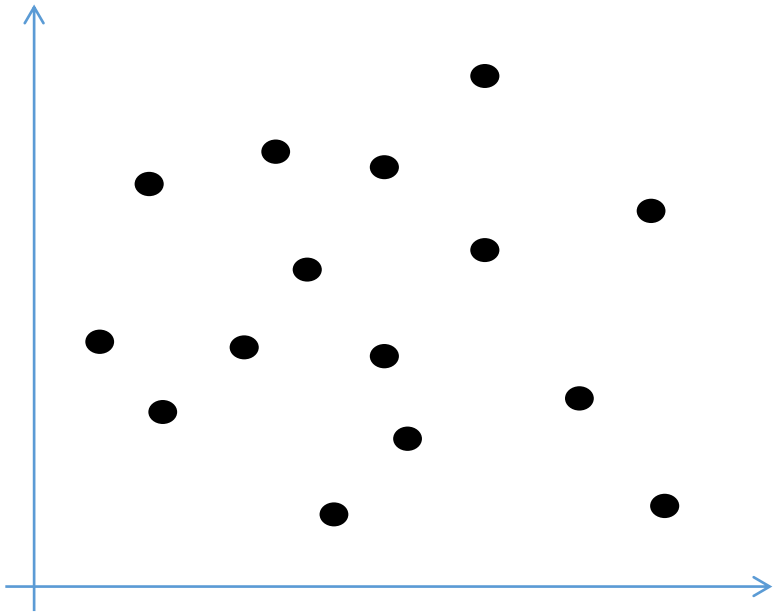
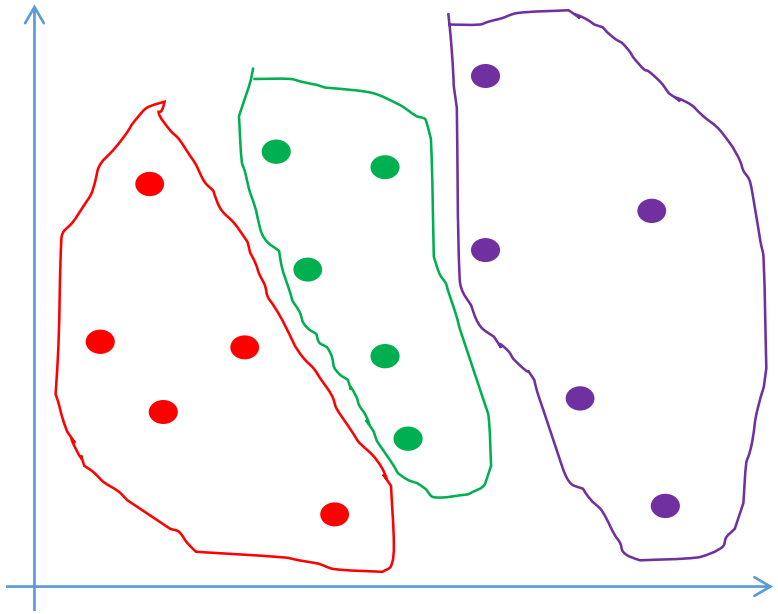
Hierarchical Clustering

Hierarchical Clustering – union between 2 nearest clusters





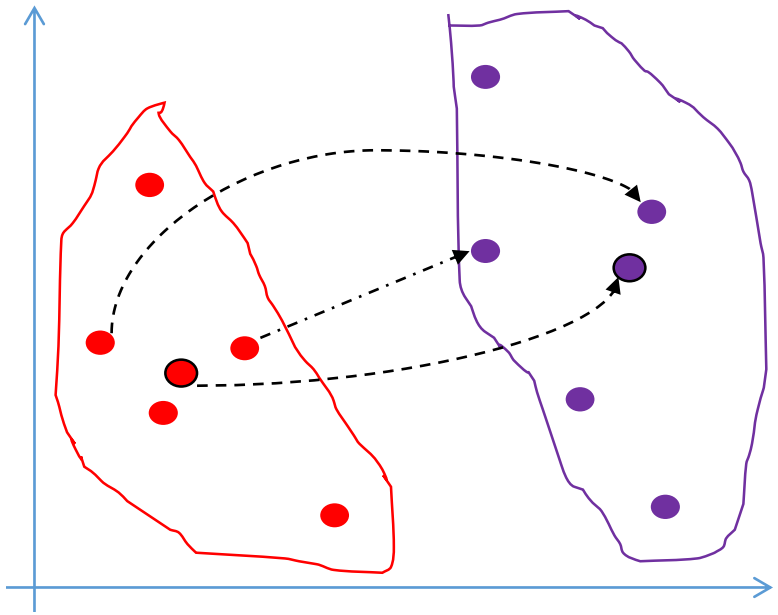
Agglomerative HC

1. Consider all data points as individual clusters : n-clusters
2. Take the 2 closest data points and make them one cluster
3. Take the 2 closest clusters and make them one cluster
4. Repeat 3 till there is one cluster

Distance between data points is calculated using the **Distance Function** formula

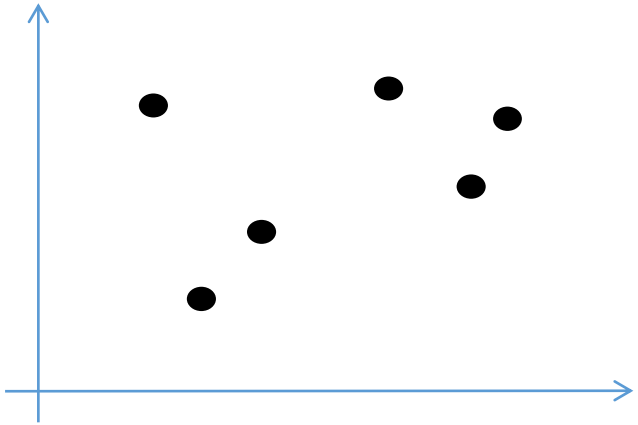
- Euclidean
- Manhattan

Distance between 2 Clusters

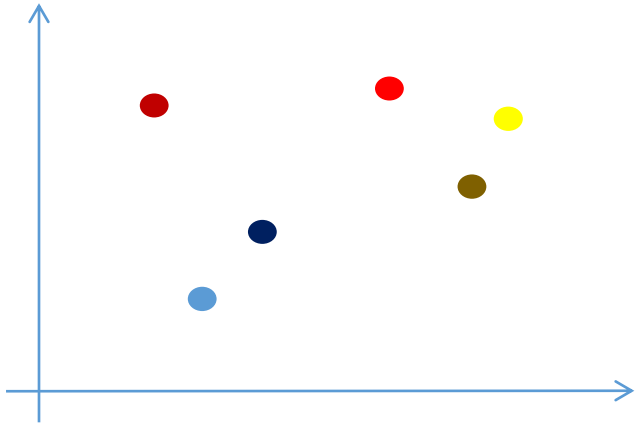


- Nearest points
- Furthest points
- Average Distance
- Centroid Distance

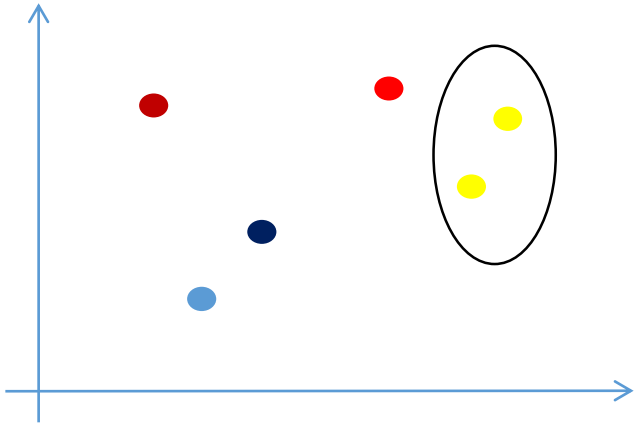
Cluster building process



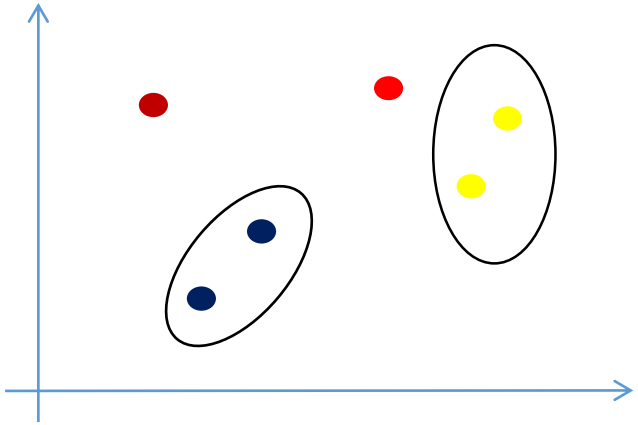
Data points = 6



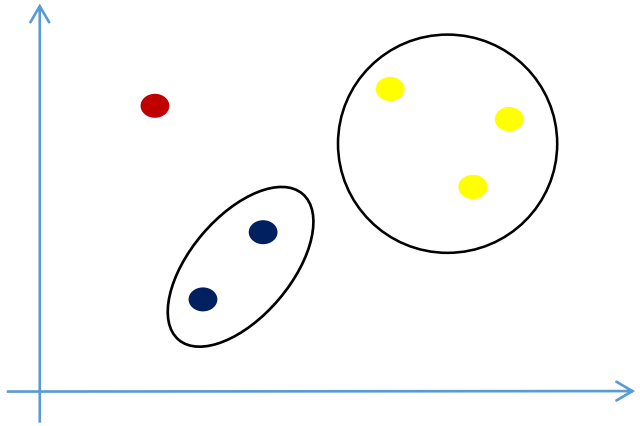
Clusters = 6



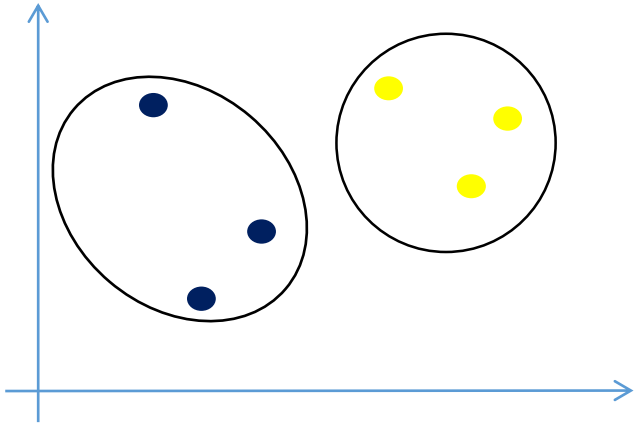
Clusters = 5



Clusters = 4

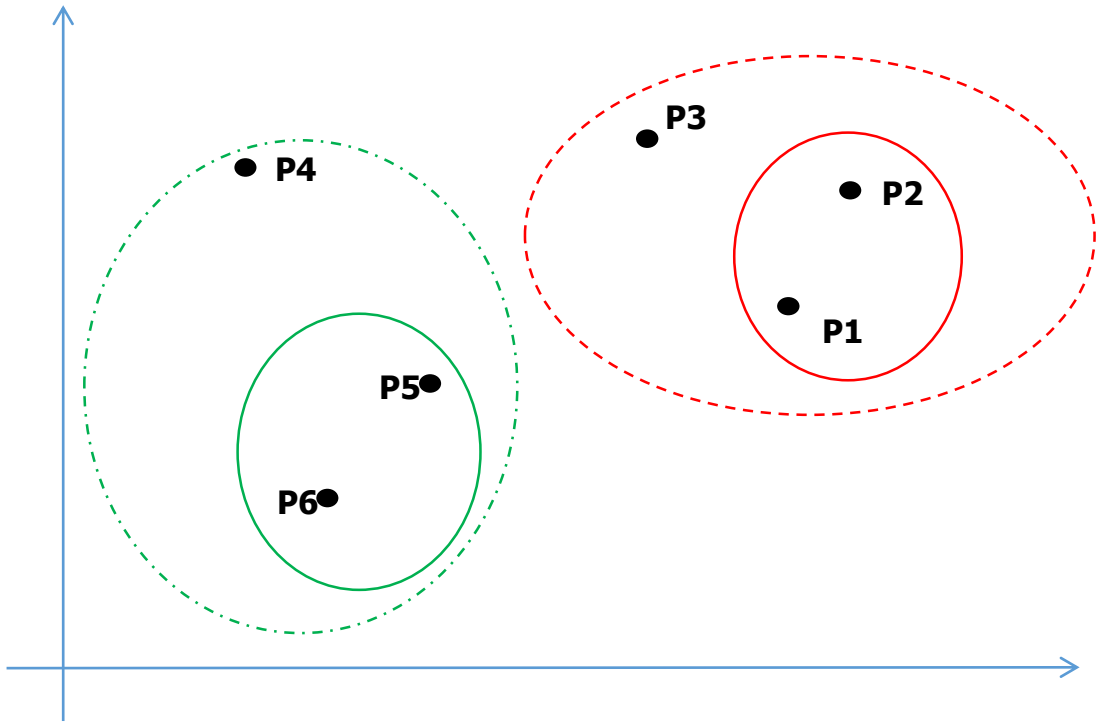


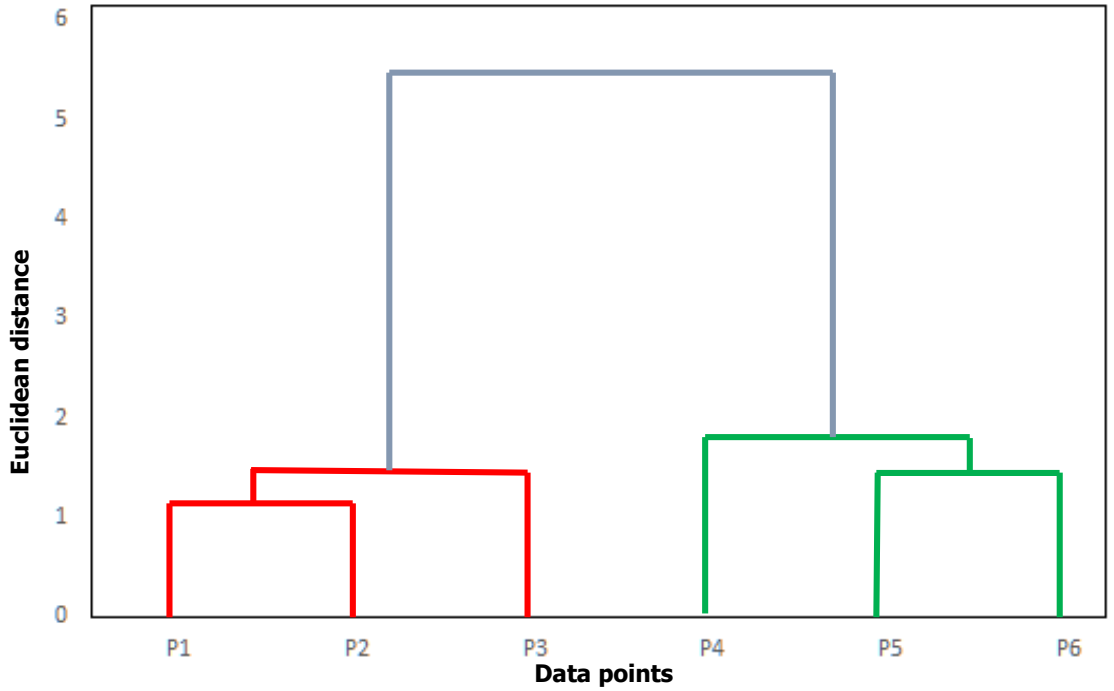
Clusters = 3



Clusters = 2

Dendrograms

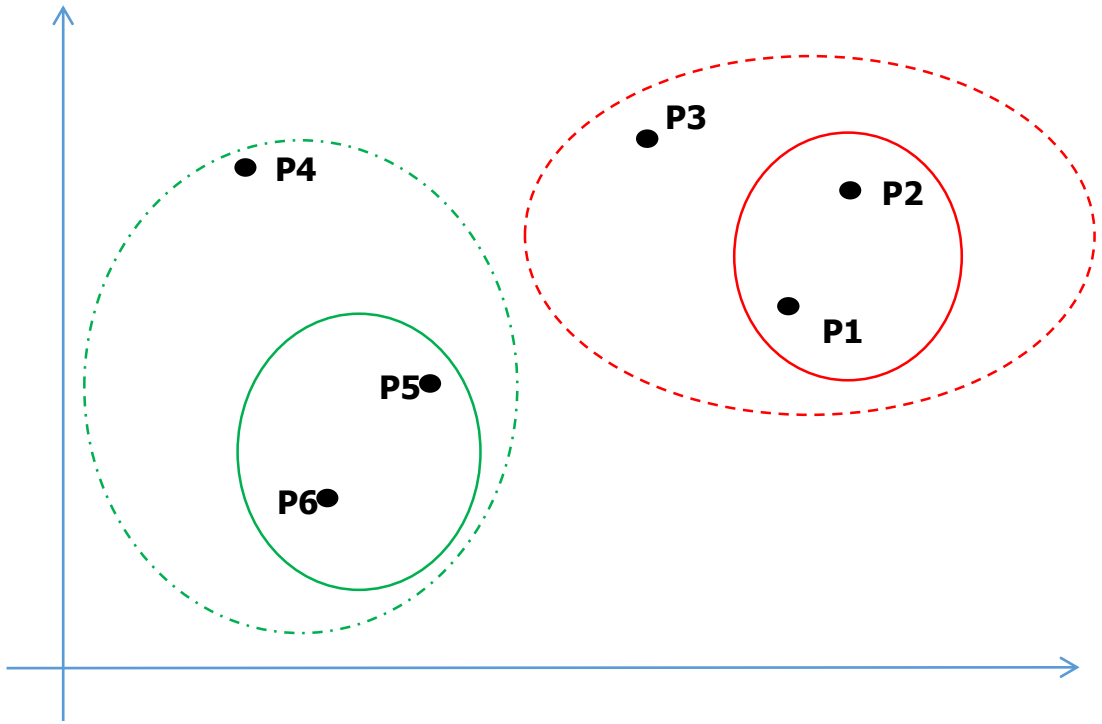


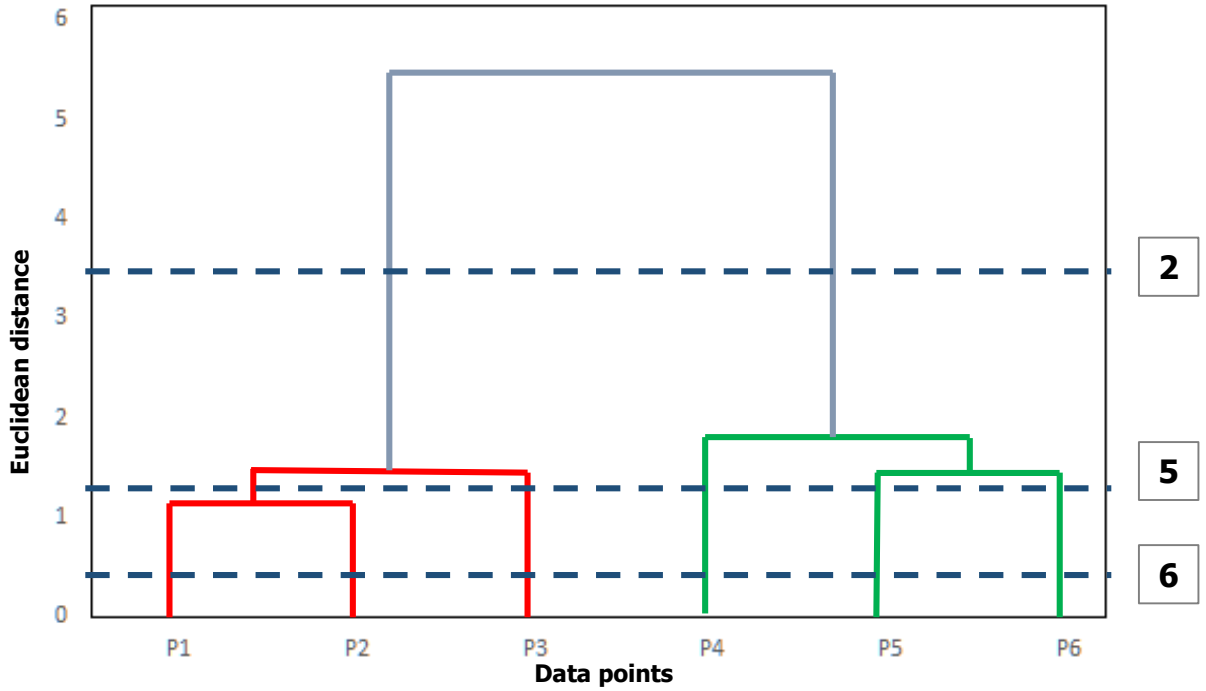


- Dendrograms are like memory of each cluster

- Height indicates dissimilarity
  Greater the height, greater the dissimilarity
- Horizontal line
  Connecting the 2 clusters

Setting threshold to form clusters





- Number of clusters = number of vertical lines crossing the threshold

Selecting the optimum number of Clusters

Approach 1 – Longest Distance

- Find the longest vertical line that does not cross any extended horizontal lines
- Mark Threshold
- Find the number of points intersecting – these indicate the number of clusters