STAT 260 Spring 2023: R Assignment 1

Due: Friday February 3rd BEFORE 11:59pm (PT) on Crowdmark.

Introduction to R: Before attempting this assignment, read and work through the **Introduction to R Assignment 1** file posted on Crowdmark. This file contains a list of all the R commands needed to complete this assignment.

Submission: Since you will need to copy your R code and the code output, your answers must be typed. The best way to complete this assignment is by using a word processor such as MS Word or Open Office Document. (I have a link posted on our Brightspace page in the "Useful Links" module where you can download Microsoft 365 for free.)

Complete Parts 1 and 2 below, copying and pasting the required R commands, R outputs, and analysis into your Word documents. Save your assignment in the PDF file format. On Crowdmark you will be asked to upload your solutions to some parts separately, so the best way to save your work would be to put your solutions to each part requiring a separate submission upload on separate pages in your Word file. To convert your Word document to a PDF use the "Save As" feature - PDF is one of the output options there. When uploading your work to Crowdmark put your submission in the first upload area, then drag and drop the pages for the other parts into the proper submission areas.

Upload your files for submission to the assignment on Crowdmark, before Friday February 3rd at 11:59pm (PT). You should keep an electronic copy of your assignment for your records until the end of the semester.

Late policy: Late assignments will be accepted until the final cutoff of 11:59pm on Sunday February 5th. Solutions submitted within 1 hour of the Friday deadline will have a 5% late penalty applied within Crowdmark. Solutions submitted after 1 hour of the Friday deadline but before the final Sunday cutoff will have a 20% late penalty applied. Solutions submitted after the final Sunday cutoff will be graded for feedback, but marks will not be awarded.

Note: For each of the following, carry out your calculations **only** using R or RStudio. Copy and paste your command(s) and the output into your document as indicated. Format your solutions in a somewhat formal way, as if you are writing a lab report (that is, use somewhat formal language and complete sentences).

Part 1 We wish to compare the midterm scores of two sections of a particular stats class. Both sections wrote exams out of 50 possible points. Collections of students from each of the two sections were randomly sampled and their exam scores are summarized below.

Morning Section Marks:

37	39	27	33	29	32	39	40	40	50
39	40	33	39	38	29	24	31	27	36
30	36	40	39	30	41	41	34	32	40
31	32	38	39	33	32	39			

Afternoon Section Marks:

38	36	40	37	42	38	37	41	43	39
40	36	37	34	41	36	39	37	40	38
35	34	38	42	39	41	40	41	37	41
37	41	35	38	41	36				

Note: These observations are also available as single columns in a text file on Crowdmark, if you wish to make use of the scan() function along with the copy and paste capabilities on your computer.

(a) **[2 marks]** Create a histogram of the marks from the morning section of the class. The title and x-axis for the histogram should have appropriate labels. Copy and paste the relevant command from the R Console Window and the resulting histogram into your Word document. (You do not need to include the code for how you stored the data into R, just include the one line of code for the histogram.)

(b) **[2 marks]** Create one side-by-side boxplot of the two sets of marks (i.e. both boxplots on the same axes). The picture should have an appropriate title, the x- and y-axes for the boxplot should have appropriate labels, and the two groups should be labelled. (The code to change the title and axes for a boxplot is the same as how to change the title and axes for a histogram.) Copy and paste this boxplot and the line of code used to create it into your Word document. The boxplots themselves may be either horizontal or vertical (your choice).

(c) **[2 marks]** Use R to calculate the mean and standard deviation of the marks for both the morning and afternoon sections. Copy and paste the relevant commands and output from the R Console Window into your document. Write a short statement summarizing the values of your R output.

(d) **[1 mark]** Answer the following question: Which class appears to have performed better on the test? Write a few sentences explaining your opinion. You should make reference to the some of the relevant features of the two data sets (e.g. the mean or median, the spread of the data, minimum/maximum values, etc.). Use your results from both parts (b) and (c) to support your statement. You may wish to run the **summary** command in R for each class section to gather information about the maximum and minimum values.

Part 2 We are interested in analyzing the relationship between the height and the volume of timber produced by a felled black cherry tree. There is a built in data set in R that we will access for this purpose. Use the command **attach(trees)** to import the **trees** data set. This data set contains the height (measured in feet) and the volume of timber (measured in cubic feet) measurements for 31 trees (i.e. we have bivariate data here). The height measurements are already stored in a vector called **Height** and the volume measurements are already stored in a vector called **Volume** (i.e. attaching the **trees** dataset will also store the data lists of **Height** and **Volume** for you, so you can now use these two lists of data and call upon them by name). (Note also that vector names are case sensitive when calling upon the vectors of **Height** and **Volume**.)

(a) **[2 marks]** Create a scatterplot to compare the tree height to the volume of timber. (Hint: we potentially believe that tree height would influence volume of timber, so choose which data set will represent x and y wisely.) Your plot should have an appropriate title and the x- and y-axes should be labelled appropriately. Copy and paste the relevant commands and output from the R Console Window into your Word document.

(b) **[1 mark]** Describe the relationship (if any) between height and timber volume that this scatterplot shows. (e.g. Is it linear or not? Positive or negative? A strong or weak relationship?)

(c) **[1 mark]** Use the **cor** function to compute the correlation coefficient for height and timber volume. (Copy and paste the relevant commands and output from the R Console Window into your Word document.) Does this value agree with your answer about the relationship between height and timber volume? Explain.

(d) **[2 marks]** The **trees** data set also contains information about the girth of the tree (measured in inches). This information is stored in the **Girth** vector. Create a scatterplot to compare girth to the volume of timber, and then use the **cor** function to compute the correlation coefficient for girth and timber volume. (Copy and paste the R code and output for the scatterplot and the correlation coefficient calculation.)

(e) **[1 mark]** According to your work in the previous parts of this question, is height or girth a better linear predictor of timber volume? Explain.