

Capstone Project - 4

Netflix Movies And TV Shows Clustering Unsupervised Machine Learning

Project by:

Debashish Das Lucky Jain Vivek Katolkar

Table Of Contents

- 1. Defining problem statement
- 2. Data Cleaning & visualization
- 4. Data Preprocessing
- 5. Feature Selection
- 6. Applying different clustering methods
- 7. Applying Clustering Models
- 8. Conclusion

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

Data Summary

- **show_id**: Unique ID for every Movie / Tv Show
- **type**: 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 : Generes
- description: The Summary description



Basic Data Exploration

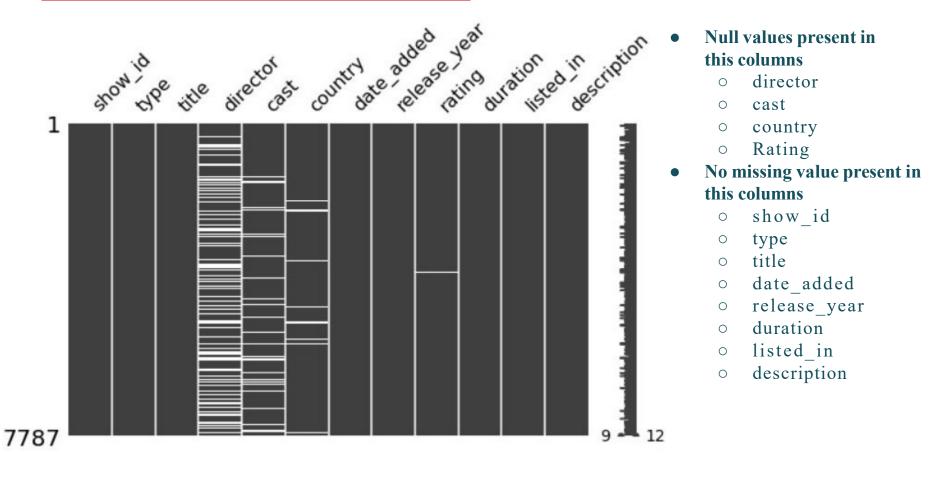
- The dataset has 7787 observations and 12 features (columns).
- The dataset consists of eleven textual columns and one numeric column('release year')
- No Duplicate values.

Dataset Shape: (7787, 12)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7787 entries, 0 to 7786
Data columns (total 10 columns):
                  Non-Null Count Dtype
    Column
    show id
                  7787 non-null
                                  object
    type
                  7787 non-null
                                  object
    title
                  7787 non-null
                                  object
    country
                  7280 non-null
                                  object
    date added
                  7777 non-null
                                  object
                                  int64
    release year 7787 non-null
                  7780 non-null
    rating
                                  object
    duration
                  7787 non-null
                                  object
    listed in
                 7787 non-null
                                  object
    description 7787 non-null
                                  object
dtypes: int64(1), object(9)
memory usage: 6.0 MB
```



EDA (Checking NAN Values)



Data Cleaning



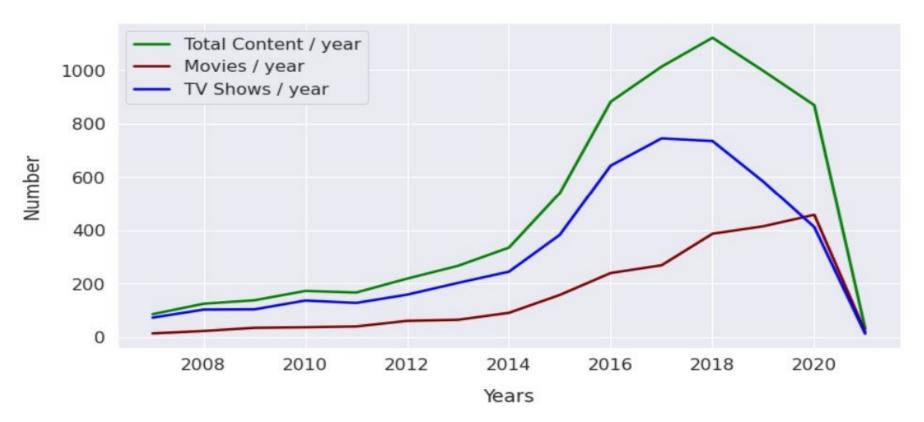
- Removing unnecessary columns like 'director', 'cast'.
- Dropping all the NaN containing date_added observations(Only 10 observations was there)
- Created 4 new columns
 - No_of_categories based on listed_in
 - $\circ \quad Date_added_month\ based\ on\ date_added$

	listed_in	no_of_category
0	International TV Shows, TV Dramas, TV Sci-Fi &	3
1	Dramas, International Movies	2
2	Horror Movies, International Movies	2
3	Action & Adventure, Independent Movies, Sci-Fi	3
4	Dramas	1

	December	October	January	November	March	September	August	April	July	June	May	February
date_added_month	817	780	746	730	661	614	612	596	592	538	537	466

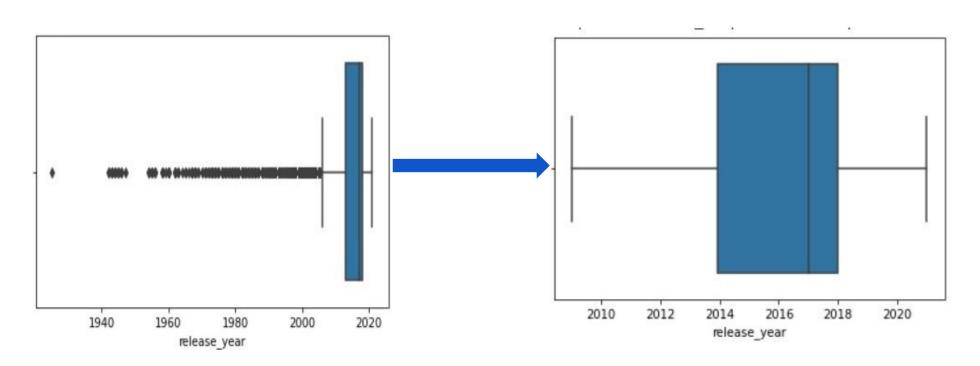
Production Yearly Growth





• Can you say what's the reason of that boom growth??

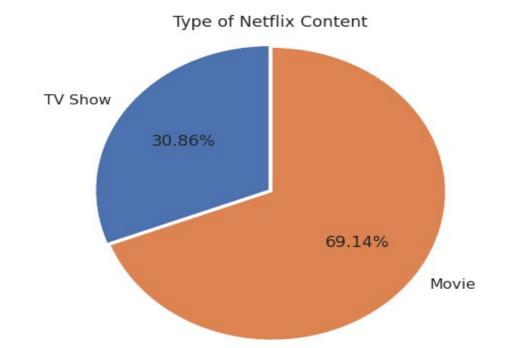
Checking Outliers



Replaced outliers values with mean value of release_year

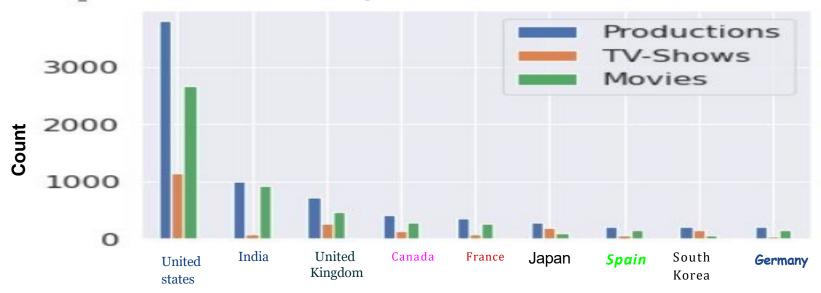


Tv shows or Movies ??

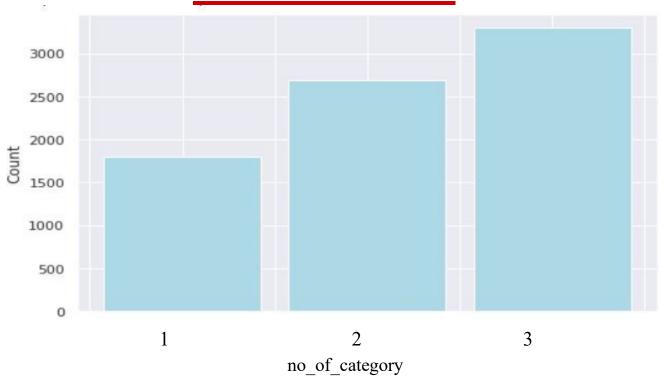


- Most of the contents are Movies
- Less than ⅓ content are Tv Shows

Countries Producing Most No Of Contents

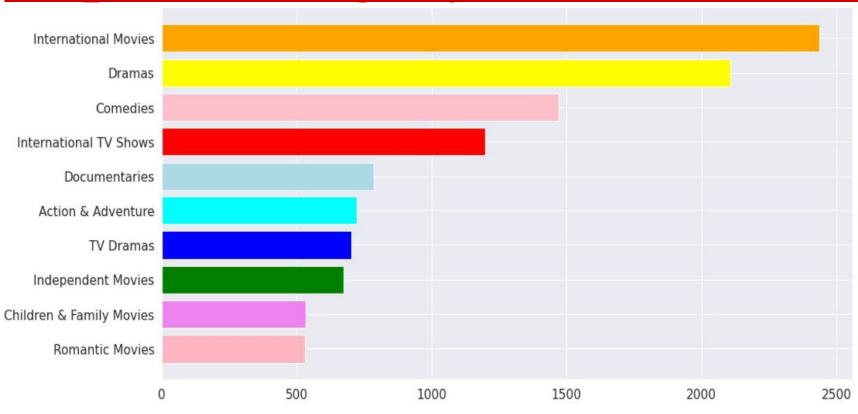


- United states have the most number of content and then india and so on
- We can conclude that except **Japan** other countries are producing movies more than TV-Shows

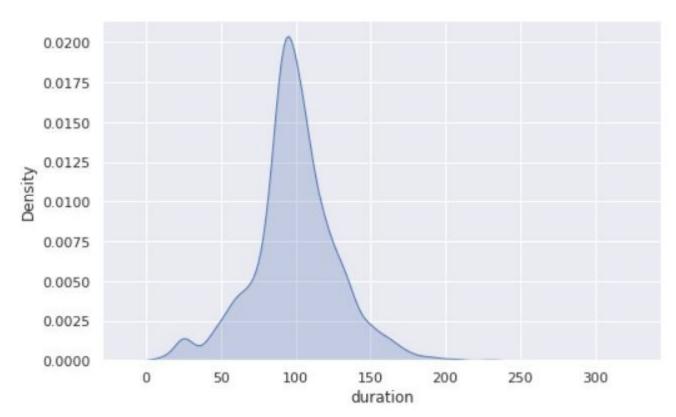


Most of the movies are belonging to 3 categories

Top 10 Category For Contents Al



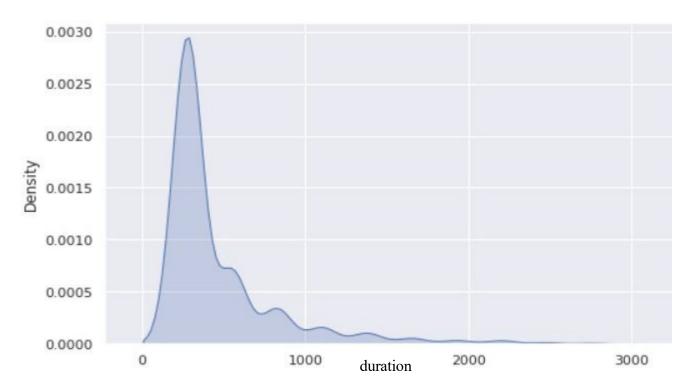
Movie Wise Density Plot



Most movies are about 70 to 120 min duration for movies

Tv-shows Wise Density Plot

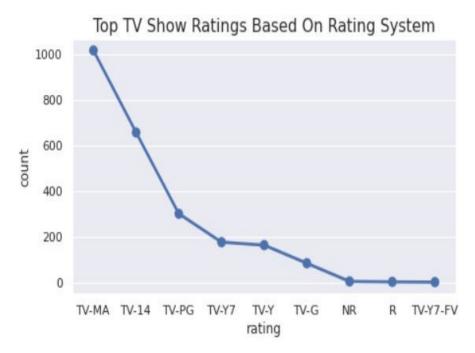




- Most contents are about 0 to 750 min duration for Tv shows
- There are very few shows which is having more than 1000 mins. (may be the no of episodes/ seasons are more)

TOP Content Based On Rating







Most of the contents got ratings like

- TV-MA (For Mature Audiences)
- TV-14 (May be unsuitable for children under 14)
- TV-PG (Parental Guidance Suggested)
- NR (Not Rated)

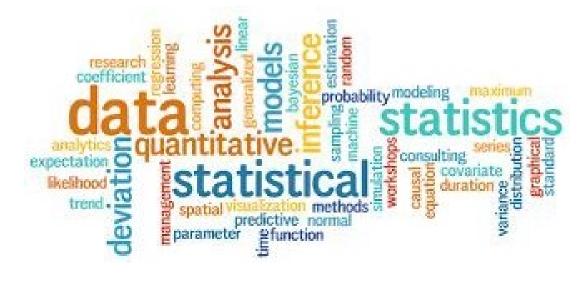


WordCloud

What Is a Word Cloud?

A word cloud (also known as a tag cloud) is a visual representation of words. Cloud creators are used to highlight popular words and phrases based on frequency and relevance. They provide you with quick and simple visual insights that can lead to more in-depth analyses.





AI

Applying WordCloud On Title



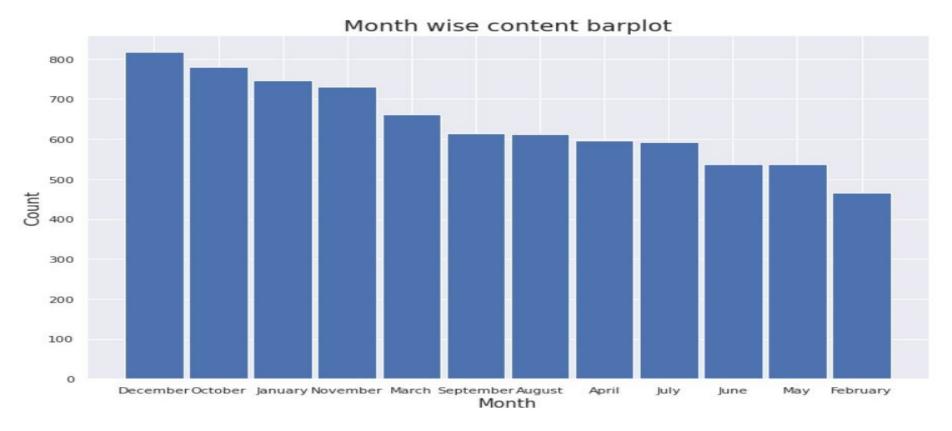
Most occurred words present in Title are:-

- Love
- Man
- World
- Story
- Christmas
- Girl
- Day



Barplot Based On Release Month



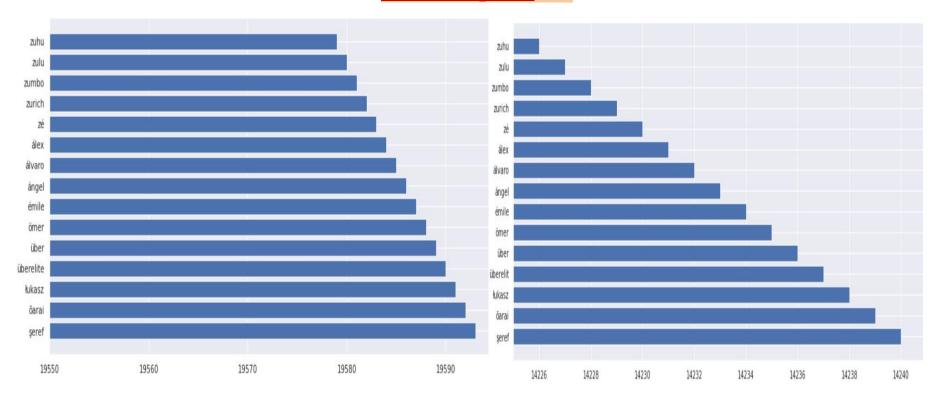


• We can say that December is the holiday season and it also has Christmas, so in that month most of the content got uploaded.



Before & After Stemming Most Occurred Words

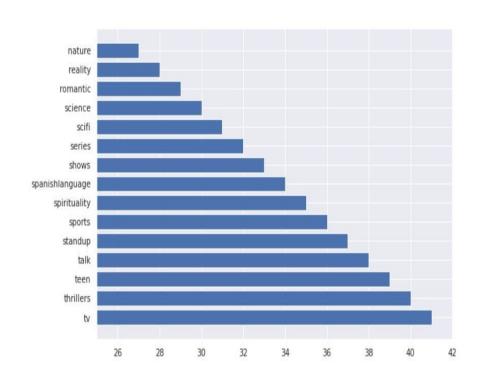
In Description

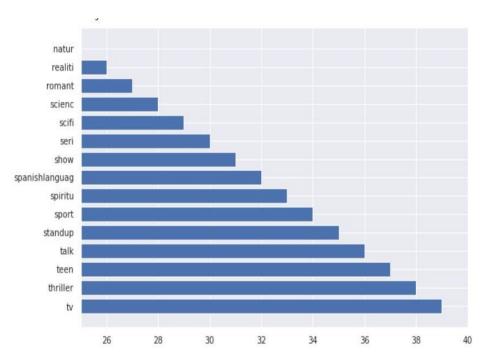




Before & After Stemming Most Occurred Words

In listed In





Feature Selection & ML Algo Used

- Only selected 3 features, to do clustering
 - o no of category
 - Length(description)
 - Length(listed-in)
- Using StandardScaler
- Used 5 algo to find out best k value
 - 1. Silhouette score
 - o 2. Elbow Method
 - o 3. DBSCAN
 - 4. Dendrogram
 - 5. AgglomerativeClustering

1. Silhouette Score

Silhouette Coefficient Formula

$$S = \frac{(b-a)}{max(a,b)}$$
.

- **mean intra-cluster distance(a):-** Mean distance between the observation and all other data points in the same cluster.
- mean nearest-cluster distance (b) :- Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be called a.

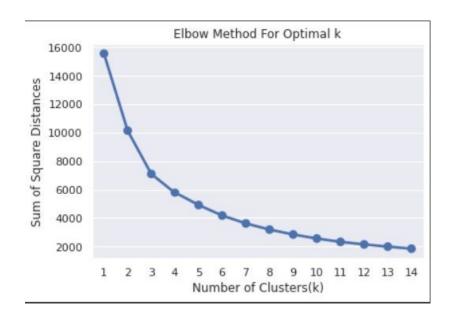
The value of the silhouette coefficient is between [-1, 1]

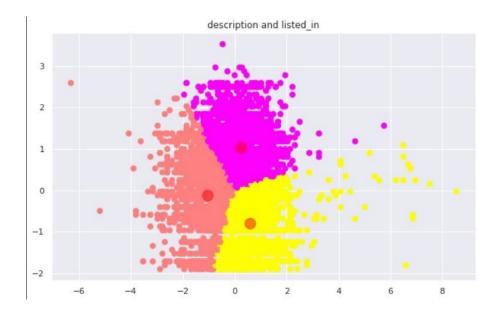
- If score is **1 denotes the best** meaning that the data point i is very compact within the cluster to which it belongs and far away from the other clusters.
- The worst value is -1
- If score is 0 denotes overlapping clusters

	n clusters	silhouette score
0	2	0.425
3	5	0.396
1	3	0.394
4	6	0.393
2	4	0.385
8	10	0.375
5	7	0.373
7	9	0.366
9	11	0.364
13	15	0.361
6	8	0.360
10	12	0.360
12	14	0.352
11	13	0.349

2. Elbow Method



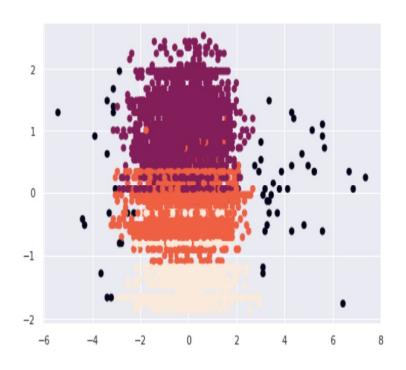




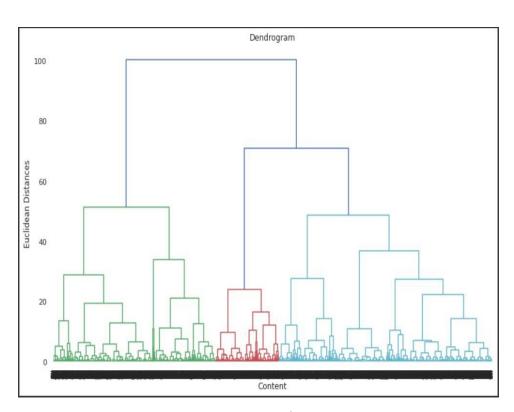
The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-15) and then for each value of k computes WCSS value. By default, the distortion score is computed, the sum of square distances from each point to its assigned center.



3 & 4 DBSCAN & Dendrogram



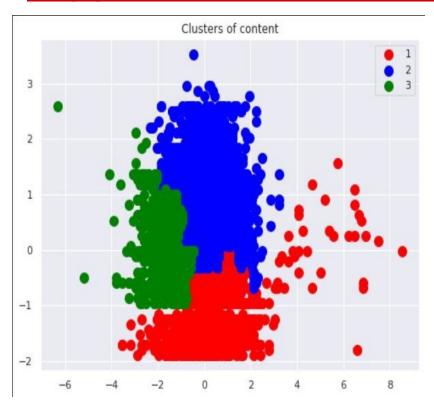
DBSCAN(Density-based spatial clustering) the black dots denoted noise



Dendrogram

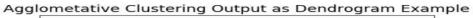
Agglomerative Clustering

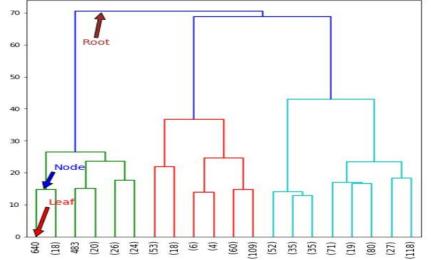




Steps: -

- 1. Each data point is assigned as a single cluster.
- 2. Determine the distance measurement and calculate the distance matrix.
- 3. Determine the linkage criteria to merge the clusters.
- 4. Update the distance matrix.
- 5. Repeat the process until every data point become one cluster.





Conclusion

Director and cast contains a large number of null values so we will drop these 2 columns.
In this dataset there are two types of contents where 30.86% includes TV shows and the remaining 69.14% carries Movies.
We have reached a conclusion from our analysis from the content added over years that Netflix is focusing movies and TV shows (Fom 2016 data we get to know that Movies is increased by 80% and TV shows is increased by 73% compare).
From the dataset insights we can conclude that the most number of TV Shows released in 2017 and for Movies it is 2020.
On Netflix USA has the largest number of contents. And most of the countries preferred to produce movies more than TV shows.
Most of the movies are belonging to 3 categories.
TOP 3 content categories are International movies, dramas, comedies.
In text analysis (NLP) I used stop words, removed punctuations , stemming & TF-IDF vectorizer and other functions of NLP.
Applied different clustering models like Kmeans, hierarchical, Agglomerative clustering, DBSCAN 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.

