

# Capstone Project-4

## NETFLIX MOVIES AND TV SHOWS CLUSTERING

Abhijeet Kulkarni , Kundan Lal  
Pankaj Ganjare , Mohd sharik

# Content:

- Introduction
- Problem Statement
- Data Description
- Null Value
- Exploratory Data Analysis
- Outlier detection
- NLP
- K- Modes
- Spectral Clustering
- Implementing D.B Scan
- Cluster Analysis



# Introduction

## Netflix:

Netflix is a company that manages a large collection of TV shows and movies, streaming it anytime via online. This business is profitable because users make a monthly payment to access the platform. However, customers can cancel their subscriptions at any time.

## Methodology:

- Unsupervised Machine Learning (Clustering)

## Database:

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

# 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.



**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

 The data was collected from Flixable which is third party Netflix search engine. The dataset consists of movies and TV shows data till 2019. The dataset has 7787 rows of data.

 The dataset consists of eleven textual columns and one numeric column.

## 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

# Data Description

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

# Null Value

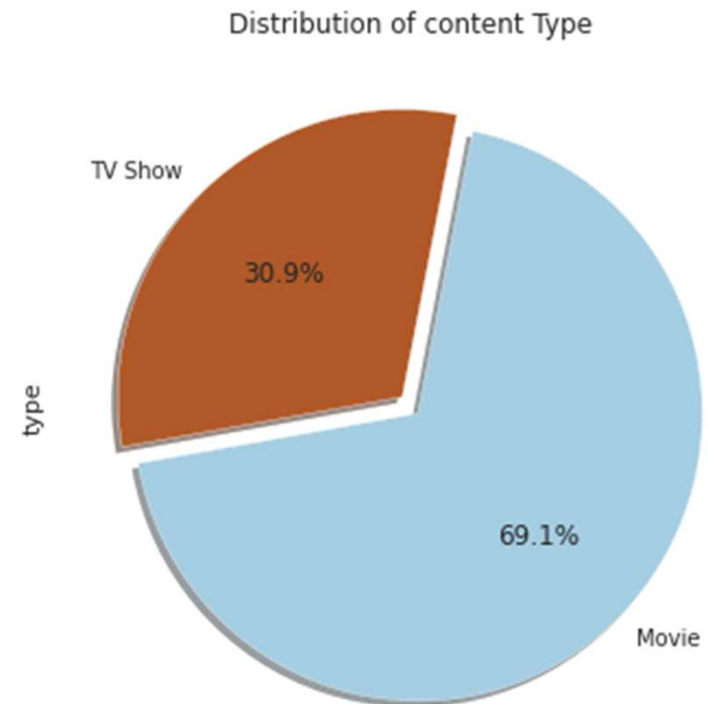


## Null Value Treatment:

- **Director** feature have more than **30.68%** of null values. Filling null values by 'unknown'.
- **Country** feature have **6.51%** of null values. Filling null values by mode of feature.
- **Cast feature** have **9.22%** of null values. Filling null values by 'unknown'.
- **Rating** feature have **0.09%** of null values. Filling null values by mode of feature.
- **Date\_added** feature have **0.13%** of null values. Dropping rows corresponding to null values.

## EDA

From plot we can say that there is 30.9% content from TV shows and 69.1% from Movie.





## Top 15 countries based on content type

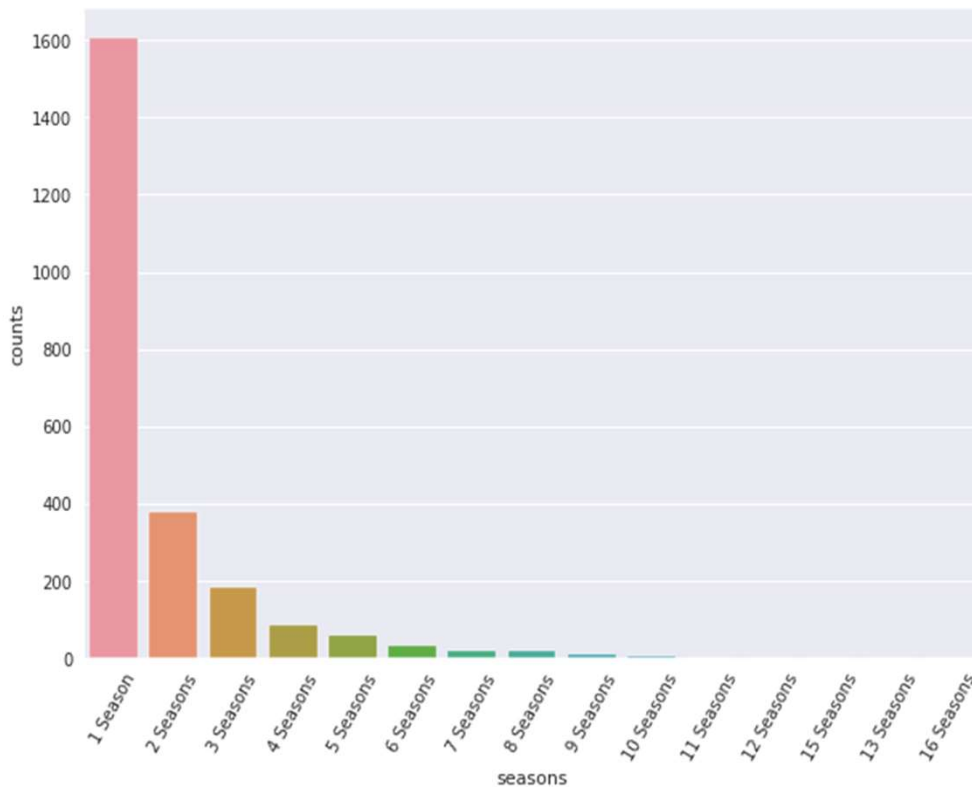
## Top 10 directors who have directed most movies

	country	shows	counts	country	movies	counts
0	United States		860	United States		2427
1	United Kingdom		255	India		915
2	Japan		182	United Kingdom		466
3	South Korea		157	Canada		286
4	Canada		126	France		265
5	France		84	Spain		158
6	India		75	Germany		157
7	Taiwan		70	Japan		103
8	Australia		58	China		102
9	Spain		57	Mexico		101
10	Mexico		53	Egypt		97
11	China		45	Hong Kong		97
12	Germany		42	Australia		84
13	Brazil		29	Turkey		80
14	Colombia		28	Philippines		77

	director	counts
0	Raúl Campos, Jan Suter	18
1	Marcus Raboy	16
2	Jay Karas	14
3	Cathy Garcia-Molina	13
4	Jay Chapman	12
5	Youssef Chahine	12
6	Martin Scorsese	12
7	Steven Spielberg	10
8	David Dhawan	9
9	Hakan Algül	8

## seasons distribution for TV shows

There are only few TV shows which crossed 8 season most of seasons have released in one 1 season.



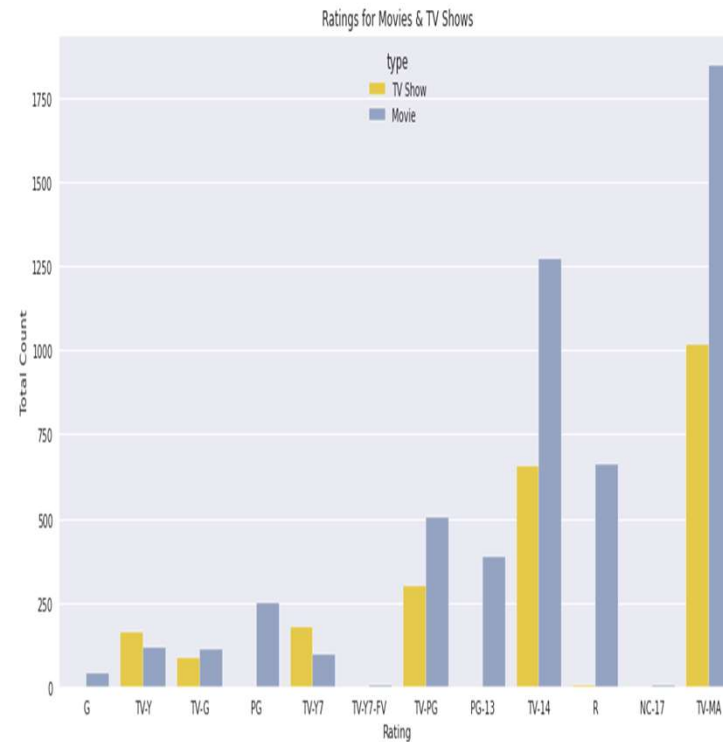
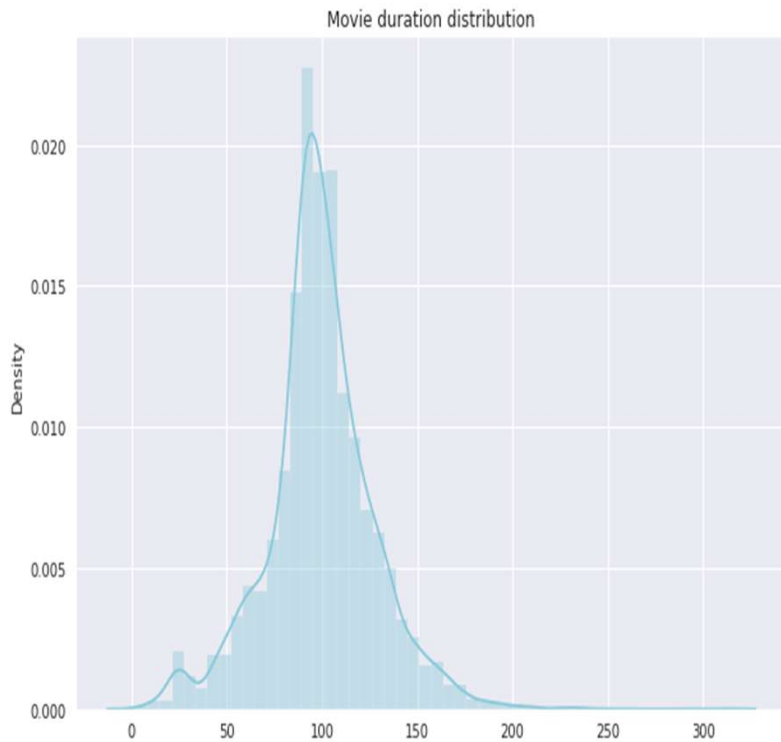
	seasons	counts
0	1 Season	1606
1	2 Seasons	378
2	3 Seasons	183
3	4 Seasons	86
4	5 Seasons	57
5	6 Seasons	30
6	7 Seasons	19
7	8 Seasons	18
8	9 Seasons	8
9	10 Seasons	5
10	11 Seasons	2
11	12 Seasons	2
12	15 Seasons	2
13	13 Seasons	1
14	16 Seasons	1

### Movie duration distribution

Distribution is positively skewed and there are very less moves which crossed 175 mins.

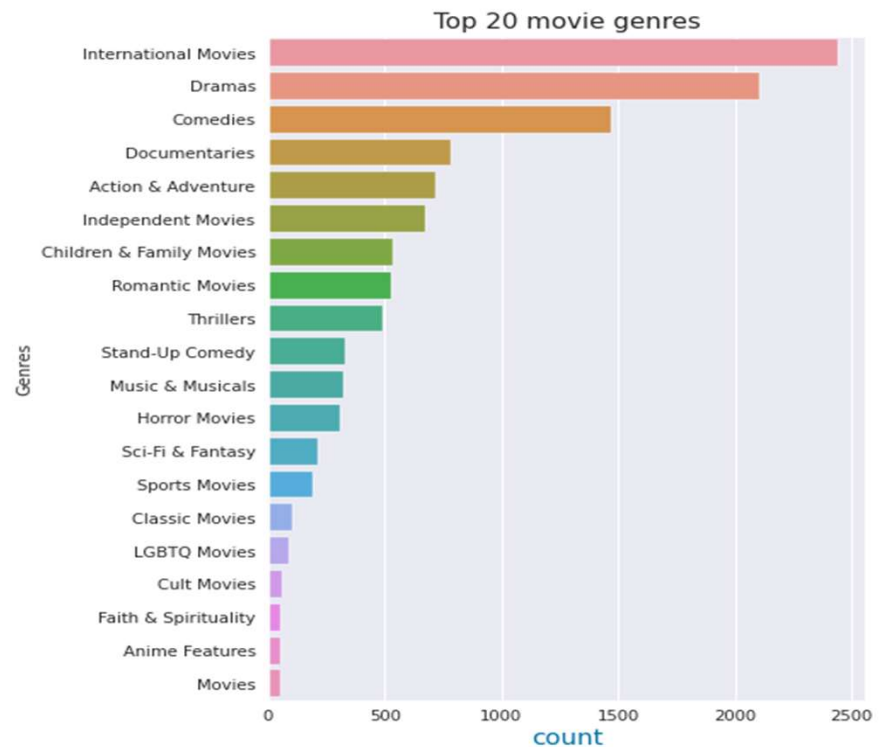
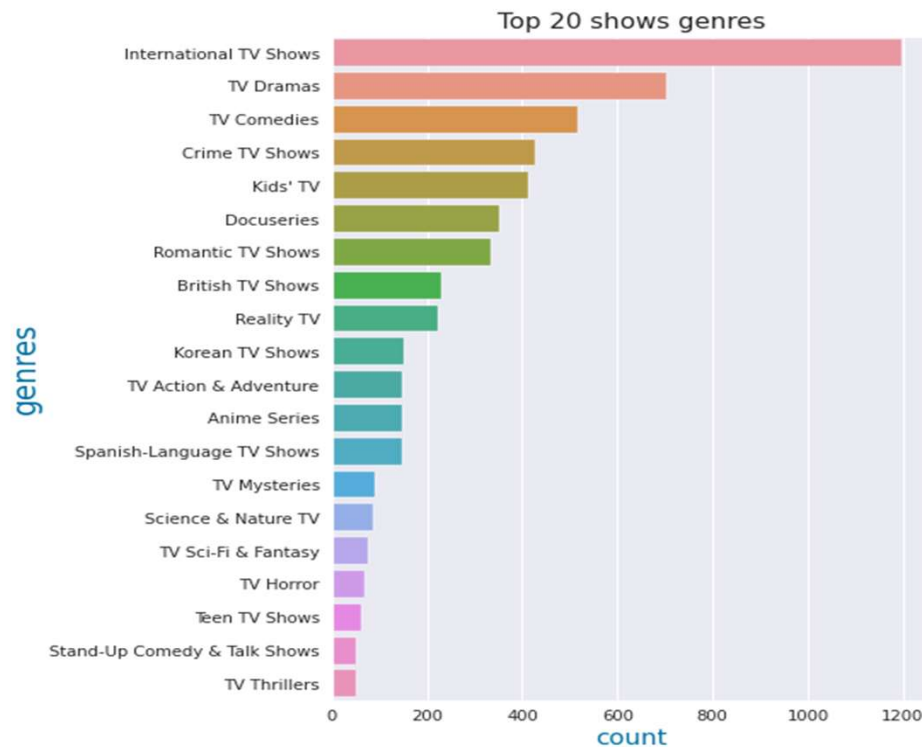
### Distribution of various ratings

Most contents are adult rated on Netflix and there are very less G-rated contents



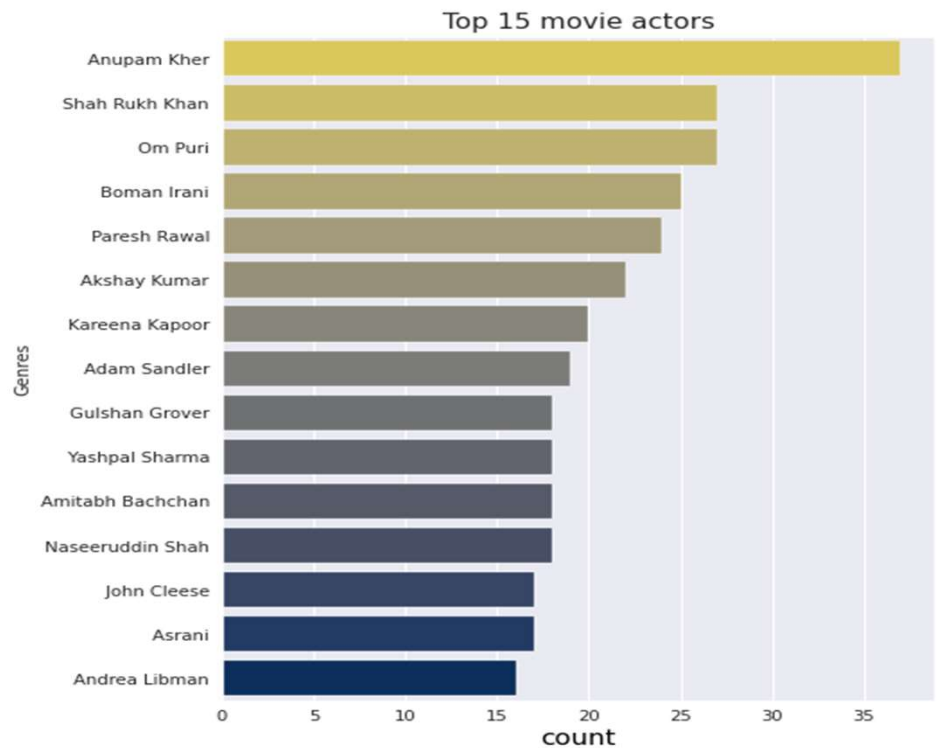
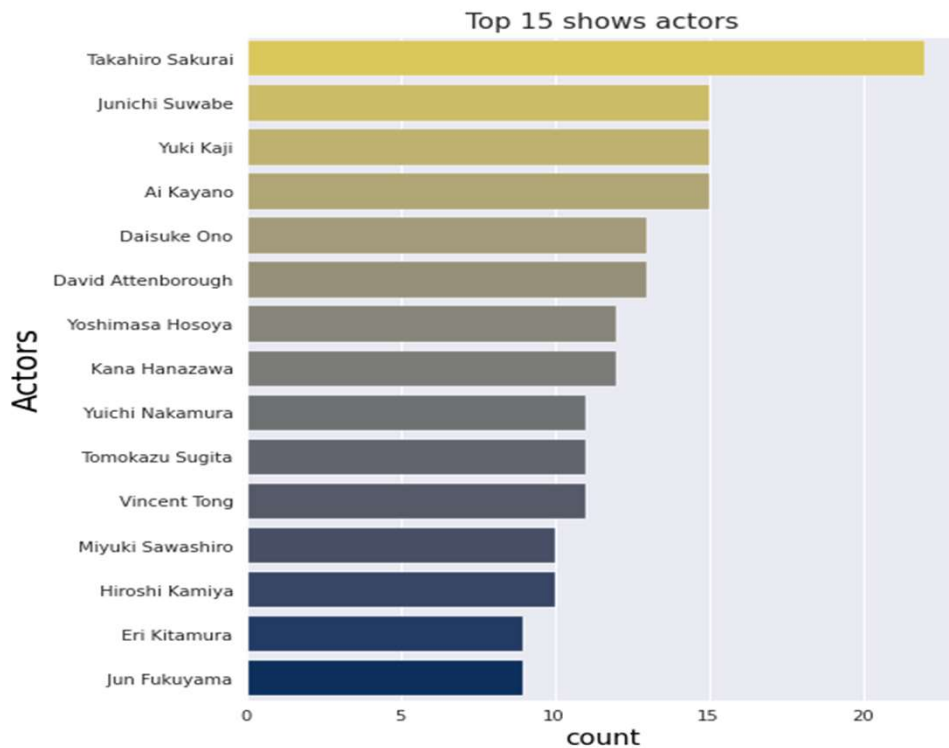
## Top 20 movie genres in movies and TV shows

In both type international genres is on top in the list followed by drama and comedies respectively.



### Top actors who worked in most content in movies and TV shows.

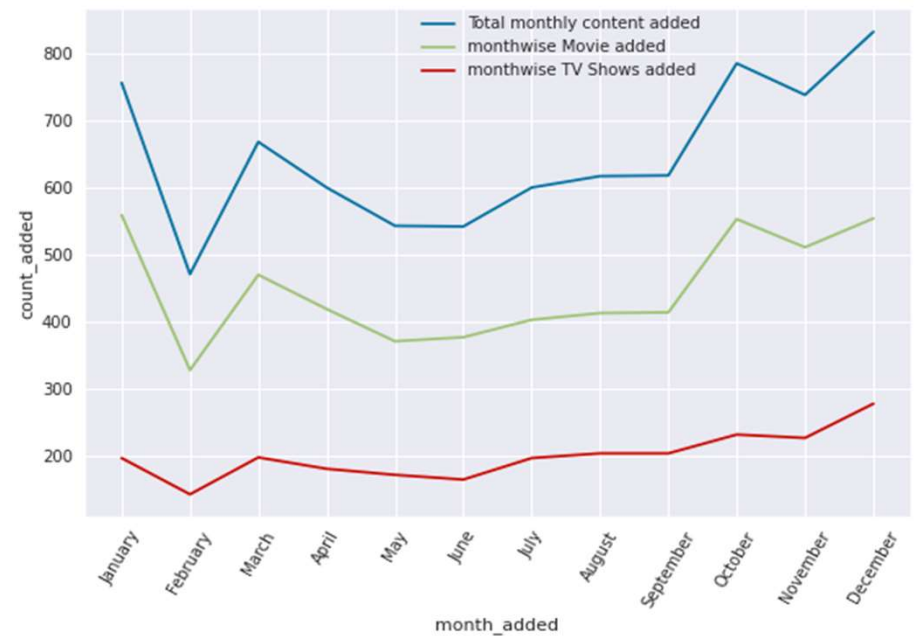
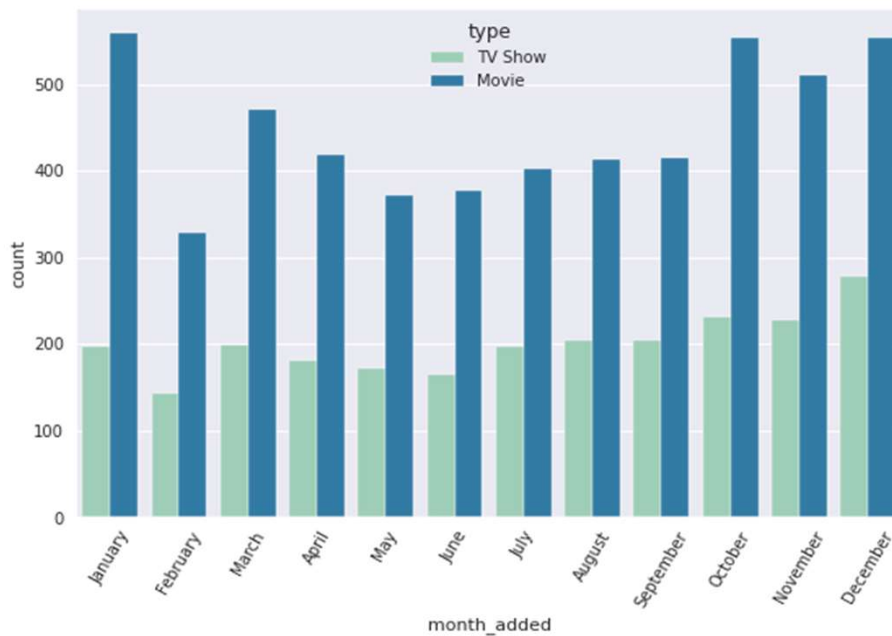
actors who were part of most shows and movies which gave us that Takahiro sukurai has worked in most shows and Anupam kher has worked in most movies.



## Month wise content distribution

With the help of month wise analysis we have found that most content is added on the platform in the month of Nov, Dec, and Jan.

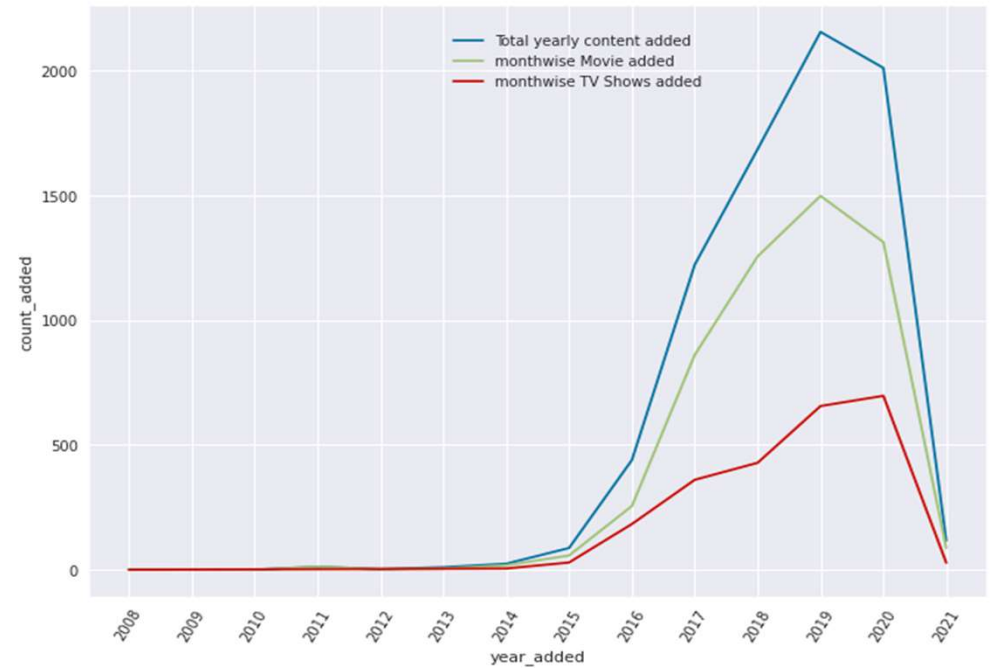
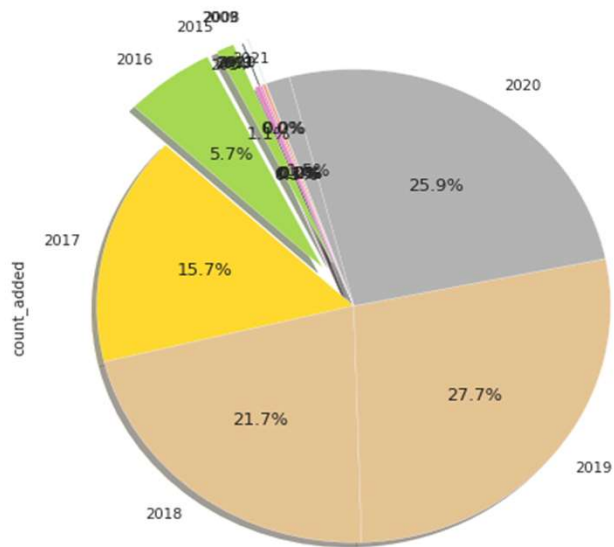
monthwise content added plots



## Year wise content distribution

In this distribution we have found that in recent year content is added on the platform exponentially. also we have found that between 2019 to 2020 Netflix has added more shows in comparison of movies

monthwise content added plots



# Outlier detection

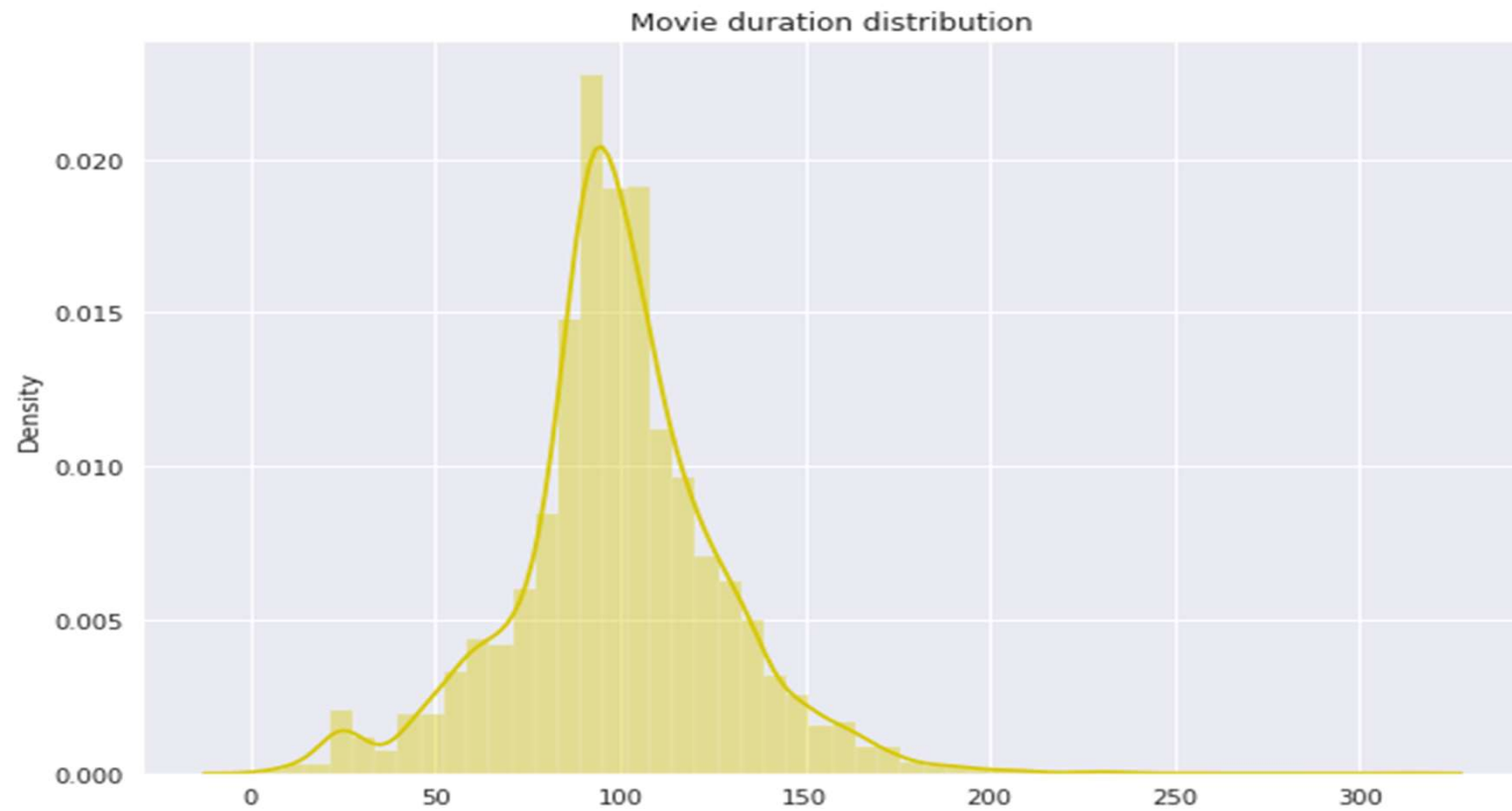
- ❑ After performing EDA on the dataset we have get rid of outliers since outliers can lead to bad clustering hence anomalies and outliers must be removed from the dataset.
- ❑ the nature of columns and values of these columns most of the columns content a lot unique values hence we cannot detect outliers however we check for the outliers in country and duration column if exist.
- ❑ basis on this data we cannot find outlier rows in country column



# Outliers in duration

- ❑ we have two content types (TV shows and movies ) and the values are different in both these types hence we will perform these tasks separately in two individual data frames. i.e shows\_df & movie\_df.
- ❑ detecting outlier detection in shows\_df as in this data frame duration columns contains categorical values i.e seasons
- ❑ based on this result we can assume that there are less than 1 percent shows which are 7 seasons or long hence we can consider these seasons as outliers
- ❑ we will be using isolation forest technique to detect outlier in duration column of movies\_df

# Movie duration distribution



# Applying NLP

- ❑ we have developed an NLP algorithm that can give us similar content. however, based on the nature of the content like movie/ TV show we have developed two function so that we can get similar movies for a given and TV shows . To achive this we have used TF-IDF vectorizer and cosine similarity. the details of both these are given above when we used them.
- ❑ Since we have used the TF-IDF vectorizer, calculating the dot product will directly give us the cosine similarity score. Therefore, we will use sklearn's `linear_kernel()` instead of `cosine_similarities()` since it is faster.

- We will be using the cosine similarity to calculate a numeric quantity that denotes the similarity between two movies. We use the cosine similarity score since it is independent of magnitude and is relatively easy and fast to calculate. Mathematically, it is defined as follows:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}},$$

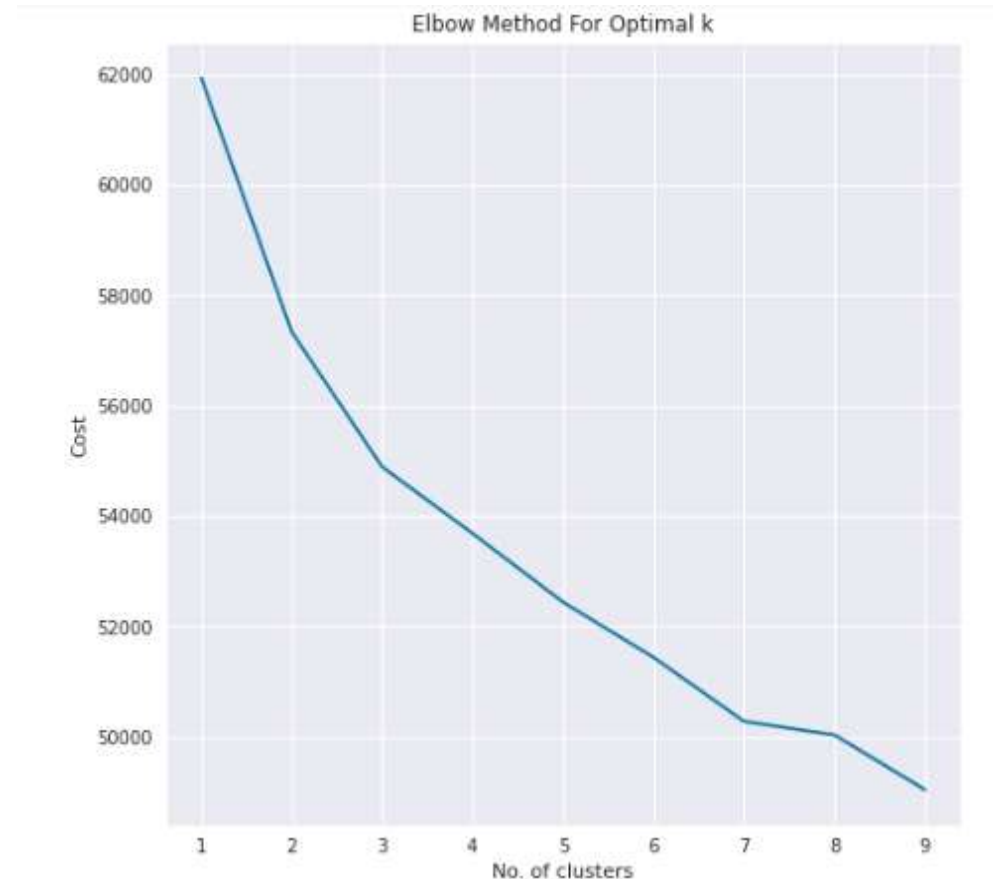
# K-modes Clustering

## *Why K-modes ?*

*Since we had categorical data and K-modes clustering algorithm was established for categorical data. It works on the modes of categorical data. i.e most occurred observations and build clusters for most similar points.*

*3 clusters are best no. according to K-Modes:*

*An Elbow plot was plotted using K-modes algorithm to decide the best suitable no. of cluster for the given dataset. Using this plot we can say that the most suitable no. of clusters is 3 .*

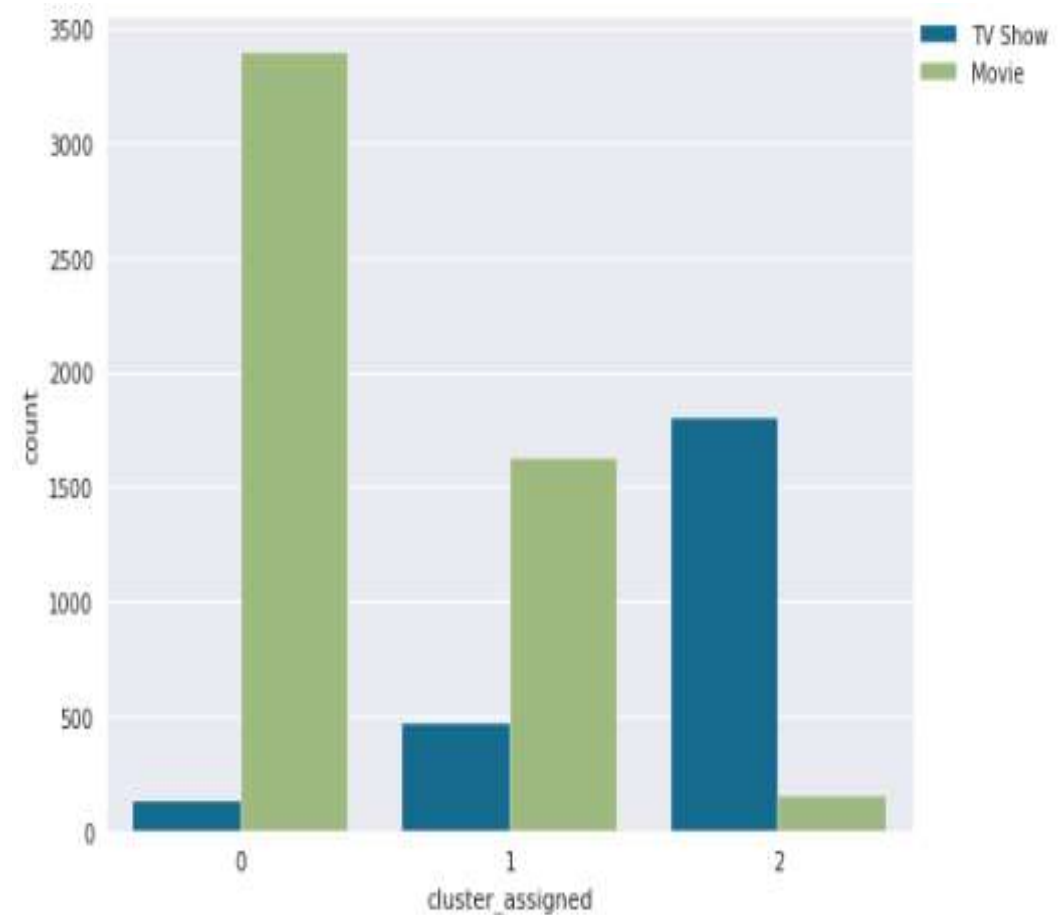


# K-modes Clustering

*How the no. of points(content) falls in each cluster ?*

*After algorithm implementation a count plot was plotted to visualize no. of points falls in each cluster.*

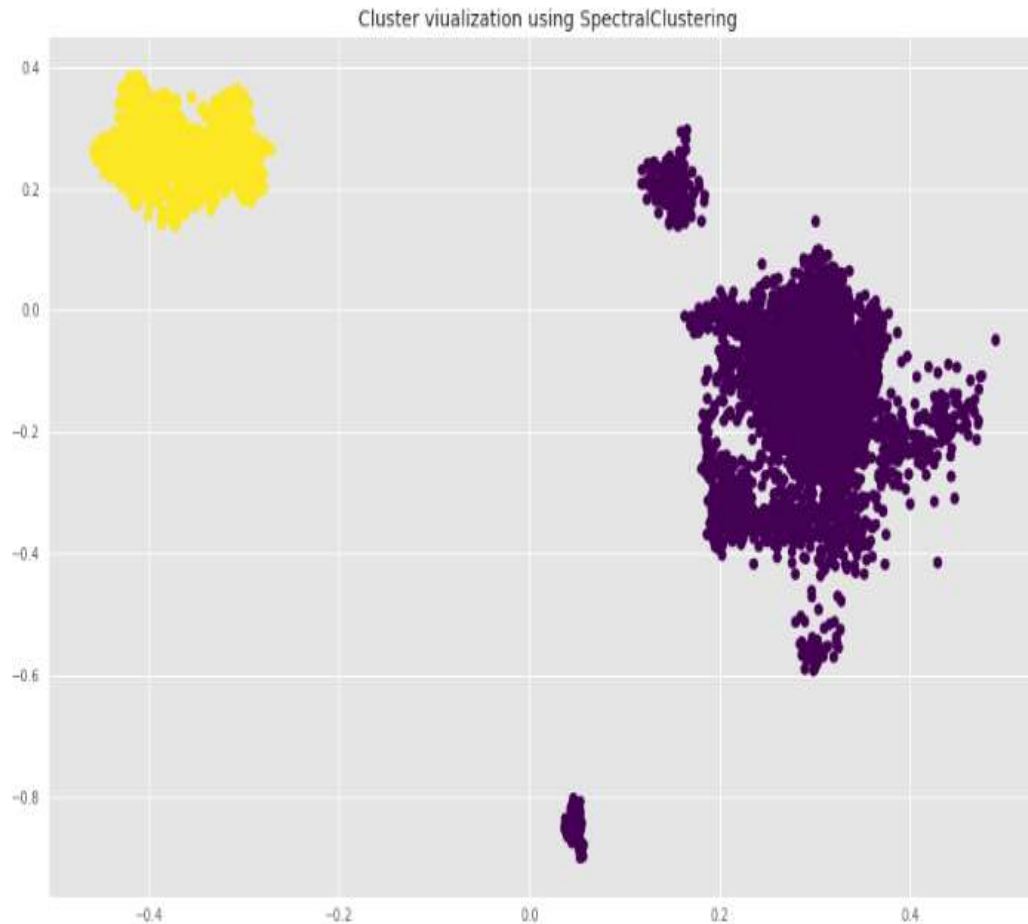
- 1. Mostly movies are in cluster 0*
- 2. Mostly movies are in cluster1 but the ratio has changed from cluster 0*
- 3. Mostly TV shows are in cluster 2 having only few movies*



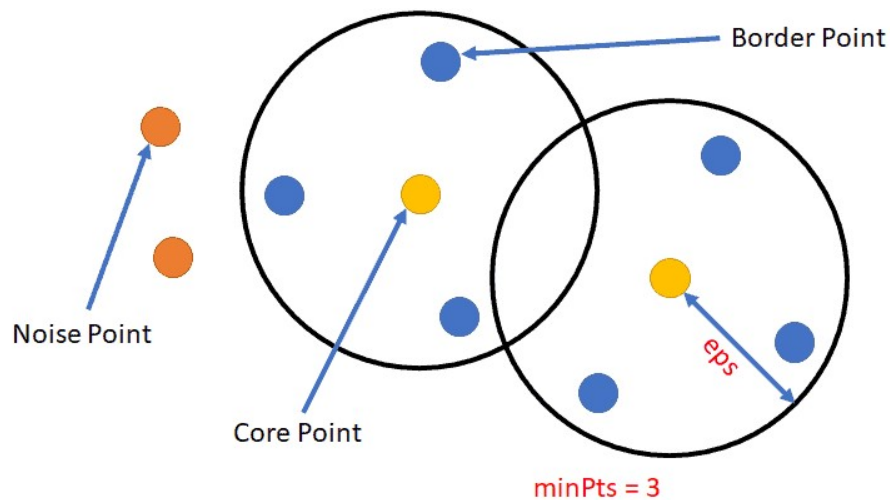
# Spectral Clustering

*Using Spectral clustering on transformed data we were able to built 2 clusters which is a different results from other algorithms.*

*We can see that how the clusters are separate from each other.*



## Implementing D. B Scan Algorithm



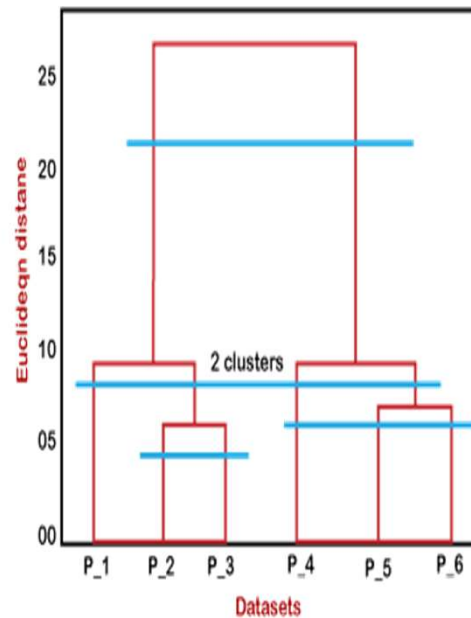
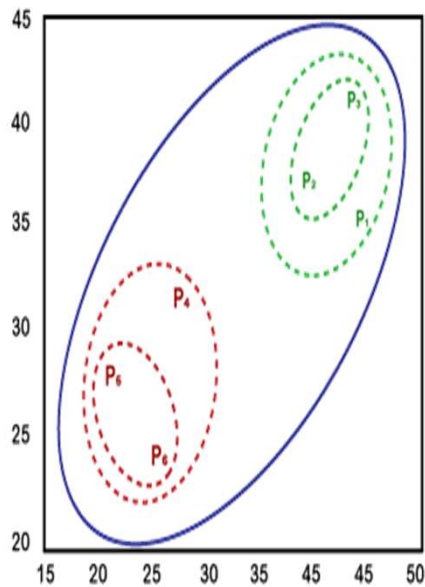
**DBSCAN** is an unsupervised learning algorithm for clustering. This algorithm is Density-based Spatial Clustering of Applications with Noise. It's a density based clustering non-parametric algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions.



## D. B Scan Hyper-parameters.

Epsilon	Epsilon is the local radius for expanding clusters. Think of it as a step size - DBSCAN never takes a step larger than this, but by doing multiple steps DBSCAN clusters can become much larger than eps .
Min Sample	Min_Samples refers to the number of neighboring points required for a point to be considered as a dense region, or a valid cluster. Default is 5 .
Algorithm	The algorithm to be used , usually its set as “auto” .
Leaf Size	Leaf_size value ( <b>default = 30</b> ) , It is Leaf size passed to BallTree or cKDTree. This can affect the speed of the construction and query, as well as the memory required to store the tree.
Random State	random state is a model hyper parameter used to control the randomness involved in machine learning models.
	No of Clusters formed 4

# Implementing Hierarchical Clustering



**Dendograms** are used to visually represent clustering operations, specifically agglomerative and divisive hierarchical clustering.

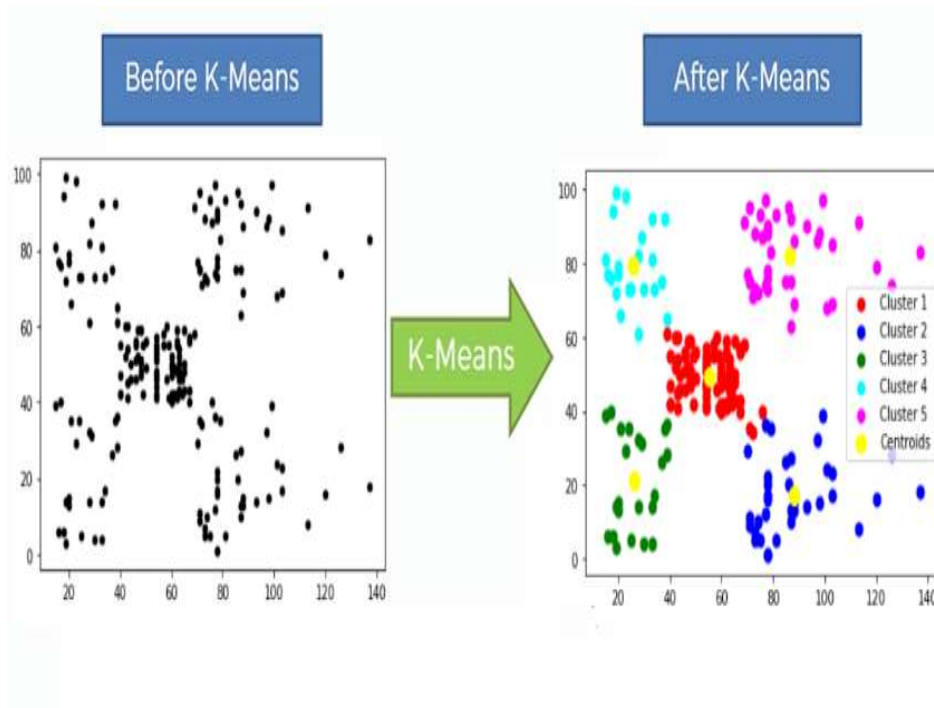
**Agglomerative hierarchical** clustering is where the elements start off in their own cluster and then are repeatedly combined by some criteria until all are in one cluster.

**Divisive hierarchical** clustering is where all elements start off in the same cluster, and then repeatedly broken by some criteria until all are in their own cluster.

## Hierarchical (Agglomerative clustering) Hyper-parameters.

n_clusters	locate the largest vertical difference between nodes, and in the middle pass an horizontal line. The number of vertical lines intersecting it is the optimal number of clusters.
Affinity	Euclidean , . _____ . , distance between 2 data points on a plane.
Linkage	<b>Average-linkage</b> and <b>complete-linkage</b> are the two most popular distance metrics in hierarchical clustering.
random_state	random state is a model hyperparameter used to control the randomness involved in machine learning models. No. Of Clusters formed - 3

# Implementing K-Means Clustering



K-means clustering uses “centroids”, **K** different randomly-initiated points in the data, and assigns every data point to the nearest centroid.

it aims to partition data into **k** clusters in a way that data points in the same cluster are similar and data points in the different clusters are farther apart.

## K-Means Clustering Algorithm Hyper-parameters.

silhouette score	<p>Silhouette score is used to evaluate the quality of clusters created using clustering algorithms like K-Means in terms of how well samples are clustered with other samples that are similar to each other.</p> <p>*1 being the best . Range -1 to 1 .</p>
n_clusters	<p>The number of clusters formed with the same number of centroids .</p> <p>No. of Clusters Formed - 3</p>

# Conclusion

- ⌚ Data set contains 7787 rows and 12 columns in that cast and director features contains large number of missing values so we can drop it.
- ⌚ We have two types of content TV shows and Movies (30.9% contains TV shows and 69.1% contains Movies).
- ⌚ Most films were released in the years 2018, 2019, and 2020 and united states have the maximum content on Netflix.
- ⌚ The months of October, November, December and January had the largest number of films and Tv-shows released.
- ⌚ The USA, India, the United Kingdom, Canada, and Egypt are the top five producer countries.
- ⌚ For the clustering algorithm, we utilized type, director, nation, released year, genre, and year.
- ⌚ Applied different clustering models Kmeans, hierarchical, Agglomerative clustering on data we got the best cluster arrangements.
- ⌚ We cut vertical lines with a horizontal line to obtain the number of clusters in Agglomerative Clustering. There were four clusters, with an average silhouette score of 0.17296314851287742.
- ⌚ The final model we used was k-means clustering, which consisted of 2,3,4,5,6 clusters. 4 numbers of clusters gives us good fitting.
- ⌚ After applying K - means optimal value of number of clusters is 5

**Thank you**