DATA1002/1902 (Sem2, 2019) Project Stage 1

Due: 11:59pm on Friday October 11, 2019 (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. The group must all be enrolled in the same unit (either all from DATA1002, or all from DATA1902).

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.). If possible, also make sure that you have some common free times so you can get together to work on the project. Another goal is diversity in skills (eg someone good at coding, someone good at writing). Finally, it is very important to be clear with one another how much time and effort you will devote to the work (you don't have to all give the same effort, as long as everyone knows that each can be relied on to do what they commit to). If necessary, the tutor may rearrange group membership.

The group members should report all their unikeys to the tutor, and unit staff will then join them as members of an official group on Canvas.

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 delivering on their tasks). If necessary, the coordinator will split a group, and leave anyone who didn't participate effectively, in a group by themselves (they will need to achieve all the outcomes on their own).

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 to ensure high-quality data that can be usefully analysed; we expect you to write Python code that does whatever transforming and cleaning is appropriate, The details of this aspect all vary a lot, depending on the data you obtained. For example, you might have several CSV files or alternatively you may have a JSON file; the work needed may be removing instances that have corrupted or missing values, or correcting obvious spelling mistakes, etc). In any case, you are required to get the data to be fairly clean; for some data sets, you need to clean the data, in others that were carefully curated before you got them, you would at least check that the data is clean.

Finally, we ask you to show some very simple analysis, that 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, and that you have the required skills in Python coding.

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. Note that this project stage is not a huge amount of work; it can all easily be done by one person.

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. The marks will be associated with the report.

• Submit a written report on your work, in pdf. This should be submitted through Turnitin, via the link in the Canvas site. The report should be targeted at a tutor or lecturer whose goal is to see what you did, so they can allocate a mark. The report should have a three-section structure that corresponds to the marking scheme: a section that describes the data source(s), the format/contents of the data, the rights associated with the data; a section that describes the initial transformation and cleaning that you did (include here the parts of Python code that you used, or a description that is detailed enough to be followed); and a section that describes and explains some simple analysis that you have done (again, show the code and also the output of the analysis). There is no required minimum or maximum length for the report; write whatever is needed to show the reader that you have earned the marks, and don't say more than that!

• Submit a copy of the raw data as you obtained it. This should be submitted through the Canvas 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. This should be submitted through the Canvas system, as a single file.

• Submit a copy of the processing Python code you wrote for cleaning and analysis. This should be submitted through the Canvas system, as a single file.

Marking: Here is the mark scheme for this assignment. Note that all members of the group receive the same score.

The marking of each of the components will depend on the volume and diversity of data you have. For volume, we will consider the number of "values": for the most common case, rectangular data eg CSV, the contents of a field for an item would be a value. So if you have 100 rows of data, each with 5 attributes, that would be 500 values. For JSON data, the keys don't count, and the values count based on their atomic (string, numnbver etc) components: so if one attribute's value somewhere is a list of 5 numbers, that counts as 5 values; if it is a dictionary with 7 keys, each associated to a string, that counts as 7 values. For diversity, what matters is whether there are truly independent sources of the data. If you get several data sets, but they are all from the Australian census, that only counts as one source; similarly, if you get datasets that are all from the World Bank, they care considered only one source. But if you get some data from Australian census, and some from US Census, that counts as two sources. In each component of the marking, the score you can get is capped depending on the volume and diversity.

• **To gain a Pass mark in any component, your data must have at least 100 values (we say this is a "simple" data set). To gain a Distinction level mark in any component, you must have at least 500 values, and they must come from at least two independent sources (a "medium" dataset). To gain full marks in any component, you must have used at least 3 sources where there are at least 1000 values from *each* of these sources (a "complex" dataset).** To be considered for full marks, there must be a real challenge in relating the data values in the three sets. It is not enough to simply take datasets 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 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 you have at least a simple dataset with genuine data, 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, your dataset is at least medium scale, 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 is complex.

• 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 a simple dataset that is able to be used for analysis (it may still have data quality problems, but not so much as to prevent analysis). At least one aspect of data quality must have been checked and (if there is some problem) it has been handled, but at this level, the checking and handling could be done by manual inspection rather than by code. A distinction score indicates that you have passed and also that the data is at least medium scale, and that you have carefully examined the source data set for data quality and format difficulties, and that you have code that automatically checks for, and deals with, several of these issues. Full marks is awarded if, in addition, you have been able to effectively and automatically integrate the data from a complex dataset (have code that transforms related data from the different sources into common formats and conventions, so the connections can be used in your analysis).

• There are 2 marks for the simple analysis work (as described in Section 3 of the report, and evidenced in the submitted code). A pass score is awarded if you have written Python code which runs on the dataset, and correctly reports on at least one suitable summary statistic (such as the highest value, or the number of different values) for one attribute of the dataset. A distinction score is given if your code gives at least one useful aggregate for each of the attributes in the data (which must be at least medium scale), and furthermore, the summaries include several different kinds (eg one may be the max, another the average, yet another the number of distinct values). Full marks would be awarded if, in addition to the above, your code gives statistics that break down the data from a complex dataset in sensible ways (eg, if the data contains a state attribute, it reports the summaries for data from each state separately, as well as the overall summaries).

Late work: As announced in CUSP, late work (without approved special consideration or arrangements) suffers a penalty of 5% of the available marks (that is, 0.25 marks), for each calendar day after the due date. No late work will be accepted more than 10 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.