WASDI Data Scientist Test

This document presents the test to be executed for the Data Scientist Position at WASDI Sarl.

Test Rules

A list of exercises can be found in the next section: the candidate is invited to implement at least one exercise. The more will be implemented the more the test will be evaluated.

The list of exercises is NOT ordered by priority.

The exercises must be implemented in Python 3.x.

The candidate can use any development tool and use her/his own internet connection.

The candidate must answer to this e-mail in 2 hours, attaching a zip file with the implemented Source Code.

Exercises List

Credits: some exercises are taken from NYU - https://github.com/nyu-cds

NOx Ground truth

The following data structure represents the reading of NOX, a pollutant, from various ground stations dislocated in the sample territory. The values report the read value and the sensor tolerance for each station with the following structure:

[[[STATION NAME] , [NOX VALUE] , [SENSOR TOLERANCE]], [...]]

The sensor tolerance is reported as integer number and represents the +/- percentage error that the reading can present. (E.g. given reading of 100 an error value of 5% lead to real possible values between 95 and 105)

```
data = [['Albany'] , [75] , [0]  ] , [['Allegany'] , ['83'] , ['2']  ] , [['Bronx'] , ['82'] ,
['3']  ] , [['Broome'] , ['67'] , ['5']  ] , [['Cattaraugus'] , ['95'] , ['5']  ] , [['Cayuga']
, ['80'] , ['4']  ] , [['Chautauqua'] , ['78'] , ['4']  ] , [['Chemung'] , ['91'] , ['3']  ] ,
[['Chenango'] , ['92'] , ['0']  ] , [['Clinton'] , ['58'] , ['1']  ] , [['Columbia'] , ['52']
, ['1'] ] , [['Cortland'] , ['59'] , ['3']  ] , [['Delaware'] , ['70'] , ['0'] ] , [['Dutchess']
, ['98'] , ['0']  ] , [['Erie'] , ['54'] , ['5']  ] , [['Essex'] , ['83'] , ['3']  ] ,
[['Franklin'] , ['83'] , ['5']  ] , [['Fulton'] , ['84'] , ['2']  ] , [['Genesee'] , ['84'] ,
['5']  ] , [['Greene'] , ['73'] , ['1']  ] , [['Hamilton'] , ['55'] , ['1']  ] , [['Herkimer']
, ['88'] , ['0']  ] , [['Jefferson'] , ['67'] , ['1']  ] , [['Kings'] , ['100'] , ['4']  ] ,
[['Lewis'] , ['65'] , ['2']  ] , [['Livingston'] , ['61'] , ['1']  ] , [['Madison'] , ['85'] ,
['0']  ] , [['Monroe'] , ['73'] , ['4']  ] , [['Montgomery'] , ['59'] , ['2']  ] , [['Nassau']
, ['88'] , ['0']  ] , [['NewYork'] , ['78'] , ['1']  ] , [['Niagara'] , ['73'] , ['4']  ] ,
[['Oneida'] , ['93'] , ['1']  ] , [['Onondaga'] , ['66'] , ['5']  ] , [['Ontario'] , ['60'] ,
['1']  ] , [['Orange'] , ['70'] , ['4']  ] , [['Orleans'] , ['75'] , ['5']  ] , [['Oswego'] ,
['61'] , ['2']  ] , [['Otsego'] , ['61'] , ['2']  ] , [['Putnam'] , ['63'] , ['2']  ] ,
[['Queens'] , ['99'] , ['5']  ] , [['Rensselaer'] , ['67'] , ['1']  ] , [['Richmond'] , ['67']
, ['3']  ] , [['Rockland'] , ['91'] , ['3']  ] , [['St.Lawrence'] , ['69'] , ['2']  ] ,
[['Saratoga'] , ['61'] , ['5']  ] , [['Schenectady'] , ['86'] , ['4']  ] , [['Schoharie'] ,
['53'] , ['3']  ] , [['Schuyler'] , ['73'] , ['5']  ] , [['Seneca'] , ['53'] , ['2']  ] ,
[['Steuben'] , ['77'] , ['1']  ] , [['Suffolk'] , ['71'] , ['1']  ] , [['Sullivan'] , ['78'] ,
['0']  ] , [['Tioga'] , ['89'] , ['5']  ] , [['Tompkins'] , ['64'] , ['1']  ] , [['Ulster'] ,
['57'] , ['2']  ] , [['Warren'] , ['91'] , ['5']  ] , [['Washington'] , ['66'] , ['0']  ] ,
[['Wayne'] , ['68'] , ['5']  ] , [['Westchester'] , ['86'] , ['1']  ] , [['Wyoming'] , ['59']
, ['2']  ] , [['Yates'] , ['55'] , ['2']  ]
```

For each of the following questions, write a script to obtain the answer:

- How many stations are there?
- What was the NOx reading from the 5th station?
- What was the NOx reading from the last station?

- What is the total sum of NOx on sites with the station's name beginning with the letter W? (don't just identify these stations by eye, in the real world there could be hundreds or thousands)

- What is the average NOx reading, considering the maximum error possible for each station?

Order Sentinel 3 Products

After a search of Sentinel-3 Products in our area and period of interest, we got these results.

```
data =
['S3A_SL_2_WST____20230701T003542_20230701T021641_20230702T115546_6059_100_287______MAR_O_NT_0
03',
'S3B_SL_2_WST____20230701T013741_20230701T031840_20230702T121228_6059_081_145______MAR_O_NT_00
3',
'S3A_SL_2_WST____20230701T021641_20230701T035740_20230702T134535_6059_100_288______MAR_O_NT_00
3',
'S3B_SL_2_WST____20230701T031840_20230701T045939_20230702T135808_6059_081_146______MAR_O_NT_00
3',
'S3A_SL_2_WST____20230701T035740_20230701T053839_20230702T152705_6059_100_289______MAR_O_NT_00
3',
'S3B_SL_2_WST____20230701T045939_20230701T064039_20230702T135812_6059_081_147______MAR_O_NT_00
3',
'S3A_SL_2_WST____20230701T053839_20230701T071939_20230702T170939_6059_100_290______MAR_O_NT_00
3',
'S3B_SL_2_WST____20230701T064039_20230701T082138_20230702T170001_6059_081_148______MAR_O_NT_00
3',
'S3A_SL_2_WST____20230701T071939_20230701T090038_20230702T184839_6059_100_291______MAR_O_NT_00
3',
'S3B_SL_2_WST____20230701T082138_20230701T100237_20230702T184618_6059_081_149______MAR_O_NT_00
3',
'S3A_SL_2_WST____20230701T090038_20230701T104137_20230702T191020_6059_100_292______MAR_O_NT_00
3',
'S3B_SL_2_FRP____20230701T092150_20230701T092650_20230701T113427_0299_081_150______MAR_O_NR_00
2',
'S3B_SL_2_AOD____20230701T092152_20230701T092651_20230701T113543_0299_081_150______MAR_O_NR_00
3',
'S3B_SL_1_RBT____20230701T092255_20230701T092555_20230702T054816_0180_081_150_2160_PS2_O_NT_00
4',
'S3B_SL_2_LST____20230701T092255_20230701T092555_20230701T113003_0180_081_150_2160_PS2_O_NR_00
4',
'S3B_SL_2_FRP____20230701T092255_20230701T092555_20230702T055613_0180_081_150_2160_PS2_O_NT_00
4',
'S3B_SL_2_WST____20230701T092255_20230701T092555_20230701T113157_0180_081_150_2160_MAR_O_NR_00
3',
'S3B_SL_2_LST____20230701T092255_20230701T092555_20230702T060749_0180_081_150_2160_PS2_O_NT_00
4',
'S3B_SL_1_RBT____20230701T092255_20230701T092555_20230701T112528_0180_081_150_2160_PS2_O_NR_00
4',
'S3A_SL_2_FRP____20230701T100051_20230701T100551_20230701T121412_0299_100_293______MAR_O_NR_00
```

```
2',
'S3A_SL_2_AOD____20230701T100052_20230701T100551_20230701T121522_0299_100_293______MAR_O_NR_00
3',
'S3A_SL_2_FRP____20230701T100155_20230701T100455_20230702T190119_0179_100_293_2160_PS1_O_NT_00
4',
'S3A_SL_1_RBT____20230701T100155_20230701T100455_20230702T185513_0179_100_293_2160_PS1_O_NT_00
4',
'S3A_SL_2_LST____20230701T100155_20230701T100455_20230702T190540_0179_100_293_2160_PS1_O_NT_00
4',
'S3A_SL_2_WST____20230701T100155_20230701T100455_20230701T121337_0179_100_293_2160_MAR_O_NR_00
3',
'S3A_SL_2_LST____20230701T100155_20230701T100455_20230701T121407_0179_100_293_2160_PS1_O_NR_00
4',
'S3A_SL_1_RBT____20230701T100155_20230701T100455_20230701T121102_0179_100_293_2160_PS1_O_NR_00
4',
'S3B_SL_2_WST____20230701T200833_20230701T214932_20230703T064744_6059_081_156______MAR_O_NT_00
3',
'S3B_SL_2_FRP____20230701T204150_20230701T204650_20230701T231301_0299_081_157______MAR_O_NR_00
2',
'S3B_SL_2_AOD____20230701T204152_20230701T204649_20230701T231345_0299_081_157______MAR_O_NR_00
3']
```

The naming convention of Sentinel-3 SLSTR is described here:

https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-3-slstr/naming-convention

Create a script to group the images found by data Type ID and, for each unique dta type, order the images lexicographically by sensing start time in ascending order.

Manipulate Tif

In the received zip there is a GeoTIFFfile called sample_rgb.tif. The image has 3 bands.

From the original image create two new GeoTIFF images:

1. One GeoTIFF Image with only one band that is a combination of bands: (B1-B2)/(B1+B2+B3)
2. One GeoTIFF Image with the standard deviation of the values of the three input bands

Anomalies

Write a Python script to analyze student exam scores stored in a nested dictionary and determine their performance relative to the national average.

The script should take as input:

- An outer dictionary representing exam scores for each student in the class, where the keys are student names and the values are inner dictionaries. The inner dictionaries have the course name as keys and a list of scores as values. Each course may have multiple tests, and the list contains all the scores for those tests.

- Another dictionary containing reference values for the courses across the entire country. The keys are the course names, and the values are dictionaries containing the mean and standard deviation of scores for that course nationwide.

The script should calculate:

For each course taken by the students:

- The average score for the class across all tests in that course.
- The difference between the class average and the national average.

Moreover, show ways of identifying outliers among the students.