PS132B Problem Set 1

Due 12:29PM Tuesday February 7, 2023

Please submit this assignment by uploading your R Markdown code file (`.Rmd`) AND either an html (`.html`) or pdf (`.pdf`) output onto bCourses before the due time, with easy-to-recognize file names (e.g., `pset1_KirkBansak.Rmd`). Your homework will be graded based on completeness, accuracy, and readability of both code and written answers.

The point allocation in this problem set is given by:

Q1.1	Q1.2	Q1.3	Q1.4	Q1.5
10	10	10	10	10

Q2.1	Q2.2	Q2.3	Q2.4	Q3	Total
10	10	10	10	10	100

Part 1: Basic Operations in R

1. Creating a Vector. Create a numeric vector that is the sequence of all integers between 1 and 1000 and assign this vector the name `vec1`.

2. Sampling. Create another vector of the same 1000 integers but whose order is randomized. You should do this by randomly drawing from the vector `vec1`, and label your new vector `vec2`. Hint: Use the `sample()` function. Remember that you can look up the documentation for any function to better understand how to use it. To do so for the `sample()` function, enter `?sample` into the R console.

3. Creating a Data Frame. Bind these two vectors together in a data frame, and call the data frame `dat`. Make sure that the first column of `dat` corresponds to `vec1` and the second column corresponds to `vec2`.

4. Comparing Variables. Compute the correlation between the two variables in `dat`.

5. Comment on whether you expected this correlation to be large or small (i.e. close to zero), and why.

Part 2: Preparing to Work with Real Data

Find the data file named `data_health_synth_small.csv` on bCourses, and save it onto your computer in the same folder/directory as the `R Markdown` file you are creating for this assignment. This is a small portion of the synthetic dataset from Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science* Vol. 366, No. 6464 (2019). We will be reading that paper later in the semester, when we discuss the idea of algorithmic bias and other issues at the intersection between machine learning, policy, and ethics. The data contain measurements of individual (de-identified) patients in a hospital system.

1. Reading in Data. Use the `read.csv()` function to read the dataset into `R` as a data frame called `hdat`.

2. Data Size. Report how many rows and how many columns there are in the data set, and explain what the rows and columns represent (i.e. each row corresponds to what, and each column corresponds to what).

3. Summarize Data. This small dataset contains the following variables.

   > `cost`: Total medical expenditures over the year, rounded to the nearest 100.
   >
   > `race`: Patient race. The paper focuses on racial bias across Black and White patients, and so the data contain only two racial categories.
   >
   > `female`: Indicator for identification with female gender.
   >
   > `bps_mean`: Mean systolic blood pressure over the year.

   Use the `summary()` function to compute summary statistics about the data, and in words, report some of the results.

4. Compute the mean `cost` across the different racial groups, and report your findings in words.

Part 3: Produce Final Output with R Markdown

Produce either an `html` file or `pdf` file output that contains your code and answers to the above questions. You should be writing your answers in an R Markdown `Rmd` file. You should be using `RStudio`, in which case you can simply use the "Knit" button to render the output. Be sure to use headers so that it is clear which answer corresponds to which question. You will turn in both your `html`/`pdf` and `Rmd` files.

More information on using can be found here:
`https://rmarkdown.rstudio.com/authoring_quick_tour.html`