NETFLIX MOVIES AND TV SHOWS CLUSTERING

by

Ajinkya Shingote

.

ABSTRACT
Netflix is a company that manages a large collection of TV shows and movies, streaming it anytime via online. This business is profitable because users make a monthly payment to access the platform. However, customers can cancel their subscriptions at any time. Therefore, the company must keep the users hooked on the platform and not lose their interest. This is where recommendation systems start to play an important role, providing valuable suggestions to users is essential.

1. Introduction

Netflix's recommendation system helps them Increase their popularity among service providers as they help increase the number of items sold, offer a diverse selection of items, increase user satisfaction, as well as user loyalty to the company, and they are very helpful in getting a better understanding of what the user wants. Then it's easier to get the user to make better decisions from a wide variety of movie products. With over 139 million paid subscribers (total viewer pool -300 million) across 190 countries, 15,400 titles across its regional libraries and 112 Emmy Award Nominations in 2018 — Netflix is the world's leading Internet television network and the most-valued largest streaming service in the world. The amazing digital success story of Netflix is incomplete without the mention of its recommender systems that focus on personalization. There are several methods to create a list of recommendations according to your preferences. You can use (Collaborative-filtering) and(Content-based Filtering) for recommendation.

2. 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.

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.

3. Objective

Netflix Recommender recommends Netflix movies and TV shows based on a user's favourite movie or TV show. It uses a Natural Language Processing (NLP) model and a K-Means Clustering model to make these recommendations. These models use information about movies and TV shows such as their plot descriptions and genres to make suggestions. The motivation behind this project is to develop a deeper understanding of recommender systems and create a model that can perform Clustering on comparable material by matching text-based attributes. Specifically, thinking about how Netflix create algorithms to tailor content based on user interests and behavior

4. Approach

As the problem statement says, understanding what type of content is available in different countries and Is Netflix increasingly focused on TV rather than movies in recent years we have to do clustering on similar content by matching text-based features. For that I used Affinity Propagation, Agglomerative Clustering.

5. 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

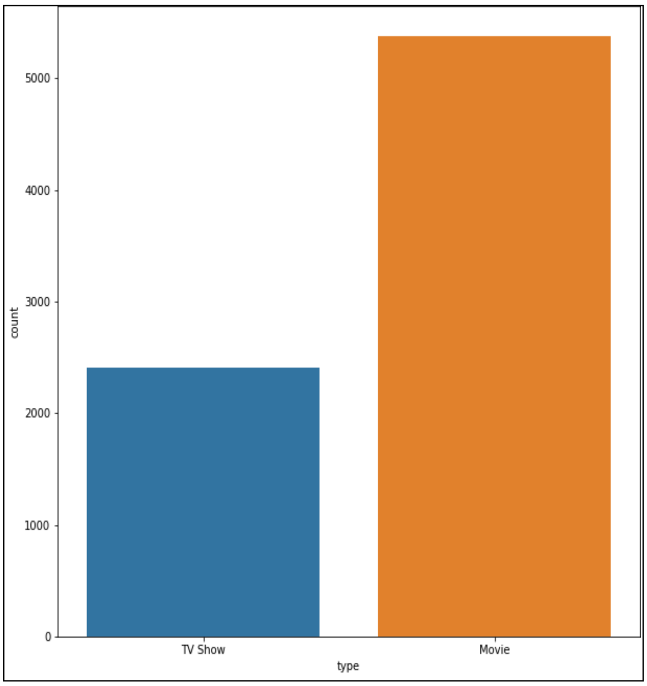
- Pandas: Extensively used to load and wrangle with the dataset.
- Matplotlib: Used for visualization.
- Seaborn: Used for visualization.
- Nltk: It is a toolkit build for working with NLP.
- Datetime: Used for analyzing the date variable.
- Warnings: For filtering and ignoring the warnings.
- NumPy: For some math operations in predictions.
- Word cloud: Visual representation of text data.
- Sklearn: For the purpose of analysis and prediction

6.Exploratory Data Analysis:

Exploratory Data Analysis (EDA) as the name suggests, is used to analyze and investigate datasets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions. It also helps to understand the relationship between the variables (if any) and it will be useful for feature engineering. It helps to understand data well before making any assumptions, to identify obvious errors, as well as better understand patterns within data, detect outliers, anomalous events, find interesting relations among the variables but Before EDA missing values are handled and duplicate values are treated.

Univariate analysis before clustering

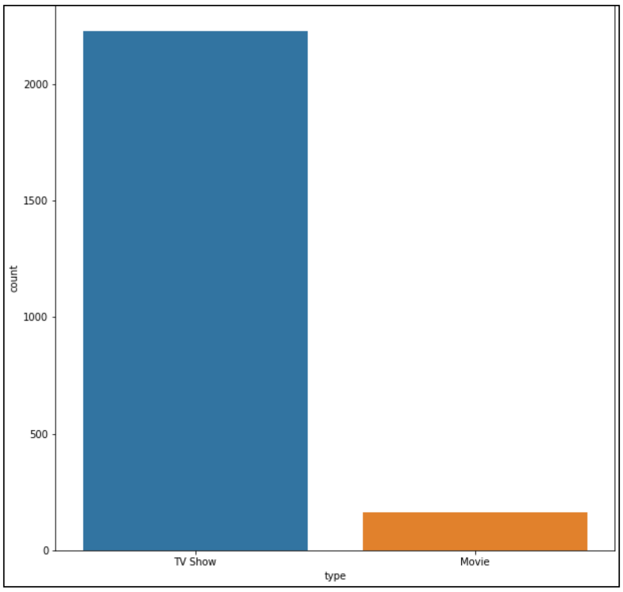
6.1 Type column:



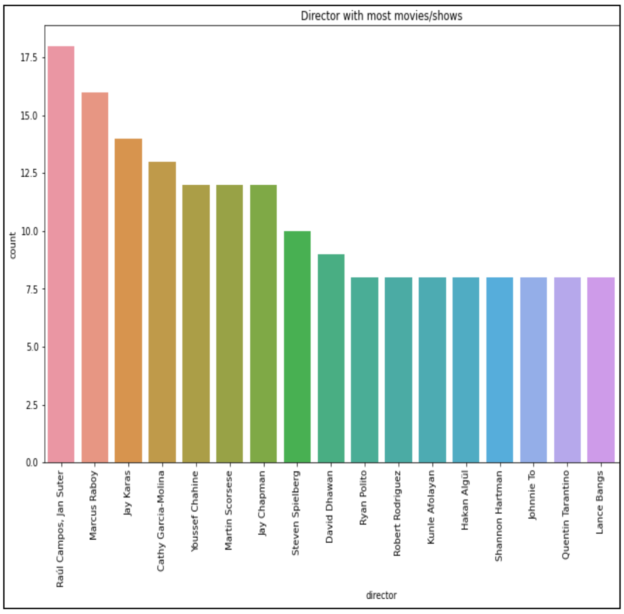
The above figure is count plot of movie and tv shows and the from the visualization we can draw the conclusion that almost 70% of datapoints belong to movie, rest 30% to TV show.

6.2 Director:

The below figure is count plot of null values of director in TV shows and movies, from the visualization we can draw the conclusion that most of the missing columns of director are for TV shows.



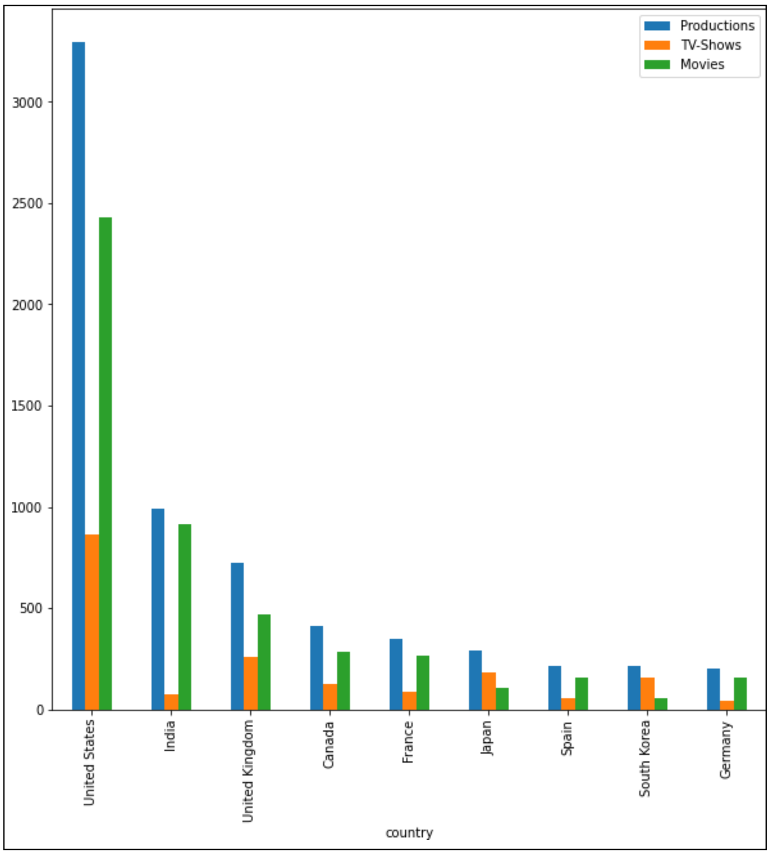
6.3 Top 20 directors



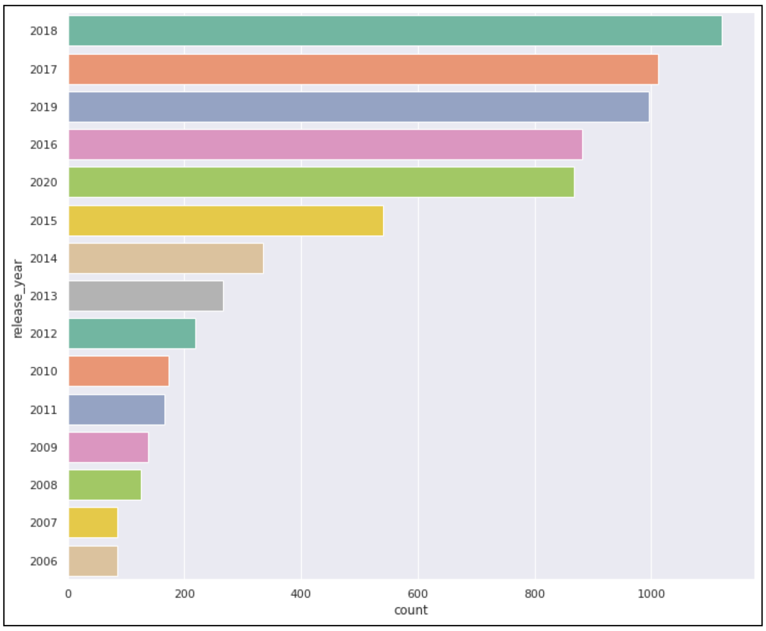
The above figure is a bar plot of count vs director which tells about director with the greatest number of movies or tv shows. The visualization depicts that Raul Campos, Jan Suter are the director with most tv shows or movies followed by Marcus Raboy.

6.4 Top countries with more no of productions:

The top countries with Tv shows and movies along with production is plotted and the from visualization it is clear that United States tops the chart followed by India and Germany has the least TV shows and movies with a smaller number of productions.

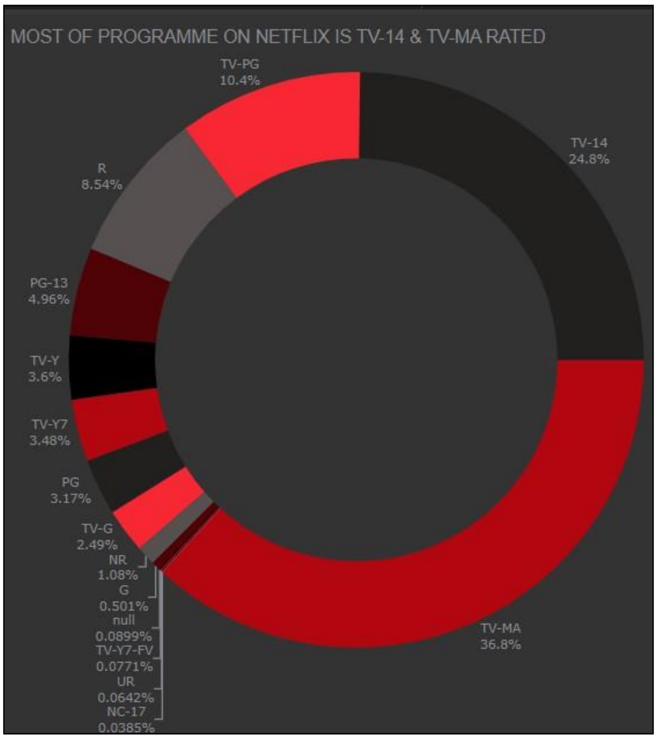


6.5 Count plot for release over years



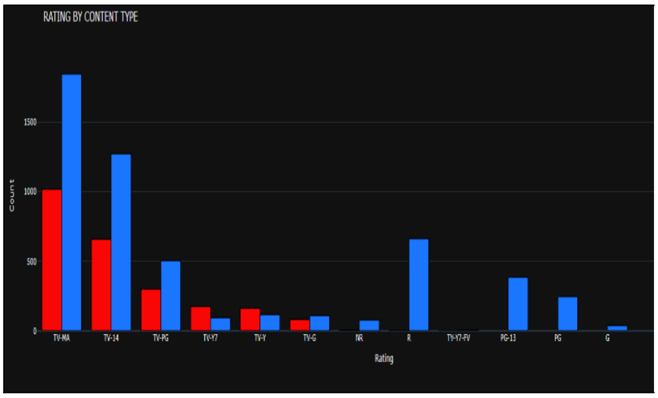
From the count plot it is clear that year 2018 is the year with the greatest number of movies and TV shows with an approximate count value of 1400 and the least movies and TV shows in the year 2006.

6.6 Rating column distribution by Pie chart



From the pie chart it is clear that most of the programs on Netflix are TV-14 and TV-MA rated we can say that more content with mature content is available on Netflix.

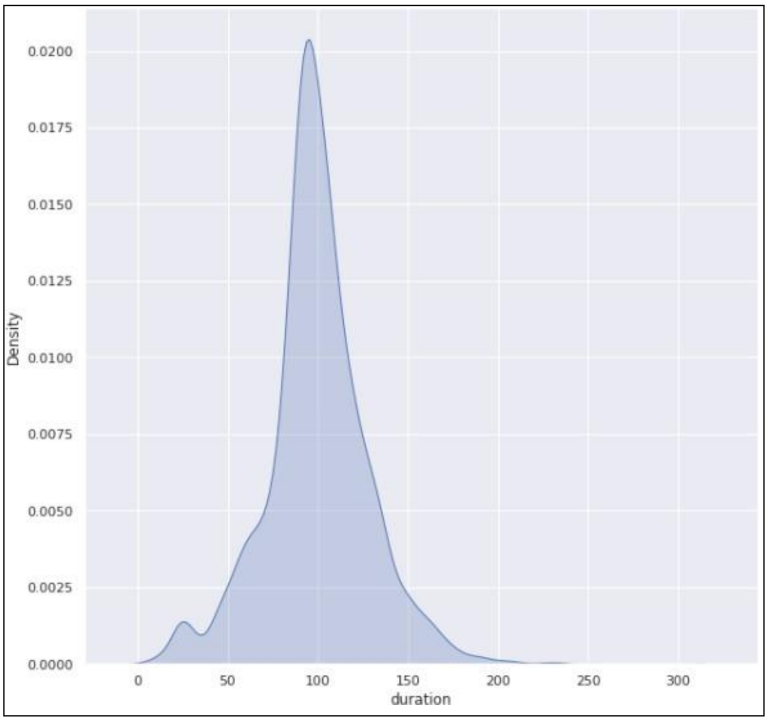
6.7 Rating distribution by content type



The above plot depicts rating distribution with content type and TV-MA and TV-14 tops the chart more number of movies than TV shows.

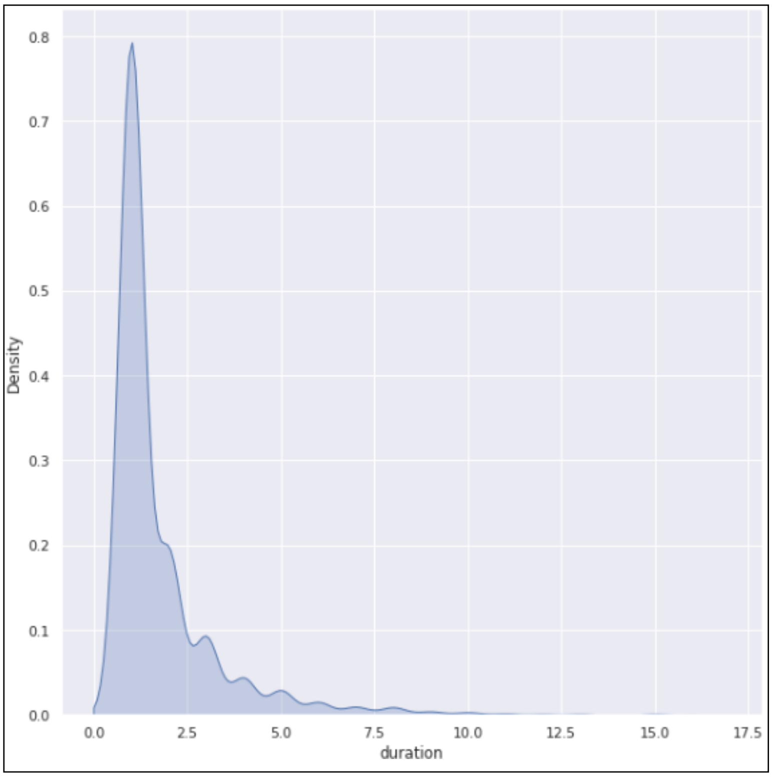
6.8 Netflix movie duration distribution

The below plot is a density plot for duration of movies and from the plot it is clear that most of the content is about 70 to 120 minutes duration for movies.

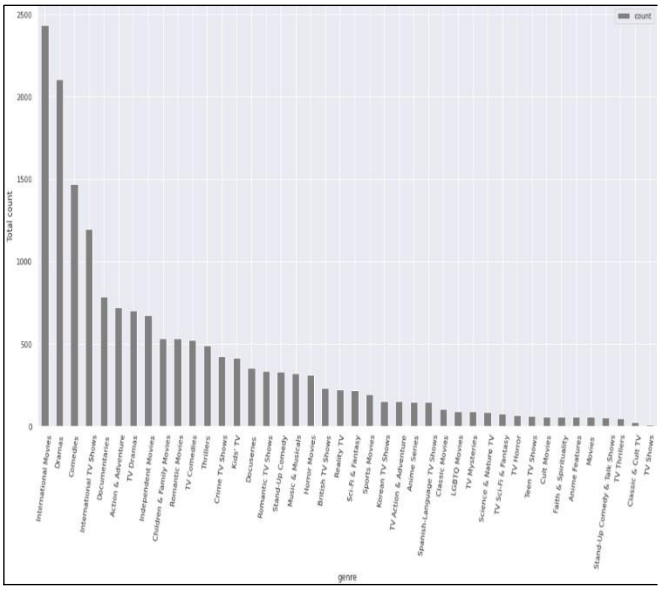


6.9 Netflix show duration distribution

The below plot is a density plot for duration of show for no of seasons and from the plot it is clear that most of the shows are 1 to 2 seasons long.

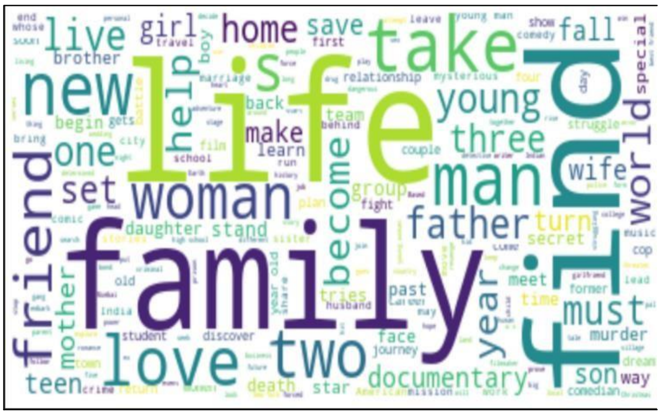


6.10 Genre wise count distribution



From the above count plot for genre distribution, it is pretty evident that 'international movies' is the genre with highest count followed by 'Dramas' and the least is tv shows.

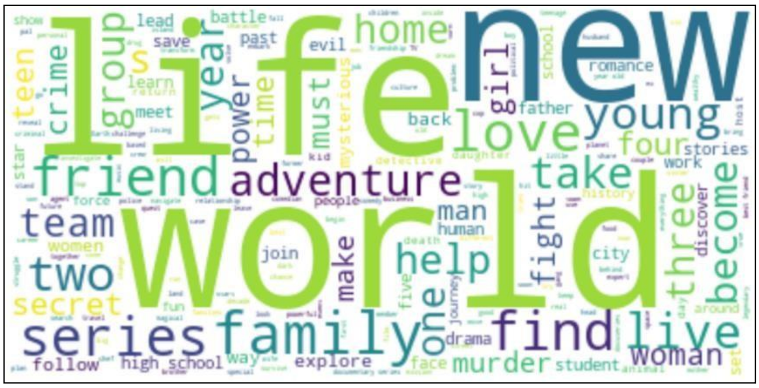
6.11Word cloud for movie on description column



From the word cloud for movies most words like life, family popped up.

6.12Word cloud for shows on description column

From the word cloud for shows most words like 'life','world','new','adventure','friend','family' popped up.



Apart from univariate analysis bivariate analysis such as

- Country vs Genre
- Country vs rating
- Country vs type
- Country vs year added
- Country vs top directors
- Country vs cast
- Country vs release year

7. Hypothesis from the data visualized

Hypothesis testing is done to confirm our observation about a population using sample data within a desired level of error. Through hypothesis testing, we can determine whether we have enough statistical evidence to conclude whether a hypothesis about a population is true or not. We conducted hypothesis testing to gain insight into the duration of movies and content with respect to various variables.

8.TFIDF vectorization:

TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is a very common algorithm to transform text into a meaningful representation of numbers which is used to fit a machine learning algorithm for prediction. I used this to be apply the data into Sigmoid Kernel so as to find the most similar movies. However, I used

Count vectorizer that gave better results for clustering the given Dataset.

9. Data preprocessing:

Removing punctuation: Punctuation has no meaning in clustering, so removing punctuation helps to get rid of useless bits of data or noise.

Removing stop words: Stop-words are basically a collection of commonly used words in any language, not just English. If we remove words that are very commonly used in a given language, we can focus on important words instead.

Lemmatization: It is the process of grouping together the inflected forms of a word so they can be analyzed as a single item with the intended meaning.

10. 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.

• Indirectly clustering tasks (tasks where clustering helps to gain good results): recommender systems, search engines, etc.

• **Directly clustering tasks**: customer segmentation, image segmentation, etc.

Building a clustering model

Clustering models allow you to categorize records into a certain number of clusters. This can help you identify natural groups in your data.

Clustering models focus on identifying groups of similar records and labeling the records according to the group to which they belong. This is done without the benefit of prior knowledge about the groups and their characteristics. In fact, you may not even know exactly how many groups to look for.

This is what distinguishes clustering models from the other machine-learning techniques—there is no predefined output or target field for the model to predict.

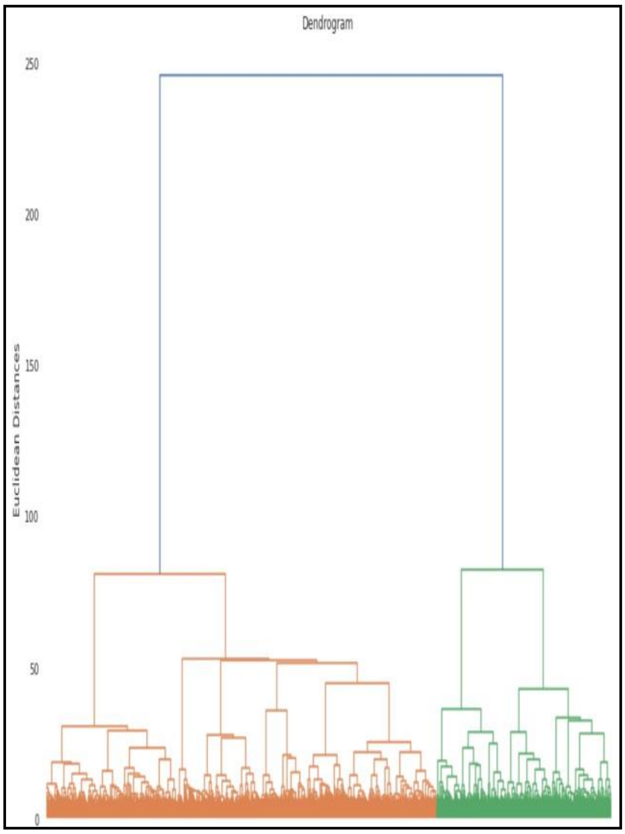
These models are often referred to as **unsupervised learning** models, since there is no external standard by which to judge the model's classification performance.

11. Cluster model implementation

11.1 Agglomerative Clustering

Also known as bottom-up approach or hierarchical agglomerative clustering (HAC). A structure that is more informative than the unstructured set of clusters returned by flat clustering. This clustering algorithm does not require us to prespecify the number of clusters. Bottom-up algorithms treat each data as a singleton cluster at the outset and then successively agglomerates pairs of clusters until all clusters have been merged into a single cluster that contains all data.

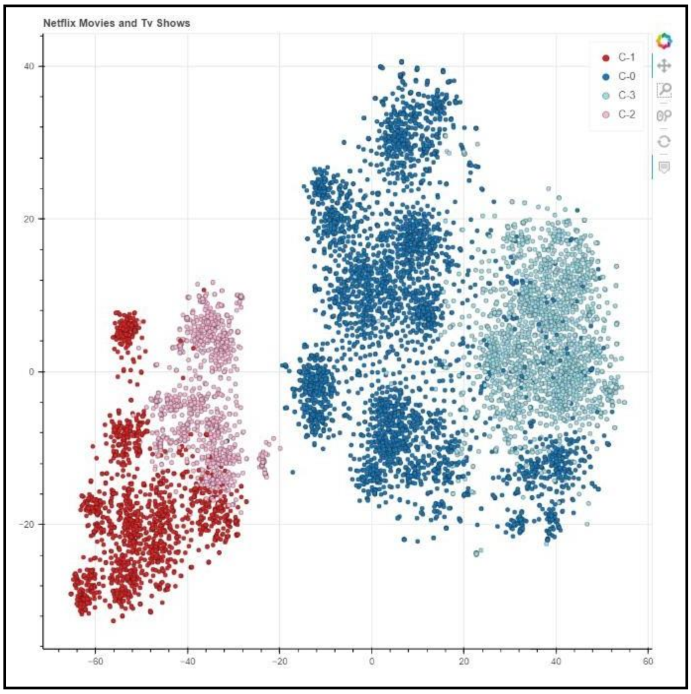
The agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. ... Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects. I used this approach to cluster our data. It was analyzed that based on the dendrogram figure.



We used this approach to cluster our data. It was analyzed that based on the dendrogram figure.

11.2 Bokeh visualization

Visualization is absolutely essential in data analysis, as it allows you to directly feed your data into a powerful neural network for unsupervised learning: your brain. It will allow you to find features and issues in your dataset. That's where interactivity is a must but bokeh will bring us a whole new set of possibilities. For example, for truly interactive plotting, and it can display big data. We can even set up a bokeh server to display data continuously in a dashboard, while it's being recorded.



Conclusion:

1. Exploratory Data Analysis was done for all the attributes to study the deep insights from the Given Dataset
2. Univariate & multivariate analysis
3. Visualized Data, inferred insights
4. Analysed various trends in Countries and the corresponding analysis was visualized to get a clear picture of the analysis.
5. TV shows or Movies? Of course, over the period of time the trend has been moving towards Netflix series instead of movies. I tried to analyze this with graphical representation as well on yearly basis.
6. I used TFIDF vectorizer and sigmoid kernel in order to recommend movies based on the similarities in the textual attributes.
7. Identified 4 distinct clusters and used visualizations to dive deeper into the clustered data.