Project 2019

Programming for Data Analysis

Due: last commit on or before December 13th

This document contains the instructions for Project 2019 for Programming for Data Analysis. Please be advised that all students are bound by the Quality Assurance Framework [3] at GMIT which includes the Code of Student Conduct and the Policy on Plagiarism. The onus is on the student to ensure they do not, even inadvertently, break the rules. A clean and comprehensive git history (see below) is the best way to demonstrate to the examiner that your submission is your own work. It is, however, expected that you draw on works that are not your own to build your submission and you should systematically reference those works to enhance your submission.

Problem statement

For this project you must create a data set by simulating a real-world phenomenon of your choosing. You may pick any phenomenon you wish – you might pick one that is of interest to you in your personal or professional life. Then, rather than collect data related to the phenomenon, you should model and synthesise such data using Python. We suggest you use the `numpy.random` package for this purpose.

Specifically, in this project you should:

- Choose a real-world phenomenon that can be measured and for which you could collect at least one-hundred data points across at least four different variables.
- Investigate the types of variables involved, their likely distributions, and their relationships with each other.
- Synthesise/simulate a data set as closely matching their properties as possible.
- Detail your research and implement the simulation in a Jupyter notebook – the data set itself can simply be displayed in an output cell within the notebook.

Note that this project is about simulation – you must synthesise a data set. Some students may already have some real-world data sets in their own files. It is okay to base your synthesised data set on these should you wish (please reference it if you do), but the main task in this project is to create a synthesised data set. The next section gives an example project idea.

Example project idea

As a lecturer I might pick the real-world phenomenon of the performance of students studying a ten-credit module. After some research, I decide that the most interesting variable related to this is the mark a student receives in the module - this is going to be one of my variables (*grade*).

Upon investigation of the problem, I find that the number of hours on average a student studies per week (*hours*), the number of times they log onto Moodle in the first three weeks of term (*logins*), and their previous level of degree qualification (*qual*) are closely related to *grade*. The *hours* and *grade* variables will be non-negative real number with two decimal places, *logins* will be a non-zero integer and *qual* will be a categorical variable with four possible values: *none*, *bachelors*, *masters*, or *phd*.

After some online research, I find that full-time post-graduate students study on average four hours per week with a standard deviation of a quarter of an hour and that a normal distribution is an acceptable model of such a variable. Likewise, I investigate the other four variables, and I also look at the relationships between the variables. I devise an algorithm (or method) to generate such a data set, simulating values of the four variables for two-hundred students. I detail all this work in my notebook, and then I add some code in to generate a data set with those properties.

Submission

You must use the version control software Git [1] to track your work and you will submit your project by providing a URL to your git repository. It is suggested you use GitHub [2] for this purpose and that you consider making your repository publicly available so that prospective employers may view it. However, should you wish to, you may restrict general public access to your repository so long as you give permission to the lecturer to view it. Furthermore, any git repository URL to which you provide access to the lecturer will suffice – you don't have to use GitHub. You must submit the URL of your git repository using the link on the course Moodle page before the deadline. You can do this at any time, as the last commit before the deadline will be used as your submission for this project.

Any submission that does not have a full and incremental git history with informative commit messages over the course of the project timeline will be accorded a proportionate mark. It is expected that your repository will have at least tens of commits, with each commit relating to a reasonably small unit of work. In the last week of term, or at any other time, you may be asked by the lecturer to explain the contents of your git repository. While it is encouraged that students will engage in peer learning, any unreferenced documentation and software that is contained in your submission must have been written by you. You can show this by having a long incremental commit history and by being able to explain your code.

Minimum standard

The minimum standard for this project is a git repository containing a README, a gitignore file and a Jupyter notebook. The README need only contain an explanation of what is contained in the repository and how to run the Jupyter notebook. Your notebook should contain the main body of work and should list all references used in completing the project.

A good submission will be clearly organised and contain concise explanations of the particularities of the data set. The analysis contained within the notebook will be well conceived, interesting, and well researched. Note that part of this project is about the use of Jupyter notebooks and so you should make use of all the functionality available in the software including images, links, code and plots. You may use any Python libraries that you wish, whether they have been discussed in class or not.

Marking scheme

This project will be worth 50% of your mark for this module. The following marking scheme will be used to mark the project out of 100%. Students should note, however, that in certain circumstances the examiner's overall impression of the project may influence marks in each individual component.

25%	**Research**	Investigation of the data set as demonstrated by references, background information, and approach.
25%	**Development**	Clear, well-written, and efficient code with appropriate comments.
25%	**Consistency**	Good planning and pragmatic attitude to work as evidenced by commit history.
25%	**Documentation**	Concise descriptions and plots of variables in the data set.

Advice for students

- Your git commit history should be extensive. A reasonable unit of work for a single commit is a small function, or a handful of comments, or a small change that fixes a bug. If you are well organised you will find it easier to determine the size of a reasonable commit, and it will show in your git history.
- Using information, code and data from outside sources is sometimes acceptable — so long as it is licensed to permit this, you clearly reference the source, and the overall project is substantially your own work. Using a source that does not meet these three conditions could jeopardise your mark.

- You must be able to explain your project during and after its completion. Bear this in mind when you are writing your README. If you had trouble understanding something in the first place, you will likely have trouble explaining it a couple of weeks later. Write a short explanation of it in your README, so that you can jog your memory later.
- Everyone is susceptible to procrastination and disorganisation. You are expected to be aware of this and take reasonable measures to avoid them. The best way to do this is to draw up an initial straight-forward project plan and keep it updated. You can show the examiner that you have done this in several ways. The easiest is to summarise the project plan in your README. Another way is to use a to-do list like GitHub Issues.
- Students have problems with projects from time to time. Some of these are unavoidable, such as external factors relating to family issues or illness. In such cases allowances can sometimes be made. Other problems are preventable, such as missing the submission deadline because you are having internet connectivity issues five minutes before it. Students should be able to show that up until an issue arose they had completed a reasonable and proportionate amount of work and took reasonable steps to avoid preventable issues.
- Go easy on yourself - this is one project in one module. It will not define you or your life. A higher overall course mark should not be determined by a single project, but rather your performance in all your work in all your modules. Here, you are just trying to demonstrate to yourself, to the examiners, and to prospective future employers, that you can take a reasonably straight-forward problem and solve it within a few weeks.

References

[1] Software Freedom Conservancy. Git.
`https://git-scm.com/`.

[2] Inc. GitHub. Github.
`https://github.com/`.

[3] GMIT. Quality assurance framework.
`https://www.gmit.ie/general/quality-assurance-framework`.