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's Name, Email and Contribution:
 	
Contributor:
 	
• Name: Ajith R
• Email: rubeshajith@gmail.com
• Integrating IMDB dataset
• Data Pre-processing
  • Null Value check
  • Column Rename
  • Creating new variables
  • Column removal
• Exploratory Data Analysis
• Hypothesis Testing
• NLP operations
• Principal Component Analysis
• UMAP
• DBScan
• Hierarchical Clustering
  • Agglomerative
• K Means Clustering
• Evaluation metrics:
  • Elbow Method
  • Silhoutte score

Please paste the GitHub Repo link.
Github Link:- https://github.com/rubeshajith/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)

First the integration of IMDB datasets were done. Then data preprocessing was done like null value check, changing column name, creating new columns etc. Next Exploratory Data Analysis was done. In clustering part we did cluster by description column for Hypothesis Testing purpose and other text based features separately. In Clustering for Hypothesis testing we used NLP techniques like stop words, removed punctuations and TF-IDF vectorizer for data prep, further we used PCA for Dimensional reduction. In Hypothesis Testing after executing clustering part I picked one in each of the top IMDB rated TV show and movie to compare them and conclude which had the top rating in each clusters, so we found out that in the total of 26 clusters TV show's IMDB rating was high in 23 clusters, further making assumptions in hypothesis testing we declared that people tend to watch TV show more than movies. In second Clustering part we used UMAP as dimensionality reduction and implemented various models like DBSCAN, hierarchical, Agglomerative clustering, K-means with Elbow , K-means with silhouette and found best number of clusters accordingly.