# Assignment 5

## Aim:-

Implement K-means clustering/hierarchical clustering on sales-data-sample.csv dataset, determine the number of clusters using the elbow method.

## Requirements:-

Python, Jupyter notebook, python installations, python libraries- pandas, sklearn, matplotlib.

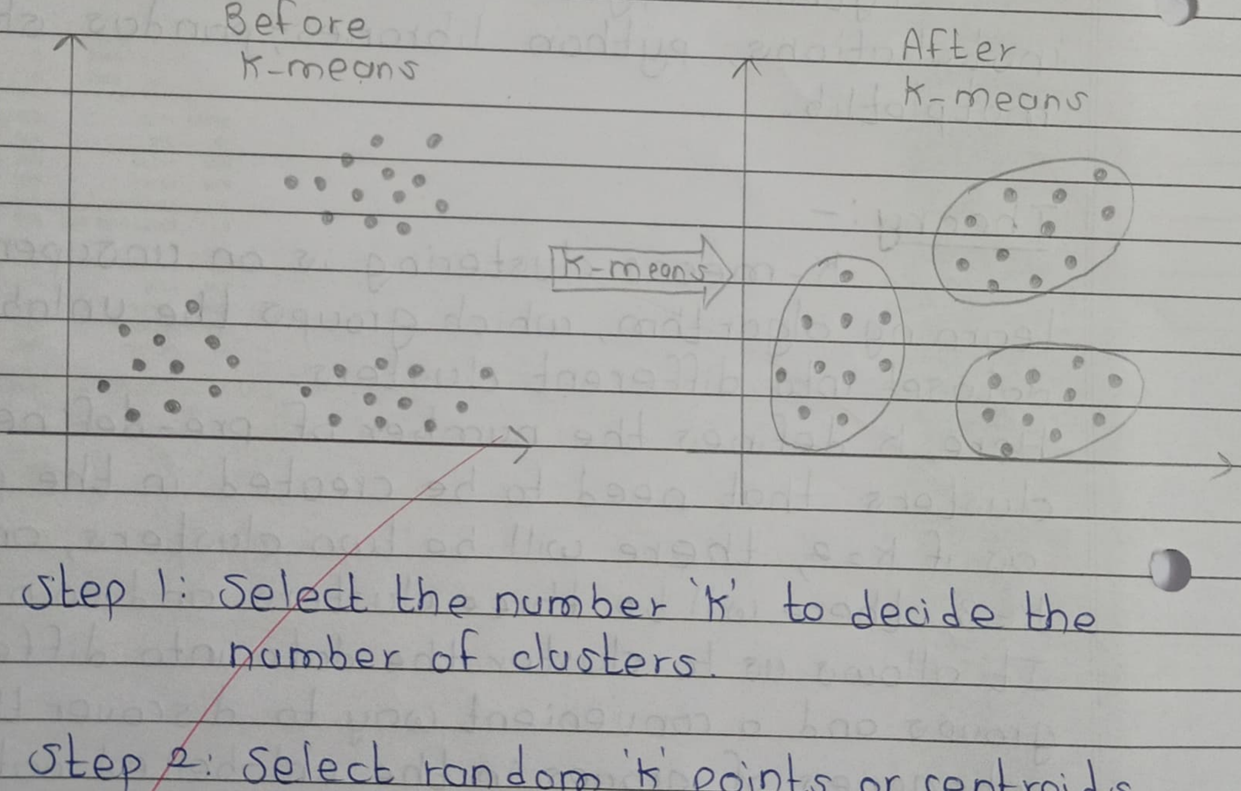
## Theory:-

K-means clustering is an unsupervised learning algorithm, which groups the unlabelled dataset into different clusters.
Here, K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters and so on.
It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabelled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The K-means clustering algorithm mainly perform two tasks :-

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest K-center. Those data points which are near to the particular K-center, create a cluster.



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

Step 2: Select random 'K' points or centroids.

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

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

Step 5: Repeat the third step.

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

Step 7: The model is ready.

Data Pre-processing:

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model.

Steps:-

1) Getting the dataset.
2) Importing libraries.
3) finding missing data.
4) Encoding categorical data.
5) Splitting dataset into training and testing

Data Transformation:

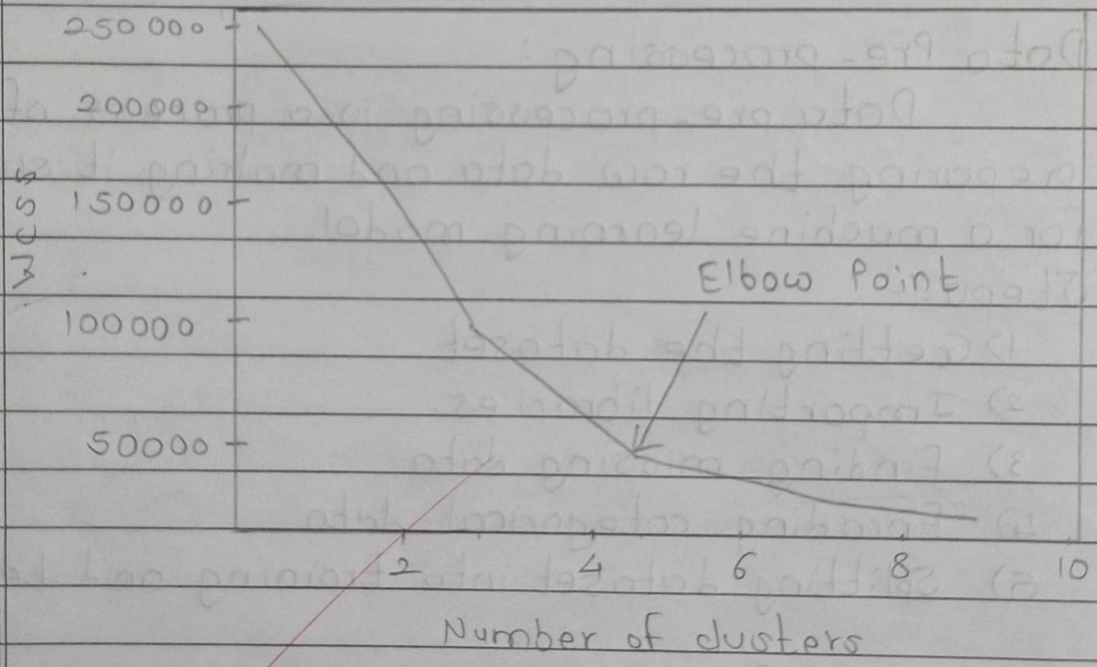
Data transformation is the process of converting raw data into a format that would be more suitable for model building and also data discovery in general.

Data Reduction:

The number of input features, variables or columns present in a given dataset is known as dimensionality and process of to reduce these features is called dimensionality reduction.

## Elbow Method:

In Elbow method, we are actually varying the no. of clusters (K) for each value of K, we are calculating WCSS.
WCSS is the sum of squared distance between each point and the centroid in a cluster.



**Conclusion:-**

Hence, we have successfully implemented K-means clustering algorithm.