Experiment. No. 11

Title : Implement k-means clustering / hierachical clustering on sales_data_sample.csv dataset. Determine the number of clusters using the elbow method

objective : To understand how to use unsupervised learni -ng to segment different - different clusters or groups and used to them to train your model to predict future things.

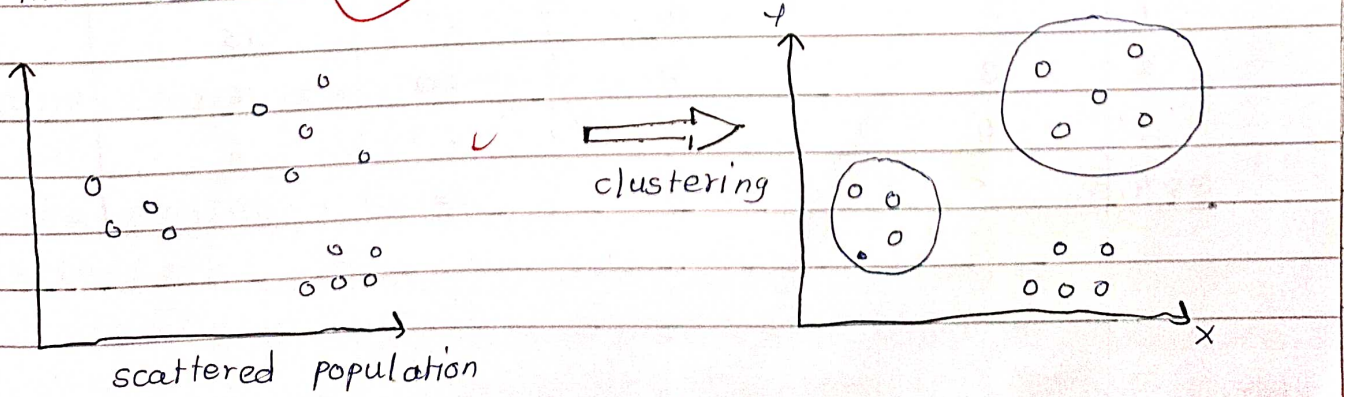
Dataset Description : The data includes the following features :

1. customer ID , 2. customer Gender , 3. Customer Age ,
4. Annual income of the customer
5. spending score of the customer.

Theory -

clustering :

clustering algorithms try to find natural clusters in data, the various aspects of how the algorithms to cluster data can be tuned and modified. clustering is based on the principle that terms within the same cluster must be similar to each other. The data is grouped in such a way that related elements are close to each other.



Use of clustering:

1. Marketing
2. Real Estate
3. Bookstore & Library management
4. Document Analysis.

K-means clustering:

K-means clustering is an unsupervised machine learning algorithm that divides the given data into given number of clusters. Here the "K" is the given number of predefined clusters, that need to be creat

It is a centroid based algorithm in which each cluster is associated with a centroid. The main idea is to reduce the distance between the data points and their respective cluster centroid

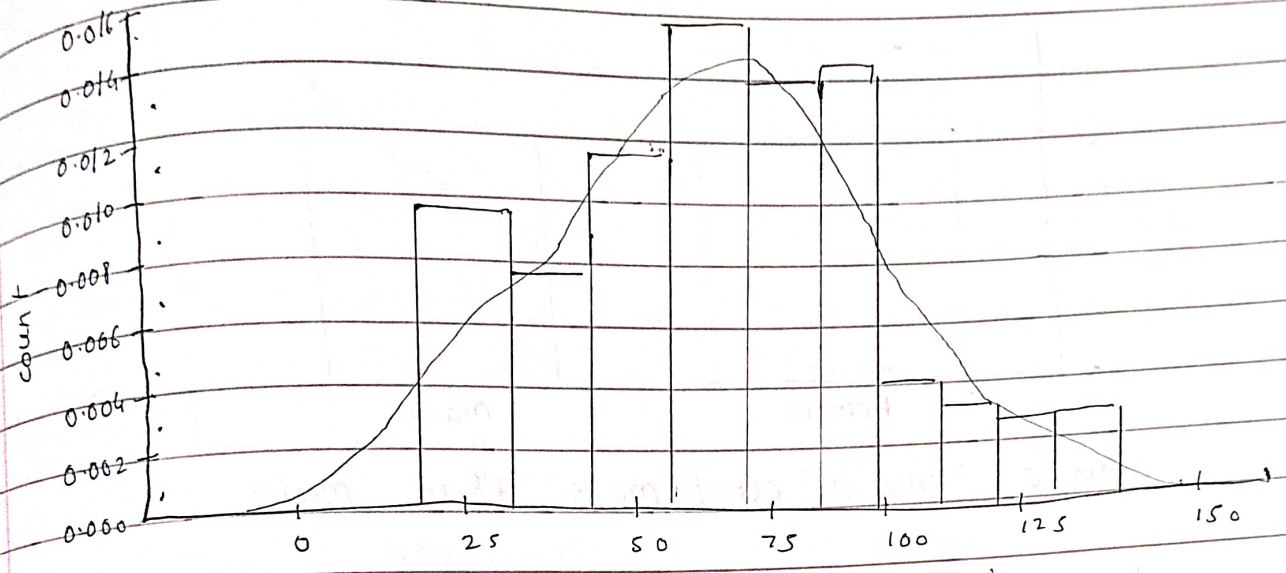
The algorithm takes raw unlabelled data as a input & divides the dataset into clusters and the process is repeated until the best clusters are found.

The data has 200 entries, that is data from 200 customers

so look at the data

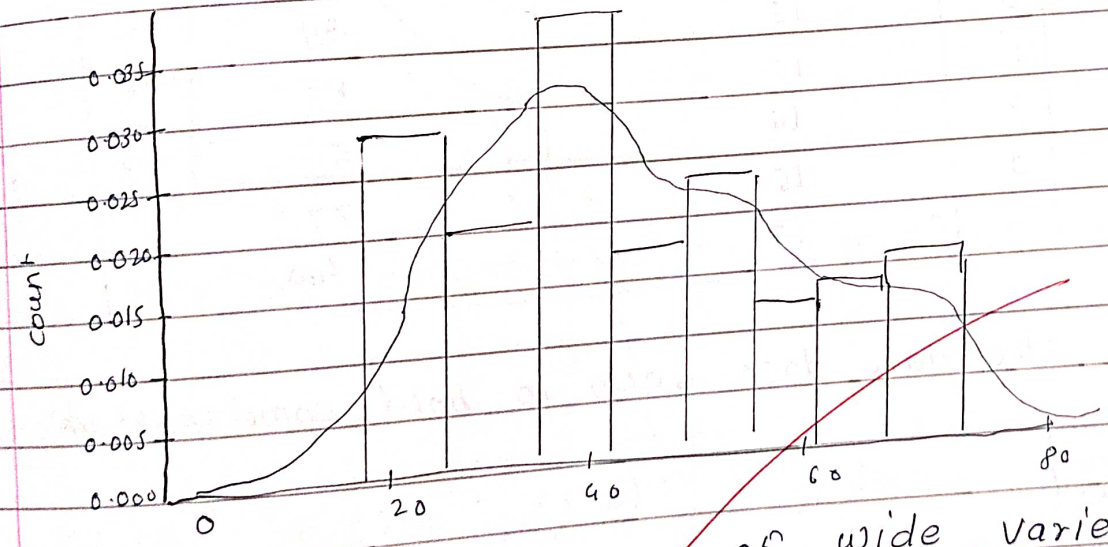
	CustomerID	Gender	Age	AI($)	SS($-100)
0	1	m	19	15	39
1	2	m	21	15	81
2	3	F	20	16	6
3	4	F	23	16	77
4	5	F	31	17	40

Annual Income Distributor :



most of the annual income falls between 50k to 80k

Age Distribution :

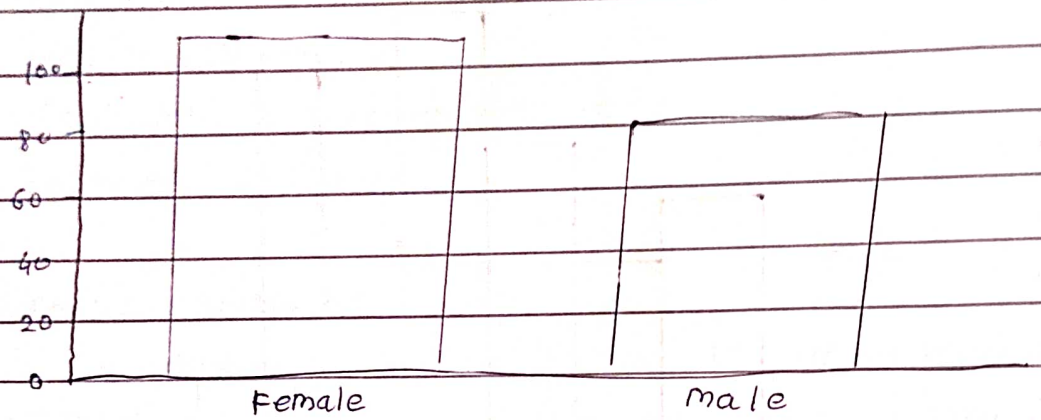


There are customer of wide variety of ages

spending score distribution :

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

Gender Analysis



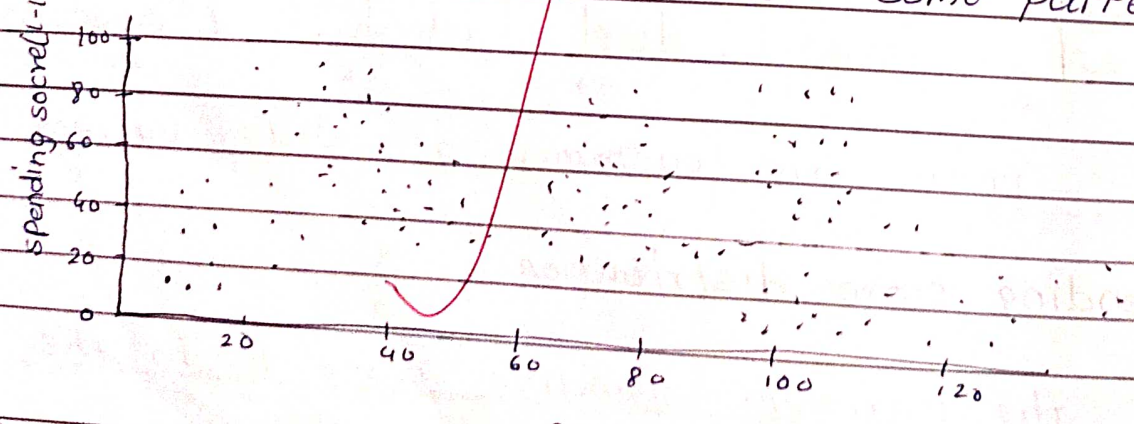
more Female customers than male

clustering based on 2 features :

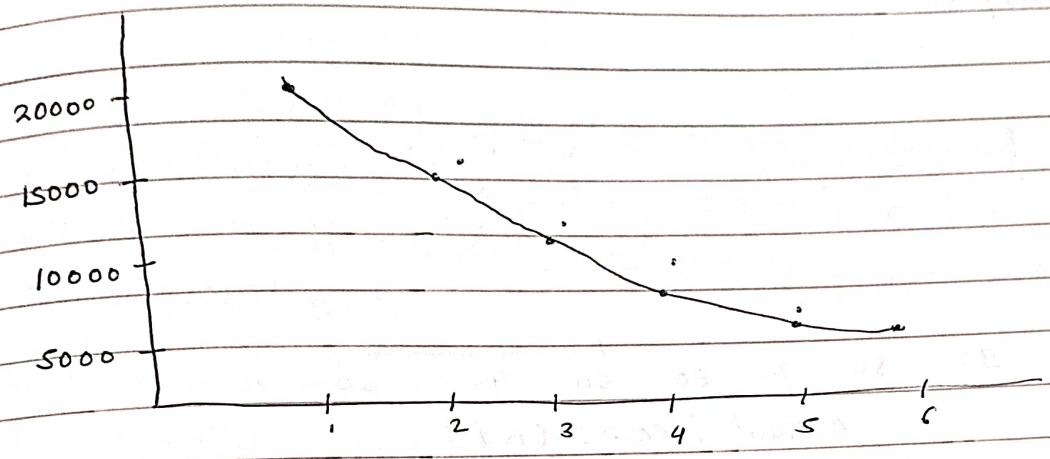
First, we work with two features only annual income and spending score.

	Annual income	spending score
0	15	39
1	15	81
2	16	6
3	16	77
4	17	40

The data does seem to hold some patterns

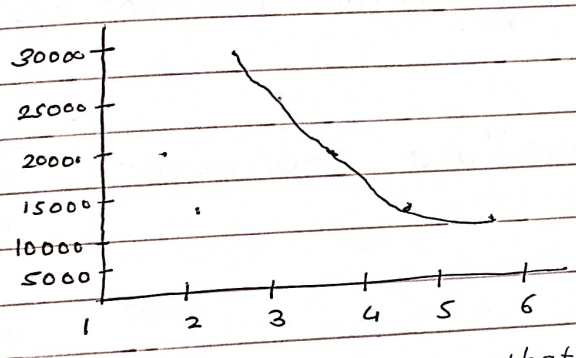


The plot :



This is known as the elbow graph

K-means clustering on the basis of 3D Data :

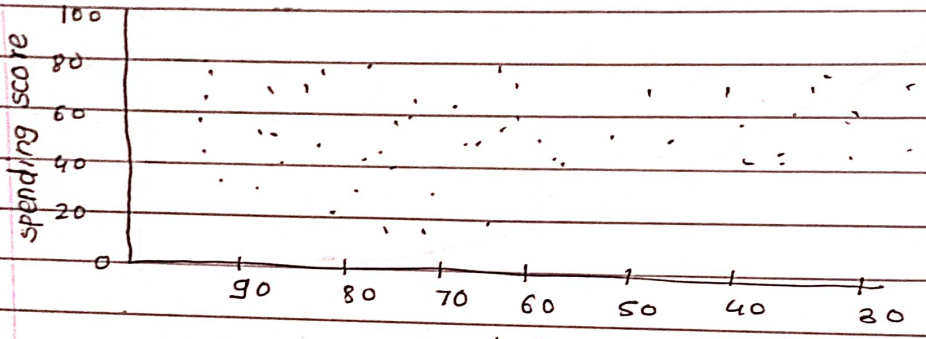


Here can assume that K = 5 , will be a good value

The data :

CU-ID	G	Age	AI($)	SS (1-100)	label
1	m	19	15	39	5
2	m	21	15	81	3
3	F	20	16	6	4
4	F	23	16	77	3
5	F	31	17	40	5

The output:



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

Conclusion:

In this way, that we implemented Kmeans clustering using dataset.