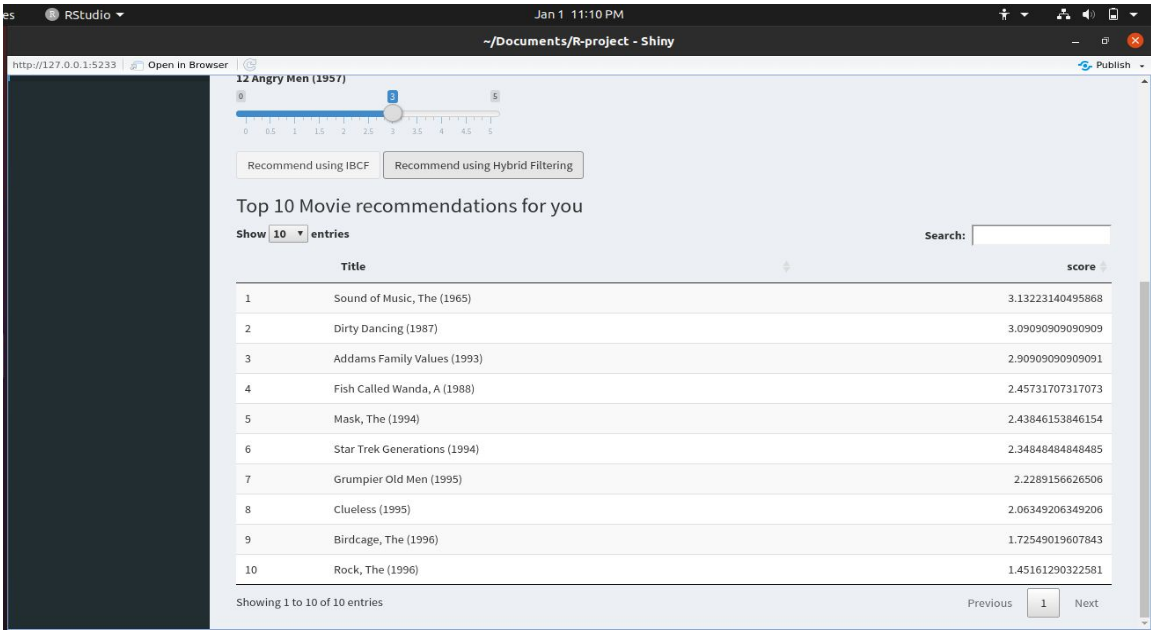
Recommendend Movies

Show 10 entries Search:

	title	score
1	Margaret's Museum (1995)	4.12012012012012
2	Jerky Boys, The (1995)	3.92553191489362
3	Party Girl (1995)	3.9029304029304
4	Now and Then (1995)	3.90118577075099
5	Haunted World of Edward D. Wood Jr., The (1996)	3.89160839160839
6	Unforgettable (1996)	3.86285714285714
7	Wild Bill (1995)	3.86190476190476
8	Black Sheep (1996)	3.83724832214765
9	Man of the Year (1995)	3.83035714285714
10	Baby-Sitters Club, The (1995)	3.8102766798419

Showing 1 to 10 of 20 entries Previous 1 2 Next

Movie recommendation using Demographic Filtering, movies with the highest demographic score are recommended first.Movie recommendation using Item-based Collaborative Filtering using Pearson method, users can rate upto 10 movies from the multiple select dropdown of movie titles.



Top 10 Movie recommendations for you

Show 10 entries Search:

	Title	score
1	Sound of Music, The (1965)	3.13223140495868
2	Dirty Dancing (1987)	3.09090909090909
3	Addams Family Values (1993)	2.90909090909091
4	Fish Called Wanda, A (1988)	2.45731707317073
5	Mask, The (1994)	2.43846153846154
6	Star Trek Generations (1994)	2.34848484848485
7	Grumpier Old Men (1995)	2.2289156626506
8	Clueless (1995)	2.06349206349206
9	Birdcage, The (1996)	1.72549019607843
10	Rock, The (1996)	1.45161290322581

Showing 1 to 10 of 10 entries

Top 10 movie recommendations using Hybrid filtering, the movies are recommended using IBCF Pearson and Demographic filtering.

III NEW LEARNINGS FROM THE SELF STUDY PROJECT

We have gained a better insight on the phases involved in the Data Analytics Life Cycle process through this project. Data extraction and manipulation on huge amounts of data has added to the experience of the learning journey. The application of several data analysis techniques has helped us in systematically extracting information from, or otherwise dealing with data sets that are too large or complex to be dealt with by traditional data-processing application software.

We have learned many useful R packages through this project. This includes the Tidyverse packages, such as ggplot2, dplyr, tibble and forcats, and other packages such as DT, reshape2, shiny and recommenderlab. Analysis of the MovieLens dataset in this project has helped us learn about recommendation systems and various recommendation algorithms such as Demographic Filtering, User-based Collaborative Filtering and Item-based Collaborative Filtering and compare the same through ROC and PR curves. We have also learnt about the pros and cons of such recommendation algorithms. The use of ShinyR in this project has taught us the process of Web app development using R. Through this project, we feel confident in tackling and solving some of the prevailing problems in the field of Data Science.

IV FUTURE ENHANCEMENTS

We propose the following future enhancements for this project :

a. The use of more sophisticated recommendation algorithms such as those used in Hybrid recommendation systems.

b. The use of a larger dataset to develop a more accurate recommendation system.

c. Real-time recommendations to users by using their browsing history.

d. This system could also be extended to recommend web series that are available on various popular platforms to the users.

V REFERENCES

1. https://towardsdatascience.com/intro-to-recommender-system-collaborative-filtering-64a238194a26
2. https://grouplens.org/datasets/movielens/
3. https://shiny.rstudio.com/tutorial/
4. https://rstudio.github.io/shinydashboard/get_started.html
5. https://www.rdocumentation.org/packages/recommenderlab/versions/0.2-6
6. https://www.rdocumentation.org/packages/dplyr/versions/0.7.8
7. https://www.kaggle.com/accountstatus/google-play-store-apps-eda-and-reccomendations
8. https://data-flair.training/blogs/data-science-r-movie-recommendation/