Intro to Data & Probability - Project Rubric

IMPORTANT: If the analysis is completed using software other than R, or not written up using R Markdown, the project should receive a 0 regardless of its content.

Part 1: Data (3 points)

* 2 pt for correct reasoning for generabizability – Answer should discuss whether random sampling was used. Learners might discuss any reservations, those should be well justified.
* 1 pt for correct reasoning for causality – Answer should discuss whether random assignment was used.

Part 2: Research questions (3 points)

* Should be phrased in a non-causal way (1 pt)
* Should be well defined / not vague (1 pt)
* Is clear why this is of interest to the author / audience (1 pt)

Part 3: EDA (10 points)

* 3 pts for plots
  + Plots should address the research questions (1 pt)
  + Plots should be constructed correctly (1 pt)
  + Plots should be formatted well – size not too large, not too small, etc. (1 pt)
* 3 pts for summary statistics
  + Summary statistics should address the research questions (1 pt)
  + Summary statistics should be calculated correctly (1 pt)
  + Summary statistics should be formatted well – not taking up pages and pages, etc. (1 pt)
* 4 pts for narrative
  + Each plot and/or R output should be accompanied by a narrative (1 pt)
  + Narrative should interpret the visuals / R output correctly (1 pts)
  + Narrative should address the research question (2 pts)

Inference (28 points)

Statistical inference via hypothesis testing and/or confidence interval.

* State hypotheses
* Check conditions
* State the method(s) to be used and why and how
* Perform inference
* Interpret results

If applicable, state whether results from various methods agree It is your responsibility to figure out the appropriate methodology. What techniques you use to conduct inference will depend on the type of data you’re using, and your sample size. All of you should conduct at least a hypothesis test, and report the associated p-value and the conclusion. Those of you comparing two means, two medians, or two proportions should also calculate a confidence interval for the parameter of interest. Those of you working with categorical variables with more than two levels will need to use methods like ANOVA and chi-square testing for which there is no associated confidence interval, and that’s ok. If your data fails some conditions and you can’t use a theoretical method, then you should use an appropriate simulation based method.

* If you can use both theoretical and simulation based methods, then choose one and stick with it. You don’t have to do both. However if you can’t use both, then you need to decide which is appropriate.
* If you can do both a hypothesis test and a confidence interval, do both, and comment on agreement of the results from the two methods. However if your variables do not lend themselves to a confidence interval, that’s ok.
* It’s essential to make sure the method you’re using is appropriate for the dataset and the research question you’re working with.
* Hypotheses stated clearly and match research question (4 pts)
* Conditions checked in context of data (4 pts)
* Appropriate methods stated and described (4 pts)
* Correct code and output (6 pts)
* Correct interpretations and conclusions (7 pts)
* Reasoning for why CI is/is not also included (3 pts)

Overall (6 points)

* Uploaded HTML document resulting from the Rmd template: 1 pt
* Organization: 1 pts
* Readability of the text: 2 pts
* Readability of the code: 2 pts