6.1: Sourcing Open Data

Data Set Link

1. Data Source and Collection

Found this date set on the Kaggle website. The document was produced by The University of Oxford and 'Our World in Data'. It is updated daily.

2. Data Content

The dataset uses the most recent official numbers from governments and health ministries worldwide. Population estimates for per-capita metrics are based on the United Nations World Population Prospects. Income groups are based on the World Bank classification.

3. Reason for collection

For a long time, I couldn't decide what to choose. I chose it because I wanted to see for myself what the numbers could tell me about this pandemic, because you often heard a lot of fake news, and sometimes it was hard to distinguish truth from false.

4. Data Profile

The data set contains 166326 rows and 67 columns

Variable	Time Component		Data Structured	Qualitative			Quantitative	
	Time-Invariant	Time-Variant	Structured	Nominal	Ordinal	Binary	Discrete	Continuous
iso_code	x		Yes	x				
continent	x		Yes	x				
location	x		Yes	x				
date		X	Yes					x
total_cases		X	Yes				x	
new_cases		X	Yes				x	
new_cases_smoothed		X	Yes				x	
total_deaths		X	Yes				X	
new_deaths		X	Yes				X	
new_deaths_smoothed		X	Yes				X	
total_cases_per_million		X	Yes				X	
new_cases_per_million		X	Yes				X	
new_cases_smoothed_per_million		X	Yes				X	
total_deaths_per_million		X	Yes				X	
new_deaths_per_million		X	Yes				X	

new_deaths_smoothed_per_million		X	Yes				X	
reproduction_rate		X	Yes				X	
icu_patients		X	Yes				X	
icu_patients_per_million		X	Yes				X	
hosp_patients		X	Yes				X	
hosp_patients _per-million		X	Yes				X	
weekly_icu_admissions		X	Yes				X	
weekly_icu_admissions_per_million		X	Yes				X	
weekly_icu_admissions_per_million		X	Yes				X	
weekly_hosp_admissions		X	Yes				X	
weekly_hosp_admissions_per_million		X	Yes				X	
new_tests		X	Yes				X	
total_tests		X	Yes				X	
total_tests_per_thousand		X	Yes				X	
new_tests_smoothed		X	Yes				X	
positive_rate		X	Yes				X	
tests_per_case		X	Yes				X	
tests_units		X	Yes					
total_vaccinations		X	Yes				X	
people_vaccinated		X	Yes				X	
people_fully_vaccinated		X	Yes				X	
total_boosters		X	Yes				X	
new_vaccinations		X	Yes				X	
new_vaccinations_smoothed		X	Yes				X	
total_vaccinations_per_hundred		X	Yes				X	
people_vaccinated_per_hundred		X	Yes				X	
people_fully_vaccinated_per_hundred		X	Yes				X	
total_boosters_per_hundred		X	Yes				X	
new_vaccinations_smoothed_per_million		X	Yes				X	
new_people_vaccinated_smoothed		X	Yes				X	
new_people_vaccinated_smoothed_per_hundred		X	Yes				X	
stringency_index		X	Yes					X
population		X	Yes				X	
population_density		X	Yes				X	
median_age		X	Yes					X
aged_65_older		X	Yes					X
aged_70_older		X	Yes					X
gdp_per_capita		X	Yes					X
extreme_poverty		X	Yes				X	

cardiovasc_death_rate		X	Yes				X	
diabetes_prevalence		X	Yes				X	
female_smokers		X	Yes				X	
male_smokers		X	Yes				X	
handwashing_facilities		X	Yes				X	
hospital_beds_per_thousand		X	Yes				X	
life_expectancy		X	Yes					X
human_development_index		X	Yes					X
excess_mortality_cumulative_absolute		X	Yes				X	
excess_mortality_cumulative		X	Yes				X	
excess_mortality		X	Yes				X	
excess_mortality_cumulative_per_million		X	Yes				X	

5.Data Cleaning

Columns dropped:

icu_patients

hosp_patients

hosp_patients_per_million

weekly_icu_admissions

weekly_hosp_admissions

weekly_hosp_admissions_per_million

new_tests

total_tests

total_tests_per_thousand

new_tests_smoothed_per_thousand

total_boosters_per_hundred

excess_mortality_cumulative_per_million

new_cases_smoothed_per_million

new_people_vaccinated_smoothed_per_hundred

new_deaths_smoothed_per_million

extreme_poverty

icu_patients_per_million

new_tests_smoothed

tests_units

new_vaccinations_smoothed

excess_mortality_cumulative_absolute

excess_mortality_cumulative

excess_mortality

weekly_icu_admissions_per_million

Duplicates:

```
In [13]: # checking fo duplicates
         df1[df1.duplicated()]
```

Out[13]:

Country_Code	continent	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	total_cas

no issues here

Mixed Data Types

Continent and tests_units

Dealing with missing values

Out[69]:

iso_code	0
continent	0
location	0
date	0
total_cases	3033
new_cases	3193
new_cases_smoothed	5176
total_deaths	20875
new_deaths	20839
new_deaths_smoothed	22936
total_cases_per_million	3791
new_cases_per_million	3951
total_deaths_per_million	21620
new_deaths_per_million	21584
reproduction_rate	40506
new_tests_per_thousand	99009
positive_rate	87671
tests_per_case	88242
total_vaccinations	121132
people_vaccinated	123339
people_fully_vaccinated	126085
total_boosters	148787
new_vaccinations	128879
total_vaccinations_per_hundred	121132
people_vaccinated_per_hundred	123339
people_fully_vaccinated_per_hundred	126085
new_vaccinations_smoothed_per_million	81928
new_people_vaccinated_smoothed	83238
stringency_index	36254
population	1075
population_density	18398
median_age	28495
aged_65_older	29989
aged_70_older	29234
gdp_per_capita	27822
cardiovasc_death_rate	29548
diabetes_prevalence	22377
female_smokers	60276

new_tests_per_thousand

positive_rate

tests_per_case

total_vaccinations

people_vaccinated

people_fully_vaccinated

total_boosters

new_vaccinations

total_vaccinations_per_hundred

people_fully_vaccinated_per_hundred

people_vaccinated_per_hundred

new_vaccinations_smoothed_per_million

new_people_vaccinated_smoothed

new_tests_per_thousand

Missing values in the above mentioned columns have been replaced with 0 and the rest with the mean

6.Limitation and Ethics

There are no ethical concerns here. Gaps in the columns are a bigger problem. This is closely related to data collection. In highly developed countries, the statistics are collected very carefully. On the other hand, in poor countries, the data collected is very incomplete. This is because data was collected there for a variety of reasons. In some cases, there were no tests available, or the tests were carried out in urbanized areas and in rural areas no one collected data.

7.Questions:

Where were the most infections?

How strong is the link between age and mortality?

Has the increase in the number of sick people always been the same?

Despite the gaps in the data, is there a chance to compare the results of the 5 richest countries with the 5 poorest?