Capstone Project - 4

Netflix Movies and TV shows Clustering

Team members

Kajal Dhun
Navinkumar Sambari
Tanu Rajput

Problem Statement

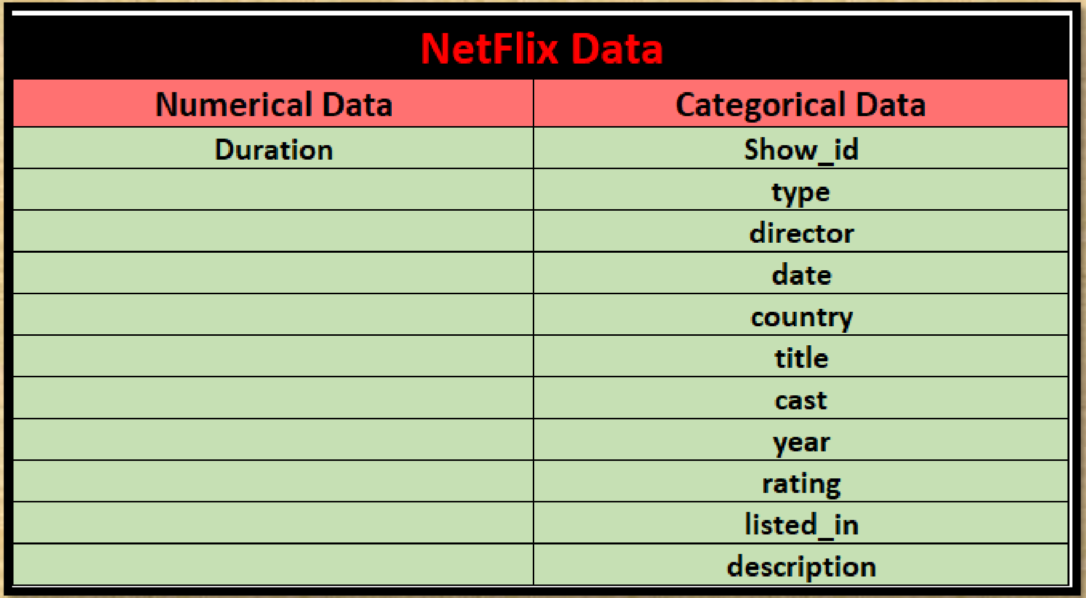
This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine.

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. Itwill be interesting to explore what all other insights can be obtained from the same dataset.

Content

1. Description of the data
2. Exploratory Data Analysis
3. Feature Engineering and its selection
4. Data Preprocessing
5. Clustering methods
6. Clustering Models
7. Conclusion

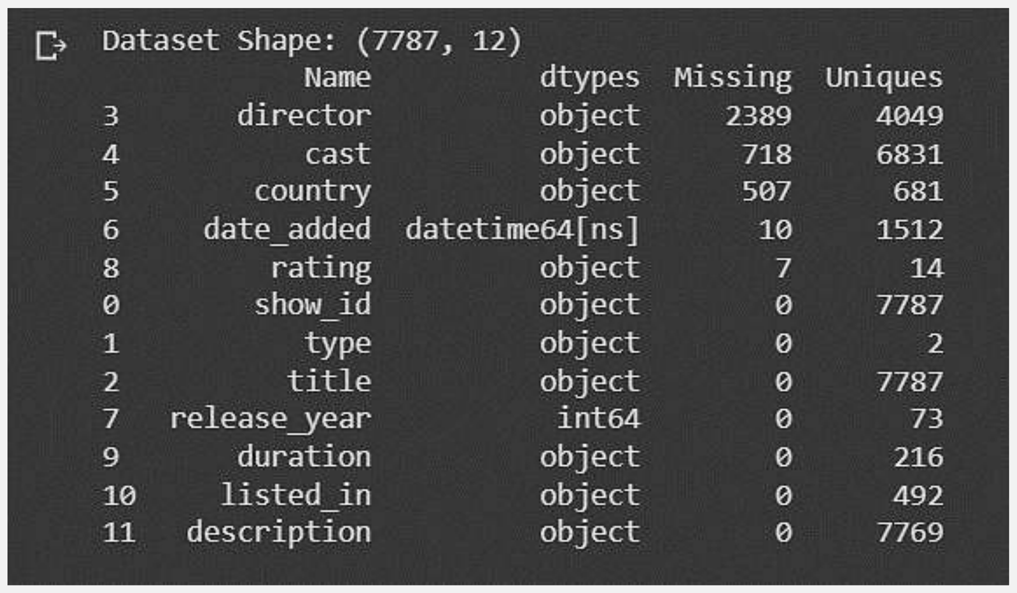
Data Description



NetFlix Data	
Numerical Data	Categorical Data
Duration	Show_id
	type
	director
	date
	country
	title
	cast
	year
	rating
	listed_in
	description

Data Description

- This dataset contains 7787 rows and 12 columns.
- Speaking about missing values, here, around 5 columns contains null values.
- One column, viz., director has huge missing values.

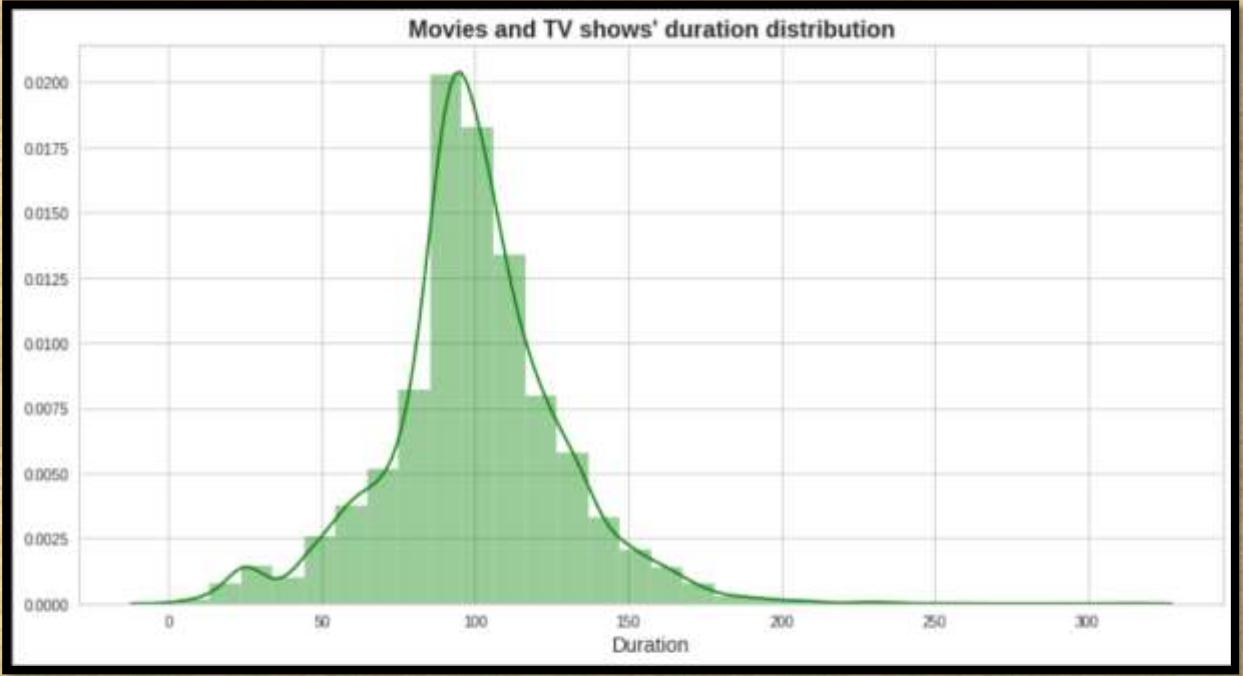


```
Dataset Shape: (7787, 12)
           Name          dtypes  Missing  Uniques
3      director          object     2389     4049
4          cast          object      718     6831
5       country          object      507      681
6    date_added  datetime64[ns]       10     1512
8        rating          object        7       14
0       show_id          object        0     7787
1          type          object        0        2
2         title          object        0     7787
7  release_year           int64        0       73
9      duration          object        0      216
10    listed_in          object        0      492
11  description          object        0     7769
```

Exploratory Data Analysis

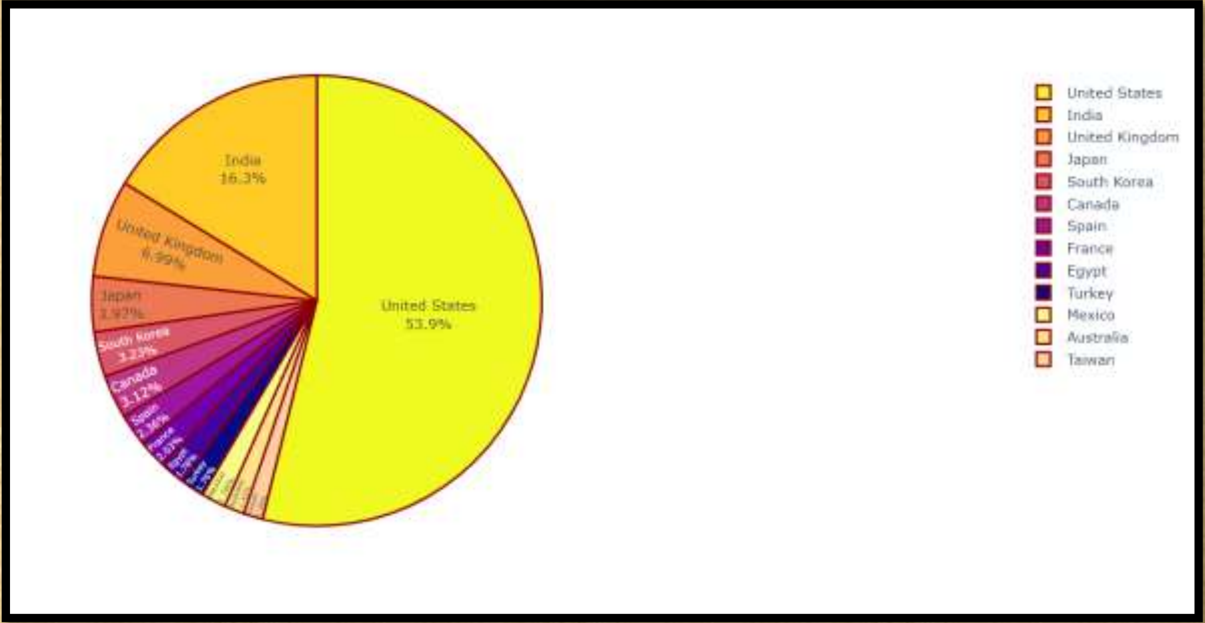
Duration Distribution

- This shape looks like a bell curve approximately. And we can see that most of the movies/ shows have the duration around 60mins to 150mins



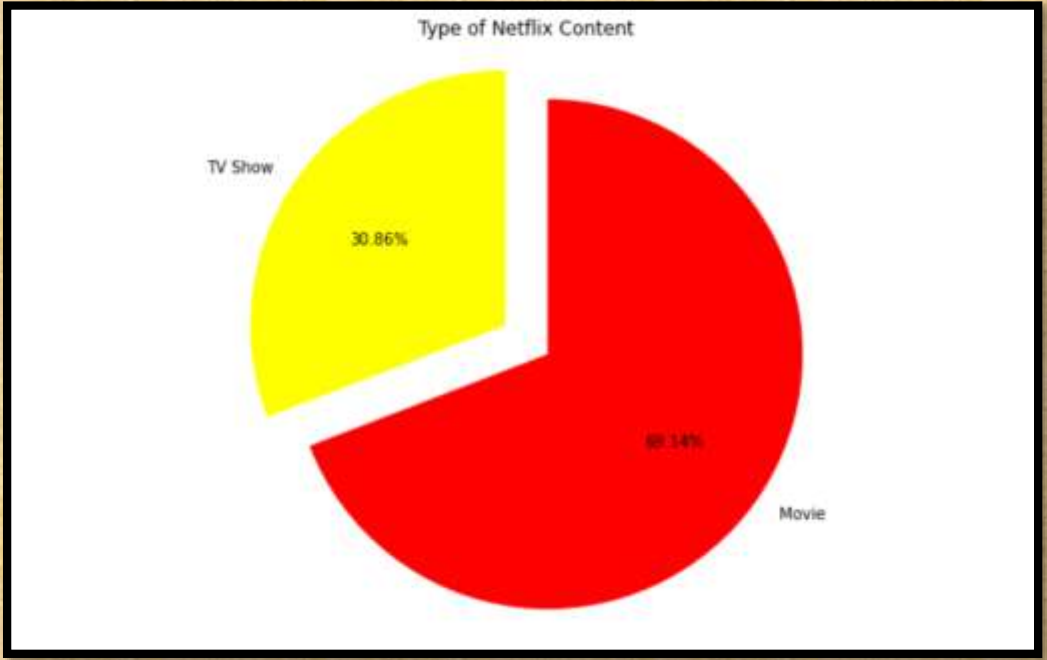
Country wise Movies/Shows produced.

- This pie chart tells us that 53.9% of the producers who produced Netflix movies/shows are from United States followed by India around 16.3%



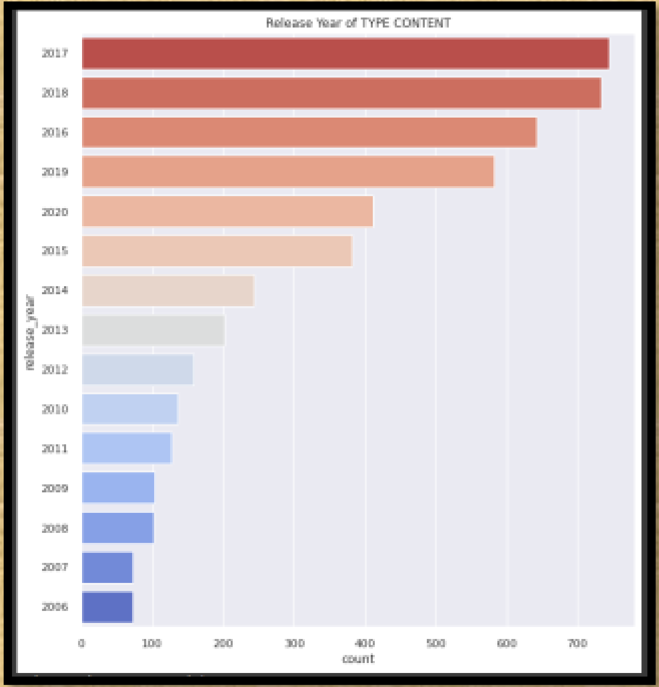
Types of Netflix's content

- Around 69.14% of the data is of Movie content and 30.86% of the data is of Netflix TV shows.

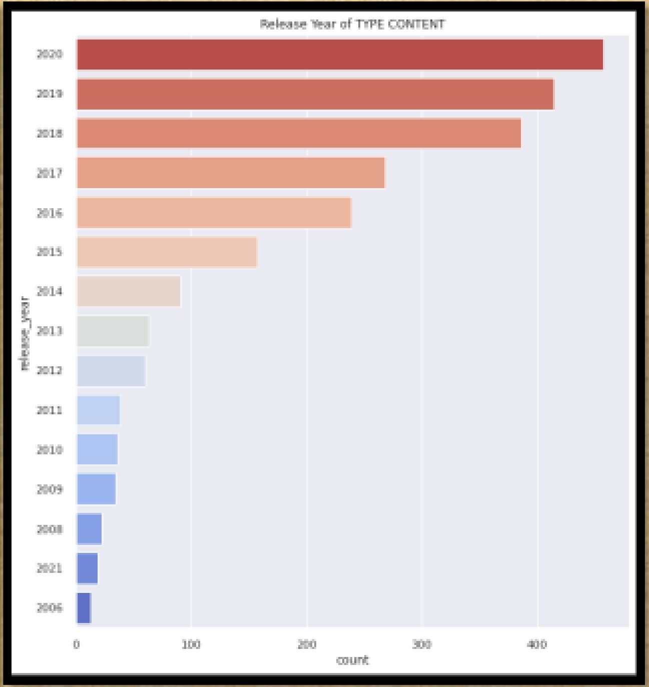


Release year of Type Content

- Count of movies is high in the year 2017 than slight decrease in the count in 2018.
- Count of Shows is high in 2020 year.

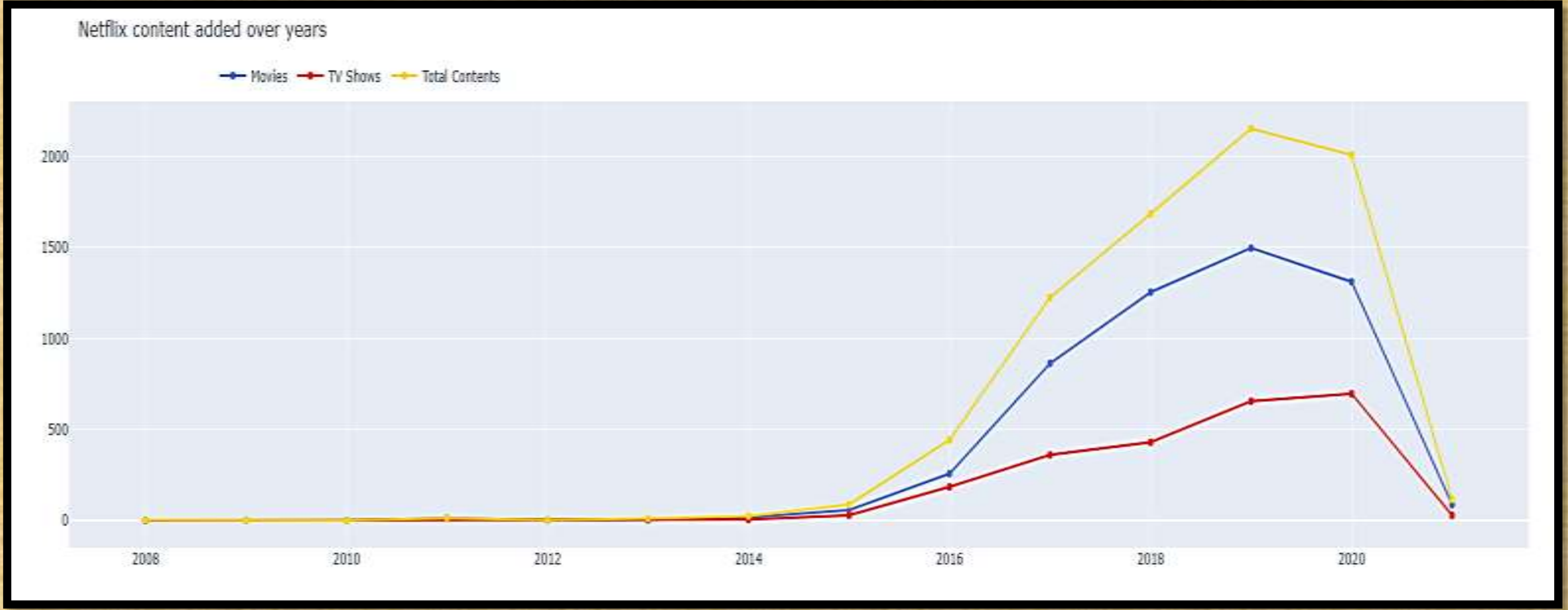


Movies

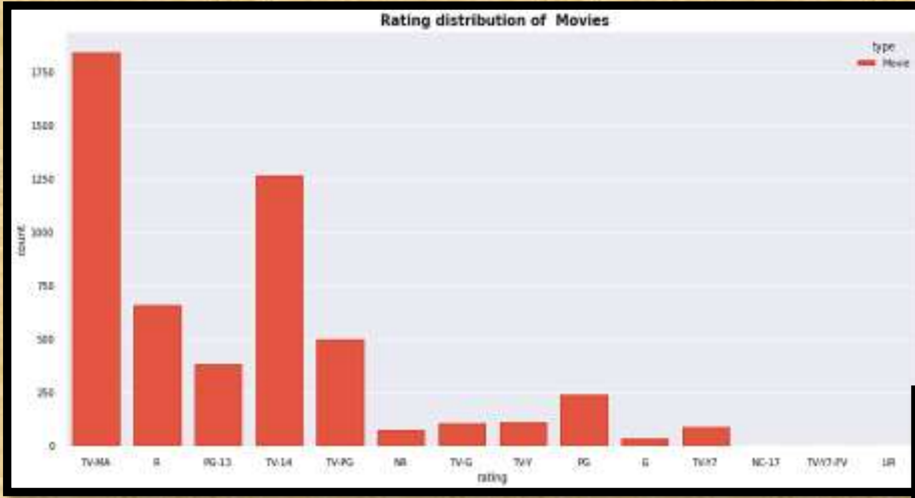


Shows

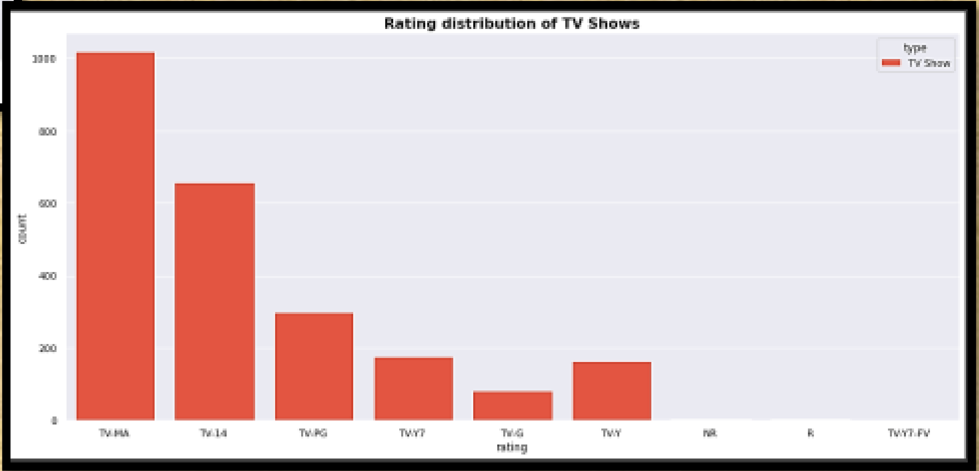
Netflix content added over years



Rating distribution

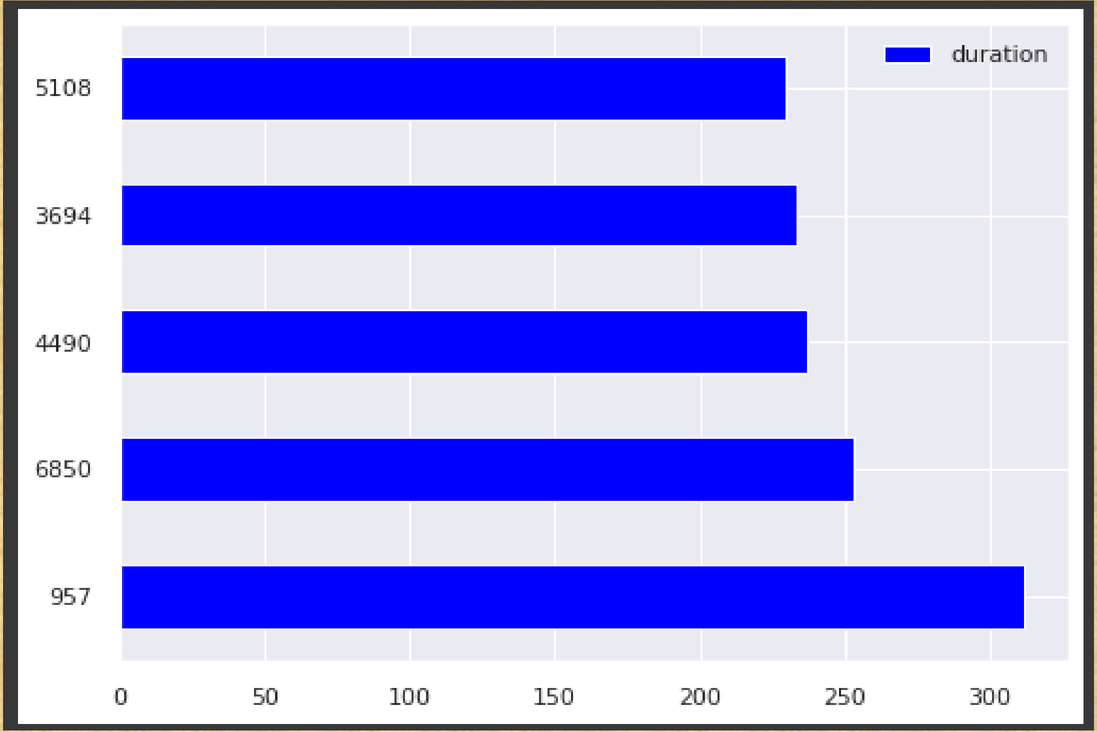


Both Movies and Shows have highest count of ‘TV-MA’ = TV Mature Audience rating, i.e., more than 1750 TV-MA rating contain Movies and more than 1000 TV-MA rating contain shows.



Top 5 movies with highest movies length

- Movie id 957 have the duration more than 300mins
- 5th position is Movie Id 5108 which have the duration length is around 230mins.



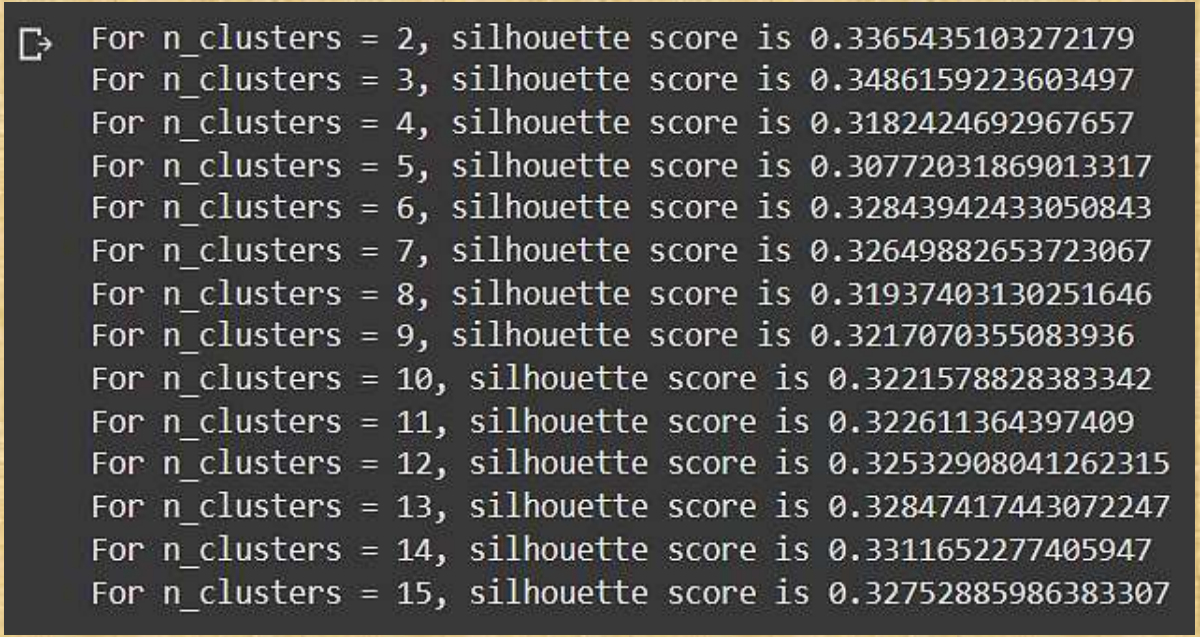
Data Preprocessing

1. Working on the text based features (description, listed_in)
2. Removing punctuations and stop words from text features
3. Stemming process applied for those text features.

Clustering Methods

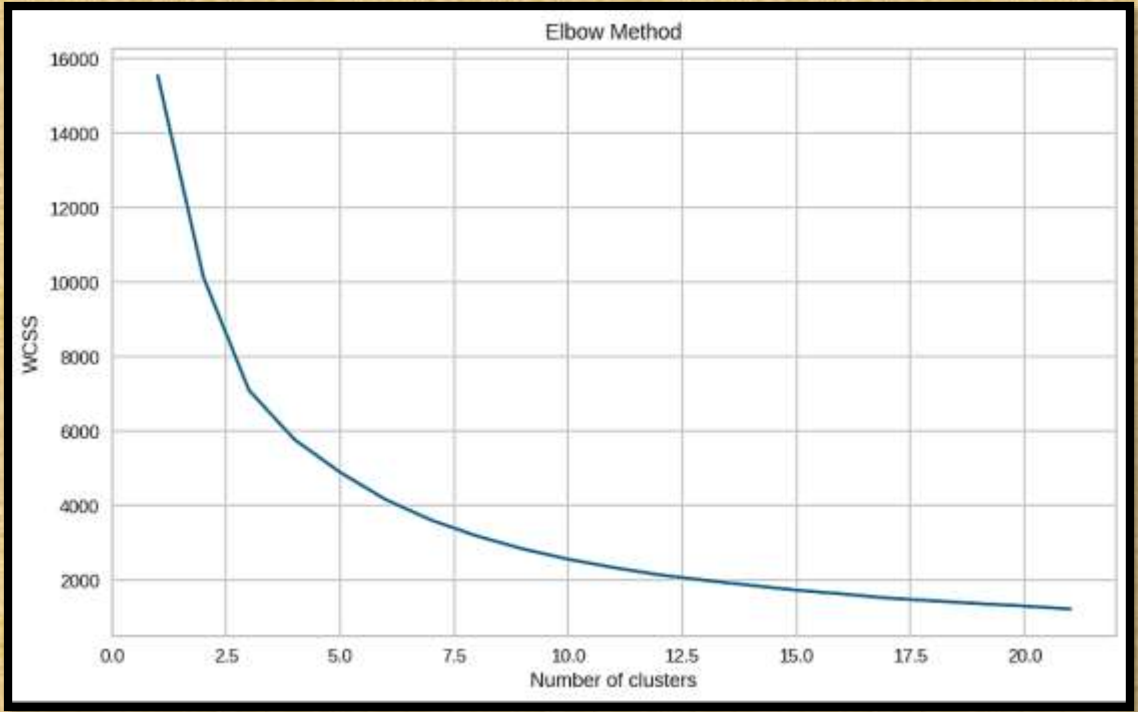
Silhouette Method

Best score = 0.34861 for n_clusters = 3

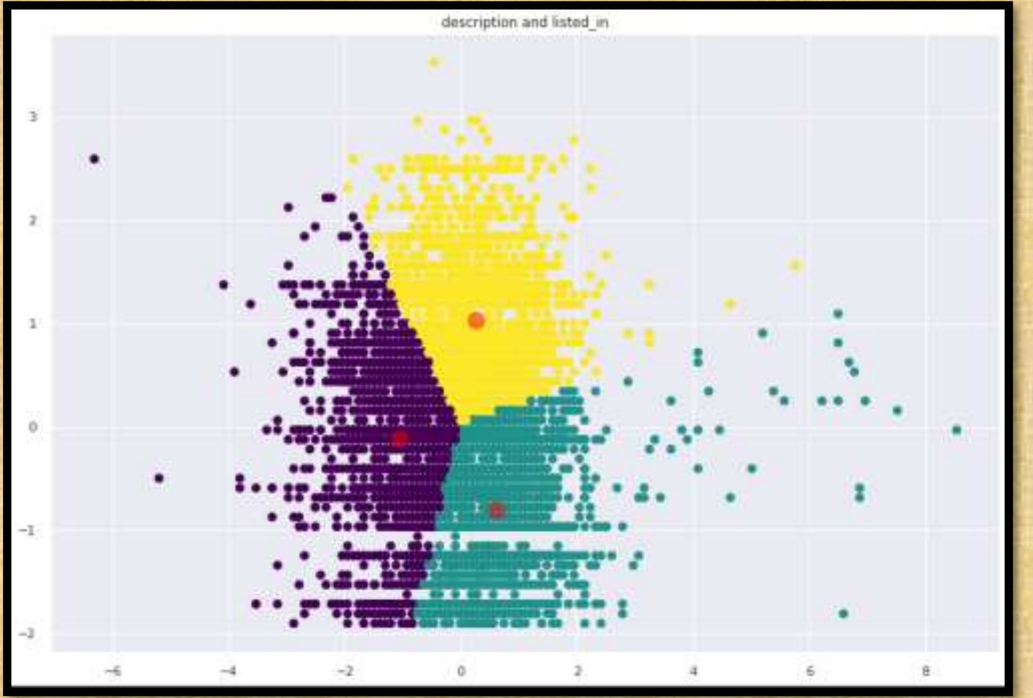


```
For n_clusters = 2, silhouette score is 0.3365435103272179
For n_clusters = 3, silhouette score is 0.3486159223603497
For n_clusters = 4, silhouette score is 0.318242469296765 7
For n_clusters = 5, silhouette score is 0.30772031869013317
For n_clusters = 6, silhouette score is 0.32843942433050843
For n_clusters = 7, silhouette score is 0.32649882653723067
For n_clusters = 8, silhouette score is 0.31937403130251646
For n_clusters = 9, silhouette score is 0.3217070355083936
For n_clusters = 10, silhouette score is 0.32215788283833342
For n_clusters = 11, silhouette score is 0.322611364397409
For n_clusters = 12, silhouette score is 0.32532908041262315
For n_clusters = 13, silhouette score is 0.32847417443072247
For n_clusters = 14, silhouette score is 0.331165227740594 7
For n_clusters = 15, silhouette score is 0.32752885986383307
```

Elbow Method

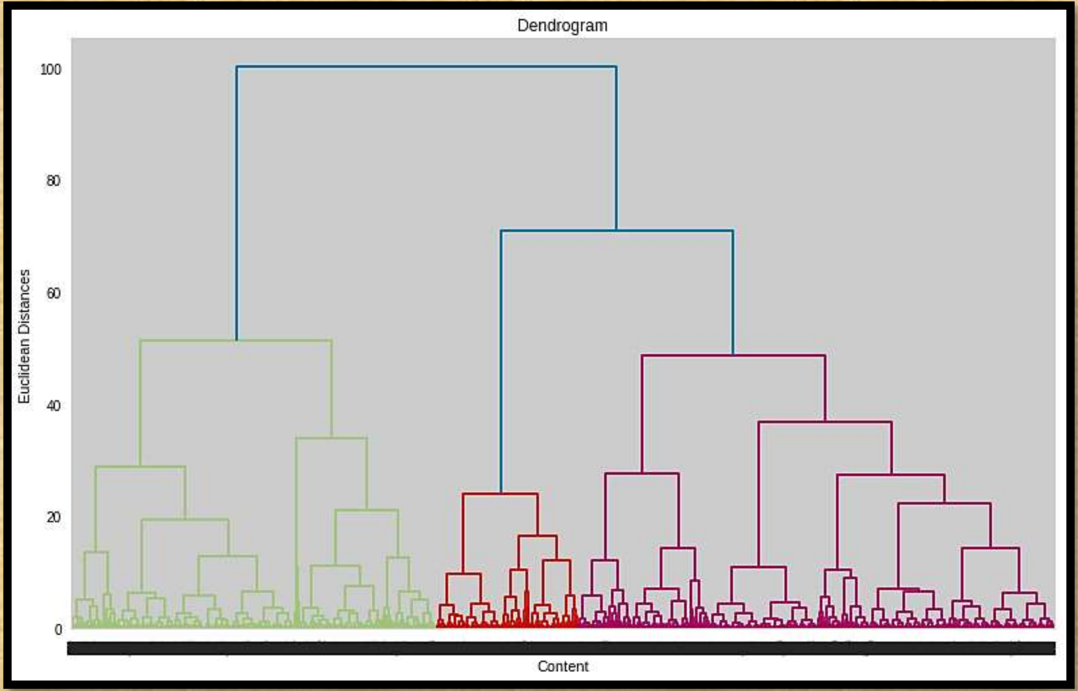


K mean Clustering



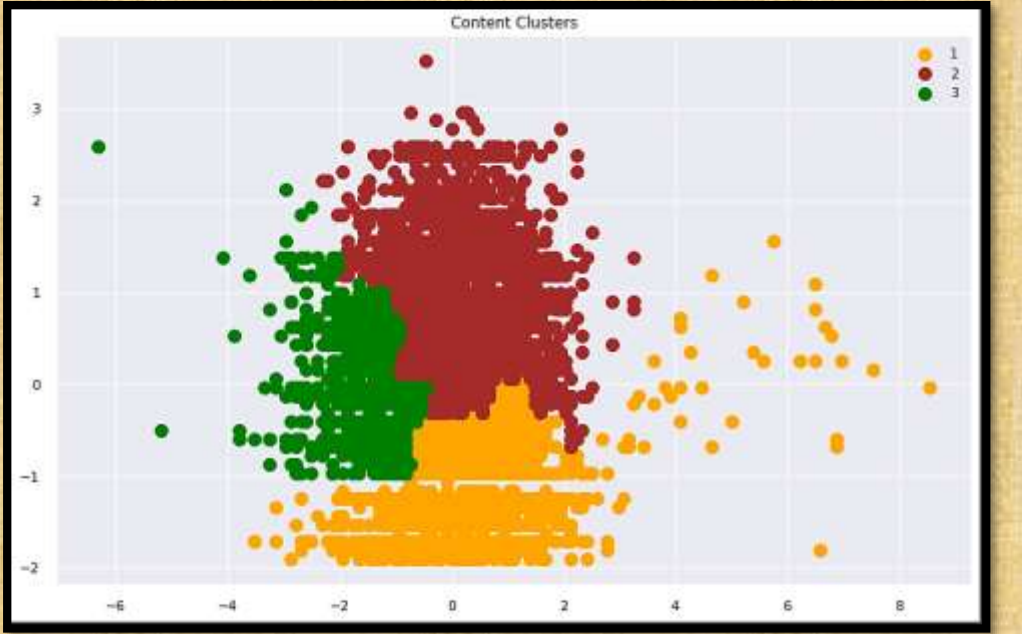
K = 3

Hierarchical clustering



Dendrogram

Hierarchical clustering



Agglomerative Clustering

Conclusion

After exploring the dataset, the percentage distribution of the types of content in Netflix is 69.14% contains MOVIES and 30.86% contains TV SHOWS.

With analysing the content added over years we got to know that in recent years Netflix is focusing on MOVIES than TV SHOWS, i.e., movies is increased by 80% and Tv shows is increased by approx. 70% compare to 2016 data.

Applying the silhouette score method for n range clusters, we got the best score which is 0.348 for 3 clusters.

Speaking about other different cluster methods, K Mean, hierarchical, agglomerative clustering on data we got the best cluster arrangements.

Optimal number of cluster = 3.

THANK YOU