Introduction to Basic R: Cluster Analysis

Introduction to Bioinformatics

Presented by Tran Ba Thien

March 19 2023

Machine learning algorithms

Type of outcome	Supervised learning	Un-supervised learning
Categorical /Discrete	Classification (Logistic regression, SVM,…)	Clustering (Hierarchical, K-means,…)
Continuous	Regression (Linear regression, Rigde/LASSO,…)	Dimensionality reduction (PCA, t-SNE, UMAP,…)

CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

SUPERVISED

Predict a category

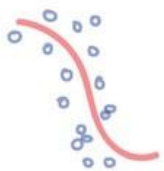
CLASSIFICATION

«Divide the socks by color»

Predict a number

REGRESSION

«Divide the ties by length»



Data is not labeled in any way

UNSUPERVISED

Divide by similarity

CLUSTERING

«Split up similar clothing into stacks»

Find hidden dependencies

DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»

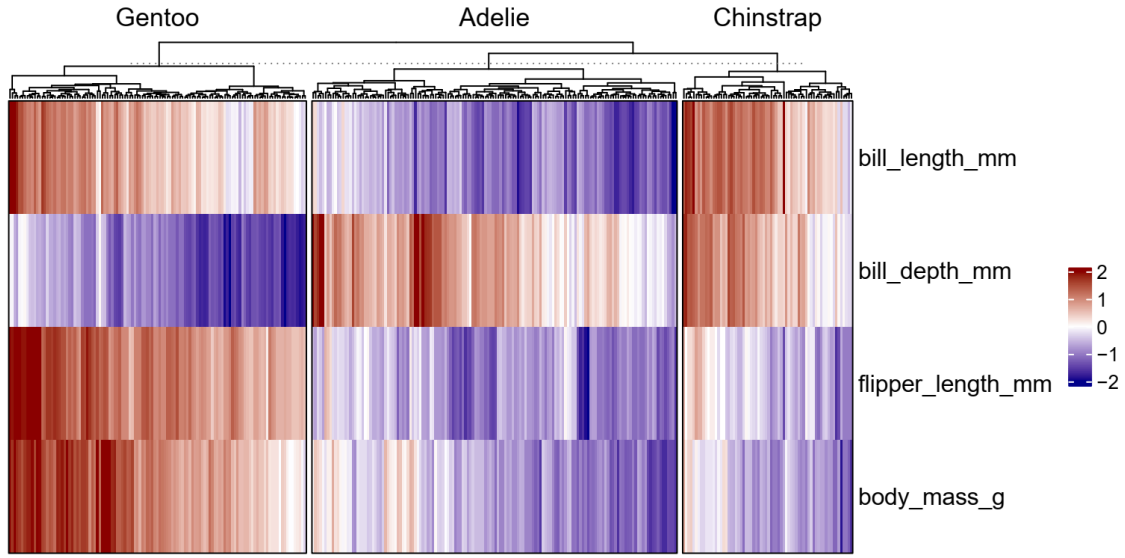
Identify sequences

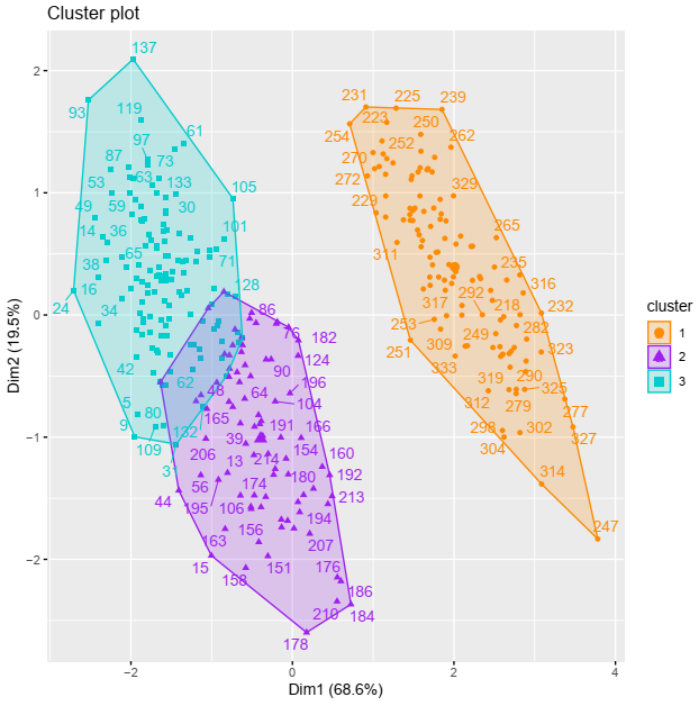
ASSOCIATION

«Find what clothes I often wear together»

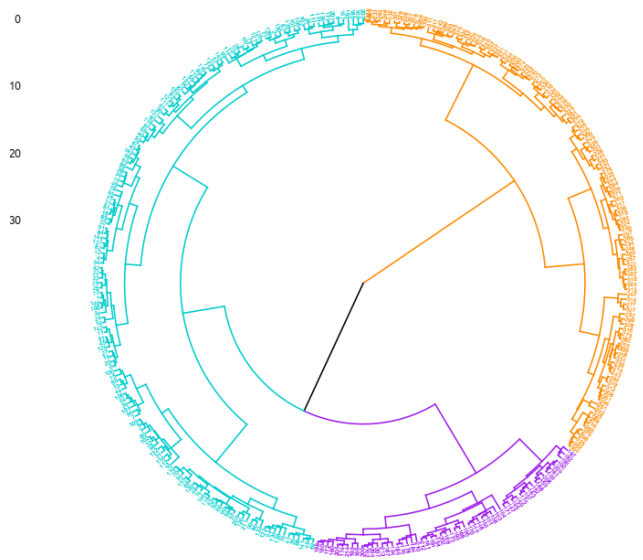
Content

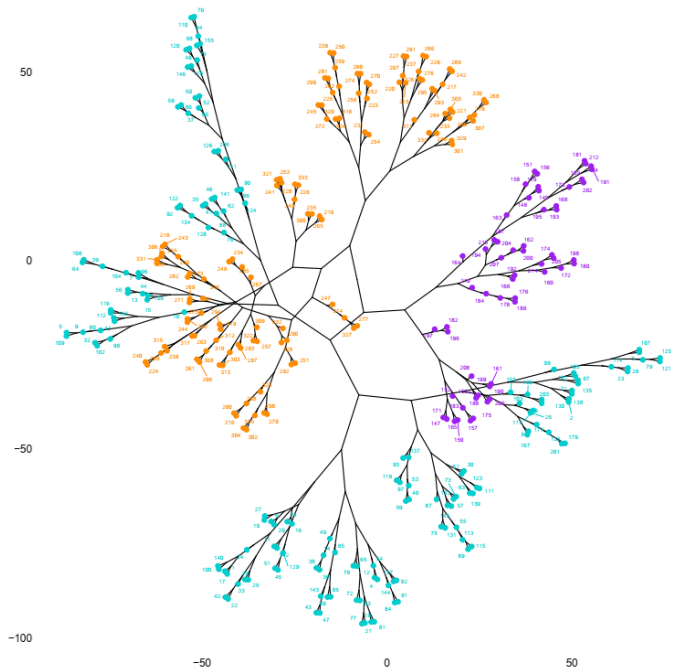
- What is cluster analysis?
- Hierarchical clustering
- K-means clustering


Gentoo
Adelie
Chinstrap
bill_length_mm
bill_depth_mm
flipper_length_mm
body_mass_g
2
1
0
−1
−2


Cluster plot
Dim2 (19.5%)
Dim1 (68.6%)
cluster
1
2
3

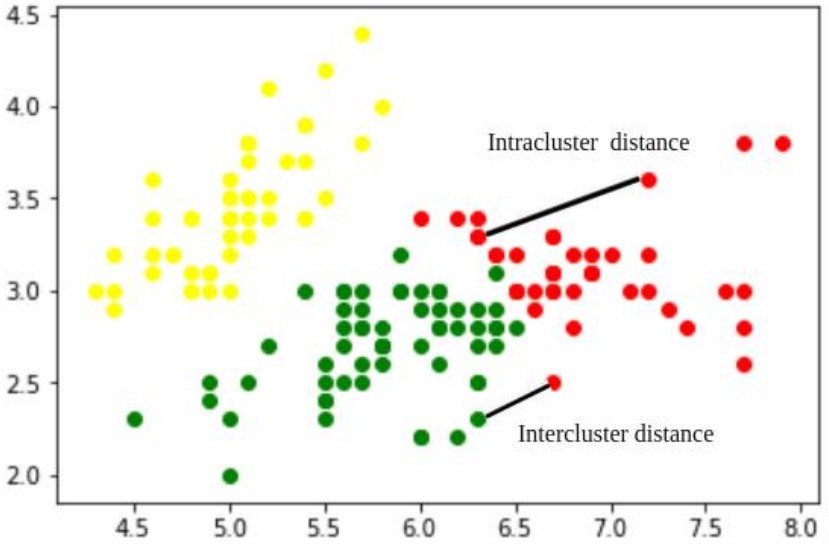

Cluster Dendrogram
Rectangle
Distance





What is clustering?

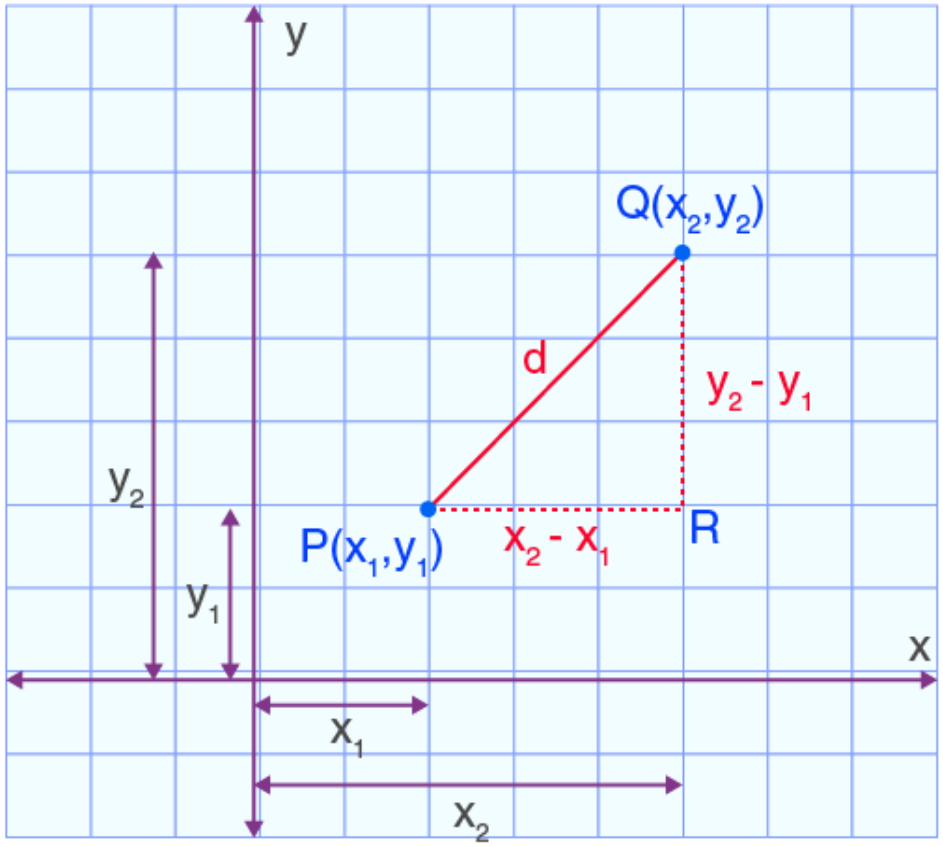
- The organization of unlabeled data into similarity groups called clusters.
- A cluster is a collection of data items which are “similar” between them, and “dissimilar” to data items in other clusters.
- The similarity or dissimilarity between objects or data points is calculated using a distance metric.



Measurement of distance metric

- Manhattan distance
- Minkowski distance
- Pearson correlation distance
- Binary distance
- Euclidean distance:

$$d_{(Q,P)} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



Example of distance metric

ID	Flipper Length (mm)	Body Mass (g)
A	187	3350
B	184	3325
C	187	3250
D	224	5650
E	202	3875

Distance between A and B:

$$d_{(A,B)} = \sqrt{(187 - 184)^2 + (3350 - 3325)^2} = 25.18$$

Distance between A and D?

Example of distance metric

ID	Flipper Length (mm)	Body Mass (g)
A	187	3350
B	184	3325
C	187	3250
D	224	5650
E	202	3875

Distance between A and B:

$$d_{(A,B)} = \sqrt{(187 - 184)^2 + (3350 - 3325)^2} = 25.18$$

Distance between A and D:

$$d_{(A,D)} = \sqrt{(187 - 224)^2 + (3350 - 5650)^2} = 2300.29$$

Example of distance metric

ID	Flipper Length (mm)	Body Mass (g)
A	187	3350
B	184	3325
C	187	3250
D	224	5650
E	202	3875

$$x_{scaled} = \frac{x - mean}{sd}$$

Scaling

ID	Flipper Length (mm)	Body Mass (g)
A	-0.58	-0.53
B	-0.76	-0.56
C	-0.58	-0.63
D	1.62	1.73
E	0.31	-0.01

Distance between A and B:

$$d_{(A,B)} = \sqrt{(187 - 184)^2 + (3350 - 3325)^2}$$
$$= 25.18$$

Distance between scaled A and scaled B?

Distance between A and D:

$$d_{(A,B)} = \sqrt{(187 - 224)^2 + (3350 - 5650)^2}$$
$$= 2300.29$$

Distance between scaled A and scaled D?

Example of distance metric

ID	Flipper Length (mm)	Body Mass (g)
A	187	3350
B	184	3325
C	187	3250
D	224	5650
E	202	3875

$$x_{scaled} = \frac{x - mean}{sd}$$

Scaling

ID	Flipper Length (mm)	Body Mass (g)
A	-0.58	-0.53
B	-0.76	-0.56
C	-0.58	-0.63
D	1.62	1.73
E	0.31	-0.01

Distance between A and B:

$$d_{(A,B)} = \sqrt{(187 - 184)^2 + (3350 - 3325)^2} = 25.18$$

Distance between A and D:

$$d_{(A,B)} = \sqrt{(187 - 224)^2 + (3350 - 5650)^2} = 2300.29$$

Distance between scaled A and scaled B:

$$d_{(A,B)} = \sqrt{\big((-0.58) - (-0.76)\big)^2 + \big((-0.53) - (-0.56)\big)^2} = 0.18$$

Distance between scaled A and scaled D:

$$d_{(A,D)} = \sqrt{\big((-0.58) - 1.62\big)^2 + \big((-0.53) - 1.73\big)^2} = 3.15$$

Example of distance metric

```
> subDT[,c(5,6)]
  flipper_length_mm body_mass_g
A               187        3350
B               184        3325
C               187        3250
D               224        5650
E               202        3875
> dist(subDT[,c(5,6)], method = "euclidean")
           A          B          C          D
B   25.17936
C  100.00000   75.05998
D 2300.29759 2325.34406 2400.28519
E  525.21424  550.29447  625.17997 1775.13633
```

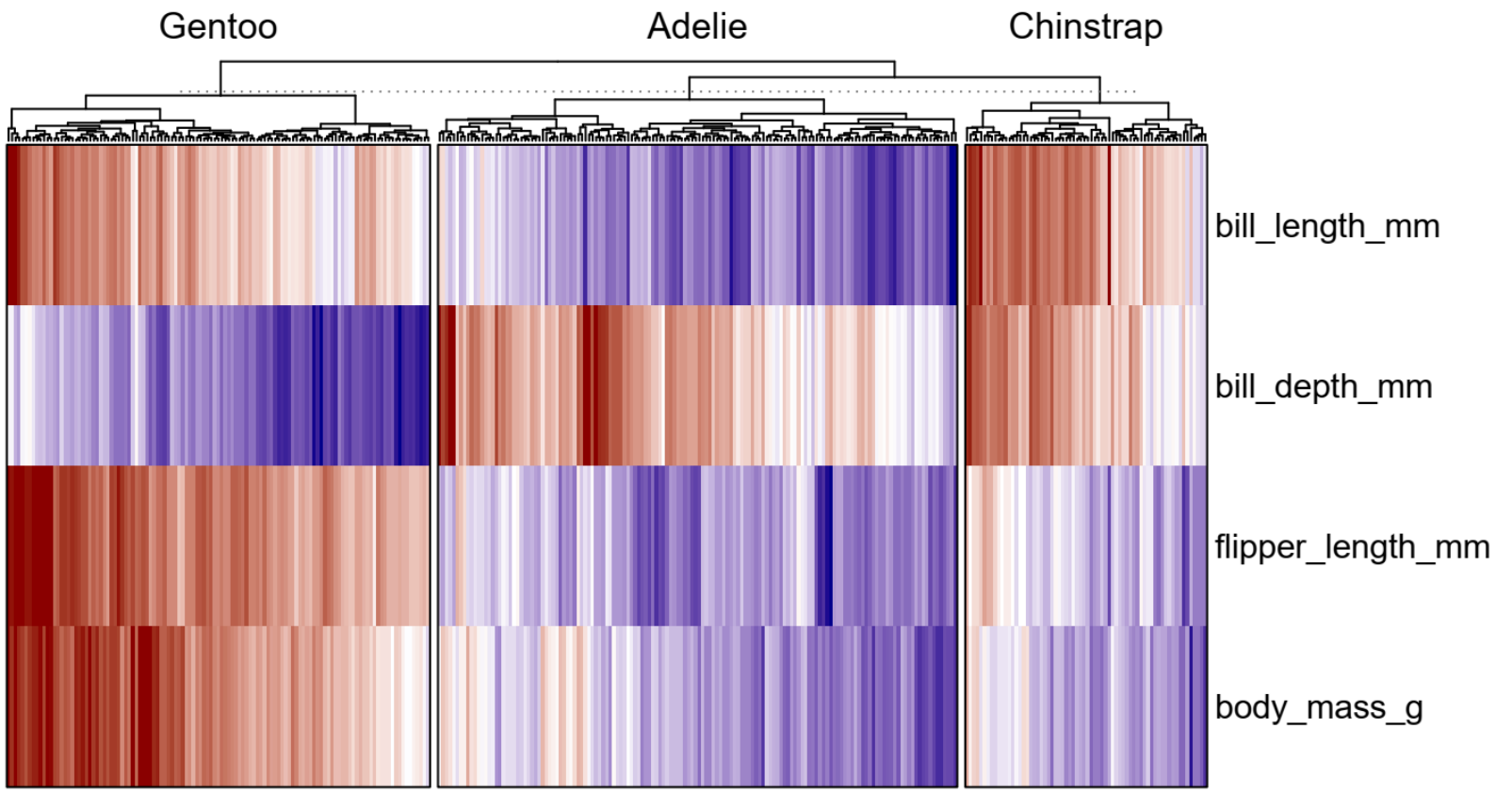
Other method

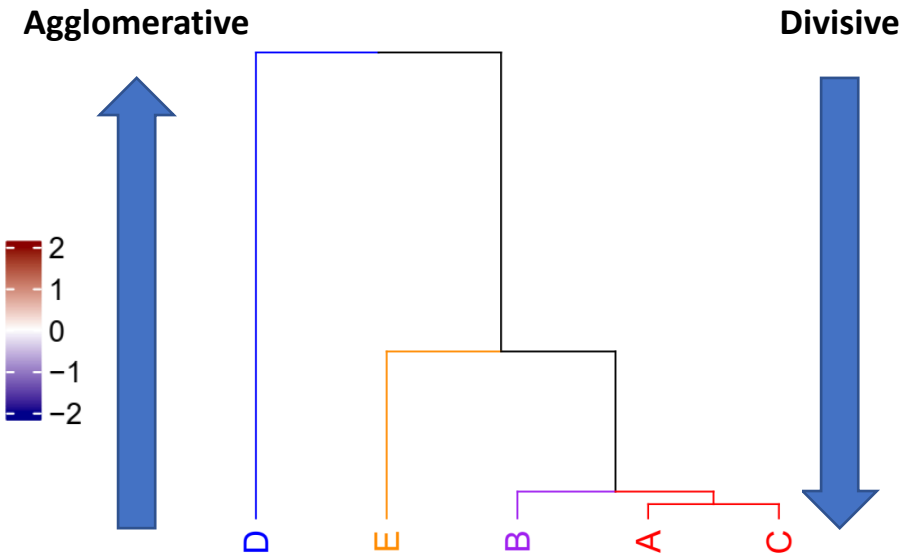
```
> dist(scale(subDT[,c(5,6)]), method = "manhattan")
           A          B          C          D
B 0.20369907
C 0.09855386 0.25297600
D 4.47515293 4.67885200 4.57370679
E 1.41271080 1.61640987 1.51126466 3.06244213
> dist(scale(subDT[,c(5,6)]), method = "minkowski")
           A          B          C          D
B 0.18074776
C 0.09855386 0.19371677
D 3.16467972 3.30914572 3.23600098
E 1.03405914 1.20335836 1.08672731 2.18733167
```

```
> scale(subDT[,c(5,6)])
  flipper_length_mm body_mass_g
A        -0.5849313 -0.53219085
B        -0.7639919 -0.55682931
C        -0.5849313 -0.63074471
D         1.6234828  1.73454795
E         0.3103717 -0.01478308
attr(,"scaled:center")
flipper_length_mm       body_mass_g
            196.8            3890.0
attr(,"scaled:scale")
flipper_length_mm       body_mass_g
          16.7541          1014.6736
> dist(scale(subDT[,c(5,6)]), method = "euclidean")
           A          B          C          D
B 0.18074776
C 0.09855386 0.19371677
D 3.16467972 3.30914572 3.23600098
E 1.03405914 1.20335836 1.08672731 2.18733167
```

Hierarchical clustering

- Grouping similar data points together in a hierarchy of clusters.
- Two types: agglomerative and divisive.



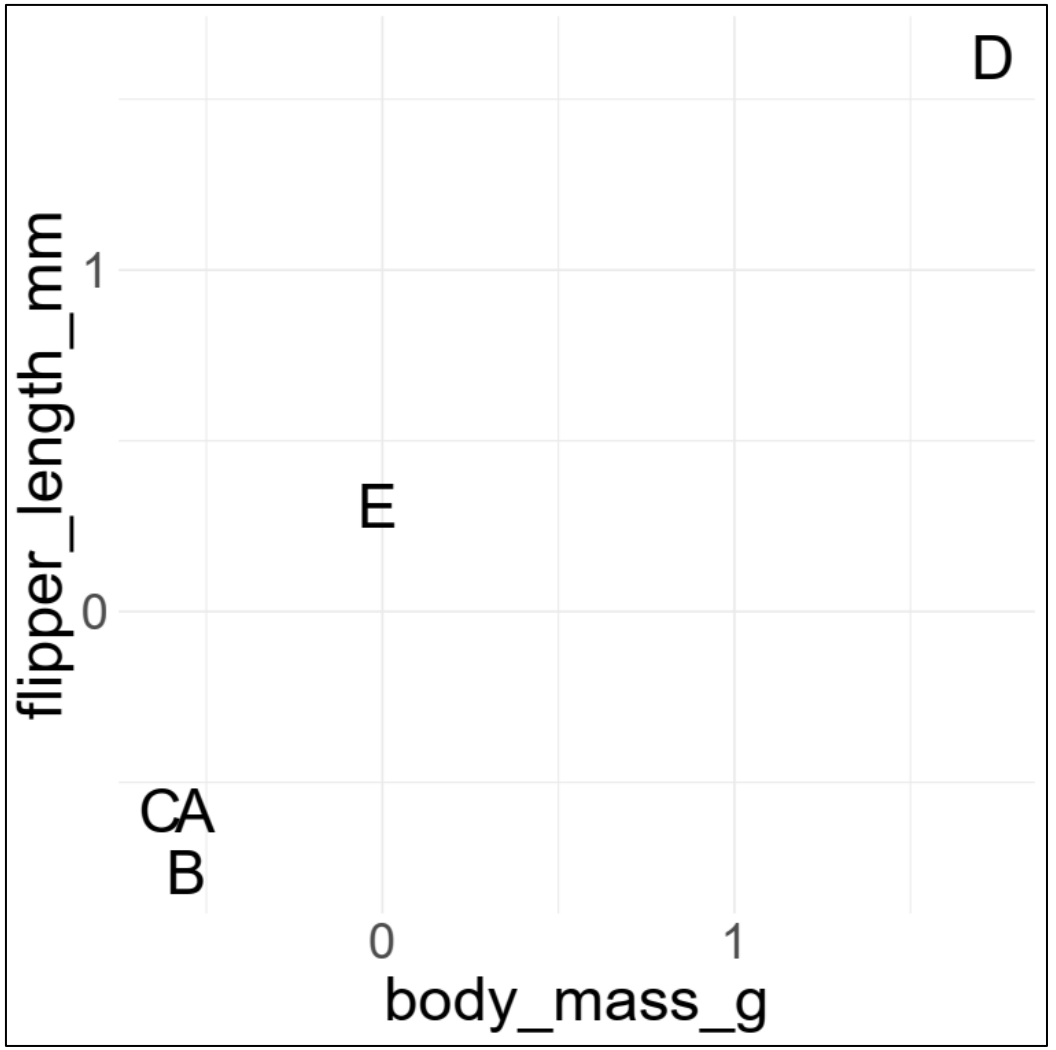


How hierarchical clustering work?

ID	Flipper Length (mm)	Body Mass (g)
A	-0.58	-0.53
B	-0.76	-0.56
C	-0.58	-0.63
D	1.62	1.73
E	0.31	-0.01

0					
	A	B	C	D	E
A	0	0.18	0.10	3.16	1.03
B		0	0.19	3.31	1.20
C			0	3.24	1.09
D				0	2.19
E					0

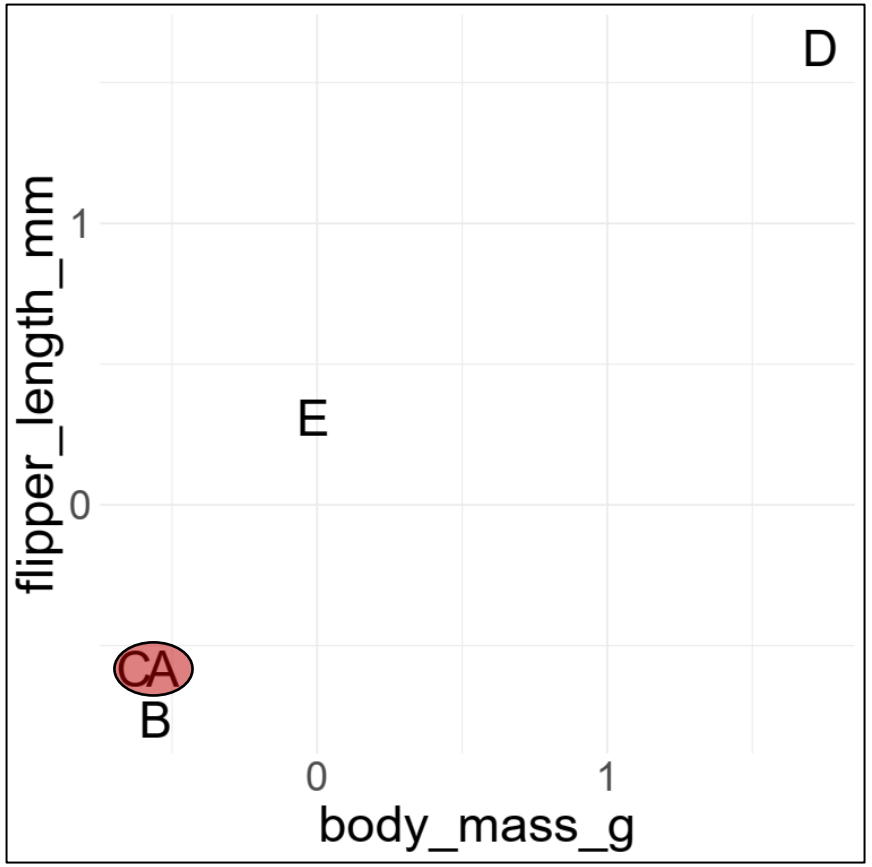
Euclidean distance



How hierarchical clustering work?

0					
	A	B	C	D	E
A	0	0.18	0.10	3.16	1.03
B		0	0.19	3.31	1.20
C			0	3.24	1.09
D				0	2.19
E					0

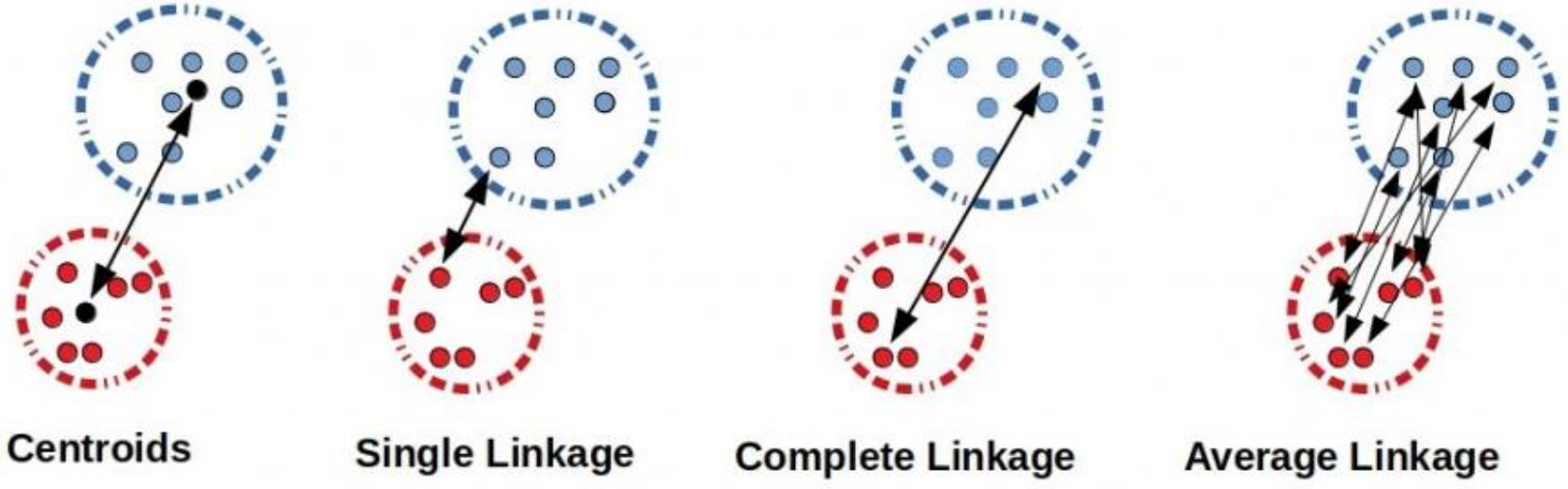
1				
	AC	B	D	E
AC	0	?	?	?
B		0	3.31	1.20
D			0	2.19
E				0



Hierarchical clustering

Distance metric depends on the linkage (dissimilarity between 2 clusters) method used to merge or split clusters:

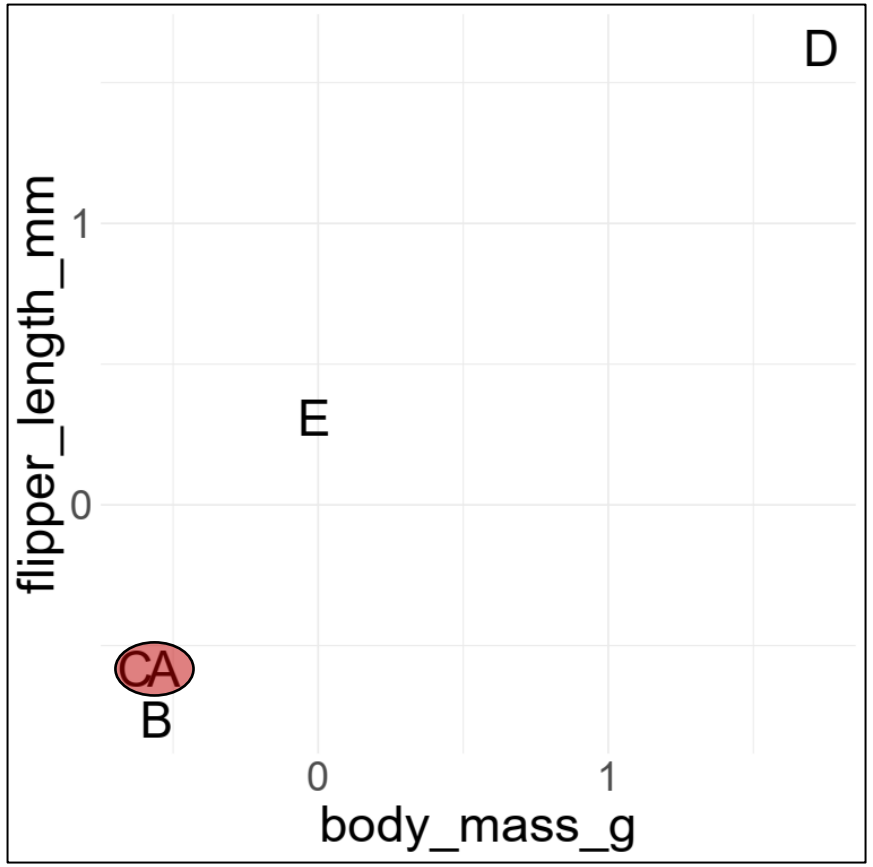
- Centroid Linkage: the distance between their mean vectors of the points in each cluster
- Single Linkage: the shortest distance between any two data points in the two clusters.
- Complete Linkage: the maximum distance between any two data points in the two clusters.
- Average Linkage: the average distance between all possible pairs of data points in the two clusters.



How hierarchical clustering work?

0					
	A	B	C	D	E
A	0	0.18	0.10	3.16	1.03
B		0	0.19	3.31	1.20
C			0	3.24	1.09
D				0	2.19
E					0

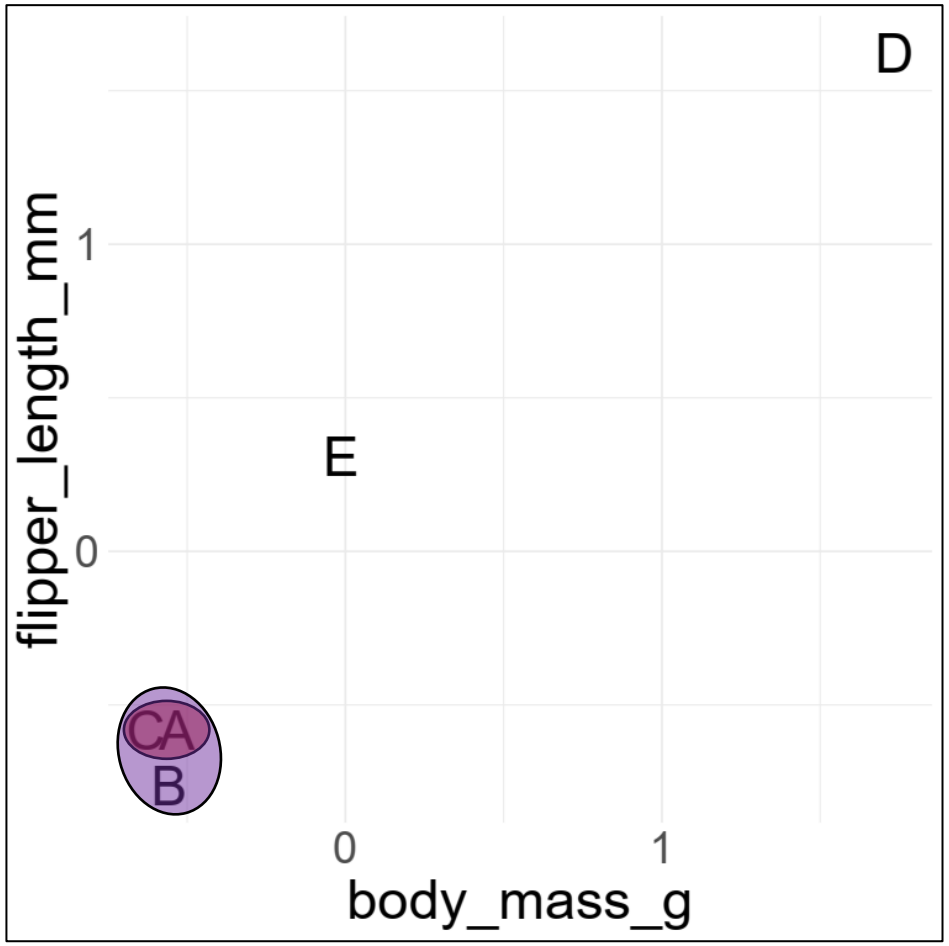
1				
	AC	B	D	E
AC	0	0.19	3.24	1.09
B		0	3.31	1.20
D			0	2.19
E				0

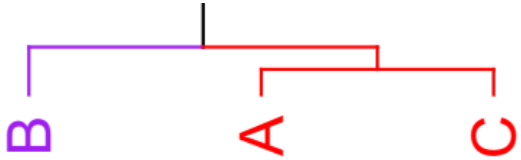


How hierarchical clustering work?

1				
	AC	B	D	E
AC	0	0.19	3.24	1.09
B		0	3.31	1.20
D			0	2.19
E				0

2			
	ACB	D	E
ACB	0	3.31	1.20
D		0	2.19
E			0

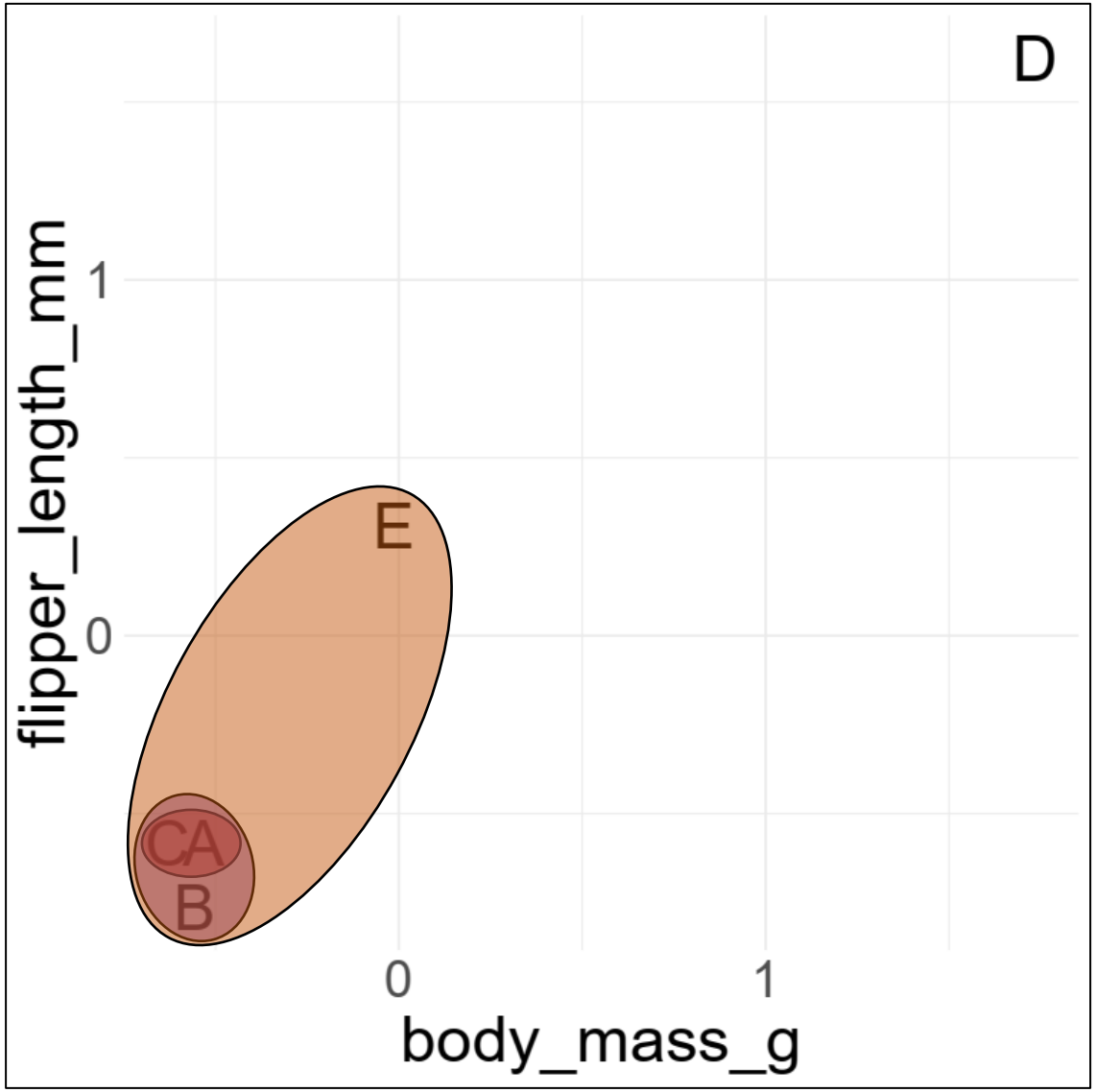


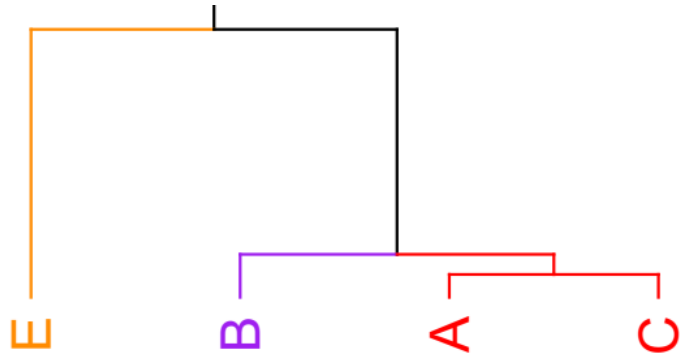


How hierarchical clustering work?

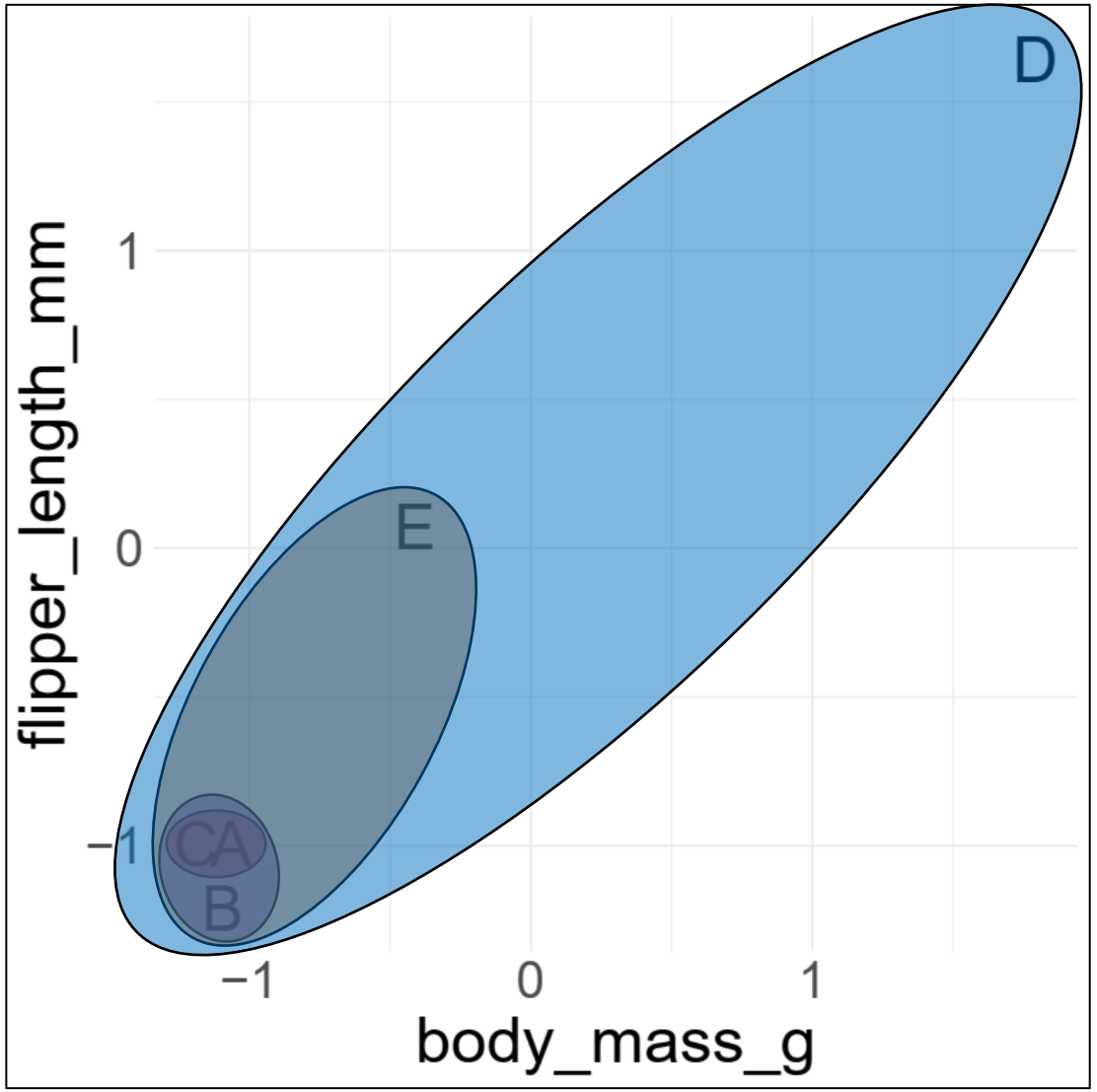
2			
	ACB	D	E
ACB	0	3.31	1.20
D		0	2.19
E			0

3		
	ACBE	D
ACBE	0	3.31
D		0

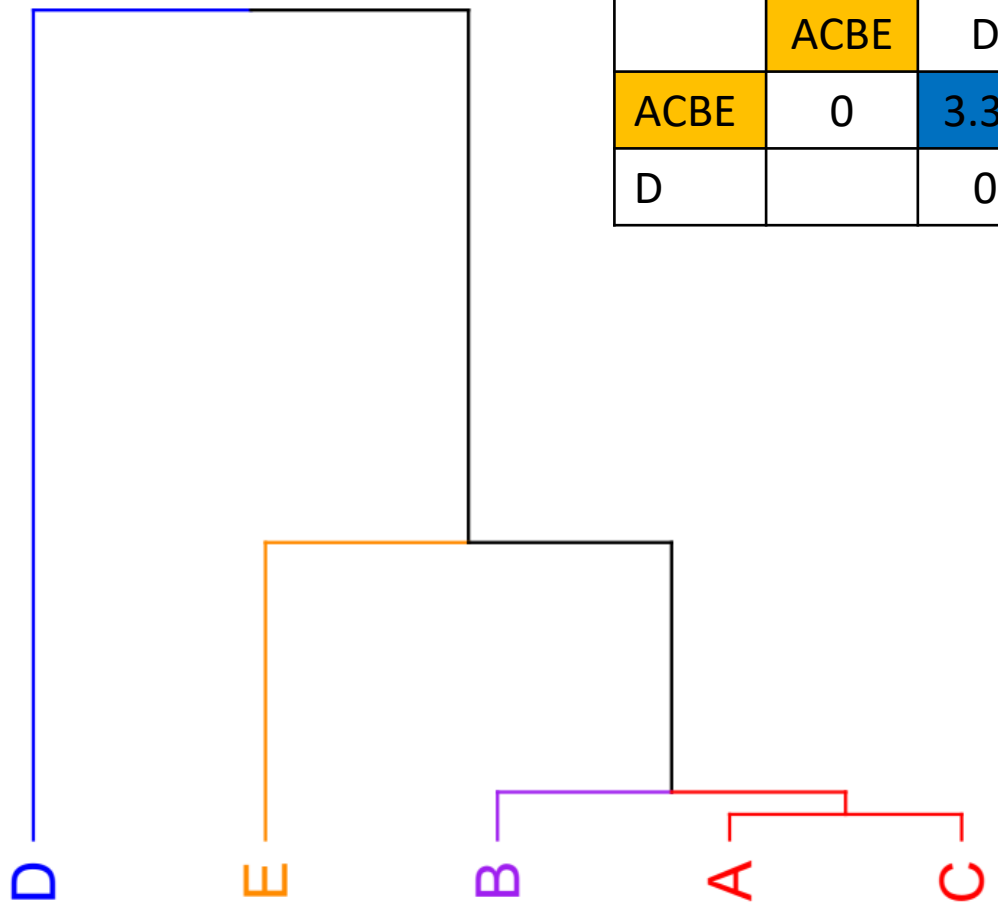




How hierarchical clustering work?



3		
	ACBE	D
ACBE	0	3.31
D		0

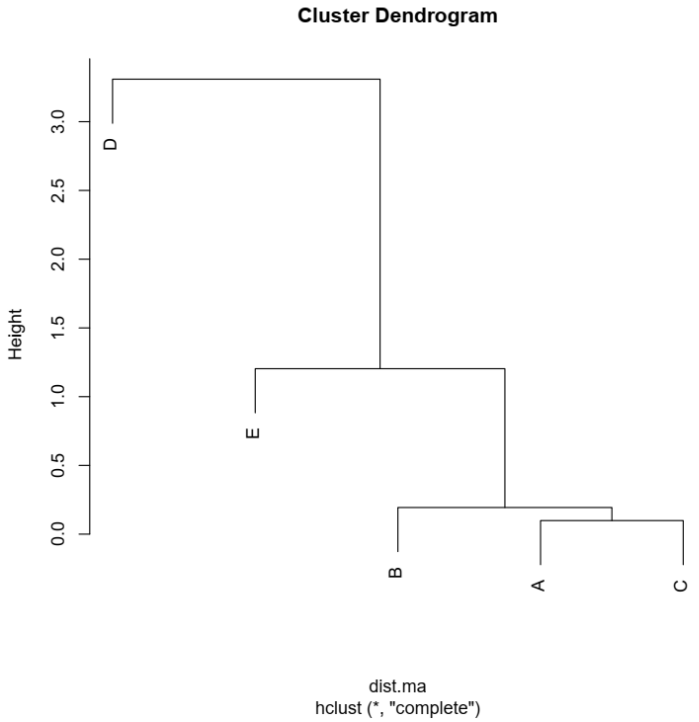


Hierarchical clustering in R

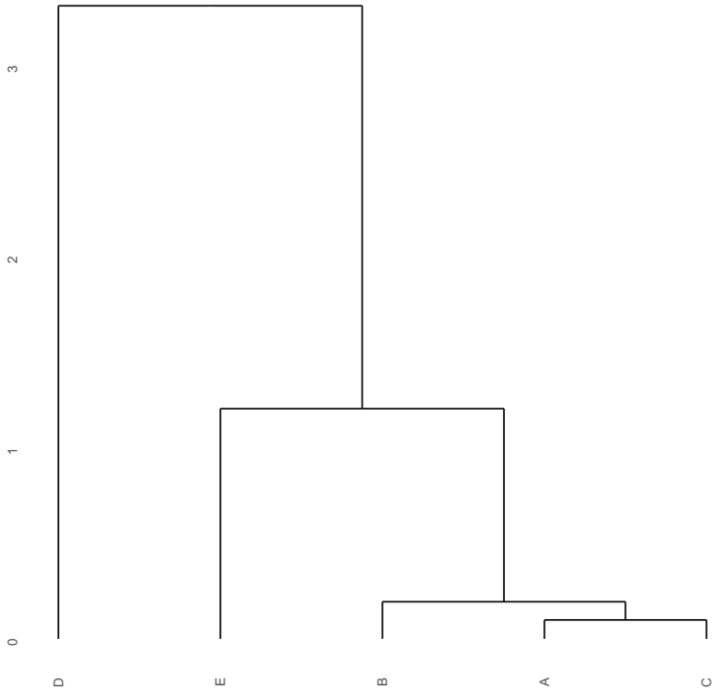
dist.ma <- dist(scale(subDT[,5:6]), method = "euclidean")

hc <- hclust(dist.ma, method="complete")

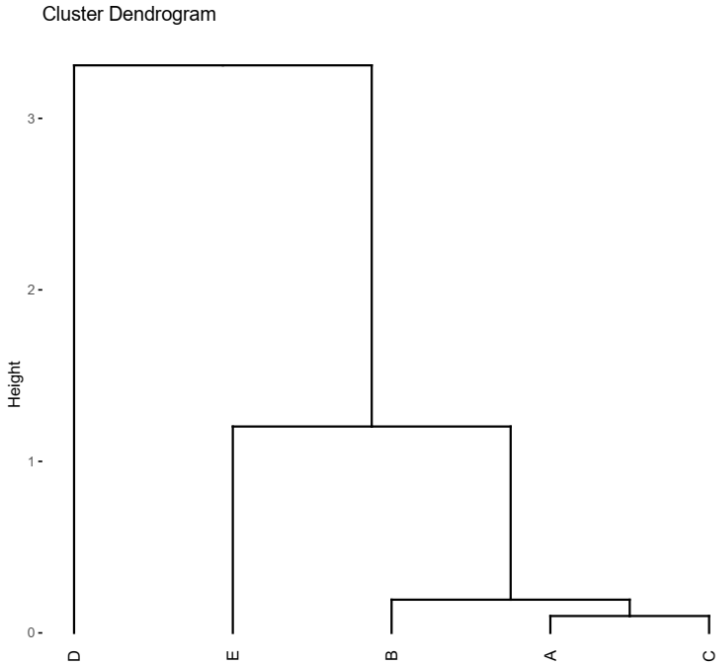
plot(hc)



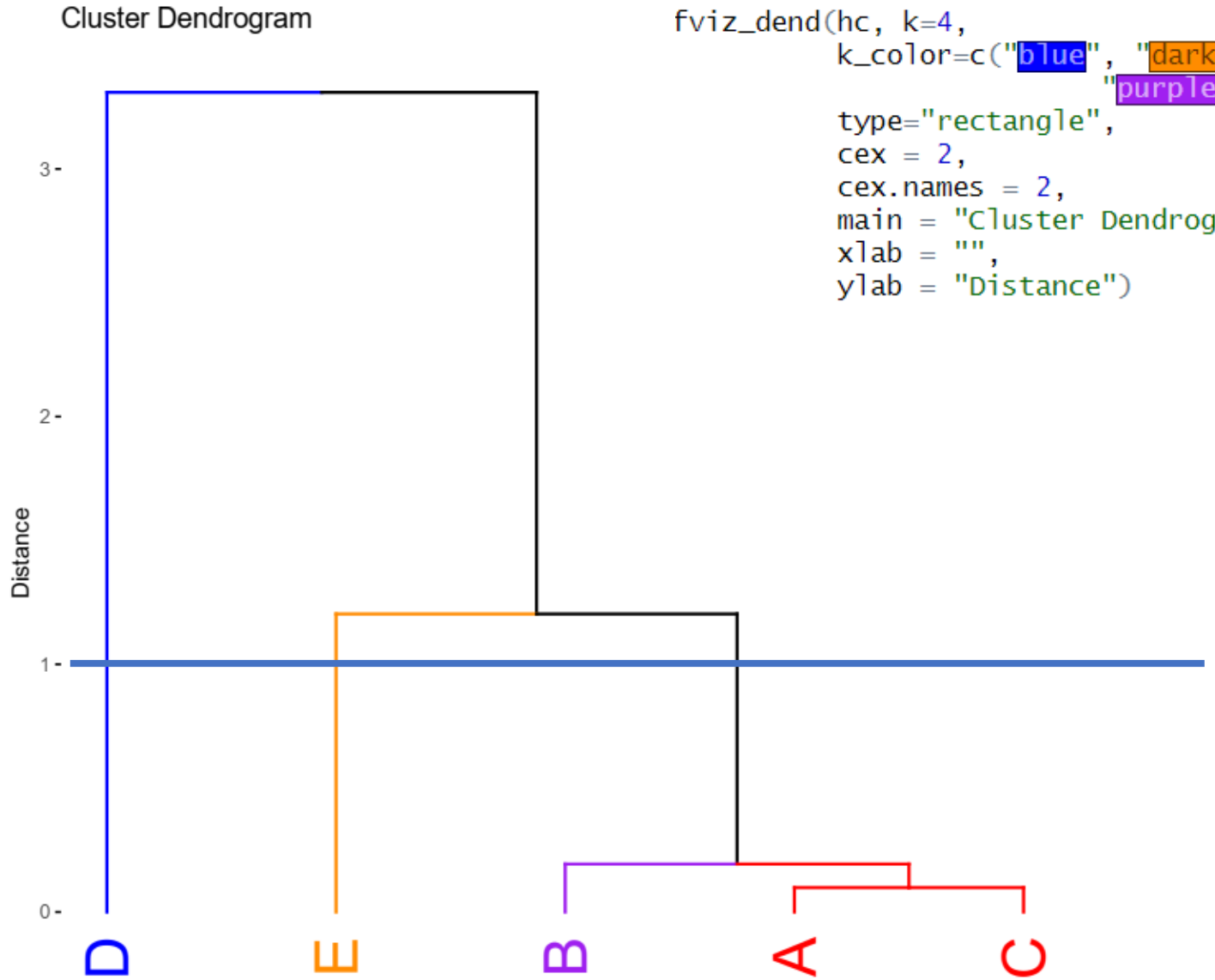
library(ggdendro)
ggdendrogram(hc)



library(factoextra)
fviz_dend(hc)



Hierarchical clustering in R

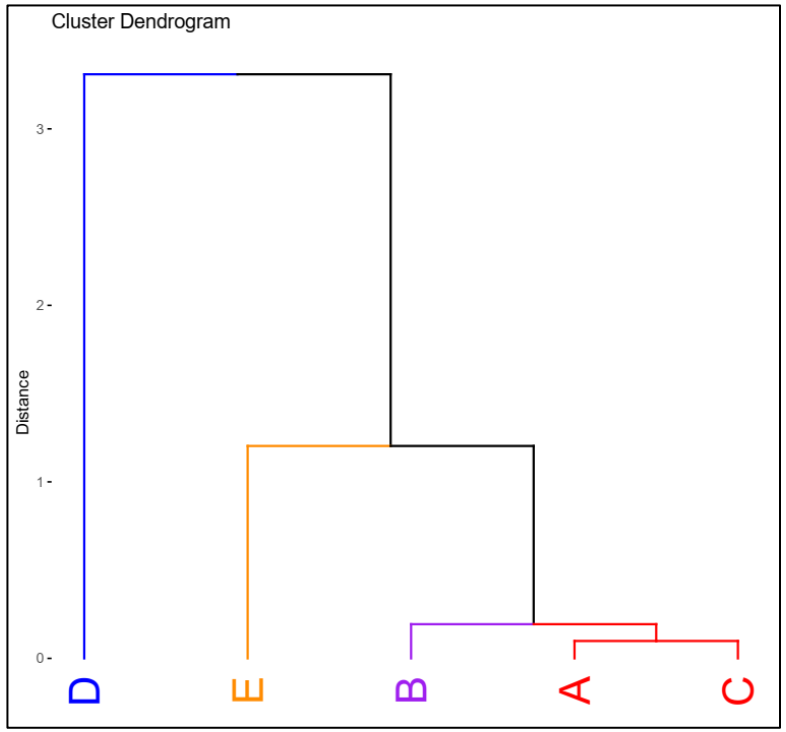



```
fviz_dend(hc, k=4,
          k_color=c("blue", "darkorange",
                    "purple", "red"),
          type="rectangle",
          cex = 2,
          cex.names = 2,
          main = "Cluster Dendrogram",
          xlab = "",
          ylab = "Distance")
```

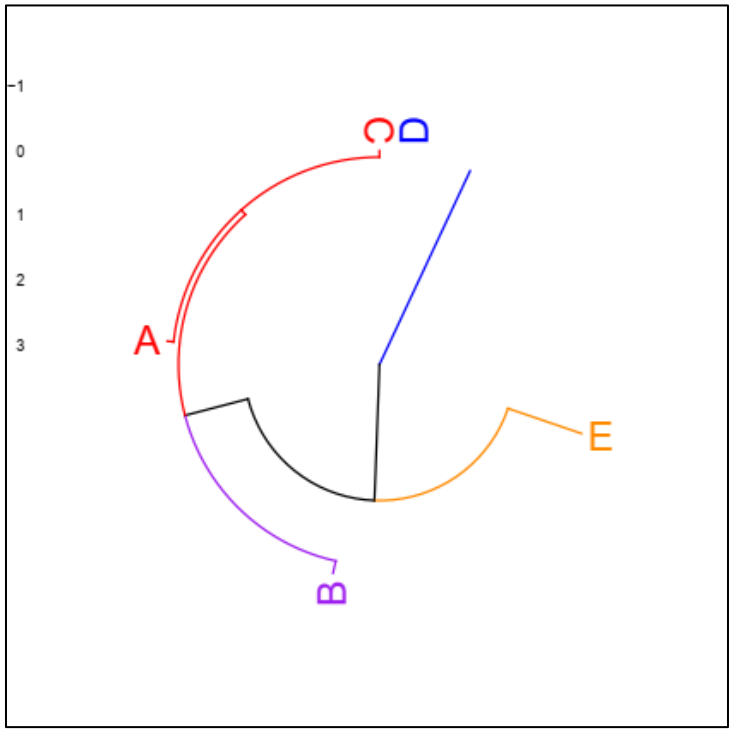
fviz_dend(hc, k=4, k_color=c("blue", "darkorange", "purple", "red"), type="rectangle", cex = 2, cex.names = 2, main = "Cluster Dendrogram", xlab = "", ylab = "Distance")

Hierarchical clustering in R

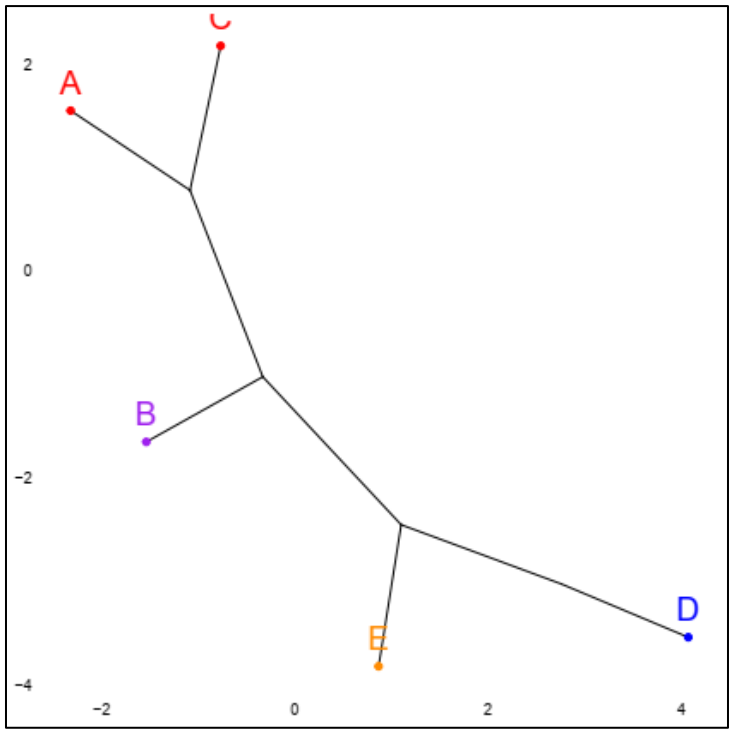
fviz_dend(type="rectangle", …)



fviz_dend(type="circular", …)



fviz_dend(type="phylogenic", …)

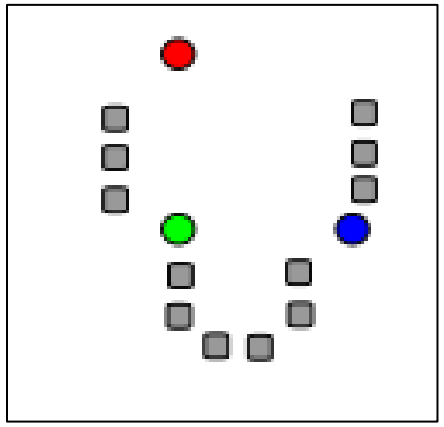


K-means clustering

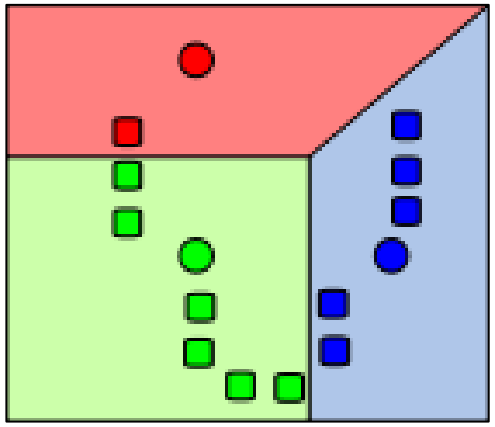
- A method of grouping data points into clusters based on their similarities.
- The "K" in K-means refers to the number of clusters that the algorithm will try to form.

Image Segmentation Using K-Means. classification of an image into… | by Aditri Srivastava | Analytics Vidhya | Medium

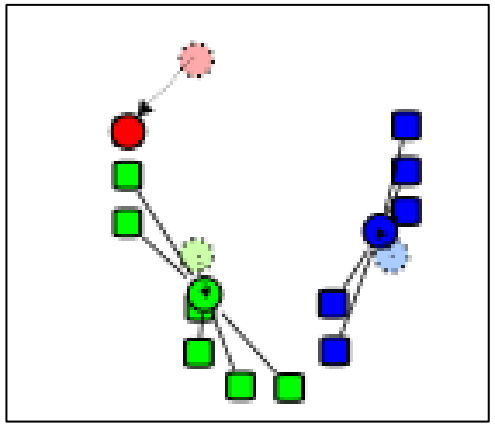
How K-means clustering work?



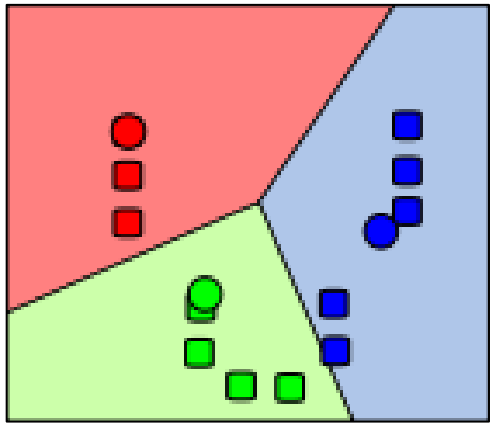
1. *k* initial "means“ are randomly generated within the data domain.



2. *k* clusters are created by associating every observation with the nearest mean.



3. The centroid of each of the *k* clusters becomes the new mean.



4. Steps 2 and 3 are repeated until convergence has been reached.

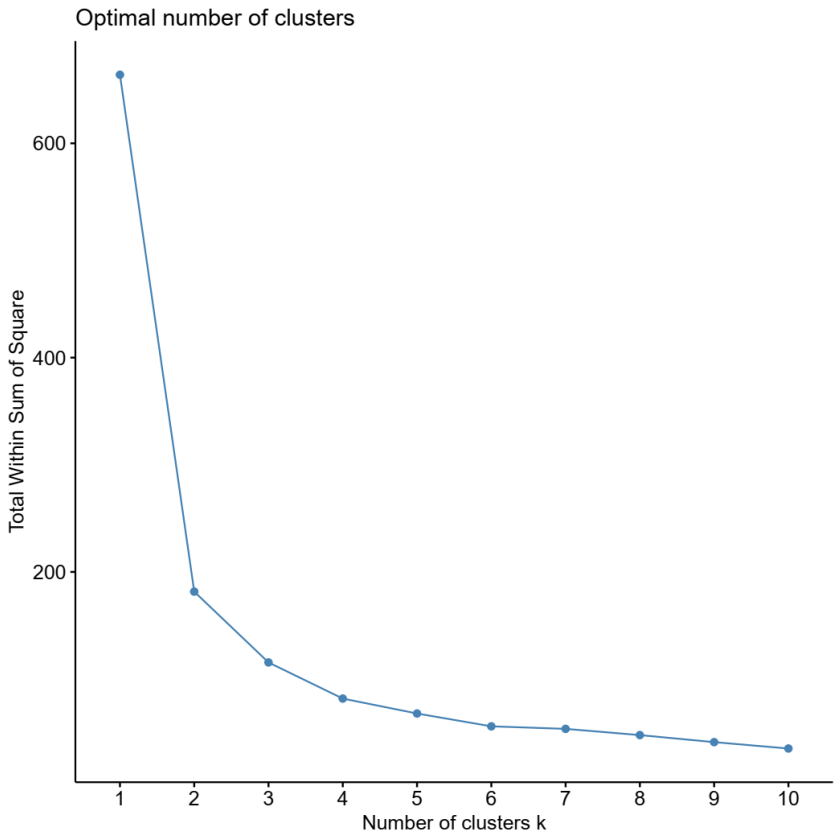
How many k is optimal?

Common methods:

- Elbow method
- Average silhouette method
- Gap statistic method

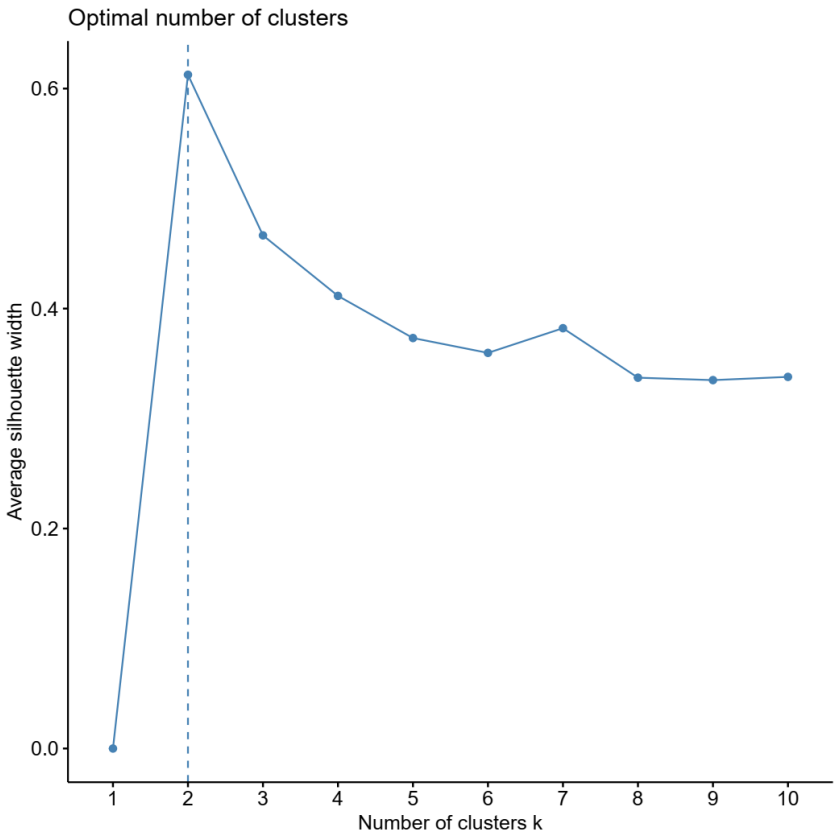
How many k is optimal?

Elbow method



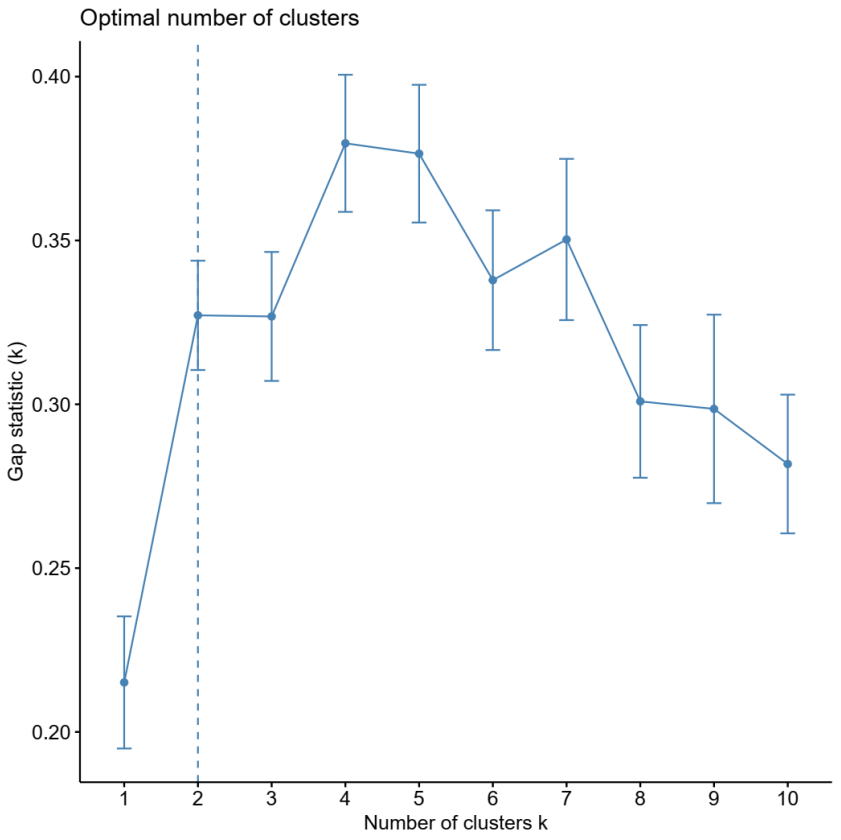
fviz_nbclust(scale(dt[,5:6]), kmeans, method="wss")

Average silhouette method



fviz_nbclust(scale(dt[,5:6]), kmeans, method="silhouette")

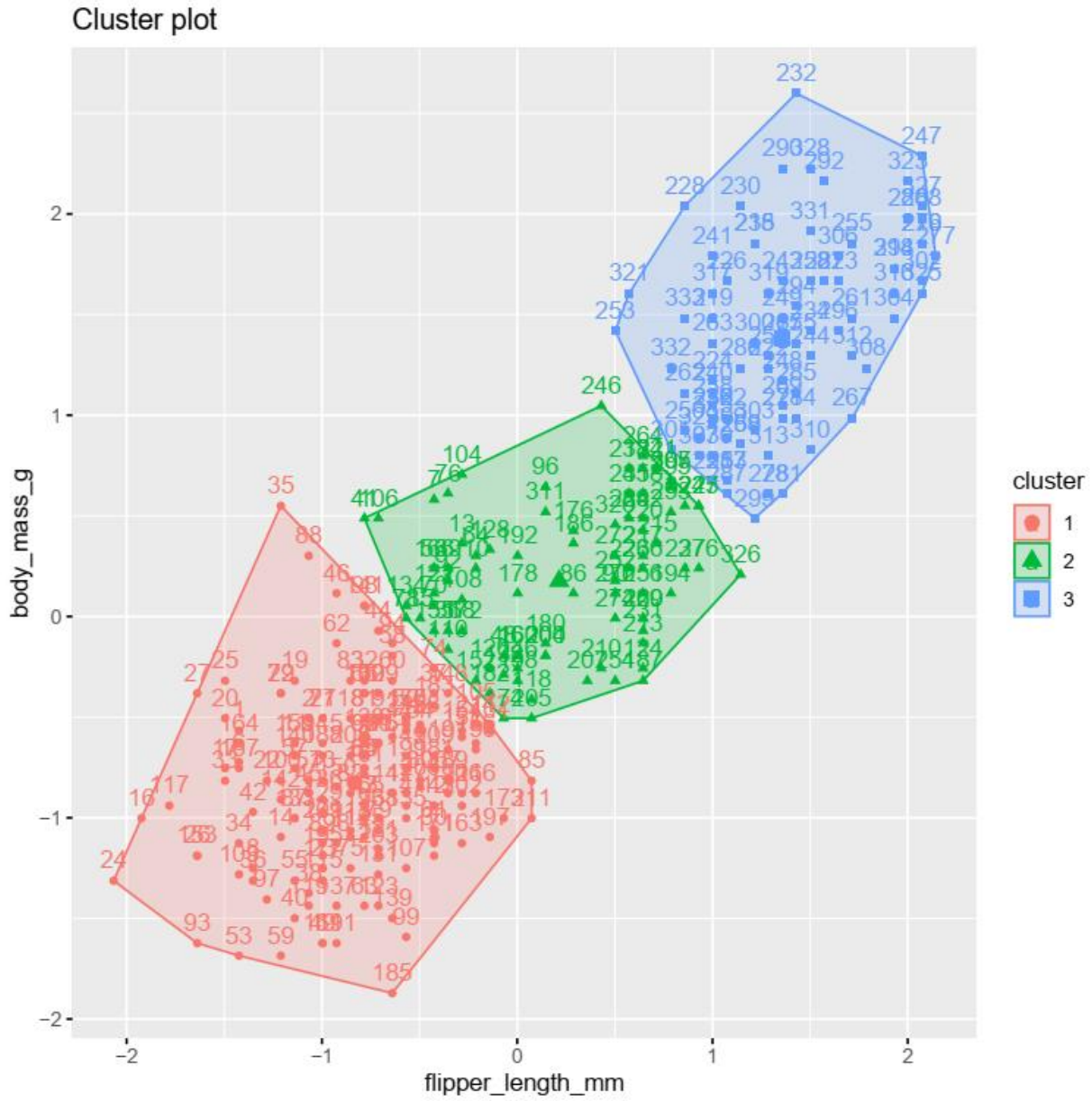
Gap statistic method



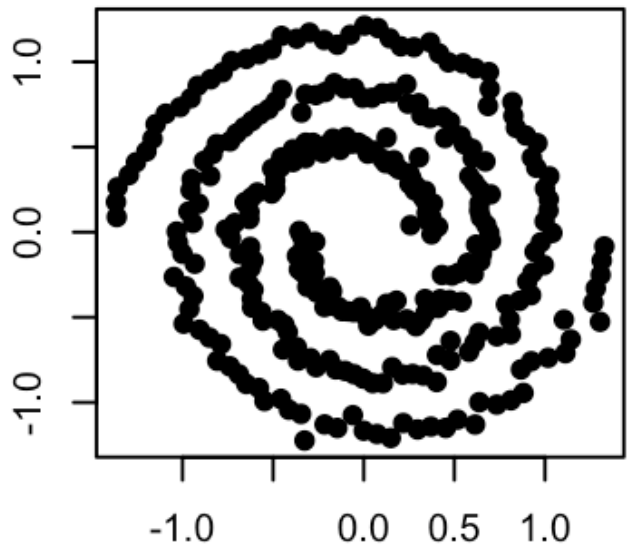
fviz_nbclust(scale(dt[,5:6]), kmeans, method="gap_stat")

(99) ML26: unsupervied clustering (K-mean, Hierarchical clustering) - YouTube

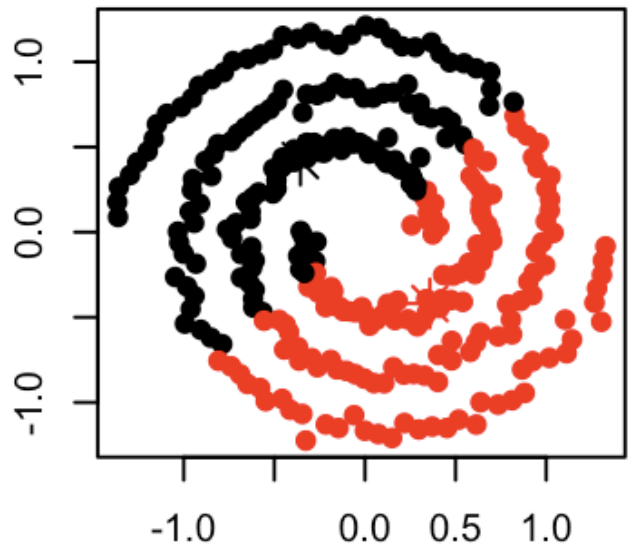
km = kmeans(scale(dt[,5:6]), center=2, nstart=20)
fviz_cluster(km, scale(dt[,5:6]), ellipse.type="convex")



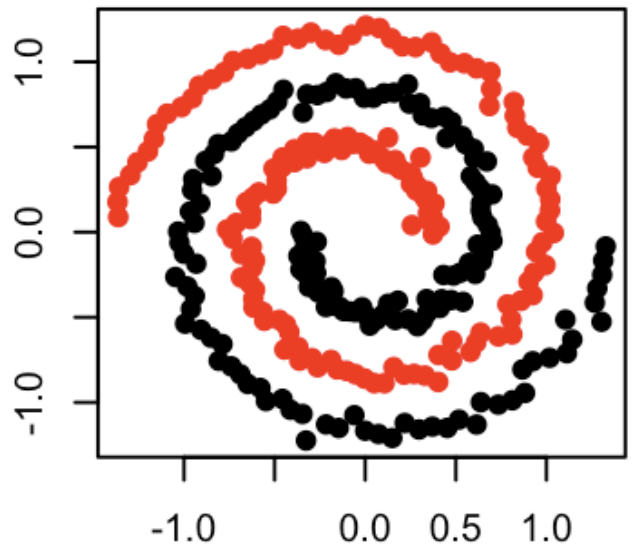
km3 = kmeans(scale(dt[,5:6]), center=3, nstart=20)
fviz_cluster(km3, scale(dt[,5:6]), ellipse.type="convex")



K-means



Spectral clustering



Summary

Name	Function in R	You want to:
Scaling	scale()	
Distance Matrix Computation	dist()	
Visualization of Dendrogram	fviz_dend()	
Hierarchical clustering	hclust()	View at once the clusterings obtained for each possible number of clusters, from 1 to n
Visualizing the Optimal Number of Clusters	fviz_nbclust()	
K-means clustering	kmeans()	Partition the observations into a pre-specified number of clusters
Visualize Clustering Results	fviz_cluster()	

```
library(palmerpenguins)
library(datasets)
library(ComplexHeatmap)
library(circlize)
library(factoextra)
library(ggdendro)
library(ggplot2)
library(ggpubr)
library(ggrepel)
library(tidyverse)
library(NbClust)
library(cluster)
```

Homework

Download the BetaMatrix.tsv data from: https://drive.google.com/file/d/1tOdeLpEzhEcsDPU6Vz_dIQsV0UZy0bz0/view?usp=share_link

Base on value of 200 CpG sites, do the following requests:

1. Draw dendrogram.
2. How many optimal *k* clusters does each method (Elbow, average silhouette, gap statistic method) tell you?
3. Draw clustering results from K-means clustering algorithm.

Thanks for listening!
Have a nice week!