Introduction to Data Science
Homework 4: Due Wednesday September 26 at 2:00pm

Exercises:

1. Go through section 9 Functions of the R Programming course in the `swirl` package, then answer the following questions:

   (a) What is a function?

   (b) What does the `Sys.Data()` function do? How many input arguments are required?

   (c) What are the two "slogans" for R stated by John Chambers?

   (d) How do you see the source code for an R function?

   (e) Why would having default arguments by useful?

   (f) What does the `args` function do? Give an example of its use.

   (g) Explain why one might want to pass a function as an argument to another function.

   (h) What is an easy way to return the last element of an arbitrary vector?

   (i) What does the `paste` function do?

   (j) What is the significance of the "dot-dot-dot" argument for a function in R?

2. Write an R function that inputs a vector and computes the mean of the vector. Save your function in an R script called `my_mean_func.R`. Be sure to test your function and make sure it is working correctly.

3. Write an R function that inputs two whole numbers and returns the remainder after dividing the first by the second. Save your function in an R script called `my_remain_func.R`. Be sure to test your function and make sure it is working correctly.

4. Read the first three sections of Chapter 4 Scores and Rankings from *The Data Science Design Manual* (remember that this is available through the library) and answer the following questions.

   (a) What is a "scoring function?"

   (b) What is a "score" according to the definition given in section 4.2?

   (c) Describe an approach or approaches to building effective scoring systems and evaluating a scoring system.

   (d) What is a ranking? Provide some examples.

   (e) What are the characteristics of a good scoring function?

   (f) Describe Z-scores and normalization.

5. Find or make up a formula for some kind of score. Write an R function that implements your formula. Apply your function to some data that is either real or simulated. Discuss whether your scoring function is good or not.

6. Using the `flights` data form the `nycflights13` package, find all flights that
   (a) Had an arrival delay of two or more hours
   (b) Flew to Houston (IAH or HOU)
   (c) Were operated by United, American, or Delta
   (d) Departed in summer (July, August, and September)
   (e) Arrived more than two hours late, but didn?t leave late
   (f) Were delayed by at least an hour, but made up over 30 minutes in flight
   (g) Departed between midnight and 6am (inclusive)
7. Another useful `dplyr` filtering helper is `between()`. What does it do? Can you use it to simplify the code needed to answer the previous challenges?
8. How many flights have a missing dep_time? What other variables are missing? What might these rows represent?
9. How could you use arrange() to sort all missing values to the start? (Hint: use is.na()).
10. Sort `flights` to find the most delayed flights. Find the flights that left earliest.
11. Sort `flights` to find the fastest flights.