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 Role:-

- Ajinkya dakhale :
  Email- Ajinkya.dakhale2408@gmail.com
  Contribution :
  - data loading
  - data summary
  - EDA
  - Understanding what type content is available in different countries
  - stemming
  - technical doc

  Harshjyot singh:
  Email-hs9158695878@gmail.com
  Contribution:
  - handling Null values
  - EDA analysis Bi variate
  - Finding if netflix has increasingly focused on TV shows rather than movies in recent years.
  - Removing punctuation
  - summary

  Suvir kapse:
  Email- suvirkapse@gmail.com
  Contribution:
  - Data cleaning
  - EDA
  - vectorization
  - applying k means clustering
  - ppt

Please paste the GitHub Repo link.

Github Link:- https://github.com/Harshjyot-Singh-Chawla/-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)

Introduction :

One of the top OTT platforms, both domestically and outside, is Netflix. Netflix oversees a vast library of TV series and films, which are available for internet streaming at any time. The diversity of content and proper user recommendations are two factors that determine the success of OTT platforms. As a result of people paying a monthly fee to access the site, this business is lucrative.

PROBLEM STATEMENT:

This dataset consists of TV shows and movies available on Netflix as of 2019. The dataset is collected from Fixable 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.
The main objectives of this study are to perform exploratory analysis and discover useful insights from the dataset, to comprehend the types of content that are offered in various countries, to determine whether Netflix has been increasingly focusing on TV rather than movies in recent years, and finally to perform clustering of related content by comparing text-based features from the dataset.

APPROACH:

- The first step includes loading of dataset and then inspecting the data through which we get to know the summary or description of data, shape of data, null value count,duplicate values in the data and about the data types of column.We found that there are null values in some column. We then did imputations of filling features with 'unkown' value which has more than 30% of null values such as in country , director and cast column.
- Second step includes doing EDA and data visualization onall categorical features and obtain their relationships by doing univariate and multivariate analysis. Here, we discovered that Netflix's TV material is much less prevalent than its movie content. We can see that most of the Netflix content is meant for grownups. Kids and teenagers have less content available to them. Next, we conduct analysis to provide answers to questions like: Is Netflix increasingly focusing on TV rather than movies in recent years?, Understanding what kind of content is available in various countries.
- Data preparation involves stripping out punctuation and commas as well as using stemming to break down words into their most basic forms, which may or may not be recognised as words by the language.
- After doing feature engineering, we used the k-means clustering approach to verify the model's effectiveness. We then determined the number of clusters using the elbow method and Silhouette's coefficient.

CONCLUSION:

- We've done null value treatment, feature engineering, and EDA since loading the dataset,then completed assigned tasks.
- Till 2019 we can see there is gradual rise in both movies and TV show but in 2020 we can see that is a bit drop in movies whereas the growth of TV shows remains the same.So yes we can say Netflix has increasingly focusing on TV rather than movies in recent years.
- The majority of countries have access to adult content programming, which is one of the several categories of content that are available in various nations. This might be the case given that it clearly states that it is just for adult audiences, and the Netflix audience likes

this type of stuff.

- In terms of production, the United States and India are ranked first and second, respectively followed by the UK and Canada. When a country produces a lot of stuff, it also consumes a lot of it.
- To discover the ideal number of clusters, we used the elbow curve and silhouette score of K-Means Clustering to build five clusters.
- The most clusters, according to our findings, are in cluster number 3.