CLUSTERING ANALYSIS ON NETFLIX MOVIES AND TV SHOWS

Karan Tiwari

Abstract

Netflix is the most widely used media and video streaming platform. It has around 200 million users World Wide. It offers a large library of movies and TV series that can be accessed at any time via internet services. Netflix works on a subscription based model , where users get unlimited access to content with a paid subscription and users can cancel their subscription whenever they want that’s why Netflix always needs to provide the best content to retain their subscriber because of this, it's essential to have a recommendation system that gives customers helpful suggestions.

Problem Statement

This dataset consists of TV shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine.

In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The

streaming service’s number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same dataset.

In this project, you are required to do

1: Exploratory Data Analysis .

2: Understanding what type content is available in different countries .

3: Is Netflix has increasingly focusing on TV rather than movies in recent years.

4: Clustering similar content by matching text-based features **.**

Objective

The project's main goal is to do Exploratory Data Analysis and implement a model that can perform Clustering on comparable material by matching text-based attributes.

Data Description

The dataset has 7787 rows and 12 attributes to work with.

1. We have null values in the dataset.
2. Changed the format of the Date.
3. Added some columns which are extracted from the Date column.

Attribute Information:

1. show_id : Unique ID for every Movie / TV Show
2. type : Identifier - A Movie or TV Show

3. title : Title of the Movie / TV Show
4. director : Director of the Movie
5. cast : Actors involved in the movie / show
6. country : Country where the movie / show was produced
7. date_added : Date it was added on Netflix
8. release_year : Actual Release Year of the movie / show
9. rating : TV Rating of the movie / show
10. duration : Total Duration - in minutes or number of seasons
11. listed_in : Genre
12. description: The Summary description

Challenges

The following are the challenges faced in the data analysis:

Tackling with null value
Conversion of Date time features
Feature engineering
Model Implementation

Approach

As per the problem statement, First we will do EDA to get business insights and then we have to cluster similar content by matching text features ,So Text processing is required than for clustering we will use K-means Clustering.

Tools Used

This project is performed on Google collaboratory and Python was used throughout the entire project. The following libraries were imported for data analysis and visualization:.

- Pandas: Extensively used to load and wrangle with the dataset.

- Matplotlib: Used for visualization.
- Seaborn: Used for visualization.
- Datetime: Used for analyzing the date variable.
- Warnings: For filtering and ignoring the warnings.
- Numpy: For some math operations in predictions.
- Sklearn: For the purpose of analysis and prediction.

Feature Engineering

I compile the text features like country**,** listed_in, description, rating_TV, type and cast in one column called combined_text_feature.

Text Processing

Steps involved

1: Remove punctuations

2: Remove Stopwords

3: Convert text into lower case

4: Stemming:

The process of reducing a word to its stem that affixes to suffixes and prefixes or to the roots of words known as "lemmas" is known as stemming. Text processing relies heavily on stemming

.

5: Vectorization

TF-IDF(Term frequency-inverse document frequency)

The phrase frequency-inverse document frequency statistic measures the significance of a word to a document in a corpus or

collection. In information retrieval, text mining, and user modeling searches, it is frequently employed as a weighting factor. To account for the fact that certain words are used more frequently than others overall, the TF-IDF value rises according to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the term.

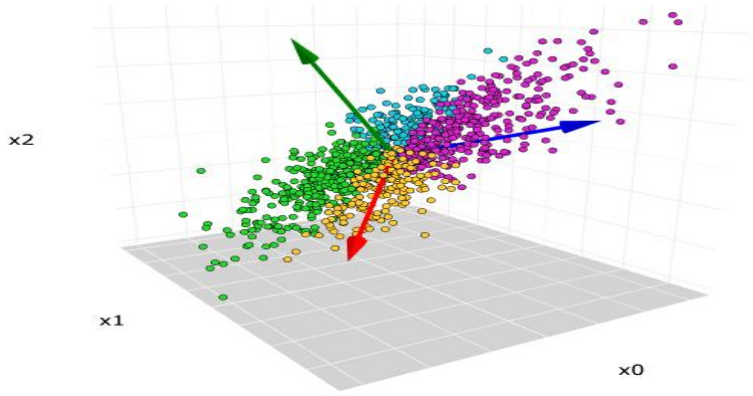
$$TF(t,d) = \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d}$$

$$IDF(t) = log\frac{N}{1+df}$$

$$TF - IDF(t,d) = TF(t,d) * IDF(t)$$

PCA (Principal Component Analysis)

PCA is a statistical approach that allows you to summarize the information contained in huge data tables, it reduces the dimensionality by converting large set of variables into a smaller one which still contains most of the information of the large set.



Building a clustering model

Clustering models enable you to group records into a set number of clusters. This can aid in the identification of natural groups in your data.

Clustering models are concerned with detecting groups of similar records and categorizing the records accordingly. This is done without any prior knowledge of the groupings or their features. Indeed, you might not even know how many organisations to look for. Clustering methods differ from other machine-learning approaches in that there is no specified output or target field for the algorithm to forecast. Because there is no external benchmark against which to measure the model's classification performance, these models are frequently referred to as unsupervised learning models.

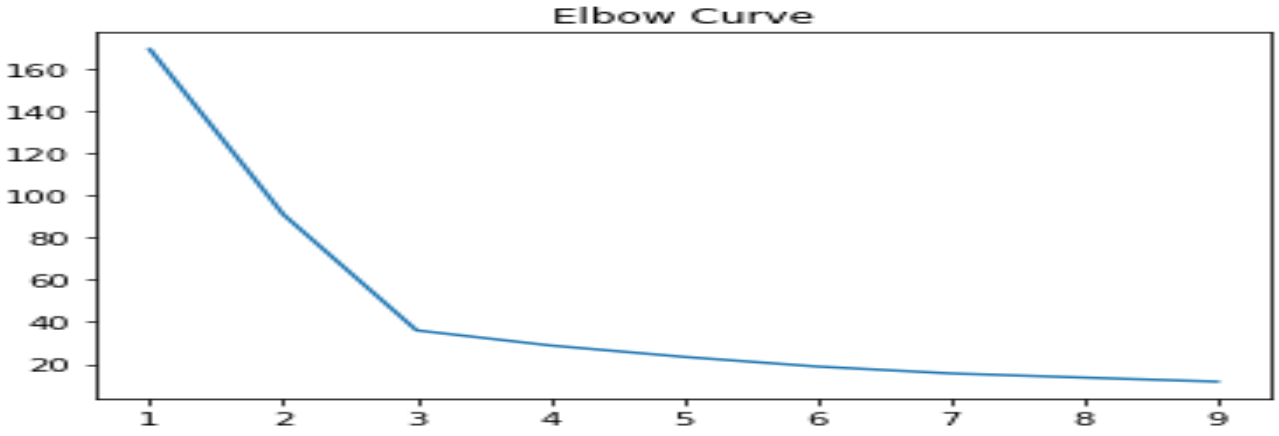
Optimal number of Cluster

Optimal number of cluster can be get through

1 Elbow curve

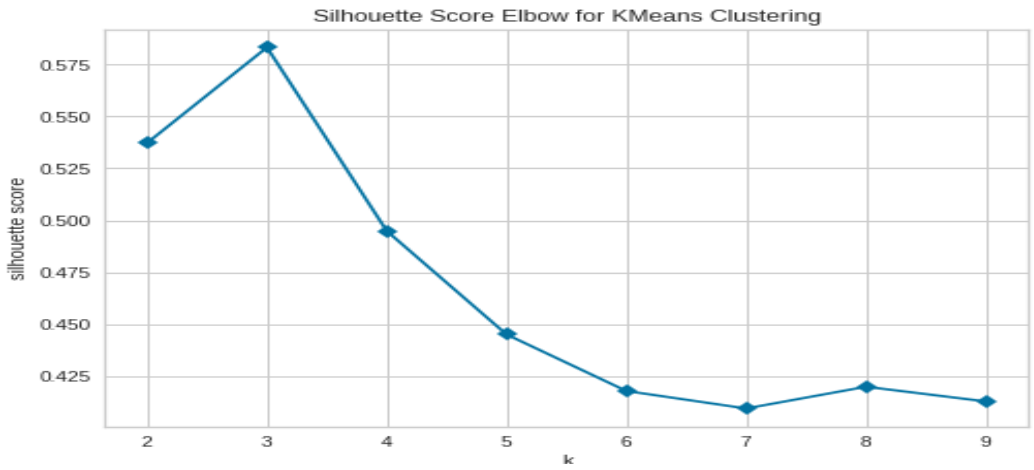
The Elbow Method is an empirical technique for determining how many clusters are best for a dataset. With this technique, we choose a range of potential k values, and then we perform K-Means clustering using each of the potential k values. Calculate the average separation between each point in a cluster and its

centroid, then plot that information. Choose the value of k at which the average distance abruptly decreases. The average SSE falls as the number of clusters (k) rises.



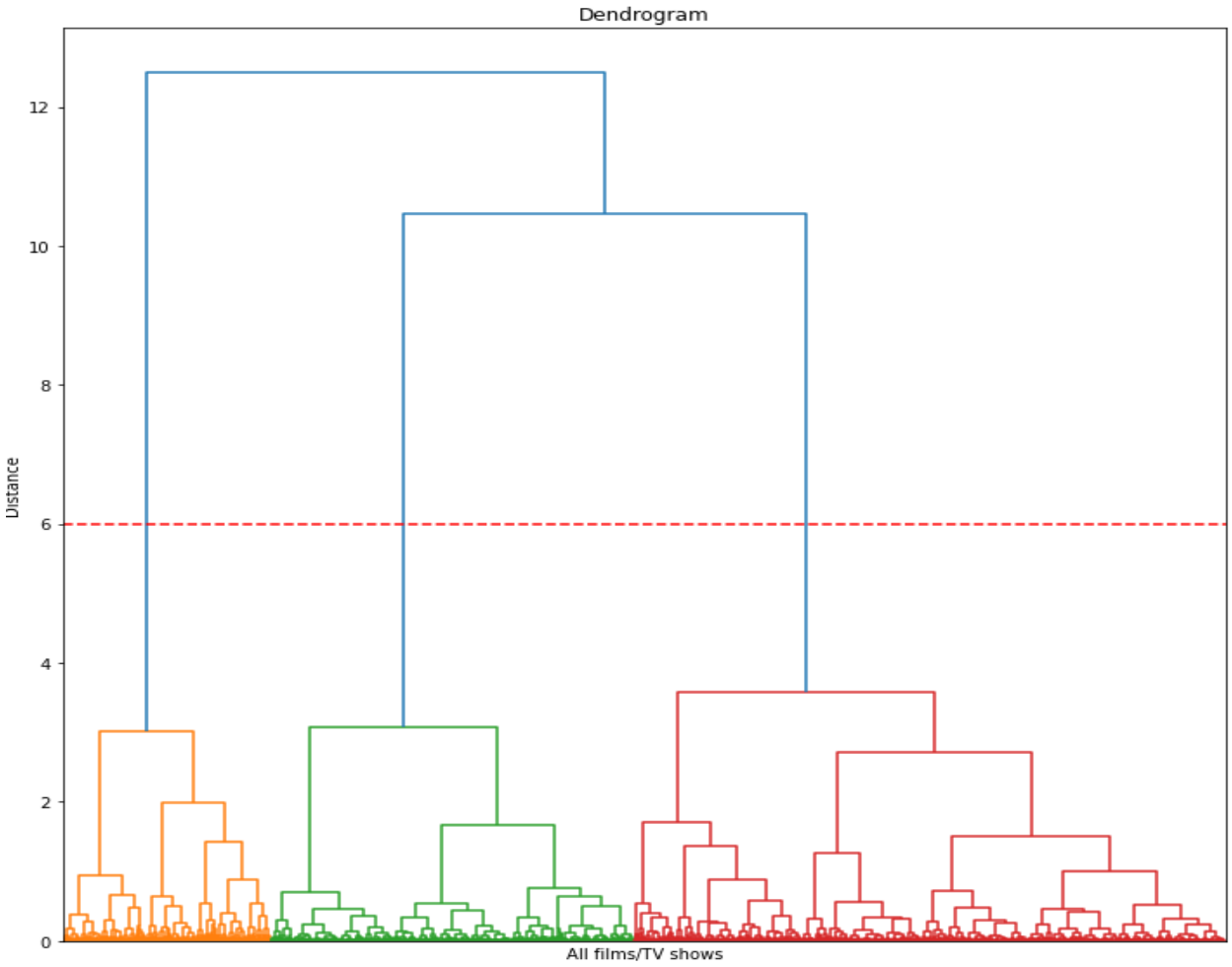
2 Silhouette Score

The silhouette value quantifies how similar an object is to its own cluster (cohesion) when compared to other clusters (separation). The silhouette has a value between 1 and +1, with a high value indicating that the item is well matched to its own cluster but poorly matched to nearby clusters. If the majority of the items have a high value, the clustering setup is acceptable. If a large number of points have a low or negative value, the clustering configuration may have too many or too few clusters.



3 Dendrogram

A dendrogram is a diagram that depicts the object's hierarchical connection. It is typically produced as a result of hierarchical clustering. A dendrogram is mostly used to determine the best approach to assign things to clusters.

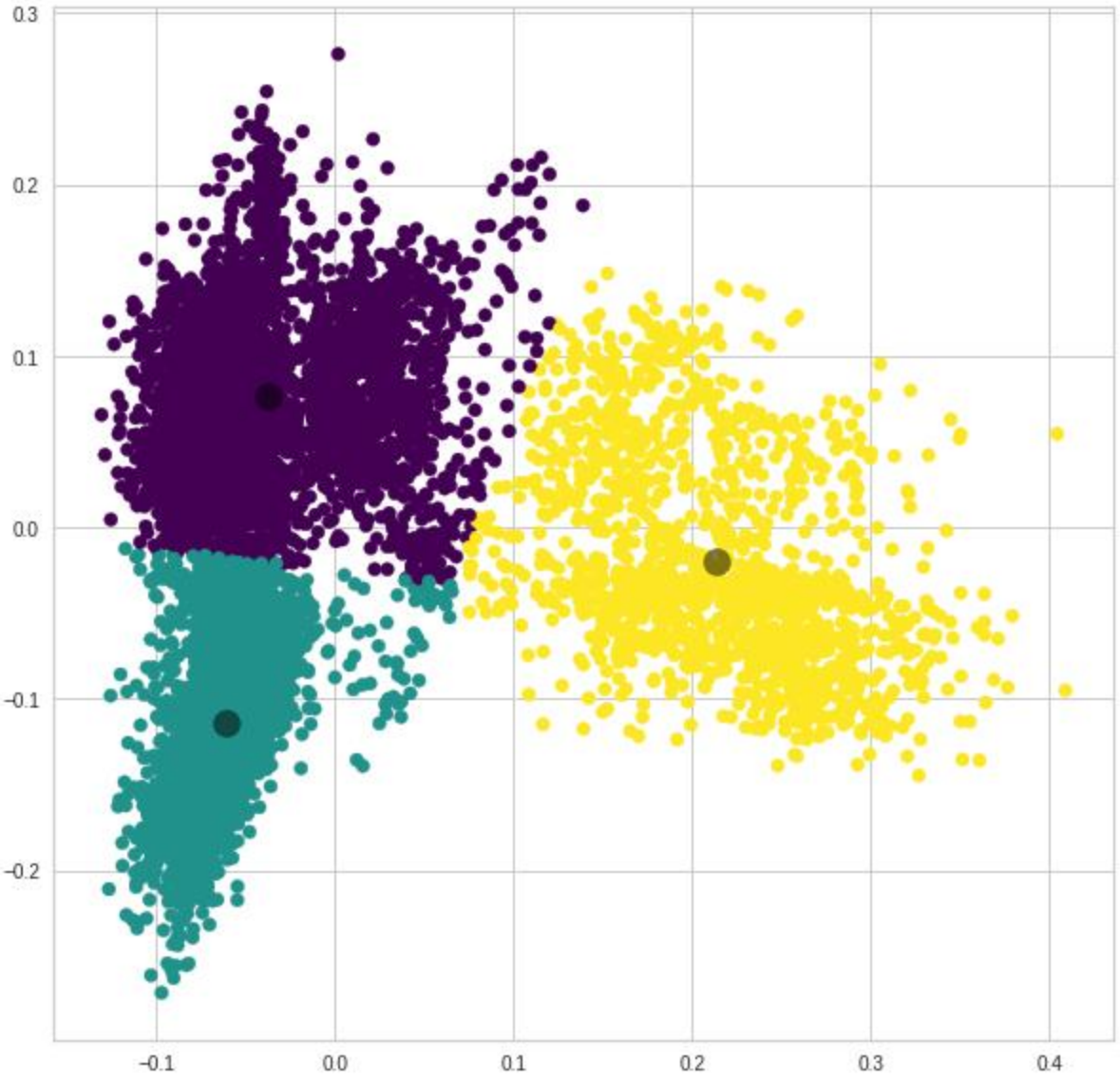


Assume we cut vertical lines with a horizontal line to obtain the number of clusters. Number of clusters = 3

Model Implementation

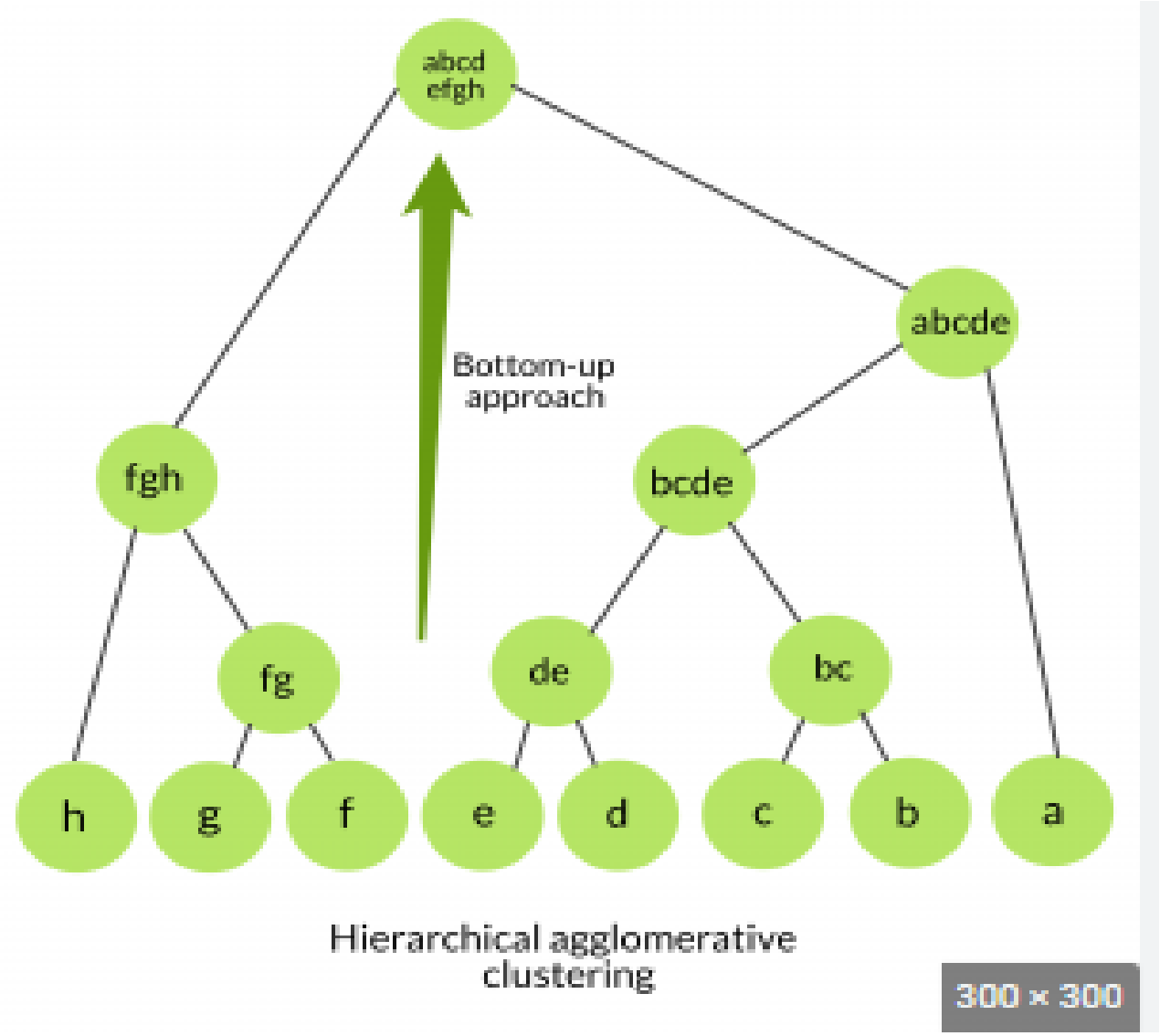
- **K-means Clustering**

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.



- **Agglomerative Clustering**

Agglomerative clustering is the most common type of hierarchical clustering, and it is used to group objects into clusters based on their similarity. The algorithm starts by treating each object as a singleton cluster. Then, one by one, pairs of clusters are merged until all clusters have been merged into one large cluster containing all objects. The dendrogram that results is a tree-based representation of the objects.

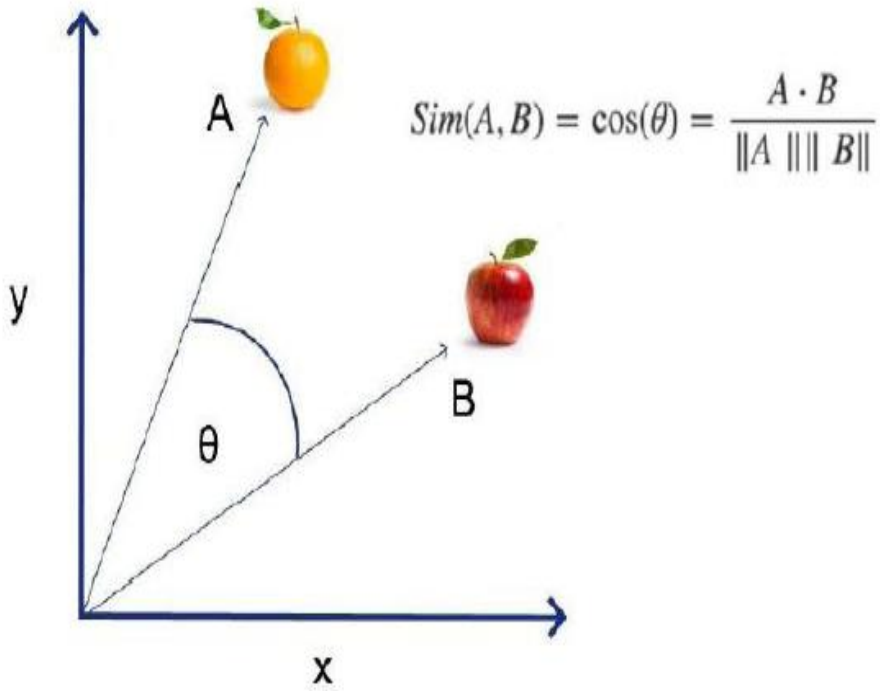


Recommendation System

We obtained recommendations using Cosine similarity.

Cosine similarity

Cosine similarity is a statistic for determining how similar two vectors are. It specifically assesses the similarity in vector direction or orientation while ignoring variations in size or scale. Both vectors must be in the same inner product space, which means they must create a scalar via inner product multiplication. The cosine of the angle between two vectors is used to calculate their similarity.



Conclusion

That's all! We had completed our exercise.

Starting with data loading, we've done EDA, null value handling, encoding of category columns, feature selection, clustering, and finally a recommendation system.