Capstone Project - 4

Netflix Movies and TV Shows Clustering

Team Members

1 Arvind Kale

2 Priyanka Jain

3 Mohd Sakib Quraishi

4 Nitish kumar

Contents :

- Introduction
- Problem Statement
- Data Description
- Null Value
- Exploratory Data Analysis
- Data Cleaning
- Topic modelling
- Model Implementation
- Data Pre-processing
- Model Implementation
- K- Means
- Cluster Analysis

Introduction

Netflix is a media distribution company. It started with DVD distribution via mail, but has evolved substantially over the course of its existence. Today, Netflix is focused on streaming video. Some of its content is licensed, and some of the content is produced in-house. Netflix originally focused on movies, but today television shows are probably the more common format. Netflix works on a subscription model, where users get unlimited access to content with a paid subscription.

Problem Statement

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flexible 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 services 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.

Let us try to solve.

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

Data Description

- **The data was collected from Flexible which is third party Netflix search engine. The dataset consists of movies and TV shows data till 2019. The dataset has 7787 rows of data.**

- **The dataset consists of eleven textual columns and one numeric Column.**

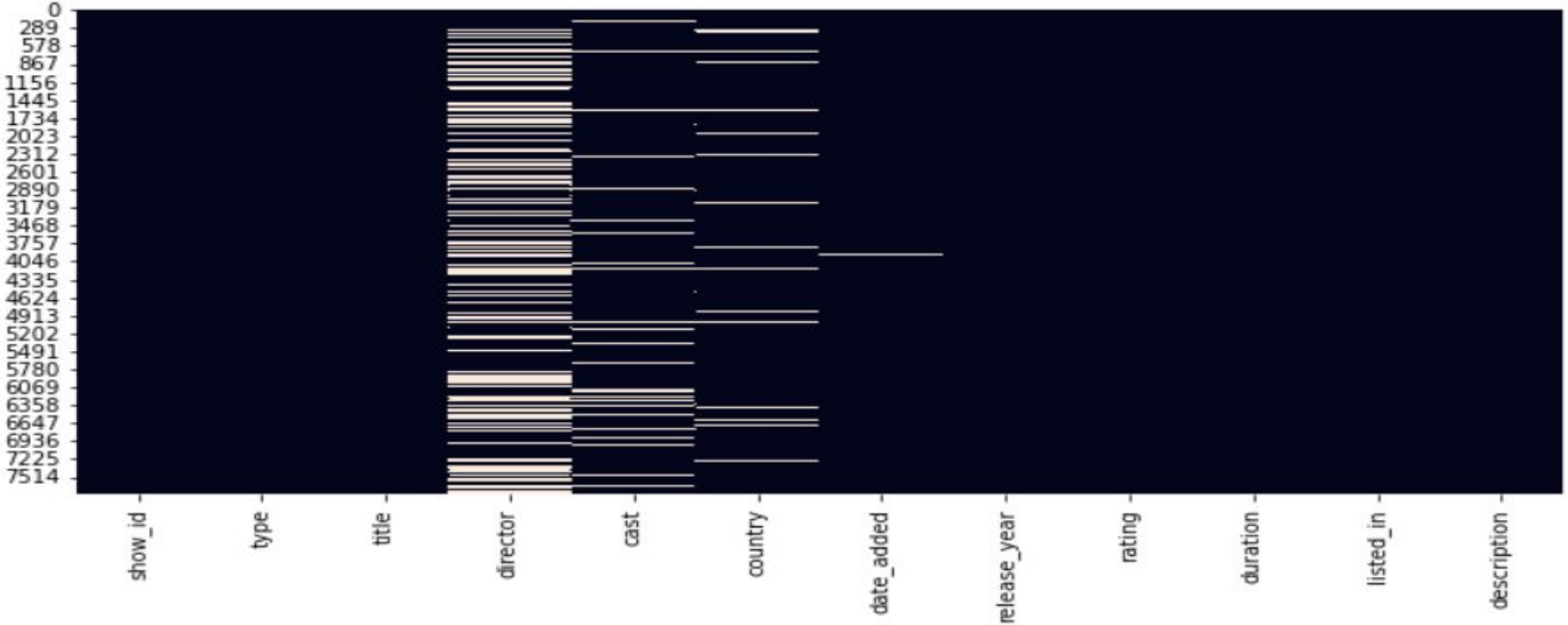
Attribute Information :

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

Data Description

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

Null Value



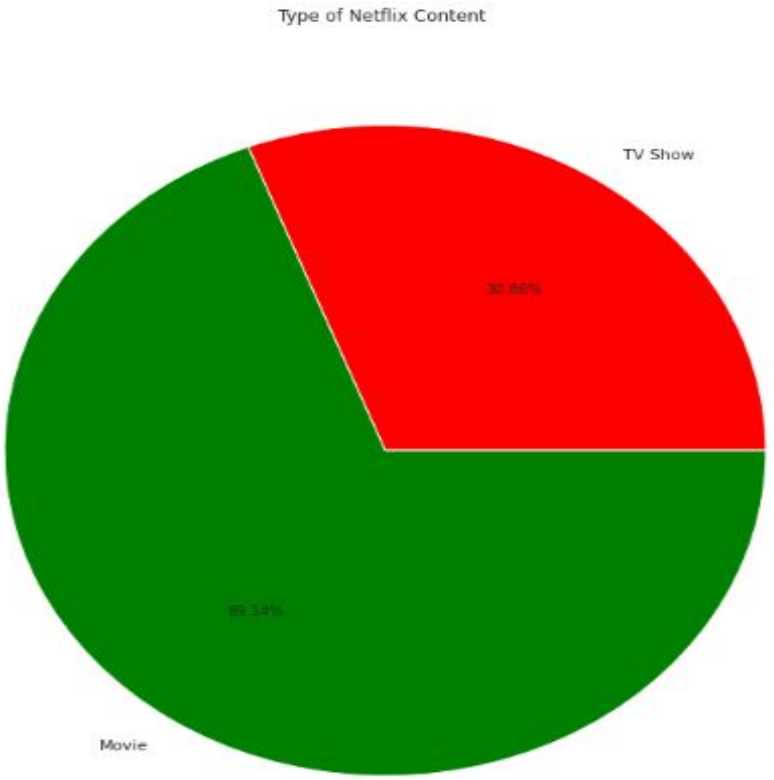
Column director,cast,country and date_added contains null values.

Null Value Treatment:

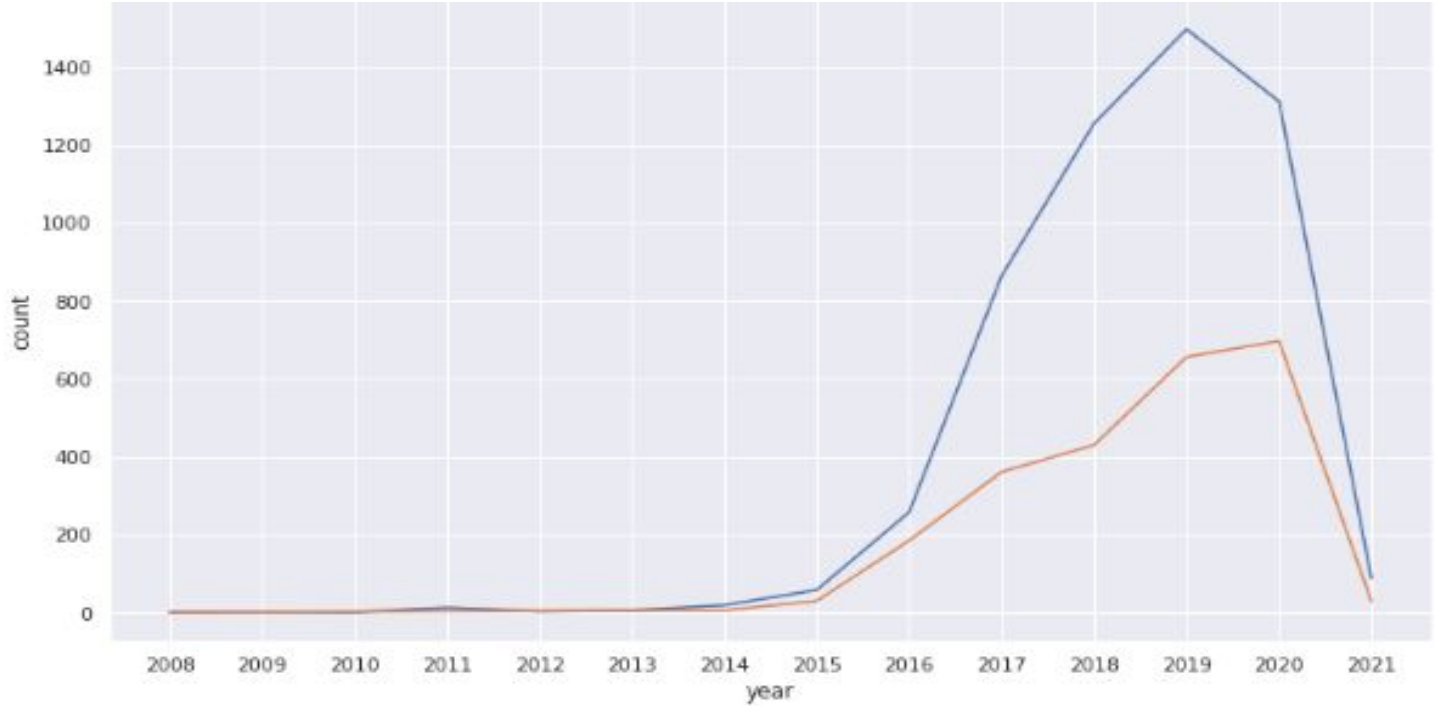
- **Director** feature have more than **30.68%** of null values. Filling null values by ‘unknown’.
- **Country** feature have **6.51%** of null values. Filling null values by mode of feature.
- **Cast feature** have **9.22%** of null values. Filling null values by ‘unknown’.
- **Rating** feature have **0.09%** of null values. Filling null values by mode of feature.
- **Date_added** feature have **0.13%** of null values. Dropping rows corresponding to null values.

Exploratory Data Analysis

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



EDA of number of movie and tv show in a year



Blue- TV show

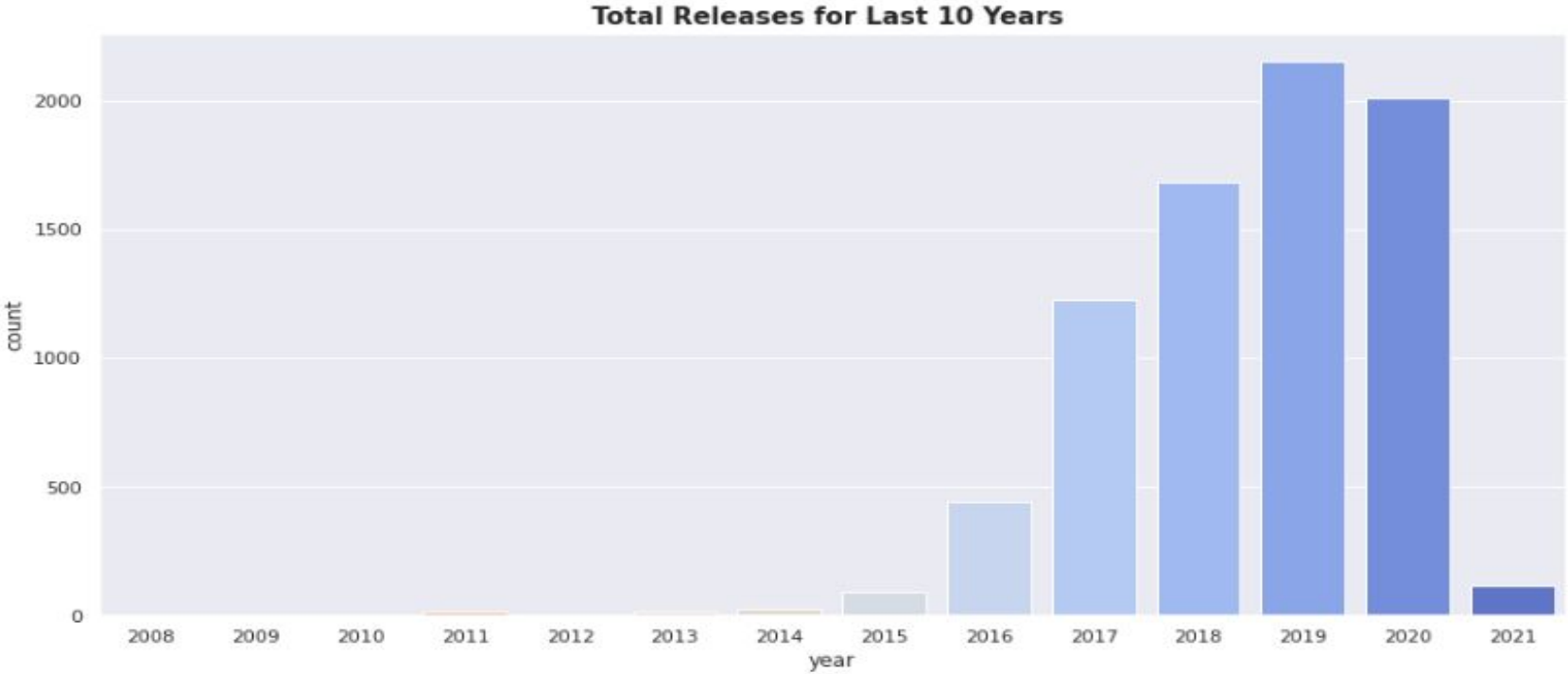
Orange- Movies

Growth in the number of movies on Netflix is much higher than tv shows

From 2015 we can see a noticeable addition in the number of movies and tv shows uploaded by Netflix on its platform.

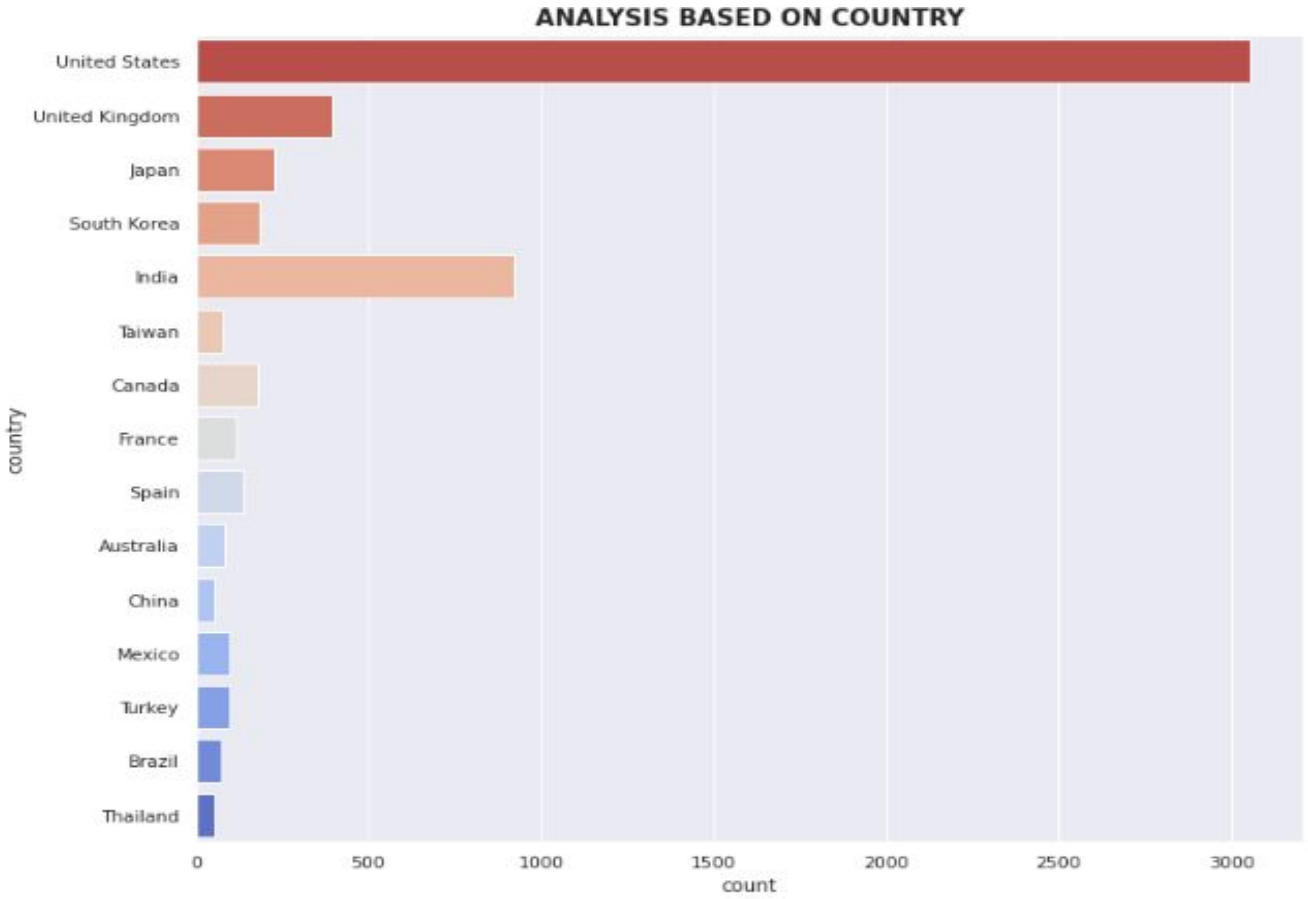
.

EDA Of Total release for last 10 years



The number of release have significantly increased after 2015 and have dropped in 2021 because of Covid 19

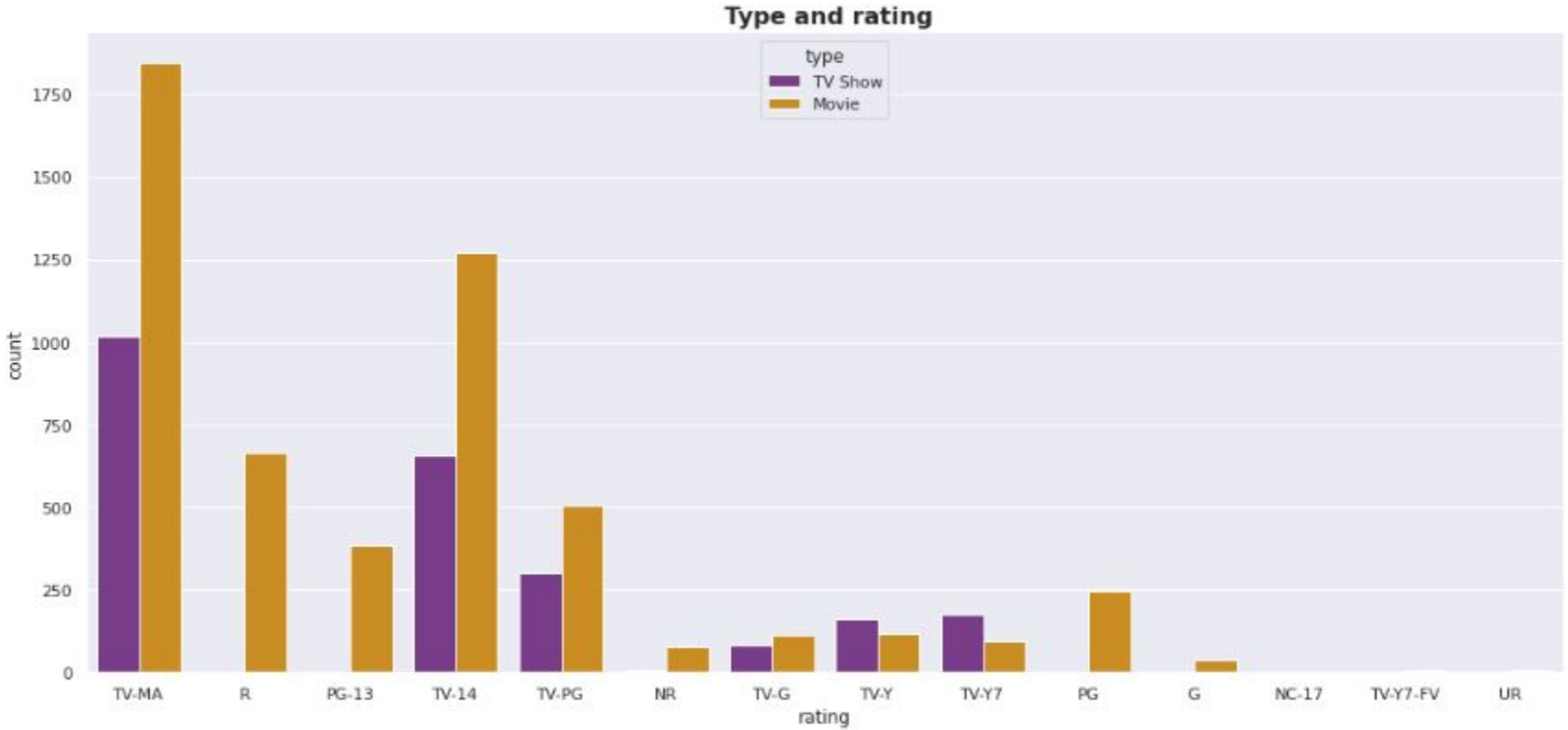
ANALYSIS Based ON country



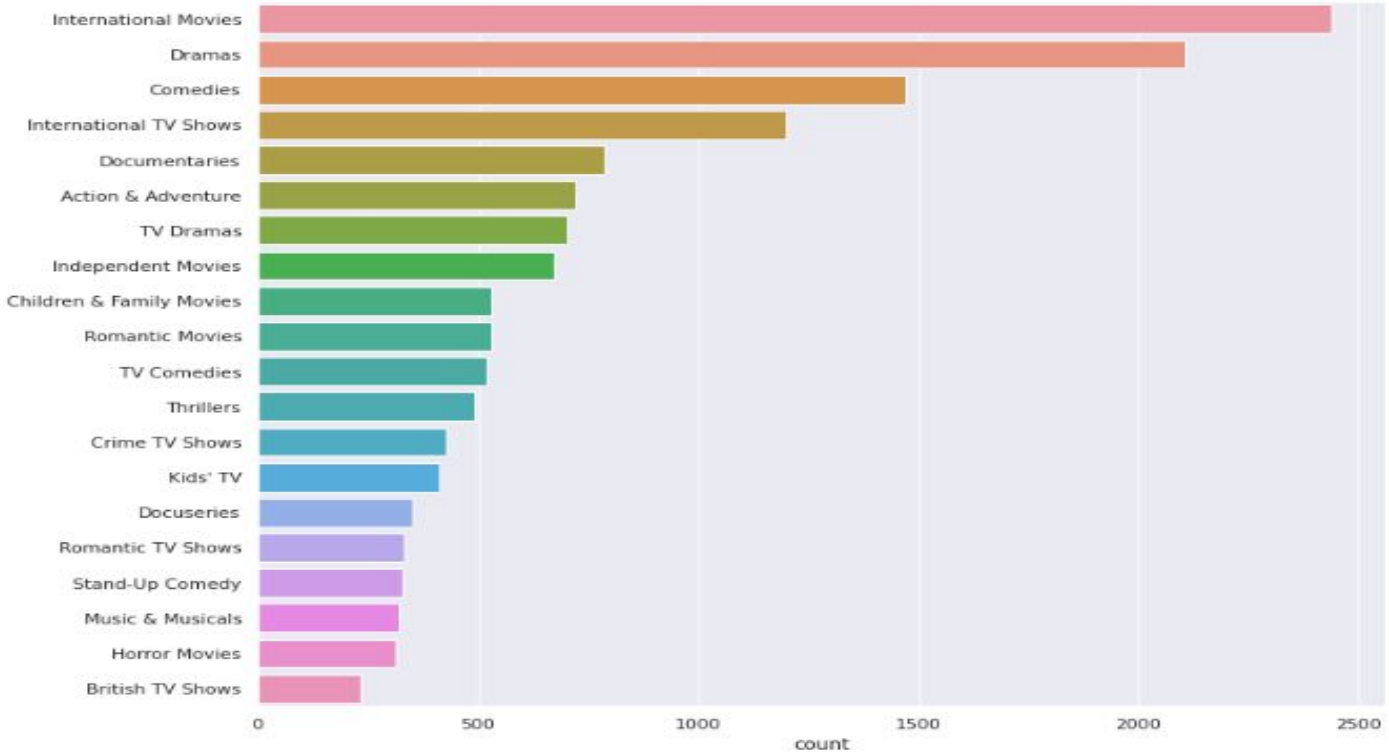
United states have the most number of content

India have second highest content on Netflix

EDA Of Rating wise content count

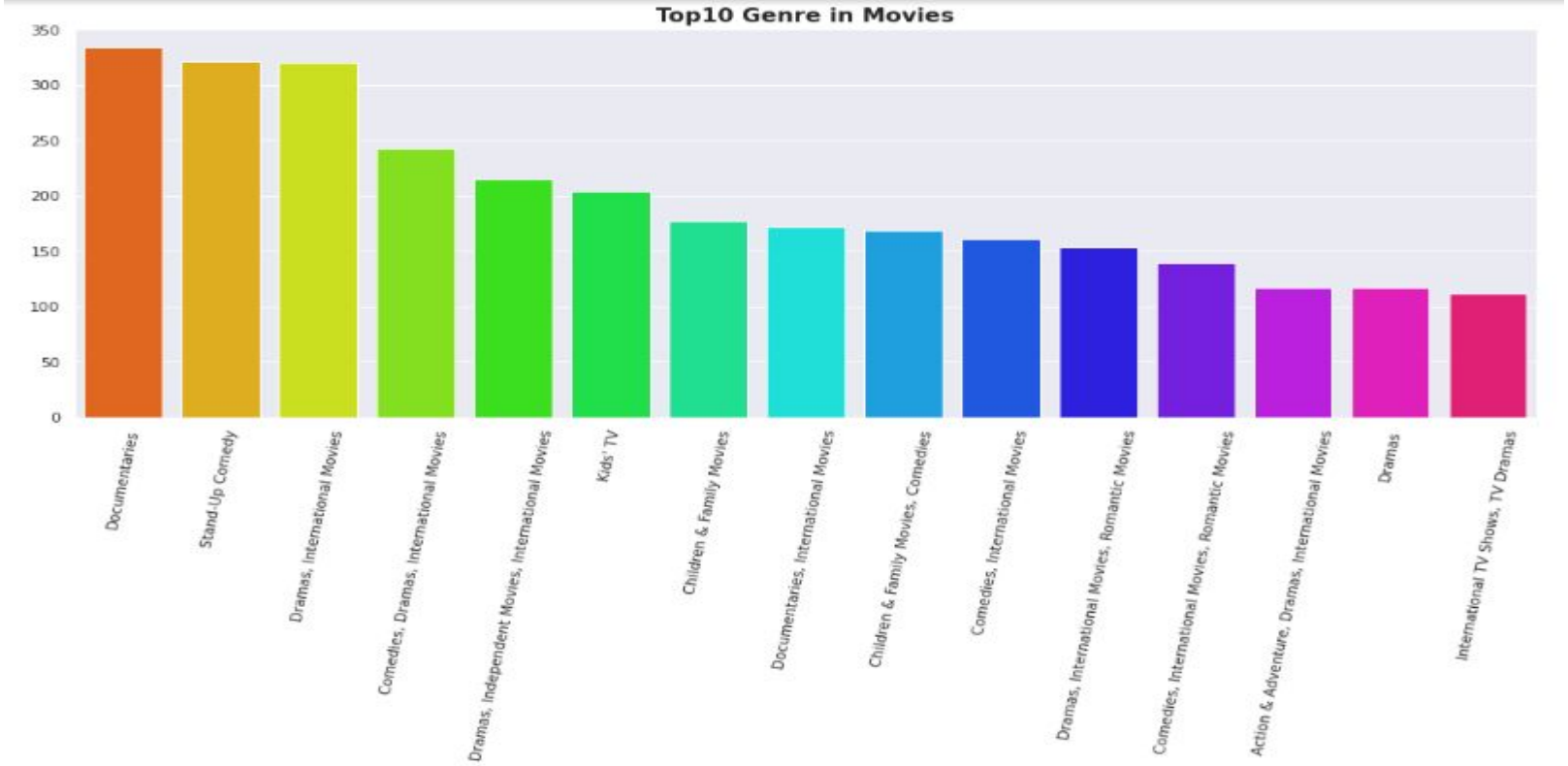


Content wise Analysis



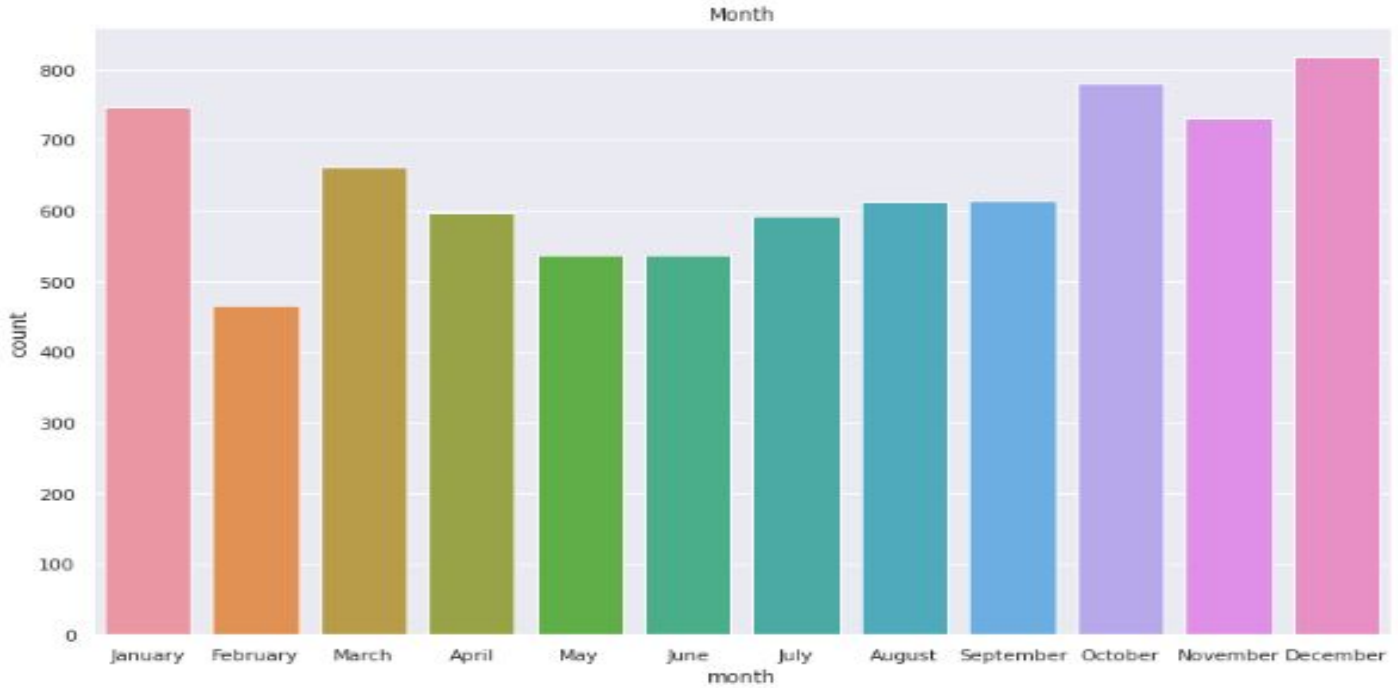
International movies and drama has highest number of content

Top genre added in netflix



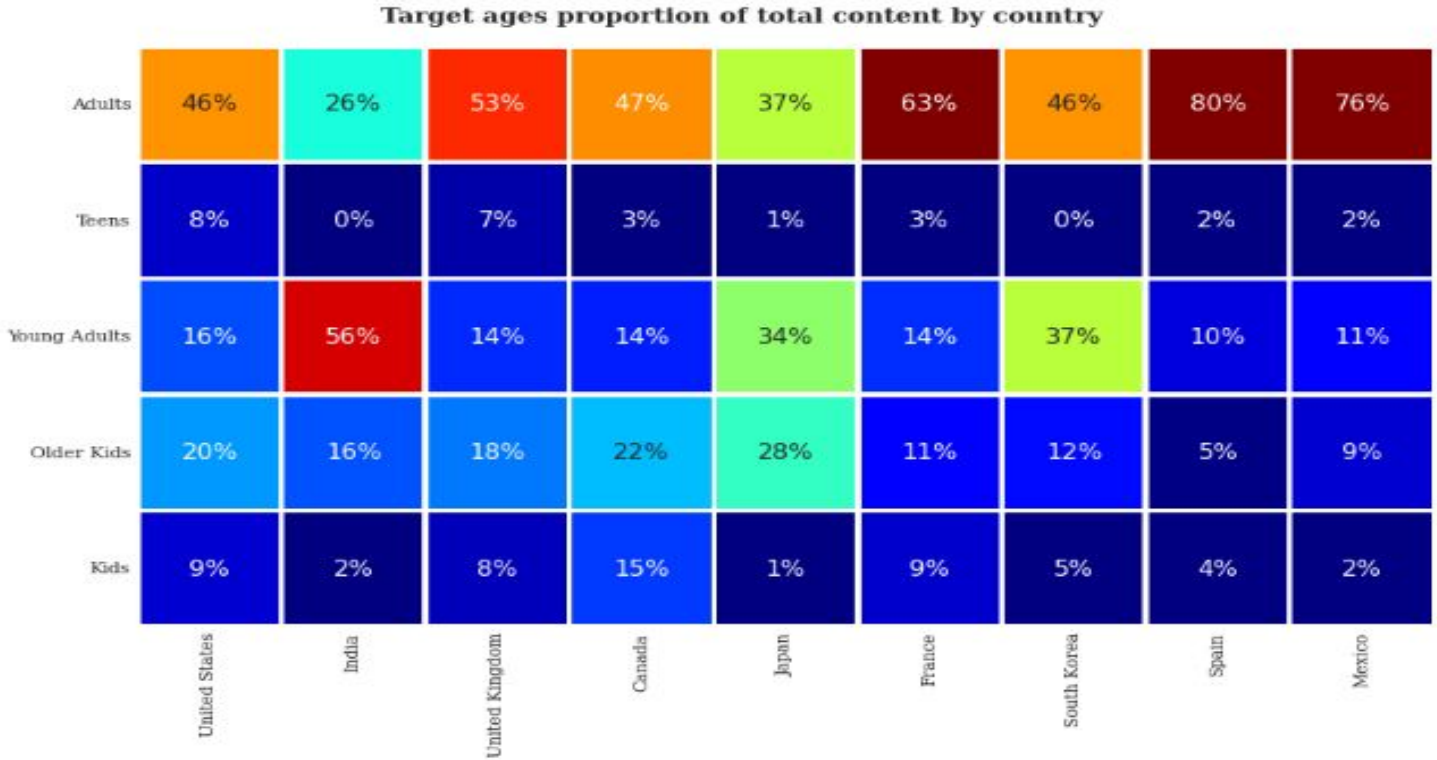
Drama is the most popular genre followed by comedy

Month wise analysis of Releases movie



we can see that, In netflix maximum content added in December and minimum in february

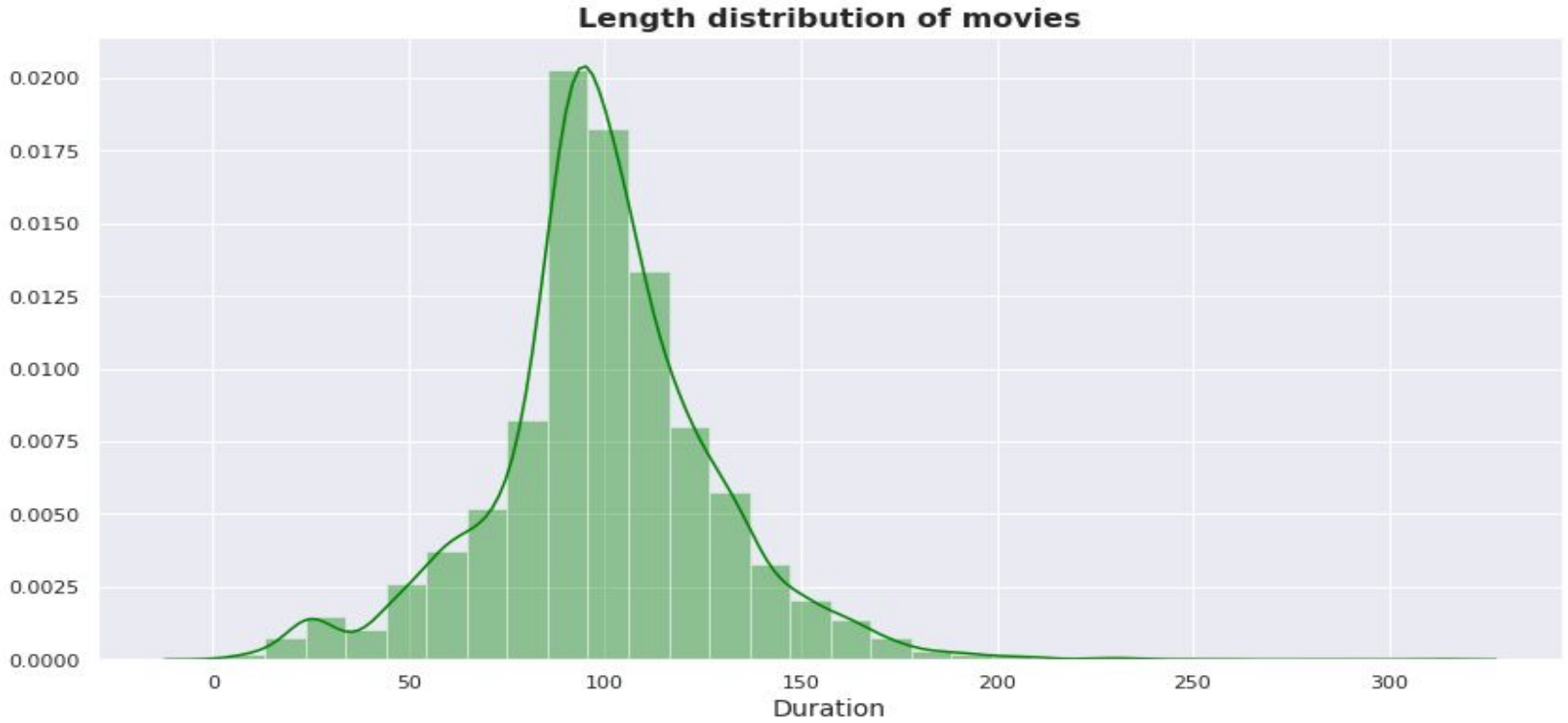
Correlation Heatmap



Available movies and tv show for different age groups in top 10 countries

This shows that the available content in different countries is maximum for adults

EDA Of Duration distribution of Movies



This graph shows Most of the movies last for 90 to 120 minutes

Data Preprocessing

- **Label Encoding-** refers to converting the labels into a numeric form so as to convert them into the machine-readable form.
- **Lemmatization-** Lemmatization, unlike Stemming, reduces the inflected words properly ensuring that the root word belongs to the language. In Lemmatization root word is called Lemma. ... For example, runs, running, ran are all forms of the word run, therefore run is the lemma of all these words.
- **Removing Stop words -** To remove stop words from a sentence, you can divide your text into words and then remove the word if it exits in the list of stop words provided by NLTK.
- **Tf - idf Vectorization -** TF-IDF stands for "Term Frequency — Inverse Document Frequency". This is a technique to quantify a word in documents, we generally compute a weight to each word which signifies the importance of the word in the document and corpus. This method is a widely used technique in Information Retrieval and Text Mining.
- **Min-max Scaling -** For each value in a feature, MinMaxScaler subtracts the minimum value in the feature and then divides by the range. It preserves shape of original distribution.

Topic Modeling (LDA and LSA)

- **Latent Semantic Analysis**(LSA) is used to find the hidden topics represented by the document or text. This hidden topics then are used for clustering the similar documents together. LSA is an unsupervised algorithm and hence we don't know the actual topic of the document.
- In natural language processing, the **Latent Dirichlet Allocatio**n (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.

Creating Clusters:

AI

What is clustering?

Clustering is the task of dividing the data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

How to cluster similar data?

To create clusters we will use the K-Means Clustering; which is an iterative process in which the dataset is grouped into k number of predefined non-overlapping clusters or subgroups, making the inner points of the cluster as similar as possible while trying to keep the clusters at distinct space it allocates the data points to a cluster so that the sum of the squared distance between the clusters centroid and the data point is at a minimum.

Silhouette Score Method

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.

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.**

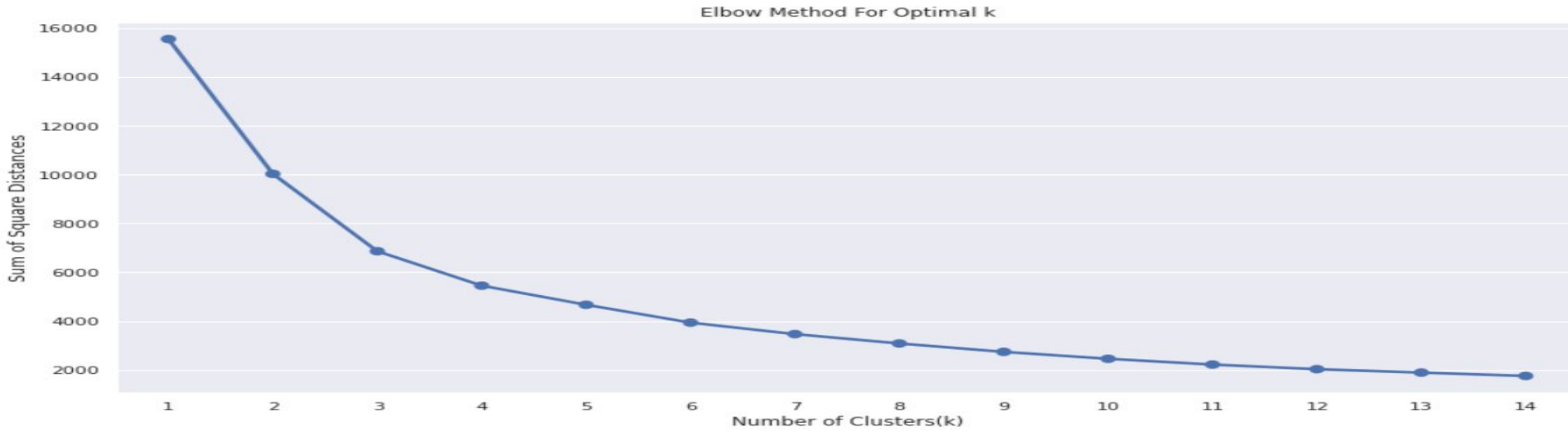
$$S = \frac{(b - a)}{max(a, b)}$$

*Using this method in our dataset we found the optimal number of cluster is equal to 3.

Elbow curve to find optimal value of cluster k:

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

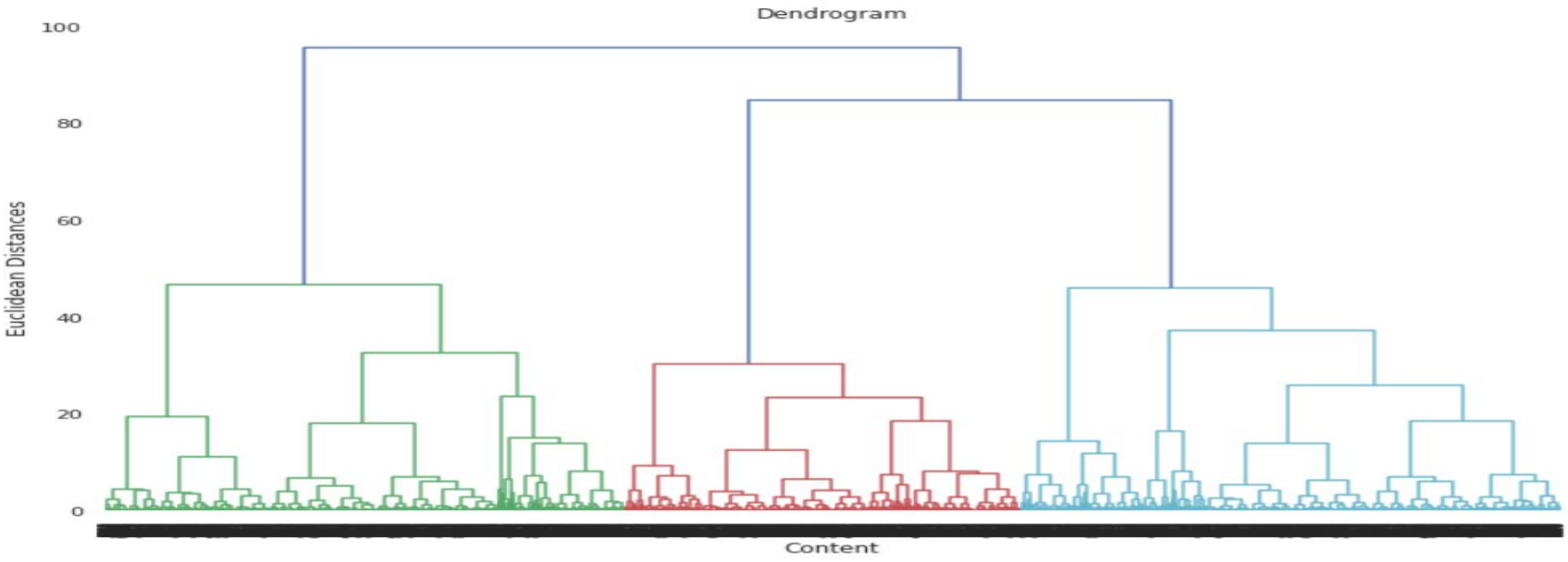
To determine the optimal number of clusters, we have to select the value of k at the “elbow” ie the point after which the distortion/inertia start decreasing in a linear fashion. Thus for the given data, we conclude that the optimal number of clusters for the data is **3**.



Dendrogram to find the optimal number of clusters(Hierarchical Clustering)

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

No. of Cluster = 3



Conclusion:

1. Data set contains 7787 rows and 12 columns in that cast and director features contains large number of missing values so we can drop it and we have 10 feature for the further implementation

2. We have two types of content TV shows and Movies (30.86% contains TV shows and 69.14% contains Movies)

3. Jan Suter is the most popular directors on Netflix with the most titles are mainly international as well.

4. Anupam Kher and Shahrukh Khan are the most popular actor

5. By analysing the content added over years we get to know that in recent years netflix is focusi ng movies than TV shows (movies is increased by 80% and TV shows is increased by 73% co mpare to 2016 data)

Conclusion(continued):

6. The most number of the movies and TV shows release in 2017 and 2020 respectively a nd united nation have the maximum content on netflix
7. On Netflix, Dramas genre contains the maximum content among all of the genres and t he most of the content added in december month and less content in february
8. By applying the silhouette score method for n range clusters on dataset we got best sco re which is 0.348 for 3 clusters it means content explained well on their own clusters, by using elbow method after k = 3 curve gets linear it means k = 3 will be the best cluster

9. Applied different clustering models Kmeans, hierarchical, Agglomerative clustering on data we got the best cluster arrangments

10. By applying different clustering algorithms to our dataset .we get the optimal number o f cluster is equal to 3

Thank You!