

Exploratory Data Analysis on World Population (1970-2022)

```
In [417]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)
```

```
In [377]: data = pd.read_csv("C:/Users/zahid/Downloads/WorldPopulation2023 EDA/world_p
```

```
In [378]: data
```

Out[378]:

	Rank	CCA3	Country/Territory	Capital	Continent	2022 Population	2020 Population	2015 Population	Po
0	36	AFG	Afghanistan	Kabul	Asia	41128771	38972230	33753499	281
1	138	ALB	Albania	Tirana	Europe	2842321	2866849	2882481	29
2	34	DZA	Algeria	Algiers	Africa	44903225	43451666	39543154	358
3	213	ASM	American Samoa	Pago Pago	Oceania	44273	46189	51368	1
4	203	AND	Andorra	Andorra la Vella	Europe	79824	77700	71746	1
...
229	226	WLF	Wallis and Futuna	Mata-Utu	Oceania	11572	11655	12182	1
230	172	ESH	Western Sahara	El Aaiún	Africa	575986	556048	491824	1
231	46	YEM	Yemen	Sanaa	Asia	33696614	32284046	28516545	281
232	63	ZMB	Zambia	Lusaka	Africa	20017675	18927715	16248230	1
233	74	ZWE	Zimbabwe	Harare	Africa	16320537	15669666	14154937	1

234 rows × 17 columns

```
In [379]: data.head(5)
```

```
Out[379]:
```

	Rank	CCA3	Country/Territory	Capital	Continent	2022 Population	2020 Population	2015 Population	Popu
0	36	AFG	Afghanistan	Kabul	Asia	41128771	38972230	33753499	281
1	138	ALB	Albania	Tirana	Europe	2842321	2866849	2882481	29
2	34	DZA	Algeria	Algiers	Africa	44903225	43451666	39543154	358
3	213	ASM	American Samoa	Pago Pago	Oceania	44273	46189	51368	1
4	203	AND	Andorra	Andorra la Vella	Europe	79824	77700	71746	1

```
In [380]: data.tail(5)
```

Out[380]:

	Rank	CCA3	Country/Territory	Capital	Continent	2022 Population	2020 Population	2015 Population	Pop
229	226	WLF	Wallis and Futuna	Mata-Utu	Oceania	11572	11655	12182	
230	172	ESH	Western Sahara	El Aaiún	Africa	575986	556048	491824	
231	46	YEM	Yemen	Sanaa	Asia	33696614	32284046	28516545	24
232	63	ZMB	Zambia	Lusaka	Africa	20017675	18927715	16248230	13
233	74	ZWE	Zimbabwe	Harare	Africa	16320537	15669666	14154937	12

◀ ▶

```
In [385]: # Continent Data
```

```
continent_data = data.groupby(by='Continent').sum()
continent_data.head(10)
```

Out[385]:

Continent	Rank	2022 Population	2020 Population	2015 Population	2010 Population	2000 Population	1990 Population	I
Africa	5253	1426730932	1360671810	1201102442	1055228072	818946032	638150629	
Asia	3878	4721383274	4663086535	4458250182	4220041327	3735089604	3210563577	2
Europe	6225	743147538	745792196	741535608	735613934	726093423	720320797	
North America	6437	600296136	594236593	570383850	542720651	486069584	421266425	
Oceania	4336	45038554	43933426	40403283	37102764	31222778	26743822	
South America	1366	436816608	431530043	413134396	393078250	349634282	297146415	

◀ ▶

```
In [386]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 234 entries, 0 to 233
Data columns (total 17 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Rank             234 non-null    int64  
 1   CCA3            234 non-null    object 
 2   Country/Territory  234 non-null  object 
 3   Capital          234 non-null    object 
 4   Continent        234 non-null    object 
 5   2022 Population  234 non-null    int64  
 6   2020 Population  234 non-null    int64  
 7   2015 Population  234 non-null    int64  
 8   2010 Population  234 non-null    int64  
 9   2000 Population  234 non-null    int64  
 10  1990 Population  234 non-null    int64  
 11  1980 Population  234 non-null    int64  
 12  1970 Population  234 non-null    int64  
 13  Area (km²)      234 non-null    int64  
 14  Density (per km²) 234 non-null  float64 
 15  Growth Rate     234 non-null    float64 
 16  World Population Percentage 234 non-null  float64 
dtypes: float64(3), int64(10), object(4)
memory usage: 31.2+ KB
```

```
In [387]: data.dtypes
```

```
Out[387]: Rank          int64
CC43           object
Country/Territory    object
Capital          object
Continent         object
2022 Population   int64
2020 Population   int64
2015 Population   int64
2010 Population   int64
2000 Population   int64
1990 Population   int64
1980 Population   int64
1970 Population   int64
Area (km²)        int64
Density (per km²) float64
Growth Rate       float64
World Population Percentage float64
dtype: object
```

```
In [388]: data.describe()
```

```
Out[388]:
```

	Rank	2022 Population	2020 Population	2015 Population	2010 Population	2000 Population	P
count	234.000000	2.340000e+02	2.340000e+02	2.340000e+02	2.340000e+02	2.340000e+02	2.34
mean	117.500000	3.407441e+07	3.350107e+07	3.172996e+07	2.984524e+07	2.626947e+07	2.27
std	67.694165	1.367664e+08	1.355899e+08	1.304050e+08	1.242185e+08	1.116982e+08	9.78
min	1.000000	5.100000e+02	5.200000e+02	5.640000e+02	5.960000e+02	6.510000e+02	7.00
25%	59.250000	4.197385e+05	4.152845e+05	4.046760e+05	3.931490e+05	3.272420e+05	2.64
50%	117.500000	5.559944e+06	5.493074e+06	5.307400e+06	4.942770e+06	4.292907e+06	3.82
75%	175.750000	2.247650e+07	2.144798e+07	1.973085e+07	1.915957e+07	1.576230e+07	1.18
max	234.000000	1.425887e+09	1.424930e+09	1.393715e+09	1.348191e+09	1.264099e+09	1.15



```
In [389]: data.describe(include="object")
```

```
Out[389]:
```

	CC43	Country/Territory	Capital	Continent
count	234	234	234	234
unique	234	234	234	6
top	AFG	Afghanistan	Kabul	Africa
freq	1	1	1	57

```
In [390]: data.describe(include="all")
```

Out[390]:

	Rank	CCA3	Country/Territory	Capital	Continent	2022 Population	2020 Population	2015 Population	2010 Population	2000 Population	1990 Population	1980 Population	1970 Population	Area (km²)	Density (per km²)	Growth Rate	World Population Percentage
count	234.000000	234		234	234	234	2.340000e+02	2.340000e+02	2.340000e+02	2.340000e+02	2.340000e+02						
unique		Nan	234		234	234	6	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan
top		Nan	AFG	Afghanistan	Kabul	Africa	Nan	Nan	Nan	Nan	Nan						
freq		Nan	1		1	57	Nan	Nan	Nan	Nan	Nan						
mean	117.500000	Nan		Nan	Nan	NaN	3.407441e+07	3.350107e+07	3.300000e+07	3.250000e+07	3.200000e+07	3.150000e+07	3.100000e+07	3.050000e+07	3.000000e+07	2.950000e+07	2.900000e+07
std	67.694165	Nan		Nan	Nan	NaN	1.367664e+08	1.355899e+08	1.350000e+08	1.345000e+08	1.340000e+08	1.335000e+08	1.330000e+08	1.325000e+08	1.320000e+08	1.315000e+08	1.310000e+08
min	1.000000	Nan		Nan	Nan	NaN	5.100000e+02	5.200000e+02	5.300000e+02	5.400000e+02	5.500000e+02	5.600000e+02	5.700000e+02	5.800000e+02	5.900000e+02	6.000000e+02	6.100000e+02
25%	59.250000	Nan		Nan	Nan	NaN	4.197385e+05	4.152845e+05	4.150000e+05	4.145000e+05	4.140000e+05	4.135000e+05	4.130000e+05	4.125000e+05	4.120000e+05	4.115000e+05	4.110000e+05
50%	117.500000	Nan		Nan	Nan	NaN	5.559944e+06	5.493074e+06	5.490000e+06	5.485000e+06	5.480000e+06	5.475000e+06	5.470000e+06	5.465000e+06	5.460000e+06	5.455000e+06	5.450000e+06
75%	175.750000	Nan		Nan	Nan	NaN	2.247650e+07	2.144798e+07	2.140000e+07	2.135000e+07	2.130000e+07	2.125000e+07	2.120000e+07	2.115000e+07	2.110000e+07	2.105000e+07	2.100000e+07
max	234.000000	Nan		Nan	Nan	NaN	1.425887e+09	1.424930e+09	1.424000e+09	1.423000e+09	1.422000e+09	1.421000e+09	1.420000e+09	1.419000e+09	1.418000e+09	1.417000e+09	1.416000e+09



```
In [391]: data.shape
```

Out[391]: (234, 17)

```
In [392]: data.index
```

Out[392]: RangeIndex(start=0, stop=234, step=1)

```
In [393]: data.columns
```

```
Out[393]: Index(['Rank', 'CCA3', 'Country/Territory', 'Capital', 'Continent', '2022 Population', '2020 Population', '2015 Population', '2010 Population', '2000 Population', '1990 Population', '1980 Population', '1970 Population', 'Area (km²)', 'Density (per km²)', '²)', 'Growth Rate', 'World Population Percentage'], dtype='object')
```

In [394]: `data.nunique`

Out[394]: <bound method DataFrame.nunique of

	Rank	CCA3	Country/Territory			
0	36	AFG	Afghanistan	Kabul	Asia	
1	138	ALB	Albania	Tirana	Europe	
2	34	DZA	Algeria	Algiers	Africa	
3	213	ASM	American Samoa	Pago Pago	Oceania	
4	203	AND	Andorra	Andorra la Vella	Europe	
..	
229	226	WLF	Wallis and Futuna	Mata-Utu	Oceania	
230	172	ESH	Western Sahara	El Aaiún	Africa	
231	46	YEM	Yemen	Sanaa	Asia	
232	63	ZMB	Zambia	Lusaka	Africa	
233	74	ZWE	Zimbabwe	Harare	Africa	
	2022	Population	2020 Population	2015 Population	2010 Population	\
0		41128771	38972230	33753499	28189672	
1		2842321	2866849	2882481	2913399	
2		44903225	43451666	39543154	35856344	
3		44273	46189	51368	54849	
4		79824	77700	71746	71519	
..	
229		11572	11655	12182	13142	
230		575986	556048	491824	413296	
231		33696614	32284046	28516545	24743946	
232		20017675	18927715	16248230	13792086	
233		16320537	15669666	14154937	12839771	
	2000	Population	1990 Population	1980 Population	1970 Population	\
0		19542982	10694796	12486631	10752971	
1		3182021	3295066	2941651	2324731	
2		30774621	25518074	18739378	13795915	
3		58230	47818	32886	27075	
4		66097	53569	35611	19860	
..	
229		14723	13454	11315	9377	
230		270375	178529	116775	76371	
231		18628700	13375121	9204938	6843607	
232		9891136	7686401	5720438	4281671	
233		11834676	10113893	7049926	5202918	
	Area (km ²)	Density (per km ²)	Growth Rate	World Population	Percentage	
0	652230	63.0587	1.0257		0.52	
1	28748	98.8702	0.9957		0.04	
2	2381741	18.8531	1.0164		0.56	
3	199	222.4774	0.9831		0.00	
4	468	170.5641	1.0100		0.00	
..	
229	142	81.4930	0.9953		0.00	
230	266000	2.1654	1.0184		0.01	
231	527968	63.8232	1.0217		0.42	
232	752612	26.5976	1.0280		0.25	
233	390757	41.7665	1.0204		0.20	

[234 rows x 17 columns]>

```
In [395]: data.unique()
```

```
Out[395]: Rank           234  
CC43            234  
Country/Territory  234  
Capital          234  
Continent         6  
2022 Population   234  
2020 Population   234  
2015 Population   234  
2010 Population   234  
2000 Population   234  
1990 Population   234  
1980 Population   234  
1970 Population   234  
Area (km²)        233  
Density (per km²) 234  
Growth Rate       180  
World Population Percentage 70  
dtype: int64
```

```
In [396]: data.isnull().sum()
```

```
Out[396]: Rank           0  
CC43            0  
Country/Territory  0  
Capital          0  
Continent         0  
2022 Population   0  
2020 Population   0  
2015 Population   0  
2010 Population   0  
2000 Population   0  
1990 Population   0  
1980 Population   0  
1970 Population   0  
Area (km²)        0  
Density (per km²) 0  
Growth Rate       0  
World Population Percentage 0  
dtype: int64
```

```
In [397]: data["Capital"].describe()
```

```
Out[397]: count    234  
unique     234  
top       Kabul  
freq       1  
Name: Capital, dtype: object
```

Insights

```
In [398]: data.mean()
```

```
Out[398]: Rank           1.175000e+02  
2022 Population   3.407441e+07  
2020 Population   3.350107e+07  
2015 Population   3.172996e+07  
2010 Population   2.984524e+07  
2000 Population   2.626947e+07  
1990 Population   2.271022e+07  
1980 Population   1.898462e+07  
1970 Population   1.578691e+07  
Area (km²)        5.814494e+05  
Density (per km²) 4.521270e+02  
Growth Rate       1.009577e+00  
World Population Percentage 4.270513e-01  
dtype: float64
```

```
In [399]: data.mode()
```

Out[399]:

	Rank	CCA3	Country/Territory	Capital	Continent	2022 Population	2020 Population	2018 Population
0	1	ABW	Afghanistan	Abu Dhabi	Africa	510	520	530
1	2	AFG	Albania	Abuja	NaN	1871	1827	1827
2	3	AGO	Algeria	Accra	NaN	1934	1942	1942
3	4	AIA	American Samoa	Addis Ababa	NaN	3780	3747	3747
4	5	ALB	Andorra	Algiers	NaN	4390	4500	4500
...
229	230	WSM	Wallis and Futuna	Yamoussoukro	NaN	235824862	227196741	210969
230	231	YEM	Western Sahara	Yaounde	NaN	275501339	271857970	259091
231	232	ZAF	Yemen	Yaren	NaN	338289857	335942003	324607
232	233	ZMB	Zambia	Yerevan	NaN	1417173173	1396387127	1322866
233	234	ZWE	Zimbabwe	Zagreb	NaN	1425887337	1424929781	1393715

234 rows × 17 columns



```
In [400]: data.median()
```

```
Out[400]: Rank                1.175000e+02
2022 Population        5.559944e+06
2020 Population        5.493074e+06
2015 Population         5.307400e+06
2010 Population         4.942770e+06
2000 Population         4.292907e+06
1990 Population         3.825410e+06
1980 Population         3.141146e+06
1970 Population         2.604830e+06
Area (km²)              8.119950e+04
Density (per km²)       9.534675e+01
Growth Rate              1.007900e+00
World Population Percentage 7.000000e-02
dtype: float64
```

```
In [401]: # Total World Population of 2022
data['2022 Population'].sum()
```

Out[401]: 7973413042

Missing / Duplicates Values

```
In [402]: #check missing values  
data.isnull().sum()
```

```
Out[402]: Rank          0  
CCA3           0  
Country/Territory  0  
Capital          0  
Continent         0  
2022 Population  0  
2020 Population  0  
2015 Population  0  
2010 Population  0  
2000 Population  0  
1990 Population  0  
1980 Population  0  
1970 Population  0  
Area (km²)        0  
Density (per km²) 0  
Growth Rate       0  
World Population Percentage 0  
dtype: int64
```

```
In [403]: # Duplicates Values  
data.duplicated().sum()
```

```
Out[403]: 0
```

```
In [404]: #rename the column 'Country/Territory' to 'Country'  
data.rename(columns={'Country/Territory':'Country'}, inplace = True)  
  
data.head()
```

```
Out[404]:
```

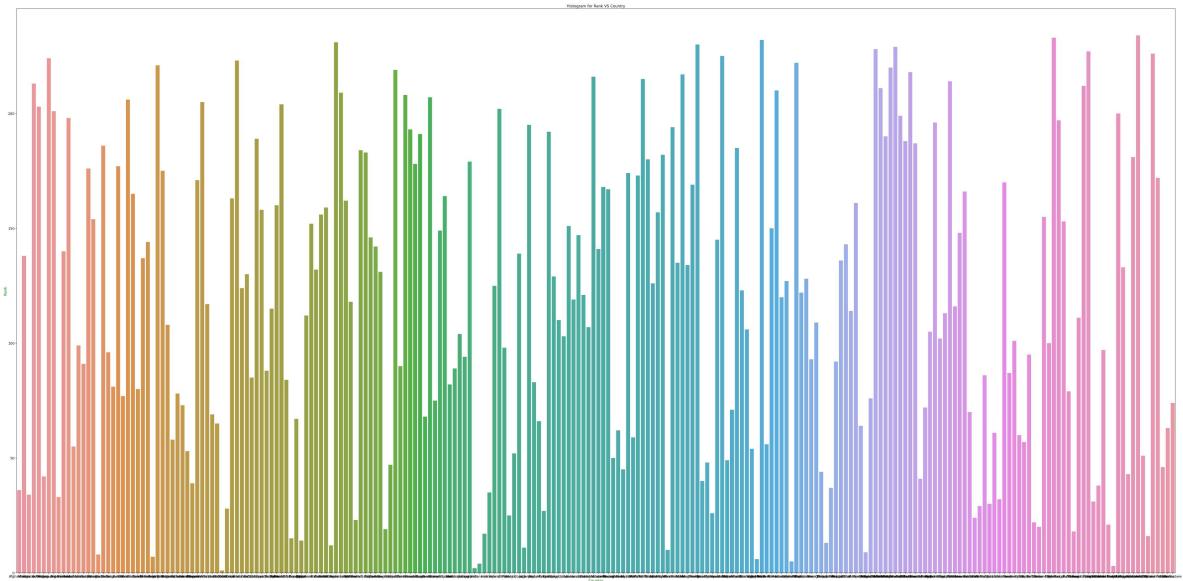
	Rank	CCA3	Country	Capital	Continent	2022 Population	2020 Population	2015 Population	2010 Population
0	36	AFG	Afghanistan	Kabul	Asia	41128771	38972230	33753499	28189672
1	138	ALB	Albania	Tirana	Europe	2842321	2866849	2882481	2913399
2	34	DZA	Algeria	Algiers	Africa	44903225	43451666	39543154	35856344
3	213	ASM	American Samoa	Pago Pago	Oceania	44273	46189	51368	54849
4	203	AND	Andorra	Andorra la Vella	Europe	79824	77700	71746	71519



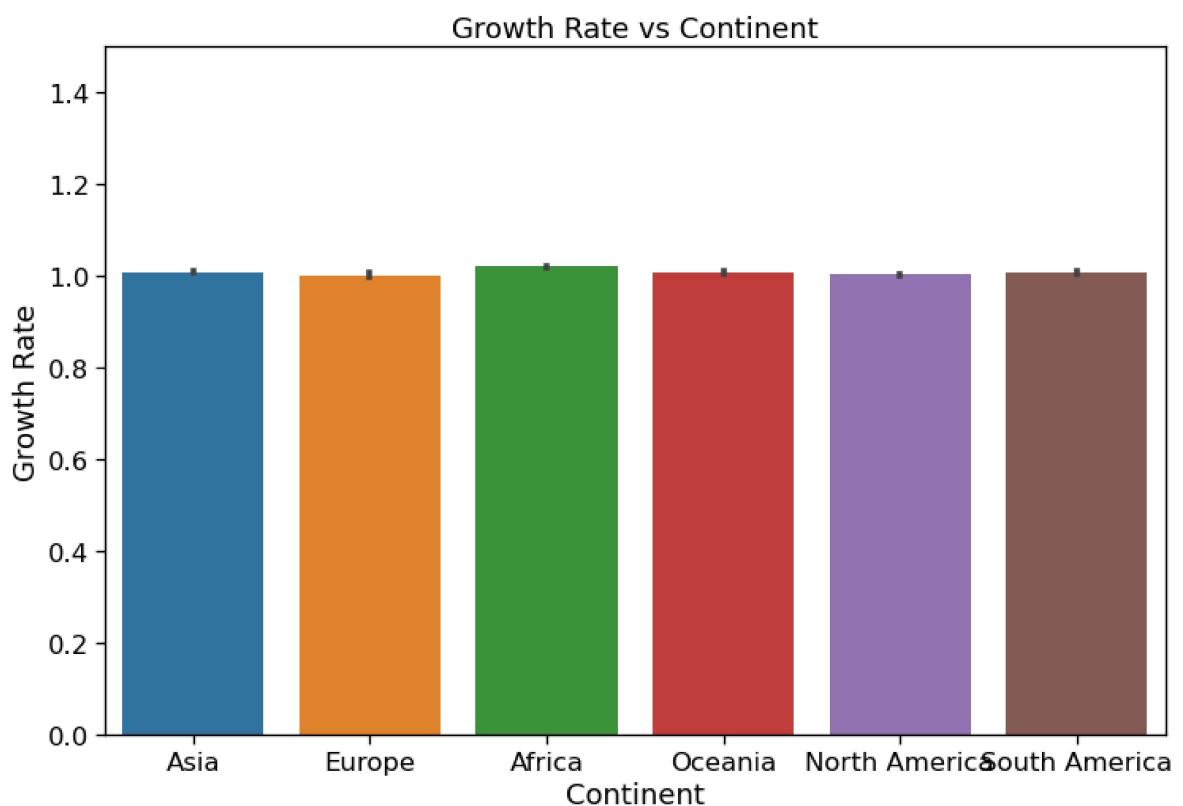
```
In [ ]:
```

Visualization

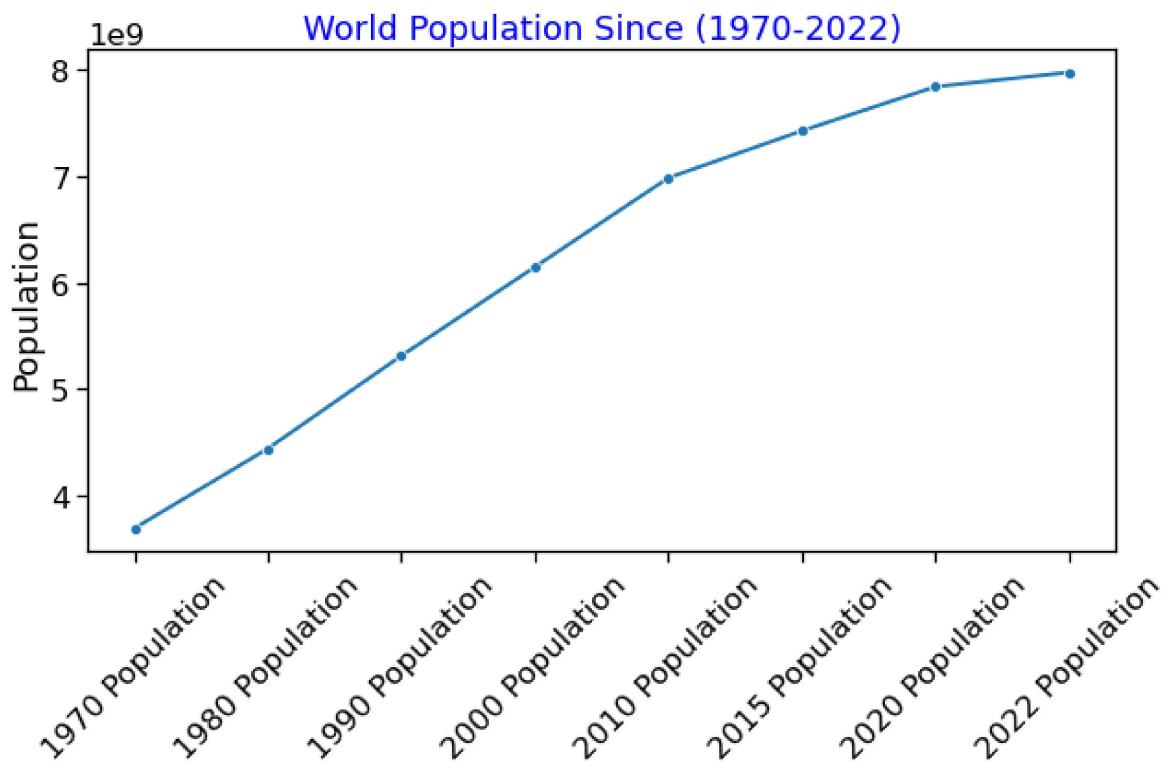
```
In [405]: plt.figure(figsize=(100,50))
plt.title("Histogram for Rank VS Country")
plt.xlabel("Country", color="green")
plt.ylabel("Rank" , color="green")
sns.barplot(data=data, x='Country', y="Rank")
plt.show()
```



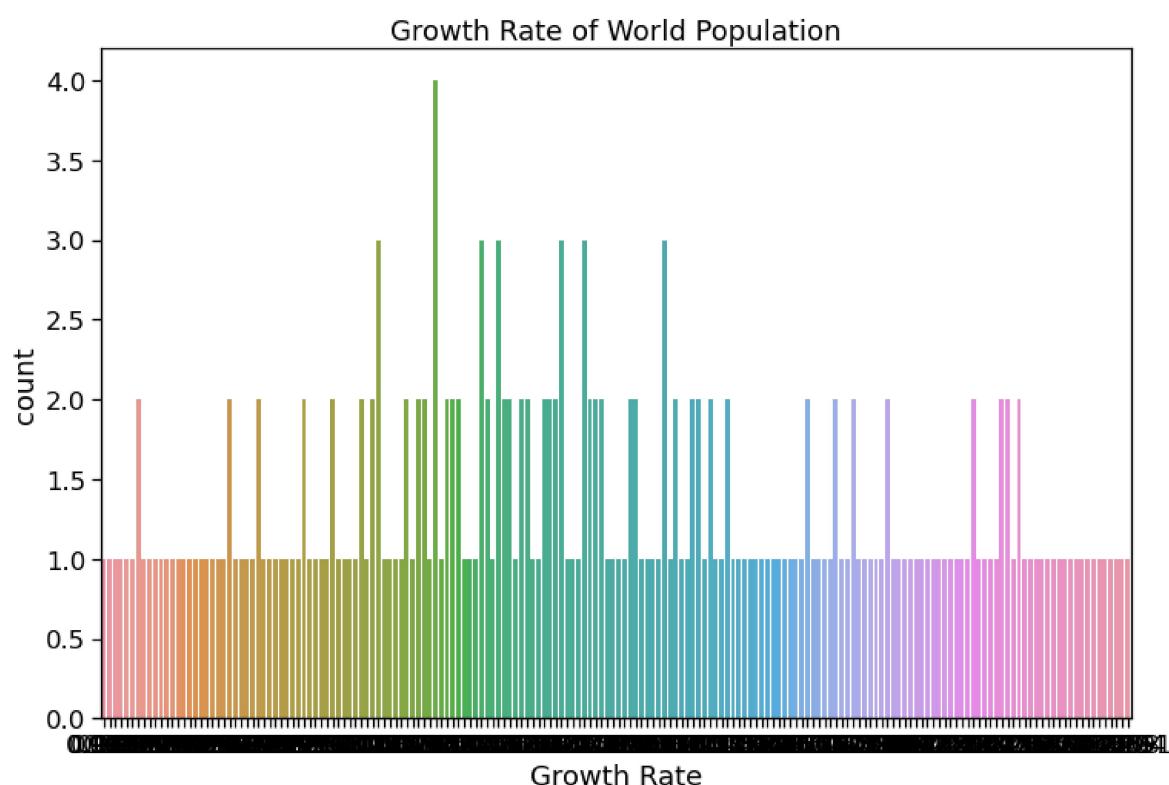
```
In [419]: plt.figure(figsize=(12,8))
plt.title("Growth Rate vs Continent")
# plt.hist(data=data, x='Growth Rate')
sns.barplot(data=data, y='Growth Rate', x = "Continent")
plt.ylim(0, 1.50)
plt.show()
```



```
In [407]: # plot World population 1970-2022
plt.subplots(figsize=(10,5))
Since = data.iloc[:,5:13].sum()[:-1]
sns.lineplot(x=Since.index, y=Since.values, marker="o")
plt.xticks(rotation=45)
plt.ylabel("Population")
plt.title("World Population Since (1970-2022)", color='Blue')
plt.show()
```

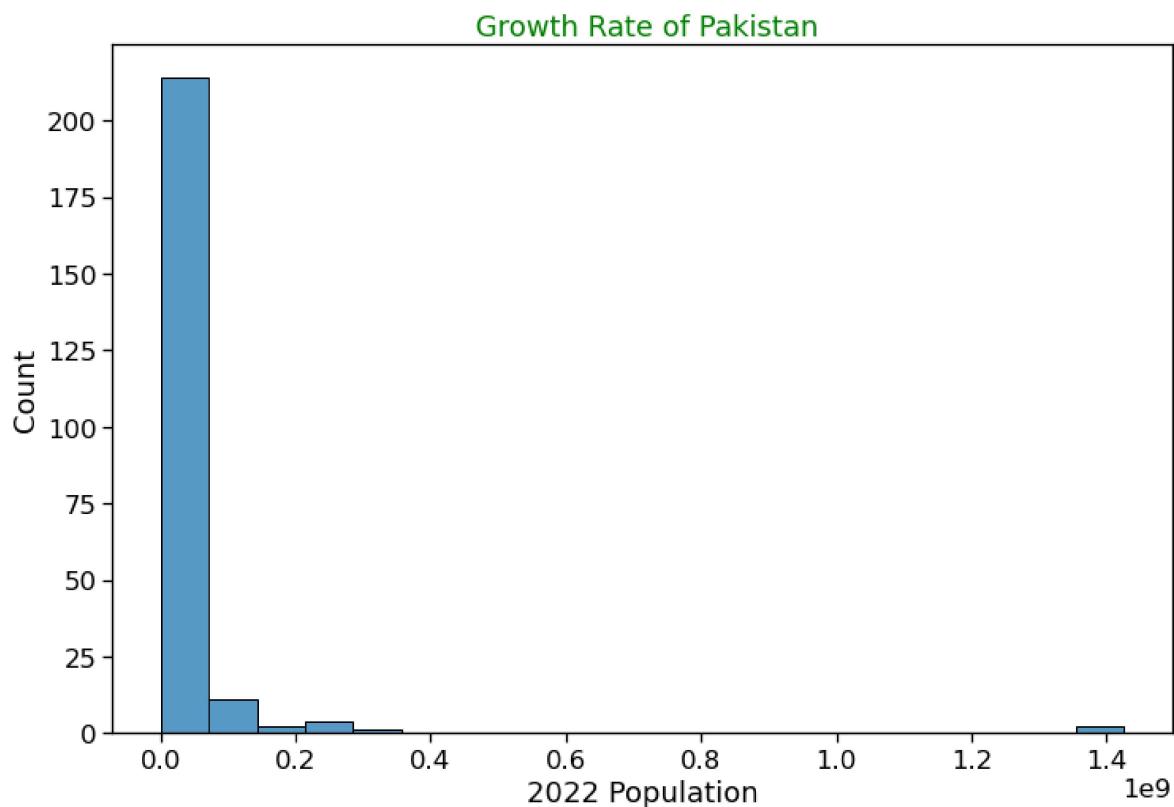


```
In [408]: plt.figure(figsize=(12,8))
plt.title("Growth Rate of World Population")
sns.countplot(data=data, x='Growth Rate')
# plt.ylim(0, 5_0)
plt.show()
```



```
In [409]: pakistan_data = data[data['Country'] == 'Pakistan']

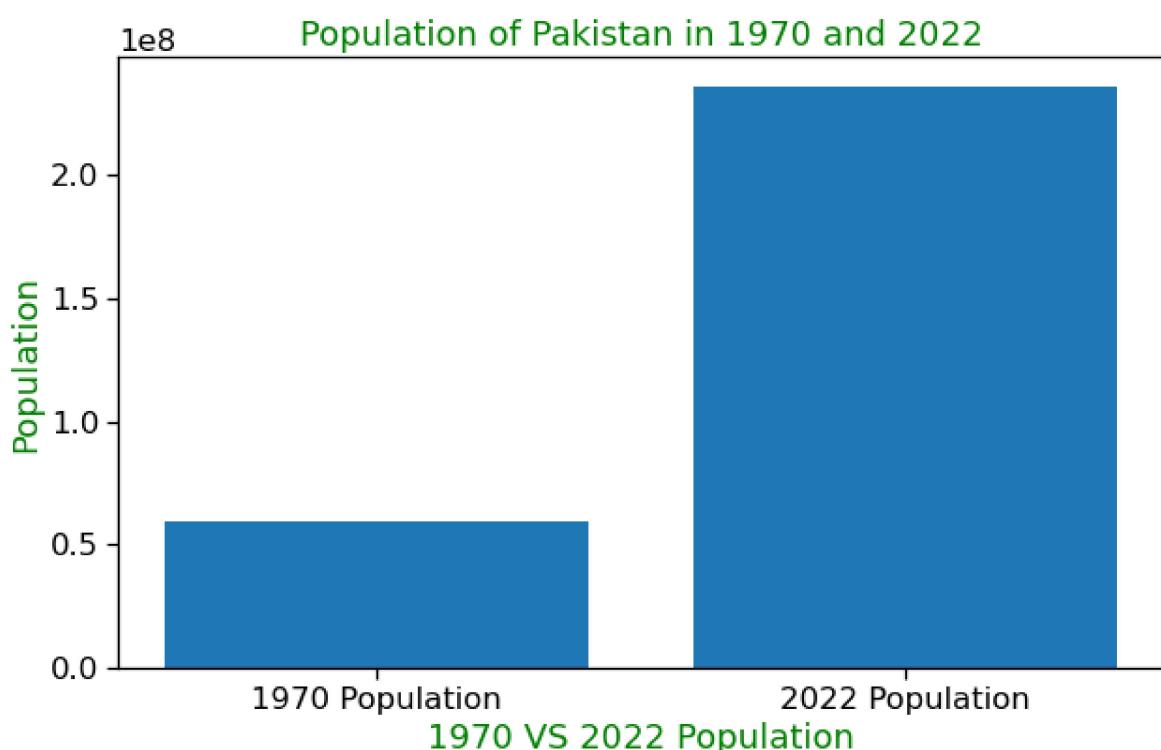
plt.figure(figsize=(12,8))
plt.title(" Growth Rate of Pakistan", color='green')
sns.histplot(data=data, x='2022 Population', bins=20)
plt.show()
```



```
In [410]: pakistan_data = data[data['Country'] == 'Pakistan']

pop_1970 = pakistan_data['1970 Population'].iloc[0]
pop_2022 = pakistan_data['2022 Population'].iloc[0]

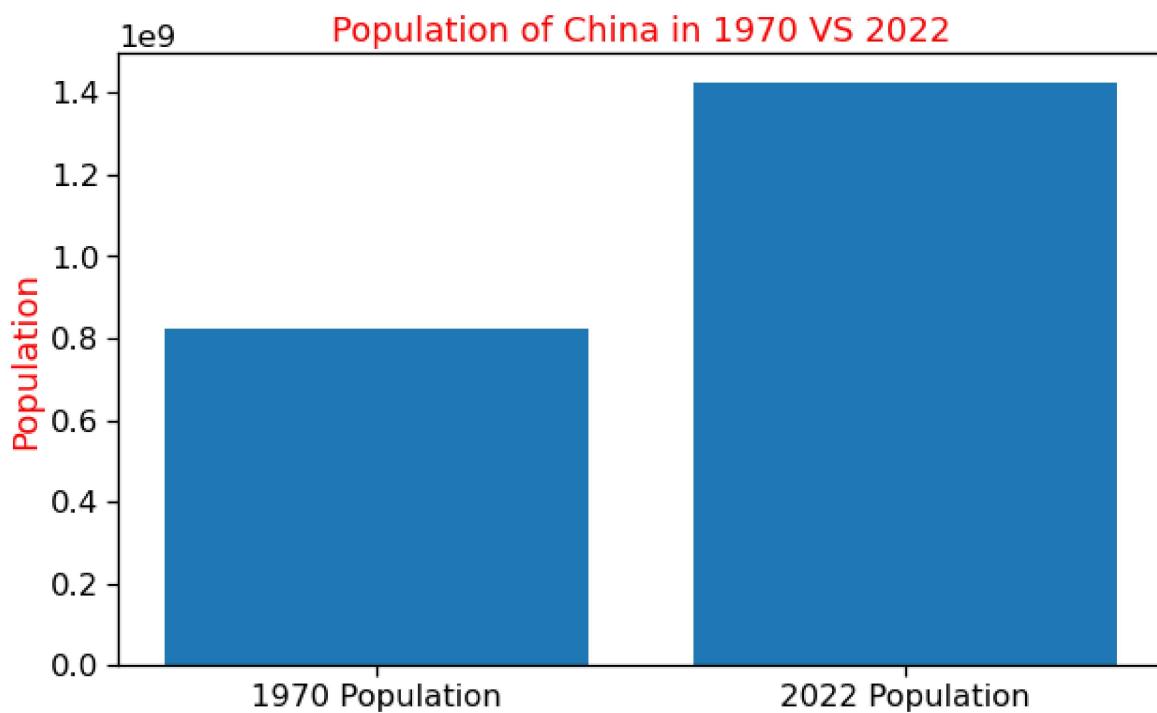
plt.figure(figsize=(10,6))
plt.bar(x=['1970 Population', '2022 Population'], height=[pop_1970, pop_2022])
plt.ylabel('Population', color = 'Green')
plt.xlabel('1970 VS 2022 Population', color = 'Green')
plt.title('Population of Pakistan in 1970 and 2022', color = 'Green')
plt.show()
```



```
In [411]: china_data = data[data['Country'] == 'China']

pop_1970 = china_data['1970 Population'].iloc[0]
pop_2022 = china_data['2022 Population'].iloc[0]

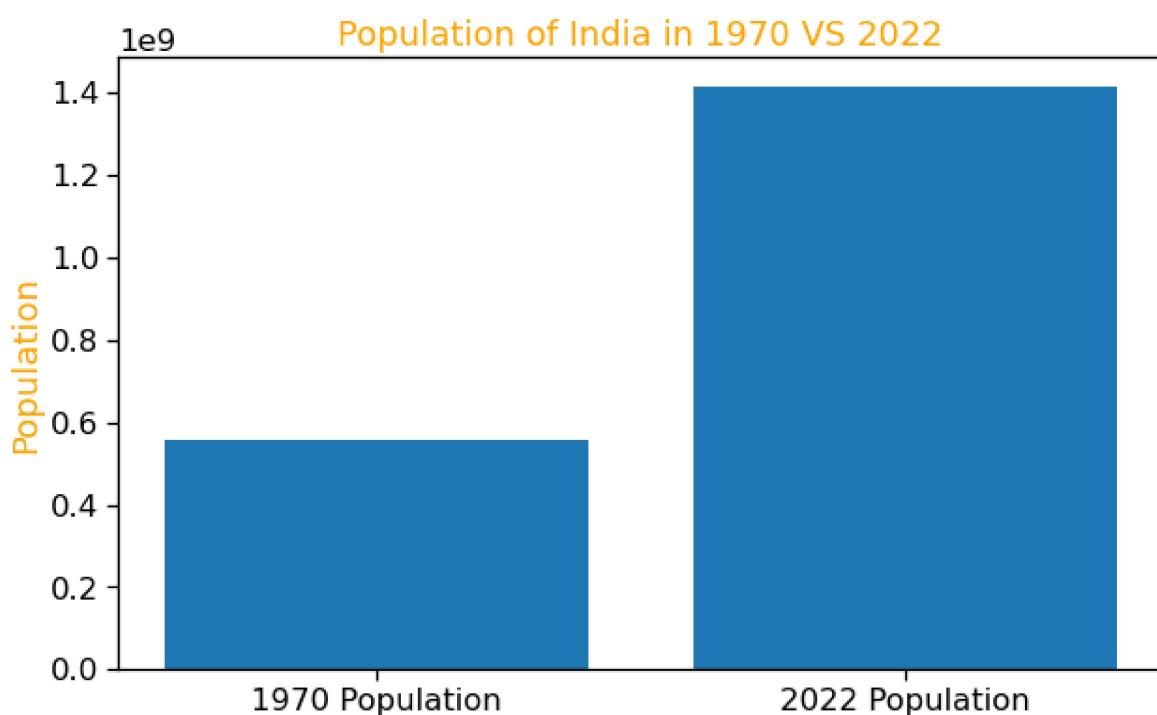
plt.figure(figsize=(10,6))
plt.bar(x=['1970 Population', '2022 Population'], height=[pop_1970, pop_2022])
plt.ylabel('Population' , color = 'red')
plt.title('Population of China in 1970 VS 2022' , color = 'Red')
plt.show()
```



```
In [412]: india_data = data[data['Country'] == 'India']

pop_1970 = india_data['1970 Population'].iloc[0]
pop_2022 = india_data['2022 Population'].iloc[0]

plt.figure(figsize=(10,6))
plt.bar(x=['1970 Population', '2022 Population'], height=[pop_1970, pop_2022])
plt.ylabel('Population' , color='orange')
plt.title('Population of India in 1970 VS 2022' , color='orange')
plt.show()
```



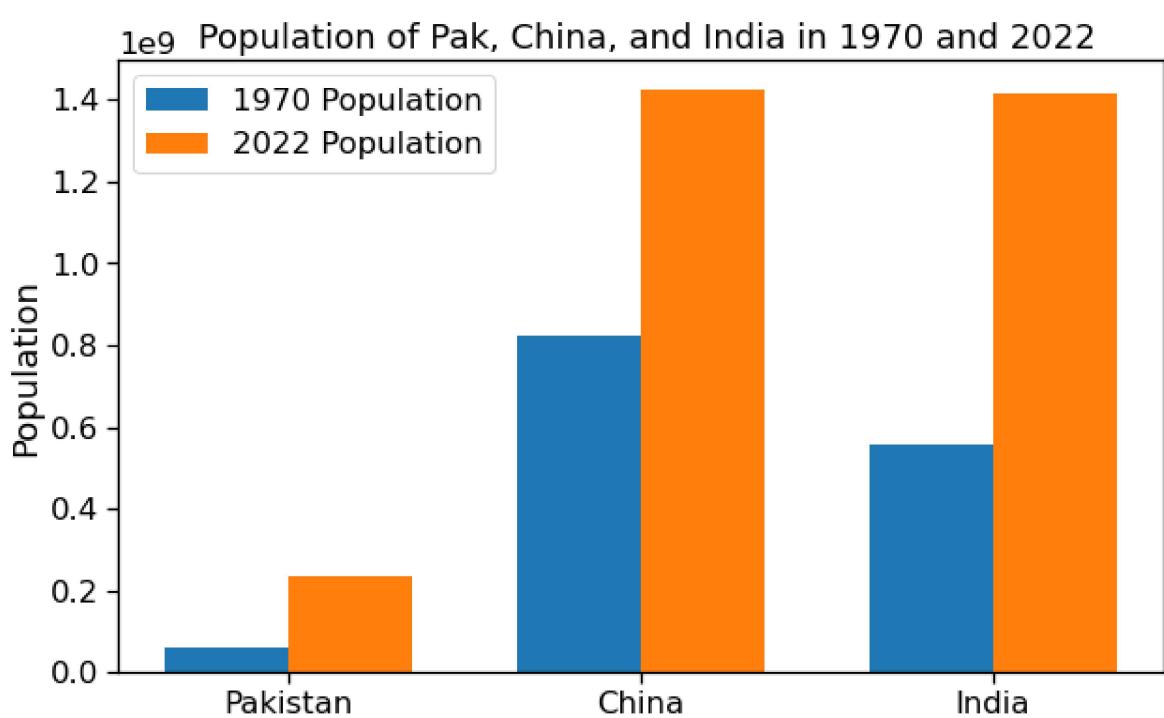
```
In [413]: pakistan_data = data[data['Country'] == 'Pakistan']
china_data = data[data['Country'] == 'China']
india_data = data[data['Country'] == 'India']

pop_1970_pakistan = pakistan_data['1970 Population'].iloc[0] # Pak Population
pop_2022_pakistan = pakistan_data['2022 Population'].iloc[0] # Pak Population

pop_1970_china = china_data['1970 Population'].iloc[0] # china Population of
pop_2022_china = china_data['2022 Population'].iloc[0] # china Population of

pop_1970_india = india_data['1970 Population'].iloc[0] # india Population of
pop_2022_india = india_data['2022 Population'].iloc[0] # india Population of

fig, ax = plt.subplots(figsize=(10,6))
bar_width = 0.35
ax.bar(x=np.arange(3)-bar_width/2, height=[pop_1970_pakistan, pop_1970_china,
ax.bar(x=np.arange(3)+bar_width/2, height=[pop_2022_pakistan, pop_2022_china,
ax.set_ylabel('Population')
ax.set_xticks(np.arange(3))
ax.set_xticklabels(['Pakistan', 'China', 'India'])
ax.set_title(' Population of Pak, China, and India in 1970 and 2022')
ax.legend()
plt.show()
```

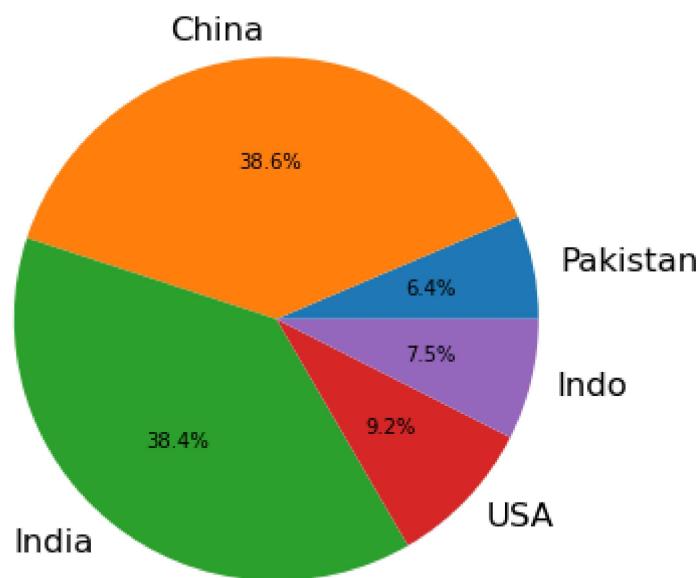


```
In [414]: pakistan_data = data[data['Country'] == 'Pakistan']
china_data = data[data['Country'] == 'China']
india_data = data[data['Country'] == 'India']
usa_data = data[data['Country'] == 'United States']
indo_data = data[data['Country'] == 'Indonesia']

# Population values for each country
pop_pakistan = pakistan_data['2022 Population'].iloc[0]
pop_china = china_data['2022 Population'].iloc[0]
pop_india = india_data['2022 Population'].iloc[0]
pop_usa = usa_data['2022 Population'].iloc[0]
pop_indo = indo_data['2022 Population'].iloc[0]

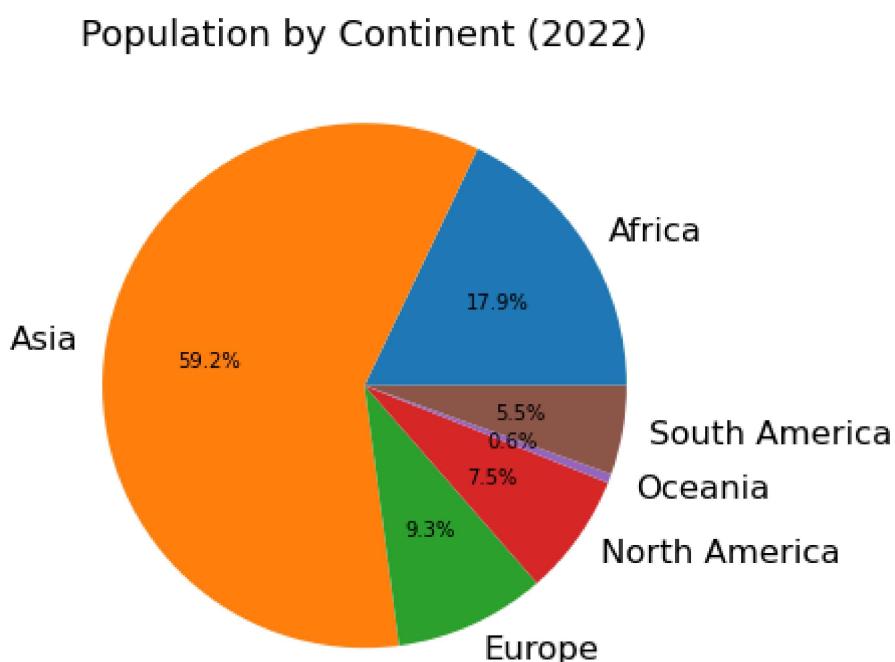
fig, ax = plt.subplots(figsize=(10,6))
labels = ['Pakistan', 'China', 'India', 'USA', 'Indo']
sizes = [pop_pakistan, pop_china, pop_india, pop_usa, pop_indo]
ax.pie(sizes, labels=labels, autopct='%1.1f%%')
ax.set_title('Top-5 Countries population in 2022')
plt.show()
```

Top-5 Countries population in 2022



```
In [415]: # Group the data by continent and sum the population values for each continent
continent_data = data.groupby('Continent').sum()

# Create a pie chart showing the population of each continent
fig, ax = plt.subplots(figsize=(8, 6))
ax.pie(continent_data['2022 Population'], labels=continent_data.index, autopct='%.1f%%')
ax.set_title('Population by Continent (2022)')
plt.show()
```

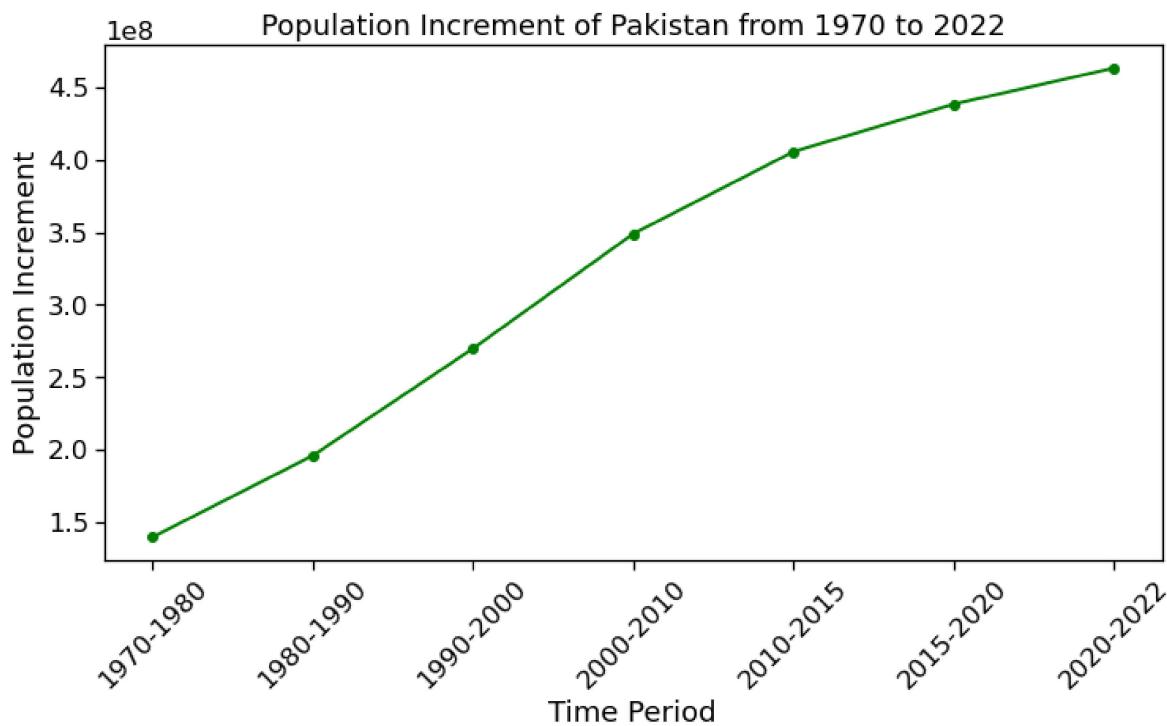


```
In [416]: pakistan_data = data[data['Country'] == 'Pakistan']

# Get the population values yearly
pop_1970 = pakistan_data['1970 Population'].iloc[0]
pop_1980 = pakistan_data['1980 Population'].iloc[0]
pop_1990 = pakistan_data['1990 Population'].iloc[0]
pop_2000 = pakistan_data['2000 Population'].iloc[0]
pop_2010 = pakistan_data['2010 Population'].iloc[0]
pop_2015 = pakistan_data['2015 Population'].iloc[0]
pop_2020 = pakistan_data['2020 Population'].iloc[0]
pop_2022 = pakistan_data['2022 Population'].iloc[0]

# Calculate the population incremented yearly
increment_1970_1980 = pop_1980 + pop_1970
increment_1980_1990 = pop_1990 + pop_1980
increment_1990_2000 = pop_2000 + pop_1990
increment_2000_2010 = pop_2010 + pop_2000
increment_2010_2015 = pop_2015 + pop_2010
increment_2015_2020 = pop_2020 + pop_2015
increment_2020_2022 = pop_2022 + pop_2020

plt.figure(figsize=(12,6))
plt.plot(['1970-1980', '1980-1990', '1990-2000', '2000-2010', '2010-2015', '2015-2020', '2020-2022'],
         [increment_1970_1980, increment_1980_1990, increment_1990_2000, increment_2000_2010,
          increment_2010_2015, increment_2015_2020, increment_2020_2022], marker='o')
plt.xlabel('Time Period')
plt.xticks(rotation=45)
plt.ylabel('Population Increment')
plt.title('Population Increment of Pakistan from 1970 to 2022')
plt.show()
```



Follow:

Github: <https://www.github.com/ZaidArman> (<https://www.github.com/ZaidArman>)

Linkedin: <https://www.linkedin.com/in/zaid-ullah07> (<https://www.linkedin.com/in/zaid-ullah07>)

Twitter: <https://www.twitter.com/ZaidArman> (<https://www.twitter.com/ZaidArman>)

Instagram: https://www.instagram.com/zaid_arman7
https://www.instagram.com/zaid_arman7

Facebook: <https://www.facebook.com/profile.php?id=100011010551170>
<https://www.facebook.com/profile.php?id=100011010551170>

