Covid-19 Variant Data

Sam Altshuler
PID: A59010373

7/14/2022

Load in correct libraries

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

Load in the Covid-19 Variant data

Read in the data from a csv. Data downloaded from the California Health and Human Services website.

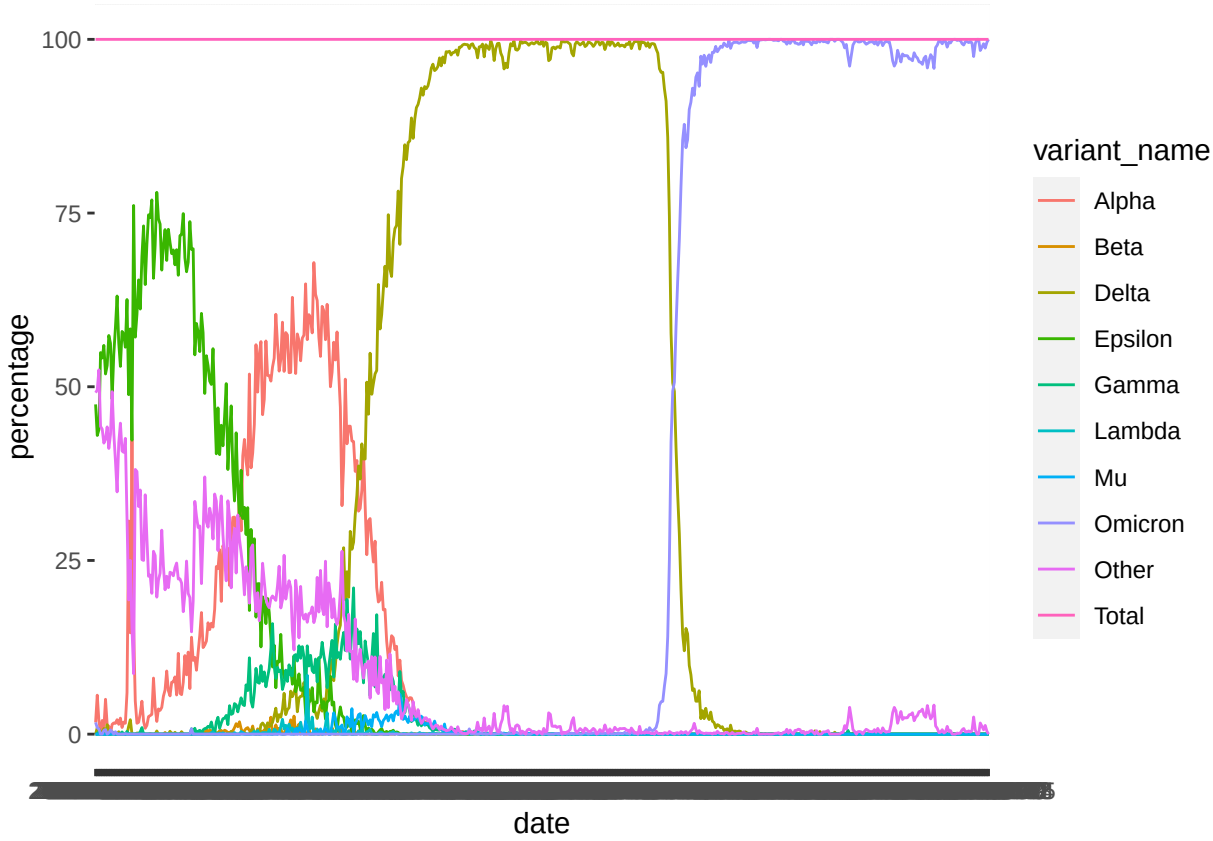
```
total_data <- read.csv("covid19_variants.csv")
# Preliminary look at the data
head(total_data)
```

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

Start to graph the variants over time

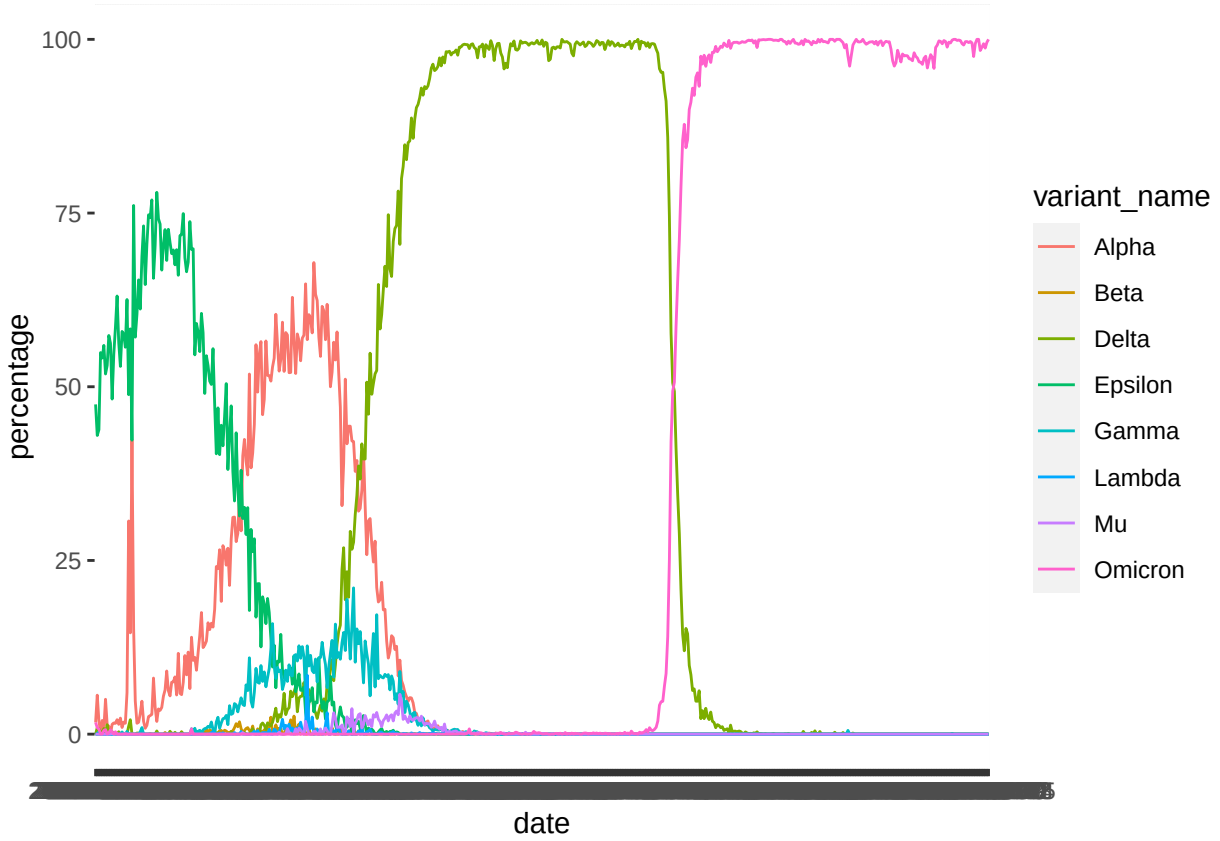
Next we can preliminarily graph the data using ggplot and `geom_line()` to see what it looks like.

```
ggplot(total_data, aes(x = date, y = percentage, group = variant_name, color = variant_name))+
  geom_line()
```



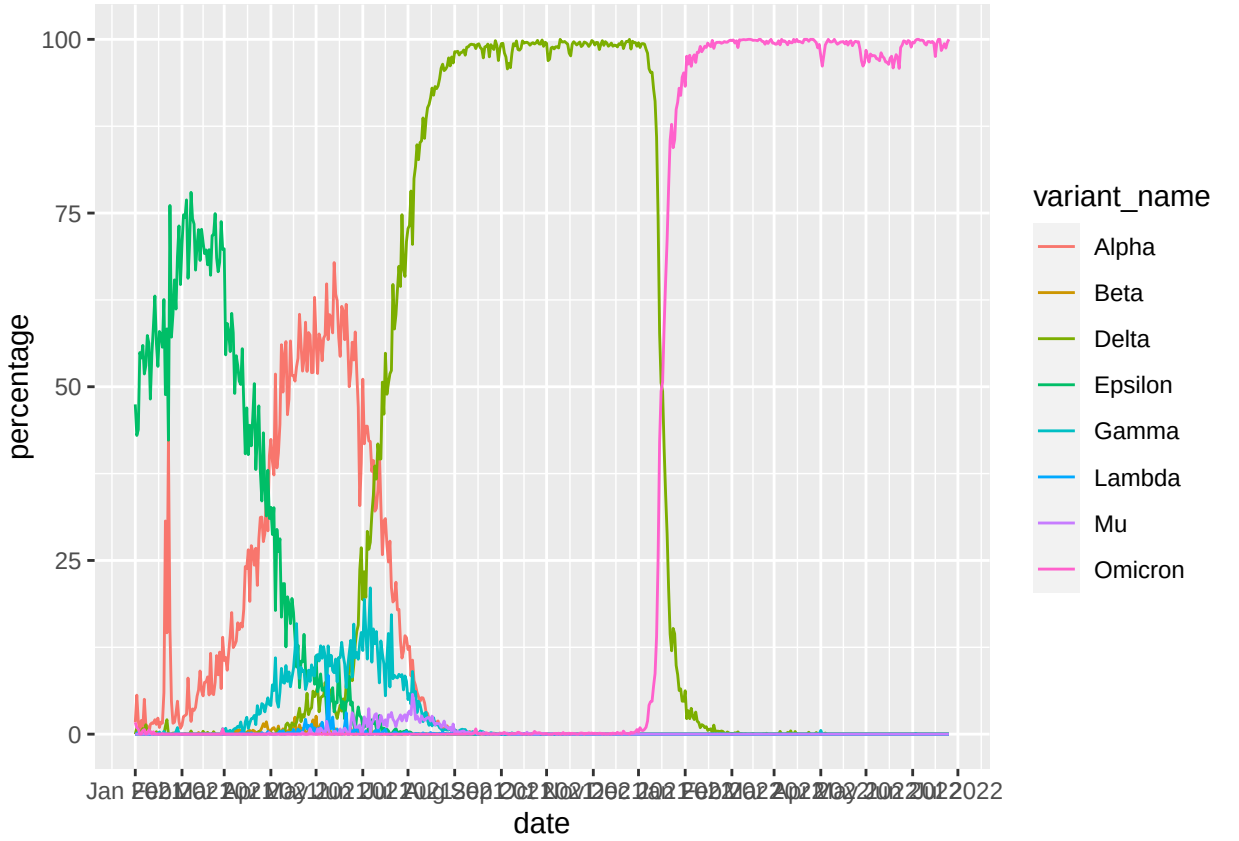
Now this looks similar to example graph, but the next step is to remove all rows that have "Total" variants as this row will always be 100%. We also need to remove the "other" rows.

```
# Filter out all rows that have a variant name of "Total"
no_total <- total_data %>% filter(variant_name != "Total" )

# Filter out all rows with a variant name of "Other"
variants <- no_total %>% filter(variant_name != "Other")

# Graph to ensure that total and other are no longer represented on the graph
ggplot(variants, aes(x = date, y = percentage, group = variant_name, color = variant_name))+
  geom_line()
```



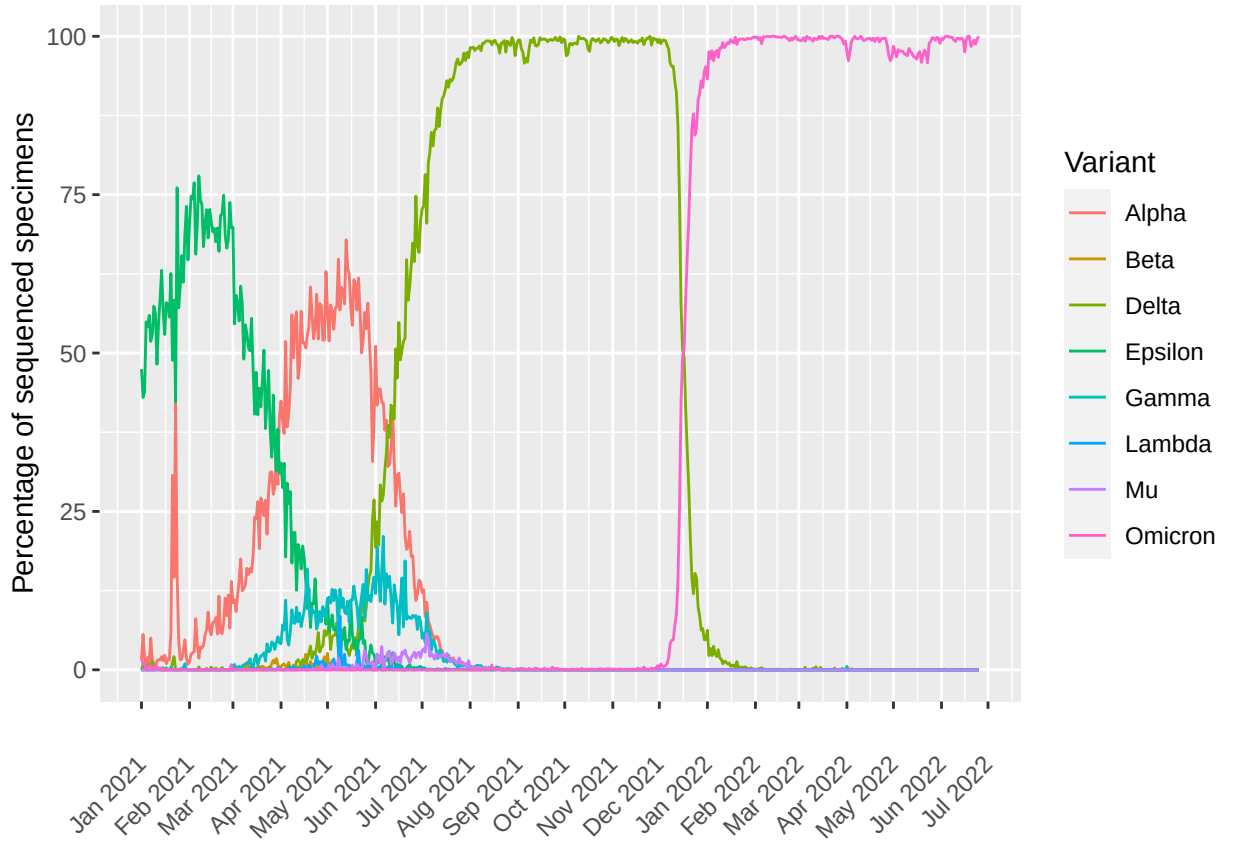
Next we want to change the x axis to 1 month increments that show the 3 letters of the month with the year. This is done by using the lubridate package. But first, we need to reformat the date column into a format that lubridate can use.

```
# Convert each date in the date column into the lubridate date class.
variants$date <- as_date(variants$date)

#graph again, with the X axis broken into 1 month increments and displaying the month and year
ggplot(variants, aes(x = date, y = percentage, group = variant_name, color = variant_name))+
  geom_line()+
  scale_x_date(date_breaks = "1 month", date_labels = "%b %Y")
```



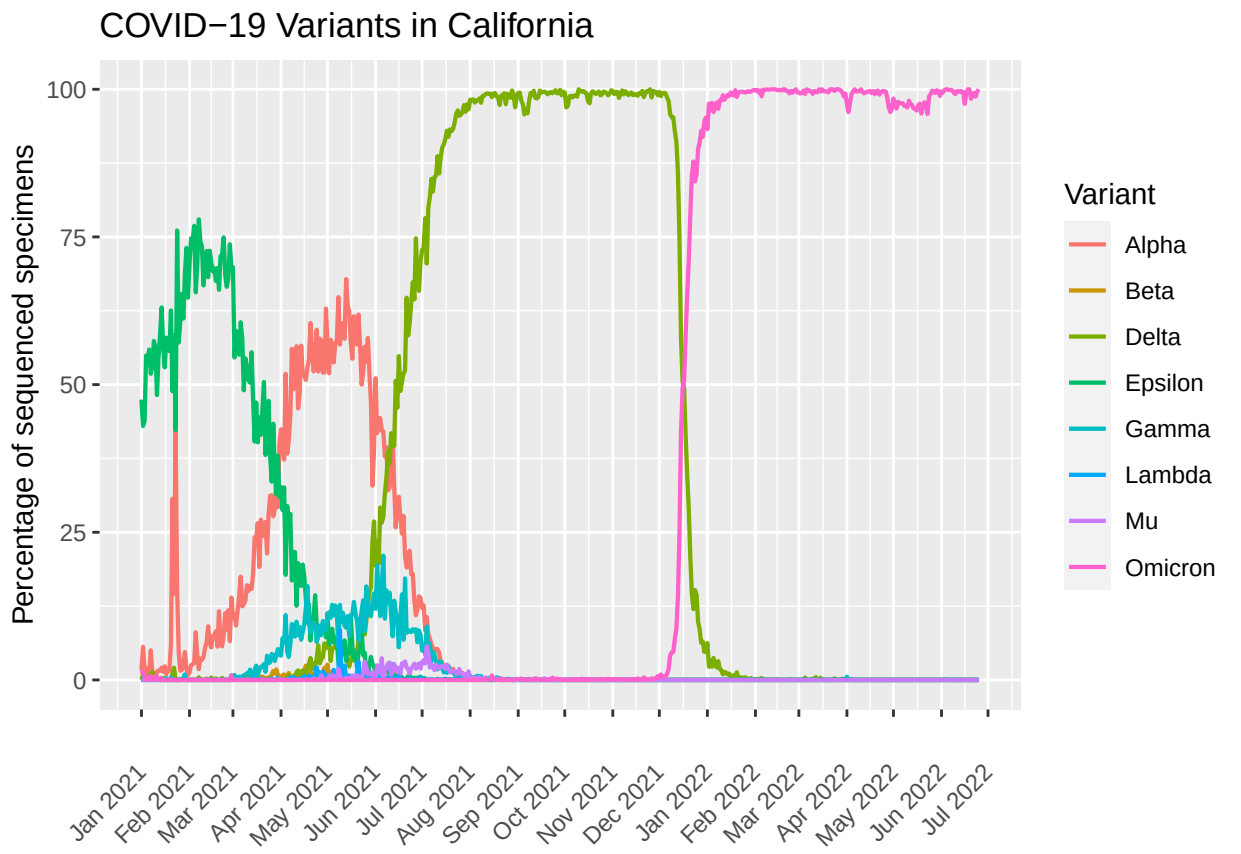
Now that our graph now has the correct axes, let's make it more readable by angling the x axis text so it doesn't overlap. We will also get rid of the X axis title since dates are already pretty descriptive of the X axis. Lastly, we will add a title to the legend.

```
ggplot(variants, aes(x = date, y = percentage, group = variant_name, color = variant_name))+
  geom_line()+
  scale_x_date(date_breaks = "1 month", date_labels = "%b %Y")+
  labs(y ="Percentage of sequenced specimens", x = "", color = "Variant" )+
  # Angle the x axis text so that they don't overlap
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust=1))
```



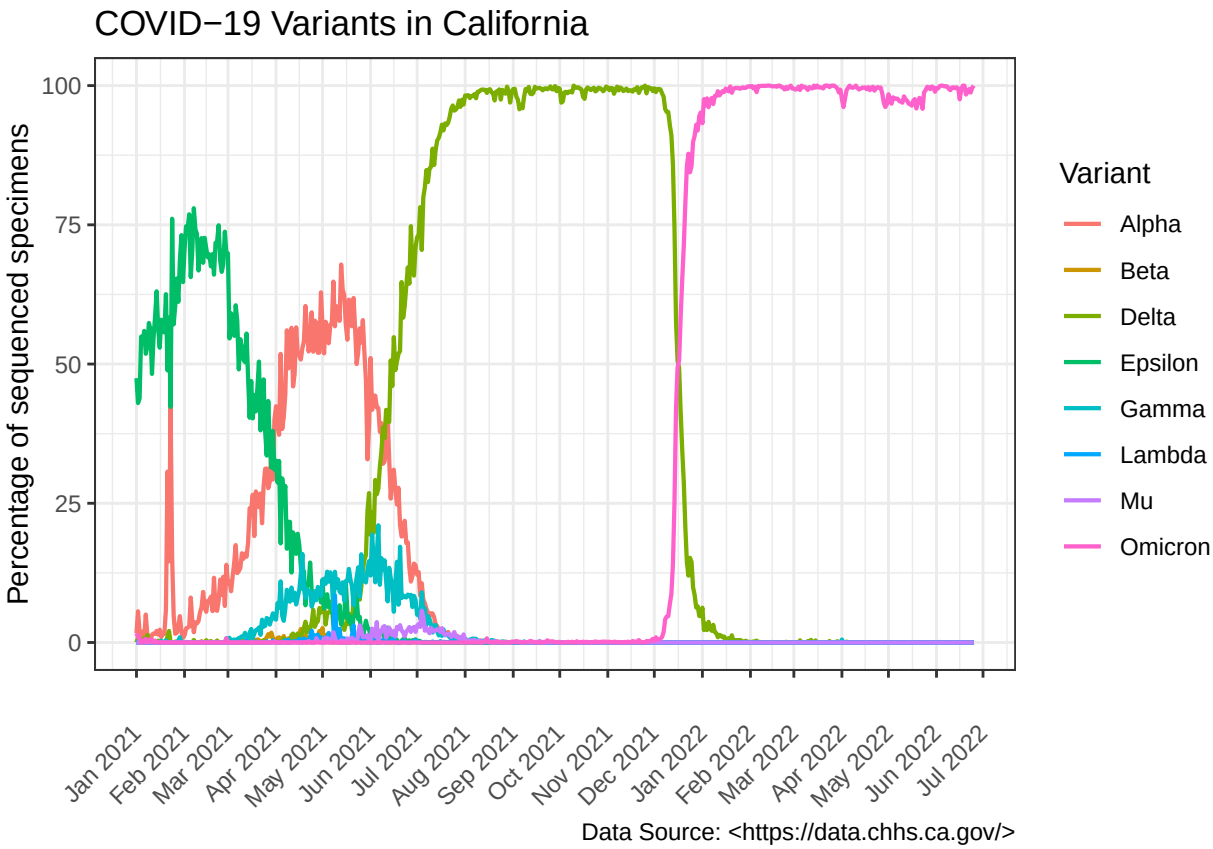
Now that the labels and axes are nice and readable, let's make the lines a bit thicker to be easier to see and add a title to the graph.

```
ggplot(variants, aes(x = date, y = percentage, group = variant_name, color = variant_name))+
  geom_line(size = 0.75)+ #change the width of the lines
  scale_x_date(date_breaks = "1 month", date_labels = "%b %Y")+
  labs(y ="Percentage of sequenced specimens", x = "", color = "Variant",
       title = "COVID-19 Variants in California")+
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust=1))
```



Lastly, we need to add a caption to state where the data is from. This will be the final product!

```
ggplot(variants, aes(x = date, y = percentage, group = variant_name, color = variant_name))+
  geom_line(size = 0.75)+ #change the width of the slide
  scale_x_date(date_breaks = "1 month", date_labels = "%b %Y")+
  labs(y ="Percentage of sequenced specimens", x = "", color = "Variant",
       title = "COVID-19 Variants in California",
       # Adds in a caption on the bottom right of the graph to show the data source
       caption = "Data Source: <https://data.chhs.ca.gov/>")+
  theme_bw()+
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust=1))
```


COVID−19 Variants in California
Percentage of sequenced specimens
100
75
50
25
0
Jan 2021
Feb 2021
Mar 2021
Apr 2021
May 2021
Jun 2021
Jul 2021
Aug 2021
Sep 2021
Oct 2021
Nov 2021
Dec 2021
Jan 2022
Feb 2022
Mar 2022
Apr 2022
May 2022
Jun 2022
Jul 2022
Data Source: <https://data.chhs.ca.gov/>
Variant
Alpha
Beta
Delta
Epsilon
Gamma
Lambda
Mu
Omicron