Topic : Netflix Movies and TV shows Clustering

Mihir Kulkarni, Ram Bakale

Data Science Trainees

Alma Better

Abstract:

Netflix is a company that manages a large collection of TV shows and movies, streaming it anytime via online. This business is profitable because users make a monthly payment to access the platform. However, customers can cancel their subscriptions at any time. Therefore, the company must keep the users hooked on the platform and not lose their interest. This is where recommendation systems start to play an important role, providing valuable suggestions to users is essential.

Introduction:

Netflix's recommendation system helps them increase their popularity among service providers as they help increase the number of items sold, offer a diverse selection of items, increase user satisfaction, as well as user loyalty to the company, and they are very helpful in getting a better understanding of what the user wants. Then it's easier to get the user to make better decisions from a wide variety of movie products. With over 139 million paid subscribers (total viewer pool -300 million) across 190 countries, 15,400 titles across its regional libraries and 112 Emmy Award Nominations in 2018 — Netflix is the world's leading Internet television network and the most-valued largest streaming service in the world. The amazing digital success story of Netflix is incomplete without the mention of its

recommender systems that focus on personalization. There are several methods to create a list of recommendations according to your preferences. You can use (Collaborative-filtering) and (Content-based Filtering) for recommendation.

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.

Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

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 increasingly focusing on TV rather than movies in recent years.
4. Clustering similar content by matching text-based features

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

- Importing the libraries

```
import pandas as pd

import numpy as np

import missingno as msno

import seaborn as sns

import matplotlib.pyplot as plt

import plotly.express as px

import plotly.graph_objects as go
```

```
from plotly.subplots import make_subplots
import matplotlib.cm as cm
%matplotlib inline
# Importing Date & Time util modules
from dateutil.parser import parse
import collections as c
import re
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
#for nlp
from sklearn import preprocessing
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.model_selection import train_test_split, KFold
from nltk.stem.snowball import SnowballStemmer
from sklearn.metrics import silhouette_score
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_samples
import scipy.cluster.hierarchy as sch
import warnings
warnings.filterwarnings('ignore')
# Importing Date & Time util modules
```

```
from dateutil.parser import parse

# Xplotter

from xplotter.insights import *

from xplotter.formatter import format_spines
```

Loading the data

```
from google.colab import drive
drive.mount('/content/drive')
# Importing Dataset
    file_path =pd.read_csv('/content/drive/MyDrive/NETFLIX MOVIES AND
    TV SHOWS CLUSTERING.csv')
 # top 5 rows of the given dataset
data.head()

# bottom 5 rows of the given dataset

data.tail()
```

Preprocessing the dataset

In the real world the data has a lot of missing values and it is due to data corruption or failure to record the data. For that purpose it is very important to handle the missing values. Also many machine learning algorithms do not support missing values, that's why we check missing values first.

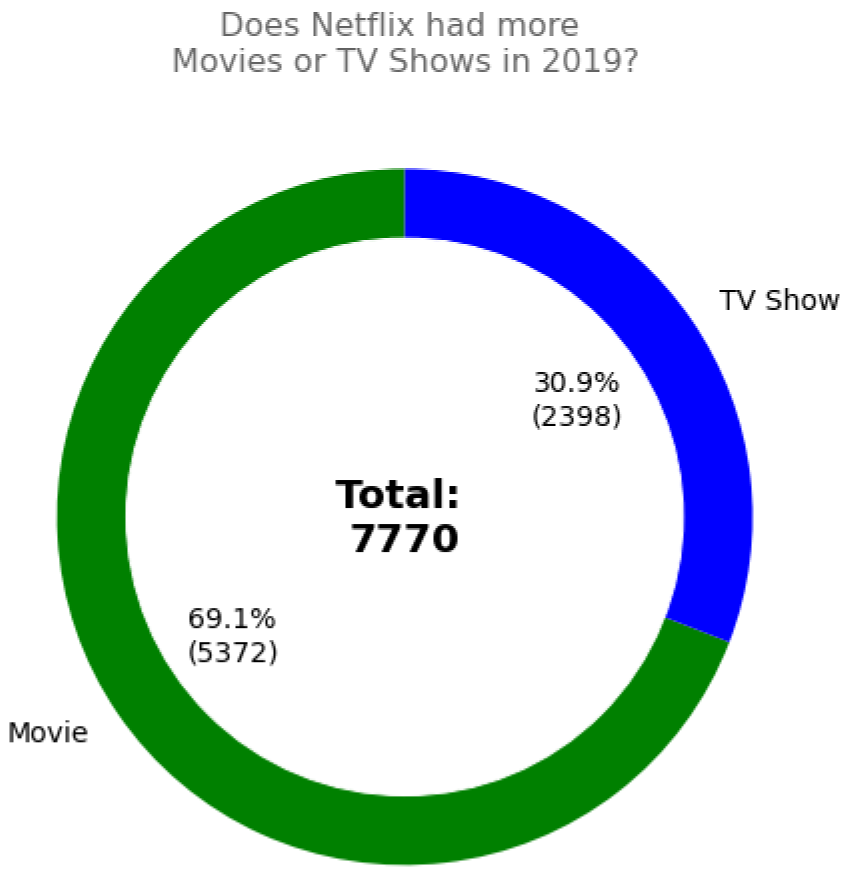
```
# Checking the count of duplicate values in dataset.
```

There are no duplicate values in our dataset.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) as the name suggests, is used to analyze and investigate datasets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions. It also helps to understand the relationship between the variables (if any) and it will be useful for feature engineering. It helps to understand data well before making any assumptions, to identify obvious errors, as well as better understand patterns within data.

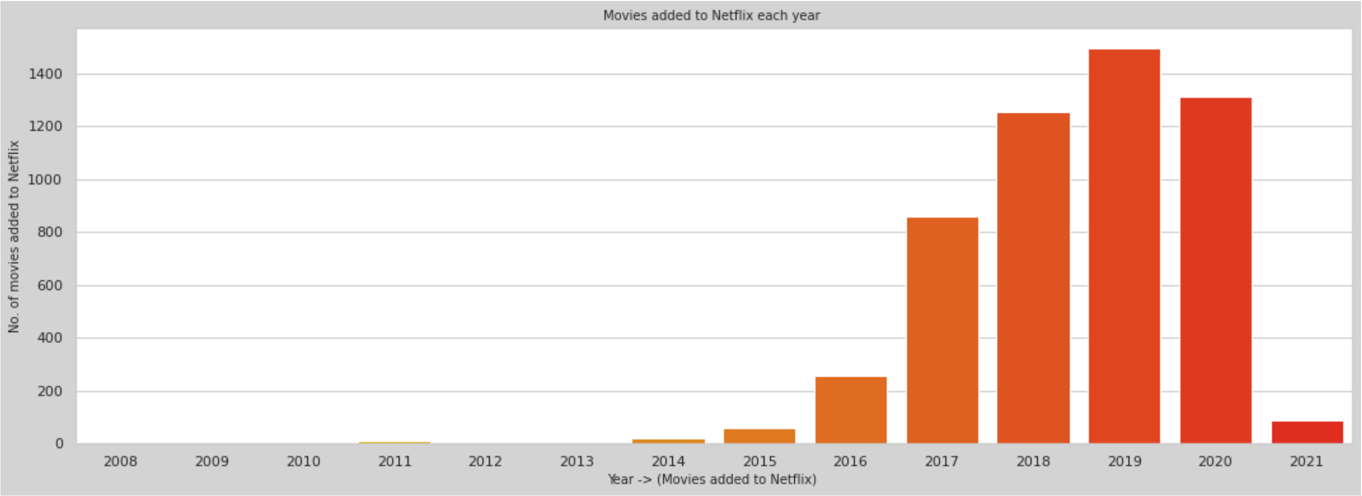
Distribution of Movies & TV shows



Netflix has 69% of its content as movies

Movies are clearly more popular on Netflix than TV shows.

Movies added to Netflix in each year



We can see from the graph that Netflix has started adding content since 2014.

The popularity of OTT has boomed in the last 5 years.

Highest number of movies and tv shows added in the year 2019.

Natural Language Processing (NLP) Model:

For the NLP portion of this project, I will first convert all plot descriptions to word vectors so they can be processed by the NLP model. Then, the similarity between all word vectors will be calculated using cosine similarity (measures the angle between two vectors, resulting in a score

between -1 and 1, corresponding to complete opposites or perfectly similar vectors.

Tfidf vectorization:

TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is a very common algorithm to transform text into a meaningful representation of numbers which is used to fit a machine learning algorithm for prediction.

We have also utilized the PCA because it can help us improve performance at a very low cost of model accuracy. Other benefits of PCA include reduction of noise in the data, feature selection (to a certain extent), and the ability to produce independent, uncorrelated features of the data.

So, it's essential to transform our text into tfidf vectorizer, then convert it into an array so that we can fit into our model.

- ## Finding number of clusters

The goal is to separate groups with similar characteristics and assign them to clusters.

We used the Elbow method and the Silhouette score to do so.

- **Fitting into model**

In this task, we have implemented a K means clustering algorithm. K-means is a technique for data clustering that may be used for unsupervised machine learning. It is capable of classifying unlabeled data into a predetermined number of clusters based on similarities (k).

Data Preprocessing:

Removing Punctuation-:: Punctuations does not carry any meaning in clustering, so removing punctuations helps to get rid of unhelpful parts of the data, or noise.

Removing stop-words-: Stop-words are basically a set of commonly used words in any language, not just in English. If we remove the words that are very commonly used in a given language, we can focus on the important words instead.

Stemming-: Stemming is the process of removing a part of a word, or reducing a word to its stem or root. Applying stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language.

Clustering-:

Clustering (also called cluster analysis) is a task of grouping similar instances into clusters. More formally, clustering is the task of grouping the population of unlabeled data points into clusters in a way that data points in the same cluster are more similar to each other than to data points in other clusters. The clustering task is probably the most important in unsupervised learning, since it has many applications.

We have used 5 clustering algorithms:

1.Silhouette score

2.Elbow Method

3.DBSCAN

4.Dendrogram

5.Agglomerative clustering

Building a clustering model -:

Clustering models allow you to categorize records into a certain number of clusters. This can help you identify natural groups in your data.

Clustering models focus on identifying groups of similar records and labeling the records according to the group to which they belong. This is done without the benefit of prior knowledge about the groups and their characteristics. In fact, you may not even know exactly how many groups to look for.

This is what distinguishes clustering models from the other machine-learning techniques—there is no predefined output or target field for the model to predict.

These models are often referred to as unsupervised machine learning models, since there is no external standard by which to judge the model's classification performance.

1 silhouette score-:

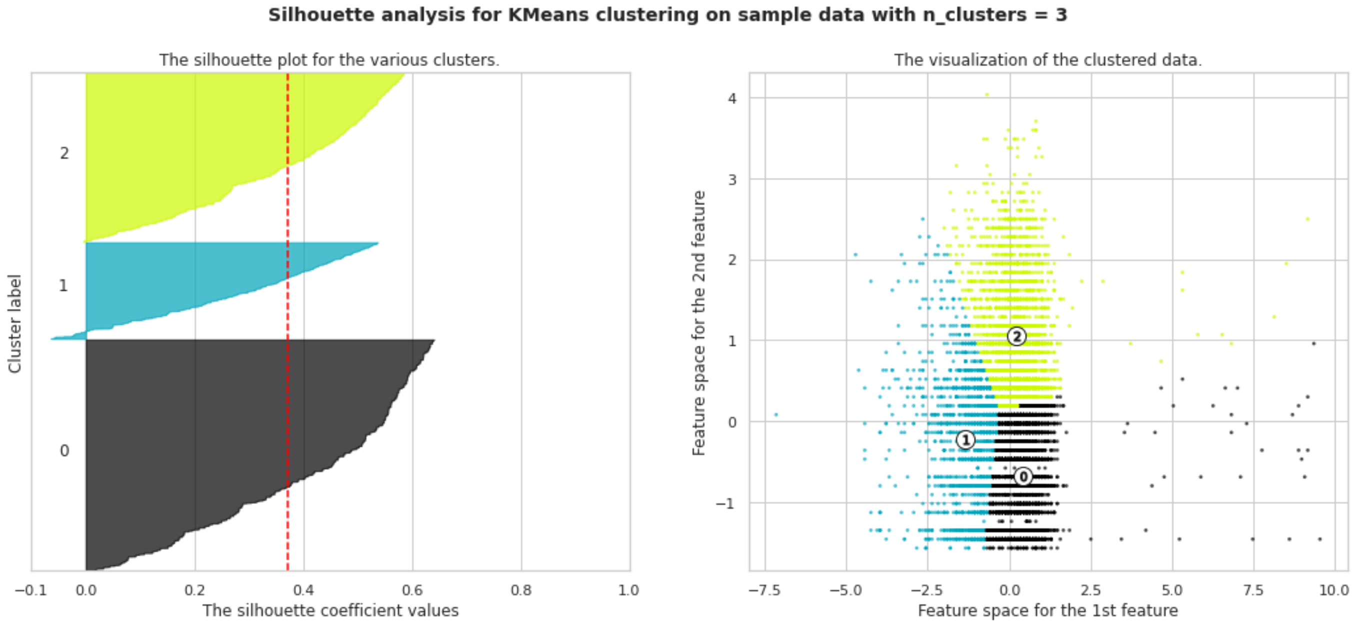
Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1. 1: Means clusters are well apart from each other and clearly distinguished. ... a= average intra-cluster distance i.e., the average distance between each point within a cluster.

If the ground truth labels are not known, the evaluation must be performed utilizing the model itself. The Silhouette Coefficient is an example of such an evaluation, where a more increased Silhouette Coefficient score correlates to a model with better-defined clusters. The Silhouette Coefficient is determined for each sample and consists of two scores.

- Mean distance between the observation and all other data points in the same cluster. This distance can also be called a mean intra-cluster distance. The mean distance is denoted by a.
- Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be called a mean nearest-cluster distance. The mean distance is denoted by b.

The Silhouette Coefficient *s* for a single sample is then given as: **(b - a) / max(a, b)**

Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters.



Silhouette analysis for KMeans clustering on sample data with n_clusters = 3

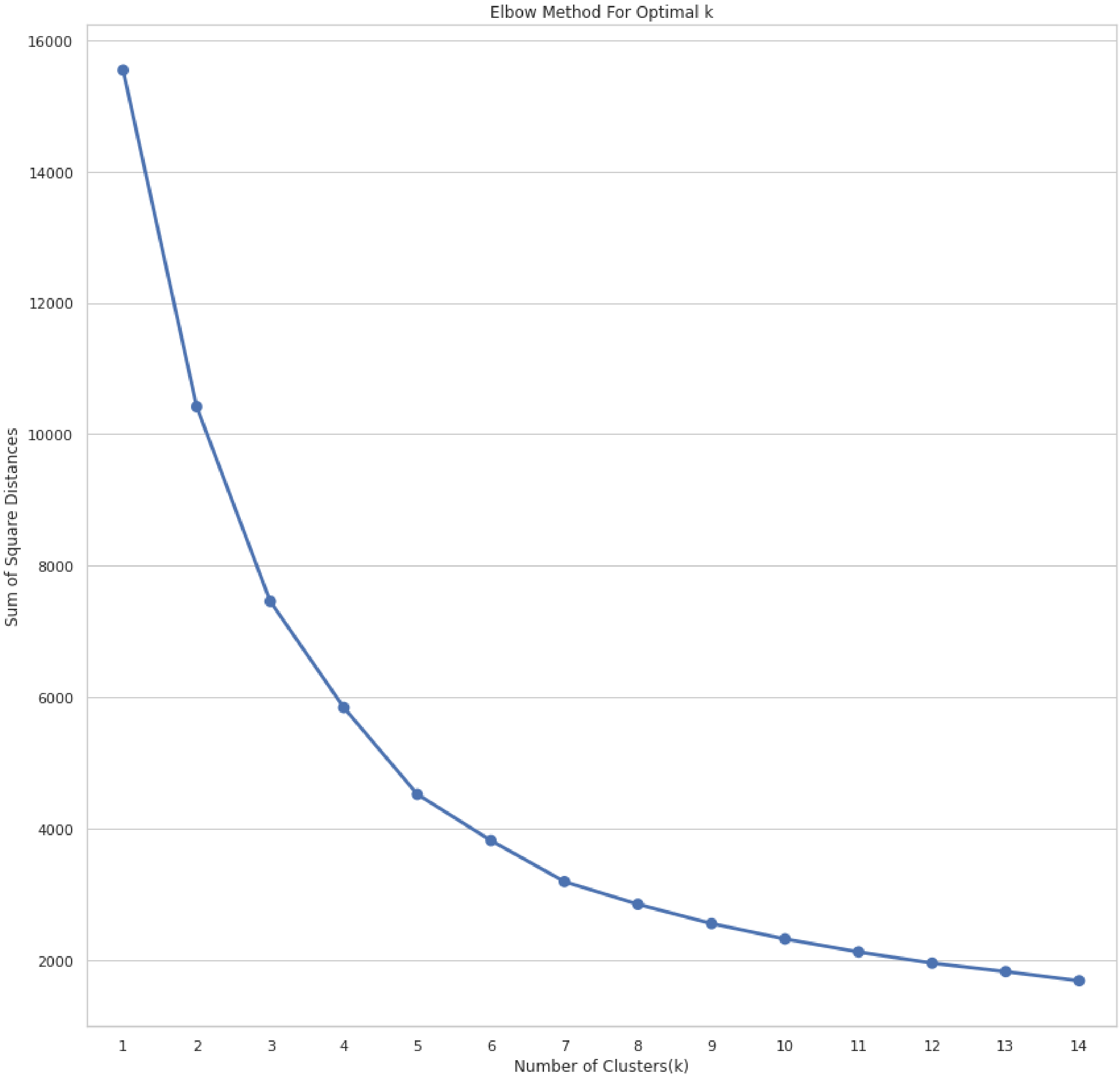
The silhouette plot for the various clusters.

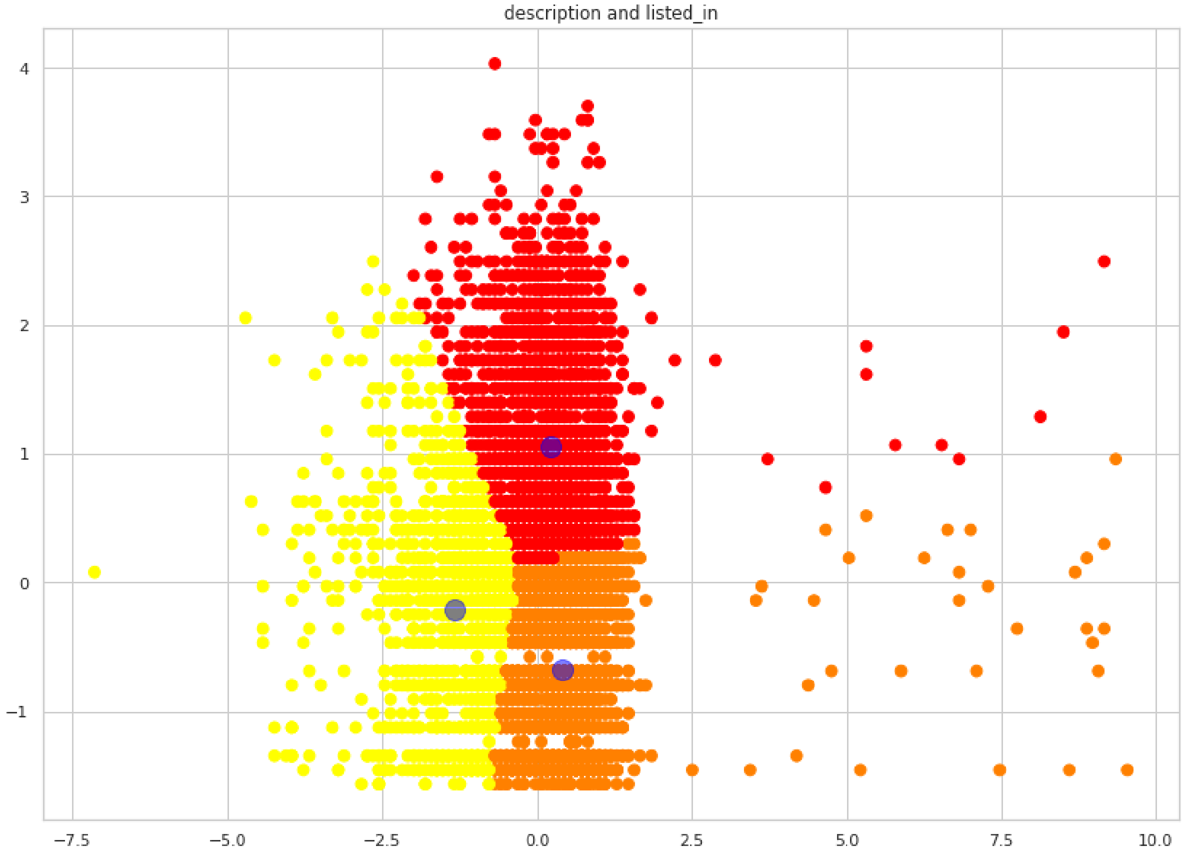
The visualization of the clustered data.

2.Elbow Curve:

The Elbow Curve is one of the most popular methods to determine this optimal value of k.

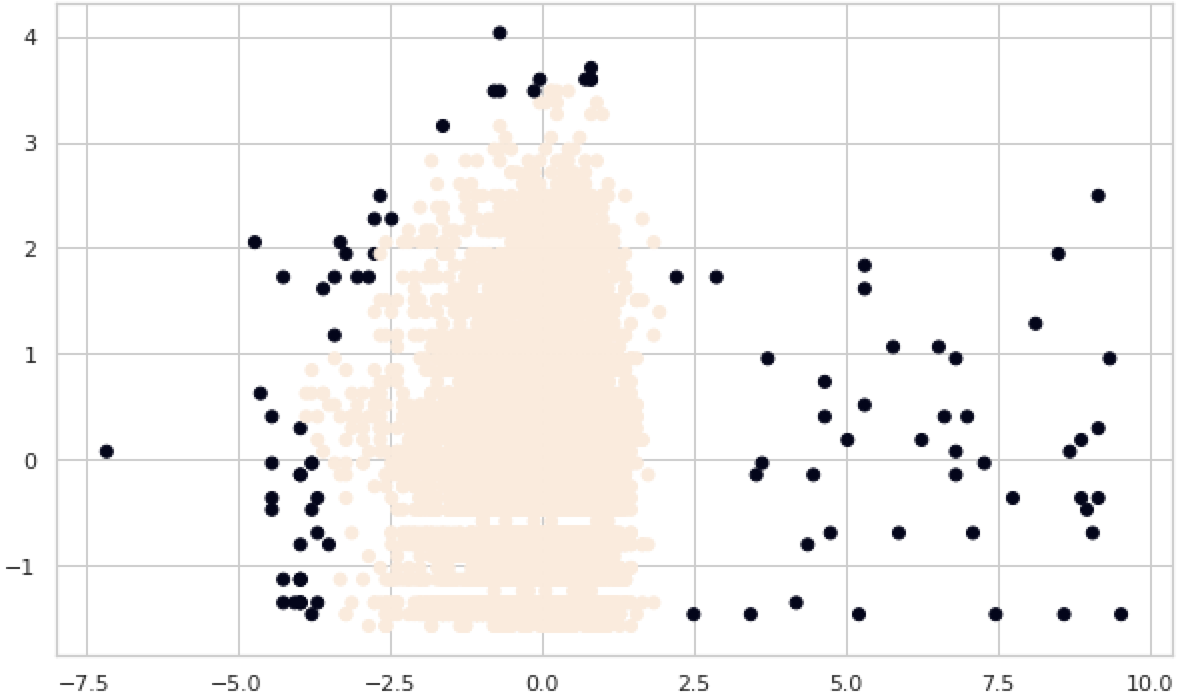
The elbow curve uses the sum of squared distance (SSE) to choose an ideal value of k based on the distance between the data points and their assigned clusters.


Elbow Method For Optimal k
Sum of Square Distances
16000
14000
12000
10000
8000
6000
4000
2000
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Number of Clusters(k)



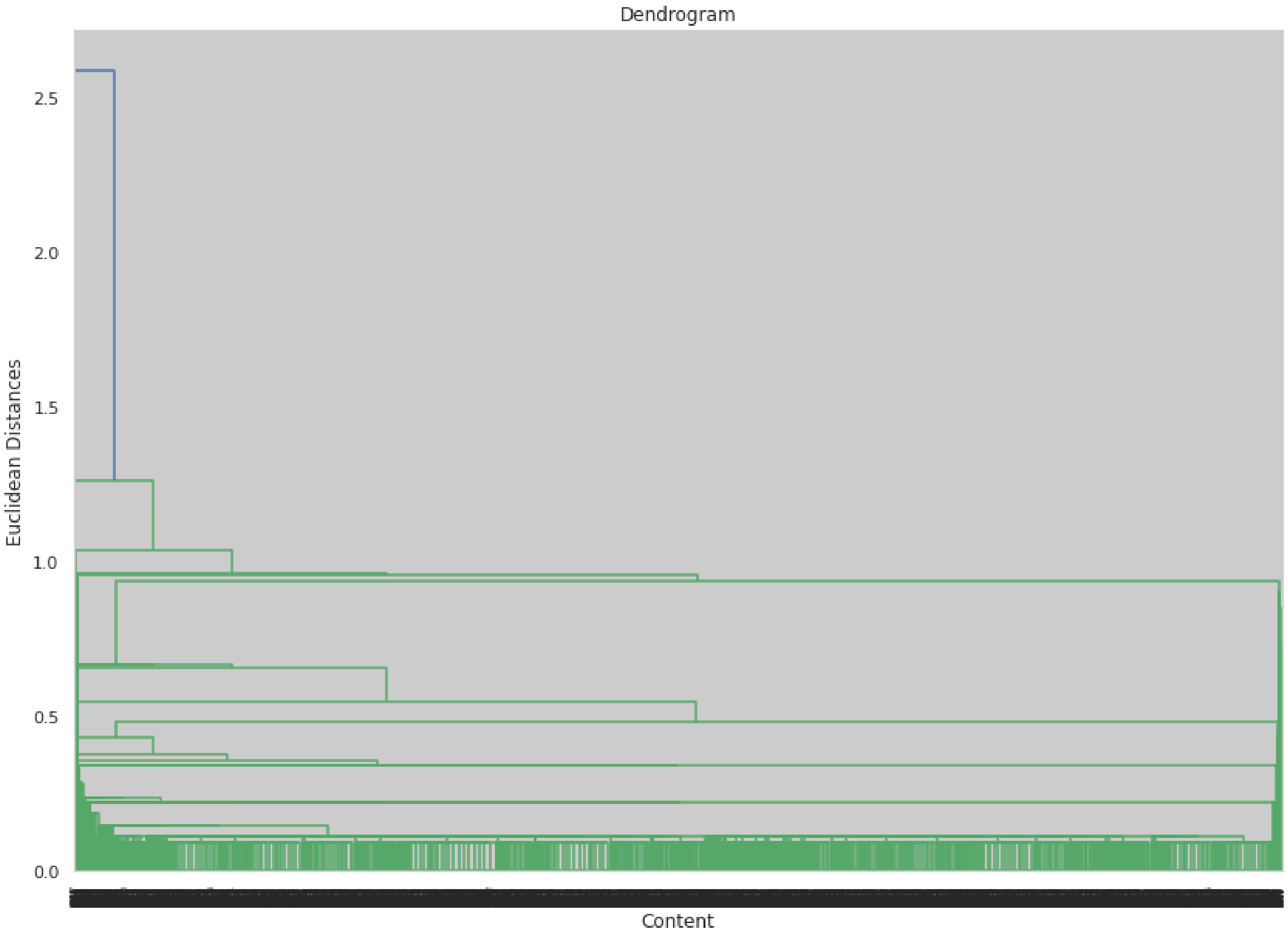
3. DBSCAN-:

DBSCAN is a clustering method that is used in machine learning to separate clusters of high density from clusters of low density. Given that DBSCAN is a density based clustering algorithm, it does a great job of seeking areas in the data that have a high density of observations, versus areas of the data that are not very dense with observations. DBSCAN can sort data into clusters of varying shapes as well, another strong advantage.



4.Dendrogram

A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to clusters.

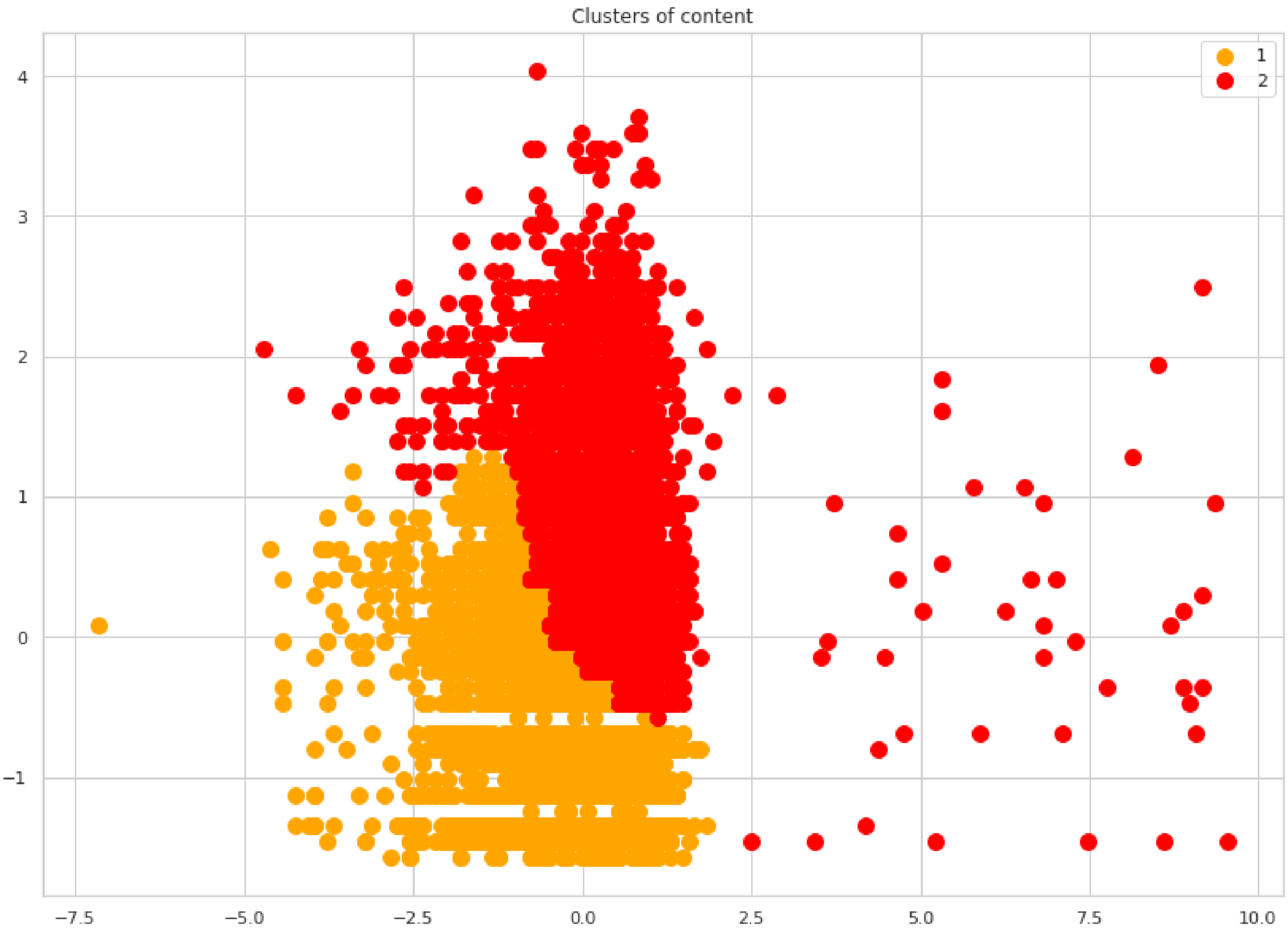


The number of clusters will be number of vertical lines which are intersected by the line drawn using the threshold

No. of Clusters = 3.

5.Agglomerative Clustering-:

The agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects.



Dimensionality Reduction-:

In machine learning classification problems, there are often too many factors on the basis of which the final classification is done. These factors are basically variables called features. The higher the number of features, the harder it gets to visualize the training set and then work on it. Sometimes, most of these features are correlated, and hence redundant. This is where dimensionality reduction algorithms come into play. Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be divided into feature selection and feature extraction.

Cosine Similarity-:

Cosine similarity measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. It is often used to measure document similarity in text analysis.
Cosine similarity is a measure of similarity that can be used to compare documents or, say, give a ranking of documents with respect to a given vector of query words. Let *x* and *y* be two vectors for comparison.

Conclusion-:

1.The dataset contains 7787 rows and 12 columns, cast and director columns have a lot of missing values so we dropped them and we have 10 features for the further analysis.

2.We have two types of content: movies and TV shows.

3.Netflix has 69% of its content as movies, so we can say that movies are clearly more popular on Netflix than TV shows.

4.For a mature audience, there is much more movie content than TV shows. However, for the younger audience (under the age of 17),there are more TV shows than movies.

5.Netflix has started adding content since 2014,highest number of movies and tv shows added in the year 2019,there is consistent content addition to netflix across the year.

6.The average duration of a movie on netflix is 90 minutes.

7.With respect to available content,the United States is on the top.India is at second followed by the UK and Canada. China is not even close to the top.

8.In terms of genres, Dramas is on the top followed by Comedies and Documentaries.

9.Number of movies added to netflix is higher than that of TV shows. In 2019, netflix added 1497 movies and 656 TV shows. So there we cannot conclude that Netflix has switched focus from movies to TV shows.

10. Principal component analysis was performed inorder to reduce the higher dimensionality which improved the silhouette coefficient to 0.34118.

11.Clusters are identified for each of the records in the dataset.

12.Recommendation based on cosine similarity is done.