**Abstract:**

Netflix is the world's greatest on-line streaming carrier provider, with over 220 million subscribers as of 2022-Q2. It is essential that they correctly cluster the indicates that are hosted on their platform in order to decorate the consumer experience, thereby stopping subscriber churn.

We will be capable to recognize the suggests that are comparable to and special from one some other through growing clusters, which may also be leveraged to provide the shoppers customized exhibit tips relying on their preferences.

The purpose of this task is to classify/group the Netflix indicates into sure clusters such that the suggests inside a cluster are comparable to every different and the suggests in exceptional clusters are numerous to every other.

**Introduction:**

This pocket book ambitions to allocate the improvement associated to exploratory evaluation of insights associated to Netflix Movies and TV Shows. Also, this pocket book makes use of the equipment introduced on xplotter python bundle made via myself and posted on PyPI repository. This is a real precise effort for coding beneficial features for making the Exploratory Data Analysis system a lot greater less complicated for Data Scientists and Data Analysis thru delivering charts customization and matplotlib/seaborn plots with a little few trace of code. I surely hope you all experience it!

**Data Description:**

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flexible 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.

Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

**Attribute Information:**

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

Date\_added: Date it was added on Netflix

Release\_year: Actual Release year of the movie / show

Rating: TV Rating of the movie / show

Duration: Total Duration - in minutes or number of seasons

Listed\_in: Genre

Description: The Summary description

**Pre-processing:**

1. Finding Missing value

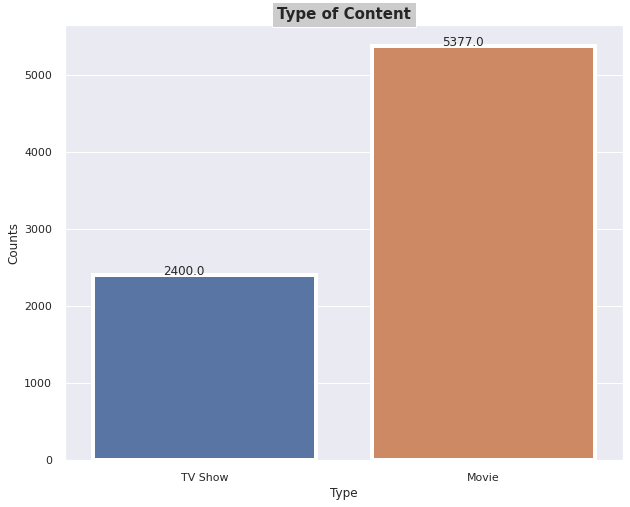
We have got Directors with 2389 null values in it then cast with 718 null values and country with 507 null values in it, also we have date\_added and rating with 10 and 7 null values in it.

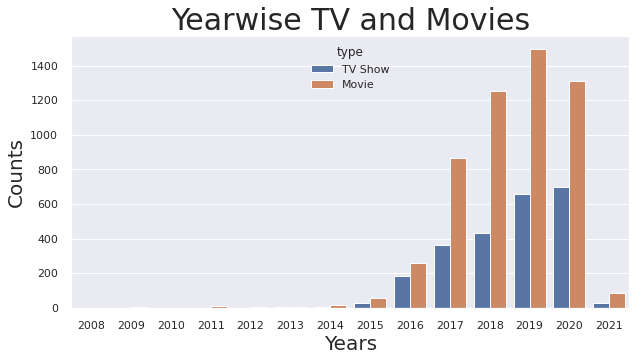
This says there are total 3671 null values in the data.

We replaced the null values in Directors, Cast, Country and Rating column with NaN.

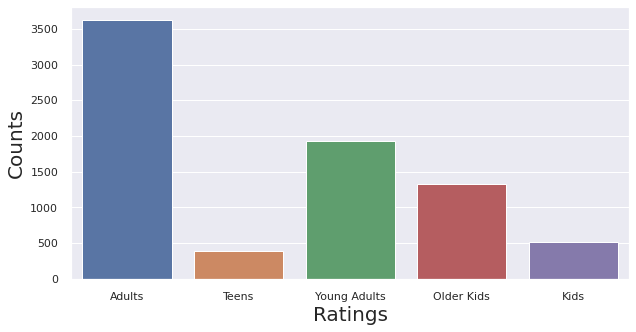
Then we dropped the date\_added column from the data since it was less important for the project.

**Exploratory Data Analysis:**

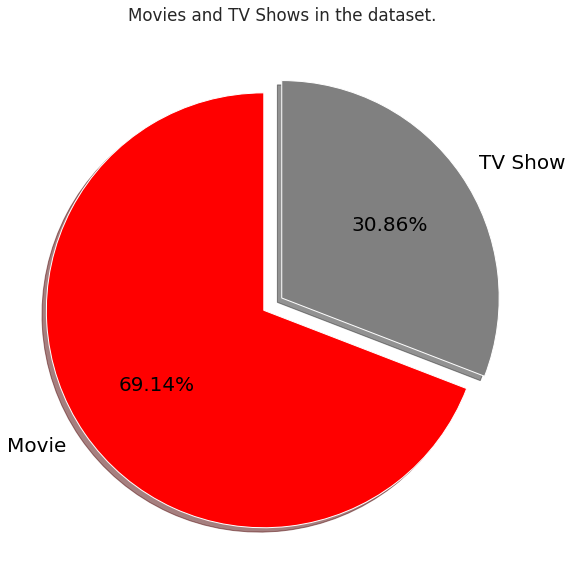
* By doing Data Analysis we find that United States has the highest content on the platform.
* Highest number of genres on the platform is Drama with 1384 content and comedy with 1074 content on the platform.
* There is total 5377 Movies on the platform and 2400 TV Shows. 
* TV Shows are less compare to Movies every year.

****

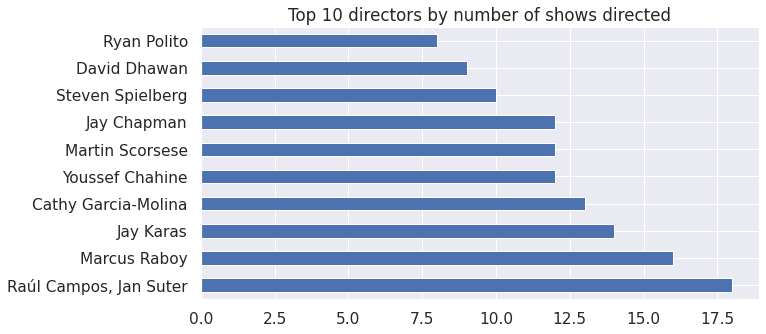
* Adults content is comparatively high from all other contents available in Netflix.

****

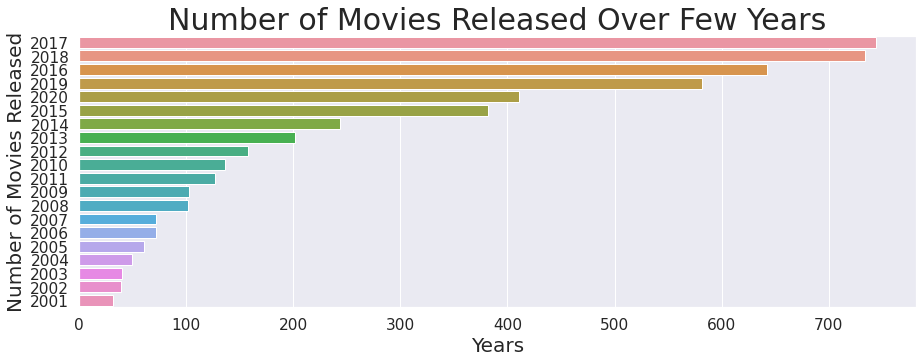
* Around 70% of contents available on Netflix are Movies and 30% of TV Shows.

****

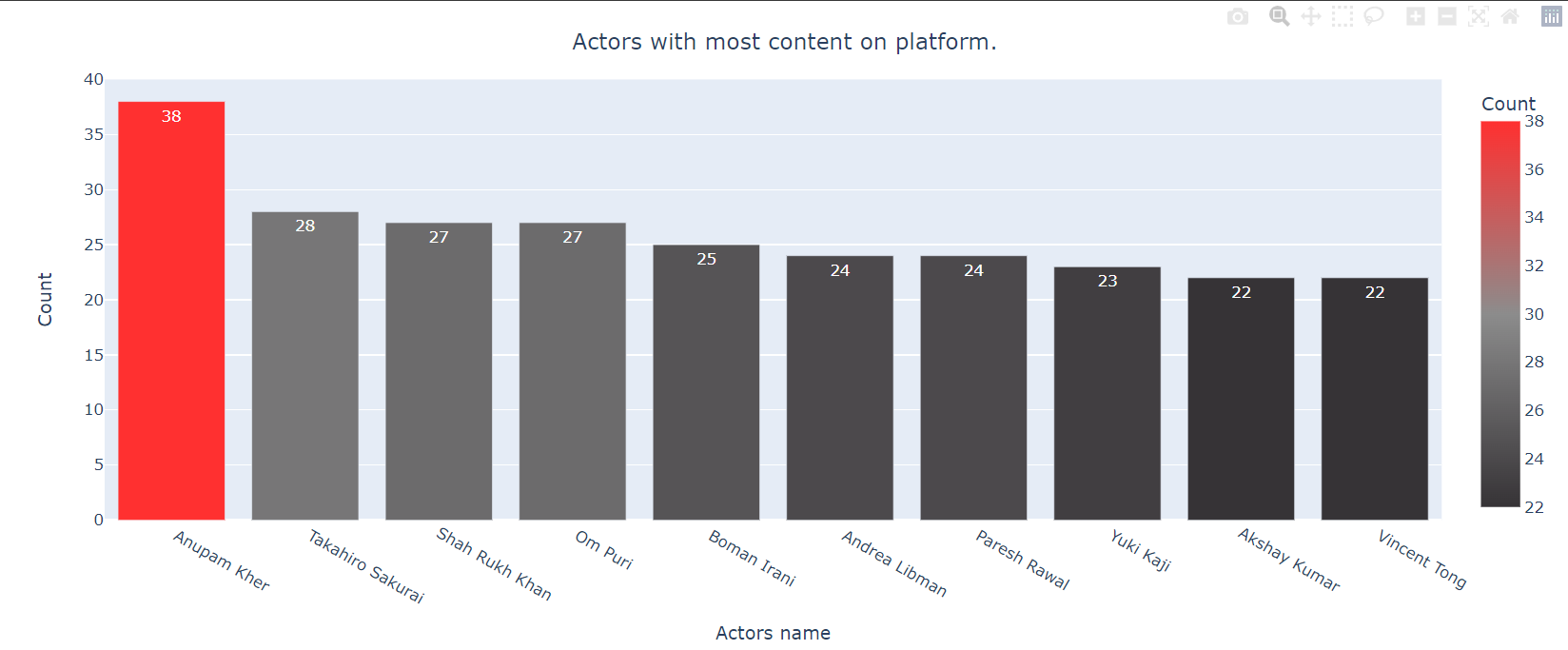
* We have done the analysis of top 10 directors of the shows available on Netflix. Where Raul Campos and Jan Suter has have directed over 17 shows.

****

* In 20 Years over 700 Movies were released in 2017 and 2018.

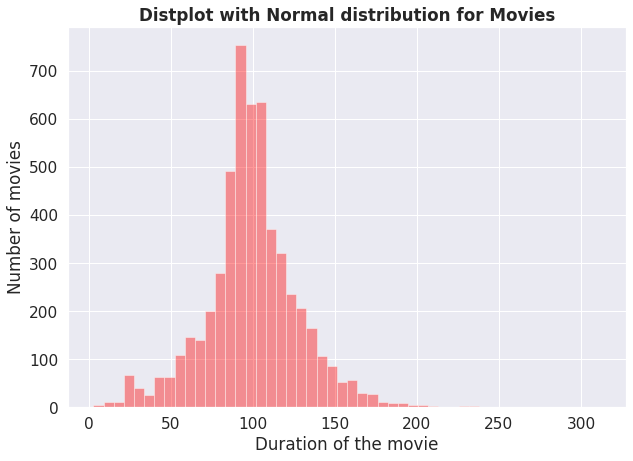
****

* Among all the actors’ contents available on Netflix, Anupam Kher has the highest contents on Netflix. He has 38 contents available on Netflix.

****

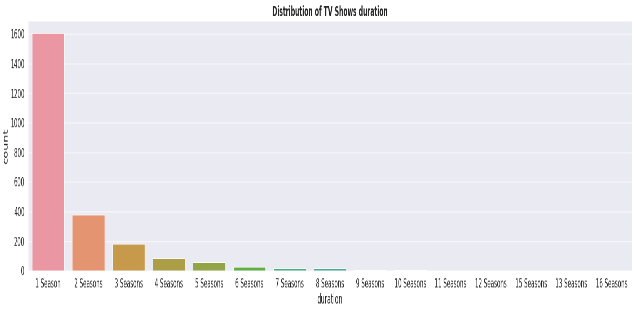
* We have done the distribution plot for checking the duration of the movies.

By the analysis we get that most of the movies have the duration of 50 to 150 mins.

****

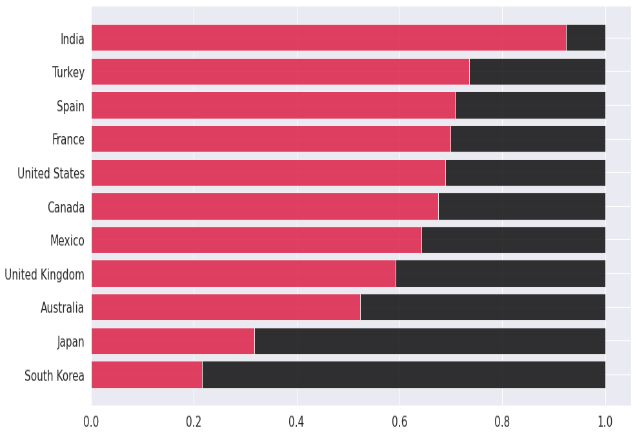
* We have done the count plot for getting the duration of TV Shows available on Netflix.

By analysis on the we get that most of the TV Shows available on Netflix have the duration of only 1 Season.

****

* There is a basic subplot for finding which country has the highest number of Movies on Netflix.

By this we get that India has the highest number of Movies on Netflix.

****

* **Data Pre-processing:**
* **Stopwords:**

Stop words are a set of commonly used words in a language. Examples of stop words in English are “a”, “the”, “is”, “are” and etc. Stop words are commonly used in Text Mining and Natural Language Processing (NLP) to eliminate words that are so commonly used that they carry very little useful information.

For example, in the context of a search system, if your search query is “what is a stop word?”, you want the search system to focus on surfacing documents that talk about stop word over documents that talk about what is a.

This can be done by maintaining a list of stop words (which can be manually or automatically curated) and preventing all words from your stop word list from being analysed. In this example, the words what is a could be eliminated, leaving only the words: stop word. This ensures that documents that are topically relevant have a high rank in your search results.

* **Lemmatization:**

Lemmatization is also used for the same purpose but it is more accurate than stemming. Example: If the word ‘go’ is in its past tense ‘went’ in the given sentence, Stemming can't arrive at 'go' from 'went'. But Lemmatization can.

Lemmatization is the method to normalize the text documents. The main goal of the text normalization is to keep the vocabulary small and remove the noise (unwanted stuff) which helps to improve the accuracy of many language modelling tasks.

* **Word Vectorization:**

Word Embeddings or Word vectorization is a methodology in NLP to map words or phrases from vocabulary to a corresponding vector of real numbers which used to find word predictions, word similarities/semantics.

* Word embeddings help in the following use cases.
* Compute similar words
* Text classifications
* Document clustering/grouping
* Feature extraction for text classifications
* Natural language processing.

After the words are converted as vectors, we need to use some techniques such as Euclidean distance, Cosine Similarity to identify similar words.

* **Dimensionality Reduction:**

Dimensionality reduction is a machine learning (ML) or statistical technique of reducing the number of random variables in a problem by obtaining a set of principal variables. This process can be carried out using a number of methods that simplify the modelling of complex problems, eliminate redundancy and reduce the possibility of the model overfitting and thereby including results that do not belong.

The process of dimensionality reduction is divided into two components, feature selection and feature extraction. In feature selection, smaller subsets of features are chosen from a set of many dimensional data to represent the model by filtering, wrapping or embedding. Feature extraction reduces the number of dimensions in a dataset in order to model variables and perform component analysis.

* **Clustering Analysis:**

In machine learning too, we often group examples as a first step to understand a subject (data set) in a machine learning system. Grouping unlabeled examples is called clustering.

As the examples are unlabeled, clustering relies on unsupervised machine learning. If the examples are labeled, then clustering becomes classification. For a more detailed discussion of supervised and unsupervised methods see Introduction to Machine Learning Problem Framing.

* **K- Means Clustering:**

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

* **Hierarchical Clustering:**

In Hierarchical Clustering, the aim is to produce a hierarchical series of nested clusters. A diagram called Dendrogram (A Dendrogram is a tree-like diagram that statistics the sequences of merges or splits) graphically represents this hierarchy and is an inverted tree that describes the order in which factors are merged (bottom-up view) or clusters are broken up (top-down view).

* **Agglomerative Clustering:**

Initially consider every data point as an individual Cluster and at every step, merge the nearest pairs of the cluster. (It is a bottom-up method). At first, every dataset is considered as an individual entity or cluster. At every iteration, the clusters merge with different clusters until one cluster is formed.

* **Word Cloud:**

Word clouds (also known as text clouds or tag clouds) work in a simple way: the more a specific word appears in a source of textual data (such as a speech, blog post, or database), the bigger and bolder it appears in the word cloud.

A word cloud is a collection, or cluster, of words depicted in different sizes. The bigger and bolder the word appears, the more often it’s mentioned within a given text and the more important it is.

* **Content Based Recommendation System:**

A Content-Based Recommender works by the data that we take from the user, either explicitly (rating) or implicitly (clicking on a link). By the data we create a user profile, which is then used to suggest to the user, as the user provides more input or take more actions on the recommendation, the engine becomes more accurate.

This type of recommender system is hugely dependent on the inputs provided by users, some common examples included Google, Wikipedia, etc. For example, when a user searches for a group of keywords, then Google displays all the items consisting of those keywords. The below video explains how a content-based recommender works.

**Conclusion:**

1. Dataset contains total 7787 rows and 12 columns.
2. The dataset contains total 3631 null values.
3. United States contains highest contents on Netflix i.e., 3055 followed by India with 923 contents on Netflix.
4. Netflix contains 5377 Movies and 2400 TV Shows.
5. 2019 was the year which has the highest number of Movies and 2020 was the year which has the highest number of TV Shows.
6. In 2021 the count of TV shows and Movies certainly decreased.

Contents available on Netflix is more Adult followed by Young Adults.

1. Netflix contains around 1400 Drama content and around 1100 comedies
2. On Netflix actors which has highest content is Anupam Kher with 38 contents followed by Takahiro Sakurai with 28 contents.
3. Most of the Movies has the duration between 50 to 150 minutes.
4. Most of the TV Shows has the duration on 1 season.
5. India has the highest number of Movies on Netflix.
6. We used Principal Component Analysis (PCA) to handle the curse of dimensionality. 4000 components were able to capture more than 80% of variance, and hence, the number of components were restricted to 4000.
7. Then clusters were built using the Agglomerative clustering algorithm, and the optimal number of clusters came out to be 12. This was obtained after visualizing the dendrogram.
8. A content-based recommender system was built using the similarity matrix obtained after using cosine similarity. This recommender system will make 10 recommendations to the user based on the type of show they watched.