Hierarchical Clustering 101

Bert Gollnick

Hierarchical Clustering

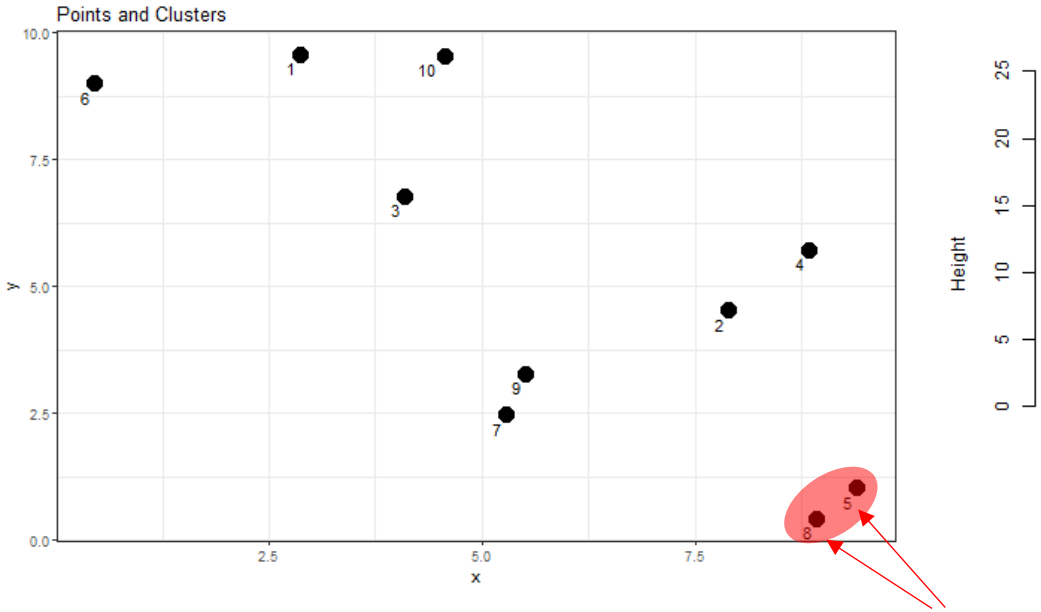
Introduction

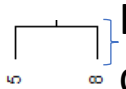
- Algorithm for building **clusters** based on **hierarchy**
- Assesses similarity (distance) of observations
- Result is visualised as dendrogram
- Bottom-up approach
  - each point individual cluster
  - gradually join points
  - Start with most similar points
- Top-down approach
  - All points in one cluster
  - Split clusters until desired cluster number reached
- Depends on two important parameters
  - Distance metric
  - Linkage

Hierarchical Clustering

How the dendrogram is created

- Find closest pair of points





Height represents distance

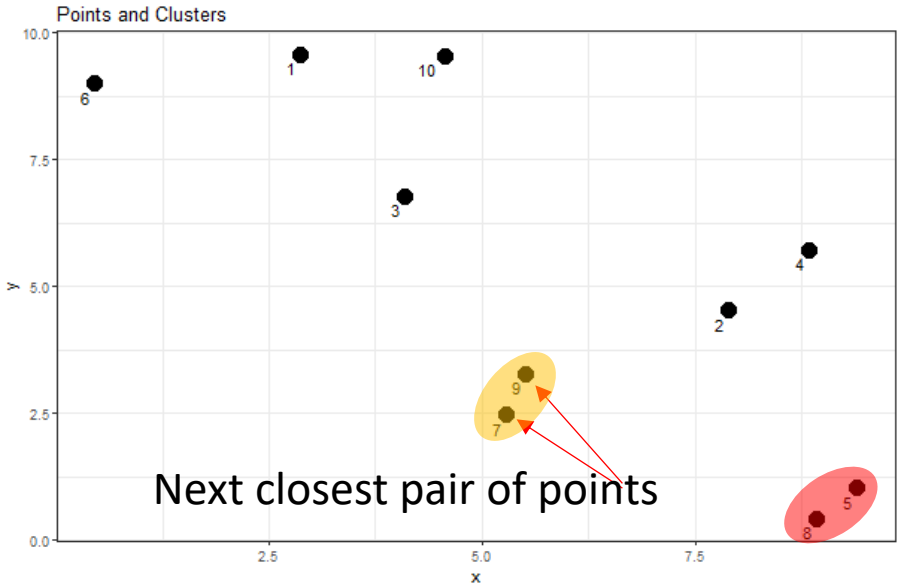
distance_mat
hclust (*, "ward.D")

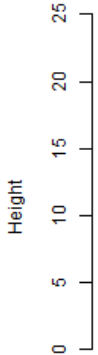
closest pair of points: 5 and 8

Hierarchical Clustering

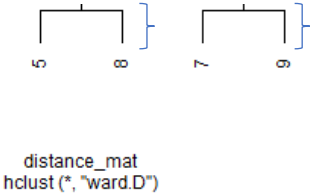
How the dendrogram is created

- Then find next closest pair of points





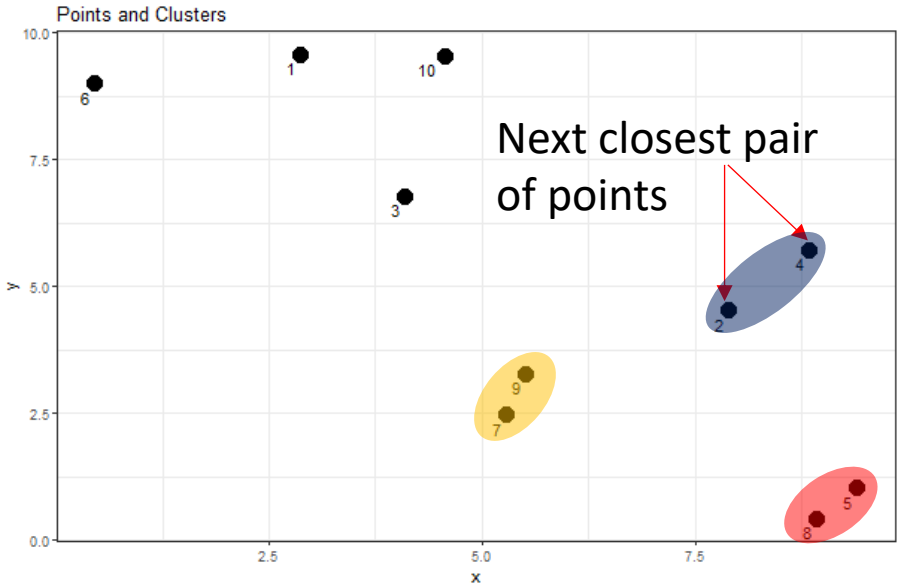
Similar height → Similar distance

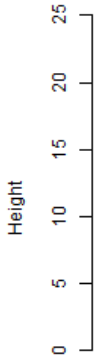


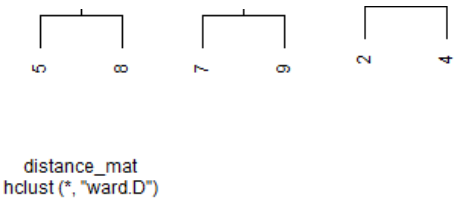
Hierarchical Clustering

How the dendrogram is created

- Then find next closest pair of points



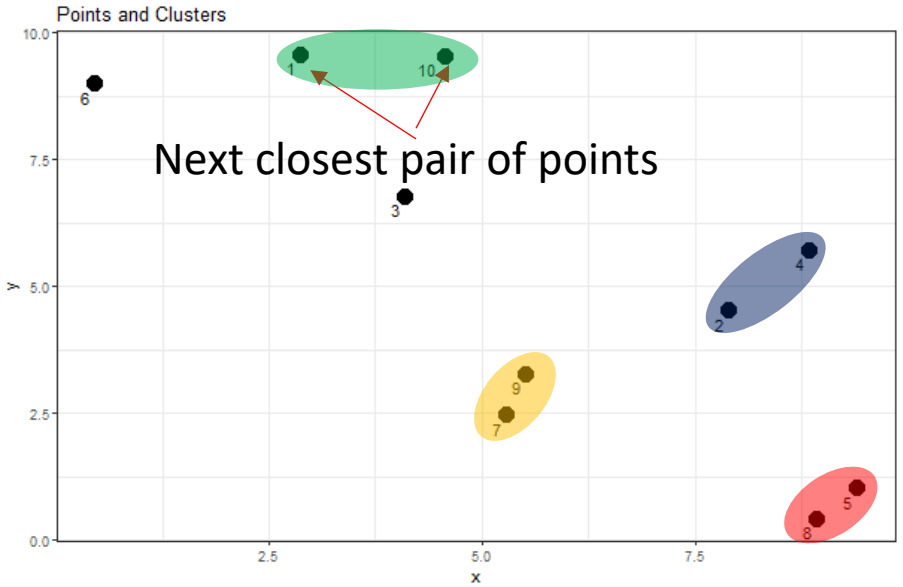


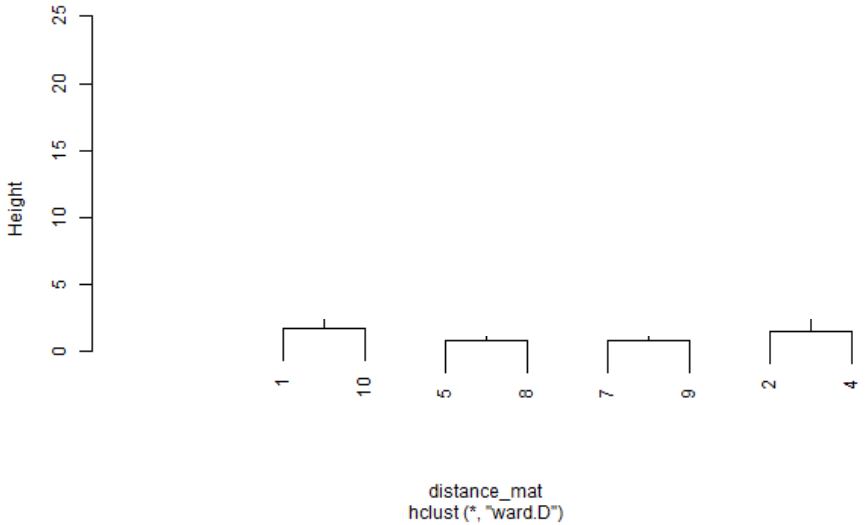


Hierarchical Clustering

How the dendrogram is created

- Then find next closest pair of points

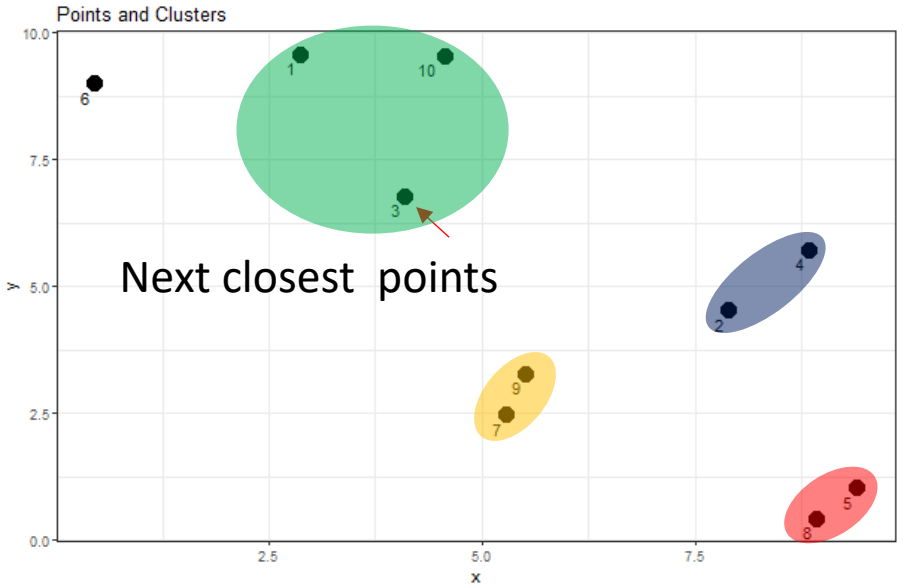


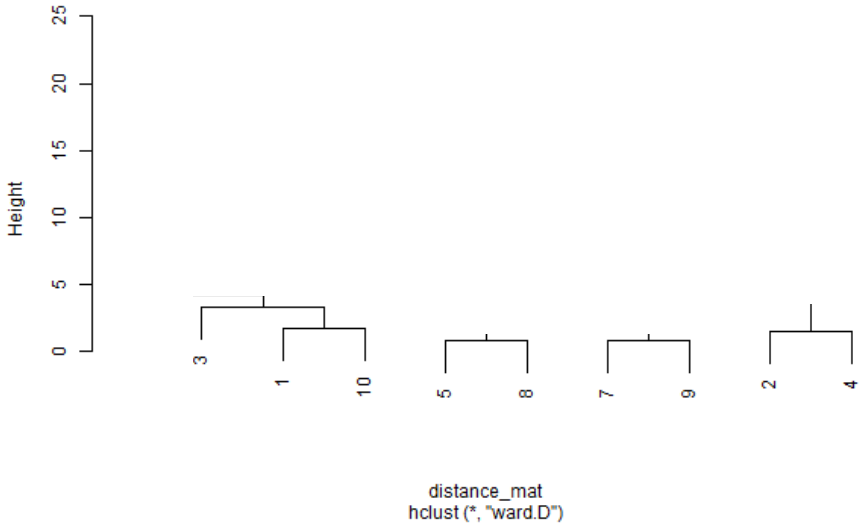


Hierarchical Clustering

How the dendrogram is created

- Then find next closest pair of points

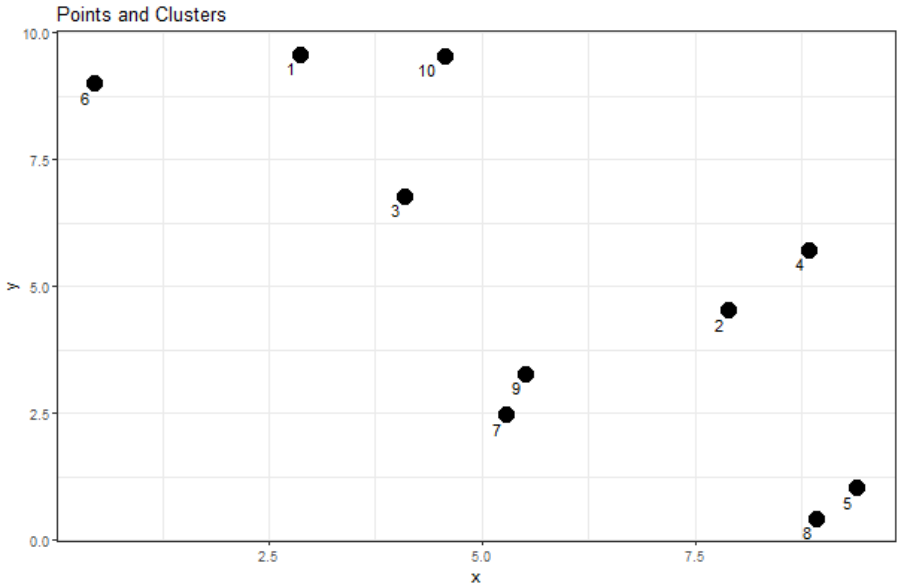


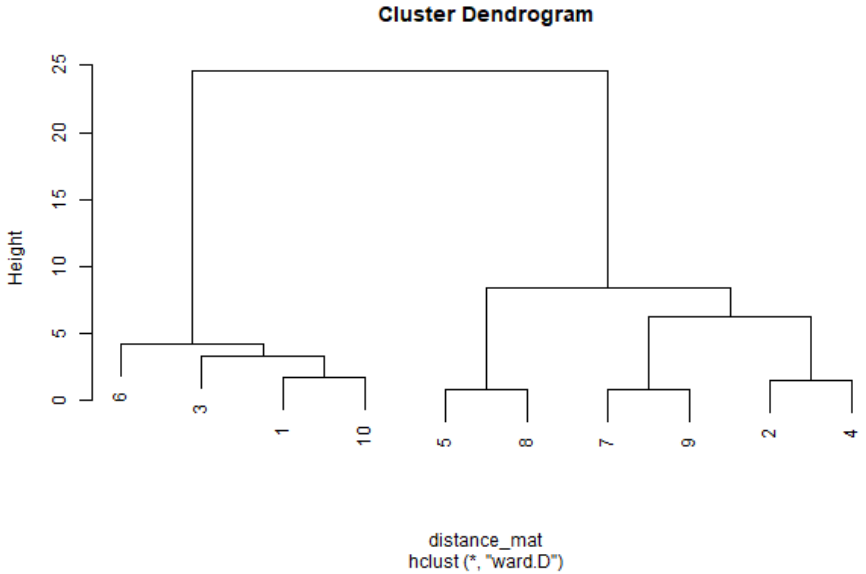


Hierarchical Clustering

How the dendrogram is created

- Finally dendrogram is created

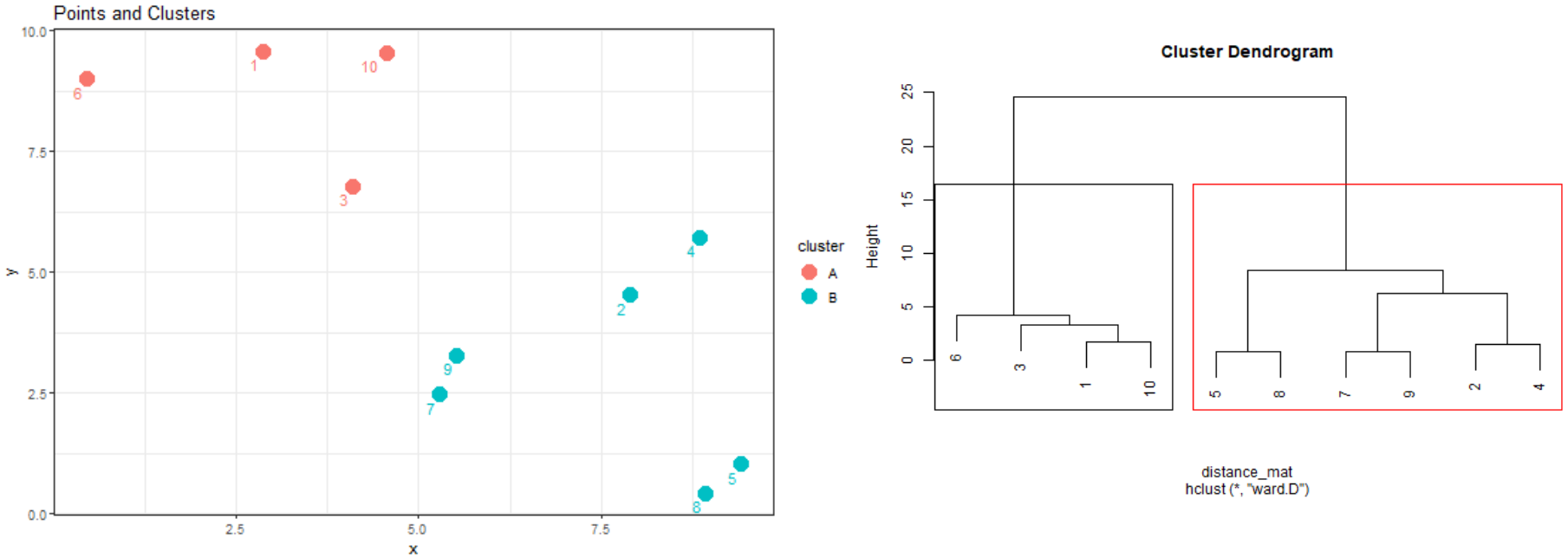




Hierarchical Clustering

Number of Clusters

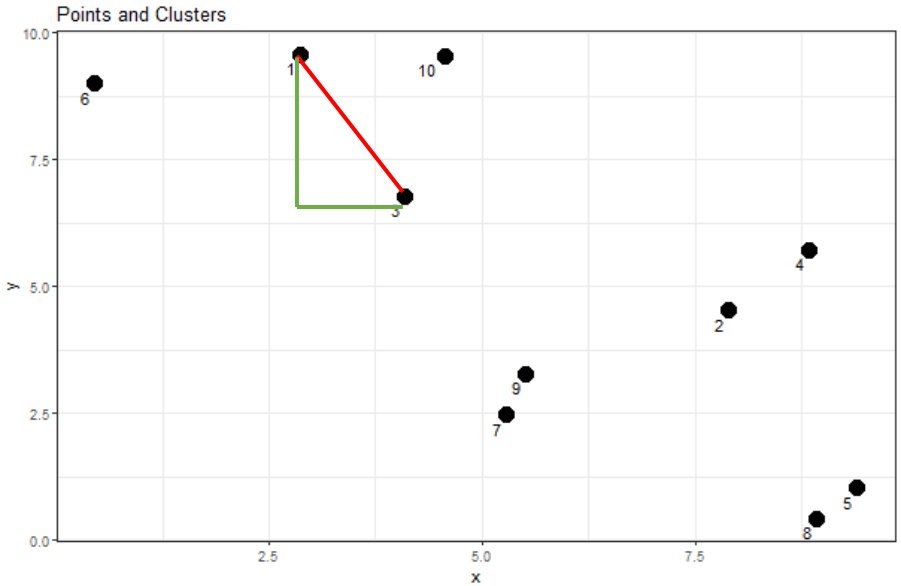
- Search longest branch, cut at the middle



Hierarchical Clustering

Distance Metrics

- Finally dendrogram is created



Most used:

- Euclidean Distance
- Manhattan Distance

Other distance metrics:

- Canberra
- Maximum
- Minkowski
- …

Hierarchical Clustering

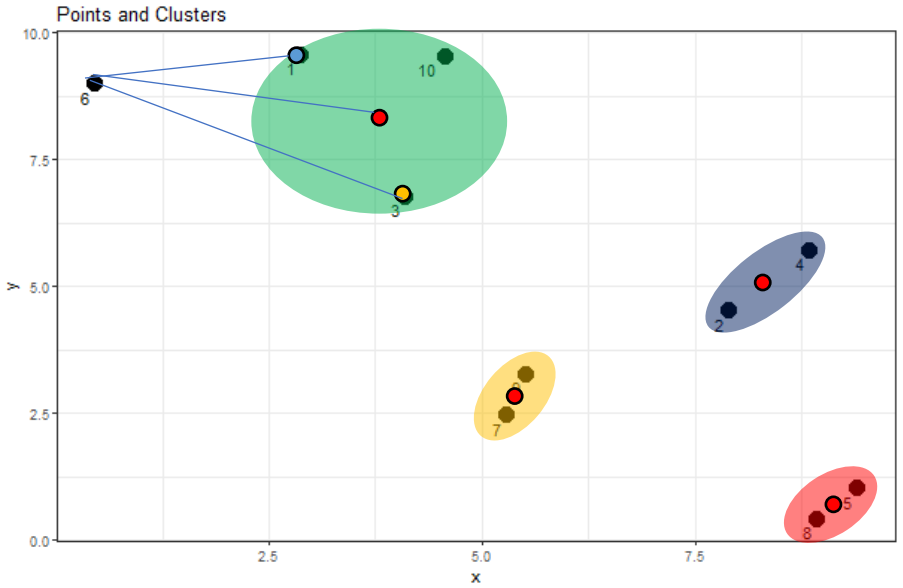
Linkage

- How to assess proximity from point / cluster to an existing cluster?

- Centroid
- Single Linkage
- Complete Linkage

Further methods:

- Ward
- Median
- Mcquitty
- …



Hierarchical Clustering

Difference kmeans and Hierarchical Clustering

Parameter	Kmeans	HC
Complexity	O(n)	$O(n^2)$
Reproducibility	No	yes
Setting of Cluster Number	pre-knowledge required	No pre-knowledge required

Hierarchical Clustering

Advantages / Disadvantages

\+

- Easy to understand
- Provides informative hierarchy
- Simpler decision on number of clusters
- User only needs to define distance met ric AND linkage

\-

- Computational effort
- Sensitive to outliers
- Sensitive to noise
- Not easily applicable for data with nu merical and categorical variables
- Missing data