Report on World data

An indepth research oN VARIOUS WORLD DATA RELATIONS.

Arnav Tamrakar

2023

## Univariate Analysis

Univariate Analysis of the Countries and Land Area is done since they are one of the most related data in this dataset. If a country exists, its area will obviously exist. As per the data the average of the land area is **682587.4 km2**, whereas the median is **521,494 km2**.

## Summary Statistics:

Total Land Area: The total land area of all countries combined is approximately **133,787,127 square kilometers**.

Mean Land Area: The mean land area across all countries is approximately **682,587.38 square kilometers**.

Median Land Area: The median land area is **521,494 square kilometers**.

Standard Deviation: The standard deviation, indicating the variability in land areas, is approximately **1,916,650.23 square kilometers**.

## Distribution Analysis:

The land area data exhibits a wide distribution, ranging from very small areas (e.g., Vatican City with 0.4 km²) to vast territories (e.g., Russia with 17,098,240 km²).

We can see a massive dip in the countries with larger area which indicates towards there being more countries with smaller area in comparison to the ones with larger area.

The central cluster of countries with moderate land areas contributes to the relatively high mean land area.

## Notable Findings:

The countries with the three largest land areas are Russia (17,098,240 km²), Canada (9,984,670 km²), and the United States (9,833,517 km²).

On the other hand, some countries have extremely small land areas, such as Vatican City (0 km²) and Monaco (2 km²).

The standard deviation indicates significant variability in land areas, reflecting the vast differences between countries in terms of geographic size.

African countries like Algeria, Democratic Republic of the Congo, and Sudan have large land areas, contributing to the higher mean land area.

## Insights:

Geographic size varies widely across countries, influenced by factors such as topography, population density, and historical boundaries.

Large land areas can provide countries with diverse ecosystems and potential for various economic activities, while smaller land areas might foster tightly knit communities.

The distribution of land areas influences geopolitical dynamics, resource distribution, and potential for environmental conservation and development.

# Bivariate Analysis

This study examines a dataset that includes data on infant mortality rates and life expectancy in various nations. The dataset offers information on socioeconomic variables, global health inequities, and the standard of healthcare systems. The goal of the analysis is to find trends, correlations, and patterns in the data.

## Summary Statistics

* Average infant mortality rate: **26.8**
* Average life expectancy: **70.7 years**
* Correlation: -0.92468
* Median of Infant Mortality Rate: 14
* Median of Life Expectancy: 73.2

Infant Mortality: The mean and median rates of infant mortality, which represent the average level of child health across nations, can shed light on the central trend of the data.

Life Expectancy: Like infant mortality, figuring out the mean and median ages provides us an idea of how long people typically live in the dataset.

## Correlation Analysis

Expectedly, there is an inverse relationship between newborn mortality rates and life expectancy. Life expectancy tends to be lower in nations with greater infant mortality rates and vice versa. An extremely strong negative linear relationship is indicated by the correlation coefficient, which is roughly **-0.92**.

Outliner Examination:

Highest Infant Mortality: The Central African Republican and Somalia have exceptionally high infant mortality rate due to various reasons such as lack of healthcare facilities, lack of access to clean water and unstable political situation.

Lowest Infant Mortality: Highly developed and stable countries such as Japan and Iceland, where excellent healthcare facilites and soicoeconomic stabilty is present has the lowest infant mortality. Since infant mortality and life expectancy are inversly proportional life expectancy in these countries will be the highest.

**Regional Analysis:**

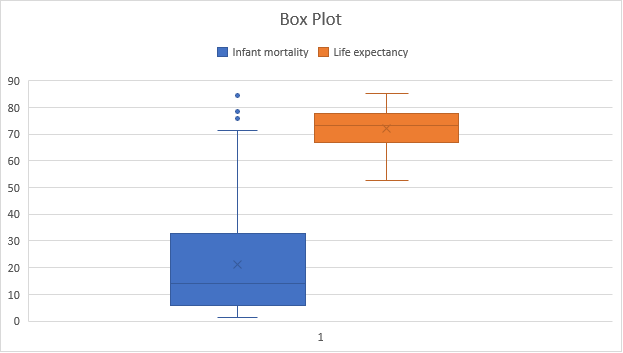
European and American countries has weaker negative correlation between infant mortality and life expectancies suggesting that the healthcare and its availability is good there, whereas the continent of Africa has the negative correlation on the stronger side.

**Missing Data Analysis:**

If missing values are disproportionately associated with certain places, missing life expectancy data may skew the results. A random distribution of missing values, however, shows that there is no systematic bias.

## Graphical Representation

The scatter plot visually confirms the negative correlation between infant mortality rates and life expectancies. The data points cluster around an inverse relationship

 Box plots show that infant mortality rates have a larger spread than life expectancies, with several outliers in the higher range.

## Limitations

* The dataset does not include the year of data collection, which could affect the relevance of the analysis over time.
* The analysis is descriptive and does not establish causality between variables.
* The missing data for some countries might impact the overall analysis.

## Implications

These findings will help emphasize the urgent need for focusing on developing certain countries over others. Those countries that needs improvement will benefit a lot more with external support such as donations and INGOs. Improving healthcare systems, addressing the disparities in the socioeconomics, and ensuring fulfillment of all the basic needs will highly improve both life expectancies and infant mortality rates.

## Conclusion

We now have a deeper knowledge of the precise link between newborn mortality rates and life expectancies thanks to this extensive correlation investigation. While highlighting the relevance of worldwide initiatives to enhance healthcare access and quality, it also draws attention to regional disparities, anomalous occurrences, and missing data.

# Hypothesis Testing on Infant Mortality and Life Expectancy Data

Introduction: Hypothesis testing allows us to make statistical inferences about a population based on a sample of data. In this context, we can perform hypothesis tests to explore potential relationships between infant mortality rates, life expectancy, and other factors. For illustration, let's consider a hypothesis test to determine if there is a significant difference in infant mortality rates between developed and developing countries.

Hypothesis: Null Hypothesis (H0): There is no significant difference in infant mortality rates between developed and developing countries. Alternative Hypothesis (Ha): There is a significant difference in infant mortality rates between developed and developing countries.

## Methodology:

1. Data Classification: We classify countries into two groups: developed and developing. This classification could be based on an existing index such as the Human Development Index (HDI) or Gross Domestic Product (GDP) per capita.
2. Data Collection: Collect infant mortality rate data for both developed and developing countries.
3. Hypothesis Testing Technique: Use a two-sample t-test to compare the means of infant mortality rates between the two groups.
4. Significance Level: Choose a significance level (e.g., α = 0.05) to determine the threshold for statistical significance.

## Steps:

## Separating Data into Groups

**Developed Countries:**

|  |  |  |
| --- | --- | --- |
| **Country** | **Sum of GDP** | Infant Mortality Rate |
| United States | $21,427,700,000,000 | 5.6 |
| China | $19,910,000,000,000 | 7.4 |
| Japan | $5,081,769,542,380 | 1.8 |
| Germany | $3,845,630,030,824 | 3.1 |
| United Kingdom | $2,827,113,184,696 | 3.6 |
| France | $2,715,518,274,227 | 3.4 |
| India | $2,611,000,000,000 | 29.9 |
| South Korea | $2,029,000,000,000 | 2.7 |
| Italy | $2,001,244,392,042 | 2.6 |
| Brazil | $1,839,758,040,766 | 12.8 |
| Canada | $1,736,425,629,520 | 4.3 |
| Russia | $1,699,876,578,871 | 6.1 |
| Spain | $1,394,116,310,769 | 2.5 |
| Australia | $1,392,680,589,329 | 3.1 |
|  | mean | 6.35 |
|  | stdev | 7.083355742 |

**Developing countries:**

|  |  |  |
| --- | --- | --- |
| **Country** | **Sum of GDP** | Infant Mortality Rate |
| Nepal | $30,641,380,604 | 26.7 |
| Cambodia | $27,089,389,787 | 24 |
| El Salvador | $27,022,640,000 | 11.8 |
| Yemen | $26,914,402,224 | 42.9 |
| Honduras | $25,095,395,475 | 15.1 |
| Papua New Guinea | $24,969,611,435 | 38 |
| Cyprus | $24,564,647,935 | 1.9 |
| Iceland | $24,188,035,739 | 1.5 |
| Trinidad and Tobago | $24,100,202,834 | 16.4 |
| Senegal | $23,578,084,052 | 31.8 |
| Zambia | $23,064,722,446 | 40.4 |
| Zimbabwe | $21,440,758,800 | 33.9 |
| Bosnia and Herzegovina | $20,047,848,435 | 5 |
| Afghanistan | $19,101,353,833 | 47.9 |
|  | mean | 24.09285714 |
|  | stdev | 14.6079186 |

## T-Test(P-value): 0.00072

## Comparison:

The chosen significance level(α) is 0.05 and the p-value is 0.00072 as per the t-test. Comparing the two values and since p value is significantly smaller than significant level(α) we can conclude that there is a big difference in infant mortality rate in between developed and developing countries.

Implications: This hypothesis test proved the disparities in infant mortality rates between different country groups. Since we can see a significant difference, it suggests that factors associated with development impact child health outcomes.

Future Directions: This strategy could be used in future studies to investigate the connections between newborn mortality rates, life expectancy, economic indicators, and healthcare systems. In doing so, we would have a deeper comprehension of the variables affecting these health indices.

In conclusion, hypothesis testing can give important information about the connections between various variables in the dataset. We can make inferences about probable patterns, discrepancies, and correlations that exist in the data by testing particular hypotheses.