

Report On World population



|  |  |  |
| --- | --- | --- |
| **S. NO.** | **Topic** | **Page No.** |
| 1 | Introduction   * Business Problems We are trying to solve * Main goal * About Dataset   + Table1- DataSet Preview   + Table 2 - Columns in Dataset   + Basic Exploration   + Informations of Dataset | 2  2  2  2  3  3  3  4 |
| 2 | EDA and Pre-EDA   * Analysis of the Data * Dataset Summary * Null values of Dataset * Population Continent wise * Analysis – BoxPlot   + Plot 1- Boxplot   + Plot 2 – Analysis of Rank   + Plot 3- Analysis of CCA3   + Plot 4- Analysis of Country   + Plot 5- Analysis of Capital   + Plot 6- Analysis of Continent   + Plot 7- Analysis of 2022 Population   + Plot 8 - Analysis of 2020 Population   + Plot 9 - Analysis of 2015 Population   + Plot 10 - Analysis of 2010 Population   + Plot 11 - Analysis of 2000 Population   + Plot 12 - Analysis of 1990 Population   + Plot 13 - Analysis of 1980 Population   + Plot 14 - Analysis of 1970 Population   + Plot 15 - Analysis of Area   + Plot 16 - Analysis of Density   + Plot 17 – Analysis of Growth Rate   + Plot 18 – World Population Percentage   + Plot 33 – Visualizing the continents   + Population in 2022 per continent: | 5  5  5  6  7  8  8  8  8  8  8  8  8  8  8  8  8  9  9  9  9  9  9  9  10 |
| 3 | Data Cleaning and Preprocessing | 11 |

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.

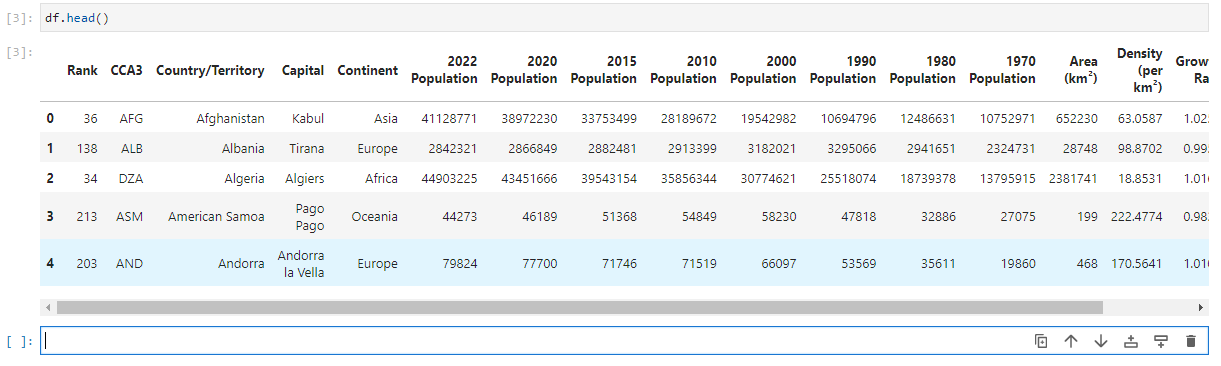
* **Rank:** Rank by population
* **CCA3:** 3 digit Country/Territories code
* **Country:** Name of the Country/Territories
* **Capital:** Name of the Capital
* **Continent:** Name of the Continent
* **2022 Population:** Population of the Country/Territories in the year 2022
* **2020 Population:** Population 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.**

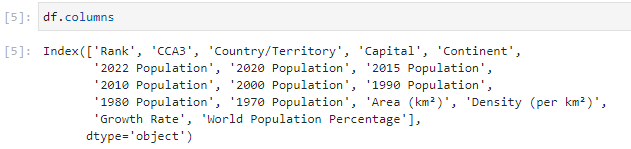
Shape Of The Dataset : (234, 17)

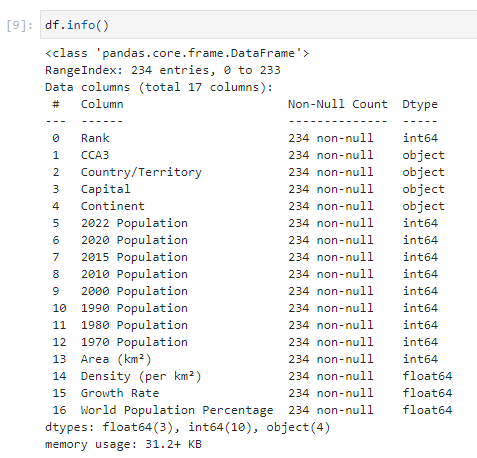
Glimpse Of The Dataset :



**Table1- DataSet Preview**

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

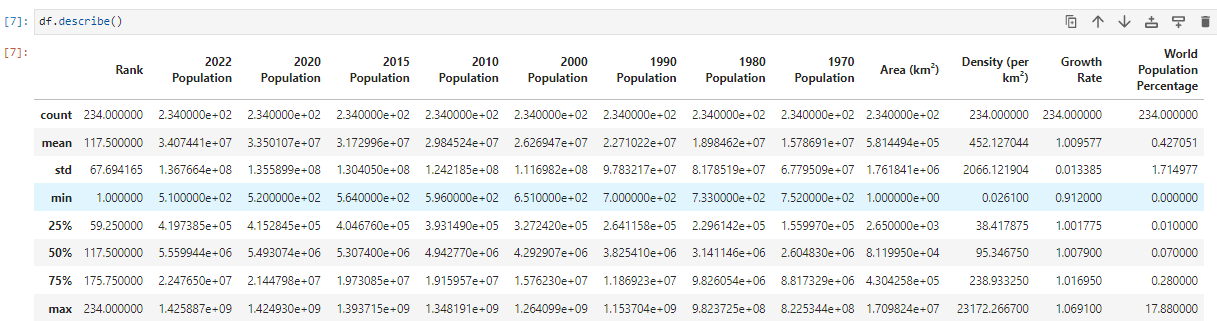


**Information of the Dataset:**  


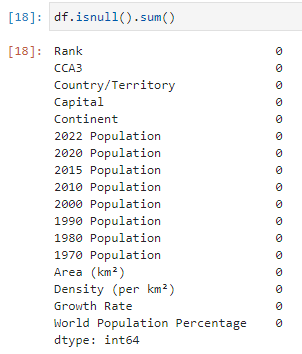
1. **EDA & Business Implication**

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:**



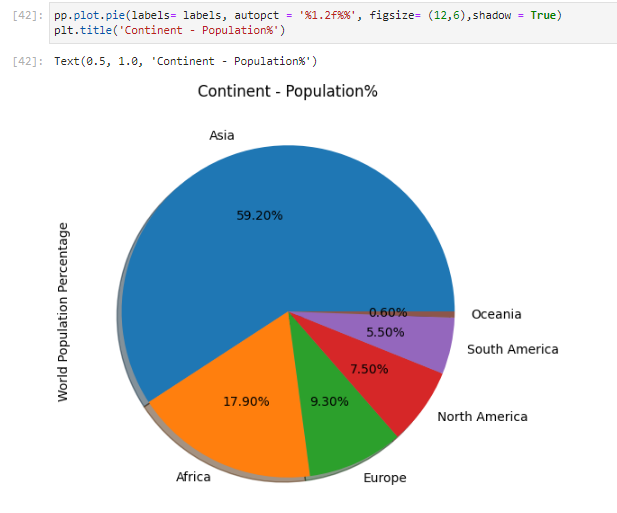
**Null Values of the Dataset:**

****

**Insights:**

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

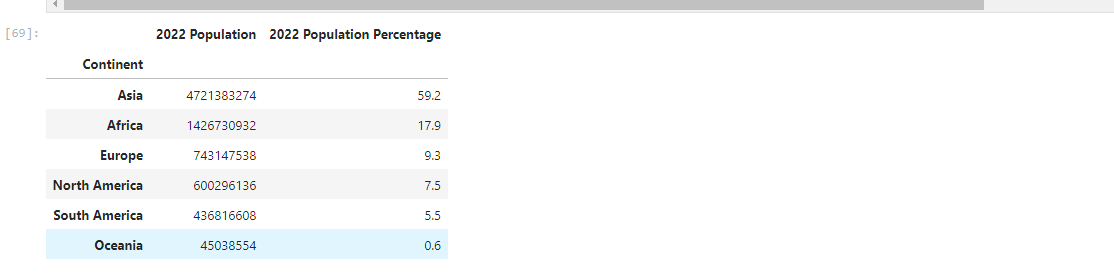
**Population Continent wise:**

****

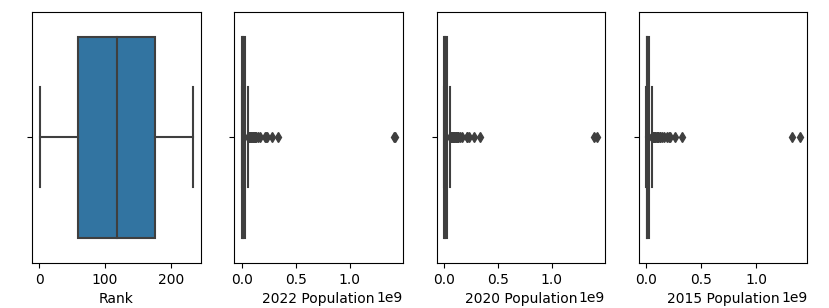
**Observation:**

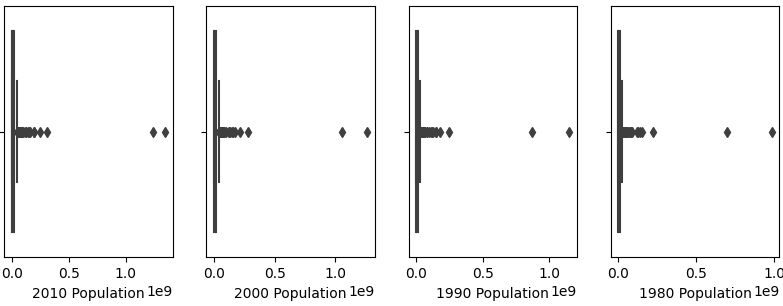
* Asia is the most densely populated continent with 59.20% followed by Africa, Europe and others.
* Oceania is the least populated with 0.6% followed by South America, North America and others.

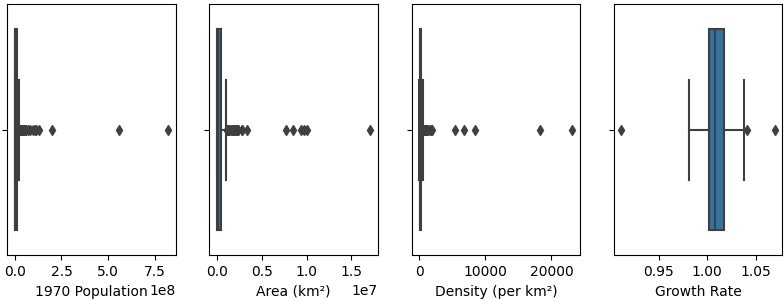
Continent contribution by Population in 2022:



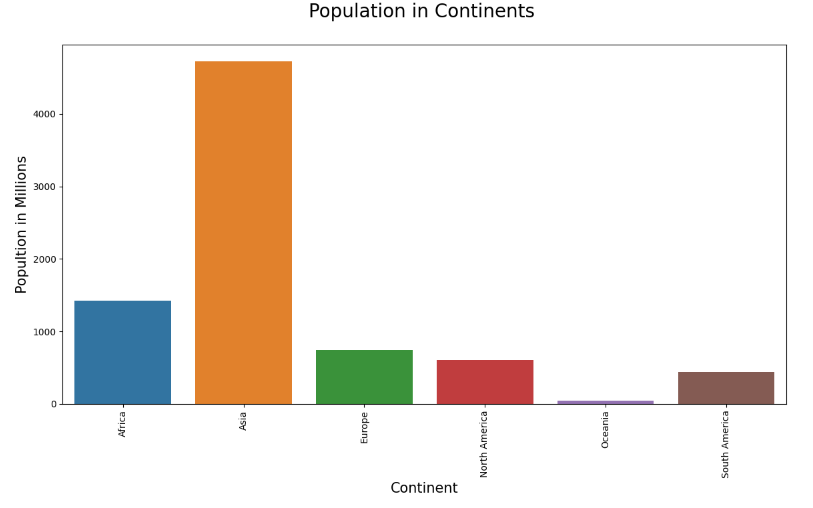
Analysis-Boxplot







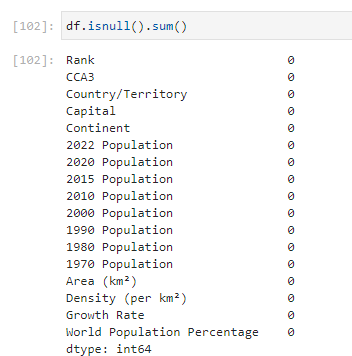


**Population in 2022 per continent:**  


1. **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:-**



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.