Lecture 4: Data Structures in R

Problem Set for Lecture 4

These questions will assist in bolstering your understanding of the material in Lecture 4. Emphasis should be placed on having the correct code/output as well as communication. The deliverable should be a knitted RMarkdown document to pdf without any code running off the page. You will be able to present these in office hours as an oral assessment interview along with the problem set for Lecture 2 and Lecture 3.

***** First, make sure you have read the Lecture Material's writeup *****

1. Answer the following questions relating to dataframes.

   (a) Create a dataframe in R relating to students in a class by following the instructions below:

      i. Create a vector of 100 random variables from the normal distribution with a mean of 75 and a standard deviation of 10. Round all of the numbers to the nearest integer and call this vector "grade".

      ii. Create a vector of 100 random samples from the list: Math, Computer Science, Data Science and Accounting and call this vector "major". Alter the probabilities when sampling so that each major is sampled at the rate of whatever you think is realistic.

      iii. Create a vector of 100 random samples from the list: Male and Female. Name this vector "gender".

      iv. Combine the vectors into a single dataframe with the student's major, then their gender, then their grade. Do not display the whole dataframe, only the first few observations.

   (b) Produce a count for each major, gender, and major/gender combination. Each count can be done with a single line of code.

   (c) Display all students who are male computer science students

   (d) Calculate the average grade for students who are either Data Science majors or Accounting majors

   (e) Display all students who are either female Accounting majors or male students who scored below 70.

   (f) Display all students who scored below 70 and are either female Accounting majors or male students.

   (g) Display all of the even numbered rows and odd numbered columns

2. Answer the following questions relating to matrices by using the built-in `mtcars` dataset in R.

   (a) Create a new dataframe called `cars` that contains the `mtcars` dataset. Then convert it to a matrix called `car_mat` using the `as.matrix()` function. The matrix should contain the following columns (in this order): `mpg`, `hp`, `wt`, `qsec`.

   (b) Determine the number of rows and columns in `car_mat`. Then display the first 4 row names and the column names.

   (c) Display the submatrix consisting of rows 3 through 8 and columns 2 through 4.

   (d) Create a new vector called `hp_per_wt` defined as horsepower per weight using the columns in `car_mat`. Then add it as a new column to the right side of `car_mat`.

   (e) Determine which rows of `car_mat` have horsepower greater than 200. Display only the row names of those cars.

3. Answer the following questions relating to lists by using the built-in `mtcars` dataset in R.

   (a) Create a vector called `high_hp` that contains the row names of all cars with `hp` greater than 200.

   (b) Create a list called `car_report` with the following named components (in this order):

   - `data` : the first 8 rows of `mtcars`
   - `mean_mpg` : the mean of `mpg` (handle missing values correctly)
   - `median_wt` : the median of `wt` (handle missing values correctly)
   - `high_hp` : the `high_hp` vector from part (a)

   (c) Use the `str()` function to display the structure of `car_report`. Based on the output, state which component is a dataframe, which components are numbers, and which component is a character vector.

   (d) Extract the `high_hp` component two different ways: (1) using the name and (2) using its position in the list.

   (e) Display the `mpg` column within the `data` variable while referencing the `car_report` list.

   (f) Add a new component to the list called `notes` that contains a character vector with two short sentences explaining what a list is used for in R.