# Capstone Project-4

## NETFLIX MOVIES & TV SHOWS CLUSTERING

**Team Members**

*1)chandan Kumar Raxit*
*2)Deepak Kumar Jena*

# CONTENT

- Introduction
- Problem Statement
- Data Summary
- Data wrangling
- EDA
- Text Preprocessing
- K-means Clustering
- Feature Selection & ML algo used
- Conclusion

# Introduction

- Netflix is a prominent OTT platform with a wide variety of content to view from a variety of nations and genres, so keep an eye on it. 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.
- The idea of this project is to analyze and perform clustering to determine various patterns related to the content available in Netflix. Based on the attributes related to the Tv shows or movies, we will be implementing different clustering algorithms which comes under unsupervised Machine learning category.

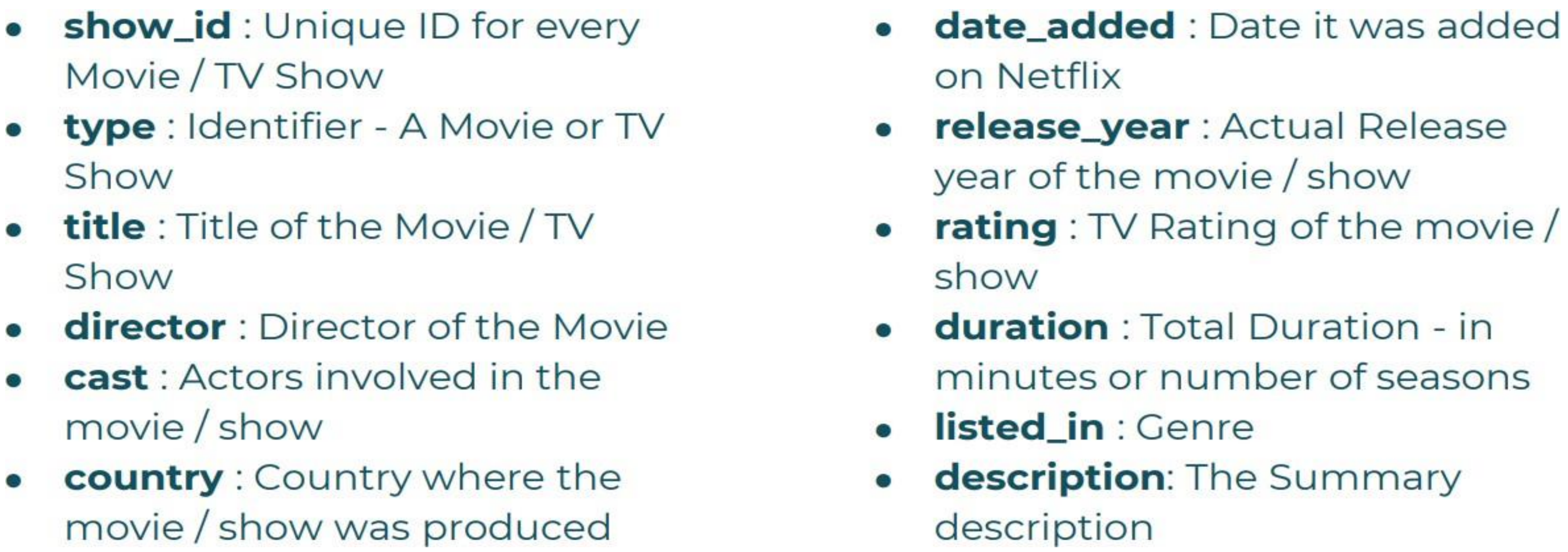
# PROBLEM STATEMENT

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

- 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

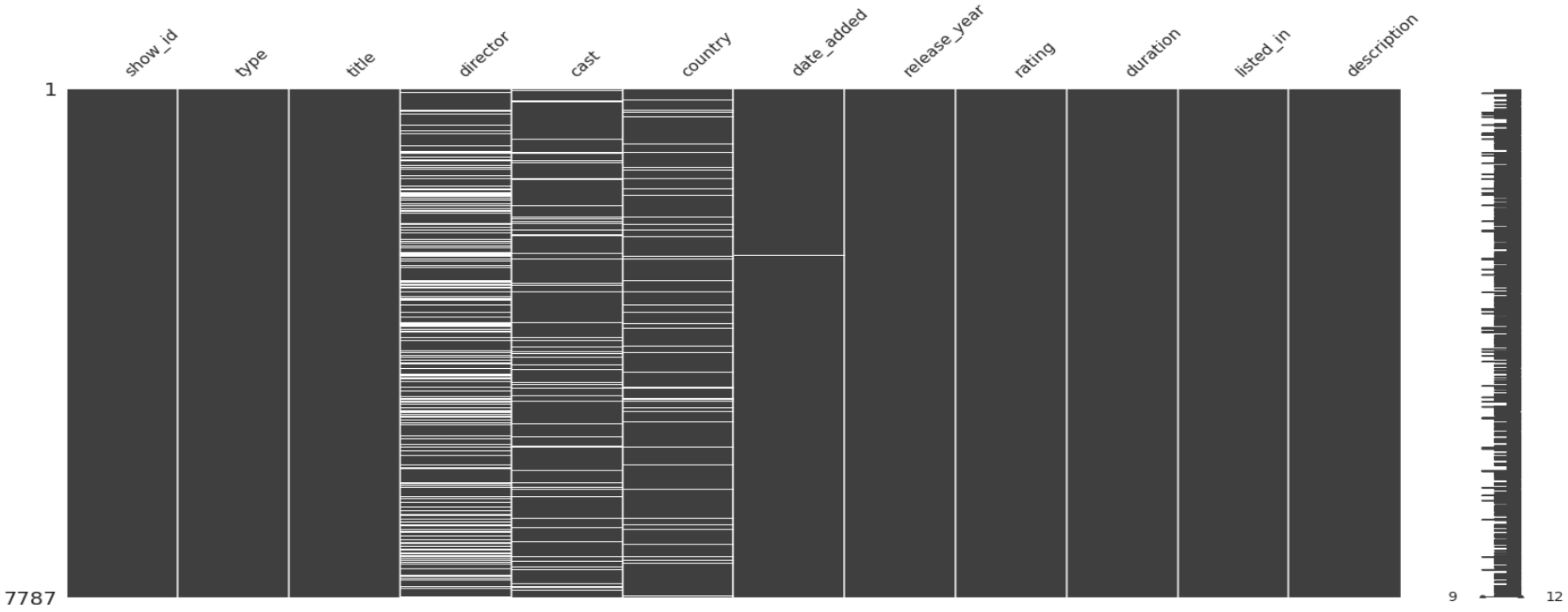
# DATA SUMMARY

**The dataset has 7787 rows and 12 columns.**



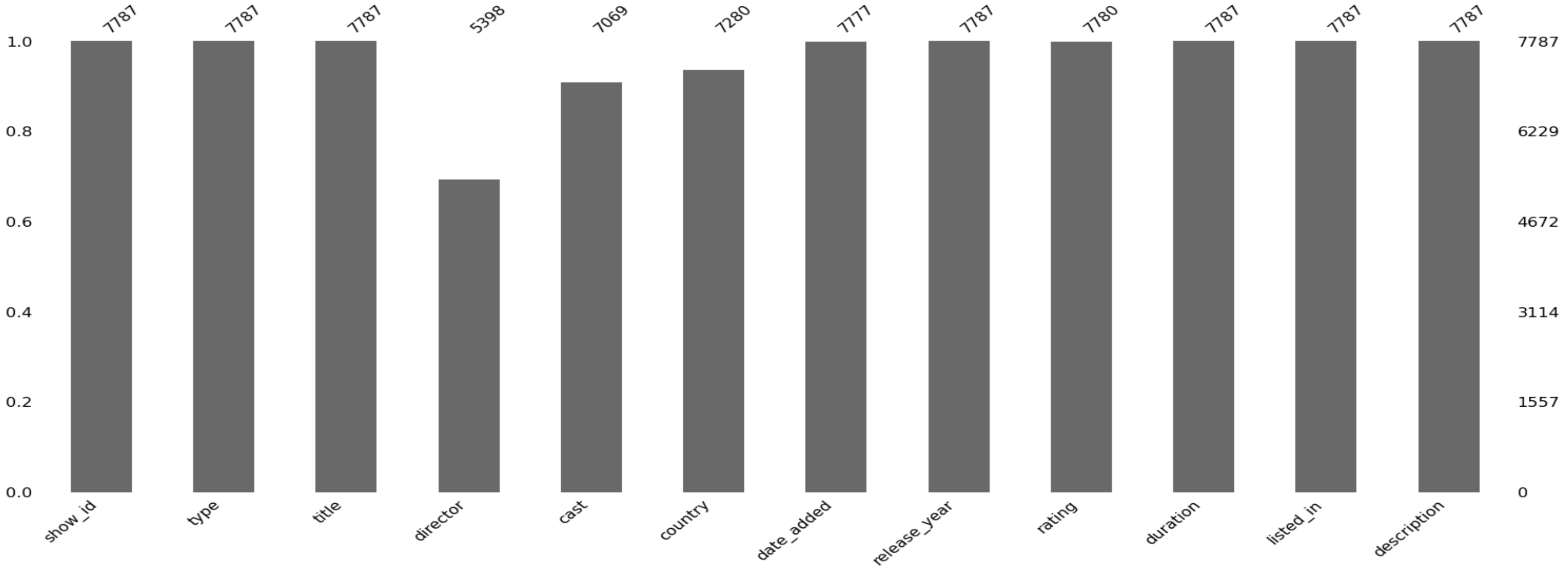
- **show_id** : Unique ID for every Movie / TV Show
- **type** : Identifier - A Movie or TV Show
- **title** : Title of the Movie / TV Show
- **director** : Director of the Movie
- **cast** : Actors involved in the movie / show
- **country** : Country where the movie / show was produced
- **date_added**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

# DATA WRANGLING



Missing values -

- “Director" has the most missing value followed by "cast" and "country".
- There are few missing value in "date_added" and "rating".



Missing values -

- “Director" has the most missing value followed by "cast" and "country".
- There are few missing value in "date_added" and "rating".

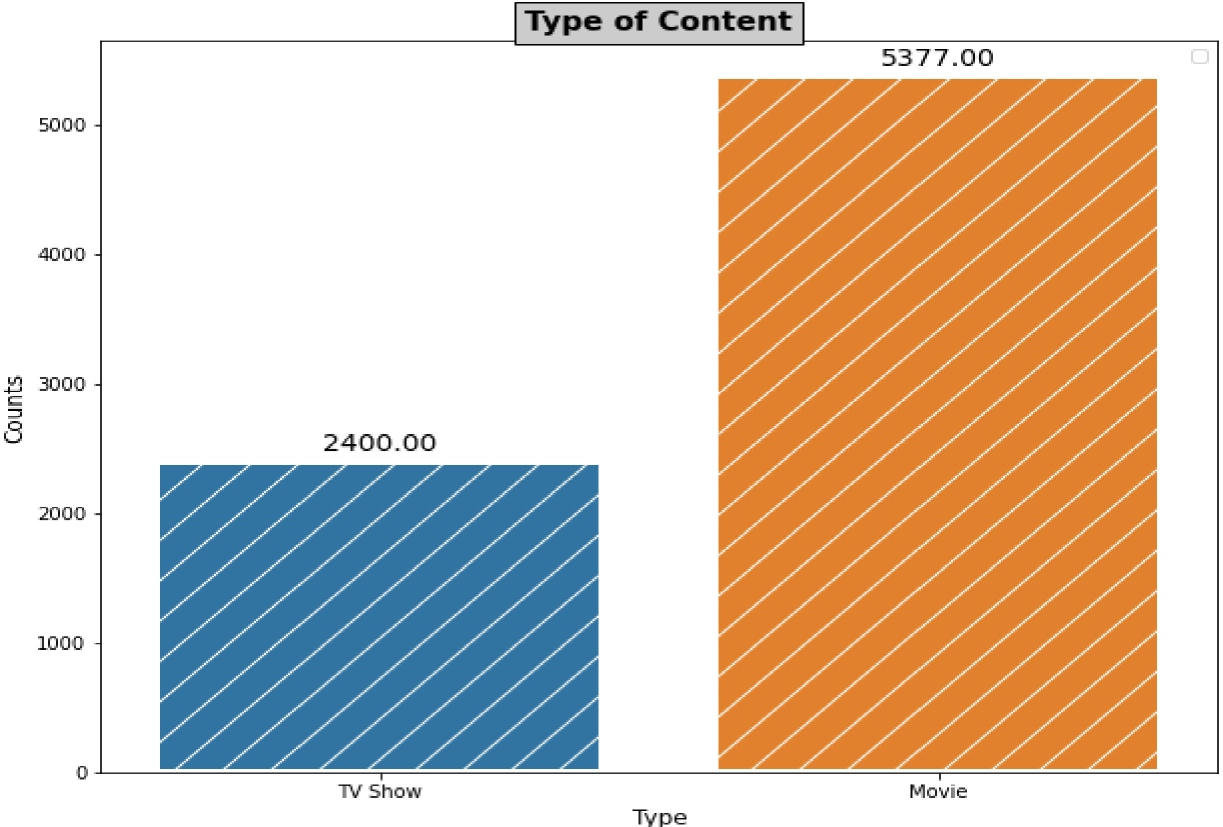
# Data Cleaning

**Null Value Treatment:**

- ***Director*** feature have more than ***30.68%*** of null values. Filling null values by ‵unknown′.
- ***Country*** feature have ***6.51%*** of null values. Filling null values by mode of feature.
- ***Cast*** feature have ***9.22%*** of null values. Filling null values by ‘unknown′.
- ***Rating*** feature have ***0.09%*** of null values. Filling null values by mode of feature.
- ***Date_added*** feature have ***0.13%*** of null values. Dropping rows corresponding to null values.

# Exploratory Data Analysis(EDA)

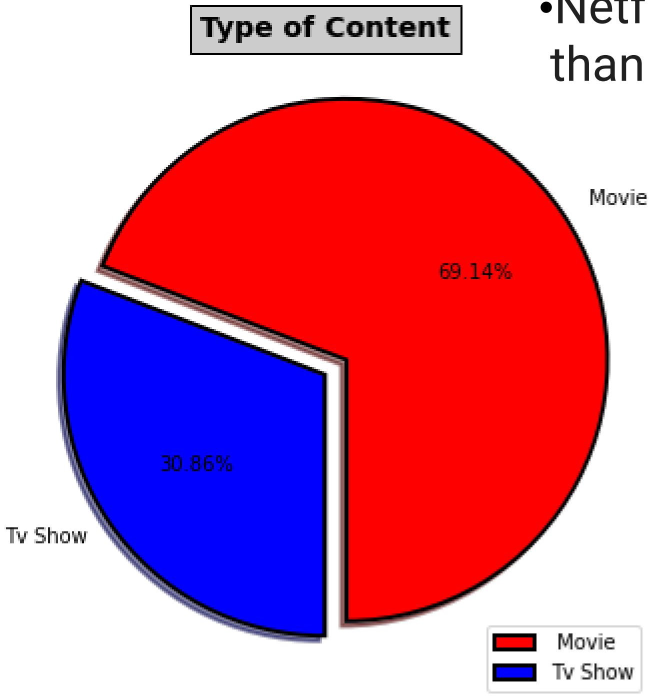
## Type of content available on Netflix

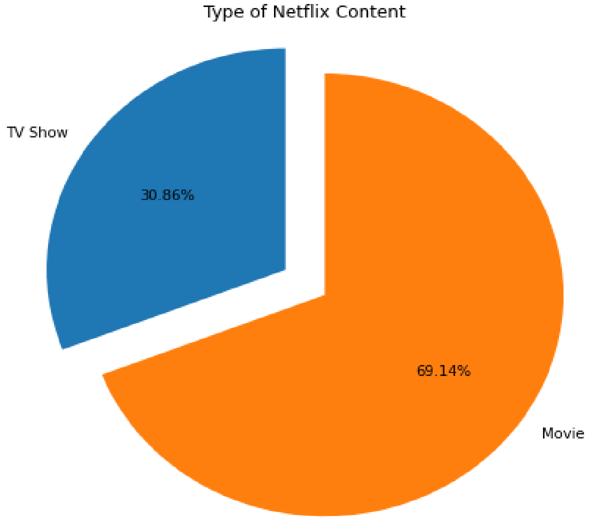


# Type of content available on Netflix

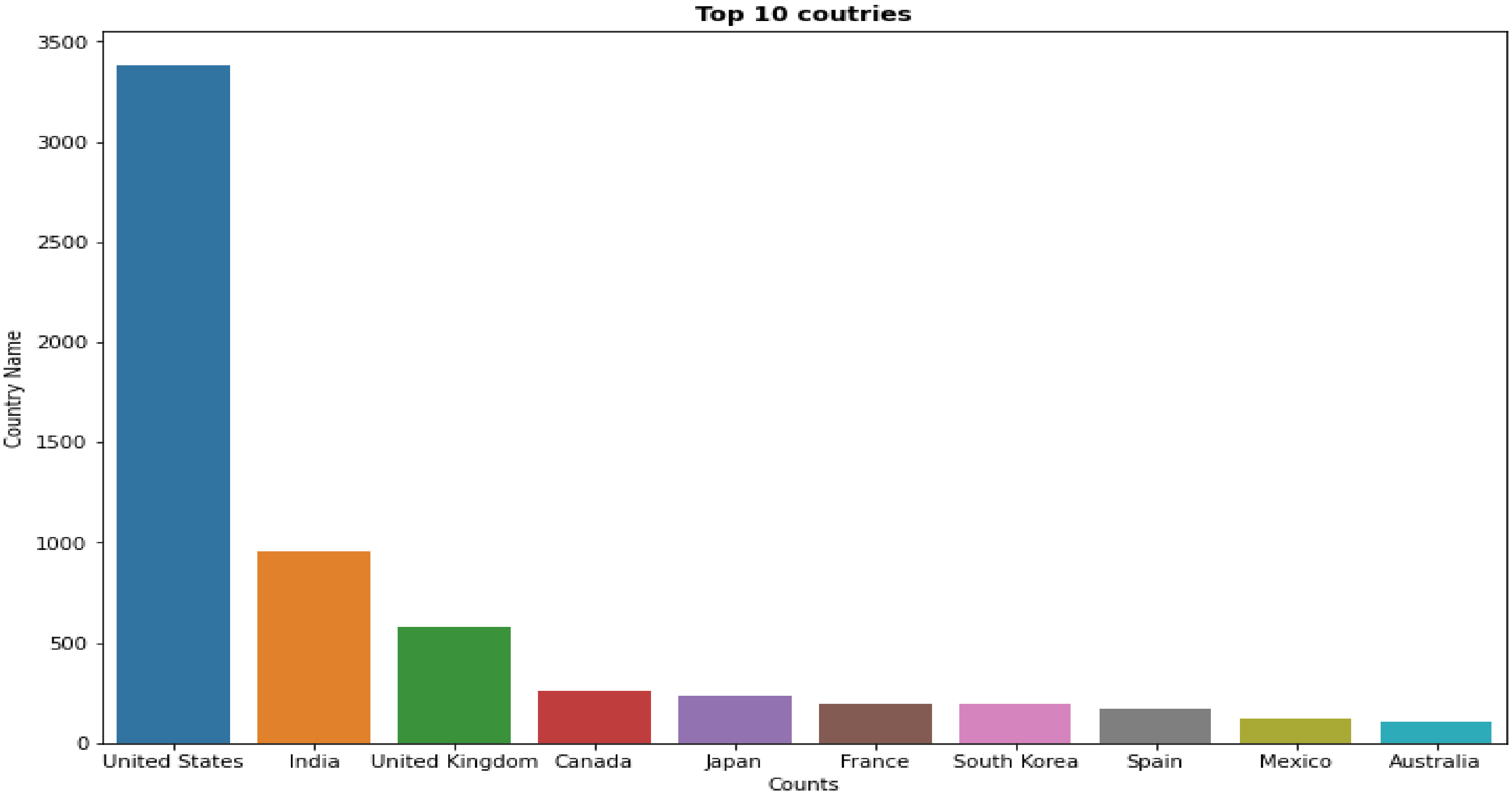
- It is evident that there are more movies on Netflix than TV shows.

- Netflix has 5377 movies, which is more than double the quantity of TV shows.

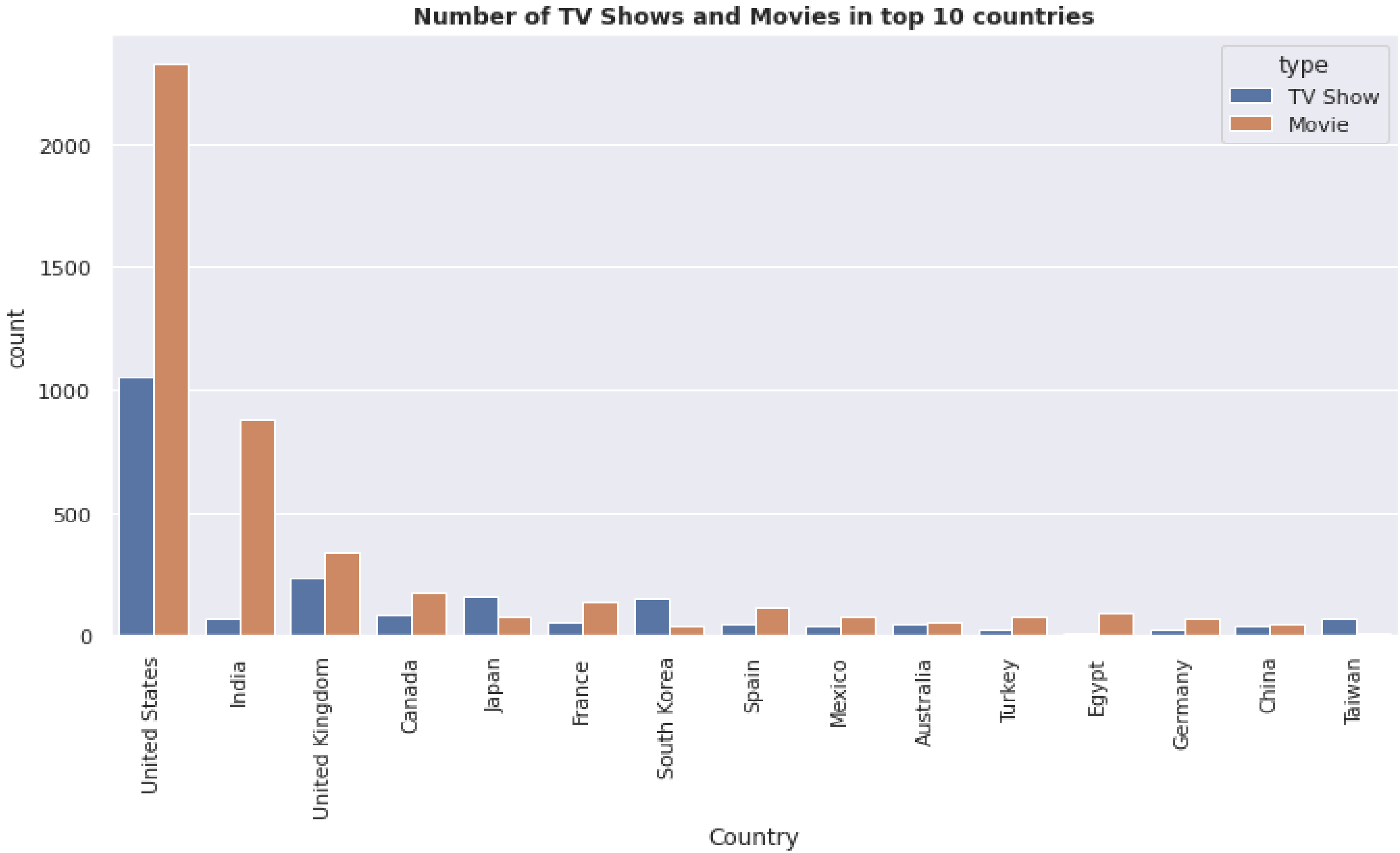




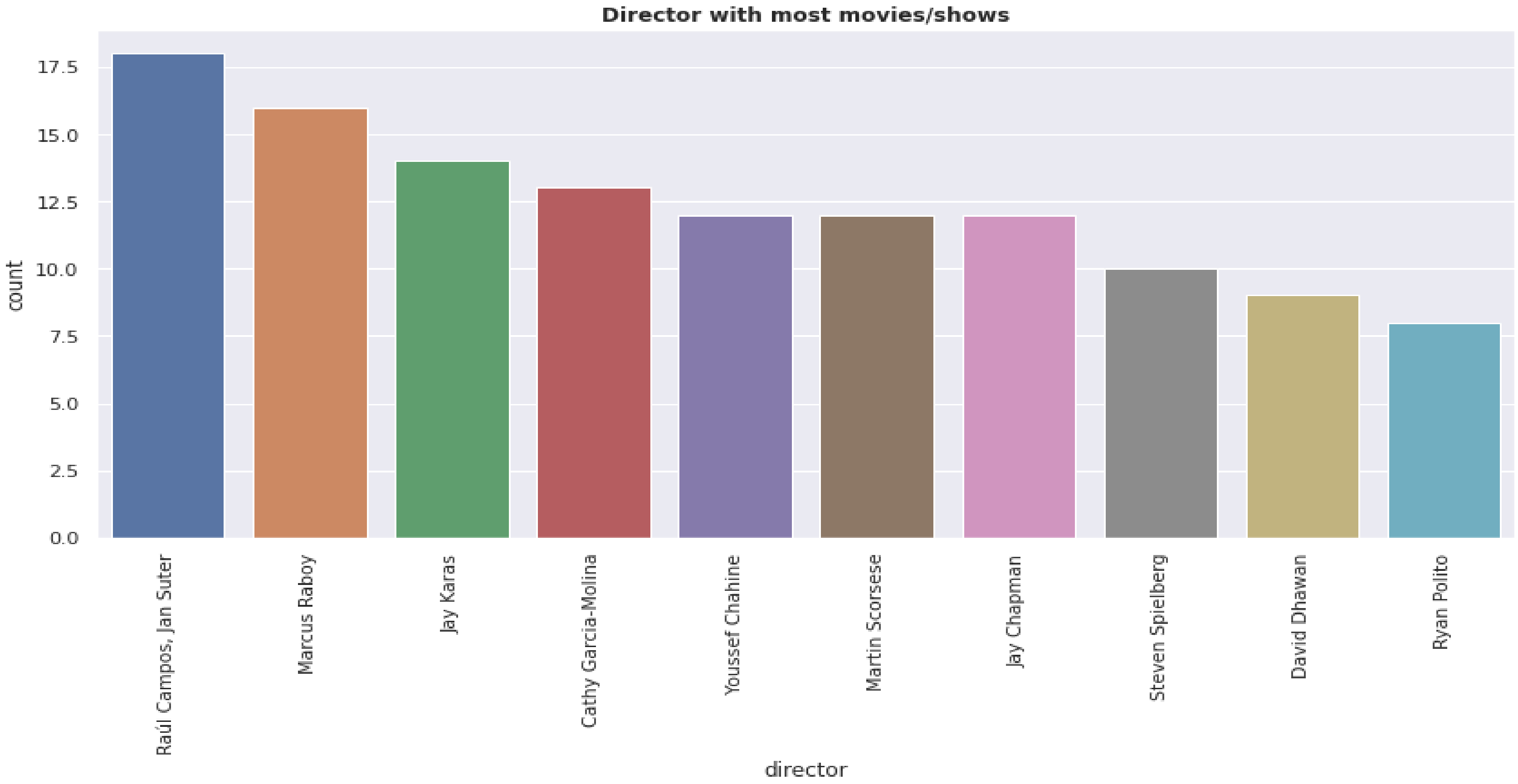
# Top 10 countries on Netflix



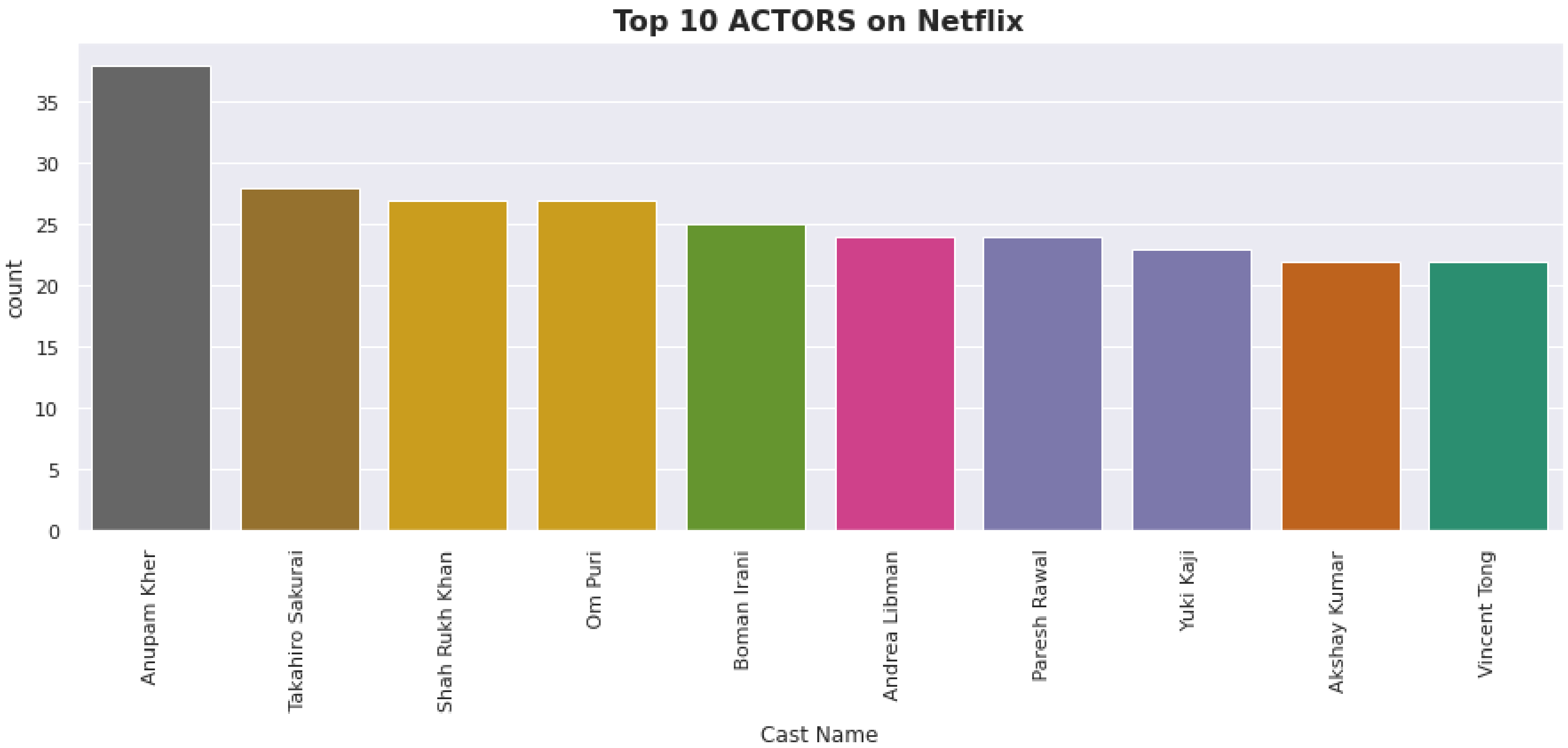
# Number of TV Shows and Movies content in top 10 countries with maximum content

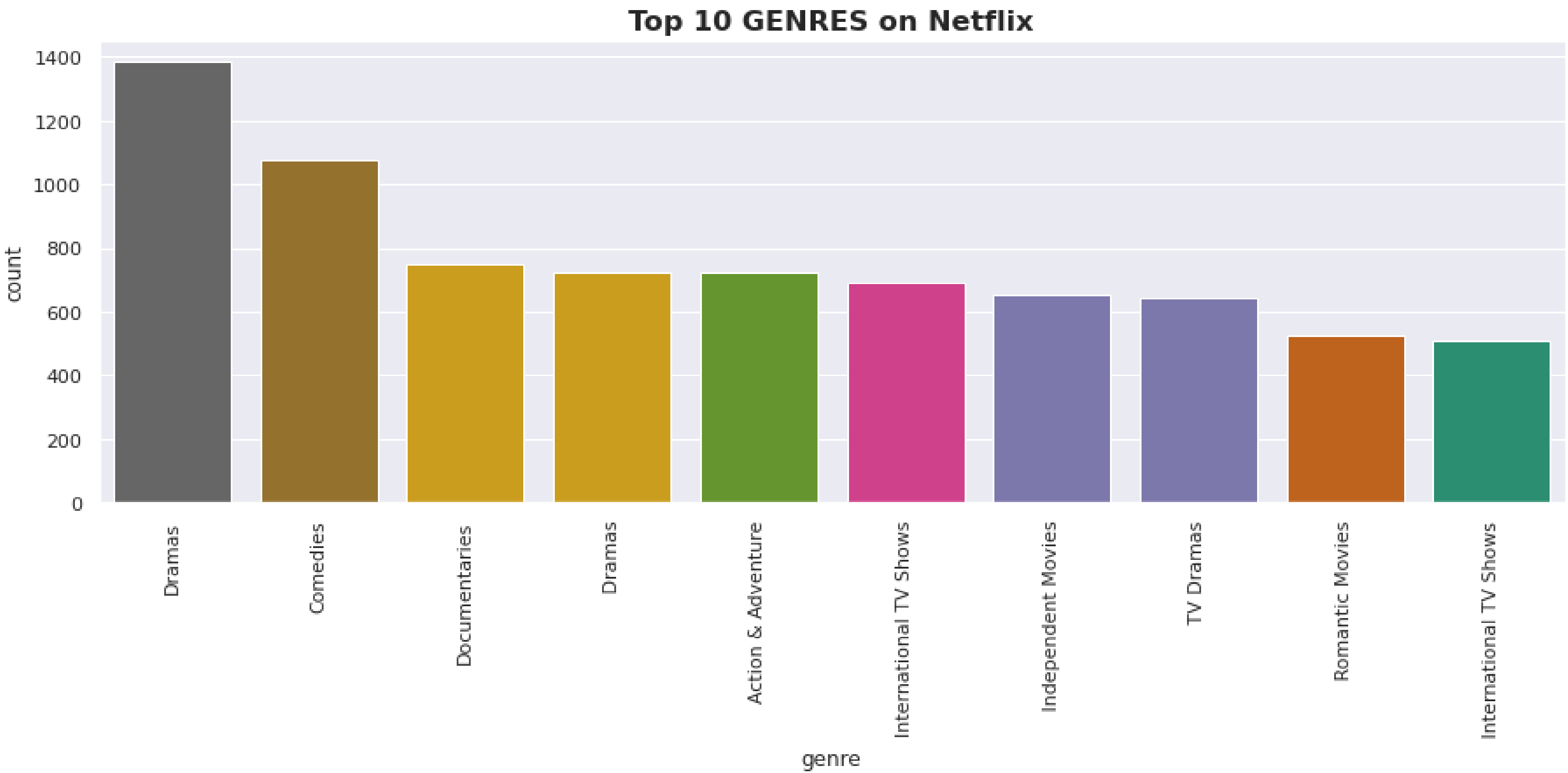


# Director with most movies/shows

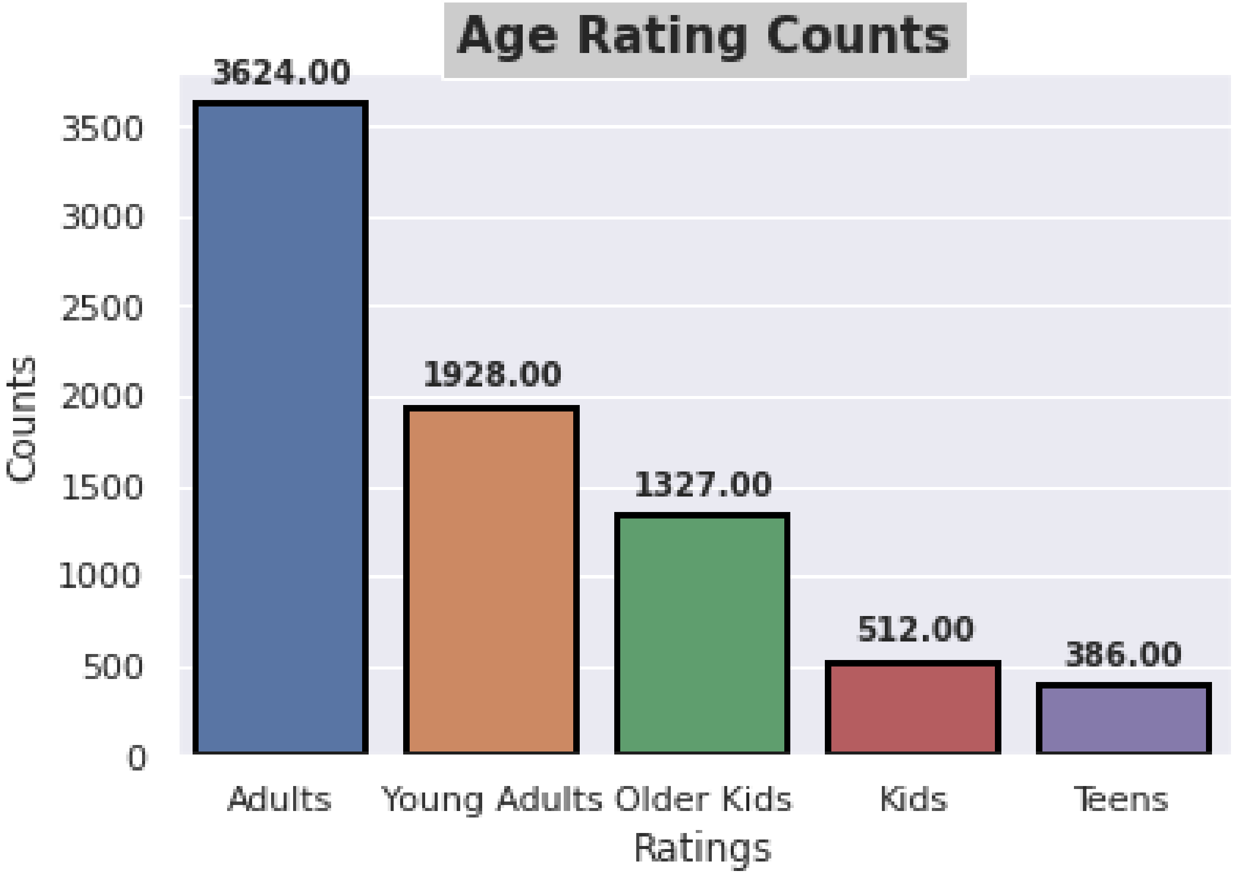


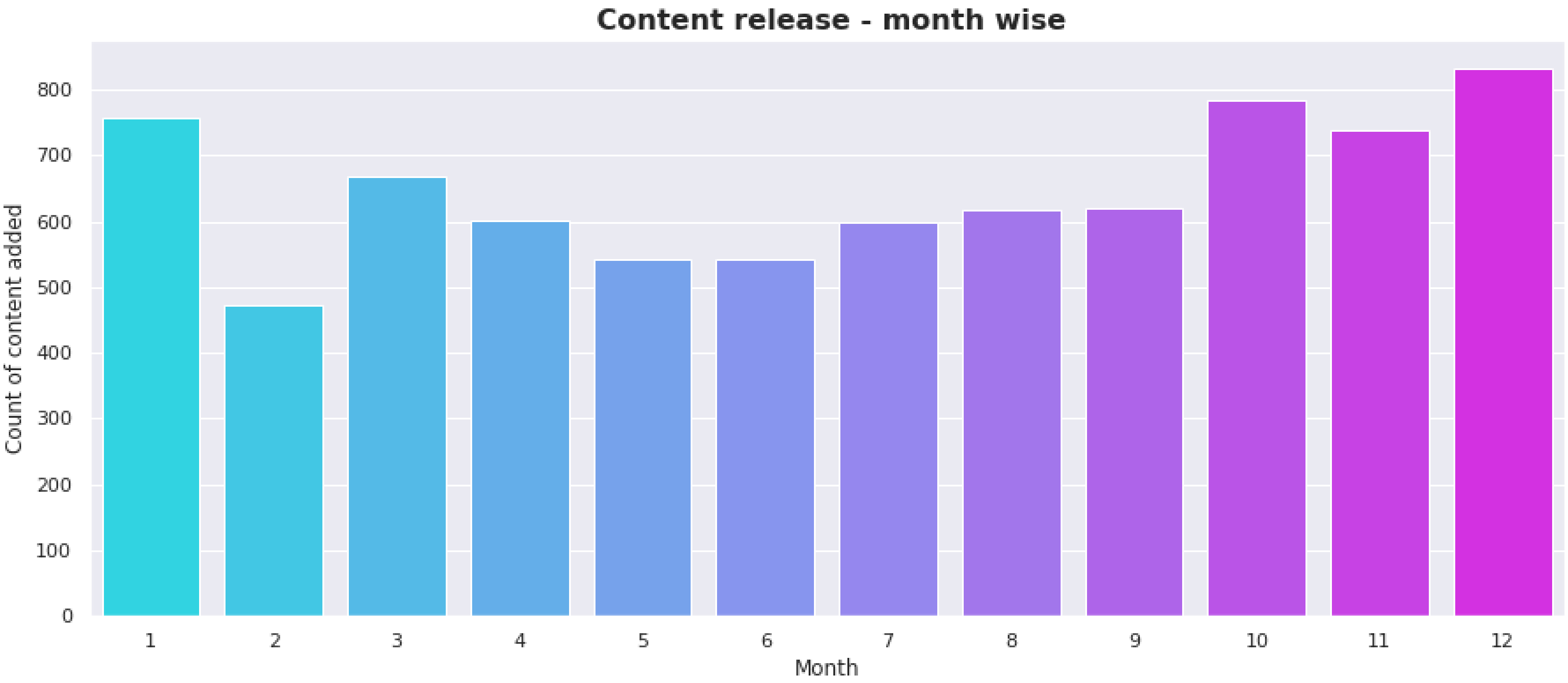
# Top 10 ACTORS on Netflix

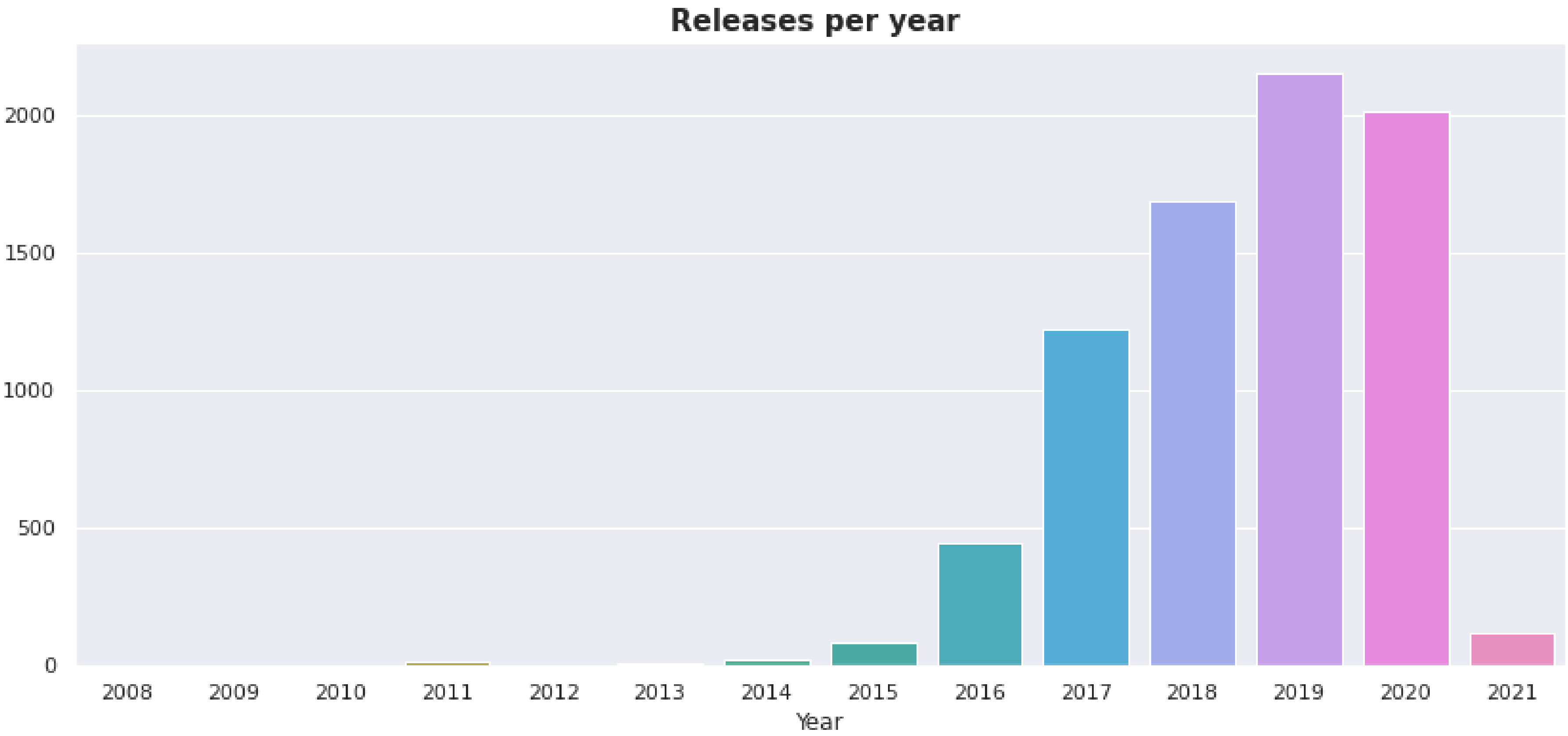




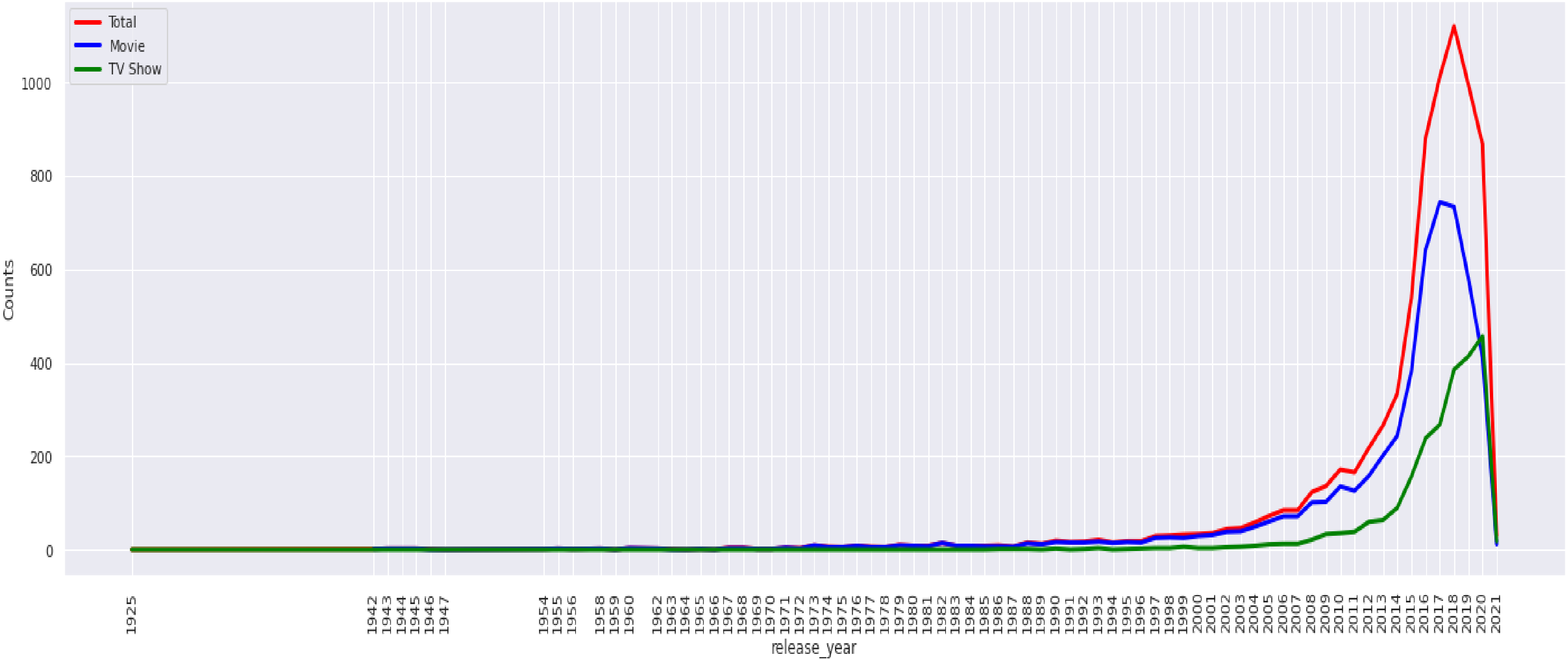
Drama is the most popular genre followed by comedy.


Age Rating Counts
3624.00
1928.00
1327.00
512.00
386.00
3500
3000
2500
2000
1500
1000
500
0
Counts
Adults
Young Adults
Older Kids
Kids
Teens
Ratings

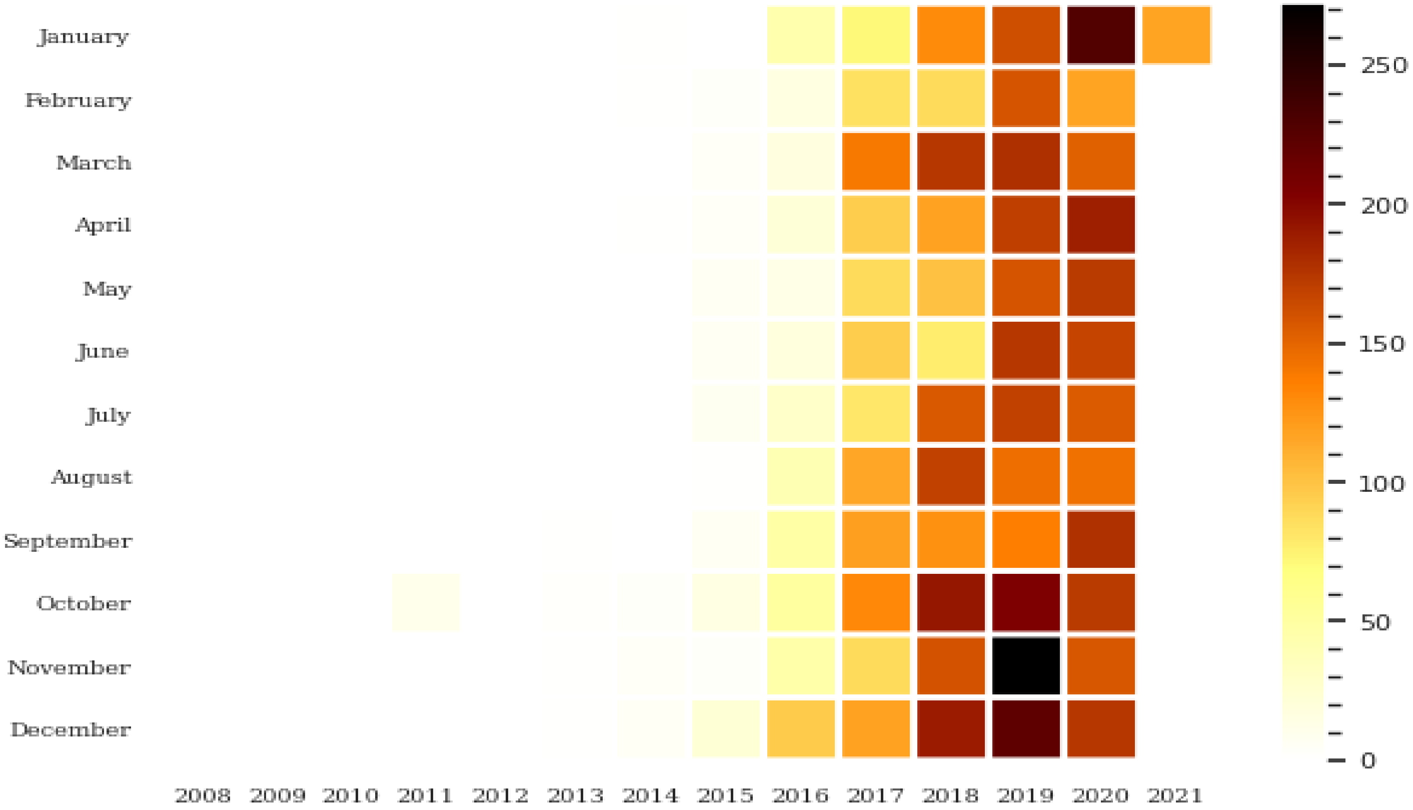

Content release - month wise
Count of content added
0
100
200
300
400
500
600
700
800
1
2
3
4
5
6
7
8
9
10
11
12
Month

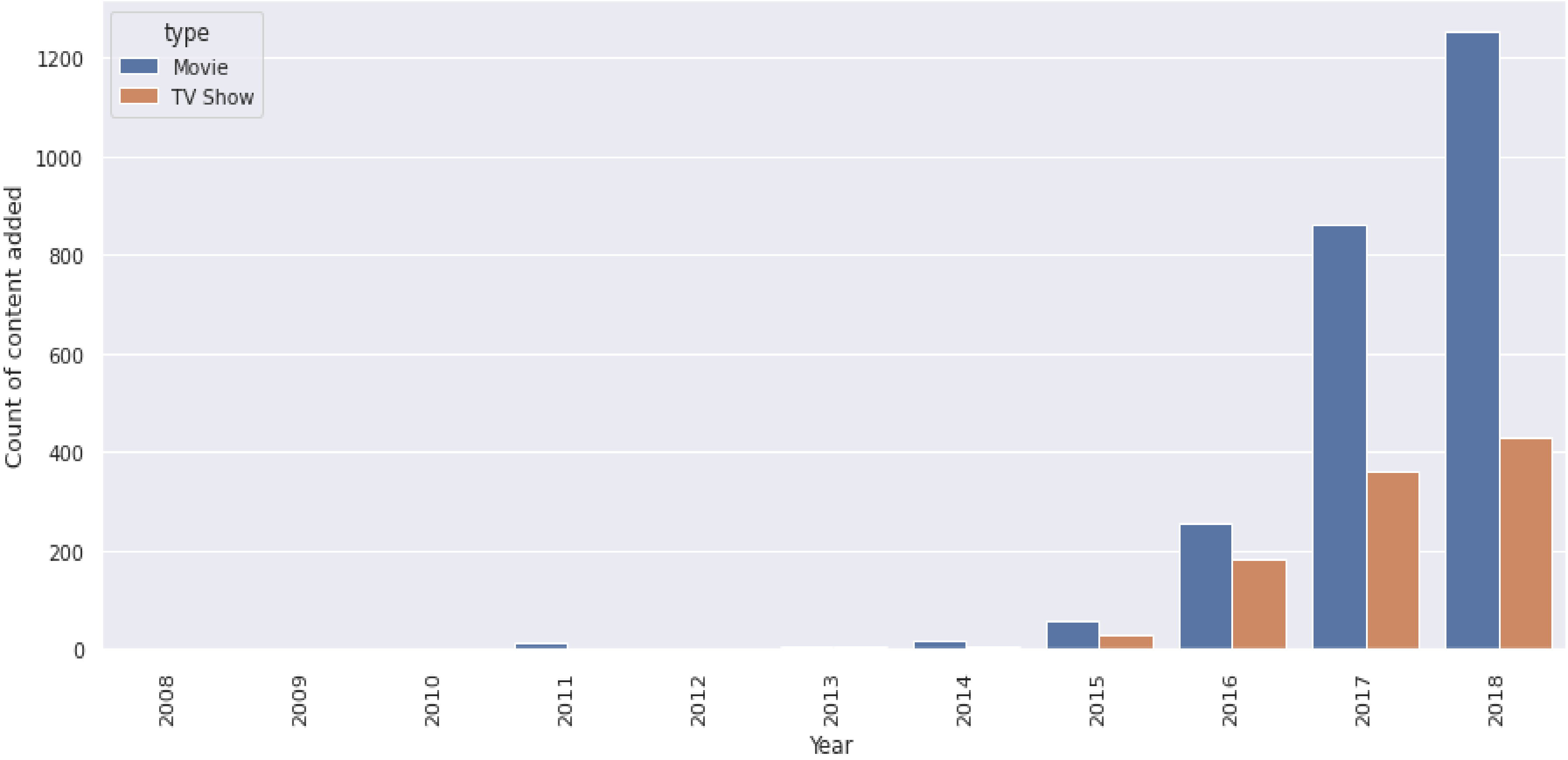

Releases per year
2000
1500
1000
500
0
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
Year

# Trend of year-wise content release



# Netflix Contents Update




type
🟦 Movie
🟧 TV Show
Count of content added
0
200
400
600
800
1000
1200
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
Year

# Text Pre-processing for Clustering

## 1. Removing Punctuation:

- Punctuations does not carry any meaning in clustering.
- So, removing punctuations helps to get rid of unhelpful parts of the data, or noise.

## 2. Removing Stopwords:

- Stopwords are basically a set of commonly used words in any language, not just in English.
- If we remove the words that are very commonly used in a given language, we can focus on the important words instead.

## 3. Stemming:

- Stemming is the process of removing a part of a word, or reducing a word to its stem or root.

- Applying stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language.

# K-Means Clustering

K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre defined distinct non overlapping subgroups where each data point belongs to only one group

1.. **Vectorization:**

- Here we have textual data
- Clustering algorithms cannot understand textual data
- So, we use vectorization technique to convert textual data to numerical vectors.

So,weusevectorizationtechniquetoconverttextualdatatonumericalvectors.

## 2. Elbow Curve:

- The Elbow Curve is one of the most popular methods to determine this optimal value of k.
- The elbow curve uses the sum of squared distance (SSE) to choose an ideal value of k based on the distance between the data points and their assigned clusters.

## 3. Silhouette score :

- Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K Means in terms of how well samples are clustered with other samples that are similar to each other.

# Feature Selection & ML algo used

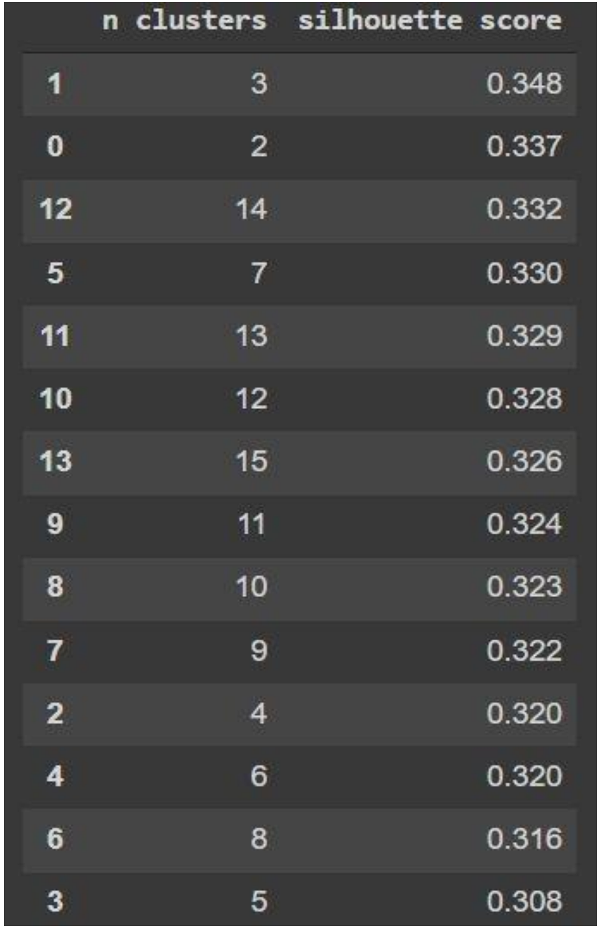
- Only selected 3 features , to do clustering
  - no_of_category
  - Length(description)
  - Length(listed-in)

- Using StandardScaler

- Used 5 algo to find out best k value

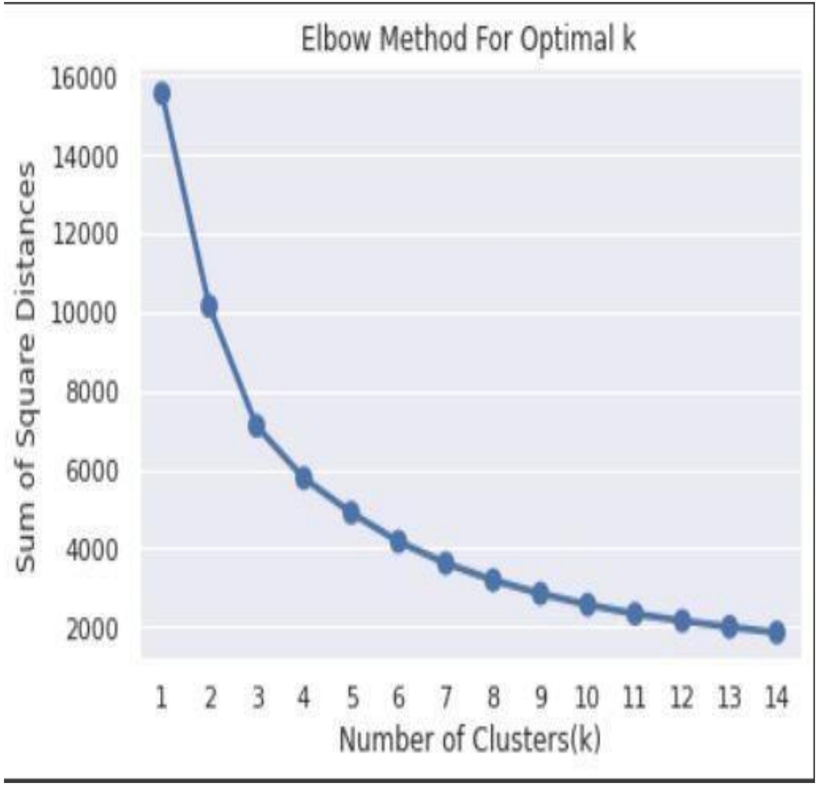
  1. Silhouette score
  2. Elbow Method
  3. DBSCAN
  4. Dendrogram
  5. Agglomerative Clustering

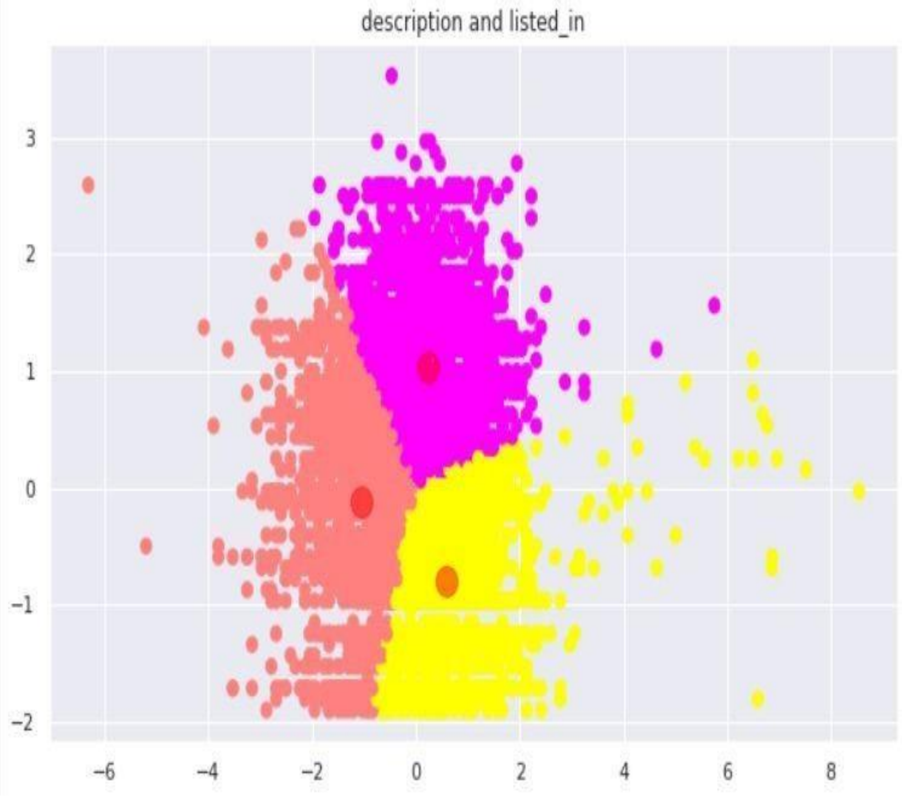
# 1. Silhouette Score



| | 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 |

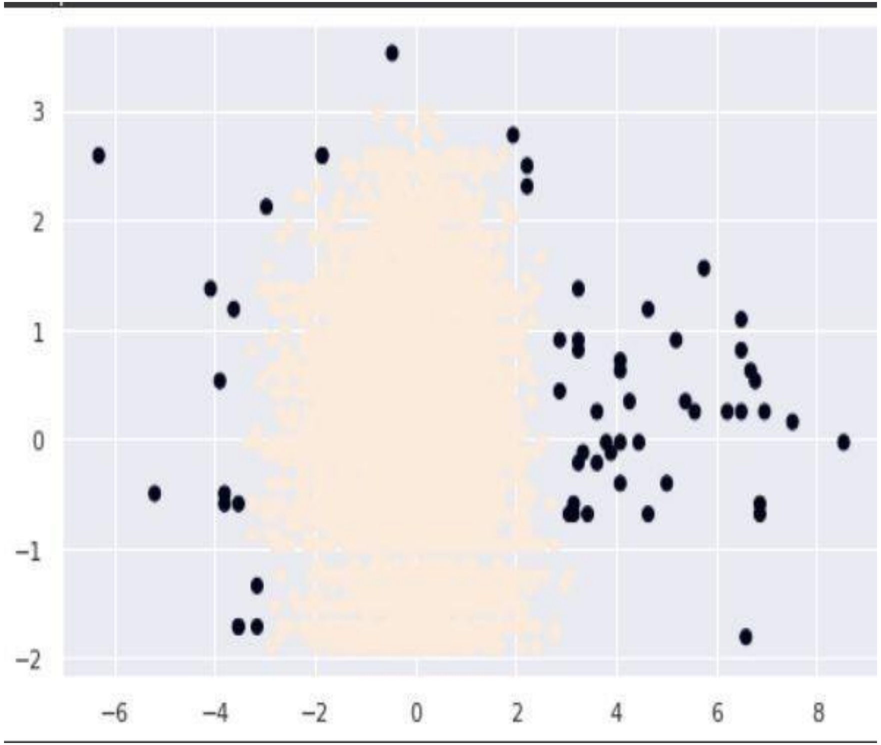
# 2. Elbow Method

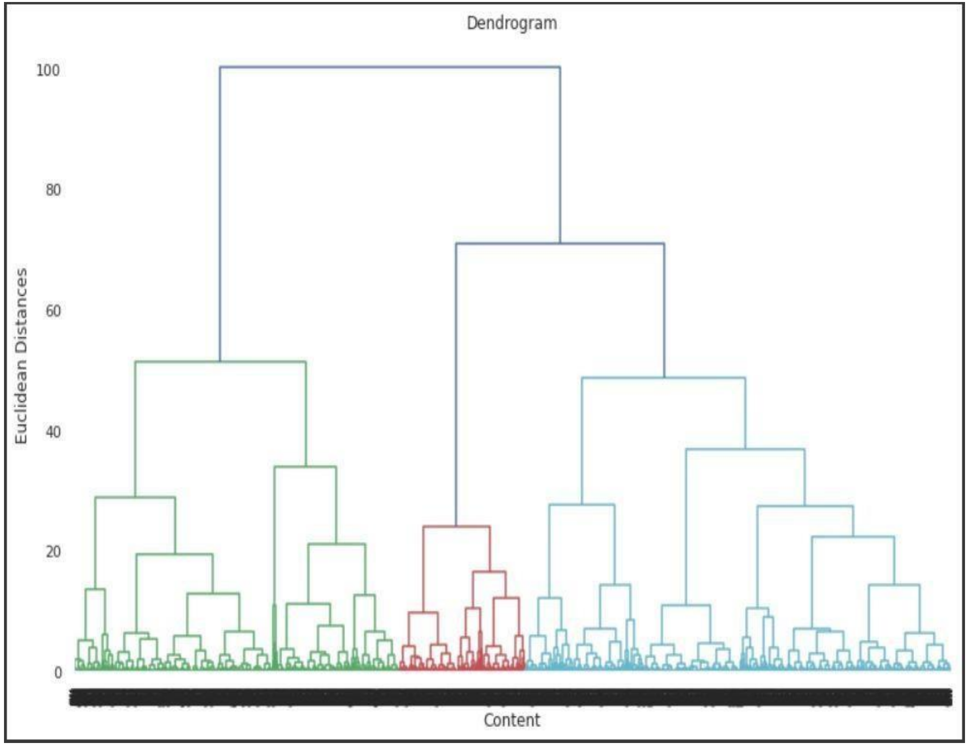




## 3 . DBSCAN

## 4. Dendrogram





*DBSCAN*

*Dendrogram*

# 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 got 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 and TV shows release in 2017 and 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 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.
- Applied different clustering models Kmeans, hierarchical, Agglomerative clustering on data we got the best cluster arrangments.
- By applying different clustering algorithms to our dataset ,we got the optimal number of cluster is equal to 3.

# Thank you