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

sample dataset: https://www.kaggle.com/datasets/sanjanchaudhari/population-dataset/download?datasetVersionNumber=1

In [1]:

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

In [2]:

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

In [3]:

```
df.head(5)
```

Out[3]:

	CCA3	Name	year 2022	year 2020	year 2015	year 2010	year 2000	year 1990	year 1980	year 1970	Area(sqkm)	Density (persqkm)	GrowthRate	W Popul Perce
0	CN	China	1425887	1424930	1393715	1348191	1264099	1153704	982372	822534	Area (km²)	146.8933	1.00	1
1	IN	India	1417173	1396387	1322867	1240614	1059634	870452	696828	557501	Area (km²)	431.0675	1.01	1
2	US	United States	338290	335942	324608	311183	282399	248084	223140	200328	Area (km²)	36.0935	1.00	4
3	ID	Indonesia	275501	271858	259092	244016	214072	182160	148177	115228	Area (km²)	144.6529	1.01	3
4	PK	Pakistan	235825	227197	210969	194454	154370	115414	80624	59291	Area (km²)	267.4018	1.02	2

In [4]:

```
df.describe()
```

Out[4]:

	year 2022	year 2020	year 2015	year 2010	year 2000	year 1990	year 1980	year 1970	Density (persqkm)
count	2.340000e+02	2.340000e+02	2.340000e+02	2.340000e+02	2.340000e+02	2.340000e+02	234.000000	234.000000	234.000000
mean	3.407441e+04	3.350109e+04	3.172995e+04	2.984523e+04	2.626947e+04	2.271024e+04	18984.645299	15786.876068	452.127044
std	1.367664e+05	1.355899e+05	1.304050e+05	1.242185e+05	1.116982e+05	9.783216e+04	81785.136077	67795.064322	2066.121904
min	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000	1.000000	0.026100
25%	4.197500e+02	4.150000e+02	4.045000e+02	3.930000e+02	3.272500e+02	2.642500e+02	229.500000	155.750000	38.417875
50%	5.560000e+03	5.493000e+03	5.307000e+03	4.943000e+03	4.293000e+03	3.825500e+03	3141.000000	2604.500000	95.346750
75%	2.247675e+04	2.144825e+04	1.973075e+04	1.915950e+04	1.576225e+04	1.186950e+04	9826.000000	8817.500000	238.933250
max	1.425887e+06	1.424930e+06	1.393715e+06	1.348191e+06	1.264099e+06	1.153704e+06	982372.000000	822534.000000	23172.266700

In [5]:

```
df.shape
```

Out[5]:

```
(234, 15)
```

In [6]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 234 entries, 0 to 233
Data columns (total 15 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   CCA3                         233 non-null    object 
 1   Name                         234 non-null    object 
 2   year 2022                    234 non-null    int64  
 3   year 2020                    234 non-null    int64  
 4   year 2015                    234 non-null    int64  
 5   year 2010                    234 non-null    int64  
 6   year 2000                    234 non-null    int64  
 7   year 1990                    234 non-null    int64  
 8   year 1980                    234 non-null    int64  
 9   year 1970                    234 non-null    int64  
 10  Area(sqkm)                   234 non-null    object 
 11  Density (persqkm)            234 non-null    float64
 12  GrowthRate                   234 non-null    float64
 13  World Population Percentage  234 non-null    object 
 14  Rank                         234 non-null    int64  
dtypes: float64(2), int64(9), object(4)
memory usage: 27.6+ KB
```

```
In [7]: df.isnull().sum()
```

```
Out[7]: CCA3                          1
        Name                          0
        year 2022                     0
        year 2020                     0
        year 2015                     0
        year 2010                     0
        year 2000                     0
        year 1990                     0
        year 1980                     0
        year 1970                     0
        Area(sqkm)                    0
        Density (persqkm)             0
        GrowthRate                    0
        World Population Percentage   0
        Rank                          0
        dtype: int64
```

```
In [8]: df['GrowthRate'].unique()
```

```
Out[8]: array([1.  , 1.01, 1.02, 0.99, 1.03, 1.04, 0.91, 0.98, 1.07])
```

```
In [9]: df.isna()
```

Out[9]:

	CCA3	Name	year 2022	year 2020	year 2015	year 2010	year 2000	year 1990	year 1980	year 1970	Area(sqkm)	Density (persqkm)	GrowthRate	World Population Percentage	Rank
0	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
229	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
230	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
231	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
232	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
233	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False

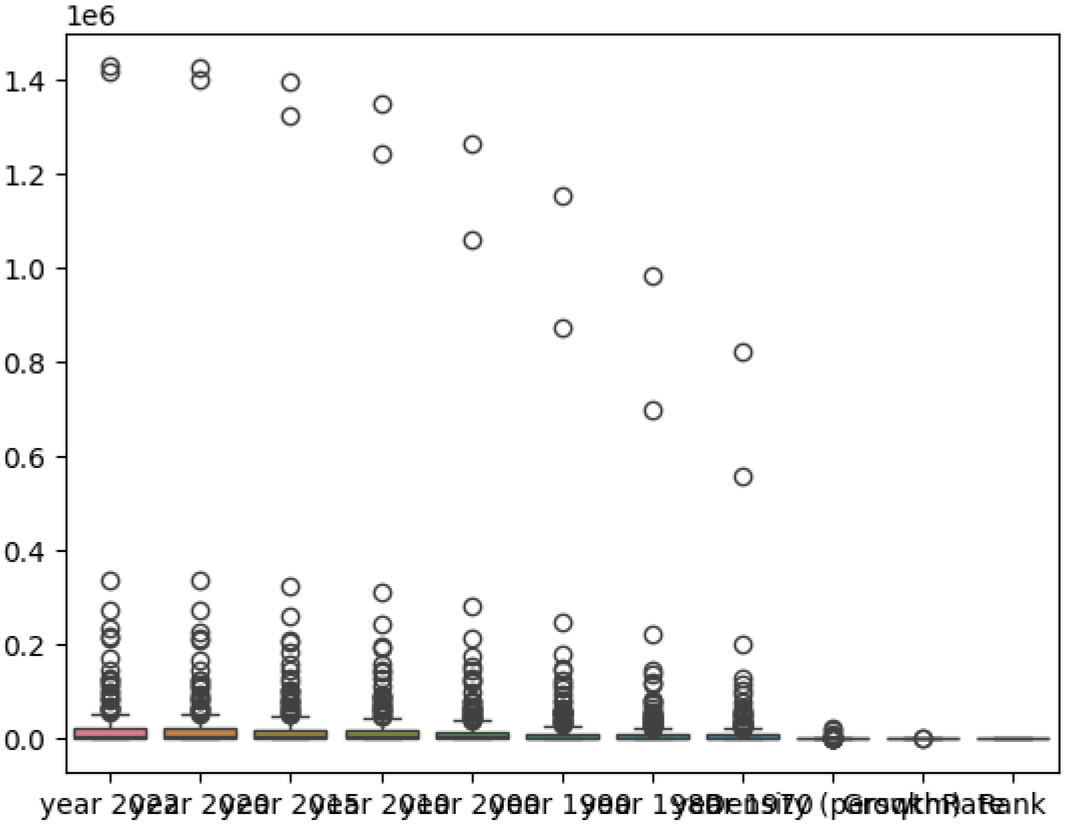
234 rows × 15 columns

In [10]:

```
sns.boxplot(data=df,palette='rainbow',)
```

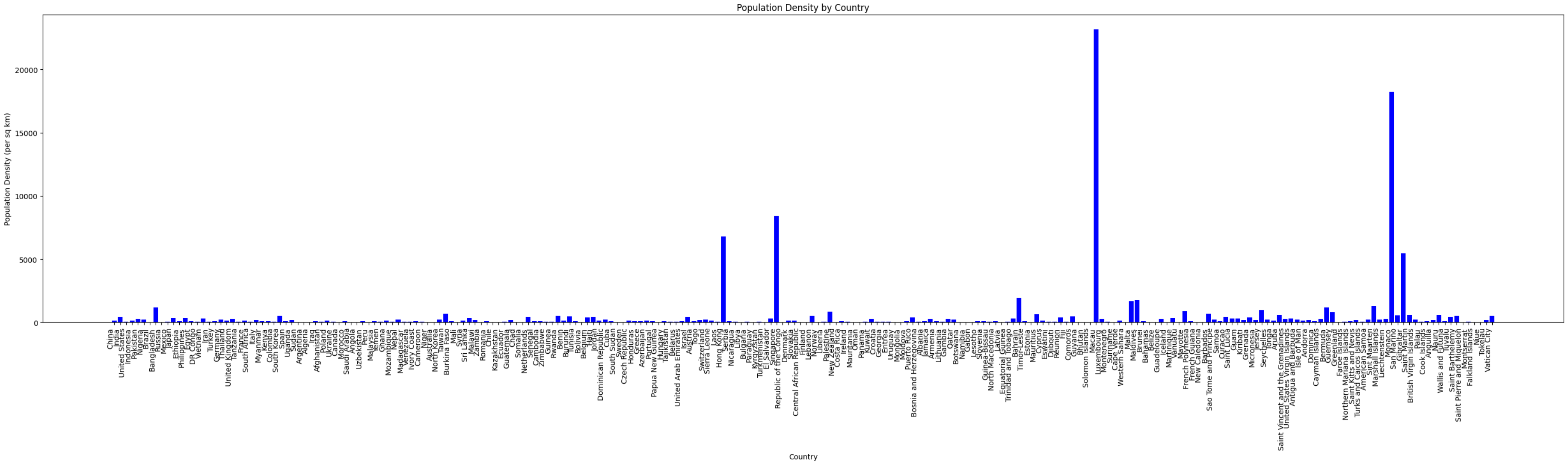
Out[10]:

```
<Axes: >
```



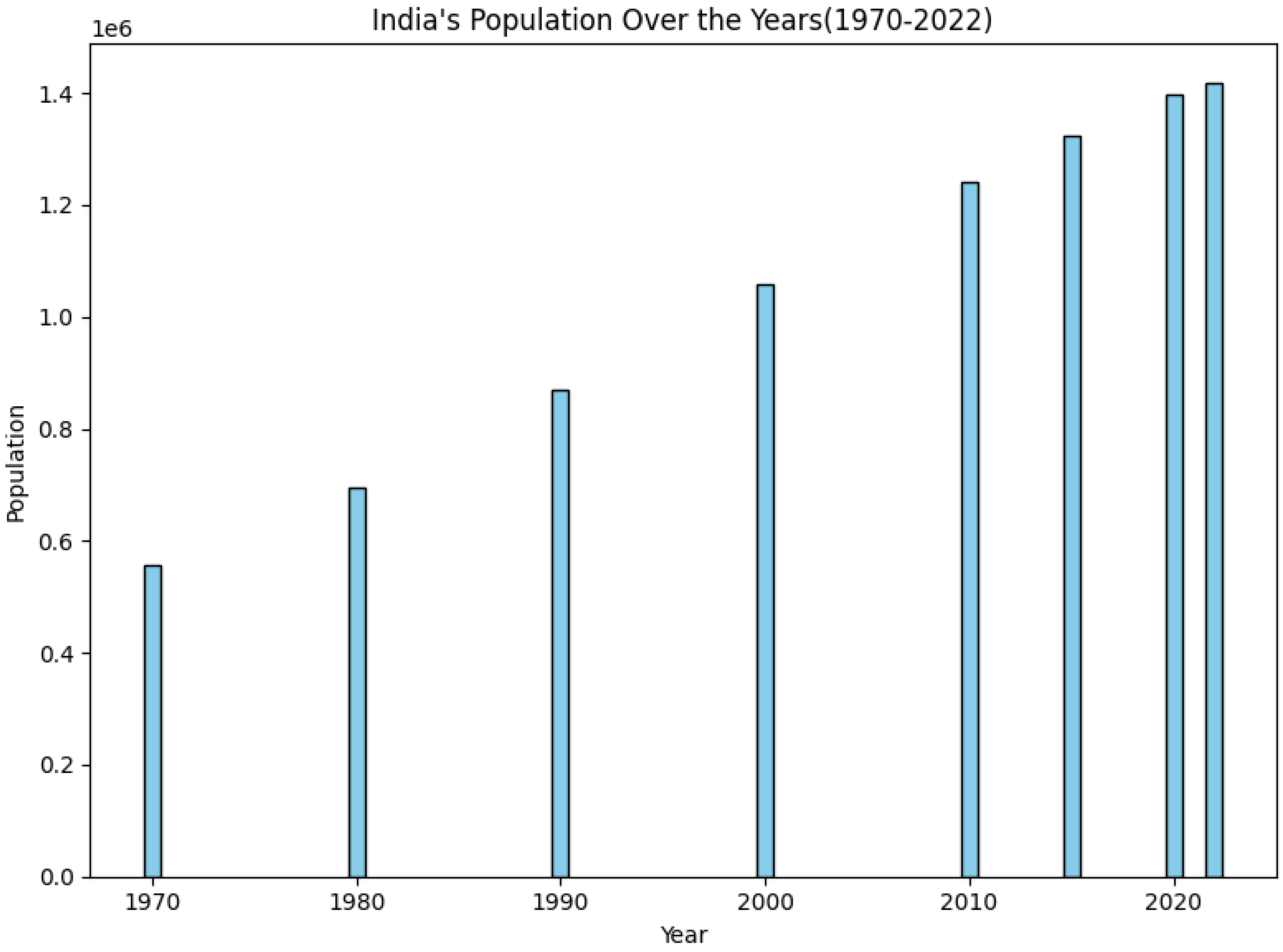
In [44]:

```
import matplotlib.pyplot as plt

plt.figure(figsize=(30, 9))
plt.bar(df['Name'], df['Density (persqkm)'], color='blue')
plt.title('Population Density by Country')
plt.xlabel('Country')
plt.ylabel('Population Density (per sq km)')
plt.xticks(rotation=90, ha='right')
plt.tight_layout()
plt.show()
```



In [55]:

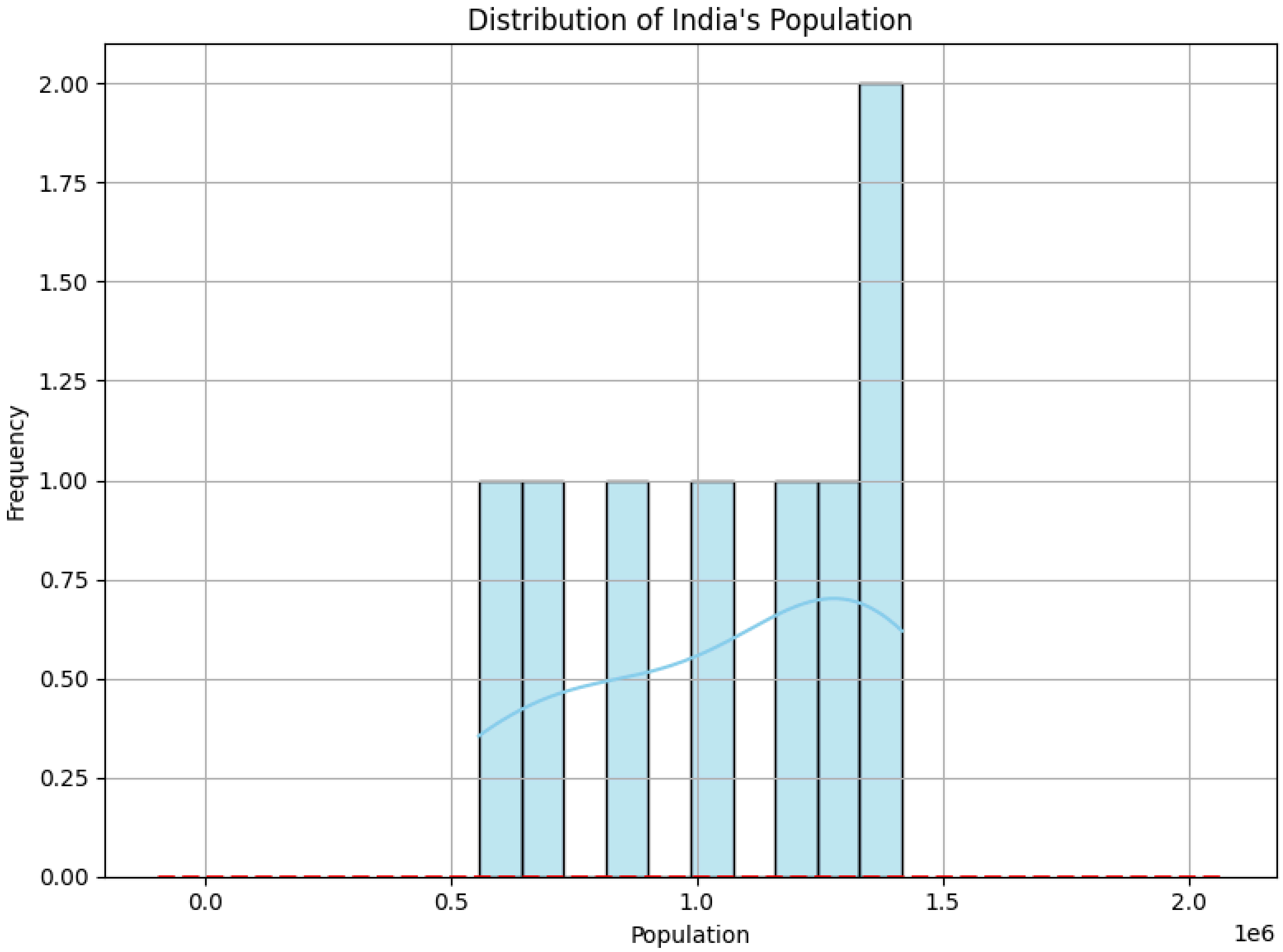
```
import matplotlib.pyplot as plt

data = {
    'Year': [2022, 2020, 2015, 2010, 2000, 1990, 1980, 1970],
    'Population': [1417173,1396387,1322867,1240614,1059634,870452,696828,557501]
}

data['Year'] = data['Year'][::-1]
data['Population'] = data['Population'][::-1]

# Plotting India's population over the years
plt.figure(figsize=(8, 6))
plt.bar(data['Year'], data['Population'], color='skyblue',edgecolor='black')
plt.title("India's Population Over the Years(1970-2022)")
plt.xlabel('Year')
plt.ylabel('Population')
plt.xticks(rotation=0)
plt.tight_layout()
plt.show()
```



In [76]:

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

# Sample population dataset with India's population for multiple years
```

```
data = {
    'Year': [2022, 2020, 2015, 2010, 2000, 1990, 1980, 1970],
    'Population': [1417173,1396387,1322867,1240614,1059634,870452,696828,557501]
}

data['Year'] = data['Year'][::-1]
data['Population'] = data['Population'][::-1]

df = pd.DataFrame(data)

plt.figure(figsize=(8, 6))

sns.histplot(data=df, x='Population', bins=10, kde=True, color='skyblue')

# Adding titles and labels
plt.title('Distribution of India\'s Population')
plt.xlabel('Population')
plt.ylabel('Frequency')

sns.kdeplot(data=df, x='Population', color='red', linewidth=2, linestyle='--')

plt.grid(True)
plt.tight_layout()
plt.show()
```


Distribution of India's Population
Frequency
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
1.5
2.0
Population
1e6

```
In [ ]:
```