JSC370 Final Project 2025

Final Project

Due Date and Grading Scheme

- **Due date April 30th, 2025** (no exceptions, this is the last day of the exam period).
- A 5-minute recorded presentation (10%).
- Written report downloadable from your project GitHub website. The report should be single spaced and a minimum length of 6-7 pages including figures and tables (90%).

Details

Learning Objective: The main objective is to apply the skills learned in JSC370 to explore a research question(s) or hypothesis(es) by analyzing and interpreting a dataset of your choice.

Narrative: Through this project you will showcase a data science project on your personal GitHub website. Using the dataset from your midterm, make sure you have formulated a clear and concise question(s)/hypothesis(es) to answer. You will apply the skills learned throughout the semester to answer this question.

Deliverables: 1) A 5-minute presentation where you walk through your website and main findings. Upload the video file or link to Quercus or your website; 2) A github website for your project. It will a summary of the project and interactive visualizations; 3) A written report with embedded tables and figures that is submitted as a PDF to a final project-specific github repository *and* as a downloadable link on your website. Please see the checklist below for additional details.

The report should have the following sections (elaborate from what was written in your midterm):

Introduction provide background on the topic and data, and a formulated question(s) and/or hypothesis(es). This does not require references to journal articles, but should include a narrative of the topic you studied.

Methods include how and where the data were acquired, how you cleaned and wrangled the data, what tools you used for data exploration, description of what models and/or statistical tests, and training/testing you used. Please include some statistical details on the model(s) (e.g. if you used random forest, briefly describe what is a random forest) and test statistics (e.g. R^2, RMSE, MAE).

Results provide final, publication ready tables and figures from your analysis, refer to your website if needed (interactive plots).

Conclusions and Summary interpret your findings in a bigger picture way and point out any limitations.

In your report, please do not include: code (so make sure `echo = FALSE`), unformatted output,dataset summaries (e.g. output from head(), str(), etc.).

Summarize model output in tables (e.g. use kable to show summary statistics, R^2, RMSE, coefficients, cross validation results, etc.) and figures (e.g. variable importance, scatterplot of test obs vs pred, etc.).

Checklist for Final Project

1. **Create a website** (HTML document and all the required files, including figures). It should feature:
   1. A brief description of the project.

   2. Interactive visualizations, each with a caption that describes the plot.
   3. A link to the PDF version of the report (i.e. a link to “Download the report.”).
   4. Your home page should feature no more than five interactive tables and figures. If you want to include more to showcase your data, please do so by adding extra pages to the website.

      The actual analysis should be included in the PDF report. **The PDF report can refer to interactive visualizations included in the website**.

2. **Upload** source code, data, website files, and PDF report, to your GitHub repository.
3. Make sure that the **website** actually **works**, i.e., figures and interactive visualizations are properly rendered when visiting the website.
4. **Have a README.md file** in the upper level of the repository. This file provides general information about the project, like title, brief description, etc.
5. The **README.md file links to the website**, e.g. https://username.github.io/JSC370-finalproject, .
6. Have a **“data” folder** with either the data or instructions about how to acquire it. If you are not providing a dataset, you should provide instructions in a README.md within that folder.
7. Your document should be **fully reproducible**, meaning you don’t have paths to files not shared in the repository. For example, if you are loading a dataset with “fread,” it should be something like “fread(“https://data.com/dataset.csv”),” so you are directly using data online or on your repo, or “fread(“data/dataset.csv”),” so the data needed has been shared on GitHub and is contained in the path “data/dataset.csv.”