Capstone Project: 04

Netflix Movies and TV Shows Clustering

By

Pankaj R.Beldar

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.

Problem Statement

In this project, you are required to do

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

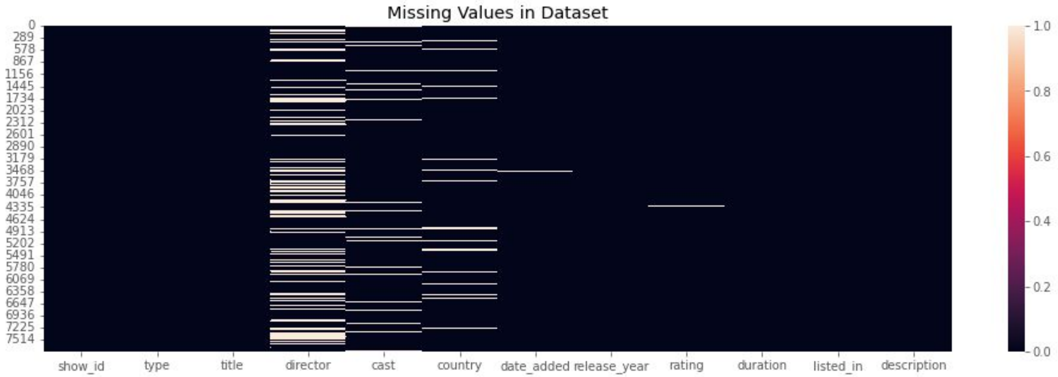
Attribute Information

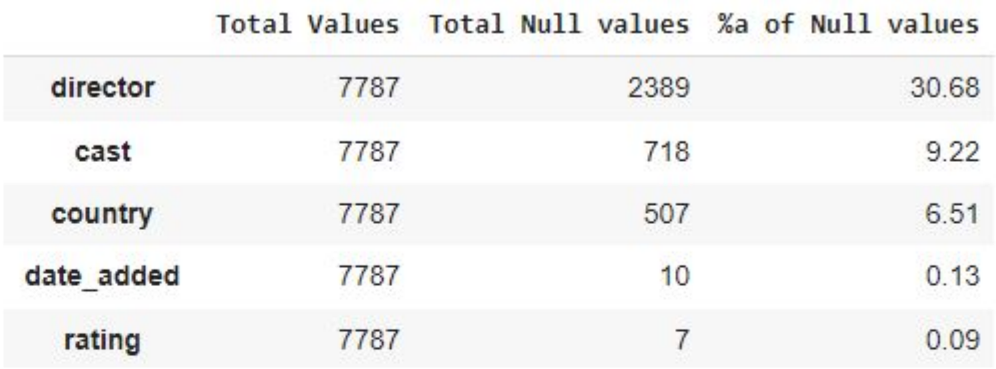
1. **show_id** : Unique ID for every Movie / Tv Show
2. **type** : Identifier - A Movie or TV Show
3. **title** : Title of the Movie / Tv Show
4. **director**: Director of the Movie
5. **cast** : Actors involved in the movie / show
6. **country** : Country where the movie / show was produced
7. **date_added** : 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

Milestones

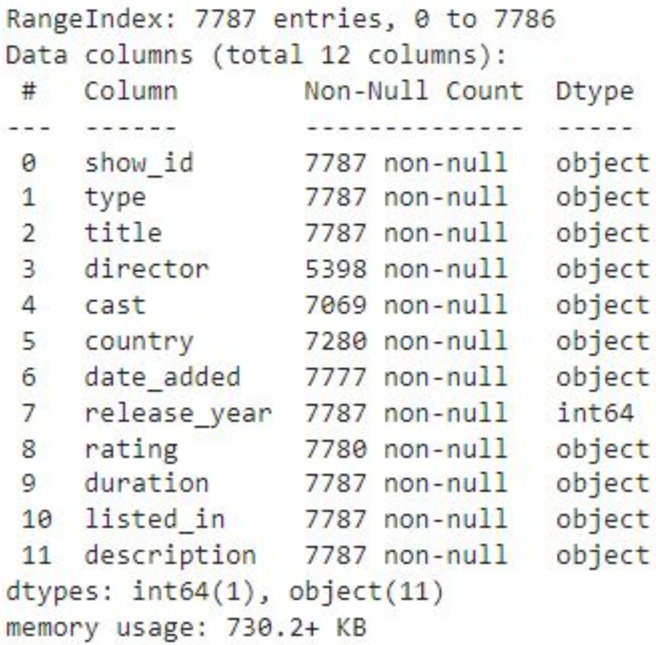
1. Importing Libraries
2. Import Data
3. Data Overview
4. Data Cleaning
5. Data Visualization with EDA
6. Text Processing
7. Model Selection and Hyper Parameter Tuning
   PART A: Modelling with Word2vec
   PART B: Modelling with CountVectorizer/tfidfVectorizer
8. Recommendation System
9. Conclusion

Data Overview





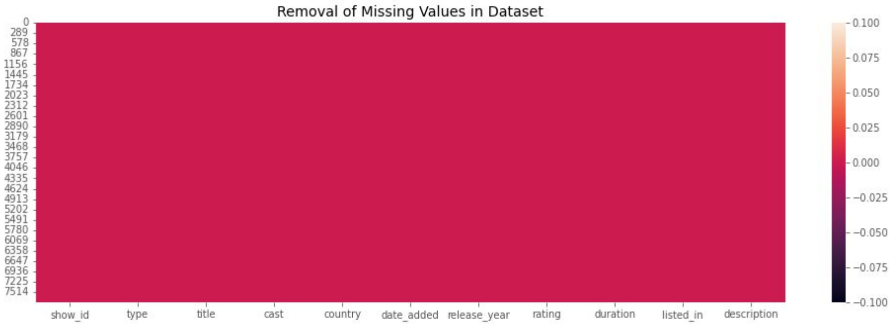
	Total Values	Total Null values	%a of Null values
director	7787	2389	30.68
cast	7787	718	9.22
country	7787	507	6.51
date_added	7787	10	0.13
rating	7787	7	0.09

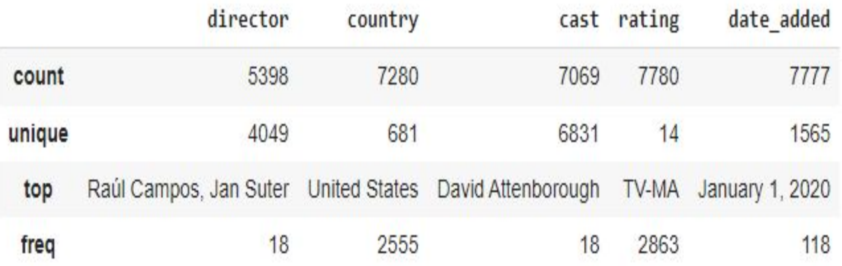


```
RangeIndex: 7787 entries, 0 to 7786
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       7787 non-null   object
 1   type          7787 non-null   object
 2   title         7787 non-null   object
 3   director      5398 non-null   object
 4   cast          7069 non-null   object
 5   country       7280 non-null   object
 6   date_added    7777 non-null   object
 7   release_year  7787 non-null   int64 
 8   rating        7780 non-null   object
 9   duration      7787 non-null   object
 10  listed_in     7787 non-null   object
 11  description   7787 non-null   object
dtypes: int64(1), object(11)
memory usage: 730.2+ KB
```

Handling Missing Values

1. The attribute 'director','cast','country','date_added','rating' consists of missing values.
2. To tackle with missing values , we will replace 'country' and 'rating' missing values by most frequent entity that is 'United States' and 'TV-MA' respectively.
3. missing values in 'cast' by 'unknown'.
4. There are around 30.68 % values are missing in 'director', hence we decide to drop it.

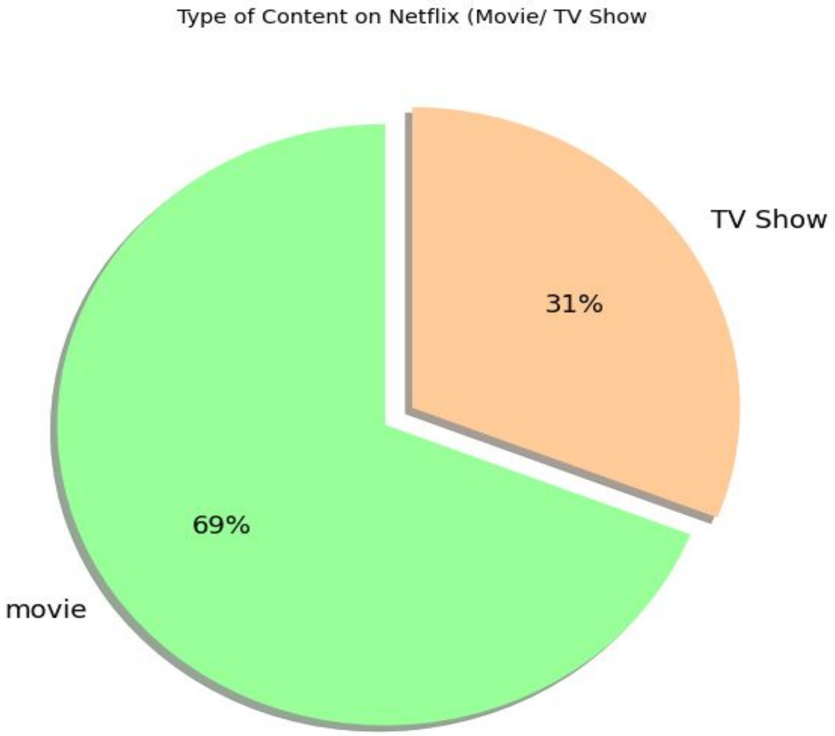


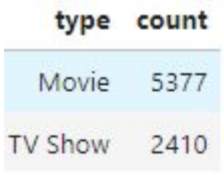


	director	country	cast	rating	date_added
count	5398	7280	7069	7780	7777
unique	4049	681	6831	14	1565
top	Raúl Campos, Jan Suter	United States	David Attenborough	TV-MA	January 1, 2020
freq	18	2555	18	2863	118

Exploratory Data Analysis

Type of content on Netflix

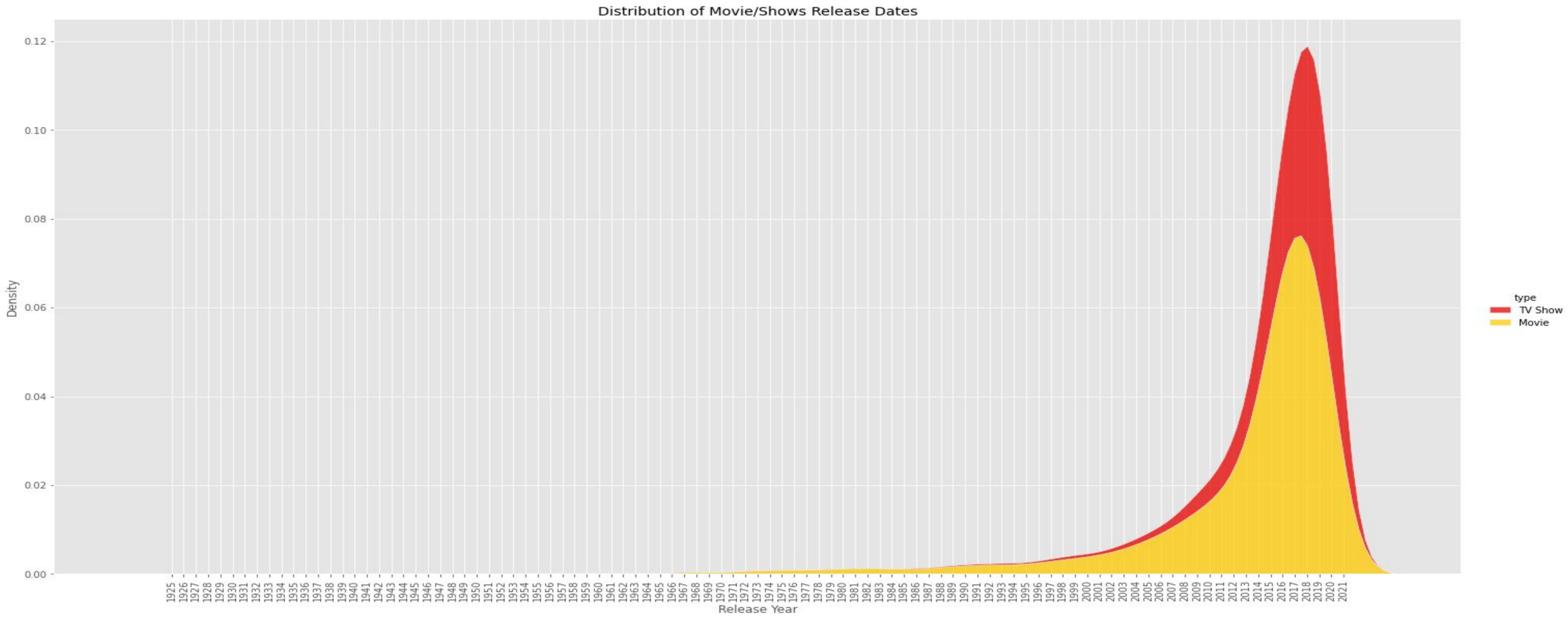




type	count
Movie	5377
TV Show	2410

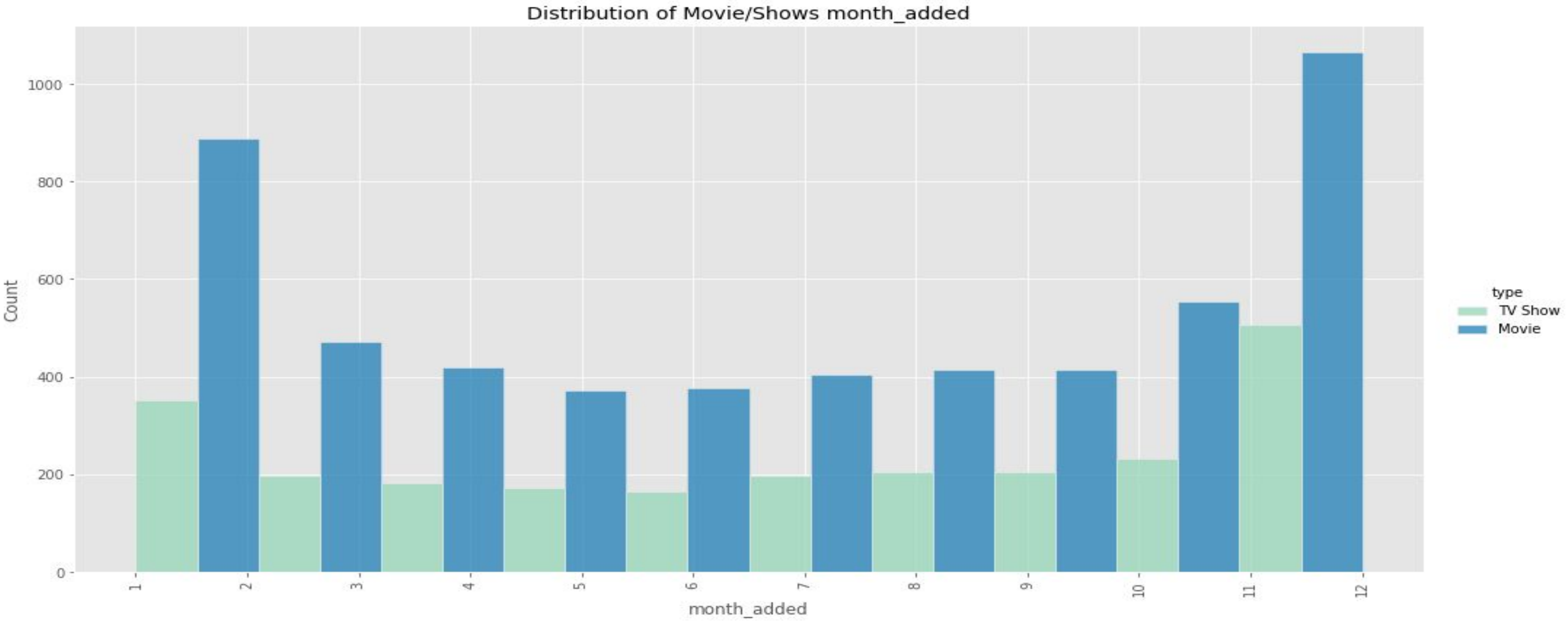
Exploratory Data Analysis

Distribution of Movie/Shows Release year



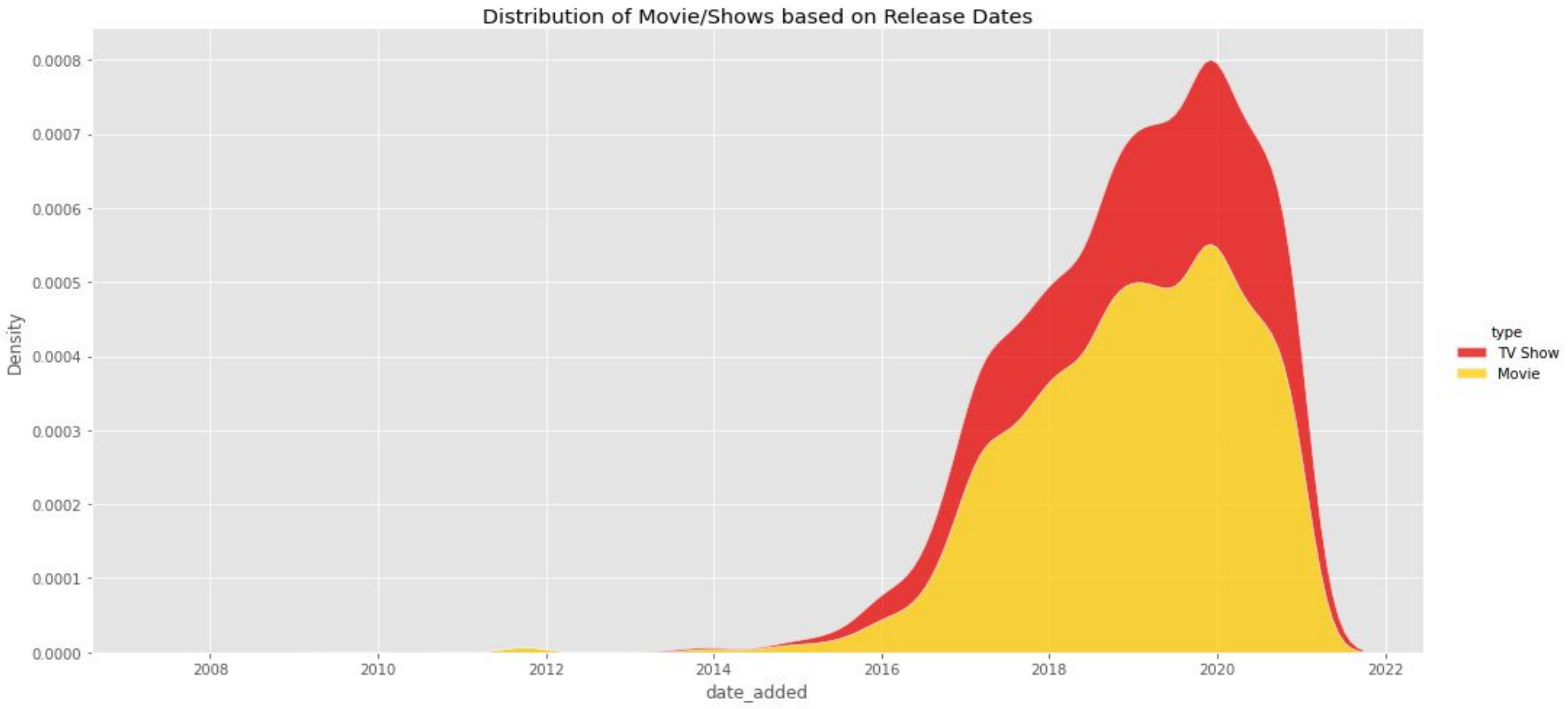
Exploratory Data Analysis

Distribution of Movie/Shows based on month added



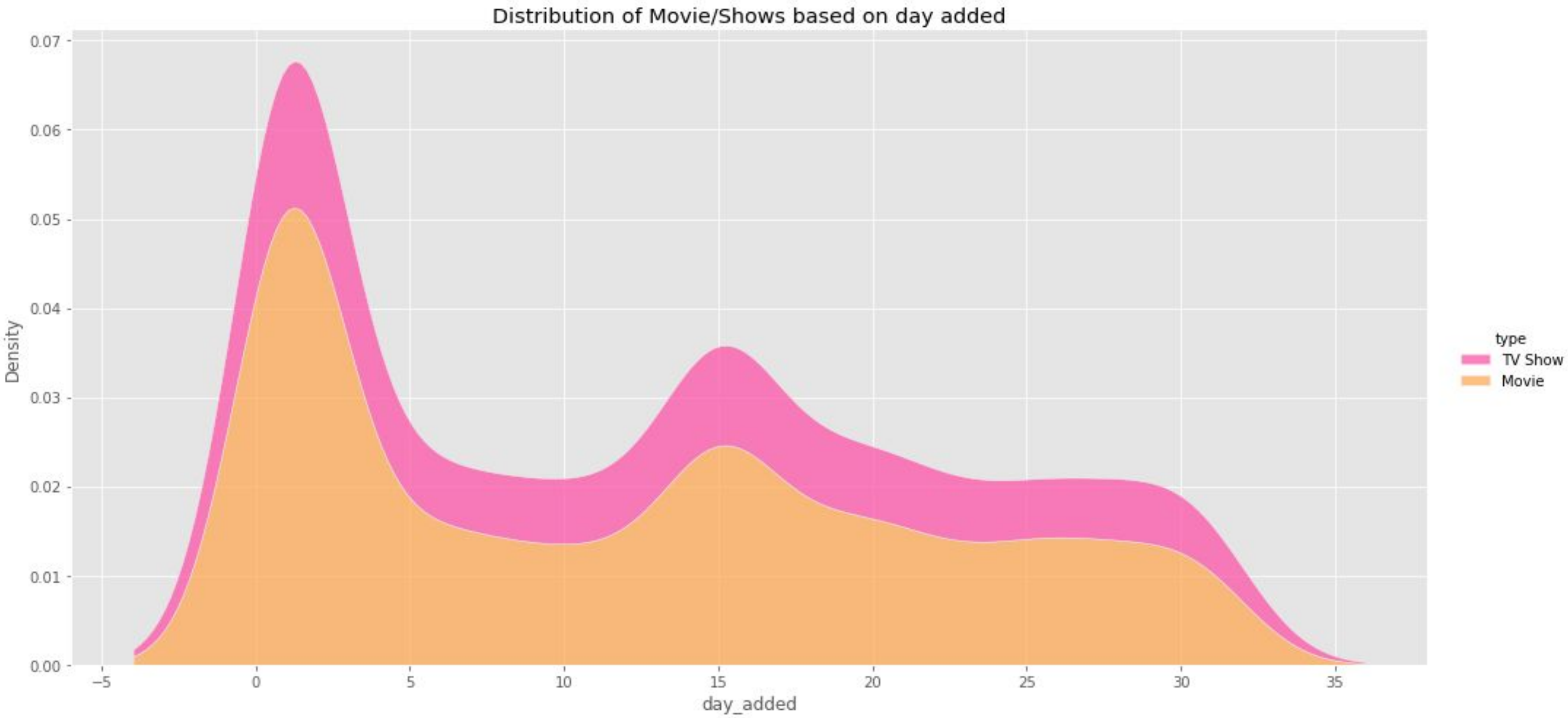
Exploratory Data Analysis

Distribution of Movie/Shows based on date added



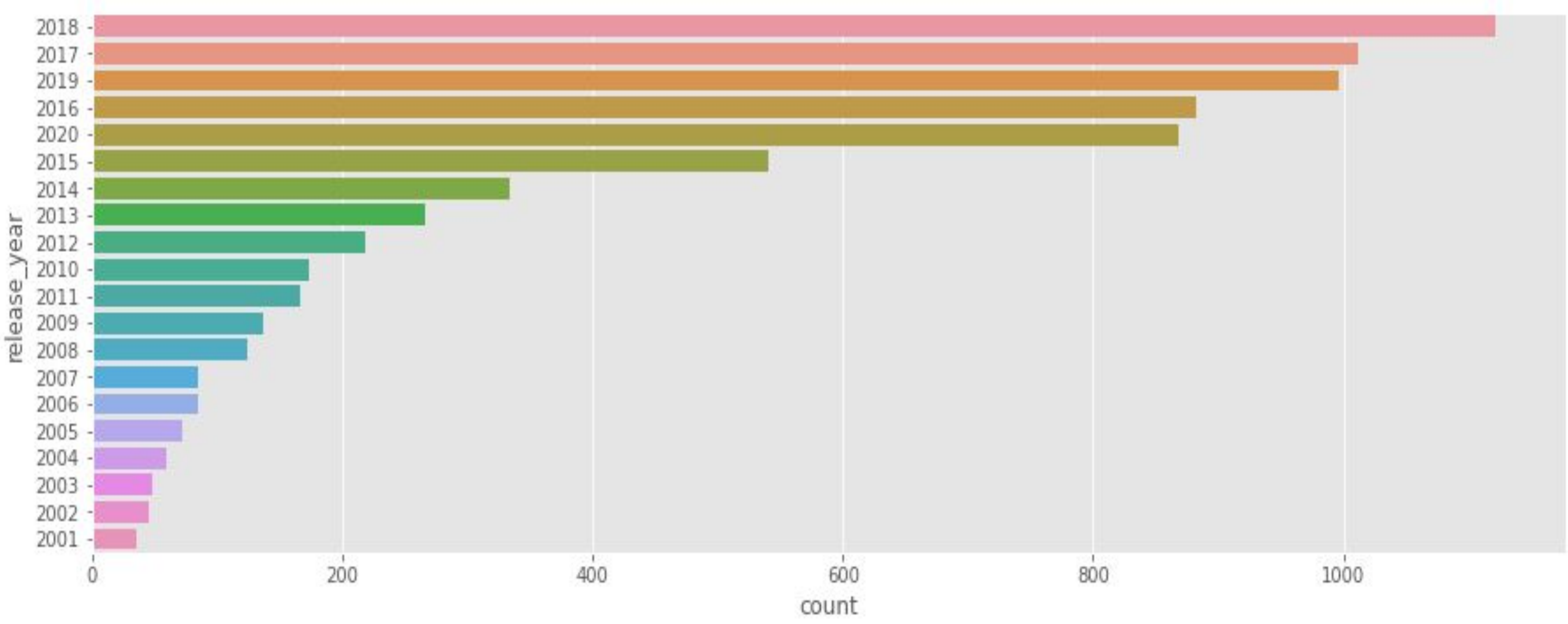
Exploratory Data Analysis

Distribution of Movie/Shows based on day added



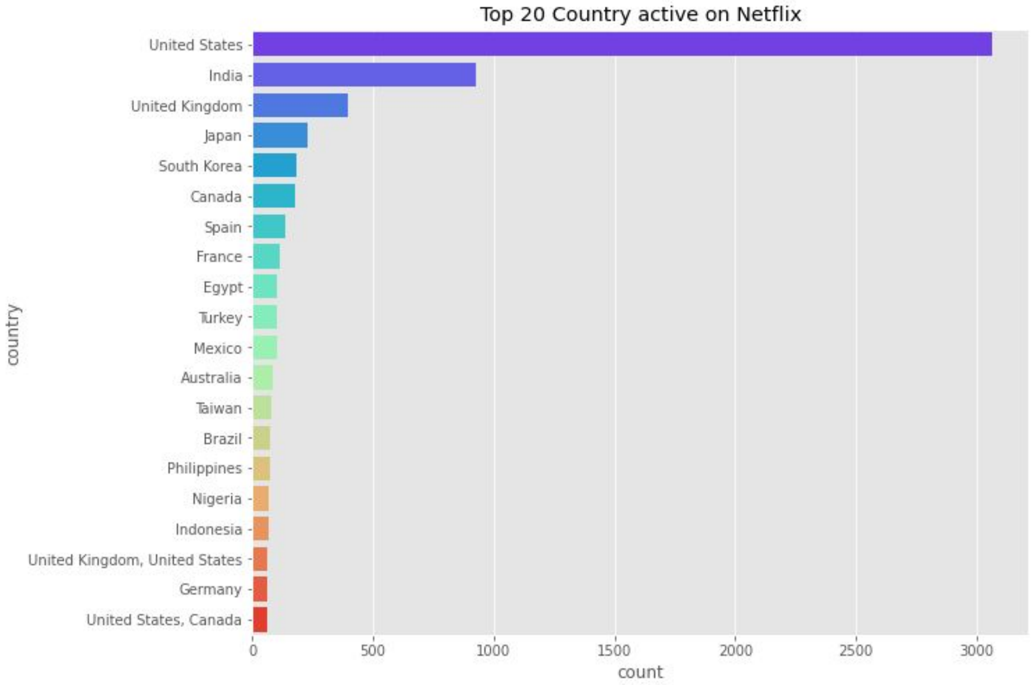
Exploratory Data Analysis

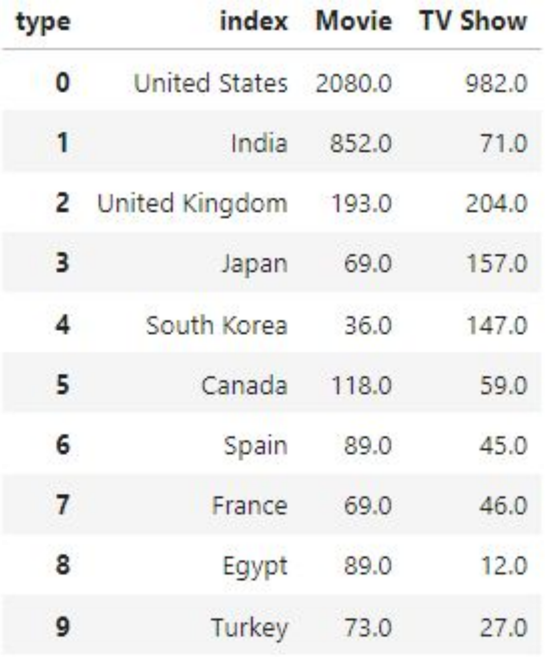
Analysing how many movies released per year in last 15 years



Exploratory Data Analysis

Top 10 Countries that produced content on Netflix

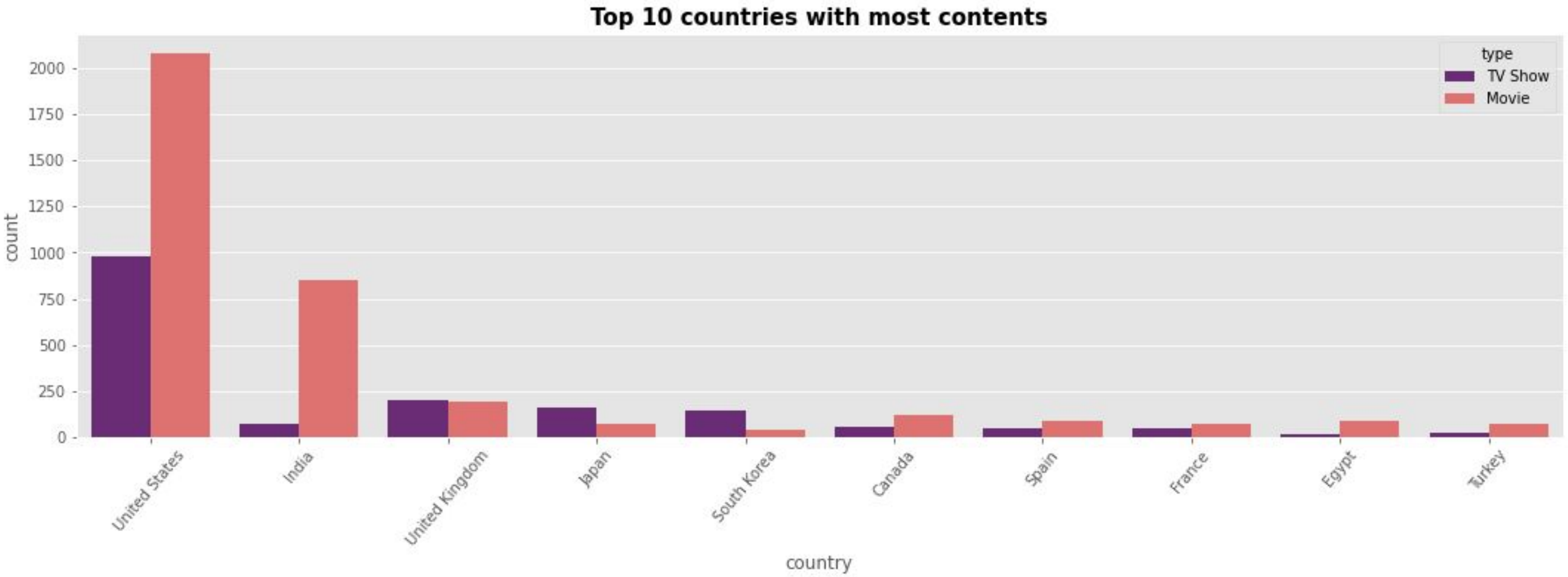




type	index	Movie	TV Show
0	United States	2080.0	982.0
1	India	852.0	71.0
2	United Kingdom	193.0	204.0
3	Japan	69.0	157.0
4	South Korea	36.0	147.0
5	Canada	118.0	59.0
6	Spain	89.0	45.0
7	France	69.0	46.0
8	Egypt	89.0	12.0
9	Turkey	73.0	27.0

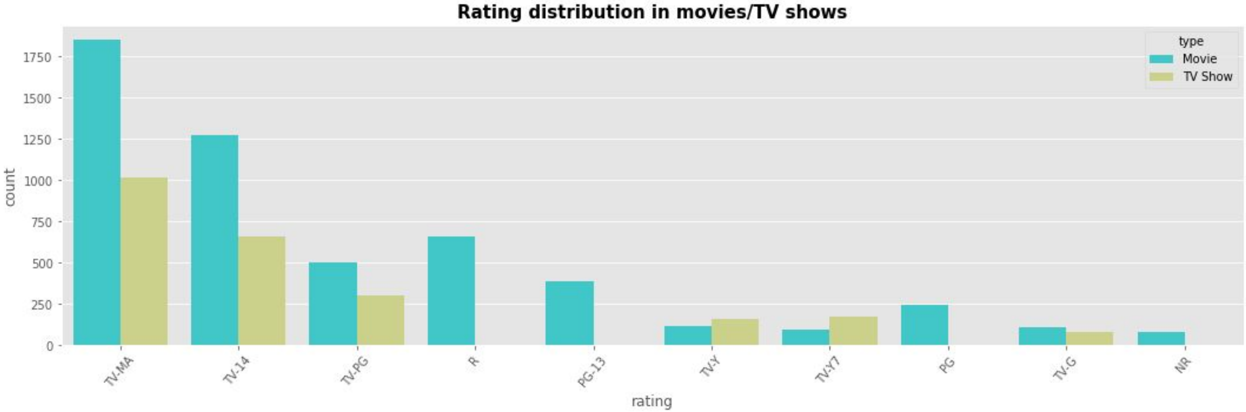
Exploratory Data Analysis

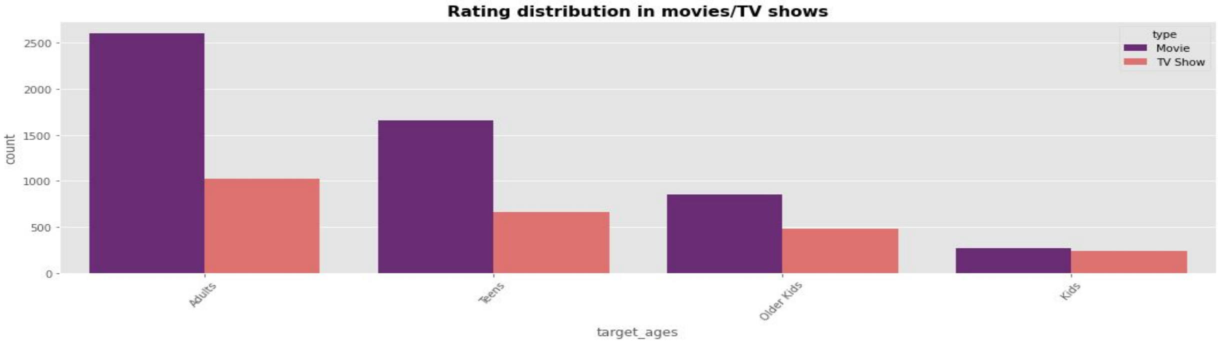
What kind of content is available in different countries in recent years?



Exploratory Data Analysis

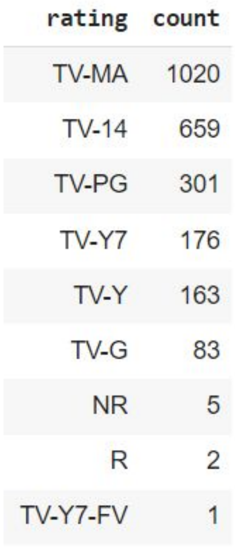
Assigning the Ratings into grouped categories


Rating distribution in movies/TV shows


Rating distribution in movies/TV shows

1. Little Kids: G, TV-Y, TV-G
   Older Kids: PG, TV-Y7, TV-Y7-FV, TV-PG
   Teens: PG-13, TV-14
   Adults: R, NC-17, TV-MA

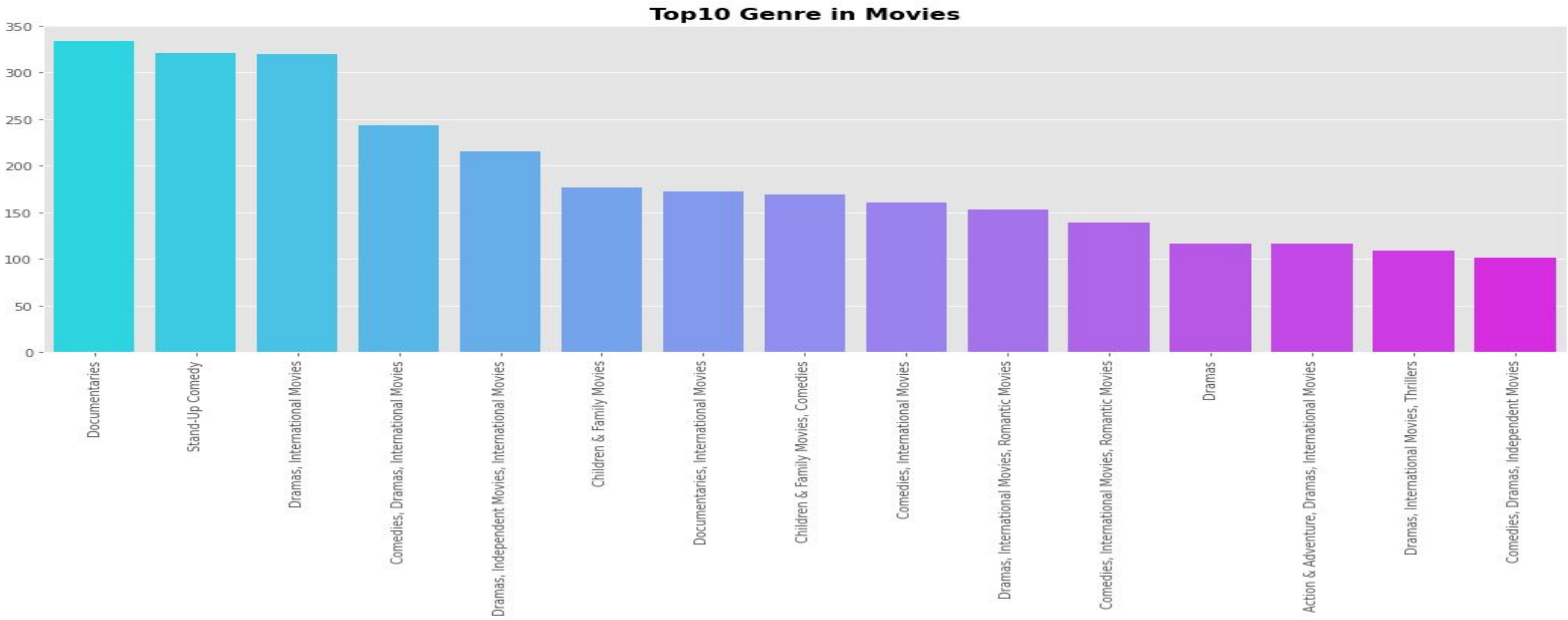
Popular Tv shows Ratings



rating	count
TV-MA	1020
TV-14	659
TV-PG	301
TV-Y7	176
TV-Y	163
TV-G	83
NR	5
R	2
TV-Y7-FV	1

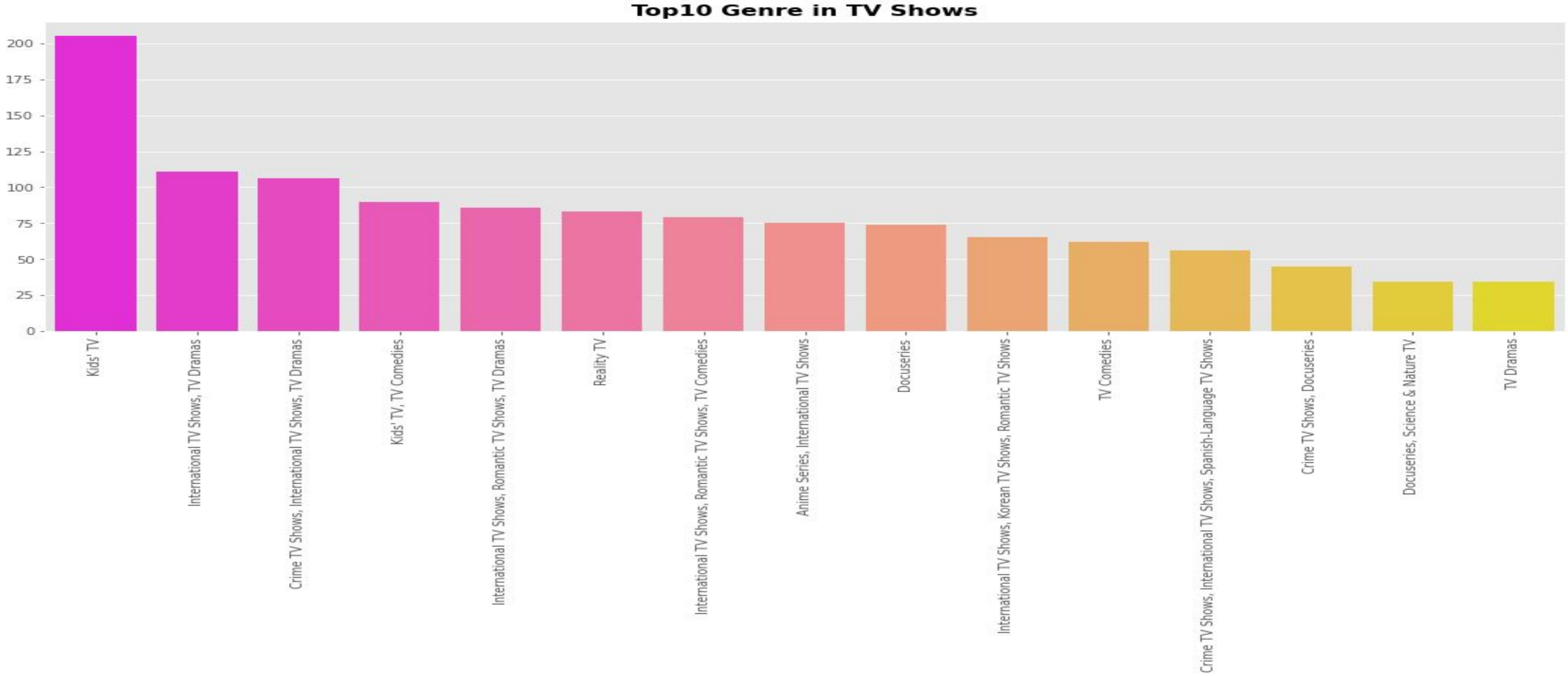
Exploratory Data Analysis

Top 10 Genre in movies

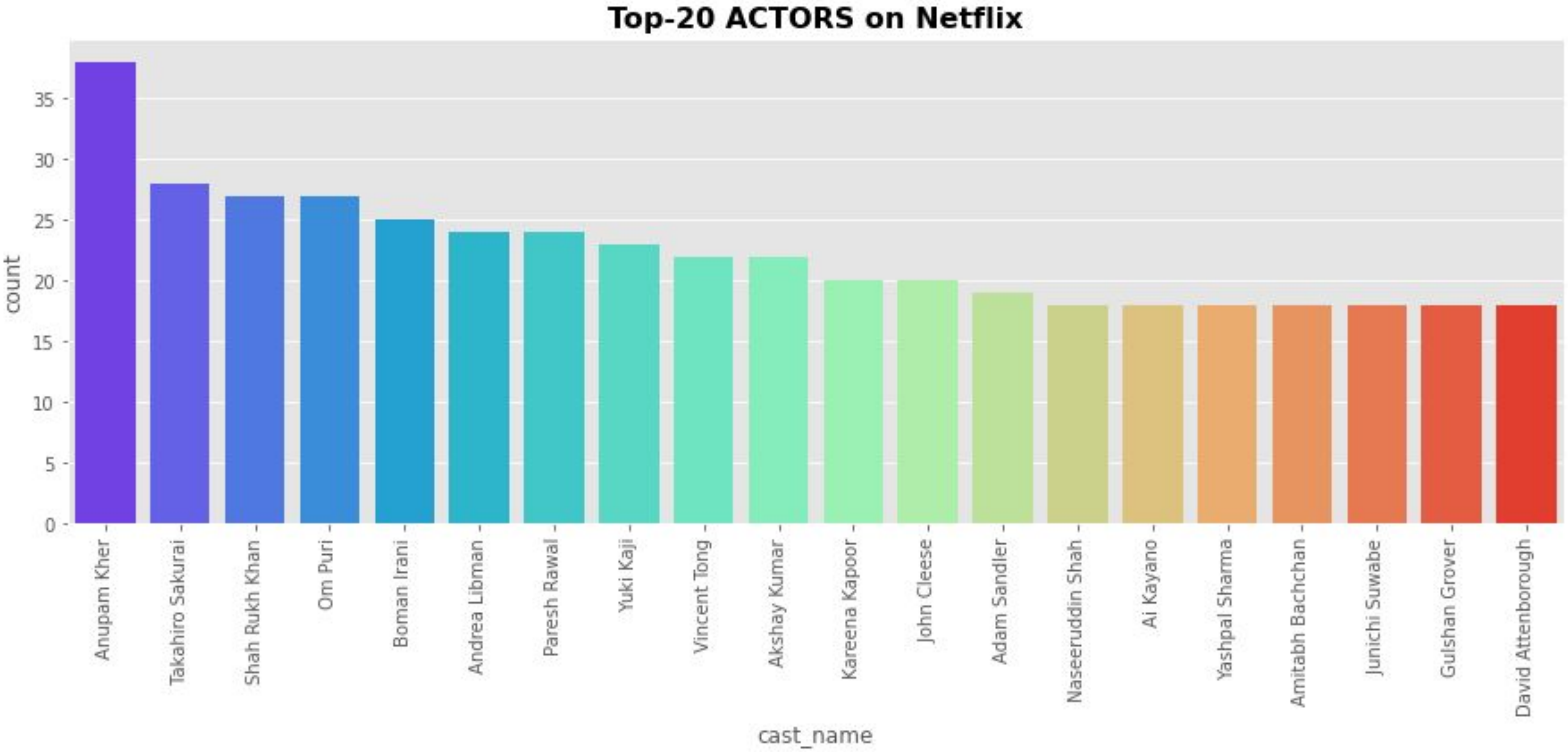


Exploratory Data Analysis

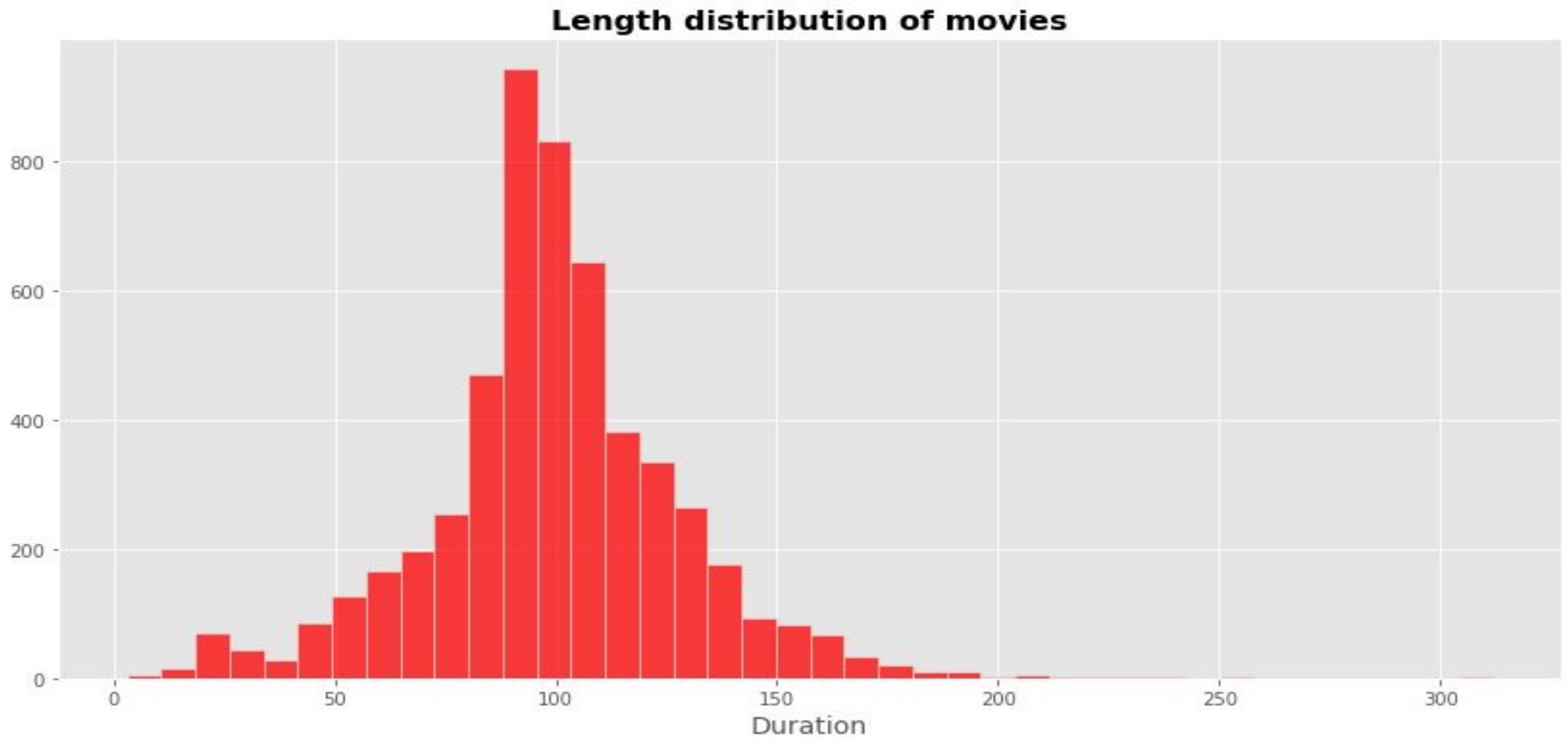
Top 10 Genre in TV Shows



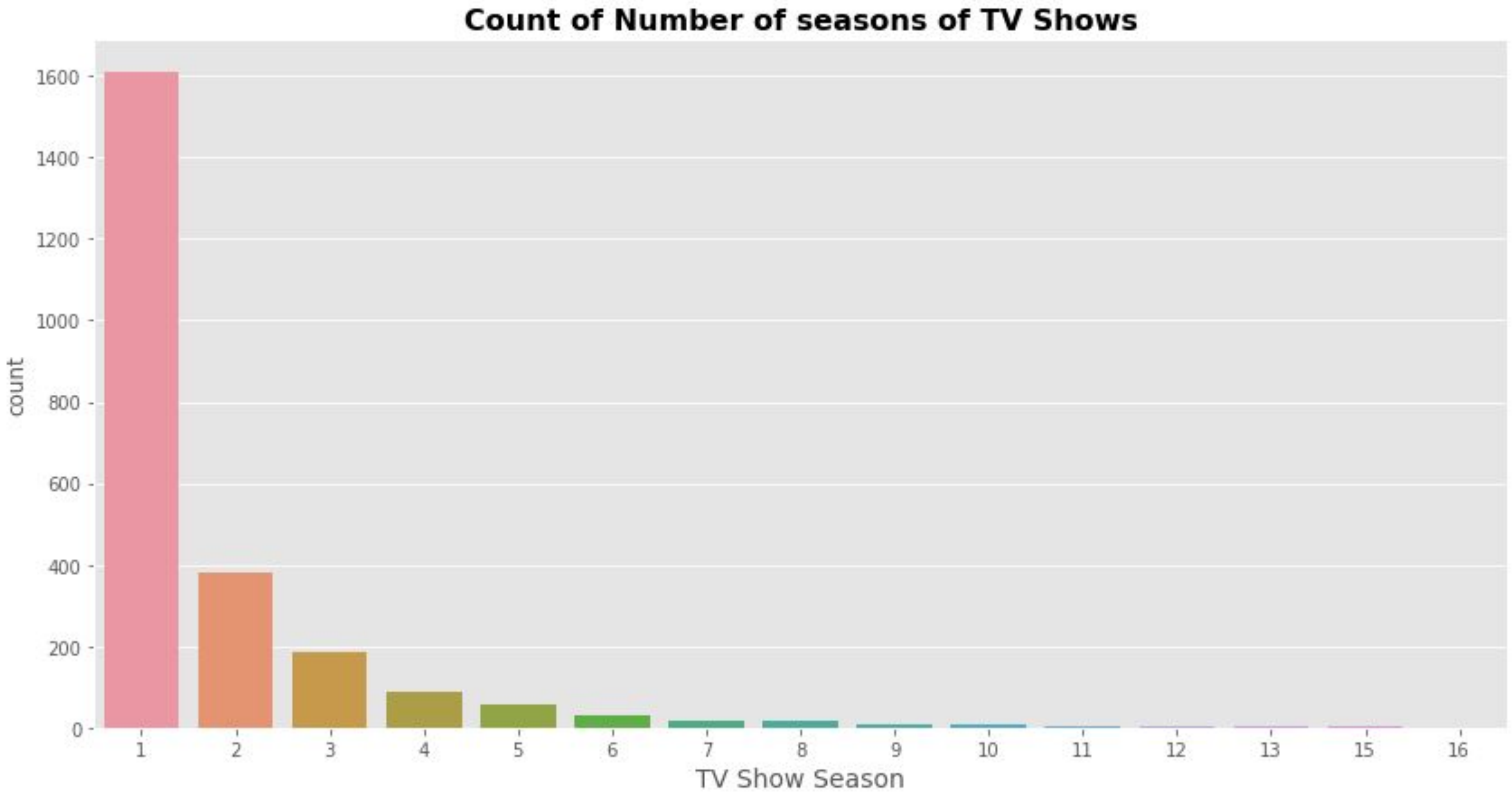
Exploratory Data Analysis



Exploratory Data Analysis

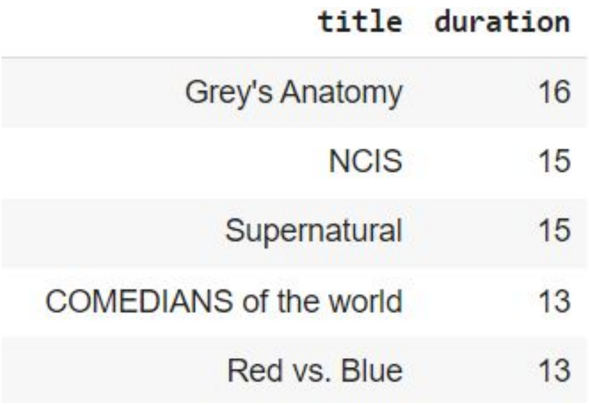


Exploratory Data Analysis



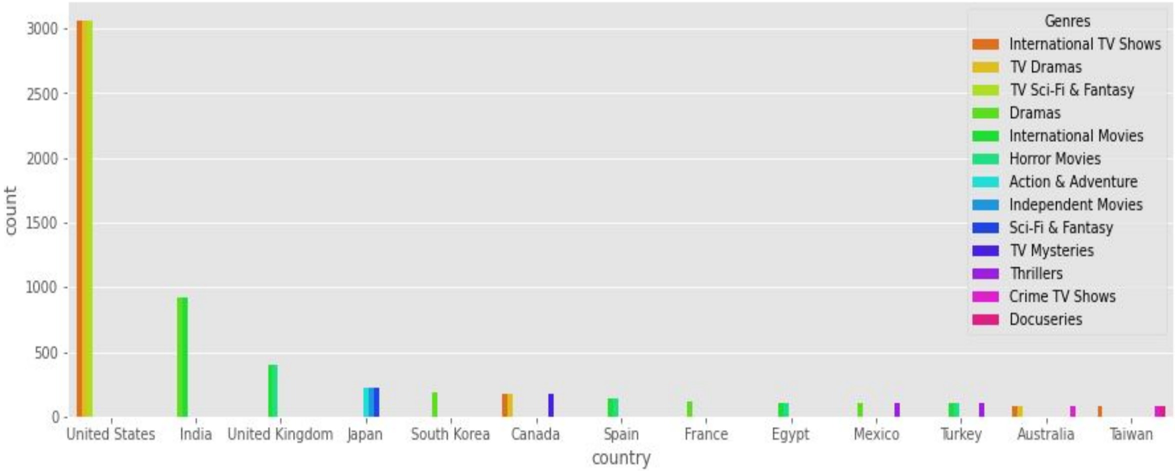
Exploratory Data Analysis

Longest TV Shows based on Number of Seasons



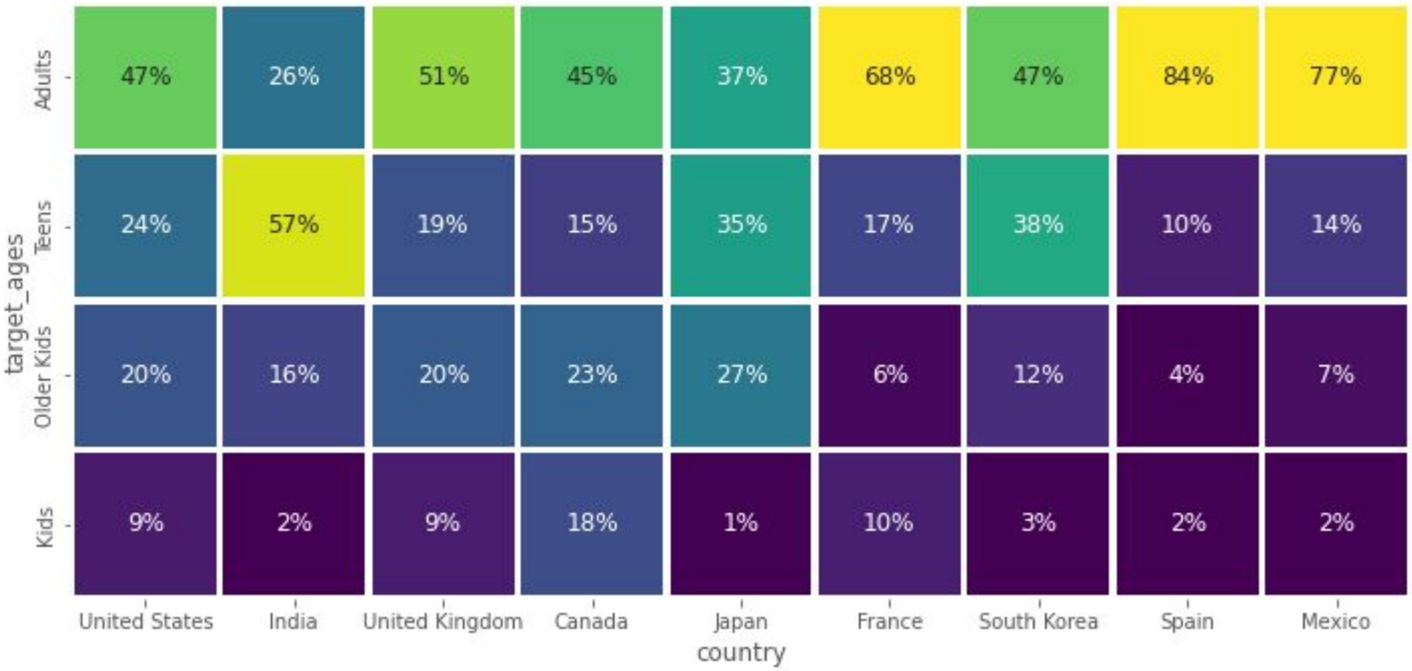
title	duration
Grey's Anatomy	16
NCIS	15
Supernatural	15
COMEDIANS of the world	13
Red vs. Blue	13

Understanding Content Produced in Different Countries



Exploratory Data Analysis

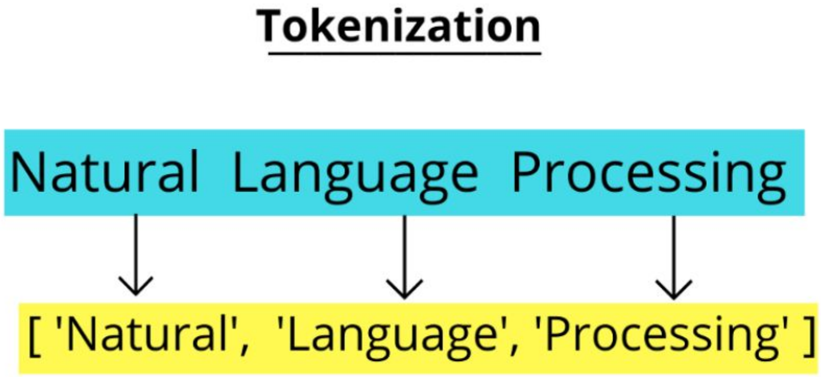
Country wise Content Production in Heatmap



Data Pre Processing

In data cleaning we have done following things-

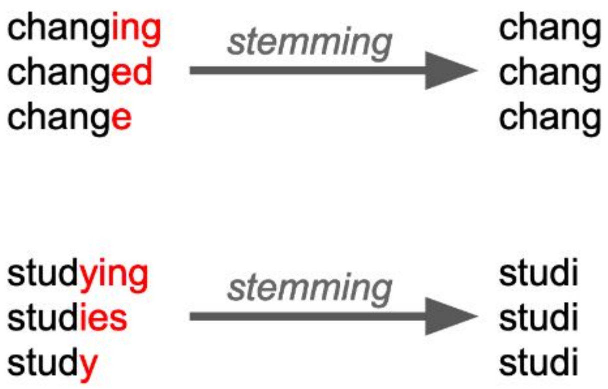
1. Removed Punctuations
2. Removed Stopwords
3. Removed Short words
4. Convert text to Lower Case
5. Stemming
6. Tokenizing
7. Lemmatization



Tokenization

Natural Language Processing

['Natural', 'Language', 'Processing']



changing
changed
change

stemming

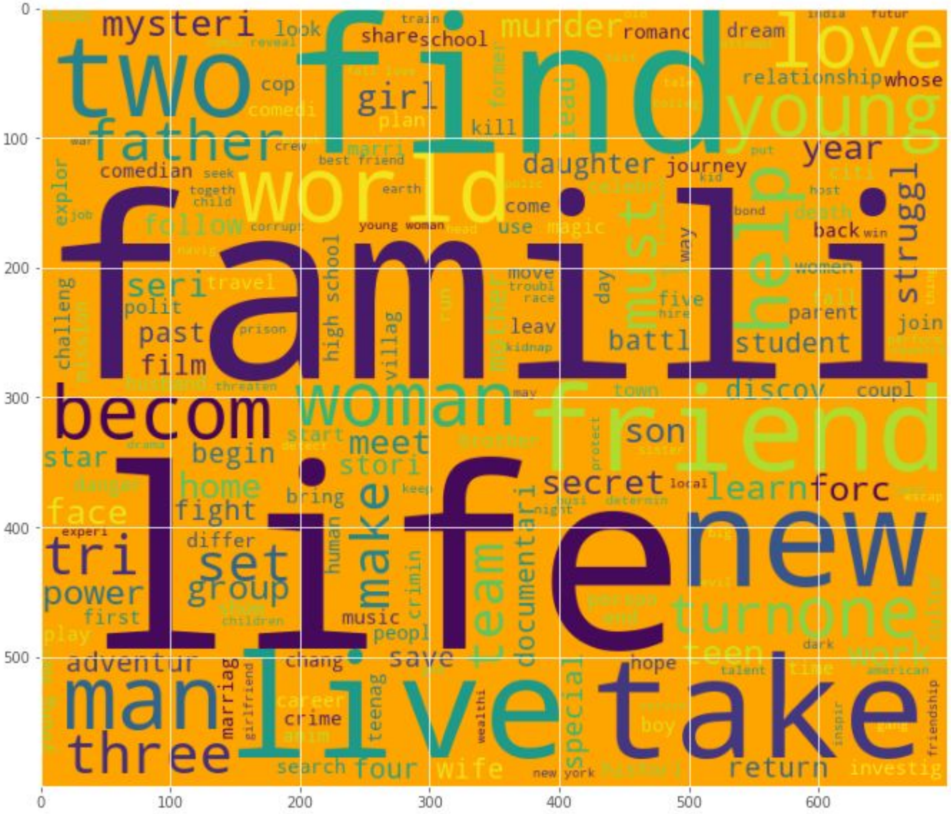
chang
chang
chang

studying
studies
study

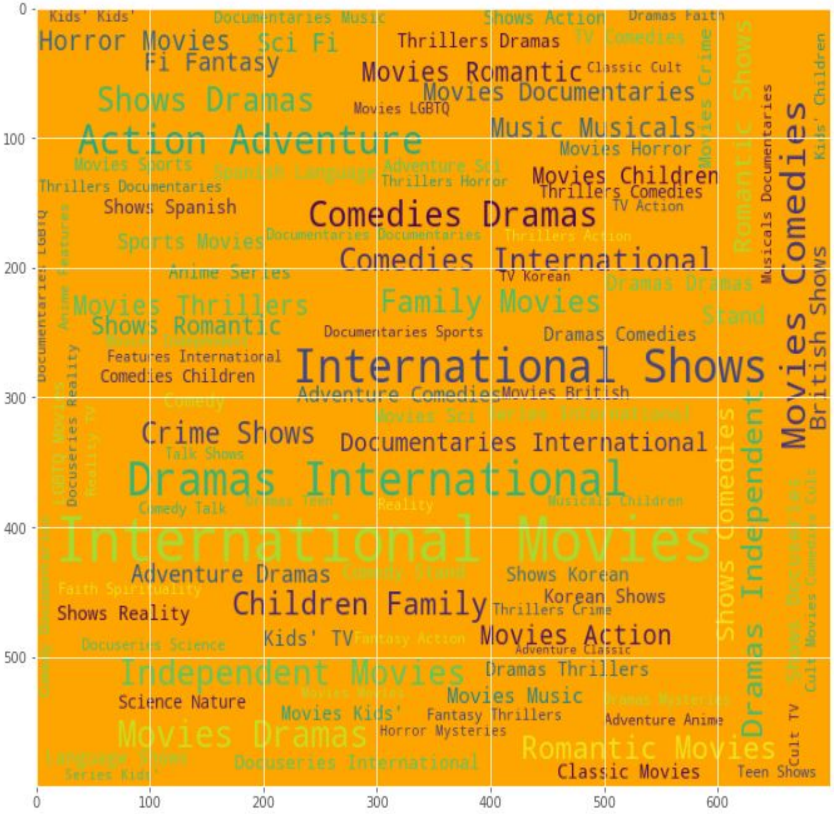
stemming

studi
studi
studi

Bag of Words

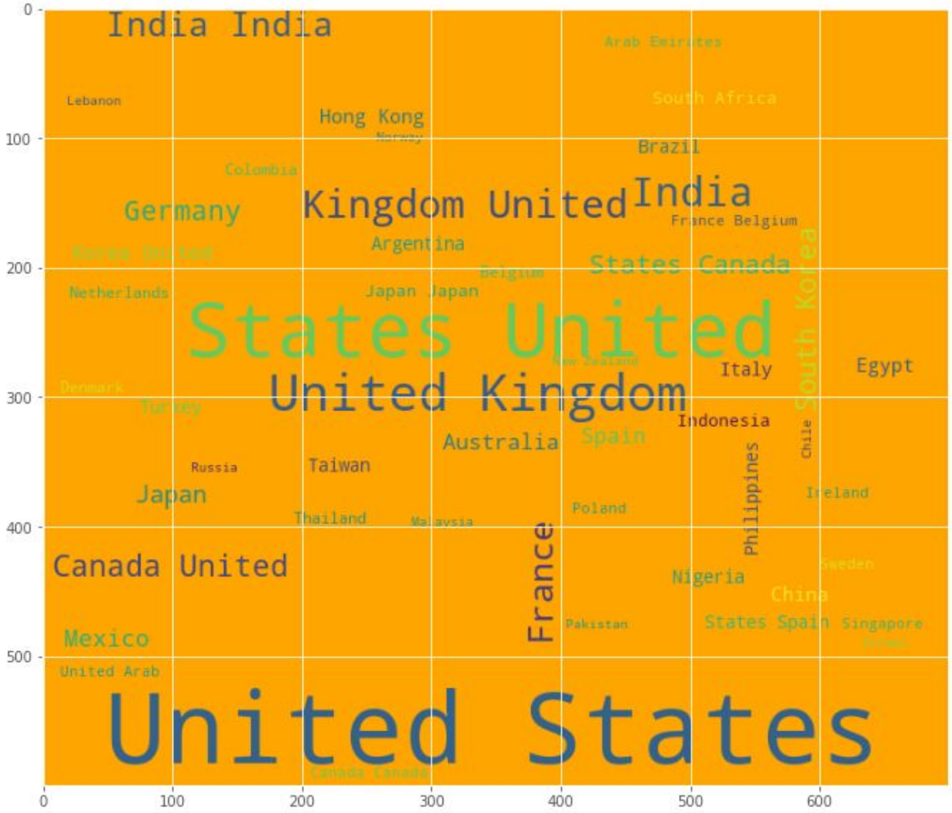


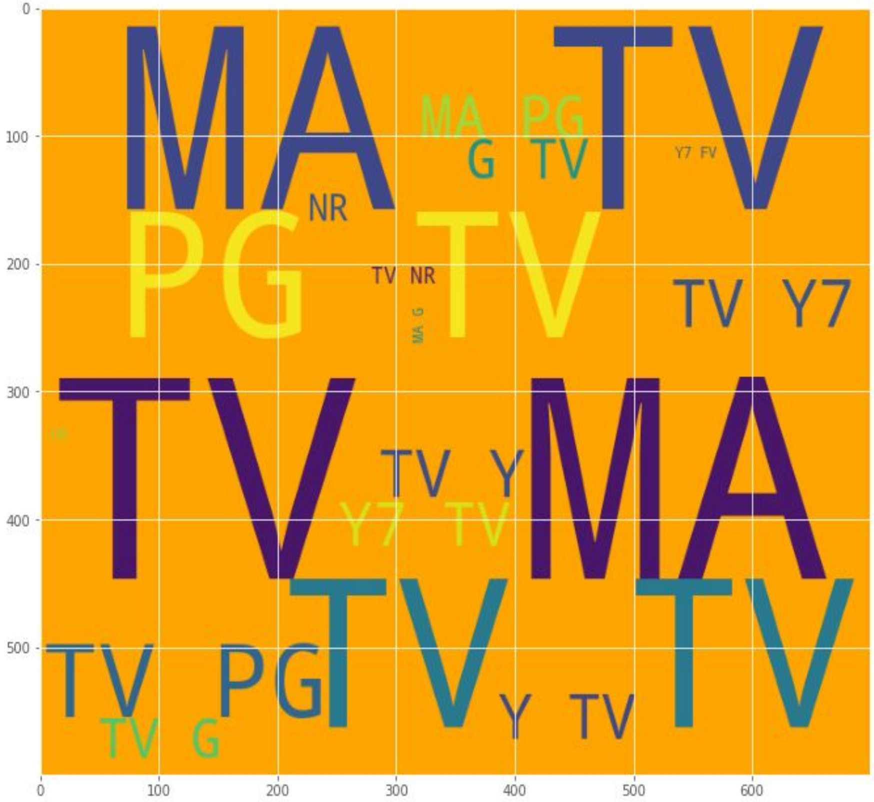
0
100
200
300
400
500
600
mysteri
look
train
india
futur
share
school
murder
romanc
dream
love
two
find
reveal
cop
girl
relationship
whose
young
comedi
plan
former
kill
father
crew
war
marri
best friend
daughter
journey
year
comedian
seek
world
earth
togeth
child
explor
job
follow
corrupt
young woman
come
use
magic
bond
death
host
back
win
struggl
famili
help
must
way
women
seri
travel
polit
past
prison
high school
villag
move
troubl
race
day
five
hire
parent
join
challeng
mission
film
leav
kidnap
battl
student
husband
threaten
becom
woman
may
town
discov
coupl
friend
son
star
drama
start
meet
begin
stori
brother
sister
protect
detect
home
danger
secret
local
learn
forc
bring
keep
face
fight
busi
determin
night
experi
life
new
tri
differ
set
group
human
make
team
documentari
crimin
power
first
music
turn
one
play
peopl
dark
adventur
chang
save
hope
teen
talent
time
american
man
marriag
girlfriend
live
take
career
crime
teenag
wealthi
special
boy
inspir
gang
friendship
three
search
four
wife
new york
return
investig



0
100
200
300
400
500
600
Kids' Kids'
Documentaries Music
Shows Action
Dramas Faith
Horror Movies
Sci Fi
Thrillers Dramas
TV Comedies
Fi Fantasy
Movies Romantic
Classic Cult
Movies Documentaries
Shows Dramas
Movies LGBTQ
Music Musicals
Movies Crime
Romantic Shows
Action Adventure
Movies Horror
Movies Sports
Spanish Language
Adventure Sci
Thrillers Horror
Movies Children
Thrillers Documentaries
Thrillers Comedies
Shows Spanish
TV Action
Comedies Dramas
Musicals Documentaries
Movies Comedies
British Shows
Kids' Children
Sports Movies
Documentaries Documentaries
Thrillers Action
Documentaries LGBTQ
Anime Features
Comedies International
TV Korean
Anime Series
Dramas Dramas
Movies Thrillers
Family Movies
Stand
Shows Romantic
Documentaries Sports
Dramas Comedies
Features International
Comedies Children
Reality TV
International Shows
Adventure Comedies
Movies British
Comedy
Movies Sci
Series International
Docuseries Reality
LGBTQ Movies
Crime Shows
Talk Shows
Documentaries International
Shows Independent
Dramas International
Comedy Talk
Dramas Teen
Reality
Musicals Children
International Movies
Adventure Dramas
Comedy Stand
Shows Korean
Faith Spirituality
Korean Shows
Children Family
Thrillers Crime
Shows Reality
Kids' TV
Fantasy Action
Movies Action
Docuseries Science
Adventure Classic
Independent Movies
Dramas Thrillers
Shows Comedies
Dramas Independent
Science Nature
Movies Music
Movies Kids'
Fantasy Thrillers
Adventure Anime
Dramas Mysteries
Horror Mysteries
Movies Dramas
Romantic Movies
Cult TV
Language Shows
Series Kids'
Docuseries International
Classic Movies
Teen Shows

Bag of Words

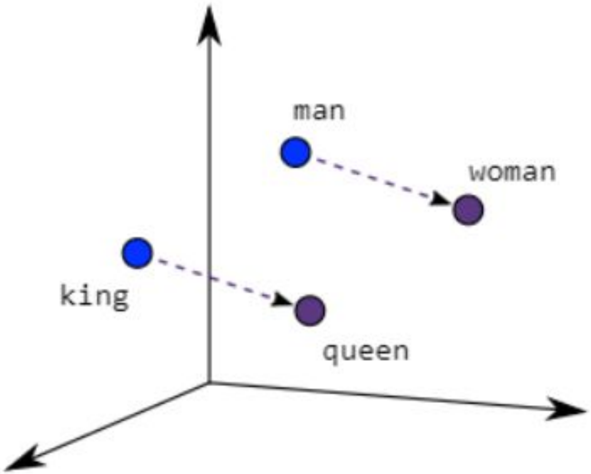

India India
Arab Emirates
Lebanon
South Africa
Hong Kong
Norway
Brazil
Colombia
Germany
Kingdom United
India
France Belgium
Argentina
Korea United
Belgium
States Canada
Netherlands
Japan Japan
South Korea
States United
New Zealand
Italy
Egypt
Denmark
United Kingdom
Turkey
Indonesia
Australia
Spain
Chile
Russia
Taiwan
Philippines
Japan
Ireland
Poland
Thailand
Malaysia
France
Canada United
Sweden
Nigeria
China
Pakistan
States Spain
Singapore
Mexico
United Arab
United States
Canada Canada
0
100
200
300
400
500
600


MA
TV
MA
PG
G
TV
Y7 FV
NR
PG
TV
TV NR
MA G
TV
Y7
TV
MA
TV
Y
Y7
TV
TV
PG
TV
Y
TV
TV
TV
G
0
100
200
300
400
500
600

Feature Extraction

Word2vec

Word2vec is a technique for natural language processing published in 2013. The word2vec algorithm uses a neural network model to learn word associations from a large corpus of text. Once trained, such a model can detect synonymous words or suggest additional words for a partial sentence. Word2Vec creates vectors of the words that are distributed numerical representations of word features – these word features could comprise of words that represent the context of the individual words present in our vocabulary. Word embeddings eventually help in establishing the association of a word with another similar meaning word through the created vectors.



Feature Extraction

CountVectorizer- CountVectorizer means breaking down a sentence or any text into words by performing preprocessing tasks like converting all words to lowercase, thus removing special characters. In NLP models can't understand textual data they only accept numbers, so this textual data needs to be vectorized.

TfidfVectorizer - Transforms text to feature vectors that can be used as input to estimator. vocabulary_ Is a dictionary that converts each token (word) to feature index in the matrix, each unique token gets a feature index.

Feature Extraction

PART A: Modelling with Word2Vec For Word Embeddings

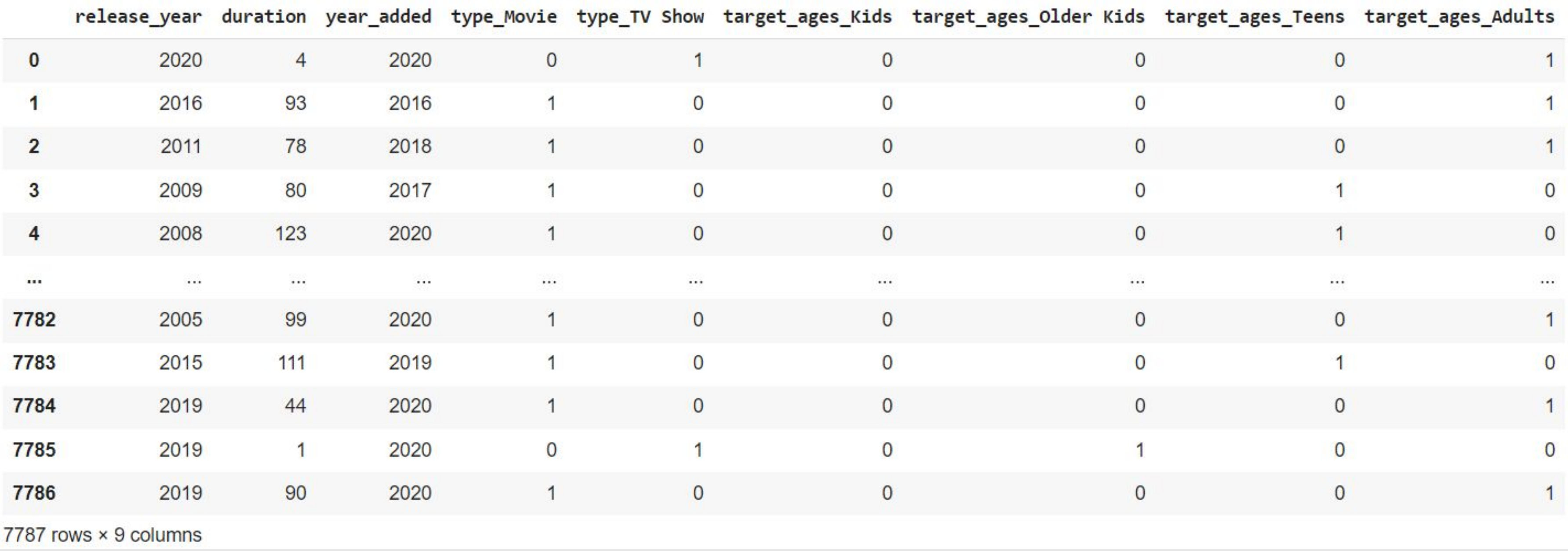
word2vec-google-news-300

Word2Vec

Pre-trained vectors trained on a part of the Google News dataset (about 100 billion words). The model contains 300-dimensional vectors for 3 million words and phrases. The phrases were obtained using a simple data-driven approach described in 'Distributed Representations of Words and Phrases and their Compositionality'

Feature Extraction

Encoding Categorical Variables- One Hot Encoding

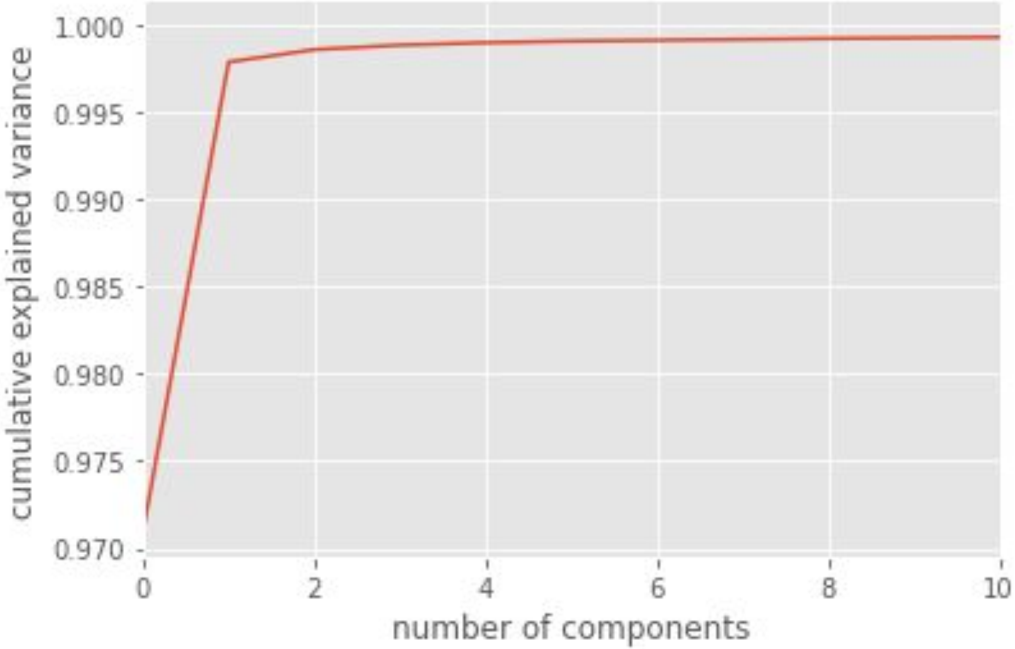


	release_year	duration	year_added	type_Movie	type_TV Show	target_ages_Kids	target_ages_Older Kids	target_ages_Teens	target_ages_Adults
0	2020	4	2020	0	1	0	0	0	1
1	2016	93	2016	1	0	0	0	0	1
2	2011	78	2018	1	0	0	0	0	1
3	2009	80	2017	1	0	0	0	1	0
4	2008	123	2020	1	0	0	0	1	0
...	...	...	...	...	...	...	...	...	...
7782	2005	99	2020	1	0	0	0	0	1
7783	2015	111	2019	1	0	0	0	1	0
7784	2019	44	2020	1	0	0	0	0	1
7785	2019	1	2020	0	1	0	1	0	0
7786	2019	90	2020	1	0	0	0	0	1

7787 rows × 9 columns

Feature Selection

Principal Component Analysis



'type', 'country',
'release_year',
'duration', 'listed_in',
'target_ages',
'new_description',
'Genres',
'new_country'

Modeling of Clusters

K-Means Clustering

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters.

It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

Agglomerative clustering algorithm

Agglomerative Clustering is **a type of hierarchical clustering algorithm**. It is an unsupervised machine learning technique that divides the population into several clusters such that data points in the same cluster are more similar and data points in different clusters are dissimilar.

Affinity Propagation

affinity propagation: An algorithm that **identifies exemplars among data points and forms clusters of data points around these exemplars**. It operates by simultaneously considering all data point as potential exemplars and exchanging messages between data points until a good set of exemplars and clusters emerges.

Evaluation Criteria

Calculation of Silhouette score

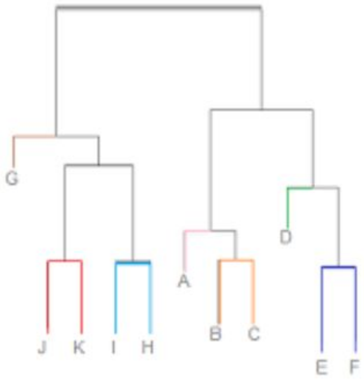
Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters. To calculate the Silhouette score for each observation/data point, the following distances need to be found out for each observations belonging to all the clusters: 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.

Elbow Method to get number of clusters

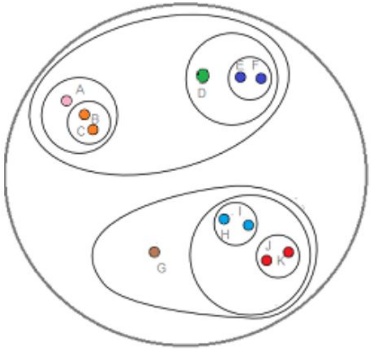
The K-Elbow Visualizer implements the “elbow” method of selecting the optimal number of clusters for K-means clustering.The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center.When these overall metrics for each model are plotted, it is possible to visually determine the best value for k. If the line chart looks like an arm, then the “elbow” (the point of inflection on the curve) is the best value of k. The “arm” can be either up or down, but if there is a strong inflection point, it is a good indication that the underlying model fits best at that point.

Evaluation Criteria

Dendrogram: Hierarchical clustering is where you build a cluster tree (a dendrogram) to represent data, where each group (or “node”) links to two or more successor groups. The groups are nested and organized as a tree, which ideally ends up as a meaningful classification scheme.

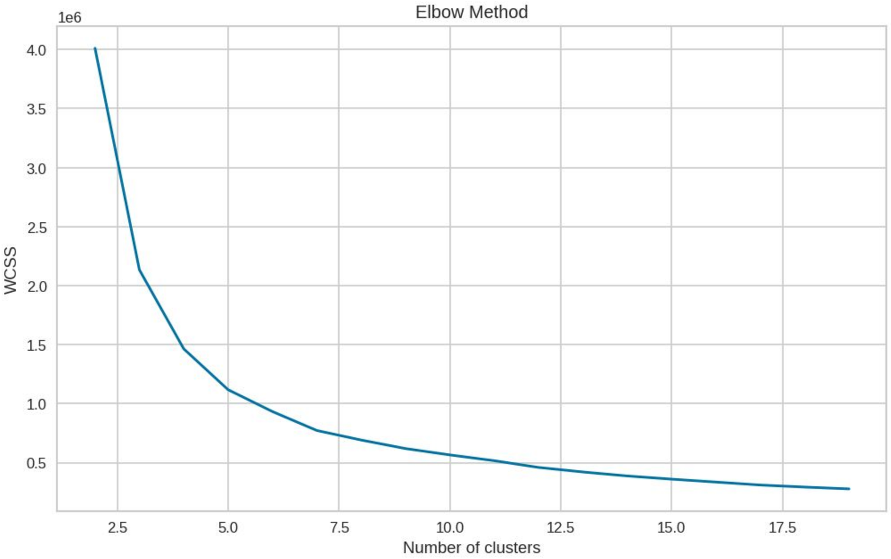


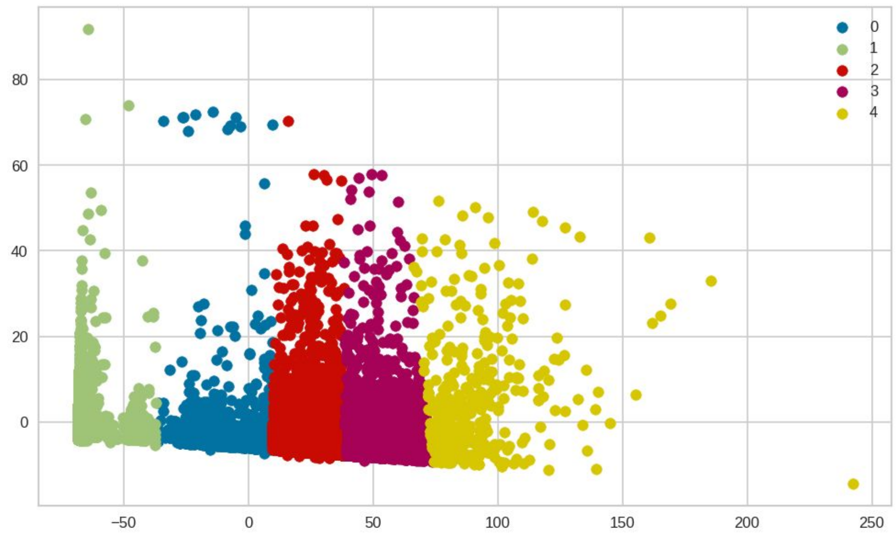
Each node in the cluster tree contains a group of similar data; Nodes group on the graph next to other, similar nodes. Clusters at one level join with clusters in the next level up, using a degree of similarity; The process carries on until all nodes are in the tree, which gives a visual snapshot of the data contained in the whole set. The total number of clusters is *not* predetermined before you start the tree creation.



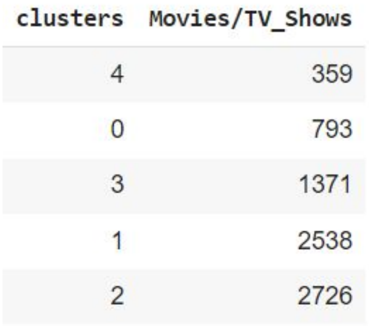
K Mean with Elbow Method

Optimum Clusters: 5
silhouette_score is : 0.5237

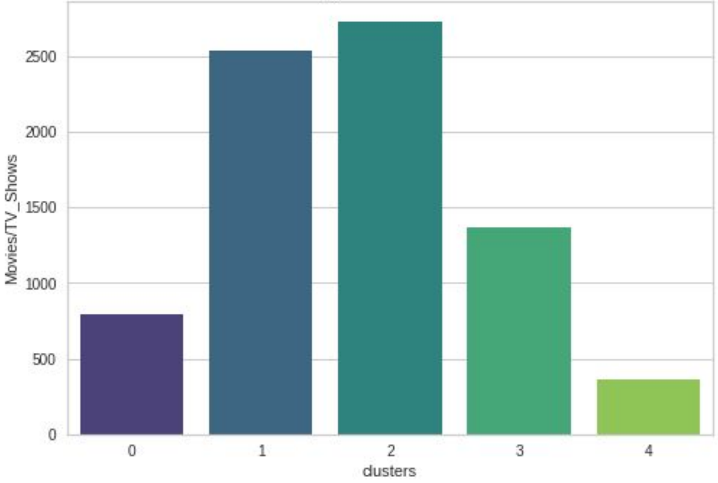




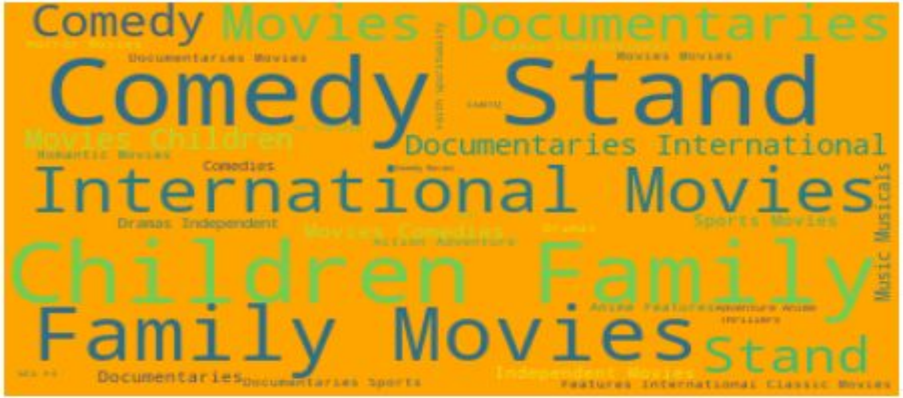
K Mean with Elbow Method



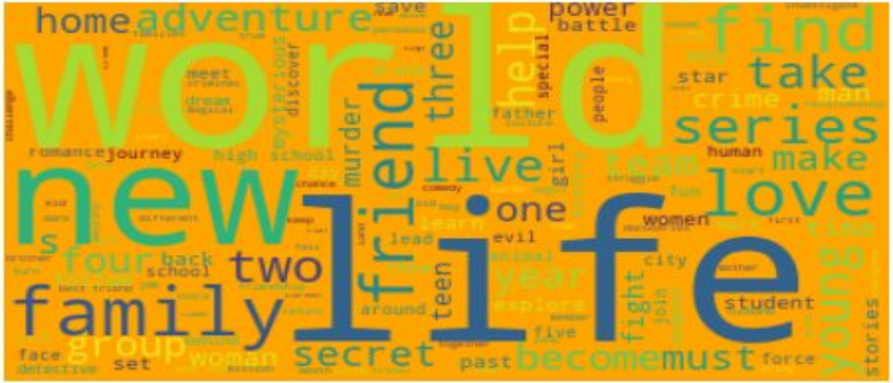
clusters	Movies/TV_Shows
4	359
0	793
3	1371
1	2538
2	2726



Counting content in each cluster

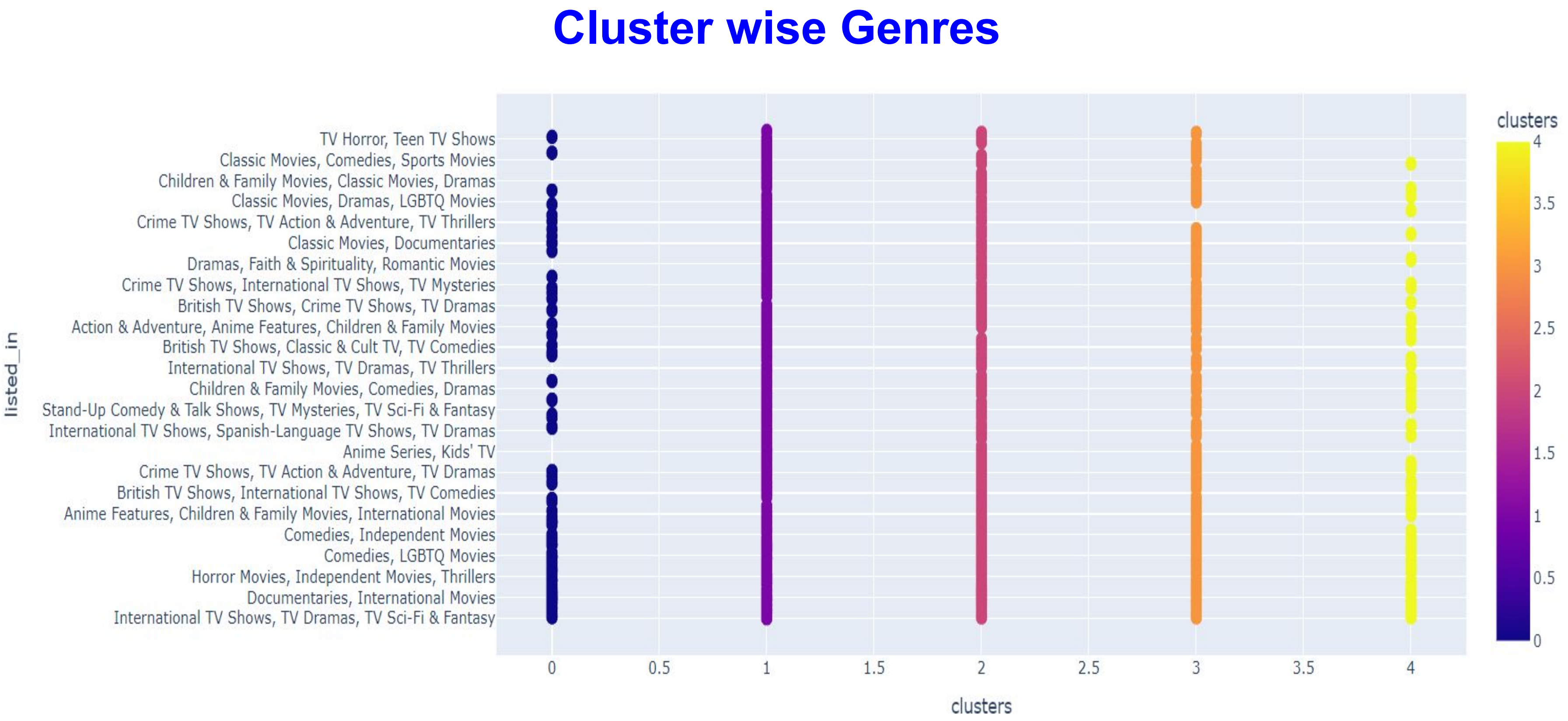


Genres in Cluster: 0

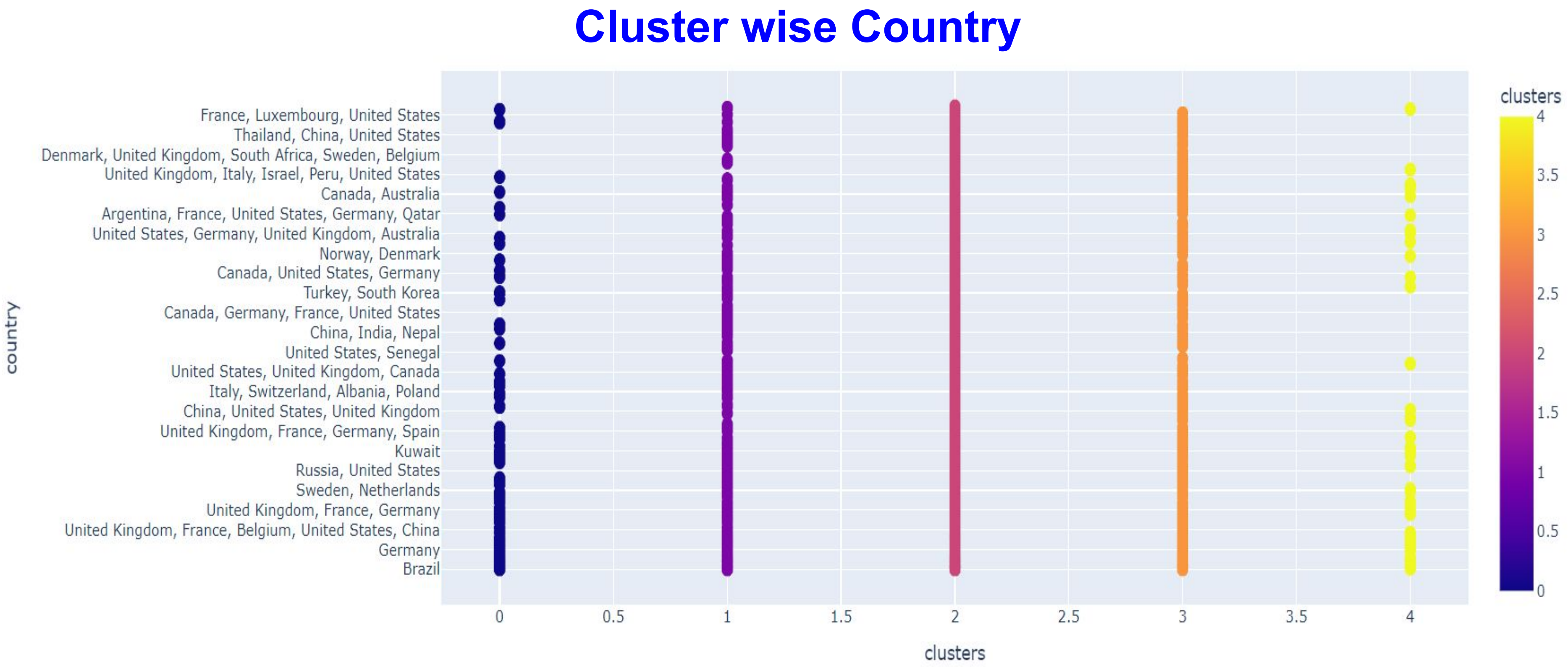


Description in Cluster: 1

K Mean with Elbow Method

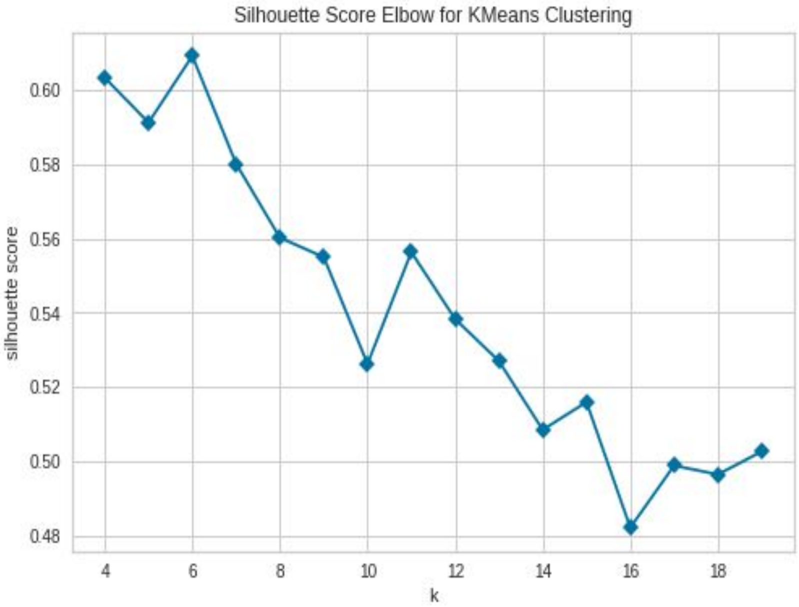


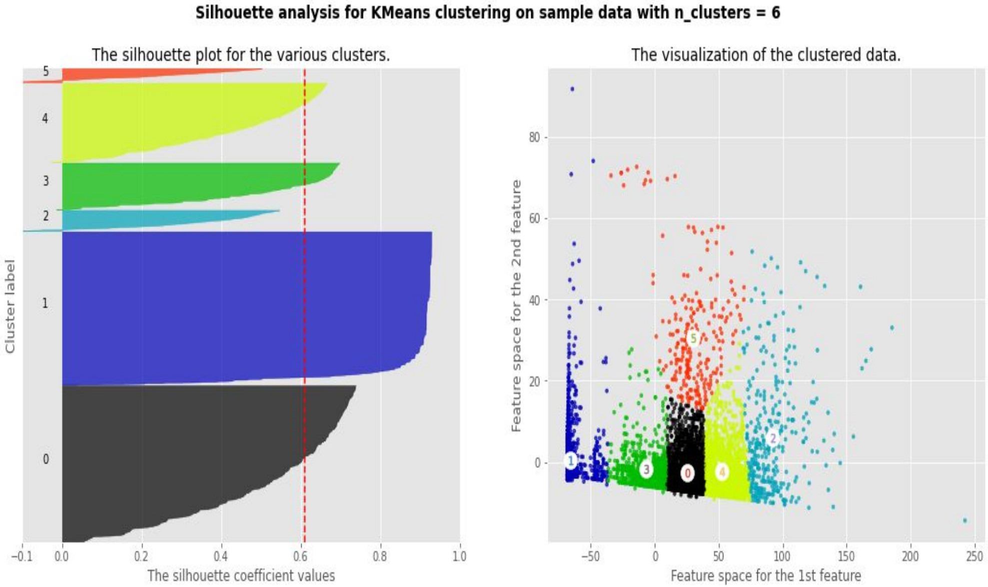
K Mean with Elbow Method



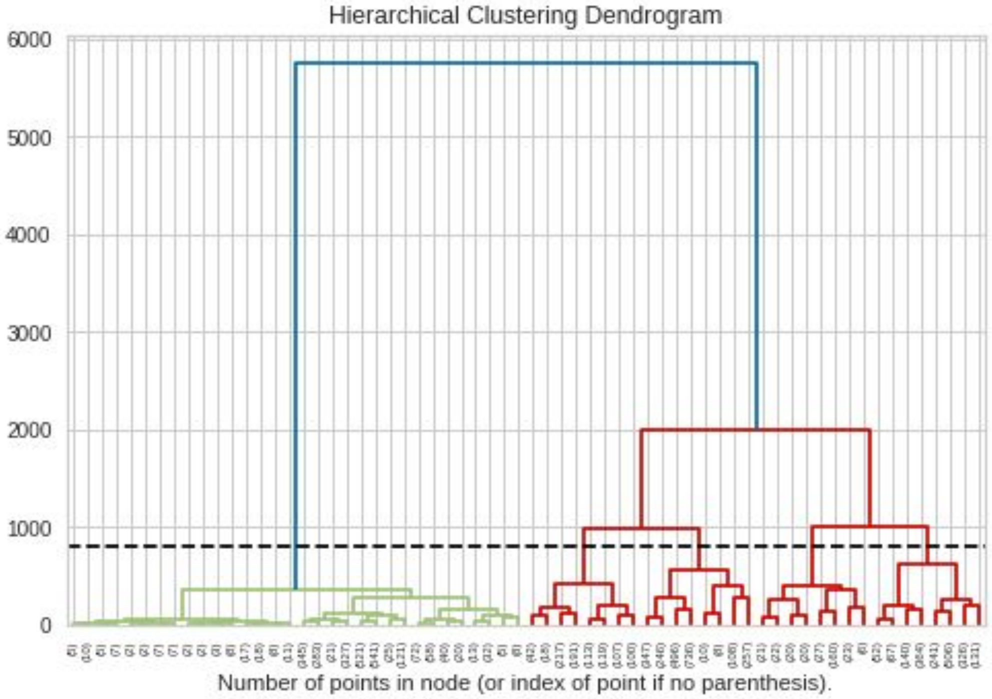
K Mean with Silhouette Score

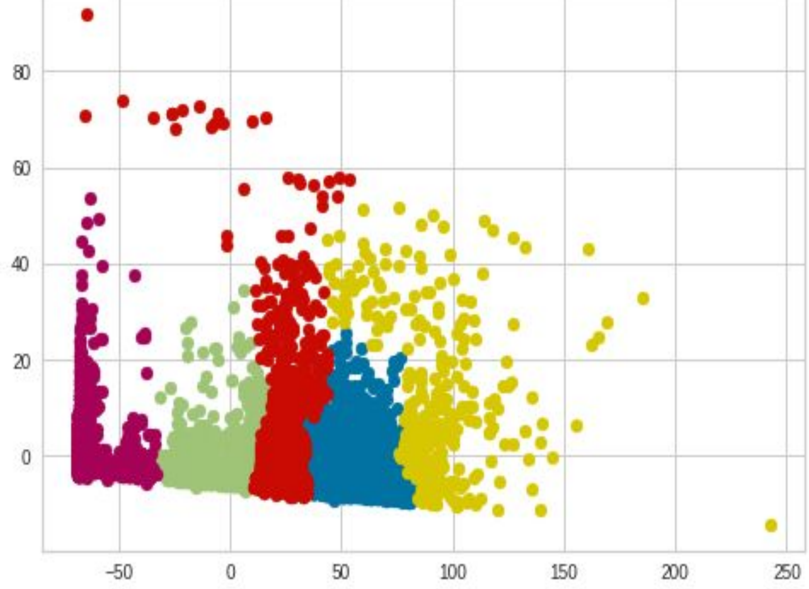
Optimum Clusters: 6
For n_clusters = 6 The average silhouette_score is : 0.6094





Agglomerative clustering using Dendrogram

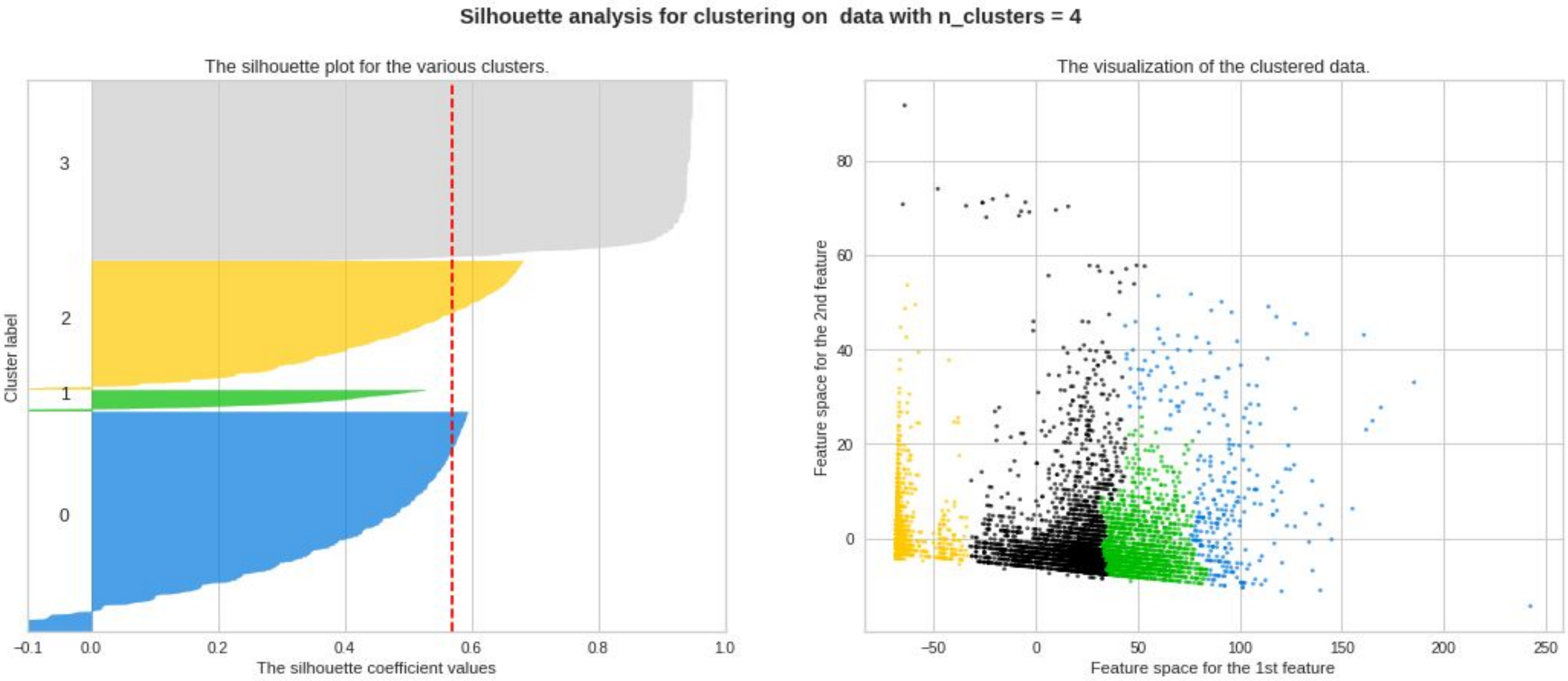




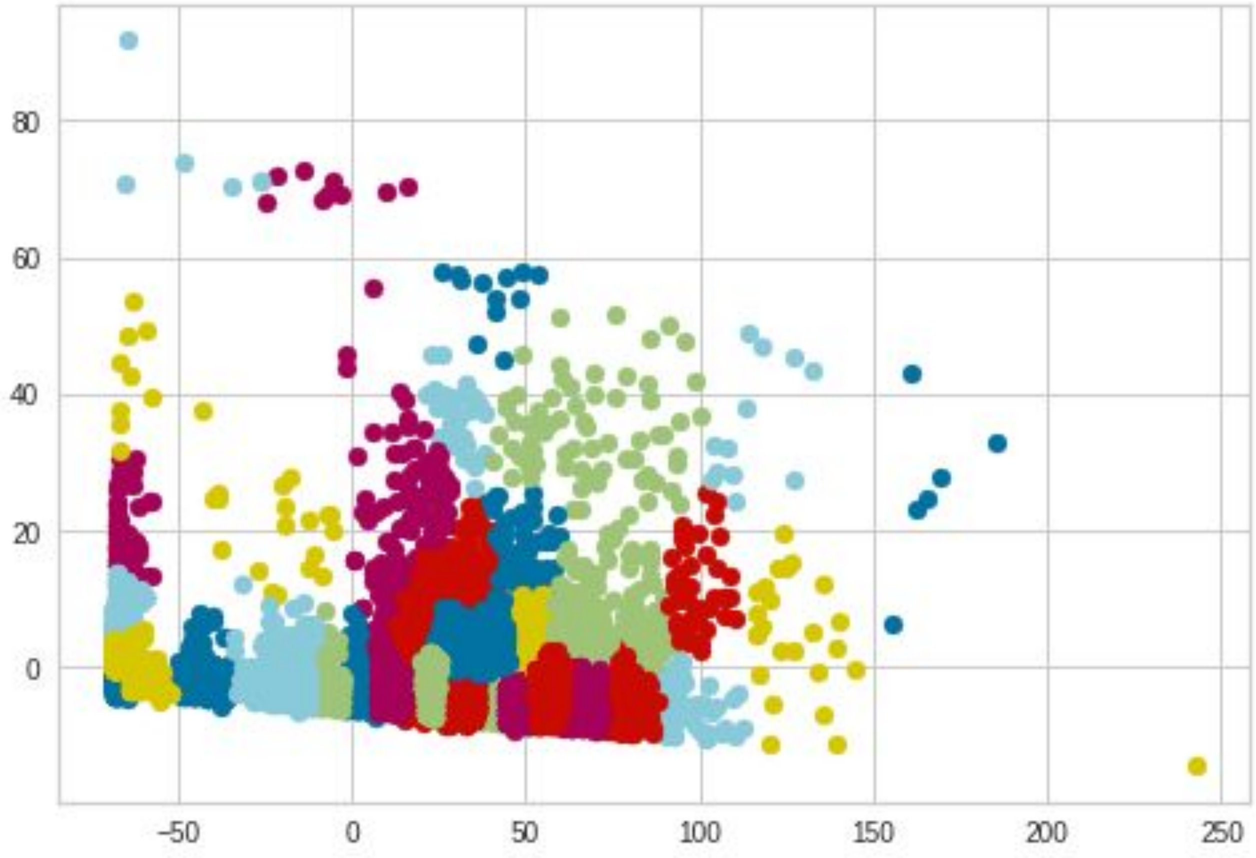
Silhouette Coefficient: 0.561
davies_bouldin_score 0.683

Agglomerative clustering using silhouette_score

For n_clusters = 4 The average silhouette_score is : 0.56898



Affinity Propagation Clustering



Silhouette Coefficient: 0.406

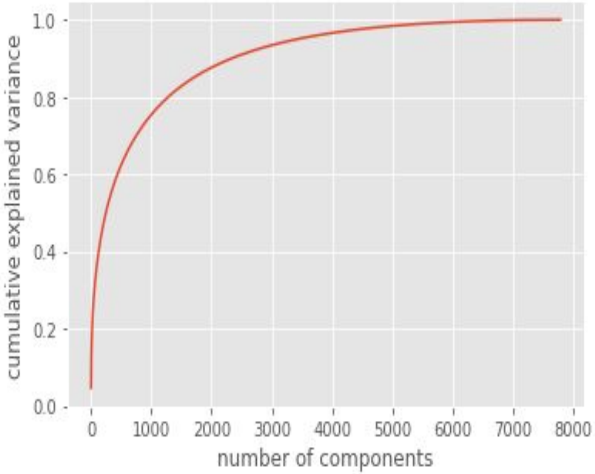
davies_bouldin_score

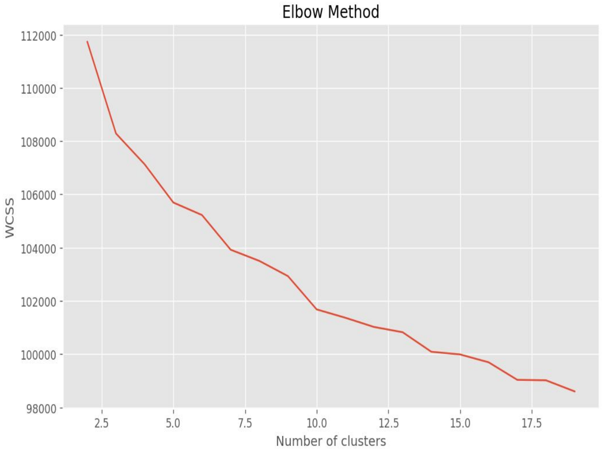
0.753

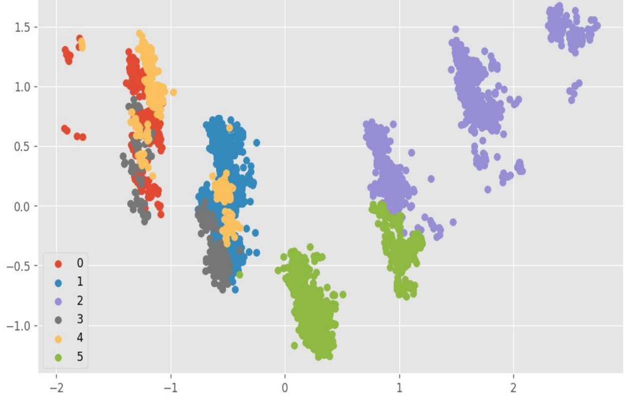
Clusters: 5

PART B: Modelling with CountVectorizer and TfidfVectorizer

K Mean with CountVectorizer and TfidfVectorizer Elbow Method







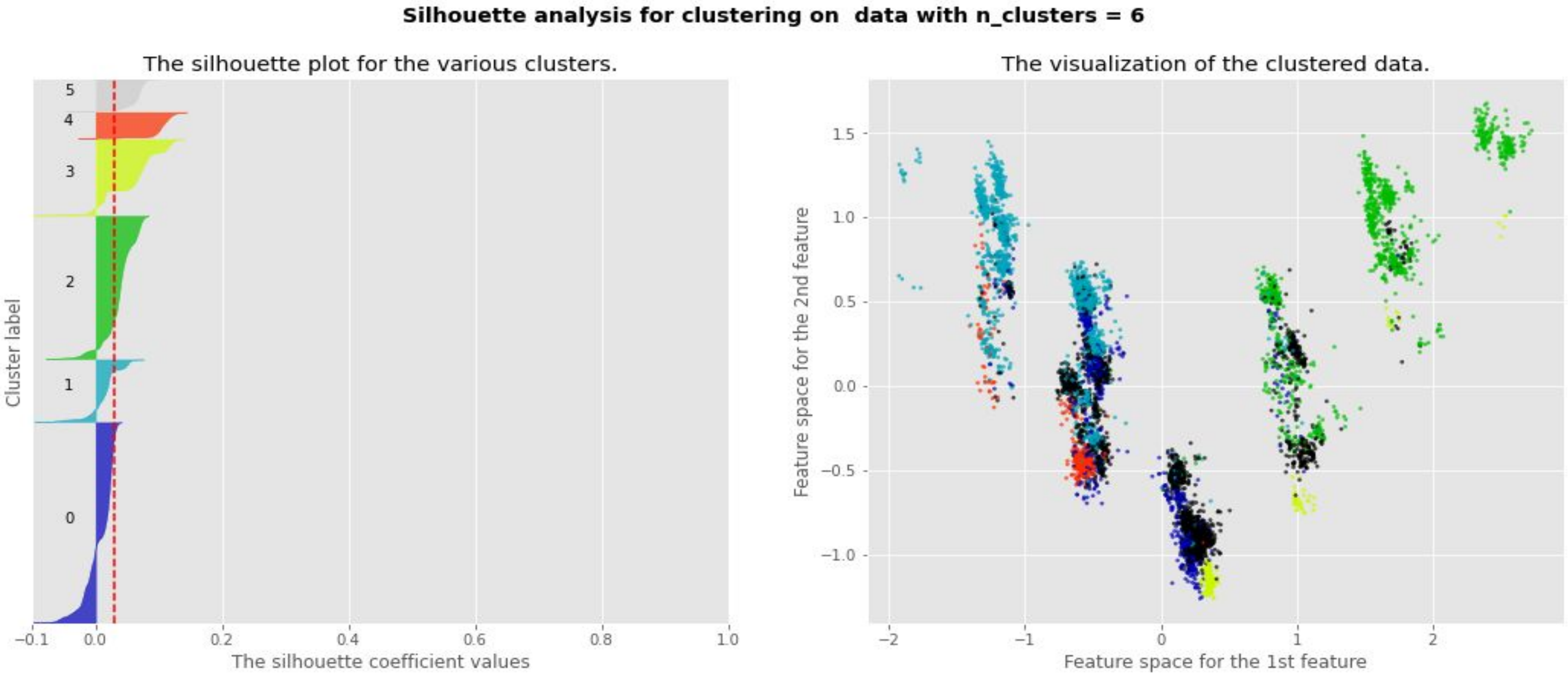
PCA with 2000
Component

Elbow method for 6 clusters

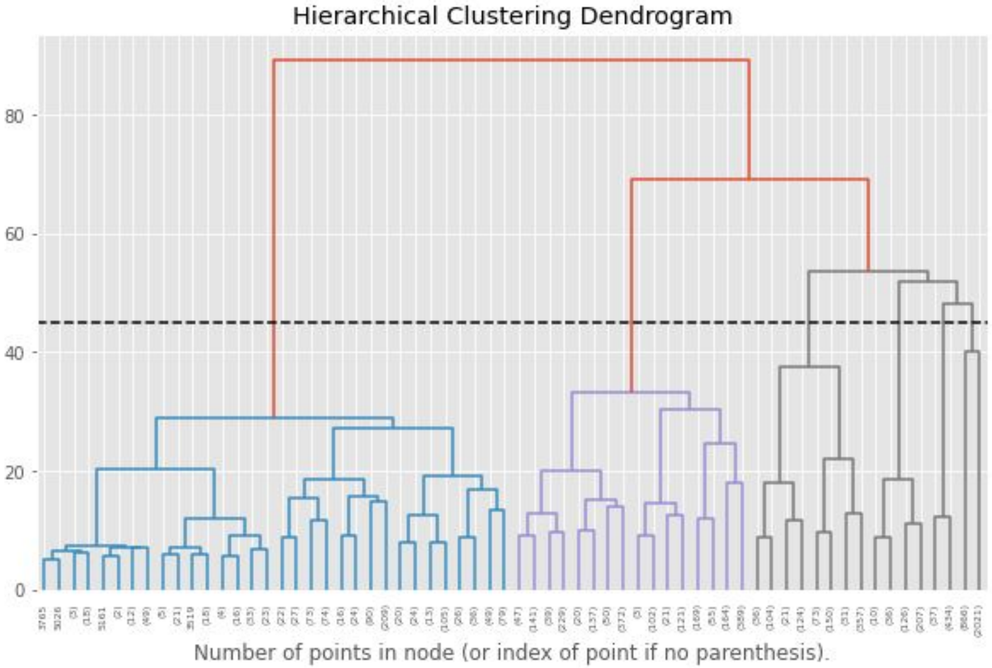
K mean Clusters

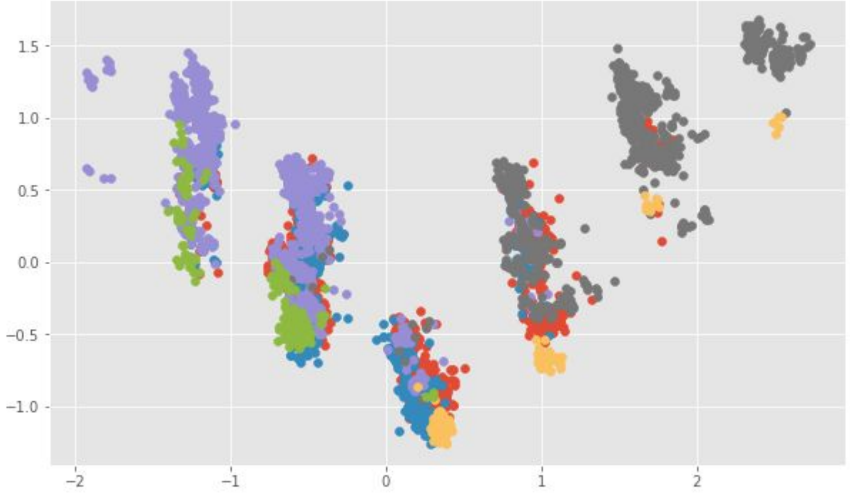
Agglomerative Clustering with CountVectorizer and TfidfVectorizer

For n_clusters = 6 The average silhouette_score is : 0.03082420009559484



Agglomerative Clustering with Dendrogram and CountVectorizer



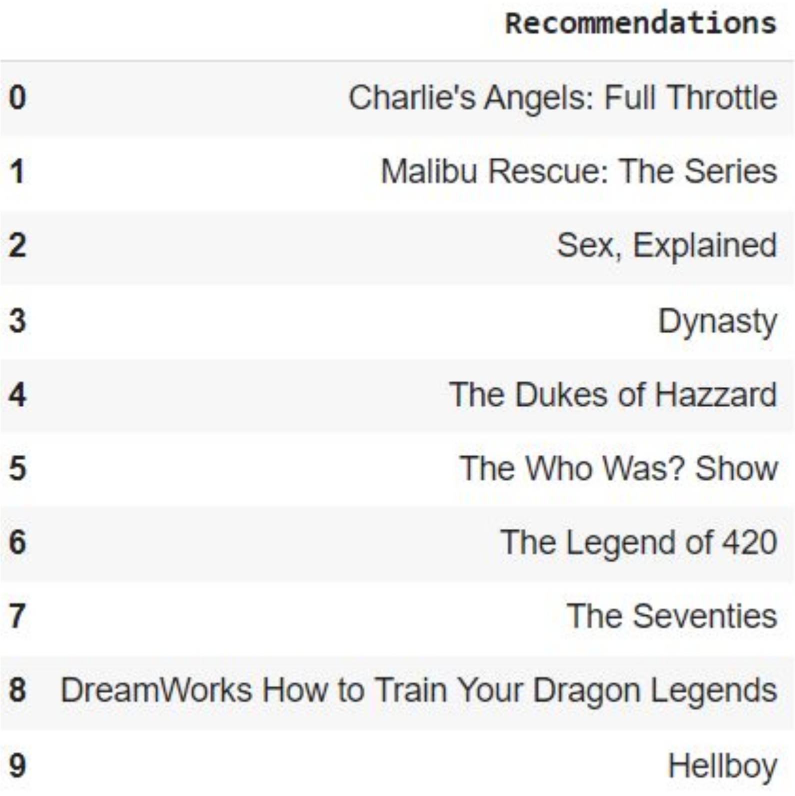


Silhouette Coefficient: 0.031

davies_bouldin_score 4.650

Clusters: 6

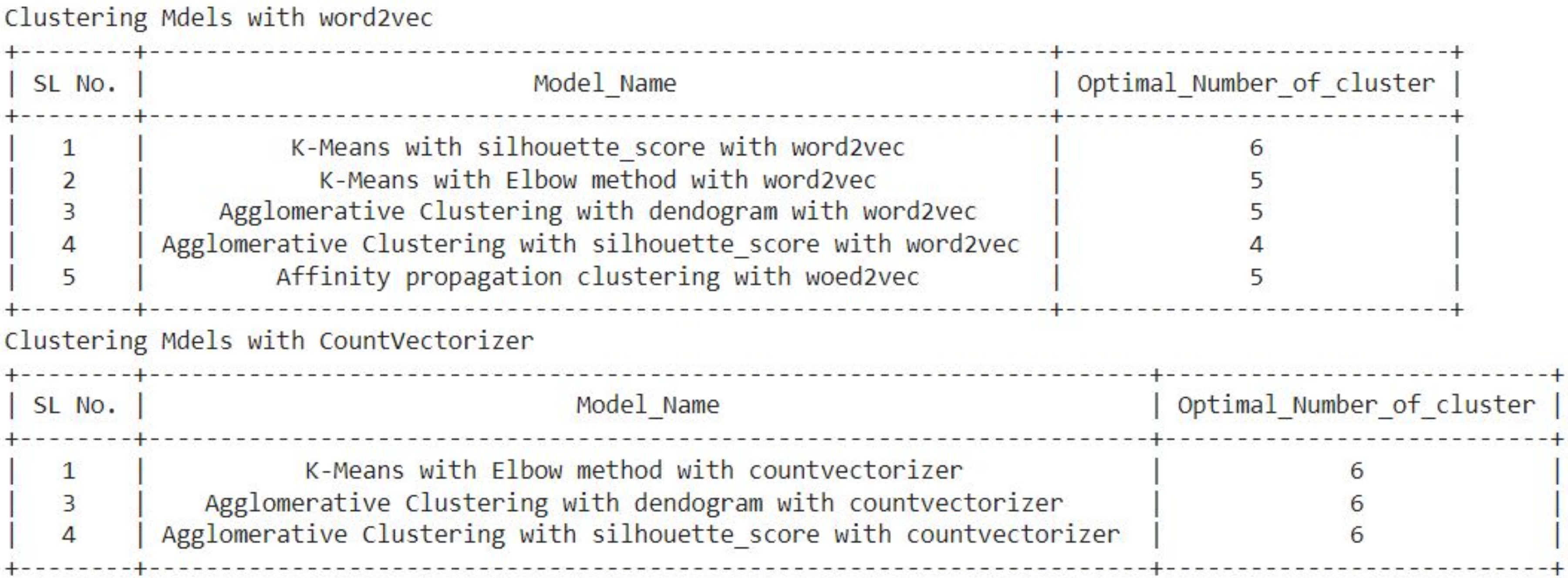
Recommendation System



	Recommendations
0	Charlie's Angels: Full Throttle
1	Malibu Rescue: The Series
2	Sex, Explained
3	Dynasty
4	The Dukes of Hazzard
5	The Who Was? Show
6	The Legend of 420
7	The Seventies
8	DreamWorks How to Train Your Dragon Legends
9	Hellboy

Cosine similarity **measures the similarity between two vectors of an inner product space**. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. It is often used to measure document similarity in text analysis

Summary



```
Clustering Mdels with word2vec
+--------+--------------------------------------------------------------+---------------------------+
| SL No. |                          Model_Name                          | Optimal_Number_of_cluster |
+--------+--------------------------------------------------------------+---------------------------+
|   1    |         K-Means with silhouette_score with word2vec          |             6             |
|   2    |           K-Means with Elbow method with word2vec            |             5             |
|   3    |    Agglomerative Clustering with dendogram with word2vec     |             5             |
|   4    | Agglomerative Clustering with silhouette_score with word2vec |             4             |
|   5    |        Affinity propagation clustering with woed2vec         |             5             |
+--------+--------------------------------------------------------------+---------------------------+
Clustering Mdels with CountVectorizer
+--------+---------------------------------------------------------------------+---------------------------+
| SL No. |                             Model_Name                              | Optimal_Number_of_cluster |
+--------+---------------------------------------------------------------------+---------------------------+
|   1    |           K-Means with Elbow method with countvectorizer            |             6             |
|   3    |    Agglomerative Clustering with dendogram with countvectorizer     |             6             |
|   4    | Agglomerative Clustering with silhouette_score with countvectorizer |             6             |
+--------+---------------------------------------------------------------------+---------------------------+
```

Conclusion

1. The attribute **'director','cast','country','date_added','rating' consists of missing values.** To tackle with missing values , we will replace 'country' and 'rating' missing values by most frequent entity that is 'United States' and 'TV-MA' respectively. missing values in 'cast' by 'unknown'. There are around **30.68 % values are missing in 'director'**, hence we decide to drop it.
2. **69% of the content available on Netflix are movies; the remaining 31% are TV Shows.Netflix has 5377 movies**, which is more than double the quantity of TV shows. In recents year more number of TV Shows are released as compared to Movies on Netflix. Less number of TV shows and Movies were released in 2020-2021 due to coronavirus pandemic. **Most of the Movies/TV Shows were added in the month of December and January.**
3. **Number of Movies added on Netflix are more as compared to TV Shows throughout the year. In recent few year more number of TV Shows were added on NetFlix as compared to Movies , We can say Netflix is more focusing on TV Shows than Movies.**
4. **United States, India, United Kingdom, Japan, South Korea, Canada, Spain, France,Egypt and Turkey are the Top 10 countries** which produces most of the content on Netflix. United States produced most of the content on Netflix also number of movies released are more than TV Shows in United States. In India, Canada, Spain, France, Egypt and Turkey , Most of the content on Netflix is Movies. United Kingdom has almost equal production of Movies and TV Shows. In Japan and South Korea, Number of TV Shows are available on Netflix on large scale.

Conclusion

1. It is observed that, in each category, Quantity of Movies is more than the Quantity of TV Shows.**The Availability of the Adult Content is more on Netflix and Least for the Kids**.
2. Popular Movies ratings are **TV-MA, TV-14, R, TV-PG, PG-14 and PG**. It is observed that Adults and Teens are mostly active on Netflix. Popular TV Shows ratings are **TV-MA, TV-14, R, TV-PG, PG-14 and PG**.
3. **Top 5 Genres in 'TV Shows' are Kid's TV, TV Dramas ,TV Crime Shows, TV Comedies, TV Romantic. Top 5 Genres in 'Movies' are Documentaries, Stand up Comedy, Dramas and International Movies, Comedies and Independent Movies.**It is observed that **1608 TV Shows has only one season**. The count of longest running TV Shows is very less.
4. Famous Actors on Netflix based on the Frequency of their occurrence on screen are Anupam Kher, Takahiro Sakurai Shah Rukh Khan, Om Puri and Boman Irani and so on. Most of the **Movies/TV Shows has duration around 100 min.**
5. United States is producing maximum International TV Shows, TV Dramas, Sci-fi and Fantasy TV shows, International Movies. **India, UK, Spain ,Egypt,Mexico and Turkey is having most of the Content as Dramas and International Movies.**
6. **It is observed that content available for kids is less as compared to other categories. Content available for Adults is more in almost every country except India. In India, Most of the content is available for Teens. Netflix should focus on Teen and Adult Contents to generate maximum revenue. Spain and Mexico is producing highest Adult Content on Netflix almost 84% and 77% respectively.**

Conclusion

Clustering with Word2vec

1. K-Means with 0.6092 silhouette_score with word2vec has optimum number of clusters as 6

2. K-Means with Elbow method with word2vec has 5 optimum clusters.

3.Agglomerative Clustering with dendrogram with word2vec has 5 optimum clusters

4.Agglomerative Clustering with 0.53 silhouette_score with word2vec gives 4 clusters

5. Affinity propagation clustering with word2vec has 5 optimum clusters

Clustering with CountVectorizer

1. It is observed that , after using CountVectorizer and tfidfVectorizer, we get the less silhouette_score as 0.032

2. Hence we can say word2vec word embedding method is more suitable for our model.

Winner Model: K-Means with word2vec with 6 optimum clusters with 0.6092 silhouette_score

References:

1. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html
2. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html
3. https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html
4. https://help.netflix.com/en/node/2064
5. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html
6. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AffinityPropagation.html

Thank You