Netflix Movies and TV Shows Clustering

Sourav Chowdhury(Data Science Trainee)
AlmaBetter, Bangalore

Abstract:

With the advent of streaming platforms, there's no doubt that Netflix has become one of the important platforms for streaming. The dataset that we have used for EDA and clustering has been collected by Flixable, a third-party Netflix search engine. There are 12 features and around 7787 observations in the dataset and are mostly textual features.
Through univariate and multivariate analysis, we found trends that will help in understanding what content is being consumed country-wise, depending on some categorical features like rating, type, genres, cast, directors, etc. Clustering was performed along with NLP on textual columns and then a mini recommendation system was built out of it.

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, we are doing –
1. Exploratory Data Analysis.
2. Understanding what type of content is available in different countries.
3. Is Netflix increasingly focused on TV rather than movies in recent years?

4. Clustering similar content by matching text-based features Our goal here is to make an unsupervised clustering model, which will help in garnering insights on Netflix and how its content is being consumed.
A brief summary of the dataset is given below:
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 content
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

Steps involved:

- Importing libraries
- Exploratory Data Analysis
- Data Preparation
- Clustering
    1. Implementing K-means
    2. Implementing Hierarchical clustering

Introduction:

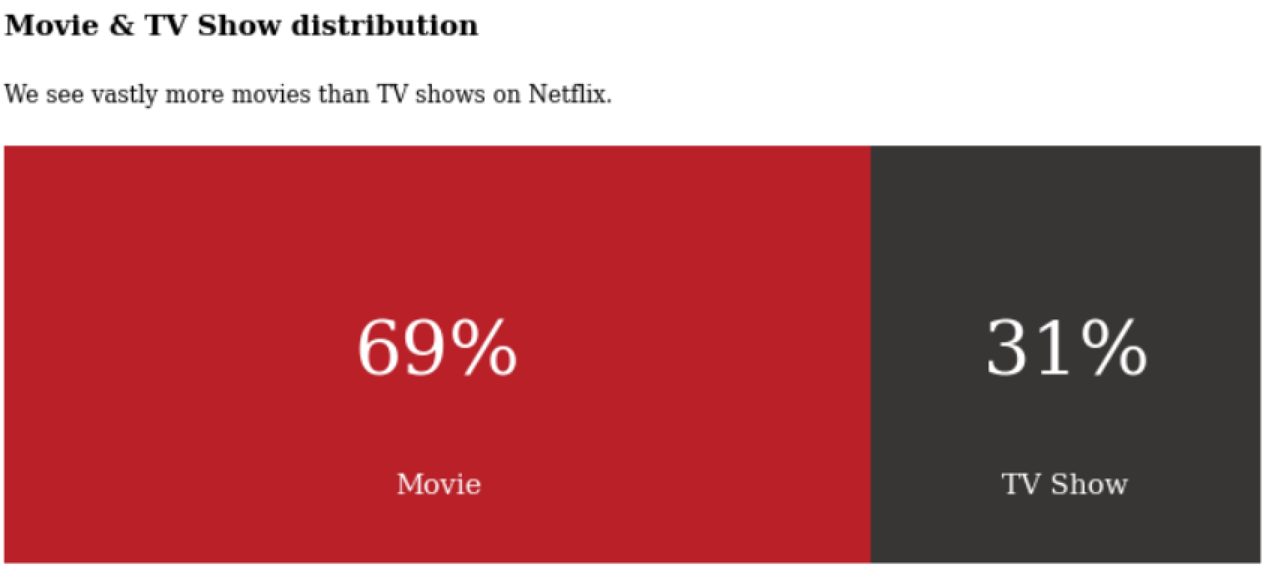
Unsupervised Learning is a machine learning technique in which the models are not supervised by the training set instead we find hidden patterns and insights from the given data. It is a machine learning technique in which models are trained on the unlabeled data set without any supervision.
A cluster is a collection of elements that are similar to each other but dissimilar to the elements belonging to other clusters. Clustering can be done using various kinds of distances such as Euclidean distance, Manhattan distance, geometric distance, etc. We can do different kinds of clustering based on the data pattern in space such as hierarchical clustering, K-means clustering, etc.

Exploratory Data Analysis:

The first step involved in the analysis is to load the dataset into the pandas data frame. Before exploring the data using different libraries available in python we should check if the dataset is ready to run the operations on it.

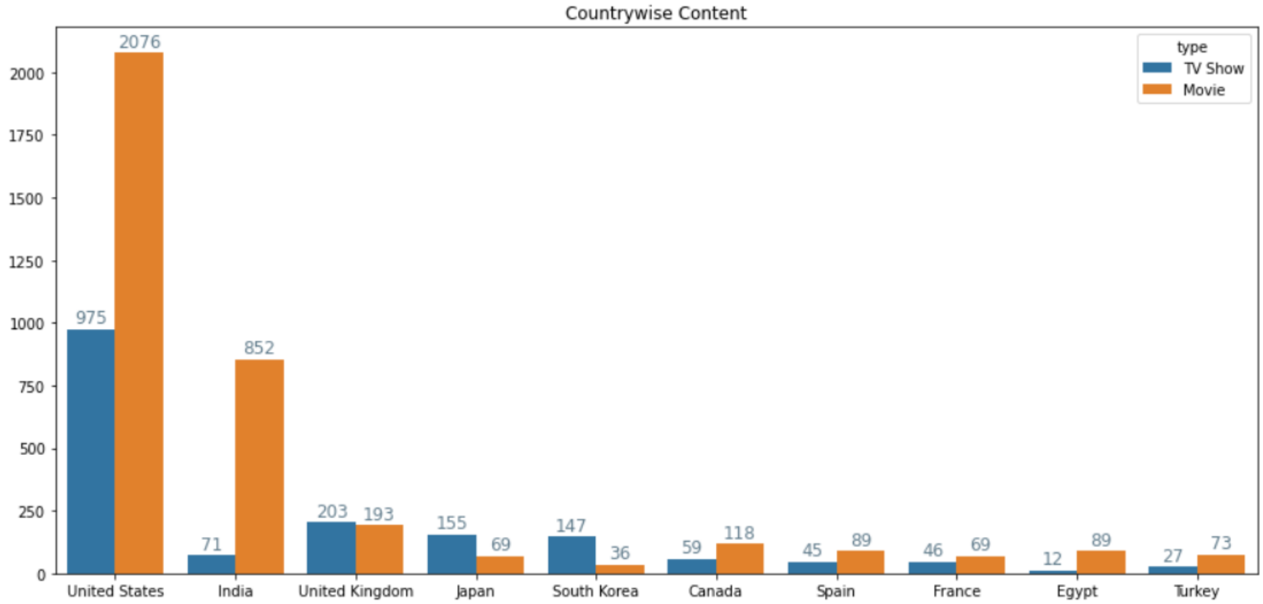
- **Data Cleaning**: Data Cleaning is one of the important steps before we start building models, in fact, there will be a significant increase in Model Performance when we have a clean, rich dataset. So here, we decided to replace null values with text, “No Data”.
    - There are 2689 null values in Director column
    - There are 718 null values in cast column
    - There are 507 null values in country column
    - There are 10 null values in date added column
    - There are 7 null values in rating column

Type of Content available:

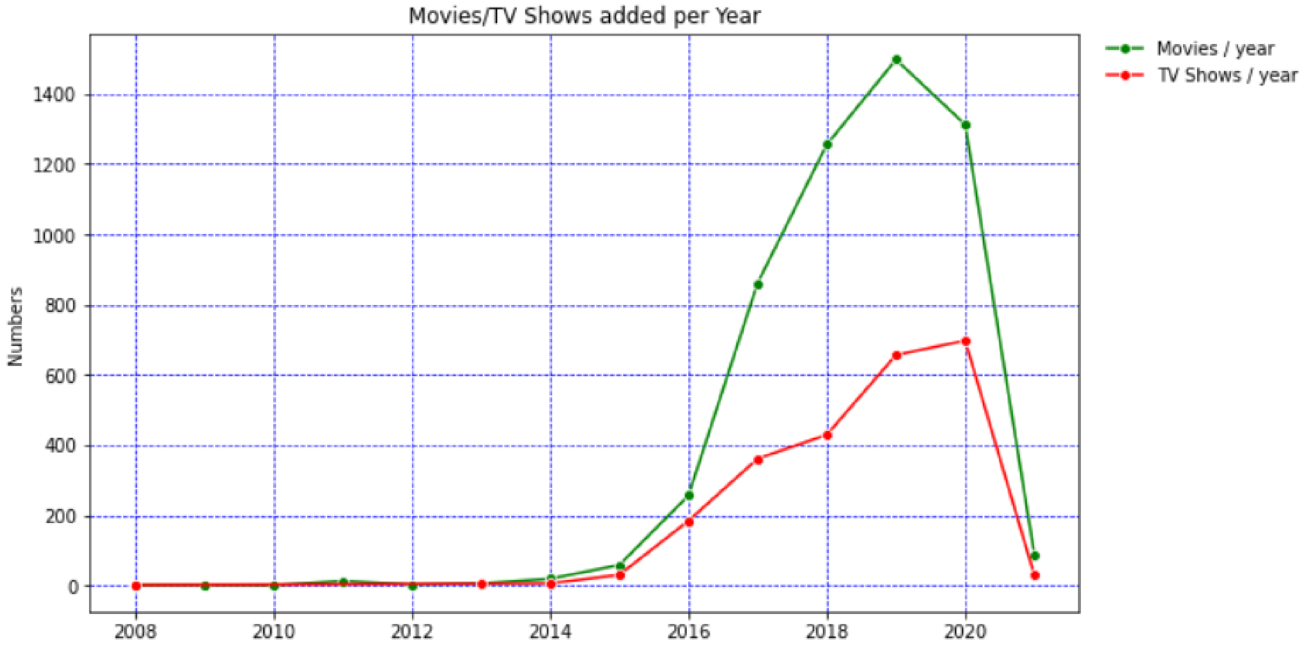


Yearly number of Movies on Netflix is more than the number of TV Shows. Almost %69 are movies while the remaining 31% are TV Shows.

Country wise Type of content:



➢Is Netflix increasingly focused on TV rather than Movies in recent years?



Yes, Netflix is increasingly focusing on TV Shows now, which is clear from the graph, from 2019 to 2020, there was a decreasing trend of Movies. The TV shows from 2019 to 2020 remains constant.

Data Preparation:

1. **Text Pre-processing:** Text preprocessing is the process of preparing text data so that machines can use the same to perform tasks like analysis, predictions, etc.

   a.) Removing Punctuations
   b.) Removing Stopwords
   c.) Stemming
   d.) Creating New Variables for Text Length

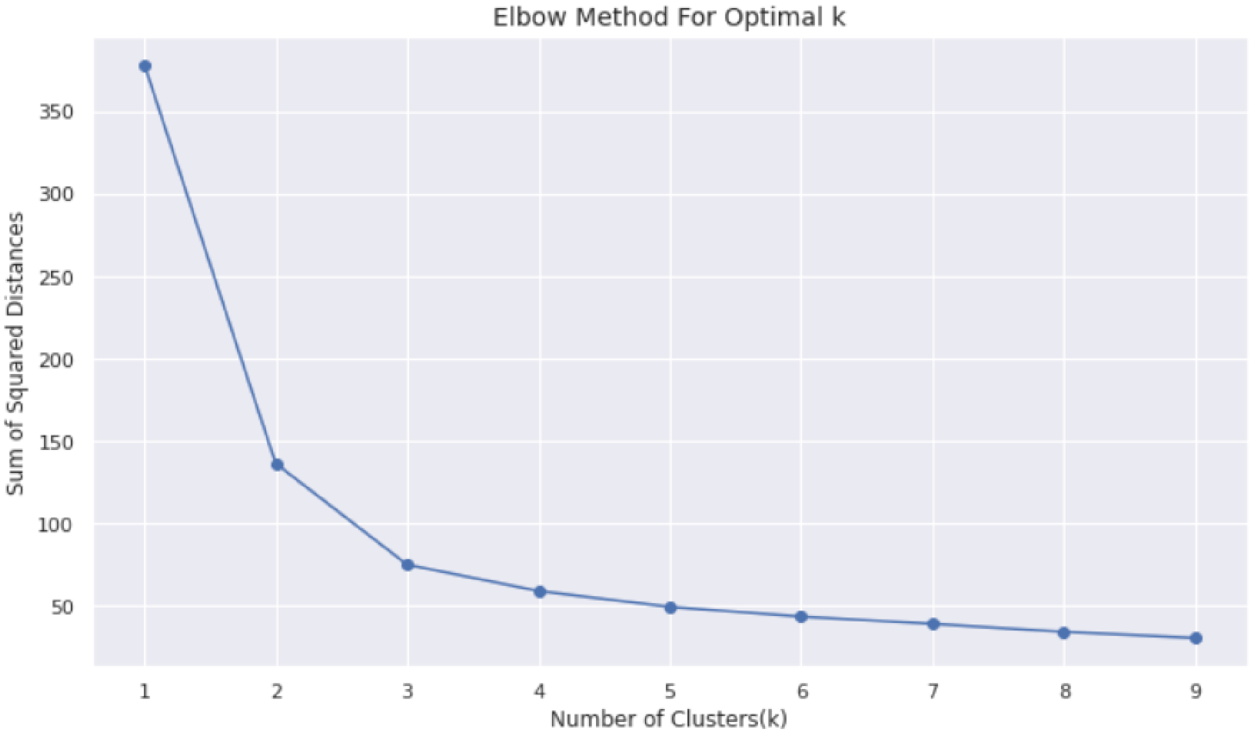
2. **Rescaling the data**

   Using Standardardization Rescaled the Description length and Listed in length of the data.

Clustering:

1. 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. We created the sample data using build blobs and used range_n_clusters to specify the number of clusters we wanted to utilize in k means.



Here we will see the output of 2,3,4,5,6 and 7 number of clusters.

Silhouette_score:

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. To calculate the Silhouette score for each observation/data point, the following distances need to be found out for each observation belonging to all the clusters.

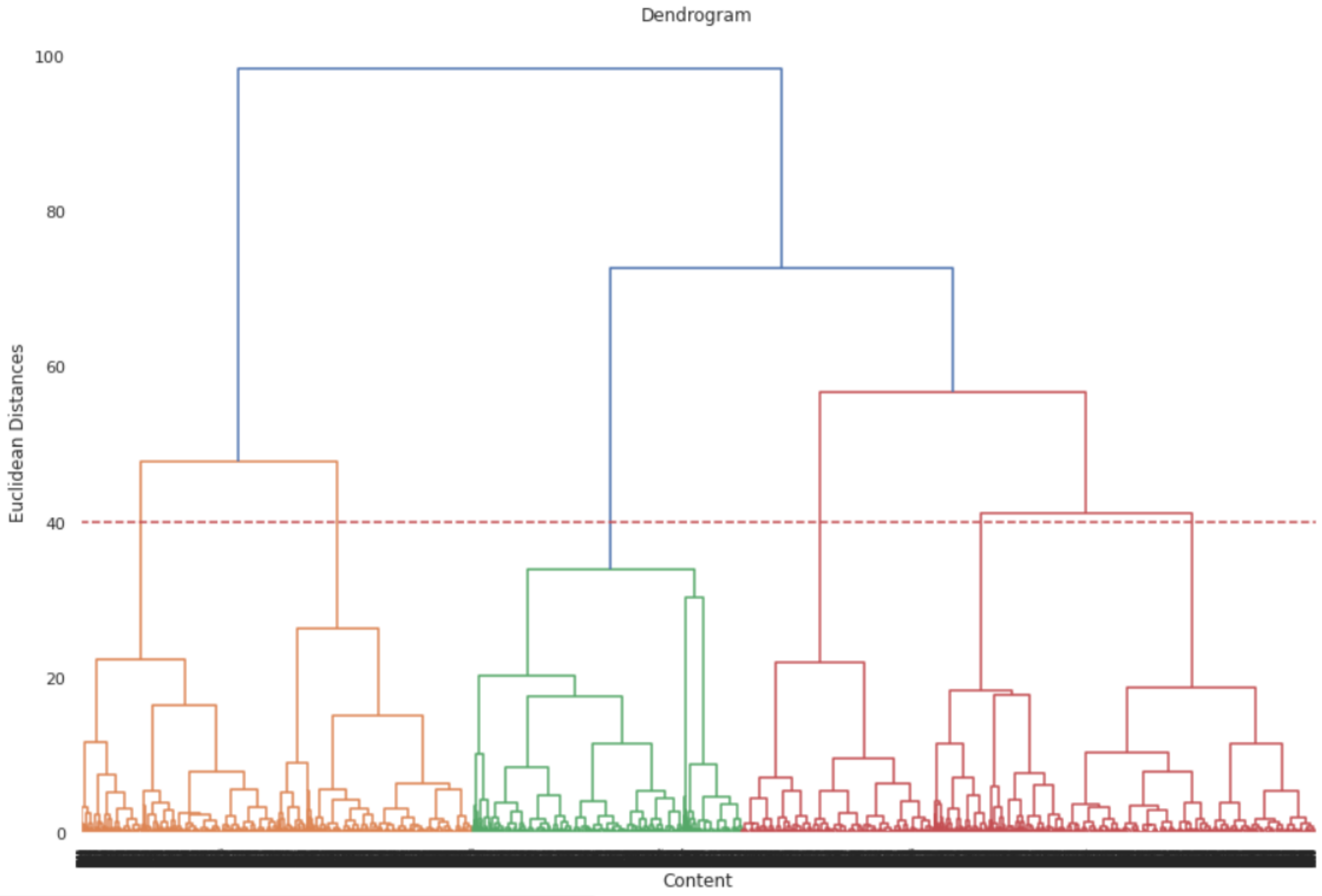
```
For n_clusters = 2, silhouette score is 0.35512756429120607
For n_clusters = 3, silhouette score is 0.3558545073559524

For n_clusters = 4, silhouette score is 0.3281130746442887

For n_clusters = 5, silhouette score is 0.33585923304123133

For n_clusters = 6, silhouette score is 0.3572596839713048

For n_clusters = 7, silhouette score is 0.35485725266624235

For n_clusters = 8, silhouette score is 0.35381821107922423
```

So in this model the 6 clusters are giving the best result. So we will consider 6 clusters as optimum clusters.

2. Hierarchical clustering:

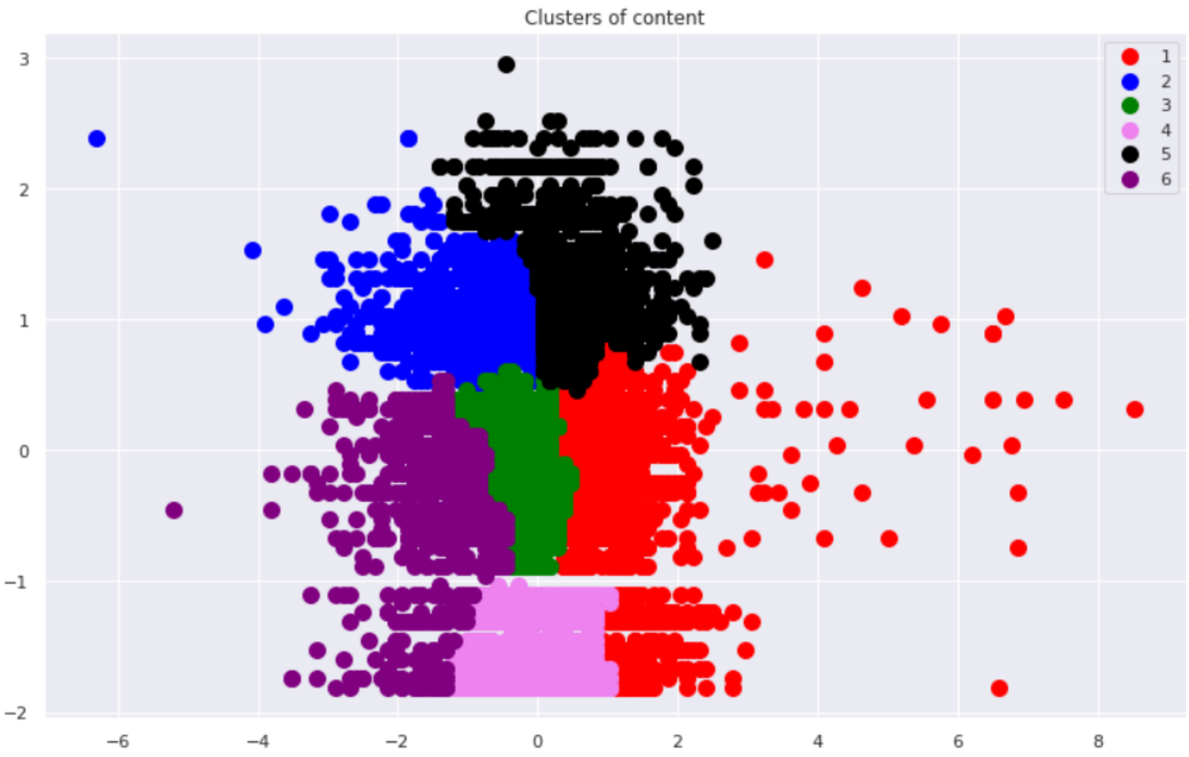
Hierarchical clustering is used to group the unlabeled datasets into a cluster and also known as hierarchical cluster analysis. In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the dendrogram. Two types of Hierarchical Clustering:

1. Agglomerative: Agglomerative is a bottom-up approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.
2. Divisive: Divisive algorithm is the reverse of the agglomerative algorithm as it is a top-down approach.



The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold, here threshold is at y=40. So we consider, no. of Cluster = 6.

The following graph shows the clustering using the agglomerative approach.



Silhouette_score:

```
For n_clusters = 2, silhouette score is 0.31733940335920435

For n_clusters = 3, silhouette score is 0.28576591463080314

For n_clusters = 4, silhouette score is 0.2970786181831897

For n_clusters = 5, silhouette score is 0.2818170622516962

For n_clusters = 6, silhouette score is 0.3004594443021899

For n_clusters = 7, silhouette score is 0.31843009608053685

For n_clusters = 8, silhouette score is 0.3102634162499398
```

Thus from the hierarchical clustering, the dataset can be grouped into 7 clusters.

Conclusion:

After preprocessing the data, I started with EDA to understand the trends and features of the provided dataset. Major findings from EDA are as follows:

- There is 69% movie content and 31% TV show content in the dataset.
- The United States accounts for the majority of the content created on Netflix, numbering 3051 titles. India is the second largest with 923 titles.
- All the top 10 countries producing the highest amount of content have a higher proportion of movies than TV shows except Japan, South Korea and United Kingdom.
- Netflix is increasingly focusing on TV Shows now, which is clear from the graph, from 2019 to 2020, there was a decreasing trend of Movies. The TV shows from 2019 to 2020 remains constant.
- TV-MA, TV-14 and TV-PG are the top content ratings to which the highest number of content belongs.
- Among the top 10 countries with highest content volume, the United States and India have the highest volume of content appropriate for teens and most other countries have highest volume of content appropriate for adults.
- Content onboarding on Netflix started increasing from 2015 and reached its peak in 2019, with a drastic downfall in 2021. The downfall can be attributed to Covid pandemic.
- International movies is the highest content genre in movies and International TV Shows is the highest one in TV shows.
- TV shows with one season are highest in number.

Findings after applying k- means and Hierarchical clustering:

- K-means clustering is used to form the clusters of clusters of the content. To find out the optimal value of k, the Elbow and Silhouette method was used. K=6 was an appropriate value and can be visualized after clustering the content using k-means clustering.
- Hierarchical clustering implementation gives a perfect score of silhouette at 7 clusters,the same as visualized using a dendrogram.

References:

[1] Applied Science Article MDPI

[2] GeeksforGeeks

[3] Wikipedia

[4] DataCamp