CSC343 Term Project
Phase 1: Dataset and Relational Schema
By: Shuotong Li, Hannah Zhang

1. Domain

With the current outburst and control measures of the global pandemic, our group settled on a common interest in investigating relationships within the COVID-19 open data source. Thus, the domain chosen for this project is the global COVID-19 dataset that includes information such as a country's total cases, deaths, vaccinations, and potentially the country's demographic data.

2. Dataset Description

2. 1 Link and Citation to the Dataset

Hannah Ritchie, Edouard Mathieu, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Esteban Ortiz-Ospina, Joe Hasell, Bobbie Macdonald, Diana Beltekian and Max Roser (2020) - "Coronavirus Pandemic (COVID-19)". *Published online at OurWorldInData.org.* Retrieved from: 'https://ourworldindata.org/coronavirus' [Online Resource]

2.2 Relevant Informations

The dataset provided by Our World in Data (OWID) (referenced in section 2.1) has a lot of relevant information to our project. OWID's dataset on the coronavirus pandemic is updated daily. Within the data, it contains information such as Cases, Deaths, Vaccinations, age distribution, country's GDP, etc. Given that this dataset has synthesized a lot of information, it is enough to work with one dataset for now.

2.3 Learning you have to do to interpret the data

Some of the data provided include the country's GDP, population information, health and diseases record. In order to get better insight into any potential relationship, a basic understanding of the field mentioned above is necessary for this project. In addition, we also have to learn in-depth relational algebra and database management software to interpret the data.

2.4 Cleaning up needed to use the data

The selected dataset was set up to record various atrributes for a country related to the pandemic. However, some countries are not able to provide certain information. Such as in the dataset, we noticed that the "total_vaccinated" column is empty for Afghanistan. In addition, the starting date of the vaccination process is different between countries, and the reporting interval of data varies from daily to annually. All these differences require our team to clean up the data before using them. Some potential methods for cleaning up are altering the data to showcase weekly updates, omitting some countries from our analysis, or set up different comparison groups so that in each group the data can provide informative insights into the relationship of interest.

Another clean-up needed is to exclude some redundant information. For example, in the OWID dataset, large countries are reported separately. However, the continental information on the covid

cases is registered again, with many other empty columns. Therefore, it is reasonable to exclude the continental information from our analysis.

3. Three investigative questions

3.1 Question 1

Does the rising number of the vaccinated population affect the number of COVID-19 cases and the case fatality rate? What about vaccination's influence on the outburst of different coronavirus variants? (The case fatality rate refers to the number of diagnosed cases over the number of deaths)

3.2 Question 2

What is the relationship between a country's GDP, HDI level, coronavirus death counts and hospitalization rate? More developed countries usually have better healthcare systems (hospital beds, life expectancy) and easier access to personal protection equipment (PPE).

3.3 Question 3

How do smoking, diabetes and cardiovascular disease influence COVID-19 infection severity? It is known that smokers are more prone to COVID-19, and cigarettes also cause cardiovascular disease, what are the relationships between these data?

4. Schema

4.1 Relational schema

Continent(<u>continentName</u>, countryName)

Country(<u>iso_code, countryName</u>, population, gdp_per_capita)

CoronaData(<u>iso_code, date</u>, total_cases, total_deaths, reproduction_rate, total_tests, total_vaccination, peope_vaccinated, people_fully_vaccinated, stringency_index)

MedicalInfo(<u>iso_code</u>, cardiovasc_death_rate, diabetes_prevalence, hospital_beds_per_thousand, life_expectancy, female_smokers, male_smokers, handwashing_facilities)

DemographicInfo(<u>iso_code</u>, population_density, median_age, aged_65_old, aged_70_old, mortality_rate, human_development_index)

Continent[countryName] ⊆Country[countryName]
CoronaData[iso_code] ⊆Country[iso_code]
MedicalInfo[iso_code] ⊆CoronaData[iso_code]
DemographicInfo[iso_code] ⊆CoronaData[iso_code]

4.2 Data dictionary

The Attributes are listed in order by their column locations in the original dataset.

Attribute	Description	Type	Required
iso_code	ISO code for a given location	3 char	Yes
continent	Continent where the location is located	String	Yes
location	Location name, country or region	String	Yes
date	Date of the data	String	Yes
total_cases	Total number of COVID-19 infections reported	int	Yes
new_cases	Daily reported new COVID-19 infection	int	No
new_cases_smoothed	7-day rolling average of new_cases	float	No
total_deaths	Total number of death due to COVID-19	int	Yes
new_deaths	Daily new COVID-19 related death	int	No
new_deaths_smoothed	7-day rolling average of new_deaths	float	No
total_cases_per_million	Total_cases per million people	float	No
new_cases_per_million	new_cases per million people	float	No
new_cases_smoothed_per_million	new_cases_smoothed per million people	float	No
total_deaths_per_million	total_deaths per million people	float	No
new_deaths_per_million	new_deaths per million people	float	No
new_deaths_smoothed_per_million	new_deaths_smoothed per million population	float	No
reproduction_rate	R_0 of COVID-19	float	Yes
icu_patients	Current patients in ICU	int	No
icu_patients_per_million	icu_patients per million people	float	No
hosp_patients	Current patients in hospital	int	No
hosp_patients_per_million	hosp_patients per million people	float	No
weekly_icu_admissions	ICU admissions per week	int	No
weekly_icu_admissions_per_million	weekly_icu_admissions per million	float	No

	people		
weekly_hosp_admissions	Hospital admission per week	int	No
weekly_hosp_admissions per_million	weekly_hosp_admissions per million people	float	No
new_tests	Daily new COVID-19 test performed	int	No
total_tests	Total COVID-10 test performed	int	Yes
total_tests_per_thousand	per thousand population	float	No
new_tests_per_thousand	per thousand population	float	No
new_tests_smoothed	7-day rolling average of	float	No
new_tests_smoothed_per_thousand	per thousand people	float	No
positive_rate	percentage of positive test over all tests	float	No
tests_per_case	test over case number	float	No
tests_units	Description of the test	string	No
total_vaccinations	Total vaccine shot administered	int	Yes
people_vaccinated	Number of people who have at least some vaccination, including partially and fully vaccinated.	int	Yes
people_fully_vaccinated	Number of people who are fully vaccinated.	int	Yes
total_boosters	Total number of booster vaccine shots	int	No
new_vaccinations	New vaccine shot administered per day	int	No
new_vaccinations_smoothed	7-day rolling average of new_vaccinations	float	No
total_vaccinations_per_hundred	total_vaccinations per hundred people	float	No
people_vaccinated_per_hundred	people_vaccinated per hundred people	float	No
people_fully_vaccinated_per_hundred	people_fully_vaccinated per hundred people	float	No
total_boosters_per_hundred	total_boosters per hundred population	float	No
new_vaccinations_smoothed_per_million	7 day rolling average of new vaccinations per million people	float	No
stringency_index	stringency index is an index for how	float	Yes

	serious the COVID-19 pandemic is of a location		
population	The population of a location	int	Yes
population_density	Population density of a location in people per km^2	float	Yes
median_age	the median age over the entire population	float	Yes
aged_65_older	percentage of people who are 65 and older	float	Yes
aged_70_older	percentage of people who are 70 and older	float	Yes
gdp_per_capita	gross domestic production per capita	float	Yes
extreme_poverty	percentage of the population in extreme poverty	float	No
cardiovasc_death_rate	The cardiovasc death rate in per cent mille (100,000s)	float	Yes
diabetes_prevalence	percentage of diabetes patient over population	float	Yes
female_smokers	percentage of females that smokes	float	Yes
male_smokers	percentage of males that smokes	float	Yes
handwashing_facilities	percentage of people who have access to handwashing facilities	float	Yes
hospital_beds_per_thousand	number of hospital beds per thousand people	float	Yes
life_expectancy	The life expectancy of a location	float	Yes
human_development_index	HDI, usually assess how developed a location is	float	Yes
excess_mortality_cumulative_absolute	Total amount of calculated cumulative excess mortality	float	No
excess_mortality_cumulative	Cumulative excess mortality percentage	float	No
excess_mortality	excess mortality percentage in a given year	float	No
excess_mortality_cumulative_per_million	amount of calculated cumulative excess mortality per million people	float	No

4.3 Justification of design

The dataset is divided to minimize redundancy and provide more efficient access by grouping related information together. OWID COVID-19 dataset is undivided and contains excessive redundant data that are not updated daily.

The core Schema of the OWID dataset is CoronaData. These data are updated on a daily basis. The attributes in CoronaData are selected to optimize for minimum resource consumption: Many attributes such as per day data, data per million etc., can be derived using only a few key attributes (total data, population data) with only a performance penalty of O(1). On the other hand, if per day data is used to determine the total data, it has a complexity of O(n), where n is the number of days passed.

Static data, such as population, median age, mortality rate, are usually updated yearly or even longer due to the difficulty or non-essential to update them sooner. We have not yet thoroughly examined the data for every individual country to confirm that MedicalInfo and DemographicInfo are invariable, as we haven't studied the means to do so yet. However, primarily examining the datasets confirms these data are static for Canada, and given the fact the dataset only started recording on Jan 22, 2020, at this point, we would assume these data are also invariable for all other countries,

The static data in the dataset are useful in researching the correlation between the COVID-19 pandemic and a country's ability to resist or reduce the stringency of Covid-19. The number of attributes is high and involves two different areas of study, thus the static data are further divided into MedicalInfo and Demographic info.