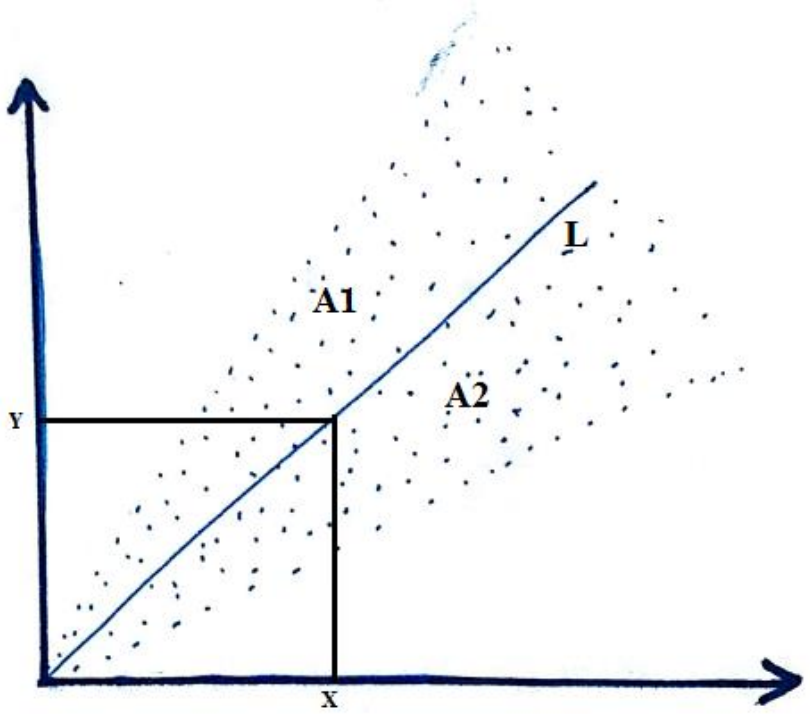
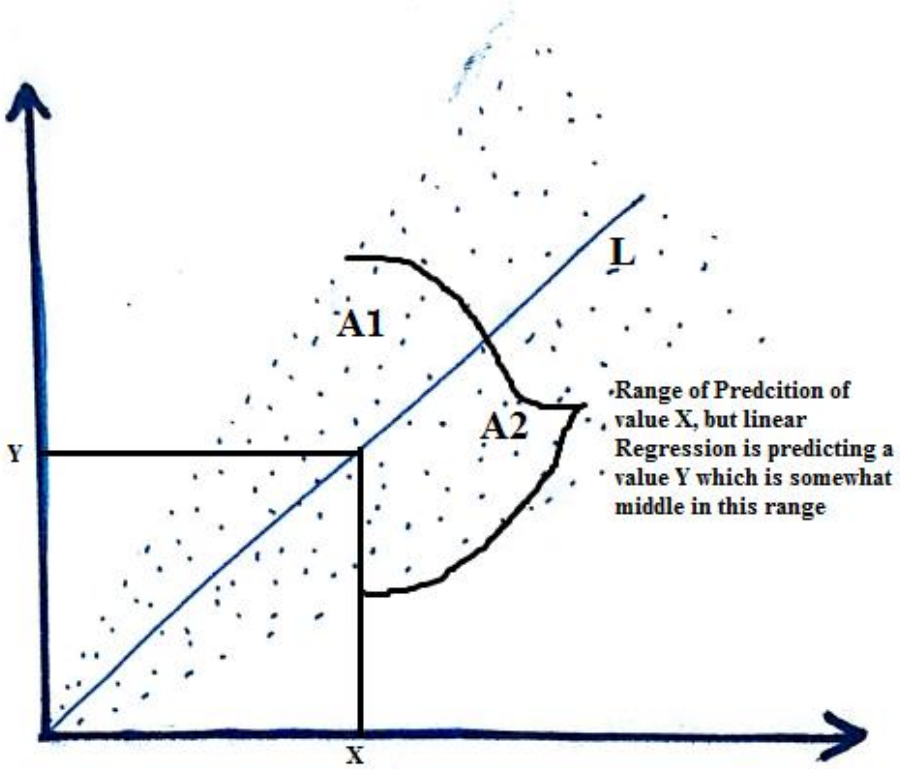
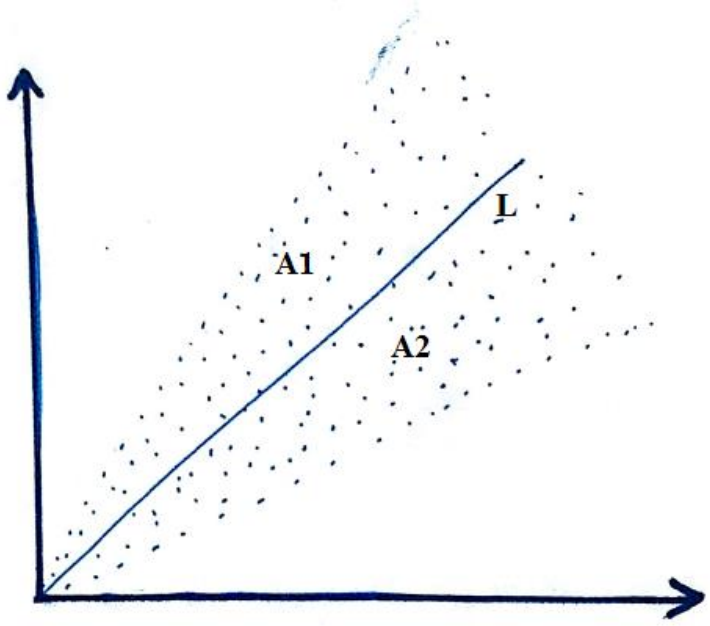
1. As Linear Regression Fits only one line on the training data, for the following figure it will predict the value Y for X as shown in figure below.



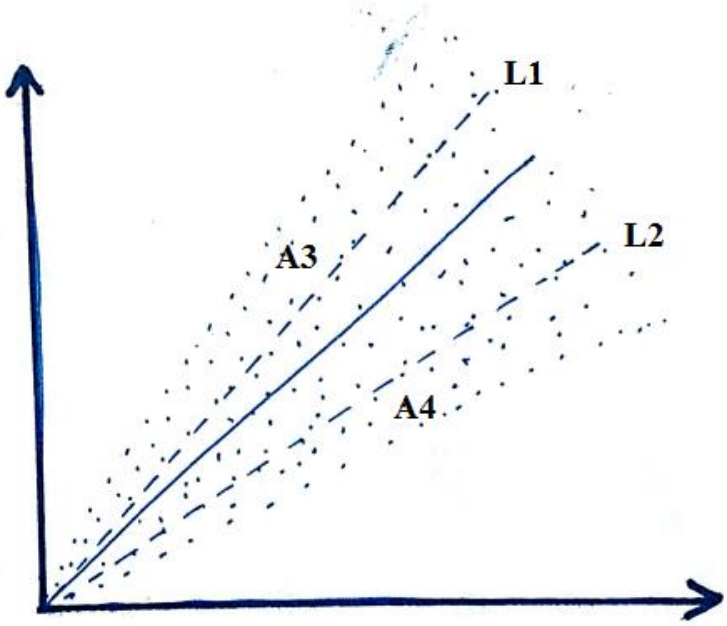
2. But as the range for prediction of X is more as shown in Figure below, the Y is being predicted in somewhat middle of that range so this is a bit of disadvantage of using Linear Regression. So to handle this disadvantage we fit two lines instead of one line to the training data.



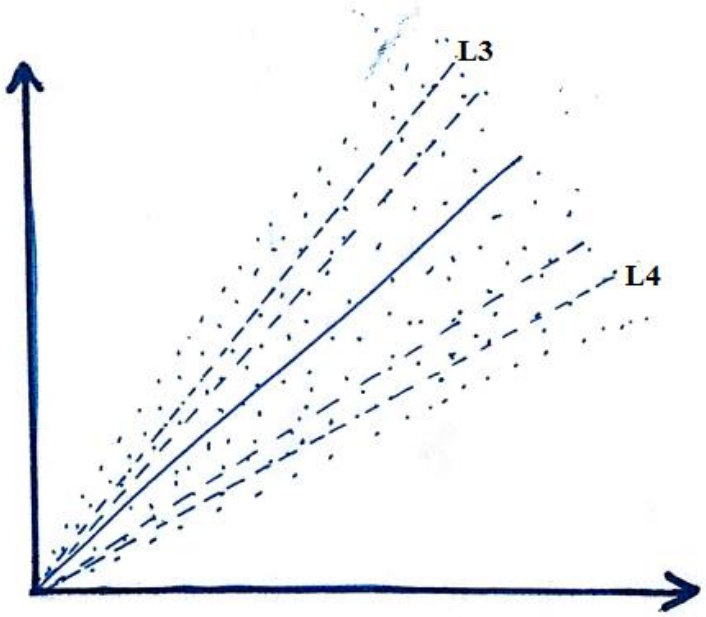
3. In Step 1 we used simple Linear Regression in order to fit line L on the training Data.(figure below)



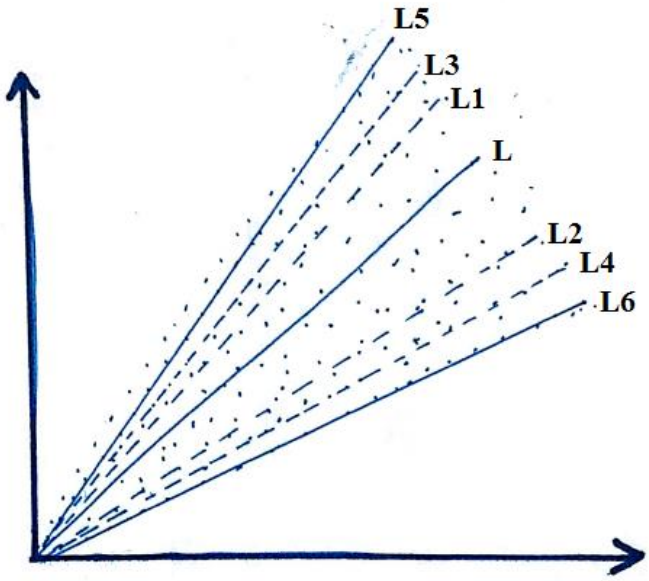
4. For step 2 we used simple Linear Regression for Data points of Area A1 and Area A2 individually to fit line L1 and L2. (figure below)



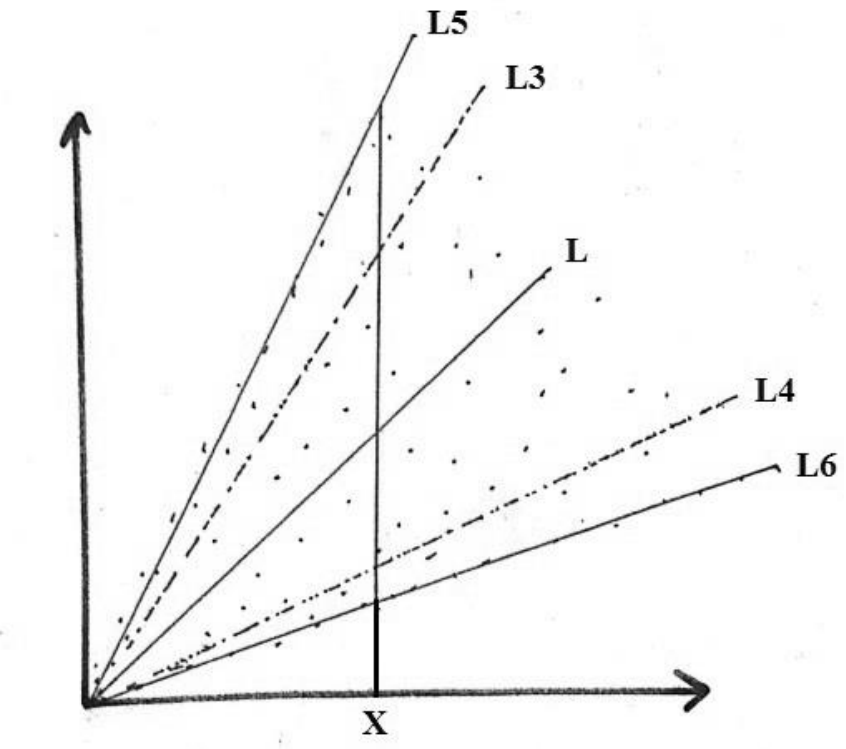
5. For step 3 we used simple Linear Regression for Data points of Area A3 and Area A4 individually to fit line L3 and L4. (figure below)



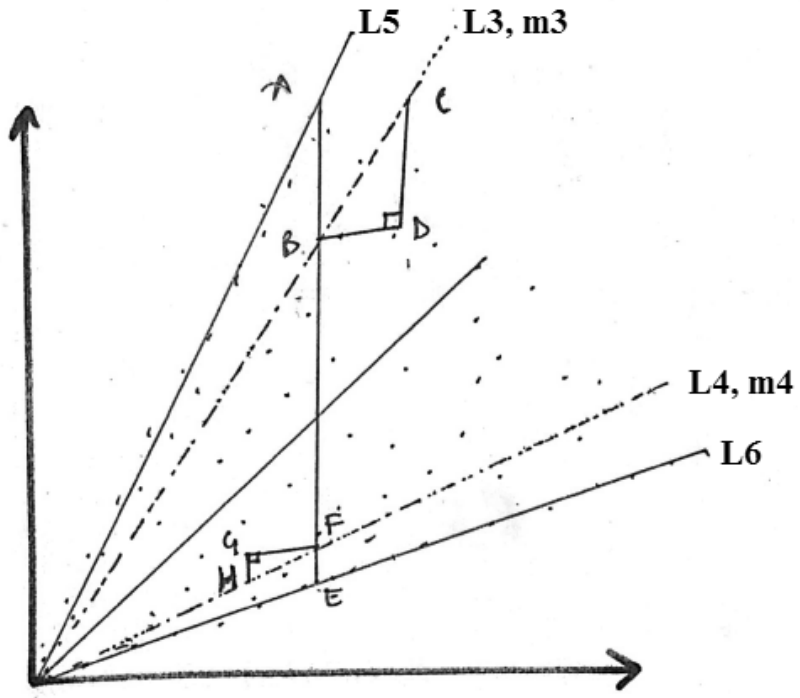
6. These steps, that is step 1, step 2, step 3 are just iterations, so suppose we want to fit L3 and L4 to our data then we input the parameter NO_OF_ITERATION=3. After the step 3 we have fitted two lines that is L3 and L4 to our training data.
7. Now we try to fit the whole data between two lines L5 and L6 in Step 4 by using large iterations like 100. Note: the main two lines that we have fitted to our training data are L3 and L4. L5 and L6 are just to help us in finding the prediction.



8. Now after Step 4 as shown in the Figure below we are left with L3, L4: Main fitted lines, L5, L6: Lines helping us in predicting. **(**as previous figure was small and I could not write each thing in that small figure hence I used new figure to show the details but actual figure remains the same)



9. Now suppose slope of **L3** is **m3** and slope of **L4** is **m4.** Considering both **Figures below** suppose **AB=CD=d1** and **GH=EF=d2.** In this, we need to find AB, CD, BD, EF, GH and GF.



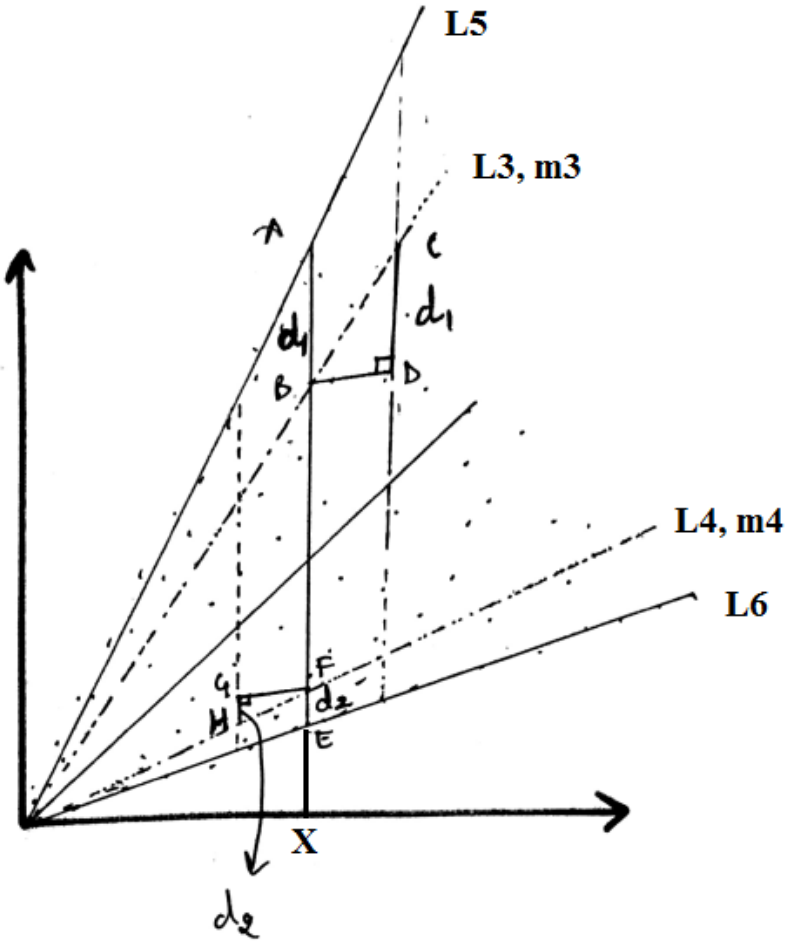


Figure 6

- Now from equations of L3, L4, L5, L6 and x=X, we can find distance between **AB=CD=d1** and **EF=GH=d2**.
- Now for finding BD:

$$tanB = m3 = \frac{d1}{BD}$$

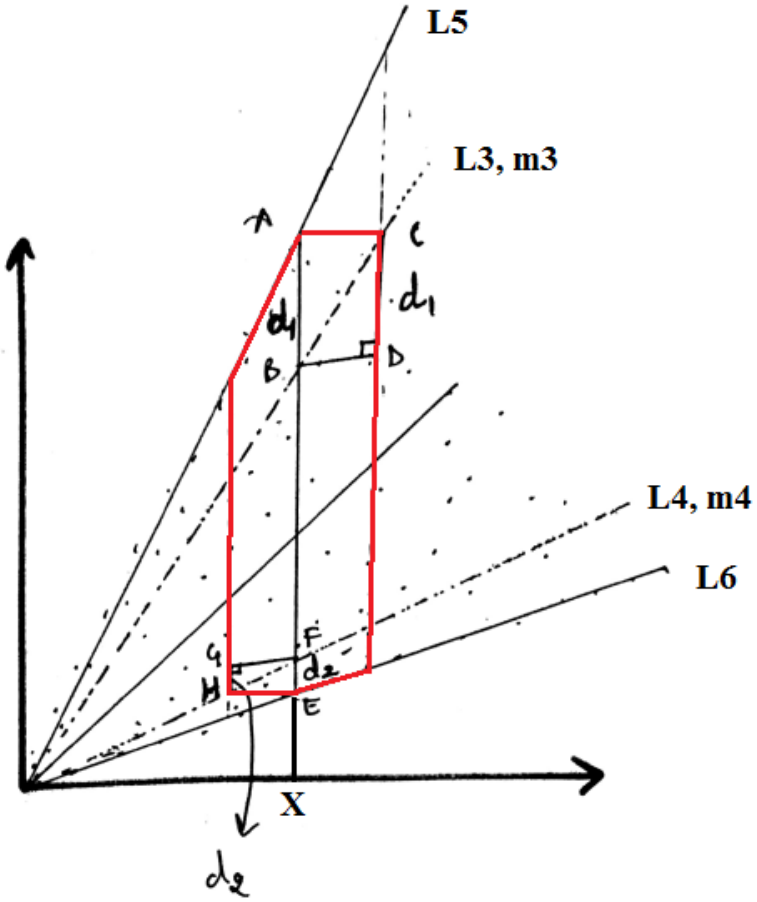
$$BD = d1/m3$$

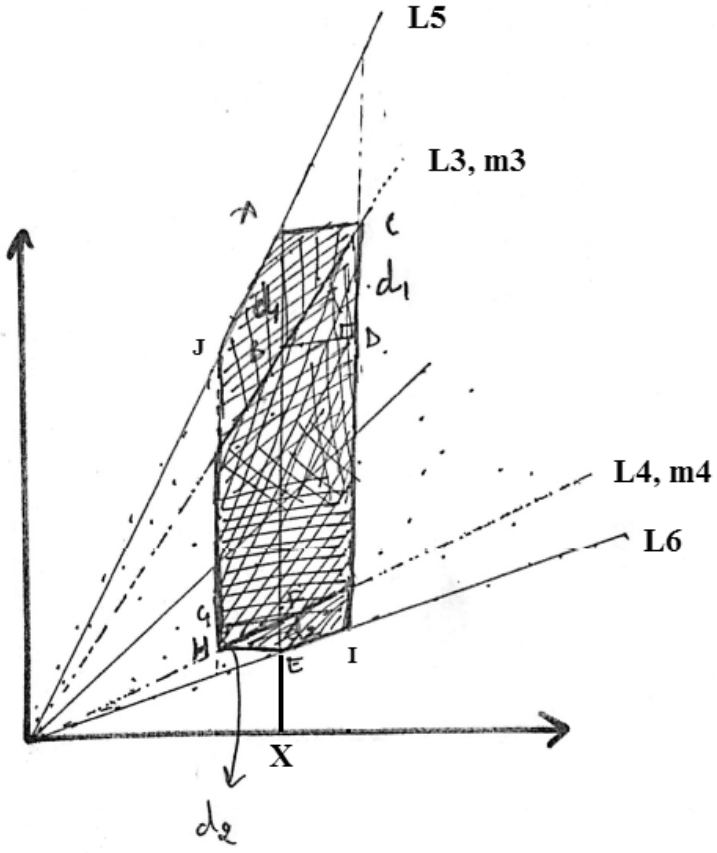
- For Finding GF:

$$tanF = m4 = \frac{d2}{GF}$$

$$GF = d2/m4$$

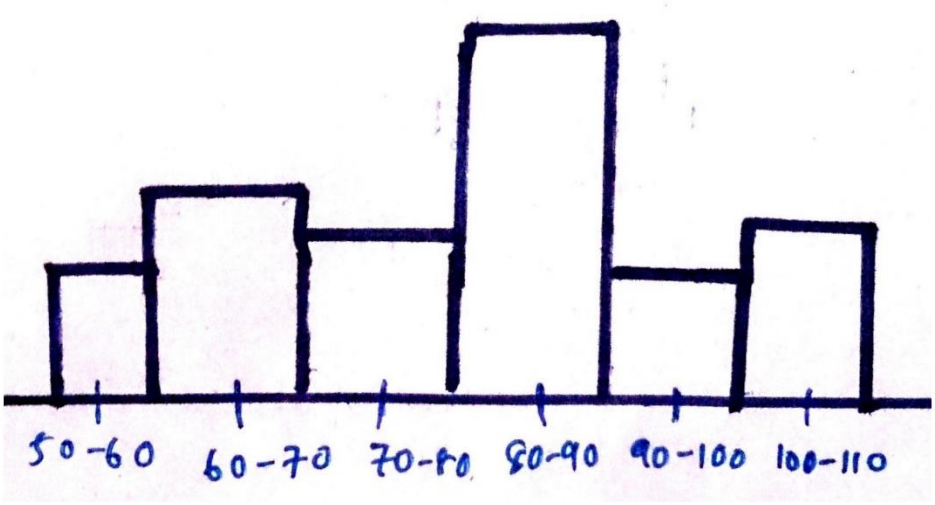
10. Now we have BD and GF, we make and Area as shown in **Figure below** by using points A, C, I, E, H, J where both A, C, I, E and A, E, H, J are trapezium**.**


L5
L3, m3
L4, m4
L6
A
C
d_1
d_1
B
D
F
G
d_2
H
E
X
d_2


L5
L3, m3
L4, m4
L6
A
C
d_1
d_1
J
D
G
H
E
I
X
d_2

Note: In this we considered **L3** and **L4** by using NO_OF_ITERATIONS=3 and not large iterations like NO_OF_ITERATIONS=50 (because for iterations like 50 we get thinner area) and made an Area out of these lines, to avoid overfitting, The Area which we made in the above figure consists of the data points that are in the range between y-coordinate of **L5** and **L6** at **x=X.**

11. Now we make certain intervals, these intervals are defined by the user. Suppose interval = 10
12. Take all the data points in that area and plot the Histogram between number of points and intervals as shown in **Figure below**.



Intervals

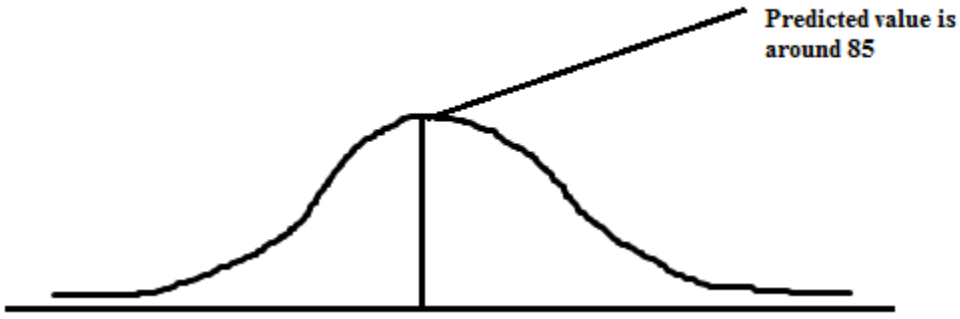
Now we can see that between 80-90 there is most number of data points hence predicted value is (80+90)/2=85

Note:

- No of iterations are defined by the user
- Interval size if defined by the user

Note:

- We can also make Gaussian distribution instead of making Histogram as shown below. In this case we do not need to define the interval



- We can also put more weight to the points closer to the line x=X in order to calculate Gaussian distribution like in locally weighted Regression the below figure depicts that.