Question 15

Chloe J. Welch

7/4/2022

Q15. [10pt] Obtain the most recently dated COVID-19 Variant Data from the California Health and Human Services (CHHS) open data site:

https://data.chhs.ca.gov/dataset/covid-19-variant-data

Upload to gradescope a PDF format report generated from an Rmarkdown document that demonstrates reading the above CSV file and generating the below visualization of this data.

NB. You can chose how to make this plot and whether you want to make improvements or stylistic changes. However, you are strongly encouraged to use the ggplot2, lubridate and dplyr packages for this task. Please make sure your name and PID number is on the first page and that your report contains all of your code, text description/narrative text of why you doing a particular task/code chunk and the resulting figure.

We will begin by loading the necessary packages to analyze this data set. These were already installed in BGGN 213 during the fall quarter.

```
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Next, we will import the COVID-19 variant data.

```
covid.data <- read.csv("covid19_variants.csv")
head(covid.data)
```

```
##         date       area area_type variant_name specimens percentage
## 1 2021-01-01 California     State        Alpha         1       1.69
## 2 2021-01-01 California     State           Mu         0       0.00
## 3 2021-01-01 California     State        Other        29      49.15
## 4 2021-01-01 California     State        Delta         0       0.00
## 5 2021-01-01 California     State         Beta         0       0.00
## 6 2021-01-01 California     State        Total        59     100.00
##   specimens_7d_avg percentage_7d_avg
## 1               NA                NA
## 2               NA                NA
## 3               NA                NA
## 4               NA                NA
## 5               NA                NA
## 6               NA                NA
```

We can use `lubridate` to help make dealing with the dates of the data set a little simpler.

```
covid.data$date <- ymd(covid.data$date)
head(covid.data$date)
```

```
## [1] "2021-01-01" "2021-01-01" "2021-01-01" "2021-01-01" "2021-01-01"
## [6] "2021-01-01"
```

Next, let's filter out the entries we do not need.

```
filtered.data <- filter(covid.data, variant_name != "Total" & variant_name != "Other")
head(filtered.data)
```

```
##         date       area area_type variant_name specimens percentage
## 1 2021-01-01 California     State        Alpha         1       1.69
## 2 2021-01-01 California     State           Mu         0       0.00
## 3 2021-01-01 California     State        Delta         0       0.00
## 4 2021-01-01 California     State         Beta         0       0.00
## 5 2021-01-01 California     State        Gamma         0       0.00
## 6 2021-01-01 California     State      Epsilon        28      47.46
##   specimens_7d_avg percentage_7d_avg
## 1               NA                NA
## 2               NA                NA
## 3               NA                NA
## 4               NA                NA
## 5               NA                NA
## 6               NA                NA
```

Let's use `ggplot2` to plot this data. The title of the plot will be "COVID-19 Variants in California" with the dates on the x-axis and the percentage of sequenced specimens on the y-axis. The date breaks are equal to 1 month so each individual month is clearly represented. The different colors represent the different variants of COVID-19 and are denoted in the legend to the right of the plot.

```
covid.data.plot <- ggplot(filtered.data, aes(date, percentage)) +
  geom_line(aes(color = variant_name)) +
  labs(x="", y="Percentage of sequenced specimens",
       title= "COVID-19 Variants in California", color="") +
  scale_x_date(date_breaks="1 month", date_labels="%b %Y")

covid.data.plot + theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```