CAPSTONE PROJECT- 4

NETFLIX MOVIES AND TV SHOWS CLUSTERING

Presented by:

Minal Kharbade
Deveshya Gupta

CONTENT

- Problem Statement
- Data Description
- Exploratory Data Analysis
- Data processing
- Clustering Methods
- Clustering Models
- Conclusion

PROBLEM STATEMENT

- Dataset consist of TV shows and movies available on Netflix as of 2019.
- In 2018 they released a interesting report which shows that number of TV shows on netflix has nearly tripled since 2010.
- The streaming services no of movies has decreased by more than 2000 titles since 2010 , while its no of Tv shows has nearly tripled. It will be interesting to explore what all the insights can be obtained from the same dataset.
- In this project we are required to do

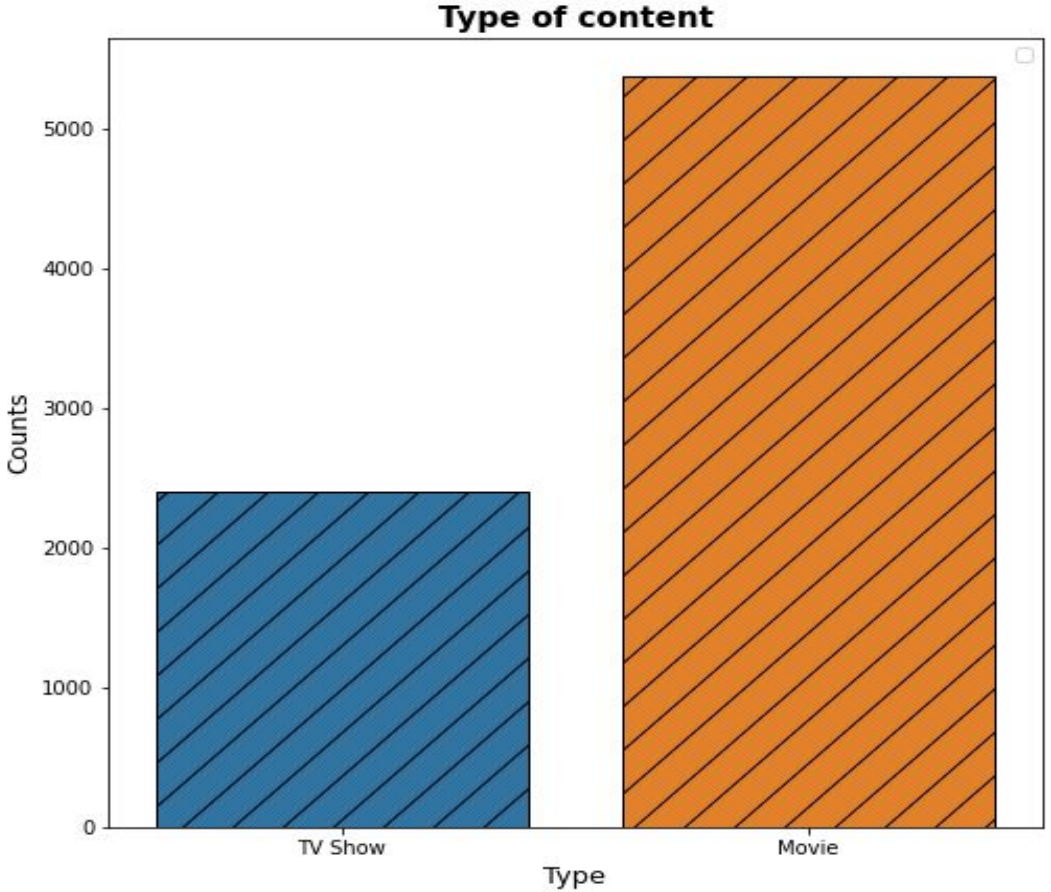
1. Exploratory Data Analysis
2. Understanding what type of content is available in different countries
3. Is Netflix has increasingly focusing on TV rather than movies in recent years?
4. Clustering similar content by matching text based features

DATA DESCRIPTION

- 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 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 : Genere
- description: The Summary description

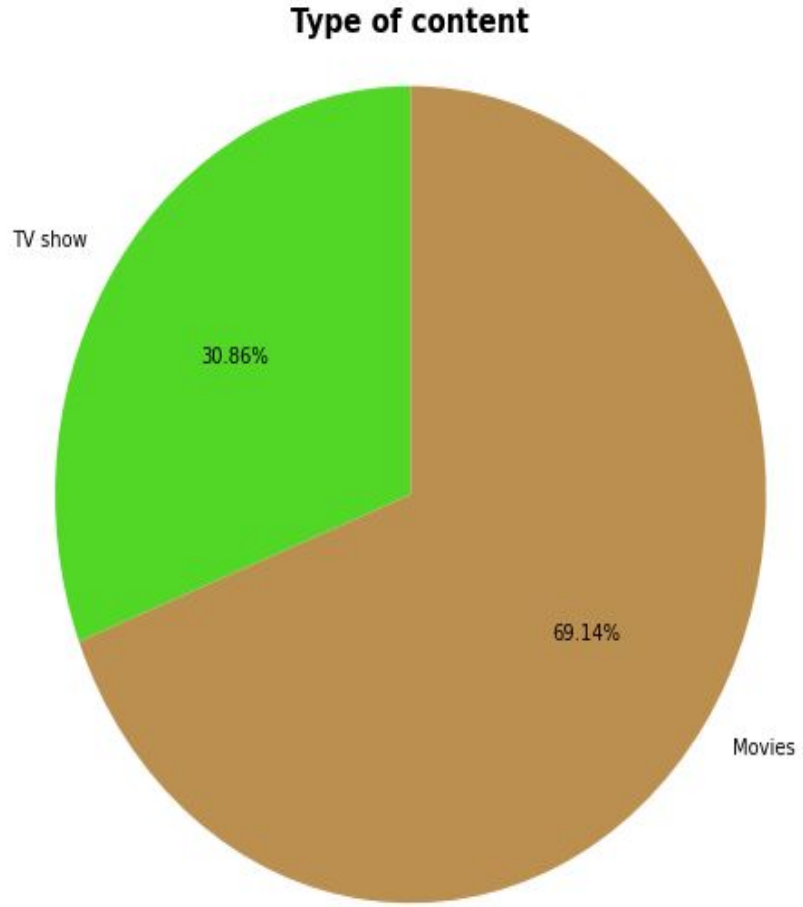
EXPLORATORY DATA ANALYSIS(EDA)

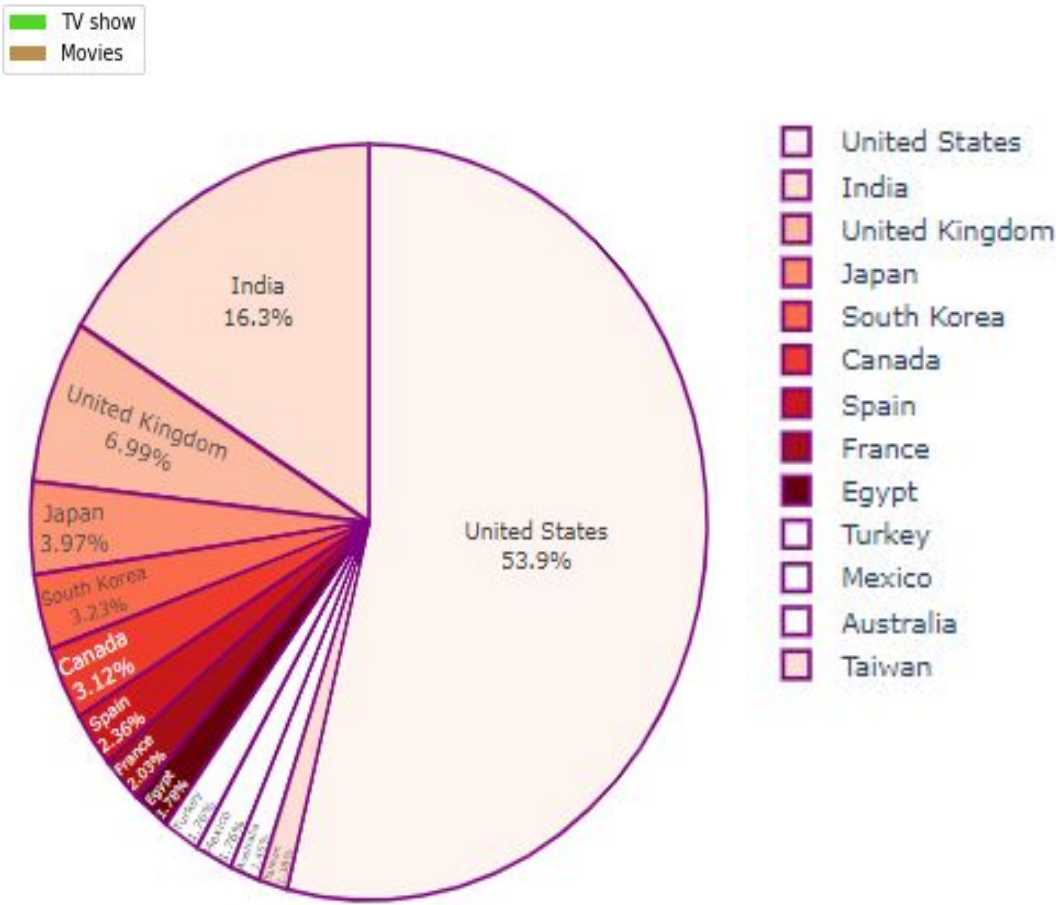
- **Analysing what type of content present in Netflix:**



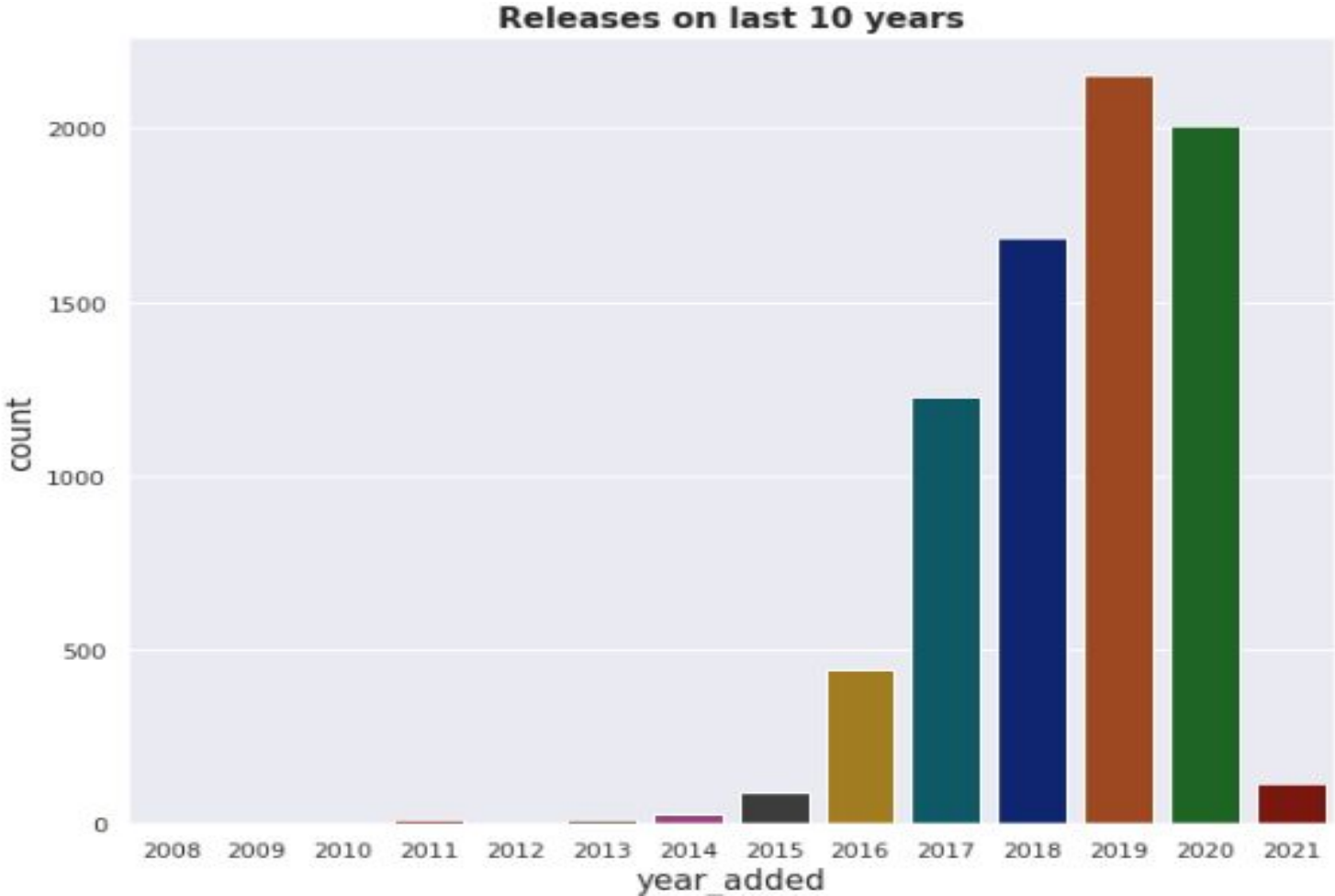
From this chart it is clearly visible that movies are more present in Netflix.

❑ Different type of content present in Netflix and different countries:

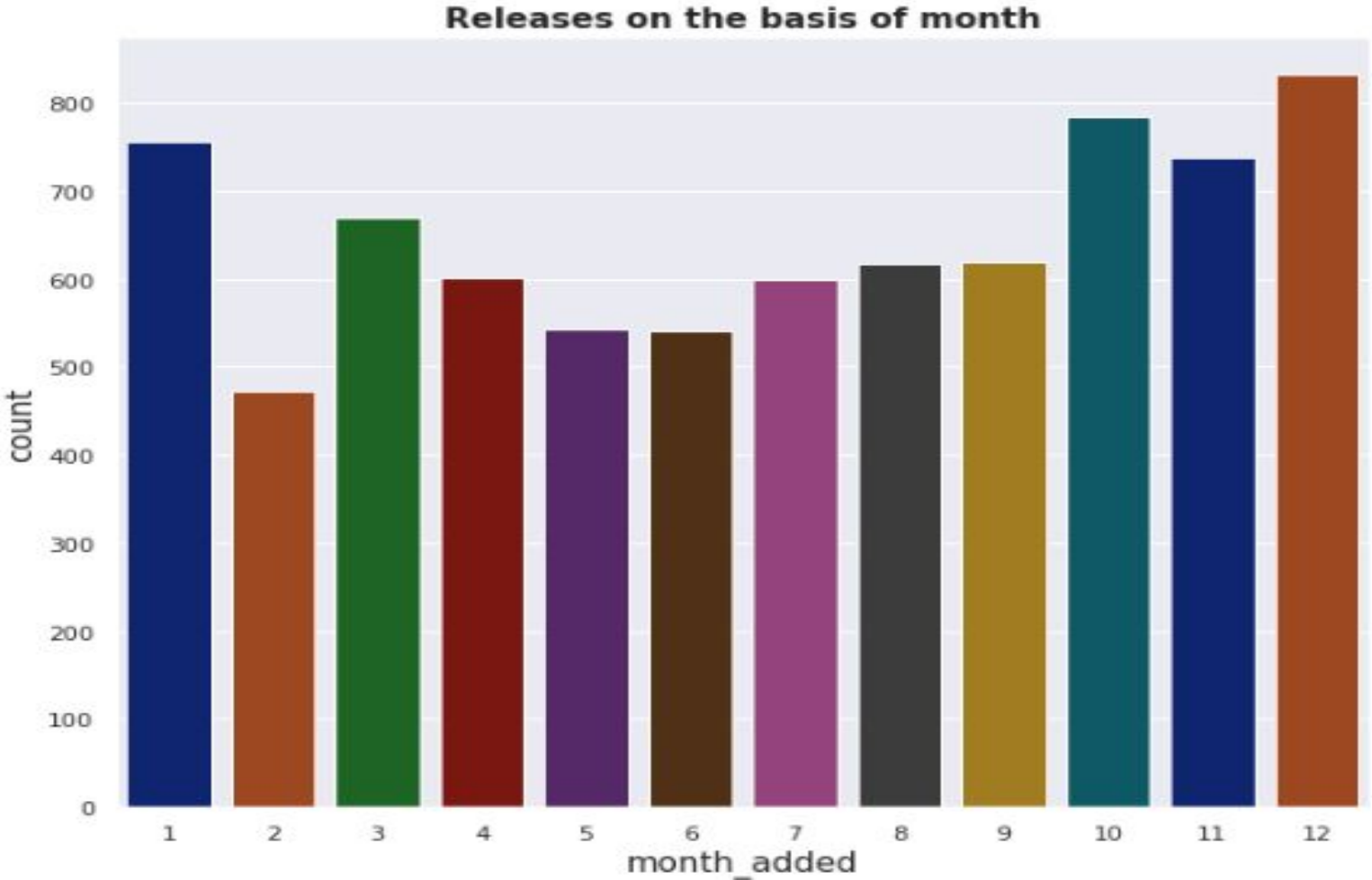




❑ Year wise Analysis:

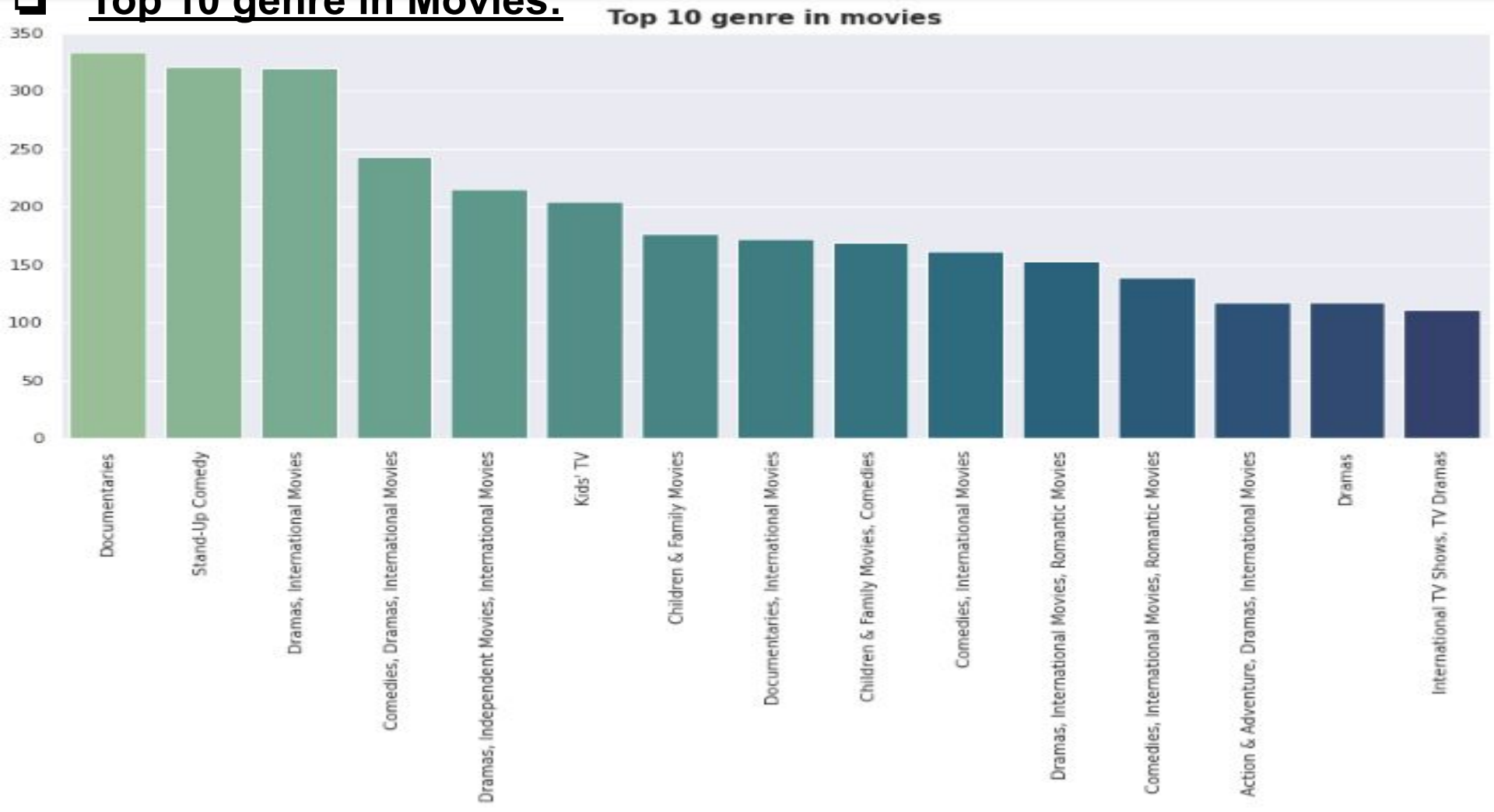


❑ Month wise Analysis:

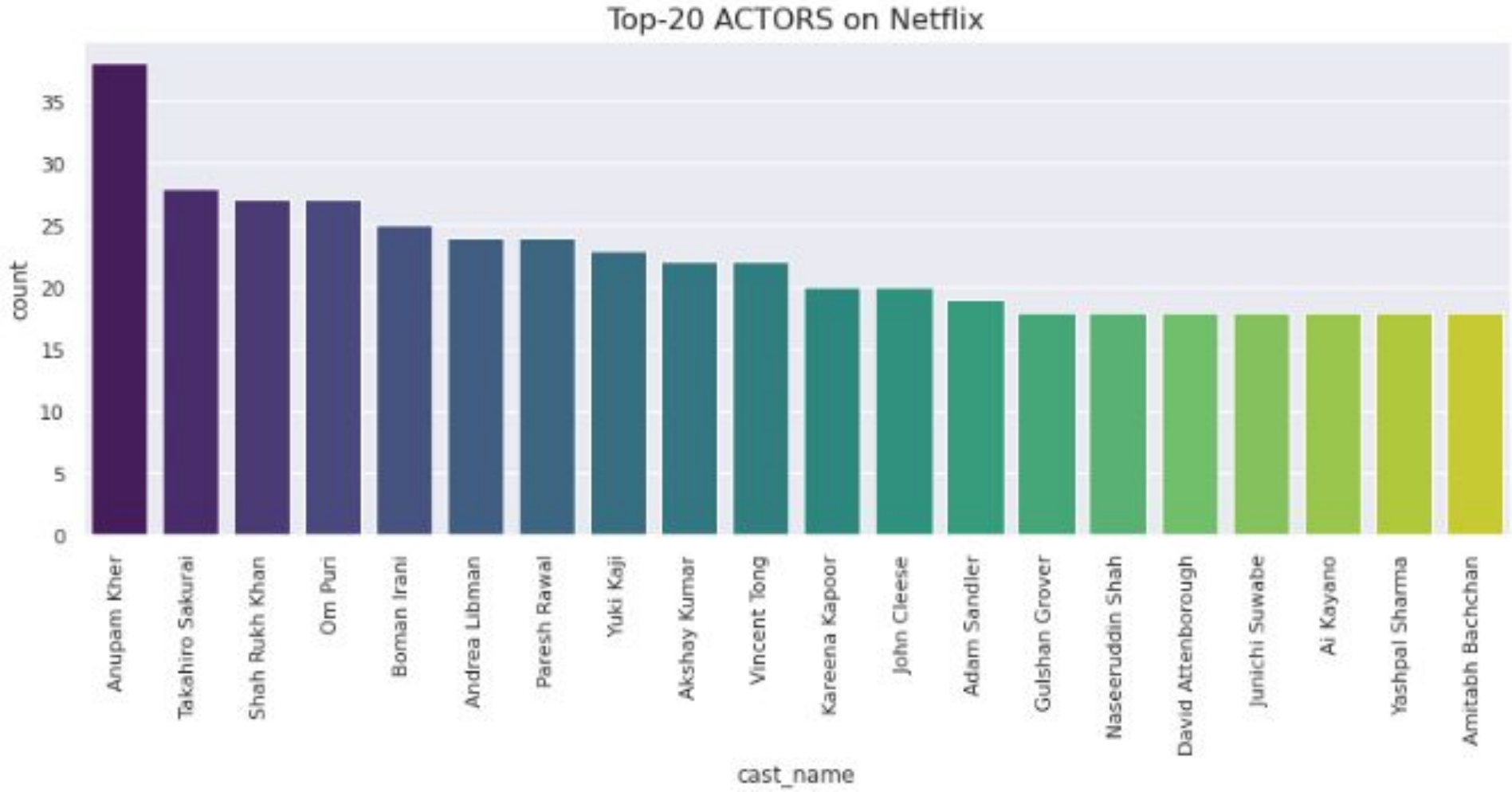


❑ Type and Rating:

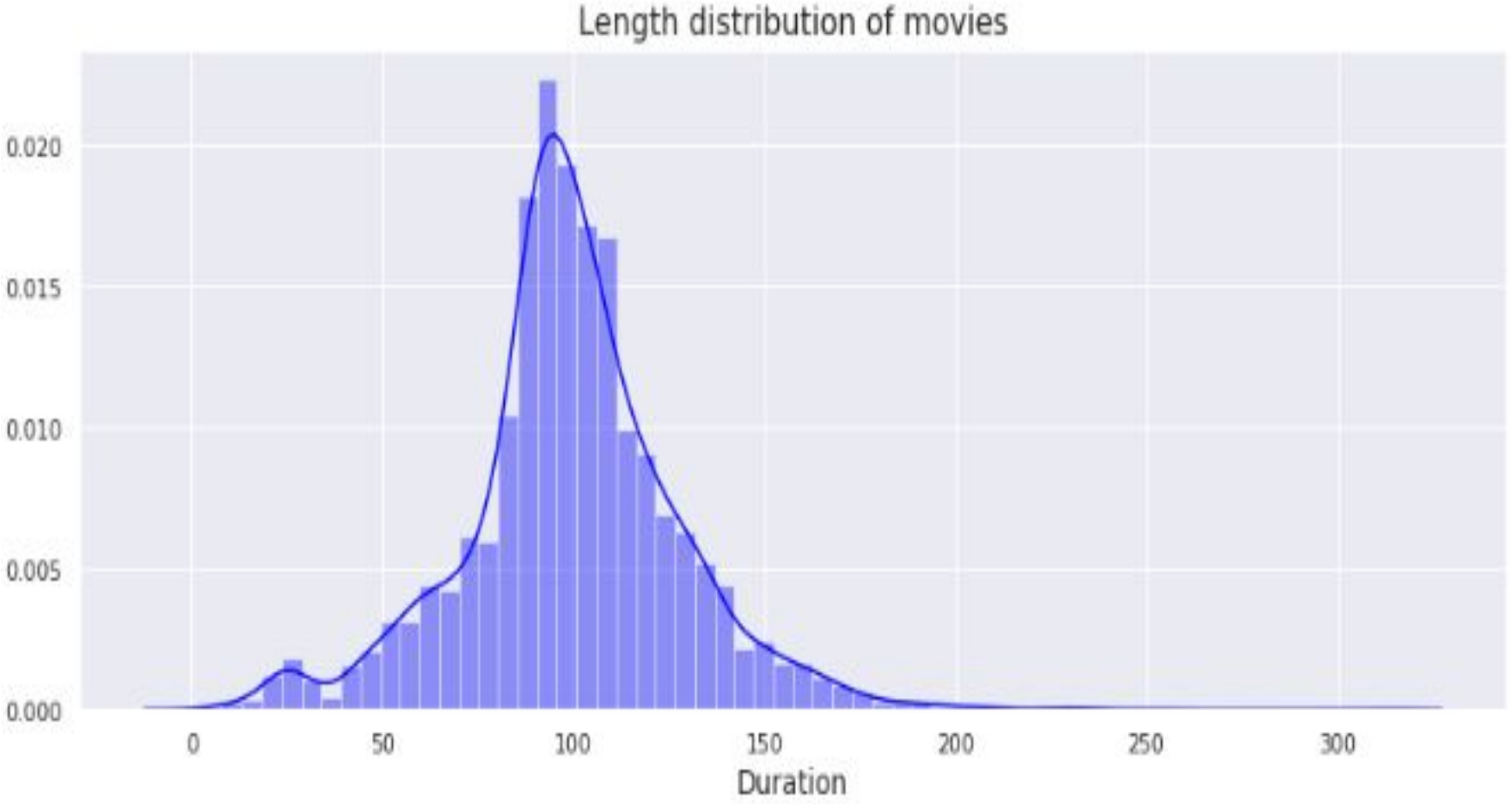
❑ Top 10 genre in Movies:



❑ Top 20 actors on Netflix:



❑ Length distribution of movies:



❑ Topic Modeling:

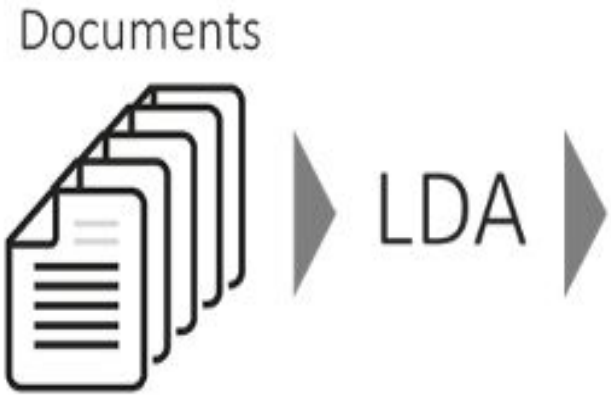
Latent Semantic Analysis (LSA)

LSA stands for latent semantic analysis is one of the foundational technique used in topic modeling .Latent semantic analysis is a natural language processing method that analyzes relationship between set of documents and term contain within. The core idea is to take matrix of documents and term and try to decompose it into separate into two matrices

- Document topic matrix
- Document term matrix

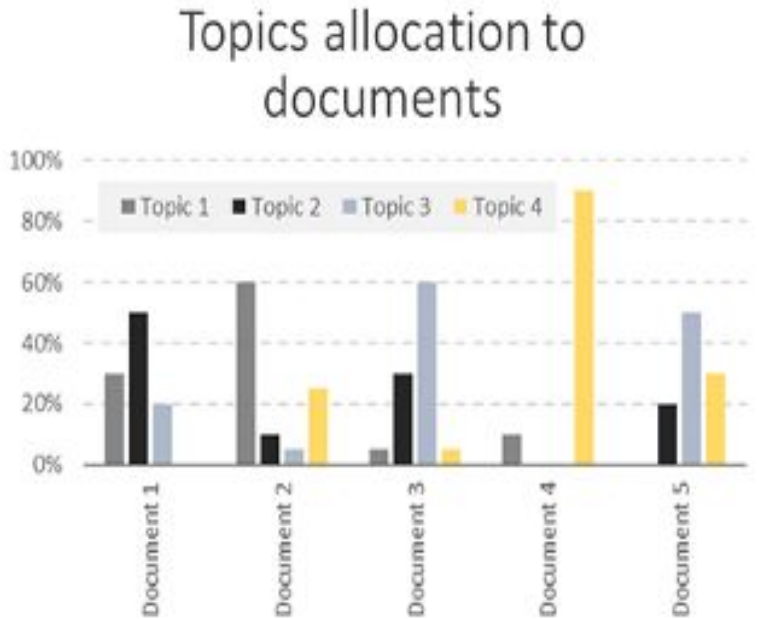
Latent Dirichlet Allocation (LDA)

LDA is general probabilistic model that assumes each topic is mixture over and underlying set of words and each document is a mixture of over set of probabilities.



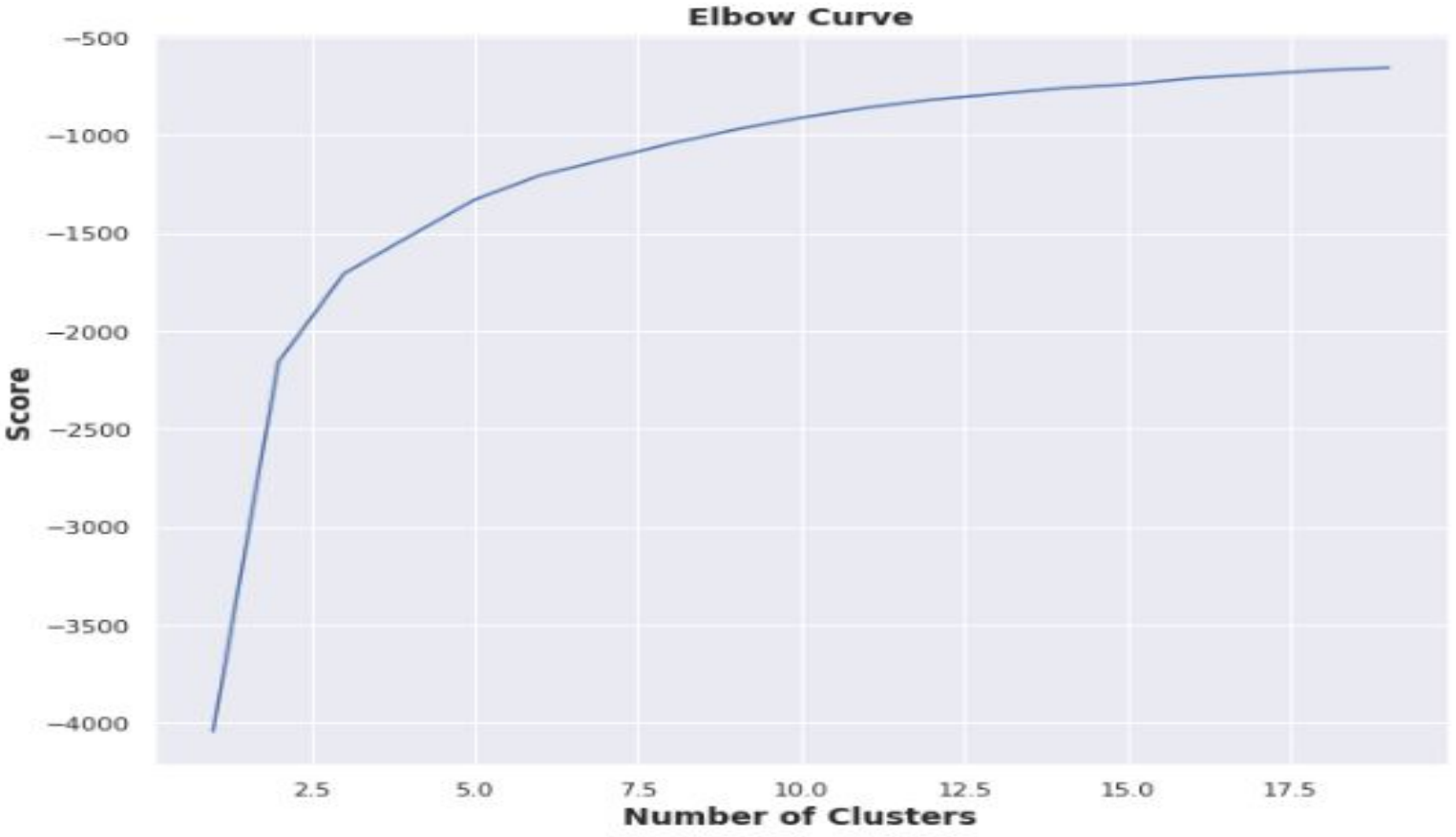
Creation of topics

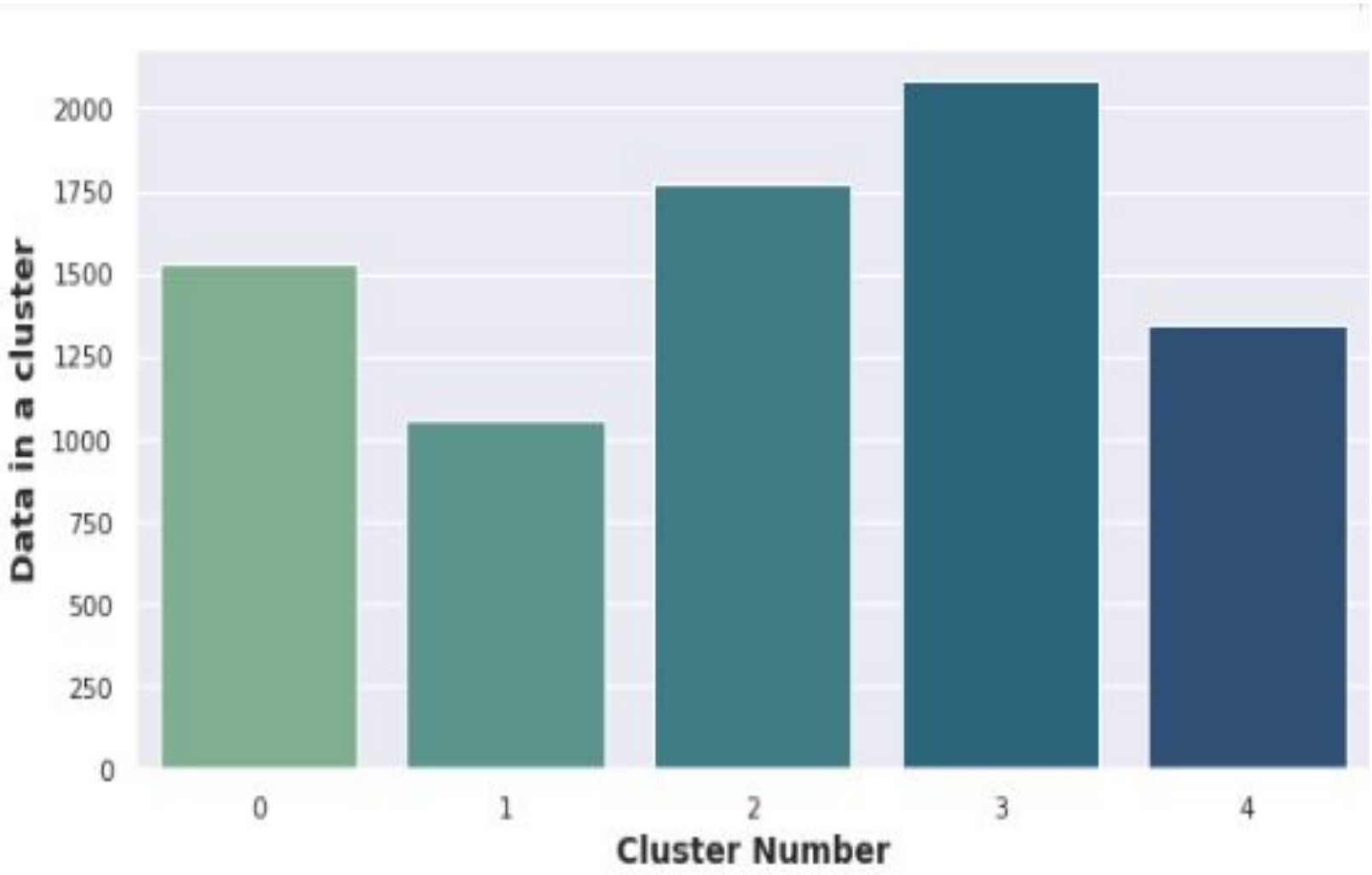
	weight	words
Topic 1	3%	flower
	2%	rose
	1%	plant
...		
Topic 2	2%	company
	1%	wage
	1%	employee



❑ K-mean Clustering:

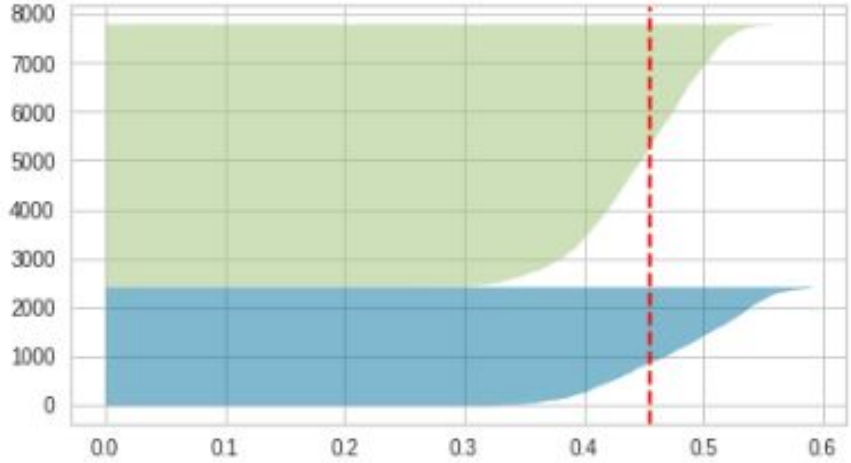
Elbow method:

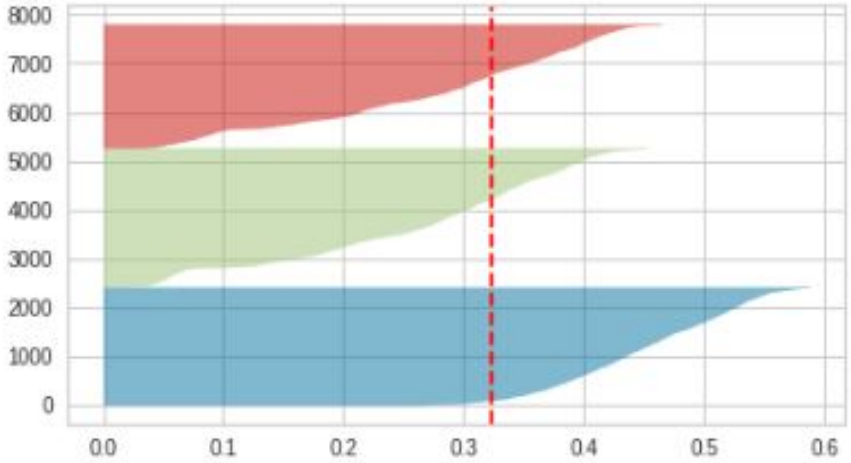


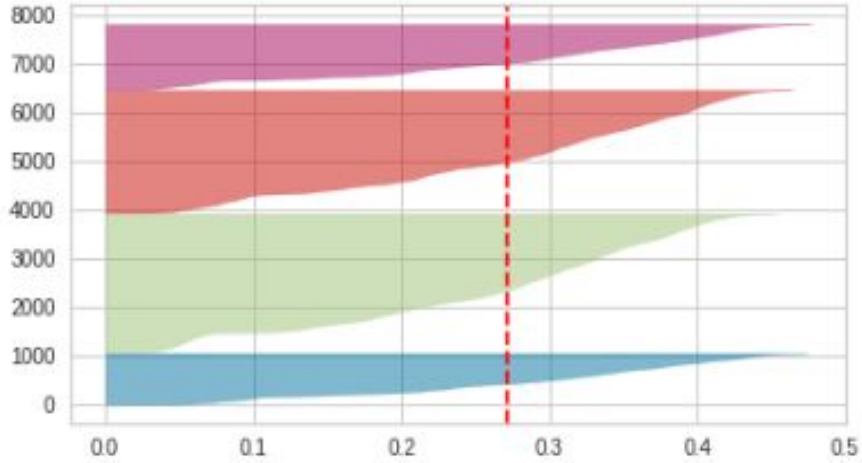

Data in a cluster
2000
1750
1500
1250
1000
750
500
250
0
0
1
2
3
4
Cluster Number

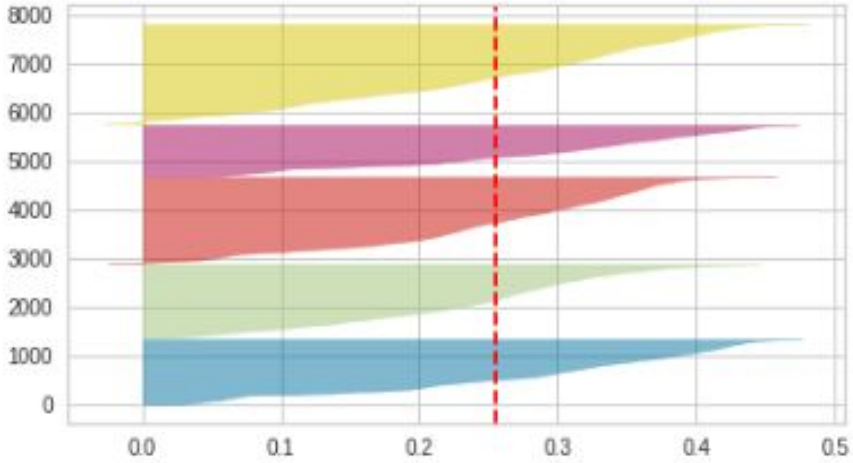

clusters
segment k means
0
1
2
3
4
comp 1
comp 2
0
50
100
150
200
250
300
0
100
200
300
400
500

Silhouette method:









CONCLUSION

- From the analysis we can say that there are 70% movies and 30% TV shows on Netflix.
- US has 54% content on Netflix and then followed by india with 16%.
- LSA and LDA has sorted much more similar titles in a group of genre.
- Recommendation system works well with description column.
- After applying k-means clustering the optimal value of no of cluster is 5.
- Silhouette score for a set of sample data points is used to measure how dense and well separated cluster are.

THANK YOU