**Project Report on World Statistics Dataset Analysis**

1. **Executive summary and purpose of report:**

As part of this project, we analyzed world population dataset of 261 countries, studied correlations between various datapoints and trends on how data has changed over past years. We used techniques like classification tree, cluster analysis, regression tree to draw predictive findings on country’s development state, agricultural growth, and life expectancy. Comparative study between developed and developing countries highlighted facts that ready availability of natural (land, water) and technological (electricity) resources, higher GDP, lower population density, lower birth and death rates, better utilization of land tend to help in economic growth and make a developed country. Developing countries, on the other hand, have higher dependency on agriculture as profession and as a means of income.

Below are the key takeaways from this report based on comparative study:

1. Form policies/strategies around agriculture, population, resources use. E.g.: reducing population density and improving availability of electricity in developing countries.
2. Provide recommendations to governments based on few focus points for improvements. E.g.: Very high infant mortality rate in developing countries by improving medical facilities.
3. **Motivation and overview of data:**

**Motivation:** World data analysis is extremely important in urban and regional planning all over the world. Factors like GDP, population, economic growth, employment, birth and death rates, land utilization can be utilized in planning and implementation decisions in various countries over the world. Based on characteristics of localities, their specific needs can be understood, and steps can be undertaken by the governments to implement administrative policies.

**Overview of data:** Dataset published by World Bank contains 318,420 data points from 1960 till 2020 for total 261 countries data (52 developed and 209 developing) like population, access to electricity, mortality, birth rates, employment, agricultural land, GDP, life expectancy data.

1. **Data dictionary: Table. 1**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. No.** | **Column Name** | **Type** | **Comment** |
| 1 | Country Name | Text | Name of the country for which data is collected |
| 2 | Year | Number | Data belongs to which year |
| 3 | Development status | Text | Developed/Developing country classification |
| 4 | Access to Electricity | Number | % Population of the country having access to electricity |
| 5 | Access to Electricity, Rural | Number | % Rural Population having access to electricity |
| 6 | Agriculture land | Number | Land under Agriculture in sq. km. |
| 7 | Agricultural raw material exports | Number | % Of Raw material exports |
| 8 | Agricultural raw material imports | Number | % Of Raw material imports |
| 9 | Birth rate, crude | Number | Per 1000 people |
| 10 | Death rate, crude | Number | Per 1000 people |
| 11 | Employment in agriculture | Number | % Of Total employment |
| 12 | Employment in agriculture, female | Number | % Of Total Female employment |
| 13 | Employment in agriculture, male | Number | % Of Total Male employment |
| 14 | Forest area | Number | Area of land under forests, in sq. km. |
| 15 | GDP Per capita | Number | Total GDP/Total population |
| 16 | Land area | Number | Total land area of country in sq. km. |
| 17 | Mortality rate, infant | Number | Infant mortality rate, per 1000 |
| 18 | Population, total | Number | Total population of the country |
| 19 | Population Density | Number | Total population/Total land area in sq. km. |
| 20 | Life Expectancy | Number | How long do people live in the country in years |

1. **Data cleaning needs, steps, and methods:**
2. Case wise data deletion: Initial years data for Employment in agriculture, both female, male, Access to electricity was found missing for few records. Without impacting mean and standard deviation values, we ignored missing rows and continued analysis using data.
3. Mean substitution technique: In the columns, Agricultural land and Agricultural imports and exports, we found intermediate values in between missing at some places. Replaced the intermediate values with the mean value of the overall trend.
4. Last observation carried forward: In columns like Access to electricity, Birth and Death Rate, replaced missing value at the end of the year range with the last observation (2019-2020).
5. Deletion of completely missing data at random: Deleted Average precipitation and Rural land area columns due to randomly missing data which couldn’t be utilized for analysis.
6. Additional columns introduced: Added life expectancy column to analyze correlation with birth, death, and mortality rates and prediction of average life expectancy in future.
7. **Methodology:**
8. **Classification tree to predict future development status of country:** Classification tree is used since outcome is categorical variable. We used classification trees to layout the factors that impact the development state of country and predicted the future development state using the input variables: 1) GDP per capita (quantitative), 2) population density (quantitative), 3) access to electricity (quantitative) 4) year (quantitative), outcome variable being Development status (qualitative) having 2 possible outcomes as developing (0) or developed (1). We used JMP from SAS tool to build the classification tree. Refer **Fig. A.1 to A.4 in attached Appendix document.**
9. **Cluster analysis to understand and highlight agricultural trends:** Cluster analysis is used to find natural groupings, or clusters within data without prespecifying groups of countries. We used cluster analysis for deeper analysis on agricultural data and understanding how the agricultural advancements, imports, exports differ between developed and developing countries. We used JMP from SAS and Minitab tools to build the classification tree. We divided the analysis in 3 years, with gap of 2 decades: 1975, 1995 and 2015, to understand changing trends over 6 decades. Refer **Fig. A.5 to A.10 in attached Appendix document.**
10. **Regression tree analysis for prediction of life expectancy:** Regression tree creates a binary tree by splitting the data on the predictor values recursively. We used regression tree technique to predict the average life expectancy (outcome is continuous) using inputs 1) Birth rate, 2) Death Rate, 3) Infant mortality rate, 4) Year. The regression tree is built using Minitab tool. Refer **Fig. A.11 to A.14 in attached Appendix document.**
11. **Results and caveats**
12. **Preliminary and exploratory analysis observations:**
13. On basisof % of people having ‘Access to Electricity’, developed countries have close to 100% people, developing countries have between 50% to 100% people who have the access. Higher value of resource availability promotes development. Refer **Fig.** **1**
14. Birth rate, death rate and mortality rate of the world overall are falling over years indicating the world overall has been improving over the medical facilities. Birth and death rates in developed countries is half of that in developing. Refer **Fig. 2**
15. Infant mortality rates in developing countries are worrisome as those are as high as 60%, which indicates that developing countries can work on this area.Refer **Fig. 3**
16. GDP per capita is an indicator of better economic status and growth & development of country. Currently, average GDP per capita of the developed countries is close to 3.5 times of that of developed countries. Refer **Fig. 4**
17. Developed countries have more agricultural land, higher population who are employed in agriculture as profession, including men and women. Developed countries are usually not proficient or reliant on agriculture. Refer **Fig. 5**
18. Life expectancy is improving, increased up to 79- and 70-years respectively in developed and developing countries. Refer **Fig. 6**

Chart, bar chart, waterfall chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

**Fig. 1** **Access to electricity (% population)** **Fig. 2 Birth Rate (per 1000 people)**

Chart, bar chart

Description automatically generated Chart, line chart

Description automatically generated

**Fig.3** **Death & mortality Rate (per 1000 people)** **Fig. 4 GDP per capita (in USD)**

Chart, table

Description automatically generatedChart, bar chart

Description automatically generated

**Fig.5** **Employment in agriculture** **Fig.6** **Life expectancy**

1. **Classification tree ‘development state’ prediction**: The analysis using classification tree indicates that access to electricity and GDP/capita are very important factors for development. Developed countries have more than 99.95% population having access to electricity. In general, for a country to be Developed, they should have higher GDP per capita than USD 13,000 and low population density, usually than 1700. Refer **Fig. 7**

Graphical user interface, application

Description automatically generated**Fig. 5** Terminal**Fig. 7 Leaf Nodes with outcome as: ‘Developed’**

Graphical user interface, application, table, Excel

Description automatically generatedMisclassification rate of the CTA is low at 11%. Observed outliers: Developed countries with very high population density (>1500 people per sq. km.) are shown. They are developed despite high population density, due to stronger economic backing. Refer **Fig. 8**

-Singapore (7950 people per sq. km.)

-Hongkong (7096 people per sq. km.)

-Malta (1514 people per sq. km.)

-Bahrain (2017 people per sq. km.)

**Fig. 8 Developed countries population density outliers**

**Prediction output for India in 2050**: currently a developing country:

**Below are projected values as per sources found.** Refer **Table. 2** below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **GDP/capita** | **Population Density** | **Access to electricity %** | **Year** | **Prediction as per CTA** | **Comment** |
| USD 53,000 | > 500 | 100% | 2050 | Developed | \* India is projected to be the world's second-largest economy by 2050 \* |

1. **Cluster analysis ‘Agricultural observations and grouping’:** Observations indicate that agriculture continues to be focus on developing countries and not much in developed countries. Usually, developed countries have very low land area under agriculture. Average agricultural imports are 2% for developed and 5-7% for developing. Average agricultural exports are 2% for developed and 7-8% for developing.

From the list of developing countries, India & Brazil form a unique cluster with high available agricultural lands and imports and exports, along with Kazakhstan, Russia, Saudi Arabia. We can conclude that these are agriculture proficient countries. (Agricultural land > 2,700,000 sq.km., exports > 5% in 2015).

However, we came across below outliers, identifying that the countries USA, Australia & China are developed countries that have higher land under agriculture, imports, and exports, which is not a common trend in developed countries. Australia has the highest amounts of agricultural exports than most of the developed countries (3%). China has the highest amounts of agricultural imports than most of the developed countries (4%). **Refer Fig. 9 and 10** below.

Graphical user interface, application, table, Excel, PowerPoint

Description automatically generated Graphical user interface, application, table, Excel, PowerPoint

Description automatically generated

**Fig. 9 Agriculture in developed countries - 1975 Fig. 10 Agriculture in developed countries - 2015**

1. **Regression tree technique observations**: Results of the regression tree analysis indicate that developed countries have maximum life expectancy more than 80 years, which is improved basis lower death and mortality rates. Increasing value of life expectancy indicates better medical facilities and overall improvement in the life expectancy of mankind. In developing countries, maximum value is closer to 76 years. Regression tree provides the results with 90% adjusted R-square values. Below **Fig.** **9** and **Fig. 10** provide forecast for next 30 years highlighting that life expectancy values in 2018 for all the countries. 5 developing countries with lowest life expectancy in 2018 (< 55 years)

* Central African Republic, Chad, Lesotho, Nigeria, Sierra Leone

**Graphical user interface, table, Excel

Description automatically generatedGraphical user interface, application, table, Excel

Description automatically generated**

**Fig. 9 Life expectancy: developed countries**  **Fig. 10 Life expectancy: developing countries**

**Leaf report from Regression tree analysis: Table 3 and 4 below**

|  |  |
| --- | --- |
| **Leaf rule** | **Mean life expectancy years** |
| Birth rate <= 15.3515, Death rate <= 3.802, 8.55 < Mortality rate <= 49.9 | 42.44 |
| Birth rate <= 13.2695, Death rate <= 4.4495, 4.8 < Mortality rate <= 8.55 | 48.96 |
| 82.7 < Mortality rate <= 83.9 | 50.00 |
| Mortality rate > 83.9 | 59.41 |
| Birth rate <= 13.2695, Death rate <= 4.4495, Mortality rate <= 4.8 | 59.46 |
| Birth rate > 30.605, 49.9 < Mortality rate <= 82.7 | 62.39 |
| Birth rate <= 30.605, 49.9 < Mortality rate <= 82.7 | 66.23 |
| Death rate > 3.802, 20.65 < Mortality rate <= 49.9 | 69.58 |
| Death rate > 3.802, 13.85 < Mortality rate <= 20.65 | 71.54 |
| Birth rate > 15.3515, Death rate <= 3.802, 8.55 < Mortality rate <= 49.9 | 73.24 |
| Death rate > 3.802, 8.55 < Mortality rate <= 13.85 | 73.90 |
| Death rate > 11.15, 4.45 < Mortality rate <= 8.55 | 74.73 |
| Birth rate > 13.2695, Death rate <= 4.4495, Mortality rate <= 8.55 | 76.47 |
| 4.4495 < Death rate <= 11.15, 4.45 < Mortality rate <= 8.55 | 77.42 |
| Death rate > 4.4495, Mortality rate <= 4.45 | 80.35 |

**Table 3 - Leaf report – prediction of life expectancy for developed countries**

|  |  |
| --- | --- |
| **Leaf rule** | **Mean life expectancy years** |
| Death rate > 37.4215, Mortality rate > 64.85 | 20.32 |
| 23.978 < Death rate <= 37.4215, 64.85 < Mortality rate <= 226 | 39.10 |
| 20.7415 < Death rate <= 23.978, 64.85 < Mortality rate <= 226 | 43.72 |
| 18.275 < Death rate <= 20.7415, Mortality rate > 64.85 | 46.55 |
| 16.0365 < Death rate <= 18.275, Mortality rate > 64.85 | 49.86 |
| 14.035 < Death rate <= 16.0365, Mortality rate > 64.85 | 52.56 |
| 11.9695 < Death rate <= 14.035, Mortality rate > 64.85 | 55.28 |
| Death rate > 10.123, 34.25 < Mortality rate <= 64.85 | 57.96 |
| 10.176 < Death rate <= 11.9695, Mortality rate > 64.85 | 58.26 |
| 20.7415 < Death rate <= 37.4215, Mortality rate > 226 | 58.81 |
| Death rate <= 10.176, Mortality rate > 64.85 | 62.21 |
| 8.307 < Death rate <= 10.123, 34.25 < Mortality rate <= 64.85 | 62.41 |
| 6.703 < Death rate <= 8.307, 34.25 < Mortality rate <= 64.85 | 65.24 |
| Death rate <= 6.703, 34.25 < Mortality rate <= 64.85 | 68.10 |
| 23.85 < Mortality rate <= 34.25 | 69.25 |
| 15.65 < Mortality rate <= 23.85 | 71.30 |
| 10.05 < Mortality rate <= 15.65 | 72.92 |
| Mortality rate <= 10.05 | 76.06 |

**Table 4- Leaf report – prediction of life expectancy for developing countries**

1. **Conclusions and implications**

Comparative study between the developed and developing countries revealed the below.

1. Higher GDP, ready availability of resources like electricity to majority of population, lower population of country contribute to economic growth and development of country as the resources shared between lesser population, increase per head usability.
2. In general, ~ 100% population having electricity access, GDP higher than USD 13,000 and population density lower than 1700 identify a developed country.
3. Birth and death rate improvements suggest enhanced medical facilities have helped these figures progress all over the world, at a faster pace in developed countries. Average life expectancy trend in the world is improving, which indicates evolution of public healthcare.
4. Developing countries should focus to improve on infant mortality rates by providing better medical facilities, developed countries should have moral obligation to help developing countries.
5. Developing countries have Agriculture as a major population which also provides employment to population. They also have higher agricultural imports and exports. Developed countries have lesser reliance on agriculture as a profession, with exceptions of USA, Australia, China who also are agricultural proficient countries, being developed.

**Potential future research:**

1. It appears that the high life expectancy, low population, and high GDP are relevant factors, relating welfare state, the economic conditions, and the demographic situation together. More research in these areas could potentially correlate these factors and provide recommendations to developing countries to enable growth and development.
2. Population analysis can be done at a detailed level, adding migration data to correlate thesize, structure, and distribution of population. This could help in predicting the population trends in future and countries can be equipped with facilities in advance.
3. **References:**

* **Dataset Source:**

<https://www.kaggle.com/mutindafestus/world-statistics-dataset-from-world-bank>

* **Other data sources:**

<https://www.un.org/en/development/desa/policy/wesp/wesp_current/2014wesp_country_classification.pdf>