CS 620–Introduction to Data Science and Analytics, HW3

In this problem, you'll get practice working with pandas DataFrames, reading them into and out of memory, changing their contents, and performing aggregation operations. For this problem, you'll need to download the celebrated iris data set, available as a .csv file. Note: for the sake of consistency, please use this version of the CSV, and not one from elsewhere.

1. (20 pts) Download the iris data set from the link above. Please include this file in your submission. Read iris.csv into Python as a pandas DataFrame. Note that the CSV file includes column headers. Answer the following questions (print statements in your code) by computing the results.
    a. How many data points are there in this data set?
    b. What are the data types of the columns?
    c. What are the column names? The column names correspond to flower species names, as well as four basic measurements one can make of a flower: the width and length of its petals and the width and length of its sepal (the part of the pant that supports and protects the flower itself).
    d. How many species of flower are included in the data?

2. (10 pts) Use the Seaborn and Matplotlib to generate a scatter matrix plot (pair plot) for the four features (petal width, petal length, sepal width, and sepal length). Each point in the scatter plot should be colored according to its species. Include this plot file in your submission.
3. (20 pts) The iris dataset is commonly used in machine learning as a proving ground for clustering and classification algorithms. Some researchers have found it useful to use two additional features, called Petal ratio and Sepal ratio, defined as the ratio of the petal length to petal width and the ratio of the sepal length to sepal width, respectively. Add two columns to your DataFrame corresponding to these two new features. Name these columns PetalRatio and SepalRatio, respectively.
4. (10 pts) Save your extended iris DataFrame to a CSV file called iris_extended.csv. Please include this file in your submission.
5. (20 pts) Use a pandas aggregate operation to determine the mean, standard deviation, minimum, and maximum of the petal and sepal ratio for each of the three species in the data set. Note: you should be able to get all of these numbers in a single table using a well-chosen group-by or aggregate operation.
6. **(Optional for Data Science Majors)** (20 pts) Use the Seaborn and Matplotlib to generate a plot showing the distribution of petal ratio and sepal ratio for each of the three species. Your plot should have two subplots, one for the petal ratio and one for the sepal ratio. You may choose the details of your plots (i.e., how to handle outliers, displaying mean vs median, etc) however you think is best. Please include labels on your x- and y-axes and give an appropriate title to your plot and subplots. Include this plot file in your submission.

What to turn in: Submit your Lastname-hw3.zip file to Canvas. **Lastname-hw3.zip** should contain the plot(s), dataset files(irs.csv, irs_extended.csv) and the Lastname-hw3.py which should include the header:

CS620
HW3
@author: <Your Name and UIN>