Project #1 (150 pts)

SDS 192— Introduction to Data Science, Fall 2025

Background

In this project, you will be analyzing a particular dataset and developing meaningful and clear visualizations that detail various trends and patterns within the data that inform the audience about particular issues. This will be framed within a workflow, in which you will import the data, write code to analyze and create plots, and format your files so that they are reproducible. By the end of this project, you will be able to:

- **Navigate** different forms of data documentation.
- **Recognize** differences in variable types and how they get assigned to certain R objects.
- **Examine** trends and patterns within the data through exploratory plotting in R.
- **Create** clear and informative plots that highlight particular patterns in the data.
- **Summarize** findings and ethical issues within data and its resources through a short video and report.

The project will contain the following parts:

- **Project Proposal** (5 pts; due Wednesday, 10/8 at 11:59pm)
- **Video Summary** (25 pts; due Wednesday, 10/22 at 11:59pm)
- **Project Repository** (120 pts; due Wednesday, 10/22 at 11:59pm)

Working Time

At least one Friday class and parts of some lectures will be dedicated to working on the project. Aside from these times, you will be tasked with working on this project outside of class.

As questions/concerns arise during the project, ask your instructor for assistance and guidance. There will be some hints and pointers that I can give to assist your group through the project.

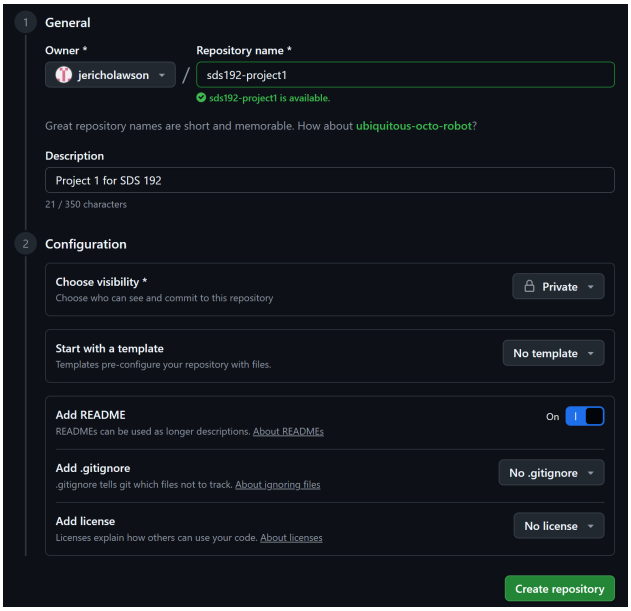
Instructions

1. **Choose a dataset.** You will choose one of the following datasets to base your project on. The datasets are the following:

   - Mental Health Care in the Last 4 Weeks
   - Chicago Microlending Institute (CMI) Microloans
   - MTA Daily Ridership Data: 2020 - 2025
   - Emergency Department Volume and Capacity

2. **Create a repository on GitHub.** You will need to go to your GitHub profile. On your profile, you will create a new repository by clicking "New" at the top-right hand corner of the screen.

   From there, you will put in the following specifications for your new repository:



   Then, create the repository.

3. **Create new project in RStudio.** You will now create a new project in RStudio based around your new repository. To do so, go to File → New Project → Version Control → Git. Copy the new repository URL and paste it into the first field. Name your project "project1" and place the new repository into a new folder on your laptop. Then, click "Create Project".

4. **Import data into R.** You will need to download the .csv file from the data link above and import the data into R. Additionally, some of the datasets may contain missing observations, which will need to be omitted before you continue this project (depending on which variables you are looking at). More on this later. You will find the following pieces of code to be helpful:

   - `read.csv(filepath/name, header = TRUE)`
   - `na.omit(object)`

   You will need to place the .csv file you are referencing into a "data" folder in your project folder. Any references to your .csv file from code should have a file path of "data/yourData.csv".

5. **Get to know the data set.** This will involve a combination of looking at the metadata and a snippet of the data itself. Consider the following questions as you examine the data:

   - Who is the creator/maintainer of the data?
   - What do the observations represent? How many variables are there?
   - How many observations and variables are in the data set?
   - What are the variables in there? What are the possible values in each variable?
   - Identify the response or variable of interest in this data set.
     - This should be numerical.
   - Are there missing values in the data set? If so, how many and is there a reason why this is the case?
     - If it appears to be random, you can omit the missing observations. If not, you will need to subset your data.

6. **Determine which variables to analyze.** You will need to identify the main (continuous) variable of interest in your dataset. Once you do so, pick five more variables you want to analyze alongside the variable of interest. For these five variables, you must follow these rules:

   - At least one variable must be categorical
   - At least one variable must be continuous

   Note that you may need to discretize a variable or two, which involves making a new variable based on other variables.

7. **Prepare the data set for analysis.** Some data wrangling may need to be done, particularly if you are looking at certain subsets of the data or missing data exists. Your code should properly subset the data so that you are able to complete the data visualizations.

8. **Create the data visualizations.** You will need to create four visualizations, which should (1) facilitate a discussion of the variable of interest based on its own distribution/subset or when compared to other variables and (2) be concise, informative, and adhere to the principles of data visualization. For grading purposes, you should adhere to the following:

    - At least three types of graphs need to be represented by your visualizations (e.g. bar, scatter, histogram, boxplot)
    - The main variable of interest should be represented in some capacity in at least two plots. The other variables just need to be used at least once.
    - Each plot should adhere to the five principles of context (see Lecture #4 for details)
    - Each aesthetic or stylistic choice should add to the interpretability of the plot.

9. **Analyze each of the visualizations.** For each of the plots, you will consider the following questions:

    - What can be said about the data? Trends? Patterns? Characteristics?
    - What does this say about the situation at hand? How does this further your understanding of the topic?

    You will also consider the pitfalls of your analysis and data as a whole. This will include exposing potential biases, exclusions, and data collection procedures.

Components

Project Proposal* (*5 pts; due Wednesday, 10/8 at 11:59pm via Google Forms)

For this component, you will need to fill out a Google form that answers which dataset you want to use and what problem/question you want to try to solve. This question should address a certain variable the author wants to explore in relation to other variables.

This task must be done by the due date mentioned above, using the link here or on Moodle. A project late day may be used here if needed.

Recording* (*25 pts; due Wednesday, 10/22 at 11:59pm on Moodle)

The recording will consist of you walking through each of your four data visualizations. In this 2-3 minute recording, you will talk briefly about the dataset, present the visualizations, and summarize what you found.

The best (and preferable) way to creating this recording will be to record yourself in a Zoom meeting while sharing a slide deck of your plots and brief bullet points. Once you've created the recording, you can choose to turn in the following via Moodle:

- A unlisted YouTube link
- A .mp4 recording straight from Zoom

The slides/visuals you present should be almost exclusively the plots you've generated. If you do have some prose on your slides, it should be at most one bullet point for a slide if accompanied with a visual and three bullet points other wise.

You will be graded based on the proficiency of the following criteria:

1. (5 pts) Delivery: points are talked about clearly; done within 2-3 minutes
2. (5 pts) Organization: slides are reduced to visualizations and minimal bullet-points; easy for the audience to interpret
3. (15 pts) Content: recording goes through the following:
    - Context, with a brief introduction into the data set and problem
    - Visualizations, with succinct and correct interpretations
    - Analysis, with proper mention of stakeholders and impact

Submit the recording on Moodle under "Project #1: Recording" by Wednesday, 10/15 at 11:59pm.

This task must be done by the due date mentioned above on Moodle. A project late day may be used here if needed.

Project Repository (120 pts; due Wednesday, 10/22 at 11:59pm via GitHub)

The project repository will contain all of the work completed. This includes all code, the dataset, a README.md file, and a 2-3 page report of all work (in a report.md file). As such, the final version of your repository will have the following elements:

1. `data` folder
    - The .csv file of the data set you chose
    - The .csv file of the revised data set you used
2. `R` folder
    - At least one .qmd file that includes all code used to complete the data cleaning and visualizations
3. `plots` folder
    - Place all generated plots here.
4. `README.md` file
5. `Report` file

A detailed account of what goes in each of the parts and how it should look is seen next.

1. Data Folder (5 pts)

You will have a folder named "data" that contains 1) the .csv file of the dataset you initially chose and 2) the .csv file of the dataset that contains relevant observations and variables you chose.

The updated data set should contain no missing observations and have only the variables you analyzed.

2. R Folder (5 pts)

You will have a folder named "R" that contains all code necessary to run through data cleaning and visualization. At a minimum, you just need one .qmd file in here, meaning

that all code can be placed into a single file. However, if you feel the need to divide your code up, you can do so here. Your code files should be named accordingly.

R code needs to be organized properly, have proper variable names, and contain clean and simplified code. To this regard, any written code should be the product of your own and not through any use of generative AI. Penalties from the syllabus will apply here.

3. README File (5 pts)

In many repositories, the role of the README file is to provide the viewer a basic summary of what the repository is for and how it is structured. You can find more details about what a README file entails at the following link.

In your README file, you will need a name, a one-paragraph description, an author, and an overview of the structure. This can be done in a .md file (seen in RStudio), where ## creates headers in the file.

4. Plots (40 pts)

In the `plots` folder, you will place all four generated visualizations here. Keep the file names to the plots simple, such as naming them `plot1.png`.

For each plot, you will be graded on the following:

- (5 pts) Appropriate variables plotted; adheres to basic principles in guideline #8
- (5 pts) Context is provided in clear and concise manner; adhers to basic principles in guideline #8

5. Report (65 pts)

This report will contain all of the analysis and discussion related to the question you are trying to answer with the dataset.

Formatting-wise, your report should:

- (3 pts) Contain 2-3 pages (with no title, abstract, or table of contents required)
- (2 pts) Be organized into multiple, sensible sections

The report should have the following:

- (10 pts) Introduction

  - An introduction into the chosen dataset, including the dimensions, variables, and information of the dataset; adhering to questions mentioned in guideline #5.
  - The question you're trying to answer.
- (10 pts each) Analysis (of each plot)
  - Should include your analysis of the plot, adhering to guideline #9.
- (10 pts) Conclusion
  - Summarize your findings, examine pitfalls in data and process.

You will need to share this repository to me (@jericholawson) on GitHub. This task must be done by the due date mentioned above. A project late day may be used here if needed.

Tips

- **Start early and plan accordingly.** This will make it easier to space out the workload, especially since we're in the middle of the semester.
- **Talk to Jericho about pointers or concerns.** Feel free to email me or come to office hours regarding some recommendations on your project.
- **Feel free to use outside sources discussed in class, but you must internalize and explain what you're doing.** For this project, using outside sources for consultation is fine, including forums and cheat-sheets. Using large-language models to analyze or create plots is prohibited. Violations will be subjected to potential consequences, as outlined in the syllabus.