AUT University
School of Engineering, Computer and Mathematical Sciences
Project
COMP828: Statistical Programming for Data Science

The purpose of this project is to assess your analytical and computing skills on the materials covered. In this project you will source and analyse a dataset of your choice.

Total Possible Marks: 100 marks, which contribute 60% towards your final grade in this paper.

Deadline: 11:59pm, Wednesday, June 5, 2024

Submission Requirements:

- The project should be completed in **Rmarkdown** or **Quarto** and should be compiled as PDF documents.
- For all parts, both the source file (.Rmd or .Qmd) and the PDF file should be submitted and will be marked.
- The source files (.Rmd or .Qmd) should be able to be compiled by the lecturers, so ensure that all relevant supplementary files (e.g. data files) are uploaded. When you compile your source file, the data file/s should be located in the same directory as your source file.
- The specific requirements for each part of the project are detailed in the coming sections.
- Submissions which contain large quantities of unnecessary code, output or text will be penalized.

Late Assignments: Failure to submit the assignment on time will result in a penalty in accordance with the DCT late assignment policy (5% per day up to a maximum of 5 days). If extenuating circumstances (e.g. illness) prevent the timely submission of your assignment you can apply for special consideration. You may also apply for special consideration if such circumstances result in your submission being incomplete. Applications for special consideration should be submitted via Canvas.

Originality/Plagiarism: This assignment is an **individual piece of work**. You are encouraged to discuss the assignment with your lecturers and classmates, however, the work you submit must be your own. Assignments that show similarities to work submitted by other students will be investigated for **plagiarism** and treated very seriously. Plagiarism software, such as TurnItIn, may be used to electronically compare submissions to those of other students and to documents on the internet. Talk to the lecturer if you have any questions about this requirement.

Question 1 Dataset (20 marks)

(a) **Background information about the problem/dataset** (approx 150 - 250 words) (6 marks)

Marking Criteria

- The context of the problem/dataset is clearly described.
- All sources are appropriately referenced.

(b) **Description of dataset**, including (8 marks)

- Source/Location (e.g. url)

- File type (e.g. csv)
- List of variables (description and type)
- Screenshot of dataset
- Type of data cleaning expected

Marking Criteria

- The dataset is described clearly and all elements listed above are included.

(c) 3 proposed research questions/objectives (6 marks)

Marking Criteria

- Questions/objectives are appropriate and clearly stated, and can be answered using data in the dataset sourced.

Notes

- There may not be any R commands required in this question.
- For ideas about where to find a dataset, refer to the "Data Sources" video on Canvas.
- For help with R Markdown, refer to the R Markdown Examples in the "Additional Resources" page on Canvas.
- You are welcome to discuss your idea for the project with the lecturers prior to starting your project.

Question 2 Data Import and Cleaning (20 marks)

Explain and perform the steps required to import, clean and tidy your dataset. This may include, but is not limited to, the following tasks:

- importing the data
- tidying the variable names
- ensuring the data follows the three rules of "Tidy Data"
- creating new variables
- changing the format or type of variables
- renaming variables

Marking Criteria

- Rmd/Qmd file can run without error on any computer.
- All data cleaning etc. is reproducible.
- All data import/cleaning/tidying tasks are completed in R using tidyverse functions (where available).
- The required data import/cleaning/tidying tasks have been successfully completed.
- For all data import/cleaning/tidying tasks, appropriate explanations, code and output/results are included in the PDF file.
- Only necessary code and output are included in the Rmd/Qmd and PDF files.

Question 3 Data Analysis/Report (60 marks)

(a) For each question/objective from Question 1 (c), write at most 2 pages (including graphs, tables, etc. and approx 200 words) to address the following: (45 marks)

- State the question/objective
- Describe the analysis undertaken and show the results.
- The analysis must include at least one graph and some summary statistics.
- At least one objective/research question must include a statistical model/method (e.g. t-test, linear regression, etc.).
- Discuss the results of the analysis

Note: Professionally formatted reports do not usually contain code or raw output. Therefore, this section of the PDF file should **not** contain any R code or raw output. Use chunk options like `echo = FALSE` and `results = "hide"`, functions like `kableExtra::kable` and `xtable::xtable` to create nicely formatted tables, and in-text R commands like `` `r mean()` `` to refer to results.

Marking Criteria

- Appropriate graphs and summary statistics are chosen and correctly used to answer each question/objective.
- Figures have appropriate titles and labels.
- Numerical results are rounded appropriately.
- An appropriate statistical model/method is chosen and applied correctly to answer at least one question/objective.
- All analysis is reproducible and included in the RMarkdown/Quarto file.
- The analysis shown PDF file is nicely formatted and does not include any R code or raw output.
- The results are discussed clearly, and correct and insightful comments are made.

(b) Conclusion (0.5-1 page) (10 marks)

Summarize your findings and discussion implications and limitations of your analysis, and opportunities for future work.

Marking Criteria

- Findings are summarized and the implications are discussed.
- Limitations of the analysis are discussed.
- Opportunities for future work are identified.

(c) Presentation (5 marks)

The report should be professionally presented and well-written.

Marking Criteria

- The report does not have any spelling and grammatical errors.
- The report is professional presented and well-written.

Notes

- All sources must be referenced.
- It is not necessary to reference the course notes.
- R and R packages should be referenced.[1]
- No marks are given for this section, but you will be penalized in the relevant section if sources are not correctly referenced.

[1] For more information refer to: https://www.r-bloggers.com/2018/08/how-to-cite-packages/