Final Year B. Tech. (CSE) – I : 2021-22
4CS462 : PE2 - Data Mining Lab
Assignment No. 2

Group id: DM21G12

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

Title : Design the data analysis tool/program to perform the given tasks

Objective/Aim :

1. Data upload (std. format like .csv, excel etc.) and view
2. Calculate and show the measures of central tendency for uploaded data : mean , median , mode , midrange , variance and standard deviation
3. Calculate and show the dispersion of data : range , quartiles , interquartile range , five-number summary
4. Graphical display of above calculated statistical description of data (provide the facility - UI form to choose different attributes from uploaded data set) :
    a. Quantile plot
    b. Quantile-quantile (q-q) plot
    c. Histogram
    d. Scatter plot
    e. Boxplot

Introduction:

We used Tkinter library for designing GUI. Tkinter is the most commonly used library for developing GUI (Graphical User Interface) in Python. It is a standard Python interface to the Tk GUI toolkit shipped with Python. As Tk and Tkinter are available on most of the Unix platforms as well as on the Windows system, developing GUI applications with Tkinter becomes the fastest and easiest.

Theory/Algorithms:

- CSV : A comma-separated values file is a delimited text file that uses a comma to separate values. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format.

- The measures of central tendency :

    In statistics, a central tendency (or measure of central tendency) is a central or typical value for a probability distribution. It may also be called a center or location of the distribution. Colloquially, measures of central tendency are often called averages. The term central tendency dates from the late 1920s.

The most common measures of central tendency are the arithmetic mean, the median, and the mode. A middle tendency can be calculated for either a finite set of values or for a theoretical distribution, such as the normal distribution. Occasionally authors use central tendency to denote "the tendency of quantitative data to cluster around some central value."

The central tendency of a distribution is typically contrasted with its dispersion or variability; dispersion and central tendency are the often characterized properties of distributions. Analysis may judge whether data has a strong or a weak central tendency based on its dispersion.

The following may be applied to one-dimensional data. Depending on the circumstances, it may be appropriate to transform the data before calculating a central tendency. Examples are squaring the values or taking logarithms. Whether a transformation is appropriate and what it should be, depend heavily on the data being analyzed.

1) Arithmetic mean or simply, mean : the sum of all measurements divided by the number of observations in the data set.

   The arithmetic mean is the simplest and most widely used measure of a mean, or average. It simply involves taking the sum of a group of numbers, then dividing that sum by the count of the numbers used in the series. For example, take the numbers 34, 44, 56, and 78. The sum is 212. The arithmetic mean is 212 divided by four, or 53.

2) Median : The middle value that separates the higher half from the lower half of the data set. The median and the mode are the only measures of central tendency that can be used for ordinal data, in which values are ranked relative to each other but are not measured absolutely.

   Median is the middle number in a sorted list of numbers. To determine the median value in a sequence of numbers, the numbers must first be sorted, or arranged, in value order from lowest to highest or highest to lowest. The median can be used to determine an approximate average, or mean, but is not to be confused with the actual mean.

   If there is an odd amount of numbers, the median value is the number that is in the middle, with the same amount of numbers below and above.
   If there is an even amount of numbers in the list, the middle pair must be determined, added together, and divided by two to find the median value.
   The median is sometimes used as opposed to the mean when there are outliers in the sequence that might skew the average of the values. The median of a sequence can be less affected by outliers than the mean.

   Median Example

To find the median value in a list with an odd amount of numbers, one would find the number that is in the middle with an equal amount of numbers on either side of the median. To find the median, first arrange the numbers in order, usually from lowest to highest.

For example, in a data set of {3, 13, 2, 34, 11, 26, 47}, the sorted order becomes {2, 3, 11, 13, 26, 34, 47}. The median is the number in the middle {2, 3, 11, 13, 26, 34, 47}, which in this instance is 13 since there are three numbers on either side.

To find the median value in a list with an even amount of numbers, one must determine the middle pair, add them, and divide by two. Again, arrange the numbers in order from lowest to highest.

For example, in a data set of {3, 13, 2, 34, 11, 17, 27, 47}, the sorted order becomes {2, 3, 11, 13, 17, 27, 34, 47}. The median is the average of the two numbers in the middle {2, 3, 11, 13, 17, 26 34, 47}, which in this case is fifteen {(13 + 17) ÷ 2 = 15}.

3) Mode : The mode is the value that appears most frequently in a data set. A set of data may have one mode, more than one mode, or no mode at all. Other popular measures of central tendency include the mean, or the average of a set, and the median, the middle value in a set.

   Examples of the Mode
   For example, in the following list of numbers, 16 is the mode since it appears more times in the set than any other number:

   3, 3, 6, 9, 16, 16, 16, 27, 27, 37, 48
   A set of numbers can have more than one mode (this is known as bimodal if there are two modes) if there are multiple numbers that occur with equal frequency, and more times than the others in the set.

   3, 3, 3, 9, 16, 16, 16, 27, 37, 48
   In the above example, both the number 3 and the number 16 are modes as they each occur three times and no other number occurs more often.

   If no number in a set of numbers occurs more than once, that set has no mode:

   3, 6, 9, 16, 27, 37, 48
   A set of numbers with two modes is bimodal, a set of numbers with three modes is trimodal, and any set of numbers with more than one mode is multimodal.

4) Geometric Mean : The nth root of the product of the data values, where there are n of these. This measure is valid only for data that are measured absolutely on a strictly positive scale.

5) Harmonic mean : The reciprocal of the arithmetic mean of the reciprocals of the data values. This measure too is valid only for data that are measured absolutely on a strictly positive scale.

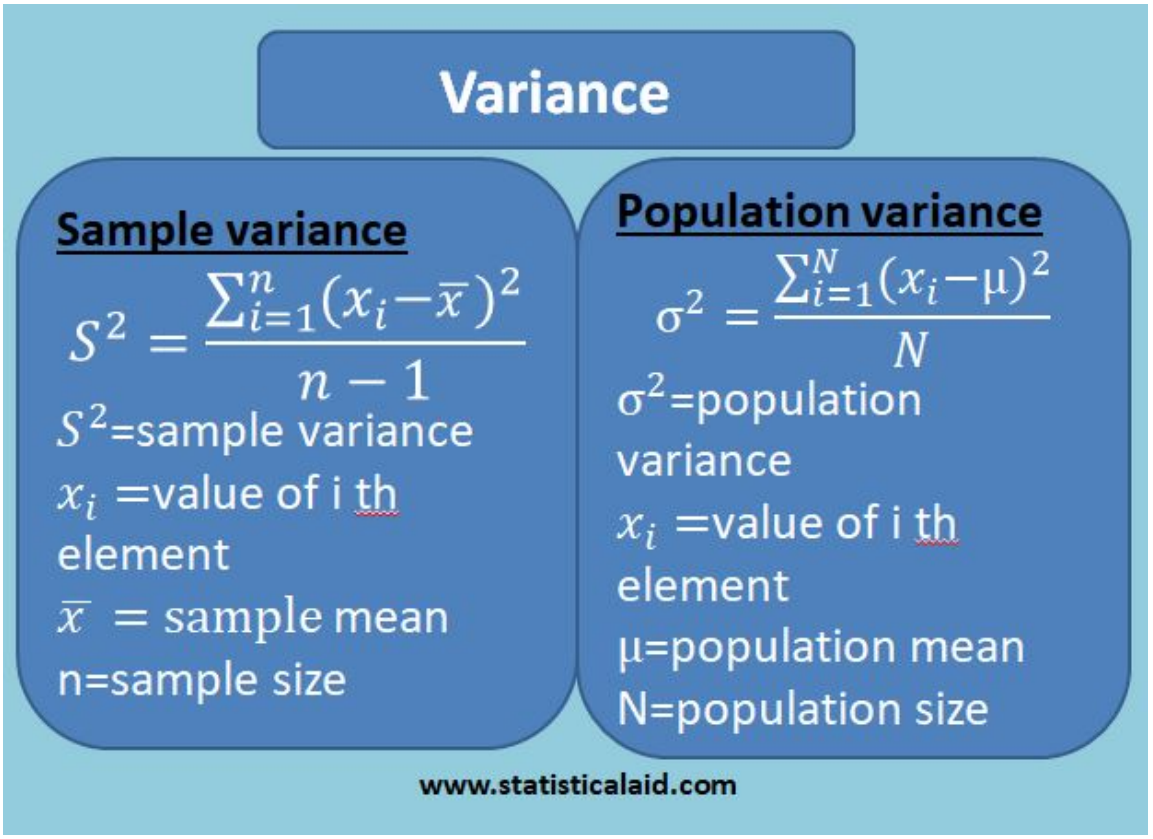
6) Weighted arithmetic mean : An arithmetic mean that incorporates weighting to certain data elements.

7) Interquartile mean : A truncated mean based on data within the interquartile range.

8) Midrange : Midrange in layman terms is the middle of any data set or the simply the average, mean of the data. A midrange is a statistical tool which is also known as the measure of center in statistics. Along with the existence of the midrange formula means, medium, average, mode, and range are also known as the measure of central tendency. The midrange of the data set is simply the value between the biggest value and the lowest value. In order to find the midrange of the data set the value is then divided by 2 after summing the lowest value present in the data set with the highest value present in the data set.

9) Variance : The term variance refers to a statistical measurement of the spread between numbers in a data set. More specifically, variance measures how far each number in the set is from the mean and thus from every other number in the set. Variance is often depicted by this symbol: σ2. It is used by both analysts and traders to determine volatility and market security. The square root of the variance is the standard deviation (σ), which helps determine the consistency of an investment’s returns over a period of time.

10) Standard Deviation : A standard deviation is a statistic that measures the dispersion of a dataset relative to its mean. The standard deviation is calculated as the square root of variance by determining each data point's deviation relative to the mean. If the data points are further from the mean, there is a higher deviation within the data set; thus, the more spread out the data, the higher the standard deviation.



11) Range : The difference between the largest value and the smallest value.

12) Quartiles : A quartile is a statistical term that describes a division of observations into four defined intervals based on the values of the data and how they compare to the entire set of observations.
Suppose the distribution of math scores in a class of 19 students in ascending order is:

59, 60, 65, 65, 68, 69, 70, 72, 75, 75, 76, 77, 81, 82, 84, 87, 90, 95, 98
First, mark down the median, Q2, which in this case is the 10th value: 75.

Q1 is the central point between the smallest score and the median. In this case, Q1 falls between the first and fifth score: 68. (Note that the median can also be included when calculating Q1 or Q3 for an odd set of values. If we were to include the median on either side of the middle point, then Q1 will be the middle value between the first and 10th score, which is the average of the fifth and sixth score—(fifth + sixth)/2 = (68 + 69)/2 = 68.5).

Q3 is the middle value between Q2 and the highest score: 84. (Or if you include the median, Q3 = (82 + 84)/2 = 83).

Now that we have our quartiles, let's interpret their numbers. A score of 68 (Q1) represents the first quartile and is the 25th percentile. 68 is the median of the lower half of the score set in the available data—that is, the median of the scores from 59 to 75.

Q1 tells us that 25% of the scores are less than 68 and 75% of the class scores are greater. Q2 (the median) is the 50th percentile and shows that 50% of the scores are less than 75, and 50% of the scores are above 75. Finally, Q3, the 75th percentile, reveals that 25% of the scores are greater and 75% are less than 84.

13) Interquartile Range : The interquartile range is a measure of where the "middle fifty" is in a data set. Where a range is a measure of where the beginning and end are in a set, an interquartile range is a measure of where the bulk of the values lie. That's why it's preferred over many other measures of spread when reporting things like school performance or SAT scores.

    The interquartile range formula is the first quartile subtracted from the third quartile:

    IQR = Q3 – Q1.

14) Five-number summary : The five number summary includes 5 items:

    - The minimum.
    - Q1 (the first quartile, or the 25% mark).
    - The median.
    - Q3 (the third quartile, or the 75% mark).
    - The maximum.

    The five number summary gives you a rough idea about what your data set looks like. for example, you'll have your lowest value (the minimum) and the highest value (the maximum). Although it's useful in itself, the main reason you'll want to find a five-number summary is to find more useful statistics, like the interquartile range, sometimes called the middle fifty.

Documentation :

Tkinter tutorial provides basic and advanced concepts of Python Tkinter. Our Tkinter tutorial is designed for beginners and professionals.

Python provides the standard library Tkinter for creating the graphical user interface for desktop based applications.

Developing desktop based applications with python Tkinter is not a complex task. An empty Tkinter top-level window can be created by using the following steps.

1. import the Tkinter module.
2. Create the main application window.
3. Add the widgets like labels, buttons, frames, etc. to the window.
4. Call the main event loop so that the actions can take place on the user's computer screen.

Example :

```
# !/usr/bin/python3
from tkinter import *
#creating the application main window.
top = Tk()
#Entering the event main loop
top.mainloop()
```

Output : 

Python Tkinter Button

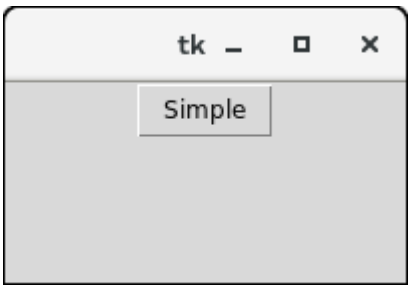
The button widget is used to add various types of buttons to the python application. Python allows us to configure the look of the button according to our requirements. Various options can be set or reset depending upon the requirements.

We can also associate a method or function with a button which is called when the button is pressed.

The syntax to use the button widget is given below.

Syntax : W = Button(parent, options)

```
#python application to create a simple button

from tkinter import *


top = Tk()

top.geometry("200x100")

b = Button(top,text = "Simple")

b.pack()

top.mainaloop()
```

Output : 

Procedure:

Steps to Import a CSV File into Python using Pandas

Step 1: Capture the File Path

Firstly, capture the full path where your CSV file is stored.

For example, let's suppose that a CSV file is stored under the following path:

C:\Users\Ron\Desktop\Clients.csv

You'll need to modify the Python code below to reflect the path where the CSV file is stored on your computer. Don't forget to include the:

File name (as highlighted in green). You may choose a different file name, but make sure that the file name specified in the code matches with the actual file name
File extension (as highlighted in blue). The file extension should always be '.csv' when importing CSV files

Step 2: Apply the Python code

Type/copy the following code into Python, while making the necessary changes to your path.

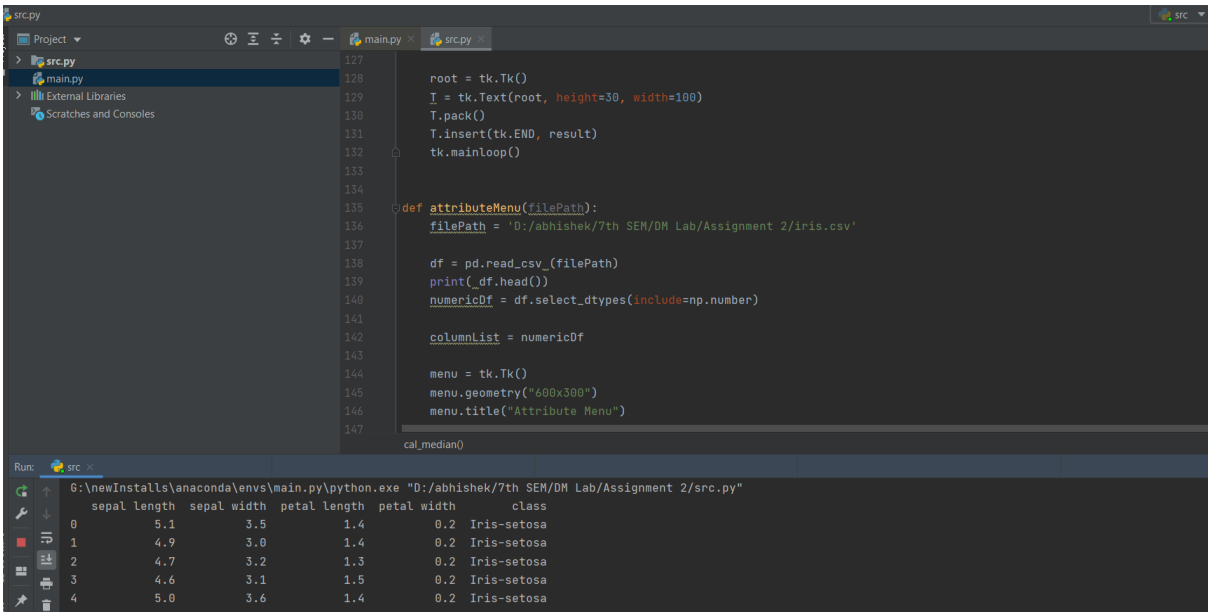
Here is the code for our example (you can find additional comments within the code itself):

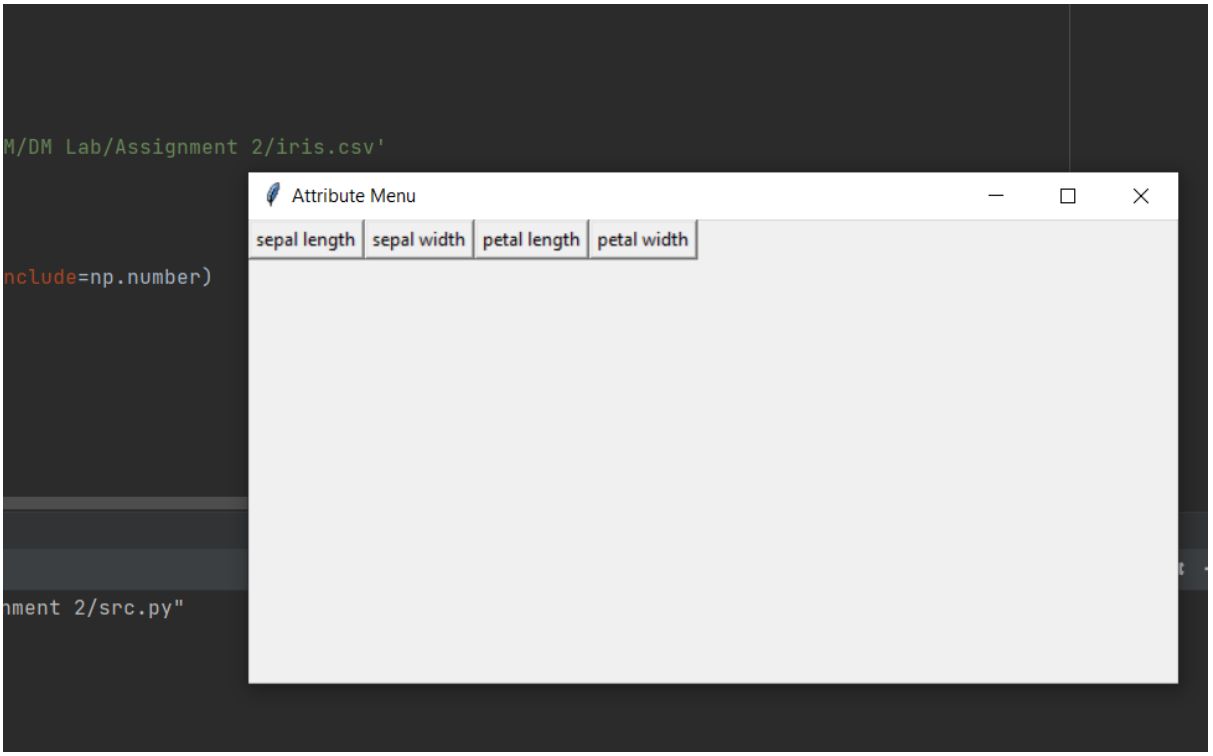
```
import pandas as pd

df = pd.read_csv (r'C:\Users\Ron\Desktop\Clients.csv')   #read the csv file (put 'r' before the path string to address any special characters in the path, such as '\'). Don't forget to put the file name at the end of the path + ".csv"
print (df)
```

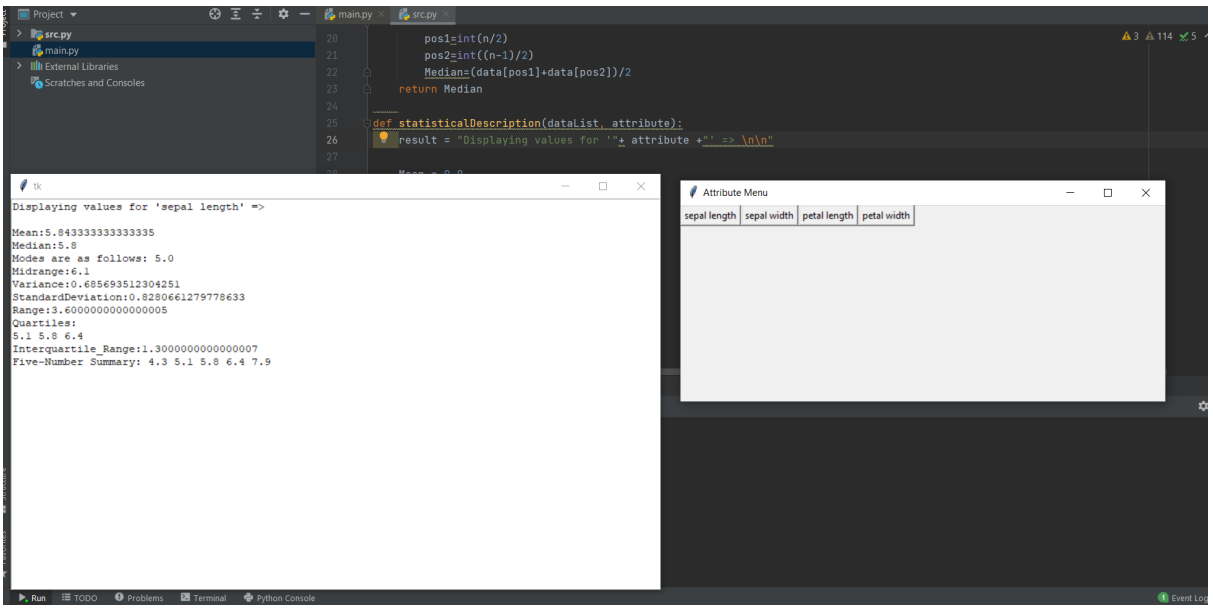
Step 3: Run the Code

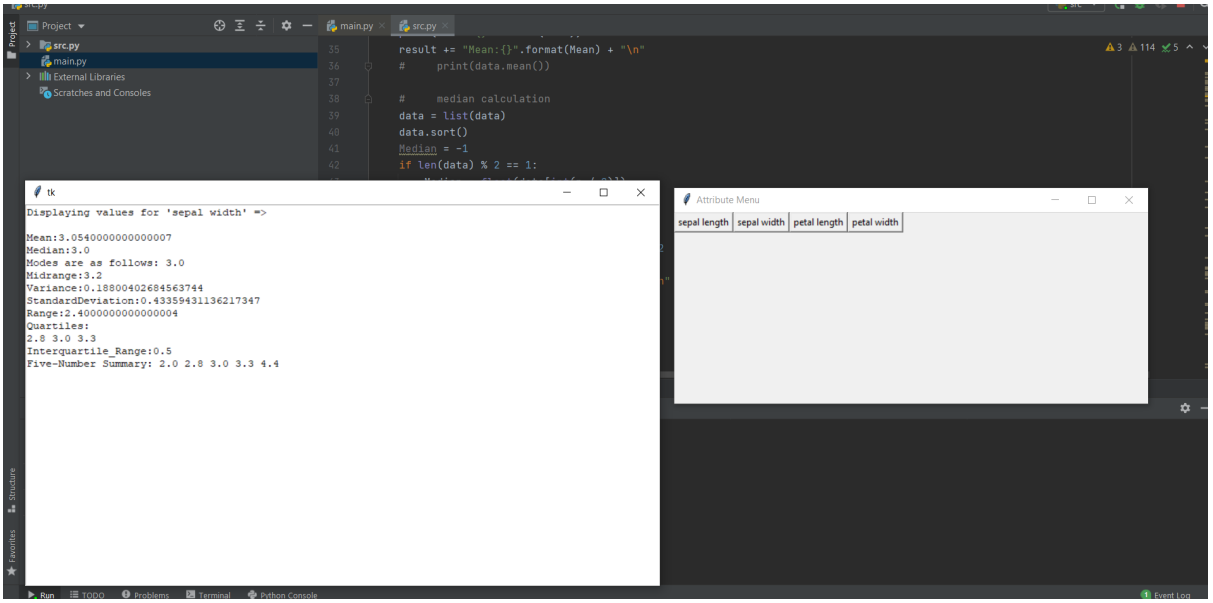
Finally, run the Python code and you'll get:

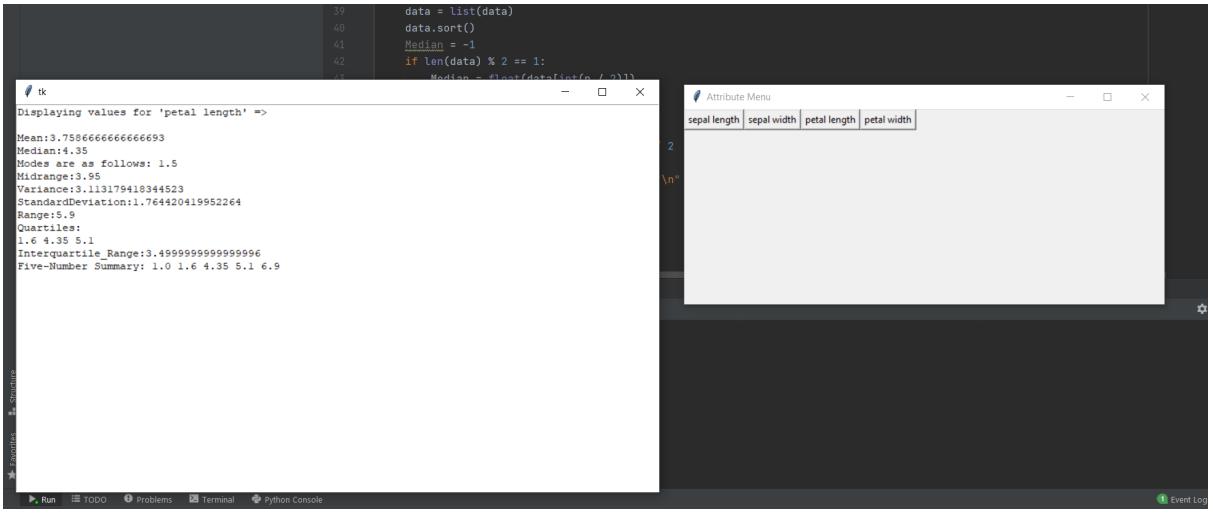
Result/Observations/Screenshots:

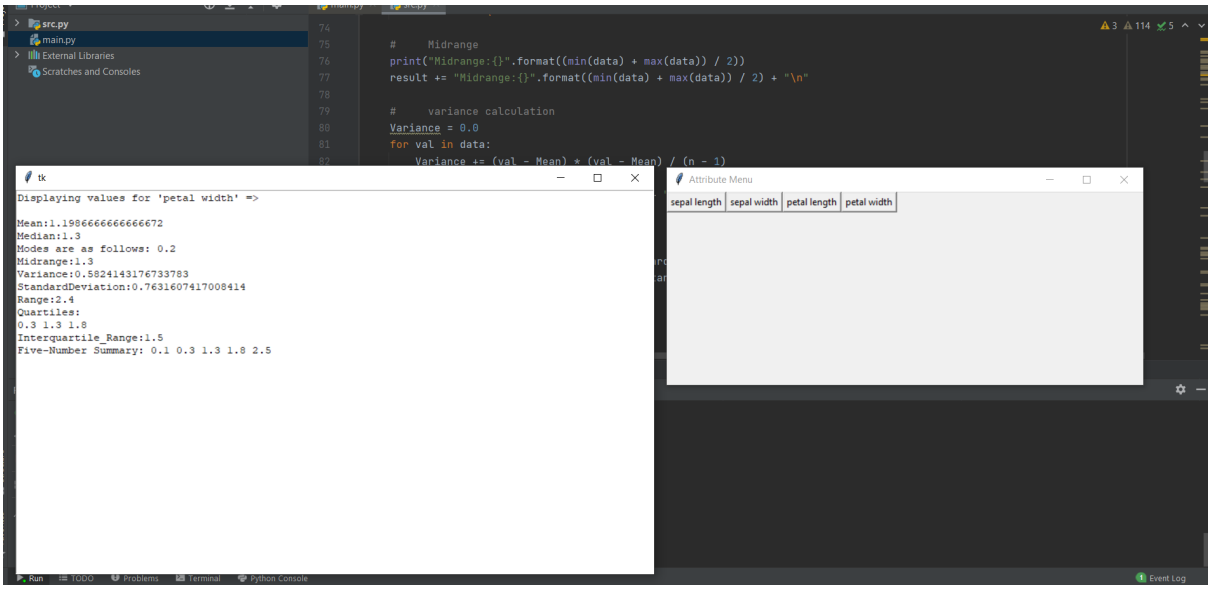


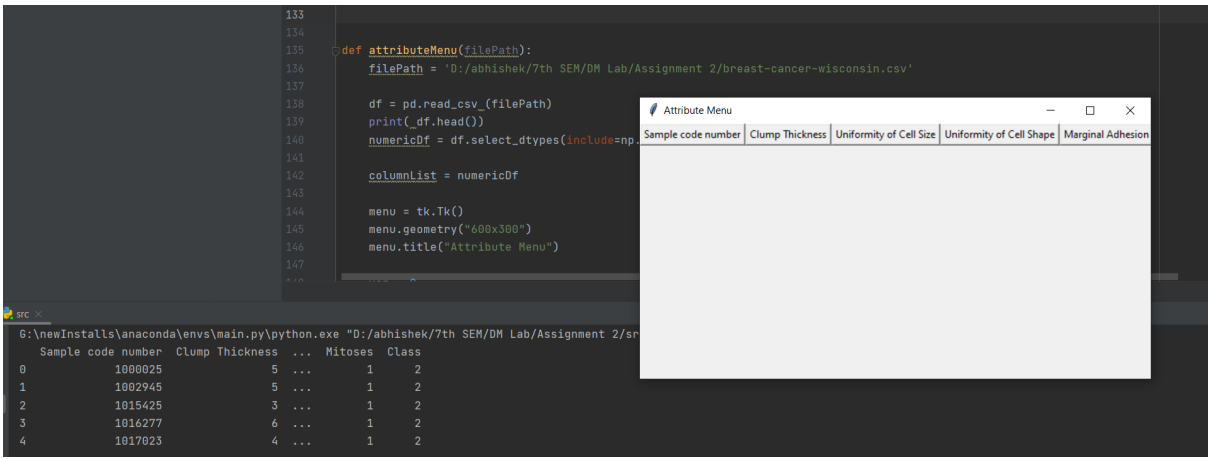


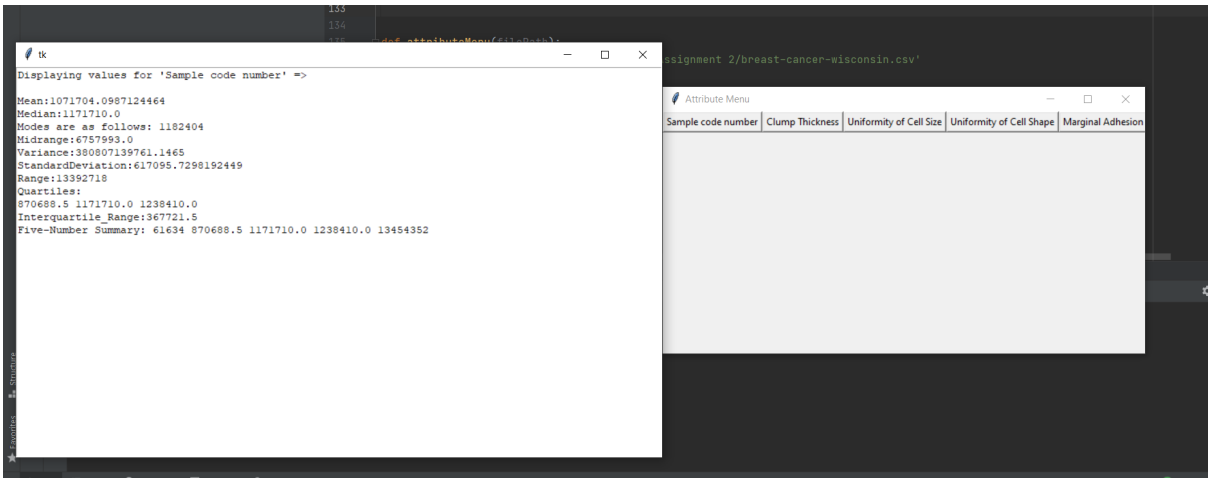

pos1=int(n/2)
pos2=int((n-1)/2)
Median=(data[pos1]+data[pos2])/2
return Median
def statisticalDescription(dataList, attribute):
result = "Displaying values for '"+ attribute +"' => \n\n"
tk
Displaying values for 'sepal length' =>
Mean:5.843333333333335
Median:5.8
Modes are as follows: 5.0
Midrange:6.1
Variance:0.685693512304251
StandardDeviation:0.8280661279778633
Range:3.6000000000000005
Quartiles:
5.1 5.8 6.4
Interquartile_Range:1.3000000000000007
Five-Number Summary: 4.3 5.1 5.8 6.4 7.9
Attribute Menu
sepal length
sepal width
petal length
petal width

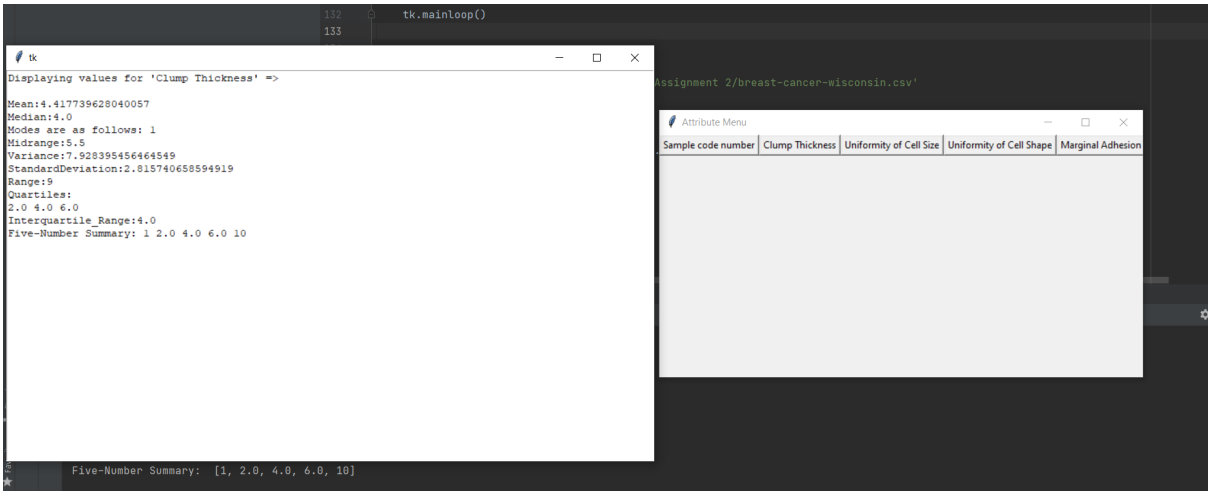

result += "Mean:{}".format(Mean) + "\n"
# print(data.mean())
# median calculation
data = list(data)
data.sort()
Median = -1
if len(data) % 2 == 1:
tk
Displaying values for 'sepal width' =>
Mean:3.0540000000000007
Median:3.0
Modes are as follows: 3.0
Midrange:3.2
Variance:0.18800402684563744
StandardDeviation:0.43359431136217347
Range:2.4000000000000004
Quartiles:
2.8 3.0 3.3
Interquartile_Range:0.5
Five-Number Summary: 2.0 2.8 3.0 3.3 4.4
Attribute Menu
sepal length
sepal width
petal length
petal width

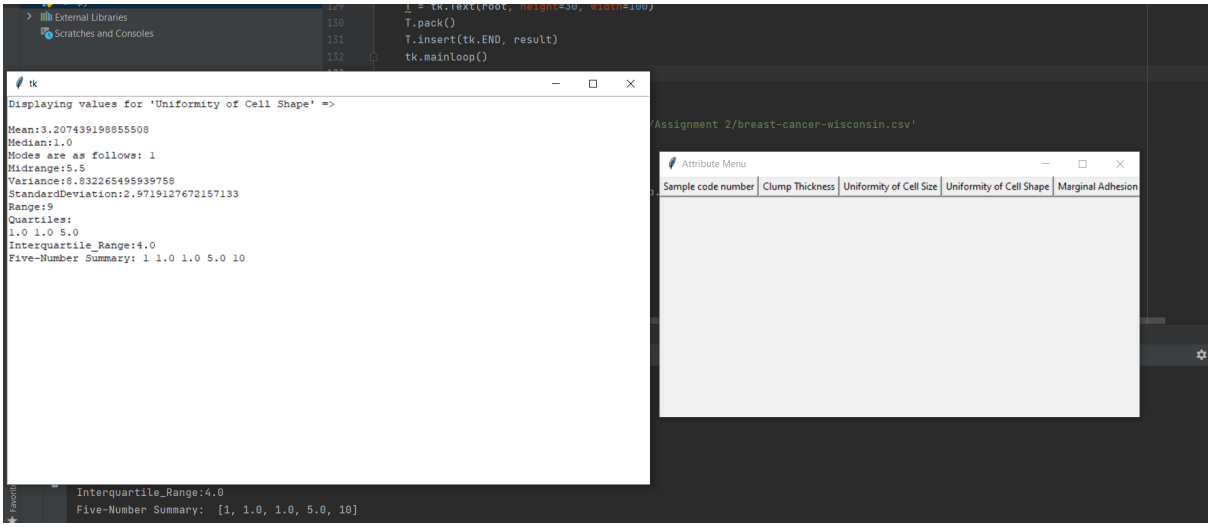

data = list(data)
data.sort()
Median = -1
if len(data) % 2 == 1:
tk
Displaying values for 'petal length' =>
Mean:3.7586666666666693
Median:4.35
Modes are as follows: 1.5
Midrange:3.95
Variance:3.1131794183445223
StandardDeviation:1.764420419952264
Range:5.9
Quartiles:
1.6 4.35 5.1
Interquartile_Range:3.4999999999999996
Five-Number Summary: 1.0 1.6 4.35 5.1 6.9
Attribute Menu
sepal length
sepal width
petal length
petal width

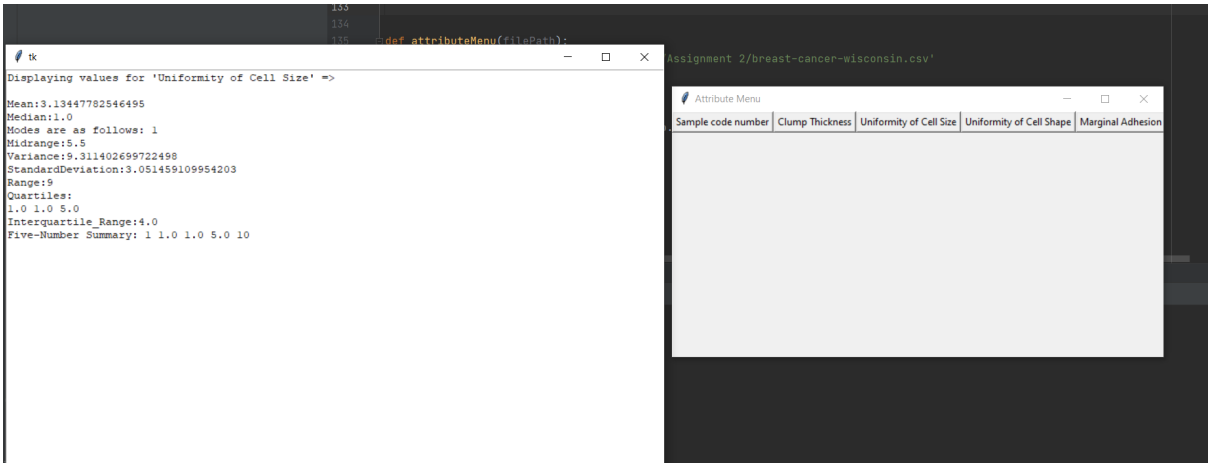

src.py
main.py
External Libraries
Scratches and Consoles
#   Midrange
print("Midrange:{}".format((min(data) + max(data)) / 2))
result += "Midrange:{}".format((min(data) + max(data)) / 2) + "\n"
#   variance calculation
Variance = 0.0
for val in data:
Variance += (val - Mean) * (val - Mean) / (n - 1)
tk
Displaying values for 'petal width' =>
Mean:1.1986666666666672
Median:1.3
Modes are as follows: 0.2
Midrange:1.3
Variance:0.5824143176733783
StandardDeviation:0.7631607417008414
Range:2.4
Quartiles:
0.3 1.3 1.8
Interquartile_Range:1.5
Five-Number Summary: 0.1 0.3 1.3 1.8 2.5
Attribute Menu
sepal length
sepal width
petal length
petal width
Run
TODO
Problems
Terminal
Python Console
Event Log

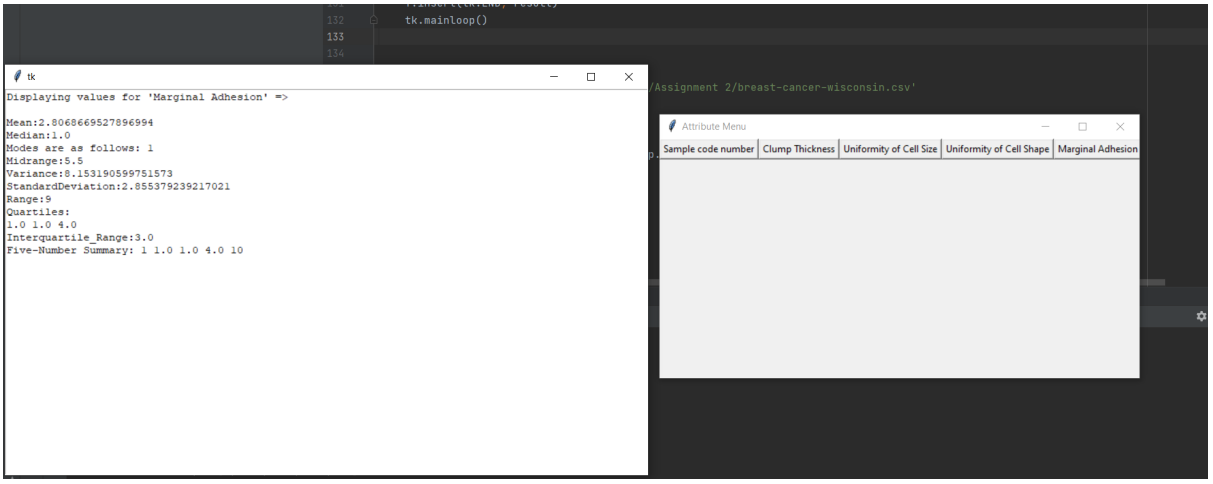

def attributeMenu(filePath):
filePath = 'D:/abhishek/7th SEM/DM Lab/Assignment 2/breast-cancer-wisconsin.csv'
df = pd.read_csv(filePath)
print(df.head())
numericDf = df.select_dtypes(include=np
columnList = numericDf
menu = tk.Tk()
menu.geometry("600x300")
menu.title("Attribute Menu")
Attribute Menu
Sample code number
Clump Thickness
Uniformity of Cell Size
Uniformity of Cell Shape
Marginal Adhesion
src
G:\newInstalls\anaconda\envs\main.py\python.exe "D:/abhishek/7th SEM/DM Lab/Assignment 2/sr
Sample code number Clump Thickness ... Mitoses Class
0 1000025 5 ... 1 2
1 1002945 5 ... 1 2
2 1015425 3 ... 1 2
3 1016277 6 ... 1 2
4 1017023 4 ... 1 2


tk
Displaying values for 'Sample code number' =>
Mean:1071704.0987124464
Median:1171710.0
Modes are as follows: 1182404
Midrange:6757993.0
Variance:380807139761.1465
StandardDeviation:617095.7298192449
Range:13392718
Quartiles:
870688.5 1171710.0 1238410.0
Interquartile_Range:367721.5
Five-Number Summary: 61634 870688.5 1171710.0 1238410.0 13454352
Attribute Menu
Sample code number
Clump Thickness
Uniformity of Cell Size
Uniformity of Cell Shape
Marginal Adhesion


tk.mainloop()
tk
Displaying values for 'Clump Thickness' =>
Mean:4.417739628040057
Median:4.0
Modes are as follows: 1
Midrange:5.5
Variance:7.928395456464549
StandardDeviation:2.815740658594919
Range:9
Quartiles:
2.0 4.0 6.0
Interquartile_Range:4.0
Five-Number Summary: 1 2.0 4.0 6.0 10
Assignment 2/breast-cancer-wisconsin.csv'
Attribute Menu
Sample code number
Clump Thickness
Uniformity of Cell Size
Uniformity of Cell Shape
Marginal Adhesion
Five-Number Summary: [1, 2.0, 4.0, 6.0, 10]


External Libraries
Scratches and Consoles
T.pack()
T.insert(tk.END, result)
tk.mainloop()
tk
Displaying values for 'Uniformity of Cell Shape' =>
Mean:3.207439198855508
Median:1.0
Modes are as follows: 1
Midrange:5.5
Variance:8.832265495939758
StandardDeviation:2.9719127672157133
Range:9
Quartiles:
1.0 1.0 5.0
Interquartile_Range:4.0
Five-Number Summary: 1 1.0 1.0 5.0 10
'Assignment 2/breast-cancer-wisconsin.csv'
Attribute Menu
Sample code number
Clump Thickness
Uniformity of Cell Size
Uniformity of Cell Shape
Marginal Adhesion
Interquartile_Range:4.0
Five-Number Summary: [1, 1.0, 1.0, 5.0, 10]


def attributeMenu(filePath):
tk
Displaying values for 'Uniformity of Cell Size' =>
Mean:3.13447782546495
Median:1.0
Modes are as follows: 1
Midrange:5.5
Variance:9.311402699722498
StandardDeviation:3.051459109954203
Range:9
Quartiles:
1.0 1.0 5.0
Interquartile_Range:4.0
Five-Number Summary: 1 1.0 1.0 5.0 10
'Assignment 2/breast-cancer-wisconsin.csv'
Attribute Menu
Sample code number
Clump Thickness
Uniformity of Cell Size
Uniformity of Cell Shape
Marginal Adhesion



Conclusion :

Designed the data analysis tool/program to perform the given tasks

References:

- https://prwatech.in/blog/data-science/measures-of-central-tendency-tutorial/
- https://www.simplilearn.com/tutorials/data-analytics-tutorial/measures-of-central-tendency?source=frs_recommended_resource_clicked
- https://www.geeksforgeeks.org/python-tkinter-tutorial/
- https://www.w3schools.com/python/pandas/default.asp
- https://www.tutorialspoint.com/matplotlib/index.htm