CS1951A Capstone: Covid-19 Vaccination and Outcomes During the Covid-19 Pandemic

Introduction

Our team was interested in the impact of Covid-19 vaccination rates in the United States on the number of cases and number of deaths during the Covid-19 pandemic. We wanted to explore whether there was a difference in number of new cases before and after vaccines were made available; whether there was a difference in case and death rates for states that are otherwise similar but have significantly different vaccination rates; and whether we could predict weekly cases given the previous weekly cases, number of vaccinations, total cases, and state demographic information.

Data

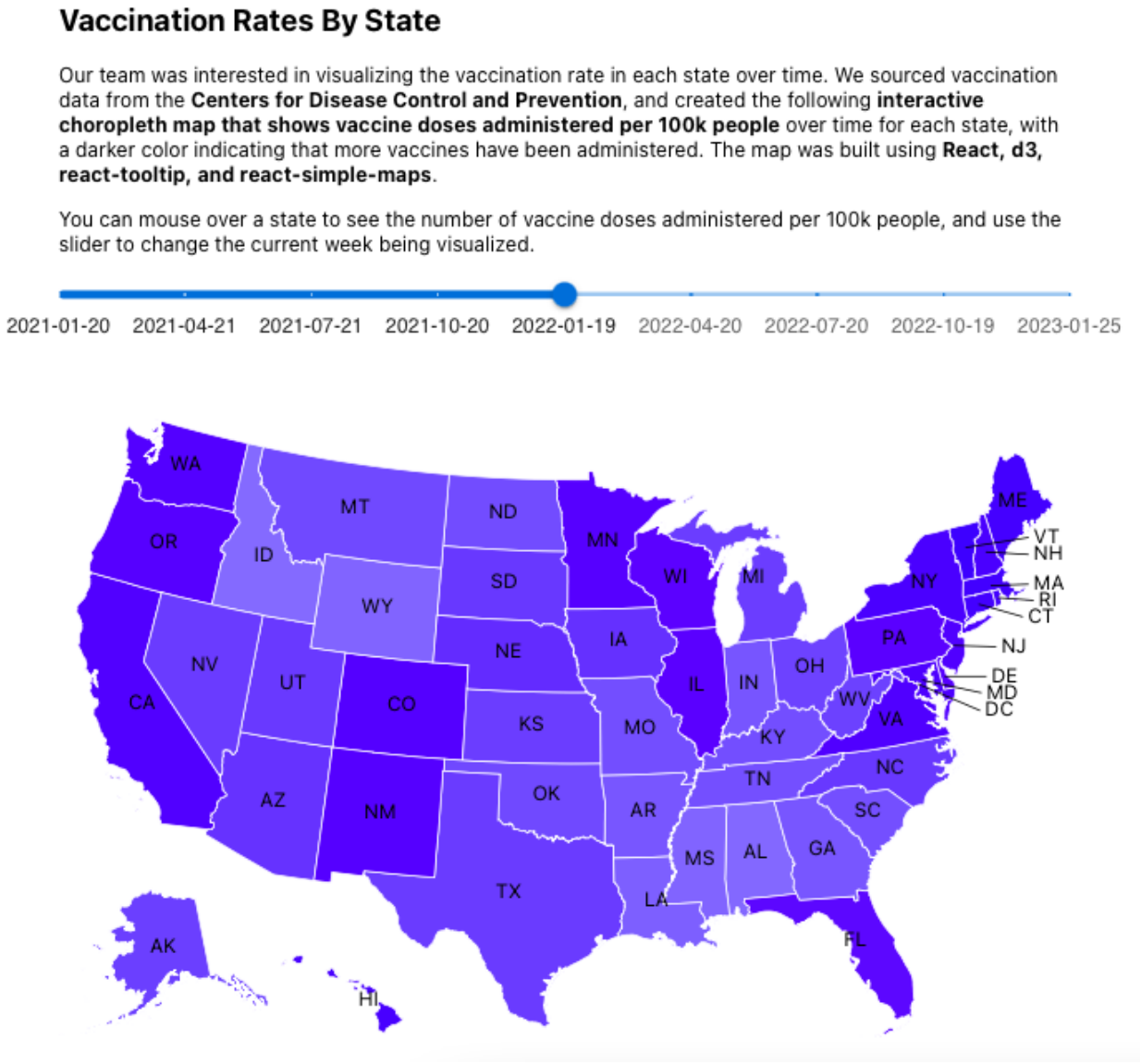
Our analysis required the use of several different datasets. Our most important dataset included data about case, vaccination, and death rates by US state for each week from late 2020 to early 2023. We sourced this data from the Centers for Disease Control and Prevention. To identify similar states, we used a dataset that included age distribution, race/ethnicity distribution, income distribution, and other demographic characteristics for each state. We sourced this data from the American Community Survey.

Findings

- We first performed a paired t-test to test the null hypothesis that the growth in the number of new weekly cases per 100k people before vaccines were made available is the same as the mean growth in the number of new weekly cases per 100k people after vaccines were made available. We rejected the null hypothesis.
- Next, we performed a k-means clustering analysis on US states to identify groups of similar states by age, income, educational attainment, etc. Using the results of the k-means analysis, we identified two states that have similar demographics but significantly different vaccination rates: Texas and California.
- We then used a paired t-test to test the null hypotheses that the mean number of weekly cases/deaths per 100k people in California is the same as the number of mean weekly cases/deaths per 100k people in Texas. We rejected the null hypotheses.
- Next, we used linear regression to see we could predict weekly cases given the previous weekly cases, vaccinations, and total cases. We are not very confident that we can accurately predict weekly cases with this information; our model's performance is better than chance, with test MSE 19,072. We then built a more advanced linear regression model that incorporated state demographic data, which performed better, with test MSE 11,328.

Interactive Component

We built an interactive web app using React that allows users to visualize vaccination rates for each state by week using a choropleth map; visualize cases for each state by week using a choropleth map; and to enter input parameters and run our linear regression model to predict next week's cases for a given state.



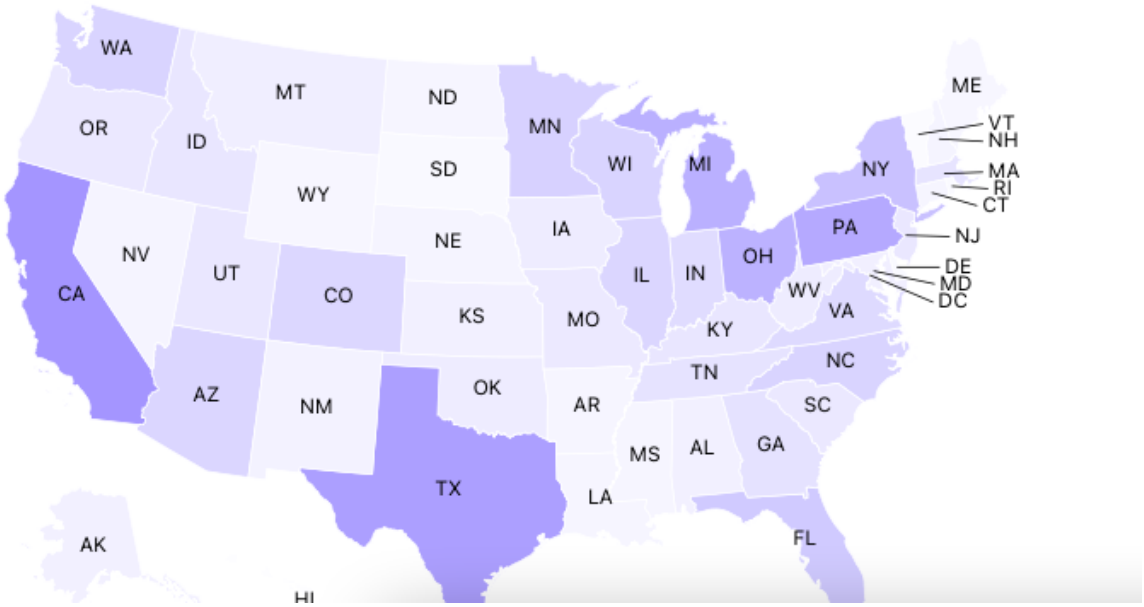
Weekly Cases By State

Our team was also interested in visualizing the weekly cases in each state over time. We sourced weekly case data from the **Centers for Disease Control and Prevention**, and created the following **interactive choropleth map that shows weekly cases** over time for each state, with a darker color indicating more cases for that week. The map was built using **React, d3, react-tooltip, and react-simple-maps**.

You can mouse over a state to see the number of weekly cases, and use the slider to change the current week being visualized.

One difficulty we faced with this particular visualization is the significant uptick in cases for the week of 2022-01-19. If we used the maximum for this week as an upper bound for our color scale, this would mean most weeks wouldn't have noticeable differences because the color range is so compressed. Instead, we opted to use a maximum based on other weeks and allow for compression for high values for the week of 2022-01-19.

2021-01-20 2021-04-21 2021-07-21 2021-10-20 2022-01-19 2022-04-20 2022-07-20 2022-10-19 2023-01-25



Linear Regression For Predicting Weekly Cases

Finally our team was interested in the impact of Covid-19 vaccination rates in the United States on number of cases and number of deaths during the Covid-19 pandemic. In this analysis, we wanted to explore whether we could predict weekly cases given the previous weekly cases, vaccinations, total cases, and state demographic information

We sourced vaccination data from the **Centers for Disease Control and Prevention** and state demographic information from the **American Community Survey**, cleaning and processing the data using **Pandas** before performing linear regression using **scikit-learn**.

Using the below calculator built in **React** with **Material UI**, you can input a state, vaccines administered per 100k people, this week's cases, and total cases to run our regression model to predict next week's cases.

Please note that this is a simple model and results may be inaccurate or nonsensical (e.g. predicting fewer than zero cases).