# CSCI 2410 Introduction to Data Analytics Using Python Homework Assignment #2

**HW Programming #2: - Python programming**

Tasks: Read and write data files in Python, perform simple searches on the dataset, obtain simple statistics and make two dimensional plots of selective data contents.

Assignment Instructions:

1. **[15%]** Read in the ‘iris.xlsx’ dataset from your computer’s local directory (You need to first download the file from the Canvas of this class), then display the data content of specific row (or rows), column (or columns), and cell (or cells).

Python libraries needed: “pandas”

Python libraries “xlrd” and “openpyxl” are needed for dealing with Excel files However, these library do not need to be explicitly imported in your .py script.



A screen shot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

2**. [30%]** (2.1) Search the data content by giving the row and column numbers (Note: the row and A screenshot of a computer code

Description automatically generatedA screen shot of a computer code

Description automatically generatedA computer screen shot of a code

Description automatically generatedA computer screen shot of a program code

Description automatically generatedA computer screen shot of a program code

Description automatically generated column numbers of dataset start with 0 in Python).

(2.2) Search the dataset for a specifically giving value, say “3.3”, in the content; then modify the value by replacing it with a string ‘NaN’ in the dataset; count and display the number of times the value A screenshot of a computer program

Description automatically generatedappears (so is replaced) in the dataset.

A screenshot of a computer

Description automatically generated

(2.3) Write the modified dataset to a new .xlsx file (be sure to use a new file name).

A computer screen shot of a black background

Description automatically generatedA computer screen shot of text

Description automatically generatedA screenshot of a computer

Description automatically generated Python libraries needed: NumPy, pandas.

(The loop statement, conditional statement, and logic operators are needed in these scripts) Note: step 2 requires doing a search on the dataset (2.1 index search, 2.2 keyword search).

A screenshot of a computer

Description automatically generatedA computer screen shot of text

Description automatically generatedA computer screen shot of a program

Description automatically generatedA computer screen shot of a program code

Description automatically generated3. **[25%]** (3.1) Obtain the basic statistical parameters (‘mean’, ‘median’, ‘standard deviation’, ‘variance’, ‘correlation’) by applying corresponding ‘pandas’ functions **on each of** the three. ‘Species’ of iris in the dataset.

(3.2) Find the 5 minimum values and the 5 maximum values for each of the data attributes, i.e., the first 4 columns) of the dataset. Print out these values **along with** each of their associated ‘Species.’ (i.e., print in pairs of ‘value’ – ‘Species’)

A screenshot of a computer

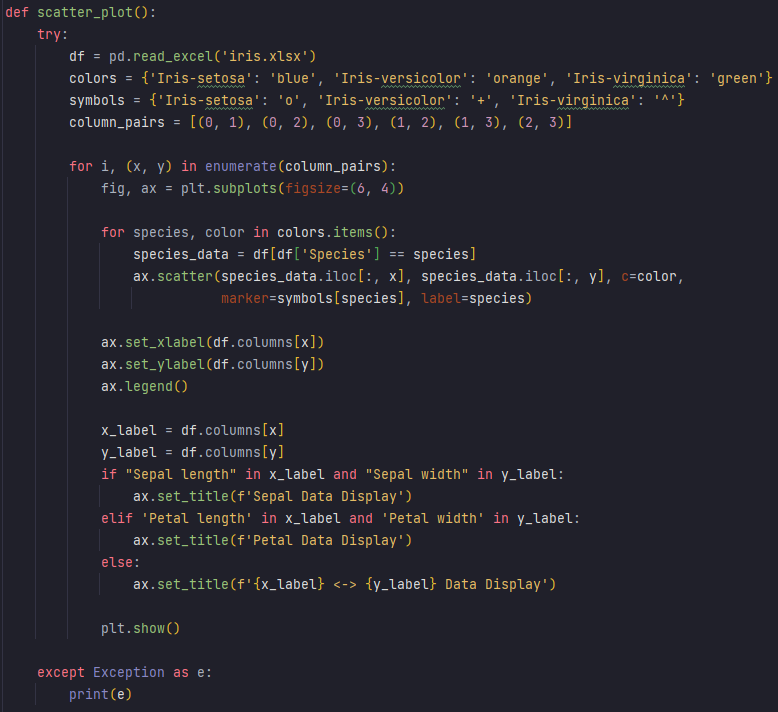
Description automatically generatedA computer screen shot of a program

Description automatically generatedA computer screen shot of a program code

Description automatically generatedPython libraries needed: panda

**4. [30%]** Plot in two dimension the dataset in pairs of the data attributes (i.e., the first 4 columns).

You are asked to plot the data attributes in pairs, such as column 1 – column 2, column 1 – column 3, …, i.e., two attributes a time due to the limitation of two-dimensional plots. Plot the data points of different classes (categories of Iris) in different colors and different marks or symbols.

Python libraries needed: pandas, matplotlib.pyplot

A diagram of different colored dots

Description automatically generatedA chart of different colored dots and numbers

Description automatically generated with medium confidence

A chart of different sizes and colors

Description automatically generatedA chart with green and orange dots

Description automatically generated

A graph of different sizes and colors

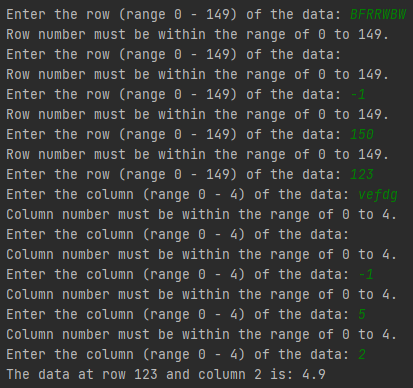
Description automatically generatedA chart with different colored dots

Description automatically generated

Test Case 1: Input Validation for Searching Specific Data.

An empty input for row number.

User input "BFRRWBW" (not a valid integer) resulted in a response indicating that the row number must be within the range of 0 to 149.



Input "-1" for the column (below the valid range)

Input "150" for row (above the valid range)

Input "-1" for row (below the valid range) produced a response indicating the valid range.

An empty input for the column produced the same response indicating the valid column range.

User input "123" for the row and "vefdg" (not a valid integer) for the column resulted in a response indicating that the column number must be within the range of 0 to 4.

Input "5" for the column (above the valid range)

Finally, when valid inputs "123" for the row and "2" for the column were provided, the program successfully retrieved and displayed the data at row 123 and column 2, which was 4.9.

Input "5" for the sheet number, which is above the valid range, results in a response indicating that the sheet number must be within the range of 1 to 4.

Test Case 2: Input Validation for Searching 5 Min and Max Values

A screenshot of a computer

Description automatically generated

Input "3.4," which is not a valid integer

An empty input for the sheet number

Finally, when a valid input "2" for the sheet number is provided, the program successfully proceeds with the analysis. It displays the length (rows) of the dataset in "Sheet2" (which is 50 rows) and proceeds to find and display the 5 minimum and 5 maximum values for each of the four columns ("Sepal length," "Sepal width," "Petal length," and "Petal width") within "Sheet2."

Input "three," which is not a valid integer, results in a response indicating that the sheet number must be within the range of 1 to 4.

Input "-1" for the sheet number, which is below the valid range.

Test Case 3: Input Validation for Modifying Data in a Dataset

Input "-0.2," which is below the valid range, also produces a response indicating the valid range.

Input "10.1," which exceeds the valid range, results in a response indicating that the float value should be between 0.0 and 10.0.

A screenshot of a computer

Description automatically generated

Input "-1" for the sheet number, which is below the valid range, results in a response indicating that the sheet number must be within the range of 1 to 4.

Following the modifications, the program displays the updated dataset, indicating that the "iris.xlsx" data has been modified and written to a file named "modified-iris.xlsx."

When a valid input "0.1" is provided, the program successfully identifies and displays six instances where the value 0.1 is found in the dataset, along with the corresponding row and column. It then changes each of these values to NaN. A total of 6 occurrences of the value 0.1 are found in the dataset.

Input "five," which is not a valid float, results in a response indicating that the float value should be between 0.0 and 10.0.

Test Case 4: Input Validation for Selecting the Dataset for Statistical Analysis

A screenshot of a computer

Description automatically generated

Input "1.5," which is not a valid integer.

An empty input for the sheet number.

Input "five," which is not a valid integer.

Input "5" for the sheet number, which exceeds the valid range.

Finally, when a valid input "2" for the sheet number is provided, the program proceeds with the analysis of "Sheet2." It displays the length (rows) of the dataset in "Sheet2" (which is 50 rows) and calculates various statistics, including means, variances, standard deviations, and correlation coefficients for the columns in "Sheet2."

# Requirements for the Submission of Programming/Homework Assignments

1. Well-documented program list (the .py files)

**20% of total points** if no .py file submitted.

Done.

1. Three annotated program test and run examples (screenshots) that **show different and representative test cases** with **input, output, and the parameter settings of the program runs clearly marked/annotated**. You can do the annotations by
   1. Pasting the screenshots into a WORD document,

Done

* 1. Editing on the WORD document pages for the required marks and annotations,

Done

* 1. Converting the document to pdf for submission (it is ok to submit the WORD file directly without converting to pdf).

Done

**20% of total points** will be taken off if run examples are not representative.

**20% of total points** will be taken off if run examples are not clearly marked/annotated.

1. A discussion page.
2. Hardware and software used by your program.

I completed this assignment using my personal computer with PyCharm Professional Version: 2023.2.1.

1. Features of your program, e.g., data structures, algorithms, programming styles, etc.

The program script is designed to interact with an Excel dataset named "iris.xlsx," which contains multiple sheets representing iris flower data. It includes functions for user input validation, such as validating row and column numbers and selecting a specific sheet from the dataset. Additionally, the program can search for specific values in the dataset, replace them with NaN if desired, and calculate various statistics (mean, median, standard deviation, and variance) for specified columns. It also computes correlation coefficients between columns. Furthermore, the program can find the 5 smallest and 5 largest values for selected columns and generate scatter plots to visualize relationships between different columns, using different colors and markers to distinguish iris species. Error handling is implemented throughout the program, and it reads data from "iris.xlsx" while saving modified data to a new file called "modified-iris.xlsx."

1. Problems you encountered during your work.

Generating the two dimensional plots was difficult.

1. Assigned discussion problems if there is any.

None

1. Fill in the following table and submit it along with your above submissions.

|  |  |  |  |
| --- | --- | --- | --- |
| Total (approximate) time spent on the assignment | 23 hours | Total (approximate) time for the correction part | 2 hours |
| Problems and difficulties encountered | Generating two dimensional plots. | | |
| Reflections (good and bad) on the assignment | Good: learned a lot from the modules  Bad: annotating the word file is time consuming . | | |
| Any comments and suggestions | None | | |

**20% of total points** will be taken off if no discussion page is submitted.