DUE: March 26, 2019 at 11:59 PM

ANA1002 – Module 8 Assignment /32

Create and submit an R script which, when run, will print the answers to the following questions and output any graphics. Your R script must include a title with your name and student number and comments for each question number.

1. **(18 marks)** Read the *airquality.csv* data set into R. This gives the ozone level, solar radiation, wind speed, and temperature (in degrees Fahrenheit) over five months.
    a) How many missing values are present for each variable?
    b) Create a data frame of complete cases and find the mean temperature using listwise deletion.
    c) Find the mean temperature using pairwise deletion.
    d) Which rows contain missing temperature values?
    e) Create a missing variable plot and an upset plot for the air quality data. Interpret EACH plot to explain why there are differences in the mean temperature using listwise and pairwise deletion.
    f) Create a missing variable plot to examine any trends in missing variables BY month. Interpret your plot. Suggest a reason why the missing values may be higher in some months than others.
    g) How many of the ozone values are outliers (using the default 1.5 IQR setting)? What are the ozone outlier values? Create a new data frame called ozone.complete that has all rows with ozone outliers removed.

DUE: March 26, 2019 at 11:59 PM

2. **(14 marks)** Read the demographics dataset into R. This dataset gives the state, year, voting status (1 = vote, 0 = no vote), income (levels 1 to 17), education (levels 1 to 4), age, gender (0 = male, 1 = female), and age group (child, adult, elderly).
    a) Use the editrules package to create a set of rules so that age is between 0 and 125, vote is either 0 or 1, year = 2000, and female is either 0 or 1.
    b) How many times is each rule violated?
    c) Create a plot of the edit violations and interpret EACH part of the plot (edit violation frequency and edit violations per record).
    d) How any duplicate records are there in the demographics database? Why is the value fairly large? How would you modify the data collection practice to solve this issue?
    e) Create a data frame with all duplicate entries removed.

Save your R Script as: Last Name, First Name Module 8

Upload your R Script to the "Module 8 Assignment" dropbox on Moodle before March 26, 2019 at 11:59 PM.