CLUSTERING ANALYSIS ON

NETFLIX MOVIES AND TV SHOWS

By :Lucky Jain, Debashish Das, VivekKatolkar

Data Science Trainee   at Almabetter

ABSTRACT:-

Netflix is an American subscriptionstreaming 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. The monthly subscription by user makes Netflix a profitable business and the flexibility in subscription users can cancel it anytime. So to engage customers to this platform 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.

Introduction: -

Netflix's recommendation system gives the idea to them about the popularity of their services provides as it helps to increase the sold the subscriptions as more as possible, which offers a varieties of items for selections, this help to get them a user satisfaction,and their loyalty to platform and get them a better understanding of what the user wants.

Then it's easier to get the user to make better decisions from a wide variety of movie products.

With over 139 million paid subscribers(total viewer pool -300 million) across 190 countries, 15,400 titles across its regional libraries and 112 Emmy Award Nominations in 2018 — Netflix is the leading Internet television network and the most-valued largest streaming service in the world. The success behind the amazing story of Netflix is incomplete without the mention of its recommender systems that focus on personalization according to users. According to your preferences, there are several methods to create a list of recommendations. You can use (Collaborative-filtering) and (Content-based Filtering) for recommendation.

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 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.

In this project, we are evaluating as below-

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

Objective

The project's main goal is to create a model that can perform Clustering on comparable material by matching text-based attributes.

Dataset Peeping

The dataset has 7787 rows and 12 attributes to work with.

1. We have NaN values in the dataset.
2. Changed the format of the Date.
3. Added some columns which are extracted from the Date column.

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

Challenges Faced

The following are the challenges faced in the data analysis:

- ➢ Conversion of Datetime features, categorical features.
- ➢ Feature engineering
- ➢ Model Implementation

Approach

As per the problem statement, understanding what type of content is available in different countries and Is Netflix increasingly focused on TV rather than movies in recent years we have to do clustering on similar content by matching text-based features. For that we used Affinity Propagation, Agglomerative Clustering, and K-means Clustering.

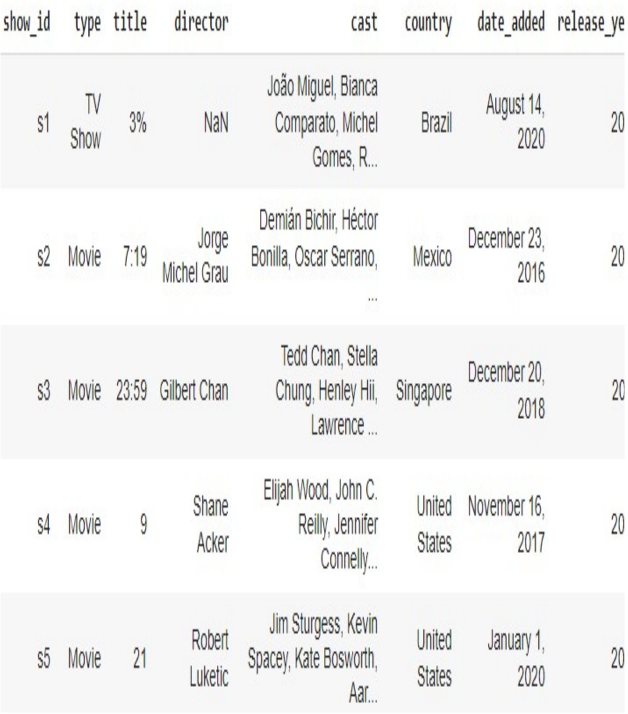
Tools Used

The whole project was done using python, in google collab. Following libraries were used for analyzing the data and visualizing it and to build the model to predict the bike count required at each hour for the stable supply of rental bikes.

- Pandas: Extensively used to load and wrangle with the dataset.

- Matplotlib: Used for visualization.
- Seaborn: Used for visualization.
- Datetime: Used for analyzing the date variable.
- Warnings: For filtering and ignoring the warnings.
- Numpy: For some math operations in predictions.
- Sklearn: For the purpose of analysis and prediction.
- Datetime: For reading the date.

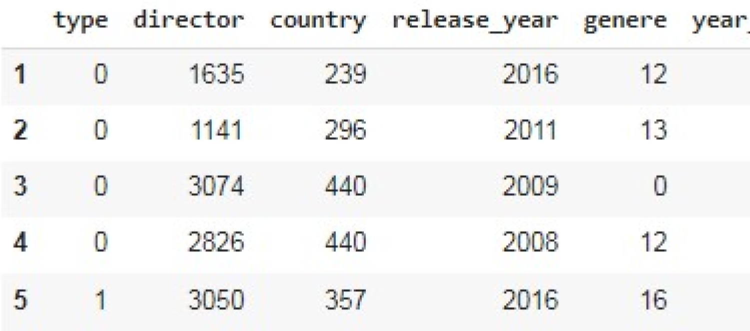
The below table shows the dataset in the form of Pandas DataFrame:-



show_id	type	title	director	cast	country	date_added	release_ye
s1	TV Show	3%	NaN	João Miguel, Bianca Comparato, Michel Gomes, R...	Brazil	August 14, 2020	20
s2	Movie	7:19	Jorge Michel Grau	Demián Bichir, Héctor Bonilla, Oscar Serrano, ...	Mexico	December 23, 2016	20
s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella Chung, Henley Hii, Lawrence ...	Singapore	December 20, 2018	20
s4	Movie	9	Shane Acker	Elijah Wood, John C. Reilly, Jennifer Connelly...	United States	November 16, 2017	20
s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin Spacey, Kate Bosworth, Aar...	United States	January 1, 2020	20

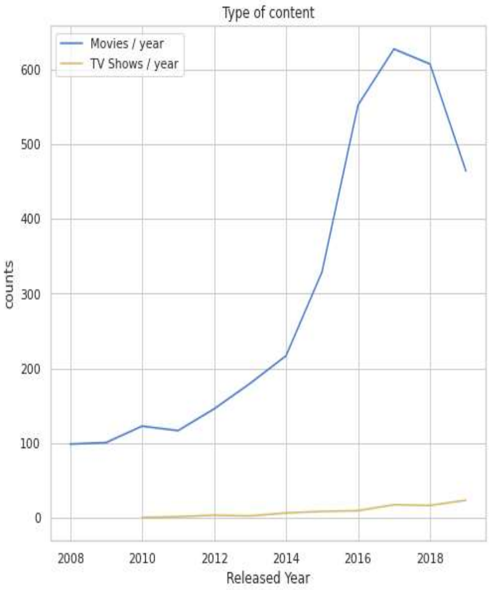
Feature Engineering

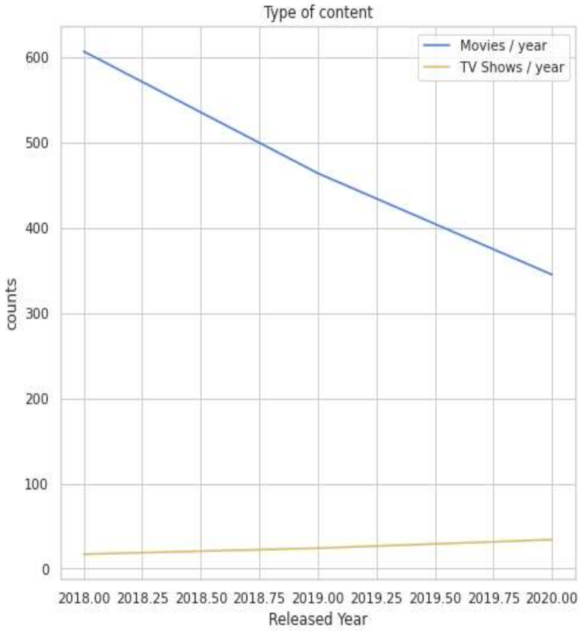
- There are too much classes, so we just obtain the first 50 (the most common 50)
- Unify some of the similar types(genre)
- Make a dictionary with similar content by matching text-based features that we are going to use in clustering.



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

Hypothesis Evaluation-



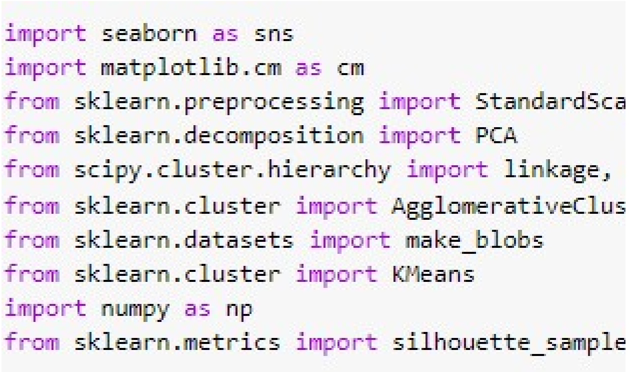


Hypothesis from the data visualized-

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.

Importing required libraries for clustering



```
import seaborn as sns
import matplotlib.cm as cm
from sklearn.preprocessing import StandardSca
from sklearn.decomposition import PCA
from scipy.cluster.hierarchy import linkage,
from sklearn.cluster import AgglomerativeClus
from sklearn.datasets import make_blobs
from sklearn.cluster import KMeans
import numpy as np
from sklearn.metrics import silhouette_sample
```

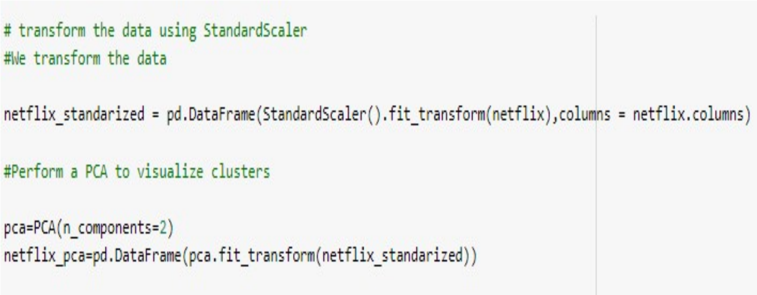
Building a clustering model-

Clustering models allow you to categorize records into a certain number of clusters. This can help you identify natural groups in your data.

Clustering models focus on identifying groups of similar records and labeling the records according to the group to which they belong. This is done without the benefit of prior knowledge about the groups and their characteristics. In fact, you may not even know exactly how many groups to look for. This is what distinguishes clustering models from the other machine-learning techniques—there is no predefined output or target field for the model to predict. These models are often referred to as **unsupervised learning** models, since there is no external standard by which to judge the model's classification performance.

Scaling the data

We used StandardScaler to transform 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))
```

Metrics:Silhouette Coefficient or silhouette score-

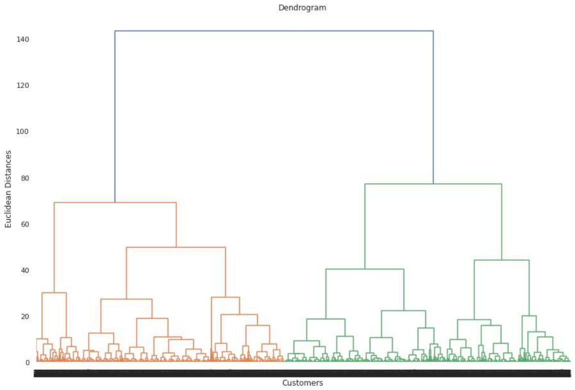
Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.

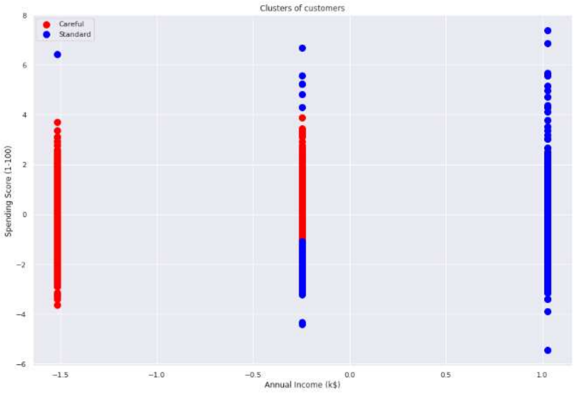
Model Implementation

1. **Agglomerative Clustering-**

The agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity.Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects.
We used a dendrogram to find the number of clusters.

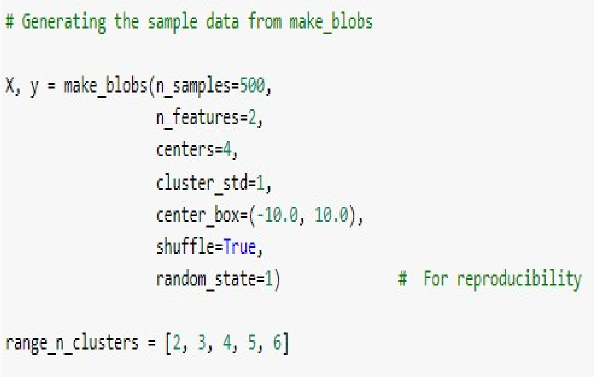


Assume we cut vertical lines with a horizontal line to obtain the number of clusters. **Number of clusters = 2**



2. K-means Clustering

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition*n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.
We created the sample data using build blobs and used range_n_clusters to specify the number of clusters we wanted to utilize in k means.



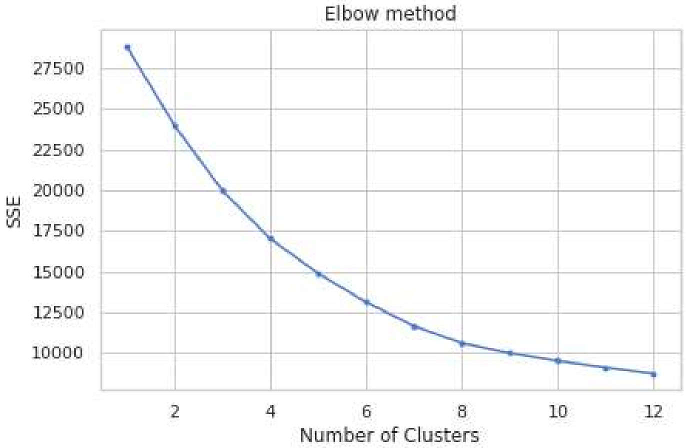
```
# Generating the sample data from make_blobs

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)           #  For reproducibility

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

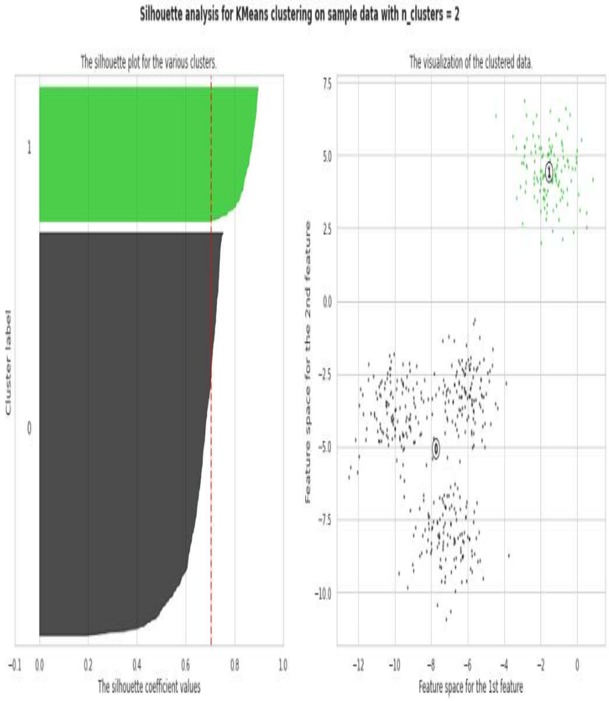
Elbow Method-

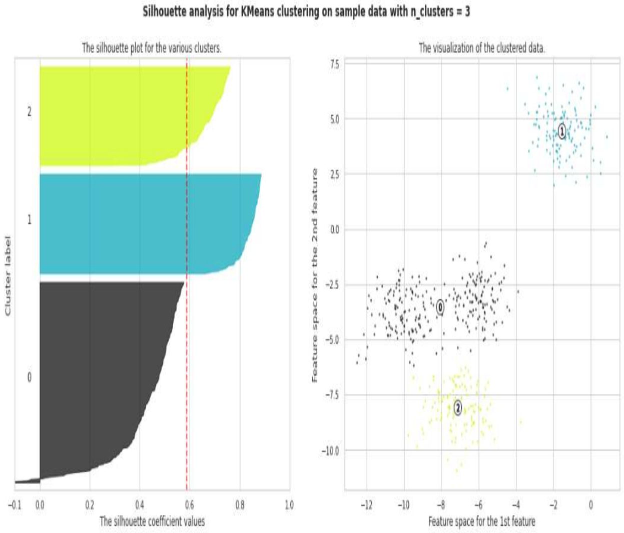
The Elbow Method is an empirical method to find the optimal number of clusters for a dataset. In this method, we pick a range of candidate values of k, then apply K-Means clustering using each of the values of k. Find the average distance of each point in a cluster to its centroid, and represent it in a plot. Pick the value of k, where the average distance falls suddenly.

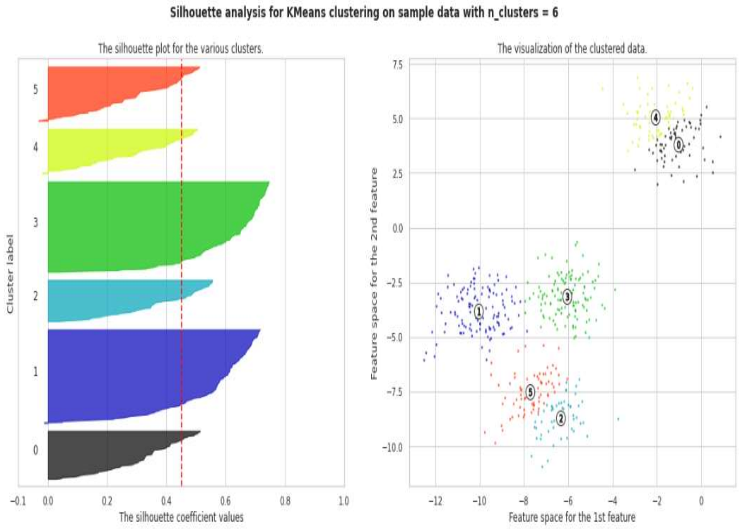


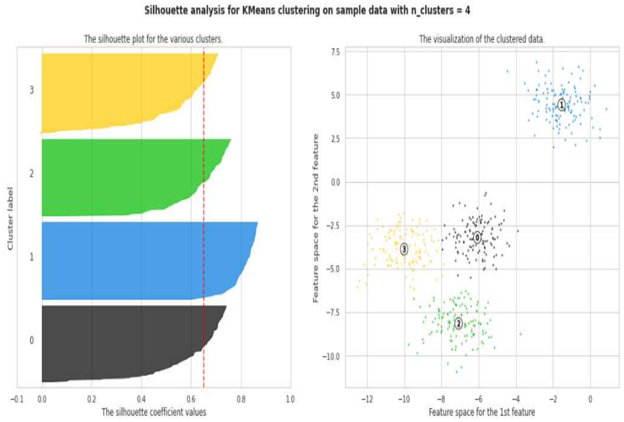
With an increase in the number of clusters (k), the average SSE decreases. To select the best value of k we use Silhouette score as below-

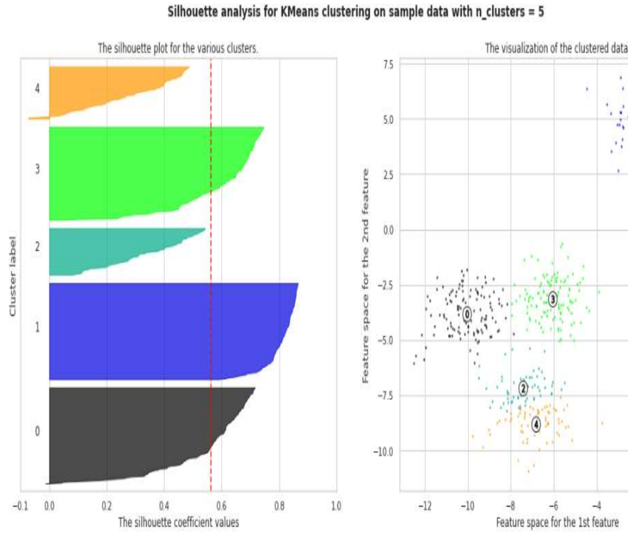
Silhouette score and visualization-









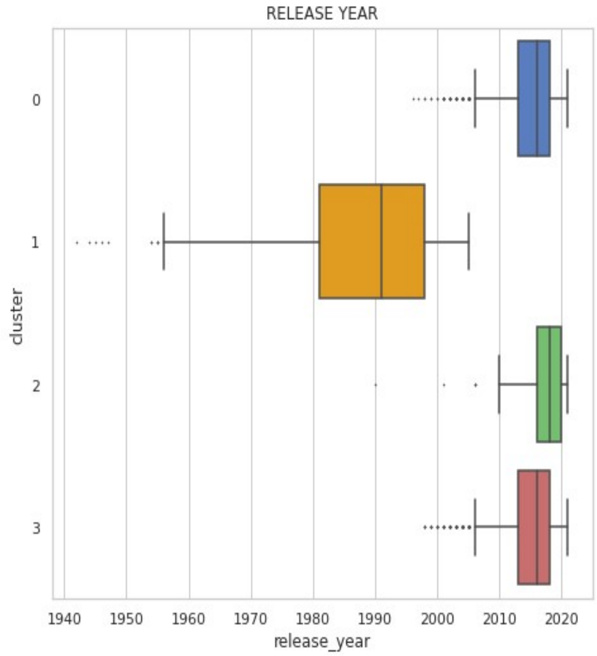


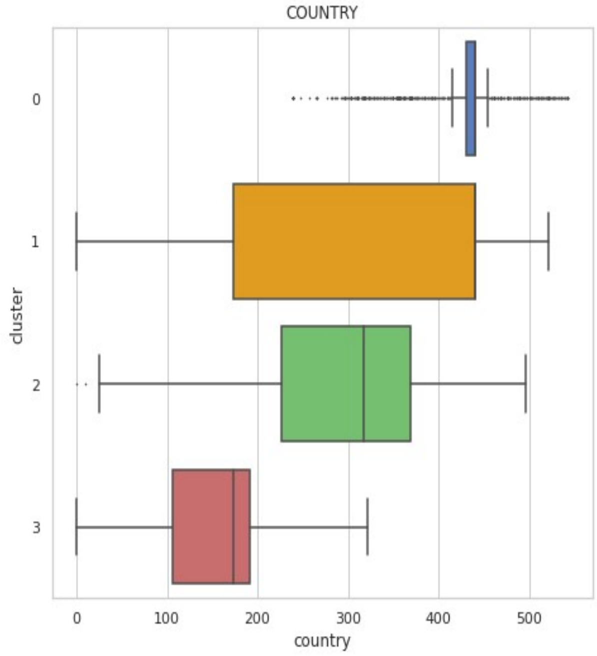
For n_clusters = 2, silhouette score Is 0.42541313028836003
For n_clusters = 3, silhouette score is 0.3940500353920696
For n clusters = 4, silhouette score is 0.38498095158183726
For n_clusters = 5, silhouette score is 0.3962372504377786
For n_clusters = 6, silhouette score is 0.3925658868286329

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

We also plot some boxplots for our clusters-

A boxplot is a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.





<u>Conclusion-</u>

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 on 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 producer 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 standardized form into scalar.**
9. **Implemented model is based on the K-means clustering algorithm consisting of 2,3,4,5,6 clusters.**

   Silhouette Analysis score for K-means :

   - **For n_clusters = 2, silhouette score Is 0.42541313028836003**
   - **For n_clusters = 3, silhouette score is 0.3940500353920696**
   - **For n clusters = 4, silhouette score is 0.38498095158183726**
   - **For n_clusters = 5, silhouette score is 0.3962372504377786**
   - **For n_clusters = 6, silhouette score is 0.3925658868286329**

11. **After clustering, we can say that our alternative hypothesis is that the number of TV shows launched in the previous few years is NOT growing.**
12. **Our second alternative hypothesis is the number of TV shows added to Netflix is higher.**

Thank you...!!