Capstone Project Submission

Team Member's Name, Email, and Contribution:

1. **Divyesh Dhanani (divyeshdhanani143@gmail.com)**

- Data Wrangling
  - ➢ Comprehending each column in the dataset
  - ➢ Raising the bar of questions from the dataset
  - ➢ Gaining worthful inference from the dataset for analytical conclusions.
  - ➢ Cleaning the clusters in Data
  - ➢ Dropping the null value columns
  - ➢ Replacing the minimum number of null values with str.
- Exploratory Data Analysis
  - ➢ Statistical Based Analysis
  - ➢ Univariate and Bivariate Analysis
  - ➢ Categorical Analysis
  - ➢ Profitability Analysis
  - ➢ Data Visualization (Plotting the insights in the graph using: Bar Plot, Box plot, Scatter Plot, Pie charts, Heatmaps, etc.)
- Clustering
  - ➢ K-means clustering
  - ➢ Silhouette Method
  - ➢ Elbow Method
  - ➢ Cosine-similarity
- Inference Gained
  - ➢ Conclusion

2. **Mrityunjay Singh Chandel (mrityunjaychandel98@gmail.com)**

- Data Wrangling
  - ➢ Comprehending each column in the dataset
  - ➢ Raising the bar of questions from the dataset
  - ➢ Gaining worthful inference from the dataset for analytical conclusions.
  - ➢ Cleaning the clusters in Data
  - ➢ Dropping the null value columns
  - ➢ Replacing the minimum number of null values with str.
- Exploratory Data Analysis
  - ➢ Statistical Based Analysis
  - ➢ Univariate and Bivariate Analysis
  - ➢ Categorical Analysis
  - ➢ Profitability Analysis
  - ➢ Data Visualization (Plotting the insights in the graph using: Bar Plot, Box plot, Scatter Plot, Pie charts, Heatmaps, etc.)
- Clustering
  - ➢ K-means clustering
  - ➢ Silhouette Method
  - ➢ Elbow Method
  - ➢ Cosine-similarity
- Inference Gained
  - ➢ Conclusion

3. **Nimisha Jadhav (jnimisha21@gmail.com)**

- Data Wrangling
  - ➢ Comprehending each column in the dataset
  - ➢ Raising the bar of questions from the dataset
  - ➢ Gaining worthful inference from the dataset for analytical conclusions.
  - ➢ Cleaning the clusters in Data
  - ➢ Dropping the null value columns
  - ➢ Replacing the minimum number of null values with str.
- Exploratory Data Analysis
  - ➢ Statistical Based Analysis
  - ➢ Univariate and Bivariate Analysis
  - ➢ Categorical Analysis
  - ➢ Profitability Analysis
  - ➢ Data Visualization (Plotting the insights in the graph using: Bar Plot, Box plot, Scatter Plot, Pie charts, Heatmaps, etc.)
- Clustering
  - ➢ K-means clustering
  - ➢ Silhouette Method
  - ➢ Elbow Method
  - ➢ Cosine-similarity
- Inference Gained
  - ➢ Conclusion

4. **Sagar Tikmani (sagartikmani900@gmail.com)**

- Data Wrangling
  - ➢ Comprehending each column in the dataset
  - ➢ Raising the bar of questions from the dataset
  - ➢ Gaining worthful inference from the dataset for analytical conclusions.
  - ➢ Cleaning the clusters in Data
  - ➢ Dropping the null value columns
  - ➢ Replacing the minimum number of null values with str.
- Exploratory Data Analysis
  - ➢ Statistical Based Analysis
  - ➢ Univariate and Bivariate Analysis
  - ➢ Categorical Analysis
  - ➢ Profitability Analysis
  - ➢ Data Visualization (Plotting the insights in the graph using: Bar Plot, Box plot, Scatter Plot, Pie charts, Heatmaps, etc.)
- Clustering
  - ➢ K-means clustering
  - ➢ Silhouette Method
  - ➢ Elbow Method
  - ➢ Cosine-similarity
- Inference Gained
  - ➢ Conclusion

5. **Vishal Chakrabarty (vishalchakrabarty20@gmail.com)**

- Data Wrangling
  - ➢ Comprehending each column in the dataset
  - ➢ Raising the bar of questions from the dataset
  - ➢ Gaining worthful inference from the dataset for analytical conclusions.
  - ➢ Cleaning the clusters in Data
  - ➢ Dropping the null value columns
  - ➢ Replacing the minimum number of null values with str.
- Exploratory Data Analysis
  - ➢ Statistical Based Analysis
  - ➢ Univariate and Bivariate Analysis
  - ➢ Categorical Analysis
  - ➢ Profitability Analysis
  - ➢ Data Visualization (Plotting the insights in the graph using: Bar Plot, Box plot, Scatter Plot, Pie charts, Heatmaps, etc.)
- Clustering
  - ➢ K-means clustering
  - ➢ Silhouette Method
  - ➢ Elbow Method
  - ➢ Cosine-similarity
- Inference Gained
  - ➢ Conclusion

Github Link: https://github.com/divyeshdhanani14/Netflix_Movies_TV_Shows_Capstone

NETFLIX MOVIES AND TV SHOWS CLUSTERING

Netflix is an OTT Platform that contains an extensive collection of TV Shows and Movies that we can access anytime online with a suitable internet connection and a subscription plan. Users can cancel their subscriptions anytime. So, the Company needs to provide content according to the users, so that the users stay hooked and don't cancel the subscription. With the help of Recommender Systems, it provides suggestions suitable to the users.

In order to explain clustering we will be building a recommendations system for the users to continue their subscriptions. We will consider the following information about the viewers which includes show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, and description. Our dataset is composed of 7787 rows and 12 Feature columns. We checked the duplicate, null, and outliers and feature-scaled the outliers for further analysis.

In this project, we have defined the following objectives - exploratory data analysis, Understanding what type of content is available in different countries, whether Netflix has increasingly focused on TV rather than movies in recent years, and Clustering similar content by matching text-based features.

As the first step before creating the cluster for the recommender system, we will be inspecting the data, cleansing it, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Further, step two involves Exploratory data analysis which refers to the critical process of performing initial investigations so as to discover patterns such as the Growth of Content over years. The following inputs were identified through exploratory data analysis - Analysis based on a country, Genre vs Country, Creating the word cloud to see which appears the most in the titles and description of the movies and the TV shows, duration, top directors, actors, and TV show ratings whose content is available on Netflix. Data Pre-processing step includes creating clusters for our data using text columns, creating functions to remove the punctuation and stop words and dimensionality reduction using PCA. The k-means clustering method is used to determine the best value for K center points or centroids by an iterative process. This method assigns each data point to its closest k-center. Those data points which are near the particular k-center, create a cluster. In order to select the optimum value Silhouette Score and Elbow method were used. By applying the Silhouette Score Method, we found the optimum value of K = 10. Using the given data set a simple recommender system was also created using the cosine similarity and recommendations for TV Shows and movies were obtained. 68.1 % of the content available on Netflix are movies and 30.9 % of the content are TV shows. Relative growth is observed here in the number of movies on Netflix rather than Tv shows.