```
.file/github  ==> assign.M.1.assignment.1.covid.data
.institution  ==> University of Tennessee
     .course  ==> COSC.526 Intro. to Data Mining
```

assign.M1.Assignment.1.covid.data

0. Problem summary

For this problem, you will be working with COVID-19 sequence processing data from Kaggle. The dataset contains data about the processing of COVID-19 sequences by different countries over time. It comes as a Comma-Separated Value (CSV) file. It contains the following 6 columns:

- **location:** the country for which the information is provided
- **date:** the date of the data entry
- **variant:** the COVID-19 variant for the data entry
- **num_sequences:** the number of sequences processed (for the country, variant, and date)
- **num_sequences_total:** the total number of sequences available (for the country, variant, and date)
- **perc_sequences:** the percentage of the available number of sequences that were processed (Note: this value is out of 100)
- **note:** each row (or data entry) in the dataset represents the processing of one variant by one country on one day.

0.1 Codebook and data files

File Name	Purpose\Description
https://github.com/cosc-526/cosc.526.home.page/blob/main/code_notebook_cosc_526.ipynb <save your own copy!>	Course Codebook in Jupyter Notebook name = code.notebook.cosc.526.ipynb
d.M.1.10.assignment.covid.data.variants.csv	Course github of source data
https://www.kaggle.com/yamqwe/omicron-covid19-variant-daily-cases?select=covid-variants.csv	Kaggle data homepage

1. Problem 1

The United States experienced 3 main variants of COVID-19, including

1. Alpha
2. Delta
3. Omicron

However, the World Health Organization documented other variants.

A. Determine the other variants.
B. Sort the variant names alphanumerically.
C. Exclude catch-all categories called others and non_who

2. Problem 2

Which variant of COVID-19 has the most sequences processed?

3. Problem 3

Which country did the best at processing sequences across all variants including the “catch-all” categories?

- hint: output is one country

4. Problem 4

Part A

Which country did the best at processing sequences across the Alpha, Delta, and Omicron variants only?

- hint: output is one country

Part B

Determine the ranking of the United States at processing sequences across the Alpha, Delta, and Omicron variants only?

- hint: output is one country

5. Problem 5c

Determine each country’s total number of processed sequences for the Omicron variant on December 27, 2021. Sort the output from highest number of processed sequences to the smallest number of processed sequences. Each element in the output should include both the name of the country and the number of processed sequences.

6. Problem 6

Determine the percentage of processed sequences for the Alpha, Delta, and Omicron variants in the United States.

7. Additional resources

- https://github.com/cosc-526/cosc.526.home.page
- Jupyter Community Forum