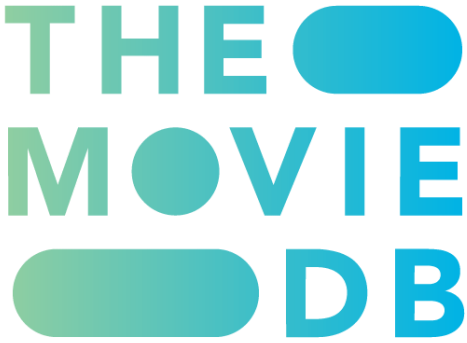


The Movie Database Analysis

TEAM # 18

Name	id
Khaled Mahmoud	1180105
Ahmed Mostafa Abdelrahman	1180123
Ali Haitham	1180048

Table of Contents

Problem Description

Provided is a dataset containing data about movies, their titles, actors, directors, genres, keywords and other information like budget, revenue and popularity. Together with a ratings file defining each rating with user id, movie id and rating.

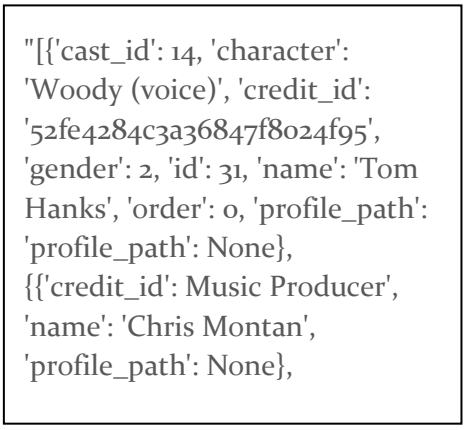
As previously established in the proposal the objective is

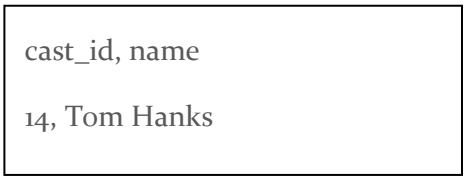
- Movie recommendation system based on
  1. Users' ratings
  2. Movies features
- Movie style (genre) recommendation for production companies

Project Pipeline
Analysis and Solution

DATA PREPROCESSING

1. Parsing





2. Cleaning

On investigating the ratings and metadata of movies, it turned out there are movie ids not in the metadata, that suggests that the data is synthetic yet, only the ones in common were extracted from both to carry out an analysis

3. Integrity

Comparing the vote averages in the metadata with the ratings' averages acquired, it shows a great difference which supports the concern raised earlier of the data being synthetic

Note: it is confirmed that the ratings data comes from different distribution, the project regards this and tries to overcome it through distribution analysis, yet one can only exploit the ratings only through a machine learning model, yet comparing the results of the content-based and collaborative filtering will not be possible

DATA VISULIZATION

Selected visualizations are shown in the report, others can be seen in the jupyter notebooks provided with the submission, some of these visualizations are for exploring data with the ratings, on acquiring the ratings or any ratings synonymous with metadata these visualizations would gain immense meaning

RECOMENATION SYSTEM

1. Histogram of movies

The data distribution is critical to check on when working across multiple years, checking on the data distribution gives insights on the data

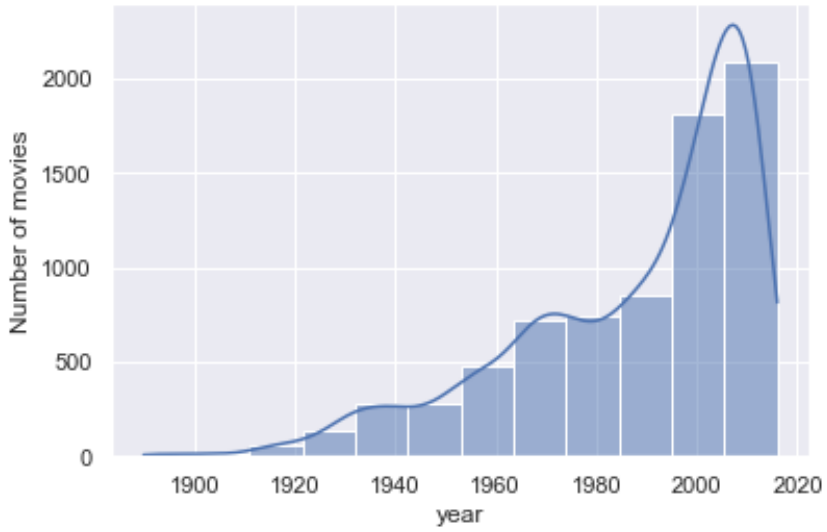


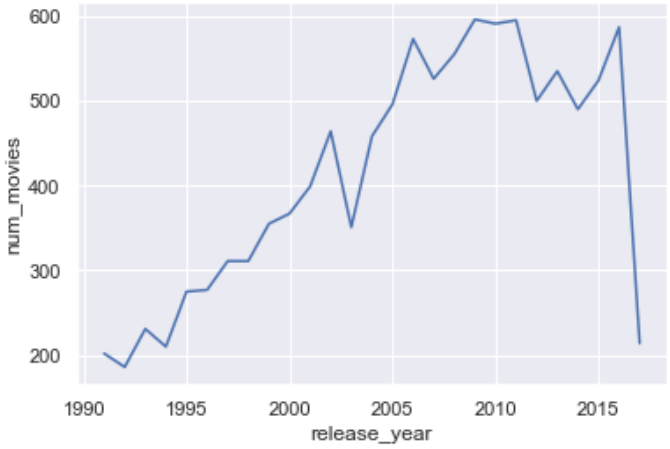
Figure 1: Number of movies per year

The number of movies increase as time goes by; this makes you wonder if the users rate the later movies more heavily than others, an interest metric which is the total votes within a year, it doesn't matter which ratings, no publicity is bad puplicity

DIRECTION OF MOVIES

2. Number of movies since the 90s

2017 data is apparently missing, yet they are the most recent ones and not to be discarded



c

Figure 2 Movies in the previous 27 years to 2017

3. Genres Profit and Budget

The movies have multiple genres assigned to them, so to eliminate the confusion, each movie with the group of genres assigned to them, then covering the per movie profit to catch the effect of the genre in all movies, with the average budget to make them

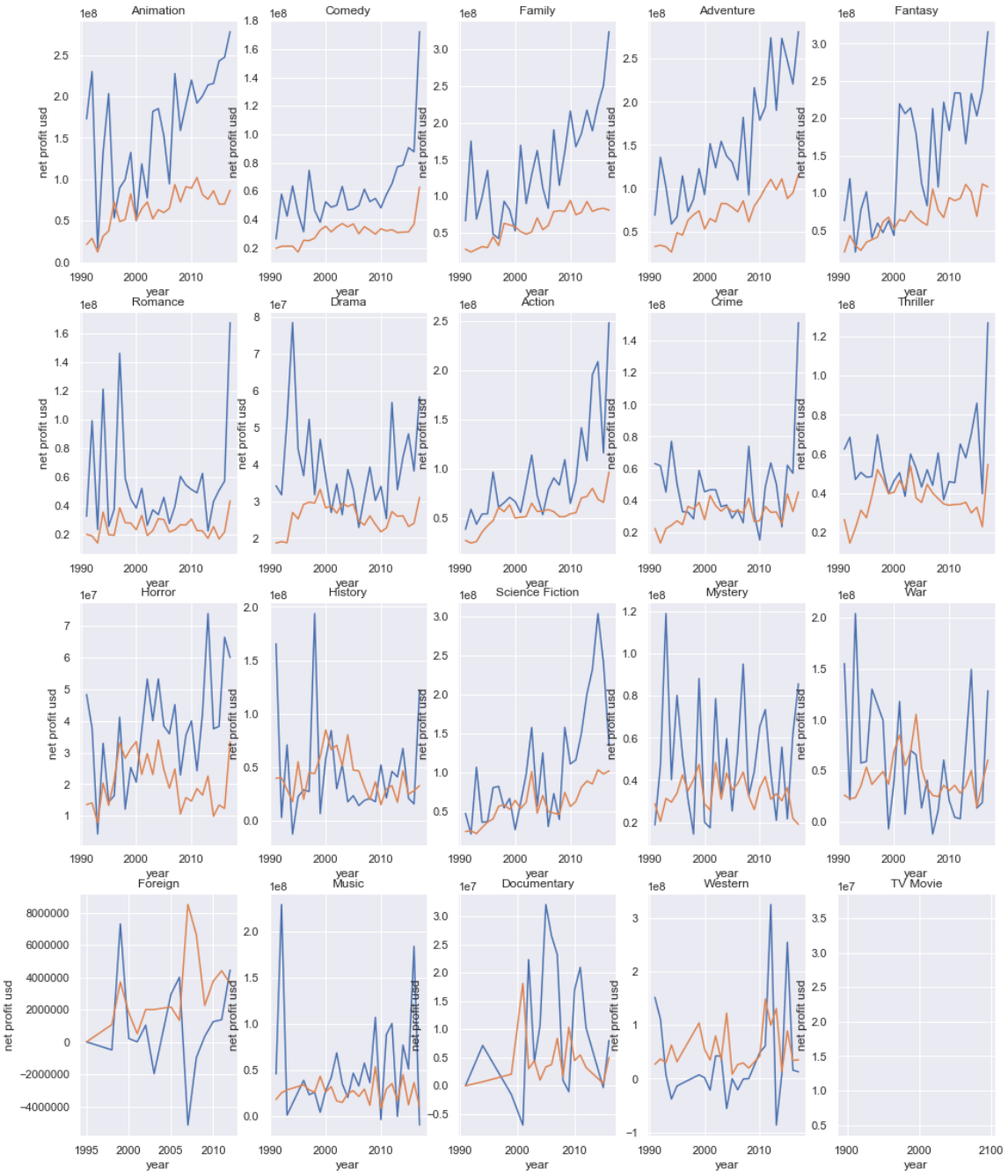


Figure 3 Genres average profit and budget over the years

4. Movies shares in a year

Share in movies is an indication on the trend of production in a year, if 50% of movies are comedy for instance, this is a strong indication that the comedy is a money maker that the movie houses are shifting towards to exploit it

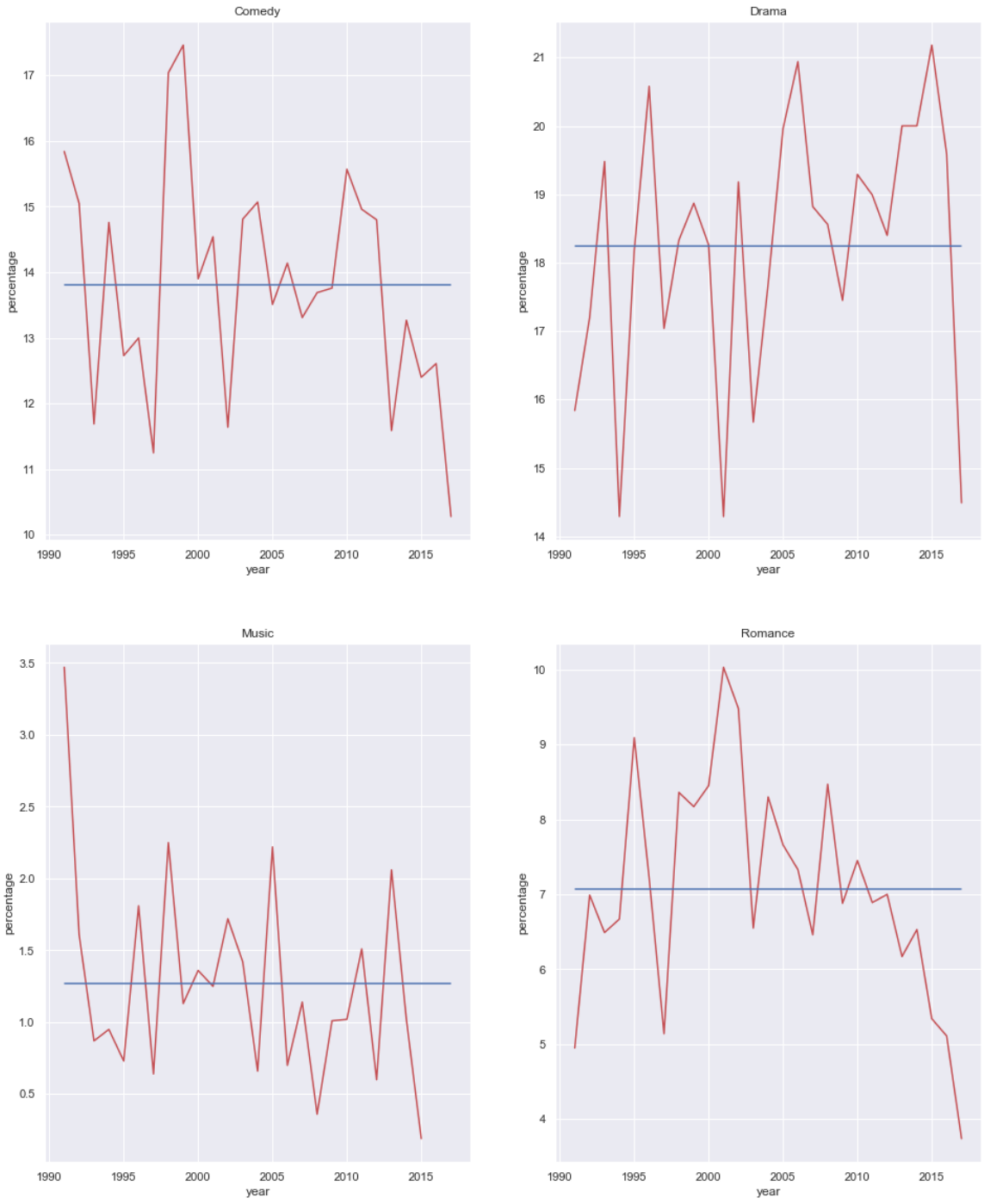


Figure 4 falling genres over the previous years

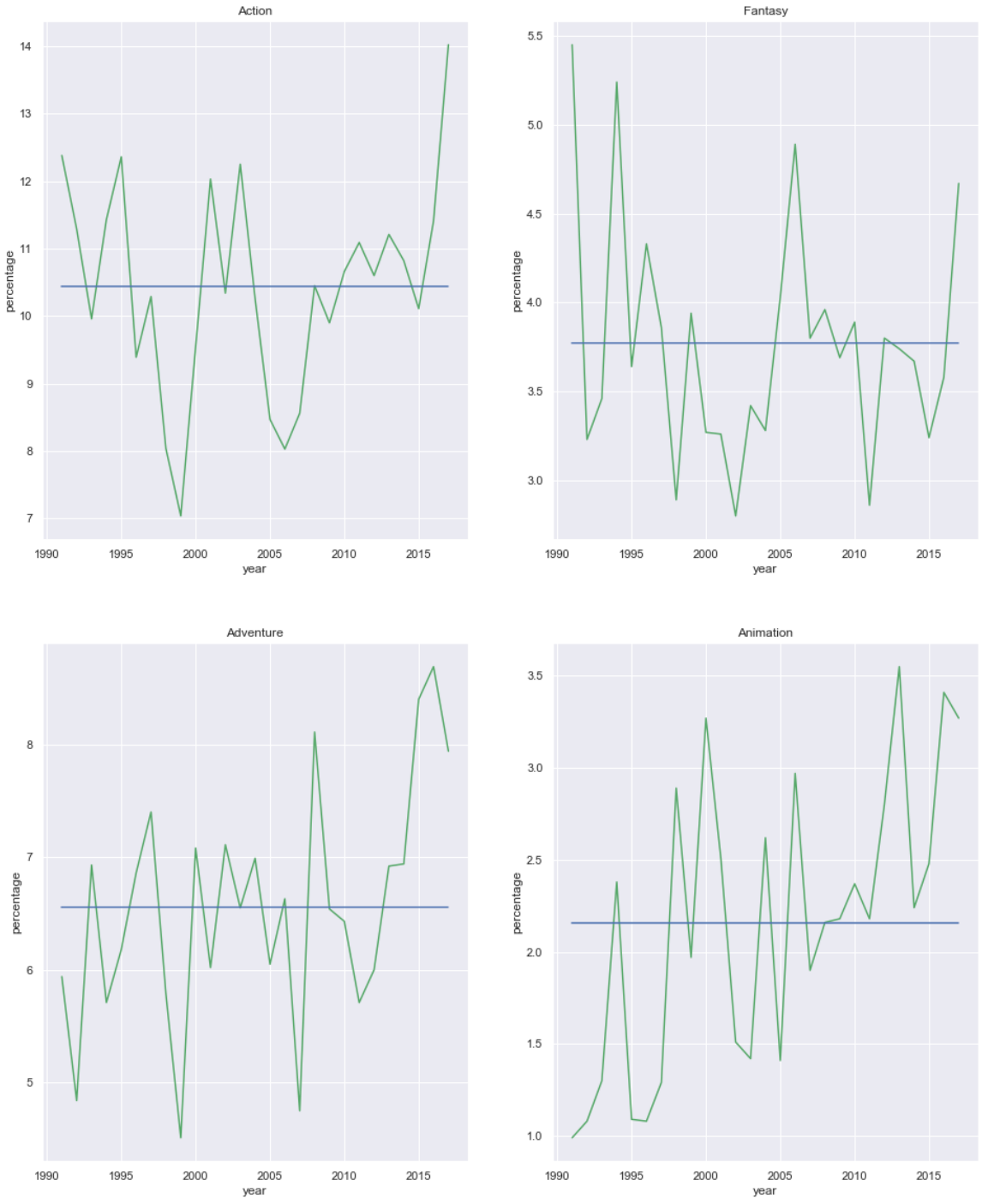


Figure 5 Thriving genres in previous years

A genre is said to be bad if it is simply declining or even it has large variance in its share, it said to be thriving if it is over at the past years and shows certain stability

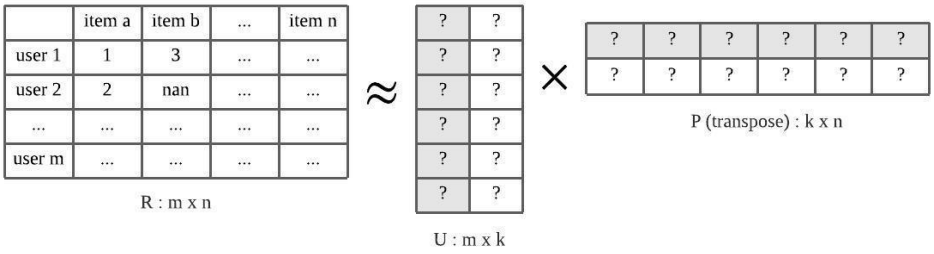
Results and Evaluation

In this section there will be an assessment for the results of the analysis along with the conclusions

RECOMMENDATION SYSTEM

Collaborative filtering

Used an alternating least squares model, which works basically by approximating the original data and inferring the missing values in the process



The model has 3 parameters, a 5% validation spilt of the 11M records were left aside

Tuning the max iterations and the learning parameter it was found to have 10 and 0.05 respectively, the rank (k) is left to the pySpark library to tune

Learning rate	Max iterations	RMSE
0.01	5	0.883
0.01	10	0.875
0.05	5	0.871
0.05	10	0.844

The 0.85 stars margin is adequate as it is all comparative to other movies, so ALS is chosen to be applied to the ratings data. The first model (0.01,5) has scored root mean squared error of 0.750 on 11M records of 7.5k movies and 25k users due to machine limitations

Content-Based Filtering

Using the attributes (cast, genre, keywords, director), getting similarities between movies and each other

Vectors and cosine rules were used and results are close to google results

It was found that google is based on some vivid heuristic so a heuristic approach is also considered

DIRECTION OF MOVIES

Conclusion

After the study of the history of the markets and the evaluations with the visualizations for mentioned the following conclusions were reached

1. Animation movies share in the yearly movies is increasing, it is usually funded by big companies, yet the independent scene is breaking into the cinematic market and due to the advance in technology some motivated individuals with remarkable talent are bringing ideas that draws more and more audience to animated features, the main target used to be children and their families but it has upgraded to appeal to everyone

2. Action and Adventure movies are somewhat costly yet they attract audience in a stable manner, an action movie is where you can go to make a movie if you have a good budget, the only increase in actions share is in 2008 during the economic crisis so it is fair to think that these movies have consistent audience

3. Fantasy movies require a large sense of innovation and their variations in the market share reflects the unpredictability they have; they present high risk high reward

4. Comedy and Drama despite being opposites in content, yet they are similar as genres in the audience perception, they represent a large portion of the movies created yet they are not enough on their own, adding an element of another genre of the previously mentioned, other than that they are not known for box office hits, Forrest Gump for instance combines drama with adventure

5. There is a decline in interest in Romance, so they are rebranded as in Romcoms or in historian movies, or in musicals like La La Land, a classic romantic movie will not do it in the revenue you would have to spice it up a little

6. Musicals impose a great risk and have a great element of randomness, the well-craftsmanship of the writing, dancing and singing does not necessarily attract audience, which makes jeopardizes the effort put in them

Citations to previous experiences and knowledge of the historic events and cinema trends

MAP REDUCE

Map reduce was applied multiple times over the course of the project, map-reduce that resulted in these two visualizations, after getting sum and count using pyspark map reduce.

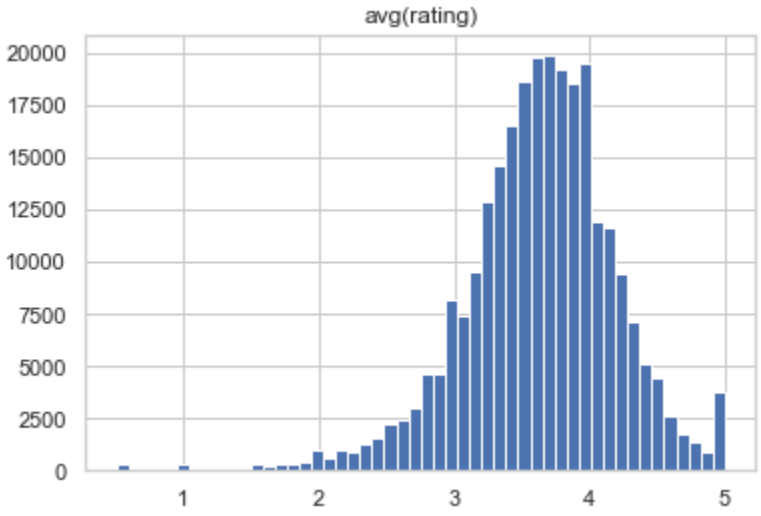


Figure 6 User average ratings

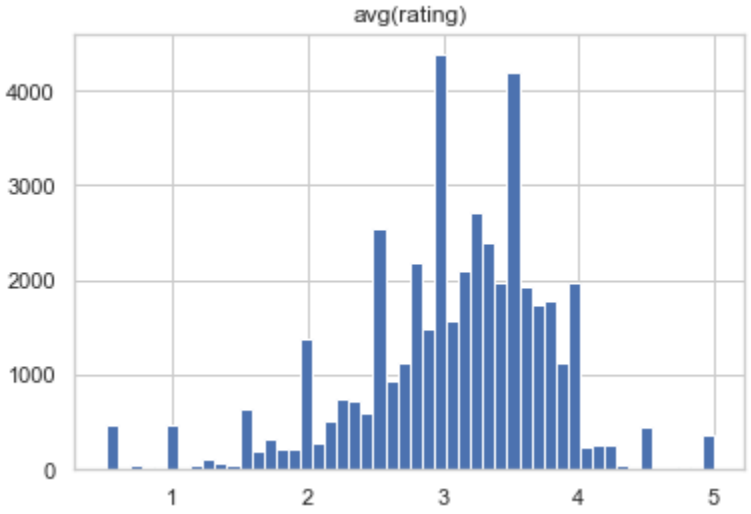


Figure 7 Movies average ratings

Unsuccessful trials

There have been many unsuccessful trials throughout the project

During collaborative filtering, there was machine limitations on applying the tuned model on the whole training set, a stack overflow happened so it was a compromise to use the simplest model which turned out acceptable within constraints

Future Work

There are many ambitions and thoughts that arisen on working on this project

1. Scrapping a better dataset form another website, to compare content-based filtering and collaborative filtering using a third-party similarity as ground truth

2. Grouping actors together with genres to see how well do they do with each genre

3. Time Series Analysis for a similar collection of movies for a certain company

Limitations

The data has two apparent limitations one for each project part

1. The ratings not being consistent with metadata

2. Genres are not underlying for each movie and with no upper limit

The data is also outdated of the effect of streaming services on the cinematic industry revenues and the kind of movies people try to go and watch blockbuster movies and wait for the rather simpler movies that rely on situations and stories to be available online to watch on the comfort of their couch or even beds