

**Impact of Economic Growth and Education in Infant Mortality in the 21st Century**

Mano de una persona

Descripción generada automáticamente



**October 2022**



**Carlos Chávez, Yohana Sandoval, Daniel Duran, Elder Giraldo, Fernán Marsiglia, Juan Alzate, Marta Ordoñez**

**Introduction 3**

[Description Problem 3](#_heading=h.3dy6vkm)

Project Scoping4

Datasets5

**Data Curation 6**

**Exploratory Data Analysis 9**

[Visualizations 9](#_heading=h.30j0zll)

**Extended Data Analysis 15**

[**Dashboard 17**](#_heading=h.2et92p0)

[**Conclusions 1**](#_heading=h.tyjcwt)**9**

# Description Project

Infant mortality is the death of an infant under the age of 1. The infant mortality

rate is the number of infant deaths for every 1,000 live births.

The infant mortality rate is directly related to what is spent on healthcare services and this is

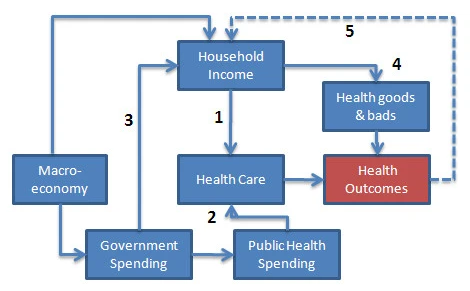
determined by the state's ability to raise funds through taxes, social security, or other means. This,

in turn, affects the ability of a national government to adequately finance basic services in order to evaluate whether the mortality of age groups is determined by economic growth and

education levels in a given country or region. Thus, we propose to analyze variables related

to economic growth (see Datasets/Economic Development) and education (see Datasets/Education).

In addition, this can help us understand how these variables influence the mortality of age groups in countries and their regions.



**Health and wealth: simplified causal relationships**.

# Project Scoping

Economic growth and education may reflect the effects of other socio-economic variables that may

also be related to mortality. For this reason, it is important to conduct comprehensive and

interdisciplinary analyses that consider as much information as possible. Such information can be

relevant to governmental decision-making on the direction of each country and/or region. This can

enable the design and implementation of public policies aligned with the needs of the population

and the objectives and goals of sustainable development (Agenda 2030).

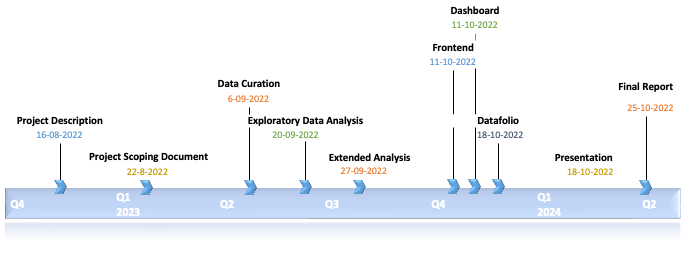
Global, regional, and national statistics on population. Also, health and mortality indicators are

essential in order to assess development and health progress to guide resource allocation. Specifically,

this data is used to monitor progress towards health-related targets within the Sustainable

Development Goals (SDGs).

In the following graph we can see graphically the life of the project in the last months:



**. Schedule of activities**

# Datasets

# For the project, data was chosen focusing on 3 fronts. In the following table we can see what they are and where the description of the tables was extracted from.

| **Dataset** | **Source** | **Description** |
| --- | --- | --- |
| **Mortality** | <https://platform.who.int/mortality/themes/theme-details/mdb/noncommunicable-diseases> | **WHO** mortality database contains mortality percentages around the world. |
| **Education** | <https://data.worldbank.org/indicator> | **The World Bank** Education database contains some of the most important measures that are part of the direct and indirect measures of education. |
| Economic growth | <https://data.worldbank.org/indicator> | **The World Bank** Economic Growth database contains some of the most important measures in the field of economic and economic growth. Most of the variables are related to petrol or energy as an indirect measure of economic growth. |

1. **Mortality**

The WHO Mortality Database is a compilation of mortality data. It is organized by country and area, year, sex, age and cause of death. It is transmitted annually by national authorities from their civil registration

and a vital statistics system. The Mortality Database comprises data from 1950 to date.

**Data Profile**:

* After loading the data frame, the first thing that was done was checking the name of the variables

and their meaning. After that, we got rid of the rows and columns we don’t need in the analysis,

for example the years between 1960 and 1999, due to the fact that the analysis is only of the 21st century. The columns that contained codes (short for regions or countries) were also dropped.

* Once we had the data needed, some statistical summaries were checked grouped by countries.

For example, we made sure all the mortality values were positive and that the values by age

were always slower than the overall total.

**Data Table Schema**:

| **Field** | **Type** | **Description** |
| --- | --- | --- |
| **Region\_Name** | STRING - Category | Country Name. |
| **Year** | STRING - Category | Year (YYYY-formatted) from 2000 until 2019 |
| **Sex** | STRING - Category | 4 categories:  All, Male, Female and Unknown |
| **Age\_Group** | STRING - Category | [All], [45-49], [85+], [80-84], [75-79], [70-74], [65-69], [60-64], [55-59], [50-54], [40-44], [0], [35-39], [30-34], [25-29], [20-24]  [15-19], [10-14], [5-9], [1-4] and [Unknown] |
| **Number\_of\_deaths** | NUMERIC - Integer | The absolute number of deaths per 100 000 inhabitants |
| **Standardized\_death\_rate** | NUMERIC - Float | The rate that is a weighted average of the age-specific mortality rates per 100 000 people |
| **Death\_rate** | NUMERIC - Float | Death rate per group (absolute) |

1. **Economic growth**

Economic growth is a dataset that contains information on the relevant economic variables for

all the countries of the world. The variables and derivatives that were considered useful for the

purpose of our project are: GPD, energy use, exports, oil consumption and unemployment rate. All

the datasets were taken directly from the World Bank page.

The first four rows of the tables were not considered as they all have a default header of that length.

The GDP information is very complete and has almost none NaN items. Therefore, no extensive

cleaning was necessary.

**Data Profile**:

* After inspecting the columns of the tables, there was a column that contained a code that was

short for the variable the table was measuring. It was dropped as well as the years (between

1960 and 1999) and the tracking code.

* Once we had the columns needed, we had a look at the arrangement of the DataFrame and

decided that we had to change the table to a long format instead of a wide format, so that we

had the actual value in a column and the years in a single columns instead of a column for a single year.

* Once we had the format set, we made some grouping by *country* and checked some

statistical summaries such as min and max value as well as percentage of nulls and outliers.

For example checking whether there were negative values in the variable unemployment

or fossil\_fuel\_consumption, because in those variables only positive values make sense.

| **Field** | **Type** | **Description** |
| --- | --- | --- |
| **region** | STRING - Category | Regions of the world according to the model of the world Bank which is a variant of the OECD classification |
| **income** | STRING - Category | Income of the country. HIGH, MIDDLE or LOW INCOME |
| **country** | STRING - Category | Country Name. |
| **year** | DATE - datetime | Year (YYYY-formatted) from 2000 until 2021 |
| **energy\_use** | NUMERIC - Float | Use of energy measured in kg of oil equivalent per capita. |
| **exports** | NUMERIC - Float | Percentage of exports of goods and services in the GDP. |
| **fossil fuel consumption** | NUMERIC - Float | Fossil fuel energy consumption (% of total) |
| **gdp\_per\_capita** | NUMERIC - Float | Annual growth of gross domestic product per capita |
| **gdp\_growth** | NUMERIC - Float | Annual growth of gross domestic product |
| **inflation** | NUMERIC - Float | Inflation, GDP deflator (annual %) |
| **ren\_energy\_consumption** | NUMERIC - Float | Renewable energy consumption (% of total final energy consumption) |
| **revenue\_excluding\_grants** | NUMERIC - Float | Revenue, excluding grants (% of GDP) |
| **unemployment** | NUMERIC - Float | Unemployment, total (% of total labor force) (modeled ILO estimate) |

**Education**

Education is a dataset that contains relevant information about education index and measures for

all the countries of the world.  The Education DataFrame was created from different tables taken

from the World Bank Open Data.

Some studies show that the level of education is a predictor of the mortality

([Hummer RA, Hernandez EM](https://docs.google.com/document/d/11CID0jdI_xju2dBIHSeRufV20-7-x0aG3tkAdW-icDo/edit#bookmark=id.jqhvoeifryw0)) and therefore the following 6 relevant variables were considered

in the Education DataFrame: children out of school, Government expenditure on education,

Literacy rate youth total, Primary completion rate total, pupil teacher ratio primary, Literacy rate

total audit..

**Data Profile**:

* After inspecting the columns of the tables, there was a column that contained a code that was

short for the variable the table was measuring. It was dropped as well as the years (between

1960 and 1999) and the tracking code.

* Once we had the columns needed, we had a look at the arrangement of the DataFrame and

decided that we had to change the table to a long format instead of a wide format, so that

we had the actual value in a column and the years in a single columns instead of a column

for a single year.

* Once we had the format set, we made some grouping by *country* and checked some

statistical summaries such as min and max value as well as percentage of nulls and outliers.

For example checking whether there were negative values in the variable number of children

out of school, which would not have made sense.

| **Field** | **Type** | **Description** |
| --- | --- | --- |
| **region** | STRING - Category | Regions of the world according to the model of the world Bank which is a variant of the OECD classification |
| **income** | STRING - Category | Income of the country. HIGH, MIDDLE or LOW INCOME |
| **country** | STRING - Category | Country Name. |
| **year** | DATE - datetime | Year (YYYY-formatted) from 2000 until 2021 |
| **chil\_out\_school** | FLOAT | Children out of school (% of primary school age) |
| **gov\_expenditure\_education** | FLOAT | Government expenditure on education, total (% of GDP) |
| **literacy\_adult\_rate** | FLOAT | Literacy rate, adult total (% of people ages 15 and above) |
| **literacy\_youth\_rate** | FLOAT | Literacy rate, youth total (% of people ages 15 and above) |
| **primary\_completion\_rate** | FLOAT | Primary completion rate, total (% of relevant age group) |
| **pupil\_teacher\_ratio** | FLOAT | Pupil-teacher ratio, primary |

# Visualizations

The exploratory data analysis showed the following characteristics:

* Low correlation between variables.
* There was an atypical behavior of the variables in which they are displayed in a solitary way. In the cases they are grouped, they show a completely opposite trend, perhaps evidencing a simpson’s paradox.
* When models of linear and multiple regression are utilized with the variables, the results do not

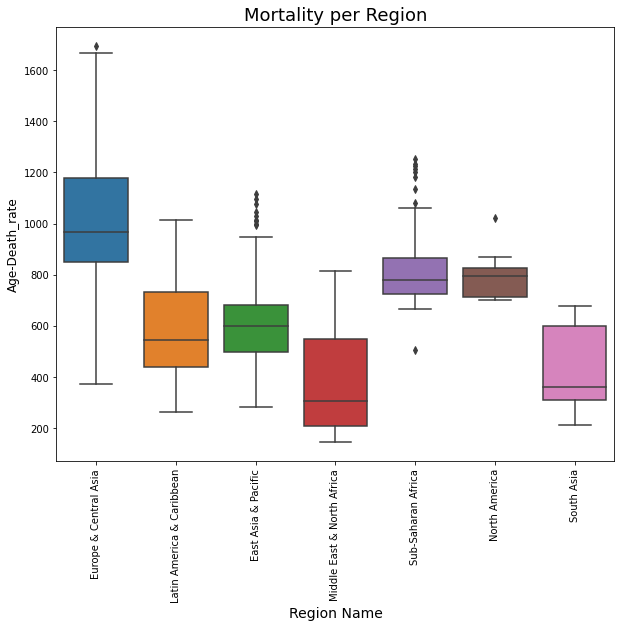
appear to fit with the models.

Gráfico de rectángulos

Descripción generada automáticamente

The heat map shows that the positive and negative correlations found are located within the

independent and dependent variables.



These box or whisker plots expresses the following characteristics of mortality by region:

* Europe and Central Asia have a right-skewed distribution, most countries in this region have

age-specific mortality rates between Q1 (interval 830) and Q3 (interval 1180) have a significant dispersion, the mean is close to interval 1000. Shows outliers after interval 1600.

* Latin America and the Caribbean whose distribution has a positive bias. The data of the

mortality rate by age are concentrated between Q1 (interval 530) and Q3 (interval 780),

presenting greater dispersion from Q2. It does not present outliers.

* East Asia and the Pacific have a slightly left-skewed distribution, their mortality rate is

concentrated between Q1 (interval 500) and Q3 (interval 700). Approaching the 1000 interval, outliers are displayed.

* The Middle East and North Africa have a positive bias and greater dispersion in the upper

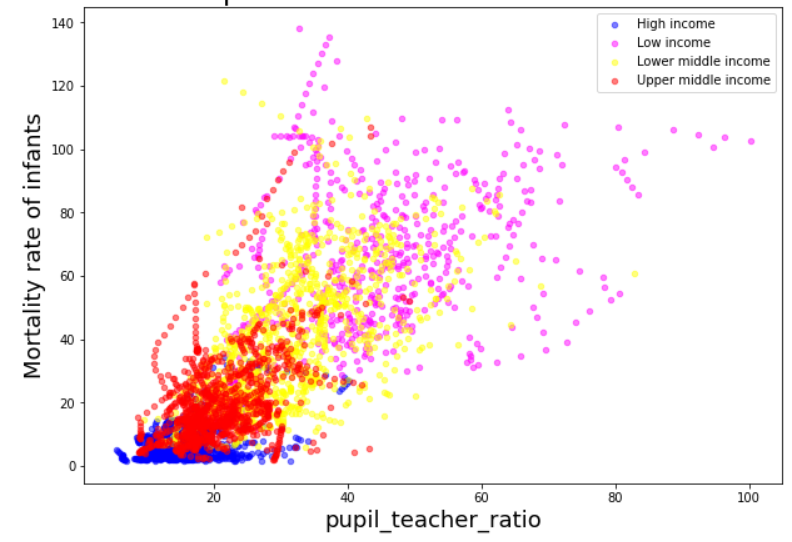
whisker. The mortality rate is concentrated between Q1 (interval 200) and Q3 (interval 570).

It does not present outliers.

* Sub-Saharan Africa is slightly skewed to the right, being more scattered in the tail, and

atypical values are observed before the minimum value and after the maximum value.

* North America has a negative or left bias, the mean is observed near Q3, it presents atypical values after the maximum value.



This graphic shows the relationship between the income of the countries and the infant mortality rate.

We may observe that high and upper middle income countries have a low mortality rate, while low and lower middle income countries have high mortality rates.

After concatenating the dataframes, we tested how the independent variables education and economy are

correlated with the new dependent variable mortality in infants. The following correlations were

obtained:



Aplicación

Descripción generada automáticamente con confianza bajaInterfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente con confianza media

We can see that the variables that correlate directly with infant mortality are pupil teacher ratio, energy

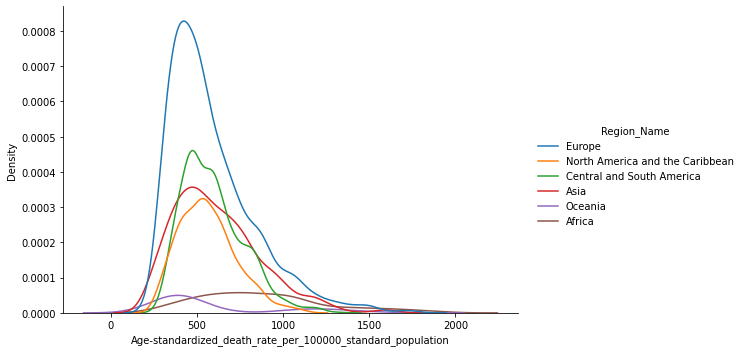
consumption, child out of school and inflation and surprisingly we see that it is indirectly related to primary completion rate, literacy and GDP.

The next whisker plots expresses the following characteristics of infant mortality rate by region:

* The region of the world with the highest infant mortality is Africa, followed by Asia, the Midwest and Latin America.
* Europe presents, on average, the lowest values ​​of infant mortality worldwide, but the USA presents the least variation of this indicator.
* The greatest variation in infant mortality is denoted in East Asia and this can be explained because it has a large population, and geographically and economically there are very diverse countries, which causes so much difference.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente



The frequency diagram that visualizes the Standardized mortality rate by age per 100,000 inhabitants shows

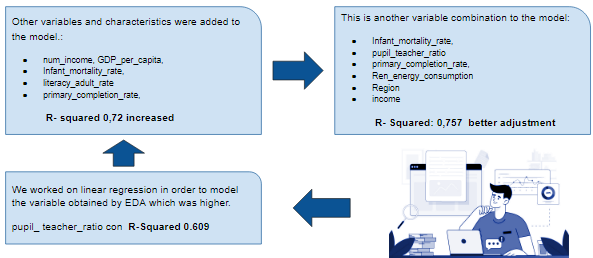
the following characteristics for the regions:

* Europe is skewed to the right with a leptokurtic data distribution where the mortality rate is concentrated between the intervals 0 and 1000. The mortality rate from the interval 1250 is high and there are atypical values beyond the interval 2000.
* North America and the Caribbean is skewed slightly to the right, the mortality rate of most countries is concentrated between 0 and 1000.
* Central and South America are skewed to the right with data concentrated between 0 and 1000, but with atypical values that are close to the 1500 interval.
* Asia’s distribution is observed to be skewed to the right, its data on the mortality rate are between the intervals 0 and 1250. In addition, it has atypical values that exceed the interval 2000 and that indicate a high mortality rate.
* Oceania data is observed skewed to the right, its distribution is platykurtic and its values are concentrated between 0 and 550, but from this interval atypical data are displayed that exceed the 2000 interval.
* Africa is skewed positively or to the right, its data presents a greater dispersion which is expressed in its platykurtic or flattened distribution.

In the modeling part, certain combinations were made to find a relationship between the variables,

the results allow the analysis to be extrapolated to real life and see how the multiple combined

variables affect each other.



* In the first correlation we can see that the GDP, the illiteracy rate, the grade of primary school completed, and infant mortality are correlated with a r square of 0.72, which indicates that they

are directly proportional to each other.

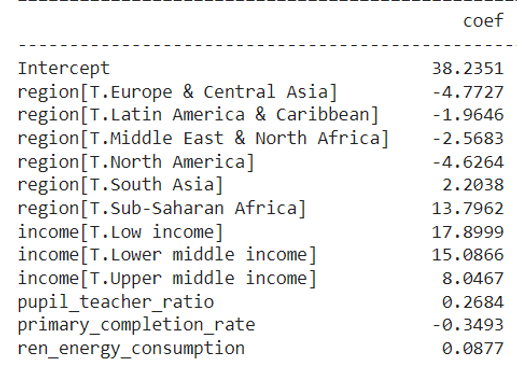
* While infant mortality is directly proportional to the teacher ratio, energy consumption, income

and primary completion rate with a r square of 0,757.

* The variable most closely related to the infant mortality rate is the pupil teacher ratio.

Perhaps the most useful information that we can analyze at a geographical level is how correlated

infant mortality is depending on the region, which we can answer with the following graph:

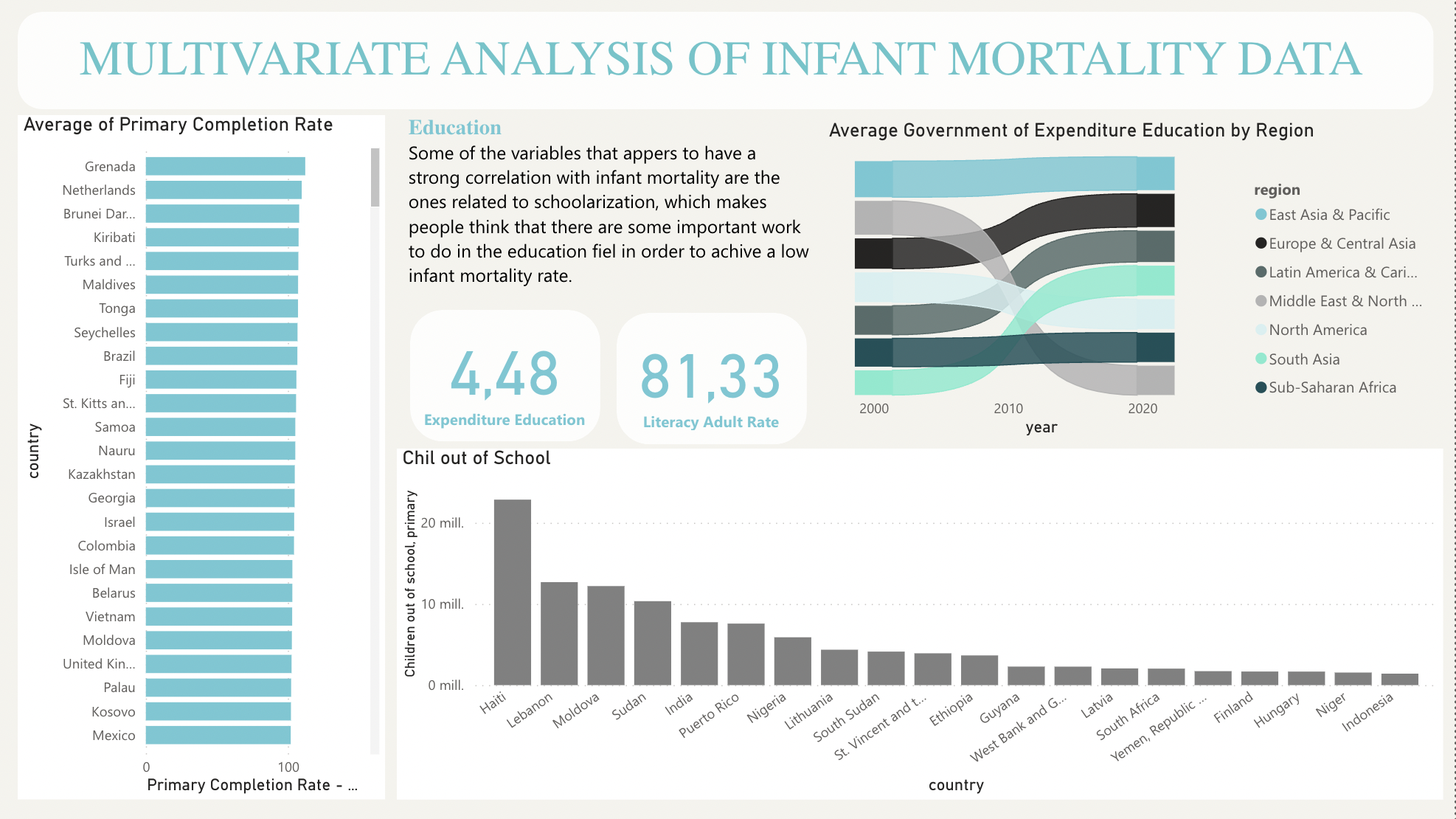
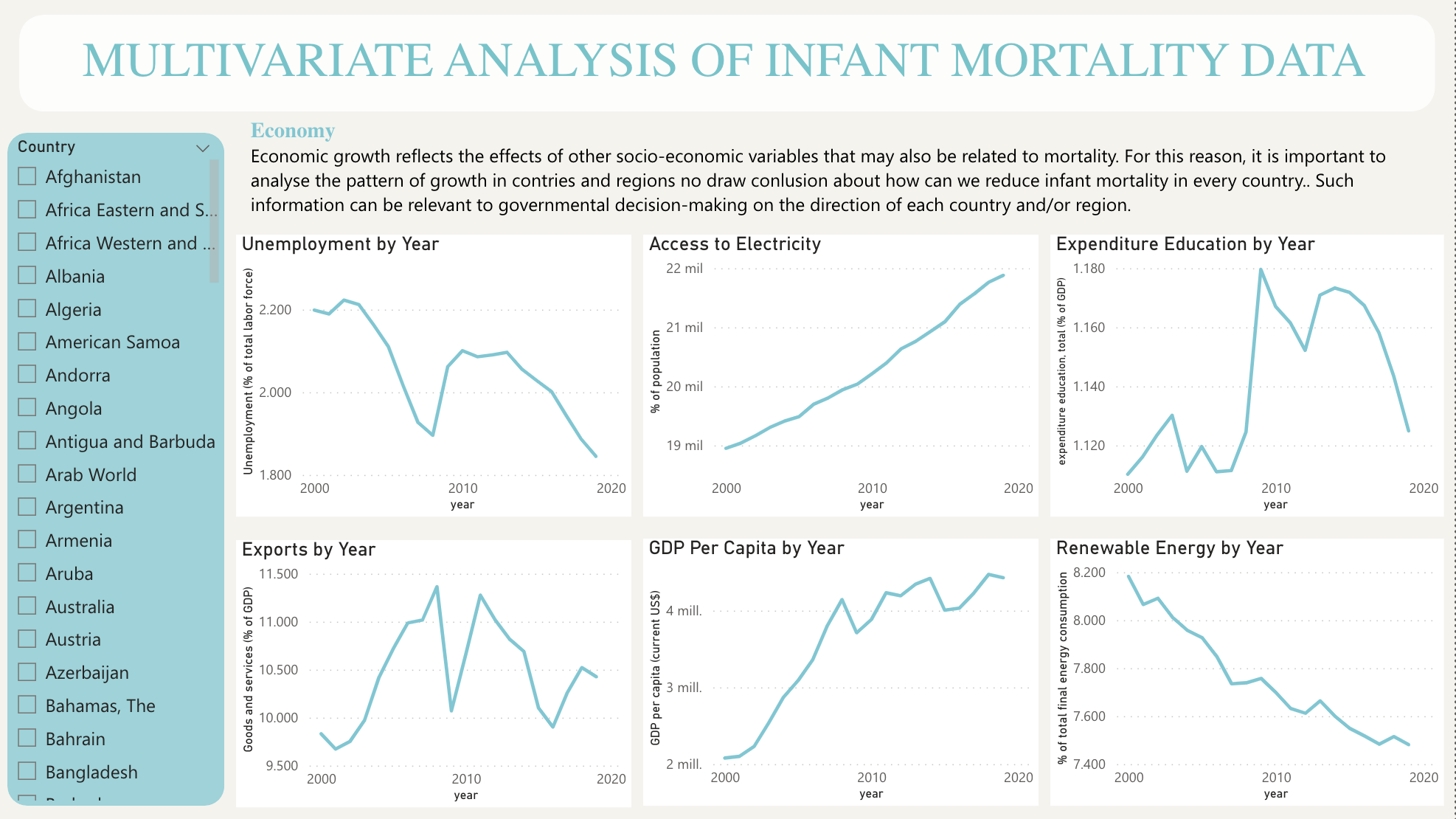
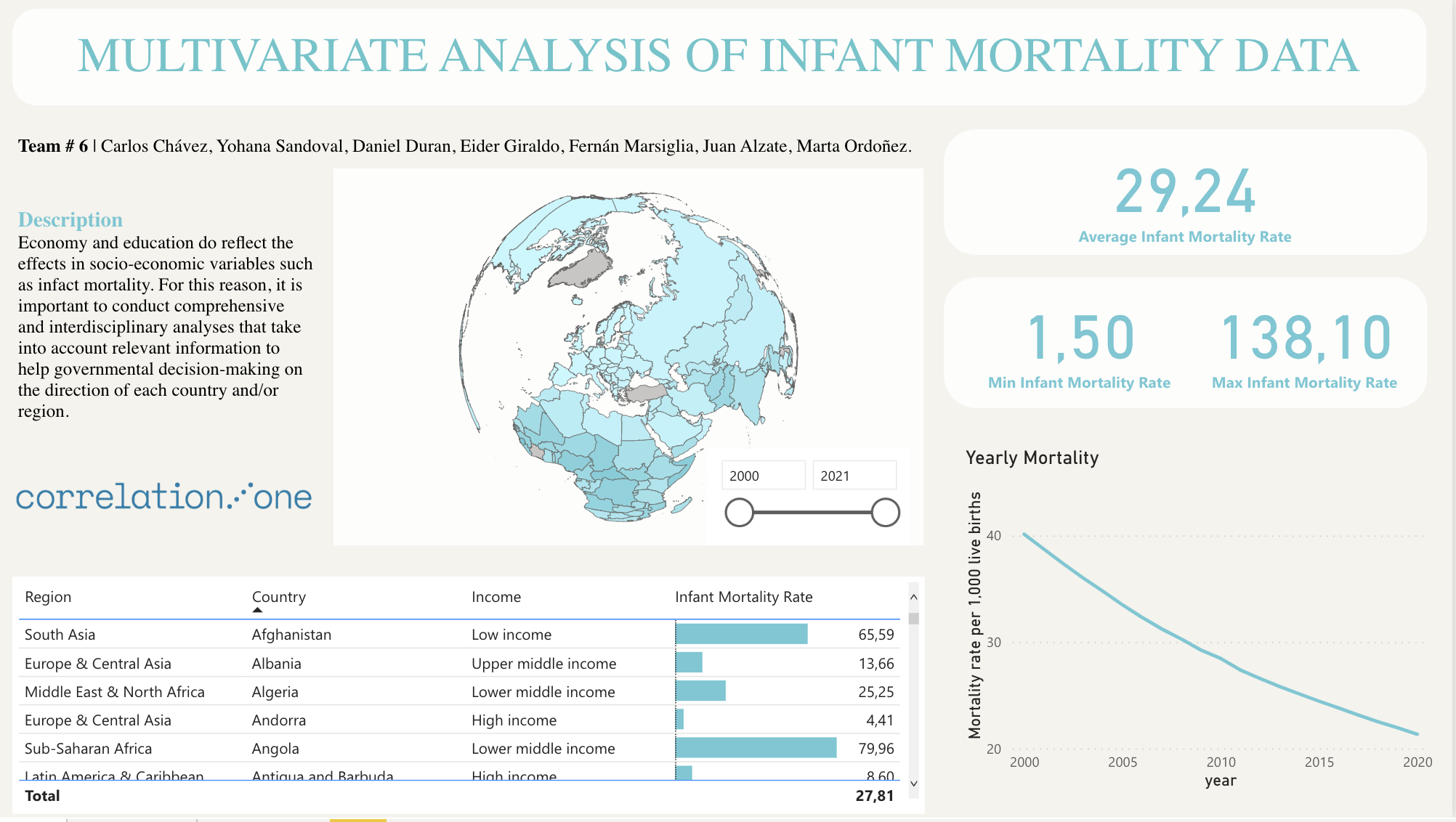


The positive coefficients indicate direct proportionality, therefore we can confirm that the regions of

the world that are most impacted by a negative infant mortality rate are Africa with 13,79 and East

Asia with 2,2.

The linear regression models showed an increase in the adjustment as more variables were added to the calculation. We can also notice that after the initial mistake was corrected, we started to obtain results that were closer to what we initially expected, i.e., countries with more resources had less mortality rate than those with less economical power.



# Conclusions

Infant mortality is inversely proportional to the income level of a country or region. In turn, the countries or regions with the highest pupil-teacher ratio tend to reduce the cases of infant mortality. On the other hand there is also a linear relationship between education and mortality. The more educated a country is, the less infant mortality it will have. However, the impact of these two variables is quite different in the mortality of infants. There is a stronger relationship between the lack of education and infant mortality whereas the impact of the economy is considerably lower in the same variable. It tells us how important education is when a country wants to reduce those mortality rates and where should they focus to achieve that goal.

# Recommendations

Some information might depend on the country and the year and is not totally independent as assumed in the linear model. For such a reason, we propose to carry out a generalized linear mixed model (GLMMs) due to the fact that there may be some repeated measures such as country and time and the inter subject variance could be analyzed differently.

It is also important to consider in future studies the role of some variables that were not taken into account in this project, such as population of the country, age of the mother, blood type and father's presence during pregnancy among others. This is because according to some studies, they may play an important role in mortality and are usually not considered.

# Limitations

There are more variables that affect the study, so a more complete analysis would require much more information on the internal behavior of the states.

All the analysis is based on the data provided. Any bias in the collection of the data may have resulted in erroneous conclusions. For example, In many low-income countries, information about mortality is not reliable. It is mainly because either the system for recording such information does not work properly or it does not exist which can lead to some bias in any analysis. It is also important to consider that there are many ways of analyzing the same data and it depends on the spertice, interests and resources, for that reason, there might be other more specific and complex approaches that were not considered by this team.