

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.

Objectives:

- 1. Import and manipulate a .csv file
- 2. Assess your Python Programming Skills
- => Other assignments are more challenging. Use this to assess your skills.
- => Attempt to solve the problems without searching for online assistance.
- => Prepare questions for class discussion to help source additional tools.

Codebook and data files

File Name	Purpose\Description
<pre>https://github.com/cosc-526/cosc.526.home. page/blob/main/code notebook cosc 526.ipyn</pre>	Course Codebook in Jupyter Notebook
b <save copy!="" own="" your=""></save>	<pre>name = code.notebook.cosc.526.ipynb</pre>
d.M.1.10.assignment.covid.data.variants.cs ⊻	Course github of source data
<pre>https://www.kaggle.com/yamqwe/omicron-covid19 -variant-daily-cases?select=covid-variants.cs </pre>	i) Kaggle data homepageii) grab an api key from this page if using that method to import data

note.1: the codebook is formatted differently and below highlights expected outcomes.

note.2: the instructions below are an overview and more details are in the notebook.

note.3: perform your work in the notebook, and export and submit as a .pdf file.

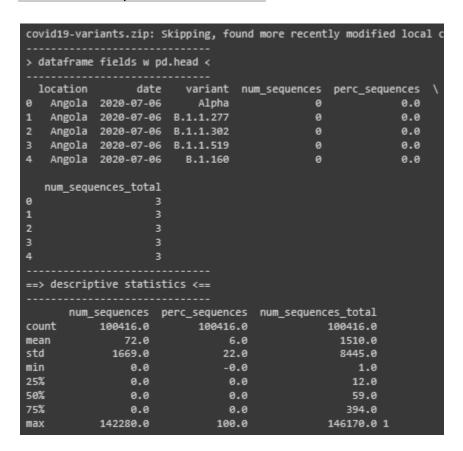


0.Task.0 - Import, inspect, and view descriptive statistics

Import data and view descriptive statistics with the pandas library.

• grab data from **Github** URL, .csv. or kaggle api

Task.0 - Expected outcome:



1.Task.1 - Find uncommon variants

- A. The 3 main variants of COVID-19 that we've experienced in the US are:
 - * `Alpha`
 - * `Delta`
 - * `Omicron`

B. Assignment Tasks

- 1. What other variants are recognized by the World Health Organization (WHO)
- in this dataset?
- 2. Sort the variant names alphanumerically and store in a list.
- 3. Exclude in output `on_who` and `others` from `variant`.

Task.1 - Expected outcome:



2.Task.2 - Description => Most sequences?

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

Task.2 - Expected outcome:

3.Task.3 - Description => Best sequence processing?

1. Which country processed sequences the best of all variants including the "catch-all" categories?

Task.3 - Expected outcome:

4. Task. 4 - Find best country and ranking

4A - Description => Find Best Country at Processing Specific Sequences

- 1. Which country did the best at processing sequences across the Alpha, Delta, and Omicron variants?
- 2. Output is the name of a single country.

Task.4a - Expected outcome:

4B - Description => Find the US Ranking when Processing Specific Sequences

- 1. Determine the ranking of the US at processing sequences across the Alpha, Delta, and Omicron variants.
- Store the rankings as an integer.
- The best country has a ranking of 1, but indexing in Python starts at 0.

Task.4b - Expected outcome:

5.Task.5 Description => Number of Processed Sequences Per Country

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



Task.5 - Expected outcome:

6.Task.6 - Description => Store outcomes in a dictionary

- 1. Determine the percentage of processed sequences for the Alpha, Delta, and Omicron variants in the US.
- 2. Store the result as a dictionary where keys are variant names and values are percentages.

Task.6 - Expected outcome:

7. Task.7 - Description => Challenges ?

Report any assignment challenges

Additional resources

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

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

