**Capstone Project Submission**

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| **1. Lucky Jain Email-id :-** [jainluckycool@gmail.com](mailto:jainluckycool@gmail.com)  **Contribution-**   * EDA on type and ratings and to evaluate the highest rating with content preferred by the audience. * Analysis on the countries and cast in the show using word cloud visualization. * Analysis done to observe Netflix has increasingly focusing on TV shows rather than movies in recent years   or not.   * Data preparation using nlp. * Model implementation for agglomerative clustering. * Correlation matrix. * Elbow method & Box-plot for K-means clustering   **2. Debashish Das Email-id :- devashishdas40@gmail.com**  **Contribution -**   * Checking the NAN values in the dataset. * EDA on highest watched genre on Netflix. * Hypothesis findings based on release year and date feature. * Data Preprocessing (USING NLTK) * Standardized the data into scalar transformation. * Model implementation for K-means Clustering. * Data engineering on the basis of clustering text based features * Silhouette score method & Box-plot for K-means clustering * Dbscan   **3. Vivek Katolkar Email-id:- vivekkatolkar7@gmail.com**  **Contribution -**   * Scatter plot with type feature and with clusters after fitting K-means clustering model. * EDA on highest watched genre on Netflix. * Converting the date into date-time format. * Understanding the type of content available in different countries * dendogram |
| **GitHub link:** [**mailto:https://github.com/luckyj23/Netflix-Movies-and-TV-Shows-Clustering-**](mailto:https://github.com/luckyj23/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 is by far the most widely used media and video streaming service. It includes over 8000+ movies with tv shows worldwide. Currently, Netflix has more than 200 million subscribers worldwide. NETFLIX is also the most widely utilized entertainment platform worldwide. It offers a vast library of films and TV series that may be seen at any time through internet services.

Netflix is without a doubt the most used media and video streaming service. There are more than 8000 films and television programmers included. At the moment, Netflix has over 200 million subscribers worldwide. Additionally, NETFLIX is the most widely used entertainment service worldwide. Through online services, it offers a vast library of movies and TV series that may be accessed at any time.

The tabular dataset includes listings for all Netflix movies and TV shows, together with information about the actors, directors, ratings, release year, duration, and other factors. It has 7787 rows and 12 columns.

Our project's main objective is to build a model that can cluster similar data by matching text-based features.

According to the problem statement, the question arises that, understanding what type of content is available in different countries and Is Netflix increasingly focused on TV rather than movies in recent years we have to do clustering on similar content by matching text-based features. For that we have used K-means Clustering.

To obtain relevant data free of NAN values (null values), we will manipulate the raw data. We will next look at the dataset's summary statistics. With feature engineering, feature scaling, and the removal of unnecessary columns, we prepared a dataset. To use the model in the study, the data was transformed into a common scalar form.

We executed the exploratory data analysis. Reviewing all the data processing then the model was trained to form Collections. In which we remark as below- k-means clustering model gave insights of silhouette analysis consisting of 2,3,4,5,6 clusters.

For n\_clusters = 2 The average silhouette score is 0.42541313028836003

For n\_clusters = 3 The average silhouette score is: 0.3940500353920696

For n\_clusters = 4 The average silhouette score is: 0.38498095158183726

For n\_clusters = 5 The average silhouette score is: 0.3962372504377786

For n\_clusters = 6 The average silhouette score is: 0.392565886828632

In the end, we plot boxplot to predict the hypothesis -

* After clustering, we can state that the number of TV series that have been released over the past few years is not increasing, which is our alternative hypothesis.
* Our second alternative hypothesis is the number of TV shows added to Netflix is high.

We calculated that whereas movies make up 97.2 percent of the total, TV series only make up 2.8 percent. When compared to films made in the last 10 years, Netflix has added a lot more movies and TV episodes in the past years, but the numbers are still small. In many nations, people choose to watch movies over TV programming.

|  |
| --- |
| **Drive link:** <mailto:https://drive.google.com/drive/folders/15bbyV8bl3GJ9OKMJVBO5ptgK5dNgLpJ5?usp=sharing> |