Shivani Kushwaha

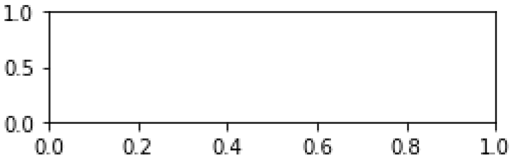
Contact No. 9453422900

Mail id : shivanikush1797@gmail.com

Task 1: Load the dataset in your environment.

```
import pandas as pd

# Load the dataset
df = pd.read_excel(r'asbl_data_analyst_interview_assignment_netflix.xlsx')

# Print the first five rows of the dataset
print(df.head())
```

```
      Type                    Title          Director  \
0    Movie   Dick Johnson Is Dead  Kirsten Johnson
1  TV Show            Blood & Water               NaN
2  TV Show                Ganglands   Julien Leclercq
3  TV Show  Jailbirds New Orleans                NaN
4  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

  Release_year                                           Genres
0         2020                                    Documentaries
1         2021    International TV Shows, TV Dramas, TV Mysteries
2         2021   Crime TV Shows, International TV Shows, TV Act...
3         2021                           Docuseries, Reality TV
4         2021   International TV Shows, Romantic TV Shows, TV ...
```



Task 2: Perform EDA (exploratory data analysis) on the dataset.

1. Check the size of the dataset and the data types of each attribute.

```
# Print the shape of the dataset
print(df.shape)

# Print the data types of each attribute
print(df.dtypes)
```

```
(8807, 7)
Type            object
Title           object
Director        object
Cast            object
Country         object
Release_year     int64
Genres          object
dtype: object
```

2. Check for missing values and handle them if necessary.

```
# Check for missing values
print(df.isnull().sum())

# Handle missing values if necessary
# For example, you can drop the rows with missing values
df.dropna(inplace=True)
```

```
Type               0
Title              0
Director        2634
Cast             825
Country          831
Release_year       0
Genres             0
dtype: int64
```

3. Check the distribution of numerical attributes.

```
# Print the summary statistics of numerical attributes
print(df.describe())
```

```
       Release_year
count   5336.000000
mean    2012.743253
```

```
std        9.622570
min     1942.000000
25%     2011.000000
50%     2016.000000
75%     2018.000000
max     2021.000000
```

4. Check the frequency distribution of categorical attributes.

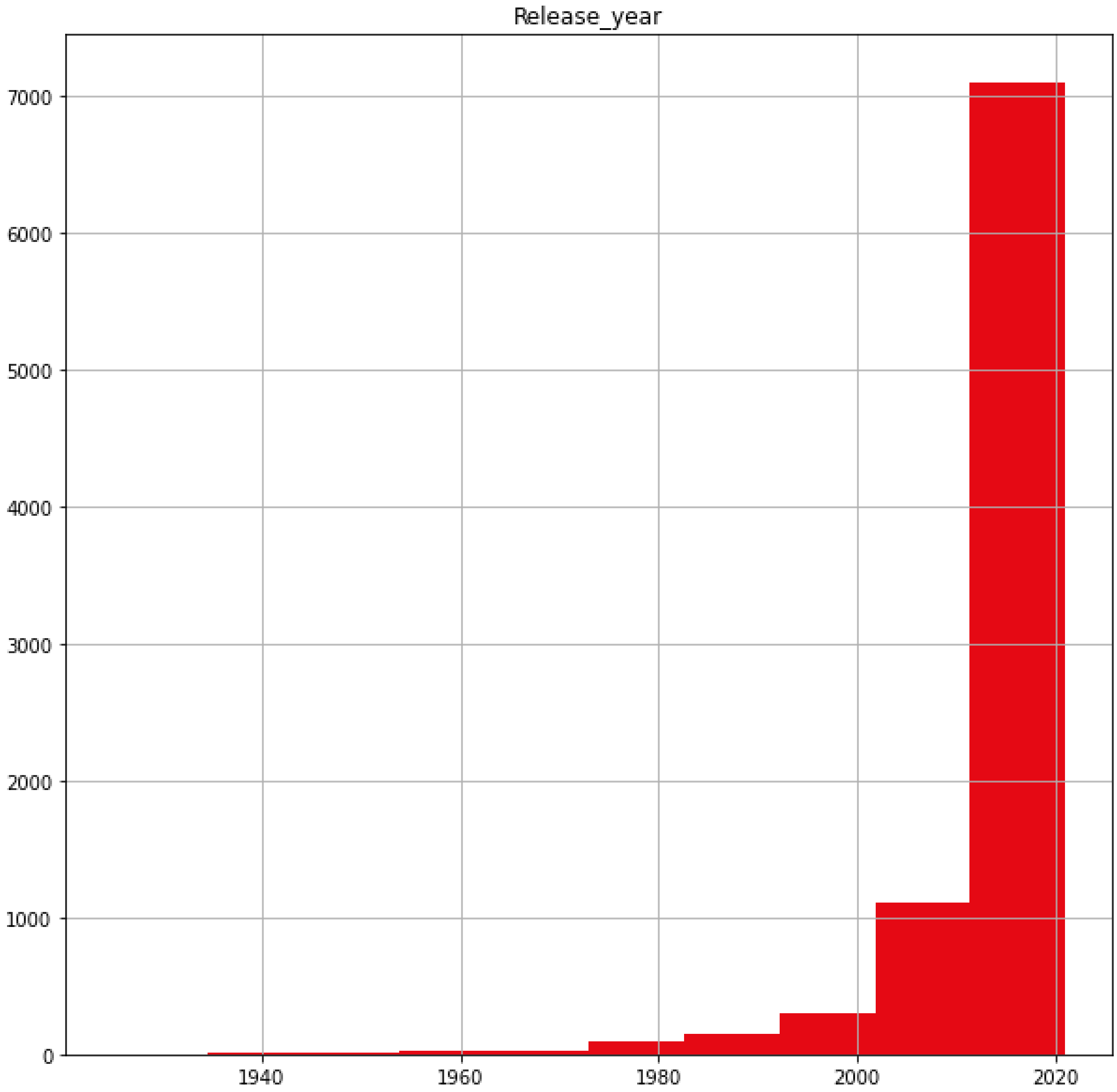
```
# Print the frequency distribution of categorical attributes
print(df['Type'].value_counts())
print(df['Country'].value_counts())
print(df['Genres'].value_counts())
```

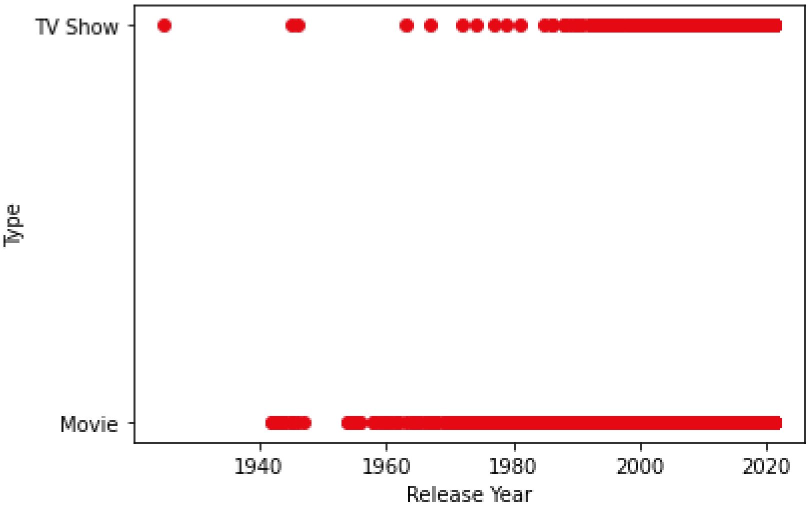
```
Movie      5189
TV Show     147
Name: Type, dtype: int64
United States                                   1849
India                                            875
United Kingdom                                   183
Canada                                           107
Spain                                             91
                                                ... 
Uruguay, Guatemala                                 1
Romania, Bulgaria, Hungary                         1
Philippines, United States                         1
India, United Kingdom, Canada, United States       1
United Arab Emirates, Jordan                       1
Name: Country, Length: 604, dtype: int64
Dramas, International Movies                               336
Stand-Up Comedy                                            286
Comedies, Dramas, International Movies                     257
Dramas, Independent Movies, International Movies           243
Children & Family Movies, Comedies                         179
                                                          ... 
Comedies, Documentaries                                      1
International TV Shows, Romantic TV Shows, TV Mysteries      1
Horror Movies, International Movies, Sci-Fi & Fantasy        1
Reality TV                                                   1
Cult Movies, Dramas, Thrillers                               1
Name: Genres, Length: 335, dtype: int64
```

4. Visualize the data to identify patterns and relationships.

```
# Visualize the distribution of numerical attributes using histograms
import matplotlib.pyplot as plt
df.hist(figsize=(10,10),color = "#E50914")
plt.show()

# Visualize the relationship between attributes using scatter plots
plt.scatter(df['Release_year'], df['Type'], color = '#E50914')
```

```
    plt.xlabel('Release Year')
    plt.ylabel('Type')
    plt.show()
```

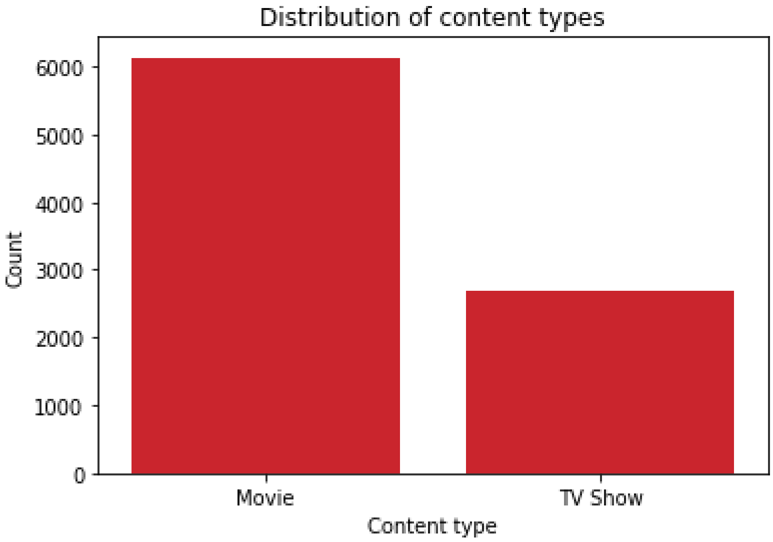




Task 3: Plot some meaningful graphs here which convey some insights and those insights businesses can use to further increase their revenue and attract more customers. Make sure that the insights found must be backed up by data and share some recommendations for the stakeholders.

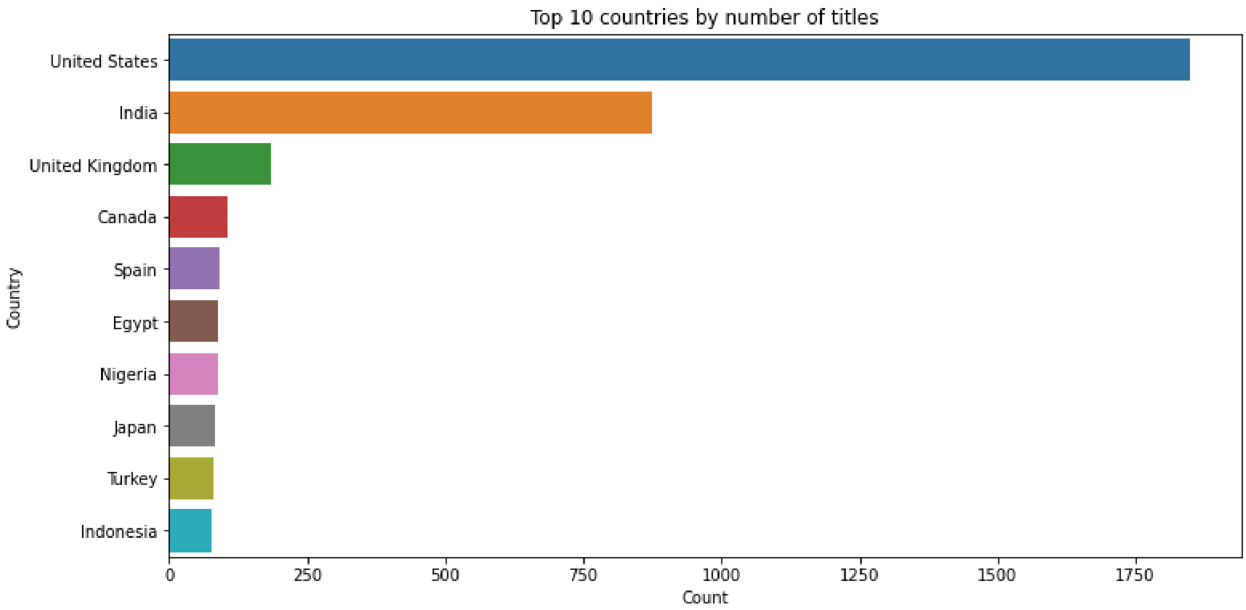
1. Distribution of content types

```
import seaborn as sns
import matplotlib.pyplot as plt

sns.countplot(x="Type", data=df, color = "#E50914")
plt.title("Distribution of content types")
plt.xlabel("Content type")
plt.ylabel("Count")
plt.show()
```



2. Top 10 countries by number of titles

```
plt.figure(figsize=(12,6))
sns.countplot(y="Country", data=df, order=df["Country"].value_counts().iloc[:10].ir
plt.title("Top 10 countries by number of titles")
plt.xlabel("Count")
plt.ylabel("Country")
plt.show()
```



3. Top 10 directors by number of titles

```
plt.figure(figsize=(12,6))
sns.countplot(y="Director", data=df, order=df["Director"].value_counts().iloc[:10].
plt.title("Top 10 directors by number of titles")
plt.xlabel("Count")
plt.ylabel("Director")
plt.show()
```

Top 10 directors by number of titles
Raúl Campos, Jan Suter
Marcus Raboy

Based on these graphs, here are some insights and recommendations for stakeholders

1. The majority of content on Netflix is movies, so it might be a good idea to focus on producing more movies to attract more customers.

2. The United States, India, and the United Kingdom are the top countries with the most content on Netflix, so Netflix could consider producing more content from these countries to cater to their audience.

3. The top directors with the most titles on Netflix are Raúl Campos, Jan Suter, and Marcus Raboy. Netflix could collaborate more with these directors to produce more content for their platform.

Task 4: Assignments Questions

a. Which are the top 5 directors who produce most of the movies only?

```
import pandas as pd

df = pd.read_excel(r'asbl_data_analyst_interview_assignment_netflix.xlsx')

# Filter for movies only
movies = df[df['Type'] == 'Movie']

# Group by director and count number of movies
director_counts = movies.groupby('Director')['Title'].count().reset_index()

# Sort by count of movies and select top 5
top_directors = director_counts.sort_values('Title', ascending=False).head(5)

print(top_directors)
```

```
                    Director  Title
3252           Rajiv Chilaka     19
3303  Raúl Campos, Jan Suter     18
3885             Suhas Kadav     16
2492            Marcus Raboy     15
1716              Jay Karas      14
```

b. Which are the top 5 genres which are liked by people or here liking means listed on the portal of Netflix (you can find a count for each genre and list the top 5 genres) for movies and TV shows?

```
import pandas as pd

df = pd.read_excel(r'asbl_data_analyst_interview_assignment_netflix.xlsx')

# Group by genre and count number of titles
genre_counts = df.groupby('Genres')['Title'].count().reset_index()

# Sort by count of titles and select top 5
top_genres = genre_counts.sort_values('Title', ascending=False).head(5)

print(top_genres)
```

```
                                               Genres  Title
326                      Dramas, International Movies    362
274                                      Documentaries    359
470                                    Stand-Up Comedy    334
200            Comedies, Dramas, International Movies    274
319  Dramas, Independent Movies, International Movies    252
```

c. Which 2 directors should Netflix collaborate with more based on the increase in their movies or tv shows over the past years?**

```
import pandas as pd

df = pd.read_excel(r'asbl_data_analyst_interview_assignment_netflix.xlsx')

# Filter for movies only
movies = df[df['Type'] == 'Movie']

# Group by director and release year, and count number of titles
director_year_counts = movies.groupby(['Director', 'Release_year'])['Title'].count

# Pivot the data so that each director has a row with columns for each release yea
pivot_table = director_year_counts.pivot(index='Director', columns='Release_year',

# Calculate the percentage increase in titles for each director between the first
percent_increase = (pivot_table.max(axis=1) - pivot_table.min(axis=1)) / pivot_tabl

# Select the top 2 directors with the highest percentage increase
top_directors = percent_increase.nlargest(2)

print(top_directors)
```

```
    Director
    A. L. Vijay     inf
    A. Raajdheep    inf
    dtype: float64
```

d. Which are the top 10 actors who are liked by people and have the most content on the Netflix OTT platform.

```
actor_count = df.groupby('Cast')['Title'].count()
top_actors = actor_count.sort_values(ascending=False)[:10]
print(top_actors)
```

```
    Cast
    David Attenborough
    Vatsal Dubey, Julie Tejwani, Rupa Bhimani, Jigna Bhardwaj, Rajesh Kava, Mousa
    Samuel West
    Jeff Dunham
    Kevin Hart
    Craig Sechler
    Michela Luci, Jamie Watson, Eric Peterson, Anna Claire Bartlam, Nicolas Aqui,
    David Spade, London Hughes, Fortune Feimster
    Jim Gaffigan
    Bill Burr
    Name: Title, dtype: int64
```

e. Which 2 actors should Netflix collaborate with more based on the increase in their movies or tv shows over the past years?

```
actor_year_count = df.groupby(['Cast', 'Release_year'])['Title'].count().reset_inde
actor_year_diff = actor_year_count.groupby('Cast')['Release_year'].apply(lambda x:
top_collaborators = actor_year_diff.sort_values(ascending=False)[:2]
print(top_collaborators)
```

```
Cast
Bob Ross        30
Sam Kinison     29
Name: Release_year, dtype: int64
```

✓ 0s completed at 23:29 ● ×