Project Guide

Data Analysis STAT 4620/5620

Important Information

Component	Length	Date	% of final grade
Proposal	500 words	5pm Fri 31 Jan	5%
Workshop	10 minutes	In class 10/12 Feb	5%
Presentation	10 minutes	In class 24/26 Mar	10%
Report	5000 words	5pm Wed 16 Apr	25%

Reminder

- **Undergraduates**: An additional 10% of your final grade will be placed on either the assignments, project presentation, project report, or final exam based on whichever helps each student the most.
- **Graduates**: An additional 10% of your final grade will be placed on either the project presentation, project report, or final exam based on whichever helps each student the most.

Project Overview

The project provides the opportunity to practice and demonstrate your mastery of the skills needed to apply data science to real world scenarios. Data analysis projects require a variety of skills to successfully complete, including presentation skills, collaboration skills, and the learning objectives listed in the course syllabus:

- a capacity to recognize important features of data,
- understanding and proficiency with a variety of statistical techniques,
- proficiency coding in R and familiarity with important R packages, and
- familiarity with version control using Git and GitHub

Students will work in groups of 2-4 and complete a data analysis project. In rare cases students will be allowed to complete individual project. The project begins by deciding on a research question of interest and finding data to answer this question. The research question should be simple enough to ask in a single question but deep enough that students will be able to showcase their mastery of the course content. After identifying important features of the data and possible data analytical tools to use, groups will present their preliminary ideas in a workshop format. In the workshop the class as a whole will brainstorm ideas, and students are expected to give collaborative feedback to other groups and be receptive to ideas they receive from others. At the end of the course, groups will present their final analysis to the class and prepare a shareable project report and R package hosted on a GitHub repository.

Proposal

- **Due date**: 5pm Friday 31 January
- **Description** (~500 word document): Groups will submit a document with a list of 2 or 3 research questions. Each question should be a single sentence. For each question groups will describe the data they will use to answer each question, the sources from which they will access the data, and a brief description of the data analysis tools that will be relevant to the analysis. Students coming in to the class with pre-existing projects are encouraged to use those projects for this course. In such cases the proposal need only include 1 research question but should also include a description of any previous analysis done on the project and what additional analyses will be done as part of this course.
- **Rubric** (5% of final grade)
  - *Research Questions*
    0. Cannot be formulated as a single concise question. *Ex:*
       Cannot be answered by a statistical analysis.
       Asks only about simple summary statistics of the data.
    1. Is a concise question but is unclear.
       Statistical analysis can be applied to parts of the question.
    2. Is a single concise question.
       Can be answered by statistical analysis.
       Is of practical importance to the applied domain.
  - *Data*
    0. No attempt to find data.
    1. Sources are given but not correctly cited.
       The data are shown but not described.
    2. Sources are correctly cited.
       Variable are described and measurement units are included.
       The data are adequate to fully answer the question
  - *Data analysis tools*
    0. No description of tools.
    1. Tools are described but not how they are relevant to the question or data.
    2. The tools described are appropriate to the question and data, *or*
       Important features of the data are described along with an explanation of how they might violate assumptions of the tools learned in the course so far.

Workshop

- **Due date**: In class 10/12 February
- **Description (10 minute presentation)**: Groups will give a 5 minute presentation of their project proposal to the class followed by a 5 minute question-and-answer/brainstorming session. The presentation should contain enough information for the rest of the class to understand the reserach question, the basic structure of the data, and the intended tools needed for the analysis. Each student in a group is expected to give at least some of the presentation. After each presentation, students are expected to provide feedback to the group. Feedback is expected to be respectful of the presenters even if you might disagree with certain choices made by the presenters. Useful feedback can include (but is not limited to):
  - explaining why you think a tool is appropriate or not for the data and research question,
  - trying to clarify whether or not the data meets certain assumptions such as normality or independence,
  - suggestion of different tools to use, and
  - explanations of why you think a presented visualisation is effective or not.
- **Rubric** (5% of final grade)
  - *Research question and data*
    0. Data are not described and visualisations are not shown.
       Data sources are not identified.
       The research question is not clearly communicated.
    1. Data are only briefly described.
       Visualisations of the data are shown but not properly explained.
    2. The structure of the data is clearly explained.
       Data sources are clearly identified.
       Informative visualisations of the data are shown and explained.
       The research question is clearly explained.
  - *Data analysis tools*
    0. No description of tools
    1. Tools are described but not how they are relevant to the question or data.
    2. Potential tools and how they relate to the data and question are explained, or
       Important features of the data are described along with an explanation of how they might violate assumptions of the tools learned so far in the course.
  - *Feedback*
    0. Is not respectful of other students' feedback or suggestions.
       Does not give feedback to other groups.
    1. Is respectful of other students' feedback and suggestions.
       Gives at least one piece of feedback to another group.

Presentation

- **Due date**: In class 24/26 March
- **Description (10 minute presentation)**: Groups will give a 10 minutes presentation of their data analysis. The presentation should include a statement of the research question, a description of the data, at least 2 visualisations, a description of the analysis tools, and the results of their analysis. While the project may not be complete at this stage, most of the analysis should be close to completion and any remaining steps should be explained at the end of the presentation.
- **Rubric** (10% of final grade)
  - *Research Question and Data*
    0. The research question is not clearly articulated.
       Data are not described.
       Data sources are not identified.
    1. Data are only briefly described.
    2. The research question is clearly articulated.
       The structure of the data is clearly explained.
       Data sources are clearly identified.
  - *Visualisation*
    0. No visualisations are shown.
    1. Visualisations are shown but are not adequately explained.
    2. Visualisations are shown and clearly explained.
  - *Data Analysis Tools*
    0. No analysis is done or only elementary statistical tools are used.
       No explanation of the tools used is given.
    1.
    2. Tools are used but are not relevant to the research question and data.
    3.
    4. The tools used are relevant to the research question and data.
       The tools used are at a similar level as those covered in the course.
  - *Results*
    0. No attempt to answer the research question is given.
       The results presented are not justified by the analysis.
    1.
    2. The research question is answered but the analysis is flawed or incomplete.
       The research question is only partially answered.
    3.
    4. The research question is answered and fully justified by the analysis, *or*
       the analysis shows that the research question cannot be answered using the data and the analysis answers as much of the research question as is supported by the data.

Report

- **Due date**: 5pm Wednseday 16 April
- **Format (5000 words + R Package + GitHub Repository)**: Groups will collaborate to create R packages that encapsulate their data analysis project using Git and GitHub to provide version control and to track changes to the project. The R package should include all the data used in the project and at least one function, and should be fully documented. The R package should also include a vignette written in R Markdown/Quarto acting as the project report. The structure of the report will be outlined at the end of this document. The GitHub repository should track the history of the R package as each groups works on and adds to their project. Each member of the group is expected to add at least one commit to the repository with a commit message summarizing the changes made by the commit. The commits added by each member also document how much work each student has contributed to the project so it is imperative that everyone commits their own work to the repository.
- **Rubric** (25% of final grade)
  - *R Package*
    - (0.5) The package contains at least one function.
    - (0.5) All data used in the project is available in the package.
    - (0.5) All functions and data in the package are documented and exported.
    - (0.5) The package DESCRIPTION file includes all of the packages that you use in the analysis.
    - (1.0) The project report is included as a package vignette written in R Markdown / Quarto and is able to be successfully compiled.
  - *GitHub Repository*
    0. The repository consists of a single commit containing the entire project.
    1. There are regular commits to the repository but the commit messages do not adequately summarize the changes.
    2. There are regular commits to the repository and the commit messages adequately summarize the changes.
  - *Research Question and Data*
    0. The research question is not clearly articulated.
       Data are not described.
       Data sources are not identified.
    1. Data are only briefly described.
    2. The research question is clearly articulated.
       The structure of the data is clearly explained.
       Data sources are clearly identified.
  - *Visualisation*
    0. No visualisations are shown.
    1. Visualisations are shown but are not adequately explained.

2. Visualisations are shown and clearly explained.

– *Data Analysis Tools*

0. No analysis is done or only elementary statistical tools are used.
   No explanation of the tools used is given.
1.
2. Tools are used but are not relevant to the research question and data.
3.
4. The tools used are relevant to the research question and data.
   The tools used are at a similar level as those covered in the course.

– *Results*

0. No attempt to answer the research question is given.
   The results presented are not justified by the analysis.
1.
2. The research question is answered but the justification by the analysis is flawed or incomplete.
   The research question is only partially answered with no explanation of why it is not fully answered.
3.
4. The research question is answered and fully justified by the analysis, *or*
   the analysis shows that the research question cannot be answered using the data and the analysis answers as much of the research question as is supported by the data.

– *Structure*

0. The report does not follow the structure outlined below.
1.
2. The report follows the structure outlined below.

- **Outline**

1. Front Matter
   – List each member of your group
   – Include "Data Analysis STAT 4620/5620 Winter 24-25"
   – Include a link to the project's GitHub repository
2. Abstract (150 words)
3. Keywords (max 10)
4. Introduction
   – State the research question and provide any necessary background information
5. Data Description
   – Describe the data being careful to define each variable, including measurement units if applicable.
   – Provide any exploratory visualisations relevant to the research question
6. Methods

- Provide a brief description of the analytical tools you will used
- Discuss why the method is appropriate in the context of the research question and the data, including any assumptions required by the method.

7. Analysis
   - Show the code used to actually apply the methods to the data.
   - Include model validation.
8. Results
   - Present and discuss the outputs of your methods.
   - Include any visualisations of the analysis.
9. Conclusions
   - Discuss the results of the previous section in the context of the research question.
   - Give an answer to your research question and discuss any uncertainty in your results.
   - If appropriate, discuss why the research question cannot be fully answered using the data and techniques available.
10. References
    - Cite data sources.
    - Cite literature and other sources useful for explaining the background information, research question, and analytical tools.