Create A Research Dataset

Introduction

The late 2019 COVID-19 epidemic has had a significant effect on cultures all across the world. From January 2020 to December 2022, a period of three years, people around the world struggled to understand the intricate interactions between many circumstances that caused the occurrence of COVID-19. The number of cases, severity, and reactions to the virus during this time period varied drastically by area, and the reasons for these differences are complex. This study dives into the economical, political, and other influencing elements that have been crucial in deciding whether particular regions have had higher or fewer rates of COVID-19. Thus, the topic chosen to create a research dataset is to examine socioeconomic, political or other factors that contributed to either higher or lower incidences of covid 19 between Jan 2020 and Dec 2022.

Recording Video:
https://drive.google.com/file/d/1lgsFoGvdQqkfn6DYJnZsdPq8Y3B66Wep/view?usp=share_link

Dataset Selection

To create a research dataset the following datasets have been selected:

1. COVID19 vaccine rate in the US by State.
2. COIVD19 cases surveillance by state
3. Facility-level data for hospital utilization aggregated on a weekly basis
4. Testing of COVID 19 and outcomes for US States
5. GDP in the US by state from 2020 to 2022
6. Population in the US by state from 2020 to 2022

Google Drive Link:
https://docs.google.com/spreadsheets/d/1nB3aztK1eVvrapERsmrmpzyVG17tpOTK_RibstPgtXE/edit#gid=0

From all the above data sets, the columns of choice are appended into a final research dataset which could be further analyzed to fetch any kind of information or perform research to derive answers.

The process of integration

The "state_testing_and_outcome" dataset, "COVID-19 cases surveillance" dataset, and "COVID-19 vaccine rate" dataset all include daily data that were recorded. Additionally, weekly data is kept in the file "Facility-level data for hospital utilization".

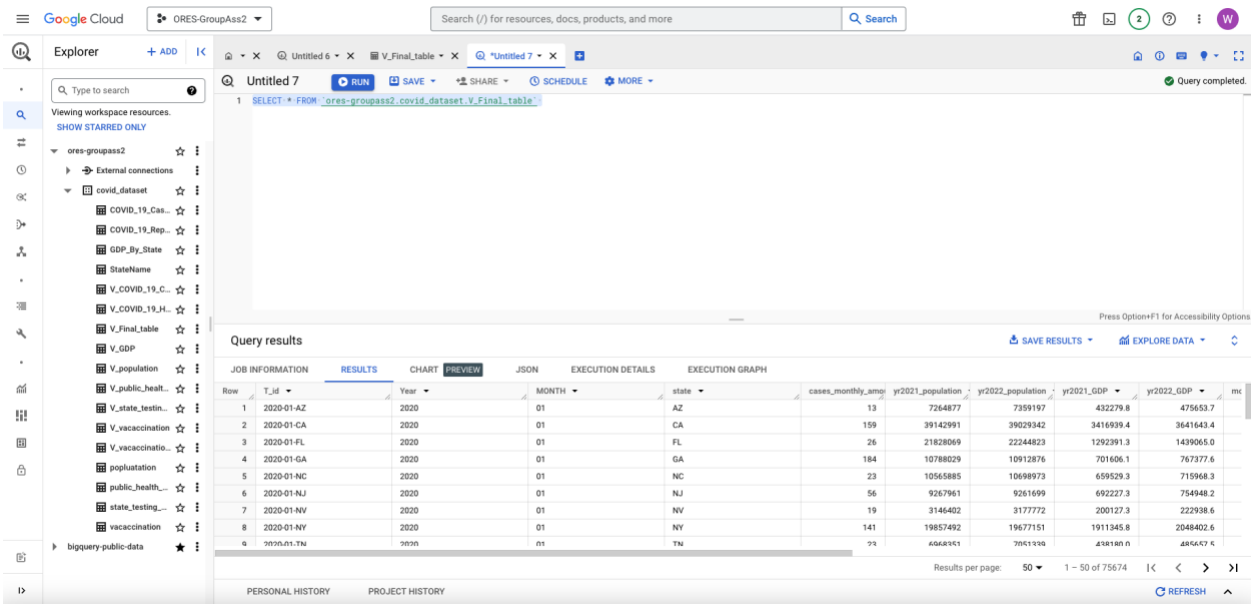
However, yearly statistics were kept on "GDP in the US by state" and "Population in the US by state" As a result, the initial step is to as closely as possible match all data with the month dimension of time. The "state_testing_and_outcome", "COVID-19 cases surveillance", and "COVID-19 vaccine rate" datasets are combined by month. The "Population in the US by state" and "GDP in the US by state" data sets were shown by month. As a result, all of the tables could be pooled.

The second stage involves making every state name column in every dataset the same. As a result, we choose to convert all state fields for full names to abbreviation states. For instance, all datasets would change "Missouri" to "MO". We created a mapping table called "StateName" in this stage. It would be used to shorten the complete name of the state to its acronym.

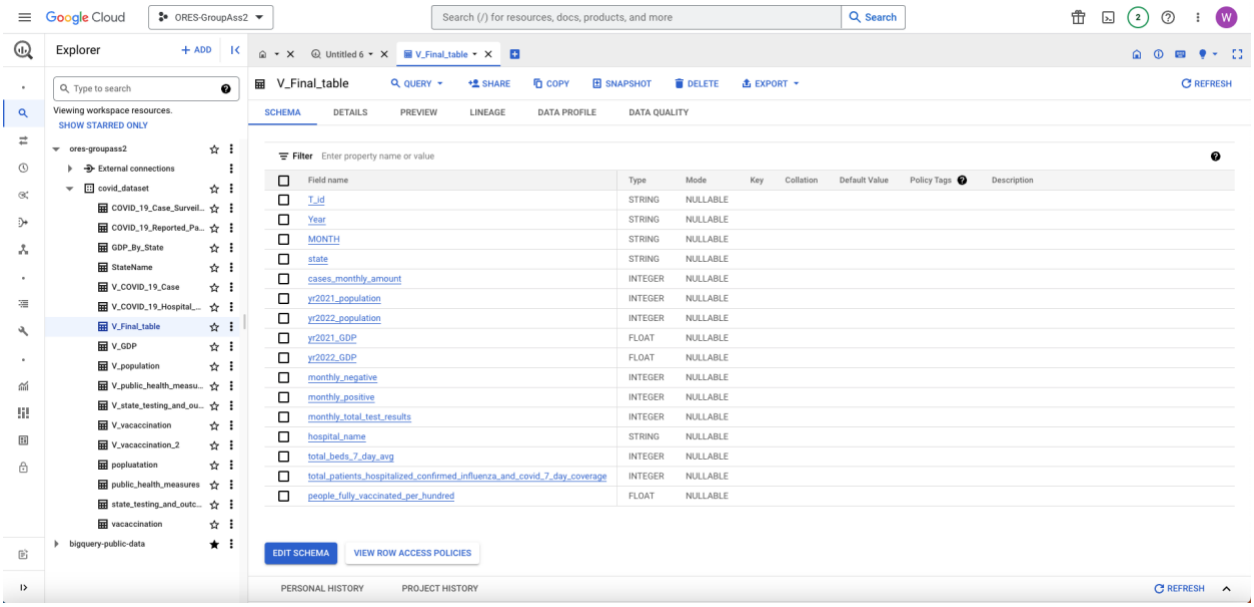
Lastly, according to the number of records in each dataset, we decided “COIVD19 cases surveillance” as the left table, and other datasets would be retrieved information to this left one. Below is the process of combination to final table.

```sql
create table `ores-groupass2.covid_dataset.V_Final_table`
as

SELECT distinct (CONCAT(cast (T2.Year as string),'-', cast (T2.Month as string),'-' ,T2.state)) As T_id,
T2.Year, T2.MONTH, T2.state, sum(T2.case_total_count) AS cases_monthly_amount,
T6.yr2021 AS yr2021_population, T6.yr2022 AS yr2022_population, --T1.total_beds_7_day_avg,
T7.yr2021_GDP, T7.yr2022_GDP,T3.monthly_negative, T3.monthly_positive, T3.monthly_total_test_results,
T5.hospital_name,T5.total_beds_7_day_avg,T5.total_patients_hospitalized_confirmed_influenza_and_covid_7_day_coverage,
T1.people_fully_vaccinated_per_hundred

FROM
`ores-groupass2.covid_dataset.V_COVID_19_Case`  AS T2
left join `ores-groupass2.covid_dataset.V_vacaccination_2` AS T1
ON cast(T1.Year as string)  = T2.Year
AND cast( T1.MONTH as string) = T2.MONTH
AND T1.Postal =T2.state
left join `ores-groupass2.covid_dataset.V_state_testing_and_outcomes` AS T3
ON T2.Year =  cast (T3.year as string)
AND T2.MONTH =cast (T3.month  as string)
AND T2.state =T3.state
/*
Response too large to return. Consider specifying a destination table in your job configuration. For more details
*/
left join `ores-groupass2.covid_dataset.V_COVID_19_Hospital_Capacity_by_Facility` AS T5
ON T2.Year =  cast (T5.year as string)
AND T2.MONTH =cast (T5.month  as string)
AND T2.state =T5.state
left join `ores-groupass2.covid_dataset.V_population` AS T6
ON T2.state = T6.postal
left join `ores-groupass2.covid_dataset.V_GDP` AS T7
ON T2.state = T7.Postal
--where T2.age_group != '65+ years' AND monthly_negative is not null
group by T2.Year, T2.MONTH, T2.state, T6.yr2021, T6.yr2022 , --T1.total_beds_7_day_avg,
T7.yr2021_GDP, T7.yr2022_GDP,T3.monthly_negative, T3.monthly_positive, T3.monthly_total_test_results,
T5.hospital_name,T5.total_beds_7_day_avg,T5.total_patients_hospitalized_confirmed_influenza_and_covid_7_day_coverage,
T1.people_fully_vaccinated_per_hundred
```

There is the result of the final table. It shows 75674 number of rows.



There is the schema and detail of the final table.



Summary

This final dataset provides information about covid19 case surveillance, testing and outcomes of the disease, hospital facilities, vaccination, and GDP of the states. This dataset provides an opportunity to researchers, public healthcare officials to obtain valuable insights that can be helpful in knowing a few important questions like highest number of testing encounters by state, medical and hospital facilities by state, total population and number of people vaccinated and their effects, economic status etc..

References

BEA Interactive Data Application. (n.d.). https://apps.bea.gov/iTable/?reqid=70&step=1&acrdn=2#eyJhcHBpZCI6NzAsInN0ZXBzIjpbMSwyOSwyNSwzMSwyNiwyNywzMF0sImRhdGEiOltbIlRhYmxlSWQiLCI5Il0sWyJTdGF0ZSIsWyIwIl1dLFsiQXJlYSIsWyJYWCJdXSxbIlN0YXRpc3RpYyIsWyIxIl1dLFsiVW5pdF9vZl9tZWFzdXJlIiwiTGV2ZWxzIl0sWyJZZWFyIixbIjIwMjIiLCIyMDIxIiwiMjAyMCJdXSxbIllYXJCZWdpbiIsIi0xIl0sWyJZZWFyX0VuZCIsIi0xIl1dfQ==

COVID-19 Case Surveillance Public Use Data with Geography | Data | Centers for Disease Control and Prevention. (2023, October 6). https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4

COVID-19 reported patient impact and hospital capacity by facility | HealthData.gov. (n.d.). https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/anag-cw7u/explore

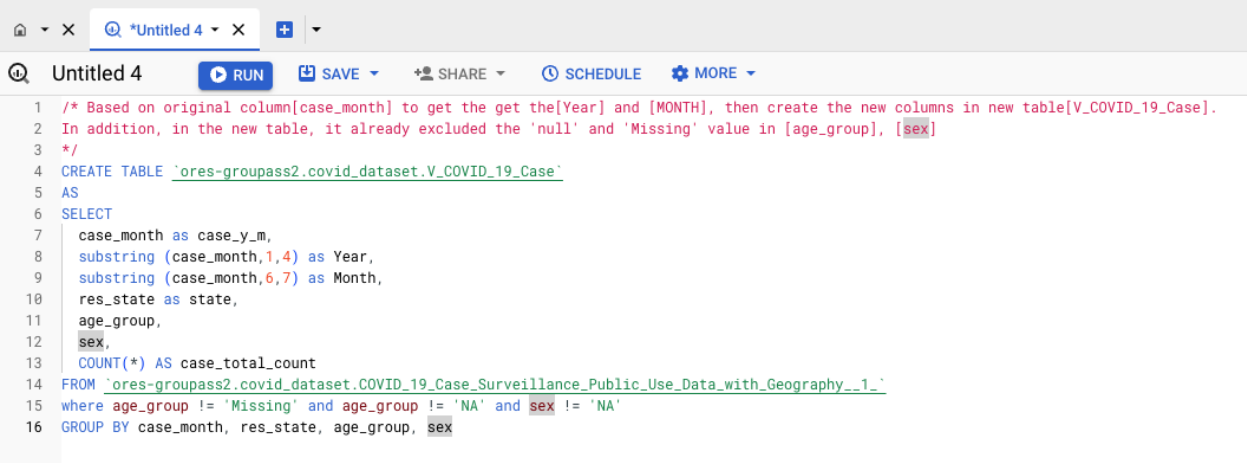
Google Cloud console. (n.d.). https://console.cloud.google.com/marketplace/product/bigquery-public-datasets/covid19-public-data-program?project=keerthana-bq-slu&pli=1

USA COVID-19 vaccinations. (2023, September 19). Kaggle. https://www.kaggle.com/datasets/paultimothymooney/usa-covid19-vaccinations/

US Census Bureau. (2023, June 13). State Population Totals and Components of Change: 2020-2022. Census.gov. https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html

Appendix

I. Process of adjusting dataset" COIVD19 cases surveillance" and insert to new table " V_COVID_19_Case"




```sql
/* Based on original column[case_month] to get the get the[Year] and [MONTH], then create the new columns in new table[V_COVID_19_Case].
In addition, in the new table, it already excluded the 'null' and 'Missing' value in [age_group], [sex]
*/
CREATE TABLE `ores-groupass2.covid_dataset.V_COVID_19_Case`
AS
SELECT
  case_month as case_y_m,
  substring (case_month,1,4) as Year,
  substring (case_month,6,7) as Month,
  res_state as state,
  age_group,
  sex,
  COUNT(*) AS case_total_count
FROM `ores-groupass2.covid_dataset.COVID_19_Case_Surveillance_Public_Use_Data_with_Geography__1_`
where age_group != 'Missing' and age_group != 'NA' and sex != 'NA'
GROUP BY case_month, res_state, age_group, sex
```

II. Process of adjusting dataset" state_testing_and_outcomes" and insert to new table " V_state_testing_and_outcomes"

```sql
/*Based on original column[date] to get the get the[year] and [month], then create the new columns in new table[V_state_testing_and_outcomes].
In addition, in the new table, it had already aggregate [death],[total_test_results], [negative] ,[positive],[positive_tests_antibody] by month
*/
CREATE TABLE `ores-groupass2.covid_dataset.V_state_testing_and_outcomes`
AS

SELECT
  EXTRACT(YEAR FROM date) AS year,
  EXTRACT(MONTH FROM date) AS month,
  state,
  SUM(death) AS mothly_total_dealth,
  SUM(total_test_results) AS monthly_total_test_results,
  SUM(negative) AS monthly_negative,
  SUM(positive) AS monthly_positive,
  SUM(positive_tests_antibody) AS monthly_positive_tests_antibody
FROM `ores-groupass2.covid_dataset.state_testing_and_outcomes`
GROUP BY year, month, state
ORDER BY year, month, state;


---drop table `ores-groupass2.covid_dataset.V_state_testing_and_outcomes`
```

III. Process of adjusting dataset vacaccination" and insert to new table " V_vacaccination"

```sql
/* Explain: part1: use CTE to get [year], [month], [day] column
Based on original column[date] to get the get the[year],[month],and [day], then create the new columns in new table[V_state_testing_and_outcomes]. In addition, in the new table, it had already aggregate
[death],[total_test_results], [negative] ,[positive],[positive_tests_antibody] by month
*/
--part2: use part1 result to cacluate the last day of each month and mark it as 'Last Day', because it can shows the monthly data rather than daily data.
CREATE TABLE `ores-groupass2.covid_dataset.V_vacaccination`

AS
WITH VaccinationData AS (
  SELECT
    V.date,
    EXTRACT(YEAR FROM V.date) AS year,
    EXTRACT(MONTH FROM V.date) AS month,
    EXTRACT(DAY FROM V.date) AS day,
    V.location,
    V.people_fully_vaccinated_per_hundred,
    V.total_vaccinations_per_hundred,
    V.total_boosters_per_hundred,
    ROW_NUMBER() OVER (PARTITION BY EXTRACT(YEAR FROM V.date), EXTRACT(MONTH FROM V.date) ORDER BY V.date DESC) AS row_num
  FROM `ores-groupass2.covid_dataset.vacaccination` AS V
)
SELECT
  date,
  year,
  month,

  CASE
    WHEN row_num = 1 THEN 'Last Day'
    ELSE 'Not Last Day'
  END AS day_type,

  --location,
  T2.Postal,
  people_fully_vaccinated_per_hundred,
  total_vaccinations_per_hundred,
  total_boosters_per_hundred
FROM VaccinationData AS T1
RIGHT JOIN `ores-groupass2.covid_dataset.StateName` AS T2
```

IV. Process of adjusting dataset COVID_19_Hospital_Capacity_by_Facility" and insert to new table "V_COVID_19_Reported_Patient_Impact_and_Hospital_Capacity_by_Facility"

```sql
-- Based on [collection_week] to get the get the[Year] and [MONTH], then create the new columns in new table[V_COVID_19_Hospital_Capacity_by_Facility]

Create table `ores-groupass2.covid_dataset.V_COVID_19_Hospital_Capacity_by_Facility`
AS
  SELECT K.collection_week,
  cast (EXTRACT ( YEAR FROM K.collection_week) as string) as Year  ,
  cast (EXTRACT ( MONTH FROM K.collection_week) as string) as MONTH
  , K.state, K.hospital_name, K.total_beds_7_day_avg, K.total_patients_hospitalized_confirmed_influenza_and_covid_7_day_coverage
  FROM `ores-groupass2.covid_dataset.COVID_19_Reported_Patient_Impact_and_Hospital_Capacity_by_Facility` AS K
```

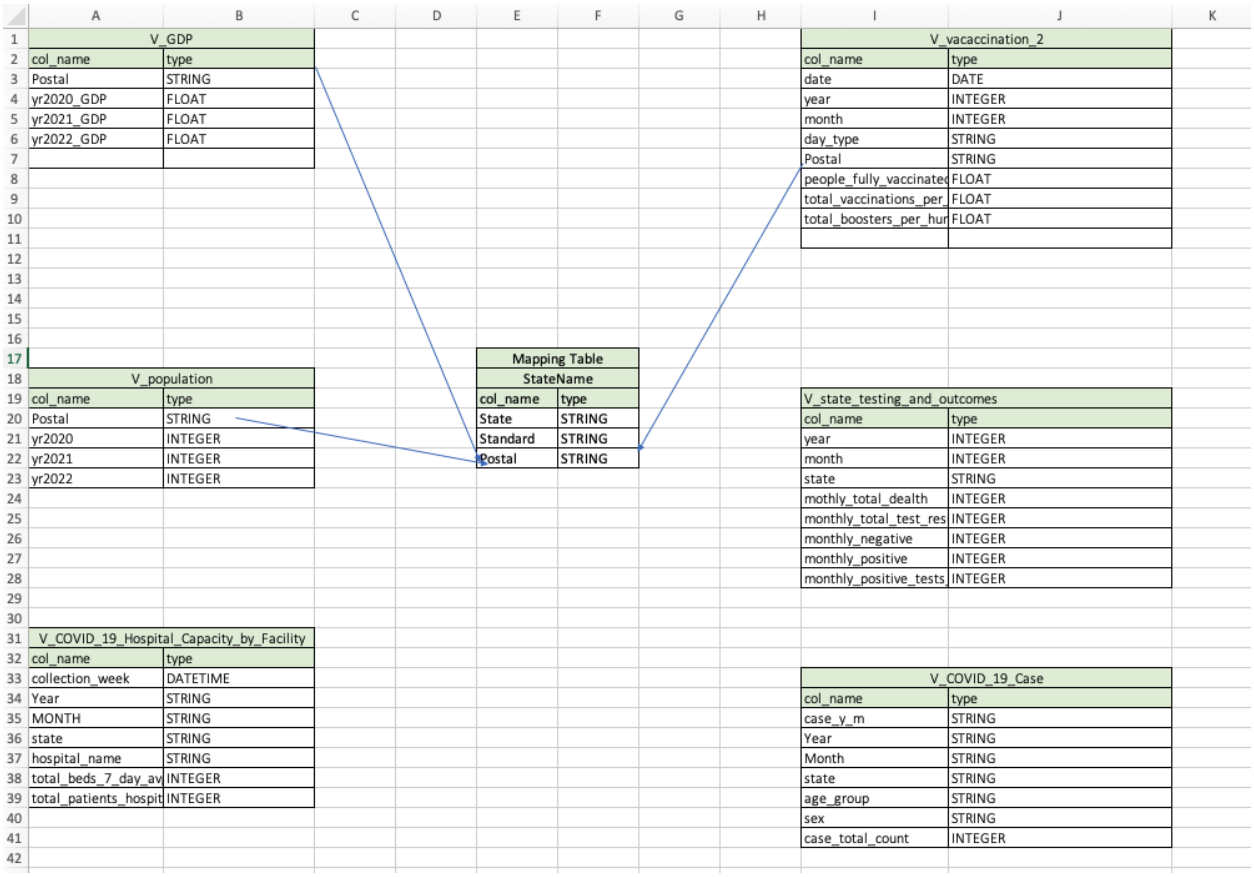
V. Process of adjusting dataset "popluatation " and insert to new table " V_population"

```sql
---  to convert original table [location] to state abbreviations [Postal] by using mapping table [StateName], then create it as the new table [V_population]

CREATE TABLE `ores-groupass2.covid_dataset.V_population`
AS
SELECT B.Postal, A.yr2020,  A.yr2021, A.yr2022
FROM `ores-groupass2.covid_dataset.popluatation` AS A
join  `ores-groupass2.covid_dataset.StateName`  AS B
On A.locatation = B.State
```

VI. Process of adjusting dataset "GDP_By_State " and insert to new table " V_GDP"

```sql
---  to convert original table [GeoName ] to state abbreviations [Postal] by using mapping table [StateName], then create it as the new table [V_GDP]

Create Table `ores-groupass2.covid_dataset.V_GDP`
AS
SELECT B.Postal, A.yr2020_GDP, A.yr2021_GDP, A.yr2022_GDP
FROM `ores-groupass2.covid_dataset.GDP_By_State`  AS A
join `ores-groupass2.covid_dataset.StateName` AS B
ON A.GeoName = B.State
--LIMIT 1000
```

VII. State Name Mapping Diagram



V_GDP	
col_name	type
Postal	STRING
yr2020_GDP	FLOAT
yr2021_GDP	FLOAT
yr2022_GDP	FLOAT

V_vacaccination_2	
col_name	type
date	DATE
year	INTEGER
month	INTEGER
day_type	STRING
Postal	STRING
people_fully_vaccinated	FLOAT
total_vaccinations_per_	FLOAT
total_boosters_per_hur	FLOAT

Mapping Table	
StateName	
col_name	type
State	STRING
Standard	STRING
Postal	STRING

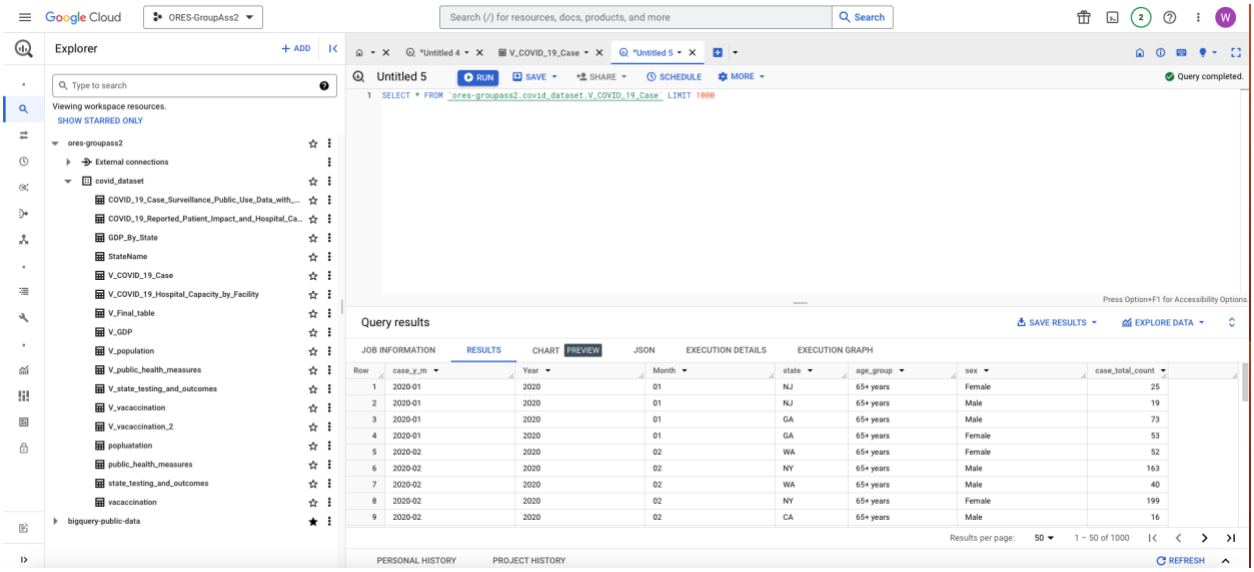
V_population	
col_name	type
Postal	STRING
yr2020	INTEGER
yr2021	INTEGER
yr2022	INTEGER

V_state_testing_and_outcomes	
col_name	type
year	INTEGER
month	INTEGER
state	STRING
mothly_total_dealth	INTEGER
monthly_total_test_res	INTEGER
monthly_negative	INTEGER
monthly_positive	INTEGER
monthly_positive_tests	INTEGER

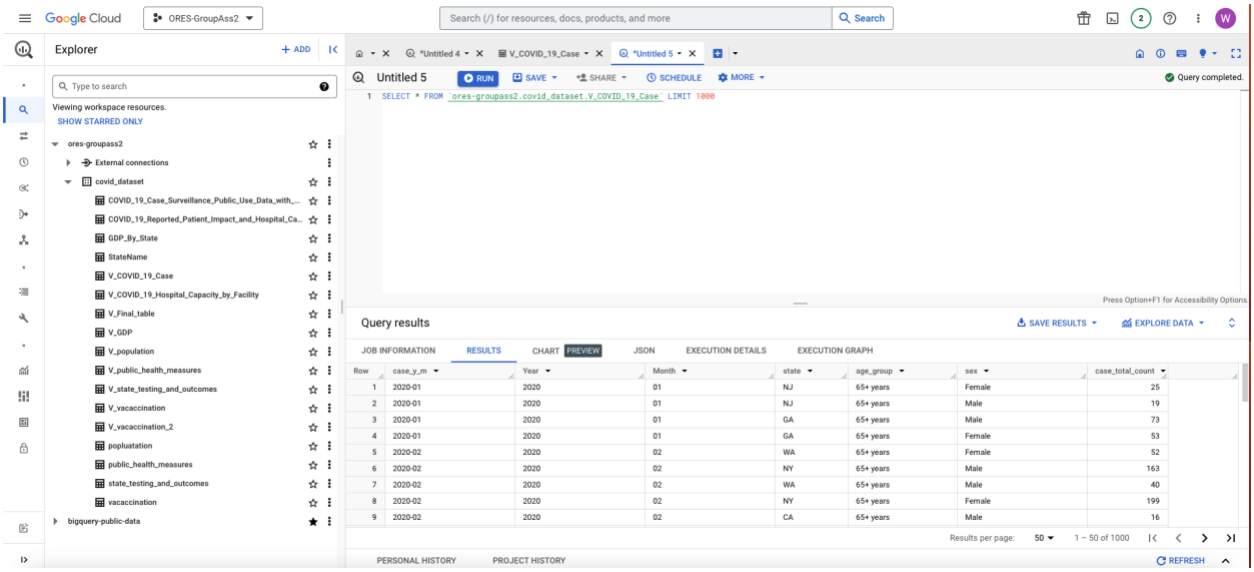
V_COVID_19_Hospital_Capacity_by_Facility	
col_name	type
collection_week	DATETIME
Year	STRING
MONTH	STRING
state	STRING
hospital_name	STRING
total_beds_7_day_av	INTEGER
total_patients_hospit	INTEGER

V_COVID_19_Case	
col_name	type
case_y_m	STRING
Year	STRING
Month	STRING
state	STRING
age_group	STRING
sex	STRING
case_total_count	INTEGER

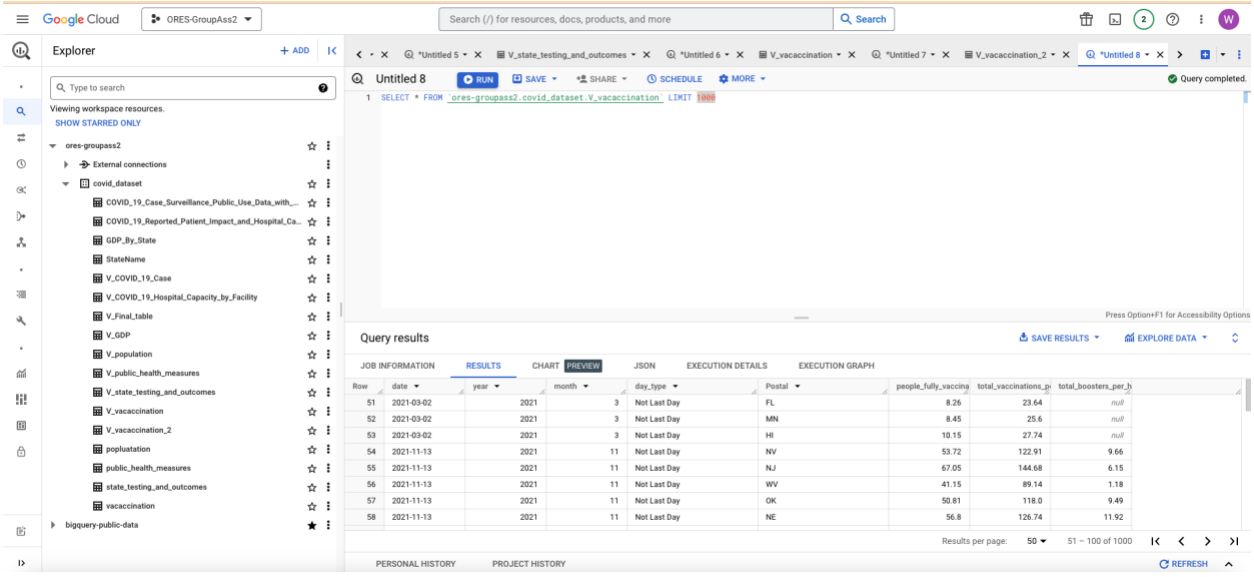
VIII. Result of table “ V_COVID_19_Case”



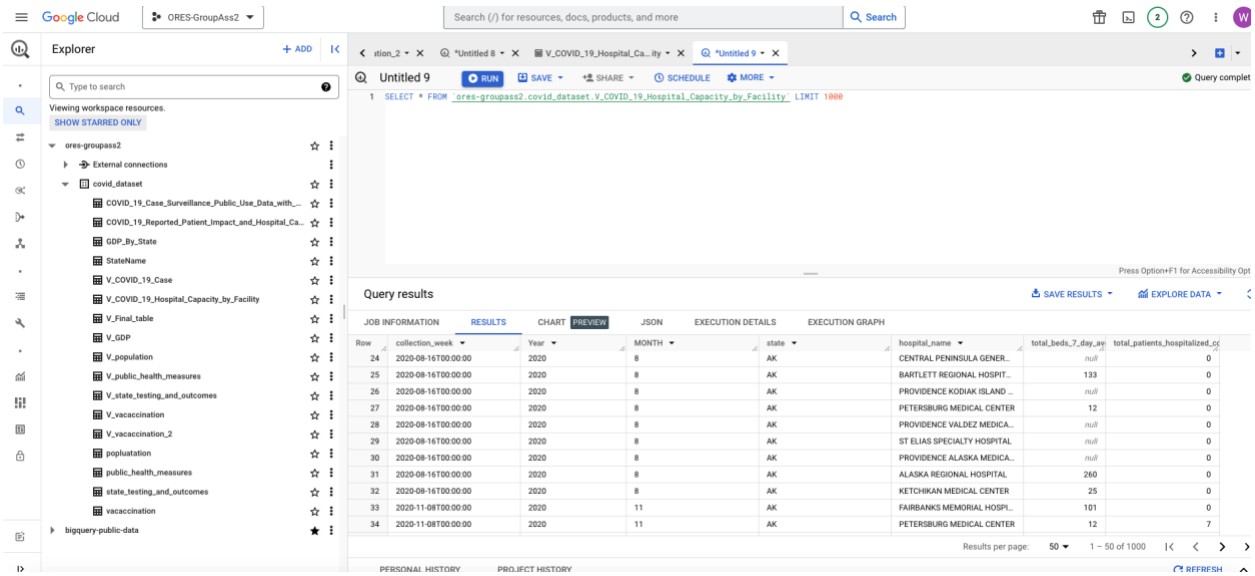
IX. Result of table “V_state_testing_and_outcomes”



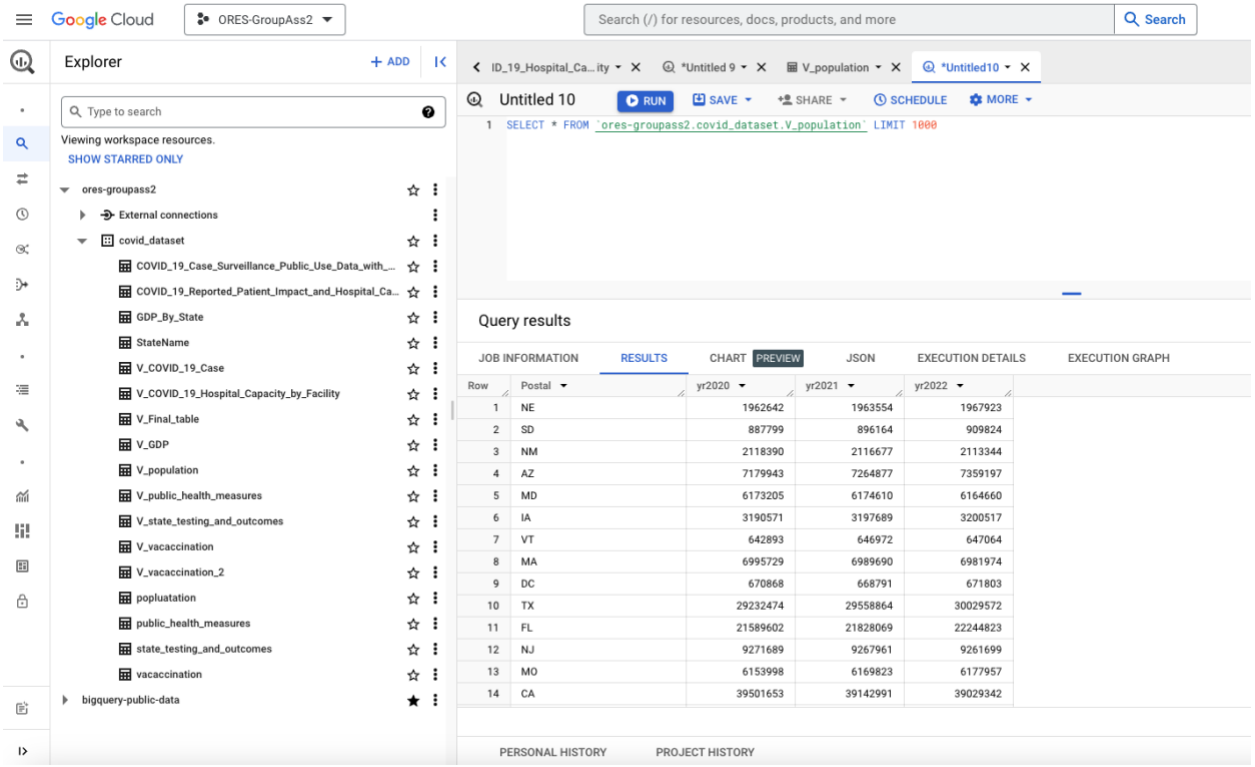
X. Result of table “V_vacaccination”



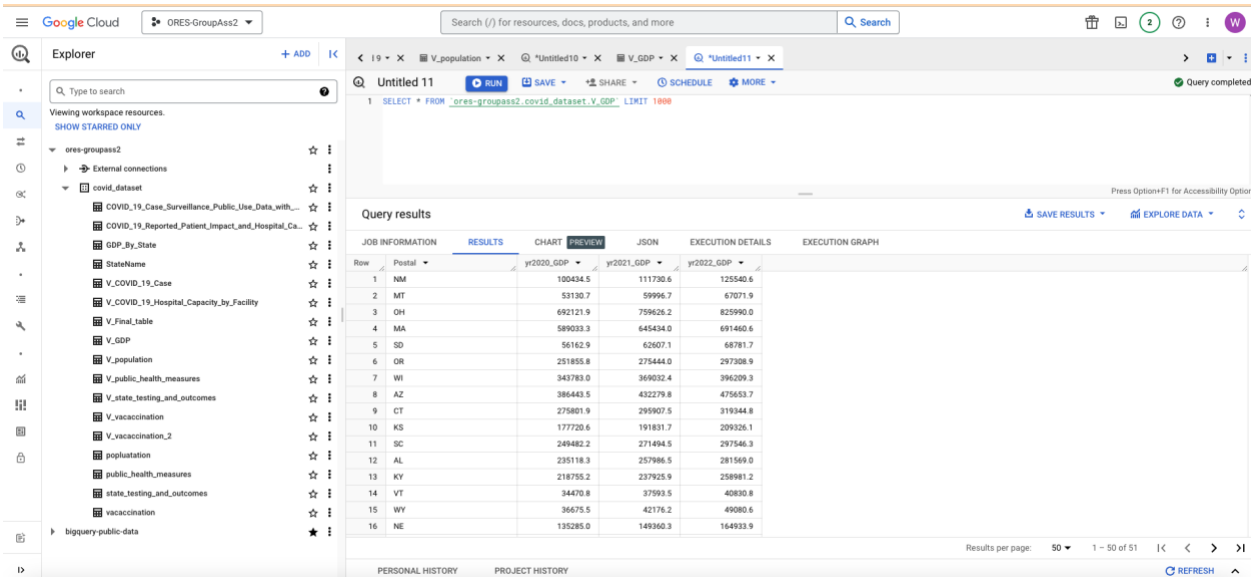
XI. Result of table “V_COVID_19_Reported_Patient_Impact_and_Hospital_Capacity_by_Facility”



XII. Result of table “V_population”



XIII. Result of table “V_GDP”



IV. Process of joining to “V_Final_table”

```sql
create table `ores-groupass2.covid_dataset.V_Final_table`
as

SELECT distinct (CONCAT(cast (T2.Year as string),'-', cast (T2.Month as string),'-' ,T2.state)) As T_id,
T2.Year, T2.MONTH, T2.state, sum(T2.case_total_count) AS cases_monthly_amount,
T6.yr2021 AS yr2021_population, T6.yr2022 AS yr2022_population, --T1.total_beds_7_day_avg,
T7.yr2021_GDP, T7.yr2022_GDP,T3.monthly_negative, T3.monthly_positive, T3.monthly_total_test_results,
T5.hospital_name,T5.total_beds_7_day_avg,T5.total_patients_hospitalized_confirmed_influenza_and_covid_7_day_coverage,
T1.people_fully_vaccinated_per_hundred

FROM
 `ores-groupass2.covid_dataset.V_COVID_19_Case`  AS T2
left join `ores-groupass2.covid_dataset.V_vacaccination_2` AS T1
ON cast(T1.Year as string)  = T2.Year
AND cast( T1.MONTH as string) = T2.MONTH
AND T1.Postal =T2.state
left join `ores-groupass2.covid_dataset.V_state_testing_and_outcomes` AS T3
ON T2.Year =  cast (T3.year as string)
AND T2.MONTH =cast (T3.month  as string)
AND T2.state =T3.state
/*
Response too large to return. Consider specifying a destination table in your job configuration. For more details
*/
left join `ores-groupass2.covid_dataset.V_COVID_19_Hospital_Capacity_by_Facility` AS T5
ON T2.Year =  cast (T5.year as string)
AND T2.MONTH =cast (T5.month  as string)
AND T2.state =T5.state
left join `ores-groupass2.covid_dataset.V_population` AS T6
ON T2.state = T6.postal
left join `ores-groupass2.covid_dataset.V_GDP` AS T7
ON T2.state = T7.Postal
--where T2.age_group != '65+ years' AND monthly_negative is not null
group by T2.Year, T2.MONTH, T2.state, T6.yr2021, T6.yr2022 , --T1.total_beds_7_day_avg,
T7.yr2021_GDP, T7.yr2022_GDP,T3.monthly_negative, T3.monthly_positive, T3.monthly_total_test_results,
T5.hospital_name,T5.total_beds_7_day_avg,T5.total_patients_hospitalized_confirmed_influenza_and_covid_7_day_coverage,
T1.people_fully_vaccinated_per_hundred
```

V. Result of table “V_Final_table”

Group assignment2
Group3