ZIAUDDIN UNIVERSITY

ASSIGNMENT NO: 1

SUBMITTED BY:

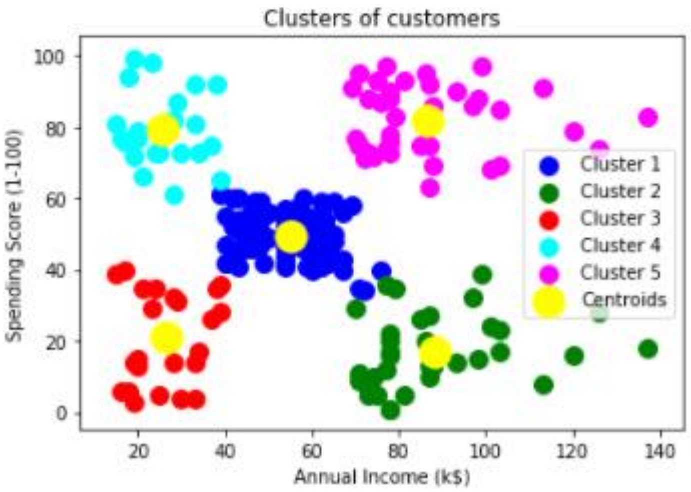
MIRZA ABDULLAH BAIG
4-23/2019/004

SUBMITTED TO:

SIR SALMAN AKBAR
DATA MINING

INDEX

Example of K-Median Algorithm:



The output image is clearly showing the five different clusters with different colors. The clusters are formed between two parameters of the dataset; Annual income of customer and Spending. We can change the colors and labels as per the requirement or choice. We can also observe some points from the above patterns, which are given below:

- **Cluster1** shows the customers with average salary and average spending so we can categorize these customers as **balanced.**

- **Cluster2** shows the customer has a high income but low spending, so we can categorize them as **careful.**

- **Cluster3** shows the low income and also low spending so they can be categorized as **sensible.**

- **Cluster4** shows the customers with low income with very high spending so they can be categorized as **careless.**

- **Cluster5** shows the customers with high income and high spending so they can be categorized as **target**, and these customers can be the most profitable customers for the mall owner.

Algorithm of K-Median:

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be other from the input dataset).

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

Step-4: Calculate the median and place a new centroid of each cluster.

Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.

Step-7: The model is ready.

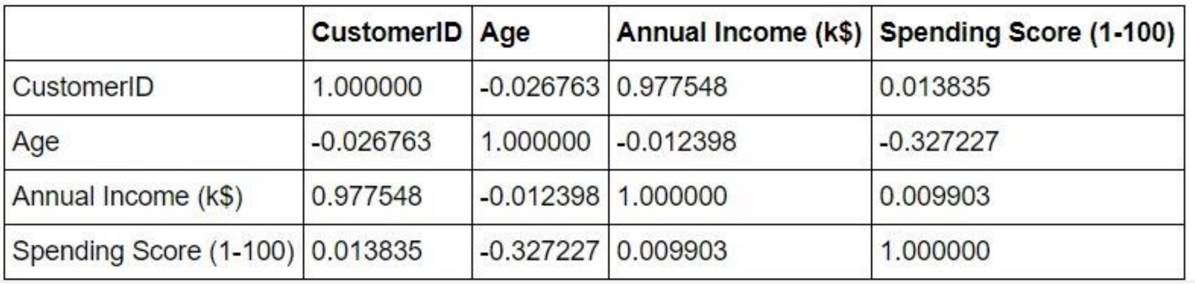
Realia of K-Median Algorithm:

Mall Customer Data: Implementation of K-Median:

Mall Customer data is an interesting dataset that has hypothetical customer data. It puts you in the shoes of the owner of a supermarket. You have customer data, and on this basis of the data, you have to divide the customers into various groups.

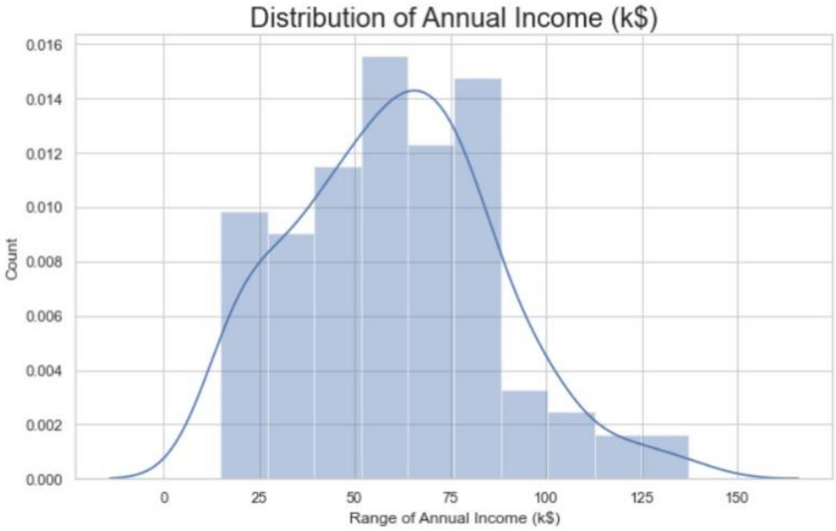
The data includes the following features:

1. Customer ID
2. Customer Gender
3. Customer Age
4. Annual Income of the customer (in Thousand Dollars)
5. Spending score of the customer (based on customer behaviour and spending nature)

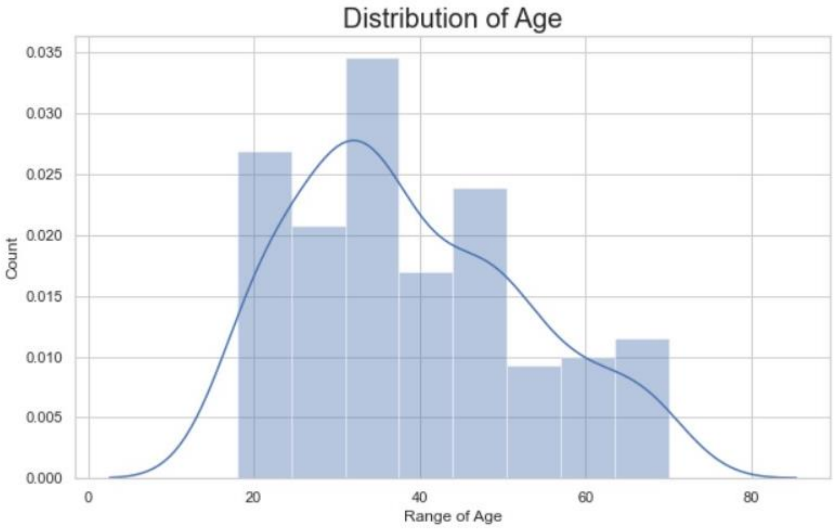


	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
CustomerID	1.000000	-0.026763	0.977548	0.013835
Age	-0.026763	1.000000	-0.012398	-0.327227
Annual Income (k$)	0.977548	-0.012398	1.000000	0.009903
Spending Score (1-100)	0.013835	-0.327227	0.009903	1.000000

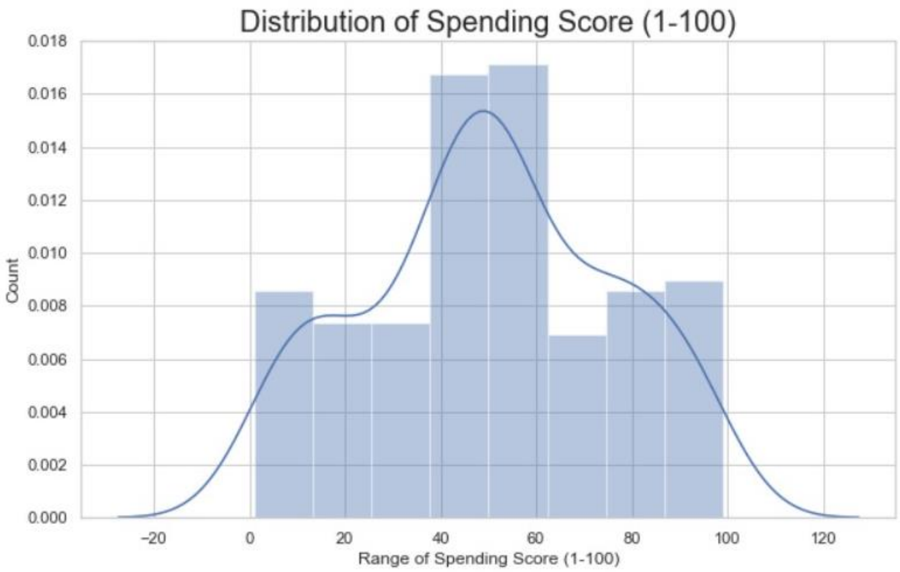
The data seems to be interesting. Let us look at the data distribution.



Most of the annual income falls between 50K to 85K.

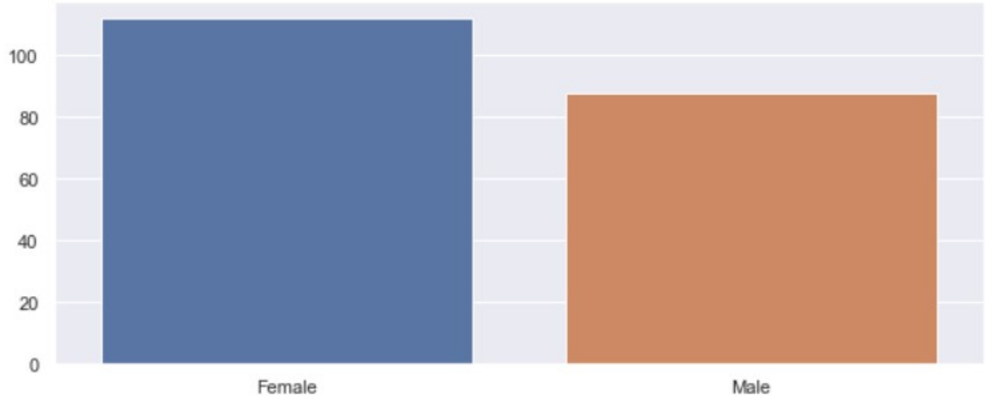


There are customers of a wide variety of ages.



The maximum spending score is in the range of 40 to 60.

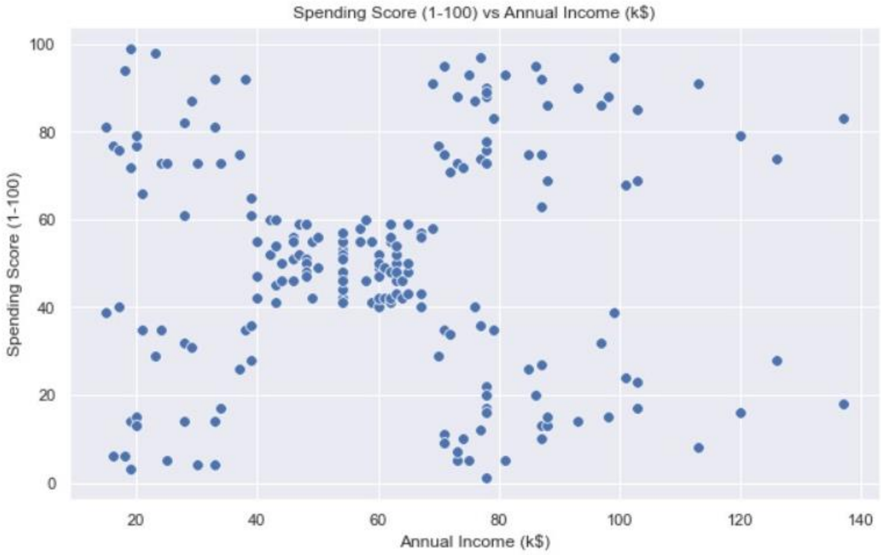
Gender Analysis:



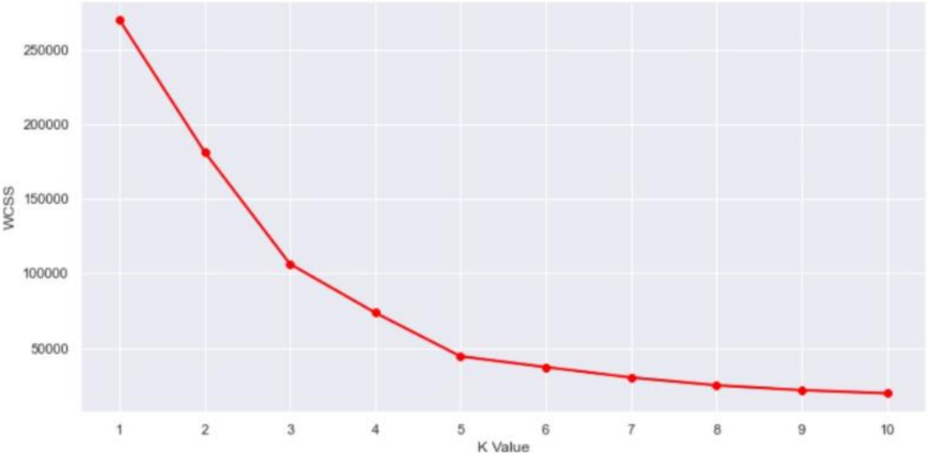
More female customers than male.

MAKING CLUSTERS:

- We take just the Annual Income and Spending score
- Scatterplot of the input data

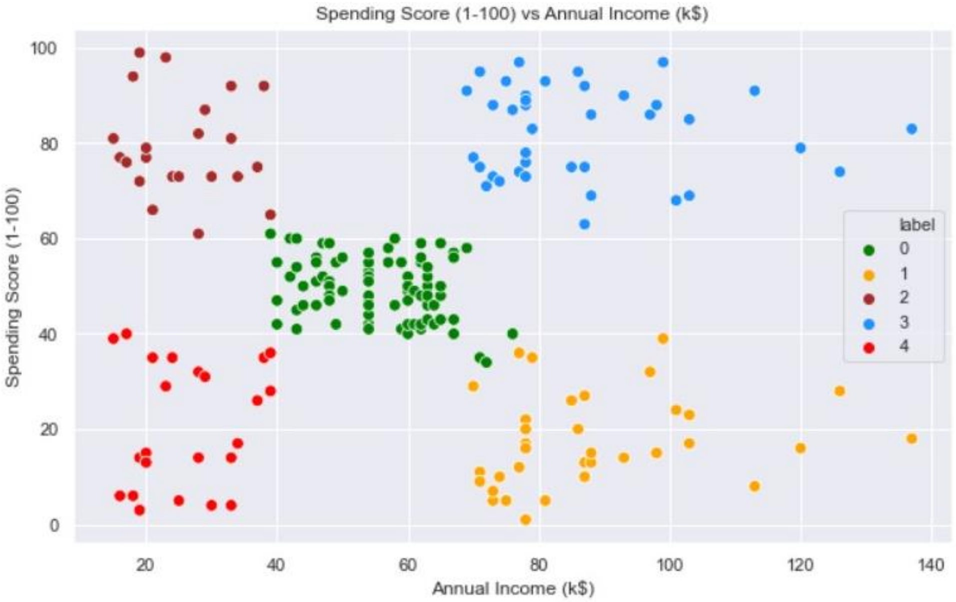


The plot:



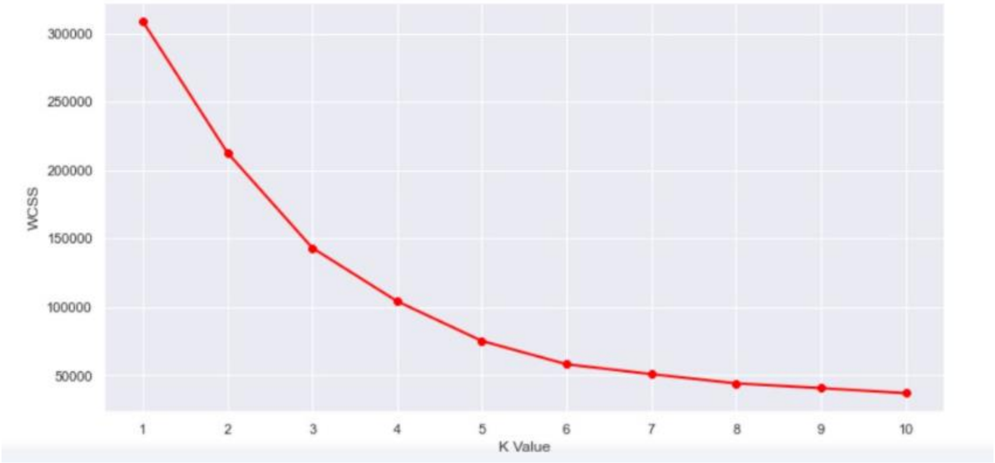
This is known as the elbow graph, the x-axis being the number of clusters, the number of clusters is taken at the elbow joint point. This point is the point where making clusters is most relevant as here the value of WCSS suddenly stops decreasing. Here in the graph, after 5 the drop is minimal, so we take 5 to be the number of clusters.

- Taking 5 clusters
- Scatterplot of the clusters



We can clearly see that 5 different clusters have been formed from the data. The red cluster is the customers with the least income and least spending score, similarly, the blue cluster is the customers with the most income and most spending score.

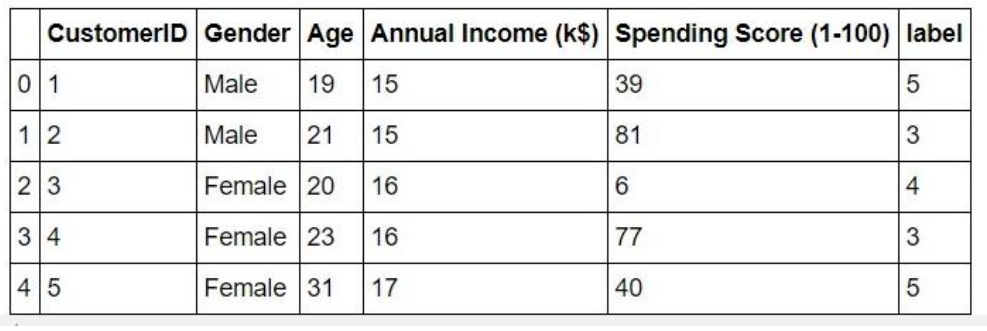
- Now, we shall be working on 3 types of data. Apart from the spending score and annual income of customers, we shall also take in the age of the customers.
- Now we calculate the Within Cluster Sum of Squared Errors (WSS) for different values of k.



Here can assume that K=5 will be a good value.

- We choose the k for which WSS starts to diminish

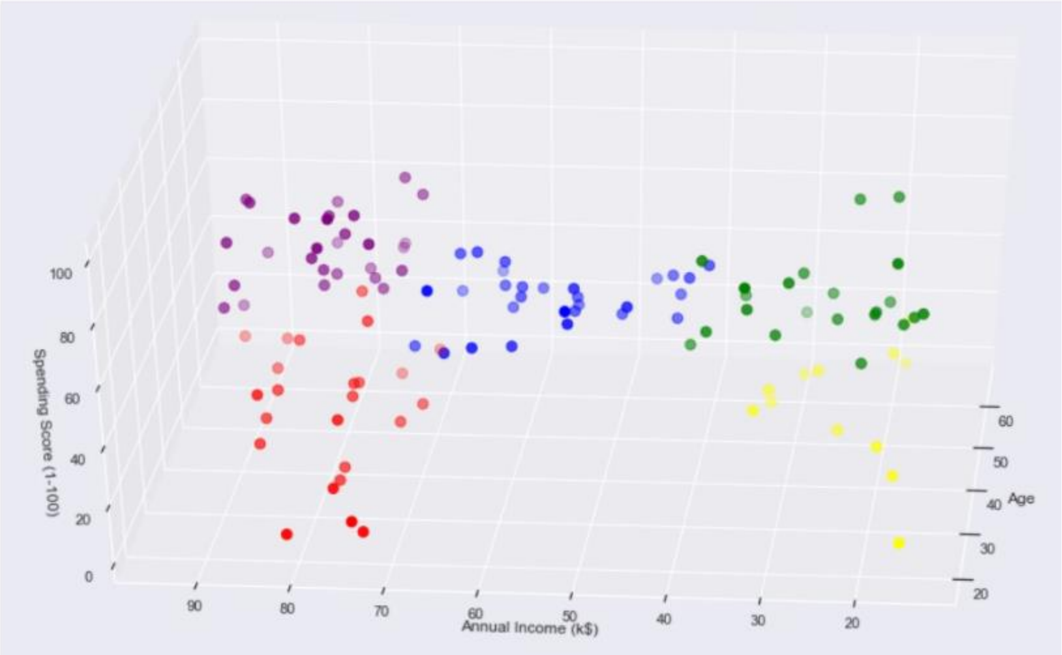
The data:



	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)	label
0	1	Male	19	15	39	5
1	2	Male	21	15	81	3
2	3	Female	20	16	6	4
3	4	Female	23	16	77	3
4	5	Female	31	17	40	5

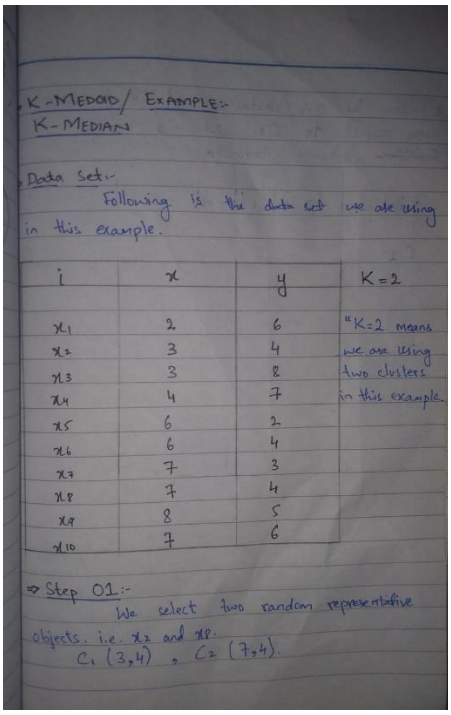
Now we plot it.

The Output:



So, we used K-Median clustering to understand customer data. K-Median is a good clustering algorithm. Almost all the clusters have similar density. It is also fast and efficient in terms of computational cost.

STEP BY STEP CALCULATION OF K-MEDIAN ALGORITHM:



K-MEDOID/ EXAMPLE:-
K-MEDIAN

Data Set:-

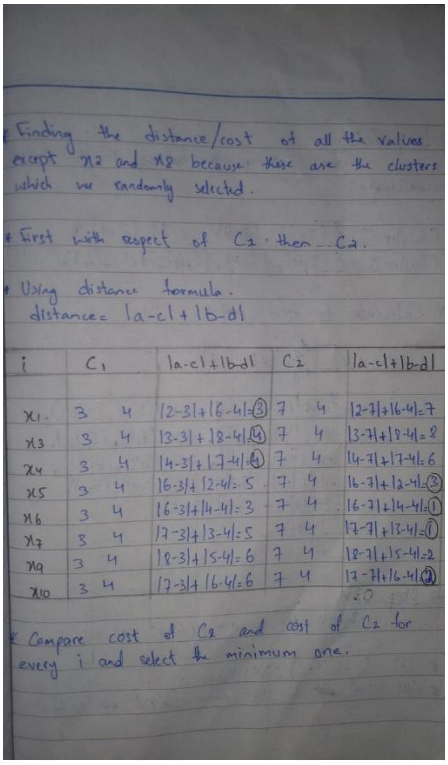
Following is the data set we are using in this example.

i	x	y	K=2
x_1	2	6	"K=2 means
x_2	3	4	we are using
x_3	3	8	two clusters
x_4	4	7	in this example.
x_5	6	2	
x_6	6	4	
x_7	7	3	
x_8	7	4	
x_9	8	5	
x_{10}	7	6	

⇒ Step 01:-

We select two random representative objects. i.e. x_2 and x_8.

C_1 (3,4) , C_2 (7,4).



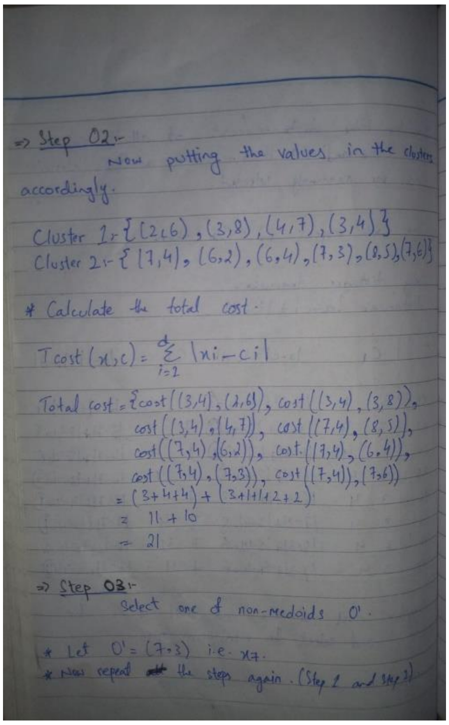
* Finding the distance/cost of all the values except x_2 and x_8 because these are the clusters which we randomly selected.

* First with respect of C_1 · then C_2.

* Using distance formula.
distance= $|a-c| + |b-d|$

i	C_1		$\lvert a-c\rvert+\lvert b-d\rvert$	C_2		$\lvert a-c\rvert+\lvert b-d\rvert$
x_1	3	4	$\lvert 2-3\rvert+\lvert 6-4\rvert=$ ③	7	4	$\lvert 2-7\rvert+\lvert 6-4\rvert=7$
x_3	3	4	$\lvert 3-3\rvert+\lvert 8-4\rvert=$ ④	7	4	$\lvert 3-7\rvert+\lvert 8-4\rvert=8$
x_4	3	4	$\lvert 4-3\rvert+\lvert 7-4\rvert=$ ④	7	4	$\lvert 4-7\rvert+\lvert 7-4\rvert=6$
x_5	3	4	$\lvert 6-3\rvert+\lvert 2-4\rvert=5$	7	4	$\lvert 6-7\rvert+\lvert 2-4\rvert=$ ③
x_6	3	4	$\lvert 6-3\rvert+\lvert 4-4\rvert=3$	7	4	$\lvert 6-7\rvert+\lvert 4-4\rvert=$ ①
x_7	3	4	$\lvert 7-3\rvert+\lvert 3-4\rvert=5$	7	4	$\lvert 7-7\rvert+\lvert 3-4\rvert=$ ①
x_9	3	4	$\lvert 8-3\rvert+\lvert 5-4\rvert=6$	7	4	$\lvert 8-7\rvert+\lvert 5-4\rvert=2$
x_{10}	3	4	$\lvert 7-3\rvert+\lvert 6-4\rvert=6$	7	4	$\lvert 7-7\rvert+\lvert 6-4\rvert=$ ②

* Compare cost of C_1 and cost of C_2 for every i and select the minimum one.



=> **Step 02:-**

Now putting the values in the clusters accordingly.

Cluster 1:- {(2,6), (3,8), (4,7), (3,4)}
Cluster 2:- {(7,4), (6,2), (6,4), (7,3), (8,5), (7,6)}

* Calculate the total cost.

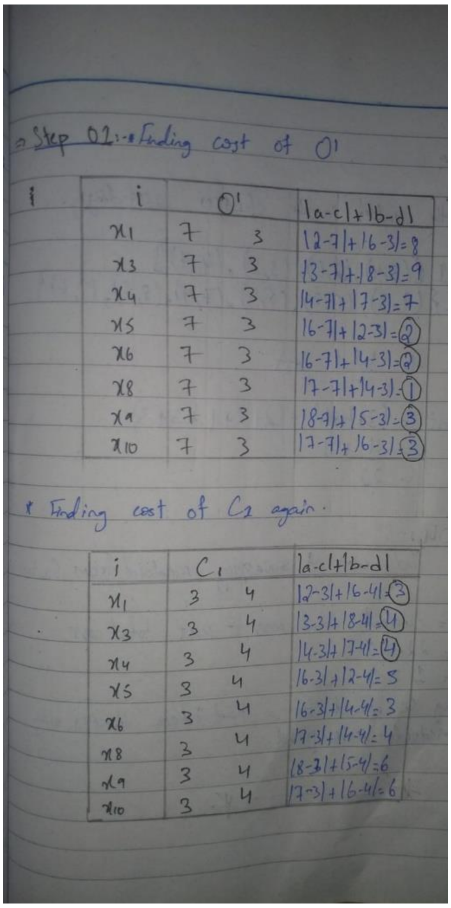
$$T\text{cost}(x,c) = \sum_{i=1}^{d} |x_i - c_i|$$

Total cost = {cost((3,4), (2,6)), cost((3,4), (3,8)),
cost((3,4), (4,7)), cost((7,4), (8,5)),
cost((7,4), (6,2)), cost((7,4), (6,4)),
cost((7,4), (7,3)), cost((7,4)), (7,6))
= (3+4+4) + (3+1+1+2+2)
= 11 + 10
= 21

=> **Step 03:-**

Select one of non-medoids O'.

* Let O' = (7,3) i.e. x_7.
* Now repeat ~~all~~ the steps again. (Step 1 and Step 2)



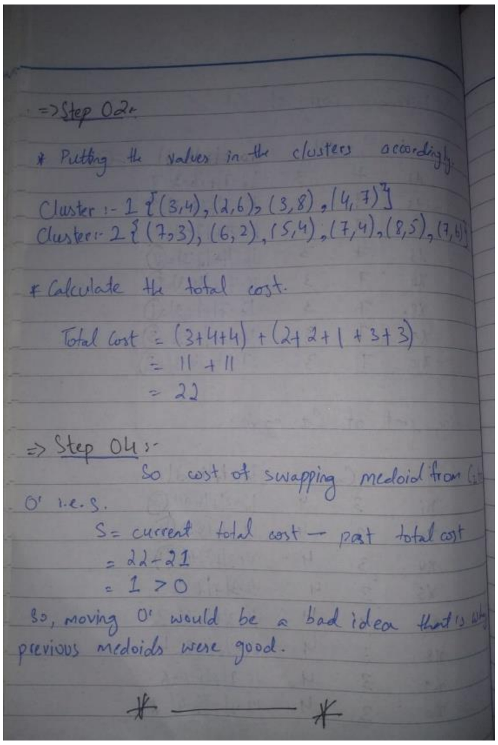
=> **Step 02:-** * Finding cost of O'

i

i	O'		$\lvert a-c\rvert+\lvert b-d\rvert$
x_1	7	3	$\lvert 2-7\rvert+\lvert 6-3\rvert=8$
x_3	7	3	$\lvert 3-7\rvert+\lvert 8-3\rvert=9$
x_4	7	3	$\lvert 4-7\rvert+\lvert 7-3\rvert=7$
x_5	7	3	$\lvert 6-7\rvert+\lvert 2-3\rvert=\boxed{2}$
x_6	7	3	$\lvert 6-7\rvert+\lvert 4-3\rvert=\boxed{2}$
x_8	7	3	$\lvert 7-7\rvert+\lvert 4-3\rvert=\boxed{1}$
x_9	7	3	$\lvert 8-7\rvert+\lvert 5-3\rvert=\boxed{3}$
x_{10}	7	3	$\lvert 7-7\rvert+\lvert 6-3\rvert=\boxed{3}$

* Finding cost of C_1 again.

i	C_1		$\lvert a-c\rvert+\lvert b-d\rvert$
x_1	3	4	$\lvert 2-3\rvert+\lvert 6-4\rvert=\boxed{3}$
x_3	3	4	$\lvert 3-3\rvert+\lvert 8-4\rvert=\boxed{4}$
x_4	3	4	$\lvert 4-3\rvert+\lvert 7-4\rvert=\boxed{4}$
x_5	3	4	$\lvert 6-3\rvert+\lvert 2-4\rvert=5$
x_6	3	4	$\lvert 6-3\rvert+\lvert 4-4\rvert=3$
x_8	3	4	$\lvert 7-3\rvert+\lvert 4-4\rvert=4$
x_9	3	4	$\lvert 8-3\rvert+\lvert 5-4\rvert=6$
x_{10}	3	4	$\lvert 7-3\rvert+\lvert 6-4\rvert=6$



=> Step 02:-

* Putting the values in the clusters accordingly.

Cluster :- 1 {(3,4), (2,6), (3,8), (4,7)}
Cluster :- 2 {(7,3), (6,2), (5,4), (7,4), (8,5), (7,6)}

* Calculate the total cost.

Total Cost = (3+4+4) + (2+2+1+3+3)
= 11 + 11
= 22

=> Step 04:-

So cost of swapping medoid from C_2 to O' i.e. S.

S = current total cost − past total cost
= 22 − 21
= 1 > 0

So, moving O' would be a bad idea that's why previous medoids were good.

* ——— *