MONASH University

ETC3250/5250: Introduction to Machine Learning

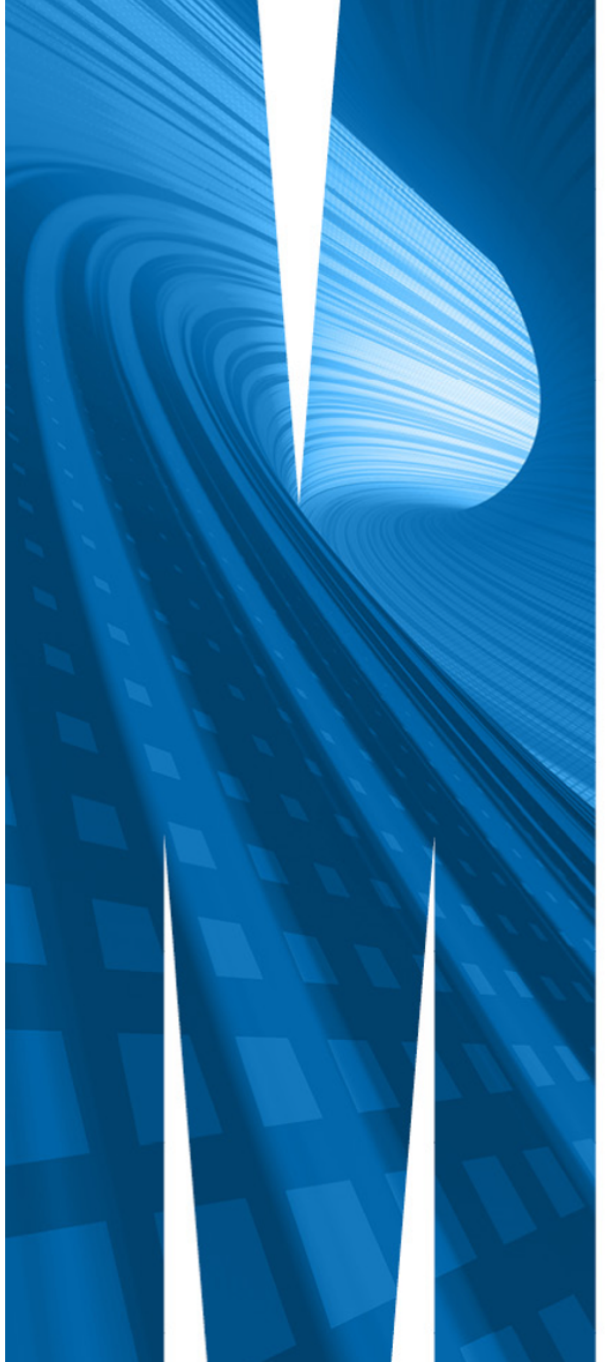
k-means clustering

Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

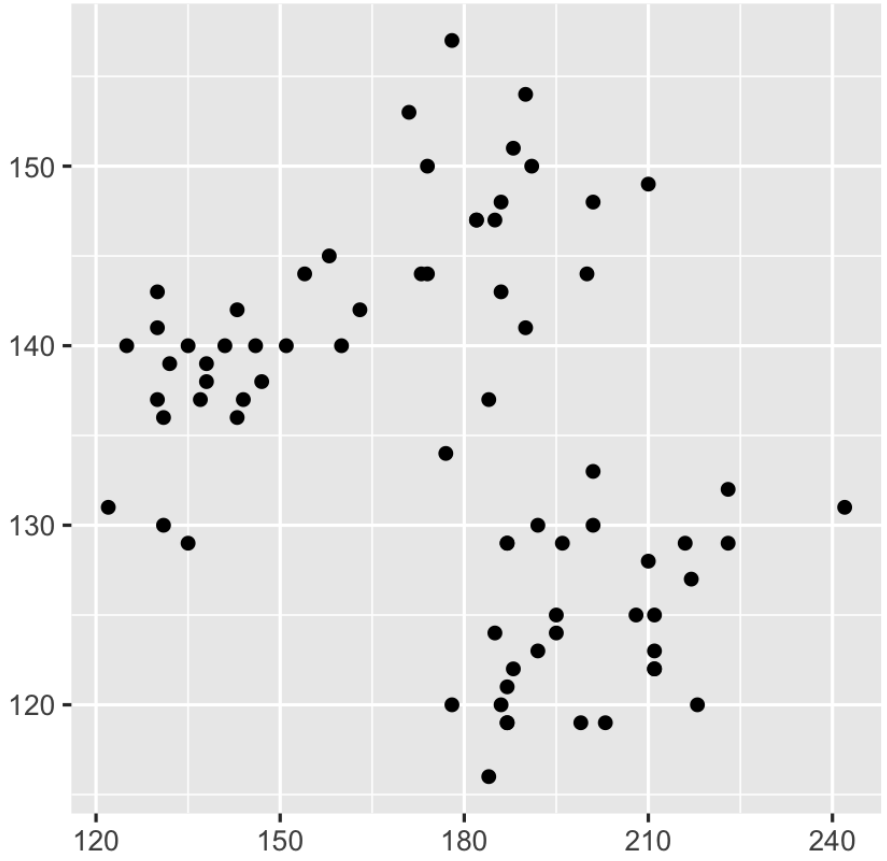
Week 10a



Cluster analysis

- The aim of cluster analysis is to group cases (objects) according to their similarity on the variables. It is also often called unsupervised classification, meaning that classification is the ultimate goal, but the classes (groups) are not known ahead of time.
- Hence the first task in cluster analysis is to construct the class information. To determine closeness we start with measuring the interpoint distances.

Cluster this!



k-means clustering - algorithm

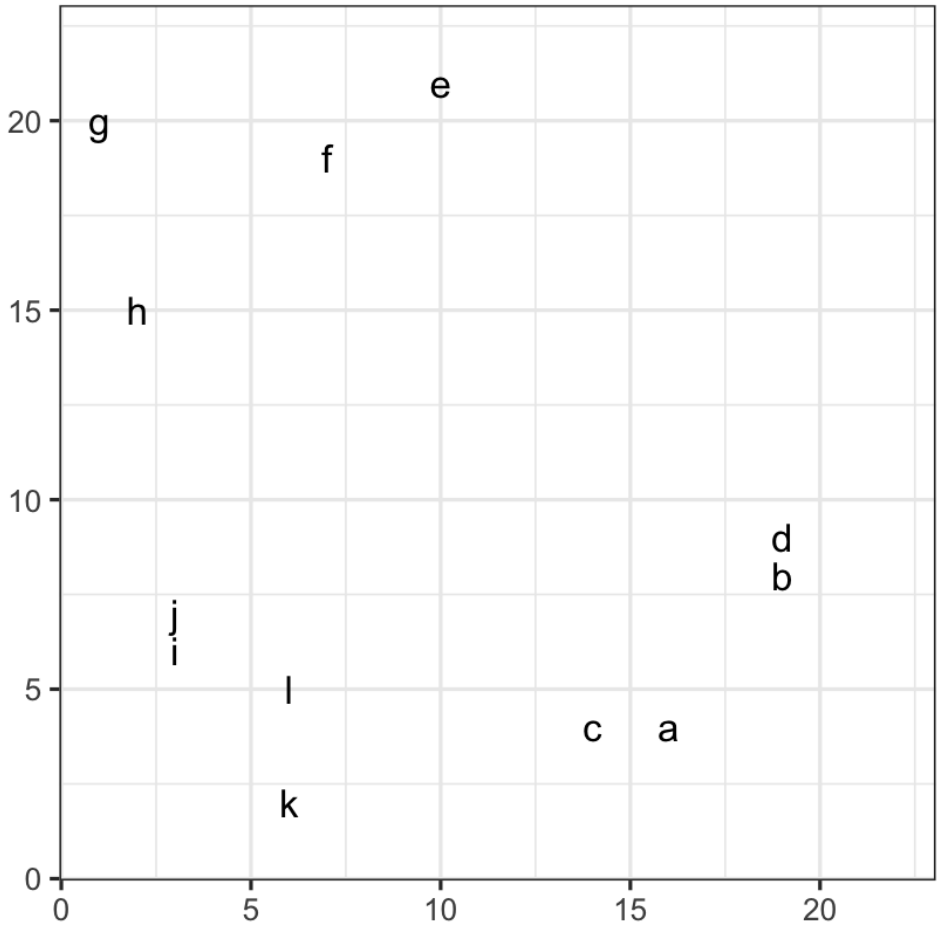
This is an iterative procedure. To use it the number of clusters, k must be decided first. The stages of the iteration are:

- Initialize by either (a) partitioning the data into k groups, and compute the k group means or (b) an initial set of k points as the first estimate of the cluster means (seed points).
- Loop over all observations reassigning them to the group with the closest mean.
- Recompute group means.
- Iterate steps 2 and 3 until convergence.

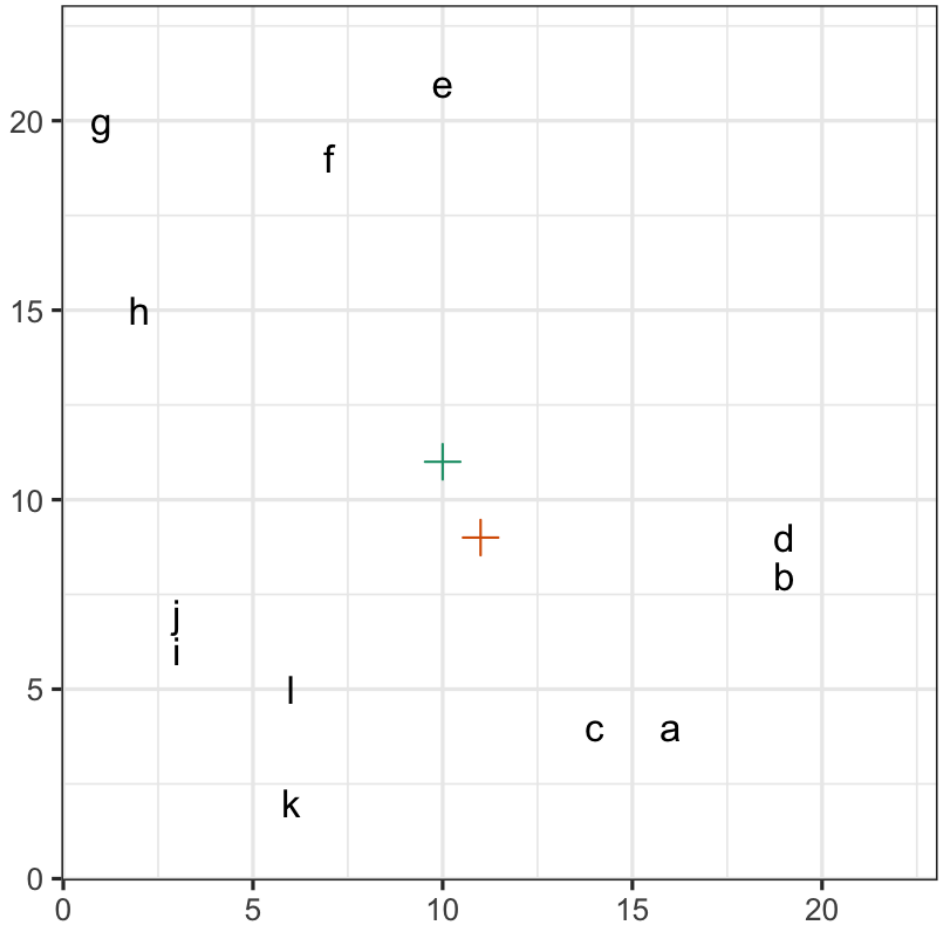
Thean C. Lim's blog post

Some data 🦒

lbl	x1	x2
a	16	4
b	19	8
c	14	4
d	19	9
e	10	21
f	7	19
g	1	20
h	2	15
i	3	6
j	3	7
k	6	2
l	6	5



Select $k = 2$ set initial seed means $x_1 = (10, 11)$, $x_2 = (11, 9)$



Compute distances between each observation and each mean.

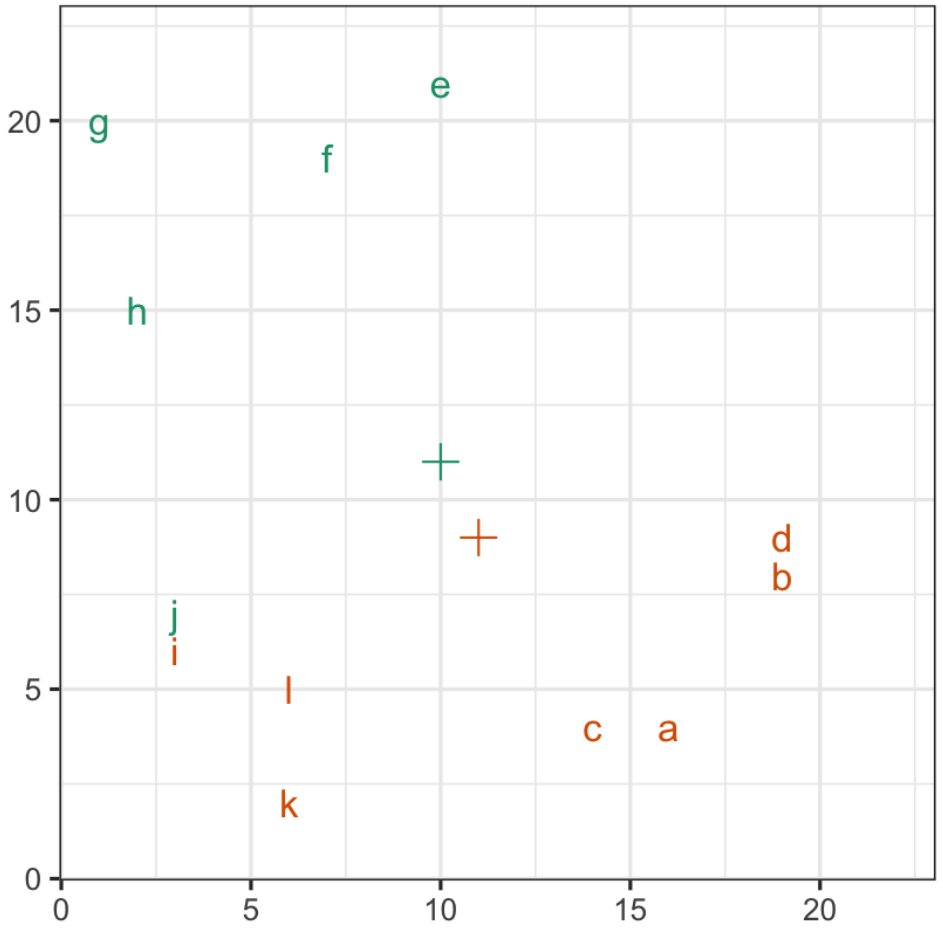
lbl	x1	x2	d1	d2
a	16	4	9.2	7.1
b	19	8	9.5	8.1
c	14	4	8.1	5.8
d	19	9	9.2	8.0
e	10	21	10.0	12.0
f	7	19	8.5	10.8
g	1	20	12.7	14.9
h	2	15	8.9	10.8
i	3	6	8.6	8.5
j	3	7	8.1	8.2
k	6	2	9.8	8.6
l	6	5	7.2	6.4

Assign the cluster membership

lbl	x1	x2	d1	d2	cl
a	16	4	9.2	7.1	2
b	19	8	9.5	8.1	2
c	14	4	8.1	5.8	2
d	19	9	9.2	8.0	2
e	10	21	10.0	12.0	1
f	7	19	8.5	10.8	1
g	1	20	12.7	14.9	1
h	2	15	8.9	10.8	1
i	3	6	8.6	8.5	2
j	3	7	8.1	8.2	1
k	6	2	9.8	8.6	2
l	6	5	7.2	6.4	2

Assign the cluster membership

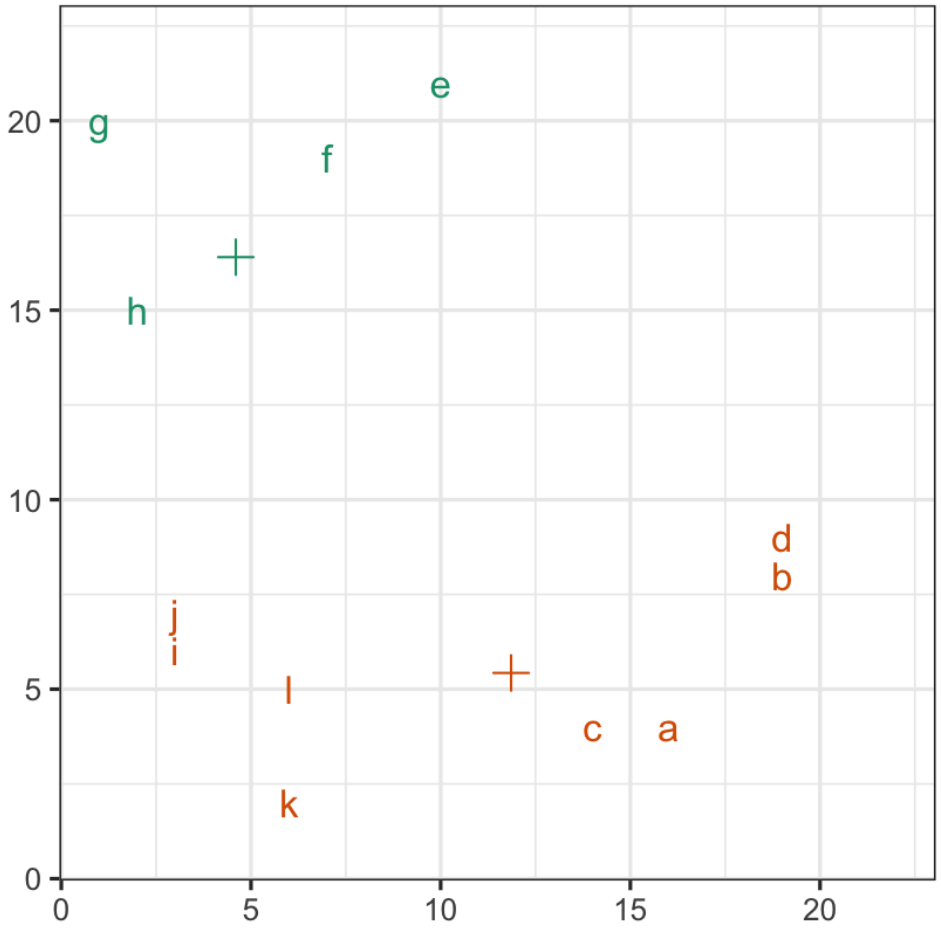
lbl	x1	x2	d1	d2	cl
a	16	4	9.2	7.1	2
b	19	8	9.5	8.1	2
c	14	4	8.1	5.8	2
d	19	9	9.2	8.0	2
e	10	21	10.0	12.0	1
f	7	19	8.5	10.8	1
g	1	20	12.7	14.9	1
h	2	15	8.9	10.8	1
i	3	6	8.6	8.5	2
j	3	7	8.1	8.2	1
k	6	2	9.8	8.6	2
l	6	5	7.2	6.4	2



Recompute means, and re-assign the cluster membership

$x_1(5, 16)$, $x_2(12, 5)$

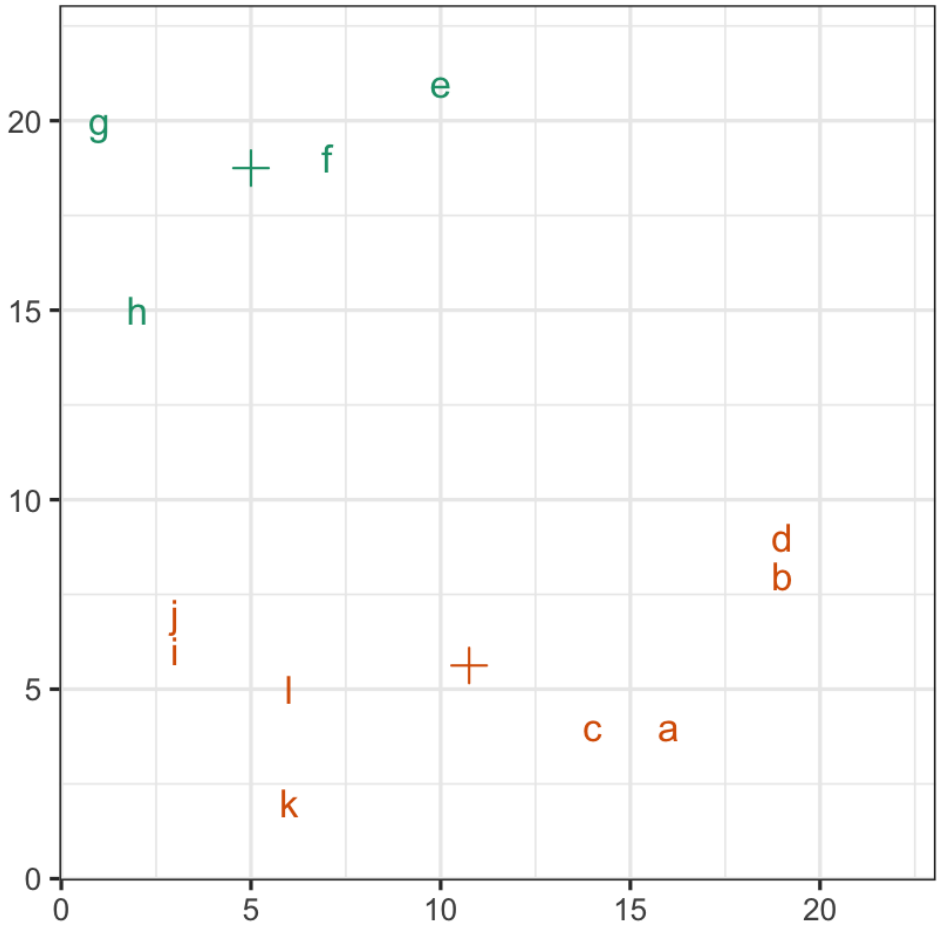
lbl	x1	x2	d1	d2	cl
a	16	4	16.8	4.4	2
b	19	8	16.7	7.6	2
c	14	4	15.6	2.6	2
d	19	9	16.2	8.0	2
e	10	21	7.1	15.7	1
f	7	19	3.5	14.4	1
g	1	20	5.1	18.2	1
h	2	15	3.0	13.7	1
i	3	6	10.5	8.9	2
j	3	7	9.5	9.0	2
k	6	2	14.5	6.8	2



Recompute means, and re-assign the cluster membership

$x_1(5, 19)$, $x_2(11, 6)$

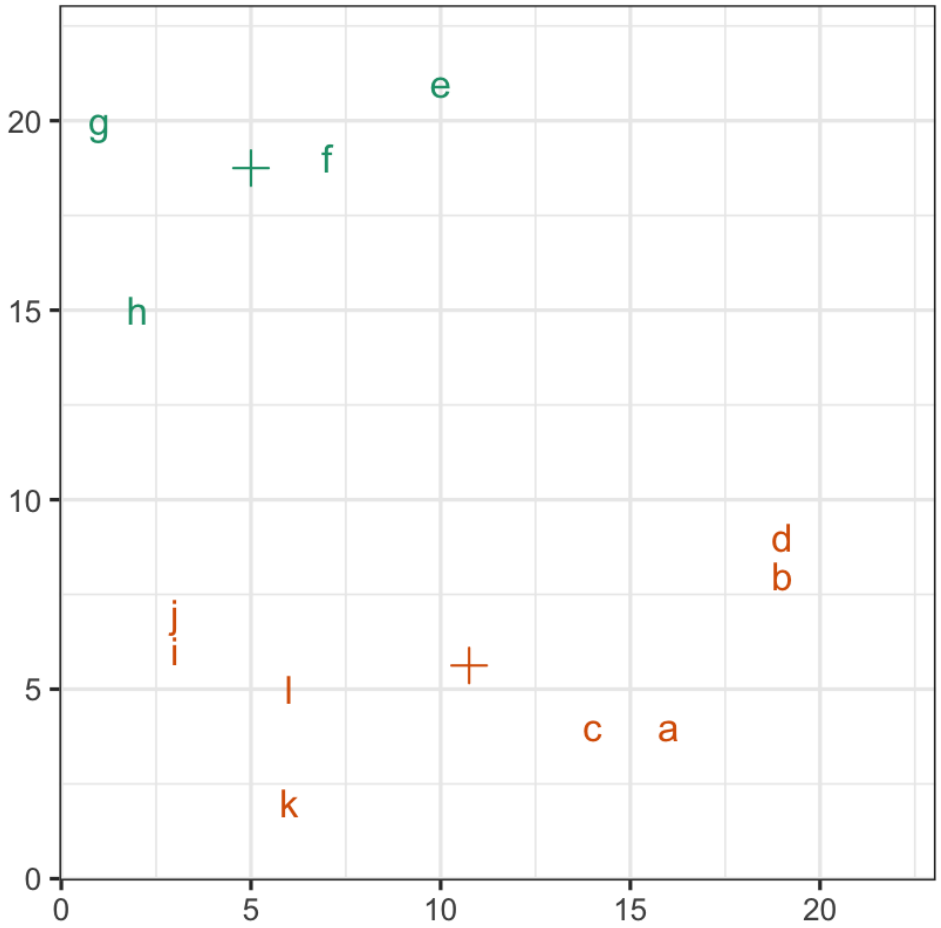
lbl	x1	x2	d1	d2	cl
a	16	4	18.4	5.5	2
b	19	8	17.7	8.6	2
c	14	4	17.3	3.6	2
d	19	9	17.1	8.9	2
e	10	21	5.5	15.4	1
f	7	19	2.0	13.9	1
g	1	20	4.2	17.4	1
h	2	15	4.8	12.8	1
i	3	6	12.9	7.8	2
j	3	7	11.9	7.9	2
k	6	2	16.8	6.0	2



Recompute means, and re-assign the cluster membership

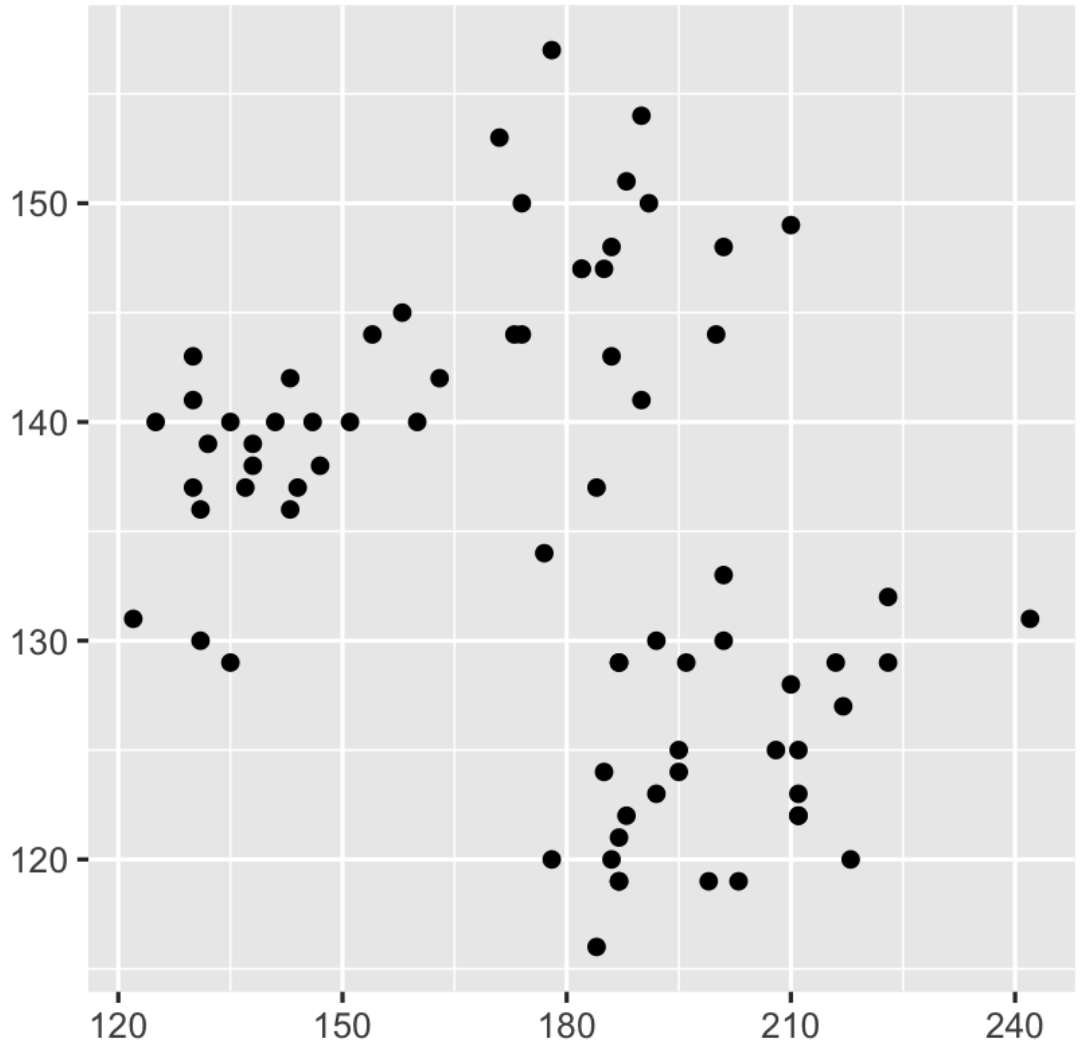
$x_1(5, 19)$, $x_2(11, 6)$

lbl	x1	x2	d1	d2	cl
a	16	4	18.4	5.5	2
b	19	8	17.7	8.6	2
c	14	4	17.3	3.6	2
d	19	9	17.1	8.9	2
e	10	21	5.5	15.4	1
f	7	19	2.0	13.9	1
g	1	20	4.2	17.4	1
h	2	15	4.8	12.8	1
i	3	6	12.9	7.8	2
j	3	7	11.9	7.9	2
k	6	2	16.8	6.0	2

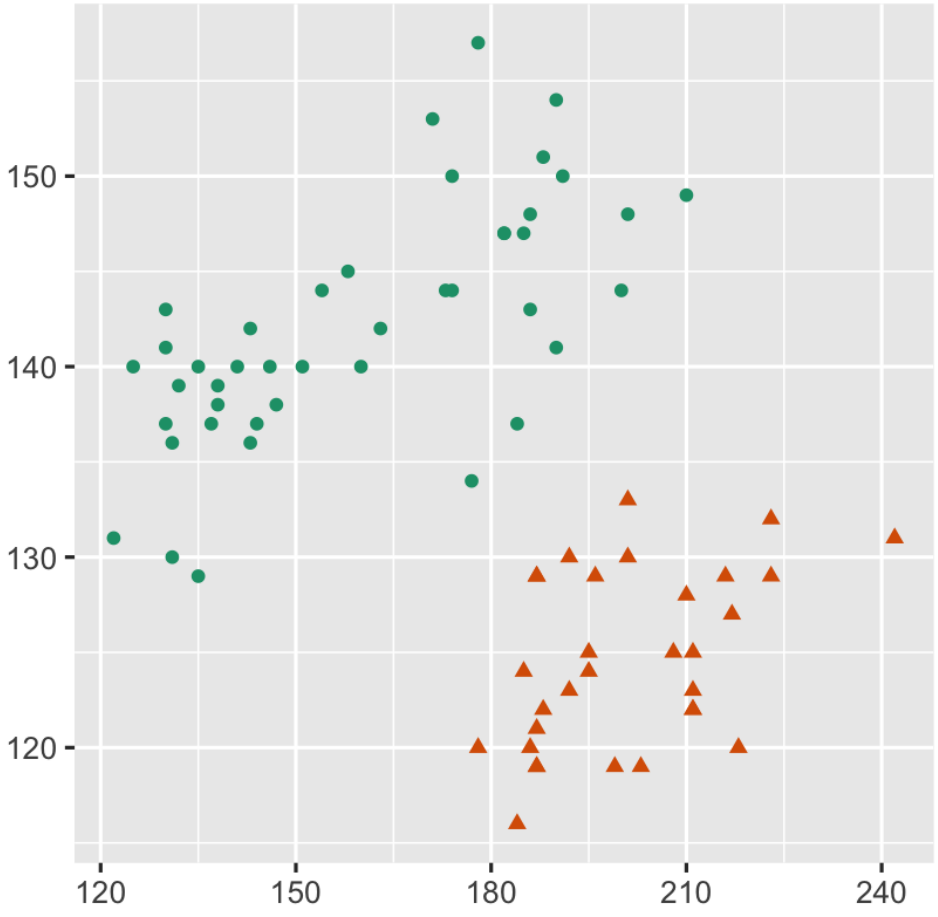


Watch it animate

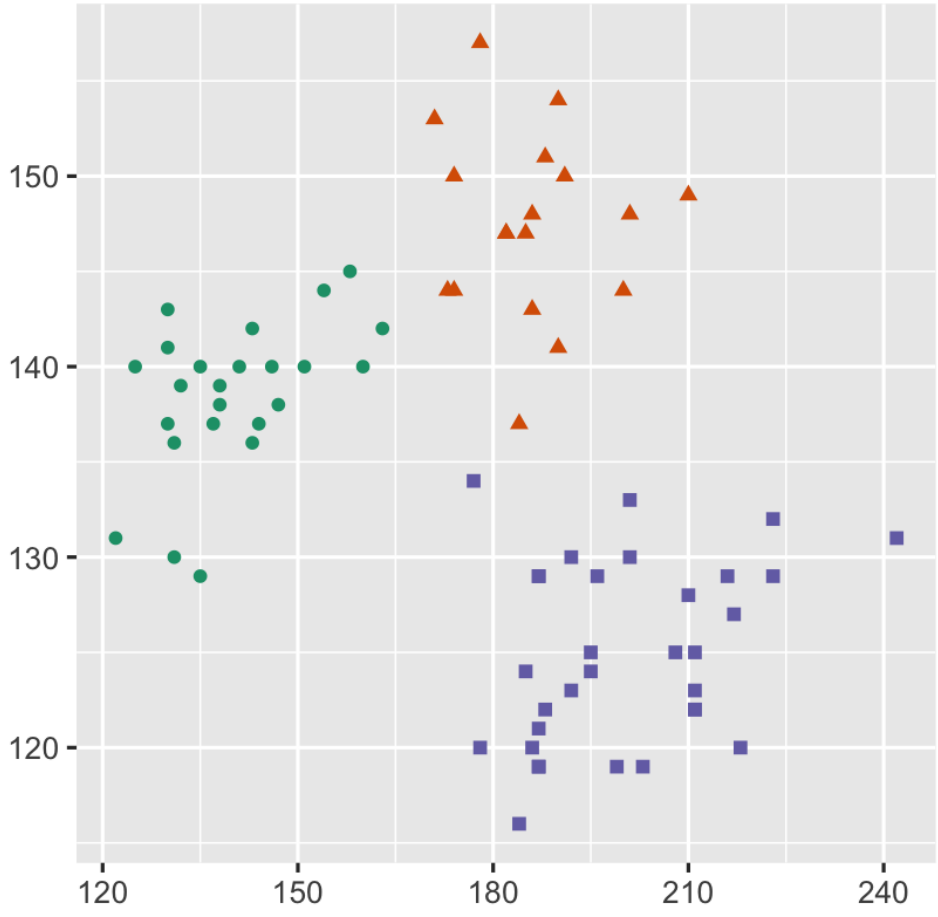
Example



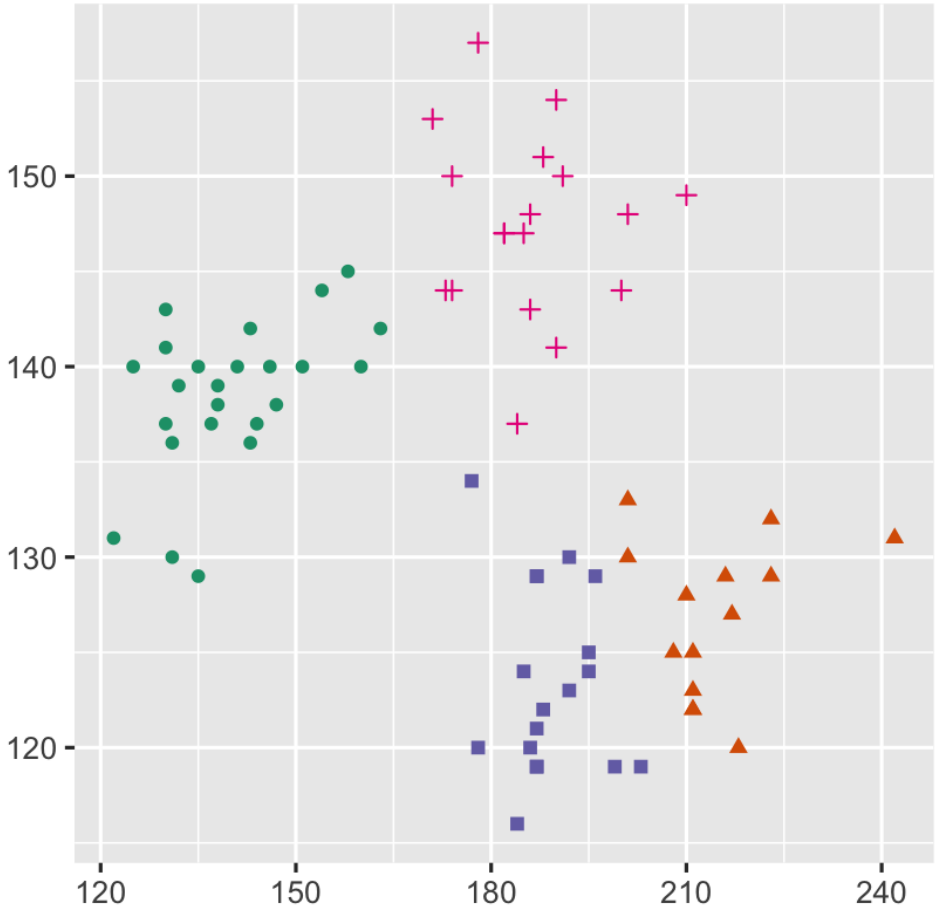
$k = 2$


150
140
130
120
120
150
180
210
240

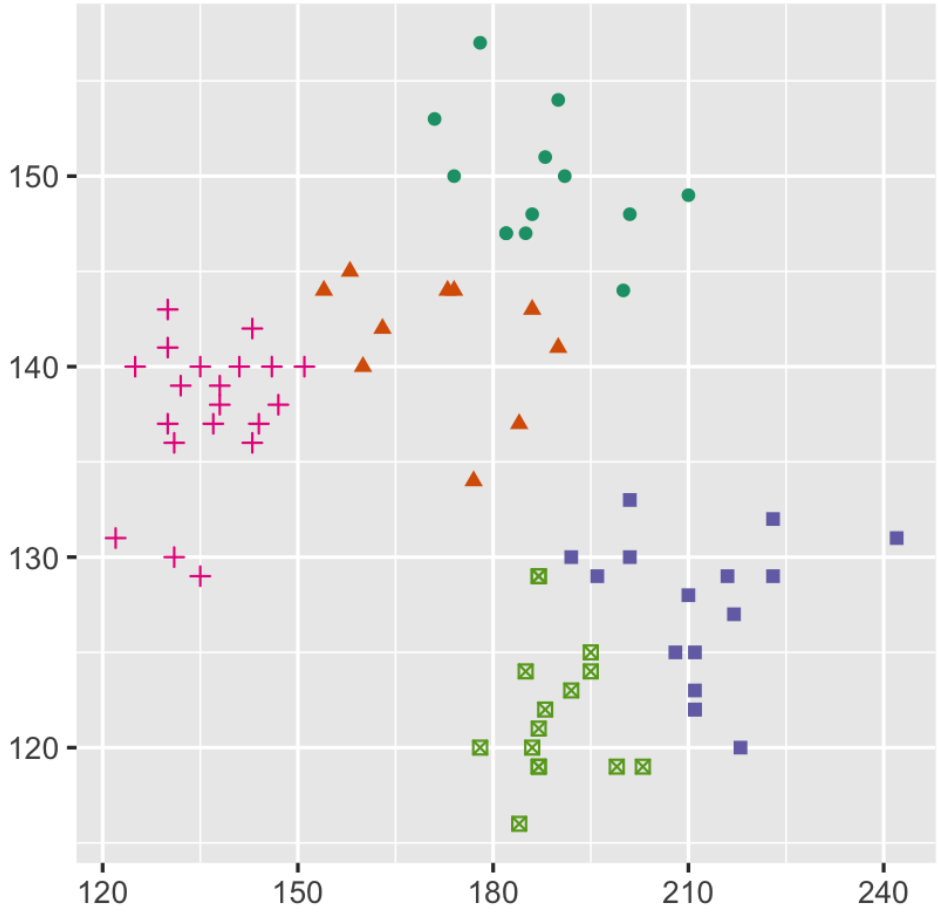
$k = 3$


150
140
130
120
120
150
180
210
240

$k = 4$


150
140
130
120
120
150
180
210
240

$k = 5$


150
140
130
120
120
150
180
210
240

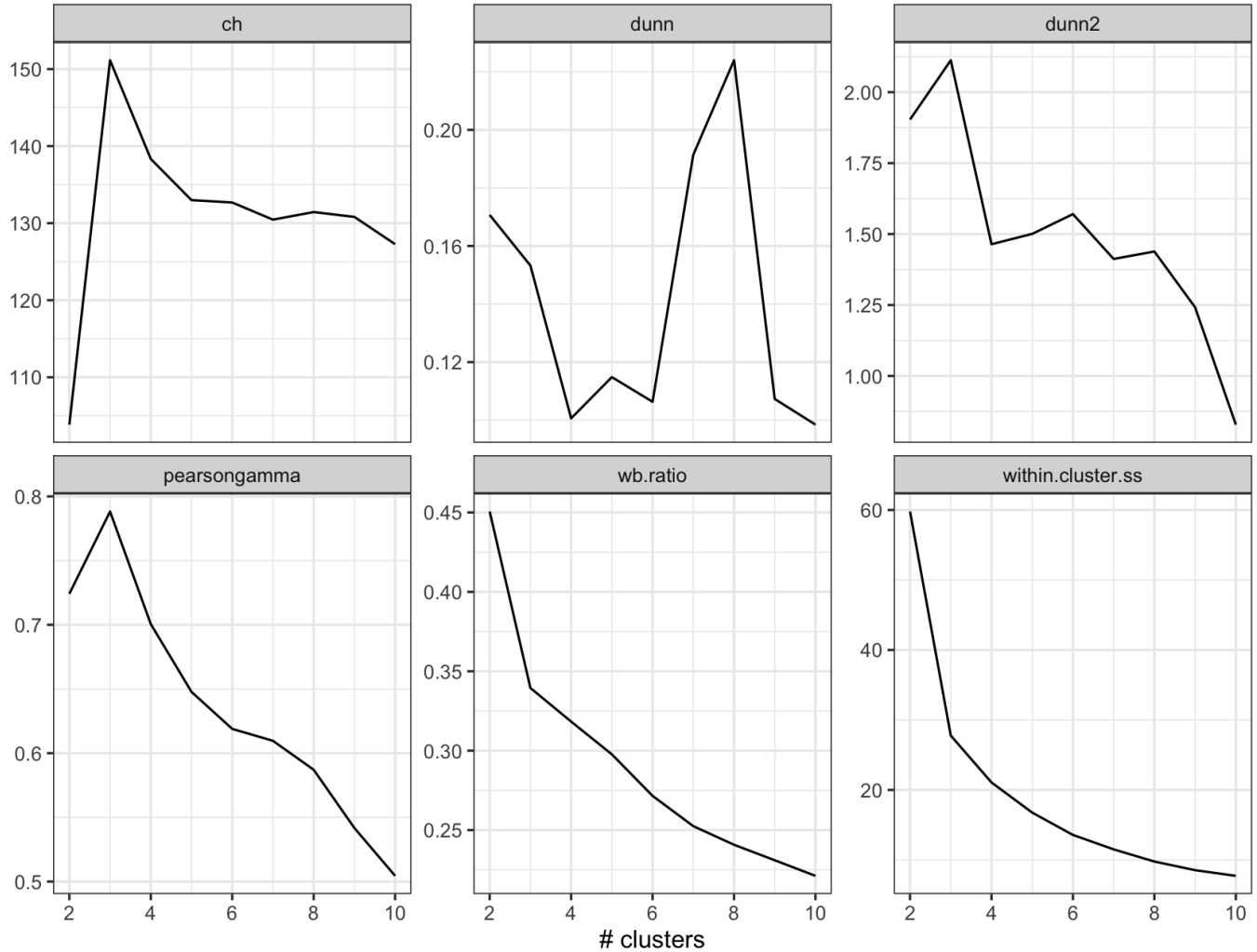
Choosing k

Cluster statistics

- WBRatio: average within/average between want it to be low, but always drops for each additional cluster so look for large drops
- Hubert Gamma: (s+ - s-)/(s+ + s-) where ~~sum~~ of number of within < between, ~~sum~~ of number within > between, want this to be high
- Dunn: smallest distance between points from different clusters/maximum distance of points within any cluster, want this to be high
- Calinski-Harabasz Index: ~~Want this to~~ be high

$$\frac{\sum_{i=1}^{p} B_{ii}/(k-1)}{\sum_{i=1}^{p} W_{ii}/(n-k)}$$

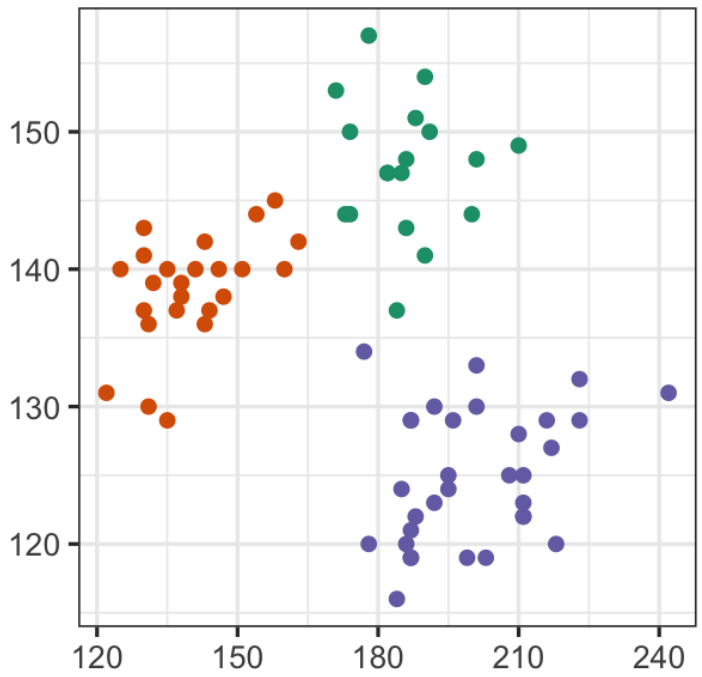
Choosing k

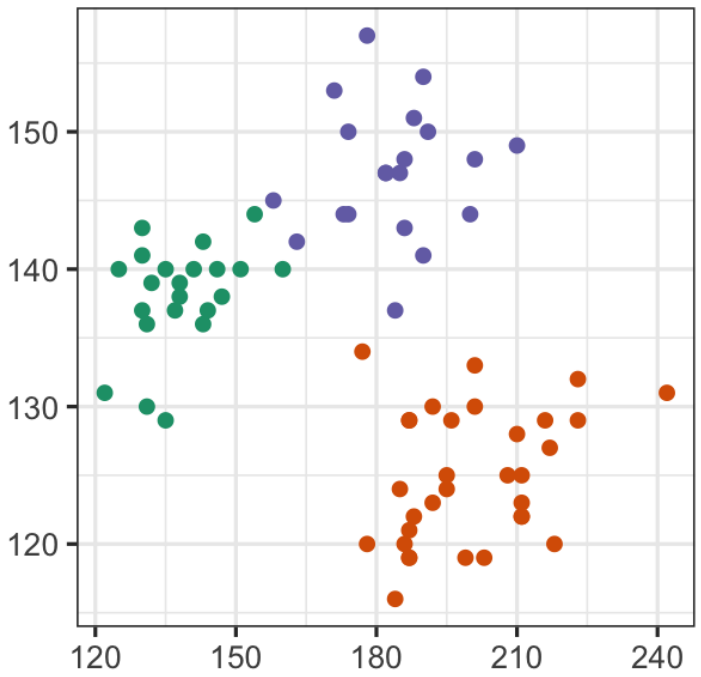

ch
dunn
dunn2
pearsongamma
wb.ratio
within.cluster.ss
150
140
130
120
110
0.20
0.16
0.12
2.00
1.75
1.50
1.25
1.00
0.8
0.7
0.6
0.5
0.45
0.40
0.35
0.30
0.25
60
40
20
2
4
6
8
10
clusters

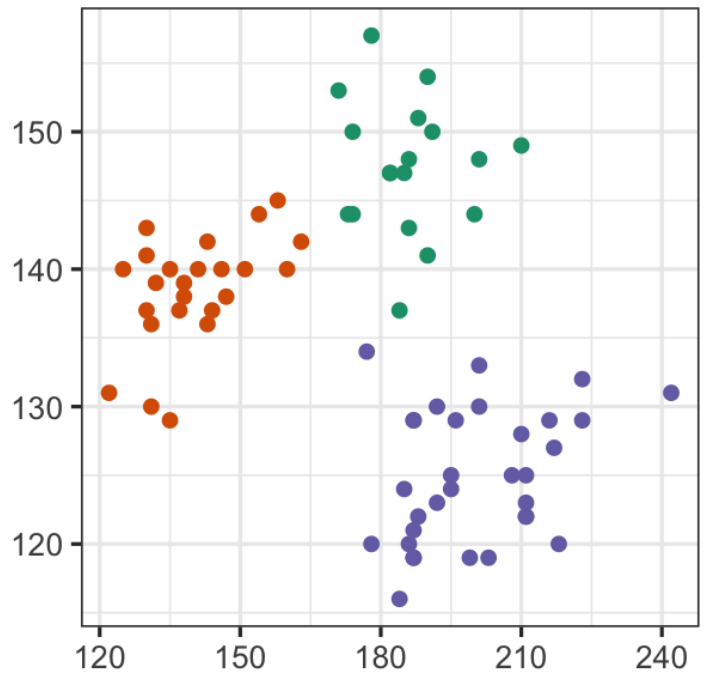
k-means caveats

Effect of seed

- The k-means algorithm can yield quite different results depending on the initial seed.
- Example runs used 5 random starts, and used the `within.cluster.ss` metric to decide on the best solution.







Interpoint distance measures

Euclidean

- Cluster analysis depends on the interpoint distances, points close together should be grouped together
- Euclidean distance was used for the example. Let $A = (x_{a1}, x_{a2}, \ldots, x_{ap}), B = (x_{b1}, x_{b2}, \ldots, x_{bp})$

$$d_{EUC}(A, B) = \sqrt{\sum_{j=1}^{p}(x_{aj} - x_{bj})^2}$$
$$= ((X_A - X_B)^T (X_A - X_B))$$

Other distance metrics

- Mahalanobis (or statistical) distance

$$\sqrt{((X_A - X_B)^T S^{-1} (X_A - X_B))}$$

- Manhattan:

$$\sum_{j=1}^{p} |(X_{aj} - X_{bj})|$$

- Minkowski:

$$(\sum_{j=1}^{p} |(X_{aj} - X_{bj})|^m)^{1/m}$$

Distances for count data

- Canberra:

$$\frac{1}{n_z} \sum_{j=1}^{p} \frac{X_{aj} - X_{bj}}{X_{aj} + X_{bj}}$$

- Bray-Curtis:

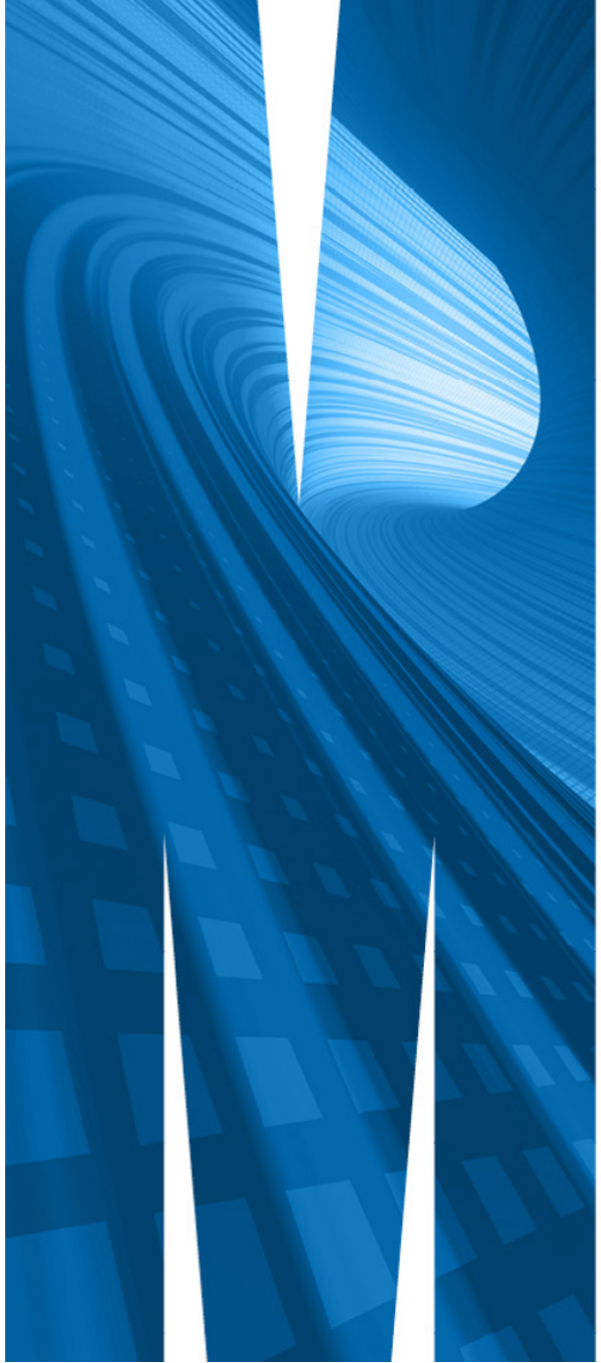
$$\frac{\sum_{j=1}^{p} |X_{aj} - X_{bj}|}{\sum_{j=1}^{p} (X_{aj} + X_{bj})}$$

Interpoint distance measures - Euclidean

Rules for any metric to be a distance

1. $d(A, B) \geq 0$
2. $d(A, A) = 0$
3. $d(A, B) = d(B, A)$
4. Metric dissimilarity satisfies $d(A, B) \leq d(A, C) + d(C, B)$ and ultrametric dissimilarity satisfies $d(A, B) \leq max\{d(A, C), d(C, B)\}$

MONASH University



Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 10a