GEA1000 Quantitative Reasoning with Data

Group Project

AY2021/2022, Semester 2

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

1 Project

There are two parts to project work this semester.

1.1 Part A: Theory Work (15%)

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 module. To scaffold your evaluation, we have formulated a series of guiding questions. (See *Appendix A*)

1.2 Part B: Data Work (15%)

You will be assigned a data set based on your project group number. A series of guiding questions will be provided to help you make sense of it. You are to use Excel and/or Radiant for your work. (See *Appendix B*)

2 Suggested Timeline

Here is a timeline of events to note.

Week 6	Release of Part A Topics & Part B Data sets
Week 7-10	Decide and start working on Part A Topics Start working on Part B Data sets
Odd week groups Week 10	Online submission of report/slides for Part A and Part B *(See Items 3-4)* ***Deadline: 12pm, 25 Mar 2022 (Friday)***
Week 11	Presentation using submitted slides
Even week groups Week 11	Online submission of report/slides for Part A and Part B *(See Items 3-4)* ***Deadline: 12pm, 1 Apr 2022 (Friday)***
Week 12	Presentation using submitted slides
Reading Week	Peer- and self-evaluation *(See Item 7)* ***Deadline: 12pm, 22 Apr 2022 (Friday)***

3 Report

Your report should answer the questions in both *Appendix A and Appendix B*. 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 Part A and Part B together in a SINGLE report that should not exceed eight pages in length, inclusive of tables and graphs (cover page, if any, will not be included).

Submit your report by 12pm, 25 Mar 2022 (Friday of Week 10) for odd week groups, or by 12pm, 1 Apr 2022 (Friday of Week 11) for even week groups.

4 Presentation Slides

You are to create a set of no more than 15 slides, to summarise and present work done in both Part A and Part B. You should organise the contents of your slides with bullet points, charts and/or tables.

Submit your slides by 12pm, 25 Mar 2022 (Friday of Week 10) for odd week groups, or by 12pm, 1 Apr 2022 (Friday of Week 11) for even week groups.

5 Presentation

In your fifth tutorial session in Week 11 or 12, you will give a presentation based on the slides you have submitted, to summarise and present the work done in both Part A and Part B. Students will be graded on content and delivery.

For Part B, your tutor may ask you to demonstrate how your group used software to answer certain parts of the exercise.

Each group will have

- 20 minutes for presentation (inclusive of Q & A).

6 Submission

You must submit both the report and slides in PDF format to the LumiNUS submission folder designated for your project group. The size of each document should not exceed 10 MB.

You must name your report:
[tutorial group code]-[project group number]-report.pdf

The slides should follow a similar naming scheme.

So if you are in ProjectGroup 4 of TutorialGroup D01, your submissions should be named as follows:
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.***

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 throughout 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.

Appendix A: Part A Questions

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

Section 1

1. What is the aim of the study?
2. Quote one sentence from the primary source that represents the main finding in the study.
3. Who were the subjects studied? How many subjects were there?
4. What do you think was the target population for this study?

Section 2i

Do this portion if the study is a controlled experiment:

5a. State the main response variables.

5b. How were the subjects assigned to control and treatment groups? Was treatment assignment blind to the subjects? How about the assessors?

5c. If the authors of your chosen study reported a table that contains the baseline characteristics of the control and treatment groups, what do you think is the purpose?

Section 2ii

Do this portion if the study is an observational study:

5a. State the main exposure and response variables. When were they assessed?

5b. What potential confounders were controlled for by the researchers? For up to two such variables, explain briefly why they are potential confounders.

Section 3

6. Can you infer what sampling scheme was employed? Why so? If the response rate is not reported for this study, calculate it if you are able to or explain why if you are not able to.
7. If the main finding involved a test of hypothesis, what is the null hypothesis for that test? Do they have sufficient evidence to reject that null hypothesis?
8. Is there a potential confounder which is not controlled for? Explain how it may be a potential confounder.
9. Are there sources of potential bias or imprecision in this study? Have these been acknowledged by the researchers?
10. In view of questions 8 and 9, to what extent can the findings be generalised to either the target population or a subpopulation?

Appendix B: Part B Questions

You are conducting a study on consumer expenditure in the US. You are provided with a data set containing data collected for a particular quarter from the Consumer Expenditure Survey (CE). CE is a nationwide household survey conducted on the entire US civilian non-institutional population. It includes people living in houses, condominiums, apartments, and group quarters such as college dormitories. It excludes military personnel living overseas or on base, nursing home residents, and people in prisons. The civilian non-institutional population represents more than 98 percent of the total US population.

The selection of households for the survey begins with the definition and selection of primary sampling units (PSUs). PSUs are small clusters of counties grouped together based on the following criteria. 91 PSUs based on population number were included:

- 23 "S" PSUs, which are urban areas
- 52 "N" PSUs, which are mid-urban areas
- 16 "R" PSUs, which are rural areas

The 23 "S" PSUs were all selected. The 52 "N" and 16 "R" PSUs are smaller areas that will be sampled from, with their probabilities being proportional to their populations. 10 "N" and 5 "R" PSUs were selected in the end. All the selected clusters were combined and the list was arranged in alphabetic order by owners' names. Out of the entire list, only the first 12000 entries were selected and contacted for an interview to collect data on large and recurring expenditures that consumers can be expected to recall for a period of 3 months or longer, such as rent and utilities. There was a response rate of 67.1%.

Using the data set provided, you are tasked to write a report on the consumer expenditure patterns in the US. You should use the following 2 documents on LumiNUS.

- `expenditure.csv`: Consolidated data set containing the expenditure for a particular quarter for all project groups (simplified version of the survey results)
- `variable_description.xlsx`: Description of the corresponding variables provided in the data set

Your group's data set depends on your project group number.
For instance, if you are in Project Group 4, you should filter the value 4 under the variable `Group` in `expenditure.csv`.

In answering the following questions, you are expected to include **data visualisations** (such as charts and plots) and **describe how you use Microsoft Excel / Radiant / any other softwares** to answer the questions, where applicable.

1. Provide basic information about your data set prior to data cleaning, inclusive of the number of data points used, the variables that you will be using in later analysis and information on at least 4 of these variables. Other information you may include are gender proportions and age range, etc.

1389

2. Perform Exploratory Data Analysis on your data set prior to data cleaning. Produce at least 2 different (types of) graphical plots (bar plots, scatter plots, histograms, boxplots, etc) involving one or more variables, to highlight anything interesting (or of note) that you discover through the EDA process. For each plot, write a short note describing what you have discovered.

3. Describe your data cleaning and transformation process. Indicate any changes to the number of data points, variable types and combination of variables if any. This process should be guided by the analysis that you plan to do later. For data cleaning, there is no single right way but do pay attention to missing or ambiguous values. Data visualisations may help you too.

4. Determine whether there is a positive association in your data set between being employed and having a high level of entertainment expenditure. The amount of expenditure is considered "high" if it is above or equal to the median and "low" if it is below the median.

5. Determine which type of expenditure is most strongly correlated with income, and provide the regression equation, using income as the independent variable.

6. The US government reported that due to the COVID-19 situation, more people are staying at home and the utilities expenditure have increased for this particular quarter. The average utilities expenditure was USD$400 in the previous quarter. Perform a suitable hypothesis test at 5% level of significance to determine if there is any statistical evidence of an increase in utilities expenditure this particular quarter. Utilities expenditure is the sum of the expenditures of electricity, water, natural gas and telephone. What are the assumptions made for this conclusion?

7. Perform 1 additional analysis for the association in the data set between two variables of your interest. You are allowed to use variables that have been discussed in earlier questions, as long as the association of the two variables has not been previously discussed.

8. Discuss 2 possible limitations of using this data set for analysis. Briefly explain how that arises, and how they affect the interpretation of your results. You can think about the data collection process, the variables or the actual recorded values.