6.1 Sourcing Open Data

Data Source

Summary: The "Novel Corona Virus 2019 Dataset" is an external source found on Kaggle. The data is collected from the Center of Systems Science and Engineering at Johns Hopkins University, and is considered trustworthy.

Data collection: The data is collected through departments of health services in every state of the USA, as well as government data from countries all over the world. Other aggregated data sources used in the dataset is the World Health Organization, European Centre for Disease and Prevention Control, and CDC.

Data content: The dataset contains serial numbers, date of observations (from January 2020-April 2021), Province/States, Countries/Regions, last updates, confirmed cases of virus till that date, number of deaths till that date and number of recovered till that date.

Resources

https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

https://github.com/CSSEGISandData/COVID-19

I chose this data after spending several hours searching open data sources to fit my interests for the project. With the criteria for the project, I could not find a fitting dataset, as they either did not include geographical features, not enough rows or did not contain continuous and categorical variables. However, since we are all still impacted by this pandemic this dataset is highly relevant. I thought it would be interesting to learn more about the effects of it and compare countries and regions.

Data Profile

Data Cleaning

The original dataset had 306,429 row and 8 columns. I have not decided yet if I want to make some of the rows into columns, such as observations dates (perhaps only having each month instead of every day for each month). Also, maybe adding and moving each province/state into each country, except for the USA to compare states (could add an extra column/row for USA states combined).

Dropping columns: So far, I have removed the column "Last update" and "Serial number" as I am not using this data.

Renaming columns: I renamed the column “ObservationDate” to “Date of observation”.

Mixed-type columns: Only the column “Province/State” was mixed-type. I did not change anything yet, because I still haven’t decided if I will be using this column or add it into the Country/Region column.

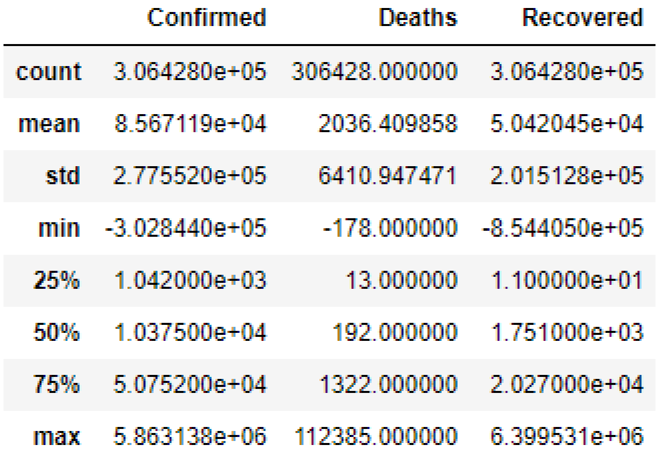
Missing values: Province/State has 78100 missing values. I understand that this is because not all Countries/Regions are divided into states. As stated above, I have not decided to change this column yet.

Duplicates: 1 duplicate was found and removed.

Basic Descriptive Statistics

The cleaned data now has 6 columns, and 306,428 rows.

I calculated descriptive statistics for the “Confirmed”, “Deaths” and “Recovered” columns in Jupyter notebook using describe:



	Confirmed	Deaths	Recovered
count	3.064280e+05	306428.000000	3.064280e+05
mean	8.567119e+04	2036.409858	5.042045e+04
std	2.775520e+05	6410.947471	2.015128e+05
min	-3.028440e+05	-178.000000	-8.544050e+05
25%	1.042000e+03	13.000000	1.100000e+01
50%	1.037500e+04	192.000000	1.751000e+03
75%	5.075200e+04	1322.000000	2.027000e+04
max	5.863138e+06	112385.000000	6.399531e+06

Limitations and Ethical Considerations

A limitation to this dataset is that we only have access to covid cases from January 2020-April 2021. It does not contain any data regarding later confirmed cases, deaths or recoveries. As this is still an ongoing pandemic we cannot compare to the latest data.

Another limitation is that we do not have any data on the newer variants of covid-19 and can therefore not compare these variants to the first one or each other. The variants are in fact not mentioned at all, so we do not know if any of these cases involves the original variant or different variants.

It would also have been interesting if vaccines were included in the dataset to see if that had an impact on the cases.

For ethical consideration I am not concerned, as this is all public information. The data does not contain any personal identifiable information.

Define questions:

1. How has covid spread throughout the world?
2. Which month(s) did covid cases peak?
3. Which countries have had the most and least cases and deaths?
4. How are different states in the US effected by Covid-19?