Capstone Project

NETFLIX MOVIES AND TV SHOWS CLUSTERING

Team Members

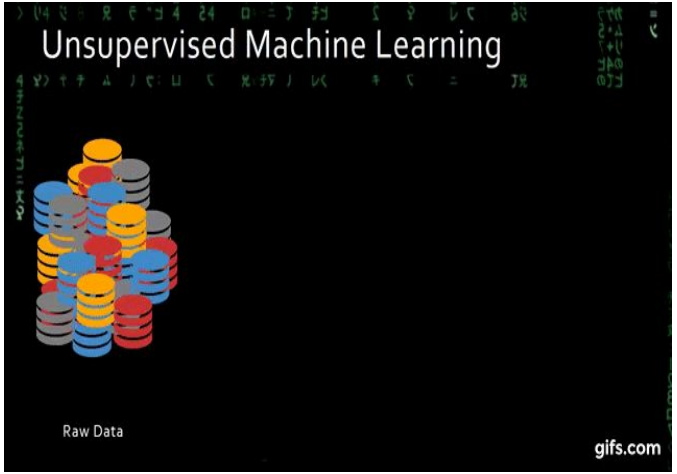
Bhavesh Rikame
AlmaBetter Data Science Trainee

Anannya Sagar Das
AlmaBetter Data Science Trainee

Kunal Borse
AlmaBetter Data Science Trainee

Itisha Jain
AlmaBetter Data Science Trainee

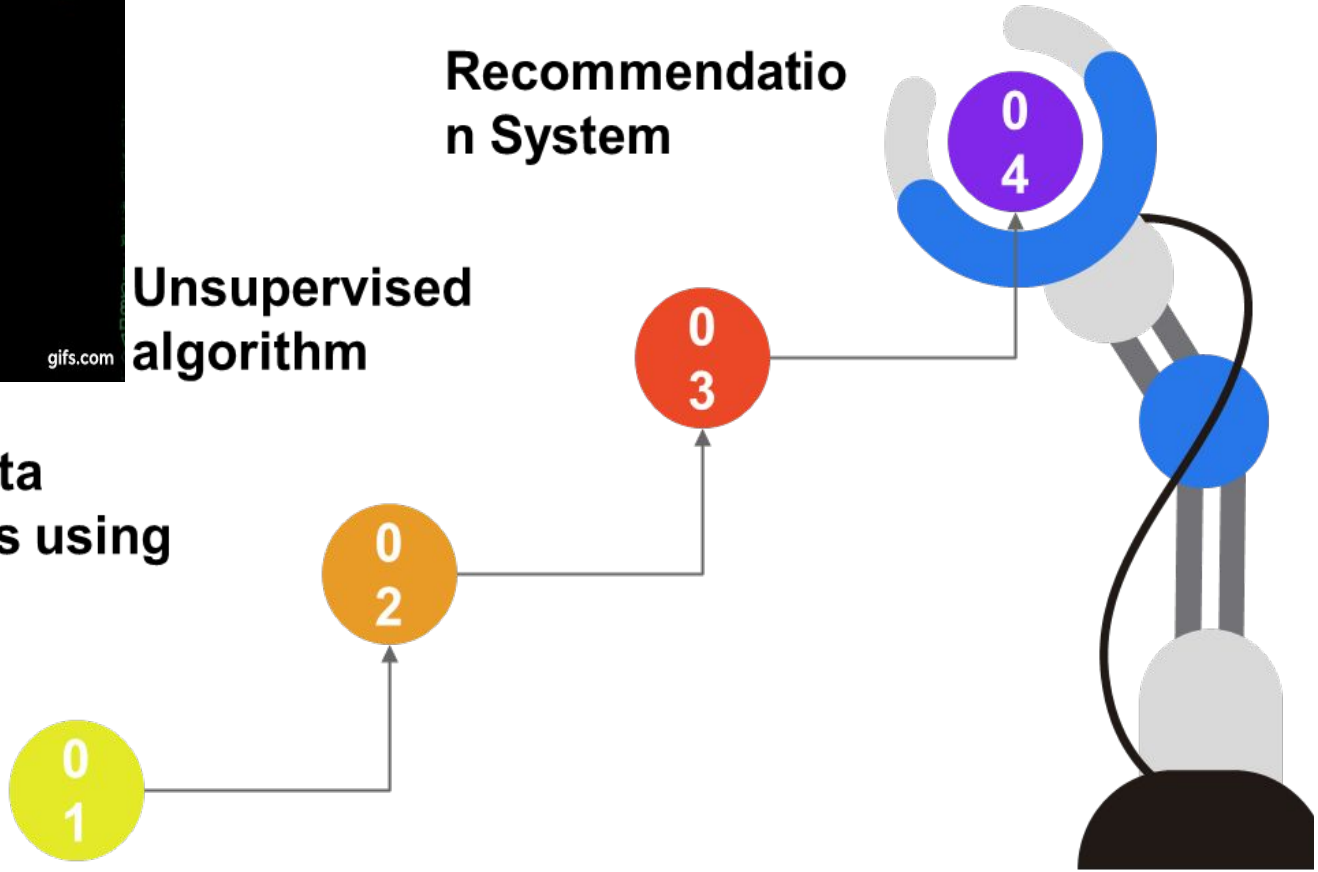
Roadmap



Unsupervised algorithm

Text data analysis using NLP

EDA



Know About Your Data

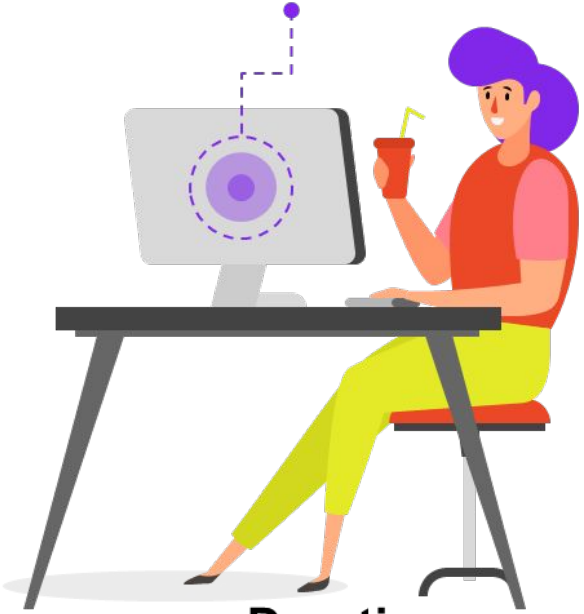
01 Show id

Type 02

Rating

03 Title

Director 04



05 Cast

Country 06

07 Date added

Release year 08

Duration

Exploratory Data Analysis

Univariate Analysis

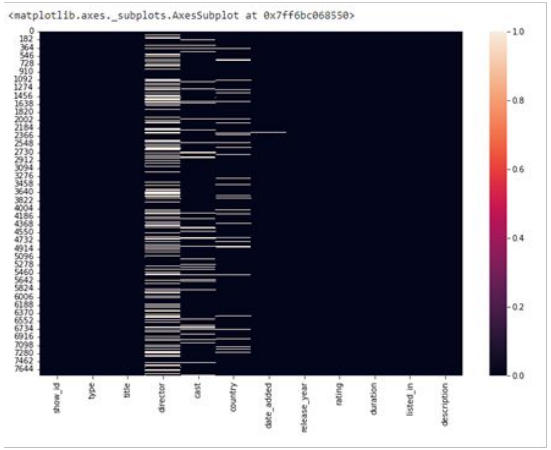
- We performed analysis on eachand every column to get insights about that perticular column
- E.G,Type,Title ,Director etc

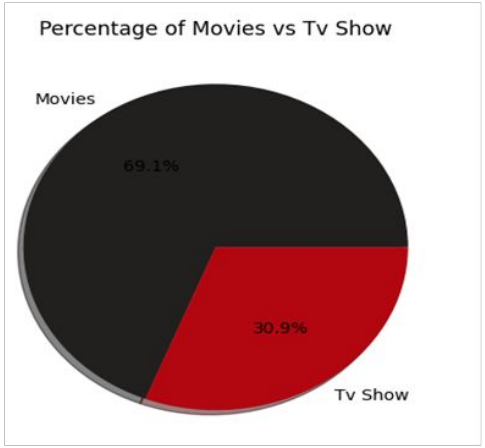
Vs

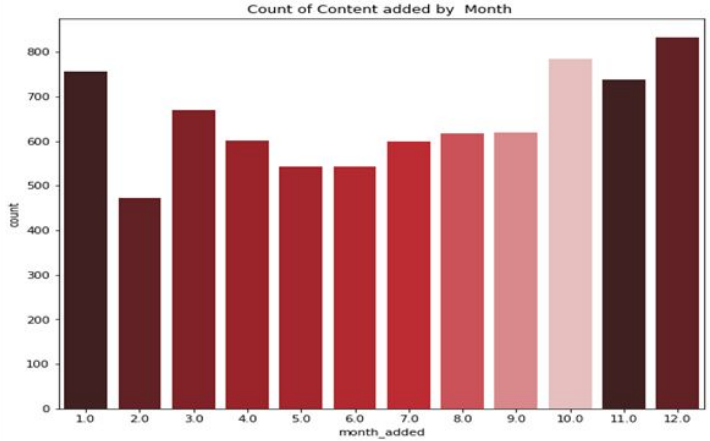
Bivariate Analysis

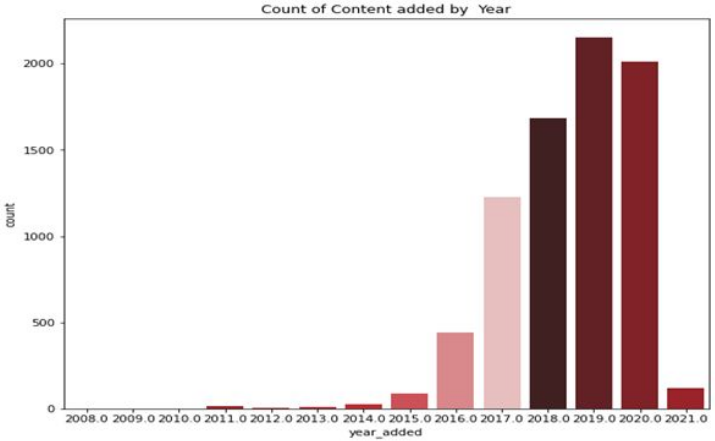
country vs genre
country vs rating
country vs type
country vs year_added
country vs top directors
country vs top cast
country vs release year

Univariate analysis

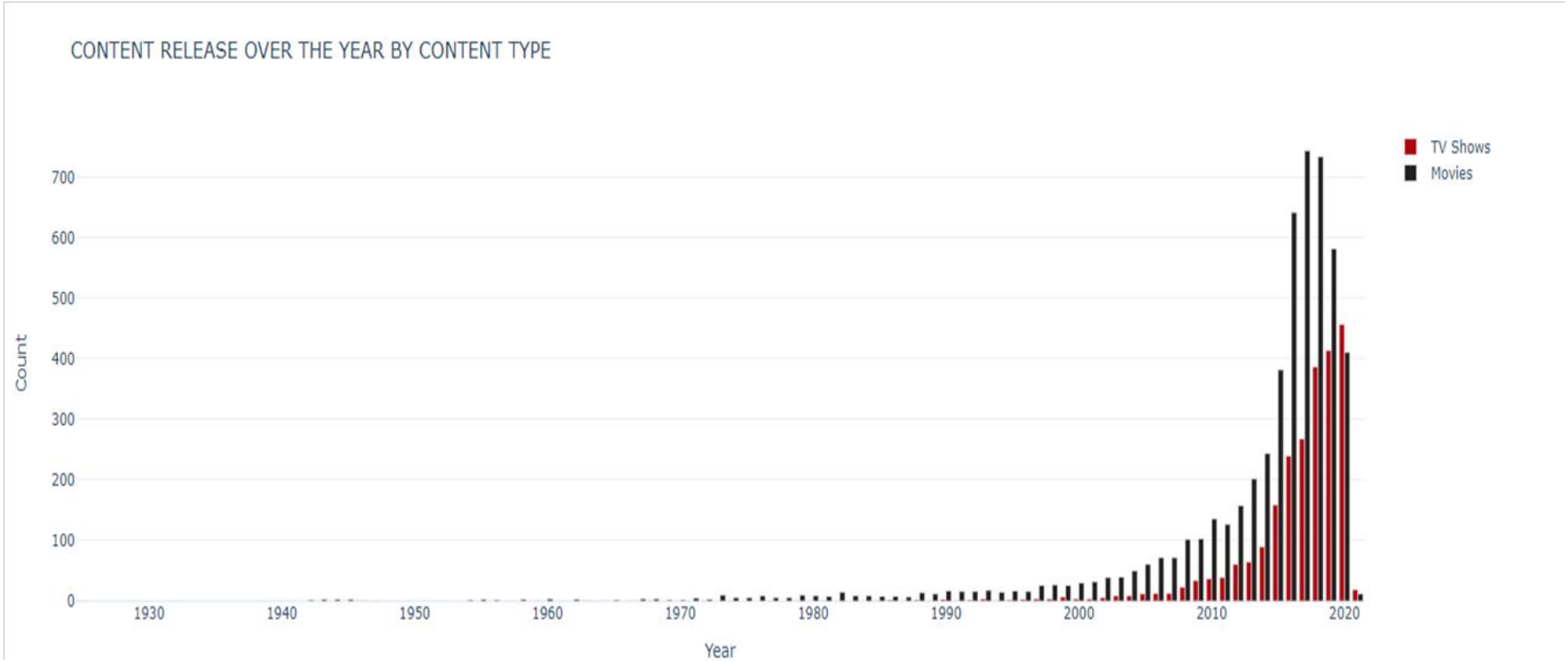




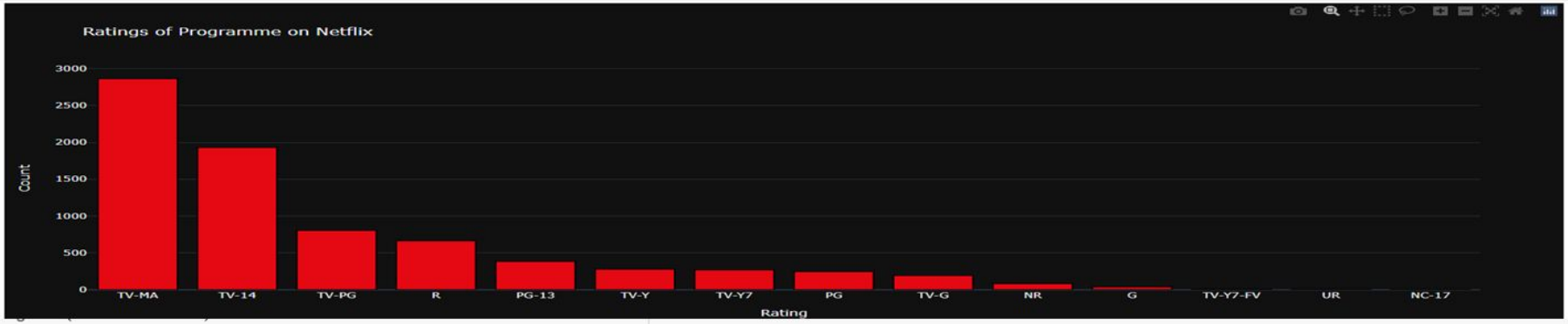


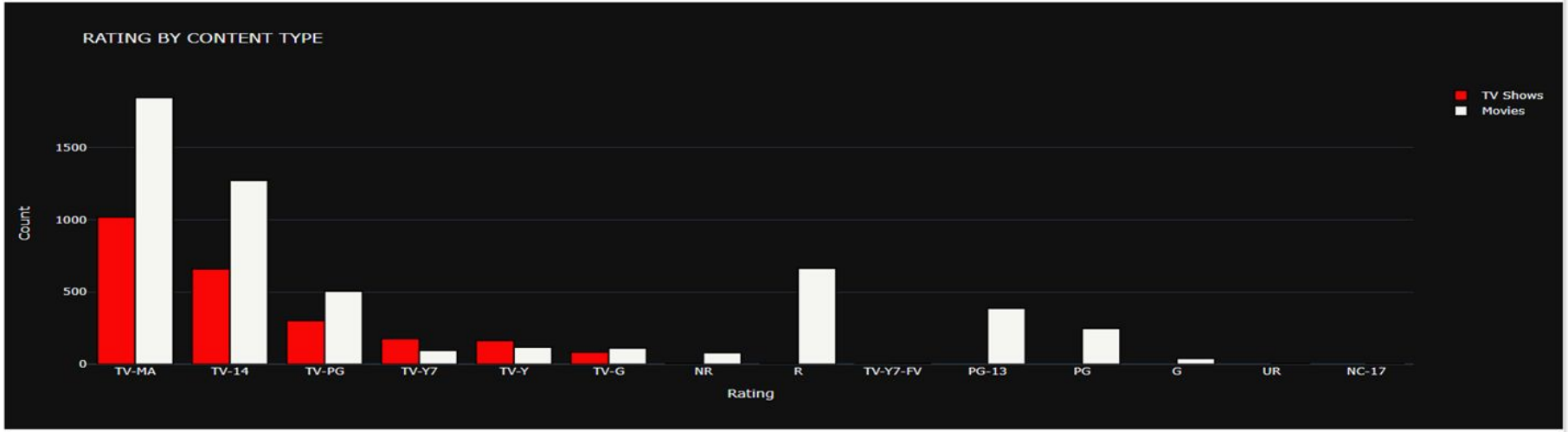


Release over the years (Content type)

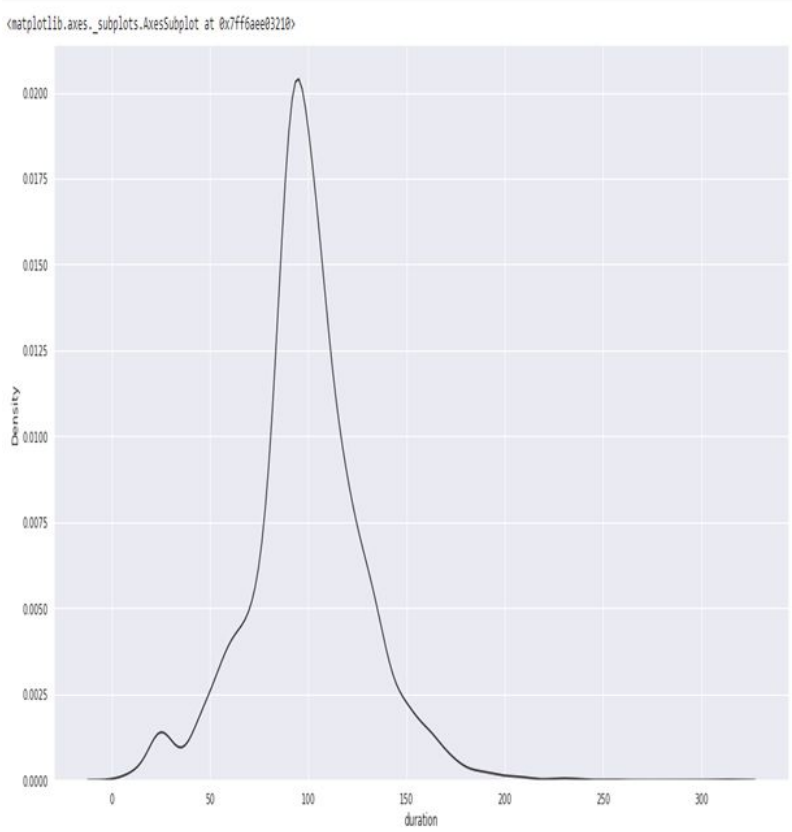


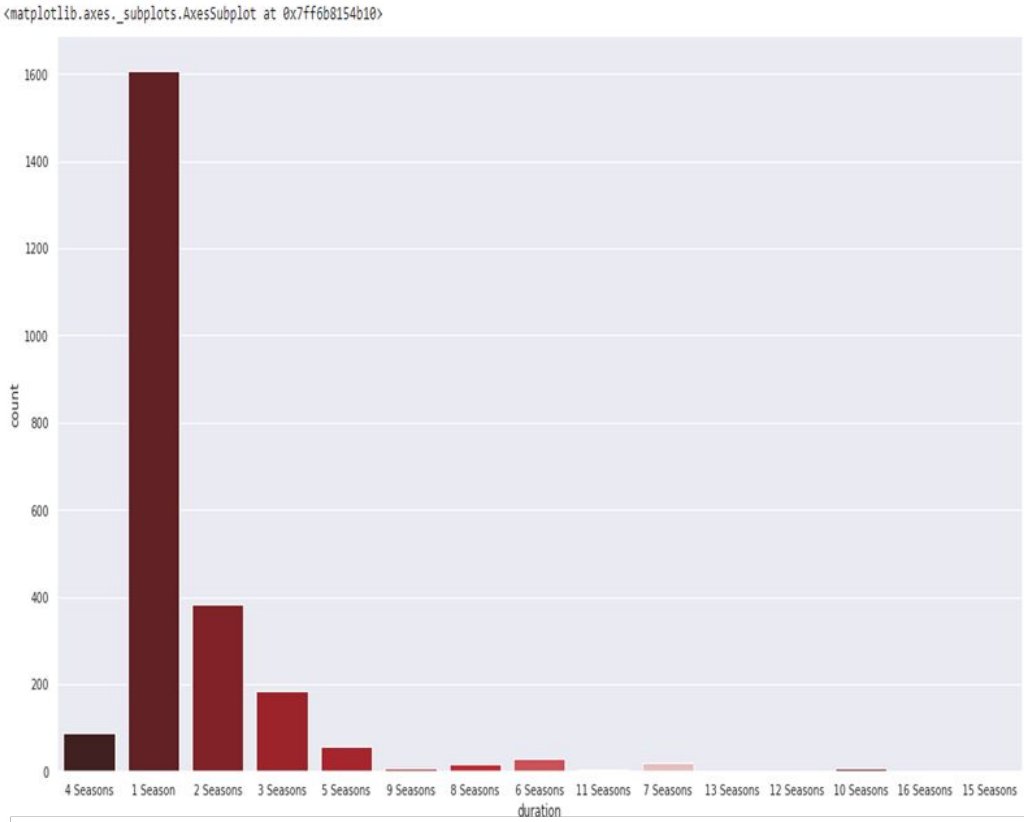
Ratings


Ratings of Programme on Netflix
Count
3000
2500
2000
1500
1000
500
0
TV-MA
TV-14
TV-PG
R
PG-13
TV-Y
TV-Y7
PG
TV-G
NR
G
TV-Y7-FV
UR
NC-17
Rating

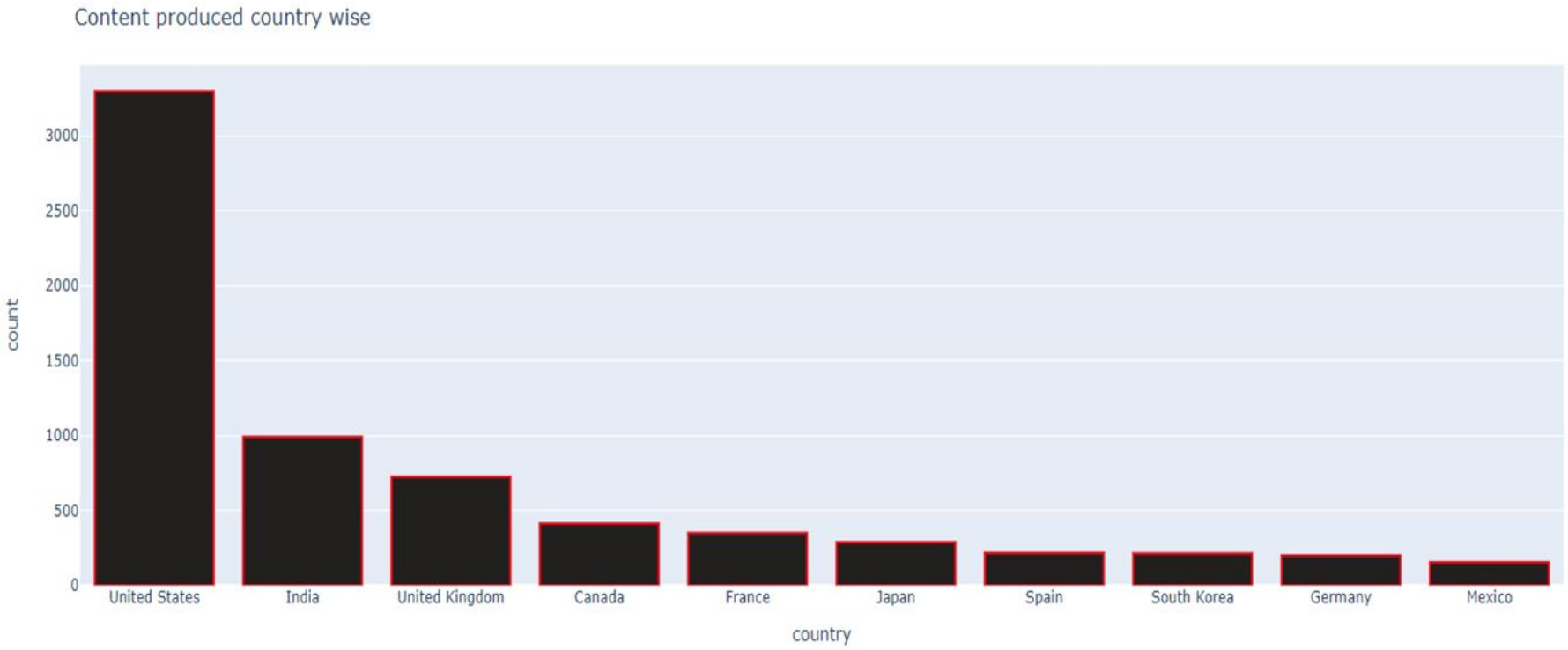

RATING BY CONTENT TYPE
TV Shows
Movies
Count
1500
1000
500
0
TV-MA
TV-14
TV-PG
TV-Y7
TV-Y
TV-G
NR
R
TV-Y7-FV
PG-13
PG
G
UR
NC-17
Rating

Duration of Movie/TV shows

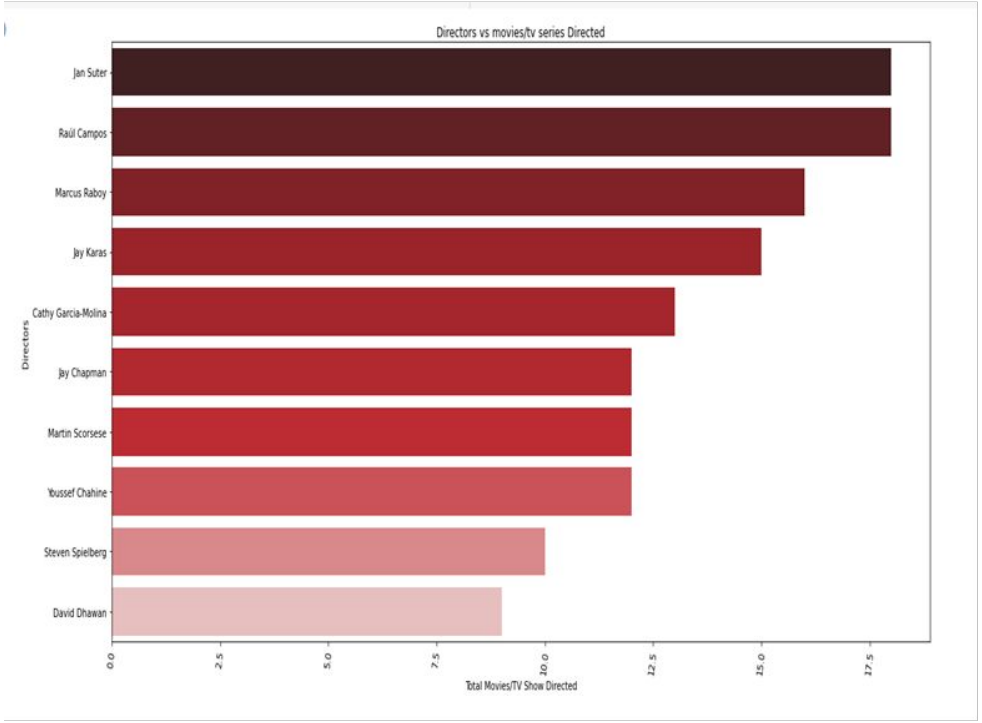


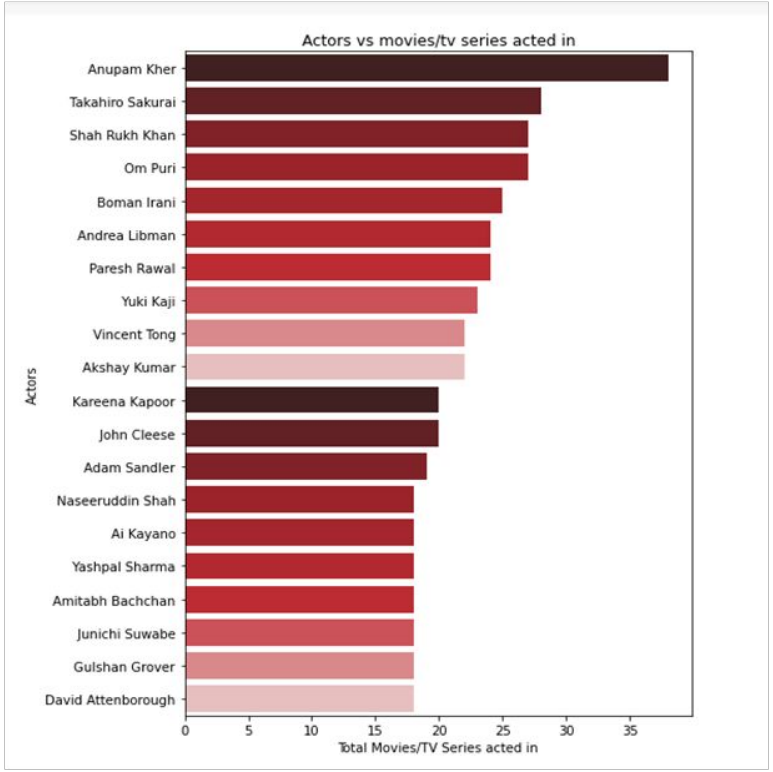


Content Produced Country Wise



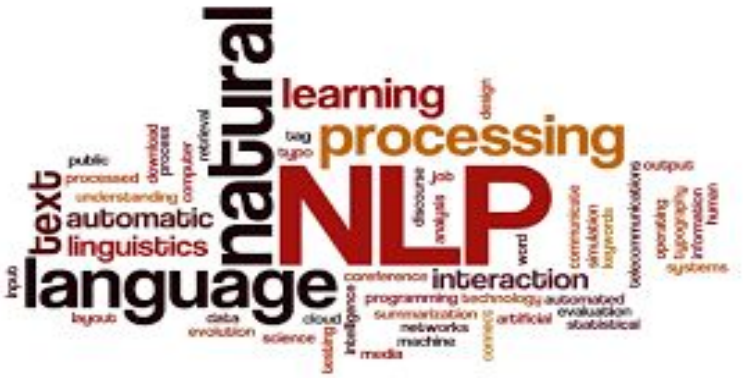
Directors and Actors





Hypothesis : Movies across the globe are mostly from India as there are more actors whose names come under top 20 movies count.

Natural Language Processing



Can Computers Understand Language?

As long as computers have been around, programmers have been trying to write programs that understand languages like English. The reason is pretty obvious,humans have been writing things down for thousands of years and it would be really helpful if a computer could read and understand all that data.

Computers can't yet truly understand English in the way that humans do, but they can already do a lot! In certain limited areas, what you can do with NLP already seems like magic.

NLP Pipeline: Step by step

Step 1: Sentence Segmentation

Step 2: Word Tokenization

Step 3: Predicting Parts of Speech for Each Token

Step 4: Text Lemmatization

Step 5: Identifying Stop Words

Step 6: Dependency Parsing

Step 7: TF-IDF Vectorization

Principal Component Analysis

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Step by Step Explanation of PCA

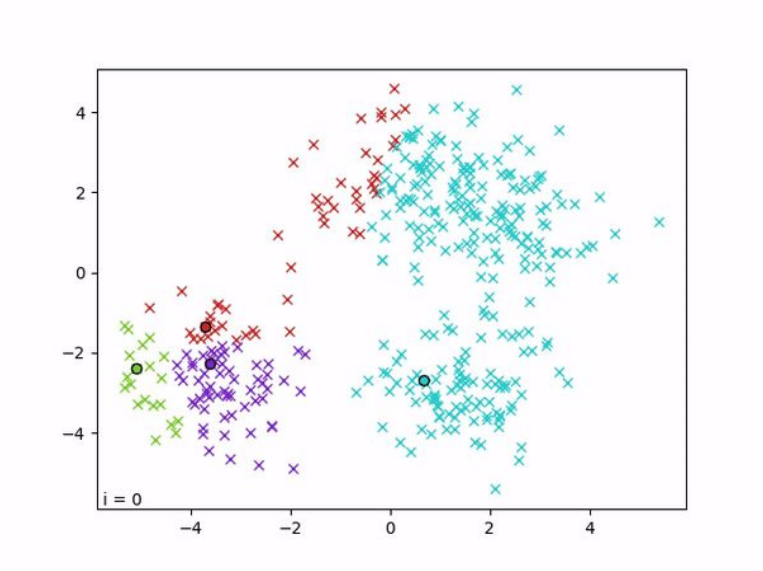
Step 1: Standardization

Step 2: Covariance Matrix Computation

Step 3: Compute The Eigenvectors And Eigenvalues Of The Covariance Matrix To Identify The Principal Components

Step 4: Feature Vector
Step 5: Recast The Data Along The Principal Components Axes

K-Means Clustering



K-Means is probably the most well-known clustering algorithm. K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.

The objective of K-means is simple: group similar data points together and discover underlying patterns.

How the K-means algorithm works :

1.Determine the value “K”, the value “K” represents the number of clusters.
2.Randomly select K distinct centroid (new data points as cluster initialization)

3. Measure the distance (euclidean distance) between each point and the centroid.

4. Assign the each point to the nearest cluster

5 Calculate the mean of each cluster as new centroid.

6. Repeat step 3–5 with the new center of cluster.

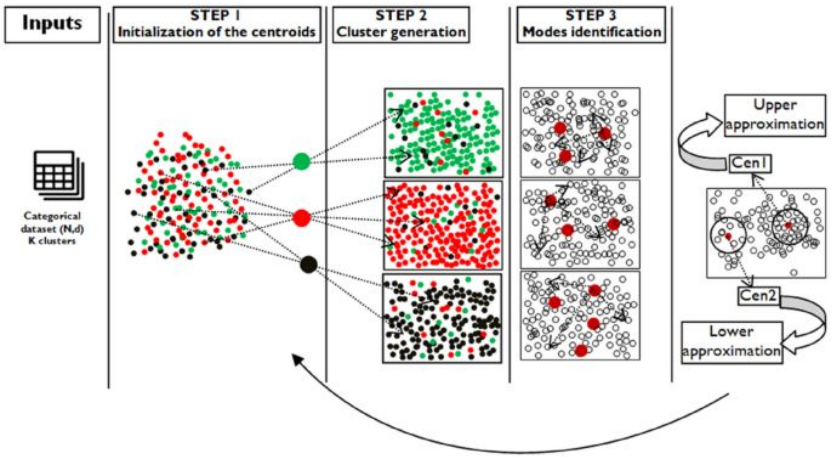
7.Repeat until stop:
Convergence. (No further changes)
Maximum number of iterations.

K-Modes Clustering

Numerical | Categorical

K-Means
(based on distance between two numerical data)

K-Modes
(based on number of differences between components of each attributes)



KModes clustering is one of the unsupervised Machine Learning algorithms that is used to cluster categorical variables.

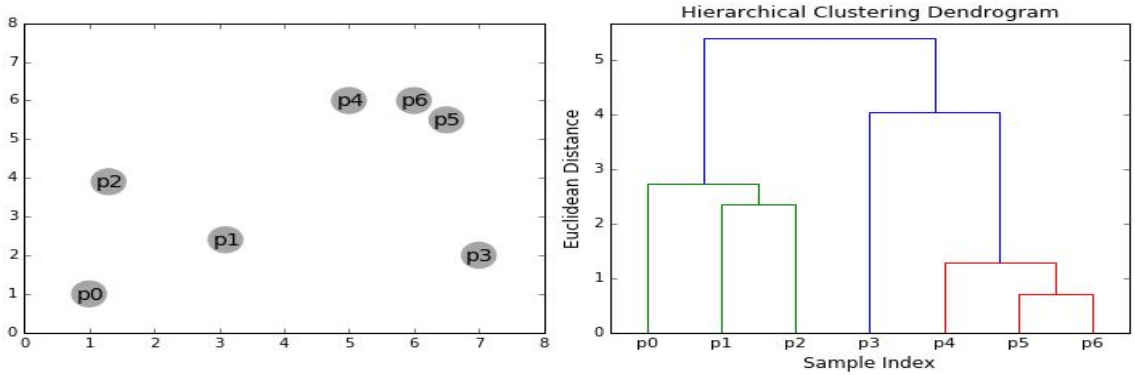
For categorical data points, we cannot calculate the distance. So we go for the K-Modes algorithm. It uses the dissimilarities between the data points. The lesser the dissimilarities the more similar our data points are. It uses Modes instead of means.

How does K-Modes Clustering work?

- Pick K observations at random and use them as clusters or defining points of a cluster also known as leaders.
- Calculate the dissimilarities and assign each observation to its closest cluster.
- Define new modes for the clusters.
- Repeat the first and third step until there is no re-assignment required.

Hierarchical Clustering

Hierarchical clustering is an alternative to prototype-based clustering algorithms. The main advantage of Hierarchical clustering is that we do not need to specify the number of clusters, it will find it by itself. In addition, it enables the plotting of dendrograms. Dendrograms are visualizations of binary hierarchical clustering.



Observations that fuse at the bottom are similar while those that are at the top are quite different. With dendrograms, conclusions are made based on the location of the vertical axis rather than on the horizontal one.

Finding optimal numbner of Clusters

Machine learning
Venus has a beautiful name, but it's hot

The model
Despite being red, Mars is a cold place

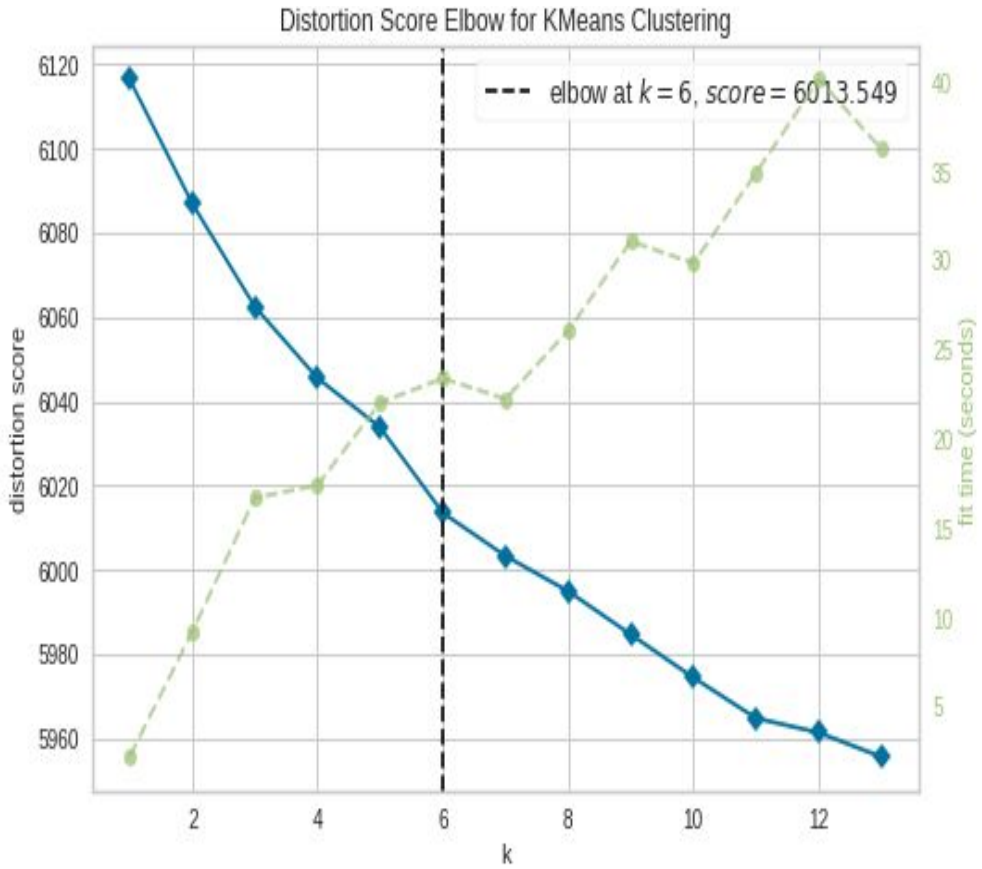
Elbow Method
The Earth is the third planet from the Sun

Dendogram
Mercury is the closest planet to the Sun

Finding Optimal number of clusters

01 Elbow Method

WCSS is the sum of squared distance between each point and the centroid in a cluster.



Finding Optimal number of clusters

02 Silhouette Score

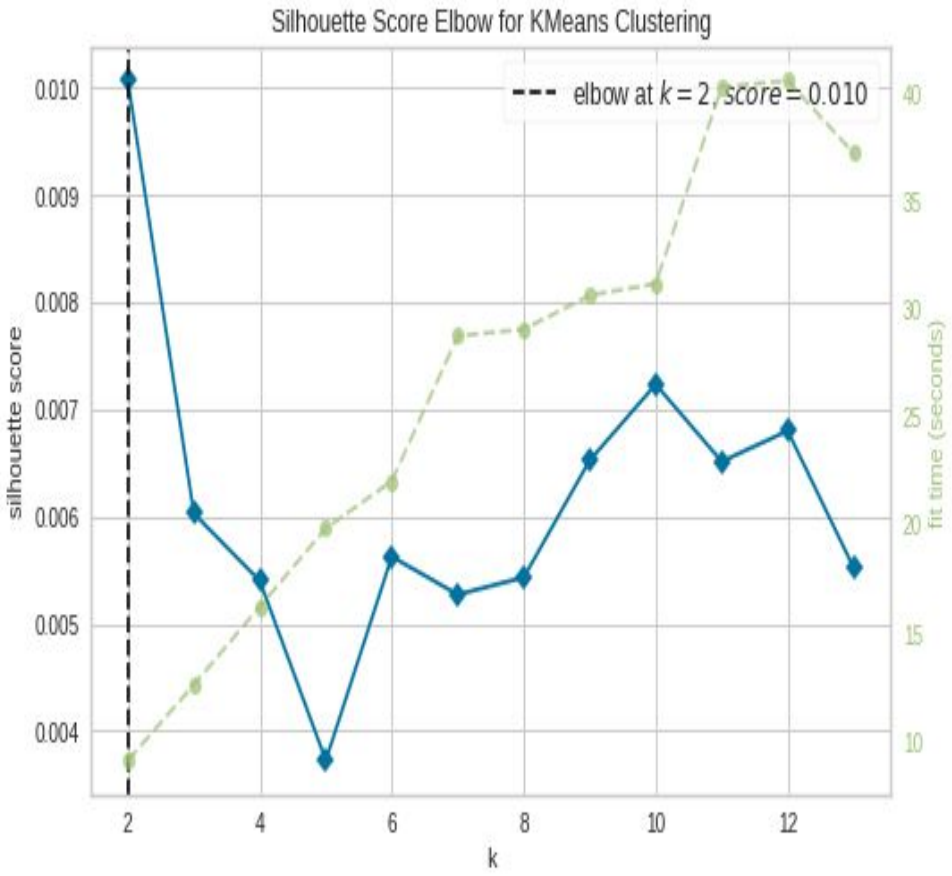
measures how similar a data point is within-cluster (cohesion) compared to other clusters (separation)

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

S(i) is the silhouette coefficient of the data point i.

a(i) is the average distance between i and all the other data points in the cluster to which i belongs.

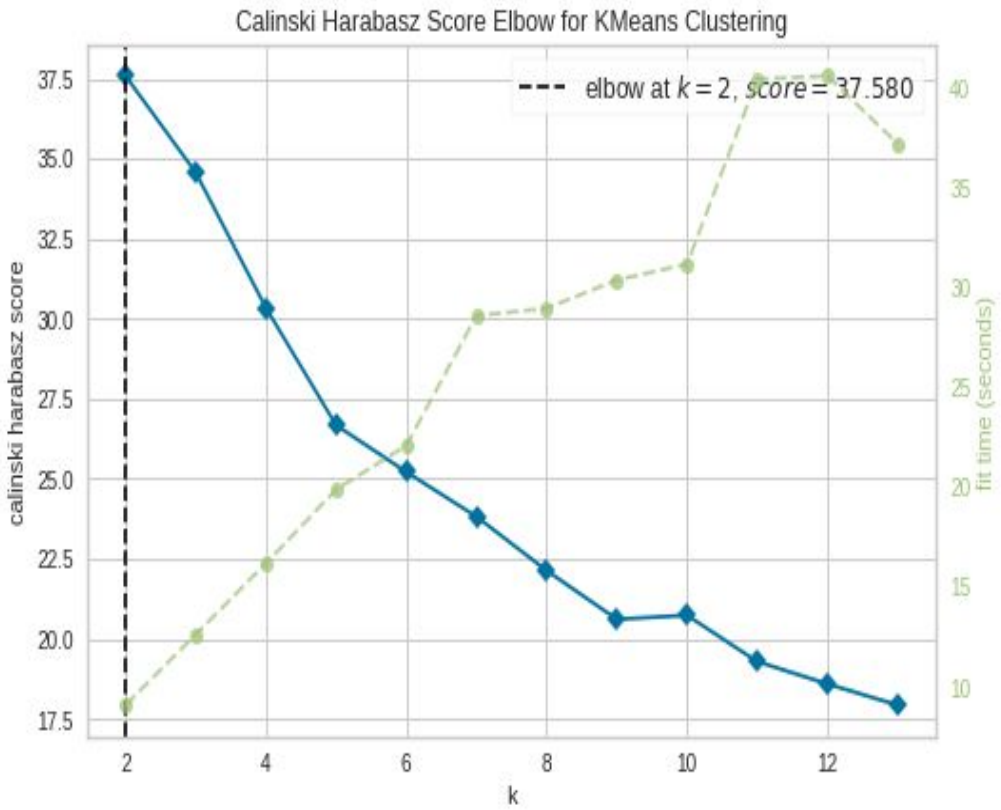
b(i) is the average distance from i to all clusters to which i does not belong.



Finding Optimal number of clusters

03 Calinski-Harabasz

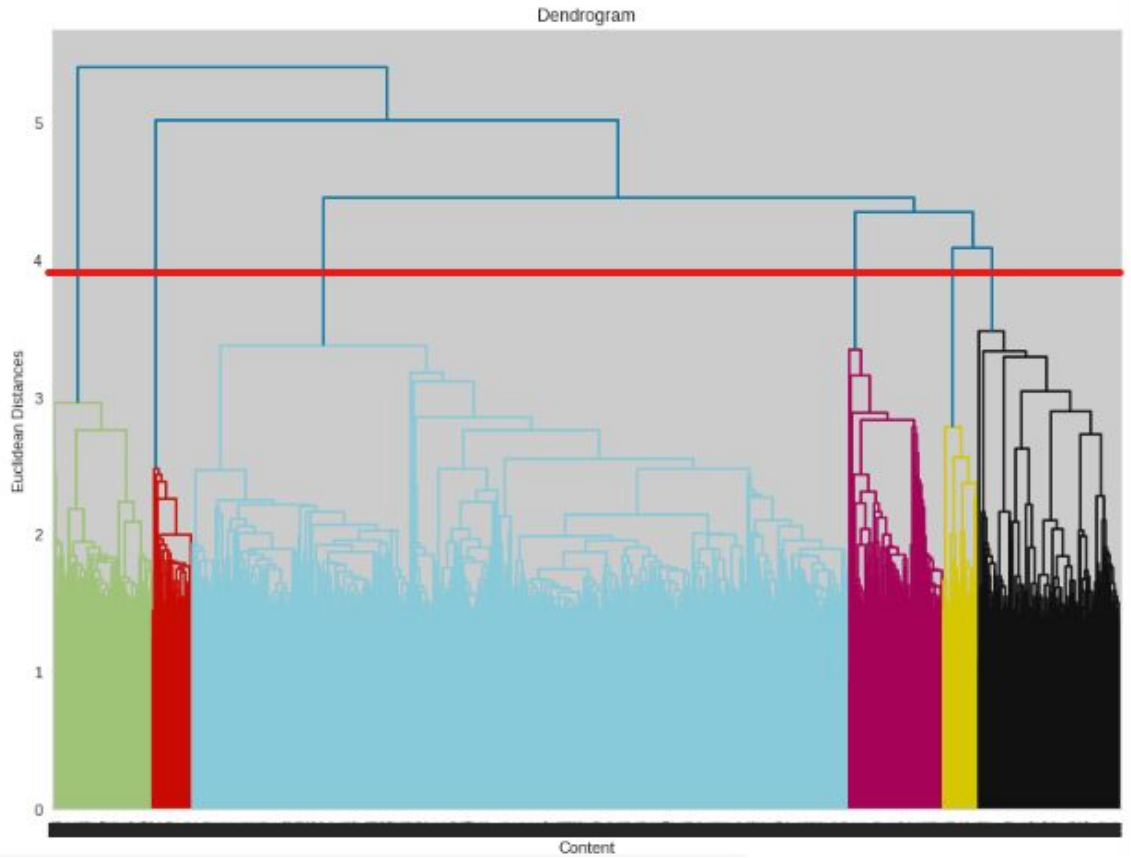
It is the ratio of the sum of **between-clusters dispersion** and of **inter-cluster dispersion** for all clusters.



Finding Optimal number of clusters

04 Dendrogram

is a tree-like chart that shows the sequences of merges or splits of clusters.



Clusters

KModes

Clusters

cluster : 0
Documentaries Documentaries

cluster : 1
Dramas InternationalMovies

cluster : 2
Children FamilyMovies

cluster : 3
Children FamilyMovies

cluster : 4
InternationalTVShows TVDramas

cluster : 5
Stand UpComedy

Recommendation systems

01

Collaborative filtering

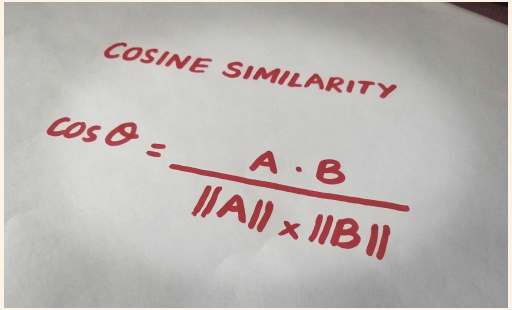
02

Content Based Filtering

Recommendation Using Cosine Similarities

- Cosine similarity is a metric used to measure how similar two items are. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The output value ranges from 0–1. 0 means no similarity, where as 1 means that both the items are 100% similar.

- The python Cosine Similarity or cosine kernel, computes similarity as the normalized dot product of input samples X and Y. We will use the sklearn cosine_similarity to find the cos θ for the two vectors in the count matrix.



Conclusion

- The Data set contains 7787 rows and 12 columns. There are missing values in columns director, cast, date added and release_year. In order to not lose important information we have replaced the missing values with “ ” using the .fillna(“ ”) method.
- There are two types of content: TV shows and Movies (30.86% contains TV shows and 69.14% contains Movies) meaning there are more movies than TV shows on Netflix.
- By analyzing the date added and release year column with respect to content types it can be observed that over the years Netflix is focusing more on 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 States have the maximum content on Netflix
- International Movies make up the top most genre, and the most of the content is added during the months of October to January.
- The words that occurred most of the time in the ‘title’ column are LOVE, MAN, WORLD, CHRISTMAS. We can infer from the above words that there are more movies/tv shows of the romantic genre and also Christmas movies/tv shows.
- The most number of content in the NETFLIX were directed by "Jan Suter", followed by "Raul Campos", "Marcus Roby
- India has the most number of actors whose names come under top 20 movies actors count. India produces the most movies or TV shows across the globe.
- The duration of the movies are about 70 to 120 mins. And most shows on Netflix are of 1 season.
- Looking at methods of finding optimal clusters like the Elbow Method, Dendrogram, Silhouette Method, we could see that the clusters of 6 are optimal.
- Taking the number of clusters as 6 we applied different clustering models for instance Kmeans, K-Modes. Further we went ahead to apply Hierarchical Agglomerative clustering on data and we got the best cluster arrangements.
- We also labeled the clusters using 6 as the optimal number of clusters. For the Kmode algorithm the 14 rating variables were clustered properly, each cluster comprising mostly one kind of rated content except cluster 1 and 6 comprising Movies and Tv Shows of various ratings. Under K Means algorithm the clusters are: Cluster 1 -Documentaries and Musical Documentaries, Cluster 2- Dramas and International Movies, Cluster 3- Children and Family Movies, Cluster 4-Children and Family Movies, Cluster 5-International Tv Shows and Tv Dramas, Cluster 6-Stand Up Comedy and Talk Shows.