1. Part 1 - Data Loading
    1) Manually Upload files - required for the project: Cases, Vaccine, Vaccine P, Timeseries (Time table), Population
    2) Stack file based on year
    3) Spark

2. Part 2 - Database Design
    1) Create and clean the “Cases” Table
    2) Create and clean the “Times” Table
    3) Create and clean the “VaccineP” Table
    4) Create and clean the “Vaccine” Table
    5) Merge the “Vaccine” table and “VaccineP” table into the “Vaccine”

        * VaccineP Table: About people vaccinated in the US with timeline (ex, people_fully_vaccinated, people_partially_vaccinated)

        * Vaccine Table: About vaccine data in the US with timeline (ex, vaccine_type, doses_alloc, doses_shipped, doses_admin, stage_one_doses, stage_two_doses)
    6) State Dimension
        1) Create and add data to “State” table - with columns “State_ID”, “FIPS”, “Province_state”, “latitude”, “longitude”
        2) Add a foreign key constraint to a table: State_ID Foreign Key for 3 tables (“cases”, “times”, “vaccine”)
    7) Time Dimension
        1) Create and add data to the “date” table - with columns “Date_ID”, “hour”, “day”, “year”, “month_of_year_str”, “monthe_of_year”, “day_of_month”, “day_of_week_str”, “day_of_week”, “hour_of_day”, “quarter_of_year”
        2) Add a foreign key constraint to a table: Date_ID Foreign Key for 3 tables (“cases”, “times”, “vaccine”)
    8) Check Final Tables created - “cases”, “vaccine”, “state”, “date”, “times”

3. Part 4 - Questions

(Q1) What’s the percentage of people who are affected by COVID based on the state?

(Q2) What are the “confirmed vs recovery vs death” cases by the state?

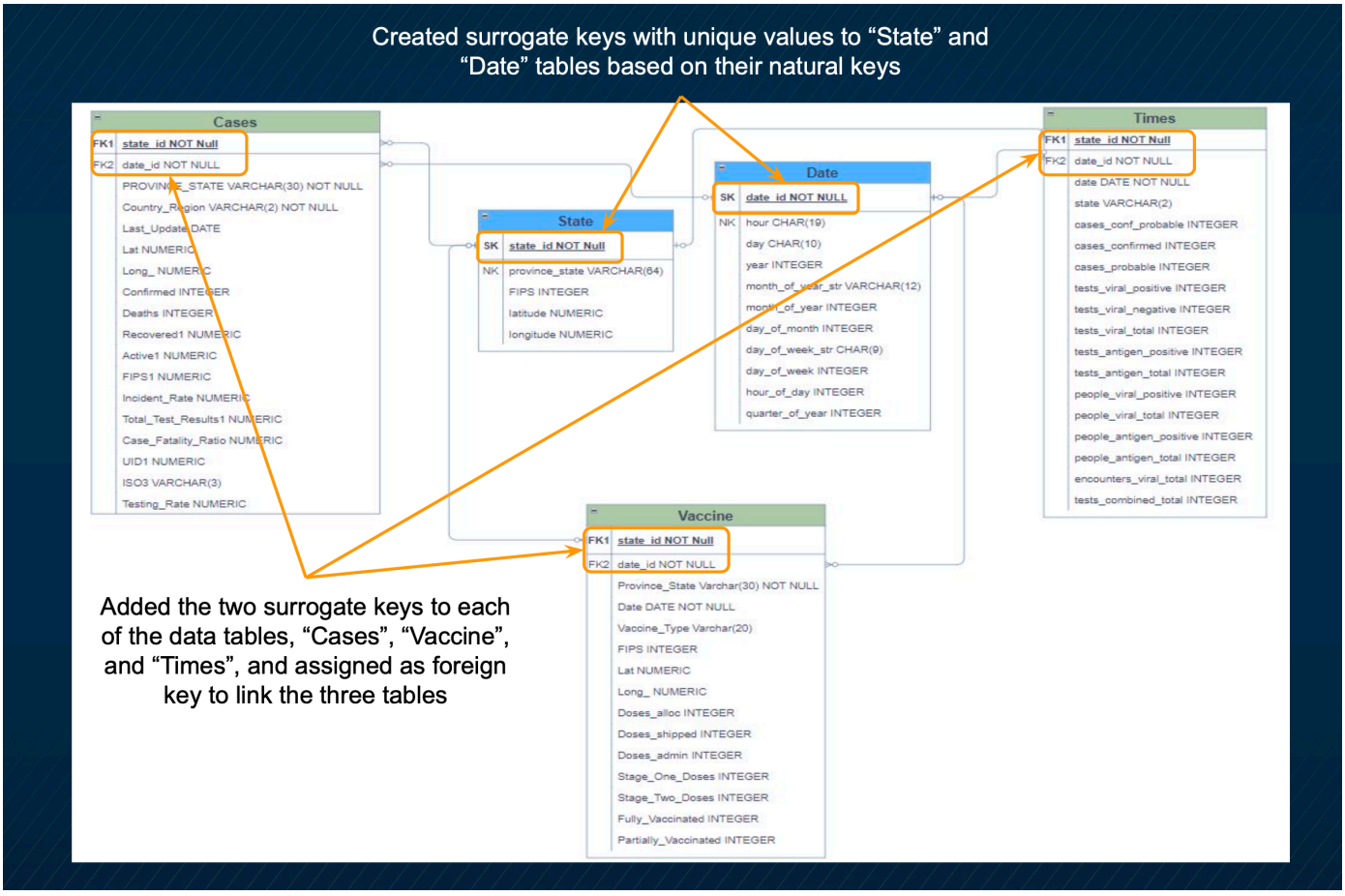
(Q3) Do vaccines have an actual effect on covid cases? (by location, by year)

(Q4) Do testing rates affect confirmed cases? (by state)

(Q5) Were there any changes in confirmed cases before/after restrictions?

4. Part 5 - Dimensional Model

- **Star Schema**



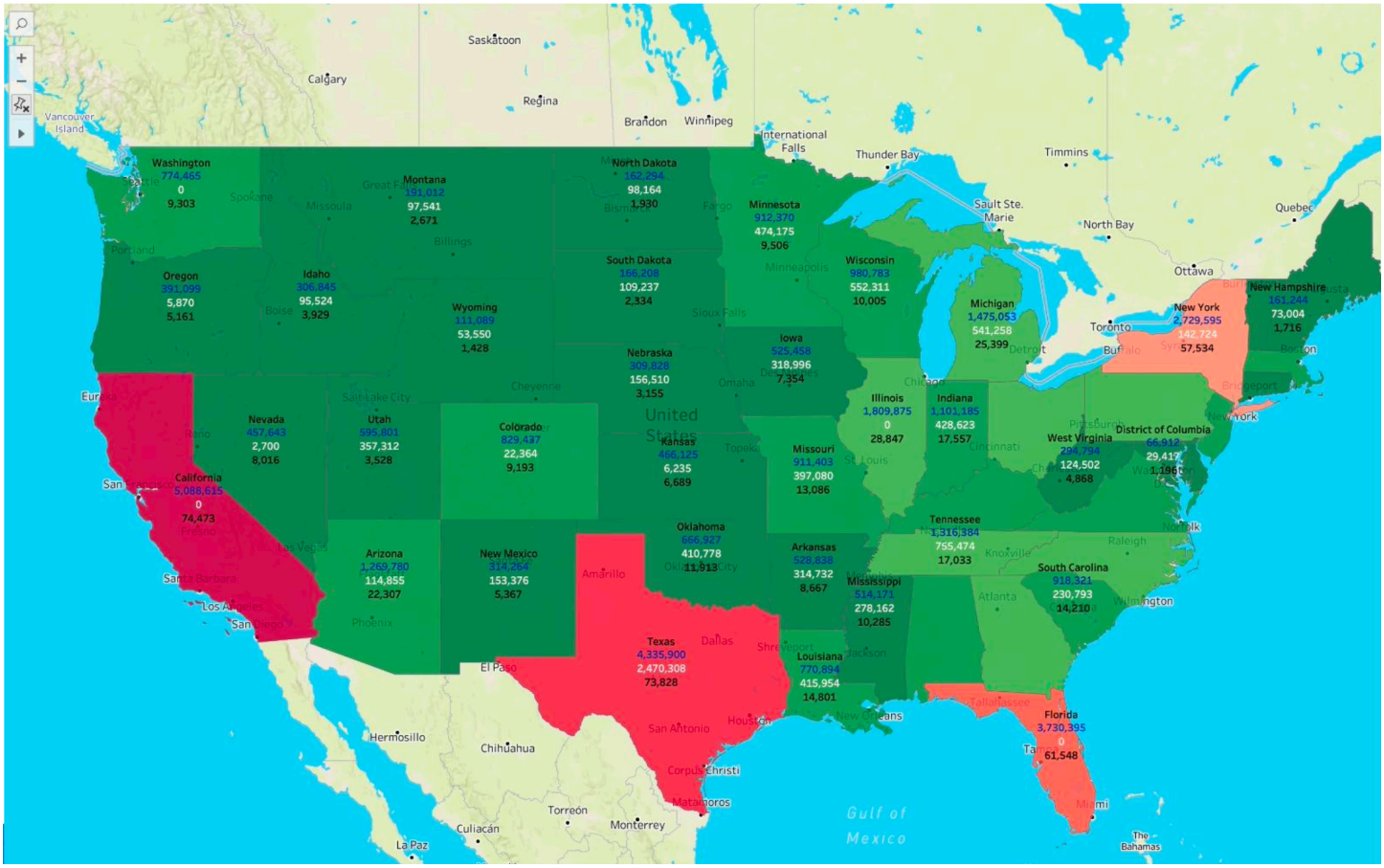
5. Part 6 - Queries to answer questions

Q1) What’s the percentage of people who are affected by COVID based on the state?

fips1	province_state	population	total_cases_confirmed	percentage
38	North Dakota	762062	162294	21.2967
2	Alaska	731545	150996	20.6407
47	Tennessee	6829174	1316384	19.2759
56	Wyoming	578759	111089	19.1943
46	South Dakota	884659	166208	18.7878
49	Utah	3205958	595801	18.5842
44	Rhode Island	1059361	191763	18.1018
30	Montana	1068778	191012	17.8720
45	South Carolina	5148714	918321	17.8359
21	Kentucky	4467673	785926	17.5914

As we can see from the above table North Dakota and Alaska have higher covid affected people since they are states with lower populations under 1 million people. All top 10 states with the highest covid percentage of covid confirmed cases besides Tennessee are all under 5 million population.

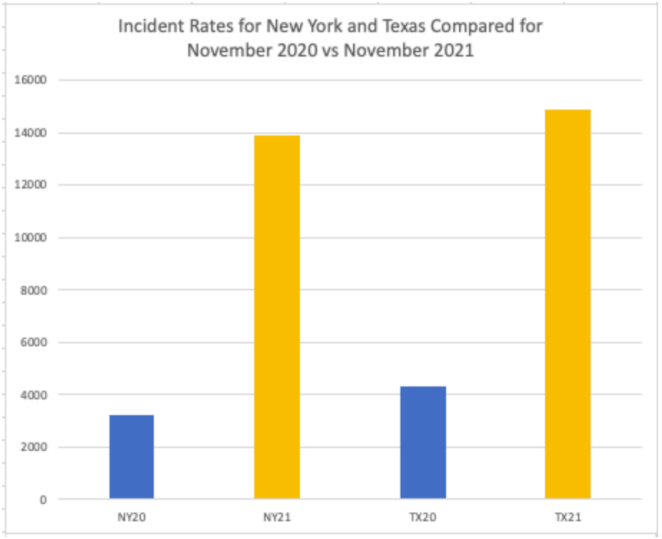
Q2. What is the Confirmed vs Death vs Recovery Cases by the state?

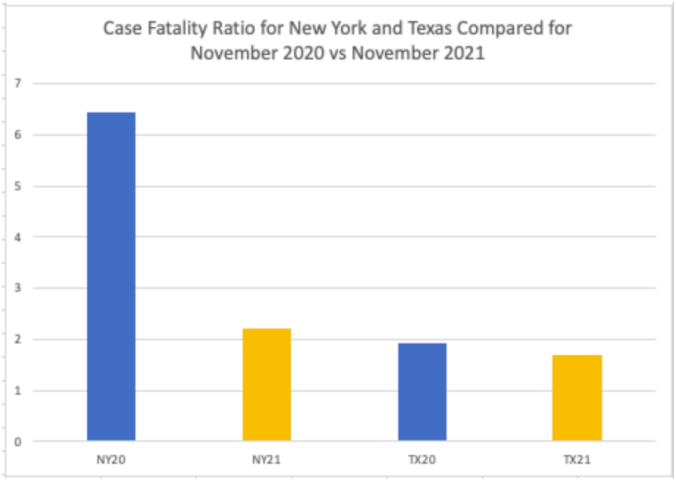


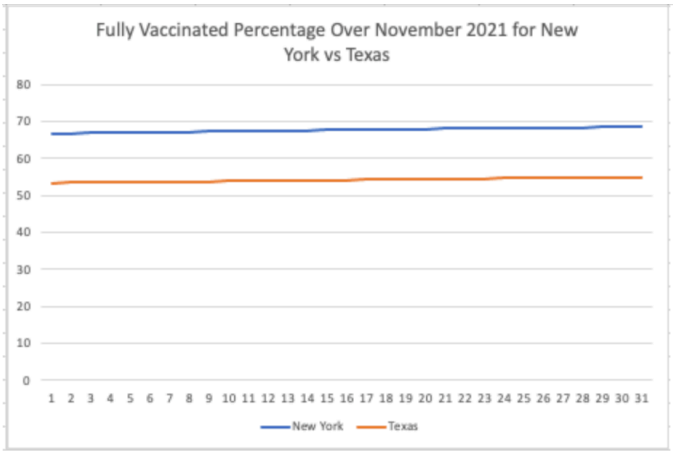
- From the heat map in the graph above, we see the highest confirmed cases (blue text) in California then Texas followed by Florida.
- We can also quickly glance at recovered (white text) and notice that California, Florida, Washington, and Illinois are 0. Which could mean there is some missing data in our data set.
- Overall, the North West side of the USA had fewer cases than the rest of the MAP.
- We observe that the highest cases recorded were in populated states which are located near the coast.
- During the Pandemic global sea freight charges escalated to prices like never before. It is interesting to see that covid cases were highest near major ports in the USA. Due to fewer people coming to work and safety precautions it could have caused delays at the port, leading to over-congestion to this day.

Q3. Do vaccines have an actual effect on covid cases? (by location, by year)

We are going to look at two opposite (in terms of vaccine and testing outlook) states, New York and Texas as two locations to compare year on year after the vaccine rollout.
The date is selected for year-on-year analysis, pre-vaccine approval, and rollout but also needed to be “early” enough that we had data to look at for this month/year. The vaccine was approved by the FDA on December 11, 2020, however, we cannot use data for December 11 2021 so shifted our analysis to the month of November for 2020 and 2021.







1) Average incident rate and Average case fatality rate

First, we looked at the average incident rate and average case fatality rate before vaccination in New York and Texas in November 2020, then we looked at the average incident rate and average case fatality ratio after vaccination in November 2021 for the two states.

(Results) The results of the above analysis are as follows.

- 2021 analysis shows that for the month of November 2021, New York had an average accident rate of 13,887.21 and an average fatality rate of 2.2. It is much lower than last year, which could indicate that the vaccine is actually having a positive impact.
- Similarly, we can see that the average accident rate in Texas in November 2021 is 14,900.34 and the fatality rate is 1.7, while the incident fatality rate is much higher in Texas than the same month in 2020, but the fatality rate is low.

2) Comparison of average number of confirmed cases and average number of deaths for vaccination status
Next, we looked at the number of vaccinations in the two weeks of November 2021, and compared the average number of confirmed cases and the average number of deaths in November in the two years 2020 and 2021. Next, after looking at the number of vaccinations in the two weeks in November 2021, we compared the average number of confirmed cases and the average number of deaths in November in both years. This is with the knowledge that vaccinations are at zero for November 2020. So we will look at the fully vaccinated and partially vaccinated numbers compared to the state populations to find a percentage of fully vaccinated status, firstly looking at New York and then Texas.

(Results) After looking at the fully vaccinated percentage of New York and Texas in November 2021, it appeared as follows.
- First, New York's higher vaccination rates than Texas also result in much lower death rates when comparing the two states. Overall, New York's overall vaccination status averaged 67.69%, and when accounting for the population, the fatality rate fell from 6.44 to 2.2 from November 2020 to November 2021. This is a positive effect of vaccination, which has the effect of reducing mortality.
- Texas is the same but on a smaller scale. The average complete vaccination rate fell from 1.91 to 1.7 from November 2020 to 2021, with an average complete vaccination rate of about 54.18%, lower than that of New York State.
- Overall accident rates have skyrocketed, but that's to be expected as we learn how to deal with and live with the pandemic while we get vaccinated out of it. The real metric we're looking for to judge a vaccine's effectiveness is lethality.

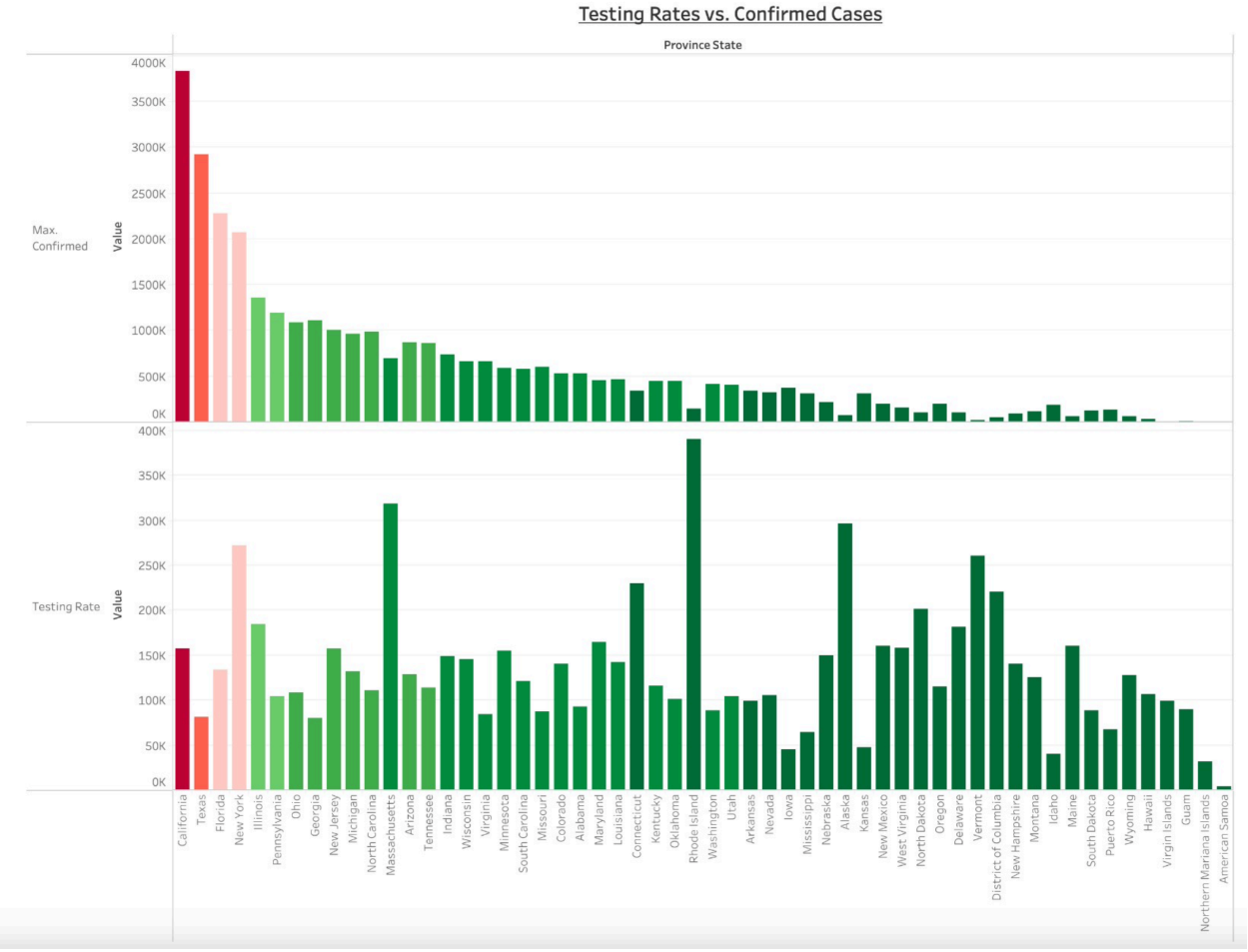
Q4) Do testing rates affect confirmed cases by the state?

To answer this question in a fully comprehensible form, we manipulated the data from different perspectives to come to the following conclusions: Testing rates are impacting confirmed cases on a state-by-state basis. We added the population to the data evaluation because the testing rate is directly related to the population of each state.

1. Process

We first identified the top 10 states with the highest testing rates and looked at their average testing rates. Next, we identified the top 10 states with the lowest testing rates and looked at their average testing rates. To provide a deeper understanding of the data, we decided to correlate testing rates with confirmed cases and find the mean of those correlations. We looked at the average testing rates of the top 10 states with the fewest confirmed cases and the average testing rates of the top 10 states with the most confirmed cases. And finally, we looked at the average test rates for all states.

2. Results



- The top 10 states with the highest testing rates are primarily occupied by the Northeastern states. With an estimated percentage of 568,595.97, Rhode Island leads the way, while Massachusetts and New York also top the list. We also can see the notorious difference between the value of the average inspection rate (399711.72) and Rhode Island's inspection rate (389,907.22). This can be explained by the exponential spread of the virus in New York, New Jersey and Rhode Island in late 2020.

- Looking at the top 10 states with the lowest testing rates, we can conclude that the states in the middle are less likely to be tested, as they mainly show states like Iowa, Idaho, and Kansas. We can also see the notorious disparity between the two states, with high and low testing rates.
- The average testing rate value for the top 10 states with the fewest confirmed cases is 225,534.99, which is higher than the average testing rate provided in the calculation above. In addition, the average inspection rate of the top 10 states with the most confirmed cases was 211,926.65, which is lower than the opposite value, but it can be judged that there is no significant difference in inspection rates for both the least confirmed cases and the most confirmed cases.
- The final calculation was to make sure that the inspection rates do not change significantly with the number of confirmed rates. Therefore, it can be concluded that the number of confirmed cases by country is not directly related to the testing rate.

Q5. Were there any changes in confirmed cases before/after restrictions?

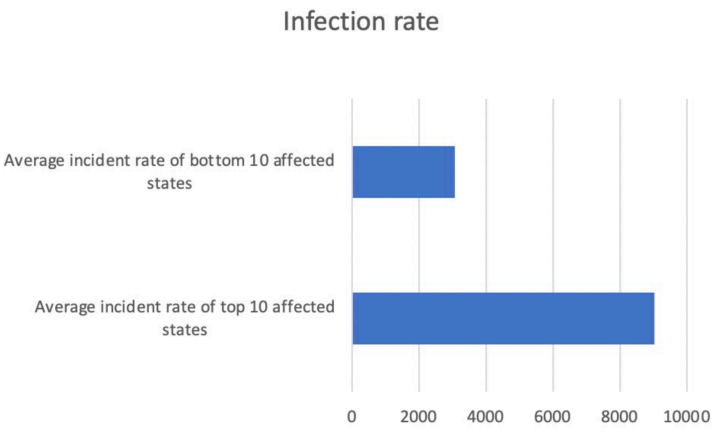
(All figures are as of December 31, 2020 (Date_Id 354) and the Incident rate is the total confirmed cases per 100,000 population)

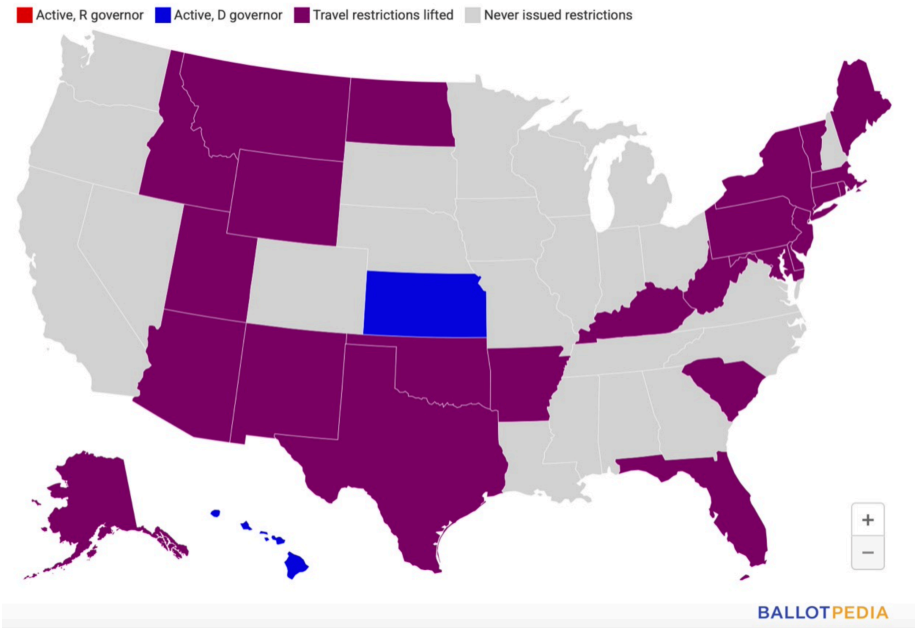
1. Process

First, We identified the top 10 states with the lowest and the top 10 highest incidence rates and examined the average incident rate for each. We then looked at the mean incidence rates in states with restrictions and those without restrictions.

2. Results

- The states with low incident rates (confirmed cases per 100,000 people) are Vermont, Hawaii, Maine, Oregon, New Hampshire, Washington, Virginia, District of Columbia, Maryland, and West Virginia. And their average accident rate is about 3085.7. Ten states have some sort of travel restriction. However, Oregon, New Hampshire, Washington, and Virginia had no travel restrictions. So it's safe to say that some sort of travel ban has reduced the number of COVID-19 cases in the first few months of the pandemic.
- The states with high incident rates(confirmed cases per 100,000 people) are North Dakota, South Dakota, Iowa, Wisconsin, Nebraska, Tennessee, Utah, Rhode Island, Idoha, and Kansas. Their average incidence is about 9039 cases. Only 4 out of 10 states have any travel restrictions in place. States with travel restrictions are North Dakota, Utah, Rhode Island, and Idaho. So it's safe to say that a travel ban of some sort reduced the number of COVID-19 cases in the first few months of the outbreak. The reason for the high incidence even with travel restrictions in place may be due to the high population density of the East Coast.
- Graphical representation comparing the top 10 and bottom 10 states as per average incident rate





- The overall data of states with restrictions shows an average incident rate of around 8130. Also, the overall data of states with no restriction have a higher average incident rate of around 8282. However, the small difference in incidence rates between states with and without restrictions suggests a small correlation between travel restrictions and COVID-19 infection rates.