2022 Sem2

Project Stage 1

Due: 11:59pm on Sunday at the end of week 6

Value: 5% of the unit

Note: In order not to have a long and complicated instructions, some of the details are included in a separated companion note, The project may sound elaborate, but the work you need to do is not actually very much. It should be easy to fit into the provided two weeks of your time, as long as you interact frequently and apply any feedback from the tutors. 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 category "Group Report", and sub-category either "Group Formation" or "Stage 1" as needed.

GROUPS

RULES

1. This assignment is done in groups of 3 or 4.
    - Under exceptional circumstances a group of 5 members may be created by the unit coordinator.
    - The tutor has the final decision of group allocation.
2. All students in a group must be attending the same lab session and will be reporting progress to your lab tutors.
3. There is work required from each member separately, but the project is handed in as a combined effort, and it is marked as a whole: all members of the group will get the same mark for the assessment.

GROUP FORMATION PROCEDURE

1. In week 4 lab, you should form a group.
2. Groups seem to work best when everyone is compatible in their level of drive (e.g., 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").
3. Exchange names and contact information (e.g., which social media platforms you prefer for coordinating).
4. You also need to arrange when to get together (physical or virtually).
    - At least one meeting per week in addition to your scheduled lab session, but more frequent coordination is even better.
5. One of the group members should
    - write all the member's unikeys on paper *if* you are in an on-campus lab OR
    - report the unikeys to the tutor through the Zoom chat window *if* you are in a remote group,

6. The unit staff will place them as members of an official group on Canvas.
7. If necessary, the lab tutor may rearrange group membership *if necessary.*

DISPUTE RESOLUTION

If 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 your lab tutor and the unit coordinator (josiah.poon@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).
- 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).
  - This option is only available up until Friday Week 5, which is the last day with time to resolve the issue before the due date.
  - For any group issues that arise after Week 5, you will need to try to resolve the problem on your own, and you will continue to be treated as a single group which all get the same mark for this Stage, based on whatever is submitted (though you should still let the coordinator know about them).

THE PROJECT WORK FOR STAGE 1:

SUMMARY

Task	Description	Together/individual
1	Identify Topic	Together
2	Obtain Datasets and Metadata	Individual (but communicate with other members)
3	Ensure Data Quality	Individual
4	Produce Summaries	Individual
5	Integrate Datasets	Together
6	Write Report	Together

TASK 1: IDENTIFY TOPIC:

1. Each member of the group will obtain and process a separate data set
   - these sets must all be relevant to a single topic or question (same for all members)
   - the question or issue is not simply a factual matter, but instead looks at relationships where insights might be impactful for some stakeholder groups
2. We realize that you may not find data that completely resolves the issue you are targeted at, but all the data should at least be potentially able to provide some insights.

Some examples can be found in the companion notes.

TASK 2: OBTAIN DATASETS AND METADATA:

1. Each member needs to obtain a *different* dataset that can contribute to the group's exploration of the topic.
2. You need to keep (and provide to us) a copy the data as you originally obtained it.
3. You need to state relevant metadata, about this dataset, including a data dictionary (which indicates which attributes there are, and what each means), and provenance (giving the whole chain, from the original source of the data, through any intermediate collections, up to the place where you got it from [and the date you obtained it]).

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 *so long as you have permission* to use it, and permission to reveal it to the markers. You can read from the companion notes if you want to get a grade beyond Pass mark.

TASK 3: ENSURE DATA QUALITY:

1. Each member then needs to work with their dataset to ensure high-quality data that can be usefully analysed
2. Transforming and cleaning data are possible process. The details of this aspect all vary a lot, depending on the data you obtained.

Some examples can be found in the companion notes. The data needed to be cleaned. If the data sources were carefully curated already, you would at least write a program that checks that the data is clean. At the end of this part of the work, you will have a dataset which should be high-quality.

TASK 4: PRODUCE SOME DATA SUMMARIES:

Each member must use Python to produce some very simple analysis of his/her cleaned dataset, that calculates some aggregate summaries of some of the attributes. Examples can be found in the companion notes.

TASK 5: INTEGRATE THE DATA SETS:

You need to bring your datasets to the group to create a single combined dataset. The effort of this can vary a lot, depending on the commonality between the datasets.

One common situation is when the datasets describe the same sort of information, about different examples of the same kind of entities. Examples can be found in the companion notes.

Warning: if the datasets have different kinds of information for different kinds of entities, then it is hard to see how they can be usefully brought to bear on a single topic. For example, it doesn't make much sense tries to combine weather data from cities in NSW, with demographic data about different countries.

TASK 6: WRITE A REPORT:

Working together as a group, you need to produce a report. The structure of the report is described below in detail, as the report is the main basic for grading in this project. The report has sections for each member's separate work, as well as a combined introduction that explains the topic or issue, and a combined discussion of the data integration activity.

GROUP PROCESS

During the project, you need to manage the work among the group members. It is essential that *you do NOT allocate a different type of job to each person*. That is, don't get one member to find the datasets and metadata, another to clean them all, another to summarize each. 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. Above all, this interferes with the purpose of the project, to assist every student in learning what this unit is trying to teach; such learning involves doing all the different activities of data science. Instead, *we insist that every person do each activity, on a separate data set*. We intend for the members to compare regularly and learn from one another. Make sure to quickly report any difficulty in working together to the unit coordinator as described above.

WHAT TO SUBMIT, AND HOW:

1. There are **three deliverables** in Stage 1 of the Project.
2. All three should be submitted by one person, on behalf of the whole group.
3. The overall mark from this stage will appear under report submission in Canvas gradebook.

SUBMIT A STAGE 1 WRITTEN REPORT ON YOUR WORK, AS A PDF.

1. This should be submitted via the link in the Canvas site.
2. The report should have a front page, that gives the group name, and lists the members involved (giving their SID and unikey, not their name)

The body of the report has structure as follows (this corresponds to the marking scheme):

1. A section that (i) describes the topic or question that you are interested to explore, (ii) it indicates some groups of stakeholders and says how they will be helped by understanding this topic or answering this question, and (iii) it includes a short discussion of the relevance to this issue of the data you have obtained.
2. One section for each dataset you are working with (the section should state the SID/unikey of the group member who did the work for this dataset). In this section, there should be three subsections
    a. a subsection that explains the dataset, including clear statements of the relevant metadata. You need to clearly state the format and structure of the data. Please give a data dictionary listing the different attributes and their meaning. Also clearly identify the provenance: show the chain of transmission, from the original collector of the data to the place from which you obtained it; along with this, clearly show the rights or

restrictions for use that are associated with the data. This subsection should give your thoughts on any strengths or limitations of the dataset, for the purpose of investigating the topic or question of interest.

b. a subsection that describes how you ensured data quality in the dataset. If there were any quality problems, describe what they were and how you cleaned the data; even if there were no problems, you need to describe what problems you checked for, and how you did the checks. If the cleaning was done with a spreadsheet, describe the steps clearly; if the cleaning or checking was done by Python code, then include the code in your report.

c. a subsection that describes and explains some simple analysis that you have done using Python code (show the code and also the output of the analysis).

3. A section that describes the data integration, that produced a single dataset from the separate cleaned ones described above. This section should give a clear account of the schema of the combined dataset, and a clear description of how the integration as performed. If you did the integration with a spreadsheet, state the steps, in turn; if you did it with Python, include the code.

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:

- 2 pages for each dataset (including the relevant code)
- 1 paragraph for topic (roughly), and
- 1 page for integration.

SUBMIT A COPY OF THE STAGE 1 PER-MEMBER DATASETS.

This should be submitted through the Canvas system, as a single zip or tar.gz file. You should put all things in a single folder, with subfolders for each member. The subfolder for a member should contain

a. the raw dataset as it was obtained from the source
b. any spreadsheet or Python code used for cleaning/checking
c. the Python code to calculate some summaries, and
d. the clean version of the dataset (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 include in your submission).

You then compress the top folder (with all these subfolders and their contents), then submit the single compressed file.

SUBMIT A COPY OF THE STAGE 1 INTEGRATED DATASET

You will use this in later work. This should be submitted through the Canvas system, as a single file. Note that the integrated dataset needs to come with metadata which describes at least:

- the provenance of the data,
- any licence or other restrictions on use of the data,

- description of all the changes you did between the original datasets 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 integrated data, or if the data is divided among several files, then you need to put the multiple files into a folder and then compress the folder into one file for submission on Canvas.

MARKING

Here is the marking scheme for this assignment. The score (out of five) is the sum of separate scores for each of five components. Note that all members of the group receive the same score. You can find the full marking details from the companion notes.

IDENTIFYING THE TOPIC [1 POINT]

This component of assessment is based on the corresponding section of the report.

DATASETS AND METADATA [1 POINT]

This component is assessed based on the corresponding subsections of all the separate dataset sections of the report; the uploaded data and code may be checked by the marker as supporting evidence for claims made in the report

ENSURING DATA QUALITY [1 POINT]

This component is assessed based on the corresponding subsections of all the separate dataset sections of the report; the uploaded data and code may be checked by the marker as supporting evidence for claims made in the report.

PRODUCE SOME SIMPLE DATA SUMMARIES [1 POINT]

This component is assessed based on the corresponding subsections of all the separate dataset sections of the report; the uploaded data and code may be checked by the marker as supporting evidence for claims made in the report.

INTEGRATE THE DATASETS [1 POINT]

This component is assessed based on the corresponding section of the report; the uploaded data and code may be checked by the marker as supporting evidence for claims made in the report.

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.