MATH224 - Data Analysis Project

Part I: Exploring the Dataset & Identifying Research Questions
Total 30 Points Possible
Due Date: 03/06/2022

Instructions: Use the R Markdown report outline template your created in RStudio cloud during the lab to write a progress report that addresses the following tasks. Knit your R Markdown code as a **PDF** report and submit your report (which should contain all code lines and answers to the questions below) under "Data Analysis Project - Part I" in Blackboard by the due date.

1. **(3 points)** Explore the dataset assigned to you for your project to answer the following questions. You can use the `glimpse` function from the `tidyverse` package to peek into the dataset and find answers to these questions.

   a. How many cases (instances/rows) are in your dataset?
   b. How many variables (attributes/columns) are in your dataset?
   c. Does your dataset contain missing values? Which variables contain missing values?

2. Identify **at least two** potential research questions that you plan to answer using your dataset. It is **strongly recommended** that you define at least 1 question that can be answered using data visualization and correlations, and at least 1 question that needs some sort of predictive modeling:

   a. **(6 points)** Write each research question and the corresponding hypothesis(es) to be tested.
   b. **(2 points)** For each research question, identify and list the relevant variables that will be used in the analysis. So, the expected format here is like this:
      - Question 1: *list of the relevant variables for question 1*
      - Question 2: *list of the relevant variables for question 2*
      - Question 3 (if applicable): *list of the relevant variables for question 3*
   c. **(2 points)** Identify the main response (also known as dependent) variable in the data.
   d. **(5 points)** Compute and report summary statistics (e.g., mean, sd, and five number summary) for summarizing the distribution of the response variable identified in (c).
   e. **(6 points)** Create at least three graphs displaying the distributions of the response variable identified in (c) and other relevant variables identified in (b). For example, if the variable is categorical, report a bar chart, while for quantitative variables, report histogram, dotplot or boxplot.
   f. **(6 points)** Create at least three graphs displaying the association between the response variable in (c) and three explanatory variables from (b). For example, you can use scatter plot to show relationship between response variable and quantitative explanatory variables, while side-by-side boxplots can be used to show association between response variable and categorical explanatory variables.