**Capstone Project**

**NETFLIX**

**Shajad**

**Contents:**

* **Abstract**
* **Introduction**
* **Problem Statement**
* **Steps Involved**
* **Data Visualization**
* **Methods and algorithm that we used**
* **Conclusion**
* **References**
* **Abstract:** This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flexable 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.

**Data Description**

Most of the fields are self-explanatory. these points are:-

* **Show id --> Unique id for every movie/tv show**
* **type --> identifier -A movie or tv show**
* **title --> Title of the movie and tv show**
* **director --> director of the show**
* **cast --> Actors involved**
* **country --> Country of production**
* **Date added --> date it was added at Netflix**
* **Release year --> Actual release year of the show**
* **rating --> TV raring of the show**
* **duration --> Total duration in minutes or no. of the seasons**
* **Listed in --> Genre**
* **description -->The summary description**

**Introduction:**

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

**Problem Statement:**

This data set contains information of Tv shows and movies available on Netflix. that information includes such as Duration, Added date, Release year. Directors, Description etc.

Explore and analyze the data to discover important factors that are most likesd by the customres.

We will tackle the problem statement in the following steps:

Step 1: Data Overview.

Step 2: Data Preparation and Cleaning.

Step 3: Visualizing the Univariate and Bivariate features.

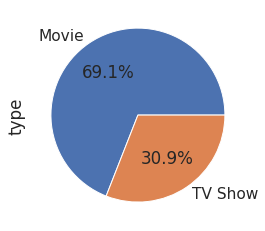
Step 4: Correlation Analysis.

Step 5: Concluding Analysis.

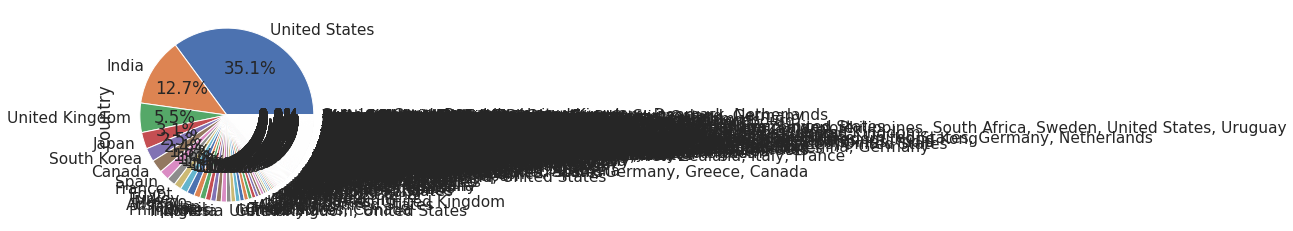
**Steps Involved:**

* **Importing Packages:** Importing the various libraries that will help us analyze our dataset properly with visual graphs.
* **Data Overview :** We load the data and go over the basic features, shape and datatypes of various variables.
* **Data Preparation and Cleaning:** We use various features of python to create combined features of date and time and other variables which can be simplified. We also drop columns that we used to create the new columns. Finally we take care of outliers.
* **Univariate and Bivariate Analysis:** We use seaborn and matplotlib to plot graphs starting with one variable graphs and then plotting two or more variable graphs to understand the variables and it spread and range with its frequency
* **Concluding Analysis:** Here after plotting and analyzing all the data we finally make predictions and remarks about our analyzes.

**Data Visualization:**

****

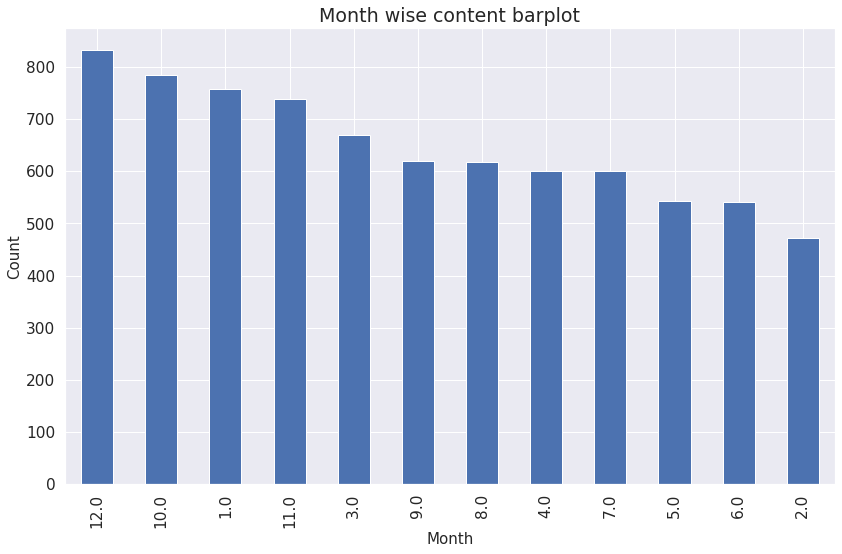
**In this data there are 69% are movies and 30.9% are tv shows present in this data**

****

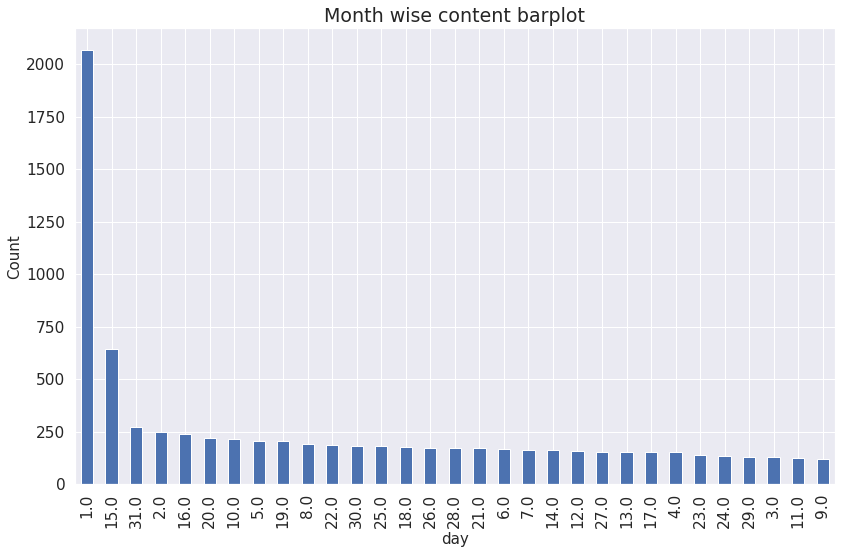
* **according to this plot**
* **there are 35.1% movies and tv shows are from the United State.**
* **12.7% movies and tv shows are from the India.**
* **5.5% tv shows and movies from the United Kingdom.**
* **3.1% tv shows and movies are from the Japan.**
* **2.5% movies and shows are from the South Korea**

****

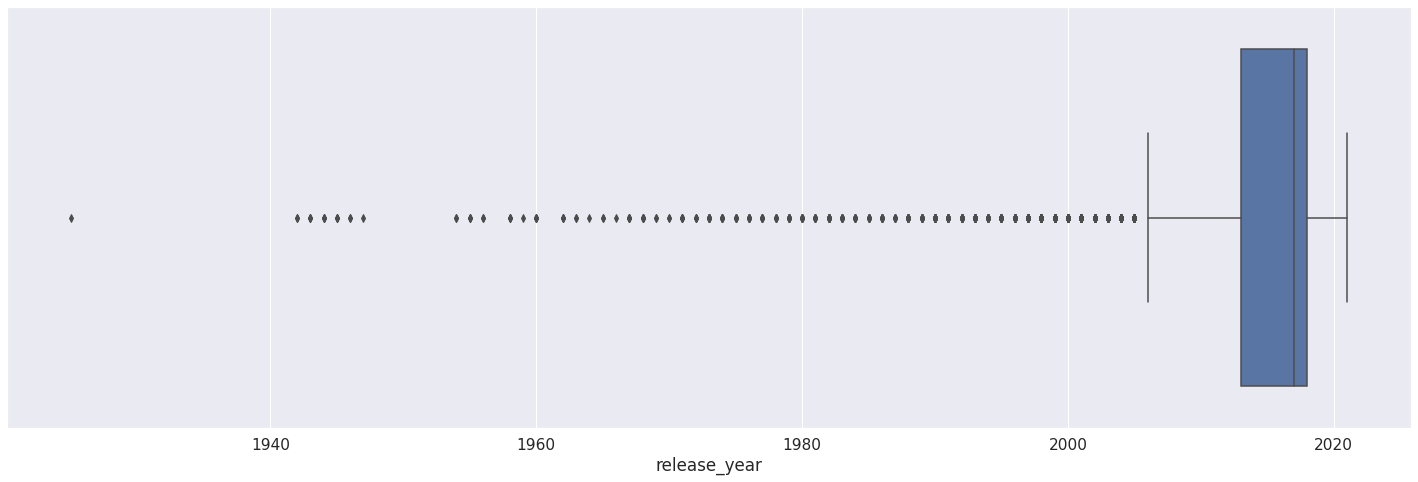
* **According to this lot we show that more then 2000 movies produce in 2019**
* **2000 movies and shows produce in 2020**
* **1600 movies and shows produce in 2018 aproximatly.**
* **1200 movies and shows are produce in 2017**
* **less them 500 shows and movies produce per year before 2017**

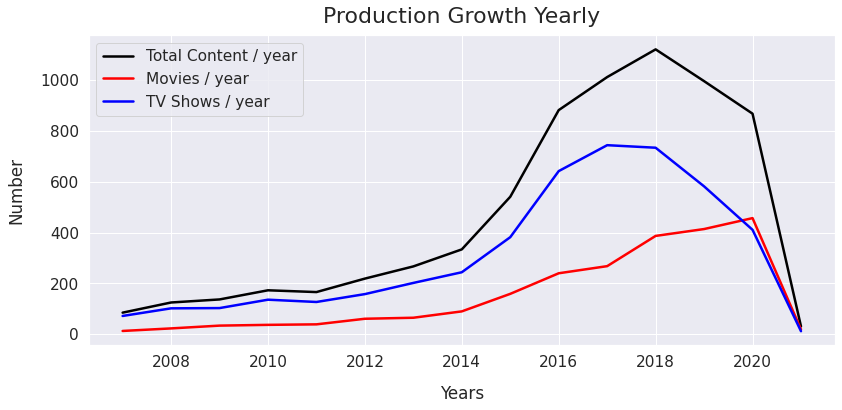
****

**in this plot we show that in December more then 800 movies and shows added on net flix and then November between 750-800 movies and shows are added on Netflix and then January 750 shows and movies are added on January and February is the least movies and shows added month.**

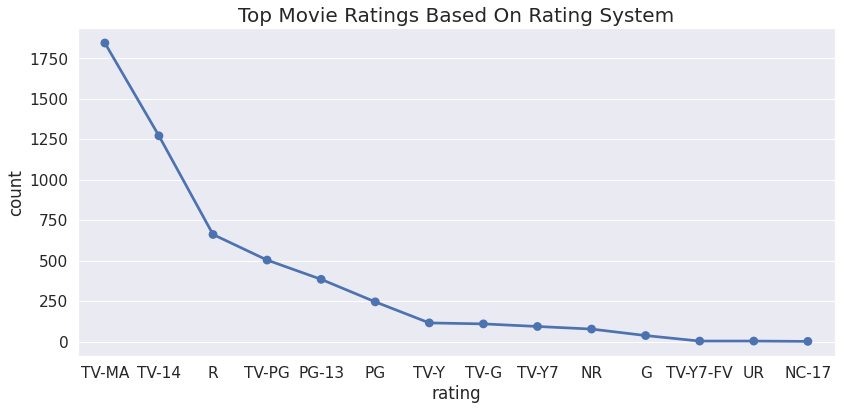
****

**according to this plot 1st days of every month to most movies and shows are added on the Netflix and then the 15th day of month the movies and the shows added on the Netflix between the 650to 700. and then the last day of the month is most released day of Netflix it is between the 200 to 250.**

**According to this plot we see that after the 2012 to 2021. is founded as outliers and we delete this outliers.**

****

**there is an boost coming from 2014 to 2021**

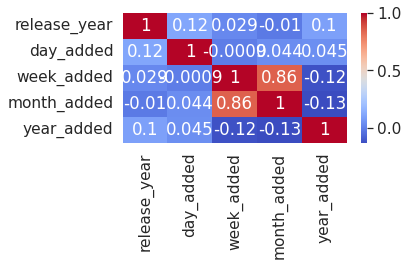
****

**TV-MA (For Mature Audiences)1900 movies come under this rating.**

**TV-14 ( May be unsuitable for children under 14 ) 1250 movies come under this rating.**

**TV-PG ( Parental Guidance Suggested ) 500 movies come under this rating.**

**NR ( Not Rated ) 150 movies come under this rating.**

****

* **In this correlation heatmap we find that the 86% correlation between the week added and the month added.**

**Conclusion:**

This project is about the Credit card default prediction analysis. When we start this project, it is very difficult to understand and look very complex. In first step we upload the data and then we start to find the duplicate values and nulls values and missing values etc. in this dataset we found the some values are missing in Cast and Directors columns so we decide to remove that colums also in release year there are sum values also missing we fill that values with mean and we also fill the country columns null values with mode in. After completion of 1st step in 2nd step we plot the data in form of graph, pie plot and bar plot and line plot and also box plot, and in box plot we find some outlier so we remove the outliers. And we also find some hidden information by plot the graph. In this dataset there are two types of contents 30.86% is Tv shows and 69.14% is movies includes. We have reached a conclusion from our analysis from the content added over years that Netflix is focusing movies and TV shows (From 2016 data we get to know that Movies is increased by 80% and TV shows is increased by 73% compare). From the dataset insights we can conclude that, most number of TV Shows released in 2017 and for Movies it is 2020. On Netflix USA has the largest number of contents. And most of the countries preferred to produce movies more than TV shows. And then India comers on number 2 hos produce the largest movies and the tv show. Most of the movies are belonging to 3 categories. TOP 3 content categories are, International movies, dramas, comedies. there are 35.1% movies and tv shows are from the United State. 12.7% movies and tv shows are from the India. 5.5% tv shows and movies from the United Kingdom.3.1% tv shows and movies are from the Japan. 2.5% movies and shows are from the South Korea. We also divide the data into month year and the day of month and add that columns into the data frame and we also create some Data frame in our data set. In text analysis (NLP) I used stop words, removed punctuations, stemming & TF-IDF vectorizer and other functions of NLP. Applied different clustering models like K-means, hierarchical, Agglomerative clustering, DBSCAN on data we got the best cluster arrangements. By applying different clustering algorithms to our dataset . we get the optimal number of cluster is equal to 3 .So we decide to take the as no of cluster in this data because it is best no in our conclusion.

**References:**

* GeeksforGeeks
* Stackoverflow
* Almabetter
* Youtube
* Github