Sara_Herrera_Q15

Sara Herrera (PID:A59011948)

7/13/2022

Name: Sara Herrera de la Mata

PID: A59011948

Covid-19 Variants in California

Libraries

```
# Calling the libraries that will be used to generate the plot
library(ggplot2) # For plots
library(lubridate) # For date-time data
```

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

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

```
library(dplyr) # For data manipulation
```

```
## Warning: package 'dplyr' was built under R version 4.1.2
```

```
##
## 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
```

Read file

```r
# To read the .csv file downloaded from the California Health and Human Services (CHHS) open data site
data <- read.csv("covid19_variants.csv")
```

Filter data

I'll use 'dplyr' to filter out unnecessary rows (containing the values 'Other' and 'Total' in the 'variant_name' column) and change the format of the date values in the 'date' column to date instead of character.

```r
# To remove the rows that contain "variant_name" = "Other", "Total" (as they're not included in the plo
# To change the date values from "character" to "date" format
clean.data <- data %>% filter(!variant_name %in% c('Other', 'Total')) %>% mutate(date = as.Date(date))
```

Assigning variables

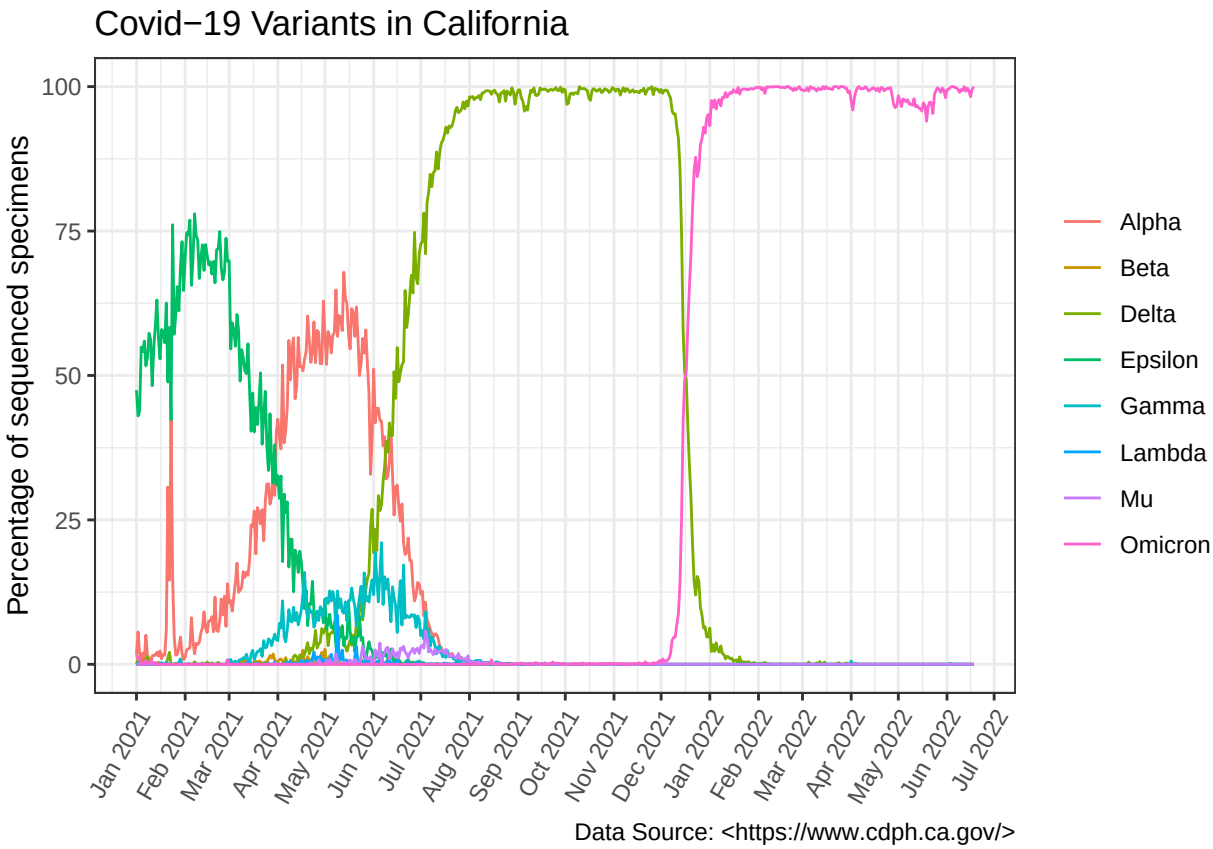
We use this to simplify things when writing code, substituting longer names to simpler and shorter words.

```r
# Assigning variables
Percentage <- clean.data$percentage
Date <- clean.data$date
Variant <- clean.data$variant_name
```

Plotting the data

I'll use 'ggplot2' and I'll first specify the data frame and the variables that we're interested in, coloring each Covid-19 variant with a different color (each variant will be a line in the graph). Then, I'll select the type of graph I would like to represent the data in, in this case I'll use a line chart. I'll modify the theme of the plot so that it resembles the example and I'll modify the dates with 'lubricate'. In addition, I'll add labels to my plot so that it's easily understandable and a caption at the bottom.

```r
# Making the plot
plot <- ggplot(data = clean.data, aes(x=Date, y=Percentage, color=Variant)) +
  # Select the data frame, x and y axis variables, and the variable by which the data will be colored b
  geom_line(aes(group=Variant)) +
  # To make a line chart, where every line will be a category of the "Variant" column, meaning a differ
  theme_bw() +
  # Specifies the theme used for the plot: in this case a white background and thin grey grid lines
  scale_x_date(date_breaks = "1 month", date_labels =  "%b %Y") +
  # Specifies how the dates in the x-axis will be represented, separated by 1 month gaps, and with the
  theme(axis.text.x=element_text(angle=60, hjust=1)) +
  # Specifies that the label for the dates in the x-axis is angled
  labs(title = "Covid-19 Variants in California", x = NULL, color = NULL, caption = "Data Source: <https
  # To label the title of the plot, remove the x-axis label to only keep the dates, remove the label of
  ylab("Percentage of sequenced specimens") # To add a y-axis label
plot
```


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://www.cdph.ca.gov/>
Alpha
Beta
Delta
Epsilon
Gamma
Lambda
Mu
Omicron