**Human Development Index**

The Human Development Index (HDI) is a composite statistic of [life expectancy](https://en.wikipedia.org/wiki/Life_expectancy), [education](https://en.wikipedia.org/wiki/Education), and [income per capita](https://en.wikipedia.org/wiki/Income_per_capita) indicators, which are used to rank countries into four (Very High, High, Medium, Low) tiers of [human development](https://en.wikipedia.org/wiki/Human_development_(humanity)).

**Comprehension of data:**

**Output labels:**

* **Human Development Index (HDI) Label:**

Categorization of a country into equal categories of, “Very High”, “High”, “Medium” or “Low” based on their human development index value.

* **Human Development Index (HDI):**

HDI is a score measuring human development based on factors such as life expectancy, knowledge and standard of living.

**Social data**

* **Education (Primary, Secondary, Tertiary):** Ratio of female to male education in each country. Ratios closer to or above one shows countries giving importance to female education.
* **Mortality (for infants, for adults):** Number of deaths per 1000 people. It shows how well equipped a country is with its health and medical services
* **Demographics:** Information about population, median age, unemployment rate. It shows distribution of resources.

**Economic data:**

Economic factors shows the resources available to the country and how they are utilizing them

* **GDP per capita:** Purchasing power in dollar.
* **Foreign Direct Investment:** Equity capital and reinvestment of earnings expressed as a percentage of gross domestic product (GDP)
* **Remittance Inflows:** Earnings and material resources received by the country through their citizens residing in foreign countries.
* **Public expenditure on education:** Total public expenditure (current and capital) on education, expressed as a percentage of gross domestic product (GDP)
* **Total expenditure on health:** Total expenditure (public and private) on health services

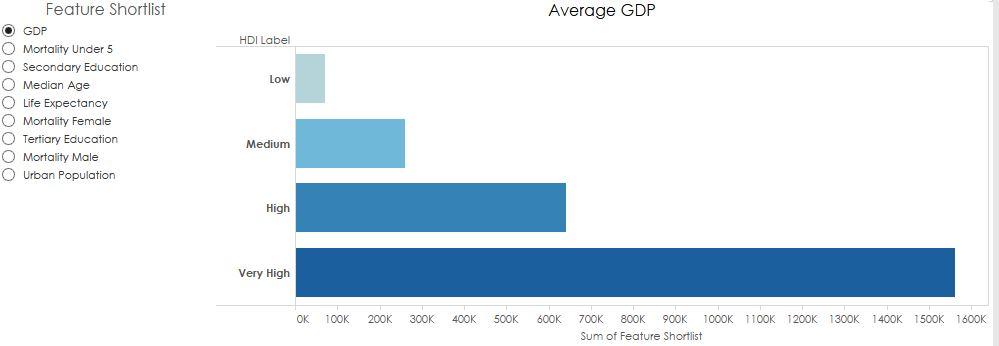
**Executive Summary**

The great development challenge of the 21st century is to safeguard the right of generations today and in the future to live healthy and fulfilling lives. Our analysis focuses on the most important factors which affects a country’s HDI. Our report also offers important new contributions to the global dialogue on this challenge, showing how if we focus on some specific features for a country we can boost its HDI from low to medium or high.

Below is the snapshot which focuses on providing the current trends of HDI indexes and labels (low, medium, high and very high) around the globe. As evident, most of the countries in Africa region have low or medium HDI levels whereas almost all western countries have High and Very High HDI labels. Talking about Asia, we see that many countries are in High and medium range of HDI labels.



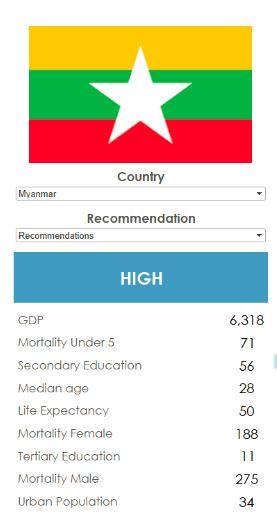
Out of the 17 features, the 9 main features that affects HDI are GDP, Mortality under 5, secondary education, Median Age, Life Expectancy, Mortality Female, Tertiary education, Mortality Male and Urban population. The graph also shows the distribution of Average GDP across the countries classified into Low, Medium, High and Very High labels of HDI.



**Specific recommendations for countries:-**

Myanmar currently has low HDI label. In order to increase its HDI label to High from low :-

* It should assist in infrastructure development, especially related to libraries and science facilities
* It should create awareness about the importance of secondary education among people and aim at increasing the participation by at least 5%.
* It should focus on increasing the number of schools and colleges across the country to make them accessible to more people.
* It should come up with schemes to offer secondary education for free for under privileged people.
* It should provide easy loan facilities and give scholarships for higher education

Kenya also currently has low HDI label. In order to increase its HDI label to Medium to low :-

**GDP**

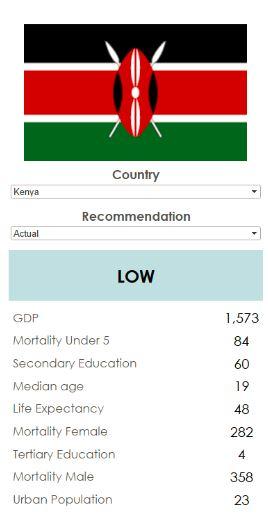
* As a share of gross domestic product (GDP), Kenya’s manufacturing sector has been stagnant in recent years. Kenya needs to increase the competitiveness of the manufacturing sector so that it can grow, export, and create much-needed jobs.
* Decreasing the cost of doing business by foreign countries addressing critical issues related to energy, access to finance, and cross-border trade will increase GDP significantly.
* Being the world’s leading exporter of black tea which is one of Kenya’s top foreign exchange earners. It should try to increase the agricultural land to produce more tea.

**Secondary Education**

* It should come up with schemes to offer secondary education for free for under privileged people.
* It should focus on increasing the number of schools and colleges across the country to make them accessible to more people.
* It should provide easy loan facilities and give scholarships for higher education

**Mortality Under 5**

* It should provide better access to medicines, health facilities, water, and sanitation facilities to people.
* Create awareness among parents by educating them about the importance of immunizations to be given to child after birth.
* Promoting breastfeeding which strengthen children and immune them from diseases.

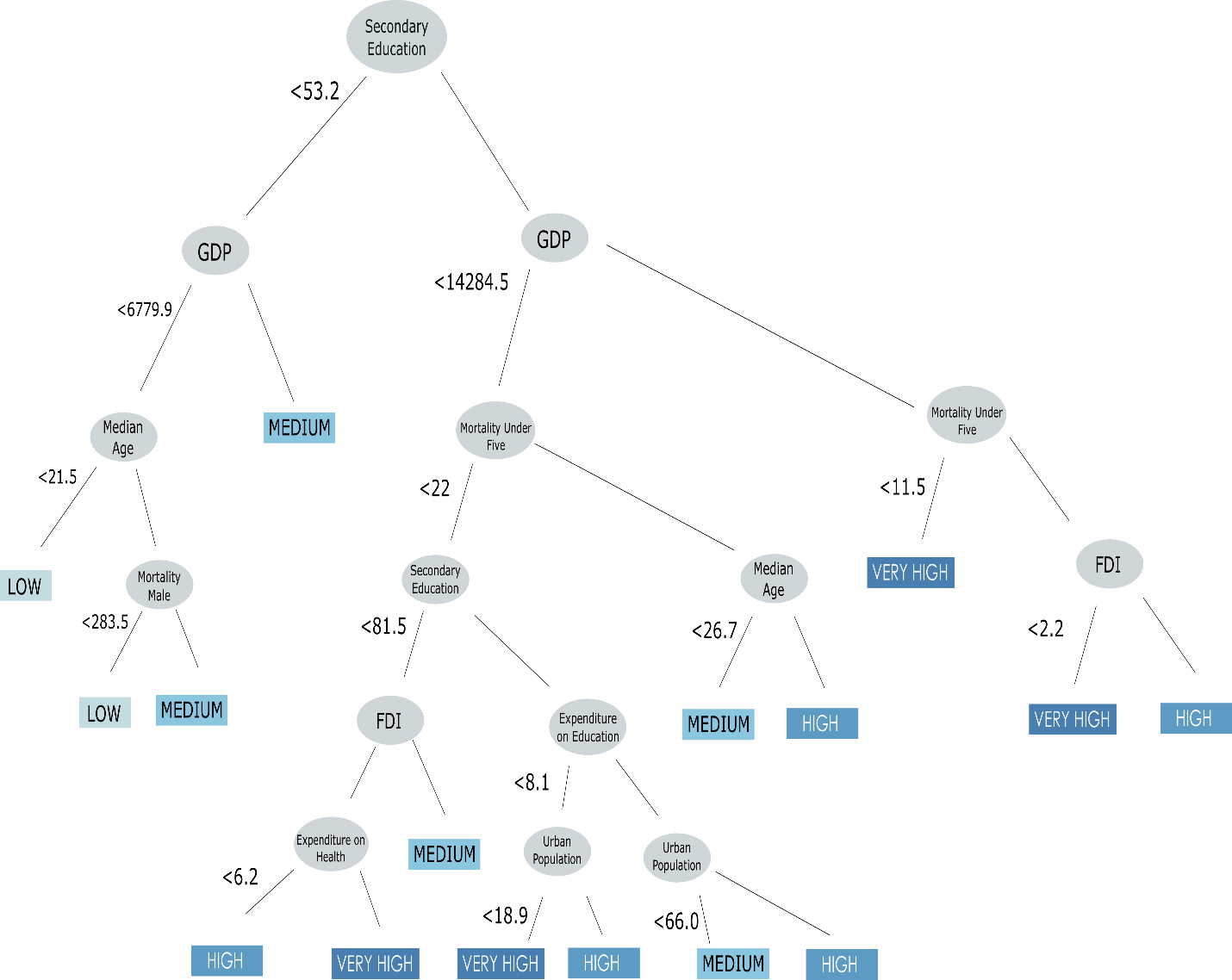
**Detailed Recommendations**

**Data Modelling**

**Final Model**

After trying all the models using 10 fold cross validation, we came to the conclusion that classification models are giving better results than regression models.

The classification model which gives us the highest accuracy is Random Forest with 84.32% accuracy



Of all the classification trees, the most stable tree which gave the lowest error was selected. This tree identifies that major factors that leads to low Human Development Index are low median age and mortality.

Correspondingly, the model also identified that greatest factor which distinguished countries with high and very high HDI was mortality among infant.

**Methodology for Simulation**

The best decision tree model helped us to run simulations for different countries by progressively incrementing the best features by a value of 1%. We selected the below enumerated features for our focussed study:

1. Secondary Education
2. Gross Domestic Product (GDP)
3. Median Age
4. Mortality rate under age 5

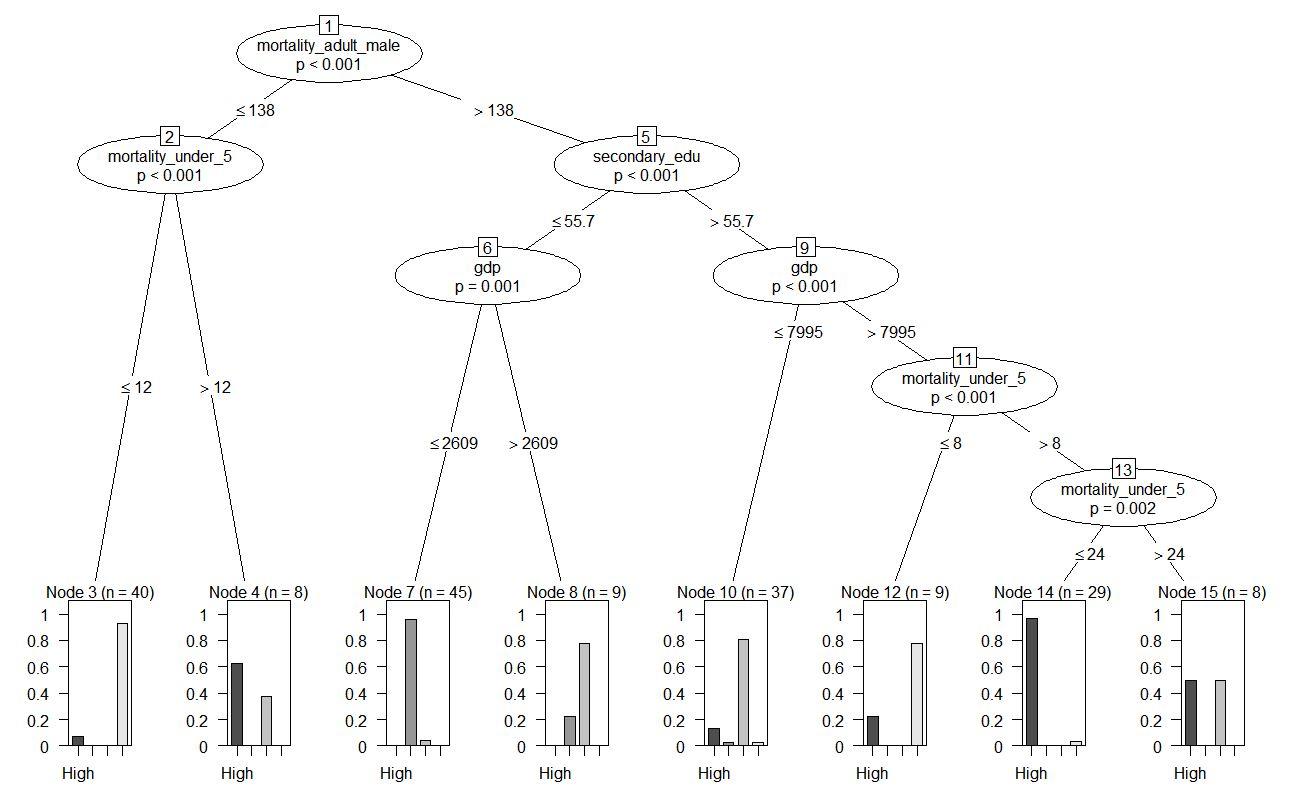
Iteratively we performed simulation for all the countries using the above features by taking them one at a time or in combination. Below are some of the key results we found

1. 5% increase in either of the best feature secondary education, Gross Domestic product Median Age or Mortality under 5 promoted Kenya to Medium level HDI.
2. 5% increase in Secondary Education promoted Myanmar to High level HDI.

**Other Models**

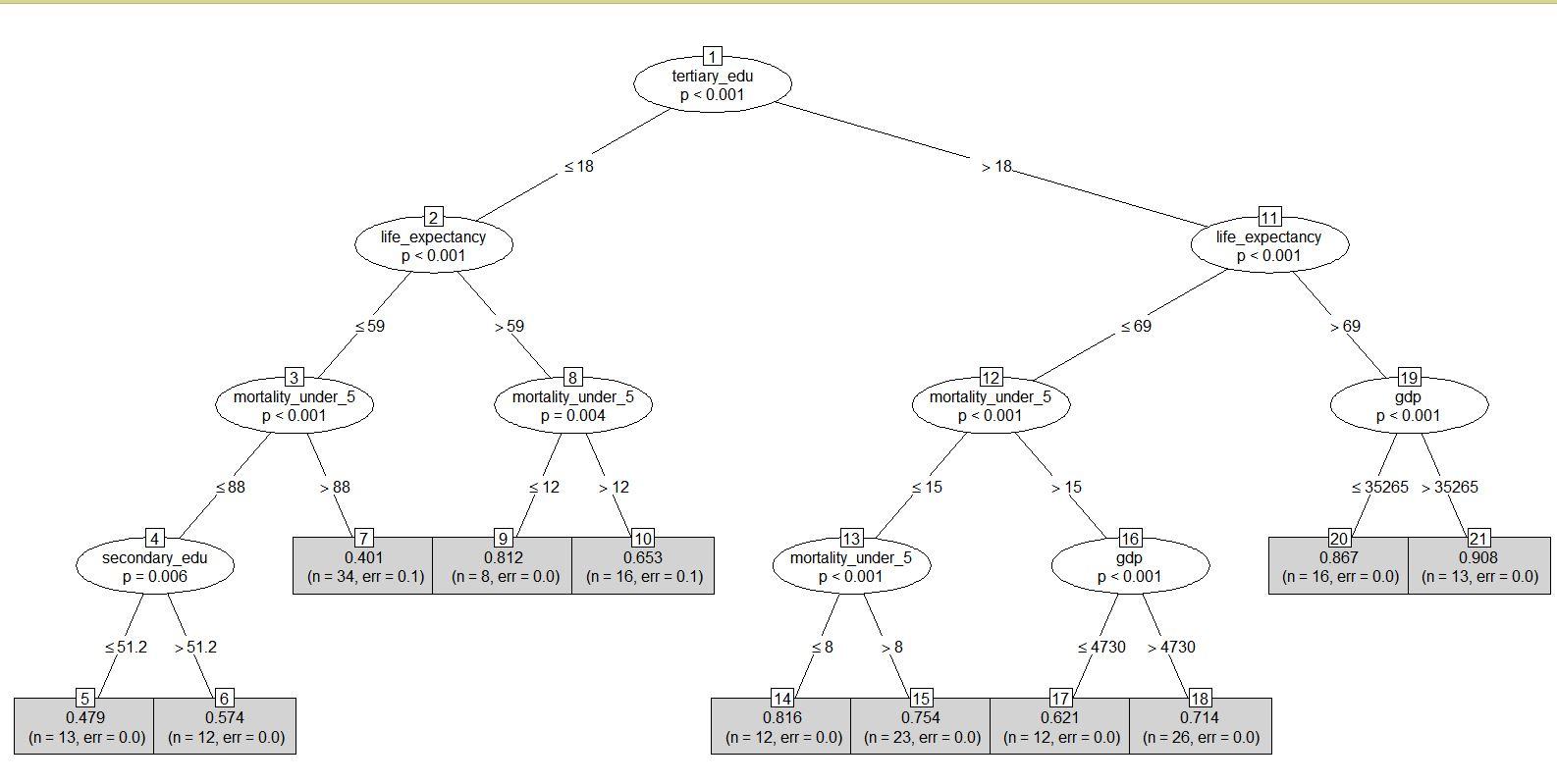
|  |  |  |
| --- | --- | --- |
| **Model Name** | **Accuracy** | **R2** |
| Random Forest | 84.32% | - |
| Regression Forest | 82.70% | 0.94 |
| Decision Tree | 78.92% | - |
| Regression Tree | 79.46% | 0.89 |
| Linear Regression | 76.75% | 0.88 |
| KNN | 75.13% | - |

**Decision Tree**



A Decision Tree (shown above), used as a classification model, identified that major factors that led to low Human Development Index are high mortality among adult males, low secondary education and low GDP.

Correspondingly, the model also identified that the greatest factor that distinguished countries with very high HDI is mortality rate of less than 12 for infants of age under 5.

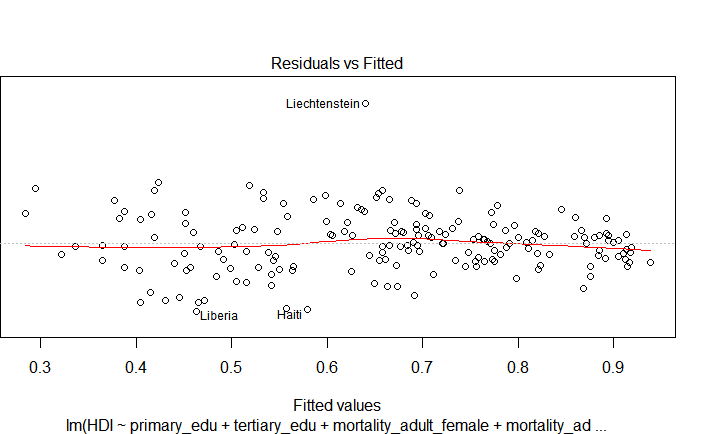
**Regression Tree** 

A decision Tree (shown above), used as a regression model, identified that major factors that led to low Human Development Index are low tertiary education, low life expectancy, low mortality under 5 and low secondary education.

Correspondingly, the model also identified that the greatest factor that distinguished countries with very high HDI is high tertiary education, high life expectancy and high gdp.

**Linear Regression**

As seen in the plot it seem to follow a linear trend, the model does not fit the data well. The vertical width of scatter tends to increase across the fitted values so we can assume that we have slight variation in variance. We can also see some of the outliers in countries like Liberia, Halti and Liechtenstein.



**Risks and Mitigation**

**Risk: Missing Values**

Out of 185 rows, 75 rows had missing value of one or more attribute. Therefore rather that deleting the rows with missing values in order to preserve the data we adopted the below two approaches to handle missing values:-

**Mitigation:**

1. Substitute the missing data with values from the reports published by the UN[[1]](#footnote-0) and the World Bank[[2]](#footnote-1).
2. Categorize the remaining missing values based on whether the data is missing randomly or systematically.
   1. **Random Missing Values**

These missing values did not follow any particular pattern or trend and hence facilitated imputation using the information available from other similar countries. The imputation is performed using the ‘Multivariate Imputation by Chained Equations’ (MICE) algorithm.

* 1. **Systematic Missing Values**

These missing values followed an identifiable trend and did not allow imputation without the risk of bias. For example, ‘Unemployment Rate’ has 25 missing values, 22 of which belong to countries categorized as Low HDI. In addition, this variable was not found significant in our preliminary analysis using basic linear regression and KNN classification. We therefore choose to exclude this variable from further analysis. (Highlighted in yellow)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Attributes for interpreting HDI** | **Missing Value Pattern** | **Total Number of Missing Values** | **Substituted Using Outside Sources** | **Imputed Using MICE** |
| Primary Education | Random | 5 | 1 | 4 |
| Secondary Education | Random | 8 | 2 | 6 |
| Tertiary Education | Random | 22 | 4 | 18 |
| Female Mortality | Random | 1 | 0 | 1 |
| Male Mortality | Random | 1 | 0 | 1 |
| Life Expectancy | Random | 1 | 0 | 1 |
| Median Age | Random | 8 | 0 | 8 |
| Unemployment Rate | Systematic | 25 | 0 | Column Ignored |
| Population Density | Random | 2 | 2 | 0 |
| GDP | Random | 13 | 9 | 4 |
| FDI | Random | 13 | 9 | 4 |
| Remittance Inflows | Random | 26 | 8 | 18 |
| Expenditure on Education | Random | 20 | 3 | 17 |
| Expenditure on Health | Random | 2 | 0 | 2 |
| Total |  | 122 | 38 | 84 |
| Proportions |  |  | 31.14% | 68.86% |

Total Proportion of Missing Values Substituted using Outside Sources :- **31.14%**

Total Proportion of Missing Values Substituted using MICE :- **68.86%**

**Risk: Limitation of data size**

**Mitigation**:

* 1. This model can be extended yearly when each year's data is available to provide better prediction. Additionally, with data of multiple years, a time series analysis could be performed to give better correlations.

**Next Steps**

**Team Contribution**

|  |  |
| --- | --- |
| **Team Member** | **Contribution** |
| Ankit Gupta | Missing Value analysis and simulations |
| Dhruv Bhogle | Data Modelling |
| Payal Fofadiya | Recommendations and Video |
| Prateek Biswas | Data Visualization |
| Sahil Nyati | Documentation |

1. <http://www.un.org/en/development/desa/population/index.shtml> [↑](#footnote-ref-0)
2. <http://data.worldbank.org/indicator/> [↑](#footnote-ref-1)