**Data Mining (CSE542)**

**Homework 01**

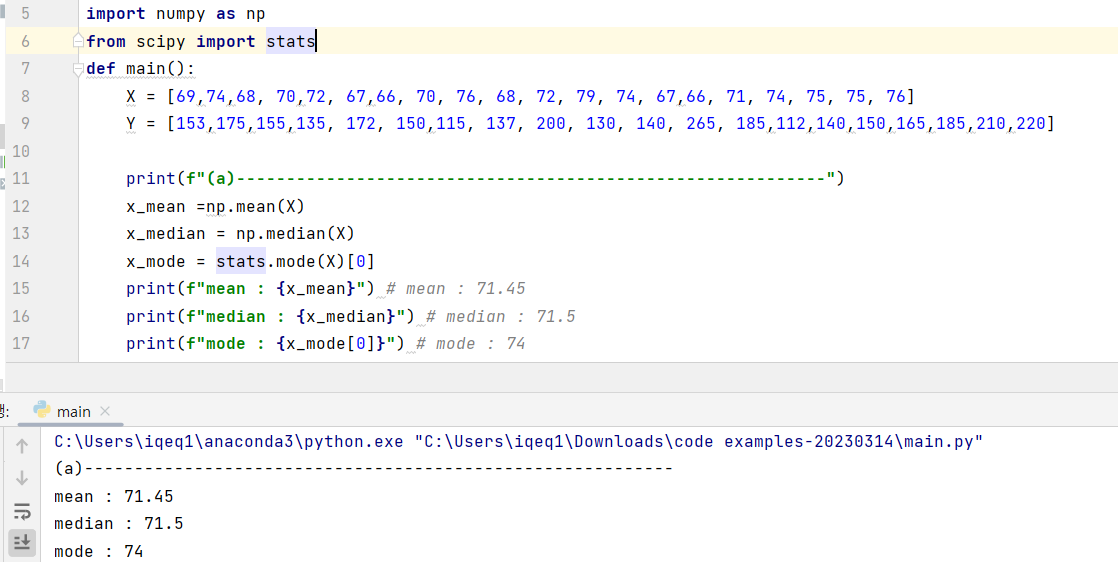
**ID: \_\_\_\_\_\_\_\_ Name: \_\_\_\_\_\_\_조원석\_\_\_\_\_\_\_ Date:\_\_’23/03/19\_\_\_\_\_\_**

**Task-1:** Let X and Y be two random variables, denoting age and weight, respectively. Consider a random sample of size n = 20 from these two variables

X =(69,74,68, 70,72, 67,66, 70, 76, 68, 72, 79, 74, 67,66, 71, 74, 75, 75, 76)

Y =(153,175,155,135, 172, 150,115, 137, 200, 130, 140, 265, 185,112,140,150,165,185,210,220)

(a) Find the mean, median, and mode for X.



- mean : 71.45

=>

median : 71.5

=> [66, 66, 67, 67, 68, 68, 69, 70, 70, **71, 72**, 72, 74, 74, 74, 75, 75, 76, 76, 79]

[01, 02, 03, 04, 05, 06, 07, 08, 09, **10, 11**, 12, 13, 14, 15, 16, 17, 18, 19, 20]

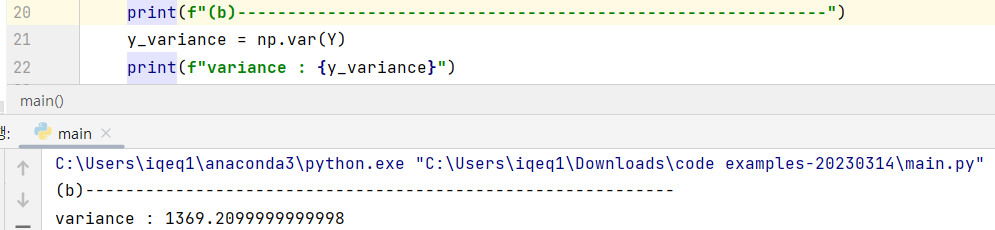
= 71.5

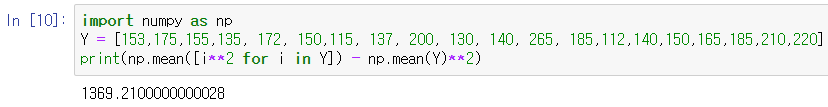
- mode : 74 ( do 3 times )

=> [66, 66, 67, 67, 68, 68, 69, 70, 70, 71, 72, 72, **74, 74, 74**, 75, 75, 76, 76, 79]

(b) What is the variance for Y?



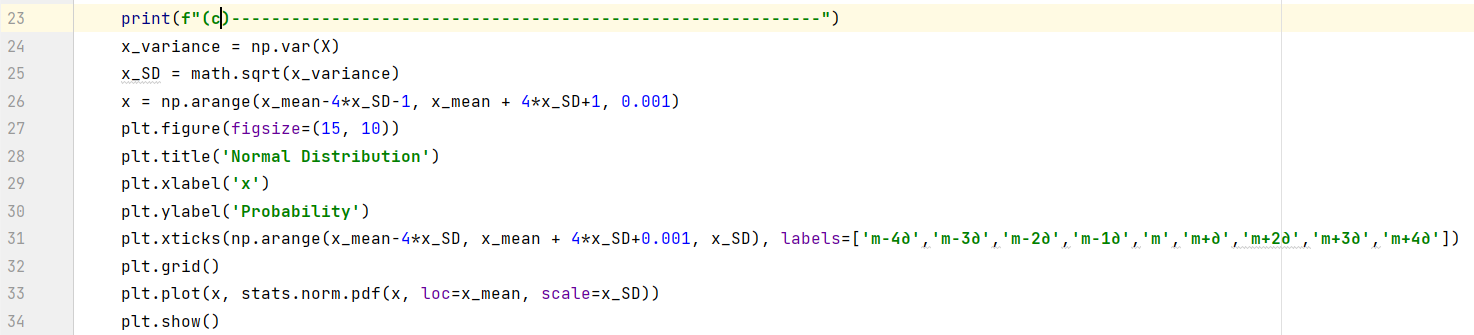


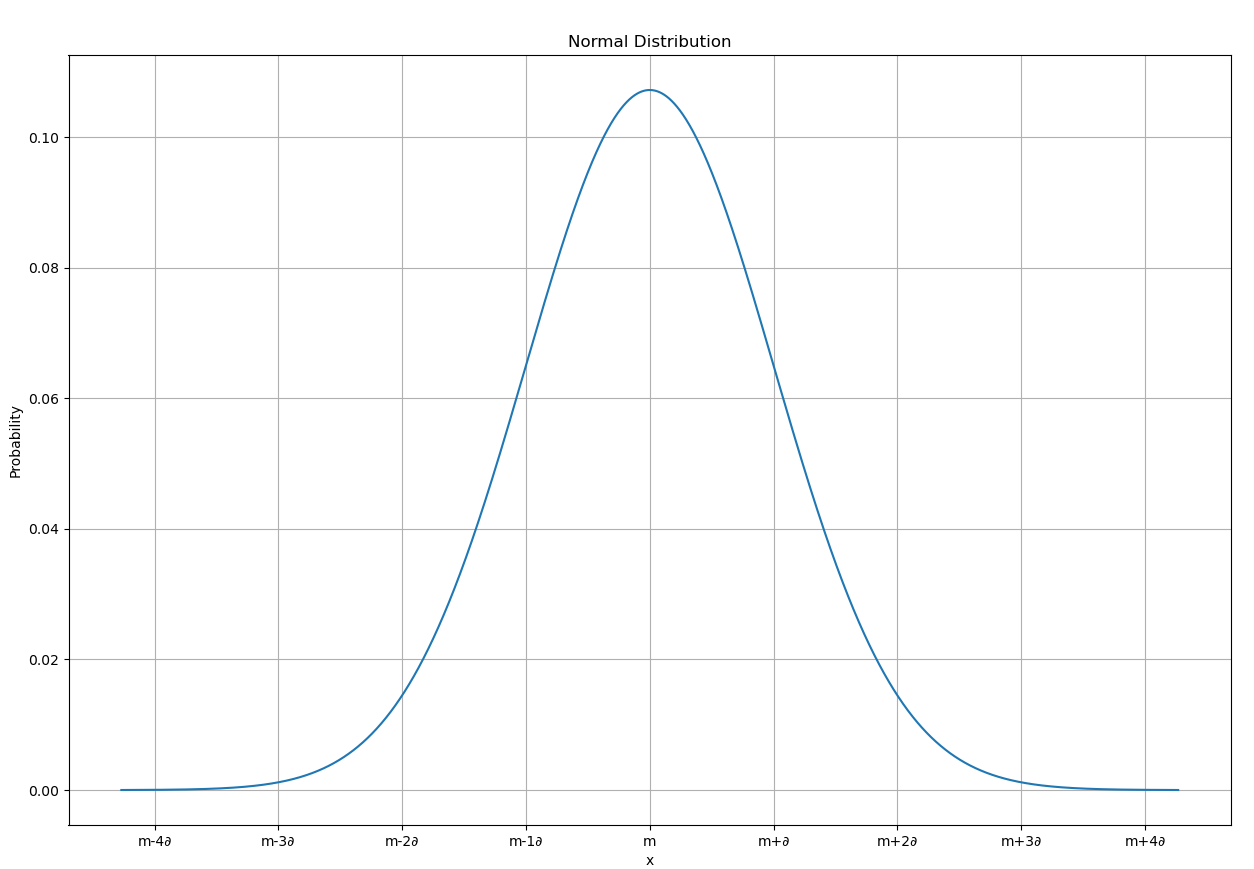


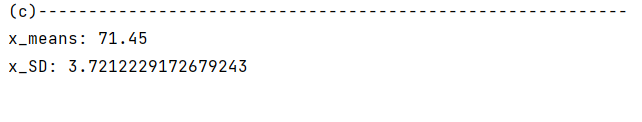
Y = [153,175,155,135, 172, 150,115, 137, 200, 130, 140, 265, 185,112,140,150,165,185,210,220]

- y\_variance ≒ 1369

(c) Plot the normal distribution for X.



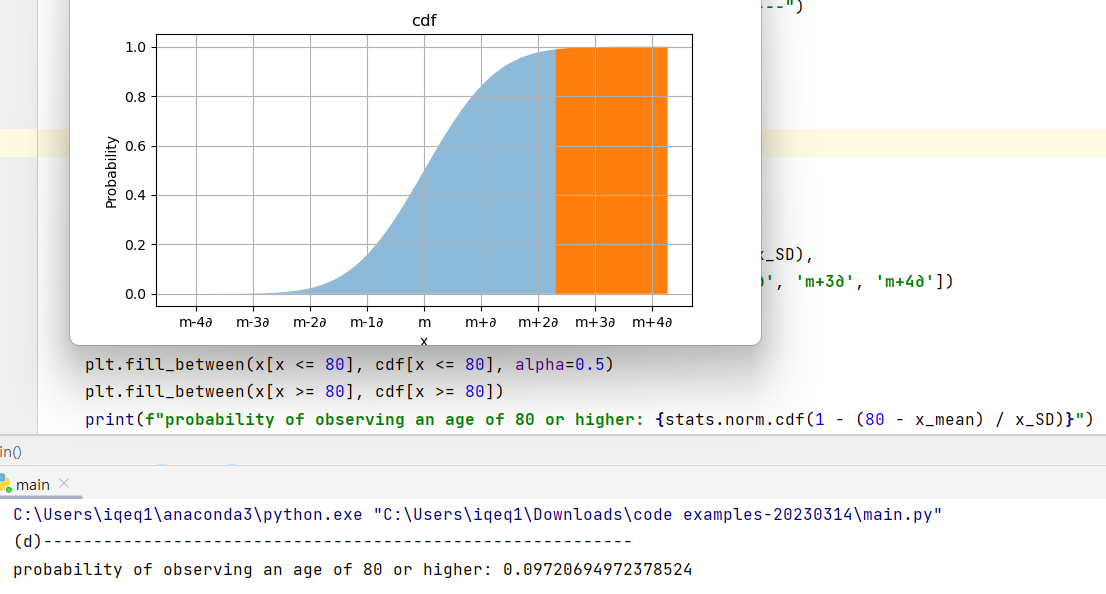




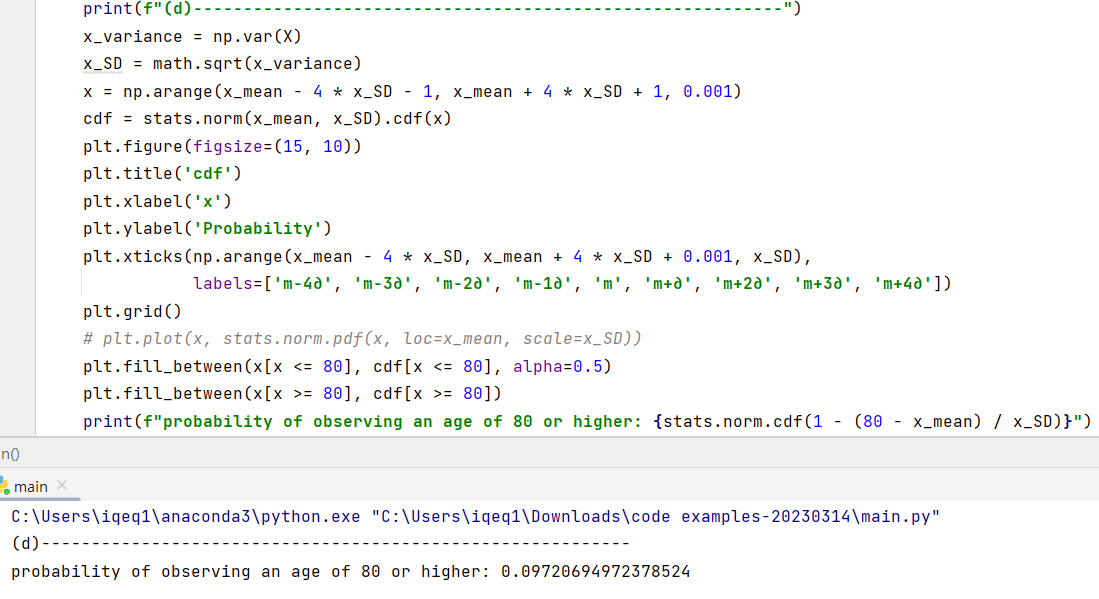
m=x\_means, ∂=x\_SD

(d) What is the probability of observing an age of 80 or higher?

- Calculate by z-score method and graph

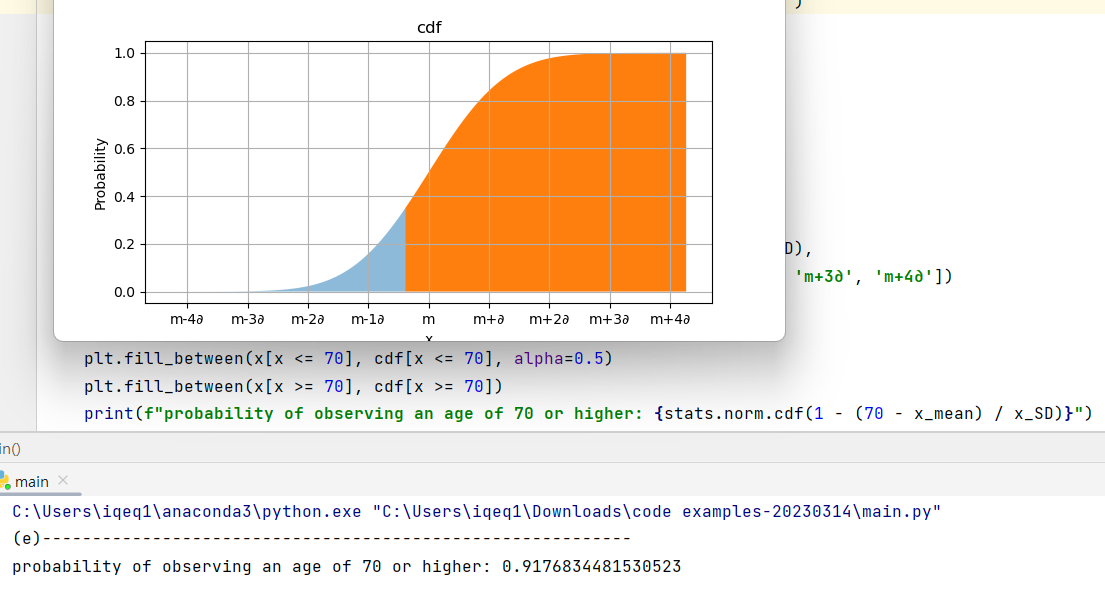


－ source code

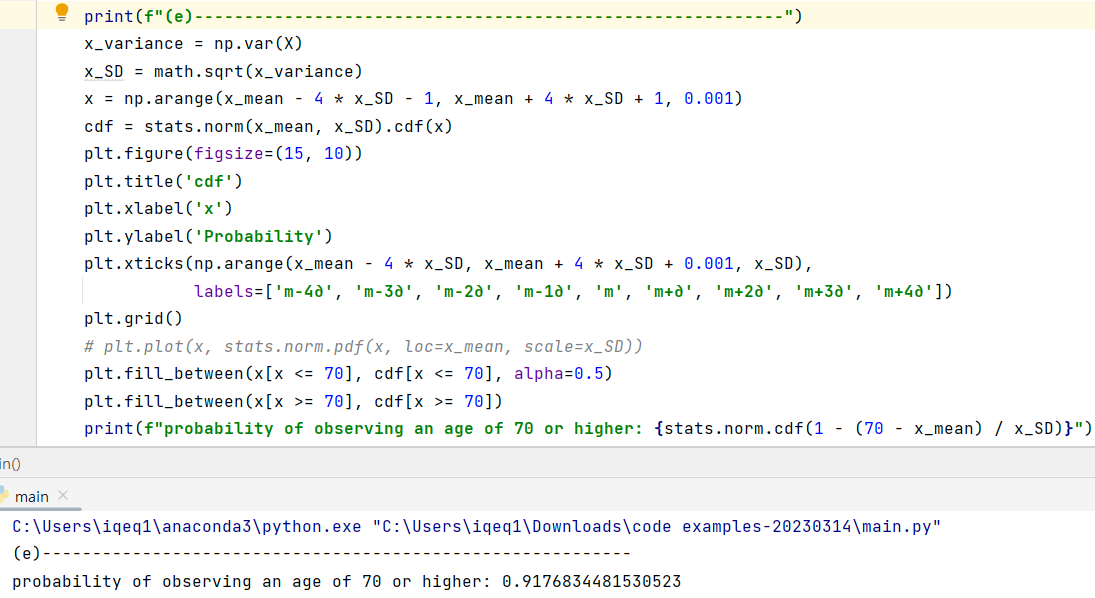


(e) What is the probability of observing an age of 70 or higher?

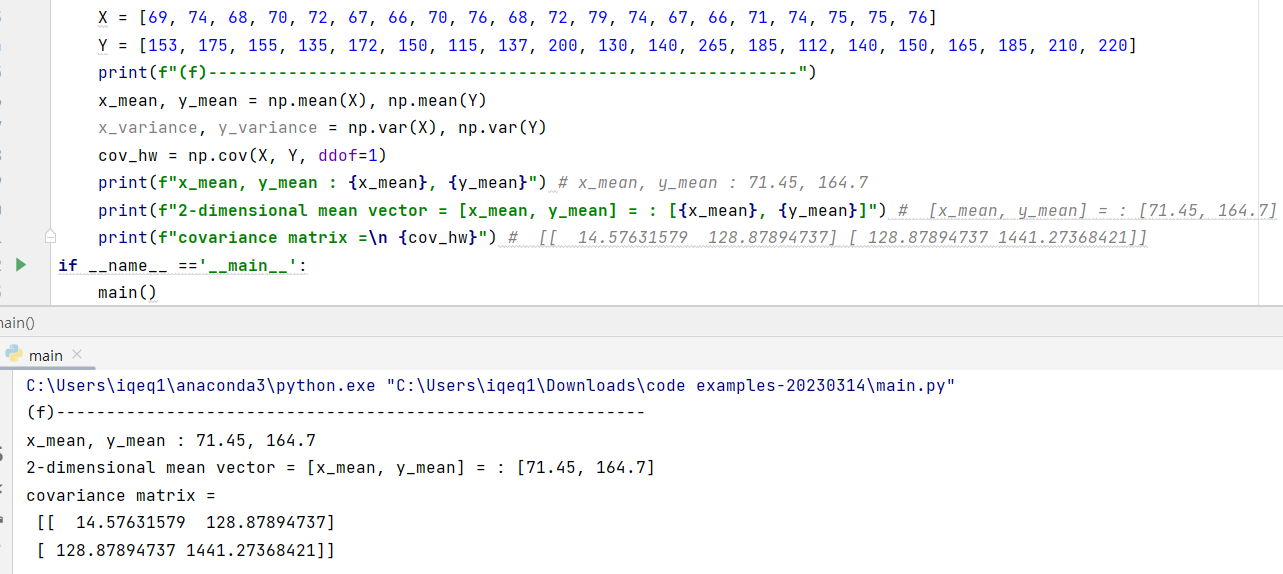
- Calculate by z-score method and graph



－ source code



(f) Find the 2-dimensional mean vector and the covariance matrix for these two variables.

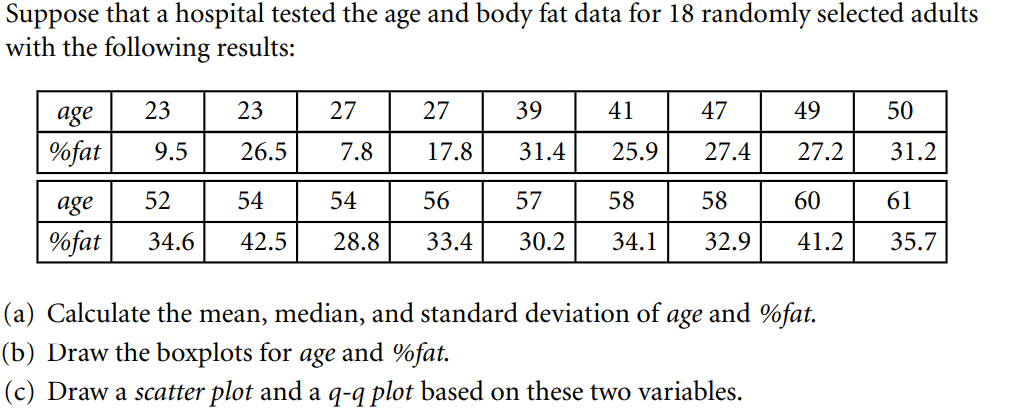
****

-covariance matrix

**≒ [ [ 14.58 128.88 ]**

**[ 128.88 1441.27]]**

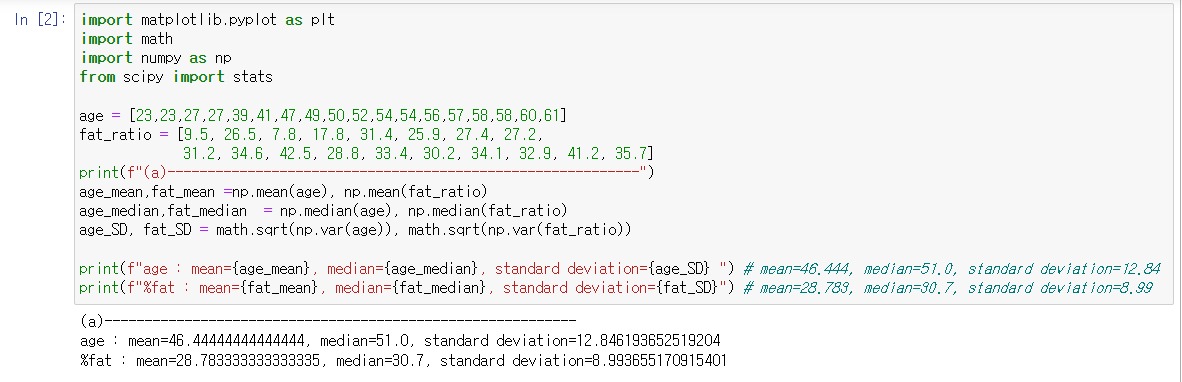
**Task-2**

****

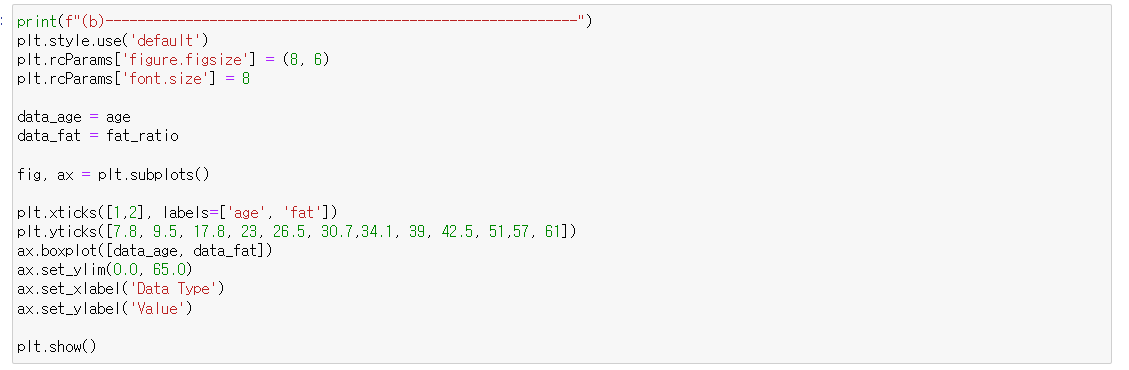
**Ans:**

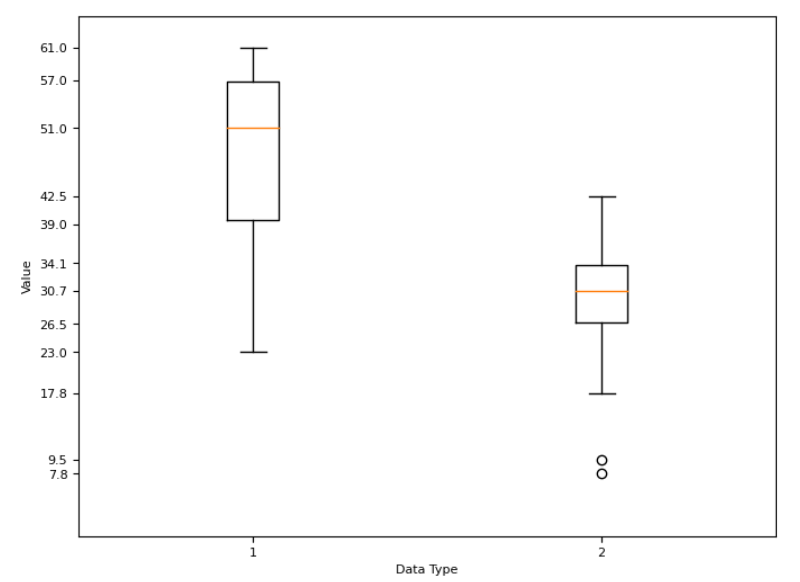
1. age: mean=46.44, median=51, standard deviation=12.85

%fat: mean=28.78, median=30.7, standard deviation=8.99

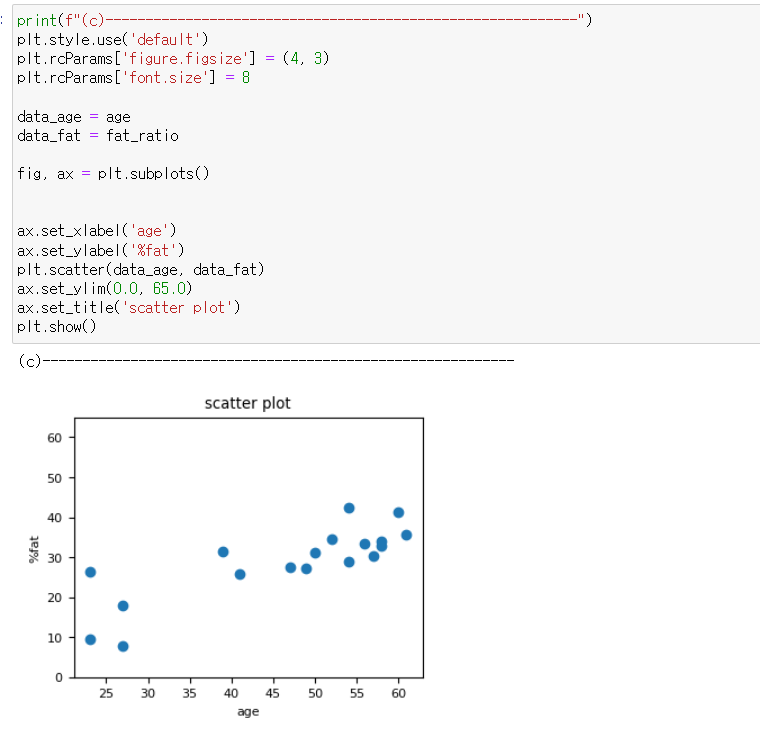


1. 1 : age 2 : %fat

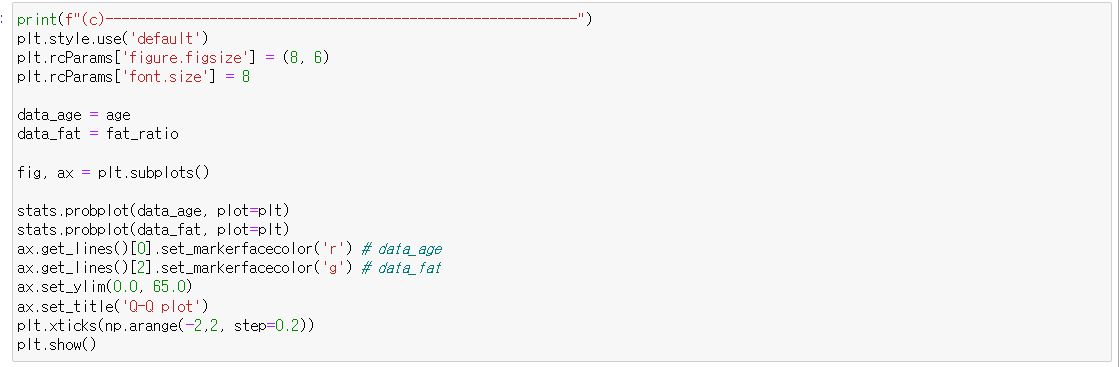


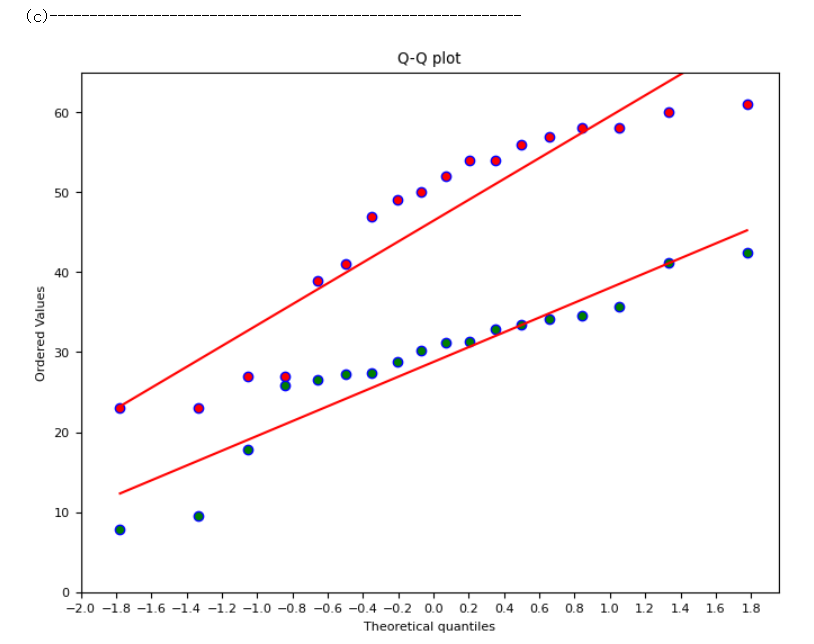


* scatter plot

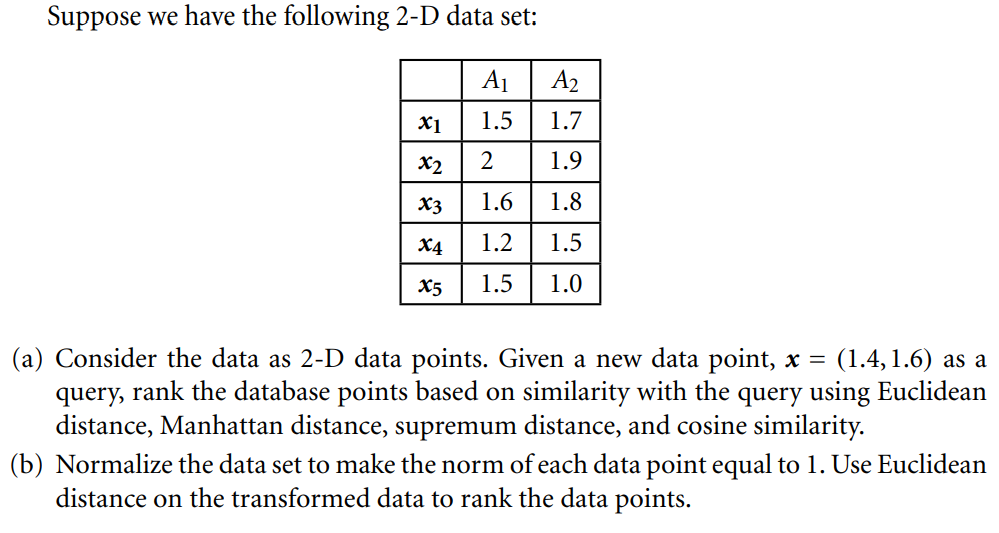


* q-q plot





**Task-3**

****

**Ans:**

**Write your Answer here**

(a)

- Euclidean distance( ED(A,B) )

x1, x4, x3, x5, x2

- Manhattan distance( MD(A,B) )

x1, x4, x3, x5, x2

- supremum distance( SD(A,B) )

x1, x3=x4, 2=x5

- cosine similarity ( CS(A, B) )



x1,x3,x4,x2,x5

(b)

- Normalizing the data

x : ( 0.6585, 0.7526 )

x1: ( 0.6616, 0.7498 )

x2: ( 0.7250, 0.6887 )

x3: ( 0.6644, 0.7474 )

x4: ( 0.6247, 0.7809 )

x5: ( 0.8321, 0.5547 )

- New eucildean distance

x1: ( 0.0041 )

x2: ( 0.0922 )

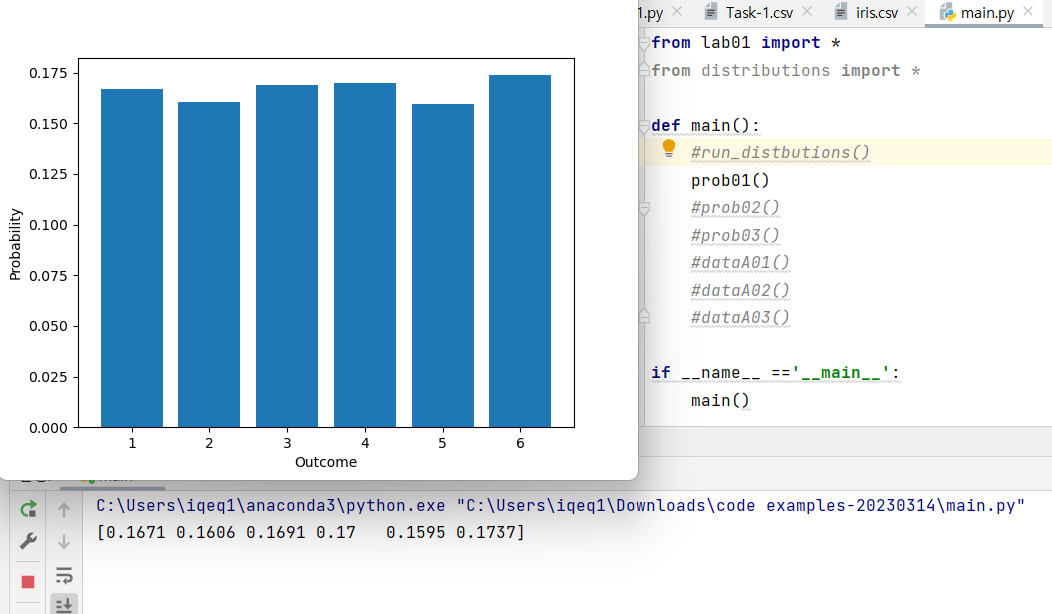
x3: ( 0.0078 )

x4: ( 0.0441 )

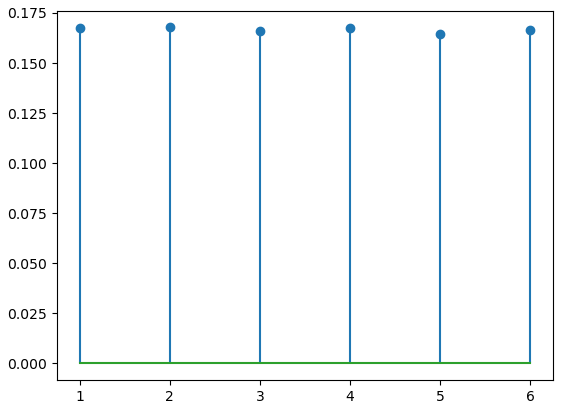
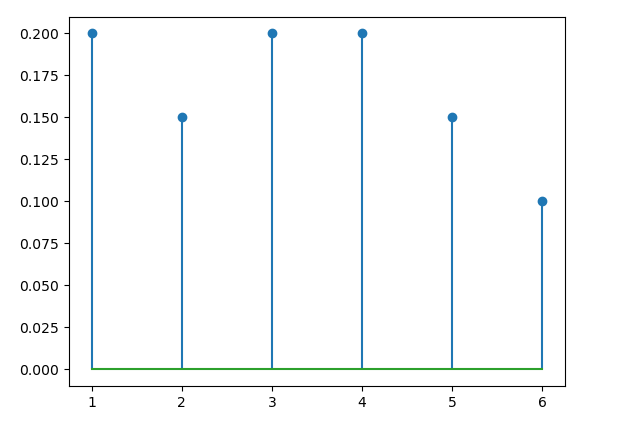
x5: ( 0.2632 )

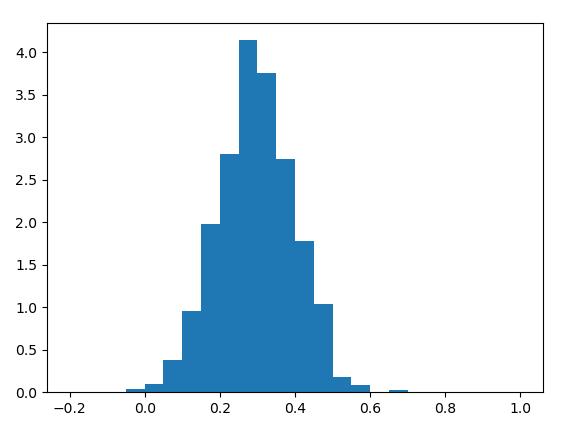
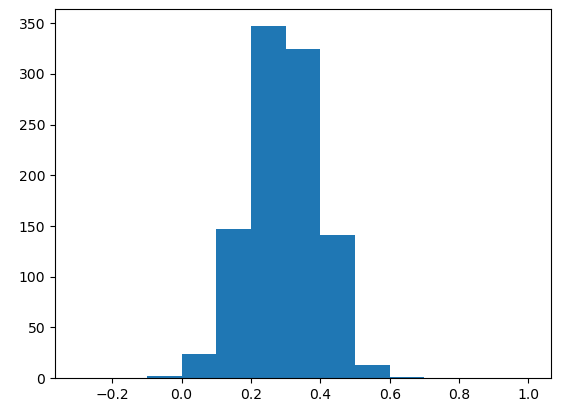
**Task-4: Run the Python code and record the results.**

=prob01()

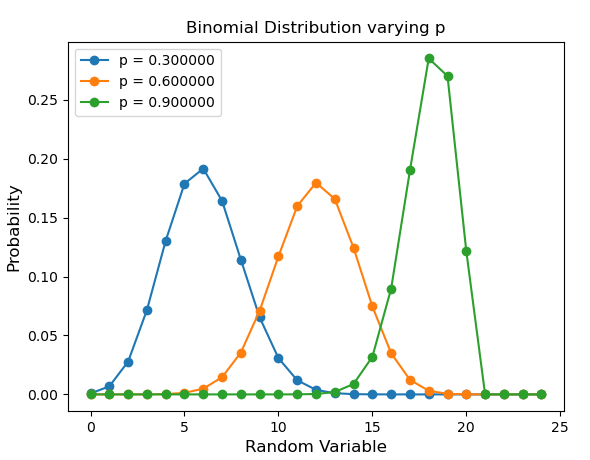
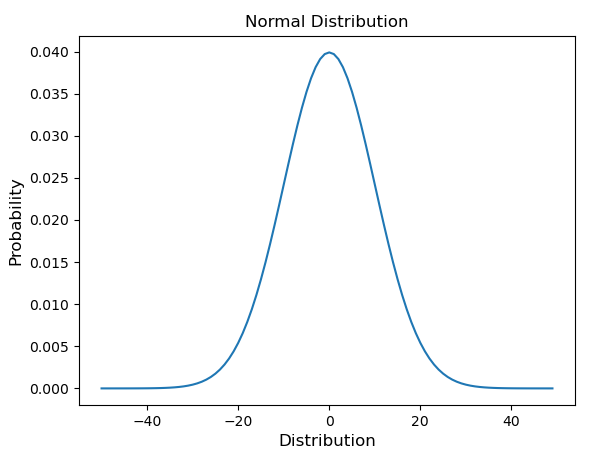


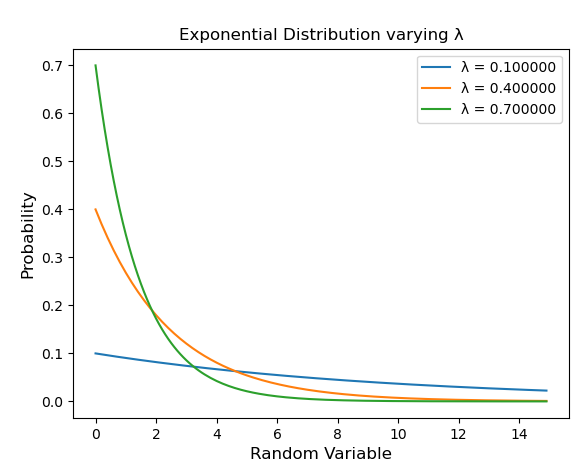
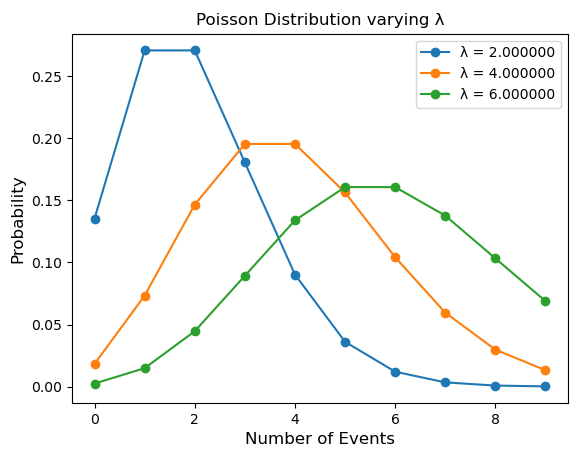
=prob02()

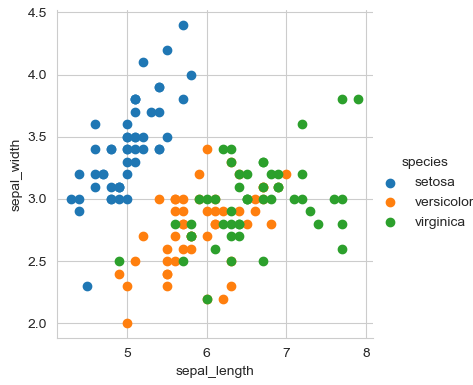
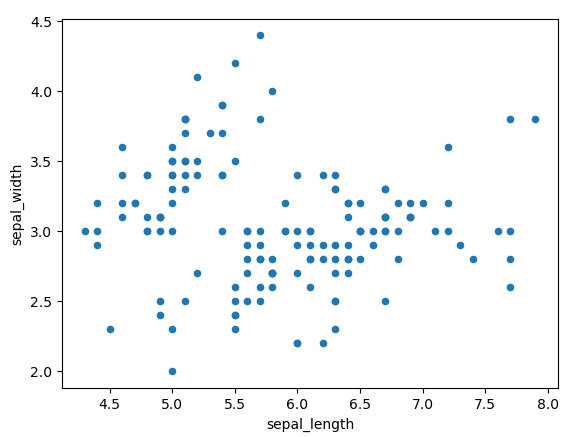


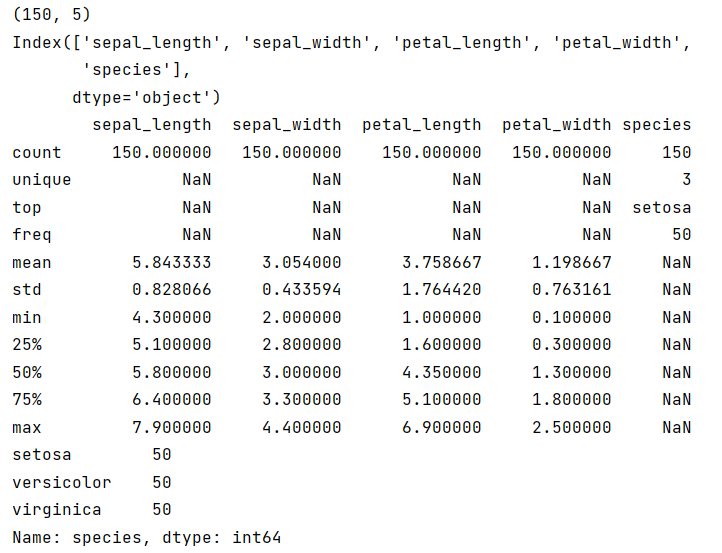
****

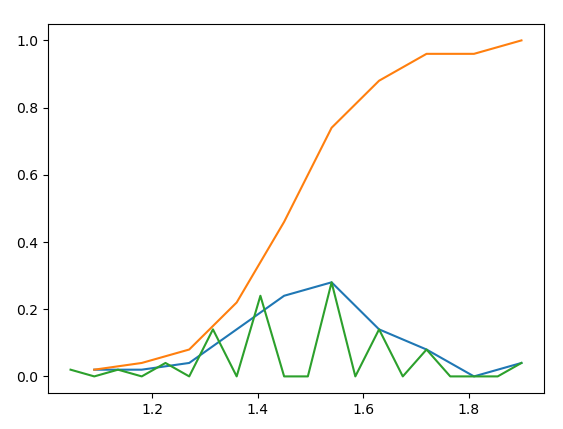
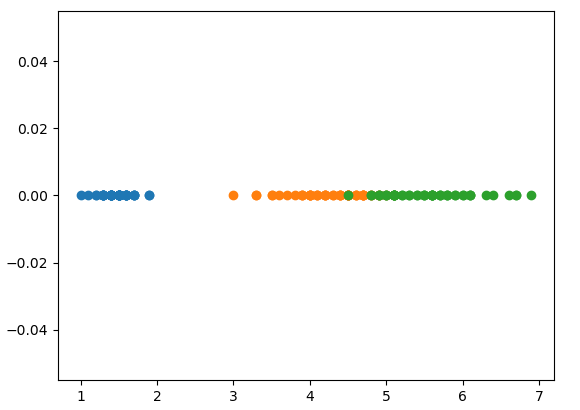
=prob03()

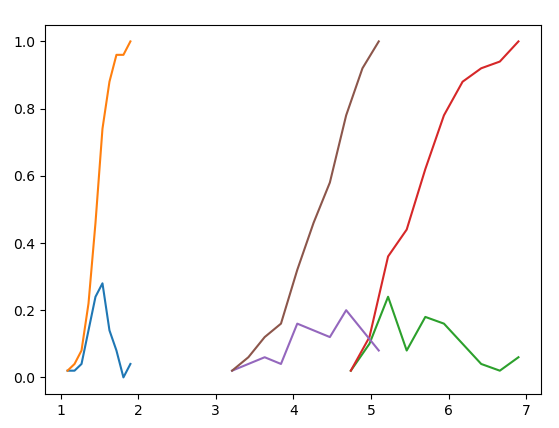
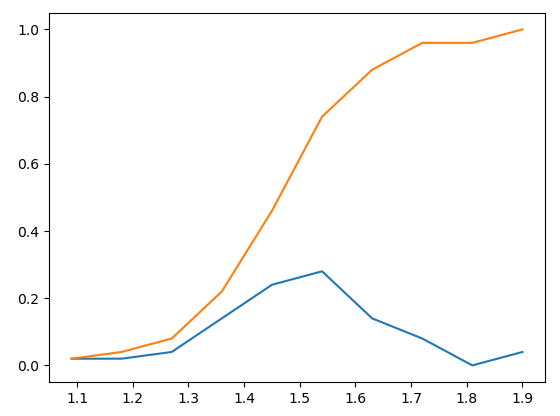
****

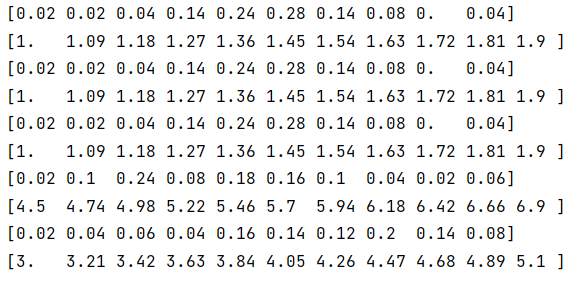
****=dataA01()

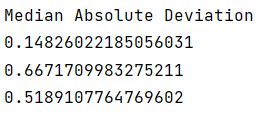
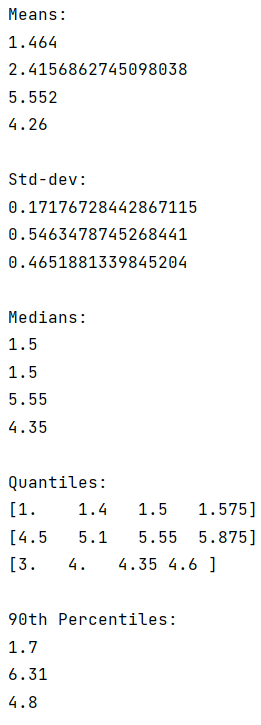
****

****=dataA02()





****=dataA03()

****

**Comments**

Write any comments about the task. For example, what you have learnt from the task, it was helpful in understanding concepts etc.

**Thank you for the fun assignment.**

**Thanks to that, I was able to review linear algebra and pyplot.**