Nanyang Technological University
Nanyang Business School

BC2402 – Designing and Developing Databases
Semester 1, 2021

Group Project

COVID'19

The New Normal – Vaccinations and Re-opening

1. INTRODUCTION

Case Background

"23.8% of the world population has received at least one dose of a COVID-19 vaccine.

3.16 billion doses have been administered globally, and 38.69 million are now administered each day.

Only 1% of people in low-income countries have received at least one dose."[1]

In a straw poll he conducted via CNA Insider's Instagram account, 23 per cent of the 300 respondents said no to vaccinating their children.

Shaheera Effendi, for example, is "still thinking" about whether to bring her elder daughter, 13, for vaccination.

"She tends to get sick very easily, so I'm quite worried … she'd have worse side effects," said the mother of two, who cited the rare cases of heart inflammation in teen vaccine recipients being investigated in some countries.

She also has friends who have questioned whether children need a vaccine to fight the novel coronavirus.

Associate Professor Thoon, who has done research on adverse effects following immunisation, said Singapore has also been "carefully evaluating" the presence of heart inflammation in young adults and adolescents — and "there've been some signals, but they're also very, very small".

From data overseas, the reported rate so far is 1.6 per 100,000 adolescents vaccinated, he cited. "Most of these reports have mentioned that these individuals have recovered without any long-term side effects, and they're still on follow-up.[2]

[1] https://ourworldindata.org/covid-vaccinations
[2] https://www.channelnewsasia.com/news/cnainsider/side-effect-infection-risk-covid-19-school-life-vaccine-children-15135060

2. Dataset

A. What do you have?

In this project, we utilized two sources of data:

1. COVID-19 World Vaccination Progress on Kaggle (used in the individual assignment)
2. Coronavirus (COVID-19) Vaccinations on Our World in Data

1. COVID-19 World Vaccination Progress on Kaggle

We are utilizing the above 2 datasets (i.e., country_vaccinations and country_vaccinations_by_manufacturer), which are used in the individual assignment, in the group project. For details on the datasets, please refer to the individual assignment document.

2. Our World in Data

Our complete COVID-19 dataset is a collection of the COVID-19 data maintained by Our World in Data. We will update it daily throughout the duration of the COVID-19 pandemic.

The dataset is built based on multiple sources, as follows:

- Confirmed cases and deaths: our data comes from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). We discuss how and when JHU collects and publishes this data here. The cases & deaths dataset is updated daily. Note: the number of cases or deaths reported by any institution—including JHU, the WHO, the ECDC and others—on a given day does not necessarily represent the actual number on that date. This is because of the long reporting chain that exists between a new case/death and its inclusion in statistics. This also means that negative values in cases and deaths can sometimes appear when a country corrects historical data, because it had previously overestimated the number of cases/deaths. Alternatively, large changes can sometimes (although rarely) be made to a country's entire time series if JHU decides (and has access to the necessary data) to correct values retrospectively.
- Hospitalizations and intensive care unit (ICU) admissions: our data comes from the European Centre for Disease Prevention and Control (ECDC) for a select number of European countries; the government of the United Kingdom; the Department of Health & Human Services for the United States; the COVID-19 Tracker for Canada. Unfortunately, we are unable to provide data on hospitalizations for other countries: there is currently no global, aggregated database on COVID-19 hospitalization, and our team at Our World in Data does not have the capacity to build such a dataset.
- Testing for COVID-19: this data is collected by the Our World in Data team from official reports; you can find further details in our post on COVID-19 testing, including our checklist of questions to understand testing data, information on geographical and temporal coverage, and detailed country-by-country source information. The testing dataset is updated around twice a week.

- Vaccinations against COVID-19: this data is collected by the Our World in Data team from official reports.

The JSON version is split by country ISO code, with static variables and an array of daily records.

The variables represent all of our main data related to confirmed cases, deaths, hospitalizations, and testing, as well as other variables of potential interest.

As of 3 June 2021, the columns are:
iso_code, continent, location, date, total_cases, new_cases, new_cases_smoothed, total_deaths, new_deaths, new_deaths_smoothed, total_cases_per_million, new_cases_per_million, new_cases_smoothed_per_million, total_deaths_per_million, new_deaths_per_million, new_deaths_smoothed_per_million, reproduction_rate, icu_patients, icu_patients_per_million, hosp_patients, hosp_patients_per_million, weekly_icu_admissions, weekly_icu_admissions_per_million, weekly_hosp_admissions, weekly_hosp_admissions_per_million, total_tests, new_tests, total_tests_per_thousand, new_tests_per_thousand, new_tests_smoothed, new_tests_smoothed_per_thousand, positive_rate, tests_per_case, tests_units, total_vaccinations, people_vaccinated, people_fully_vaccinated, new_vaccinations, new_vaccinations_smoothed, total_vaccinations_per_hundred, people_vaccinated_per_hundred, people_fully_vaccinated_per_hundred, new_vaccinations_smoothed_per_million, stringency_index, population, population_density, median_age, aged_65_older, aged_70_older, gdp_per_capita, extreme_poverty, cardiovasc_death_rate, diabetes_prevalence, female_smokers, male_smokers, handwashing_facilities, hospital_beds_per_thousand, life_expectancy, human_development_index, excess_mortality

2. Project Deliverables

The due date for the group project is 12 November 2021, 23:59 (23:59 hrs NTULearn server time)

There are two key deliverables (and one set of optional deliverables), namely

A. 1 x project report
B. 1 x presentation
C. [optional] database implementations (e.g., relational database and nonrelational database)

A. DATABASE IMPLEMENTATIONS

To allow you to focus on query development, you are provided with both mySQL and MongoDB implementations of the datasets. SQL and noSQL queries can be developed entirely based on the provided database implementations.

However, given the structural complexities (e.g., an array of documents) of the databases, you may find that the queries can be overly complex. In such a case, you can rework the database implementations (e.g., changing the JSON structures, joining tables) as you deem necessary. Do take note of the following limitations:

1. The mySQL database must have at least 2 tables. There is no maximum limit in the number of tables in the mySQL database.
2. The MongoDB database must not have more than 4 collections. There is no minimum limit in the number of collections in the MongoDB database.

A.1 mySQL database implementation

The specific deliverables are:

- Instructions on deploying the mySQL database (i.e., steps to import the .sql package, which contains the schema and records)
- SQL statements (with expected outputs) for queries (i.e., in a sql or text file) on Appendix B.

A.2 MongoDB database implementation

The specific deliverables are:

- Instructions on deploying the MongoDB database
- noSQL statements (with expected outputs) for queries (i.e., in a document file) as depicted in Appendix C.

B. PROJECT REPORT

The report should contain the following:

1. A cover page that includes a title, and names as well as matric numbers of each team member
2. (Optional) Inconsistence between relational data model and nonrelational data model (if changes in modeling is required for the nonrelational model)
3. Recommendation to WHO (i.e., comparisons between relational database and nonrelational database implementations, recommendation and justification)

C. PRESENTATION

Your team is expected to deliver a video-recorded presentation (which must be made available via YouTube), in which the team is expected to:

1. Present the relational database design (i.e., a brief discussion of the team's ERD)
2. Demonstrate the relational database implementation (i.e., discuss the team's approach in resolving a SQL query in Appendix D)
3. (optional) Present the nonrelational database design (i.e., a brief discussion on the differences in design between relational and nonrelational implementation)
4. Demonstrate the nonrelational database implementation (i.e., discussion on a noSQL in Appendix D)

The entire presentation **MUST** be within 20 minutes (video duration beyond 20 minutes will be ignored). Each member is expected to contribute equally in the presentation.

3. SUBMISSION

A submission folder will be made available on NTULearn. Please zip the files and make a single file submission. One member will complete the submission on behalf of the group.

The following files must be submitted to complete this group project:

A. [optional] Database implementations (which include database dump, script files, etc.)
B. Project report (in pdf format)
C. YouTube URL (in a text file, please ensure the URL is viewable)
D. Completed Task Allocation form

The submission must be made by 12 November 2021, 23:59. Do note that video processing and YouTube uploading can be computationally intensive and bandwidth demanding. Please ensure ample time for processing and uploading the presentation video.

Appendix A

column	description
iso_code	ISO 3166-1 alpha-3 – three-letter country codes
continent	Continent of the geographical location
location	Geographical location
date	Date of observation
total_cases	Total confirmed cases of COVID-19
new_cases	New confirmed cases of COVID-19
new_cases_smoothed	New confirmed cases of COVID-19 (7-day smoothed)
total_deaths	Total deaths attributed to COVID-19
new_deaths	New deaths attributed to COVID-19
new_deaths_smoothed	New deaths attributed to COVID-19 (7-day smoothed)
total_cases_per_million	Total confirmed cases of COVID-19 per 1,000,000 people
new_cases_per_million	New confirmed cases of COVID-19 per 1,000,000 people
new_cases_smoothed_per_million	New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people
total_deaths_per_million	Total deaths attributed to COVID-19 per 1,000,000 people
new_deaths_per_million	New deaths attributed to COVID-19 per 1,000,000 people
new_deaths_smoothed_per_million	New deaths attributed to COVID-19 (7-day smoothed) per 1,000,000 people
reproduction_rate	Real-time estimate of the effective reproduction rate (R) of COVID-19. See https://github.com/crondonm/TrackingR/tree/main/Estimates-Database
icu_patients	Number of COVID-19 patients in intensive care units (ICUs) on a given day
icu_patients_per_million	Number of COVID-19 patients in intensive care units (ICUs) on a given day per 1,000,000 people
hosp_patients	Number of COVID-19 patients in hospital on a given day
hosp_patients_per_million	Number of COVID-19 patients in hospital on a given day per 1,000,000 people
weekly_icu_admissions	Number of COVID-19 patients newly admitted to intensive care units (ICUs) in a given week
weekly_icu_admissions_per_million	Number of COVID-19 patients newly admitted to intensive care units (ICUs) in a given week per 1,000,000 people
weekly_hosp_admissions	Number of COVID-19 patients newly admitted to hospitals in a given week
weekly_hosp_admissions_per_million	Number of COVID-19 patients newly admitted to hospitals in a given week per 1,000,000 people

total_tests	Total tests for COVID-19
new_tests	New tests for COVID-19 (only calculated for consecutive days)
total_tests_per_thousand	Total tests for COVID-19 per 1,000 people
new_tests_per_thousand	New tests for COVID-19 per 1,000 people
new_tests_smoothed	New tests for COVID-19 (7-day smoothed). For countries that don't report testing data on a daily basis, we assume that testing changed equally on a daily basis over any periods in which no data was reported. This produces a complete series of daily figures, which is then averaged over a rolling 7-day window
new_tests_smoothed_per_thousand	New tests for COVID-19 (7-day smoothed) per 1,000 people
positive_rate	The share of COVID-19 tests that are positive, given as a rolling 7-day average (this is the inverse of tests_per_case)
tests_per_case	Tests conducted per new confirmed case of COVID-19, given as a rolling 7-day average (this is the inverse of positive_rate)
tests_units	Units used by the location to report its testing data
total_vaccinations	Total number of COVID-19 vaccination doses administered
people_vaccinated	Total number of people who received at least one vaccine dose
people_fully_vaccinated	Total number of people who received all doses prescribed by the vaccination protocol
new_vaccinations	New COVID-19 vaccination doses administered (only calculated for consecutive days)
new_vaccinations_smoothed	New COVID-19 vaccination doses administered (7-day smoothed). For countries that don't report vaccination data on a daily basis, we assume that vaccination changed equally on a daily basis over any periods in which no data was reported. This produces a complete series of daily figures, which is then averaged over a rolling 7-day window
total_vaccinations_per_hundred	Total number of COVID-19 vaccination doses administered per 100 people in the total population
people_vaccinated_per_hundred	Total number of people who received at least one vaccine dose per 100 people in the total population
people_fully_vaccinated_per_hundred	Total number of people who received all doses prescribed by the vaccination protocol per 100 people in the total population
new_vaccinations_smoothed_per_million	New COVID-19 vaccination doses administered (7-day smoothed) per 1,000,000 people in the total population
stringency_index	Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response)
population	Population in 2020
population_density	Number of people divided by land area, measured in square kilometers, most recent year available

median_age	Median age of the population, UN projection for 2020
aged_65_older	Share of the population that is 65 years and older, most recent year available
aged_70_older	Share of the population that is 70 years and older in 2015
gdp_per_capita	Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available
extreme_poverty	Share of the population living in extreme poverty, most recent year available since 2010
cardiovasc_death_rate	Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people)
diabetes_prevalence	Diabetes prevalence (% of population aged 20 to 79) in 2017
female_smokers	Share of women who smoke, most recent year available
male_smokers	Share of men who smoke, most recent year available
handwashing_facilities	Share of the population with basic handwashing facilities on premises, most recent year available
hospital_beds_per_thousand	Hospital beds per 1,000 people, most recent year available since 2010
life_expectancy	Life expectancy at birth in 2019
human_development_index	A composite index measuring average achievement in three basic dimensions of human development—a long and healthy life, knowledge and a decent standard of living. Values for 2019, imported from http://hdr.undp.org/en/indicators/137506
excess_mortality	Excess m

Source: https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-codebook.csv

Appendix B – 10 queries (SQL)

1. What is the total population in Asia?

2. What is the total population among the ten ASEAN countries?

3. Generate a list of unique data sources (source_name).

4. Specific to Singapore, display the daily total_vaccinations starting (inclusive) March-1 2021 through (inclusive) May-31 2021.

5. When is the first batch of vaccinations recorded in Singapore?

6. Based on the date identified in (5), specific to Singapore, compute the total number of new cases thereafter.
   For instance, if the date identified in (5) is Jan-1 2021, the total number of new cases will be the sum of new cases starting from (inclusive) Jan-1 to the last date in the dataset.

7. Compute the total number of new cases in Singapore before the date identified in (5).
   For instance, if the date identified in (5) is Jan-1 2021 and the first date recorded (in Singapore) in the dataset is Feb-1 2020, the total number of new cases will be the sum of new cases starting from (inclusive) Feb-1 2020 through (inclusive) Dec-31 2020.

8. Herd immunity estimation. On a daily basis, specific to Germany, calculate the percentage of new cases (i.e., percentage of new cases = new cases / populations) and total vaccinations on each available vaccine in relation to its population.

9. Vaccination Drivers. Specific to Germany, based on each daily new case, display the total vaccinations of each available vaccines after 20 days, 30 days, and 40 days.

10. Vaccination Effects. Specific to Germany, on a daily basis, based on the total number of accumulated vaccinations (sum of total_vaccinations of each vaccine in a day), generate the daily new cases after 21 days, 60 days, and 120 days.

Note: Please be aware that a value (e.g., daily vaccinations) can be retrieved from multiple tables. The key purpose of this intended data diversity (and perhaps, inconsistency) is to illustrate real-life challenges in different data perspective, quality, and consistency. Since you are not provided with a (set of) specific tables to construct the above queries, you are advised to make your assessment in retrieving data from the most relevant tables.

Appendix C – 20 queries (noSQL)

1. Display a list of total vaccinations per day in Singapore.
   [source table: country_vaccinations]

2. Display the sum of daily vaccinations among ASEAN countries.
   [source table: country_vaccinations]

3. Identify the maximum daily vaccinations per million of each country. Sort the list based on daily vaccinations per million in a descending order.
   [source table: country_vaccinations]

4. Which is the most administrated vaccine? Display a list of total administration (i.e., sum of total vaccinations) per vaccine.
   [source table: country_vaccinations_by_manufacturer]

5. Italy has commenced administrating various vaccines to its populations as a vaccine becomes available. Identify the first dates of each vaccine being administrated, then compute the difference in days between the earliest date and the 4th date.
   [source table: country_vaccinations_by_manufacturer]

6. What is the country with the most types of administrated vaccine?
   [source table: country_vaccinations_by_manufacturer]

7. What are the countries that have fully vaccinated more than 60% of its people? For each country, display the vaccines administrated.
   [source table: country_vaccinations]

8. Monthly vaccination insight – display the monthly total vaccination amount of each vaccine per month in the United States.
   [source table: country_vaccinations_by_manufacturer]

9. Days to 50 percent. Compute the number of days (i.e., using the first available date on records of a country) that each country takes to go above the 50% threshold of vaccination administration (i.e., total_vaccinations_per_hundred > 50)
   [source table: country_vaccinations]

10. Compute the global total of vaccinations per vaccine.
    [source table: country_vaccinations_by_manufacturer]

[10 additional questions are presented on the next page]

Note: Please be aware that a value (e.g., daily vaccinations) can be retrieved from multiple collections. The key purpose of this intended data diversity (and perhaps, inconsistency) is to illustrate real-life challenges in different data perspective, quality, and consistency. Since you are not provided with a (set of) specific collections to construct the above queries, you are advised to make your assessment in retrieving data from the most relevant collections.

11. What is the total population in Asia?

12. What is the total population among the ten ASEAN countries?

13. Generate a list of unique data sources (source_name).

14. Specific to Singapore, display the daily total_vaccinations starting (inclusive) March-1 2021 through (inclusive) May-31 2021.

15. When is the first batch of vaccinations recorded in Singapore?

16. Based on the date identified in (5), specific to Singapore, compute the total number of new cases thereafter.
    For instance, if the date identified in (5) is Jan-1 2021, the total number of new cases will be the sum of new cases starting from (inclusive) Jan-1 to the last date in the dataset.

17. Compute the total number of new cases in Singapore before the date identified in (5).
    For instance, if the date identified in (5) is Jan-1 2021 and the first date recorded (in Singapore) in the dataset is Feb-1 2020, the total number of new cases will be the sum of new cases starting from (inclusive) Feb-1 2020 through (inclusive) Dec-31 2020.

18. Herd immunity estimation. On a daily basis, specific to Germany, calculate the percentage of new cases and total vaccinations on each available vaccine in relation to its population.

19. Vaccination Drivers. Specific to Germany, based on each daily new case, display the total vaccinations of each available vaccines after 20 days, 30 days, and 40 days.

20. Vaccination Effects. Specific to Germany, on a daily basis, based on the total number of accumulated vaccinations (sum of total_vaccinations of each vaccine in a day), generate the daily new cases after 21 days, 60 days, and 120 days.