CSE 40647/60647: Data Science **Fall 2021**

Homework Programming Assignment 1: Data Processing

Handed Out: August 24, 2021 *Due: September 12, 2021 11:55pm*

This is Graduate Version for students who register the 60000-level section. The undergraduate version does not have Question 3 (the last question). Here the total number of points is 60. Your score will be normalized to [0, 50].
Save and submit your solution file as *NETID-hw1-programming.zip*. The zip file has *NETID-hw1-programming.pdf* and (saving *hw1.ipynb* as) *NETID-hw1-programming.ipynb*.

1 Incremental sample mean and variance (30 points)

Suppose the population size is $N = 1,000,000$. We sample $n = 9$ examples x_i $(1 \leq i \leq n)$ from the data. Suppose the mean value of the sample data is $\mu = 10$ and the variance is $v = 18$. Now we sample one more example $x_{n+1} = 20$ from the data. So the sample size is $n + 1 = 10$. The task is to incrementally calculate the sample mean $\mu' = f(\mu, n, x_{n+1})$ and sample variance $v' = g(v, \mu, n, x_{n+1})$. Note that the result doesn't depend on x_i $(1 \leq i \leq n)$.

** Function f is not allowed to be used or duplicated in g, and μ' is not allowed to be used in g. Actually, it will make your g function look simpler if avoid using f or μ'.

1.1 [12 points] Derive and write the mathematical functions of $f(\cdot)$ and $g(\cdot)$ in pdf.

1.2 [10 points] Complete the functions $f(\cdot)$ and $g(\cdot)$ in ipynb.

1.3 [4 points] Run the codes to obtain the new mean value and new variance in ipynb.

1.4 [4 points] Write the results μ' and v' in pdf.

2 Correlation analysis (20 points)

Analyze data in *data-faculty.csv*, **NOT** *data-faculty-small.csv*. This file has 103 rows of data.

2.1 [5 points] Describe the mean value, median, Q1, Q3, and variance of the feature "Count" (i.e., score of CS ranking).

2.2 [5 points] Normalize the feature "Count" by MIN-MAX and print the normalized feature values.

2.3 [5 points] Normalize the feature "Count" by Z SCORE and print the normalized feature values.

2.4 [5 points] Calculate the correlation coefficient ρ between the original (not the normalized) "Count" and "Faculty".

Perform the tasks in ipynb. Present the results in ipynb.

3 Data integration and cleaning (10 points)

[10 points] Write a piece of code to integrate

- *data-faculty.csv*: 103 rows, column "Faculty;"
- *data-graduate.csv*: 11 rows, column "#Graduate;"
- *data-tuition.csv*: 9 rows, column "Tuition;"
- *data-salary.csv*: 10 rows, column "Early Career Salary;"

and generate a cleaned dataset that has as many rows as possible. The format is like *data-early_salary-small.csv* Implement your solution in ipynb. Present the results in ipynb.