Capstone Project

NETFLIX MOVIES AND TV SHOWS CLUSTERING

BY – AYUSH SHARMA

Agenda

- ❑ Problem Statement
- ❑ Data Description
- ❑ EDA
- ❑ K-MEANS ALGORITHM.
  - > Silhouette Analysis
  - > Cluster with most data points.
- ❑ Conclusion

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 Description

- The Dataset provided consisted of 7787 rows and 12 columns.

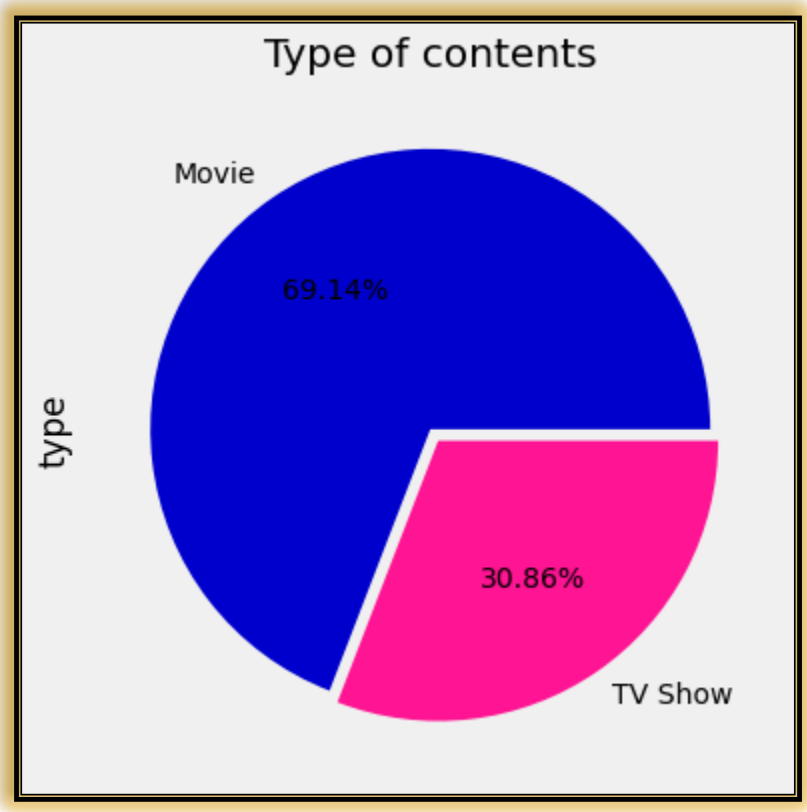
- There were zero duplicates values in the Dataset.

Columns

- Show Id
- Type
- Title
- Director
- Cast
- Country
- Release year
- Listed in
- Date added
- Description
- Ratings
- Duration

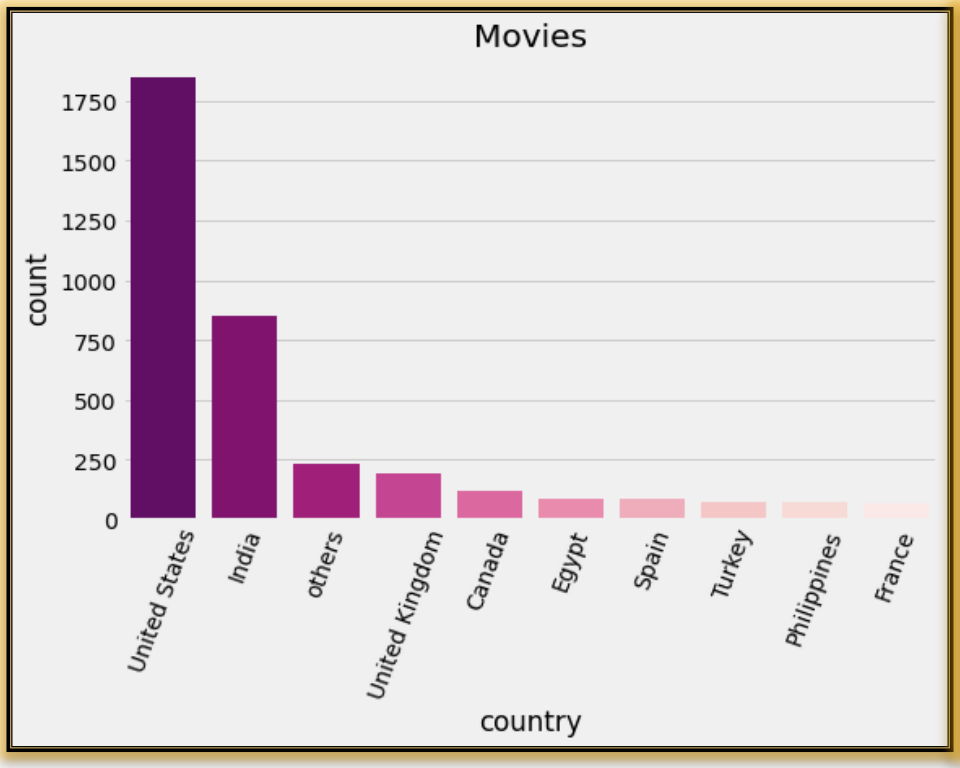
EXPLORATORY DATA ANALYSIS (EDA)

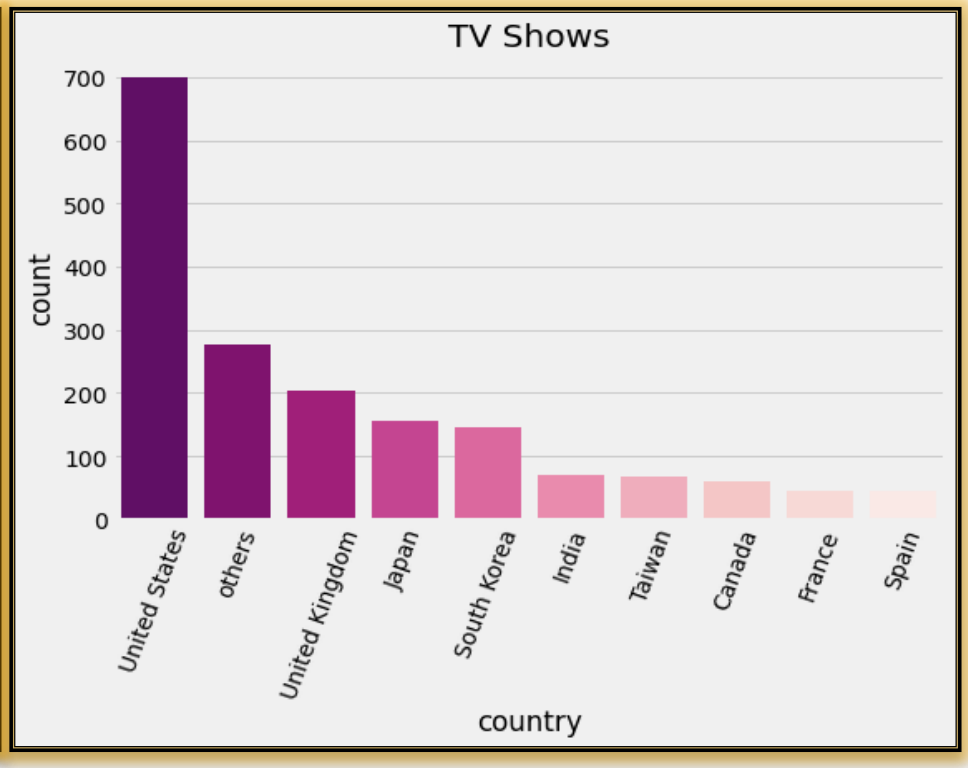
Type of Content available



From above plot we can clearly infer that Netflix has more number of Movies than TV Shows.

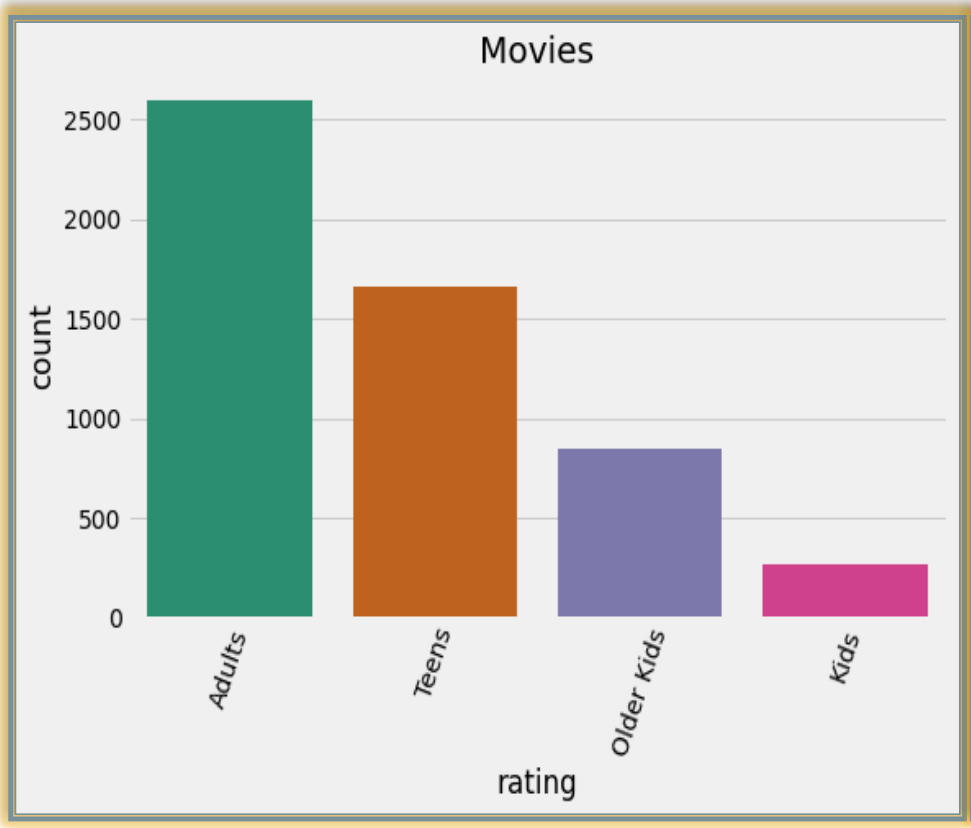
Country Wise Content

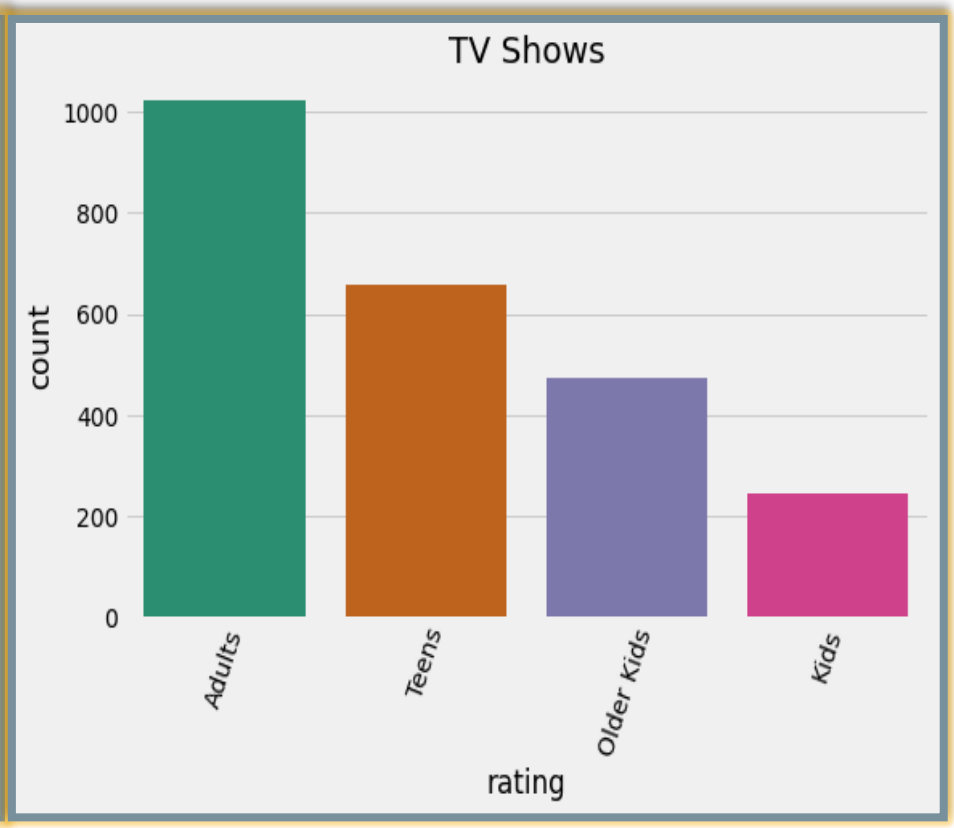




USA has most number of Movies and TV Shows on Netflix

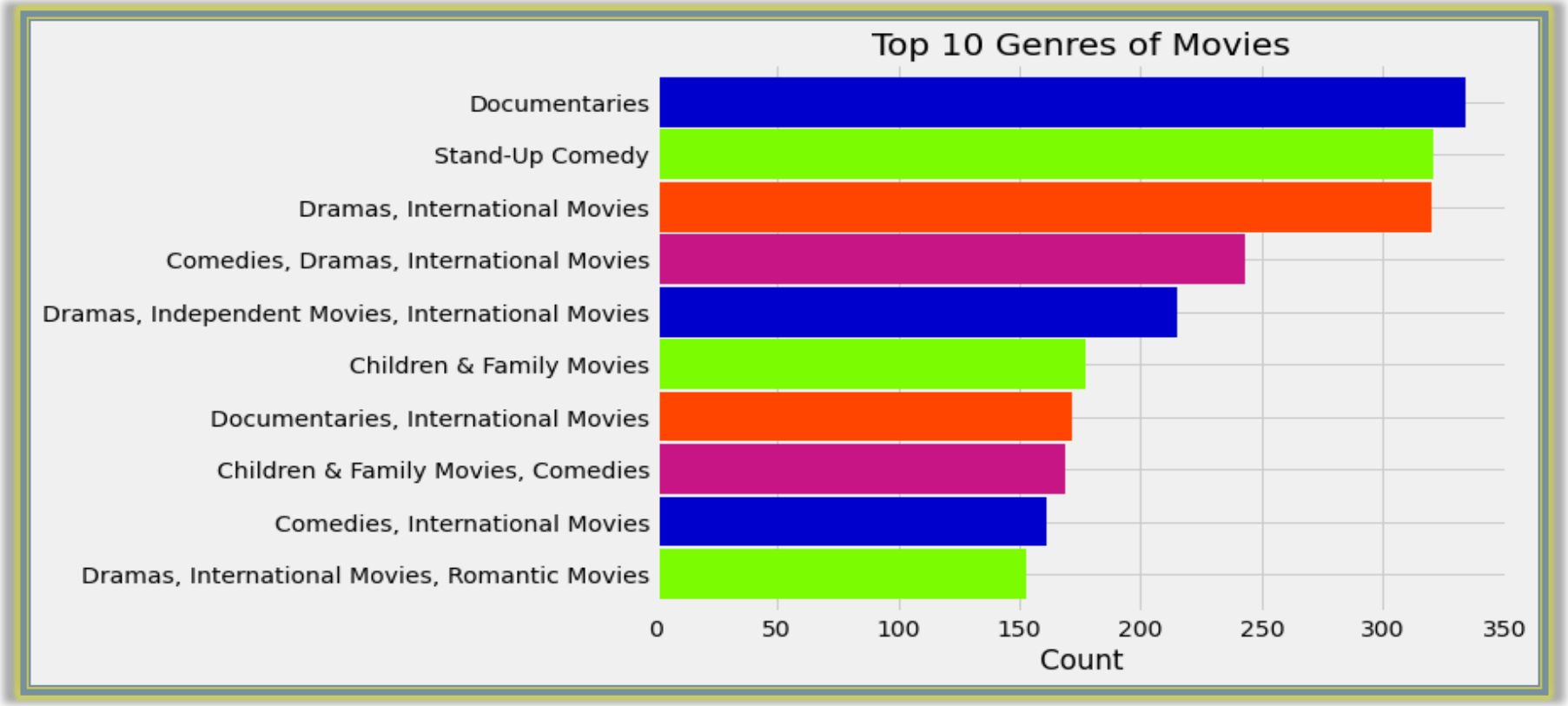
Ratings of the Contents





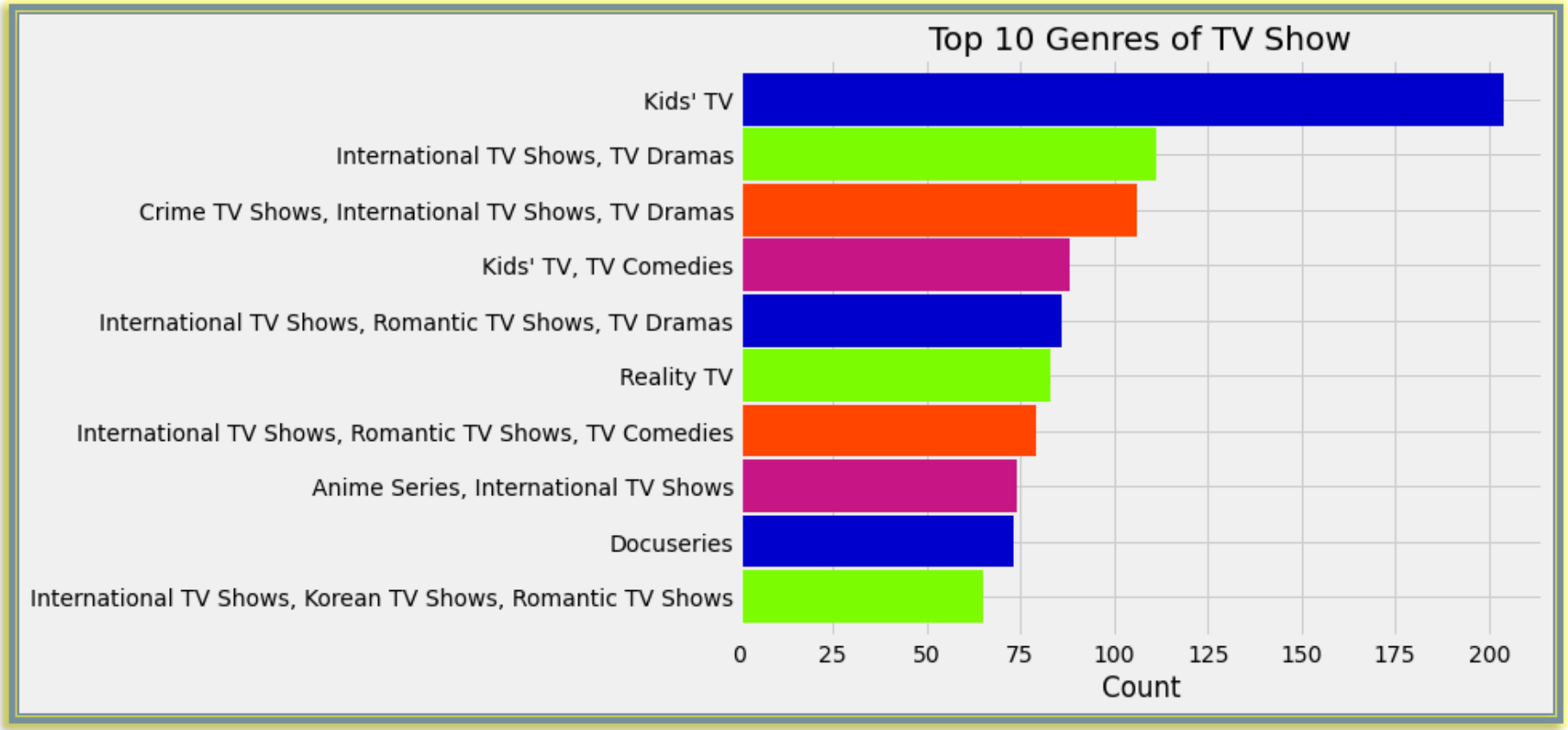
Most of content on Netflix is for adult audience followed by teens.

Top Genres for Movies



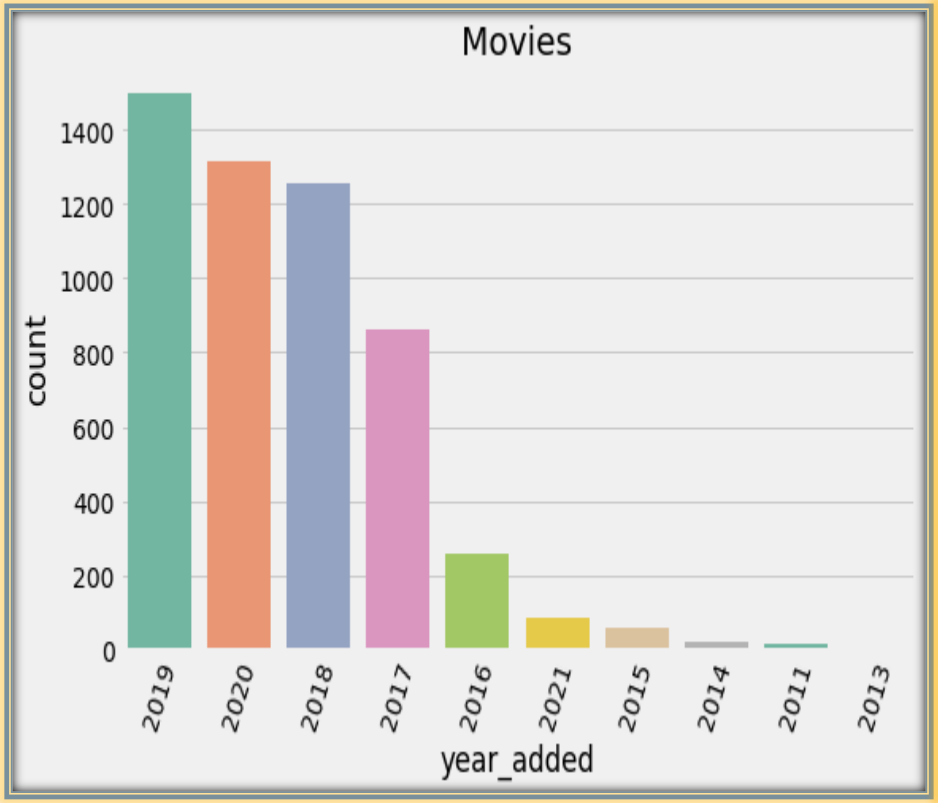
The top most genre for Movies on Netflix is Documentaries

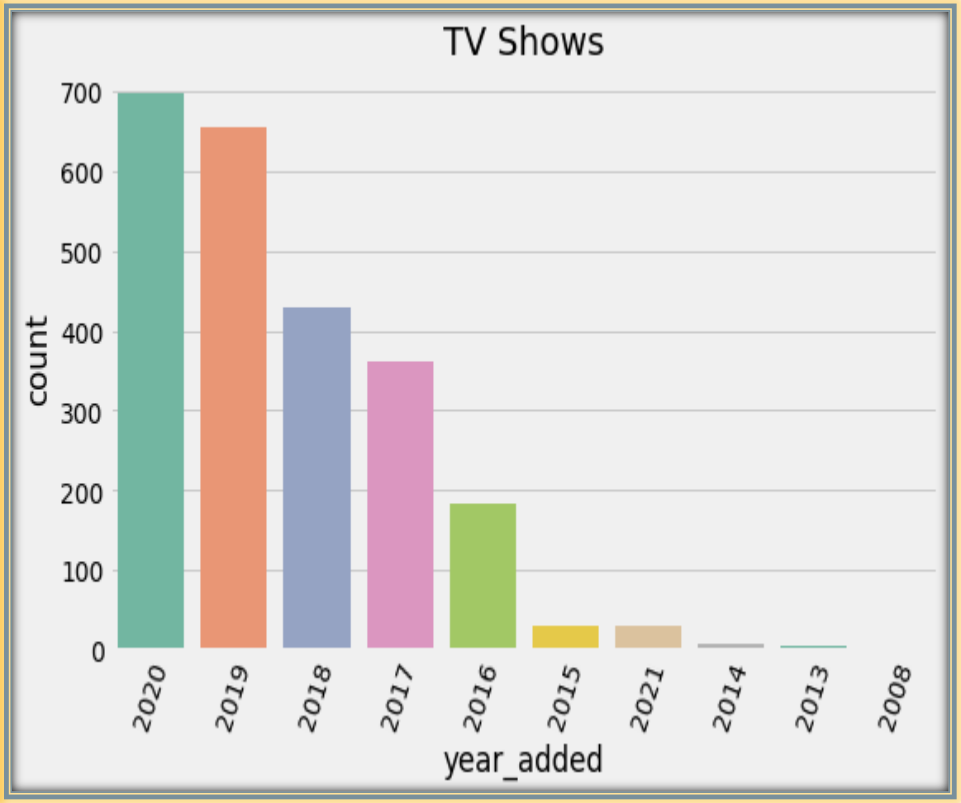
Top Genres for TV Shows



The top most genre for TV Show on Netflix is Kid's TV

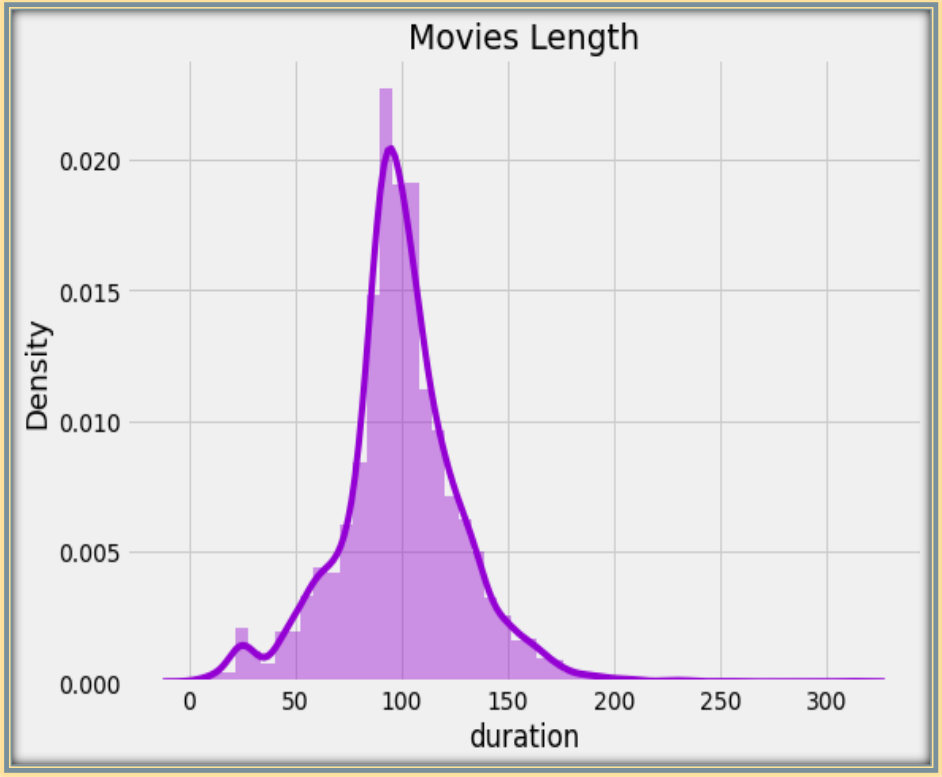
Movies and TV Shows according to year added

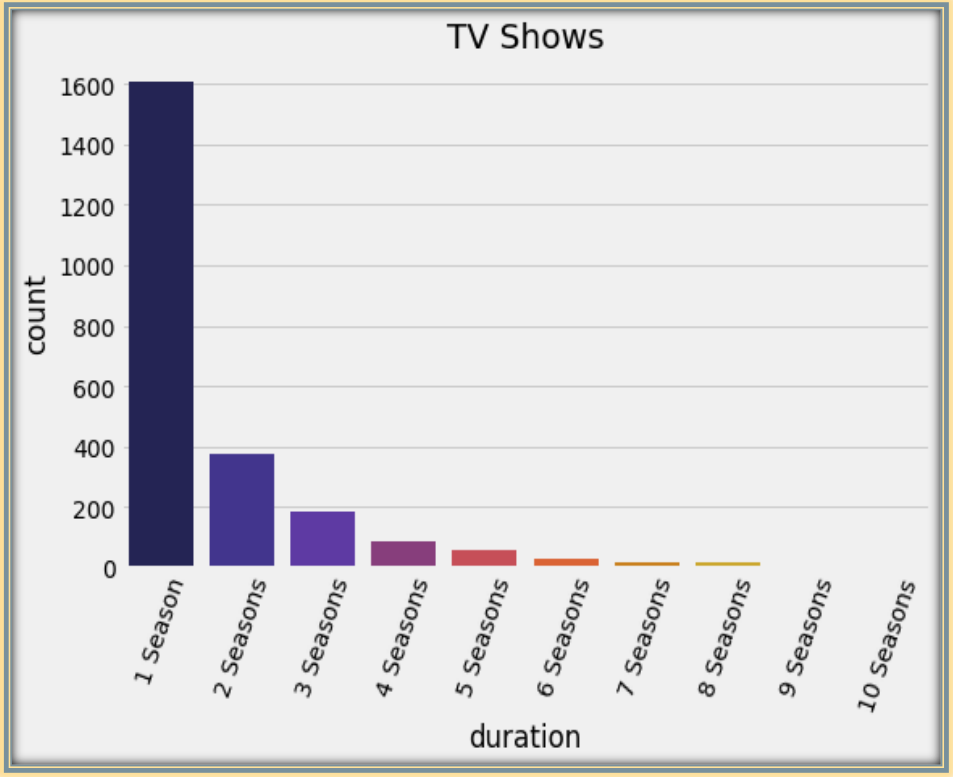




We can clearly see that the most number of movies were added in the year 2019 and most number of TV shows were added in the year 2020.

Movies and TV Shows duration

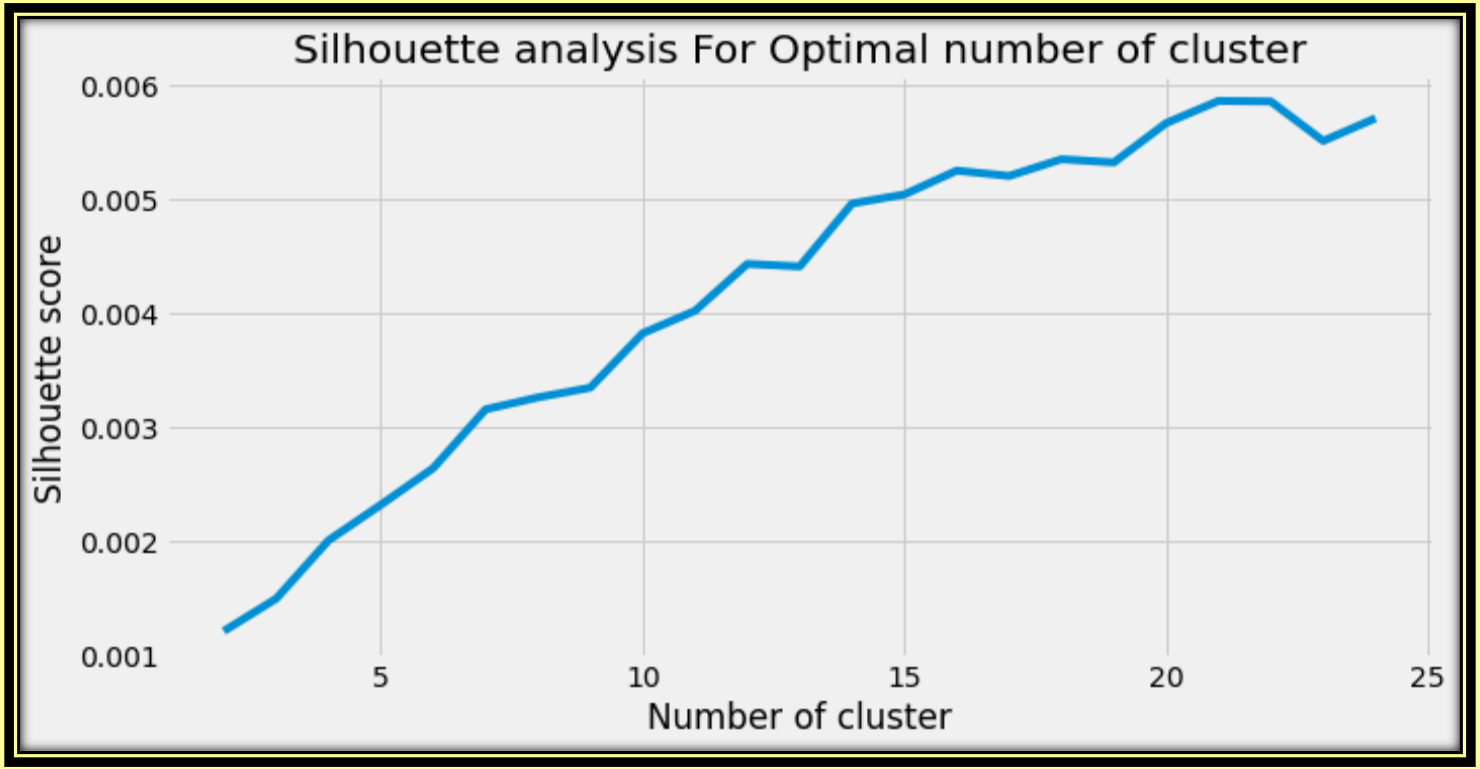




- **Runtime of movies on Netflix is between 80-120 min. approx. and most TV shows on Netflix have only one season.**

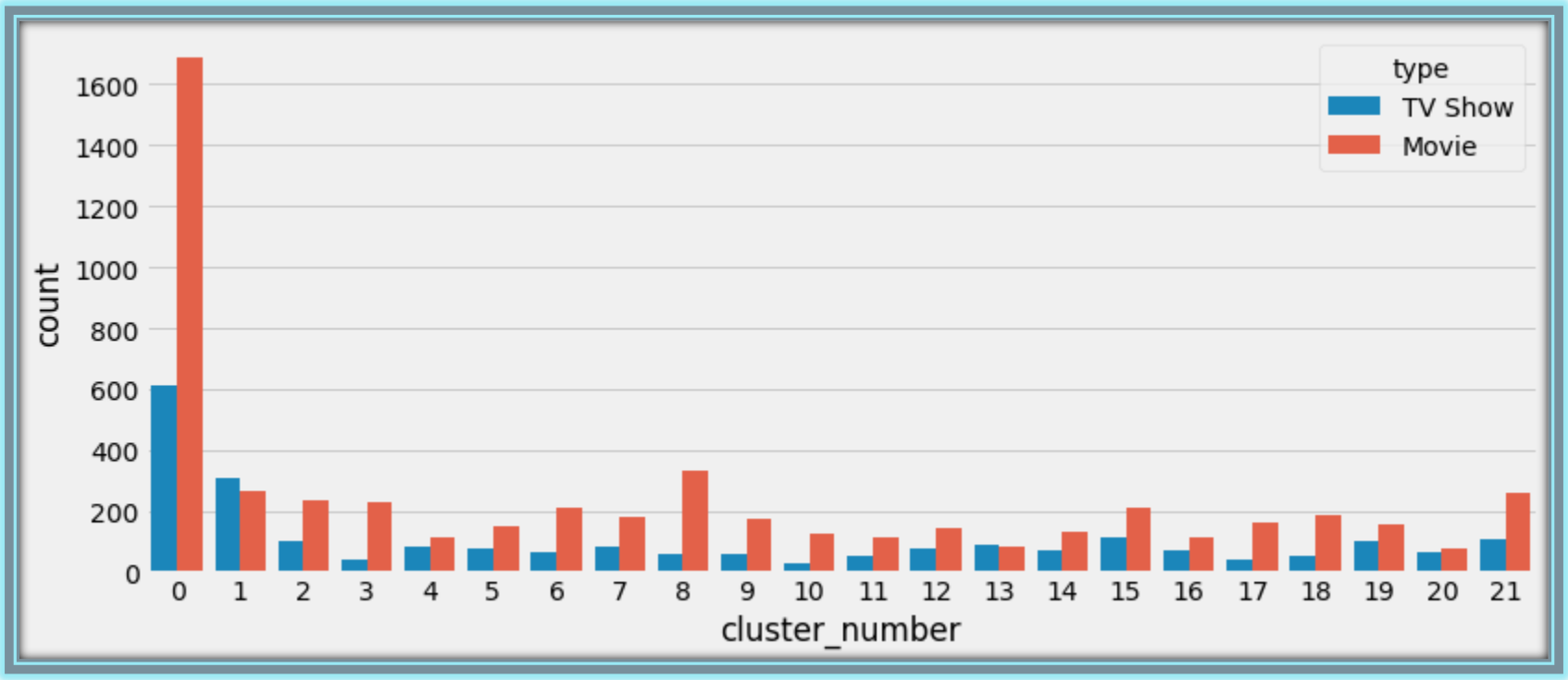
K-MEANS ALGORITHM

Silhouette Analysis



- **With Silhouette Analysis we found that building 22 clusters would be optimal for our dataset.**

Cluster with maximum data point



- We see that cluster number 0 has most number of data points.

Conclusion

- To conclude I just want to say that we were able to get some valuable insights after performing EDA and after that we implemented K-Means clustering algorithms with 22 clusters. Most number of data points were in cluster number 0.

THANK YOU