DATA ANALYTICS REFERENCE DOCUMENT	
Document Title:	Document Title
Document No.:	1540308947
Author(s):	Rita Raher, Gerhard van der Linde
Contributor(s):	

REVISION HISTORY

Revision	Details of Modification(s)	Reason for modification	Date	By
0	Draft release	Document description here	2018/10/23 15:35	Gerhard van der Linde

52465 Programming Project Remit

Project 2018 - Programming for Data Analysis

Due: last commit on or before December 14th

This document contains the instructions for Project 2018 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.