DUE: April 23rd, 2019 at 11:59 PM

ANA1002 – Module 12 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. **(19 marks)** Read the *normality.csv* data set into R
   a) Identify the minimum, maximum, mean, and median of the distribution. Based on the mean and median, do you think the distribution is skewed? If so, in which direction? (5 marks)
   b) Calculate the skewness and kurtosis for the distribution, and interpret the values. (4 marks)
   c) Create a histogram and boxplot of the data. Describe the distribution based on these plots (make sure to comment on each plot!). (4 marks)
   d) Create a QQ-plot for the data with a line that can be used to assess normality. Interpret the plot. (3 marks)
   e) Use the Shapiro-Wilk test to determine if the data is normally distributed at the $\alpha = 0.05$ level. Give the results of the test and your interpretation. (3 marks)
2. **(11 marks)** Import the *array.csv* data set into R.
   a) Calculate the skewness of each column of data in the array data set (HINT: use the apply function). Determine the maximum and minimum skewness values. Which column of data is the most positively skewed? Which is the most negatively skewed? (5 marks)
   b) Suppose we want to identify which columns are highly skewed so they can be removed from the array. Use R to give a list of all columns in which the absolute magnitude of the skew is greater than or equal to 0.3. How many columns meet this condition? (4 marks)
   c) Create a subset of the data that only contains the normal columns (with skew less than 0.30). (2 marks)

DUE: April 23rd, 2019 at 11:59 PM

Save your R Script as: Last Name, First Name Module 12

Upload your R Script to the "Module 12 Assignment" dropbox on Moodle before April 23rd, 2019 at 11:59 PM.