TASK 1 DUE DATE: MONDAY, FEBRUARY 17, 2025 by 11:59 PM on Gradescope
DUE DATE: FRIDAY, MARCH 14, 2025 by 11:59 PM on Gradescope
LATE DUE DATE: SUNDAY, MARCH 16, 2025 by 11:59 PM on Gradescope with penalty of 10 percentage points per day late

For this project, you will be given some data regarding a sample of gym users. You will conduct various types of analysis and summarize your results in a report. You should create your report in RMarkdown. Your report should read like a professional report that could be read by someone without an extensive statistics background. Answers should be given in complete sentences and paragraphs. Graphs should be appropriately labeled. When statistics are computed, include an explanation of what information can be determined from those values. For the report itself, hide chunks of code where appropriate to make your output cleaner. Submit both your Markdown file and the knitted code as a **PDF** document. (You may create an HTML file, and when you open it, then Print to PDF.)

Submit both files to Gradescope.

When submitting Project 1, make sure you do the following things:

1. Use only methods used in this course. Code or responses from outside sources, including AI, will be considered Academically Dishonest. Do not use AI to improve your grammar or spelling.
2. Any outside sources should be properly cited. Geeks for Geeks and StackExchange (and Chegg, CourseHero, AI etc) are not acceptable sources.
3. Clearly identify where each Task starts in your final document.
4. Knit your code to create a HTML file. Convert the HTML file to a PDF. Submit this PDF.
5. Submit your .Rmd code (what you used to create the HTML file).
6. All values and graphs, not just code, needs to be provided in the PDF document. Values as comments in the code will be ignored.
7. All values and code related to a particular task are in the region identified with that task in the code or final document, and not elsewhere. For example, if you need to create a table in Task 2, we don't want to have to look in Task 4 to find the table.
8. All text and explanation written should be the "white space" in the .RMD document, and not as comments in the code.

   For example,
   ```{r}
   ## DISCUSSION SHOULD NOT BE HERE.
   ```

   Discussion should be HERE

Part of these projects is to have you practice writing a report. If you were to hand this document to your manager at work, they would want something that looks like a report, not just a bunch of code without any output.

Unless the Task asks you to code something, we will be looking in your PDF file for your answer, not the RMD file.

Dataset Information:
The data set *gym.csv* contains data for a number of gym members. There is a header in the dataset. The variables are:

- **Age** - Age of the gym member.
- **Weight** - Member's weight in kilograms.
- **Height** - Member's Height in kilograms.
- **Max_BPM** -Maximum heart rate (beats per minute) during workout sessions.
- **Avg_BPM** - Average heart rate (beats per minute) during workout sessions.
- **Resting_BPM** -Heart rate at rest before workout.
- **Session_Duration** - Duration of each workout session in hours.
- **Calories_Burned** - Total calories burned during each session.
- **Workout_Type** - Type of workout performed (e.g., Cardio, Strength, Yoga, HIIT).
- **Fat_Percentage** - Body fat percentage of the member.
- **Water_Intake** - Daily water intake during workouts.
- **Workout_Frequency** - Number of workout sessions per week.
- **Experience_Level** - Level of experience, from beginner (1) to expert (3).
- **BMI** - Body Mass Index
- **Gender** - 1 if the member identifies as gender 1, 0 if they do not.

Task to Get Started:

Task 1) Watch the videos on Blackboard about R Markdown under the "RMarkdown" content area. Then create your own R Markdown file for this project. In your file:

- Remove the default code from the created document ("This is an Rmarkdown document...", summary(cars), Including Plots, etc).
- Create headers for the additional tasks. The header should be bold and use a font size that is substantially larger than the standard font. The header should be separated from the written content.
- Suppose we were creating a scatterplot and added a vertical line to it, like in the below code:

```
plot(x,y)
abline(x = 2)
```

  What do we need to add to the code in R Markdown to make sure the code runs and the file gets knitted?
- In a sentence or two, explain whether you are going to knit directly to a PDF file or if you are going knit to an .html file. If you are creating an .html file, explain how you will create a PDF of your output.
- Explain the difference between your .PDF file and your .RMD file.

Tasks on the Statistical Analysis of the Data:

Task 2) Convert `Workout_Type`, `Experience_Level` and `Gender` to factors in the data frame. Make the factor ordered where appropriate.

Task 3) A quantitative variable that is of interest is `Avg_BPM`. Write a paragraph summarizing and describing the variable. The explanations should be such that a person with limited statistical knowledge can understand.

- You should initially check for NA values.
- Include relevant graphs (minimally a histogram and a boxplot). Be sure your graphs are adequately labeled.
- Include detailed descriptions of the graphs, including shape, presence of outliers, etc.
- Include appropriate descriptive statistics (minimally measuring the center and spread).
- Explain what those statistics describe about the data.

Task 4) A categorical variable that is of interest is `Workout_Type`. Write a paragraph summarizing and describing the variable. The explanations should be such that a person with limited statistical knowledge can understand.

- Include at least one relevant graph. Be sure your graph is adequately labeled.
- Include a detailed description of the graph.
- Include appropriate descriptive statistics (such as frequency from a table).
- Explain what those statistics describe about the data.

Task 5) One would like to know if `Workout_Type` varies by `Experience_Level`. Write a paragraph explaining whether or not you think this variable varies by `Experience_Level` and detailing how you reached your conclusion.

- Include graphs (minimally side by side barplots) for each group that support your conclusions. Be sure your graphs are adequately labeled.
- Compute appropriate conditional distributions and explain how these distributions support your conclusion.

Task 6) A variable of interest is `Calories_Burned`. It would be helpful to know if this variable is normally distributed. Write a paragraph explaining why one should or should not assume the variable is normally distributed. Explain in a way that a person with limited statistical knowledge would understand.

- Include all necessary graphs (at least one of a histogram or a Q-Q Plot with reference line). Explain the implications of the graph(s) in regards to normality of the variable.
- Calculate the skew and kurtosis. Explain the implications of the values reported in regards to the normality of the variable.
- Perform the Shapiro-Wilk Test. State your hypotheses, your decision, and conclusion in regards to the normality of the variable. Use a 1% significance level.

Task 7) We want to divide the members into groups based on their `BMI` using the definitions provided by the CDC.

- Create a new ordered factor called `bmi_class` that takes on the value of "Underweight" if `BMI` is less than 18.5, "Healthy" if `BMI` is at least 18.5 and less than 25, "Overweight" if `BMI` is at least 25 and less than 30, and "Obese" if the `BMI` is at least 30. Please provide the code in your project, but you do not need to comment on it.
- Write a paragraph explaining whether or not you think `Water_Intake` varies by `bmi_class` and detail how you reached your conclusion.
  - Include graphs (minimally side by side boxplots) for each group that support your conclusions. Be sure your graph is adequately labeled.
  - Include appropriate summary statistics (minimally measuring the center and spread) for each group that support your conclusions.

Task 8) One wants to test whether the population mean `Calories_Burned` greater than 890.

- Conduct a significance test at a 2% significance level to determine if this is the case and describe your results in a paragraph. Your paragraph should include hypotheses tested, the p-value, the decision you made, and your conclusion. Your conclusion should be understandable to a person with limited statistical knowledge.
- Include an appropriate confidence interval or confidence bound and explain how that supports your conclusion.

Task 9) Determine if there is enough evidence the mean of `Session_Duration` is different for those who identify as gender 1 and those who do not.

- Conduct a significance test at a 3% significance level to determine if the mean of `Session_Duration` varies by gender. Describe your results in a paragraph. Your paragraph should include hypotheses tested, the p-value, the decision you made, and your conclusion. Your conclusion should be understandable to a person with limited statistical knowledge. *If you are making any assumptions, be sure to include any tests to validate those assumptions (including hypotheses, p-values, decisions, and conclusions), also at a 3% significance level.*
- Include an appropriate confidence interval or confidence bound and explain how that supports your conclusion.