STA 106 Exam I Project, due
Tuesday, Feb 6^{th} onto Gradescope

Read the following instructions carefully:

- **You may work in your assigned group.**
- **You are not allowed to discuss the questions with anyone other than the instructor or TA and your group mate.**
- **Any outside help beyond that from the instructor or TA is considered plagiarism. This including asking a tutor, your classmates (for example, comparing answers), posting the questions to homework help sites, etc. Should we believe you have sought outside help, you will be reported to the Student Judicial Affairs office.**
- **You are allowed to use or modify your previous functions, or the instructors functions that are posted online.**
- **Do not share answers, or specific values for calculations, particularly on Piazza**.
- You may ask clarifying questions about code and general approach on Piazza, but do not give away any numerical answers. If you are concerned you may be giving something away, email me or the TA's directly.

All teams will pick exactly one of the following datasets to explore with Single Factor ANOVA.

Question 1

The goal of this experiment was to see if different nests for sparrows on Kent Island attracted different size sparrows. The data is in the file `sparrow.csv`, which has the following columns:

Column 1: `Treatment` - What type of nest, with `control` (not manipulated), `enlarged` (manipulated to be a larger nest than normal), or `reduced` (manipulated to be a smaller next than normal).

Column 2: `Weight` - The weight of the bird in grams

In addition to hypothesis testing, we are interested in the following confidence intervals:

- A confidence interval for the nest that tends to have the largest sparrow.
- A confidence interval comparing the control nest to the enlarged nest.
- A confidence interval comparing the control nest to the reduced nest.

Question 2

The data set contains information on 76 people who undertook one of three diets (referred to as diet A, B and C). The aim of the study was to see which diet was best for losing weight. The data is in the file `loseit.csv`, with the following columns.

Column 1: `Diet` - Which diet they were on, with values `A, B, C`.

Column 2: `Loss` - The difference (in pounds) of their weight at the beginning of the program, and their weight after 6 months. A positive number therefore suggests they lost weight, while a negative suggests they gained weight.

In addition to hypothesis testing, we are interested in all pairwise confidence intervals for differences in means.

Question 3

This data was collected as part of the SENIC project, and the overall goal is to assess if the length of stay for patients differs between geological regions. The data is found in the file `senic.csv`, with the following columns:

Column 1: `Length`: The average length of stay for patients at this hospital (in days).

Column 2: `Region`: The region the hospital was in, with values `NC` (North Central), `NE` (North East), `W` (West), and `S` (South)

In addition to hypothesis testing, we are interested in all pairwise confidence intervals for differences in means, and if you believe any regions could be combined.

The Report Format

You or your team will turn in a short report. This means you should write in full sentences, and have the following sections for each question, while being **as specific as you can** about your results. **There should not be any "copy and pasted" R code in this report. You must format the results you get from R.**

I. Introduction. State the question you are trying to answer, why it is a question of interest (why might we be interested in the answer), and what approach you are going to take (just the name of the approach).

II. Summary of your data. This should include things like plots (histograms, boxplots) including the interpretation of the plots, and summary values such as sample means and standard deviations. You should have an idea about the trend of the data from this section.

III. Diagnostics. You should discuss your assumptions here, and if you believe they are violated. Perform diagnostics for the model. Remove outliers if necessary. You do not need to do transformation of variables.

IV. Analysis. Report back the model fit, confidence intervals, test-statistic/s, and p-value/s, nulls and alternatives, power calculations, etc. You may use tables here, but be sure that you organize your work. Remember to write your results in full sentences where possible.

V. Interpretation. State your conclusion, and what inference you may draw from your corresponding tests or confidence intervals. These should all be in terms of your problem.

VI. Conclusion. Summarize briefly your findings. Here you do not have to re-iterate your numeric values, but summarize all relevant conclusions.

Details

Your report should be the following format:

i. Typed.

ii. A title page including your name/s, the name of the class, and the name of your instructor (me).

iii. Double-sided pages.

iv. An appendix of your R code used to produce the results. **Do not include in R code in the body of your report.**

Feel free to make your cover page "unique" so that it is easy to find when I hand them back.

Notice: your project will be graded as a group effort. This means that you are responsible for your own work, and your partner's work. I will not assign two different grades to one project.

Notice: In the submission process, one person (team leader preferably) will submit the team work "included" all group members.

Important Note: Let us know in the front page if there is a conflict of interest. For example, if a group member did not fully participate.

Grading Rubric

I. (7 points) Introduction

II. (7 points) Summary of your data

III. (7 points) Diagnostic

IV. (10 points) Analysis

V. (9 points) Interpretation

VI. (6 points) Conclusion

VII. (5 points) formatting