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

Data Curation  
  
Team #6 | Carlos Chávez, Yohana Sandoval, Daniel Duran, Eider Giraldo, Fernán Marsiglia, Juan Álzate, Marta Ordoñez, Mateo Moreno, José López.

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

**Dataset Details**: initially the dataset had a total of 300,258 rows, 12 columns and was taking approximately 27.5+ MB of space.

After the data curation was done, the Dataset ended up with a total of 125,895 rows, 8 columns and it takes approximately 8.6+ MB of space.

**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 Data Wrangling**:

* In this specific table we did not merge any specific dataset, however we did a transformation of the year column to a date time format and we did not change the missing values because they were already treated as [unknown] in the dataset.
* There were quite a few outliers, however when we inspected the data carefully and grouped by country the outliers ended up being normal data points.

**Data Sourcing**:

**Name:** WHO - World Health Organization

**Description:** It is a compilation of mortality data; it is organized by country and area, year, sex, age and cause of death.

**Size**: 8.6+ MB

**URL:**<https://platform.who.int/mortality/themes/theme-details/mdb/noncommunicable-diseases>

**Method:** Taken from a URL in a GitHub repository.

**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) |

2. 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.

**Dataset Details**: Initially there were 2394 rows / 24 columns 467.6 kb. At the end 5896 filas / 13 columns y 506.8 kb

**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.

**Data Data Wrangling**:

* The different tables were concatenated one on top of each other (by columns) due to the fact that they all had the [World Bank Data](#mckhy5pd3ig) format.
* After that we changed the wide format for a long format. We grouped the date by both, years and countries and carried out a linear interpolation to replace the missing values in every column.
* Finally , the name of the variables were slightly changed to have a better control of the dataframe and make it more understandable.

**Data Sourcing**:

**Name:** [World Bank Data](#mckhy5pd3ig)

**Description:** It is a compilation of economic index datasets; it is organized by country, region, income and year and contains information about 9 economical growth variables.

**Size**: 506.8+ MB

**URL:** <https://databank.worldbank.org/>

**Method:** Taken from a URL in a GitHub repository.

| 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) |

3. 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](#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 adult total.

All datasets were taken directly from the World Bank Open Data page, loaded into [GITHUB repository](#4a2pcr7tmikh) and then loaded in a GoogleColab project.

**Dataset Details**: Initially there were 1596 rows / 24 columns 311.7 kb. At the end 5852 rows/ 10 columns 365.9 kb

**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.

**Data Data Wrangling**:

* The different tables were concatenated one on top of each other (by columns) due to the fact that they all had the [World Bank Data](#mckhy5pd3ig) format.
* After we changed the wide format into a long format, we grouped the date by both, years and countries and carried out a linear interpolation to replace the missing values in every column.
* There were some missing values in the whole period for countries like Venezuela or Afghanistan, in those cases the missing values were left as they were and the team is considering getting rid of those countries for the analysis or else work with the regions rather than the countries.
* Finally , the name of the variables were slightly changed to have a better control of the dataframe and make it more understandable.

**Data Sourcing**:

**Description:** It is a compilation of educational index datasets; it is organized by country, region, income and year and contains 6 columns with information about education.

**Size**: 365.9+ MB

**URL:** <https://databank.worldbank.org/>

**Method:** Taken from a URL in a GitHub repository.

| 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 |

**References**

1. Hummer RA, Hernandez EM. *The Effect of Educational Attainment on Adult Mortality in the United States*. Popul Bull. 2013 Jun;68(1):1-16. PMID: 25995521; PMCID: PMC4435622.

1. <https://github.com/juanpaat/DS4A-Medellin.git>

1. <https://databank.worldbank.org/>

1. <https://platform.who.int/mortality/themes/theme-details/mdb/noncommunicable-diseases>