INFO2950 Final Project - Phase 2: Data Description

1. Total confirmed cases of COVID-19 data

**What are the observations (rows) and the attributes (columns)?**

* The observations are groups of populations in the world (divided by country, continent, region, and income level). The attributes are entity, code (alpha3 code of each country), date, and total confirmed cases.

**Why was this dataset created? Who funded the creation of the dataset?**

* The original dataset of “Total Confirmed cases of COVID-19”, founded on *Our World in Data*, was collected and created by the European Center for Disease Prevention and Control (ECDC), who publishes daily statistics on COVID-19 from around the world.  This dataset is used to help public health professionals and healthcare providers monitor the spread of COVID-19 around the world and support better understanding of illness and disease severity associated with the disease.

**If people are involved, were they aware of the data collection and if so, what purpose did they expect the data to be used for?**

* In the data collection process, patients should be aware, to some extent, that results regarding their testing will be used for statistical means to better track the spread and status of the disease.

**What processes might have influenced what data was observed and recorded and what was not?**

* According to the WHO guidelines, cases of COVID-19 are broadly defined under three systems: suspected case, probable case, and confirmed case. Here, we use the data of confirmed cases, which are based on a positive laboratory confirmation of the test, because they have a higher degree of certainty and they provide standardized comparisons between countries.
* The WHO and ECDC report case figures submitted by national governments and reporting institutions, who gather laboratory testing results form state, local, and territorial public health departments, and commercial laboratories.

**Identify any potential problems with your dataset.**

* Since each country has slightly different criteria for how cases are defined and reported, there might be some inaccuracy and discrepancy in the data presented. For example, the U.S. CDC only reported the daily number of confirmed cases until 14th April, and then it switched to reporting the sum of confirmed and probable cases. This transition in reporting method will potentially lead to inconsistency in the data, which might limit the significance of the analyses.

**What preprocessing was done, and how did the data come to be in the form that you are using?**

* *Our World in Data* preprocessed the data by standardizing names of countries and regions to its own standard entity names, and they also corrected or discarded inconsistencies detected in the original data.

**Where can your raw source data be found, if applicable?**

* <https://github.com/owid/covid-19-data>

1. Age demographics data

**What are the observations (rows) and the attributes (columns)?**

* Median Age (median-age.csv)
  + There are four columns: entity, country code, year, the median age
* Age breakdown (population-by-broad-age-group.csv)
  + There are eight columns: entity, country code, year, under-5s, 15-24 years, 25-64 years, 65+ years, 5-14 years.

**Why was this dataset created?  Who funded the creation of the dataset?**

* The original dataset of median age and age breakdown, founded on Our World in Data, was prepared and created by the Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat.

**If people are involved, were they aware of the data collection and if so, what purpose did they expect the data to be used for?**

* The principal data source related to median age and age breakdown was based on the information from 1,690 population and housing censuses for 235 countries or areas, including 236 censuses conducted since 2010. Participants were aware of the data collection and were acknowledged that the data they provided was one of the primary sources of data needed for formulating, implementing and monitoring policies and programs aimed at inclusive socioeconomic development and environmental sustainability.

**What processes might have influenced what data was observed and recorded and what was not?**

* Most countries conduct a population census once per decade. In some countries, population registers based on administrative data systems are sufficiently well developed to serve as a basis for population estimates.

**Identify any potential problems with your dataset.**

* The Population Division of the United Nations estimates historical demographic trends for the period from 1950 to the present and projects future population trends out to 2100. The estimates are based on all available sources of data on population size and levels of fertility, mortality, and international migration for 235 distinct countries or areas comprising the total population of the world.
* Therefore, some countries might not have a population census. Also, since the data are all self-reported, by filling out the census, there might be some discrepancy between real life situations and what is shown on the data.
* I aimed to choose the latest data available, so I had median age data in the year of 2020, and age breakdown data in the year of 2015. There might be some inaccuracies when using the 2015 data, but since population demographics won’t have that big of a change in 5 years, it is considered acceptable in this case.

**What preprocessing was done, and how did the data come to be in the form that you are using?**

* The Population Division used the [cohort-component method](https://www.un.org/en/development/desa/population/publications/manual/projection/sex-age.asp) to ensure internal consistency by age and sex and over time, and between the three demographic components of change (fertility, mortality, and migration) and the enumerated population. The cohort-component method was also used to project population trends until 2100 using a variety of demographic assumptions concerning the components of population change.

**Where can your raw source data be found, if applicable?**

* <https://population.un.org/wpp/Download/Standard/Population/>

1. Economics indicator data

**What are the observations (rows) and the attributes (columns)?**

* GDP per capita (API\_NY.GDP.PCAP.PP.CD\_DS2\_en\_csv\_v2\_988619.csv)
  + The columns are country name, country code, indicator name, indicator code, and columns for each year from 1960 to 2019.
* Unemployment (unemployment.csv)
  + The columns are country name, country code, indicator name, indicator code, and columns for each year from 1960 to 2019.

**Why was this dataset created?  Who funded the creation of the dataset?**

* The original dataset of GDP per capita and unemployment were founded on *The World Bank*. At the World Bank, the Development Data Group coordinates statistical and data work and maintains a number of macro, financial and sector databases. Working closely with the Bank’s [regions](http://www.worldbank.org/en/country) and Global Practices, the group is guided by professional standards in the collection, compilation, and dissemination of data to ensure that all data users can have confidence in the quality and integrity of the data produced.

**If people are involved, were they aware of the data collection and if so, what purpose did they expect the data to be used for?**

* People are aware that governments/organizations are collecting and analyzing data to calculate GDP and unemployment rate. These are common economic indicators, so people know that their data are the sources for measuring the well-being of the country.

**What processes might have influenced what data was observed and recorded and what was not?**

* Many factors affect data availability, reliability, and comparability. Statistical systems in many of the poorest countries are limited; statistical methods, coverage, practices, and definitions differ widely; and cross-country and intertemporal comparisons involve complex technical and conceptual problems that cannot be resolved unequivocally. Data relevant at the national level may not be suitable for standardized international use due to methodological concerns or the lack of clear documentation.

**Identify any potential problems with your dataset.**

* Delays in reporting data and the use of old surveys as the basis for current estimates may further compromise the quality of data reported. Data coverage may not be complete because of special circumstances affecting the collection and reporting of data, such as problems stemming from conflicts. Considerable effort has been made to standardize the data, but full comparability cannot be assured, so care must be taken in interpreting the indicators.
* I aimed to choose the latest data available, so I had GDP data in the year of 2018, and unemployment data in the year of 2019. There might be some inconsistency between the years.

**What preprocessing was done, and how did the data come to be in the form that you are using?**

* The World Bank uses aggregation rules to compile it’s data. The aggregation rules are intended to yield estimates for a consistent set of economies from one period to the next and for all indicators. Small differences between sums of subgroup aggregates and overall totals and averages may occur because of the approximations used.
* Details about methodologies that the World Bank used to compile data: <https://datahelpdesk.worldbank.org/knowledgebase/articles/906531-methodologies>

**Where can your raw source data be found, if applicable?**

* GDP: <https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD?end=2018&most_recent_value_desc=true&start=2018&view=map>
* Unemployment rate: <https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS?view=map>

1. Healthcare system indicator data

**What are the observations (rows) and the attributes (columns)?**

* My dataset is consisted of 6 health system indicators (columns) among 264 countries (rows) , including number of hospital beds per 1000 people, health expenditure per capita, health expenditure out of pocket per capita, number of physicians per 1000 people, and people with basic hand-washing facilities including soap and water (% of population), and testing coverage (total tests per thousand).

**Why was this dataset created? Who funded the creation of the dataset?**

* First 5 sets of the original data were found on The World Dank and from World Health Organization's Global Health database, including rows—country name, country code, indicator name, indicator code, values of the indicators from 1960 to 2019, and an unnamed column. The rows represent information regarding each country. Testing coverage data was found on Our World in Data. The first 5 factors may be recorded to track each country's well-being of health system, while the testing coverage data has been closely monitored to record as well as facilitate covid-19 fighting progress.

**If people are involved, were they aware of the data collection and if so, what purpose did they expect the data to be used for?**

* The first 5 sets of data may be collected by governments as people are somewhat aware every year while testing coverage may be collected under people’s awareness during this pandemic. From the first 5 indicators we may access each country’s health system’s preparedness while testing coverage data may affect the reported confirmed rate in spite of the real confirmed rate.

**What preprocessing was done, and how did the data come to be in the form that you are using?**

* Preprocessing and processing procedures are similar to the aforementioned.

**What processes might have influenced what data was observed and recorded and what was not? Identify any potential problems of the dataset.**

* For the first 5 health indicators. I aimed to take the latest data on the basis of most countries included. Hence, I chose 2015 data for hospital beds and 2017 data for the other 4 factors. This, as a result, may lead to inaccuracy by overlooking some countries' health system improvements in the most recent years, but relative strength among countries regarding health systems should not vary vastly in such a short period of time, so our indicators are still useful for comparisons. In addition, we use testing coverage data on April 29 to parallel with the chosen date of confirmed rates.

**Where can your raw source data be found, if applicable?**

* **<https://docs.google.com/document/d/1GtycyIeDMVE-42pVjp4G0_oxVhCYK9Y41hz5bgb5dYU/edit?usp=sharing>**