Homework 1: Introduction to Statistical Computing Using R

Instructor: Hanxiang Peng
STAT 42100: Modern Statistical Modeling Using R and SAS
STAT 52100: Statistical Computing
Due date: 09/06/2022 (Tuesday)

Instruction: Collaboration on homework assignments is acceptable, but all write-ups must be done independently and clearly indicate the submitter's understanding of the material. Unclear or disorganized homework may have points removed, even if the content is correct. **No hardcopy of the typed report needs to be turned in. The pdf file should be submitted electronically on canvas.** An assignment handed in after the deadline is late, and may or may not be accepted. My solutions to the assignment questions will be available when everyone has handed in their assignment.

Edit the program(s) and output together into a single document, showing the lines of code and relevant output produced by SAS/R. Your answers must be easy for the grader to find. A simple structure is, for each part of each question in order, to put these three things:

- Your code
- Your output
- Your answers and explanation

For SAS, the code and output are naturally separate, so this order is good; for R, the code and output can be intermingled, and that is OK. If your assignment is disorganized or otherwise difficult for the grader to deal with, you can expect to lose marks.

You are reminded that work handed in with your name on it must be entirely your own work. It is as if you have signed your name under it. If it was done wholly or partly by someone else, you have committed an academic offence, and you can expect to be asked to explain yourself. The same applies if you allow someone else to copy your work. The graders will be watching out for assignments that look suspiciously similar to each other (or to my solutions). Besides which, if you do not do your own assignments, you will do badly on the exams.

Problem 1 The quality of orange juice produced by a manufacturer (identity unknown) is constantly being mon- itored. The manufacturer has developed a "sweetness index" for its orange juice, for which a higher value means sweeter juice. Is the sweetness index related to a chemical measure such as the amount of water-soluble pectin (parts per million) in the orange juice? Data were obtained from 24 production runs, and the sweetness and pectin content were measured for each run. The data are in *ojuice.txt*. Download to save it somewhere in your computer.

1. The data values are separated by a space. Use the appropriate Tidyverse function to read the data into a "tibble".
2. Print out the whole data to see whether you have 24 runs.
3. The juice manufacturer was interested in whether there was a relationship between sweetness and pectin. To assess this, draw a scatterplot. Does it look as if there is any kind of a relationship? (I think sweetness is the outcome variable and pectin is explanatory, so draw your scatterplot appropriately.)
4. Obtain a histogram of the pectin values, using 10 bins for your histogram. Comment briefly on the shape of the histogram. Is it approximately symmetric, skewed to the left, skewed to the right or something else? (By "comment briefly" I mean "say in a few words why you gave the answer you did.")
5. Obtain a boxplot for the pectin values to re-confirm the outliers.

Problem 2 Beginning accounting students need to learn to learn to audit in a computerized environment. A sample of beginning accounting students took each of two tests: the Computer Attitude Scale (CAS) and the Computer Anxiety Rating Scale (CARS). A higher score in each indicates greater anxiety around computers. The test scores are scaled to be between 0 and 5. Also noted was each student's gender. The data *compatt.txt* (download from canvas please) values are separated by spaces.

(a) Read the data into R. Do you have what you expected? Explain briefly.

(b) How many males and females were there in the sample?

(c) Do the CAS scores tend to be higher for females or for males? Draw a suitable graph to help you decide, and come to a conclusion.

(d) Find the median CAS scores for each gender. Does this support what you saw on your plot? Explain briefly.

(e) Find the mean and standard deviation of both CAS and CARS scores (for all the students combined, ie. not separated by gender) without naming those columns explicitly.

(f) You might be wondering whether the test scores are related. Draw a scatterplot for this and color points by gender.