First Year Exam

Sneha Patel A12125893

Obtaining and opening the data

First, go to https://data.chhs.ca.gov/dataset/covid-19-variant-data and download the .csv file titled "Covid-19 Variant Data". You can move the file to the same directory as your Quarto document for ease of access. Next, we will read the file, and check the first few lines of its contents.

```
c19 <- read.csv("covid19_variants.csv")

head(c19)
```

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

Visualizing the data

We want to graph the data to show the percentage of each variant in all specimens sequenced by month. For this, we will need `ggplot2, lubridate, and dplyr`

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

```
Loading required package: timechange
```

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

```
The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

```r
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
```

The `date` column of our data is type . We want to change it to using a `lubridate` function.

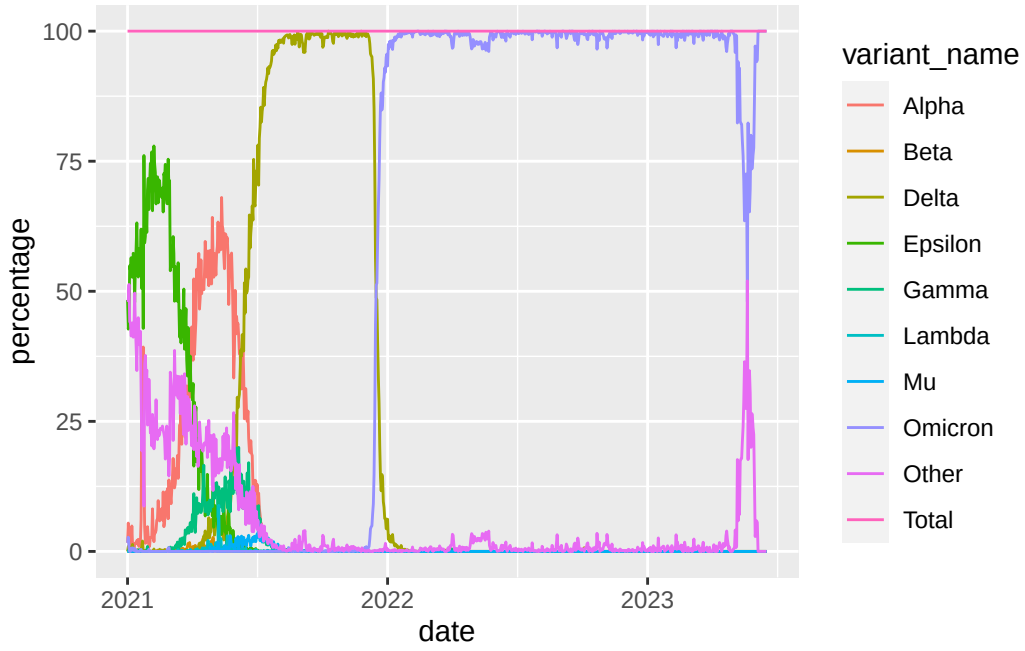
```r
c19$date <- ymd(c19$date)
head(c19)
```

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

Let's plot what we have so far!

```
plot1 <- ggplot(data = c19) + aes(x = date, y = percentage, col=variant_name) + geom_line(
plot1
```



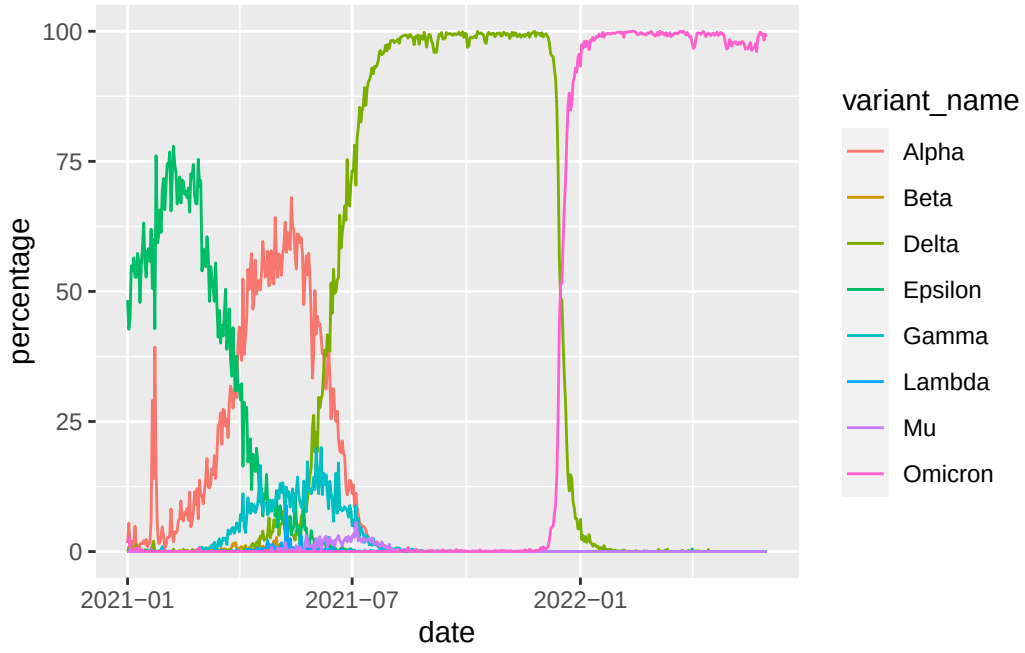
We want to remove data that shows the total number, and other variants. We also only want our data up until May 2022. We will do that with a `dplyr` function.

```
c19_variants <- c19 %>%
  filter((variant_name != "Other") & (variant_name != "Total"))

c19_variants <- c19_variants %>%
  filter(date < "2022-06-01")
```

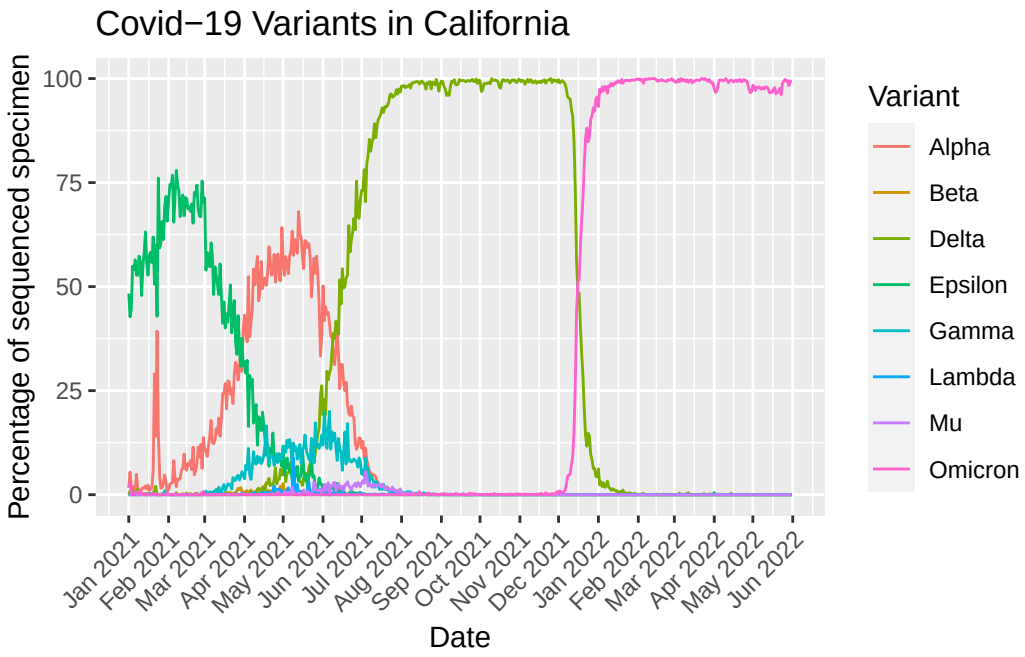
Let's plot what we have so far!

```
p <- ggplot(data = c19_variants) + aes(x = date, y = percentage, col=variant_name) + geom_
p
```



This looks great, but lets change the title, axis labels, legend label, and the ticks on the x-axis to show us each month.

```
p <- p + labs(title = "Covid-19 Variants in California", y = "Percentage of sequenced spec
p
```



WOW! Beautiful :)