COVID-19 Hospital Data:

Data Analysis and Visualizations using Python and R

Team Polaris: Caroline Wills, Militha Madur, Beth Gronski, & Kalkidan Tamirat

Table of contents

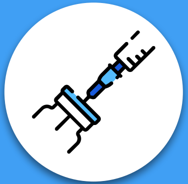
Introduction

For our analysis, we chose the Health prompt, which was to better leverage COVID-19 data of hospitals in the US by answering analytical questions that sparked our interest as data scientists.

Data information

Author

This data was gathered by HealthData.gov, and compiled from the U.S. Department of Health and Human Services as well as state partners



Dates

This dataset provides hospital COVID-19 information from Dec. 2019 to Feb. 2023. In our analysis we are focusing on data from 2021 on.

Location

This data is collected from the US, and US territories. In some analyses, we only observe mainland data, but in others we include all territories.

Numbers

In this data, numbers are recorded for the number of hospital beds used by COVID-19 patients, vaccine status of healthcare providers, and other case information.

Limitations and Implications of Data

Limitations:

- Within this dataset, there was lots of NaN values, especially for vaccination status of healthcare providers
- For 4 or less patients, those values were listed as '-999,999', so sums and averages may be off from actual value. Depending on the analysis, we replaced these values with 1 or NaN.

Implications:

- This information brings attention to how COVID is affecting hospitals, especially considering that many in the US have returned to pre-COVID activities.[2]

01
Data Cleaning

How did we clean the data for easier use?

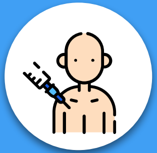
Cleaning the Data in Python

```python
covid = pd.read_csv("COVID-19_Reported_Patient_Impact_and_Hospital_Capacity_by_Facility.csv",
                    usecols=["state", "hospital_pk", "collection_week", "hospital_name", "hospital_subtype",
                             "address", "city", "zip", "fips_code", "is_metro_micro",
                             "total_personnel_covid_vaccinated_doses_none_7_day",
                             "total_personnel_covid_vaccinated_doses_one_7_day",
                             "total_personnel_covid_vaccinated_doses_all_7_day",
                             "total_adult_patients_hospitalized_confirmed_and_suspected_covid_7_day_avg",
                             "total_patients_hospitalized_confirmed_influenza_7_day_avg",
                             "total_patients_hospitalized_confirmed_influenza_7_day_sum",
                             "total_patients_hospitalized_confirmed_influenza_and_covid_7_day_sum",
                             "total_patients_hospitalized_confirmed_influenza_and_covid_7_day_avg",
                             "total_adult_patients_hospitalized_confirmed_and_suspected_covid_7_day_avg",
                             "total_adult_patients_hospitalized_confirmed_covid_7_day_avg",
                             "total_adult_patients_hospitalized_confirmed_covid_7_day_sum",
                             "total_adult_patients_hospitalized_confirmed_and_suspected_covid_7_day_sum",
                             "total_pediatric_patients_hospitalized_confirmed_covid_7_day_sum",
                             "previous_day_admission_adult_covid_confirmed_18-19_7_day_sum",
                             "previous_day_admission_adult_covid_confirmed_20-29_7_day_sum",
                             "previous_day_admission_adult_covid_confirmed_30-39_7_day_sum",
                             "previous_day_admission_adult_covid_confirmed_40-49_7_day_sum",
                             "previous_day_admission_adult_covid_confirmed_50-59_7_day_sum",
                             "previous_day_admission_adult_covid_confirmed_60-69_7_day_sum",
                             "previous_day_admission_adult_covid_confirmed_70-79_7_day_sum",
                             "previous_day_admission_adult_covid_confirmed_80+_7_day_sum",
                             "previous_day_admission_pediatric_covid_confirmed_7_day_sum"])
```
✓ 5.8s

In python, using the pandas package we used the function "usecols" to select which columns we'd like to use for our analysis.

First

Extracting columns for all analysis



Second

Dropping NaN values from selected columns

Third

Individual analysis cleaning for specific visualizations

Cleaning the Data

```
covid.isna().sum()
```

✓ 0.7s

```
Output exceeds the size limit. Open the full output data in a text editor
hospital_pk                                                          0
collection_week                                                      0
state                                                                0
hospital_name                                                        0
address                                                            445
city                                                               445
zip                                                                 20
hospital_subtype                                                     0
fips_code                                                          926
is_metro_micro                                                       0
total_adult_patients_hospitalized_confirmed_and_suspected_covid_7_day_avg    66809
total_adult_patients_hospitalized_confirmed_covid_7_day_avg          66880
total_patients_hospitalized_confirmed_influenza_7_day_avg           162778
total_patients_hospitalized_confirmed_influenza_and_covid_7_day_avg 320111
total_adult_patients_hospitalized_confirmed_and_suspected_covid_7_day_sum    66809
total_adult_patients_hospitalized_confirmed_covid_7_day_sum          66880
total_pediatric_patients_hospitalized_confirmed_covid_7_day_sum      72531
total_patients_hospitalized_confirmed_influenza_7_day_sum           162778
total_patients_hospitalized_confirmed_influenza_and_covid_7_day_sum 320111
previous_day_admission_adult_covid_confirmed_18-19_7_day_sum         86057
previous_day_admission_adult_covid_confirmed_20-29_7_day_sum         96025
previous_day_admission_adult_covid_confirmed_30-39_7_day_sum         95913
previous_day_admission_adult_covid_confirmed_40-49_7_day_sum         95619
previous_day_admission_adult_covid_confirmed_50-59_7_day_sum         94635
previous_day_admission_adult_covid_confirmed_60-69_7_day_sum         93439
...
previous_day_admission_pediatric_covid_confirmed_7_day_sum           77213
total_personnel_covid_vaccinated_doses_none_7_day                   477347
total_personnel_covid_vaccinated_doses_one_7_day                    477001
total_personnel_covid_vaccinated_doses_all_7_day                    476468
dtype: int64
```

We then summed the amount of NaN values for each column, and discovered that there were numerous missing values in our data.

Within group discussion, we debated ways to treat the NaN values and how to go about accurate analysis considering how much information was missing.

The dataset in particular has a total of 742255 rows, and we decided there was enough information left over after dropping the NaN values to continue with our analysis.

We originally debatting splitting the data by analysis, but because this project is meant to be holistic, we decided we would all work with the same version of the data and dropped all the NaN values.

```
covid = covid.dropna()
covid.isna().sum()
```

✓ 1.0s

Cleaning the Data

```
covid.shape
✓ 0.0s

(208883, 31)
```

```
covid.to_csv("covid.csv")
✓ 5.3s
```

The new shape of our dataset still contained 208,883 columns, which we deemed significant in creating an analysis of the data. We also remain very diligent in expressing that data is missing and this is not a representative analysis of every hospital in the dataset.

We then export the new dataset into a csv file and this allowed every group member access to the same data to start with when beginning their individual analysis

Cleaning the Data in R

```
# Getting COVID Data
setwd('/Users/carolinewills/Desktop/COVID Datathon')
covid_data <- read.csv('covid_filtered.csv')

# Filter data for hospital characteristics and confirmed covid-19 hospitalizations
covid_filtered_df <- covid_data %>% select(hospital_subtype, is_metro_micro, city,
total_adult_patients_hospitalized_confirmed_covid_7_day_sum, total_pediatric_patients_hospitalized_confirmed_covid_7_day_sum) %>%
mutate(total_patients_hospitalized_confirmed_covid_7_day_sum = total_pediatric_patients_hospitalized_confirmed_covid_7_day_sum +
total_adult_patients_hospitalized_confirmed_covid_7_day_sum)

# Group data by city of hospital and sum of COVID hospitalizations for adult and pediatric patients
covid_filtered_df_city <- covid_filtered_df %>% group_by(city) %>%
summarise(total_patients_covid_hospitalized_sum=sum(total_patients_hospitalized_confirmed_covid_7_day_sum),
total_adult_patients_covid_hospitalized_sum=sum(total_adult_patients_hospitalized_confirmed_covid_7_day_sum),total_pediatric_pati
nts_covid_hospitalized_sum=sum(total_pediatric_patients_hospitalized_confirmed_covid_7_day_sum))

# Filter COVID hospitalizations by Seattle, New York, Chicago, Los Angeles, and Atlanta
covid_filtered_df_select_cities <- covid_filtered_df_city %>% filter(city %in% c('SEATTLE', 'NEW YORK', 'LOS ANGELES', 'CHICAGO',
'ATLANTA'))
```

After the data was cleaned in Python, depending on the individual analysis, we grouped the data depending on the trends we were analyzing.

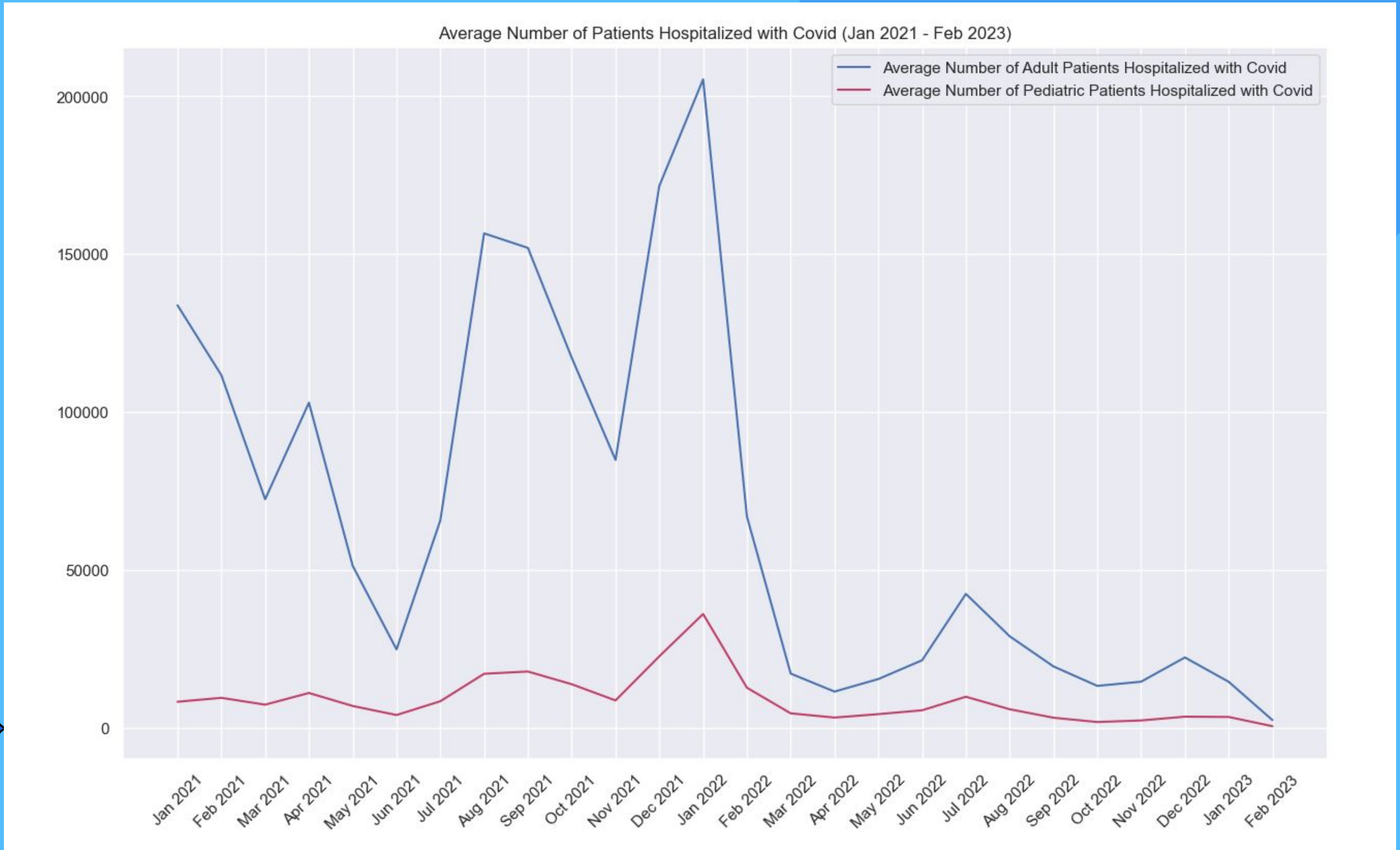
In order to visualize trends in COVID-19 cases per age group and city, we calculated aggregate sums of confirmed case totals per each independent variable.

02

What are the trends in adult and pediatric COVID-19 Cases?

How we track the cases of COVID for adults and pediatric patients overtime

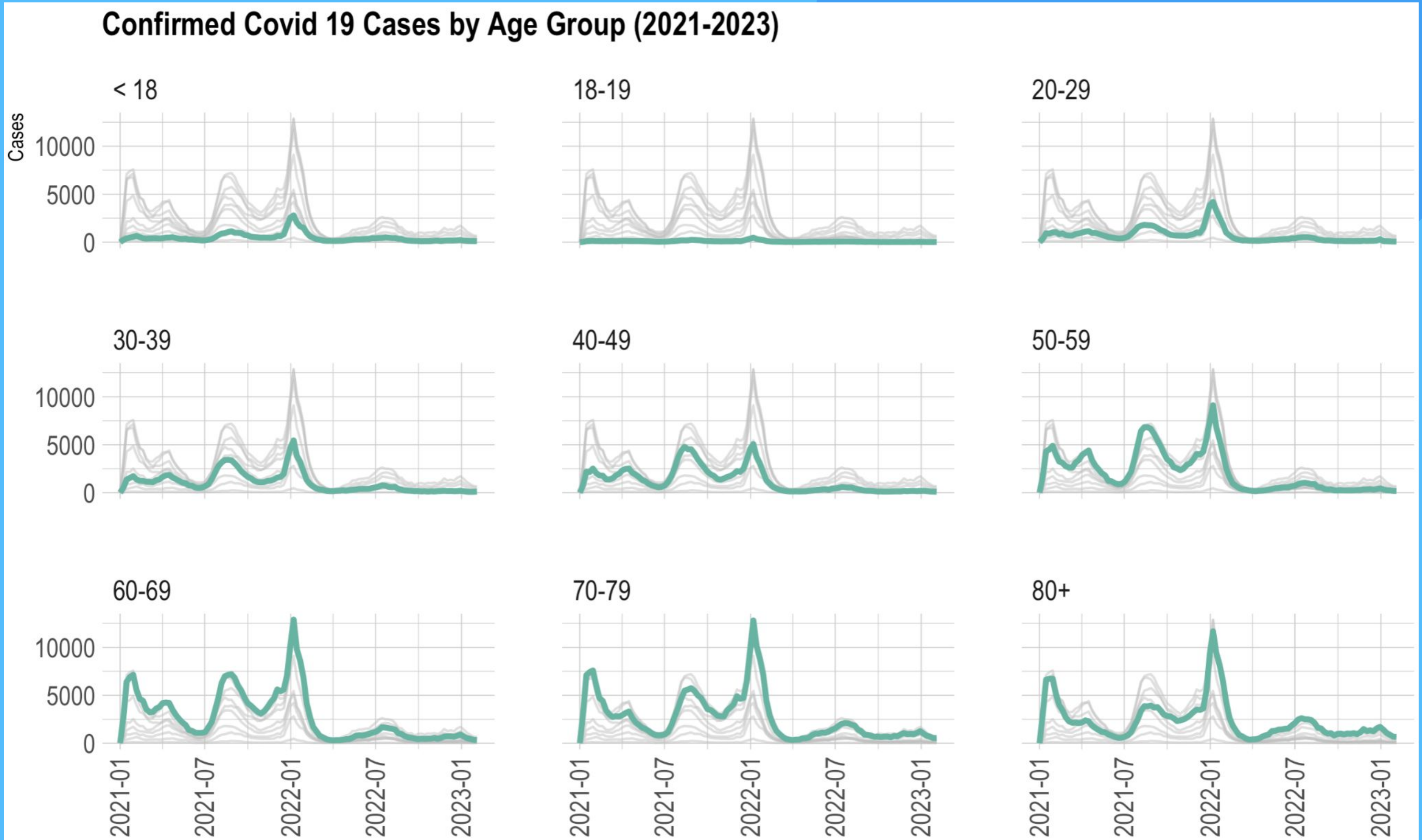
Average Number of COVID Cases Overtime



In this graph, we track the average number of COVID cases for adults(blue) and pediatric (pink) patients from January 2021 to February 2023.

Using the data from the Health Services that takes the average of 7-day reported COVID cases, we combine the available data under their respective month and year.

Total COVID Cases per Age Group



Key Insights:

Highest number of confirmed COVID cases across all age groups occurred during the Week of Jan 07 2022

Max Confirmed Cases per Age group:

Pediatric: 2795
18-19: 470
20-29: 4194
30-39: 5450
40-49: 5069
60-69: 12884
70-79: 1279
80+: 1279

03

What were the total COVID hospitalizations in different cities in the US from 2021 to 2022?

COVID Hospitalizations in Select Cities from January 2021 to September 2022

Key Insights:

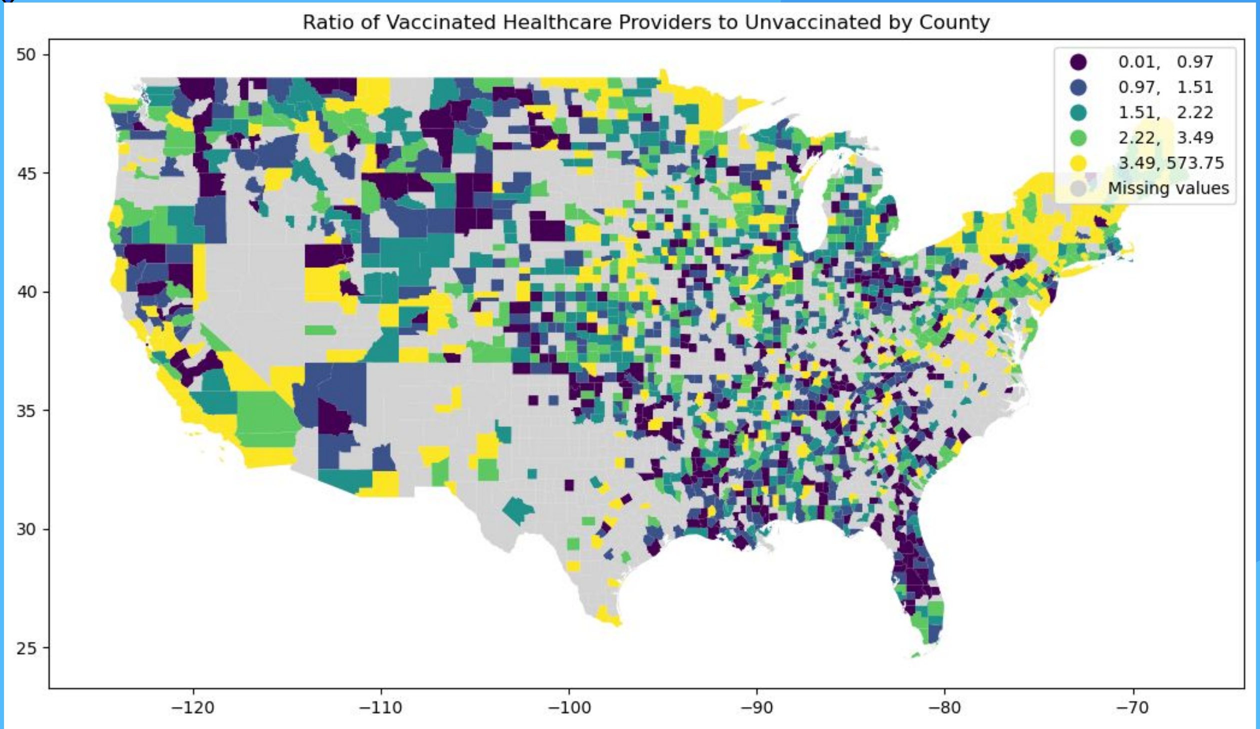
- From this visualization, the total number of COVID cases reported by the HHS TeleTracking, varies greatly by cities. In Seattle, there was a total of 535 cases whereas in Los Angeles there was a total of 139,042 cases reported over a 10 month period.
- The average total number of COVID hospitalizations between Atlanta, Chicago, Los Angeles, New York, and Seattle over 10 months from 2021 to 2022 is 85,990 cases.

04 Healthcare Provider Vaccination Status

Ratio and difference of healthcare providers who are and aren’t vaccinated mapped in the United States

What trends do you notice in the vaccination status of healthcare personnel by counties and states? How does the ratio and difference in vaccinated and unvaccinated change regionally?

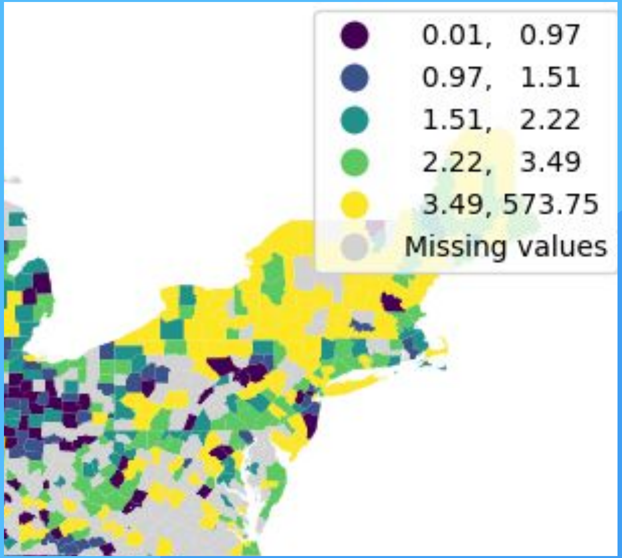
Ratio of Vaccinated Healthcare Providers



In this graph we have all the counties within the mainland United States. US Territories, Alaska, and Hawaii were excluded to create an easily interpretable graph.

In this analysis, higher ratios (in yellow) mean there is larger proportion of vaccinated healthcare providers to unvaccinated. The dark purple is the opposite, where there is a larger proportion of unvaccinated providers.

Ratio of Vaccinated Healthcare Providers



Zoom in on New York counties for up close analysis

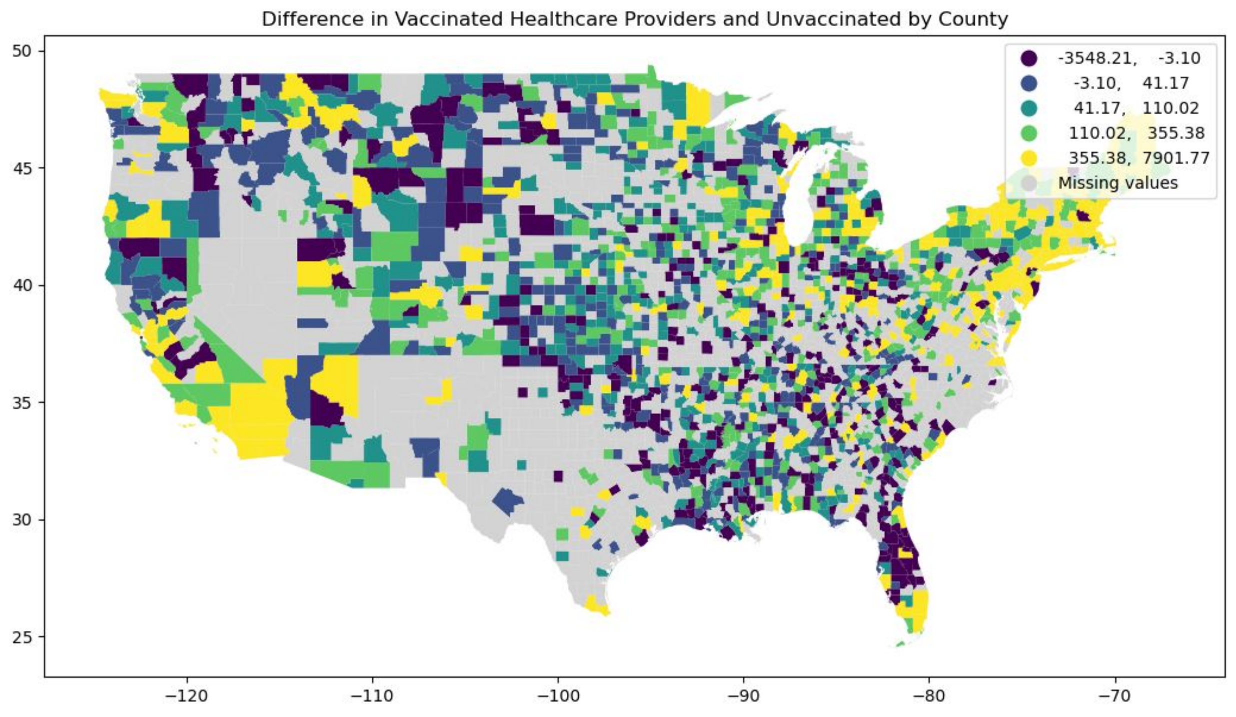
Within these values, we can see counties in California, New York, and other north & northeastern states having larger proportions of vaccinated healthcare providers.

This does not surprise anyone, as these are more liberal states which historically during the pandemic have been more likely to receive the vaccine.

Additionally, many states, including the ones mentioned, have vaccine mandates in place. For example, New York has a policy requiring vaccination in all healthcare settings and if employees are not vaccinated they will be fired [3].

This analysis also displays lots of missing values, which does not create a comprehensive analysis, but it is more accurate. Missing values are displayed in 'grey' on the graph.

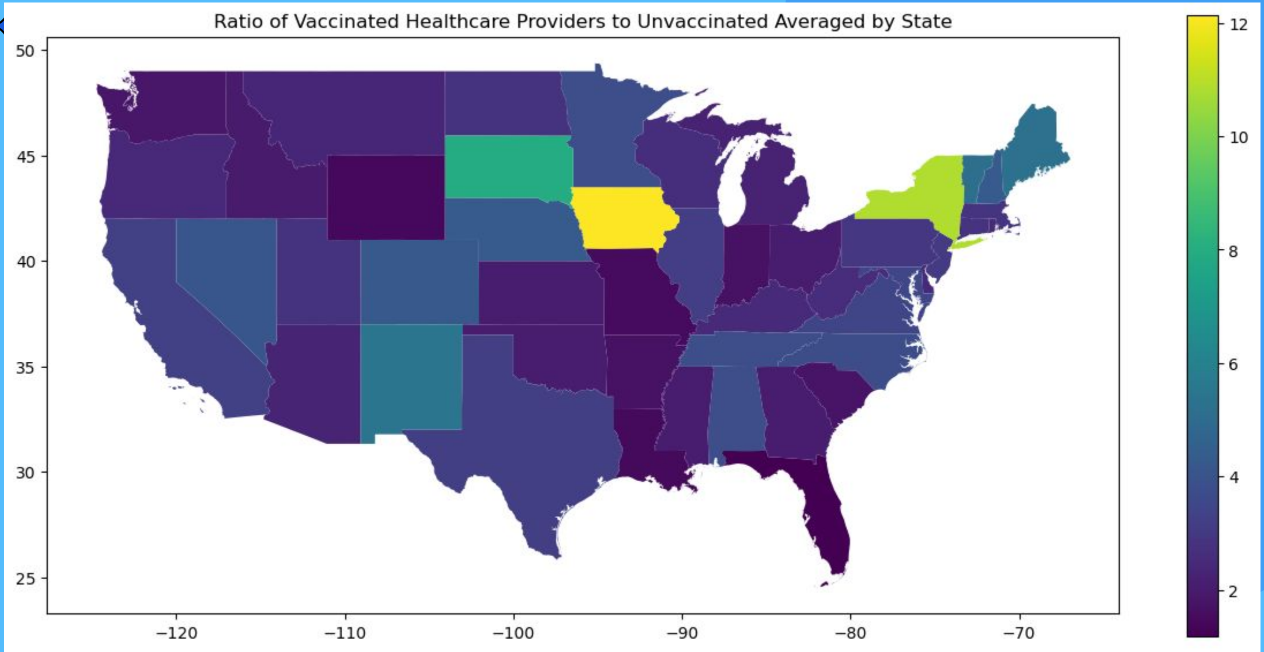
Difference of Vaccinated Healthcare Providers



In this graph, we are again examining counties within the US, but we are looking in the difference of vaccinated healthcare providers and unvaccinated.

In this case it is vaccinated - unvaccinated. Like the ratio, those with a higher number of vaccinated healthcare providers are in bright yellow, and areas is more unvaccinated are in dark purple. This graph is very similar to the last.

Average Ratio of Vaccinated Healthcare Providers



In this graph, we are again examining the ratio of vaccinated to unvaccinated healthcare providers, but we have now averaged this ratio by state.

As you can see New York and the northeast are still pretty light, but Iowa and South Dakota now stand out which was harder to see in the earlier graphs. In fact, with all the county lines drawn it is hard to discern Iowa and South Dakota at all, so this creates a more readable graph.

Ratio of Vaccinated Healthcare Providers

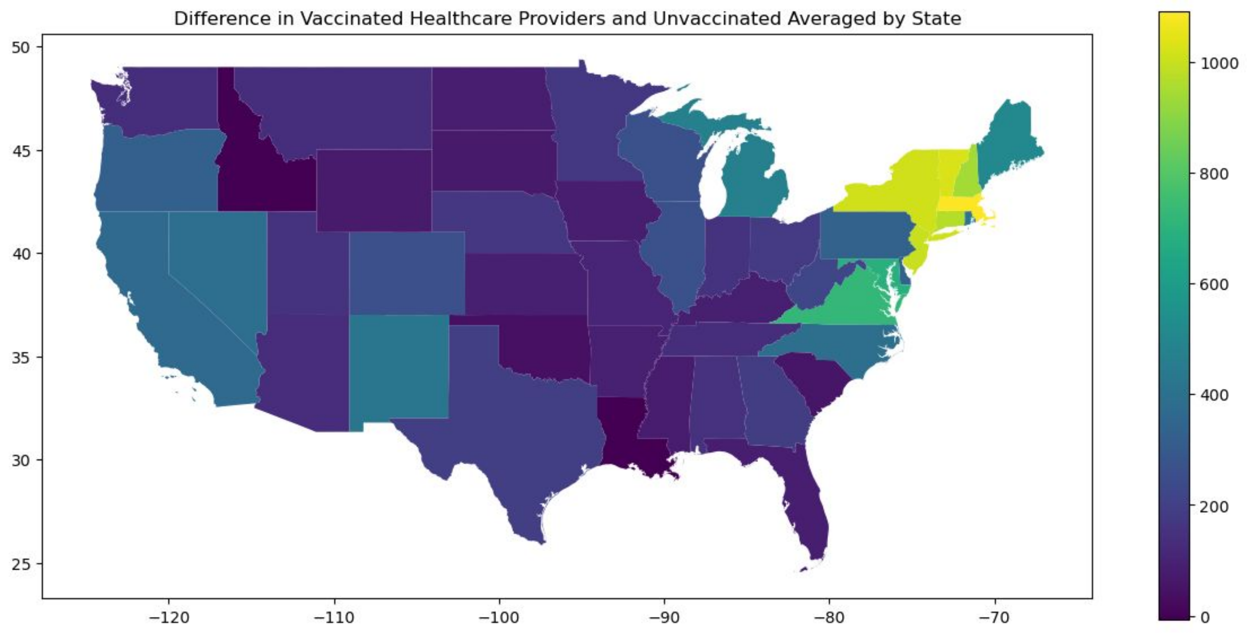
Zoom in on Iowa and South Dakota within the last graph

This analysis has caveats for how the numbers are created. For each county we are getting a ratio of vaccinated healthcare providers to unvaccinated, but dividing vaccinated/unvaccinated. For each county we get a proportion.

Within this graph we are taking every county for that state and we are averaging these proportions to get a more representative value for the whole state. Anytime you take an average, these numbers are easily influenced by outliers, and even more so when there is minimal data.

We cannot be sure how representative these values are without the missing data. However, this graph is much easier to read for those evaluating the analysis. When creating graphs there is always a balance of readability and accuracy.

Average Difference of Vaccinated Healthcare Providers



In this graph, we are again examining the difference of vaccinated to unvaccinated healthcare providers, but we have now averaged this difference by state.

This graph already seems more representative of our county evaluation by the northeast lighting up in bright yellow. Additionally California looks slightly lighter in color in comparison to the previous graph. Same caveats apply from the last graph in consideration to averages.

https://github.com/efgronski/hospital_covid_data

GitHub Repository of Code

References

- [1] Dataset: https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/anag-cw7u
- [2] https://www.ipsos.com/en-us/news-polls/axios-ipsos-coronavirus-index
- [3] https://leadingage.org/workforce-vaccine-mandates-state-who-who-isnt-and-how/

Thanks!

Team Polaris: Caroline Wills, Militha Madur, Beth Gronski, & Kalkidan Tamirat