R Notebook

Code

Hide

```
library(dplyr)
library(tidyverse)
library(ggplot2)
library(RColorBrewer)
library(Stat2Data)
library(maps)
library(USAboundaries)
library(sf)
library(vcd)
library(lubridate)
```

Hide

```
require(devtools)
devtools::install_github("ropensci/USAboundariesData")
```

```
Skipping install of 'USAboundariesData' from a github remote, the SHA1 (a3db4fb6) has no
t changed since last install.
  Use `force = TRUE` to force installation
```

Load COVID Data

Hide

```
#us_state_vaccinations <- read.csv(file.choose())
us_state_vaccinations
```

date <chr>	location <chr>	total_vaccinations <dbl>	total_distributed <dbl>	people_vaccinated <dbl>
2021-01-12	Alabama	78134	377025	70861
2021-01-13	Alabama	84040	378975	74792
2021-01-14	Alabama	92300	435350	80480
2021-01-15	Alabama	100567	444650	86956
2021-01-16	Alabama	*NA*	*NA*	*NA*
2021-01-17	Alabama	*NA*	*NA*	*NA*
2021-01-18	Alabama	*NA*	*NA*	*NA*
2021-01-19	Alabama	130795	444650	114319
2021-01-20	Alabama	139200	483275	121113
2021-01-21	Alabama	165919	493125	144429

1-10 of 8,148 rows | 1-5 of 14 columns Previous **1** 2 3 4 5 6 … 100 Next

Hide

```
#us_states <- read.csv(file.choose())
us_states
```

date <chr>	**state** <chr>	**fips** <int>	**cases** <int>	**deaths** <int>
2020-01-21	Washington	53	1	0
2020-01-22	Washington	53	1	0
2020-01-23	Washington	53	1	0
2020-01-24	Illinois	17	1	0
2020-01-24	Washington	53	1	0
2020-01-25	California	6	1	0
2020-01-25	Illinois	17	1	0
2020-01-25	Washington	53	1	0
2020-01-26	Arizona	4	1	0
2020-01-26	California	6	2	0

1-10 of 24,269 rows Previous **1** 2 3 4 5 6 … 100 Next

Hide

```
#us_counties <- read.csv(file.choose())
us_counties
```

date <chr>	**county** <chr>	**state** <chr>	**fips** <int>	**cases** <int>	**deaths** <int>
2020-01-21	Snohomish	Washington	53061	1	0
2020-01-22	Snohomish	Washington	53061	1	0
2020-01-23	Snohomish	Washington	53061	1	0
2020-01-24	Cook	Illinois	17031	1	0
2020-01-24	Snohomish	Washington	53061	1	0
2020-01-25	Orange	California	6059	1	0
2020-01-25	Cook	Illinois	17031	1	0
2020-01-25	Snohomish	Washington	53061	1	0
2020-01-26	Maricopa	Arizona	4013	1	0

date <chr>	county <chr>	state <chr>	fips <int>	cases <int>	deaths <int>
2020-01-26	Los Angeles	California	6037	1	0

1-10 of 1,329,487 rows Previous **1** 2 3 4 5 6 … 100 Next

Hide

```
#country_vaccinations_by_manufacturer <- read.csv(file.choose())
country_vaccinations_by_manufacturer
```

location <chr>	date <chr>	vaccine <chr>	total_vaccinations <int>
Chile	2020-12-24	Pfizer/BioNTech	420
Chile	2020-12-25	Pfizer/BioNTech	5198
Chile	2020-12-26	Pfizer/BioNTech	8338
Chile	2020-12-27	Pfizer/BioNTech	8649
Chile	2020-12-28	Pfizer/BioNTech	8649
Chile	2020-12-29	Pfizer/BioNTech	8649
Chile	2020-12-30	Pfizer/BioNTech	8649
Chile	2020-12-31	Pfizer/BioNTech	8649
Chile	2021-01-01	Pfizer/BioNTech	8649
Chile	2021-01-02	Pfizer/BioNTech	8649

1-10 of 3,808 rows Previous **1** 2 3 4 5 6 … 100 Next

Hide

```
#country_vaccinations <- read.csv(file.choose())
country_vaccinations
```

country <chr>	iso_code <chr>	date <chr>	total_vaccinations <dbl>	people_vaccinated <dbl>	people_fully_
Afghanistan	AFG	2021-02-22	0	0	
Afghanistan	AFG	2021-02-23	*NA*	*NA*	
Afghanistan	AFG	2021-02-24	*NA*	*NA*	
Afghanistan	AFG	2021-02-25	*NA*	*NA*	
Afghanistan	AFG	2021-02-26	*NA*	*NA*	
Afghanistan	AFG	2021-02-27	*NA*	*NA*	

country <chr>	iso_code <chr>	date <chr>	total_vaccinations <dbl>	people_vaccinated <dbl>	people_fully_
Afghanistan	AFG	2021-02-28	8200	8200	
Afghanistan	AFG	2021-03-01	*NA*	*NA*	
Afghanistan	AFG	2021-03-02	*NA*	*NA*	
Afghanistan	AFG	2021-03-03	*NA*	*NA*	

1-10 of 17,607 rows | 1-6 of 15 columns Previous **1** 2 3 4 5 6 … 100 Next

Hide

```
#owid_covid_data <- read.csv(file.choose())
owid_covid_data
```

iso_code <chr>	continent <chr>	location <chr>	date <chr>	total_cases <dbl>	new_ca… <dbl>	new_cases_smoothed <dbl>
AFG	Asia	Afghanistan	2020-02-24	1	1	*NA*
AFG	Asia	Afghanistan	2020-02-25	1	0	*NA*
AFG	Asia	Afghanistan	2020-02-26	1	0	*NA*
AFG	Asia	Afghanistan	2020-02-27	1	0	*NA*
AFG	Asia	Afghanistan	2020-02-28	1	0	*NA*
AFG	Asia	Afghanistan	2020-02-29	1	0	0.143
AFG	Asia	Afghanistan	2020-03-01	1	0	0.143
AFG	Asia	Afghanistan	2020-03-02	1	0	0.000
AFG	Asia	Afghanistan	2020-03-03	2	1	0.143
AFG	Asia	Afghanistan	2020-03-04	4	2	0.429

1-10 of 89,357 rows | 1-8 of 59 columns Previous **1** 2 3 4 5 6 … 100 Next

The overall goal of my project is to explore how COVID deaths and cases effected vaccination rates in each state.

Plot #1

I am looking to explore this data by first looking at vaccination rates in each state. I wanted to begin by creating a chloropleth. I thought that this would be an interesting and unique way to visualize data from all of the different states. I will be creating a chloropleth which shows how many people per 100 people are vaccinated in each state.

Hide

```
us_state_vaccinations1 <- us_state_vaccinations %>%
  filter(date == "2021-05-16")
```

Hide

```
states <- us_states("2000-01-01")
```

Hide

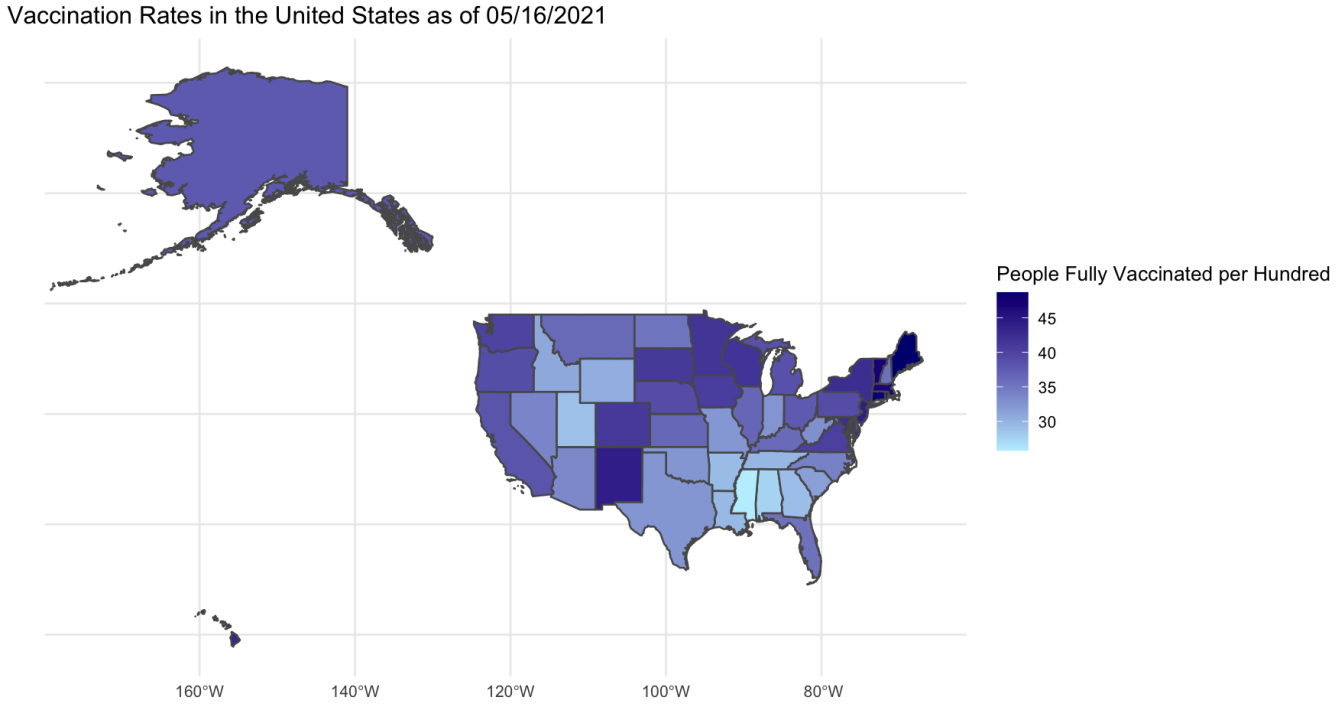
```
us_state_vaccinations2 <- us_state_vaccinations1 %>%
  mutate(location = ifelse(location == "New York State", "New York", location))
```

Hide

```
covidmap1 <- merge(states, us_state_vaccinations2, by.x = "state_name", by.y = "locatio
n")
```

Hide

```
ggplot(covidmap1) + geom_sf(mapping = aes(fill = people_fully_vaccinated_per_hundred)) +
coord_sf(crs = st_crs(4269)) + ggtitle("Vaccination Rates in the United States as of 05/
16/2021") + scale_fill_gradient(low = "lightblue1", high = "navy") + labs(fill = "People
Fully Vaccinated per Hundred") + theme_minimal()
```



A principal from FDV which I included in this figure was from Chapter 4, color scales. I used a sequential color scale. I chose for my chloropleth to color the states with a low vaccination rate light blue, and the states with a high vaccination rate dark blue. I created my chloropleth so that all the values inbetween the highest and lowest vaccination rates were a shade of blue which fell between light and dark blue.

I also used concepts from FDV chapter 15 which discusses chloropleths in particular. I used dark colors to represent the higher vaccination rates, and lighter colors to represent the lower vaccination rates. I used a continuous color scale for my chloropleth.

I also used labels in my chloropleth. This realted to FDV chapter 22. It is necessary to include a title in almost all plots, so I included one in this plot. I also included a legend title because it is not evident from just the plot title what the data in the chloropleth is representing.

Plot #2

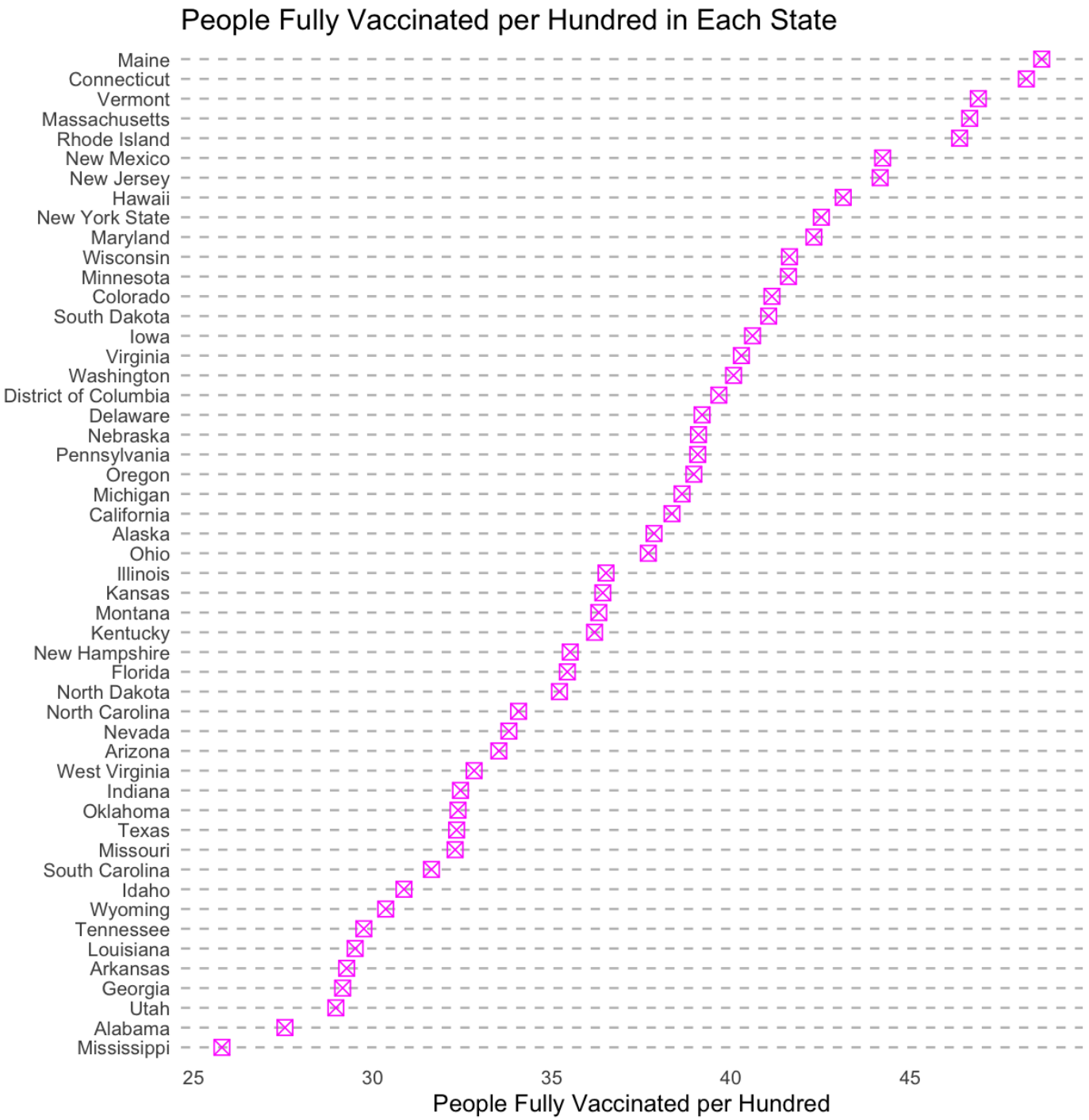
Although the chloropleth is visually appealing and an interesting way to look at the data, I am hoping to obtain a less vague idea of which states had the highest and lowest vaccination rate. To do this, I will create a cleveland dot plot.

Hide

```
#Removing locations which do not store people_fully_vaccinated_per_hundred data or are n
ot part of the US
us_state_vaccinations3 <- us_state_vaccinations1 %>%
  filter(location != "Veterans Health") %>%
  filter(location != "Long Term Care") %>%
  filter(location != "Dept of Defense") %>%
  filter(location != "Bureau of Prisons") %>%
  filter(location != "Virgin Islands") %>%
  filter(location != "Republic of Palau") %>%
  filter(location != "Puerto Rico") %>%
  filter(location != "Northern Mariana Islands") %>%
  filter(location != "Marshall Islands") %>%
  filter(location != "Indian Health Svc") %>%
  filter(location != "Guam") %>%
  filter(location != "Federated States of Micronesia") %>%
  filter(location != "American Samoa") %>%
  filter(location != "United States")
```

Hide

```
ggplot(us_state_vaccinations3, aes(people_fully_vaccinated_per_hundred, reorder(locatio
n, people_fully_vaccinated_per_hundred))) + geom_point(col = "magenta", size = 3, pch =
7) + theme_minimal() + theme(panel.grid.major.x = element_blank(), panel.grid.minor.x =
 element_blank(), panel.grid.major.y = element_line(colour = "grey", linetype = "dashed"
)) + ylab(" ") + xlab("People Fully Vaccinated per Hundred") + ggtitle("People Fully Vac
cinated per Hundred in Each State")
```



This plot provides similar information as the chloropleth, however, it is much easier to visualize the exact data value for each state. From this plot I was also able to determine the descending order of vaccination rate in each state.

I used several principles from FDV such as in my first plot. I thought that I effectively used color because I chose colors like black and grey for the labels and dashed line, and then I chose a bright pink for the data points so that they stuck out from the rest of plot. I also only added a x axis label because the y axis was self explanatory, as each state is labeled next to the appropriate row. I also included a appropriatly capitalized title. I also incorporated the minimal theme because I thought it made my plot look neat and allow the pink data points to really pop out.

Plot #3:

After obsrving that Maine has the highest vaccination rate, I wanted to look into what may have affected this. I was curious if the number of deaths in Maine was higher than in other states. I chose to compare Maine with Mississippi which had the lowest vaccination rate. I chose to convey this information in a color coded bar plot.

Hide

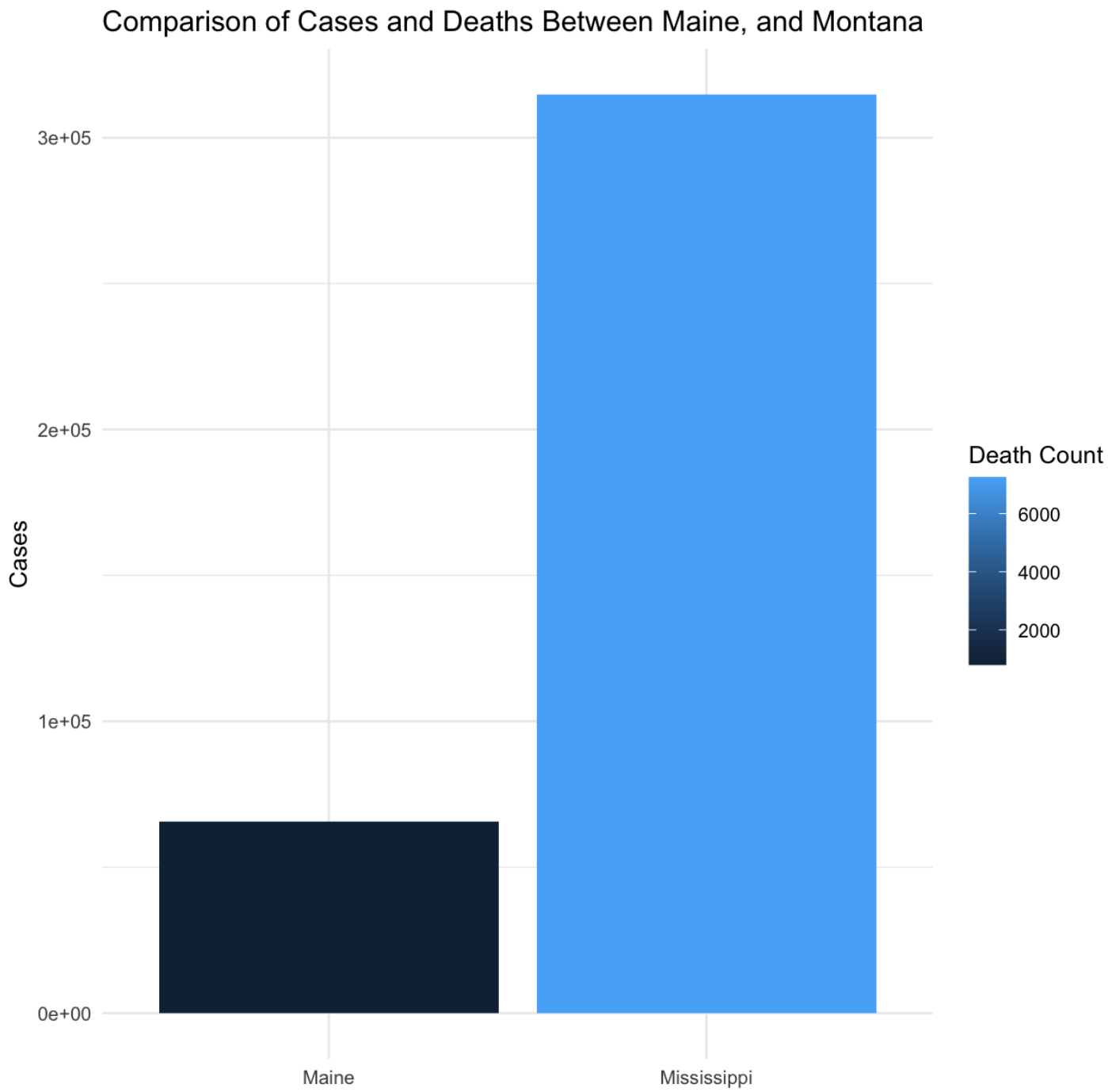
```
us_states1 <- us_states %>%
  filter(date == "2021-05-16") %>%
  filter(state %in% c("Mississippi", "Maine"))
us_states1
```

date <chr>	state <chr>	fips <int>	cases <int>	deaths <int>
2021-05-16	Maine	23	65715	802
2021-05-16	Mississippi	28	314710	7254

2 rows

Hide

```
ggplot(us_states1, aes(x = state, y = cases, fill = deaths)) + geom_col() + theme_minima
l() + ggtitle("Comparison of Cases and Deaths Between Maine, and Montana") + ylab("Case
s") + xlab(" ") + labs(fill = "Death Count")
```



It is interesting to visualize that although Maine had signifgantly lower cases and deaths than Mississippi, they still had a much higher vaccination rate. Going forward, I hope to look into whether region effects whether or not a person got vaccinated, since Maine and Mississippi are in different regions.

I used several concepts from FDV in this plot. I effectively used color because the there is a clear distinction between death count in Maine and Mississippi. I also properly capitalized all legends, titles, and axes. I also incorporated the minimal theme as I did in my previous plots, as I like how it makes the plot look neat and organized.

Plot #4

I wanted to explore whether which region a state was in may have effected whether or not an individual got vaccinated. I created a function which input each state into one of the 5 regions in the US. I then decided to convey this information using a time series.

Hide

```
regions <- list(
  West = c("Hawaii", "California", "Nevada", "Utah", "Colorado", "Wyoming", "Idaho", "Or
egon", "Washington", "Montana", "Alaska"),
  Southwest = c("Arizona", "Texas", "New Mexico", "Oklahoma"),
  Southeast = c("Louisiana", "Arkansas", "Mississippi", "Tennessee", "Alabama", "Georgi
a", "South Carolina", "North Carolina", "Florida", "Kentucky", "West Virginia", "Virgini
a"),
  Midwest = c("North Dakota", "South Dakota", "Nebraska", "Kansas", "Minnesota", "Iowa",
"Missouri", "Illinois", "Wisconsin", "Michigan", "Ohio", "Indiana"),
  Northeast = c("Pennsylvania", "Maryland", "Delaware", "New Jersey", "Connecticut", "Rh
ode Island", "Massachusetts", "New York State", "Vermont", "New Hampshire", "Maine", "Di
strict of Columbia")
)

convert <- function(x, data) {
  out <- x[NA]
  for(nm in names (data)) {
    ind <- x %in% data[[nm]]
    out[ind] <- nm
  }
  return(out)
}
```

Hide

```
us_state_vaccinations4 <- us_state_vaccinations %>%
  mutate(date = ymd(date)) %>%
  mutate(Region = convert(location, regions)) %>%
  group_by(date, Region) %>%
  summarise(people_fully_vaccinated_per_hundred = mean(people_fully_vaccinated_per_hundr
ed, na.rm = TRUE))
```

```
`summarise()` regrouping output by 'date' (override with `.groups` argument)
```

Hide

```
us_state_vaccinations4
```

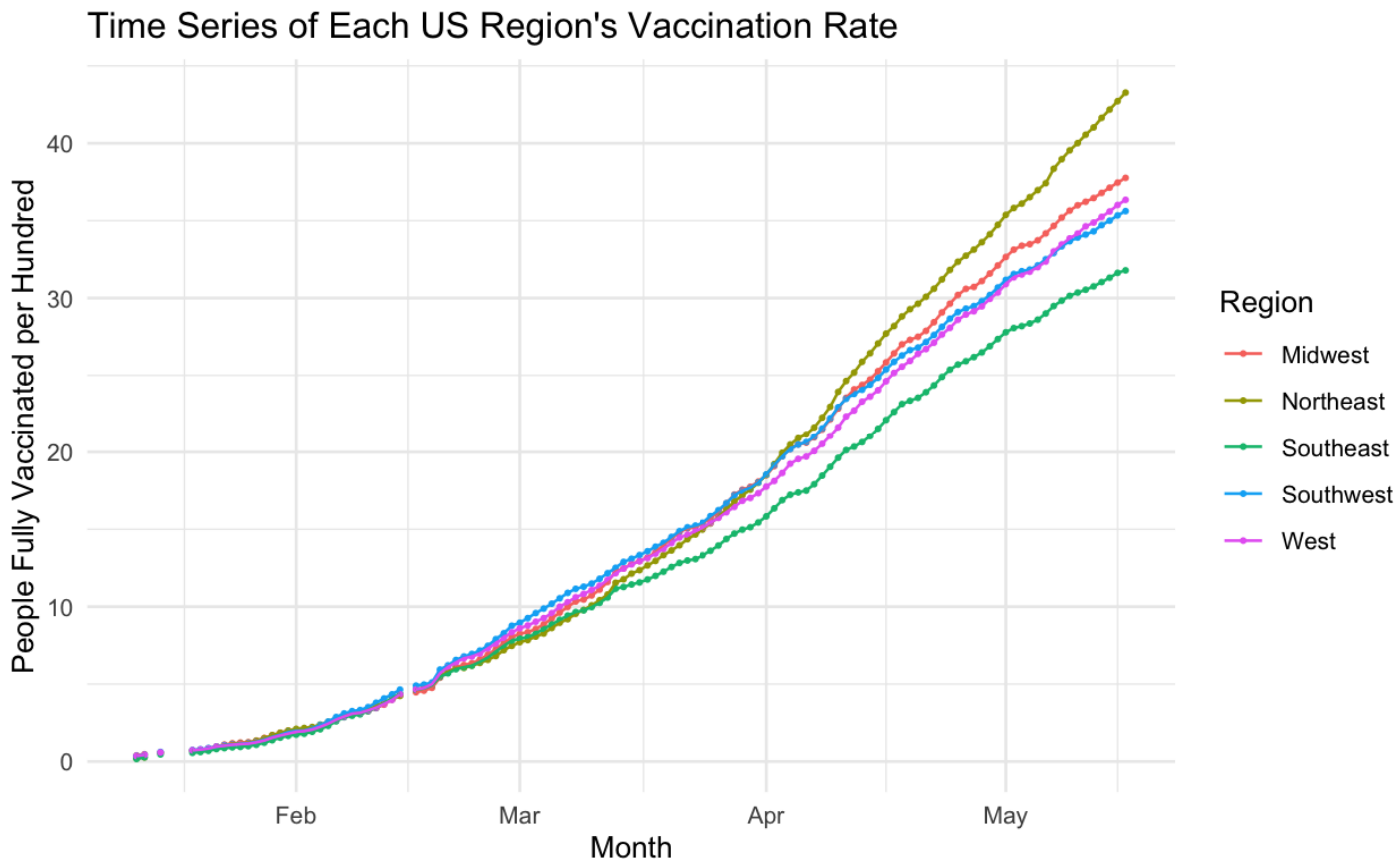
date	Region	people_fully_vaccinated_per_hundred
<date>	<chr>	<dbl>
2020-12-20	*NA*	NaN
2020-12-21	*NA*	NaN
2020-12-22	*NA*	NaN
2020-12-23	*NA*	NaN
2020-12-24	*NA*	NaN
2020-12-25	*NA*	NaN

date	Region	people_fully_vaccinated_per_hundred
<date>	<chr>	<dbl>
2020-12-26	*NA*	NaN
2020-12-27	*NA*	NaN
2020-12-28	*NA*	NaN
2020-12-29	*NA*	NaN

1-10 of 773 rows Previous **1** 2 3 4 5 6 ... 78 Next

Hide

```
ggplot(us_state_vaccinations4 %>%
         filter(Region %in% c("Northeast", "Southwest", "West", "Midwest", "Southeast"
)), aes(as.Date(date), people_fully_vaccinated_per_hundred, col = Region)) + geom_point
(size = .5) + theme_minimal() + geom_line() + xlab("Month") + ylab("People Fully Vaccina
ted per Hundred") + ggtitle("Time Series of Each US Region's Vaccination Rate")
```



I found this plot very interesting. My hypothesis was correct that vaccination rate differed in each region. You can see that in the Northeast, the vaccination rate reached almost 50% by the end of May, whereas in the Southeast, the vaccination rate was barely above 30%. This is a reasonable explation as to why Maine and Mississippi had such a large difference in cases, deaths, and vaccination rate.

I used several concepts from FDV in this chapter. I effectively used color, colorcoding each region in the legend. I appropriately capitalized all titles, legends, and axes. I incorporated a legend because it isnt apparent from just the plot what the different colors mean. I also incorporated the minimal theme as I did in all my previous plots.

Plot #5

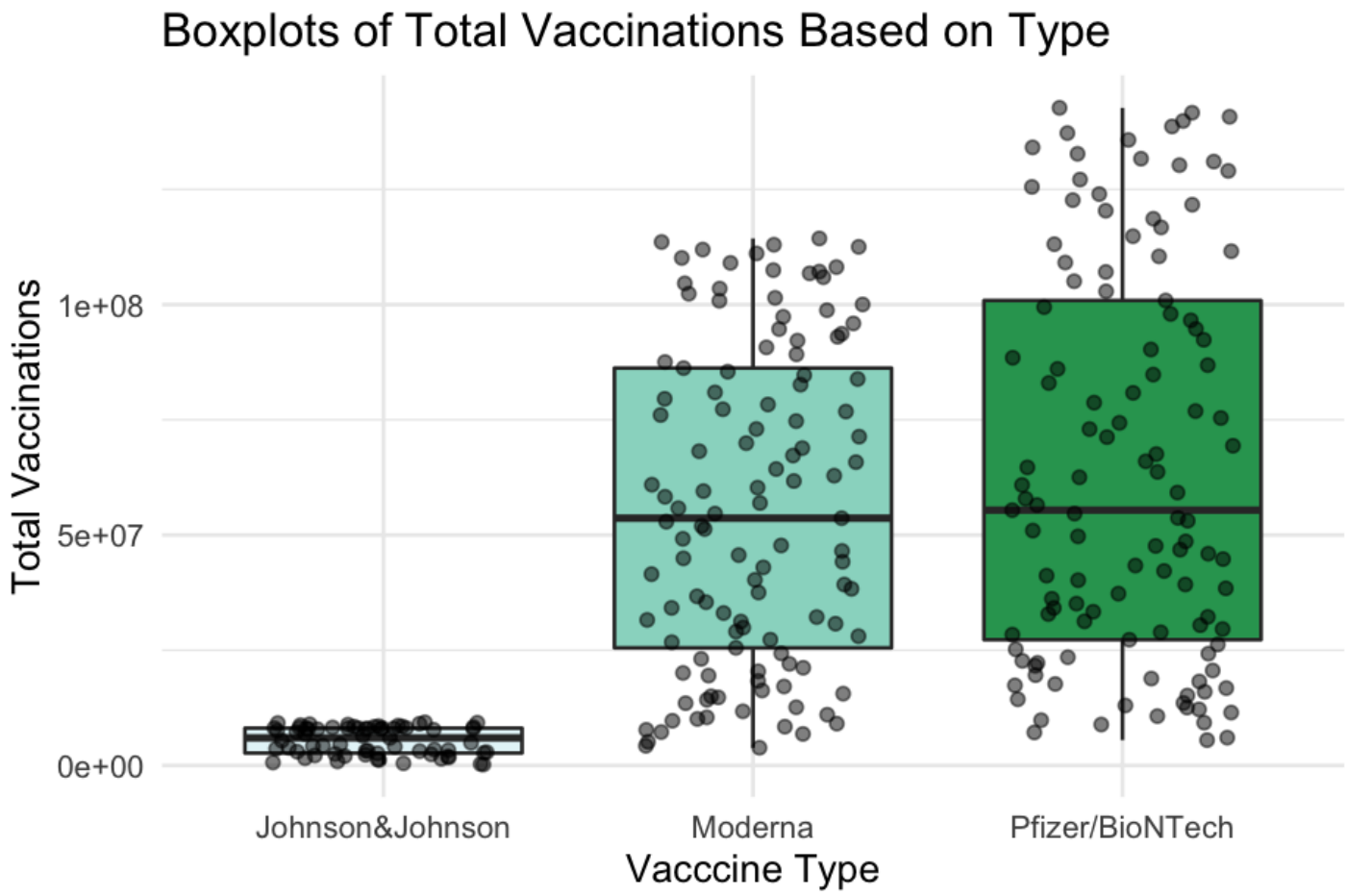
For my last plot I took a bit of a turn from what I analyzed in my previous plots. I was curious about the different vaccine brands and was hoping to analyze data that concerns the three different types of vaccines that were distributed in the US. I will create a plot which displays 3 boxplots that provide data about the total number of vaccinations produced for the US.

Hide

```
country_vaccinations_by_manufacturer1 <- country_vaccinations_by_manufacturer %>%
  filter(location == "United States")
country_vaccinations_by_manufacturer1
```

Hide

```
ggplot(country_vaccinations_by_manufacturer1, aes(x = vaccine, y = total_vaccinations))
+ geom_boxplot(fill = brewer.pal(3, "BuGn")) + geom_jitter(width = .3, alpha = .5, col =
"black") + theme_minimal() + ggtitle("Boxplots of Total Vaccinations Based on Type") + y
lab("Total Vaccinations") + xlab("Vacccine Type")
```



I found this plot very interesting because it really highlights how much more Moderna and Pfizer vaccines were used in the US vs the Johnson and Johnson vaccine. I chose to include the cloud points in this visualisation because I thought it helped to better understand the boxplots and add to the visual appearance. I thought it was interesting to see this data that has been being collected since the vaccine began getting distributed. These boxplots show that overall since these vaccine have begun distribution, Moderna and Pfizer have similar distributions, where as Johnson and Johnson is much different.

I used several concepts from FDV in this plot. For my use of color I utilized color brewer and chose a palette which I thought was visually appealing. I also used labels in which I made sure to capitalize appropriatly and provide useful information. I did not need to include a legend in this plot because I labeled the three different vaccine types under the boxplots.