Covid 19 Variant Data

Hetian Su, A14553328

Import Covid 19 Dataset

```
# Read in the csv file and inspect the overall info and structure of the file
covid_file <- "~/BGGN213/BGGN213_github/covid19_variants.csv"
covid_data = read.csv(covid_file)
head(covid_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
```

Generate Figre

Load relevant packages and preprocess the dataset

```
# generate plot with ggplot2
#install.packages('ggplot2')
library(ggplot2)
```

```
# process the dataset with dplyr
# install.packages('dplyr')
library(dplyr)

# use dplyr filter function to remove record of "total" and 'Others"
strain_data <- covid_data %>% filter(!variant_name %in% c('Total', 'Other'))
```

Make the figure utilizing ggplot2

```
# set the r system time display to english
Sys.setlocale('LC_TIME', 'English')
```

```
[1] "English_United States.1252"
```

```
# make the plot using ggplot
ggplot(strain_data)+

  # convert the date to r date format and map to x
  # map percentage data to y
  # group and color the lines by strain
  aes(x=as.Date(date),
      y=percentage,
      group=variant_name, color=variant_name)+

  # use the line plot format
  geom_line()+

  # create an annotation for showing data source
  # specify x coordinate in date format
  # specify y coordinate
  # reduce the font size
  annotate('text',
           x=as.Date('12/01/2021', format='%m/%d/%Y'),
           y=-35,
           label='Data Source:<https://www.cdph.ca.gov/>',
           size=3)+

  # let the plot focus on y range from 0 to 100
  # turn off clip to show the annotation
  coord_cartesian(ylim=c(0,100),
                  clip = 'off')+
```

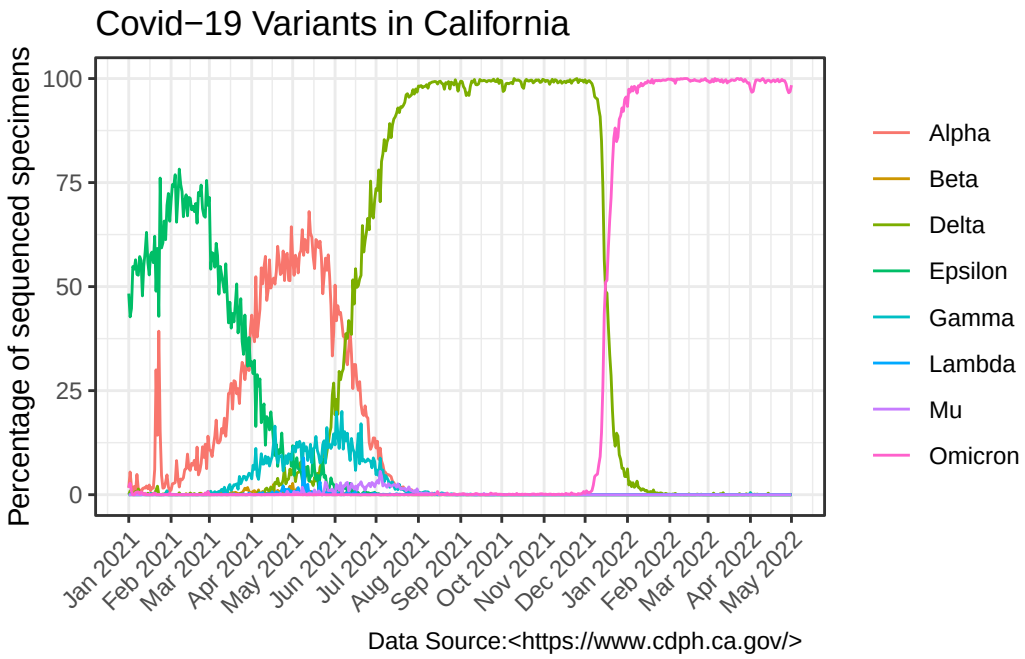
```
# specify y label
ylab('Percentage of sequenced specimens')+

# remove x label
# show xticklabel every 1 month
# specify xticklabel format
# specify x range in data format
scale_x_date(name = '',
             date_breaks = '1 month',
             date_labels = ('%b %Y'),
             limits = as.Date(c('01/01/2021','05/01/2022'),format='%m/%d/%Y'))+

# specify title of the figure
labs(title='Covid-19 Variants in California', ylab)+

# set theme to black and white
theme_bw()+

# rotate and reposition xticklabels
# remove the legend label
theme(axis.text.x = (element_text(angle = 45, vjust = 1, hjust = 1)),
      legend.title = element_blank())
```


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
Data Source:<https://www.cdph.ca.gov/>