DATA EXPLORATION AND VISUALISATION
USING PYTHON - PART1

WEB SCRAPING AND ITS ANALYSIS

WEBSITE SCRAPED - WORLDOMETERS
INFO OF COUNTRIES IN THE WORLD BY POPULATION(2023)

BY-
RISHIKA ARYA
045046

PROJECT OBJECTIVES

- Population Analysis and Visualization: Analysing the distribution of the world's population across countries and exploring demographic factors such as median age, fertility rates, and net migration.

- Correlation Analysis: Investigating the correlations between different demographic factors to measure the degree of relatedness and change between two or more variables.

- Visualization: Creating visual representations such as bar graphs, histogram, pie charts and scatter plots to present the data findings effectively.

GENERAL DESCRIPTION OF THE DATA

The data has been scraped from the Worldometers.info – Countries in the world by population(2023) https://www.worldometers.info/world-population/population-by-country/ which includes the following data :

- Country (or Dependency): This variable contains the names of 234 countries or dependent territories. It serves as the primary identifier for each observation in the dataset.

- Population (2023): This variable provides the estimated population of each country for the year 2023. It represents the total number of individuals living in the country.

- Yearly Change: This variable indicates the annual percentage change in population for each country. It reflects the population growth or decline over time.

- Net Change: This variable represents the net change in population, typically calculated as the difference between births and deaths, along with other factors like migration.

- Density (P/Km²): This variable indicates the population density, often measured as the number of people per square kilometer of land area. It helps assess how crowded or sparsely populated a country is.

- Land Area (Km²): This variable provides the land area of each country in square kilometer. It's essential for calculating the land area and understanding a country's geographical size.

- Migrants (net): This variable represents the net migration rate, indicating the balance between people immigrating to and emigrating from a country. It can affect population growth.

- Fertility Rate: This variable shows the fertility rate, which is the average number of children born to a woman during her lifetime. It's a key demographic indicator.

- Median Age: This variable represents the median age of a country's population, providing insights into the age distribution.

- Urban Pop %: This variable indicates the percentage of a country's population living in urban areas. It reflects urbanization trends.

- World Share: This variable shows the country's share of the global population. It's often expressed as a percentage.

Here is the dataset that has been scraped:

	#	Country (or dependency)	Population (2023)	Yearly Change	Net Change	Density (P/Km²)	Land Area (Km²)	Migrants (net)	Fert. Rate	Med. Age	Urban Pop %	World Share
0	1	India	1,428,627,663	0.81 %	11,454,490	481	2,973,190	-486,136	2.0	28	36 %	17.76 %
1	2	China	1,425,671,352	-0.02 %	-215,985	152	9,388,211	-310,220	1.2	39	65 %	17.72 %
2	3	United States	339,996,563	0.50 %	1,706,706	37	9,147,420	999,700	1.7	38	83 %	4.23 %
3	4	Indonesia	277,534,122	0.74 %	2,032,783	153	1,811,570	-49,997	2.1	30	59 %	3.45 %
4	5	Pakistan	240,485,658	1.98 %	4,660,796	312	770,880	-165,988	3.3	21	35 %	2.99 %
...	...	...	...	...	...	...	...	...	...	...	...	...
229	230	Montserrat	4,386	-0.09 %	-4	44	100	0	1.6	44	11 %	0.00 %
230	231	Falkland Islands	3,791	0.29 %	11	0	12,170	0	1.6	40	62 %	0.00 %
231	232	Niue	1,935	0.05 %	1	7	260	0	2.4	36	41 %	0.00 %
232	233	Tokelau	1,893	1.18 %	22	189	10	0	2.6	27	0 %	0.00 %
233	234	Holy See	518	1.57 %	8	1,295	0	0			N.A.	0.00 %

234 rows × 12 columns

Libraries which are used in this project are :

```
import pandas as pd
import requests
from bs4 import BeautifulSoup
import matplotlib.pyplot as plt
import numpy as np
```

Columns in the Table:

```
Index(['#', 'Country (or dependency)', 'Population (2023)', 'Yearly Change',
       'Net Change', 'Density (P/Km²)', 'Land Area (Km²)', 'Migrants (net)',
       'Fert. Rate', 'Med. Age', 'Urban Pop %', 'World Share'],
```

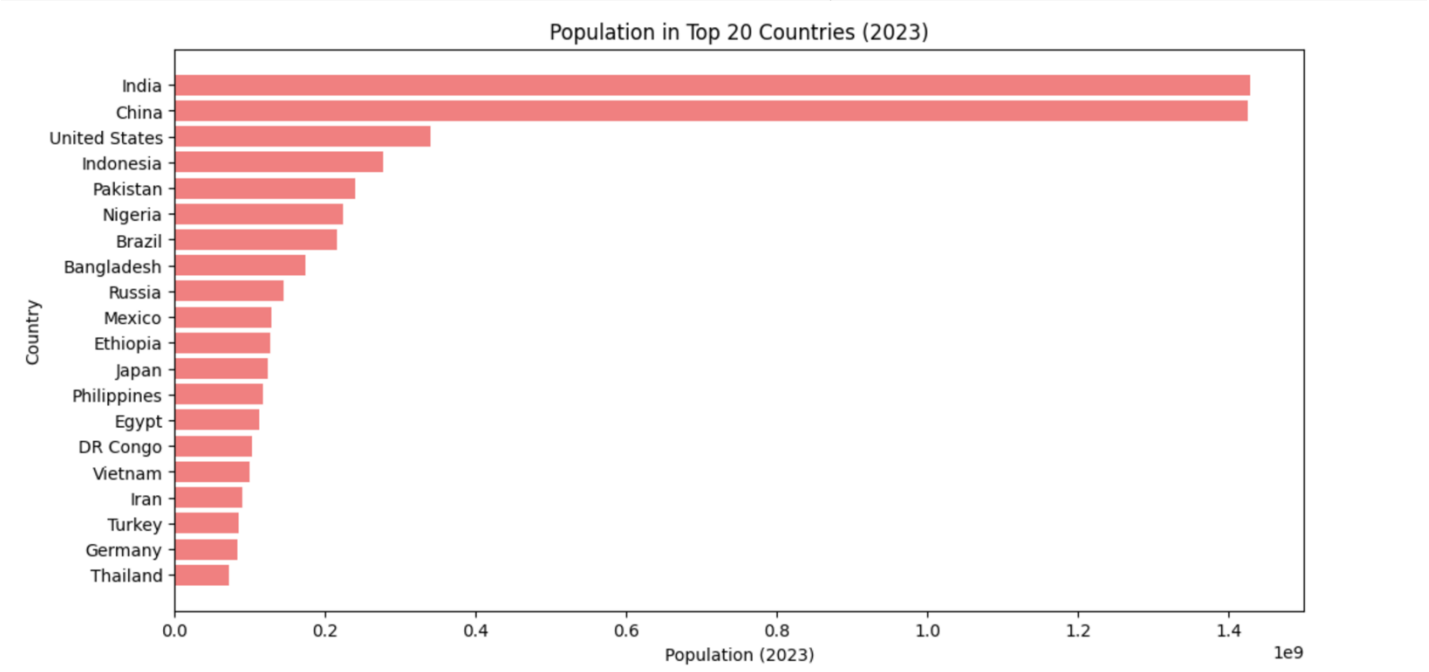
ANALYSIS AND FINDINGS

Population Analysis:
The mean of the population for the year 2023, represents the average population size. The median population provides a measure of central tendency and indicates the middle value in the population distribution. The mode of the dataset represents the most frequently

occurring population size among the countries. To assess the variability in population sizes, we computed the standard deviation and variance. The standard deviation illustrates the degree of dispersion from the mean population. A higher standard deviation suggests greater variability among countries. The variance quantifies the spread of population data points. Finally, the total population encompassing all countries, was determined to know the population size of all the given countries. I identified the country with the largest population in 2023, which is India and the country with the smallest population is Holy See. I further explored the population data by creating a bar chart. The resulting bar chart, shown below, clearly illustrates the distribution of populations among the top 20 countries and I can find that after India, China has second highest population and United States has the third highest Population by 2023.

```
Total Population: 8043901603
Mean Population (2023): 34375647.876068376
Median Population (2023): 5643895.0
Mode Population: 518.0
Standard Deviation of Population (2023): 137386102.42132795
Variance of Population (2023): 1.8874941138523612e+16
```

```
Largest Population: 1428627663.0 (Country: India)
Smallest Population: 518.0 (Country: Holy See)
```



Yearly Change and Net Change Analysis:

Yearly Change represents the annual change in population for the countries included in the dataset to understand the annual variations in population. Net Change represents the difference between births and deaths, which contributes to the overall population growth or decline to gain a comprehensive understanding of population dynamics. Descriptive

statistics has been found out for both Yearly change and Net change. Average yearly change is around 0.97% and Average net change is around 30000000.

```
Descriptive Statistics for Yearly Change:
count    234.000000
mean       0.965470
std        1.242413
min       -7.450000
25%        0.222500
50%        0.805000
75%        1.685000
max        4.980000
```

```
Descriptive Statistics for Net Change:
count    2.340000e+02
mean     3.000230e+05
std      1.001815e+06
min     -2.957105e+06
25%      2.360000e+02
50%      2.860150e+04
75%      2.236855e+05
max      1.145449e+07
```

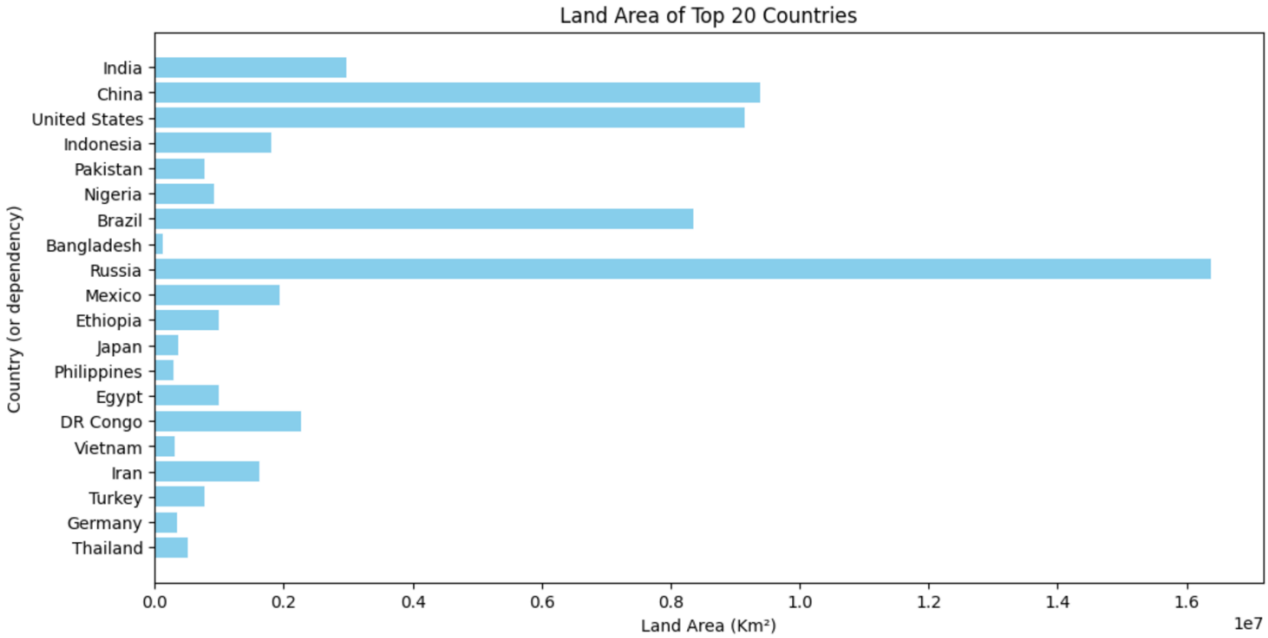
Furthermore, After conducting a correlation analysis between yearly change and net change in population, we found that the correlation coefficient is approximately 0.33. This positive correlation indicates a moderate, positive linear relationship between the two variables. It suggests that, on average, as the yearly population change increases for a country, there tends to be a corresponding increase in net population change, and vice versa. However, the relationship is not strong, indicating that other factors likely influence these changes as well.

```
Correlation between Net Change and Yearly Change: 0.33
```

Land Area (km^2) Analysis:
Land Area represents the total land area of the countries in the dataset. Descriptive statistics were calculated to understand the distribution of land area among the countries, and the countries with the highest and lowest land areas were identified. Country with the highest land area is Russia and the country with the lowest land area is Holy See. Furthermore, A bar chart is created to visualize the land area distribution among the top 20 countries in the dataset. The chart provides a clear comparison of land areas, allowing us to identify countries with the largest land areas in the dataset. After Russia, China has the second highest land area and United States has the third largest land area.

```
Mean Land Area: 555956.81 Km²
Median Land Area: 79720.00 Km²
Mode Land Area(s): [460.0]
```

```
Country with the Highest Land Area: Russia
Country with the Lowest Land Area: Holy See
```

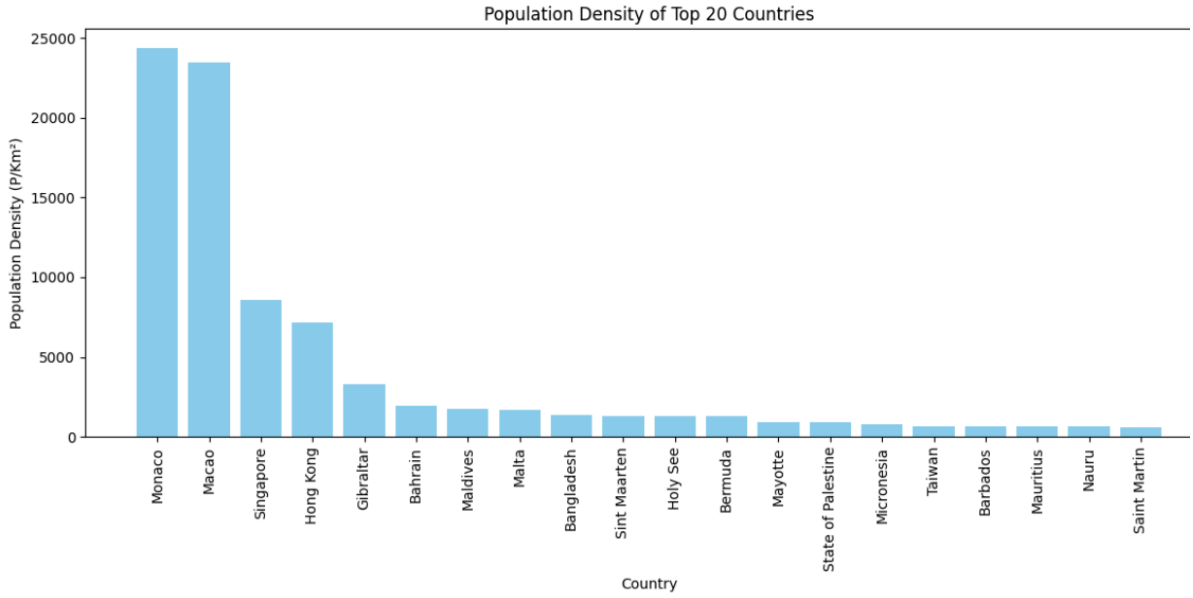


Density (P/km^2) Analysis:

Population density measures the number of people living per square kilometer of land, is analysed for the dataset. Descriptive statistics has been found out to understand the population density across the given countries. On an average, 477. 41 p/km^2 is the population density of all the given countries in the dataset. The highest population density is found in Monaco. On the other hand, Greenland has the lowest population density for 2023. Furthermore, To gain a better understanding of the population density distribution among countries, a graphical representation is created for the top 20 countries. After Monaco, Macao has the second highest population density and Singapore has the third highest population density.

```
Mean Density: 477.41 P/Km²
Median Density: 96.50 P/Km²
Mode Density: 4.0 P/Km²
```

```
Country with the Highest Population Density: Monaco
Country with the Lowest Population Density: Greenland
```



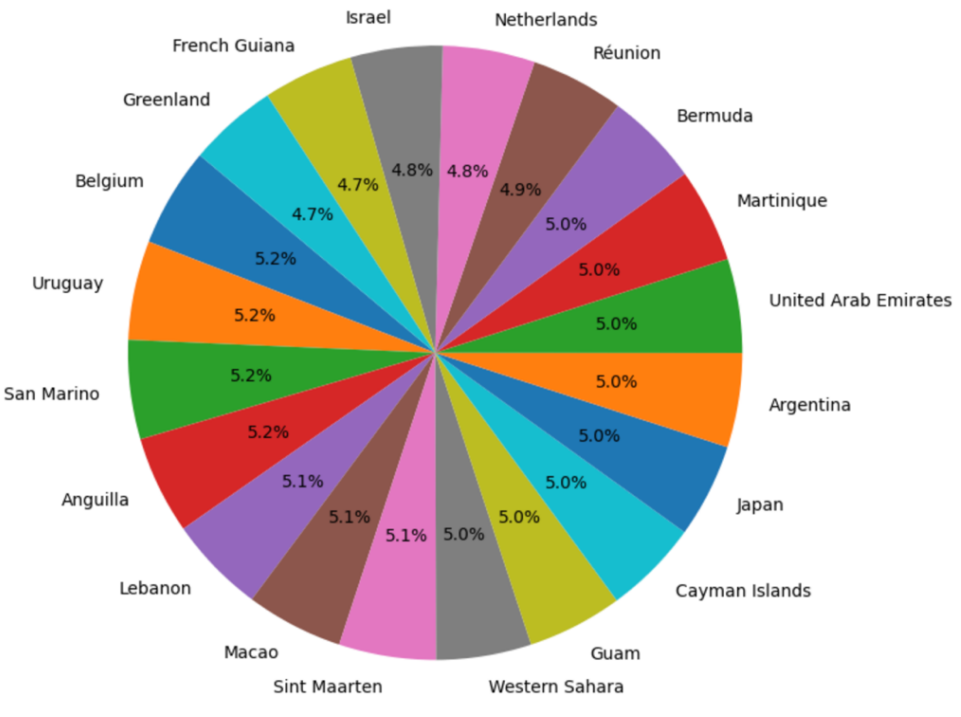
Correlation Analysis between Land Area and Density :
The correlation coefficient measures the strength and direction of the linear relationship between land area and population density. I conducted a correlation analysis to explore the relationship between land area (Km²) and population density (P/Km²). The correlation coefficient between these two variables was found to be approximately -0.06. A negative correlation coefficient suggests that as land area increases, population density tends to decrease. The weak negative correlation may imply that larger countries, in terms of land area, tend to have a relatively lower population density. This indicates that there is a very slight tendency for areas with larger land areas to have slightly lower population densities.

```
Correlation between Density and Land Area: -0.06
```

Urban Population Analysis :
Urban population is a demographic indicator that provides insights into the level of urbanization within a given area. Descriptive Statistics for urban population offer valuable insights into the distribution and characteristics of urban population percentages in the dataset. On an average, countries have approximately 60% of urban population. Countries with higher urban population percentages may have more developed urban infrastructures, while those with lower percentages might rely more on rural lifestyles. My analysis revealed that Belgium holds the top position with the highest percentage of urban population. On the other end, Saint Martin has the lowest urban population percentage.

```
Descriptive Statistics for Urban Population (%):
count    216.000000
mean      59.550926
std       23.826450
min        0.000000
25%       40.750000
50%       61.500000
75%       79.250000
max       99.000000
```

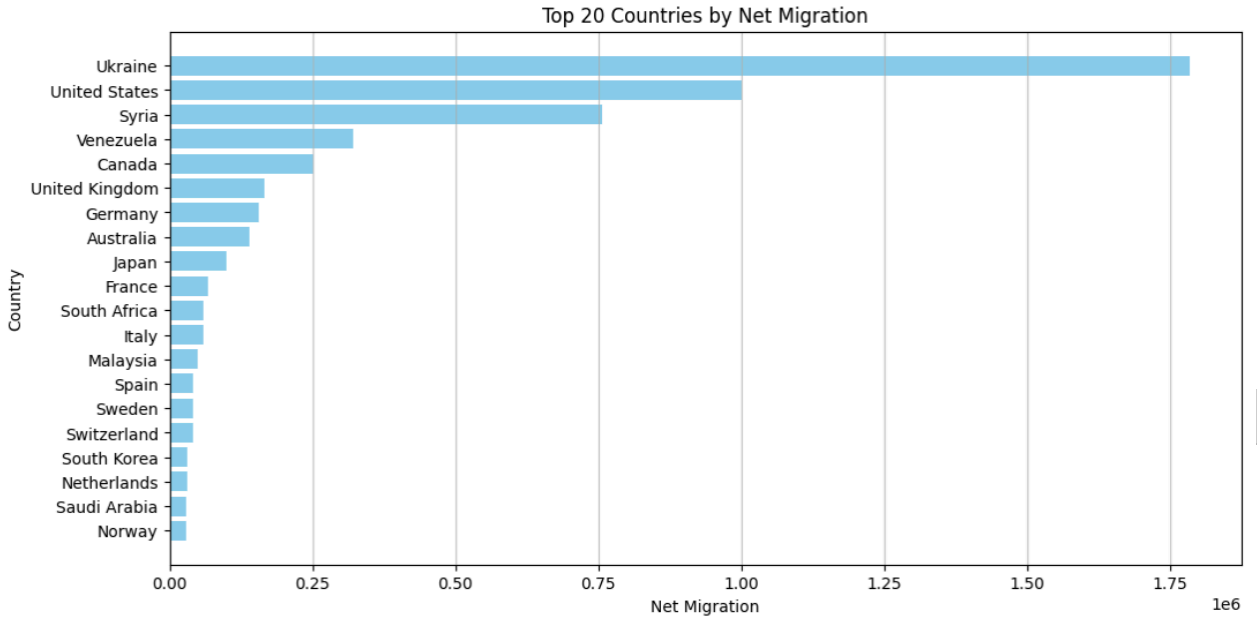
```
Country with Highest Urban Population: Belgium (99.0%)
Country with Lowest Urban Population: Saint Martin (0.0%)
```



Net Migration Analysis:
Net migration refers to the difference between the number of people immigrating (moving into a country) and the number of people emigrating (leaving a country) during a specific time period, typically a year. Descriptive analysis depicts the patterns and variations in population movements, whether it be due to immigration or emigration. It is found that the mean net migration is approximately 130000, indicating that, on average, 1300000 people are added or lost from a country's population due to migration. Furthermore, bar graph shows the net migration of top 20 countries, Ukraine being on the top, United States on the second and Syria on the third position. On the other hand, Poland has the lowest Net Migration.

```
Descriptive Statistics for Net Migration:
count    2.340000e+02
mean     1.301282e+01
std      1.698334e+05
min     -9.104750e+05
25%     -9.776750e+03
50%     -5.000000e+02
75%      4.750000e+02
max      1.784718e+06
```

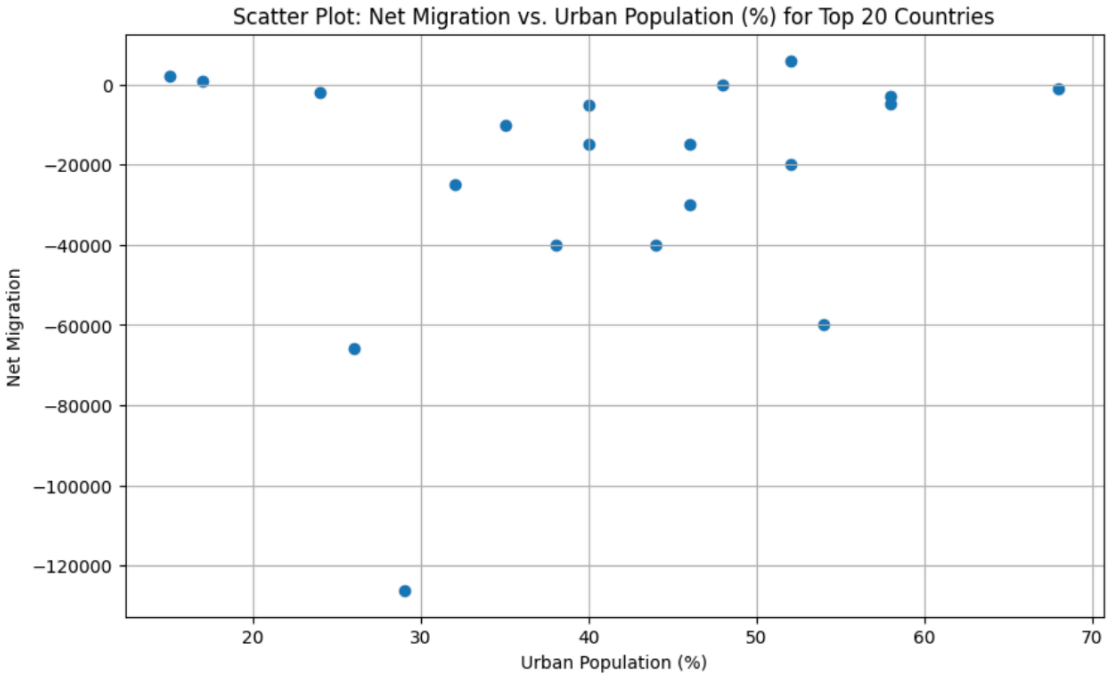
```
Country with the Highest Net Migration: Ukraine
Country with the Lowest Net Migration: Poland
```



Correlation Analysis between Urban Population and Net Migration:
The correlation coefficient quantifies the strength and direction of the linear relationship between two variables. I have calculated the correlation coefficient between Net Migration and Urban Population and revealed a correlation coefficient of 0.17. A coefficient of 0.17 suggests a positive but relatively weak correlation between net migration and the percentage of the population living in urban areas. A positive correlation implies that as net migration increases (more people moving into a country than leaving), there is a tendency for a higher urban population percentage. However, the strength of this relationship is not

particularly strong, indicating that other factors may also play a significant role in determining urbanization patterns within these countries.

```
Correlation between Net Migrants and Urban Population (%): 0.17
```

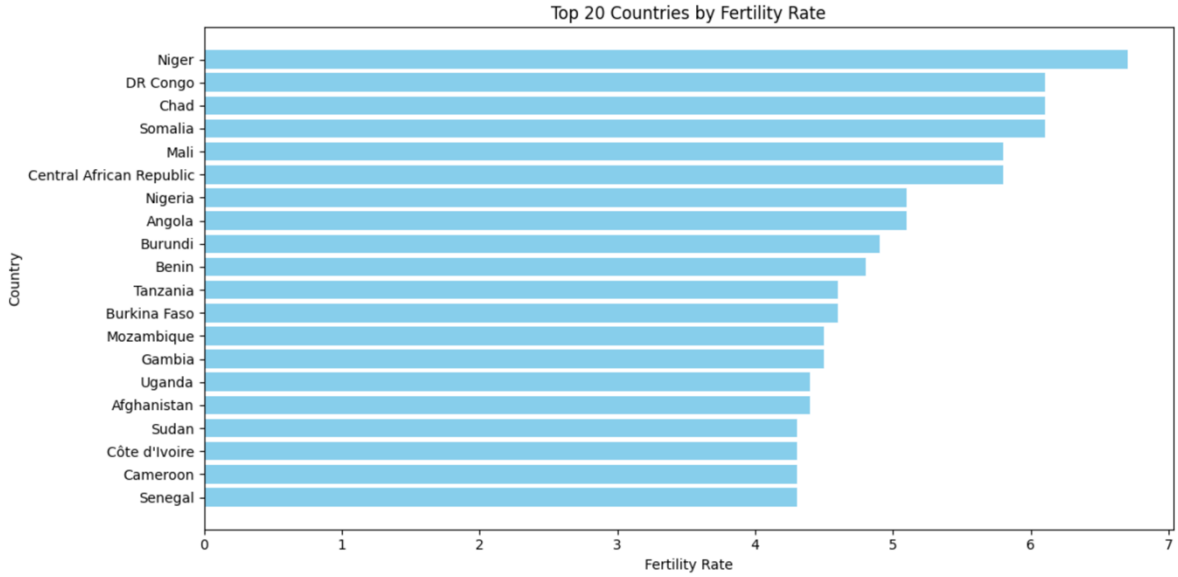


Fertility Rate Analysis :

Fertility rate is a critical demographic indicator that represents the average number of children born to a woman during her lifetime in a specific country or region. Descriptive analysis has been done to understand the valuable insights from mean, median and mode. The average number of children born to women in the countries included in the dataset is 2.41. The country with the highest fertility rate is Niger and the country with the lowest fertility rate is Hong Kong. Furthermore, In the analysis of fertility rates, a bar graph is created to visualize the fertility rates of the top 20 countries. The bar graph clearly illustrates the variations in fertility rates among these countries, allowing for easy comparison. After Niger, DR congo has the second highest fertility rate and Chad has the third highest fertility rate.

```
Mean Fertility Rate: 2.41
Median Fertility Rate: 2.00
Mode Fertility Rate: 1.60
```

```
Highest Fertility Country: Niger
Lowest Fertility Country: Hong Kong
```

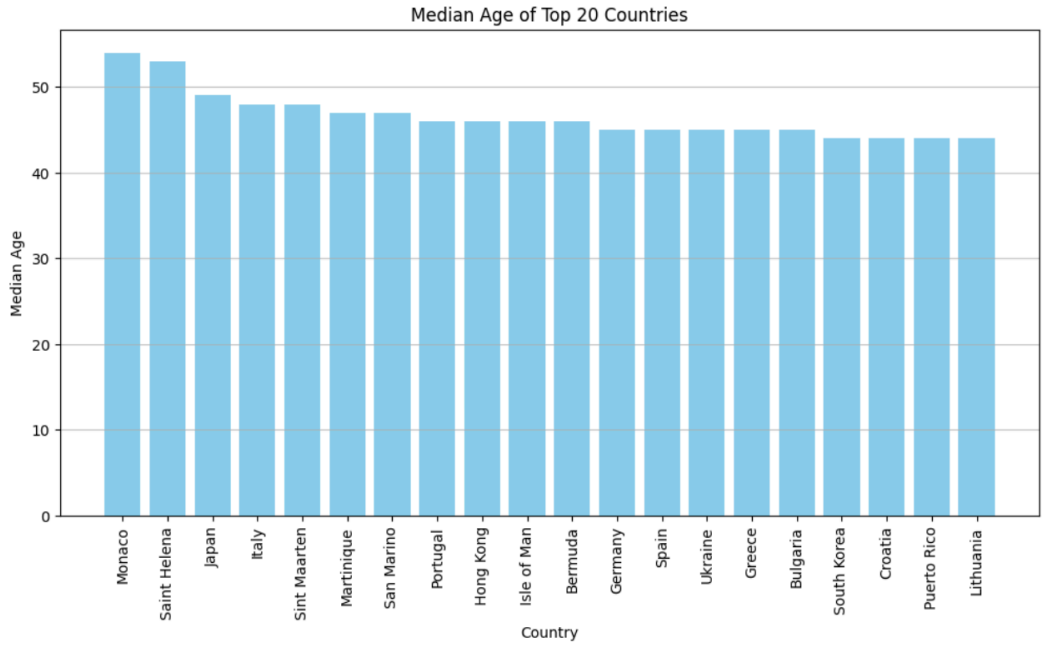


Median Age Analysis:

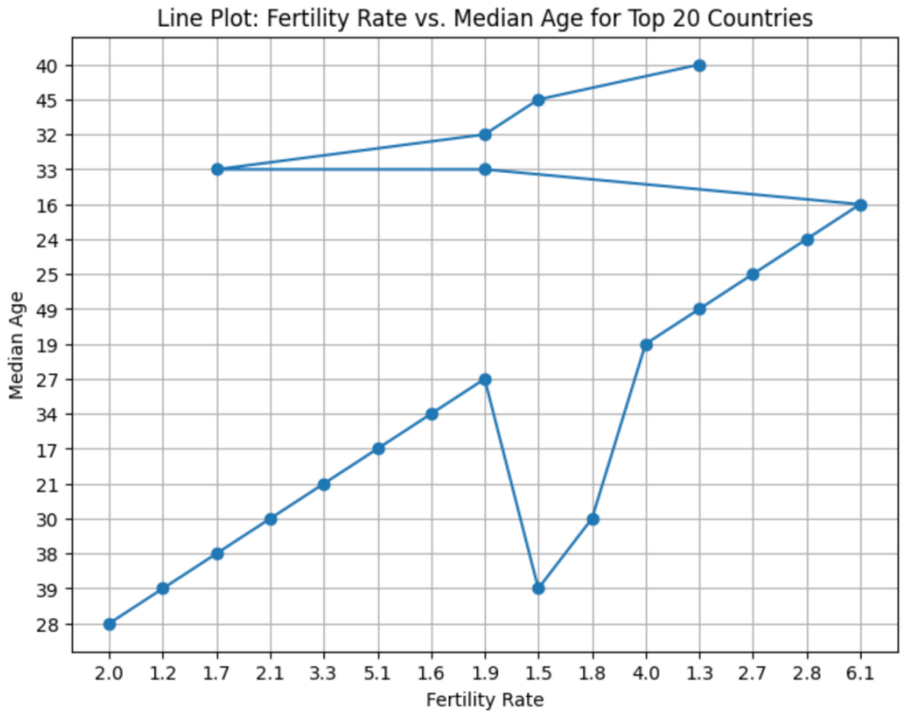
The median age is a demographic measure that represents the age at which half of the population is older and half is younger. It's the middle value of a dataset when all ages are arranged in ascending order. Descriptive analysis has been done to understand the valuable insights from mean, median and mode. The average media age of the population in the dataset is 32 years. The country with the highest Median age is Monaco and the country with the lowest Median age is Niger. Furthermore, In the analysis of median age, a bar graph is created to visualize the median age of the top 20 countries. After Monaco, Saint Helena has the second highest fertility rate and Japan has the third highest fertility rate.

```
Mean Median Age: 31.31
Median Median Age: 32.00
Mode Median Age: 40.00
```

```
Country with the Highest Median Age: Monaco
Country with the Lowest Median Age: Niger
```

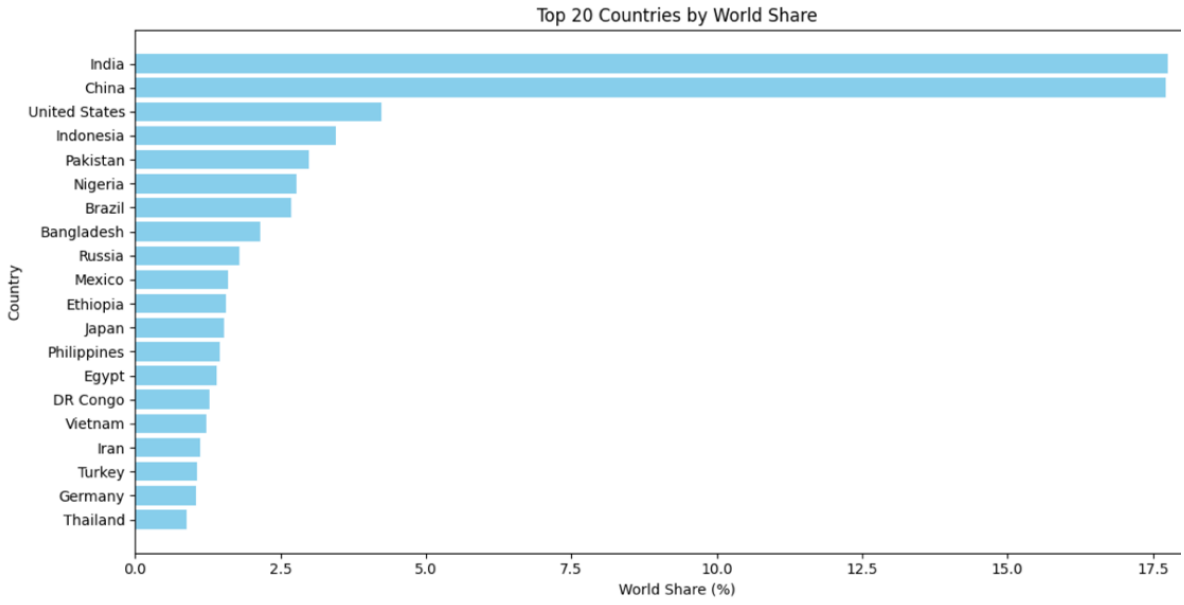


Correlation Analysis between Fertility Rate and Median Age :
For this visual representation, we created a line plot with fertility rate on the y-axis and median age on the x-axis. Top 20 countries in the dataset is represented as a data point on the plot. The resulting lines connect these data points, allowing us to observe trends in fertility rates concerning changes in median age.



World Share Analysis:
World Share refers to the percentage of the world's total population that a particular country or region represents. The country with the highest world share in the dataset is India which makes up 17.76% of the global population. Conversely, the country with the lowest world share is Guadeloupe whose global share is negligible. Graphical representation of world share of top 20 countries shows that after India, China is on the second position and United States on the third Position.

```
Country with the Highest World Share: India
Highest World Share Percentage: 17.76
Country with the Lowest World Share: Guadeloupe
Lowest World Share Percentage: 0.0
```



MANAGERIAL IMPLICATIONS

Analysing the data from Worldometers info of countries in the world by population including various demographic factors (yearly change, net change, land area, density, migrants, etc) can provide valuable managerial implications for various stakeholders, including policymakers, businesses, and researchers.
Here are some potential managerial implications:

Population Trends and Growth Rates:

- Understanding the population growth trends can help policymakers plan for future infrastructure and resource allocation.
- Businesses can use population growth data to identify potential markets for expansion or areas with declining populations for market saturation.

Fertility Rates:

- High fertility rates in certain countries indicate the need for family planning and healthcare services.
- Low fertility rates may signal potential labour shortages, impacting businesses and workforce planning.

Median Age:

- A high median age suggests an aging population, which can have implications for healthcare services, retirement planning, and labour markets.
- A low median age indicates a youthful population, which can be an opportunity for businesses targeting younger consumers.

Urbanization:

- High urbanization rates may require investments in urban infrastructure, transportation, and housing.
- Businesses may focus marketing efforts on urban areas with higher consumer density.

Net Migration:

- Positive net migration indicate economic opportunities, while negative net migration may signal economic challenges.
- Policymakers can use this data to address immigration policies and labour force issues.

Land Area and Population Density:

- High population density countries require more efficient land use planning and public services.
- Lower population density countries may need incentives for economic development and infrastructure improvement.

World Share:

- The distribution of world share among countries can influence international trade agreements and global diplomatic relations.

- Businesses can target markets based on the share of the global population.

Correlations:

- Identifying correlations between demographic factors can inform policy decisions and business strategies. For example, the correlation between fertility rates and median age can help plan for future workforce needs.

These insights and implications can guide decision-making processes at various levels, from national policies to business strategies, ultimately contributing to better resource allocation, economic development, and the well-being of populations.