```
.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

A.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.

B.Assignment Objectives:

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

C.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 master copy!**	Course Codebook in Jupyter Notebook name = code.notebook.cosc.526.ipynb
data.M.1.assignment.1.covid.data.csv	Source data on github
https://www.kaggle.com/yamqwe/omicron-covid19-variant-daily-cases?select=covid-variants.csv	i) Kaggle data homepage ii) 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.

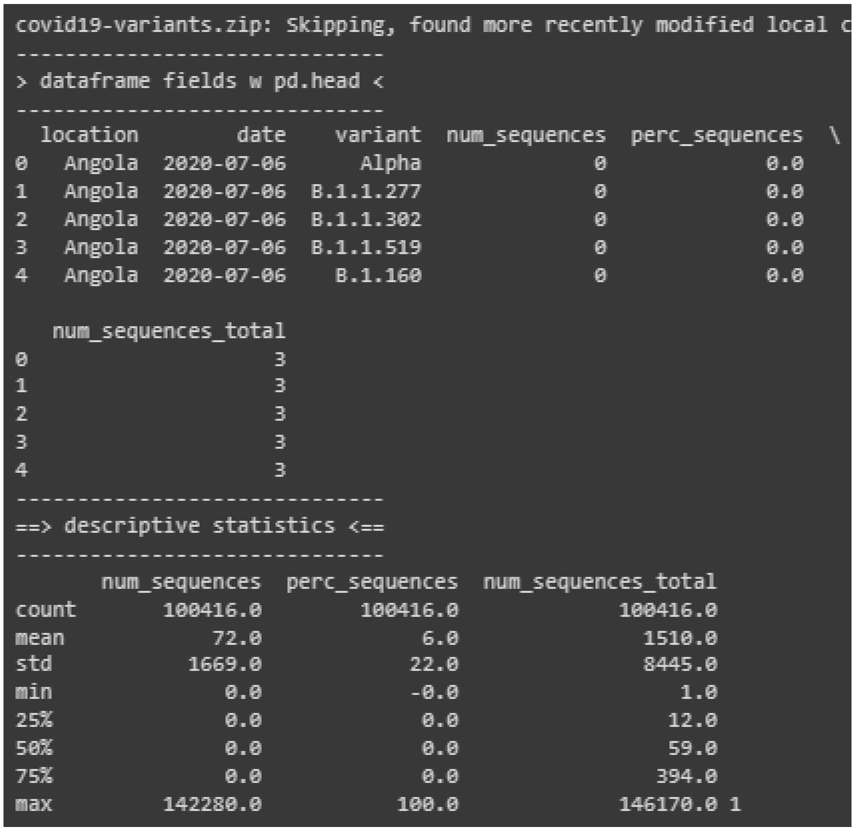
note.3: perform your work in the notebook, and export and submit as a .pdf file.

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

1. Import data and view df.head() and descriptive statistics with pandas.
2. read the data from the class repository or use a kaggle api key
   - url="https://raw.githubusercontent.com/cosc-526/cosc.526.home.page/main/d.data.M.1.assignment.1.covid.data.csv"
   - api address = > !kaggle datasets download -d gpreda/covid19-variants

Note: This assignment does not require you to have a Kaggle API key, but it is important to be aware of this standard method of data retrieval. You may be required to use it later in the course, so it is recommended that you keep a record of exercise and assignment solutions provided by your professor."

Task.0 - Expected outcome:



```
covid19-variants.zip: Skipping, found more recently modified local c
------------------------------
> dataframe fields w pd.head <
------------------------------
  location        date    variant  num_sequences  perc_sequences  \
0   Angola  2020-07-06      Alpha              0             0.0
1   Angola  2020-07-06  B.1.1.277              0             0.0
2   Angola  2020-07-06  B.1.1.302              0             0.0
3   Angola  2020-07-06  B.1.1.519              0             0.0
4   Angola  2020-07-06    B.1.160              0             0.0

   num_sequences_total
0                    3
1                    3
2                    3
3                    3
4                    3
------------------------------
==> descriptive statistics <==
------------------------------
       num_sequences  perc_sequences  num_sequences_total
count       100416.0        100416.0             100416.0
mean            72.0             6.0               1510.0
std           1669.0            22.0               8445.0
min              0.0            -0.0                  1.0
25%              0.0             0.0                 12.0
50%              0.0             0.0                 59.0
75%              0.0             0.0                394.0
max         142280.0           100.0             146170.0 1
```

Problem.1 - Description => Find uncommon variants

The 3 main variants of COVID-19 that we've experienced in the US are:

- `Alpha`
- `Delta`
- `Omicron`

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:

Problem.2 - Description => Most variant sequences?

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

Task.2 - Expected outcome:

Problem.4 - Description => Best sequence processing?

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

Task.3 - Expected outcome:

Problem.4 - Description => 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.
2. Store the rankings as an integer.
3. The best country has a ranking of 1, but indexing in Python starts at 0.

Task.4b - Expected outcome:

Problem.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:

Problem.5 - 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:

Problem.7 - Description => Challenges ?

1. Report any assignment challenges

Additional resources

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

10. Additional resources

Resources [H3]

- Anaconda for windows
- Anaconda for mac-os
- Anaconda for Linux
- Install scientific packages.
- Reading: Software installation
- Reading: Using GitHub - class file exchange

Additional Resources

- Anaconda installation documentation.
- Jupyter Notebook documentation (including get started guides).
- Jupyter Discourse Forum.
  - Search here for tips, tricks, and solutions.
- Python Package Index (pypi)
- Spyder IDE - an alternative programming environment to Notebooks called an *integrated development environment* (IDE). Spyder is a sister environment to Notebooks providing an interactive console to view data, variables, and outcomes. It is not covered in the course but works alongside Jupyter Notebooks.

- GitHub - a place for storing files, searching for ideas, and framing interactive Jupyter Notebooks environments.
    - All data mining and machine learning scientists should have a page!
-