GEA1000 Quantitative Reasoning with Data

Group Project

AY2024/2025, Semester 2

This document details the nature of the group project work for this course.

1 Project

There are three parts to project work this semester.

1.1 Part A: Theory Work

You are to evaluate a quantitative study. You will choose a topic (in the form of a journal article) from a provided list. You should critically evaluate the article using the concepts taught in this course. To scaffold your evaluation, we have formulated a series of guiding questions.

1.2 Part B: Data Work

You will be assigned a data set based on your project group number. A series of questions will be provided to guide your exploratory data analysis. You can use Excel or Radiant or any other software for your work.

1.3 Part C: Presentation Work

You will be tasked to present on a work related to the data set provided in Part B. Q&A on work done in Parts A and B will also be conducted.

2 Timeline

Here is a timeline of events to note.

Week 5	Release of Part A Topics and Part B data sets.
Week 5-10	Choose a Part A topic and work on Parts A, B and C of the project
Odd week groups Week 10	Online submission of report and presentation slides for Part C *(See Items 3-4)* ***Deadline: 2359h, 28 March 2025 (Friday)***
Week 11	Presentation using submitted presentation slides during tutorial time slot
Even week groups Week 11	Online submission of report and presentation slides for Part C *(See Items 3-4)* ***Deadline: 2359h, 4 April 2025 (Friday)***
Week 12	Presentation using submitted presentation slides during tutorial time slot
Reading Week	Peer- and self-evaluation *(See Item 7)* ***Deadline: 2359h, 25 April 2025 (Friday)***

3 Report

Your report should answer all questions in Sections 8 and 9 of this document. They are to be answered with reference to the question numbers. You are not required to repeat the questions in your report.

Please observe the usual academic integrity protocol, and use either the Arial or Calibri font at size 11, set on A4 size paper with normal margins of 2.54 cm.

Type out your answers to both Parts A and B together in a **SINGLE** report that should not exceed **15** pages in length (ideally, Part A should only be about 2-3 pages), inclusive of tables and graphs (cover page, references or appendix, if any, will not be included in the page limit).

You are to include an **executive summary** for Part B of your report. This should be a concise overview (fewer than 150 words) that contains your key findings and proposed solutions to the problem statement.

Important: For Part B, you **must** include sufficient details/images of how you solve each question using the software of your choice. As a guideline, to get the relevant marks, the details provided should be enough for someone else to replicate what you did.

Submit your report by 2359h, 28 March 2025 (Friday of Week 10) for odd week groups, or by 2359h, 4 April 2025 (Friday of Week 11) for even week groups.

4 Presentation Slides

You are to create a set of slides to present your work for Part C.

Submit your slides by 2359h, 28 March 2025 (Friday of Week 10) for odd week groups, or by 2359h, 4 April 2025 (Friday of Week 11) for even week groups.

5 Presentation

In your fifth tutorial session in Week 11 or 12, you will give a 10-minute presentation on the work done in Part C, based on the slides you have submitted. Details on how you are graded is described in Section 10 of this document.

Q&A on work done in Parts A and B will also be conducted.

6 Submission

Nominate **two different students** from your group to perform the submission of your group's report and slides to the Canvas submission folder designated for your project group. One student is to submit the report, while the other is to submit the slides.

Both the report and slides should be submitted in PDF format, and each document should not exceed 10 MB. Name your report and slides following the conventions below:

- **[tutorial group code]-[project group number]-report.pdf** for the report; and
- **[tutorial group code]-[project group number]-slides.pdf** for the slides.

For instance, if you are in Project Group 4 of Tutorial Group D01, your submissions should be named as **D01-4-report.pdf** and **D01-4-slides.pdf**

Please also ensure that all group members' full names and matriculation numbers are clearly written in all documents. ***Note that all submissions will be subjected to plagiarism checks.***

6.1 Acknowledging your use of AI

If you completed any work with the aid of an AI tool, you should acknowledge the use. Using the output of an AI tool without proper acknowledgment is equivalent to lifting or paraphrasing a paragraph from a source without citation and attracts the same sanctions.

You can give this acknowledgment at the end of the assignment explaining, e.g., which AI tools were used, in which parts of the process they were used, what were the prompts used to generate results, and what you did with the outputs to add value. One way this can be done is in a tabular form added at the end of your report (for example, in the appendix and it will not be included in the page limit) as shown below:

AI Tool Used	Prompt and output	How the output is used in the assignment

Alternatively, if an AI tool was used to generate a more extensive set of intermediate outputs that were then developed into a final product, you can also preserve a full transcript of the relevant interactions with the AI in the appendix for submission with your assignment.

7 Peer- and Self-Evaluation

We will solicit your evaluation of every individual in the project group, including yourself. The inputs from a group will potentially have a bearing on the project score of every individual in the group.

For this purpose, an online evaluation form will be open to all students during Reading Week up until 2359h, 25 April 2025 (Friday of Reading Week). You can choose not to respond to the evaluation. However, we will take that to mean you think everyone in your project group contributed equally.

8 Part A: Theory Work

While answering these questions, you should **think critically about the answers and their implications**. We look for evidence that you can apply the concepts covered in the course, and do not expect you to possess domain knowledge of the article's topic to list all possible implications.

Section 1

1. What was the aim of the study?
2. State the main independent and dependent variables and describe how they were measured.
3. a) Summarise the main finding(s) of the study in one sentence. You may choose to quote a sentence directly from the journal article.

   b) What do you think was the target population for the study?

   c) If a main finding of the study involved a test of hypothesis, what were the null and alternative hypotheses for that test? Did the authors have sufficient evidence to reject the null hypothesis?

Section 2

4. If the study was a **controlled experiment**:
   - How were the subjects assigned to treatment and control groups? Was treatment assignment blind to the subjects? How about the assessors?
   - Did the authors of the study report the baseline characteristics of the control and treatment groups? If they did, what do you think was the purpose?

   If the study was an **observational study**:
   - What potential confounders were controlled for by the researchers? For any **one** of the variables stated, explain how this variable may be a potential confounder.
   - Is there a potential confounder which was not controlled for? Explain how it may be a potential confounder.
5. Describe the selection process of subjects used by the researchers. Additionally, include answers to the following, or explain why it is not possible to do so:
   - Who were the subjects studied, and how many subjects were there?
   - To what extent did the sampling frame cover the target population?
   - Was probability sampling employed?
   - What was the response rate of the study?

   Based on your description above, evaluate the generalisability of the study's findings to the target population.

6. Conclude your evaluation of the study as a whole: to what extent was it well-conducted? In your response, you may reflect on the following:

   - In addition to what was discussed in your answer to Question 5, what are the strengths and weaknesses of the study?
   - How do the strengths/weaknesses of the study aid/compromise its conclusion?
   - How do these strengths and weaknesses weigh against one another?

9 Part B: Data Work

Exploratory data analysis is an iterative, investigative approach to data that seeks to explore trends and relationships, formulate questions and obtain answers, and ultimately learn from data. In this part of the project, you will explore a data set and utilise it to generate conclusions and guide decisions. The following sections provide a basic framework for you to apply exploratory data analysis.

We look for evidence that you can apply the concepts covered in the course, and do not expect you to possess knowledge to the level of an expert on mental health/depression in your work.

Brief

Depression is a global mental health challenge that affects millions of people across all age groups, yet it remains widely misunderstood and often stigmatised. This project invites you to explore the complexities of depression, shedding light on its causes, effects, and potential solutions.

It is currently the summer vacation. You and your groupmates are interns with NUS Health & Wellbeing, Office of the President at the National University of Singapore. Over the past semester, your mentor had observed many students who describe school work to be "highly stressful", with some of them even describing themselves as "depressed". Your mentor is worried, and has asked your group for ideas on forming early detection systems for students at risk of falling into depression, or policies to reduce the risk of students falling into depression.

As a start, your mentor has given your group a data set from a study on depression conducted in India and has encouraged your group to explore the data set to find out interesting trends **amongst students** that may be learnt upon to determine how policies regarding depression may be crafted in NUS. Your mentor has provided your group with some basic questions about the data set, and has tasked your group to prepare a report on these questions.

Project Instructions

Download the data set assigned to your group from Canvas. For example, if you are in Project Group 1, you will use "group1.csv"; if you are in Project Group 2, you will use "group2.csv", and so on. Excel workbook versions of the same data sets have also been provided.

Work through the following sections and present your responses in a report, beginning with an **executive summary of fewer than 150 words** that highlights your key findings and proposed solution to form early detection systems for students at risk of falling into depression, or policies to reduce the risk of NUS students falling into depression. **You are encouraged to comment beyond answering questions verbatim** to demonstrate your insight. You may also find these sections helpful for your preparation for Part C of this project.

An appendix that gives a brief description of each variable in your data set is also provided. Refer to the last section of this instruction document "Appendix" for more details.

Reminder: In your **report**, you **must** include sufficient details/images of how you solve each question using the software of your choice. As a guideline, to get the relevant marks, the details provided should be enough for someone else to replicate what you did.

Section 1: Data Cleaning and Comprehension

We begin by looking at the data set as a whole. After verifying the source of data (i.e. data set comes from a reliable source), we can describe the characteristics of the data set, such as its size and variables. We may also seek to generate summary statistics for variables to have a sense of the central tendencies and spread of these variables, and to look out for anomalies.

This data set appears to be a compilation of survey responses studying the possible risk factors of depression. We begin with a cursory inspection of the data set.

1. The size of the data set allows us to gauge the generalisability of the conclusions we make. A larger sample better mimics the properties of the population of interest.

   State the number of observations in the data set and specify the types of variables present (categorical/numerical/neither).

2. We next investigate the data set to determine if data cleaning is required. This involves searching for missing data points, wrongly recorded data, or anomalies.

   Do a quick scan of the data set and point out any interesting or unusual observations that could affect any subsequent data analysis. How would you deal with this data?

Section 2: Data Visualisation and Analysis

The next step is to explore variables in the data set and their relationships. Visualisations in the form of plots can provide a broad overview of the distributions of the variables. Contingency tables and scatter plots can also help with the analysis of the associations between variables.

3. Summary statistics condense a large data set into a single set of numbers, and highlights unusual observations. We may also review a large data set easily with the use of visualisations.

   (a) Treating `CGPA` as a numerical variable, provide its 5-number summary, along with its mean and standard deviation. Based on these statistics alone, what can you infer about the participants of the study?

   (b) Plot an appropriate visualisation for the variable `CGPA`. Are there outliers or interesting or unusual observations when analysing the `CGPA` variable? If there are outliers, how many of them are there, and how will you treat them?

4. Different visualisations are used for different purposes. Some are better at teasing out the relationship between quantitative variables, while others are more effective in showing the distribution of a quantitative variable across multiple categories.

   It is commonly said that hard work begets success. With an appropriate visualisation, analyse the relationship between time spent studying and the measure of academic success. What is the predicted CGPA of a person who studies an average of 10 hours a day, if linear regression is used? Comment on the suitability of the linear regression model in answering the question: “Does hard work beget success?”.

5. It is also expected that studying for longer hours increases the stress felt by the student. With the help of an appropriate plot, compare the distribution of the number of hours spent studying across the various levels of academic pressure.

6. Beyond exploring relationships between variables within a sample, it is possible to generalise the findings to a larger population.

   Using data from your sample, construct a 95% confidence interval to estimate the rate of depression amongst students in India. What can you say about this estimate?

7. As you explore more relationships between the variables, you might stumble across interesting observations. Upon knowing that you were examining a data set on depression amongst students in India, a fellow intern remarked, "the key to getting rid of depression for all students in India is to get lots of sleep!".

   Infer from the data set if there is an association between sleeping for more than 8 hours a day and not being depressed **amongst all students in India**. Address the statement made by the intern, including any limitations or assumptions that you made.

Section 3: Thinking with Data and Reflecting on Data

At this stage, you have explored a few variables and made use of data to ask and answer questions. Let's dive deeper for a second round of data analysis!

8. a) Having a well-designed research question is a critical beginning to any data-driven research problem. Craft a research question that you can answer with the help of the data set. This research question should open discussions to new ideas and new perspectives as compared to what has been studied above. State the **motivations** behind why you are interested in the question and **what you may expect to learn** from answering the question.

   b) Answer the research question that you have crafted in part 8(a) using the given data set and the concepts that have been introduced in this course.

Ultimately, we would like to synthesise useful insights to cap off our learning from data.

9. Suppose you were able to reconduct the study that produced the data set that you have been given.

   If you were able to collect **one** more piece of information from each subject in the current data set (i.e. you can have one more variable of information in the data set), what would it be? Why? Explain how you would like to measure that variable, and any considerations you may have when interpreting information from that variable.

10 Part C: Presentation Work

After reading your group's report, your mentor now wants you to help the department generate some ideas to detect or reduce the risk of students falling into depression.

Based on your findings in Part B, propose a policy to serve as an early detection system for NUS students at risk of falling into depression, or a policy to reduce the risk of NUS students falling into depression.

Your proposal should be guided by lessons learnt from your data set:

- What have you learnt from the data set, and how has it informed you in crafting your policy?
- How would you know if your policy has been successful?
- What are some pitfalls of your policy, or gaps in your work that you foresee?

Your mentor has requested for your group to prepare a 10-minute presentation to share your proposed policy. The presentation should include the problem statement, your group's proposed solution, data-backed evidence for the proposed solution and some considerations on the feasibility of the solution.

The presentation will be assessed based on:

- **Communication skills**: Being able to communicate your policy succinctly and effectively to a general audience (i.e. an audience with potentially no prior knowledge about the topic at hand);
- **Competency**: Being able to support your proposed policy with evidence from the given data set in Part B, and being able to use the appropriate tools to showcase the evidence in a clear and succinct manner;
- **Delivery**: Being able to engage and excite the audience to be interested in your work.

It is **not required** that every member speaks during the presentation. Your group may send the best candidate(s) to present. To ensure fairness, those who are not presenting should contribute to the project in other areas as deemed fit by the group.

There will be a Q&A session after your presentation to address queries about your group's policy, and your work on Parts A and B.

Appendix

The following lists a brief description of the variables present in the data set:

- `id`: Identifier variable for a respondent.
- `Gender`: Respondent’s gender.
- `Age`: Respondent’s age.
- `City`: City of residence of the respondent.
- `Profession`: Respondent’s main profession.
- `Academic_Pressure`: Pressure felt by the respondent from past/present studies, rated from 0 (no pressure) to 5 (very high pressure).
- `Work_Pressure`: Pressure felt by the respondent from work (if respondent works), rated from 0 (no pressure) to 5 (very high pressure). “0” is keyed in if a respondent does not work.
- `CGPA`: Respondent’s cumulative grade point average, which is calculated as follows: for each subject/course that a student completes, he/she receives a grade ranging from 0 to 10 points. The CGPA is the sum of all such points accumulated by a student, divided by the number of subjects/courses taken by the student.
- `Study_Satisfaction` : Satisfaction felt by the respondent from past/present studies, rated from 0 (no/little satisfaction) to 5 (immense satisfaction).
- `Job_Satisfaction` : Satisfaction felt by the respondent from work (if respondent works), rated from 0 (no/little satisfaction) to 5 (immense satisfaction). “0” is keyed in if a respondent does not work.
- `Sleep_Duration`: Number of hours a respondent sleeps on average per day.
- `Dietary_Habits`: Level of healthiness of daily diet, as described by respondent.
- `Degree`: Respondent’s current level of education.
- `Have_you_ever_had_suicidal_thoughts_`: Whether the respondent has ever had suicidal thoughts.
- `Work_Study_Hours`: Average number of hours spent on work/study per work/study day.
- `Financial_Stress`: Stress felt by the respondent about finances, rated from 0 (no/little stress) to 5 (very high stress).
- `Family_History_of_Mental_Illness`: Whether the respondent’s immediate family has had any form of mental illness.
- `Depression`: Whether the respondent feels that they are depressed. ”0” indicates that a respondent does not feel that he/she is depressed; while ”1” indicates that a respondent feels that he/she is depressed.