Capstone Project - 4

Netflix Movies and TV shows Clustering

Done By

Vinit Ladse

Pratiksha Kharode

Gaurav Bhakte

Content :

1. Defining problem statement
2. EDA and feature engineering
3. Feature Selection
4. Data Preprocessing
5. Applying different clustering methods
6. Applying Clustering Models
7. Conclusion

Problem Statement:

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 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. 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 has increasingly focusing on TV rather than movies in recent years.
4. Clustering similar content by matching text-based features.

Data Summary:

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.

8.release_year : Actual Release year of the movie / show.

9.rating : TV Rating of the movie / show.

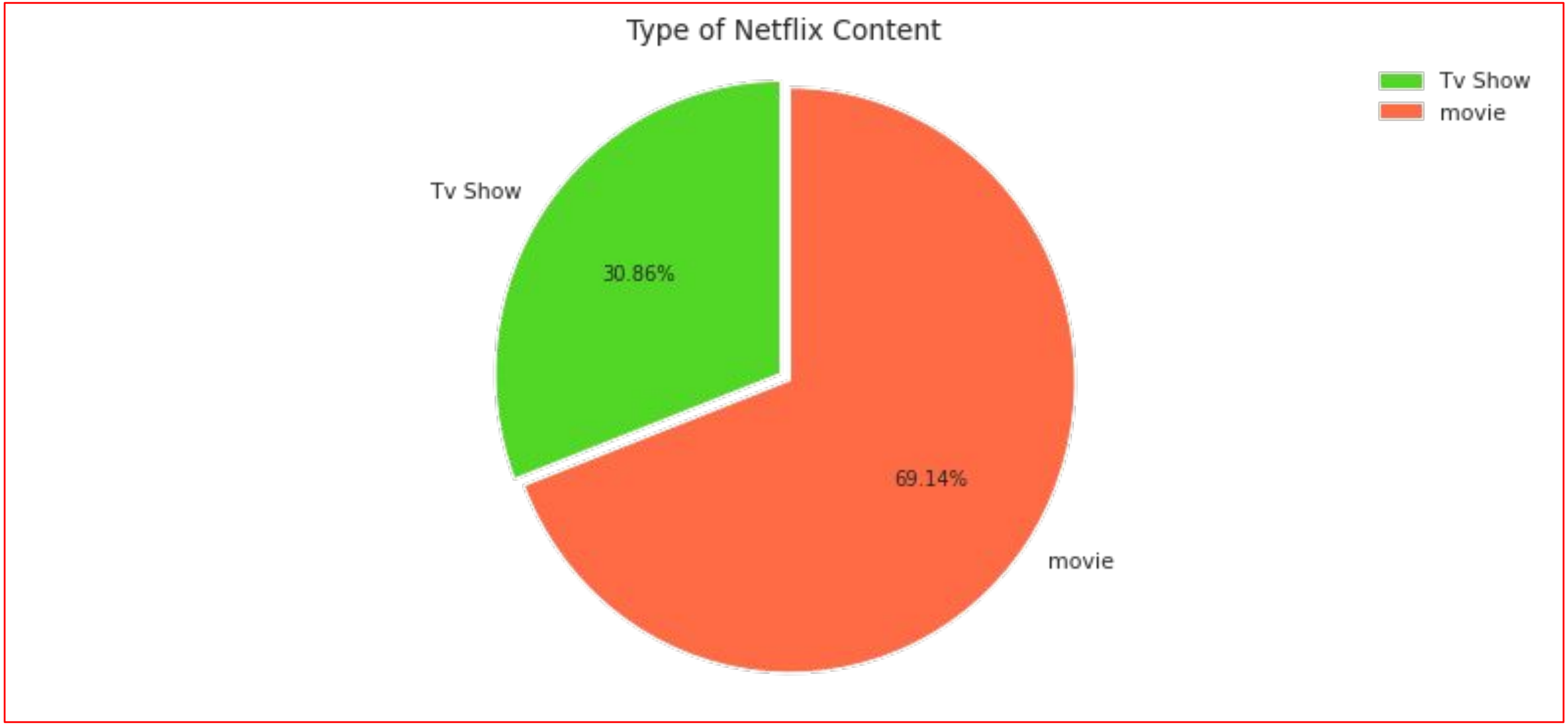
10.duration : Total Duration - in minutes or number of seasons.

11.listed_in : Genres.

12.description: The Summary description.

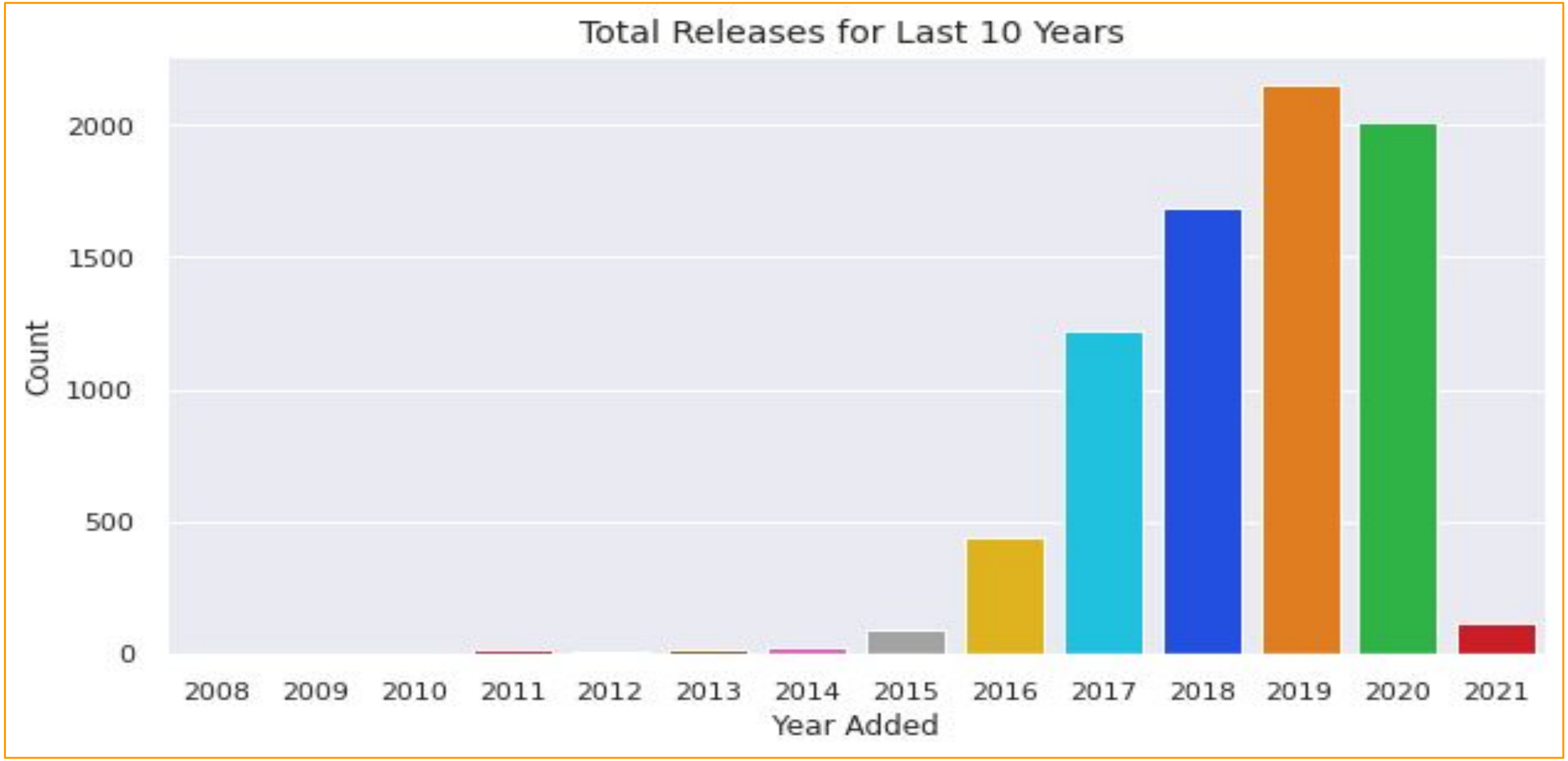
Exploratory Data Analysis:

Different types of content present In the Netflix



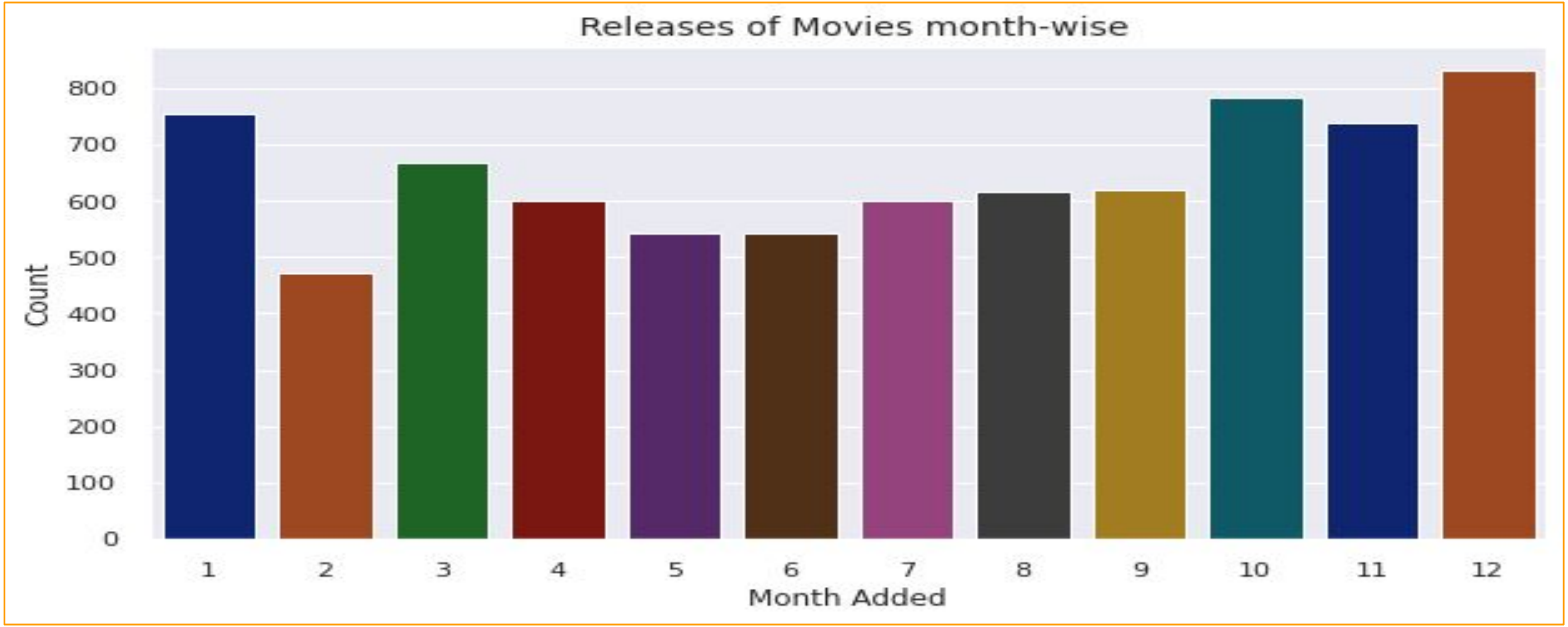
Exploratory Data Analysis:

Year wise Analysis :-

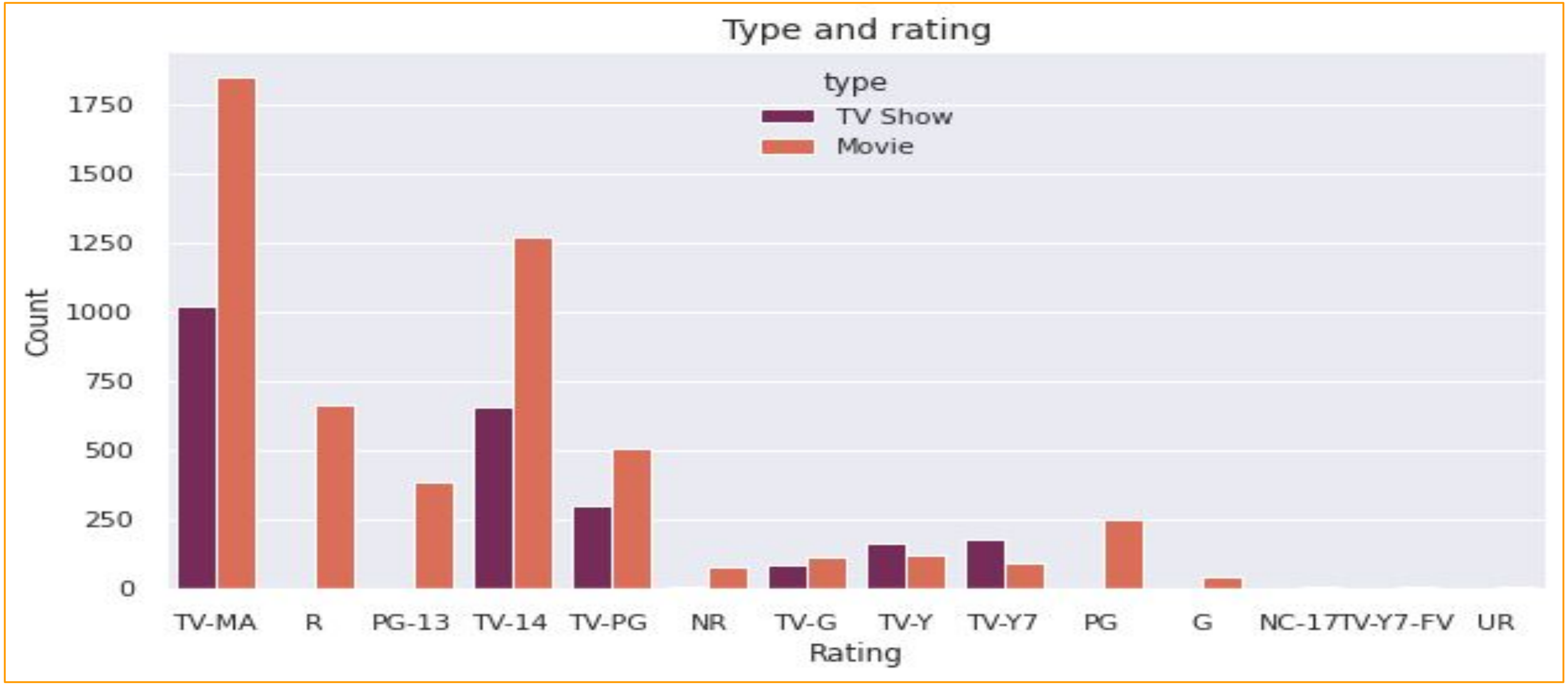


Exploratory Data Analysis:

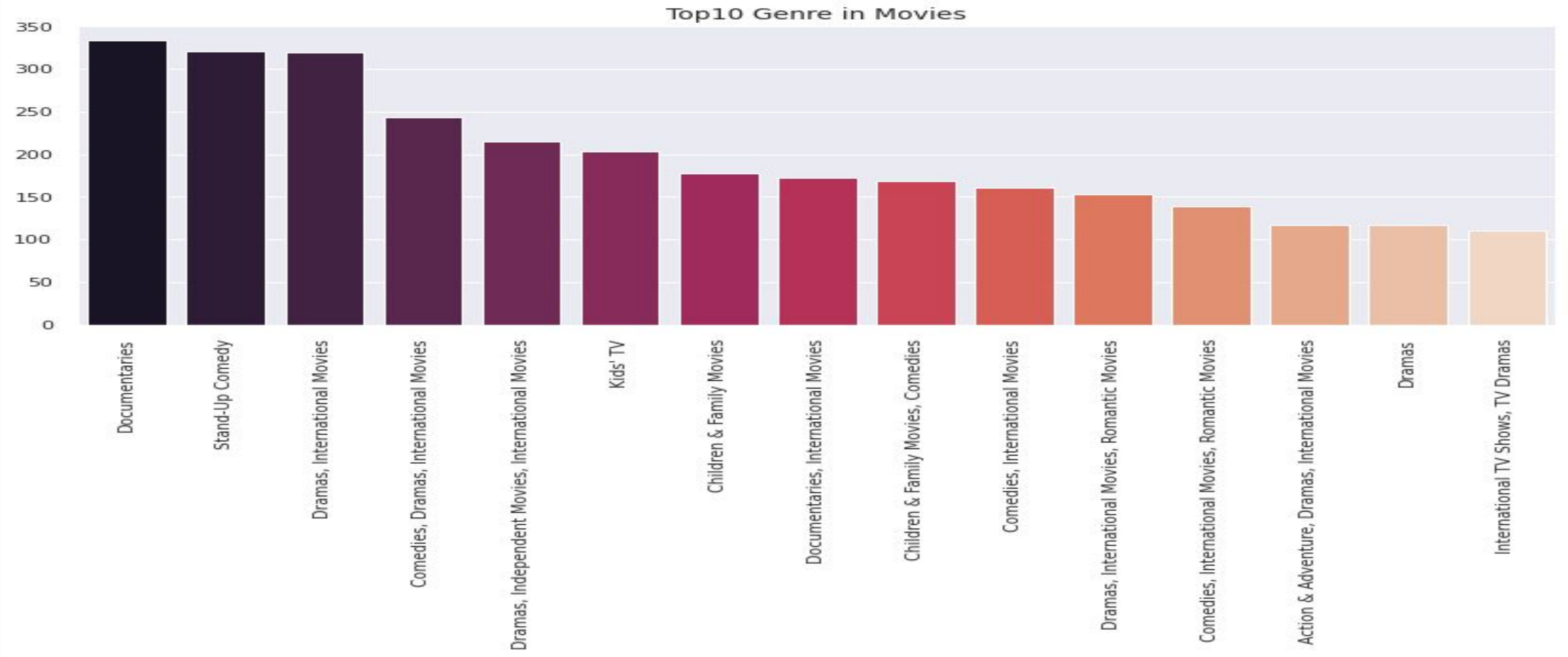
Month wise Analysis :-



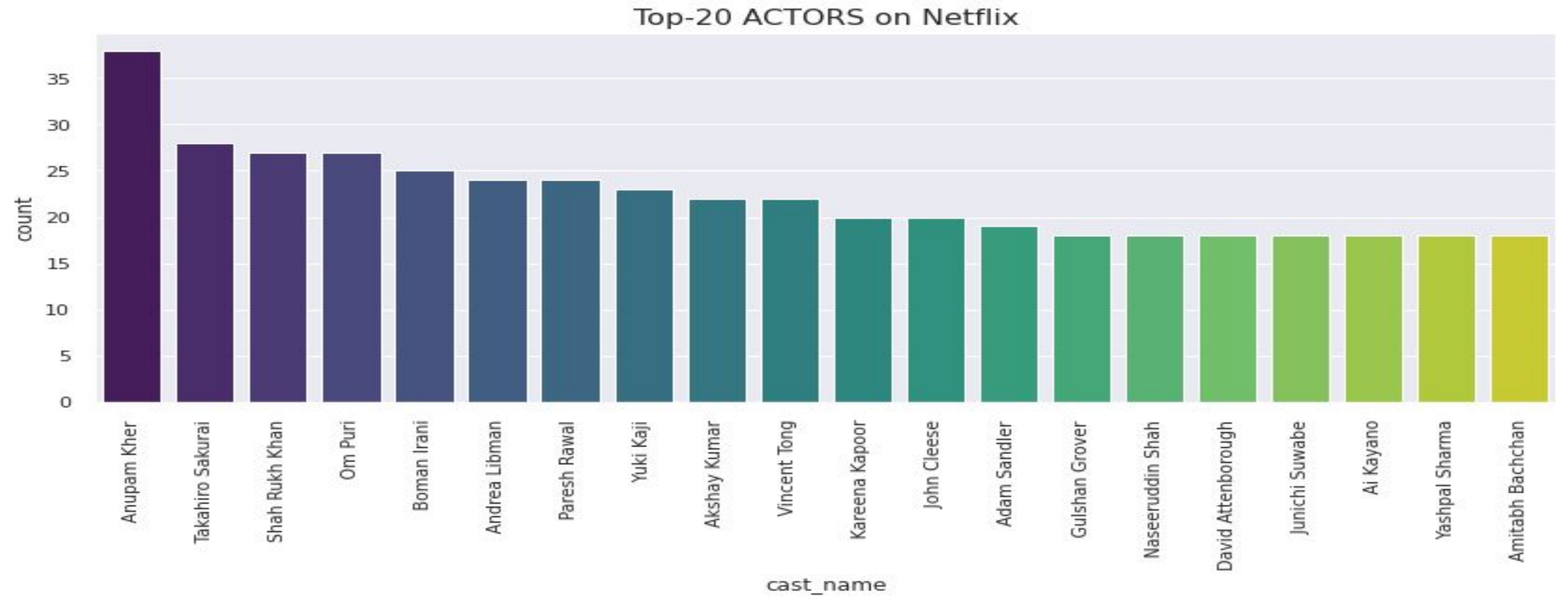
Exploratory Data Analysis:



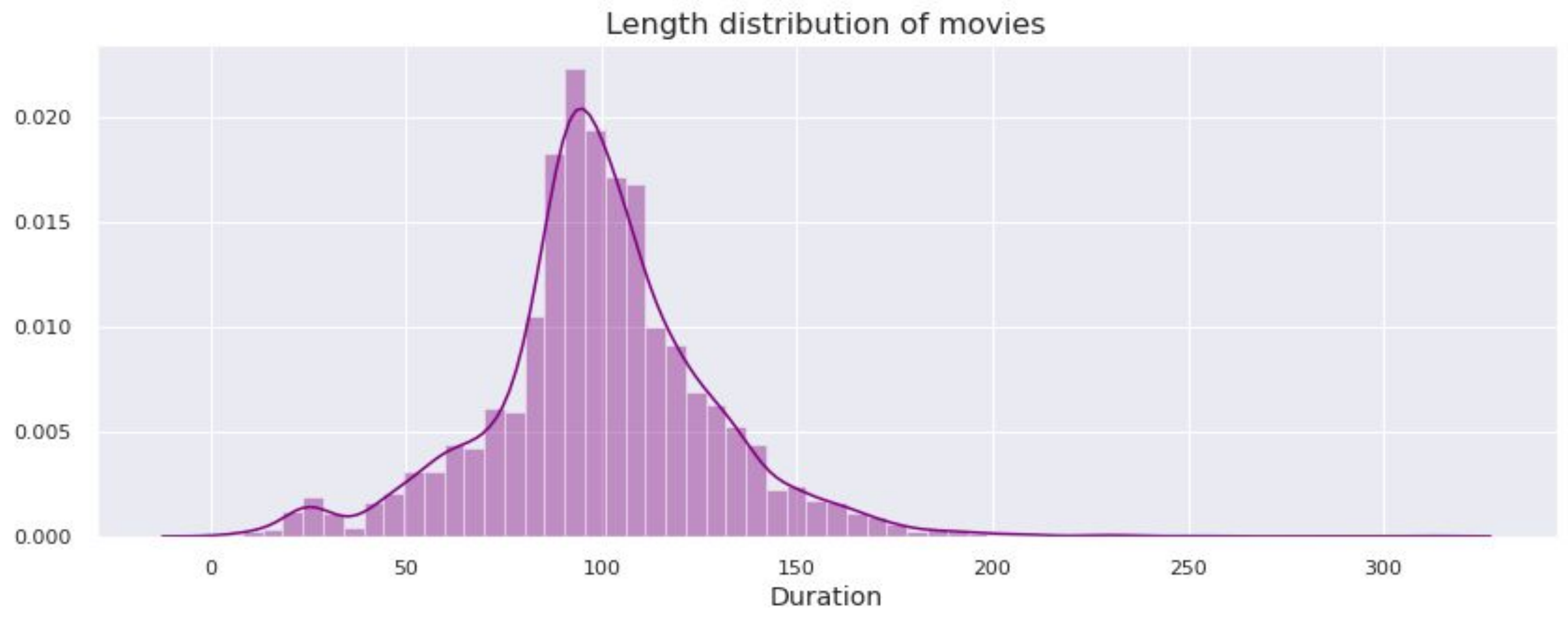
Exploratory Data Analysis:



Exploratory Data Analysis:



Exploratory Data Analysis:



Topic Modeling

Latent Semantic Analysis (LSA)

LSA, which stands for Latent Semantic Analysis, is one of the foundational techniques used in topic modeling. The core idea is to take a matrix of documents and terms and try to decompose it into separate two matrices –

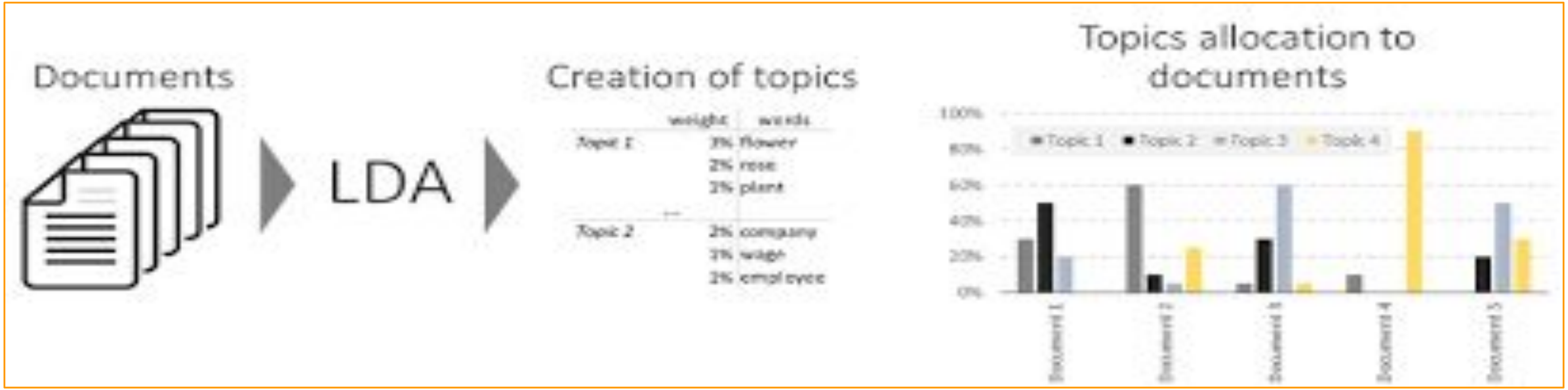
- A document-topic matrix
- A topic-term matrix.

Therefore, the learning of LSA for latent topics includes matrix decomposition on the document-term matrix using Singular value decomposition. It is typically used as a dimension reduction or noise-reducing technique.

Topic Modeling

Latent Dirichlet Allocation(LDA)

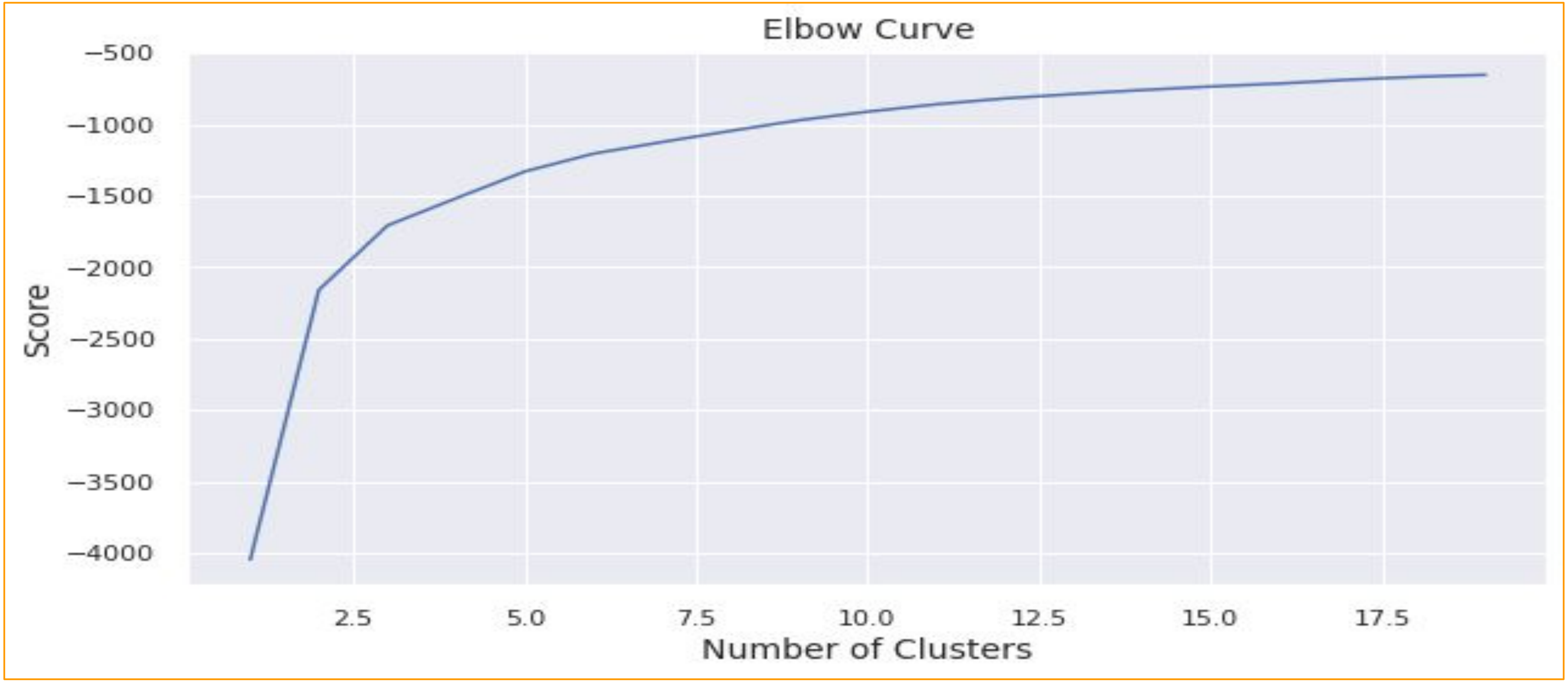
LDA is a generative probabilistic model that assumes each topic is a mixture over an underlying set of words, and each document is a mixture of over a set of topic probabilities.

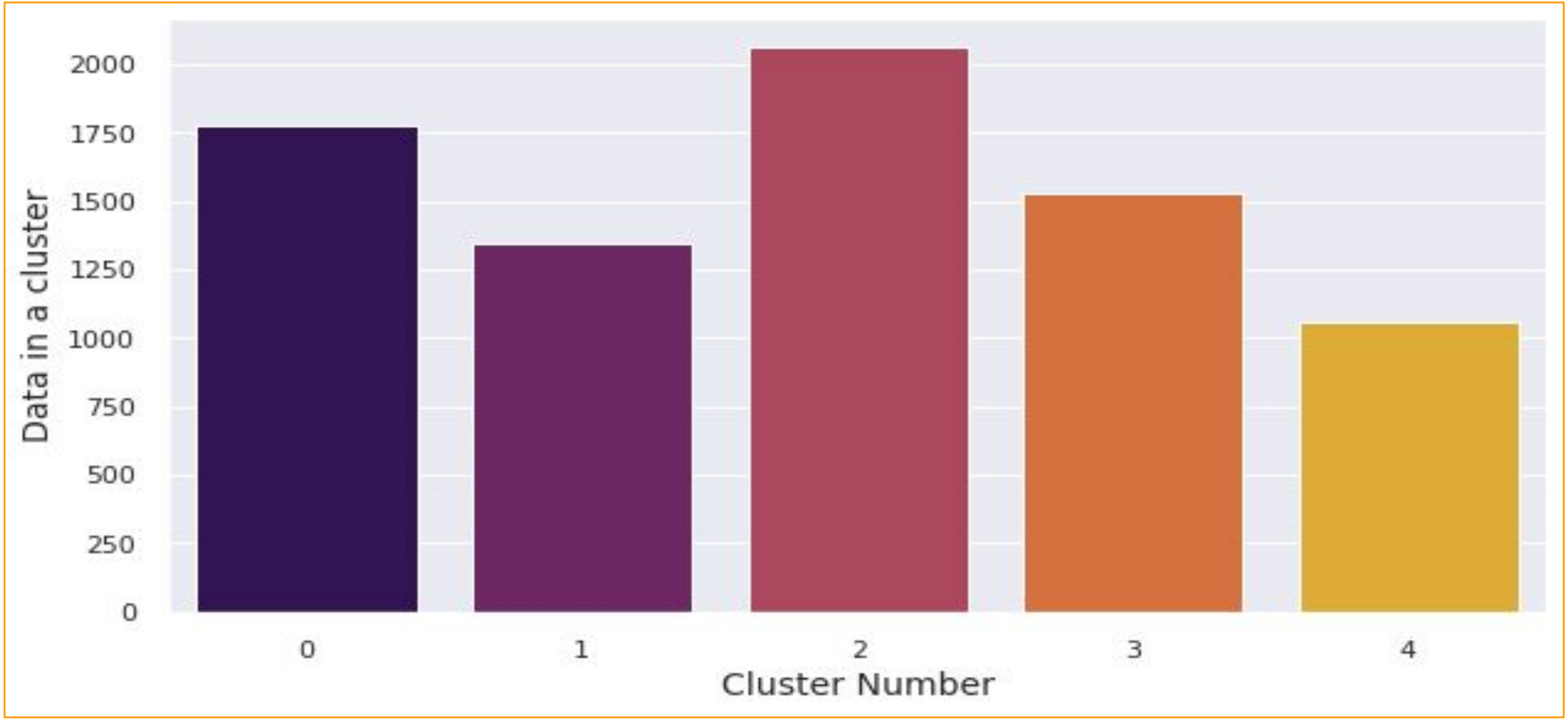


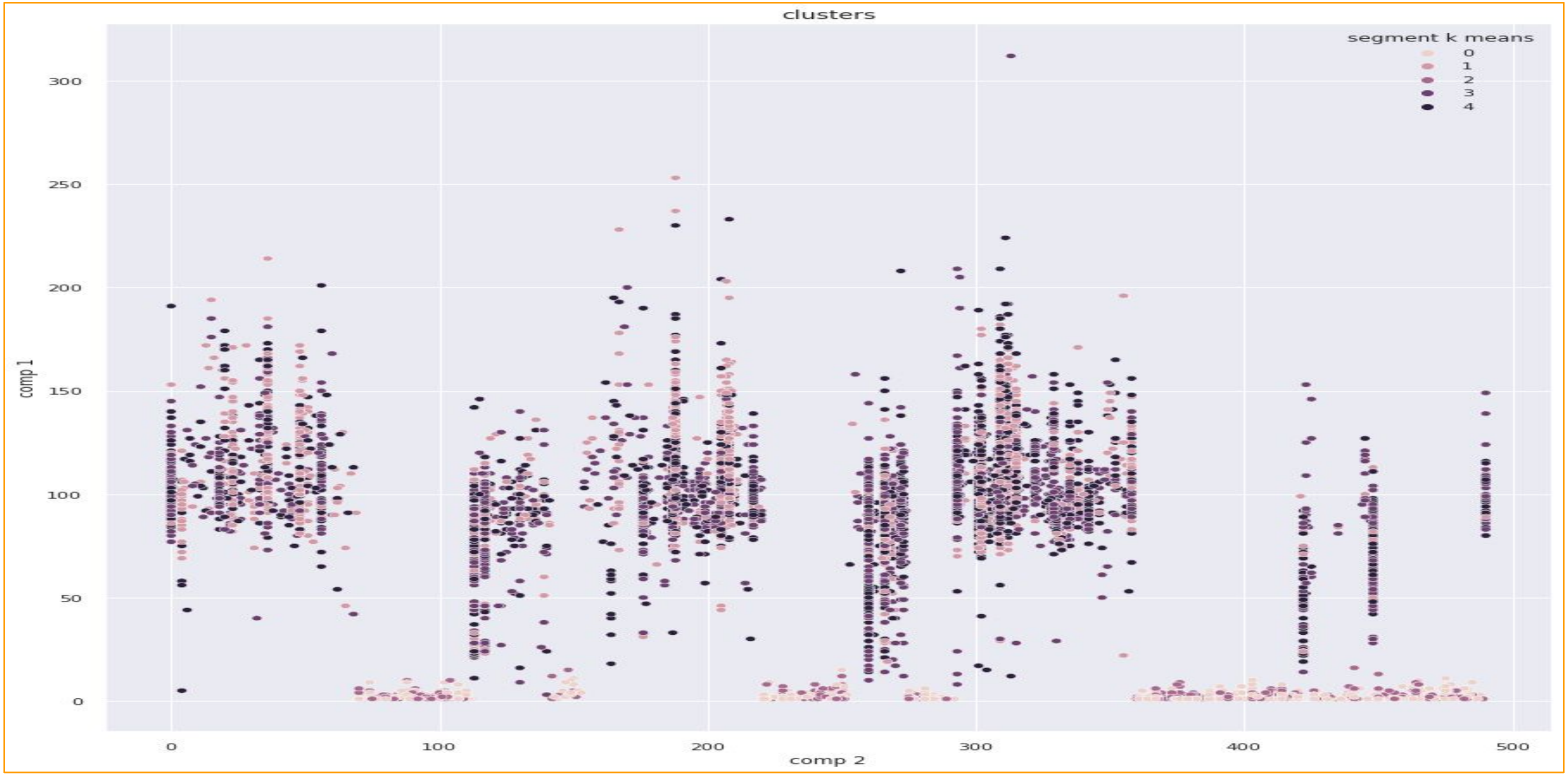
	weight	words
Topic 1	3%	flower
	2%	rose
	2%	plant
Topic 2	2%	company
	3%	wage
	2%	employee

K- Mean Clustering:

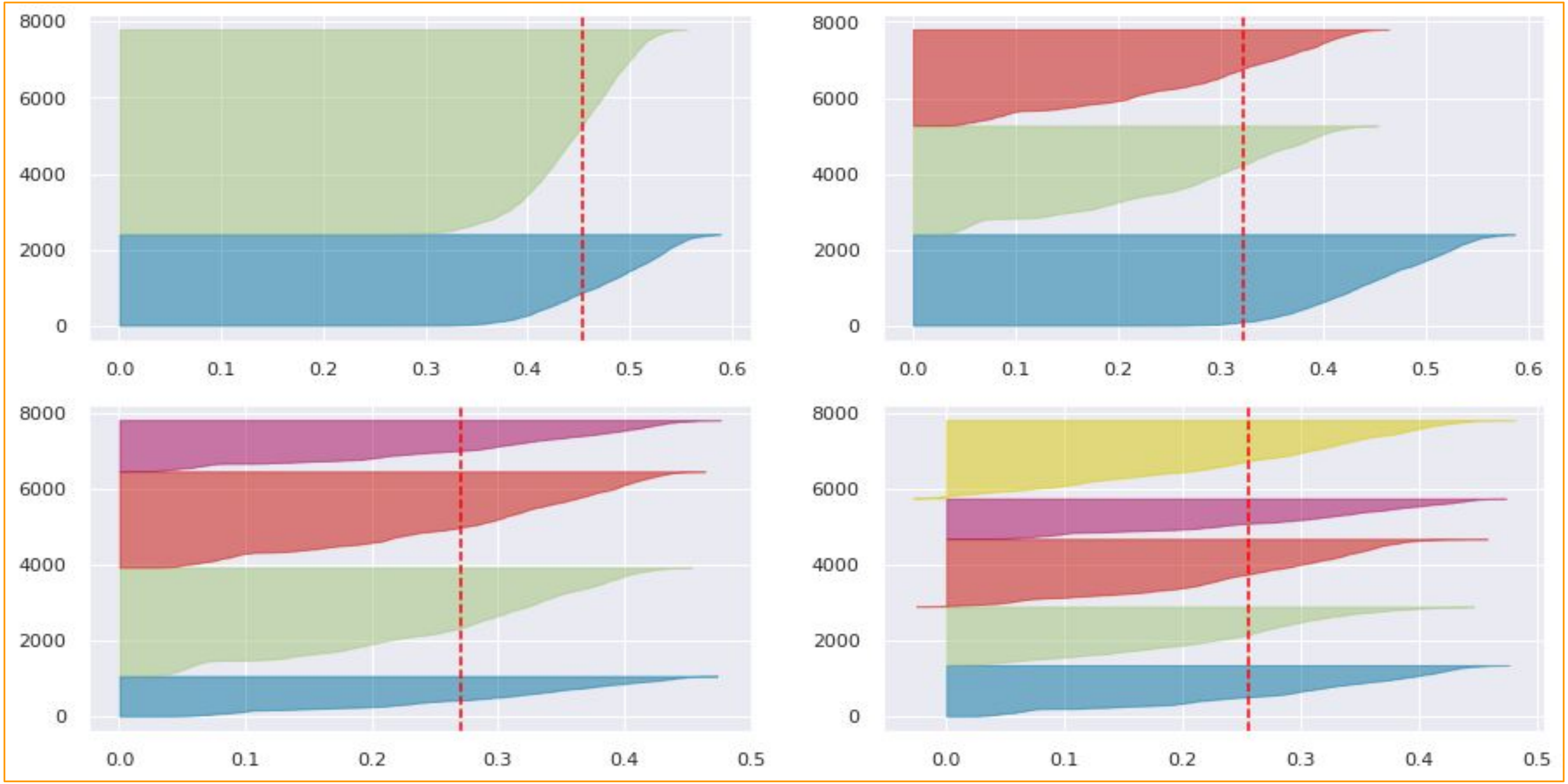
Elbow Method:




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


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

Silhouette Method



Conclusion:

- There are about 70% movies and 30% TV shows on Netflix.
- The United States has the highest number of content on Netflix by a huge margin followed by India.
- LDA and LSA has sorted much more similar titles in a group of genre.
- Recommendation system works perfectly well with description column.
- After applying K - means optimal value of number of clusters is 5.
- Silhouette score for a set of sample data points is used to measure how dense and well-separated the clusters are.

Thank You