Assignment 3

Course Name : Programming in Python

Course Code : 1010043230

Name : Kushwaha Chirag Singh Devendra Singh

Enrollment No.: 2301031800049

Division : C / C1

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import os

# Create folder for graphs
if not os.path.exists("visuals"):
    os.makedirs("visuals")

# Load the dataset
df = pd.read_csv("netflix_titles.csv")

# Display first 5 rows
print(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...
```

```
# Check for null values
print("\nMissing values:\n", df.isnull().sum())

# Fill missing values for simplicity
df.fillna("Unknown", inplace=True)

# Check basic stats
print("\nData Info:")
print(df.info())
```

```
Missing values:
 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

Data 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      8807 non-null   object
 4   cast          8807 non-null   object
 5   country       8807 non-null   object
 6   date_added    8807 non-null   object
```

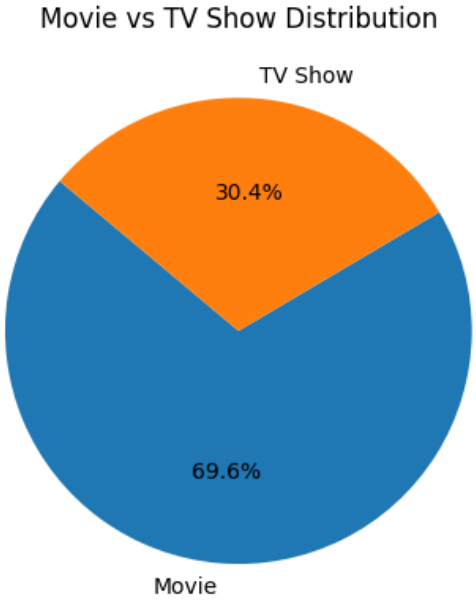
```
 7   release_year  8807 non-null   int64
 8   rating        8807 non-null   object
 9   duration      8807 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: int64(1), object(11)
memory usage: 825.8+ KB
None
```

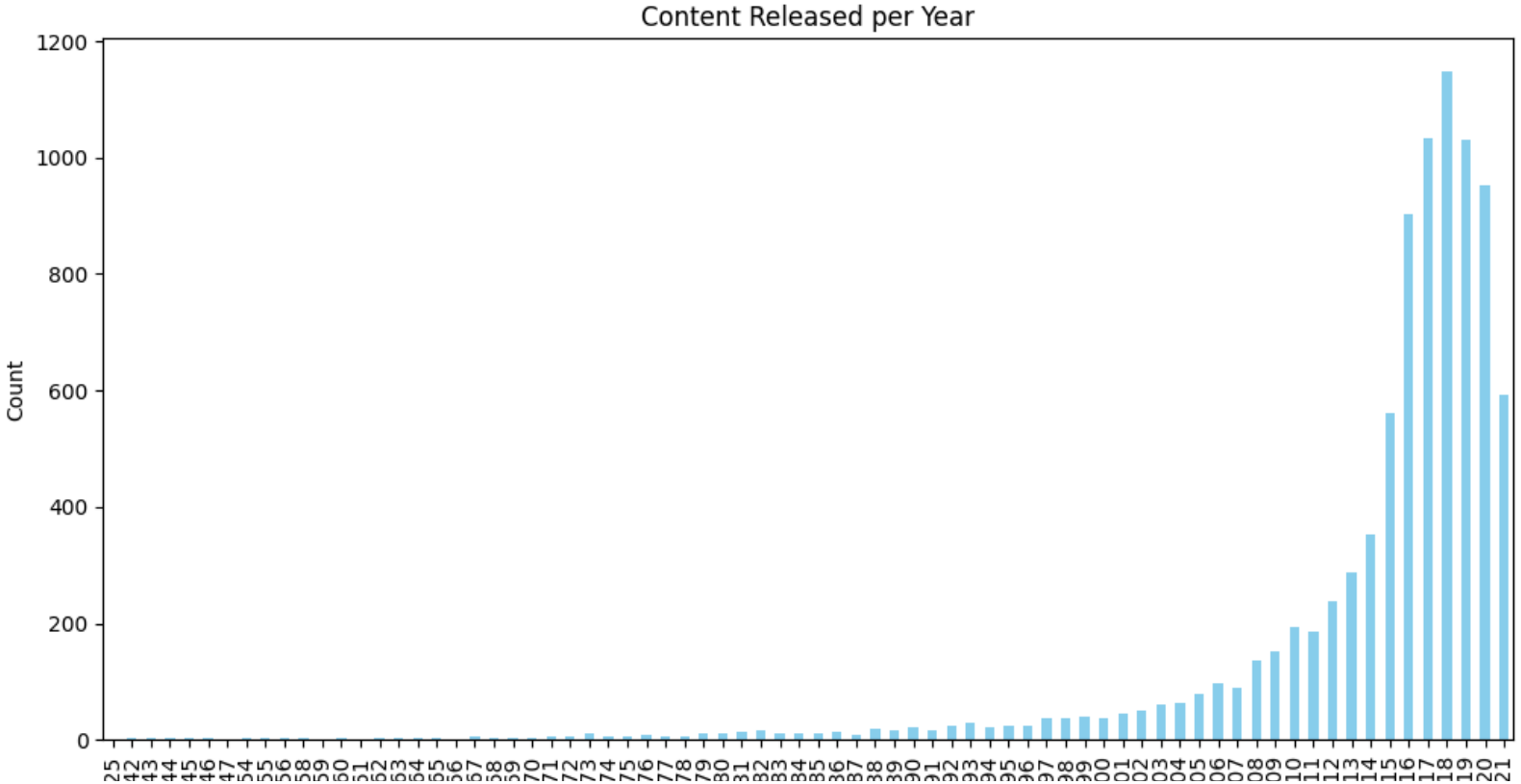
```
# Count movies vs TV shows
type_counts = df['type'].value_counts()
print("\nType Counts:\n", type_counts)

# Content released per year
content_per_year = df['release_year'].value_counts().sort_index()

# Most frequent countries
top_countries = df['country'].value_counts().head(10)

# Average duration of Movies
movie_durations = df[df['type'] == 'Movie']['duration'].str.replace(' min', '').replace("Unknown", np.nan).dropna().astype(int)
print("\nAverage Movie Duration:", np.mean(movie_durations), "minutes")
```

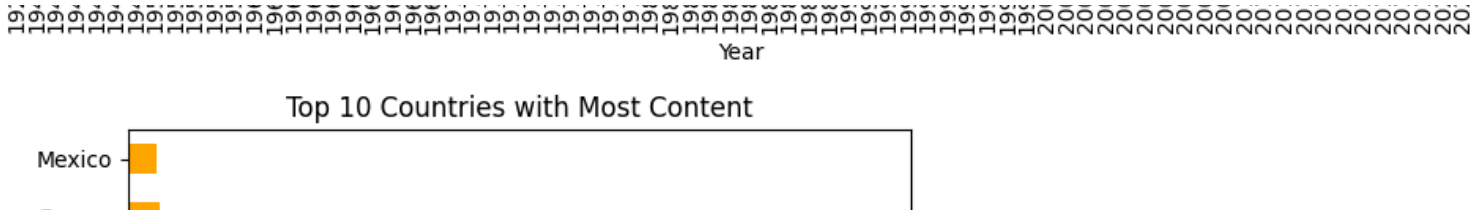
```
Type Counts:
 type
Movie      6131
TV Show    2676
Name: count, dtype: int64

Average Movie Duration: 99.57718668407311 minutes
```

```
# Plot: Type Distribution
type_counts.plot(kind='pie', autopct='%1.1f%%', startangle=140)
plt.title('Movie vs TV Show Distribution')
plt.ylabel("")
plt.savefig("visuals/type_distribution.png")
plt.show()
plt.clf()

# Plot: Content Released per Year
content_per_year.plot(kind='bar', figsize=(12, 6), color='skyblue')
plt.title('Content Released per Year')
plt.xlabel('Year')
plt.ylabel('Count')
plt.savefig("visuals/content_by_year.png")
plt.show()
plt.clf()

# Plot: Top 10 Countries
```

```
    # Plot: Top 10 Countries
    top_countries.plot(kind='barh', color='orange')
    plt.title('Top 10 Countries with Most Content')
    plt.xlabel('Count')
    plt.savefig("visuals/top_countries.png")
    plt.show()
    plt.clf()
```


Movie vs TV Show Distribution
TV Show
30.4%
69.6%
Movie


Content Released per Year
Count
1200
1000
800
600
400
200
0


Year
Top 10 Countries with Most Content
Mexico