Vaccination Rates in NJ

Meredith Lou

3/27/2022

How do vaccination rates compare in three different states?

I was interested in this data because it is very relevant to the current pandemic we are living through. Furthermore, I were curious to compare and contrast vaccination rates in three states/districts and within the counties of those states (for those applicable). I chose New Jersey (where I'm from), DC (where I go to school), and Texas (politically different)

Description of Data

I got my data from the CDC website, titled "Covid 19 Vaccinations in the United States, County". There are 66 columns and 1.57M rows. The column variables include data on the proportion of people who were administered their first dose, those who had their series complete (fully vaccinated, or two doses), and those who were boosted. It also had the census data on the total population of each county in 2019. The variables for the rows are each county of the 50 states, plus DC.

Plan

I filtered the data to include only my 3 chosen states: New Jersey, Texas, and DC. I had to compare rates, as the three states had vastly different populations. This means that I had to calculate vaccination rates ourselves, even though the data provided rates, as I could not just aggregate vaccination rates across all the counties for each state. In terms of visualizations, I wanted to construct choropleths to visualize the differences within the counties, time series to compare the three states, and a bar graph to summarize my findings and compare the three vaccination options (one dose, fully vaccinated, and boosted).

Design

For the choropleths, I chose to use blue to represent the scale of NJ and red for TX to underline the political connotations underneath my findings. I did not know what to do with DC as it does not have any counties but I ultimately had no choice but to leave it out of the choropleths. Not long after examining the dataset, I realized that CDC only provides data for October 21st, 2021 and on for Texas. This meant that I had to adjust the design of our visualizations. I decided to keep it consistent across all three states and use October as the start for the time series. I contemplated between doing a paneled bar plot or a stacked bar plot to compare first dose, full series, and booster rates across all three states. Ultimately, I decided a stacked bar plot would be the better option as it would clutter the graph less.

Preprocessing

Choropleth

In order to make the CDC data compatible with map data from the maps library, I had to use the strings library to take "county" out of each county name and make all county names lowercase. Only then was I able to join the two datasets for the choropleth. I soon found that in order to make the scales consistent across the two before and afters, I would have to do a log transformation on the percentage vaccinated.

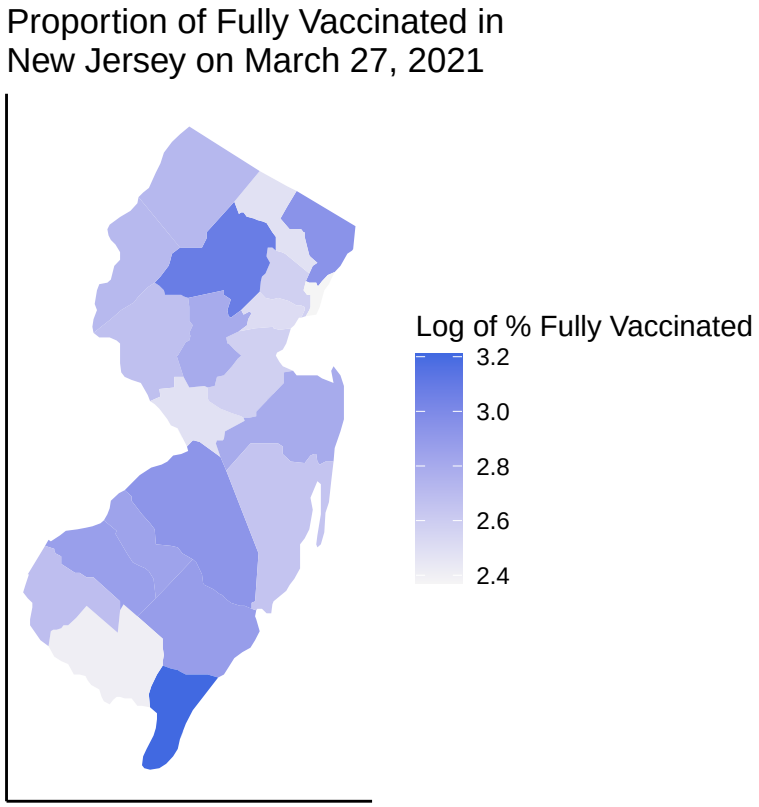
```
vacc_county <- mutate(vacc_county,
                      subregion = tolower(substr(Recip_County,1,str_length(Recip_County)-7)))
nj_county <- map_data("county", "new jersey")
nj_map <- left_join(nj_county,
                    filter(vacc_county, Recip_State == "NJ"), by="subregion")
```

```
## Warning in left_join(nj_county, filter(vacc_county, Recip_State == "NJ"), : Detected an unexpected ma
## i Row 1 of `x` matches multiple rows in `y`.
## i Row 6 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

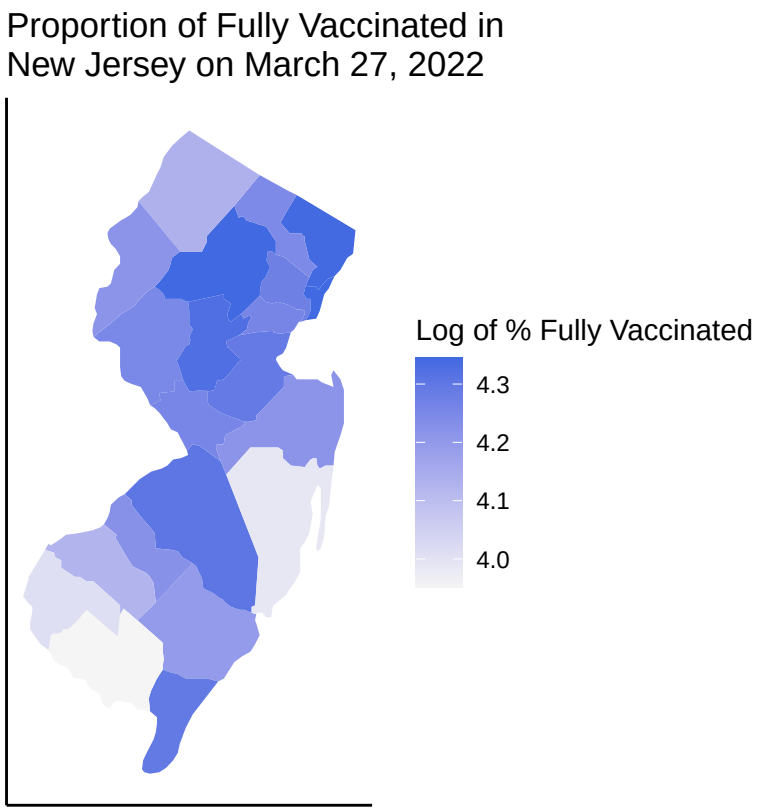
```
tx_county <- map_data("county", "texas")
tx_map <- left_join(tx_county,
                    filter(vacc_county, Recip_State == "TX"), by="subregion")
```

```
## Warning in left_join(tx_county, filter(vacc_county, Recip_State == "TX"), : Detected an unexpected ma
## i Row 1 of `x` matches multiple rows in `y`.
## i Row 166 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

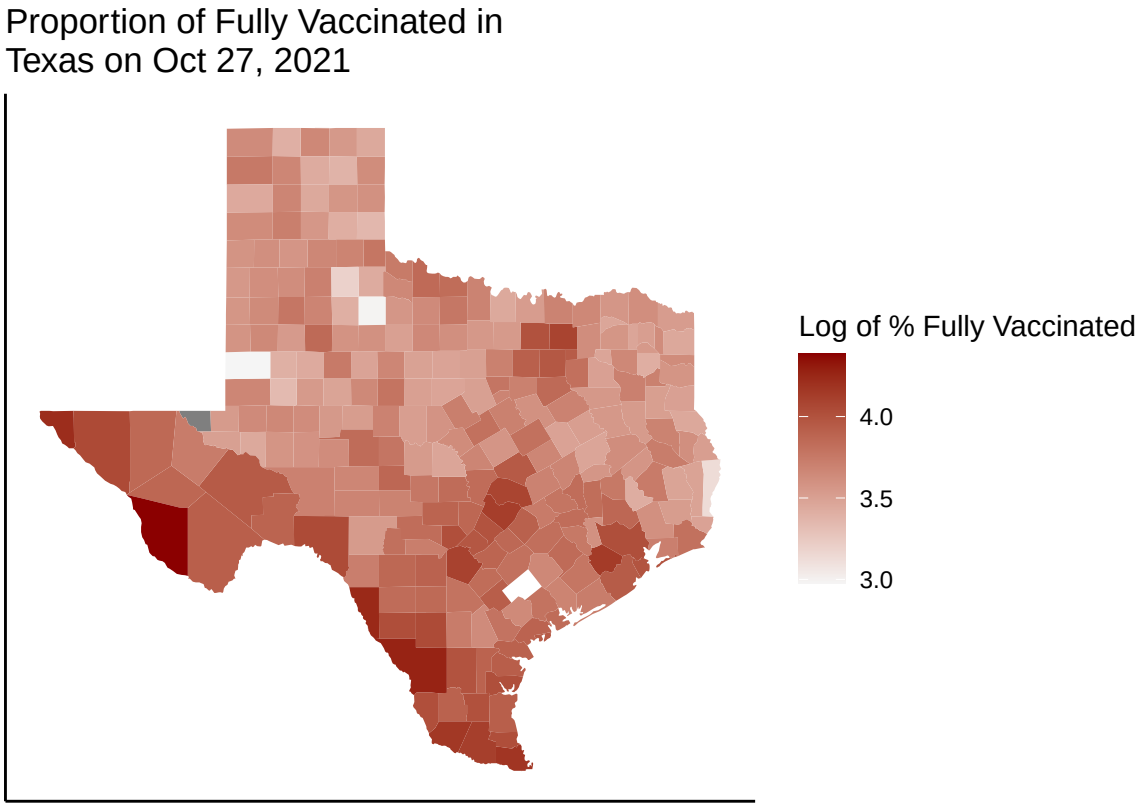
```
#NJ
ggplot() +
  geom_polygon(data=filter(nj_map, Date == '03/27/2021'),
    aes(x = long, y = lat, group = group, fill = log(Series_Complete_Pop_Pct))) +
    coord_map() +
    labs(x= "", y = "",
         title = "Proportion of Fully Vaccinated in \nNew Jersey on March 27, 2021") +
    theme_classic() + theme(axis.ticks.y = element_blank(),
                            axis.text.y = element_blank(),
                            axis.ticks.x = element_blank(),
                            axis.text.x = element_blank()) +
  scale_fill_gradient(name = "Log of % Fully Vaccinated",
                      low = "whitesmoke", high = "royalblue")
```



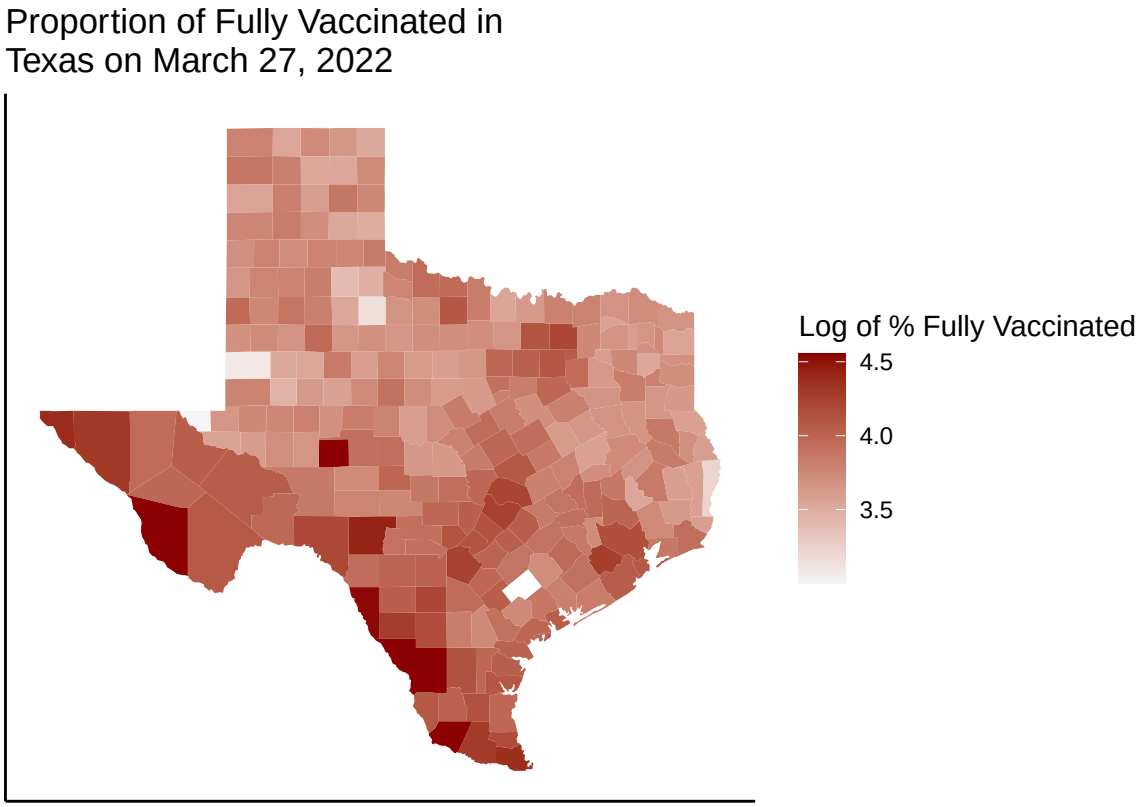

```
ggplot() +
  geom_polygon(data=filter(nj_map, Date == '03/27/2022'),
    aes(x = long, y = lat, group = group, fill = log(Series_Complete_Pop_Pct))) +
    coord_map() +
  labs(x = "", y = "",
       title = "Proportion of Fully Vaccinated in \nNew Jersey on March 27, 2022") +
  theme_classic() + theme(axis.ticks.y = element_blank(),
                          axis.text.y = element_blank(),
                          axis.ticks.x = element_blank(),
                          axis.text.x = element_blank()) +
  scale_fill_gradient(name = "Log of % Fully Vaccinated",
                      low = "whitesmoke", high = "royalblue")
```




```
#TX
ggplot() +
  geom_polygon(data=filter(tx_map, Date == '10/27/2021'),
    aes(x = long, y = lat, group = group, fill = log(Series_Complete_Pop_Pct))) +
    coord_map() +
  labs(x = "", y = "",
       title = "Proportion of Fully Vaccinated in \nTexas on Oct 27, 2021") +
  theme_classic() + theme(axis.ticks.y = element_blank(),
                          axis.text.y = element_blank(),
                          axis.ticks.x = element_blank(),
                          axis.text.x = element_blank()) +
  scale_fill_gradient(name = "Log of % Fully Vaccinated",
                      low = "whitesmoke", high = "darkred")
```




```
ggplot() +
  geom_polygon(data=filter(tx_map, Date == '03/27/2022'),
    aes(x = long, y = lat, group = group, fill = log(Series_Complete_Pop_Pct))) +
    coord_map() +
  labs(x = "", y = "",
       title = "Proportion of Fully Vaccinated in \nTexas on March 27, 2022") +
  theme_classic() + theme(axis.ticks.y = element_blank(),
                          axis.text.y = element_blank(),
                          axis.ticks.x = element_blank(),
                          axis.text.x = element_blank()) +
  scale_fill_gradient(name = "Log of % Fully Vaccinated",
                      low = "whitesmoke", high = "darkred")
```



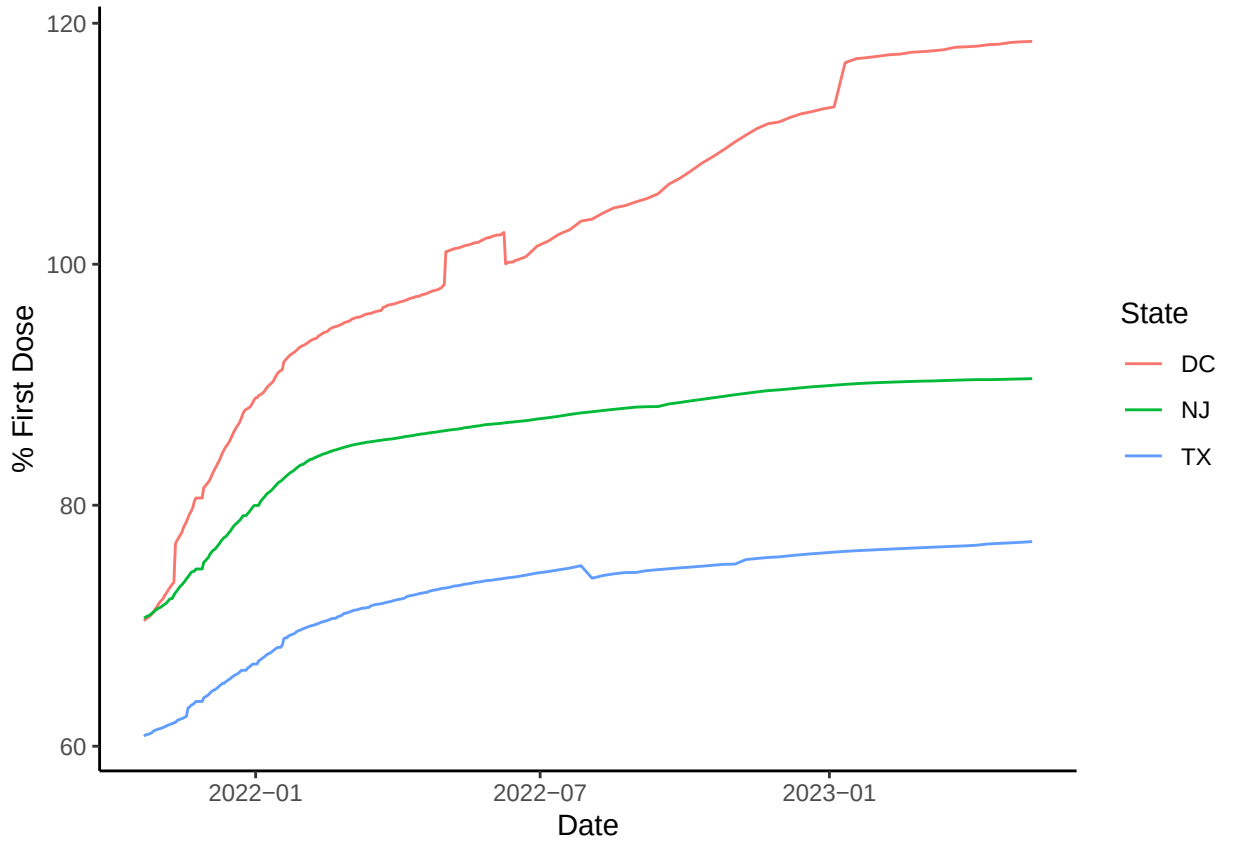
Time Series

As stated before, I had to create my own vaccination rates for each state as the CDC data only provided rates for each county. Since boosters weren't widely distributed until December, I filtered the data for that graph for December and on.

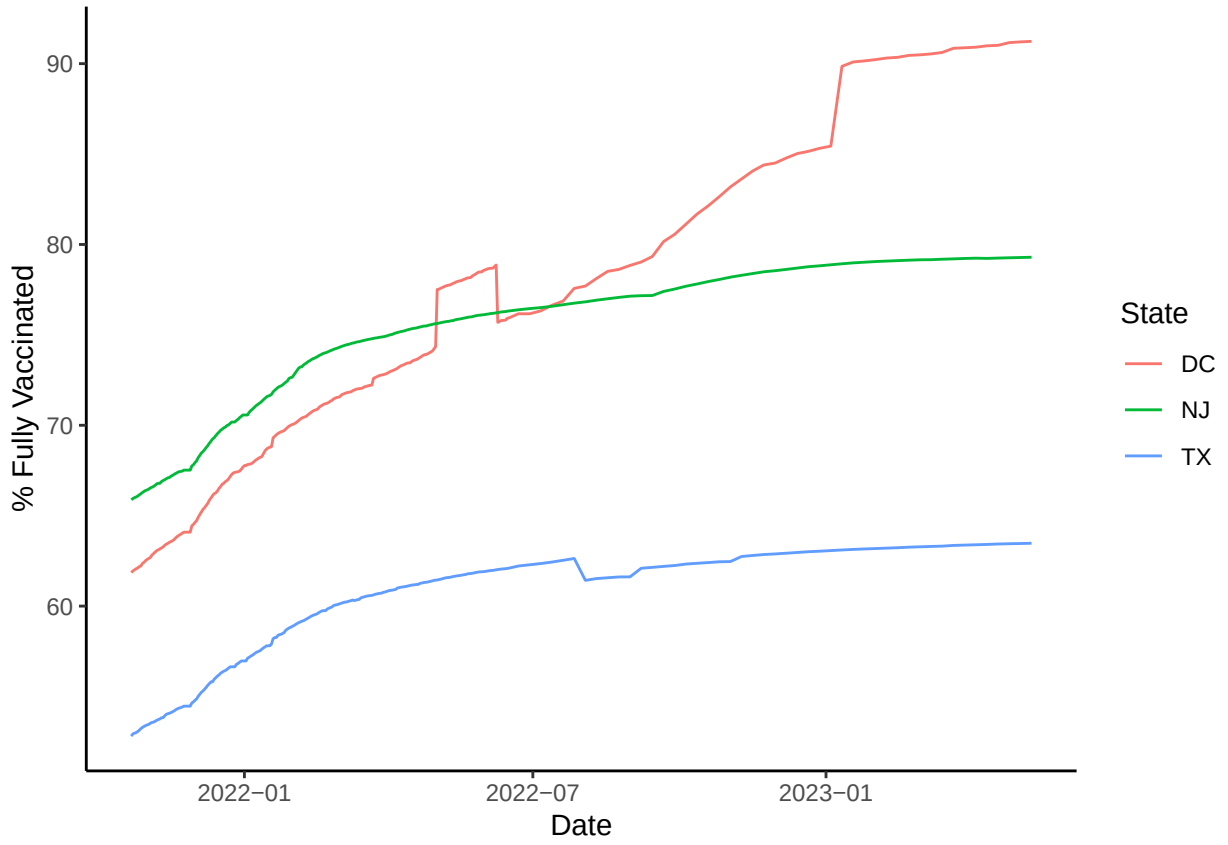
```
vacc_state <- summarize(group_by(vacc_county, Date, Recip_State),
                                One_Dose = sum(Administered_Dose1_Recip,na.rm=T)/sum(Census2019,na.rm=T
                                Fully_Vaccinated = sum(Series_Complete_Yes,na.rm=T)/sum(Census2019,na.r
                                Boosted = sum(Booster_Doses,na.rm=T)/sum(Census2019,na.rm=T)*100)
```

```
## `summarise()` has grouped output by 'Date'. You can override using the
## `.groups` argument.
```

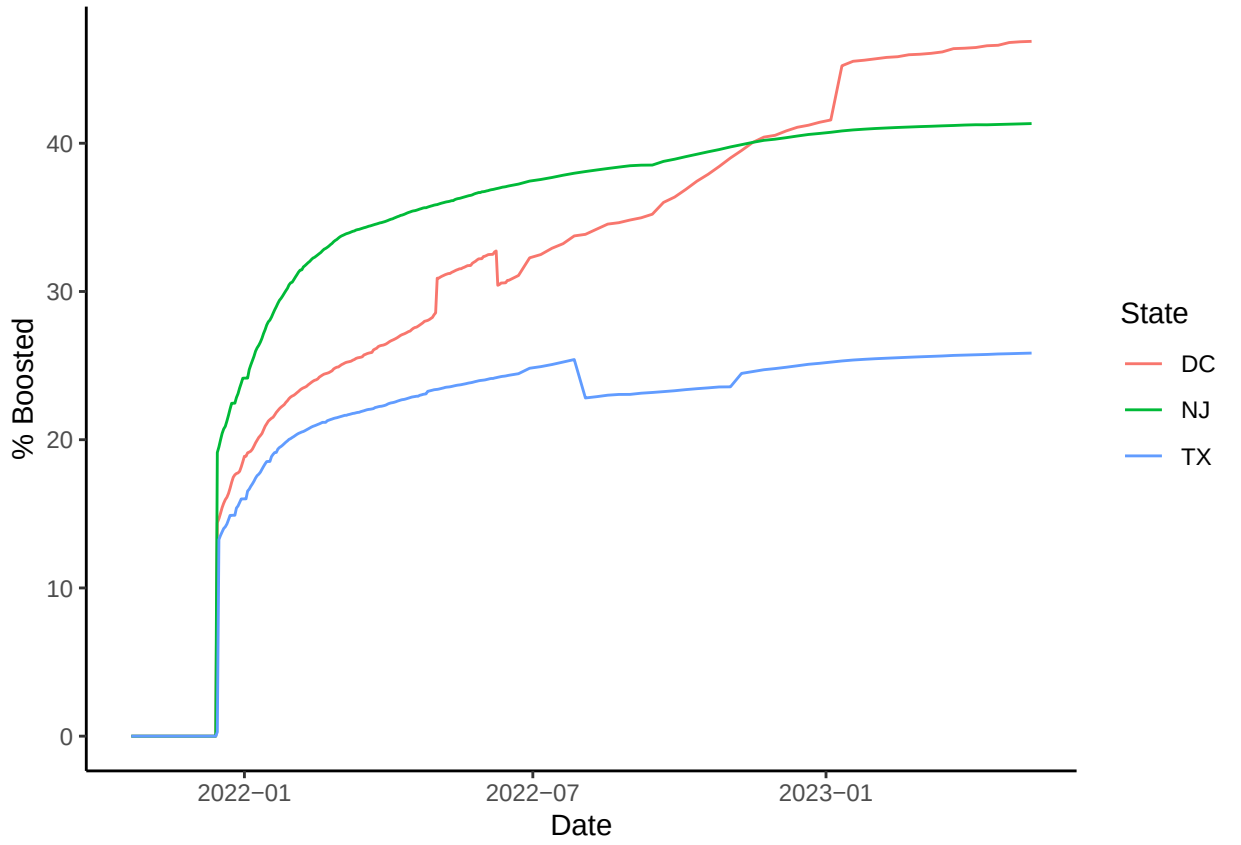
```
vacc_state$Date <- as.character(vacc_state$Date)
vacc_state <- mutate(vacc_state, Date = mdy(Date, truncated=0))
vacc_state <- filter(vacc_state, Date > "2021-10-21")
ggplot(data = vacc_state) + geom_line(aes(x=Date, y=One_Dose, color = Recip_State)) +
  labs(y = "% First Dose", color = "State") + theme_classic()
```




```
ggplot(data = vacc_state) + geom_line(aes(x=Date, y=Fully_Vaccinated, color = Recip_State)) +
  labs(y = "% Fully Vaccinated", color = "State") + theme_classic()
```




```
ggplot(data = vacc_state) + geom_line(aes(x=Date, y=Boosted, color = Recip_State)) +
  labs(y = "% Boosted", color = "State") + theme_classic()
```



Bar Graph

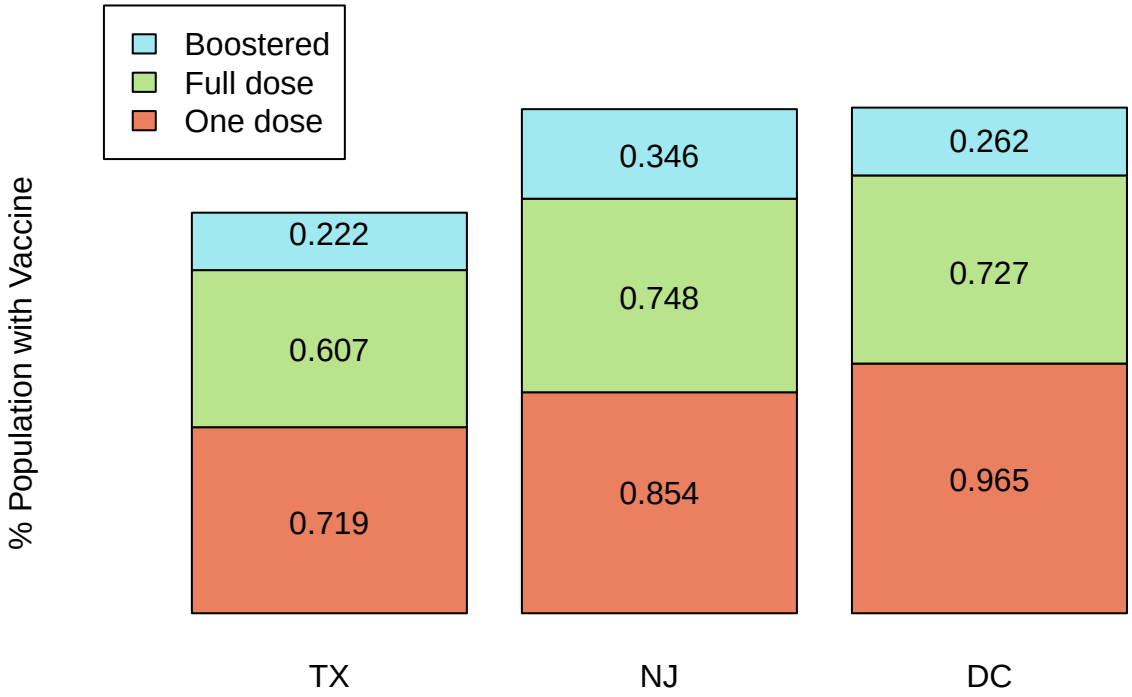
In order to present the data in a bar graph, I had to create a new matrix.

```
vacc_table <- matrix(ncol = 3, nrow=3)
colnames(vacc_table) <- c("TX", "NJ", "DC")
rownames(vacc_table) <- c("One dose","Full dose","Boostered")
vacc_table[1,1]=0.719
vacc_table[2,1]=0.607
vacc_table[3,1]=0.222
vacc_table[1,2]=0.854
vacc_table[2,2]=0.748
vacc_table[3,2]=0.346
vacc_table[1,3]=0.965
vacc_table[2,3]=0.727
vacc_table[3,3]=0.262

barplot(vacc_table, ylab = "% Population with Vaccine",
        yaxt = 'n', legend.text = TRUE,
        args.legend = list(x = "topleft",
                           inset = c(-.05, -.2)), col = c("#eb8060", "#b9e38d", "#a1e9f0", "#d9b1f0"))

text(vacc_table[1,1], 0.719, x=0.7, y=0.2, cex=1, pos=3, adj = c(0.5,0.5))
text(vacc_table[2,1], 0.607, x=0.7, y=0.9, cex=1,pos=3)
text(vacc_table[3,1], 0.222, x=0.7, y=1.35, cex=1,pos=3)
```

```
text(vacc_table[1,2], 0.854, x=1.9, y=0.35, cex=1,pos=3)
text(vacc_table[2,2], 0.748, x=1.9, y=1.1, cex=1,pos=3)
text(vacc_table[3,2], 0.346, x=1.9, y=1.65, cex=1,pos=3)
text(vacc_table[1,3], 0.965, x=3.1, y=0.4, cex=1,pos=3)
text(vacc_table[2,3], 0.727, x=3.1, y=1.2, cex=1,pos=3)
text(vacc_table[3,3], 0.262, x=3.1, y=1.71, cex=1,pos=3)
```



Conclusions

The bar graph summarizes my findings in one visualization and answering my research question. It shows the proportion of vaccination status in Texas, New Jersey, and DC side by side, and the viewer can easily compare and contrast the three. It is also easy to compare and contrast the data within the state, and compare the proportion of people who have one dose, are fully vaccinated, and those who are boosted. The time series with the proportion of first dose over time demonstrates that while DC and New Jersey started off with the same proportion of those who had their first dose in early November 2021, by mid November, DC jumps ahead of New Jersey. This jump could be attributed to a loss of follow up in DC, where DC failed to bring in the same number of people for their second dose. This could be due to the fact that New Jersey has a smaller low income population than DC, and thus more people in DC would not be able to follow up on their second dose due to inability to take off of work, find childcare, etc. The time series also shows that Texas has a much lower overall proportion of those who have their first dose compared to both DC and New Jersey - this is consistent across all three time series graphs.