# Project Stage 2: Data Analysis

## Due date: 10:00pm Monday of StuVac (2017-06-12)

*This task is worth* 10% *of your final assessment.*

## Project Stage 2

For this task, you need to do some interesting analysis on a data set. We expect that you will use the data set and tools you already worked with in Stage 1 (but use at least 100 data items, even if you had fewer in Stage 1). If you want to do the extra work to find a different data set, clean it, etc, you are allowed to do so. Alternatively, you may request a data set from us, and we will supply one that has been cleaned (but it may not interest you particularly).

There are two deliverables in this Stage of the Project.

- Submit a written report on your work, as a PDF document. This should be submitted through Turnitin, via the link in the eLearning site. The report should have two distinct parts. Part One is aimed at a general audience that is interested in the domain (for example, if your data set is about pulsars, assume the readers are like those of a popular science article on pulsars). In this part, you should focus on the insight gained from the analysis: describe the domain situation, the origin of the data you used, and then present what your analysis has revealed about the domain. You should include well-chosen visual displays of the summarised data, along with associated textual discussion. If you choose to do so, you make also make available a web site where information is available; in that case, the written report should include the URL of your site. Part Two of your report is aimed at people with interest in IT approaches to data analysis (such as other students in info1903!); this should explain how you did the processing (what tools you used both for analysis and for presentation) – you should include the key aspects of the code, queries, or formulas from the spreadsheet. It should also explain why you made these choices (including things you tried that did not work out, and what you learned from those unsuccessful attempts).
- Submit a copy of the source code that you wrote to perform the analysis you have done. This should be submitted through the eLearning system, as a single archive, compressed file. The nature of the file will vary depending on the tool you chose: if you are processing with Excel, then you might submit a spreadsheet; if you are processing with Unix, submit an archived directory that contains the data file(s), the processing pipeline; if you are processing with SQL, submit an archived directory that contains the data file(s), and the processing sequence of queries; if you are processing with Python, submit an archived directory that contains the data file(s), and the Python program. If you have used multiple tools in your processing, submit an archived directory that contains the data set(s) and all the source of the processing you did.

Here is the marking scheme for this assignment. The marker's evaluation will be made principally on the basis of your report; the submitted data and analysis processing will be considered as evidence to check or clarify statements made in the report.

- There are 3 marks for the outcome of your analysis (as described in Part 1 of the report). A pass (adequate) score indicates that your report delivers an analysis that explores the relationship between at least two aspects or attributes of of the data, using at least 100 items of data. A distinction level score (good work) is awarded if, in addition to the above, you explore the connections among at least four aspects or attributes of the data. Full marks (excellent work) indicates that you have achieved all the distinction-level requirements and in addition, that your exploration shows something interesting and not trivial about some connection among these aspects.

- There are 3 marks for the way you carried out the analysis (as described in Part 2 of the report, and evidenced in the submitted data files and processing). A pass score is awarded if you have written an analysis that correctly produces some meaningful outputs involving at least two attributes, and is understandable for your peers (either it is internally self-evident, or well explained in comments and the report). A distinction score is given if you reach the pass level, and also your processing is well automated, so the whole analysis can be redone for changed data sets with only a command or two). Full marks would be awarded for doing the above where your analysis uses tools with capabilities beyond those taught in INFO1903 (such as more sophisticated libraries, statistics packages, etc).
- There are 4 marks for the way you communicate your work (as captured in the report). A pass score indicates that the intended audience could obtain the main features of a message involving at least two aspects or attributes of the data, without excessive effort or confusion (as part of this, you need to include some helpful visual presentations, and have a clear structure in your report, including a summary of the main conclusions). A distinction score indicates that the intended audience would find good support in your report for finding the main features of the message, involving at least four attributes or aspects of the data, and also that they could obtain a deeper or broader understanding as well. Full marks is awarded if, in addition, you provide the capacity for the user to explore these aspects of the data themselves (for example, through an associated web site with interactive functionality).

## Academic Honesty

This assignment forms part of the assessment for INFO1903. As such, the work you submit must be your own (except where you explicitly acknowledge sources). Please make sure that you have understood the University's Academic Integrity expectations. By submitting work, you will be declaring that you are aware of the policy, that the work is your own, and that you understand that the University may reproduce the work and communicate it to others, in order to run similarity detection software. We urge that if you haven't yet done so, you complete AHEM1001, the online module on Academic Integrity that the University is offering.