**Capstone Project Submission Netflix Movies and TV Shows Clustering**

|  |
| --- |
| **Summary -** |
| As movie industry is evolving into streaming platforms, there’s no doubt that Netflix has become one of the important platforms for streaming. The dataset that we have used for EDA and clustering has been collected by Flixable, a third-party Netflix search engine. There are 12 features and around 7700 observations in the dataset and are mostly textual features  Our main objectives of this project are to do exploratory analysis and find useful insights from dataset, to understand what type content is available in different countries, also to find out is Netflix has increasingly focused on TV rather than movies in recent years and at last to do clustering of similar  content by matching text-based features from dataset.  I started this project with the intention to obtain some useful insights related to the type of Netflix content. For this, I performed exploratory data analysis on our data after cleaning and making it easy to analyse. This analysis helped me to understand the trend. I found that most of the content on Netflix are of TV-MA and TV-14 rating. USA and India are two countries producing the maximum number of contents. Documentaries and stand up are top genre in terms of number of contents they have on platform. Further I found number of movies on Netflix outnumbers TV-shows.  My next job was to make an unsupervised clustering model. For this, I processed the text by removing no use characters like - stop words, punctuation and did stemming. After getting the length for each text feature I rescaled them for generalization and started applying algorithms. First, I used Hierarchical clustering and used dendrogram to find out number of clusters. To increase the model performance, PCA was also implemented. Then applied K-means clustering. In order to find appropriate cluster number, I used elbow method. Then I found out that total 5 clusters are good for this data. Then implemented wordcloud for the clusters.  This was all about my analysis and training that I did, and based on result and my inferences I make the following conclusion:   * Majority of content available on Netflix are movies. * Most of the TV Shows and Movies are added in the month October, November, December and January. * United States and India are the highest content producing countries. * Large number of contents are for mature audiences. * Over past 5 years immense amount of TV Shows and Movies were released. * High percentage of TV-MA rating shows that Mature Content is more popular on Netflix. * TV Shows rarely go above 5 seasons and average time of a movie is around 90 to 120 minutes. * It was found that the optimal number of clusters was 5. Therefore total 5 distinct clusters were created using K-means Clustering Algorithm. * Documentaries, Family and Children Movies, International TV Shows, International Movies and Drama, Comedy Shows are the data represented in the clusters. |
| **Contributor Role** |
| Ritik Vaidande ([vr171k@gmail.com](mailto:vr171k@gmail.com))   * Understanding Data   Understanding different column  Having overview of data   * Exploratory Data Analysis * Content type analysis * Content added analysis * Release analysis * Visualizations * Bar graphs, Elbowplot, Wordcloud * Data Preparation * PCA * Modelling * Hierarchical Clustering * K-Means Clustering * Conclusions |
| **GitHub Repo and Drive Link** |
| Github Link:-  <https://github.com/vaidande/Netflix-Movies-and-TV-Shows-Clustering>  Drive Link:-  https://drive.google.com/drive/folders/1XLYhmO92KOuw6fhQcpETrqGByak3K8Y\_?usp=sharing |