Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Jai Harish S (ppjairam2@gmail.com):

a. Handling Null Values
b. Feature engineering
c. EDA
d. TFIDF
e. Evaluation Metrics
f. Conclusion

Bhaskar subanji (bysubanji@gmail.com):

a. Basic analysis
b. EDA
c. Silhouette score
d. Hypothesis testing
e. Implementation of PCA
f. Conclusion

Pranil Satish Thorat (pranilthorat@gmail.com):

a. EDA
b. Elbow method
c. Silhouette score
d. Implementation of K Means clustering
e. Interactive scatterplot of the cluster
f. Conclusion

Please paste the GitHub Repo link.

GitHub Link: https://github.com/bysubanji/Netflix_movies_and_Tv_shows_clustering

Drive link:
https://drive.google.com/drive/folders/1XK0F3-49eTpA6lLWRu9zGox6qBOzt3Qr?usp=sharing

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

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 also Checked for Null values and treated them.
- Analyzing all the variables of the data set and identifying the solution for given tasks.
- Performed the Exploratory data analysis and tried to address the given tasks with the help of visualization graphs by getting insights from analysis.
- 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, Calinski-harabasz score and Davies-Bouldin index to identify the best fit Model.

CONCLUSION

- We've done null value treatment, feature engineering, and EDA since loading the dataset then completed assigned tasks.
- Anupam Kher has acted in more Indian films than anyone else and the United States and India are the two countries where Netflix is most popular.
- Concluded that Netflix is increasingly focusing on movies rather than TV shows, especially after 2014.
- Among different types of content available in different countries, content TV-MA is available in the majority of countries. This could be because it shows that it is just for adult audiences, and the Netflix audience enjoys content like this.
- We've also explained different clusters based on their content; Defined 28 clusters and enforced the K-means clustering algorithm and cluster number nine has the most clusters; we've also plotted a scatter plot in which we may interact with similar content about that cluster.