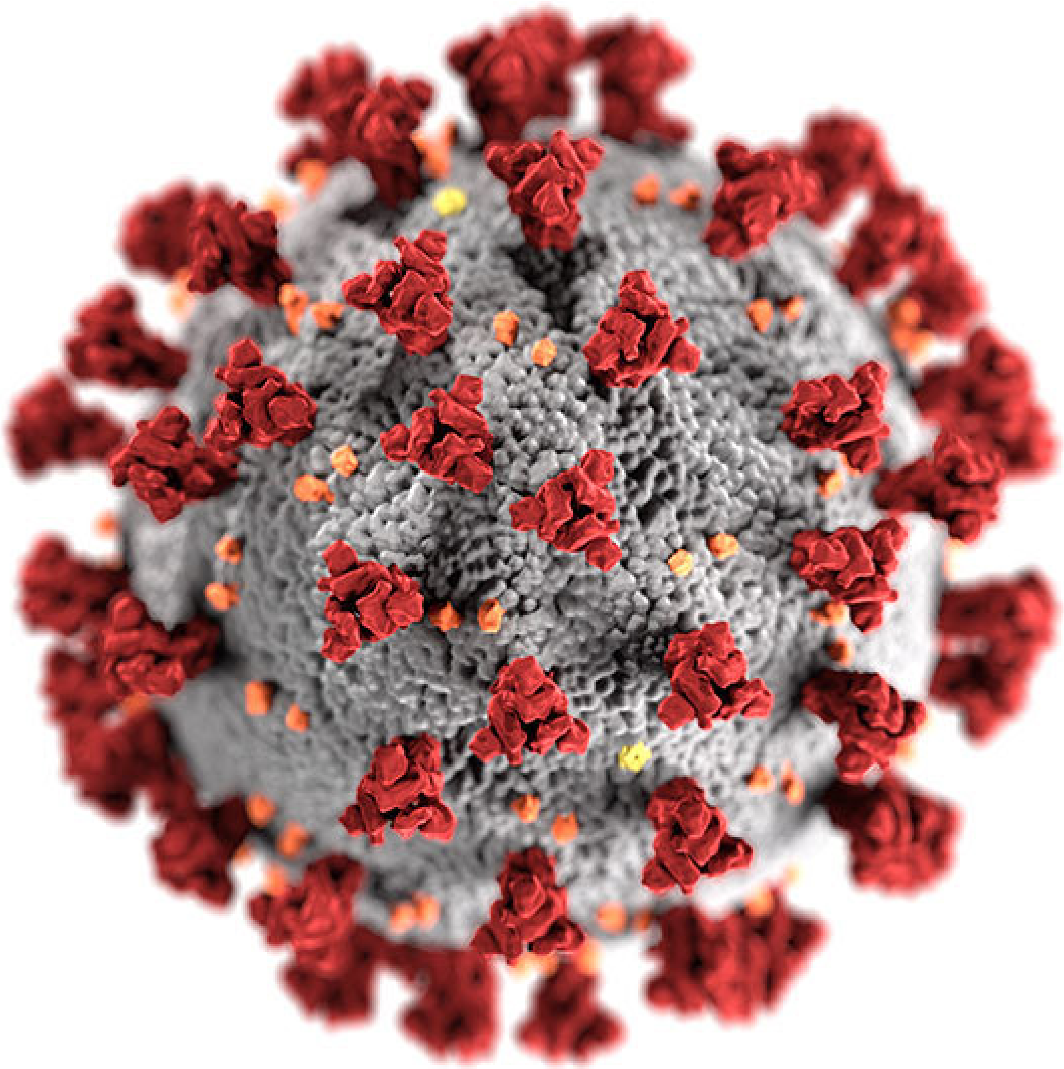
Covid-19 Analysis across countries

Case Report



Submitted by:
Ridha Altaf- 1591952
Imad Nizami- 1953306
Arjun Sareen- 1964736
Ifraz Ahmed- 1855952

TABLE OF CONTENT

Topic Name **Page Number**

INTRODUCTION:

Our case presentation covers the impeccable pandemic of Covid – 19 that has struck globally. While we're making history. Why not analyze it? The data we chose not only covers the impact of the virus internationally but globally describing the day to day effects of the disease in each continent along with their counties. Since the data itself is very large and sparsely distributed we chose HBase as the datastore and Hive as the interface for analysis of our dataset. We decided to focus on two continents: Asia and North America. Using the dataset, we are hoping to find a correlation between attributes such as the impact of covid on population density versus a population of a country in the designated continent. Furthermore, we plan to analyze the stringency index of countries within the continent to understand how they responded firsthand to the virus and the impact of the response on the total number of cases in each country of the continent. We also take a deeper dive into the data to further study the statistics of the total deaths in each country, followed by the number of hospital beds being used and the number of testing facilities in the country, and how those attributes affected the continent individually. While also studying the relation of age group and the population with diabetes in response to the virus. All of these questions are the basis of the analysis that truly helps us understand the impact of the virus on an international level. Our results below explain how each continent was affected during this time of crisis.

ABOUT THE DATA:

The dataset, "rearc-covid-19-world-cases-deaths-testing/" was retrieved from Amazon web services cloud front. "https://dj2taa9i652rf.cloudfront.net/. it consists of 50,000 + rows and 41 columns. The dataset goes the in-depth analysis of the coronavirus and how it has impacted globally. It shows the emerging cases, deaths, hand wash facilities, and the impact on the virus in each continent specifying countries in each continent by each day. Furthermore, the data shows the day to day impact of the virus using specific dates and averages from each country to provide an accurate representation of the spread. Below is a data dictionary that will help describe the columns included in the dataset. The dataset in and of itself is very sparse and randomly scattered which makes it a perfect match to use HBase and Hive.

DATA DICTIONARY	
ATTRIBUTE	**DESCRIPTION**
iso_code	ISO 3166-1 alpha-3 â€“ three-letter country codes
continent	The 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
total_tests	Total tests for COVID-19
new_tests	New tests for COVID-19
new_tests_smoothed	New tests for COVID-19 (7-day smoothed). For countries that don't report testing data daily, we assume that testing changed equally daily over any period 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_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_per_thousand	New tests for COVID-19 (7-day smoothed) per 1,000 people
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)
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_units	Units used by the location to report its testing data
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	The 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	The 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 summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable, and have a decent standard of living

METHODOLOGY:

Our group decided to use HBase as our data store of choice because it can handle sparse data and null values better than a traditional RDBMS. To connect to HBase, we used the PuTTY client and PSFTP to load the dataset into the HBase environment. The steps our group used for ETL (extract, transform, load) include the following:

ETL Process

1. **Download data**

- We used Amazon S3 (Simple Storage Service) Explorer to the Covid dataset https://dj2taa9i652rf.cloudfront.net/ using the 'rearc-covid-19-world-cases-deaths-testing/' table.
- Removed column headers directly from excel.
- Added ID column directly from excel.

2. **Create Table in HBase**

- Logged into HBase through PuTTY and created the table by specifying the table name and only ColumnFamilies.
  - No qualifiers yet.
- Create 'table_name', 'CF1', 'CF2', 'CF3', 'CF4'
- Create 'covid19', 'location_date', 'cases', 'tests', 'population_demographics'

3. **Import to HBase Server**

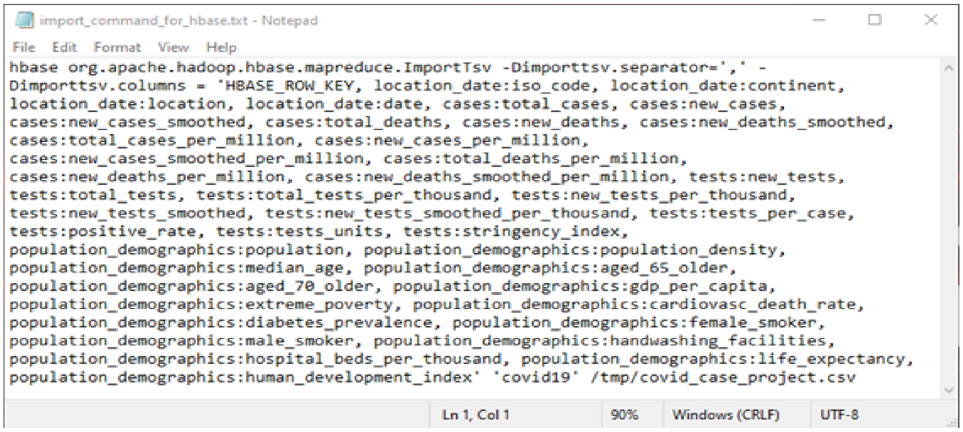
- Used PSFTP to import from local machine to the HBase server.
- Used the Unix terminal 'put' command.

4. **Copy from HBase into HDFS**

- Used -copyFromLocal command in HDFS <file name> <target directory>

5. **Import from HDFS into HBase Cluster**

- This is the step where we specified the Qualifiers for each Column Family

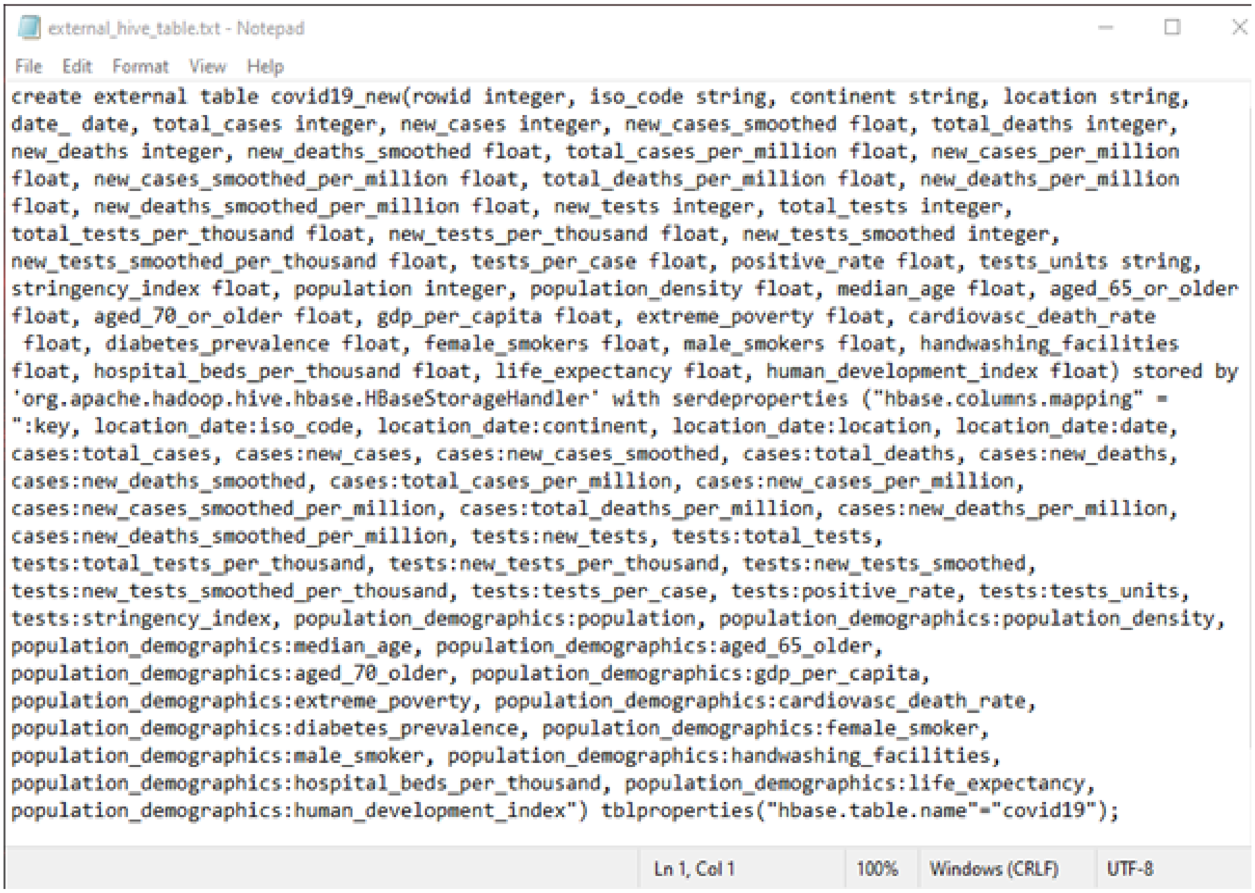


import_command_for_hbase.txt - Notepad
File Edit Format View Help
Ln 1, Col 1
90%
Windows (CRLF)
UTF-8

```
hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.separator=',' -
Dimporttsv.columns = 'HBASE_ROW_KEY, location_date:iso_code, location_date:continent,
location_date:location, location_date:date, cases:total_cases, cases:new_cases,
cases:new_cases_smoothed, cases:total_deaths, cases:new_deaths, cases:new_deaths_smoothed,
cases:total_cases_per_million, cases:new_cases_per_million,
cases:new_cases_smoothed_per_million, cases:total_deaths_per_million,
cases:new_deaths_per_million, cases:new_deaths_smoothed_per_million, tests:new_tests,
tests:total_tests, tests:total_tests_per_thousand, tests:new_tests_per_thousand,
tests:new_tests_smoothed, tests:new_tests_smoothed_per_thousand, tests:tests_per_case,
tests:positive_rate, tests:tests_units, tests:stringency_index,
population_demographics:population, population_demographics:population_density,
population_demographics:median_age, population_demographics:aged_65_older,
population_demographics:aged_70_older, population_demographics:gdp_per_capita,
population_demographics:extreme_poverty, population_demographics:cardiovasc_death_rate,
population_demographics:diabetes_prevalence, population_demographics:female_smoker,
population_demographics:male_smoker, population_demographics:handwashing_facilities,
population_demographics:hospital_beds_per_thousand, population_demographics:life_expectancy,
population_demographics:human_development_index' 'covid19' /tmp/covid_case_project.csv
```

1. **Create and Link Hive Table to HBase Table**

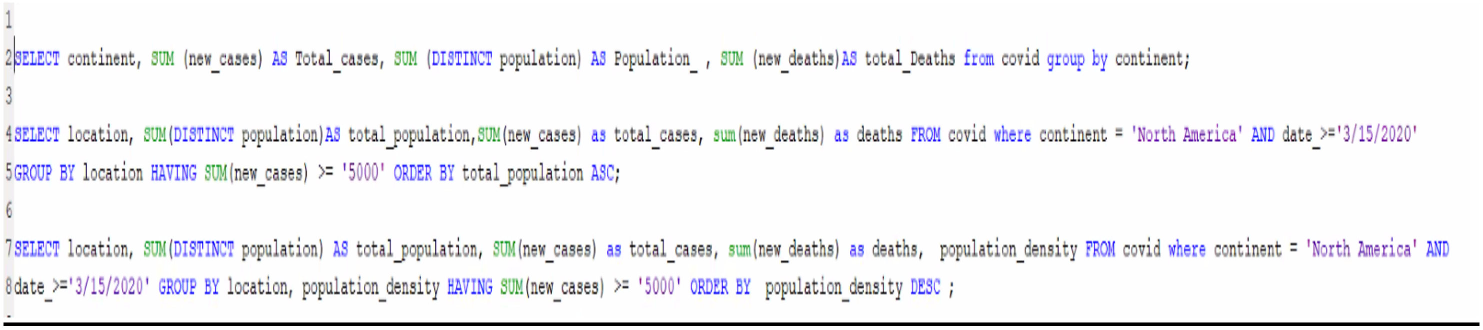
- Logged into hive shell to create an external table in Hive
- The external table must-have attribute names similar to the qualifiers from HBase but not necessary.
    - As long as the **order** of column headers in the hive table match with the HBase table, the data will be mapped to the correct column.



```
create external table covid19_new(rowid integer, iso_code string, continent string, location string,
date_ date, total_cases integer, new_cases integer, new_cases_smoothed float, total_deaths integer,
new_deaths integer, new_deaths_smoothed float, total_cases_per_million float, new_cases_per_million
float, new_cases_smoothed_per_million float, total_deaths_per_million float, new_deaths_per_million
float, new_deaths_smoothed_per_million float, new_tests integer, total_tests integer,
total_tests_per_thousand float, new_tests_per_thousand float, new_tests_smoothed integer,
new_tests_smoothed_per_thousand float, tests_per_case float, positive_rate float, tests_units string,
stringency_index float, population integer, population_density float, median_age float, aged_65_or_older
float, aged_70_or_older float, gdp_per_capita float, extreme_poverty float, cardiovasc_death_rate
 float, diabetes_prevalence float, female_smokers float, male_smokers float, handwashing_facilities
float, hospital_beds_per_thousand float, life_expectancy float, human_development_index float) stored by
'org.apache.hadoop.hive.hbase.HBaseStorageHandler' with serdeproperties ("hbase.columns.mapping" =
":key, location_date:iso_code, location_date:continent, location_date:location, location_date:date,
cases:total_cases, cases:new_cases, cases:new_cases_smoothed, cases:total_deaths, cases:new_deaths,
cases:new_deaths_smoothed, cases:total_cases_per_million, cases:new_cases_per_million,
cases:new_cases_smoothed_per_million, cases:total_deaths_per_million, cases:new_deaths_per_million,
cases:new_deaths_smoothed_per_million, tests:new_tests, tests:total_tests,
tests:total_tests_per_thousand, tests:new_tests_per_thousand, tests:new_tests_smoothed,
tests:new_tests_smoothed_per_thousand, tests:tests_per_case, tests:positive_rate, tests:tests_units,
tests:stringency_index, population_demographics:population, population_demographics:population_density,
population_demographics:median_age, population_demographics:aged_65_older,
population_demographics:aged_70_older, population_demographics:gdp_per_capita,
population_demographics:extreme_poverty, population_demographics:cardiovasc_death_rate,
population_demographics:diabetes_prevalence, population_demographics:female_smoker,
population_demographics:male_smoker, population_demographics:handwashing_facilities,
population_demographics:hospital_beds_per_thousand, population_demographics:life_expectancy,
population_demographics:human_development_index") tblproperties("hbase.table.name"="covid19");
```

○

Query for analysis



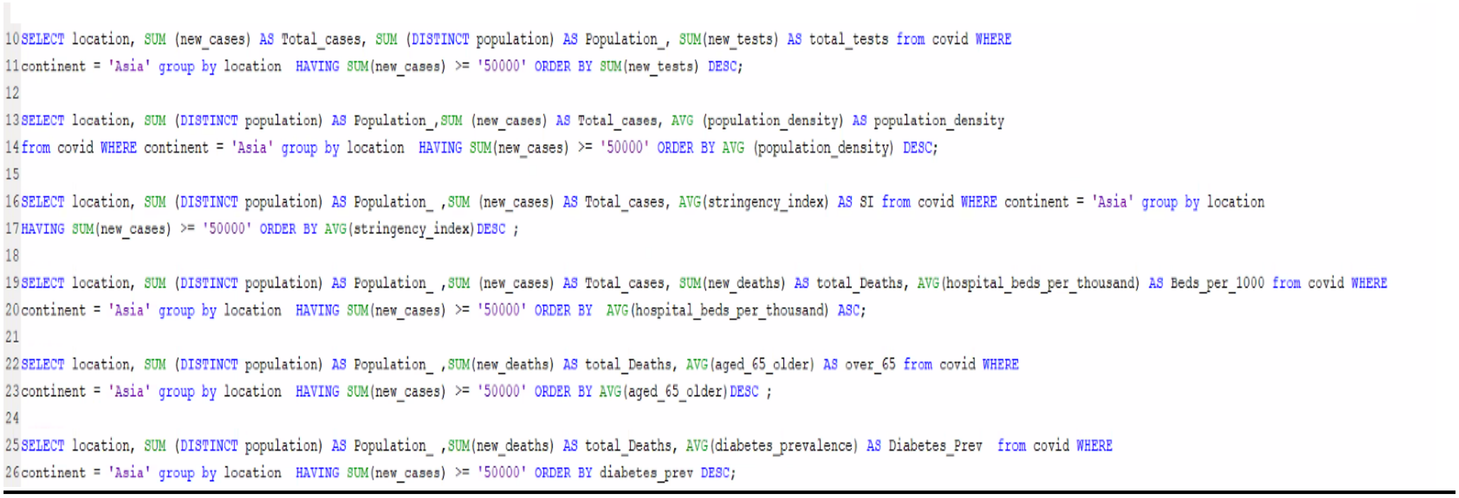
```sql

SELECT continent, SUM (new_cases) AS Total_cases, SUM (DISTINCT population) AS Population_ , SUM (new_deaths)AS total_Deaths from covid group by continent;

SELECT location, SUM(DISTINCT population)AS total_population,SUM(new_cases) as total_cases, sum(new_deaths) as deaths FROM covid where continent = 'North America' AND date >='3/15/2020'
GROUP BY location HAVING SUM(new_cases) >= '5000' ORDER BY total_population ASC;

SELECT location, SUM(DISTINCT population) AS total_population, SUM(new_cases) as total_cases, sum(new_deaths) as deaths,  population_density FROM covid where continent = 'North America' AND
date >='3/15/2020' GROUP BY location, population_density HAVING SUM(new_cases) >= '5000' ORDER BY  population_density DESC ;
```

Fig. Code for Covid Analysis in North America



```sql
SELECT location, SUM (new_cases) AS Total_cases, SUM (DISTINCT population) AS Population_, SUM(new_tests) AS total_tests from covid WHERE
continent = 'Asia' group by location  HAVING SUM(new_cases) >= '50000' ORDER BY SUM(new_tests) DESC;

SELECT location, SUM (DISTINCT population) AS Population_,SUM (new_cases) AS Total_cases, AVG (population_density) AS population_density
from covid WHERE continent = 'Asia' group by location  HAVING SUM(new_cases) >= '50000' ORDER BY AVG (population_density) DESC;

SELECT location, SUM (DISTINCT population) AS Population_ ,SUM (new_cases) AS Total_cases, AVG(stringency_index) AS SI from covid WHERE continent = 'Asia' group by location
HAVING SUM(new_cases) >= '50000' ORDER BY AVG(stringency_index)DESC ;

SELECT location, SUM (DISTINCT population) AS Population_ ,SUM (new_cases) AS Total_cases, SUM(new_deaths) AS total_Deaths, AVG(hospital_beds_per_thousand) AS Beds_per_1000 from covid WHERE
continent = 'Asia' group by location  HAVING SUM(new_cases) >= '50000' ORDER BY  AVG(hospital_beds_per_thousand) ASC;

SELECT location, SUM (DISTINCT population) AS Population_ ,SUM(new_deaths) AS total_Deaths, AVG(aged_65_older) AS over_65 from covid WHERE
continent = 'Asia' group by location  HAVING SUM(new_cases) >= '50000' ORDER BY AVG(aged_65_older)DESC ;

SELECT location, SUM (DISTINCT population) AS Population_ ,SUM(new_deaths) AS total_Deaths, AVG(diabetes_prevalence) AS Diabetes_Prev  from covid WHERE
continent = 'Asia' group by location  HAVING SUM(new_cases) >= '50000' ORDER BY diabetes_prev DESC;
```

Fig. Code for Covid Analysis in Asia

Functions and clauses used in our Analysis:

- SUM ()
- AVERAGE ()
- GROUP BY ()
- DISTINCT
- ORDER BY ()
- COUNT ()
- WHERE
- HAVING

ANALYSIS:

ANALYSIS ASSUMPTIONS:

1. We only focused our study on Asia and North America because the data from these two continents stood out and we were able to find more measures adopted by countries in these continents.
2. We only considered countries in Asia with Covid cases of 50,000 or more, because we didn't have any tangible data for other attributes for countries with cases less than 50,000.
3. For our analysis of the North American continent, we analyzed data for cases and other attributes after the 15th of March as cases started merging after mid of March. We also selected countries that had more than 5000 cases as countries less than that had a smaller amount of coherent data.
4. In our study, we discarded the US and China to avoid skewness in our charts and also to highlight other countries we already knew about.

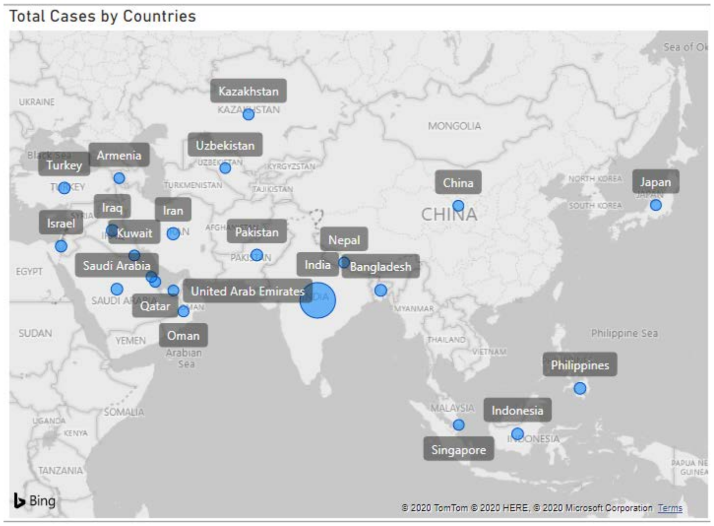
ANALYSIS RESULTS:

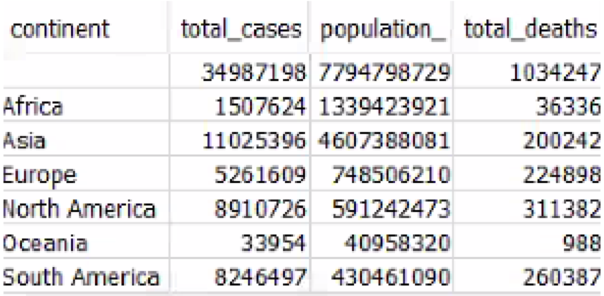
1. We wanted to have a basic idea of which continent was affected the most by the Covid pandemic. By using our first query, we were able to have a proper idea of the numbers. Asia was affected the most, followed by North America. However, what stood out to us, was that even though North America had the third-lowest population, it had the second-highest cases and the most deaths.

   Asia and Africa had the lowest death rate.

   Asia had the highest total cases of almost any continent. Asia's population is equivalent to almost **60%** of the world's total population and with it being the epicenter of the Covid pandemic, cases were at an all-time high.

   In **North America**, sheer negligence and slow response in the US, the country with the highest population in this continent, led to a spike in coronavirus cases, even when the North American population was equivalent to the **4.73%** of the world's population.





continent	total_cases	population_	total_deaths
	34987198	7794798729	1034247
Africa	1507624	1339423921	36336
Asia	11025396	4607388081	200242
Europe	5261609	748506210	224898
North America	8910726	591242473	311382
Oceania	33954	40958320	988
South America	8246497	430461090	260387

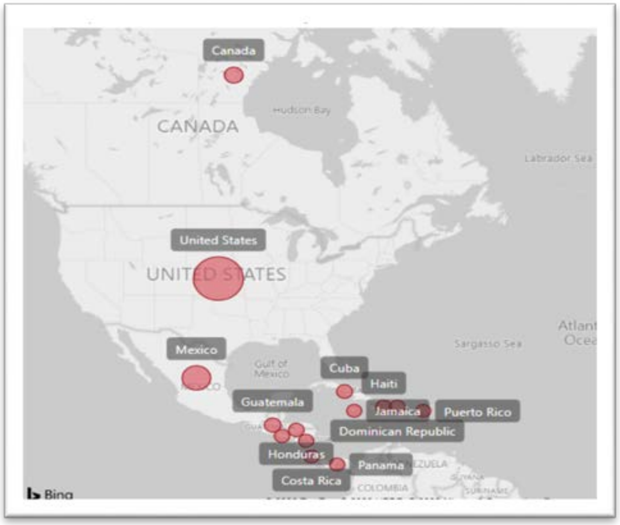
2. We focused on the North American continent. **Haiti** and **Cuba** were the best performing countries when comparing **population to total cases**. The **Dominican Republic** ranked last along with **Panama**, even when Panama has half the population of Cuba and Haiti.

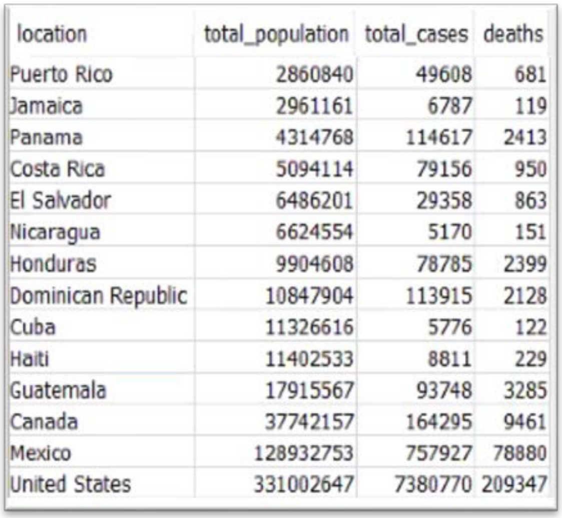
   Cuba had several advantages when the coronavirus pandemic began. Free universal healthcare, the world's highest ratio of doctors to population, and positive health indicators played a key role in controlling the spread of the virus. Cuba's reaction to the pandemic was swift, preparing quarantine facilities along with proper lockdown procedures helped curb the spread of the virus.

The **Haitian** government took proactive steps in combatting the virus. Additionally, Haitian authorities also shut down the border with the **Dominican Republic**. These efforts were undertaken to prevent the rapid spread of COVID-19 in Haiti by minimizing movement to and from the country and by identifying any possibly infected persons before entry.

The **Dominican Republic** suffered heavily. It's alarming spikes in cases forced Haiti to close off its borders from its neighbor, thereby hampering the import and export business. Its weak medical infrastructure hit maximum capacity early on, and as a result, Covid-19 spread. A vulnerable government means no one was in charge and as a result, it was very hard to implement any proper procedures to combat the coronavirus.

Panama was hit the worst. Its weak economy was crippled further when the government declared a state of emergency. Instead of implementing proper, conventional procedures for social distancing and quarantines, Panama instead opted to separate men from women, where women were allowed to leave their homes on someday, and men were allowed out on other days.



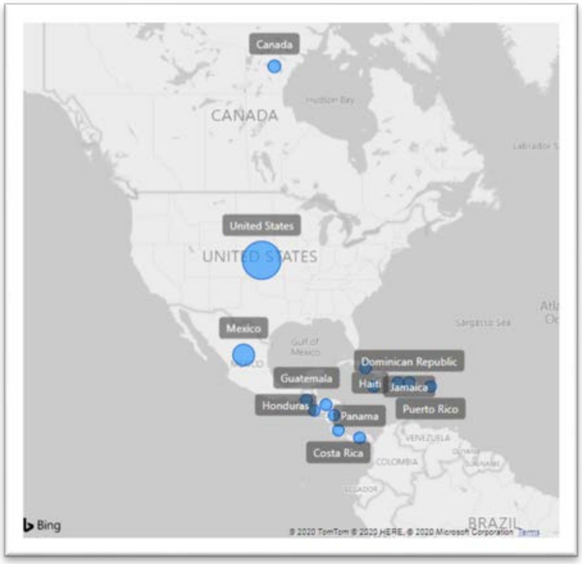


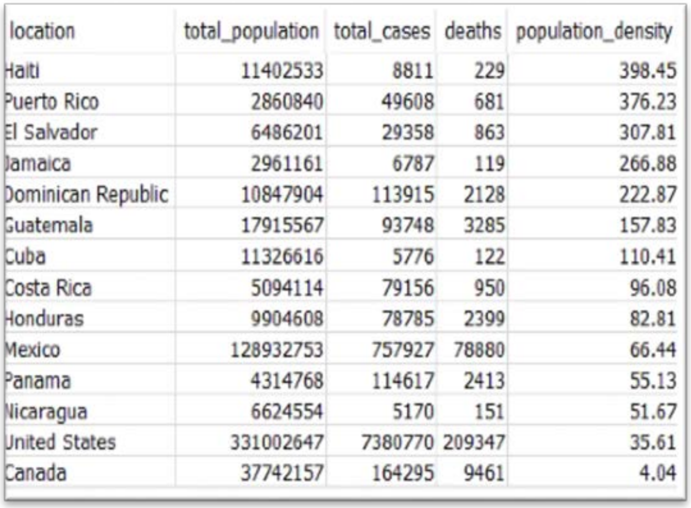
location	total_population	total_cases	deaths
Puerto Rico	2860840	49608	681
Jamaica	2961161	6787	119
Panama	4314768	114617	2413
Costa Rica	5094114	79156	950
El Salvador	6486201	29358	863
Nicaragua	6624554	5170	151
Honduras	9904608	78785	2399
Dominican Republic	10847904	113915	2128
Cuba	11326616	5776	122
Haiti	11402533	8811	229
Guatemala	17915567	93748	3285
Canada	37742157	164295	9461
Mexico	128932753	757927	78880
United States	331002647	7380770	209347

3. We then decided to compare population density in North America and total cases in each country.

   Haiti had the highest **population density** of any nation in North America, at 398.45, and yet they were able to minimize the spread of covid-19. A proper response such as communication and lockdown, closing off its borders early on, and by identifying people infected by the virus early on helped.

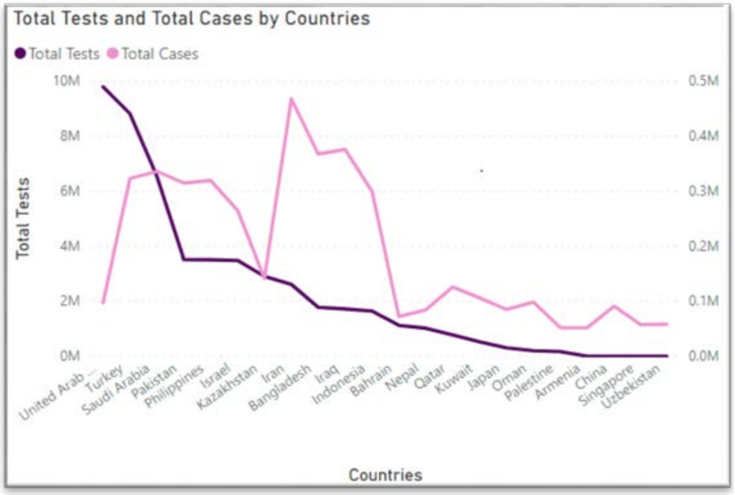
   Canada has the **lowest density** of any country in North America, 4.04, however, we considered it to be an outlier because almost half of the country's population is concentrated in three cities i.e. Vancouver, Toronto, and Montreal, thus making it harder to have a tangible data.

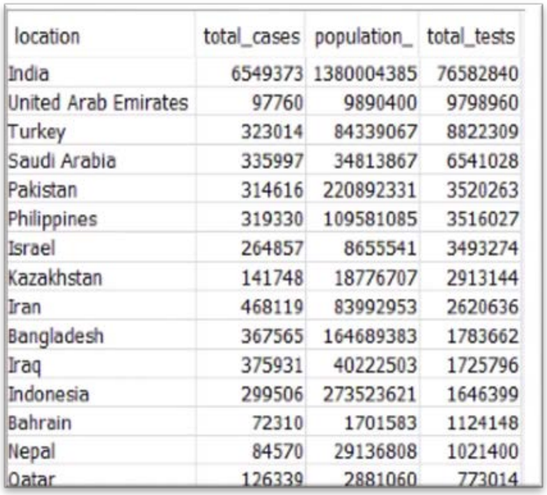




location	total_population	total_cases	deaths	population_density
Haiti	11402533	8811	229	398.45
Puerto Rico	2860840	49608	681	376.23
El Salvador	6486201	29358	863	307.81
Jamaica	2961161	6787	119	266.88
Dominican Republic	10847904	113915	2128	222.87
Guatemala	17915567	93748	3285	157.83
Cuba	11326616	5776	122	110.41
Costa Rica	5094114	79156	950	96.08
Honduras	9904608	78785	2399	82.81
Mexico	128932753	757927	78880	66.44
Panama	4314768	114617	2413	55.13
Nicaragua	6624554	5170	151	51.67
United States	331002647	7380770	209347	35.61
Canada	37742157	164295	9461	4.04

4. We then decided to find a relation between tests and cases in Asia.
UAE was second in **total tests** done across the country at **9 million tests** ever since the pandemic began. Due to quick and efficient testing, total cases in UAE were less than 100,000 for a population of more than 10 million. For a population of around 10 million, 9 million tests were the major contributing factor in subsiding the virus spread. The country's crackdown on censorship helps stifle the spread of Covid 19 and implementing procedures for testing and healthcare. The country closed its border as early as possible and enforced rules for social distancing and sanitation.

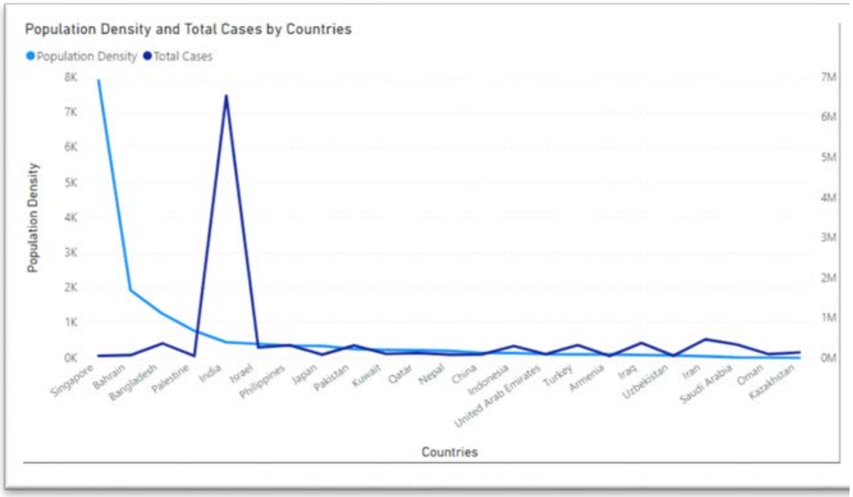


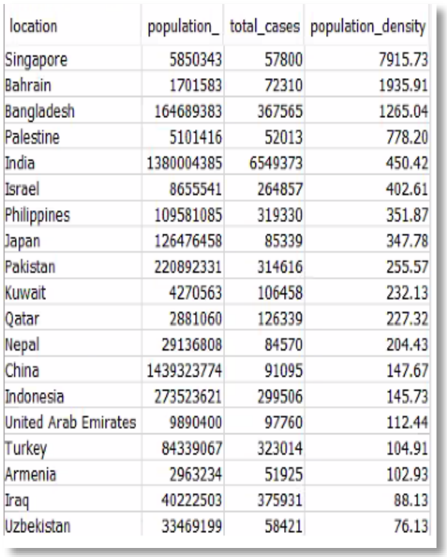


location	total_cases	population_	total_tests
India	6549373	1380004385	76582840
United Arab Emirates	97760	9890400	9798960
Turkey	323014	84339067	8822309
Saudi Arabia	335997	34813867	6541028
Pakistan	314616	220892331	3520263
Philippines	319330	109581085	3516027
Israel	264857	8655541	3493274
Kazakhstan	141748	18776707	2913144
Iran	468119	83992953	2620636
Bangladesh	367565	164689383	1783662
Iraq	375931	40222503	1725796
Indonesia	299506	273523621	1646399
Bahrain	72310	1701583	1124148
Nepal	84570	29136808	1021400
Qatar	126339	2881060	773014

5. After narrowing down the most affected continents, we wanted to check whether **population density** played a role in the spread of Covid-19. Our results varied.
Singapore has the highest **population dens**ity of any country in Asia, at 7915, yet due to targeted and timely measures and proper communication, they were able to limit the country's cases. The Singaporean government implemented plans to cover the costs of testing and treatment, scaling screening, and tests for every citizen. Singapore's aggressive contract tracing allowed the country to quickly, identify, and isolate new cases due to strict quarantine orders.
Kazakhstan on the other hand has the lowest **population density**, yet due to negligence had three times the cases like that of Singapore. The country's slow response to a deadly pneumonia outbreak was

detrimental to the population, coupled with the country lifting its lockdown early, which led to an increase in its total cases.

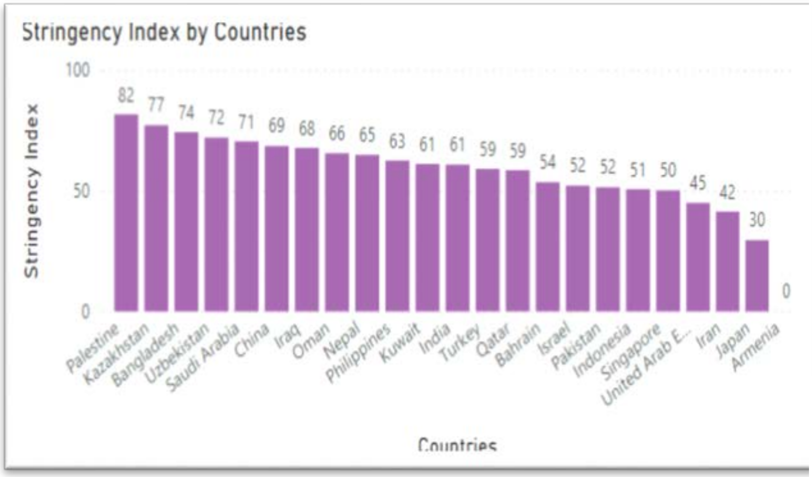


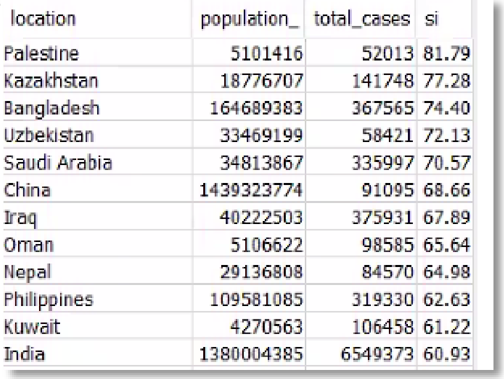


location	population_	total_cases	population_density
Singapore	5850343	57800	7915.73
Bahrain	1701583	72310	1935.91
Bangladesh	164689383	367565	1265.04
Palestine	5101416	52013	778.20
India	1380004385	6549373	450.42
Israel	8655541	264857	402.61
Philippines	109581085	319330	351.87
Japan	126476458	85339	347.78
Pakistan	220892331	314616	255.57
Kuwait	4270563	106458	232.13
Qatar	2881060	126339	227.32
Nepal	29136808	84570	204.43
China	1439323774	91095	147.67
Indonesia	273523621	299506	145.73
United Arab Emirates	9890400	97760	112.44
Turkey	84339067	323014	104.91
Armenia	2963234	51925	102.93
Iraq	40222503	375931	88.13
Uzbekistan	33469199	58421	76.13

6. We then decided to focus on Stringency Index. Stringency Index is a composite measure, rescaled to a value from 0 to 100 (100 being the strictest, based on nine response indicators including school closures, workplace closures, and travel bans.

   Palestine had the highest **stringency index.** As a result, the country, with a population of over 50 million, had a total of 52,000 cases, one of the lowest in Asia. The country's timely response to the pandemic and Palestine's strategic communication with regular briefings to the population was key in maintaining control over the spread of the virus.



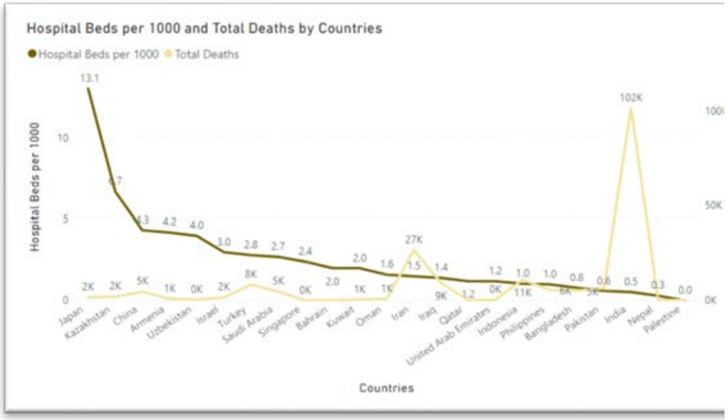


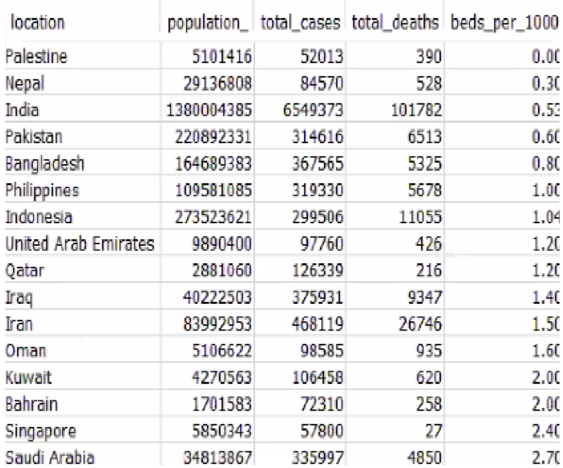
location	population_	total_cases	si
Palestine	5101416	52013	81.79
Kazakhstan	18776707	141748	77.28
Bangladesh	164689383	367565	74.40
Uzbekistan	33469199	58421	72.13
Saudi Arabia	34813867	335997	70.57
China	1439323774	91095	68.66
Iraq	40222503	375931	67.89
Oman	5106622	98585	65.64
Nepal	29136808	84570	64.98
Philippines	109581085	319330	62.63
Kuwait	4270563	106458	61.22
India	1380004385	6549373	60.93

7. Next, we compared the average number of hospital beds across a country and death across countries that were in Asia.

   India had the highest **death** of any country in Asia. We believe that scarce hospital beds, at 0.53 hospital beds per thousand, highly contributed to this factor as India ranked third last in average hospital beds. With a large population and ever-increasing cases in the country, patients don't have proper access to the necessary treatments. Testing can take days and hospitals are full. Weak health infrastructure and mismanagement was the main reason behind the sharp rise in deaths in India.

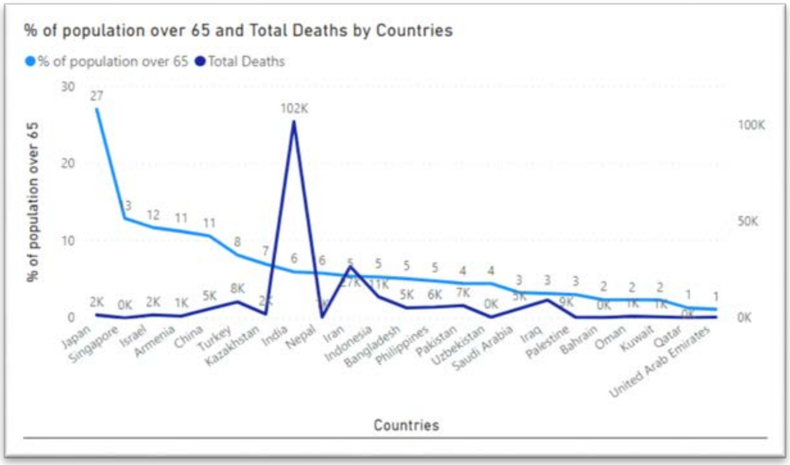
   Japan on the other hand had the **highest number of beds** in Asia, hitting numbers as high as 13.05 beds thousand. As a result, deaths have been far lower than initially predicted. Even though Japan has a population of 126 million, the total cases in Japan were just above 85000.

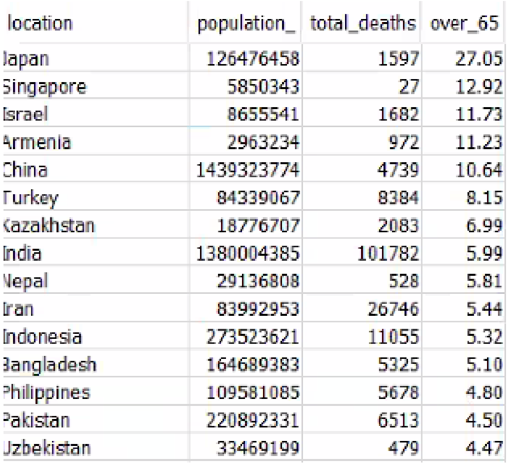




location	population_	total_cases	total_deaths	beds_per_1000
Palestine	5101416	52013	390	0.0
Nepal	29136808	84570	528	0.3
India	1380004385	6549373	101782	0.5
Pakistan	220892331	314616	6513	0.6
Bangladesh	164689383	367565	5325	0.8
Philippines	109581085	319330	5678	1.0
Indonesia	273523621	299506	11055	1.04
United Arab Emirates	9890400	97760	426	1.2
Qatar	2881060	126339	216	1.2
Iraq	40222503	375931	9347	1.4
Iran	83992953	468119	26746	1.5
Oman	5106622	98585	935	1.6
Kuwait	4270563	106458	620	2.0
Bahrain	1701583	72310	258	2.0
Singapore	5850343	57800	27	2.4
Saudi Arabia	34813867	335997	4850	2.7

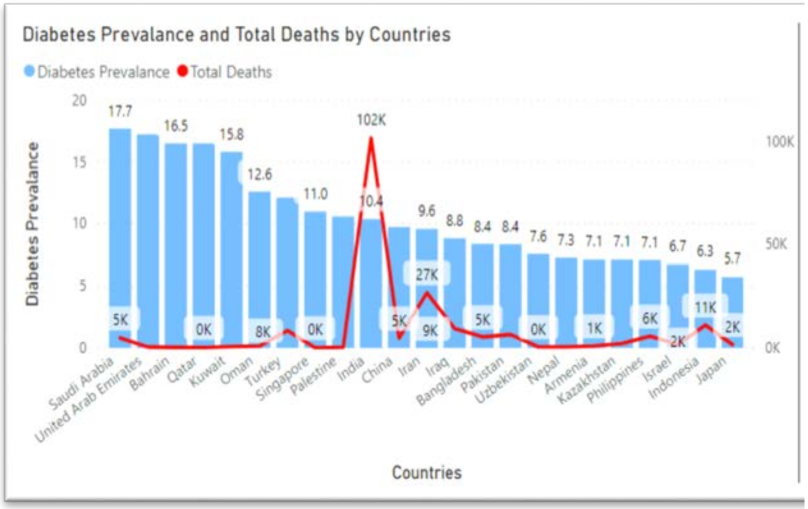
8. In our next query, we compared each country’s average population of 65 and older to total deaths in each country.
   Japan hads more elderly per capita (**over 1/3rd of the population is older than 65**) than any other country in the world but it had a far lower death rate when compared to other countries in Asia. Due to the country’s timely response such as early and timely lockdown procedures and the government enforcing its people to stay at home, Japan saw few excess cases.
   The country drastically reduced the virus’s transmission and coupled with Japan’s excellent medical infrastructure; the death rate was as low as possible. Also, according to studies, there is a possibility that the initial SARS pandemic might have strengthened the east Asian population’s immunity to the coronavirus.

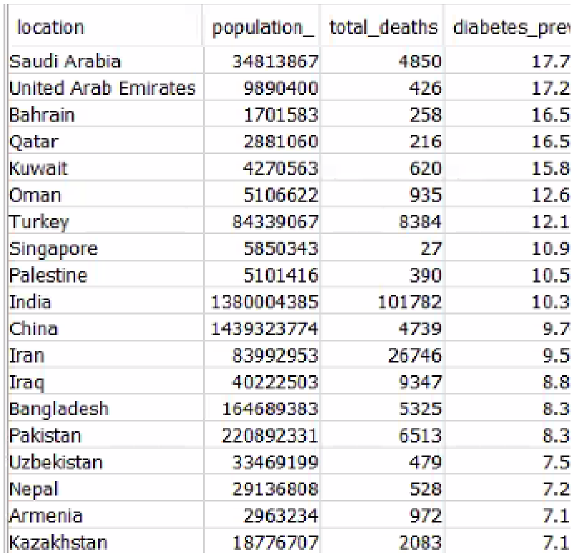




location	population_	total_deaths	over_65
Japan	126476458	1597	27.05
Singapore	5850343	27	12.92
Israel	8655541	1682	11.73
Armenia	2963234	972	11.23
China	1439323774	4739	10.64
Turkey	84339067	8384	8.15
Kazakhstan	18776707	2083	6.99
India	1380004385	101782	5.99
Nepal	29136808	528	5.81
Iran	83992953	26746	5.44
Indonesia	273523621	11055	5.32
Bangladesh	164689383	5325	5.10
Philippines	109581085	5678	4.80
Pakistan	220892331	6513	4.50
Uzbekistan	33469199	479	4.47

9. We then decided to find a correlation between diabetes prevalence and total death.
   Middle eastern countries, like **Saudi Arabia, Qatar, Bahrain**, had the highest **prevalence of diabetes** in any country. Yet the total death was far lower than in other countries. Our study indicated that the prevalence of diabetes among patients affected by COVID-19 is not higher than that observed in the general population, thus suggesting that diabetes is not a risk factor for SARS-CoV-2 infection.
   Middle eastern countries, even though they have the highest diabetes prevalence in Asia, were able to control Covid effectively due to tighter lockdown procedures and rapid testing. By closing down their borders early, the Middle East was able to isolate its cases. Saudi Arabia, for example, closed down its border early in the pandemic and restricted its yearly pilgrimage to the locals only.





location	population_	total_deaths	diabetes_pre
Saudi Arabia	34813867	4850	17.7
United Arab Emirates	9890400	426	17.2
Bahrain	1701583	258	16.5
Qatar	2881060	216	16.5
Kuwait	4270563	620	15.8
Oman	5106622	935	12.6
Turkey	84339067	8384	12.1
Singapore	5850343	27	10.9
Palestine	5101416	390	10.5
India	1380004385	101782	10.3
China	1439323774	4739	9.7
Iran	83992953	26746	9.5
Iraq	40222503	9347	8.8
Bangladesh	164689383	5325	8.3
Pakistan	220892331	6513	8.3
Uzbekistan	33469199	479	7.5
Nepal	29136808	528	7.2
Armenia	2963234	972	7.1
Kazakhstan	18776707	2083	7.1

Comparison of DBMS used

Query	HIVE APACHE HBASE	PostgreSQL
Describe covid;	.184 seconds	4.3 seconds
select * from covid where pk = "21195";	.8 seconds	.146 seconds
SELECT continent, SUM(DISTINCT population)AS total_population,SUM(new_cases) as total_cases, sum(new_deaths) as deaths FROM covid GROUP BY continent ORDER BY total_population ASC;	17 seconds	.18 seconds

For most of our queries, PostgreSQL had better performance (speed) than HBase. This is because all the data is stored on a single storage device and processed on a single processor in Postgres. HBase leverages a distributed architecture for storage and the MapReduce framework for processing. Although our dataset contained over 50,000 rows, it is not considered real big data. HBase's performance gains are only experienced when the dataset is so large, it cannot be stored on a single disk, and cannot be processed on a single computer, thus making the comparison with PostgreSQL impossible in the context of big data.

CHALLENGES:

Problem

- When creating the external table in Hive, we could not use ‘date’ as an attribute name because it is used as a data type in Hive

Solution

- Changed from ‘date’ to ‘date_’ and the import command ran successfully

Problem

- Importing the data into PostgreSQL was a hassle.

Solution

- The data had a lot of null values, which we had to manually eliminate else it was affecting our analysis.

Problem

- We had to deal with a lot of noise, such as redundancy, null values

Solution

- We smoothed out the data to get a proper estimation of our attributes.

Problem

- Difficulty in finding the right attributes

Solution

- We had to run queries for almost every attribute, all 40 of them, and then select the ones that had a better correlation and answered our questions.

Problem

- Several countries had apostrophes and accents in their names.

Solution

- Instead of thinking of them as characters, Postgres was instead using them as codes. For example, Côte d'Ivoire.

CONCLUSION:

- We emphasized our study on countries located in the Asian and North American continent. These two continents were the worst hit by the pandemic and we were able to gather more relevant data.
- We wanted to find out why some countries performed much better than others, even though they shared the same borders.
- Some countries were more densely populated than others, like Singapore, yet due to their efficiency and properly following protocol they were able to minimize the spread of Covid-19. This showed us that a competent government to implement procedures and enforce policies can go a long way in combatting the spread.
- Stringency was another attribute that played a large role in subsiding the transmission of the coronavirus. Palestine, a war-torn country, did a much better job than others. The government can impact the spread of the virus.
- However, the government isn’t just enough. People need to follow the rules. Citizens living in the Middle East and East Asian countries got tested as soon as the virus began. They followed the quarantine and lockdown procedures and hence minimized the spread of Covid-19.
- Cuba and Japan have been a shining example of what proper and universal healthcare and advanced medical facilities can achieve. Both countries have done a great job in containing the virus.
- Just like every DBMS we have learned about so far, HBase is not superior or inferior to any other. HBase can perform the same functions and queries like any other technology. The value in HBase resides in its ability to handle big data through HDFS and MapReduce. With that said, this project gave us good exposure to HBase and Hive but did not necessarily improve the ETL or analysis process of our 50,000 rows (9MB) dataset.