Australian National University

STUDENT NUMBER: _____

Research School of Finance, Actuarial Studies and Statistics

Final Project

Semester 2, 2025

STAT3016/4116/6016 - Introduction to Bayesian Data Analysis

Final Project Due Date: Thursday 30th October 11:59pm, AEDT
Weighting: 60%

INSTRUCTIONS TO STUDENTS:

1. The final project is mandatory and individual-based
2. The assignment must be submitted using the online submission on the course Canvas site under the module 'Final Project'. Submit your assignment online as a single PDF document. Please attach your computer code as an appendix to your final project submission.
3. No late assignments will be accepted without prior permission before the due date and time from the course convenor. An assignment submitted after the due date without an approved extension from the course convenor will receive a mark of zero.
4. University policies on plagiarism will be strictly enforced. Be sure that the work you submit is a result of your own efforts. The submission facility Turnitin will provide a similarity score after matching your submission against other student submissions and external sources.

USE OF GENERATIVE ARTIFICIAL INTELLIGENCE (GenAI) TOOLS:

This assessment requires you to analyse data in R or another statistical software package of your choice. You can use generative AI software to assist you with writing code and editing your written work. Use of AI tools for the above permitted tasks must be acknowledged in the following way at the end of your final project report.

I acknowledge the use of [*insert name of AI tool*] to prepare my final project. I used AI to [*list the tasks you used AI for e.g writing code*].

IN ADDITION, you must also prepare an Appendix which shows screenshots of all your AI prompts and the output generated where the output forms part of your submission.
Failure to include this Appendix will result in LOSS of marks.

- You should note that the material generated by AI programs may be inaccurate, incomplete, or otherwise problematic. Thus use of AI may result in a lower quality product with AI unable to produce the sophistication required for the final project. AI tools should be used with caution and proper citation. AI is **not** a replacement for your own thinking and research.
- It is very important that you do not use AI to merely 'do' your assignment for you. Final project submissions that have been generated entirely by AI are not permitted and will be treated as plagiarism and a breach of ANU's Academic Integrity Rule.
- If the outputs of generative AI software form a substantial part of your submission and are not appropriately attributed, teaching staff will determine whether the omission is significant. If so, you may also be asked to explain your understanding of your submission in an oral in-person presentation. If you are unable to satisfactorily demonstrate your understanding of your submission, you may be referred to the ANU Registrar for investigation of potential academic misconduct.

GENERAL DESCRIPTION

For the final project you will analyse a dataset of your choice using any appropriate Bayesian method(s) we have discussed in class. You may also implement Bayesian models that we have not discussed in class. You will need to formulate your own research question(s), and apply your knowledge of Bayesian statistics and computational strategies to answer your chosen question(s).

The dataset could be one of academic or personal interest to you, and that fits one or more of the Bayesian methods discussed in class. Your chosen dataset should be a real data set which you have not analysed before and which has not been analysed in a textbook. Once you have chosen your real data set, it is recommended that you confirm with the lecturer that your choice of data set is suitable for the final project in terms of complexity (number of records and number of variables) and potential research questions that may be answered using Bayesian methods.

The following websites may be useful to you in finding a data set for your final project:

- UC Irvine Machine Learning Repository `https://archive.ics.uci.edu/`
- Kaggle `https://www.kaggle.com/datasets`

As a default option you may choose to analyse the data set available for download here `https://www150.statcan.gc.ca/n1/pub/72m0003x/72m0003x2024001-eng.htm`. This is the public use microdata file for the Canadian Income Survey and provides information on the income and income sources of Canadians, along with their individual and household characteristics. The data set contains a mix of continuous and discrete variables and presents a variety of options to demonstrate your knowledge of Bayesian methods.

PROJECT PROPOSAL

You may submit a short project proposal to the course lecturer no later than Friday 26 September 2025 (or earlier). The project proposal is not graded. It exists primarily for you to get feedback on your project idea and to make sure you have started thinking about your project. The proposal should comprise up to one page addressing the following questions:

- What data set are you using?
- What are the main questions to be addressed?
- What variables in the data set will you use?
- What are your initial thoughts on appropriate models/distributions?
- What questions and/or concerns do you have about the project?

WRITTEN REPORT

You must submit a written report to communicate your project findings. Your report must be submitted electronically via Turnitin on the course website.

Please include the following sections in your report:

- **Introduction**: Explain the motivation behind your research question and state the dataset you are using (including its source). What are the main questions to be addressed? Why is this research question of interest to you?
- **Methodology**: Describe the variables to be used in your analysis. Specify and justify your choice of prior distribution(s) and sampling model(s) and derive your posterior distribution(s). Be sure to define all notation and state any assumptions you make. Provide step by step details of any sampling algorithm you implement, (that is, the sequence of draws at each iteration). Be sure to specify the number of iterations, burn-in period or thinning interval, and choice of proposal distributions (where applicable). Specify what computer software you used to implement your Bayesian model, and whether you wrote your own code, used an existing package in that computer software and/or used a GenAI tool to write code. Even if you use an existing package or GenAI tool, you must still write down the step by step sequence of draws of your sampling algorithm using correct mathematical notation to demonstrate that you know what the algorithm is doing.
- **Results**: Describe the main findings of your analysis. Be sure to relate your discussion on your results back to your original research question(s). Include graphs if appropriate, MCMC convergence diagnostics and model checking results.
- **Conclusions**: Summarize the main findings of your project. What did you learn? What are the key points that a reader should take away? Discuss any limitations of your analysis, for example, did you need to make any simplifying assumptions when deriving your model or hack your code to get something working? Briefly describe any next steps that you would take to extend or improve your analysis if you had more time or additional resources.
- **Appendices**: Attach the main computer code files for your analysis. If applicable you may also include any detailed mathematical derivations in the appendices that do not need to be contained in the main body of the report. If you used GenAI tool(s), attach screenshots of all prompts and output used as part of your submission.
- **Reference List**: If applicable. Please use the Harvard referencing style `https://www.anu.edu.au/students/academic-skills/academic-integrity/referencing/harvard`

Total length guide: 15 pages (excluding appendices but including graphs).

For copyright and plagiarism reasons, no sample projects will be made available to students.

PROJECT GRADING GUIDELINES

In addition to correct specification of your model and the depth of your analysis, the marker will also be looking for the following in your report:

1. **Consistency**: Did you answer your question(s) of interest?
2. **Clarity**: Is it easy for the reader to understand what you did and the arguments you made? Is the report logically structured?
3. **Relevancy**: Did you use Bayesian statistical techniques wisely to address your question?
4. **Interest**: Did you tackle a challenging, interesting question?
5. **Methodology**: Are all components of your model clearly described? Have you defined all your notation? Have you included step by step details of your MCMC algorithm? In other words, is it straightforward for a person to implement your model after reading your report?

Some tips:

- Talk to the teaching staff for advice.
- If you are using techniques we learned in class, you do not need to re-explain the theory behind the techniques. If you are using techniques that we did not cover in class, the techniques should be clearly explained in your report.

Final Project grade breakdown (subject to minor changes):

Item	**Total marks available**
Consistency/Clarity/Graphical Displays	4
Relevancy	3
Interest	8
Methodology	8
Results + Discussion	8
Model Diagnostics	5
Conclusions	4
Total	**40**