# DATA1002/1902 (Sem2, 2020) Project Stage 1

Due: 11:59pm on Sunday October 25, 2020 (end of week 8)
Value: 5% of the unit
Draft worth 0% is due 11:59pm on Tuesday October (week 7)

*Note: these instructions are long and somewhat complicated, but the work you need to do is not actually very much. It should be easy to fit into the provided four weeks of your time, as long as you interact frequently and apply any feedback from the tutors. You need to manage your group's time carefully, as you will want to focus in week 6 on the Practice Python Coding Test. Don't wait till near the due date to start! If anything in the instructions is unclear or confusing, please ask about it on Edstem, using the tag "Project".*

This assignment is done in **groups** (we expect **3 or 4 students in most groups**, but it may happen that sometimes a group is smaller, if there are not enough students in a lab with a particular interest, or if a larger group was split after being unable to work well together; a group of 5 members will only be allowed with approval from the unit coordinator, and is only to happen in special circumstances such as adding someone who had missed allocation, to an already formed group of 4). All students in a group must 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). Note: all members of the group will get the same mark for this assessment.

**Group formation procedure**: In week 4 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.). 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: groups seem to work best when everyone is compatible in their level of drive (eg if someone is eager for High Distinction grades, they can find it challenging working with a partner whose attitude is "near enough is good enough"). While you don't have to all give the same effort, a minimal requirement is that everyone knows that each person can be relied on to do what they commit to.

When you think you have a group who all want to work together, you need to settle some things among the members. Exchange names and contact information (eg which social media platforms you prefer for co-ordinating). You also need to arrange when to get together (virtually): at least one meeting per week (in addition to your scheduled lab session) is vital, but more frequent coordination is even better. Almost all students (except those in DATA1902) have a timetabled slot at 4pm on Fridays, for holding a group meeting, so that is usually a possibility. Please tell your fellow group-members if you can't participate in that time; if that happens, you need to negotiate another time for your meeting, that works for everyone in the group – this needs to be settled before the end of the lab.

One of the group members should report all the member's unikeys to the tutor through the Zoom chat window, and unit staff will then place them as members of an official group on Canvas. **If necessary, the tutor may rearrange group membership;** this is most often needed when someone, who is left out of any other group, gets added to an existing group, but the tutor is also allowed to split a 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 (including anyone you are complaining about). 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:**

Summary of the work to do:

- Obtain a suitable data set
- Ensure that you have data which has good quality and is clean from serious errors

- Produce a few summaries (aggregates) of some attributes
- Write a report and submit it

Details of the work: You need to obtain some data set(s). 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 obtained several CSV files or alternatively you may have one JSON file; the work needed may be removing instances that have corrupted or missing values, or filling in those missing values in some sensible way; you may be correcting obvious spelling mistakes, or changing data formats in one source to match those used in another, etc; you may be putting data from different sources in a single file and removing duplicates. 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 other cases where your data sources were carefully curated already, you would at least write a program that checks that the data is clean (for example, by showing there is no missing data, or that every entry has an appropriate value for that attribute). At the end of this part of the work, you will have dataset(s) which should be high-quality; you also need to produce helpful metadata about this data, describing the sources, any changes that you made, the meaning of each attribute, etc.

Finally, we ask you to show some very simple analysis, that reports on some aggregate summaries of some of the attributes. This is not intended to be a detailed exploration of the data (that will come in Stage Two), but it is 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.

**Group process:** During the project, you need to manage the work among the group members. We advise that you do NOT allocate a different *type* of 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. Also, make sure to quickly report any difficulty in working together, to the unit coordinator as described above.

**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 from this stage will appear in canvas gradebook as being associated with the report submission; the other submissions have no marks appearing for them in Canvas, but they can be used as evidence in determining the mark for the stage.

• Submit a **Stage 1 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:

1. a section that describes the data source(s), the format/contents of the data, the rights associated with the data, and some comment on any strengths or limitations of the dataset;

2. 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); this section should end with a brief explanation of where (in the submitted material) is found the metadata for the resulting dataset(s) which you will use in the rest of your project;

3. 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! For most groups, three or four pages

should be plenty.

• Submit a copy of the **Stage 1 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 on Canvas.

• Submit a copy of the **Stage 1 clean dataset** as you will use it in later work. This should be submitted through the Canvas system, as a single file. Even if your source data was in fact completely clean, you need to submit it again as a separate deliverable here. Note that the clean dataset needs to come with metadata which describes at least:

- the sources of data,
- any licence or other restrictions on use of the data,
- description of any changes you did between the original data and the final dataset; and
- the meaning of each attribute, what format or units are used, etc

If the metadata is in a separate file from the data, or if the data is divided among several files, then you need to compress all the files into one file for submission on Canvas.

• Submit a copy of the **Stage 1 Python code** you wrote for cleaning and for analysis. This should be submitted through the Canvas system, as a single file. If you did parts of the processing separately, for example, one program to deal with missing values, another to fix date formats, and another to produce a simple aggregate, then you can take the separate files for each task, and compress them into a single file to submit on Canvas. If you have used Grok or some other browser-based approach to running your code, make sure you download a copy of the code to have a file you can submit on Canvas.

**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, number 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. To determine 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 are 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 in total, 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 (as we defined that above).

• There are 2 marks for the work on producing a high-quality data set to support later processing (as described in Section 2 of the report, and as evidenced in the submitted code and also in any changes between the raw data set and the cleaned data set). A pass score indicates that at least one aspect of data quality in a simple dataset has been automatically checked

and (if there is some problem) it has been handled. A distinction score indicates that the data is at least medium scale, and that you have Python code that automatically checks for, and handles in a sensible way, several different kinds of data quality and format difficulties (eg it must not only check for missing data entries, it must not only check for values which are out-of-valid-range). Full marks is awarded if you achieved Distinction and, 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 aggregate 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 data has medium scale, your code gives *at least four useful summaries* and furthermore, at least one of which the summaries must be a grouped-aggregate (for example, if the data contains a state attribute, it might report the summaries for some other attribute from each state separately). Full marks would be awarded if, in addition to the above, your dataset is complex, and also your code gives at least four sensible grouped-aggregate statistics, where these don't all use the same attribute to define the groups, and they also don't all give the same kind of aggregate (eg one might be a sum in each group, and another a maximum in each group).

**Late work:** As announced in the unit outline, late work (without approved special consideration or arrangements) suffers a penalty of 5% of the maximum marks, for each calendar day after the due date. That is, we subtract 0.25 marks per day from what you would otherwise get for the work. 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 request a clean data set, for a domain chosen by the instructor, to use in the later stages of the project.

**Draft submission:** The draft is not worth marks, but its purpose is to be the basis for feedback from your tutors during the week 7 lab session. Submit (at the Canvas link for the draft) a document that is really a draft, that is, structured and formatted like the final report, but with some parts missing, not properly proof-read, maybe some parts are just done as bullet-points rather than proper paragraphs, etc. Your draft should have:

- Part 1 should be fairly complete though perhaps needing polishing: the data sources should be described, and the contents and formats.
- Part 2 should have at least got to the level of describing how you check for some issues, though maybe you haven't yet handled any problems found.
- Part 3 should at least indicate which summaries you plan to include by the final due date.