NETFLIX MOVIES AND TV SHOWS CLUSTERING

TECHNICAL DOCUMENT

BY – AYUSH SHARMA

PROBLEM STATEMENT

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine.

In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same dataset.

Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

DATA OVERVIEW

The dataset provided consisted of 7787 rows and 12 columns

we are given the following columns in our data-:

1. show_id : Unique ID for every Movie / Tv Show
2. type: Identifier - A Movie or TV Show
3. title: Title of the Movie / Tv Show
4. director: Director of the Movie
5. cast: Actors involved in the movie / show
6. country: Country where the movie / show was produced
7. date_added: Date it was added on Netflix

8. release_year: Actual Release year of the movie / show

9. rating: TV Rating of the movie / show

10. duration: Total Duration - in minutes or number of seasons

10. listed_in: Genere

11. description: The Summary description

PROBLEMS FACED

The problem that I faced was that the value in rating column were difficult to understand so in order to make it understandable I googled the ratings and converted them into 4 age groups Kids, Older kids, Teens and Adults.

Performing EDA (exploratory data analysis)

A. Exploring head and tail of the data to get insights on the given data.

B. Looking for null values and handling them.

C. Creating new columns like year added and month added from date_added.

D. Creating more columns in our dataset which would be helpful for creating model.

E. Plotting columns like country, release year, year added, ratings, duration, listed in with respect to type to gain insights

F. Perform NLP tasks for description column.

K-MEANS IMPLEMENTATION

G. Using silhouette analysis to get optimal value of k. we got k equals to 22.

H. Implemented K-MEANS Algorithm with 22 clusters.

I. Seeing which cluster contains maximum number of data points.

conclusion

To conclude I want to say that throughout the entire duration of the project we initially did data inspection and then cleaned the data accordingly. Them performed basic EDA and were able draw out some relevant insights like Netflix has more number of movies than TV Shows, most movies are of duration between 80-120 minutes whereas most TV shows had only one season and many more. After that we implemented K-MEANS algorithm and formed 22 clusters. From which cluster number 0 had maximum number of data points.