Netflix

Author: Yash Kant Sharma

Importing Libraries

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

import warnings
warnings.filterwarnings('ignore')
```

Loading Datset

In [2]:

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

In [83]:

```
df.head()
```

Out[83]:

	show_id	type	title	director	cast	country	date_added	release_year	rat
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	David Attenborough	United States	September 25, 2021	2020	I
1	s2	TV Show	Blood & Water	Rajiv Chilaka	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	United States	September 24, 2021	2021	
3	s4	TV Show	Jailbirds New Orleans	Rajiv Chilaka	David Attenborough	United States	September 24, 2021	2021	
4	s5	TV Show	Kota Factory	Rajiv Chilaka	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	

In [4]:

```
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
```

There are Total 12 Columns in which 11 columns are object dtype and 1 column int dtype

In [5]:

```
df.shape
```

Out[5]:

```
(8807, 12)
```

Tere are Total 8807 rows and 12 columns

Computing Total No. of Missing Values and the Percentage of Missing Values.

In [6]:

```
df.isnull().sum()
```

Out[6]:

```
show_id            0
type               0
title              0
director        2634
cast             825
country          831
date_added        10
release_year       0
rating             4
duration           3
listed_in          0
description        0
dtype: int64
```

In [7]:

```python
df.isnull().sum()/df.shape[0]*100
```

Out[7]:

```
show_id          0.000000
type             0.000000
title            0.000000
director        29.908028
cast             9.367549
country          9.435676
date_added       0.113546
release_year     0.000000
rating           0.045418
duration         0.034064
listed_in        0.000000
description      0.000000
dtype: float64
```

In [8]:

```python
for i in df:
    print(i,'...............',df[i].unique(),':::::::::::::::::::::::
```

```
 'United Kingdom, France, Poland, Germany, United States'
 'Ireland, Switzerland, United Kingdom, France, United States'
 'United Kingdom, South Africa, France'
 'Ireland, United Kingdom, France, Germany' 'Russia, United States'
 'United Kingdom, United States, France' 'United Kingdom,'
 'United States, India, United Kingdom' 'Kenya' 'Spain, Argentina'
 'India, United Kingdom, France, Qatar' 'Belgium, France'
 'Argentina, Chile' 'United States, Thailand' 'Chile, Brazil'
 'United States, Colombia' 'Canada, United States, United Kingdom'
 'Uruguay' 'Luxembourg' 'United States, Cambodia, Romania' 'Banglades
h'
 'Spain, Belgium, United States'
 'United Kingdom, United States, Australia'
 'Canada, United States, France' 'Portugal, United States'
 'Portugal, Spain' 'India, United States' 'United Kingdom, Ireland'
 'United Kingdom, Spain, United States' 'Hungary, United States'
 'United States, South Korea' 'Canada, United States, Cayman Islands'
 'India, France' 'France, Canada' 'Canada, Hungary, United States'
 'Norway' 'Canada, United Kingdom, United States'
 'United Kingdom, Germany, France, United States' 'Denmark, United Sta
```

In [9]:

```python
df['director'].value_counts().head()
```

Out[9]:

```
Rajiv Chilaka             19
Raúl Campos, Jan Suter    18
Marcus Raboy              16
Suhas Kadav               16
Jay Karas                 14
Name: director, dtype: int64
```

In [10]:

```
df['cast'].value_counts().head()
```

Out[10]:

```
David Attenborough
19
Vatsal Dubey, Julie Tejwani, Rupa Bhimani, Jigna Bhardwaj, Rajesh Kava, M
ousam, Swapnil    14
Samuel West
10
Jeff Dunham
7
David Spade, London Hughes, Fortune Feimster
6
Name: cast, dtype: int64
```

In [11]:

```
df['country'].value_counts().head()
```

Out[11]:

```
United States     2818
India              972
United Kingdom     419
Japan              245
South Korea        199
Name: country, dtype: int64
```

In [12]:

```
df['director'].fillna(df['director'].mode()[0],inplace=True)
df['cast'].fillna(df['cast'].mode()[0],inplace=True)
df['country'].fillna(df['country'].mode()[0],inplace=True)
```

In [13]:

```
df.dropna(inplace=True)
```

In [14]:

```
df.isnull().sum()
```

Out[14]:

```
show_id         0
type            0
title           0
director        0
cast            0
country         0
date_added      0
release_year    0
rating          0
duration        0
listed_in       0
description     0
dtype: int64
```

In [15]:

```
df.head()
```

Out[15]:

	show_id	type	title	director	cast	country	date_added	release_year	rat
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	David Attenborough	United States	September 25, 2021	2020	I
1	s2	TV Show	Blood & Water	Rajiv Chilaka	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	United States	September 24, 2021	2021	
3	s4	TV Show	Jailbirds New Orleans	Rajiv Chilaka	David Attenborough	United States	September 24, 2021	2021	
4	s5	TV Show	Kota Factory	Rajiv Chilaka	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	

Insight 1:How many movies or shows are present in this dataset?

In [117]:

```
x=df['type'].value_counts().reset_index()
x=x.rename(columns={'index':'Type','type':'Count'})
x
```

Out[117]:

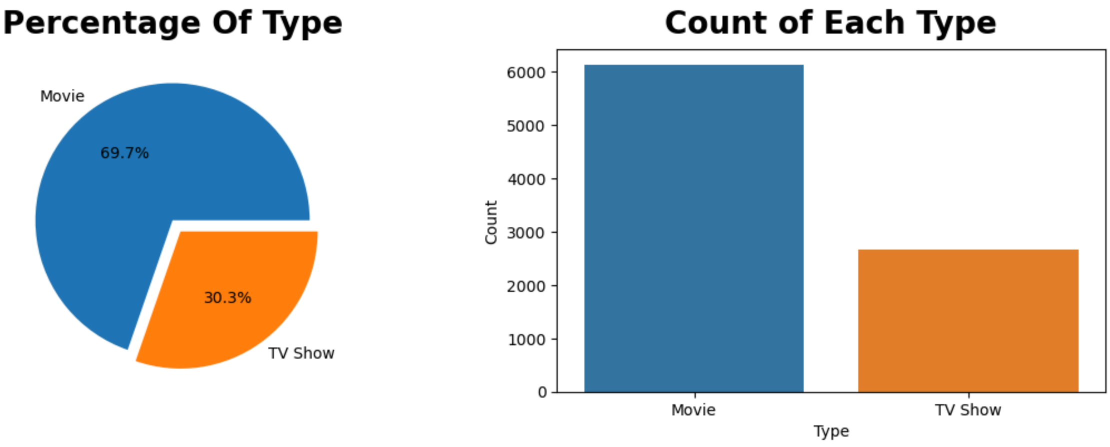
	Type	Count
0	Movie	6126
1	TV Show	2664

In [118]:

```
plt.figure(figsize=[14,4])

# Pie chart
plt.subplot(1,2,1)
plt.pie(x=x['Count'], labels=x['Type'], autopct='%1.1f%%',explode=[0,0
plt.title('Percentage Of Type',fontweight="black",size=20,pad=10)

# Bar plot
plt.subplot(1,2,2)
sns.barplot(data=x, x='Type', y='Count')
plt.title('Count of Each Type',fontweight="black",size=20,pad=10);
```



In [38]:

```
df.head()
```

Out[38]:

	show_id	type	title	director	cast	country	date_added	release_year	rat
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	David Attenborough	United States	September 25, 2021	2020	I
1	s2	TV Show	Blood & Water	Rajiv Chilaka	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	United States	September 24, 2021	2021	
3	s4	TV Show	Jailbirds New Orleans	Rajiv Chilaka	David Attenborough	United States	September 24, 2021	2021	
4	s5	TV Show	Kota Factory	Rajiv Chilaka	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	

Insight 2: Which director has made the maximum number of movies?

In [99]:

```
x=df['director'].value_counts().head().reset_index()
x
```

Out[99]:

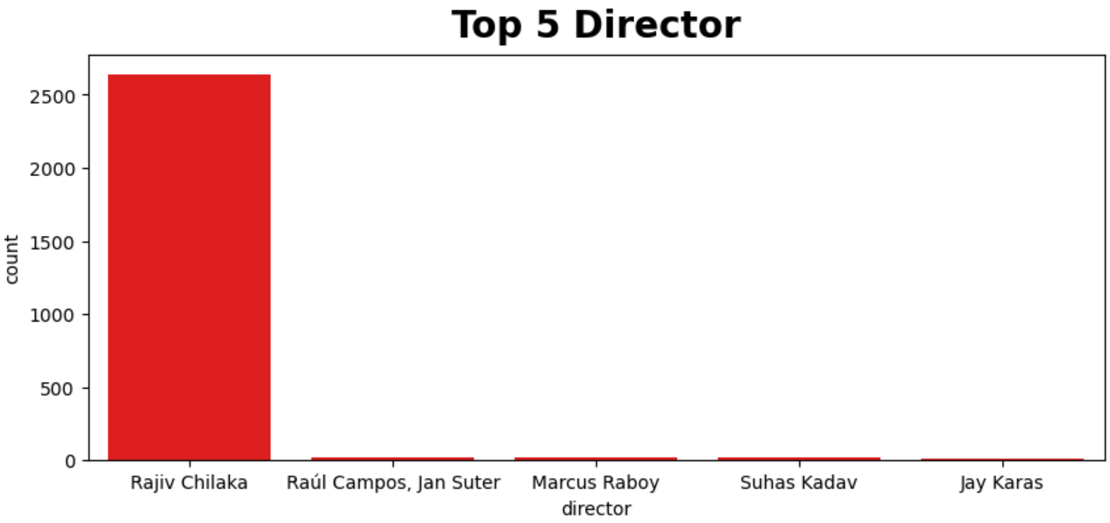
	index	director
0	Rajiv Chilaka	2640
1	Raúl Campos, Jan Suter	18
2	Marcus Raboy	16
3	Suhas Kadav	16
4	Jay Karas	14

According To this df['director'].value_counts() ------>

1. Rajeev Chilaka has making highest movies than others

In [104]:

```
plt.figure(figsize=[10,4])
sns.barplot(data=x,x=x['index'],y=x['director'],color='Red')
plt.xlabel('director')
plt.ylabel('count')
plt.title('Top 5 Director',fontweight='black',size=20,pad=10);
```



Insight 3: In which country have the maximum number of movies been made?

In [121]:

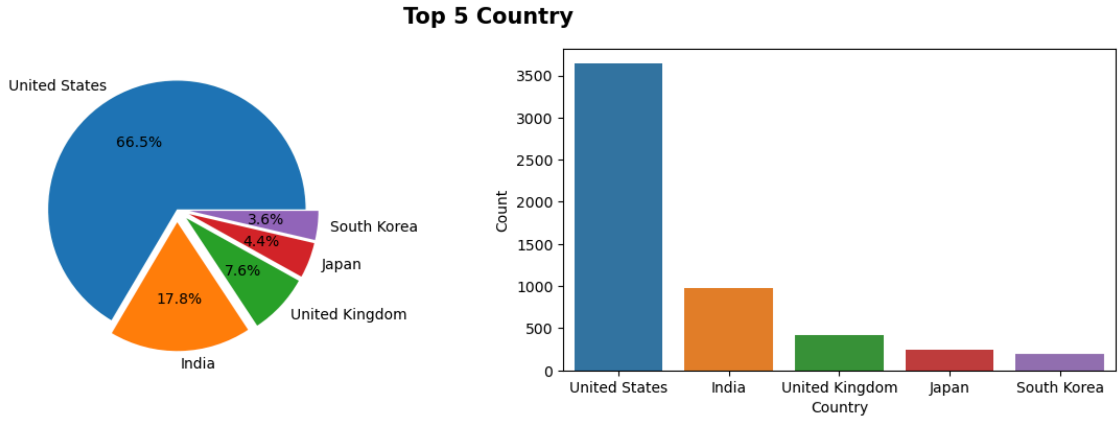
```
x=df['country'].value_counts().head().reset_index()
x
```

Out[121]:

	index	country
0	United States	3638
1	India	972
2	United Kingdom	418
3	Japan	243
4	South Korea	199

In [122]:

```
plt.figure(figsize=[15,4])
plt.subplot(1,2,1)
plt.pie(x=x['country'],labels=x['index'],autopct='%0.1f%%',explode=[0,
plt.subplot(1,2,2)
sns.barplot(data=x,x=x['index'],y=x['country'])
plt.xlabel('Country')
plt.ylabel('Count')
plt.suptitle('Top 5 Country',size=15,fontweight='black');
```



In [54]:

```
df.head()
```

Out[54]:

	show_id	type	title	director	cast	country	date_added	release_year	rat
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	David Attenborough	United States	September 25, 2021	2020	I
1	s2	TV Show	Blood & Water	Rajiv Chilaka	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	United States	September 24, 2021	2021	
3	s4	TV Show	Jailbirds New Orleans	Rajiv Chilaka	David Attenborough	United States	September 24, 2021	2021	
4	s5	TV Show	Kota Factory	Rajiv Chilaka	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	

Insight 4: Which year has released the highest number of movies or shows?

In [126]:

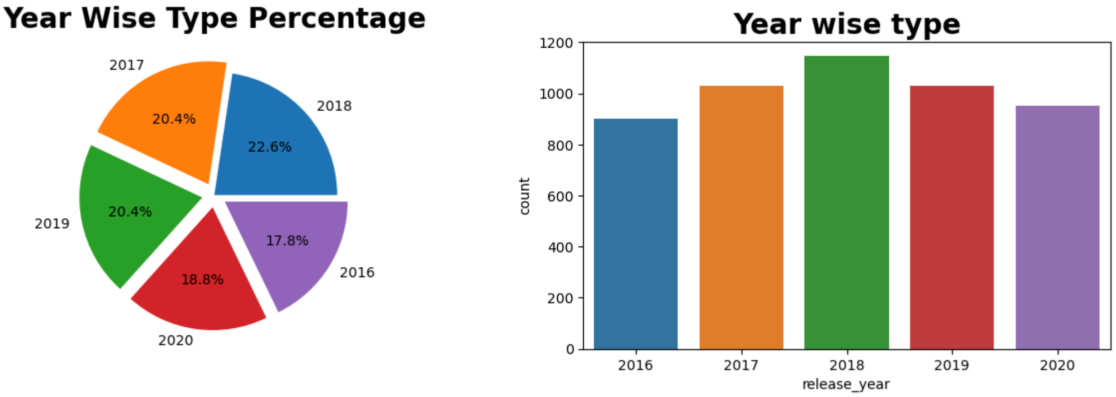
```
x=df['release_year'].value_counts().head().reset_index()
x=x.rename(columns={'index':'release_year','release_year':'count'})
x
```

Out[126]:

	release_year	count
0	2018	1146
1	2017	1030
2	2019	1030
3	2020	953
4	2016	901

In [130]:

```
plt.figure(figsize=[15,4])
plt.subplot(1,2,1)
plt.pie(x=x['count'],labels=x['release_year'],autopct='%0.1f%%',explod
plt.title('Year Wise Type Percentage',fontweight='black',size=20,pad=1
plt.subplot(1,2,2)
sns.barplot(data=x,x=x['release_year'],y=x['count'])
plt.title('Year wise type',fontsize=20,fontweight='black');
```



Insight 5: Highest Rating

In [124]:

```
x=df['rating'].value_counts().head().reset_index()
x=x.rename(columns={'index':'rating','rating':'count'})
x
```

Out[124]:

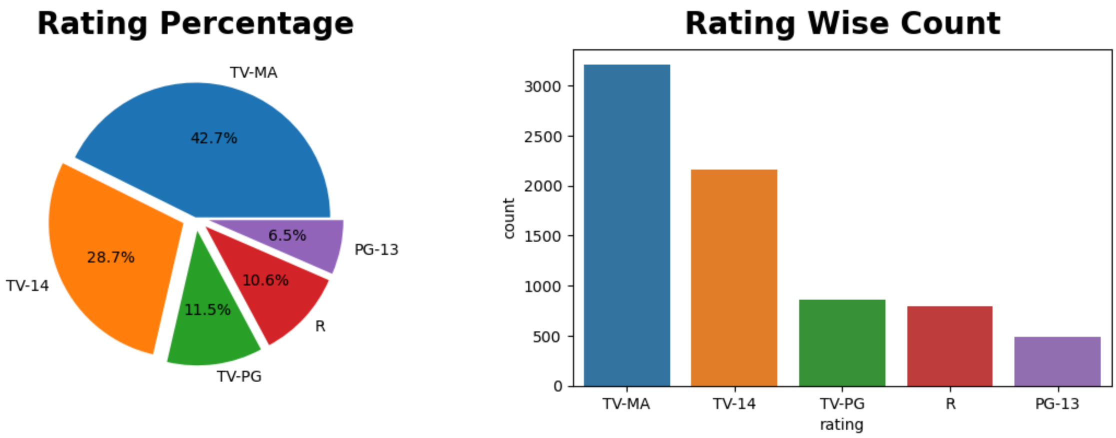
	rating	count
0	TV-MA	3205
1	TV-14	2157
2	TV-PG	861
3	R	799
4	PG-13	490

In [125]:

```
plt.figure(figsize=(14,4))
plt.subplot(1,2,1)
plt.pie(data=x,x=x['count'],labels='rating',autopct='%0.01f%%',explode
plt.title("Rating Percentage",fontweight="black",size=20,pad=10)
plt.subplot(1,2,2)
sns.barplot(data=x,x=x['rating'],y=x['count'])
plt.title("Rating Wise Count",fontweight="black",size=20,pad=10)
```

Out[125]:

```
Text(0.5, 1.0, 'Rating Wise Count')
```


Rating Percentage
TV-MA
TV-14
PG-13
R
TV-PG
Rating Wise Count
count
TV-MA
TV-14
TV-PG
R
PG-13
rating