**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| RAHUL SINGH WALDIA  [rahulwaldia2@gmail.ciom](mailto:rahulwaldia2@gmail.ciom)   1. DATA CLEANING 2. DATA VISUALIZATION 3. DATA CLUSTERING 4. DATA ANALYSIS |
| **Please paste the GitHub Repo link.** |
| Github Link:- <https://github.com/Link/to/Repo> |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| The aim of this project is to utilize Python for data visualization and cluastering purposes. Netflix data from AlmaBetter is used for this project. In this project I will be exploring the Movies and TV shows in Netflix The Netflix Content Type summary indicates that there are 69.05% of observations are Movies compared to 30.95% TV Shows. Moreover, Rating categories shows a wide range of movie and TV Show rating in the Netflix data set.  The Duration summary table Movies duration is measured in minutes and TV Shows duration is measured in number of seasons. It can be seen that average duration for a movie is 99 minutes, and the average length of a TV Show is around two seasons. Netflix content distribution shows the existence of more movies than TV shows. Movies duration is measured in minutes, the below distribution has a normal distribution, it seems that the average duration of a movies is around 100 minutes with is close to the duration mean from the summary table. TV shows distribution shows that many shows have one season, moreover the distribution has a right long tail and it is skewed. Graphs shows that United States is the largest contributor to the Netflix content. In addition to the data summary, description column contains a large amount of information about the details of a content, and it provides a great amount of information about the movie or the TV show. It is interesting to see how the words are related to each other such as mother, daughter, son and others. Curve clearly indicate that there has been very few movies and TV shows released before 2000. As the animated line graphs show that many movies and TV shows were released after 2000.It can be seen that in 1960s to 2000s the duration of movie is longer but then it decreases. Based on the visualization and text analysis of the Netflix data, it can be seen that movies have a larger part in the Netflix content. Furthermore, the data summaries and visualization indicated that Netflix started contributing it Content since 2008 and United states was the only country back in time. Moreover, Netflix movies and TV shows are rated and these ratings are standardized for different audience. From duration and ranting categories analysis it can be seen that movies which children can watch, on average are shorter. |