**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

| **Team Member’s Name, Email and Contribution:** |
| --- |
| 1. Ayush Goyal [erayushgoyal96@gamil.com](mailto:erayushgoyal96@gamil.com)  2. M Sameer Ahamed [sameerm8095@gmail.com](mailto:sameerm8095@gmail.com)  3. Nitesh bhowmick [nitesh.gnit@gmail.com](mailto:nitesh.gnit@gmail.com) |
| **Please paste the GitHub Repo link.** |
| **https://github.com/Nitesh7179/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)** |
| **Netflix movies and Tv show clustering project is done by a group of 3 members-M. Sameer, Ahamed, Ayush Goyal, Nitesh Bhowmick. In this project, we got NETFLIX MOVIES AND TV SHOWS CLUSTERING as a CSV file.** **As we downloaded the data as a CSV file from the Almabetter Capstone project dashboard we encoded the file in the colab notebook by mounting the drive. All members of the group participated throughout the project with great effort.**  **The cleaning of data was done and created the newly cleaned dataframe consists of columns which were required for analysis. Each and every column were compared to gain the knowledge for analysis. Worked individually gaining some insights by doing some EDA. The first difficulty was the missing data in the dataset & one Column Name, so we renamed the columns by using a dictionary format. In the dataset, there are # columns which contain more NaN Values, so I filled them up with required values & there are another 2 columns which contain fewer NaN Values, so, I removed those particular rows which contain Nan values. From the graphs, we cleared the Type of shows, ratings, and Production Growth based on the type of content & release\_year, Genre, Duration, Country, Title & Cast.**  **After that, I created a model with 2 different Clustering Algorithms & we conclude that K Means is best for identification than Hierarchical as the evaluation metrics also indicate the same.**   * **From elbow and silhouette score, optimal of 26 clusters formed, K Means is best for identification than Hierarchical as the evaluation metrics also indicate the same. In kmean cluster 0 has the highest number of data points and is evenly distributed for other clusters** * **Netflix has 5372 movies and 2398 TV shows, there are more number movies on Netflix than TV shows.** * **TV-MA has the highest number of ratings for tv shows i.e adult ratings** * **Highest number of movies released in 2017 and 2018**   **highest number of movies released in 2020 The number of movies on Netflix is growing significantly faster than the number of TV shows. We saw a huge increase in the number of movies and television episodes after 2015. there is a significant drop in the number of movies and television episodes produced after 2020. It appears that Netflix has focused more attention on**  **increasing Movie content than TV Shows. Movies have increased much more dramatically than TV shows**   * **The most content is added to Netflix from October to January** * **Documentaries are the top most genre on Netflix which is followed by standup comedy and Drams and international movies** * **kids tv is the top most TV show genre in Netflix** * **most of the movies have a duration of between 50 to 150** * **highest number of tv\_shows consisting of single season** * **Those movies that have a rating of NC-17 have the longest average duration.**   **When it comes to movies having a TV-Y rating, they have the shortest runtime on average**   * **United States has the highest number of content on Netflix, followed by India** * **India has the highest number of movies on Netflix** * **30% of movies are released on Netflix.** * **70% of movies added on Netflix were released earlier in a different mode.**     **Contributors Roles:**   1. **Ayush Goyal:**   1. Data Wrangling:  1. work on data handling.  2. Visualizing based on Distplot with normal distribution for movies.  3. Visualizing based on the count.  4. Deploy & Run k –means clustering.  5. Visualizing based on Top 10 movies and Tv shows Ratings.  **2. M Sameer Ahamed:**  1. Data Wrangling:  1. work on dendrogram algorithm.  2. Visualizing based on Tv show rating.  3. Visualizing based on movie rating.  4. Visualizing based on production growth yearly.  5. Visualizing based on top 10 genre movies.  **3. Nitesh Bhowmick:**  1. Data Wrangling:  1. work on data handling.  2. Visualizing based on top 10 genre Tv shows.  3. Visualizing based on the top 15 countries with the most countries.  4. Visualizing Common titles using a word cloud.  5. Run & Deploy an agglomerative clustering Model. |