# Capstone Project Submission

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

| **Team Member's Name, Email and Contribution:** |
| --- |
| Member Name:<br>• Shaik Ahmad Basha ahmadshaik982basha@gmail.com<br><br>Contribution:<br>• Exploring Data<br>• Data Wrangling<br>• Data Cleaning<br>• Checking for Null Values and Duplicated Values<br>• Handling Null Values<br>• Performed EDA on Dataset<br>• Text Preprocessing (tokenization, lowercasing, removing stop words and punctuations, stemming)<br>• Text Vectorization<br>• Dimensionality Reduction<br>• Finding similar movies using Cosine Similarity<br>• Determining optimal k value using elbow method and silhouette analysis<br>• K means clustering<br>• Hierarchical clustering |
| **Please paste the GitHub Repo link.** |
| GitHub Link: -https://github.com/ahmedshaik982/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.

The main objective behind the project is to explore the dataset to find some key understandings and to cluster similar content by using machine learning clustering algorithms

**Approaches:**

The first step imported all the necessary libraries like NumPy, Pandas etc. and then collected the data. I started with understanding the data like what are the columns and their meanings and data types.

After that, the second step is data preprocessing. Data preprocessing is a process where raw data is converted into clean data. The dataset has null values. After handling the null values, the next step is to analyze the data using Exploratory data analysis techniques.

And then I moved on text preprocessing where I have done tokenization, lowercasing, removed stop words and punctuations and stemming. And then I converted words to vectors by using TF-IDF Vectorizer. After converting, the data has so many columns, so in order to reduce dimensionality, I applied Principal component analysis.

And the last step is to cluster the data using clustering algorithms. And find the similar movies / TV shows using cosine similarity.

**Conclusions:**

There are more movies that tv shows on Netflix.

Jan Suter and Raul Campos directed most number of movies. Alastair Fothergill and Ken Burns directed most number of TV Shows.

Anupam Kher acted in most number of movies followed by Shah Rukh Khan. Takahiro Sakurai acted in most number of TV Shows followed by Junichi Suwabe.

United Stated produced most number of movies/Tv Shows followed by India.

In the year 2018 most number of movies and TV Shows are released. Most number of movies are released in the month of December and January.

In the recent years, Netflix has started to increase the TV Shows content as you can see the graph is increasing gradually.

Most number of movies and TV Shows belongs the rating category of 'TV-MA' (adults).

Most number of movies / TV Shows are from genre of 'International Movies' followed by dramas.

Steven Spielberg , David Dhawan and Cathy Garcia-Molina are the directors who makes content mostly for Teens and Older Kids.

By applying silhouette analysis, for k means, the optimal k value is 6. And for k = 6, silhouette score is 0.5001759759427156.

For hierarchical clustering, from dendrogram we can choose optimal k value as 3. And silhouette score is 0.5049348034764315.