DATA1002 (Sem2, 2018) Project Stage 2

Due: 11pm on Friday October 26, 2018 (week 12)
Value: 10% of the unit

This stage is done in the same group as you worked with for Stage 1. If for any reason you want to change group membership, you should urgently contact the unit coordinator alan.fekete@sydney.edu.au .

If, during the course of the assignment work, there is a dispute among group members that you can't resolve, or that will impact your group's capacity to complete the task well, you need to inform the unit coordinator.
Make sure that your email names the group, and is explicit about the difficulty; also make sure this email is copied to all the members of the group. We need to know about problems in time to help fix them, so set early deadlines for group members, and deal with non-performance promptly (don't wait till a few days before the work is due, to complain that someone is not doing their share).
If necessary, the coordinator will split a group, and leave anyone who doesn't participate effectively in a group by themself.

The project work for this stage:

For this task, you need to do some interesting analysis on a data set. We expect that most groups will use the data set and tools you already worked with 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).

You may also choose to change tool from Python to Excel or vice-versa, or even you can decide to do this stage with a mixture of tools (for example, you might use Pandas for the analysis, but then use Excel to produce the charts).

What to submit, and how:

There are two deliverables in this Stage of the Project. ***Each should be submitted by only one person, on behalf of the whole group.***

Submission 1: Report

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.

Report Part One:

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.

Report Part Two:

Aimed at people with interest in IT approaches to data analysis *(such as other students in data1002!);*

- 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.

Submission 2: Your Source Code

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 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 single spreadsheet;
- if you are processing with Python, submit a compressed (archived) directory that contains the data file(s), and the Python program.
- If you have used multiple tools in your processing, submit a compressed (archived) directory that contains the data set(s) and a spreadsheet and all the Python source.

Marking:

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

Outcome of analysis: 3 Marks

(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 the data.
    - The phrase "explore the relationship" could mean seeing if there is a trend that describes how one attribute's value is influenced by the values of other attributes,
    - or it could mean deciding whether the distribution of values of one attribute is different among different subsets of the data, defined by the values of other attributes, etc.
- A distinction level score (good work) is awarded if,
    - you have used at least 100 data items in your analysis, and also
    - your analysis explores connections that (among them) involve at least four aspects or attributes of the data.
- Full marks (excellent work) indicates that you have
    - used at least 500 data items, and that you
    - have a sensible exploration of how at least four attributes are related together [that is, you haven't just considered pairwise relationships among the four, but really a four-way relationship].

Methodology of Analysis: 3 Marks

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 your processing correctly produces some meaningful outputs
    - involving at least two attributes, and
    - it is clear that the calculations are correct (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 very well automated, so the whole analysis and chart production can be redone for changed data sets with only a command or two).
- Full marks would be awarded for doing the above, and also your analysis uses features of Python or Excel that go beyond those taught in DATA1002 (such as more sophisticated libraries, statistics packages, etc).

Communication of Analysis Results: 2 Marks

The way you communicate the results of your analysis (as shown in Part 1 of the report).

- A pass score indicates that the intended audience could gain knowledge of some feature of the data, without excessive effort or confusion
    - (as part of this, you need to include some visual presentations that are helpful; the report should also be explicit about the properties the readers should observe).
- A distinction score indicates that the report is well-targeted to make it *easy* for the intended audience to gain *understanding* of several aspects of the data
    - (this includes clearly linking your writing to the audience's background and aims;
    - it also requires that the charts draw attention to important properties of that data; as well, the writing must provide a convincing justification of any claims made about the data).
- Full marks is awarded for a report that meets all the Distinction criteria, and
    - also it conveys a sophisticated understanding of a complex and non-obvious relationship that is found in the data.

Communication of techniques: 2 Marks

The way you communicate the techniques used for doing your analysis and charting (as shown in Part 2 of the report).

- A pass score indicates that the intended audience could gain knowledge of what you did, without excessive effort or confusion
    - (as part of this, you need to include clear descriptions of the computations).
- A distinction score indicates that the report is well-targeted to make it *easy* for the intended audience to gain *understanding* of the techniques you used and of the lessons you learned about the techniques
    - (this includes clearly linking your writing to the audience's background and aims; it also requires that the report reflect on strengths and limitations of the techniques used).
- Full marks is awarded for a report that
    - meets all the Distinction criteria, and also it
    - conveys a sophisticated understanding of some technique that goes beyond what was taught in data1002.

Advice:

During the project, you need to manage the work among the group members. We recommend that you do NOT allocate a different kind of work to each person. That is, don't get one member to write code, another to produce graphs, another to write text, etc.

Instead, we recommend that every person do each activity (perhaps for exploring the relationships of a different group of attributes). This will be important for preparing each member for the final exam.

In choosing between a spreadsheet and Python code, as the tool to use, you need to consider the capability of each tool and also your expertise.

In general, spreadsheets make simple things easier, but it is harder to extend one to more sophisticated tasks. So to get a pass mark, it may be easier in Excel, but this may make it harder to reach full marks. One possible approach is to do a passing solution quickly in a spreadsheet, and then try to redo it (with more sophistication) in Python; that way, if the Python doesn't work, you still have something that can get reasonable marks.

Note that you can use a different tool in Stage 2, from what you used in Stage1. You can also combine tools; perhaps doing analysis in Python but then outputting summary data that you import into a spreadsheet to produce charts.

Late work:

As announced in CUSP: Late work (without approved special consideration or arrangements) suffers a penalty of 10% of the available marks, on each calendar day after the due date.

No late work will be accepted more than 5 calendar days after the due date. If this stage is missed or badly done, the group can be given a clean data set, for a domain chosen by the instructor, to use in the rest of the project.