Yugine Lama- Yugineruit2012@gmail.com

Melanie Li- [limelanie01@gmail.com](mailto:limelanie01@gmail.com)

Salma Elhassa [salmae309@gmail.com](mailto:salmae309@gmail.com)

Falande Alexandre [Alex.falande@gmail.com](mailto:Alex.falande@gmail.com)

**Topic:**

Covid-19 vaccinations on US data.

**Problem statement:**

Since the COVID vaccine was created, much data has been collected. The data ranges from how many people have been vaccinated, the rate of vaccination, and how has COVID subsided since the vaccine rollout. Many people tend to underestimate or overestimate the impact COVID has had worldwide, thus it is important to highlight the data to the worldwide audience but we have focused on the US.

This data will help businesses and governments predict when they can start operating businesses with less restrictions, reopening schools, and transitioning to live similar to pre-covid.

**1. Information about your data**

**a. Why did you choose the data set you were working with?**

**Answer:** Since the COVID vaccine was created, much data has been collected. The data ranges from how many people have been vaccinated, the rate of vaccination, and how has COVID subsided since the vaccine rollout. Many people tend to underestimate or overestimate the impact COVID has done worldwide, thus it is important to highlight the data to the worldwide audience.

**b. What was your dataset about? What were some of the important columns in your data and why are they important?**

**Answer:** So, our dataset is about US daily-covid-vaccine-dose-administered. The data gives the details about each state's’ daily data starting from 21st December, 2020 until 12th May 2021. They are important because it gives detail about the number of vaccination records and keeps track.

**c. What kind of things did you want to explore about your data?**

**Answer:** We wanted to explore more about the age group as well as find areas that do not have access to vaccination as well. We are also trying to find vaccination on the income aspect.

**3. What if anything, did the project teach you? Do you have any suggestions to improve this project? What issues did you face when trying to answer your questions/copying the data (or any other issues)?**

Answer: This project not only helped us in finding the real data of COVID vaccination in the US but also equally helped us in answering the some curiosities we had. We compared each state's records and how many were administered so far. We were able to find the number of vaccinations compared to cases that each state has/had.

**4. If you had unlimited time and resources (i.e. you can collect your own dataset and use tools other than SQL) what project would you pursue?**

Answer: We enjoyed doing this project but if we had more time and resources we would have worked on the same project with more details such as collecting data based on the brand's name and such. We wanted to find which brands of vaccine were used most in all states.

**5. Did you enjoy this class? What did you like? What did you dislike? Do you have any advice for improvements or other suggestions? If you have multiple authors, please separate by author)**

**Responses:**

Yugine: Yes, very much. We were made sure if we understood what was taught in class, which was one of the best parts. Since there was a recorded lecture it was even better when we missed any part during class time. One thing I disliked was the last minute 3 assignments. The more participating in class must be encouraged if it's still going to be online then.

Falande: I really enjoyed this class and the coding language, this is my 2nd time taking a comp sci and this language was very simple and usual to me as I am working in a lab. I just disliked how the last 3 assignments were so close in due dates I wish they were given earlier and more spread apart so I could plan accordingly with all my other work.

Melanie: Yes, I enjoyed this class. I really liked how we explored different ways to use sql in order to solve the problems we have in mind. This class helped us critically think and navigate through various sources to assist us in the various assignments we have. Like the others, I also dislike the short time frame we had for the assignments because it seems to be back to back. Instead, I think it would be much better if each of those projects were given to us with a month difference.

Salma: I thoroughly enjoyed this class. In fact I was able to apply it to other things. I took part in a blind study at the Microbe Directory in Dr. Christopher Mason’s lab at Weill Cornell. There I contributed to the directory (a sql database) by specifying various microbe’s taxonomy including optimal growth pH, morphology, microbiome, symmetry, mobility, and secondary structures. These taxonomic classifications are eventually utilized by other researchers to link to meta genomic taxonomic analyses. Maybe the pace was a little quick, but i eventually caught up so it was not too bad.

**2. For each of the questions you asked, please answer the following: Please also write the question text. If you had multiple authors, kindly indicate which author answered which questions.**

Answer: We have asked questions and answered all of them as a group rather than making individual responses. Our questions are as follows:

**https://ourworldindata.org/grapher/vaccination-coverage-by-income-in**

**Our Question 1:** Which areas have lower/higher access to vaccines based on income?

**a. Why did you ask this question? Was there any information you wanted to learn?**

We asked this question in order to view if there is a disparity between the low and high income class when it comes to the access to the covid vaccines.

**b. Did you succeed in answering the question? If you did succeed, what did you learn? If you weren’t successful, can you propose a method which might help solve the question (i.e. different data, or using a different tool than SQL etc.)?**

Yes, we did succeed in answering the question. We created a solution code, and used a different data set using the union function to merge more than one dataset.

**c. Which SQL tool(s) did you use to solve the question (i.e. aggregates, joins, cases etc.)?**

We used an aggregate function in order to compare the average vaccination rate between the two social classes. The union function can be used, however, it would depend if the columns are the same in number. This would enable the date to be compared to side by side to vaccination rates in various states in the U.S., specifically comparing the low and high income classes.

**Our Question 2:** Which day had the lowest vaccination record and in which state?

**a. Why did you ask this question? Was there any information you wanted to learn?**

We ask this question because we wanted to know at which location had the lowest vaccination record even after the vaccine for the pandemic had been introduced and when did the least amount of vaccination occur.

**b. Did you succeed in answering the question? If you did succeed, what did you learn? If you weren’t successful, can you propose a method which might help solve the question (i.e. different data, or using a different tool than SQL etc.)?**

For this question, we provided two solutions because the first solution gave us a negative number, which shouldn’t be possible. From the result it gave us, we learned that on May 11, 2021, New Hampshire had a daily vaccination of -678. So, we included another statement for a non-negative response to our question. In our second solution, we learned that on January 13, 2021, Guam had a daily vaccination of 0.

**c. Which SQL tool(s) did you use to solve the question (i.e. aggregates, joins, cases etc.)?**

Although we didn’t use any functions, we used the group by statement to group the result by the said columns. We also ordered it by listing the daily\_vaccinations in an ascending order and limited the result to one, which would only show us the least amount of vaccination in the order of the entity. Due to the unexpected outcome in the first solution, we decided to include a where statement into our second solution. That statement allowed us to determine the location where the daily vaccination would be greater than or equal to 0, which would not give us a negative result like the first one did.

**https://covid.cdc.gov/covid-data-tracker/#vaccinations**

**Our Question 3:** which areas have been/not been vaccinated based on political affinity?

**a. Why did you ask this question? Was there any information you wanted to learn?**

We ask this question because we are aware that during the last President’s time in office, a lot of controversy struck in terms of vaccination. As a result of his republican affiliation, his supporters were less reluctant to get the vaccine. With the new Presidency, we want to analyze the difference in vaccination rate in those red states and blue states and those in between.

**b. Did you succeed in answering the question? If you did succeed, what did you learn? If you weren’t successful, can you propose a method which might help solve the question (i.e. different data, or using a different tool than SQL etc.)?**

Yes, we did succeed in answering the question. We used a different dataset that has a column on political affiliation and includes democrats, republicans, third party, or none of the above.

**c. Which SQL tool(s) did you use to solve the question (i.e. aggregates, joins, cases etc.)?**

We can use the case statement and join function in order to not include income and only include those who are vaccinated and if those people are affiliated with political groups such as democratic, republican, moderate, or something else.

**Our Question 4- How many vaccine hub per area to find out if everyone has access to vaccines?**

**a. Why did you ask this question? Was there any information you wanted to learn?**

We wanted to know how many hubs were there in each state and based on that we could see the ratio of the number of population in each state and the rate of people getting vaccinations.

**b. Did you succeed in answering the question? If you did succeed, what did you learn? If you weren’t successful, can you propose a method which might help solve the question (i.e. different data, or using a different tool than SQL etc.)?**

No, we were not successful in getting this answered. We were a bit out of time as well as lack of resources so couldn’t solve this. We could have solved this, if we found another dataset that includes a vaccine hub column.

**c. Which SQL tool(s) did you use to solve the question (i.e. aggregates, joins, cases etc.)?**

Question could not be answered.

**Our Question 5-** Which state has the most vaccinations up to this day (May 12, 2021) according to the dataset?

**a. Why did you ask this question? Was there any information you wanted to learn?**

We ask this question because we want to be aware of which state is using the vaccines the most and is taking advantage of the vaccine distribution in order to help others in need and help them be safe.

**b. Did you succeed in answering the question? If you did succeed, what did you learn? If you weren’t successful, can you propose a method which might help solve the question (i.e. different data, or using a different tool than SQL etc.)?**

Initially, we weren’t succeeding in answering the question, because the first solution gave us the United States as the state with the most vaccinations up to this day. What we want to find is a state, not a country, so we decided to include another statement. This resulted in our second solution, from this we learned that California is the state that has the most vaccination, totaling up to 32,448,015.

**c. Which SQL tool(s) did you use to solve the question (i.e. aggregates, joins, cases etc.)?**

We used an aggregate function to find the total amount of vaccinations for each location in the dataset. By doing that, we are able to list the locations according to the sum of vaccinations from highest to lowest amount. We also used the where statement to allow us to determine the location that has the most total vaccination that purposely excludes the United States because it’s not a state.

**Our Question 6:** How can we encourage more people to get the vaccine?

**a. Why did you ask this question? Was there any information you wanted to learn?**

We can use our overall analysis of data. As you can see from our queries of the data set we can put faith within the American people that the usage of vaccines has substantially aided in the decrease of covid cases. This is in addition to the fact that we do not see any adverse effect of the vaccine as more Americans continue to receive it. Overall, the fact that we have covid deaths, but not covid vaccine deaths should be reason enough for people to get vaccinated.

**b. Did you succeed in answering the question? If you did succeed, what did you learn? If you weren’t successful, can you propose a method which might help solve the question (i.e. different data, or using a different tool than SQL etc.)?**

No. This was more like open ended questions so we just thought of not solving this because different people have different responses.

**c. Which SQL tool(s) did you use to solve the question (i.e. aggregates, joins, cases etc.)?**

Question could not be answered.

**Our Question 7: Which day had the highest vaccination record in New York?**

**a. Why did you ask this question? Was there any information you wanted to learn?**

We ask this question because we found it interesting and at the same time we wanted to know how early on or later on in the covid situation people decided to take the vaccine. In addition, we decided to choose New York State simply because we are currently residing in this area.

**b. Did you succeed in answering the question? If you did succeed, what did you learn? If you weren’t successful, can you propose a method which might help solve the question (i.e. different data, or using a different tool than SQL etc.)?**

Yes, we succeeded in answering the question. We learned that April 4, 2021, is the day in New York State that people took the most vaccinations out of all the other days. The specific amount of vaccinations taken on that day according to the dataset is 266,430.

**c. Which SQL tool(s) did you use to solve the question (i.e. aggregates, joins, cases etc.)?**

Although we didn’t use any functions, we used the group by statement to group the result by the said columns. We also ordered it by listing the daily\_vaccinations in a descending order and limited the result to one, which would only show us the most amount of vaccination out of the New York State entity.

**Our Question 8:** **Which entity in the dataset should have an increase of access to vaccinations?**

**a. Why did you ask this question? Was there any information you wanted to learn?**

We ask this question because we believe it’s important to distribute and increase the access to the vaccines efficiently. In other words, we want to learn which location needs it the most.

**b. Did you succeed in answering the question? If you did succeed, what did you learn? If you weren’t successful, can you propose a method which might help solve the question (i.e. different data, or using a different tool than SQL etc.)?**

Yes, we succeeded in answering the question. The code lists the location that has the least vaccinations to the location that has the most vaccinations. If we want the result to show only one entity, we will have to add “limit 1;” after “asc” on the last line of the code. We learned that the Republic of Palau has the least average amount of vaccination out of all entities listed in the dataset. The average vaccination of that location is 174.15.

**c. Which SQL tool(s) did you use to solve the question (i.e. aggregates, joins, cases etc.)?**

We used an aggregate function to find the average amount of vaccinations for each location in the dataset. By doing that, we are able to list the locations according to the average vaccination from least to highest amount. We also used a scalar function, specifically round(), which helped us round and limit the average amount to the hundredths place.

**Our Question 9.** How many Americans are currently fully vaccinated as of May 12th in the United States?

**a. Why did you ask this question? Was there any information you wanted to learn?**

We asked this question as we wanted to know the people’s thoughts about vaccination. There have always been various groups of people who are for or against anything so although many people are willing to get vaccines there are many who aren’t. So, based on different states we wanted to know the fully vaccinated population until today.

**b. Did you succeed in answering the question? If you did succeed, what did you learn? If you weren’t successful, can you propose a method which might help solve the question (i.e. different data, or using a different tool than SQL etc.)?**

Yes, we succeed in answering this question. We learned that the United States has a sum of 517,759,580 people who have been vaccinated until May 12, 2021.

**c. Which SQL tool(s) did you use to solve the question (i.e. aggregates, joins, cases etc.)?**

We used an aggregate function to find the total amount of vaccinations for the United States in the dataset. We used the where statement in order to specifically call for the United States and did not find the sum for the other entities because we noticed the dataset not only has the states, but other locations as well.

**Our Question 10. How many people under 18 have been vaccinated ?**

**a. Why did you ask this question? Was there any information you wanted to learn?**

We asked this question because now that everyone is allowed to get vaccination who are age 18 and above, we wanted to learn what under-aged populations are being vaccinated**.**

**b. Did you succeed in answering the question? If you did succeed, what did you learn? If you weren’t successful, can you propose a method which might help solve the question (i.e. different data, or using a different tool than SQL etc.)?**

No, we did not succeed in answering the question. Even though we have a solution code for this question, we might have been successful if we were using a different dataset that includes that age of people who have been vaccinated.

**c. Which SQL tool(s) did you use to solve the question (i.e. aggregates, joins, cases etc.)?**

We used the count function so we would be able to count those who vaccinated for each age group. Especially, for our solution code, we will be only seeing the result of the amount of people who have been vaccinated below the age of 18.

**Our Question 11:** What is the average rate of vaccination per day?

**a. Why did you ask this question? Was there any information you wanted to learn?**

We asked this question in order to know how many people are constantly getting vaccinations based on the average vaccination in our dataset.

**b. Did you succeed in answering the question? If you did succeed, what did you learn? If you weren’t successful, can you propose a method which might help solve the question (i.e. different data, or using a different tool than SQL etc.)?**

Yes, we succeeded in answering the question. This is an active result, that will update each day. We learned that the first day of receiving the vaccine (December, 21, 2020) from all the entities in the dataset had an average of 57,909.00 vaccinations and that the most recent day (May 12, 2021) has an average of 66,943.09 vaccinations.

**c. Which SQL tool(s) did you use to solve the question (i.e. aggregates, joins, cases etc.)?**

We used an aggregate function to find the average amount of vaccinations from all the locations in the dataset. By doing that, we are able to list the average vaccinations according to the dates of when it first got distributed until today (May 12, 2021). We also used a scalar function, specifically round(), which helped us round and limit the average amount to the hundredths place.

**Our Question 12. Which age group is more likely to get vaccinated?**

**a. Why did you ask this question? Was there any information you wanted to learn?**

After the vaccination was available there was an age group to get the vaccine and now that all 18 years and plus can get the vaccine, we wanted to find which age group got the vaccine the most as well as find who are more likely to get vaccinated.

**b. Did you succeed in answering the question? If you did succeed, what did you learn? If you weren’t successful, can you propose a method which might help solve the question (i.e. different data, or using a different tool than SQL etc.)?**

No, we did not succeed in answering the question. Even though we have a solution code for this question, we might have been successful if we were using a different dataset that includes that age of people who have been vaccinated.

**c. Which SQL tool(s) did you use to solve the question (i.e. aggregates, joins, cases etc.)?**

We used an aggregate function to find the average amount of vaccinations for each age group. We also used the group by statement and order by statement. By doing that, we are able to list the age groups according to the average vaccinations ordered from highest to lowest. We also used a scalar function, specifically round(), which helped us round and limit the average amount of vaccinations to the hundredths place.

**Our Attempts:**

1 - which areas have lower/higher access to vaccines based on income?

create table vaccination\_rates(

age integer,

name varchar,

vaccination\_status varchar,

income integer,

state varchar

);

\copy vaccination\_rates(age, name, vaccination\_status, income,state) FROM 'C:\Users\Salma\OneDrive\Desktop\Comp Sci- 232 (Spring Term 2021)\our-world-in-data-vaccination-coverage-by-income.csv' DELIMITER ',' CSV HEADER

SELECT

AVG(ALL vaccination\_status),

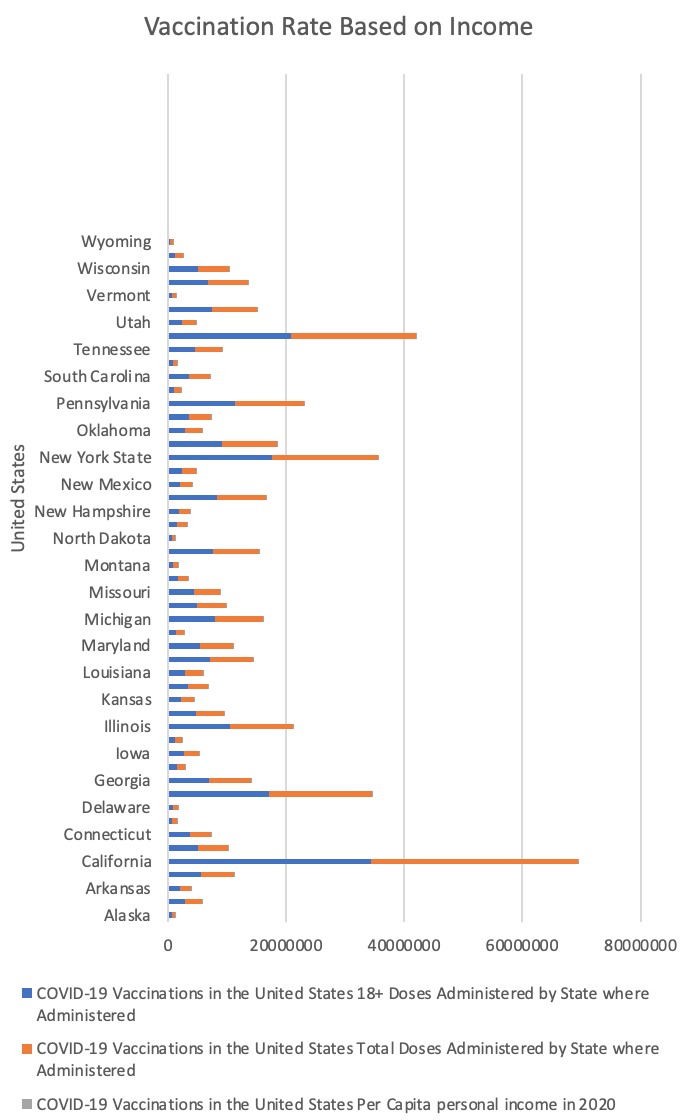
avg\_vaccination\_rate

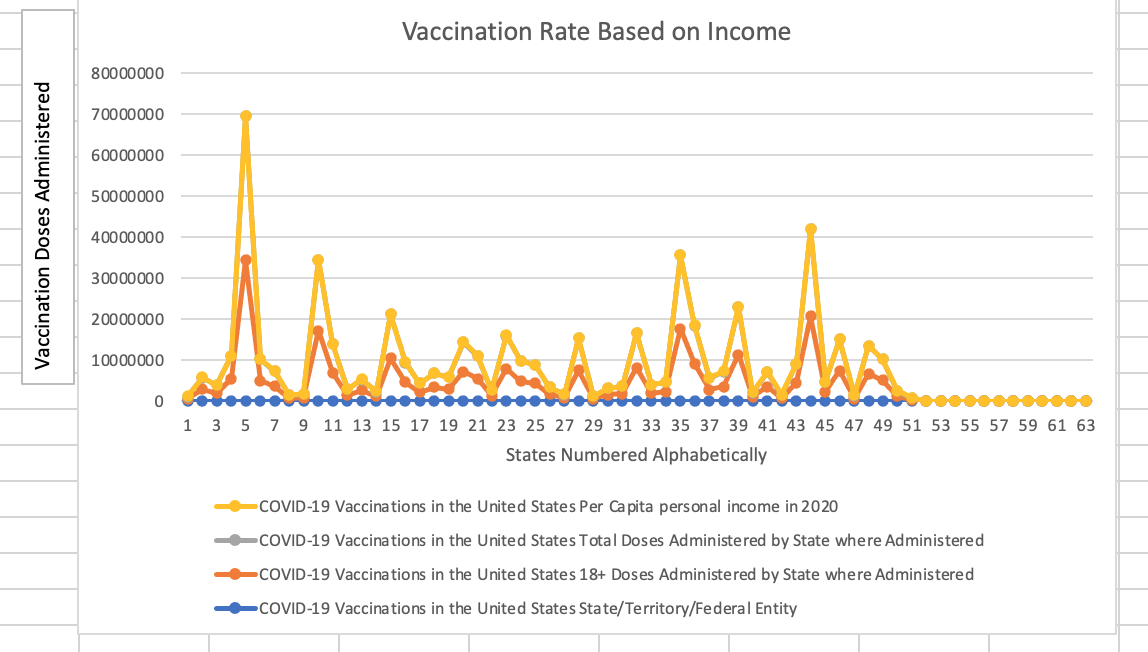
FROM

vaccination\_rates

GROUP BY

income, state;





2 - Which day had the lowest vaccination record and in which state?

create table us\_vaccination(

entity varchar,

code varchar,

day date,

daily\_vaccinations int

);

\copy us\_vaccination(entity, code, day, daily\_vaccinations) FROM 'C:\Users\Melanie\OneDrive\Desktop\Comp Sci- 232 (Spring Term 2021)\us-daily-covid-vaccine-doses-administered.csv' DELIMITER ',' CSV HEADER

--FIRST SOLUTION

select entity, daily\_vaccinations, day

from us\_vaccination

group by entity, day, daily\_vaccinations

order by daily\_vaccinations, entity asc limit 1;

--SECOND SOLUTION

select entity, daily\_vaccinations, day

from us\_vaccination

where daily\_vaccinations >= 0

group by entity, day, daily\_vaccinations

order by daily\_vaccinations, entity asc limit 1;

3 - which areas have been/not been vaccinated based on political affinity?

ALTER TABLE vaccination\_rates

ADD political\_affiliation varchar:

\copy vaccination\_rates(age, name, vaccination\_status, income,state,political\_affiliation) FROM 'C:\Users\Salma\OneDrive\Desktop\Comp Sci- 232 (Spring Term 2021)\covid-cdc-gov-covid-data-tracker.csv' DELIMITER ',' CSV HEADER

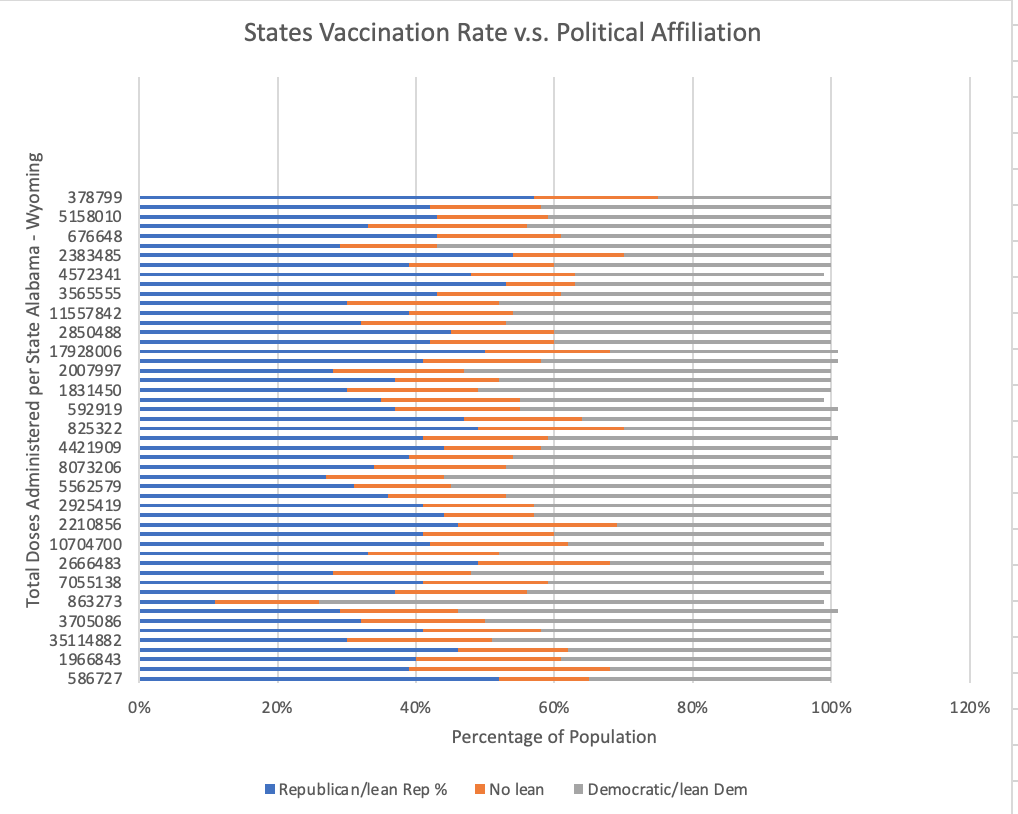
SELECT vaccination\_rates, political\_affiliation

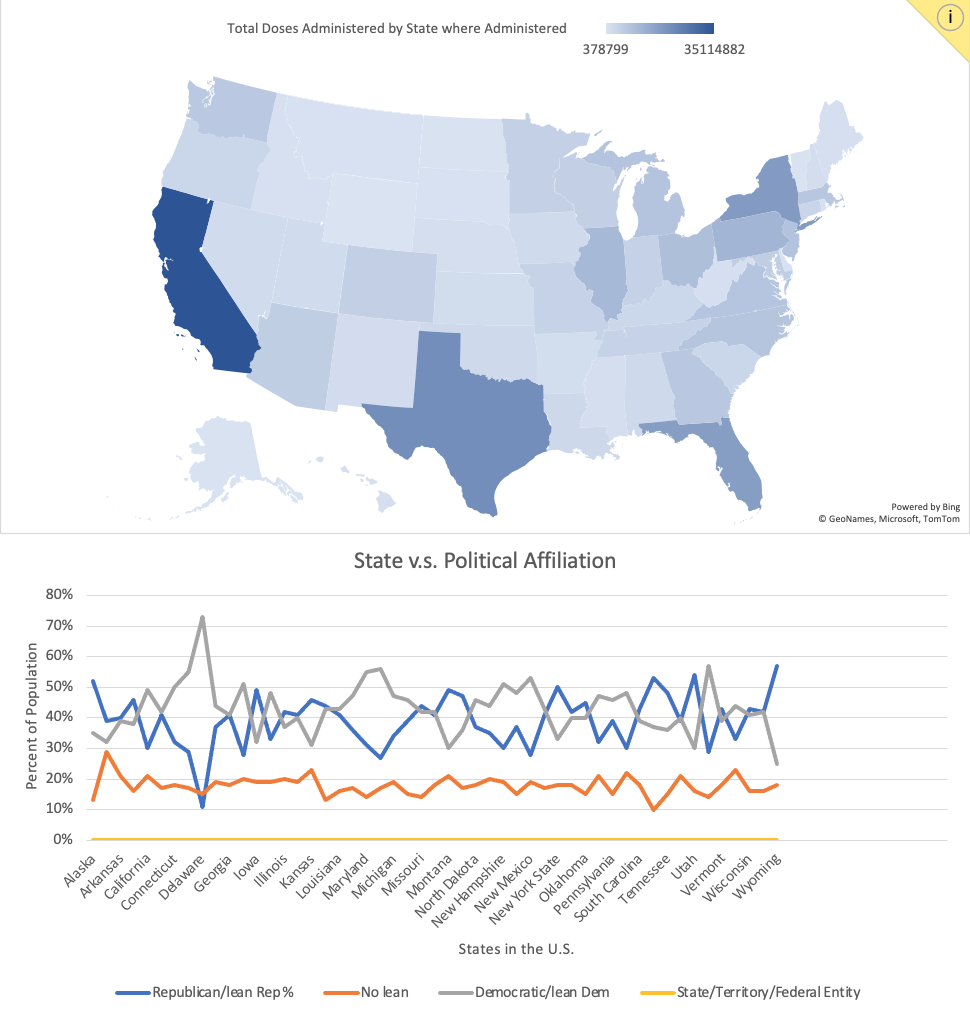
FROM table\_name

WHERE democratic OR republican OR third\_party OR none\_of\_the\_above

GROUP BY

political\_affiliation, state;





4- how many vaccine hub per area to find out if everyone has access to vaccines

5 - Which state has the most vaccinations up to this day (May 12, 2021) according to the dataset?

create table us\_vaccination(

entity varchar,

code varchar,

day date,

daily\_vaccinations int

);

\copy us\_vaccination(entity, code, day, daily\_vaccinations) FROM 'C:\Users\Melanie\OneDrive\Desktop\Comp Sci- 232 (Spring Term 2021)\us-daily-covid-vaccine-doses-administered.csv' DELIMITER ',' CSV HEADER

--FIRST SOLUTION

select entity,

round(sum(daily\_vaccinations), 2) as sum\_vaccine

from us\_vaccination

group by entity

order by sum\_vaccine desc limit 1;

--SECOND SOLUTION

select entity,

round(sum(daily\_vaccinations), 2) as sum\_vaccine

from us\_vaccination

where entity != 'United States'

group by entity

order by sum\_vaccine desc limit 1;

6- how can we encourage more people to get the vaccine?

We can use our overall analysis of data. As you can see from our queries of the data set we can put faith within the American people that the usage of vaccines has substantially aided in the decrease of covid cases. This is in addition to the fact that we do not see any adverse effect of the vaccine as more Americans continue to receive it. Overall, the fact that we have covid deaths, but not covid vaccine deaths should be reason enough for people to get vaccinated.

7- Which day had the highest vaccination record in New York?

create table us\_vaccination(

entity varchar,

code varchar,

day date,

daily\_vaccinations int

);

\copy us\_vaccination(entity, code, day, daily\_vaccinations) FROM 'C:\Users\Melanie\OneDrive\Desktop\Comp Sci- 232 (Spring Term 2021)\us-daily-covid-vaccine-doses-administered.csv' DELIMITER ',' CSV HEADER

select entity, daily\_vaccinations, day

from us\_vaccination

where entity = 'New York State'

group by entity, day, daily\_vaccinations

order by daily\_vaccinations desc limit 1;

8- Which entity in the dataset should have an increase of access to vaccinations?

create table us\_vaccination(

entity varchar,

code varchar,

day date,

daily\_vaccinations int

);

\copy us\_vaccination(entity, code, day, daily\_vaccinations) FROM 'C:\Users\Melanie\OneDrive\Desktop\Comp Sci- 232 (Spring Term 2021)\us-daily-covid-vaccine-doses-administered.csv' DELIMITER ',' CSV HEADER

select entity,

round(avg(daily\_vaccinations), 2) as avg\_vaccine

from us\_vaccination group by entity order by avg\_vaccine asc;

9. How many Americans are currently fully vaccinated as of May 12th in the United States?

create table us\_vaccination(

entity varchar,

code varchar,

day date,

daily\_vaccinations int

);

\copy us\_vaccination(entity, code, day, daily\_vaccinations) FROM ‘'C:\Users\Melanie\OneDrive\Desktop\Comp Sci- 232 (Spring Term 2021)\us-daily-covid-vaccine-doses-administered.csv' DELIMITER ',' CSV HEADER

SELECT entity, sum(daily\_vaccinations)

FROM us\_vaccination

where entity = 'United States'

group by entity;

10. How many people under 18 have been vaccinated ?

create table us\_vaccination(

entity varchar,

code varchar,

day date,

daily\_vaccinations int,

age int

);

SELECT age, COUNT(daily\_vaccinations)

FROM us\_vaccination

WHERE age < 18

group by age;

11. What is the average rate of vaccination per day?

create table us\_vaccination(

entity varchar,

code varchar,

day date,

daily\_vaccinations int

);

\copy us\_vaccination(entity, code, day, daily\_vaccinations) FROM 'C:\Users\Melanie\OneDrive\Desktop\Comp Sci- 232 (Spring Term 2021)\us-daily-covid-vaccine-doses-administered.csv' DELIMITER ',' CSV HEADER

select day,

round(avg(daily\_vaccinations), 2) as avg\_vaccine

from us\_vaccination

group by day

order by day asc, avg\_vaccine;

12. Which age group is more likely to get vaccinated?

create table us\_vaccination(

entity varchar,

code varchar,

age int,

day date,

daily\_vaccinations int

);

select entity, age,

round(avg(daily\_vaccinations), 2) as avg\_vaccine

from us\_vaccination

group by age

order by avg\_vaccine desc, age;