JUNE 15, 2022

CM3010 Databases and Advanced Data Techniques

MID TERM REPORT

CHIA YI QUAN
Student Number: 210220223

Table of Contents

Stage 1. Find and critique a dataset

1.1 Choosing A Source Of Data

For this coursework, I have chosen an open data from Our World In Data and the dataset I used is the coronavirus dataset. This is the link where the csv file is downloaded from. The reason why I picked this dataset is because coronavirus have been with us for the past 2+ years and there have been many variants of the virus and what made this dataset interesting is that we can examine the number of cases, death rate, vaccination and other data by country. Because the data is in csv, we could use the database to help us organize the data better which will allow us to visualize the data easily. The main objective of this database is to allow us to better visualize the coronavirus cases, the relationship of the death rate and vaccination around the world by country, continent, and as a world.

1.2 Assessing The Dataset

Quality

Metrics	Source	Updated	Countries
Vaccinations	Official data collated by the Our World in Data team	Daily	218
Tests & positivity	Official data collated by the Our World in Data team	Weekly	193
Hospital & ICU	Official data collated by the Our World in Data team	Daily	47
Confirmed cases	JHU CSSE COVID-19 Data	Daily	217
Confirmed deaths	JHU CSSE COVID-19 Data	Daily	217
Reproduction rate	Arroyo-Marioli F, Bullano F, Kucinskas S, Rondón-Moreno C	Daily	192
Policy responses	Oxford COVID-19 Government Response Tracker	Daily	187
Other variables of interest	International organizations (UN, World Bank, OECD, IHME...)	Fixed	241

The source can be found here and the full list here. The quality of the data is quite reliable as the data collected came from different places CCSE from Johns Hopkins University, United Nations, World Bank, Global Burden of Disease, Blavatnik School of Government, etc. The organization(Our World In Data) researches poverty, disease, hunger, climate change, war, existential risks, and inequality and data that are released by them have been used by journals as a source like The Washington Post, The New York Times and The Economist(cited from wiki).

Detail

The data source has a lot of details like country population, name, the covid cases and deaths, a country excess mortality, hospital and icu, reproduction rate, policy responses for each country, covid tests, vaccination data and other miscellaneous data of the country. Because there are a lot of data provided, most of them are

helpful but some aren't that helpful. However, because this data consisted of countries around the world, there are some country data that are not up to date, so it is a bit difficult to do 100% analysis of all country.

Documentation

The clarity of the data is quite straightforward. The documentation can be found here and it have stated on every field what it description. The documentation is relatively easy to find because the csv file can only be downloaded from their github repository and on that page, they have displayed the README.md with all fields and it's description.

Interrelation

If we just look at covid cases and deaths, the data will not mean anything but because this dataset provided a lot of other data, linking with other columns will create a more useful analysis. Stringency index can be linked with the covid cases and deaths which will let us analyse whether those policy in place decreases daily new case. Vaccination data and covid data would show us insight whether this vaccine have any effect of slowing/improve the covid situation. Using hospitalize data(covid patient hospitalized and covid icu patient) and covid data will show us is all the covid cases is very severe that people who infected with coronavirus need to be hospitalized or not.

Use

This data can be used for analysis in whether certain policy decision or certain factor can slow down the growth rate of the covid cases as new vaccine is being developed to counter this virus. Question like if the country has high chronic diseases(like diabetes) will it contribute to a higher death rate. However, the data does not record all the chronic disease for each days/month instead it records only the latest value. Other than that there are a lot of data that are missing due to some country not reporting or is not up to date.

Discoverability

The data of my chosen domain is relatively easy to find. There will other website they provide the csv data like gov.sg provide covid cases in Singapore, kaggle provide the covid cases csv however, the author stopped updating the csv. There are other alternative that I have not explored but the current source I used have all the information and additional dataset which can be made a good use with it.

Licensing

So as mentioned in the first section that the data was found in Our World In Data, under the licensing section, the data is open access license.

Reuse our work freely

All visualizations, data, and code produced by Our World in Data are completely open access under the Creative Commons BY license. You have the permission to use, distribute, and reproduce these in any medium, provided the source and authors are credited.

The data produced by third parties and made available by Our World in Data is subject to the license terms from the original third-party authors. We will always indicate the original source of the data in our documentation, so you should always check the license of any such third-party data before use and redistribution.

All of our charts can be embedded in any site.

Although this dataset is open access license, there are data that are from third-party authors who might have different licensing. However, even though the data is open access, we should always include the citations of the source when we use the data.

1.3 Explanation Of Interest

As mentioned from the first section, what made this dataset interesting is because we are still experiencing the pandemic when I was doing this coursework. So, the few questions I am interested is

1. Does more people get vaccinated reduces the covid cases or deaths?
2. Which top 10 country have the lowest total covid cases against the total population of the country?
3. As the cases increase, people who are vaccinated increases, do more people get hospitalized?

Stage 2. Model your data

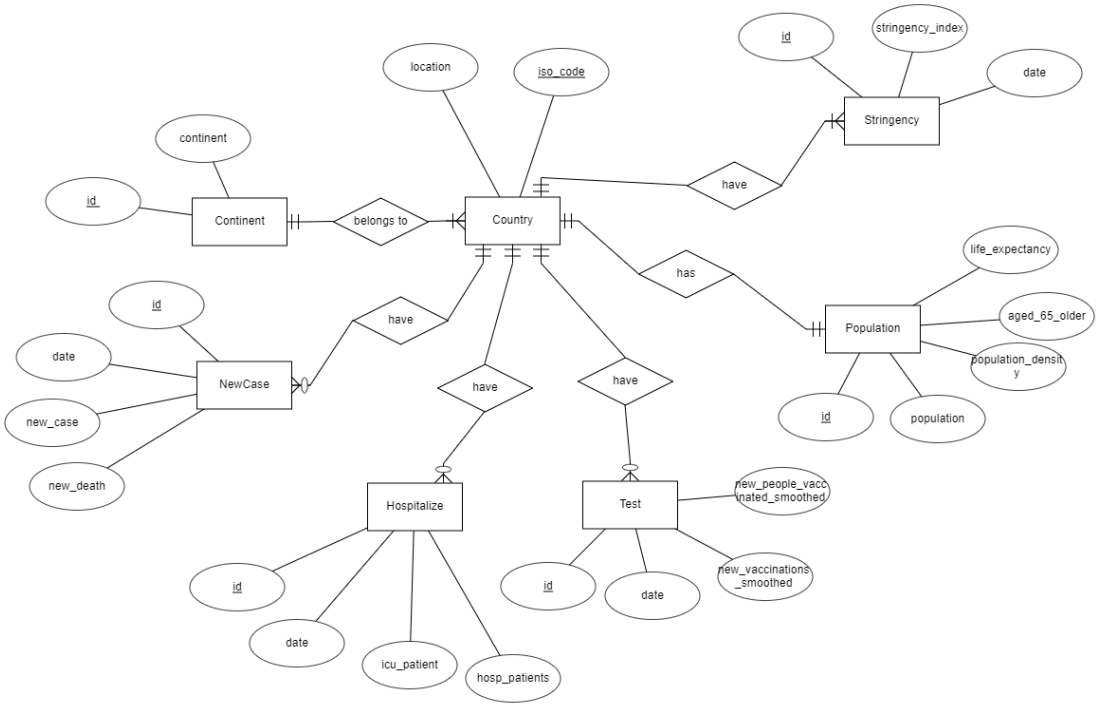


Fig 1.1 ER Model

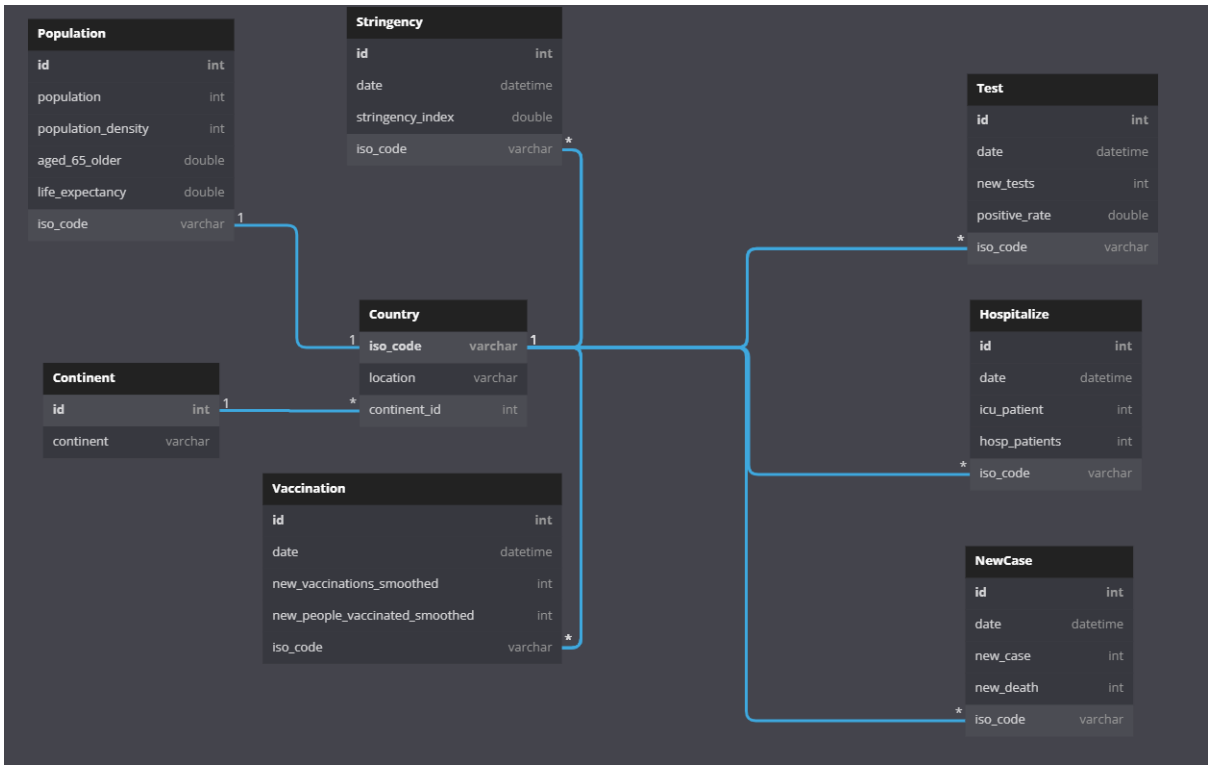


Fig 1.2 Database Table

Fig 1.1 is the er model, Fig 1.2 is the database table and all the table below are the description of all the columns that are in the csv. For this coursework I did not use all

the columns in the csv provided. Most of the table uses iso_code as the foreign key which is the primary key in Country table.

Confirmed cases

Variable	Description
total_cases	Total confirmed cases of COVID-19. Counts can include probable cases, where reported.
new_cases	New confirmed cases of COVID-19. Counts can include probable cases, where reported. In rare cases where our source reports a negative daily change due to a data correction, we set this metric to NA.
new_cases_smoothed	New confirmed cases of COVID-19 (7-day smoothed). Counts can include probable cases, where reported.
total_cases_per_million	Total confirmed cases of COVID-19 per 1,000,000 people. Counts can include probable cases, where reported.
new_cases_per_million	New confirmed cases of COVID-19 per 1,000,000 people. Counts can include probable cases, where reported.
new_cases_smoothed_per_million	New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people. Counts can include probable cases, where reported.

From the above table, total_cases, new_cases_smoothed, total_cases_per_million, new_cases_per_million, new_cases_smoothed_per_million I did not use is because storing new_cases in the database would allow us to get the total number of cases so all the total_cases column can be left out. new_cases_smoothed is around the same as new_cases which makes all the other column irrelevant.

Confirmed deaths

Variable	Description
total_deaths	Total deaths attributed to COVID-19. Counts can include probable deaths, where reported.
new_deaths	New deaths attributed to COVID-19. Counts can include probable deaths, where reported. In rare cases where our source reports a negative daily change due to a data correction, we set this metric to NA.
new_deaths_smoothed	New deaths attributed to COVID-19 (7-day smoothed). Counts can include probable deaths, where reported.
total_deaths_per_million	Total deaths attributed to COVID-19 per 1,000,000 people. Counts can include probable deaths, where reported.
new_deaths_per_million	New deaths attributed to COVID-19 per 1,000,000 people. Counts can include probable deaths, where reported.
new_deaths_smoothed_per_million	New deaths attributed to COVID-19 (7-day smoothed) per 1,000,000 people. Counts can include probable deaths, where reported.

From the above table, total_deaths, new_deaths_smoothed, total_deaths_per_million, new_deaths_per_million, new_deaths_smoothed_per_million I did not use because storing new_deaths would allow us to get the total number of deaths so all the total columns can be left out. new_deaths_smoothed is around the same as new_deaths which make other columns irrelevant.

Excess mortality

Variable	Description
excess_mortality	Percentage difference between the reported number of weekly or monthly deaths in 2020–2021 and the projected number of deaths for the same period based on previous years. For more information, see https://github.com/owid/covid-19-data/tree/master/public/data/excess_mortality
excess_mortality_cumulative	Percentage difference between the cumulative number of deaths since 1 January 2020 and the cumulative projected deaths for the same period based on previous years. For more information, see https://github.com/owid/covid-19-data/tree/master/public/data/excess_mortality
excess_mortality_cumulative_absolute	Cumulative difference between the reported number of deaths since 1 January 2020 and the projected number of deaths for the same period based on previous years. For more information, see https://github.com/owid/covid-19-data/tree/master/public/data/excess_mortality
excess_mortality_cumulative_per_million	Cumulative difference between the reported number of deaths since 1 January 2020 and the projected number of deaths for the same period based on previous years, per million people. For more information, see https://github.com/owid/covid-19-data/tree/master/public/data/excess_mortality

From the above table, all the column I did not use as I didn't find it relevant with the analysis I trying to do.

Hospital & ICU

Variable	Description
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 (reporting date and the preceeding 6 days)
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 (reporting date and the preceeding 6 days)
weekly_hosp_admissions	Number of COVID-19 patients newly admitted to hospitals in a given week (reporting date and the preceeding 6 days)
weekly_hosp_admissions_per_million	Number of COVID-19 patients newly admitted to hospitals in a given week per 1,000,000 people (reporting date and the preceeding 6 days)

From the above table, only icu_patients and hosp_patients are used because weekly_icu_admissions and weekly_hosp_admissions can be calculated with icu_patients and hosp_patients. For the other columns that are irrelevant as they mean the same thing as the used columns.

Policy responses

Variable	Description
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)

From the above table, stringency_index data is used.

Reproduction rate

Variable	Description
`reproduction_rate`	Real-time estimate of the effective reproduction rate (R) of COVID-19. See https://github.com/crondonm/TrackingR/tree/main/Estimates-Database

For the above table, I did not use the reproduction_rate because it is irrelevant to my analysis which I was doing.

Tests & positivity

As of 23 June 2022, we will no longer add new data points to our COVID-19 testing dataset. You can read more at #2667.

Variable	Description
`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

From the above table, total_tests data are not used as new_tests are able to calculate it. total_tests_per_thousand, new_tests_per_thousand, new_tests_smoothed, new_tests_smoothed_per thousand all became relevant when new_tests is used. tests_unit are irrelevant for this case and since positive_rate are used tests_per_case has become irrelevant.

Vaccinations

Variable	Description
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 initial vaccination protocol
total_boosters	Total number of COVID-19 vaccination booster doses administered (doses administered beyond the number 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 initial vaccination protocol per 100 people in the total population
total_boosters_per_hundred	Total number of COVID-19 vaccination booster doses administered 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
new_people_vaccinated_smoothed	Daily number of people receiving their first vaccine dose (7-day smoothed)
new_people_vaccinated_smoothed_per_hundred	Daily number of people receiving their first vaccine dose (7-day smoothed) per 100 people in the total population

From the above table, total_vaccinations, total_vaccinations_per_hundred is irrelevant as new_vaccinations column is used which can calculate the total. new_vaccinations_smoothed and new_vaccinations_smoothed_per_million become irrelevant as new_vaccinations column are used. people_fully_vaccinated and people_fully_vaccinated_per_hundred become irrelevant as it stored the total number of people only. new_people_vaccinated_smoothed_per_hundred is irrelevant as new_people_vaccinated_smoothed is used instead. total_boosters and total_boosters_per_hundred are not used in this case as only the total number are stored.

Others

Variable	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
population	Population (latest available values). See https://github.com/owid/covid-19-data/blob/master/scripts/input/un/population_latest.csv for full list of sources
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

From the table above, median_age, aged_70_older, gdp_per_capita, extreme_poverty, female_smokers, male_smokers, human_development_index, handwashing_facilities, hospital_beds_per_thousand are not used as it is not relevant to the analysis that I am doing. cardiovasc_death_rate and diabetes_prevalence are not used as the data is from 2017 which are irrelevant to the analysis that I am doing.

I believe my database is in third normal form, but I am not very sure whether it is in boyce-codd normal form. Starting from country table, with continent_id it will be able to retrieve the continent the country belongs to. For other tables, it uses iso_code as a foreign key so with iso_code from Country table, it can retrieve any data from any table with just iso_code. Most of the columns are prime attribute to their individual table.

Stage 3. Create the database

3.1 Building Of Database Structure In MySQL

Creating Database Command

CREATE DATABASE Covid;

Creating Of User And Granting Of The Appropriate Permission

CREATE USER 'covidUser'@'127.0.0.1' IDENTIFIED WITH mysql_native_password BY 'password';

GRANT ALTER, SELECT, INSERT, CREATE, DROP, REFERENCES on Covid.* TO 'covidUser'@'127.0.0.1' WITH GRANT OPTION;

FLUSH PRIVILEGES;

** Reasoning for granting ALTER, SELECT, INSERT, CREATE, DROP, REFERENCES to covidUser and not just the SELECT option is because I wrote script to migrate the data into the database therefore this option are needed.**

Creating Of Tables In Database

Tables	Query
Continent	CREATE TABLE `Continent` (`id` int AUTO_INCREMENT PRIMARY KEY, `continent` varchar(255));
Country	CREATE TABLE `Country` (`iso_code` varchar(255) PRIMARY KEY, `location` varchar(255), `continent_id` int);

NewCase	CREATE TABLE `NewCase` (
	  `id` int AUTO_INCREMENT PRIMARY KEY,
	  `date` datetime,
	  `new_case` int,
	  `new_death` int,
	  `iso_code` varchar(255)
	);

Hospitalize	CREATE TABLE `Hospitalize` (
	  `id` int AUTO_INCREMENT PRIMARY KEY,
	  `date` datetime,
	  `icu_patient` int,
	  `hosp_patients` int,
	  `iso_code` varchar(255)
	);

Test	CREATE TABLE `Test` (
	  `id` int AUTO_INCREMENT PRIMARY KEY,
	  `date` datetime,
	  `new_tests` int,
	  `positive_rate` double,
	  `iso_code` varchar(255)
	);

Vaccination	CREATE TABLE `Vaccination` (
	  `id` int AUTO_INCREMENT PRIMARY KEY,
	  `date` datetime,
	  `new_vaccinations_smoothed` int,
	  `new_people_vaccinated_smoothed` int,
	  `iso_code` varchar(255)
	);

Population	CREATE TABLE `Population` (
	  `id` int AUTO_INCREMENT PRIMARY KEY,
	  `population` int,
	  `population_density` int,
	  `aged_65_older` double,
	  `life_expectancy` double,
	  `iso_code` varchar(255)
	);

Stringency	CREATE TABLE `Stringency` (
	  `id` int AUTO_INCREMENT PRIMARY KEY,
	  `date` datetime,
	  `stringency_index` double,
	  `iso_code` varchar(255)
	);

Adding Foreign Key Constraint To Tables

ALTER TABLE `Country` ADD FOREIGN KEY (`continent_id`) REFERENCES `Continent` (`id`);

ALTER TABLE `NewCase` ADD FOREIGN KEY (`iso_code`) REFERENCES `Country` (`iso_code`);

ALTER TABLE `Hospitalize` ADD FOREIGN KEY (`iso_code`) REFERENCES `Country` (`iso_code`);

ALTER TABLE `Test` ADD FOREIGN KEY (`iso_code`) REFERENCES `Country` (`iso_code`);

ALTER TABLE `Vaccination` ADD FOREIGN KEY (`iso_code`) REFERENCES `Country` (`iso_code`);

ALTER TABLE `Population` ADD FOREIGN KEY (`iso_code`) REFERENCES `Country` (`iso_code`);

ALTER TABLE `Stringency` ADD FOREIGN KEY (`iso_code`) REFERENCES `Country` (`iso_code`);

3.2 Populating Data From CSV To MySQL

To populate data from csv to MySQL I chose to write a nodejs script. The script can be found mid-term/scripts/migration.js. I will go through briefly on the code that is in the migration.js. The code that are before the createTable function are setup code to use mysql library, reading the csv file data, and setting up the connection to connect to the database.

createTable function is to create all the tables and it's references, the only different is that I added **IF NOT EXISTS** keyword so that if the table did not delete successfully it won't crash the script when it tries to create the table again.

deleteTable function is to delete all the tables that are in Covid database. It will select all the tables from information_schema.tables first then loop through the list of tables to delete except the Country and Continent table. Reasoning is that the primary key are used as foreign key in other table so deleting other tables first then Country table then Continent table would not throw error because of foreign key.

The run function is the main function of this script so when the script run, it will call deleteTable function and createTable function.

```
const continents = new Set();
let countryData = {};
const caseData = [];
const hospitalizeData = [];
const testData = [];
const vaccinationData = [];
const stringencyData = [];
let populationData = {};
```

The above screenshot is to create all the variable named according to their table name. The code after declaration of the variables is a for loop to access each row of data in the csv file. The first few if is to check for if the current row is the header in the csv which is the first row, if the first column is empty and if the iso_code begin with OWID_ (those are continent data which we don't want to include). If those condition are met it will skip those rows. The continents variable type is a set because we only 1 unique record of the continent. For countryData and populationData it is type object, the reason for this is that we only want 1 record for each country. For countryData, every country will only have 1 entry for iso_code, location and continent_id. For populationData, all the data that are provided in the csv are duplicated data(latest number only) and they are all the same value throughout the csv for each country so there isn't a point to store multiple entry of it and just store 1 record for each country. For the remaining table data, it will store all the entry as most of the data is updated daily. One thing to note is that because the

csv have a lot of missing data which are blank for those number field so 0 will be inserted in.

After processing all the data, continents will be converted from set to array and is being inserted into the Continent table first. This is because country need to use the Continent table primary key as foreign key reference. After continents are being inserted, it will be retrieved out and transformed to an object with the continent name as the key and the id as value. countryData will then be iterated and changing the continent name in countryData to the id of the continent that is in the database. After that the countryData will be inserted into the database first because other tables uses iso_code as the foreign key so it have to exist in the database first. populationData will have it's key removed because we stored in object but because mysql library only accept array as the value so the key have to be removed. After removing the key from populationData, we will then insert all the data for the remaining tables.

To run the script, just open the terminal and **cd mid-term** directory and run the following command in the terminal **npm run migrate** and the script will run and do its job. Benefit of using a script is that if there is a need to migrate the newer csv data simply just put in the latest csv into the scripts folder and run the script which will save a lot of time. Another benefit of using a script is that it is more maintainable when there is a need to create a new table or update the newer data as there is no need to create a table to store all the data in the csv before storing it to different tables. The last benefit is that you can export this project to any pc and run locally by creating the same database name and user(can be found in dbconfig.js) in mysql and run **npm run migrate** in the terminal in the project directory which will migrate all the data into the mysql automatically.

3.3 List SQL commands that answer questions identified in Stage 1/Step 3.

1. Does more people get vaccinated reduces the covid cases or deaths?

   Query:

   ```
   WITH newcase_query AS (SELECT DATE_FORMAT(`date`,'%Y-%m') AS date,
   FORMAT(SUM(new_case),0) AS 'Monthly Total Case',
   FORMAT(SUM(new_death),0) AS 'Monthly Total Death' FROM NewCase
   GROUP BY DATE_FORMAT(`date`,'%Y-%m')), vaccination_query AS (SELECT
   DATE_FORMAT(`date`,'%Y-%m') AS date,
   FORMAT(SUM(new_vaccinations_smoothed),0) AS 'Monthly Total
   Vaccination', FORMAT(SUM(new_people_vaccinated_smoothed),0) AS
   'Monthly Total Unique Vaccinated People' FROM Vaccination GROUP BY
   DATE_FORMAT(`date`,'%Y-%m')) SELECT * FROM newcase_query LEFT JOIN
   vaccination_query USING (date) ORDER BY date;
   ```

Result:

date	Monthly Total Case	Monthly Total Death	Monthly Total Vaccination	Monthly Total Unique Vaccinated People
2020-01	9,370	196	0	0
2020-02	75,393	2,724	0	0
2020-03	783,398	41,725	0	0
2020-04	2,441,126	195,882	0	0
2020-05	2,906,732	152,123	0	0
2020-06	4,314,179	143,814	0	0
2020-07	7,112,144	175,038	0	0
2020-08	7,938,857	183,054	0	0
2020-09	8,501,563	164,550	0	0
2020-10	12,104,834	181,544	0	0
2020-11	17,271,368	276,896	0	0
2020-12	19,531,080	355,791	8,681,300	5,547,822
2021-01	19,547,470	417,875	87,670,229	53,240,643
2021-02	11,249,736	309,940	159,226,131	84,800,664
2021-03	14,830,386	299,817	328,486,853	190,350,684
2021-04	22,532,828	379,018	521,456,730	254,599,305
2021-05	19,700,040	379,184	774,239,333	242,648,602
2021-06	11,476,813	271,546	1,126,058,339	476,608,572
2021-07	15,736,865	261,519	1,059,291,456	555,869,167
2021-08	19,850,387	300,850	1,193,866,187	614,751,121
2021-09	15,922,609	265,793	968,110,515	466,423,482
2021-10	13,098,230	217,247	804,490,870	399,948,113
2021-11	15,640,830	216,051	936,306,493	323,002,992
2021-12	25,603,725	219,364	1,137,104,173	271,387,958
2022-01	89,370,887	238,737	965,948,100	246,788,092
2022-02	58,102,353	278,772	661,389,176	155,291,859
2022-03	51,441,412	177,802	522,548,259	107,187,490
2022-04	25,002,558	87,738	339,303,766	70,481,582
2022-05	16,191,815	53,007	242,863,460	50,812,802
2022-06	11,738,548	31,156	145,917,485	19,395,798

This is the result for monthly world total case, total death, total vaccination and total unique vaccinated people. From the result, we can see that before the vaccination is release, the covid case and death has been increasing monthly. When there are vaccines, there was a shortage in the beginning and not a lot of people are able to get vaccinated. In Singapore, the government prioritized elderly who are above 65-year-old to get vaccinated first as they are in the high-risk category. From the data we can see that starting from June 2020 there is more vaccine created and more people are vaccinated. The unique vaccinated people means that new people getting their first vaccine dose. We can see that there are a lot of people who are vaccinated but the monthly total case has been increasing and still a lot of cases but then the death rate is not that high. In conclusion, more people get vaccinated does not mean that cases will decrease but death rate will decrease for sure.

2. Which top 10 country have the lowest total covid cases against the total population of the country?

Query:

```
WITH newCase_query AS (SELECT SUM(new_case) AS total_case, iso_code FROM NewCase GROUP BY iso_code), country_query AS (SELECT iso_code, location FROM Country), population_query AS (SELECT iso_code, population, population_density FROM Population) SELECT country_query.location AS Country,
FORMAT((newCase_query.total_case/population_query.population),10) AS 'Total Covid Case Against Population', FORMAT(population,0) AS Population,
```

population_query.population_density AS 'Population Density' FROM newCase_query LEFT JOIN country_query USING (iso_code) LEFT JOIN population_query USING (iso_code) WHERE (newCase_query.total_case/population_query.population) > 0 ORDER BY (newCase_query.total_case/population_query.population) LIMIT 10;

Result:

Country	Total Covid Case Against Population	Population	Population Density
North Korea	0.0000000380	25,887,045	211
Macao	0.0001609980	658,391	20546
Marshall Islands	0.0003186950	59,618	295
Micronesia (country)	0.0003268670	116,255	150
Niger	0.0003595180	25,130,810	16
Chad	0.0004389000	16,914,985	11
Yemen	0.0003878230	30,490,639	53
Tanzania	0.0005750710	61,498,438	64
China	0.0006143130	1,444,216,102	147
Saint Helena	0.0006562750	6,095	0

From the result, this are the top 10 country that have and is sorted according to the total covid cases the country has divided by its population. The smaller the ratio would mean the country are better at controlling this pandemic. However, this can be a flaw in the dataset too because there are country that did not report their covid cases and death which make it inaccurate to say that this are the top 10 country that have the lowest total covid cases against the population.

3. As the cases increase, people who are vaccinated increases, do more people get hospitalized?

Query:

WITH newcase_query AS (SELECT DATE_FORMAT(`date`,'%Y-%m') AS date, FORMAT(SUM(new_case),0) AS 'Monthly Total Case' FROM NewCase GROUP BY DATE_FORMAT(`date`,'%Y-%m')), hospitalize_query AS (SELECT DATE_FORMAT(`date`,'%Y-%m') AS date, FORMAT(SUM(icu_patient),0) AS 'Monthly Total ICU Patient', FORMAT(SUM(hosp_patients),0) AS 'Monthly Total Hospital Patients' FROM Hospitalize GROUP BY DATE_FORMAT(`date`,'%Y-%m')), vaccination_query AS (SELECT DATE_FORMAT(`date`,'%Y-%m') AS date, FORMAT(SUM(new_vaccinations_smoothed),0) AS 'Monthly Total Vaccination', FORMAT(SUM(new_people_vaccinated_smoothed),0) AS 'Monthly Total Unique Vaccinated People' FROM Vaccination GROUP BY DATE_FORMAT(`date`,'%Y-%m')) SELECT * FROM newcase_query LEFT JOIN

hospitalize_query USING (date) LEFT JOIN vaccination_query USING (date) ORDER BY date;

Result:

date	Monthly Total Case	Monthly Total ICU Patient	Monthly Total Hospital Patients	Monthly Total Vaccination	Monthly Total Unique Vaccinated People
2020-01	9,370	0	2	0	0
2020-02	75,393	346	1,691	0	0
2020-03	783,398	139,388	807,622	0	0
2020-04	2,441,126	586,946	3,028,278	0	0
2020-05	2,906,732	277,699	1,820,039	0	0
2020-06	4,314,179	142,183	827,117	0	0
2020-07	7,112,144	356,381	1,368,310	0	0
2020-08	7,938,857	543,697	1,873,969	0	0
2020-09	8,501,563	500,693	1,643,747	0	0
2020-10	12,104,834	767,197	3,196,550	0	0
2020-11	17,271,368	1,343,784	7,196,904	0	0
2020-12	19,531,080	1,592,796	8,545,051	8,681,300	5,547,822
2021-01	19,547,470	1,775,374	9,224,968	87,670,229	53,240,643
2021-02	11,249,736	1,256,597	5,914,532	159,226,131	84,800,664
2021-03	14,830,386	1,127,174	5,906,948	328,486,853	190,350,684
2021-04	22,532,828	1,265,359	6,180,684	521,456,730	254,599,305
2021-05	19,700,040	1,097,823	3,827,031	774,239,333	242,648,602
2021-06	11,476,813	712,845	1,720,420	1,126,058,339	476,608,572
2021-07	15,736,865	614,129	1,925,969	1,059,291,456	555,869,167
2021-08	19,850,387	1,090,018	4,253,585	1,193,866,187	614,751,121
2021-09	15,922,609	1,149,437	4,625,982	968,110,515	466,423,482
2021-10	13,098,230	879,988	4,022,843	804,490,870	399,948,113
2021-11	15,640,830	825,566	4,344,784	936,306,493	323,002,992
2021-12	25,603,725	1,099,252	5,350,944	1,137,104,173	271,387,958
2022-01	89,370,887	1,374,358	8,931,299	965,948,100	246,788,092
2022-02	58,102,353	895,911	6,557,695	661,389,176	155,291,859
2022-03	51,441,412	487,570	3,897,354	522,548,259	107,187,490
2022-04	25,002,558	297,629	3,024,399	339,303,766	70,481,582
2022-05	16,191,815	213,413	2,451,686	242,863,460	50,812,802
2022-06	11,738,548	118,341	1,299,038	145,917,485	19,395,798

Based on the result, we can see that the monthly total admitted icu patient because of covid for the whole world increases when the vaccine was not ready, and the monthly total hospital patient admitted because of covid increases. As more vaccine is available to the world and more people vaccinated, we can see that the monthly total icu patient who admitted because of covid, and monthly total hospital patient admitted because of covid have decreased significantly as the time goes.

So, the key takeaway from the 3 question I have gone through is that when there is vaccine, we should take it so that we will have lesser risk to death, personal hygiene and following country protocol is a must because if there is still a covid case in the world, it will still spread so taking care of ourselves also play a part of taking care of others.

Stage 4. Create a simple web application

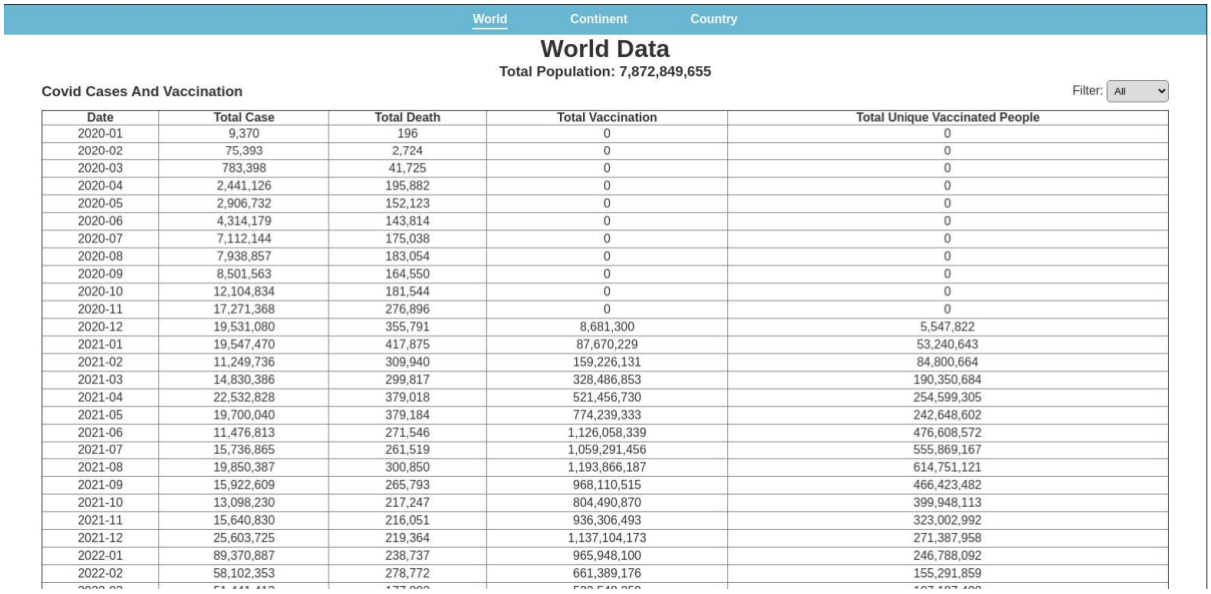
Instruction To Run The Web Application

My Cousera Lab URL is:
https://hub.labs.coursera.org:443/connect/sharedsgrfbvwc?forceRefresh=false&path=%2F%3Ffolder%3D%2Fhome%2Fcoder%2Fproject

- **cd mid-term**
- Open the terminal and run **npm start**
- Visit localhost:8088/

****Note: Coursera Browser Preview doesn't allow select option to be displayed, use the down and up arrow key to navigate to different option and use the enter button to confirm the select option****

Covid Cases And Vaccination Table



Date	Total Case	Total Death	Total Vaccination	Total Unique Vaccinated People
2020-01	9,370	196	0	0
2020-02	75,393	2,724	0	0
2020-03	783,398	41,725	0	0
2020-04	2,441,126	195,882	0	0
2020-05	2,906,732	152,123	0	0
2020-06	4,314,179	143,814	0	0
2020-07	7,112,144	175,038	0	0
2020-08	7,938,857	183,054	0	0
2020-09	8,501,563	164,550	0	0
2020-10	12,104,834	181,544	0	0
2020-11	17,271,368	276,896	0	0
2020-12	19,531,080	355,791	8,681,300	5,547,822
2021-01	19,547,470	417,875	87,670,229	53,240,643
2021-02	11,249,736	309,940	159,226,131	84,800,664
2021-03	14,830,386	299,817	328,486,853	190,350,684
2021-04	22,532,828	379,018	521,456,730	254,599,305
2021-05	19,700,040	379,184	774,239,333	242,648,602
2021-06	11,476,813	271,546	1,126,058,339	476,608,572
2021-07	15,736,865	261,519	1,059,291,456	555,869,167
2021-08	19,850,387	300,850	1,193,866,187	614,751,121
2021-09	15,922,609	265,793	968,110,515	466,423,482
2021-10	13,098,230	217,247	804,490,870	399,948,113
2021-11	15,640,830	216,051	936,306,493	323,002,992
2021-12	25,603,725	219,364	1,137,104,173	271,387,958
2022-01	89,370,887	238,737	965,948,100	246,788,092
2022-02	58,102,353	278,772	661,389,176	155,291,859

World Page

World Continent Country

Continent Data

Continents: Asia

Asia Population: 4,678,503,132

Covid Cases And Vaccination Filter: All

Date	Total Case	Total Death	Total Vaccination	Total Unique Vaccinated People
2020-01	9,333	196	0	0
2020-02	73,839	2,691	0	0
2020-03	96,833	4,240	0	0
2020-04	342,990	10,272	0	0
2020-05	618,468	12,125	0	0
2020-06	1,166,691	26,144	0	0
2020-07	1,991,775	39,750	0	0
2020-08	2,818,014	46,431	0	0
2020-09	3,555,356	52,406	0	0
2020-10	3,046,598	48,216	0	0
2020-11	3,077,099	48,918	0	0
2020-12	3,101,894	45,603	3,685,785	670,262
2021-01	2,398,667	35,474	32,662,456	7,922,381
2021-02	1,909,888	25,658	58,372,132	22,156,997
2021-03	3,484,925	31,819	140,257,444	62,201,475
2021-04	11,017,797	92,762	266,005,455	97,124,593
2021-05	11,741,131	165,888	495,465,069	78,813,445
2021-06	4,556,963	98,070	830,871,787	315,459,120
2021-07	6,250,643	104,619	750,580,458	391,679,601
2021-08	8,096,119	138,677	889,947,627	465,054,242
2021-09	5,794,363	88,884	717,480,365	358,173,703
2021-10	3,695,766	49,145	545,738,159	302,591,474

Continent Page

World Continent Country

Country Data

Country: Afghanistan

Population: 39,835,428 **Population Density: 54** **Age 65 And Above Population: 2.581% (1,028,152)** **Life Expectancy: 64.83**

Covid Cases And Vaccination Filter: All

Date	Total Case	Total Death	Total Vaccination	Total Unique Vaccinated People
2020-02	5	0	0	0
2020-03	163	4	0	0
2020-04	1,661	56	0	0
2020-05	13,353	194	0	0
2020-06	16,265	485	0	0
2020-07	5,183	536	0	0
2020-08	1,620	131	0	0
2020-09	1,106	56	0	0
2020-10	1,980	71	0	0
2020-11	4,881	230	0	0
2020-12	6,115	426	0	0
2021-01	2,693	211	0	0
2021-02	691	43	8,202	8,202
2021-03	740	41	85,894	85,894
2021-04	3,291	141	219,606	204,969
2021-05	12,093	319	285,838	184,688
2021-06	46,821	1,927	242,899	172,466
2021-07	28,495	1,837	337,196	102,550
2021-08	6,066	410	591,825	17,662

Country Page

The above table is about covid cases and vaccinations which address the question of Does more people get vaccinated reduces the covid cases or deaths?

For the table in world page, the table can filter by month which will show the daily data. This was achieved by changing DATE_FORMAT(`date`,'%Y-%m') AS date to DATE_FORMAT(`date`,'%Y-%m-%d') AS date and adding a WHERE DATE_FORMAT(`date`,'%Y-%m')=? and changed GROUP BY DATE_FORMAT(`date`,'%Y-%m') to GROUP BY DATE_FORMAT(`date`,'%Y-%m-%d') for the query in 3.3 first question query.

For continent page, the table can be filter by month to show the daily data and by continent. To achieve that, I have added the WHERE iso_code IN (SELECT iso_code FROM Country WHERE continent_id=? To the query in 3.3 first question query. And

to display daily data I changed DATE_FORMAT(`date`,'%Y-%m') AS date to DATE_FORMAT(`date`,'%Y-%m-%d') AS date and added WHERE DATE_FORMAT(`date`,'%Y-%m')=? AND iso_code IN (SELECT iso_code FROM Country WHERE continent_id=?) GROUP BY DATE_FORMAT(`date`,'%Y-%m-%d') for the query in 3.3 first question query.

For country page, the table can be filter by month which will show the daily data and by country. To achieve that, I have added the WHERE iso_code=? To the query in 3.3 first question query. And to display daily data I changed DATE_FORMAT(`date`,'%Y-%m') AS date to DATE_FORMAT(`date`,'%Y-%m-%d') AS date and added WHERE DATE_FORMAT(`date`,'%Y-%m')=? AND iso_code=? GROUP BY DATE_FORMAT(`date`,'%Y-%m-%d') for the query in 3.3 first question query.

Covid Cases And Hospitalization

Covid Cases And Hospitalization

Filter: All

Date	Total Case	Total Death	Total ICU Patient	Total Hospitalize Patient
2020-01	9,370	196	0	2
2020-02	75,393	2,724	346	1,691
2020-03	783,398	41,725	139,388	807,622
2020-04	2,441,126	195,882	586,946	3,028,278
2020-05	2,906,732	152,123	277,699	1,820,039
2020-06	4,314,179	143,814	142,183	827,117
2020-07	7,112,144	175,038	356,381	1,368,310
2020-08	7,938,857	183,054	543,697	1,873,969
2020-09	8,501,563	164,550	500,693	1,643,747
2020-10	12,104,834	181,544	767,197	3,196,550
2020-11	17,271,368	276,896	1,343,784	7,196,904
2020-12	19,531,080	355,791	1,592,796	8,545,051
2021-01	19,547,470	417,875	1,775,374	9,224,968
2021-02	11,249,736	309,940	1,256,597	5,914,532
2021-03	14,830,386	299,817	1,127,174	5,906,948
2021-04	22,532,828	379,018	1,265,359	6,180,684
2021-05	19,700,040	379,184	1,097,823	3,827,031
2021-06	11,476,813	271,546	712,845	1,720,420
2021-07	15,736,865	261,519	614,129	1,925,969
2021-08	19,850,387	300,850	1,090,018	4,253,585
2021-09	15,922,609	265,793	1,149,437	4,625,982
2021-10	13,098,230	217,247	879,988	4,022,843
2021-11	15,640,830	216,051	825,566	4,344,784
2021-12	25,603,725	219,364	1,099,252	5,350,944
2022-01	89,370,887	238,737	1,374,358	8,931,299
2022-02	58,102,353	278,772	895,911	6,557,695
2022-03	51,441,412	177,802	487,570	3,897,354
2022-04	25,002,558	87,738	297,629	3,024,399

World Page

Covid Cases And Hospitalization

Filter: All

Date	Total Case	Total Death	Total ICU Patient	Total Hospitalize Patient
2020-01	9,333	196	0	0
2020-02	73,839	2,691	0	0
2020-03	96,833	4,240	954	19,024
2020-04	342,990	10,272	3,614	62,899
2020-05	618,468	12,125	1,982	33,783
2020-06	1,166,691	26,144	1,167	15,204
2020-07	1,991,775	39,750	998	33,358
2020-08	2,818,014	46,431	1,855	56,404
2020-09	3,555,356	52,406	11,187	68,521
2020-10	3,046,598	48,216	14,212	92,812
2020-11	3,077,099	48,918	11,746	110,651
2020-12	3,101,894	45,603	20,503	163,574
2021-01	2,398,667	35,474	36,248	245,488
2021-02	1,909,888	25,658	28,562	178,937
2021-03	3,484,925	31,819	20,726	123,561
2021-04	11,017,797	92,762	20,160	122,637
2021-05	11,741,131	165,888	37,383	235,552
2021-06	4,556,963	98,070	45,168	260,272
2021-07	6,250,643	104,619	51,855	349,109
2021-08	8,096,119	138,677	70,297	543,093
2021-09	5,794,363	88,884	64,082	435,442
2021-10	3,695,766	49,145	40,542	214,577
2021-11	2,799,526	42,840	34,379	144,751
2021-12	2,528,915	40,899	43,173	113,180
2022-01	15,854,431	37,467	32,183	174,189
2022-02	16,504,393	56,715	32,966	316,997
2022-03	22,838,210	53,670	56,608	326,799
2022-04	8,593,541	21,466	37,741	123,278
2022-05	4,526,863	7,721	15,175	67,057

Continent Page

Covid Cases And Hospitalization

Filter: All

Date	Total Case	Total Death	Total ICU Patient	Total Hospitalize Patient
2020-02	5	0	0	0
2020-03	163	4	0	0
2020-04	1,661	56	0	0
2020-05	13,353	194	0	0
2020-06	16,265	485	0	0
2020-07	5,183	536	0	0
2020-08	1,620	131	0	0
2020-09	1,106	56	0	0
2020-10	1,980	71	0	0
2020-11	4,881	230	0	0
2020-12	6,115	426	0	0
2021-01	2,693	211	0	0
2021-02	691	43	0	0
2021-03	740	41	0	0
2021-04	3,291	141	0	0
2021-05	12,093	319	0	0
2021-06	46,821	1,927	0	0
2021-07	28,495	1,837	0	0
2021-08	6,066	410	0	0
2021-09	1,954	86	0	0
2021-10	1,076	76	0	0
2021-11	1,039	28	0	0
2021-12	795	48	0	0
2022-01	4,848	58	0	0
2022-02	10,733	184	0	0
2022-03	4,088	72	0	0

Country Page

The above table is the Covid Cases And Hospitalization table which address the question of As the cases increase, people who are vaccinated increases, do more people get hospitalized?

For the table in world page, the table can filter by month which will show the daily data. This was achieved by changing DATE_FORMAT(`date`,'%Y-%m') AS date to DATE_FORMAT(`date`,'%Y-%m-%d') AS date and adding WHERE DATE_FORMAT(`date`,'%Y-%m')=? and changed GROUP BY DATE_FORMAT(`date`,'%Y-%m') to GROUP BY DATE_FORMAT(`date`,'%Y-%m-%d') for the query in 3.3 third question query.

For the table in continent page, the table can be filter by month which will show the daily data and by continent. To achieve that, I have added the WHERE iso_code IN (SELECT iso_code FROM Country WHERE continent_id=? To the query in 3.3 third question query. And to display daily data I changed DATE_FORMAT(`date`,'%Y-%m') AS date to DATE_FORMAT(`date`,'%Y-%m-%d') AS date and added WHERE DATE_FORMAT(`date`,'%Y-%m')=? AND iso_code IN (SELECT iso_code FROM Country WHERE continent_id=?) GROUP BY DATE_FORMAT(`date`,'%Y-%m-%d') for the query in 3.3 third question query.

For the table in country page, the table can be filter by month which will show the daily data and by continent. To achieve that, I have added a WHERE iso_code=? To the query in 3.3 third question query. And to display daily data I changed DATE_FORMAT(`date`,'%Y-%m') AS date to DATE_FORMAT(`date`,'%Y-%m-%d') AS date and added WHERE DATE_FORMAT(`date`,'%Y-%m')=? AND iso_code=? GROUP BY DATE_FORMAT(`date`,'%Y-%m-%d') for the query in 3.3 third question query.

Covid Cases Against Population

Country	Population	Total Covid Case	Total Covid Case Against Population	Population Density
North Korea	25,887,045	1	0.0000000380	211
Macao	658,391	106	0.0001609980	20546
Micronesia (country)	116,255	38	0.0003268670	150
Marshall Islands	59,618	19	0.0003186950	295
Niger	25,130,810	9,035	0.0003595180	16
Chad	16,914,985	7,424	0.0004389000	11
Yemen	30,490,639	11,825	0.0003878230	53
China	1,444,216,102	887,202	0.0006143130	147
Tanzania	61,498,438	35,366	0.0005750710	64
Saint Helena	6,095	4	0.0006562750	0
Sierra Leone	8,141,343	7,693	0.0009449300	104
Burkina Faso	21,497,097	21,044	0.0009789220	70
Democratic Republic of Congo	92,377,986	90,561	0.0009803300	35
Nigeria	211,400,704	256,741	0.0012144750	209
Sudan	44,909,351	62,554	0.0013928940	23
Liberia	5,180,208	7,619	0.0014707900	49
Mali	20,855,724	31,147	0.0014934500	15
Somalia	16,359,500	26,748	0.0016350130	23
South Sudan	11,381,377	17,697	0.0015549080	0
Tajikistan	9,749,625	17,786	0.0018242750	64
Benin	12,451,031	27,331	0.0021950790	99
Nicaragua	6,702,379	14,619	0.0021811650	51
Madagascar	28,427,333	65,016	0.0022870940	43
Eritrea	3,601,462	9,784	0.0027166740	44
Guinea	13,497,237	36,817	0.0027277430	51
Haiti	11,541,683	31,226	0.0027054970	398
Angola	33,933,611	99,761	0.0029398870	23

This is the Covid Cases Against Population which is addressing the question of Which top 10 country have the lowest total covid cases against the total population of the country? However, for this query I removed the LIMIT 10 from 3.3 second question query which will display for all the country.

Test

Filter: 2020-02

Date	Daily Case	Daily Death	Daily Test	Positive Test	Positive Rate(%)
2020-02-24	5	0	0	0	0
2020-02-25	0	0	0	0	0
2020-02-26	0	0	0	0	0
2020-02-27	0	0	0	0	0
2020-02-28	0	0	0	0	0
2020-02-29	0	0	0	0	0
Total	5	0	0	0	

The last table which is Test table that can be found in Country page. This table shows the daily covid cases and death along side with the daily test and daily positive rate from the test conducted. This table will provide us insight on how many were tested daily and its positive rate. This are the query used for the table.

WITH covid_query AS (SELECT DATE_FORMAT(`date`,'%Y-%m-%d') as date, FORMAT(SUM(new_case),0) as total_case, FORMAT(SUM(new_death),0) as total_death FROM NewCase WHERE DATE_FORMAT(`date`,'%Y-%m')=? AND iso_code = ? GROUP BY DATE_FORMAT(`date`,'%Y-%m-%d')), test_query AS (SELECT DATE_FORMAT(`date`,'%Y-%m-%d') as date, FORMAT(new_tests,0) as new_tests, FORMAT((positive_rate*new_tests),0) as total_positive_tests, positive_rate FROM Test WHERE DATE_FORMAT(`date`,'%Y-%m')=? AND iso_code = ?) SELECT * FROM covid_query LEFT JOIN test_query USING (date) ORDER BY date;

This query will total up all the new case and death by day, select the daily new tests, the tests that are positive and the positive rate from the Test table. The two query is joined into 1 table by using date.

```
SELECT FORMAT(SUM(new_tests),0) as total_tests,
FORMAT((SUM(positive_rate)*SUM(new_tests)),0) as total_positive_tests FROM
Test WHERE DATE_FORMAT(`date`,'%Y-%m')=? AND iso_code = ? GROUP BY
DATE_FORMAT(`date`,'%Y-%m')
```

This query will return us the sum of new test and the sum of positive rate * new test which will be grouped by month. This value is displayed in the last row of the table to show the total for the entire month.

This report is 25 page long so I would not explain in detail but be able to filter the date for different table, I have created a few api to get data based on the different criteria, mustachejs is unable to do that so I had to tap into javascript and use fetch function to call the api I created and changes the table data after getting the result.

In conclusion, this coursework is fruitful as it brushes up my concept and allow me to explore different area of interest of data rather than the school providing one. I apologize as certain part of the report I did not explain clearly and my bad English. But thank you for your time reading this lengthy report.