Capstone Project-4

NETFLIX MOVIES AND TV SHOWS CLUSTERING

Kartika Sharma

Netflix and chill !

- Problem Statement
- EDA
- Data Pre-processing
- Topic modelling
- Recommendation system
- K- Means Clustering
- Cluster Analysis

Too many choices

- 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 : 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

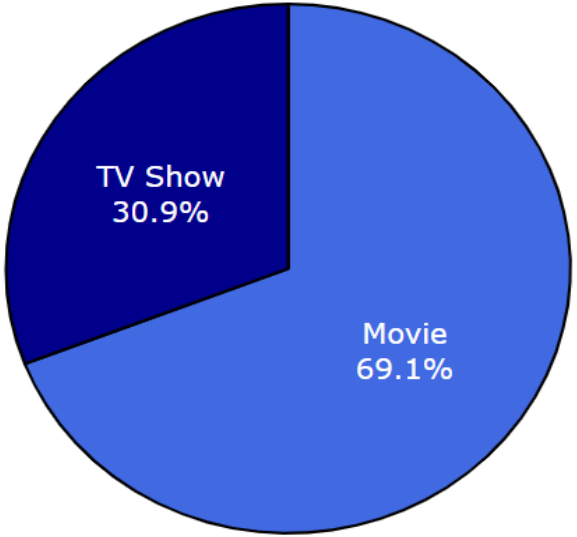
Data Summary

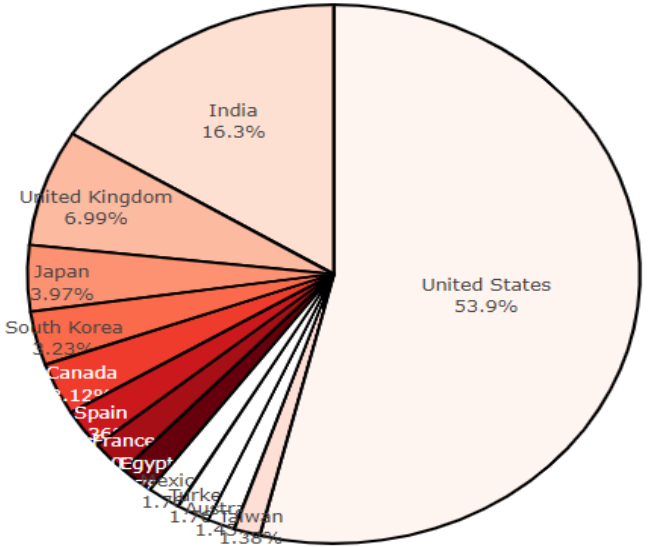
- date_added : Date it was added on Netflix
- release_year : Actual Releaseyear of the movie / show
- rating : TV Rating of the movie / show
- duration : Total Duration - in minutes or number of seasons
- listed_in : Genere
- description: The Summary description

Null Values

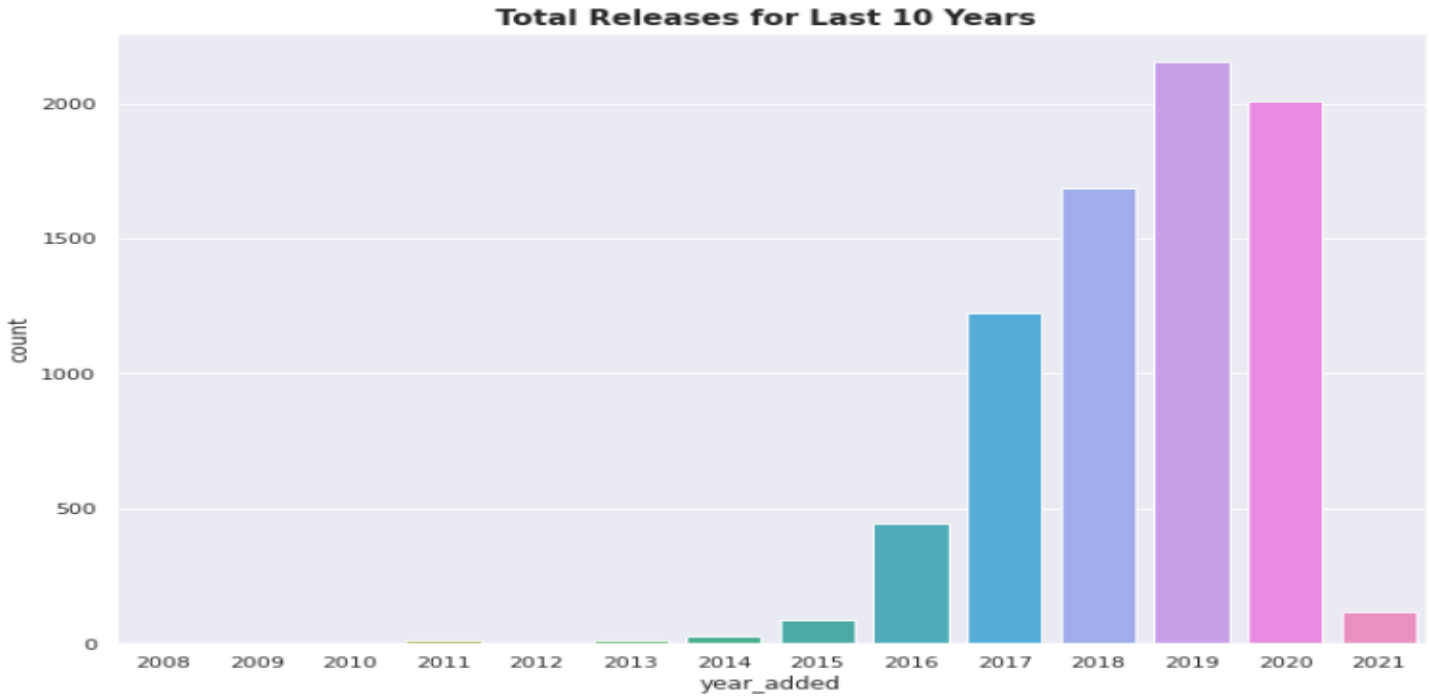
- Delete if more than 8% (cast and Director)
- Delete rows with missing date (<0.5%)
- Fill categoricals with frequent values

EDA

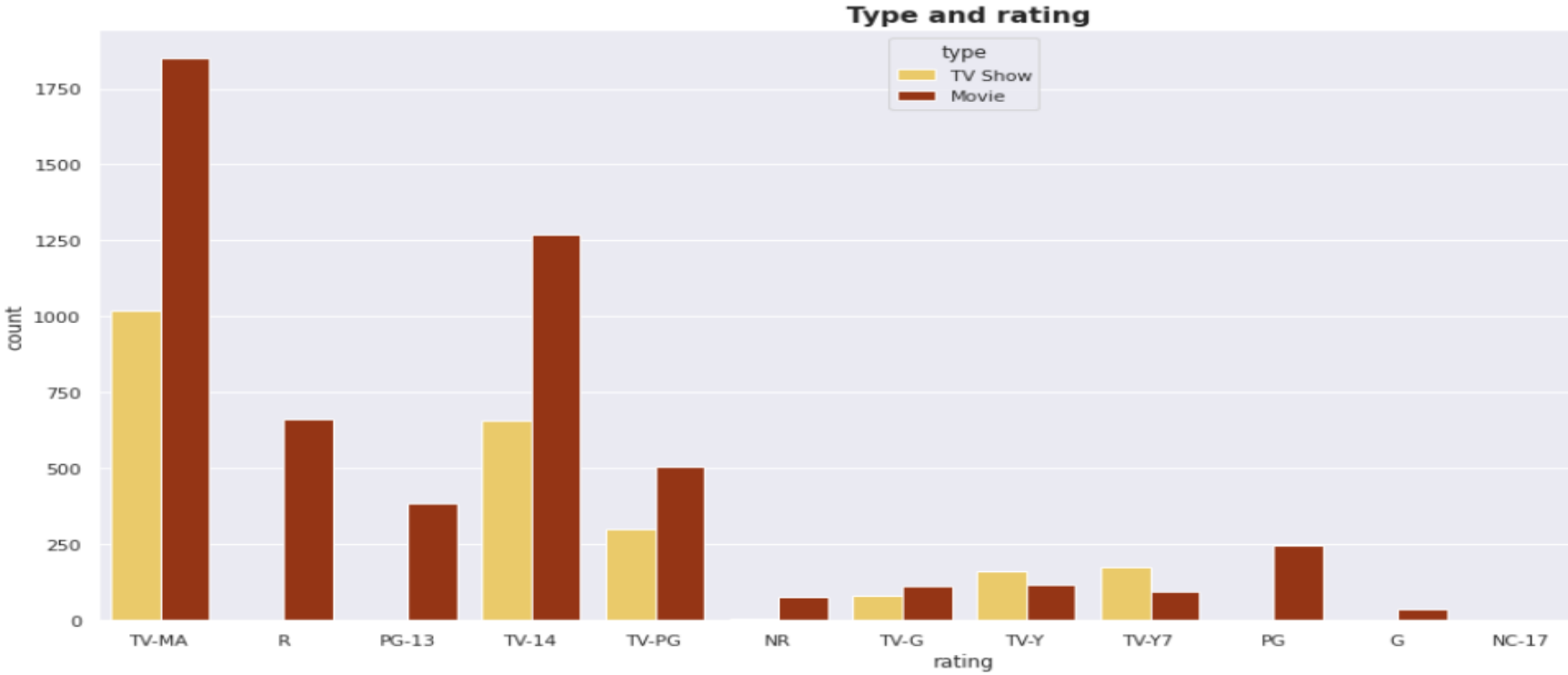

TV Show
30.9%
Movie
69.1%


India
16.3%
United Kingdom
6.99%
Japan
3.97%
United States
53.9%
Canada
Spain
France
Egypt
Taiwan
1.38%

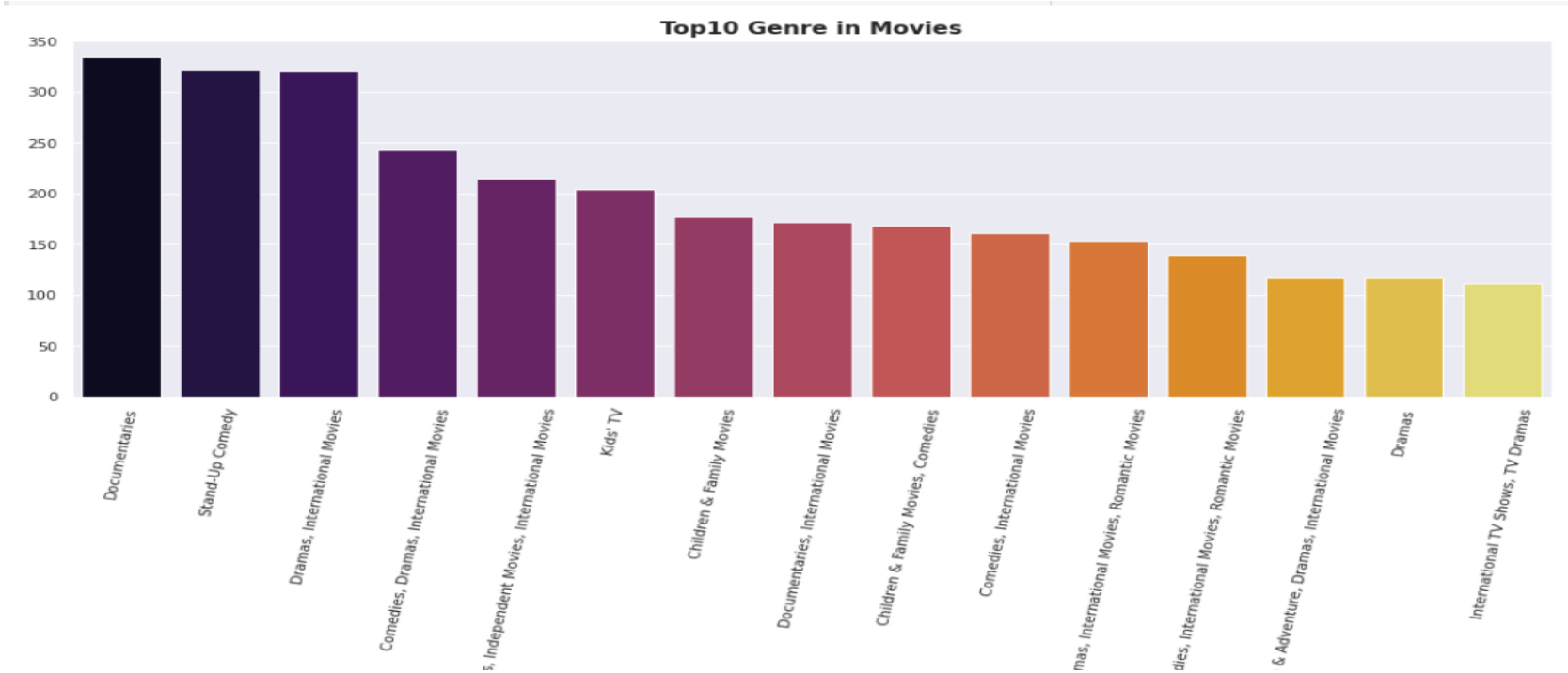
EDA


Total Releases for Last 10 Years
count
2000
1500
1000
500
0
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
year_added

EDA


Type and rating
type
TV Show
Movie
count
1750
1500
1250
1000
750
500
250
0
TV-MA
R
PG-13
TV-14
TV-PG
NR
TV-G
TV-Y
TV-Y7
PG
G
NC-17
rating

EDA)


Top10 Genre in Movies
350
300
250
200
150
100
50
0
Documentaries
Stand-Up Comedy
Dramas, International Movies
Comedies, Dramas, International Movies
, Independent Movies, International Movies
Kids' TV
Children & Family Movies
Documentaries, International Movies
Children & Family Movies, Comedies
Comedies, International Movies
mas, International Movies, Romantic Movies
dies, International Movies, Romantic Movies
& Adventure, Dramas, International Movies
Dramas
International TV Shows, TV Dramas

Data Cleaning

- Label Encoding
- Lemmatisation- Lemmatization, unlike Stemming, reduces the inflected words properly ensuring that the root word belongs to the language. In Lemmatization root word is called Lemma. ... For example, runs, running, ran are all forms of the word run, therefore run is the lemma of all these words.
- Removing Stop words - To remove stop words from a sentence, you can divide your text into words and then remove the word if it exits in the list of stop words provided by NLTK.
- Tf - idf Vectorization - TF-IDF stands for “Term Frequency — Inverse Document Frequency”. This is a technique to quantify a word in documents, we generally compute a weight to each word which signifies the importance of the word in the document and corpus. This method is a widely used technique in Information Retrieval and Text Mining.
- Min-max Scaling - For each value in a feature, MinMaxScaler subtracts the minimum value in the feature and then divides by the range. It preserves shape of original distribution.

Topic Modelling (LDA and LSA)

- **Latent Semantic Analysis**(LSA) is used to find the hidden topics represented by the document or text. This hidden topics then are used for clustering the similar documents together. LSA is an unsupervised algorithm and hence we don't know the actual topic of the document.
- In natural language processing, the Latent Dirichlet Allocation (LDA) is **a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar**.

Topic Modelling (LDA and LSA)

NETFLIX Genre 0:
international shows movies dramas comedies romantic family life independent young

NETFLIX Genre 1:
shows crime british spanish language korean docuseries series reality romantic

NETFLIX Genre 2:
adventure action fi sci kids stand family children world series

NETFLIX Genre 3:
stand special comedy comedian comic talk family take show life

NETFLIX Genre 4:
family children movies shows save friend christmas comedieswhen music kids

NETFLIX Genre 5:
documentaries documentary music world docuseries series moviesthis sports life international

NETFLIX Genre 6:
comedies romantic friend kids love life school best high series

NETFLIX Genre 7:
series kids independent docuseries dramas friend science nature anime world

NETFLIX Genre 8:
horror movies fi sci romantic series reality kids thrillersa docuseries

Recommendation

- A recommender system, or a recommendation system (sometimes replacing 'system' with a synonym such as platform or engine), is a subclass of information filtering system that seeks to predict the "rating" or "preference" a user would give to an item.

```
1  print('IF YOU WATCHED CRIMINAL MINDS,YOU WILL LIKE\n\n',recommended_movies_and_shows('Criminal Minds'))
```

```
IF YOU WATCHED CRIMINAL MINDS,YOU WILL LIKE

 4281    Mundeyan Ton Bachke Rahin
1538              Criminal: France
1540               Criminal: Spain
3868                Mahjong Heroes
1303             Chef & My Fridge
Name: title, dtype: object
```

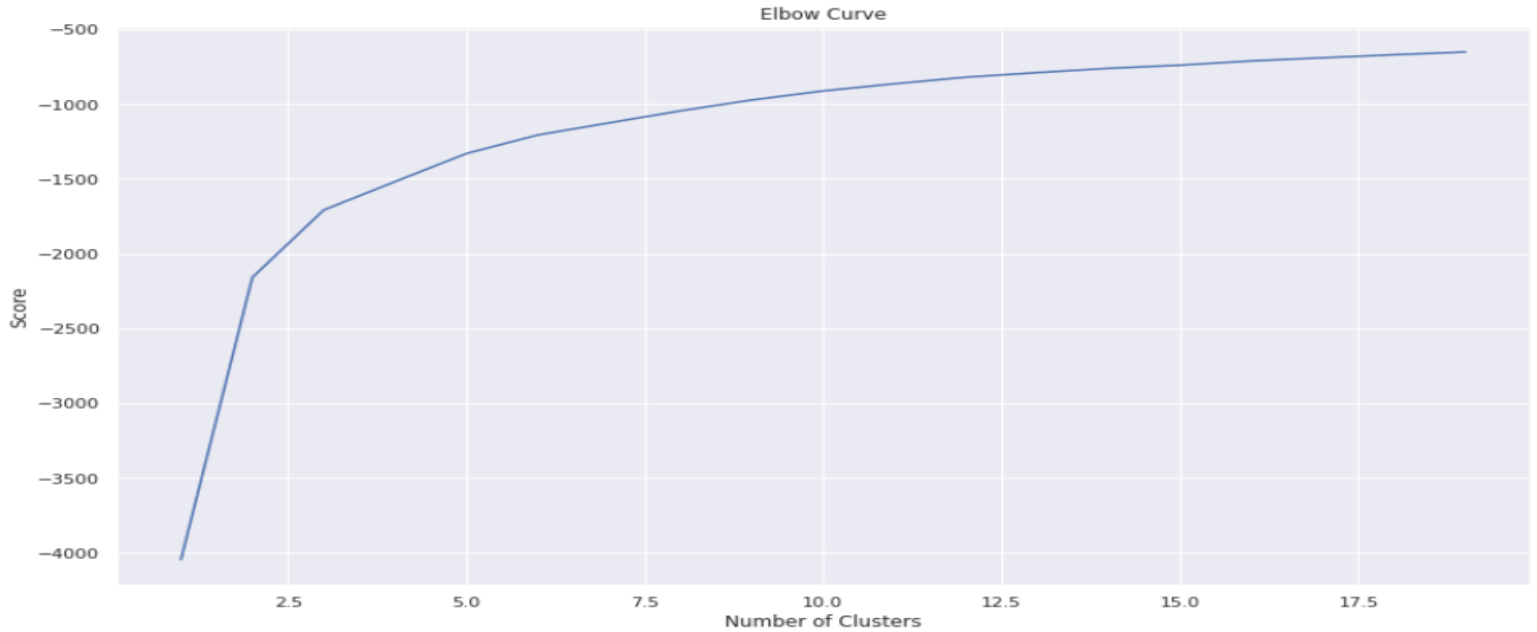
K - Means

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

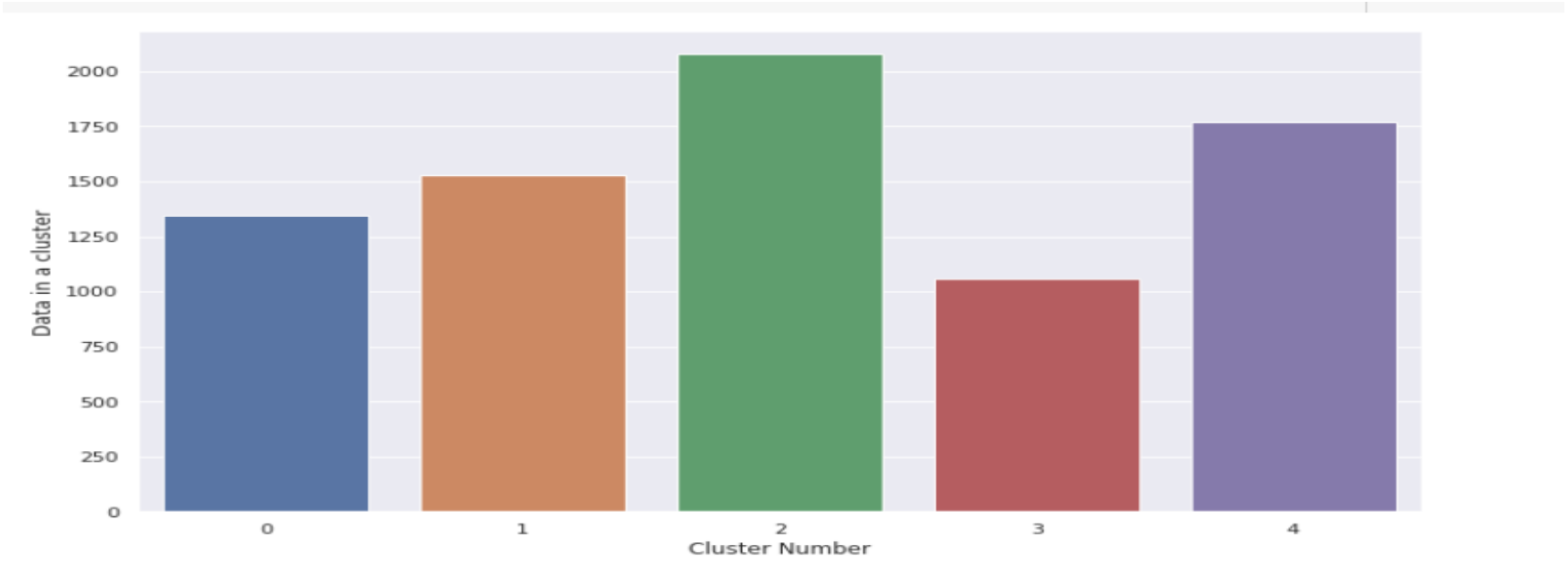
It halts creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved

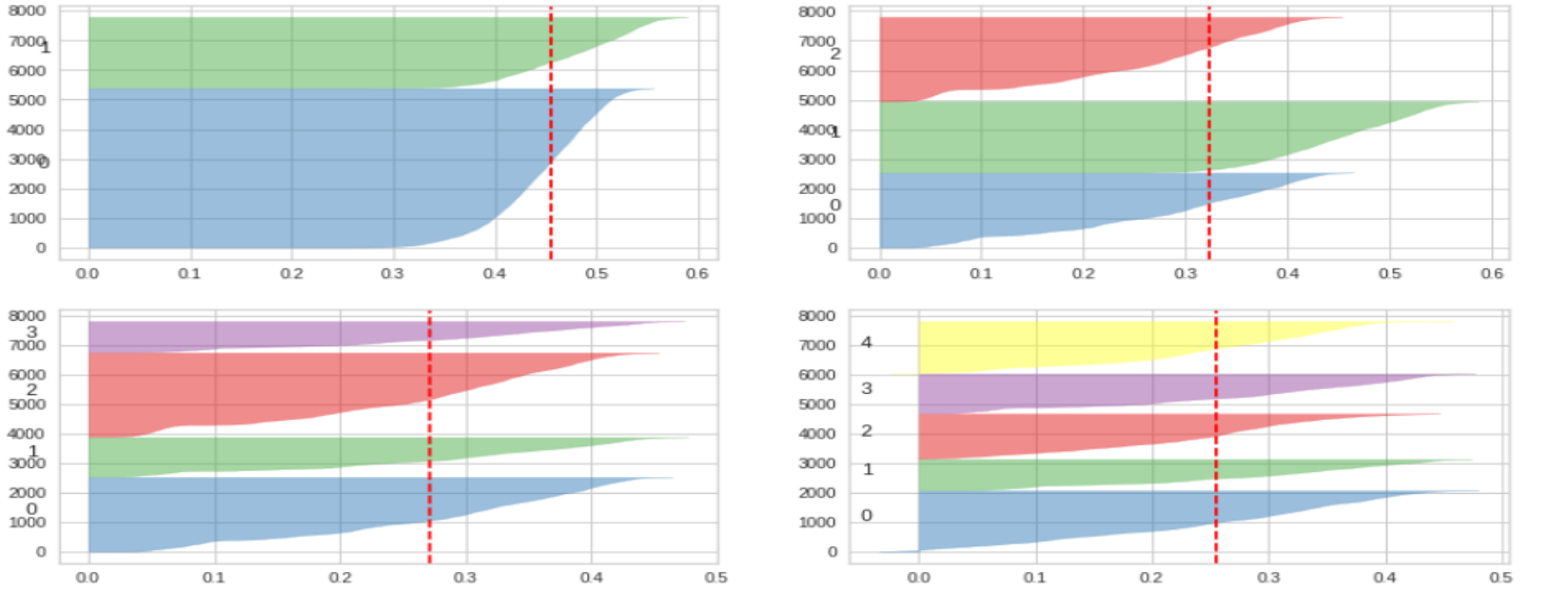
K-Means


Elbow Curve
−500
−1000
−1500
−2000
−2500
−3000
−3500
−4000
Score
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Number of Clusters

Clusters


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

Silhouette Analysis



Conclusion

- 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.

Q & A