Capstone Project

NETFLIX MOVIES AND TV SHOWS CLUSTERING

By : Sanjay Verma
Poonam Shevkar

Flow of the Presentation

- Introduction
- Problem Statement
- Data Preparation
- Exploratory Data Analysis
- Feature engineering
- Model Implementation
- Conclusion

Introduction

Netflix is an American subscription streaming service and production company. It is the one of the largest Platform which provides the collection of TV shows and movies, streaming via online means. Netflix must keep their content interesting that can hook users on the platform. That's why the recommendation system which provides valuable suggestions to users is essential.

❖ **Database:**

- Netflix Movies and TV Shows
- 7787 rows and 12 columns
- Data from last decade

Methodology:Unsupervised Machine Learning (Clustering)

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.

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

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

Data Preparation

- **Data processing:** At first phase checked for null Values and changed the datetime containing column in dataset.

- **EDA:** Exploratory analysis was done on the Features selected in the first phase.

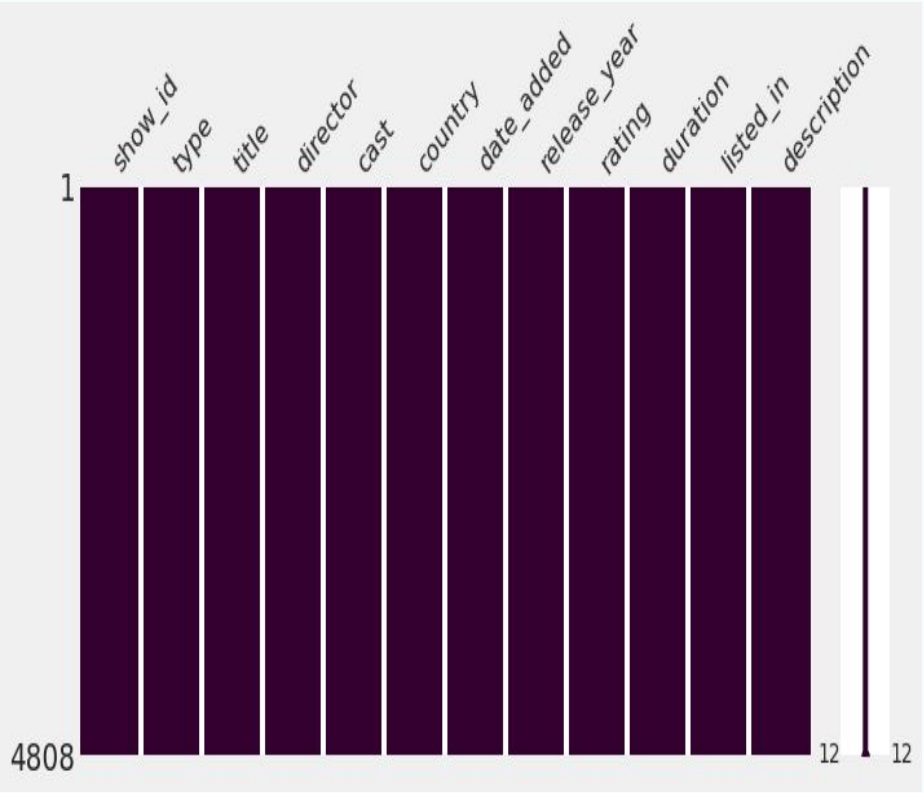
- **Feature engineering:** Unify some of the similar Types (genre) and Make a dictionary with matching Text based features that we are going to use in clustering.

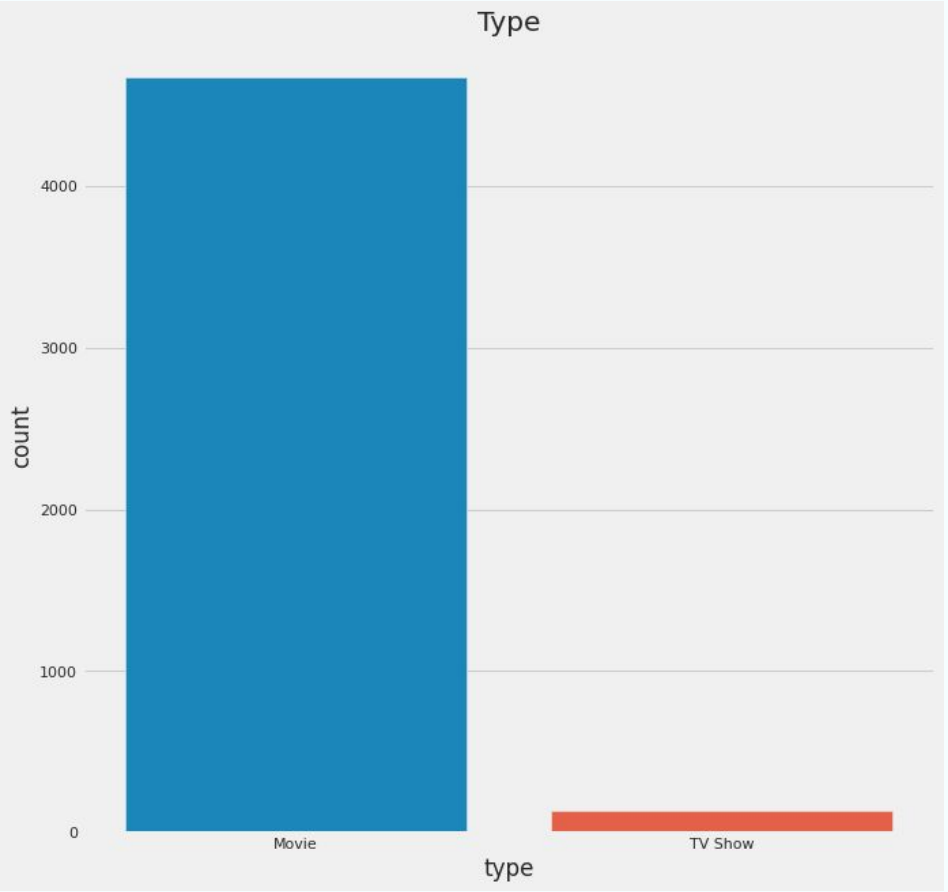
- **Prepare Dataframe:** Delete some features, we prepare new dataframe to feed the clustering algorithms.

- **Create a model:** Finally in this part created models based on clustering.

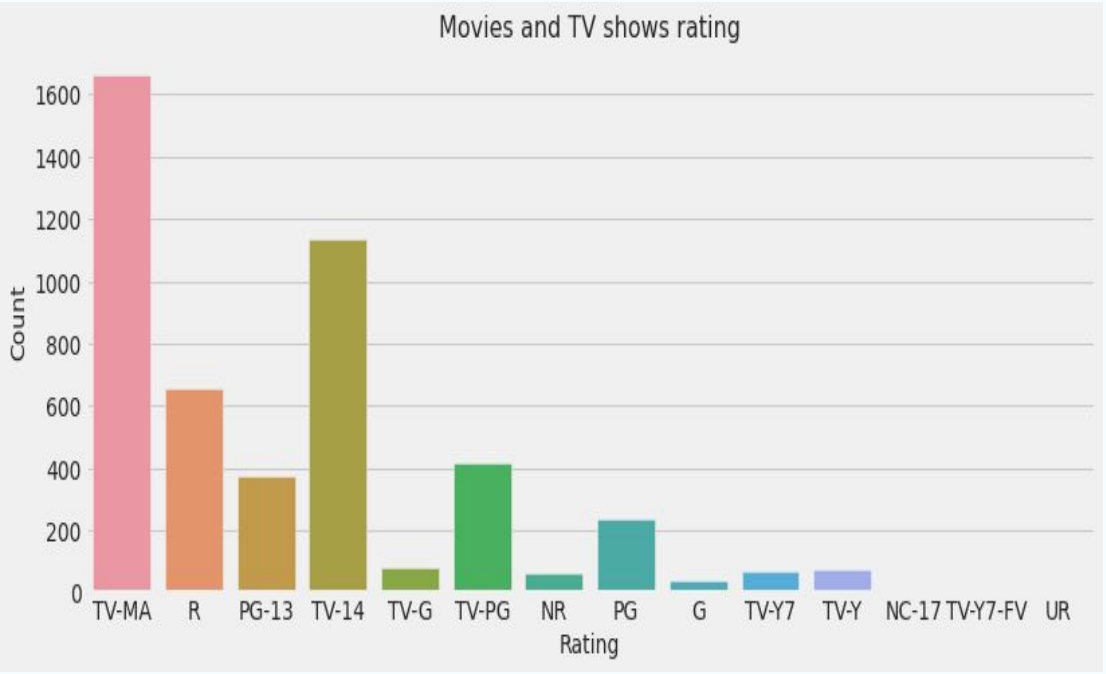
Exploratory Data Analysis

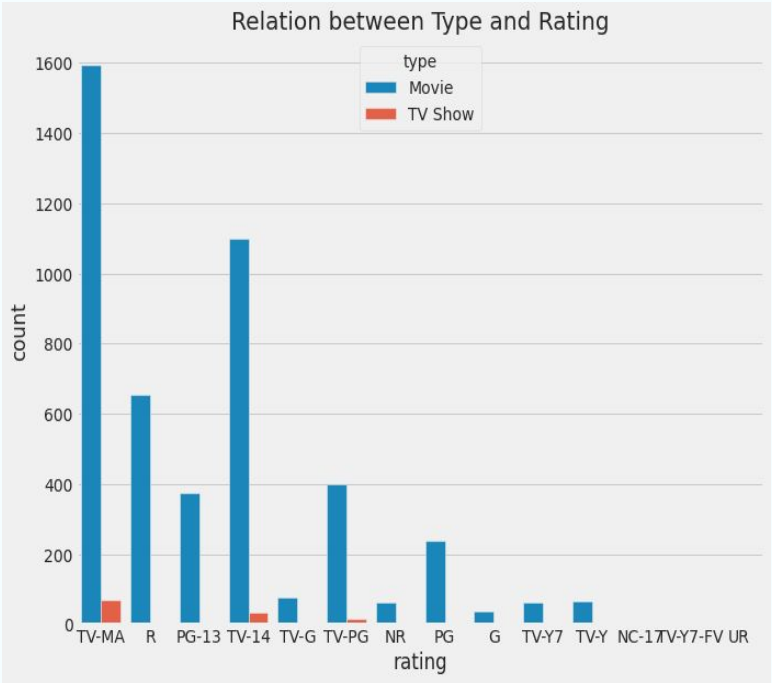
- ❖ Check Null Values
- ❖ Visualization on features





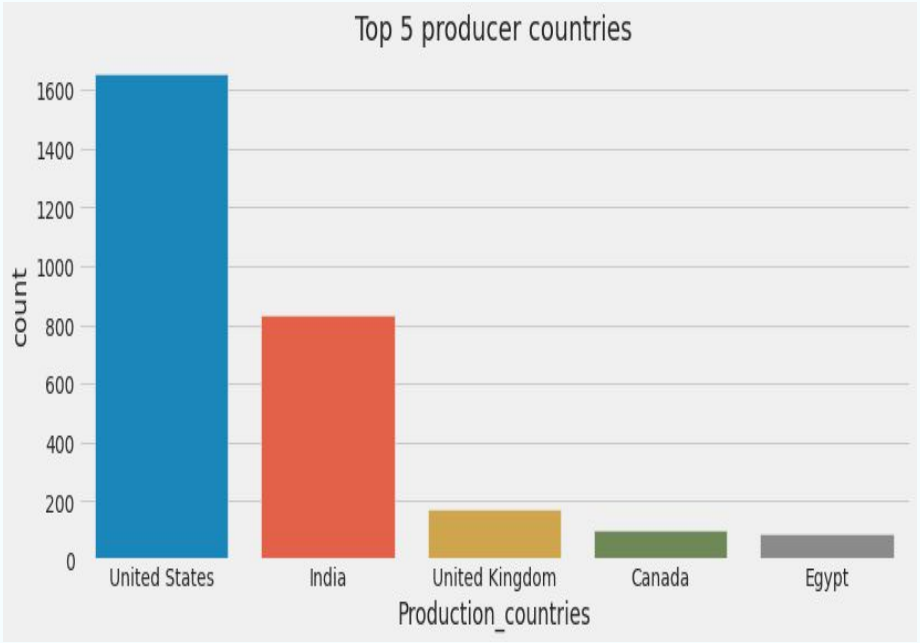
Exploratory Data Analysis





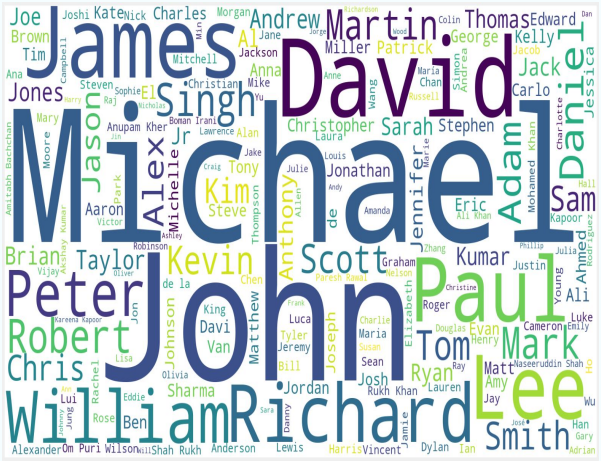
The audience prefers TV-MA and TV-14 shows more and the least preferred rating shows are Nc-17. Most of the content watched by the audience is for a mature audience. The TV-MA rating is a type of rating given by the TV parental guidelines to a television program.The second largest type of rating watched by the audience is TV-14 which is inappropriate for children younger than age 14. The conclusion is drawn here is most of the audience is of mature age.

Exploratory Data Analysis



Top 5 Producer Countries

Countries

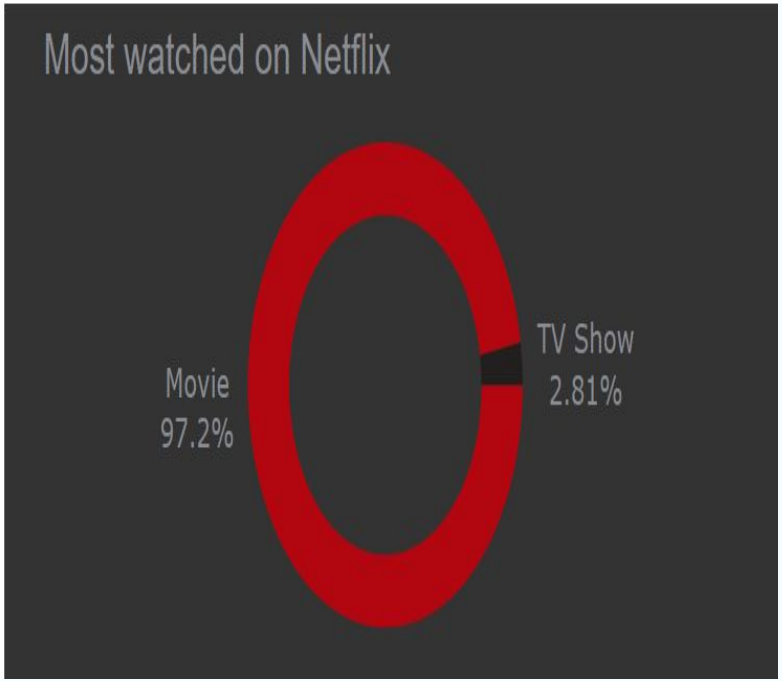


Cast in the Shows

Exploratory Data Analysis

Drama is the highest preferred show by the audience then comes the comedy show and action show, the least preferred show is of LGBTQ movies.

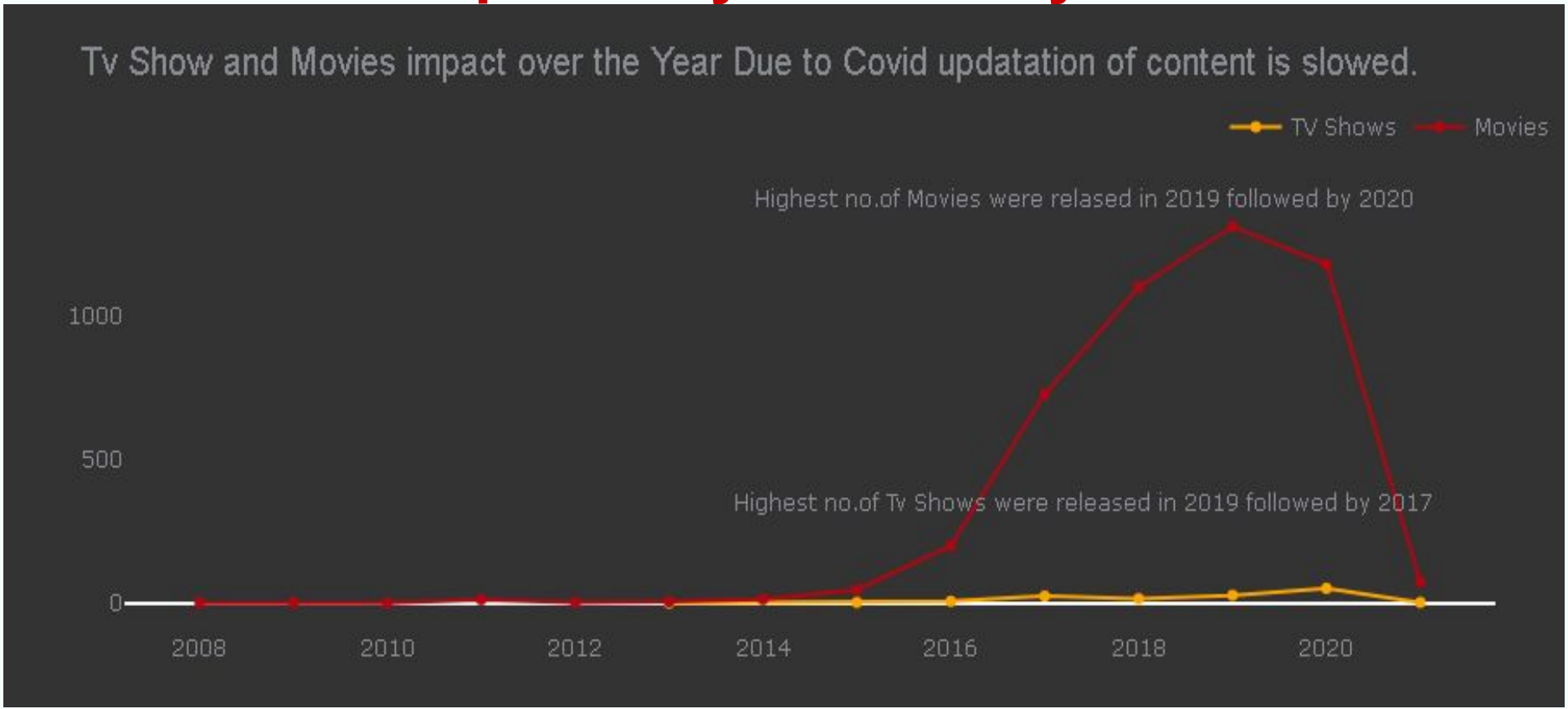
Exploratory Data Analysis



title	type	country		
#Alive	Movie	South Korea	27	2020-09-08
#AnneFrank - Parallel Stories	Movie	Italy	28	2020-07-01
#FriendButMarried	Movie	Indonesia	31	2020-05-21
#FriendButMarried 2	Movie	Indonesia	32	2020-06-28
#Roxy	Movie	Canada	34	2019-04-10
#Selfie	Movie	Romania	36	2019-06-01
#Selfie 69	Movie	Romania	37	2019-06-01
#realityhigh	Movie	United States	33	2017-09-08
1 Chance 2 Dance	Movie	United States	46	2017-07-01
1 Mile to You	Movie	United States	47	2017-07-07
10 Days in Sun City	Movie	South Africa, Nigeria	48	2019-10-18
10 jours en or	Movie	France	49	2017-07-01
10,000 B.C.	Movie	United States, South Africa	50	2019-06-01
100 Meters	Movie	Portugal, Spain	54	2017-03-10
100% Halal	Movie	Indonesia	57	2021-01-07
1000 Rupee Note	Movie	India	59	2016-12-01
12 ROUND GUN	Movie	United States	60	2019-03-14
122	Movie	Egypt	6	2020-06-01
13 Cameras	Movie	United States	62	2016-08-13

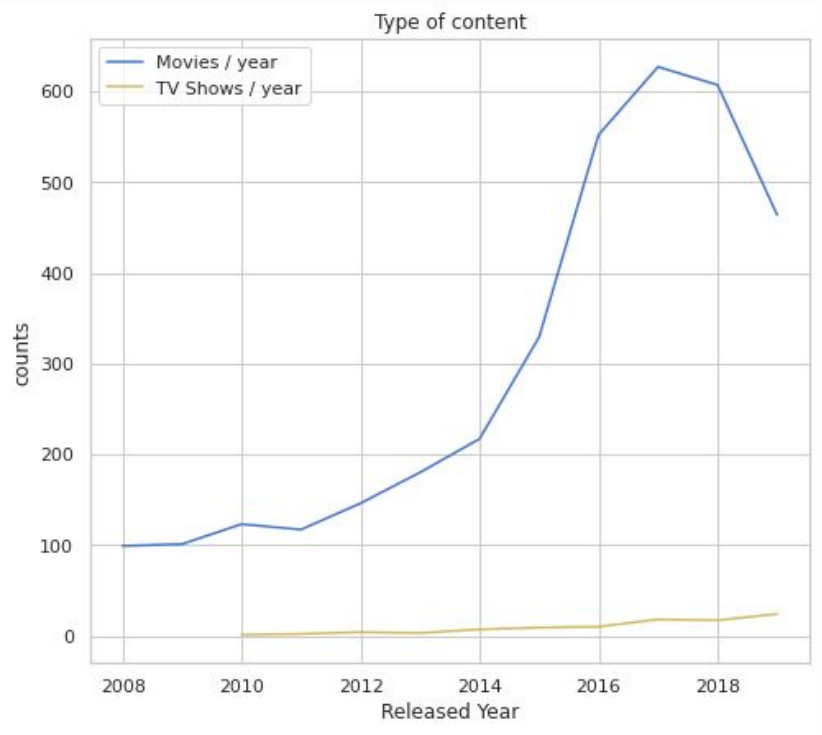
Here, we are understanding what type content is available in different countries.Because the quantity of movies outnumbers the number of TV series. It appears that movies are most widely available in various countries.

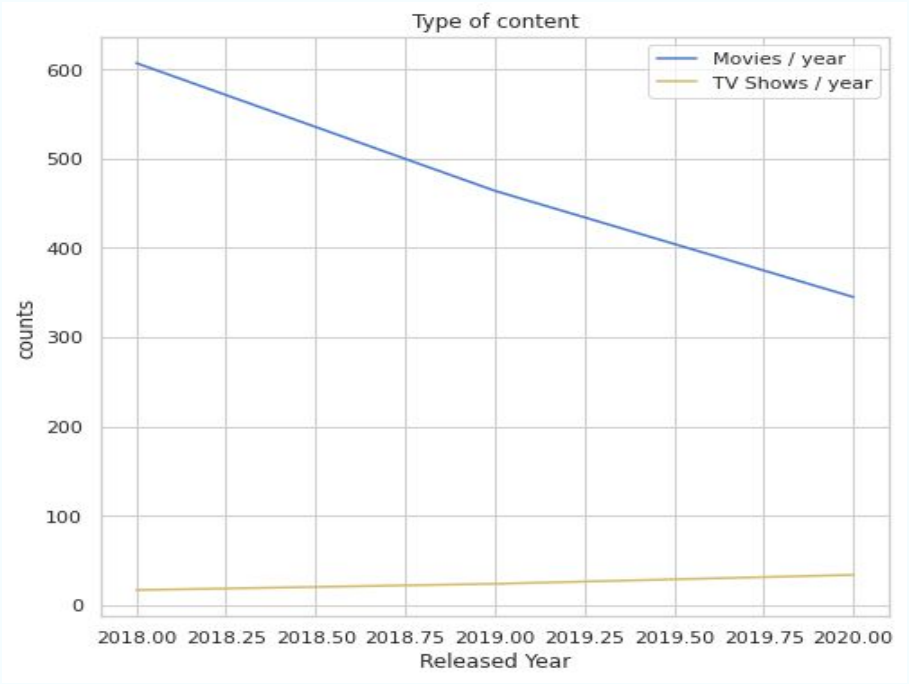
Exploratory Data Analysis



After the year 2019 covid came that badly affects Netflix for producing content. Movies have exponential growth from the start but due to covid, it is going downwards.

Hypothesis from Visualization of Type of content & Released Year-





Hypothesis from the data visualized-

1. According to the first graph, the number of TV shows launched in the previous few years is growing.

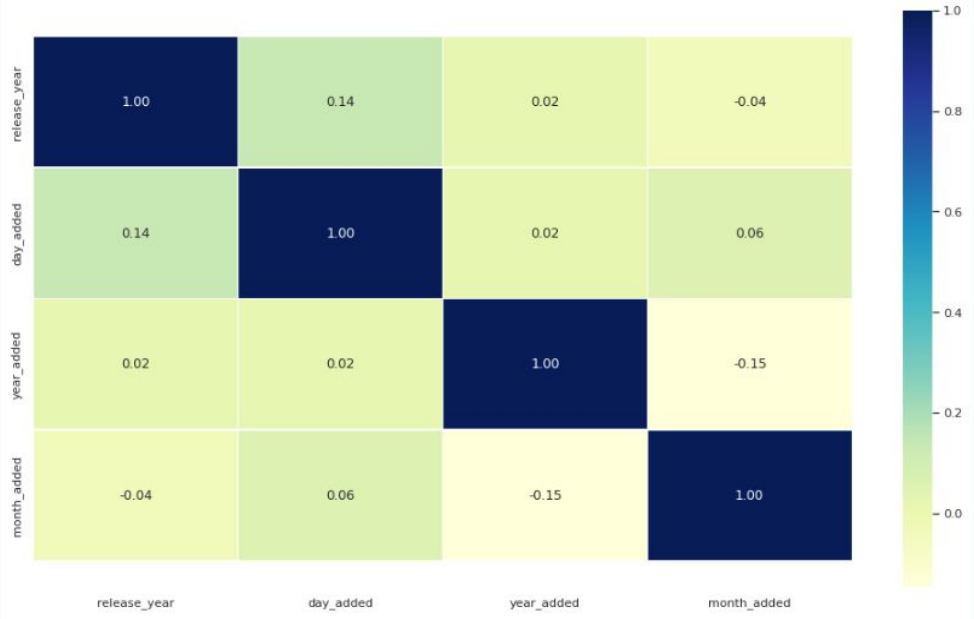
2. According to the second graph, the number of TV shows added to Netflix is stable.

Clustering Some feature based and correlation Matrix-

	type	director	country	release_year	genere	year_added
1	0	1635	239	2016	12	2016
2	0	1141	296	2011	13	2018
3	0	3074	440	2009	0	2017
4	0	2826	440	2008	12	2020
5	1	3050	357	2016	16	2017

Data engineering done based on text based feature by removing unnecessary columns

Correlation matrix



Data Preparation

➢ Feature engineering

➢ Select required columns

➢ Import required libraries for clustering

Importing required libraries for clustering

```
[ ]  import seaborn as sns
     import matplotlib.cm as cm
     from sklearn.preprocessing import StandardScaler
     from sklearn.decomposition import PCA
     from scipy.cluster.hierarchy import linkage, dendrogram
     from sklearn.datasets import make_blobs
     from sklearn.cluster import KMeans
     import numpy as np
     from sklearn.metrics import silhouette_samples, silhouette_score
```

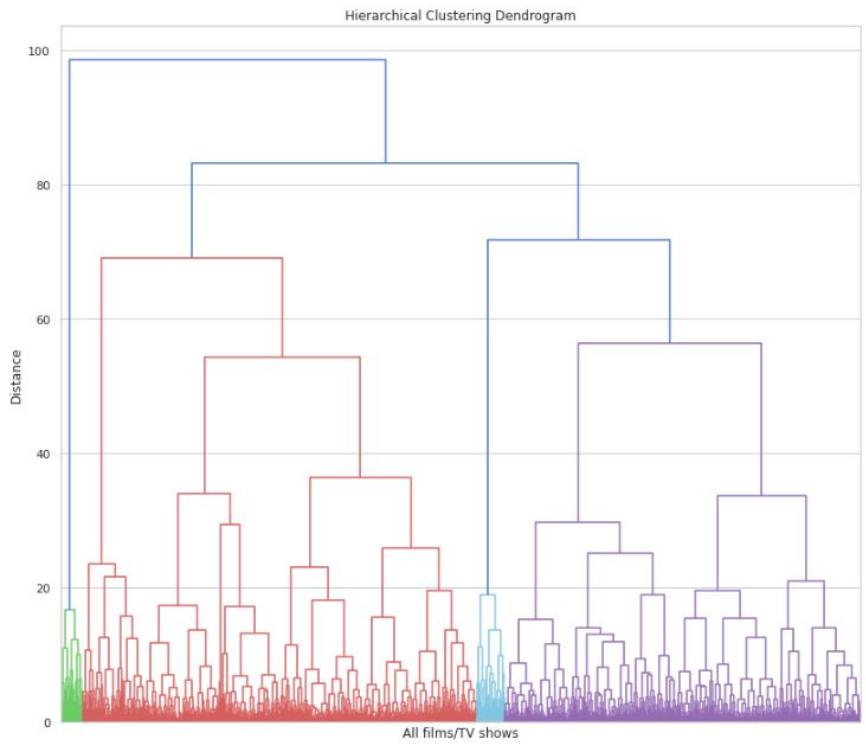
Scaling the data

```
[ ]  # transform the data using StandardScaler
     #We transform the data

     netflix_standarized = pd.DataFrame(StandardScaler().fit_transform(netflix),columns = netflix.columns)

     #Perform a PCA to visualize clusters

     pca=PCA(n_components=2)
     netflix_pca=pd.DataFrame(pca.fit_transform(netflix_standarized))
```

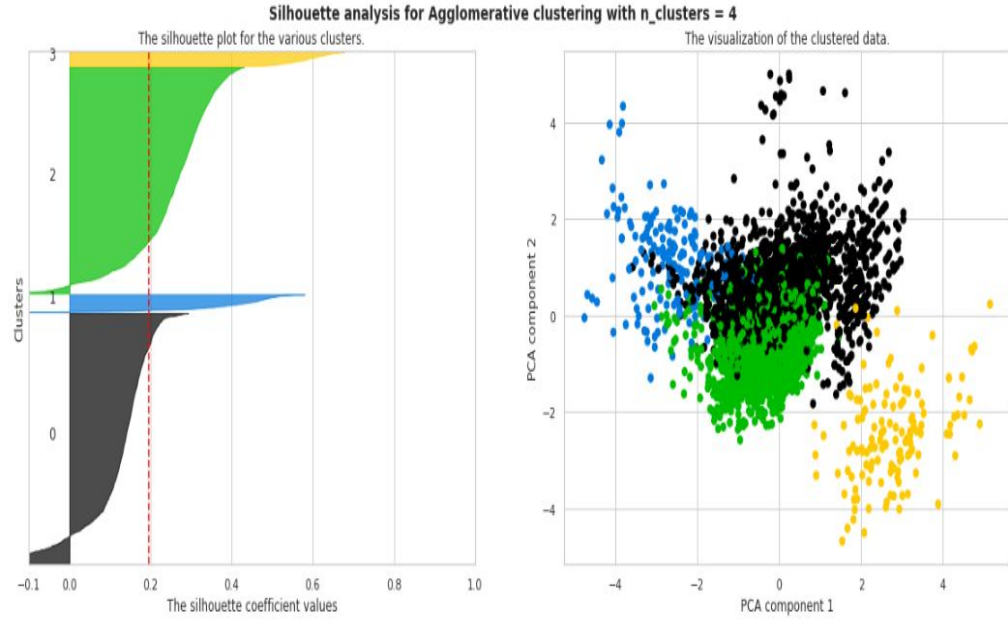
➢ StandardScaler was used to scale the data.

MODEL IMPLEMENTATION

Agglomerative Clustering-



Assume we cut vertical lines with a horizontal line to obtain the number of clusters. n=4 clusters.

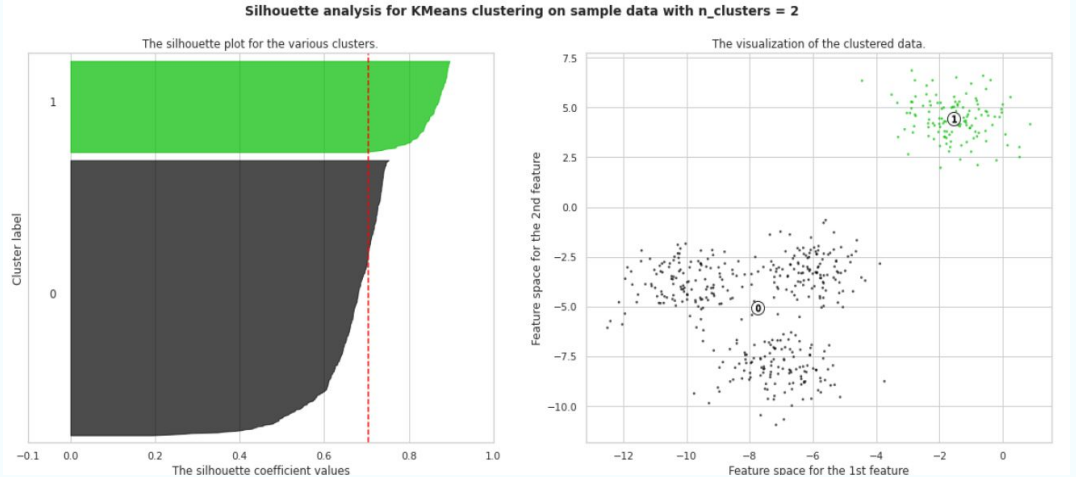


For clusters =4 , The average silhouette_score is : **0.19676189959151683** which is not good.

MODEL IMPLEMENTATION

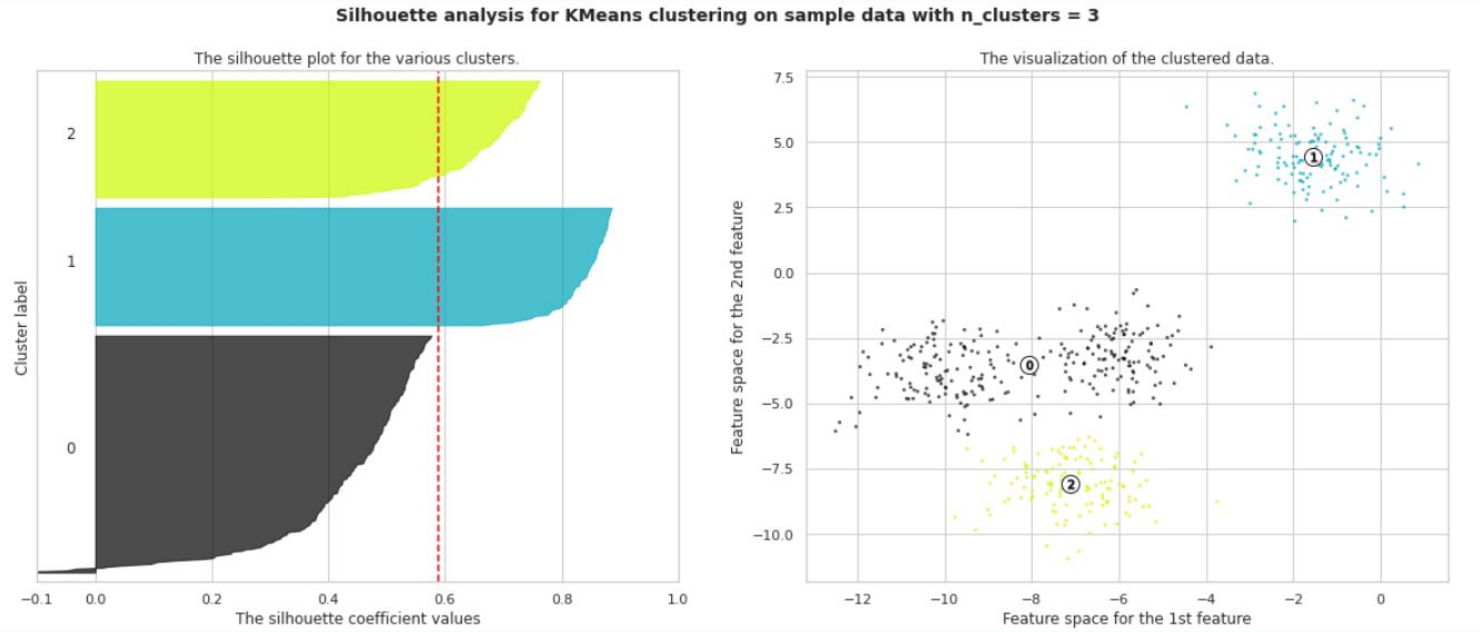
K-means Clustering-

```
# Generating the sample data from make_blobs

X, y = make_blobs(n_samples=500,
                  n_features=2,
                  centers=4,
                  cluster_std=1,
                  center_box=(-10.0, 10.0),
                  shuffle=True,
                  random_state=1)

range_n_clusters = [2, 3, 4, 5, 6]
```




```
For n_clusters = 2 The average silhouette_score is : 0.7049787496083262
For n_clusters = 3 The average silhouette_score is : 0.5882004012129721
For n_clusters = 4 The average silhouette_score is : 0.6505186632729437
For n_clusters = 5 The average silhouette_score is : 0.56376469026194
For n_clusters = 6 The average silhouette_score is : 0.4504666294372765
```

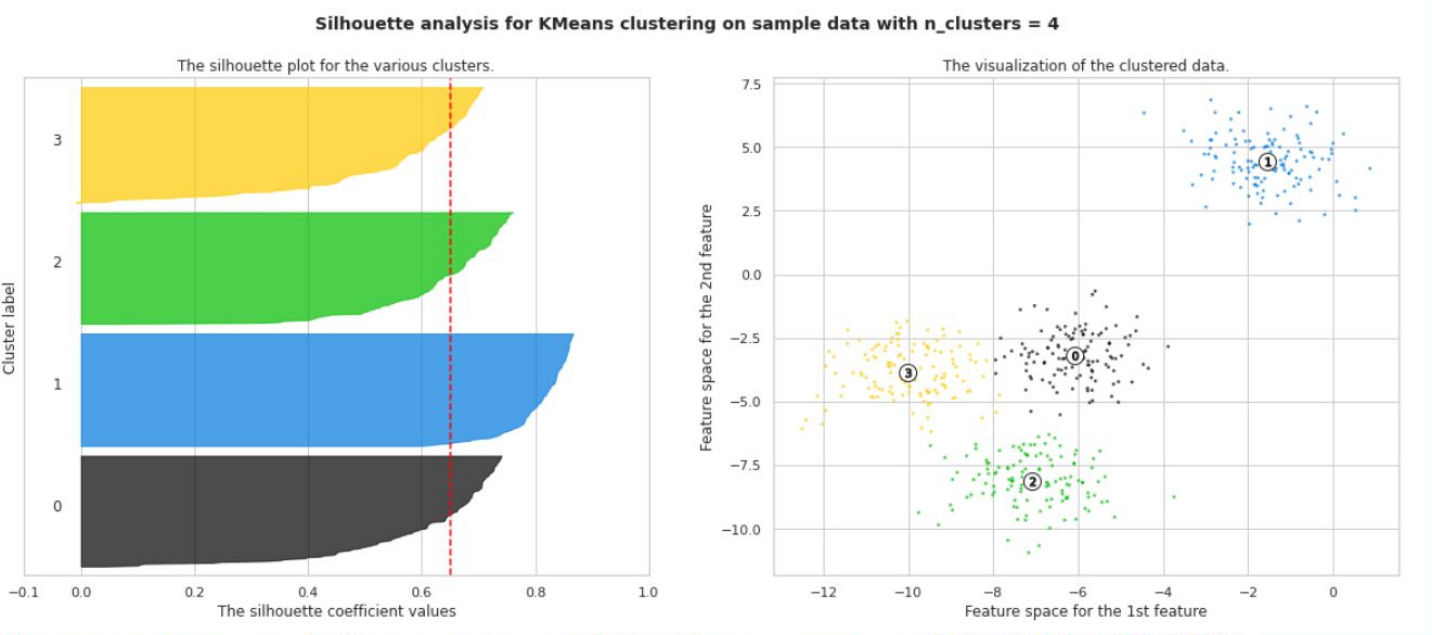
MODEL IMPLEMENTATION




```
For n_clusters = 2 The average silhouette_score is : 0.7049787496083262
For n_clusters = 3 The average silhouette_score is : 0.5882004012129721
For n_clusters = 4 The average silhouette_score is : 0.6505186632729437
For n_clusters = 5 The average silhouette_score is : 0.56376469026194
For n_clusters = 6 The average silhouette_score is : 0.4504666294372765
```

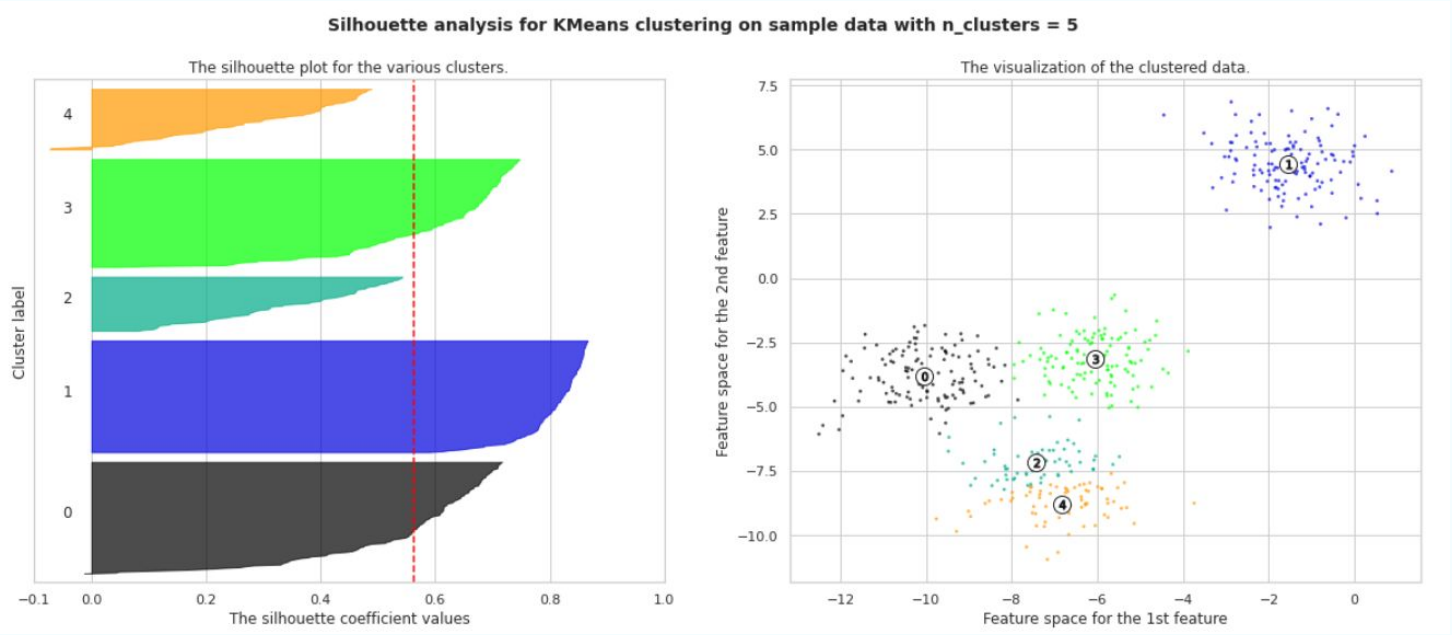
MODEL IMPLEMENTATION

K-means Clustering



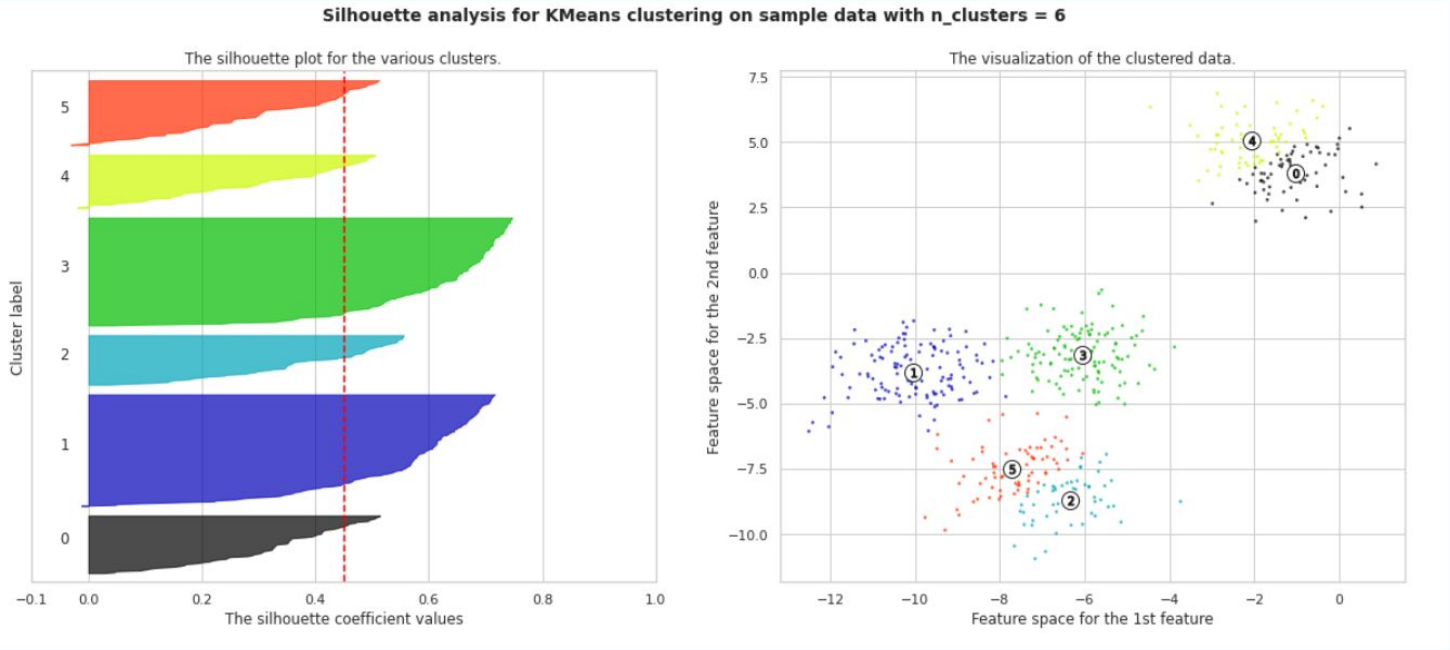

```
For n_clusters = 2 The average silhouette_score is : 0.7049787496083262
For n_clusters = 3 The average silhouette_score is : 0.5882004012129721
For n_clusters = 4 The average silhouette_score is : 0.6505186632729437
For n_clusters = 5 The average silhouette_score is : 0.56376469026194
For n_clusters = 6 The average silhouette_score is : 0.4504666294372765
```

MODEL IMPLEMENTATION



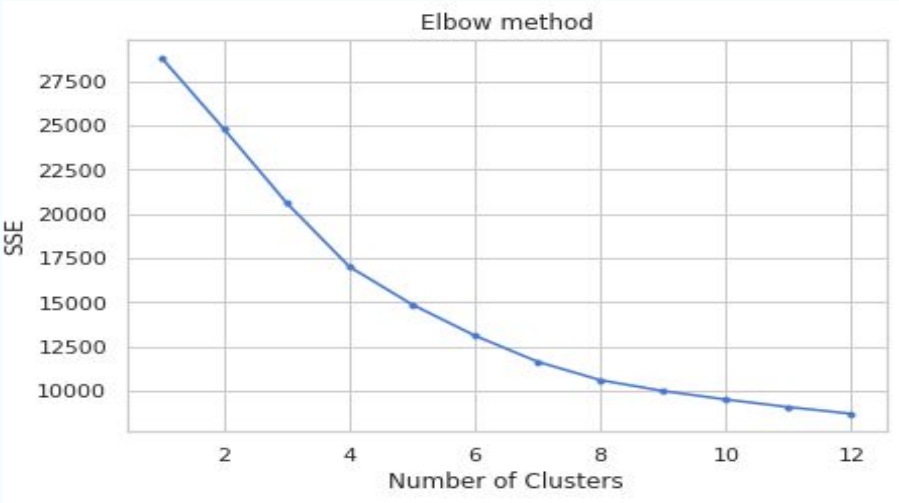

```
For n_clusters = 2 The average silhouette_score is : 0.7049787496083262
For n_clusters = 3 The average silhouette_score is : 0.5882004012129721
For n_clusters = 4 The average silhouette_score is : 0.6505186632729437
For n_clusters = 5 The average silhouette_score is : 0.56376469026194
For n_clusters = 6 The average silhouette_score is : 0.4504666294372765
```

MODEL IMPLEMENTATION




```
For n_clusters = 2 The average silhouette_score is : 0.7049787496083262
For n_clusters = 3 The average silhouette_score is : 0.5882004012129721
For n_clusters = 4 The average silhouette_score is : 0.6505186632729437
For n_clusters = 5 The average silhouette_score is : 0.56376469026194
For n_clusters = 6 The average silhouette_score is : 0.4504666294372765
```

Graphical Representation For K- means clustering

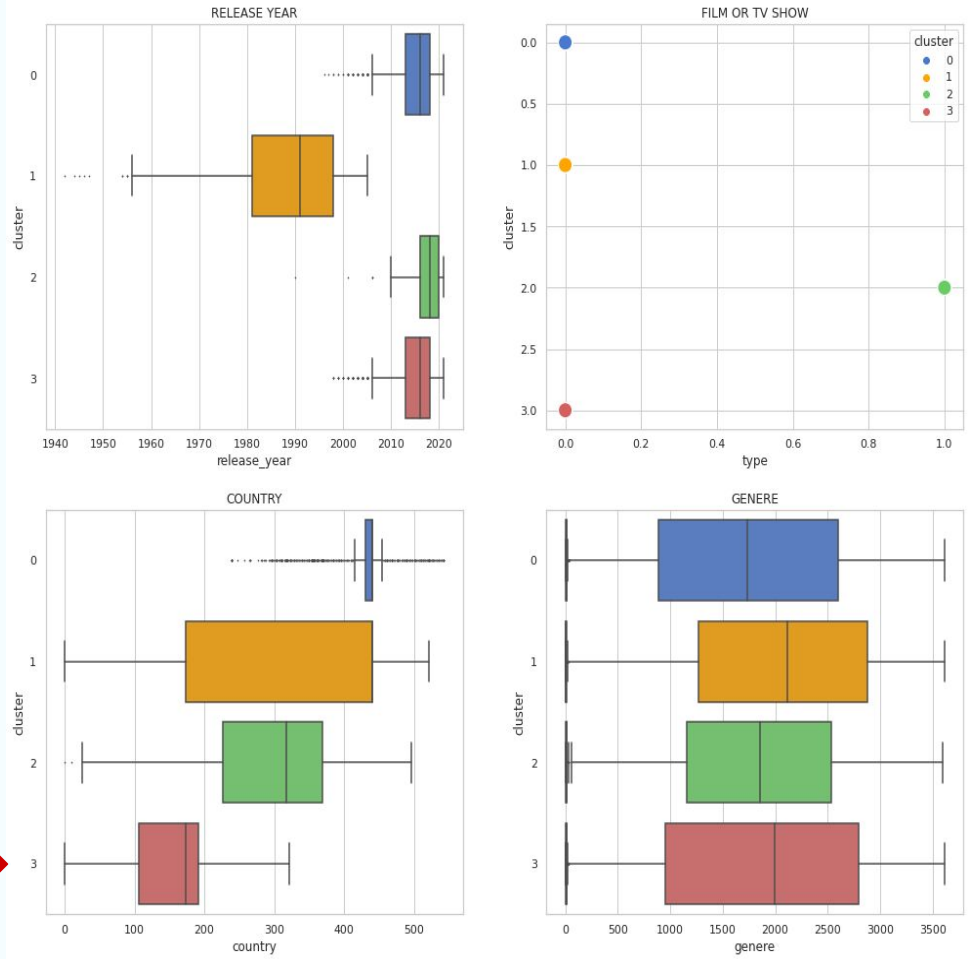


With an increase in the number of clusters (k),

the average SSE decreases.

Box plot for -

Release year,Type,Country & Genere



Conclusion

1. Most films were released in the years 2018, 2019, and 2020.
2. TV shows account for 2.8 percent of the total, while movies account for 97.2 percent.
3. Dramas is a genre that is mostly watched in Netflix and as per audience preference international movies are mostly watched.
4. The largest count of Netflix content is made with a “TV-14” rating,
5. The United States, India, the United Kingdom, Canada, and Egypt are the top five producers countries.
6. Netflix has added a lot more movies and TV episodes in the previous years, but the numbers are still low when compared to movies released in the last ten years.
7. Movies are mostly watched in various countries rather than TV shows.
8. We performed data engineering to remove the unnecessary variables and to convert the data into standard form into scalar.

Interpretation based on models-

1. A dendrogram was used to determine the number of clusters in Agglomerative Clustering.
2. For improvement in implemented model **K-means clustering** algorithm is utilised to get better results which is consist of 2,3,4,5,6 clusters.
3. After clustering, we can say that our alternative hypotheses is number of TV shows launched in the previous few years is NOT growing.
4. Our second alternative hypotheses is number of TV shows added to Netflix is higher.

Thank You !!..