BSDS 100: Intro to Data Science with R
Assignment 5
Due 11/18 at 11:59pm

Directions: Write a single R markdown (.Rmd) file that answers each of these questions and produces a knitted .pdf which holds your responses. Make sure that all code can be successfully run on any computer. Put your name and date at the top of the document. For all questions that require written responses, write the answer (numbered appropriately) in markdown, not as a comment in the code. Turn in the .Rmd file and .pdf file on Canvas. Late assignments are not accepted.

Playlists: Here is a playlist from me to help you get started on the assignment:

Fall Time Mix: https://tinyurl.com/57j3v6y8

1. Read **Sections 19.1, 19.2, 19.3, 19.4** and **19.5** in the *R for Data Science* text on Functions here:

   https://r4ds.had.co.nz/functions.html

2. Answer the questions posed as exercises in these Sections, including:

   - **Section 19.4.4** numbers 1 - 6
   - **Section 19.5.5** numbers 1 - 4

3. On our course website you will find the dataset `fitness.csv`, which has two columns:

   Column 1: `Tread`: The typical amount of time an individual spends training at high intensity on the treadmill during a workout, in minutes (X).

   Column 2: `Run`: The time it took to complete a 10 kilometer run (in minutes) (Y).

   The goal is to see if an athlete's treadmill time has a linear relationship with their 10 kilometer run time.

   (a) Plot a scatter plot of the data and add the estimated regression line.

   (b) Write down the equation for the estimated regression line.

   (c) Interpret the slope in terms of the problem. Specify its units.

   (d) Based on your plot from (a), would it make sense to interpret the intercept in this case? Explain.

   (e) Calculate and report the value of the sum of squared residuals. What are the units of this quantity?

   (f) If a subject's high intensity treadmill training time decreased by two minutes, what is the expected amount of time their 10 kilometer run would increase by?

   (g) Calculate and report the residual associated with the subject in the first row of your dataset. What does this value mean? Does it imply the line over- or underestimated the 10k run time?

   (h) Predict the value of an individual's runtime when, when their training time is 35 minutes. Do you have any issues with making this prediction?

   (i) Predict the value of an individual's runtime when, when their training time is 8.5 minutes. Do you have any issues with making this prediction?