**GLOBAL DATASET 2023**

**Introduction**

For the Capstone Project of the Coding Temple Data Analytics Program, I have selected Global Country Information Dataset 2023 analysis case study.

In order to answer key business questions, I will follow the steps of the data analysis process taught in the courses, which can be used in many other projects as well: **Ask, Prepare, Process, Analyze, Share .**Data cleaning, validation and initial exploration of datasets done using R

* Further analysis done using Excel and R in a Excel Worksheet and R notebook
* Data visualization done using Tableau Public

# **Background**

In this case study, I would analyse the Global Country Information Dataset 2023 and investigate the relationship between countries and the agricultural land which would provide more information about food security . I would analyse healthcare metrics such as infant mortality and life expectancy to assess overall well-being and I would be investigating educational enrolment rates,gdp and unemployment rates . Finally, I would be analysing the defence and military power of countries across the world .

I would like to check which countries have a high life expectancy ,agricultural land,tertiary educational enrolment rates ,infant mortality ,unemployment rates and ,gdp and military power

Hypothesis 1:

**Question**

Is there a correlation between birth rate and infant mortality

**Null Hypothesis**

There is no correlation between birth rate and infant mortality

**Alternate Hypothesis**

There is correlation between birth rate and infant mortality

**Test**

A correlation test would be performed and then reject or fail to reject one of the hypothesis

**Conclusion**

Analyse the results and draw conclusions depending on whether the hypothesis is rejected or not

Hypothesis 2:

**Question**

Is there a correlation between birth rate and life expectancy

**Null Hypothesis**

There is no correlation between birth rate and life expectancy

**Alternate Hypothesis**

There is correlation between birth rate and life expectancy

**Test**

A correlation test would be performed and then reject or fail to reject one of the hypothesis

**Conclusion**

Analyse the results and draw conclusions depending on whether the hypothesis is rejected or not

Hypothesis 3:

**Question**

Is there a correlation between population and armed forces size

**Null Hypothesis**

There is no correlation between population and armed forces size

**Alternate Hypothesis**

There is correlation between population and armed forces size

**Test**

A correlation test would be performed and then reject or fail to reject one of the hypothesis

**Conclusion**

Analyse the results and draw conclusions depending on whether the hypothesis is rejected or not

**Prepare**

Data Source:

<https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023>

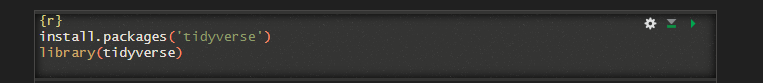
Data Organization:

The excel file consists of 195 rows and 35 columns

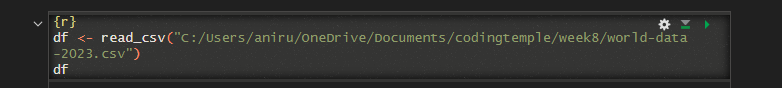
**Process/Tidy**

I have used R for data validation and cleaning

* Installed tidyverse package



* Read the data from .csv to .rmd format and created a data frame



* Removed columns which were not necessary for my analysis and had a lot of missing data



* Created a new data frame df1

A screenshot of a computer

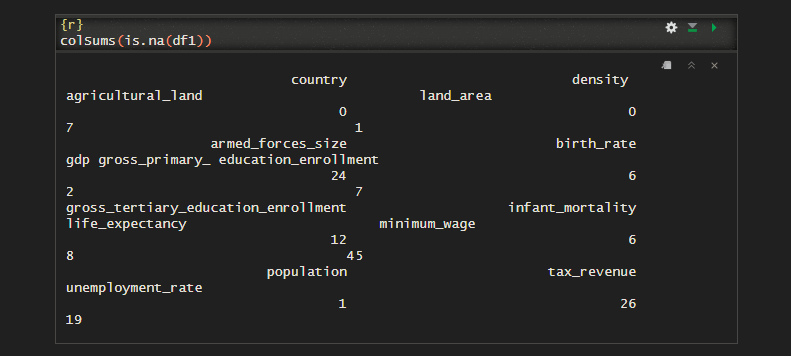
Description automatically generated

* Renamed column names

A screen shot of a computer screen

Description automatically generated

* Checking number of null values in each column

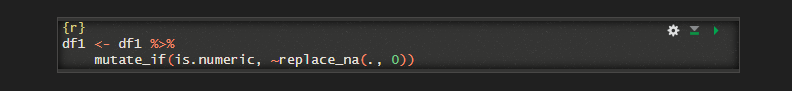


* Sum of the number of null values in the dataframe df1

A black rectangular object with a white stripe

Description automatically generated

* Replaced all the null values of numeric columns to zero



* Null values in all the numeric columns have been removed

A screenshot of a computer

Description automatically generated

* Exported to a .csv file



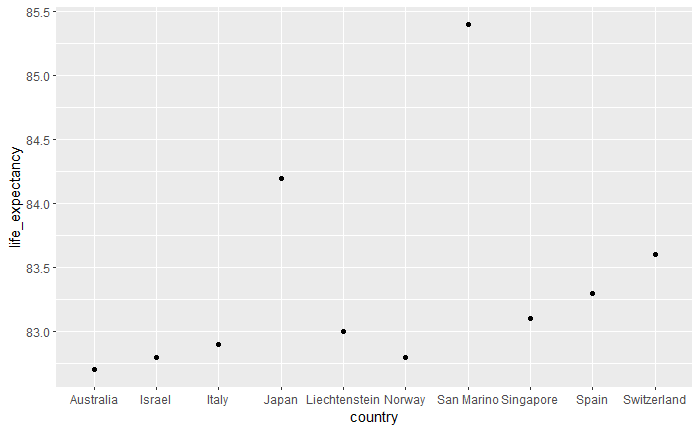
**Explore**

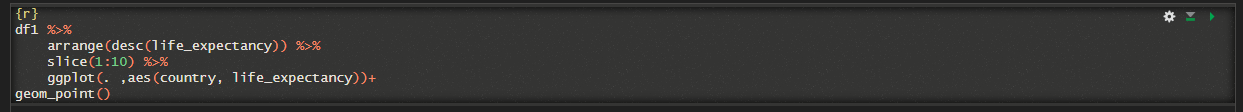
* Exploring descriptive statistics for all numeric columns of the data frame

A black screen with white text

Description automatically generated

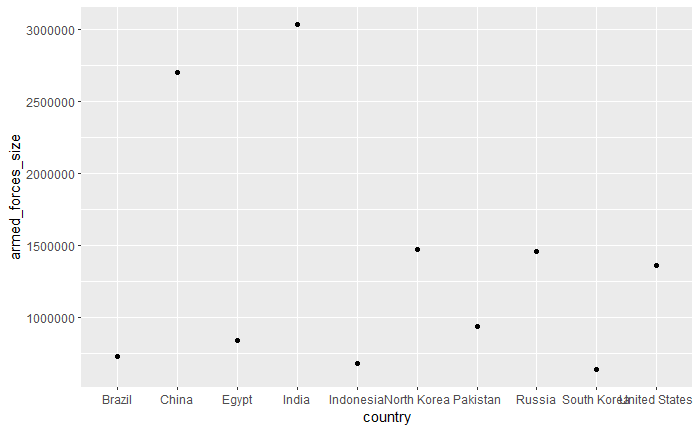
* Top ten countries having highest life expectancy

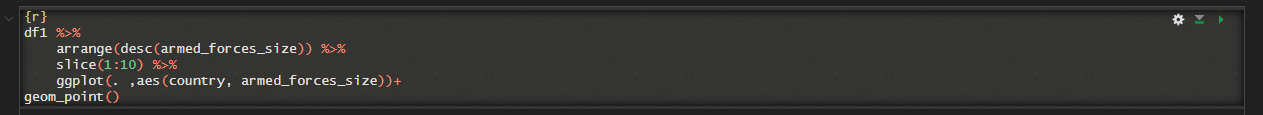




As per the scatter plot, San Marino has the highest life expectancy with an average of around 85.4 years .

* Top ten countries having the largest armed forces size





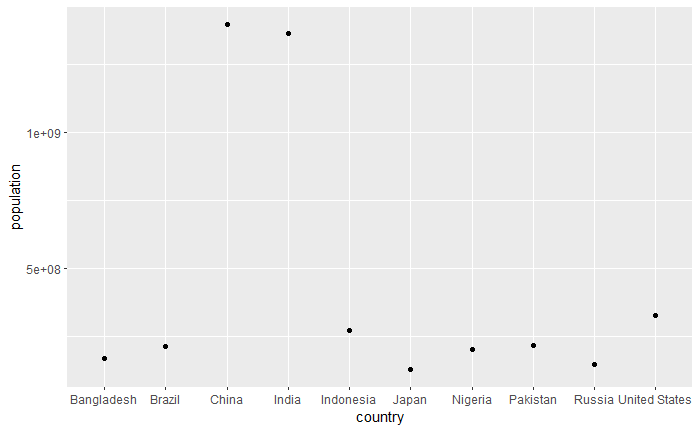
India has the largest armed forces size of 3 million active personnel followed by China which has around 2.7 million active personnel

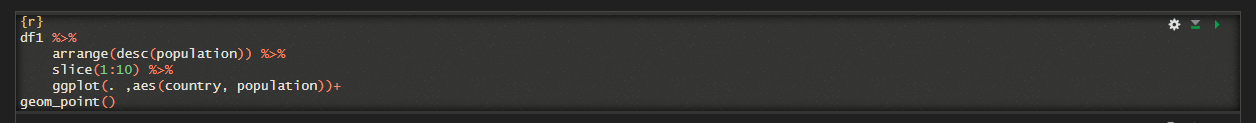
* Top ten countries with the highest unemployment rate

|  |  |
| --- | --- |
| **Country** | **Percentage of unemployment\_rate** |
| South Africa | 28% |
| Lesotho | 23% |
| Saint Lucia | 21% |
| Namibia | 20% |
| Gabon | 20% |
| Saint Vincent and the Grenadines | 19% |
| Libya | 19% |
| Bosnia and Herzegovina | 18% |
| Botswana | 18% |
| Greece | 17% |

As per the bar graph above, South Africa has the highest unemployment rate which is at 28% followed by Lesotho at 23%

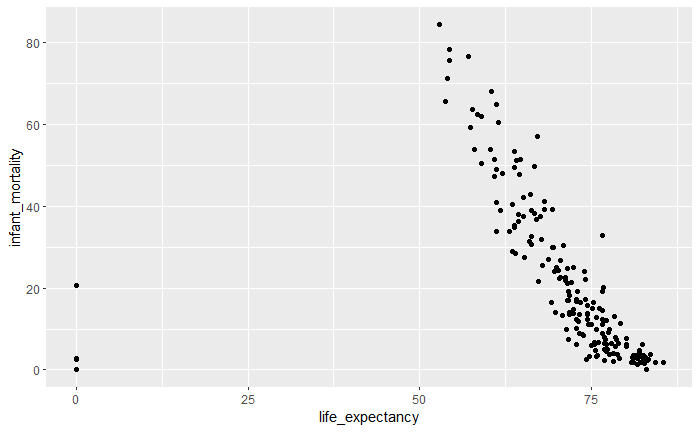
* Top ten countries with the highest population

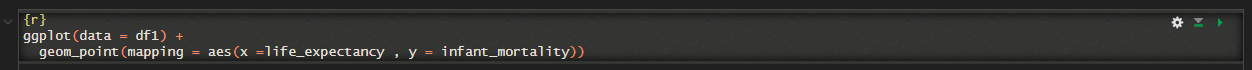




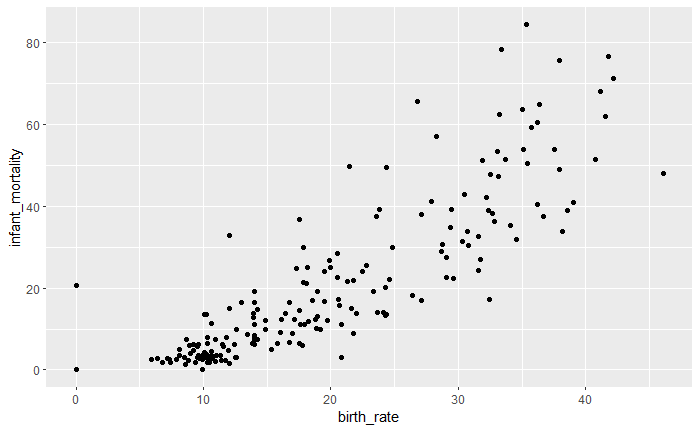
China has the highest population in the world with a little over one billion followed by India

Infant mortality vs Life expectancy





Birth rate vs Infant mortality





Country by % of agricultural land

|  |  |
| --- | --- |
| **Country** | **Average of agricultural land** |
| Uruguay | 83% |
| Saudi Arabia | 81% |
| Kazakhstan | 80% |
| South Africa | 80% |
| Burundi | 79% |
| Nigeria | 78% |
| Lesotho | 78% |
| El Salvador | 76% |
| Syria | 76% |
| Eritrea | 75% |

As per the pivot table, Uruguay has the highest percentage of agricultural land in the world at 83% followed closely by Saudi Arabia at 81%

Percentage of population enrolled in armed forces by country

|  |  |  |  |
| --- | --- | --- | --- |
| **country** | **armed forces** | **population** | **% enrolled** |
| North Korea | 1469000 | 25666161 | 572% |
| Eritrea | 202000 | 6333135 | 319% |
| Israel | 178000 | 9053300 | 197% |
| Montenegro | 12000 | 622137 | 193% |
| Brunei | 8000 | 433285 | 185% |
| Laos | 129000 | 7169455 | 180% |
| South Sudan | 185000 | 11062113 | 167% |
| Armenia | 49000 | 2957731 | 166% |
| Belarus | 155000 | 9466856 | 164% |
| Sri Lanka | 317000 | 21803000 | 145% |

As per the pivot table above, North Korea has the highest enrollment percentage when it comes to their population enrolling in the armed forces

**Analyze**

In this step, we will perform statistical tests that will test the relationships between various quantitative and continuous numerical data .

We would be using Pearson’s correlation testing to test various hypothesis , so that we get a clear understanding of the correlation or relationship strength between two variables

Pearson’s correlation test between infant mortality and birth rate

A screenshot of a computer

Description automatically generated

* The p- value is 2.2e-16
* r value or correlation coefficient value is 0.8583599
* This means the p-value is a lot less than 0.05, which means we can reject the null hypothesis β = 0. Hence, there is a significant relationship between the variables
* +.70 or higher - Very strong positive relationship or correlation
* Therefore infant mortality and birth rate have a very strong positive relationship or correlation

Pearson’s correlation test between life expectancy and birth rate

A black screen with white text

Description automatically generated

* This means we retain the null hypothesis and reject the alternative hypothesis as p - value is greater than 0.05.
* No or negligible relationship between both the variables because correlation coefficient is -0.0512559

Pearson’s correlation test between population and armed forces size

A black screen with a white text

Description automatically generated

* The p-value is 2.2e-16
* r value or correlation coefficient value is 0.8583599
* This means the p-value is a lot less than 0.05, which means we can reject the null hypothesis that β = 0. Hence, there is a significant relationship between the variables
* +.70 or higher - Very strong positive relationship or correlation
* Therefore population and armed forces size have a very strong positive relationship or correlation

**Share**

Here are dashboards of visualizations of different quantitative variables/data which are either continuous or discrete in nature . These visualizations would provide a clear understanding of relationships between different variables

A screenshot of a graph

Description automatically generated

A blue and white chart

Description automatically generated with medium confidence

The links to the Tableau Dashboards are listed below which would provide more information on the relationships between different variables

<https://public.tableau.com/app/profile/anirudh.r3844/viz/global_data_dashboard1/Dashboard1?publish=yes>

<https://public.tableau.com/app/profile/anirudh.r3844/viz/global_data_dashboard2/Dashboard2?publish=yes>

**Conclusion**

* San Marino has the highest life expectancy with an average of around 85.4 years
* Uruguay has the highest percentage of agricultural land in the world at 83%
* China has the highest population in the world with a little over one billion
* India has the largest armed forces size of 3 million active personnel
* South Africa has the highest unemployment rate which is at 28%
* North Korea has the highest enrollment percentage when it comes to their population enrolling in the armed forces
* Infant mortality and birth rate have a very strong positive relationship or correlation
* Population and armed forces size have a very strong positive relationship or correlation