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
2. M Sameer Ahamed	sameerm8095@gmail.com
3. Nitesh bhowmick	nitesh.gnit@gmail.com

Please paste the GitHub Repo link.

https://github.com/ Ayugoyal/-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)

Netflix movies and Tv show clustering project is done by 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 CSV file from Almabetter Capstone project dasboard we encoded the file in colab notebook trough mounting the drive. All members from the group participated throughout the project with great efforts.

The cleaning of data was done and created the new 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 doing some EDA.The first difficulty was the missing data in the dataset & one Column Name, so we renamed the columns by using dictionary format. In the dataset tehre are # columns which contain more NaN Values, so i filled up with requird values & another there are another 2 columns which contains less NaN Values, so, i remove those particular rows which contain Nan values . From graphs we cleared the Type of shows, ratings, Production Growth based on type of the content & release_year, Genre, Duration, Country, Title & Cast.

After that I created model with 2 different Clustering Algorithims & we conclude that K Means is best for identification than Hierarchical as the evaluation metrics also indicates the same.

- From elbow and sillhoute score ,optimal of 26 clusters formed , K Means is best for identification than Hierarchical as the evaluation metrics also indicates the same. In kmean cluster 0 has the highest number of datapoints and evnly distributed for other cluster
- 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 in Netflix which is fllowed by standup comedy and Drams and international movies
- kids tv is the top most TV show genre in Netflix
- most of the movies have duration of between 50 to 150
- highest number of tv_shows consistig 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

- unitated states has the highest number of content on the netflix ,followed by india
- india has highest number of movies in Netflix
- 30% movies released on Netflix.
- 70% movies added on Netflix were released earlier by different mode.

Contributors Roles:

1. Ayush Goyal:

1. Data Wrangling:

1. work on data handing

2. Visualizing based on Distplot with normal distribution for movies

3. Visualizing based on count

4. Deploy & Run k –means clustering

5. Visualizing based on Top 10 movies and Tv shows Rating

2. M Sameer Ahamed:

1. Data Wrangling:
1. work on dendogram 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 movie

3. Nitesh Bhowmick:

1. Data Wrangling:
1. work on aggomerative clustering
2. Visualizing based on top 10 genre Tv show
3. Visualizing based on top 15 countries with most countries
4. visualizing based on production growth