# Capstone Project Submission

| **Team Member's Name, Email, and Contribution:** |
| --- |
| Name : Vishal Mahadev Potdar<br><br>Email : Potdar15v@gmail.com<br><br>Contribution : Individual |
| **GitHub Repo link.** |
| Github Link:- https://github.com/VishalPotdar-ds/Unsupervised-ML---Netflix-Movies-and-TV-Shows-Clustering.git |
| **Summary :** |

# Netflix Movies and Tv Shows Clustering

The Netflix Movies and TV Shows Clustering project aims to cluster similar movies and tv shows available on Netflix into different clusters based on their content. The project begins with data collected from the third-party Netflix search engine, which contains information about more than 7,000 movies and TV shows available on the platform.

The data set contained details about movies and TV shows available on Netflix. Descriptive statistics were computed for each variable as part of the analysis, and visualizations were made to investigate the relationships between the various variables. We created a number of graphs, such as the scatterplot, distplot, count plot, bar plot, pair plot, heatmap, pie plot, and box plot to gain insight from the dataset.

The dataset containing 12 columns and 7787 rows. This dataset contains no values that are duplicates. The some values for the director, cast, country, date added, and rating are null. Everything else is categorical, with the exception of the numerical feature release year. Although the datatype of the date_added feature's dates incorrectly associates an object, it contains dates.

In 30.68%, 9.22%, 6.51%, 0.13%, and 0.09% of their respective features, director, cast, country, date_added, and rating had null values. These null values have been replaced by director Unavailable, Cast Unavailability, and Country Unavailable, respectively, because there are several null values for features like director, cast, and country. Due to the extremely low quantity of null values for features like date_added and rating, we eliminated those null values.

The interquartile range is successfully used to treat outliers from the variable release year. Before eliminating the feature date added, it was converted to datetime and used to generate additional features like year added, month added, and day added. Geners is the new name assigned to the listed in feature. The feature release year data type is changed from float64 to int64 because year cannot be a float.

After doing univariate, bivariate, and multivariate analyses, we discovered insights that are as follows :

- More movies (69.14%) than TV shows (30.86%) are available on Netflix.
- The majority of Netflix movies were released between 2015 and 2020, and the majority of - Netflix TV shows were released between 2018 and 2020.

- The most movies and TV shows were released for public viewing on Netflix in 2017 and 2020, respectively, out of all released years.
- From 2006 to 2019 Netflix is constantly releasing more new movies than TV shows, but in 2020, it released more TV shows than new movies, indicating that Netflix has been increasingly focusing on TV rather than movies in recent years.
- More TV shows will be released for public viewing in 2020 and 2021 than at any other time in the history of Netflix.
- The majority of TV shows and movies available on Netflix have a TV-MA rating, with a TV-14 rating coming in second.
- The majority of movies added to Netflix in 2019 and the majority of TV shows added to - Netflix in 2020.
- In 2019, Netflix added nearly one-fourth (27.71%) of all content (TV shows and movies).
- The majority of the content added to Netflix was in October and January, respectively, but almost all months throughout the year saw Netflix adding content to its platform.
- Netflix has more movies (69.14%) than TV shows (30.86%).
- The majority of movies available on Netflix are produced in the United States, with India coming in second.
- The United States and the United Kingdom are the two countries that produced the most of the TV shows that are available on Netflix.
- Raul Campos and Jan Suter directed most of the movies available on Netflix for public viewing.
- Alastair Fothergill directed most of the TV shows available on Netflix for public viewing.
- International movies and the second-most popular dramas are available on Netflix as content.
- Actors who have appeared in films and TV shows that are most available on Netflix are Lee, Michel, David, Jhon, and James.
- We see that the movie or TV show release year and day of the month on movies or TV shows added to Netflix are slightly correlated with each other.
- Based on the plot of release_year and year_added, we can conclude that Netflix is increasingly adding and releasing movies and TV shows over time.
- We can conclude from plot release_year and month_added that Netflix releases movies and TV shows throughout the all months of the year.

We processed text data from the description variable by removing punctuation, stopwords, whitespace, emails, html tags, urls, special characters, and digits. Vectorized after lemmanization and got a TFIDF matrix for feeding to the model as input. We trained various clustering algorithms, such as KMeans clustering, Hierarchical clustering, and DBSCAN.

Among all models, the K-Means clustering model has the highest Calinski-Harabasz score (9.039247). Also, the K-Means Clustering model has a silhouette_score of 0.004634, which is closer to 1 than other models,

which means the K-Means Clustering model is capable of perfectly clustering items.

We faced the following challenges when building models: [1] Identifying the number of clusters is a difficult task. [2] The Kelbow method and silhouette score method take a long time to show the results for finding the ideal number of clusters.