CS 215 ASSIGNMENT

Tathagat Verma - 180050111
Neel Aryan Gupta - 180050067

Question 3

To use PCA to get best approximate linear relationship between X & Y, we follow following procedure;

First we subtract X_mean & Y_mean from the data points.

$$X\_mean = \frac{\sum x_i}{n} \qquad Y\_mean = \frac{\sum y_i}{n}$$

Call this matrix A. $\rightarrow n \times 2$.

$$A_{i1} = x_i - X_{mean}$$
$$A_{i2} = y_i - Y_{mean}.$$

Then we compute co-variance matrix of the modified x_i's & y_i's {co-variance matrix of X, Y}

using $\longrightarrow \boxed{A^T . A / n}$

~~$\frac{\sum x_i'^2}{n}, \frac{\sum y_i'^2}{n}$~~

$x_i' = x_i - X_{mean}$

$y_i' = y_i - Y_{mean}$

Now we compute the eigen vector corresponding to the highest eigen value for the co-variance matrix.

This vector corresponds to the direction along which there is maximum variance when the points are projected along this line.

(This follows from the PCA analysis)

~~This~~

So now we can generate a line having slope ~~equal~~ (direction) same as that along this eigen vector. & the mean of the dataset ie $\left(\frac{\Sigma x_i}{n}, \frac{\Sigma y_i}{n}\right)$ lying on this line.

This is the line that will best approximate a linear relationship between X & Y. as there is maximum variance when these points are projected on the line predicted.

For the 2nd dataset, the linear approximation is not good.
This is majorly because the points of X, Y follow more likely a quadratic relationship, not linear.
PCA is not working well here because it just gives that line projecting on which we get maximum variance, so it works well when ~~that~~ data itself follows a linear relationship. which clearly isn't the case here.
Hence the quality of the linear approximation is not good.