CLUSTERING

→ Clustering refers to a very broad set of techniques for finding subgroups or cluster in a dataset.

→ When we cluster the observations of a data set, we seek to partition them into distinct groups so that the observations within each group are quite similar to each other, while observations in different groups are quite different from each other. Of course to make this concrete we must define what it means for two observations or more to be similar or different.

→ Application of clustering arise in marketing. We may have the access to a large number of measurement (eg median household income, occupation, distance from nearest urban area etc) for a large number of people. Our goal is to perform market segmentation by identifying subgroups of people who might be more receptive to a particular form of advertising or more likely to buy a product. The task of performing market segmentation amount to clustering the people in the dataset.

→ Clustering looks to find homogenous group / subgroups among the observations.

→ We will study 3 types of Clustering – i) K Means clustering
ii) Hierarchical clustering
iii) DB Scan clustering.

Practical Issue in Clustering →

1. Small decisions with Big Consequences →
   - Should the observations / features first be standardized in some way?
   - In case of hierarchical clustering
     1. What dissimilarity measure should we use?
     2. What type of linkage should we use?
     3. Where should we cut the dendogram in order to obtain clusters.
   - In case of K means clustering, how many clusters should we look in data.
2. Validating the cluster obtained →
   We need to know whether the cluster that we have found represent true subgroups in the data or whether they are result cluster of noise.
3. Robustness of the cluster →
   - Sense of robustness of the cluster should be obtained. Most importantly we should be careful how results of the cluster analysis are reported.
4. Other issue in Clustering →
   - Clustering force every observation into a cluster, the cluster may be heavily distorted due to presence of outliers that do not belong to any cluster.