Questions

1. **Correlation** is a statistical method that measures the similarity of the variation between two random vectors. The correlation coefficient (value in between -1 to +1) in between two vectors can be calculated with the help of the given formula

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\left[n\sum x_i^2 - \left(\sum x_i\right)^2\right]^{\frac{1}{2}} \left[n\sum y_i^2 - \left(\sum y_i\right)^2\right]^{\frac{1}{2}}}$$

Where, n = sample size
x and y are the sample points with index i.

When one variable increases as the other increases the correlation is positive. If one decreases as the other increases it is negative. Complete absence of correlation is represented by 0.
Keeping the definition in mind, and considering (x(i),y(i)) as a single point in 2D plane (i is from 1 to 14).

(i) Plot the variables (x and y) as scatter plots given below
(ii) Find out the correlation between the different cases of two variables (x and y) as given below and use the equation (mentioned above). Analyze the numerical value and scatter plots of the variables (i.e. if correlation is positive, y should increase with increase in x).

x	3	-1	6	12	8	10	9	13	15	-3	0	-5	-2	16
y	4	-2	5	11	6	11	7	8	9	-4	1	-6	-1	14

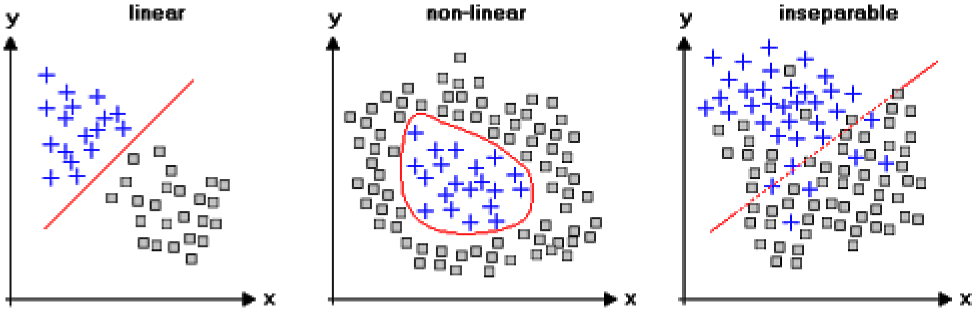
b)

x	3	-1	6	12	8	10	9	-5	-2	4	2	0	14	11
y	0	2	-3	4	0	1	2	-1	-2	2	1	-2	9	2

c)

x	3	4	2	5	8	-8	-6	-1	0	-5	4	2	6	8
y	5	-1	0	1	2	2	4	2	1	2	1	2	-4	4

2. Consider samples belonging to the data of two categories (two different classes) as given below (Each sample is 2D and each class is having 500 samples). Plot this 2D data as a scatter plot, representing the data for each category in a different colour. See if the data for two categories is linearly separated (two classes are linearly separated if a single line can separate those two classes). If not, then perform the following operation on the given samples. Calculate the mean of the individual classes. If the mean of a class lies in the first quadrant (i.e. x and y of the mean are positive) then add +k in y -dimension of each sample. If the mean of a class lies in the third quadrant , then add -k in the y-dimension of each sample. Decide k yourself. Again, plot this 2 class data and comment about the separability of the 2 classes. (Hint : See figure for data separability understanding)



3 . A survey has been conducted by a private firm in between customer care services of two banks. The bank in which the waiting time for a client to be served by a customer service representative is less will be declared as a better bank. The data for 20 customers for the waiting time (in seconds) in bank A is given below:

43.1	45.3	47.3	30.3	54.1
35.6	43.5	31.2	31.4	45.6
37.6	40.3	42.2	35.6	36.5
36.5	50.2	45.5	45.2	43.1

The data for 20 customers for the waiting time (in seconds) in bank B is given below:

33.1	35.3	37.3	32.3	44.1
31.6	33.5	35.2	30.4	35.6
32.6	42.3	40.2	36.6	38.5
36.5	30.2	42.5	50.2	40.1

Plot the histogram of the above data with 5-second bins (5-second intervals). Analyze the histogram and argue which bank is giving better services to its customer.

4 . The information about the number of children belonging to different age groups tabulated in the table given below.

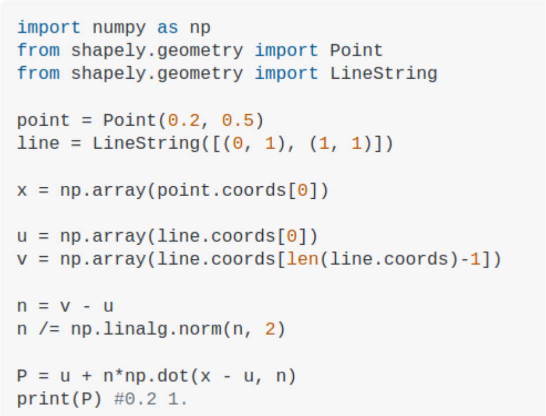
a) Plot this table as a graph, which is essentially the histogram representing the data.

b) Mark the points of mean, mode and median on the histogram, by vertical lines

Age (in years)	Number of Children
0 <= age < 2	8
2 <= age < 4	10
4 <= age < 6	18
6 <= age < 8	10
8 <= age < 10	12
10 <= age <= 12	6

5. Consider the given two class linearly separable data (i.e. the two classes can be separated with the help of a line). Project each point of this data on the line x = 0 (i.e on x-axis). After projection, plot the samples and comment on the separability of the two classes.

[Hint : The python code for the projection of a point on a line is given below. Here, the point is (0.2 , 0.5) and the line is passing through (0,1) and (1,1) and P is the projection of the point on the line.]



```python
import numpy as np
from shapely.geometry import Point
from shapely.geometry import LineString

point = Point(0.2, 0.5)
line = LineString([(0, 1), (1, 1)])

x = np.array(point.coords[0])

u = np.array(line.coords[0])
v = np.array(line.coords[len(line.coords)-1])

n = v - u
n /= np.linalg.norm(n, 2)

P = u + n*np.dot(x - u, n)
print(P) #0.2 1.
```

6. Given data (data.csv) is a single column data. Draw the box plot of this single column data to find out the range in which maximum data is lying (Interquartile range) .