Summary

Name : Raja Chowdhury

Gmail : rajachowdhury2468@gmail.com

GitHub Repo link - https://github.com/RajaChowdhury/Netflix-Movies-and-TV-Shows-Clustering---Capstone-Project.git

PROBLEM

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. Integrating, this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

APPROACH

Initially, in the 1st step imported the data set to carry out the analysis over the data set to comprehend the details of available data and Checked for Null values and treated them. Here, we found more than 30% null values in the director's column. Then, we take appropriate action for null values according to the circumstances.

Performed the Exploratory data analysis and tried to get the understanding of the data and how the content is distributed in the dataset, its type and details such as which countries are watching more and which type of content is in demand etc. has been analyzed in this step with the help of visualization graph by getting insights from analysis.

- Data preprocessing – in this we remove the punctuation and stops words also used stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language.
- We used the k-means clustering algorithm and then checked the model performance using Silhouette's coefficient and elbow method to find the number of clusters.

Analyzing all the variables of the data set and identifying the solution for given tasks. Performed hypothesis testing to get the insights on duration of movies and content with respect to different variables. After doing feature engineering and finding the number of clusters, we used the k-means algorithm and then checked the model performance using Silhouette's coefficient, to identify the best fit Model. The number of movies on Netflix is growing significantly faster than the number of TV shows. Because of covid-19, there is a significant drop in the number of movies and television episodes produced after 2019.

CONCLUSION

- Director and cast contain a large number of null values so we will drop these 2 columns.
- In this dataset there are two types of contents where 30.86% includes TV shows and the remaining 69.14% carries Movies.
- Maximum number of contents are released in winter season that might be because of the festival season li ke Christmas and due to which "Christmas" word is more occurred in the tittle of the contents.
- From the dataset insights we can conclude that the maximum number of TV Shows released in 2017-18 and for Movies it is 2019-20.
- On Netflix USA has the largest number of contents. And most of the countries preferred to produce movies more than TV shows.

- Most of the movies are belonging to 3 genre categories.
- TOP 3 content categories are international movies, dramas, comedies.
- In text analysis (NLP) I used stop words, removed punctuations, stemming & TFIDF vectorizer and other functions of NLP.
- Applied different clustering models like Kmeans, hierarchical, Agglomerative clustering, DBSCAN on data we got the best cluster arrangements.
- By applying different clustering algorithms to our dataset, we get the optimal number of clusters is equal to 3.