Data Analysis Individual Project

Deliverable 1: Dataset Selection & Description

Dataset chosen: The complete Our World in Data COVID-19 dataset as of 9/23/2020

Hyperlinks: https://ourworldindata.org/coronavirus-source-data
https://github.com/owid/covid-19-data/tree/master/public/data/
https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-codebook.csv

Dataset description

Domain of this dataset - Health

Size of the dataset, i.e. required storage:

- as a csv file, it is 9.55 MB
- as an excel workbook file, it is 7.16 MB

Metadata

Column name & data type	*Brief field description*	*Source*
iso_code - nominal	Three-letter country codes (in all caps)	International Organization for Standardization
continent - nominal	The continent of the geographical location	Our World in Data
location - nominal	Geographical location	Our World in Data
date - interval	The date of the observation in each record	Our World in Data
total_cases - ratio	Total confirmed cases of COVID-19	European Centre for Disease Prevention and Control
new_cases - ratio	New confirmed cases of COVID-19	European Centre for Disease Prevention and Control
new_cases_smoothed - ratio	New confirmed cases of COVID-19 (7-day smoothed)	European Centre for Disease Prevention and Control
total_deaths - ratio	Total deaths attributed to COVID-19	European Centre for Disease Prevention and Control
new_deaths - ratio	New deaths attributed to COVID-19	European Centre for Disease Prevention and Control
new_deaths_smoothed - ratio	New deaths attributed to COVID-19 (7-day smoothed)	European Centre for Disease Prevention and Control
total_cases_per_million - ratio	Total confirmed cases of COVID-19 per 1,000,000 people	European Centre for Disease Prevention and Control
new_cases_per_million - ratio	New confirmed cases of COVID-19 per 1,000,000 people	European Centre for Disease Prevention and Control
new_cases_smoothed_per_million - ratio	New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people	European Centre for Disease Prevention and Control
total_deaths_per_million - ratio	Total deaths attributed to COVID-19 per 1,000,000 people	European Centre for Disease Prevention and Control
new_deaths_per_million - ratio	New deaths attributed to COVID-19 per 1,000,000 people	European Centre for Disease Prevention and Control

new_deaths_smoothed_per_million - ratio	New deaths attributed to COVID-19 (7-day smoothed) per 1,000,000 people	European Centre for Disease Prevention and Control
total_tests - ratio	Total tests for COVID-19	National government reports
new_tests - ratio	New tests for COVID-19	National government reports
new_tests_smoothed - ratio	New tests for COVID-19 (7-day smoothed)	National government reports
total_tests_per_thousand - ratio	Total tests for COVID-19 per 1,000 people	National government reports
new_tests_per_thousand - ratio	New tests for COVID-19 per 1,000 people	National government reports
new_tests_smoothed_per_thousand - ratio	New tests for COVID-19 (7-day smoothed) per 1,000 people	National government reports
tests_per_case - ratio	A rolling 7-day average of tests conducted per new confirmed case of COVID-19	National government reports
positive_rate - ratio	A rolling 7-day average of the share of COVID-19 tests that are positive	National government reports
tests_units - nominal	Units used by the location to report its testing data	National government reports
stringency_index - ratio	Government Response Stringency Index from 0 to 100 (100 = strictest response)	Oxford COVID-19 Government Response Tracker
population - ratio	Population in 2020	The U.N.
population_density - ratio	Number of people divided by land area, measured in square kilometers	World Bank
median_age - ratio	Median age of the population	The U.N.
aged_65_older - ratio	Share of population that's 65 years & older	World Bank
aged_70_older - ratio	Share of the population that is 70 years and older in 2015	The U.N.
gdp_per_capita - ratio	Gross domestic product at purchasing power parity (constant 2011 international dollars)	World Bank
extreme_poverty - ratio	Share of the population living in extreme poverty	World Bank
cardiovasc_death_rate - ratio	Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people)	Global Burden of Disease Study 2017 Results, Global Burden of Disease Collaborative Network
diabetes_prevalence - ratio	Diabetes prevalence (% of population aged 20 to 79) in 2017	World Bank
female_smokers - ratio	Share of women who smoke	World Bank
male_smokers - ratio	Share of men who smoke	World Bank
handwashing_facilities - ratio	Share of the population with basic handwashing facilities	The U.N.
hospital_beds_per_thousand - ratio	Hospital beds per 1,000 people	OECD, Eurostat, World Bank, national government records and other sources
life_expectancy - ratio	Life expectancy at birth in 2019	The U.N.
human_development_index - ordinal	Summary measure of average achievement in key dimensions of human development	The U.N.

1) a. All of this data on the Coronavirus pandemic was aggregated by Our World in Data which is a project and website put together by a collaborative effort between researchers at Oxford University in England and the non-profit organization Global Change Data Lab.

b. Their purpose, broadly speaking, is to publish data and tools for analyzing that data which are publicly available in order to help make the world a better place. To lift a quote from their website's about page, "To work towards a better future, we also need to understand how and why the world is changing."

2) a. The reason they sought out and aggregated so much data on various aspects of the impacts of the Coronavirus pandemic from so many different sources is to allow curious members of the public (with or without formal training in statistics, inductive inference, epidemiology, or data analysis) to conduct their own investigations of much of the relevant data on it so they can come to their own conclusions and become more informed about it.

b. This is a big data problem because there are over 45,000 records in the dataset (volume), 42 fields (variety) on well over a hundred different countries all over the world (variety), and the figures are updated daily (velocity).

c. There are no significant privacy issues with any of this data because no individual names or other identifying information are included for any specific persons. As for data quality, there are potential issues with data consistency across different countries and regions because they may be gathered in different ways by different entities and institutions with different data collection protocols and even slightly different definitions or interpretations of the definitions of some of the variables. Another potential data quality issue is places with lest COVID-19 testing might be underestimating the true number of cases. Countries which are poor and underdeveloped without very much modern medical infrastructure might be undercounting the number of deaths from the virus because many people in these countries may get sick and die in their homes rather than in a hospital where they can be counted. All medical tests generate both false positives and false negatives, it is unclear without looking into it further which is likely to outweigh the other when it comes to COVID-19 tests. One further potential quality of data problem here is that some doctors or nurses may put COVID-19 as the only cause of death or a cause of death for persons who already had one or more underlying health condition to such an extent that COVID-19 might better be thought of as a catalyst for their death rather than the cause, but this is a tricky problem to settle epistemologically speaking.

3) What potential questions could be answered by studying this data?

a. What is the average fatality rate for COVID-19 overall (worldwide)?

What is the fatality rate for a handful of different countries?

What is the fatality rate for different States in the United States?

What is the average transmission rate worldwide?

What is the transmission rate for the United States?

Did places with very different public policy approaches to combating COVID-19 have different outcomes?

What is the relationship between population density and total cases?

Does having a higher GDP per capita lower the fatality rate?

Does "flattening the curve" have any effect (so far) on the total number of deaths in the long run?

4) For hardware, I just purchased an extra 8 GB of RAM for my laptop last week which came with 8, and I installed it on Wednesday, the 23rd of September, bringing the total up to 16 GB of memory so that I can conduct the various parts of my analysis of this data more quickly. I also bought a 250 GB external hardrive several weeks ago, if I run into any problems with data storage space on my laptop, I can use that or my 120 GB flashdrive. As for software, I plan on definitely using MySQL, Python, and R in my analysis and possibly also Tableau Public or Microsoft Excel should the opportunity for their use present itself during the analysis.