Capstone Project

Netflix Movies And TV Shows Clustering

by

Roshan Jamthe

Points for Discussion

- Problem Statement
- Data Summary
- Insights from EDA
- Feature Selection & Train Datasets
- Clustering Models
- Model Comparison
- Best Models
- Similar Content System
- Conclusions

Problem Statement

This dataset consists of TV Shows and Movies available on Netflix as of 2021. 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.

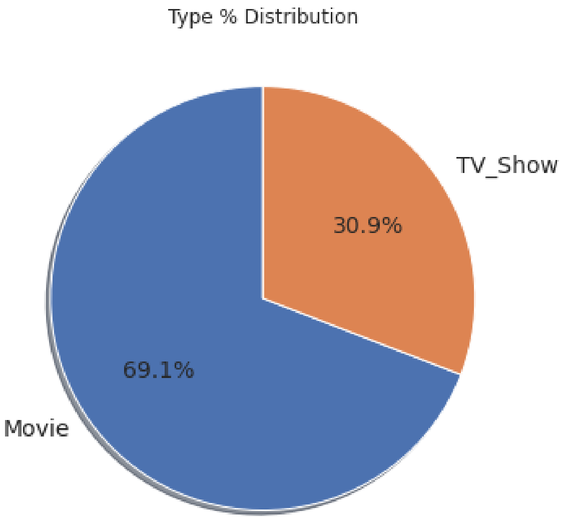
Data Summary

Sr. No.	Feature	Data Type	Details
1	show_id	object	Unique ID for every Movie / TV Show
2	type	object	Identifier - A Movie or TV Show
3	title	object	Title of the Movie / TV Show
4	director	object	Director of the Movie
5	cast	object	Actors involved in the movie / show
6	country	object	Country where the movie / show was produced
7	date_added	date	Date it was added on Netflix
8	release_year	int	Actual Release year of the movie / show
9	rating	object	TV Rating of the movie / show
10	duration	object	Total Duration - in minutes or number of seasons
11	listed_in	object	Genres
12	description	object	The Summary description

Insights from EDA

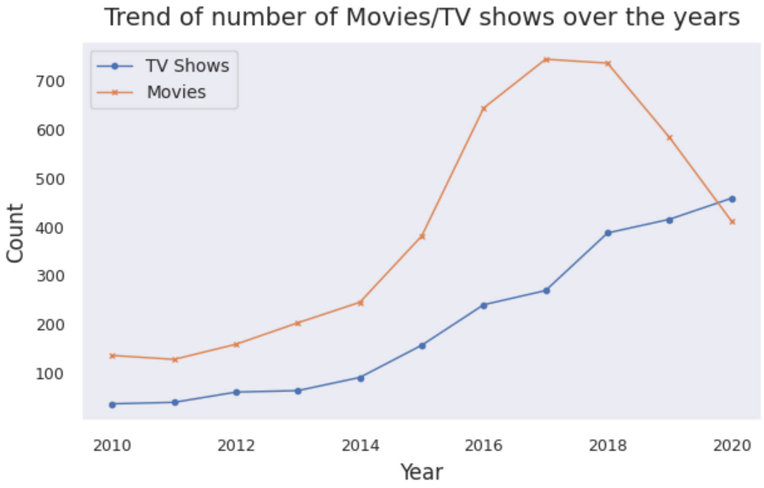
1. Number of Movies Vs Number of TV Shows

There is more than twice the number of movies than the number of tv shows.



2. Trend of content added over the years

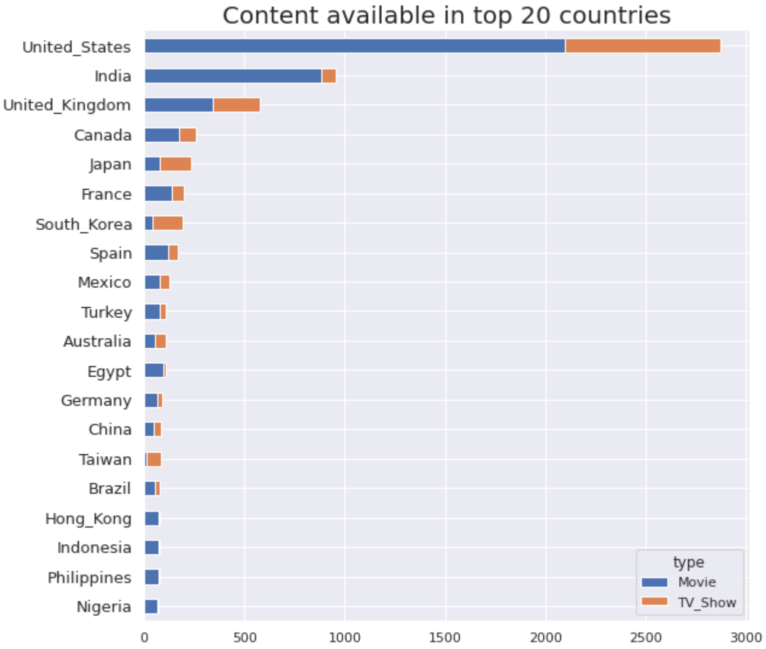
Movie addition declined aggressively after 2018, while tv shows addition displayed continuous growth. In 2020, more number of tv shows were added than movies.



Insights from EDA (Contd.)

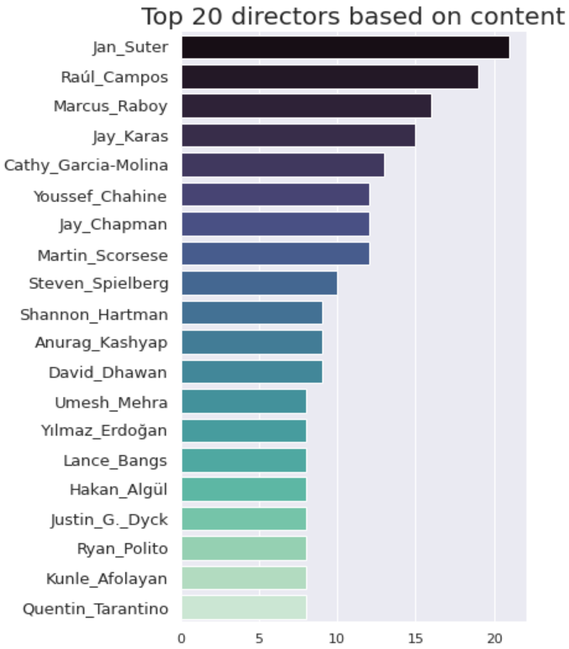
3. Content available in the top 20 countries

Netflix has over 90% of its total content from the top 12 countries. Most are from United States.



4. Top 20 most featured directors

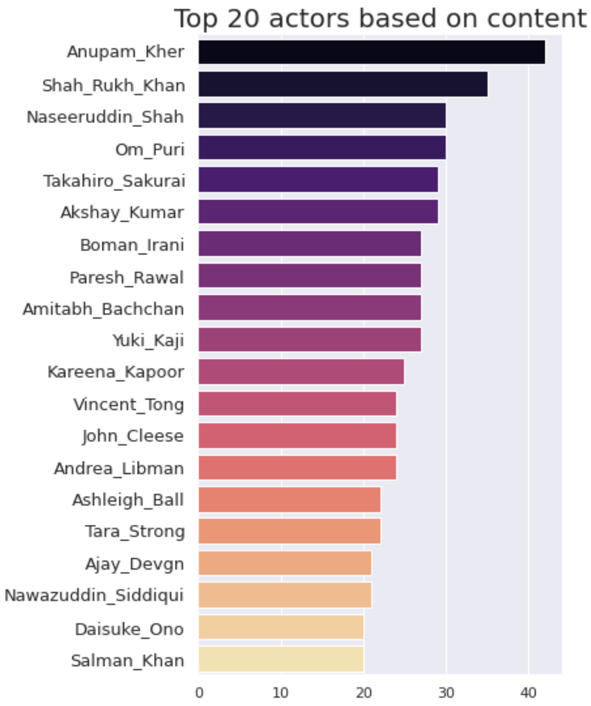
Jan Suter is the most featured director on Netflix.



Insights from EDA (Contd.)

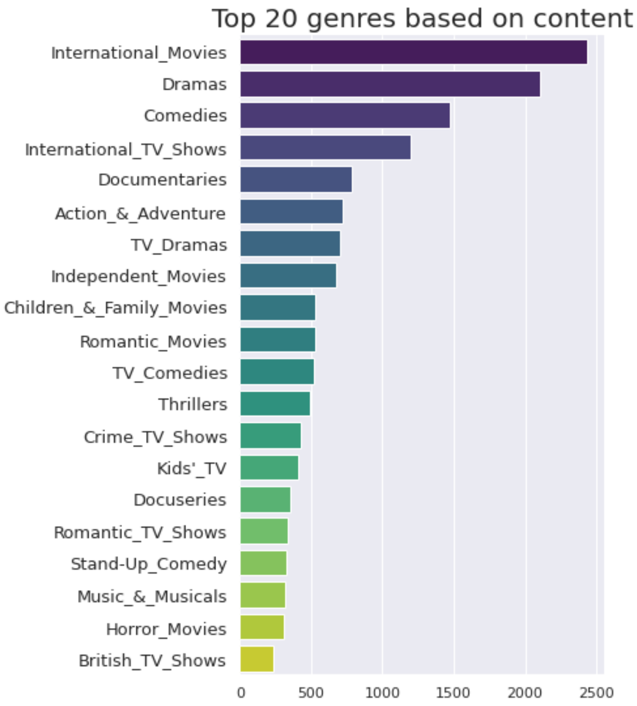
5. Top 20 most featured actors

Anupam Kher is the most featured actor on Netflix.



6. Top 20 genres

International Movies is the top genre.

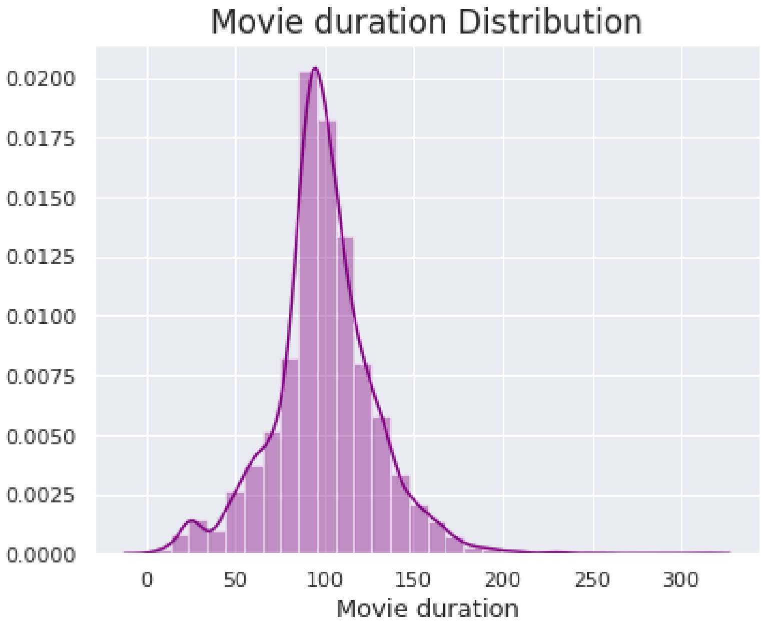


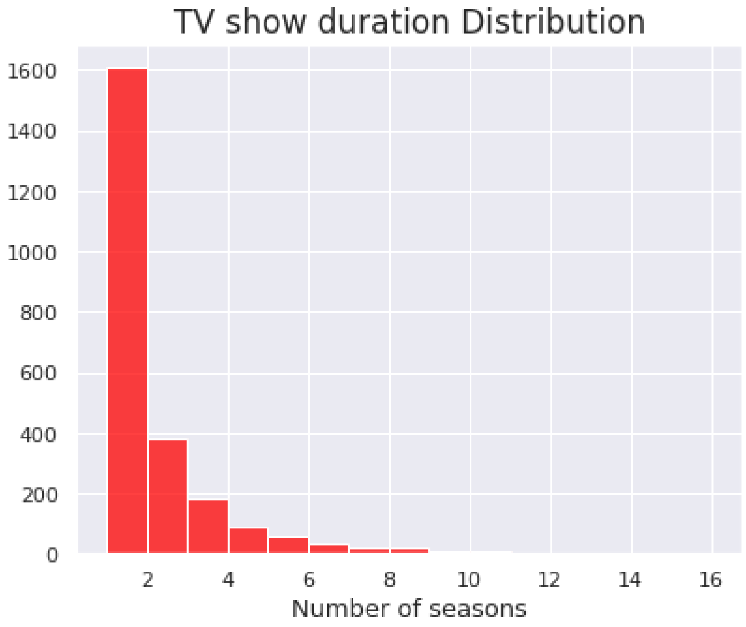
Insights from EDA (Contd.)

7. Content duration distribution

Most movies are of standard 90mins.

Most tv shows have only 1 season. Some tv shows have up to 16 seasons.

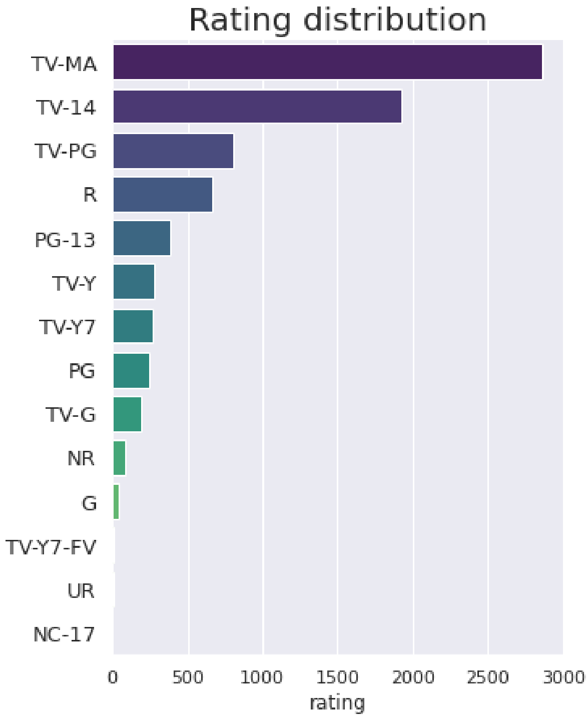




Insights from EDA (Contd.)

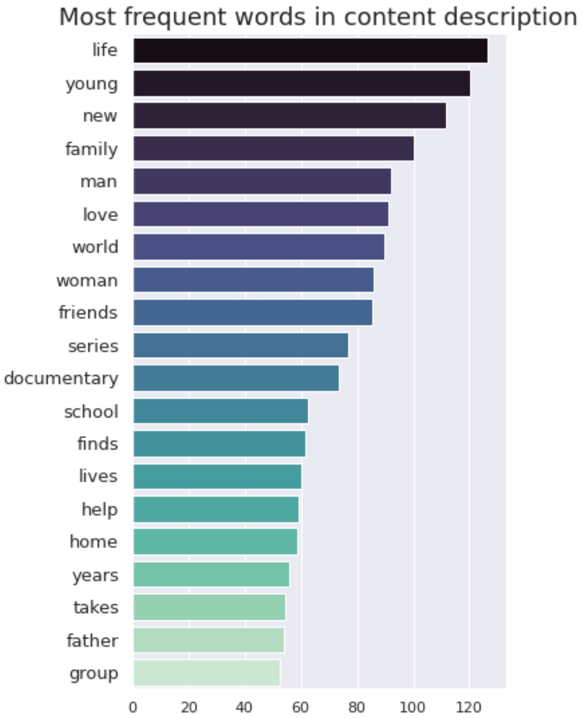
8. Rating distribution

Most content is for a mature audience only.



9. Top 20 most frequent words in content description

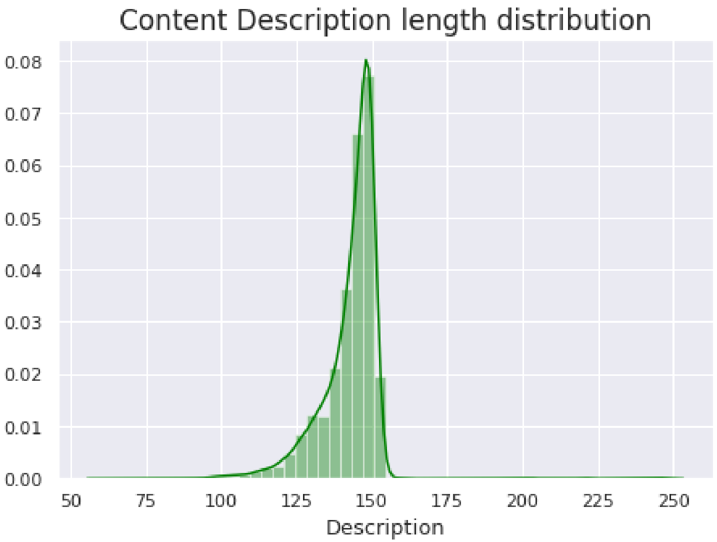
Life, young, new, family, and man are the top 5 most frequent words.

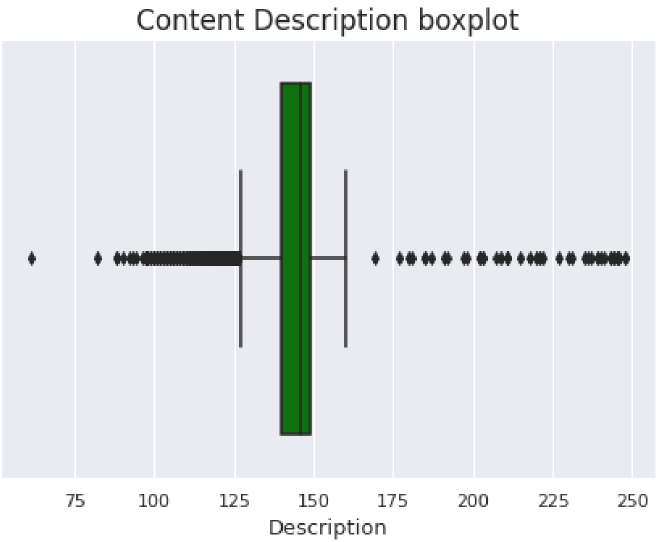


Insights from EDA (Contd.)

10. Content description length distribution

Most content is described in approximately 140 words.

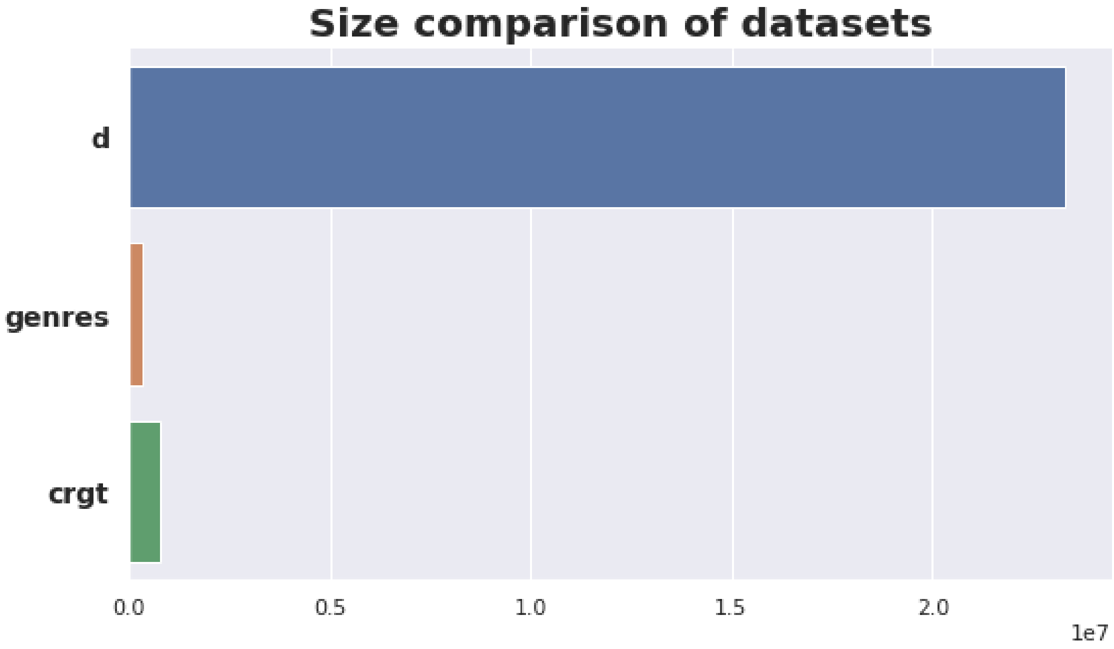




Feature Selection &Train Datasets

Text based features:

Features	type	title	cast	director	country	rating	listed_in	description
Nunique	2	7770	32836	4476	82	14	42	7770



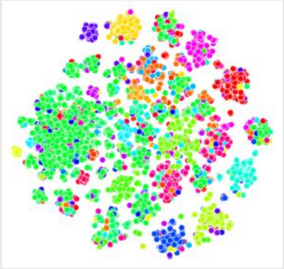
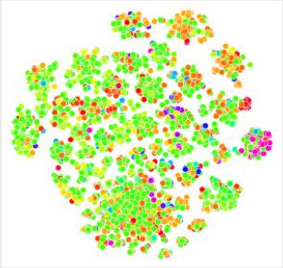
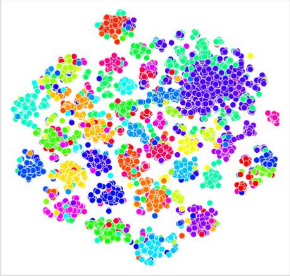
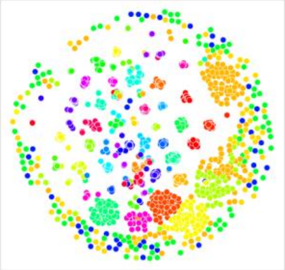
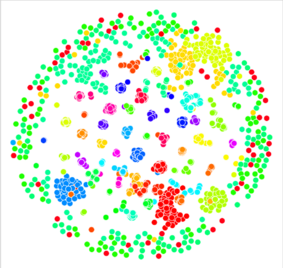
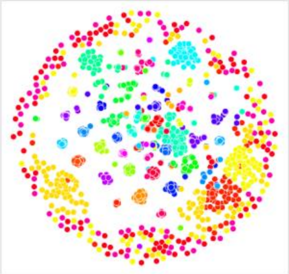
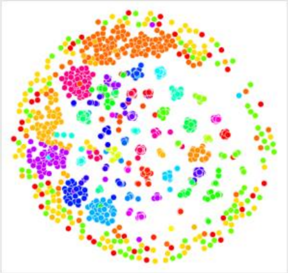
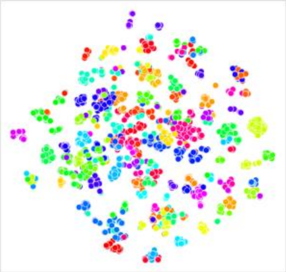
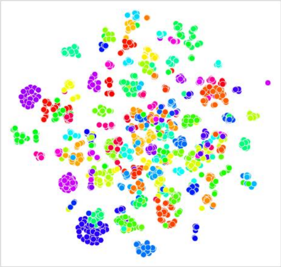
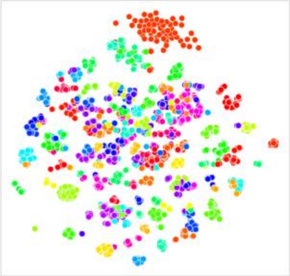
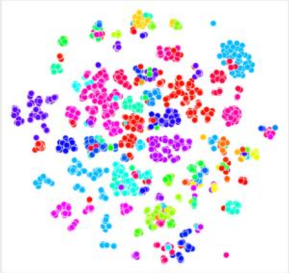
Datasets Explanation

d :
Shape = (7770, 3000)
Has pre-processed and vectorized description feature.

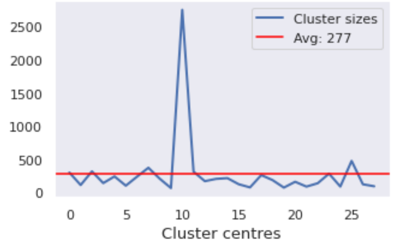
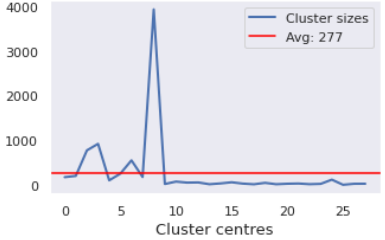
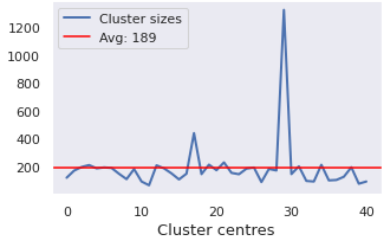
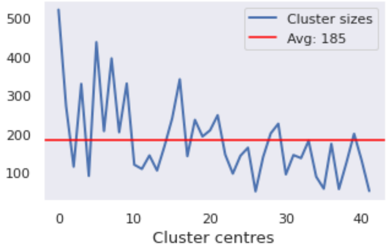
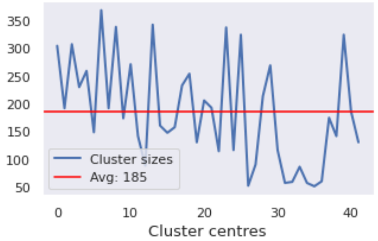
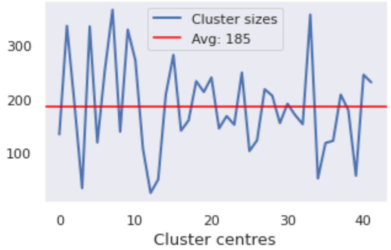
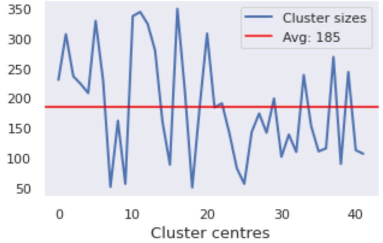
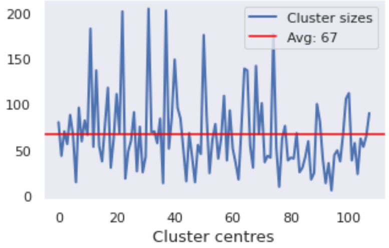
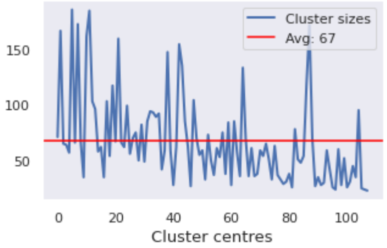
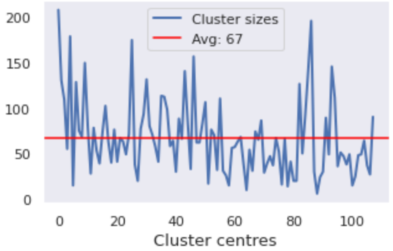
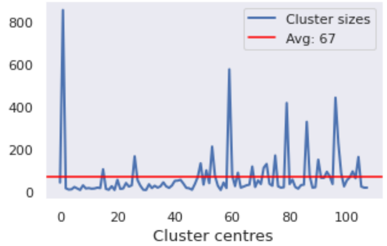
genres:
Shape = (7770, 42)
Has pre-processed and vectorized listed_in feature.

crgt :
Shape = (7265, 108)
Has pre-processed and vectorized lead_country, rating, listed_in, and type features.

Model Comparison (Cluster separation)

	MiniBatchKMeans	Ward	Gaussian Mixture	Spectral
d				
genres				
crgt				

Model Comparison (Cluster size consistency)

	MiniBatchKMeans	Ward	Gaussian Mixture	Spectral
d				
genres				
crgt				

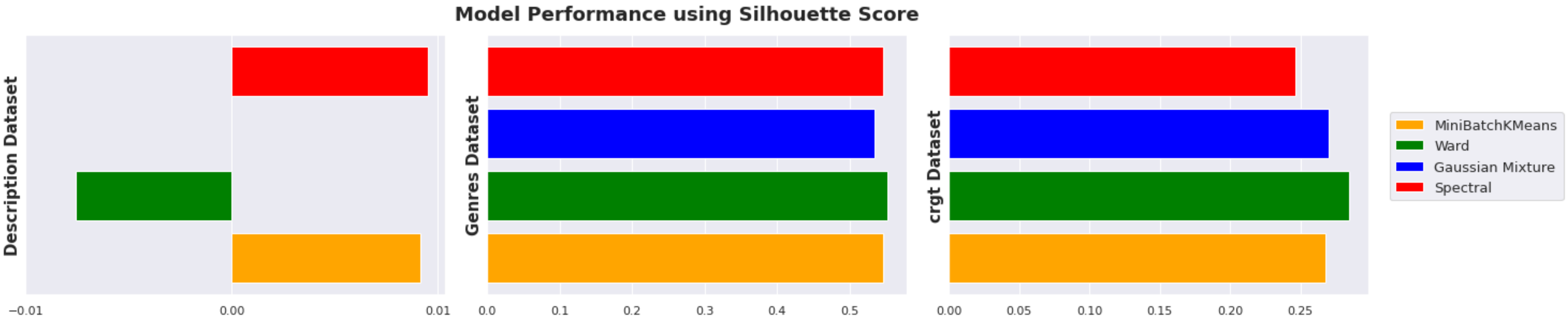
Model Performance Comparison

We selected best model for the dataset based on **Silhouette score.**

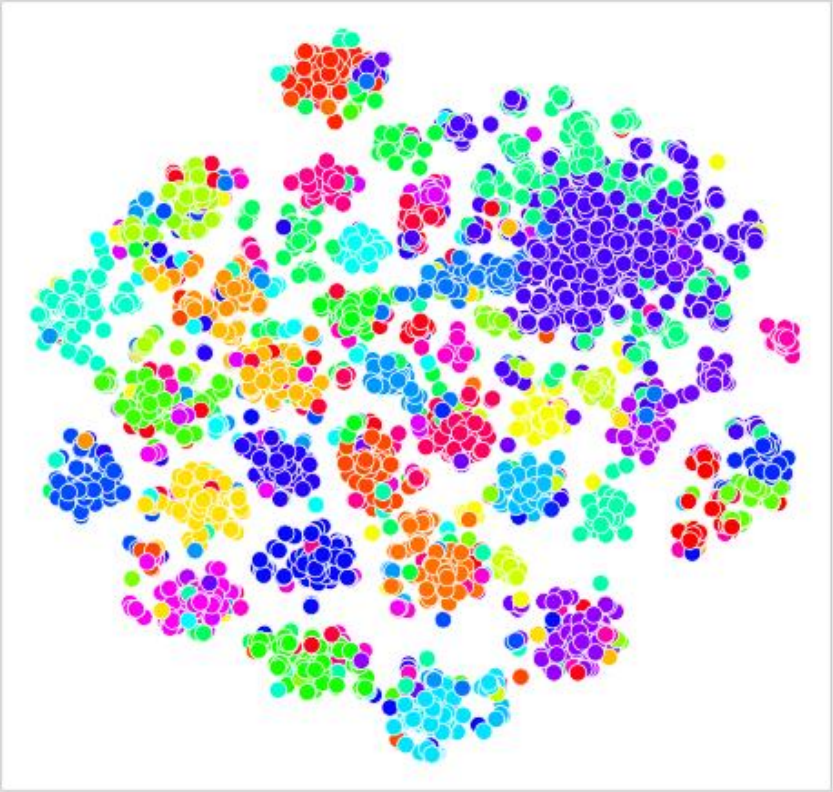
For **Description** dataset, **Spectral clustering** scored 0.00949.

For **Genres** dataset, **Ward clustering** scored 0.55162.

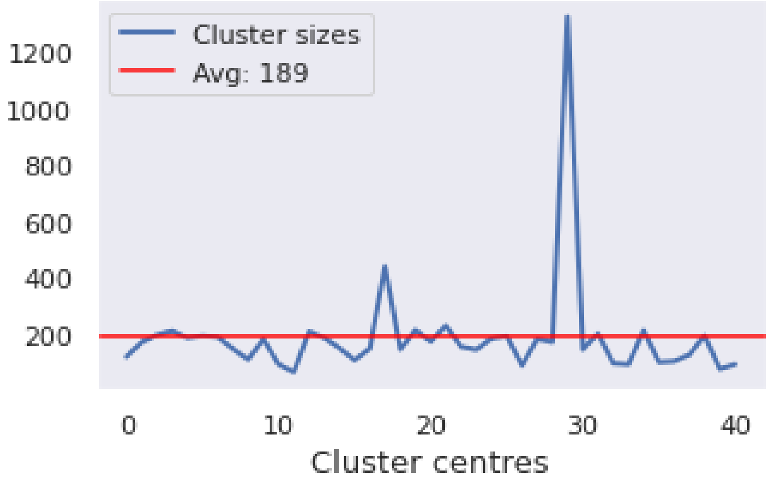
For **crgt** dataset, **Ward clustering** scored 0.28438.



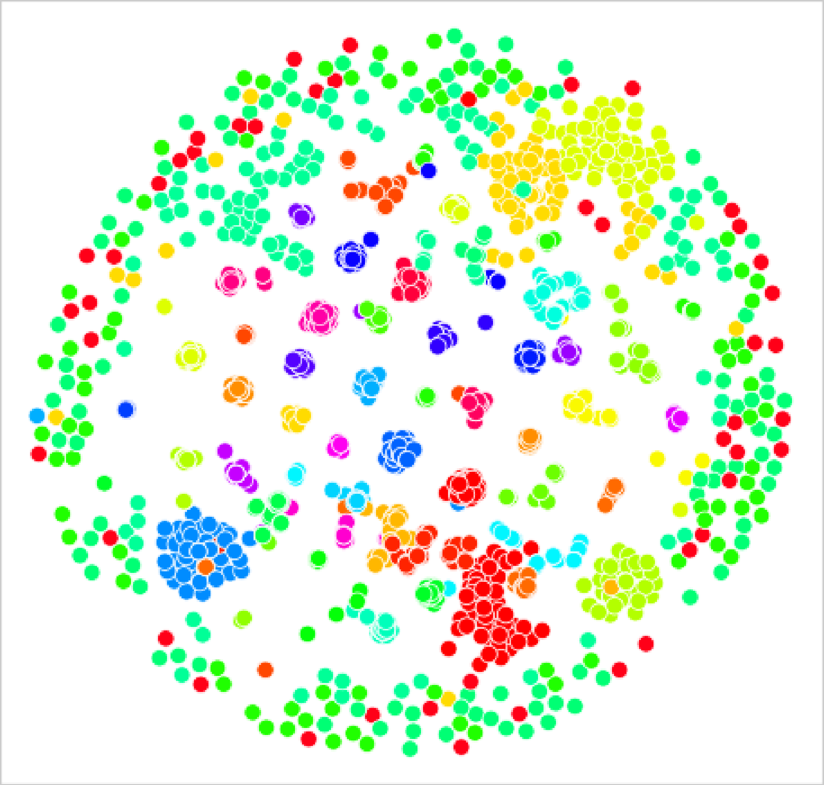
Best Model (for Description dataset)



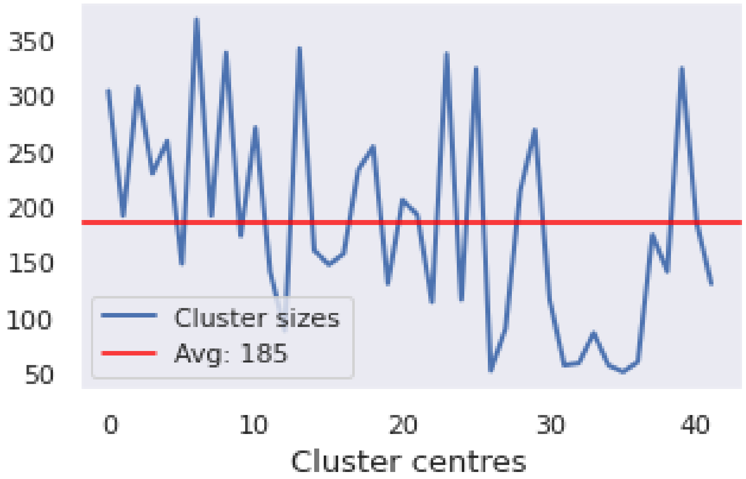
For Description dataset, the Spectral clustering worked the best with the silhouette score of 0.00949, n_clusters = 41, avg cluster size = 189, and peak cluster size = 1323.



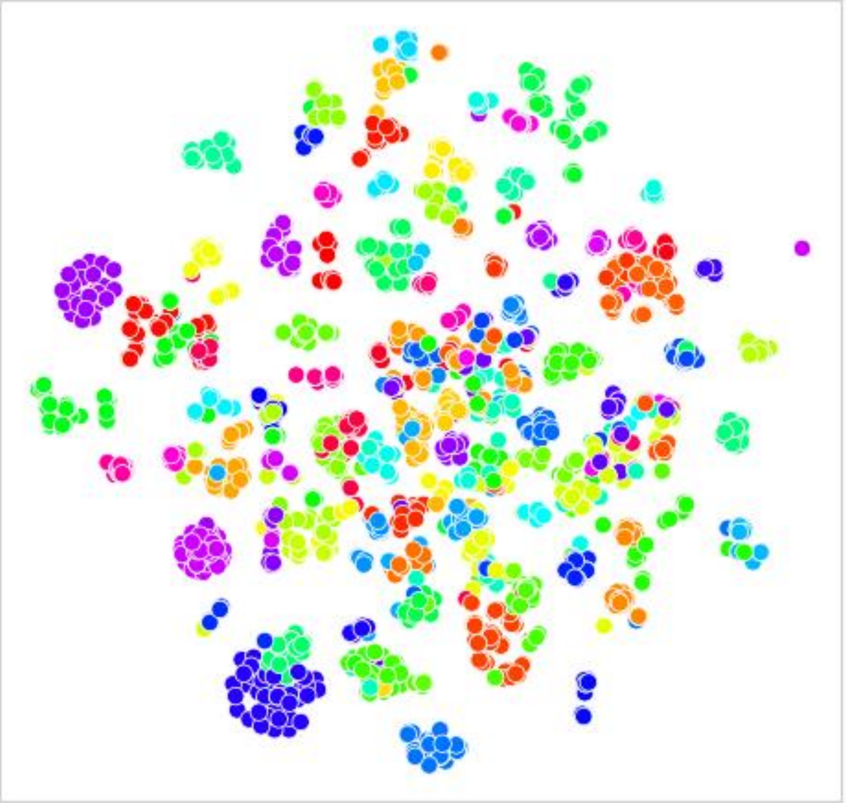
Best Model (for genres dataset)



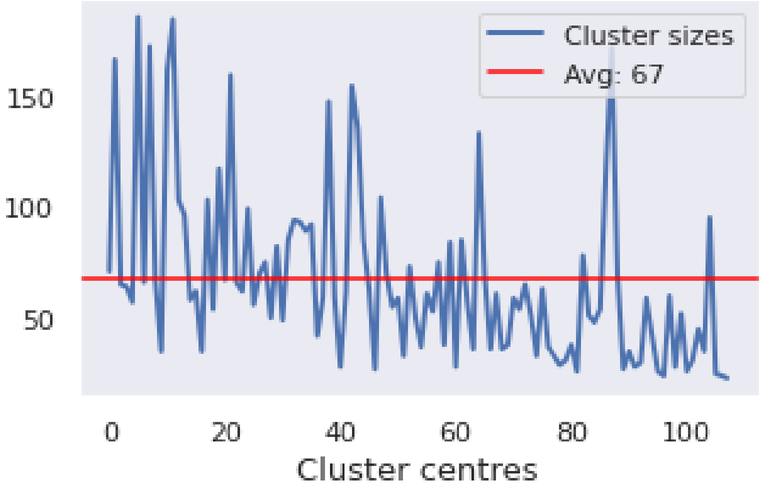
For genres dataset, the Ward clustering worked the best with the silhouette score of 0.55162, n_clusters = 42, avg cluster size =185, and peak cluster size = 367.



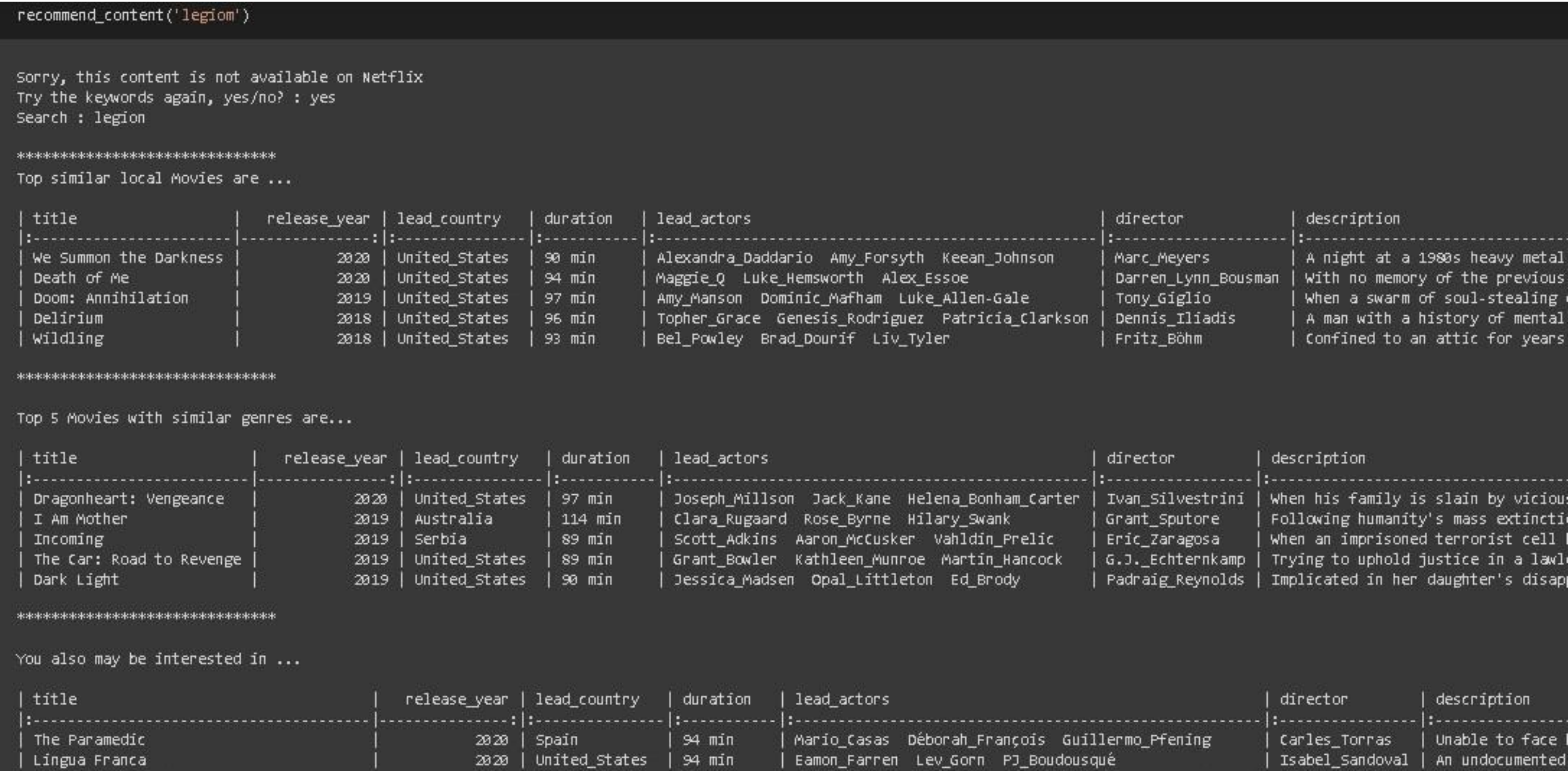
Best Model (for crgt dataset)



For crgt dataset, the ward clustering worked the best with the silhouette score of 0.28438, n_clusters = 108, avg cluster size = 67, and peak cluster size = 185



Similar Content System



```
recommend_content('legiom')
```

```
Sorry, this content is not available on Netflix
Try the keywords again, yes/no? : yes
Search : legion

******************************
Top similar local Movies are ...

| title                 |   release_year | lead_country   | duration   | lead_actors                                                  | director            | description
|:----------------------|---------------:|:---------------|:-----------|:-------------------------------------------------------------|:--------------------|:---------------------------
| We Summon the Darkness |           2020 | United_States  | 90 min     | Alexandra_Daddario  Amy_Forsyth  Keean_Johnson               | Marc_Meyers         | A night at a 1980s heavy metal
| Death of Me           |           2020 | United_States  | 94 min     | Maggie_Q  Luke_Hemsworth  Alex_Essoe                         | Darren_Lynn_Bousman | With no memory of the previous
| Doom: Annihilation    |           2019 | United_States  | 97 min     | Amy_Manson  Dominic_Mafham  Luke_Allen-Gale                  | Tony_Giglio         | When a swarm of soul-stealing
| Delirium              |           2018 | United_States  | 96 min     | Topher_Grace  Genesis_Rodriguez  Patricia_Clarkson           | Dennis_Iliadis      | A man with a history of mental
| Wildling              |           2018 | United_States  | 93 min     | Bel_Powley  Brad_Dourif  Liv_Tyler                           | Fritz_Böhm          | Confined to an attic for years

******************************

Top 5 Movies with similar genres are...

| title                    |   release_year | lead_country   | duration   | lead_actors                                            | director           | description
|:-------------------------|---------------:|:---------------|:-----------|:-------------------------------------------------------|:-------------------|:---------------------------
| Dragonheart: Vengeance   |           2020 | United_States  | 97 min     | Joseph_Millson  Jack_Kane  Helena_Bonham_Carter        | Ivan_Silvestrini   | When his family is slain by viciou
| I Am Mother              |           2019 | Australia      | 114 min    | Clara_Rugaard  Rose_Byrne  Hilary_Swank                | Grant_Sputore      | Following humanity's mass extincti
| Incoming                 |           2019 | Serbia         | 89 min     | Scott_Adkins  Aaron_McCusker  Vahldin_Prelic           | Eric_Zaragosa      | When an imprisoned terrorist cell
| The Car: Road to Revenge |           2019 | United_States  | 89 min     | Grant_Bowler  Kathleen_Munroe  Martin_Hancock          | G.J._Echternkamp   | Trying to uphold justice in a lawl
| Dark Light               |           2019 | United_States  | 90 min     | Jessica_Madsen  Opal_Littleton  Ed_Brody               | Padraig_Reynolds   | Implicated in her daughter's disap

******************************

You also may be interested in ...

| title                                |   release_year | lead_country   | duration   | lead_actors                                                     | director          | description
|:-------------------------------------|---------------:|:---------------|:-----------|:----------------------------------------------------------------|:------------------|:------------
| The Paramedic                        |           2020 | Spain          | 94 min     | Mario_Casas  Déborah_François  Guillermo_Pfening                | Carles_Torras     | Unable to face
| Lingua Franca                        |           2020 | United_States  | 94 min     | Eamon_Farren  Lev_Gorn  PJ_Boudousqué                           | Isabel_Sandoval   | An undocumented
```

Conclusions

Insights from EDA

- Movies and TV Shows on Netflix are in ratio 7/3.
- From past five years, Netflix has shifted its focus towards TV shows. Number of movies added to Netflix reduced by 45% while number of TV Shows added increased by 70%.
- Most content on Netflix is from United States followed by India and United Kingdom. Netflix has a little over 90% of its total Content from top 12 countries.
- Jan Suter is the most featured director.
- Anupam Kher is the most featured actor. Out of top 10 actors, 8 are Indians. It shows the impact of Indian actors on the World television industry.
- International movies is the top genre followed by Dramas and Comedies.
- Most movies are of length ~90 mins while most TV Shows have only one season.
- Most content on Netflix is rated for mature audience only
- Most content has description length of ~140 words.
- Life, young, new, family and man are the top 5 most frequent words in description.

Conclusions

Clustering

- Clustering on description column alone (dataset name : d)

  We found 41 clusters with best separation using Spectral clustering with the silhouette score of 0.00949. The average cluster size is 189, while the largest cluster has 1323 items.

- Clustering on listed_in (referred as genres) column alone (dataset name : genres)

  We found 42 clusters with best separation using Ward clustering with the silhouette score of 0.55162. The average cluster size is 185, while the largest cluster has 367 items.

- Clustering on lead_country, rating, listed_in and type columns (dataset name : crgt)

  We found 108 optimal clusters with best separation using Ward clustering with the silhouette score of 0.28438. The average cluster size is 67, while the largest cluster has 185 items.

- The MiniBatchKMeans was the fastest clustering algorithm and it gave decent results on all the datasets. The ward clustering was the slowest clustering algorithm but it gave better results in terms of smoother cluster size and better cluster separation on genres and crgt dataset. Spectral clustering should be a preferred choice for high dimensional data.

Thank you!