Capstone Project - 4

Netflix Movies and TV shows Clustering

Navin Kodam
Jyoti Chiluka

Content :

1. Defining problem statement
2. EDA and feature engineering
3. Feature Selection
4. Data Preprocessing
5. Applying different clustering methods
6. Applying Clustering Models
7. Conclusion

The Dilemma :

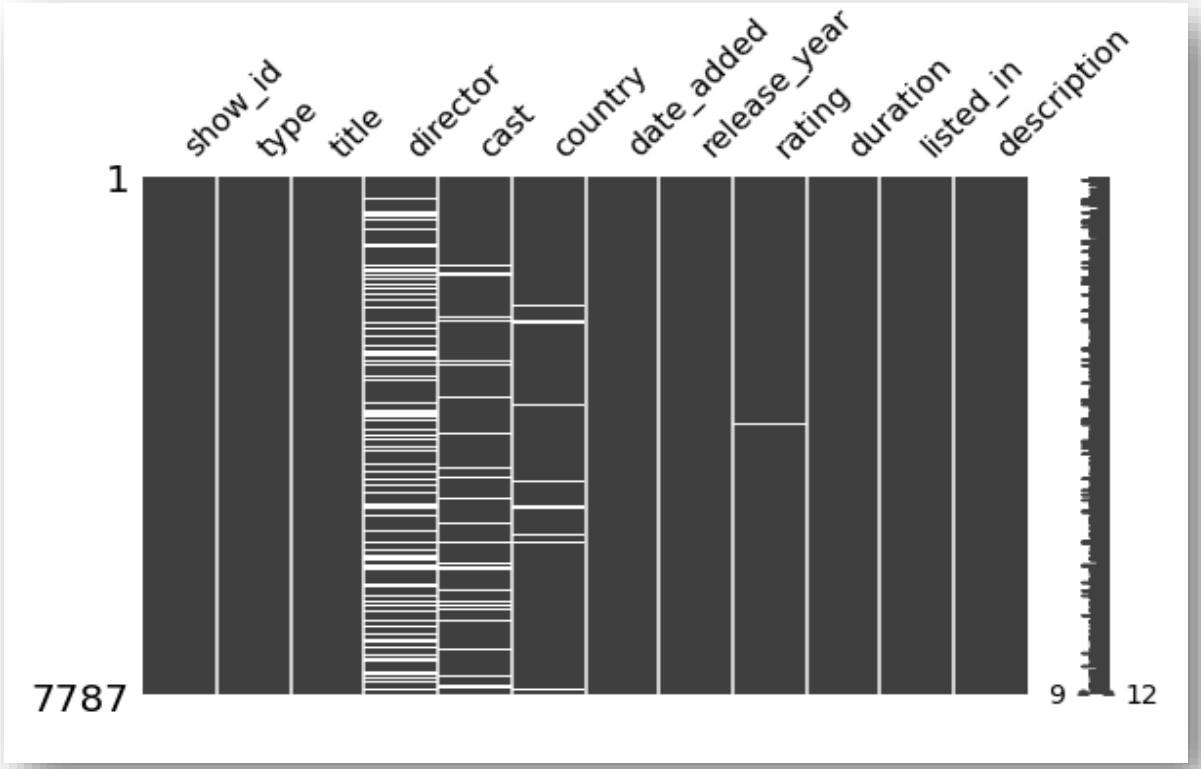
This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine.
In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming 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

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

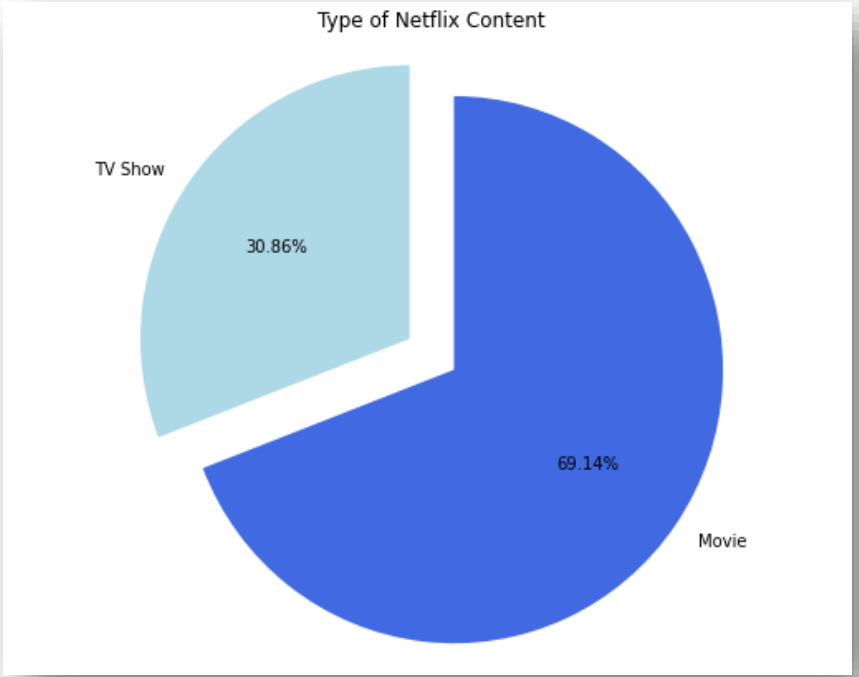
Exploratory Data Analysis

Missing data :
Here we can see that large no of null values present in director and cast features so we can drop these features



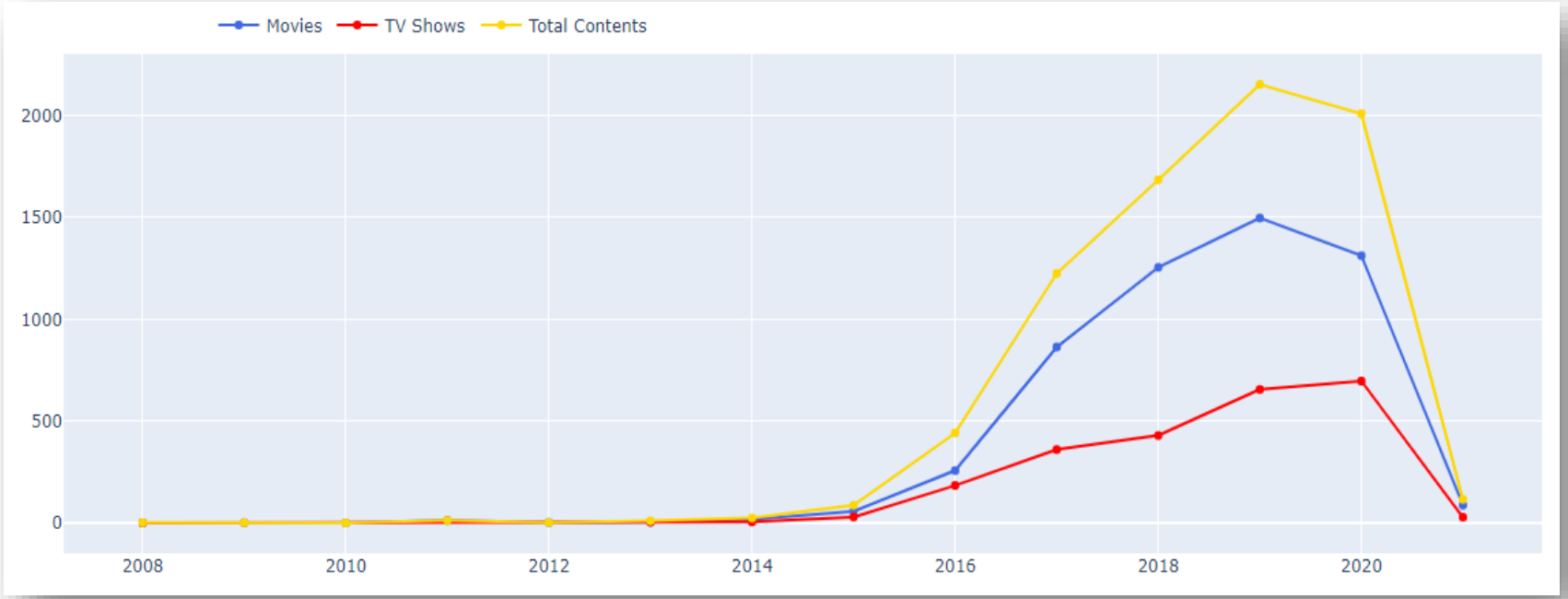
Eda contd...

Different types of content present In the Netflix

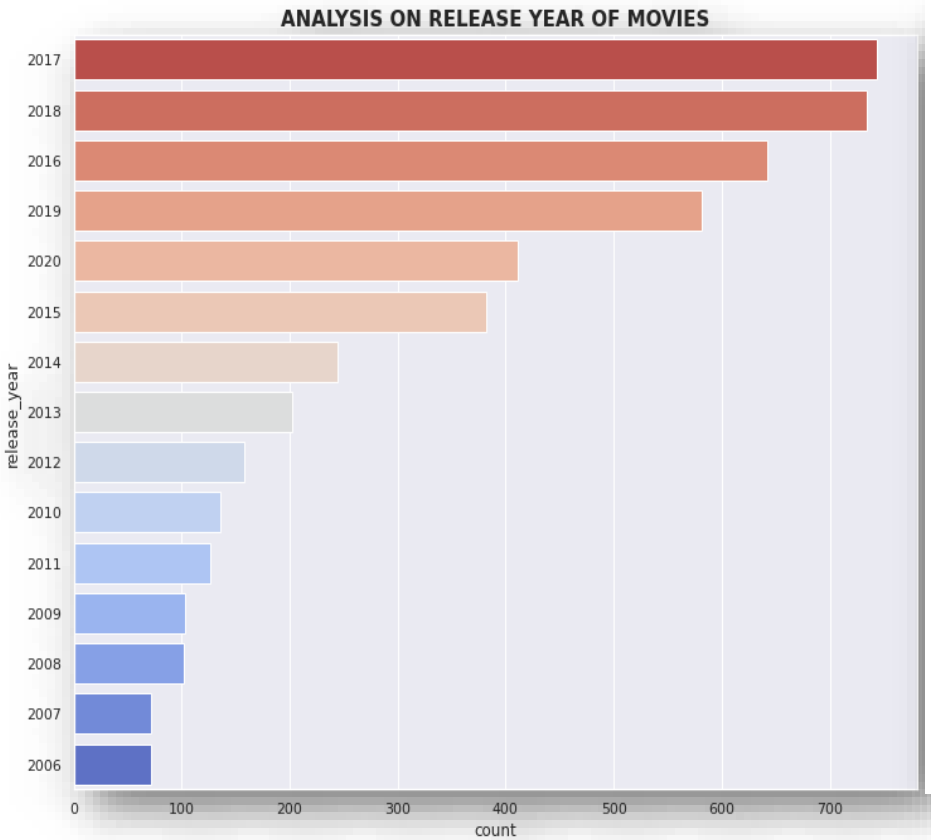


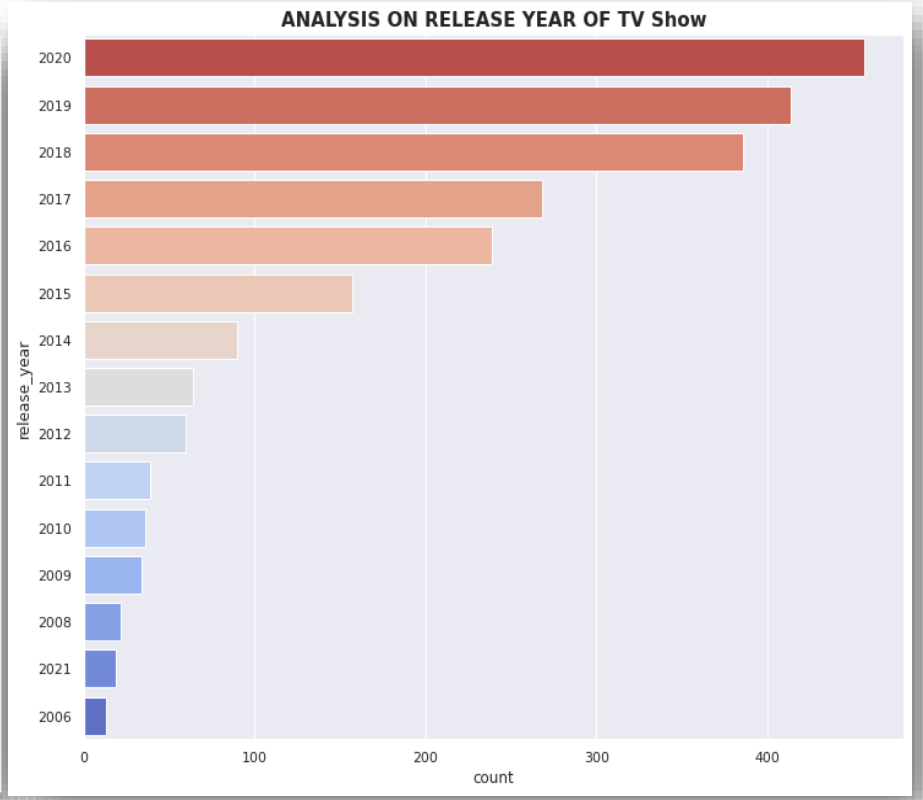
Eda contd...

Content added over the years



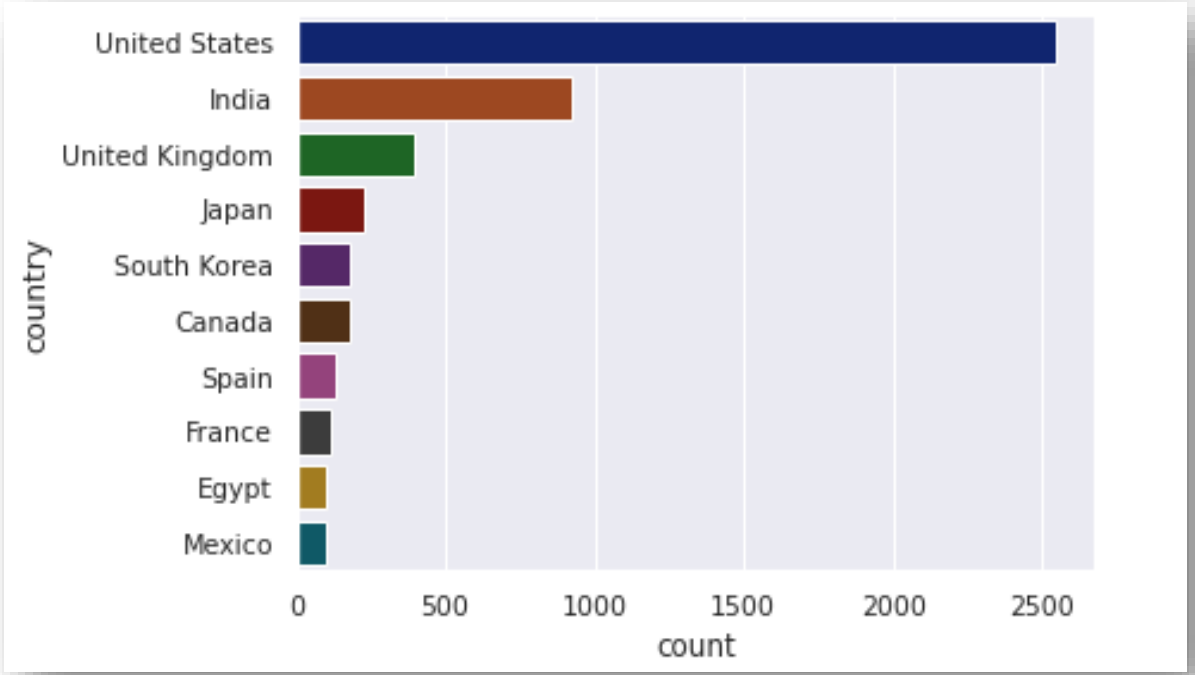
Eda contd...


ANALYSIS ON RELEASE YEAR OF MOVIES
release_year
2017
2018
2016
2019
2020
2015
2014
2013
2012
2010
2011
2009
2008
2007
2006
0
100
200
300
400
500
600
700
count


ANALYSIS ON RELEASE YEAR OF TV Show
release_year
2020
2019
2018
2017
2016
2015
2014
2013
2012
2011
2010
2009
2008
2021
2006
0
100
200
300
400
count

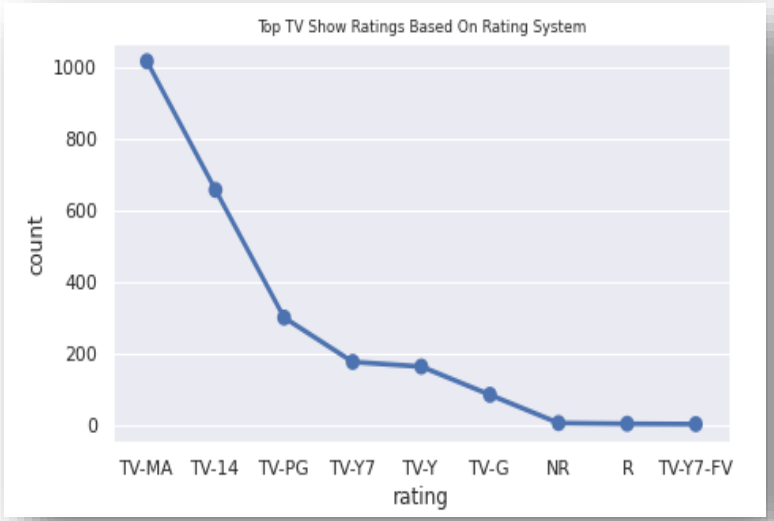
Eda contd...

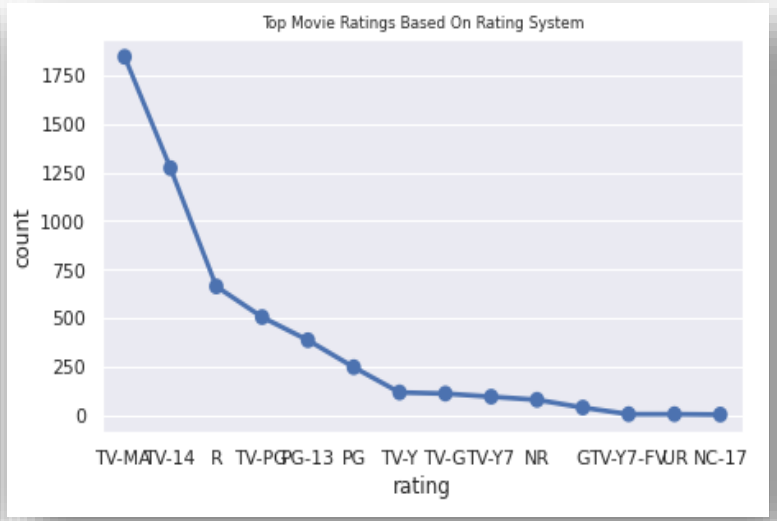
United States has the most number of content on Netflix



Eda contd...

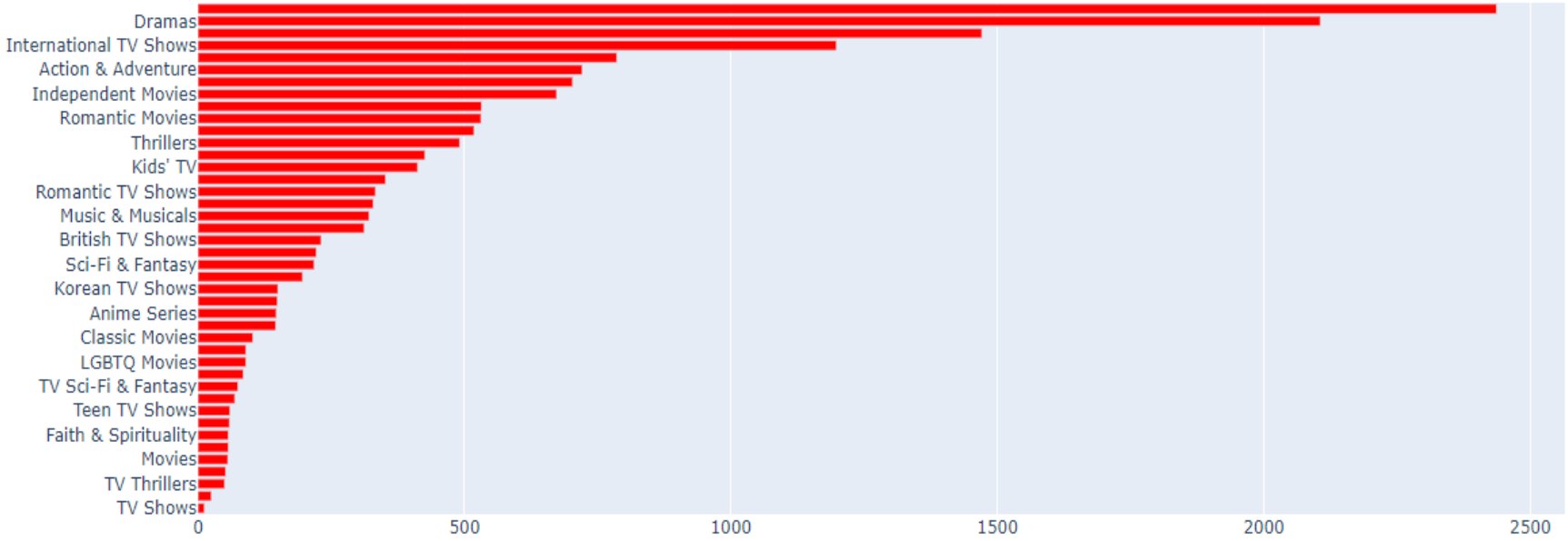
Top TV show and Movie ratings





Eda contd...

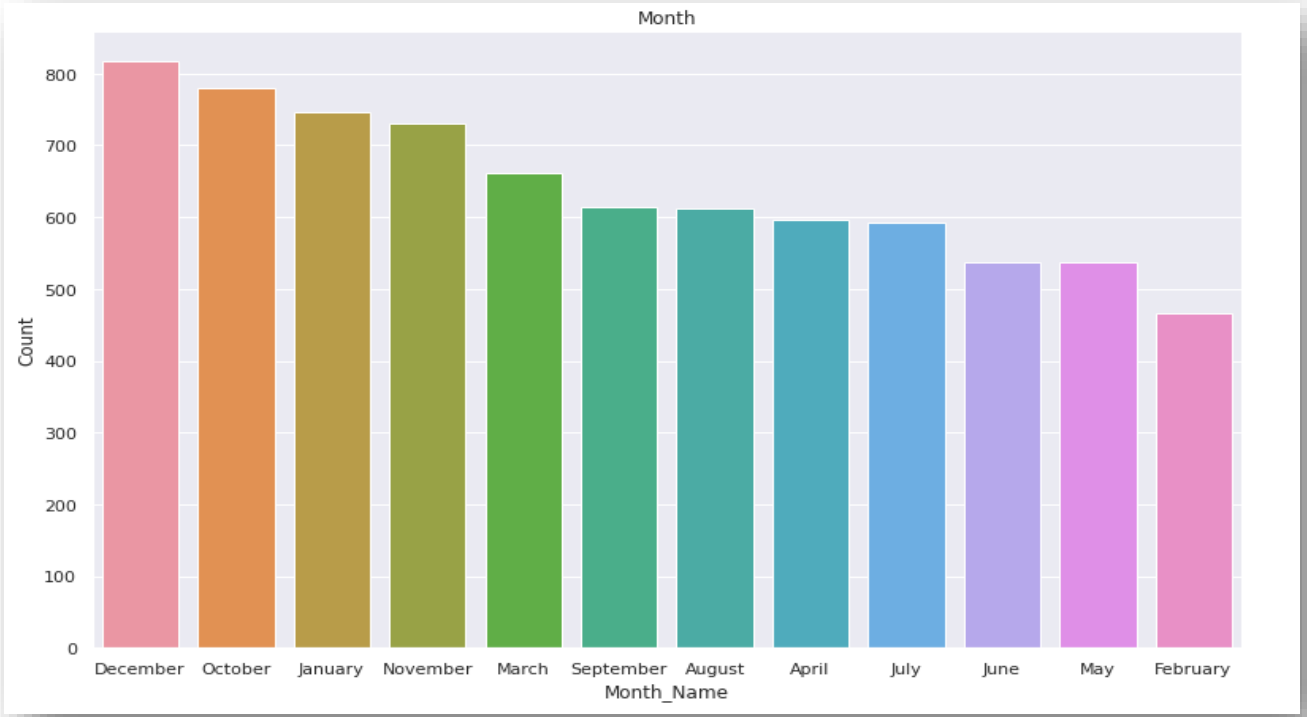
Dramas are the maximum content present in the Netflix



Eda contd...

Country wise content added

The Maximum Content added in December and minimum in February

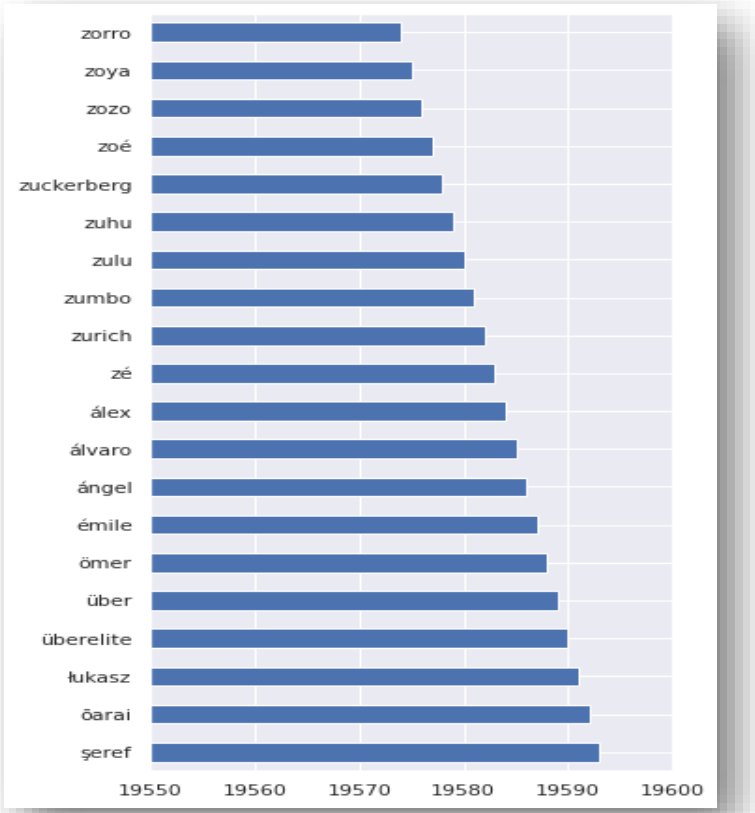


Preparing dataset for modeling(Data preprocessing)

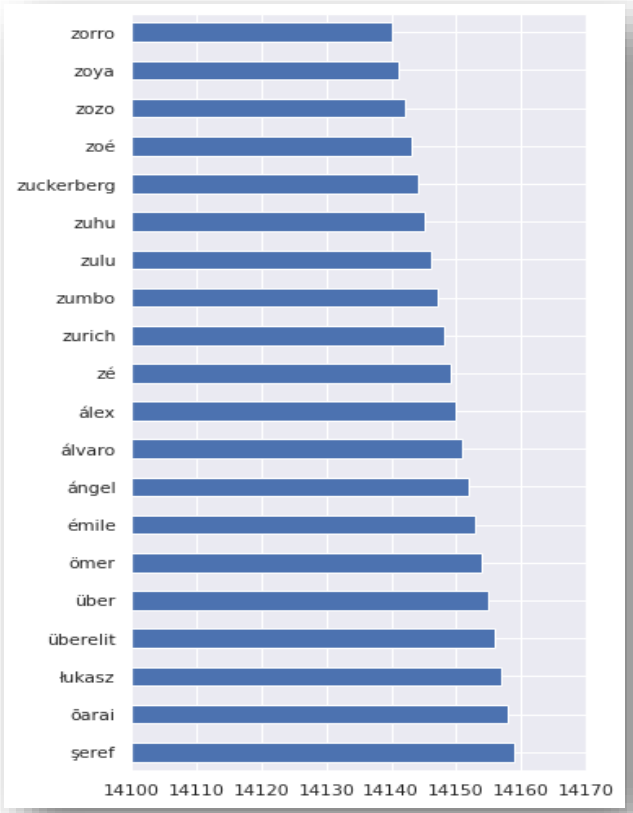
1. Working on the text based features (description, listed_in)
2. Removing punctuations and stop words from text features
3. Stemming process applied for those text features
4. Applying the count vectorizer on those update text

Data preprocessing contd...

Stemming before(description)

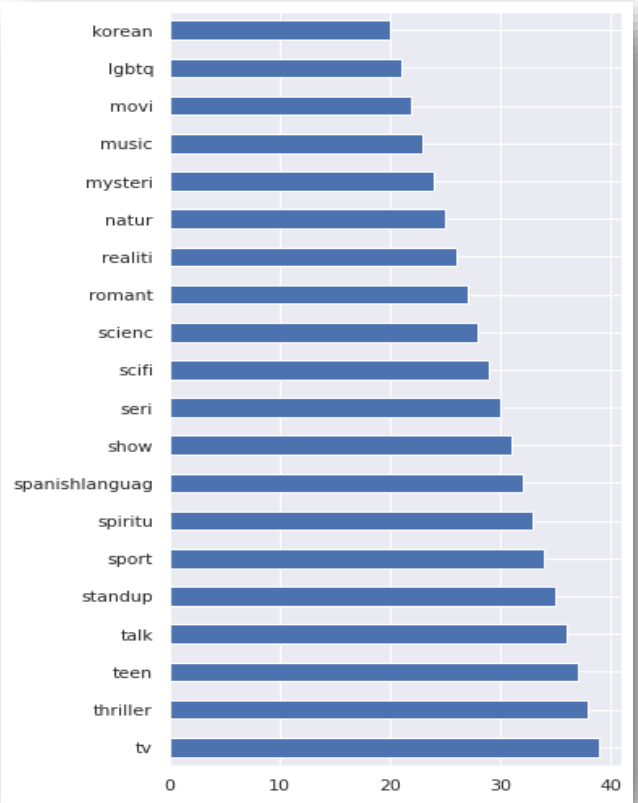


Stemming after(description)

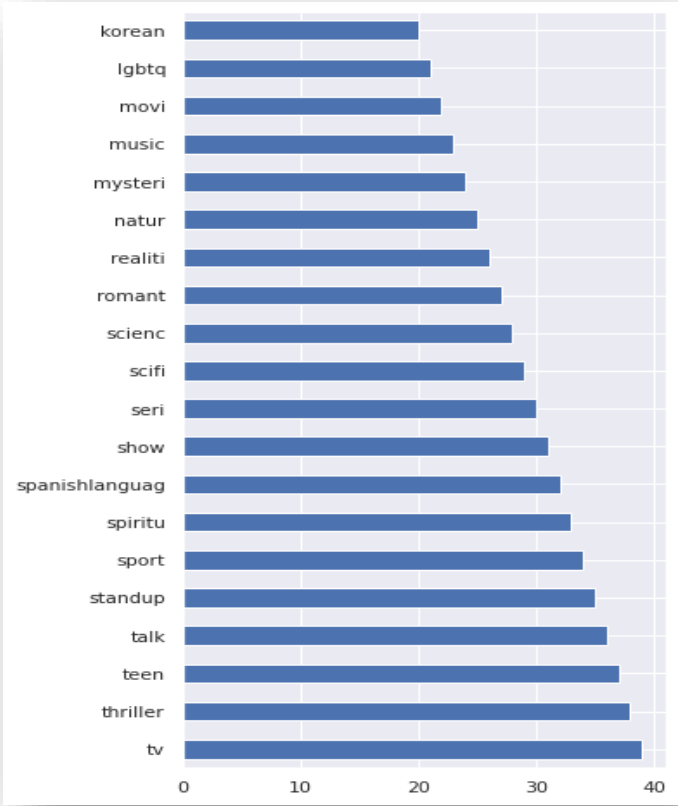


Data preprocessing contd...

Stemming before(listed_in)



Stemming after(listed_in)



Implementing clustering methods

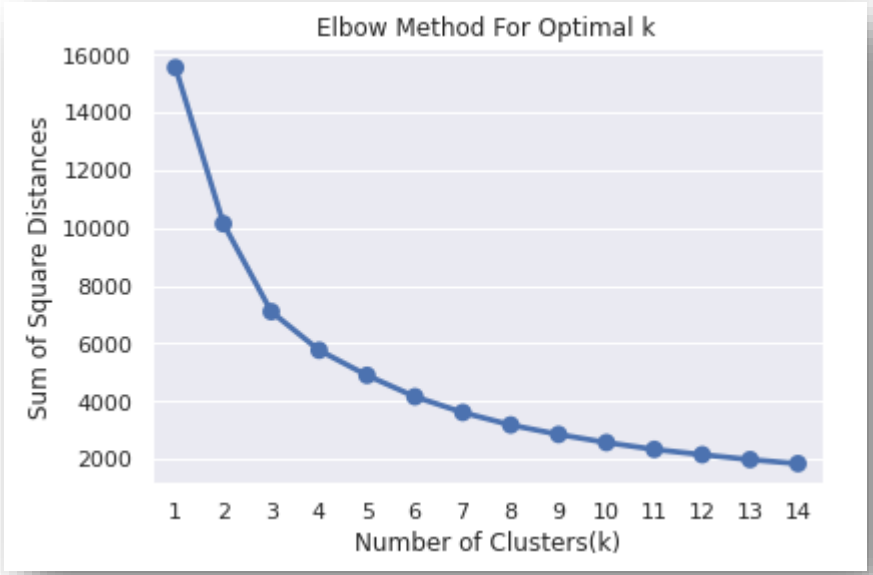
Silhouette Score Method (n range clusters)

Optimum score is 0.347 for n_clusters = 3

```
For n_clusters = 2, silhouette score is 0.3363965040356366
For n_clusters = 3, silhouette score is 0.3478645619691327
For n_clusters = 4, silhouette score is 0.31891157662083064
For n_clusters = 5, silhouette score is 0.30838620971694375
For n_clusters = 6, silhouette score is 0.3265885983437178
For n_clusters = 7, silhouette score is 0.3244775375499722
For n_clusters = 8, silhouette score is 0.32212131813075195
For n_clusters = 9, silhouette score is 0.3205876972612368
For n_clusters = 10, silhouette score is 0.3221508857145542
For n_clusters = 11, silhouette score is 0.3260472504304281
For n_clusters = 12, silhouette score is 0.32379706776790246
For n_clusters = 13, silhouette score is 0.3221718375843846
For n_clusters = 14, silhouette score is 0.3273690321995855
For n_clusters = 15, silhouette score is 0.33014747220177704
```

Implementing clustering methods

Applying elbow method

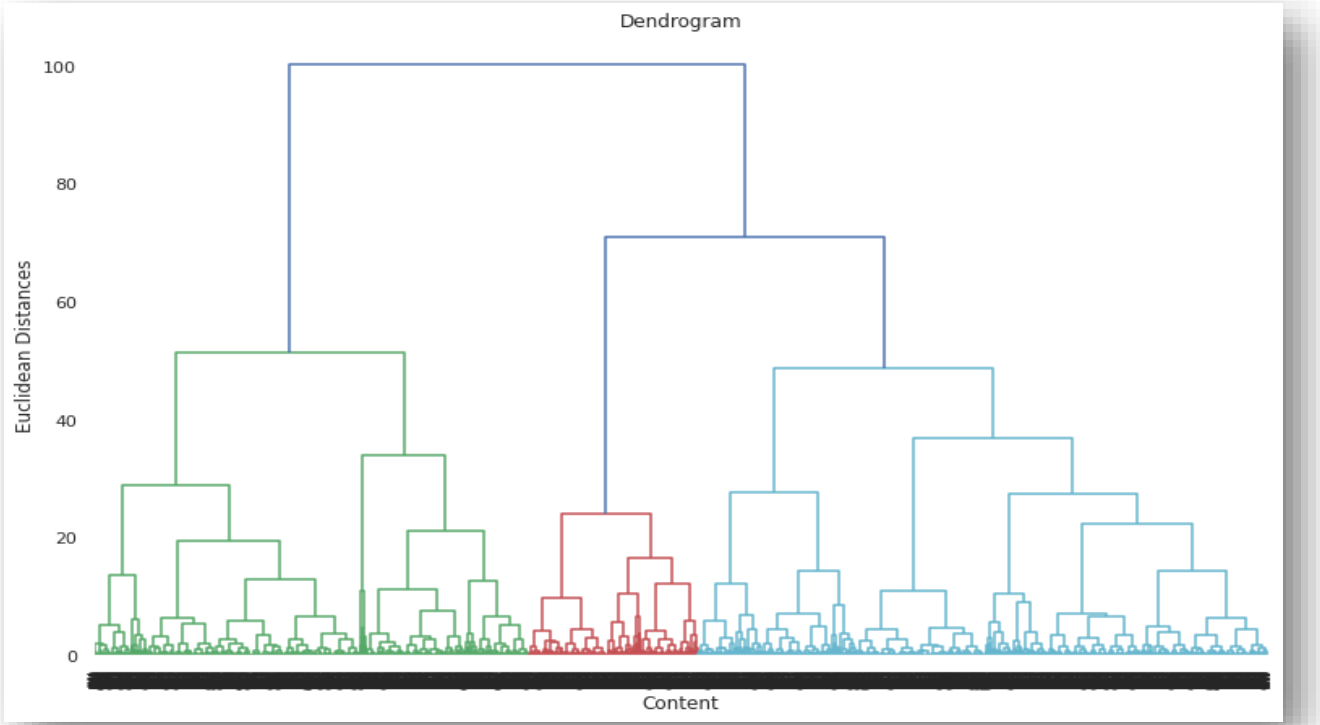


Applying KMeans clustering

For k = 3

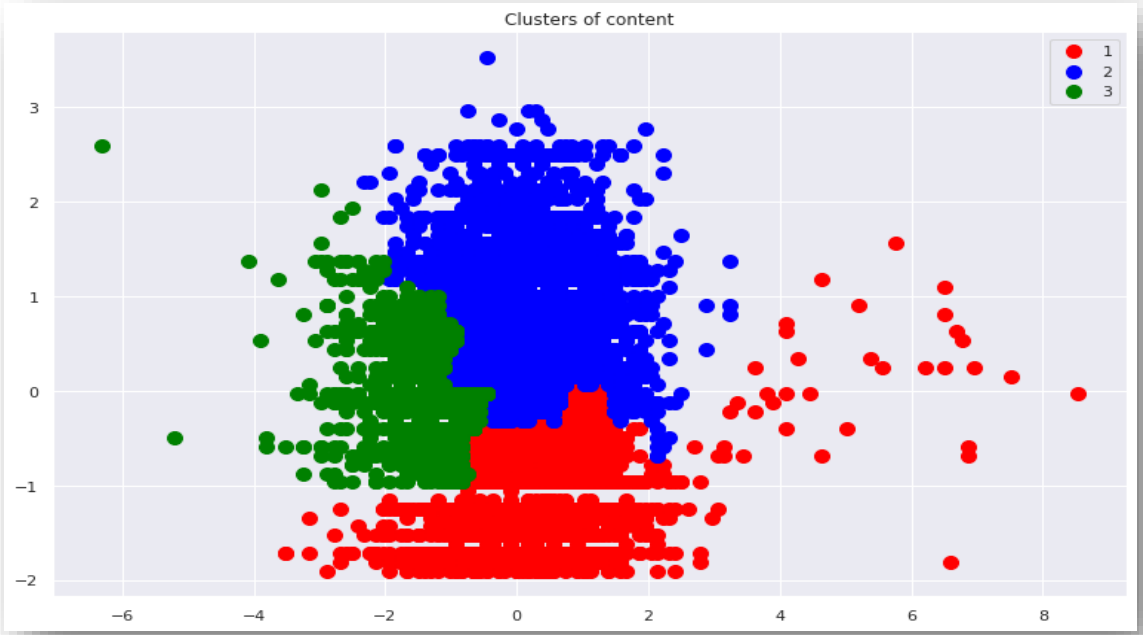
Applying Hierarchical clustering

Dendrogram to find the optimal number of clusters



Applying Hierarchical(Agglomerative clustering)

Optimal number of clusters are 3



Conclusion

- 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
- We have two types of content TV shows and Movies (30.86% contains TV shows and 69.14 % contains Movies)
- 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)
- The most number of the movies release in 2017 and TV shows in 2020 respectively and united nation have the maximum content on netflix
- 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
- By applying the silhouette score method for n range clusters and we got the best score which is 0.348 for 3 clusters it means content explained well on their own clusters, by using elbow method we also got k cluster is 3
- Applied different clustering models Kmeans, hierarchical, Agglomerative clustering on data we got the best cluster arrangements
- By applying different clustering algorithms to our dataset .we get the optimal number of cluster is equal to 3

Q & A