Date: 27/08/2022

Assignment No. 01

Title: Design the data analysis tool (GUI)

NAME: Saurabh Nagre

PRN: 2019BTECS00080

**Objective / Aim:** Calculate the measures of central tendency and dispersion of data, display the statistical description of data in graphical display.

**Introduction:** The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data Mining.

To analyze the data pattern, we use various statistical measures of dataset. In measures of central tendency we have mean, median, mode, midrange , variance and standard deviation and to analyze the dispersion of data we use range , quartiles , interquartile range, five-number summary.

**Theory / Algorithms:**

Mean: x̄ = Σfx/Σf  
where,

* x̄ = the mean value of the set of given data.
* f = frequency of each class
* x = mid-interval value of each class

Median:

For Odd number of observations, {(n+1)/2}th term

For Even number of observations, [(n/2)th term + {(n/2)+1}th]/2

where n is the number of observations

Mode:

The mode is the value that is repeatedly occurring in a given set. We can also say that the value or number in a data set, which has a high frequency or appears more frequently, is called mode or modal value

Midrange: It’s an average between highest and lowest value in data.

Variance:  Σ(Xi−¯X)2/N-1

Standard deviation: sqrt(Σ(Xi−¯X)2/N-1)

Range: It’s a difference between the highest and lowest value in data.

Quartiles:

It divides the data in three parts. To find the quartile we need to sort data. The first quartile will be at N/4, second at N/2 and third at 3N/4 position in data.

Interquartile range:

The 50% central data range is the interquartile range. The difference between the first and third quartile is the interquartile range.

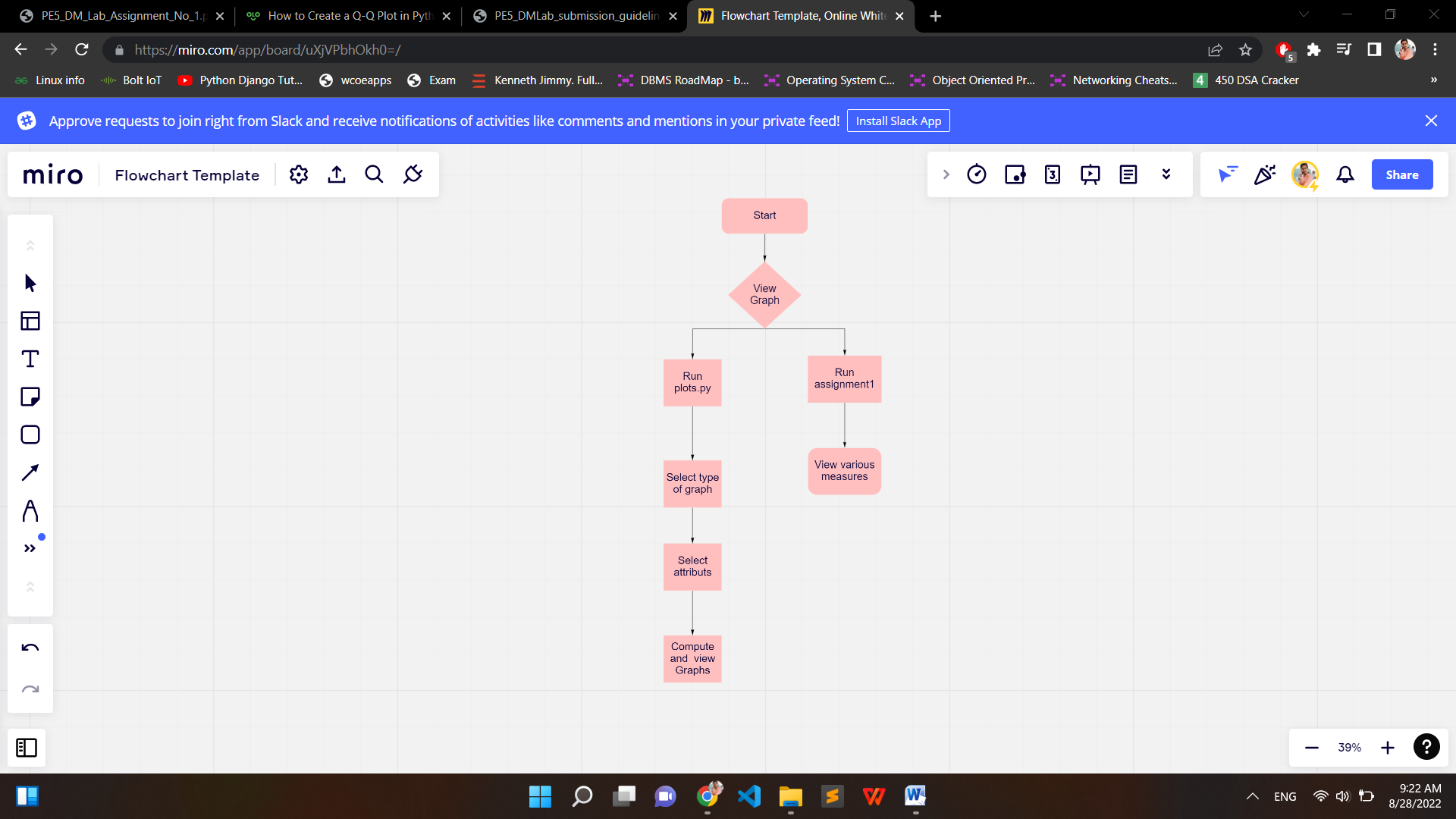
Five-number summary:

It gives the whole idea of data from the five numbers. The five numbers are minimum of data, first, second, third quartile and maximum of data.

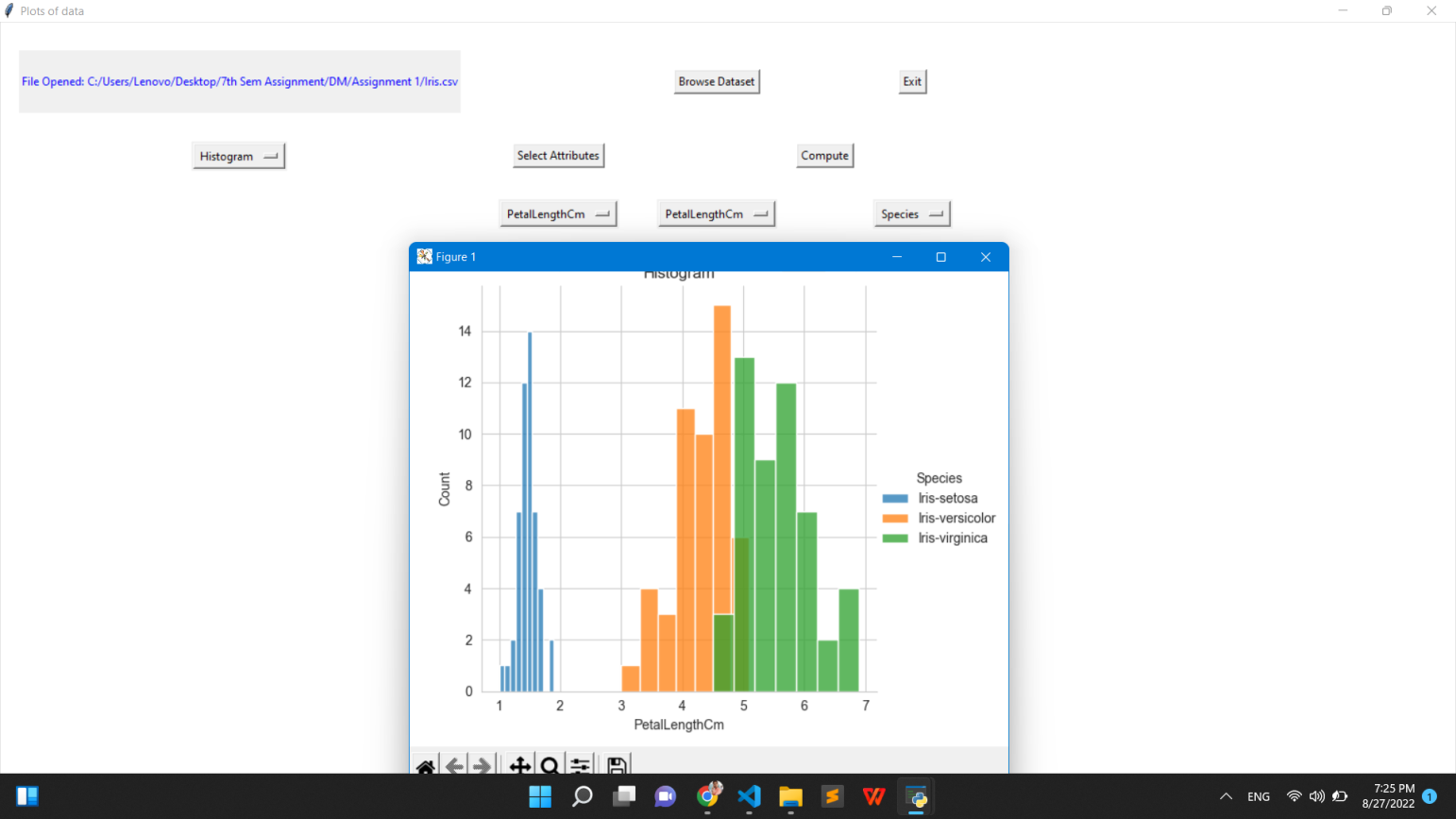
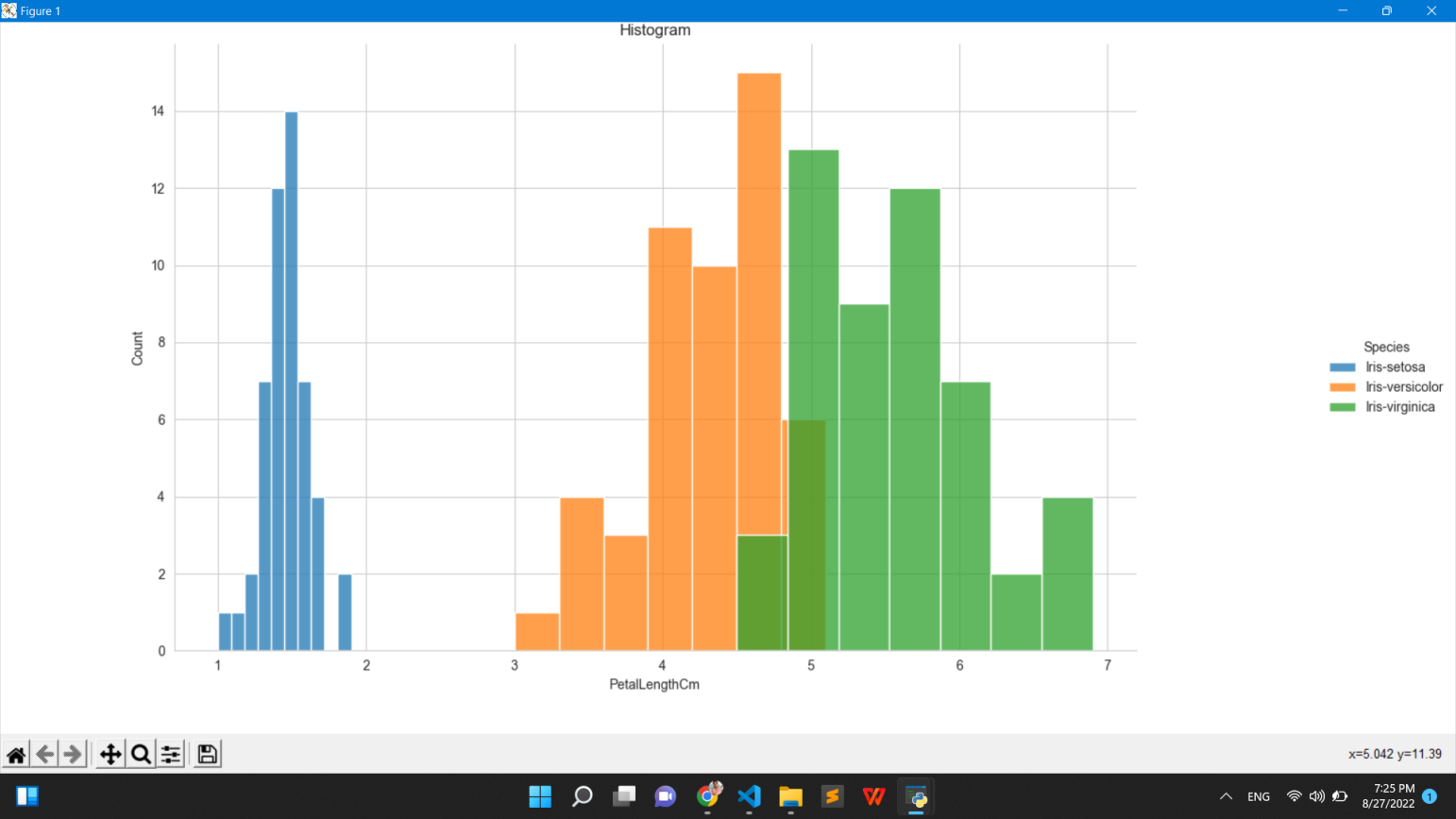
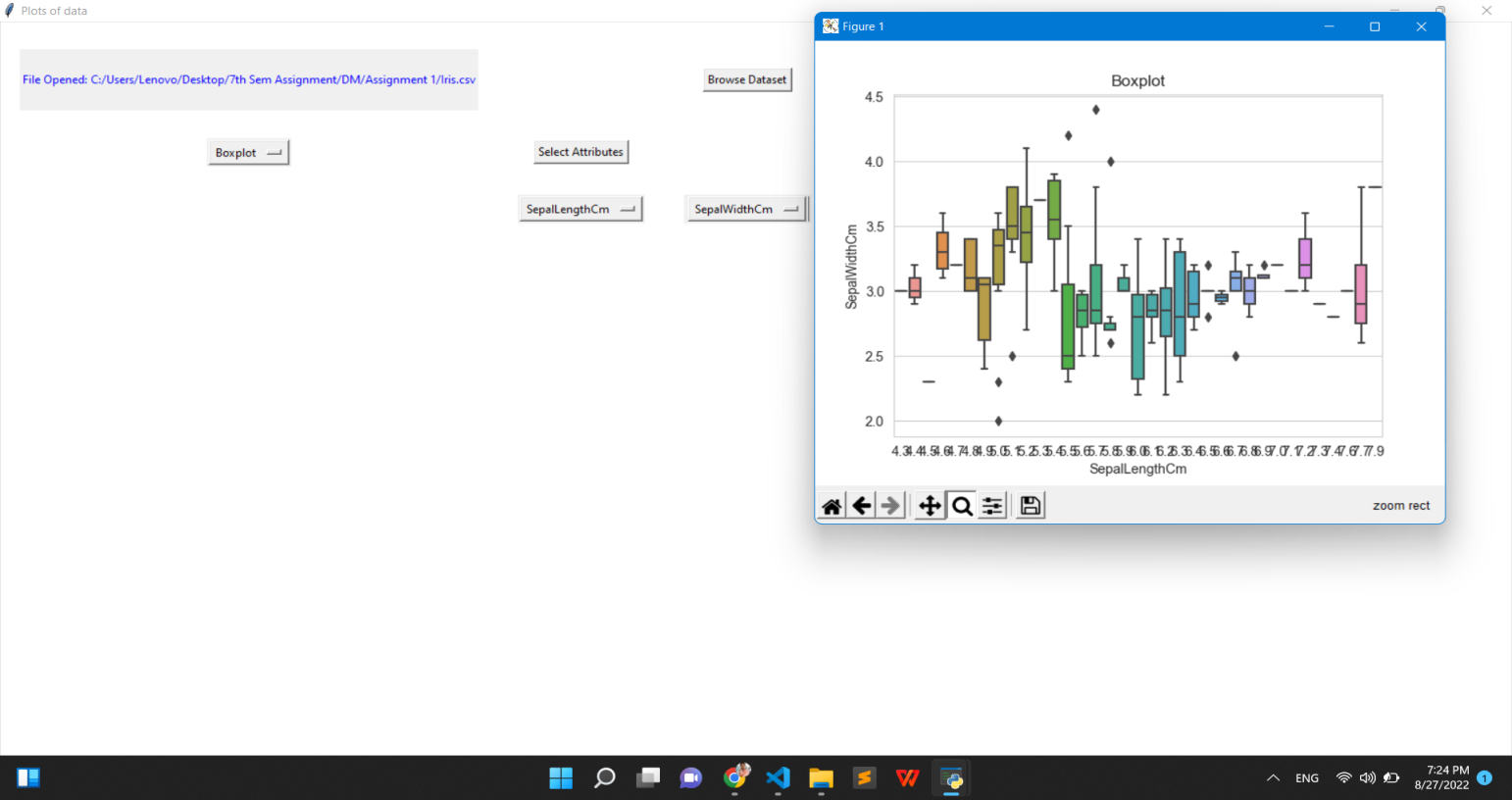
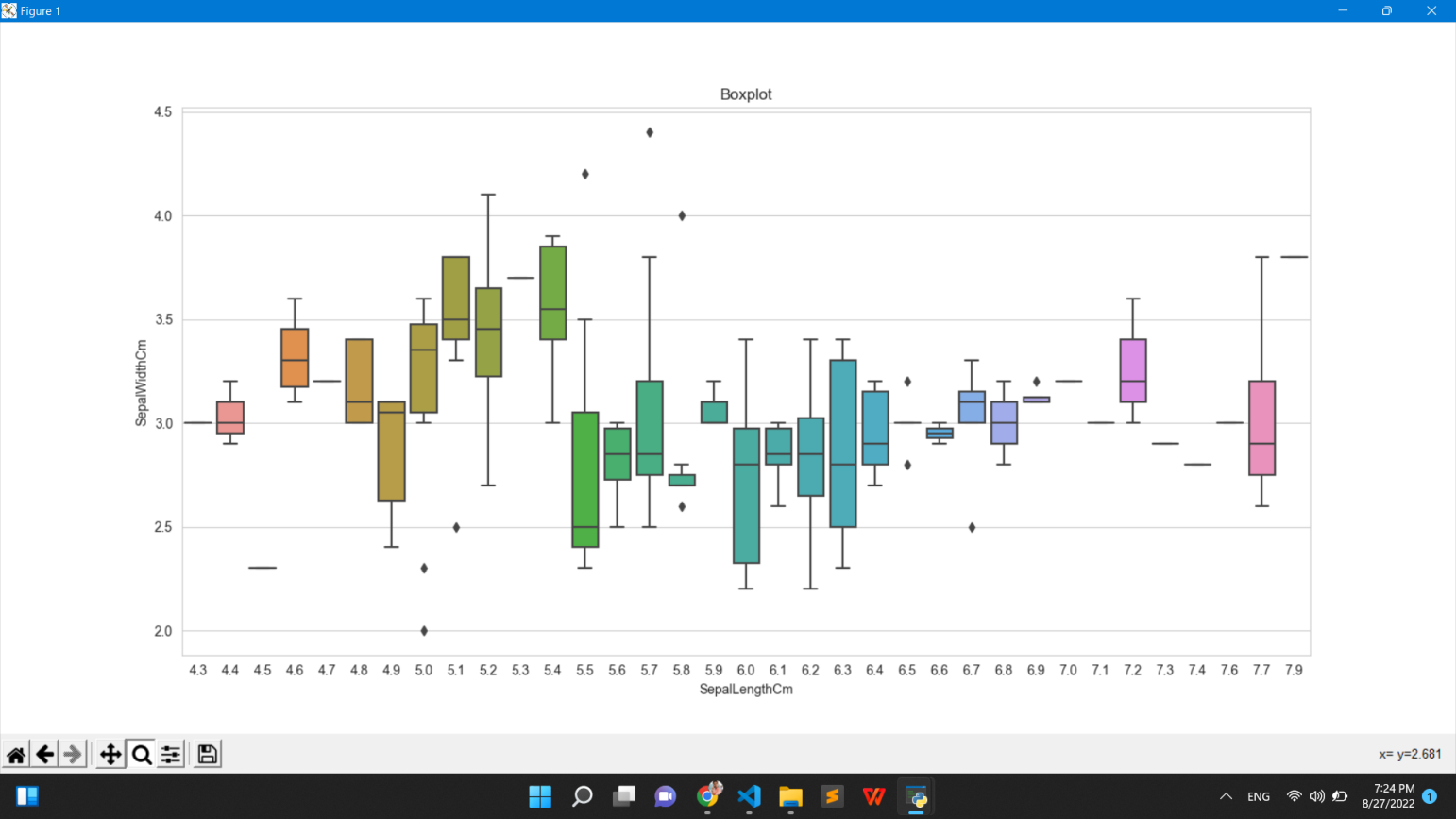
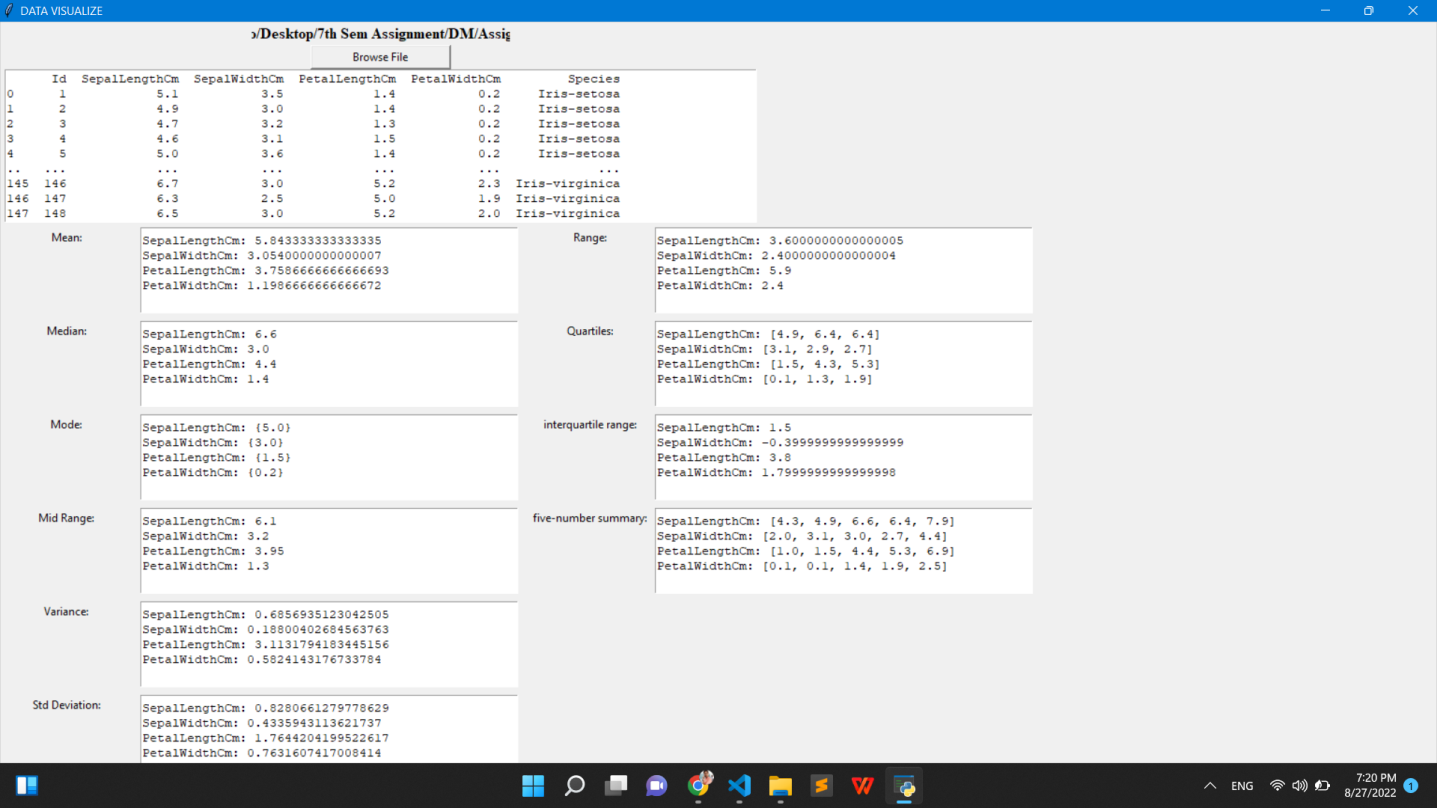
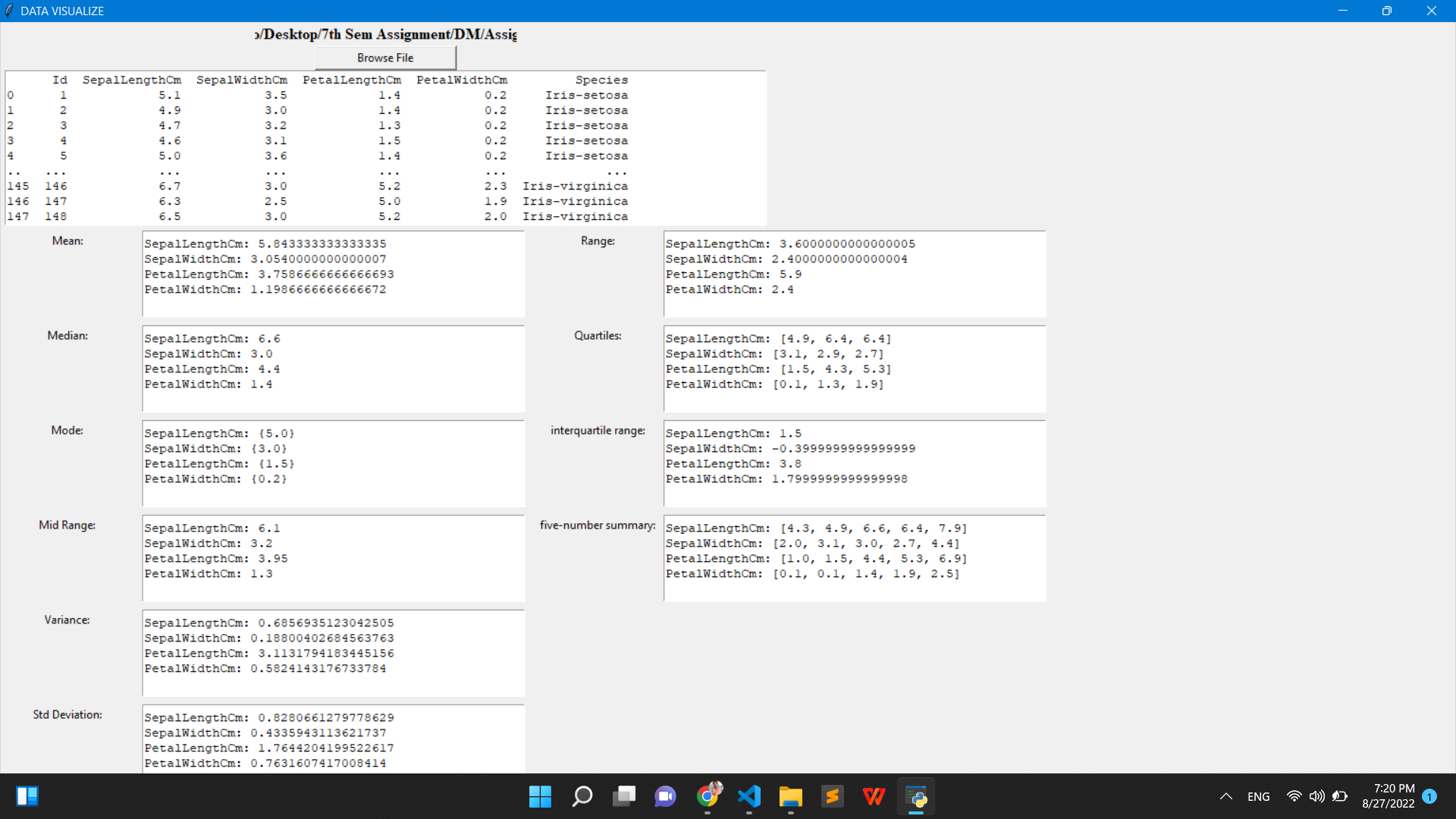
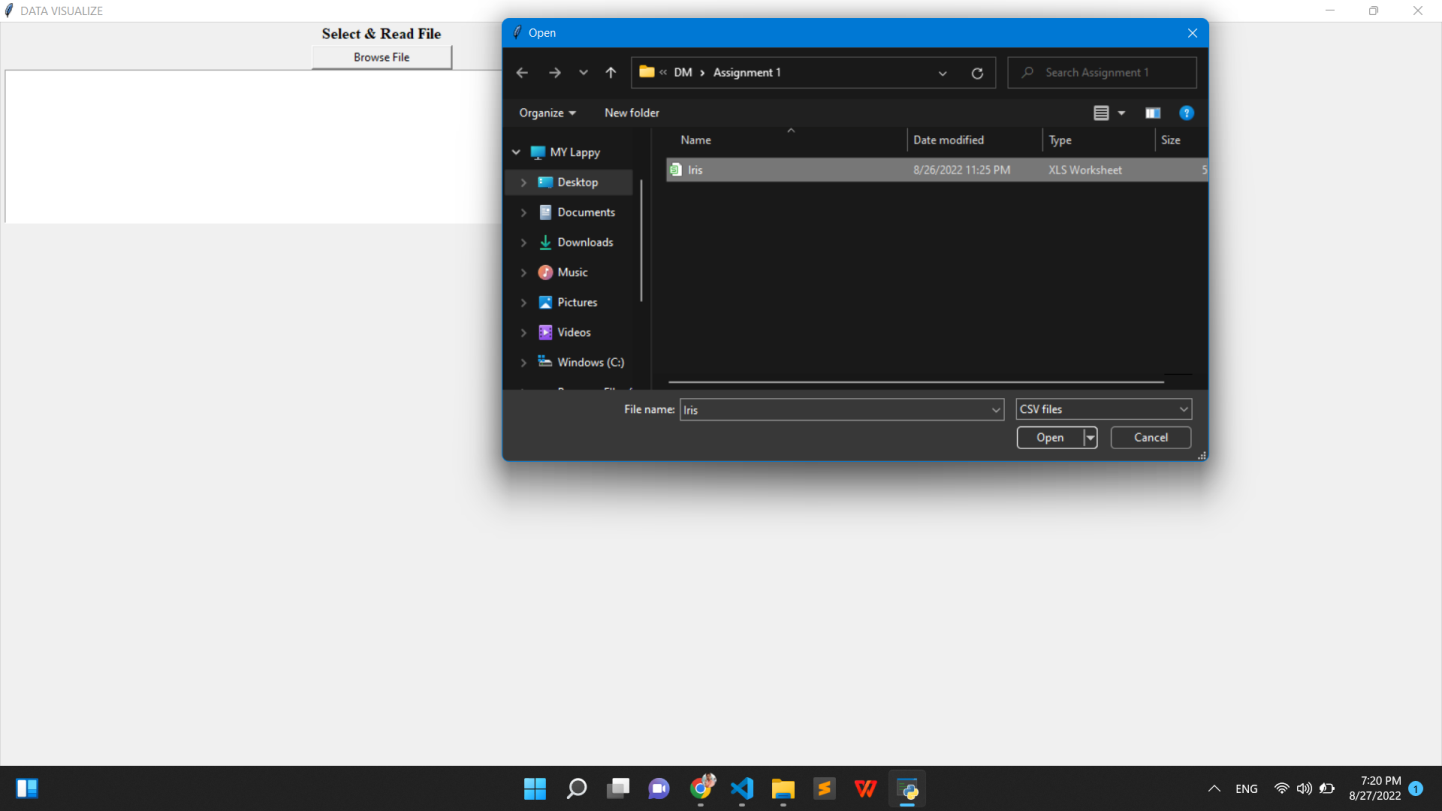
**Procedure:**

1. Select the dataset after executing the program.
2. The central tendency and dispersion of data calculations of each attribute is calculated without using in-build methods.
3. Graphs are generated after selecting the attributes.

**Functional block diagram / DFD:**



Screenshots:



Conclusion:

1. The dataset is balanced i.e. equal records are present for all three species.
2. We have four numerical columns while just one categorical column which in turn is our target column.
3. A strong correlation is present between petal width and petal length.
4. The setosa species is the most easily distinguishable because of its small feature size.

References:

1. <https://study.com/academy/lesson/quartiles-the-interquartile-range-definition-formulate-examples.html>
2. <http://www.statisticslectures.com/topics/fivenumbersummary/#:~:text=The%20five%20numbers%20are%20the,)%20value%2C%20and%20the%20maximum.&text=The%20first%20thing%20you%20might,outlier%20and%20must%20be%20removed>.
3. <https://www.statology.org/q-q-plot-python/>