Project Report on

Netflix Data Analysis

Movies & TV Shows

Submitted by:

Anshika Patel

Sudhir

In partial fulfillment of completion of the course

Advanced Diploma in IT, Networking and Cloud Computing.

Under Guidance of:

Year 2022-2023

Abstract

The "Netflix Movies & TV Shows Data Analysis" project delves into a comprehensive examination of Netflix's vast content library to extract meaningful insights and trends. With the streaming industry's escalating prominence, understanding user preferences, content distribution, and emerging patterns becomes crucial for content providers like Netflix. This project aims to unravel the dynamics of Netflix's content landscape through exploratory data analysis (EDA) techniques.

Key aspects explored include the distribution of movies and TV shows across genres and release years, user ratings, and content durations. The project investigates the evolving landscape of genres over time, discerning popular trends and patterns. Furthermore, it delves into regional variations in content preferences, accounting for linguistic and cultural nuances.

Acknowledgement

At this juncture of our journey, we wish to express our heartfelt gratitude to all those who have contributed to the creation and success of **"Netflix Data Analysis".** This project has been a labor of passion and dedication, and it would not have been possible without the unwavering support and guidance we have received.

First and foremost, we offer our thanks to the boundless creativity and inspiration that flows from the universe. We are grateful for the opportunity to embark on this venture.

We extend our sincerest appreciation to our mentors, **Mrs. Mala Mishra & Ms. Ankita Shukla**, whose wisdom and guidance have been instrumental in shaping the vision of **"Netflix Data Analysis".** Your support at every crucial turn has illuminated our path and fueled our determination to create a meaningful platform.

To our dedicated team of developers, designers, and content creators, we extend our deepest gratitude. Your tireless efforts, innovation, and creativity have breathed life into **"Netflix Data Analysis".** It is your collective dedication that has made this project a reality.

Our appreciation also goes to our colleagues and friends who provided invaluable insights and feedback during the development process. Your input has been instrumental in refining our ideas and enhancing the user experience.

We acknowledge the contributions of the broader IT community, whose open-source ethos has been a wellspring of knowledge and inspiration. The collaborative spirit of this community has been a guiding light.

Last but not least, we owe a debt of gratitude to our families and friends who have stood by us throughout this journey. Your unwavering support, encouragement, and belief in our vision have been our constant motivation.

Project Requirements

Project Name	**Netflix Data Analysis**
Languages Used	**Python & data wrangling and data Visualisation tools**
Editor	**Jupyter Notebook, Google Colab**
Web Browser	**Google Chrome, Microsoft Edge**

Team Composition and Workload Division

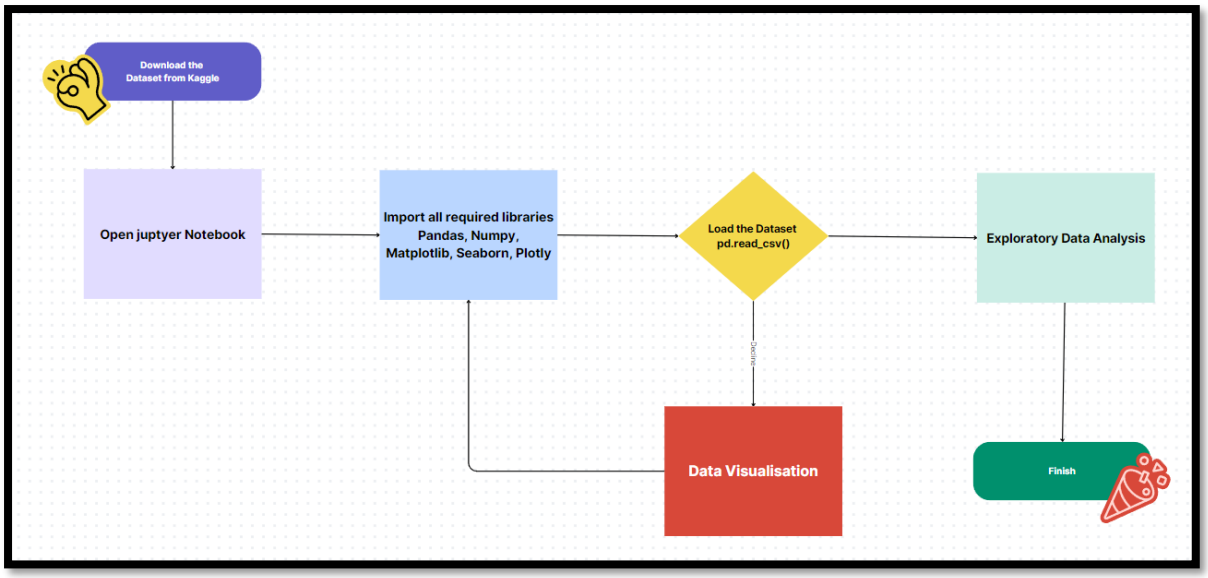
Anshika Patel	Data Analysis, Synopsis
Sudhir	Data Analysis

Table of Contents

1. Introduction to Problem

The entertainment industry has undergone a seismic transformation with the rise of streaming platforms, reshaping the way audiences consume content. As viewers increasingly turn to on-demand services, understanding the intricacies of content distribution, user preferences, and emerging trends becomes paramount for streaming giants like Netflix. This data analysis project seeks to address the evolving landscape of Netflix's content library, aiming to extract actionable insights to enhance the platform's strategic decision-making.

2. E-R Model



3. Requirements

3.1 Technology Stack

Python: High-level programming language used for server-side scripting.

Jupyter Notebook: Jupyter Notebook is an open-source web application that allows you to create and share documents containing live code, equations, visualizations, and narrative text, providing an interactive and collaborative environment for data science and analysis.

3.2 Hardware

Laptop/ Computer

3.3 Software

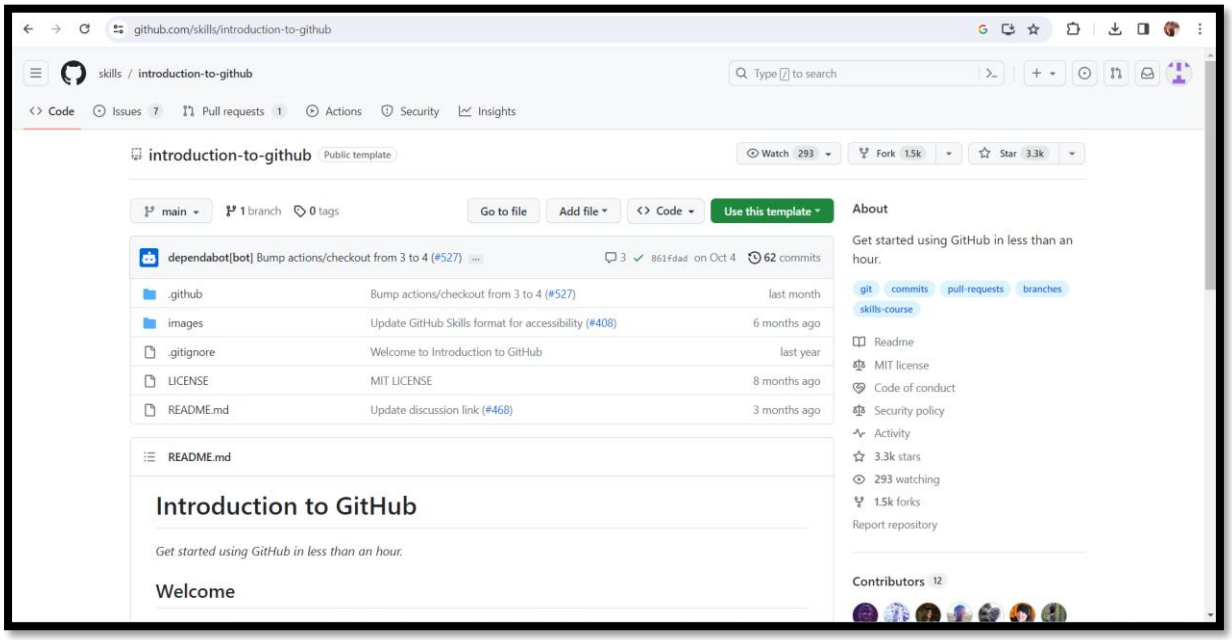
Operating System (OS)

Version Control System

Text Editors and Integrated Development Environments (IDEs)

3.4 Deployment Environment

Github



4. Introduction

The entertainment industry has undergone a profound transformation with the advent of streaming platforms, and Netflix stands at the forefront of this revolution. Originally founded in 1997 as a DVD rental-by-mail service, Netflix swiftly adapted to the digital era, evolving into a global streaming giant. Today, it is a household name, offering an extensive library of movies, TV shows, documentaries, and original content accessible to subscribers worldwide.

4.1 Background

The platform's user-friendly interface and diverse content offerings have contributed to a cultural shift where viewers have the autonomy to choose what, when, and how they watch. Streaming platforms, led by Netflix, have become the new norm, reshaping the way audiences engage with entertainment.

4.2 Objective

The primary goals of our data analysis project are to:

- Understand the Distribution of Content.
- Identify Trends.

- Explore User Preferences.

5. Data Collection

5.1 Data Source

The Netflix dataset used in this analysis was sourced from **Kaggle**. This dataset captures a comprehensive snapshot of Netflix's movies and TV shows, encompassing a range of variables that form the basis of our exploratory analysis.

Dataset Structure:

The dataset consists of [7787] rows and [12] columns.

display the rows or column

```
In [11]: print('Number of Rows',df.shape[0])
         print('Number of Columns',df.shape[1])

         Number of Rows 7787
         Number of Columns 12
```

Key variables include [list the essential variables, such as 'Title,' 'Directors,' 'Release Year,' 'Ratings,' etc.].

display all column Name

```
In [20]: df.columns

Out[20]: Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
                'release_year', 'rating', 'duration', 'listed_in', 'description'],
               dtype='object')
```

5.2 Data Cleaning

Steps Taken:

Handling Missing Values:

Identified and assessed missing values across variables.

Check Missing Values in the Dataset

```
In [17]: df.isnull().sum()

Out[17]: show_id          0
         type             0
         title            0
         director      2389
         cast           718
         country        507
         date_added      10
         release_year     0
         rating           7
         duration         0
         listed_in        0
         description      0
         dtype: int64
```

Duplicate Removal:

Checked for and removed duplicate entries to ensure data integrity.

Dealing with the dataset

```
In [22]: # replace the null value with mode

df['country'] = df['country'].fillna(df['country'].mode()[0])
df['cast'].replace(np.nan, 'No Data', inplace = True)
df['director'].replace(np.nan, 'No Data', inplace = True)

# drop columns

df.dropna(inplace = True)

# drop duplicates

df.drop_duplicates(inplace = True)
```

6. Exploratory Data Analysis (EDA)

The exploratory data analysis phase has provided foundational insights into the Netflix dataset. The distribution of content across genres, user ratings, and content durations form the basis for more in-depth analyses and strategic recommendations in subsequent sections of the project.

6.1 Overview

The dataset under consideration comprises [7787] records and [12] features, offering a comprehensive view of Netflix's movies and TV shows. Initial statistical analysis reveals [brief summary of key statistics, such as mean, median, and standard deviation], providing a foundation for further exploration.

Getting Information About our Dataset and memory Requirement)

```
In [15]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7787 entries, 0 to 7786
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       7787 non-null   object
 1   type          7787 non-null   object
 2   title         7787 non-null   object
 3   director      5398 non-null   object
 4   cast          7069 non-null   object
 5   country       7280 non-null   object
 6   date_added    7777 non-null   object
 7   release_year  7787 non-null   int64 
 8   rating        7780 non-null   object
 9   duration      7787 non-null   object
 10  listed_in     7787 non-null   object
 11  description   7787 non-null   object
dtypes: int64(1), object(11)
memory usage: 730.2+ KB
```

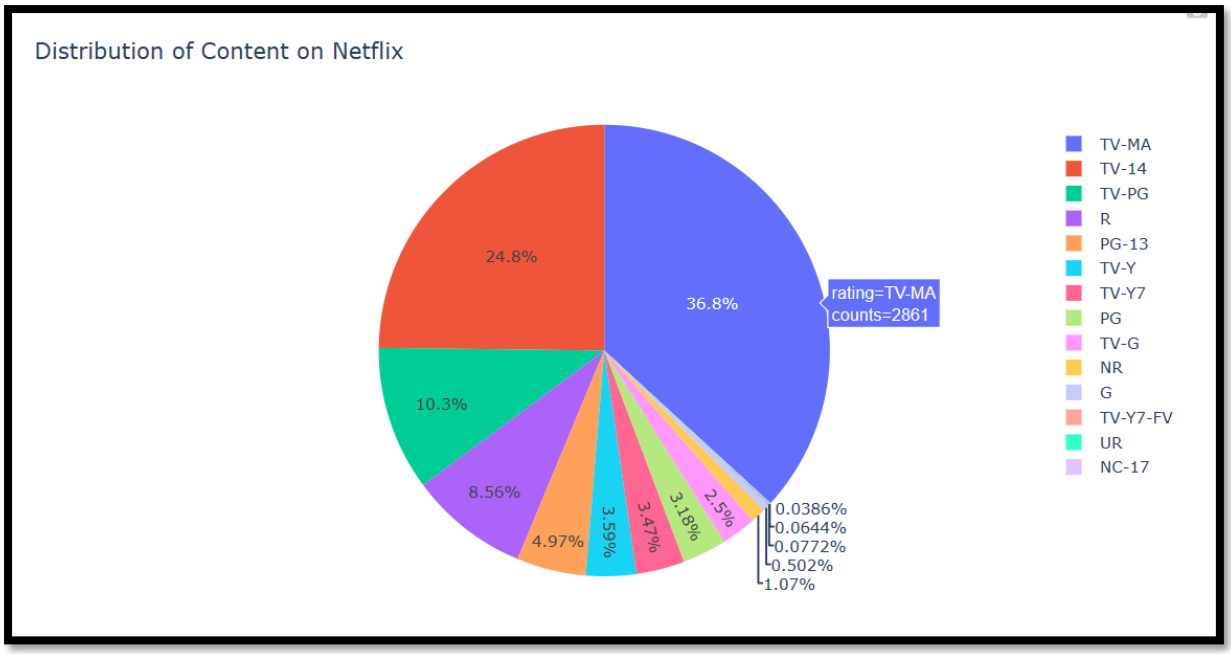
```
In [13]: df.describe()

Out[13]:
```

	release_year
count	7787.000000
mean	2013.932580
std	8.757395
min	1925.000000
25%	2013.000000
50%	2017.000000
75%	2018.000000
max	2021.000000

6.2 Explore the distribution of movies and TV shows on Netflix:

Visualizations, including pie charts, bar graphs, or heatmaps, highlight the distribution of content across genres, allowing for a quick assessment of genre popularity.

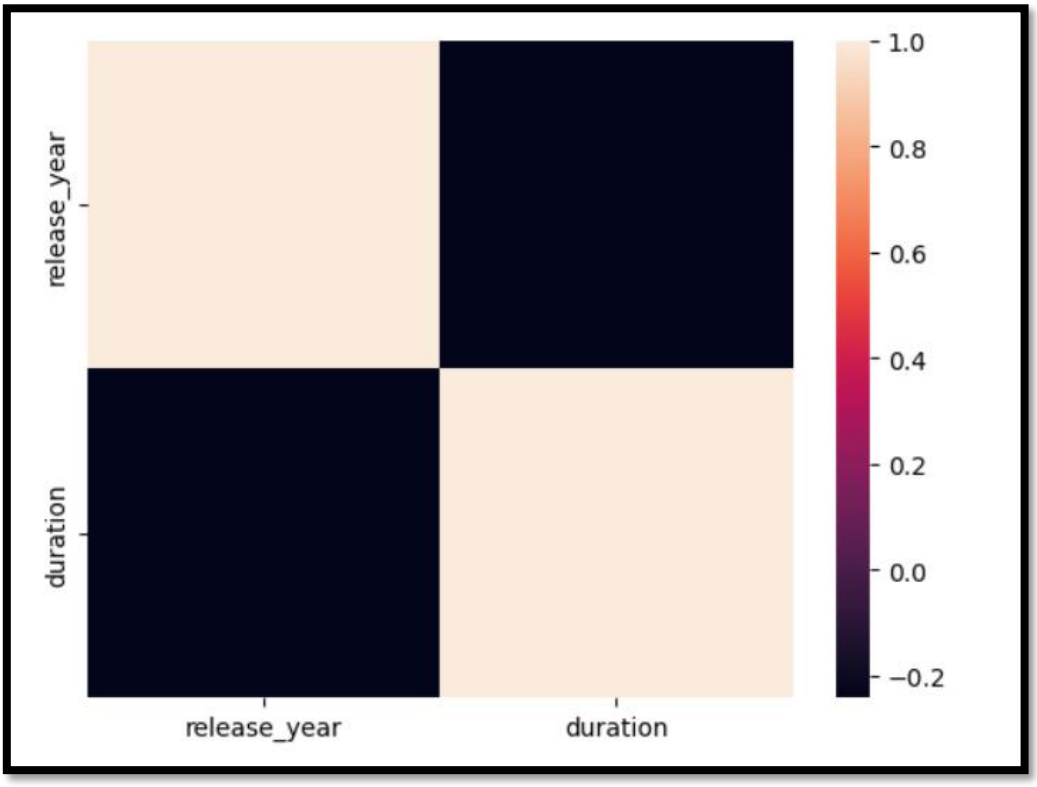


6.3 User Ratings

Analyze user ratings and reviews:

Identify the highest and lowest-rated content to understand user preferences and content quality.

Correlate user ratings with other variables (e.g., release_year, duration) to uncover patterns and potential influencers of user satisfaction.



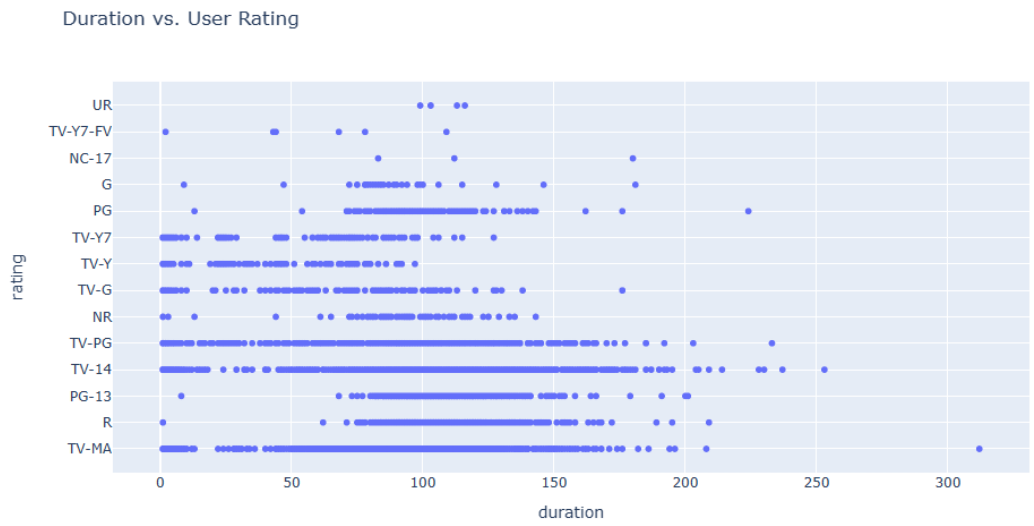
6.4 Duration Analysis

Duration Analysis

In [56]:
```
# Create a DataFrame for duration analysis
duration_df = df[['duration', 'rating']].dropna()

# Convert 'duration' to numeric (assuming it's in a format like 'X min')
duration_df['duration'] = pd.to_numeric(duration_df['duration'].str.extract('(\d+)')[0], errors='coerce')


# Scatter plot for duration vs. user rating
scatter_plot = px.scatter(duration_df, x='duration', y='rating', title='Duration vs. User Rating')
scatter_plot.show()
```



7. Overview

The data analysis project aims to investigate and derive meaningful insights from a specific dataset. It involves collecting, cleaning, and processing raw data to uncover patterns, trends, and correlations. Using statistical methods and visualization tools, the project seeks to provide a comprehensive understanding of the data, enabling informed decision-making. The analysis may involve exploring relationships between variables, identifying outliers, and creating predictive models. Throughout the project, a systematic approach is followed, including hypothesis testing and validation of results. The ultimate goal is to offer actionable recommendations or conclusions based on the data findings. The project typically employs programming languages such as Python along with tools like Jupyter Notebooks, to facilitate a transparent and reproducible analytical workflow. Overall, the data analysis project serves to extract valuable insights, enhance understanding, and support evidence-based decision-making in a given domain.

Project Module

1. Import the required libraries.

2. Load/ Read the Dataset

3. Prepare EDA

4. Do Visualizations

5. Analysing Top Actor / Directors/ Country on Netflix

6. Prepare Heatmap

7. Prepare Profile Report

8. Sample Screenshots

Exploratory Data Analysis[EDA]

Display the Top 5 rows

In [5]:
```
df.head()
```

Out[5]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	TV Show	3%	NaN	João Miguel, Bianca Comparato, Michel Gomes, R...	Brazil	August 14, 2020	2020	TV-MA	4 Seasons	International TV Shows, TV Dramas, TV Sci-Fi &...	In a future where the elite inhabit an island ...
1	s2	Movie	07:19	Jorge Michel Grau	Demián Bichir, Héctor Bonilla, Oscar Serrano, ...	Mexico	December 23, 2016	2016	TV-MA	93 min	Dramas, International Movies	After a devastating earthquake hits Mexico Cit...
2	s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella Chung, Henley Hii, Lawrence ...	Singapore	December 20, 2018	2011	R	78 min	Horror Movies, International Movies	When an army recruit is found dead, his fellow...
3	s4	Movie	9	Shane Acker	Elijah Wood, John C. Reilly, Jennifer Connelly...	United States	November 16, 2017	2009	PG-13	80 min	Action & Adventure, Independent Movies, Sci-Fi...	In a postapocalyptic world, rag-doll robots hi...
4	s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin Spacey, Kate Bosworth, Aar...	United States	January 1, 2020	2008	PG-13	123 min	Dramas	A brilliant group of students become card-coun...

Display Last 5 rows

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

Out[6]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description

Data Visulisation

Taking the count of ratings available

In [16]:
```
# which content is avialable in the most amount on the netflix
a = df.groupby(['rating']).size().reset_index(name='counts')
print(a)
```

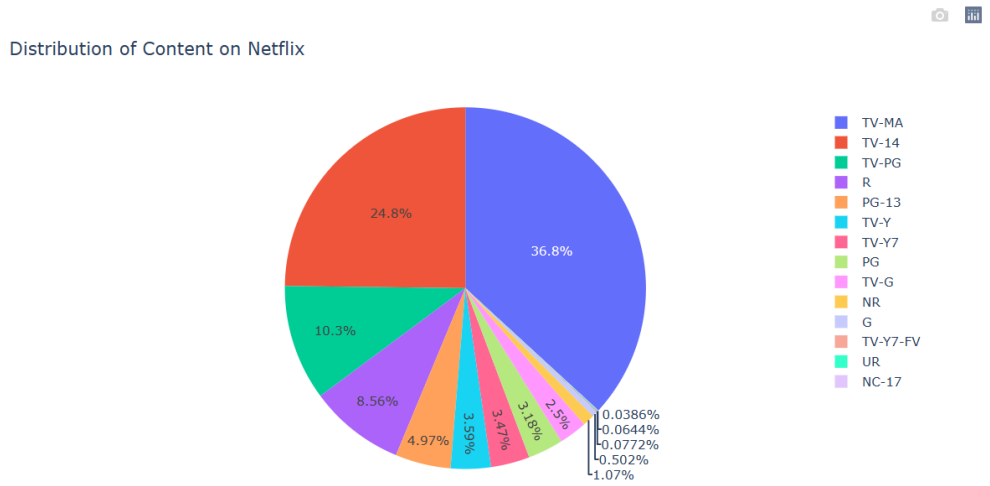
```
      rating  counts
0          G      39
1      NC-17       3
2         NR      83
3         PG     247
4      PG-13     386
5          R     665
6      TV-14    1928
7       TV-G     194
8      TV-MA    2861
9      TV-PG     804
10      TV-Y     279
11      TV-Y7     270
12  TV-Y7-FV       6
13        UR       5
```

Creating a pie chart based on content ratings

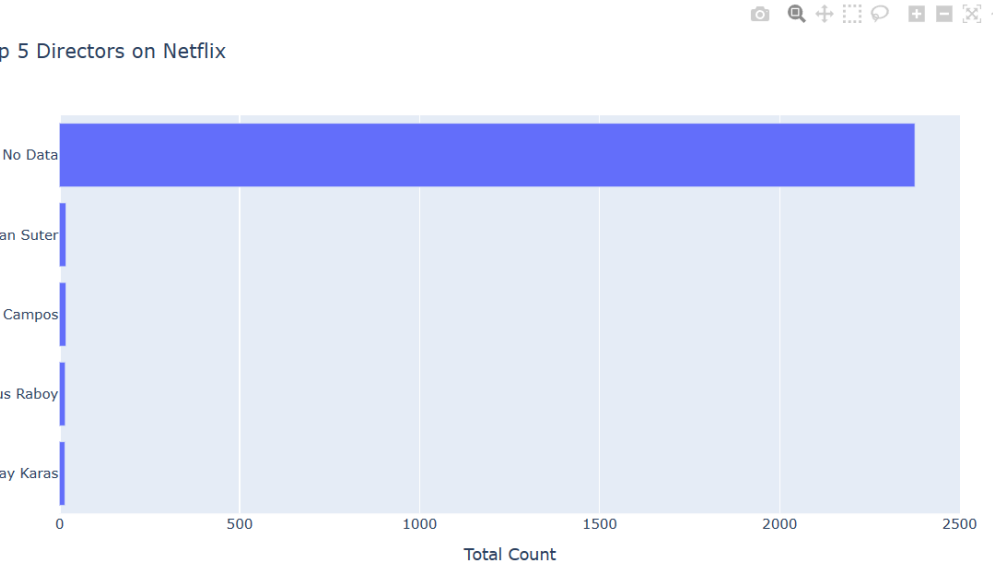
In [17]:
```
# create a pie chart
piechart = px.pie(a, values='counts', names='rating', title="Distribution of Content on Netflix")
piechart.show()
```

Creating a pie chart based on content ratings

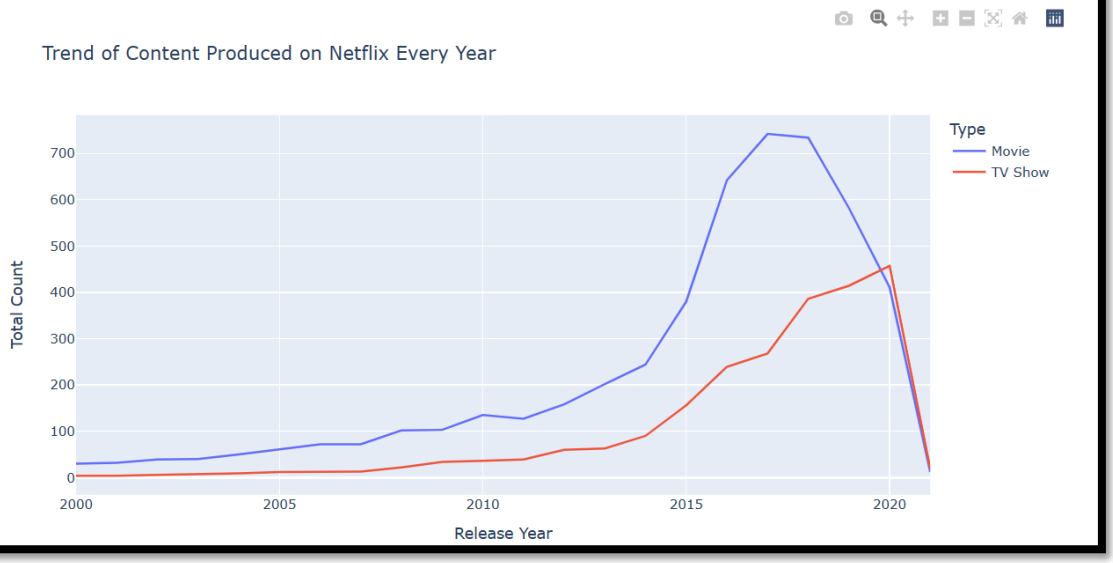
```
In [17]: # create a pie chart
         piechart = px.pie(a, values='counts', names='rating', title="Distribution of Content on Netflix")
         piechart.show()
```




```
In [31]: # horizontal bar chart
         top_5_dir = top_5_dir.sort_values(by=['Total Count'])
         barChart = px.bar(top_5_dir, x='Total Count', y='Directors', title='Top 5 Directors on Netflix')
         barChart.show()
```



```
In [35]: df2 = df2[df2['Release Year']>=2000]
         graph = px.line(df2, x='Release Year', y='Total Count', color ='Type', title='Trend of Content Produced on Netflix Every Year')
         graph.show()
```



9. Source Code

![new.jpg](attachment:new.jpg)

 # Netflix Data Analysis Movies & TV Shows

 # About this Dataset:

Netflix is one of the most popular media and video streaming platforms. They have over 8000 movies or tv shows available on their platform, as of mid-2021, they have over 200M Subscribers globally. This tabular dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration, etc.

 # Intresting Task Ideas:

Understanding what content is available on Netflix.

Analysis of what content on Netflix based on Rating.

Analysis of Top 5 Actors / Directors.

Analyzing the content produced on Netflix based on years.

Import required Libraries

 # import required library

import pandas as pd # used for data prepration

import numpy as np # mathematical operations or linear algebra

```
import matplotlib.pyplot as plt # used for data visulisation
import seaborn as sns # used for data visulisation
%matplotlib inline
import plotly.express as px # used for data visualisation
import datetime as dt
from textblob import TextBlob  # used for sentiment analysis

# import warnings
import warnings
warnings.filterwarnings("ignore")

### Load the Dataset

df = pd.read_csv('netflix_data.csv')

# read the dataset
df

# Exploratory Data Analysis[EDA]

### Display the Top 5 rows

df.head()

### Display Last 5 rows

df.tail()

### display the rows or column

print('Number of Rows',df.shape[0])
print('Number of Columns',df.shape[1])

### Display the some Statistical information about our dataset
```

```
df.describe()

### Getting Information About our Dataset (Total Number Rows, Total number of columns, datatypes of each column and memory Requirement)

df.info()

### Check Missing Values in the Dataset

df.isnull().sum()

### display all column Name

df.columns

### Dealing with the dataset

# replace the null value with mode

df['country'] = df['country'].fillna(df['country'].mode()[0])
df['cast'].replace(np.nan, 'No Data', inplace = True)
df['director'].replace(np.nan, 'No Data', inplace = True)

# drop columns

df.dropna(inplace = True)

# drop duplicates

df.drop_duplicates(inplace = True)



# check again null value
df.isnull().sum()
```

```
### Correct date format

# corect date format
df["date_added"] = pd.to_datetime(df['date_added'])

df['month_added'] = df['date_added'].dt.month
df['month_name_added'] = df['date_added'].dt.month_name()
df['year_added'] = df['date_added'].dt.year

# check the date format
df.head(3)

### Check Duplicates Values in our Dataset

# check duplicates values
df.duplicated().sum()

# Data Visulisation

## Taking the count of ratings available

# which content is avialable in the most amount on the netflix
a = df.groupby(['rating']).size().reset_index(name='counts')
print(a)

### Creating a pie chart based on content ratings

# create a pie chart
piechart = px.pie(a, values='counts', names='rating', title="Distribution of Content on Netflix")
piechart.show()

# Correlate user ratings with duration
```

```
# Convert 'Duration' to numeric (assuming it's in a format like 'X min')
df['duration'] = pd.to_numeric(df['duration'].str.extract('(\d+)')[0], errors='coerce')

# Correlation matrix
correlation_matrix = df[['rating', 'duration']].corr()

# Pairplot for visualizing correlations
sns.pairplot(df[['rating', 'duration']], kind='scatter')
plt.show()

# Heatmap for correlation matrix
correlation_matrix = df[['rating', 'duration']].corr()

sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap between User Rating and Duration')
plt.show()

sns.heatmap(df.corr())
plt.show()

# Duration Analysis

# Create a DataFrame for duration analysis
duration_df = df[['duration', 'rating']].dropna()

# Convert 'duration' to numeric (assuming it's in a format like 'X min')
duration_df['duration'] = pd.to_numeric(duration_df['duration'].str.extract('(\d+)')[0],
errors='coerce')



# Scatter plot for duration vs. user rating
scatter_plot = px.scatter(duration_df, x='duration', y='rating', title='Duration vs. User Rating')
scatter_plot.show()
```

```
# Analyzing the top 5 Directors on Netflix

# dealing with missing values in director columns
df['director']=df['director'].fillna('Director Not Specified')
df.head()

director_list = pd.DataFrame()
print(director_list)

director_list = df['director'].str.split(',', expand=True).stack()
print(director_list)

director_list = director_list.to_frame()
print(director_list)

# gave the name of coulmn
director_list.columns = ['Directors']
print(director_list)

# count of directors total content they created
directors = director_list.groupby(['Directors']).size().reset_index(name='Total Count')
print(directors)

# remove
directors = directors[directors.Directors != 'Director Not Specified']

print(directors)

directors = directors.sort_values(by=['Total Count'], ascending= False)
print(directors)

# top 5 director
top_5_dir = directors.head()
```

```
top_5_dir

# horizontal bar chart
top_5_dir = top_5_dir.sort_values(by=['Total Count'])
barChart = px.bar(top_5_dir, x='Total Count', y='Directors', title='Top 5 Directors on Netflix')
barChart.show()

# Analyzing the top 5 Actors on Netflix

# repalce NaN Values
df['cast'] = df['cast'].fillna('No cast Specified')
cast_df = pd.DataFrame()
cast_df = df['cast'].str.split(',', expand=True).stack()
cast_df = cast_df.to_frame()
cast_df.columns = ['Actor']
actors = cast_df.groupby(['Actor']).size().reset_index(name='Total Counts')
# remove no cast
actors = actors[actors.Actor != 'No cast Specified']
# sort them
actors = actors.sort_values(by=['Total Counts'], ascending=False)
top_5_actors = actors.head()
top_5_actors = top_5_actors.sort_values(by=['Total Counts'])
barChart2 = px.bar(top_5_actors, x='Total Counts', y='Actor', title='Top 5 Actors on NetFlix')
barChart2.show()

# Analyzing the content produced on Netflix based on years

# how many movies & Tv Shows re produced per year
df1= df[['type', 'release_year']]
df1 = df1.rename(columns = {'release_year':"Release Year", "type":"Type"})
df2 = df1.groupby(['Release Year', 'Type']).size().reset_index(name='Total Count')

print(df2)
```

```
df2 = df2[df2['Release Year']>=2000]

graph = px.line(df2, x='Release Year', y='Total Count', color ='Type', title='Trend of Content Produced on Netflix Every Year')

graph.show()


# Analyzing Top 10 Country on Netflix


# Quick feature engineering


# Helper column for various plots

df['count'] = 1


# Many productions have several countries listed - this will skew our results , we'll grab the first one mentioned


# Lets retrieve just the first country

df['first_country'] = df['country'].apply(lambda x: x.split(",")[0])

df['first_country'].head()


# Rating ages from this notebook: https://www.kaggle.com/andreshg/eda-beginner-to-expert-plotly (thank you!)


ratings_ages = {
    'TV-PG': 'Older Kids',
    'TV-MA': 'Adults',
    'TV-Y7-FV': 'Older Kids',
    'TV-Y7': 'Older Kids',
    'TV-14': 'Teens',
    'R': 'Adults',
    'TV-Y': 'Kids',
    'NR': 'Adults',
    'PG-13': 'Teens',
    'TV-G': 'Kids',
    'PG': 'Older Kids',
    'G': 'Kids',
```

```
    'UR': 'Adults',
    'NC-17': 'Adults'
}

df['target_ages'] = df['rating'].replace(ratings_ages)
df['target_ages'].unique()

# Genre

df['genre'] = df['listed_in'].apply(lambda x :  x.replace(' ,',',').replace(', ',',').split(','))

# Reducing name length

df['first_country'].replace('United States', 'USA', inplace=True)
df['first_country'].replace('United Kingdom', 'UK',inplace=True)
df['first_country'].replace('South Korea', 'S. Korea',inplace=True)

data = df.groupby('first_country')['count'].sum().sort_values(ascending=False)[:10]

# Plot

color_map = ['#f5f5f1' for _ in range(10)]
color_map[0] = color_map[1] = color_map[2] =  '#b20710' # color highlight

fig, ax = plt.subplots(1,1, figsize=(12, 6))
ax.bar(data.index, data, width=0.5,
       edgecolor='darkgray',
       linewidth=0.6,color=color_map)

#annotations
for i in data.index:
    ax.annotate(f"{data[i]}",
                xy=(i, data[i] + 150), #i like to change this to roughly 5% of the highest cat
                va = 'center', ha='center',fontweight='light', fontfamily='serif')
```

```
# Remove border from plot

for s in ['top', 'left', 'right']:
    ax.spines[s].set_visible(False)

# Tick labels

ax.set_xticklabels(data.index, fontfamily='serif', rotation=0)

# Title and sub-title

fig.text(0.09, 1, 'Top 10 countries on Netflix', fontsize=15, fontweight='bold', fontfamily='serif')

fig.text(0.09, 0.95, 'The three most frequent countries have been highlighted.', fontsize=12,
fontweight='light', fontfamily='serif')

fig.text(1.1, 1.01, 'Insight', fontsize=15, fontweight='bold', fontfamily='serif')

fig.text(1.1, 0.67, '''
The most prolific producers of
content for Netflix are, primarily,
the USA, with India and the UK
a significant distance behind.

It makes sense that the USA produces
the most content as, afterall,
Netflix is a US company.
'''
        , fontsize=12, fontweight='light', fontfamily='serif')

ax.grid(axis='y', linestyle='-', alpha=0.4)

grid_y_ticks = np.arange(0, 4000, 500) # y ticks, min, max, then step
```

```
ax.set_yticks(grid_y_ticks)
ax.set_axisbelow(True)

#Axis labels

#plt.xlabel("Country", fontsize=12, fontweight='light', fontfamily='serif',loc='left',y=-1.5)
#plt.ylabel("Count", fontsize=12, fontweight='light', fontfamily='serif')
 #plt.legend(loc='upper right')

# thicken the bottom line if you want to
plt.axhline(y = 0, color = 'black', linewidth = 1.3, alpha = .7)

ax.tick_params(axis='both', which='major', labelsize=12)



import matplotlib.lines as lines
l1 = lines.Line2D([1, 1], [0, 1], transform=fig.transFigure, figure=fig,color='black',lw=0.2)
fig.lines.extend([l1])

ax.tick_params(axis=u'both', which=u'both',length=0)

plt.show()

# TV & Movies is the highest rating of the dataset

# plot of rating by type
plt.figure(figsize=(10,8))
sns.countplot(df, x='rating', hue='type')
plt.title("Plot of rating by type")
plt.show()

# Sentiment Analysis of Netflix Content

df3 = df[['release_year', 'description']]
```

```
df3 = df3.rename(columns = {'release_year':"Release Year", "description":"Description"})
for index, row in df3.iterrows():
    d = row['Description']
    testimonial = TextBlob(d)
    p =testimonial.sentiment.polarity
    if p==0:
        sent = 'Neutral'
    elif p>0:
        sent = 'Positive'
    else:
        sent = 'negative'
    df3.loc[[index,2], 'Sentiment']=sent

df3 = df3.groupby(['Release Year', 'Sentiment']).size().reset_index(name='Total Count')

df3 = df3[df3['Release Year']>2005]

barGraph = px.bar(df3, x='Release Year', y= 'Total Count', color='Sentiment', title='Sentiment Analysis Of Content on Netflix ')
barGraph.show()

# Asking and Answering Questions
### Q1. Which country has the most number of titles produced?

most_titles= df.groupby('country').count().sort_values('title', ascending=False).head(5)
most_titles.reset_index(inplace=True)
most_titles

plt.bar(most_titles.country,most_titles.title)

plt.show()

#### Hence, united States produced the most number of titles

### Q2: Does Netflix have more Movies or TV Shows?
```

```
sns.countplot(x='type', data=df)
plt.show()



#### Clearly, Netflix have a higher number of movies than shows infact, the number of Movies is more than double of TV Shows


### Q3: What are the top 5 most popular ratings on netflix?


net_df_copy_rat=df['rating'].value_counts()
net_df_copy_rat = pd.DataFrame(net_df_copy_rat).reset_index()
net_df_copy_rat.columns=['rating','Nbr']
sns.barplot(x='rating', y='Nbr', data=net_df_copy_rat.head(5))
plt.show()


#### Hence, we saw TV-MA, TV-14 and TV-PG, R , PG-13 are the five most popular ratings among all of them


### Q4: Which are the top 5 Least popular genre on Netflix?


genre=df['listed_in'].value_counts().tail(5)
genre


plt.figure(figsize=(10,4))
genre.plot(kind='barh',color='red')
plt.title('5 Least popular genre on Netflix')
plt.show()


#### We saw in the bar chart above the 5 least popular netflix genre.


### Q5: Which were the top 5 years in number of titles released?


top_5 = df.groupby('release_year').count().sort_values('title', ascending=False).head(5)
top_5.reset_index(inplace=True)
```

```
top_5


plt.bar(top_5.release_year,top_5.title)

plt.show()


#### Clearly 2018 was the year highest number of netflix titles were released. followed by 2017, 2019, 2016, and 2015


! pip install ydata-Profiling

import pandas as pd

from ydata_profiling import ProfileReport


df = pd.read_csv('netflix_data.csv')

profile = ProfileReport(df, title="Profiling Report")

profile.to_file('output.html')
```

10. Future Scope

The analysis of Netflix's movies and TV shows data has provided valuable insights into current content trends and user preferences. However, there are several avenues for future research and exploration that could further enhance our understanding of the dynamic streaming landscape.

- Advanced User Behaviour Analysis.
- Personalization Algorithms.
- Collaboration with External Data Sources.
- Content Quality Assessment.

11. Conclusion

In conclusion, the analysis of Netflix's movies and TV shows data has shed light on various aspects of the streaming platform's content landscape. The exploration of content distribution, user preferences, and emerging trends provides a solid foundation for strategic decision-making. The insights gained can aid Netflix in refining its content strategy to better cater to a diverse and ever-evolving audience.

In essence, this project not only contributes to the understanding of Netflix's content landscape but also emphasizes the need for continuous data-driven strategies in the ever-evolving landscape of streaming platforms.

12. References

https://www.kaggle.com/datasets