EAS 509

Movie Clustering for Recommendation

University at Buffalo The State University of New York

Problem Statement

- The problem is the limitation of collaborative filtering techniques in movie recommendation systems to suggest movies to new users and those outside a user's usual preferences.

Objective

- To explore the use of clustering algorithms for grouping movies based on their attributes, such as genre, ratings, and other features.
- To build a movie recommendation system that suggests movies to users based on their preferences and the similarity of the movies in the same cluster.
- To evaluate the performance of different clustering algorithms in the movie recommendation system using various metrics such as silhouette score, elbow method, etc.

Need for Recommendation System

- Increasingly popular due to the large amount of available movie data and the need to personalize recommendations to improve user satisfaction.
- Collaborative filtering is a widely used technique in movie recommendation systems that uses user behavior and preferences to make recommendations.
- Content-based filtering is another popular approach that uses movie attributes to recommend movies to users based on their preferences.
- Hybrid approaches that combine collaborative and content-based filtering have also been proposed to improve recommendation accuracy.

Unsupervised Learning Techniques

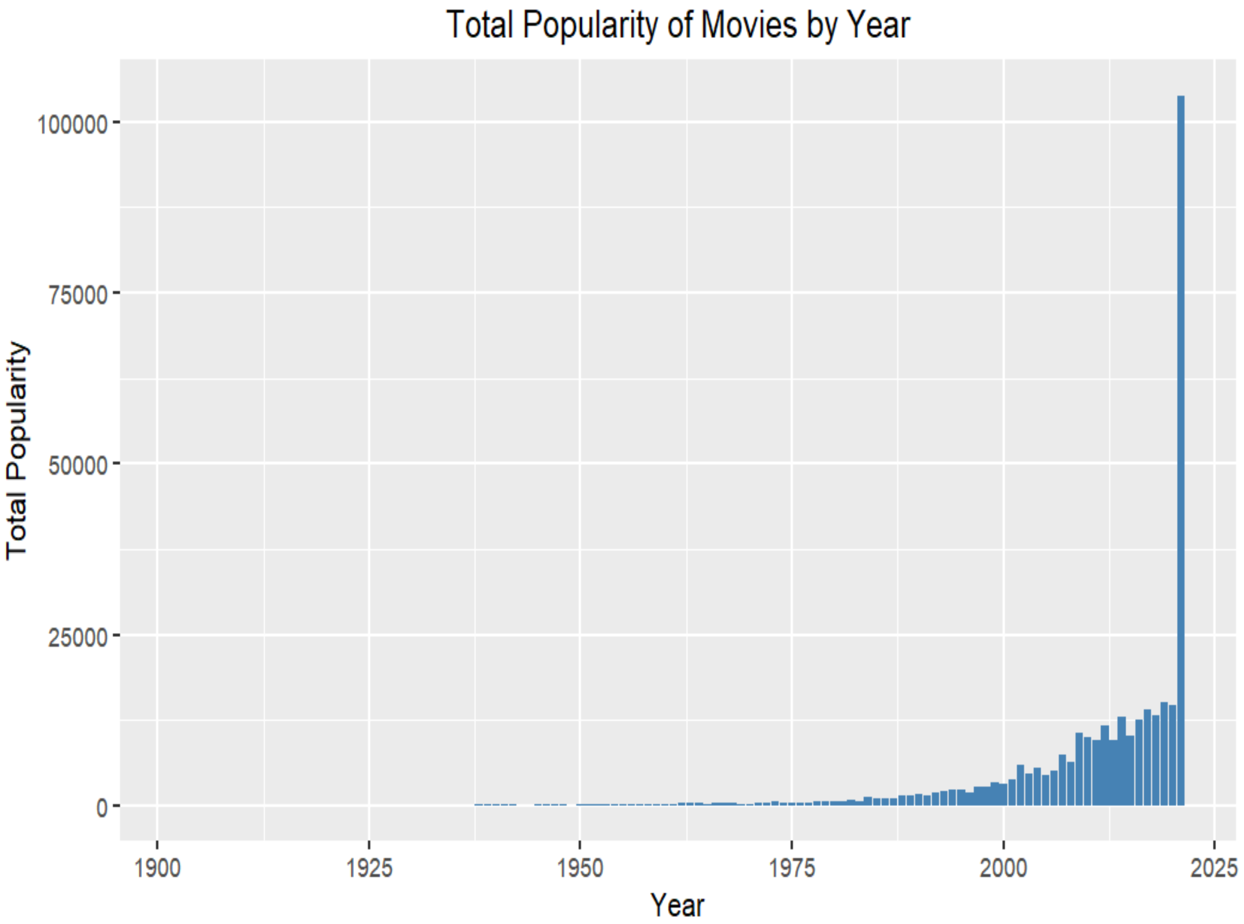
- Unsupervised learning algorithms are used in movie recommendation systems to analyze patterns and relationships in the data without the need for labeled training data.
- These algorithms include k-means clustering, hierarchical clustering, DBSCAN clustering, and principal component analysis (PCA).
- These algorithms are particularly useful in movie recommendation systems as they can identify patterns and relationships that may not be immediately apparent.

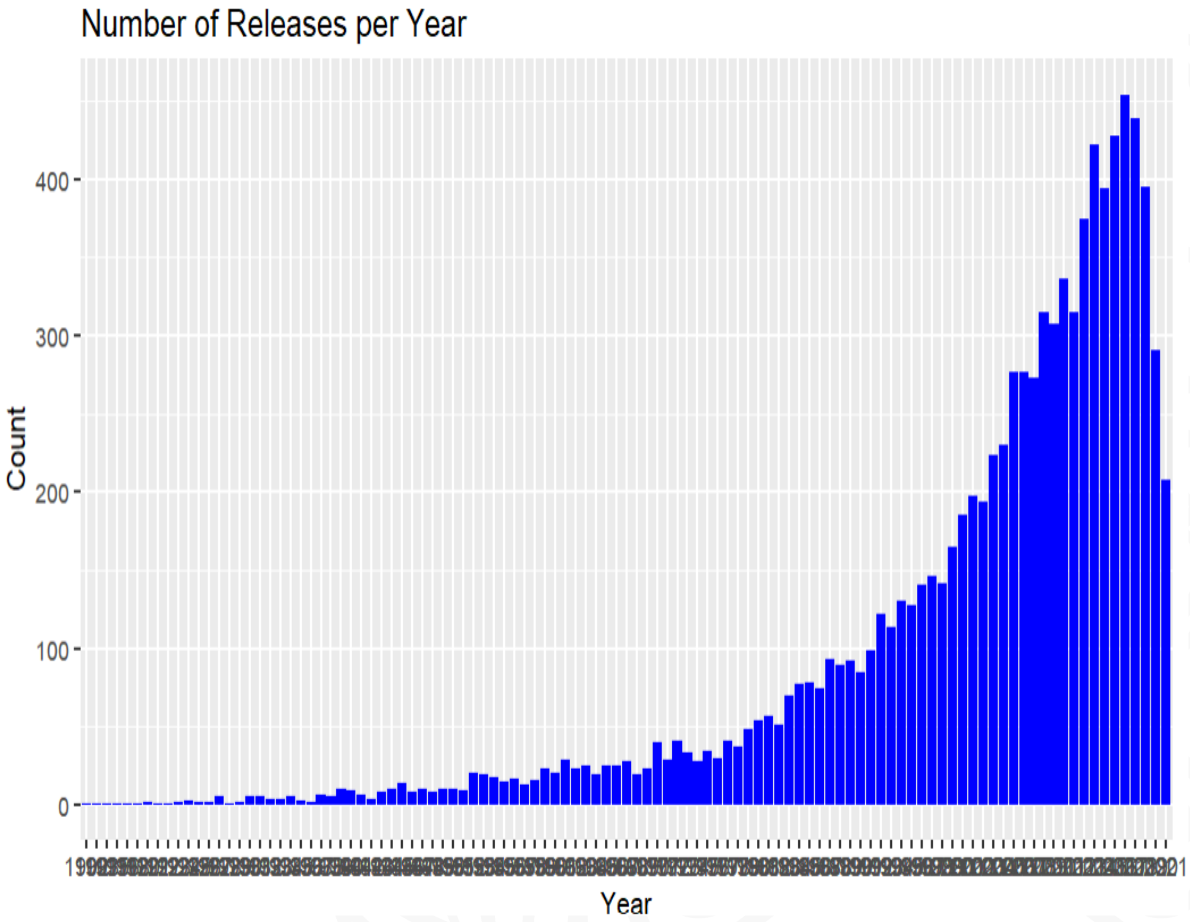
Data Description

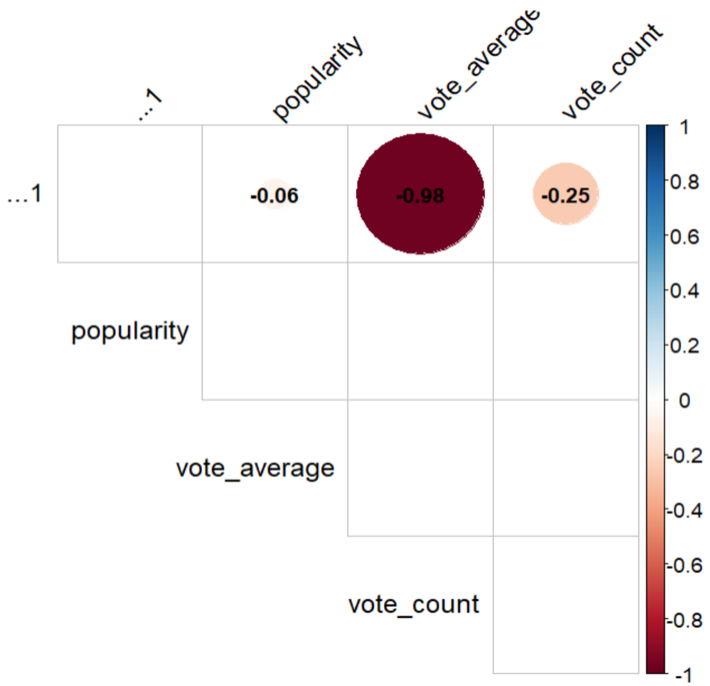
- The data has 8 columns with 9463 records. It has movie details released from 1902 to 2021.

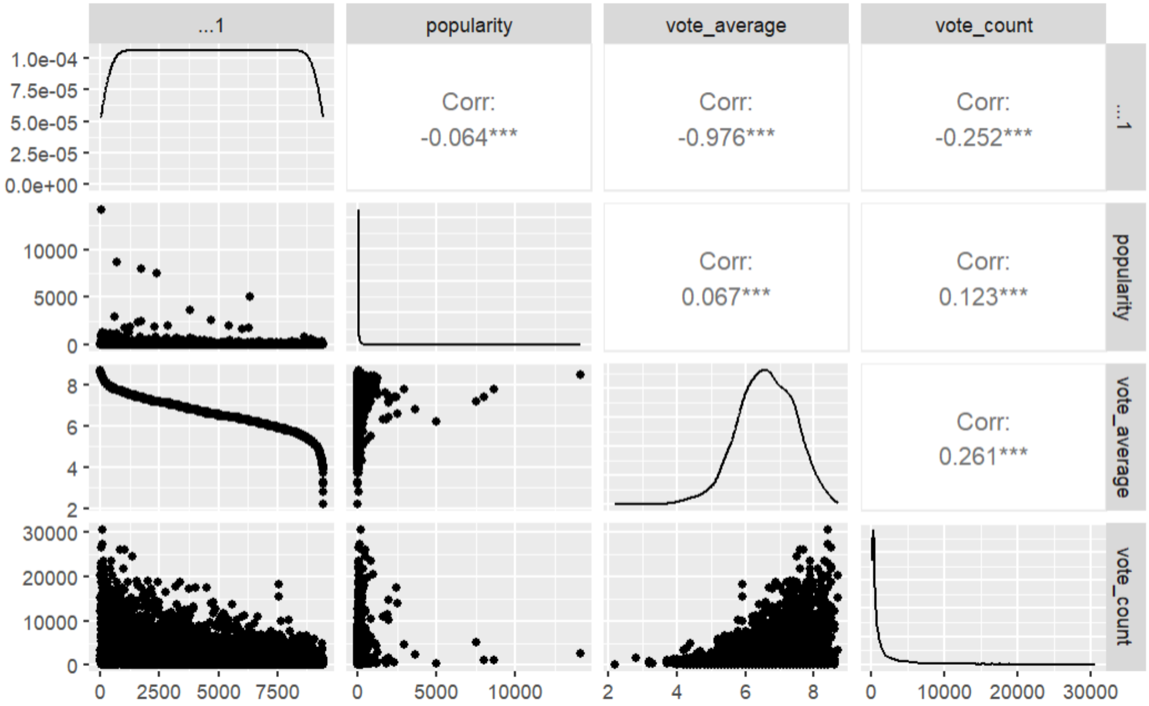
Column Name	**Description**
id	ID of the movie
title	Movie title
release_date	Date of release of the movie
overview	General overview of the movie
popularity	Movie popularity
vote_average	Average vote received
vote_count	Total vote count
video	Video

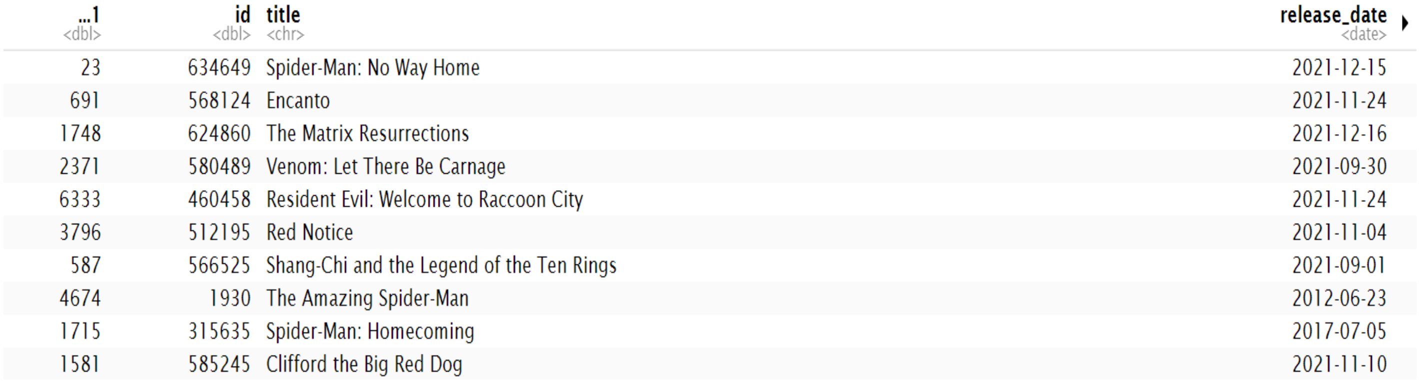
Exploratory data analysis











...1 <dbl>	id <dbl>	title <chr>	release_date <date>
23	634649	Spider-Man: No Way Home	2021-12-15
691	568124	Encanto	2021-11-24
1748	624860	The Matrix Resurrections	2021-12-16
2371	580489	Venom: Let There Be Carnage	2021-09-30
6333	460458	Resident Evil: Welcome to Raccoon City	2021-11-24
3796	512195	Red Notice	2021-11-04
587	566525	Shang-Chi and the Legend of the Ten Rings	2021-09-01
4674	1930	The Amazing Spider-Man	2012-06-23
1715	315635	Spider-Man: Homecoming	2017-07-05
1581	585245	Clifford the Big Red Dog	2021-11-10

Clustering

K-means clustering

- ➢Partition-based clustering algorithm that divides the dataset into k clusters, where k is a user-defined parameter.

- ➢The algorithm iteratively assigns each data point to the cluster with the closest centroid, and then recalculates the centroids based on the mean of the data points in each cluster.

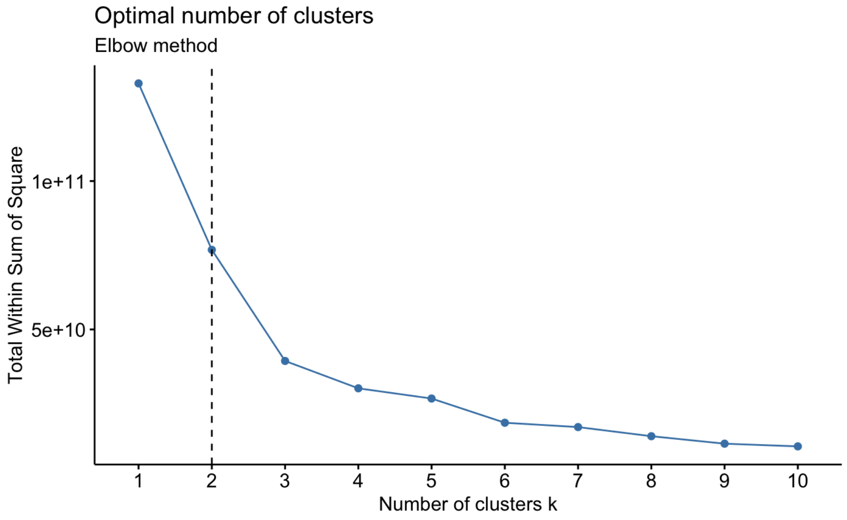
- ➢K-means clustering is simple, fast, and easy to implement, but it requires the number of clusters to be specified in advance.

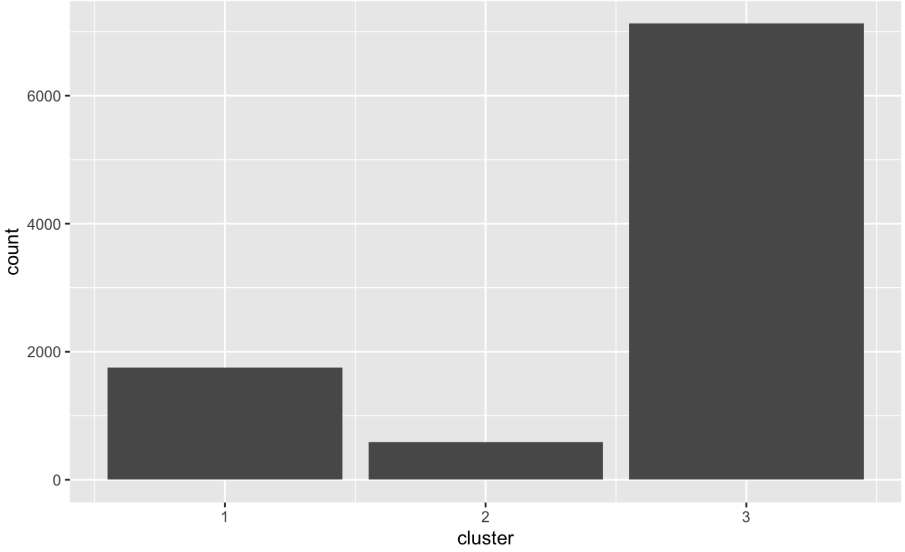
K-means clustering

- ➢From the elbow graph, although the suggested optimal number of clusters are 2, we considered 3 clusters for K means clustering.

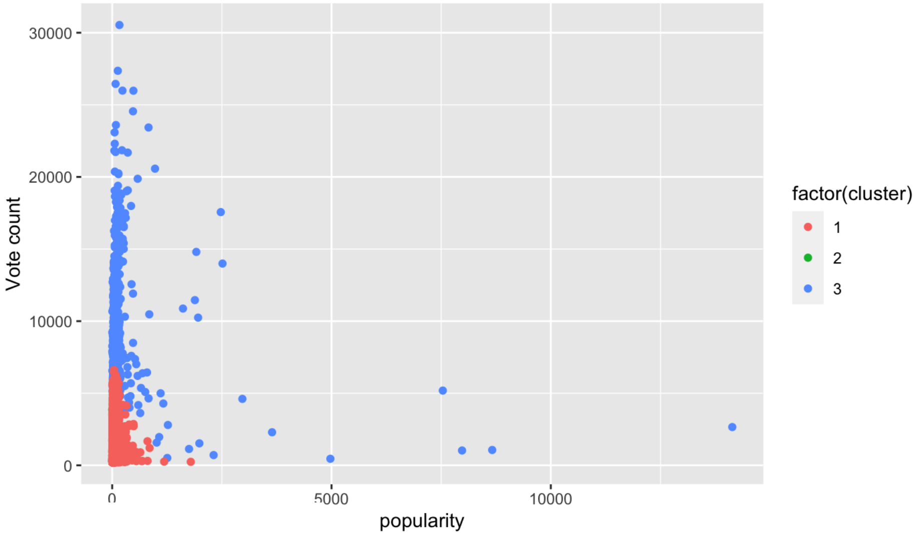
- ➢After clustering, we can observe that cluster 3 has most number of movies followed by cluster 1.

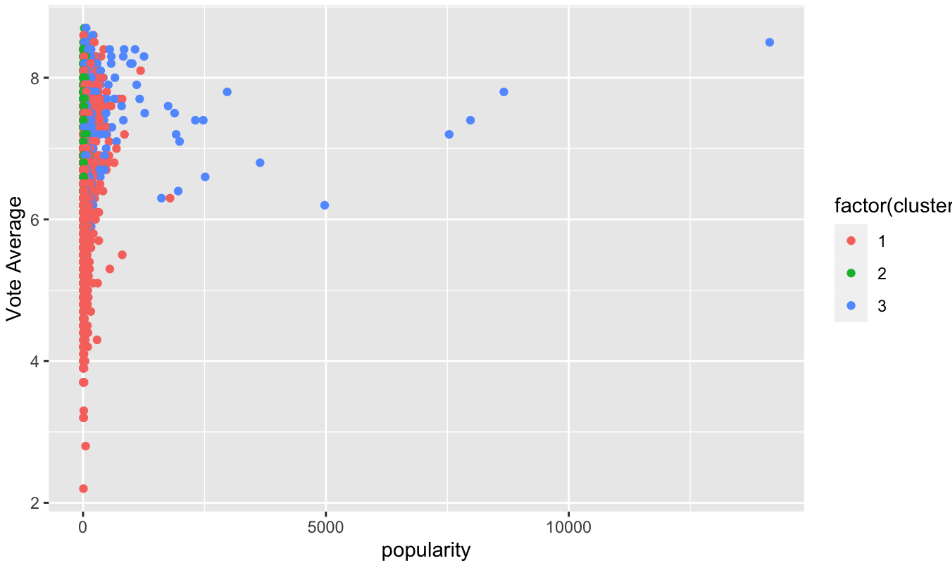
- ➢ We evaluated the clustering with between cluster sum of squares. We got 14410.69

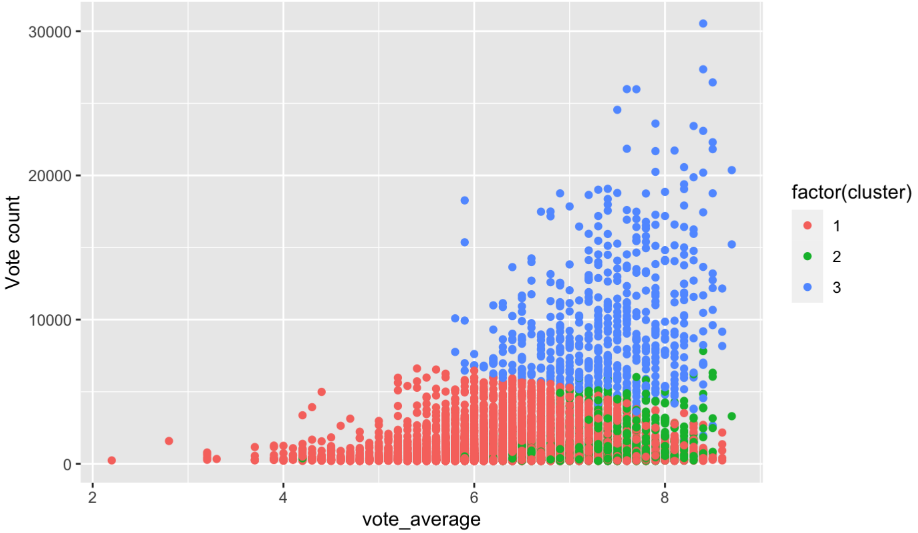


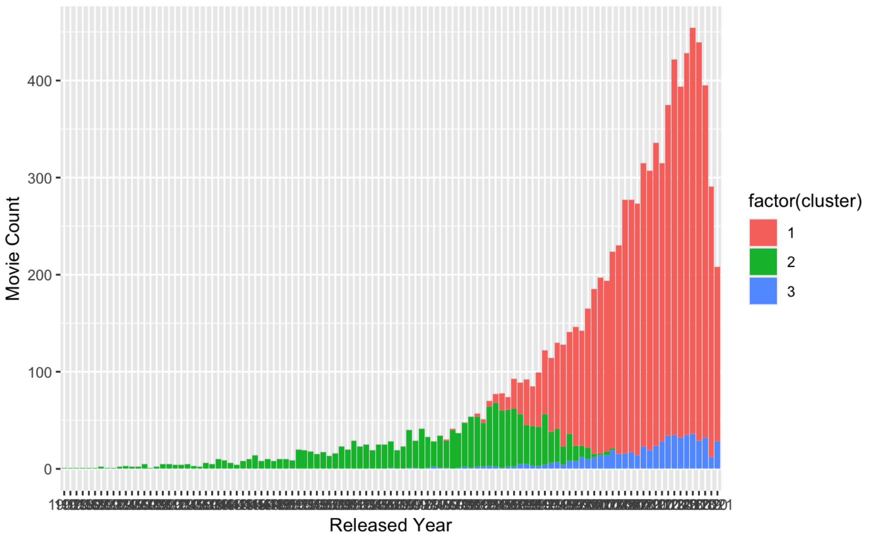


K-means clustering


Clustered movies by popularity and Vote count


Clustered movies by popularity and Vote Average


Clustered movies by popularity and Vote count



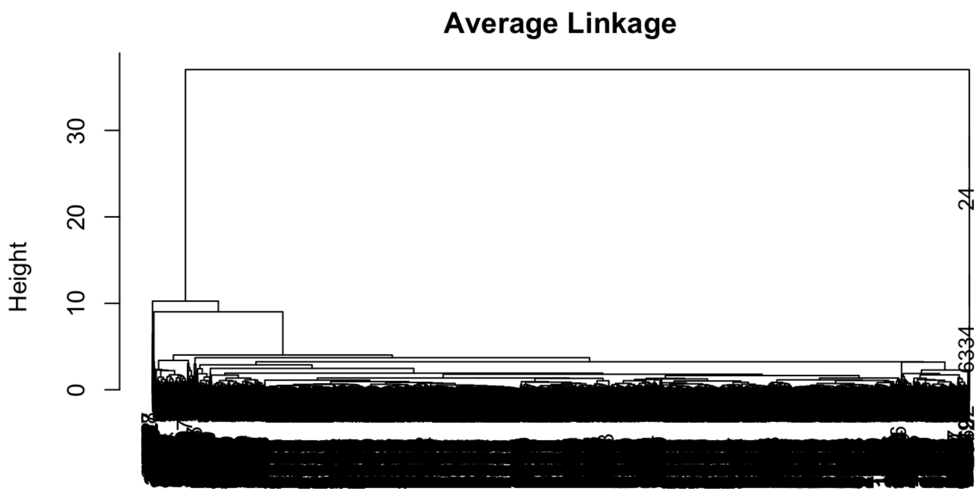
Hierarchical clustering

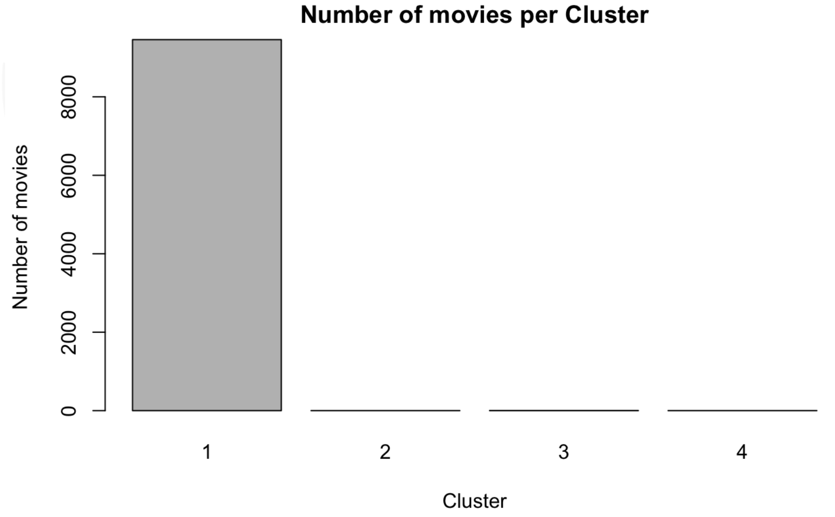
- Hierarchical clustering is another commonly used clustering algorithm in movie recommendation systems.
- It is a hierarchical clustering algorithm that creates a tree-like structure of clusters based on the similarity between data points.
- The algorithm starts with each data point in its own cluster and then merges clusters iteratively until all data points are in a single cluster.
- Hierarchical clustering can be either agglomerative (bottom-up) or divisive (top-down), and it does not require the number of clusters to be specified in advance.

Hierarchical clustering

- From the dendrogram, we used average linkage and selected number of clusters as 4.

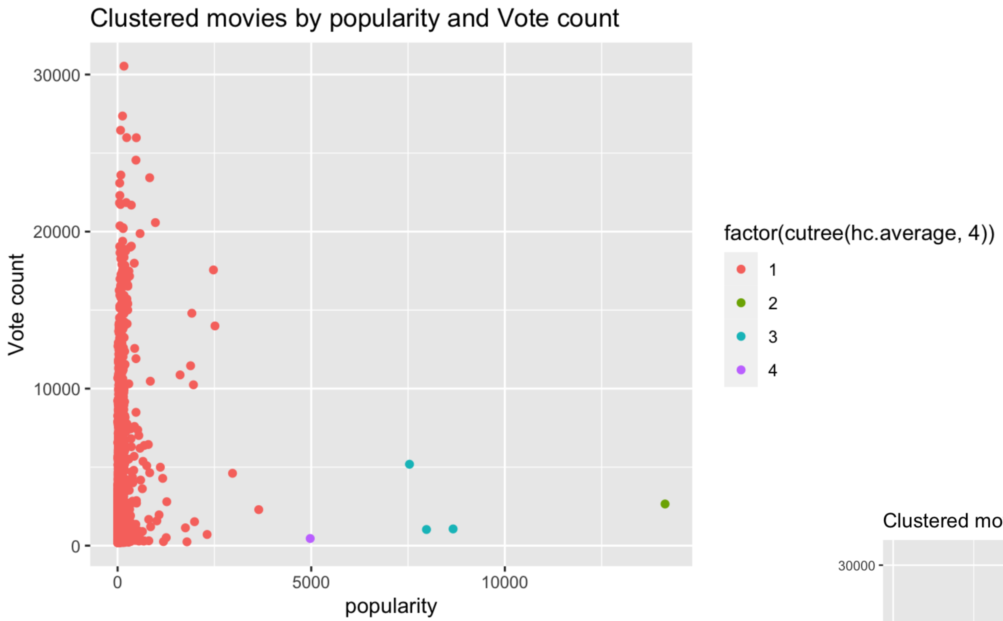
- After clustering, we can observe that cluster 1 has most number of movies followed by cluster 1.

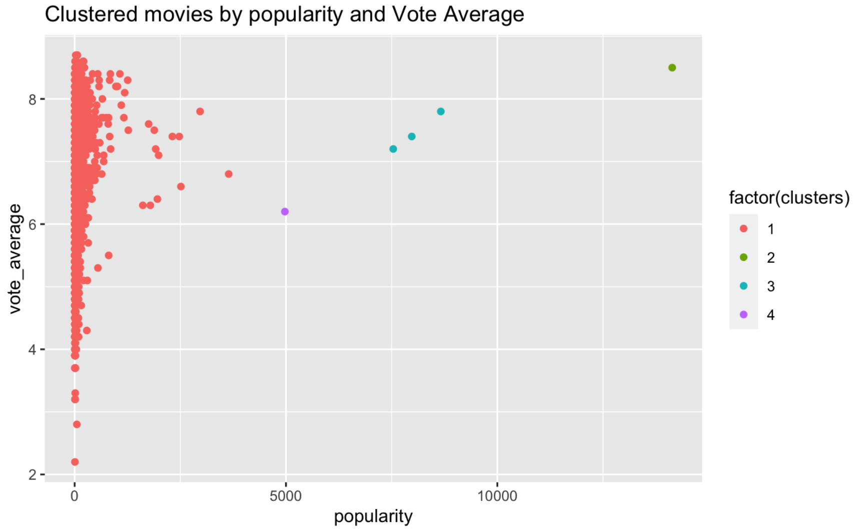


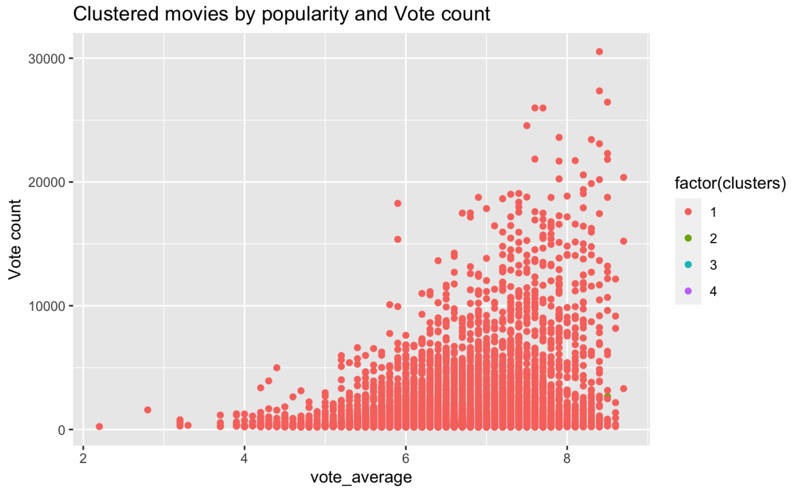


University at Buffalo The State University of New York

Hierarchical clustering







Thank You