Capstone Project

Netflix Movies and TV Shows Clustering

By – Sneha Raikar

Agenda

- Defining Problem Statement
- Data Pipeline
- Data Summary
- EDA
- Applying Model Clustering
- Conclusion

Problem Statement

- Netflix has become the dominant company in the on-demand media industry, with 167 million paying subscribers around the world.
- We have been provided a dataset collected from Flexible which is a third-party Netflix search engine.

Our job is to:-

- To perform Exploratory Data Analysis
- Understanding what type of content is available in different countries
- Is Netflix increasingly focused on TV rather than movies in recent years.
- Clustering similar content by matching text-based features

Data pipeline

- **Data Pre-processing:** After exploring and understanding our data, we did data cleaning by handling Null/missing values, and checking for duplicate values. We further changed the "date_added" variable to its appropriate DateTime format and created a new variable "year_added" by extracting year from it.

- **EDA:** We performed an exploratory analysis of data and found useful insights.

- **Creating a model:** After identifying useful features,we performed text cleaning- by removing stopwords, and punctuation and doing stemming of words. After calculating clean text lengths, we standardize those values and applied two clustering algorithms- K-means and HAC (Hierarchal Agglomerative Clustering).

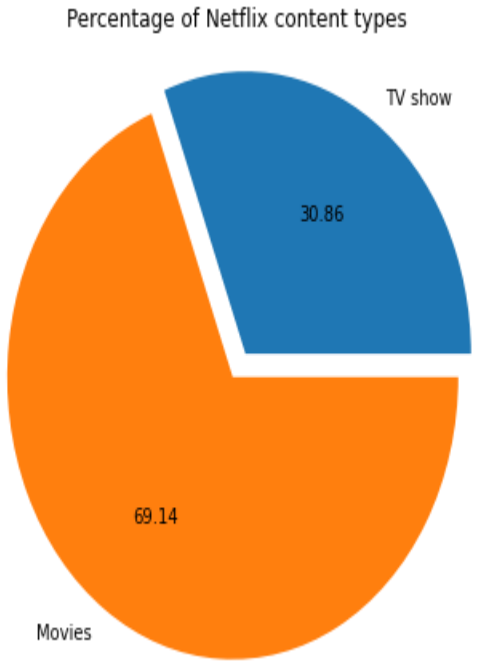
Data summary

- **show_id** : Unique ID for every Movie / Tv Show
- **type** : Shows it is a Movie or TV Show
- **title** : Gives 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 of Movies and TV Shows
- **description**: The Summary description of movies and TV shows

Exploratory Data Analysis

- Clearly number of Movies on Netflix outnumbered the number of TV Shows.
- Almost 70% content are movies while rest 30% are TV Shows.

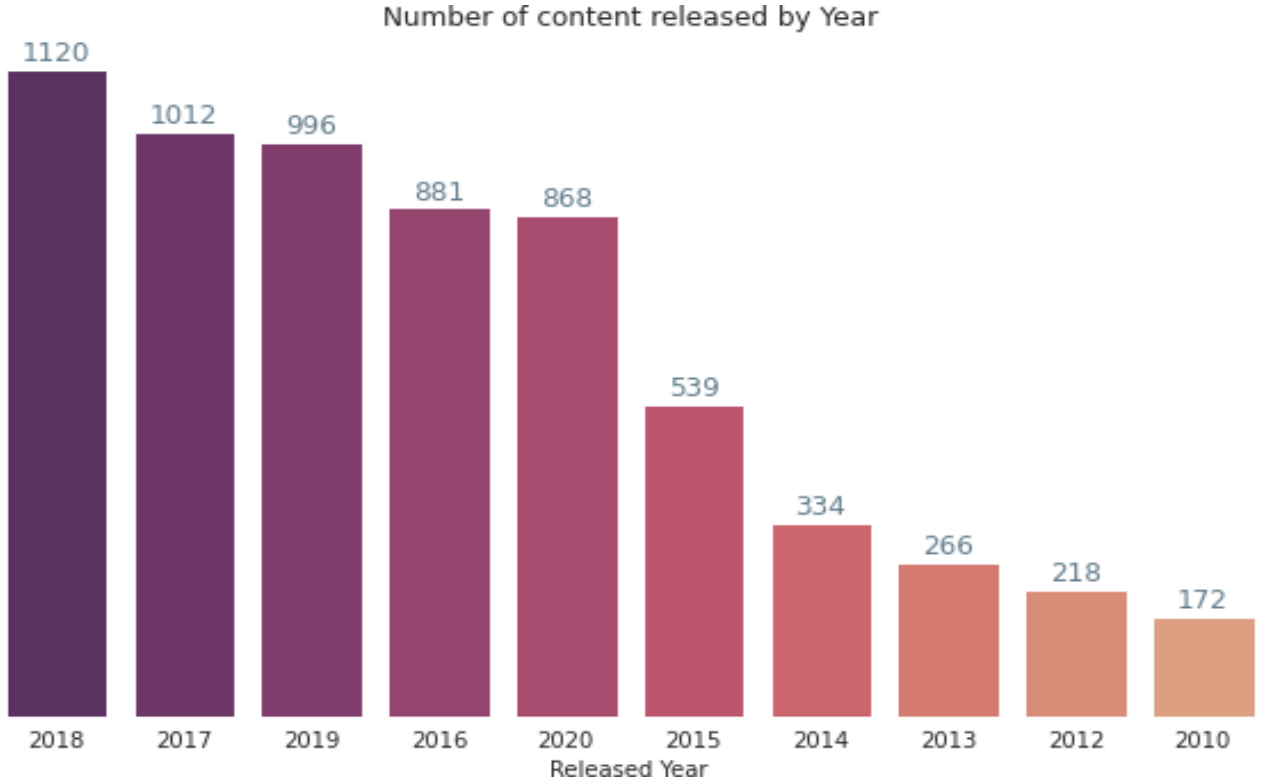
Share of TV Show and Movie in dataset



Exploratory Data Analysis

- As per the contents available on Netflix, most of them are from recent years – i.e 2018,2017 and 2019.

- The trend shows that as we go from year 2010 to 2018, the number of contents on the basis of respective year release increases.

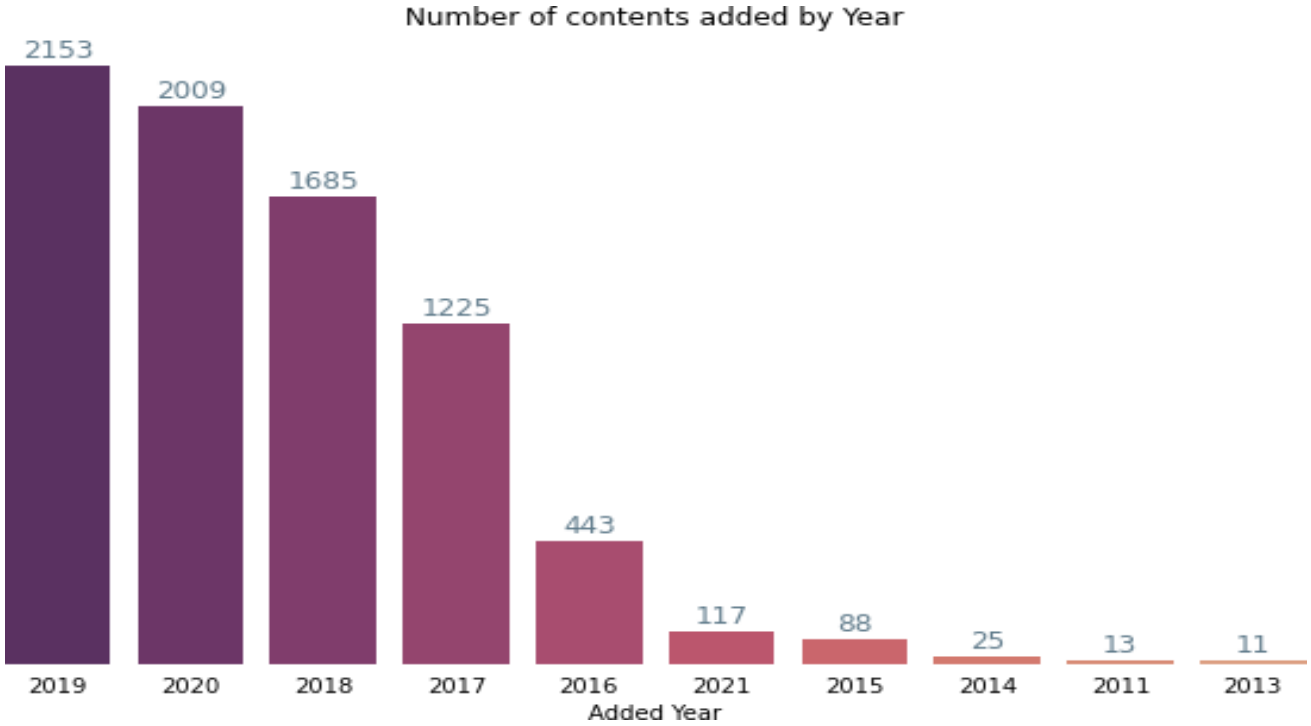


- Clearly, we can see the number of contents on the platform is increasing year on year.

- There is drastic increase in 2016-2017, this is probably Netflix was launched in India this year. and to attract Indian viewers it started adding Indian contents as well.

Exploratory Data Analysis

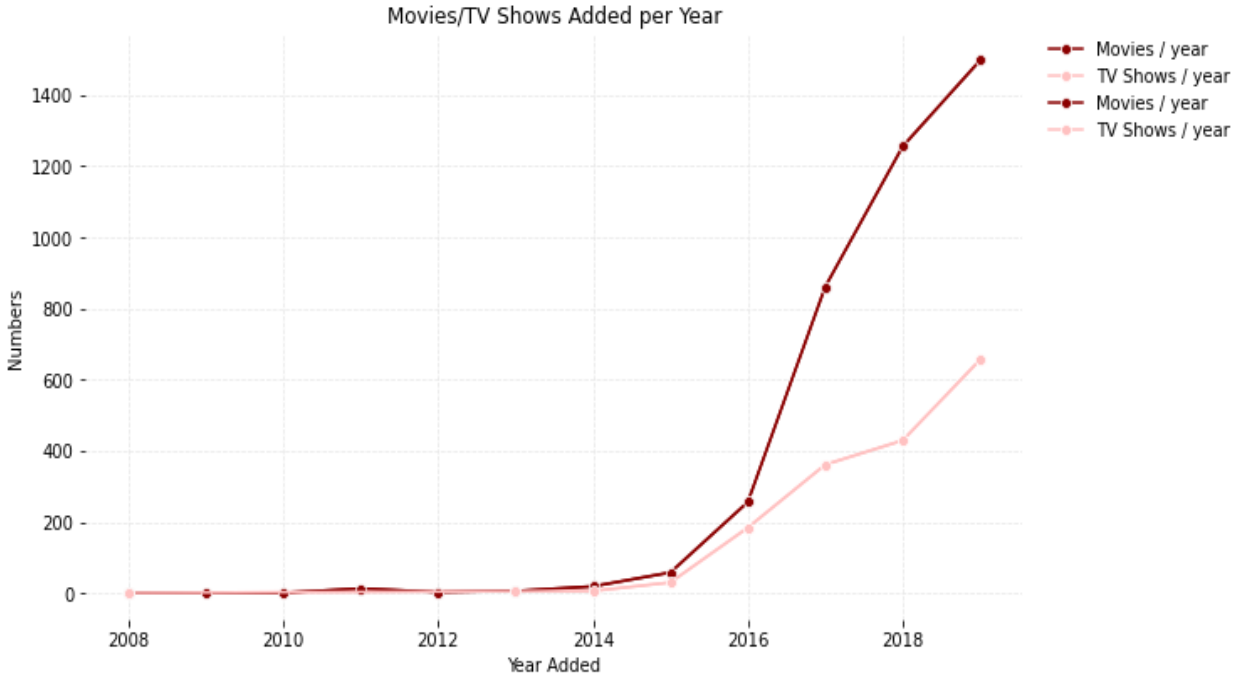
Distribution of content added on Netflix by year



Exploratory Data Analysis

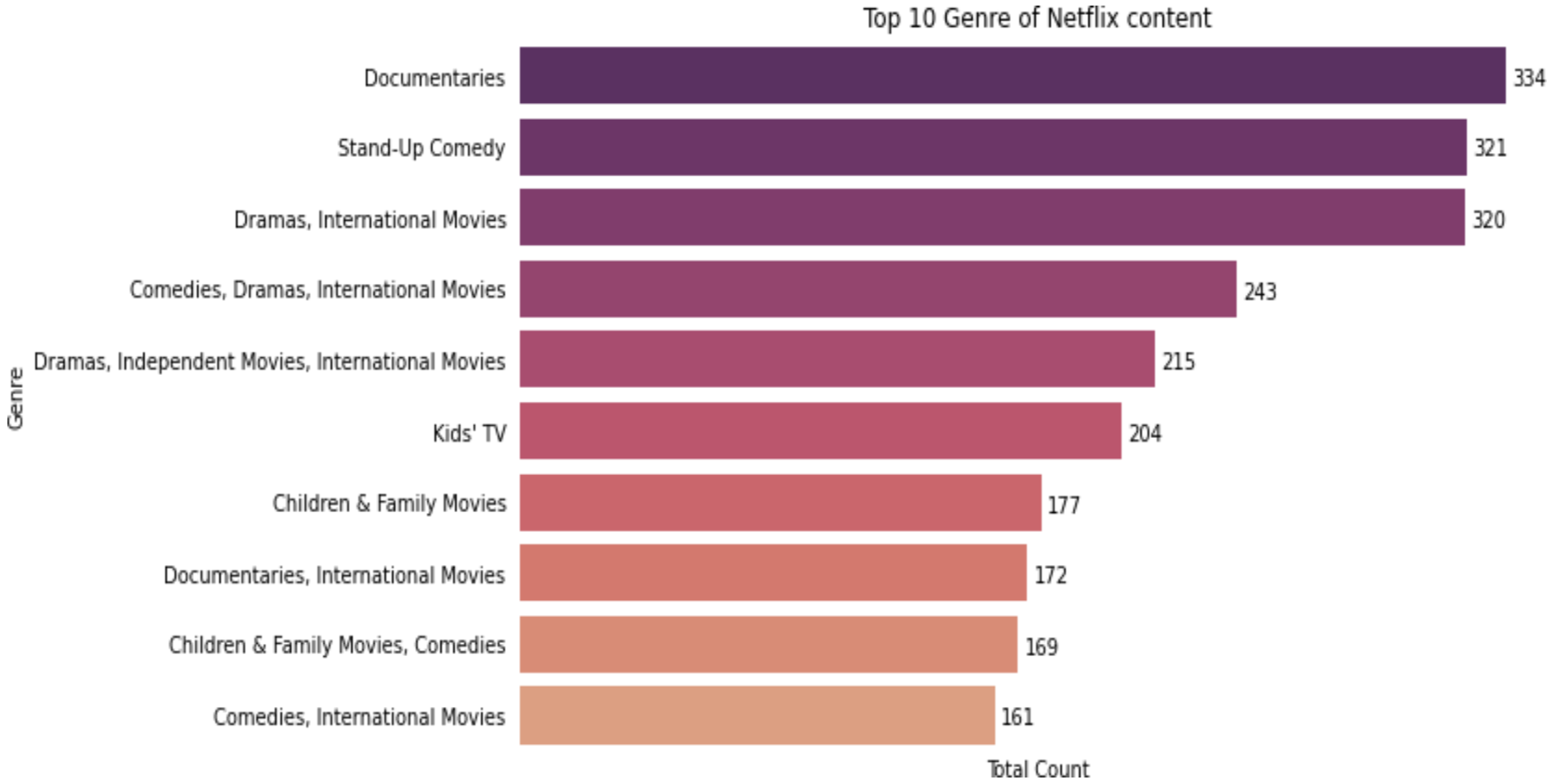
Distribution of ratings of content added on Netflix

- The plot shows with a rating of TV-MA are in the clear majority. This is followed by TV-14.
- These top two ratings have way too much numbers of contents compared to other ratings.



Exploratory Data Analysis

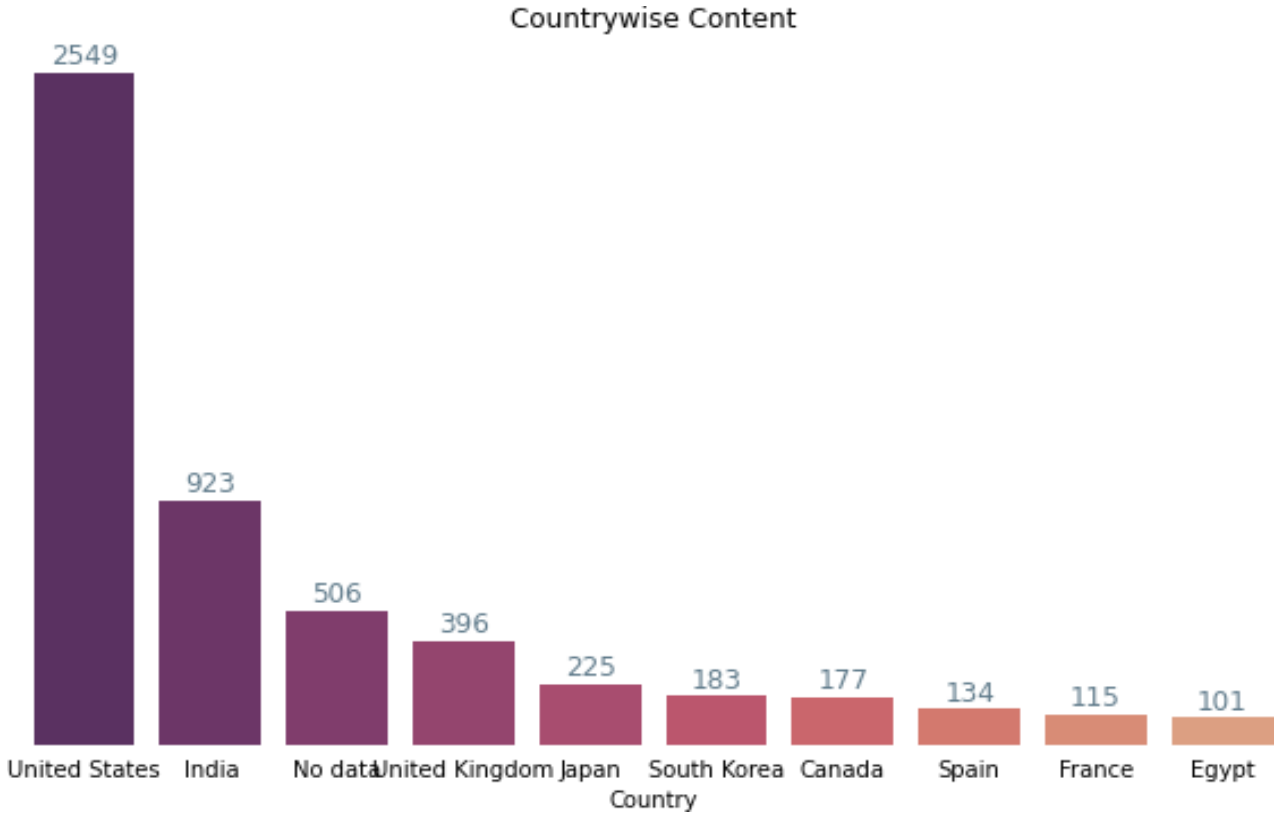
Top content genre available in Netflix



Exploratory Data Analysis

Top content producing countries

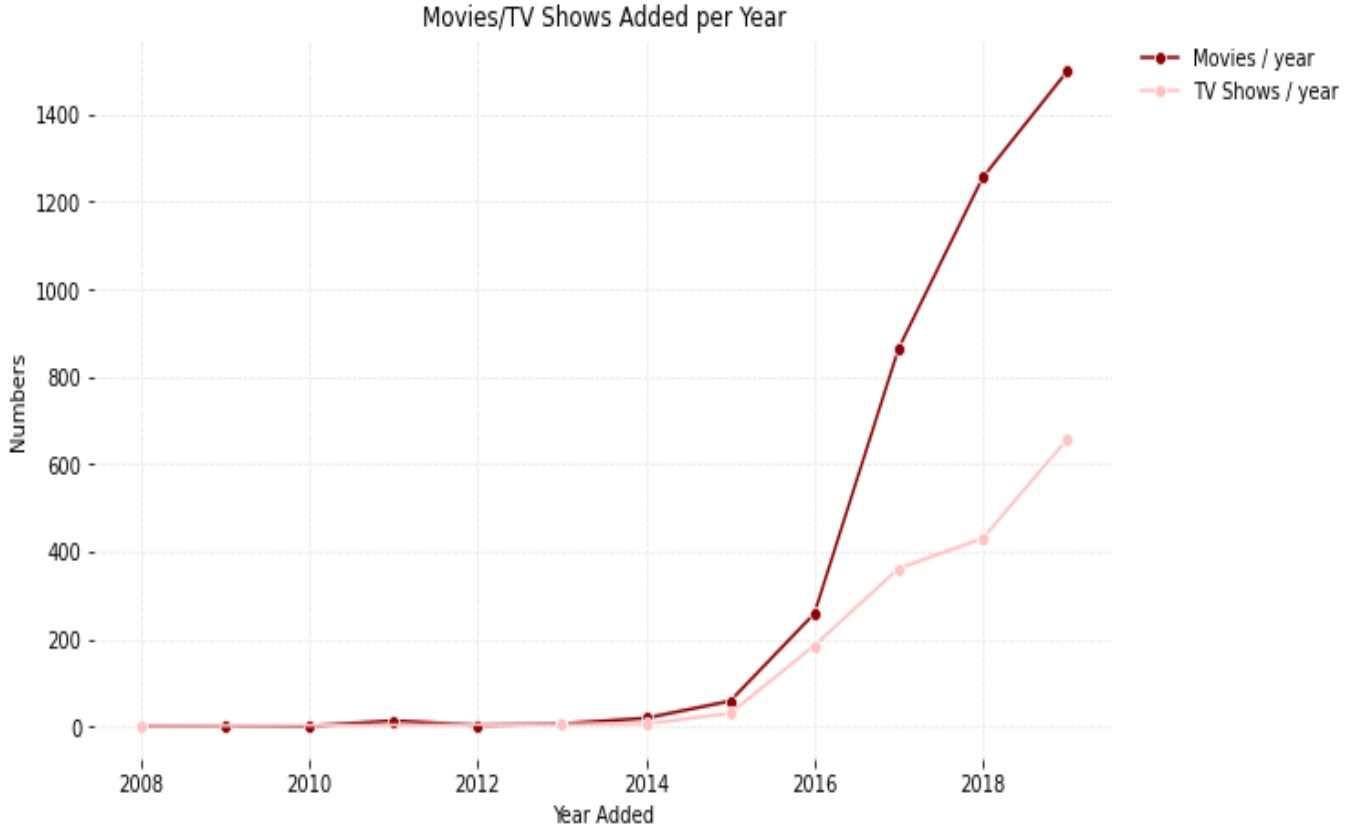
- The United States account for the majority of the content created on Netflix, numbering 2549 titles.
- India is the second largest with 923 titles.



Exploratory Data Analysis

Content added on Netflix per year

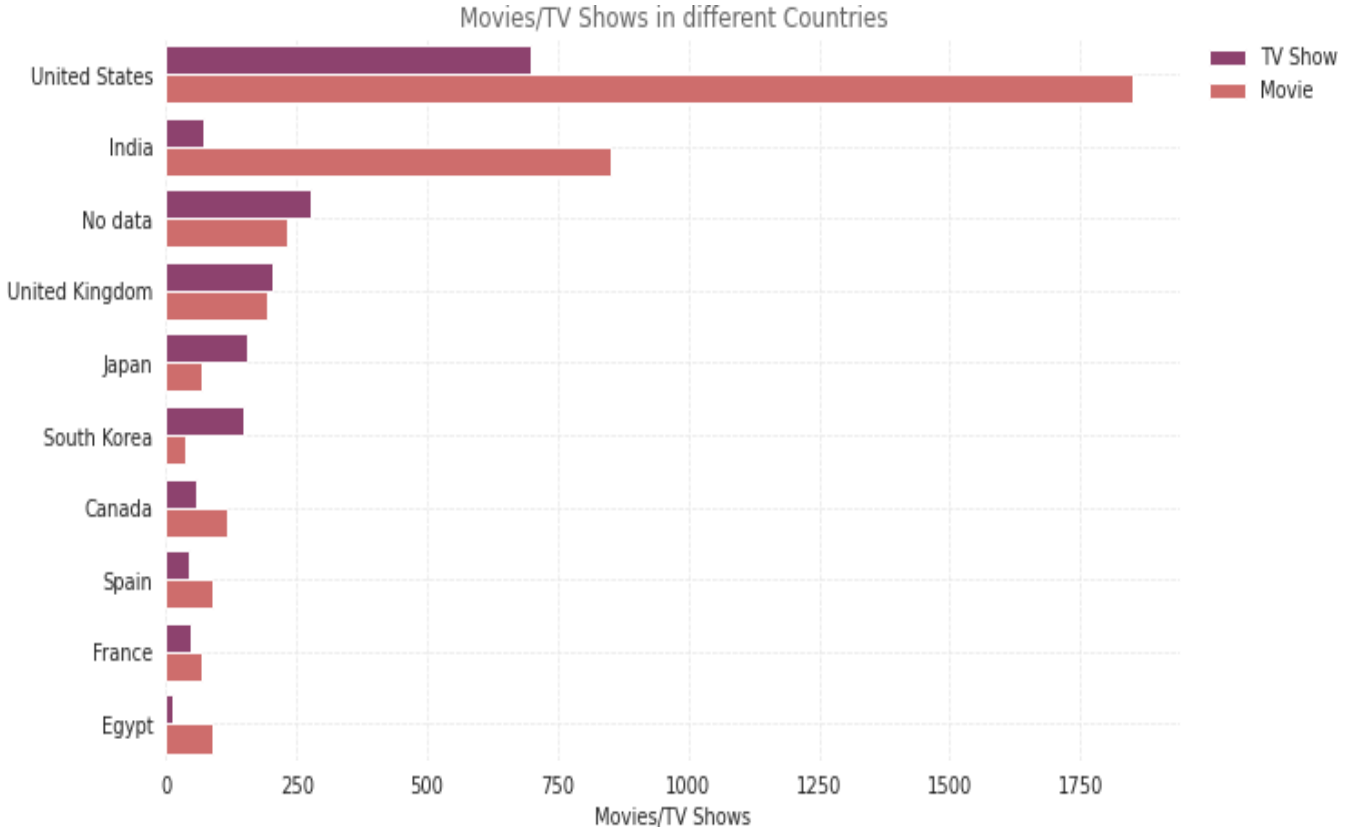
- As can be seen in this plot, both TV shows and Movies content numbers increased drastically after 2016.
- Also the number of movies added were much higher compared to TV shows number.



Exploratory Data Analysis

Distribution of Movies/TV Shows produced in various countries

- The United States accounts for the majority of the content created on Netflix.
- India is the second largest.
- In both countries, number of movies outnumbers the number of shows.



Data Pre-processing

We cannot go straight from raw text to fitting a machine learning model. We must clean text first, which means splitting it into words and handling punctuation.

For clustering we choose "description" and "Listed_in" variables. Before clustering we need to pre-process the data. So that we filtered data with following steps:

Text Example: "After an awful accident, a couple admitted to a grisly hospital are separated and must find each other to escape before death finds them."

1. Remove Punctuation

Text after removing punctuations: After an awful accident a couple admitted to a grisly hospital are separated and must find each other to escape — before death finds them.

2. Remove Stop-words

Text after removing stopwords:
awful accident couple admitted grisly hospital separated must find escape — death finds

3. Stemming

Text after removing stopwords:
aw accid coupl admit grisli hospit separ must find escap — death find

4. Length of processed text

Calculate the length of text we got from first three steps to do clustering

Applying Model

1. K-Means Clustering

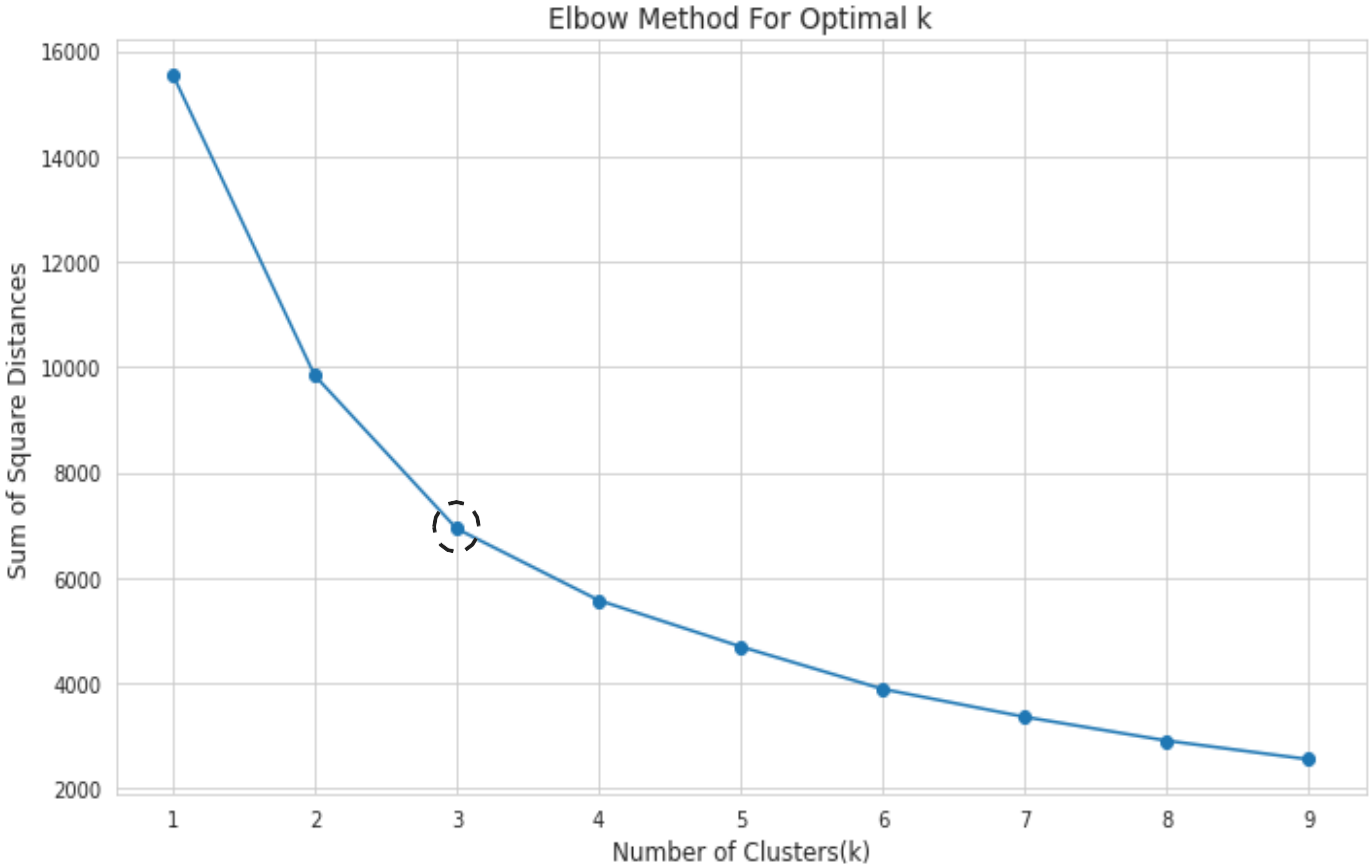
The **K-Means** algorithm searches for a predetermined number of clusters within an unlabelled multidimensional dataset.

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

To determine the optimal number of clusters, we have to select the value of k at the "elbow" i.e. the point after which the distortion/inertia start decreasing in a linear fashion.

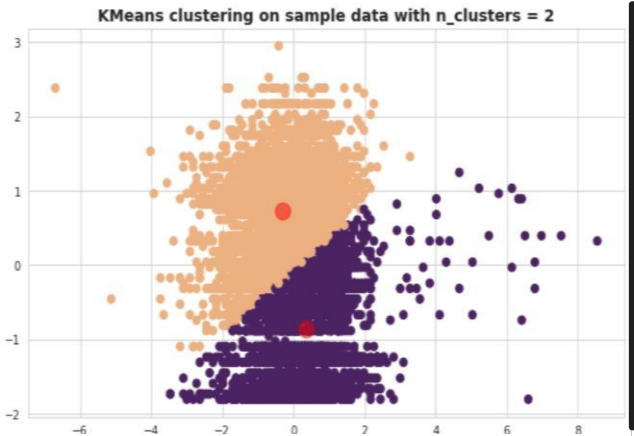
Thus from this chart we need to check, which would be the best number of clusters from 2,3,4,5, and 7.

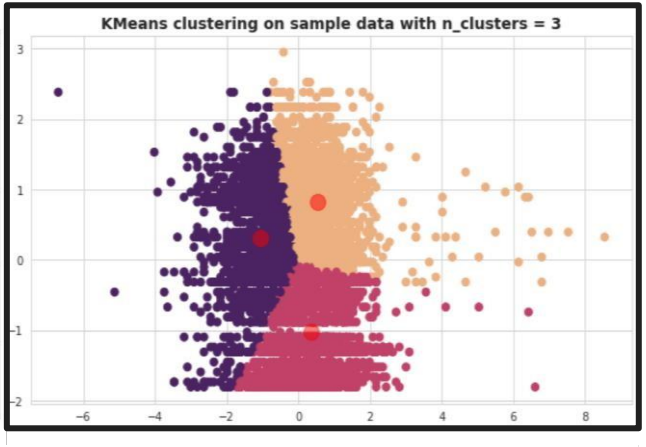
We found elbow formation at k=3.

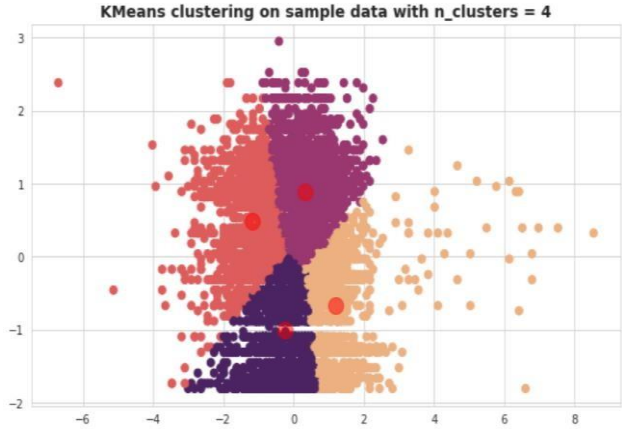


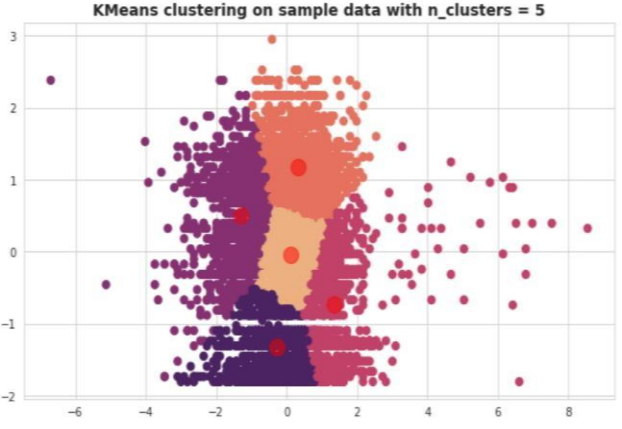
Applying Model

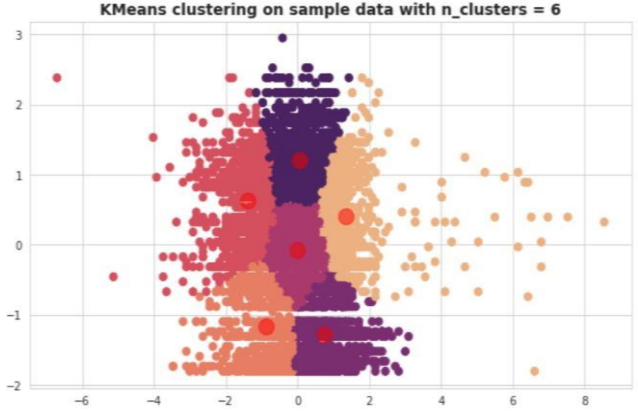
Different clusters to check optimum number of clusters

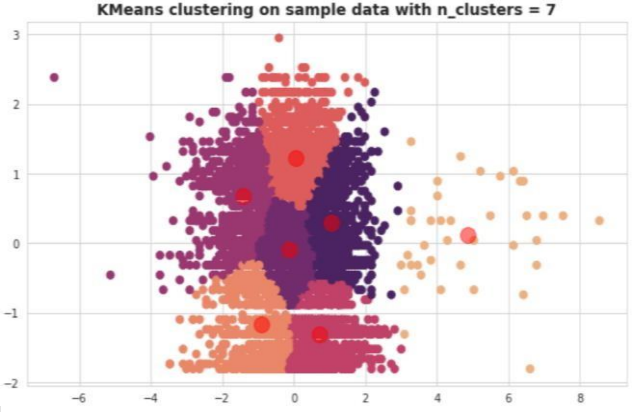












Silhouette Score for K-Means

Let's see the Silhouette scores of 2,3,4,5,6,7 and 8 number of clusters

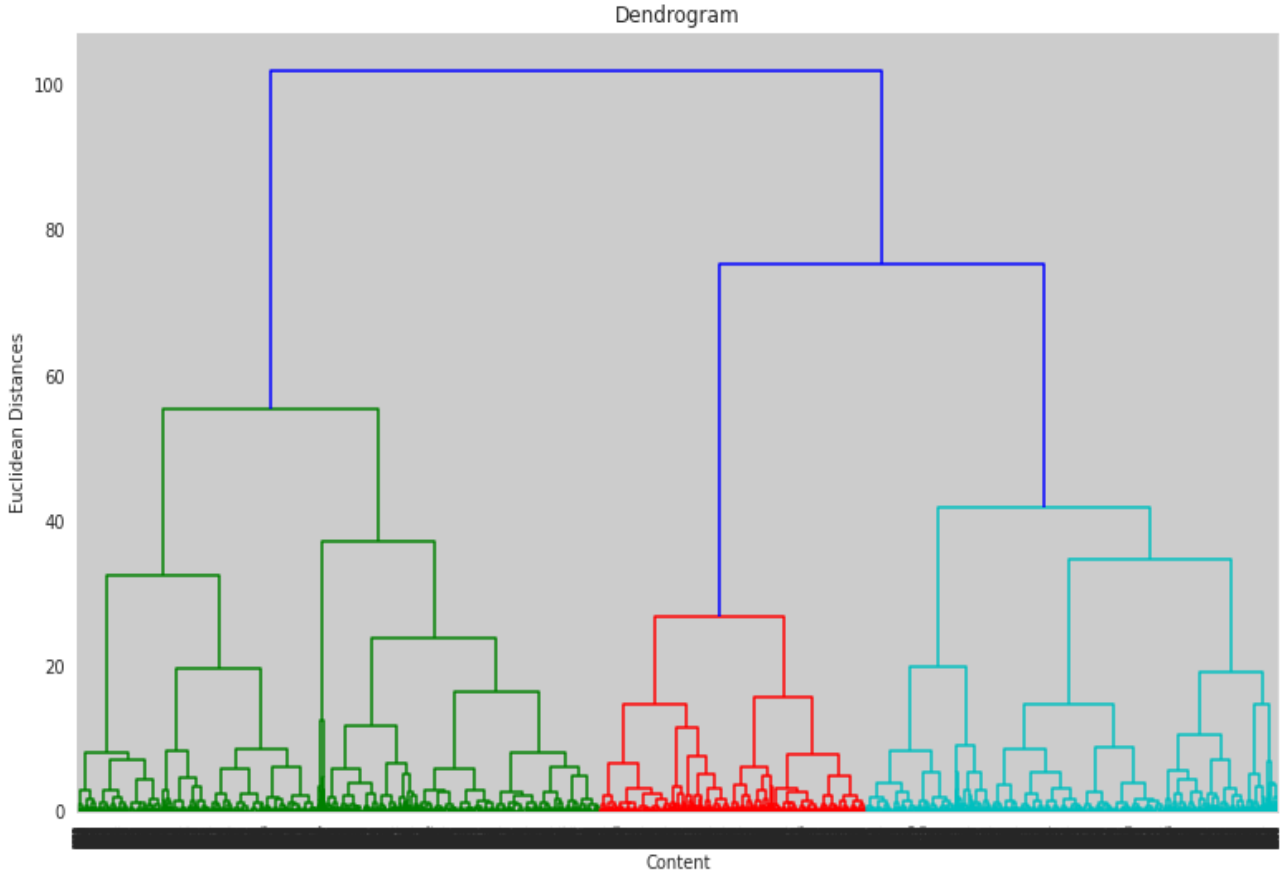
- Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other.
- The Silhouette score is calculated for each sample of different clusters.

- For n_clusters = **2**, silhouette score is **0.3551**415129079424
- For n_clusters = **3**, silhouette score is **0.3558**9698124108055
- For n_clusters = **4**, silhouette score is **0.3266**968907071311
- For n_clusters = **5**, silhouette score is **0.3355**8430881056384
- For n_clusters = **6**, silhouette score is **0.3557**380959007992
- For n_clusters = **7**, silhouette score is **0.3548**817152999796
- For n_clusters = **8**, silhouette score is **0.3522**803075712804

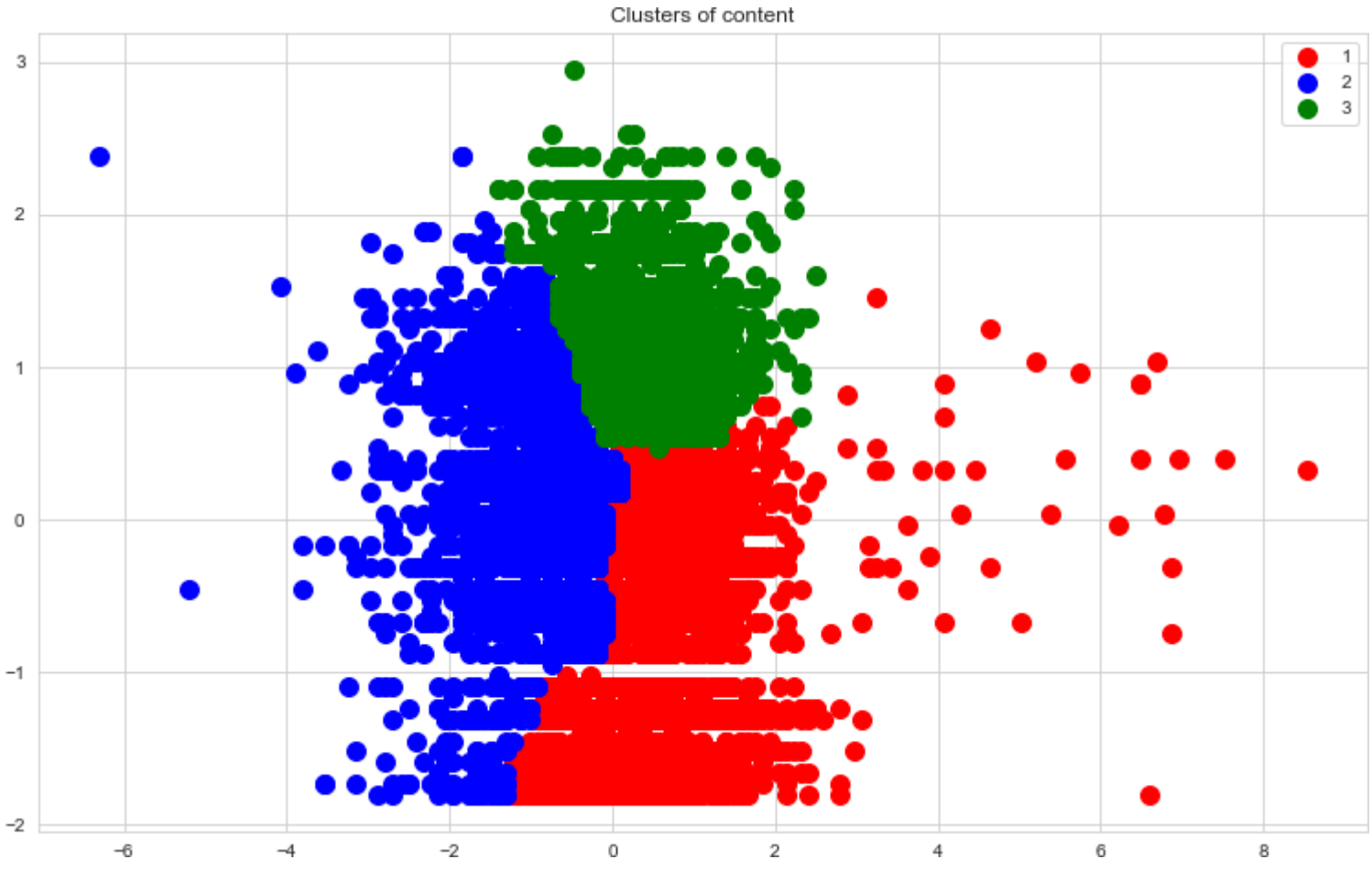
Applying Model

2. Hierarchical Agglomerative Clustering

- Hierarchical Agglomerative clustering starts with treating each observation as an individual cluster, and then iteratively merges clusters until all the data points are merged into a single cluster.
- Dendrograms are used to represent hierarchical clustering results.
- The number of appropriate clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold in this case – 3.



Hierarchical Agglomerative Clustering



Conclusion

- On the given dataset of OTT platform - "**Netflix Movies and Tv shows**" clustering was performed.
- First Data cleaning was done.Then Feature Engineering was done.Then some interesting insights were found by **Exploratory Data Analysis**.
- Majority of **content available on Netflix is Movies**.
- In recent years though many TV shows have been added, **number Movies outpower the number of TV shows**.
- United States and India top the countries that produce all of the available content on Netflx.
- TV-MA tops the graphs,indicating that mature content is more popular on Netflix.Then to perform clustering based on matching text features - Unsupervised Machine learning models were used.

Conclusion

- Then text pre-processing was done by removing unuseful characters like-stopwords,punctuation and stemming.
- Firstly **K-Means clustering** unsupervised Machine learning technique was applied.For this elbow method was used to find K value and Silhouette score of 0.35 was obtained.
- Next **Hierarchical clsutering** was applied for which dendogram was obtained.Silhouette score of 0.32 was obtained.
- So **K-Means clustering** performs better on the dataset.

Thank You