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 includes all Netflix-eligible TV series and films as of 2019.The dataset was gathered through the third-party Netflix search engine Flixable.
- ❑ The amount of TV series available on Netflix has almost tripled since 2010, according to an interesting analysis that was published in 2018.
- ❑ Since 2010, the number of movies available on the streaming service has dropped by more than 2,000, although the number of TV series has nearly tripled. Investigating what further insights may be drawn from the same dataset will be intriguing.
- ❑ Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

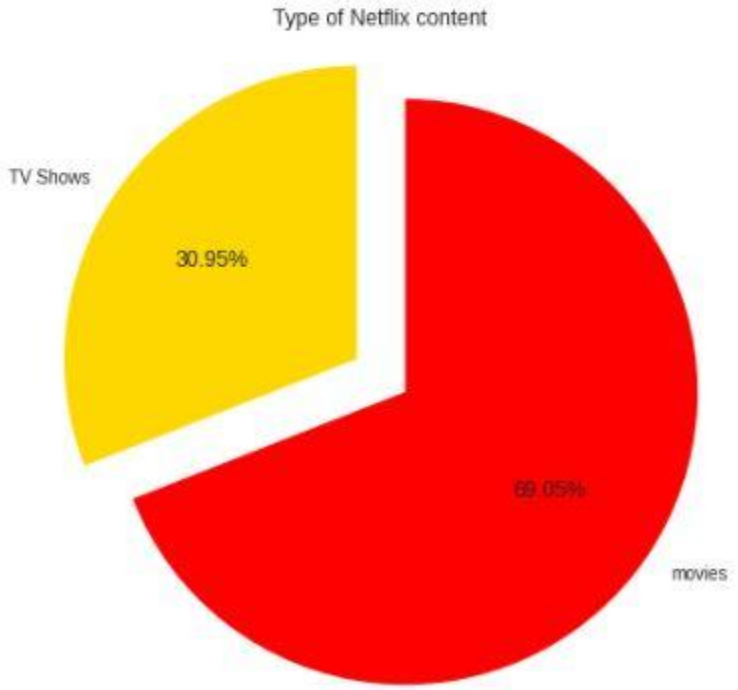
Data Description

- ❑ **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

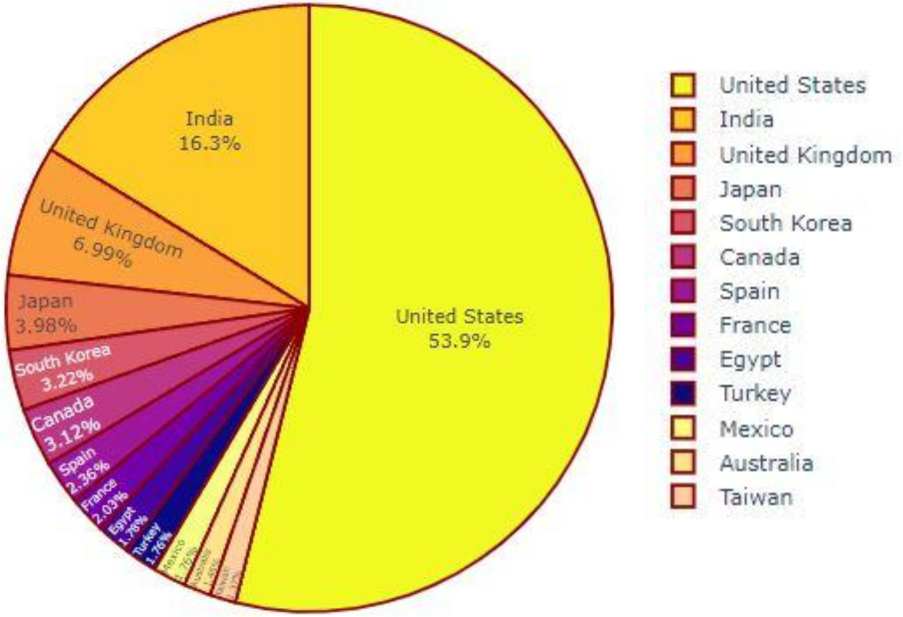


- It is evident that there are more movies on Netflix than TV shows

- Netflix boasts 69.05% of movies and 30.95% of TV shows, more than twice as many movies as TV shows.

Exploratory Data Analysis

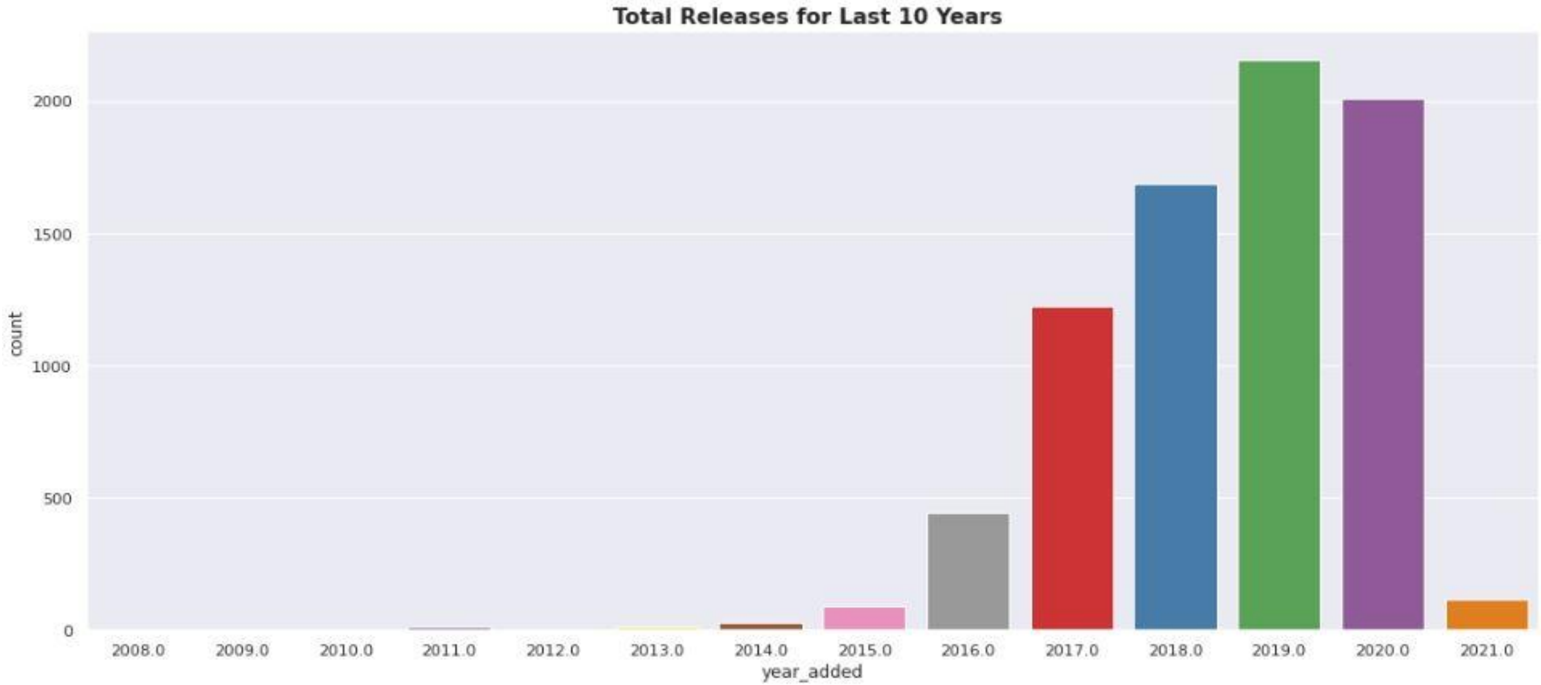
❑ Top Countries With Highest content production



- 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

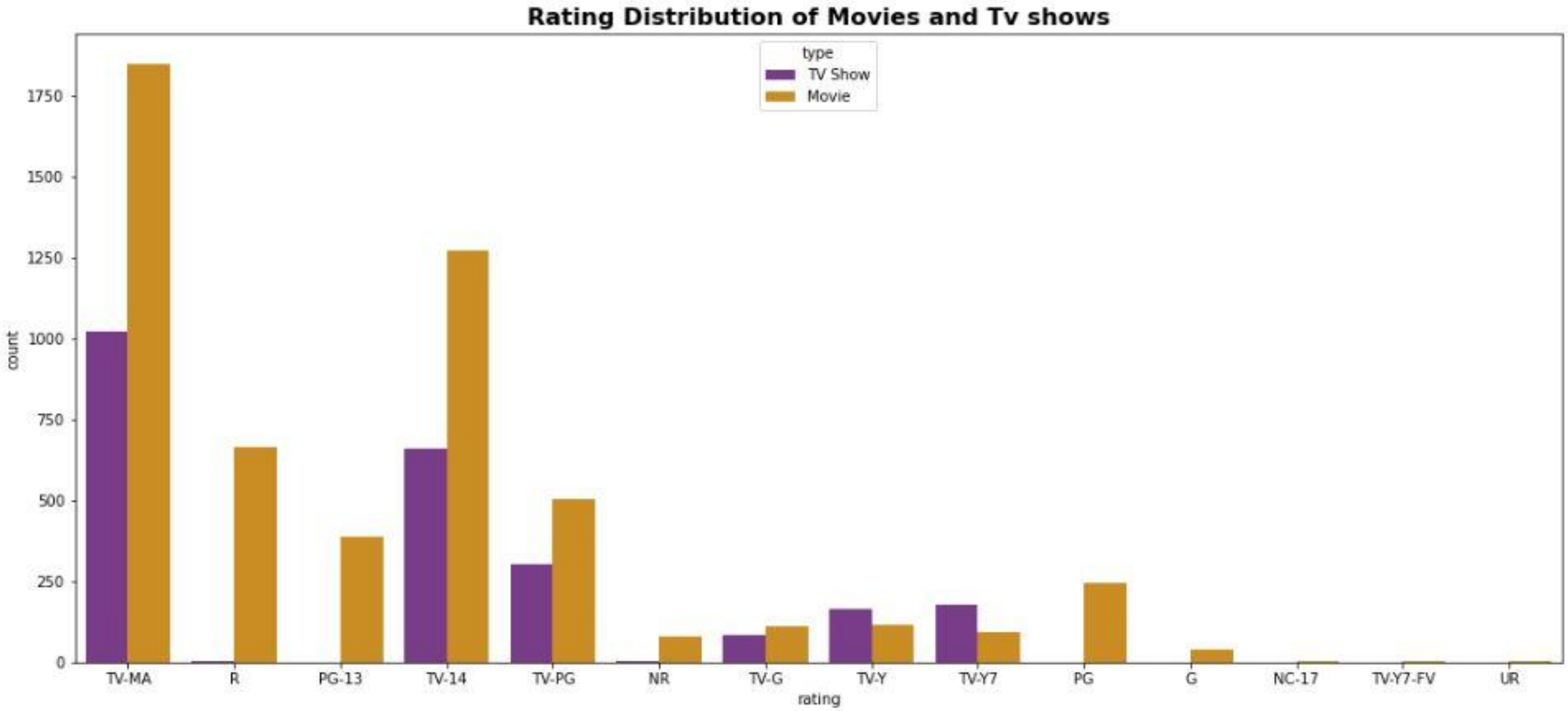
❑ Total Releases over 10 years



- **Due to COVID-19, the number of releases increased dramatically after 2015 and decreased in 2021.**

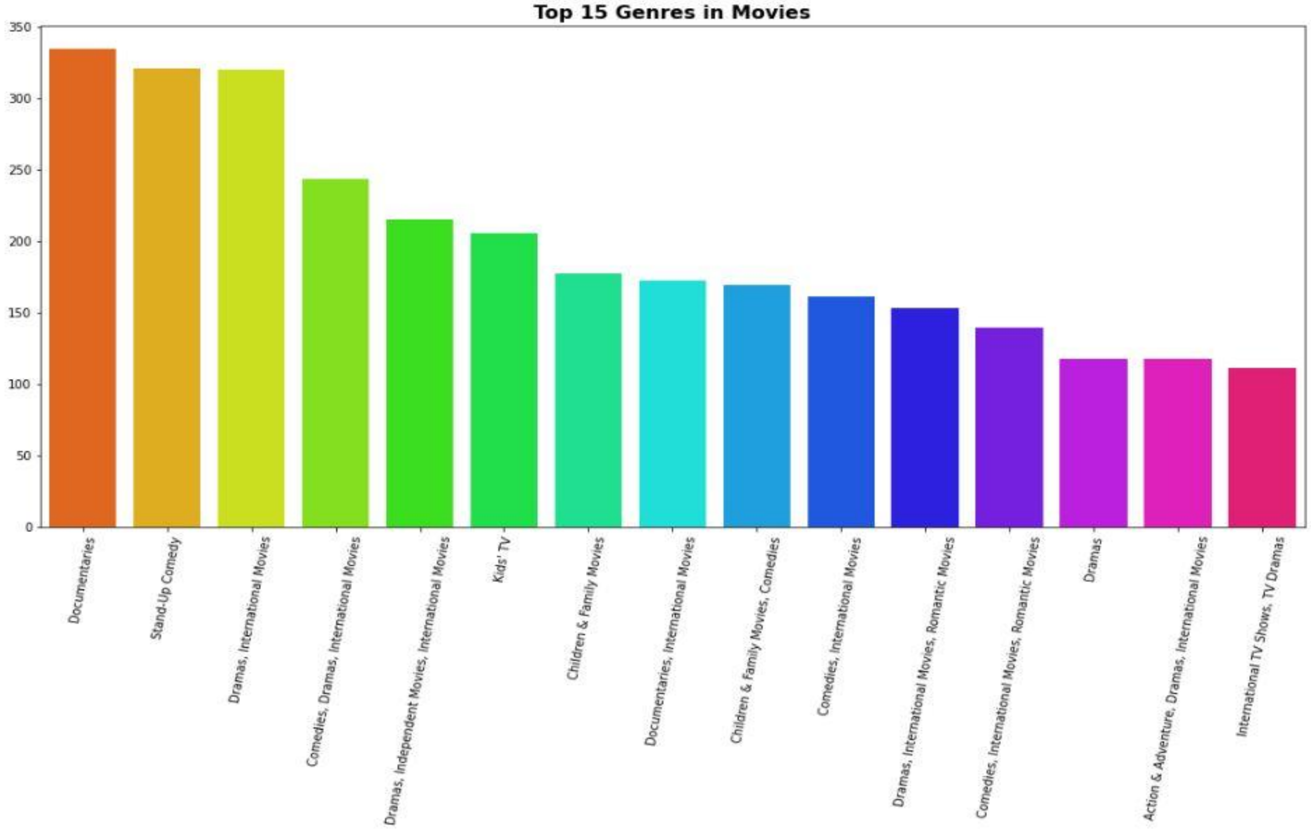
Exploratory Data Analysis

❑ Rating-wise Content count



Exploratory Data Analysis

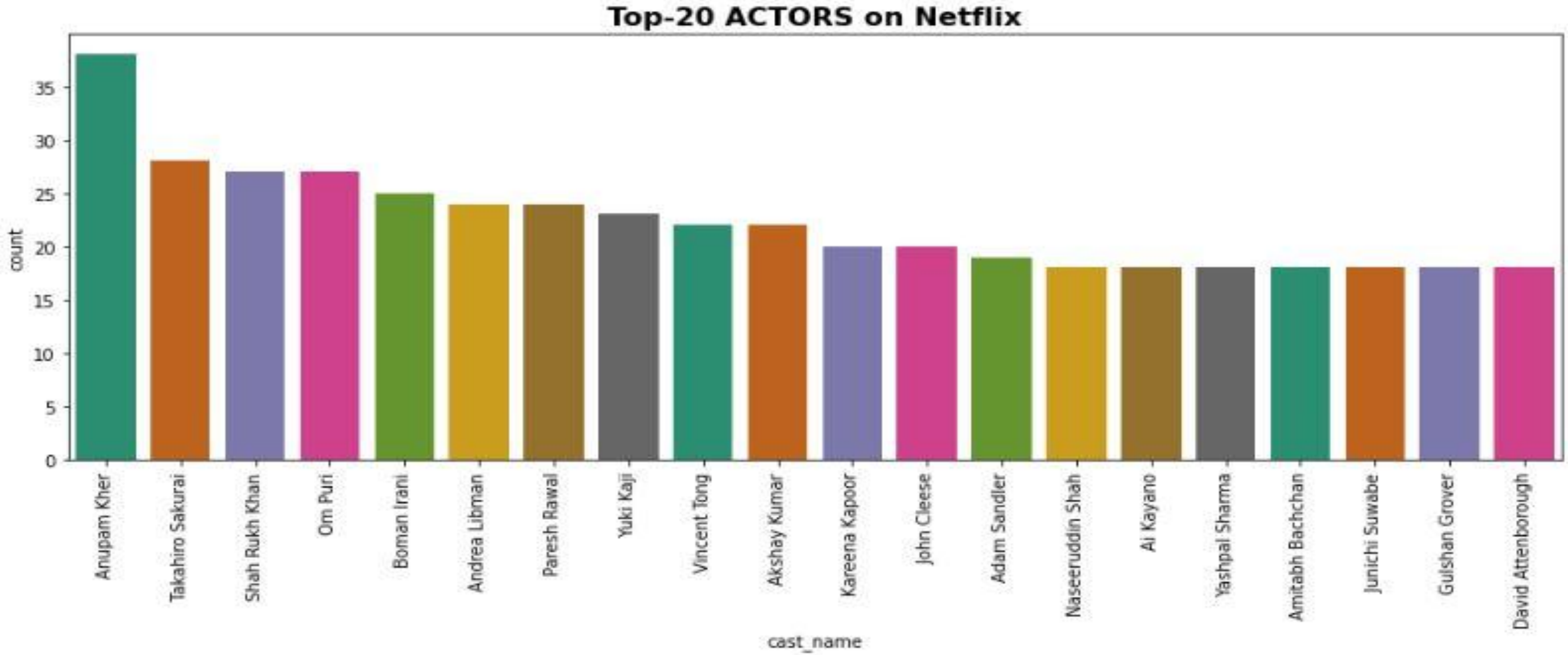
❑ Genre-wise Content count



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

Exploratory Data Analysis

❑ Top 20 Actors on Netflix



Data Cleaning

- **Stemming:** When a lemma is attached to suffixes, prefixes, or the roots of words, the word is reduced to its word stem. In plain English, stemming is the process of reducing a word to its root word or stem so that terms of the same kind are grouped together under a single stem. As an illustration, the words care, cared, and caring all share the same root word. In the processing of natural language, stemming is crucial.
- **Removing Stop-words:** Pre-processing is the process of transforming data into a form that a computer can comprehend. Filtering away pointless data is one of the main types of pre-processing. Stop words are worthless words (data) that are used in natural language processing.

  Stop Words: A stop word is a frequently used term that a search engine has been configured to ignore, both while indexing entries for searching and when retrieving them as the result of a search query. Examples of stop words include "the," "a," "an," and "in."
- **TF-IDF Vectorizer :**Term Frequency Inverse Document Frequency is referred to as TF-IDF. This is a widely popular algorithm that converts text into meaningful numerical representations that can be used to fit machine prediction algorithms. It aids us in coping with the most common words. We can punish them with it. The word counts are weighted by a measure of how frequently they appear in the documents using TfidfVectorizer.

Model Implementation

K-Means :

The KMeans Algorithm in data mining uses a first set of centroids that are chosen at random to process the learning data. These centroids serve as the starting point for each cluster, and iterative calculations are then used to optimise the position of the 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:

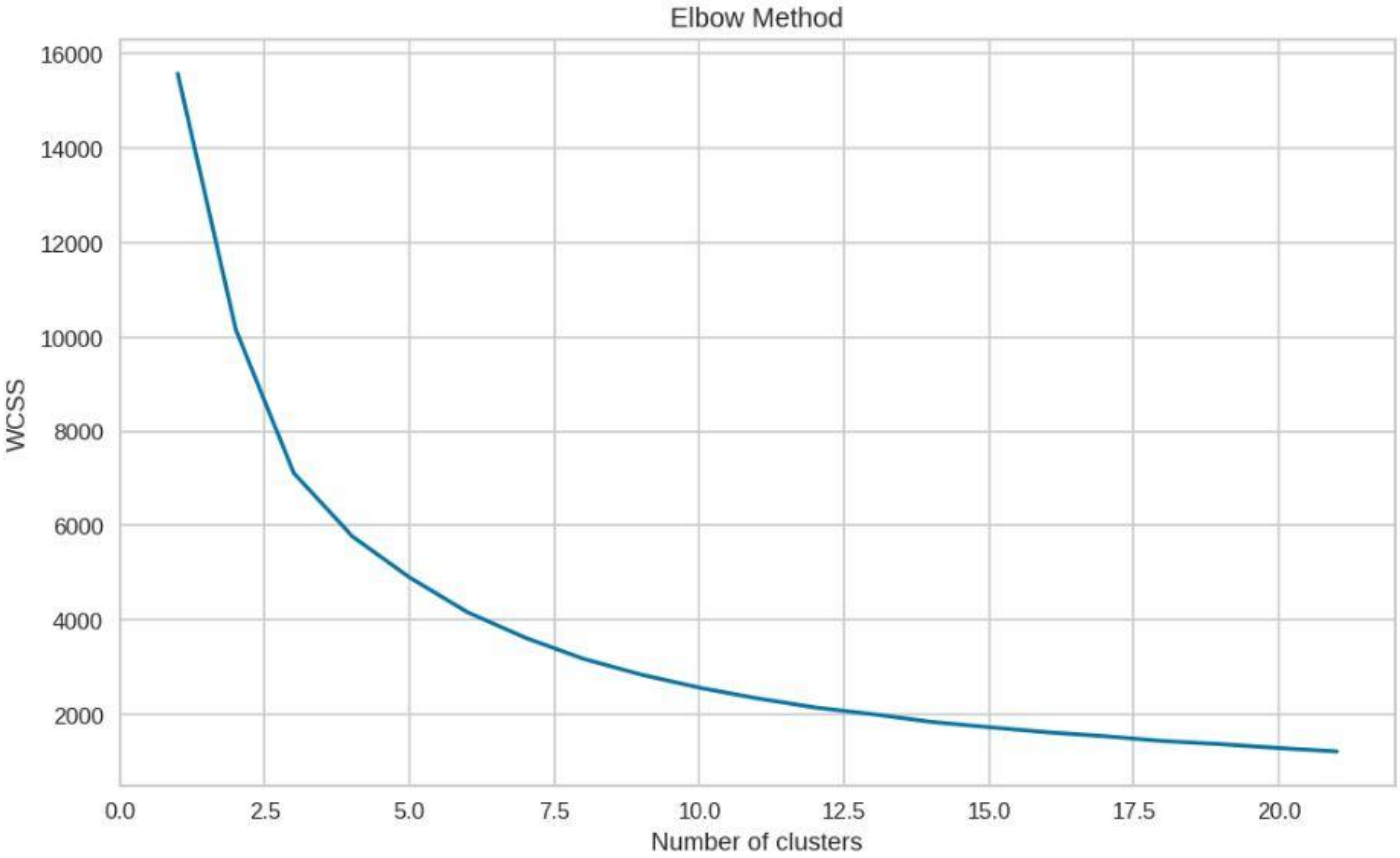
- In k-means clustering, the ideal number of clusters is established using the elbow approach.
- By using various choices of k, the elbow approach plots the value of the cost function.

2. Silhouette Score:

- A metric used to assess the efficacy of a clustering method is the silhouette coefficient, often known as the silhouette score. Its value is between -1 and 1.
- 1: Means clusters are obviously distinct from one another and spaced widely apart.

Model Implementation

Elbow curve:



Model Implementation

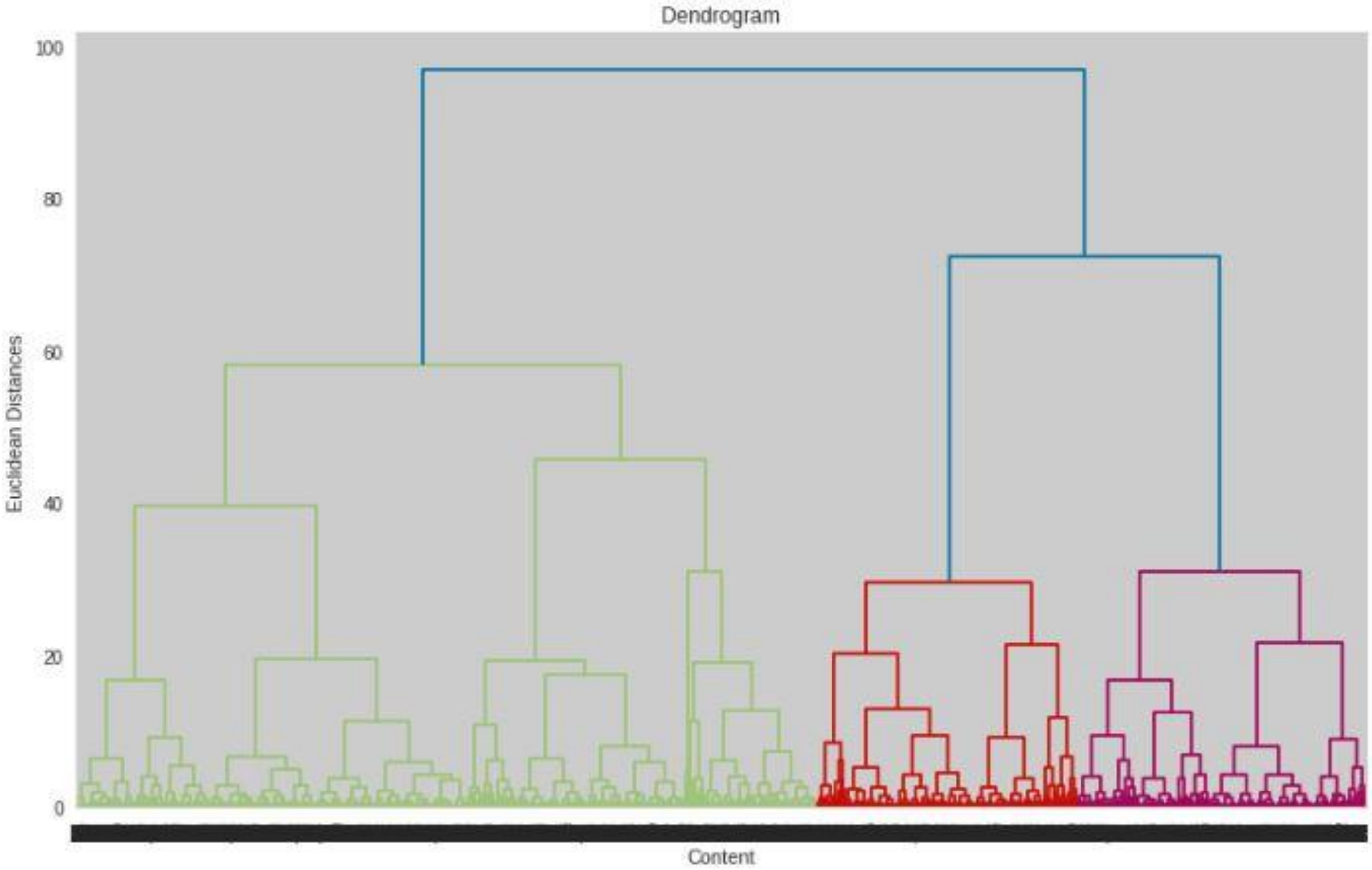
Hierarchical Clustering:

Data are grouped into groups in a tree structure in a hierarchical clustering method. Every data point is first treated as a separate cluster in a hierarchical clustering process. The following steps are then repeatedly carried out by it:

- Identify the 2 clusters which can be closest together
- Merge the two clusters that are most comparable. These procedures must be repeated until all of the clusters are combined.
- The goal of hierarchical clustering is to create a hierarchy of nested clusters. a Dendrogram, a type of graph (A Dendrogram is a tree-like diagram that statistics the sequences of merges or splits)

Model Implementation

Dendrogram:



Agglomerative Hierarchical Clustering:

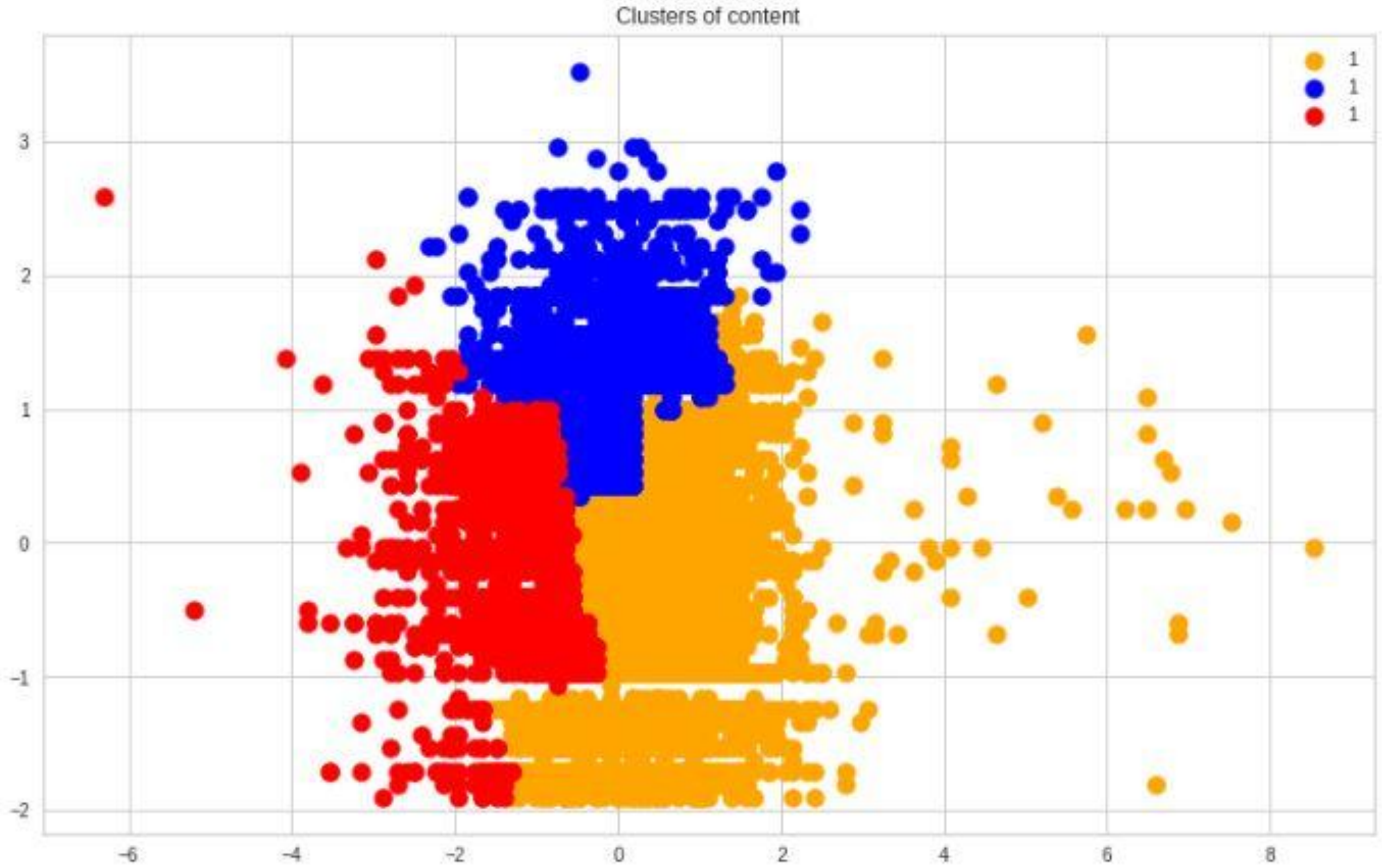
Agglomerative: At first, treat each data point as its own cluster. Then, at each step, combine the closest cluster pairs. (It uses a bottom-up approach.) Every dataset is first viewed as a distinct entity or cluster. The clusters combine with other clusters at each iteration until only one cluster remains.

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

❑ The data set consists of 7787 rows and 12 columns, and the director and cast features have a significant number of missing values. As a result, we chose to remove the director attribute as well as the duration and show id attributes because they are not necessary 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 appeared in the most Netflix movies of any actor. The most popular genre is documentaries, followed by stand-up comedy.

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

❑ Due to COVID-19, the number of releases dramatically increased after 2015 and decreased in 2021.

Conclusion

- We can see that Netflix has been focused more on movies than TV series in recent years by examining the content that has been added over time.

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

- By using the silhouette score approach for n range clusters on the dataset, we obtained the best score of 0.348 for clusters = 3, which indicates that the content was adequately explained on its 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 alternative clustering techniques, we obtained the optimum cluster arrangements using KMeans, hierarchical, and agglomerative clustering on the data.

Optimal number of cluster = 3