CAPSTONE PROJECT - 4

NETFLIX MOVIES & TV SHOWS CLUSTERING

(Unsupervised Machine Learning)

By

VIKASKUMAR SHARMA

(Cohort Tosh)

❑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.
- In this project, you are required to do :-
  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 DESCRIPTION

- Attribute Information:-

show_id : Unique ID for every Movie / Tv Show

type : Identifier - 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 Releaseyear of the movie / show

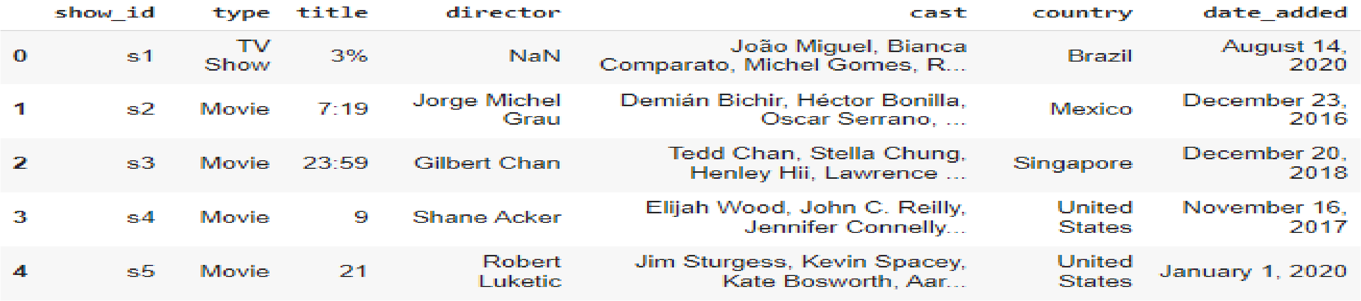
rating : TV Rating of the movie / show

duration : Total Duration - in minutes or number of seasons

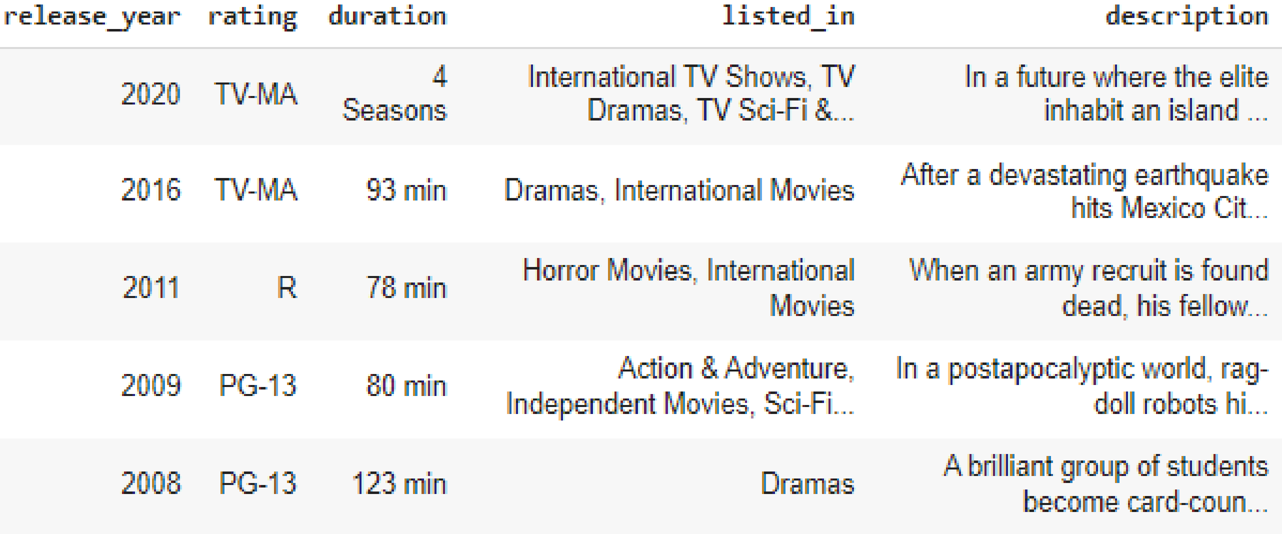
listed_in : Genere

description: The Summary description

DATA COLLECTION & UNDERSTANDING

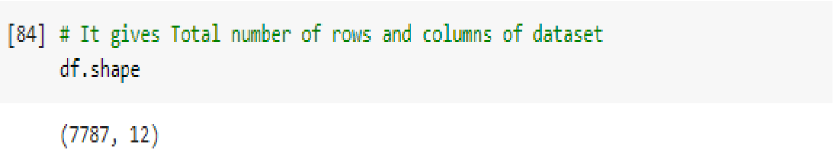


	show_id	type	title	director	cast	country	date_added
0	s1	TV Show	3%	NaN	João Miguel, Bianca Comparato, Michel Gomes, R...	Brazil	August 14, 2020
1	s2	Movie	7:19	Jorge Michel Grau	Demián Bichir, Héctor Bonilla, Oscar Serrano, ...	Mexico	December 23, 2016
2	s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella Chung, Henley Hii, Lawrence ...	Singapore	December 20, 2018
3	s4	Movie	9	Shane Acker	Elijah Wood, John C. Reilly, Jennifer Connelly...	United States	November 16, 2017
4	s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin Spacey, Kate Bosworth, Aar...	United States	January 1, 2020



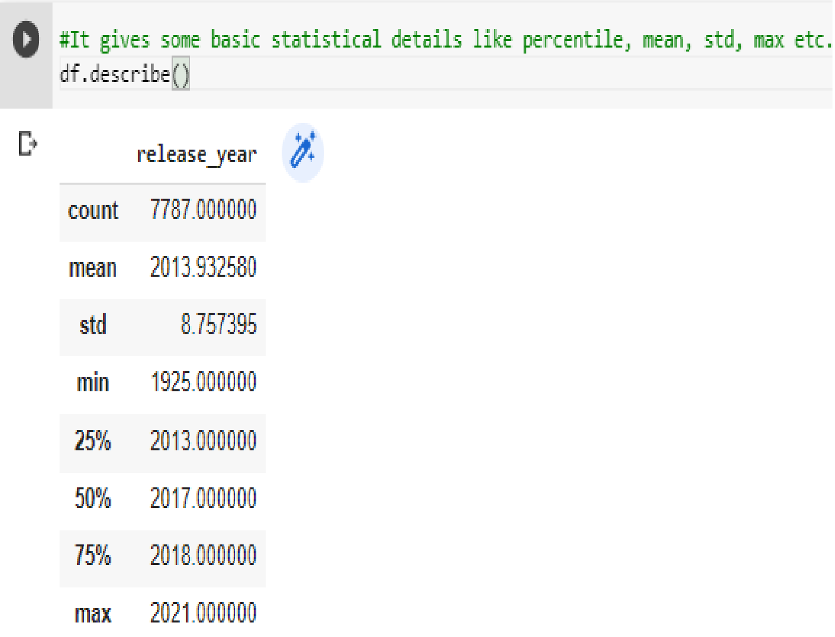
release_year	rating	duration	listed_in	description
2020	TV-MA	4 Seasons	International TV Shows, TV Dramas, TV Sci-Fi &...	In a future where the elite inhabit an island ...
2016	TV-MA	93 min	Dramas, International Movies	After a devastating earthquake hits Mexico Cit...
2011	R	78 min	Horror Movies, International Movies	When an army recruit is found dead, his fellow...
2009	PG-13	80 min	Action & Adventure, Independent Movies, Sci-Fi...	In a postapocalyptic world, rag-doll robots hi...
2008	PG-13	123 min	Dramas	A brilliant group of students become card-coun...

DATA COLLECTION & UNDERSTANDING



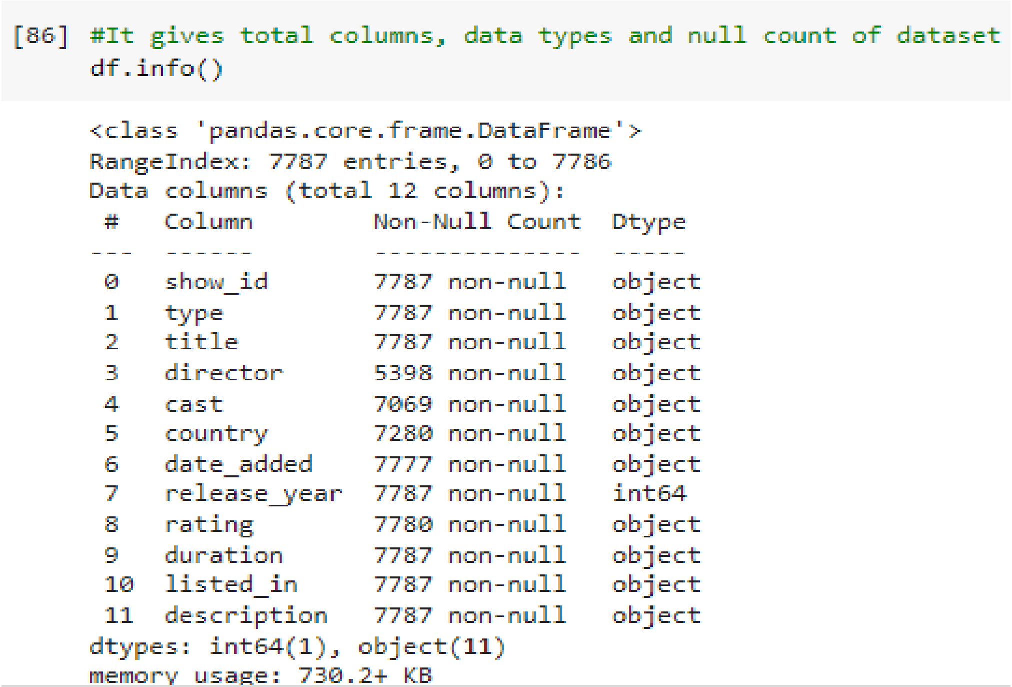
```
[84] # It gives Total number of rows and columns of dataset
     df.shape

     (7787, 12)
```

Dataset contain 7787 rows and 12 columns



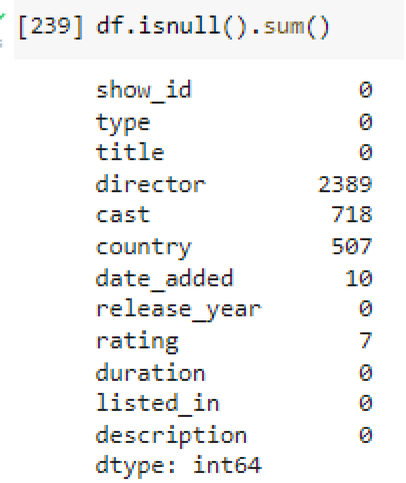
```
#It gives some basic statistical details like percentile, mean, std, max etc.
df.describe()
```

	release_year
count	7787.000000
mean	2013.932580
std	8.757395
min	1925.000000
25%	2013.000000
50%	2017.000000
75%	2018.000000
max	2021.000000

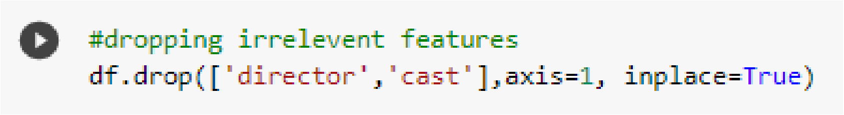


```
[86] #It gives total columns, data types and null count of dataset
     df.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 7787 entries, 0 to 7786
     Data columns (total 12 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   director      5398 non-null   object
      4   cast          7069 non-null   object
      5   country       7280 non-null   object
      6   date_added    7777 non-null   object
      7   release_year  7787 non-null   int64 
      8   rating        7780 non-null   object
      9   duration      7787 non-null   object
      10  listed_in     7787 non-null   object
      11  description   7787 non-null   object
     dtypes: int64(1), object(11)
     memory usage: 730.2+ KB
```

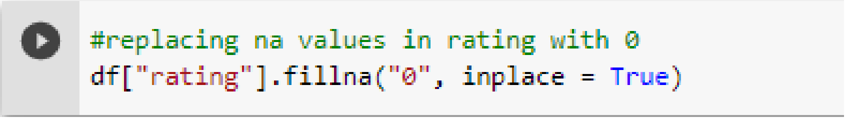
DATA CLEANING & FEATURE ENGG.



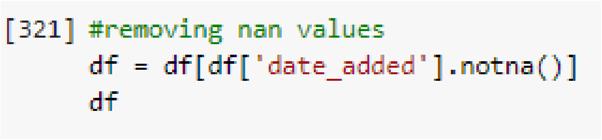
```
[239] df.isnull().sum()

      show_id            0
      type               0
      title              0
      director        2389
      cast             718
      country          507
      date_added        10
      release_year       0
      rating             7
      duration           0
      listed_in          0
      description        0
      dtype: int64
```



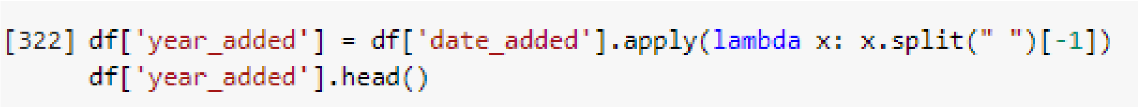
```
#dropping irrelevent features
df.drop(['director','cast'],axis=1, inplace=True)
```



```
#replacing na values in rating with 0
df["rating"].fillna("0", inplace = True)
```

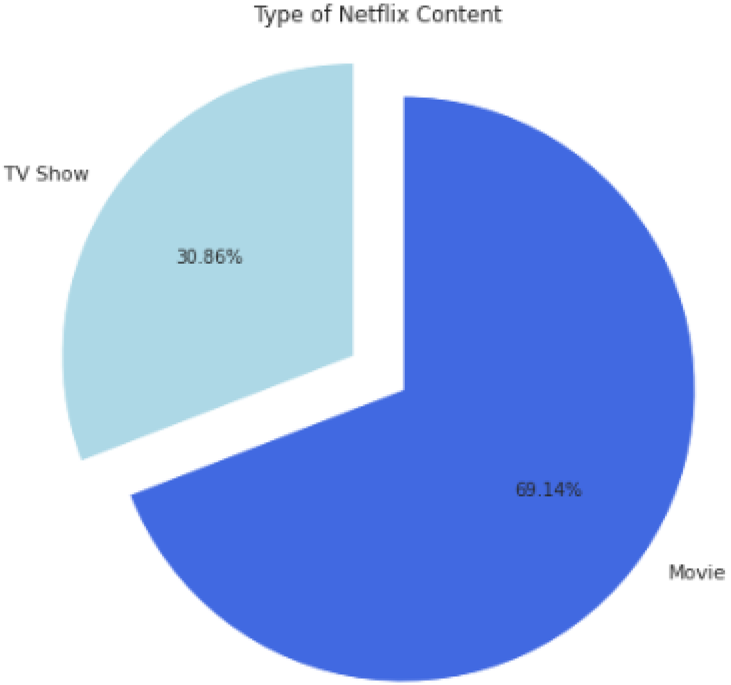


```
[321] #removing nan values
      df = df[df['date_added'].notna()]
      df
```



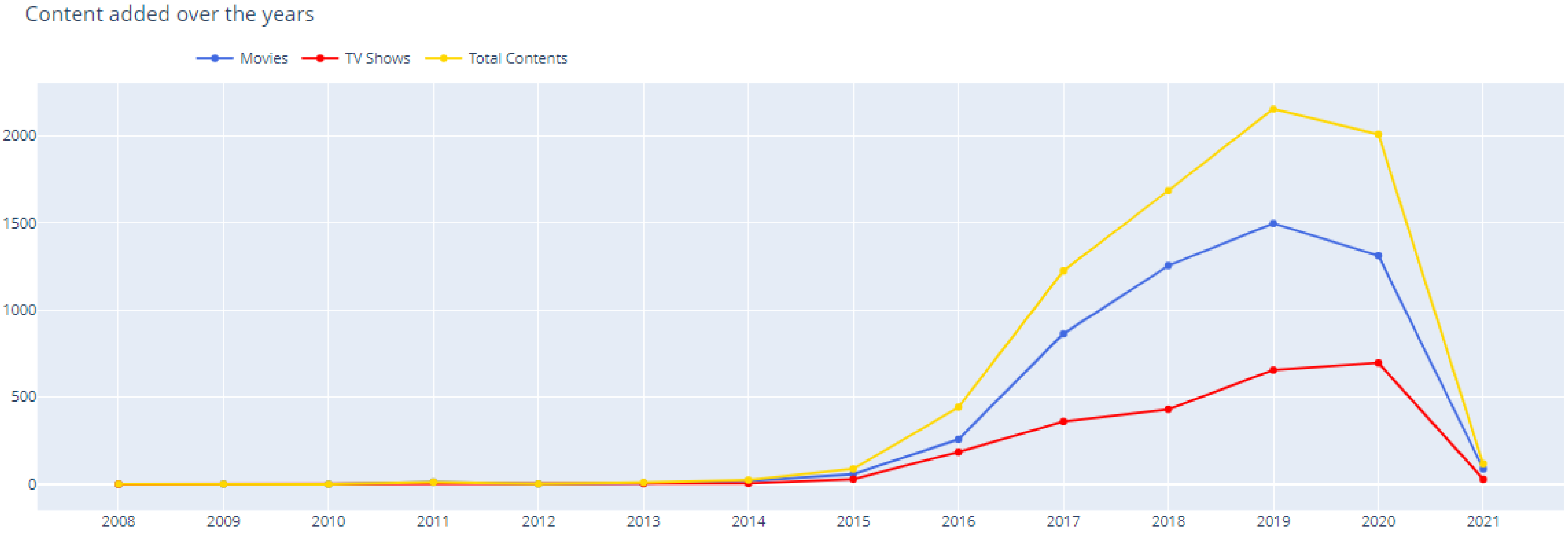
```
[322] df['year_added'] = df['date_added'].apply(lambda x: x.split(" ")[-1])
      df['year_added'].head()
```

EDA



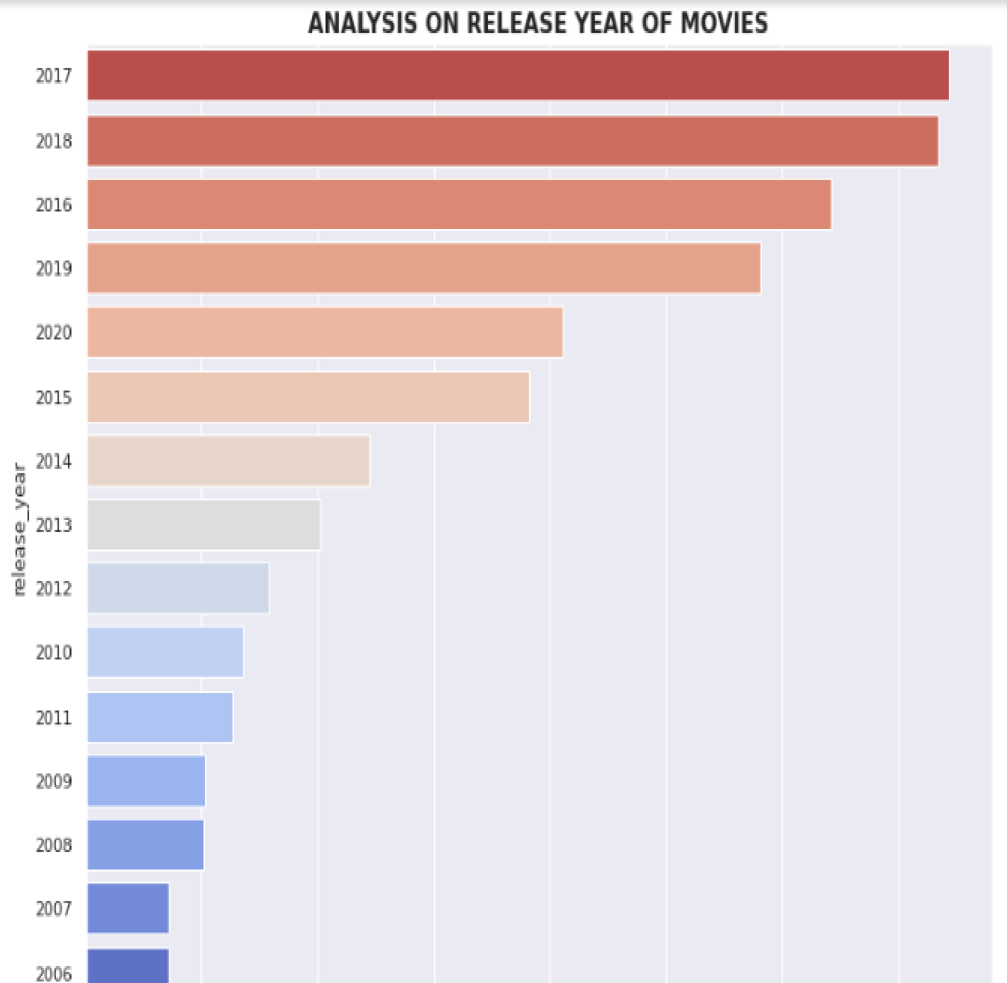
As we can see from the Pie – Chart , About 70 % of the total Data is Movies and the rest of the 30% is about TV Show.

EDA

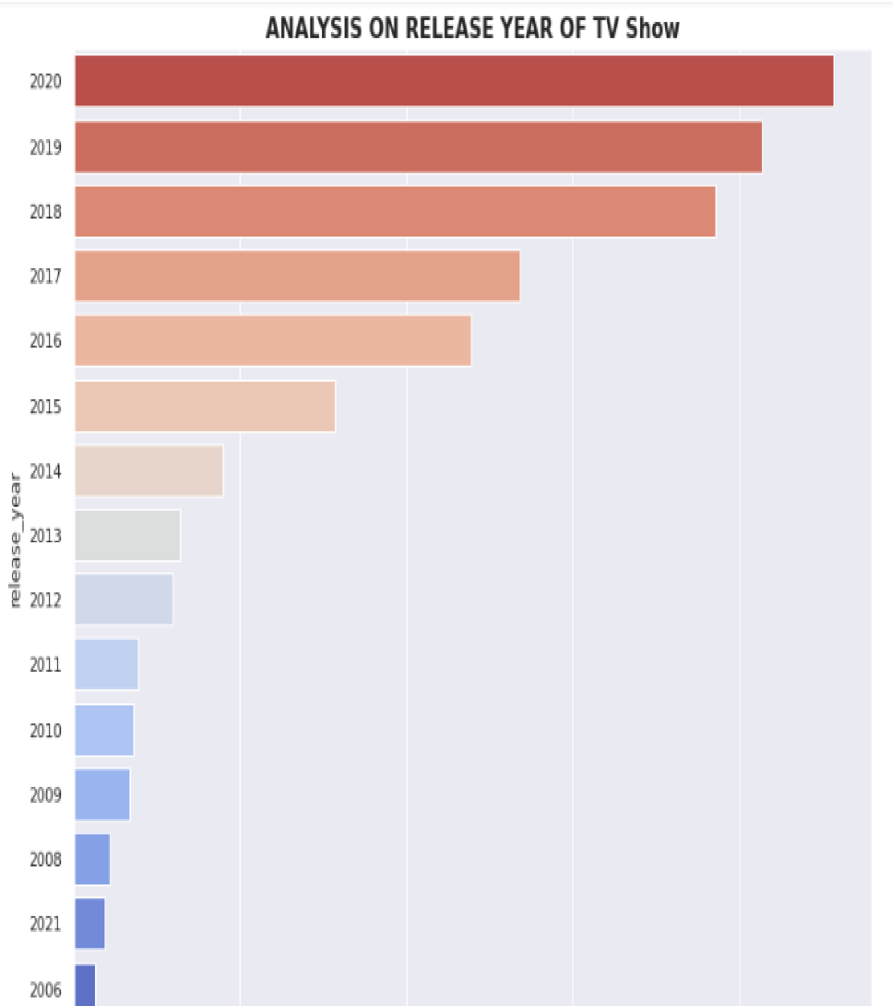


As we can see from the above line chart, the content of the Movies is more compared to the TV Shows.

EDA



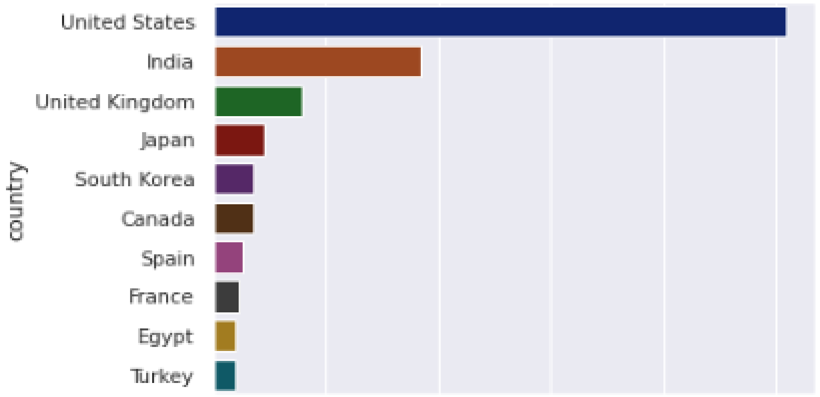
ANALYSIS ON RELEASE YEAR OF MOVIES

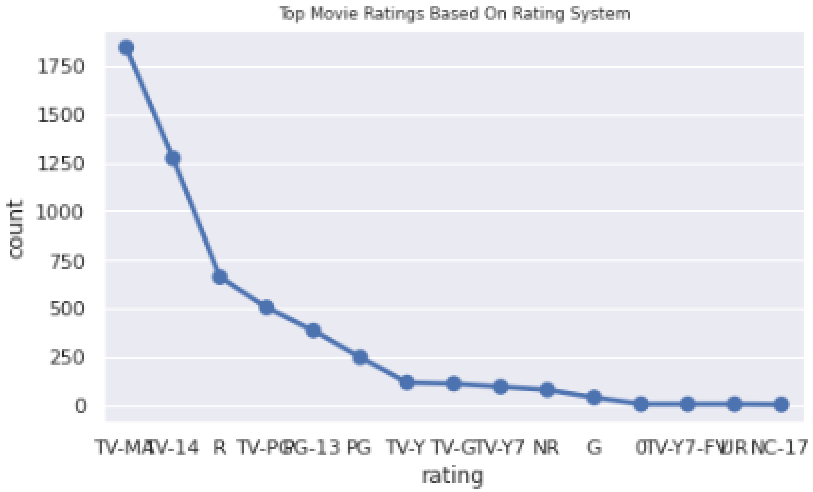


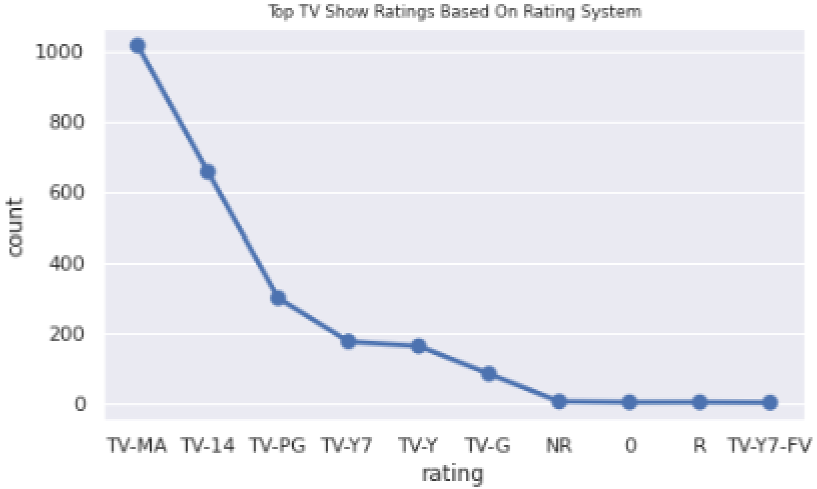
ANALYSIS ON RELEASE YEAR OF TV Show

EDA

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f044e92c7d0>
```

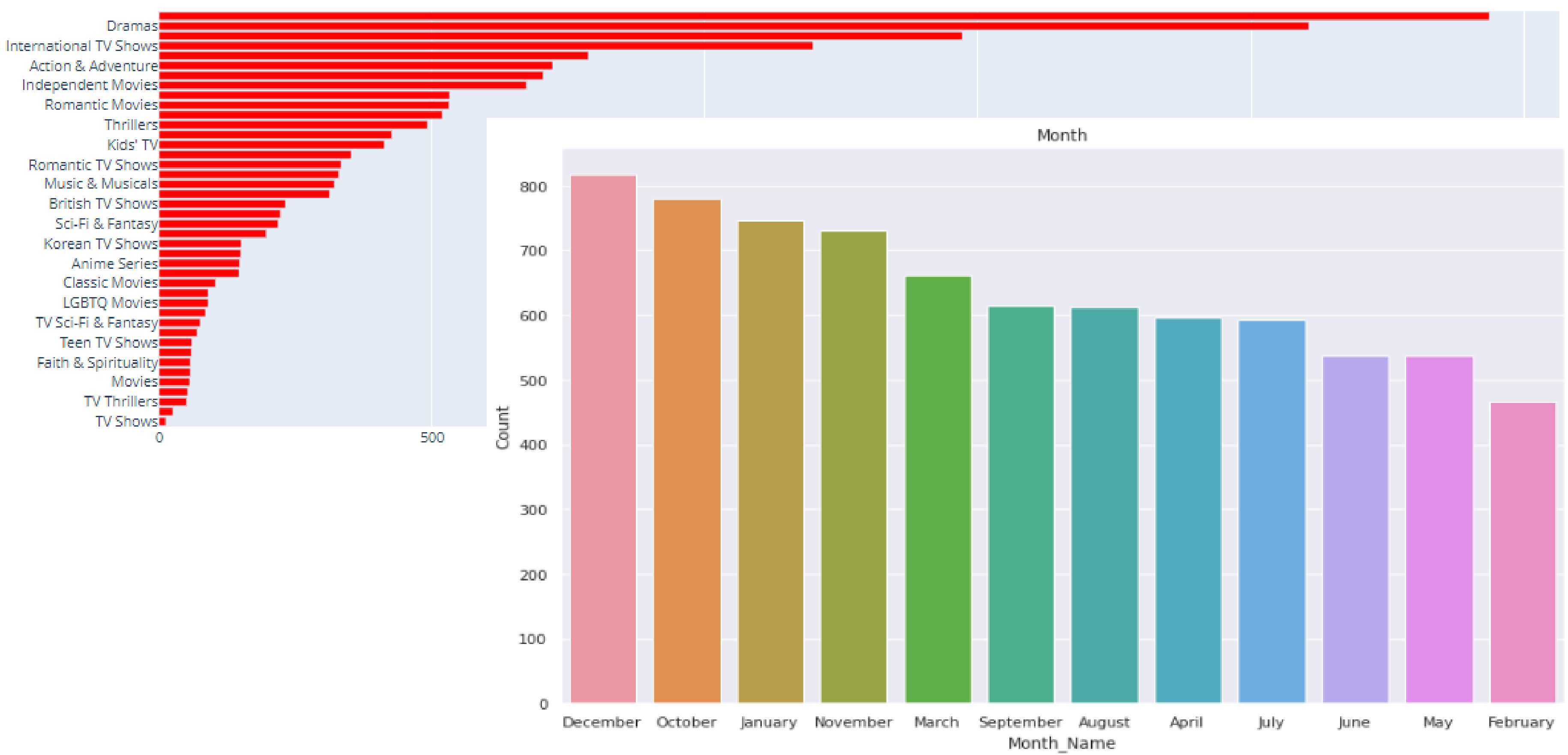



Top Movie Ratings Based On Rating System


Top TV Show Ratings Based On Rating System

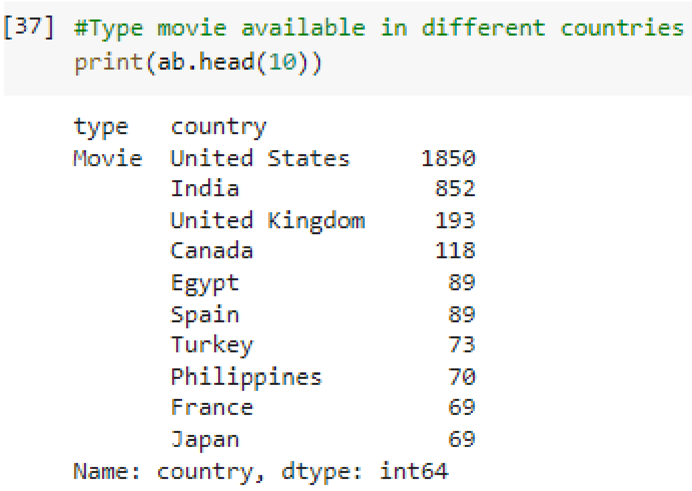
- The US,India & Uk are the top 3 countries by Content wise.
- The Top 3 TV Shows based on rating systems are TV-MA,TV-14 & TV-PG.
- The Top 3 Movies based on rating systems are TV-MA,TV-14 & R.

EDA



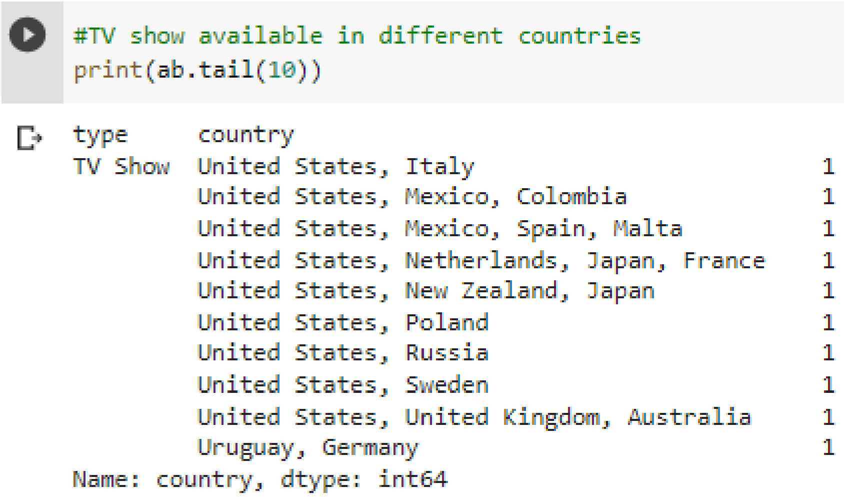
count of each Content

EDA



```
[37] #Type movie available in different countries
     print(ab.head(10))
```

```
type   country
Movie  United States     1850
       India              852
       United Kingdom     193
       Canada             118
       Egypt               89
       Spain               89
       Turkey              73
       Philippines         70
       France              69
       Japan               69
Name: country, dtype: int64
```

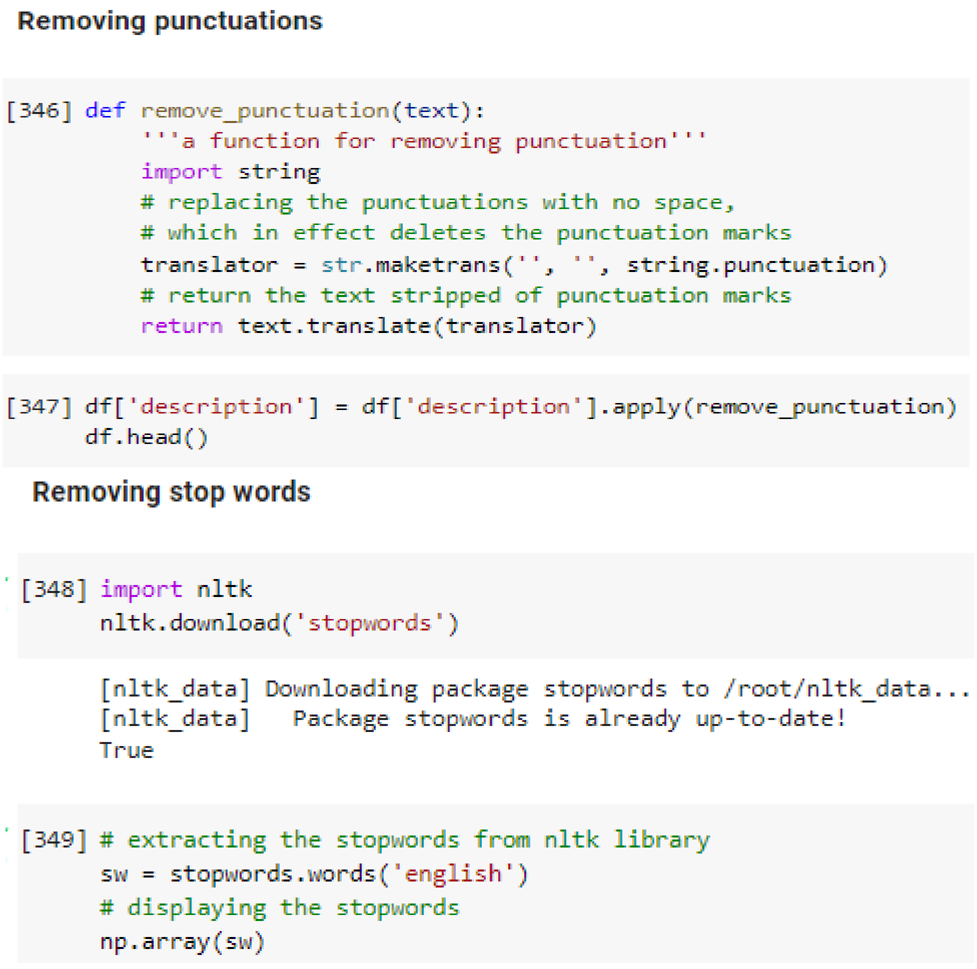


```
#TV show available in different countries
print(ab.tail(10))
```

```
type     country
TV Show  United States, Italy                               1
         United States, Mexico, Colombia                    1
         United States, Mexico, Spain, Malta                1
         United States, Netherlands, Japan, France          1
         United States, New Zealand, Japan                  1
         United States, Poland                              1
         United States, Russia                              1
         United States, Sweden                              1
         United States, United Kingdom, Australia           1
         Uruguay, Germany                                   1
Name: country, dtype: int64
```

- **The Top 10 Countries were Movies are produced are:-US, India, UK, Canada, Egypt, Spain, Turkey, Philippines, France & Japan.**
- **The Top 10 Countries were TV Shows are produced are:-US, Italy, Mexico, Colombia, Spain, Malta, Netherlands, France & Japan.**

DATA PREPROCESSING



Removing punctuations

```
[346] def remove_punctuation(text):
          '''a function for removing punctuation'''
          import string
          # replacing the punctuations with no space,
          # which in effect deletes the punctuation marks
          translator = str.maketrans('', '', string.punctuation)
          # return the text stripped of punctuation marks
          return text.translate(translator)
```

```
[347] df['description'] = df['description'].apply(remove_punctuation)
      df.head()
```

Removing stop words

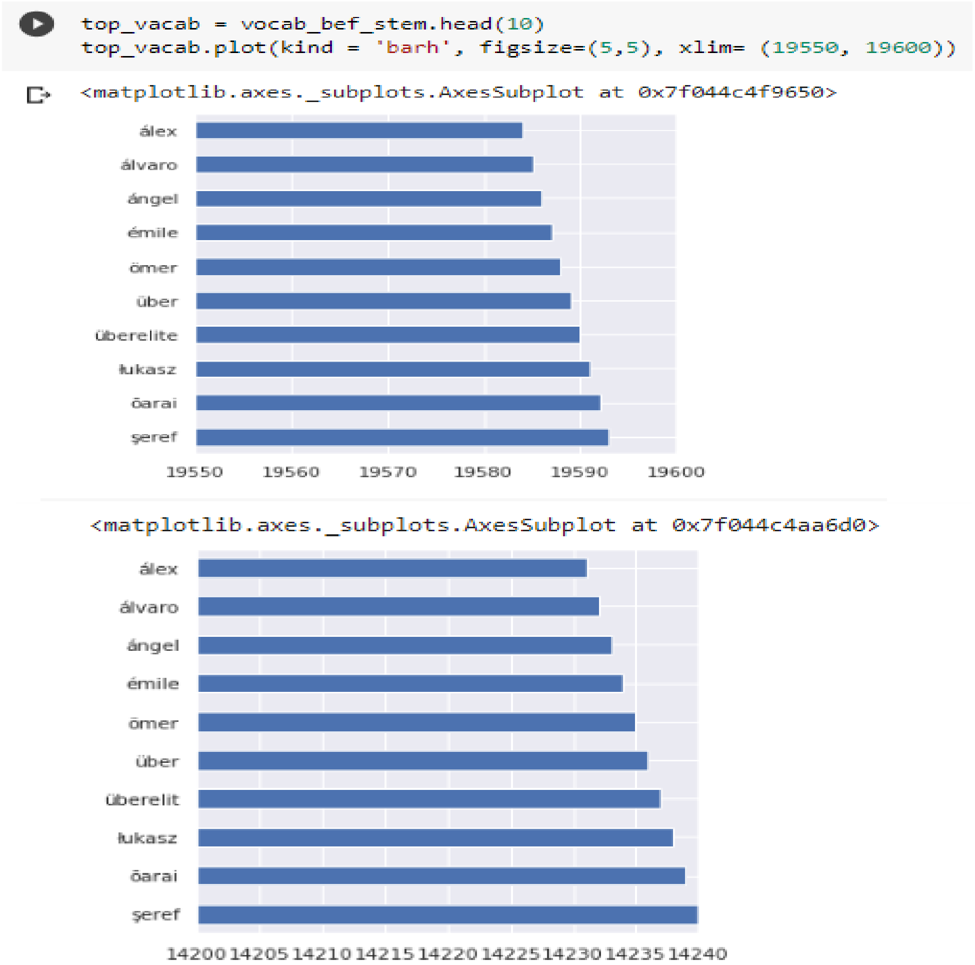
```
[348] import nltk
      nltk.download('stopwords')

      [nltk_data] Downloading package stopwords to /root/nltk_data...
      [nltk_data]   Package stopwords is already up-to-date!
      True
```

```
[349] # extracting the stopwords from nltk library
      sw = stopwords.words('english')
      # displaying the stopwords
      np.array(sw)
```

In Data Pre processing, the very first step is to remove punctuations.

The next step is to remove stopwords.

DATA PREPROCESSING



```
top_vacab = vocab_bef_stem.head(10)
top_vacab.plot(kind = 'barh', figsize=(5,5), xlim= (19550, 19600))
```

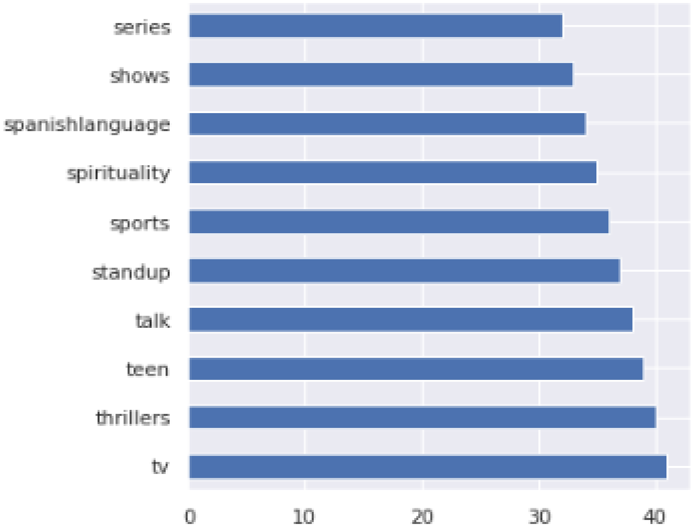
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f044c4f9650>
```

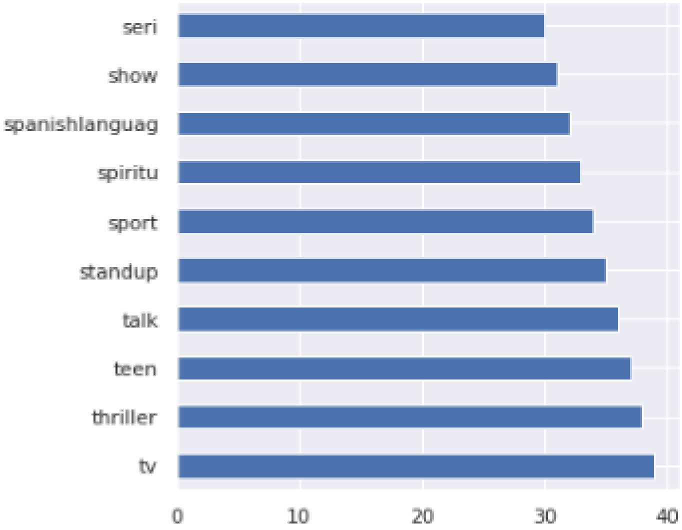
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f044c4aa6d0>
```

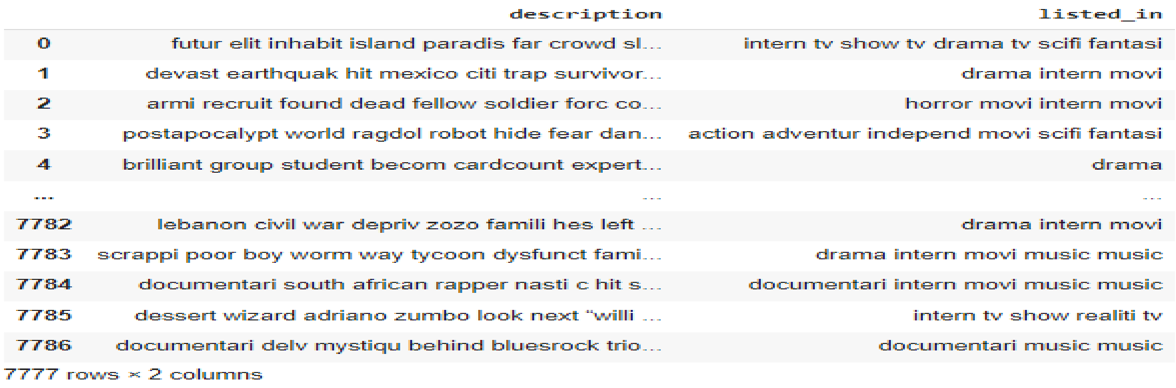
Theses are the Top 10 words before Stemming.

Theses are the Top 10 words after Stemming.

DATA PREPROCESSING



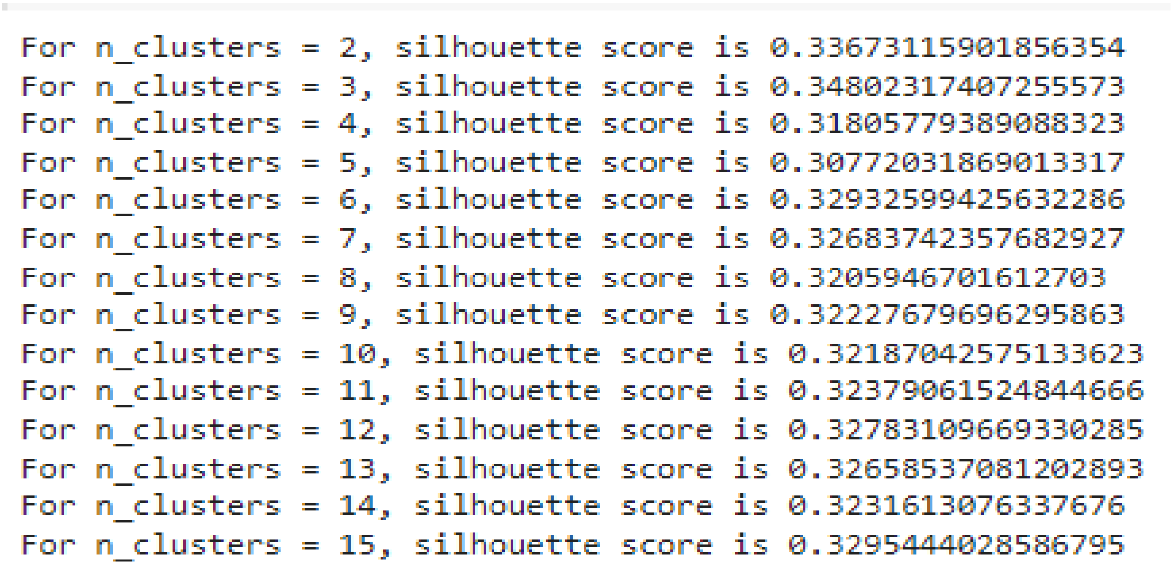




	description	listed_in
0	futur elit inhabit island paradis far crowd sl...	intern tv show tv drama tv scifi fantasi
1	devast earthquak hit mexico citi trap survivor...	drama intern movi
2	armi recruit found dead fellow soldier forc co...	horror movi intern movi
3	postapocalypt world ragdol robot hide fear dan...	action adventur independ movi scifi fantasi
4	brilliant group student becom cardcount expert...	drama
...	...	...
7782	lebanon civil war depriv zozo famili hes left ...	drama intern movi
7783	scrappi poor boy worm way tycoon dysfunct fami...	drama intern movi music music
7784	documentari south african rapper nasti c hit s...	documentari intern movi music music
7785	dessert wizard adriano zumbo look next “willi ...	intern tv show realiti tv
7786	documentari delv mystiqu behind bluesrock trio...	documentari music music

7777 rows × 2 columns

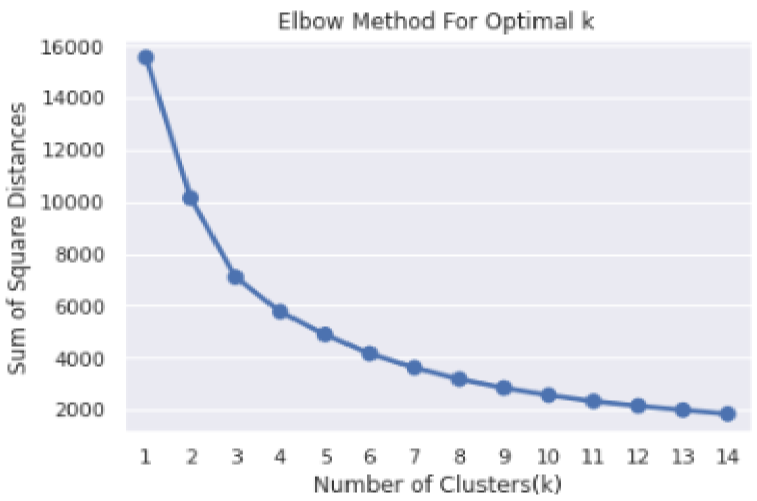
SILHOUETTE SCORE

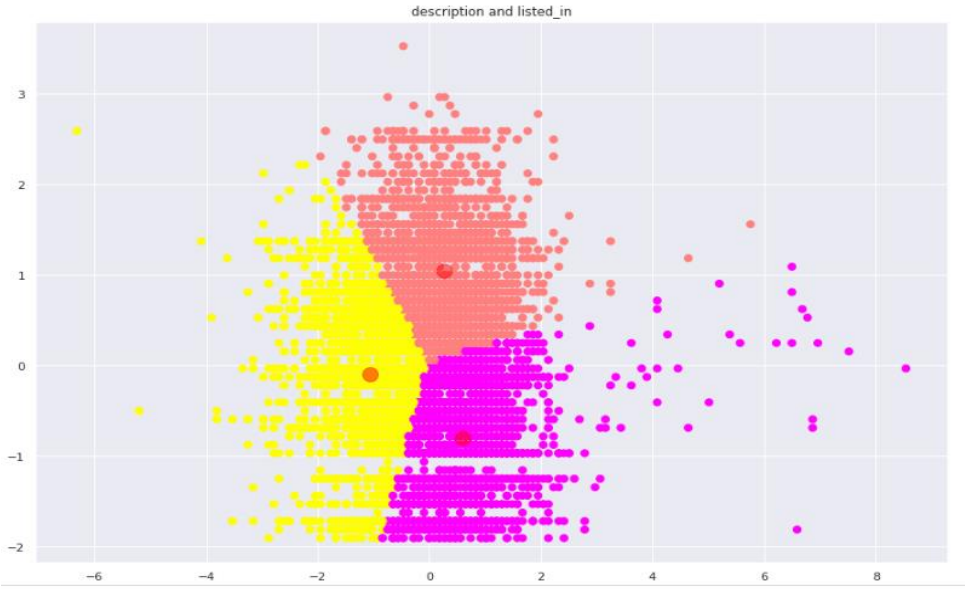


```
For n_clusters = 2, silhouette score is 0.33673115901856354
For n_clusters = 3, silhouette score is 0.3480231740725573
For n_clusters = 4, silhouette score is 0.3180577938908323
For n_clusters = 5, silhouette score is 0.3077203186901317
For n_clusters = 6, silhouette score is 0.3293259942563286
For n_clusters = 7, silhouette score is 0.3268374235768292
For n_clusters = 8, silhouette score is 0.320594670161270
For n_clusters = 9, silhouette score is 0.3222767969629586
For n_clusters = 10, silhouette score is 0.3218704257513362
For n_clusters = 11, silhouette score is 0.3237906152484466
For n_clusters = 12, silhouette score is 0.3278310966933028
For n_clusters = 13, silhouette score is 0.3265853708120289
For n_clusters = 14, silhouette score is 0.323161307633767
For n_clusters = 15, silhouette score is 0.329544402858679
```

As we can see from the above scores, the highest silhouette score is 0.348 for the number of clusters equal to 3

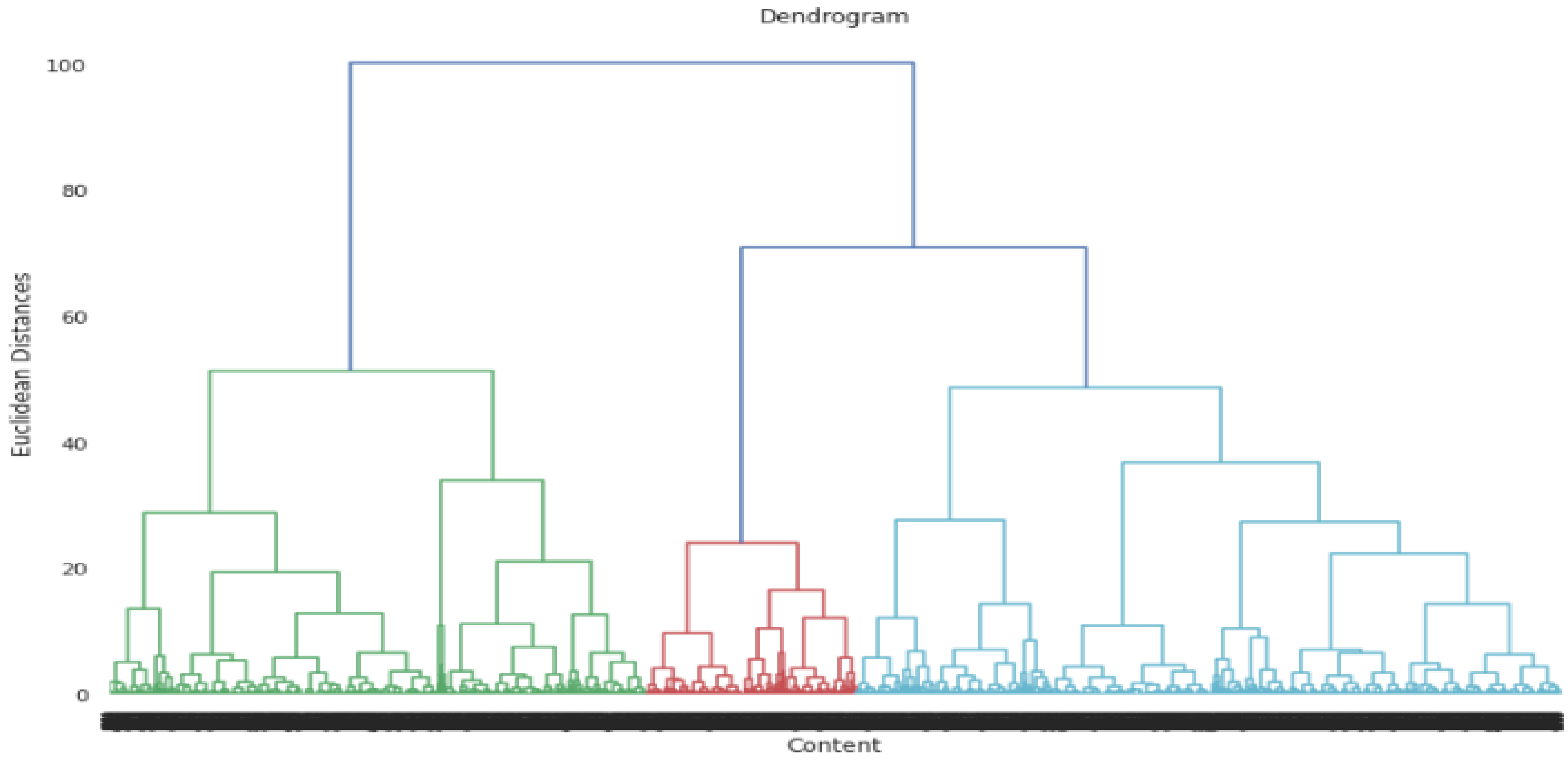
K – MEANS CLUSTERING





As we can see from the elbow method, the optimal number of clusters is also 3

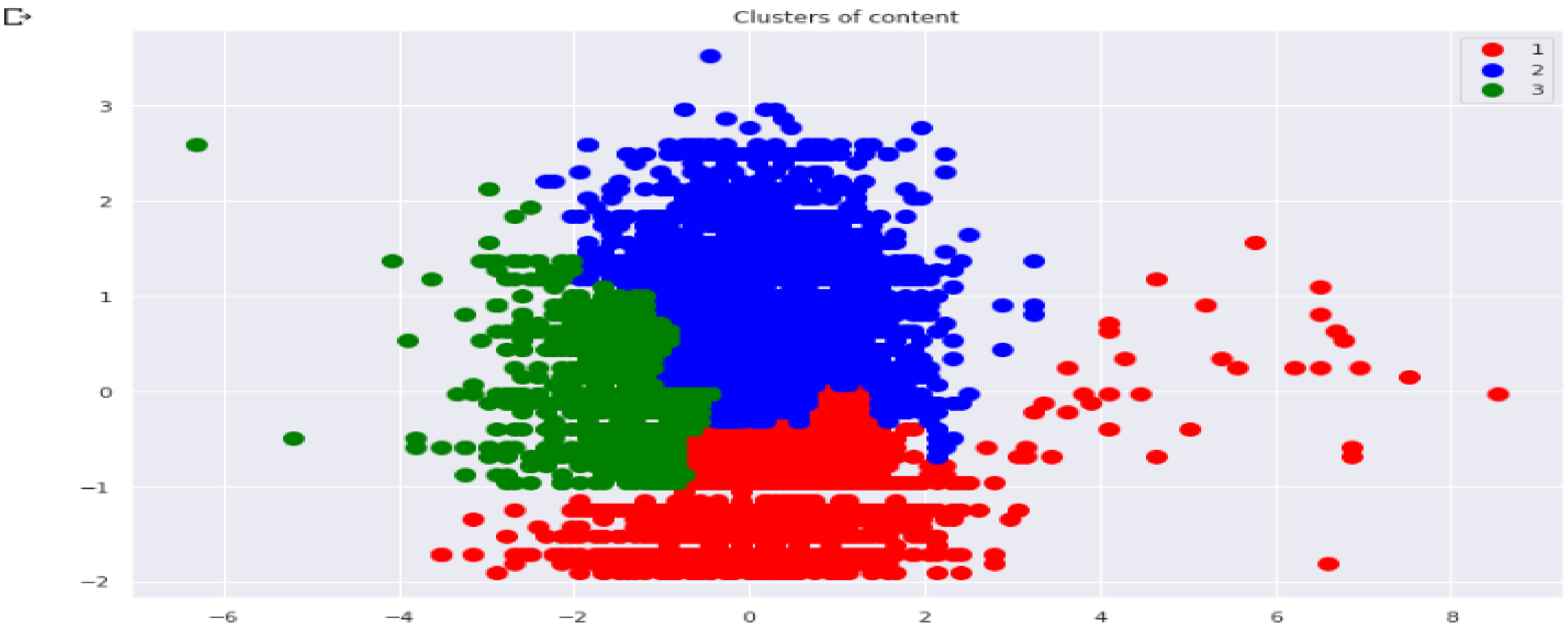
DENDROGRAM



The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold

No. of Cluster = 3

AGGLOMERATIVE HIERARCHICAL CLUSTERING



By applying different clustering algorithem to our dataset .we get the optimal number of cluster is equal to 3

CONCLUSION

1.Data set contains 7787 rows and 12 columns in that cast and director features contains large number of missing values so we can drop it and we have 10 features for the further implementation
2.We have two types of content TV shows and Movies (30.86% contains TV shows and 69.14% contains Movies)
3.By analysing the content added over years we get to know that in recent years Netflix is focusing movies than TV shows (movies is increased by 80% and TV shows is increased by 73% compare to 2016 data)
4.The most number of the movies and TV shows release in 2017 and 2020 respectively and united nation have the maximum content on Netflix
5.On Netflix, Dramas genre contains the maximum content among all of the genres and the most of the content added in December month and less content in February
6.By applying the silhouette score method for n range clusters on dataset we got best score which is 0.348 for 3 clusters it means content explained well on their own clusters, by using elbow method after k = 3 curve gets linear it means k = 3 will be the best cluster
7.Applied different clustering models K means, hierarchical, Agglomerative clustering on data we got the best cluster arrangements
8.By applying different clustering algorithms to our dataset .we get the optimal number of cluster is equal to 3

THANK YOU