# 1st Year Bioinformatics Exam

Jennifer Swindlehurst Chan

---

Name: Jennifer Swindlehurst Chan

PID: 12415558

## Covid Variant Data

From the California Health and Human Services (CHHS) open data site.

```
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

First. we load in all the packages we will use. I ended up not using lubridate.

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

Next, we upload the csv file with the data will be plotting. We can use the head() function to take a peak at what our data looks like.

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

```
        date       area area_type variant_name specimens percentage
1 2021-01-01 California     State        Alpha         1       1.67
2 2021-01-01 California     State        Other        29      48.33
3 2021-01-01 California     State        Delta         0       0.00
4 2021-01-01 California     State        Gamma         0       0.00
5 2021-01-01 California     State      Omicron         1       1.67
```

```
6 2021-01-01 California     State        Total        60     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
```

Next, we can start to filter for what we want included. We only want the data relating to variants Alpha, Beta, Delta, Epsilon, Gamma, Lambda, Mu, and Omicron. We do not want Total or Other. This subsection of the data is now saved at data_2. We can also work with the date of each measurement easier by using the as.Date function. The example graph only plots until May 2022 so we can take our data_2 and section it out further to be the data from the beginning of the data set (January 1, 2021) until May 1, 2022. This is now saved as data_3. We can use the tail() function to check that the data is cut off at May 1, 2022.

```
data$date <- as.Date(data$date)
data_2 <- dplyr::filter(data, variant_name %in% c("Alpha", "Beta",
    "Delta", "Epsilon", "Gamma", "Lambda", "Mu", "Omicron"))
data_3 <- data_2[data_2$date >= "2021-01-01" & data_2$date <=
    "2022-05-01", ]

head(data_2)
```

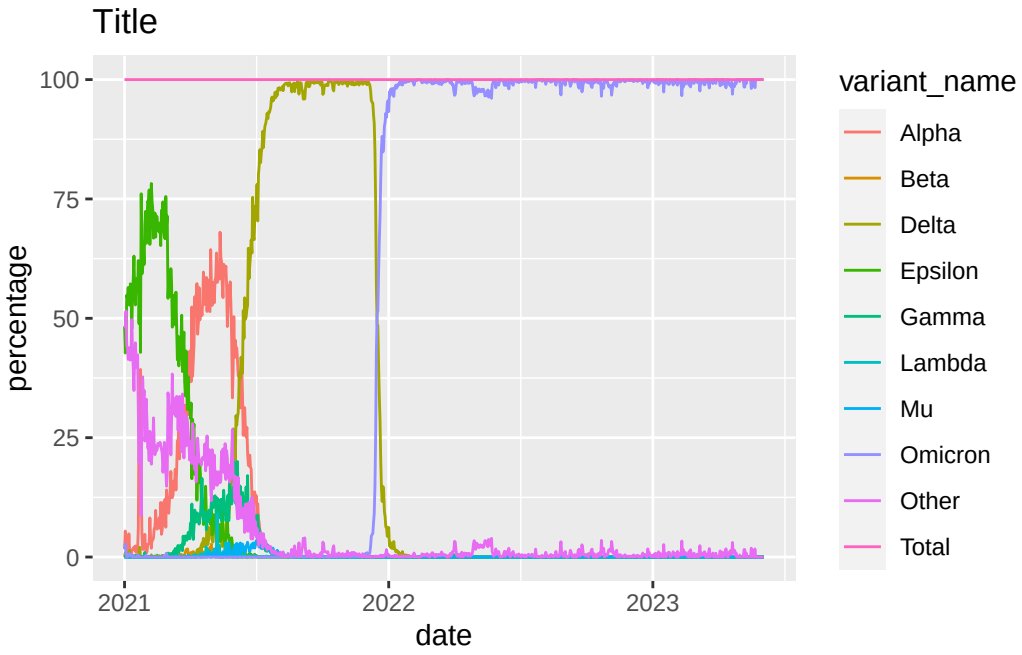
```
        date       area area_type variant_name specimens percentage
1 2021-01-01 California     State        Alpha         1       1.67
2 2021-01-01 California     State        Delta         0       0.00
3 2021-01-01 California     State        Gamma         0       0.00
4 2021-01-01 California     State      Omicron         1       1.67
5 2021-01-01 California     State         Beta         0       0.00
6 2021-01-01 California     State       Lambda         0       0.00
  specimens_7d_avg percentage_7d_avg
1               NA                NA
2               NA                NA
3               NA                NA
4               NA                NA
5               NA                NA
6               NA                NA
```

```
tail(data_3)
```

```
          date       area area_type variant_name specimens percentage
3883 2022-05-01 California     State         Beta         0        0.0
3884 2022-05-01 California     State        Gamma         0        0.0
3885 2022-05-01 California     State      Epsilon         0        0.0
3886 2022-05-01 California     State        Alpha         0        0.0
3887 2022-05-01 California     State      Omicron       492       98.4
3888 2022-05-01 California     State           Mu         0        0.0
     specimens_7d_avg percentage_7d_avg
3883           0.0000           0.00000
3884           0.0000           0.00000
3885           0.0000           0.00000
3886           0.0000           0.00000
3887         649.2857          98.37662
3888           0.0000           0.00000
```

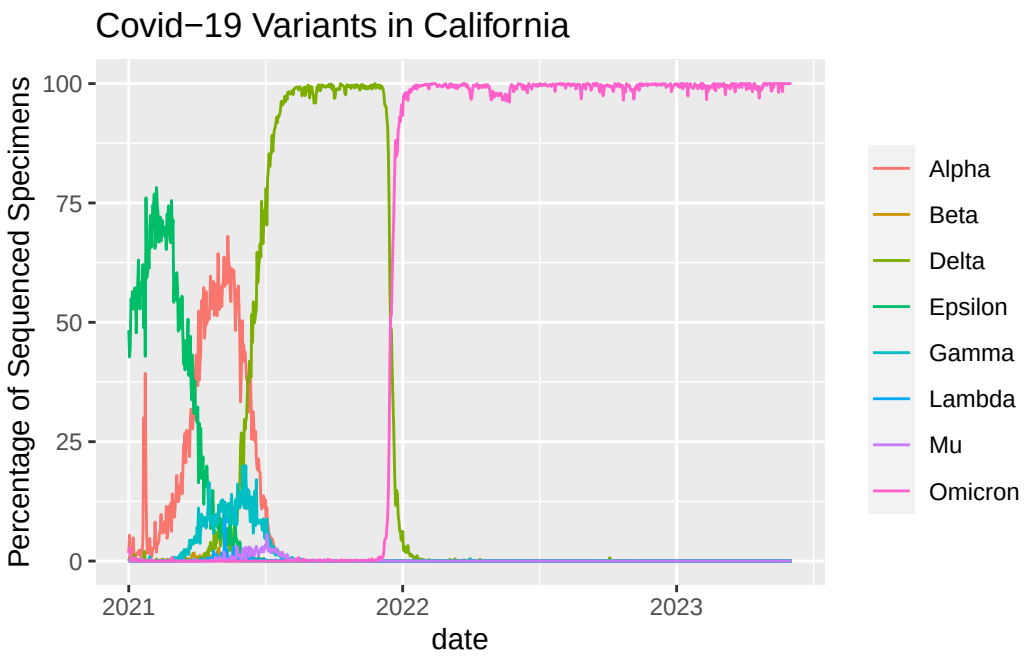
Now we can plot. Below is what the original data looked like plotted. Here, we still have the Other and Total. The data is color coordinated based on variant name information. The percentage of the variant (y-axis) is plotted across time (x-axis). This is also a generally basic plot that needs some polishing.

```
plot <- ggplot(data) + aes(date, percentage, color = variant_name) +
    geom_line() + labs(title = "Title")

plot
```



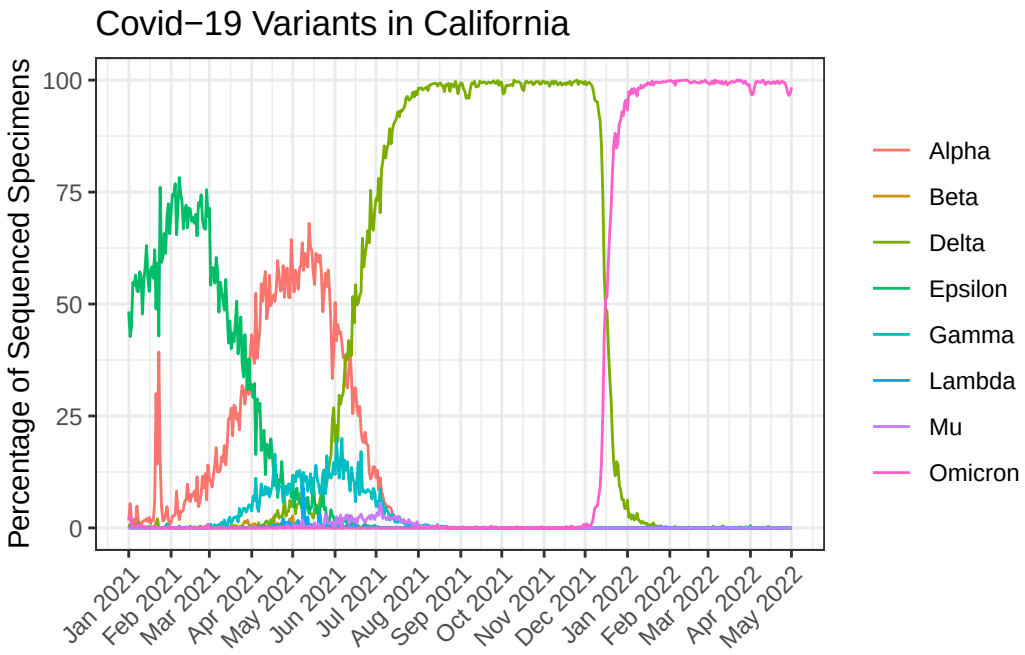

This second plot is of data_2 which has gotten rid of Other and Total. I also added a title, a y-axis label, and got rid of the color legend label.

```
plot <- ggplot(data_2) + aes(date, percentage, color = variant_name) +
    geom_line() + labs(title = "Covid-19 Variants in California",
    y = "Percentage of Sequenced Specimens", color = NULL)

plot
```




This third and final graph plots our data_3 which is just the variants we want to plot (from data_2) as well as cut off at a certain time (until May 2022). The graph is further polished to look like the example by having the bw theme. The x-axis has the month and year for each tick mark, labelled 1 month apart, with the labels set at a 45 degree angle and adjusted appropriately.

```
plot <- ggplot(data_3) + aes(date, percentage, color = variant_name) +
    geom_line() + labs(title = "Covid-19 Variants in California",
    y = "Percentage of Sequenced Specimens", color = NULL) +
    theme_bw()

plot + theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    scale_x_date(name = NULL, date_breaks = "1 month", date_labels = "%b %Y")
```


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
Alpha
Beta
Delta
Epsilon
Gamma
Lambda
Mu
Omicron