Final Assignment in DiP

Data in Python course, 1st edition (20/21)
Ver. 1.00

Your task in this assignment is to create a data analyzing program. The program has to be written in Python and is to be delivered in a form of a Python package installable with pip. You also have to attach to that package the results of profiling the execution of your code (you do not have to optimize the code, just submitting the profiling results and pointing out possible bottlenecks is enough). For the data analysis, you should use at least one of the libraries shown during the lecture (*numpy*, *pandas*,...).
Your Program should be implemented as a library implementing all important functions. Additionally, there should be a simple Python program or a Jupyter notebook using that library. If you decide to use a script then this script should use *argparse* to read the parameters (such as the output file name or path to a directory with data files) and in this case, the results of the analysis should be written to a file (specified as program argument). Your library must not contain any hardcoded paths to any data files. There should also be unit tests provided with your code.

Assignment versions

There are three possible versions of the data analysis task:

- the one described below (the canonical one),
- your own (has to be compatible with the canonical requirements and has to be accepted ahead by the teacher in your group),
- the more advanced one (about busses in Warsaw).

The canonical assignment

Using the data that can be found following the links given below:

- schools,
- inhabitants,

calculate in Python basic statistics like min/average/max number of:

- students per teacher broken down by the type of school,
- students per school broken down by their year of birth,

in each district (polish ‘gmina’) and in total for cities and rural districts.

Your code should also report possible inconsistencies in the data. The identification of individual data files, that should be used for this task, is a part of the assignment.

The advanced assignment:

Using data available on the webpage https://api.um.warszawa.pl/# collect the data about bus positions over a period (one hour). Then perform an analysis of the collected data. Sample questions that should be answered are:

- How many buses exceeded the speed of 50 km/h (the bus position is updated every minute, we can approximate the real speed by assuming that the bus is moving during the minute in a straight line). Were there any locations with a significant percent of busses exceeding that speed limit? The notion of a location should be defined in the solution, for example, it could be a specific place in a city (i.e. a street or bridge), a radius around a given geographic point.
- Analysis of punctuality of buses during the observed period (we can compare actual arrival times at the bus stops against the schedule).

The solution should be implemented in two parts. Firstly, collect data and save them in a file. Secondly, perform analysis on the collected data. That gives the flexibility of replacing one of the parts of your solution with a replacement.

Evaluation rules:

- Explanation of design decisions during the exam.
- Proper division of code into packages and modules.
- Code quality - proper names for functions and variables. Short functions. We suggest trying to use automated tools like pylint or flake to check code. We also suggest reading https://www.python.org/dev/peps/pep-0008/
- Possibility to install code as a package using `pip install ./path/to_package_directory` which should be introduced as PR to the master (or main) branch. [This topic will be presented during our classes in January]
- Code coverage with tests.

If you look for inspiration for your own task you could check this source of data:

- https://api.um.warszawa.pl/#
- https://stat.gov.pl/
- https://dane.gov.pl/pl

We wish you a merry Pythoning!