STAT-427/627 Group Project Instructions

1 Overarching Requirement

Form groups of 2-4 students and produce a

- **Project Plan by Friday of week 10:**
- Either A: a conference-style **poster and presentation** *or* B: a **project report, by week 14 class** - the last class before final exam period.

To achieve the learning outcomes, the products shall clearly demonstrate the application of a variety of sound statistical learning techniques to a real-world data set, evaluating model performance, tuning for better accuracy, and the ability to articulate correct findings and make appropriate recommendations.

The analysis shall be reproducible using the submitted information.

The deliverables shall clearly communicate your work and show cooperative and collaborative efforts towards a high-quality group product.

1.1 Learning Outcomes

Course	Learning Outcomes
427/627	• Identify appropriate statistical learning methods for a given problem involving real data. • Analyze the underlying assumptions of the methods, be able to verify them, and then propose appropriate remedies for invalid assumptions. • Create and use training and test data as appropriate to evaluate the performance of the chosen regression and/or classification techniques and analyze the results. • Illustrate results with appropriate plots and diagram. • Communicate analysis approaches, results/findings and implications in oral presentations or written reports in clear language, appropriately formatted with supporting references. • Collaborate with diverse groups under constraints of time and resources.
627-only	• Identify other possible problems with messy data, such as multicollinearity, understand their consequences, and propose appropriate solutions. • Apply cross-validation techniques to find the optimal degree of flexibility - the best subset of predictors or the optimal tuning parameters.

1.2 Rubric

The overall allocation of points is in the following table. The detailed requirements for each aspect are in the following sections.

Element	427	627
1 Project Plan	10	10

Element	427	627
2 Poster/ Report Content	30	30
3 Poster/ Report Structure and Formatting	10	10
4 Collaboration Assessment	4	4
Total	54	54

Everyone in the group may not receive the same score. A final score for an individual may be raised or reduced based on Peer Collaboration Assessments relative to "Meets Expectations".

2 Create Groups

Self-organize into groups of 2 to 4 people. Groups may include students from either 427 or 627 or both. Identify the members of the group as individual submission on Canvas.

3 Deliverables

3.1 Project Plan: Due Week 10 (10 Pts)

Deliver an approximately 2 page **plan**, in HTML or PDF format, addressing the following elements in order:

1. Project Title
2. Group Members and their course
3. Topic - 1 sentence
4. Questions of Interest: At least one for regression **and** one for classification.
5. Planned Approach: This includes framing the problem, the data, the methods, and the Intended Outcome.
    1. Literature Review: Identify at least one article or publication *per group member* relevant to the topic. It does not have to be a scholarly article but it shall provide context for the topic and questions of interest to help in framing the analysis.
    2. Data Assessment: this includes
        - Identification of the data set and its source.
        - Assessment of the data set and characteristics, e.g., number of observations and expected number and type of variables.
    3. Planned Methods: this includes
        - Identification of the expected methods for both regression and classification.

    - Identify planned approaches for assessing multicollinearity and model comparison as appropriate to the course. Any group with members from 627 must satisfy the 627-unique learning outcomes.
    - Identification of potential ethical concerns.
    - Risk Assessment and mitigation.

6. Deliverable: Option A (Poster and Presentation) or Option B (Report)

7. Schedule and Hours: A table for weeks 10-14 with the main activities, estimated hours, and expected outcomes for the week.

8. Group Member Responsibilities

9. Summary

The plan must be appropriately formatted, clear, and concise with minimal errors.

> **It's a Plan, not a Proposal.**
>
> This deliverable is called a "plan" as it is meant to reflect what the group intends to do based on your work in the initial steps in the data science life cycle. It is not a "proposal" about what you hope to do based on vague ideas.
>
> The requirements for this plan include collaborating on the elements of a typical program plan, i.e., Performance, Schedule, and Cost (hours).
>
> However, this is a plan so it is a baseline for your group activities. it is expected the group will need to adjust the plan before the end. That is okay and you do Not need to submit an updated plan.

3.2 Final Deliverable A or B: Due Class 14 (40 Pts)

3.2.1 Common Content Requirements (30 Pts)

Include the following content in either the poster or the report. Organize the content in a manner that best fits the format and intended engagement by the viewer/reader.

- 2 Pts: An "executive" summary that succinctly describes the most important findings and recommendations in a manner that entices the viewer/reader to engage more with your poster or report.
- 18 Pts: Use at least **six different statistical learning methods covered in the course** across the regression and classification questions, e.g., three for regression and three for classification.
    - You can choose from linear, logistic, polynomial regression with proper variable selection, linear or quadratic discriminant analysis, K-nearest neighbor classifier, jackknife, bootstrap, ridge regression, lasso, principal components regression, partial least squares, splines, regression and classification trees, artificial neural networks, support vector machines, clustering, or related methods.
    - Illustrate results with appropriate plots and diagrams.
    - (627-Only) Apply cross-validation techniques to find the optimal degree of flexibility, the best subset of predictors or the optimal tuning parameters.
    - Evaluate prediction performance of competing methods.

- 3 Pts: Identify the data source, describe the original data, and any challenges or choices in cleaning the data for analysis.
- 3 Pts: Identify the stakeholders in the analysis and its outcomes. Assess any ethical implications of the data (collection methods, sources, structure) or the choices made in the analysis (grouping, selection, etc.) or any other responsible data science concerns for implementation.
- 2 Pts: Summarize Findings
- 1 Pts: Offer recommendations for implementation or additional work
- 1 Pts: Identify key references.

3.2.2 Common Data Requirements

Use real-world data from any open data source of interest with a few restrictions.

- The data set shall include data from as recent as the last three years.
- You cannot have used the data set for another course.
- The data set shall not contain any personally identifiable information or protected health information about identifiable individuals.
- The data set shall not contain any classified or proprietary data.
- The data set shall be from a real-world primary source, not a machine learning-specific curated data set, i.e., not from Kaggle or other common repositories of competition or textbook data. The goal is for your data to be unique to your group and not repeat use a commonly-used data set analyzed by many.
- The data set shall have sufficient variables with sufficient variety to enable meaningful regression and classification.
- The data set shall have sufficient observations to enable meaningful findings and recommendations.

3.2.3 Structure and Formatting Requirements (10 Pts)

3.2.3.1 Option A: Presentation and Poster

Create a conference-style poster and deliver a group presentation about your project during the final class period using only the poster.

The poster shall cover the topics in the Content Requirements in Section 3.2.1.

- The presentation will be recorded.
- The poster shall be 24 inches high by 36 inches wide.
- It shall be created using Quarto, RMarkdown, or one of the Office 365, Google or Apple applications.
- The poster shall be suitable for public display. Ensure the text and all labels on graphs or tables are appropriately sized for reading from a distance of five feet.
- The analysis shall be reproducible using the submitted materials.
- **Submit a PDF or HTML of the poster** as well the original source file on Canvas. You do not need to print out the poster (but you can) as we will display it on the screen and on Zoom.
- **Submit a quarto, rmd, or python file containing all the code** and comments supporting the poster content.
- Submit a document (at most one page) describing the work performed by each group member.
- **Submit either a (zipped) file of the data** or a link to the raw data on canvas.

- Complete the release form if you want your poster available for display in DMTI.

3.2.3.2 Option B: Project Report

Create a project report discussing the data, approaches, analytical findings, and recommendations.

The report shall cover the topics in the Content Requirements in Section 3.2.1.

- The report shall be created using Quarto, RMarkdown, or one of the Office 365, Google or Apple applications.
- The report shall be formatted in a manner that is clear, concise, and appropriate for sharing with the public.
  - Use a theme to ensure consistent formatting across the document.
- The report body shall be between 8 and 10 pages long when printed to HTML or PDF.
  - Paragraphs may use 1 to 1 1/2 line spacing between lines.
  - Up to two columns may be used
  - Margins shall be 1 inch.
  - Use fonts of not less than 11 point for body text and not greater than 12.
- **Do not show any code or lengthy output in the body of the report (you can use code chunk options to control that)**. Summarize results in tables or graphs.
- All tables and figures shall have captions with the **interpretation** of the table or figure.
- Include a table of contents (does not count towards the 10 pages).
- Include an appendix (at most one page) describing the work performed by each group member.
- You may use appendices to include references and code. Appendices do not count as part of the 10 pages.
- The report shall be free of grammatical and spelling errors.
- The analysis shall be reproducible using the submitted materials.
- **Submit a PDF or HTML of the report as well the original source file on Canvas.**
- **Submit a file on Canvas containing all the code and comments** supporting the report if not included in the report source file.
- **Submit on Canvas either a (zipped) file of the data** or a link to the raw data.

3.3 Collaboration Assessment (4 Pts)

The project is a group effort where **every member is required to contribute to a meaningful level.**

- Use the plan to describe the proposed roles and responsibilities.
- Use the appendix for the document submissions (code file or project report) to describe actual roles, responsibilities, and accomplishments of each group member.
- **Complete the Collaboration Survey using the link in Canvas and the associated rubric.**
  - The standard for meeting the learning outcomes is an average score that rounds to “Meets Expectations” across all criteria and peers.
  - Individual scores for the project may be raised or reduced based on the direction and distance of one’s average score from peers relative to “Meets Expectations”.

Groups may solicit feedback or peer reviews from other class members without citation.

Extra Credit (2 points). If the group uses a shared GitHub repo, submit the URL with your code files and provide me access to the organization and repos.