Capstone Project

Project Title

Netflix Movies and TV shows Clustering

By- Arkopravo Pradhan

Content :

1. Defining problem statement
2. EDA and feature engineering
3. Feature Selection
4. Data Preprocessing
5. Applying different clustering methods
6. Applying Clustering Model
7. Conclusion

PROBLEM STATEMENT

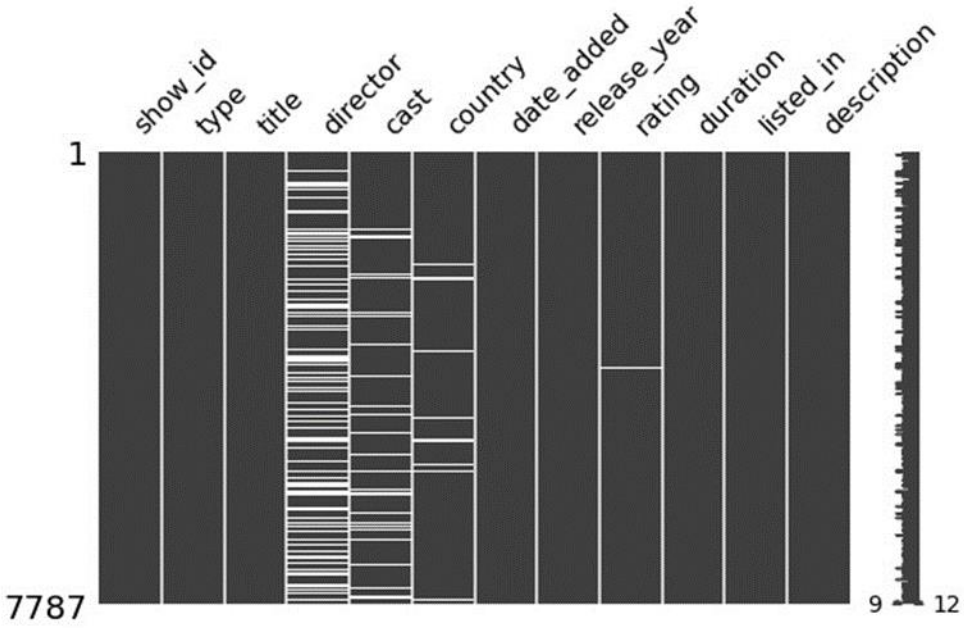
- This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine.
- In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming 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.

Data Summary

- **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 :** Genere
- **12.description:** The Summary description

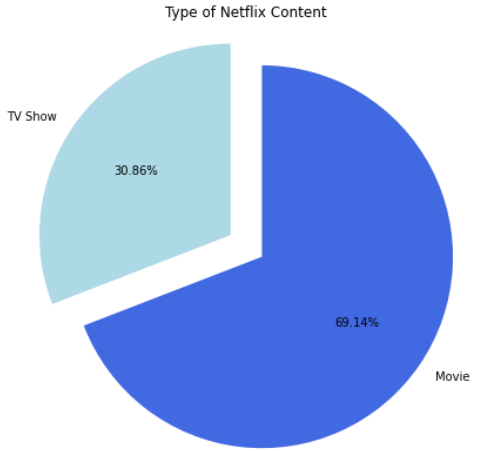
Exploratory Data Analysis

- We tried to plot the nan values of the dataset for each column and the result is director and cast contains large number of null values so we will drop it.

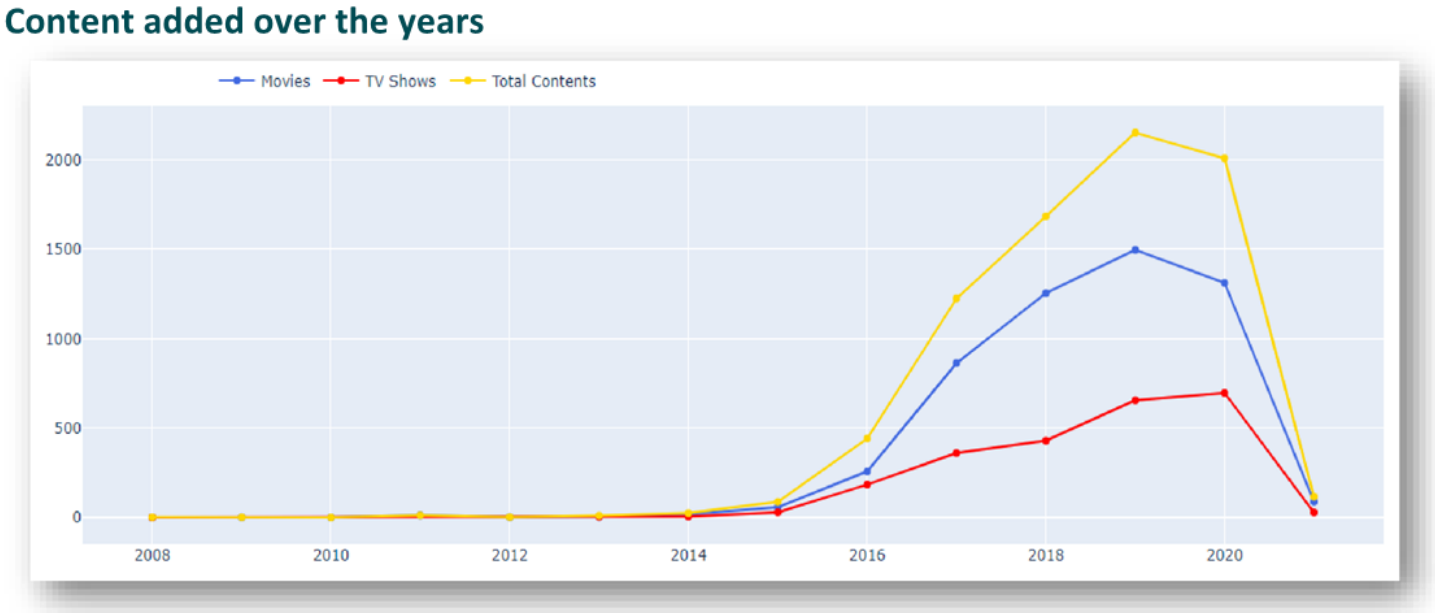


Exploratory Data Analysis

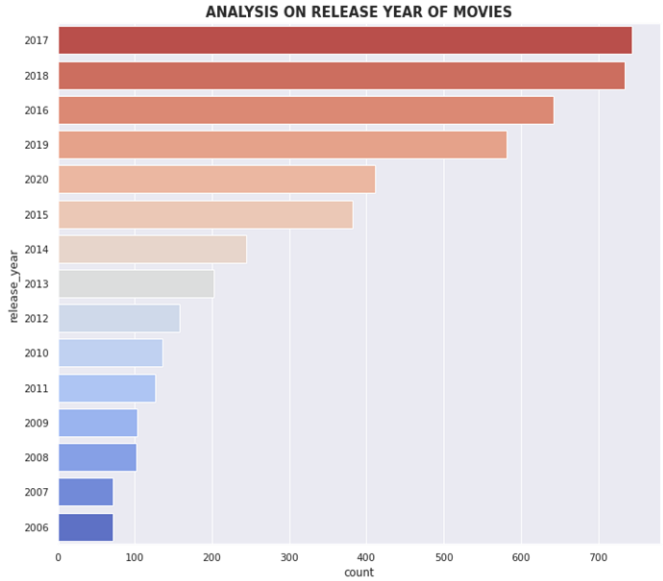
- Plotted pie plot of the type column (Netflix content). It resulted 30.86% tv show and 69.14% movie.

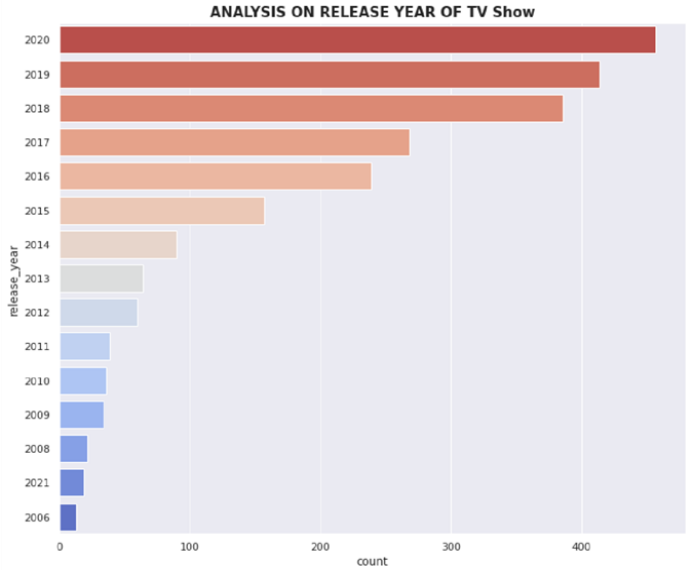


Exploratory Data Analysis



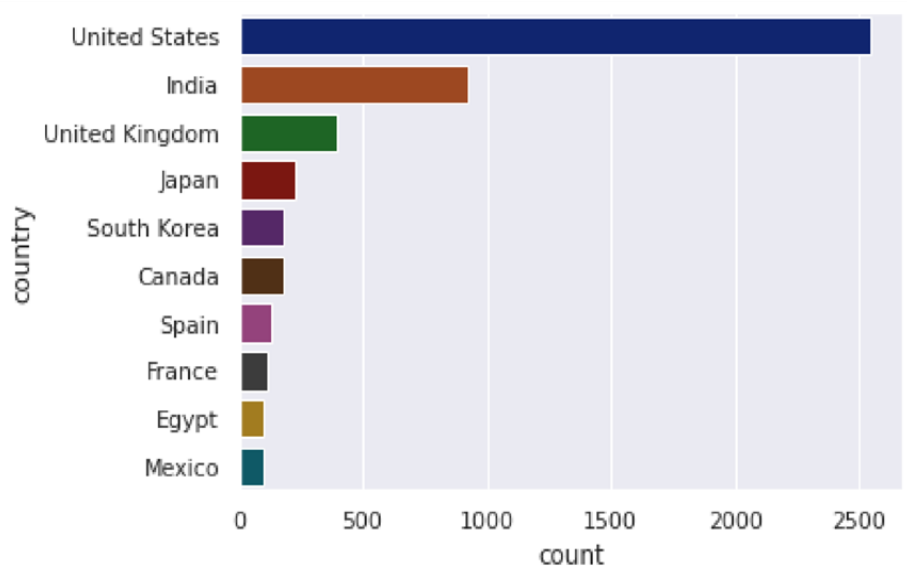
Exploratory Data Analysis



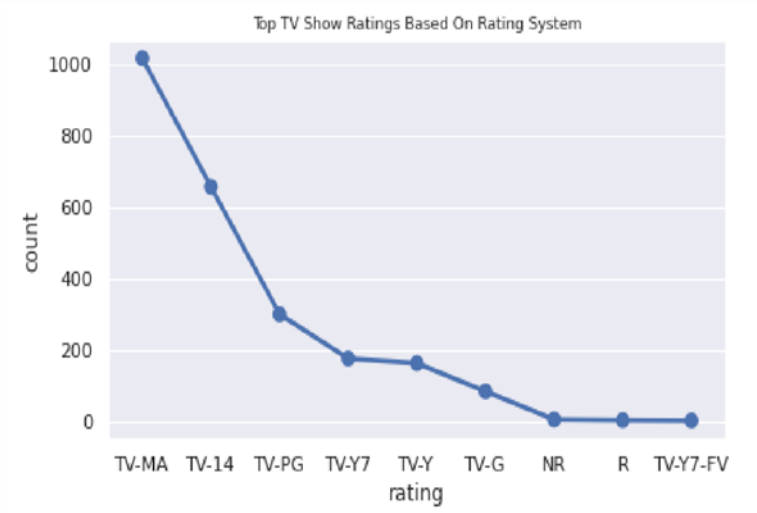


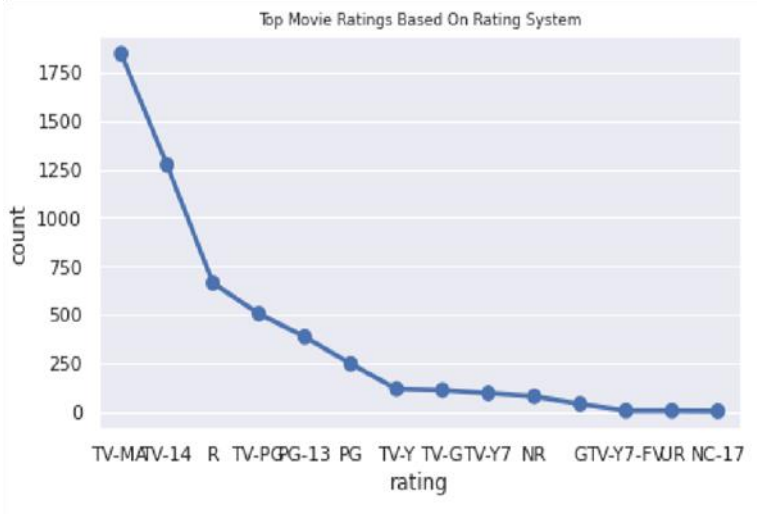
Exploratory Data Analysis

- United States have the greatest number of content and then India and so on.



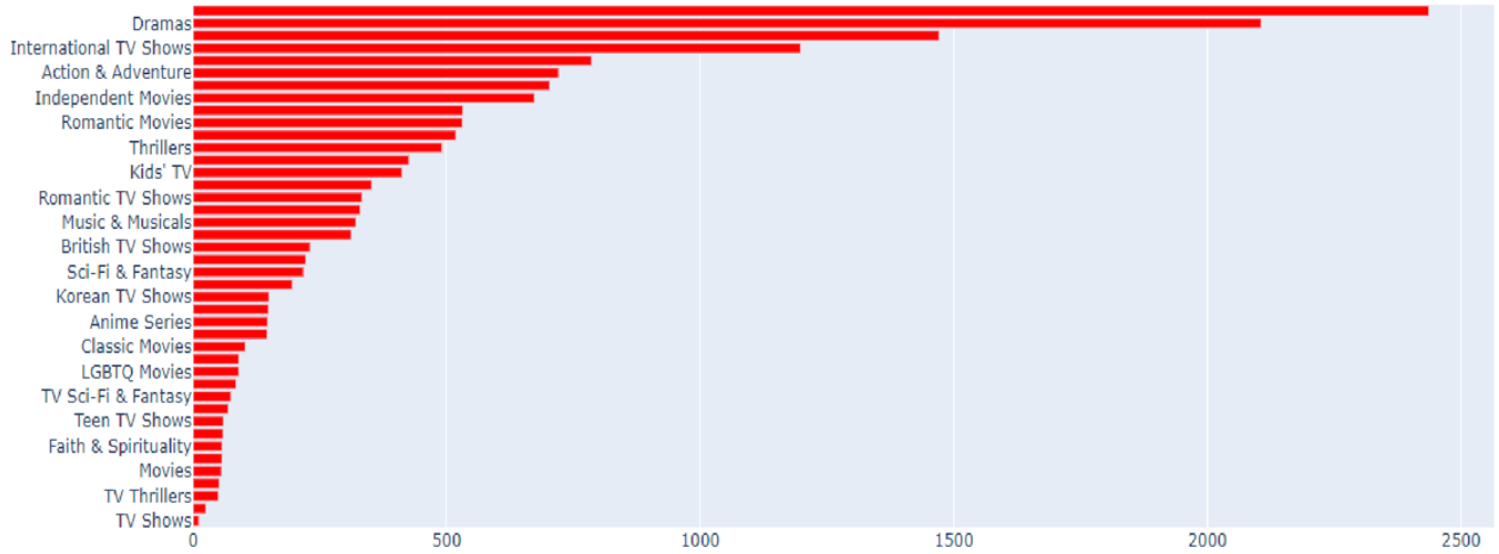
Exploratory Data Analysis





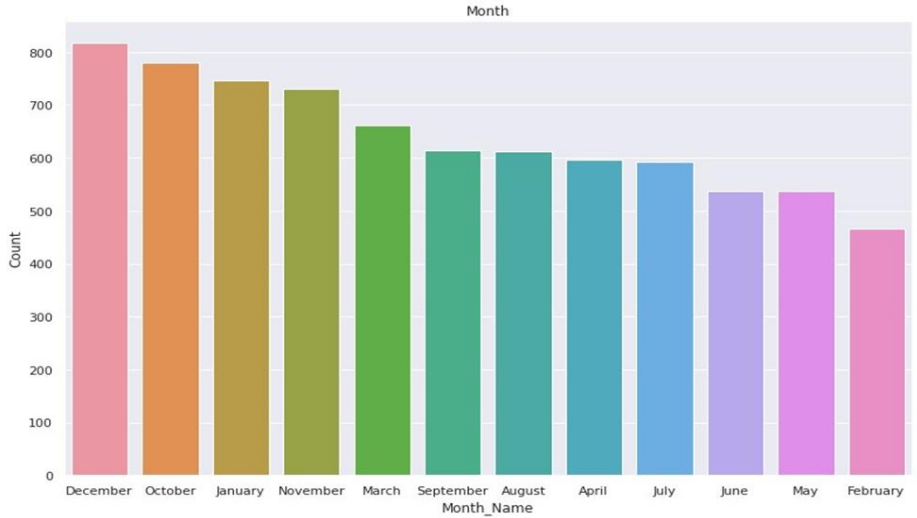
Exploratory Data Analysis

- Dramas and international tv shows are the top two contents.



Exploratory Data Analysis

- Here we can say that, In netflix maximum content added in December and minimum in february

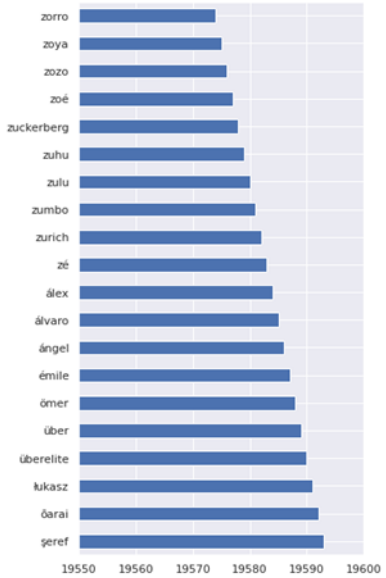


Data preprocessing

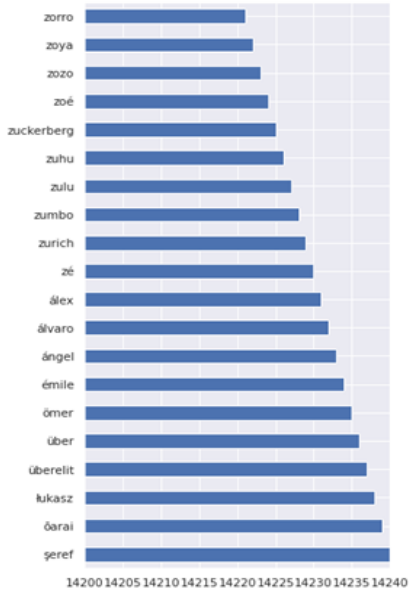
1) Working on the text based features (description, listed_in)
2) Removing punctuations and stop words from text features
3) Stemming process applied for those text features
4) Applying the TF-IDF vectorizer on those update text

Data preprocessing

Description column before stemming

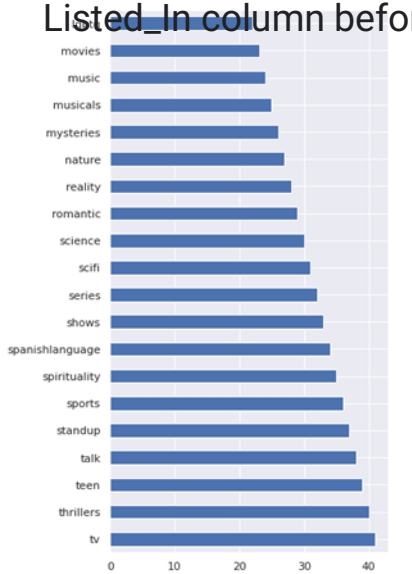


Description column after stemming



Data preprocessing

- Listed_In column before stemming



Listed_In column after stemming

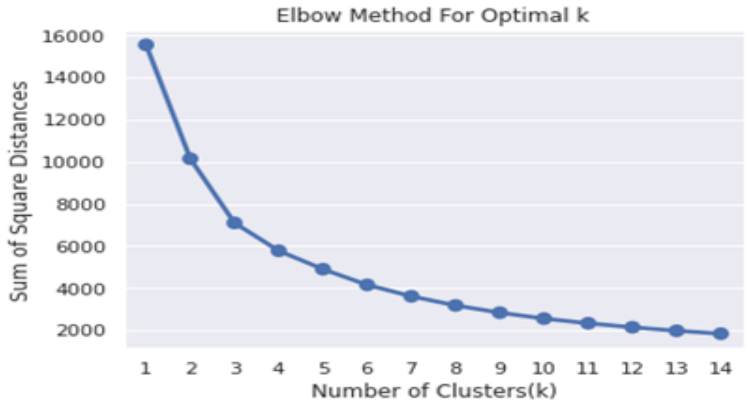
Implementing clustering methods

- **Silhouette Score Method (n range clusters) : Optimum score is 0.347 for n_clusters = 3**

```
For n_clusters = 2, silhouette score is 0.3366434960744673
For n_clusters = 3, silhouette score is 0.34802317407255573
For n_clusters = 4, silhouette score is 0.3195730674816924
For n_clusters = 5, silhouette score is 0.3068724493724028
For n_clusters = 6, silhouette score is 0.3280580079918211
For n_clusters = 7, silhouette score is 0.3276325599424118
For n_clusters = 8, silhouette score is 0.3205163824216981
For n_clusters = 9, silhouette score is 0.3221049828799175
For n_clusters = 10, silhouette score is 0.31638881237151417
For n_clusters = 11, silhouette score is 0.32735129784540845
For n_clusters = 12, silhouette score is 0.3283483265570967
For n_clusters = 13, silhouette score is 0.3248743240034122
For n_clusters = 14, silhouette score is 0.32536009493098533
For n_clusters = 15, silhouette score is 0.33046737498075635
```

Implementing clustering methods

- **Applying elbow method**

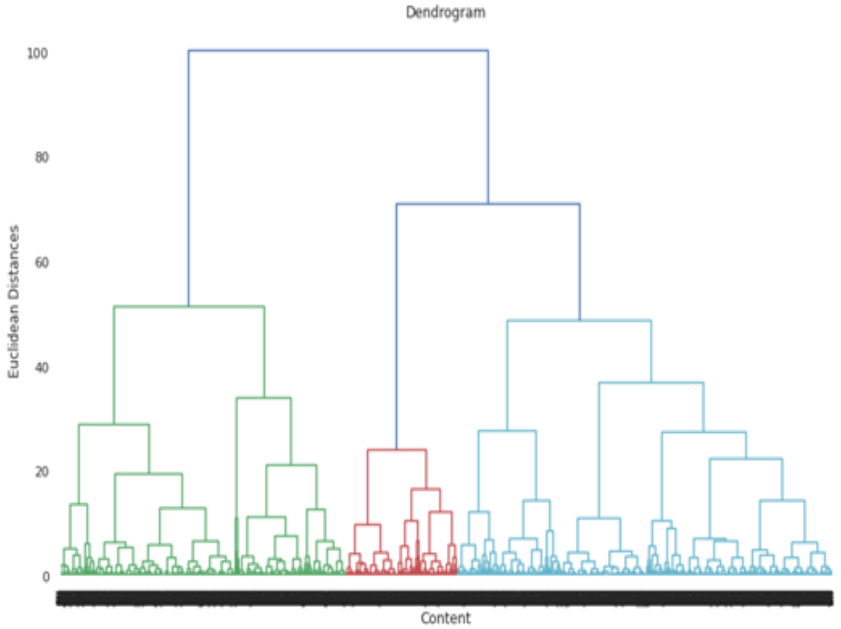


Applying KMeans clustering

- For k = 3

Applying Hierarchical clustering

- In Hierarchical Clustering, the aim is to produce a hierarchical series of nested clusters. A diagram called **Dendrogram** (A Dendrogram is a tree-like diagram that statistics the sequences of merges or splits) graphically represents this hierarchy and is an inverted tree that describes the order in which factors are merged (bottom-up view) or cluster are break up (top-down view).
- Dendrogram to find the optimal number of clusters. The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold. No. of Cluster is 3.

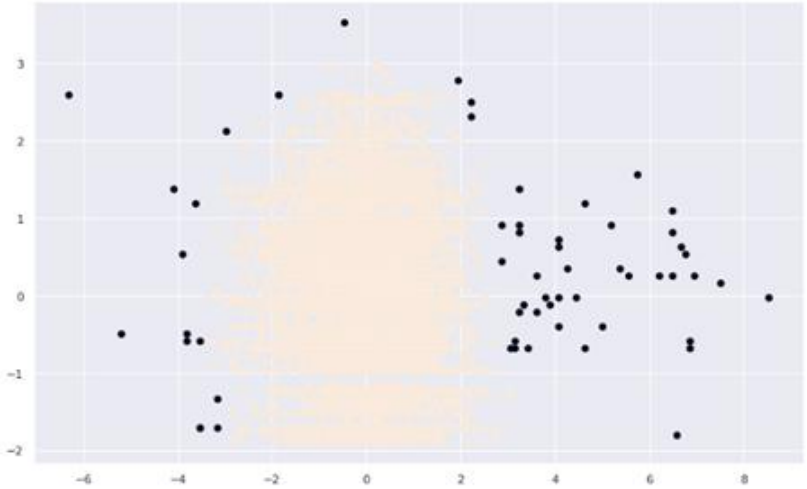


Applying Hierarchical(Agglomerative clustering)

- Optimal number of clusters are 3

DBSCAN

- Fundamentally, all clustering methods use the same approach i.e. first we calculate similarities and then we use it to cluster the data points into groups or batches.
- Here we will focus on **Density-based spatial clustering of applications with noise** (DBSCAN) clustering method.
- Clusters are dense regions in the data space, separated by regions of the lower density of points. The ***DBSCAN algorithm*** is based on this intuitive notion of “clusters” and “noise”. The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points.



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 and we have 10 features for the further implementation
- We have two types of content TV shows and Movies (30.86% contains TV shows and 69.14 % contains Movies)
- 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)
- The most number of the movies release in 2017 and TV shows in 2020 respectively and united nation have the maximum content on netflix
- On Netflix, Dramas genre contains the Maximum content among all of the genres and the most of the content added in december month and less content in february
- By applying the silhouette score method for n range clusters and we got the best score which is 0.348 for 3 clusters it means content explained well on their own clusters, by using elbow method we also got k cluster is 3
- Applied different clustering models Kmeans, hierarchical, Agglomerative clustering on data we got the best cluster arrangements
- By applying different clustering algorithms to our dataset .we get the optimal number of cluster is equal to 3

Future Work

From this clustering analysis we can create Netflix movies and tv shows recommendation system. With the help of this analysis Netflix can increase their use in different countries


ThankYou!