DUE: April 9th, 2019 at 11:59 PM

ANA1002 – Module 10 Assignment /30

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. Read the *airquality.csv* data set into R. This data set records the ozone level, solar radiation, wind speed, and temperature (in degrees Fahrenheit) over five months (from month 5 = May to month 9 = September).
   a) Read in the *"airquality.csv"* data and save it in an object called air. Examine the missing values in the data set using the **md.pattern()** command to answer the following: (4 marks)
      i. How many entries are missing altogether?
      ii. What is the variable with the most missing values?
      iii. Which sets of variables are missing at the same time? How many times are they missing?
   b) Find the mean of each of the variables in the air quality data set using pairwise deletion. (1 mark)
   c) Initialize a new data set called air.median from the air data set (for example: air.median<-air). Impute the missing solar radiation values with the **MEDIAN** of the non-missing radiations in the air.median data set. (3 marks)
   d) Initialize a new data set called air.mean from the air.median data set . Impute the missing temperature values with the mean temperature ***for the month*** that the temperature is missing from in the air.mean data set. For example, impute missing month 5 temperature values with the mean of the non-missing temperatures for month 5. (5 marks)
   e) Initialize a new data set called air.ratio from the air.mean data set. Impute the missing values of the Ozone variable using ratio imputation in the air.ratio data set (let the correlated complete variable be temperature). (4 marks)

f) Initialize a new data set called air.complete from the air.ratio data set. Use linear regression to impute the missing values of Wind using Ozone as the independent variable in the air.complete data set. (4 marks)

g) Check the air.complete data set for missing values. (If you have done the question correctly you should have no NAs remaining!) Find the mean of each of the variables in the air.complete data set. (2 marks)

h) Starting with the original *"airquality.csv"* data set, use the mice package in R to impute missing data using m = 5 and seed = 2. Save the imputed data set in an object called imputeddata. Extract each of the five sets of imputed values using the complete() function and then create a data frame called air.complete2 with all of the imputed data sets (the data frame should have 765 rows). Find the mean of each of the variables in the air.complete2 data set. (5 marks)

i) Compare the means of all of the variables in the air.complete and air.complete2 data sets from part g) and part h). Do you think one set of imputed values is better than the other? (2 marks)

Save your R Script as: Last Name, First Name Module 10

Upload your R Script to the "Module 10 Assignment" dropbox on Moodle before April 9th, 2019 at 11:59 PM.