Q10_Bioinformatics_Exam

Toheeb_Balogun; PID:A59018916

Analysis of COVID-19 Variant Data obtain from from the California Health and Human Services (CHHS) open data site

```
# Load the packages
library(ggplot2)
library(lubridate)
library(dplyr)
```

Import the data and perform exploratory data analysis

```
# read the csv file containing the most most recently dated COVID-19 Variant Data
# from the California Health and Human Services (CHHS) and store it in a data frame
# called covid_data. Also, print out the first six rows for exploratory analysis
#setwd = set working directory
setwd("C:/Users/ITSloaner/Desktop/Bioinformatics_exam")
covid_data = read.csv('covid19_variants.csv')
head(covid_data)
```

```
        date       area area_type variant_name specimens percentage
1 2021-01-01 California     State      Epsilon        29      48.33
2 2021-01-01 California     State        Other        29      48.33
3 2021-01-01 California     State        Gamma         0       0.00
4 2021-01-01 California     State        Delta         0       0.00
5 2021-01-01 California     State         Beta         0       0.00
6 2021-01-01 California     State        Alpha         1       1.67
  specimens_7d_avg percentage_7d_avg
1               NA                NA
2               NA                NA
```

```
3                NA                 NA
4                NA                 NA
5                NA                 NA
6                NA                 NA
```

Subset only columns of interest

```r
#Subset the covid dataframe by selecting only the required columns
#(date, variant_name and percentage) and store it in a new dataframe called covid_data_1
covid_data_1 = covid_data[, c(1, 4, 6)]
head(covid_data_1) # Print covid_data_1 which is the new dataframe
```

```
        date variant_name percentage
1 2021-01-01      Epsilon      48.33
2 2021-01-01        Other      48.33
3 2021-01-01        Gamma       0.00
4 2021-01-01        Delta       0.00
5 2021-01-01         Beta       0.00
6 2021-01-01        Alpha       1.67
```

Exclude 'Total' and 'Other' column

```r
#exclude other and total in the variant_name column and assign the
#result to the same dataframe
covid_data_1 <- covid_data_1 %>%
  filter(variant_name != "Other" & variant_name != "Total")
head(covid_data_1)
```

```
        date variant_name percentage
1 2021-01-01      Epsilon      48.33
2 2021-01-01        Gamma       0.00
3 2021-01-01        Delta       0.00
4 2021-01-01         Beta       0.00
5 2021-01-01        Alpha       1.67
6 2021-01-01      Omicron       1.67
```

Use of date format

```
#convert the date column in covid_data_1 to a date object using the ymd function
covid_data_1$date <- ymd(covid_data_1$date)
#Print the first-six role of the modified dataframe
head(covid_data_1)
```

```
        date variant_name percentage
1 2021-01-01      Epsilon      48.33
2 2021-01-01        Gamma       0.00
3 2021-01-01        Delta       0.00
4 2021-01-01         Beta       0.00
5 2021-01-01        Alpha       1.67
6 2021-01-01      Omicron       1.67
```

Dates filtering to include only the dates of interest

```
# Filter the date column to only include Jan 2021 to April 2022
# Convert the date column to a date object using as.Date function
# Use the between function to check if the date is within the range
covid_data_2 = covid_data_1
covid_data_2 <- covid_data_2 %>%
  filter(between(as.Date(date), as.Date("2021-01-01"), as.Date("2022-04-30")))
head(covid_data_2)
```

```
        date variant_name percentage
1 2021-01-01      Epsilon      48.33
2 2021-01-01        Gamma       0.00
3 2021-01-01        Delta       0.00
4 2021-01-01         Beta       0.00
5 2021-01-01        Alpha       1.67
6 2021-01-01      Omicron       1.67
```

Graph plotting

```
# This code chunk plot the  percentage of sequenced specimen over time
#colored by variant_name.
ggplot(covid_data_2, aes(x = date, y = percentage, color = variant_name)) +
```

```r
geom_line() +
# adjust x-axis labels and breaks
scale_x_date(date_labels = "%b. %Y", date_breaks = "1 month") +
labs(x = "Data source: <https://data.chhs.ca.gov/>",
     y = "Percentage of sequenced specimen",
     title = "Covid-19 Variants in California", color=NULL) +
theme_minimal() +
# rotate and align x-axis labels
theme(axis.text.x = element_text(angle = 55, hjust = 1),
      axis.line = element_line(), # add axis lines
      # add panel border
      panel.border = element_rect(color = "black", fill = NA),
      axis.ticks = element_line(),
       axis.title.x = element_text(hjust = 0.8, vjust=-2))
```