World Population Dataset

Introduction

I researched from Kaggle, where I found the dataset about the population of the world. The data seemed to hold the population of each country at different year instances like 2022, 2020, 2015 and so on. It also seemed to convey the topographical details of the nation and provides relation in the form of density to relate the population of the country with its topography. The said dataset contains data of over 234 countries. The original source of the dataset, being discussed, is from world population review.

Data Overview

Dataset Glossary (Column-Wise)

The actual columns with the corresponding definitions, as taken from Kaggle-

- 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 (km^2): Area size of the Country/Territories in square kilometre.
- Density (per km^2): Population Density per square kilometre.
- Growth Rate: Population Growth Rate by Country/Territories.
- World Population Percentage: The population percentage by each Country/Territories.

```
Rank                            int64
CCA3                           object
Country                        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
```

Sample Tabulation

Figure 1 describes the sample table for the data mentioned. Here we see different attributes with different types of values (both numerically and data-type wise). From the tabulation, we can see the variety of attributes mentioned, which might or not necessary for our analyses, about which will be discussed at later stage.

	Rank	CCA3	Country	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
0	36	AFG	Afghanistan	Kabul	Asia	41128771	38972230	33753499	28189672	19542982	10694796	12486631	10752971	652230	63.0587	1.0257	0.52
1	138	ALB	Albania	Tirana	Europe	2842321	2866849	2882481	2913399	3182021	3295066	2941651	2324731	28748	98.8702	0.9957	0.04
2	34	DZA	Algeria	Algiers	Africa	44903225	43451666	39543154	35856344	30774621	25518074	18739378	13795915	2381741	18.8531	1.0164	0.56
3	213	ASM	American Samoa	Pago Pago	Oceania	44273	46189	51368	54849	58230	47818	32886	27075	199	222.4774	0.9831	0.00
4	203	AND	Andorra	Andorra la Vella	Europe	79824	77700	71746	71519	66097	53569	35611	19860	468	170.5641	1.0100	0.00

Figure 1: A sample tabulation of the data as mentioned above.

Exploratory Data Analysis

It is first recommended to analyse the data and explore more about it, before coming to a decision as to how to approach the problem or how to answer certain questions (here hypothesis)

Dataset Cleaning, is it necessary?

The data available here is clear of missing value and thus the issue of data column being not useful and being problematic is not an issue.

Do we need to clean the data? There is no need for cleaning any given data. Since there are no left out columns present. Maybe for intuitive understanding, we can express the population in Billions, to offset the huge length of the population value. But nevertheless, this does not pose any disturbance to the analysis. Given these details, we find that any data column can be used to analyse the dataset.

Apart from the above comment, certain columns are having their datatype as objects, which needed to be changed to string.

Planning for data exploration

From the available data, it is clear that there are not much to clean the dataset. So we don't have to worry about that. The following steps will provide a brief illustration as to what our overall planning will be:

1. Determine the numerical columns.
2. Using box plots, we have to find what are the extremes of a particular column and is it exceeding the average of the countries at that particular year.
3. If needed, these population values can also be forcefully compressed to lie between 0-1 for normalization. Normalization also gives us range of the countries lying around the mean, which can be a qualitative information.
4. Once the data is sorted, at least from the numerical perspective, we will try the corelation function from plotly. This will tell us what attributes are actually correlated and can be used for analysis purpose.
5. From the correlation, we determine if we can fit them into some polynomials.
6. My guess is that with the data, we can fit the population for each country at any particular year into a polynomial of some degree *n*. The polynomial of one country will be compared with other countries. It might be possible that *n* might not work, so we try to use other values for *n.* Then it becomes in our interest to find if *n* has some relation with other numerical.

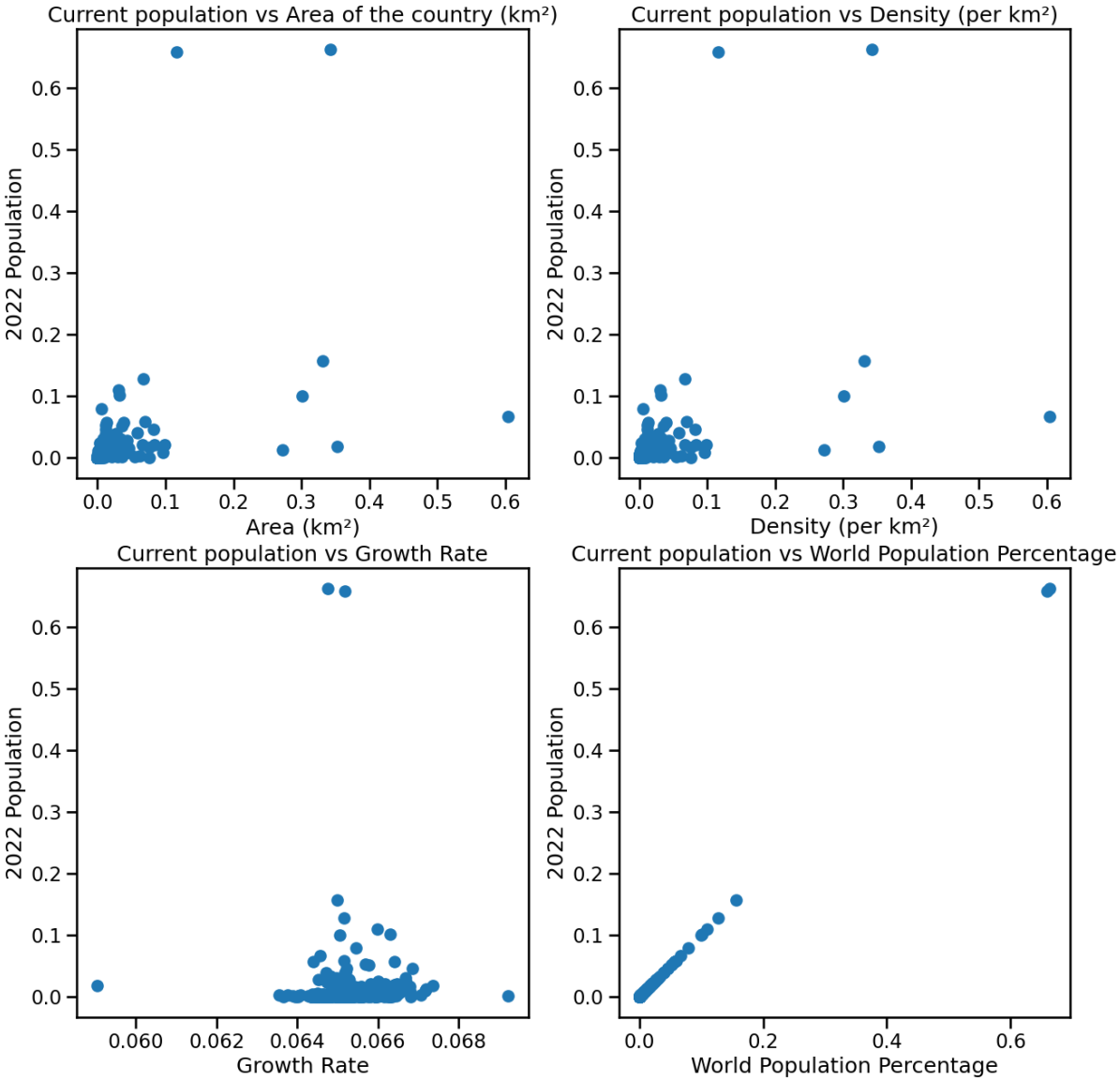
Corelations and Hypothesis

The below mentioned are some corelations, I could potentially figure from the data. More on this can be explored on further actual analysis.

- This dataset is the historical representation of the population of different countries. Thus, we correlate the population with the country name.
- We can also corelate the population with the geographical data such as the density and area of the nation.
- We can determine each country's population as a percentage of total population in the world and rank the countries accordingly.
- A dynamic plot describing the country ascending or descending can be plotted to better understand which country is taking population control seriously.
- It is also possible to discern the rate of population growth in each country over the years.

From the above corelation, I could figure that multiple hypothesis can be established, not limited to discerning the probability that a particular country could be the leading country in population, by simply making a linear relation, where years will be the index and the population will be the output. This can be done by directly fitting the data with a linear corelation equation. So, we can extend the given data to any particular year in future and find the predicted population of that country, given that other factors are constant and not varying.

My analysis



The image is self-explanatory about what it is conveying to the audience, which when compared to the previous sections text will provide further insight.

Future plan and Additional Information

The above said subtopic discussed a proper linear hypothesis. But this data can also be fitted with non-linear equation with degree of 2 or 3. But such fitting is quite arbitrary and we need to evaluate the effectiveness of the said equation corelation. This can be the future prospect with respect to this dataset.

Given everything, the dataset provides, we also need to understand if the above discussed model can even be made more effective and near to perfection. This can only be done by getting more data about the population of the country over many years. I do know that World War 1 and 2, would have had a drastic impact on the population. Thus, it is recommended that the population data, after 1950-60, is needed specifically. Apart from the usual geographical data, we also need other details such as the economical rankings of the countries over the years and many other similar rankings, to perceive if any external factors are influencing the actual population of a country.

Motivation to choose this dataset

We know in our current world; population is the main concern. For many decades, the whole world has been pointing at India and China for their outrageous population, without considering if other countries are following the same cue. This dataset, with the help of Sensex data, showcases the population of countries at different countries and tries to point out which country is having outrageous population growth in the world as of 2022. Apart from the above-mentioned reason, it also plays vital role in making us data scientists discern if population growth or decline has something do with the physical geographical factors like area and density.

References

1. https://www.kaggle.com/datasets/iamsouravbanerjee/world-population-dataset
2. https://worldpopulationreview.com/