MTHM501 Project Specification 2024

Project description

Date set: 15/10/24

Date due: midday on Friday 1/11/24

Your project will be on a topic chosen by you. It should involve analysis of a dataset (that you identify) to answer a specific question or set of questions. You should use your dataset to demonstrate your mastery of the broad topics we have covered in this course. Those topics are:

1. Producing high quality graphics
2. Producing high quality tables
3. Wrangling data
4. Hypothesis testing

In addition you should demonstrate your ability to conduct one of

1. Handling missing data
2. Hierarchical modelling

You should also demonstrate your ability to go beyond what you have been taught and conduct some self-directed learning, i.e. use some R package not covered in the course, or extend some of the ideas covered, or demonstrate a deeper understanding of a topic through your own exploration.

You should:

1. **Find some data.** You can either collect it yourself or you can source it online. As a general rule, the data you choose should have more than 1000 observations (rows) and a mixture of variable types (numeric, categorical, ordinal etc...). Feel free to merge existing datasets to create a new dataset (this would demonstrate data wrangling skills). Data is available in many places, however here are some suggestions.
   - Exeter Data Mill https://exeterdatamill.com/
   - UK Data Service https://www.ukdataservice.ac.uk/
   - Kaggle https://www.kaggle.com/datasets
   - Google Datasets https://toolbox.google.com/datasetsearch
   - UCI Machine Learning Repository https://archive.ics.uci.edu/ml/index.php
   - World Health Organisation https://apps.who.int/gho/data/node.home
   - Gapminder https://www.gapminder.org/data/

2. **You should formulate a question of interest** (e.g. do countries with higher GDP also have higher life expectancy?, where in the UK is cycle commuting most prevalent? has air pollution decreased in UK cities since lockdown?) that can be answered with the data you have, and provide the most comprehensive answer to that question that you can using the ideas covered in the course. (Note: you can either find the data first and come up with the research question later, or the reverse).

3. **Write a report summarising your findings.** Your answer should be supported by graphics and tables, as well as formal statistical analysis. You should discuss your findings and interpret your results, including the limitations of what you have found.

4. **Put your code in an appendix.** You should include an appendix containing the annotated code that you used to answer the question. This can be R code or the contents of a .Rmd file. The key is that the work be reproducible.

Your report should be no longer than 8 pages **including figures and tables,** **but excluding your appendix**. You may cite literature if you wish, but this is not required. You may use any software you like to create the report (including Word), but RMarkdown is recommended. Whatever you use you must present the report in pdf. ELE will only accept one file so whatever you present as a report must be contained in one document

Here is a suggested structure with some suggested page lengths:

1. **Declaration**: here you must declare your use or non-use of AI assistance according to University guidance https://libguides.exeter.ac.uk/referencing/generativeai

2. **Introduction**: here you should introduce the question you are interested in answering. You could give some background on the problem/question and list which topic areas you will showcase. ∼ 1 page

3. **Objectives**: here you should set out what you plan to do. You should have a research question you are hoping to answer and some objectives to help you get there. ∼ 0.5 page

4. **Data**: here you could describe the data that you will use, where you got it from and briefly describe what wrangling you may have had to do to format it correctly ∼ 0.5 page

5. **Results**: here you should present the results of your analysis. Figures and tables should be presented as you go along. You should include discussion as you go along. ∼ 4-5 pages

6. **Limitations**: here you can highlight any limitations of your work, what likely impact those limitations have and how you might address those limitations in a perfect world. ∼ 1 page

7. **Conclusion**: here you should summarise your finding succinctly ∼ 1 page

8. **Appendices**: You should include a code appendix and an appendix of AI prompts if you have used them.

Some general hints and tips:

- You should indicate which topics you are demonstrating.
- If you include a figure/table, make sure it is discussed. If it isn't, that indicates it's not needed.
- You should aim to have no more than one figure or table per page.
- Write clearly and distinctly using short sentences.

- Pay attention to how the final report looks. Make sure figure labels and legends are readable.
- Make sure there is a logical structure to the report so that a reader understands what you are doing and why you are doing it at every stage
- Your report should be a readable document, so code should be in the appendix, not in the report. Similarly, raw R output should not appear in the report.
- Get started early. Start writing as soon as you can so that problems with compiling the document or downloading R packages are sorted quickly.
- Guidance on how to reference and avoid plagiarism is available on ELE. There is a 30-60 minute module called "Academic Honesty and Plagiarism".

Marking scheme

- Demonstrating mastery of tables, figures and data wrangling - 30 marks
- Demonstrating mastery of hierarchical modelling or missing data handling - 20 marks
- Demonstrating mastery of another topic, beyond the scope of the course - 20 marks
- Demonstrating an understanding of hypothesis testing and multiple comparisons problems - 10 marks
- Presentation, clarity of communication, labelling of graphics 10 marks
- Neat, reproducible code 10 marks
- Total: 100 marks