STA 106 Winter 2021
Take Home Project II - Due Thursday, June 2^{rd} onto Grade scope

- **You may work in a group of two or three.**
- **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 R output should be **formatted**. You should **not** copy and paste directly from the R console.

You (or your group) should pick one question from each topic, for a total of two questions

Topic I: Transformation of Variables

Question 1 - Helicopter.csv

This data is over the period of a year, and is interested in if different times of day see different amount of helicopters requested for a sheriff's office.

Column 1: `Count`: The number of times a helicopter was called to an emergency in one year.

Column 2: `Shift`: The shift, with `I` (between 2AM and 8AM), `II` (between 8AM and 2PM), `III` (between 2PM to 8PM), and `IV` (between 8PM to 2AM).

Question 2 - NewHawk.csv

This data is a random sample of hawks from the larger dataset used in Homework 2. It has columns

Column 1: `Wing`: Length (in mm) of primary wing feather from tip to wrist it attaches to

Column 2: `Species`: CH=Cooper's, RT=Red-tailed, SS=Sharp-Shinned

Topic II: Two Factor ANOVA

Question 1 - Salary.csv

This data is taken from a random sample of "technology workers" from Seattle and San Francisco. *Note, this data was scraped from the web, but from what website and what year was not disclosed.*

Column 1: `Annual`: The subjects annual salary in thousands of dollars.

Column 2: `Prof`: The subjects title, with values `DS` ("Data Scientist"), `SE` ("Software Engineer"), and `BE` ("Bioinformatics Engineer").

Column 3: `Region`: `SF` for San Francisco, `S` for Seattle.

For this question, you should consider 6 confidence intervals total, 4 of which are pairwise, and two of which are contrasts (but are not pairwise). You get to pick which pairwise and contrasts you are interested in.

Question 2 - Scores.csv

This data is taken from a "high risk" population, who is undergoing treatment for psychological problems. A random sample was taken.

Column 1: `Beck`: The BECK depression score of the subject, with higher values indicating more at risk for depression. The scale is 0 to 75.

Column 2: `Drug`: The subjects' history of drug use, with values `Recent` (they have recently used recreational drugs), `Never` (they have never used recreational drugs), and `Previous` (they have previously, but not recently, used recreational drugs).

Column 3: `Treatment`: The length they have been receiving treatment, with values `Long` (a year or more), `Short` (less than a year).

For this question, you should consider 6 confidence intervals total, 4 of which are pairwise, and two of which are contrasts (but are not pairwise). You get to pick which pairwise and contrasts you are interested in. __

The Report Format (by topic)

Transformation of Variables

For this topic, the report will have very little interpretation. The goal for this topic is to transform the data given so that it best fits the ANOVA model. This report should be fairly short. The sections of the report should include:

a. A small introduction. A sentence or two is fine.

b. Plot your original data and the diagnostic plots/tests for the original data. Report the model fit of the original model.

c. You should consider removing outliers, transforming Y, or both. You should report back appropriate values (or plots) for **every combination of transformations and removing outliers you considered**. You should pick your "best" combination of transformed variables.

d. Discuss your results. Did transforming the data help? What are the downsides? Do you believe the transformed data is a better fit? What would you suggest for a client who wants to use this data set for ANOVA (which transformations / removal of outliers would you use, if any)?

Two Factor ANOVA

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:

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. If you believe assumptions are violated, note this and continue with the project. Do not consider transformation of variables for this topic.

IV. Analysis. Report back the "best" model (and how you chose that as the best), confidence intervals, test-statistic/s, and p-value/s, nulls and alternatives, 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, and any other useful statistics you may have calculated. These should all be interpreted 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, or unformatted output from R in the body of your report.**

v. Please have the following format:

```
Cover Page
Report for Topic I
Report for Topic II
Code Appendix
```

Small Notes

- "Model fit" refers to estimated values of your parameters, and this will be based on what model formation you used.
- Feel free to make your cover page "unique" so that it is easy to find when I hand them back.
- You may combine the Analysis/Interpretation section if you so choose.
- Notice: your project will be graded as a group effort (if you have two people). 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.