Project Overview

This document provides a general overview of the project expectations for AMOD 5240H. Detailed information and requirements for each component of the project will be provided.

Project Goals

During this project, you will go through the process of:

1. Selecting an appropriate dataset and understanding its context;
2. Describing your data and its variables;
3. Performing exploratory data analysis (EDA) to identify patterns and relationships;
4. Formulating research questions that can be addressed using linear regression;
5. Fitting a linear regression model (simple or multiple) to answer your research question;
6. Performing appropriate hypothesis tests (such as t-tests, z-tests, or chi-squared tests) to support your analysis;
7. Interpreting and communicating your findings through visualizations, statistical results, and written explanations.

The focus of this project is to apply the statistical methods covered in this course to a real-world dataset. The scope is intentionally manageable: you will fit linear regression models, conduct hypothesis tests, and interpret the results—skills that align with the content covered in AMOD 5240H.

Due Dates

There are two ways that you can approach this project.

Multiple Submissions with Feedback

The project contains 3 written components and 1 presentation:

1. **Part 1: Data Background and Variables** - Due October 30 (4% of final grade)
2. **Part 2: EDA and Analysis Proposal** - Due November 13 (4% of final grade)
3. **Part 3: Modelling and Analysis (Final Report)** - Due December 11 (8% of final grade)
    - This report should include all components: background, EDA, analysis, results, and discussion.
    - You can include your original Parts 1 and 2, revised based on feedback; you do not need to rewrite these sections.
4. **Presentation (Recorded)** - Due December 11 (4% of final grade)

Single Submission

1. **Final Report** - Due December 11 (16% of final grade)
    - This report should include all components: background, EDA, analysis, results, and discussion.
2. **Presentation (Recorded)** - Due December 11 (4% of final grade)

Groups

You will be working in groups of 3-4 students. These groups have already been formed and are available on Blackboard.

Component Details

Part 1: Data Background and Variables [~3 pages in length]

This writeup should be approximately **3 pages** in length and is worth **4%** of your final grade. It is due on **October 30**.

In this component, you will introduce your dataset and provide context for your reader. This is where you demonstrate that you understand how your data was collected and what the variables represent, before diving into analysis.

Your writeup should include:

- **Dataset Introduction**: Where did the data come from? What is it about? Why is it interesting or important?
- **Variable Descriptions**: Describe the key variables in your dataset. What do they measure? What are their units? Are they categorical or quantitative?
- **Research Context**: What kinds of questions could be answered with this dataset? What makes this dataset suitable for linear regression analysis?
- **Data Quality Considerations**: Are there missing values? Outliers? Any data cleaning that is required?
    - This would not be a result of EDA, but rather an initial look at the data file (if applicable) and any information provided about the data collection process.

What I'm looking for:

- Clear writing that would help a reader unfamiliar with your dataset understand it.
- Thoughtful selection of which variables to highlight (you don't need to describe every single variable if there are many).
- Evidence that you've explored and understood the structure of your data.

Part 2: EDA and Analysis Proposal [~3-4 pages in length]

This writeup should be approximately **3-4 pages** in length and is worth **4%** of your final grade. It is due on **November 13**.

In this component, you will conduct exploratory data analysis (EDA) and use your findings to propose a specific research question and analysis plan.

Your writeup should include:

- **Brief Dataset Reminder**: A short (1-2 paragraph) reminder of what your dataset is about and the key variables you'll be examining.
- **Exploratory Data Analysis**: Present the most relevant numerical and graphical summaries of your data. This might include:
  - Summary statistics for key variables
  - Visualizations showing distributions (histograms, boxplots)
  - Visualizations showing relationships (scatterplots, correlation matrices)
  - Any patterns, trends, or unusual observations you've noticed
- **Research Question - Linear Regression**: Based on your EDA, state a clear research question that can be addressed using linear regression. For example:
  - "Does [variable X] predict [variable Y]?"
  - "How do [variables X1, X2, X3] together influence [variable Y]?"
- **Research Question - Hypothesis Tests**: Identify any additional hypothesis test(s) you plan to conduct as part of your analysis. For example:
  - Comparing means between two groups (t-test)
  - Testing for associations between categorical variables (chi-squared test)
  - You are required to include at least one hypothesis test that we covered in class in addition to the linear regression analysis. The tests that we will have covered are tests for proportions, tests for means, goodness-of-fit, test of independence, and analysis of variance (ANOVA).
    Note: Hypothesis tests on the regression coefficients are considered part of the modelling process.
- **Proposed Analysis Plan**: Describe how you will analyze the data:
  - What will be your response variable (Y)?
  - What will be your predictor variable(s) (X)?
  - Will you fit a simple or multiple linear regression model?
  - What hypothesis tests do you plan to conduct? (e.g., tests on regression coefficients, overall model F-test, t-tests comparing groups, chi-squared tests for categorical associations)

- **Potential Challenges**: Identify any concerns or limitations you foresee (e.g., outliers, non-linear relationships, small sample size, multicollinearity, lack of independence).

What I'm looking for:

- Thoughtful EDA that informs your research question (not just a dump of plots and tables).
- A clear, focused research question appropriate for the scope of this project.
- A realistic analysis plan using methods covered in AMOD 5240H.
- Evidence that you're thinking ahead about potential issues.

Part 3: Modelling and Analysis - Final Report [~8-10 pages in length]

The final report should be approximately **8-10 pages** in length and is worth **8%** of your final grade. It is due on **December 11**.

This is your complete analysis report that brings together all components of your project. By this point, you will have received feedback on Parts 1 and 2, which you should incorporate into this final writeup.

Your report should include the following sections:

- **Executive Summary** (up to 1 page): A brief overview of your research question, approach, key findings, and conclusions. This should be readable on its own for someone who doesn't have time to read the full report.
- **Introduction and Background**: Introduce your dataset, provide context, and state your research question clearly. (This should be a refined version of Part 1 that is written in a way so that it flows into the rest of the report.)
- **Exploratory Data Analysis**: Present the key EDA that informed your analysis. Include relevant visualizations and summary statistics. (This should be a refined version of Part 2 that is written in a way so that it flows into the rest of the report. You may decide to add or remove some plots/tables based on feedback and how the analysis process proceeded.)
- **Methodology**: Describe your statistical approach:
  - What model did you fit? (Simple or multiple linear regression)
  - What are your response and predictor variables?
  - What assumptions does your model require?
  - What hypothesis tests did you conduct?
- **Results**: Present your findings:
  - Model fit and parameter estimates
  - Model equation
  - Diagnostic plots (residual plots, QQ-plots, etc.)

  - Results of hypothesis tests (with appropriate test statistics, p-values, confidence intervals)
  - Any additional analyses (t-tests, chi-squared tests, etc.)

- **Discussion**: Interpret your results in context:
  - What do your findings mean in practical terms?
  - Do the results make sense given what you know about the data?
  - What are the limitations of your analysis?
  - How well did your model meet the assumptions of linear regression?

- **Conclusion**: Summarize your key findings and their implications. What did you learn? What questions remain?

What I'm looking for:

- A complete, coherent analysis from start to finish.
- Proper application of linear regression modeling.
- Appropriate use and interpretation of hypothesis tests.
- Clear communication of statistical results for a general audience.
- Evidence of critical thinking about your results and their limitations.
- Professional presentation with well-labeled figures and tables.

Note: You will have already written portions of this report in Parts 1 and 2. The final report should integrate and refine that work based on feedback, and add the modeling, results, and discussion sections.

Presentation (Recorded) [5 minutes + 3 minutes for Q&A]

The presentation component is worth **4%** of your final grade. Both your slides and recorded video are due on **December 11**.

Format:

- You will create a **recorded video presentation** that is approximately **5 minutes** in length and no longer than 6 minutes.
- Your presentation should be uploaded to Crowdmark and the recorded video will need to be uploaded to Blackboard.
  - Both should be submitted no later than **December 11**.

Content:

Your presentation should provide a high-level overview of your project:

- **Introduction** (<45 seconds): What dataset did you use? Why is it interesting?
- **Research Question** (<45 seconds): What question did you try to answer?
- **Methods** (<2 minutes): What analysis did you conduct? What model did you fit?

- **Results** (<2 minutes): What did you find? Include key visualizations and statistical results.
- **Conclusions** (<45 seconds): What are your main takeaways?

The times are general guidelines to help you think about the relative lengths of the particular parts of your presentation. You should keep the presentation length to around 5 minutes and <6 minutes in total.

Technical Requirements:

- Video format: MP4 or other common format that can be uploaded to Blackboard
  - I would recommend using Zoom to record the video as it will compress the file for you.
  - Your file sizes should not be more than 20MB for a 5 minute video. (Our lecture recordings are around 2MB per minute of recording.)
- Slides: PDF or PowerPoint format
  - You can make the slides in Quarto!!
- All group members should participate in the presentation
- Ensure audio quality is clear

What I'm looking for:

- Clear, concise communication of your project
- Effective use of visualizations
- Ability to explain statistical concepts to a general audience
- Professional presentation skills
- Equal participation from all group members

Tips for Recording

- Practice your presentation beforehand and time yourselves
- Use screen recording software (e.g., Zoom, OBS Studio, PowerPoint's built-in recording feature)
- Make sure everyone's audio is clear
- Test your video file before submitting to ensure it plays correctly

Acknowledgements

- Grammarly was used to check grammar and spelling.
  - Many edits made as a result because my writing is awful.
- Claude 3.5 Sonnet was used to check for clarity.

  - Only suggestion implemented was to add the grade weights for each component.

- Qwen 2.5 Coder (1.5b-base) was "used" for autocomplete suggestions.
  - This was used sparingly, however was active; I leave the autocomplete suggestions on unless I am teaching.