R Notebook

Code

Honor Pledge:

“I understand I am allowed to discuss the data set and methodology with classmates in this class as I work on this project, but the decisions and actual work are my own. I certify that I have not done anyone else’s Final Project work for them. I acknowledge that I am not allowed to provide the data set from this assignment to any person or website outside of this class. I agree to delete this data set at the conclusion of the project.” Arjun Bhan

Hide

```
library(tidyverse)
library(Stat2Data)
library(maps)
library(USAboundaries)
library(sf)# Simple features standard.library
library(ggplot2)
library(RColorBrewer)
library(mapproj)
library(Stat2Data)
library(lubridate)
```

Hide

```
us_states <- read.csv("us-states.csv", stringsAsFactors = FALSE)
```

Hide

```
states98 <- us_states("1998-01-01")
```

Hide

```
us_states_cases<-us_states%>% group_by(state) %>% summarize(cases=sum(deaths)/sum(cases))
```

Hide

```
US_Map_Death<-merge(states98, us_states_cases, by.x = "state_name", by.y = "state")
US_Map_Death
```

```
Simple feature collection with 51 features and 19 fields
Geometry type: MULTIPOLYGON
Dimension:     XY
Bounding box:  xmin: -179.1294 ymin: 18.91228 xmax: -66.94993 ymax: 71.38961
Geodetic CRS:  WGS 84
First 10 features:
             state_name id_num                 name       id version start_date
1               Alabama      7              Alabama al_state       3 1820-12-19
2                Alaska      3               Alaska ak_state       1 1959-01-03
3               Arizona     15              Arizona az_state       1 1912-02-14
4              Arkansas     11             Arkansas ar_state       2 1840-05-21
5            California     19           California ca_state       2 1959-12-31
6              Colorado     20             Colorado co_state       1 1876-08-01
7           Connecticut     25          Connecticut ct_state       4 1804-12-31
8              Delaware     28             Delaware de_state       1 1783-09-03
9  District of Columbia     27 District of Columbia       dc       2 1846-09-07
10              Florida     39              Florida fl_state       1 1845-03-03
     end_date
1  2000-12-31
2  2000-12-31
3  2000-12-31
4  2000-12-31
5  2000-12-31
6  2000-12-31
7  2000-12-31
8  2000-12-31
9  2000-12-31
10 2000-12-31

change
1
MARION's overlap of the state of Mississippi and TUSCALOOSA's overlap of the state of Mississipp
i ended.
2
The state of Alaska was created from Alaska Territory by Presidential proclamation; Alaska Terri
tory eliminated.
3
The state of Arizona was created from Arizona Territory; Arizona Territory eliminated.
4                            Survey of boundary between the Republic of Texas and the United Sta
tes began. MILLER (original) officially became extinct and LAFAYETTE was eliminated from Texas w
hen Texas claims to the area were upheld.
5  Between 1949 and 1959 (precise date unknown) SAN FRANCISCO gained a small area of the Alameda
Naval Air Station when landfill expansion extended westward over the county line between SAN FRA
NCISCO and ALAMEDA in San Francisco Bay.
6
The state of Colorado was created from Colorado Territory; Colorado Territory eliminated.
7
HARTFORD lost part of the town of Southwick (the "Southwick Jog") to HAMPSHIRE (Mass.) when the
state boundary was adjusted.
8              The three Lower Counties, of KENT, NEW CASTLE, and SUSSEX became an independent st
ate on 4 July 1776. The name Delaware was formally adopted on 20 September 1776. The map depicts
state boundaries as of 3 September 1783.
9                           The federal government retroceded to Virginia all of the District of C
```

```
olumbia west of the Potomac River, including all of ALEXANDRIA (now ARLINGTON, Va.). ALEXANDRIA
eliminated from the District of Columbia.
10
The state of Florida was created from Florida Territory, with boundaries the same as those set i
n 1822; Florida Territory eliminated.

citation
1
(Ala. Acts 1820, 2d sess., secs. 1, 9/pp. 90, 92)
2
(Swindler, 1:208-225; Van Zandt, 165)
3  (U.S. Stat., vol. 34, pt. 1, ch. 3335[1906]/pp. 267-285; U.S. Stat., vol. 36, pt. 1, ch. 310
[1910]/pp. 557-579; U.S. Stat., vol. 37, pt. 1, res. 8[1911]/pp. 39-43; pt. 2[1912]/pp. 1728-172
9)
4
(U.S. Stat., vol. 5, ch. 75 [1844]/p. 674; Marshall, 235-236)
5
(U.S.G.S., 7.5 Minute Series, Oakland West Quadrangle, ";Edition of 1949"; and ";Edition of 195
9";)
6
(Van Zandt, 141; U.S. Stat., vol. 18, part 3 [1876], p. 474)
7
(Hooker, 25-26; Van Zandt, 69)
8
(Declaration of Independence; Swindler, 2:197)
9
(U.S. Stat., vol. 9, ch. 35 [1846]/pp. 35-37, and appendix 3/p. 1000)
10
(Swindler, 2:332; U.S. Stat., vol. 5, ch. 48 [1845], secs. 1, 5/pp. 742-743)
    start_n    end_n area_sqmi            terr_type            full_name abbr_name
1  18201219 20001231     51656                State              Alabama        AL
2  19590103 20001231    575301                State               Alaska        AK
3  19120214 20001231    113999                State              Arizona        AZ
4  18400521 20001231     53179                State             Arkansas        AR
5  19591231 20001231    158097                State           California        CA
6  18760801 20001231    104093                State             Colorado        CO
7  18041231 20001231      4975                State          Connecticut        CT
8  17830903 20001231      2013                State             Delaware        DE
9  18460907 20001231        68 District of Columbia District of Columbia        DC
10 18450303 20001231     56618                State              Florida        FL
                          name_start state_abbr state_code       cases
1               Alabama (1820-12-19)         AL         01 0.018154422
2                Alaska (1959-01-03)         AK         02 0.004728442
3               Arizona (1912-02-14)         AZ         04 0.020067687
4              Arkansas (1840-05-21)         AR         05 0.016375293
5            California (1959-12-31)         CA         06 0.015440396
6              Colorado (1876-08-01)         CO         08 0.016223184
7           Connecticut (1804-12-31)         CT         09 0.038131053
8              Delaware (1783-09-03)         DE         10 0.019449469
9  District of Columbia (1846-09-07)         DC         11 0.030534288
10              Florida (1845-03-03)         FL         12 0.016984627
                         geometry
1  MULTIPOLYGON (((-88.07462 3...
2  MULTIPOLYGON (((-179.0575 5...
```

```
3  MULTIPOLYGON (((-110.7507 3...
4  MULTIPOLYGON (((-91.67235 3...
5  MULTIPOLYGON (((-117.2326 3...
6  MULTIPOLYGON (((-106.0794 4...
7  MULTIPOLYGON (((-73.5055 41...
8  MULTIPOLYGON (((-75.07024 3...
9  MULTIPOLYGON (((-76.9094 38...
10 MULTIPOLYGON (((-81.74982 2...
```

Hide

```
count=1
for (i in US_Map_Death$cases)
{
  if (US_Map_Death$cases[count]<=.01)
  {
    US_Map_Death$Percen[count]="0-1%"
  }
  else if (US_Map_Death$cases[count]<=.02 && US_Map_Death$cases[count]>.01)
  {
    US_Map_Death$Percen[count]="1-2%"
  }
    else if (US_Map_Death$cases[count]<=.03 && US_Map_Death$cases[count]>.02)
  {
    US_Map_Death$Percen[count]="2-3%"
  }
  else if (US_Map_Death$cases[count]<=.04 && US_Map_Death$cases[count]>.03)
  {
    US_Map_Death$Percen[count]="3-4%"
  }
  else
  {
    US_Map_Death$Percen[count]="Greater than 4%"
  }
  count=count+1
}
```

Hide

```
print(US_Map_Death$Percen[1])
```

```
[1] "1-2%"
```

Hide

```
print(US_Map_Death$cases[1])
```

```
[1] 0.01815442
```

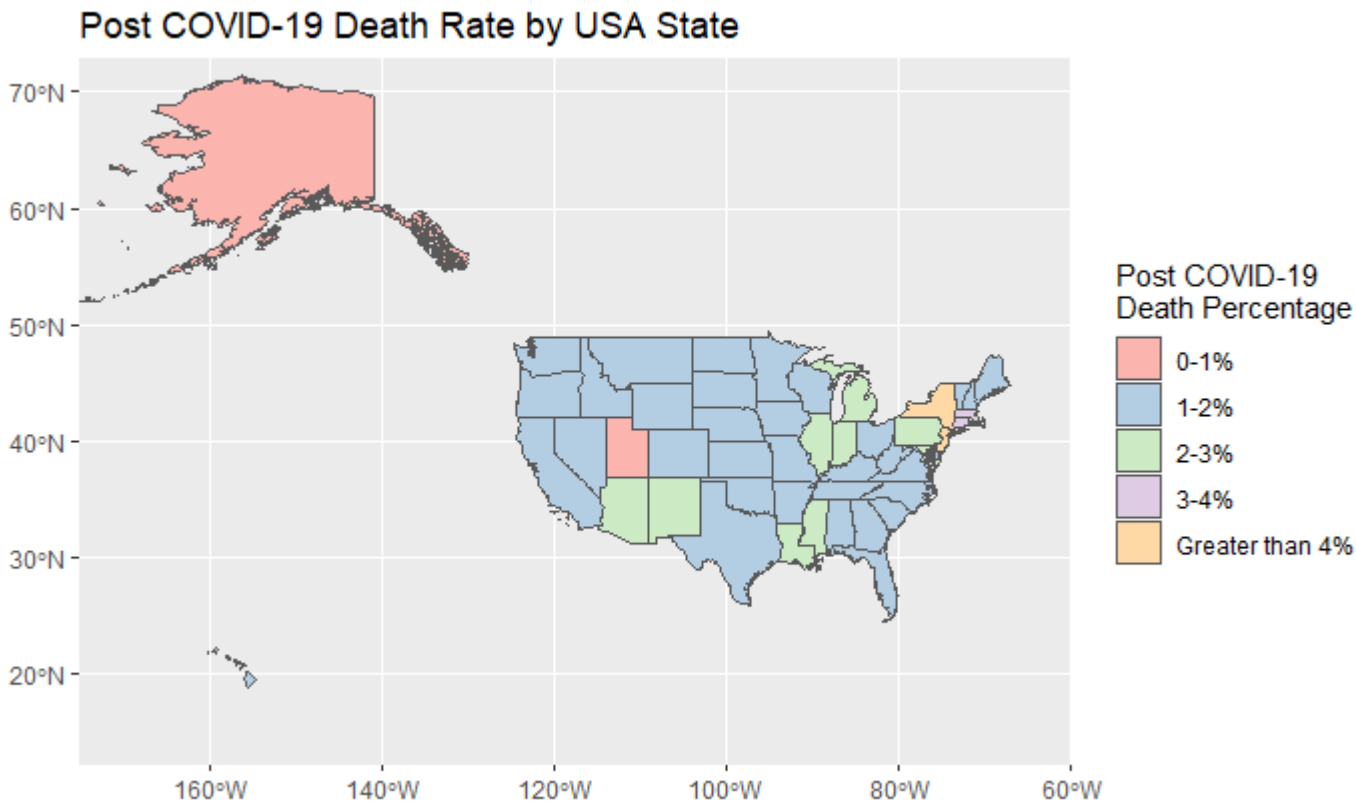
Hide

```
print(US_Map_Death$state_name[1])
```

```
[1] "Alabama"
```

Hide

```
ggplot(US_Map_Death)+geom_sf(mapping = aes(fill = factor(Percen)))+ coord_sf(xlim = c(-170, -65
), ylim = c(15,70), crs = st_crs(4269))+labs(title="Post COVID-19 Death Rate by USA State",x=ele
ment_blank(),y=element_blank(),fill="Post COVID-19\nDeath Percentage")+scale_fill_brewer(palette
="Pastel1")
```



This visual is a choropleth that shows the chances of dying if one has contracted COVID for each state in America. We can see that the residents of New York and New Jersey have the highest probability of dying once they have contracted COVID. Similarly, residents of Alaska and Utah are the least likely to die after contracting COVID. The difference between states in their post-COVID death rates could be explained by one of the following factors: difference in health care systems, strains of the virus prevalent in the state or the local climate that can affect COVID's ability to spread.

I chose not to label the x and y axis because I felt that the longitude and the latitude were unrelated to the post-COVID death rate metric and it distracted the reader from the main point of the choropleth. The reader can determine what the axes represent from other data in the graph such as the cardinal directions being shown in axes and by the visual of the map. I labeled the color bar "Post-COVID-19 Death Percentage" so the reader understands what the colors represent. I transformed the Post COVID death rate values into percentages so it would be easier for the reader to grasp immediately. I labeled the title "Post COVID-19 Death Rate by USA State" so the reader could get a general idea of what the data represents. I used a pastel1 color filter as its colors are very distinct from one another and are appealing. I chose to use a choropleth as my visual as it is effective in showing difference in data related to different states.

Based on Chapter 15 of FDV reading, I had chosen not to use a continuous color scale for the post COVID-19 death percentage. That color scale is hard for people to understand. Matching the slight color variations to the values can be difficult for the reader. For this reason, I decided binning the data values into a discrete color scale would make it easier for the reader to distinguish between states.

Hide

```
Cont_Vacs_Manu_Brand <- read.csv("country_vaccinations_by_manufacturer.csv", stringsAsFactors = FALSE)
```

Hide

```
unique(Cont_Vacs_Manu_Brand)
```

	location <chr>	date <chr>	vaccine <chr>	total_vaccinations <int>
1	Chile	2020-12-24	Pfizer/BioNTech	420
2	Chile	2020-12-25	Pfizer/BioNTech	5198
3	Chile	2020-12-26	Pfizer/BioNTech	8338
4	Chile	2020-12-27	Pfizer/BioNTech	8649
5	Chile	2020-12-28	Pfizer/BioNTech	8649
6	Chile	2020-12-29	Pfizer/BioNTech	8649
7	Chile	2020-12-30	Pfizer/BioNTech	8649
8	Chile	2020-12-31	Pfizer/BioNTech	8649
9	Chile	2021-01-01	Pfizer/BioNTech	8649
10	Chile	2021-01-02	Pfizer/BioNTech	8649

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

Hide

```
Cont_Vacs_Manu_Brand<-Cont_Vacs_Manu%>%group_by(vaccine,location) %>% summarize(VaccinesCount=round(max(total_vaccinations)/1000))
```

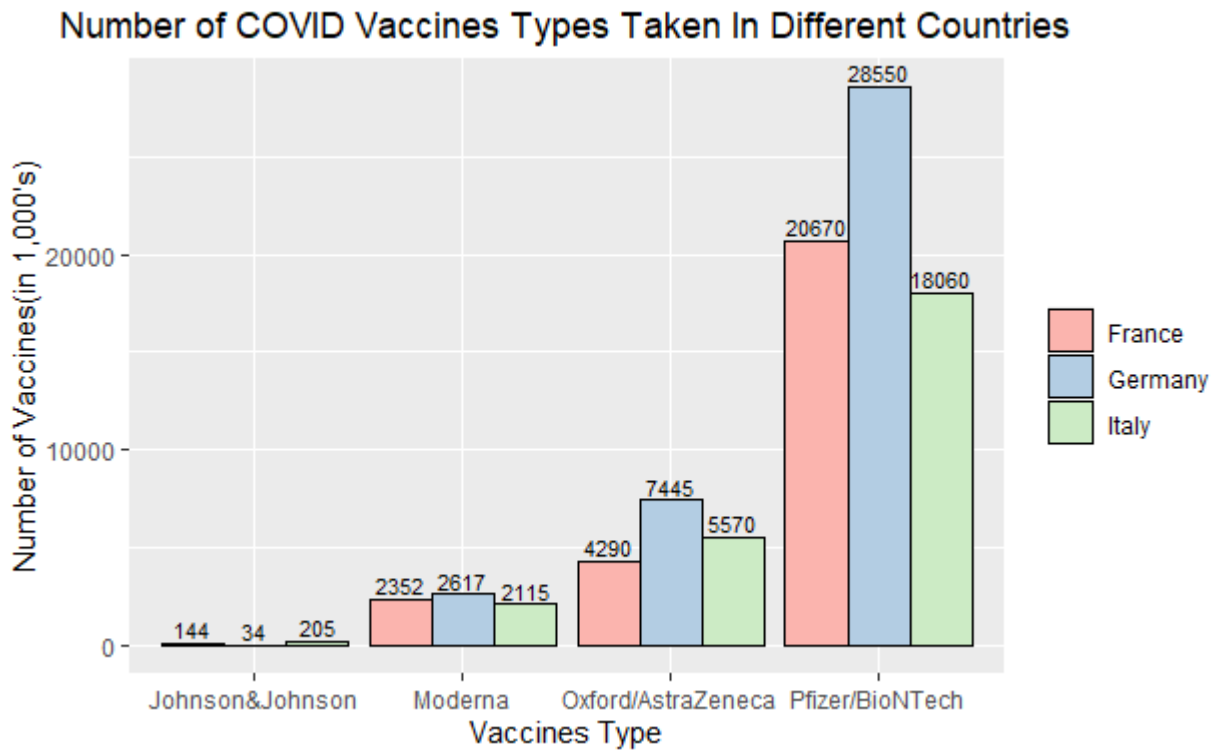
```
`summarise()` has grouped output by 'vaccine'. You can override using the `.groups` argument.
```

Hide

```
Coun_Vac<-filter(Cont_Vacs_Manu_Brand,location=="France"|location=="Germany"|location=="Italy")
```

Hide

```
ggplot(Coun_Vac, aes(x=vaccine,y=VaccinesCount,fill=location))+geom_col(position="dodge", colour
="black")+labs(title="Number of COVID Vaccines Types Taken In Different Countries",x="Vaccines T
ype" ,y="Number of Vaccines(in 1,000's)",fill=element_blank())+theme(plot.title = element_text(h
just=0.5))+scale_fill_brewer(palette="Pastel1")+geom_text(aes(label=VaccinesCount), size=3, vjus
t=-.3, position = position_dodge(.9))
```



This bar graph is a visual display of the amount of different vaccines administered in France, Germany and Italy. We can see from the graph that the number of Johnson&Johnson vaccines administered in every country shown is far less than the other vaccine brands. In contrast, Pfizer/BioNTech is the vaccine used the most by far in each of 3 country shown. This could be because of several reasons such as: the population size, date when vaccines were manufactured and made available in that country,the price for the vaccine, ability of people to pay for it and the amount of publicity for each vaccine. Italy seems to be having the least amount of vaccine administered for almost every vaccine brand and Germany appears to have the most. These difference could be because of the population size and the health facilities in each country. For example, Germany has the largest population and economy and its healthcare is among the best The exception for both of these cases is when it comes to the Johnson&Johnson vaccine in which Italy has the most vaccine administered and Germany has the least. It is possible that Italy had better access for this specific vaccine and that is why it has more of Johnson&Johnson adminstered the other countries.

I chose to label the y-axis “Number of Vaccines(in 1,000’s)” to inform the reader that the y-axis is showing number of vaccines administered in units of 1,000. I chose to label the x axis “Vaccine type” to indicate the manufacturer of the vaccine. I labeled the graph “Number of Vaccines by Type Taken In Germany, Italy and France” to allow the reader to understand the meaning of the graph quickly. I chose not to label the color bar as I felt that the audience would be able to understand the its meaning by the countries names and the title. I chose to place the number of vaccines taken by brand right above each bar so that the reader could quickly understand the numerical values represented by the height of each bar. I also altered the text size so that it would fit within the boundaries of the bar chart. The colors of the bars allow the audience to differentiate the data based on country.

Based on Chapter 3 of the FDV reading I had chosen to not stack the bar graph so that the data could be more easily compared. Also, I chose not to include more countries in my graph in order to not cluster the visual.

Hide

```
owid_covid <- read.csv("owid-covid-data.csv", stringsAsFactors = FALSE)
```

Hide

```
unique(owid_covid$continent)
```

```
[1] "Asia"          ""              "Europe"        "Africa"        "North America"
[6] "South America" "Oceania"
```

Hide

```
owid_covid<-filter(owid_covid,continent=="Europe"|continent=="Africa"|continent=="Asia"|continen
t=="South America"|continent=="North America")
owid_covid <- owid_covid%>% gather(
 male_smokers:female_smokers, key = "gender",
 value = "SmokeVal",
 na.rm = TRUE
)
owid_covid
```

	iso_code <chr>	continent <chr>	location <chr>	date <chr>	total_cases <dbl>	new_ca… <dbl>	new_cases_smoothed <dbl>
450	ALB	Europe	Albania	2020-02-25	*NA*	*NA*	*NA*
451	ALB	Europe	Albania	2020-02-26	*NA*	*NA*	*NA*
452	ALB	Europe	Albania	2020-02-27	*NA*	*NA*	*NA*
453	ALB	Europe	Albania	2020-02-28	*NA*	*NA*	*NA*
454	ALB	Europe	Albania	2020-02-29	*NA*	*NA*	*NA*
455	ALB	Europe	Albania	2020-03-01	*NA*	*NA*	*NA*
456	ALB	Europe	Albania	2020-03-02	*NA*	*NA*	*NA*
457	ALB	Europe	Albania	2020-03-03	*NA*	*NA*	*NA*
458	ALB	Europe	Albania	2020-03-04	*NA*	*NA*	*NA*
459	ALB	Europe	Albania	2020-03-05	*NA*	*NA*	*NA*

1-10 of 120,255 rows | 1-9 of 59 columns Previous **1** 2 3 4 5 6 … 100 Next

Hide

```
us_state_vac <- read.csv("us_state_vaccinations.csv", stringsAsFactors = FALSE)
```

This box plot shows the smoking values of different continents based on gender. As we can see from this graph, females has a lower smoking value on average than males for every continent. This could be because of social expectations of woman not to smoke. The difference between the average male and female smoking values is greatest for Asia. This could be because countries in this continent have harsher social expectations for woman.

I choose to not label the x-axis as the reader would be able to obtain this information from the title and by the continents being shown as data points. For the y-axis, I labeled it “Smoking Value” so that the reader would have no issues understanding what it values represent. I made the title Smoking Value based on Gender and Continent so that the reader would be able to quickly understand the visual’s meaning. I chose to use a color palette so that the reader could easily identify between the gender values (Males vs. Females). I choose to use a box plot for its ability to show the percentiles of its data. I choose to use the title “Smoking Value” based on Gender and Continent to allow the reader to quickly understand what the visual is about.

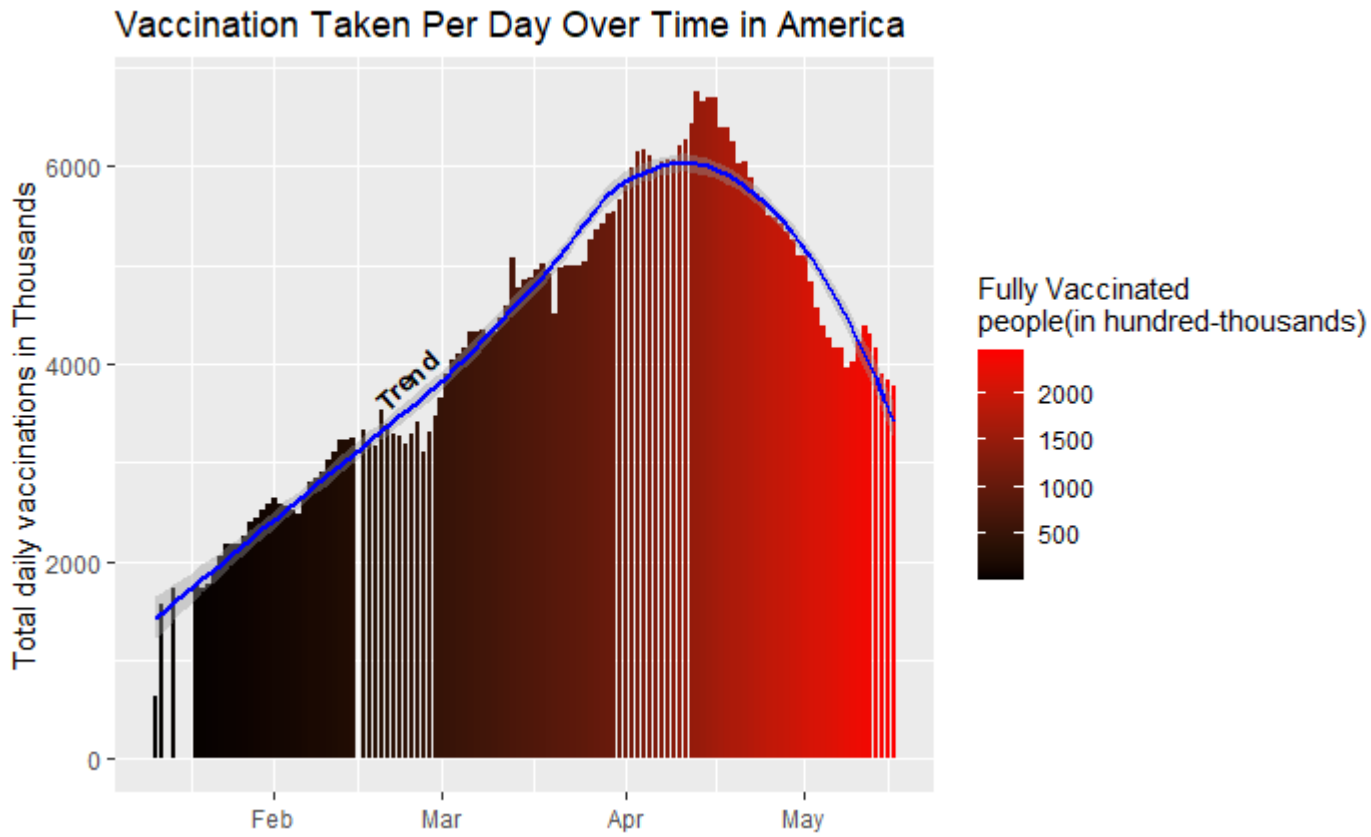
Hide

```
us_state_vac <- read.csv("us_state_vaccinations.csv", stringsAsFactors = FALSE)
```

Hide

```
us_state_vac$date<-ymd(us_state_vac$date)
us_state_vac<-us_state_vac%>%drop_na()
us_state_vac_sor<-us_state_vac%>%group_by(date) %>% summarize(tot_daily_vaccinations=sum(daily_v
accinations)/1000,people_fully_vaccinated=sum(people_fully_vaccinated)/100000)

ggplot(us_state_vac_sor,aes(x=date,y=tot_daily_vaccinations,fill=people_fully_vaccinated))+geom_
col()+geom_smooth(color="blue")+scale_fill_gradient(low="black",high="red")+labs(title="Vaccinat
ion Taken Per Day Over Time in America",x=element_blank() ,y="Total daily vaccinations in Thousa
nds",fill="Fully Vaccinated \npeople(in hundred-thousands)")+geom_text(data=us_state_vac_sor[43
,],label="Trend",angle=42,vjust=-1.7,size=3.5,fontface="bold")
```



Hide

NA

This time series shows the changes in number of vaccination taken and fully vaccinated people per day in America. We can see that there is a general positive trend of vaccinations taken per day till the middle of April in which the number of vaccinations being taken start decreasing. It is possible that for the first few months of the vaccine its usage rose due to the severity of the COVID virus. The decrease in the middle of April could because of the large number of people being already vaccinated. The color bar supports this hypothesis by showing that the number of people fully vaccinated always increases over time.

I didn't label the X-axis as I felt that the reader could indicate that it was about time from the data values that are visualized being months and by the title. I choose the Y axis to be Total daily vaccinations in Thousands and the color bar to be Fully Vaccinated people(in hundred-thousands) as I felt without those labels the reader might be confused on what the values represent. I choose to represent the number of vaccines taken per day as columns as I felt they would be an effective way of representing a large number of numeric data. I choose to plot a smooth line in the chart as the FDV reading states it is a great way to show trends in data within a time series. I choose to label the smooth line trend so that the reader would understand the meaning of the line.