**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

| **Team Member’s Name, Email and Contribution:** |
| --- |
| Name : V. Bhavya Reddy  Email : [bhavya.reddy0711@gmail.com](mailto:bhavya.reddy0711@gmail.com)  Contribution: Entire project |
| **Please paste the GitHub Repo link.** |
| Github Link:- <https://github.com/coolphotography/NETFLIX-MOVIES-AND-TV-SHOWS-CLUSTERING.git> |
| **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:**   1. I started by replacing Nan values with No Director, No Cast, and Country Not Available for director, cast, and country features respectively. Nan values were dropped from date\_added and rating features. 2. By using the 'date\_added' feature, we can obtain the features 'month\_added' and 'year\_added'. 3. As a first step, we compare the top 20 countries by type - Movies vs TV Shows. From the dataset, we select country-specific data and compute LDA and Document Term Matrix(DTM). Several plots were taken to evaluate the available content for specific countries. There were several conclusions made regarding the United States, as an example. 4. In the feature engineering - the rating values were reassigned, ordinal encoding was used on the type, the string value was dropped from the duration, etc. 5. K-Means Clustering was performed on type, director, cast, country, year\_added, month\_added, release\_year, rating, duration, listed\_in features. The silhouette\_score was used to calculate the number of clusters. A test on '13 Reasons Why' was then conducted to determine the recommendations. 6. In NLP, we used only the description to determine recommendations. 7. A K-means clustering dataset with TF-IDF vectoriser from NLP is then combined for clustering. As a result, better recommendations were made.   **Conclusion:**   1. A recommendation system with the description column works well. 2. In the case of K-means, the optimal number of clusters are 12. 3. When K-means is applied to the description sum column, the optimal number of clusters was 12. 4. Clustering with the description column had better recommendations than clustering without description. 5. The optimal number of clusters was calculated using silhouette\_score.   **EDA Observations:**   1. The most content type on Netflix is movies. 2. The largest count of Netflix content is made with a 'TV-MA' rating. 3. After 2014 the amount of content added has been increasing significantly. 4. The number of movies in 2020 have reduced compared to the previous year. However, the number of TV shows has increased. 5. While most TV seasons have only 1 season, movie lengths follow a normal distribution with a mean of 100 minutes. 6. According to the amount of content produced, the United States is the top country. 7. International Movies are a genre mostly found on Netflix. 8. In terms of titles, Jan Suter is the most popular director on Netflix. 9. Anupam Kher is the most popular Netflix cast member, according to the number of movies made. 10. In 2018, 2019, and 2020, the majority of films were released. 11. A large number of movies and TV Shows were released in October, November, December, and January. |