K-means clustering

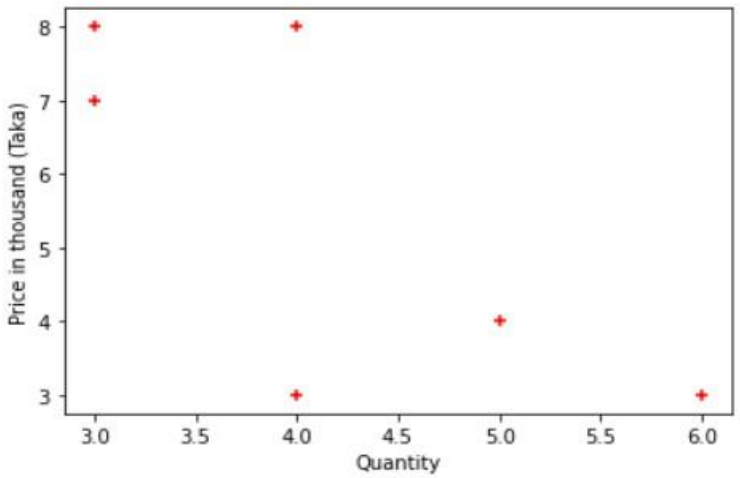
Euclidean distant: $d = \sqrt{(x2 - x1)^2 + (y2 - y1)^2}$

Elbow method: **Use for find the number of optimal cluster value. That's mean how many clusters should be best for particular datasets.**

Out[4]:

	Product	Quantity	Price(K)
0	FaceWash	3	7
1	Cream	5	4
2	Shoes	4	3
3	Bags	4	8
4	Jacket	6	3
5	Shirt	3	8

```
1 plt.scatter(x=data['Quantity'], y=data['Price(K)'], marker="+", color='red')
2 plt.xlabel("Quantity")
3 plt.ylabel("Price in thousand (Taka)")
4 plt.show()
```



Now we will calculate the which point will be gone which cluster:

At first, we pick tow data point or any two data point randomly and then from dose data point we will calculate the distant for the data point. We will take the cluster value means 'K' value. It depends on us. We will select the cluster value 2 for this dataset.

Let's assume our first cluster data point is C1 and second is C2. So, data point will be C1(3,7) and C2(5,4).

For first data point (3,7) or Facewash:

Distant from c1 to c1 = $\sqrt{(x2-x1)^2+(y2-y1)^2}$

= $\sqrt{(3-3)^2+(7-7)^2}$

= **0**

Distant from c2 = $\sqrt{(5-3)^2+(4-7)^2}$

= 4.24

C1 < C2 so first data point will go C1(3,7) cluster.

For second data point (5,4) or Cream:

Distant from c2 to c2 = $\sqrt{(5-5)^2+(4-4)^2}$

= 0

Distant from c1 = $\sqrt{(5-3)^2+(4-7)^2}$

= $\sqrt{13}$

= 3.60

C2 < C1 so first data point will go C2(5,4) cluster.

For third data point (4,3) or Shoes:

Distant from c1 = $\sqrt{(4-3)^2+(3-7)^2}$

= 4.123

Distant from c2 = $\sqrt{(4-5)^2+(3-4)^2}$

= $\sqrt{13}$

= 1.41

C1 > C2 so third data point will go C2 cluster.

Now, new centroid in for C2 cluster = $(\frac{5+4}{2}, \frac{4+3}{2})$

C2 = (4.5, 3.5)

In C2 cluster have two data point these are Cream (5,4) and Shoes (4,3). So, we took the average value of these for calculate the new centroid.

For 4th data point (4,8) or Bags:

Distant from c1 = $\sqrt{(4-3)^2 + (8-7)^2}$

= $\sqrt{2}$

= 1.41

Distant from c2 = $\sqrt{(4-4.5)^2 + (8-3.5)^2}$

= 20.50

C1 < C2 so first data point will go C1 cluster.

Now, new centroid in for C1 cluster = $(\frac{3+4}{2}, \frac{7+8}{2})$

C1 = (3.5, 7.5)

In C1 cluster have two data point these are Facewash (3,7) and Bags (4,8). So, we took the average value of these for calculate the new centroid.

For 5th data point (6,3) or Jacket:

Distant from c1 = $\sqrt{(6-3.5)^2 + (3-7.5)^2}$

= 26.5

Distant from c2 = $\sqrt{(6-4.5)^2 + (3-3.5)^2}$

= 2.50

C1 > C2 so 5th data point will go C2 cluster.

Now, new centroid in for C2 cluster = $(\frac{5+4+6}{3}, \frac{4+3+3}{3})$

C2 = (5, 3.33)

In C2 cluster have three data point these are Cream (5,4), Shoes (4,3) and Jacket (6,3). So, we took the average value of these for calculate the new centroid.

For 6th data point (3,8) or Shirt:

Distant from c1 $= \sqrt{(3-3.5)^2 + (8-7.5)^2}$

$= 0.70$

Distant from c2 $= \sqrt{(3-5)^2 + (8-3.33)^2}$

$= 2.48$

C1 < C2 so 6th data point will go C1 cluster.

Now, new centroid in for C1 cluster = $(\frac{3+4+3}{3}, \frac{7+8+8}{3})$

C1 = (3.33, 7.67)

In C1 cluster have two data point these are Facewash (3,7), Bags (4,8) and Shirt (3,8). So, we took the average value of these for calculate the new centroid.

So, our cluster C1(3.33, 7.67) and C2 is (5, 3.33)