Capstone Project-4

Netflix Movies & Tv show Clustering

Ashfaque sayyed

Contents

- Exploratory Data Analysis
- Data Cleaning & Feature Engineering
- Data Visualization
- Feature Engineering
- Model Building
- Evaluation Metrics
- Conclusion

Problem statement

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 nearlytripled. It will be interesting to explore what all other insights can be obtained from the same dataset.

Dataset:

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.

Data Summary

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

11. listed_in : Genere

12. description: The Summary description

Understanding Data

The dataset consist of 12 columns and around 7700 rows.

In this project we had following tasks:

1. Exploratory Data Analysis
2. Understanding what type content is available in different countries
3. Is Netflix has increasingly focusing on TV rather than movies in recent years.
4. Clustering similar content by matching text-based features

Data Cleaning

Null values checking

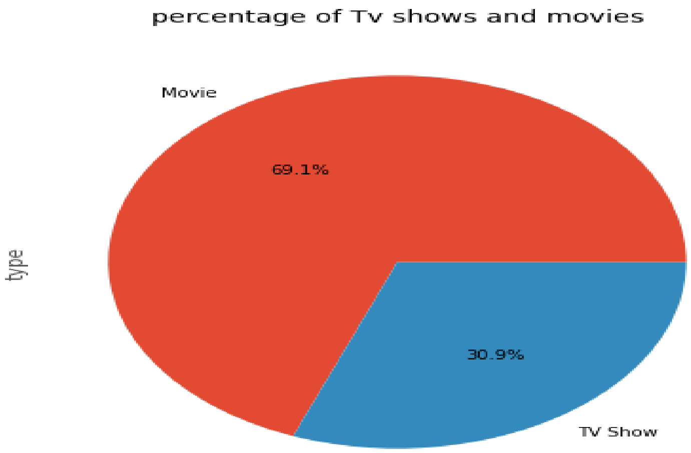
- In our dataset director, cast, country, date_added and Rating columns are having null values.
- So I filled the missing values with ‘unknown’ using fillna.

Duplicate values

- In our dataset we don’t have any duplicate records.

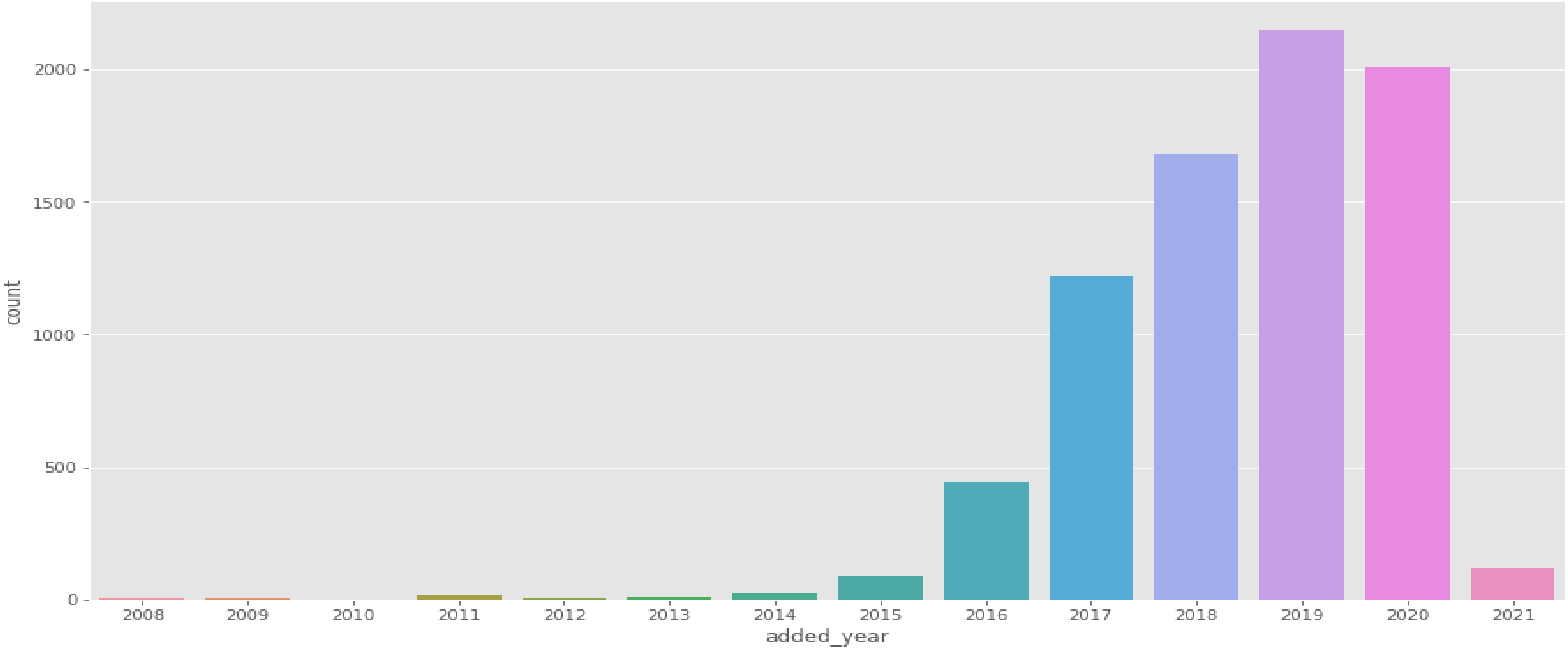
Exploratory Data Analysis

1.Data Visualization-Univariate Analysis



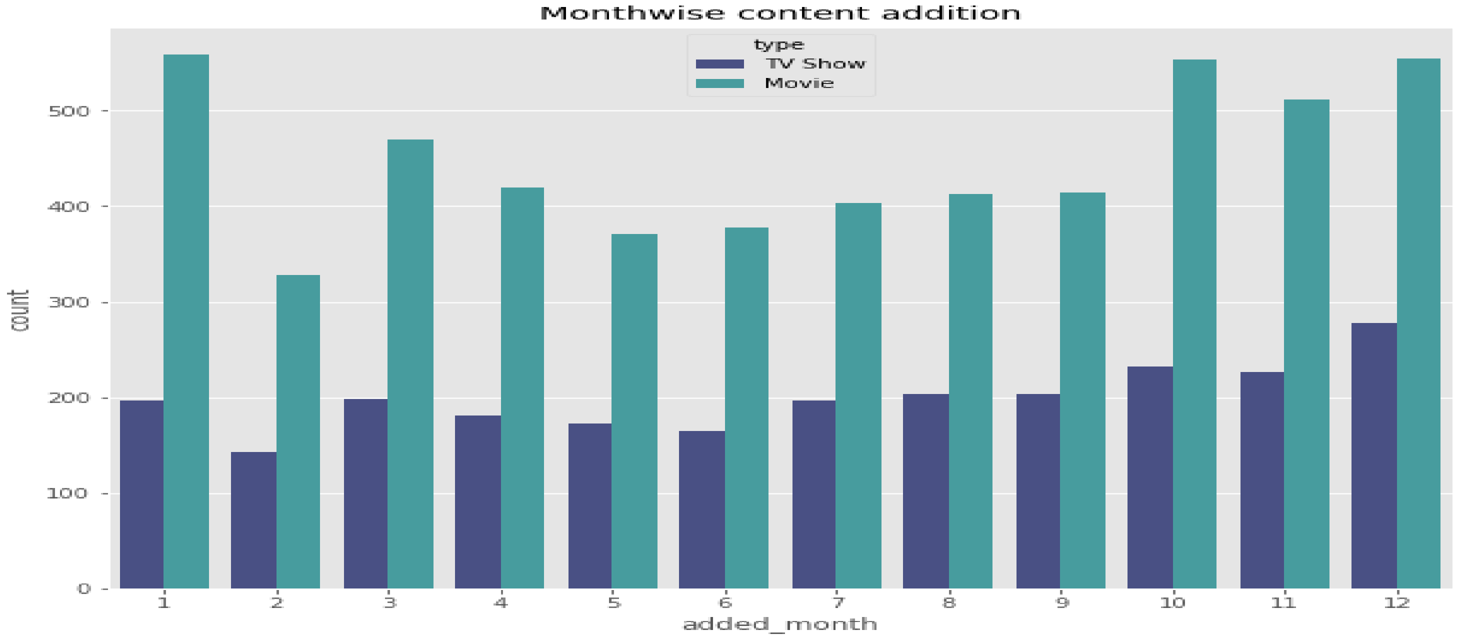
In our dataset there is around 69% content as movies and remaining 31% as TV shows. Netflix is releasing more movies than TV shows.

2.Movies added to netflix each year



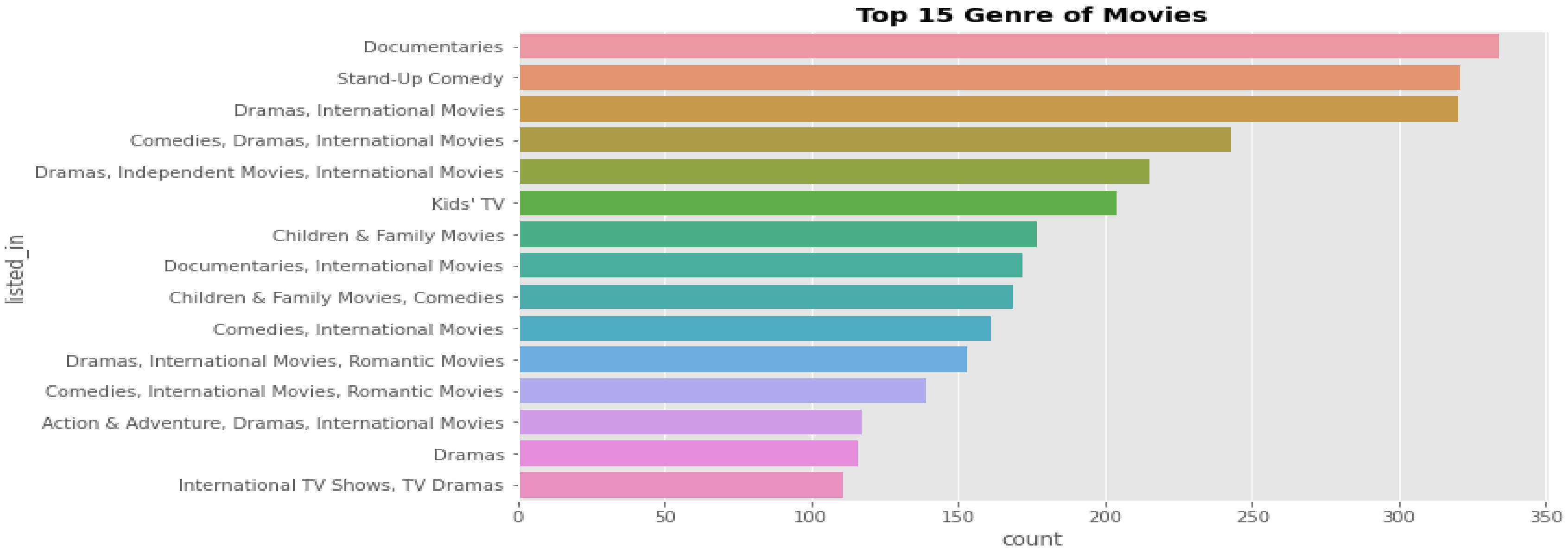
The number of shows and movies added are maximum for year 2019 and 2020, number of shows decreased for year 2021.

3.Content addition



we can clearly see that on **January**, **October** and in **December** there is more content added on netflix And in **February** very less amount of content added.

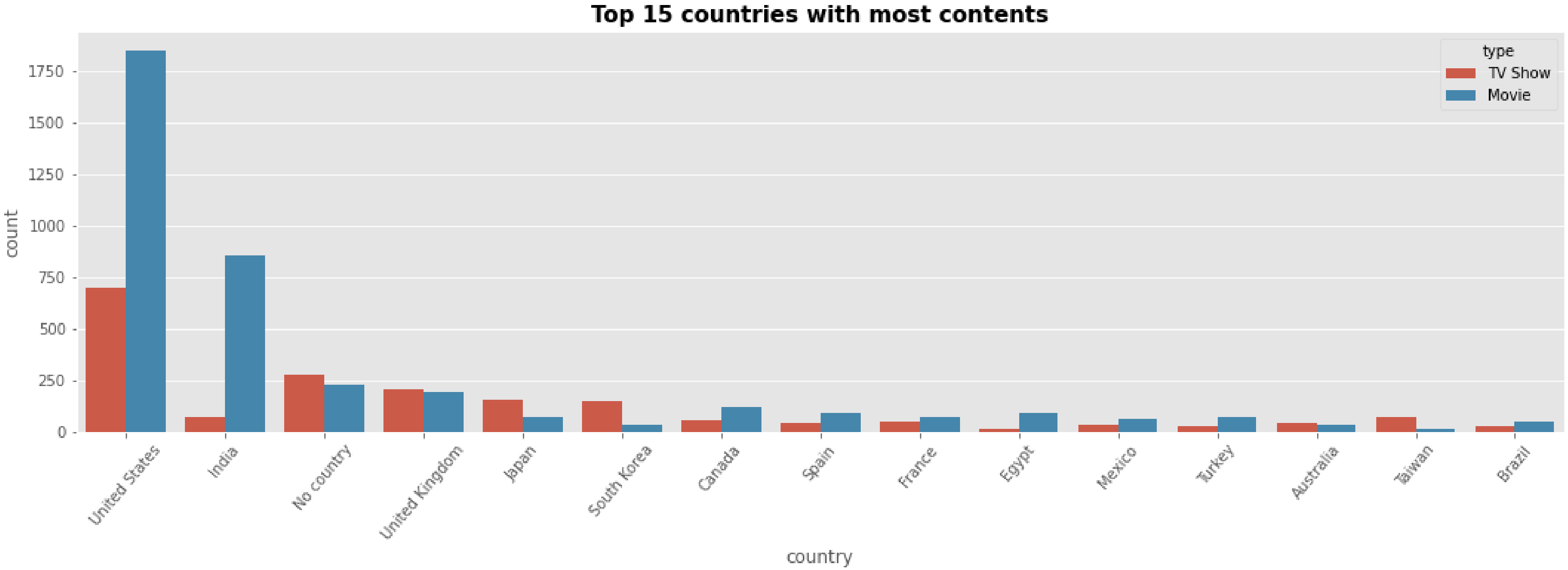
4.Top 15 Genres in Netflix



* Documentaries, stand-up comedy, Dramas and international movies are the top most genres in Netflix.

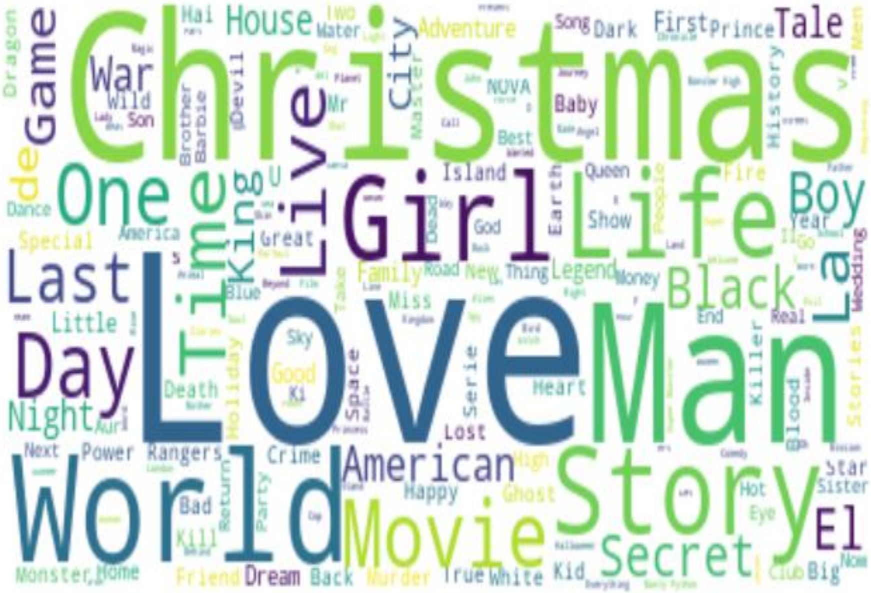
* International TV shows, tv dramas are rarest genre in Netflix.

5.Country



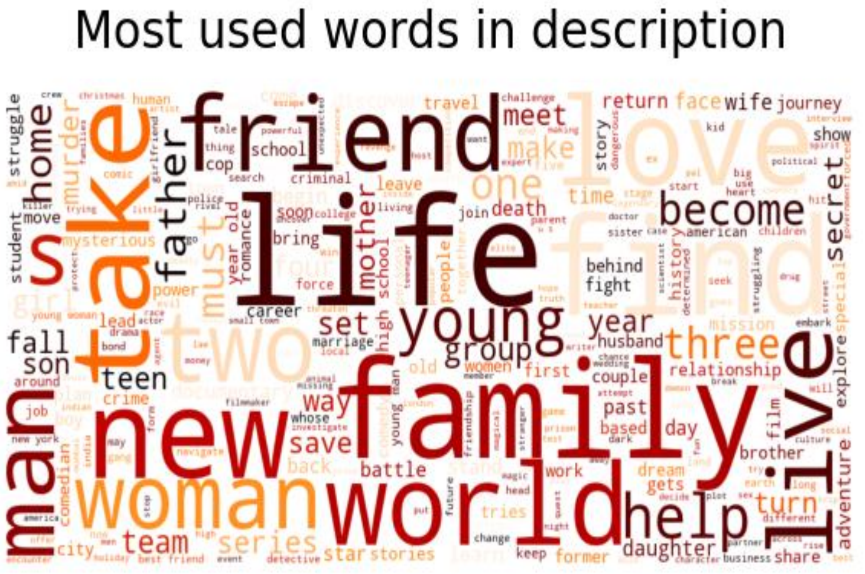
- United states has the highest number of movies and TV shows on the netflix, followed by India.
- In india, the number of movies are tremendously higher than Tv shows.

6.Word cloud



Title

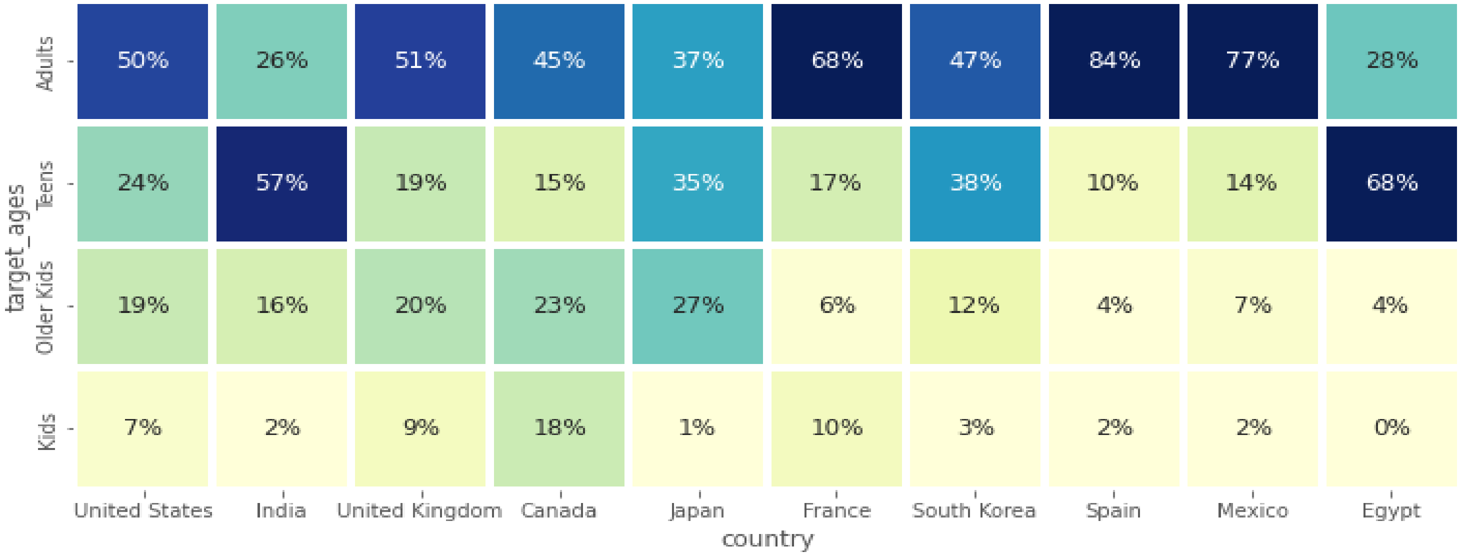
Most used words in description



Description

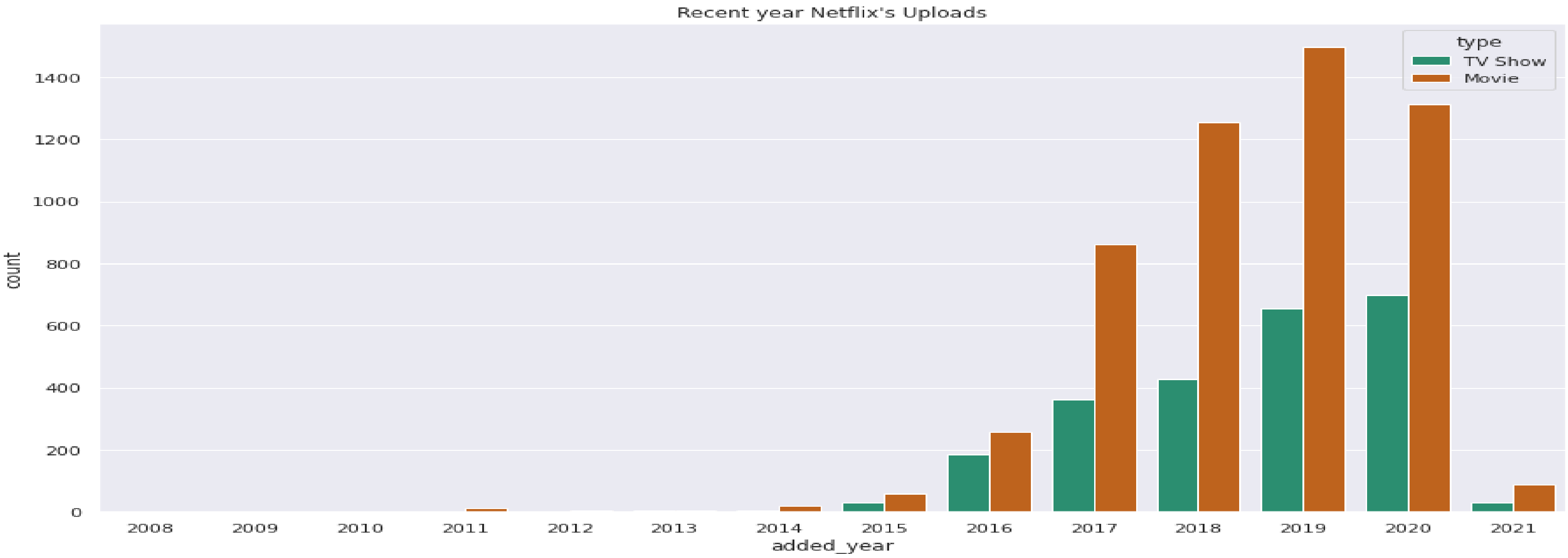
- It seems like words like 'world', 'life', 'man','Love','girl', and 'christmas' are very common in titles.
- Most occurring words in the description of the tv shows and movies are Family,life, friend,new,'take' and 'world'.

7. Heatmap



1. countries US, india, UK, canada, spain etc the netflix content is mostly adult.
2. Netflix content is more for teens in india and egypt, followed by south korea.
3. for older kids the content is maximum in japan(27%) followed by canada, united kingdom.
4. kids content is maximum in canada (18%) and there is no kids content released in egypt.

8.Is Netflix has increasingly focusing on TV rather than movies in recent years ??



- From above countplot we can clealry see that from 2017 number of Movies added increased tremendously, but at the same time TV shows added from 2017 are also increased but as comparison to Movies they are very less in numbers.

Feature Engineering

- **STEMMING -** Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma

- **TF-IDF -** Term frequency-inverse document frequency is a text vectorizer that transforms the text into a usable vector.

- **TF -** The term frequency is the number of occurrences of a specific term in a document. Term frequency represents every text from the data as a matrix whose rows are the number of documents and columns are the number of distinct terms throughout all documents.

- **IDF -** Document frequency is the number of documents containing a specific term. Document frequency indicates how common the term is. Inverse document frequency (IDF) is the weight of a term, it aims to reduce the weight of a term if the term's occurrences are scattered throughout all the documents.

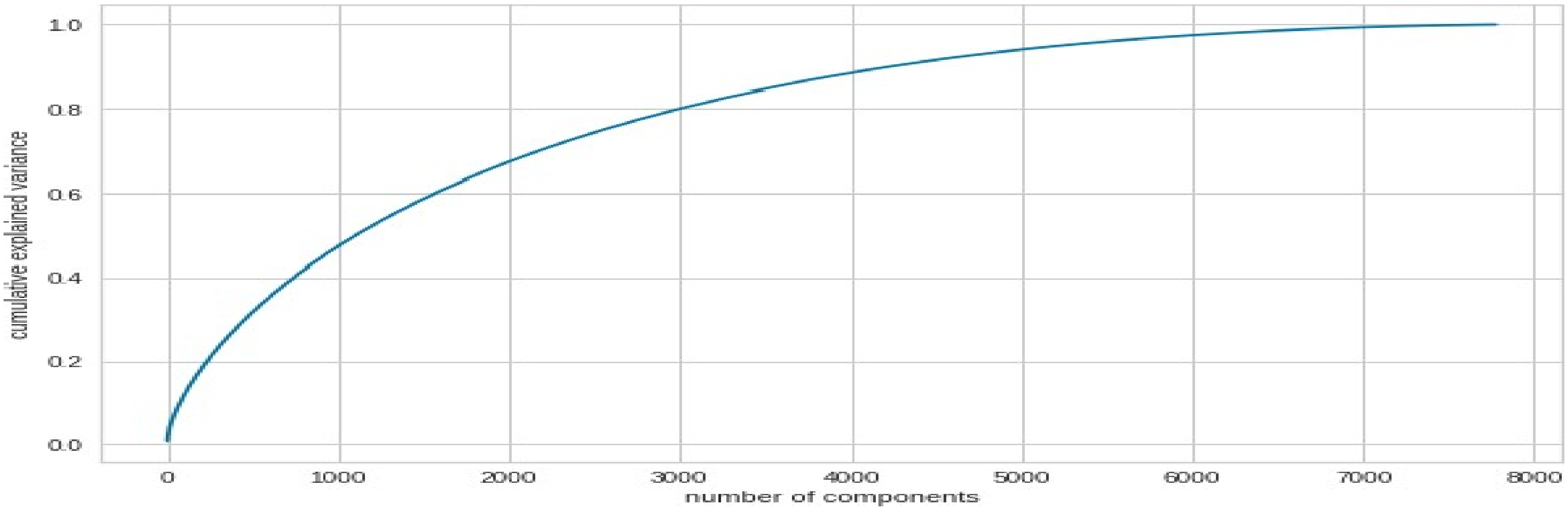
PCA- Principal Component Analysis

Principal Component Analysis is an unsupervised learning algorithm that is used forthe dimensionality reduction in machine learning. It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation.

It is a technique to draw strong patterns from the given dataset by reducing thevariances.

PCA generally tries to find the lower-dimensional surface to project thehigh-dimensional data.

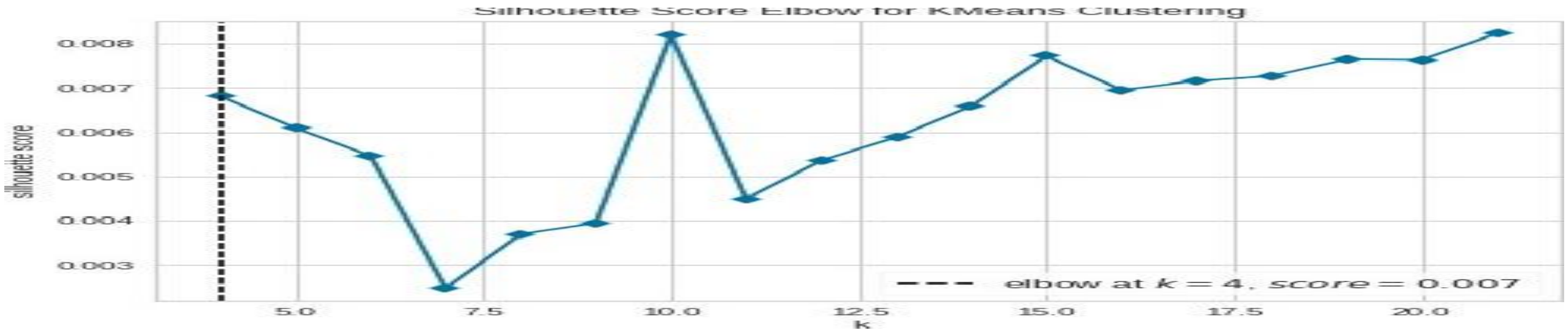
Cumulative Explained Variance



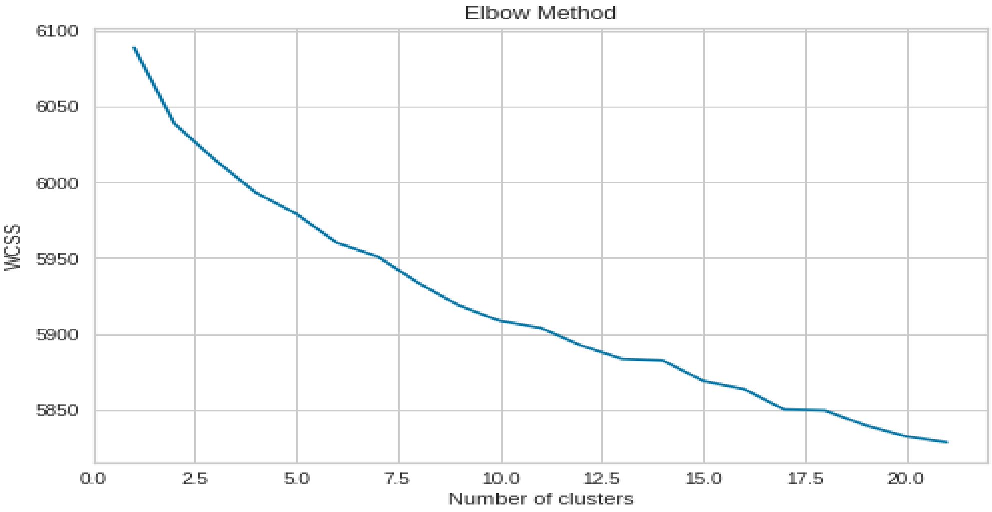
- Here we can clearly spot that 80% variance is explained by3000 components only.

5. Model Building

- **KMeans Clustering-**
- K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabelled dataset into different clusters.
- It is an iterative algorithm that divides the unlabelled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.



Elbow Method to get number of clusters

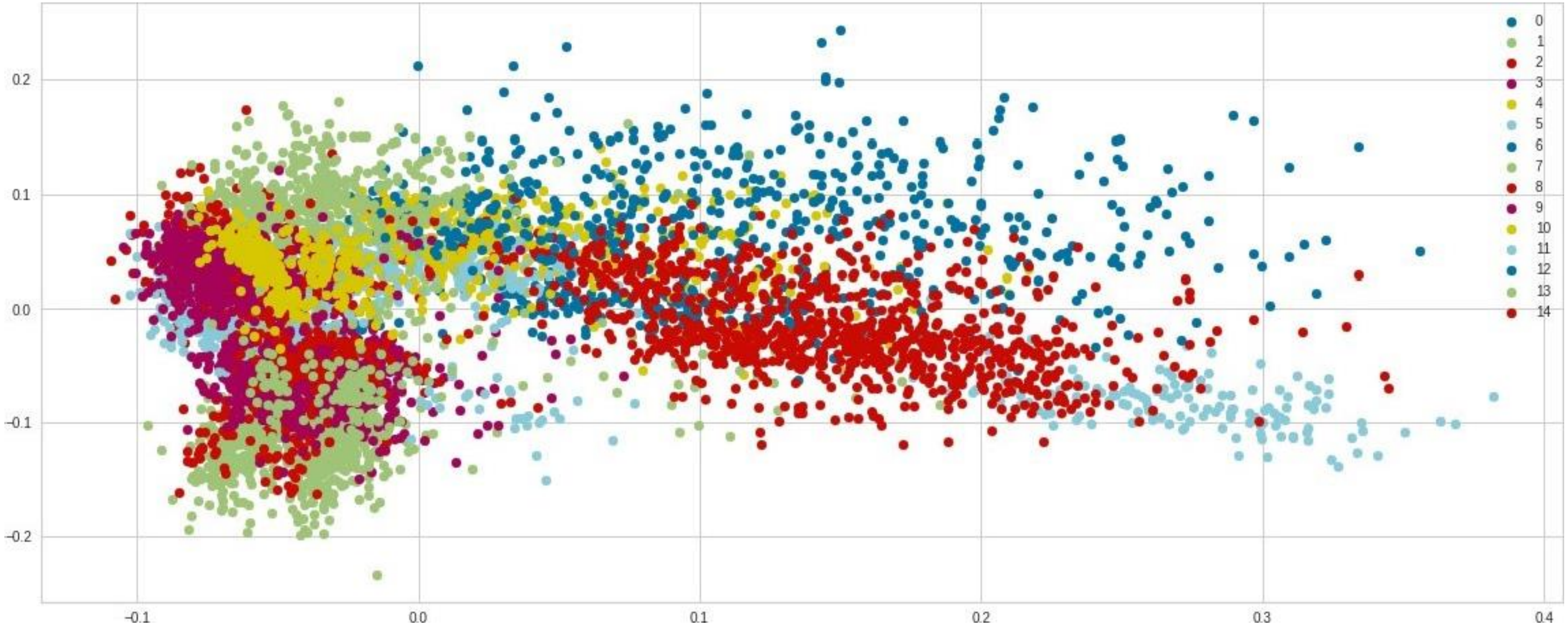


we will take no. of clusters as 15

The K-Elbow Visualizer implements the “elbow” method of selecting the optimal number of clusters for K-means clustering.

The elbow method runs k means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned centre.

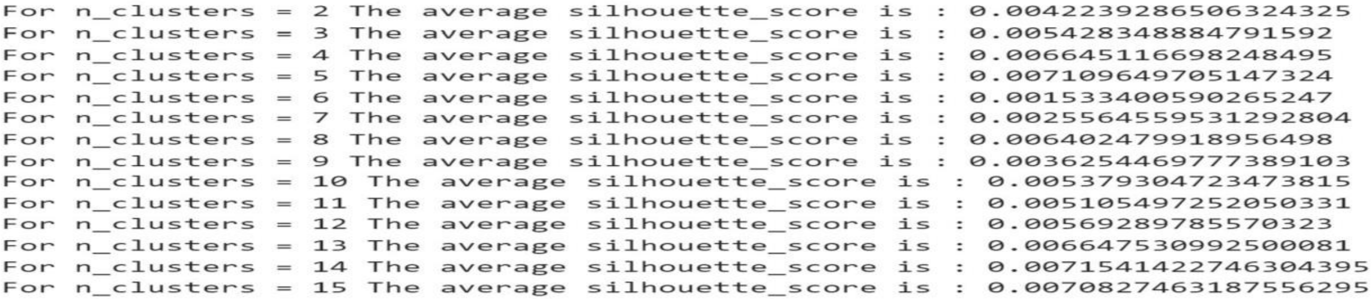
Predicted clusters visualization-



Silhauette score

Silhouette Score -

Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1,1].



```
For n_clusters = 2 The average silhouette_score is : 0.0042239286506324325
For n_clusters = 3 The average silhouette_score is : 0.005428348884791592
For n_clusters = 4 The average silhouette_score is : 0.006645116698248495
For n_clusters = 5 The average silhouette_score is : 0.007109649705147324
For n_clusters = 6 The average silhouette_score is : 0.001533400590265247
For n_clusters = 7 The average silhouette_score is : 0.0025564559531292804
For n_clusters = 8 The average silhouette_score is : 0.006402479918956498
For n_clusters = 9 The average silhouette_score is : 0.0036254469777389103
For n_clusters = 10 The average silhouette_score is : 0.005379304723473815
For n_clusters = 11 The average silhouette_score is : 0.005105497252050331
For n_clusters = 12 The average silhouette_score is : 0.00569289785570323
For n_clusters = 13 The average silhouette_score is : 0.006647530992500081
For n_clusters = 14 The average silhouette_score is : 0.0071541422746304395
For n_clusters = 15 The average silhouette_score is : 0.0070827463187556295
```

We selected number of clusters as 15 which in above calculations showing 0.00708 as silhouette score.

Conclusion

- In our dataset there is around 69% content as movies and remaining 31% as TV shows. Netflix is releasing more movies than TV shows.
- we can clearly see that on January October and in December there is more content added on netflix And in February very less amount of content added.
- Documentaries, stand-up comedy, Dramas and international movies are the top most genres in Netflix.
- United states has the highest number of content on the netflix, followed by India.
- US, india, UK, canada, spain etc netflix content is mostly adults.
- Netflix content is more for teens in india and egypt, followed by south korea.
- from 2017 number of Movies added increased tremendously, but at the same time TV shows added from 2017 are also increased but as comparison to Movies they are very less in numbers.
- In cumulative explained variance graph we got 80% of variance captured by 3000 components only, thats why we selected no. of components as 3000.
- We selected no. of clusters as 15 from Elbow method.
- Calculated silhouette score for 15 no. of clusters which was showing 0.008.
- 6. Then we plotted average silhouette score for clusters ranging from 2 to 16, and in that we get silhouette score 0.00708 for cluster=15 which is pretty close to earlier we calculated.

THANK YOU