# **Data Preparation and Exploratory Data Analysis Assignment**

# Introduction

For this assignment, you will produce descriptive statistics for a set of interval level variables in the dataset. There are two goals for the assignment. The first goal is for you to become familiar with the R code that is used to run descriptive statistics. The second is for you to learn about the specific qualities of the variables. When you use the data for the assignments later in the semester you will need to understand the characteristics of the variables and how well they will, or will not, work in a model.

## Directions

Use the lecture material from Lesson 3 and the reading material *R in Action*, Chapter 4, Section 4.4 and Chapter 7 as a guide for making changes in the dataset and producing descriptive statistics.

The completed assignment should adhere to the following guidelines:

* 1. Include your answers on the assignment document. Leave the assignment document as it is and include your answers after each question.
  2. Write your answers using complete sentences with correct punctuation, grammar, and spelling.
  3. Submit your completed assignment through the Blackboard portal in Lesson 3.

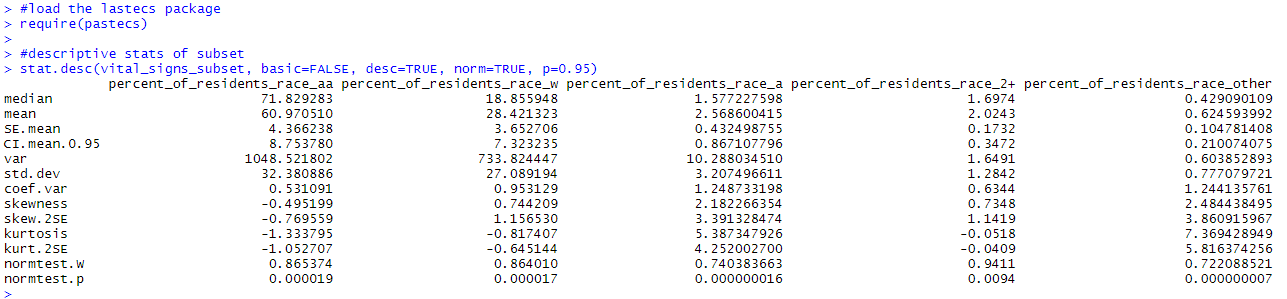
Use the Field Aliases Excel file as the guide for variable names. Open the BNIA Vital Signs Census Demographics CSV file in RStudio. Use the dataset to answer the following questions:

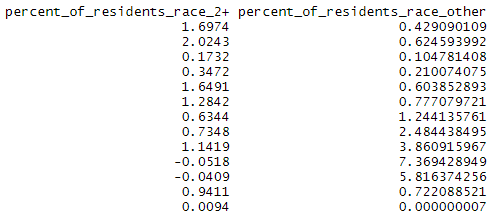
1. Change the variable names to names that are descriptive of the variables (something other than pasi17, etc.). Paste the code that you used to make the changes below. *(1 point)*

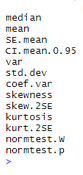


1. Create a subset of the BNIA Vital Signs data with five interval level variables of your choice. Choose five that you find interesting. Produce a complete set of descriptive statistics for the five variables with **pastecs**. Paste the code and the output below. *(2 points)*

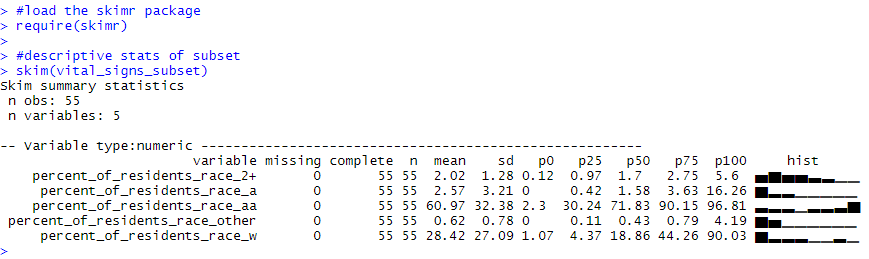








1. Use **skimr** to produce descriptive statistics and histograms. Paste the code and the output below. *(2 points)*



1. Describe each of the variables using what you observe from the descriptive statistics and the histogram. Describe the results in terms of statistical distributions and in terms of what it means in the real world (for example, one of the observations has a minimum of zero on this variable, meaning that . . .). You do not need to include a discussion of the number of observations, missing, complete, or unique. The dataset is complete so these variables are not particularly informative. You also do not need to discuss the standard errors of the estimates, or coefficient of variance. However, with those exceptions, you should make sure to use **all of the statistics** (means, skewness, kurtosis, standard deviations, the statistical significance of normality, etc.) to be as complete as possible in your discussion. *(5 points)*

percent\_of\_residence\_race\_2+:

The percentage of residents identifying with **two or more races** is 2%, this includes: White, African American, Asian, Hispanic, and Other (Hawaiian/Pacific Islander, Alaskan/Native American, Other). The resulting histogram is a tightly right-skewed distribution, with a small SD of 1.28. The corresponding kurtosis indicates a mild thin-tailed distribution of -0.0518. This variable’s normality is 0.93, suggesting the distribution of data is within normal ranges +/-1.96.

percent\_of\_residents\_race\_a

The percentage of residents identifying as **Asian** is 2.57%. The resulting histogram is a right-skewed distribution, with a relatively large SD of 3.21. The corresponding kurtosis indicates a strong fat-tailed distribution of 5.39. This variable’s normality is 0.73, suggesting the distribution of data is within normal ranges +/-1.96.

percent\_of\_residents\_race\_aa

The percentage of residents identifying as **African American** is 60.97%. The resulting histogram is a left-skewed distribution, with a moderate SD of 32.38. T The corresponding kurtosis indicates a moderate thin-tailed distribution of -1.33. This variable’s normality is 0.87, suggesting the distribution of data is within normal ranges +/-1.96.

percent\_of\_residents\_race\_other

The percentage of residents identifying as **Other** (Hawaiian/Pacific Islander, Alaskan/Native American, Other/Unspecified) is 0.62%. The resulting histogram is a right-skewed distribution, with a relatively large SD of 0.78. The corresponding kurtosis indicates a strong fat-tailed distribution of 7.97. This variable’s normality is 0.71, suggesting the distribution of data is within normal ranges +/-1.96.

percent\_of\_residents\_race\_w:

The percentage of residents identifying as **White** is 28.42%. The resulting histogram is right-skewed, with a SD of 27.09. The corresponding kurtosis indicates a moderate thin-tailed distribution of -0.81. This variable’s normality is 0.85, suggesting the distribution of data is within normal ranges +/-1.96.

## Scoring

The assignment is worth 10 points. You must have a correct and complete answer to receive full credit.

## Rubric

Questions 1-3: There are variations in how code can be written and there are several packages that can be used to produce descriptive statistics. For full credit the code and the output must use the required packages and the code must be correct.

Question 4: Content is worth 4 points, and spelling, punctuation, and grammar are worth 1 point.