Capstone Project – 4
Unsupervised M.L.

Netflix Movies and TV Shows Clustering
Unsupervised M.L.

Individual Project:

Pratik Vishwakarma

Table Of Contents

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.

Data Summary

- **show_id :** *Unique ID for every Movie / Tv Show*
- **type :** *A Movie or TV Show*
- **title :** *Title of the Movie / Tv Show*
- **director :** *Director of the Movie*
- **cast :** *Actors involved in the movie / show*
- **country :** *Country where the movie / show was produced*
- **date_added :** *Date it was added on Netflix*
- **release_year :** *Actual Release year of the movie / show*
- **rating :** *TV Rating of the movie / show*
- **duration :** *Total Duration - in minutes or number of seasons*
- **listed_in :** *Generes*
- **description:** *The Summary description*

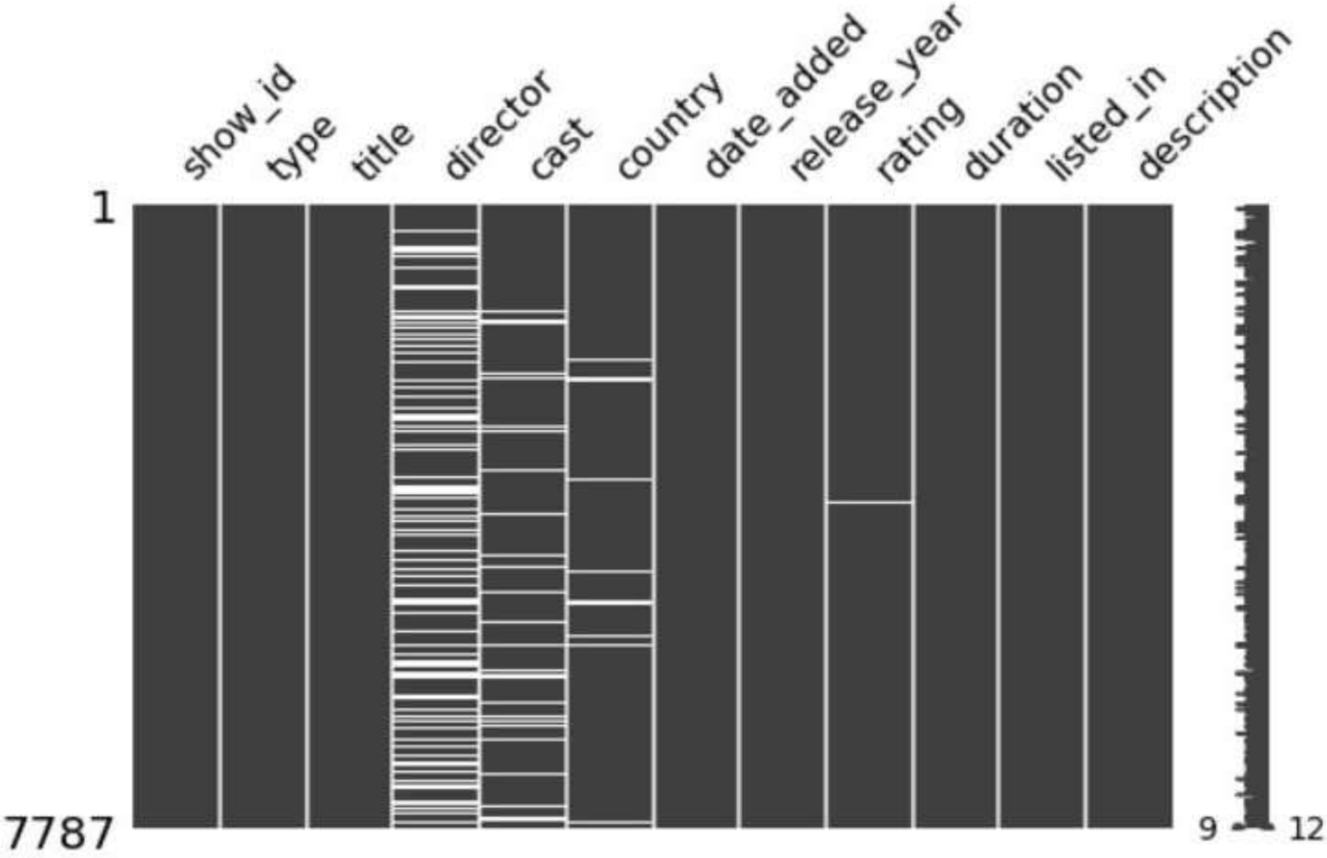
Basic Data Exploration

- *The dataset has **7787 observations and 12 features(columns).***
- *The dataset consists of **eleven textual columns and one numeric column('release_year')***
- *No Duplicate values.*

```
Dataset Shape: (7787, 12)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7787 entries, 0 to 7786
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       7787 non-null   object
 1   type          7787 non-null   object
 2   title         7787 non-null   object
 3   country       7280 non-null   object
 4   date_added    7777 non-null   object
 5   release_year  7787 non-null   int64
 6   rating        7780 non-null   object
 7   duration      7787 non-null   object
 8   listed_in     7787 non-null   object
 9   description   7787 non-null   object
dtypes: int64(1), object(9)
memory usage: 6.0 MB
```

EDA (Checking NaN values)



- **Null values present in this columns**
  - director
  - cast
  - country
  - Rating
- **No missing value present in this columns**
  - show_id
  - type
  - title
  - date_added
  - release_year
  - duration
  - listed_in
  - description

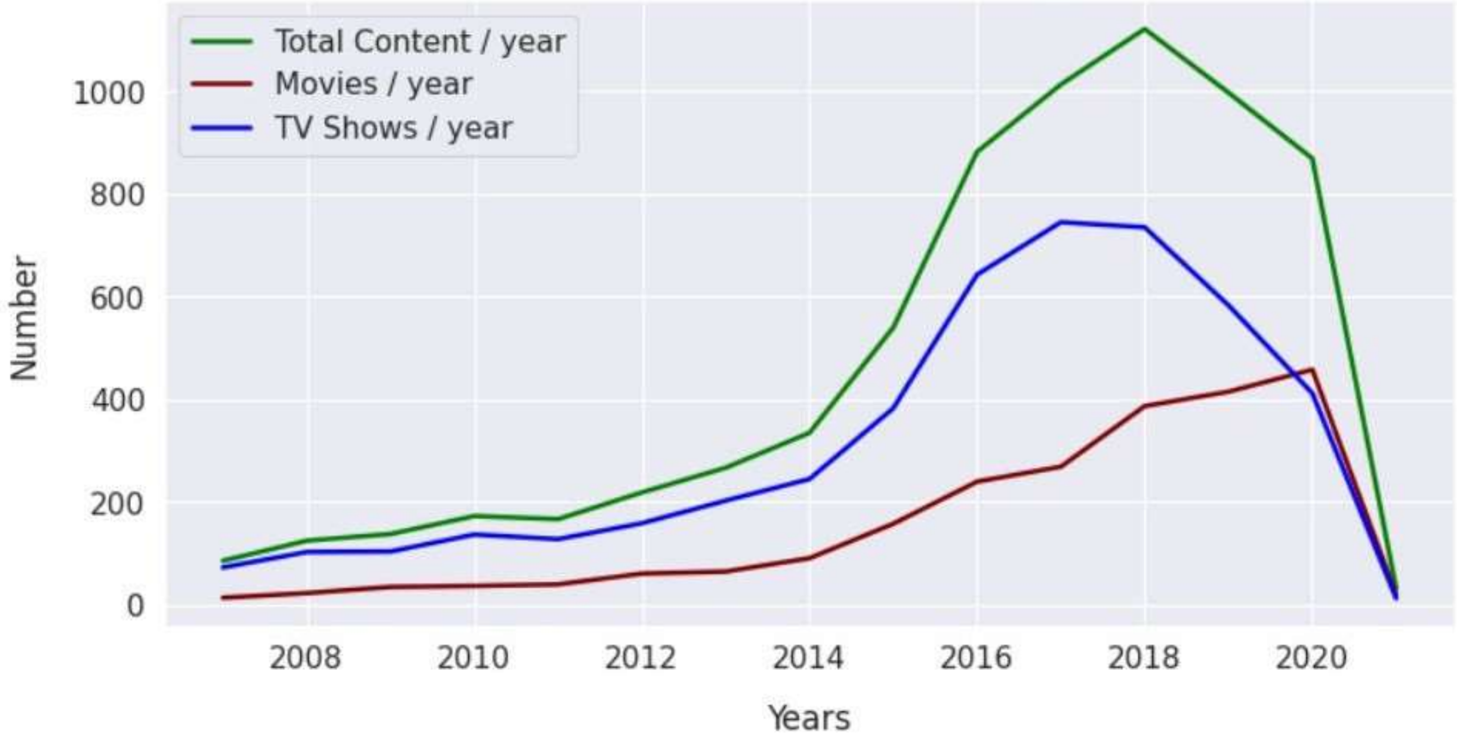
Data Cleaning

- Removing unnecessary columns ***like 'director','cast'***
- Dropping all the NaN containing date_added observations*(Only 10 observations was there)*
- Created 4 new columns
    - **No_of_categories** based on *listed_in*
    - *Date_added_month* based on *date_added*

	listed_in	no_of_category
0	International TV Shows, TV Dramas, TV Sci-Fi &...	3
1	Dramas, International Movies	2
2	Horror Movies, International Movies	2
3	Action & Adventure, Independent Movies, Sci-Fi...	3
4	Dramas	1

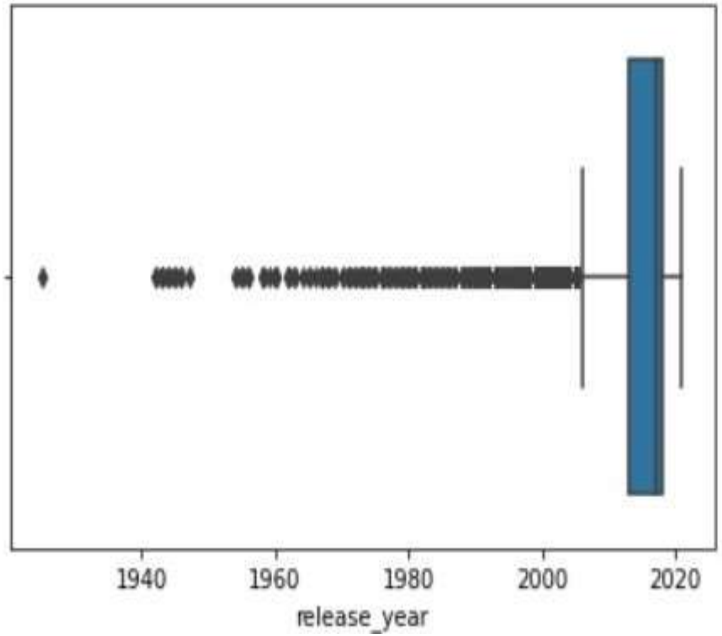
	December	October	January	November	March	September	August	April	July	June	May	February
date_added_month	817	780	746	730	661	614	612	596	592	538	537	466

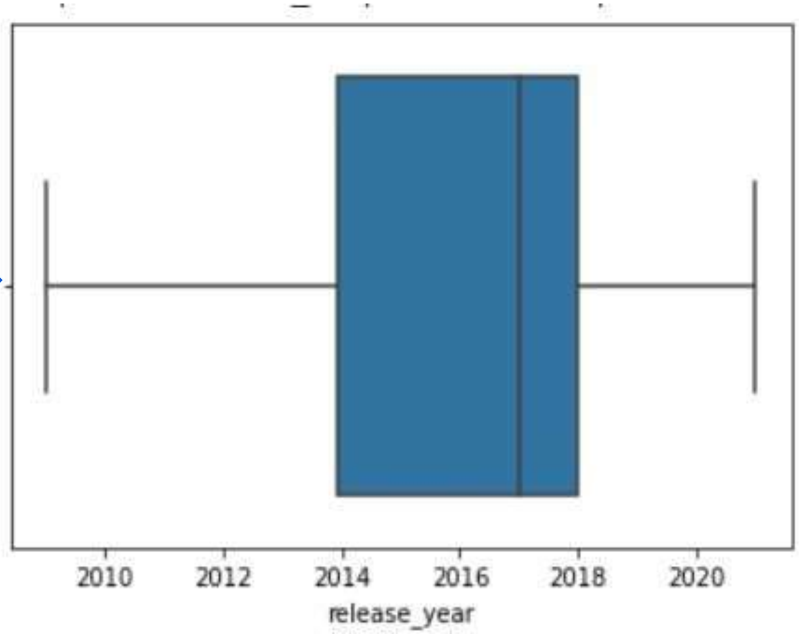
Production Yearly Growth



- Can you say what’s the reason of that boom growth??

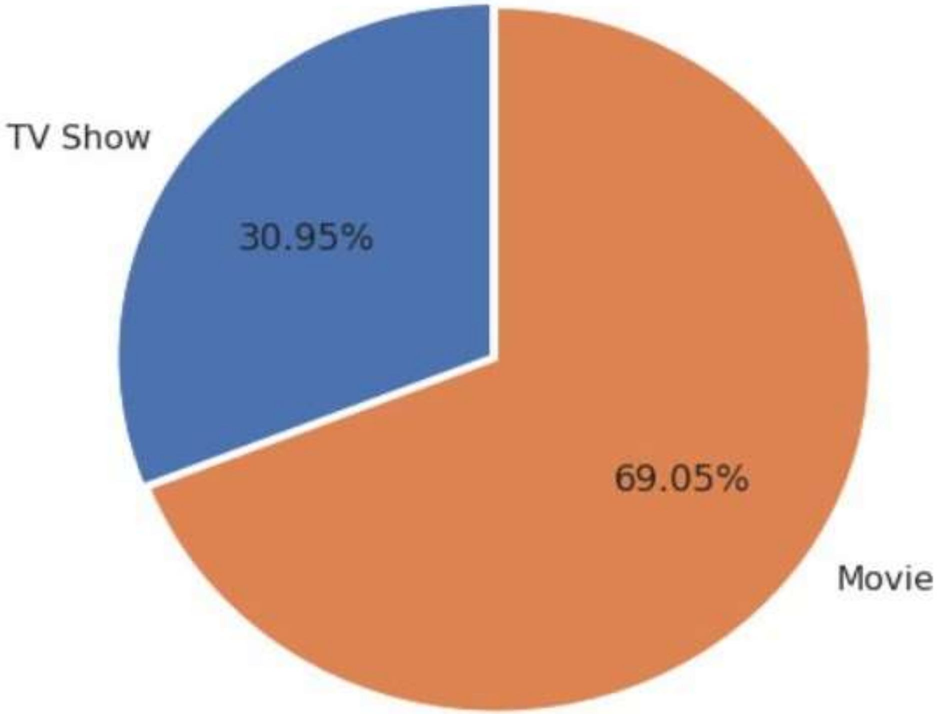
Checking Outliers





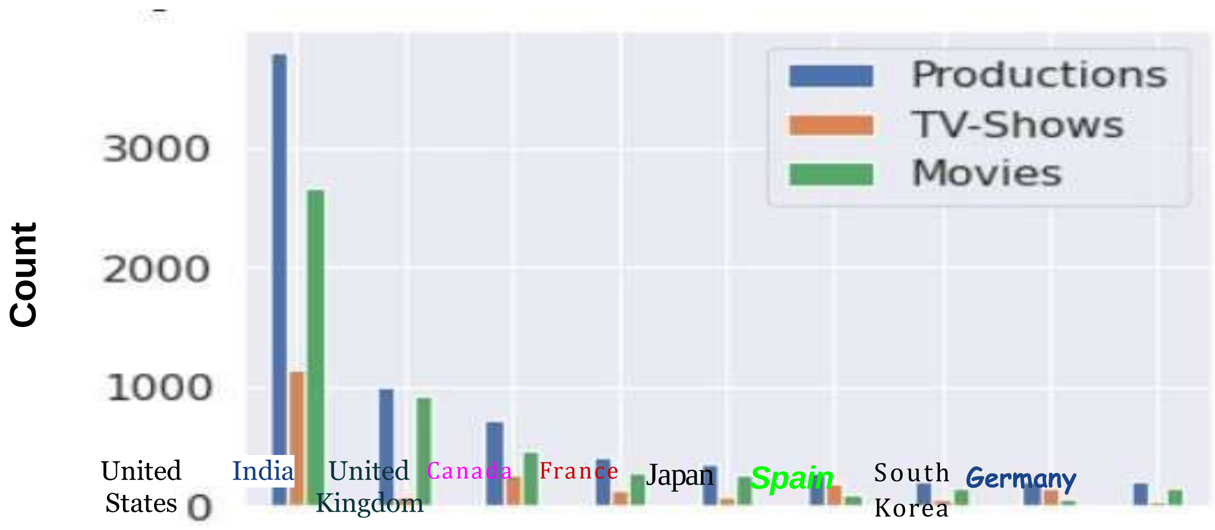
Replaced outliers values with mean value of *release_year*

Tv shows or Movies ??



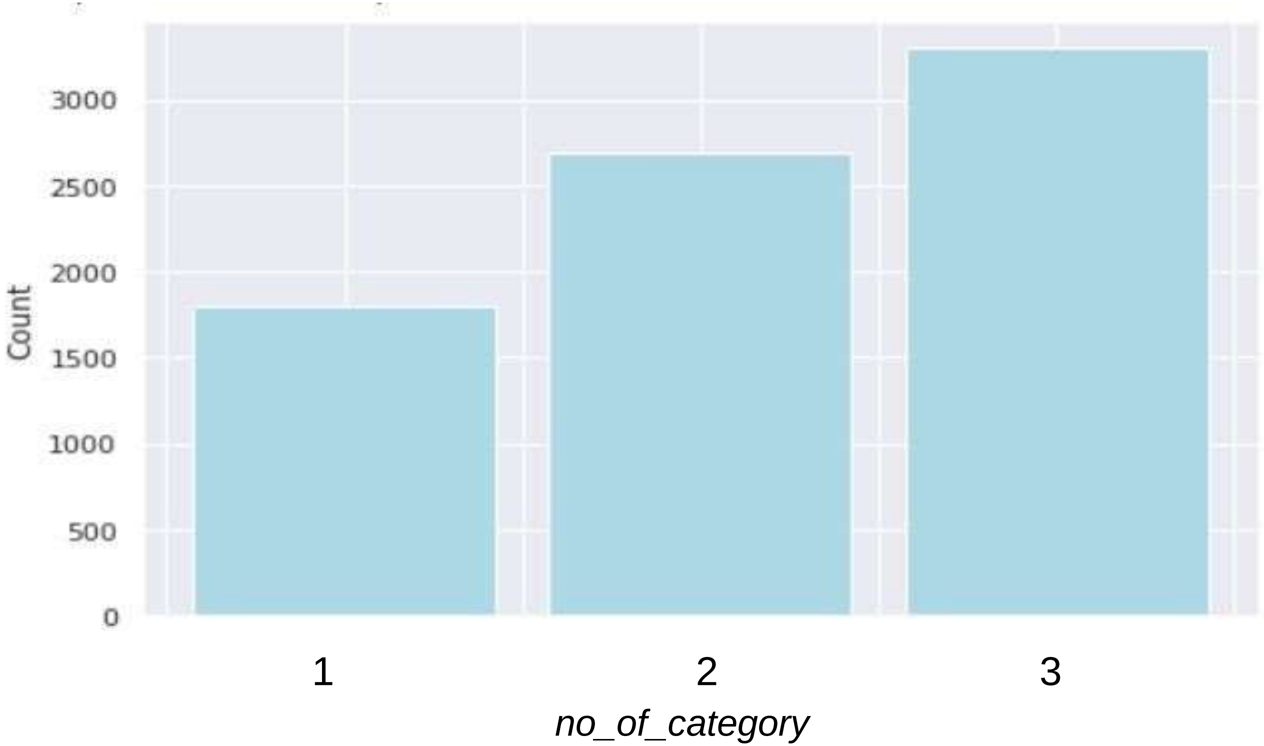
- Most of the contents are *Movies*
- *Less than ⅓ content are Tv Shows*

Countries Producing Most No of Contents



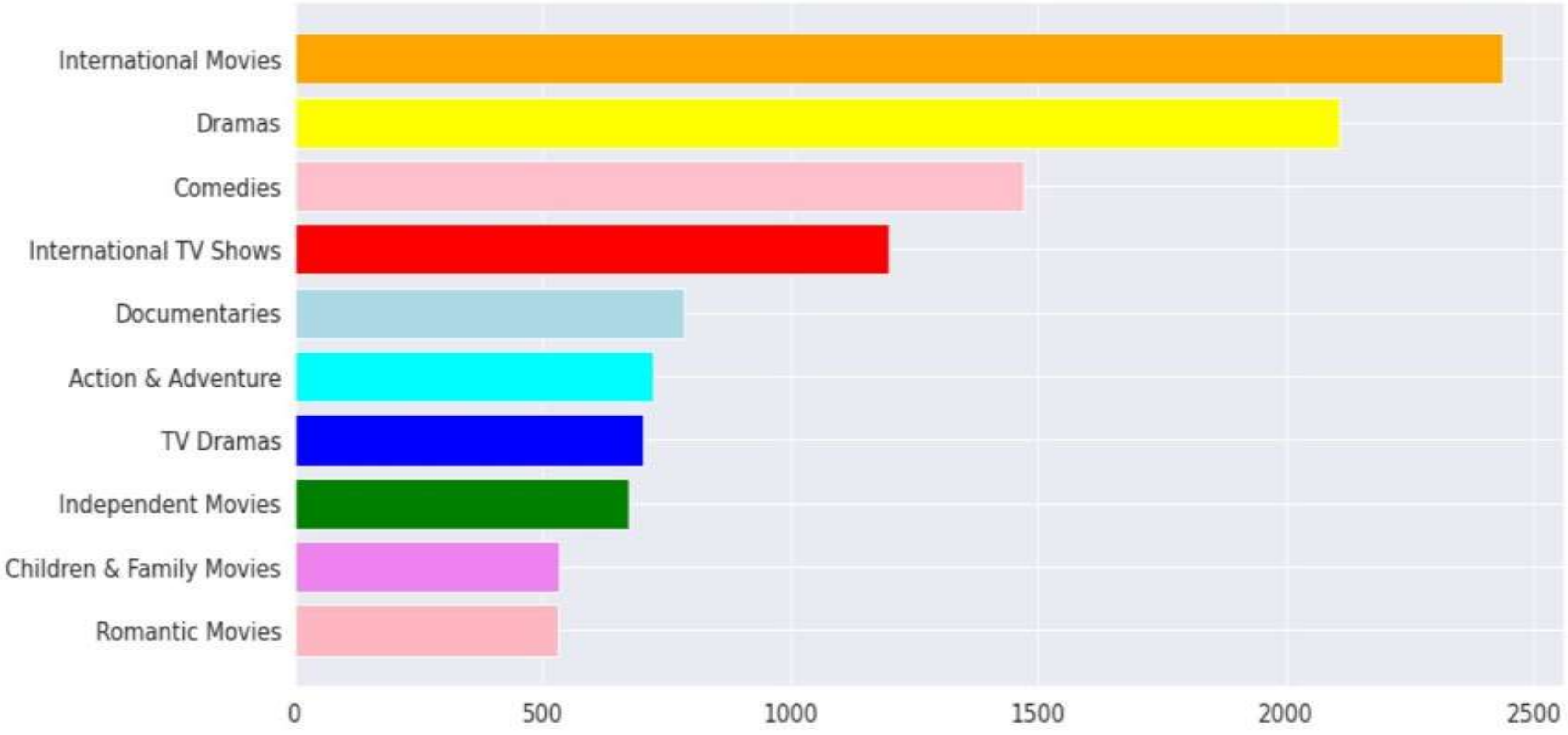
- *United states have the most number of content and then india and so on*
- We can conclude that except ***Japan*** other countries are producing movies more than TV-Shows

How many no of categories are present there in each content ?

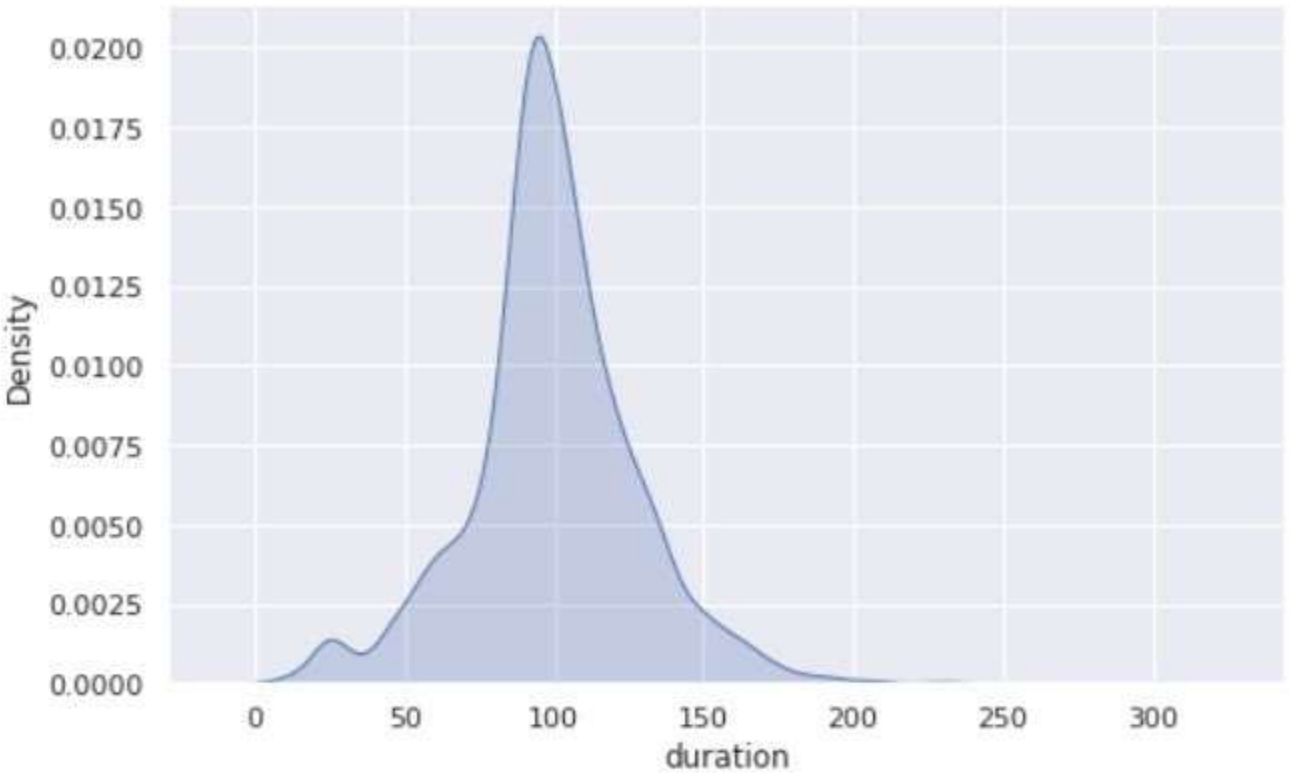


Most of the movies are belonging to 3 categories

Top 10 Category For Contents

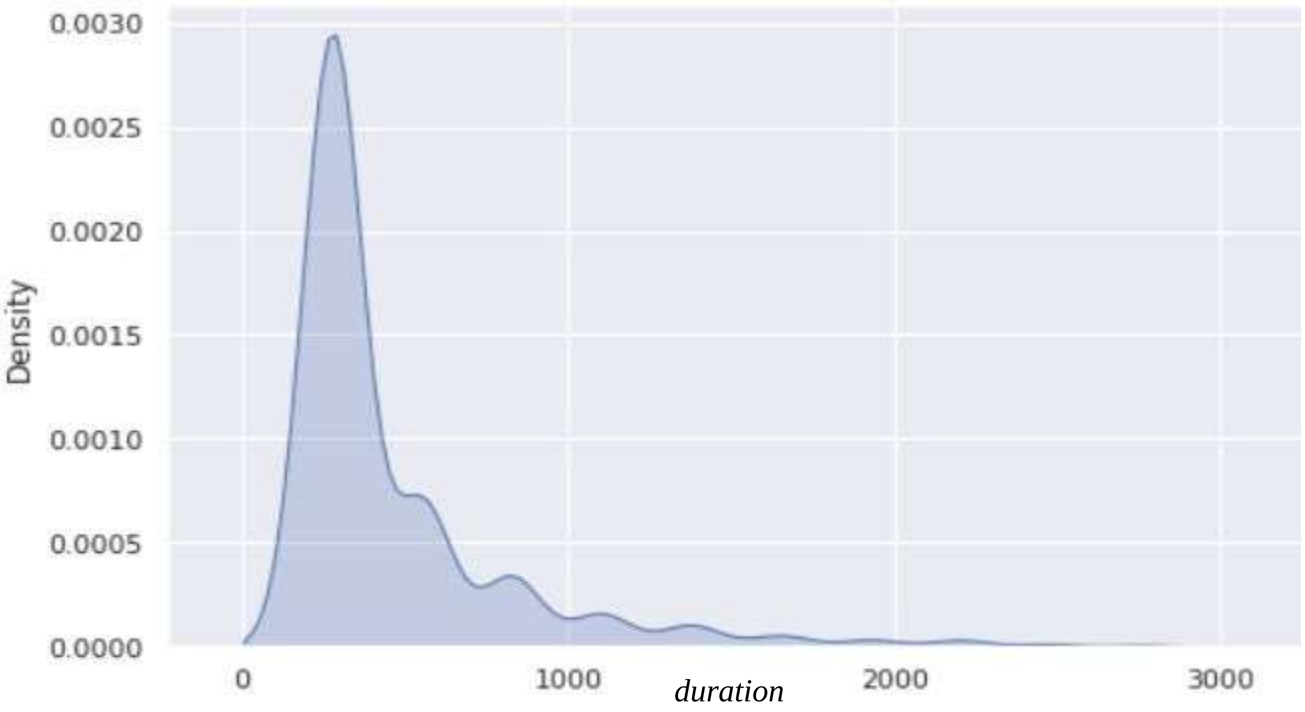


Movie Wise Density Plot



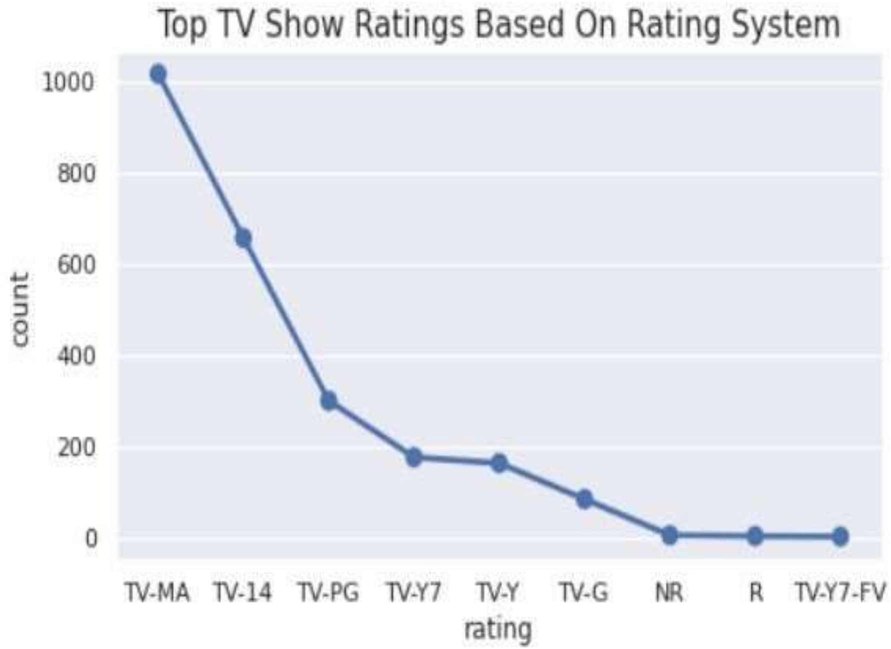
Most movies are about 70 to 120 min duration for movies

TV-Shows wise density plot



- *Most contents are about 0 to 750 min duration for movies*
- *There are very few shows which is having more than 1000 mins. (may be the no of episodes/ seasons are more)*

TOP Content Based On Rating



Most of the contents got ratings like

- **TV-MA** *(For Mature Audiences)*
- **TV-14** (May be unsuitable for children under 14)
- **TV-PG (** Parental Guidance Suggested **)**
- **NR** (Not Rated)

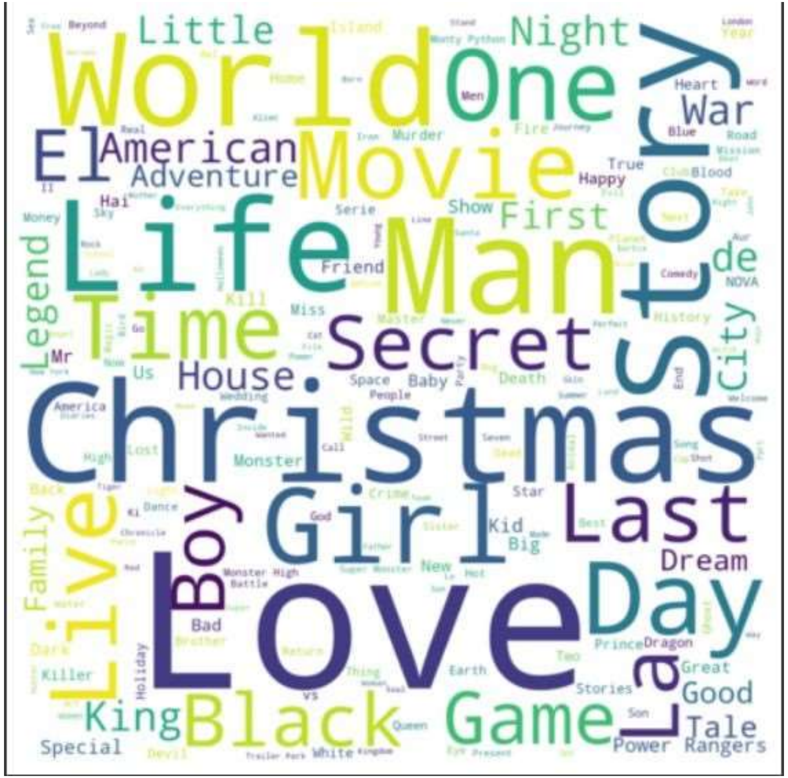
Woí dCloud

What Is a Word Cloud?

A word cloud (also known as a tag cloud) is a visual representation of words. Cloud creators are used to highlight popular words and phrases based on frequency and relevance. They provide you with quick and simple visual insights that can lead to more in-depth analyses.

Example ⟶

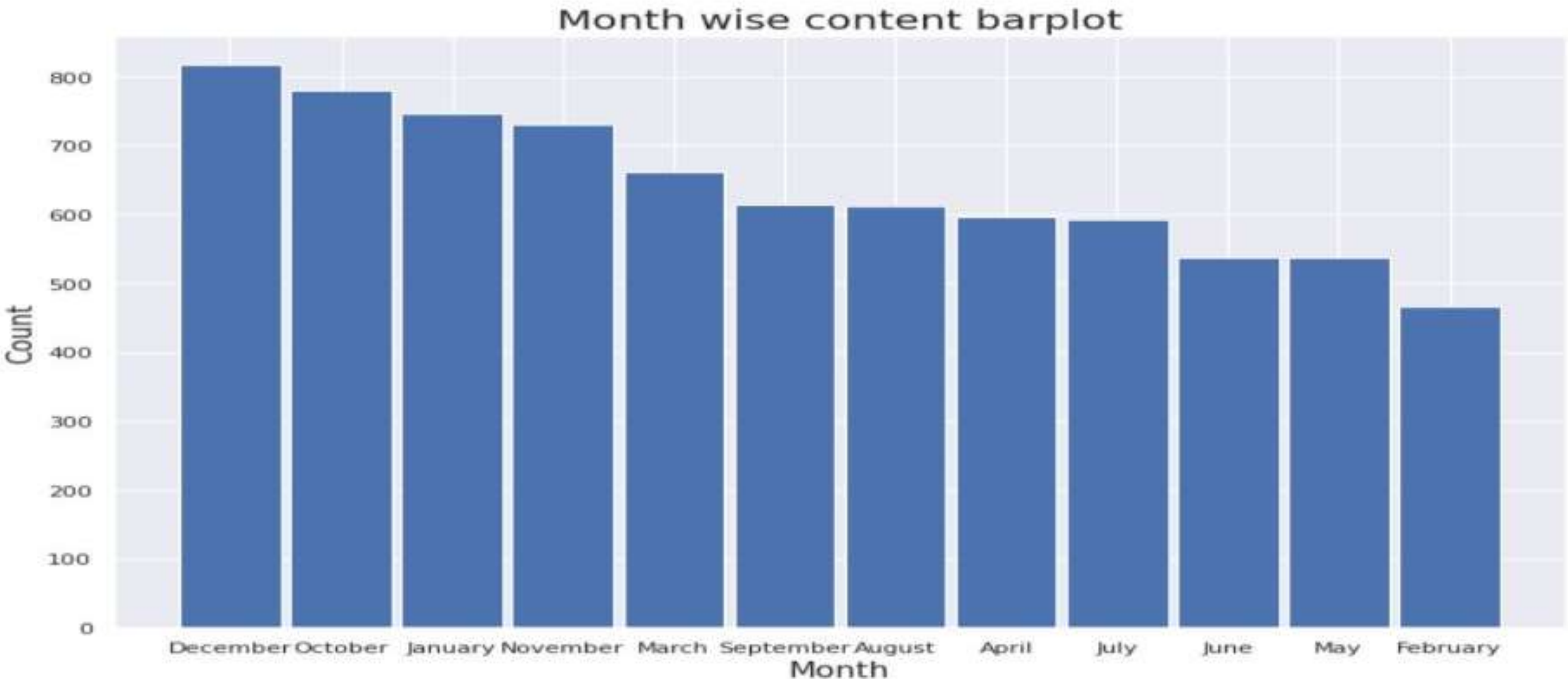
Applying *WordCloud* on Title



*Most occurred words present in **Title** are:-*

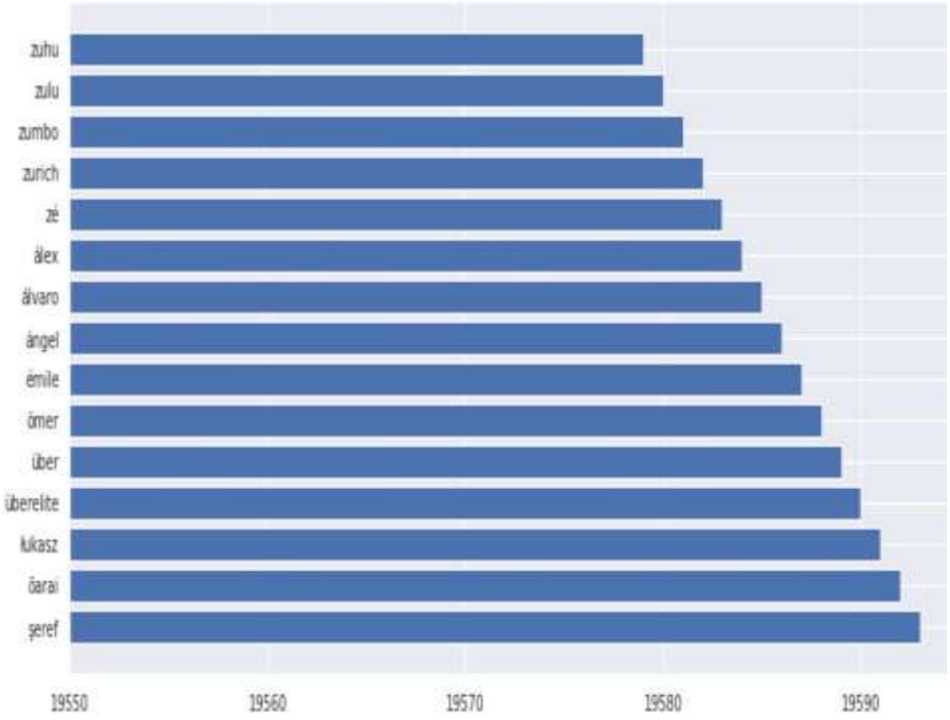
- ***Love***
- ***Man***
- ***World***
- ***Story***
- ***Christmas***
- ***Girl***
- ***Day***

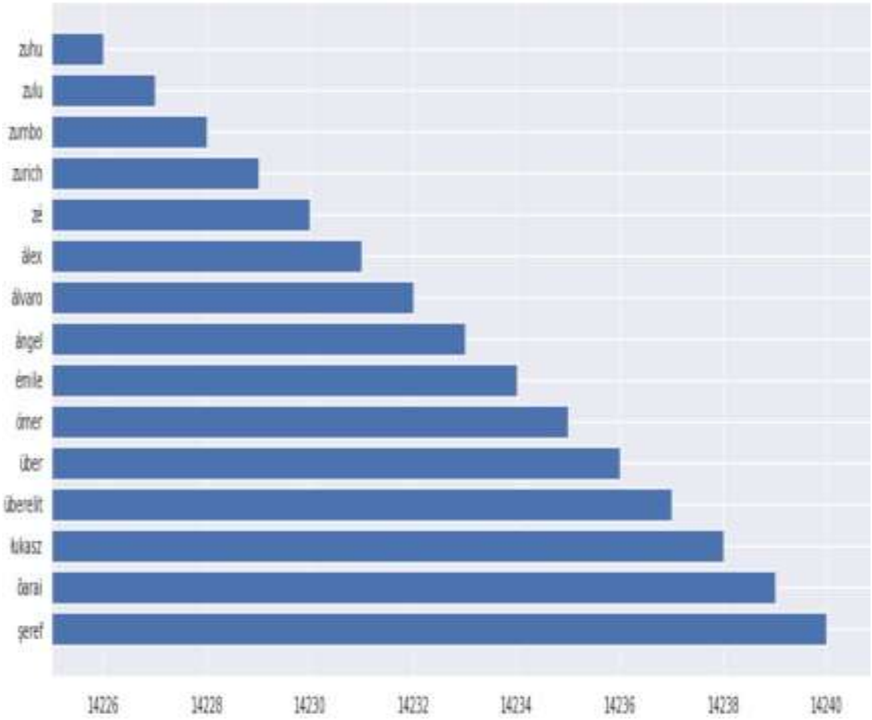
Barplot based on release month



We can say that December is the holiday season and it also has Christmas, so in that month most of the content got uploaded.

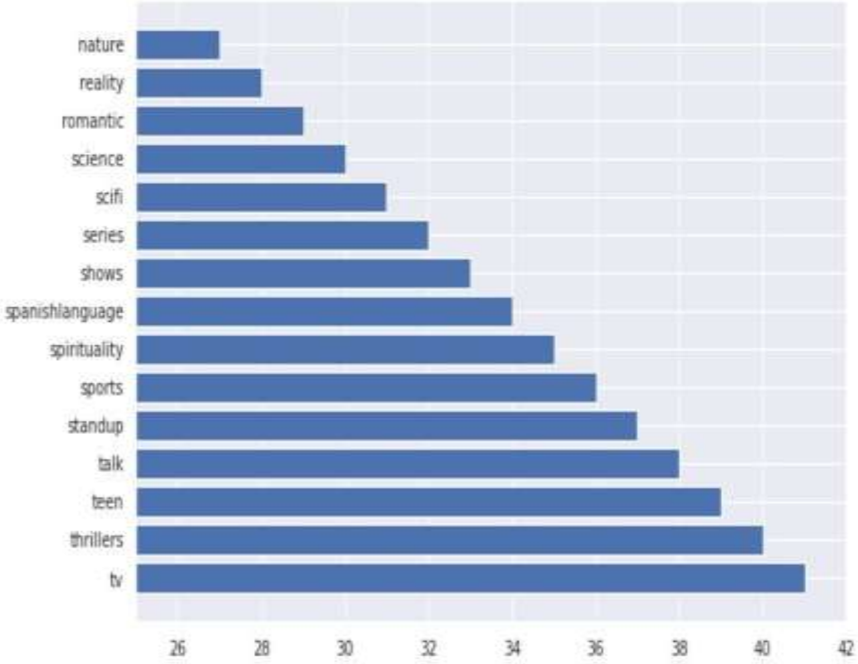
Before & After Stemming most occurred words in description

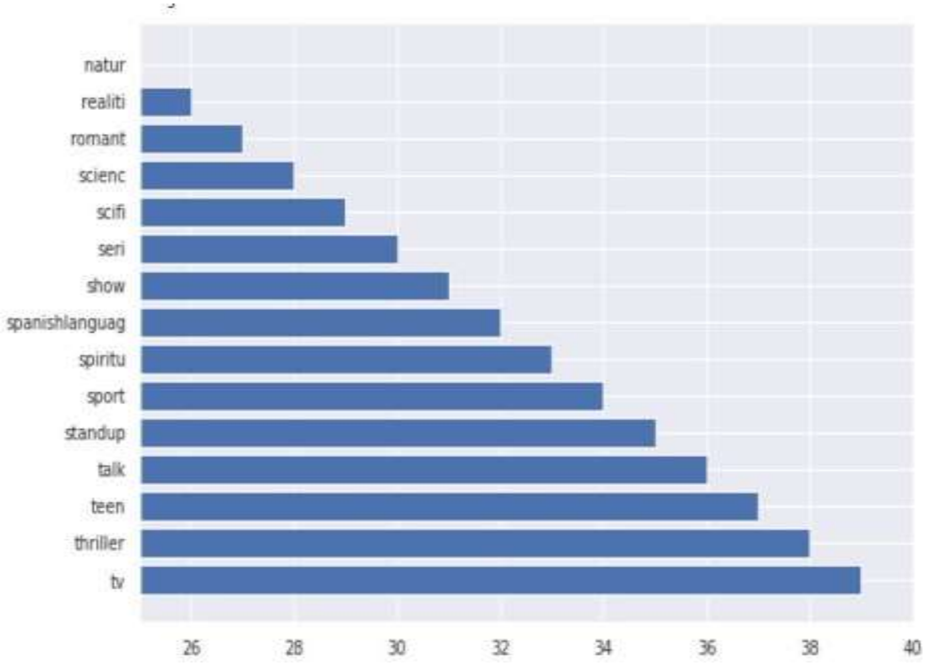




Before & After Stemming most occurred words

in listed_in





Feature Selection & ML algo used

- Only selected 3 features , to do clustering
  - *no_of_category*
  - *Length(description)*
  - *Length(listed-in)*

- *Using StandardScaler*

- *Used 5 algo to find out best k value*

  - *1. Silhouette score*
  - *2. Elbow Method*
  - *3. DBSCAN*
  - *4. Dendrogram*
  - *5. AgglomerativeClustering*

1. Silhouette Score

	n clusters	silhouette score
1	3	0.348
0	2	0.337
12	14	0.332
5	7	0.330
11	13	0.329
10	12	0.328
13	15	0.326
9	11	0.324
8	10	0.323
7	9	0.322
2	4	0.320
4	6	0.320
6	8	0.316
3	5	0.308

Silhouette Coefficient Formula

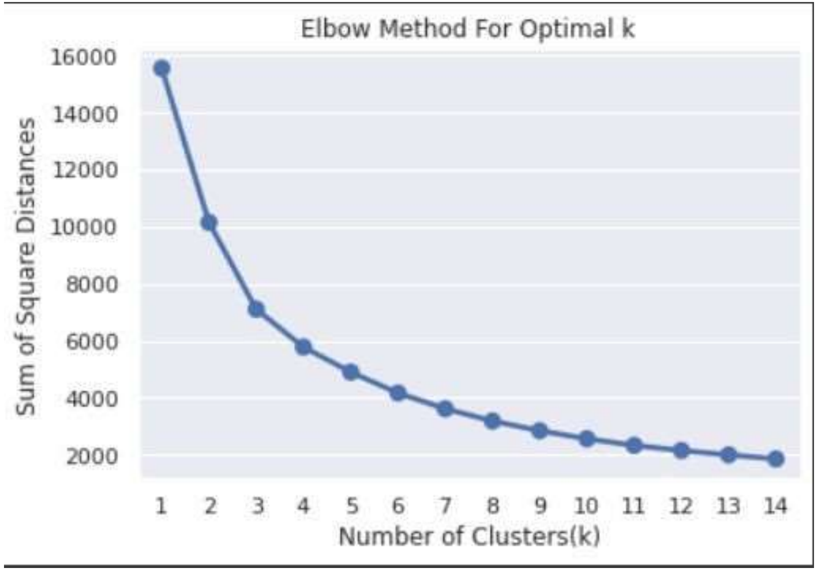
$$S = \frac{(b-a)}{max(a,b)}.$$

- **mean intra-cluster distance(a)**:- Mean distance between the observation and all other data points in the same cluster.
- **mean nearest-cluster distance (b)** :- Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be called a.

The value of the silhouette coefficient is between [-1, 1]

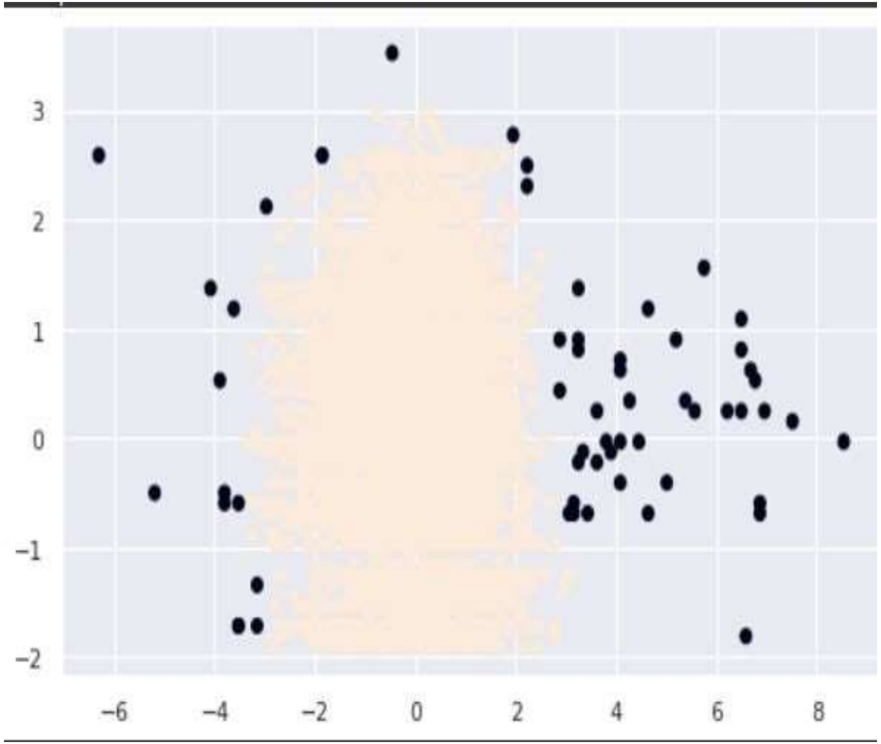
- If score is **1 denotes the best** meaning that the data point i is very compact within the cluster to which it belongs and far away from the other clusters.
- *The worst value is -1*
- If score is 0 denotes overlapping clusters

2. Elbow Method

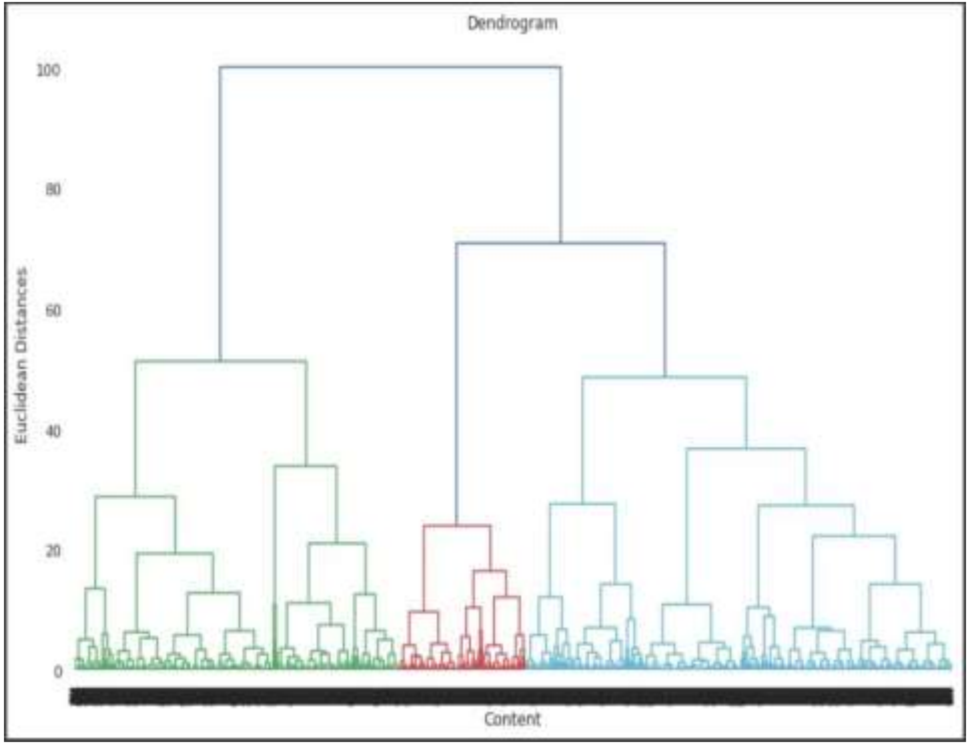


The elbow method runs k-**means clustering on the dataset for a range of values for k** (say from 1-15) and then for each value of k computes WCSS value . By default, the distortion score is computed, the sum of square distances from each point to its assigned center.

3 & 4 DBSCAN & Dendrogram

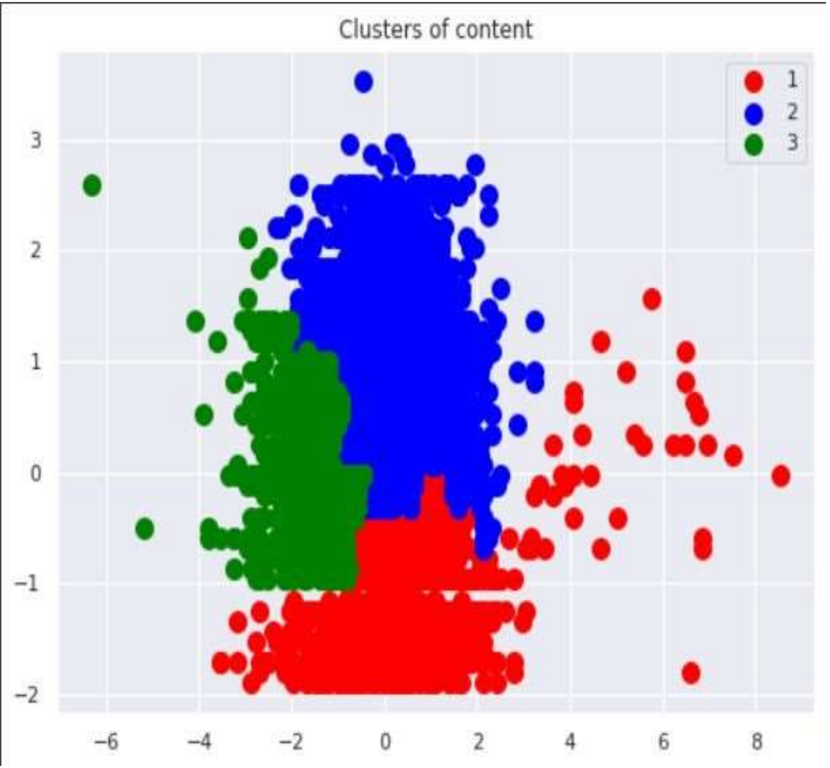


DBSCAN



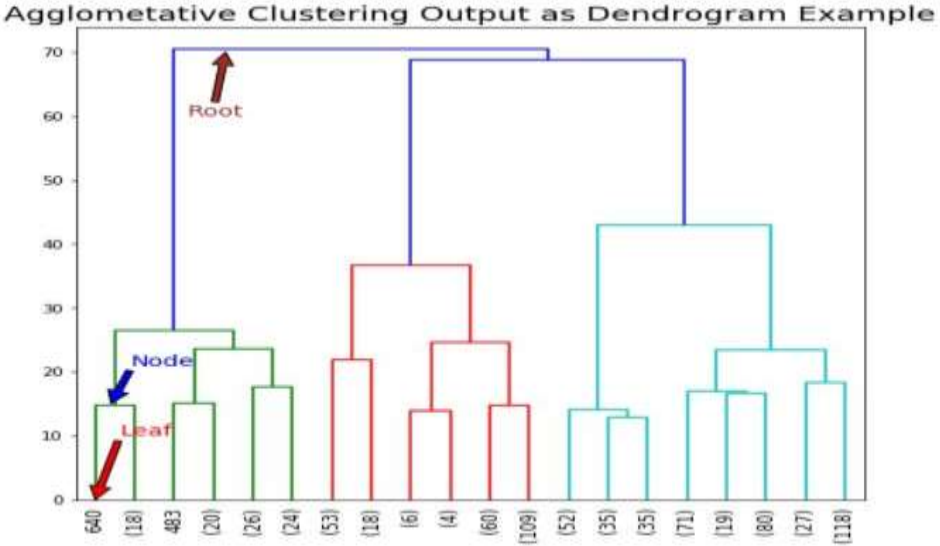
Dendrogram

5 AgglomerativeClustering



Steps: -

1. Each data point is assigned as a single cluster.
2. Determine the distance measurement and calculate the distance matrix.
3. Determine the linkage criteria to merge the clusters.
4. Update the distance matrix.
5. Repeat the process until every data point become one cluster.



Conclusion

1. Director and cast contains a large number of null values so we will drop these 2 columns .
2. In this dataset there are two types of contents where 30.86% includes TV shows and the remaining 69.14% carries Movies.
3. We have reached a conclusion from our analysis from the content added over years that Netflix is focusing movies and TV shows (Fom 2016 data we get to know that Movies is increased by 80% and TV shows is increased by 73% compare)
4. From the dataset insights we can conclude that the most number of TV Shows released in 2017 and for Movies it is 2020
5. On Netflix USA has the largest number of contents. And most of the countries preferred to produce movies more than TV shows.
6. Most of the movies are belonging to 3 categories
7. TOP 3 content categories are International movies , dramas , comedies.
8. In text analysis (NLP) I used stop words, removed punctuations , stemming & TF-IDF vectorizer and other functions of NLP.
9. Applied different clustering models like Kmeans, hierarchical, Agglomerative clustering, DBSCAN on data we got the best cluster arrangements.
10. By applying different clustering algorithms to our dataset .we get the optimal number of cluster is equal to 3

THANK YOU