**CPSC 223P Assignment #4**

Spring 2020

Due Date: 05/11

In this assignment, you will use Pandas library to perform analysis on the dataset stored in the following csv file: breast-cancer-wisconsin.csv. You can download the csv file from the Titanium.

Please write script(s) to do the following:

1. Read the csv file and covert the dataset into a DataFrame object.
2. Persist the dataset into a JASON file.

* Write the content of the DataFrame object into a JASON file. This will convert the dataset into a JASON format. You can decide which JASON format you like to convert.

1. Calculate the mean and standard deviation for every (non-categorical) column using DataFrame methods.
2. Use DataFrame Data Visualization methods to draw either the Boxplot to display the distribution function for each column of the DataFrame object. Please compare the curves generated and determined which columns have distribution functions of similar shape.
3. Use the DataFrame method to calculate the correlation coefficient between any two (continuous) columns. Also draw the Scatter Plots to demonstrate how any two columns are correlated. Use the coefficient coefficients and Scatter Plots to determine if any two columns are *positively correlated*, *negatively correlated* or *not correlated*.
4. Use the ***class*** column to group the records in the dataset and repeat step 3 and 4 for all groups.
5. Use your own words to describe the findings discovered from the above analysis.

The deliverables –

* A document summarizing your findings
* Your Python scripts