|  |
| --- |
| **Team Member’s Name, Email, and Contribution:** |
| **Name** – Biswajit Gochhayat  **Email** -gochhayatbiswajit253@gmail.com  **Contribution –** Everything (Individual Project) |
| **Please paste the GitHub Repo link.** |
| ***Github link*:-https://github.com/143biswajit/NETFLIX-MOVIES-AND-TV-SHOWS-CLUSTERING.git** |
| **Please write a summary of your Capstone project and its components. Describe the problem statement, your approaches, and your conclusions. (200-400 words)** |
| **problem statement:**   1. 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. 2. 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.   **Approaches :**  **Part1: EDA**  Following are the takeaway points from exploratory data analysis:   1. Majority of contents are TV-MA : Mature Audience only. 2. Majority of contents were added after year 2015. 3. Director column has 30% null values, out of which most of them are for TV shows and very few for movies. 4. Majority of movies & shows are of around 100mins length & 1 season resp. 5. Majority of Netflix movies have IMDB user ratings between 5-7 while for TV Shows its 6-8.   **Part2: Country-level analysis**   1. Top 3 countries based on Netflix movies are U.S., India & U.K., and based on TV Show are U.S., U.K. & Japan 2. India has produced more movies compared to shows, the majority of which fall into the International, Drama & Comedy Genre. Top actors from India are Anupam Kheer, Shahrukh Khan, and Akshaya Kumar. IMDB user ratings of the majority of movies are in the range of 5-7. 3. U.S. has also produced more movies compared to shows, the majority of which fall into comedy, drama & documentaries. Top actors from the U.S. are Adam Sandler, Samuel Jackson, and Laura Bailey. IMDB user ratings of the majority of movies are in the range of 5-7. 4. U.K. has also produced more shows compared to movies, the majority of which falls into international & British genre. Top actors from the U.K. are David Attenborough, Greg Davies & Harriet Walter. IMDB user ratings of the majority of shows are in the range of 6-8. 5. Japan has also produced more shows compared to movies, majority of which falls into Anime & international genre. IMDB user ratings of majority shows are in range of 6-8   **Part3: Movie vs Tv Show in recent years**   1. Netflix is increasing both TV show & Movies contents. 2. Netflix has focused more on movies in recent year not the other way round. 3. Netflix has added 5186 movies and 2339 shows during 2016-2020 period(last 5 years),i.e twice number of movies were added as compared to shows.   **Part4 : Clustering**  The final model we used wask-means clustering, which consisted of 2,3,4,5,6 clusters.  For n\_clusters = 2 The average silhouette\_score is : 0.7049787496083262  For n\_clusters = 3 The average silhouette\_score is : 0.5882004012129721  For n\_clusters = 4 The average silhouette\_score is : 0.6505186632729437  For n\_clusters = 5 The average silhouette\_score is : 0.56376469026194  For n\_clusters = 6 The average silhouette\_score is : 0.4504666294372765  **Conclusions :**   1. Data set contains 7787 rows and 12 columns in that cast and director features contains large number of missing values so we can drop it and we have 10 features for the further implementation. 2. We have two types of content TV shows and Movies (30.86% contains TV shows and 69.14%contains Movies). 3. The most number of the movies and TV shows release in 2017 and 2020 respectively and united nation have the maximum content on Netflix 4. On Netflix, Dramas genre contains the maximum content among all of the genres and the most of the content added in December month and less content in February. 5. By applying the silhouette score method for n range clusters on dataset we got best score which is 0.244for 4 clusters it means content explained well on their own clusters, by using elbow method after k =4 curve gets linear it means k =4 will be the best cluster 6. By applying different clustering algorithms to our dataset, we get the optimal number of cluster is equal to 4. 7. We have done null value treatment, feature engineering, and EDA since loading the dataset then completed assigned tasks. 8. Among different types of content available in different countries, content TV-MA is available in most countries. This could be because it shows that it is just for adult audiences, and the Netflix audience enjoys content like this. 9. We have also explained different clusters based on their content; Defined clusters and enforced the K-means clustering algorithm and cluster number nine has the most clusters; we have also plotted a scatter plot in which we may interact with similar content about that cluster. |