MATH224 - Data Analysis Project

Part II: Further Data Exploration and Inference
Total 30 Points Possible
Due Date: 04/15/2022

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 II" in Blackboard by the due date.

In Part I of the project, understood the dataset, identified research questions and performed exploratory data analysis. In this part of the project, you will focus on the following tasks:

1. **(5 points)** Repeat task (d) from Part I: Compute and report summary statistics (e.g., mean, standard deviation, and five number summary) for summarizing the distribution of the response variable identified in Part I (c).
2. Construct confidence interval for estimating the population mean of the main response variable (e.g., annual income, house price, body weight, etc.).
    a. **(2.5 points)** Check the assumptions needed in order to perform the procedure of constructing a confidence interval for the population mean of response variable. For example, make a histogram for the response variable and check if it is approximately normally distributed. Also report your sample size to show that the central limit theorem approximation is valid.
    b. **(5 points)** Compute and report the 95% confidence interval for the population mean of the main response variable. Use R function `t_test()` to compute the confidence interval. Show your code and output.
    c. **(2.5 points)** Interpret the confidence interval obtained in the context of your dataset.
3. **(4 points)** Repeat task (f) from Part I: create at **least two graphs** displaying the association between the response variable you identified in Part I (c) and three explanatory variables that you think might correlate with the response variable. For example, you can use scatter plots to show relationship between response variable and quantitative explanatory variables (e.g., house price and lot size in square foot), while side-by-side boxplots can be used to show association between response variable and categorical explanatory variables (e.g., house price and whether it has central air conditioning).
4. **(4 points)** Compute summary statistics (mean and standard deviation) to summarize the response variable in your data by groups defined by the levels of one categorical variable from your dataset. For example, you can compare the mean body weight for men versus women. Comment on the patterns seen from those summaries.
5. **(7 points)** Construct confidence interval for estimating the difference in the population mean of the main response variable across the levels of one categorical explanatory variable (e.g., the difference in mean body weight between men versus women).
    a. **(5 points)** Compute and report the 90% confidence interval for the difference in population means. Use R function `t_test()` to compute the confidence interval. Show your code and output.
    b. **(2 points)** Interpret the confidence interval obtained in the context of your dataset.