```
import numpy as np
import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.impute import KNNImputer
from sklearn import linear_model
from sklearn.metrics import r2_score
from sklearn.tree import DecisionTreeRegressor
from lightgbm import LGBMRegressor
from xgboost import XGBRegressor
from sklearn.metrics import mean_absolute_error,r2_score, confusion_matrix
from sklearn.model_selection import cross_val_score

import warnings
warnings.filterwarnings('ignore')

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))


netflix=pd.read_csv('/content/netflix_titles.csv')


netflix
```

	show_id	type	title	director	cast	country	
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	S
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	S
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	S
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	S
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	S
...	...	...	...	...	...	...	
5393	s5394	TV Show	Breakout	NaN	Jeanette Aw, Elvin Ng, Zhou Ying, Christopher ...	NaN	
5394	s5395	Movie	Hans Teeuwen: Real Rancour	Doesjka van Hoogdalem	Hans Teeuwen	Netherlands	
5395	s5396	TV Show	Intersection	NaN	İbrahim Çelikkol, Belçim Bilgin, Alican Yüceso...	Turkey	
5396	s5397	Movie	Lal Patthar	Sushil Majumdar	Raaj Kumar, Hema Malini, Rakhee Gulzar, Vinod ...	India	
5397	s5398	TV Sh	NaN	NaN	NaN	NaN	

5398 rows × 12 columns

```
netflix.shape
```

```
(5398, 12)
```

```
netflix.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5398 entries, 0 to 5397
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
```

```
 ---  ------        --------------  -----
  0   show_id       5398 non-null   object
  1   type          5398 non-null   object
  2   title         5397 non-null   object
  3   director      3515 non-null   object
  4   cast          4903 non-null   object
  5   country       4735 non-null   object
  6   date_added    5397 non-null   object
  7   release_year  5397 non-null   float64
  8   rating        5397 non-null   object
  9   duration      5397 non-null   object
  10  listed_in     5397 non-null   object
  11  description   5397 non-null   object
dtypes: float64(1), object(11)
memory usage: 506.2+ KB
```

```
netflix.isnull().sum()
```

```
show_id            0
type               0
title              1
director        1883
cast             495
country          663
date_added         1
release_year       1
rating             1
duration           1
listed_in          1
description        1
dtype: int64
```

```
netflix.country.fillna(value="unknown", inplace =True)
netflix.country
```

```
0       United States
1        South Africa
2             unknown
3             unknown
4               India
              ...
5393          unknown
5394      Netherlands
5395           Turkey
5396            India
5397          unknown
Name: country, Length: 5398, dtype: object
```

```
netflix.date_added.fillna(value = "unknown", inplace = True)
netflix.date_added
```

```
0       September 25, 2021
1       September 24, 2021
2       September 24, 2021
3       September 24, 2021
4       September 24, 2021
                ...
5393          July 1, 2017
5394          July 1, 2017
5395          July 1, 2017
5396          July 1, 2017
5397              unknown
Name: date_added, Length: 5398, dtype: object
```

```
netflix.isnull().sum()
```

```
show_id            0
type               0
title              1
director        1883
cast             495
country            0
date_added         0
release_year       1
rating             1
duration           1
listed_in          1
description        1
dtype: int64
```

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

```
netflix.isnull().sum()
```

```
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
```

```
#To check for the type

netflix.Movie_ID.value_counts().index
```

```
Int64Index([    1, 11845, 11851, 11850, 11849, 11848, 11847, 11846, 11844,
            11836,
            ...
             5925,  5926,  5927,  5928,  5929,  5930,  5931,  5932,  5933,
            17770],
           dtype='int64', length=17770)
```

```
netflix.country.value_counts().index
```

```
Index(['United States', 'India', 'unknown', 'Nigeria', 'United Kingdom',
       'Japan', 'Philippines', 'Spain', 'Turkey', 'Indonesia',
       ...
       'Lebanon, France', 'France, Belgium, Italy',
       'United States, China, Colombia',
       'Lebanon, United Arab Emirates, France, Switzerland, Germany',
       'Canada, United Kingdom', 'Canada, Belgium', 'China, United States',
       'Ireland, Luxembourg, Belgium', 'Spain, Thailand, United States',
       'Chile, France'],
      dtype='object', length=380)
```

```
#visualizing the type
plt.figure(figsize=(10,8))

plt.pie(netflix.type.value_counts(),
        labels = netflix.type.value_counts().index,
        labeldistance = None, autopct="%.2f",
        textprops = {'fontsize': 16,},
        colors = ['lightsteelblue','lightsalmon' ] )
plt.legend()
plt.show()
```


```
last_decade = netflix[["type", "release_year"]]
last_decade = last_decade.rename(columns = {"release_year" : "Release Year"})
last_decade = last_decade[last_decade["Release Year"]>=2010]
last_decade
```

	type	Release Year
2	TV Show	2021.0
5	TV Show	2021.0
6	Movie	2021.0
8	TV Show	2021.0
9	Movie	2021.0
...	...	...
5385	Movie	2017.0
5387	Movie	2012.0
5388	TV Show	2016.0
5389	Movie	2011.0
5394	Movie	2018.0

2807 rows × 2 columns

```
last_decade_df = last_decade.groupby("Release Year")["type"].size().reset_index()
last_decade_df = pd.DataFrame(last_decade_df)
last_decade_df
```

	Release Year	type
0	2010.0	61
1	2011.0	67
2	2012.0	91
3	2013.0	91
4	2014.0	110
5	2015.0	113
6	2016.0	168
7	2017.0	331
8	2018.0	512
9	2019.0	510
10	2020.0	497
11	2021.0	256

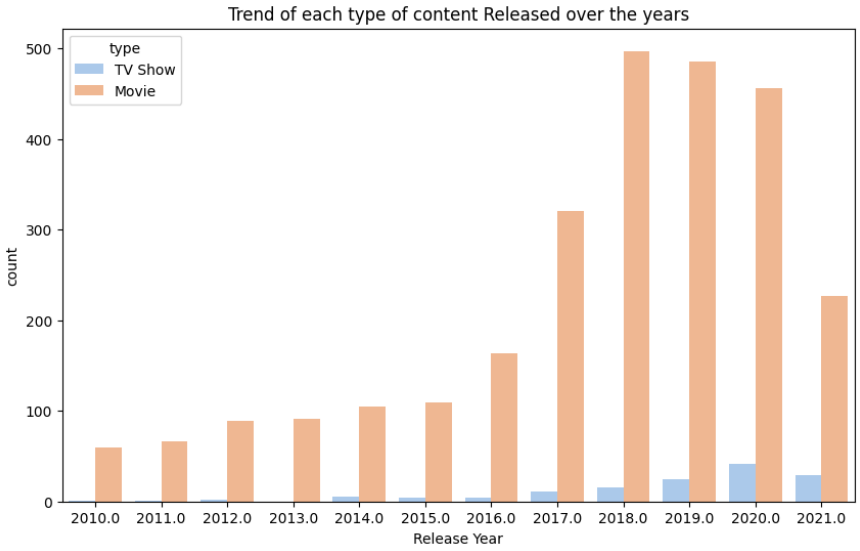
```
last_decade.groupby("Release Year")["type"].value_counts()
```

```
Release Year  type
2010.0        Movie       60
              TV Show      1
2011.0        Movie       66
              TV Show      1
2012.0        Movie       89
              TV Show      2
2013.0        Movie       91
2014.0        Movie      105
              TV Show      5
2015.0        Movie      109
              TV Show      4
2016.0        Movie      164
              TV Show      4
2017.0        Movie      320
              TV Show     11
2018.0        Movie      497
              TV Show     15
2019.0        Movie      485
              TV Show     25
```

```
2020.0        Movie      456
              TV Show     41
2021.0        Movie      227
              TV Show     29
Name: type, dtype: int64
```

```
plt.figure(figsize = (10,6))
count_plot = sns.countplot(x = "Release Year", data = last_decade, hue="type",
             palette= "pastel")
count_plot.set(title = "Trend of each type of content Released over the years")
;
```

''

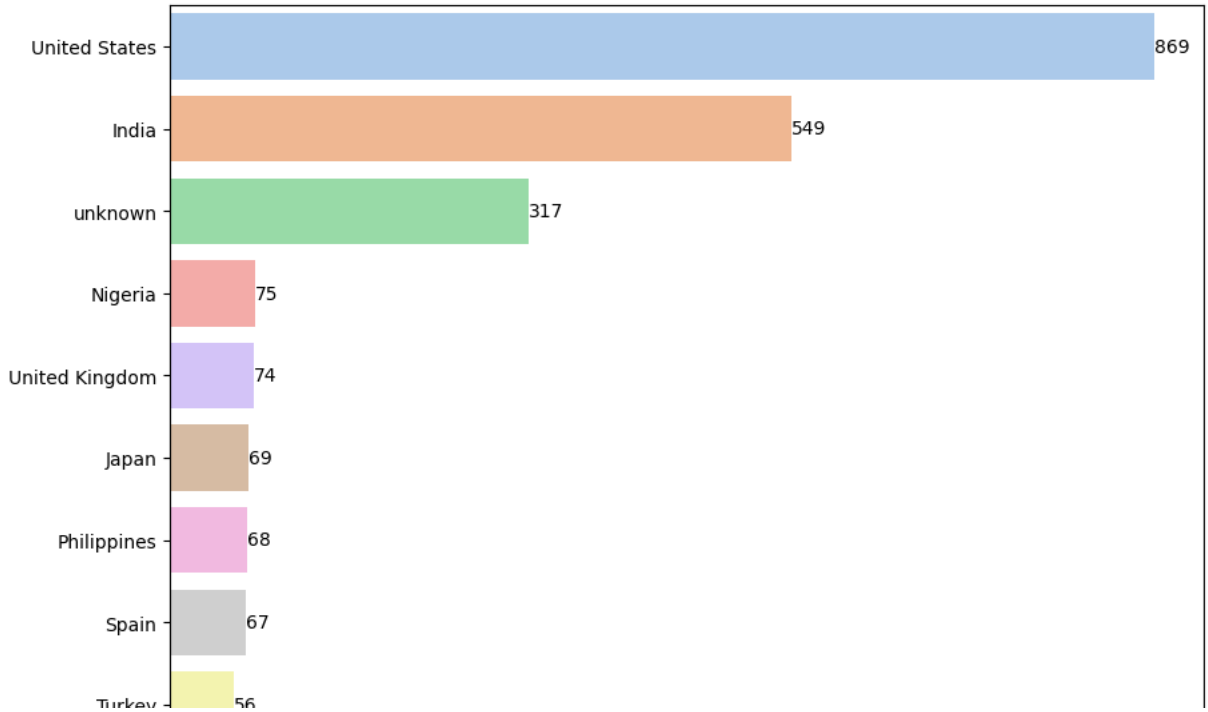



```
#Country

top_10_countries= netflix.country.value_counts().head(10)
top_10_countries = pd.DataFrame(top_10_countries)
top_10_countries
```

	country
United States	869
India	549
unknown	317
Nigeria	75
United Kingdom	74
Japan	69
Philippines	68
Spain	67
Turkey	56
Indonesia	51

```
plt.figure(figsize = (10,8))
country_plot = sns.barplot(x = netflix.country.value_counts()[:10].values,
                           y= netflix.country.value_counts()[:10].index,palette = "pastel")
for i in country_plot.containers:
    country_plot.bar_label(i);
```




```
# Rating

netflix.rating.unique()
```

```
array(['TV-MA', 'PG', 'TV-14', 'PG-13', 'TV-PG', 'TV-Y', 'R', 'TV-G',
       'TV-Y7', 'G', 'NC-17'], dtype=object)
```

```
plt.figure(figsize= (10,6))
sns.countplot(x="rating", data=netflix, palette="pastel",)
plt.title("count of Rating by Movie and Shows");
```