2023 Sem2

Group Project Stage 2

Due: 11:59pm on Sunday (end of week 9)

Value: 10% of the unit

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 two weeks of your time, if 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 "Stage 2".

GROUPS

Rules

This assignment is done in **groups of 3 or 4,** and all students in a group *must* be part of the **same lab session**. Under *VERY* exceptional circumstances a group of 5 members may be created by the unit coordinator (for example, the coordinator may be adding someone who had missed allocation to an already formed group of 4), similarly a smaller group may be created by the coordinator when dealing with group disputes as described below, or when a group is reduced in size due to member discontinuing this unit. Note: 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.

THE PROJECT WORK FOR STAGE2:

Task	Description	Group/ individual	Details
1	Identify topic and datasets	Group	The analysis done in this Stage must all be relevant to a single topic or question, which you are investigating because it matters to some stakeholders. You need to then have one or more datasets that you will analyse, to produce results that are relevant to this topic/question. You are allowed to use the same topic as in Stage 1, but you are equally free to change topic. The members of the group are allowed to all work with the same dataset, or some (or all) may choose to work

Task	Description	Group/ individual	Details
			with different datasets. These datasets are allowed to be cleaned data from Stage 1, or integrated data from Stage 1, or you may choose to obtain new/extra data. There are no requirements for particular origin or volume in the datasets for this Stage. We will make available a dataset (on a topic of our choice) and any group can use that data instead, if they prefer. Note that all members of the group must be working on the same topic/question as each other, even if they use different datasets that deal with different facets of the issue. We realize that the results you produce from analysis may not completely resolve the issue you are targeted at, but each result should at least be potentially able to provide some insights. For example, if your topic is "what influences the average level of wealth in a community?", one analysis may calculate the average wealth in communities having different levels of housing density, and a chart may show how wealth relates to percentage of people living alone in each suburb. Please make sure that your question or issue is not simply a factual matter, but instead looks at relationships where insights might be impactful for some stakeholder groups (for example, it is not a good choice of question to ask just "which country has the highest level of wealth?").
2	Choose summaries and charts to produce	Group	Each member needs to calculate one or more grouped-aggregate summaries from the dataset they are using, and they also must produce one or more charts from that dataset. The number of summaries and charts, and some constraints on what sort of attributes are used, depends on the level of score you are seeking. Details are in the marking scheme below. It is required that all the summaries be distinct from one another, and similarly each chart must be distinct. So you need to coordinate among the

Task	Description	Group/ individual	Details
			members, in case two members want to do the same calculation, one at least will need to change!
3a	Use Python to produce a few tables from parts of the data	Individual	Each member then needs to work with their chosen dataset, to produce the material for their section in Part A of the report. This will involve writing Python code to calculate one or more summaries, and running that code to get the output, this can then be formatted as a Table in the report.
3b	Produce a few charts from parts of the data; evaluate the effectiveness of each chart	Individual	Producing charts can be done either by Python code, or else in a spreadsheet such as excel. You need to describe how each chart was produced, and to evaluate its effectiveness following the approach discussed in Lectures 8A and 8B, which is based on knowing how data attributes get encoded by visual ones in your chart.
3c	Write your section in Part A of the report	Individual	Each member needs to write not more than 4 pages of report, based on what they have done in 3a & 3b above.
4	Write Part B of the report: Communicate your results for interested readers.	Group	Working together as a group, you need to write up a presentation of what your analysis has revealed about the topic. This needs to be written to communicate with readers whose focus is not on the technical details of data science, but rather they are interested in the topic itself. Your report should clearly identify relationships or trends which your analysis has shown (or suggested), and back up these statements with some of the tables or charts taken from what members produced in their individual work. We realise that your work is likely to be limited, and indeed it may be that your analysis suggests that some attributes are not related in any simple way (for example, it may be that wealth and housing density seem fairly independent of one another, or at least, that your data doesn't show any connection!) – that's ok, just be honest in saying what you

Task	Description	Group/ individual	Details
			expected, and what you found. 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 brief combined introduction that explains the topic or issue, and a combined presentation of conclusions.
5	Produce a PDF of the whole report	Group	From the combined document, you need to produce a PDF. As well, there needs to be a file which compresses a folder, within which are subfolders for each member, the subfolders contain the dataset the member worked with, and the code or spreadsheet for producing their analysis (both summaries and charts). One person submits both PDF and zipped folder, to the submission links on Canvas, on behalf of the whole group. Every member of the group will get the marks earned by the combined submission.

SUBMISSION FOR STAGE 2

1. There are **two deliverables** in Stage 2 of the Project to be submitted to Canvas site.
2. All two deliverables 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.

Deliverable	Description
Report	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), and then the body of the report has **two parts** as follows:
	Part A should be targeted at a tutor or lecturer whose goal is to see what you achieved, so they can allocate a mark.
	1. There is an initial section which briefly states the topic of interest, and the stakeholders who care about this. This is not marked as such, it is just so the marker can understand the setting for the rest of the report. (max 1 page) 2. There should be one section for each member (the section should state the

	SID/unikey of the group member who did the work reported in this section). In this section, there should be some subsections. (max 4 pages per member) a. A brief description of the dataset being used by this member; showing at least the schema of the dataset. You do not need here to describe the provenance or give a detailed data dictionary. This is not marked as such, it is just so the marker can understand the tables and charts that follow. b. One or more subsections, each giving a grouped-aggregate summary. In any subsection, you should show the Python code that calculates a summary, followed by a table that presents the output of that summary. c. One or more subsections, each giving a chart. In any subsection, you should describe how you produced the chart, followed by a display of the chart, followed by an evaluation of the effectiveness of the chart. If the chart was produced by Python, the description of how you produced it is the relevant Python code; if you used a spreadsheet to produce the chart, you should state in words the actions you took when creating the chart. 3. If the group is seeking full marks for Chart Production, there will be an extra section, with a chart which shows four attributes and their relationships. This chart may be produced jointly, or by any individual member. (if produced jointly, max 4 pages; if individual, max 2 pages per individual)
	Part B is targeted at someone who is interested in the topic or issue you are investigating.
	4. This section is jointly written by the group. It is written for readers who are interested in the general topic that you have investigated. In it, you describe the specific issue that you have been investigating, and you present some conclusions or insights about this, which you reached from your analysis. You should include some tables and charts (derived from the data and chosen from among those calculated by the members and reported in Part A), to justify or illustrate the conclusions you give. (max 2 pages per group)
	Write whatever is needed to show the reader that you have earned the marks, and don't say more than that! In most cases, the code to produce a summary or chart will be fairly short (a few dozen lines at most), and the evaluation of a chart should not take more than a half- page.
Data and Code	This should be submitted through the Canvas system, as a single zip or tar.gz file. You should put them in a single folder, with *subfolders for each member*. The subfolder for a member should contain the dataset used, the Python code to calculate some summaries, and either Python code or a spreadsheet for producing the charts. You should also include • 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. Compress the top folder (with all these subfolders and their contents), then submit the single compressed file.

MARKING

The score (out of 10) is the sum of separate scores for each of the five components. A student's overall Stage 2 mark will come from the group (40%) and individual marks (60%). However, if all members agree to share the same mark within the group, this should be made explicitly on the front page of the report.

- M1, 2 & 3 are individual marks for the person who write the individual part of Section 2 of the report.
- M4 & M5 are group marks, and all members receive that same score.

M1: GROUP AGGREGATE CALCULATIONS [2 POINTS]

This component is assessed either based on the section prepared by the specific member or based on the corresponding subsections of all the separate member sections in Part A of the report; the uploaded data and code may be checked by the marker as supporting evidence for claims made in the report.

Full marks	The Distinction criteria hold, and all the code is well-documented and clear, and each provides the overall aggregate as well as the aggregate of each group.
Distinction	Every member has written Python code that correctly computes some grouped-aggregate where the grouping is based on a nominal attribute of the data, and each member has written Python code that correctly calculates some grouped aggregate where the grouping is based on a binned quantitative attribute. All the code pieces are distinct from one another.
Pass	Every member has written Python code that correctly calculates a grouped-aggregate summary of some data. All the code pieces are distinct from one another.
Flawed	There is a correct calculation of some grouped-aggregate summary of some data.

M2: CHART PRODUCTION [2 POINTS]

This component is assessed either based on the section prepared by the specific member or the corresponding subsections of all the separate member sections in Part A of the report, as well as in the final 4-aspect chart section if that is present; the uploaded data and code may be checked by the marker as supporting evidence for claims made in the report.

Full marks	The Distinction criteria holds, and also there is at least one chart which illuminates connections that involve at least four aspects or attributes of the data that are relevant to the question, and where there is a reasonable expectation of a relationship where all four attributes interact together [not just that each pair are related, but that the way in which any two relate, is impacted by the values of the other two!]. This chart must be compelling in communicating the information to the reader (e.g. it draws the reader to easily gain a deep awareness of the patterns, especially how the relationship of some are impacted by the other attributes) and makes them keen to learn more.
Distinction	Each member produces at least two charts that accurately convey the relationship between aspects or attributes from their data that are relevant to the topic. For each member, at least one chart must show information about at least three aspects or attributes. For each member, at least one of the aspects shown among their charts must be nominal or ordinal and at least one attribute (possibly in a different chart) must be quantitative. All the charts in the report must be distinct from one another, and without serious flaws (such as distortion or misleading or missing crucial information such as axis scales).
Pass	Each member produces a chart that accurately conveys the relationship between at least two aspects or attributes from their data that are relevant to the topic. [The phrase "convey the relationship" could mean showing whether there is a trend that describes how one attribute's value is influenced by the values of other attributes, or it could mean showing whether the distribution of values of one attribute is different among different subsets of the data, defined by the values of other attributes, etc.]. All the charts in the report must be distinct from one another, and without serious flaws (such as distortion or misleading or missing crucial information such as axis scales).
Flawed	Some charts are produced.

M3: CHART EVALUATION [2 POINTS]

This component is assessed either based on the section prepared by the specific member or based on the corresponding subsections of all the separate member sections in Part A of the report; the uploaded data and code may be checked by the marker as supporting evidence for claims made in the report.

Full marks	The Distinction criteria hold, and, for each chart, there are good reflections on how well (or not) the chart design would work if much more data is obtained.
Distinction	Every member has written an evaluation for each chart in their section, which correctly documents the encoding between data attributes and visual attributes, and documents other decisions (such as style of chart, scale etc.), and it sensibly justifies the decisions in view of the effectiveness of communication. All the charts in the report must be distinct from one another.
Pass	Every member has written an evaluation for each chart in their section, which correctly documents the encoding between data attributes and visual attributes. All the charts in the report must be distinct from one another.

Flawed	Some reasonable attempts to evaluate the effectiveness of some of the charts.

M4: CONCLUSION – CONTENT [2 POINTS]
This component is assessed based on Part B of the report. Material in Part A, or the submitted data and code, may be checked by the marker as supporting evidence for claims made in the report.

Full marks	The Conclusion section has all the Distinction criteria, and discusses honestly and with insight, the limitations, and uncertainties about the results.
Distinction	The Conclusion section provides some accurate information which provides insight into important issues in the topic, supported by at least four relevant tables and at least four relevant charts; the tables and charts must include something produced by each member of the group in their sections of the report.
Pass	The Conclusion section provides some accurate information about the topic, supported by at least two relevant tables and at least two relevant charts which were each part of the earlier material in the report.
Flawed	The Conclusion section contains at least one relevant table and at least one relevant chart, as well as text about the topic.

CONCLUSION – COMMUNICATION [2 POINTS]
This component is assessed based on Part B of the report.

Full marks	The Conclusion section has all the Distinction criteria, and it draws the reader in and engages their attention with vivid and stylish prose.
Distinction	The Conclusion section makes it easy for the intended audience to gain understanding they seek. It clearly links to the readers' background and aims. The structure needs to be logical and well-organised, (for example, tables, charts, and text relate well to one another). It makes explicit what has been learned and aspects which have not been resolved.
Pass	The Conclusion section allows the intended audience to gain some knowledge of the domain, without excessive effort or confusion.
Flawed	A reasonable attempt to communicate.

GROUP PROCESS

During the project, you need to manage the work among the group members. *We insist that every person do each activity and describe what they did and found in the appropriate section of the report and in the appropriate subfolder of the compressed folder that gets submitted*. We intend for the members to compare regularly and learn from one another (as well as from tutor feedback during lab sessions). Because any member's poor work will reduce everyone's score, make sure to quickly report any difficulty in working together to the unit coordinator as described above.

DISPUTE RESOLUTION

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, 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). 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). **This option is only available up until Friday of week 8**, which is the last day with time to resolve the issue before the due date. **For any group issues that arise after this time, 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). If someone doesn't provide material required for the report, or their material is not of the agreed standard, you should still have the report show what that person did. Their section of the report may be empty if they don't produce anything, or it may have material but not enough. In such cases, please put a "Note to marker" on the front page of the report, which describes the circumstances. That way, we can consider how best to apply the marking scheme. Note that it is not expected or sensible, for other members to do the work that someone failed to deliver.

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.