Capstone Project- 4

Netflix Movies and TV Shows Clustering

(Unsupervised Machine Learning)

BY

Sk Samim Ali,
Sarath Haridas
(Cohort – Florence)

❖ Content

- Introduction
- Problem statement
- Data Description
- Exploratory Data Analysis
- Data cleaning
- Data Pre-processing
- Model Implementation
- K-Means
- Clustering Analysis
- Hierarchical Clustering
- Conclusion

❖ 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.

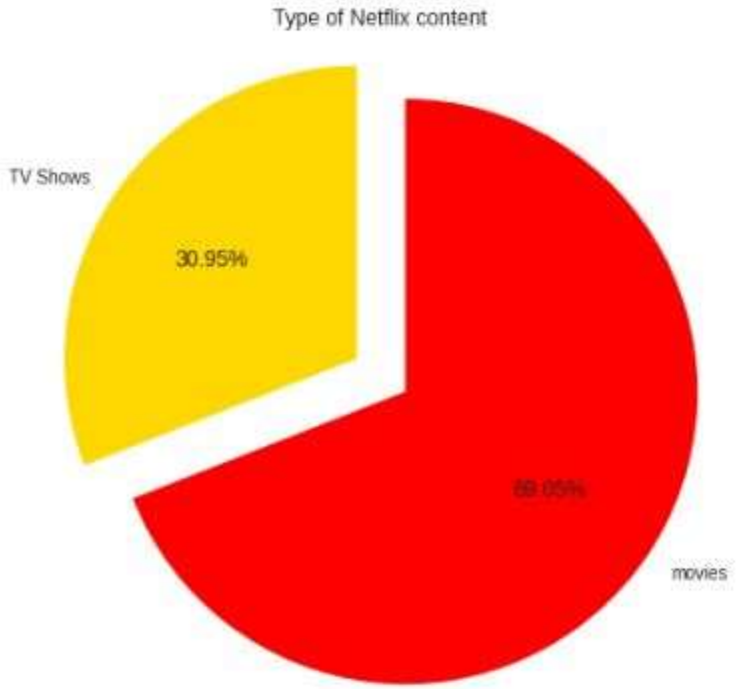
Data Description:

- **The data was collected from Flixable which is third party Netflix search engine. The dataset consist of movies and TV Shows. The Dataset has 7787 rows of Data.**
- **The Dataset consists of eleven textual columns and one Numeric Column.**

- **show_id :** Unique ID for every Movie / TV Show
- **type :** Identifier - A Movie or TV Show
- **title :**Title of the Movie / TV Show
- **director** :Director of the Movie
- **cast** : Actors involved in the movie / show
- **country :** Country where the movie / show was produced
- **date_added :**Date it was added on Netflix
- **release_year** :Actual Release year of the movie/ show
- **rating :**TV Rating of the movie / show
- **duration :** Total Duration - in minutes or number of seasons
- **listed_in** : Genre
- **description:** The Summary description

❖ Exploratory Data Analysis:

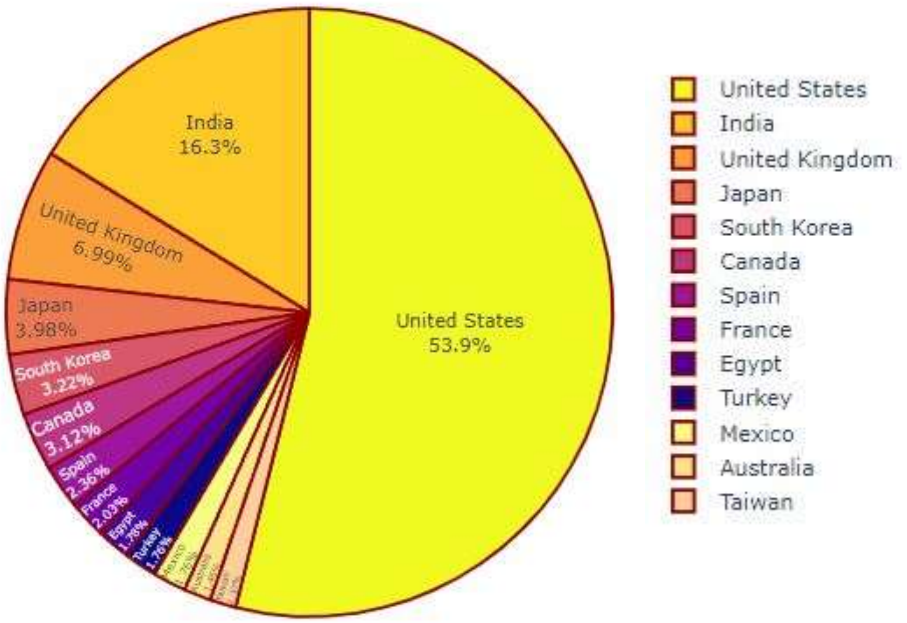
❑ Type of Content Available on Netflix



- It is evident that there are more movies on Netflix than TV shows
- Netflix has 30.95% of TV Shows and 69.05% of Movies which is more than double the quantity of TV shows.

❖ Exploratory Data Analysis:

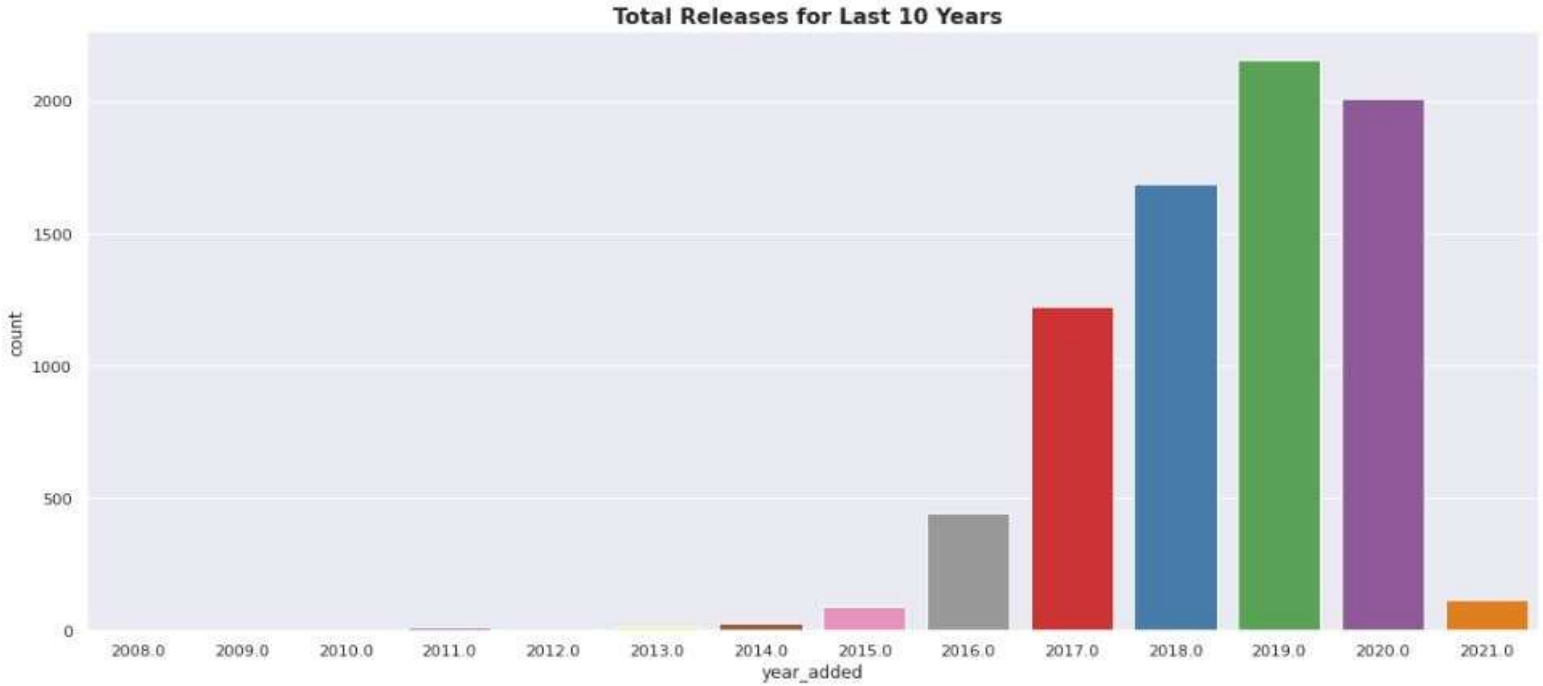
❑ Top Countries With Highest content production



- United state has the most number of content on Netflix.
- India has second highest content on Netflix.
- Australia and Taiwan has least number of Content on netflix.

❖ Exploratory Data Analysis:

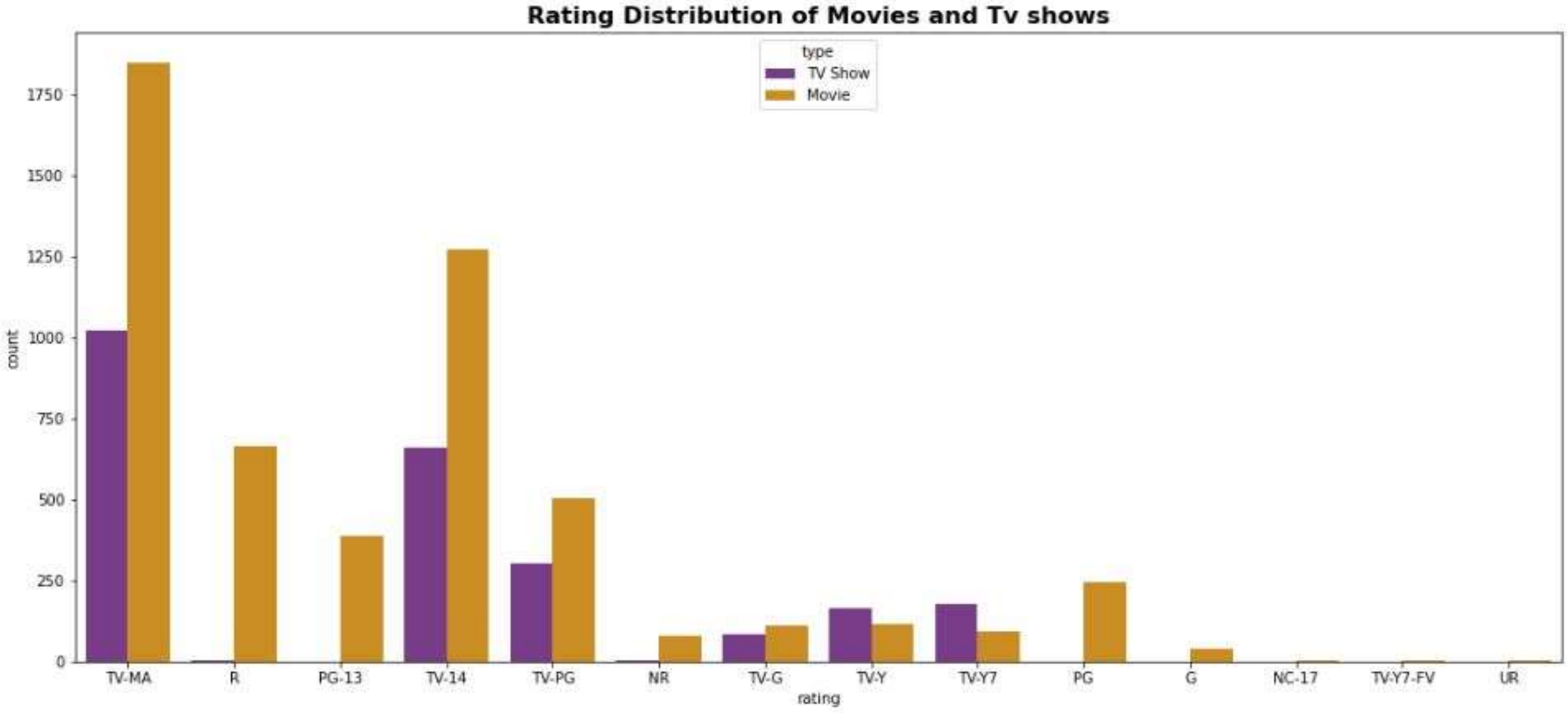
❑ Top Releases over 10 years



- **The number of release have significantly increase after 2015 and have dropped in 2021 because of COVID-19.**

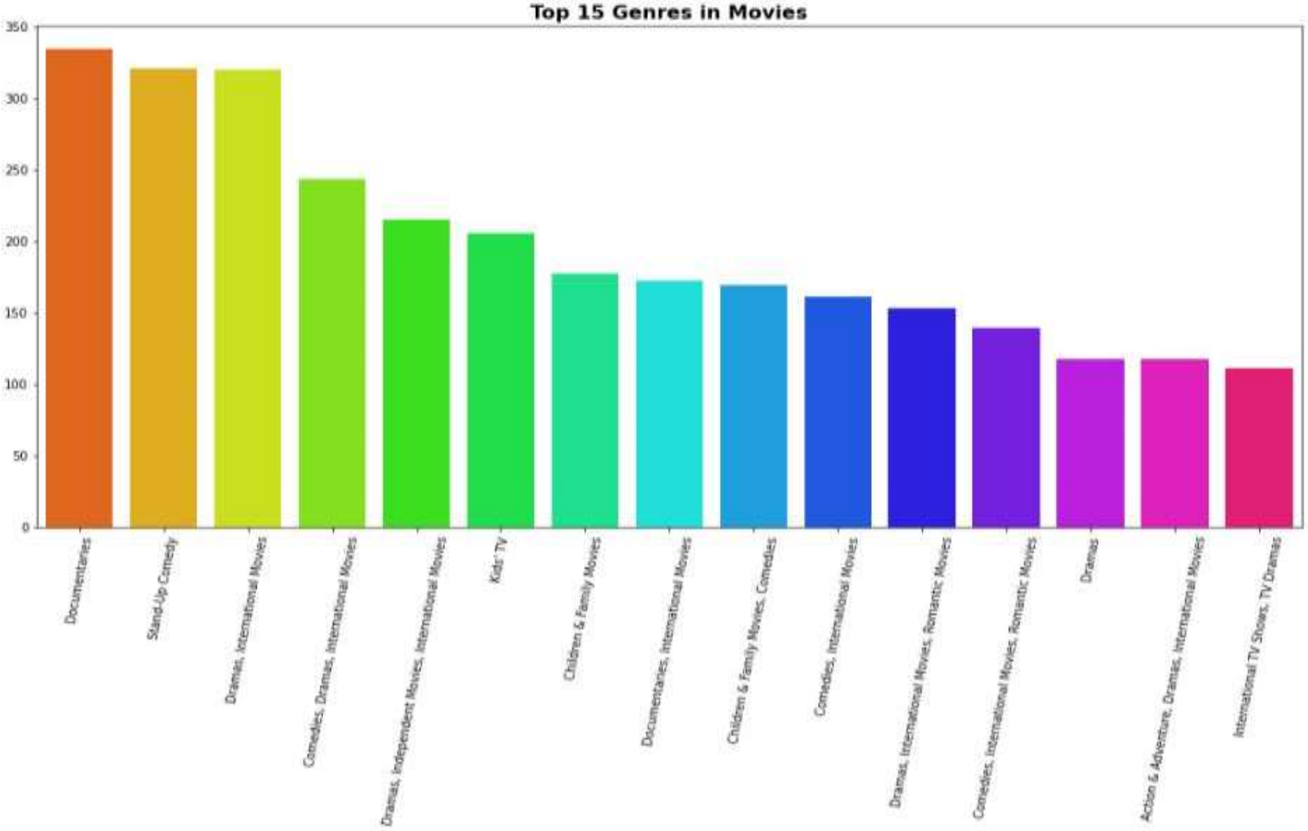
❖ Exploratory Data Analysis:

❑ Rating-wise Content count



❖ Exploratory Data Analysis:

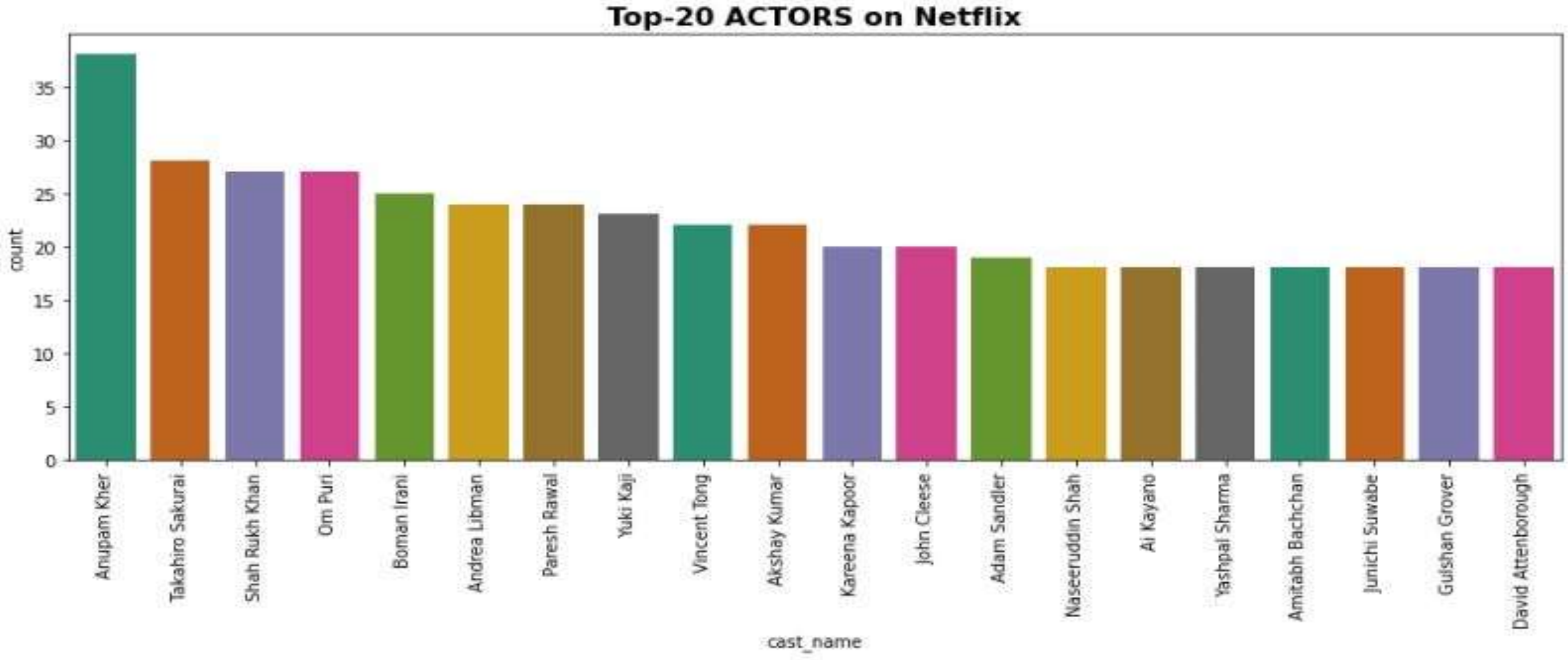
❑ Rating-wise Content count



- **Documentaries are the most popular Genre followed by the comedy**

❖ Exploratory Data Analysis:

❑ Top 20 Actors on Netflix



Data Cleaning:

- **Stemming:** It is the process of reducing the word to its word stem that affixes to suffixes and prefixes or to roots of words known as a lemma. In simple words stemming is reducing a word to its base word or stem in such a way that the words of similar kind lie under a common stem. For example – The words care, cared and caring lie under the same stem 'care'. Stemming is important in natural language processing.
- **Removing Stop-words:** The process of converting data to something a computer can understand is referred to as pre-processing. One of the major forms of pre-processing is to filter out useless data. In natural language processing, useless words (data), are referred to as stop words.
- **Stop Words:** A stop word is a commonly used word (such as "the", "a", "an", "in") that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.
- **TF-IDF Vectorizer :**TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction. It helps us in dealing with most frequent words. Using it we can penalize them. TfidfVectorizer weights the word counts by a measure of how often they appear in the documents

Model Implementation:

K-Means :

To process the Learning Data, the KMeans Algorithm in data mining start with a first group of randomly selected centroids, which are used as the beginning point for every cluster, then performs iterative (repetitive) calculations to optimize the position of centroids.

It halts creating and optimizing clusters when either :

- The Centroids have stabilized - there is no change in their values because the clustering has been successful.
- The define number of iterations has been achieved.

Model Implementation:

K-Means Clustering:

K-Means Algorithm is an iterative Algorithm that tries to partition the dataset into K pre defined distinct non overlapping subgroups where each data points belongs to only one group.

1. Elbow Curve:

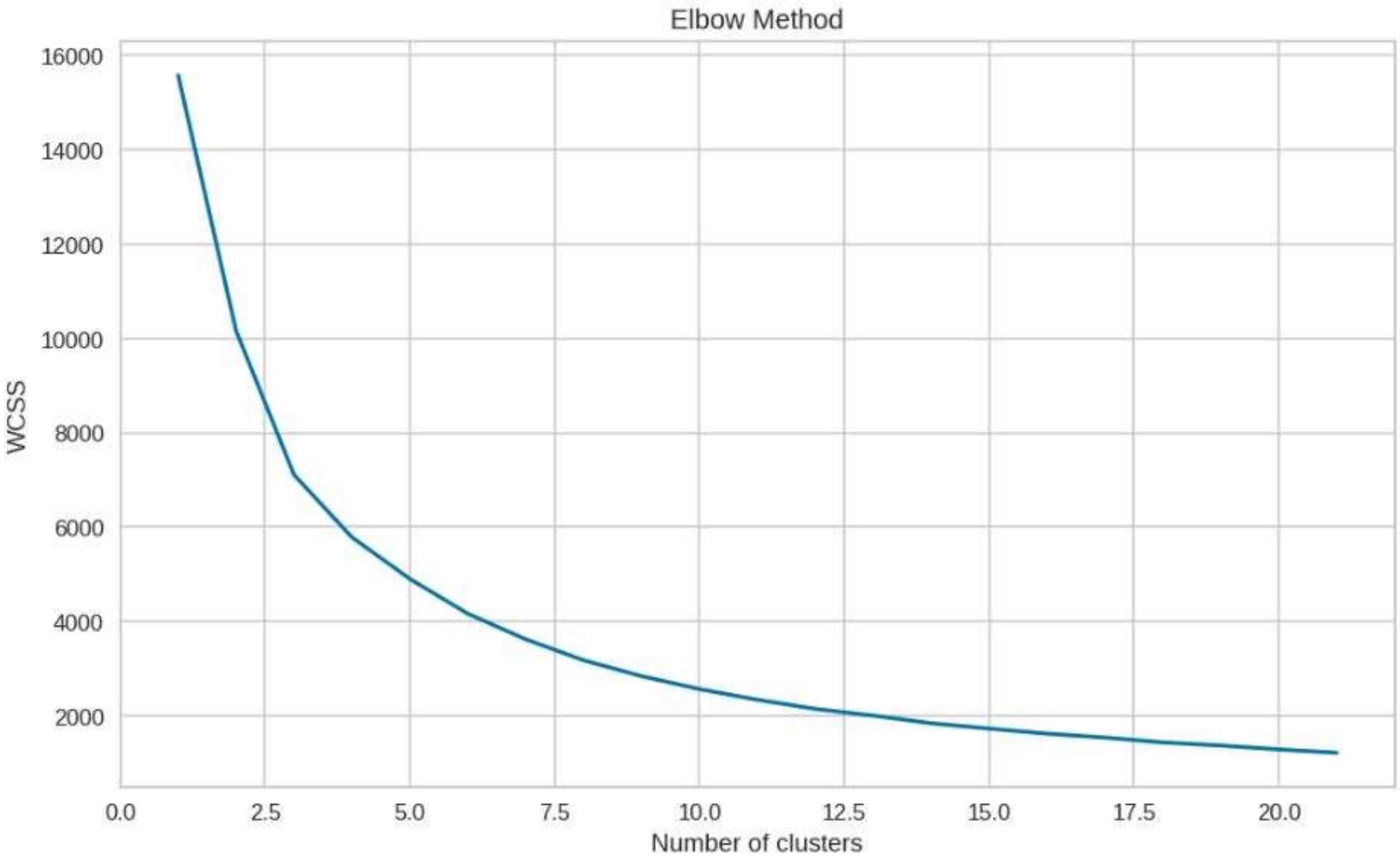
- The elbow method is used to determine the optimal number of clusters in k-means clustering.
- The elbow method plots the value of the cost function produced by different values of k.

2. 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.

❖ Model Implementation:

Elbow curve:



❖ Model Implementation:

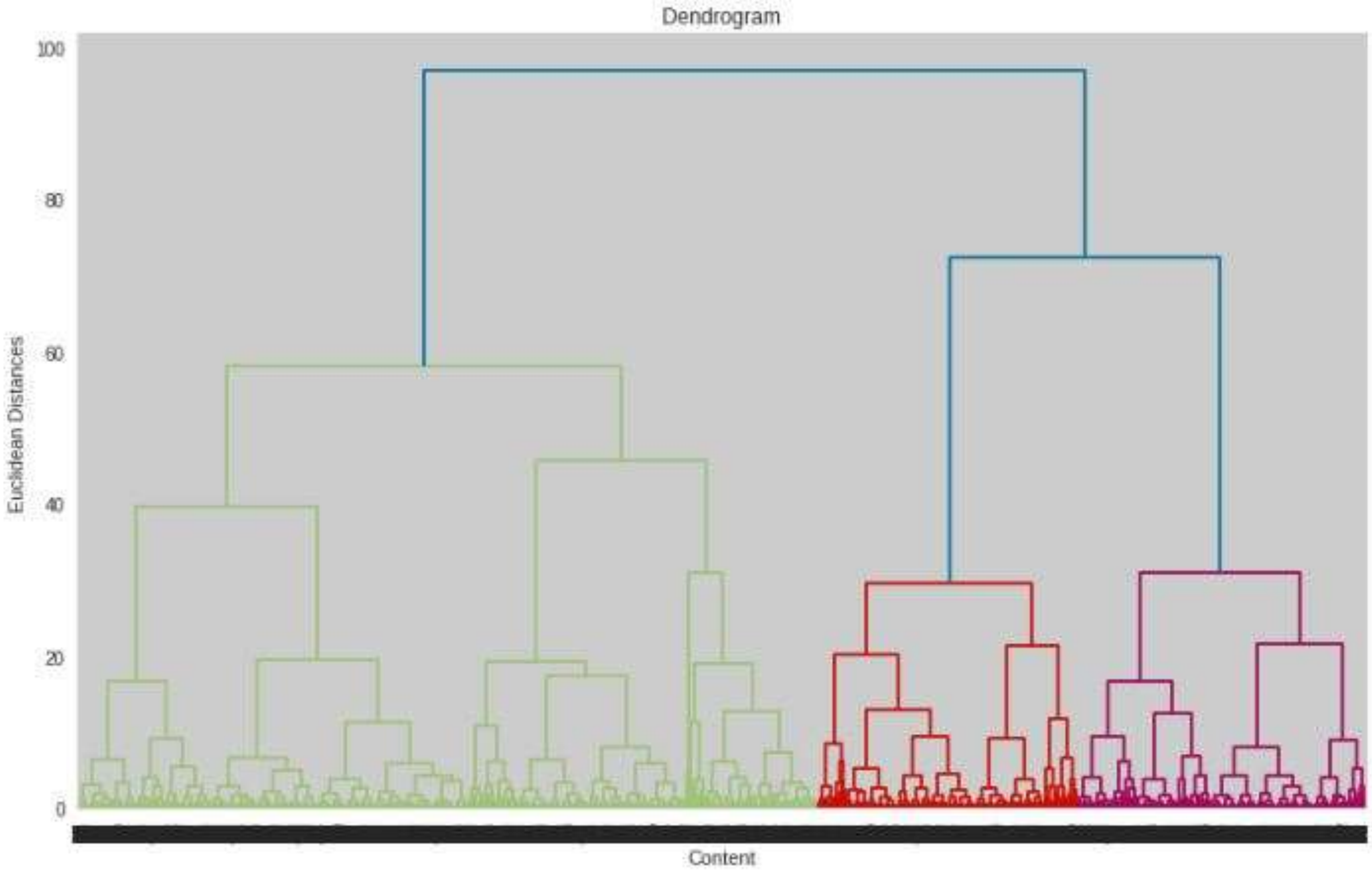
Hierarchical Clustering:

A Hierarchical clustering method works via grouping data into a tree of clusters. Hierarchical clustering begins by treating every data point as a separate cluster. Then, it repeatedly executes the subsequent steps:

- Identify the 2 clusters which can be closest together
- Merge the 2 maximum comparable clusters. We need to continue these steps until all the clusters are merged together.
- In Hierarchical Clustering, the aim is to produce a hierarchical series of nested clusters. A diagram called **Dendrogram** (A Dendrogram is a tree-like diagram that statistics the sequences of merges or splits)

❖ Model Implementation:

Dendrogram:



Model Implementation:

Agglomerative Hierarchical Clustering:

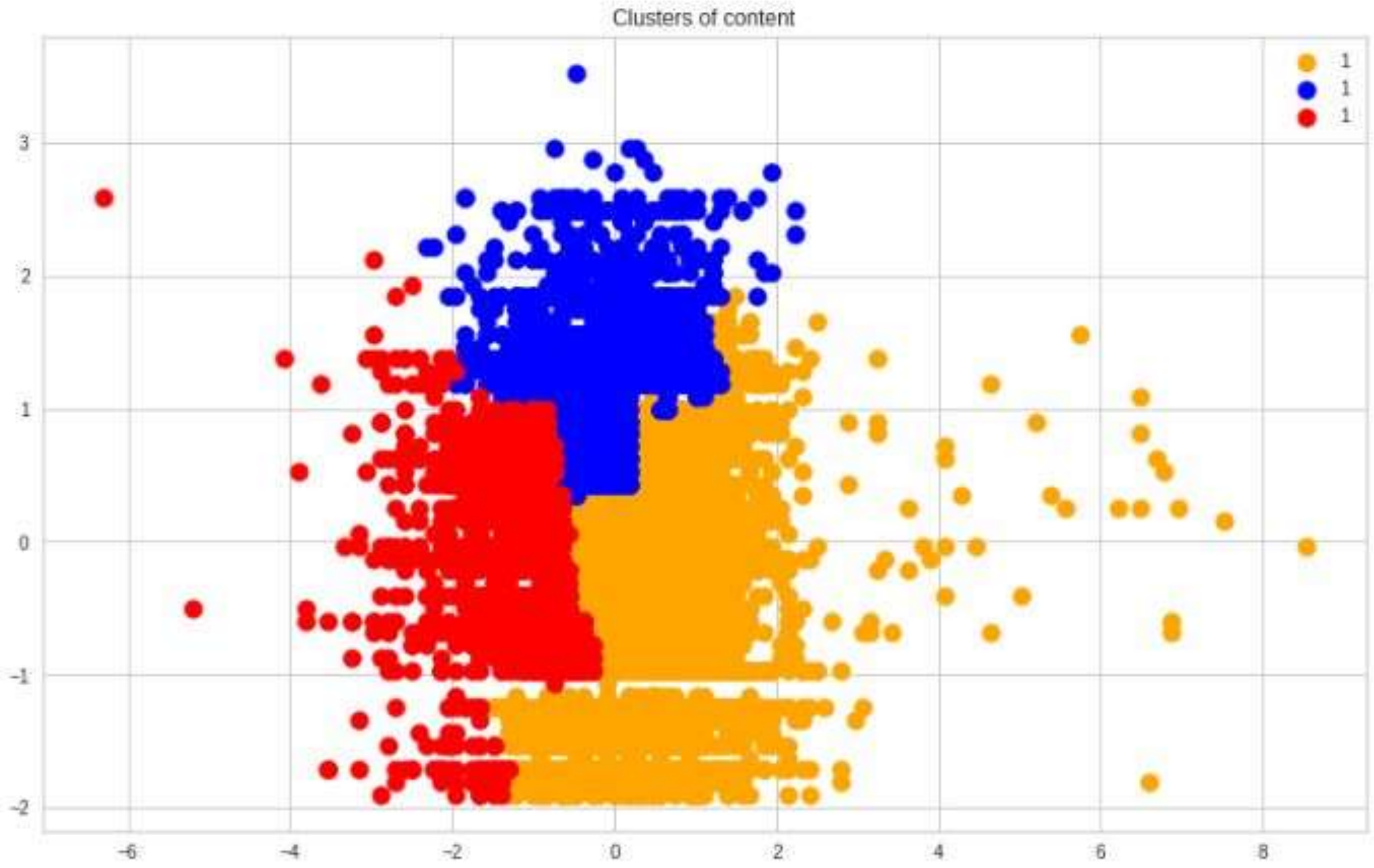
Agglomerative: Initially consider every data point as an **individual** Cluster and at every step, merge the nearest pairs of the cluster. (It is a bottom-up method). At first, every dataset is considered as an individual entity or cluster. At every iteration, the clusters merge with different clusters until one cluster is formed.

The algorithm for Agglomerative Hierarchical Clustering is:

- **Calculate the similarity of one cluster with all the other clusters (calculate proximity matrix)**
- **Consider every data point as an individual cluster**
- **Merge the clusters which are highly similar or close to each other.**
- **Recalculate the proximity matrix for each cluster**
- **Repeat Steps 3 and 4 until only a single cluster remains.**

Model Implementation:

Agglomerative Hierarchical Clustering:



❖ Conclusion:

- ❑ Data set contains 7787 rows and 12 columns in that cast and director features contains large number of missing values so we decided to drop this director attribute also duration and show_id attribute because its not useful for our model.

- ❑ We have two types of content TV shows and Movies (30.95% contains TV shows and 69.05% contains Movies)

- ❑ The United States has the highest number of content on Netflix by a huge margin followed by India.

- ❑ Anupam Kher has acted in the highest number of films on Netflix. Documentaries is the most popular genre followed by Stand-up comedy.

- ❑ Most films were released in the years 2018, 2019, and 2020.

- ❑ The number of releases have significantly increased after 2015 and have dropped in 2021 because of COVID-19.

❖ Conclusion:

- ❑ By analyzing the content added over years we get to know that in recent years Netflix is focusing movies than TV shows.

- ❑ The second thing we did was feature engineering, which involved removing certain variables and preparing a dataframe to feed the clustering algorithms.

- ❑ By applying the silhouette score method for n range clusters on dataset we got best score which is 0.348 for clusters = 3, it means content explained well on their own clusters.

- ❑ For the clustering algorithm, we utilized "description" and "listed_in" attributes

- ❑ Applied different clustering models KMeans, hierarchical, Agglomerative clustering on data we got the best cluster arrangements

- ❑ Speaking about other different cluster methods, KMeans, hierarchical, agglomerative clustering on data, we got the best cluster arrangements.

Optimal number of cluster = 3.