STAT 306: Finding Relationships in Data

2024W1 Group Project

This project will be done in groups of four by default. The project involves the analysis of a multivariate dataset of your group's choice. The analysis should include an exploratory component (i.e., visualizing, summarizing, and identifying key characteristics of the data) and a formal regression component that investigates at least one research question or feature of interest based on the data. Each group will submit one copy of their interim proposal midway through the term, and one copy of their final report, along with supplementary files, at the end of the term. Individual peer evaluations and a self-reflection will also be requested at the end of the term. All group members will receive the same grade for the interim proposal and final report by default.

Groups

A survey in the form of a Canvas quiz will be open during the third week of the course. In this survey, you will have three options that will determine how your group is formed:

1. *Work with your lab group.* After **everyone** in the group agrees to work together, **one** member of the group should complete and select this option in the survey.
2. *Choose one other student in your class to work with.* After agreeing with the other student to work together, **one** of you should complete and select this option in the survey. You and your group mate will be randomly paired with two other students in the course.
3. *No preference.* You do not need to complete the survey in this case. You will be randomly paired with three other students in the course.

If any conflicts arise from the survey (e.g., you and your named group mate(s) both complete and choose conflicting options in the survey), you will be assigned option (3) by default. You will be notified about your group the week after the survey closes. Except under special circumstances, no changes will be made to these groups after they are formed.

Data

The data may be taken from an existing source or be data the group collected themselves. The data should include a well-defined response variable and at least three possible explanatory variables. Unless permitted by the instructor, simulated data should not be used. Some motivation should be provided for studying the data in question, such as a hypothesis of interest or a question or prediction about the process generating the data. After describing the collected data, an analysis of the data should be conducted in R using methods encountered in the course. The findings, including any conclusions, are to be written up in a short report.

Interim proposal

An interim proposal will be due the evening of **Monday, November 4**. The interm report should be no longer than one page in length (including tables and figures; excluding references, if any). The font size should be 10pt at a minimum. The interim report should include the following information:

1. The source of the data being used for the project.

2. A brief description of the variables measured (including when, where, how, in what units, plus any other important information).

3. The research question or other motivation behind the analysis of the data, and any brief background or context necessary to understand the motivation.

4. An overview of who in your group will do what in your project. This does not need to be in fine detail, but should broadly outline the responsibilities undertaken by each group member.

Final report

The final report will be due the evening of **Monday, December 9**. Most final reports are expected to be six pages in length, and should not exceed eight pages in length (including tables and figures; excluding references and appendix sections, if any). The font size should be 10pt at a minimum. The report should comprise the following sections (see the rubric at the end of this document for more details):

1. *Introduction*: Describe the data in detail, with specifics on what was recorded and how, along with the motivation behind the analysis.

2. *Analysis*: Present suitable visualizations of the data and a summary of any key features. Explain and apply the chosen statistical methodology to address the question(s) of interest motivating the study.

3. *Conclusion*: Discuss findings from the analysis along with any other pertinent comments of interest. Address the initial research question(s).

4. *Appendix*: Include any other relevant information or materials in this **optional** section. Note that the grader is **not** obligated to read this section, and so any content that is to be graded should be within the main body of the report.

Each group is to submit a PDF copy of their final report on Gradescope, along with their data file and R script.

Peer evaluations

Along with your final report, you will also complete individual peer evaluations using the template provided on Canvas. In rare circumstances, the group project grades for individual members may be adjusted if a lack of cooperation or contribution is emphasized in the evaluations.

Self-reflection

Along with your final report, you will also write a self-reflection that answers the following questions:

1. What is one thing you learned about data analysis or statistical modeling from this project that was not covered in lecture?

2. What was the greatest challenge relating to data analysis or statistical modeling that you faced in this project? Were you able to overcome this challenge? Why or why not?

3. Choose **one** question from the following questions to answer:

   (a) What is one question that you still have about data analysis or statistical modeling?

   (b) At this point in time, which of the topics covered in this course are you least comfortable with? Why?

4. What is the most important takeaway about data analysis or statistical modeling that you got out of this course?

The self-reflection should be no longer than one page in length with minimum 10pt font size.

Grading

Grades for the interm proposal and final report will be assigned based on the rubrics attached at the end of this document. Grades for the peer evaluations and self-reflection will be assigned based on satisfactory completion. The group project overall is out of 45 marks and is worth 10% of your final course grade. The breakdown of the 45 marks is as follows:

Component	**Marks**
Interim proposal	6
Final report	34
Peer evaluations	1
Self-reflection	4
Total	45

Generative AI tools may be used to correct grammar or phrasing of sentences, but they are **not** to be used to generate ideas, opinions, and interpretations that should be fully yours. Suspected academic misconduct will be investigated and, if validated, will be harshly dealt with.

Interim proposal rubric

Component	Marks	Description of inadequacy	Description of adequacy
Data	2	• Source of data not specified • Variable descriptions missing or largely incomplete	• Source of data specified • Variables fully described
Motivation	2	• Background or context for data missing • Motivation for analyzing data missing or overly contrived	• Brief background or context provided for data • Reasonable motivation for analyzing data provided
Member responsibilities	2	• Group member responsibilities not outlined or are significantly unevenly distributed	• Group member responsibilities are outlined and reasonably distributed

Total marks: 6

Final report rubric

Component	Marks	Description of inadequacy	Description of adequacy
Motivation	2	• Background or context for data missing • Motivation for analyzing data missing or overly contrived	• Background or context provided for data • Reasonable motivation for analyzing data provided
Data	4	• Source of data not specified • Data collection methodology not explained nor addressed • Variable descriptions missing or largely incomplete • Performed data cleaning procedures not explained	• Source of data specified • Data collection methodology explained or addressed • Variables fully described • Performed data cleaning procedure clearly explained
Exploratory analysis	8	• Plots missing or inappropriately chosen • Relevant summary statistics missing • Interpretations of exploratory findings missing or unreasonable • Potential need for data transformation not addressed	• Plots are appropriately chosen, fully labeled, and clear • Relevant summary statistics presented and discussed • Interpretations of exploratory findings relevant and logical • Potential data transformations appropriate and explained
Statistical analysis	8	• Chosen methodology inappropriate or reasoning unclear • Assumptions of chosen methodology not addressed or unreasonable • Model selection is not addressed • Relevant outputs of regression analysis missing	• Chosen methodology appropriate for addressing study objective • Assumptions of chosen methodology addressed and reasonable • Model is appropriately chosen • Relevant outputs of regression analysis presented in a readable format
Discussion	4	• Interpretation of analysis findings missing or incorrect • Study objective not addressed • Potential limitations of study not addressed • Questionable aspects (e.g., outliers) not addressed	• Interpretation of analysis findings correct and reasonable • Study objective addressed • Potential limitations of study addressed and solutions proposed • All relevant or interesting aspects of the analysis addressed
Writing	4	• Ideas unclear and writing at times unreadable • Report does not follow proposed structure	• Ideas clearly expressed and writing largely error free • Report follows proposed structure
Data file and R script	4	• Data file missing • R script missing or does not run	• Data file submitted • R script runs and is neatly documented

Total marks: 34