NETFLIX MOVIES AND TV SHOWS CLUSTERING

Presented By
Tanay Tupe

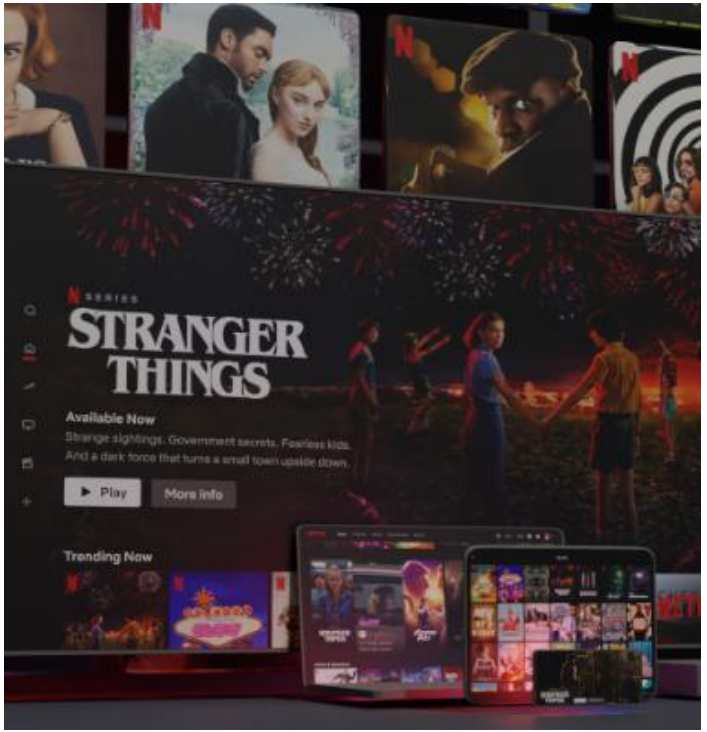
Contents

* **Business domain**
* **Project Problem Description**
* **Dataset Summery**
* **Understanding Each Variable**
* **Modules in project**
    1. Importing necessary libraries, Loading and understanding data
    2. Exploratory Data Analysis
    3. Preprocessing and Feature Engineering
    4. Clustering on Movies and TV Shows using K-means
* **Conclusion**

Business Domain

Netflix is an American subscription streaming service and production company. Launched on August 29, 1997, it offers a film and television series library through distribution deals as well as its own productions, known as Netflix Originals. As of December 31, 2021, Netflix had over 221.8 million subscribers worldwide. The company is ranked 115th on the Fortune 500

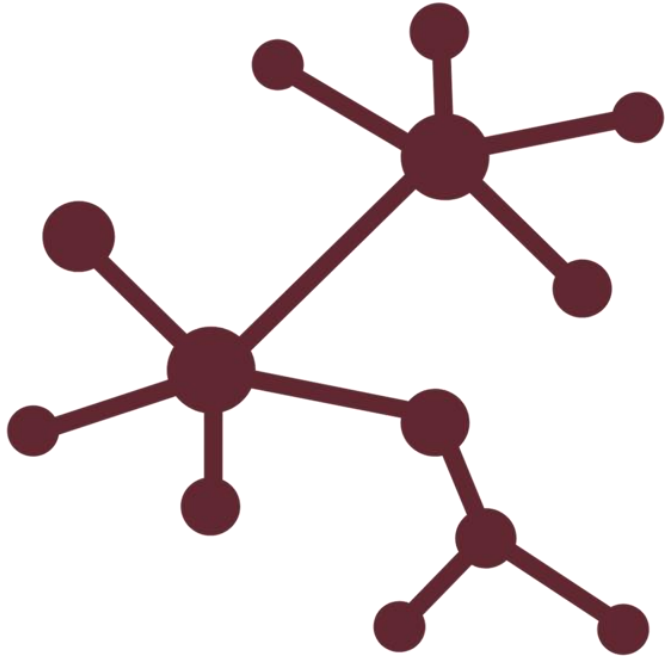
Netflix can be accessed via internet browser on computers, or via application software installed on smart TV.



Project Problem Description

In the competitive entertainment market companies want to understand consumer's viewing pattern of the content available. To understand the type of content watched by the consumer and to effectively predict what might they be interested in next.

We are trying to find out some relation or similarity between shows available on type of their content. In this problem, Our goal is to Categorize all listings from Netflix Dataset and cluster similar content together.



Dataset Summery

The dataset we have used here is “**NETFLIX MOVIES AND TV SHOWS**”.This dataset in it’s raw form has around “7787” observations and “12” columns and it’s a mix between categorical and numeric values.

Also, we have few null values and no duplicated observations throughout our dataset. Target variable that we will be categorizing on is “Listed_in” representing different genres of content displayed.

Understanding Each Variable

Let’s Get Familiar With Each feature Present In The DataSet.

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

Understanding Each Variable (Continued)

8. **release_year** - Actual Releaseyear 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

Module 1 Importing necessary libraries, Loading and understanding data.

Before anything, we'll be needing all relevant libraries as required in our analysis. Few universal libraries like Numpy, Pandas, Matplotlib, Seaborn are must needed, and are imported even before anything, but we also will be needing other libraries as we go through our analysis.

```
# importing all relevant libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.model_selection import GridSearchCV
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA

from nltk.stem.snowball import SnowballStemmer
from nltk.tokenize import RegexpTokenizer
import nltk
from nltk.corpus import stopwords

from collections import Counter
from wordcloud import WordCloud
import re
import warnings
warnings.filterwarnings('ignore')
```

Module 1 Importing necessary libraries, Loading and understanding data.

As this analysis was performed on “Google Colab” we first need to mount the notebook to our drive, where we have our “**NETFLIX MOVIES AND TV SHOWS CLUSTERING**” dataset in “CSV” format.

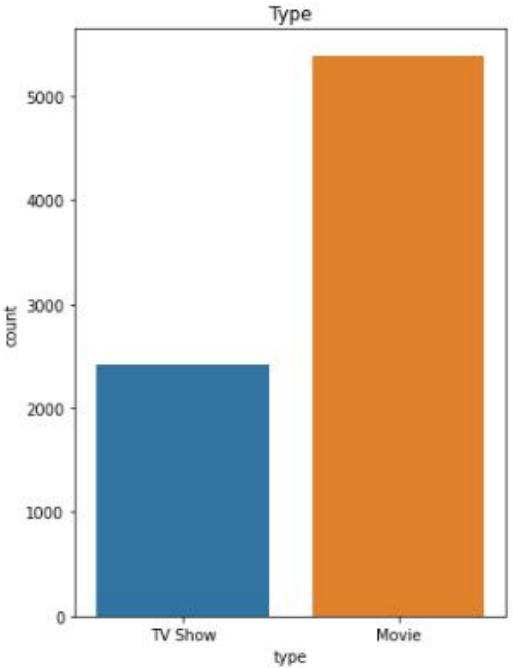
Checks were performed on dataset’s Size, Shape, general information on dataset with info() and describe() function. Where we observed number of rows(7787), features (12), the count of their null values, memory(730.2 KB) the dataset is occupying.

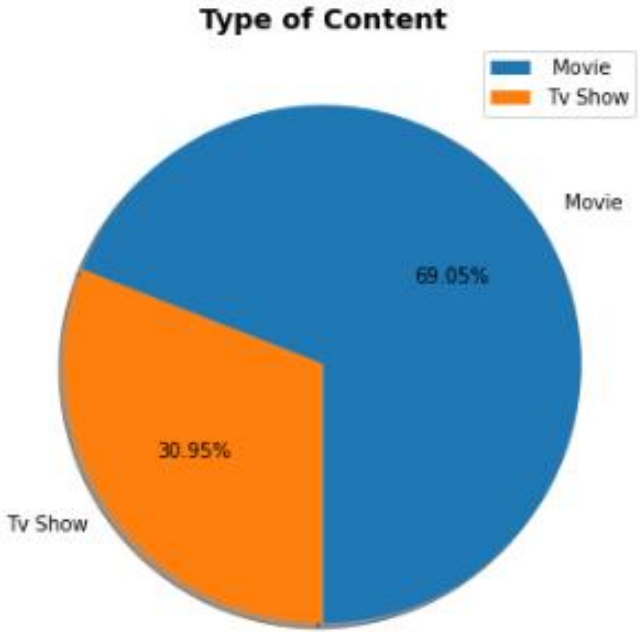
As our goal is to cluster listings on genres, the feature “Listed in” does not have any null values. This already works in our favour, but we also don’t have to check for any kind of outliers in feature like these.

we’ll be using K-means Clustering for this analysis and check for different number of clusters.

Module 2 Exploratory Data Analysis.

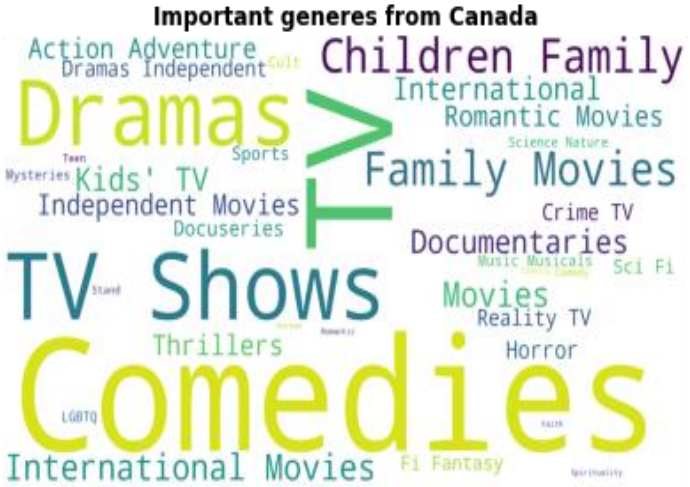
We began our EDA with analyzing difference in TV Shows and Movies. On which type of content Netflix focuses on.





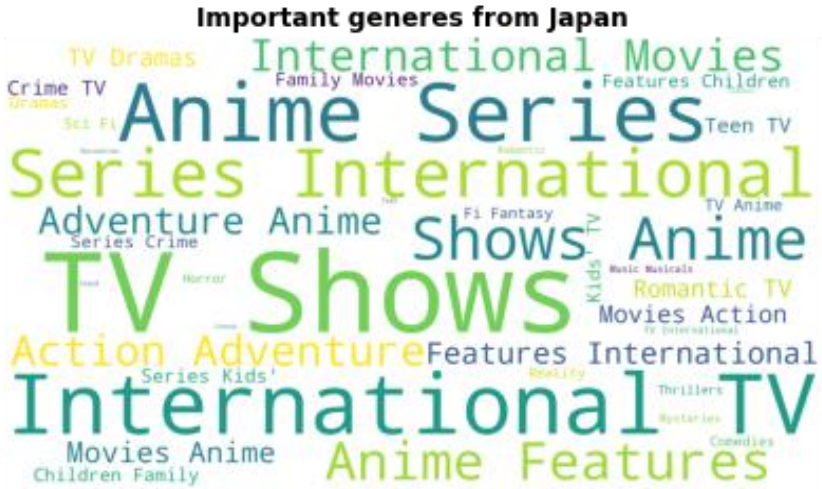
Module 2 Exploratory Data Analysis.

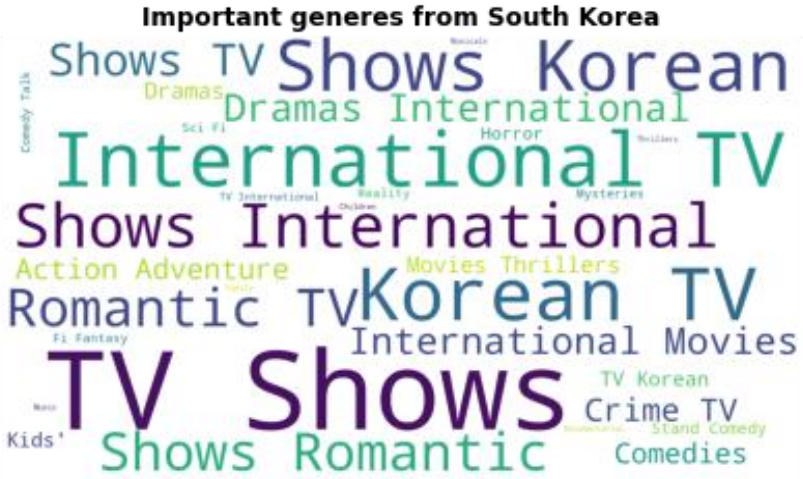
Type content available in different countries



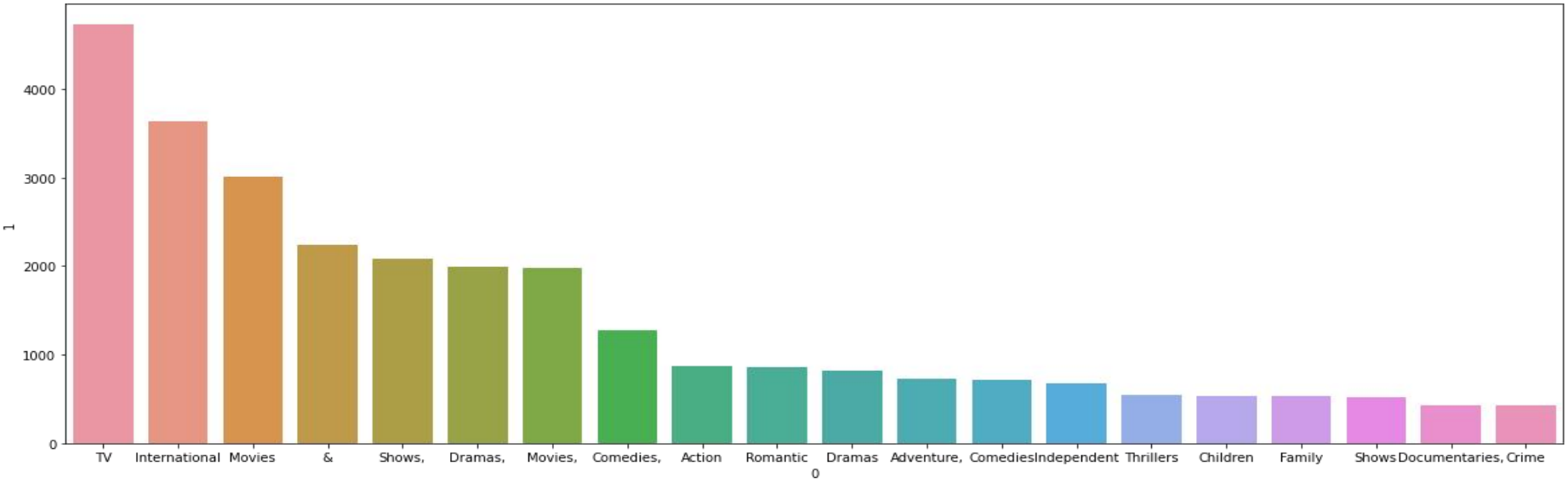
Module 2 Exploratory Data Analysis.

Type content available in different countries





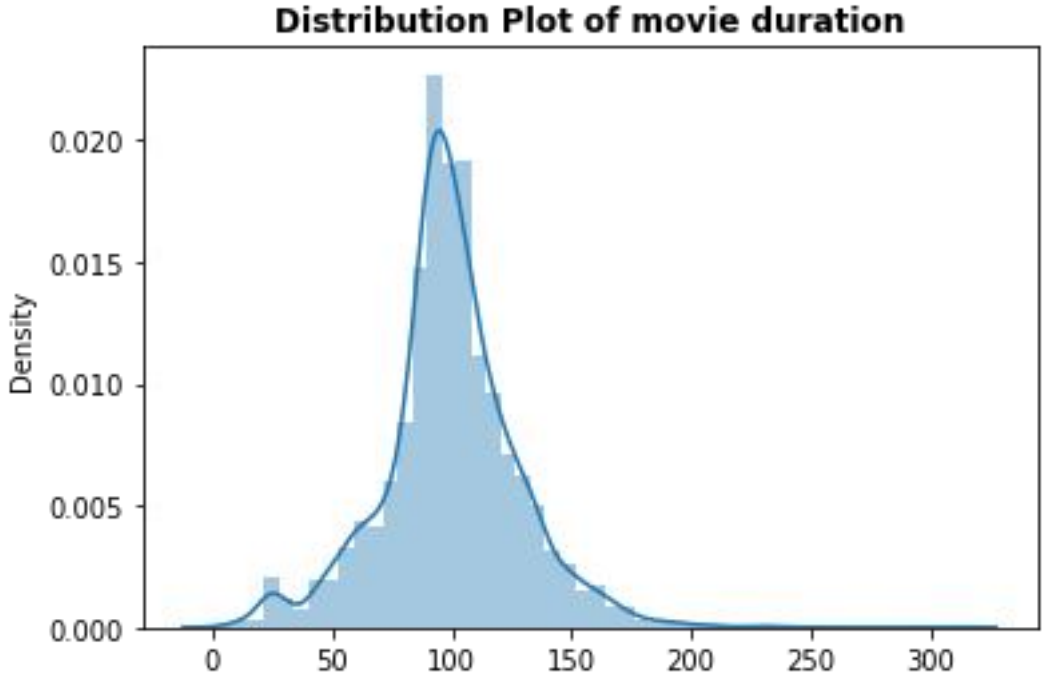
Module 2 Exploratory Data Analysis.



Top words occurred in listed_in feature

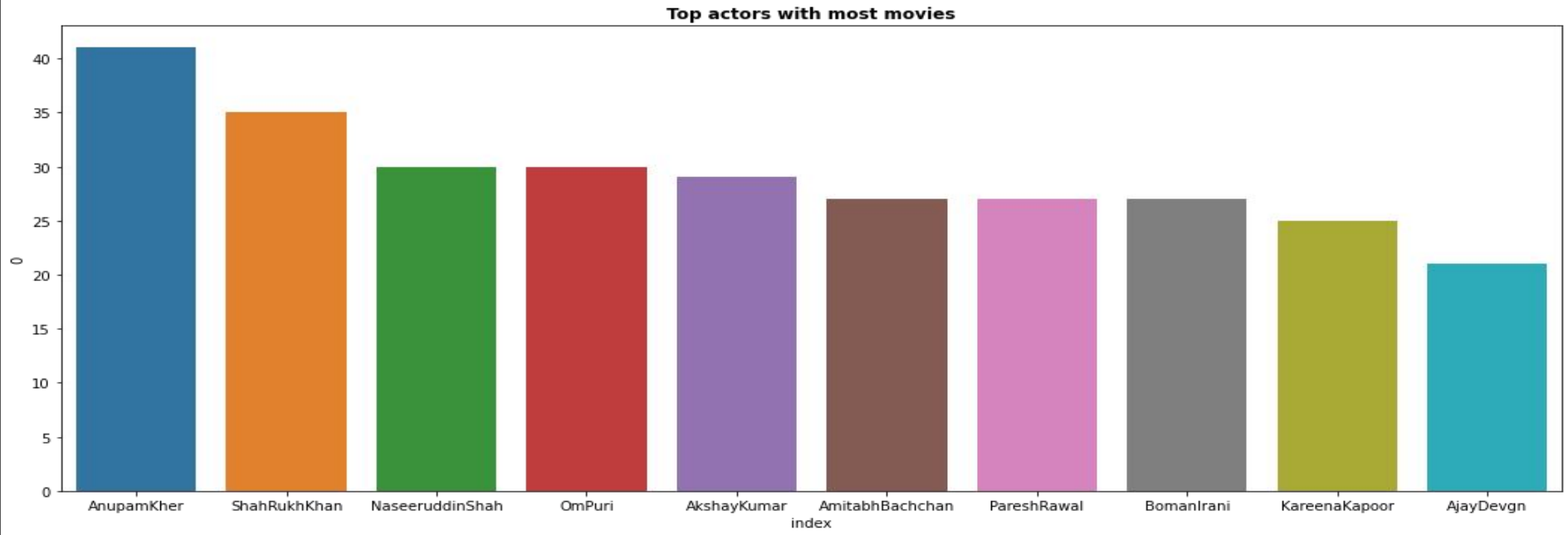
Module 2 Exploratory Data Analysis.

Let's analyze movie duration distribution



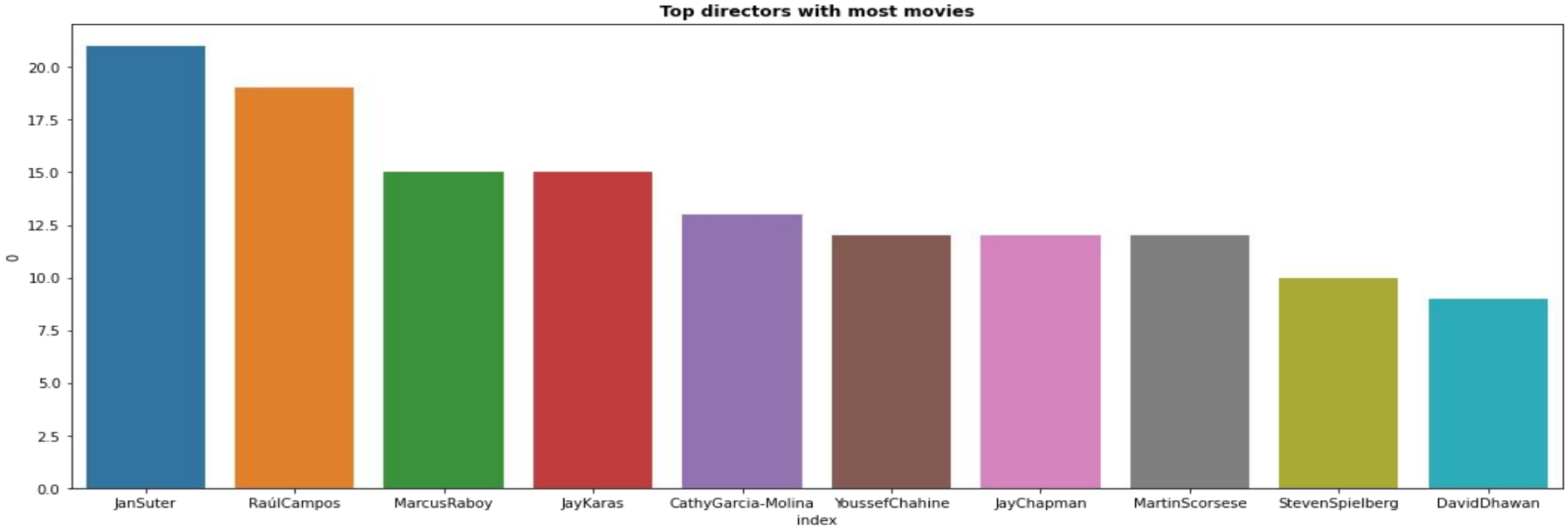
Module 2 Exploratory Data Analysis.

Let's check top actors and directors.



Module 2 Exploratory Data Analysis.

Let's check top actors and directors.



Module 3 Preprocessing and Feature Engineering

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is a crucial step while creating a machine learning model. Following are steps taken for preparing data for ML Model.

Stopwords and Punctuation elimination :
Stop word removal is one of the most commonly used preprocessing steps across different NLP applications. The idea is simply removing the words that occur commonly across all the documents in the corpus. Typically, articles and pronouns are generally classified as stop words.

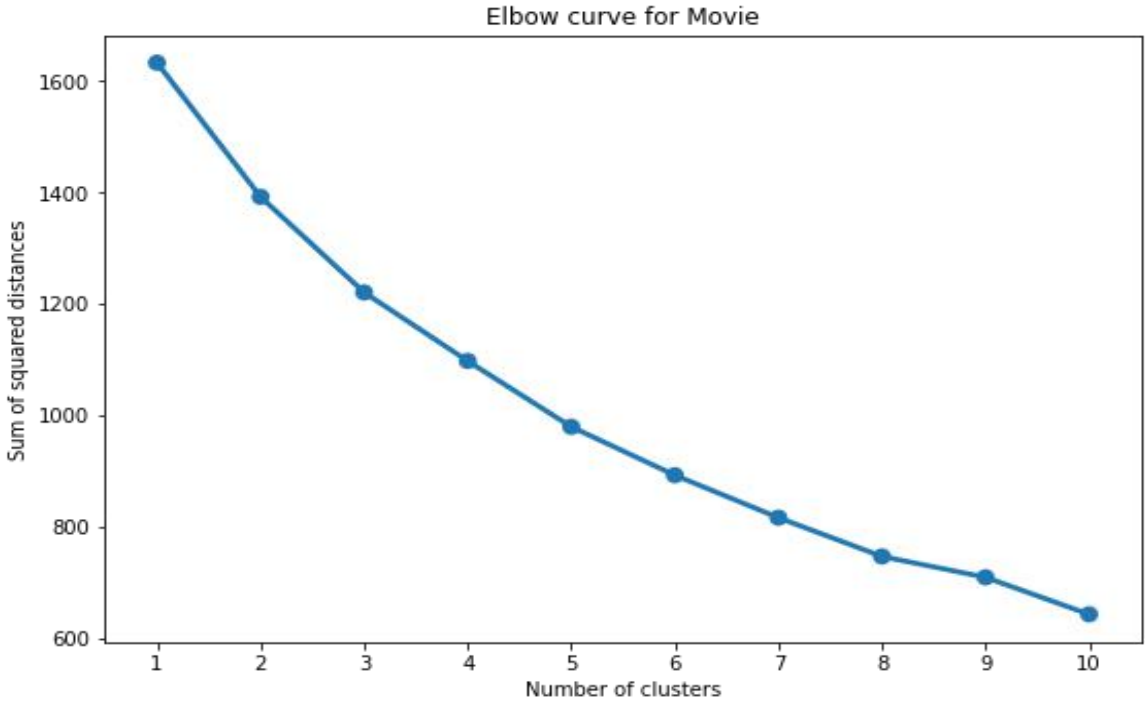
Stemming and Tokenizing :
Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. Stemming is important in natural language understanding (NLU) and natural language processing (NLP).

And finally, we divide dataset into Movie set and TV Show set to cluster on genres on both of them.

Module 4 Clustering on Movies dataset

Elbow method to get select number of clusters

It's difficult to spot a proper elbow here in the plot. We'll experiment with different clusters to have a perfect cluster number.



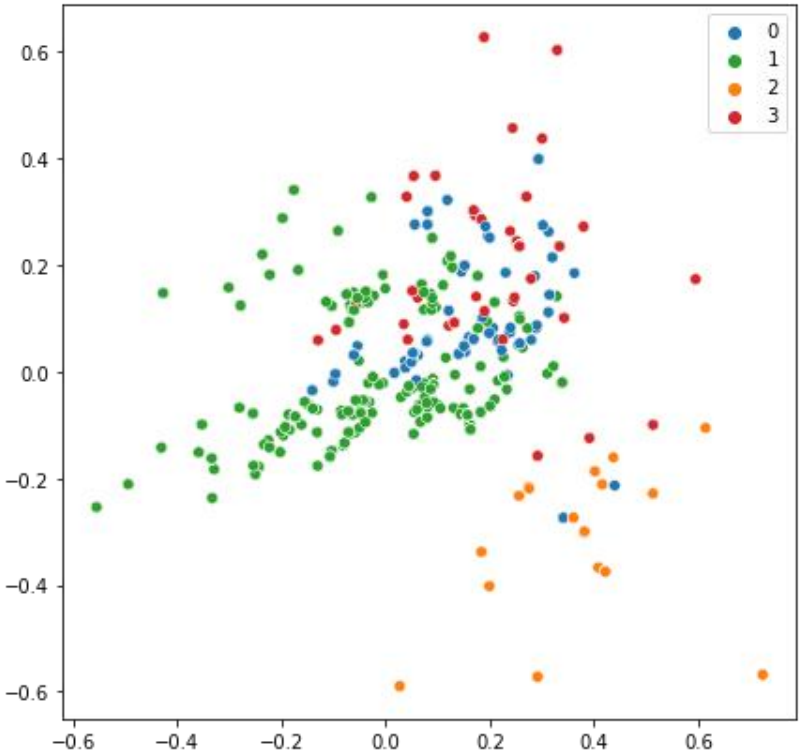
Module 4 Clustering on Movies dataset

Application of K-Means

When applying K-Means on movies dataset and tuning hyperparameters we found best classification was when K was equal to 4

Here in figure :

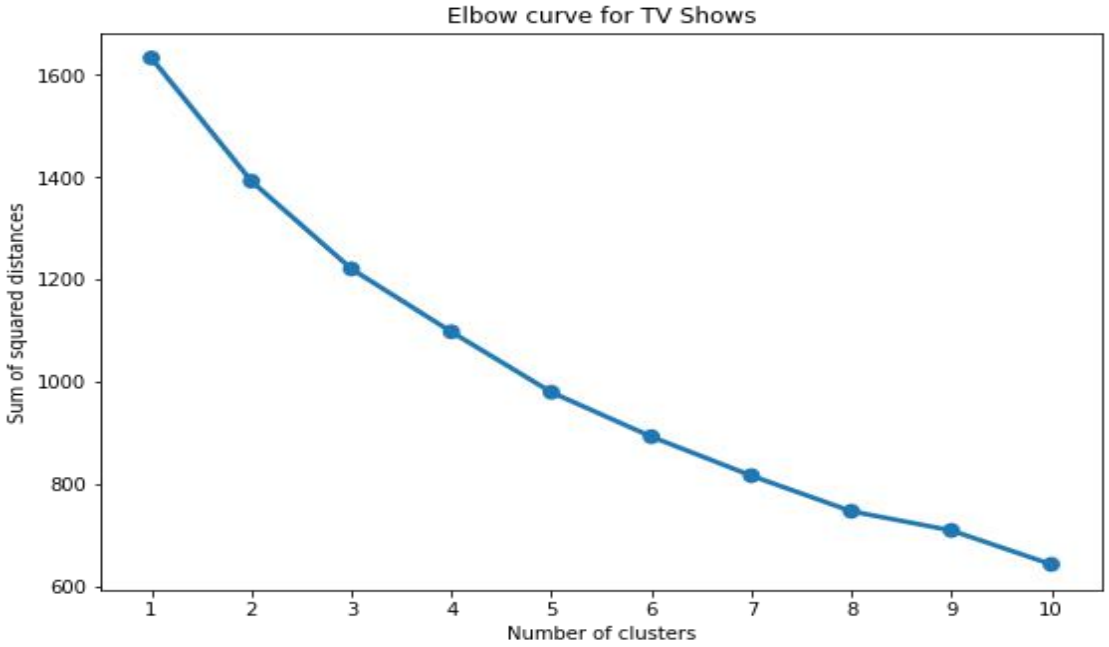
0 : International Drama, Thriller, Horror movies
1 : Documentary movies of different types
2 : Action and Adventure movies
3 : Comedy, Kids and Family movies



Module 4 Clustering on TV Shows dataset

Elbow method to get select number of clusters

Here again, It's difficult to spot a proper elbow here in the plot. We'll again have to experiment with different clusters to have a perfect cluster number.



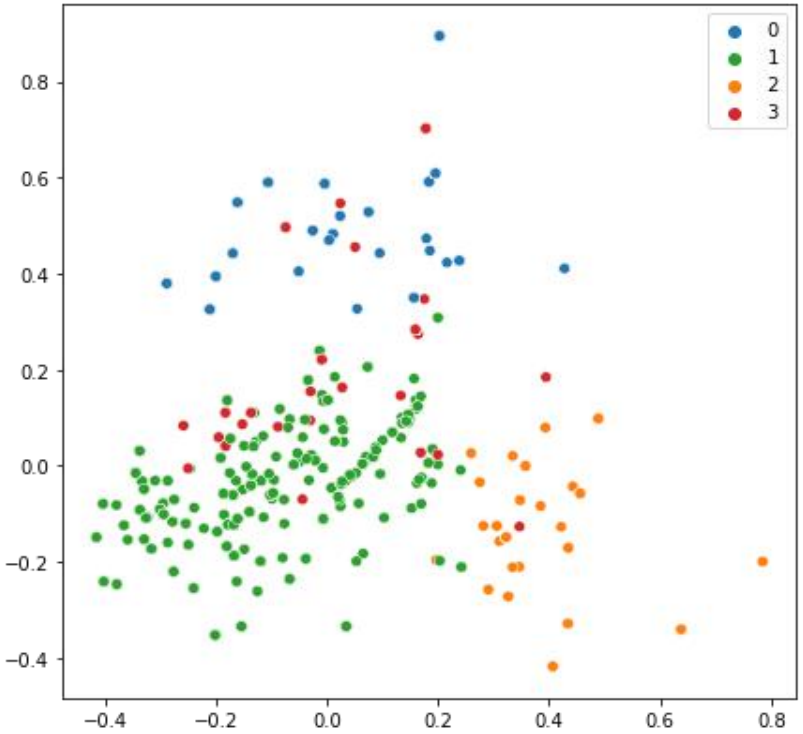
Module 4 Clustering on TV Shows dataset

Application of K-Means

When applying K-Means on TV Shows dataset and tuning hyperparameters we found best classification was when K was again equal to 4

Here in figure :

0 : Documentary Shows on different types
1 : International Drama Shows
2 : Reality based TV Shows
3 : kids and comedy Shows



Conclusion

Looking at barplot and piechart at beginning of analysis it was very evident that, Netflix focuses on movies more than TV Shows.

We had a look at different content type viewed in different countries. Where
* Canada's most watched content was comedy, TV Shows, International Movies etc.
* India's most viewed are International movies and more
* Japan watches Anime content more than any other type.

We checked for top actors which were dominated by bollywood actors like Anupam Kher, Shah Rukh Khan, Naseeruddin shah and more.

Also checked top directors in the industry, globally which were Jan Suter, Raul Campos, Marcus Raboy and more.

Conclusion

In preprocessing we performed :
* Stopword elimination
* Punctuation elimination
* Stemming of words
* Tokenization

As we separated our clustering for Movies and TV Shows we plotted two elbow curve for each. But, it was very difficult to identify a proper cluster number form both curve. Therefore we experimented with different clusters to identify best cluster number visually.

Finally, we found “4” clusters to be the best number for separation of genres in both Movies ans TV Shows. Where,
* Movies : Action, Drama, Comedy and Documentary Movies
* TV Shows : Drama, Comedy, Documentary and Reality Shows