MATH 244 Project Proposal Information

Project proposal due **Wednesday, March 5, 2025 by 11:59 PM.**

Introduction

The goal of this project is to demonstrate proficiency in data science techniques by conducting a novel analysis of a dataset of your own choosing or creation. The dataset may already exist, or you may collect your own data using a survey or by scraping the web. You will also get practice presenting your results.

Your project can be inferential in nature or can be focused more on prediction. Your project can involve regression or classification (or elements of both, but don't spread yourself too thin).

Brief Project Logistics

The final project will be done in a group of your choice.

The seven deliverables for the final project are (with associated percentage of the overall project grade):

1. A project proposal describing two datasets of interest (5%)
2. Exploratory Data Analysis (7.5%)
3. A preliminary analysis using Quarto (rough draft of the report)
    - Peer Review (15%)
4. Reproducible slides using Quarto + presentation (30%)
5. A written, reproducible report using Quarto detailing your analysis (30%)
6. Evaluation of your own and team members efforts on project (2.5%)
7. Your GitHub repo for the project (10%)

Late projects will not be accepted without prior approval.

Data Sources

To perform a successful analysis it is imperative that you choose a manageable dataset that can be analyzed using the tools we have learned in class. This means that the data should be readily accessible, not contain too many missing values, and be large enough so that multiple relationships can be explored. Your dataset must have at least 500 observations and at least ten variables (or my approval). The dataset should include a rich mix of categorical, discrete numeric, and continuous numeric data. If you have any doubts or are having trouble please reach out to me.

All analyses must be done in either R or Python (or parts in both) and your written work and analysis should be reproducible. This means that you must create a Quarto document that will render a final result. You can use scripts (for scraping data or defining useful functions) if you want, but you all analysis presented in your written report and slides should be completed in those files.

Reusing datasets from class: Do not reuse datasets or variations of datasets used in examples / homework. Also, you may not use data you analyzed in another course. If you use a dataset from the US government, then proceed with caution.

The resources below may be helpful for finding an interesting dataset but feel free to explore on your own.

- R Data Sources for Regression Analysis
- FiveThirtyEight data
- TidyTuesday
- World Health Organization
- The National Bureau of Economic Research
- International Monetary Fund
- General Social Survey
- United Nations Data
- United Nations Statistics Division
- U.K. Data
- U.S. Data
- U.S. Census Data
- European Statistics
- Statistics Canada
- Pew Research
- World Bank
- Election Studies
- Coperative Election Study
- IRIS (Repository for "research into languages, including first, second-, and beyond, and signed language learning, multilingualism, language education, language use, and language processing")
- Harvard Dataverse
- Redistricting Data Hub
- Integrated Public Use Microdata Series (IPUMS)
- Inter-university Consortium for Political and Social Research (ICPSR)
- CERN Open Data

Scraping

You are allowed to scrape data from the web with respect to forming a dataset, but before you do so, make sure that you are allowed to scrape the data! You can use `polite` or `robotstxt` to check scraping rules, but you should also see if a website's terms of use mention scraping in any way. If a website is scrapable, make sure that you scrape *politely* and respect any rules defined by a page's `robots.txt` page.

If you are having difficulty with scraping, please don't hesitate to ask.

You must have the data available for the proposal. This means that if you are scraping data, you must have it ready by the time you submit the proposal. Keep that in mind as you make a decision about what direction to take your project.

Git and Github

You will be working in one shared Github repo for this project to facilitate collaboration. This repo has the following structure to start off with:

```
|-data                          # folder in which data will be saved
|  |- README.md                 # short description of data folder
|  |- README.qmd                # .qmd file that creates README.md
|-figures                       # folder for containing figures saved to your repo
|  |- README.md                 # short description of figures folder
|-peer_review                   # folder for files associated with the peer review
|  |- README.md                 # short description of peer_review folder
|-presentation                  # folder for files associated with written report
|  |- README.md                 # short description of presentation folder
|-proposal                      # folder for files associated with proposal
|  |- README.md                 # short description of proposal folder
|-report                        # folder for files associated with written report
|  |- README.md                 # short description of report folder
|-scripts                       # folder for any scripts
|  |- README.md                 # short description of script folder
|-project.Rproj             # R Project file for final project (open first!)
|-project_info.html         # HTML version of instructions
|-README.md                     # Short description of repo
|-README.qmd                    # .qmd file that creates README.md
```

You can change this structure to a structure that makes more sense to your group but your repo should be organized; I do not want to see random files in random places. You do not have to use all of the pre-created folders (if you do not plan to use one or more, you can delete them and its associated `README.md` file); they exist purely as a tentatively suggested organizational superstructure.

Feel free to expand the `README.md` files of the subfolders if you wish; you can edit them in any text editor (including RStudio).

You are free to create any and all kinds of files necessary to complete your project. To make a new .qmd file, click File >>> New File >>> Quarto Document in RStudio. You can create .R and .py files (among many others) in a similar way.

Project components

Project Proposal

The first stage of the final project is the project proposal. The proposal is designed as a check to make sure you choose a dataset that allows you to perform an interesting analysis using the tools we have developed in class. Choose **two** substantially different datasets you are interested in analyzing. **For each**, identify the components below.

Clarifying Note: You can combine multiple datasets together to form one dataset. If you can use one *very* large dataset to answer multiple questions, then you can just describe this dataset in the proposal and present two project options. Basically, I want you to have at least two options for the final project in case one doesn't end up working out; I don't want you to fall behind with respect to work on the project.

For Each Data Set:

Introduction and Data

Identify the source of the data, when and how it was originally collected, the cases, and a general description of relevant variables.

As a reminder, your dataset must have at least 500 observations and at least ten variables (or my approval).

Research Question

In 1-2 paragraphs, describe what interests you about the chosen data and propose a research question that you would like to try to answer using it. Detail what you expect to find (you can think of these as your hypotheses, but please keep in mind that your question does not have to be inferential in nature).

Formatting

All parts of the preliminary analysis should be professionally formatted. For example, this means labeling plots and figures, and using tidyverse style.

You must comment your code!

You must suppress all warnings and messages.

You must label your chunks!

Style and format do count for this assignment, so please take the time to make sure everything looks good and your data and code are properly formatted.

Workflow and Organization

You should commit to your repo regularly as you work on your project, and you should keep your repo well organized.

You **MUST** have added your data (both datasets) to your repo.

Grading

Submit a pdf of your proposals (combined in **ONE (1)** document) by **Wednesday, March 6, 2025 by 11:59 PM.**. I will provide feedback on your proposal and help you decide which dataset you should use for your final project.

The project proposal will be graded as follows:

Total	**30 pts**
Research questions (4 pts for each research question)	8 pts
Data (4 pts for each research question)	8 pts
Formatting	5 pts
Workflow and Organization	4 pts