**CSC 630 Introduction to Data Science**

**Exercise: Calculate the Pearson’s correlation coefficient for two variables**

**Goals**

1. Use the csv module
2. Practice Python programming
3. Learn to calculate Pearson’s correlation coefficient

**Background**

Correlation tells us how strongly two variables are related and, therefore, help us better understand our data. Correlation can also help in data mining – some data mining algorithms need variables (in data mining, they’re often called *features*) that are uncorrelated. Hands-On Example 3.4, in our text, shows an example calculation of Pearson’s correlation coefficient for a data set holding the measured heights and corresponding weights of people. In this exercise, we’ll calculate the coefficient and see if our results match those shown in the example

**Procedure**

1. Open *Ex02\_Correlation.py* in your IDE or text editor.
2. Note that the skeleton code imports the csv and math modules.
3. Examine the main function. You are given the filename. Translate the pseudocode comments into Python:
   1. Call readcsv() to get the data. You are given the readcsv() function. Note that it returns a list of lines from the file.
   2. Call calc\_coeff() to calculate and return the coefficient. Send the data from the file to the function.
   3. Print the coefficient and compare it with the coefficient published in the text.
4. Write the code in calc\_coeff():
   1. Get the number of data points. You can assume that each row of the data (after the header row) holds one height and one weight value.
   2. Examine the correlation formula, given in the text. The formula uses x and y as generic variable names. In your program, height will be the x values and weight will be the y values. You can see that you’ll need:
      1. the sum of pairwise products (sum of height[i] \* weight[i] for i ranging from 0 to the number of data points)
      2. the sum of the heights
      3. the sum of the weights
      4. the sum of the squares of the heights
      5. the sum of the squares of the weights

Use a loop to get all of these values.

* 1. Calculate the coefficient. I suggest you break it down and calculate the numerator, then the two parts of the denominator, then carry out the division.
  2. The example in the text shows the values of all these calculations. It’s a good idea to print them to see if your results match the values shown in the text.

**Deliverables**

Submit your .py file on Blackboard.