CAPSTONE PROJECT 4

Netflix Movies and T.V Shows Clustering

Created by Sakshi Dhyani

Project Details

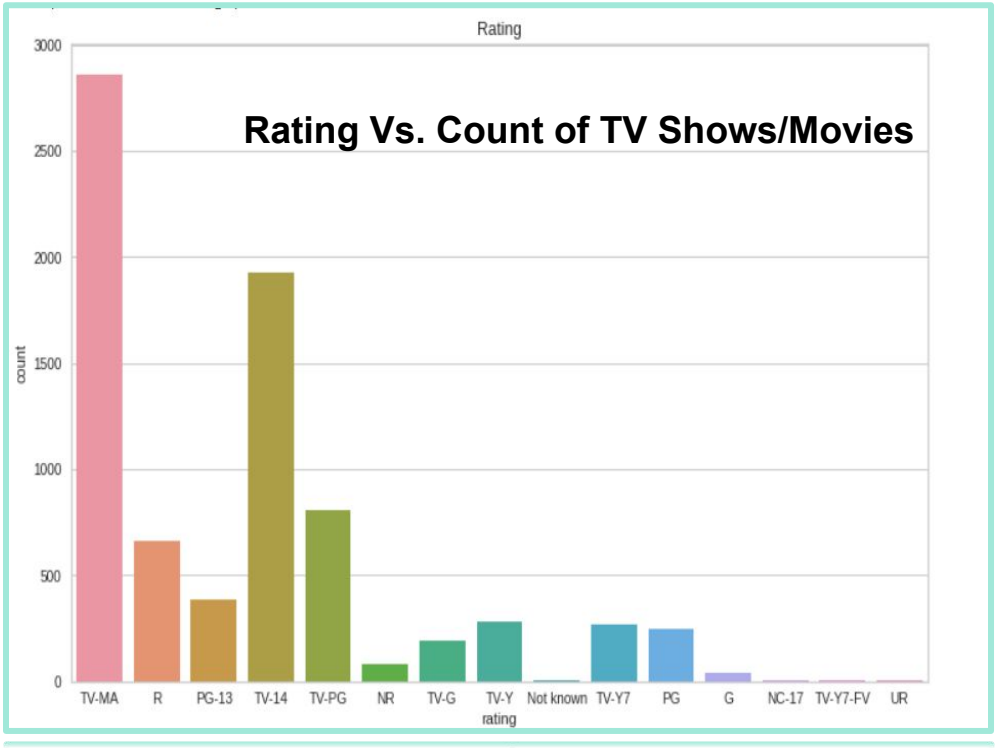
Netflix is a streaming service with wide variety of content. The idea of this project is to analyze and perform clustering to determine various patterns related to the content available in Netflix. The data is gathered from a third party engine.

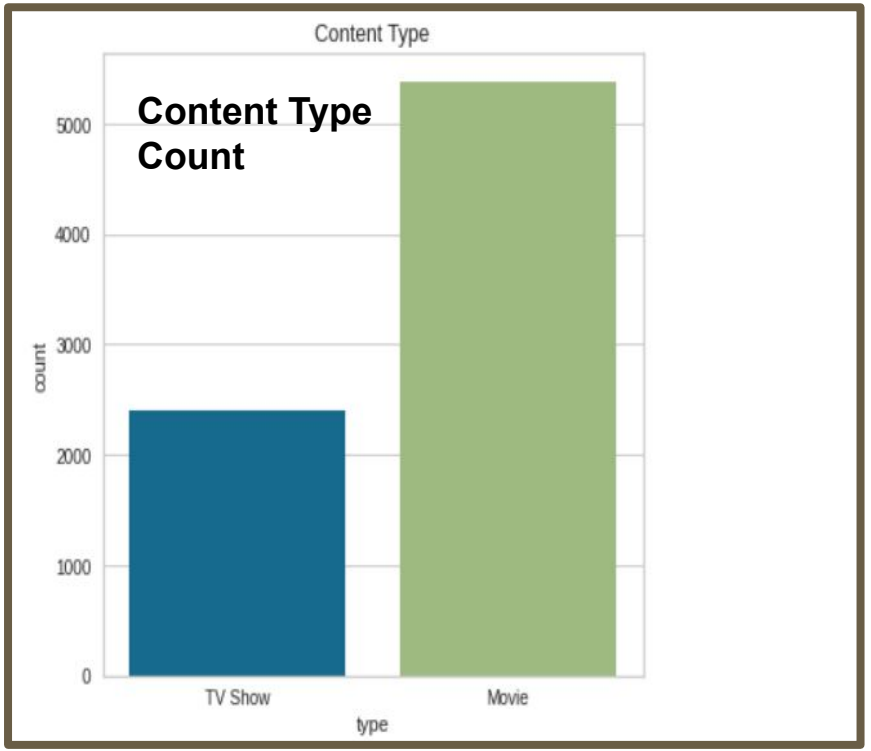
Based on the attributes related to the Tv shows or movies, we will be implementing different clustering algorithms which comes under unsupervised Machine learning category.

DataSet Details

- **show_id** : Unique ID for every Movie / Tv Show
- **type** : Identifier - A Movie or TV Show
- **title** : Title of the Movie / Tv Show
- **director** : Director of the Movie
- **cast** : Actors involved in the movie / show
- **country** : Country where the movie / show was produced
- **date_added** : Date it was added on Netflix
- **release_year** : Actual Release year of the movie / show
- **rating** : TV Rating of the movie / show
- **duration** : Total Duration - in minutes or number of seasons
- **listed_in** : Genre
- **description**: The Summary description

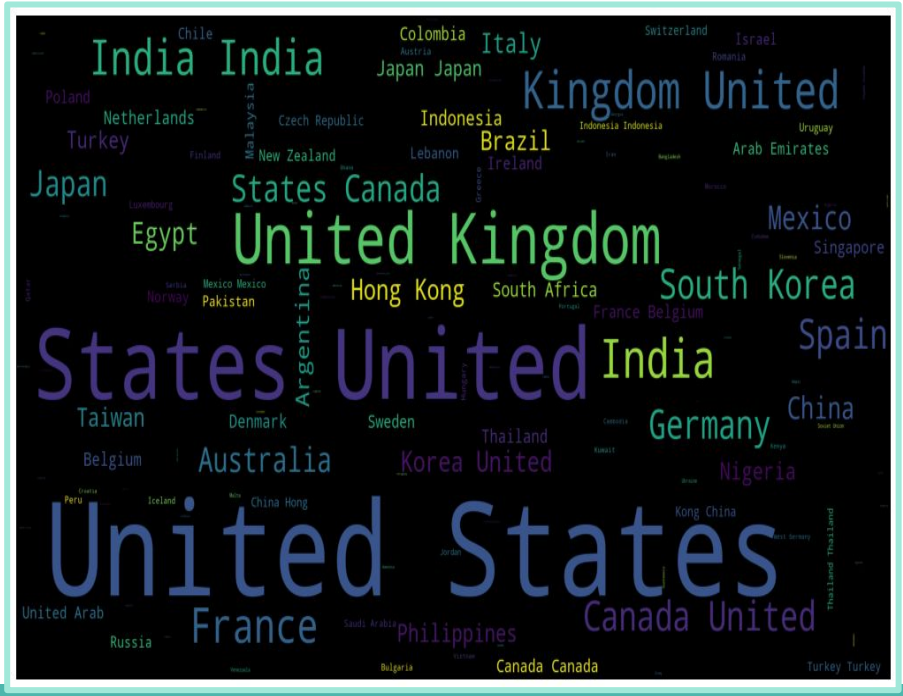
Exploratory Data Analysis





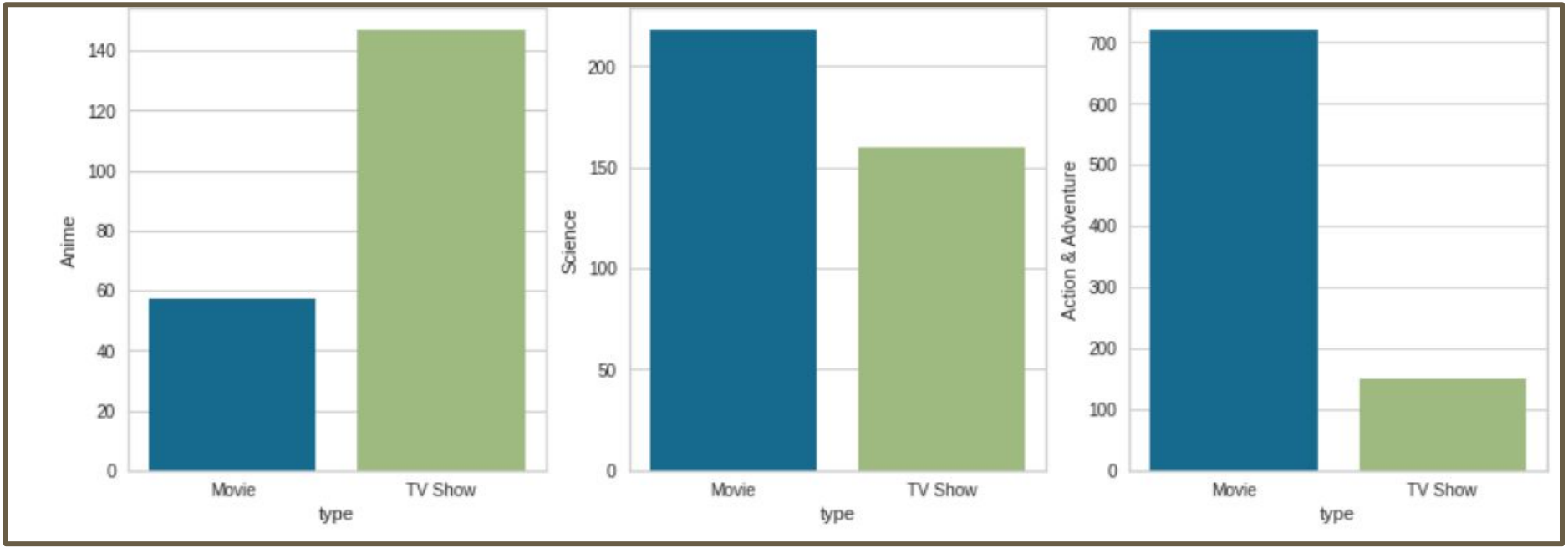
Exploratory Data Analysis

Word cloud for genres and different countries for the content in netflix



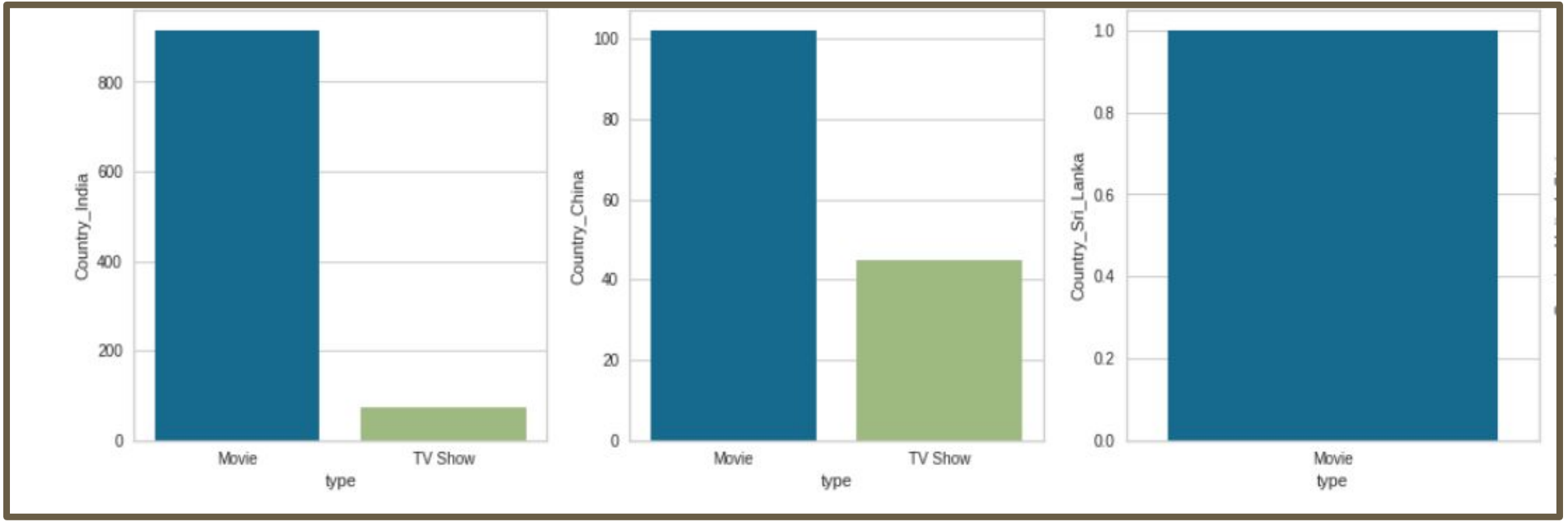
Bar Plots for different genres based on content type

Selectively showing bar plots for 3 genres Anime, Science, Action & Adventure here. Similarly for all genres, bar plots were plotted

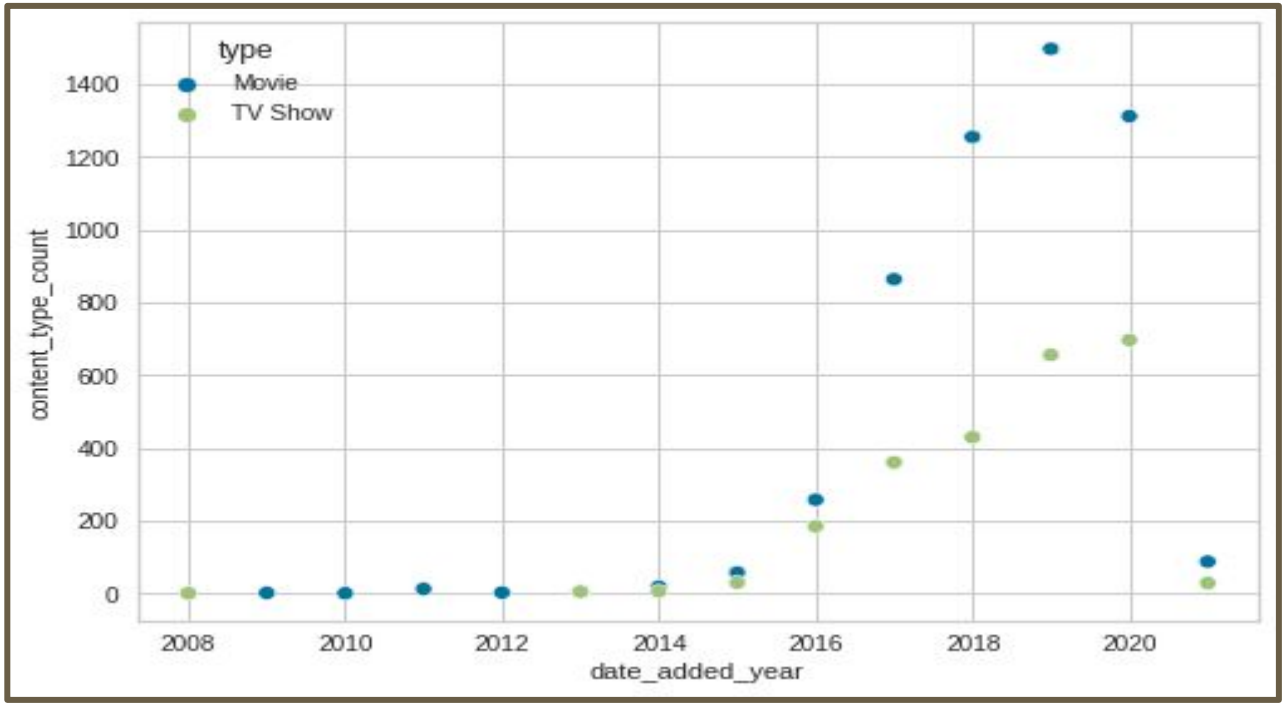


Bar Plots for different countries based on content type

Selectively showing bar plots for countries India, China and Sri Lanka. Similar bar plots were plotted for other countries as well.



Scatter Plot for content type count based on year when content was added to netflix



Implementation of Clustering Algorithms

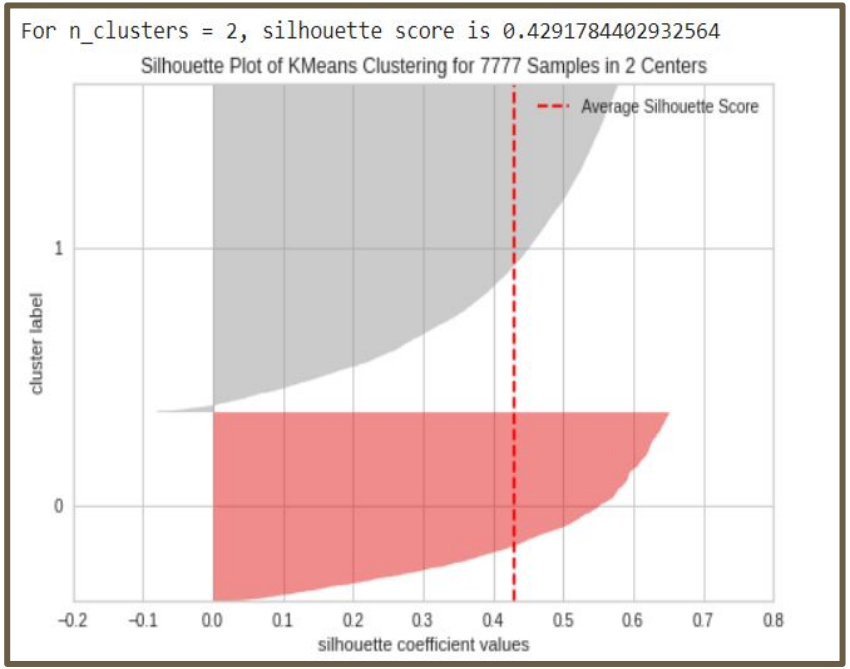
Data was converted into numerical data using label encoder. Standardization and normalization was performed on the data as well. Applied PCA, to perform dimensionality reduction. Data was converted to 2-dimensions.

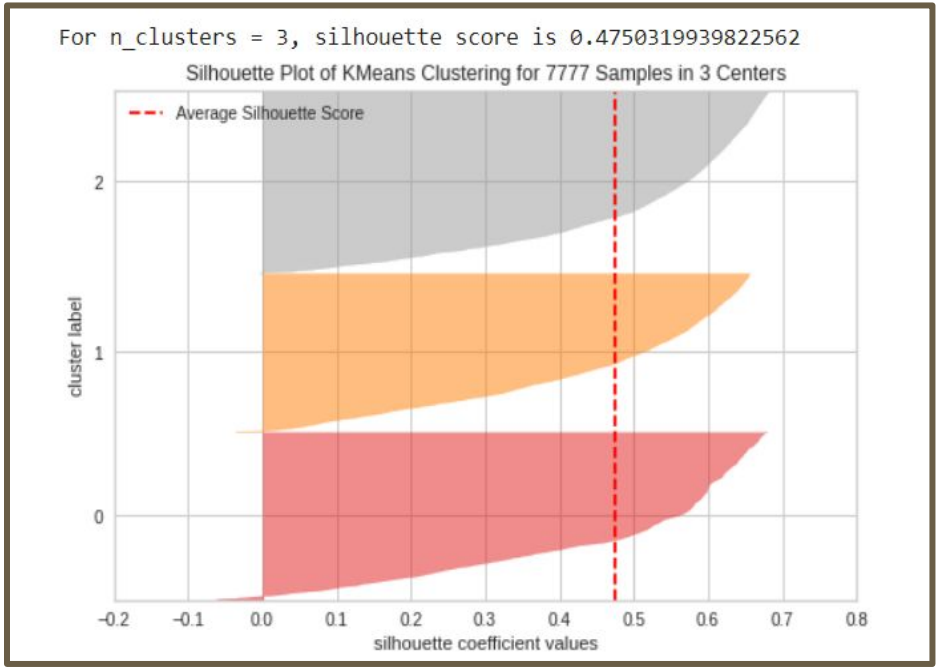
Clustering algorithms used:

- K-Means Clustering
- Gaussian Clustering
- Agglomerative Clustering

K-Means Clustering

Created silhouette score analysis plot for different number of clusters. Here showing only for 2 and 3 number of clusters.

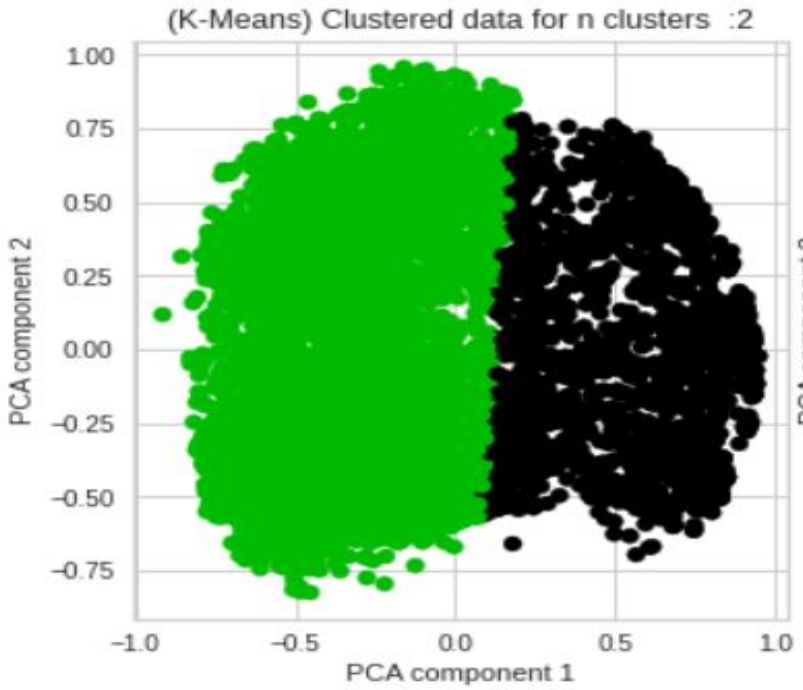


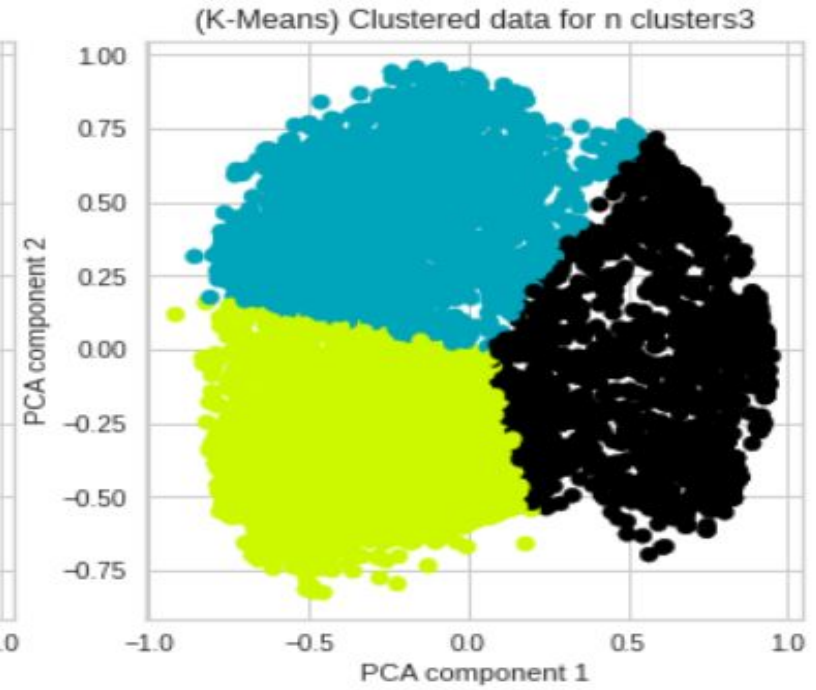


Clusters formed for K- Means

Here showing plot for n_clusters 2 and 3

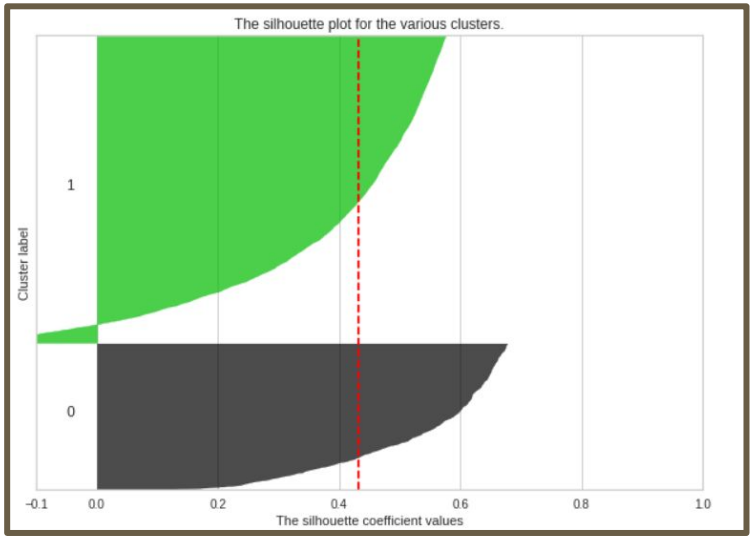
Silhouette score for no of clusters 3 is maximum in case of K-Means clustering.

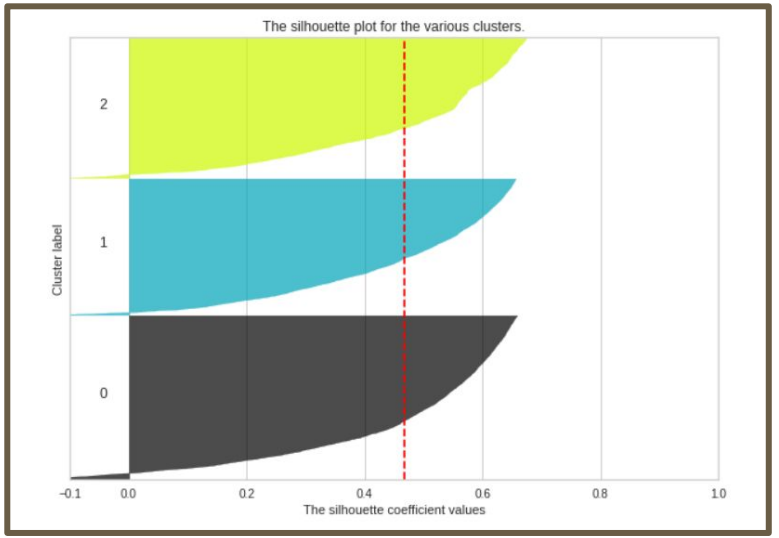




Gaussian Clustering

Created silhouette score analysis plot for different number of clusters. Here showing only for 2 and 3 number of clusters.

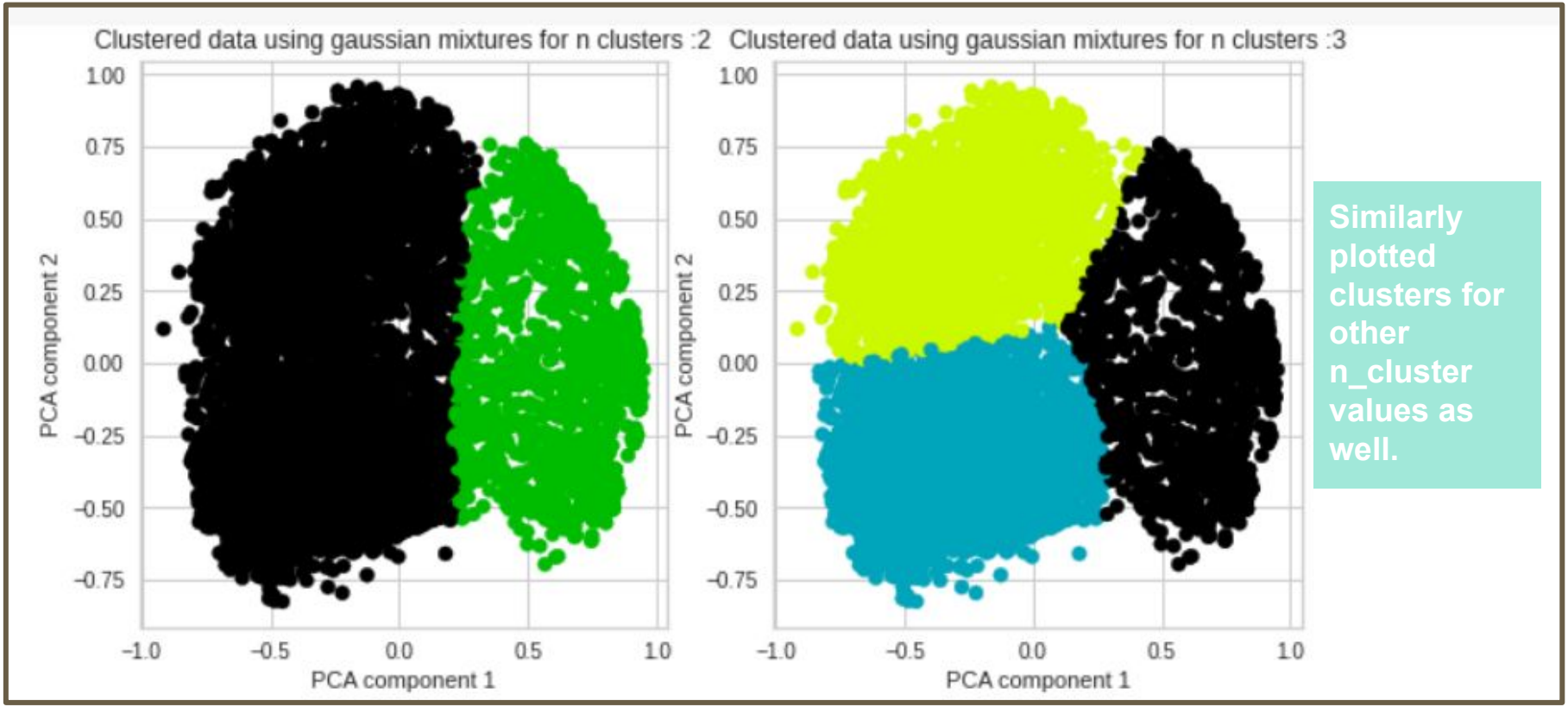





```
For n_clusters = 2 The average silhouette_score is : 0.43126921807594
For n_clusters = 3 The average silhouette_score is : 0.46807048338990626
For n_clusters = 4 The average silhouette_score is : 0.3994123519417917
For n_clusters = 5 The average silhouette_score is : 0.3988816772966538
For n_clusters = 6 The average silhouette_score is : 0.4004766201837277
For n_clusters = 7 The average silhouette_score is : 0.38913510941342827
```

Showing silhouette analysis plot for n_clusters 2 and 3 in this slide. These are plotted for other n_cluster values as well.

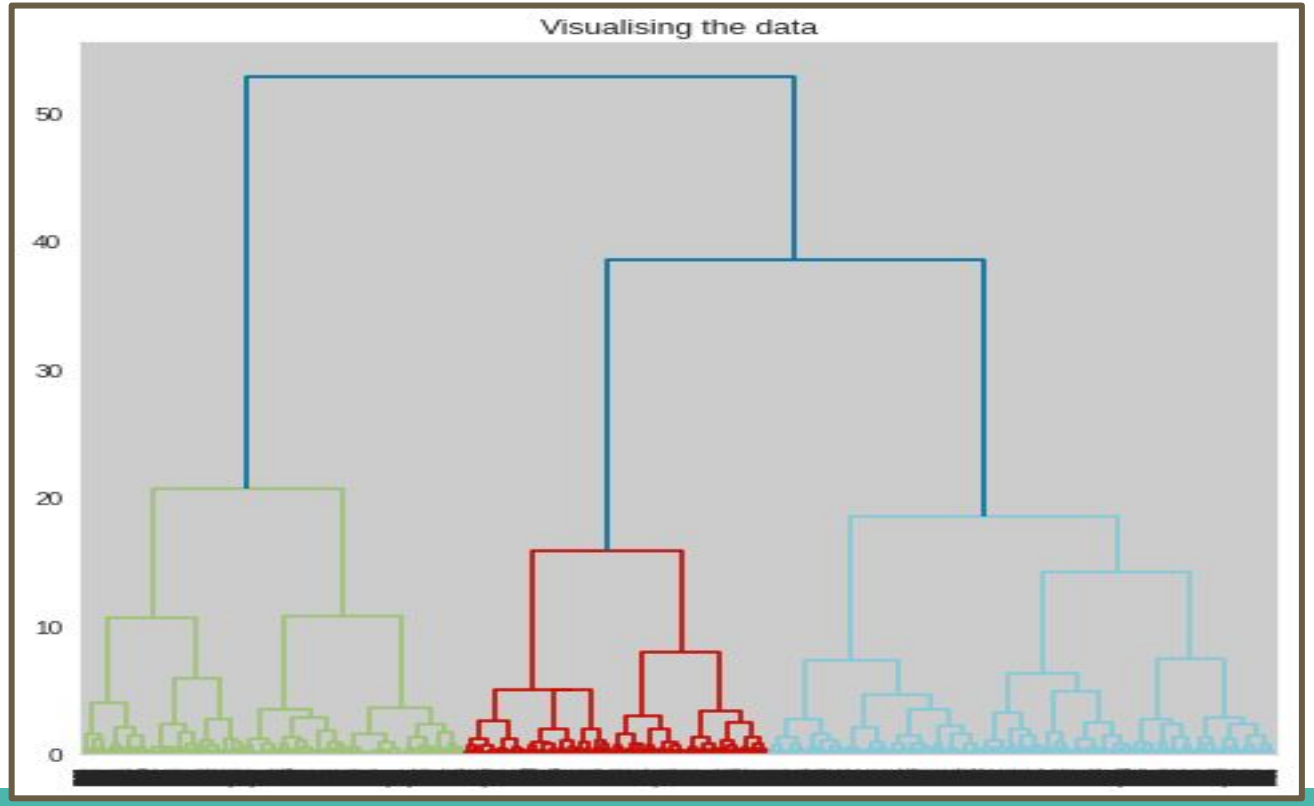
Clusters formed for Gaussian Clustering



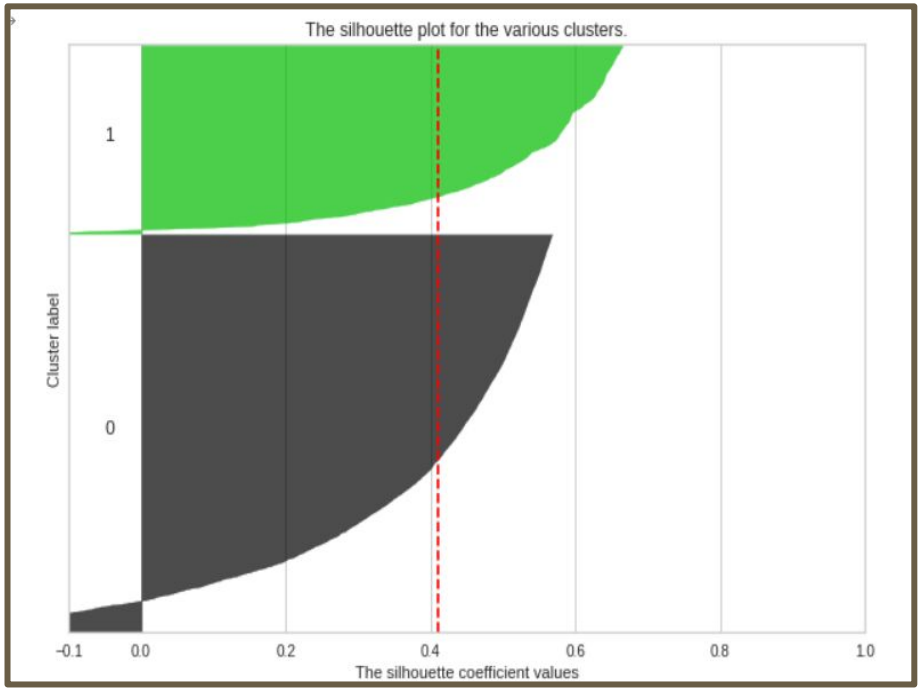
Similarly plotted clusters for other n_cluster values as well.

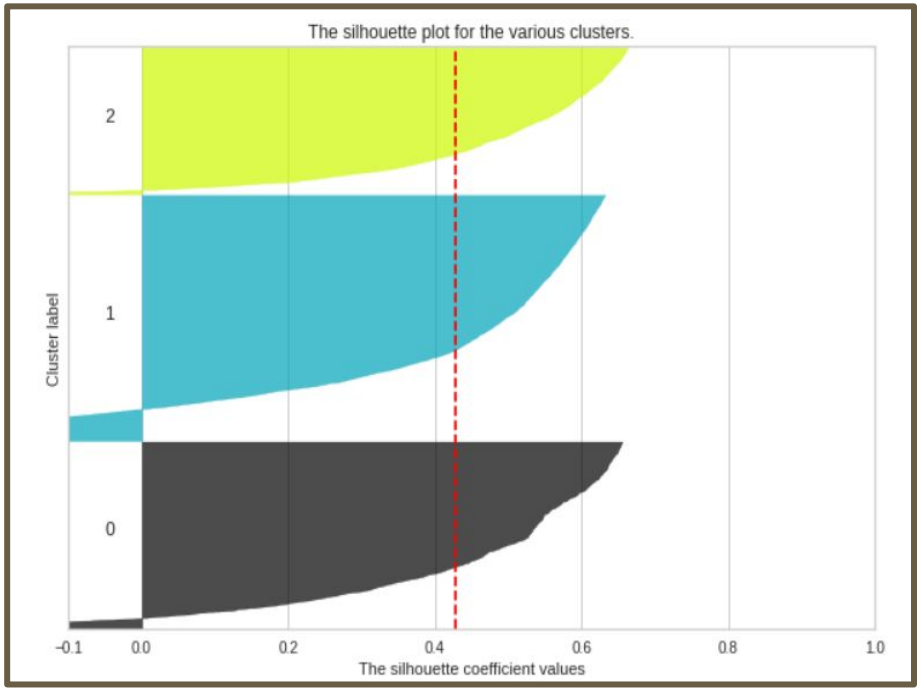
Agglomerative Clustering

Dendrogram for visualizing the data

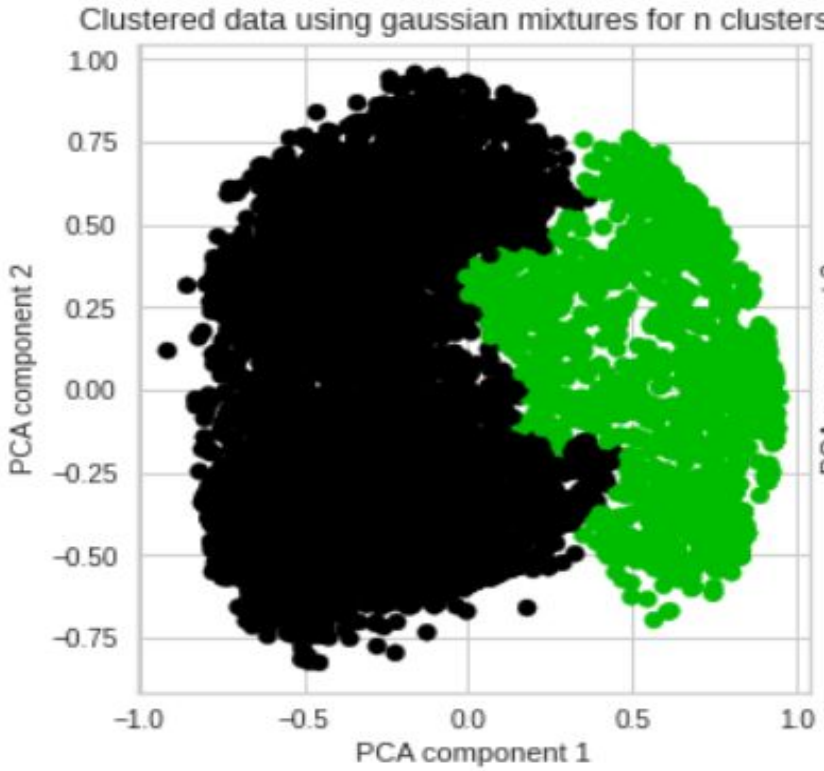


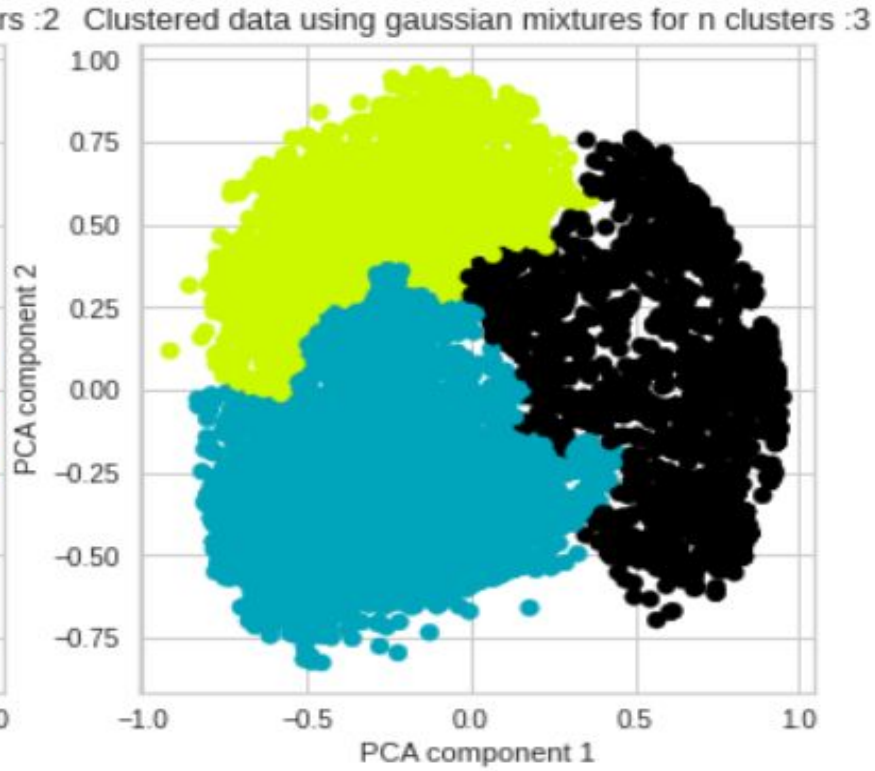
Created silhouette score analysis plot for different number of clusters for Agglomerative. Here showing only for 2 and 3 number of clusters.





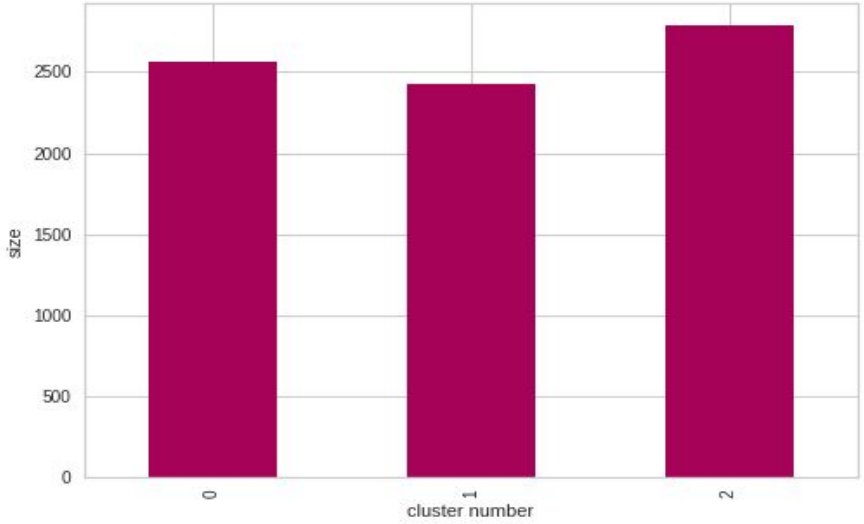
Clusters formed for Agglomerative clustering

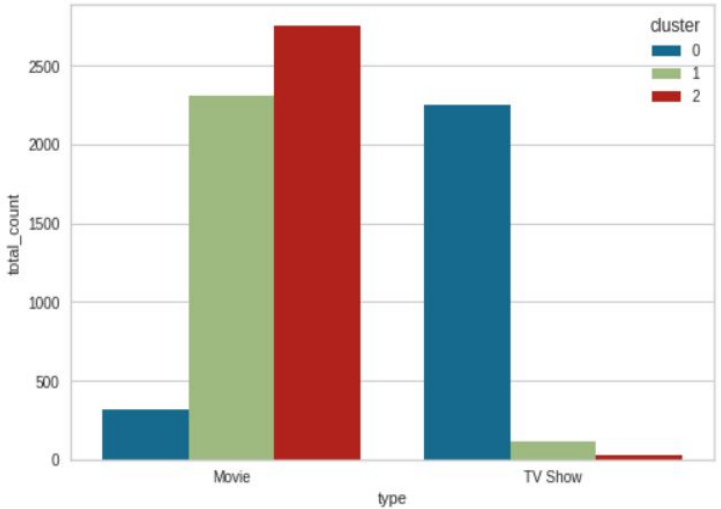




Cluster Analysis

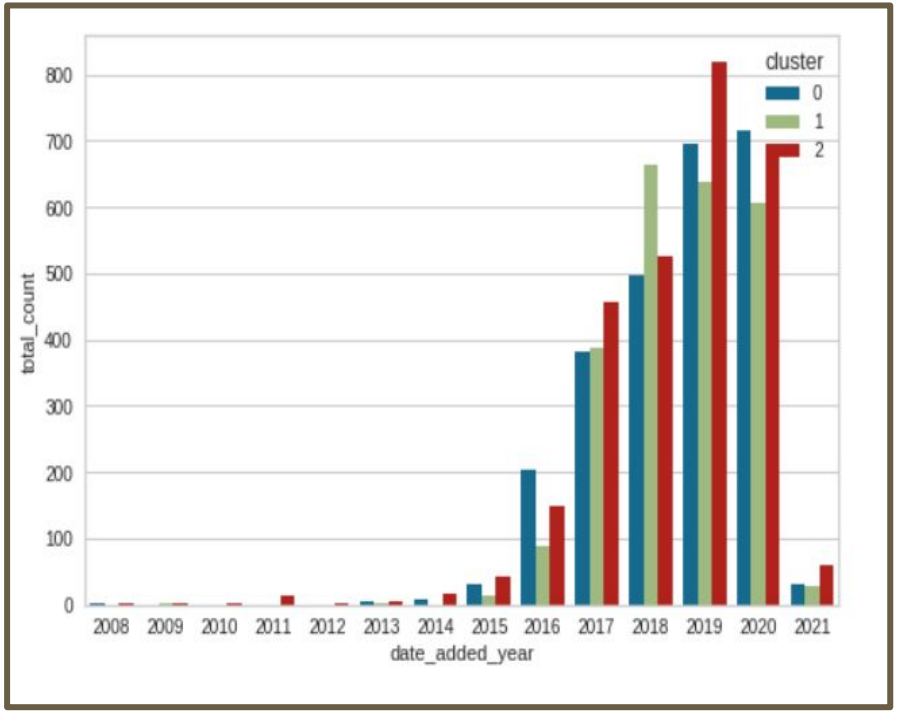
Final Model Implemented was K-Means with n_clusters =3 as it was having high silhouette score value as compared to other cases. Different graphs were plotted to understand the cluster based differences between the features. First graph shows cluster size. Second graph shows cluster wise count of movie and tv shows in Netflix.

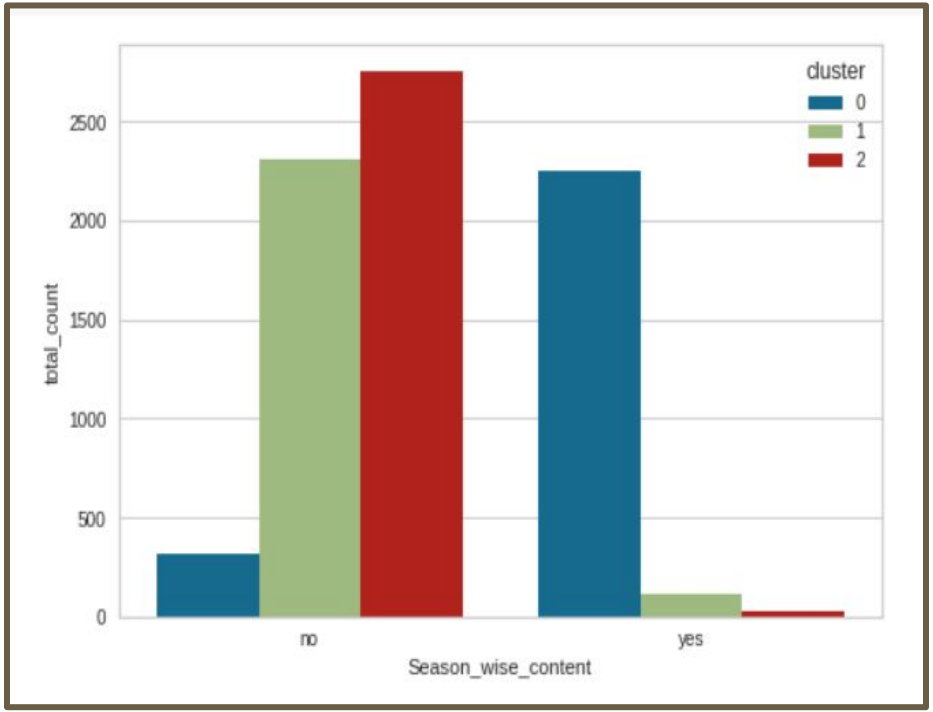




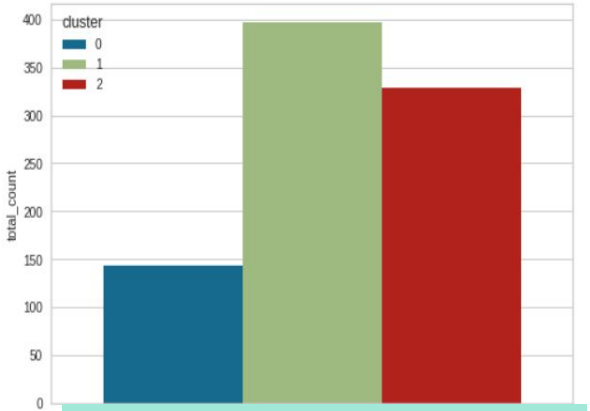
First graph shows year when content was added for different clusters.

Second graph shows cluster wise , total count of season wise content.

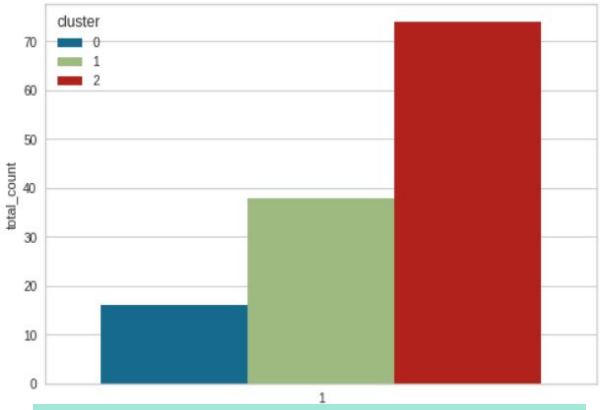




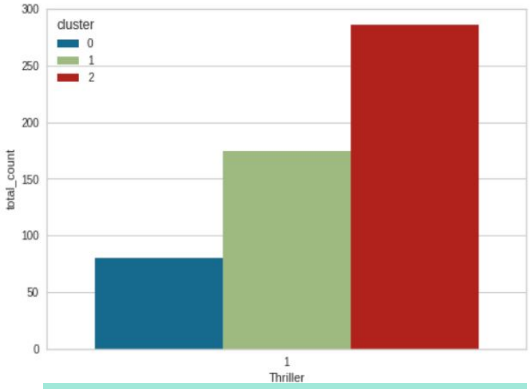
Number of Movies and Tv shows for different genres in different clusters



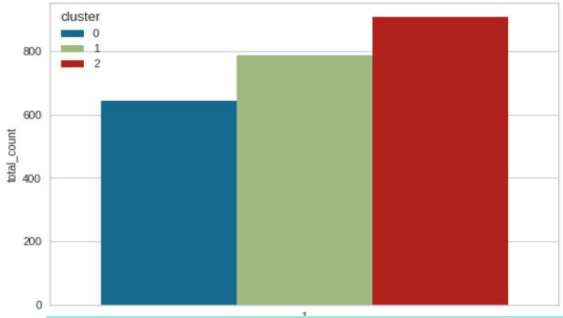
Action & Adventure



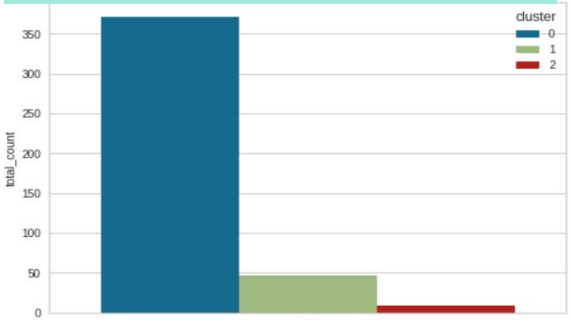
Classic



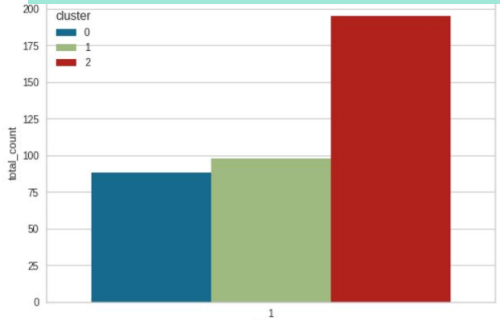
Thriller



Comedy



Crime



Horror

Conclusion

- ➔ EDA was performed.
- ➔ Standardization was performed for rescaling the data. Normalization was performed to create better clusters.
- ➔ PCA was implemented to reduce dimensionality.
- ➔ Agglomerative clustering, K-means clustering and Gaussian Clustering was implemented.
- ➔ K Means Algorithm was selected with n clusters 3 at the end.
- ➔ Silhouette score analysis helped us to understand the measure of how close each point in one cluster is to points in the neighbouring clusters. Clusters formed were plotted for visualization purpose.
- ➔ Then after implementing the most appropriate or the better one clustering algorithm out of all as per the silhouette score value, features for different clusters were compared.

THANK YOU