DATA1002 (Sem2, 2018) Project Stage 1

Due: 11pm on Friday October 5, 2018 (week 9)
Value: 5% of the unit

This assignment is done in **groups of up to 4 students** (we expect 3 or 4 students in most groups, but it may happen that sometimes a group is smaller, eg if there are not enough students in a lab). We recommend that all students in a group be attending the same lab session, so you can work together more easily.

Group formation procedure: In week 6 lab, you should form a group. In choosing who you want to work with, we suggest that you aim to be able to agree on the domain of the data you will work with (eg finance, biology, meteorology, sociology, literature, etc), and also on the tool that you will use (Python code, or a spreadsheet). If necessary, the tutor may rearrange group membership. One person in the group should go to the "People" page of the unit's Canvas site, and then on the "Groups" tab of that page, pick an empty group for stage1, and join it. The "leader" can inform the others of the group name, and the other members can then use this page and join the same group.

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, alan.fekete@sydney.edu.au. 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: You need to obtain a data set. This may be any data that interests you. We prefer that you use publicly available data (so we can check your work if we need to) but it is OK for you to work on privately-owned data as long as you have permission to use it, and permission to reveal it to the markers. As you will see in the marking scheme, if you aim for higher marks, then you should make sure that the data is sufficiently large that automated processing shows genuine

benefits, and that it is produced by combining data from at least two different sources.

You are then expected to do whatever transforming and cleaning is appropriate, to get the data so it can be analysed by the tool of your choice. The details of this aspect all vary a lot, depending on both the format of the data you obtained, and the tool you will use for processing. For example, you might have several CSV files that you plan to analyse in Excel, or alternatively you may have a JSON file and want to process it with Python. In any case, there will almost certainly be a need to do some data cleaning (such as removing instances that have corrupted or missing values, or correcting obvious spelling mistakes, etc), or at least checking that the data is clean.

Finally, we ask you to show one very simple analysis, that picks out some subset of the data and reports on some aggregate summaries. This is not intended to be a detailed exploration of the data (that will come in Stage Two), but simply a demonstration that the data is now in a form where you can work with it.

During the project, you need to manage the work among the group members. We advise that you do NOT allocate a separate job to each person. That is, don't get one member to find the data, another to clean it, another to analyse it. This would mean that work is badly spread through the time period for each person, and also it makes the outcome very vulnerable if one member is slow or doesn't do a good job, because each job depends on the previous ones. Instead, we recommend that every person do each activity, and that you compare regularly and take whichever is better (or even, find a way to combine the good features of each). So, each member should hunt for a dataset, and then everyone looks at all the datasets found, and either choose the dataset that has most potential, or even combine several datasets together. Similarly, each member should try to clean the data, and then see who found what issues, and produce a dataset that has all the aspects clean at once.

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 change your tool between Stage1 and Stage2.

What to submit, and how: There are four deliverables in this Stage of the Project. All four should be submitted by one person, on behalf of the whole group.

• Submit a written report on your work, in pdf. This should be submitted through Turnitin, via the link in the eLearning site. The report should have a three-section structure that corresponds to the marking scheme: a section that describes the data source(s), the contents of the data, and what your interest is in this; a section that describes the initial transformation and cleaning that you did (if you did this automatically, include here the code that you used, or a description that is detailed enough to be reproduced); and a section that describes and explains a simple analysis that you have done (including saying which tool you used, and illustrating the output of the analysis).

• Submit a copy of the raw data as you obtained it. This should be submitted through the eLearning system, as a single file (if you got multiple files from your sources, you need to compress them into a single file for submission)

• Submit a copy of the cleaned and transformed data set, that you will use in your tool of choice. This should be submitted through the eLearning system, as a single file.

• Submit a copy of the simple 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 submit a spreadsheet; if you are processing with Python, submit a Python program.

Marking: Here is the mark scheme for this assignment.

• There is 1 mark for the work on obtaining a dataset (as described in Section 1 of the report, and as evidenced in the submitted raw data set). A pass (adequate) score indicates that the data is genuine, that you have clearly showed where you obtained the data, that you have described the contents of the dataset (explaining clearly both the format, and the meaning of the various aspects). A distinction level score (good work) is awarded if, in addition to the above, the amount of data is at least 100 items, and your description shows clearly that you have appropriate rights to use the data in the ways that you do use it, and your explanation shows sensible reflection of the strengths and limitations of the data that you obtained. Full marks (excellent work) indicates that you have achieved all the distinction-level requirements and in addition, that your data set has at

least 500 items, and that your data set is produced by combining data from more than one source. To be considered for full marks, there must be a real challenge in relating the data values in the two sets. It is not enough to simply take two datasets from the same authority (that use the same definitions of attributes etc), nor is it ok just to use unrelated data, where there is not connection made across the information.

• There are 2 marks for the work on transforming and cleaning the data set to support later processing in the tool of your choice (as described in Section 2 of the report, and as evidenced in the changes between the raw data set and the cleaned data set). A pass score indicates that you have produced a version of the data that is able to be used by your tool. A distinction score indicates that you have passed and also that you have carefully examined the source data set for data quality and format difficulties, and that you have dealt reasonably with several of these issues. If you have found a dataset where the data is already clean, you can instead show how you check the data cleanliness and quality properties. Full marks is awarded if, in addition, your transformation and cleaning was to a substantial extent, an automated process (that is, it could be easily performed on extra data, without a lot of manual inspection or adjustment).

• There are 2 marks for the simple analysis work (as described in Section 3 of the report, and evidenced in the submitted analysis). A pass score is awarded if you have written an analysis that can produce output that correctly derives some aggregate value over some subset of the data (for example, it might give the maximum value of one attribute, among all items with a given value for some other attribute). A distinction score is given if you produce output that gives the aggregates over multiple subsets (for example, the maximum value of some attribute in each subset corresponding to a value of another attribute like an Excel pivot table). Full marks would be awarded for doing the above where your analysis combines and connects data that originated in different data sources [for example, the attribute you aggregate may be from a different source than the attribute used to determine the subset; clearly this is only possible if you used data from more than one source].

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.