Final Year B. Tech. (CSE) – I : 2021-22
4CS462 : PE2 - Data Mining Lab
Assignment No. 3

Group id: DM21G12
Group members:
Abhishek More(2018BTECS00037)
Sushil Wagh(2018BTECS00031)

Title : Use/extend the data analysis tool (GUI) developed in Assignment No. 2

Objective/Aim :
perform the following pre-processing task :

1. Correlation analysis - Chi-Square Test
   a. User should be able to choose any two attributes.
   b. Display the contingency table
   c. Show the chi-square value and conclusion whether the selected attributes are correlated or not.

2. Correlation analysis – Correlation coefficient (Pearson coefficient) & Covariance
   a) User should be able to choose any two attributes.
   b) Show the calculated values and conclusion whether the selected attributes are correlated or not.

3. Normalization using following techniques :
   a. Min-max normalization
   b. Z-Score normalization
   c. Normalization by decimal scaling

Introduction:

Correlation Analysis is statistical method that is used to discover if there is a relationship between two variables/datasets, and how strong that relationship may be.

In terms of market research this means that, correlation analysis is used to analyse quantitative data gathered from research methods such as surveys and polls, to identify whether there is any significant connections, patterns, or trends between the two.

Essentially, correlation analysis is used for spotting patterns within datasets. A positive correlation result means that both variables increase in relation to each other, while a negative correlation means that as one variable decreases, the other increases.

Theory/Algorithms:

1. Correlation analysis - Chi-Square Test

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

The subscript “c” is the degrees of freedom. “O” is your observed value and E is your expected value. It’s very rare that you’ll want to actually use this formula to find a critical chi-square value by hand. The summation symbol means that you’ll have to perform a calculation for every single data item in your data set. As you can probably imagine, the calculations can get very, very, lengthy and tedious.

There are a few variations on the chi-square statistic. Which one you use depends upon how you collected the data and which hypothesis is being tested. However, all of the variations use the same idea, which is that you are comparing your expected values with the values you actually collect. One of the most common forms can be used for contingency tables:

$$c^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

Where O is the observed value, E is the expected value and “i” is the “ith” position in the contingency table.

A low value for chi-square means there is a high correlation between your two sets of data. In theory, if your observed and expected values were equal (“no difference”) then chi-square would be zero — an event that is unlikely to happen in real life. Deciding whether a chi-square test statistic is large enough to indicate a statistically significant difference isn’t as easy it seems.

2. What Is the Correlation Coefficient?

The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement. A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. A correlation of 0.0 shows no linear relationship between the movement of the two variables.

Correlation statistics can be used in finance and investing. For example, a correlation coefficient could be calculated to determine the level of correlation between the price of crude oil and the stock price of an oil-producing company, such as Exxon Mobil Corporation. Since oil companies earn greater profits as oil prices rise, the correlation between the two variables is highly positive.

2.1 Understanding the Correlation Coefficient :

There are several types of correlation coefficients, but the one that is most common is the Pearson correlation (r). This measures the strength and direction of the linear relationship between two variables. It cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

A value of exactly 1.0 means there is a perfect positive relationship between the two variables. For a positive increase in one variable, there is also a positive increase in the second variable. A value of -1.0 means there is a perfect negative relationship between the two variables. This shows that the variables move in opposite directions—for a positive increase in one variable, there is a decrease in the second variable. If the correlation between two variables is 0, there is no linear relationship between them.

The strength of the relationship varies in degree based on the value of the correlation coefficient. For example, a value of 0.2 shows there is a positive correlation between two variables, but it is weak and likely unimportant. Analysts in some fields of study do not consider correlations important until the value surpasses at least 0.8. However, a correlation coefficient with an absolute value of 0.9 or greater would represent a very strong relationship.

2.2 KEY TAKEAWAYS :

- Correlation coefficients are used to measure the strength of the relationship between two variables.
- Pearson correlation is the one most commonly used in statistics. This measures the strength and direction of a linear relationship between two variables.
- Values always range between -1 (strong negative relationship) and +1 (strong positive relationship). Values at or close to zero imply a weak or no linear relationship.
- Correlation coefficient values less than +0.8 or greater than -0.8 are not considered significant.

3. Normalization :

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.

3.1 Need of Normalization –

Normalization is generally required when we are dealing with attributes on a different scale, otherwise, it may lead to a dilution in effectiveness of an important equally important attribute(on lower scale) because of other attribute having values on larger scale.
In simple words, when multiple attributes are there but attributes have values on different scales, this may lead to poor data models while performing data mining operations. So they are normalized to bring all the attributes on the same scale.

3.2 Methods of Data Normalization –

1. Decimal Scaling
2. Min-Max Normalization
3. z-Score Normalization(zero-mean Normalization)

Procedure:

Pearson's correlation coefficient formula

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

Covariance:

$$cov_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

$cov_{x,y}$ = covariance between variable a and y

x_i = data value of x

y_i = data value of y

$\bar{x}$ = mean of x

$\bar{y}$ = mean of y

N = number of data values

Decimal Scaling Method For Normalization –

It normalizes by moving the decimal point of values of the data. To normalize the data by this technique, we divide each value of the data by the maximum absolute value of data. The data value, vi, of data is normalized to vi' by using the formula below –

$$v_i{}' = \frac{v_i}{10^j}$$

where j is the smallest integer such that max(|vi‘|)<1.

Let the input data is: -10, 201, 301, -401, 501, 601, 701

To normalize the above data,
Step 1: Maximum absolute value in given data(m): 701
Step 2: Divide the given data by 1000 (i.e j=3)

Result: The normalized data is: -0.01, 0.201, 0.301, -0.401, 0.501, 0.601, 0.701

Min-Max Normalization –

In this technique of data normalization, linear transformation is performed on the original data. Minimum and maximum value from data is fetched and each value is replaced according to the following formula.

$$\mathbf{v}' = \frac{\mathbf{v} - \mathbf{min(A)}}{\mathbf{max(A)} - \mathbf{min(A)}}(\mathbf{new\_max(A)} - \mathbf{new\_min(A)}) + \mathbf{new\_min(A)}$$

Where A is the attribute data,
Min(A), Max(A) are the minimum and maximum absolute value of A respectively.
v’ is the new value of each entry in data.
v is the old value of each entry in data.
new_max(A), new_min(A) is the max and min value of the range(i.e boundary value of range required) respectively

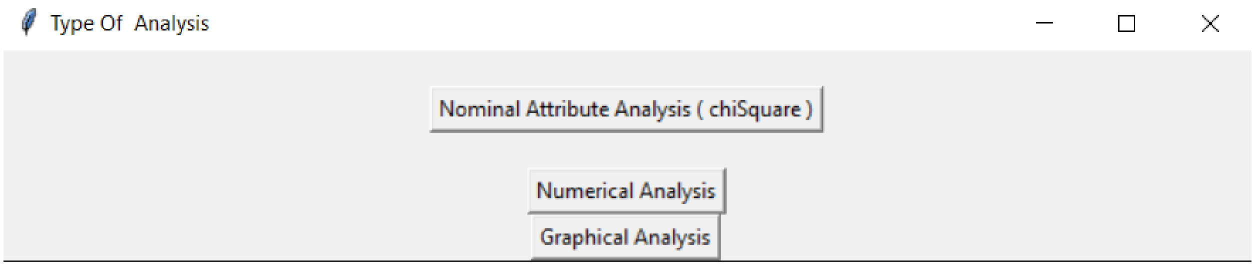
Z-score normalization –

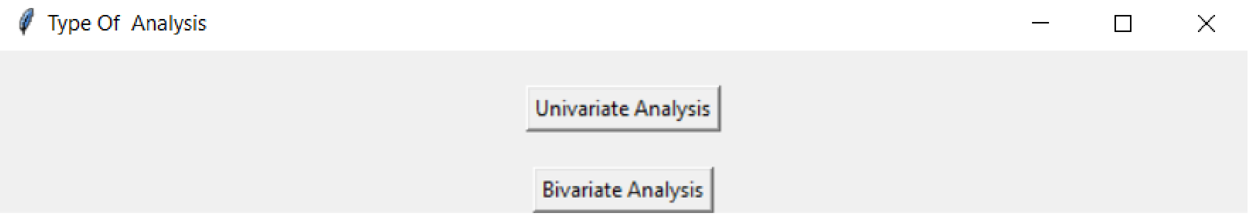
In this technique, values are normalized based on mean and standard deviation of the data A. The formula used is:

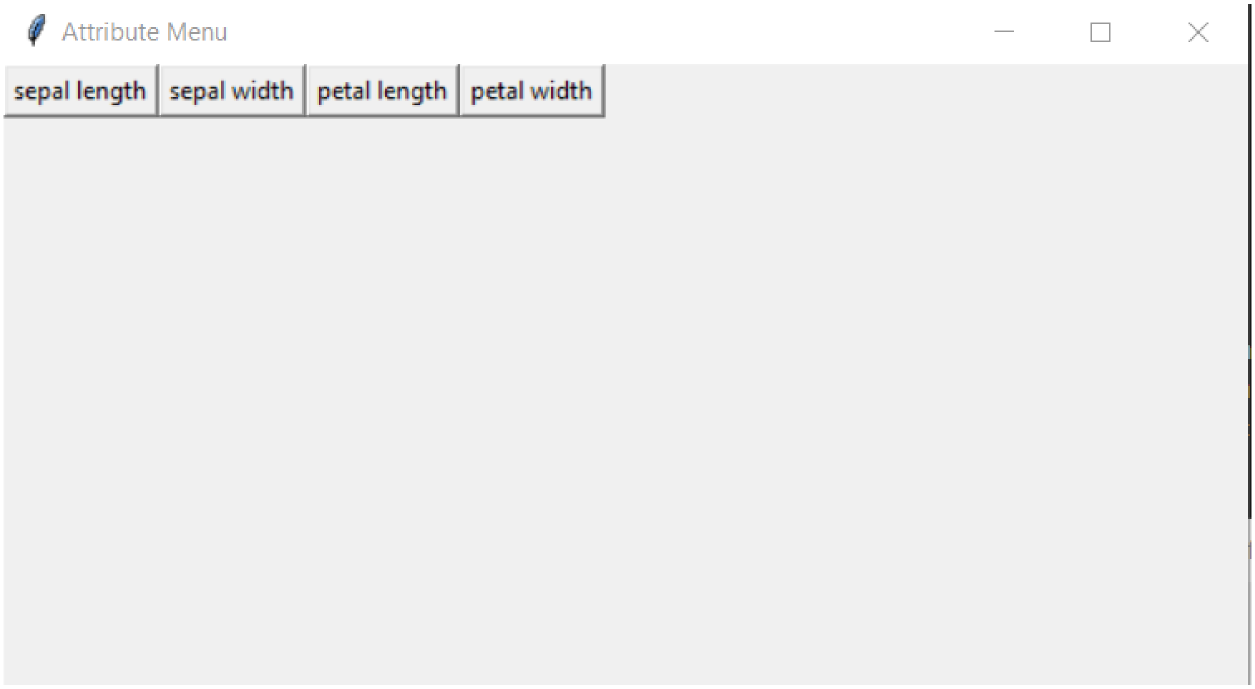
$$v' = \frac{v - \overline{A}}{\sigma_A}$$

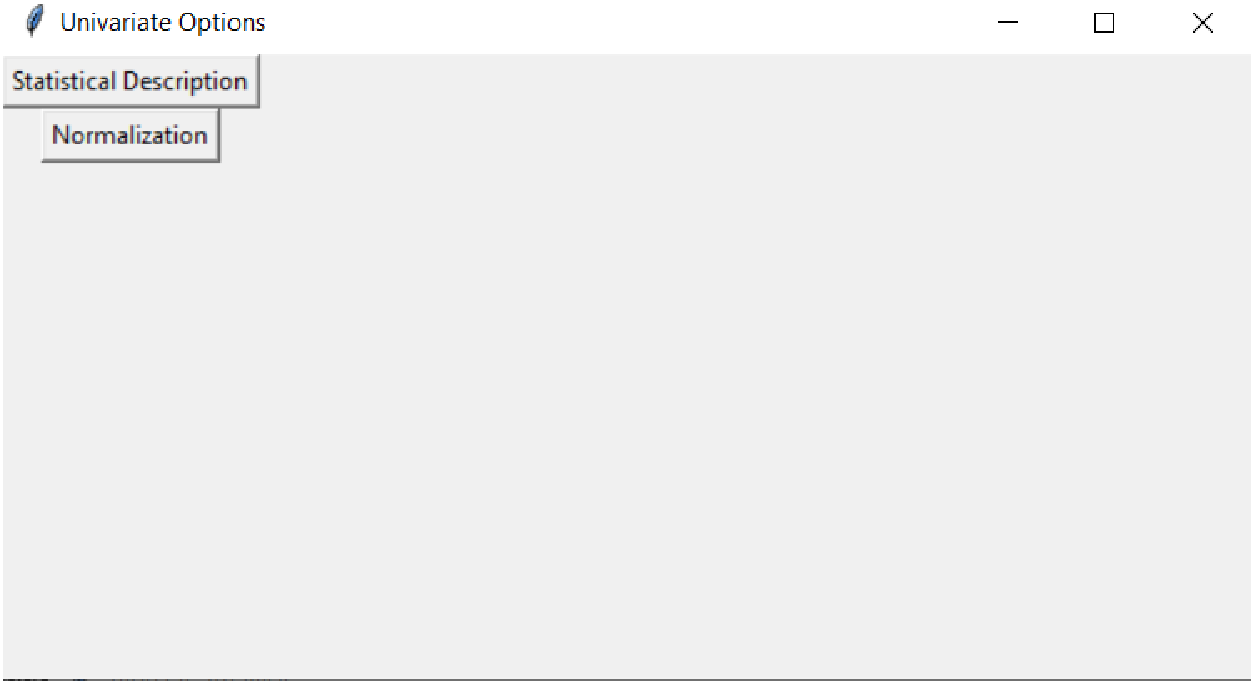
v’, v is the new and old of each entry in data respectively. σA, A is the standard deviation and mean of A respectively.

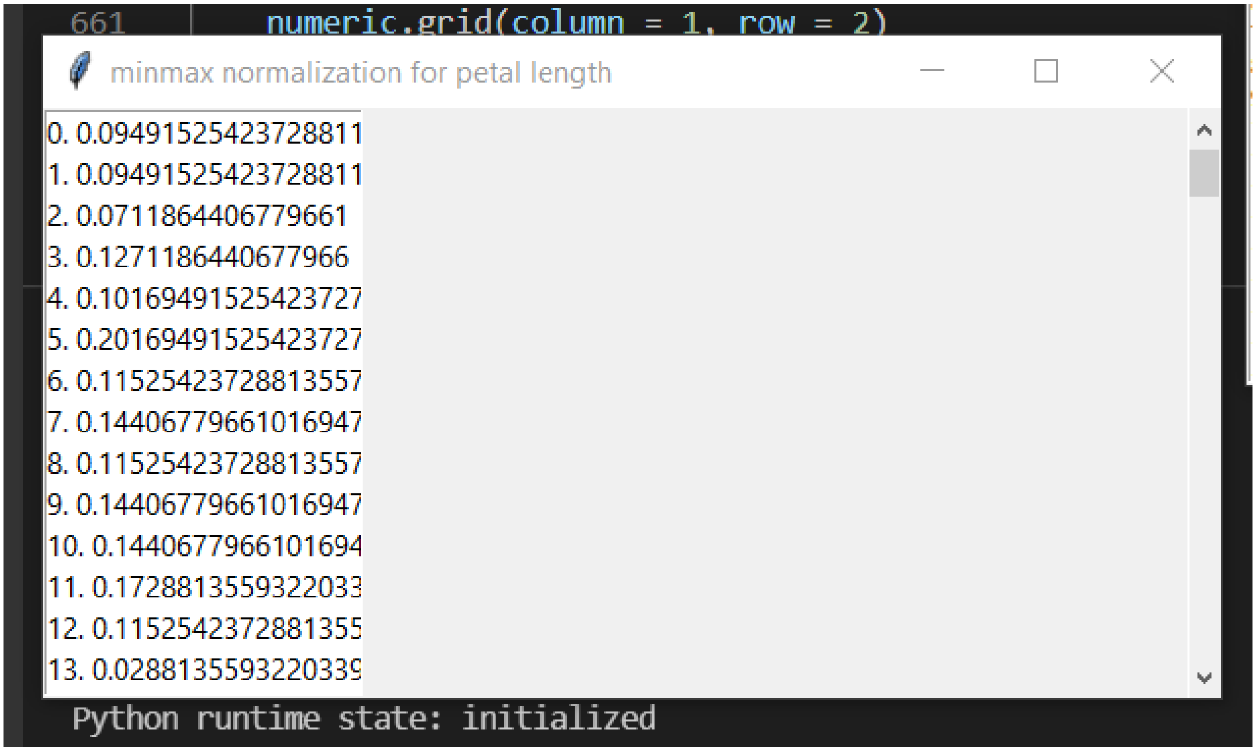
Result/Observations/Screenshots:


Type Of Analysis
Nominal Attribute Analysis (chiSquare)
Numerical Analysis
Graphical Analysis


Type Of Analysis
Univariate Analysis
Bivariate Analysis


Attribute Menu
sepal length
sepal width
petal length
petal width

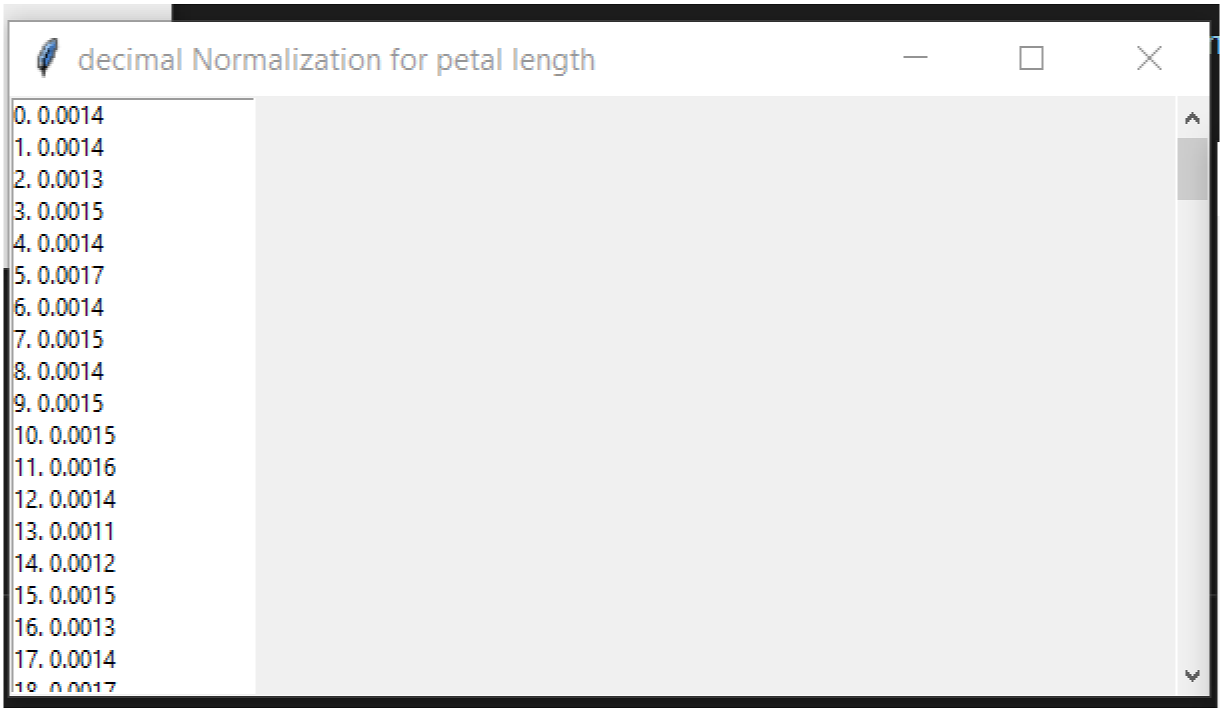




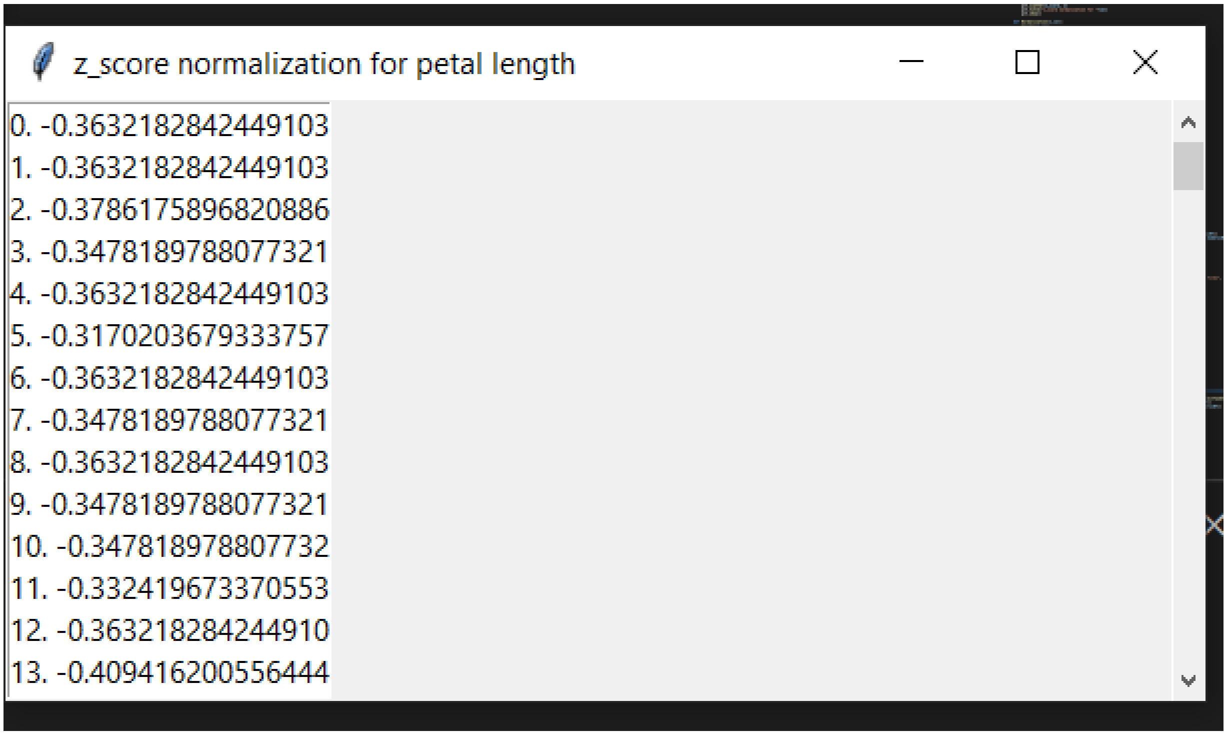
```
numeric.grid(column = 1, row = 2)
```

```
0. 0.09491525423728811
1. 0.09491525423728811
2. 0.0711864406779661
3. 0.1271186440677966
4. 0.10169491525423727
5. 0.20169491525423727
6. 0.11525423728813557
7. 0.14406779661016947
8. 0.11525423728813557
9. 0.14406779661016947
10. 0.1440677966101694
11. 0.1728813559322033
12. 0.1152542372881355
13. 0.0288135593220339
```

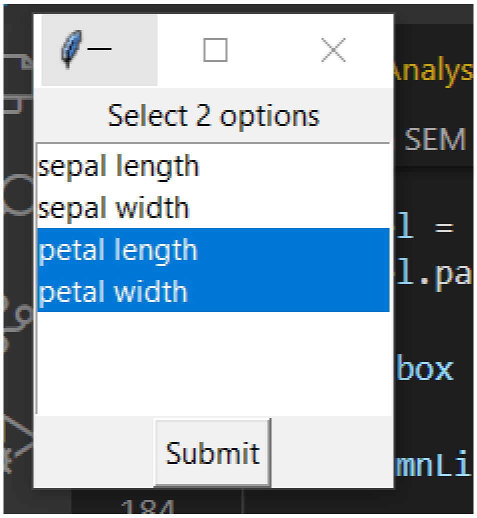
```
Python runtime state: initialized
```



```
0. 0.0014
1. 0.0014
2. 0.0013
3. 0.0015
4. 0.0014
5. 0.0017
6. 0.0014
7. 0.0015
8. 0.0014
9. 0.0015
10. 0.0015
11. 0.0016
12. 0.0014
13. 0.0011
14. 0.0012
15. 0.0015
16. 0.0013
17. 0.0014
18. 0.0017
```

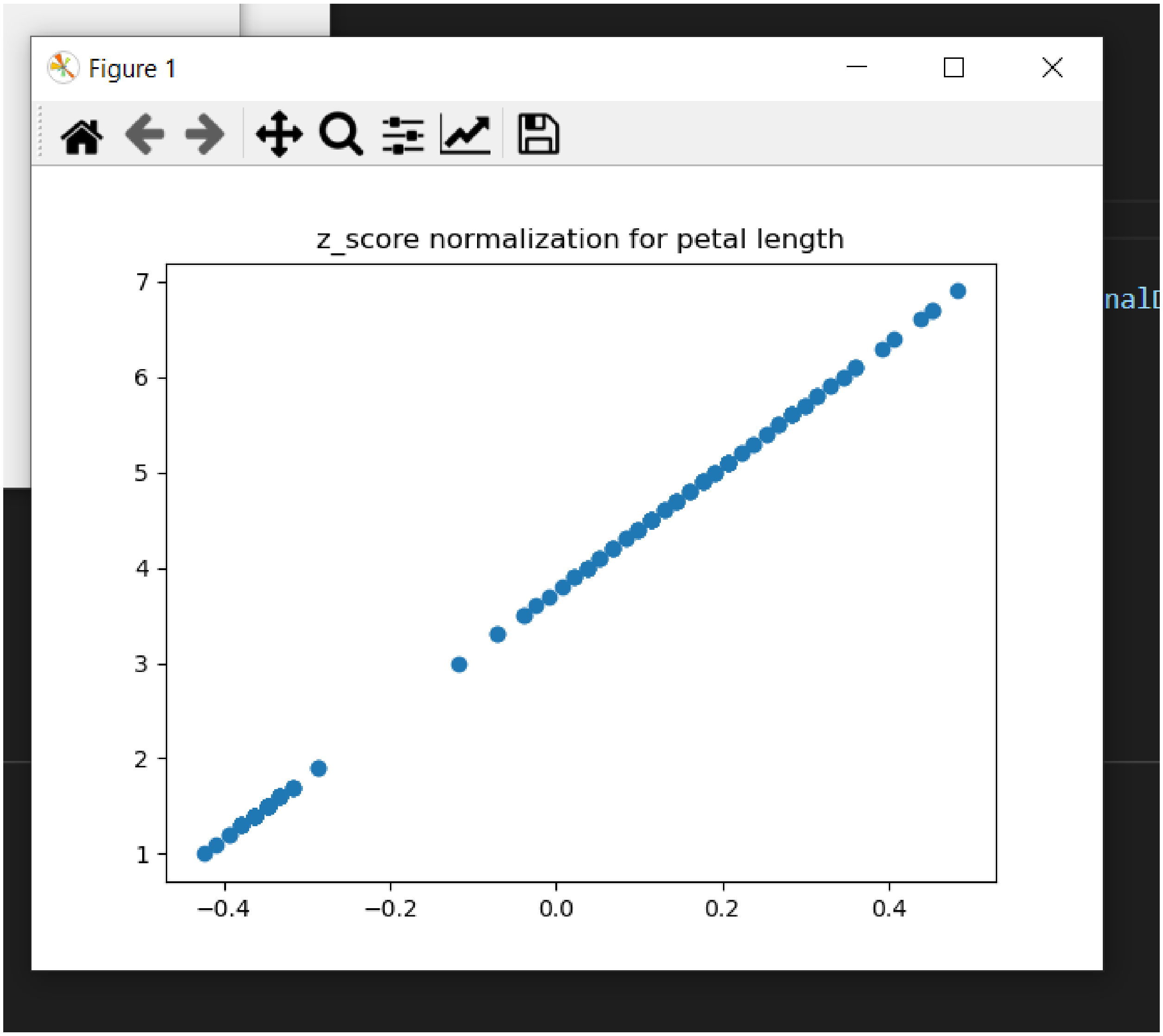


```
0. -0.3632182842449103
1. -0.3632182842449103
2. -0.3786175896820886
3. -0.3478189788077321
4. -0.3632182842449103
5. -0.3170203679333757
6. -0.3632182842449103
7. -0.3478189788077321
8. -0.3632182842449103
9. -0.3478189788077321
10. -0.347818978807732
11. -0.332419673370553
12. -0.363218284244910
13. -0.409416200556444
```



Select 2 options

- sepal length
- sepal width
- petal length
- petal width



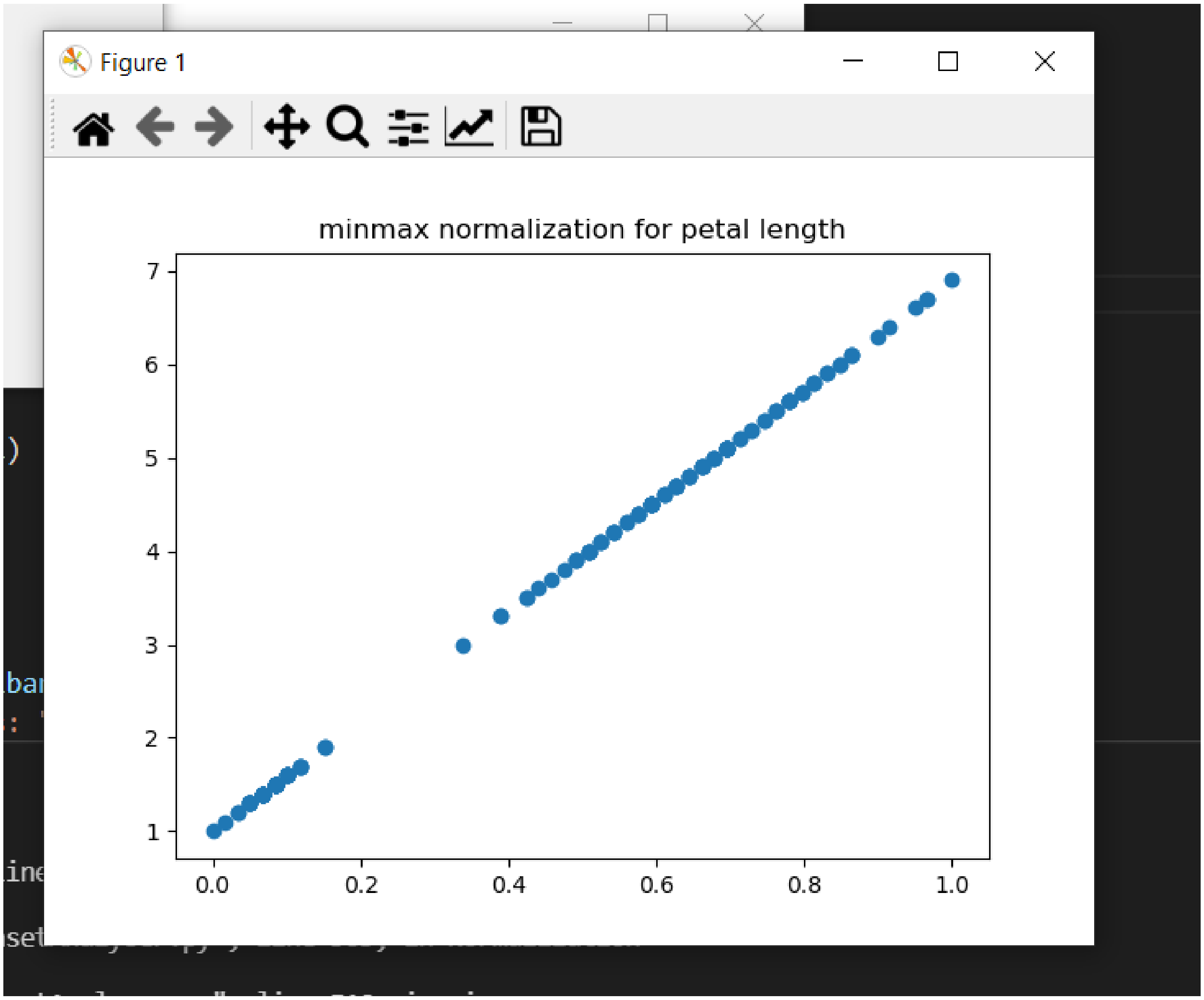


Figure 1
minmax normalization for petal length
7
6
5
4
3
2
1
0.0
0.2
0.4
0.6
0.8
1.0

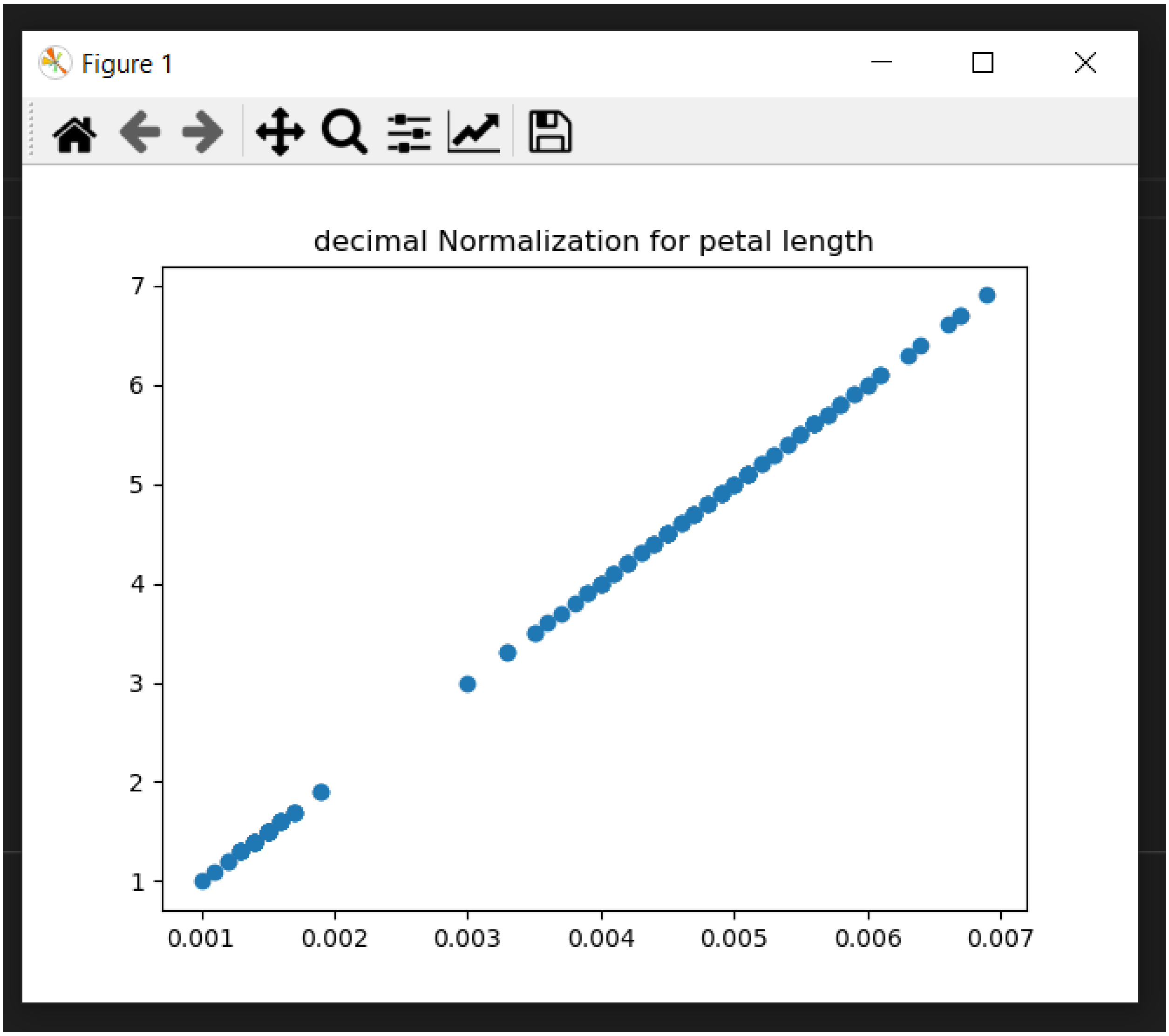
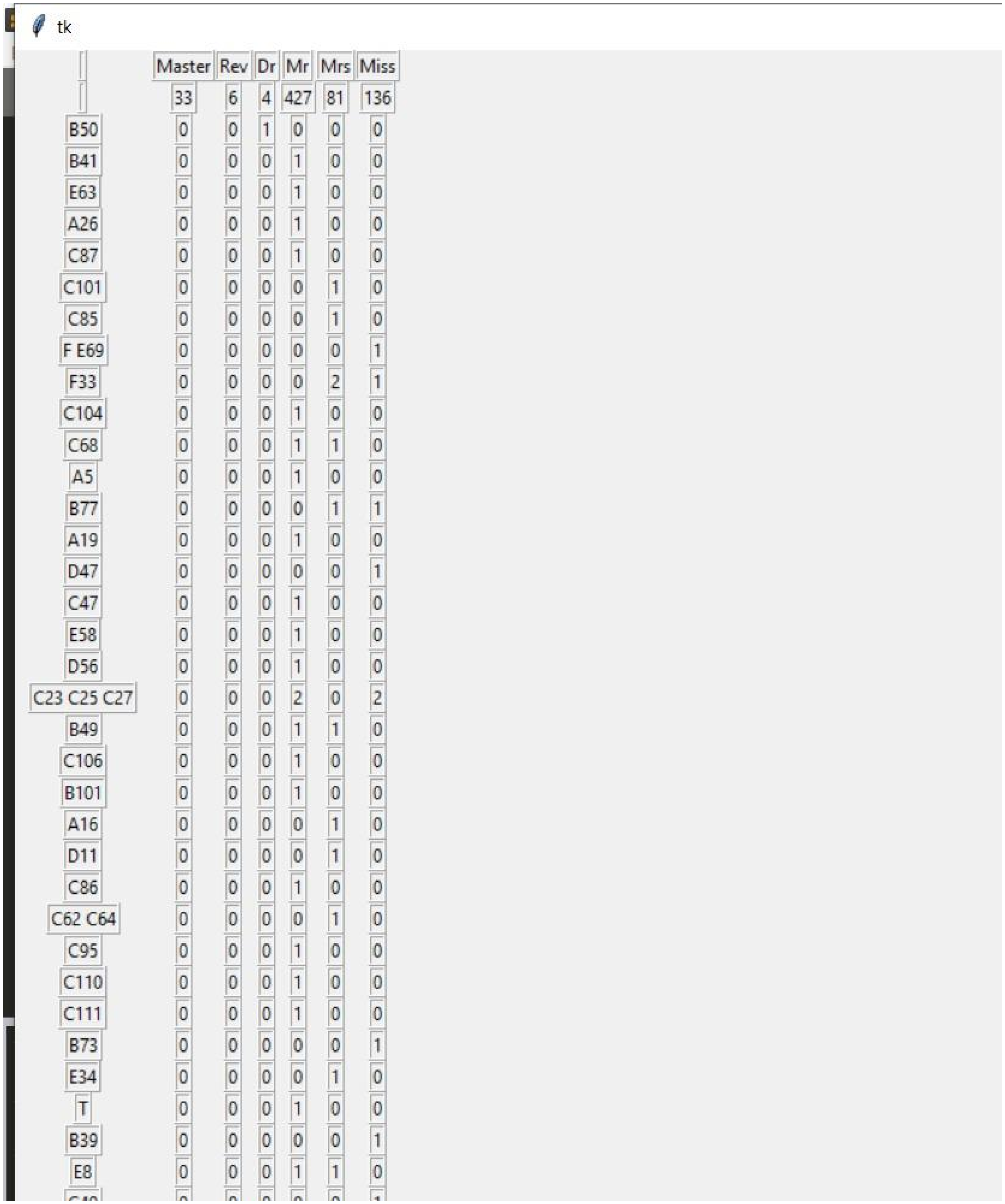
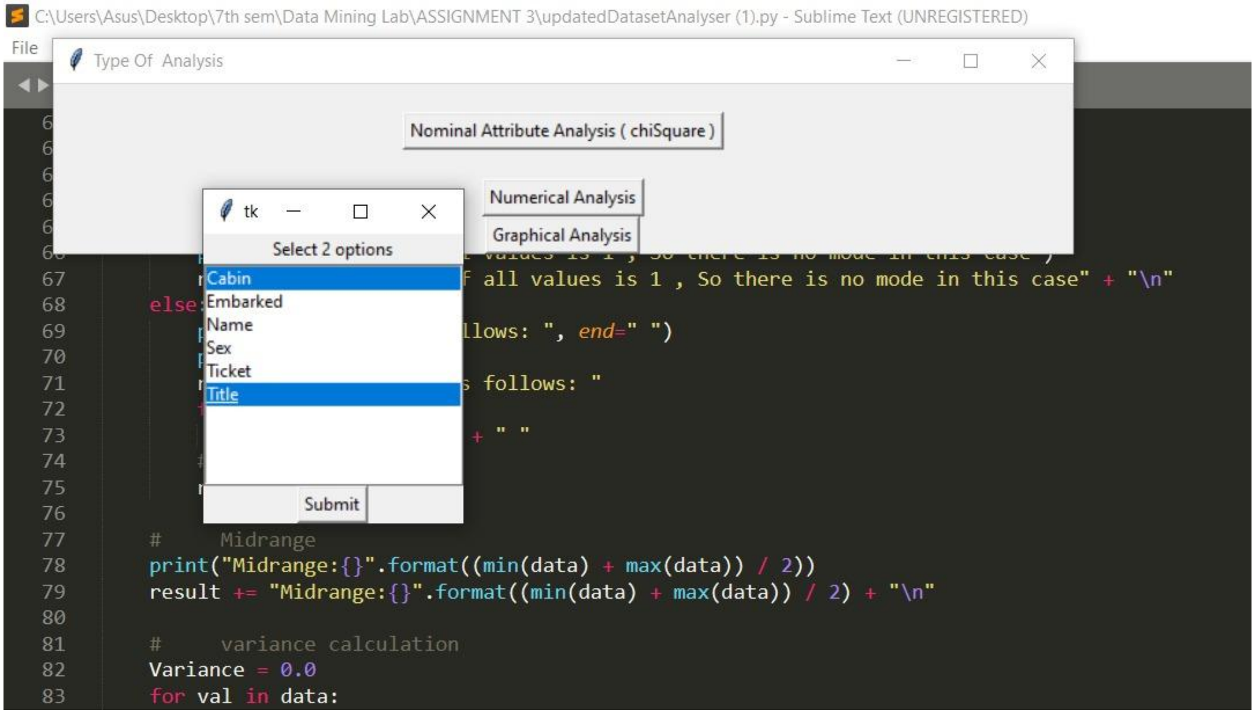


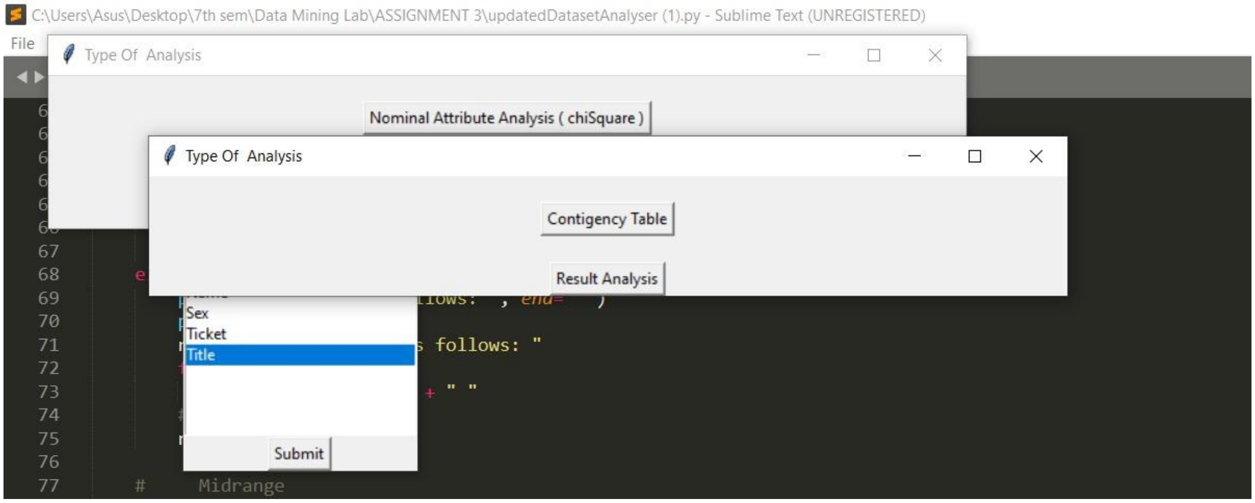
Figure 1
decimal Normalization for petal length
7
6
5
4
3
2
1
0.001
0.002
0.003
0.004
0.005
0.006
0.007



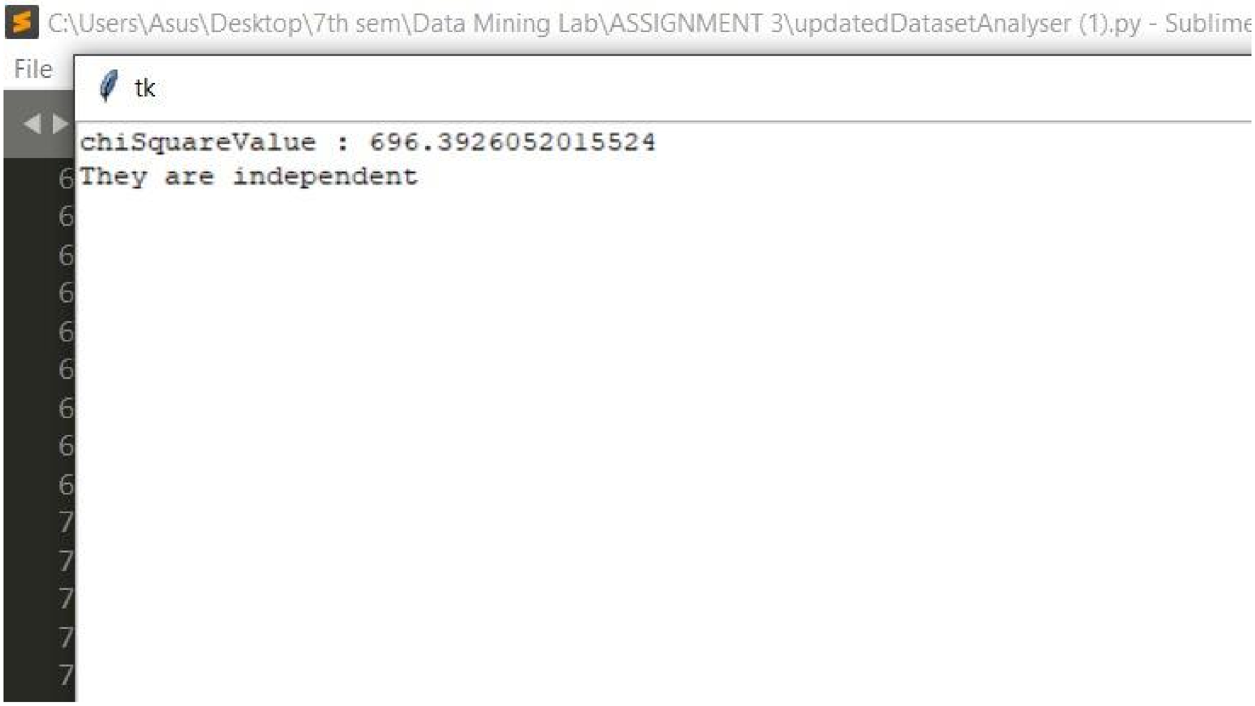
	Master	Rev	Dr	Mr	Mrs	Miss
	33	6	4	427	81	136
B50	0	0	1	0	0	0
B41	0	0	0	1	0	0
E63	0	0	0	1	0	0
A26	0	0	0	1	0	0
C87	0	0	0	1	0	0
C101	0	0	0	0	1	0
C85	0	0	0	0	1	0
F E69	0	0	0	0	0	1
F33	0	0	0	0	2	1
C104	0	0	0	1	0	0
C68	0	0	0	1	1	0
A5	0	0	0	1	0	0
B77	0	0	0	0	1	1
A19	0	0	0	1	0	0
D47	0	0	0	0	0	1
C47	0	0	0	1	0	0
E58	0	0	0	1	0	0
D56	0	0	0	1	0	0
C23 C25 C27	0	0	0	2	0	2
B49	0	0	0	1	1	0
C106	0	0	0	1	0	0
B101	0	0	0	1	0	0
A16	0	0	0	0	1	0
D11	0	0	0	0	1	0
C86	0	0	0	1	0	0
C62 C64	0	0	0	0	1	0
C95	0	0	0	1	0	0
C110	0	0	0	1	0	0
C111	0	0	0	1	0	0
B73	0	0	0	0	0	1
E34	0	0	0	0	1	0
T	0	0	0	1	0	0
B39	0	0	0	0	0	1
E8	0	0	0	1	1	0
[illegible]	0	0	0	0	0	1


C:\Users\Asus\Desktop\7th sem\Data Mining Lab\ASSIGNMENT 3\updatedDatasetAnalyser (1).py - Sublime Text (UNREGISTERED)
File
Type Of Analysis
Nominal Attribute Analysis (chiSquare)
Numerical Analysis
Graphical Analysis
tk
Select 2 options
Cabin
Embarked
Name
Sex
Ticket
Title
Submit

```
f all values is 1 , So there is no mode in this case" + "\n"
    else:
                                     llows: ", end=" ")
                                     s follows: "
                                     + " "

    #     Midrange
    print("Midrange:{}".format((min(data) + max(data)) / 2))
    result += "Midrange:{}".format((min(data) + max(data)) / 2) + "\n"

    #     variance calculation
    Variance = 0.0
    for val in data:
```


C:\Users\Asus\Desktop\7th sem\Data Mining Lab\ASSIGNMENT 3\updatedDatasetAnalyser (1).py - Sublime Text (UNREGISTERED)
File
Type Of Analysis
Nominal Attribute Analysis (chiSquare)
Type Of Analysis
Contigency Table
Result Analysis
Sex
Ticket
Title
Submit

```
                                     llows: ", end=" ")
                                     s follows: "
                                     + " "

    #     Midrange
```


C:\Users\Asus\Desktop\7th sem\Data Mining Lab\ASSIGNMENT 3\updatedDatasetAnalyser (1).py - Sublime
File
tk

```
chiSquareValue : 696.3926052015524
They are independent
```

Conclusion:

Studied and implemented the targeted concepts

References:

- Correlation Coefficient: Simple Definition, Formula, Easy Calculation Steps (statisticshowto.com)
- https://www.geeksforgeeks.org/data-normalization-in-data-mining/
- python - How to plot 1-d data at given y-value with pylab - Stack Overflow