**I535 Final Project – COVID-19 Data Ingestion, Storage, & Visualization**

For this project, I chose to focus on the course module regarding data ingestion and storage. I selected a dataset from the CDC website containing records of COVID-19 vaccinations in the United States, which can be found [here](https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Trends-in-the-United-States-N/rh2h-3yt2/about_data). This dataset contains information regarding various dosage types, trends, and geographic metadata. Date entries range from November of 2020 to May of 2023, when the end of the coronavirus pandemic was declared by the World Health Organization. (1, 2)

A close-up of a text

Description automatically generated

I chose this dataset for a few reasons. The COVID-19 pandemic occurred in recent history, and it illuminated the fact that many developed countries are ill-prepared to handle a national or worldwide pandemic. As such, the storage and analysis of COVID-19 related data likely holds a much greater degree of immediate practical value compared to information regarding smallpox or scarlet fever (although analyses of such diseases are not without their value). Because the pandemic began at a time when modern record keeping technologies were available, and occurred over a relatively short period of time, analysts can essentially review the entire history of the pandemic. This is a massive benefit compared to other prevalent international diseases such as tuberculosis, which has been detected in organic remains from roughly 17,000 years ago. (3)

The information within the initial file (shown below) is thorough, and it contains more than enough data to warrant the use of BigQuery and Tableau. However, much of the metadata surrounding the dataset is esoteric and inconsistent. Due to the large number of entries, there are few trends that are easily noticed without the aid of visualization software. As such, this led me to believe I could realistically apply the skills I learned surrounding BigQuery and GCP within the assigned labs, and that my results would be more useful within visualization software than the initial file. Thankfully, after multiple weeks of work, this assumption held correct.

A screen shot of a computer

Description automatically generated

First, I exported the dataset from the [CDC website](https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Trends-in-the-United-States-N/rh2h-3yt2/about_data) in CSV format, primarily to ensure a clean upload to BigQuery later in the process (click the “Export” button at the top right of the page to replicate this step). I also downloaded the [data dictionary](https://data.cdc.gov/api/views/rh2h-3yt2/files/6fb9b669-a258-45d4-823b-9e7b3e08397e?download=true&filename=DataDictionary_v37_05052023.xlsx), both as a reference and as an alternative set of data. One of my BigQuery tables was generated solely from the latter file. The initial CSV data contained 29 columns, many of which could be inferred from existing columns (such as cumulative totals and rolling averages). I condensed this data down to nine columns: six related to dosage types, two listing the date and location, and an added ID column to aid in SQL queries within BigQuery. To replicate this step, first open a new CSV file (using Excel or a similar application) and insert a column with the header, “ID”, then populate it with a sequential series of numbers. (This can be completed by filling in the first two columns and then dragging down the fill bar, or by using the “Fill Series” command within the Excel ribbon). Next, copy the following columns from the original file (my example is shown below):

* Date
* Location
* Administered\_Daily
* Admin\_Dose\_1\_Daily
* Series\_Complete\_Daily
* Booster\_Daily
* Second\_Booster\_50Plus\_Daily
* Bivalent\_Booster\_Daily

A screenshot of a computer

Description automatically generated

As you can see within the condensed file, each group of daily totals is linked to an abbreviated state entry. However, territories and agencies are listed as well, which is not immediately obvious given the abbreviations. For example, “VA2” and “PW” are unlikely to be recognized as “Veterans’ Affairs” and the “Republic of Palau” without referencing the data dictionary first. Similarly, those not familiar with US abbreviations could be forgiven for mistaking “MH” as Michigan or Massachusetts, instead of the Marshall Islands. Conversely, reading “MI” as the Marshall Islands would be equally understandable.

A portion of the abbreviations reference, including non-states such as the Bureau of Prisons:

A list of states with names

Description automatically generated

To address this issue, I split the data into two parts and analyzed each portion separately. For the first portion, I limited the condensed file to daily totals across the entire US, ignoring all data subsets pertaining to individual states, territories, and agencies. This was achieved by selecting the entire “Locations” column, using the “Sort & Filter” command, and choosing “Filter”. I ensured only “US” was checked within the dropdown menu for the column and copied the entire filtered sheet to a new CSV file. My result is shown below:

A screenshot of a computer

Description automatically generated

I decided to generate the second portion of data within BigQuery, using the original unaltered file from the CDC website. As such, I created two datasets within BigQuery, then uploaded both the original CSV and the cleaned file as tables. My settings for the original file are shown below. I repeated the process for the cleaned file, giving the second table a distinct name and ensuring the other database was selected.

A screenshot of a computer

Description automatically generated

A close up of text

Description automatically generated

With all necessary information now in BigQuery, I began writing queries for both portions of the dataset. For the location-focused data, I copied the abbreviations and “longnames” into VS Code and formatted them as strings within an array. Within BigQuery, I created a table to store each location and its total number of vaccinations. Then, I wrote a script to convert each abbreviation into its human-readable counterpart, retrieve the total number of vaccinations for each location, and store each matching pair of values in a table called “StateTotals”. To aid in reproduction of my results, I have included my code within a separate file that has been included within my submission. (It can also be found on GitHub.)

A screenshot of a computer

Description automatically generated

Queries for the other set of data were much simpler, although I wrote six separate queries to account for each dosage type. Each of these queries retrieved the ten dates with the highest number of administered doses for each dosage type and stored them in a table along with their ID value. This code has also been included in a separate file and can be found on GitHub.

A screenshot of a computer

Description automatically generated

At this point, I connected to my BigQuery project within Tableau, connected the tables as shown below, selected date values as the values to compare, and loaded my selections into a worksheet. I created a separate worksheet that contained only the “StateTotals” table.

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

For the first visualization, I stored the date from the initial data file within the column field and selected “Month” as the format. Then I moved the measure values for each dosage type into the row field. I selected “Bar” as the type of graph. This resulted in a visualization that shows “bursts” of administered doses across all types:

A screenshot of a computer

Description automatically generated

For the next visualization, I again selected the formatted date for the column field. Then, I stored all measure values within the rows field, and all measure names within the filters field. I selected “Line” as the type of graph. This resulted in a graph that shows a somewhat similar “triple peak” pattern across all dosage types:

A graph of a graph

Description automatically generated with medium confidence

Finally, I connected to the StateTotals table and selected a bar graph. I stored “State” within the column and filter fields. I stored the measure value for total vaccinations within the row field. However, this data included the US as a location, which unsurprisingly rendered all other values insignificant. I hid this value by right-clicking on the field, which resulted in a far more useful visualization:

A screenshot of a computer

Description automatically generated

These results reveal a few interesting patterns. Beginning with the collection of graphs for dosage types, you can see a similar pattern between high values for the initial dose and completion of the initial series. Both dosage types have a high initial value, followed by a collection of smaller values four to six weeks later. The series completion values are delayed by roughly a month compared to the initial doses, which makes sense considering what each type represents. Likely, a large amount of people received their first dose shortly after it was available, then went back to complete the dosage series after the recommended waiting period.

The pattern within the initial booster section does not appear to be replicated elsewhere. However, a single value within the daily total administrations section lines up with the initial booster values: November 29, 2021. Roughly 2.75 million boosters and 4.75 million total administered doses were recorded on this day, meaning the booster accounted for over half of all doses for the day. Similarly, the last four values for the highest daily totals fall within the full group of values for the bivalent booster, which range from October 12th of 2022 to December 14th of 2022. The last seven values for the fifty-plus booster line up perfectly with these four daily totals. Both sets of values begin on October 5th of 2022 and end on November 16th of 2022.

As mentioned before, when looking at the combined line graph for all recorded dosage types, most appear to exhibit a “triple-peak” pattern. This pattern is likely most noticeable within the daily total data, which peaks in April of 2021, December of 2021, and October 2022. These peaks make sense considering the consistent spike in coronavirus fatalities during December and January. (4) The FDA first approved a vaccine for emergency use in December of 2020 (5), so the delayed peak in the following April is reasonable considering the time needed for mass production and distribution. The subsequent peaks in December and October fall within the “winter surge” and shortly before, respectively.

This graph also reveals which dosage types are primarily responsible for peaks in the daily total. The first peak appears to be the result of individuals either receiving their first dose or completing their first series, with both groups relating to roughly 50% of the total within April 2021. The second peak is primarily related to administrations of the booster for those over the age of fifty, although many individuals completed their first dose or series near this time period as well. The last peak reveals an interesting anomaly within the dataset.

A graph of a number of months

Description automatically generated with medium confidence

The last peak is comprised of values from the bivalent booster (the red line), and this subset surpasses the daily total (the orange line) within November 2022. This comparison is obviously illogical; no individual dosage type should surpass the daily total. This paradox could simply be a result of careless record-keeping, although the reason is likely more complex. This peak occurs near the end of the coronavirus emergency, when federal requirements surrounding vaccines were gradually relaxed. (2) Records of individual vaccines may have taken priority over ensuring a balanced total near this point, considering dosage types other than the fifty-plus booster (the green line) are nearly flat by this point. In fairness, I do not have considerable evidence to support these claims beyond the graphs. They serve as reasonable guesses for now and would require more research to consider objectively.

The last graph consists of total vaccination values for each state, territory, and agency. As mentioned before, values for total US entries were eliminated. California is immediately noticeable as the state with the highest vaccination total. The next highest values in order are Texas, New York, and Florida. Of note, this pattern matches population data across the US. California has roughly 39 million inhabitants, Texas has roughly 30.5 million, Florida has roughly 22.5 million, and New York has roughly 19.5 million. Roughly 13 million people live in Penssylvania, and the values decrease by less than 1 million for the remaining states. This leads me to believe vaccination data and state population values are highly correlated. (6)

I am satisfied with the final quality of each graph, and I utilized many skills I learned within the data ingestion and storage module to achieve these results. Our assigned lab, “Ingesting New Datasets into BigQuery”, prepared me to transform each CSV file into BigQuery tables after I cleaned the initial data. I also decided to begin with static data, specifically to address the performance issues cited within the same lab. My final tables primarily served as a relational database, which was covered as a refresher within the same module of the course. I am also thankful that I designed the project with reproducibility in mind. This will allow me, or other individuals, to research the vaccinations dataset in greater detail at a later date.

**References**

1. U.S. Health and Human Services. COVID-19 Public Health Emergency (PHE). <https://www.hhs.gov/coronavirus/covid-19-public-health-emergency/index.html> . Last reviewed December 15, 2023
2. United Nations. WHO chief declares end to COVID-19 as a global health emergency. <https://news.un.org/en/story/2023/05/1136367> . May 5, 2023
3. Lawn SD, Zumla AI (July 2011). "Tuberculosis". Lancet. 378 (9785): 57–72. <https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(10)62173-3/>
4. CDC. COVID-19 can surge throughout the year. <https://www.cdc.gov/ncird/whats-new/covid-19-can-surge-throughout-the-year.html>
5. FDA. FDA Takes Key Action in Fight Against COVID-19 By Issuing Emergency Use Authorization for First COVID-19 Vaccine. <https://www.fda.gov/news-events/press-announcements/fda-takes-key-action-fight-against-covid-19-issuing-emergency-use-authorization-first-covid-19>
6. United States Census Bureau. State Population Totals and Components of Change: 2020-2023. <https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html>