Netflix Data Analysis

Rohan Shah (202218002), Karan Parashar (202218004), Rutul Patel (202218036)

Abstract

Our project work is on exploratory data analysis of the content on Netflix (OTT platform). The dataset is being taken from Kaggle, it is composed of details such as *Title, Category, Rating, etc* for all the content available on Netflix. Furthermore, we have also implemented a Pytrends API for displaying the ranking of countries by relative number of searches for word 'Netflix' on Google.

Index:

1. **Importing libraries**:

    a. **pandas**: Pandas is a Python library used to analyze data. Some of the basic operations such as creating DataFrame/s. Importing .csv or .json files can be performed using this library. Moreover, one can also perform delete operations, clean empty cells, remove duplicates, and data formatting.
        More about pandas

    b. **matplotlib.pyplot:** matplotlib library is used to plot interactive visualizations in Python and matplotlib.pyplot is a collection of functions that makes matplotlib work like MATLAB. Simply speaking, it is used for creating plots.
        More about matplotlib.pyplot

    c. **seaborn**: seaborn contains a number of patterns and plots for data visualization. It is more comfortable in handling pandas data frames which use basic sets of methods to provide beautiful graphics in python. Here the seaborn library is used for creating a heatmap of the Netflix dataset.
        More about seaborn

    d. **numpy**: NumPy (Numerical Python) is a Python library used for working with arrays. It supports all array-related operations such as indexing, iterating, join/split, search, sorting, filtering, and many more. Moreover, it is also used to create different types of distributions.
        More about NumPy

    e. **plotly.express**: plotly.express is used for creating entire figures at once and displaying them. It supports the other library for creating plots but this helps in displaying them.
        More about plotly.express

    f. **TrendReq**: It is an interface to download Google Trends data and generate reports in csv format.
        More about TrendReq

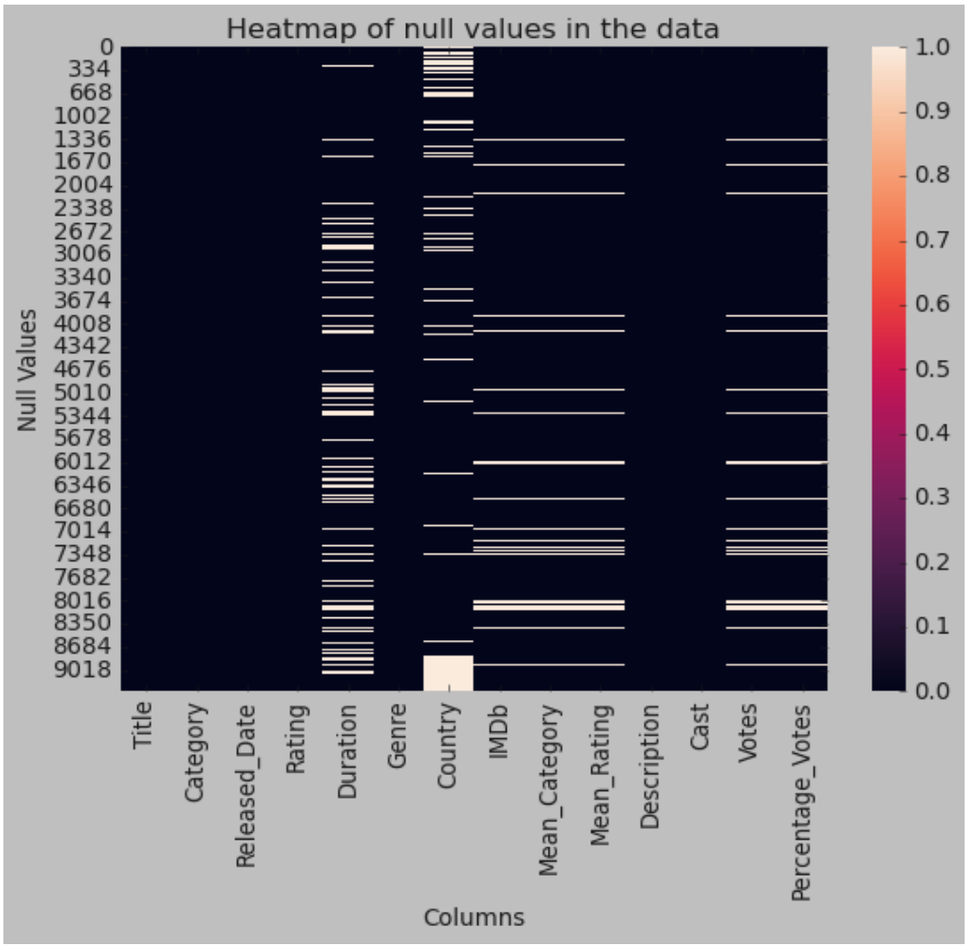
2. **Data information**: We have made a function for displaying all the basic information about the data, *such as info*, *shape (number of attributes and columns)*, *data types of each column*, *duplicate values*, *null and not-null values* in the data, *column names*, *unique records* in each column.

3. **Data visualization and interpretation**:

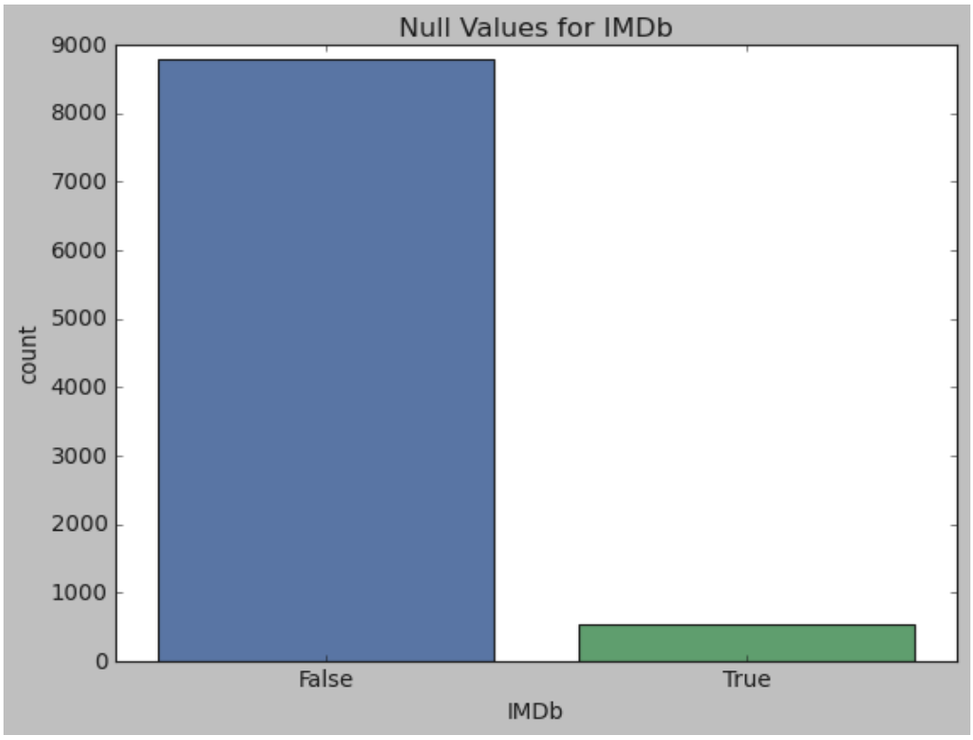
a. **Heatmap for null values**: The below heatmap shows the null values in each column and their corresponding approximate index number.\

As we can see, the columns Title, Category, Released_Date, Rating, Genre, Description, Cast and day do not contain any empty values indicated by full black stripes while the columns Duration, Country, IMDb, Mean_Category, Mean_Rating, Votes and Percentage_Votes contain null values, denoted by while lines in between.

Not the precise location of missing data but this gives us an overall understanding of the data and where the null values are highly dense.

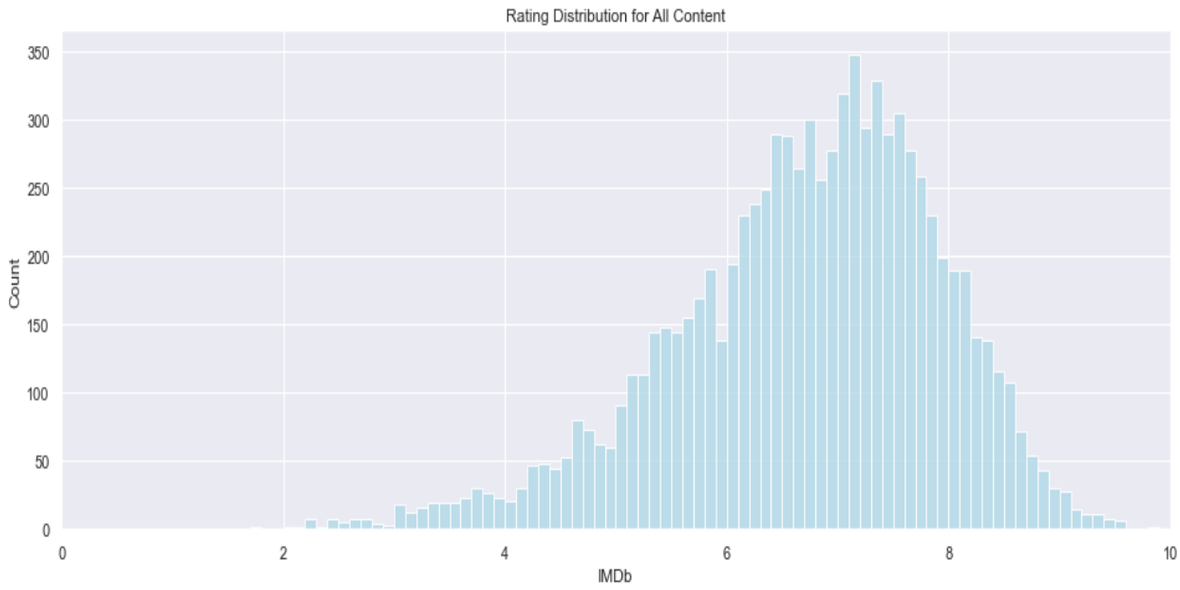


b. **Countplot of Null Values for IMDb**: The below countplot shows the number of non-empty IMDb (8784 non-empty values).



c. **Rating distribution for content**: The below graph shows the number of content on Netflix, for each IMDb value. We can conclude that the majority of the content (350 TV Shows / Movies) is rated 7.1 IMDb.

Negative skewness of the distribution indicates that there are a few contents that are rated very low on IMDb scale, which separates them from the others.



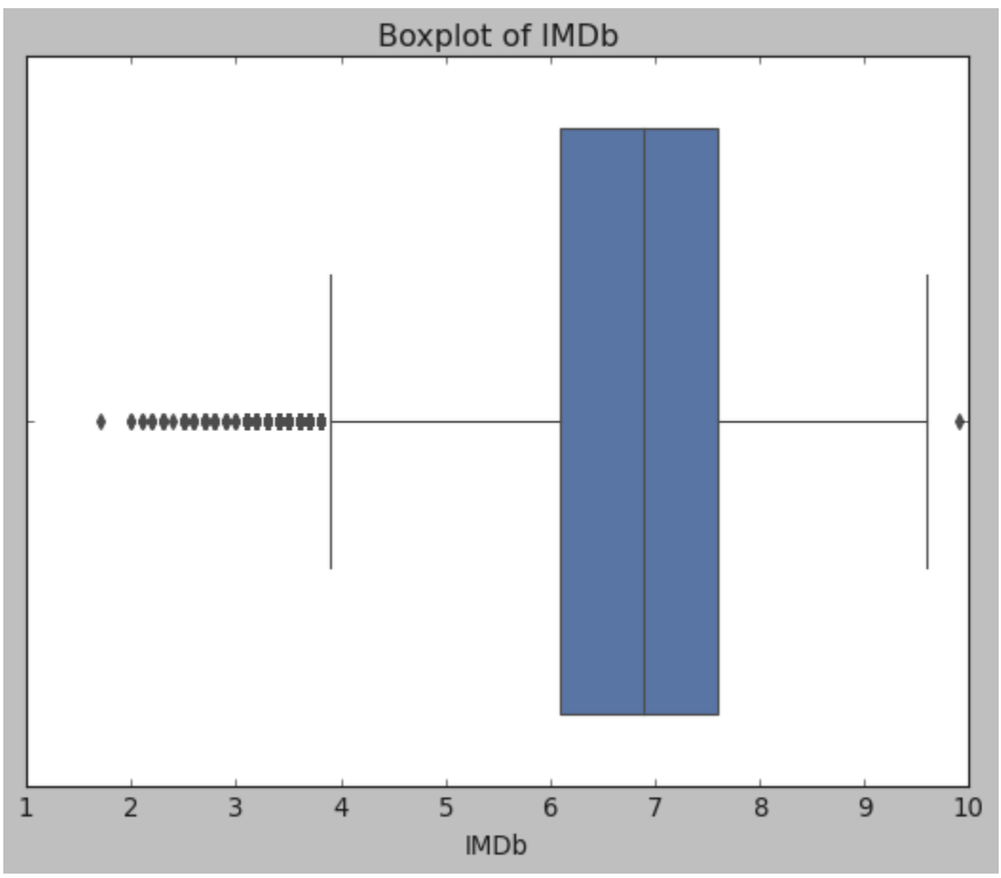
Five number summary of the above distribution

count	mean	std	min	25% (Q1)	50% (Q2)	75% (Q3)	max
8784	6.764515	1.214840	1.700000	6.100000	6.900000	7.600000	9.900000

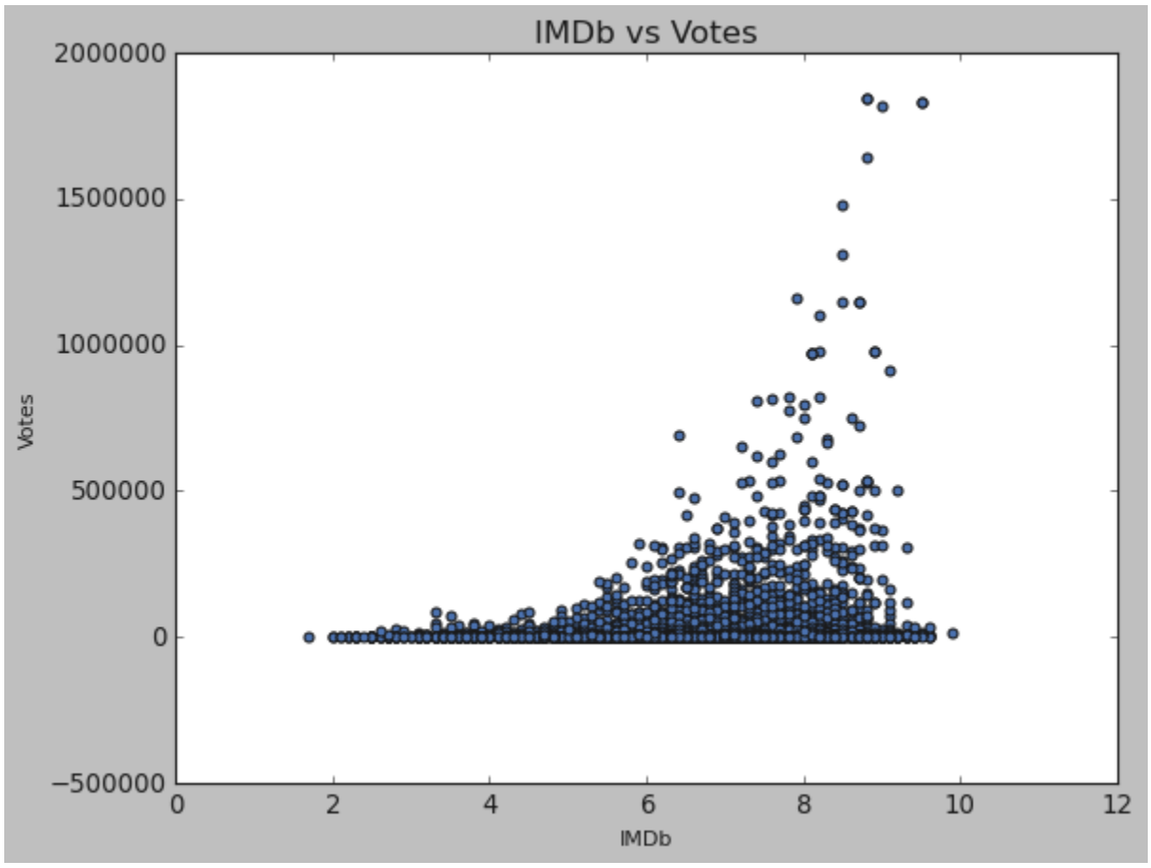
IQR (Interquartile Range) = Q3 - Q1 = 7.6 - 6.1 = 1.5 IMDb

All the content whose IMDb values are less than Q1-1.5*IQR (3.85) are outliers with very less IMDb and content whose IMDb are greater than Q3 + 1.5*IQR (9.85) are outliers with very high IMDb.

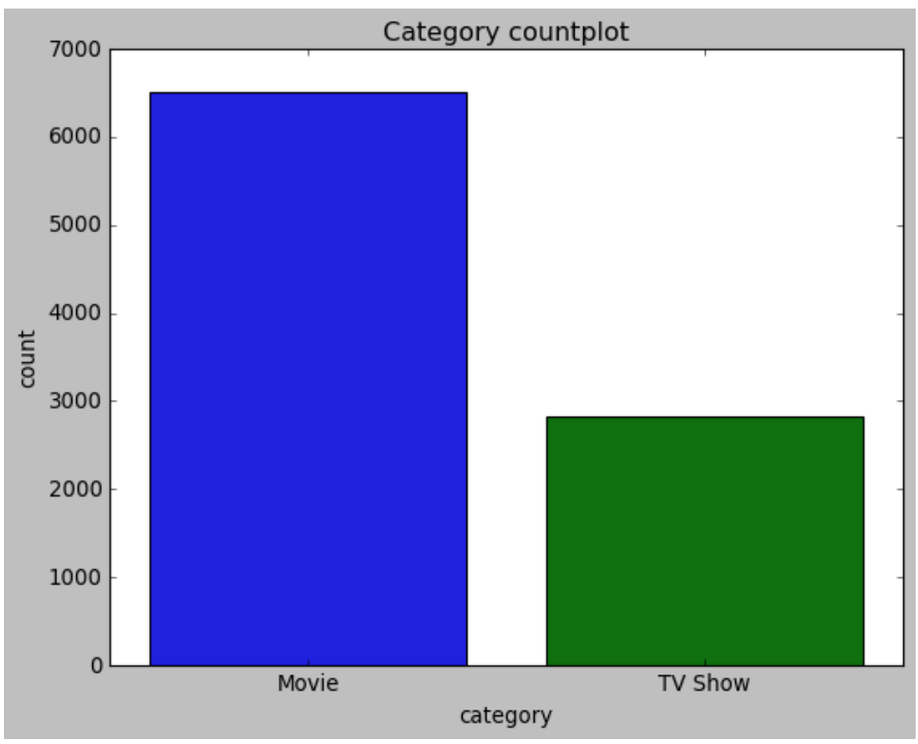
d. **Boxplot of the above five number summary of IMDb:**
As we can see that there is only one content whose IMDb is greater than 9.85 while there are comparatively many content whose IMDb is less than 3.85.



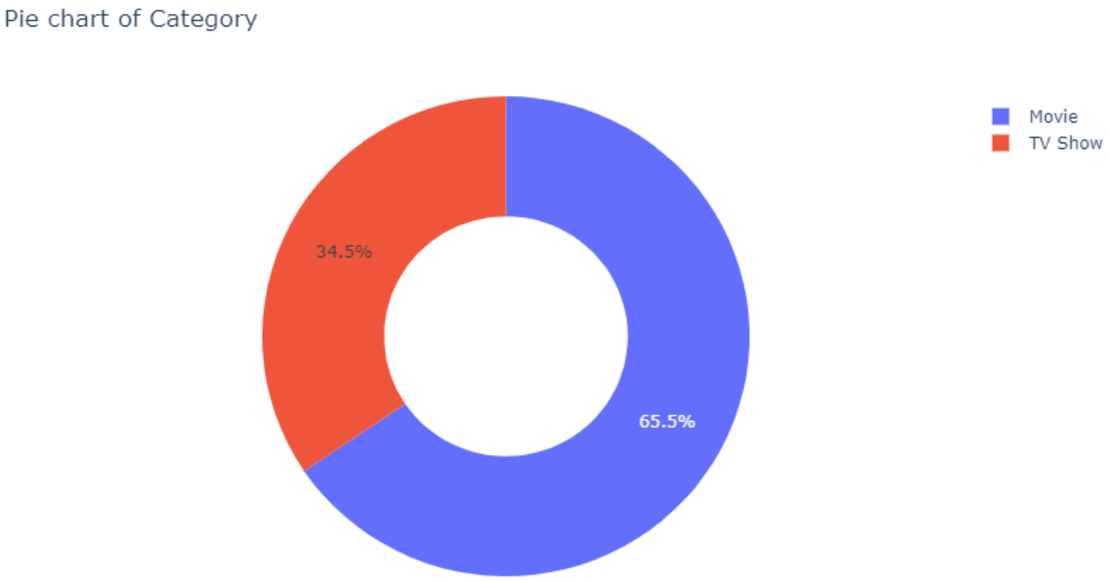
e. **Scatter plot of IMDb vs Votes:** From the below scatter plot, we can conclude that viewers tend to vote for high IMDb content.



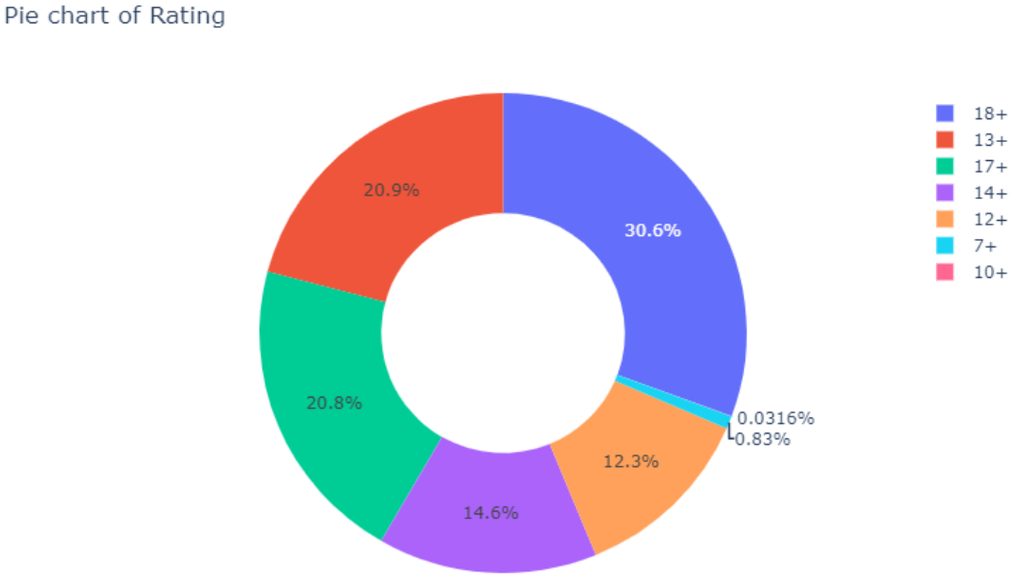
f. **Category Counts**: The data contains **6503 Movies** and **2825 TV Shows**. From the below countplot, we can conclude that there is more Movie content on Netflix, than TV Shows.



g. **Pie Charts**: It is the Pie chart representation of the above data with total number of votes for each category. **65.5%** of the total content is movies and **34.5%** of them are TV Shows.

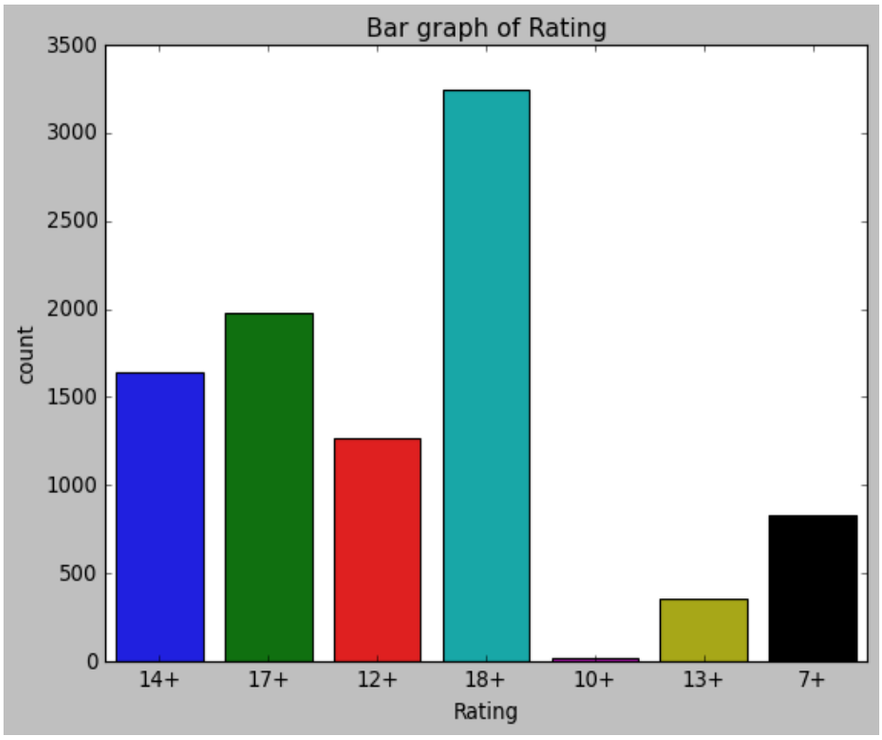


The below pie chart represents the percentage proportion of data by Rating (18+, 13+, 17+, 14+, 12+, 7+, 10+). From this visualization, we can see that majority of the content on Netflix is 18+ rated and very few (≅ 0.0316%) of total is 10+ rated. .

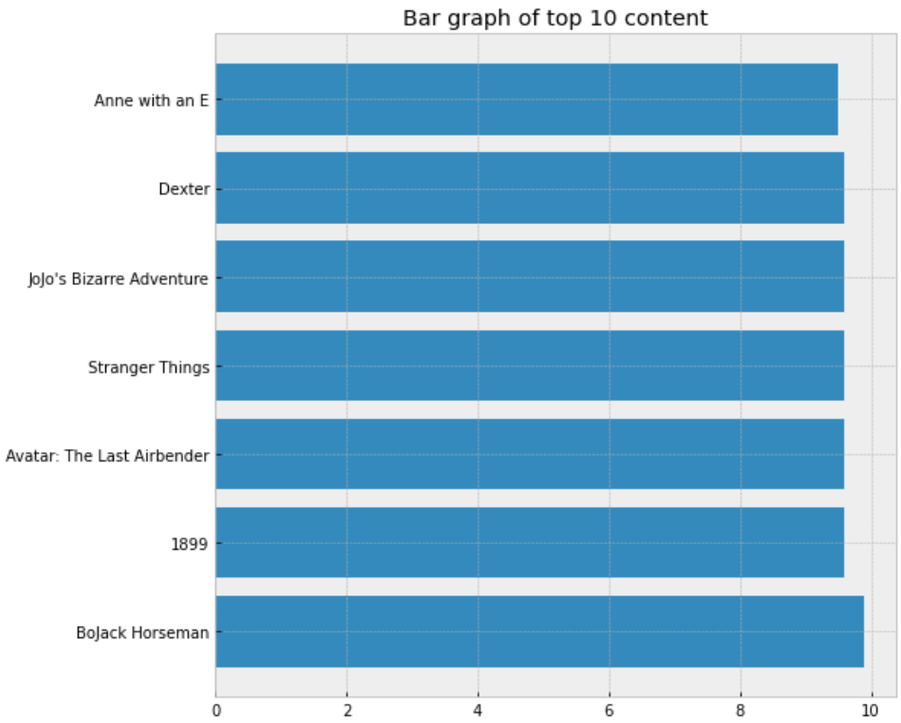


h. **Rating Counts**: The below countplot (bar graph) shows the countwise distribution of the content on the basis of the Rating.

Rating	Counts
18+	3244
17+	1979
14+	1640
12+	1268
7+	826
13+	354
10+	17



i. **Top 10 contents on Netflix:** The below bar graph shows the IMDb rating of the top 10 content on Netflix



4. **Data Filtering:**

    a. Filtering the data on the basis of the Rating given by the user.

    b. Filtering the data on the basis of Category and Country given by the user.

    c. Renaming the name of the column.

    d. Filtering the data on the basis of the Category and and Rating defined by the user.

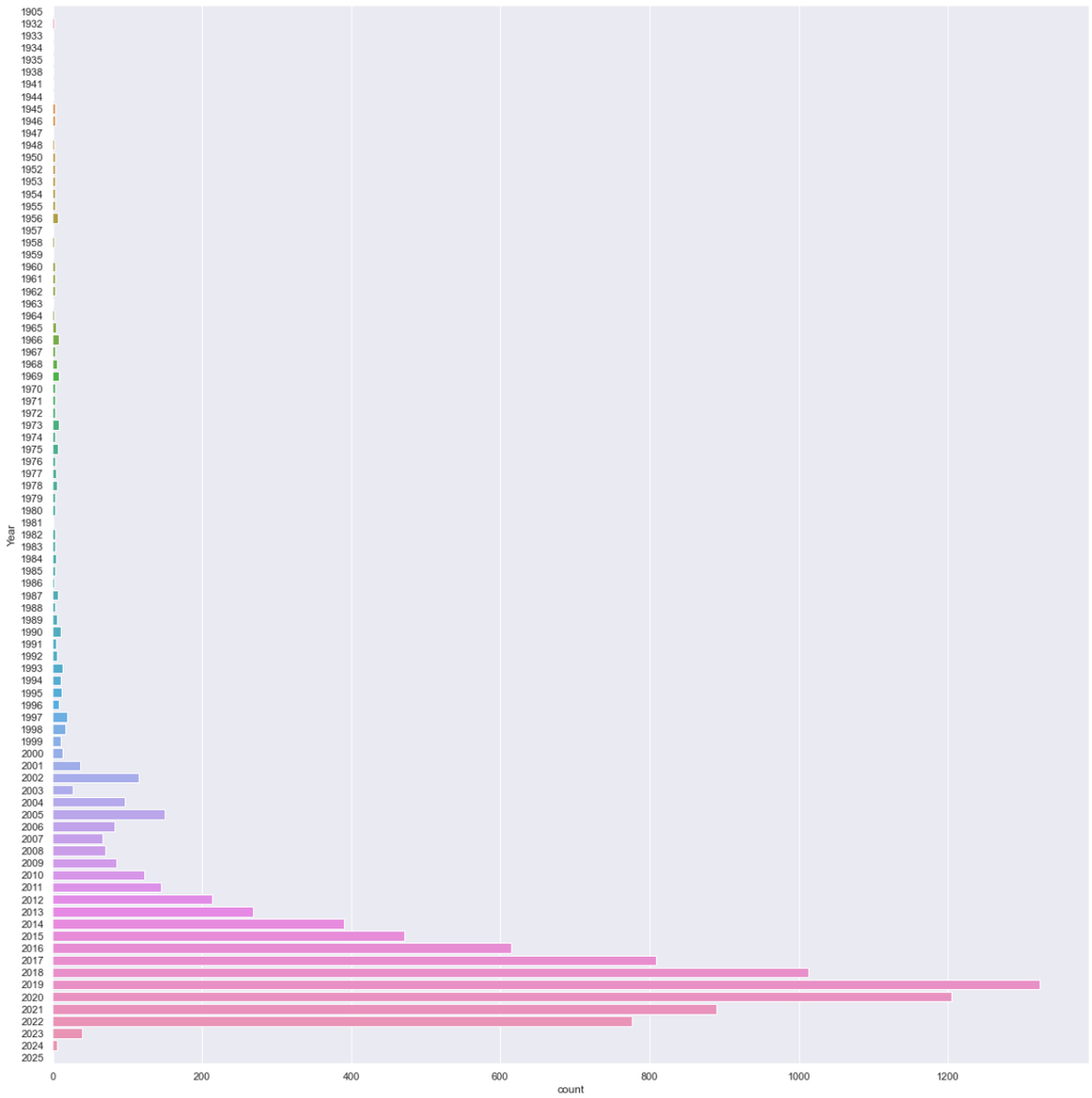
    e. Searching for the content from the title given by the user.

5. **Data Extraction**: Here we have extracted data from the Released_Date to a new column, namely ‘Date’ that contains date in formatted manner (yyyy-dd-mm). Furthermore, we have created new columns namely ‘Year’, ‘Month’, and ‘Day’ by splitting ‘Date’

    By doing so we can get a count of content released each year.

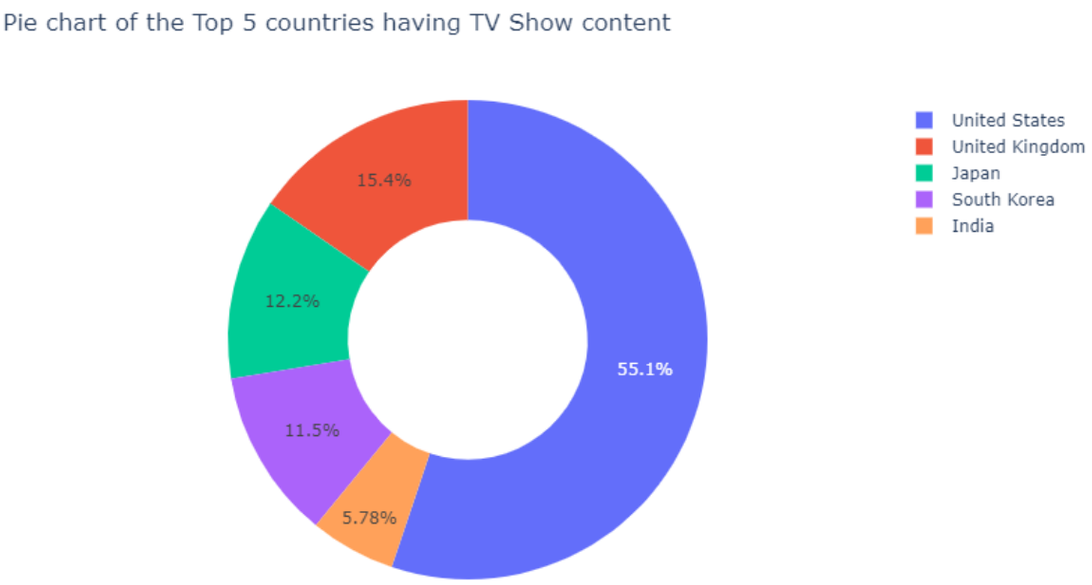
    By plotting the bar graph of the counts of released content against the Year, we can see that the majority of the content was released in the year 2019.

Year	**Counts**
2019	1323
2020	1205
2018	1013
2021	890
2017	809



Value Counts: Here, we have printed the leading five countries by number of TV Shows. As we can see, the majority of the TV Show content on Netflix are United States based.

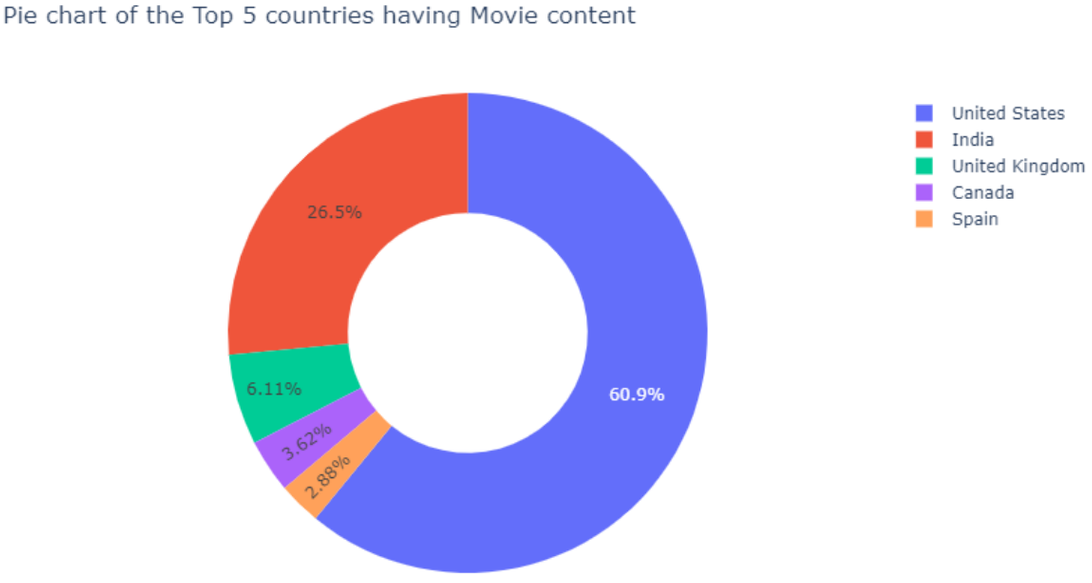
Country	**Number of TV Shows**
United States	763
United Kingdom	213
Japan	169
South Korea	159
India	80



Similarly, for the leading five countries by number of Movies.

Country	Number Movies
United States	2055
India	892
United Kingdom	206
Canada	122
Spain	97

Below pie chart is the visualization of above table showing that 60.9% of the total Movie content is United States based followed by India, contributing 26.5% of the total

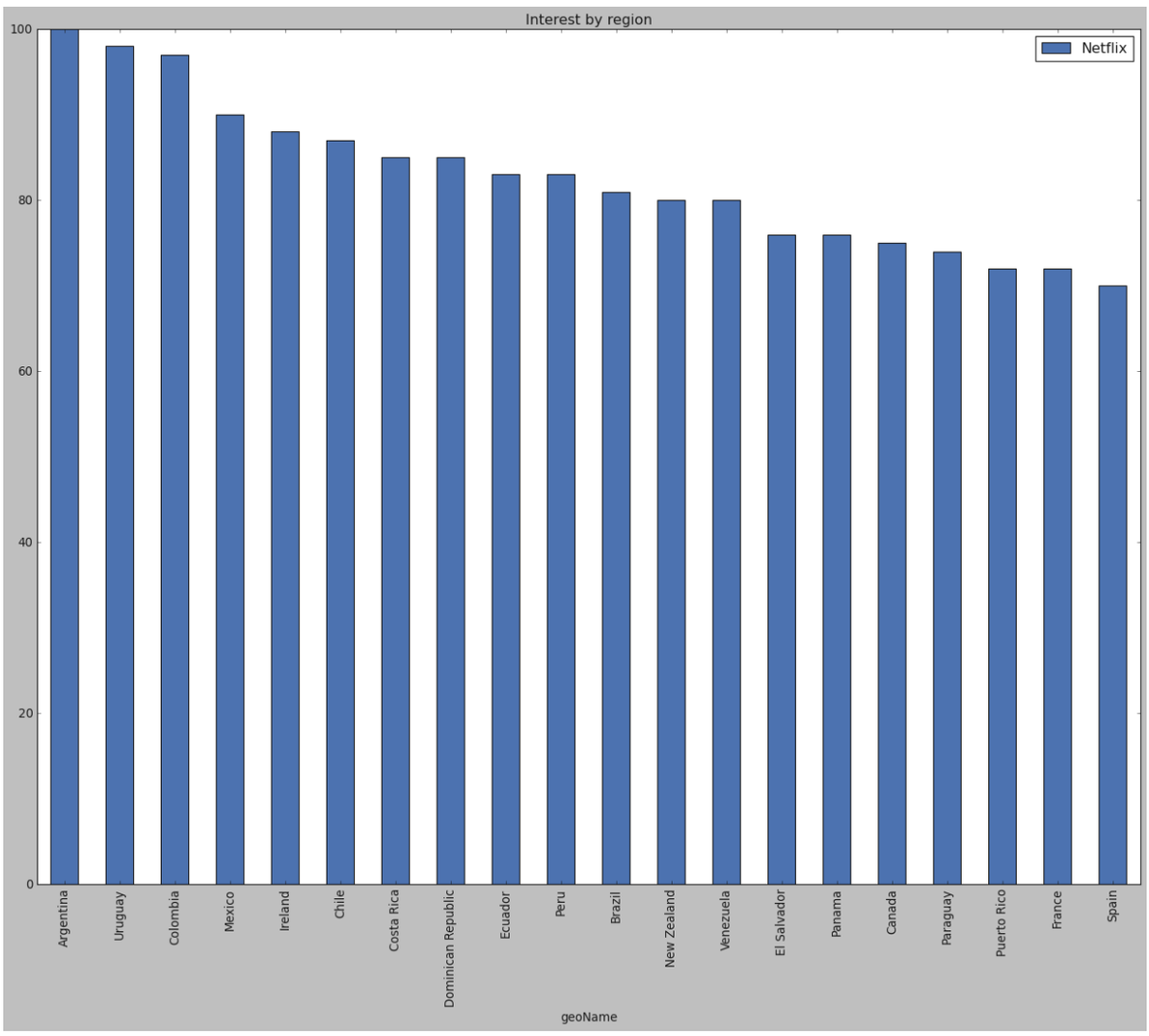


Here, we are not using the ‘Duration’ column in any visualization, therefore we are removing it using drop command.

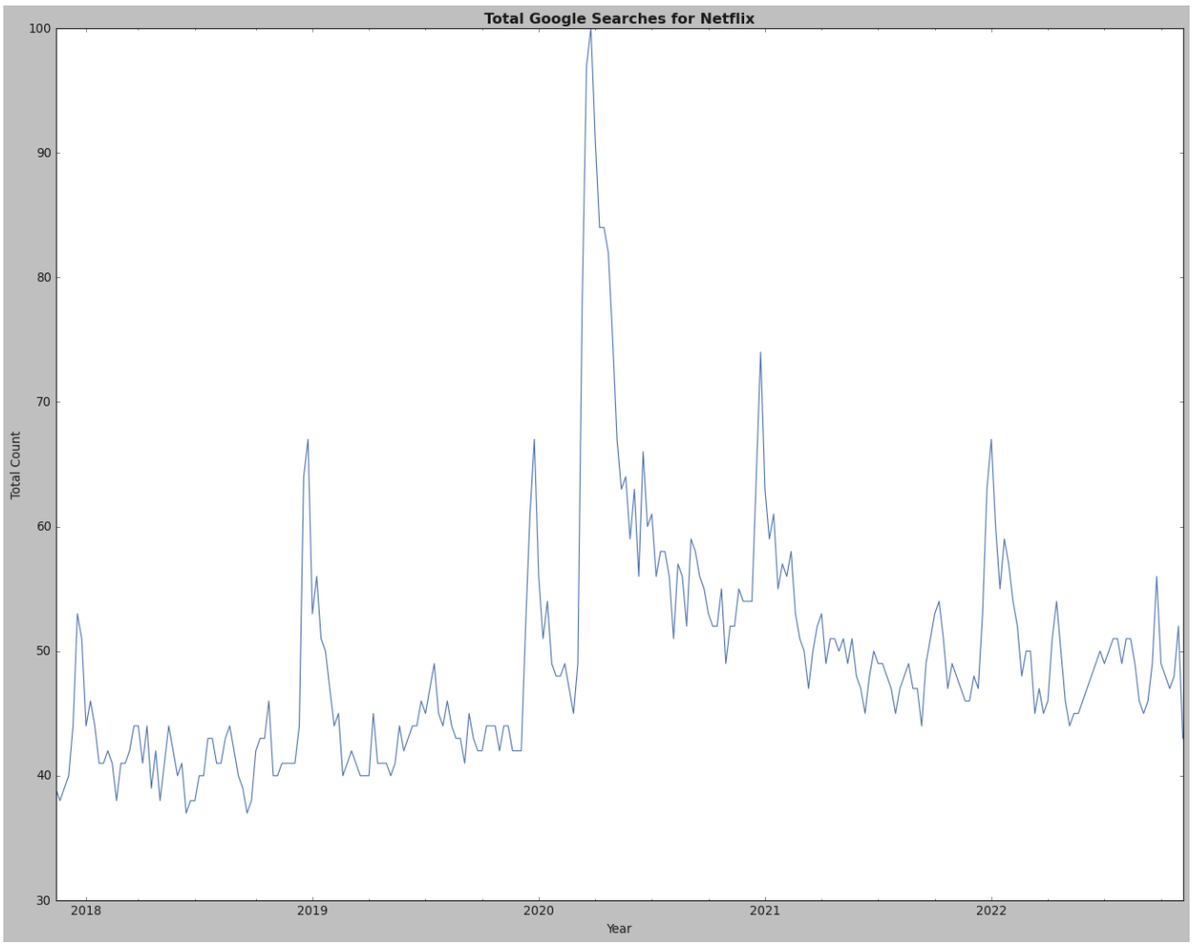
6. Netflix Search on Google:

Here, we are using an unofficial API for Google Trends, for getting the list of countries with decreasing relative search for the word ‘Netflix’ on Google.

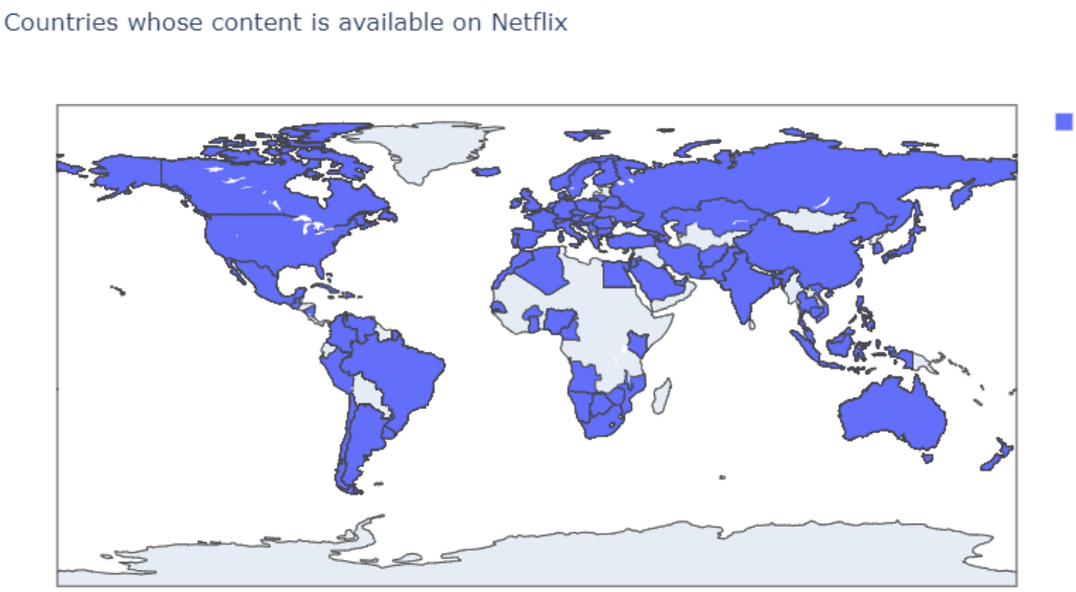
The values are calculated on a scale from 0 to 100, where 100 is the location with the most popularity as a fraction of total searches in that location, a value of 50 indicates a location which is half as popular. A value of 0 indicates a location where there is no search for the term.



Similarly, we can also plot the graph of the relative number of searches for the word ‘Netflix’ on Google, over time using interest_over_time() method of Pytrends.



Furthermore, we have also made a choropleth world map for getting a visual representation of the countries whose content is available on Netflix denoted by blue fill in it.



7. **Conclusion and further studies**: Here, we have implemented basic filtering operations and extracted meaningful insights about the data with their corresponding visualizations using different python libraries.

    For the problem of multiple countries for each content, we can filter the data by implementing one hot encoding.