Final Assignment

Data in Python course, Spring 24/25
Ver. 1.00

Your task in this assignment is to create a program for data analysis. The program must be written in Python and delivered in the 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 optimise the code, just submitting the profiling results and pointing out possible bottlenecks is enough). You should use at least one of the libraries shown during the lecture (*numpy*, *pandas*) for data analysis.

Your Program should be implemented as a library, implementing all essential functions. A simple Python program and a Jupyter notebook should also be used with that library. The script should use *argparse* to read the parameters (such as the output file name or path to a directory with data files), and the analysis results should be written to a file (specified as a program argument). Your library must not contain any hardcoded paths to any data files. Unit tests should also be provided with your code.

Assignment versions

There are two 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 canonical assignment

Use data that can be found under the following links:

- PSP (Państwowa Straż Pożarna, State Fire Service) data about events,
- Data about the area of JST (Jednostka Samorządu Terytorialnego, Local Government Unit) (note that here the code of JSTs is different),
- Data about alcohol selling concessions,
- Data about the population

Calculate basic statistics in Python, such as min/average/max/etc, of data in given sources. Then formulate hypotheses about correlations between:

- The number of people living in an area and
  - the number of fire events,
  - the number of alcohol selling companies.
- The number of alcohol selling companies and the number of fire events.
- And one possible correlation chosen by you (your hypothesis).

Your code should also report possible inconsistencies in the data. When you decide to drop part of data you need to check and report how much information was lost.

Identifying individual data files that should be used for this task is a part of the assignment. Also, the assignment includes identifying data on which information should be merged.

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 using automated tools like pylint or flake to check the code. We also suggest reading https://www.python.org/dev/peps/pep-0008/
- The possibility to install code as a package using `pip install path_to_package_directory`
- The code should be provided in a repository. Putting obsolete files in the repository can lower the grade.
- The possibility to install should be introduced as Pull Request (Merge Request) to the master (or main) branch.
- Code coverage with tests.
- The usage of libraries, which were shown during class.
- Code profiler output, and its explanation.

Important note

During the exam, you will be asked to do a minor amendment to your code. It will not be something complicated, but failing to do this amendment will result in failing the exam (regardless of the rest of the provided code). We have to be sure that you are the authors of the presented solution.

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