Netflix Data Analysis

Import Libraries & Data

```
pip install textblob
```

```
Collecting textblob
  Downloading textblob-0.19.0-py3-none-any.whl.metadata (4.4 kB)
Requirement already satisfied: nltk>=3.9 in c:\users\ritul\anaconda3\
lib\site-packages (from textblob) (3.9.1)
Requirement already satisfied: click in c:\users\ritul\anaconda3\lib\
site-packages (from nltk>=3.9->textblob) (8.1.7)
Requirement already satisfied: joblib in c:\users\ritul\anaconda3\lib\
site-packages (from nltk>=3.9->textblob) (1.4.2)
Requirement already satisfied: regex>=2021.8.3 in c:\users\ritul\
anaconda3\lib\site-packages (from nltk>=3.9->textblob) (2024.11.6)
Requirement already satisfied: tqdm in c:\users\ritul\anaconda3\lib\
site-packages (from nltk>=3.9->textblob) (4.67.1)
Requirement already satisfied: colorama in c:\users\ritul\anaconda3\
lib\site-packages (from click->nltk>=3.9->textblob) (0.4.6)
Downloading textblob-0.19.0-py3-none-any.whl (624 kB)
   ---------------------------------------- 0.0/624.3 kB ? eta -:--:--
   ---------------------------------------- 624.3/624.3 kB 7.7 MB/s
eta 0:00:00
Installing collected packages: textblob
Successfully installed textblob-0.19.0
Note: you may need to restart the kernel to use updated packages.

WARNING: Ignoring invalid distribution ~ensorflow (C:\Users\RITUL\
anaconda3\Lib\site-packages)
WARNING: Ignoring invalid distribution ~ertifi (C:\Users\RITUL\
anaconda3\Lib\site-packages)
WARNING: Ignoring invalid distribution ~ransformers (C:\Users\RITUL\
anaconda3\Lib\site-packages)
WARNING: Ignoring invalid distribution ~umpy (C:\Users\RITUL\
anaconda3\Lib\site-packages)
WARNING: Ignoring invalid distribution ~~mpy (C:\Users\RITUL\
anaconda3\Lib\site-packages)
WARNING: Ignoring invalid distribution ~ensorflow (C:\Users\RITUL\
anaconda3\Lib\site-packages)
WARNING: Ignoring invalid distribution ~ertifi (C:\Users\RITUL\
anaconda3\Lib\site-packages)
WARNING: Ignoring invalid distribution ~ransformers (C:\Users\RITUL\
anaconda3\Lib\site-packages)
WARNING: Ignoring invalid distribution ~umpy (C:\Users\RITUL\
anaconda3\Lib\site-packages)
WARNING: Ignoring invalid distribution ~~mpy (C:\Users\RITUL\
```

```
anaconda3\Lib\site-packages)
WARNING: Ignoring invalid distribution ~ensorflow (C:\Users\RITUL\
anaconda3\Lib\site-packages)
WARNING: Ignoring invalid distribution ~ertifi (C:\Users\RITUL\
anaconda3\Lib\site-packages)
WARNING: Ignoring invalid distribution ~ransformers (C:\Users\RITUL\
anaconda3\Lib\site-packages)
WARNING: Ignoring invalid distribution ~umpy (C:\Users\RITUL\
anaconda3\Lib\site-packages)
WARNING: Ignoring invalid distribution ~~mpy (C:\Users\RITUL\
anaconda3\Lib\site-packages)
```

```
import numpy as np      #Linear Algebra Operations
import pandas as pd     # Used for Data Preparation
import plotly.express as px      # Used for Data Visualization
from textblob import TextBlob    # Used for Sentimental Analysis

df = pd.read_csv('netflix_titles.csv')
```

Checking Number of Rows and Columns in Data

```
df.shape
```

```
(8807, 12)
```

Checking Content available in Dataset

```
df.head()
```

```
  show_id     type                   title         director  \
0      s1    Movie   Dick Johnson Is Dead  Kirsten Johnson
1      s2  TV Show          Blood & Water              NaN
2      s3  TV Show              Ganglands  Julien Leclercq
3      s4  TV Show  Jailbirds New Orleans              NaN
4      s5  TV Show           Kota Factory              NaN

                                                cast        country  \
0                                                NaN  United States
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...   South Africa
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...            NaN
3                                                NaN            NaN
4  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...          India

           date_added  release_year rating   duration  \
0  September 25, 2021          2020  PG-13     90 min
1  September 24, 2021          2021  TV-MA  2 Seasons
2  September 24, 2021          2021  TV-MA   1 Season
3  September 24, 2021          2021  TV-MA   1 Season
4  September 24, 2021          2021  TV-MA  2 Seasons
```

```
                                              listed_in  \
0                                         Documentaries
1    International TV Shows, TV Dramas, TV Mysteries
2  Crime TV Shows, International TV Shows, TV Act...
3                              Docuseries, Reality TV
4  International TV Shows, Romantic TV Shows, TV ...

                                           description
0  As her father nears the end of his life, filmm...
1  After crossing paths at a party, a Cape Town t...
2  To protect his family from a powerful drug lor...
3  Feuds, flirtations and toilet talk go down amo...
4  In a city of coaching centers known to train I...
```

Checking Column Names of Dataset

```
df.columns
```

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

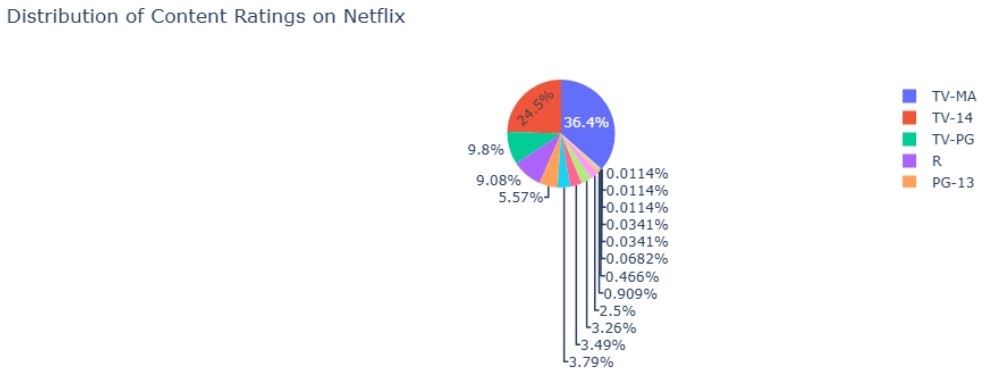
Taking the Count of Ratings available

```
x = df.groupby('rating').size().reset_index(name = 'counts')
print(x)
```

```
       rating  counts
0     66 min       1
1     74 min       1
2     84 min       1
3          G      41
4      NC-17       3
5         NR      80
6         PG     287
7      PG-13     490
8          R     799
9      TV-14    2160
10      TV-G     220
11     TV-MA    3207
12     TV-PG     863
13      TV-Y     307
14     TV-Y7     334
15  TV-Y7-FV       6
16        UR       3
```

Creating a Piechart based on Content Rating

```
pieChart = px.pie(x, values = 'counts', names = 'rating', title =
'Distribution of Content Ratings on Netflix')
pieChart.show()
```



Analyzing top 5 Directors on Netflix

```
df['director'] = df['director'].fillna('Director not Specified')
df.head()
```

```
  show_id     type                   title                director  \
0      s1    Movie   Dick Johnson Is Dead        Kirsten Johnson
1      s2  TV Show          Blood & Water  Director not Specified
2      s3  TV Show              Ganglands        Julien Leclercq
3      s4  TV Show  Jailbirds New Orleans  Director not Specified
4      s5  TV Show           Kota Factory  Director not Specified

                                                cast        country  \
0                                                NaN  United States
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...   South Africa
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...            NaN
3                                                NaN            NaN
4  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...          India

           date_added  release_year rating   duration  \
0  September 25, 2021          2020  PG-13     90 min
1  September 24, 2021          2021  TV-MA  2 Seasons
2  September 24, 2021          2021  TV-MA   1 Season
3  September 24, 2021          2021  TV-MA   1 Season
4  September 24, 2021          2021  TV-MA  2 Seasons

                                      listed_in  \
0                                  Documentaries
1    International TV Shows, TV Dramas, TV Mysteries
2  Crime TV Shows, International TV Shows, TV Act...
3                            Docuseries, Reality TV
4  International TV Shows, Romantic TV Shows, TV ...
```

```
                                          description
0  As her father nears the end of his life, filmm...
1  After crossing paths at a party, a Cape Town t...
2  To protect his family from a powerful drug lor...
3  Feuds, flirtations and toilet talk go down amo...
4  In a city of coaching centers known to train I...
```

```
directors_list = pd.DataFrame()
print(directors_list)
```

```
Empty DataFrame
Columns: []
Index: []
```

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

```
0     0           Kirsten Johnson
1     0    Director not Specified
2     0           Julien Leclercq
3     0    Director not Specified
4     0    Director not Specified
                   ...
8802  0             David Fincher
8803  0    Director not Specified
8804  0           Ruben Fleischer
8805  0              Peter Hewitt
8806  0               Mozez Singh
Length: 9612, dtype: object
```

```
directors_list = directors_list.to_frame()
print(directors_list)
```

```
                             0
0    0         Kirsten Johnson
1    0  Director not Specified
2    0         Julien Leclercq
3    0  Director not Specified
4    0  Director not Specified
...                        ...
8802 0           David Fincher
8803 0  Director not Specified
8804 0         Ruben Fleischer
8805 0            Peter Hewitt
8806 0             Mozez Singh

[9612 rows x 1 columns]
```

```
directors_list.columns = ['Director']
print(directors_list)
#directors_list.head()
```

```
                      Director
0    0         Kirsten Johnson
1    0  Director not Specified
2    0         Julien Leclercq
3    0  Director not Specified
4    0  Director not Specified
...                         ...
8802 0           David Fincher
8803 0  Director not Specified
8804 0         Ruben Fleischer
8805 0            Peter Hewitt
8806 0             Mozez Singh

[9612 rows x 1 columns]
```

```
directors =
directors_list.groupby(['Director']).size().reset_index(name = 'Total
Count')
print(directors)
```

```
                      Director  Total Count
0                Aaron Moorhead            2
1                   Aaron Woolf            1
2      Abbas Alibhai Burmawalla            1
3              Abdullah Al Noor            1
4           Abhinav Shiv Tiwari            1
...                         ...          ...
5116               Çagan Irmak            1
5117          Ísold Uggadóttir            1
5118        Óskar Thór Axelsson            1
5119           Ömer Faruk Sorak            2
5120               Şenol Sönmez            2

[5121 rows x 2 columns]
```

```
directors = directors[directors.Director != 'Director not Specified']
print(directors)
```

```
                      Director  Total Count
0                Aaron Moorhead            2
1                   Aaron Woolf            1
2      Abbas Alibhai Burmawalla            1
3              Abdullah Al Noor            1
4           Abhinav Shiv Tiwari            1
...                         ...          ...
5116               Çagan Irmak            1
5117          Ísold Uggadóttir            1
5118        Óskar Thór Axelsson            1
5119           Ömer Faruk Sorak            2
5120               Şenol Sönmez            2
```

```
[5120 rows x 2 columns]
```

```
directors= directors.sort_values(by = ['Total Count'])
print(directors)
```

```
             Director  Total Count
2560  Joanna Lombardi            1
2802       K. Subhash            1
2803     K.C. Bokadia            1
2804   K.S. Ravikumar            1
2805     KVR Mahendra            1
...                ...          ...
4652      Suhas Kadav           16
3236     Marcus Raboy           16
4068      Raúl Campos           18
261         Jan Suter           18
4021    Rajiv Chilaka           22
```

```
[5120 rows x 2 columns]
```

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

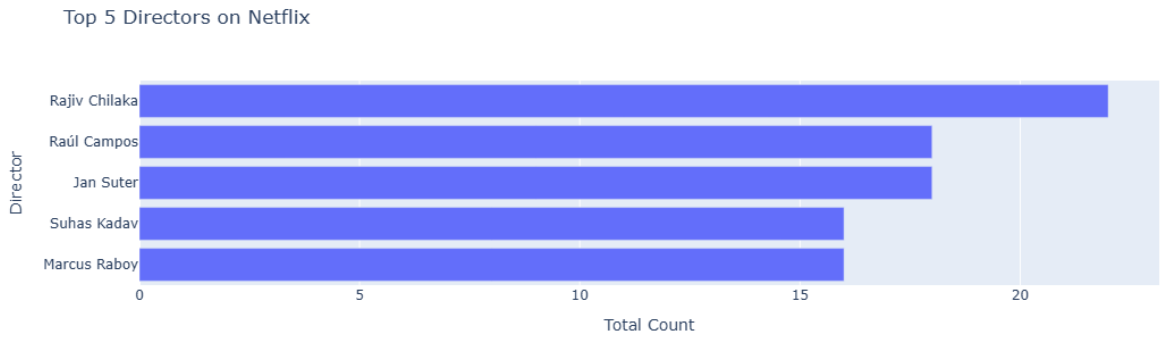
```
           Director  Total Count
4021  Rajiv Chilaka           22
261       Jan Suter           18
4068    Raúl Campos           18
3236   Marcus Raboy           16
4652    Suhas Kadav           16
...             ...          ...
3218    Marc Meyers            1
3217     Marc Levin            1
3216   Marc Francis            1
3215  Marc Fouchard            1
5014  Will Lovelace            1
```

```
[5120 rows x 2 columns]
```

```
top5Directors= directors.head()
print(top5Directors)
```

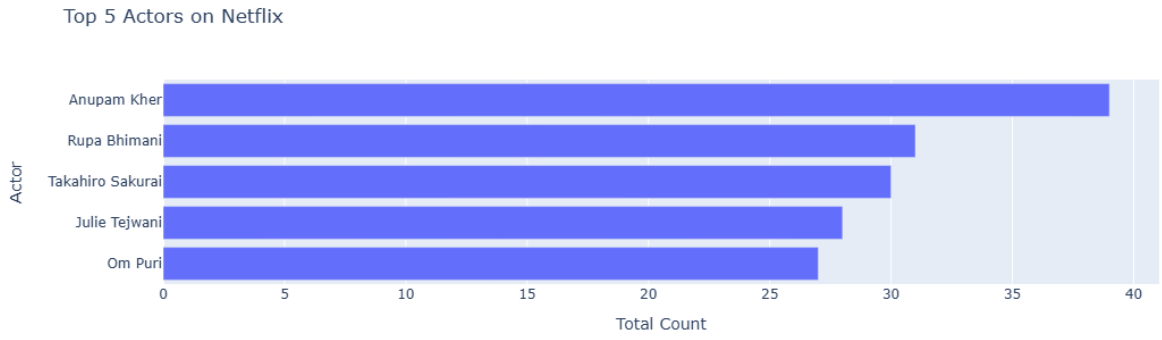
```
           Director  Total Count
4021  Rajiv Chilaka           22
261       Jan Suter           18
4068    Raúl Campos           18
3236   Marcus Raboy           16
4652    Suhas Kadav           16
```

```
top5Directors = top5Directors.sort_values(by = ['Total Count'])
barChart = px.bar(top5Directors, x = 'Total Count', y = 'Director',
title = 'Top 5 Directors on Netflix')
barChart.show()
```



Analyzing top 5 Actors on Netflix

```
df['cast'] = df['cast'].fillna('Cast not 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
Count')
actors = actors[actors.Actor != 'Cast not Specified']
actors = actors.sort_values(by = ['Total Count'], ascending = False)
top5Actors = actors.head()
top5Actors = top5Actors.sort_values(by = ['Total Count'])
barChart2 = px.bar(top5Actors, x = 'Total Count', y = 'Actor', title =
'Top 5 Actors on Netflix')
barChart2.show()
```



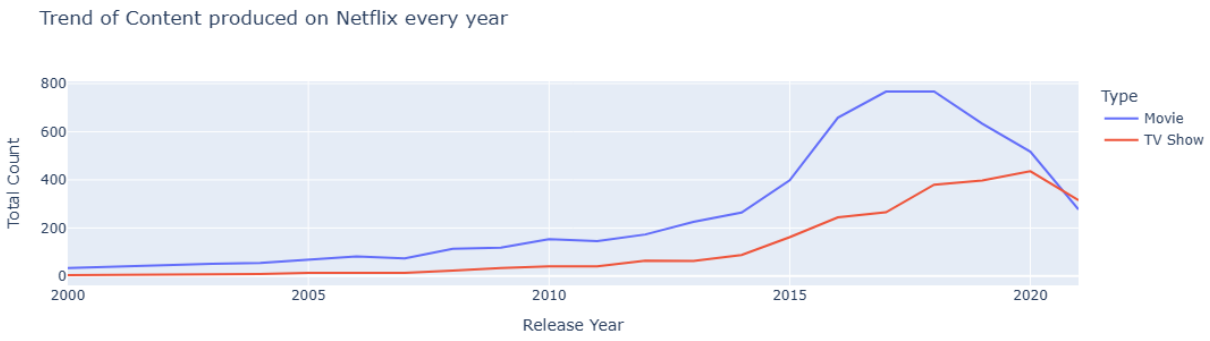
Analyzing Content produced on Netflix based on Years

```
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)
```

```
     Release Year     Type  Total Count
0            1925  TV Show            1
1            1942    Movie            2
2            1943    Movie            3
3            1944    Movie            3
4            1945    Movie            3
..            ...      ...          ...
114          2019  TV Show          397
115          2020    Movie          517
116          2020  TV Show          436
117          2021    Movie          277
118          2021  TV Show          315

[119 rows x 3 columns]
```

```
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()
```



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.at[index, '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()
```