REPORT ON WORLD POPULATION

6

1. Introduction

The current US Census Bureau world population estimate in June 2019 shows that the current global population is 7,577,130,400 people on earth, which far exceeds the world population of 7.2 billion in 2015. Our own estimate based on UN data shows the world's population surpassing 7.7 billion.

Business Problems We are Trying to Solve

- Problems we generally face while finding world population like:
    - We are not aware of general factors that is affecting the population growth rate.
    - Sudden outbreak in population.
    - Here, we are making predictions of the world population in near future in context of past population.
    - We are also making conclusion what are the factors affecting the world population.

Main Goal

- Create an analytical framework to understand
    - Key factors impacting world population.
- Develop a modelling framework
    - To estimate the future population of the world.

About Dataset

In this Dataset, we have Historical Population data for every Country/Territory in the world by different parameters like Area Size of the Country/Territory, Name of the

Continent, Name of the Capital, Density, Population Growth Rate, Ranking based on Population, World Population Percentage, etc.

- **Rank:** Rank by population
- **CCA3:** 3 digit Country/Territories code
- **Country:** Name of the Country/Territories
- **Capital:** Name of the Capital
- **Continent:** Name of the Continent
- **2022 Population:** Population of the Country/Territories in the year 2022
- **2020 Population:** Population of the Country/Territories in the year 2020
- **2015 Population:** Population of the Country/Territories in the year 2015
- **2010 Population:** Population of the Country/Territories in the year 2010
- **2000 Population:** Population of the Country/Territories in the year 2000
- **1990 Population:** Population of the Country/Territories in the year 1990
- **1980 Population:** Population of the Country/Territories in the year 1980
- **1970 Population:** Population of the Country/Territories in the year 1970
- **Area (km²):** Area size of the Country/Territories in square kilometer
- **Density (per km²):** Population density per square kilometer
- **Growth Rate:** Population growth rate by Country/Territories
- **World Population Percentage:** The population percentage by each Country/Territories

BASIC EXPLORATION:

Let's have a glimpse of the dataset.

Shape Of The Dataset : (234, 17)

Glimpse Of The Dataset :

```
[3]: df.head()
```

[3]:

	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)	Grow Ra
0	36	AFG	Afghanistan	Kabul	Asia	41128771	38972230	33753499	28189672	19542982	10694796	12486631	10752971	652230	63.0587	1.02
1	138	ALB	Albania	Tirana	Europe	2842321	2866849	2882481	2913399	3182021	3295066	2941651	2324731	28748	98.8702	0.99
2	34	DZA	Algeria	Algiers	Africa	44903225	43451666	39543154	35856344	30774621	25518074	18739378	13795915	2381741	18.8531	1.01
3	213	ASM	American Samoa	Pago Pago	Oceania	44273	46189	51368	54849	58230	47818	32886	27075	199	222.4774	0.98
4	203	AND	Andorra	Andorra la Vella	Europe	79824	77700	71746	71519	66097	53569	35611	19860	468	170.5641	1.01

```
[ ]:
```

Table1- DataSet

It has a number rows, which is 234 rows spread across 17 columns. Here is the list of columns this dataset has,

```
[5]: df.columns
```

```
[5]: 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)',
       'Growth Rate', 'World Population Percentage'],
      dtype='object')
```

Information of the Dataset:

```
[9]: df.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
```

2. EDA & Business Implication

EDA stands for exploratory data analysis where we explore our data and grab insights from it. EDA helps us in getting knowledge in form of various plots and diagrams where we can easily understand the data and its features.

Dataset Summary:

```
[7]: df.describe()
```

[7]:

	Rank	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	2.340000e+02	2.340000e+02	2.340000e+02	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	117.500000	3.407441e+07	3.350107e+07	3.172996e+07	2.984524e+07	2.626947e+07	2.271022e+07	1.898462e+07	1.578691e+07	5.814494e+05	452.127044	1.009577	0.427051
std	67.694165	1.367664e+08	1.355899e+08	1.304050e+08	1.242185e+08	1.116982e+08	9.783217e+07	8.178519e+07	6.779509e+07	1.761841e+06	2066.121904	0.013385	1.714977
min	1.000000	5.100000e+02	5.200000e+02	5.640000e+02	5.960000e+02	6.510000e+02	7.000000e+02	7.330000e+02	7.520000e+02	1.000000e+00	0.026100	0.912000	0.000000
25%	59.250000	4.197385e+05	4.152845e+05	4.046760e+05	3.931490e+05	3.272420e+05	2.641158e+05	2.296142e+05	1.559970e+05	2.650000e+03	38.417875	1.001775	0.010000
50%	117.500000	5.559944e+06	5.493074e+06	5.307400e+06	4.942770e+06	4.292907e+06	3.825410e+06	3.141146e+06	2.604830e+06	8.119950e+04	95.346750	1.007900	0.070000
75%	175.750000	2.247650e+07	2.144798e+07	1.973085e+07	1.915957e+07	1.576230e+07	1.186923e+07	9.826054e+06	8.817329e+06	4.304258e+05	238.933250	1.016950	0.280000
max	234.000000	1.425887e+09	1.424930e+09	1.393715e+09	1.348191e+09	1.264099e+09	1.153704e+09	9.823725e+08	8.225344e+08	1.709824e+07	23172.266700	1.069100	17.880000

Null Values of the Dataset:

```
[18]: df.isnull().sum()
```

```
[18]: 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
```

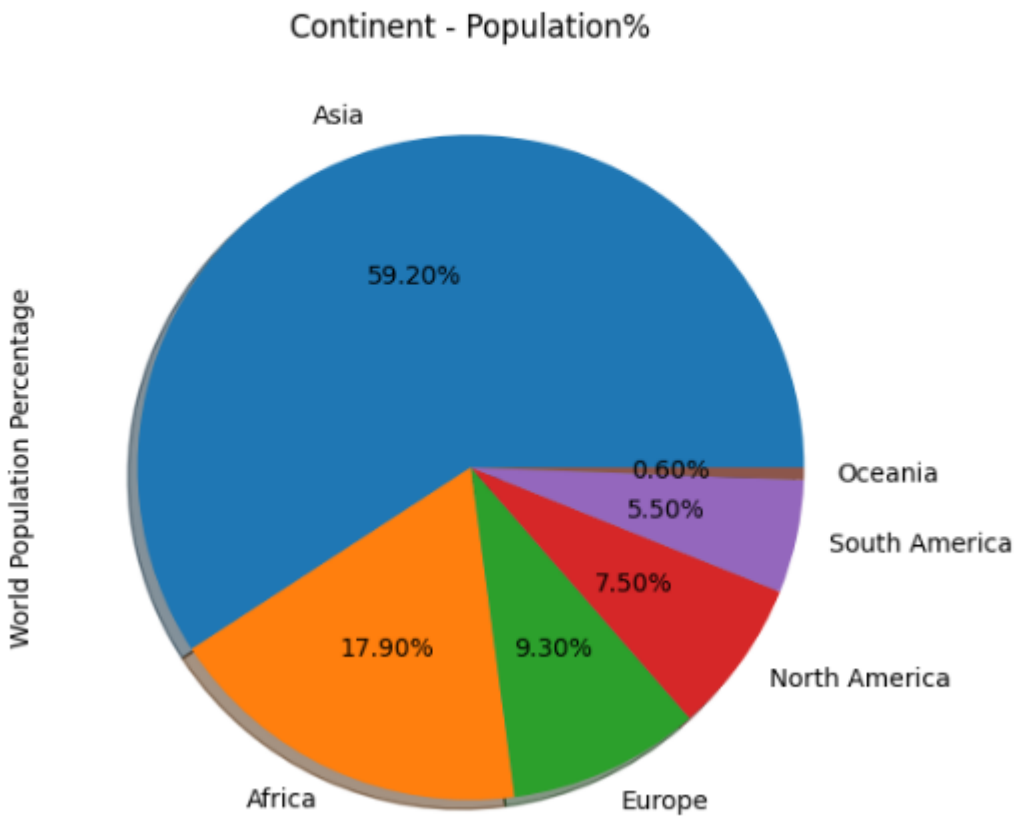
Insights:

- There is no missing values in this dataset.
- We will encode the categorical features into numerical form later.

Population Continent wise:

```
[42]: pp.plot.pie(labels= labels, autopct = '%1.2f%%', figsize= (12,6),shadow = True)
      plt.title('Continent - Population%')
```

```
[42]: Text(0.5, 1.0, 'Continent - Population%')
```



Observation:

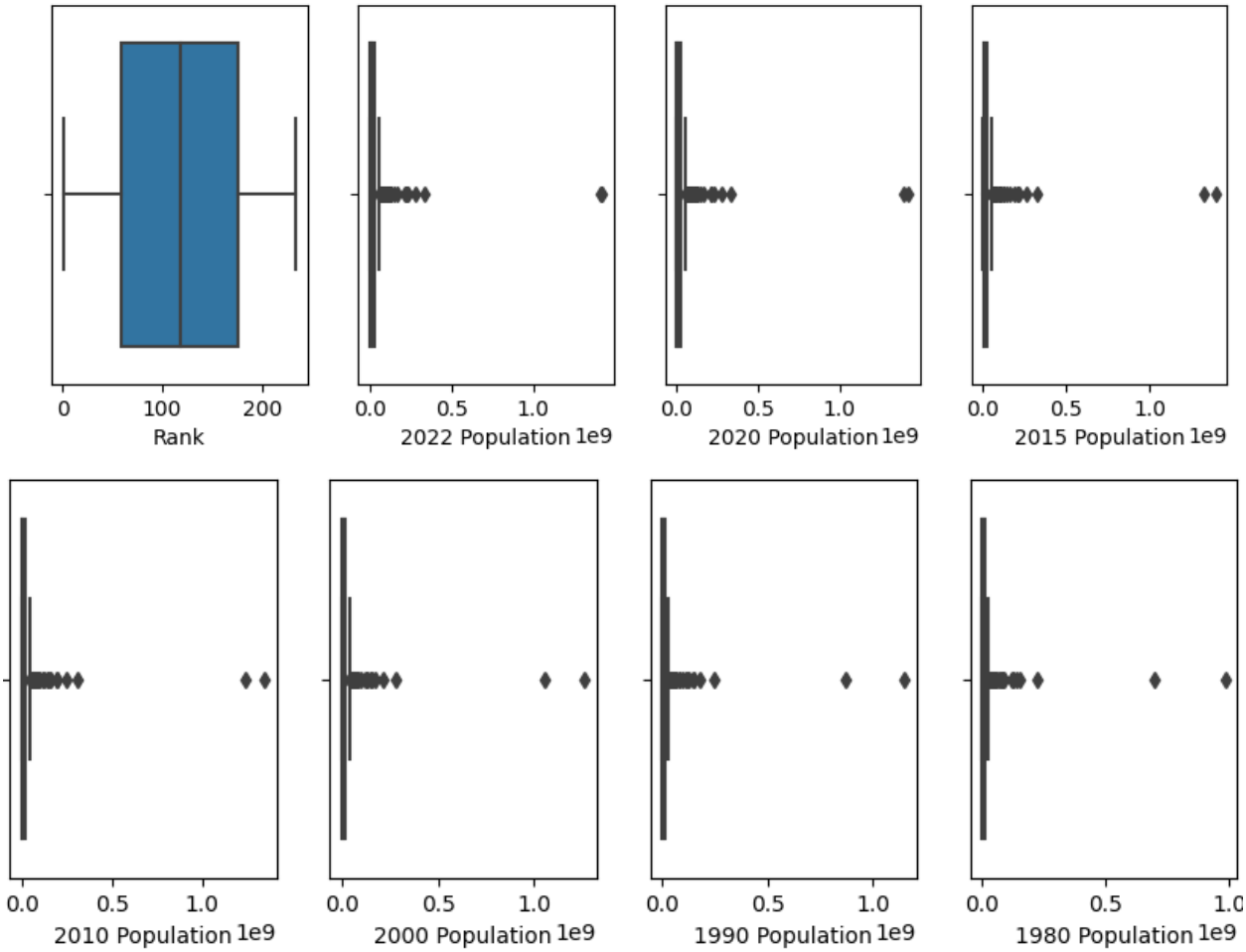
- Asia is the most densely populated continent with 59.20% followed by Africa, Europe and others.
- Oceania is the least populated with 0.6% followed by South America, North America and others.

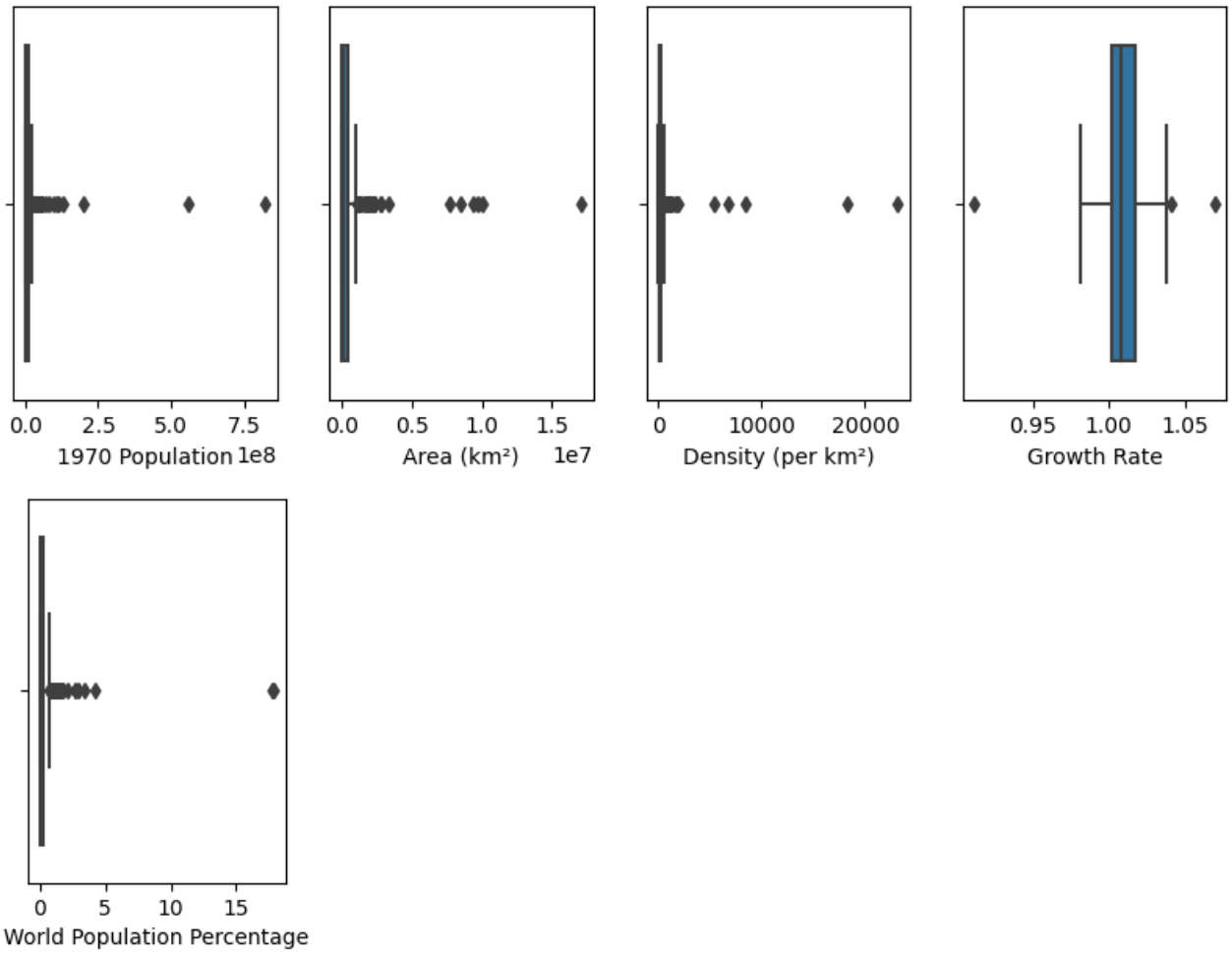
Continent contribution by Population in 2022:

[69]:

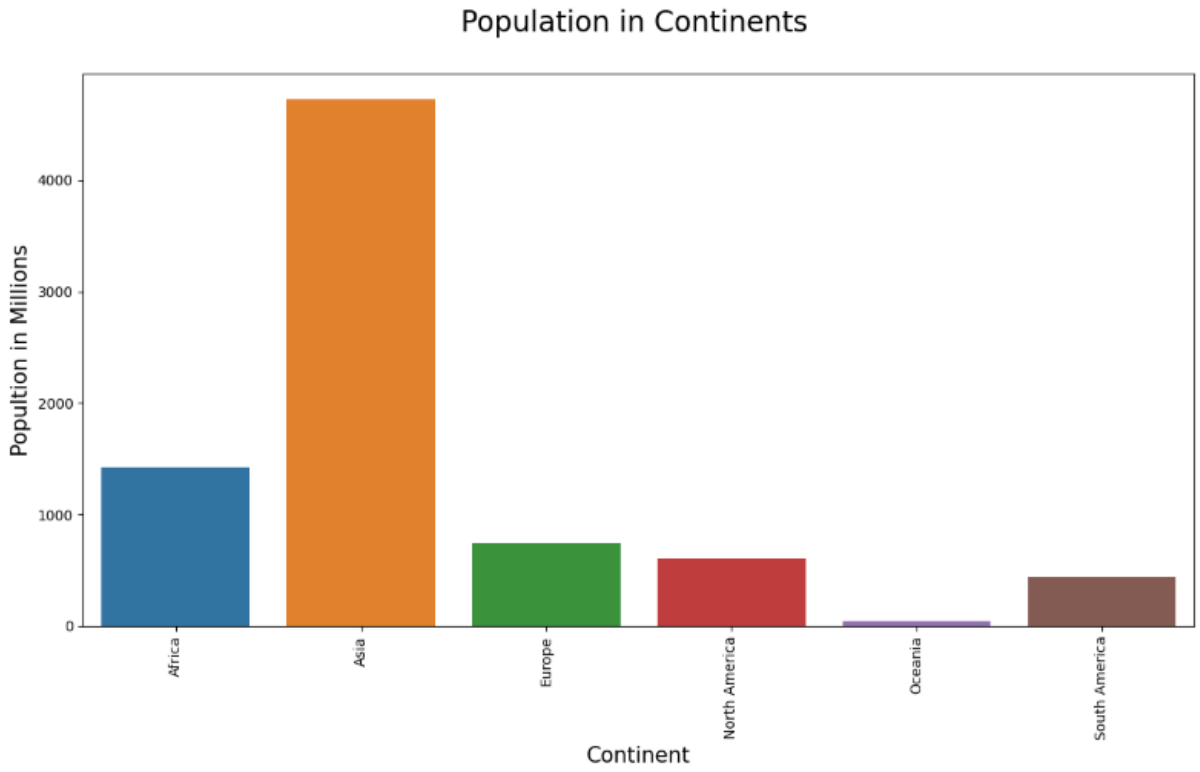
Continent	2022 Population	2022 Population Percentage
Asia	4721383274	59.2
Africa	1426730932	17.9
Europe	743147538	9.3
North America	600296136	7.5
South America	436816608	5.5
Oceania	45038554	0.6

Analysis-Boxplot




0.0
2.5
5.0
7.5
1970 Population
1e8
0.0
0.5
1.0
1.5
Area (km²)
1e7
0
10000
20000
Density (per km²)
0.95
1.00
1.05
Growth Rate
0
5
10
15
World Population Percentage

Population in 2022 per continent:



3. Data Cleaning & Pre-processing

Data Cleaning is an important phase in any data science project, if our data is clean then only we can provide it to our machine learning model. Uncleaned Data can further lead our model with low accuracy. And, If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset.

The approach used for identifying and treating missing values and outlier treatment:-

```
[102]: df.isnull().sum()

[102]: 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
```

To identify any missing values in our data set we have used **Pandas** pre built function **isnull()** to detect any missing values in our datasets. Since we have zero missing values in our data. But in case if we have missing values then our next step would how to handle a large number of missing values. One approach is, that we will delete the column if we don't need that column for further analysis. And, what if we need that column for further analysis then we have use an approach will is a predefined function in **Pandas** called fillna().
How we can fill the missing values in a **categorical** variable using **mode**. And, How we can fill the missing values in a **numerical** variable using **Median**. This is how we have an approach for identifying and handling missing values.

While performing Preprocessing and Data cleaning we have to also deal with outliers. Dealing with outliers is also a necessary step to be taken for further analysis and model building. Outliers are data points in a data set that is distant from all other observations. A data point that lies outside the overall distribution of the dataset.