k-Means Clustering

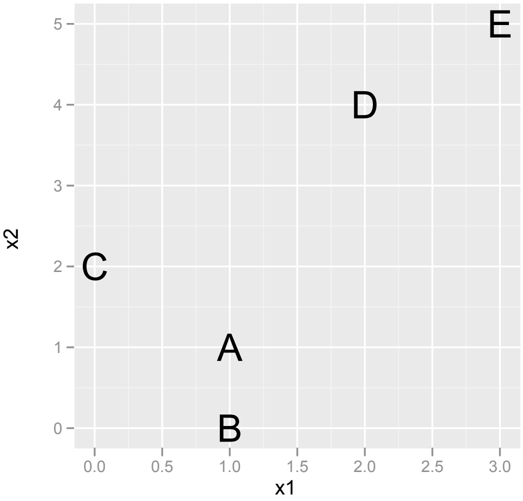
This is an iterative procedure. To use it the ______ ____________, k, must be decided first. The stages of the iteration are:

1. 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).
2. Loop over all observations ___________ them to the group with the closest mean.
3. Recompute group _________.

Iterate steps 2 and 3 until _____________.

Step 0

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5



Use k=2. Suppose A and C are randomly selected as the initial means.

Step 1.1

	i	X1	X2
$\bar{X}_1^0$	A	1	1
	B	1	0
$\bar{X}_2^0$	C	0	2
	D	2	4
	E	3	5

i		
A		1.4
B		2.2
C		0
D		2.8
E		4.2

Compute distances between each of the cluster means and all other points.

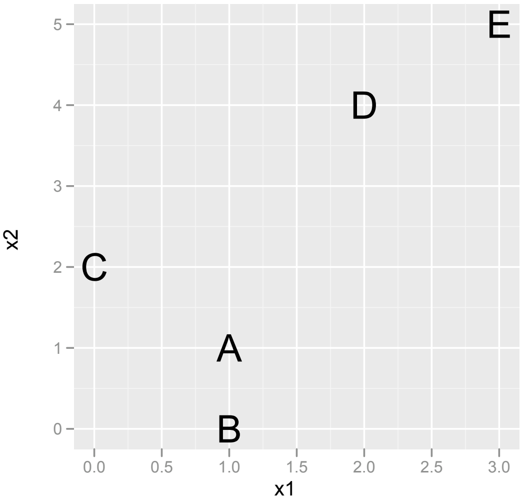
Step 1.1

i			Cluster
A	0	1.4	
B	1	2.2	
C	1.4	0	
D	3.2	2.8	
E	4.5	4.2	

i	X1	X2	
A	1	1	$\bar{X}_1^1$
B	1	0	
C	0	2	
D	2	4	$\bar{X}_2^1$
E	3	5	

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 1.1 - Plots



$\bar{X}_1^1 = (1,0.5)$

$\bar{X}_2^1 = (1.7,3.7)$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 2.1

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

$\bar{X}_1^1 = (1,0.5)$

$\bar{X}_2^1 = (1.7,3.7)$

i		
A		2.7
B		3.7
C		2.4
D		0.5
E		1.9

Compute distances between each of the cluster means and all other points.

Step 2.1

i			Cluster
A	0.5	2.7	
B	0.5	3.7	
C	1.8	2.4	
D	3.6	0.5	
E	4.9	1.9	

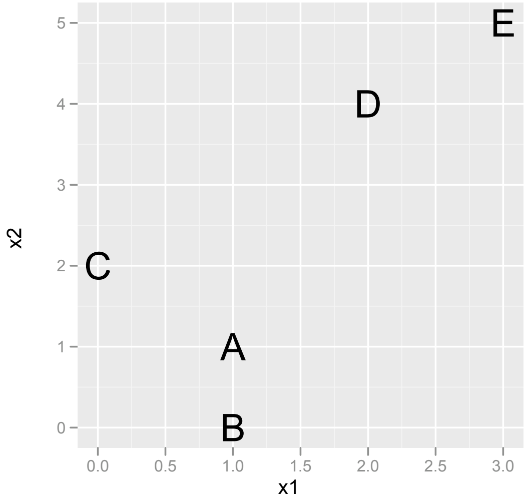
i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

$\bar{X}_1^2$

$\bar{X}_2^2$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 2.1 - Plots

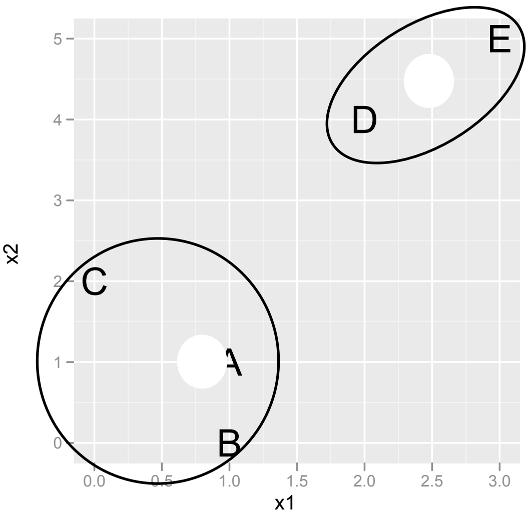


$\bar{X}_1^2 = (0.7,1)$

$\bar{X}_2^2 = (2.5,4.5)$

Assign each case to the cluster having the closest mean. Recalculate the cluster means.

Step 3



$\bar{X}_1^2 = (0.7,1)$

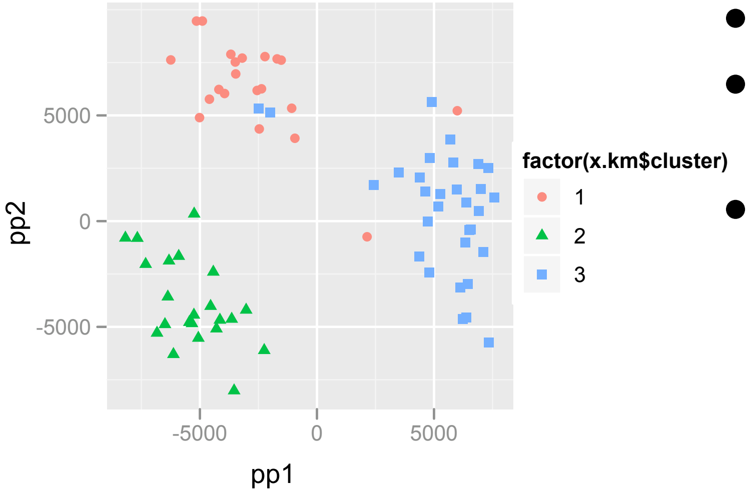
$\bar{X}_2^2 = (2.5,4.5)$

Algorithm has _________ - re-calculating distances, reassigning cases to clusters results in no change. This is the ______________.

k-Means - Initialization

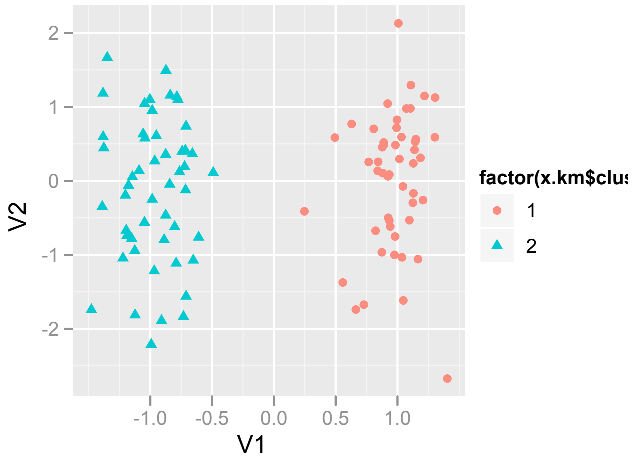
- The algorithm needs to be ___________ by choosing k initial means.
- Approaches:
  1. ___________ choose k points from the data set to act as the initial means.
  2. First do ________________, decide on k, and use the _____ of these clusters as the initial k-means.
- Initialization can _______ the final result.
- If k is not known, re-run for several ________ k.

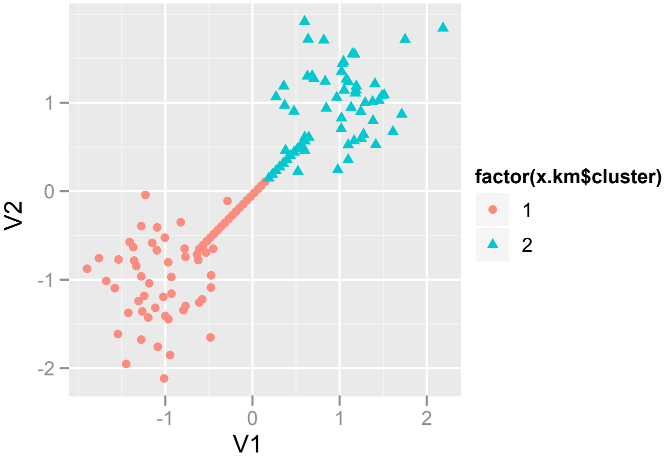
Examples



- Flea beetles
- Several cases are confused.
- Why would k-means have trouble with this data?

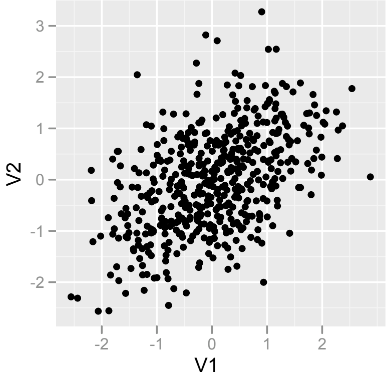
Example





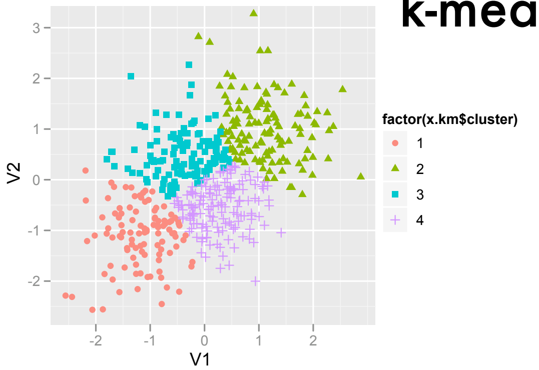
- k-means does not handle nuisance variables well, but surprisingly does well with these data sets.

Example - partitioning

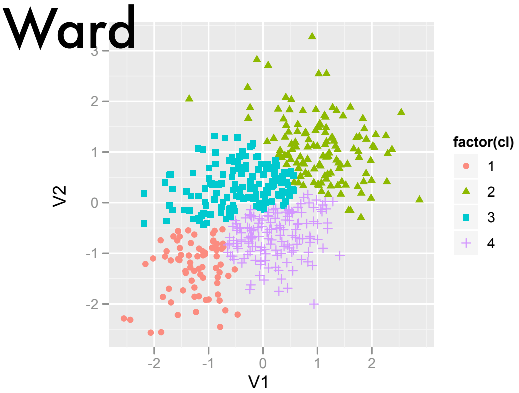


- Many clustering tasks involve __________ data into chunks.
- There may not be natural clusters.

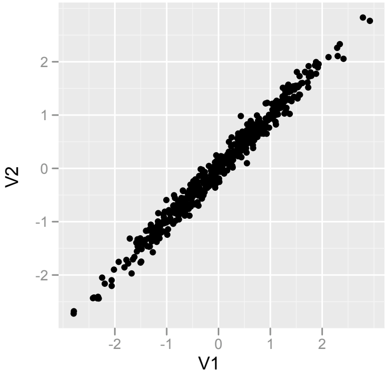
k-means



Ward

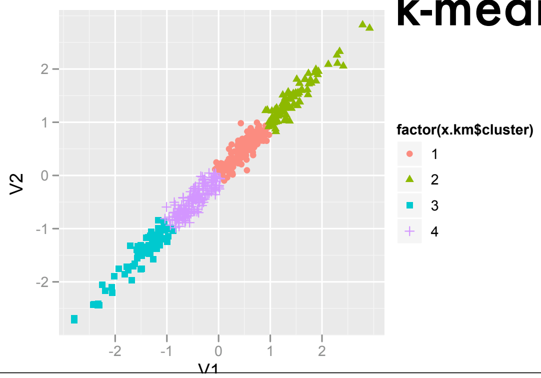


Example - partitioning

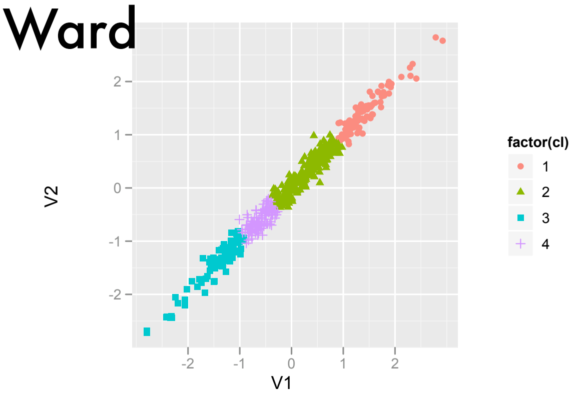


- ____________ matters in the way the data gets partitioned.

k-means



Ward



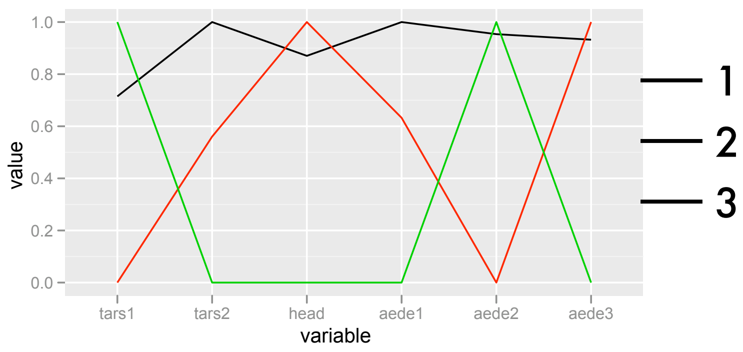
Summarizing results

Need to show how the clusters ______ from each other:

- Tabulate the _________________ for each cluster.
- Make separate plots for each cluster, using same scale
- Plot the __________ on one plot

Example

cluster	tars1	tars2	head	aede1	aede2	aede3
mean 1	183.10	129.62	51.24	146.19	14.10	104.86
sd 1	12.14	7.16	2.23	5.63	0.89	6.18
mean 2	138.23	125.09	51.59	138.27	10.09	106.59
sd 2	9.34	8.55	2.84	4.14	0.97	5.85
mean 3	201.00	119.32	48.87	124.65	14.29	81.00
sd 3	14.90	6.65	2.35	4.62	1.10	8.93

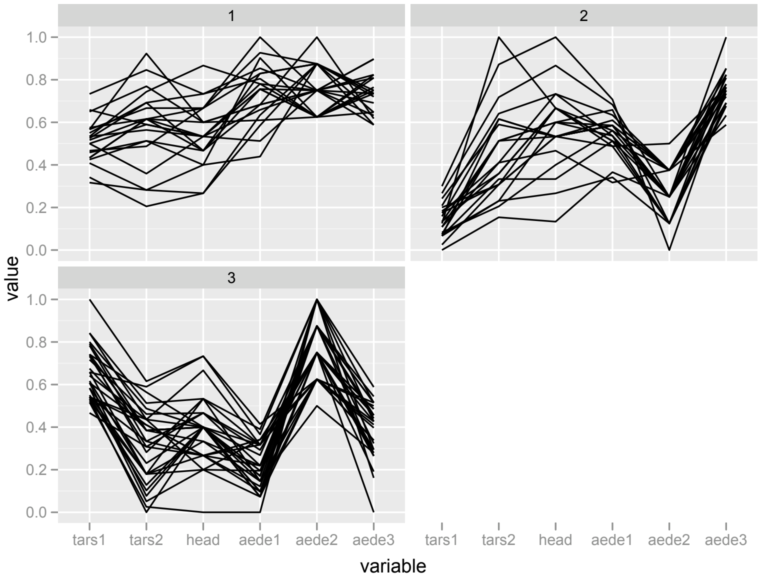


Cluster 1 has ____ values on all variables.

Cluster 2 has ___ values for tars1 and aede2, ___ values of head and aede3.

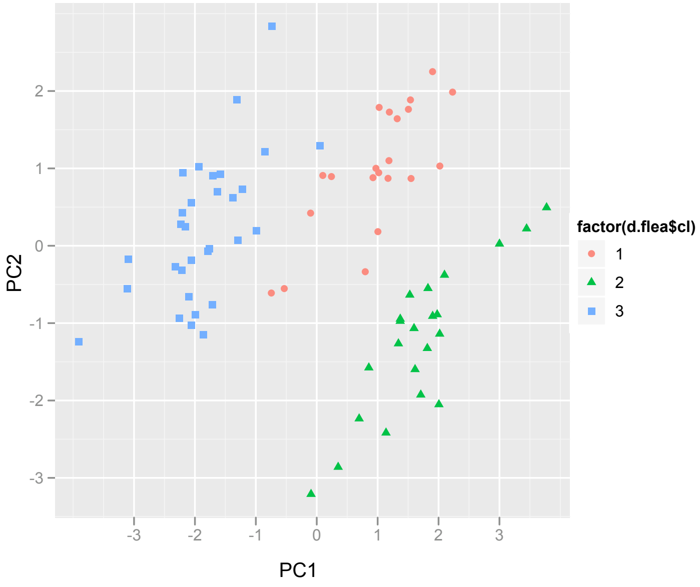
Cluster 3 has ____ values of tars 1 and aede2, but ___ values of all other variables.

Example



Plotting all of the data shows the __________ in each cluster.

Example



Plotting the clusters in a _____________ _____________ like the first two principal components can also help evaluate the clusters.