

# MTH208a: Project Guidelines

## Guidelines

We will continue to use GitHub Classroom for the Group Project. A Project repository has been created and will be sent via email to the class.

This repository will contain **ALL** information related to your project. This includes:

- All your R codes for web scraping
- All data obtained from web scraping
- Your project report files (including support documents)
- Your project presentation files

Only one repository will be made for each group. More details on how such repositories work, will be shared in class.

## Part 1

**Due date:** October 10th

The goal is to identify an application of interest to your group. Think about various sources of data and how could you collect it using web-scraping. Then write a script to collect the data. By October 10th, you are expected to:

1. Identify the dataset you want to collect. Be creative and come up with interesting applications.
2. Check to make sure you have permissions to obtain the data. APIs may be required, and it is your groups' responsibility to figure out how the scraping will work.
3. Clean and save the dataset in an R compatible format.
4. Pose a list of questions that your group thinks can be answered with the dataset.

## Part 2

Once your data has been scraped and cleaned, you are expected to produce the following outputs:

1. **R Shiny App:** An R Shiny app that helps present and summarize the data. Think carefully about which plots and widgets to include in the shiny and app. Make it interactive and interesting.
2. **R Markdown/Quarto report:** (due on 20th November at 5pm) The main outcome will be a report of the project. I expect the report to contain the following information:
  - a. Data: Describe was the dataset is, what variables does it have, how many observations.
  - b. Obtaining the data: Describe exactly how you obtained the data. Which parts were scraped, which parts were obtained via csv files.
  - c. Identify any biases in the data: Are there any sources of bias or misinformation that you can identify in the dataset?
  - d. Interesting questions to ask from the data: here, I want you to list the potential questions the data can help us answer. Do not answer the questions, just pose the questions here.
  - e. Important visualizations: Present all plots that may potentially help answer the questions posed above? Be very careful in the quality of plots produced. Make sure plots are readable and are able to tell a story.
  - f. Final conclusions: Write up some final thoughts and conclusions for the project.
  - g. References: List all resources you used for the project. This includes links to data sources.
3. **Presentation:** Each group will give a 10 minute **executive presentation** of their project. The first 5 minutes will be dedicated to explaining the dataset and the questions posed. The last 5 minutes will be for demonstrating the R shiny app. These presentations are scheduled for the last week of class.

Summary of breakdown of marks for the project (out of 100)

1. Report - 50 marks
2. Shiny App - 20 marks
3. Presentation - 20 marks
4. Code organization in GitHub - 10 marks