Data Immersion Achievement 6

**Levels and trends in child malnutrition**

Advanced Analytics & Dashboard Design

Contents

[Objective 4](#_Toc140302716)

[Why the analysis 4](#_Toc140302717)

[Key Questions and Objectives 4](#_Toc140302718)

[Scope 4](#_Toc140302719)

[Data source 5](#_Toc140302720)

[About the data and Data Types 8](#_Toc140302721)

[Data Profiling 19](#_Toc140302722)

[Data Analysis Criteria for this Project: 19](#_Toc140302723)

[Analysis Criteria 20](#_Toc140302724)

[Dashboard Requirements 21](#_Toc140302725)

[Analysis 21](#_Toc140302726)

[Top 10 countries that are most affected by Stunting 21](#_Toc140302727)

[Top 10 countries that are most affected by Wasting 22](#_Toc140302728)

[Top 10 countries that are most affected by Stunting and Wasting 23](#_Toc140302729)

[Correlation between Stunting, Wasting and Prevalence of low birth weight among new-borns 24](#_Toc140302730)

[Stunting, Wasting and Prevalence of low birth weight relationship with Income 26](#_Toc140302731)

[Pairplot on global severe wasting data 28](#_Toc140302732)

[Categorical plot depicting the severity wasting on global data among genders and total population 29](#_Toc140302733)

[Geospatial Analysis using Choropleth Map 30](#_Toc140302734)

[Geospatial Analysis on Stunting Data 30](#_Toc140302735)

[Average Stunting population observation values from 2001 to 2023 31](#_Toc140302736)

[Geospatial Analysis On Severe Wasting Data 31](#_Toc140302737)

[Supervised Machine Learning: Regression 32](#_Toc140302738)

[Linear Regression Analysis 33](#_Toc140302739)

Objective

The JME includes estimates of prevalence and numbers affected for stunting, overweight, wasting and severe wasting among children under five years of age at country, regional and global levels. Global, regional and country annual trends from 2000-2022 are available for stunting and overweight. The objective is to build an interactive dashboard that will visually showcase well-curated results of an advanced exploratory analysis conducted in Python to measure nutritional imbalance such as undernutrition (assessed from stunting, wasting and underweight) or overweight.

# Why the analysis

Every child has the right to good nutrition. Well-nourished children grow and develop to their full potential. They are better equipped to lead healthy lives, to be free from poverty, to learn and participate, and to continue thriving across the life course, with benefits that continue over generations. The past decade has seen important gains in improving maternal and child nutrition, including a one-third decline in the proportion of children suffering from stunting. Yet the triple burden of malnutrition – stunting, wasting and overweight – continues to jeopardize children’s ability to survive and thrive.

All forms of malnutrition are preventable. To stop malnutrition before it starts, children and their families need access to nutritious diets, essential services and positive practices to set them on the path to survival and thrive. But today, these vital pathways to good nutrition are under growing threat as many countries plunge deep into a global food and nutrition crisis fuelled by poverty, conflict, climate change and the enduring secondary effects of the COVID-19 pandemic. As the world responds to the crisis, urgent action is critical to protect maternal and child nutrition – especially in the most affected regions – and secure a future where the right to nutrition is a reality for every child.

# Key Questions and Objectives

* Which countries are affected most by child nutrition?
* Identify the number of deaths and affected due to mall nutrition.
* Does progress in reducing stunting equal across regions and sub-regions?
* Find the trend in the country’s income and malnutrition growth.
* SDG target by 2030, including achieving by 2025 the internationally agreed target to end all forms of malnutrition, stunting and wasting in children under five years of age. Will this target be achievable by the trends in the dataset?

Scope

SDG target by 2030, end all forms of malnutrition, including achieving, by 2025, the internationally agreed targets on stunting and wasting in children under five years of age.

# Data source

* The UNICEF, WHO and the World Bank inter-agency team update the Joint Child Malnutrition Estimates (JME) every other year.
* The JME includes estimates of prevalence and numbers affected for stunting, overweight, wasting and severe wasting among children under 5 years of age at country, regional and global level. Global, regional and country annual trends.
* Data Source: United Nations Children's Fund (UNICEF), World Health Organization (WHO), World Bank (WB)
* Data Collection: Data is collected various national organisation’s health survey departments. The list of organisations are defined here.

|  |  |
| --- | --- |
| Source | Short source for major survey types. |
| DHS | Demographic and Health Survey (from the official Demographic and Health Survey Programme available at <https://dhsprogram.com/>) |
| DHS-style or MICS-style survey | Survey with MICS or DHS in the titled but which is not and official MICS or DHS/not available at the above websites |
| LSMS | The Living Standards Measurement Study (from official World Bank Group supported surveys available at: <http://iresearch.worldbank.org/lsms/lsmssurveyFinder.htm>) |
| MICS | Multiple Indicator Cluster Survey (from official UNICEF-supported surveys available at <http://mics.unicef.org/>) |
| NNS | National Nutrition Survey |
| SMART | Standardized Monitoring and Assessment of Relief and Transitions |
| Nutrition Surveillance | Estimates originating from nutrition monitoring and surveillance systems which are nationally representative |
| ONS | Other National Survey |
| PAPFAM | Pan Arab Project for Family Health survey |

**Data Contents:** The JME includes estimates of prevalence and numbers affected for stunting, overweight, wasting and severe wasting among children under 5 years of age at country, regional and global level. Global, regional and country annual trends from 2000-2022 are available for stunting and overweight. Refer Data Profiling (section 1. About the data and data type) for more information.

* **Data Relevance:** The data is part UN’s Sustainable Development Goals (SDG) and accurate to their standards. Data can be used for study, research and analysis. Data contains relevant columns for the analysis and meets project objectives.
  + **Objective 1**: Identify the number of deaths and affected due to mall nutrition(stunting, overweight, wasting and severe wasting ).
  + **Hypothesis**: Malnutrition and mortality among children's counts reducing each year.
  + **Datasource** 
    - UNICEF\_Global\_Databases\_Mortality dataset contains mortality rate in each geographic regions with year.
      * Limitation: Data isn’t recorded every year.
    - UNICEF\_Global\_Joint-Malnutrition-Estimates contains global stunting, wasting and overweight information.
  + **Objective 2:** Which countries are affected most by child nutrition(stunting, overweight, wasting and severe wasting)?
  + **Hypothesis**: Countries with low income have the highest child malnutrition count.
  + **Datasource:**
    - UNICEF\_Global\_Joint-Malnutrition-Estimates
    - UNICEF\_Global\_Databases\_Overweight
    - UNICEF\_Global\_Databases\_Stunting
    - UNICEF\_Global\_Databases\_Databases\_Wasting

All the above table have enough columns to find the trends in child nutrition growth in each country with period over the years.

* + **Objective 3:** Does progress to reduce stunting has equal across regions and sub-regions?
  + **Hypothesis**: Progress to reduce stunting has been equal across regions and sub-regions.
  + **Datasource:** 
    - UNICEF\_Global\_Joint-Malnutrition-Estimates
    - UNICEF\_Global\_Databases\_Overweight
    - UNICEF\_Global\_Databases\_Stunting
    - UNICEF\_Global\_Databases\_Databases\_Wasting

All the above table have enough columns to find the trends in child nutrition growth in each country with period over the years.

* + **Objective 4:** Find the trend in countries income and malnutrition growth.
  + **Hypothesis**: countries income and malnutrition growth worldwide affected with countries income.
  + **Datasource:**
  + **Objective 5:** SDG target by 2030, end all forms of malnutrition, including achieving, by 2025, the internationally agreed targets on stunting and wasting in children under 5 years of age. Will this target achievable by the trends in dataset?
  + **Hypothesis**: More intensive efforts are required if the world is to achieve the global target of reducing the number of children with stunting to 89 million by 2030.
  + **Datasource:**
    - UNICEF\_Global\_Joint-Malnutrition-Estimates
    - UNICEF\_Global\_Databases\_Overweight
    - UNICEF\_Global\_Databases\_Stunting
    - UNICEF\_Global\_Databases\_Databases\_Wasting

All the above table have enough columns to find the trends in child nutrition growth in each country with period over the years.

* **Data Bias and Gaps:**

{Need more analysis}

* **Completeness of Data:**
  + The data is collected from different regions with different sources. The data collection format isn’t standardised.
  + Period of data collection isn’t the same. Some countries have data collection every year, and some are 2 or 3 years. And countries collected the data every decade.
  + Some countrie’s samples are segregated by gender, age, family status etc. Some country’s observations combined result in a percentage of the total sample.
  + Malnutrition data doesn’t contain all the 195 country’s data.

**Limitations and Ethics**

* Data is collected by UN from various countries reports.
* The data could be possibly politically biased to provide countries reputation.
* Collected data have clear transparency from which organisation is collected, in which year and survey timing. And the data can be accessed by any one globally.
* The data doesn’t contain any personal information or to target specific community.
* All the data is collected from the UNICEF WHO database and the data can be used to for research, study or analysis purpose.

## About the data and Data Types

|  |  |  |
| --- | --- | --- |
| **Wasting**: Child **wasting** is the life-threatening result of poor nutrient intake and/or recurrent illnesses. Children suffering from wasting have weakened immunity, are susceptible to long-term developmental delays and face an increased risk of death, particularly when wasting is severe. Children suffering from severe wasting require early detection, timely treatment, and care to survive. | | |
| Table:   * UNICEF\_Databases\_Severe\_Wasting * UNICEF\_Global\_Databases\_Databases\_Wasting | | |
| Column Name | Description | Type |
| ISO | Three-digit alphabetical codes International Standard ISO 3166-1 assigned by the International Organization for Standardization (ISO).  thttp://www.iso.org/iso/home/standards/country\_codes | Characters(3),  Qualitative- Nominal |
| Countries and areas | The UNICEF Global databases contain a set of 202 countries as reported on through the State of the World's Children Statistical Annex 2023 | Characters(20),  Qualitative- Nominal |
| UNICEF Regions | EAP: East Asia and the Pacific  EECA: Eastern Europe and Central Asia  ESA: East and Southern Africa  LAC: Latin America and the Caribbean  MENA: Middle East and North Africa  NA : North America  SA : South Asia  WCA: West and Central Africa  WE: Western Europe | Characters(4),  Qualitative- Nominal |
| UN Regions |  | Characters(4),  Qualitative- Nominal |
| WHO Regionns |  | Characters(4),  Qualitative- Nominal |
| World Bank Income Groups |  | Characters(4),  Qualitative- Nominal |
| Survey Years | Represents the year(s) in which the data collection (e.g. survey interviews) took place | Character  Qualitative – Ordinal |
| Year | Year assigned to each survey | Number  Quantitative-Continuous  Time variant |
| Short Source | Short source for major survey types.  Demographic and Health Survey (from the official Demographic and Health Survey Programme available at https://dhsprogram.com/) | Characters |
| Full Source Title |  | Characters |
| Latest Estimate |  | Characters |
| Estimate Type | The source of the estimate, whether as reported, reanalyzed, etc.  Reanalyzed : Dataset has been obtained, standardized and reanalyzed to conform to the definition. Lower and Upper Confidence limits are based on a 95% confidence level  External Reanalysis : Dataset reanalyzed by external partners  Reported: Point Estimate, and when available, other information including sample size, Lower and Upper Confidence limits, are based on results from report. | Characters  Qualitative- Nominal |
| National, Sex, Residence , Age Group , Wealth Quintile Wealth Quintile Grouping, Mother's Education, Area / Wealth Quintile, Area / Wealth Quintile Grouping Sex / Age Group .Subnational Region:  The above column contains all the estimate values for the above categories, For exaple National contains records Point Esimate, Lower Limit, Upper Limit, Sample Size and foot note.  **Value**:   * Percentage of children under 5 years of age falling below -3 standard deviations severe wasting from the median weight-for-height of the reference population * Percentage of children under 5 years of age falling below -2 standard deviations wasting from the median weight-for-height of the reference population  |  |  |  | | --- | --- | --- | | Point Estimate | The prevalence of the wasting (severe) in the population | Continuous-Discrete  Time variant | | Lower Limit | 95% lower confidence interval of the point estimate | Continuous-Discrete  Time variant | | Upper Limit | 95% upper confidence interval of the point estimate | Continuous-Discrete  Time variant | | Sample Size | The weighted sample size which the point estimate is based on | Continuous-Discrete  Time variant | | Footnote | A variety of notes that relate to the point estimates as follows:  a: Alternate definition (Numbered footnote contains details)  b: Age-adjusted estimate  c: Converted estimate  d: Adjusted National-Rural to National  p: Age calculated using century month codes  q: Oedema data were collected in the survey  r: Height/length modality (standing/lying) was not collected  s : Oedema data was collected in the survey and not considered in the analysis  y : Unweighted sample size is less than 25, point estimate, lower limit and upper limit have been suppressed  z : Unweighted sample size is between 25 and 49 | Qualitative- Nominal | | | |

|  |  |  |
| --- | --- | --- |
| **Stunting** is the devastating result of poor nutrition in-utero and early childhood. Children suffering from stunting may never attain their full possible height, and their brains may never develop to their full cognitive potential. These children begin their lives at a marked disadvantage with consequences continuing into adulthood: they face learning difficulties in school, earn less as adults, and face barriers to community participation. | | |
| **Table**:   * UNICEF\_Global\_Databases\_Stunting | | |
| **Column Name** | **Description** | **Type** |
| ISO | Three-digit alphabetical codes International Standard ISO 3166-1 assigned by the International Organization for Standardization (ISO).  thttp://www.iso.org/iso/home/standards/country\_codes | Characters(3),Qualitative- Nominal |
| Countries and areas | The UNICEF Global databases contain a set of 202 countries as reported on through the State of the World's Children Statistical Annex 2023 | Characters(20),  Qualitative- Nominal |
| UNICEF Regions | EAP: East Asia and the Pacific  EECA: Eastern Europe and Central Asia  ESA: East and Southern Africa  LAC: Latin America and the Caribbean  MENA: Middle East and North Africa  NA : North America  SA : South Asia  WCA: West and Central Africa  WE: Western Europe | Characters(4),  Qualitative- Nominal |
| UN Regions |  | Characters(4),  Qualitative- Nominal |
| WHO Regionns |  | Characters(4),  Qualitative- Nominal |
| World Bank Income Groups |  | Characters(4),  Qualitative- Nominal |
| Survey Years | Represents the year(s) in which the data collection (e.g. survey interviews) took place | Character  Qualitative - Ordinal |
| Year | Year assigned to each survey | Number  Quantitative-Continuous  Time variant |
| Short Source | Short source for major survey types.  Demographic and Health Survey (from the official Demographic and Health Survey Programme available at https://dhsprogram.com/) | Characters |
| Full Source Title |  | Characters |
| Latest Estimate |  | Characters |
| Estimate Type | The source of the estimate, whether as reported, reanalyzed, etc.  Reanalyzed : Dataset has been obtained, standardized and reanalyzed to conform to the definition. Lower and Upper Confidence limits are based on a 95% confidence level  External Reanalysis : Dataset reanalyzed by external partners  Reported: Point Estimate, and when available, other information including sample size, Lower and Upper Confidence limits, are based on results from report. | Characters  Qualitative- Nominal |
| National, Sex, Residence , Age Group , Wealth Quintile Wealth Quintile Grouping, Mother's Education, Area / Wealth Quintile, Area / Wealth Quintile Grouping Sex / Age Group .Subnational Region:  The above column contains all the estimate values for the above categories, For exaple National contains records Point Esimate, Lower Limit, Upper Limit, Sample Size and foot note   |  |  |  | | --- | --- | --- | | Point Estimate | The prevalence of the stunting (moderate and severe) in the population  Percentage of under-fives falling below minus 2 standard deviations (moderate and severe) from the median height-for-age of the reference population | Continuous-Discrete | | Lower Limit | 95% lower confidence interval of the point estimate  Percentage of under-fives falling below minus 2 standard deviations (moderate and severe) from the median height-for-age of the reference population | Continuous-Discrete | | Upper Limit | 95% upper confidence interval of the point estimate  Percentage of under-fives falling below minus 2 standard deviations (moderate and severe) from the median height-for-age of the reference population | Continuous-Discrete | | Sample Size | The weighted sample size which the point estimate is based on | Continuous-Discrete | | Footnote | A variety of notes that relate to the point estimates as follows:  a: Alternate definition (Numbered footnote contains details)  b: Age-adjusted estimate  c: Converted estimate  d: Adjusted National-Rural to National  p: Age calculated using century month codes  q: Oedema data were collected in the survey  r: Height/length modality (standing/lying) was not collected  s : Oedema data was collected in the survey and not considered in the analysis  y : Unweighted sample size is less than 25, point estimate, lower limit and upper limit have been suppressed  z : Unweighted sample size is between 25 and 49 | Qualitative- Nominal | | | |

|  |  |  |
| --- | --- | --- |
| **Overweight**: Childhood **overweight** occurs when children’s caloric intake from food and beverages exceeds their energy requirements. This form of malnutrition is driven by failing food systems characterised by poor affordability and access to nutritious foods, the marketing of nutrient-poor ultra-processed foods and inadequate opportunities for physical activity. | | |
| **Table**:   * UNICEF\_Global\_Databases\_Overweight | | |
| **Column Name** | **Description** | **Type** |
| ISO | Three-digit alphabetical codes International Standard ISO 3166-1 assigned by the International Organization for Standardization (ISO).  thttp://www.iso.org/iso/home/standards/country\_codes | Characters(3),Qualitative- Nominal |
| Countries and areas | The UNICEF Global databases contain a set of 202 countries as reported on through the State of the World's Children Statistical Annex 2023 | Characters(20),  Qualitative- Nominal |
| UNICEF Regions | EAP: East Asia and the Pacific  EECA: Eastern Europe and Central Asia  ESA: East and Southern Africa  LAC: Latin America and the Caribbean  MENA: Middle East and North Africa  NA : North America  SA : South Asia  WCA: West and Central Africa  WE: Western Europe | Characters(4),  Qualitative- Nominal |
| UN Regions |  | Characters(4),  Qualitative- Nominal |
| WHO Regionns |  | Characters(4),  Qualitative- Nominal |
| World Bank Income Groups |  | Characters(4),  Qualitative- Nominal |
| Survey Years | Represents the year(s) in which the data collection (e.g. survey interviews) took place | Character  Qualitative - Ordinal |
| Year | Year assigned to each survey | Number  Quantitative-Continuous  Time variant |
| Short Source | Short source for major survey types.  Demographic and Health Survey (from the official Demographic and Health Survey Programme available at https://dhsprogram.com/) | Characters |
| Full Source Title |  | Characters |
| Latest Estimate |  | Characters |
| Estimate Type | The source of the estimate, whether as reported, reanalyzed, etc.  Reanalyzed : Dataset has been obtained, standardized and reanalyzed to conform to the definition. Lower and Upper Confidence limits are based on a 95% confidence level  External Reanalysis : Dataset reanalyzed by external partners  Reported: Point Estimate, and when available, other information including sample size, Lower and Upper Confidence limits, are based on results from report. | Characters  Qualitative- Nominal |
| National, Sex, Residence , Age Group , Wealth Quintile Wealth Quintile Grouping, Mother's Education, Area / Wealth Quintile, Area / Wealth Quintile Grouping Sex / Age Group .Subnational Region:  The above column contains all the estimate values for the above categories, For exaple National contains records Point Esimate, Lower Limit, Upper Limit, Sample Size and foot note   |  |  |  | | --- | --- | --- | | Point Estimate | The prevalence of the wasting (severe) in the population | Continuous-Discrete | | Lower Limit | 95% lower confidence interval of the point estimate | Continuous-Discrete | | Upper Limit | 95% upper confidence interval of the point estimate | Continuous-Discrete | | Sample Size | The weighted sample size which the point estimate is based on | Continuous-Discrete | | Footnote | A variety of notes that relate to the point estimates as follows:  a: Alternate definition (Numbered footnote contains details)  b: Age-adjusted estimate  c: Converted estimate  d: Adjusted National-Rural to National  p: Age calculated using century month codes  q: Oedema data were collected in the survey  r: Height/length modality (standing/lying) was not collected  s : Oedema data was collected in the survey and not considered in the analysis  y : Unweighted sample size is less than 25, point estimate, lower limit and upper limit have been suppressed  z : Unweighted sample size is between 25 and 49 | Qualitative- Nominal | | | |

# Understanding the data

**GLOBAL\_DATAFLOW\_2000-2023 stunting -2SD.csv**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Geographic area | count | unique | top | freq | mean | std | min | 25% | 50% | 75% | max |
| Indicator | 3657 | 159 | Afghanistan | 23 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| Sex | 3657 | 1 | Height-for-age <-2 SD (stunting), Modeled Esti... | 3657 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| TIME\_PERIOD | 3657 | 1 | Total | 3657 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| OBS\_VALUE | 3657 | NaN | NaN | NaN | 2011 | 6.63 | 2000 | 2005 | 2011 | 2017 | 2022 |
| Unit multiplier | 3657 | NaN | NaN | NaN | 21.4 | 14.69 | 0.9 | 8.3 | 18.7 | 32.6 | 63.1 |
| Unit of measure | 3657 | 1 | Units | 3657 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| Observation Status | 3657 | 1 | % | 3657 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| Observation confidentaility | 3657 | 1 | Modelled | 3657 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| LOWER\_BOUND | 3657 | 1 | Free | 3657 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| ... | 3657 | NaN | NaN | NaN | 19.49 | 14.23 | 0.7 | 6.8 | 16.5 | 30.3 | 60.2 |
| WGTD\_SAMPL\_SIZE | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| OBS\_FOOTNOTE | 0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| SERIES\_FOOTNOTE | 247 | 3 | The estimates for the year 2020 do not account... | 155 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| DATA\_SOURCE | 3657 | 1 | UNICEF/WHO/World Bank Joint Malnutrition Estim... | 3657 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| SOURCE\_LINK | 3657 | 1 | UNICEF/WHO/World Bank Joint Malnutrition Estim... | 3657 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| CUSTODIAN | 3657 | 1 | https://data.unicef.org/topic/nutrition/malnut... | 3657 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| Time period activity related to when the data are collected | 3657 | 1 | UNICEF Division of Data Analysis Planning an... | 3657 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| REF\_PERIOD | 3657 | 1 | Other | 3657 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| COVERAGE\_TIME | 0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| Current age | 3657 | 1 | 2000-2022 | 3657 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
|  | 3657 | 1 | Under 5 years old | 3657 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |

**UNICEF\_Databases\_Severe\_Wasting.csv**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ISO | 0 | 1 | 2 | 3 | 4 |
| Countries and areas | AFG | AFG | AFG | ALB | ALB |
| UNICEF Regions | Afghanistan | Afghanistan | Afghanistan | Albania | Albania |
| UN Regions | SA | SA | SA | EECA | EECA |
| WHO Regions | Asia | Asia | Asia | Europe | Europe |
| World Bank Income Groups | EMRO | EMRO | EMRO | EURO | EURO |
| Survey Years | Low Income | Low Income | Low Income | Upper Middle Income | Upper Middle Income |
| Year\* | 2004 | 2013 | 2018 | 2005 | 2008-09 |
| Short Source | 2004 | 2013 | 2018 | 2005 | 2009 |
| Full Source Title | NNS | NNS | Other | MICS | DHS |
| ... | Summary report of the national nutrition surve... | Afghanistan National Nutrition Survey 2013 | Afghanistan Health Survey 2018 | Albania multiple indicator cluster survey 2005... | Albania demographic and health survey 2008-09.... |
| Female Point Estimate | ... | ... | ... | ... | ... |
| Urban Point Estimate | NaN | 3.5 | 1.4 | 4.2 | 4.3 |
| Rural Point Estimate | NaN | 1.9 | 1.6 | 2.2 | 6.4 |
| 0-5 months Point Estimate | NaN | 4.6 | 1.5 | 4.5 | 5.6 |
| 6-11 months Point Estimate | NaN | 6.3 | 2.8 | 7.4 | 12.7 |
| 12-13 months Point Estimate | NaN | 5.5 | 2.8 | 1 | 5.5 |
| 24-35 months Point Estimate | NaN | 4.8 | 2 | 0.9 | 5.3 |
| 36 - 47 Point Estimate | NaN | 3.5 | 1.4 | 5.2 | 5.2 |
| 48-59 Point Estimate | NaN | 2.5 | 0.7 | 3.7 | 5.9 |
| 24-59 Point Estimate | NaN | 2.8 | 0.7 | 3.8 | 5.1 |
|  | NaN | 3 | 0.9 | 4.2 | 5.4 |

**UNICEF\_Global\_Databases\_Overweight.csv**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ISO | 0 | 1 | 2 | 3 | 4 |
| Countries and areas | AFG | AFG | AFG | AFG | ALB |
| UNICEF Regions | Afghanistan | Afghanistan | Afghanistan | Afghanistan | Albania |
| UN Regions | SA | SA | SA | SA | EECA |
| WHO Regions | Asia | Asia | Asia | Asia | Europe |
| World Bank Income Groups | EMRO | EMRO | EMRO | EMRO | EURO |
| Survey Years | Low Income | Low Income | Low Income | Low Income | Upper Middle Income |
| Year\* | 1997 | 2004 | 2013 | 2018 | 2005 |
| Short Source | 1997 | 2004 | 2013 | 2018 | 2005 |
| Full Source Title | MICS | NNS | NNS | Other | MICS |
| ... | Afghanistan 1997 multiple indicator baseline (... | Summary report of the national nutrition surve... | Afghanistan National Nutrition Survey 2013 | Afghanistan Health Survey 2018 | Albania multiple indicator cluster survey 2005... |
| 0-23 Point Estimate | ... | ... | ... | ... | ... |
| 0-23 Sample Size | NaN | NaN | 6.1 | 5.3 | 22.2 |
| 24-35 Point Estimate | NaN | NaN | 1801308.75 | 7593.22 | 374.95 |
| 24-35 Sample Size | NaN | 3 | 5 | 3.7 | 22.2 |
| 36-47 Point Estimate | NaN | 182 | 836250 | 3955.94 | 207.62 |
| 36-47 Sample Size | NaN | 5.6 | 5.1 | 3.7 | 25.2 |
| 48-59 Point Estimate | NaN | 227 | 821790.25 | 4209.26 | 226.69 |
| 48-59 Sample Size | NaN | 6.4 | 4.3 | 2.7 | 30.3 |
| 24-59 Point Estimate | NaN | 241 | 787946.06 | 3670.15 | 266.07 |
| 24-59 Sample Size | NaN | NaN | 4.8 | 3.4 | 26.3 |
|  | NaN | NaN | 2445986.25 | 11835.35 | 700.38 |

**UNICEF\_Global\_Databases\_Mortality.csv**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| DATAFLOW | count | unique | top | freq | mean | std |
| REF\_AREA:Geographic area | 1967354 | 1 | UNICEF:GLOBAL\_DATAFLOW(1.0): Cross-sector indi... | 1967354 | NaN | NaN |
| INDICATOR:Indicator | 1967354 | 304 | PER: Peru | 12940 | NaN | NaN |
| SEX:Sex | 1967354 | 403 | DM\_LIFE\_EXP: Life expectancy | 56388 | NaN | NaN |
| TIME\_PERIOD:Time period | 1967354 | 3 | \_T: Total | 965843 | NaN | NaN |
| OBS\_VALUE:Observation Value | 1967354 | 274 | 2019 | 31304 | NaN | NaN |
| UNIT\_MULTIPLIER:Unit multiplier | 1967354 | 1003505 | 0 | 34802 | NaN | NaN |
| UNIT\_MEASURE:Unit of measure | 1040160 | 2 | 0: Units | 731381 | NaN | NaN |
| OBS\_STATUS:Observation Status | 1967354 | 25 | PCNT: % | 731455 | NaN | NaN |
| OBS\_CONF:Observation confidentaility | 1073371 | 8 | A: Normal value | 682080 | NaN | NaN |
| ... | 439171 | 2 | F: Free | 438867 | NaN | NaN |
| WGTD\_SAMPL\_SIZE:Weighted Sample Size | ... | ... | ... | ... | ... | ... |
| OBS\_FOOTNOTE:Observation footnote | 134793 | NaN | NaN | NaN | 1818821949 | 57102539477 |
| SERIES\_FOOTNOTE:Series footnote | 90383 | 882 | Oedema data was not considered in the analysis | 11066 | NaN | NaN |
| DATA\_SOURCE:Data Source | 959444 | 22 | Data from 236 countries/areas with population ... | 505568 | NaN | NaN |
| SOURCE\_LINK:Citation of or link to the data source | 1966488 | 4947 | UN\_IGME | 690459 | NaN | NaN |
| CUSTODIAN:Custodian | 978980 | 72 | https://population.un.org/wpp/ | 619864 | NaN | NaN |
| TIME\_PERIOD\_METHOD:Time period activity related to when the data are collected | 214345 | 6 | UNICEF Division of Data Analysis Planning an... | 159596 | NaN | NaN |
| REF\_PERIOD:Reference Period | 245328 | 3 | MOF: Middle of fieldwork | 145955 | NaN | NaN |
| COVERAGE\_TIME:The period of time for which data are provided | 662 | 1 | 2000-2021 | 662 | NaN | NaN |
| AGE:Current age | 364037 | 406 | 2000 to 2021 | 98710 | NaN | NaN |
|  | 1967354 | 70 | \_T: Total | 1629882 | NaN | NaN |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| DATAFLOW | min | 25% | 50% | 75% | max |
| REF\_AREA:Geographic area | NaN | NaN | NaN | NaN | NaN |
| INDICATOR:Indicator | NaN | NaN | NaN | NaN | NaN |
| SEX:Sex | NaN | NaN | NaN | NaN | NaN |
| TIME\_PERIOD:Time period | NaN | NaN | NaN | NaN | NaN |
| OBS\_VALUE:Observation Value | NaN | NaN | NaN | NaN | NaN |
| UNIT\_MULTIPLIER:Unit multiplier | NaN | NaN | NaN | NaN | NaN |
| UNIT\_MEASURE:Unit of measure | NaN | NaN | NaN | NaN | NaN |
| OBS\_STATUS:Observation Status | NaN | NaN | NaN | NaN | NaN |
| OBS\_CONF:Observation confidentaility | NaN | NaN | NaN | NaN | NaN |
| ... | NaN | NaN | NaN | NaN | NaN |
| WGTD\_SAMPL\_SIZE:Weighted Sample Size | ... | ... | ... | ... | ... |
| OBS\_FOOTNOTE:Observation footnote | 0 | 1377.5 | 3018.4 | 7711 | 2.43122E+12 |
| SERIES\_FOOTNOTE:Series footnote | NaN | NaN | NaN | NaN | NaN |
| DATA\_SOURCE:Data Source | NaN | NaN | NaN | NaN | NaN |
| SOURCE\_LINK:Citation of or link to the data source | NaN | NaN | NaN | NaN | NaN |
| CUSTODIAN:Custodian | NaN | NaN | NaN | NaN | NaN |
| TIME\_PERIOD\_METHOD:Time period activity related to when the data are collected | NaN | NaN | NaN | NaN | NaN |
| REF\_PERIOD:Reference Period | NaN | NaN | NaN | NaN | NaN |
| COVERAGE\_TIME:The period of time for which data are provided | NaN | NaN | NaN | NaN | NaN |
| AGE:Current age | NaN | NaN | NaN | NaN | NaN |
|  | NaN | NaN | NaN | NaN | NaN |

# Data Profiling

* **Profile and clean a data set ready for analysis**
* **Improve the integrity of data in each data set**

1. Convert Excel data into CSV format and extract relevant columns that are required for analysis.
2. Data are integrated into different .csv format under 02 data/Prepared data/01 CSV/

# Data Analysis Criteria for this Project:

|  |  |
| --- | --- |
| Be open source. -  source. - Include non-anonymized column names. | Data Source  <https://data.unicef.org/resources/dataset/malnutrition-data/> |
| Come from an authentic/authoritative | The UNICEF, WHO and the World Bank inter-agency team update the Joint Child Malnutrition Estimates (JME) every other year. And the data can be used for analysis, study or the research purpose. |
| Include non-anonymized column names. | Databases containing national country values from survey and administrative sources, disaggregated by various background characteristics such as place of residence, geographic location, age, and sex, when available. |
| Be recent (ideally, no more than 3 years old. However, this factor is not essential - if you’ve found a perfect data set for your purposes, it could be older too, but not more than 10 years old). | The UNICEF, WHO and the World Bank inter-agency team update the Joint Child Malnutrition Estimates (JME) every other year. The data collected on **May 2023** from WHO database. |
| Be recent (ideally, no more than 3 years old. However, this factor is not essential - if you’ve found a perfect data set for your purposes, it could be older too, but not more than 10 years old). | The data contains all the essentials requirement for this course UNICEF, WHO and the World Bank inter-agency team update the Joint Child Malnutrition Estimates (JME) every other year |
| Contain at least 2-3 continuous variables (apart from index variables, ID variables, dates, years, etc). | Yes, the data set contains continuous variable |
| Contain at least 2-3 categorical variables (apart from index variables, ID variables, dates, years etc). | The table contains year and geographical information. |
| Contain at least 1,500 rows. | Yes |
| Include a geographical object of some kind: for instance, a column relating to a country, continent, or something similar. If the information from the data set refers to the US, for example, there should be a column containing the names/abbreviations of the states. Note: there should be at least a couple of different values in this column. This is important for the geospatial analysis you’ll be conducting. Of course, you can also use data sets with latitude and longitude | The data is collected from worldwide and contains all regional geographical information like country and continental information. |

# Analysis Criteria

Data Analyst will be conducting the following analyses in this Achievement (note that not all results need to be included in the final dashboard):

* Exploratory analysis through visualizations (scatterplots, correlation heatmaps, pair plots and categorical plots)
* Geospatial analysis using a shapefile
* Regression analysis
* Cluster analysis
* Time-series analysis
* Analysis narrative and
* final results (presented in a dashboard)

Dashboard Requirements

As explained above, the dashboard needs to tell the story of your analytical journey. it should contain a curation of the key results that discovered as you worked through the Achievement.

Final dashboard must: -

* Be designed with a use-case in mind (answering key guidance questions).
* Be created in Tableau Public. - Be interactive.
* Adhere to visual design best practices.
* Include an introduction page that describes the project (data and purpose).
* Include relevant result(s) of initial visual exploratory analysis.
* Include an explanation for how the results of the exploratory analysis resulted in defining research questions and/or hypotheses.
* Contain a geospatial component.
* Address the defined questions/hypotheses using advanced analytical techniques.
* For example: - Regression analysis - Cluster analysis - Time-series analysis -
* Include a results summary page explaining how the results do or don’t address the initial research questions/hypotheses.
* It should innclude details on the limitations of the project.
* Include a proposal of next steps for further analysis.

# Analysis

## Top 10 countries that are most affected by Stunting

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | | Burundi | 55.27 | | Timor-Leste | 53.82 | | Yemen | 51.82 | | Guatemala | 49.34 | | Eritrea | 47.75 | | Papua New Guinea | 46.53 | | Madagascar | 46.47 | | Niger | 46.06 | | Afghanistan | 45.96 | | Malawi | 44.48 | |

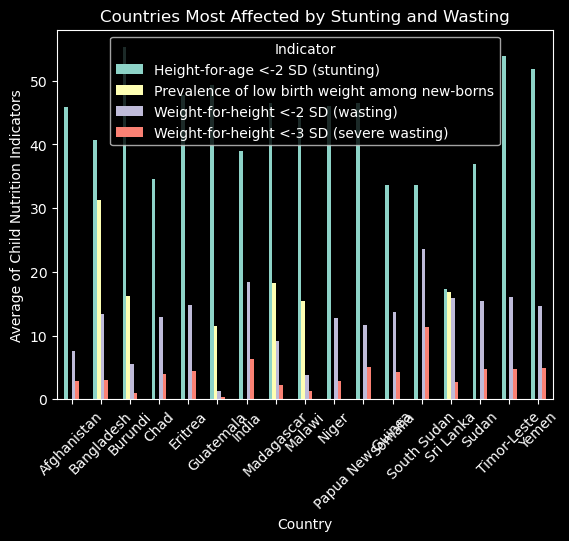
## Top 10 countries that are most affected by Wasting

Geographic area

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | | South Sudan | 23.63 | | India | 18.37 | | Timor-Leste | 15.98 | | Sri Lanka | 15.83 | | Sudan | 15.38 | | Eritrea | 14.77 | | Yemen | 14.67 | | Somalia | 13.75 | | Bangladesh | 13.37 | | Chad | 12.91 | |

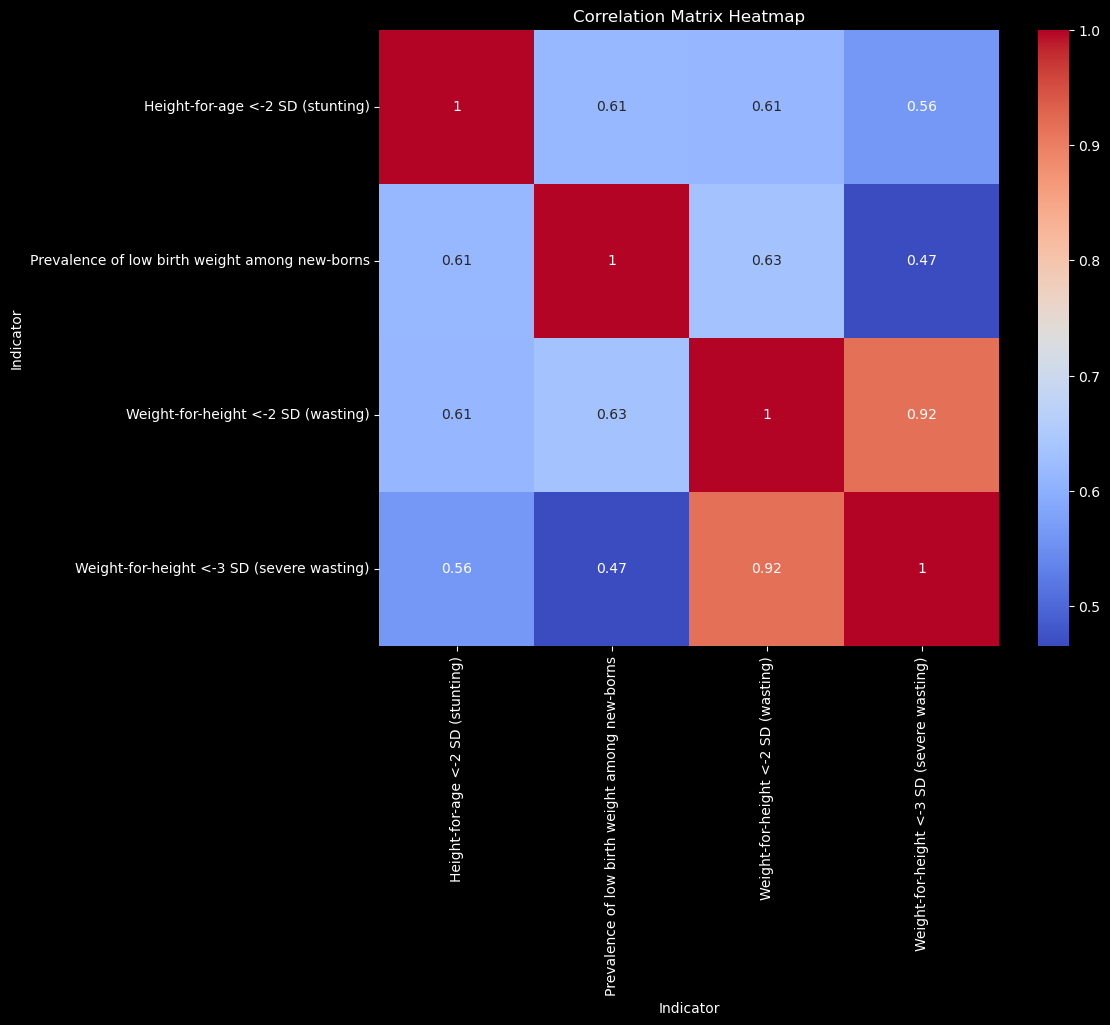
## Top 10 countries that are most affected by Stunting and Wasting

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Indicator** | **Height-for-age <-2 SD (stunting)** | **Prevalence of low birth weight among new-borns** | **Weight-for-height <-2 SD (wasting)** | **Weight-for-height <-3 SD (severe wasting)** |
| **Geographic area** |  |  |  |  |
| **Afghanistan** | 45.96 | NaN | 7.56 | 2.93 |
| **Bangladesh** | 40.76 | 31.24 | 13.37 | 3.03 |
| **Burundi** | 55.27 | 16.13 | 5.47 | 1.04 |
| **Chad** | 34.6 | NaN | 12.91 | 3.93 |
| **Eritrea** | 47.75 | NaN | 14.77 | 4.4 |
| **Guatemala** | 49.34 | 11.45 | 1.24 | 0.37 |
| **India** | 38.91 | NaN | 18.37 | 6.3 |
| **Madagascar** | 46.47 | 18.21 | 9.08 | 2.24 |
| **Malawi** | 44.48 | 15.47 | 3.86 | 1.34 |
| **Niger** | 46.06 | NaN | 12.75 | 2.86 |
| **Papua New Guinea** | 46.53 | NaN | 11.62 | 5.03 |
| **Somalia** | 33.62 | NaN | 13.75 | 4.33 |
| **South Sudan** | 33.7 | NaN | 23.63 | 11.4 |
| **Sri Lanka** | 17.3 | 16.86 | 15.83 | 2.77 |
| **Sudan** | 36.86 | NaN | 15.38 | 4.74 |
| **Timor-Leste** | 53.82 | NaN | 15.98 | 4.67 |
| **Yemen** | 51.82 | NaN | 14.67 | 4.88 |



## Correlation between Stunting, Wasting and Prevalence of low birth weight among new-borns

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Indicator** | **Height-for-age <-2 SD (stunting)** | **Prevalence of low birth weight among new-borns** | **Weight-for-height <-2 SD (wasting)** | **Weight-for-height <-3 SD (severe wasting)** |
| **Indicator** |  |  |  |  |
| **Height-for-age <-2 SD (stunting)** | 1 | 0.61 | 0.61 | 0.56 |
| **Prevalence of low birth weight among new-borns** | 0.61 | 1 | 0.63 | 0.47 |
| **Weight-for-height <-2 SD (wasting)** | 0.61 | 0.63 | 1 | 0.92 |
| **Weight-for-height <-3 SD (severe wasting)** | 0.56 | 0.47 | 0.92 | 1 |



The correlation coefficients between stunting, wasting, and the prevalence of low birth weight among newborns are as follows:

Stunting (Height-for-age <-2 SD) and the prevalence of low birth weight among newborns: The correlation coefficient is 0.61. This suggests a moderate positive correlation, indicating that there is a tendency for areas with higher stunting rates to also have higher prevalence of low birth weight among newborns.

Stunting (Height-for-age <-2 SD) and wasting (Weight-for-height <-2 SD): The correlation coefficient is 0.61. This indicates a moderate positive correlation, implying that areas with higher stunting rates also tend to have higher rates of wasting.

Stunting (Height-for-age <-2 SD) and severe wasting (Weight-for-height <-3 SD): The correlation coefficient is 0.56. This indicates a moderate positive correlation, suggesting that areas with higher stunting rates are also likely to have higher rates of severe wasting.

The prevalence of low birth weight among newborns and wasting (Weight-for-height <-2 SD): The correlation coefficient is 0.63. This indicates a moderate positive correlation, implying that areas with higher prevalence of low birth weight among newborns also tend to have higher rates of wasting.

The prevalence of low birth weight among newborns and severe wasting (Weight-for-height <-3 SD): The correlation coefficient is 0.47. This indicates a moderate positive correlation, suggesting that areas with higher prevalence of low birth weight among newborns are also likely to have higher rates of severe wasting.

Wasting (Weight-for-height <-2 SD) and severe wasting (Weight-for-height <-3 SD): The correlation coefficient is 0.92. This indicates a strong positive correlation, implying that areas with higher rates of wasting are highly likely to have higher rates of severe wasting.

In summary, the data suggests that stunting, wasting, and the prevalence of low birth weight among newborns are positively correlated, despite the fact that to varying degrees. Higher rates of stunting are associated with higher rates of low birth weight, wasting, and severe wasting. Similarly, areas with higher prevalence of low birth weight among newborns are more likely to have higher rates of wasting and severe wasting. Additionally, wasting and severe wasting are strongly correlated, indicating that they tend to occur together.

## Stunting, Wasting and Prevalence of low birth weight relationship with Income

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sever wasting among children below 5 years have impacted more with countries income.   |  |  |  | | --- | --- | --- | | **Sever wasting depending on the income** | | | | **World Bank Income Groups** | |  | | **Low Income** | **2.67** | | | **Lower Middle Income** | 2.23 | | | **Upper Middle Income** | 1.25 | | | **High Income** | 0.58 | | |  |
|  |  |
|  | |  |  | | --- | --- | | **World Bank Income Groups** | **Low-Income Estimation value** | |  |  | | **Countries and areas** |  | | **South Sudan** | 11.4 | | **Yemen** | 5.3 | | **Syrian Arab Republic** | 5.03 | | **Eritrea** | 4.87 | | **Sudan** | 4.77 | | **Somalia** | 4.35 | | **Chad** | 4 | | **Democratic Republic of the Congo** | 3.72 | | **Niger** | 3.44 | | **Guinea** | 3.2 | |
|  | Scatter plot drawn on the countries South Sudan, Yemen, Syrian Arab Republic, Eritre, Sudan. Somalia, Chad, Democratic Republic of the Congo, Niger, Guinea.  From the Scatter plot the low-income countries stunting is gradually reducing over the past 5 years. During the 2005 and 2010 Severe wasting highly affected. |

## Pairplot on global severe wasting data

|  |  |
| --- | --- |
|  |  |

## Categorical plot depicting the severity wasting on global data among genders and total population

|  |  |
| --- | --- |
|  | **Categorical plot: Severe wasting comparison between males, Females & country’s total wasting records under 5-year-old infants** |

|  |  |
| --- | --- |
| **Categorical plot: Severe wasting comparison between genders under 5-year-old infants.**  The severity among male children is higher compared to female children. Severe wasting in 2022 in male children below 5 years is 3%, and females Children’s highest severity is 2% |  |

# Geospatial Analysis using Choropleth Map

In this step, geospatial analysis is made on the Stunting data.

**Objective 3:** Does progress to reduce stunting has equal across regions

and sub-regions?

**Hypothesis:** Progress to reduce stunting has been equal across regions

and sub-regions.

**Datasource:**

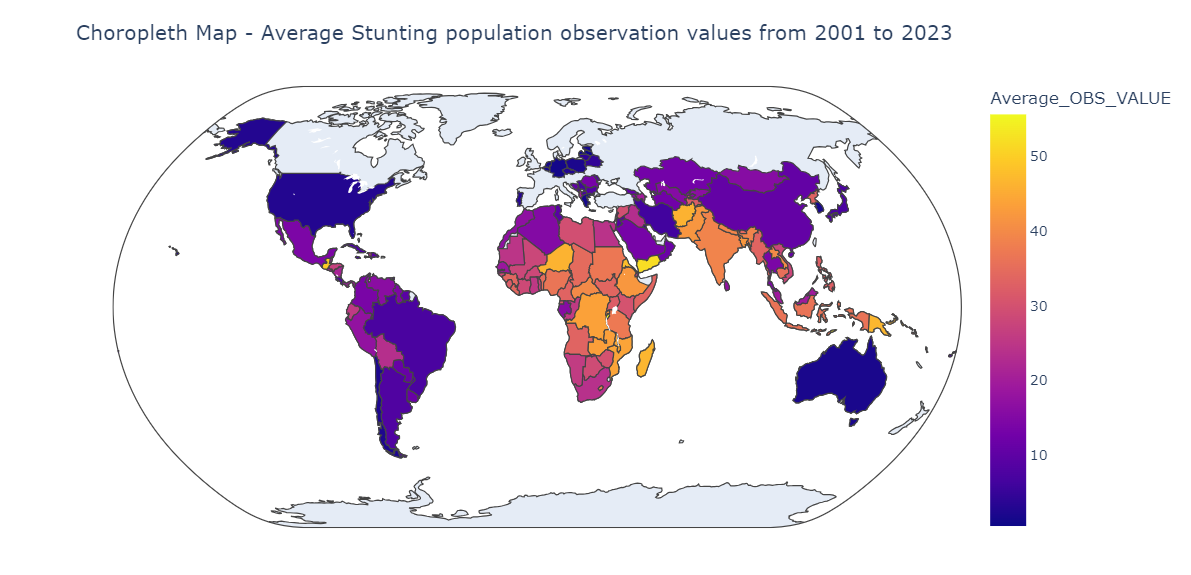
▪ UUNICEF\_Database\_wasting\_stunting.csv

## Geospatial Analysis on Stunting Data

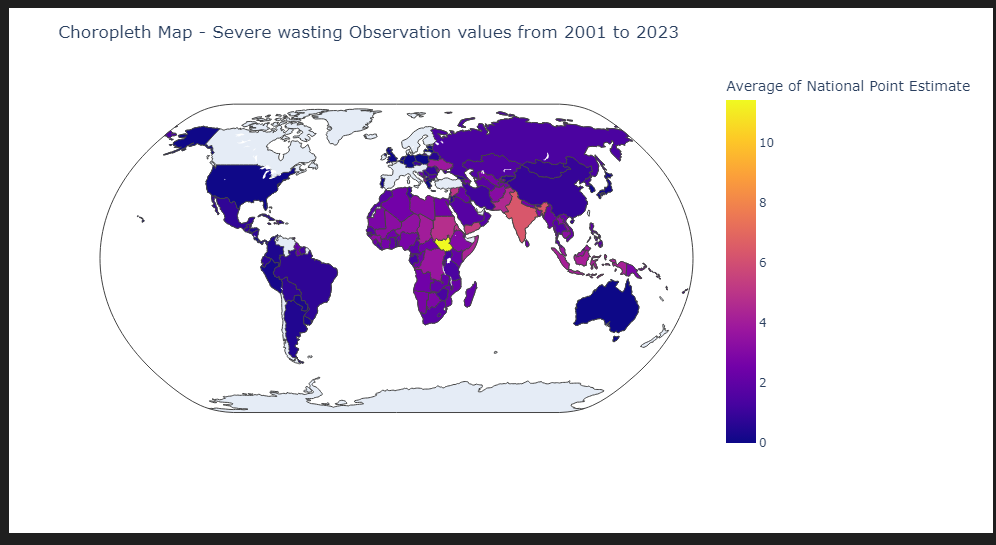
For analysis and visualisation, the Plotly library is installed. Refer to the link for the Getting Started with Plotly in Python.

[Getting started with plotly in Python](https://plotly.com/python/getting-started/#plotly-charts-in-dash)

## Average Stunting population observation values from 2001 to 2023



## Geospatial Analysis On Severe Wasting Data



From the above analysis, it is clear that nutrition deficiency isn’t the same globally. One of the reasons for malnutrition is because of low-income countries and the country’s financial stability.

Some of the most affected reasons due to Wasting are South Sudan, India, Timor-Leste, Sri Lanka, Sudan, Eritrea, Yemen, Somalia, Bangladesh, and Chad.

And some of the Stunting affected reasons are Burundi, Timor-Leste, Yemen, Guatemala, Eritrea, Papua New Guinea, and Madagascar.

**This observation provides enough evidence to conclude that malnutrition globally isn’t the same.**

# Supervised Machine Learning: Regression

**Objective**:

Identify the number of deaths and affected due to mall nutrition(stunting, overweight, wasting and severe wasting ).

**Hypothesis**:

Malnutrition among children under the age of 5 is directly related to the mortality count in the nations.

**Datasource**

UNICEF\_Global\_Databases\_Mortality dataset contains the mortality rate in each geographic region with year.

Limitation: Data isn’t recorded every year.

UNICEF\_Global\_Joint-Malnutrition-Estimates contains global stunting, wasting and overweight information.

In regression analysis, **independent variables** are also known as explanatory variables because they help explain the trends (if any) that we see in the **dependent variable**. The dependent variable is also known as the response variable because it responds to changes in the explanatory variables.

**Dependent Variables**

* Stunting Observation value collected from the Sample data

**Independent Variables**

* Mortality rate under five years of age.

## Linear Regression Analysis

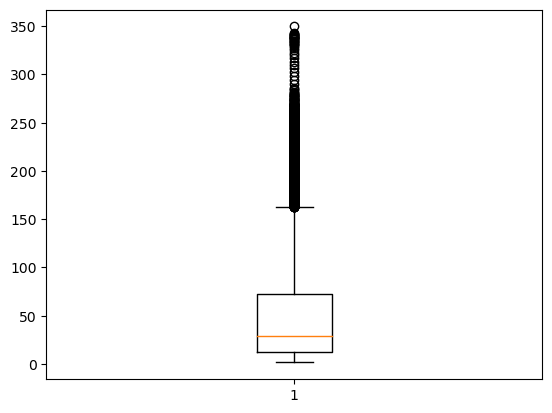
1. **Data Sourcing and Data Wrangling**

|  |  |
| --- | --- |
| **Data Source** | UNICEF\_Global\_Databases\_Mortality.csv  UNICEF\_Global\_Databases\_Stunting.xls |
| **Filter the columns** | ['REF\_AREA:Geographic area', 'TIME\_PERIOD:Time period', 'OBS\_VALUE:Observation Value']  ['REF\_AREA:Geographic area', 'INDICATOR:Indicator' = 'CME\_MRY0T4: Under-five mortality rate',  'TIME\_PERIOD:Time period', 'OBS\_VALUE:Observation Value'] |
| **Unique types in Date column** | All the values are converted to year column in YYYY format |
| **REF\_AREA:Geographic area** | Geographic data contains the name for different reasons. Also it has some extra character which requires to filter out. |
| **Data type conversion** | Convert the year column to an integer.  Convert the OBS values to float |

1. **Exploratory Analysis**

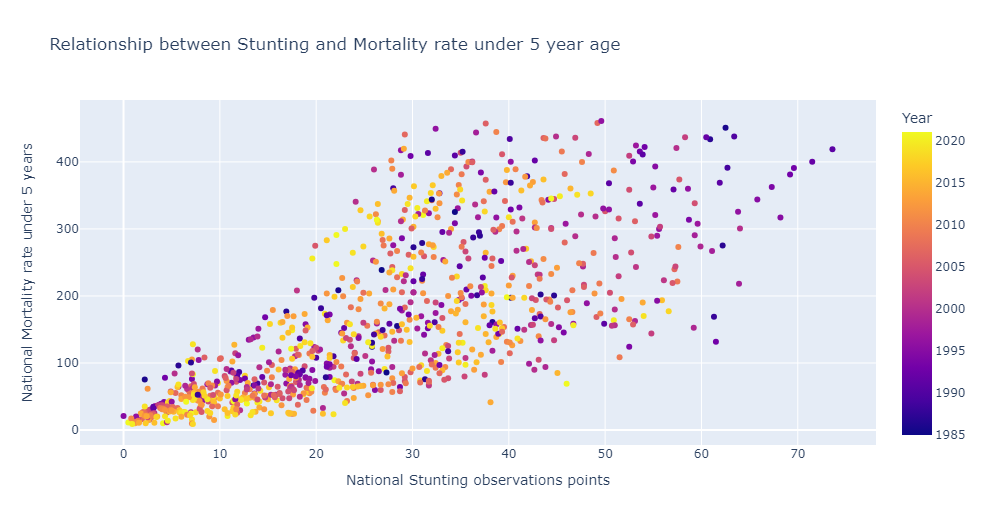
**Performed some basic exploratory analysis on the** UNICEF\_Global\_Databases\_Mortality.csv, UNICEF\_Database\_wasting\_stunting.csv.

1. Removing the outliers from the dataset



1. Updating the countries name to a proper convention by removing REF\_AREA

**Comparison with the Stunting observation dataset with the mortality data set.**

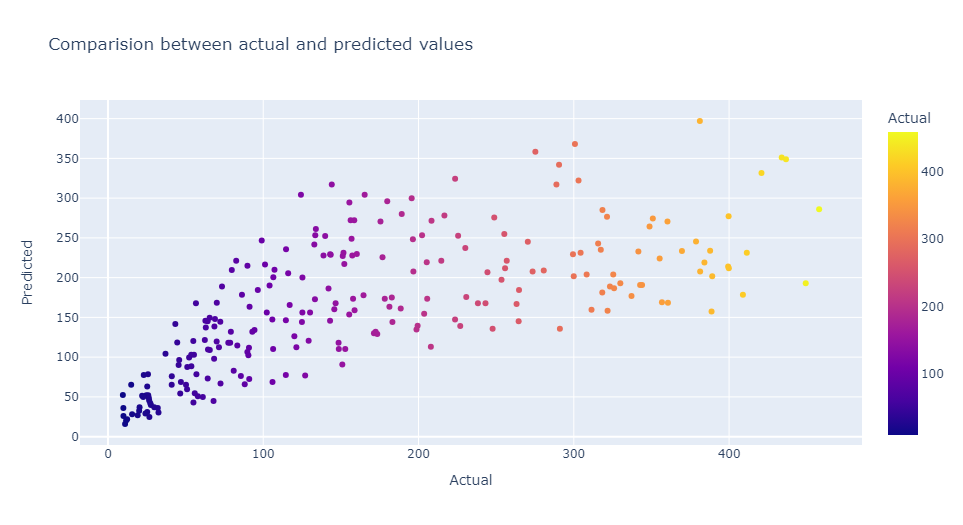


**. 3. Model training and evaluation**:

|  |
| --- |
| Coefficient |
| **National Point Estimate** 5.55 |

According to the above coefficient result increase in stunting observation by 5.55 increases one mortality rate.

1. **Testing the Linear Regression Model.**



**The prediction values are almost the same as actual. But some of the values which have higher mortality rate are deviated**

MAE: 66.28091714178167

MSE: 7137.504180438801

RMSE: 84.48375098466451

MAE represents the average magnitude of the errors between the predicted and actual values. It measures the average absolute difference between the predicted and actual values. the MAE is 66.28091714178167.

MSE calculates the average of the squared differences between the predicted and actual values. It gives higher weights to larger errors due to the squaring operation. In this case, the MSE is 7137.504180438801.

RMSE is the square root of MSE. It measures the typical magnitude of errors in the same units as the dependent variable. By taking the square root of MSE, RMSE cancels out the squaring operation, giving a more interpretable value. In this example, the RMSE is 84.48375098466451.

**The model's predictions have less deviation from the actual values. It means the number of national mortality under age five and the stunting observation is positively correlated. An increase in stunting observation values by 5.55 increases the mortality count by 1.**