Data Analysis and Visualization - Homework 5

Ronil Surve, Shubham Sharma

Net id: rs2136, ss4236

Topic: Netflix Movies & TV Shows Analysis using Tableau & Python

Dataset source: Kaggle Link

The original dataset contains 12 columns and 8807 rows. Each row represents a TV show or movie available on Netflix.

Details about the dataset:

Column names and description are:

- **show_id**: Unique ID for each show
- **type**: Type of the show (TV Show or Movie)
- **title**: Name of the show
- **director**: Name of the director(s) of the show
- **cast**: Names of the cast members
- **country**: Country where the show was produced
- **date_added**: Date when the show was added on Netflix
- **release_year**: Year of release
- **rating**: Rating of the show
- **duration**: Duration of the show (in minutes for movies or number of seasons for TV shows)
- **listed_in**: Genres of the show
- **description**: Brief summary of the show

Objective:

This project has four main objectives:

1. Create an interactive dashboard in Tableau & Python to understand the content available on Netflix in different countries. This will involve analyzing the dataset to identify trends in the types of shows available in different countries and presenting the insights using visualizations in Tableau and Python.
2. Identify similar content on Netflix by matching text-based features such as title, cast, and genre.
3. Conduct network analysis of actors and directors to find interesting insights.
4. Determine whether Netflix has a greater focus on TV shows than movies in recent years. This will involve analyzing the dataset to identify trends in the number of TV shows and movies released each year and comparing the trends over time to determine whether Netflix has shifted its focus towards TV shows.

Exploratory Data Analysis

```
In [5]: f,ax=plt.subplots(figsize=(14,7))
        sns.heatmap(df.isnull(),yticklabels=False,cbar=False,cmap='viridis',ax=ax)
        ax.set_title('Missing Values Visualization',fontsize=16,color='black')
        plt.show()
```

The yellow horizontal lines in a column means that there are some missing values in that column. So, we “director, cast, country, duration” that have some missing values
The presence of missing values in the dataset can have an impact on the analysis and modeling results. Therefore, it is important to handle missing values appropriately, either by removing them, imputing them with appropriate values, or using algorithms that can handle missing values.

In [4]:
```
# checking missing data in stack data
df_clean = df.copy()
total = df_clean.isnull().sum().sort_values(ascending = False)
percent = (df_clean.isnull().sum()/df_clean.isnull().count()*100).sort_values(ascending = False)
missing_df_clean = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
missing_df_clean
```

Out[4]:

	Total	Percent
director	2634	29.908028
country	831	9.435676
cast	825	9.367549
date_added	10	0.113546
rating	4	0.045418
duration	3	0.034064
show_id	0	0.000000
type	0	0.000000
title	0	0.000000
release_year	0	0.000000
listed_in	0	0.000000
description	0	0.000000

This information is useful for identifying which columns in the dataset have the most missing values and how much data is missing. This information can be used to decide how to handle missing values in the dataset, such as by imputing missing values or dropping rows or columns with a high percentage of missing values.

In [7]:
```
# number of unique values for each variable
df.nunique(axis=0)
```

Out[7]:
```
show_id         8807
type               2
title           8807
director        4528
cast            7692
country          748
date_added      1767
release_year      74
rating            17
duration         220
listed_in        514
description     8775
dtype: int64
```

In [8]:
```
# statistical summary of numeric variables
df.describe()
```

Out[8]:

	release_year
count	8807.000000
mean	2014.180198
std	8.819312
min	1925.000000
25%	2013.000000
50%	2017.000000
75%	2019.000000
max	2021.000000

The Year variable ranged from 1925 to 2021.

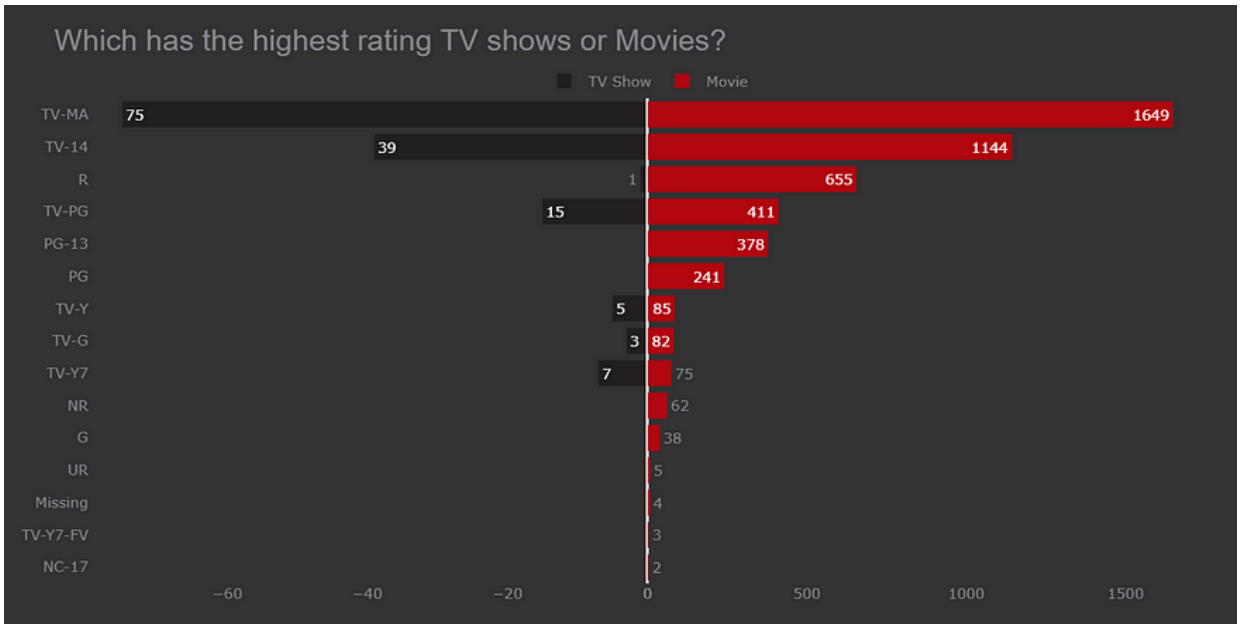
```
In [6]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64 
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: int64(1), object(11)
memory usage: 825.8+ KB
```

The dataset contains 12 columns, out of which are 5 integers, 6 strings, and 1 country.

Type of visualization intended to create:

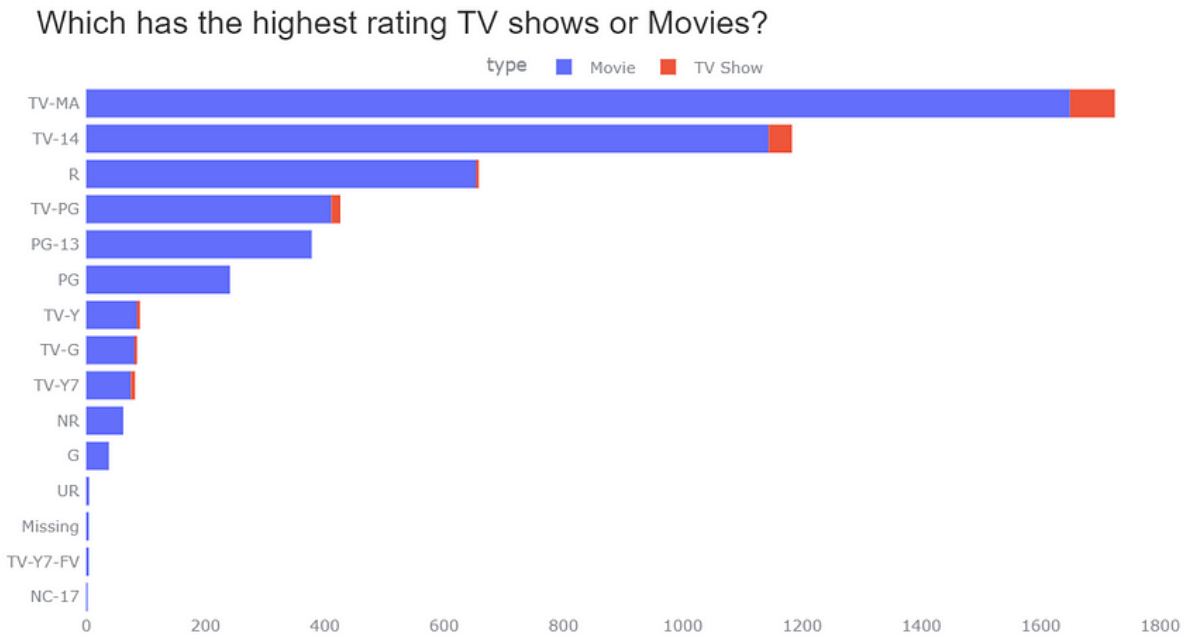
1.



https://pub.towardsai.net/tips-and-tricks-for-plotly-bar-chart-71261391c57b

We are planning to make a similar butterfly chart with the movies v/s TV shows

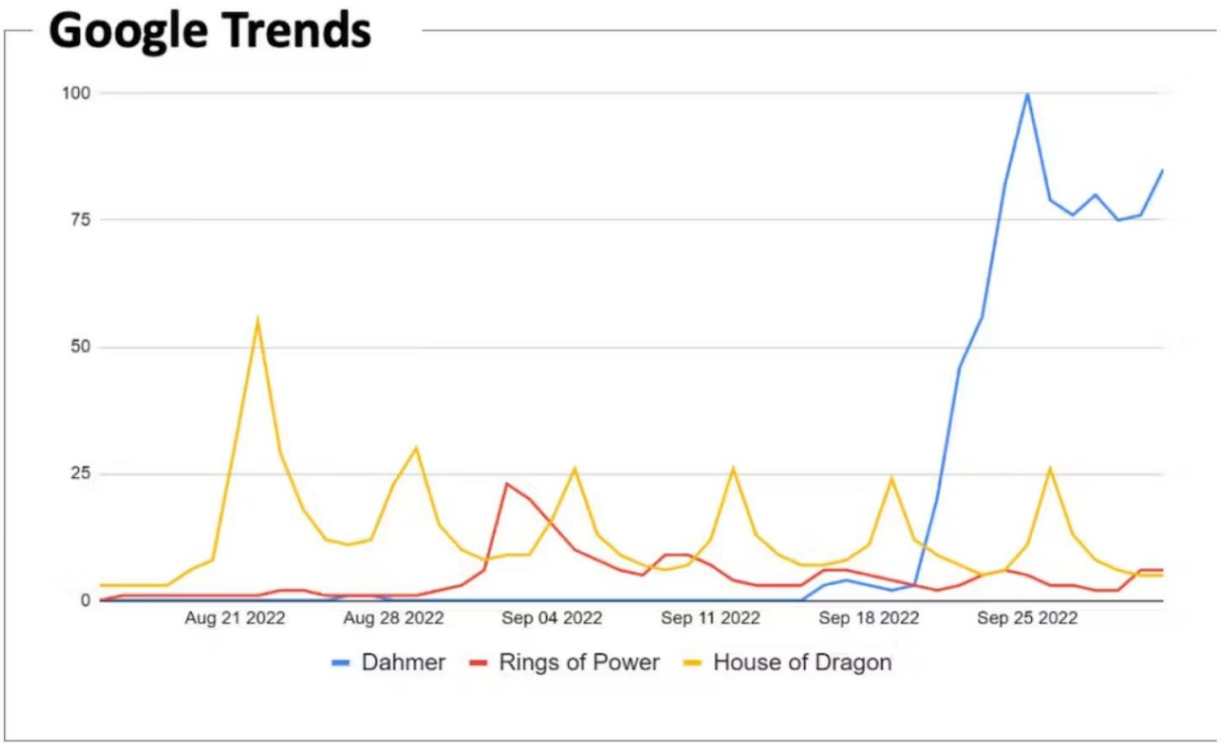
2.



https://pub.towardsai.net/tips-and-tricks-for-plotly-bar-chart-71261391c57b

Stacked bar chart that shows different preferences on movies and TV shows according to certain countries.

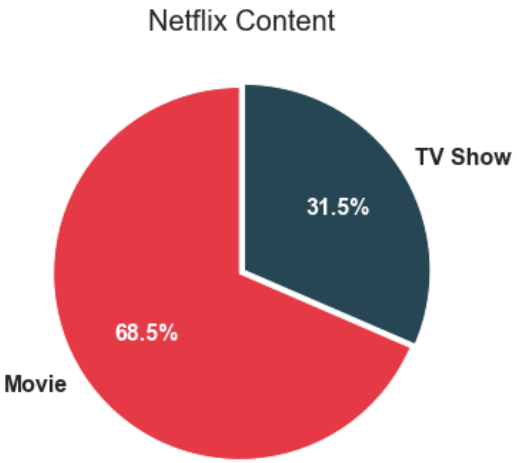
3.



https://hackernoon.com/netflix-business-strategy-in-2022

Movies and TV shows trend over time

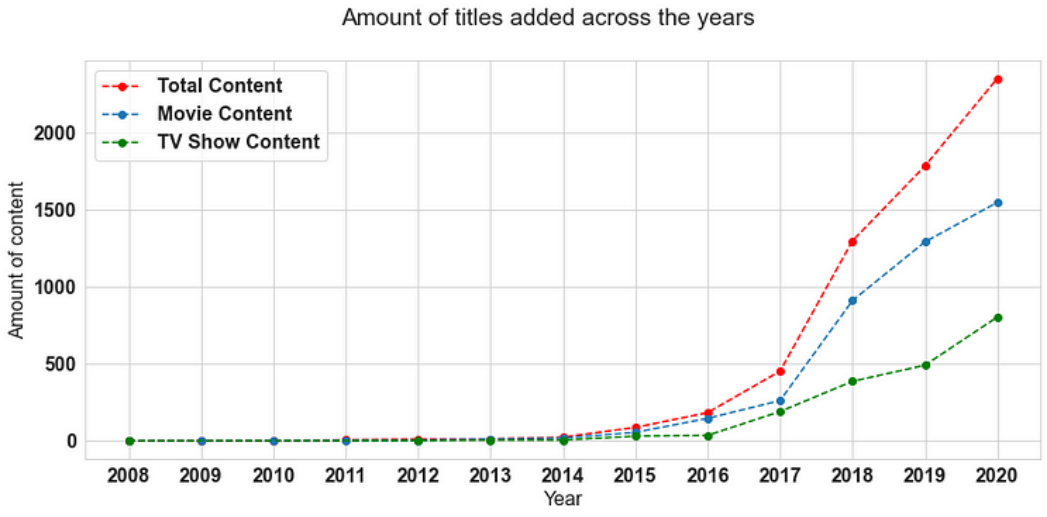
4.



https://jobymathew97.medium.com/netflix-movies-and-tv-shows-data-visualization-using-matplotlib-f1b4e91b5226

Movies v/s TV shows pie chart.

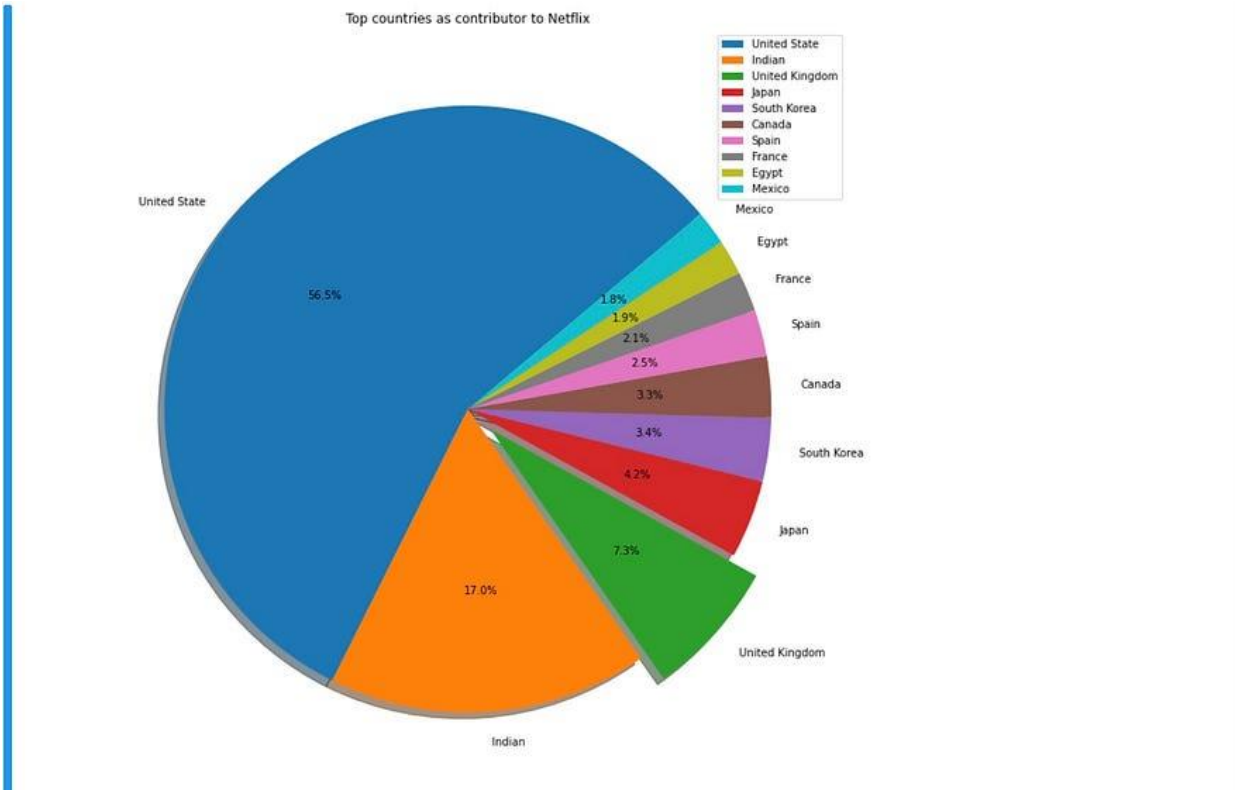
5.



https://jobymathew97.medium.com/netflix-movies-and-tv-shows-data-visualization-using-matplotlib-f1b4e91b5226

Increase in number of shows added in movies and TV shows and together yearly.

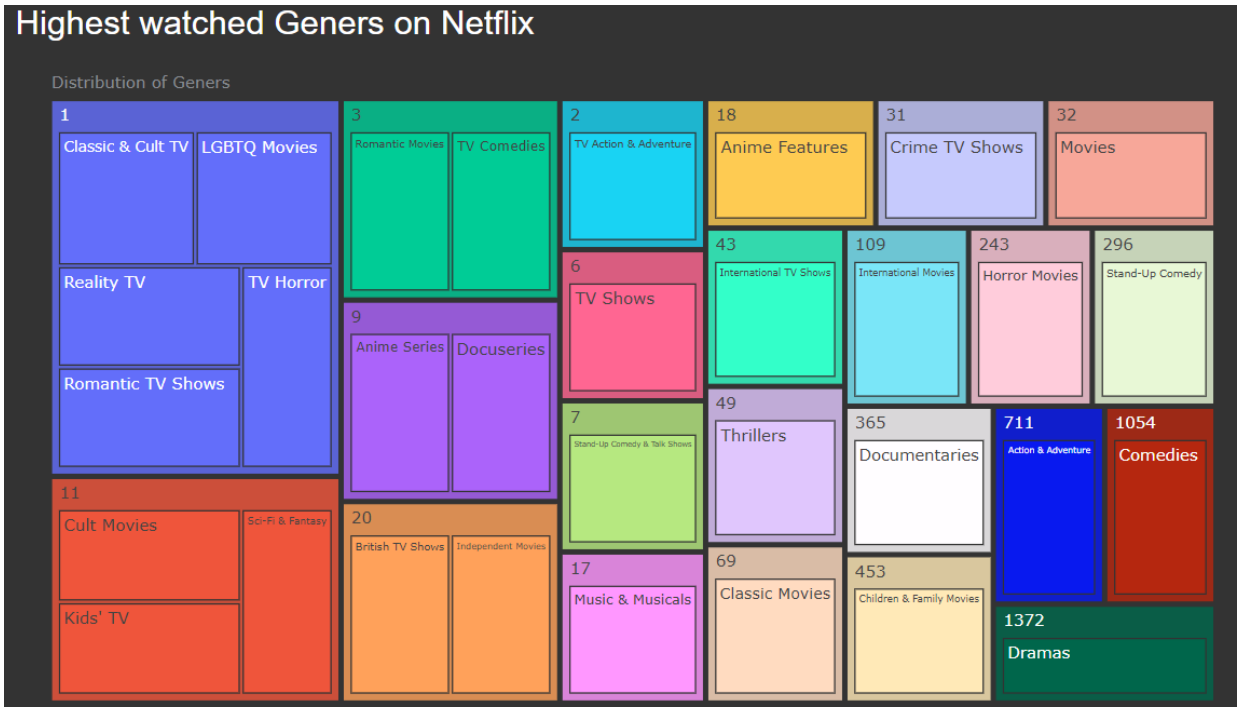
6.



https://wuraolaifeoluwa.medium.com/basic-data-wrangling-and-visualization-of-netflix-data-8b9609328f8c

Top 10 counties on Netflix.

7.



Performing EDA of Netflix Dataset with Plotly - Analytics Vidhya

Tree chart depicting most watched genres.

Challenges:

1. Data quality issues: The dataset contains missing values, inconsistent formatting, and other data quality issues that can make it challenging to create accurate and informative visualizations.
2. Geographical limitations: The dataset does not include information on the availability of shows and movies in specific countries or regions, which can limit the ability to create geographic visualizations or analyze regional trends.
3. Evolving content: The Netflix dataset is dynamic and constantly changing, with new shows and movies being added and removed from the platform regularly. This can make it challenging to create long-term trends or draw meaningful conclusions from the data.
4. Lack of context: The dataset lacks contextual information about the shows and movies, such as production budgets or marketing spend. This can make it challenging to understand the factors that drive the popularity of shows and movies on the platform.

Despite these challenges, creative data visualization techniques and data analysis strategies can be used to overcome these limitations and derive meaningful insights from the Netflix dataset.