Covid-19: Creating an Aggregate Dataset

To the Department of Health and Human Services Judging Committee:

Recently in our lifetimes, we have gone through a global pandemic as a collective, which has impacted every aspect of our daily lives, our attitudes, the economy, and our sense of interconnectedness and resilience in the face of unprecedented challenges. We live in an era where we are able to communicate with each other across the globe about the nature of the virus and develop and dispatch vaccines at a rapid pace, which has helped to contain the spread of the disease. In the age of data, we have a vast collection of data that describes everything from vaccination patterns to socioeconomic changes during this pandemic. Through our research, we hope to understand many of these patterns.

"How do community-level COVID-19 cases, hospitalizations, vaccine hesitancy, influenza vaccination rates, allocation of provider relief funds, and weather patterns interrelate and influence the trajectory and management of the COVID-19 pandemic at a regional level?" This research question encompasses a wide range of factors that have a substantial impact on the spread and management of COVID-19. By examining these interconnected aspects, we aim to discern patterns, correlations, and potential causal relationships that can inform public health strategies, resource allocation, and targeted interventions to mitigate the spread of COVID-19 and enhance public health preparedness in the face of future pandemics.

To get started, we need a comprehensive dataset that we can explore. For this reason, we have prepared an aggregate dataset by joining several publicly available datasets. Each dataset contains valuable information that we can use to understand the complex interactions among these factors and their influence on the COVID-19 pandemic's regional dynamics.

Community covid levels, fips/zips converter:

The first dataset contains information on community covid levels by county and health service area. Notable features of this dataset include covid cases per 100k, covid hospital admissions and covid inpatient bed utilization. This dataset provides granular information on COVID-19 incidence at the community level, offering critical insights into the severity and geographic distribution of the pandemic. An additional dataset we have included in the aggregate dataset is a county level fips to zips converter, which has counties and their associated zip codes

and fips codes. This information will help to join together datasets with only zip codes and those with only fips codes.

Provider relief fund:

The next dataset contains information on the Provider Relief Fund Nursing Home Quality Incentive Program that allocated $2 billion as an incentive to nursing home facilities to reduce both COVID-19 infection rates relative to their county and mortality rates against a national benchmark. Some notable features of this dataset include total payment per facility, total covid infections, county infection rate, infection performance score as well as mortality adjustment.

Flu vaccine coverage:

The next dataset contains information on influenza vaccination coverage for the general population at the national, regional, and state levels by age group and race/ethnicity. Although this dataset is not specifically related to COVID-19, it serves as a crucial piece of the puzzle by providing insights into influenza vaccination coverage. Understanding influenza vaccination rates during the pandemic is essential, as it allows us to differentiate between influenza and COVID-19 cases, enabling more accurate disease monitoring and resource allocation in the face of co-circulating respiratory viruses. Additionally, it aids in assessing the overall burden on healthcare systems during this challenging period.

Vaccine hesitancy:

The next dataset contains information on COVID-19 vaccine hesitancy rates by county. This dataset on COVID-19 vaccine hesitancy rates by county is a critical resource. It includes estimates of hesitancy, a social vulnerability index, and the percentage of the county population by race, offering a comprehensive view of vaccination challenges. Understanding vaccine hesitancy at the community level, considering social vulnerabilities and racial disparities, is essential for targeted public health campaigns and equitable vaccine distribution. This data enhances our research by illuminating the societal factors that influence vaccination decisions during the pandemic.

Covid19 weather, hospital beds and deaths:

The following three datasets, sourced from BigQuery's public datasets, are instrumental in our research effort. They provide valuable information on daily weather observations by county, aiding our understanding of the influence of weather on COVID-19 transmission and availability of hospital beds by county. Furthermore, a comprehensive summary of COVID-19

cases, including details on confirmed cases, active cases, recoveries, and fatalities by region, enables us to assess the pandemic's impact and the effectiveness of public health measures at different geographic levels. These datasets collectively enhance our ability to explore the multifaceted dynamics of the pandemic and inform evidence-based strategies for managing and mitigating its spread.

Description of final dataset and types of research it could help inform:

The dataset we've compiled from various sources holds significant potential for informing diverse research efforts. It allows researchers to analyze the complex interplay between COVID-19 incidence rates, vaccine hesitancy, influenza cases, and weather patterns, providing insights into disease dynamics and management at the regional level. This dataset also supports the evaluation of incentive programs' impact on COVID-19 infection rates and helps differentiate between influenza and COVID-19 cases for more accurate disease monitoring. Moreover, it offers the opportunity to investigate socioeconomic and racial disparities in vaccine acceptance, aiding in targeted public health interventions. Lastly, the inclusion of weather-related data facilitates the exploration of weather patterns' influence on COVID-19 incidence, contributing to our understanding of environmental factors in disease transmission.

APPENDIX

Link to corresponding presentation:

https://docs.google.com/presentation/d/1xa1cULEbszwOsKdj4klm1JaZsti29FWxK7gFBfoBBwU/edit?usp=share_link

Link to recording of presentation:

https://drive.google.com/file/d/1LJQXmXRFmp8HjtS8w4rpF1uov1bztLS-/view?usp=sharing

Link to other datasets used: Google Drive folder

https://drive.google.com/drive/folders/1FdiEGA0Qpfv-1hnI2hXqtX-zc_xgOdir?usp=share_link

Vendor Provided Datasets: BigQuery

- Covid_19_weathersource_com.county_day_history
- Covid19_aha.hospital_beds
- Covid19_jhu_csse.summary

Individual Links from where we downloaded the datasets provided in the Google Drive folder:

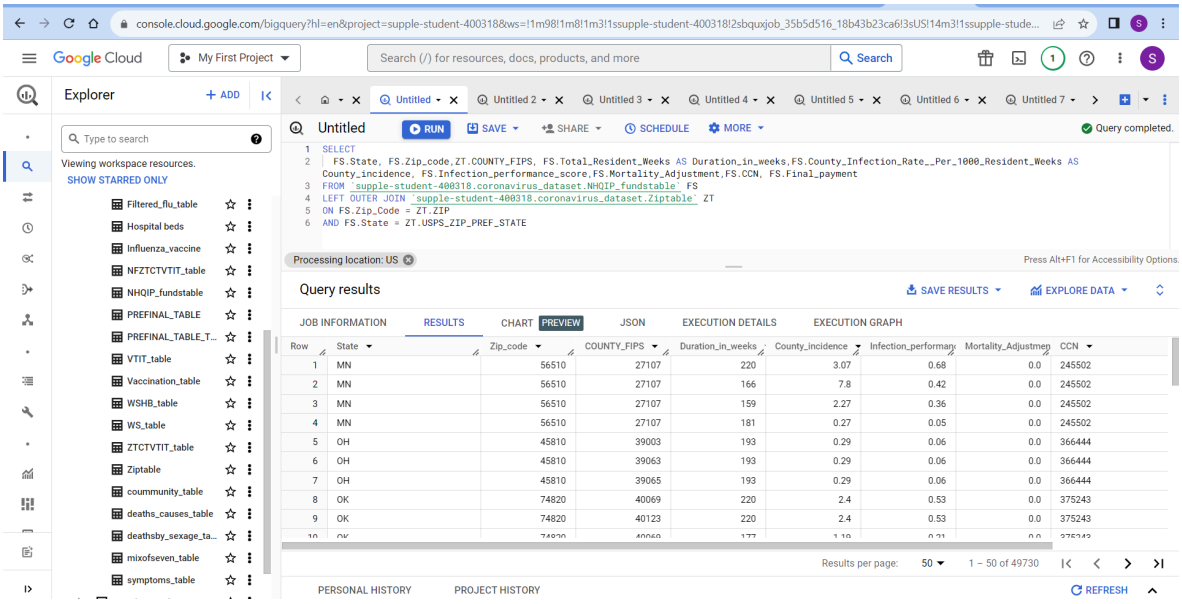
1) HUD USPS ZIP Code Crosswalk files. (n.d.). https://www.huduser.gov/apps/public/uspscrosswalk/home
2) Centers for Disease Control and Prevention. (n.d.). *United States covid-19 community levels by county*. Centers for Disease Control and Prevention. https://data.cdc.gov/Public-Health-Surveillance/United-States-COVID-19-Community-Levels-by-County/3nnm-4jni
3) Centers for Disease Control and Prevention. (n.d.-a). *Covid-19 county hesitancy*. Centers for Disease Control and Prevention. https://data.cdc.gov/Vaccinations/COVID-19-County-Hesitancy/c4bi-8ytd
4) Centers for Disease Control and Prevention. (n.d.-b). *Influenza vaccination coverage for all ages (6+ months)*. Centers for Disease Control and Prevention. https://data.cdc.gov/Flu-Vaccinations/Influenza-Vaccination-Coverage-for-All-Ages-6-Mont/vh55-3he6
5) Centers for Disease Control and Prevention. (n.d.-c). *Provider relief fund COVID-19 nursing home quality incentive program*. Centers for Disease Control and Prevention. https://data.cdc.gov/Administrative/Provider-Relief-Fund-COVID-19-Nursing-Home-Quality/bfqg-cb6d

Resources to help write Queries:

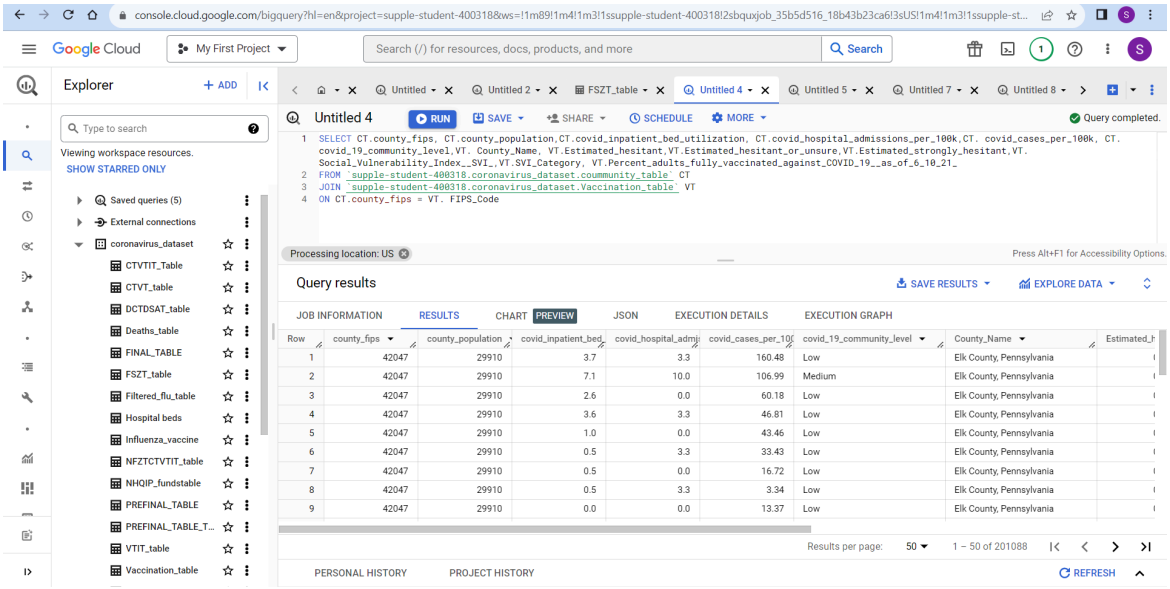
- Writing subqueries in SQL: Advanced SQL - mode. Mode Resources. (2016, May 23). https://mode.com/sql-tutorial/sql-sub-queries/

STEPS TO CREATE AN AGGREGATE DATASET, INCLUDING SCREENSHOTS OF SQL QUERIES:

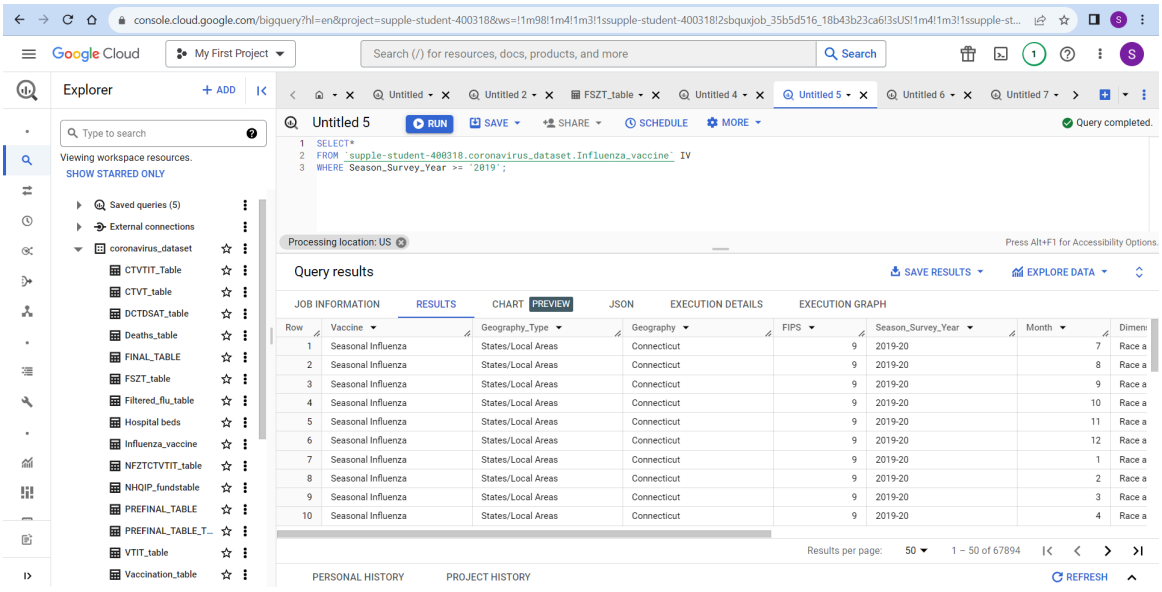
1. Search public health data repositories: CDC, BigQuery, Snowflake, HealthData.gov, for publically available datasets related to COVID-19.
    a. We initially selected Cardiovascular Disease related datasets, but seeing as these datasets did not have common keys to join on, we switched our topic to COVID-19 due to the vast amount of data available as well as county level geographic markers that could be used as a primary key to join on.
2. Create an ongoing document with the name of each table and list the columns of each.
3. Find common column names to use as a key for joining. Highlight these column names.
4. Join 2 tables at a time and analyze the resulting intermediary table. Filter tables as necessary.
    a. We performed a left outer join to integrate Provider relief funds with the county FIPS , Zip codes, using zip codes as the common key for the linkage as we wanted all the data from provider relief funds and only matched values from ZIP table then we named the resulting table "FSZT_table."



b. Next we joined community levels by county with vaccine hesitancy based on fips code as key and named table “CTVT_table”

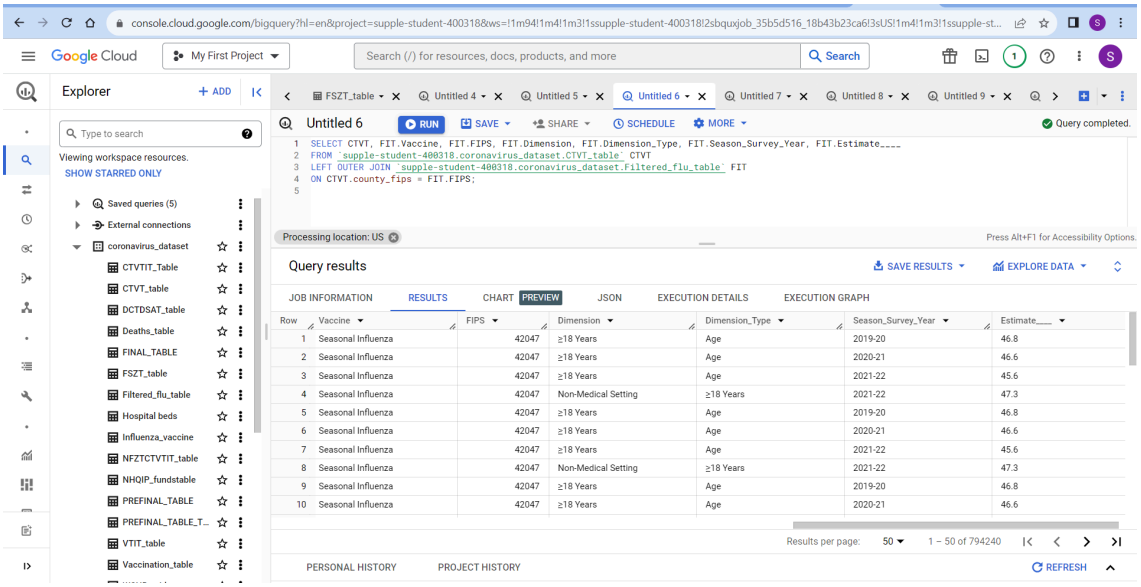


c. We filtered the influenza vaccine table using the following command, selecting data for survey years starting from 2019, and designated the resulting table as "Filtered_flu_table."

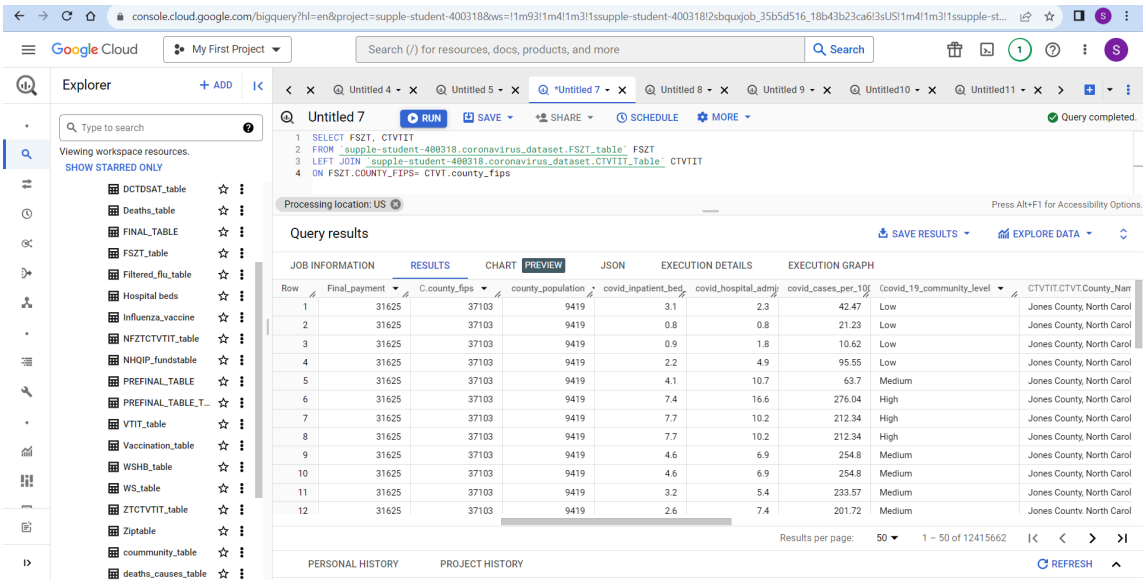


d. Later we joined “CTVT” table with filtered_flu_table using county FIPS code as common key as we wanted all the values from CTVT table and matched values

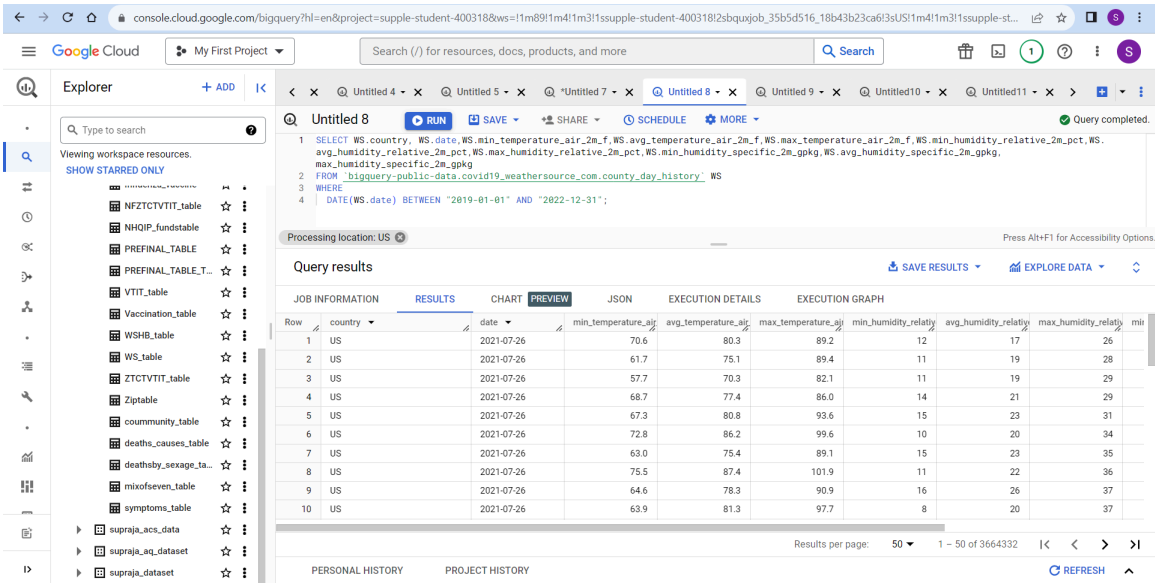
from filtered_flu_table



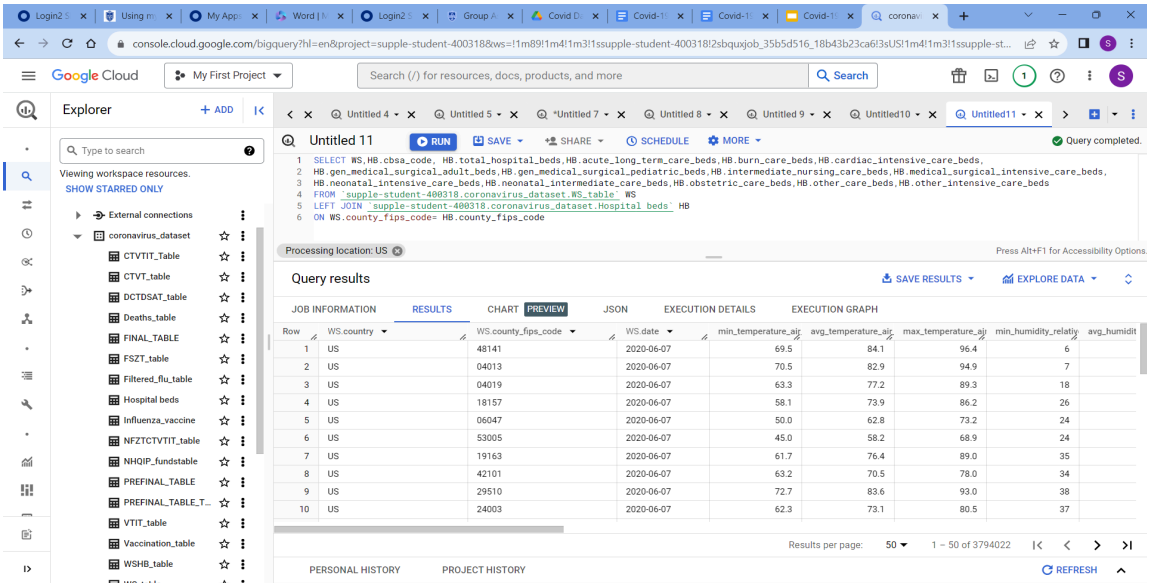
e. Then “FSZT” was added to “CTVTIT” by using county_FIPS as a common key as we wanted all the values from left table and named it as “NFZTCTVTIT”



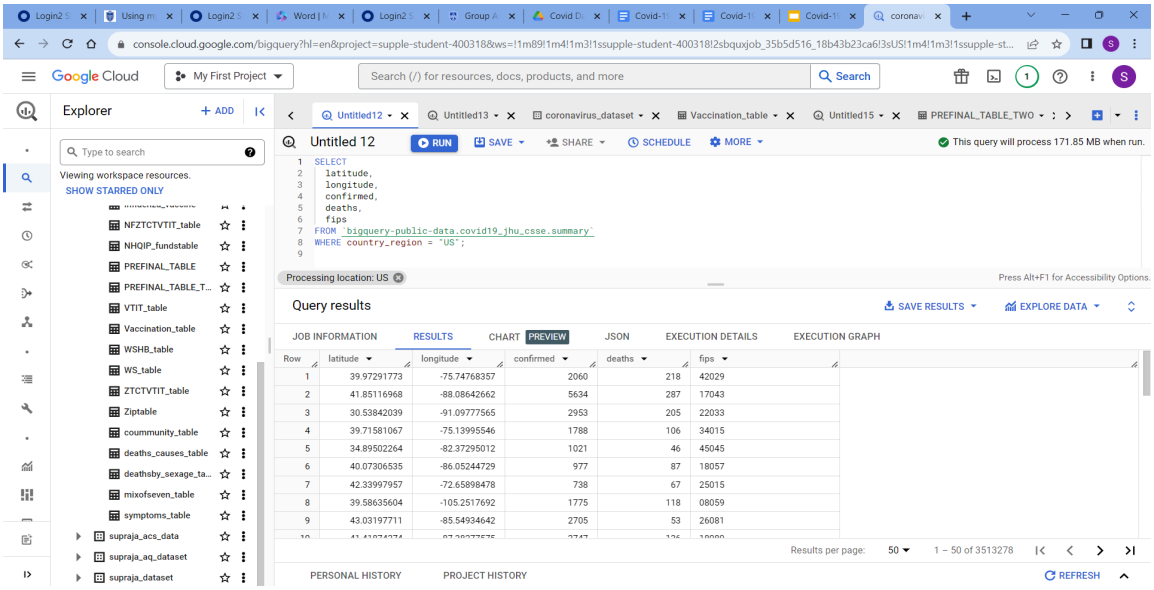
f. Next we filtered covid19 weather by using the WHERE keyword as we wanted data from 2019 and named it as “WS_table”.



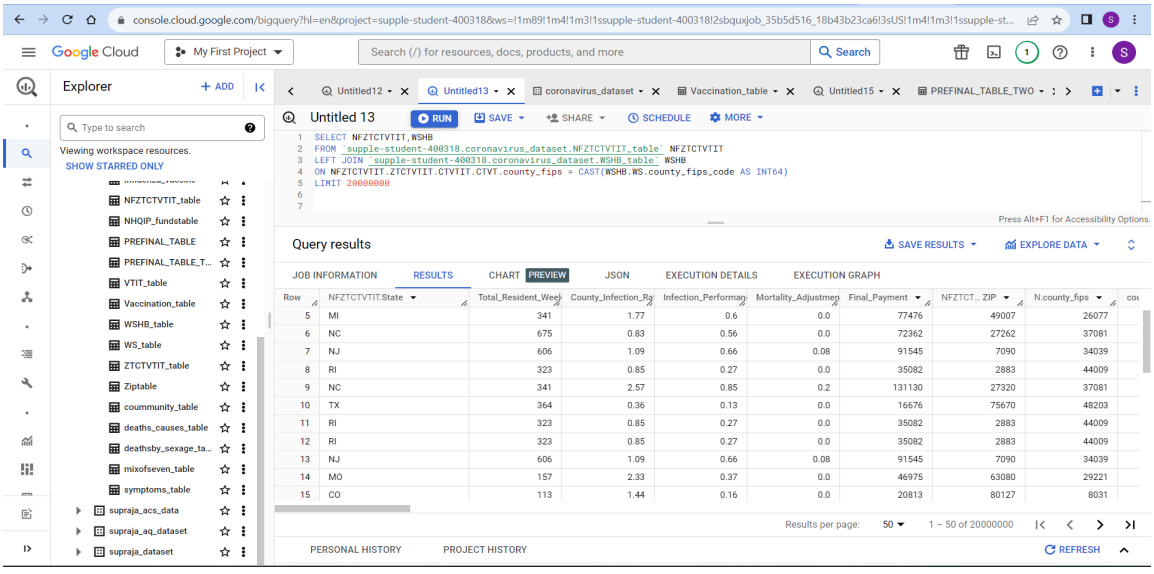
g. We joined ‘WS_table” with hospital_beds table based on county_fips_code as a common key as we wanted all the rows from left table and named resulting table as “WSHB_table”



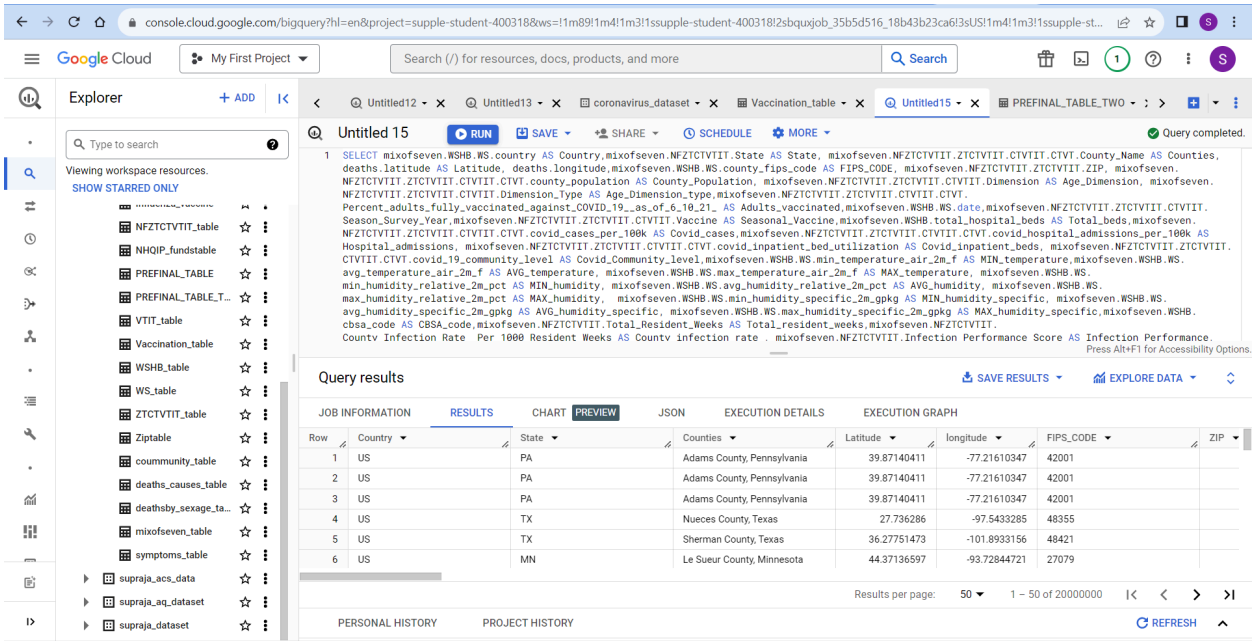
h. Later we filtered Covid19_jhu_csse.summary based on criteria and named it as deaths_table which was joined with PREFINAL_table.



i. NFZTCTVTIT was joined with WSHB to form a table named “FINAL_table”. We've limited the final dataset to 20 million rows and 57 columns, totaling 8.31 GB, to optimize query performance due to significant time delays.

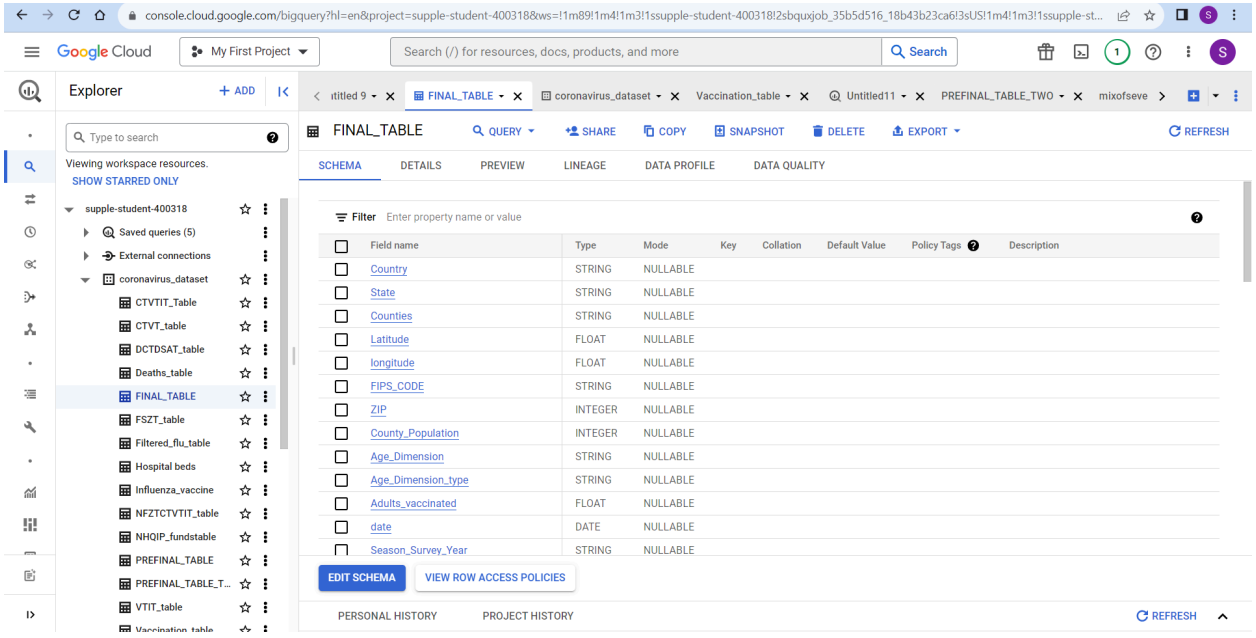


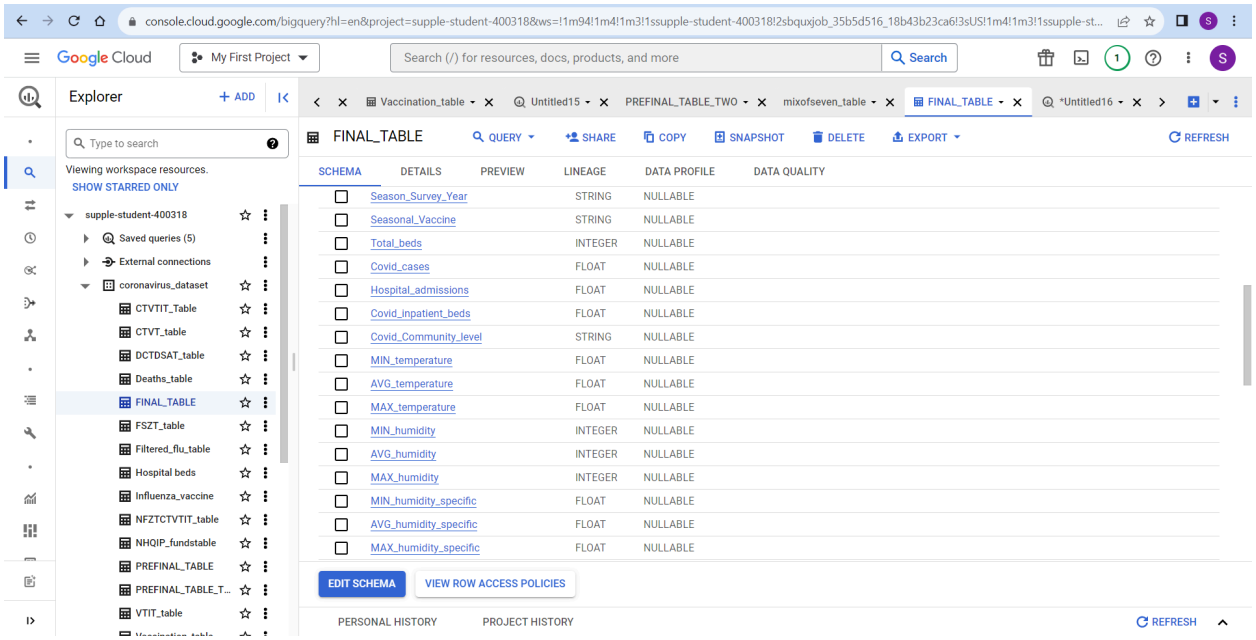
5. Finally we changed the names of the columns as per our requirements.

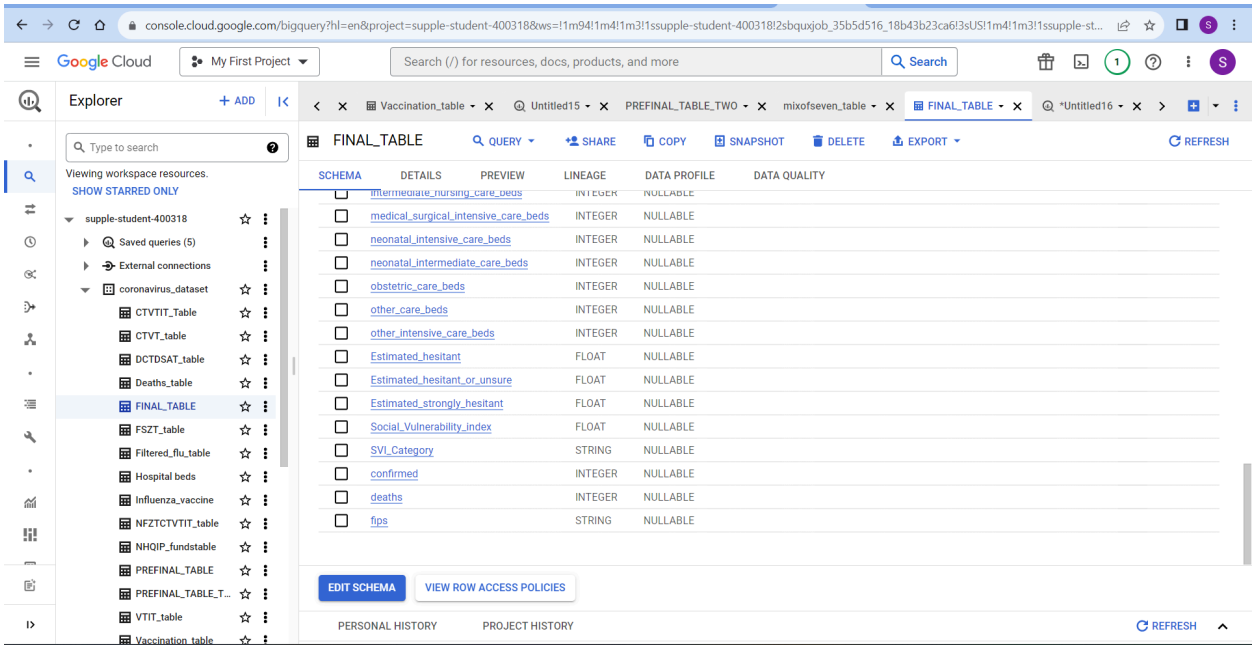


Below is the screenshot of our final data set with its schema, size, number of observations and columns:

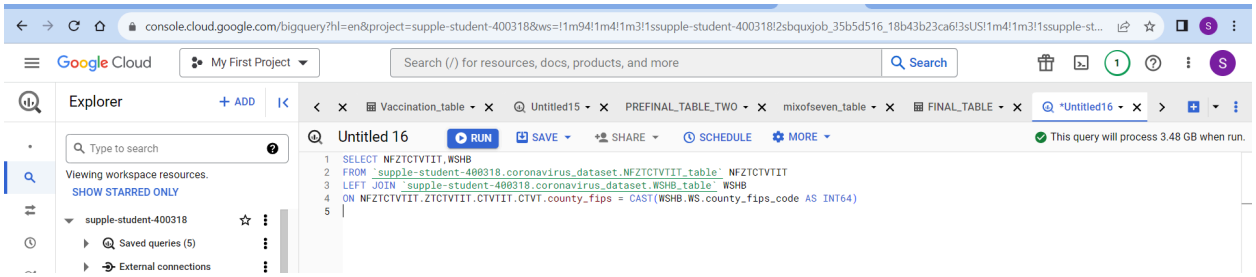
<u>Schema</u>

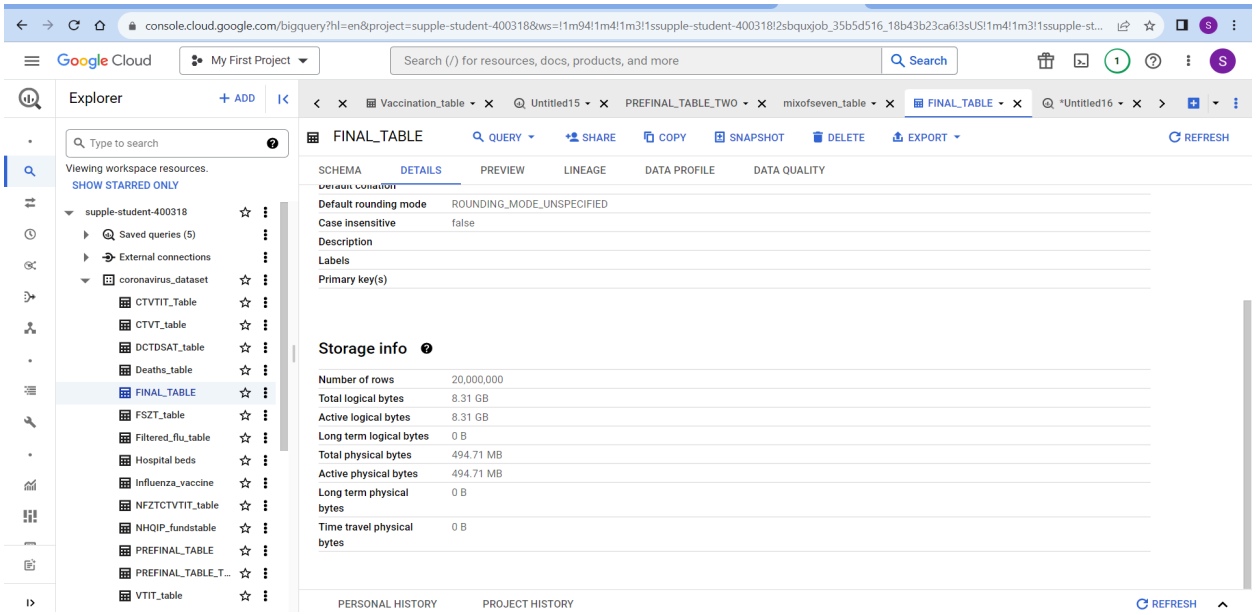



FINAL_TABLE
QUERY
SHARE
COPY
SNAPSHOT
DELETE
EXPORT
REFRESH
SCHEMA
DETAILS
PREVIEW
LINEAGE
DATA PROFILE
DATA QUALITY
Season_Survey_Year STRING NULLABLE
Seasonal_Vaccine STRING NULLABLE
Total_beds INTEGER NULLABLE
Covid_cases FLOAT NULLABLE
Hospital_admissions FLOAT NULLABLE
Covid_inpatient_beds FLOAT NULLABLE
Covid_Community_level STRING NULLABLE
MIN_temperature FLOAT NULLABLE
AVG_temperature FLOAT NULLABLE
MAX_temperature FLOAT NULLABLE
MIN_humidity INTEGER NULLABLE
AVG_humidity INTEGER NULLABLE
MAX_humidity INTEGER NULLABLE
MIN_humidity_specific FLOAT NULLABLE
AVG_humidity_specific FLOAT NULLABLE
MAX_humidity_specific FLOAT NULLABLE
EDIT SCHEMA
VIEW ROW ACCESS POLICIES
PERSONAL HISTORY
PROJECT HISTORY
REFRESH


FINAL_TABLE
QUERY
SHARE
COPY
SNAPSHOT
DELETE
EXPORT
REFRESH
SCHEMA
DETAILS
PREVIEW
LINEAGE
DATA PROFILE
DATA QUALITY
medical_surgical_intensive_care_beds INTEGER NULLABLE
neonatal_intensive_care_beds INTEGER NULLABLE
neonatal_intermediate_care_beds INTEGER NULLABLE
obstetric_care_beds INTEGER NULLABLE
other_care_beds INTEGER NULLABLE
other_intensive_care_beds INTEGER NULLABLE
Estimated_hesitant FLOAT NULLABLE
Estimated_hesitant_or_unsure FLOAT NULLABLE
Estimated_strongly_hesitant FLOAT NULLABLE
Social_Vulnerability_index FLOAT NULLABLE
SVI_Category STRING NULLABLE
confirmed INTEGER NULLABLE
deaths INTEGER NULLABLE
fips STRING NULLABLE
EDIT SCHEMA
VIEW ROW ACCESS POLICIES
PERSONAL HISTORY
PROJECT HISTORY
REFRESH

Size:


Untitled 16
RUN
SAVE
SHARE
SCHEDULE
MORE
This query will process 3.48 GB when run.
SELECT NFZTCTVTIT.WSHB
FROM `supple-student-400318.coronavirus_dataset.NFZTCTVTIT_table` NFZTCTVTIT
LEFT JOIN `supple-student-400318.coronavirus_dataset.WSHB_table` WSHB
ON NFZTCTVTIT.ZTCTVTIT.CTVTIT.CTVT.county_fips = CAST(WSHB.WS.county_fips_code AS INT64)



Preview of our Final Dataset: