# EMIS 5357 Problem Set 1

Due Date: Sep 10, 2021 (Fri)

**Instruction**: In order to receive full credit, two files should be submitted. R codes that are implemented to solve all problems should be written in a single R script file named `PS1_YourLastName.R`. A pdf file named `PS1_YourLastName.pdf` should clearly contains your answers.

### Analyzing State Data in the United States

We will be examining the dataset `StateData.csv`, which is available on Canvas and has data from the 1970s on all fifty U.S. states. Indeed, this dataset is a subset of a built-in dataset provided by R. A description of the variables in the dataset is given in Table 1. All plots should contain a proper title and axes should be clearly defined.

| **Variable** | **Description** |
|---|---|
| `Population` | Population estimate of the state in 1975. |
| `Income` | Per capita income in the state in 1974. |
| `Illiteracy` | Illiteracy rates in 1970, as a percentage of the state's population. |
| `LifeExp` | The life expectancy in years of residents of the state in 1970. |
| `Murder` | The murder and non-negligent manslaughter rate per 100,000 population in 1976. |
| `HighSchoolGrad` | The high-school graduation rate in the state in 1970. |
| `Frost` | The mean number of days with minimum temperature below freezing from 1931 to 1960. |
| `Area` | The land area (in square miles) of the state. |
| `Longitude` | The longitude of the center of the state. |
| `Latitude` | The latitude of the center of the state. |
| `Region` | The region (Northeast, South, North Central, or West) that the state belongs to. |

Table 1: Variables in the dataset `StateData.csv`.

(a) Create a scatter plot of all of the states by putting `Longitude` on the horizontal axis and `Latitude` on the vertical axis. Does the shape of the plot look like the outline of the United States?

(b) Suppose that you draw a scatter plot of two variables, where the horizontal axis is `Latitude`. Which variable, do you think, can be placed on the vertical axis such that the scatter plot will look like a line? Provide this plot with proper reasoning of your choice of the variable.

(c) Based on the dataset, compute the number of populations in each of regions as defined in `Region` variable (i) by using `which` function, and (ii) by using `tapply` function. Note that your answers for both should be the same.

(d) Create a box plot of the variable `Murder` for each `Region` (so you should have for box plots in total). Provide your observation of the box plot, e.g., which region has the highest median murder rate? which region has the safest state (lowest murder rate)? etc.

(e) Provide a box plot of the distribution of income of (i) states that have more than 55% of high school graduation rate and (ii) states that have less than 55% of high school graduation rate. Interpret your results.

(f) Using `table` function, compute the number of states that the average income exceeds $4,500 for each region.

(g) By drawing a few scatter plots while fixing the $y$-axis as the `LifeExp` variable, provide a variable that you think most significant to have negative impact on life expectancy. Does the result seem intuitive to you?

### (Optional for 5357 students; Required for 7357 students) WHO Dataset Revisited

In class, we discussed that the best way to learn R is to search from Google and give trials and errors. No one knows all of commands and functions in R; if you need something, then you should be able to find appropriate resource and apply accordingly to your implementation.

The goal of this problem is to have such experience with the WHO dataset that we covered in class. Implement an R script, with proper online resources and helps, to draw the scatter plot of GNI and Life Expectancy, in which (i) European countries are labeled with red, and (ii) Non-European countries are labeled with blue. In order to receive full credit, your script should exactly reproduce Figure 1. You would have to set legend, points should be filled, and the plot should have title and axis names.

*Hint*: Start by splitting the WHO dataset to two subsets, depending on whether `Region` value is Europe or not. You would have to search some words like "add points to a plot in R" from Google or other search engines.

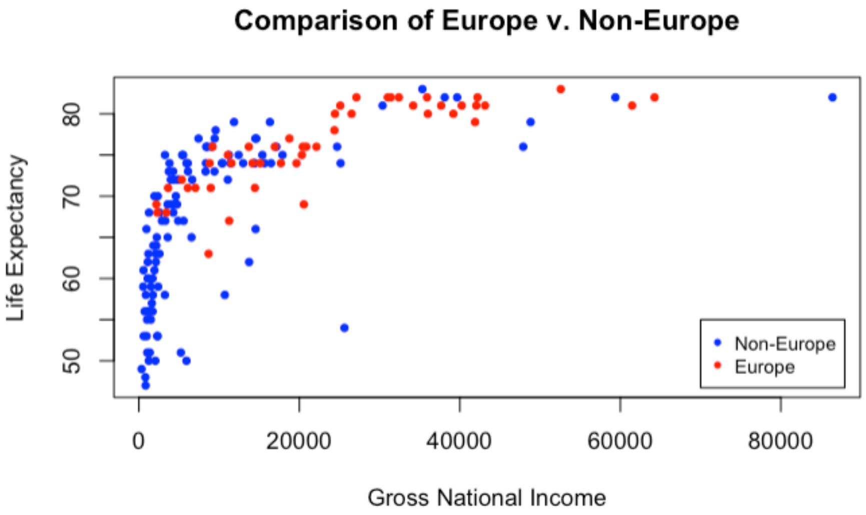

Figure 1: Scatter plot with different color labels.