Data Wrangling – Project Report

Investigating the correlation between GDP and child mortality, specifically regarding HIV/AIDS

Joe Hopkinson, 2612836
Armin Shokri Kalisa, 2612801
Eveline Kleijne , 2620873

Research question

How does a country's GDP (per capita) correlate with its child (<4 years) mortality rates, specifically regarding HIV/AIDS?

- If there is a correlation, is it specific to any given region?
- Can we observe this correlation in most countries?
- Does the population of a country have an impact on mortality rates?

Data sources

World Health Organization. (2020, January 19). *Number of deaths by country HIV/AIDS.* Retrieved from GHO | By category | Number of deaths by country - HIV/AIDS

World Bank. (2020, January 19). *GDP per capita, PPP.*
Retrieved from https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD

Data.World. (202, January 24). *World Population JSON*
Retrieved from https://data.world/edmadrigal/world-population-json

United Nations. (2020, January 24). *World Population Prospects 2019.*
Retrieved from https://population.un.org/wpp/Download/Standard/CSV/

Data wrangling methods

Data Retrieval

Our initial two datasets regarding GDP and mortality rates were both extracted via CSV format, and therefore were simple to process into data frames in order to be cleaned. Following some initial data exploration of the initial two datasets, we determined that it would be appropriate to introduce some data regarding the population of the nations in our study. Two datasets (JSON and an additional CSV) were introduced and later cleaned.

Data Cleaning

Once our datasets were processed, we could begin preparing the data for exploration. Following the initial inspections of our datasets, we determined that we must first specify a timeframe for our investigation. Our HIV dataset only included data from 2000 onwards, furthermore it soon became apparent that our GDP dataset wasn't complete preceding that year either. We concluded that a ten year period between 2007-2017 would give us a near complete pool of reliable data to work with. Not only was it required to remove the excess years, there was also present a number of redundant columns including data such as country and indicator codes that were not necessary. The countries with a number of null (NaN) entries were also removed as they would cause numerous issues regarding processing the data and the results themselves. We followed the same procedures with our later introduced population datasets.

The individual countries in the HIV dataset were initially split into three age categories so it was required to concatenate them into one single figure (below 4 years). It was also necessary to convert the GDP dataset's column names into a string value, in order to be comparable with the HIV dataset (e.g. 2007.0 to '2007'). Once a cleaned data-frame had been developed from our GDP and HIV datasets, we exported them into new CSV files for ease of use in other notebooks. Throughout the data exploration phase of our project, we also created numerous versions of the initial datasets (e.g. pivot tables / changing indexes etc.) that required some degree of manipulation to the structure of the data.

Data Exploration

The initial exploration phase of our project consisted of creating some statistics that reflected a general overview of the data that we had at our disposal. This included summing, sorting and taking the mean of various properties of our data in order to create the initial visualisations. From our bar plot of the countries with the highest total deaths over our time period, we observed that all 10 countries were African. This inspired us to create an overview of the locations of the top 50 countries with the most deaths. Through the use of the python library PyCountry, we were able to specify the continent of any given country, and therefore create a pie plot representing the continents of the top 50 countries regarding total deaths. This visualisation saw that 78% of the top 50 were African nations, and therefore our final conclusion will likely reflect the situation in this specific continent.

Once we had an overview of some general statistics of our data, we plotted a regression plot of the mean GDP against the mean number of deaths over our time frame. This returned a correlation score of only -0.188. Through this exploration, as well as the use of the PyCountry pie plot, we were able to determine that it would be difficult to create an accurate correlation score for all nations as most countries' child mortality rates (HIV relating) are relatively low; it is not possible for child mortality to fall below zero. Because of this, it was decided that we would analyse the nations with significant mortality rates in order to see a more substantial correlation. We began by overlaying the GDP over time onto the death count for the countries with the highest total deaths (2007-2017). We did this by creating two pivot tables per country for HIV and GDP, merging the two and then plotting the result. Although we observed significant correlation between the two variables, it soon occurred to us that it was difficult to compare individual countries as population was not a constant value. In order to greater reflect reality, we introduced the population dataset and plotted the top 3 total deaths per 1,000,000 citizens against their GDP's. Mozambique remained in the top 3, however South Africa and Nigeria were replaced by Lesotho and Zimbabwe. We similarly observed a general correlation for all 3 nations, however Lesotho's death per 1,000,000 did in fact rise over a three year period as its GDP continued to also rise, contrary to our hypothesis.

Our final exploration of the data was to determine a Pearson's correlation score for the top 50 countries regarding total child deaths over our time period. This would give us an impression of whether there is in fact correlation in the nations where we have significant data to work with (death count much greater than zero). This process involved taking a correlation score for each country for the ten year period, and then taking the mean of that figure. This gave us a correlation score of -0.583.

Conclusion

The initial exploratory phase of our project showed us that it would be difficult to draw conclusions regarding the correlation between GDP and child mortality rates due to HIV for countries with low child mortality rates. This is because, for a significant number of wealthier nations, since their death figures are so low, it is not possible to see the same correlation as their GDP continues to rise. It is not possible to have a negative number of deaths. We were however able to observe some degree of correlation between GDP and child mortality rates in countries with high death counts between 2007-2017, specifically the top 50. The Pearson correlation score of -0.583 allows us to suggest that there perhaps is some correlation between the variables for these countries. With 78% of the top 50 countries being African, we can suggest that the correlation that we've observed is present in that continent.

In answer to the first sub-question, we cannot conclude that this is specific to this region however, rather the fact that this is where the most usable data can be found. Similarly, it is difficult to determine if this correlation can be seen in most countries, for the reasons previously discussed. What we can conclude though is that it is not the case in every country, Turkmenistan being a clear example. Finally, we observed that the countries with the highest total deaths were not necessarily the countries with the highest total deaths per 1,000,000 residents. Although this does not suggest much regarding the impact population has on mortality, it did allow us to compare nations on a more proportionate basis.

This all being said, the visualisations we created as well as the correlation figure of -0.583 cannot, with confidence, conclude that this correlation exists and no causation can be determined. The correlation score is far from overwhelming and it is clear that there are many other factors in play that have not been taken into account.

Limitations

Our project only took into account population as an external factor regarding child mortality rates due to HIV, however there are many more that in future research would allow a more in depth analysis of the situation. The first of which is data regarding humanitarian aid or other large scale projects in dealing with HIV. These projects are often deployed in nations with low GDPs, and therefore could influence a downward trend in HIV related child deaths, independent to GDP. This is arguably also the case regarding general advancements in the capability and affordability of medical treatment for the disease. Moreover, the availability of these medicines may be determined by more geographical or political factors as opposed to simply a nation’s wealth. Different governments may not prioritise the investment in this area and therefore make less contribution to reducing the general health and well-being of the citizens, specifically regarding HIV induced child deaths. This may include the investment in new medicines or simply putting less emphasis on improving the quality of life for a countries residents. Finally, future research on this topic would be wise to take into account general inflation in GDP or the occurrences of any financial crises over the time period, in order to observe a more accurate comparison between the two variables in real terms.