Name: Maragathavalli c s

Data Science Internship Position at Prodigy Infotech

Task 1

Create a bar chart or histogram to visualize the distribution of a categorical or continuous variable, such as the distribution of ages or genders in a population

Importing necessary libraries

In [2]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Loading the Dataset

In [4]:
```
df=pd.read_excel(r"C:\Users\cskes\Downloads\P_Data_Extract_From_World_Development_Indicators.xlsx")
```

In [6]:
```
df.head()
```

Out[6]:

	Series Name	Series Code	Country Name	Country Code	1990 [YR1990]	2000 [YR2000]	2015 [YR2015]	2016 [YR2016]	2017 [YR2017]	2018 [YR2018]	2019 [YR2019]	2020 [YR2020]	[YR
0	Population, total	SP.POP.TOTL	Afghanistan	AFG	12045660	20130327	33831764	34700612	35688935	36743039	37856121	39068979	400
1	Population, total	SP.POP.TOTL	Albania	ALB	3286542	3089027	2880703	2876101	2873457	2866376	2854191	2837849	28
2	Population, total	SP.POP.TOTL	Algeria	DZA	25375810	30903893	40019529	40850721	41689299	42505035	43294546	44042091	447
3	Population, total	SP.POP.TOTL	American Samoa	ASM	46640	56855	52878	52245	51586	50908	50209	49761	
4	Population, total	SP.POP.TOTL	Andorra	AND	52597	65685	72174	72181	73763	75162	76474	77380	

Selecting and cleaning relevant columns

In [25]:
```
df_2023=df[['Country Name', '2023 [YR2023]']].copy()
df_2023.columns=['Country','Population_2023']
```

In [27]:
```
df_2023
```

Out[27]:

	Country	Population_2023
0	Afghanistan	41454761
1	Albania	2745972
2	Algeria	46164219
3	American Samoa	47521
4	Andorra	80856
...	...	...
266	NaN	NaN
267	NaN	NaN
268	NaN	NaN
269	NaN	NaN
270	NaN	NaN

271 rows × 2 columns

Convert Population to numeric and drop missing values

In [32]:
```
df_2023['Population_2023']=pd.to_numeric(df_2023['Population_2023'],errors='coerce')
df_clean=df_2023.dropna()
```

In [36]:
```
df_clean
```

Out[36]:

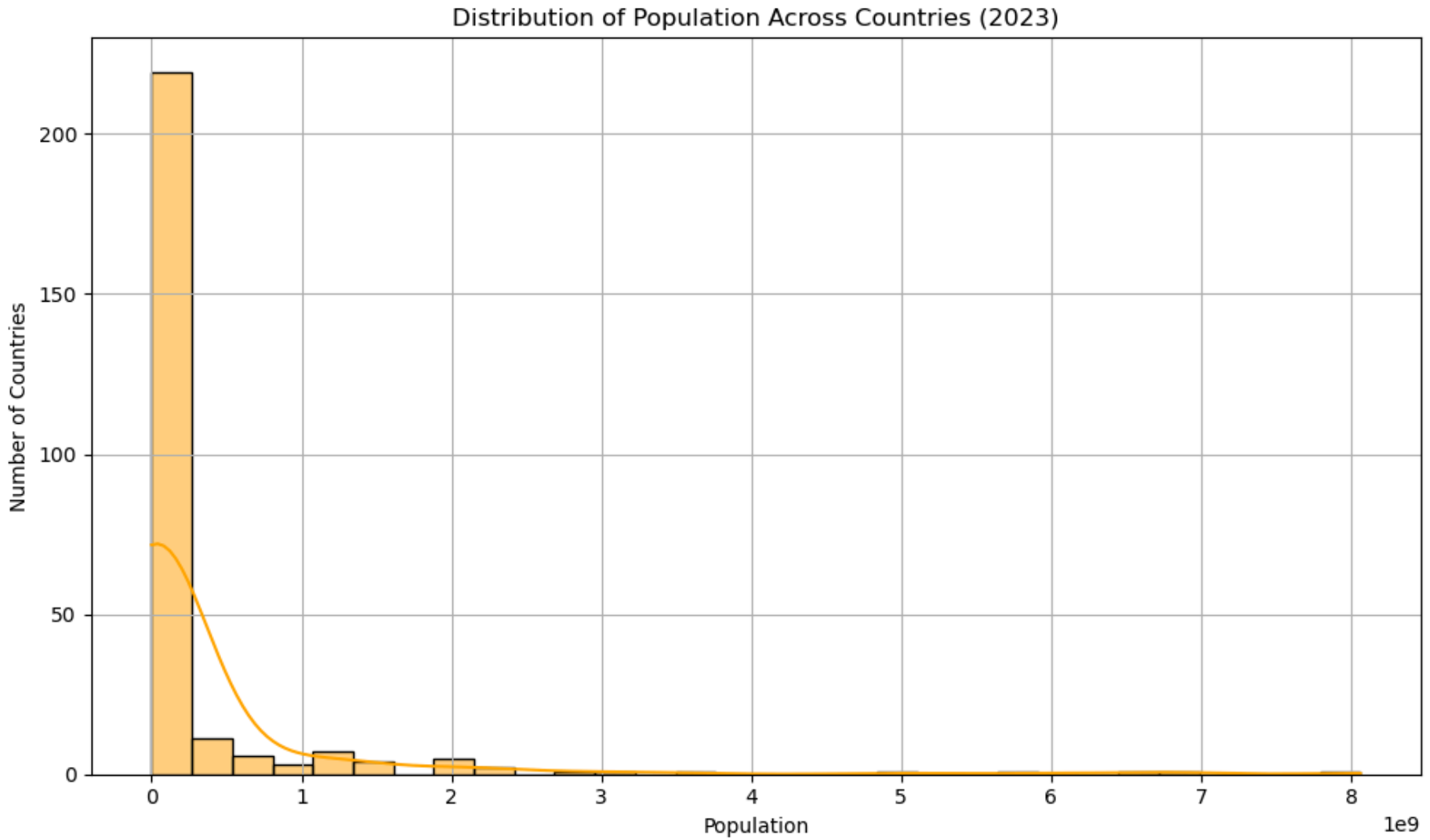
	Country	Population_2023
0	Afghanistan	4.145476e+07
1	Albania	2.745972e+06
2	Algeria	4.616422e+07
3	American Samoa	4.752100e+04
4	Andorra	8.085600e+04
...	...	...
261	Sub-Saharan Africa	1.259902e+09
262	Sub-Saharan Africa (excluding high income)	1.259783e+09
263	Sub-Saharan Africa (IDA & IBRD countries)	1.259902e+09
264	Upper middle income	2.816864e+09
265	World	8.061876e+09

265 rows × 2 columns

Histogram (Distribution of Population Across Countries(2023))

In [49]:

```
plt.figure(figsize=(10,6))
sns.histplot(df_clean['Population_2023'],bins=30,kde=True,color='orange')
plt.title("Distribution of Population Across Countries (2023)")
plt.xlabel('Population')
plt.ylabel('Number of Countries')
plt.grid(True)
plt.tight_layout()
plt.show()
```



Bar Chart (Top 10 countries by Population(2023))

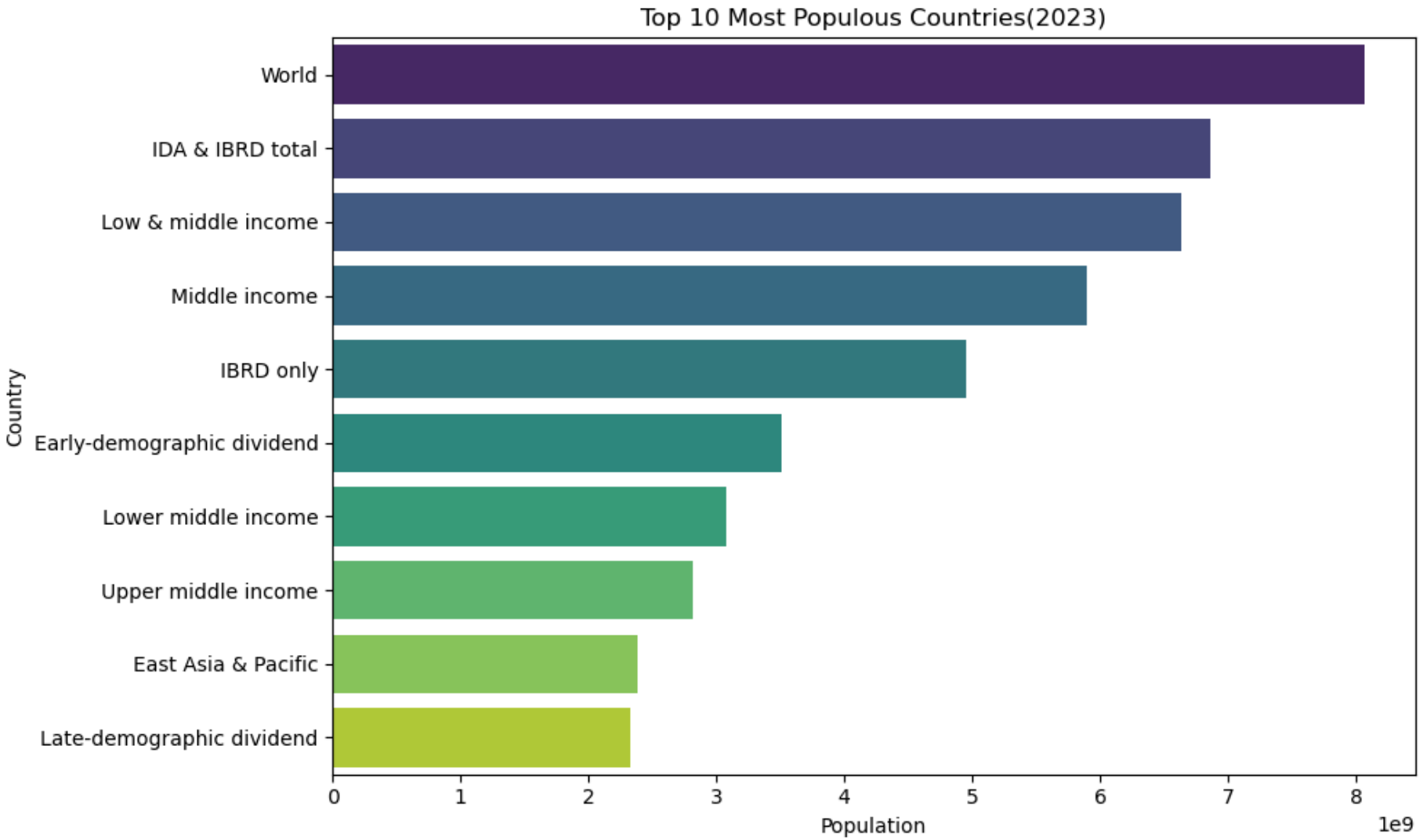
In [56]:
```
top_10=df_clean.sort_values(by='Population_2023',ascending=False).head(10)
```

In [58]:
```
plt.figure(figsize=(10,6))
sns.barplot(x='Population_2023',y='Country',data=top_10,palette='viridis')
plt.title('Top 10 Most Populous Countries(2023)')
plt.xlabel('Population')
plt.ylabel('Country')
plt.tight_layout()
plt.show()
```

```
C:\Users\cskes\AppData\Local\Temp\ipykernel_18344\373236642.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and se
t `legend=False` for the same effect.

  sns.barplot(x='Population_2023',y='Country',data=top_10,palette='viridis')
```


Top 10 Most Populous Countries(2023)
World
IDA & IBRD total
Low & middle income
Middle income
IBRD only
Early-demographic dividend
Lower middle income
Upper middle income
East Asia & Pacific
Late-demographic dividend
Country
0
1
2
3
4
5
6
7
8
Population
1e9