CAPSTONE PROJECT 4

NETFLIX MOVIES & TV-SHOWS CLUSTERING

TEAM MEMBERS

KIRTESH VERMA

PRAVIN BEJJO

SAHIL PARDESHI

❖ CONTENTS

- PROBLEM STATEMENT
- DATA OVERVIEW
- EXPLORATORY DATA ANALYSIS(EDA)
- FEATURE ENGINEERING & DATA PREPROCESSING
- MODEL BUILDING
- CONCLUSION

❖ PROBLEM STATEMENT

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 services 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.

- **We are required to do following task**

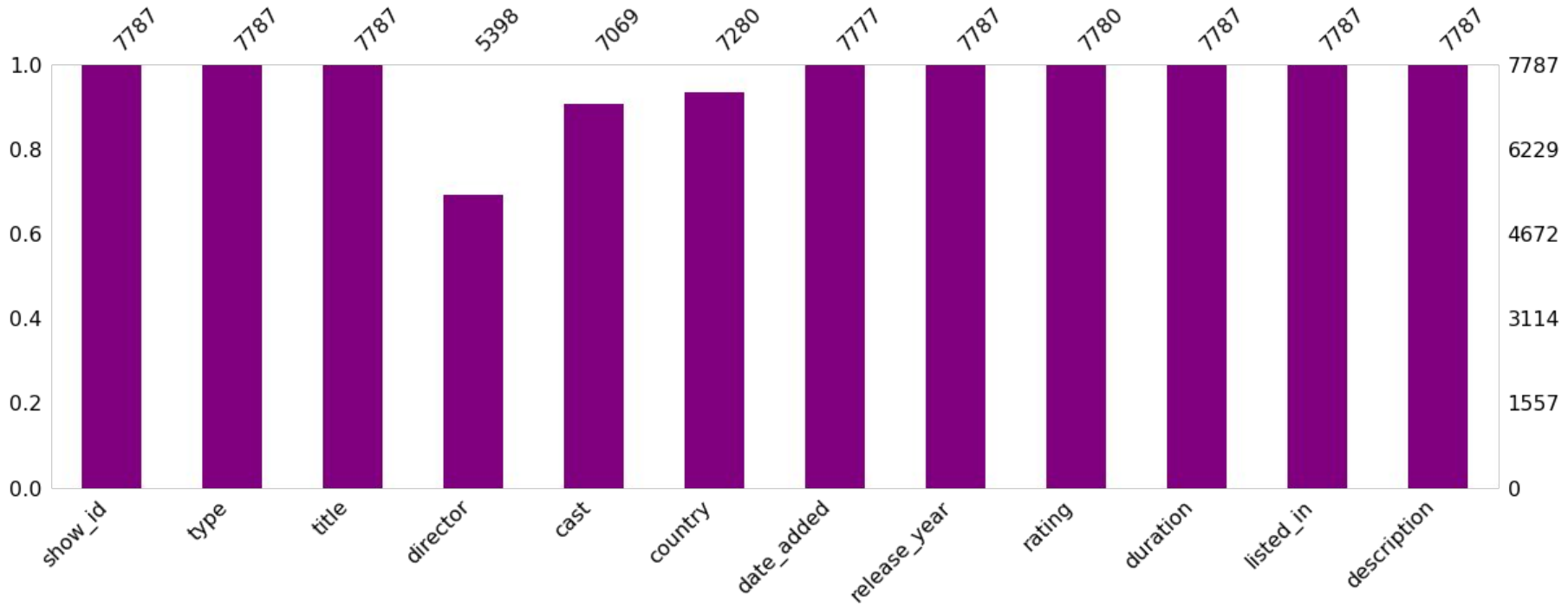
1. **Exploratory Data Analysis**
2. **Understanding what type content is available in different countries**
3. **Is Netflix has increasingly focusing on TV rather than movies in recent years.**
4. **Clustering similar content by matching text-based features**

❖ DATA OVERVIEW

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 :** Genere
12. **description:** The Summary description

```
RangeIndex: 7787 entries, 0 to 7786
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       7787 non-null   object
 1   type          7787 non-null   object
 2   title         7787 non-null   object
 3   director      5398 non-null   object
 4   cast          7069 non-null   object
 5   country       7280 non-null   object
 6   date_added    7777 non-null   object
 7   release_year  7787 non-null   int64
 8   rating        7780 non-null   object
 9   duration      7787 non-null   object
 10  listed_in     7787 non-null   object
 11  description   7787 non-null   object
dtypes: int64(1), object(11)
memory usage: 730.2+ KB
```

❖ NULL & DUPLICATE VALUES

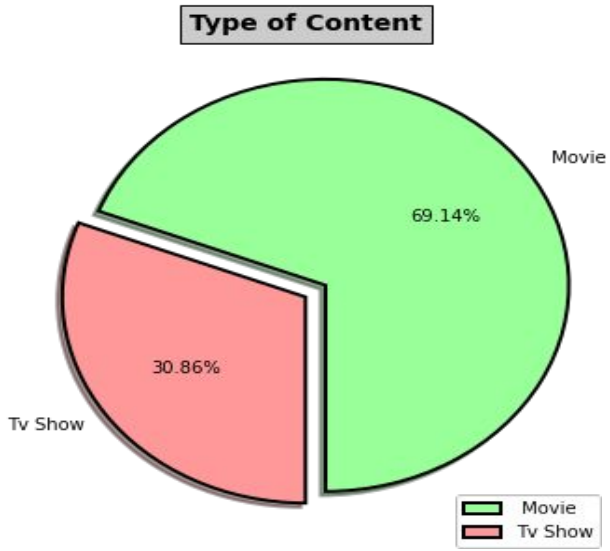


❖ DATA SAMPLE

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	TV Show	3%	NaN	João Miguel, Bianca Comparato, Michel Gomes, R...	Brazil	August 14, 2020	2020	TV-MA	4 Seasons	International TV Shows, TV Dramas, TV Sci-Fi &...	In a future where the elite inhabit an island ...
1	s2	Movie	7:19	Jorge Michel Grau	Demián Bichir, Héctor Bonilla, Oscar Serrano, ...	Mexico	December 23, 2016	2016	TV-MA	93 min	Dramas, International Movies	After a devastating earthquake hits Mexico Cit...
2	s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella Chung, Henley Hii, Lawrence ...	Singapore	December 20, 2018	2011	R	78 min	Horror Movies, International Movies	When an army recruit is found dead, his fellow...
3	s4	Movie	9	Shane Acker	Elijah Wood, John C. Reilly, Jennifer Connelly...	United States	November 16, 2017	2009	PG-13	80 min	Action & Adventure, Independent Movies, Sci-Fi...	In a postapocalyptic world, rag-doll robots hi...
4	s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin Spacey, Kate Bosworth, Aar...	United States	January 1, 2020	2008	PG-13	123 min	Dramas	A brilliant group of students become card-coun...

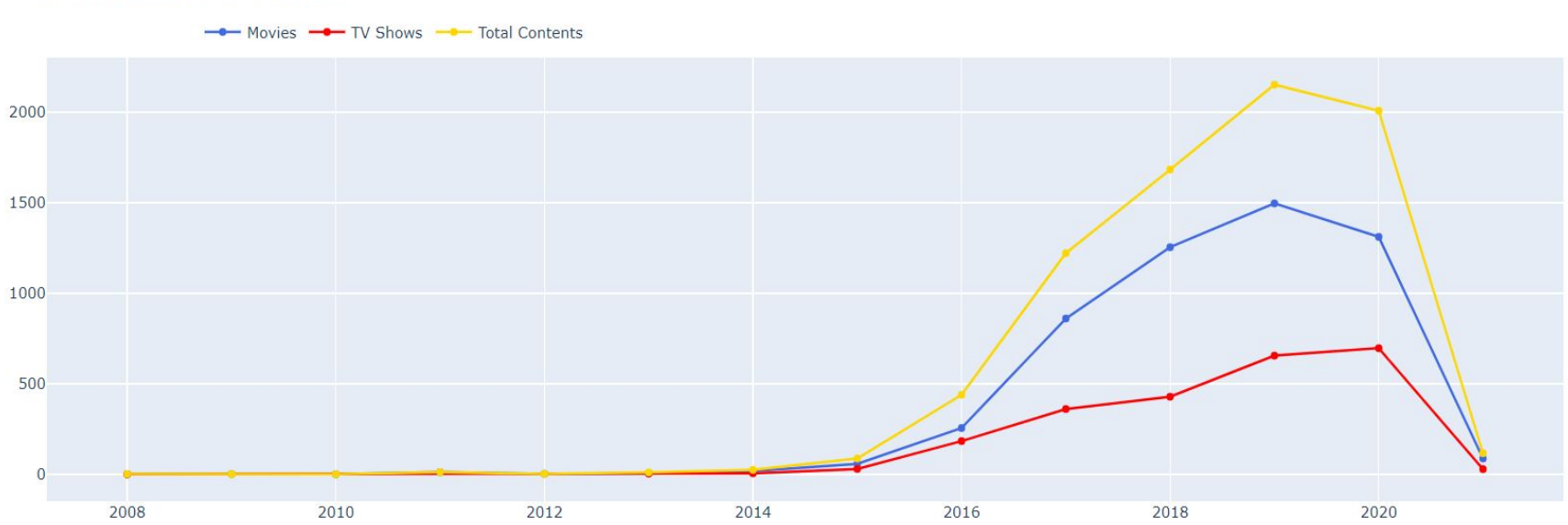
❖ EXPLORATORY DATA ANALYSIS

❖ MOVIES AND TV-SHOWS



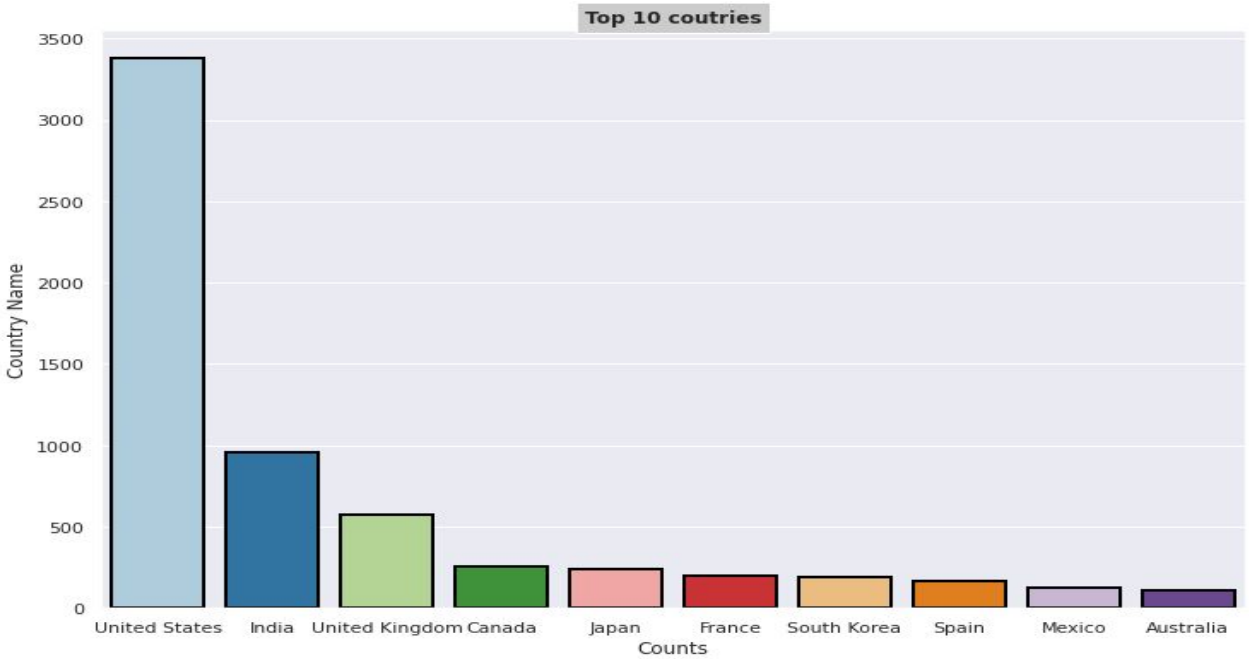
- The percentage of movies available on netflix is way greater than percentage of tv shows
- Almost 70% of content is of type- Movies while rest 30% content includes TV shows

❖ CONTENT ADDED OVER THE YEARS



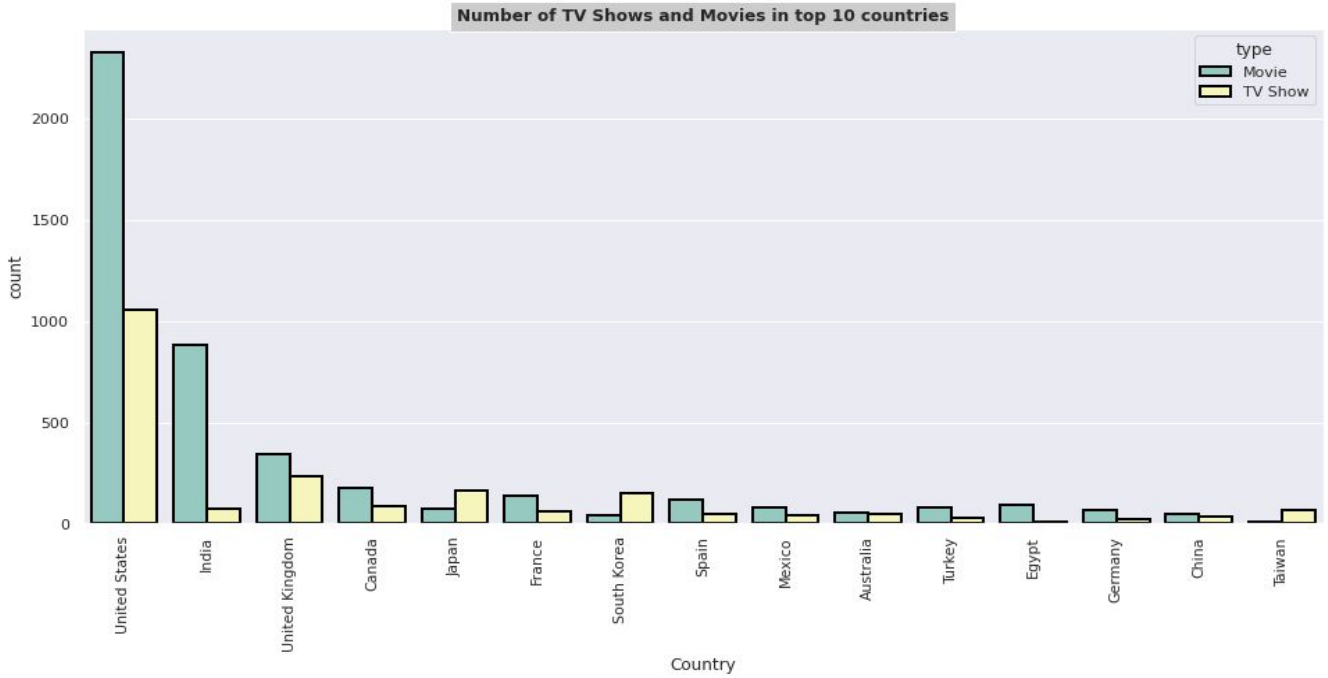
- Netflix has been producing more number of movies then Tv shows from year 2014.
- 2019 is the year of maximum movies released on netflix. A total of 1497 movies were added during this year.
- Max number of Tv shows were added during the year 2020. 697 Tv shows were released during this year.

POPULARITY



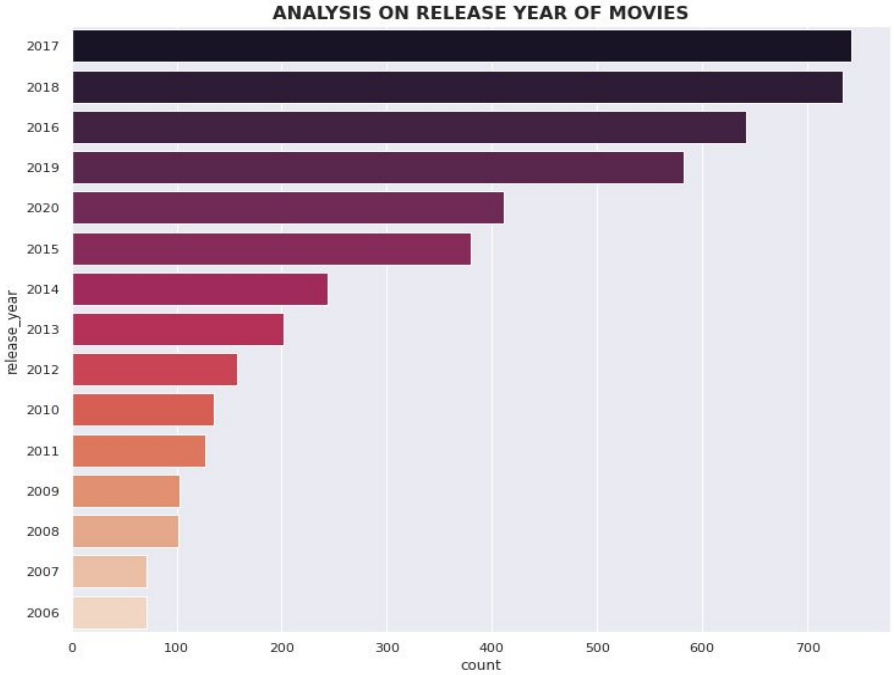
- Netflix is most popular in United States
- India lies at second place in the popularity list followed by United Kingdom which is at third position

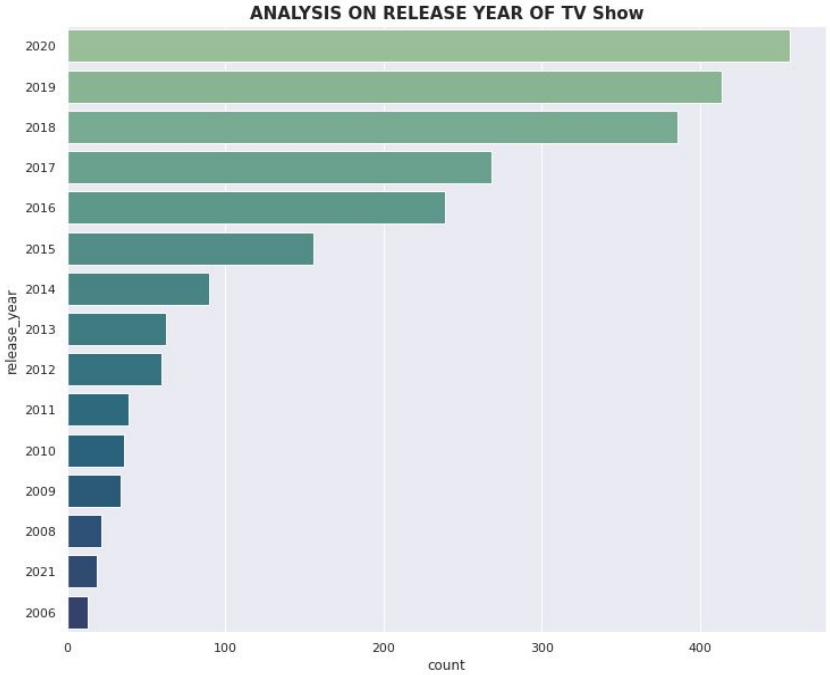
❖ CONTENT IN DIFFERENT COUNTRIES



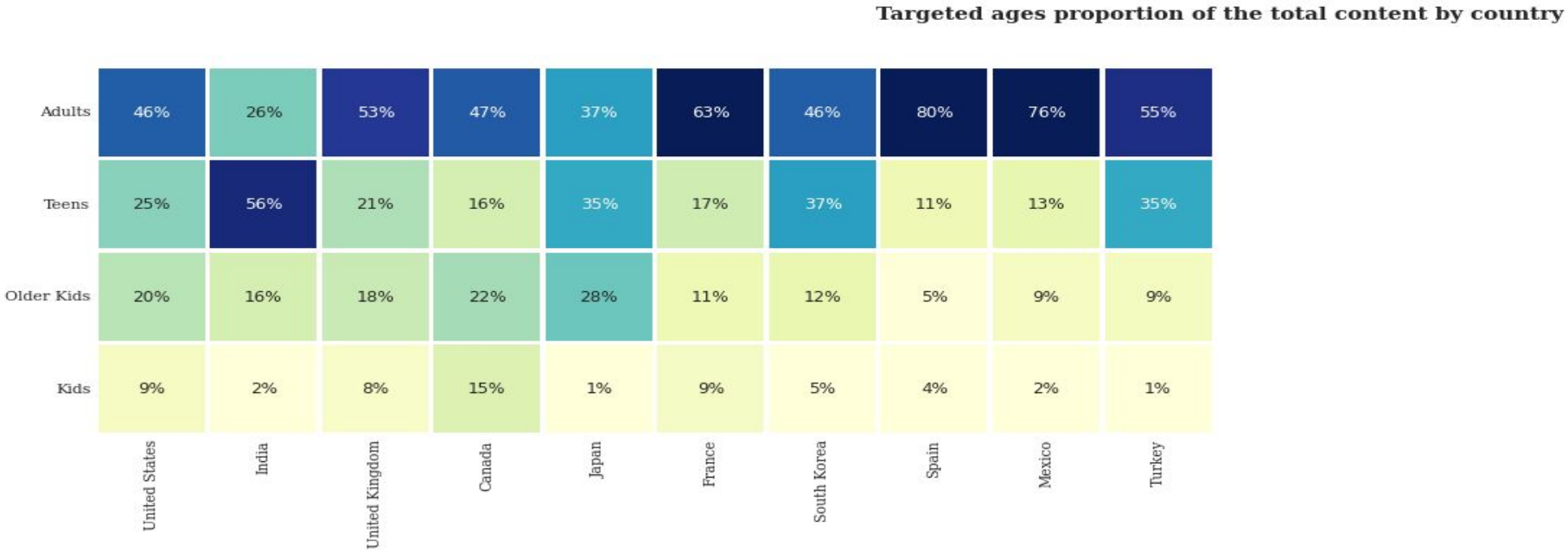
- **The majority of content on netflix is comprised of Movies in most of the countries where as in countries like Japan and South Korea the content is comprised of Tv-shows more than movies.**

❖ RELEASE YEAR



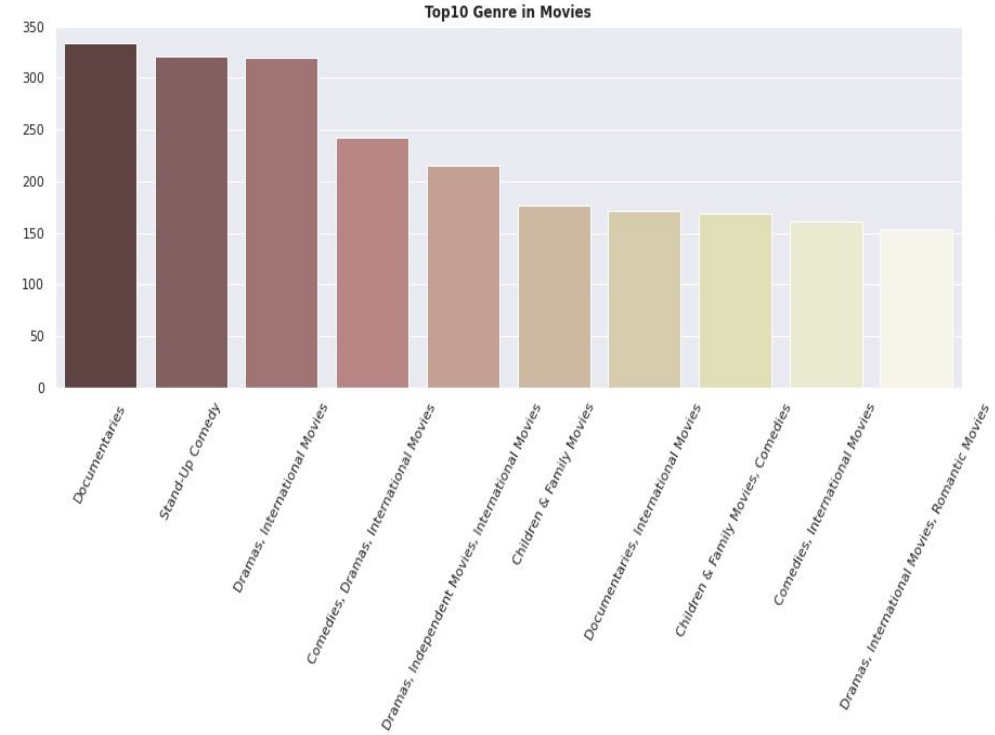


❖ TARGETED AGES

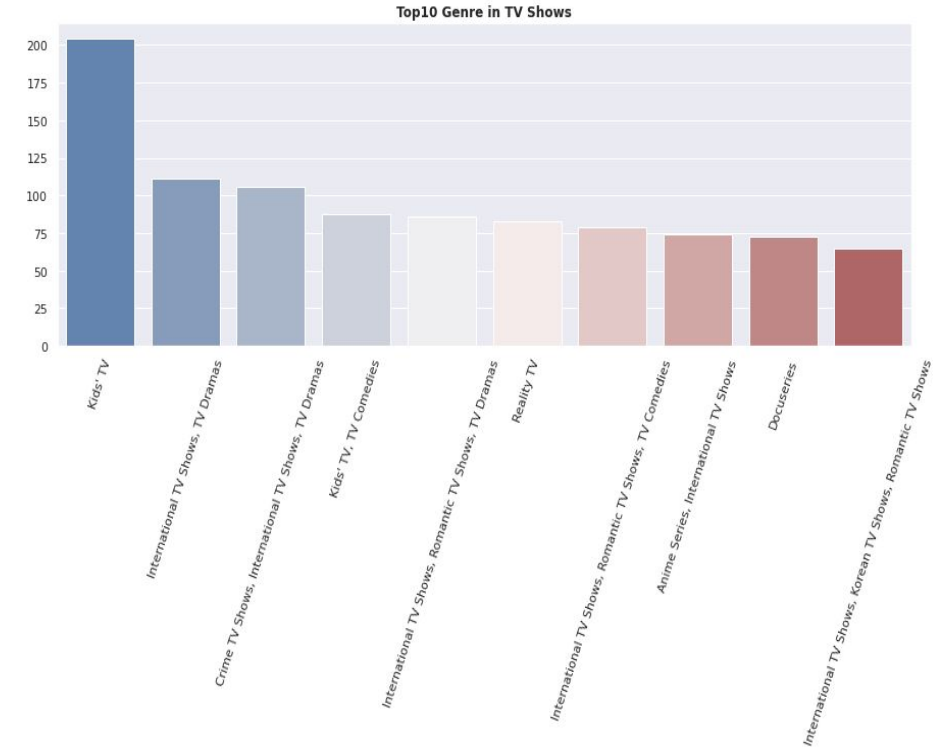


- **Content developed by netflix mostly target teen. 56% of content is for teen which is highest as compared to any other country.**
- **Content developed by netflix in Spain, Mexico and France is mostly for adults.**

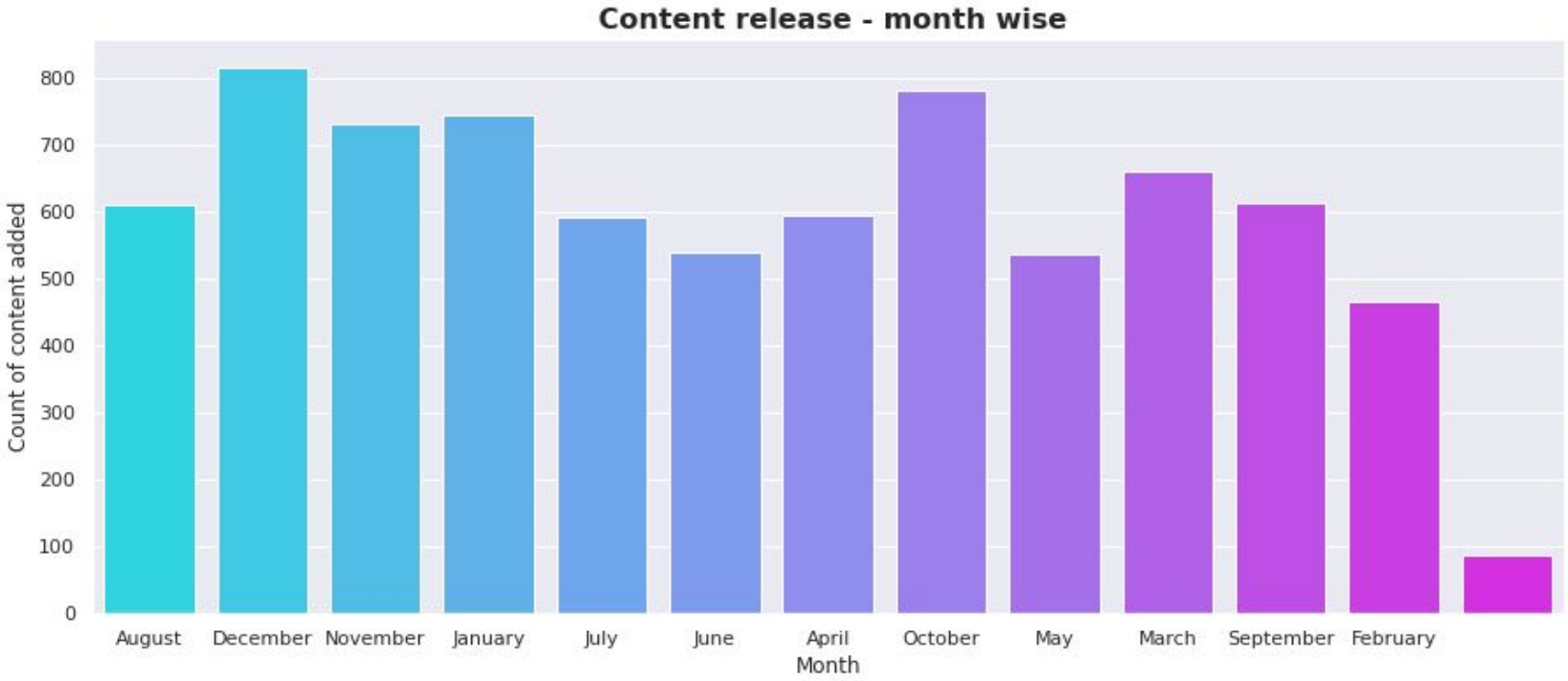
❖ TOP GENRES IN MOVIES



❖ TOP GENRES IN TV-SHOWS

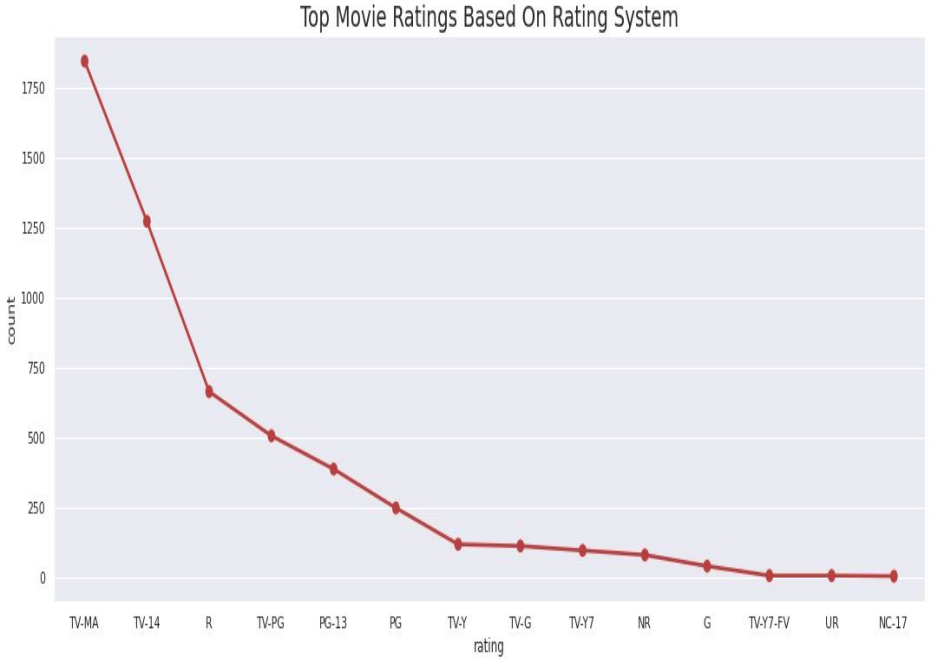


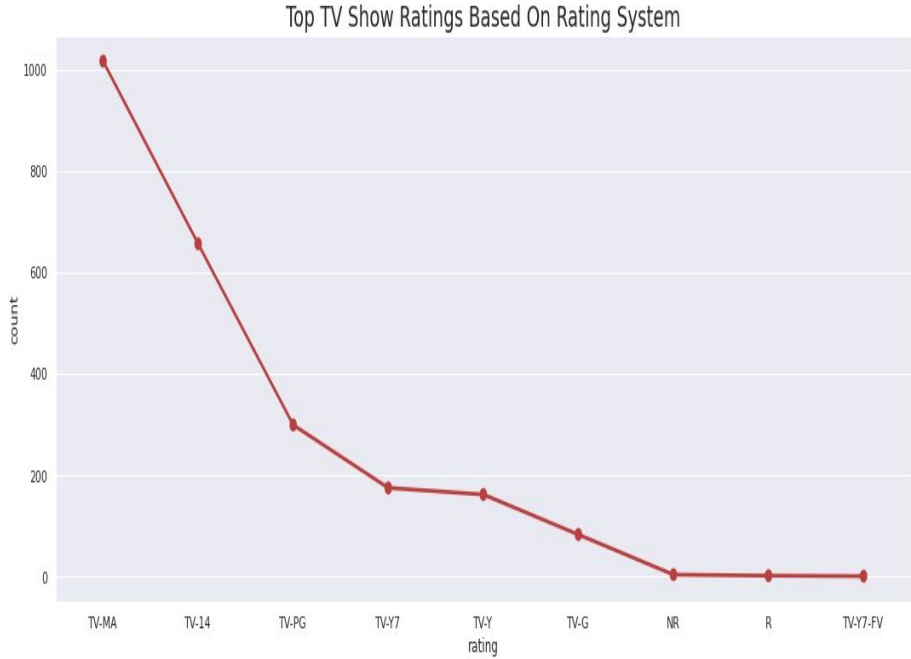
❖ CONTENT RELEASE MONTH WISE



- **Maximum content is added during the end of the year.**
- **Least amount of content is added during the mid year months like May june and july.**

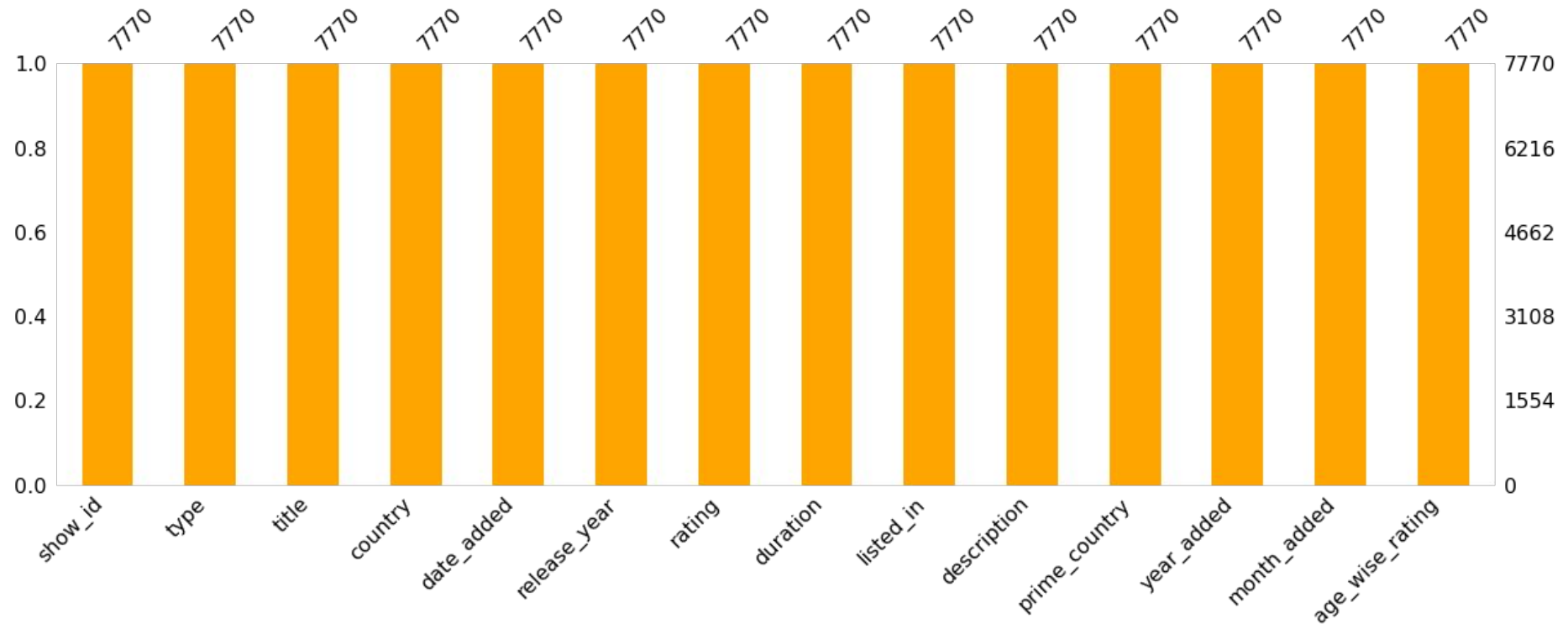
TOP MOVIES & TV-SHOWS RATINGS





- Movies and TV-shows belonging to 'TV-MA' rating are highest on netflix. It is for Mature audience only i.e. viewed by adults and therefore may be unsuitable for children under 17
- The second most popular rating is 'TV-14,' which stands for content that may be inappropriate for minors under the age of 14.

❖ DATA PREPROCESSING



❖ DATA PREPROCESSING (Cont..)

- **Feature engineering**

```
df['prime_country'] = df['country'].apply(lambda x: x.split(",")[0])
df['prime_country'].head()
```

```
df['year_added'] =df['date_added'].apply(lambda x: x.split(" ")[-1])
df['month_added'] = df['date_added'].apply(lambda x: x.split(" ")[0])
```

```
#lets convert the category accourding to age wise for better understanding and thus making eda more interpretable
age_wise_rating = {   'TV-PG': 'Older Kids','TV-MA': 'Adults','TV-Y7-FV': 'Older Kids', 'TV-Y7': 'Older Kids',
                'TV-14': 'Teens', 'R': 'Adults', 'TV-Y': 'Kids', 'NR': 'Adults','PG-13': 'Teens','TV-G': 'Kids',
                'PG': 'Older Kids','G': 'Kids','UR': 'Adults','NC-17': 'Adults'}
```

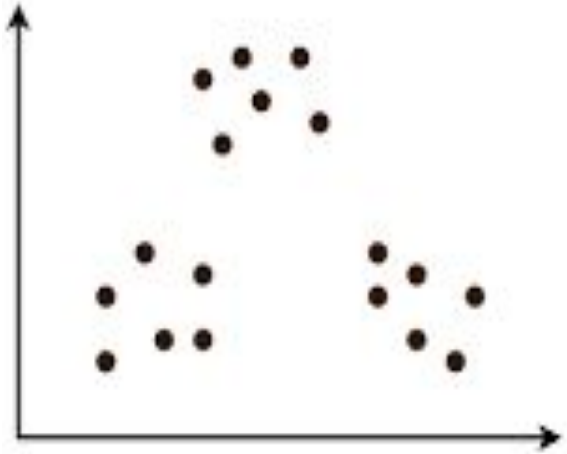
```
df['type'] = pd.Categorical(df['type'])
df['age_wise_rating'] = pd.Categorical(df['age_wise_rating'], categories=['Kids', 'Older Kids', 'Teens', 'Adults' ])
df['year_added'] = pd.to_numeric(df['year_added'])
```

❖ DATA PREPROCESSING (Conti…)

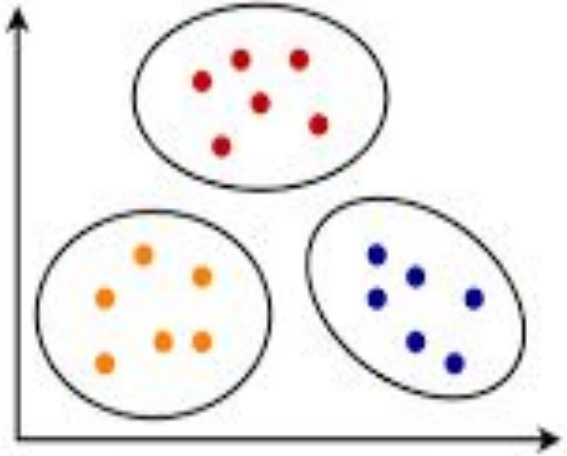
- **HANDLING TEXTUAL DATA**

	description	listed_in
0	futur elit inhabit island paradis far crowd sl...	intern tv show tv drama tv scifi fantasi
1	devast earthquak hit mexico citi trap survivor...	drama intern movi
2	armi recruit found dead fellow soldier forc co...	horror movi intern movi
3	postapocalypt world ragdol robot hide fear dan...	action adventur independ movi scifi fantasi
4	brilliant group student becom cardcount expert...	drama

❖ Model Building (k-means clustering)

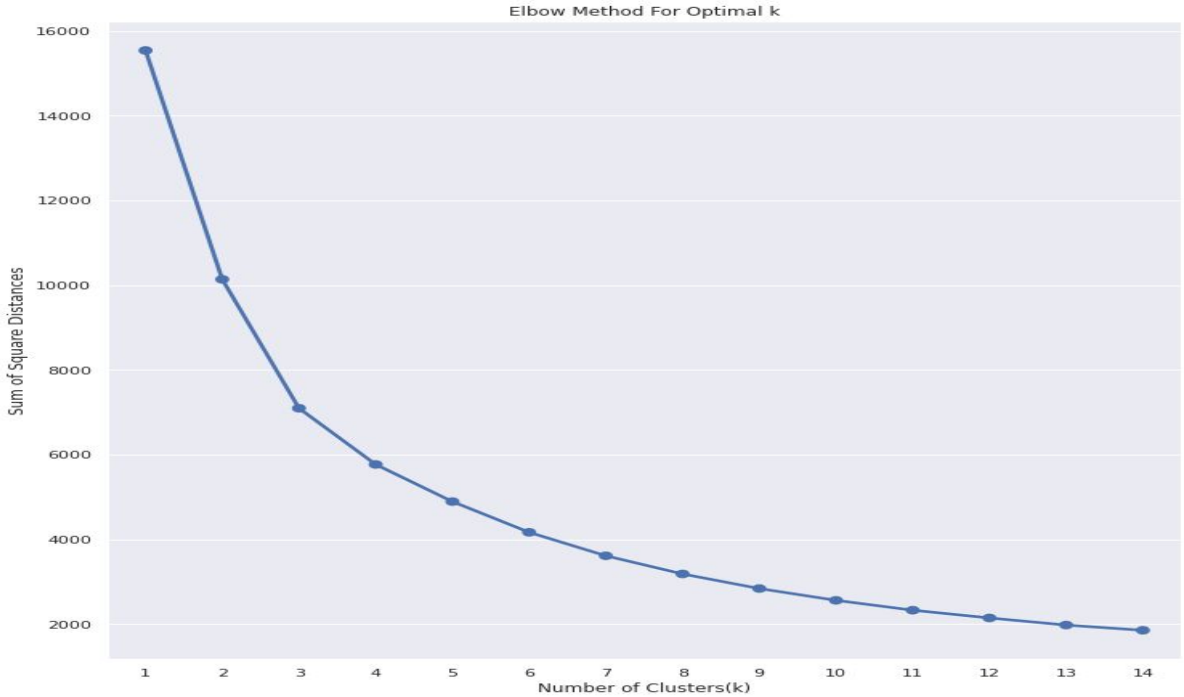


Before K-means



After K-means

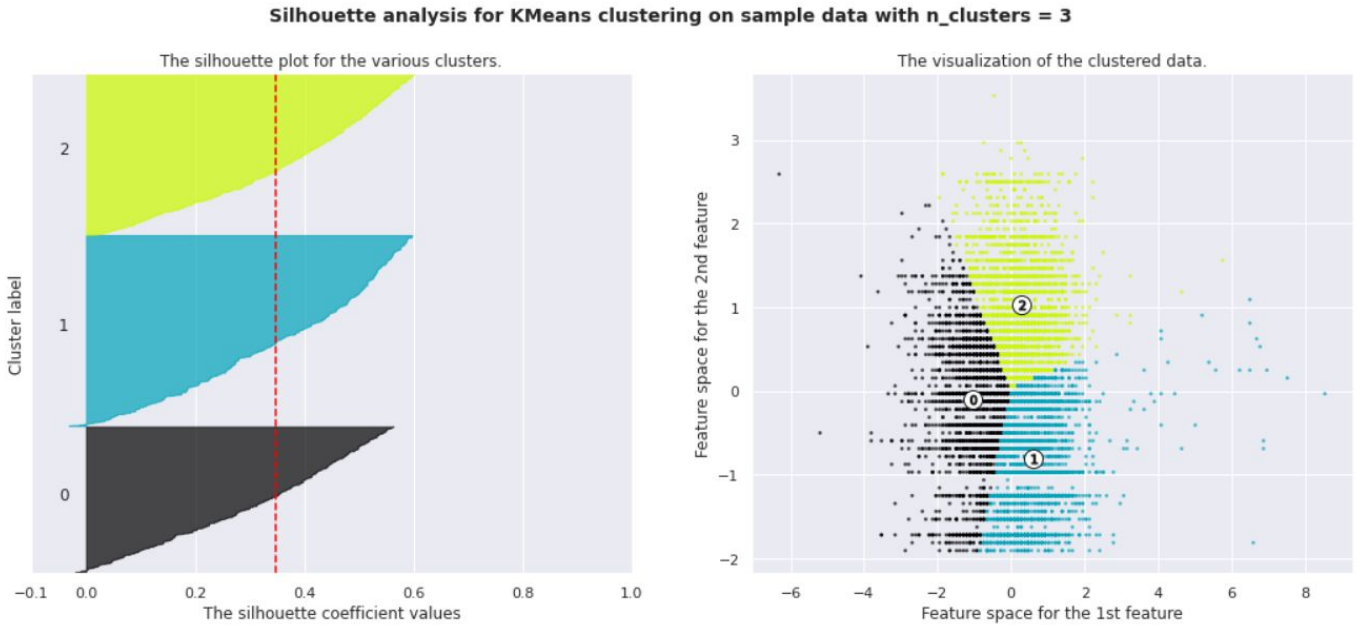
ELBOW METHOD



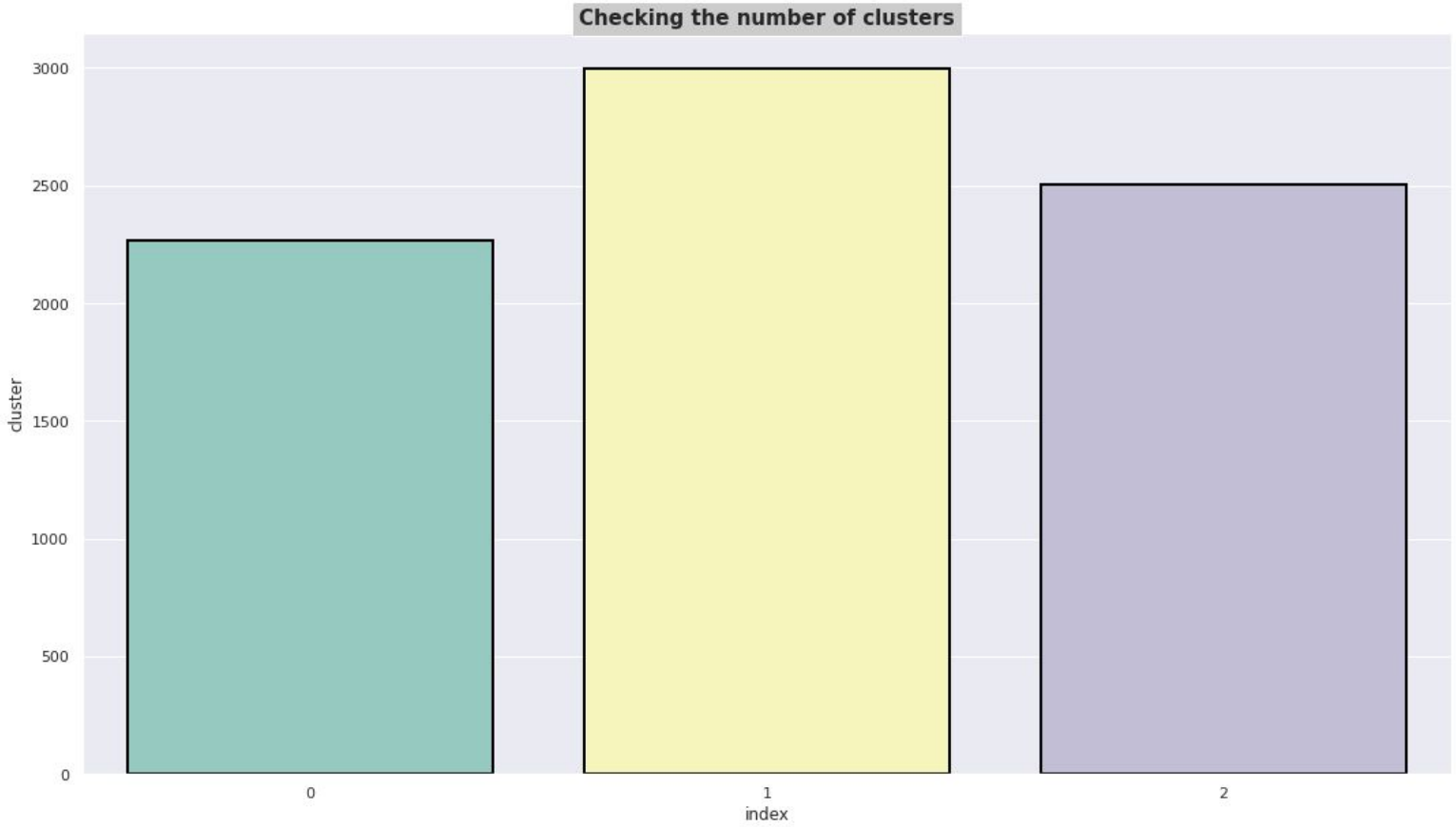
- **Number of clusters (k=3) as the error term is less and the computational cost isn't very high**

❖ SILHOUETTE SCORE

```
For n_clusters = 2, silhouette score is 0.3367485099208557
For n_clusters = 3, silhouette score is 0.3479889203807525
For n_clusters = 4, silhouette score is 0.31624113580823954
For n_clusters = 5, silhouette score is 0.3079420368105537
For n_clusters = 6, silhouette score is 0.32863886904073103
For n_clusters = 7, silhouette score is 0.32714712530542855
For n_clusters = 8, silhouette score is 0.32057164222284834
For n_clusters = 9, silhouette score is 0.3221156705055922
For n_clusters = 10, silhouette score is 0.3219237476661186
```



❖ CLUSTERS


Checking the number of clusters
3000
2500
2000
1500
1000
500
0
cluster
0
1
2
index

❖ CONCLUSION

- More movies are available on netflix than tv shows. 70% of the content is movies while the rest 30% is Tv shows.
- Netflix is most popular in United States followed by India and United Kingdom.
- Most of the content released on netflix in different cities is of type Movie except for Japan and South Korea.
- Movies and TV-shows belonging to 'TV-MA' rating are highest on netflix. It is for Mature audience only.
- Content produced by netflix in different countries is mostly targeted towards adult audience especially in countries like
- Spain and Mexico except for India where targeted audience is teen.
- Number of clusters used for K-means clustering are 3. We used elbow method to find the suitable number of clusters.
- Silhouette score for 3 clusters is 0.34 which is maximum.

THANK YOU!