London Business School

Cluster Analysis II

Data Science MAM 2021-22

Dr Kanishka Bhattacharya

Welcome back to Data Science

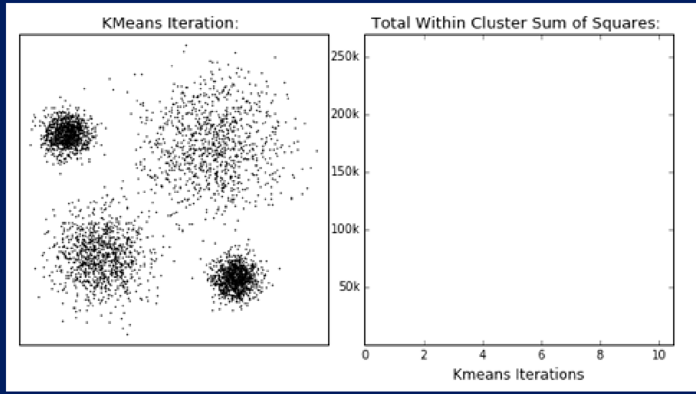
- Thanks for joining in person!

Zoom classroom etiquette

- Please turn on your cameras and mute your microphone.
- I will make warm calls.
- Use chat when instructed, otherwise please raise your virtual hand if you have questions.

Session plan

- Part 1: K-medoids and hierarchical clustering methods
- Part 2: Workshop: Clustering iPlayer users



[redacted]on Business School

Lectures and assignments

Zoom Etiquette

Turn your came

If your camera i ... ase to the facilitator

If you don't have ... g decent for £10 on eBay.

If you have othe ... in your situation

If you don’t turn ... plaining your situation, *I will call on you*

Learning objectives

Last week

What is the difference between supervised and unsupervised learning?

What is the main objective of clustering?

How does K-Means algorithm work?

- Objective
- Inputs
- Outputs

How do we determine how many clusters we have in the data?

- Elbow chart
- PCA visualization
- Comparing clustering results with different clusters
- Silhouette analysis

Learning objectives

This week

k-medoids or partitioning around medoids (PAM) algorithm

- Objective
- Inputs
- Outputs

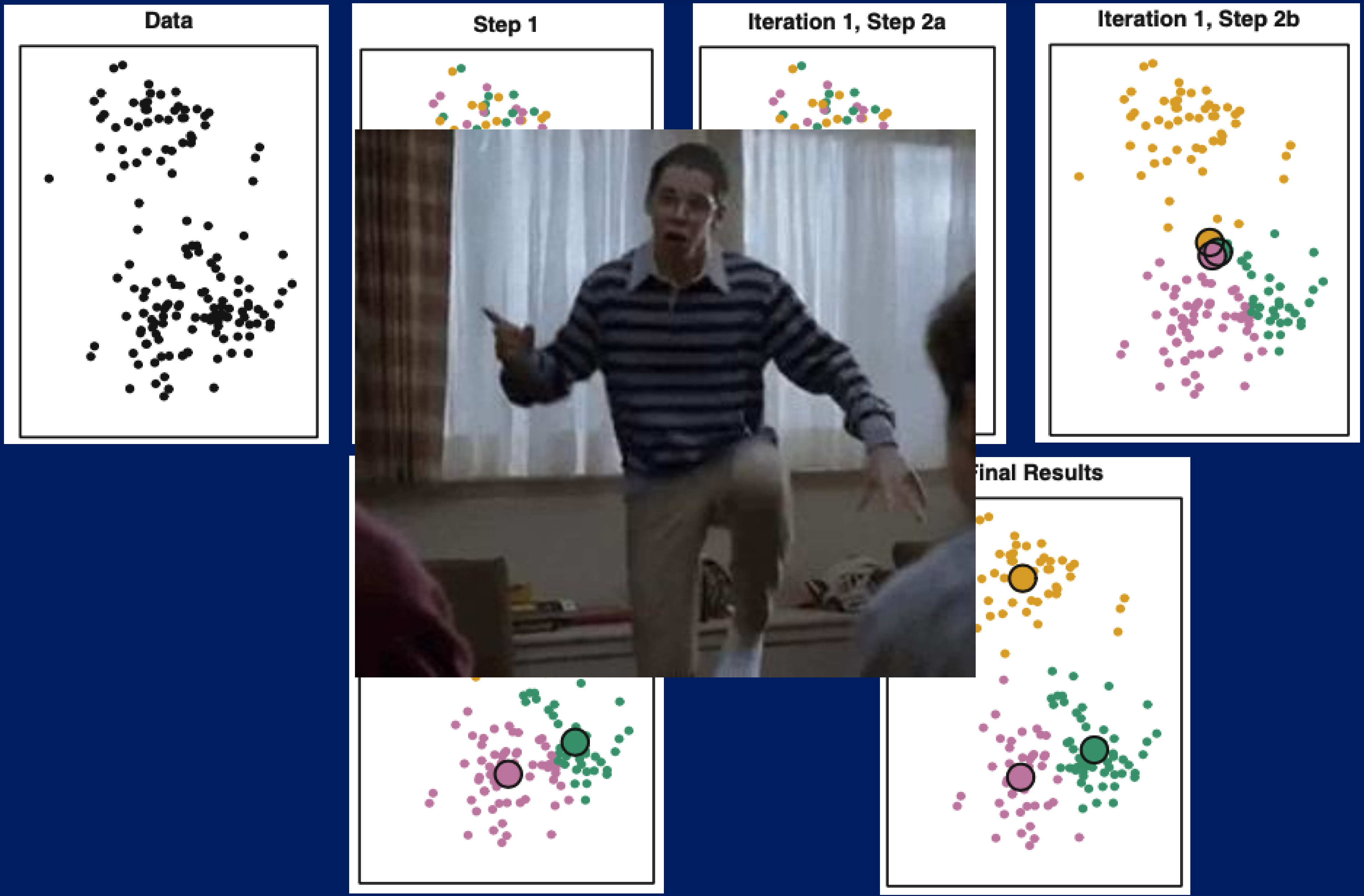
Hierarchical clustering

- Objective
- Inputs
- Outputs

Applying clustering methods in a large data set (workshop)

- Using different clustering methods
- Visualization of the results
- Choosing the best clustering results
- Presenting your findings

K-Means Clustering Algorithm



K-means Algorithm: Issues

Results are sensitive to outliers (why?)

- How to resolve this?
  - Use different distance measures such as absolute error
  - Use the median of observations instead of average.
  - Use K-Medoids (next)

K is fixed in advance

- Use elbow charts and see when there is no improvement

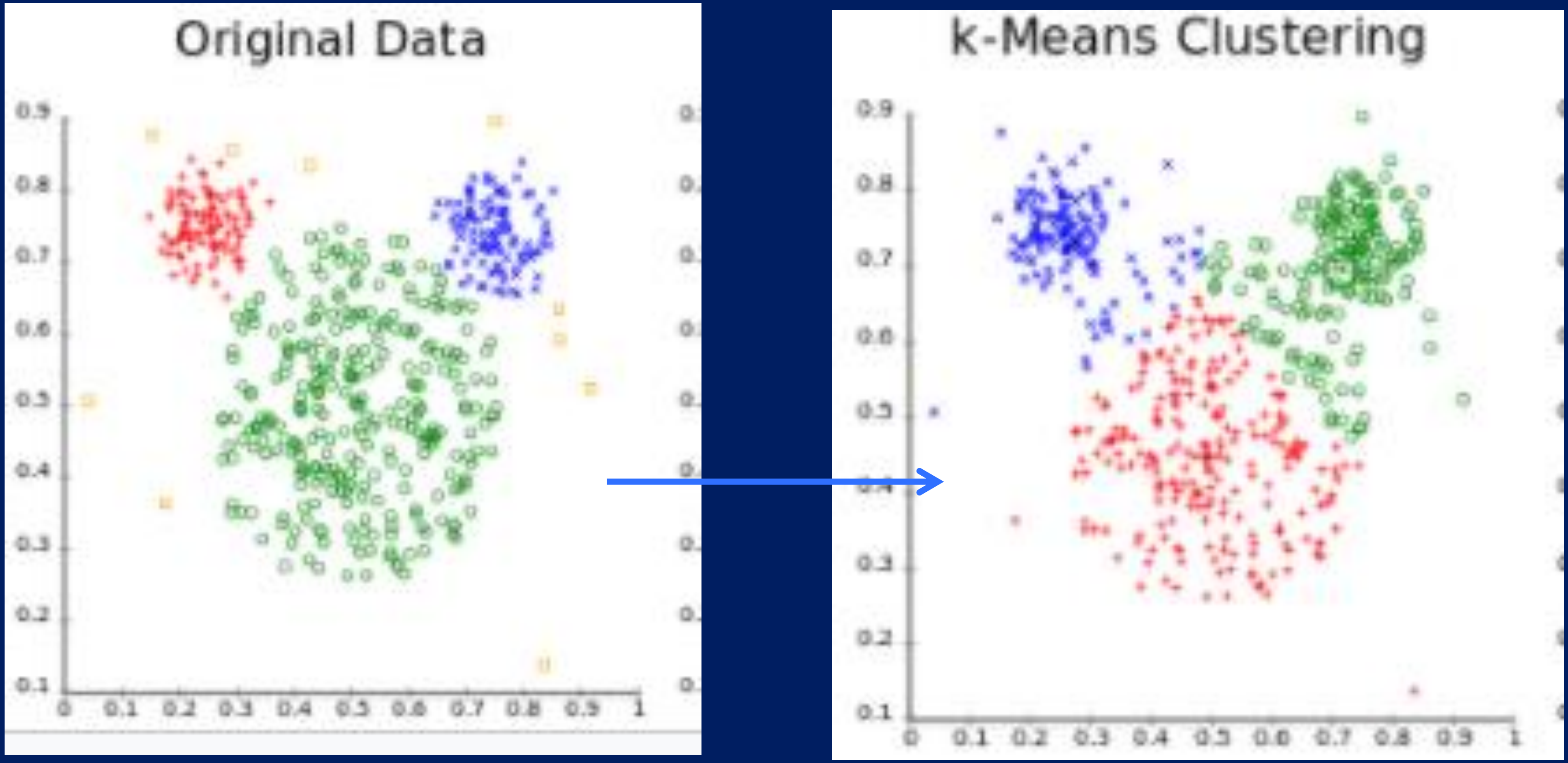
The algorithm may end up in a local optimal

- Re-start with different initial clusters and see if there is a big difference

The clusters have similar area (not cardinality)

- Use other methods to verify the results (next)

Examples



Example of k-means’ tendency to end-up with similar size clusters

Examples

Data	Result of k-means with k=2	More sensible result
{1,2,3,8,9,10,25}	{1,2,3,8} {9,10,25}	{1,2,3} {8,9,10,25}

k-medoids

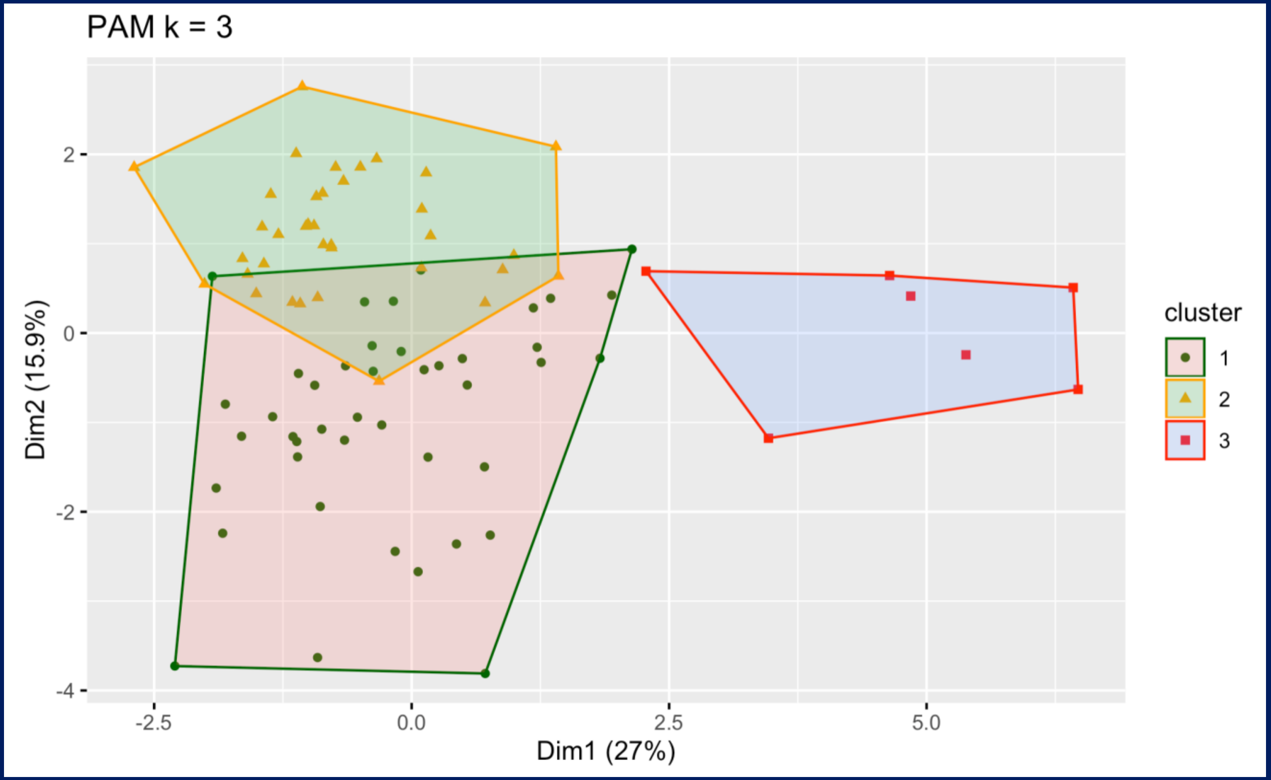
- It is very similar to K-means method
- The main idea is to use a point in the data as the center of each cluster instead of the mean values of all observations in the same cluster
  - This makes it more robust to outliers because it does not rely on the mean values. (Why?)
- Partitioning Around Medoids (PAM) Algorithm (for fixed k)

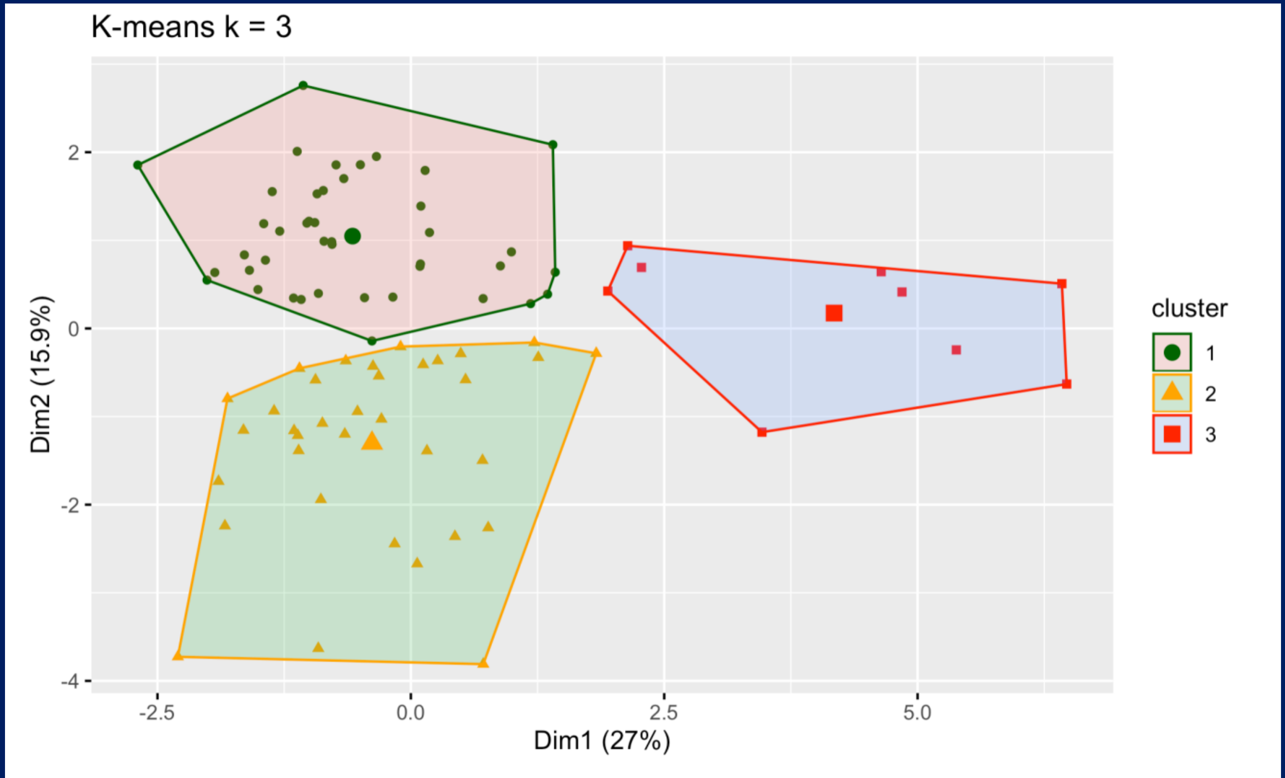
1. Arbitrarily choose *k* objects as the medoids
2. **Repeat**
3. Assign each remaining object to the cluster with the nearest medoid
4. Randomly select a point *p* that is currently not a medoid
5. Compute the total cost *S* of swapping each current medoid o_j with *p*
6. If $S<0$ swap o_j with *p* to form the new set of k medoids
7. **Until no change**

$$S = \sum_{i=1..k} \sum_{p} dist(p, o_i)$$

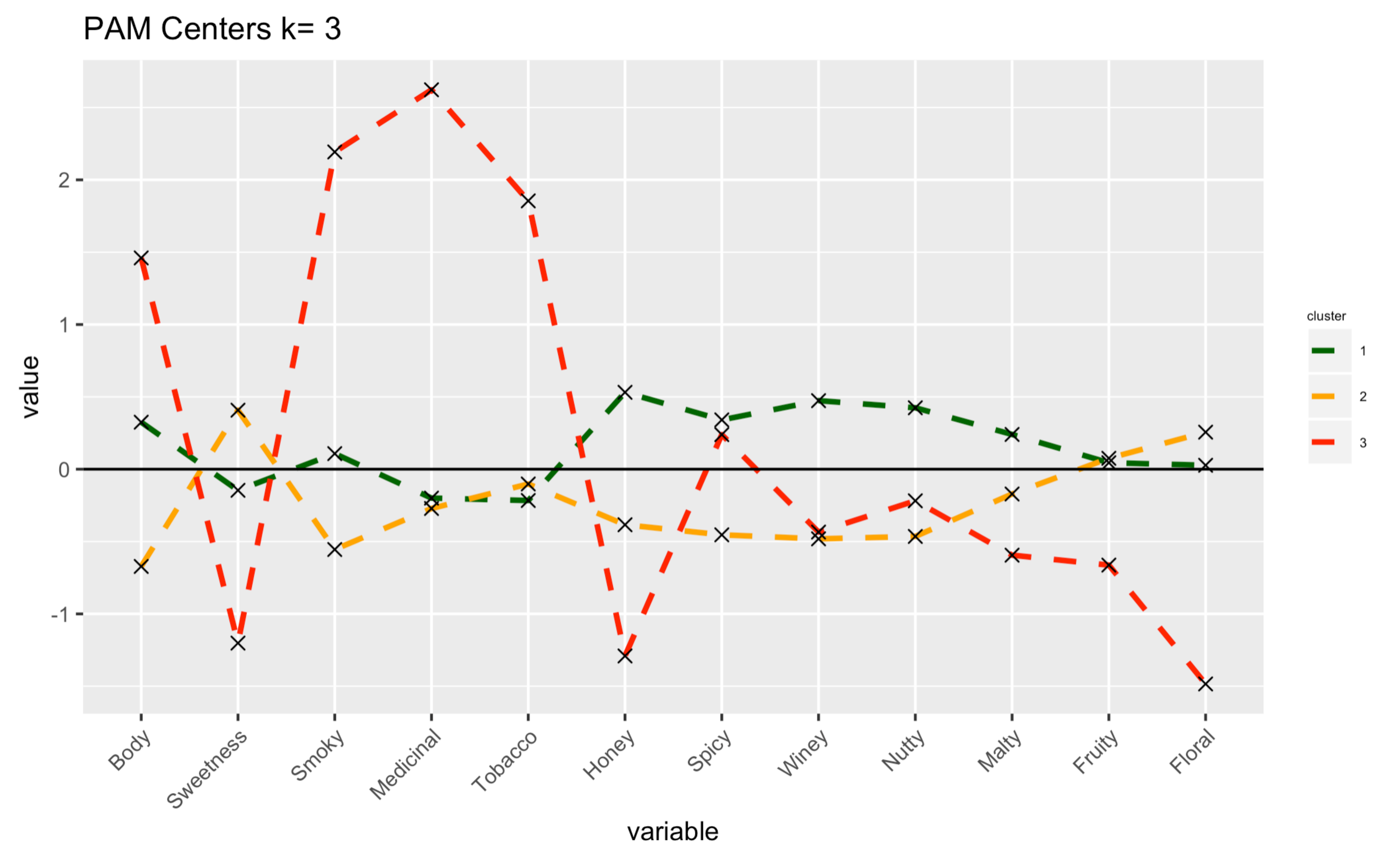
Distance between each point and its closest medoid

PAM on Whisky Data





PAM on Whisky Data



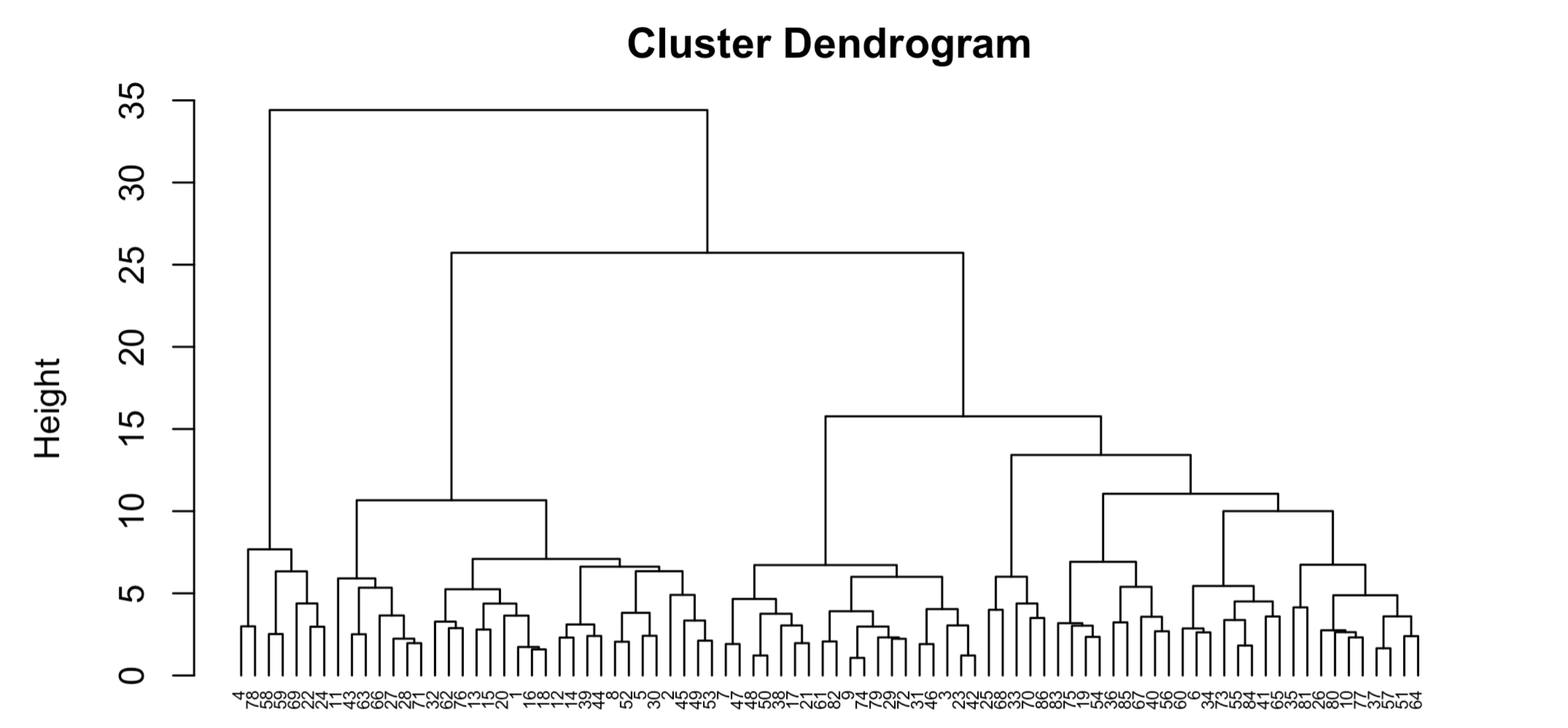
London Business School

Hierarchical Clustering

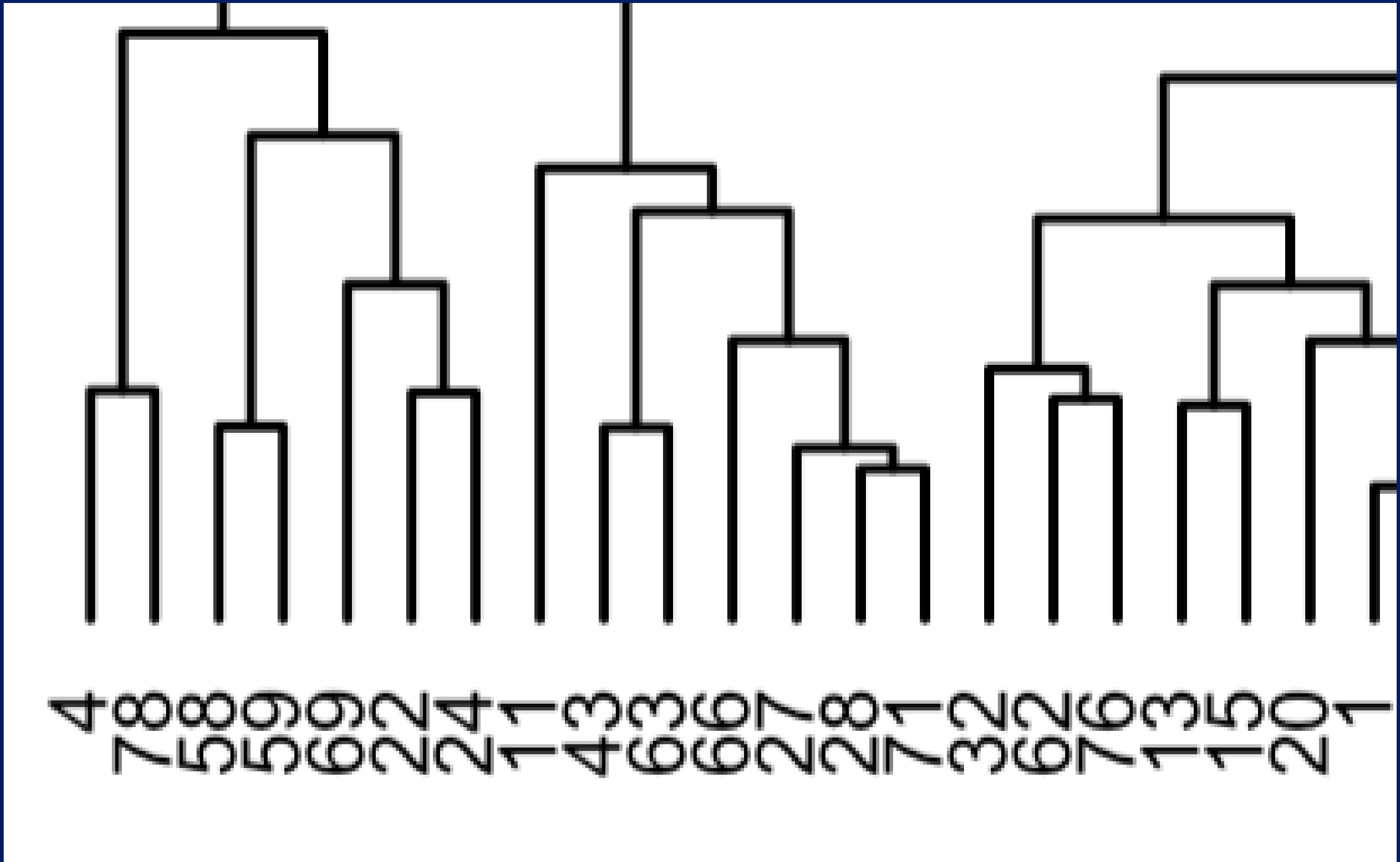
Unfortunately	• We don't know how many clusters we need before we do the analysis, hence it is in general difficult to determine k in k-means • And sometimes the clusters have significantly different sizes
Hierarchical clustering	• creates a visualization to see the impact of changing the number of clusters
The algorithm	• Start with each point in its own cluster. • Identify the closest two clusters and merge them. • Repeat. • Ends when all points are in a single cluster.

London Business School

Dendrogram: Whisky Data Set



Dendrogram: Whisky Data Set



How do we measure distance?

Between two points

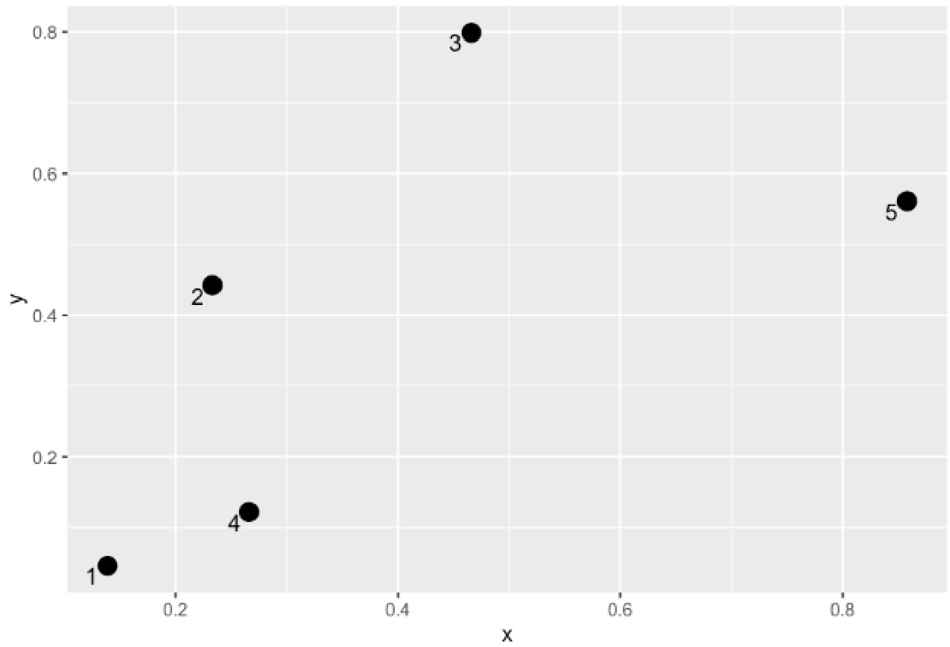
- Euclidean, Manhattan or etc.

Between two sets of points

- Complete linkage: The distance between two clusters is defined as the **maximum** value of all pairwise distances between the elements in cluster 1 and the elements in cluster 2. It tends to produce more compact clusters.
- Single linkage: The distance between two clusters is defined as the minimum value of all pairwise distances between the elements in cluster 1 and the elements in cluster 2. It tends to produce long, “loose” clusters.
- Average linkage: The distance between two clusters is defined as the average distance between the elements in cluster 1 and the elements in cluster 2.
- Ward’s minimum variance method: It minimizes the total within-cluster variance. At each step the pair of clusters with minimum between-cluster distance are merged.
- There are many more

Illustration of Hierarchical Clustering

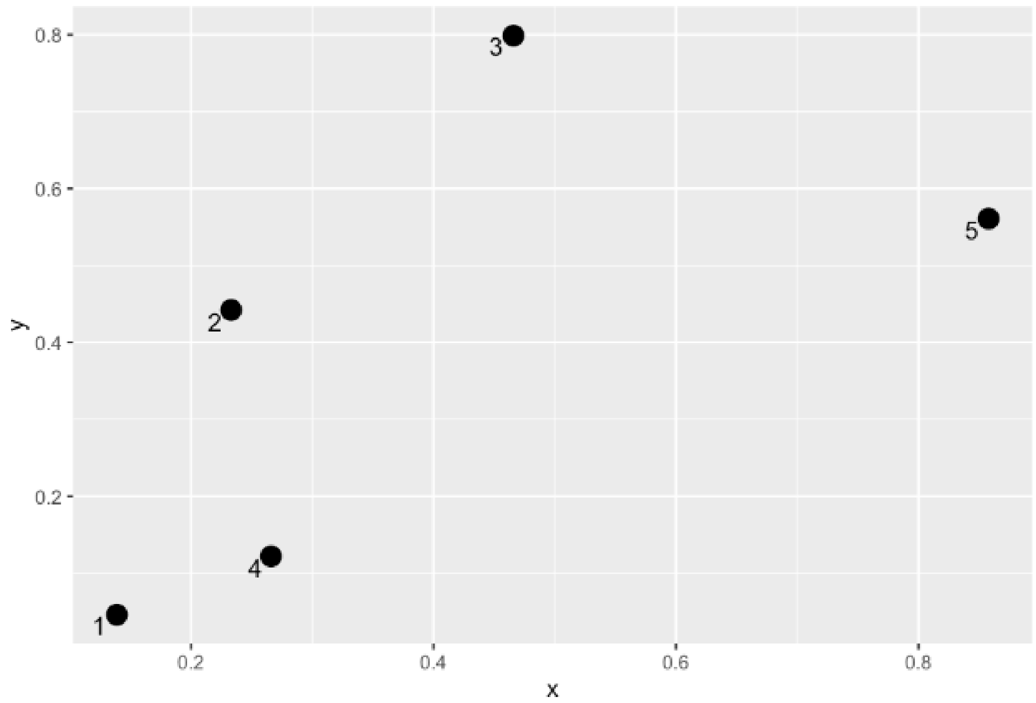
Random Data




```
res.dist <- (round(dist(data[,2:3], method = "euclidean"),2))
```

	1	2	3	4
2	0.41			
3	0.82	0.43		
4	0.15	0.32	0.71	
5	0.88	0.64	0.46	0.74

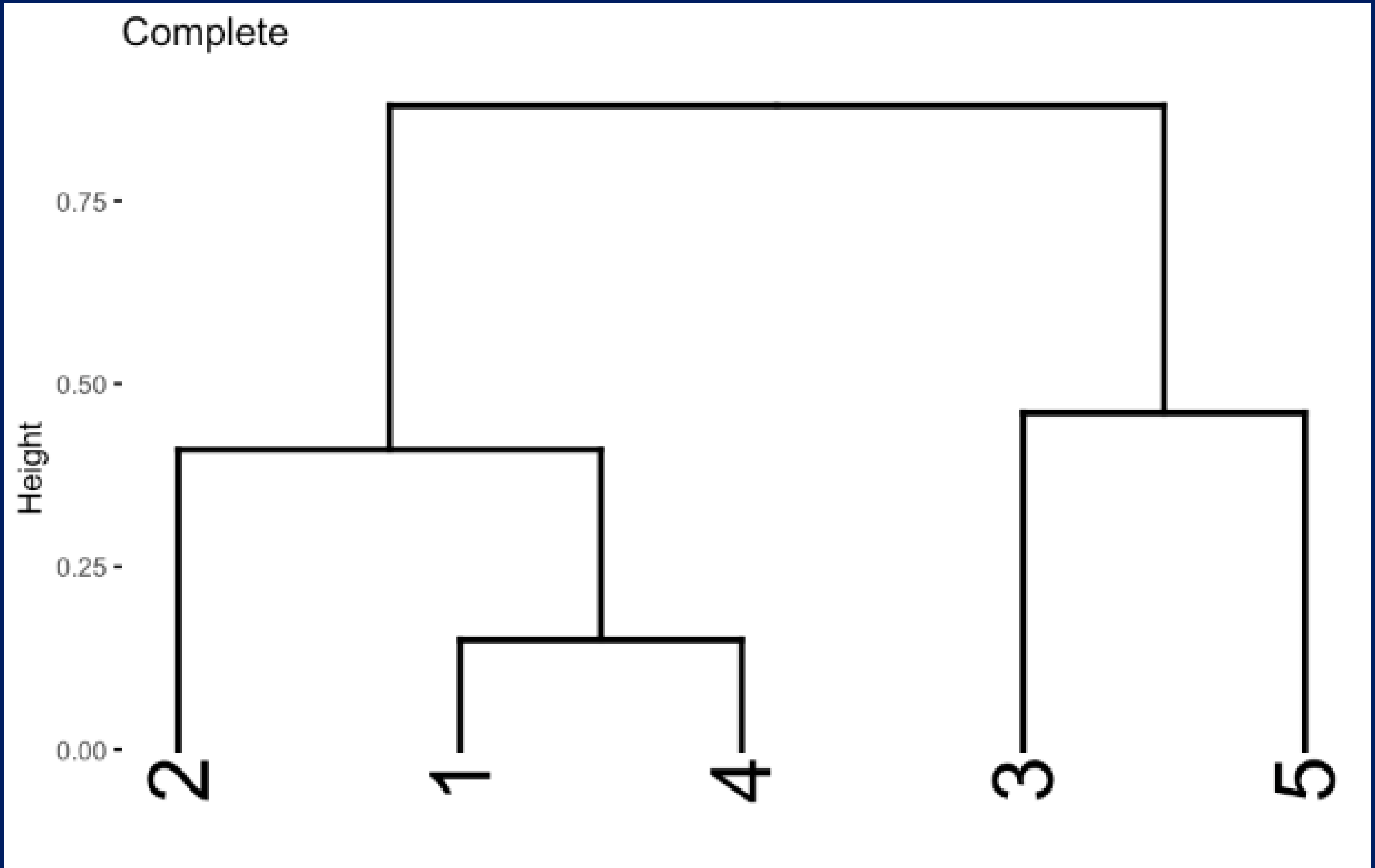
Complete Linkage



$$d(A,B) \equiv \max_{\vec{x}\in A, \vec{y}\in B} \|\vec{x}-\vec{y}\|$$

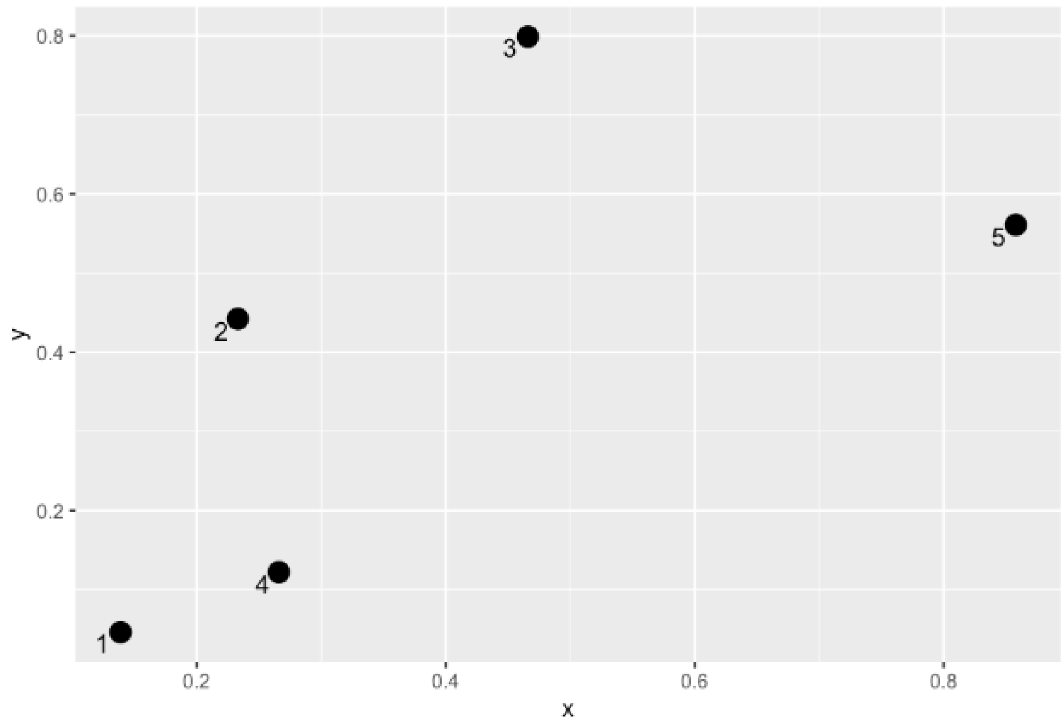
	1	2	3	4
2	0.41			
3	0.82	0.43		
4	0.15	0.32	0.71	
5	0.88	0.64	0.46	0.74

Complete Linkage



London Business School

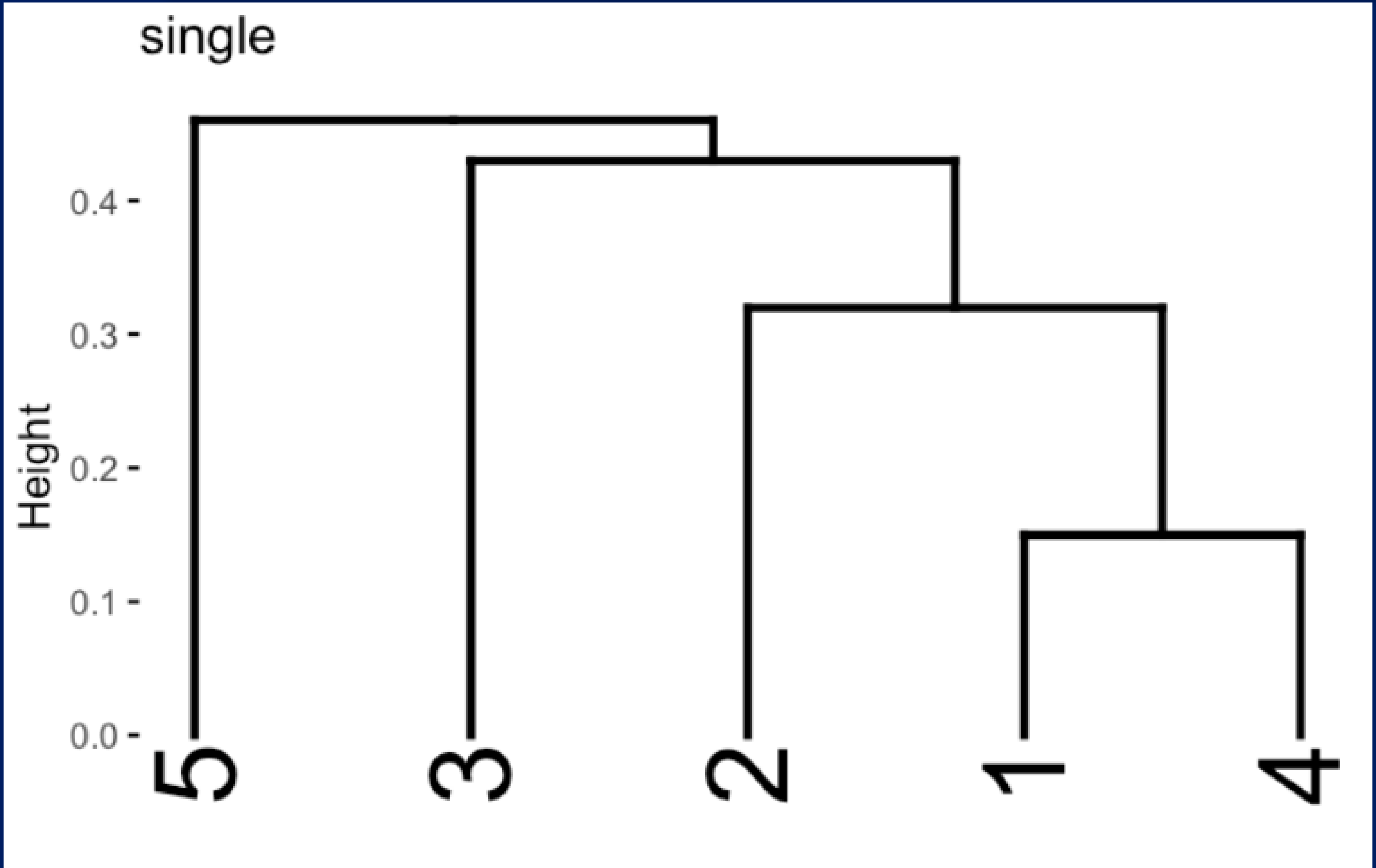
Single Linkage



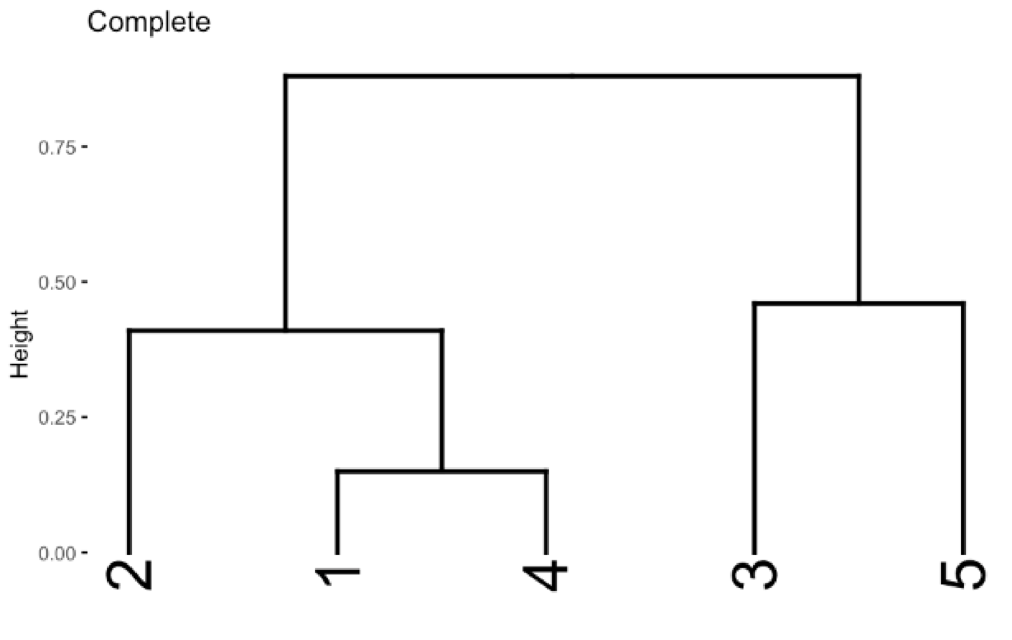
$$d(A,B) \equiv \min_{\vec{x}\in A, \vec{y}\in B} \|\vec{x} - \vec{y}\|$$

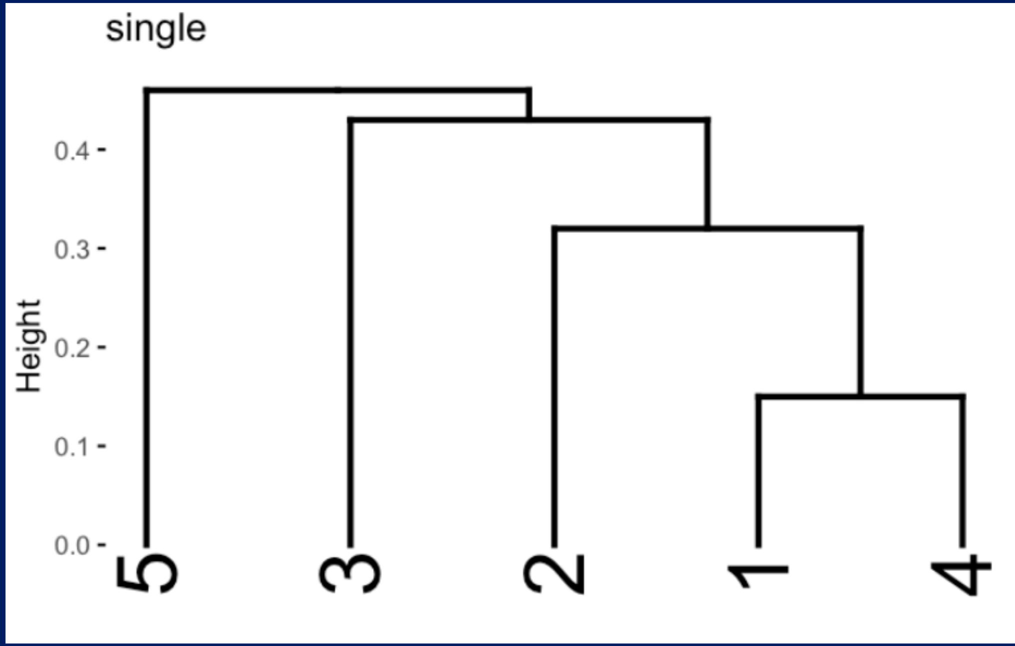
	1	2	3	4
2	0.41			
3	0.82	0.43		
4	0.15	0.32	0.71	
5	0.88	0.64	0.46	0.74

Single Linkage



Complete vs Single Linkage





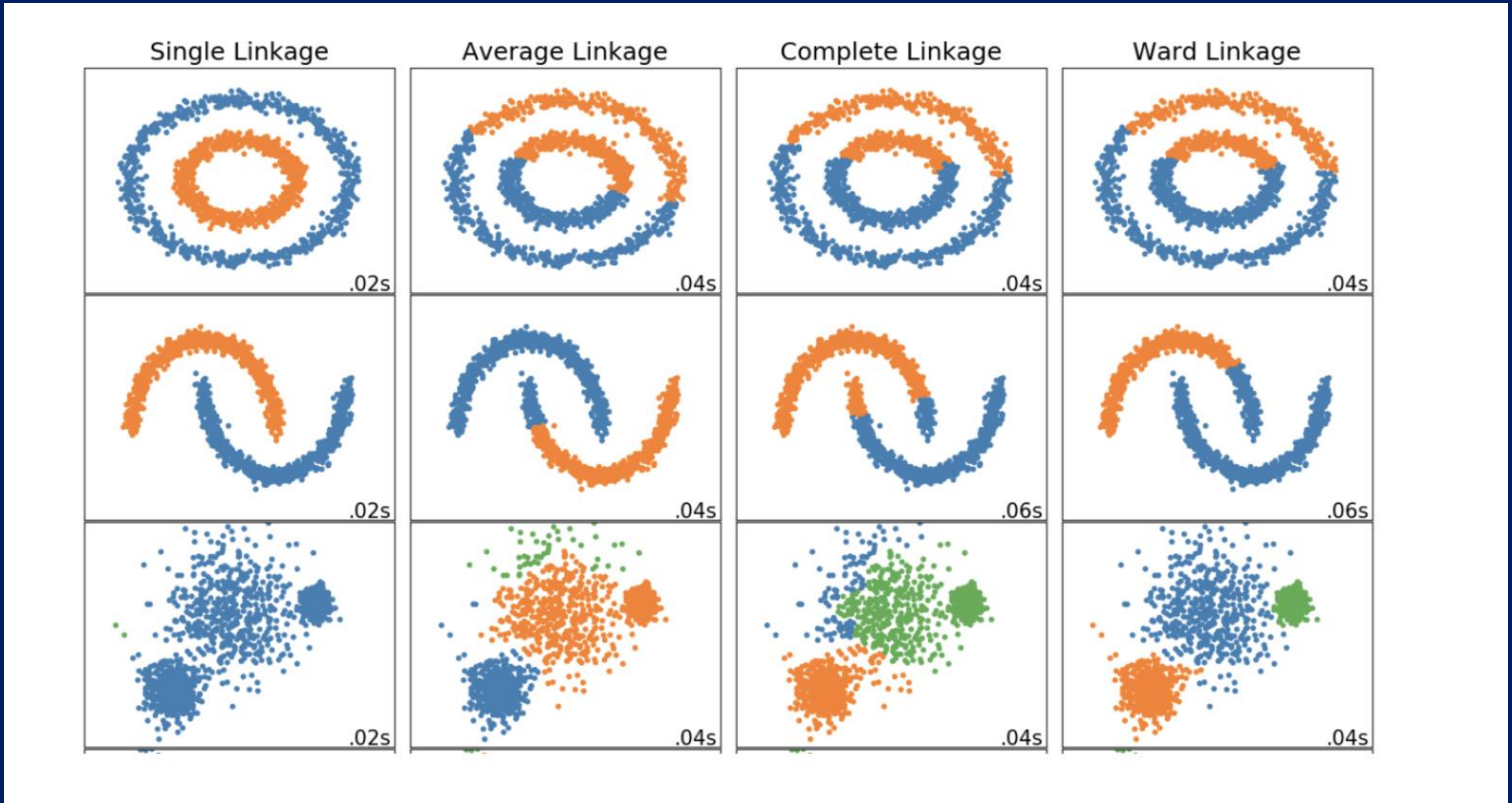
Ward's distance (linkage)

- Somewhere between k-means and hierarchical clustering
- It joins cluster pairs whose merger minimizes the increase in the total sum of squares within-group error.

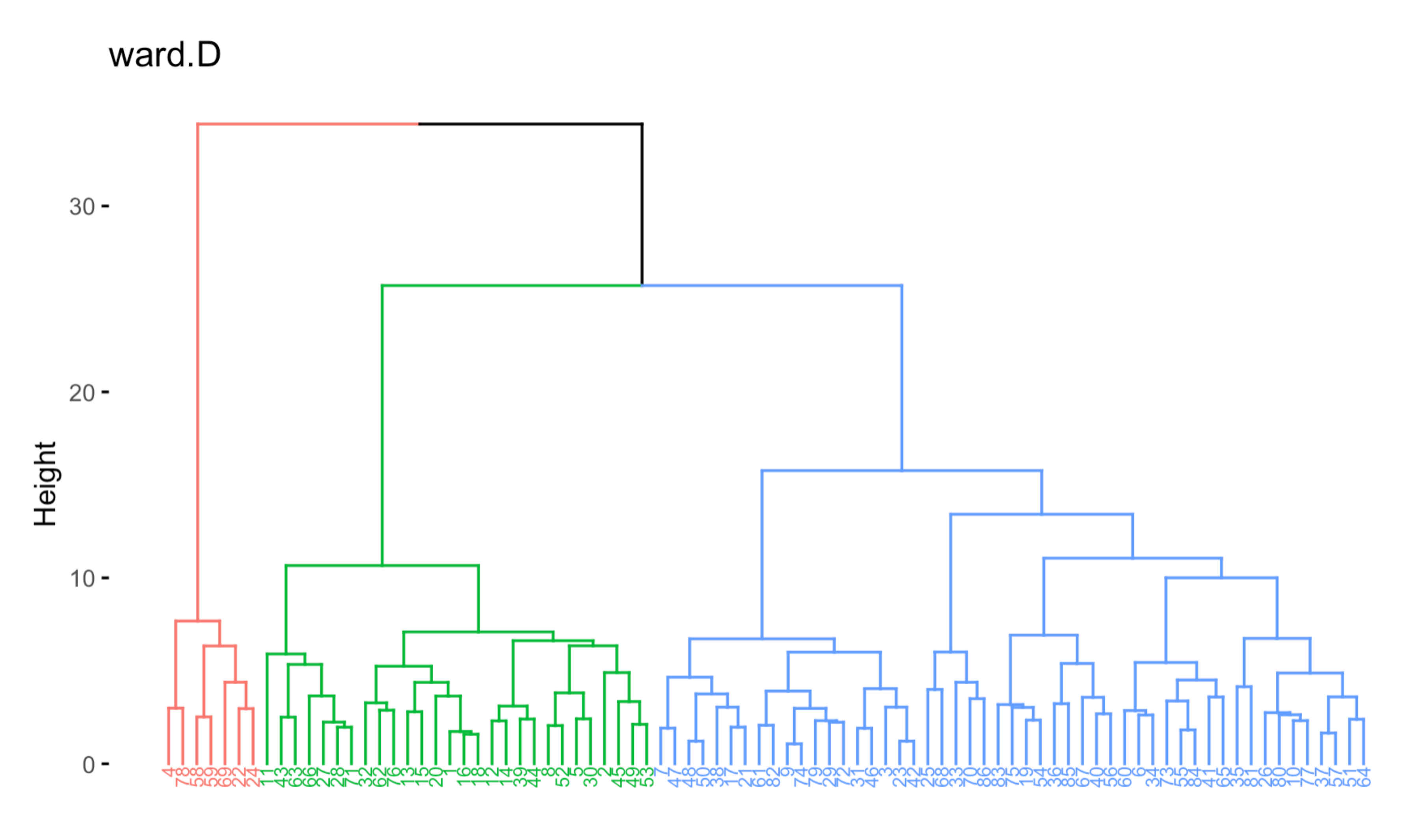
$$\Delta(A,B) = \sum_{i \in A \cup B} \|\vec{x}_i - \vec{m}_{A \cup B}\|^2 - \sum_{i \in A} \|\vec{x}_i - \vec{m}_A\|^2 - \sum_{i \in B} \|\vec{x}_i - \vec{m}_B\|^2$$

$$= \frac{n_A n_B}{n_A + n_B} \|\vec{m}_A - \vec{m}_B\|^2$$

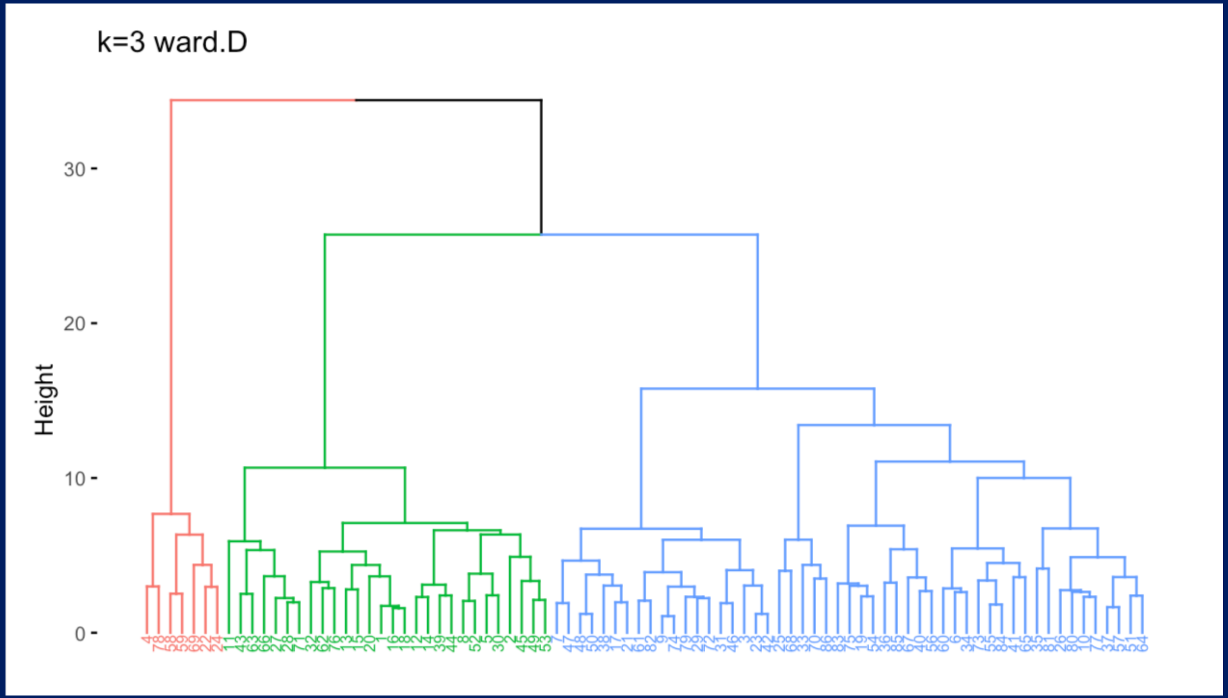
Distance measures

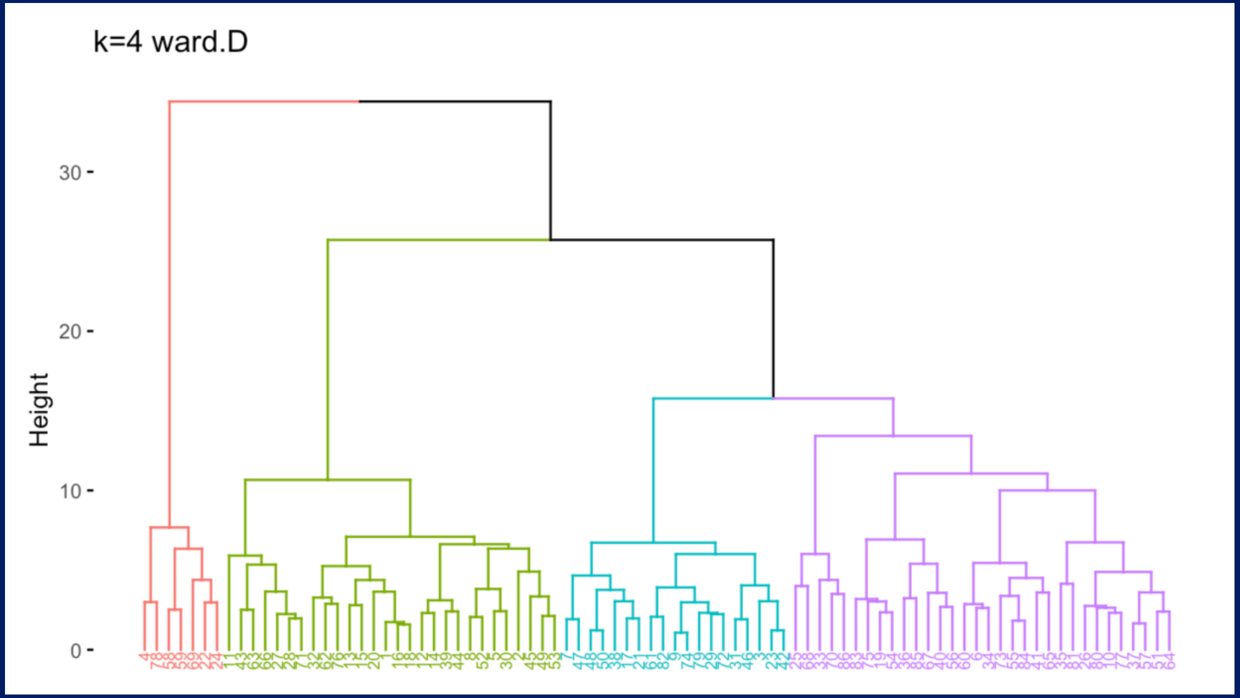


Dendrogram: Whisky Data Set

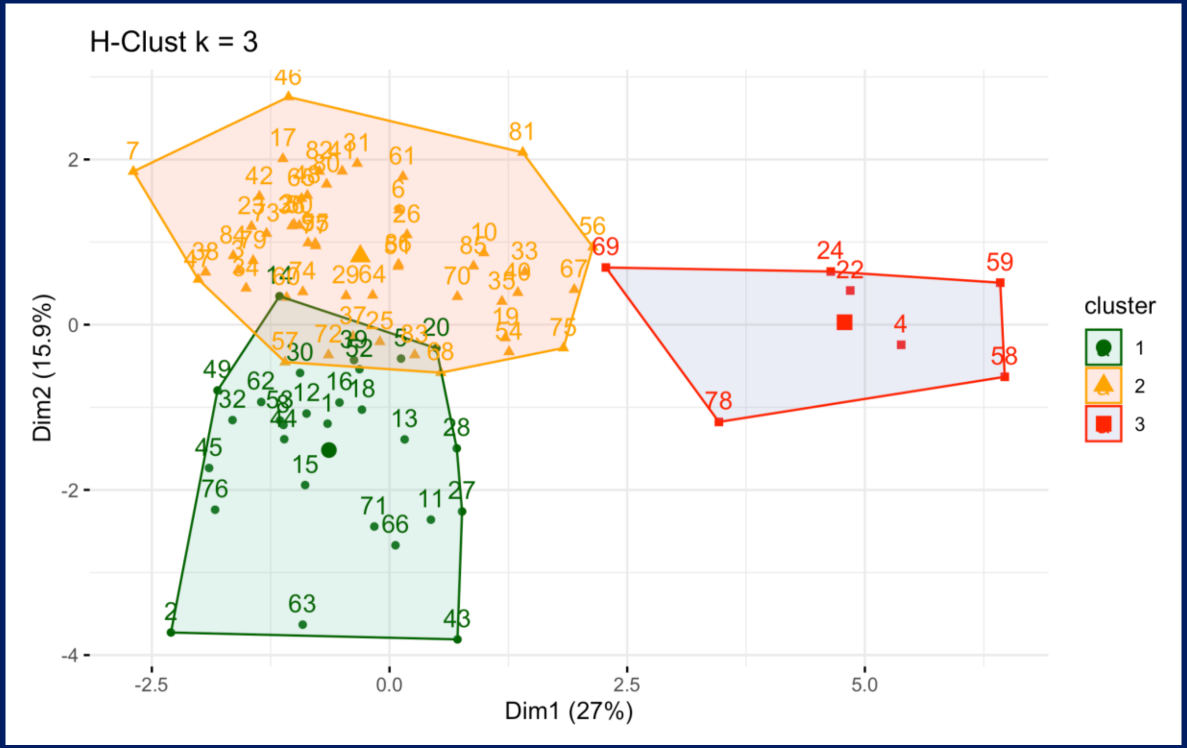


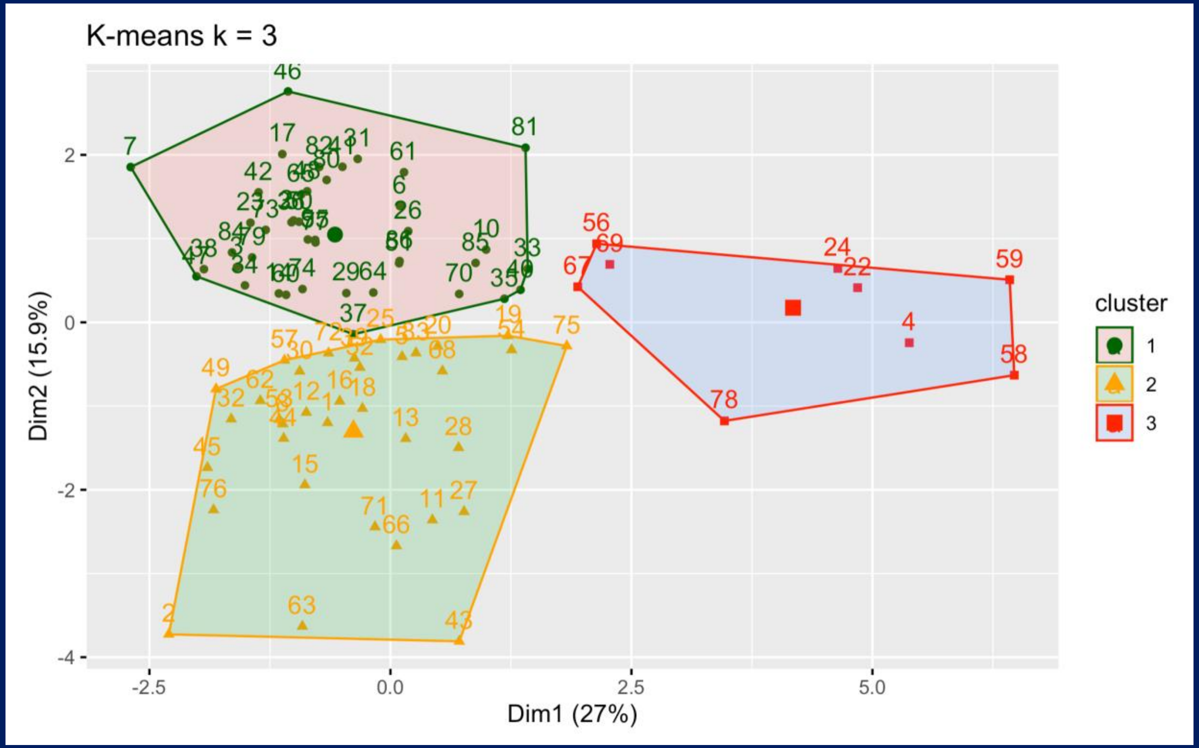
Dendrogram: Whisky Data Set





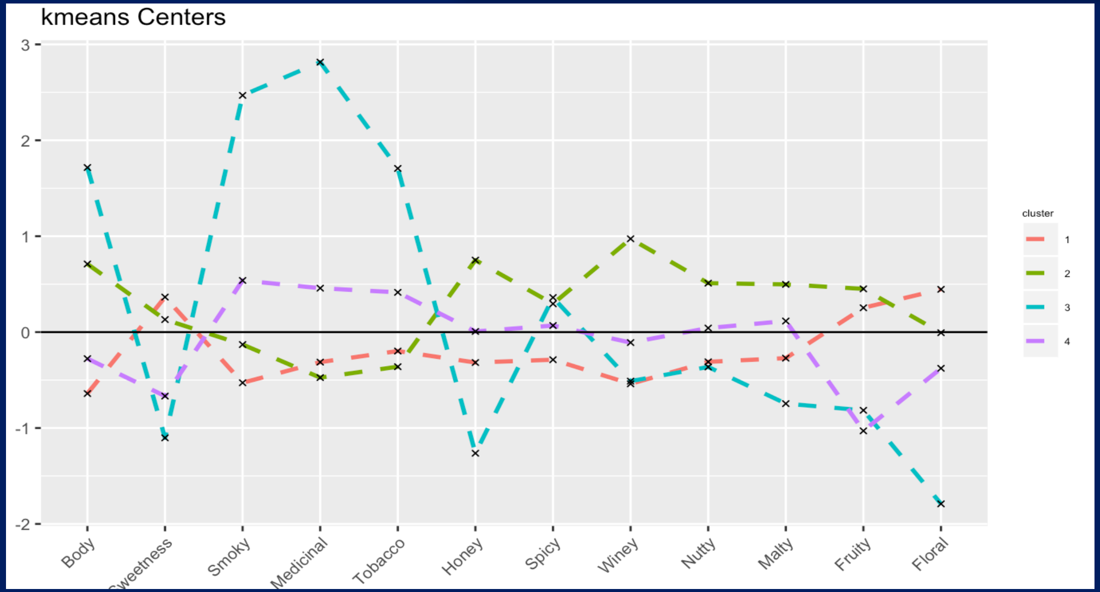
Dendrogram: Whisky Data Set

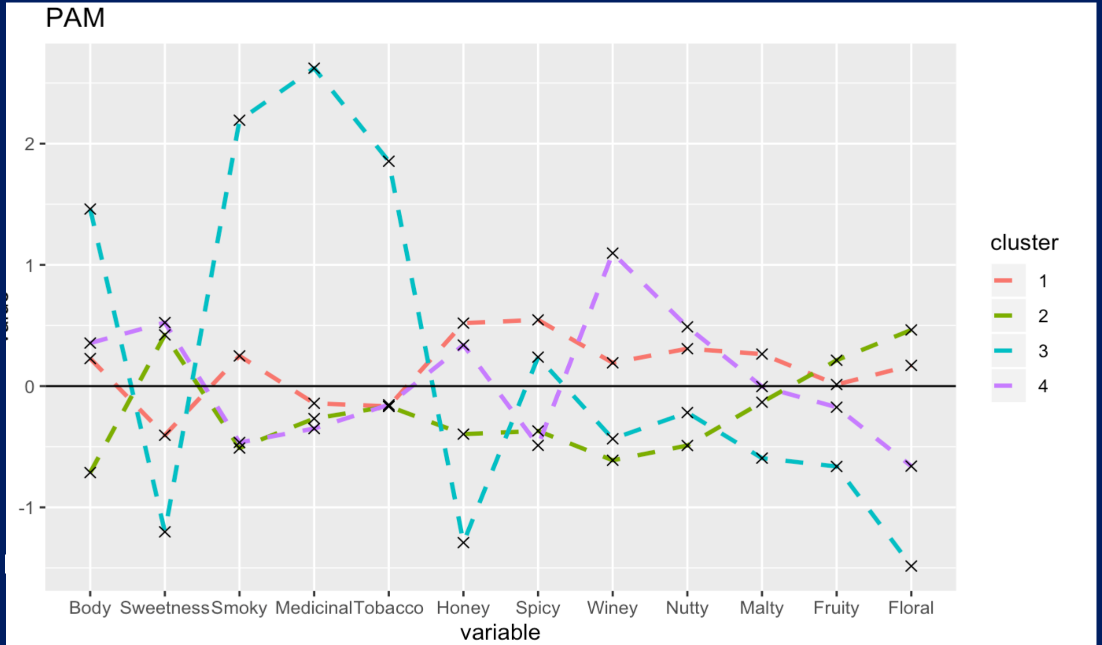


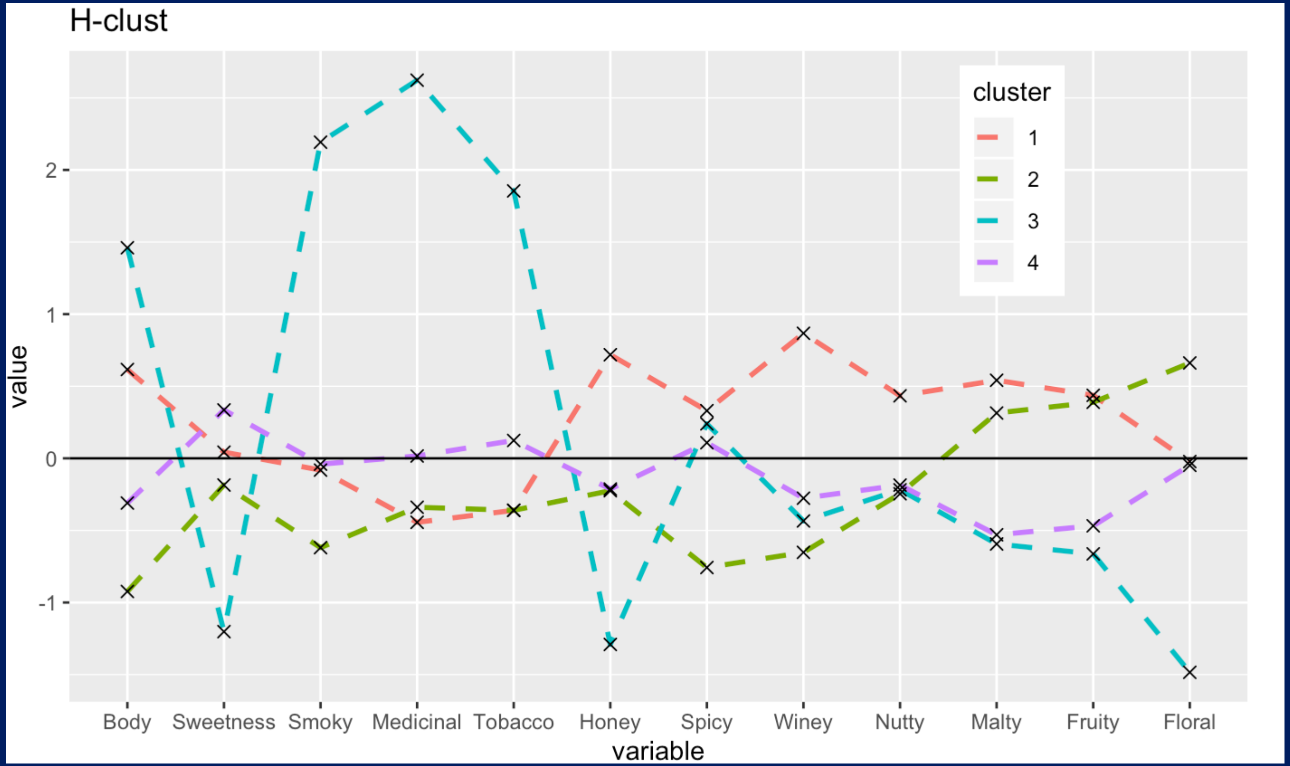


London Business School

Compare the results of clustering methods







Summary

Cluster analysis is an exploratory tool. Useful only when it produces **meaningful** clusters

We usually use a few methods to verify the accuracy of our conclusions.

Be wary of chance results; data may not have definitive “real” clusters- Texas sharpshooter fallacy

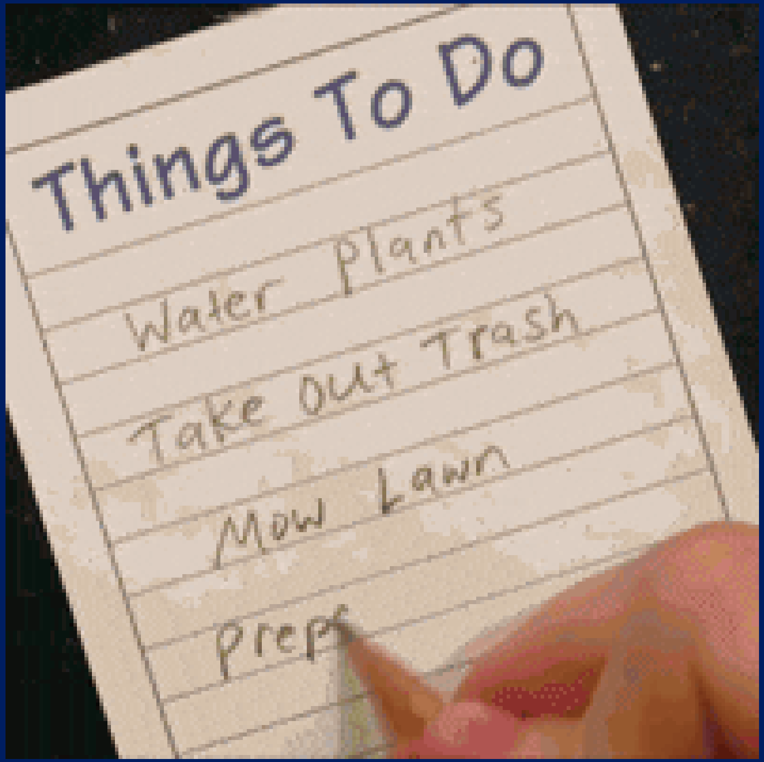
Learning objectives

This week

1. k-medoids or partitioning around medoids (PAM) algorithm
   - Objective
   - Inputs
   - Outputs
2. Hierarchical clustering
   - Objective
   - Inputs
   - Outputs
3. Applying clustering methods in a large data set (workshop)
   - Using different clustering methods
   - Visualization of the results
   - Choosing the best clustering results
   - Presenting your findings

Checklist

- Use different methods and identify the best results using visualization tools we covered last week
  - Elbow chart
  - PCA
  - Cluster centers
  - Silhouette
- Compare the clusters you found under different methods
- Those that are prevalent in multiple methods are likely to be true clusters



Next: Workshop

Clustering BBC iPlayer users

Learning outcomes

- How to carry out a clustering project on a new data set
- Using three different clustering methods
  - K-Means
  - PAM
  - Hierarchical clustering
- Determining the (optimal) number of clusters in each method
- Comparing the results of different clustering methods
- Making business recommendations based on clustering results
- Sharing your results