Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Team Member Name: Pratik m Gumble Email: pratikmgumble@gmail.com Contribution: Individually
Please paste the GitHub Repo link.
Github Link:- https://github.com/pmgumble/pmgumble-Netflix-Movies-TV-Shows-clustering
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 Statement: Dataset consists of tv shows and movies available on Netflix as of 2019. 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. Understanding what type content is available in different countries Is Netflix has increasingly focusing on TV rather than movies in recent years. **Task** 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 **Approaches:** 1. Data Inspection 2. Data processing a. NaN values detection and conversion depending upon their size b. Data Encoding 3. EDA

4. Feature Selection
5. Data Standardization
6. Apply Principal Component Analysis (PCA) for data reduction
7. Apply K Means clustering
8. Find out Proper value of K using K Elbow and Silhouette score methods.
9. Apply Hierarchical Clustering
10. Apply Agglomerative Clustering

Conclusion: We have drawn many interesting inferences from the dataset Netflix titles; here's a summary of the few of them:

- The most content type on Netflix is movies. It appears that Netflix has focused more attention on increasing Movie content than TV Shows. Movies have increased much more dramatically than TV shows
- There are about 69.1% movies and 30.9% TV shows on Netflix.
- Most films were released in the years 2018, 2019, and 2020.
- The number of releases have significantly increased after 2015 and have dropped in 2021 because of Covid 19.
- The months of October, November, December and January had the largest number of films and television series released.
- More of the content is released in holiday season - October, November, December and January.
- The United States has the highest number of content on Netflix by a huge margin followed by India.
- Raul Campos and Jan Sulter collectively have directed the most content on Netflix.
- Anupam Kher has acted in the highest number of films on Netflix. Drama is the most popular genre followed by comedy.
- International movies are the top most genre in Netflix which is flowed by standup comedy and Drams.
- Most of the movies have duration of between 50 to 150
- Highest number of tv_shows consisting of single season
- Using correlation heatmap we see that in India mostly teens watching netflix so question arises that what content teens watched.
- TV-MA has the highest number of ratings for tv shows i,e adult ratings
- In India teens mostly watched international movies.
- Principal Component analysis (PCA)reduced the number of components as 7 with approximately 99% of variance.
- For K Means clustering to find out number of k we used elbow and sillhoute score method.
- Using both the methods we found k=3 is optimal value of clustering.
- Using Hierarchical clustering method again we find out that k=3 is optimal value of clustering.