St Lawrence College, School of Business

Graduate Certificate Program, Business Analytics

**ADMN5003: Data Acquisition, Analysis and Modelling Techniques (Fall 2022)**

12 Marks

* **Open book assignment.**
* **Work individually or in a group of two students.**
* **Submit two files on BlackBoard in the folder of “Assignment 2”.**

**1. A R file, name it as “assign2R\_First name1\_First name2”. Write all your answers and coding on the same file. Clearly label question numbers.**

**2. A data file in a .csv format, name it as “assign2data\_First name1\_First name2”. This should be your original file, not your new file.**

* **10% late penalty per day**

1. Collect your own data including 2 continuous variables and 1 categorical variable with 2 categories. Save the data in a CSV file.
   1. Sample size is 24 in each variable.
   2. Assign a missing value to each continuous variable.
   3. Data in each row should be independent from data in other rows.
   4. They represent a random sample of a population.
2. Clearly define your population and describe briefly how you did random sampling. Explain what each column means. (1 mark) Write this answer in your R codes in a section called “#Question 1”.
3. Save your data in a CSV file. It should include 4 columns in the following order and 25 rows (The first row is the head of each column). (1 mark)

Column 1: Identifier of each row (eg. John, Joe, James etc, Customer #1 to #24).

Column 2: Continuous variable (dependent/response variable/output variable).

Column 3: Continuous variable (independent/predictor variable/input variable), contains a missing value.

Column 4: Categorical variable with 2 categories (independent/predictor variable/input variable)

1. Explore and summarize the data as follows: (Total 4 marks)
2. Replace the missing values with the median values of their corresponding columns. (1 mark)
3. Compute the mean, median, min, max and standard deviation for each of the continuous variables, using sapply(). (1 mark)
4. Plot a side-by-side box plot of dependent variable (column 2) as a function of categorical independent variable (column 4). Include main title and the label of two axes. (1 mark)
5. Build a scatter plot between dependent variable (column 2) and independent continuous variable (column3). Include main title and the label of two axes. (1 mark) Be mindful about which variable is in the x axis and which one is in the y axis.
6. Use the new CSV file do data partition and prediction accuracy measures. (Total 4 marks)
7. Partition the data into training (60%) and validation (40%) sets. (1 mark)
8. Fit a linear regression model to the two continuous variables in the training data. (1 mark)
9. Use the “forecast” package to compare the accuracy of the model on training and validation data. (1 mark) It there any overfitting of the model? Why? (1 mark)
10. Use the “Cereals” data to answer the following questions: (Total 2 marks)
11. Perform a Principal Component Analysis on “any 7 variables” using normalized data. Remove missing values. (1 mark)
12. How many components in your future modelling are sufficient, and why? (1 mark)