# POLI 170 Data Analysis Assignment 1

Due 3:29PM Tuesday October 12, 2021

Please submit this assignment by uploading your html (`.html`) and code (`.Rmd`) files onto Canvas before the due time, with easy-to-recognize file names (e.g., `assignment1_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 | Q1.6 | Q1.7 | Q1.8 |
|---|---|---|---|---|---|---|---|
| 5 | 5 | 5 | 10 | 5 | 5 | 5 | 5 |

| Q2.1 | Q2.2 | Q2.3 | Q2.4 | Q3 | Q4 | Total |
|---|---|---|---|---|---|---|
| 5 | 5 | 10 | 10 | 15 | 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. Lookup and Indexing. Determine which elements of column 2 of `dat` contain the numbers 2, 47, 290, and 812. Store the indices of these elements (i.e. the row numbers) in a manner of your choosing.

5. Replacement. Now replace the instances of the numbers 2, 47, and 290, and 812 in column 2 of `dat` with missing values (NA).

6. Variable Renaming. Rename the columns of `dat`. The new names, in order, should be `caseid` and `wage`.

7. Compute Summary Statistics. Calculate the mean, median, and standard deviation of `wage` and report those values. Since there are NA values, you will need to use the "`na.rm = TRUE`" argument when calculating these values.

8. Subsetting. Create a new data frame, `dat2`, that is `dat` without the missing values. In other words, delete all observations (rows) that have missing wage values.

# Part 2: Preparing to Work with Real Data

Find the data file named `data_health_synth_small.csv` on Canvas, 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 much larger 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). 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.

   `female`: Indicator for identification with female gender.

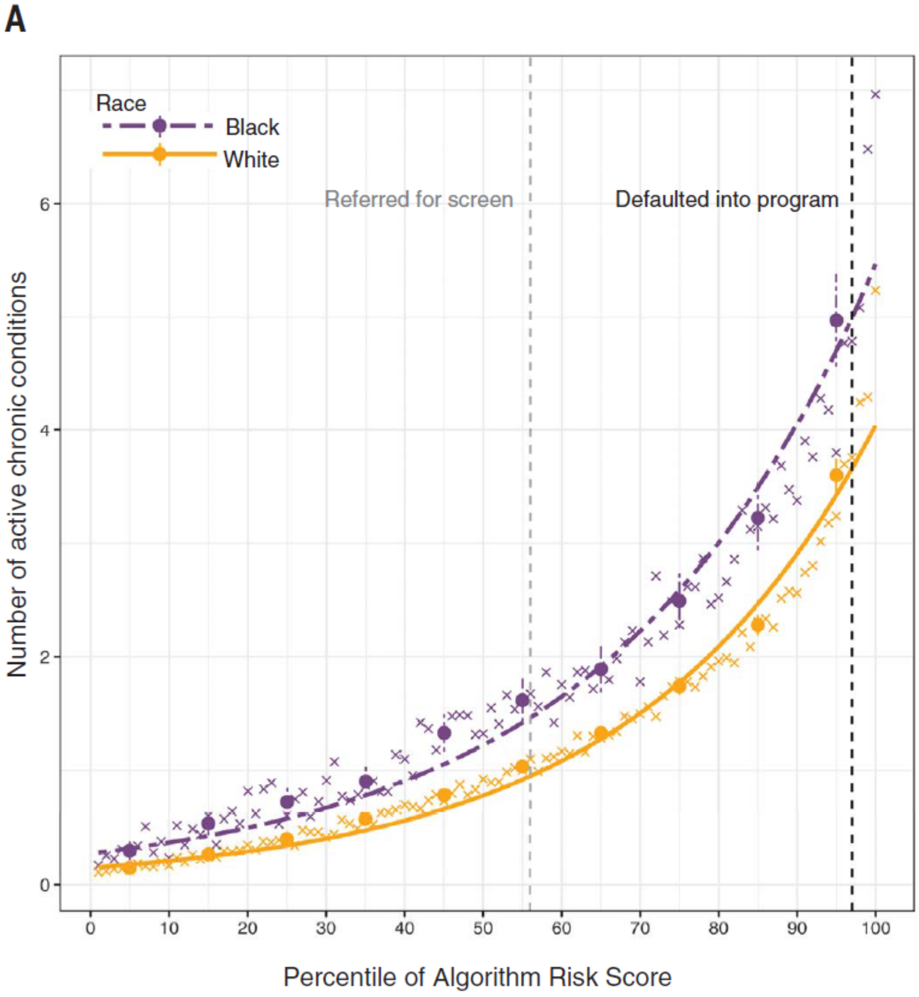
   `bps_mean`: Mean systolic blood pressure over the year.

   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.

# Part 3: Short Answer

The figure below is from Obermeyer et al. (2019). In 2-4 sentences, explain what the two curves represent and the implications of the fact that they are divergent from one another.



## Part 4: Produce Final Output with R Markdown

Produce an `html` file output that contains your code and answers to the above questions. You should be writing your answers in an R Markdown `Rmd` file. If you are using `RStudio`, you can simply use the "Knit" button to render the output. Be sure you are clearly demarcating which answer corresponds to which question. You will turn in both your `html` and `Rmd` files.

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