Unsupervised ML

NETFLIX MOVIES AND TV SHOWS CLUSTERING

TEAM DATA DEFENDERS

SHUBHAM SARTAPE

LOKESH TOKAS

SARASWAT MUKHERJEE

CHARAN

Content

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.

Data Summary

Fields	Description
show_id	Unique ID for every Movie / TV Show
type	Identifier - A Movie or TV Show
title	Title of the movie / show
director	Director of the show
cast	Actors involved
Country	Country of Production

Fields	Description
date_added	Date it was added on Netflix
release_year	Actual release year of the show
rating	TV rating of the show
duration	Total duration in minutes or number of seasons
listed_in	Genre
Description	The summary description

NULL VALUES

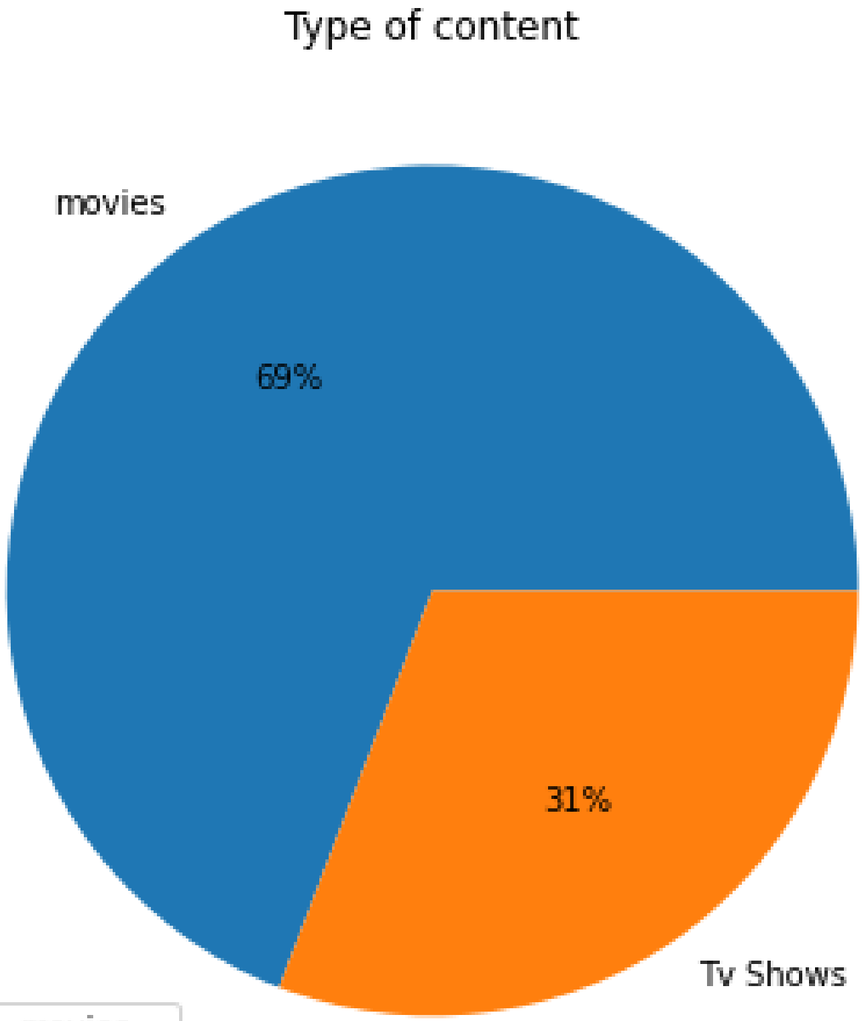
Dropping and replacing null values for different columns

- Rating feature have 0.09% of null values and Date_added feature have 0.13% of null values. Dropping null values of date_added and rating column as it has only single digit null values
- Director feature have more than 30.68% of null values. Replacing null values in director column by unknown
- Country feature have 6.51% of null values. Replacing null values in country column by mode value of it
- Cast feature have 9.22% of null values. Replacing null values in cast column by no cast

EDA

Type of Content

- There are 69% of movie content and 31% of Tv show content in netflix.
- We can see that movie content is double the number of TV shows
- Content of top 10 countries accounts for 69.54% for overall contents present

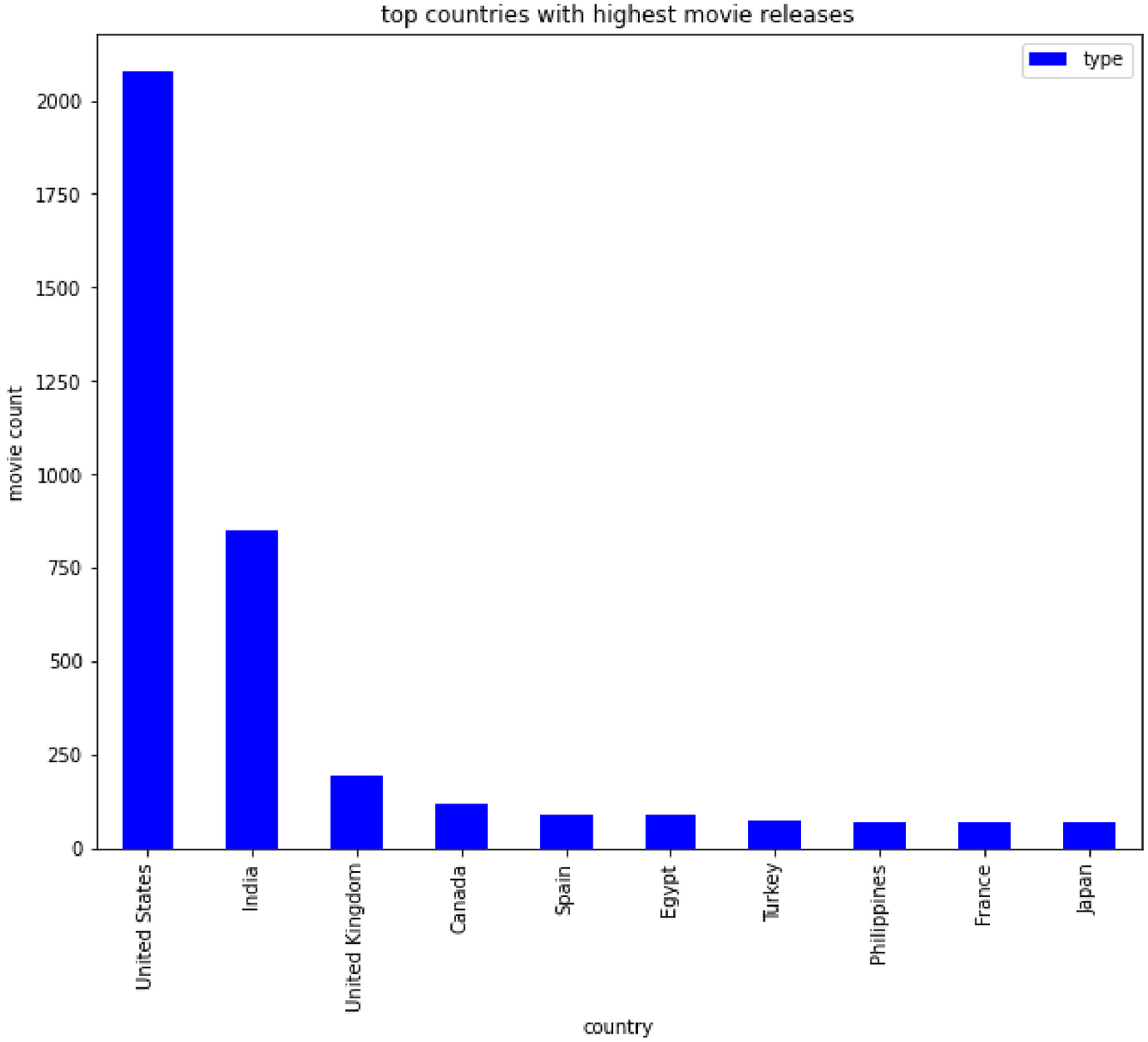


Movie	5372
TV Show	2398

EDA

Top Countries with highest movie releases

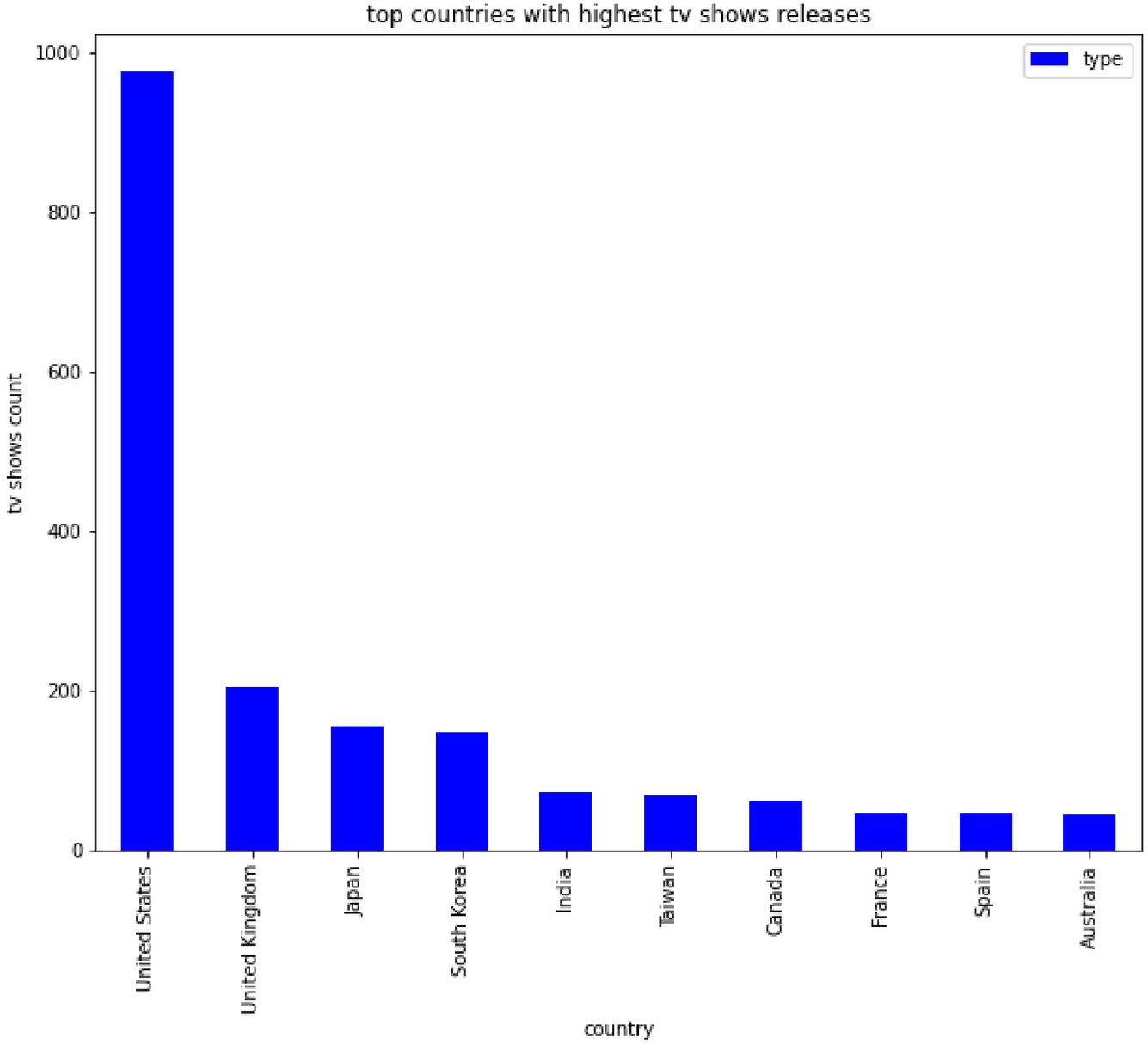
- United Nations has highest movie content released and then followed by India
- Content of top 10 countries accounts for 69.54% for overall contents present



EDA

Top Countries with highest TV Show releases

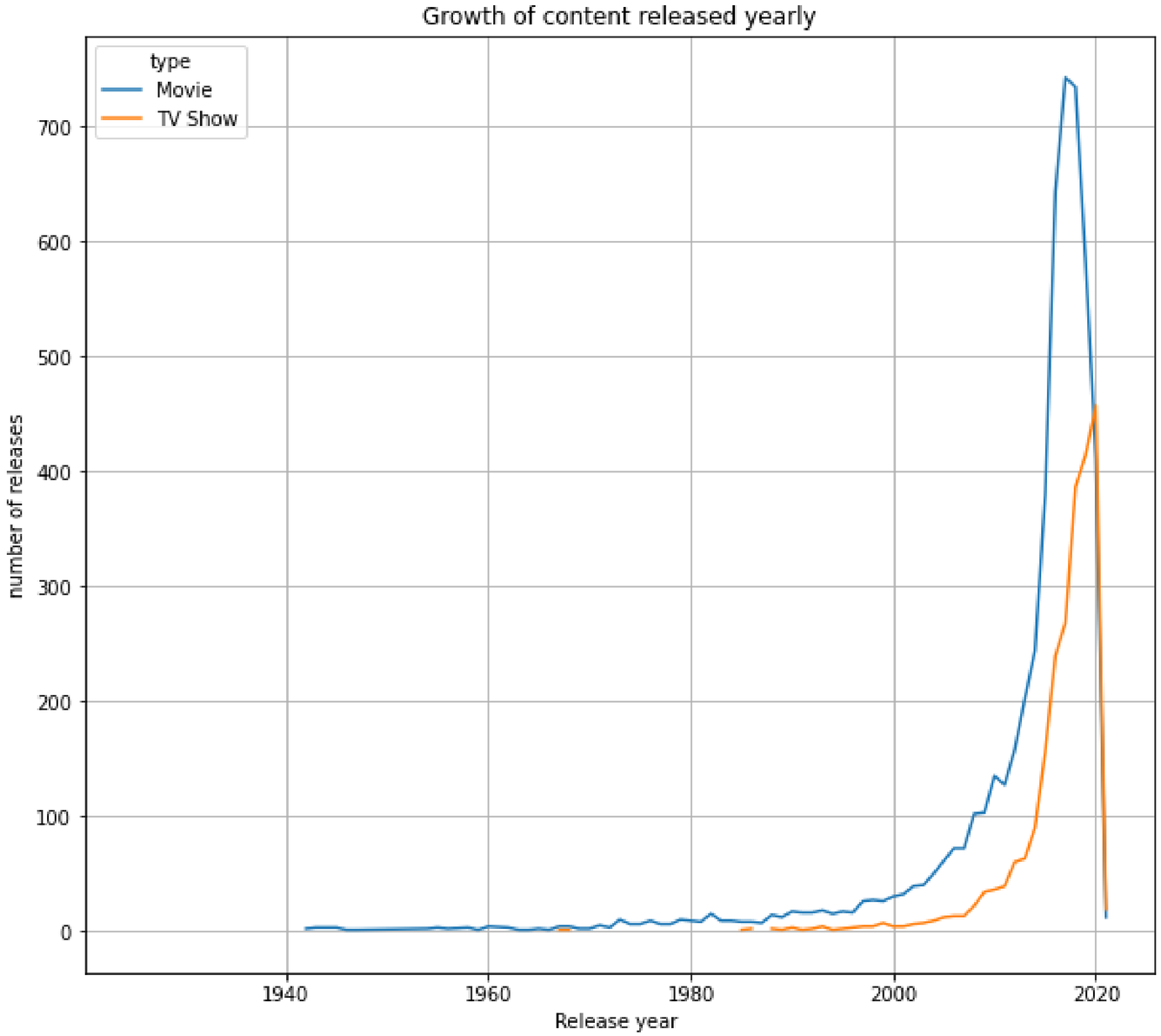
- United Nations has highest tv show content released and then followed by United Kingdom
- Content of top 10 countries accounts for 69.54% for overall contents present



EDA

Growth of content released yearly

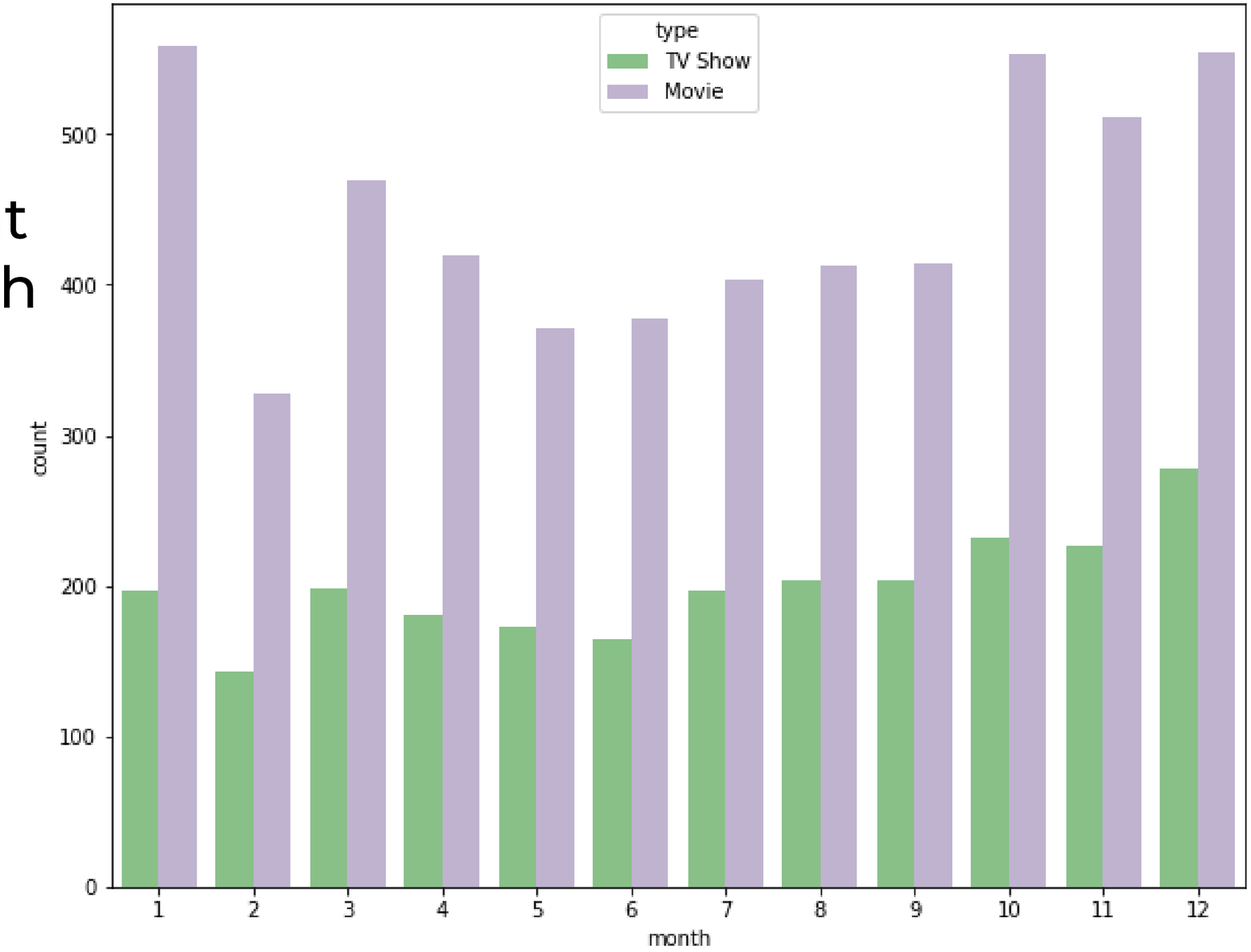
- Netflix started releasing movies after 1940 and after 1980 Tv show content was started streaming.
- we can see that there is a graduall increase in releases after 2000 where movie content releases are more then Tv show content



EDA

Monthly content released for both Movies and TV Shows

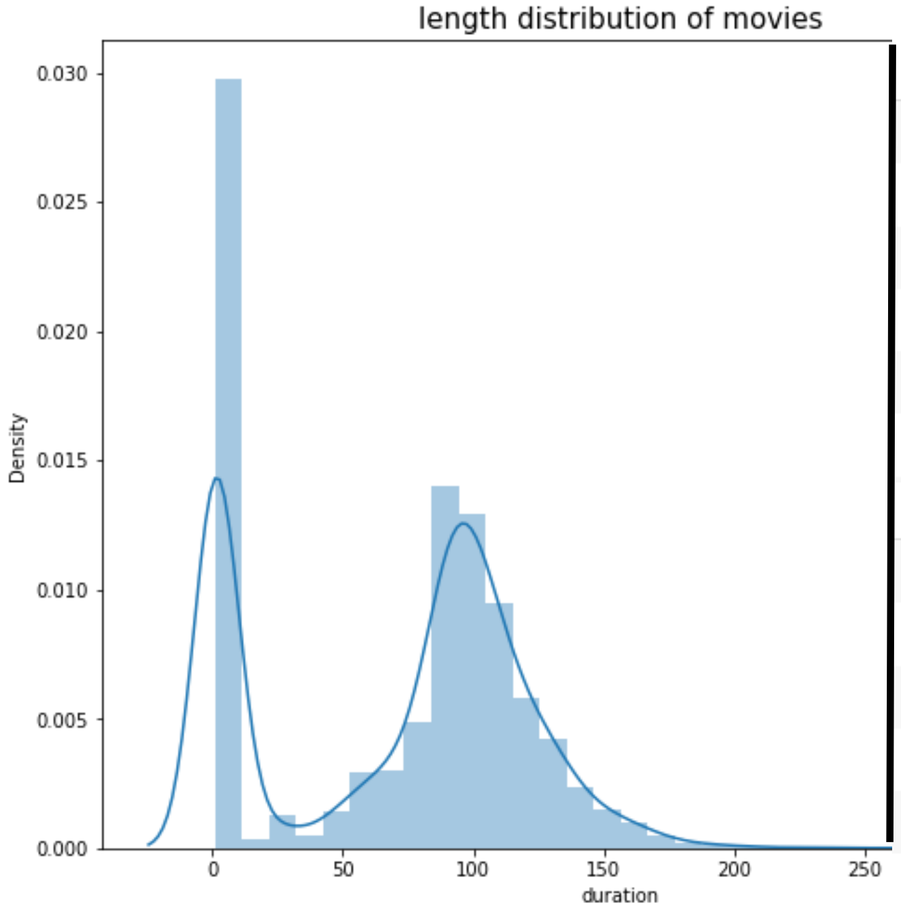
It is concluded that in the month of january, october and december maximum movies/tvshows has been released.



EDA

Top Movie/TV Show content type with respect to duration

- The mean duration of movies are 100 minutes
- Black Mirror: Bandersnatch is the longest movie which is 312 minutes long and the second is The School of Mischief .
- Grey's Anatomy is longest Tv show which has 16 seasons and followed by Supernatural which has 15 seasons

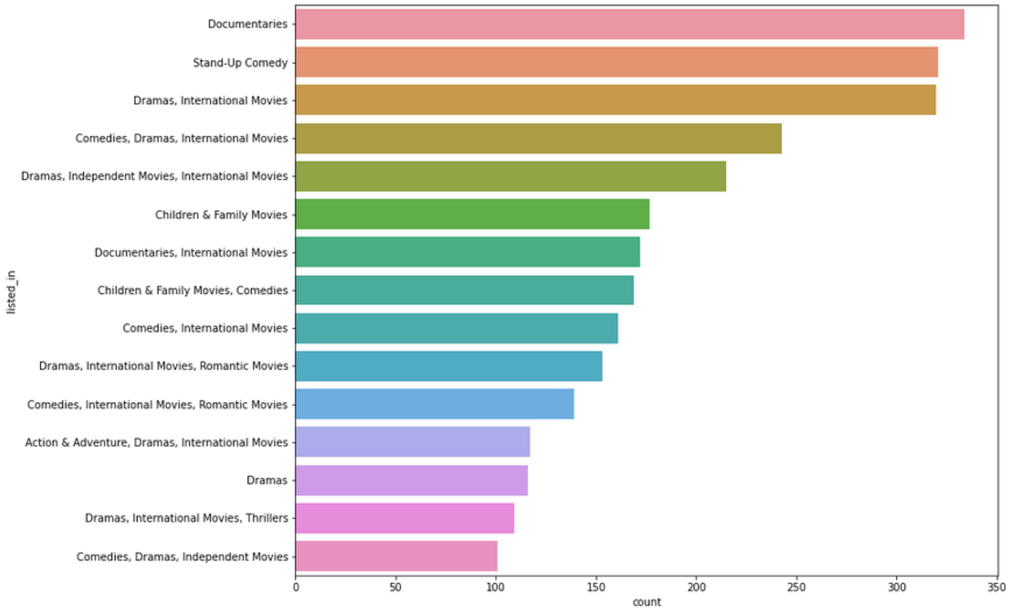


Movies	title	duration
Black Mirror: Bandersnatch		312
The School of Mischief		253
No Longer kids		237
Lock Your Girls In		233
Raya and Sakina		230
...		...

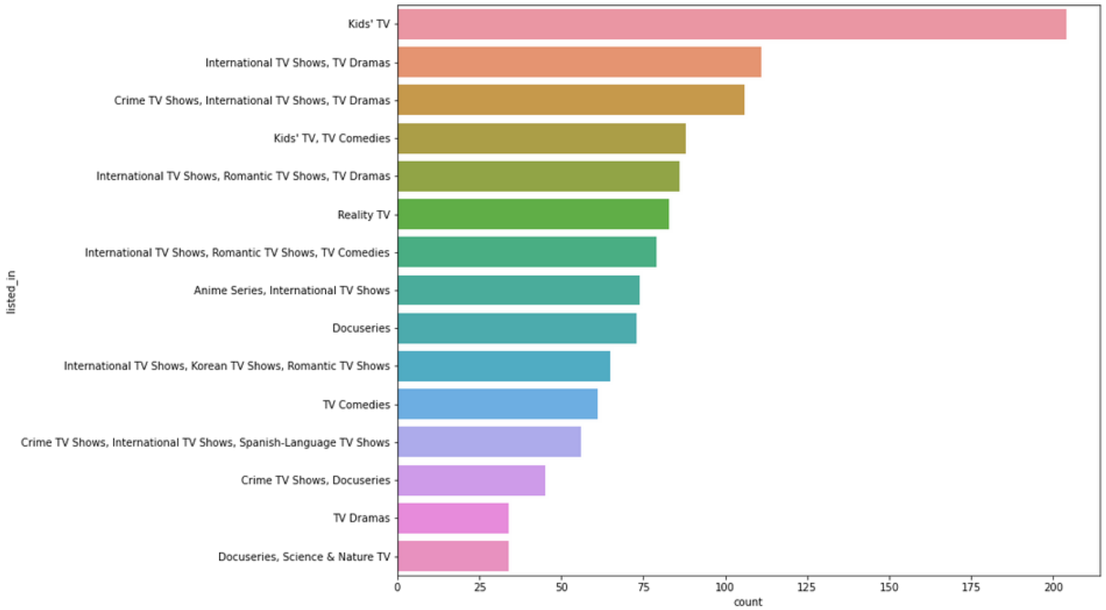
TV Shows	title	duration
Grey's Anatomy		16
Supernatural		15
NCIS		15
COMEDIANS of the world		13
Criminal Minds		12
...		...

EDA

Most occured Categories for Movies and TV Shows



Movies which are documentaries and stand up comedy has highest content in netflix

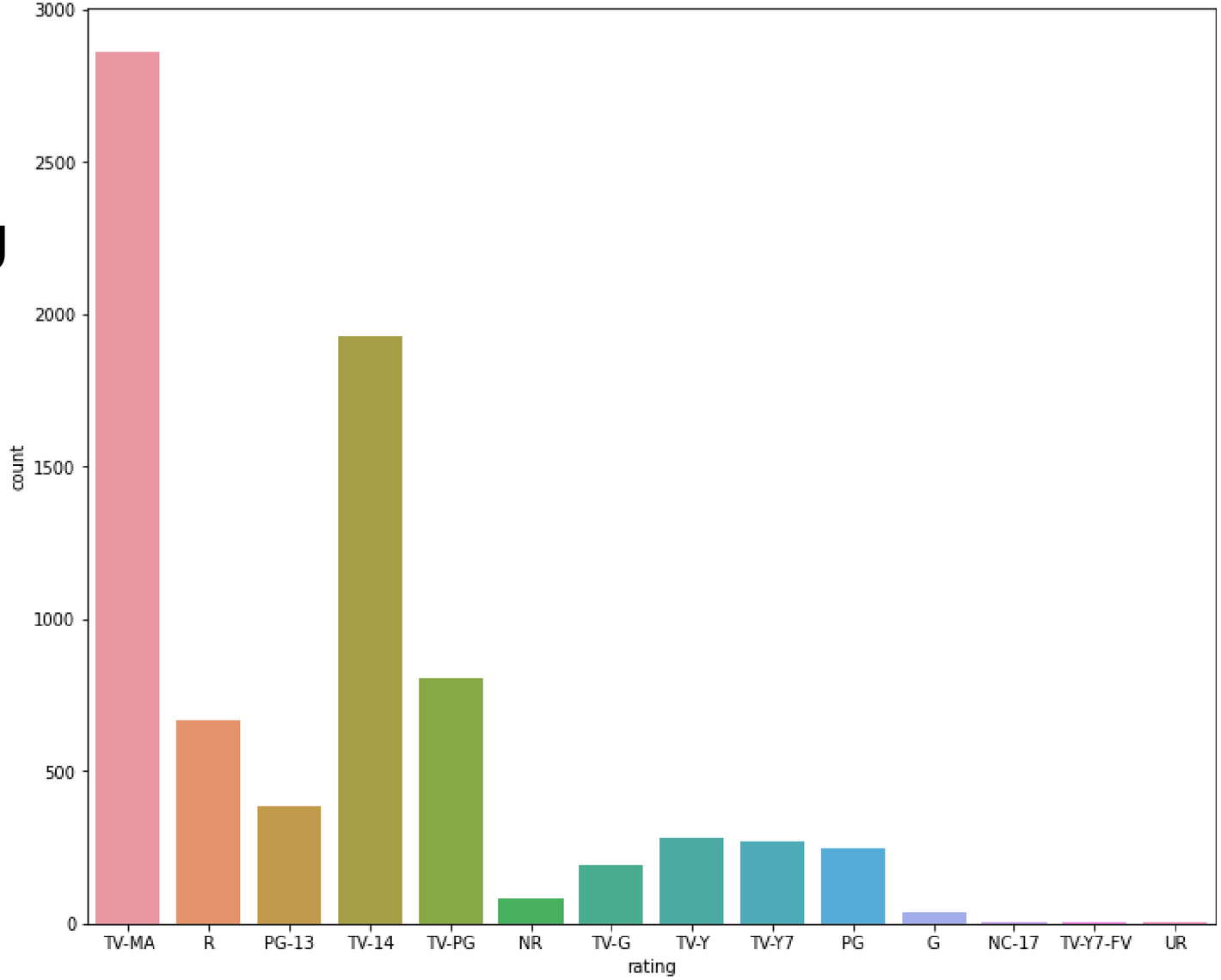


Tv shows for kids category has highest content in netflix

EDA

Most popular rating for TV Shows / Movies

Most of the contents got rating TV_MA(Mature Audience) and the second most rated contents is for Tv-14(excludes children)



Feature Engineering

- Removing punctuations for the text column

- 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.

- Stemming

Stemming is the process for reducing inflected words to their word stem (base form). A word stem is part of a word. It is sort of a normalization idea, but linguistic. For Example, stem word for 'waiting', 'waited', and 'waits' is 'wait'.

- Tokenizing the data

It is the practice of replacing a piece of sensitive or regulated data (like PII or a credit card number) with a non-sensitive counterpart, called a token, that has no inherent value. The token maps back to the sensitive data through an external data tokenization system. Data can be tokenized and de-tokenized as often as needed with approved access to the tokenization system.

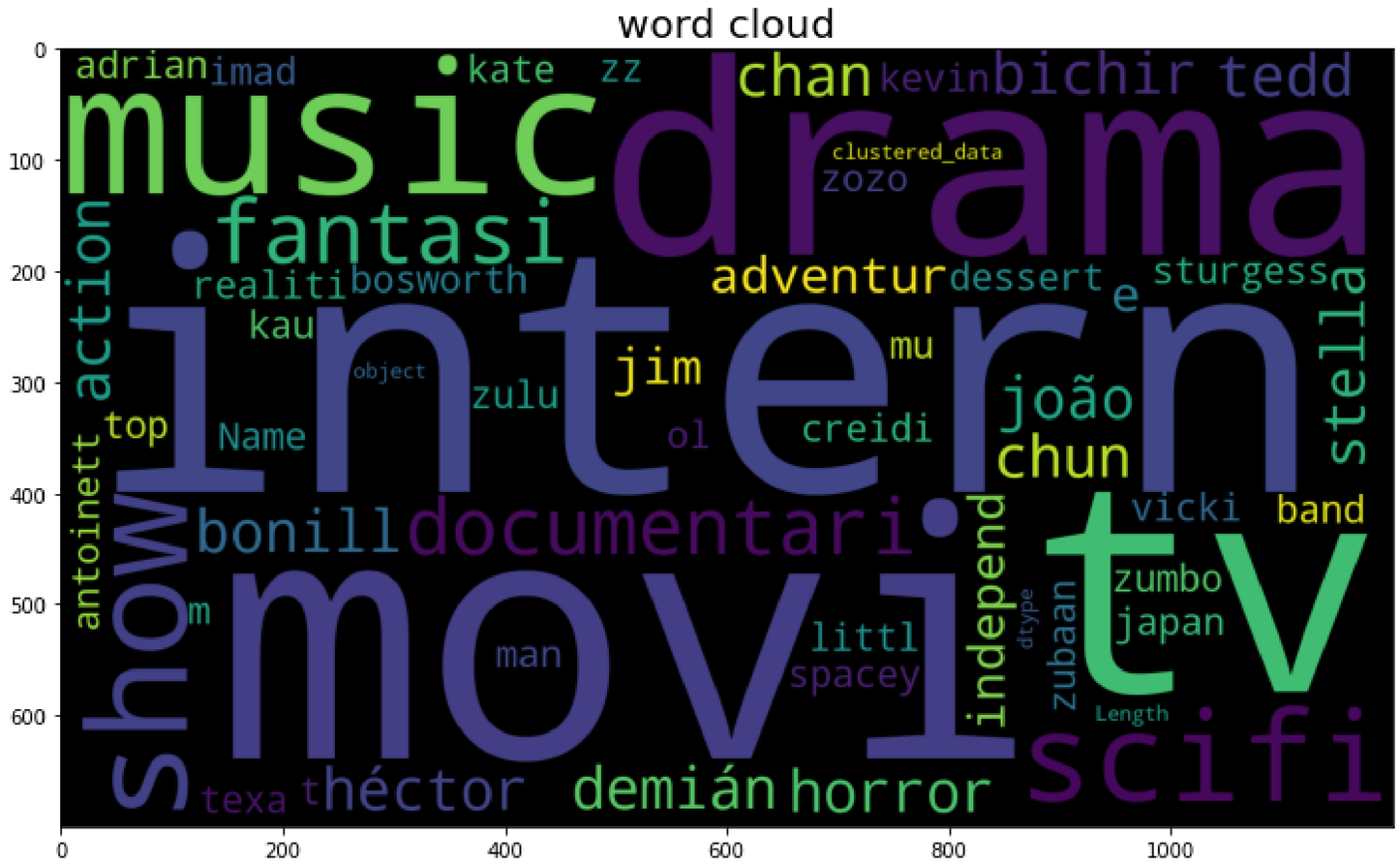
- Lemmatization

Lemmatization, unlike Stemming, reduces the inflected words properly ensuring that theroot 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.

Feature Engineering

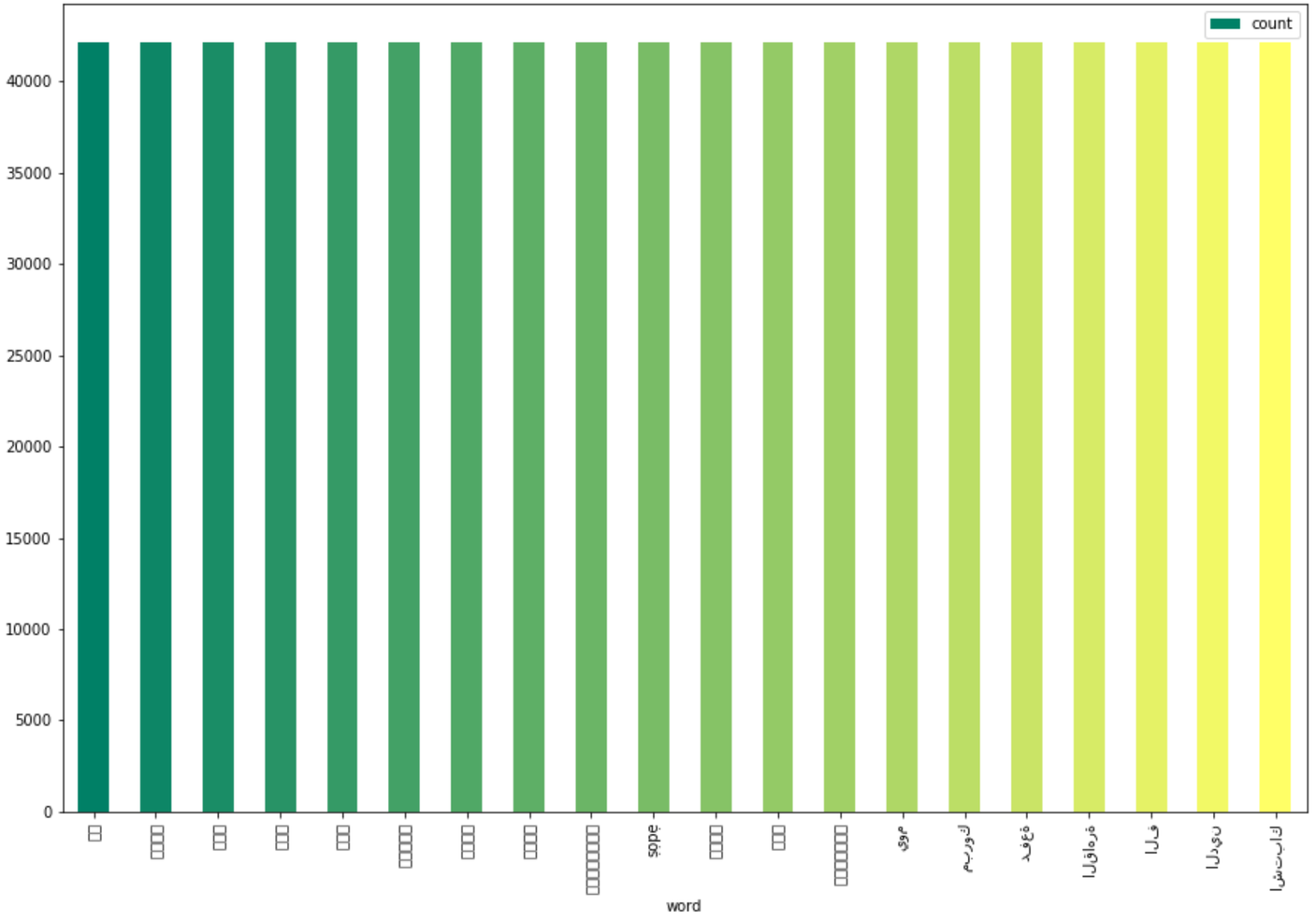
- Getting Frequency of each word by applying countvectorizer

- TFIDF
  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



Feature Engineering

Plotting top 20 words which are largely repeated



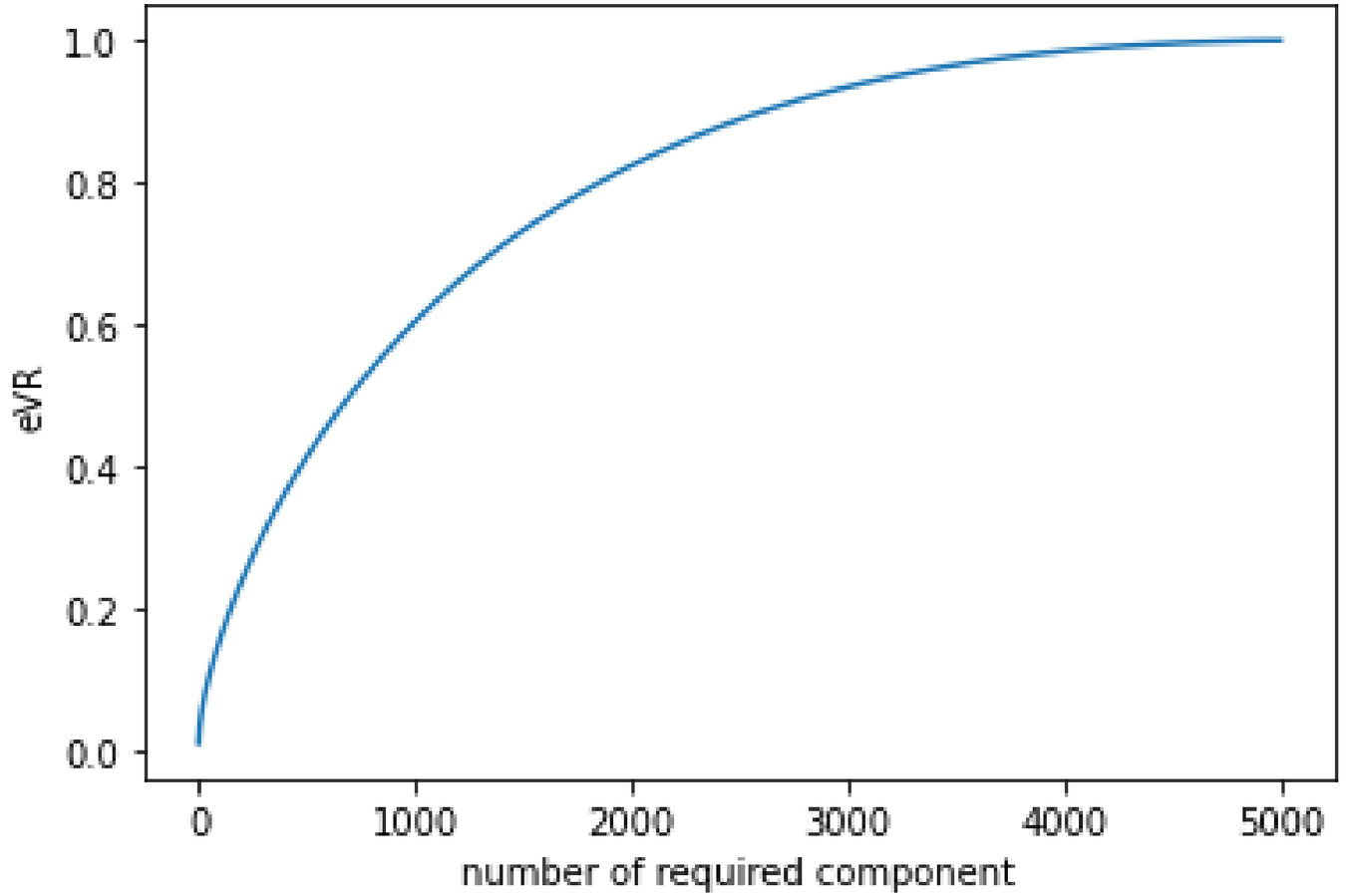
	word	count
25283	탄생	42146
25280	최강전사	42145
24533	잡는다	42144
25282	영웅의	42143
24532	반드시	42142
...	...	...
25135	aadarsh	4
34329	aacharekar	3
41290	aachal	2
39675	aabha	1
32550	aa	0

42147 rows × 2 columns

Feature Engineering

Feature component analysis for dimensionality reduction

In the selected 3000 components after pca dimensionality reduction we require 1500 components to summarize 90 percent of data



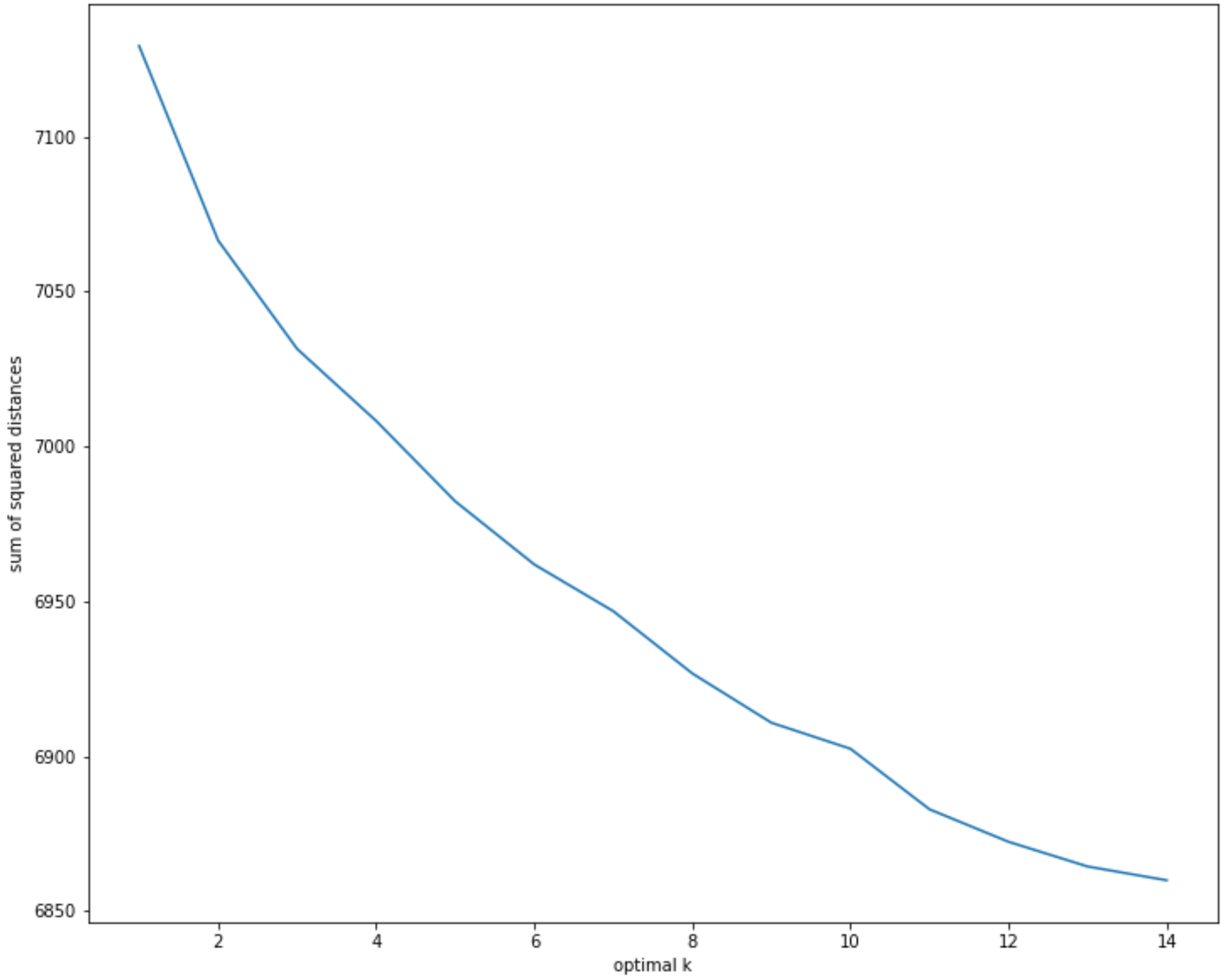
Clustering

K Means Clustering

Elbow method to find the number of clusters
Implementaing kmeans clustering
#considering range of 10 clusters

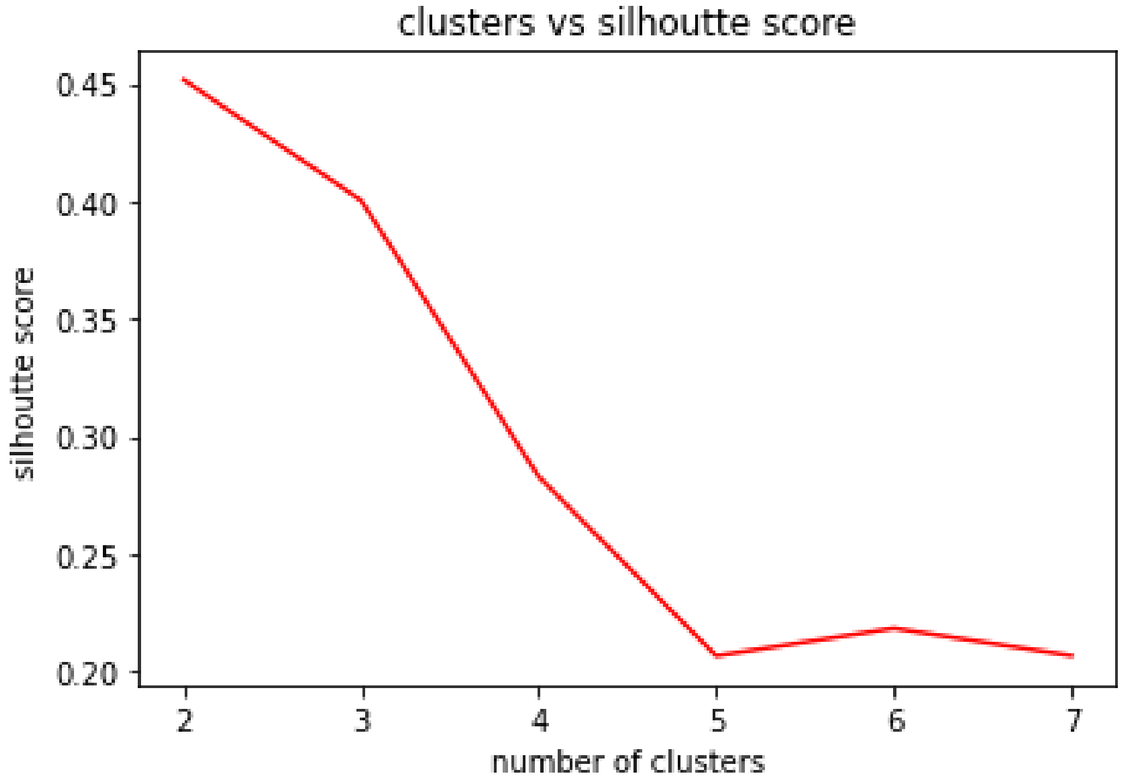
By plotting the optimal k by using Elbow method there are 7 clusters which are optimal

3	2125
5	1796
2	1677
0	744
6	713
4	379
1	336



Clustering

Silhoutte method to find the optimal clusters

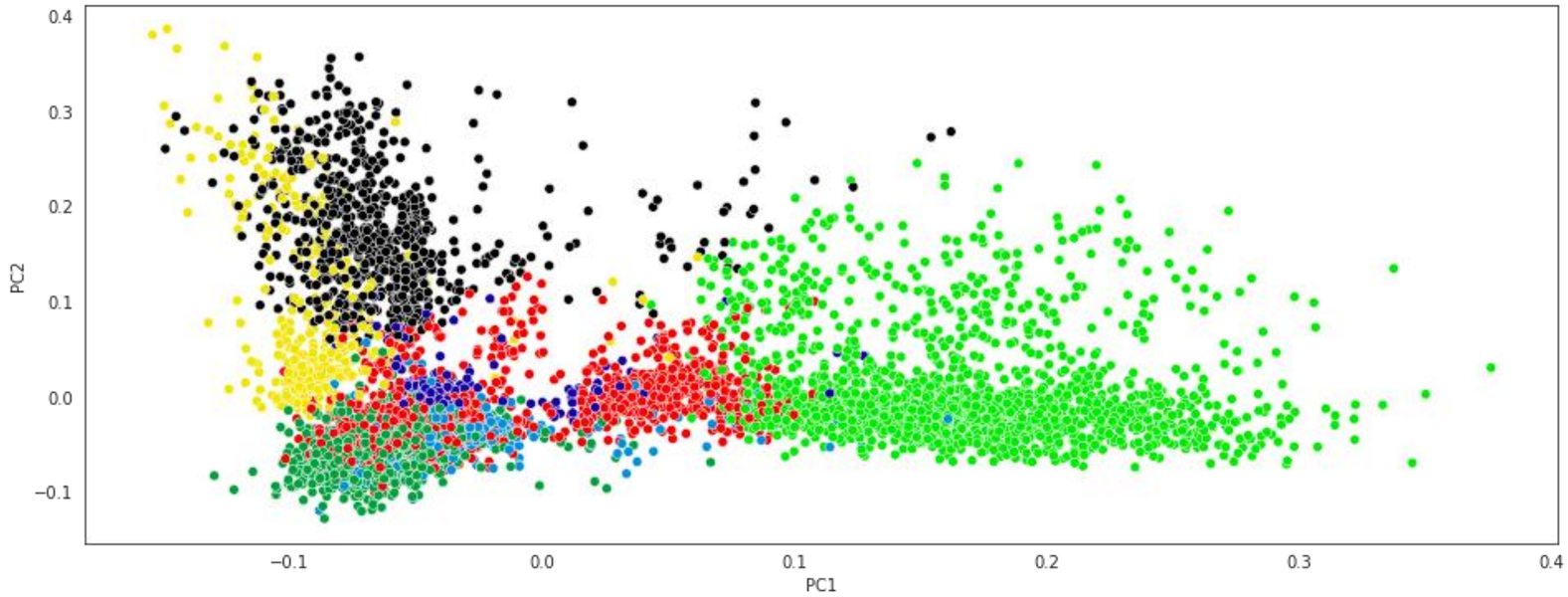



```
for n_clusters:  2
silhouette_score is: 0.4516037190029964
for n_clusters:  3
silhouette_score is: 0.4002917692372383
for n_clusters:  4
silhouette_score is: 0.2828787060690157
for n_clusters:  5
silhouette_score is: 0.20661253426556428
for n_clusters:  6
silhouette_score is: 0.218367887593708
for n_clusters:  7
silhouette_score is: 0.20670814905504448
```

Clustering

Silhoutte method to find the optimal clusters

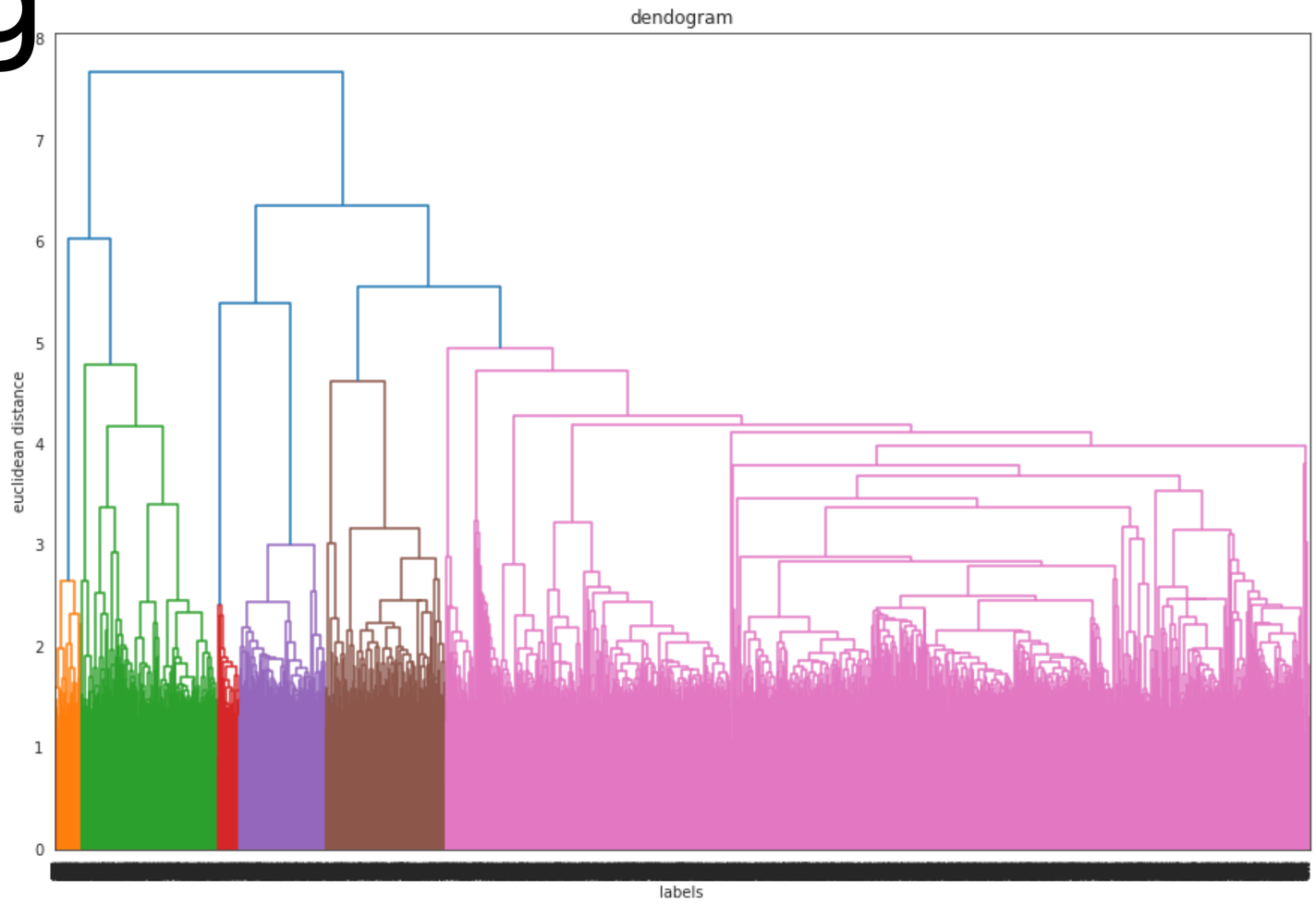
- The sihoutte score is highest for 3 clusters and the score is 0.510
- Cluster 4 has the highest text values
- As the number of clusters increases silhoutte decreases that is the quality of the cluster decreases



Clustering

Heirarchical Clustering

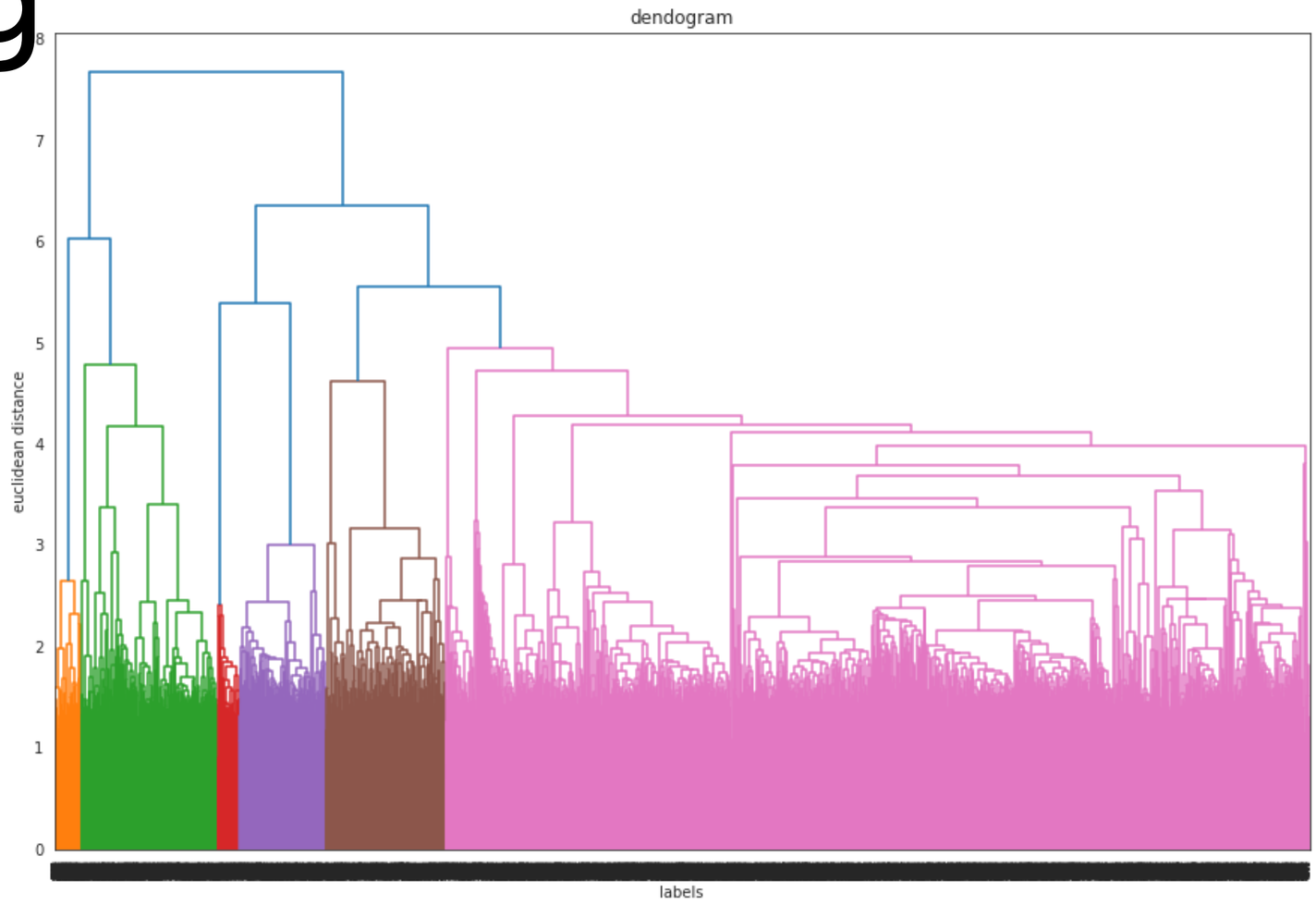
Choosing the number of clusters using dendegram
After finding the largest vertical line with maximum distance between the lines we can say that at distance 12 we can pass a horizontal line for which there are 6 clusters



Clustering

Heirarchical Clustering

- plotting the dendogram with known clusters and marking a horizontal line for the longest vertical distance
- Number of clusters are 6

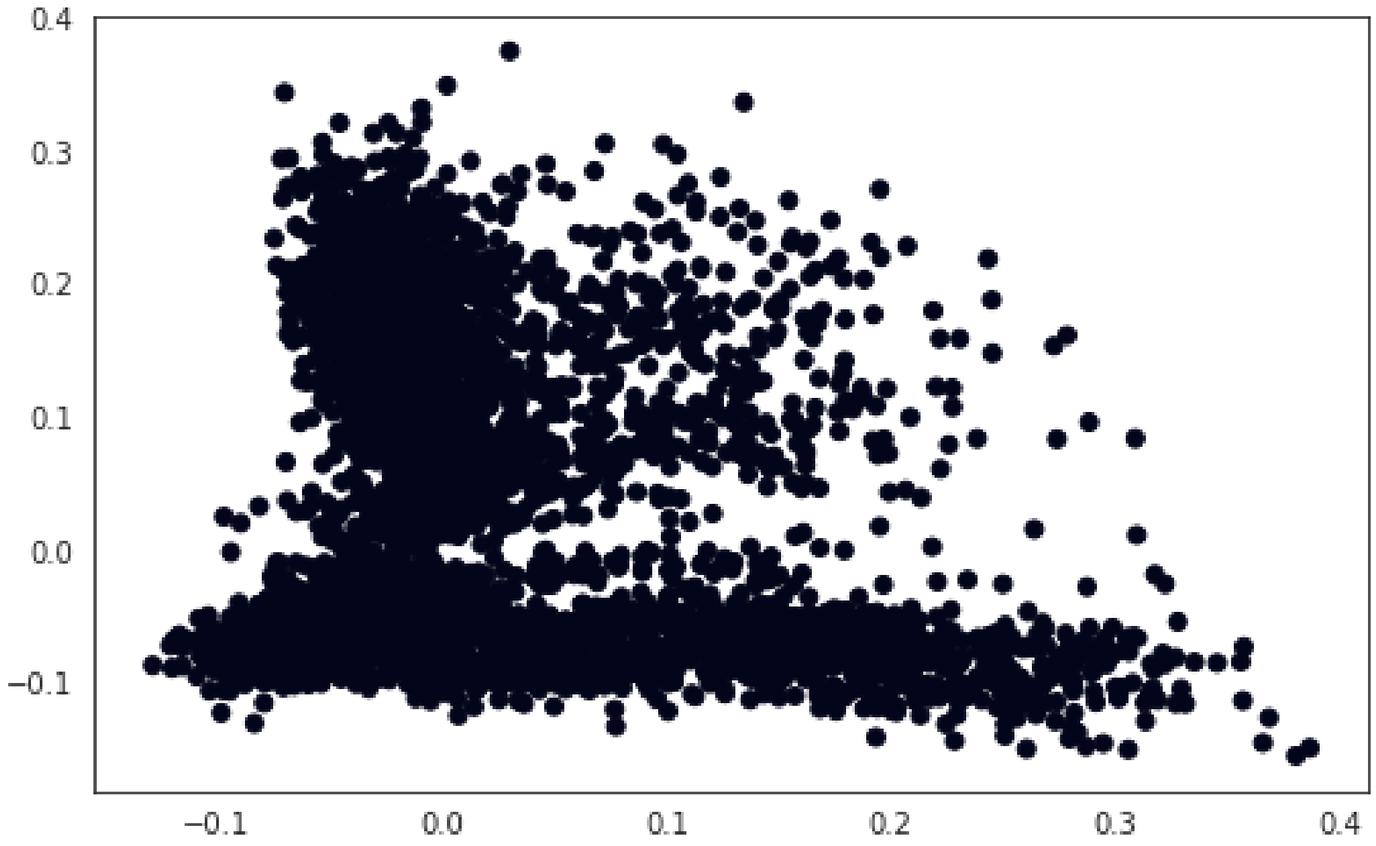


Clustering

DBSCAN

plotting the dbscan plot to check the purity of the cluster

- The labels of the data is too close to form the multiple clusters from DBSCAN clustering
- -1 value labels means that it is not able to accommodate the label to a cluster which are outliers or noise.



Conclusion

- There are 69% of movie content and 31% of Tv show content in Netflix
- Content of top 10 countries accounts for 69.54% for overall contents present
- United Nations has highest movie content released and then followed by India
- United Nations has highest tv show content released and then followed by United Kingdom
- Netflix started releasing movies after 1940 and after 1980 Tv show content was started streaming.
- There is a gradual increase in releases after 2000 where movie content releases are more then Tv show content
- In the month of January, October and December maximum movies/TV shows has been released.
- Black Mirror: Bandersnatch is the longest movie and Grey's Anatomy is longest Tv show
- Movies which are documentaries and stand-up comedy has highest content in Netflix and Tv shows for kids category has highest content in Netflix
- Most of the contents got rating TV_MA(Mature Audience)
- After vectorizing text data by using TFIDF and PCA dimensionality reduction we require 1500 components to summarize 90 percent of data
- In k-means clustering Elbow method is used to find the number of optimal clusters and there are 10 clusters after applying kmeans clustering
- To measure the quality of clusters and how well the clusters are separated, silhouette score is considered.
- In kmeans clustering the silhouette score is highest for 3 clusters and the score is 0.510
- In hierarchical clustering dendrogram plot is used to find the optimal clusters and the number of clusters came out to be 6
- The silhouette score for hierarchical cluster is highest for cluster 3 and the silhouette_score is 0.505
- By applying kmeans and hierarchical we got the best clusters and optimal cluster equal to 3

Thank You!