**NETFLIX MOVIES & TV SHOWS CLUSTERING**

# **Shubham Bhadouria**

**Data science trainee,**

**AlmaBetter, Bangalore**

**Abstract**—Netflix is one of the leading over-the-top (OTT) platforms because of its reputation for offering users a wide variety of high-quality streaming movies as well as TV Shows. The reason why Netflix's services are so popular worldwide is that the company uses recent technologies like machine learning, deep learning and Artificial Intelligence to provide consumers with more appropriate and intuitive recommendation. The Aim of this project is to form the clusters based on K mean clustering and Agglomerative clustering. I did data preprocessing, text cleaning, Exploratory Data Analysis, Vectorization, implementing Clustering Models. Dataset is analyzed with CountVectorizer and TfidfVectorizer. The optimum number of Clusters has been determined using Silhouette Score, Elbow Method and Dendrogram. We figure out Exploratory Data Analysis, understanding what type of content is available in different countries. Clustering similar content by matching text-based features

*Keywords:* clustering, countvectorizer, TFIDF vectorizer, Kmeans, Agglomerative clustering

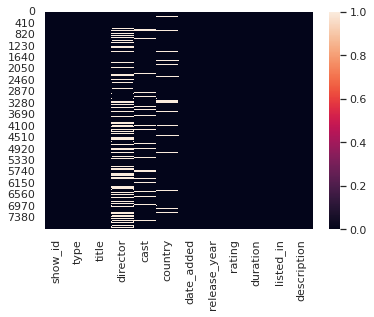
1. **Introduction**

Netflix is a streaming service that offers a wide variety of award-winning TV shows, movies, anime, documentaries and more on thousands of internet-connected devices. This business is profitable because users make a monthly payment to access the platform. However, customers can cancel their subscriptions at any time. Therefore, the company must keep the users hooked on the platform and not lose their interest. This is where recommendation systems start to play an important role, providing valuable suggestions to users is essential.

1. **Data Overview**

There are a total 7787 entities and 12 features in our dataset. About 30.67% data is missing in director, 9.22% in cast, 6.51% in country and 0.0898 % in rating.

**Missing values in dataset**



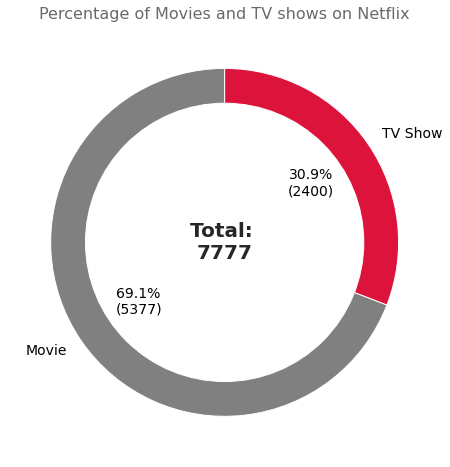
The attribute ‘director’, ’cast’, ‘country’, ‘date added', 'rating' consists of missing values.

1. **Data Cleaning**

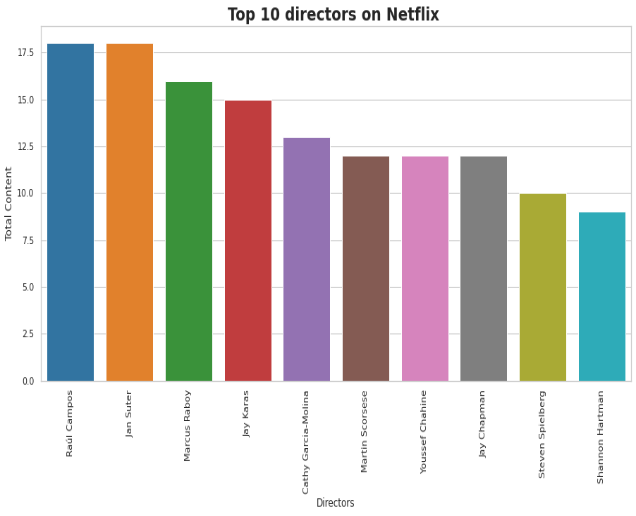
The attribute 'director', 'cast', 'country', 'date added', 'rating' consists of missing values. To tackle missing values, we replaced missing values of attributes 'country', 'director' and 'cast' by ‘Unknown’. The missing values of attribute ‘rating’ is replaced by mode values.

1. **Exploratory Data Analysis**

**Distribution of Movies and TV Shows**

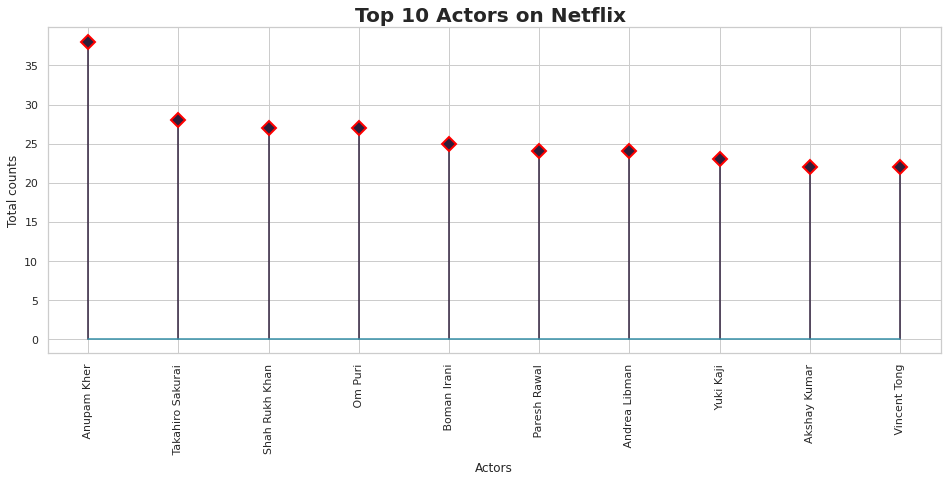
 69% of the content available on Netflix are movies; the remaining 31% are TV Shows. Netflix has 5377 movies, which is more than double the quantity of TV shows.

**Top 10 directors on the basis of the total number of shows or movies directed by them**



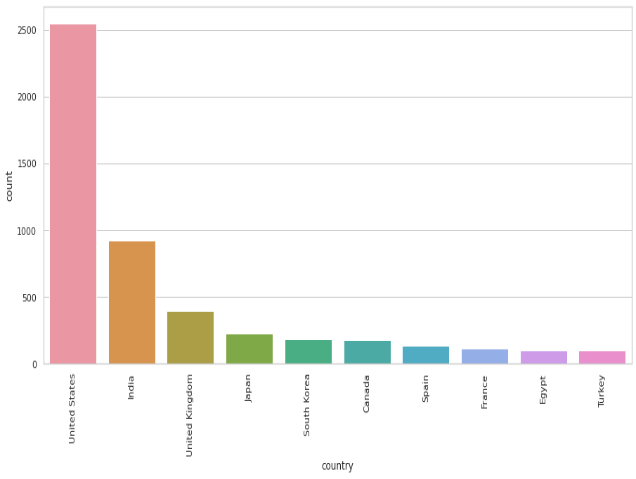
Raul Campos and Jan Suter has directed the most number of movies/Tv shows (total 18 both) followed by Marcus Raboy (total 16).

**Top actors based on the movies/Tv shows they acted.**



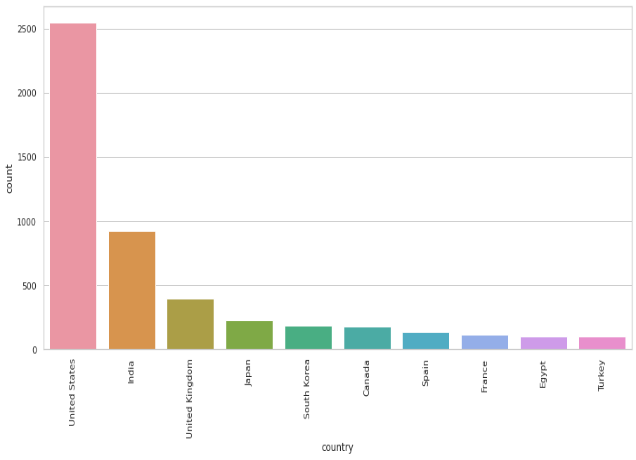
Anupam Kher has acted in the most number of movies/Tv shows (total 38) followed by Takahiro Sakurai (total 28).

**Top 10 actors based on total shows/movies in which they work**



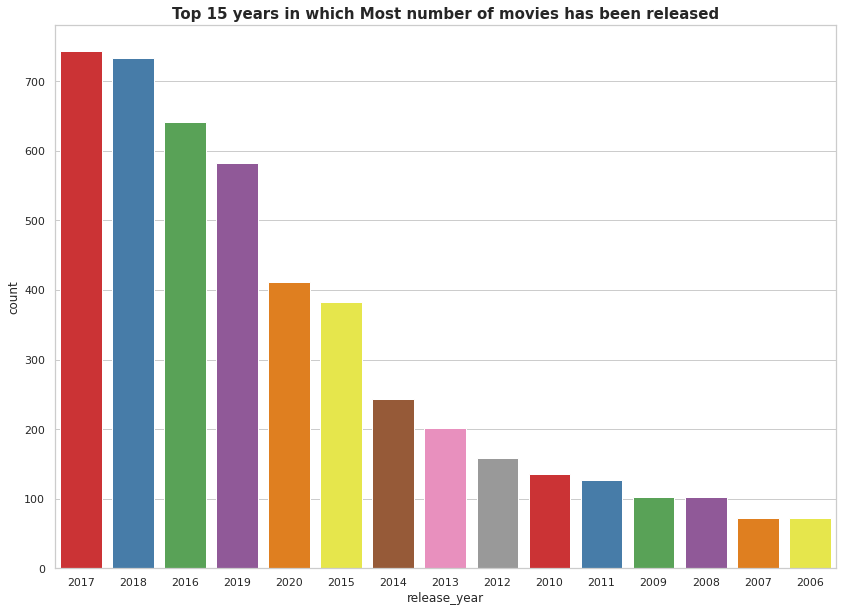
Anupam Kher has acted in the most number of movies/Tv shows (total 38) followed by Takahiro Sakurai (total 28).

**Top 10 countries where most number of movies/TV shows has been produced**



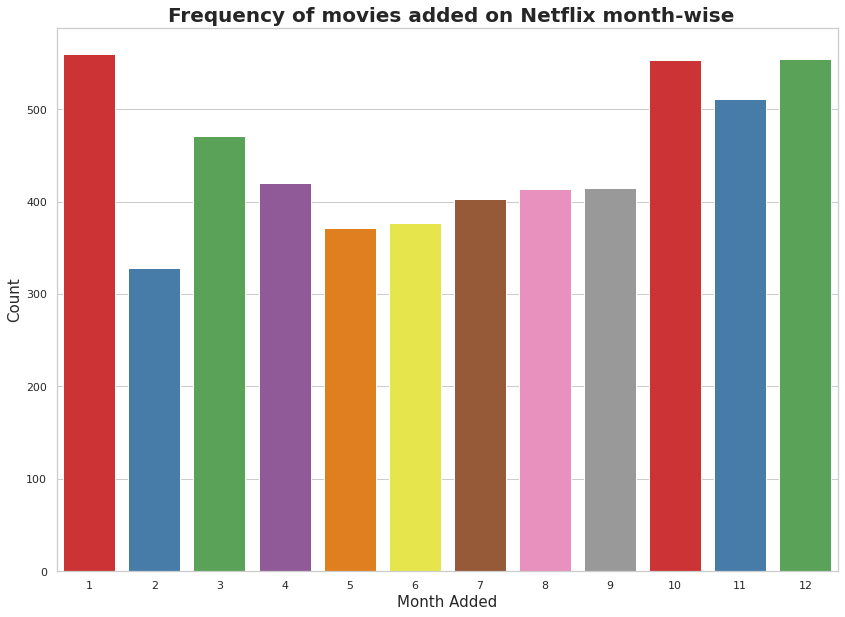
Most of Movies/TV shows listed on Netflix are produced in United States followed by India.

**Total movies released every year**

****

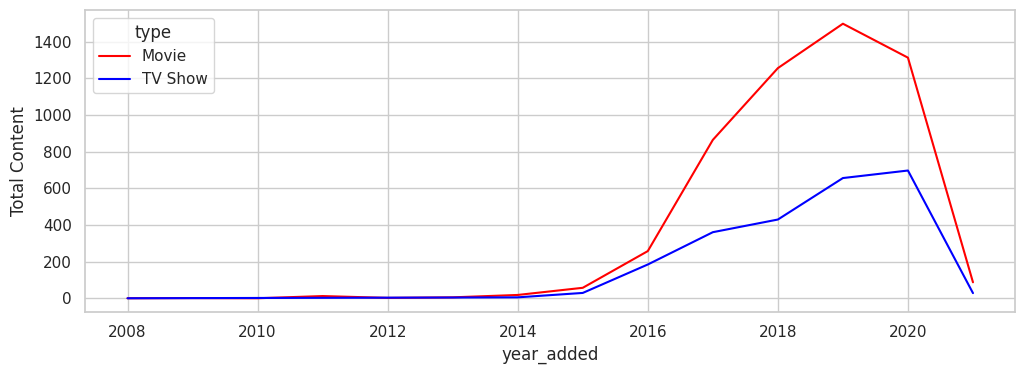
Most number of movies has been released in 2017 followed by 2018 and 2016.

**Movies added on Netflix month wise**



Most of the movies has been added in the month october, november, december and  january.

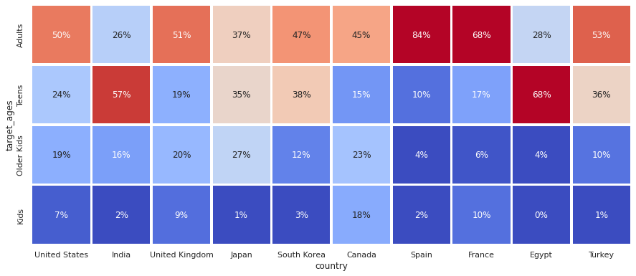
**Movies and TV shows added over the years on Netflix**

****

The number of movies and TV shows added on Netflix has been increasing significantly (almost exponentially) from 2015 to 2019 and after 2019, the number of movies and TV shows added on Netflix has been decreasing significantly due to Covid-19 pandemic.

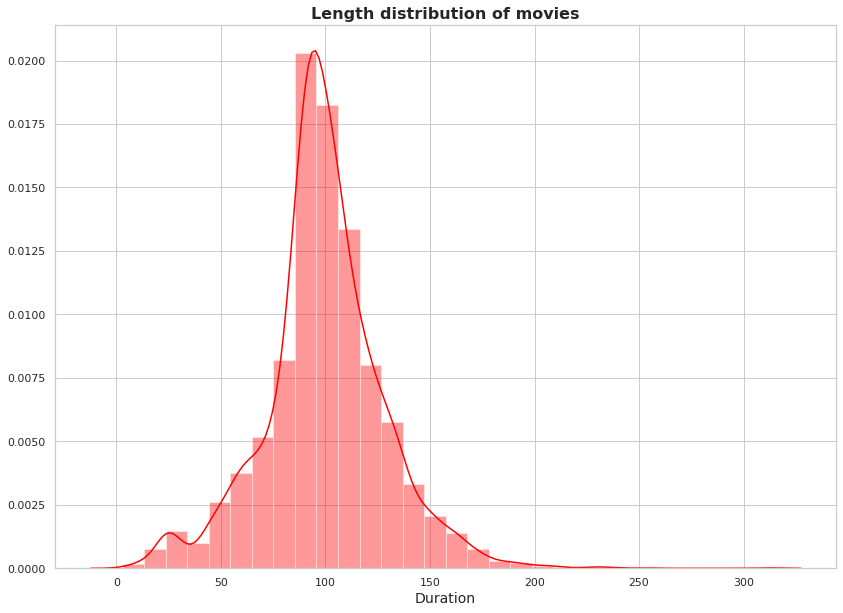
In 2019, maximum number of TV shows (total 656 shows) and movies (total 1497 movies) has been added on Netflix.

**Target audience proportion in the top ten countries where the movies/TV shows has been produced**

****

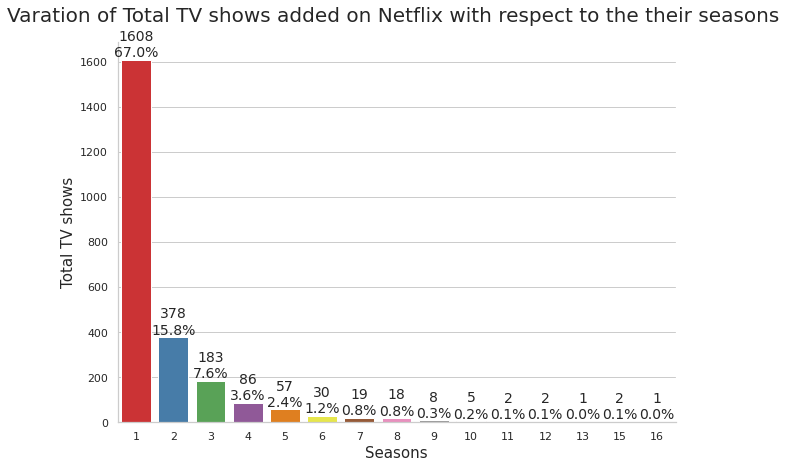
It is observed that content available for kids is less as compared to other categories. Content available for Adults is more in almost every country except India. In India, Most of the content is available for Teens. Netflix should focus on Teen and Adult Contents to generate maximum revenue.

**Length distribution of movies**

****

Average duration of movies is between 90 minutes to 120 minutes.

**Variation of total TV shows added on Netflix with respect to their seasons**

****

Most of the TV shows (approximately 1600) on Netflix has only 1 season.

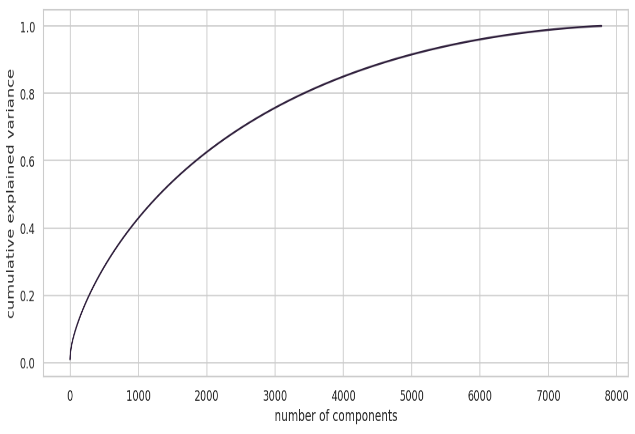
1. **Text Processing**

* Removed Punctuations
* Removed Stop words
* Removed non-ASCII characters
* Lemmatizing
* Tokenizing
* Vectorisation

After doing vectorization, to deal with the dimensionality curse we did PCA on the cluster data.

**Principal Component Analysis (PCA)**

Principal component analysis (PCA) is a technique for reducing the dimensionality of the datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance.

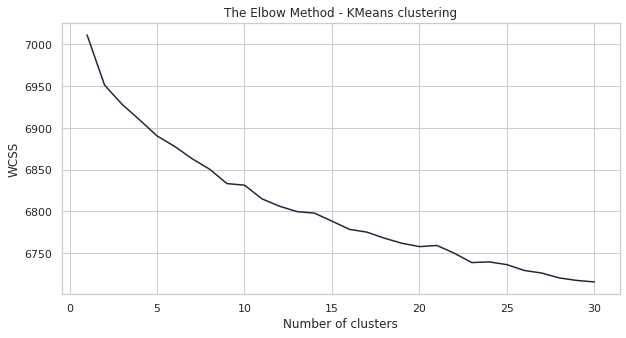
****

95% of the variance can be explained by 5000 components.

1. **Unsupervised machine learning clustering algorithms**
   1. **K means clustering algorithm**

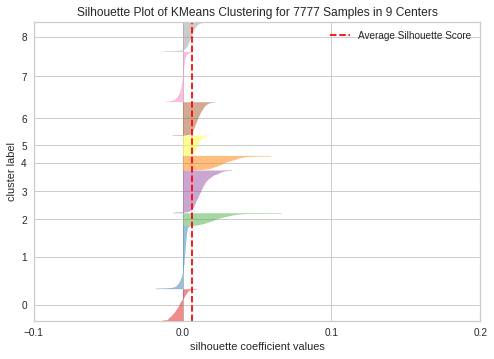
K-means clustering is a distance-based unsupervised clustering algorithm where data points that are close to each other are grouped in a given number of clusters/groups.

**6.1.1: Elbow method to find the optimum number of clusters**

****

From the above elbow plot, it is difficult to find the optimum number of clusters, but approximate optimum number of clusters can be considered as 9.

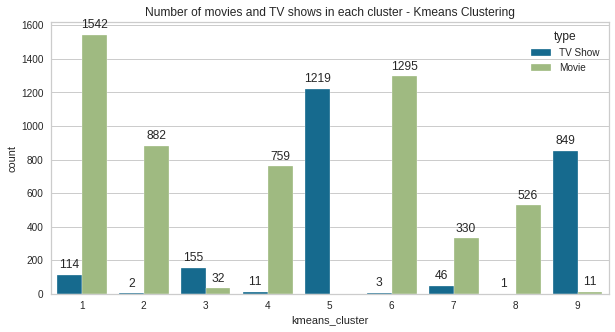
**6.1.2: Silhouette score method to find the optimum number of clusters**

****

For K = 9, silhouette score is 0.006245540833936832.

For K=9, all the clusters have a Silhouette score more than the average score of the dataset and there is not much fluctuations in the size of the clusters. So, the Silhouette plot approach gives us K = 9 as the optimal value.

**Total Movies/TV shows in each clusters**

****

The above graph shows number of movies and shows in each cluster.

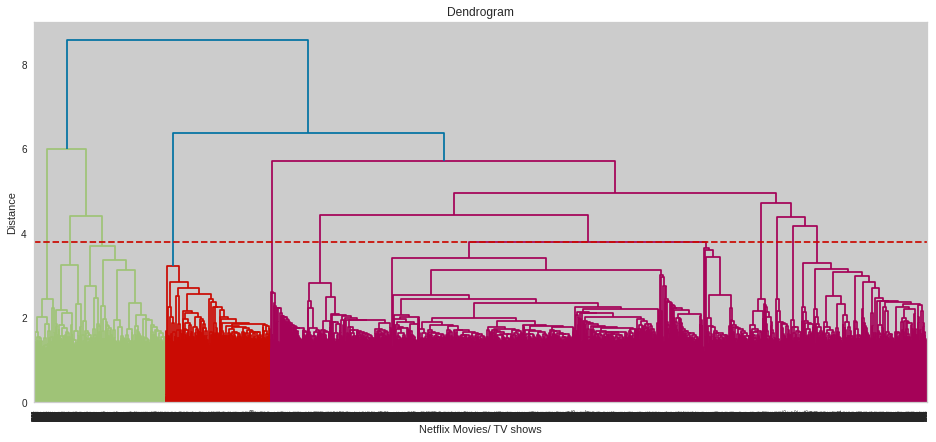
**Word cloud for movie description for cluster number 2**

****

**6.2: Hierarchal clustering**

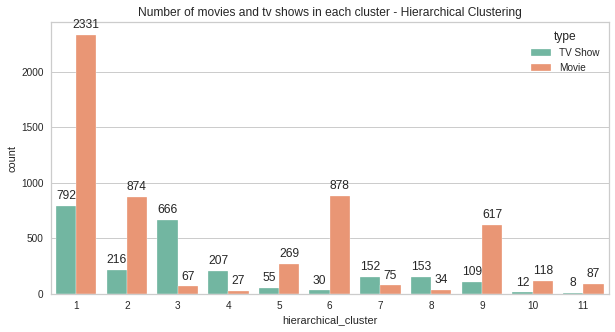
Hierarchical clustering is a method of cluster analysis that seeks to build a hierarchy of clusters.

**Dendrogram to find the optimum number of clusters**

****

As red dotted horizontal line cuts the 11 vertical lines in the dendrogram, therefore optimum number of clusters will be 11.

**Total Movies/TV shows in each clusters**

****

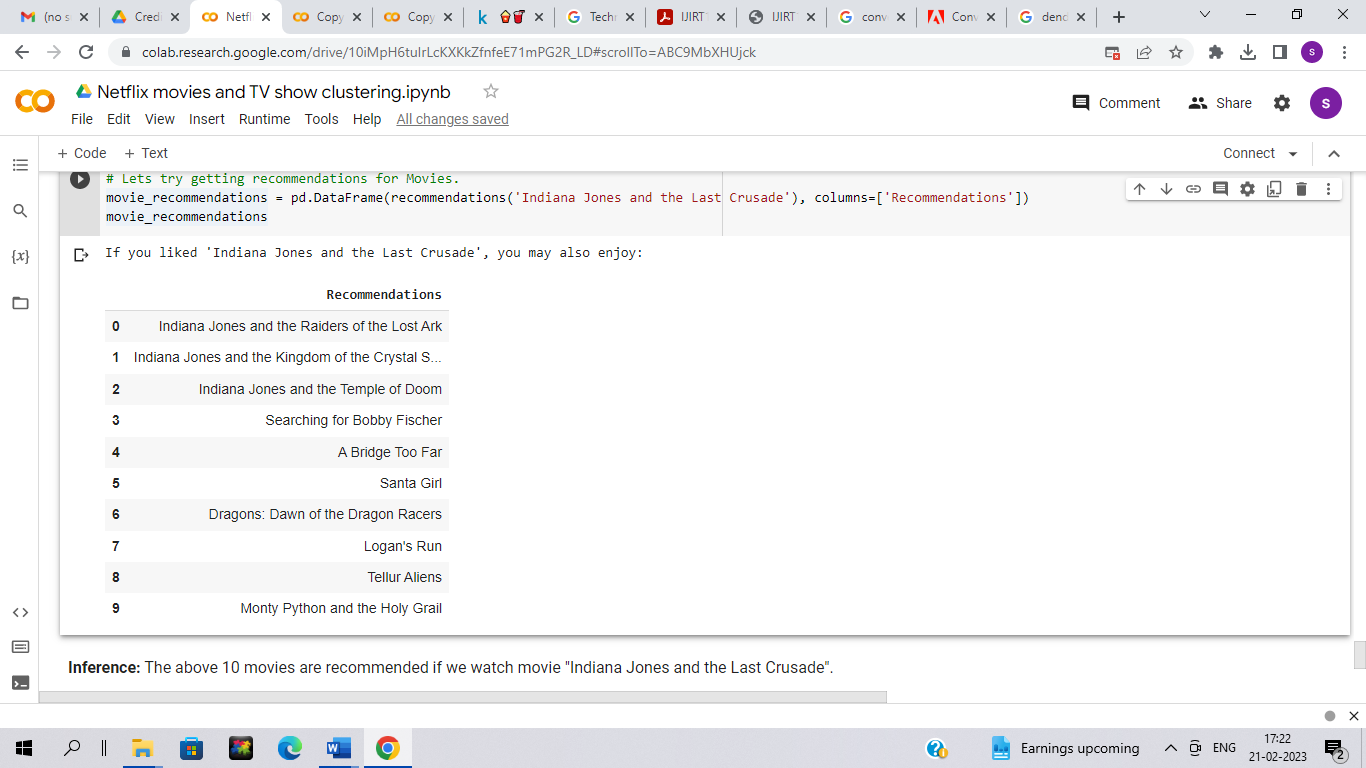
The above graph shows number of movies and shows in each cluster.

**Word cloud for movie description for cluster number 5**

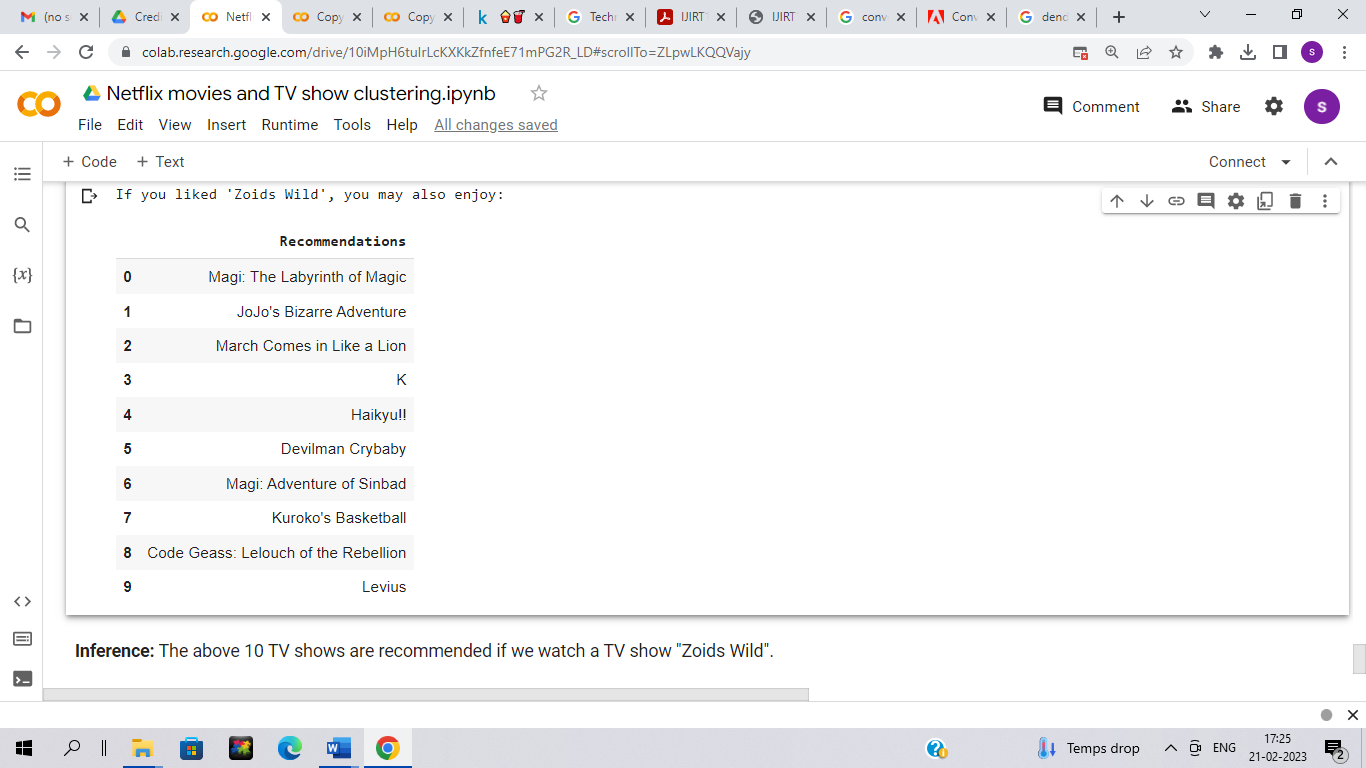
****

1. **Recommendation system**

Recommendation system is made using cosine similarity. Cosine similarity measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. It is often used to measure document similarity in text analysis.



The above 10 movies are recommended by recommender system if we watch movie "Indiana Jones and the Last Crusade".



The above 10 TV shows are recommended by recommender system if we watch a TV show "Zoids Wild".

1. **Conclusion**

The most important takeaways of this project are:

* Overall Netflix has more movies than the TV shows in a percentage of 69.1% against 30.9%. Netflix has 5377 movies, which is more than double the quantity of TV shows. Increase in the total movies added on Netflix in different years is much higher than the total TV shows added on Netflix.
* Raul Campos and Jan Suter has directed the most number of movies/Tv shows (total 18 both) followed by Marcus Raboy (total 16).
* Anupam Kher has acted in the most number of movies/Tv shows (total 38) followed by Takahiro Sakurai (total 28).
* Majority of the Movies/TV shows were produced in the United States, and the majority of the shows on Netflix were created for adults and teens means mature content is more popular on Netflix.
* Clusters are made using the k-means clustering algorithm and Agglomerative clustering algorithm.
  + With k-means clustering algorithm, optimal number of clusters came out to be 9. This was obtained through the elbow method and Silhouette score analysis.
  + With Agglomerative clustering algorithm, optimal number of clusters came out to be 11. This was obtained after visualizing the dendrogram.
* A simple content-based recommender system was created using cosine similarity which makes 10 recommendations to the user based on the type of movie/show they watched.

1. **References**

* <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html>
* <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>
* <https://towardsdatascience.com/a-complete-guide-to-recommender-system-tutorial-with-sklearn-surprise-keras-recommender-5e52e8ceace1>
* <https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html>
* https://scikit-learn.org/stable/tutorial/text\_analytics/working\_with\_text\_data.html