## **Netflix Movies and TV Shows clustering**

**Capstone Project Using Unsupervised Machine Learning**

**Monica Patel**

**(Data science Enthusiast at Alma Better Pro Member Bangalore)**

**Abstract**

Exploring datasets of Netflix for Future Release of TV shows and Movies on the Platform. In this project, we are going to explore the dataset from Kaggle and we would like to and out how long the Netflix platform takes a movie or a TV show to release on its platform, how many movies and TV shows are released in specific time frame, how many movies and TV shows are release in the recent ten years on the platform, and what were the top 10 genres that the audience of the Netflix platform liked the most. From here, we would like to apply a machine learning approach to understand the data fully and provide a great solution where the platform should be headed to. From our data analysis we conducted on R markdown, we have discovered that there were a wide variety of genres that movie directors produced worldwide and we have observed many cast members and genres they were in.

Keywords—Machinelearning,  
Explanatory Data Analysis,

Netflix,

Tvshows,Movies,  
Genre, Clustering, K Means.

**Introduction**  
Unsupervised Learning is a machine learning technique in which the models are not supervised by the training set instead we find hidden patterns and insights from the given data. It is a machine learning technique in which models are trained on  
the unlabeled data set without any supervision. A cluster is a collection of elements that are similar to each other but dissimilar to the elements belonging to other clusters. Clustering can be done using various kinds of distances such as Euclidean distance, Manhattan distance, gomer distance, etc. We can do different kinds of clustering based on the data pattern in space such as spherical clustering, K-means  
clustering, etc.

**ProblemStatement**  
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.

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 increasingly focused on TV rather than movies in recent years?  
4. Clustering similar content by matching text-based features.

Our goal here is to make an unsupervised clustering model, which will help in garnering  
insights on Netflix and how its content is being consumed.

**A brief summary of the dataset is given below:**

**The dataset provided contains 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 Content  
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**

**Tools Used**

The whole project was done using python, in google Collaboratory. Following libraries were used for analyzing the data and visualizing it and to build the model to predict the Netflix clustering

* Pandas: Extensively used to load and wrangle with the dataset.
* Matplotlib: Used for visualization.
* Seaborn: Used for visualization.
* Warnings: For filtering and ignoring the warnings.

**. Clustering:**

Clustering (also called cluster analysis) is a task of grouping similar instances into clusters. More formally, clustering is the task of grouping the population of unlabeled data points into clusters in a way that data points in the same cluster are more similar to each other than to data points in other clusters. The clustering task is probably the most important in unsupervised learning, since it has many applications.

for example:

**• Data analysis:** often a huge dataset contains several large clusters, analyzing which separately, you can come to interesting insights.

**• Anomaly detection:** as we saw before, data points located in the regions of low density can be considered as anomalies

**• Semi-supervised learning:** clustering approaches often helps you to automatically label partially labeled data for classification tasks

**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.**DBSCAN-:**  
DBSCAN is a clustering method that is used in machine learning to separate clusters of high  
density from clusters of low density. Given that DBSCAN is a density based clustering  
algorithm, it does a great job of seeking areas in the data that have a high density of  
observations, versus areas of the data that are not very dense with observations. DBSCAN can  
sort data into clusters of varying shapes as well, another strong advantage

LIBRARIES WHICH IMPORT IN THE PROJECT

import numpy as np                  # helps in working arrays and matrices

import pandas as pd                 # helps in reading dataset/ making dataframes/ manipulating dataframes

import seaborn as sns               # this is for the data visualization for heatmap \countplot\

import matplotlib.pyplot as plt

from matplotlib.ticker import FuncFormatter

%matplotlib inline

import warnings

warnings.filterwarnings("ignore")

import pylab as pl

import spacy

import sklearn

import en\_core\_web\_sm

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import missingno as msno

import matplotlib.cm as cm

from os import path

from PIL import Image

from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator

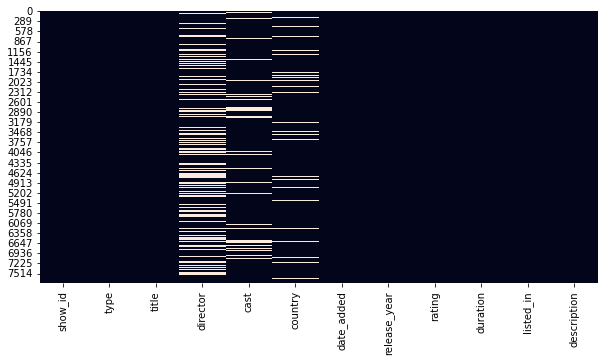
Exploratory Data Analysis The first step involved in the analysis is to load the dataset into  
the pandas data frame. Before exploring the data using different libraries available in python we should if the dataset is ready to run the operations on it.  
Data Cleaning: Data Cleaning is one of the important steps before we start building models, infact, there will be asignificant increase in Model Performance when we have a clean, rich  
dataset. So here, we decided to replace null values with an empty string.  
● There are 2389 null values in Director column  
● There are 718 null values in cast column  
● There are 507 null values in country column  
● There are 10 null values in date added column  
● There are 7 null values in rating column

DATA ANALYSIS AS A SUBJECTIVE MATTER:  
A. Requirements Gathering The data is the necessary requirement for providing inputs to any type of analysis. It can be based on the requirements and parameters based on the user. This data can be either numerical or categorical. The purpose and scope can range from supervised to unsupervised learning.  
B. Data Collection  
The required data can be collected from a wide range of sources. It can be structured based on the criteria provided by analysts to custodians of a particular data set. The data can be man-made or in the form of technological output over utility (sensor system tracking) and many other implications.  
C. Data Processing & Cleaning  
The data when purported for utilization must be processed on the level where the needs for analysis are satisfied. This includes placing data in the form of rows and columns that are  
human-understandable in nature. Further, it must be cleaned for getting rid of any redundant data, or minimalize the presence of anomalies prior to the deployment of the data for analysis. The figure given below explains these four fundamental steps in a lucrative pictogram.

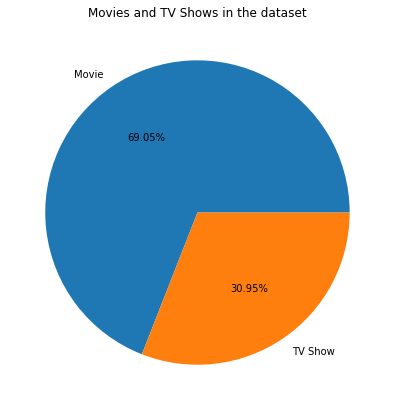
BELOW ARE THE VISUALIZATION CHARTS .

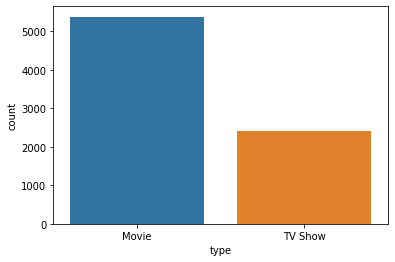
plt.figure(figsize=(10,5))

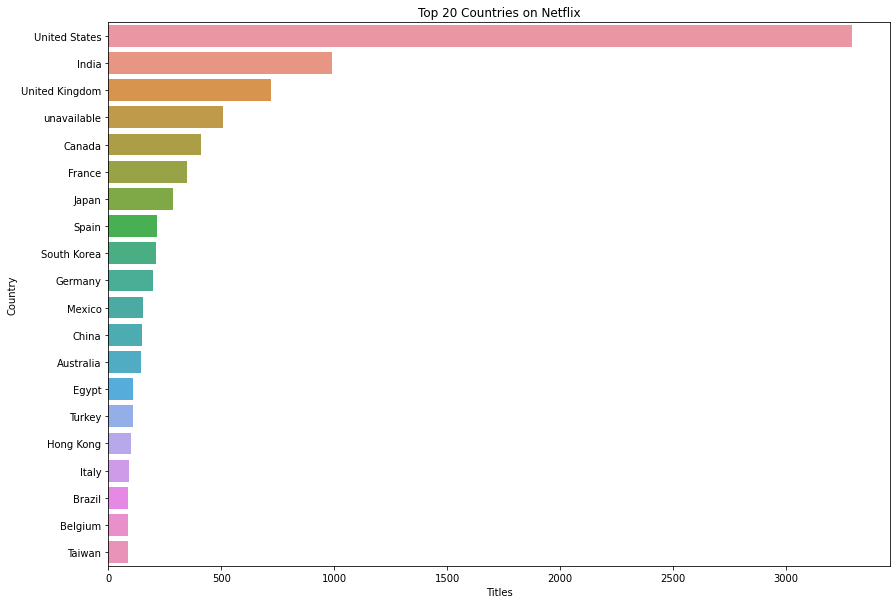
sns.heatmap(netflix\_df.isnull(), cbar=False)

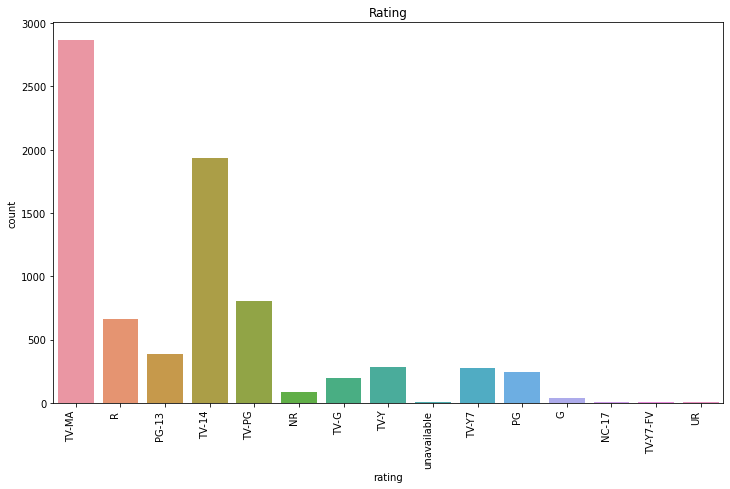


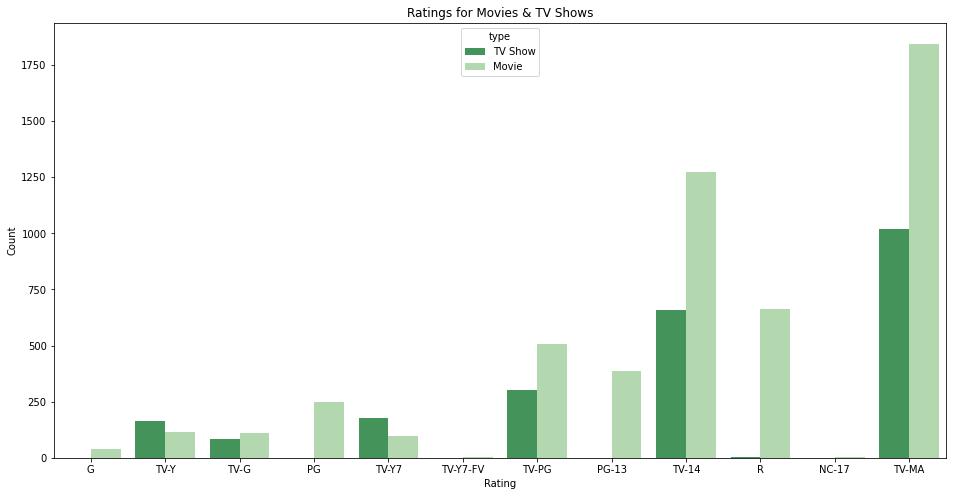
Netflix has 69% of its content as movies  
Movies are clearly more popular on Netflix than TV shows.

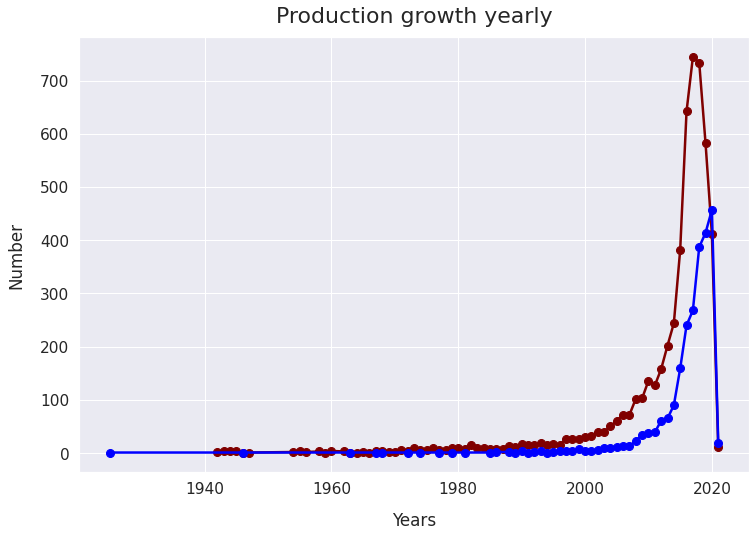


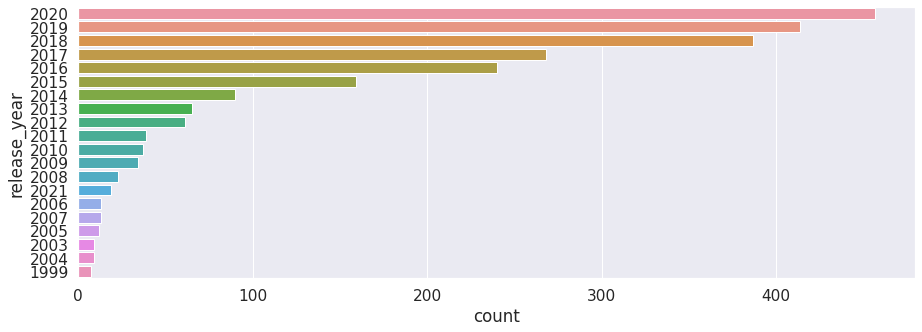
****

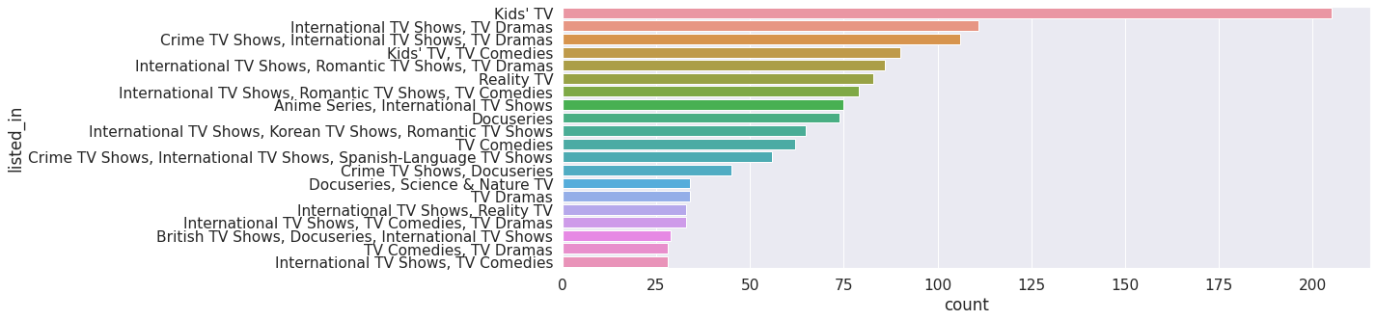
****

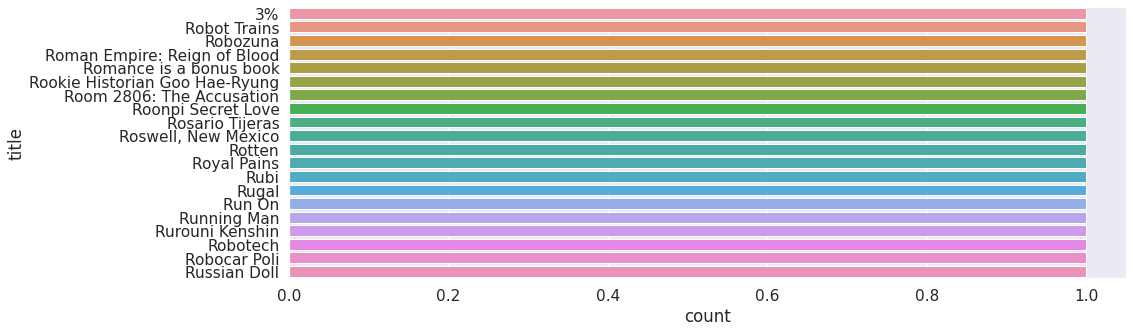
****

****

****

****

****

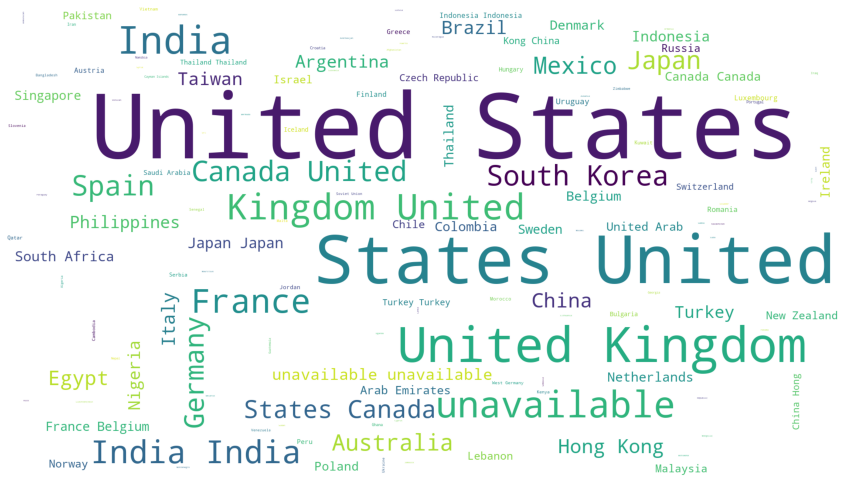
****

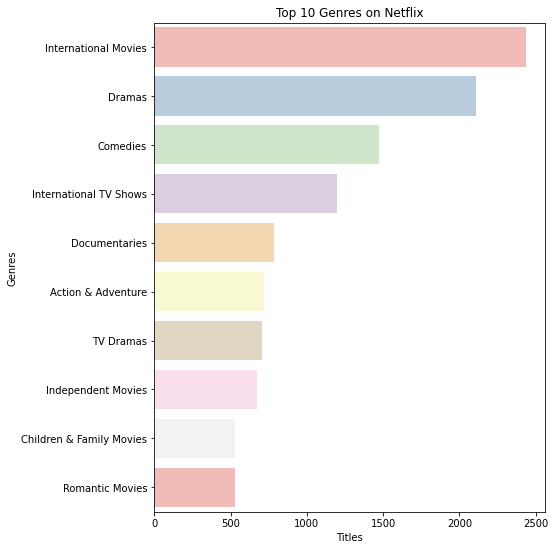
**Common titles**

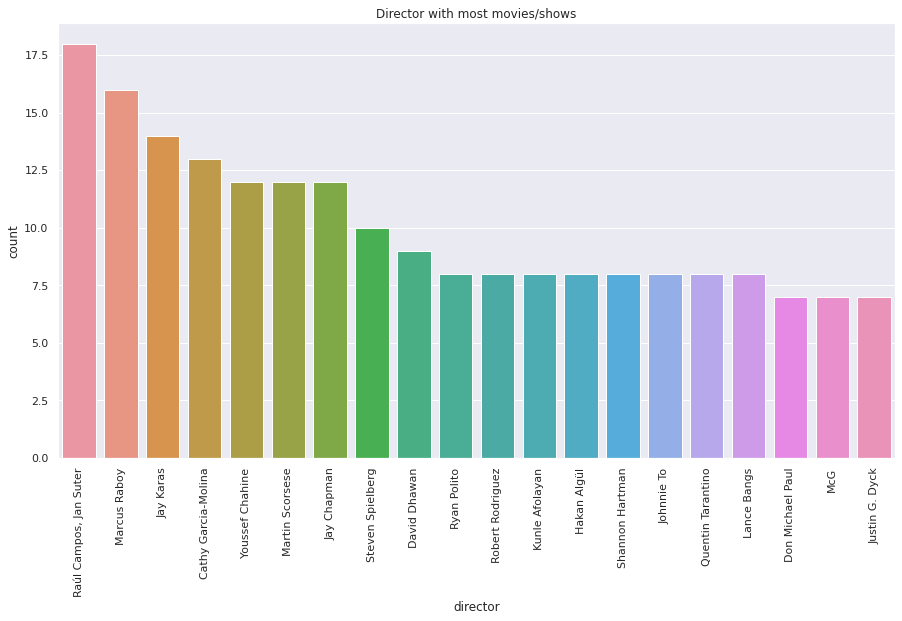
It seems like words like  
"Love", "Man", "World",  
"Story" are very common in  
titles

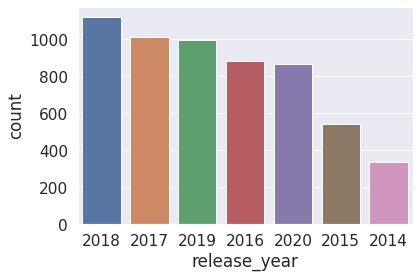
Cluster description charts

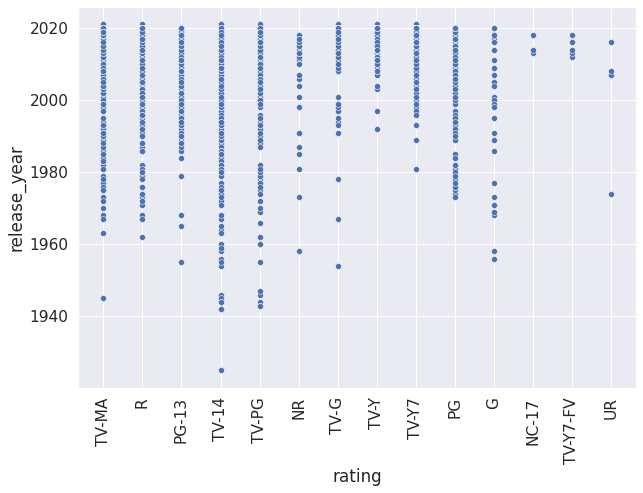
****

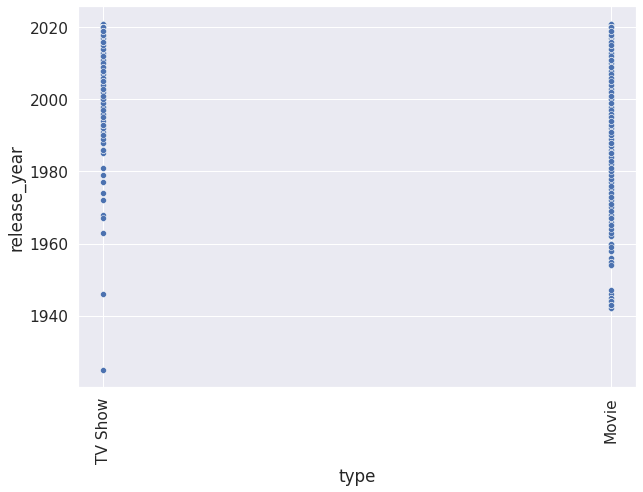
****

****

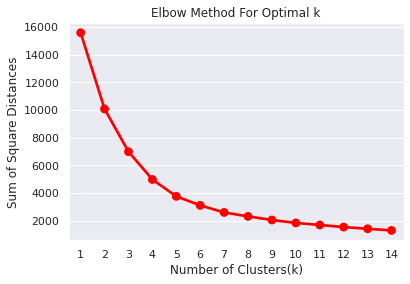
****

****

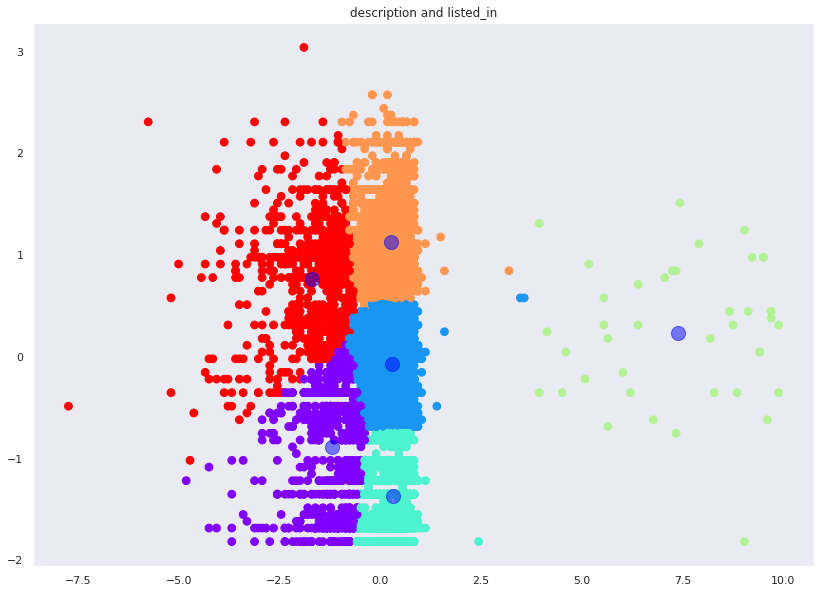
****

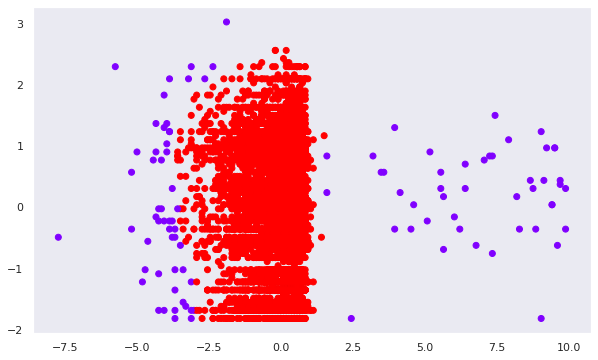
****

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

****

**DBSCAN-:**  
DBSCAN is a clustering method that is used in machine learning to separate clusters of high  
density from clusters of low density. Given that DBSCAN is a density based clustering  
algorithm, it does a great job of seeking areas in the data that have a high density of  
observations, versus areas of the data that are not very dense with observations. DBSCAN can  
sort data into clusters of varying shapes as well, another strong advantage

****

****

**Conclusion**

Tailored recommendations can be made based on information about movies and TV shows. In addition, similar models can be developed to provide valuable recommendations to consumers in other domains.

* We've done null value treatment, feature engineering, and EDA since loading the dataset then completed assigned tasks.
* 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)
* Most films were released in the years 2018, 2019, and 2020 and and united nation have the maximum content on Netflix.
* .In terms of genres, Dramas is on the top followed by Comedies and Documentaries.
* Number of movies added to netflix is higher than that of TV shows. In 2019, netflix added1497 movies and 656 TV shows.
* So there we cannot conclude that Netflix has switched focus  
  from movies to TV shows.
* .Principal component analysis was performed inorder to reduce the higher dimensionality  
  which improved the silhouette coefficient to 0.34118.
* .Clusters are identified for each of the records in the dataset.
* Recommendation based on cosine similarity is done
* **Reference  
  [1] Applied Science  
  Article MDPI  
  [2] GeeksforGeeks  
  [3] Wikipedia  
  [4] DataCamp**