Capstone Project-4

Project Title

NETFLIX MOVIES & TV SHOWS CLUSTERING

BY

Anurag Taiskar

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.
- Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

Dataset Top Row -

```
# Dataset First Look
nmc_df.head(10)
```

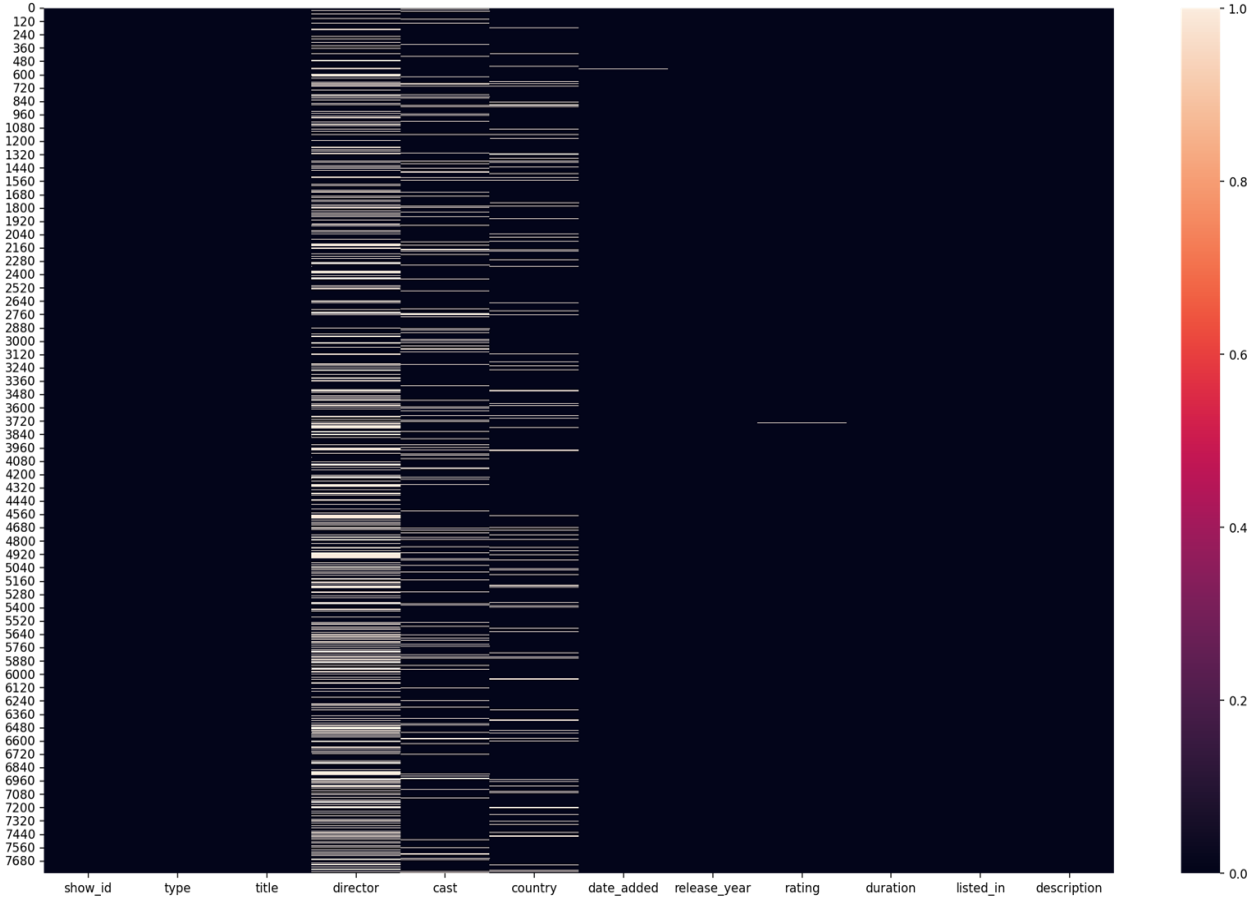
	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	TV Show	3%	NaN	João Miguel, Bianca Comparato, Michel Gomes, R...	Brazil	August 14, 2020	2020	TV-MA	4 Seasons	International TV Shows, TV Dramas, TV Sci-Fi &...	In a future where the elite inhabit an island ...
1	s2	Movie	7:19	Jorge Michel Grau	Demián Bichir, Héctor Bonilla, Oscar Serrano, ...	Mexico	December 23, 2016	2016	TV-MA	93 min	Dramas, International Movies	After a devastating earthquake hits Mexico Cit...
2	s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella Chung, Henley Hii, Lawrence ...	Singapore	December 20, 2018	2011	R	78 min	Horror Movies, International Movies	When an army recruit is found dead, his fellow...
3	s4	Movie	9	Shane Acker	Elijah Wood, John C. Reilly, Jennifer Connelly...	United States	November 16, 2017	2009	PG-13	80 min	Action & Adventure, Independent Movies, Sci-Fi...	In a postapocalyptic world, rag-doll robots hi...
4	s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin Spacey, Kate Bosworth, Aar...	United States	January 1, 2020	2008	PG-13	123 min	Dramas	A brilliant group of students become card-coun...
5	s6	TV Show	46	Serdar Akar	Erdal Beşikçioğlu, Yasemin Allen, Melis Birkan...	Turkey	July 1, 2017	2016	TV-MA	1 Season	International TV Shows, TV Dramas, TV Mysteries	A genetics professor experiments with a treatm...
6	s7	Movie	122	Yasir Al Yasiri	Amina Khalil, Ahmed Dawood, Tarek Lotfy, Ahmed...	Egypt	June 1, 2020	2019	TV-MA	95 min	Horror Movies, International Movies	After an awful accident, a couple admitted to ...
7	s8	Movie	187	Kevin Reynolds	Samuel L. Jackson, John Heard, Kelly Rowan, Cl...	United States	November 1, 2019	1997	R	119 min	Dramas	After one of his high school students attacks ...
8	s9	Movie	706	Shravan Kumar	Divya Dutta, Atul Kulkarni, Mohan Agashe, Anup...	India	April 1, 2019	2019	TV-14	118 min	Horror Movies, International Movies	When a doctor goes missing, his psychiatrist w...
9	s10	Movie	1920	Vikram Bhatt	Rajneesh Duggal, Adah Sharma, Indraneil Sengup...	India	December 15, 2017	2008	TV-MA	143 min	Horror Movies, International Movies, Thrillers	An architect and his wife move into a castle t...

Dataset Information-

```
# Dataset Info
nmc_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
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
```

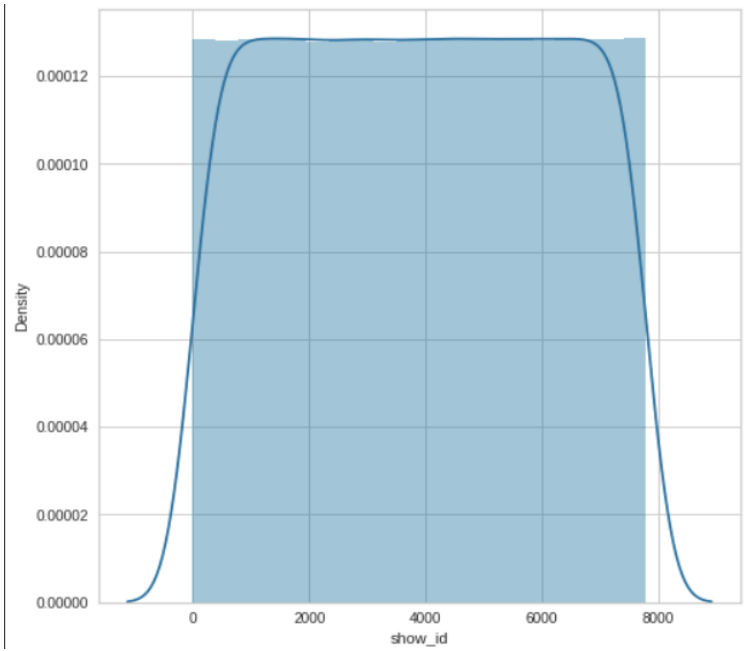
Handling Null Values and feature engineering



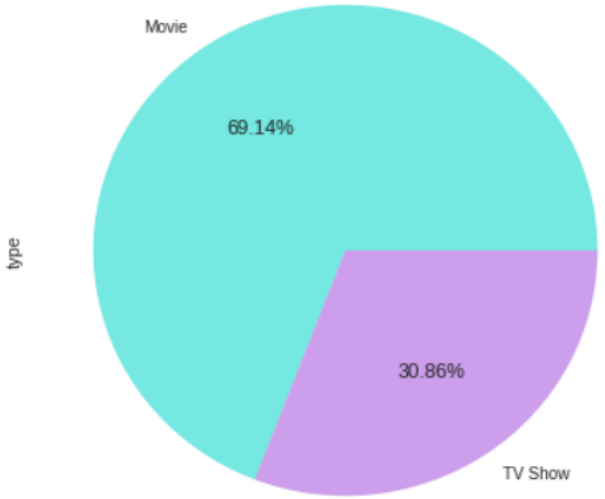
- **We checked for null values after loading the dataset and removed the null values, as well as some unnecessary columns.**

EDA (univariate)

Show id column

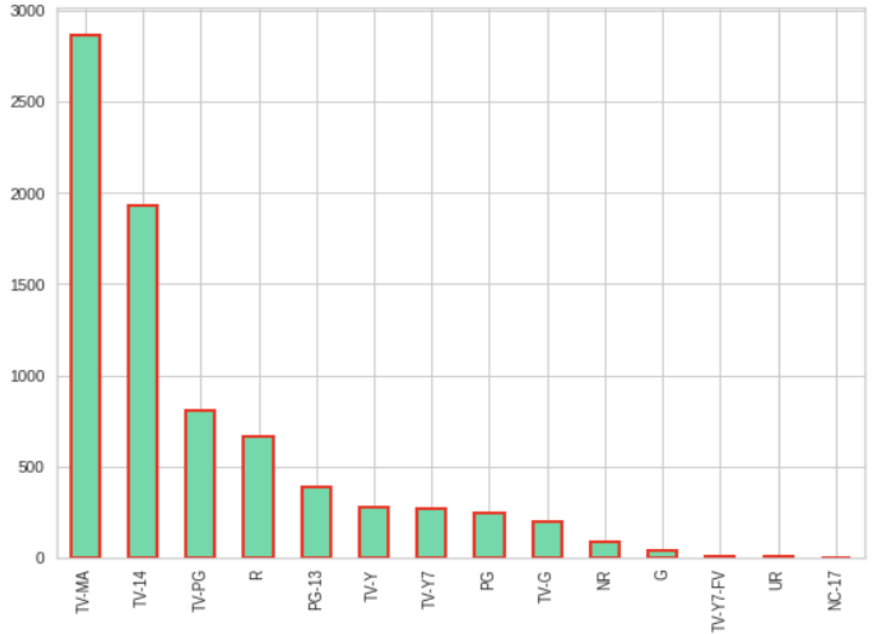


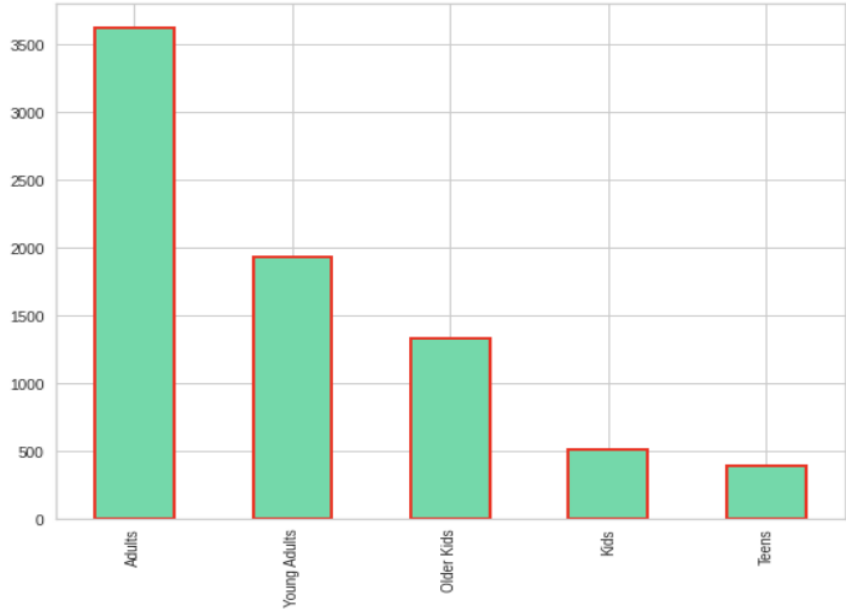
Type column



EDA (univariate)

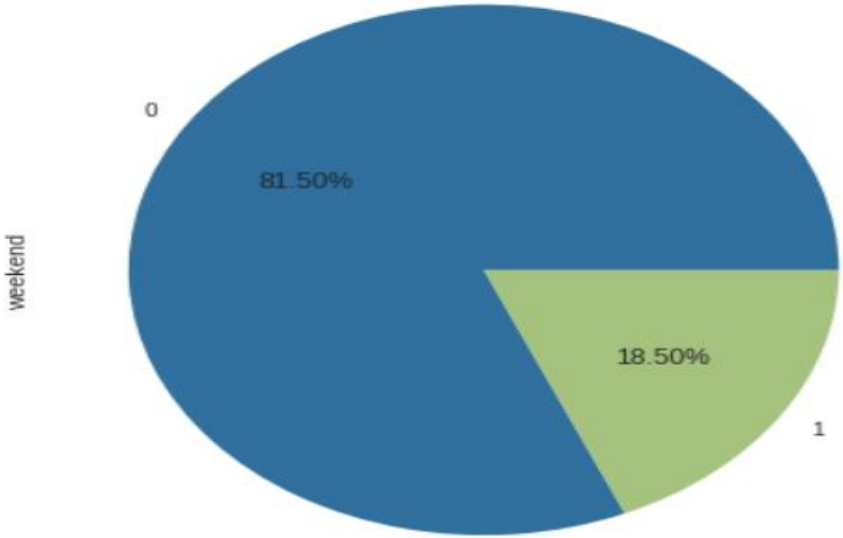
Rating Column





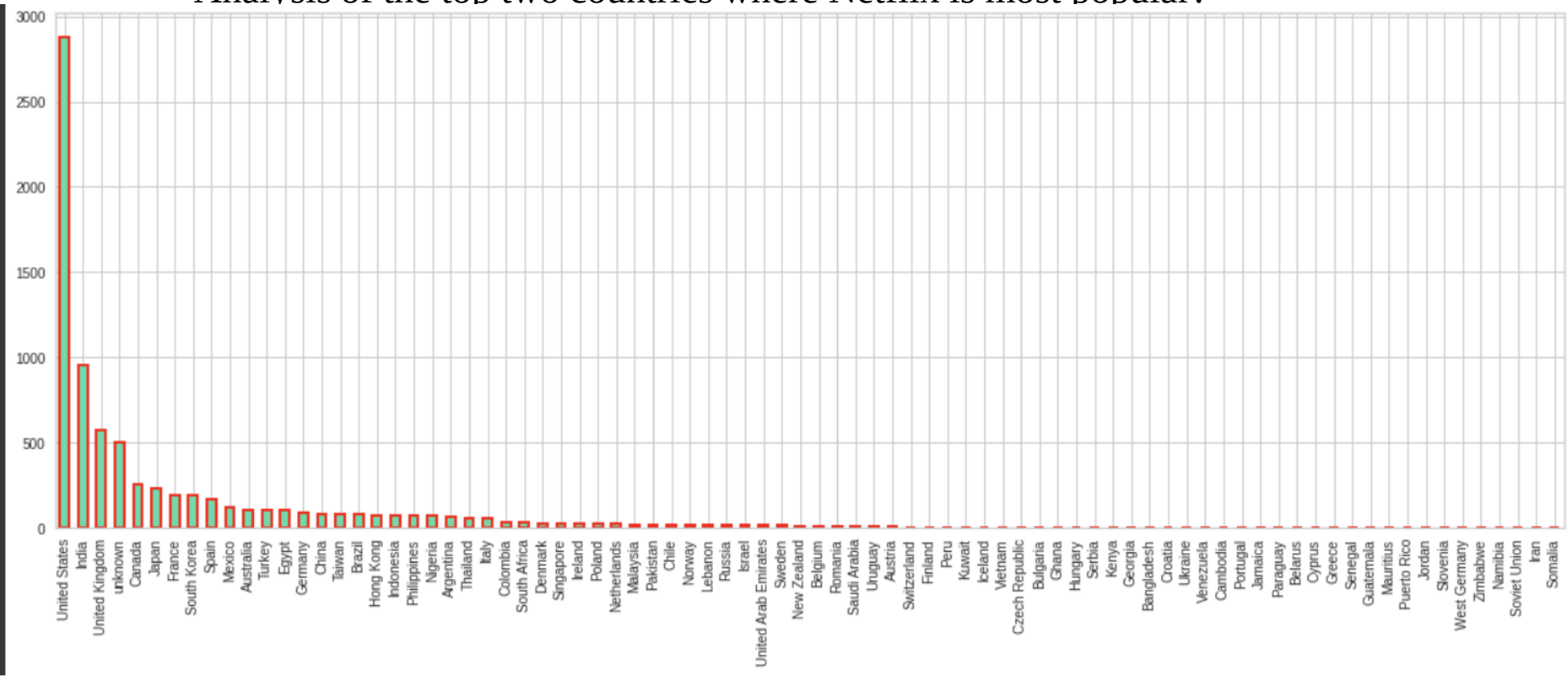
EDA (univariate)

Weekend Column



EDA cont.

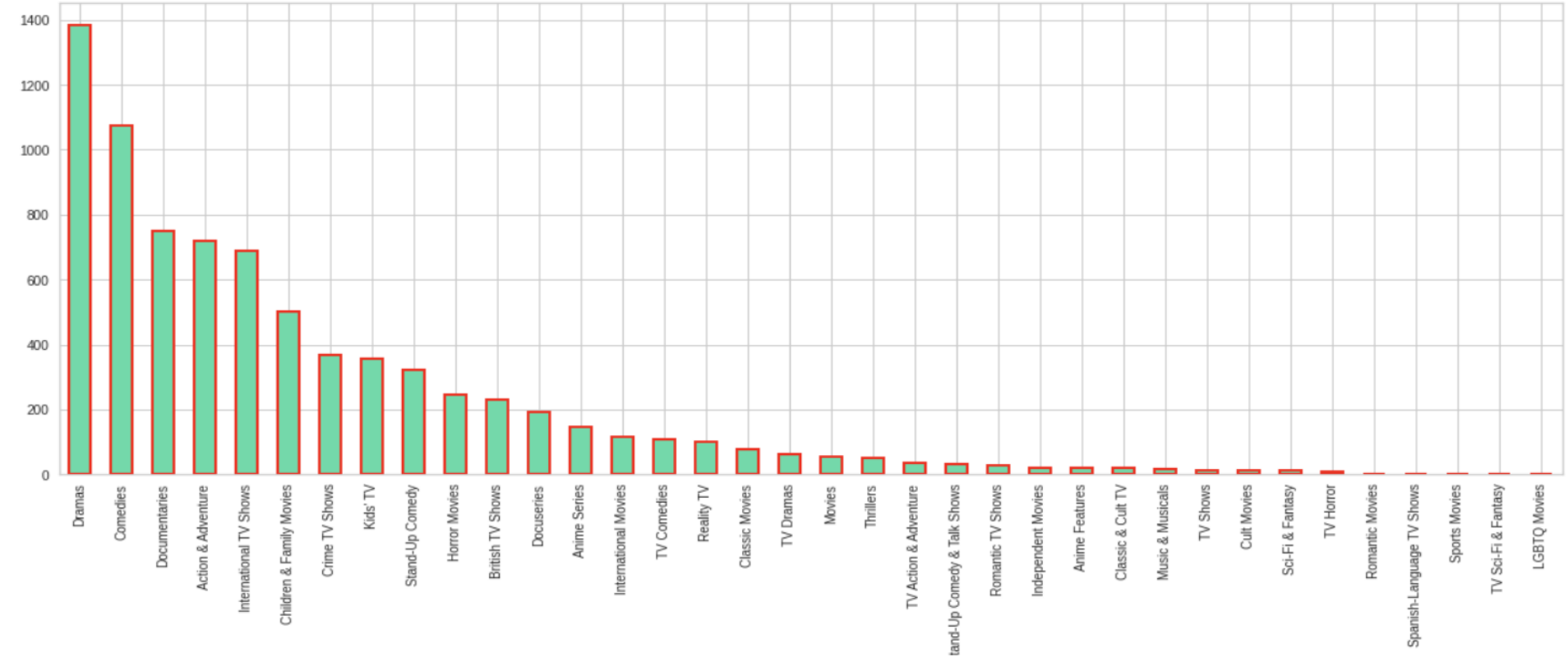
Analysis of the top two countries where Netflix is most popular?



- **As can be seen in the plot above, the United States and India are the two countries where Netflix is most popular.**

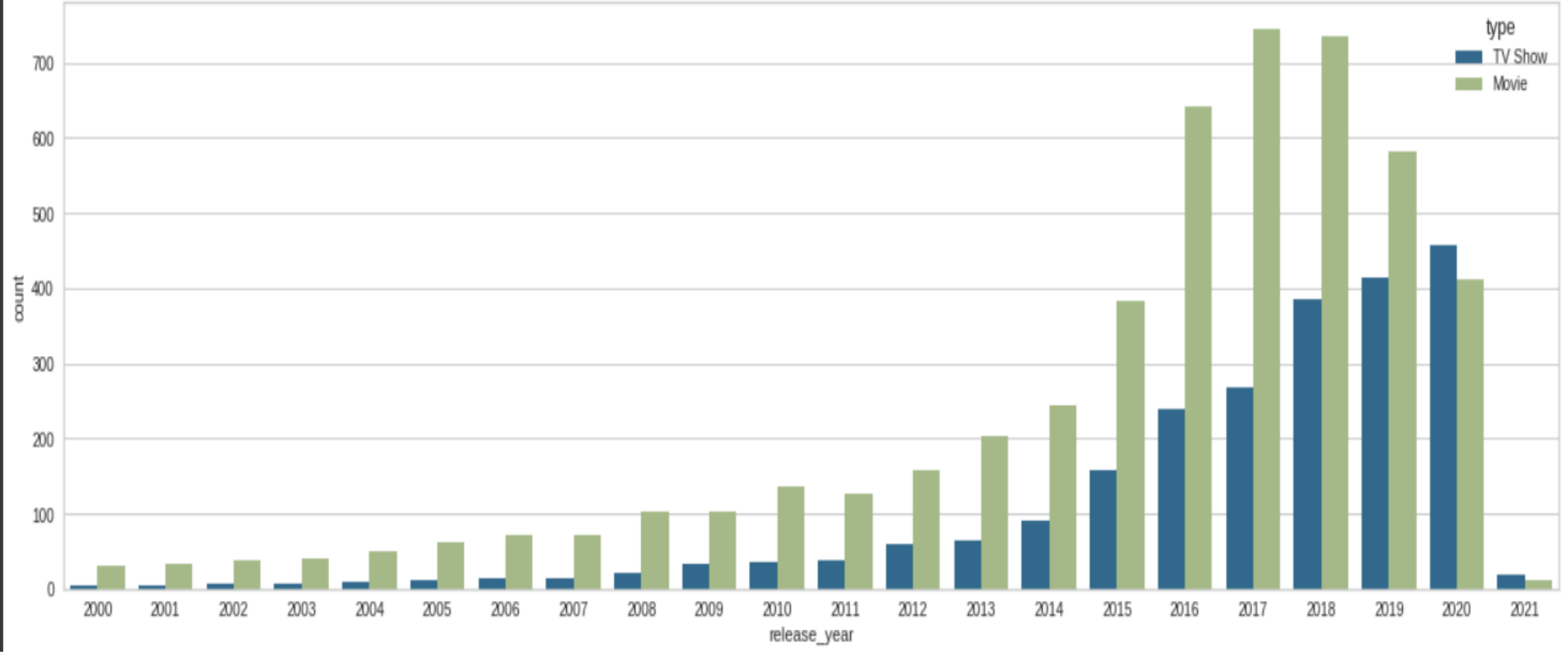
EDA (univariate)

Genre Column



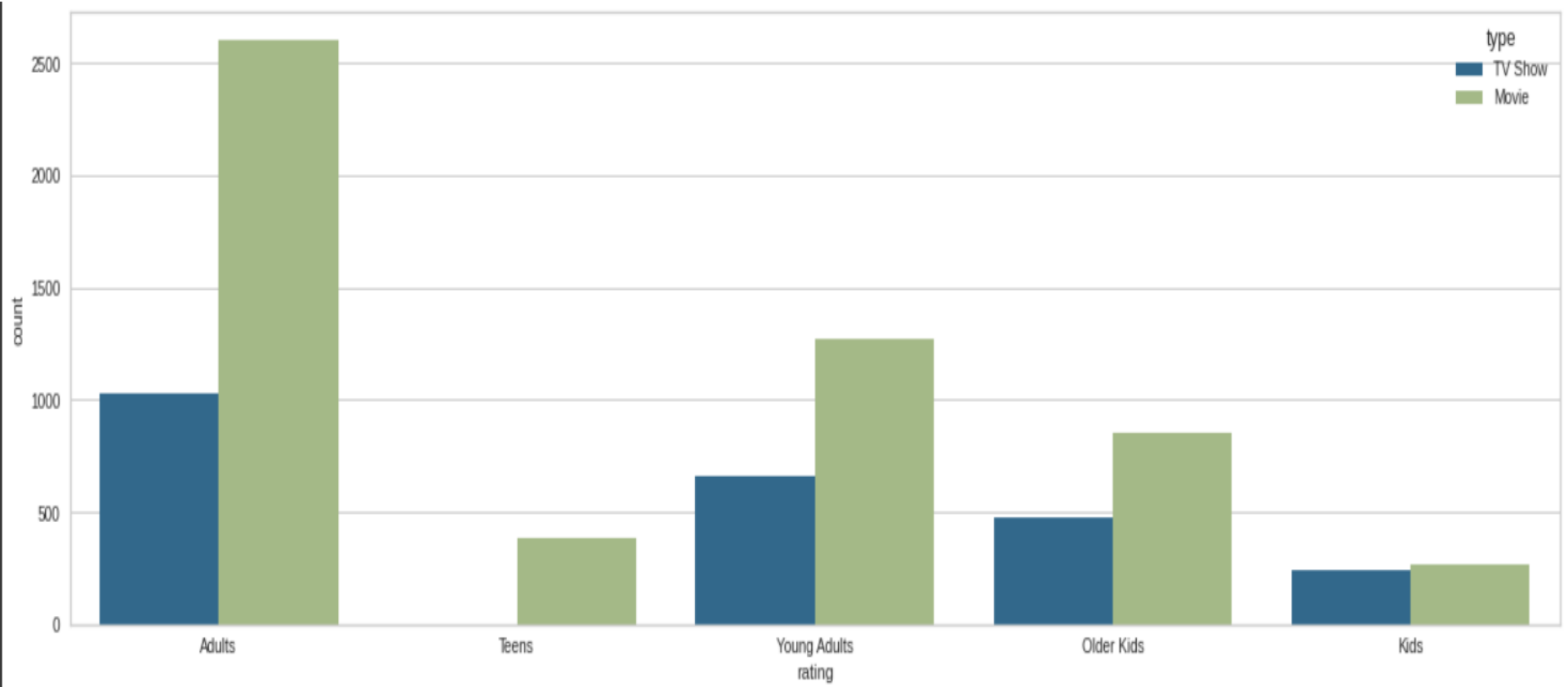
EDA (biavariate)

Year based analysis of movies and tv shows



EDA (biavariate)

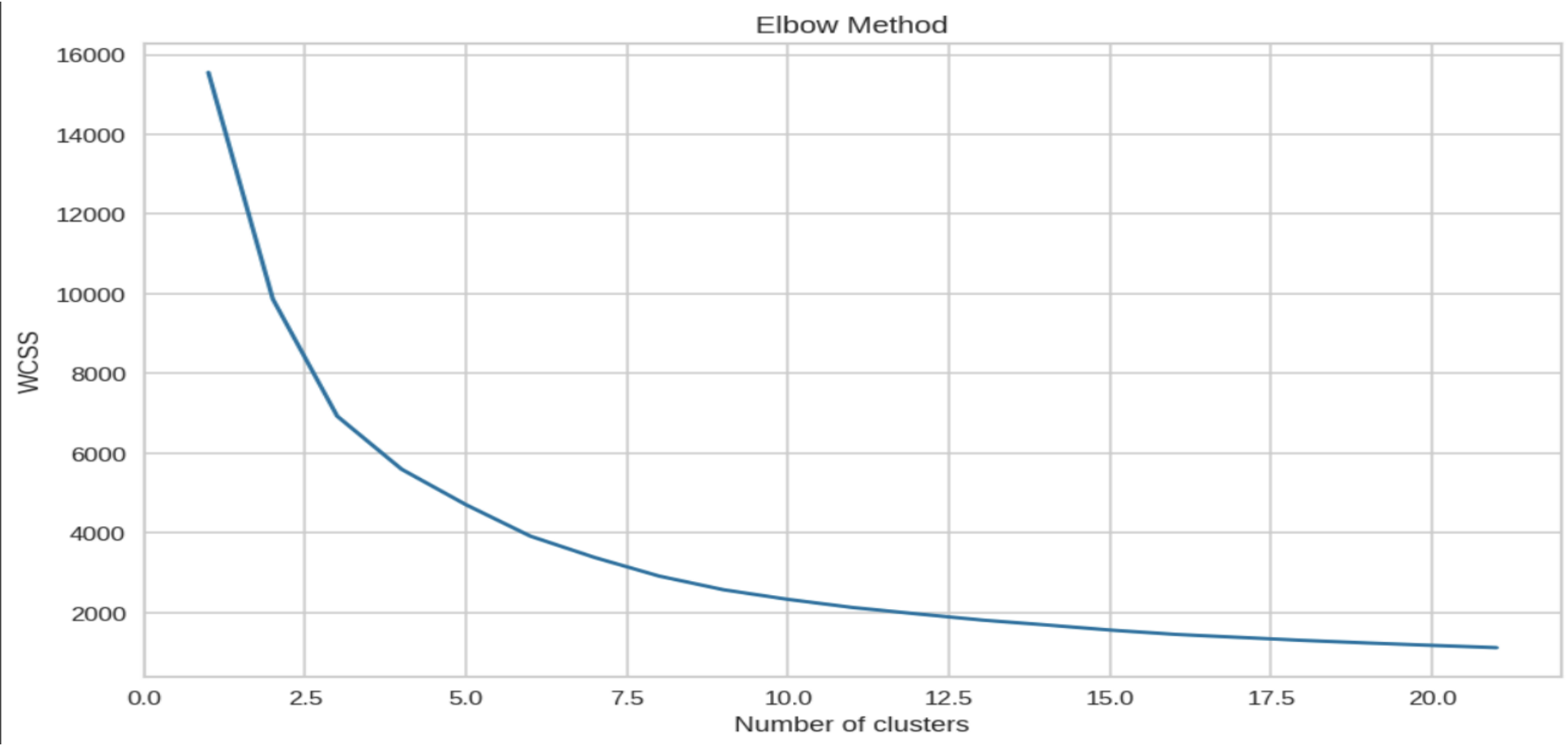
Rating based analysis of movies and tv shows



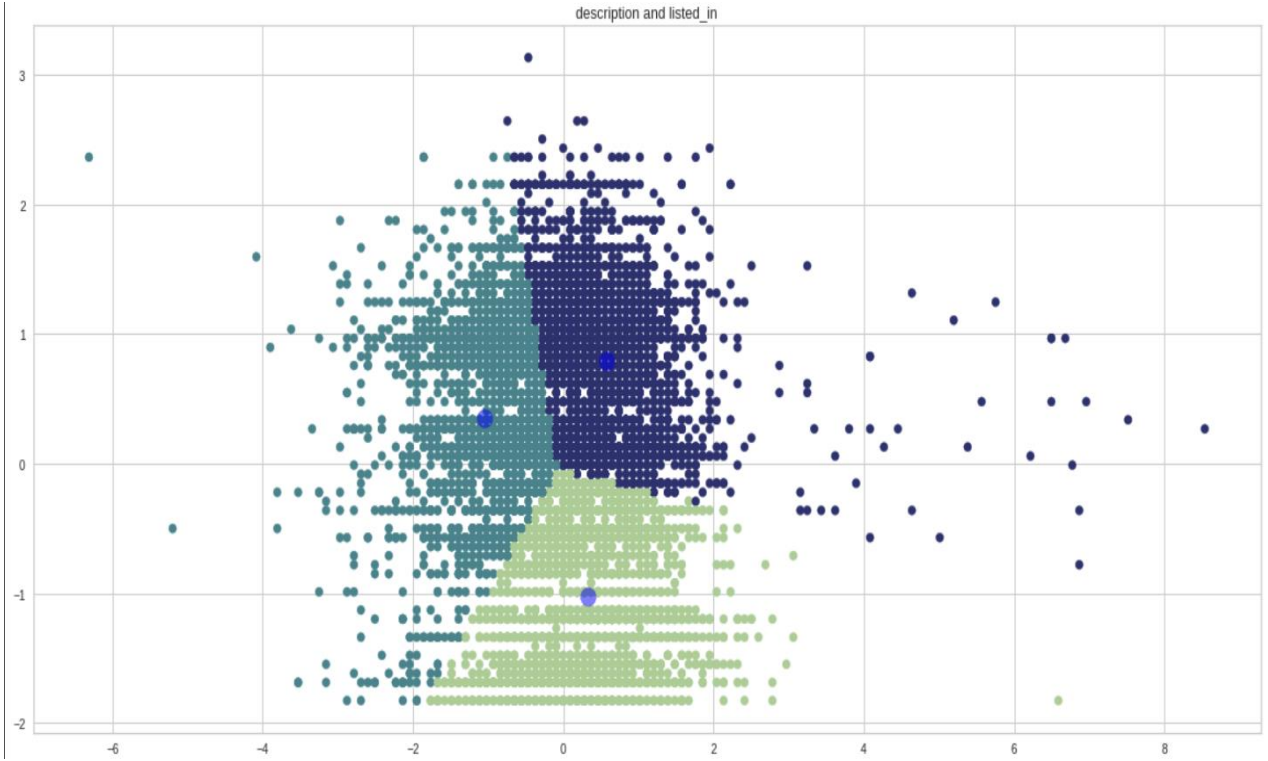
Implementation Model -

1 K-Means Clustering(Elbow method)
2 Hierarchical Clustering(Dendrogram & Agglomerative)
3 Silhouette Score

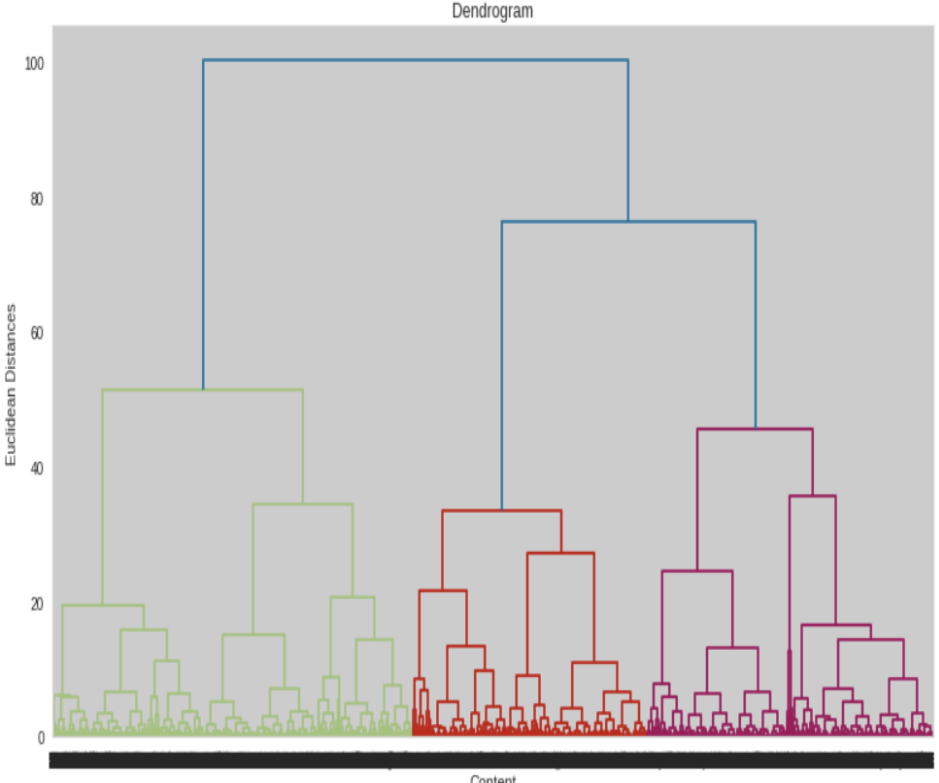
Finding Number of Clusters

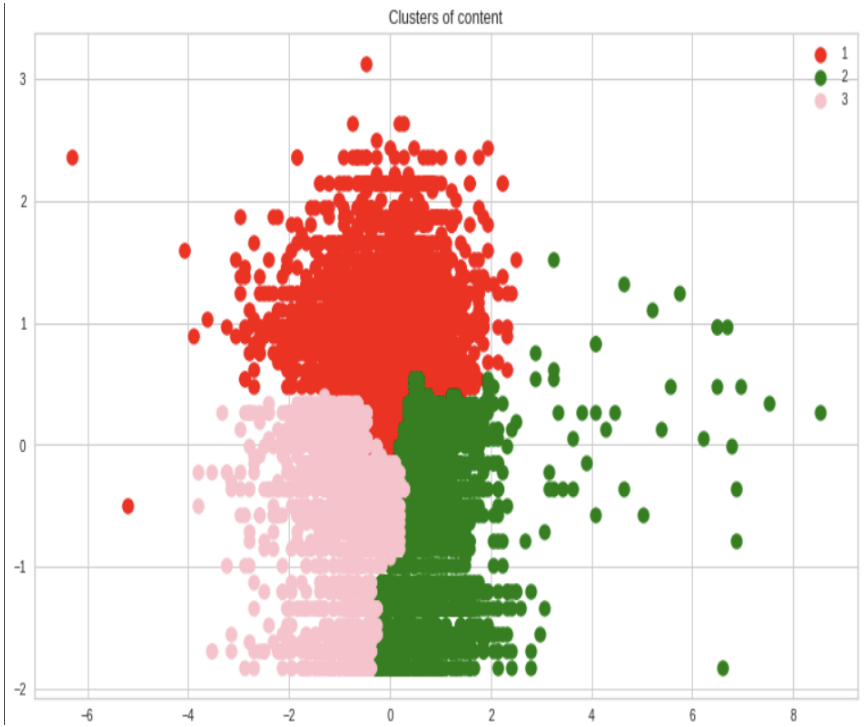


K-Means Clustering(Elbow method)

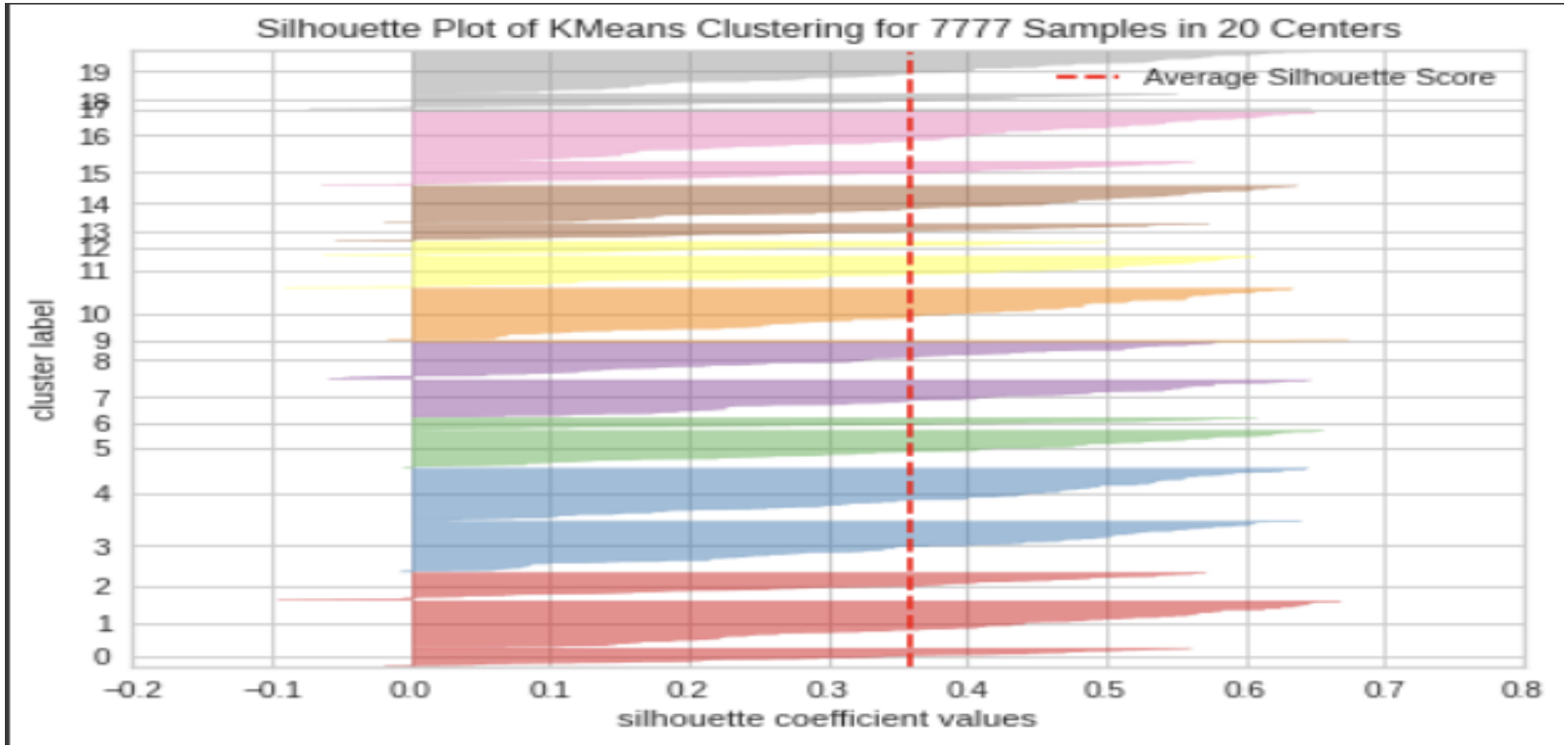


Hierarchical Clustering(Dendrogram & Agglomerative)





Silhouette Score



Conclusion

- This Netflix Dataset is a Unsupervised machine leaning dataset
- Data set contains 7787 rows and 12 columns
- Director features contains large number of missing values (more then 30%) so we decided to drop this director attribute.
- We have two types of content TV shows and Movies (30.86% contains TV shows and 69.14% contains Movies
- The growth rate of the content on Netflix is exponential
- Mostly movies are for adults sections
- US and India produce more than 70% of the content
- By analysing the content added over years we get to know that in recent years netflix is focusing movies than TV shows (movies is increased by 80% and TV shows is increased by 73% compare to 2016 data)
- By applying the silhouette score method for n range clusters on dataset we got best score which is 0.356 for clusters = 3, it means content explained well on their own clusters.
- Speaking about other different cluster methods, K mean, hierarchical, agglomerative clustering on data, we got the best cluster arrangements.
- 3 is the best cluster for this dataset

Thank you