Project: Netflix Movie Recommendations

- Objective: Develop a recommendation system for Netflix movies based on user preferences.
- Dataset: Netflix Movies & TV Shows

Project Goals:

Data Collection:

- Import Netflix dataset with movie details and user ratings.
- Ensure data consistency by cleaning movie titles and genres.

Data Exploration:

- Understand trends in popular genres and user preferences.
- Identify top-rated movies based on different criteria.

Data Preprocessing:

- Handle missing values and duplicate entries.
- Convert categorical genres into numerical representations.

Exploratory Data Analysis (EDA):

- Use bar charts to show the distribution of movie genres.
- Analyze correlations between movie ratings and release years.

Feature Selection:

- Use TF-IDF and cosine similarity to build a recommendation system.

Conclusion:

Provide personalized movie recommendations based on user viewing history.

Netflix movie recommendation Project

When a young boy vanishes, a small town uncovers a mystery involving secret experiments, terrifying supernatural forces and a strange little girl.

Abstract−Nowadays, the development of technology is very rapid, so watching movies at home has become a means of entertainment. Netflix is one of the platforms for watching movies and provides various movie titles. However, because of the many movie titles, it makes it difficult for users to determine the movie they want to watch. The solution to this problem is to provide a recommendation system that can provide movie recommendations to watch. Collaborative filtering is a method that exists in the recommendation system by providing recommendations based on the ratings given by other users. Collaborative filtering is divided into two, namely based on items (item-based) and based on users (user-based). Twitter is a social media used to write posts called tweets. For this system, tweets serve as data that will be processed into ratings. This research was conducted using k-means clustering with collaborative filtering and collaborative filtering only. By using a dataset obtained from Twitter by crawling data and added with ratings from IMDb, Rotten Tomatoes, and Metacritic. Which resulted in a dataset with 35 users, 785 movie titles, and 6184 reviews. Then preprocessing the data with text processing, polarity, and labeling. And get the dataset that will be used for this experiment. The results of this research test show that k-means clustering with collaborative filtering gets the best results with the best prediction of 2.8466, getting an MAE value of 0.5029, and an RMSE value of 0.6354

1. INTRODUCTION

The development of technology at this time has developed very rapidly, including movies which are a means of entertainment media for the community. However, because there are too many movie titles that have been circulating, it is difficult for people to determine the movie they want. Currently, people not only watch movies through theaters but also online streaming platforms, namely Netflix. Netflix is one of the online streaming platforms founded in 1999 as an online video store, has become the most widely used, and is still a rapidly growing American online streaming provider specializing in video on demand [1].

Social media is a platform that gives a big impact on the development of technology, one of the most popular social media is Twitter. Twitter is one of the social media used by various groups, by writing tweets, Twitter users usually provide information, express, and express opinions on things that are happening such as movies [2].

2 Introduction

Nowadays, many people want to watch TV-shows or -series anytime and anywhere they want. In recent years, online TV has experienced exponential growth. To be exact, regarding the Digital Democracy Survey by Deloitte, which is an annual survey about changes in the digital environment, 49% of the United States households are subscribed to one or more streaming video services in 2016, compared to 31% in 2012 [1].

An interesting aspect of this exponential growth is the difference in age and the way people watch TV-shows. As can be seen in **Fig. 1**, there is a big difference between the millennials (age between 14 and 31) and the seniors (age of 68 +) regarding watch behaviour. The millennials prefer not to watch on TV only anymore, as seniors watch on TV almost all the time [2]. Instead, the millennials often choose a mobile device.

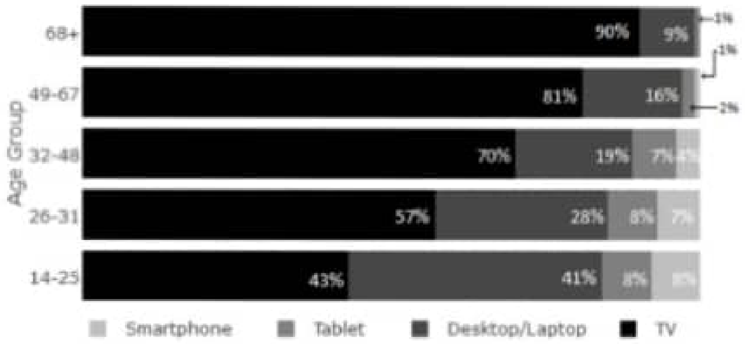
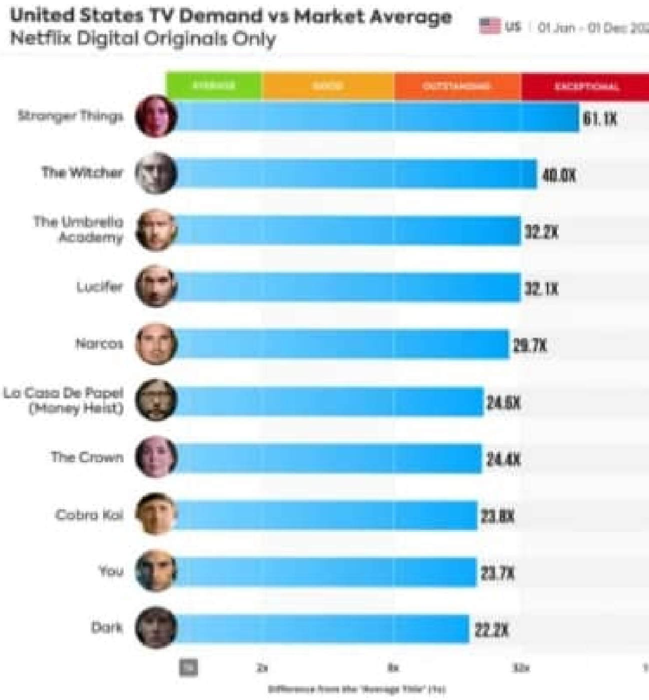
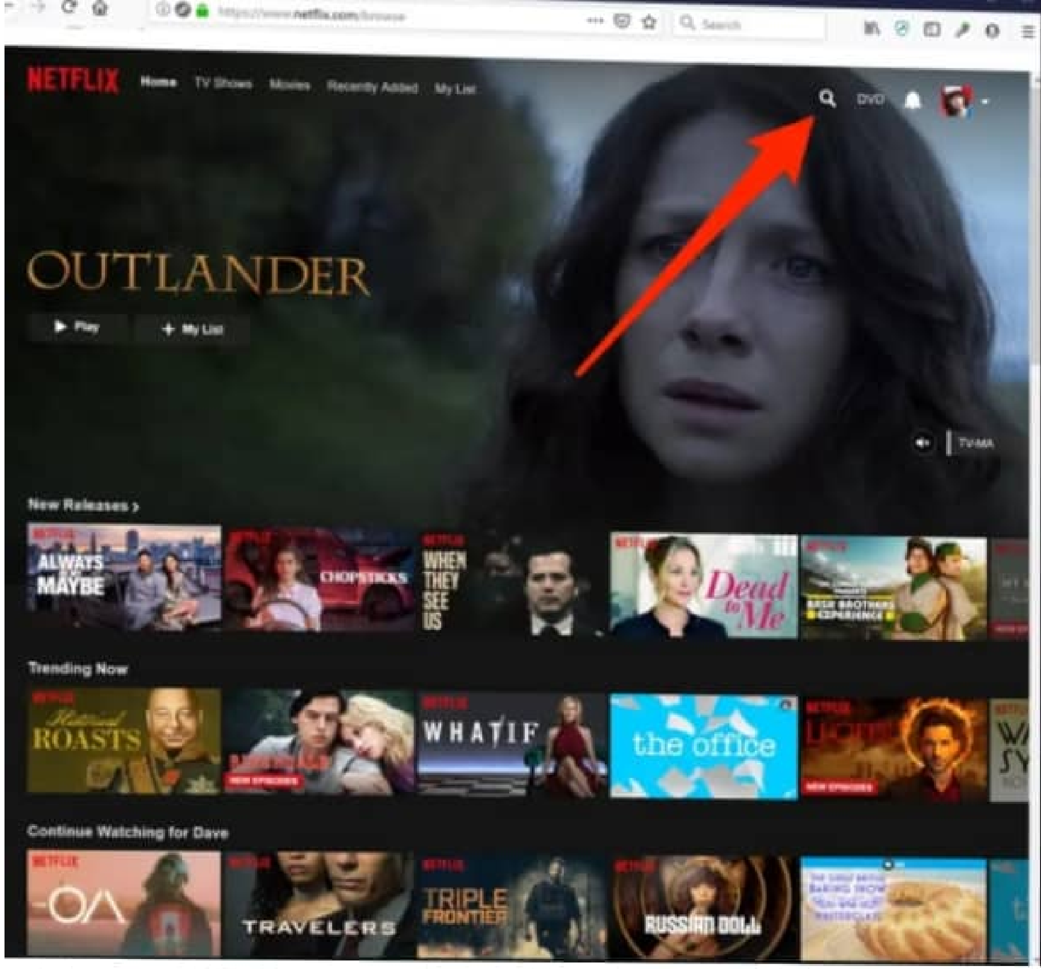


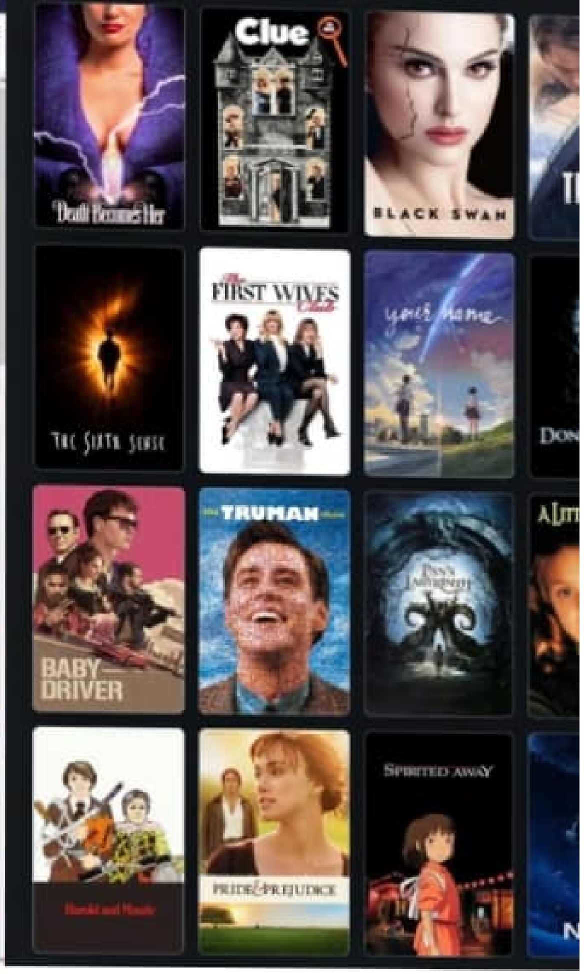
Fig. 1. Share of time spent watching TV-shows per device and age group



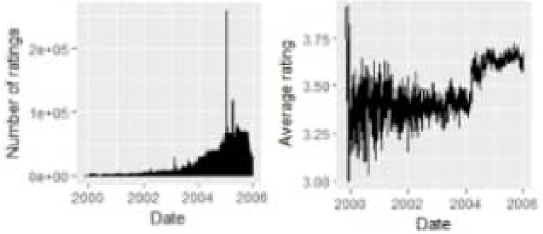
Based on the research of Arwin Halim, etc. with the title "Sistem Rekomendasi Film menggunakan Bisecting K-Means dan Collaborative Filtering". Showing the error rate on the recommendation system has been calculated using the average value of MAE combination of bisecting K-Means and user-based CF is 1.63, lower than the average value of MAE combination of bisecting K-Means and item-based. In addition to the recommended method, the distribution of rating values on the dataset also greatly affects the MAE value. This is also shown in clusters 11 and 17 with uneven distribution of rating values, which will result in a higher error value in the recommendation system [5].

Based on the research of Yessica Putri Santoso, etc. with the title "Implementasi Metode K-Means Clustering pada Sistem Rekomendasi Dosen Tetap Berdasarkan Penilaian Dosen". Shows that of the 70 data tested 39 lecturer data can be recommended as worthy of being a permanent lecturer and 31 lecturer data that is not worthy of being recommended as a permanent lecturer with an accuracy calculation result of 55.67% so it can be concluded that the K-Means algorithm is not suitable for this case [7].





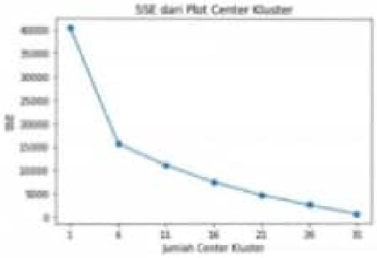
In addition, it is interesting to see if the behaviour of people changes over the years. Therefore, two other histograms have been made to provide more insight into the number of movies rated over time, and the average rating over time. As can be obtained in the graphs below (**Fig. 7**), the number of ratings increases over time, with one big outlier somewhere in 2005. An obvious cause for this is that Netflix sampled its data randomly, so that it would protect user privacy. On the right-hand side one sees the average ratings over the years. It must be noted that the average rating increases over time. Besides, the average rating becomes more stable over time. This can be explained by the fact that there are fewer movies rated in the early 2000's, which causes a higher standard deviation in the average rating.



Then, for the final step, measuring performance through errors using MAE and RMSE, where if the resulting value is close to 0, it means it is close to accurate. The results for collaborative alone get an MAE value of 2.762112297727458 and an RMSE value of 3.773938001358683.

3.3 K-Means Clustering

To get the optimal cluster, it is needed to determine it with the elbow method, where calculations will be made to get a comparison with the Sum of Square Error (SSE) of each cluster. Therefore, the number of clusters that will be taken based on the elbow position. Based on Figure 4, it can be concluded that the number of clusters is 6.



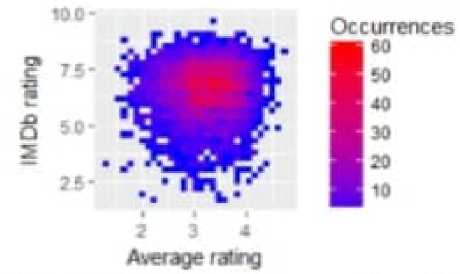
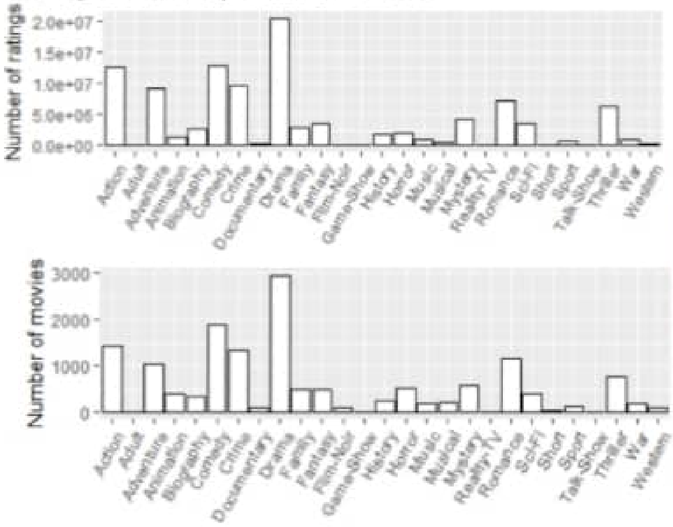
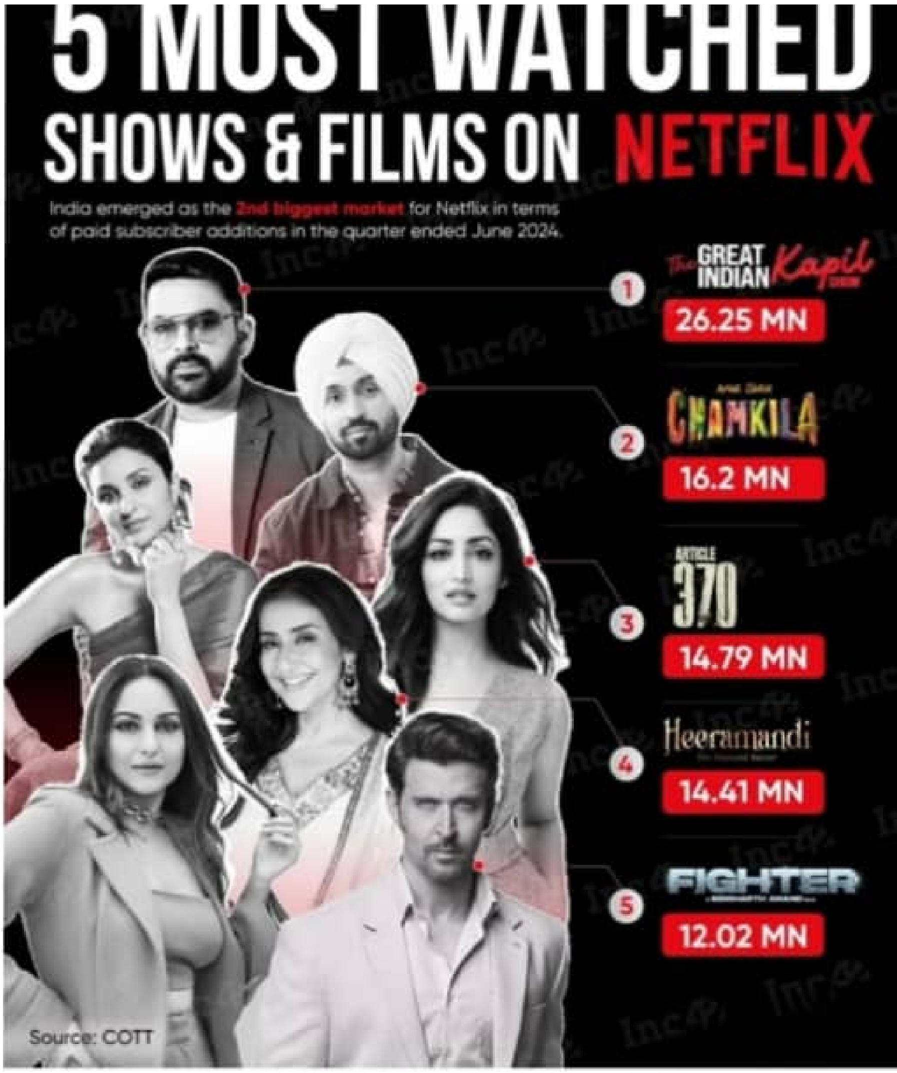


Fig. 5. Scatter plot that shows the correlation between average and IMDb rating of the movies.

Next, a more detailed inspection of the gathered movie genres is done, to check whether the externally retrieved data from IMDb is useful. In order to do a proper analysis, a closer look is taken in the number of ratings and number of movies per genre. The result of this analysis is shown in **Fig. 6**. From these figures, one can state that the most common genres are drama, comedy, action, adventure and crime. In the bottom histogram, the same genres are in the top 5 occurrences of movies.





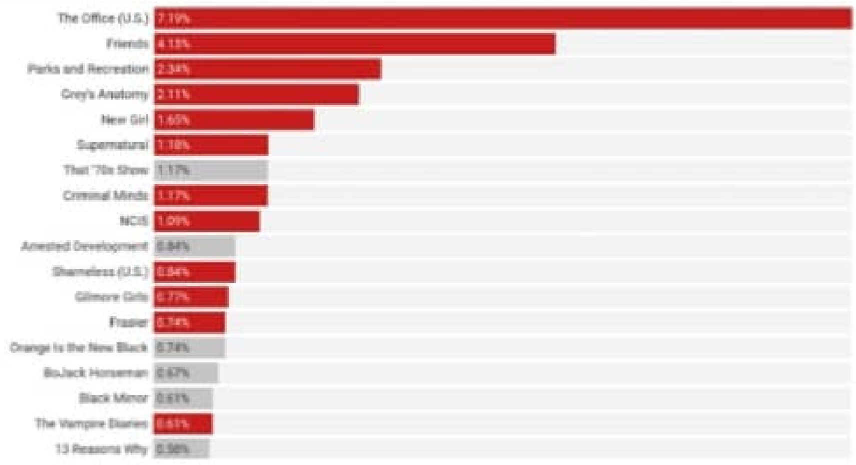
5 MOST WATCHED SHOWS & FILMS ON NETFLIX

India emerged as the 2nd biggest market for Netflix in terms of paid subscriber additions in the quarter ended June 2024.

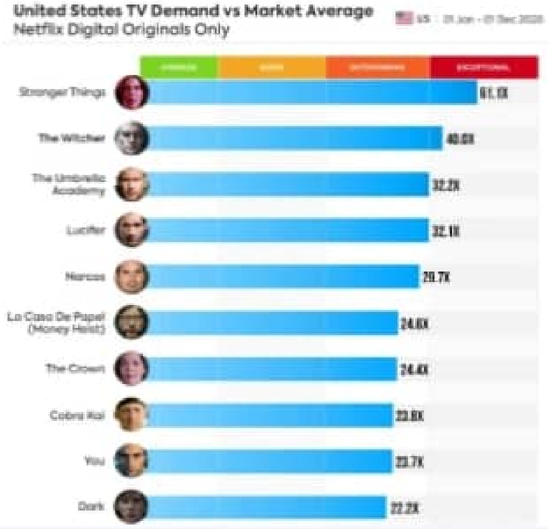
Source: COTT

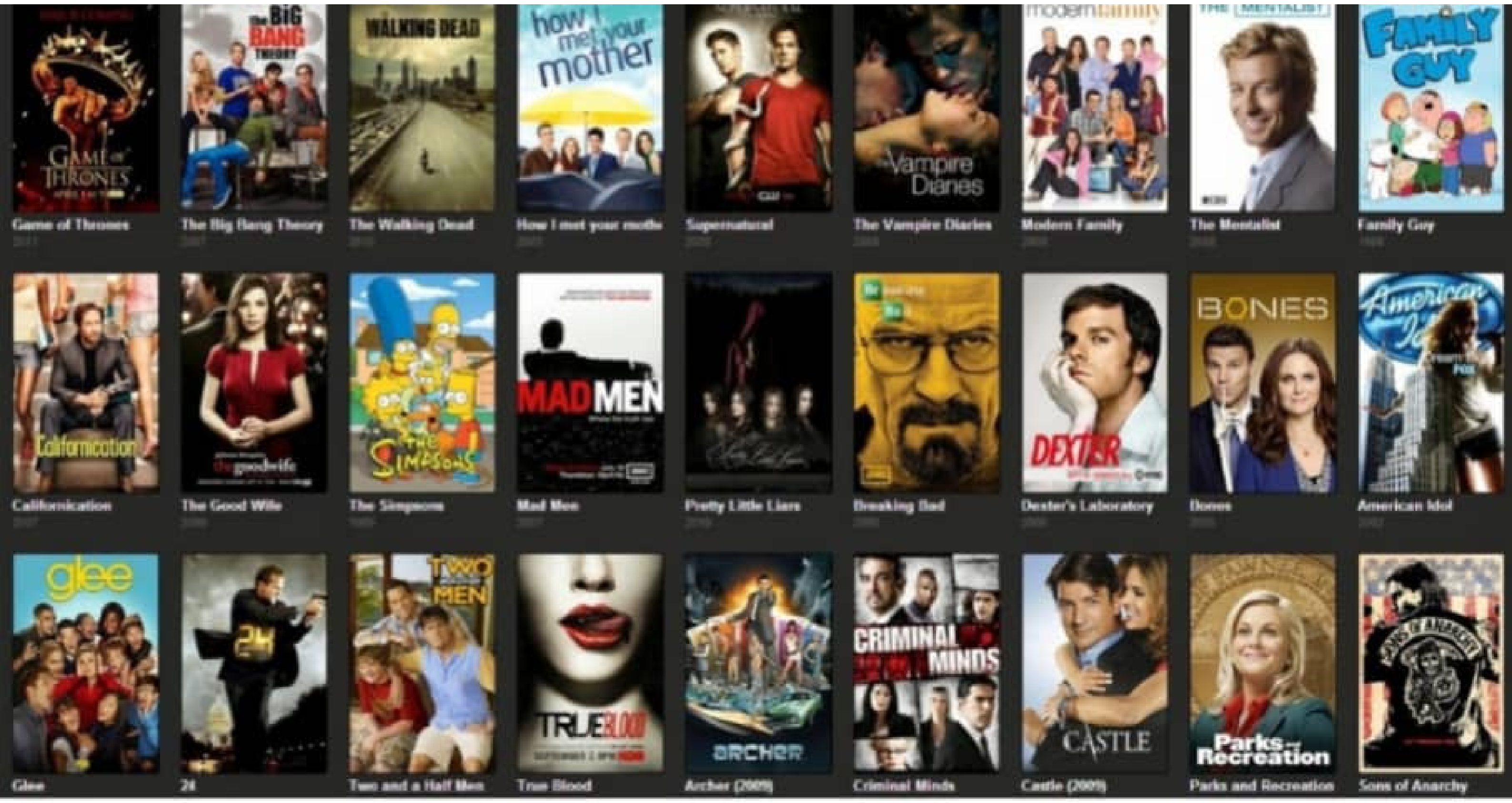
Most viewed Netflix shows, as a percentage of all Netflix views

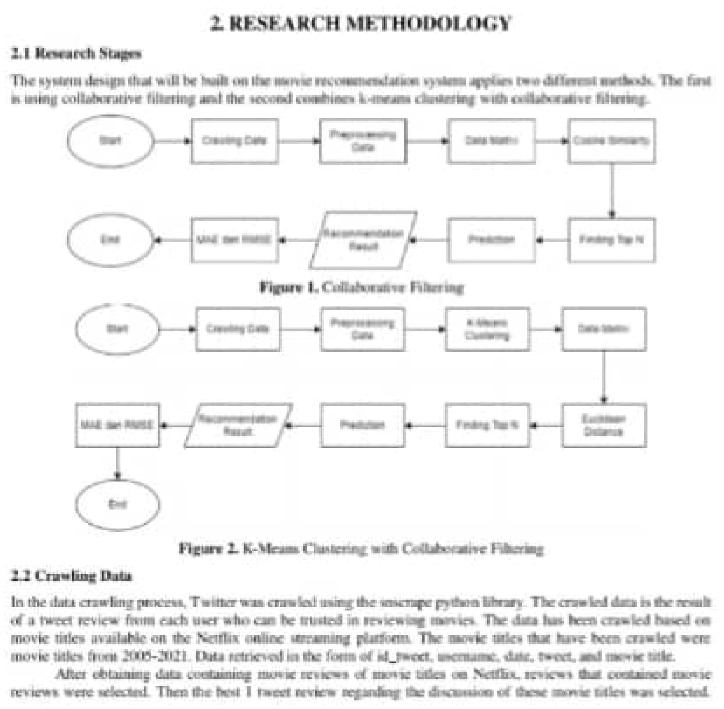
Red bars are for shows that could be taken away since they are owned by Disney, Fox, WarnerMedia or NBCU



United States TV Demand vs Market Average
Netflix Digital Originals Only







2. RESEARCH METHODOLOGY

2.1 Research Stages

The system design that will be built on the movie recommendation system applies two different methods. The first is using collaborative filtering and the second combines k-means clustering with collaborative filtering.

Start
Crawling Data
Preprocessing Data
Data Matrix
Cosine Similarity
Finding Top N
Prediction
Recommendation Result
MAE dan RMSE
End

Figure 1. Collaborative Filtering

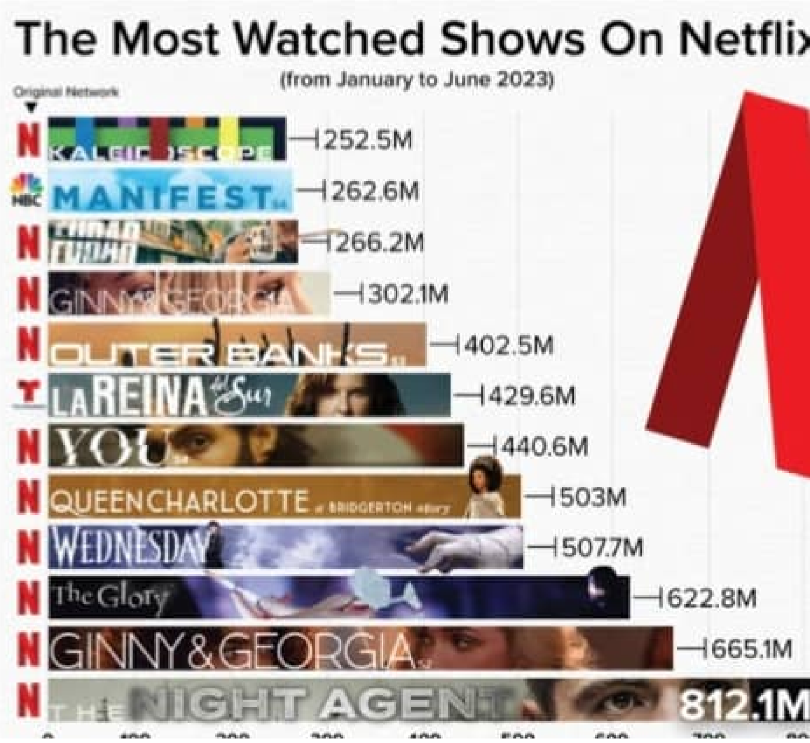
Start
Crawling Data
Preprocessing Data
K-Means Clustering
Data Matrix
Euclidean Distance
Finding Top N
Prediction
Recommendation Result
MAE dan RMSE
End

Figure 2. K-Means Clustering with Collaborative Filtering

2.2 Crawling Data

In the data crawling process, Twitter was crawled using the snscrape python library. The crawled data is the result of a tweet review from each user who can be trusted in reviewing movies. The data has been crawled based on movie titles available on the Netflix online streaming platform. The movie titles that have been crawled were movie titles from 2005-2021. Data retrieved in the form of id_tweet, username, date, tweet, and movie title.

After obtaining data containing movie reviews of movie titles on Netflix, reviews that contained movie reviews were selected. Then the best 1 tweet review regarding the discussion of these movie titles was selected.



6 Data analysis

Before one is able to make an initial recommendation based on historic ratings, one must get more insight of the data. Therefore, a data analysis is done to get more acquainted and familiar with the reduced and selected data.

To start, the distribution of ratings is important for the recommender system: it is essential for such a system that there is some kind of diversity in this distribution. As can be obtained in the first histogram below (**Fig. 4**), most common ratings are 3 or 4 stars. Besides, the second histogram on the right shows the average rating, which is between 2.5 and 4 stars.

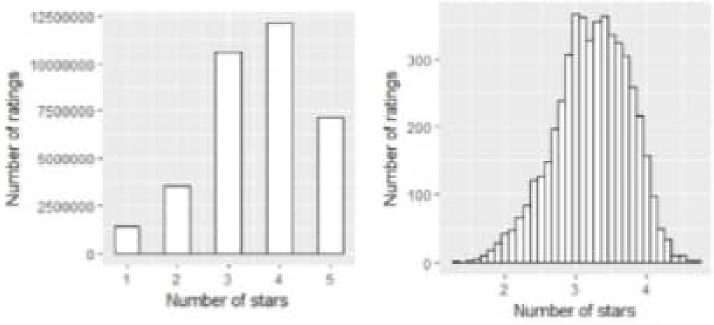
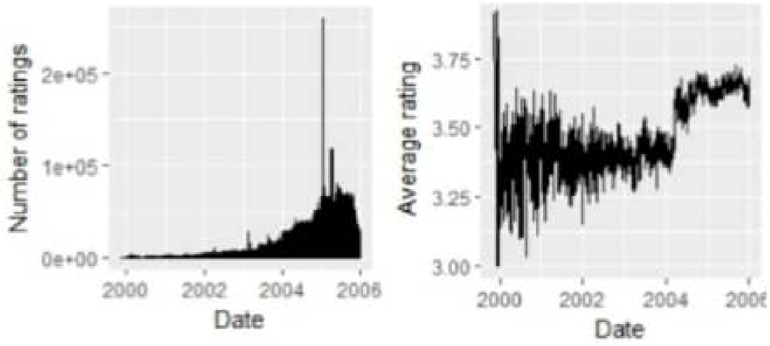
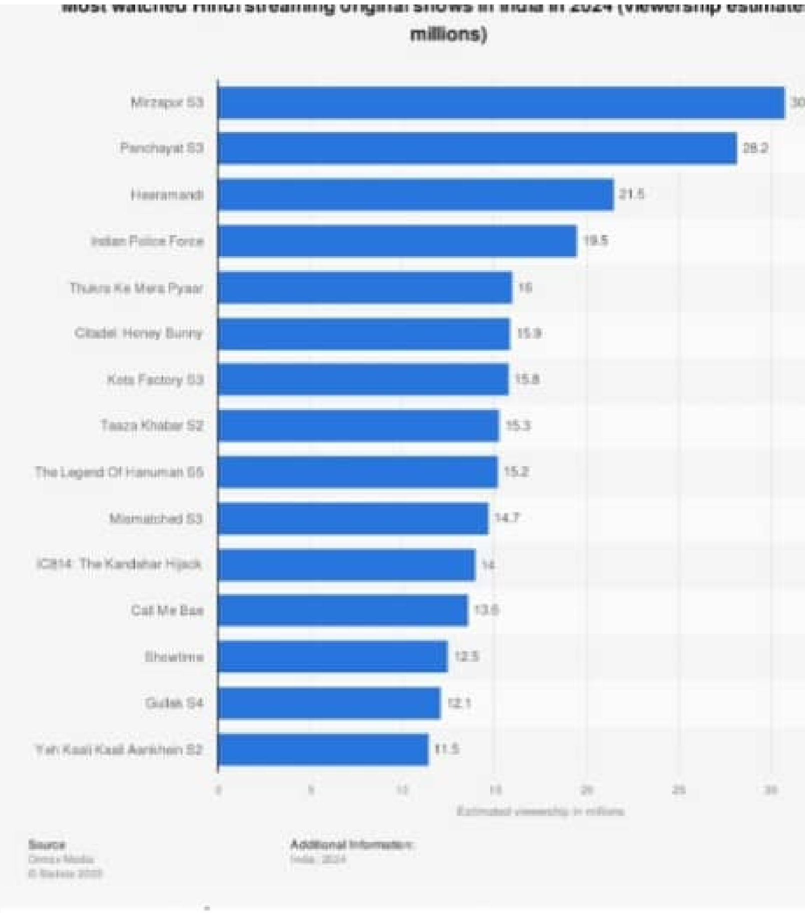


Fig. 4. Histograms that shows the distribution of the ratings (left) and average rating (right)

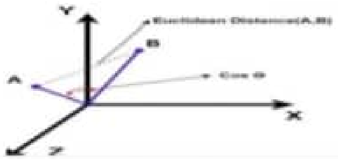
Furthermore, it is important to check the correlation between the average rating in the original dataset and the retrieved IMDb rating of the movies. In the scatter plot below (**Fig. 5**) the average rating of every movie is plotted against the IMDb rating of the corresponding movies. As can be concluded from this plot, many movies have a comparable IMDb rating, but there seems no clear correlation between these features in the movie dataset.

In addition, it is interesting to see if the behaviour of people changes over the years Therefore, two other histograms have been made to provide more insight into the number of movies rated over time, and the average rating over time. As can be obtained ir the graphs below (**Fig. 7**), the number of ratings increases over time, with one big outlier somewhere in 2005. An obvious cause for this is that Netflix sampled its data randomly, so that it would protect user privacy. On the right-hand side one sees the averag ratings over the years. It must be noted that the average rating increases over time Besides, the average rating becomes more stable over time. This can be explained by the fact that there are fewer movies rated in the early 2000's, which causes a highe standard deviation in the average rating.





Collaborative filtering is one of the methods used in recommendation systems that are used based on interactio between users and stored items which will be used to create a recommendation system [14]. Collaborative filteri is divided into two types user-based collaborative filtering and item-based collaborative filtering. User-bas collaborative filtering is a method that provides item recommendations by comparing all items across all users obtain Top-N user similarity [15]. Meanwhile, item-based collaborative filtering is a method that provides ite recommendations by looking for other items that have Top-N similarity to other items. [15]. To get Top-N us similarity and Top-N item similarity, we use the cosine similarity method. Cosine similarity is a measure used determine the similarity between two items, systematically the cosine angle between two vectors in thr dimensions. [16].



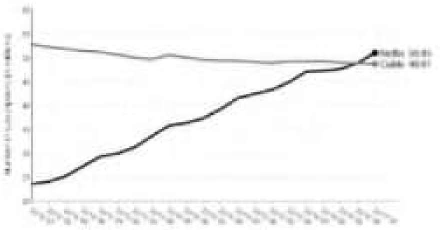
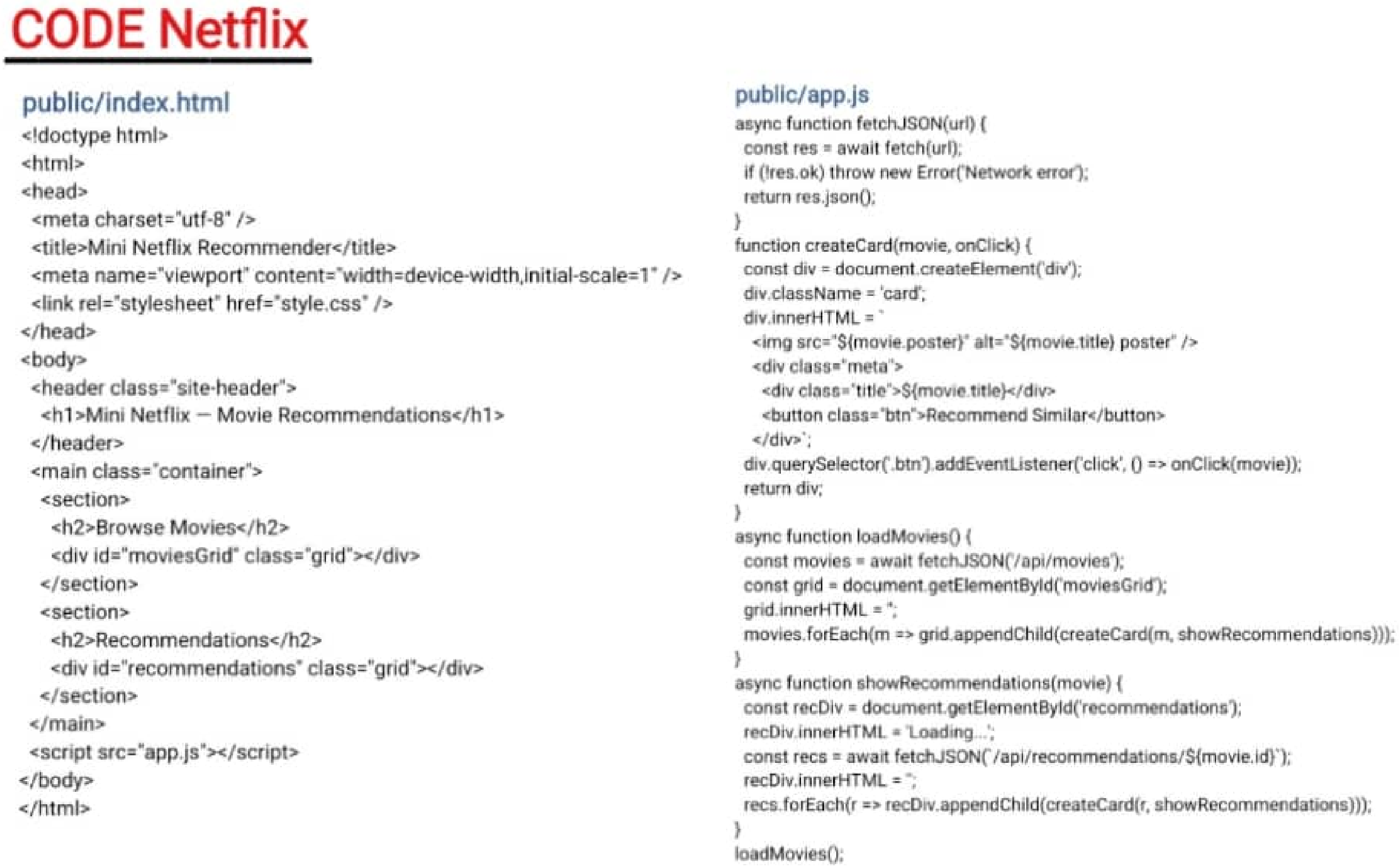


Fig. 2. Number of subscriptions in the United States

From these numbers, one can conclude that Netflix collects a lot of data, which can be used in many ways. For example, they can analyze data to increase the revenue, for marketing purposes, and to improve their customer satisfaction.



CODE Netflix

public/index.html

```html
<!doctype html>
<html>
<head>
 <meta charset="utf-8" />
 <title>Mini Netflix Recommender</title>
 <meta name="viewport" content="width=device-width,initial-scale=1" />
 <link rel="stylesheet" href="style.css" />
</head>
<body>
 <header class="site-header">
  <h1>Mini Netflix — Movie Recommendations</h1>
 </header>
 <main class="container">
  <section>
   <h2>Browse Movies</h2>
   <div id="moviesGrid" class="grid"></div>
  </section>
  <section>
   <h2>Recommendations</h2>
   <div id="recommendations" class="grid"></div>
  </section>
 </main>
 <script src="app.js"></script>
</body>
</html>
```

public/app.js

```javascript
async function fetchJSON(url) {
 const res = await fetch(url);
 if (!res.ok) throw new Error('Network error');
 return res.json();
}
function createCard(movie, onClick) {
 const div = document.createElement('div');
 div.className = 'card';
 div.innerHTML = `
  <img src="${movie.poster}" alt="${movie.title} poster" />
  <div class="meta">
   <div class="title">${movie.title}</div>
   <button class="btn">Recommend Similar</button>
  </div>`;
 div.querySelector('.btn').addEventListener('click', () => onClick(movie));
 return div;
}
async function loadMovies() {
 const movies = await fetchJSON('/api/movies');
 const grid = document.getElementById('moviesGrid');
 grid.innerHTML = '';
 movies.forEach(m => grid.appendChild(createCard(m, showRecommendations)));
}
async function showRecommendations(movie) {
 const recDiv = document.getElementById('recommendations');
 recDiv.innerHTML = 'Loading...';
 const recs = await fetchJSON(`/api/recommendations/${movie.id}`);
 recDiv.innerHTML = '';
 recs.forEach(r => recDiv.appendChild(createCard(r, showRecommendations)));
}
loadMovies();
```

4. CONCLUSION

Based on research that has been done by combining k-means clustering with collaborative filtering and collaborative filtering only, it is used for recommendation systems. By using a dataset obtained from Twitter by crawling data and added with ratings from IMDb, Rotten Tomatoes, and Metacritic. Which resulted in a dataset with 35 users, 785 movie titles, and 6184 reviews. Then preprocessing the data with text processing, polarity, and labeling. And get the dataset that will be used for this experiment. After that, testing the dataset by combining k-means clustering with collaborative filtering and collaborative filtering only. It was found that the rating prediction generated from k-means clustering with collaborative filtering has a greater result than collaborative filtering only, which is 2.8466. Then the MAE and RMSE values generated by k-means clustering with collaborative filtering are smaller with the resulting MAE value of 0.5029 and for the resulting RMSE of 0.6354 which can be interpreted as better than collaborative filtering only, because accuracy/performance can be seen from the average value of MAE and RMSE errors. If the value is closer to 0, the better the accuracy/performance obtained. Therefore, it can be concluded that k-means clustering with collaborative filtering has better results than collaborative filtering only. Therefore, it is hoped that future research can improve the performance of the recommendation system with a