Date: 09/09/2022

Assignment No. 02

Title: Design the pre-processing tool (GUI)

NAME: Saurabh Nagre

PRN: 2019BTECS00080

**Objective / Aim:** Calculate the Correlation analysis - Chi-Square test, Correlation coefficient (Pearson coefficient) & Covariance and Normalization of data in graphical display.

**Introduction**:

Correlation analysis in research is a statistical method used to measure the strength of the linear relationship between two variables and compute their association. Simply put - correlation analysis calculates the level of change in one variable due to the change in the other. A high correlation points to a strong relationship between the two variables, while a low correlation means that the variables are weakly related.

Normalization is used to scale the data of an attribute so that it falls in a smaller range, such as -1.0 to 1.0 or 0.0 to 1.0. It is generally useful for classification algorithms.

Methods of normalization are:

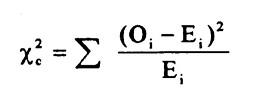
a. Min-max normalization

b. Z-Score normalization

c. Normalization by decimal scaling

**Theory / Algorithms:**

**Chi-Square test:**

[](https://www.statisticshowto.com/wp-content/uploads/2013/09/chi-square-formula.jpg)

The subscript “c” is the degrees of freedom. “O” is your observed value and E is your expected value.

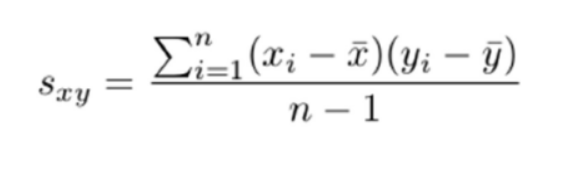
A low value for chi-square means there is a high correlation between your two sets of data.

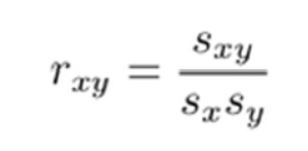
Degrees of freedom: That’s just the number of categories minus 1.

**Correlation Coefficient (Pearson Coefficient) and Covariance:**

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, i.e.:

* A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
* A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
* Zero means that for every increase, there isn’t a positive or negative increase. The two just aren’t related.



* Where, x̄ is the sample mean of x, ȳ is sample mean of y, xi and yi are the values of x and y for ith record in the sample and n is the no of records in the sample.
* Correlation between x and y can be calculated as follows:
* 
* Where sxy is the covariance between x and y, sx and sy is standard deviation of x and y respectively and rxy is correlation or Pearson coefficient.

**Normalization methods:**

1. Min-max normalization:

For all values vi of attribute, A (i = 1 … n), replace vi as,

vi = (vi-min(A))/(max(A)-min(A)).

1. Z-Score normalization:

For all values vi of attribute, A (i = 1 … n), replace vi as,

vi = (vi-mean(A))/Standard Deviation(A)

1. Normalization by decimal scaling:

For all values vi of attribute, A (i = 1 … n), replace vi as,

vi = vi /10j

where j is the smallest integer value such that max(A)/10j < 1.

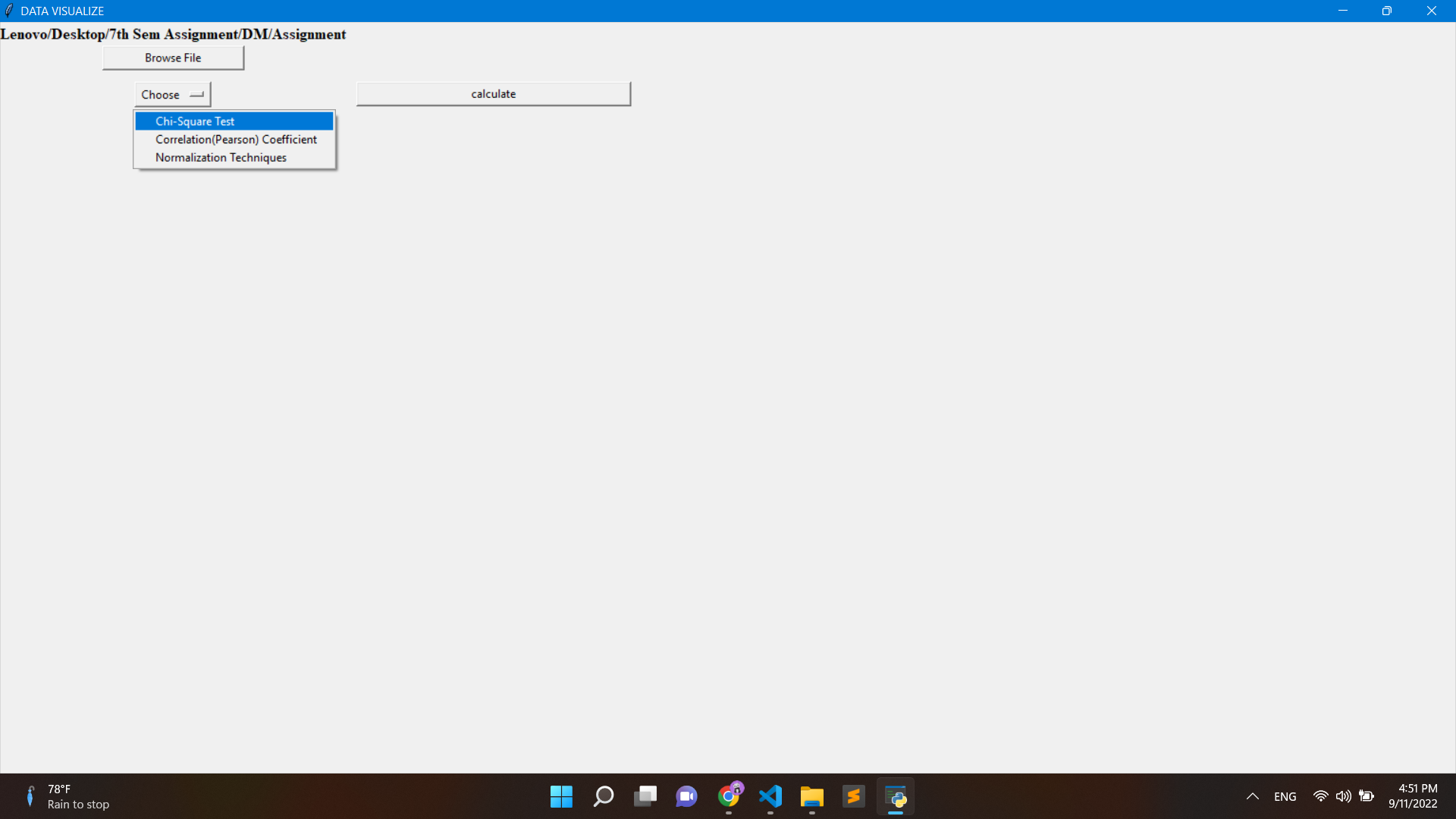
**Procedure:**

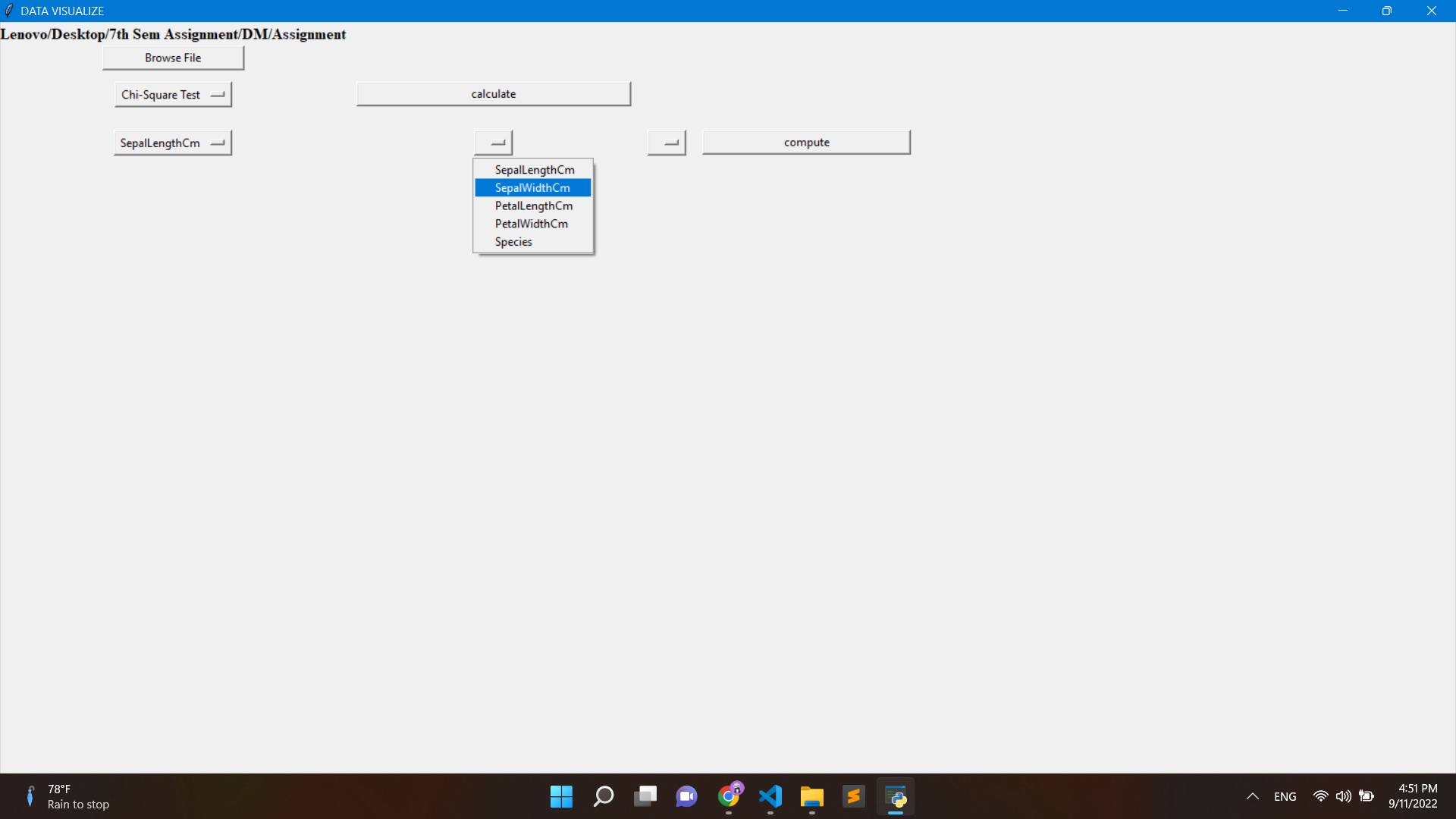
1. Use the python tkinter library, create a GUI having functionalities like selecting data file and after file selection, user should be able to select 2 attributes and target class.

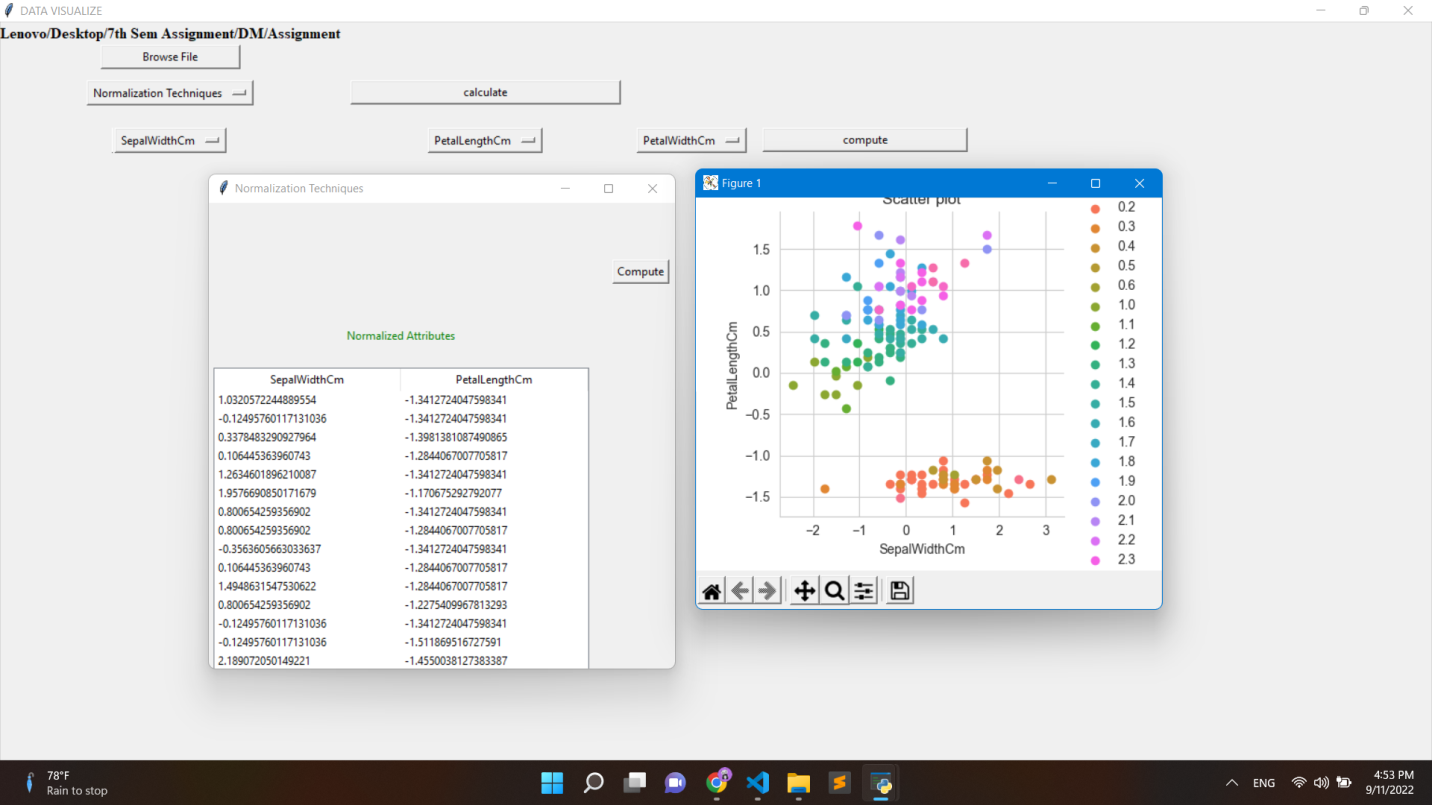
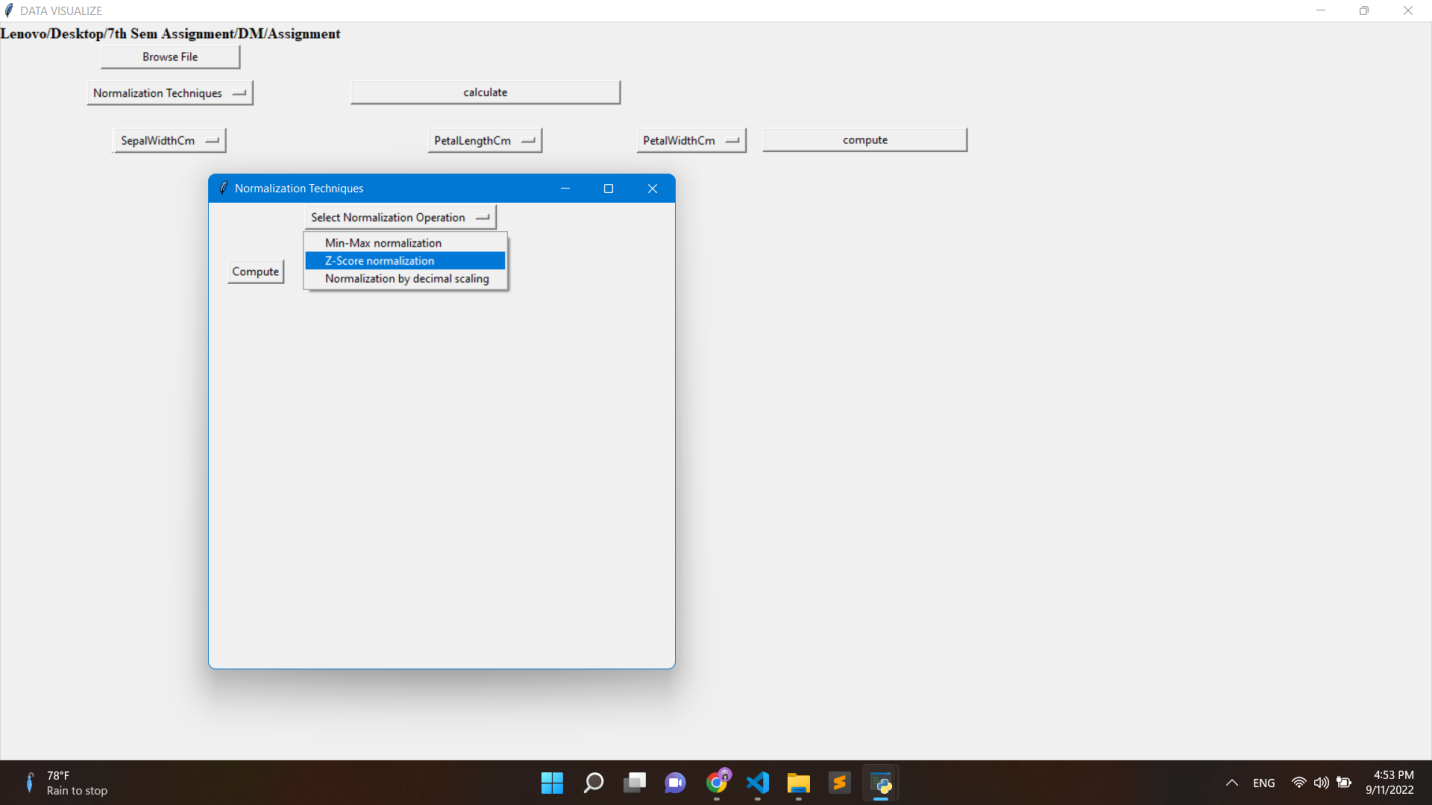
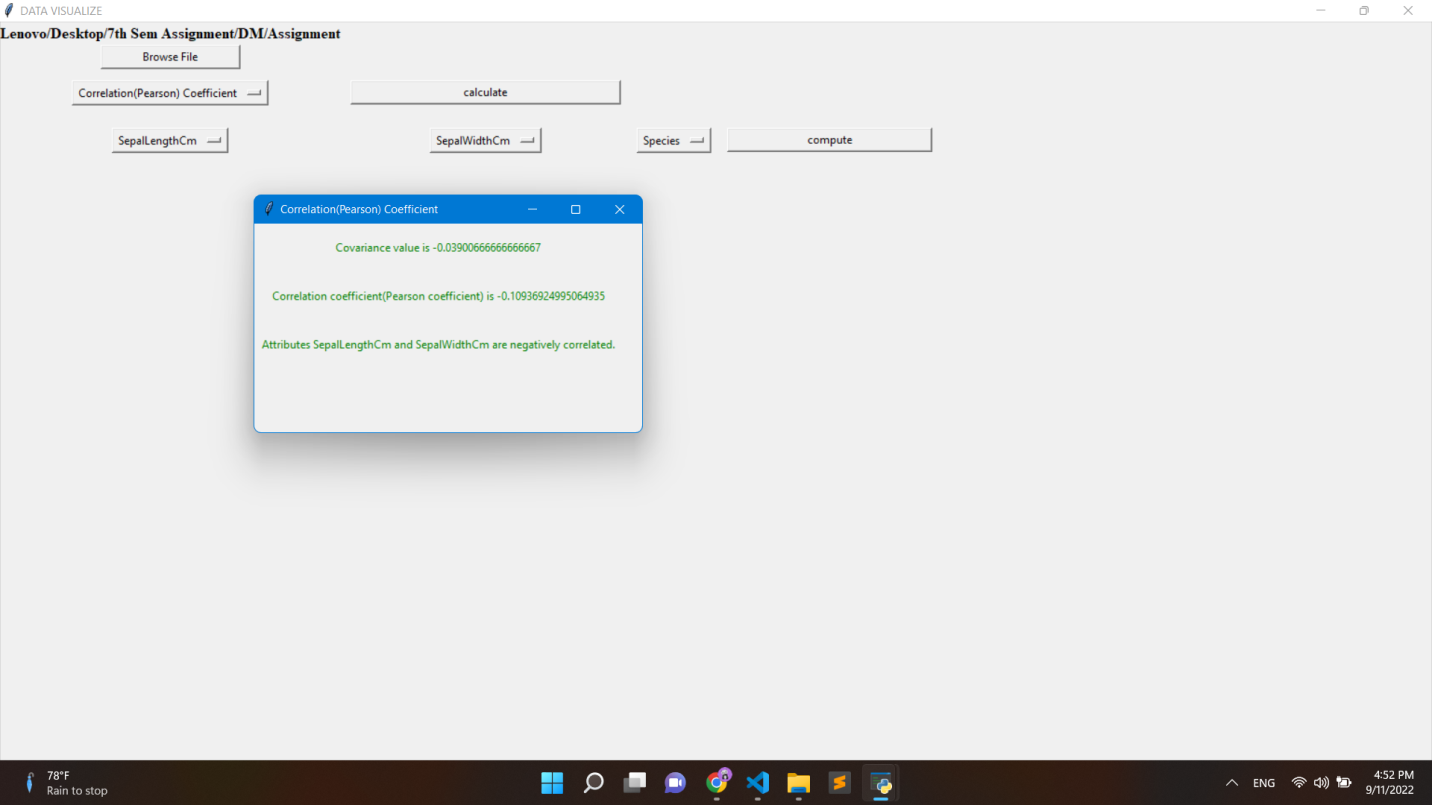
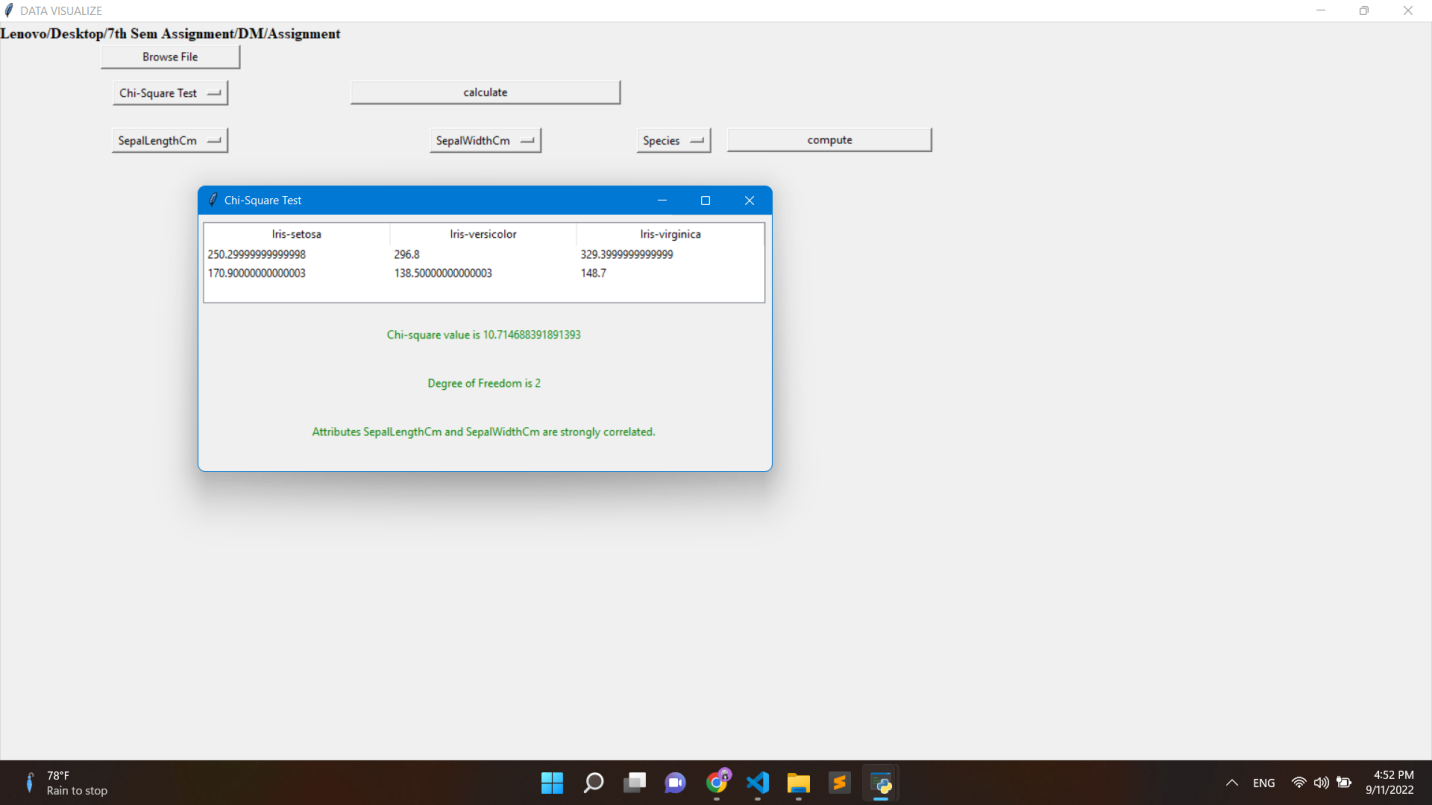
2. Apply appropriate formulae’s to user chosen method.

3. Display the result and appropriate graph to user.

**Screenshots**:







**Conclusion:**

We are able to find out relations between two attributes, if we have to reduce the dataset we can remove those columns which have less importance/effect on result.

**Reference:**

1. <https://www.statisticshowto.com/probability-and-statistics/chi-square/>
2. <https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/>