

Q15 - First Year Exam

Meg Robinson PID: A59010583

2022-07-17

Load the data and packages

```
df <- read.csv("covid19_variants.csv")
library(ggplot2)
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
```

```
library(lubridate)
```

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

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

Look at data

```
head(df)
```

```
##      date      area area_type variant_name specimens percentage
## 1 1/1/2021 California   State      Alpha         1         1.69
## 2 1/1/2021 California   State      Beta          0          0.00
## 3 1/1/2021 California   State      Mu           0          0.00
## 4 1/1/2021 California   State      Gamma         0          0.00
## 5 1/1/2021 California   State      Total        59        100.00
## 6 1/1/2021 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
```

Remove 'Total' and 'Other' variants since these aren't in the plot given to us

```
df <- df[!df$variant_name %in% c("Other", "Total"), ]
head(df)
```

```
##      date      area area_type variant_name specimens percentage
## 1 1/1/2021 California      State      Alpha         1         1.69
## 2 1/1/2021 California      State      Beta         0         0.00
## 3 1/1/2021 California      State      Mu          0         0.00
## 4 1/1/2021 California      State      Gamma        0         0.00
## 6 1/1/2021 California      State      Omicron        1         1.69
## 7 1/1/2021 California      State      Epsilon       28        47.46
##   specimens_7d_avg percentage_7d_avg
## 1                NA                NA
## 2                NA                NA
## 3                NA                NA
## 4                NA                NA
## 6                NA                NA
## 7                NA                NA
```

Change the date to a 'date' variable rather than a 'character' variable using the 'lubridate' package

```
# observe column variable in wrong format
class(df$date)
```

```
## [1] "character"
```

```
df$date[1]
```

```
## [1] "1/1/2021"
```

```
# fix above, make new column and add to df
# note some will fail to parse as csv doesn't have dates for all rows

df$new_date <- mdy(df$date)
head(df)
```

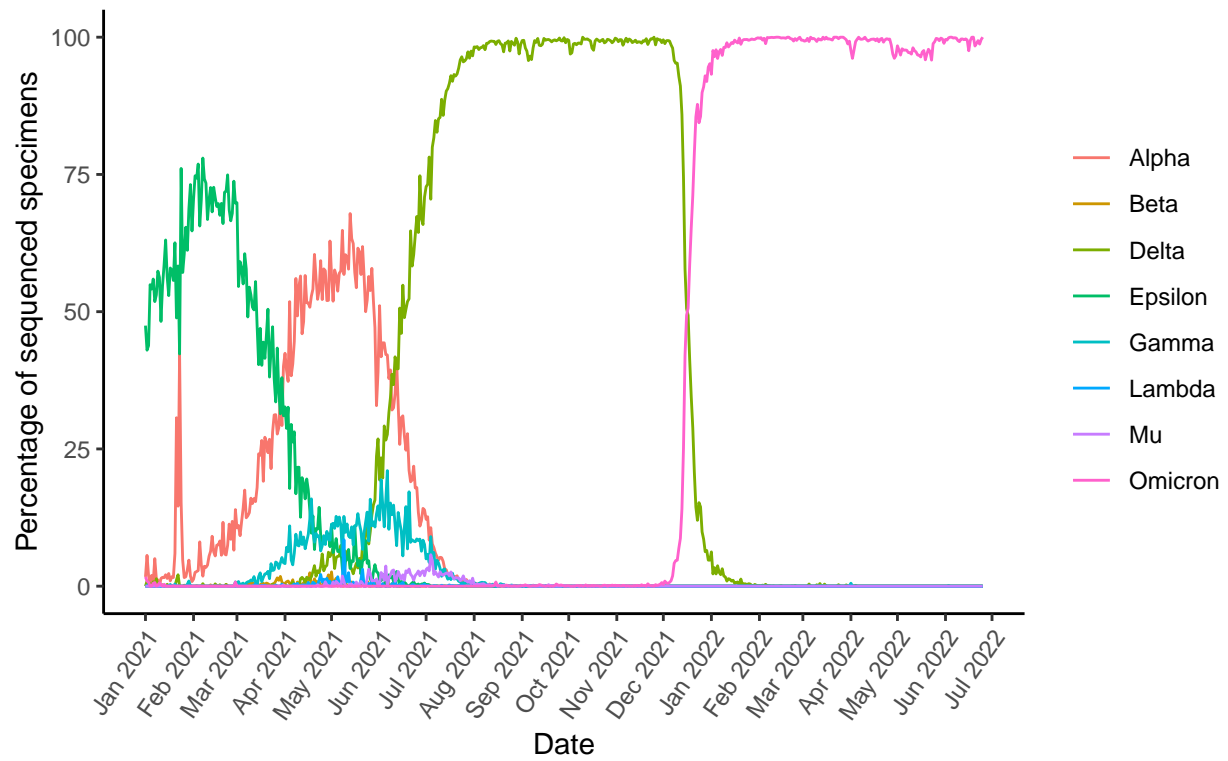
```
##      date      area area_type variant_name specimens percentage
## 1 1/1/2021 California      State      Alpha         1         1.69
## 2 1/1/2021 California      State      Beta         0         0.00
## 3 1/1/2021 California      State      Mu          0         0.00
## 4 1/1/2021 California      State      Gamma        0         0.00
## 6 1/1/2021 California      State      Omicron        1         1.69
## 7 1/1/2021 California      State      Epsilon       28        47.46
##   specimens_7d_avg percentage_7d_avg   new_date
```

```
## 1          NA          NA 2021-01-01
## 2          NA          NA 2021-01-01
## 3          NA          NA 2021-01-01
## 4          NA          NA 2021-01-01
## 6          NA          NA 2021-01-01
## 7          NA          NA 2021-01-01
```

Plot:

```
ggplot(df, aes(new_date, percentage)) +
  # group by variant (hence excluding total and other)
  geom_line(aes(color = variant_name)) +
  # set theme
  theme_classic() +
  # now, format axes
  scale_x_date(date_labels = "%b %Y", date_breaks = "1 month") +
  # add axes titles
  ylab("Percentage of sequenced specimens") +
  xlab("Date") +
  # make x axis readable (rotate), remove legend title
  theme(axis.text.x = element_text(angle = 55,
                                     hjust = 1),
        legend.position = "right",
        legend.title = element_blank()) +
  # title plot
  ggtitle("Covid-19 Variants in California") +
  # add caption to include website where data was found
  labs(caption = "Data Source: <https://data.chhs.ca.gov/dataset/covid-19-variant-data>")
```

Covid-19 Variants in California



Data Source: <<https://data.chhs.ca.gov/dataset/covid-19-variant-data>>