Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Name: Aaksh A. Salmuthe **Email:** akashsalmuthe30@gmail.com • Data Summary • Data Description • Data Analysis • Data Cleaning • Data Pre Processing • Exploratory Data Analysis • K Means Clustering: 1. Vectorization 2. Elbow Curve 3. Silhouette score • Recommender System • Conclusion
Please paste the GitHub Repo link.
Github Link:- https://github.com/AkashSalmuthe/NETFLIX-MOVIES-AND-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:** 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:** • 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 and elbow method.
- Also performed recommender system with cosine similarities.

Conclusion:

- Movies uploaded on Netflix are more than twice the TV Shows uploaded.
- TV shows and movies are increasing continuously but in 2020 there is drop in number of movies.
- From October to January, maximum number of movies and TV shows were added.
- Maximum number of movies and TV shows were either on start of the month or mid of the month.
- United State tops in the list of maximum number of movies and TV shows followed by India, UK and Japan.
- Maximum of the movies as well as TV shows are for matures content.
- Anupam Kher top from the list of casts having maximum number of movies and TV shows
- Majority of movies have running time of between 50 to 150 min
- Almost 68 of TV shows consist of single season only
- Top 3 genres are exactly same for movies and TV shows
- Dramas genres hit all over the world
- 30 movies and 50 TV shows are Netflix Originals
- Clustering done by K Means Clustering, found optimal number of clusters equal to 25 with highest Silhouette Score.
- Recommender system using cosine similarity performs well on data.