MATH 271 – Project 1

You will be working in randomly selected groups of 2:

1. Kelli McCarty & Augustus Coffey
2. Maria Pla Prahl & Justin Santiago
3. Christian Castro & Joshua Kralewski
4. Tomoko Sakishima & Peyton Taylor
5. Dylan Mattheus & Christian Morgan
6. Bence Brian & Marlon Simmons
7. Nolan Brophy & Acadia Clark

The goal of this project is to apply your knowledge of data wrangling and visualization to clean up, explore and find insights in a real-world dataset.

1. Finding and Importing a Dataset

a) Search online for a not-so-clean dataset.
b) Import the data into R.
c) Briefly describe the dataset: Where did you find it? What does it contain?

2. Data Cleaning and Wrangling

a) Explore the structure of the data and data types.
b) Use dplyr functions to clean up the data by identifying and fixing issues such as:
    - Ensure data has a tidy format.
    - Missing values.
    - Fix/create columns if necessary.
    - Filter/aggregate data if necessary.
    - Ensure variables are stored in the correct data type.
c) Clearly document each step explaining what was changed and why.

3. Distributions and Summary Statistics

a) Identify 2-3 numeric variables in the dataset and compute measures of:
    - Central tendency (e.g., mean, median, mode).
    - Spread (e.g., standard deviation, standard error, range).
b) Perform group-level summary statistics, practicing using the group_by() and summarize() functions.
c) Create visualizations to explore distributions (e.g., histograms, density plots).
d) Write a short narrative explaining what these distributions reveal about the data.

4. Developing Questions and Making Visualizations

a) Develop 2 research questions that could be addressed using the dataset. Ex:
    Are there differences in a numeric variable across categories?
    Is there a relationship between two numeric variables?

b) Use appropriate visualizations to help answer these questions (e.g., boxplots, scatter plots, bar plots, stacked bar plots).
c) Apply your ggplo2 skills by customizing your figure using colors, themes, axis titles, facets, etc.
d) Write a brief interpretation of each figure, explaining what insights it provides. (Do not run any actual statistical test to answer your question, only use data summaries and visualizations).

5. Format and Submission.

a) Format a report using R Markdown:
    - Include headers, bold text, and other formatting options we covered in class.
b) Ensure code is well-commented, and explanations are provided for each step.
c) Work with your partner to find the data and develop the code, but please write your own personal interpretations and explanations.
d) Submit both the .Rmd file and the knitted HTML report.

6. Presentation.

a) Each group will give a short 5-min talk to the class on 02/24.