BIA 5302 – FALL 2023

ASSIGNMENT #1 (5%)

# INSTRUCTIONS:

* This assignment is to be completed in groups.
* Use MLA or APA format.
* The solutions must be submitted via Blackboard through the assignment’s link.
* Due date: Midnight of 09/1.

# CASE STUDY:

BostonHousing dataset contains data collected by the US Census Service concerning housing in the area of Boston Massachusetts. It was obtained from the StatLib archive [(http://lib.st](http://lib.stat.cmu.edu/datasets/boston))a[t.cmu.edu/datasets/boston).](http://lib.stat.cmu.edu/datasets/boston)) The dataset has 167 cases.

The data was originally published by Harrison, D. and Rubinfeld, D.L. `Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.

The **BostonHousing.xlsx** dataset has 11 attributes. The dataset comes with different imperfections (missing and outliers). As described earlier, most algorithms will not process records with these imperfections.

# REQUIREMENTS:

**PART A**

Use the provided data file in the following tasks:

1. Except PTRATIO predictor, perform the necessary “Handling Missing Data” operations to the missing values and highlight them with **yellow**.
2. Find possible "outliers" in the PTRATIO predictor. The possible causes of outliers are:
   1. Typing non-numeric value.
   2. Shift in decimal place while data entry error.
   3. Genuine case of outlier.

Highlight the cells with outlier cases and state the possible cause indicating a, b, or c.

# PART B

Use the provided data file in the following tasks:

1. Substitute the missing data by NaN (not a number).
2. Write and provide Python code to implement:
   1. Omission
   2. Imputation