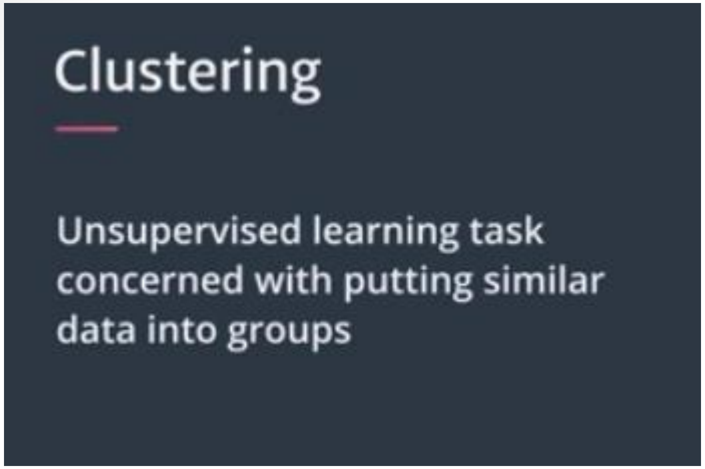
K – Means

Un-Supervised learning algorithm

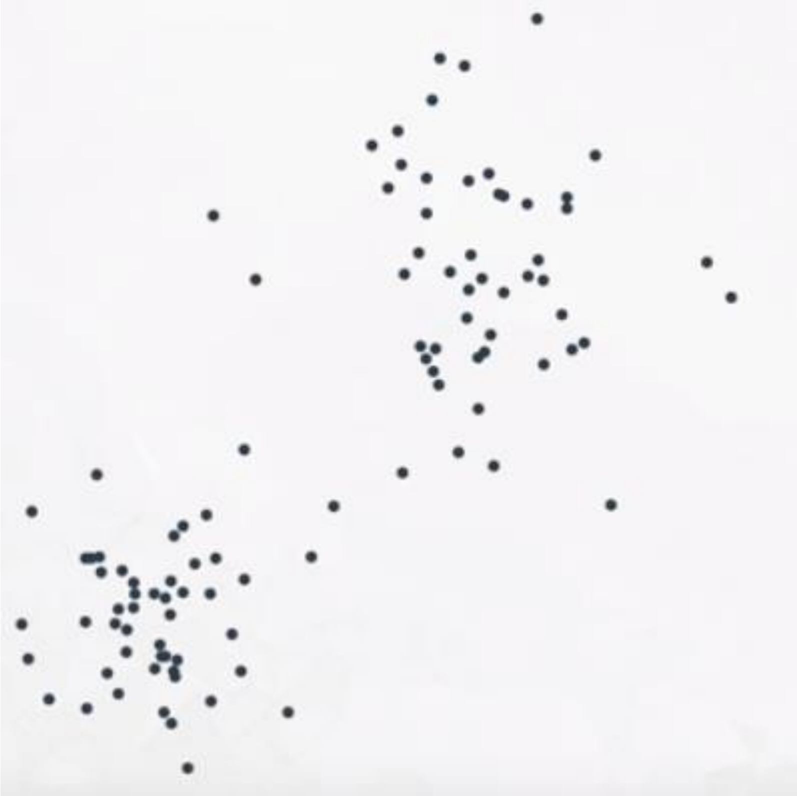
Clustering

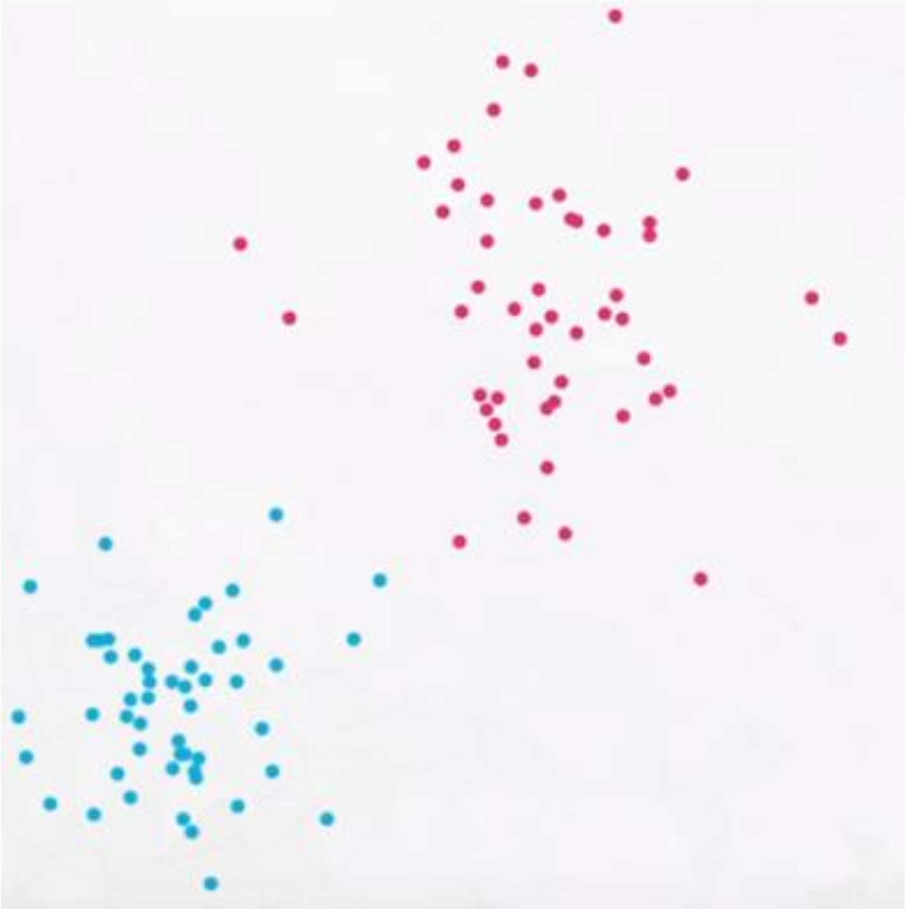
No dependant variable



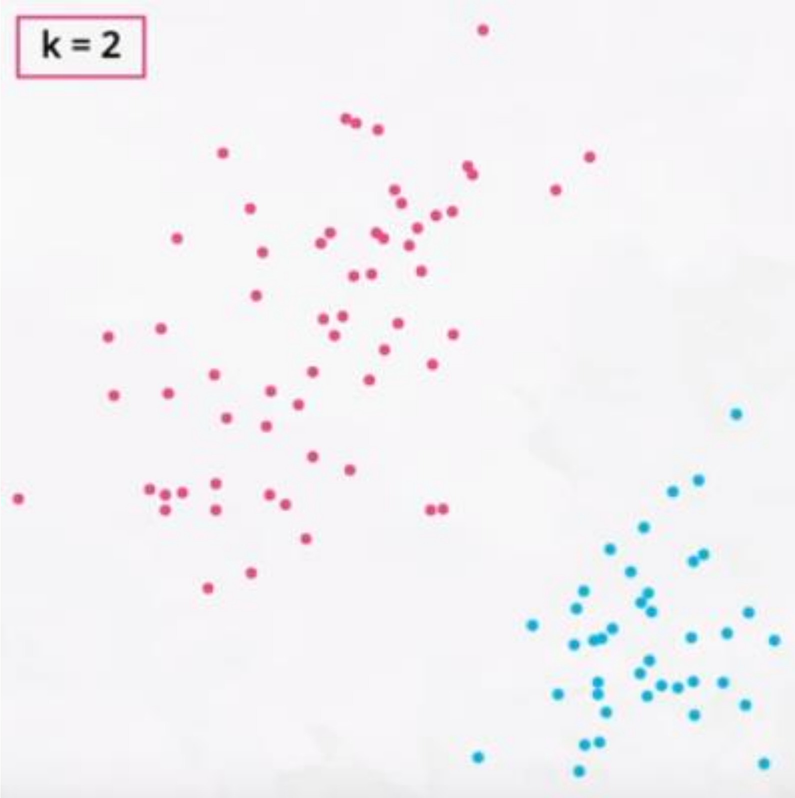
Clustering

Unsupervised learning task concerned with putting similar data into groups

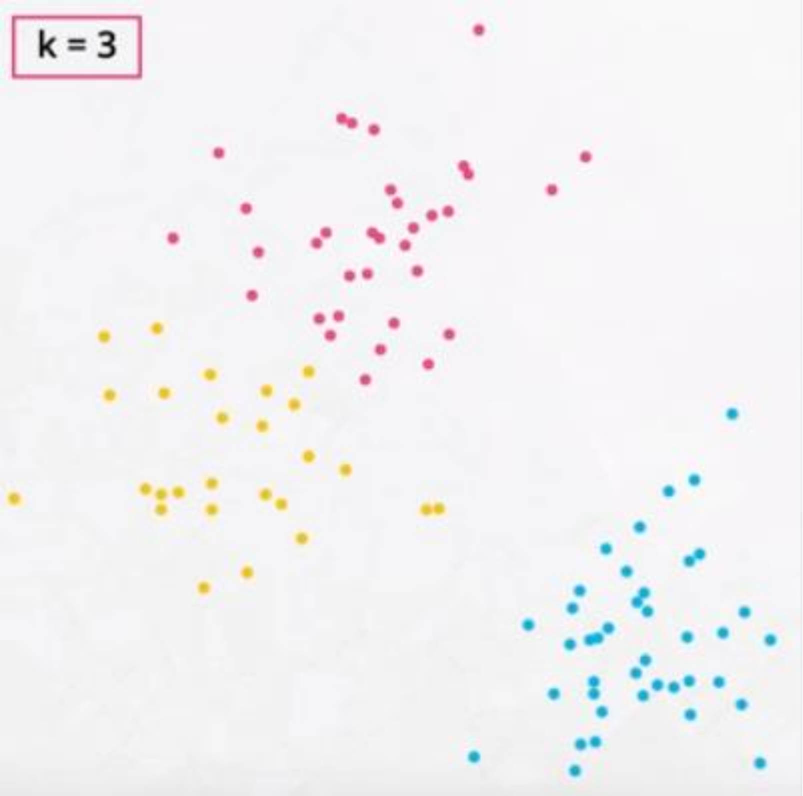




Sample Dataset - Can be 2 Groups ?

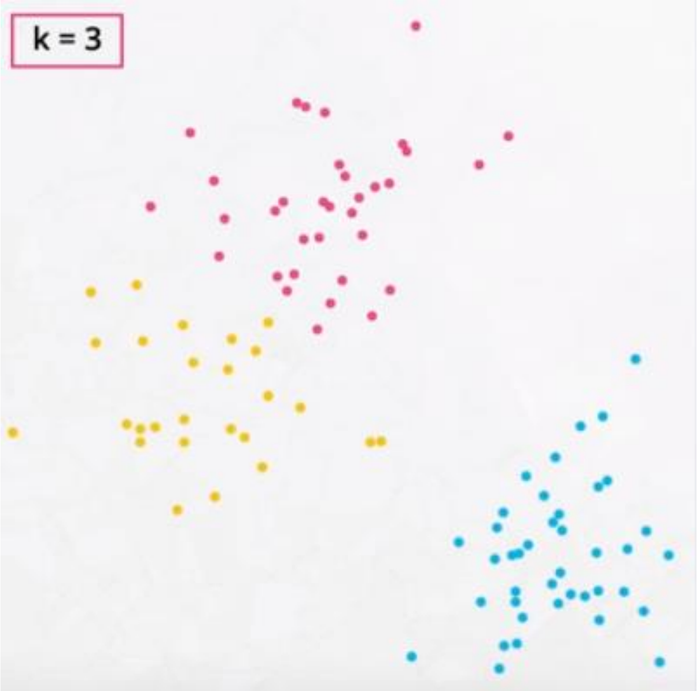


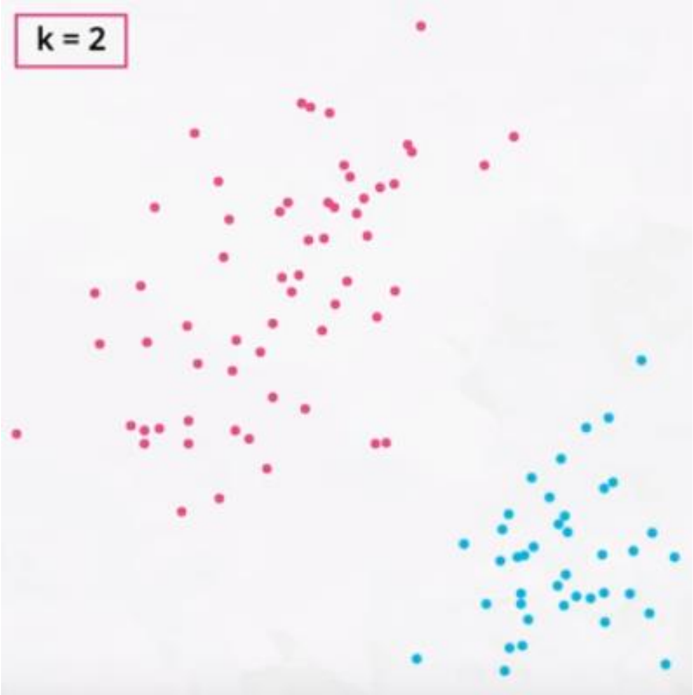
Sample Dataset - Can be 3 Groups ?



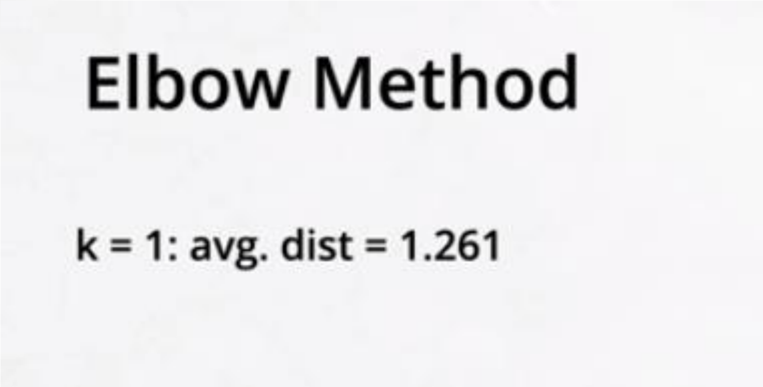
2 or 3 Groups ?

Which is better



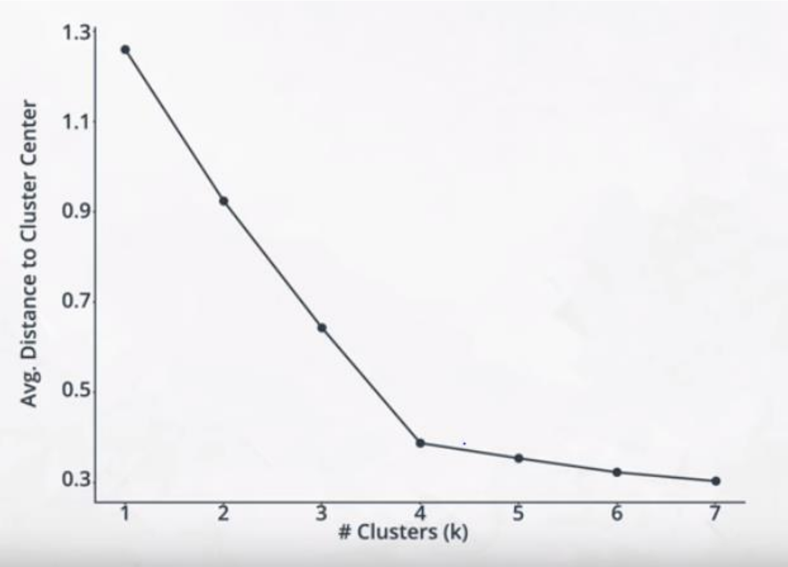


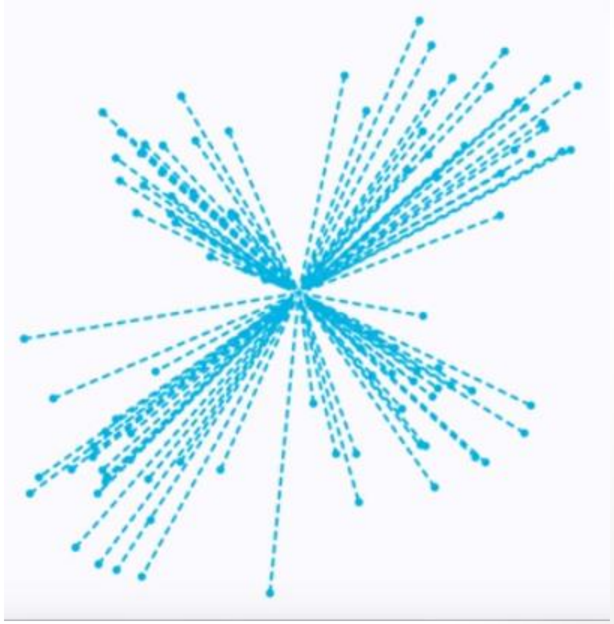
How to find optimum K value ?

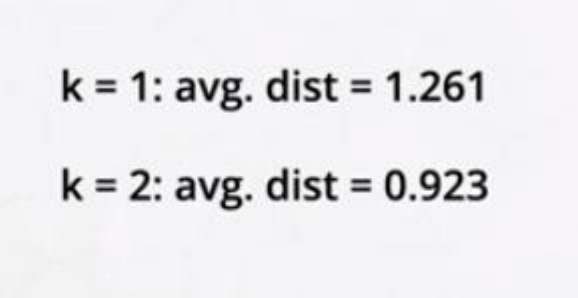


Elbow Method

k = 1: avg. dist = 1.261

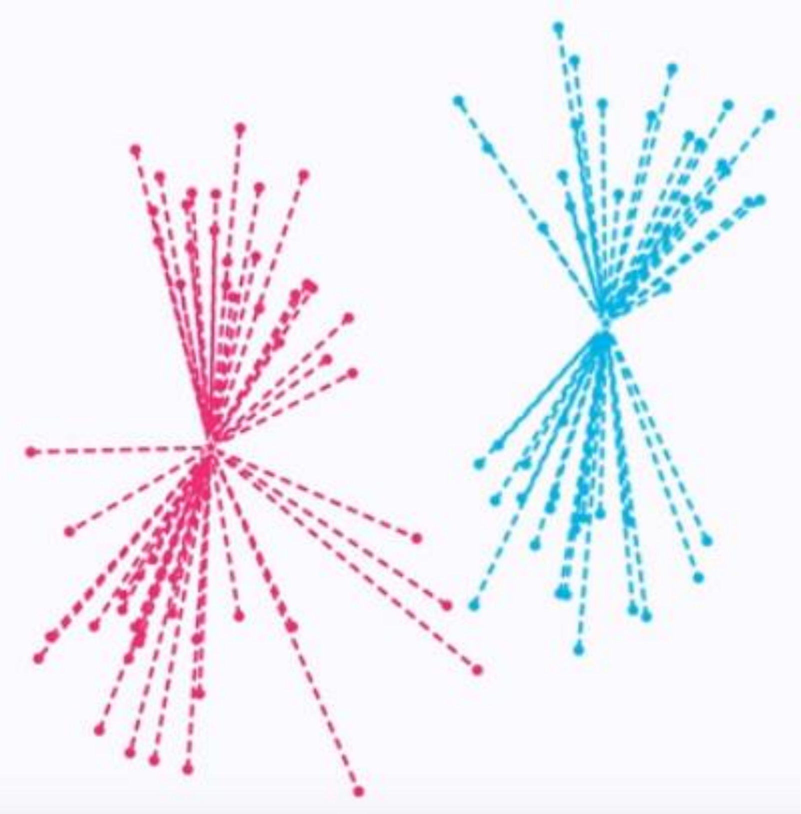


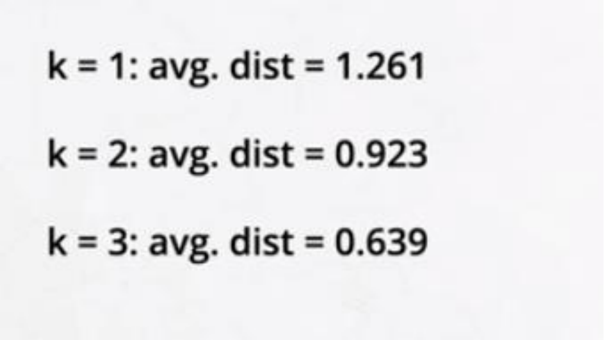




k = 1: avg. dist = 1.261

k = 2: avg. dist = 0.923

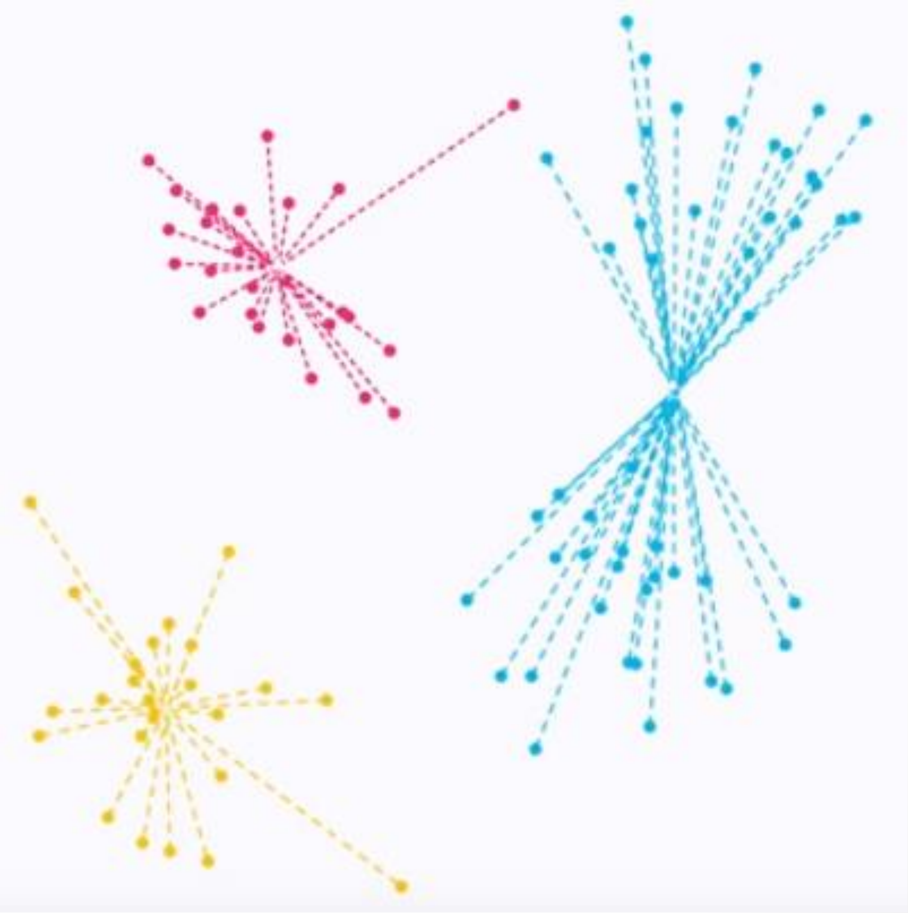


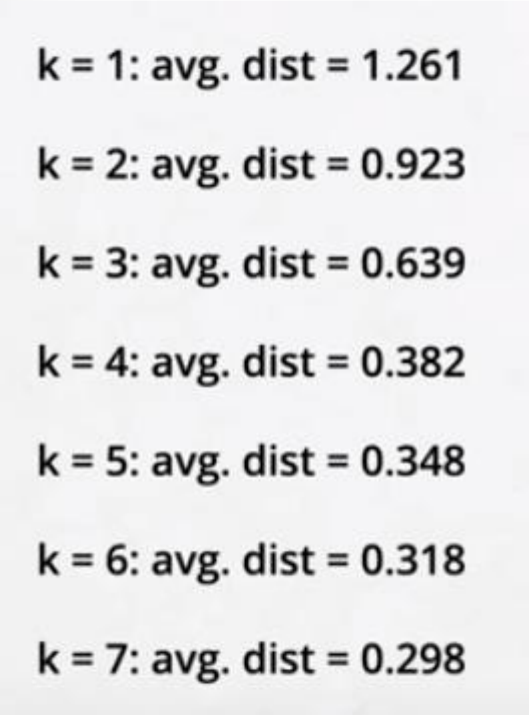


k = 1: avg. dist = 1.261

k = 2: avg. dist = 0.923

k = 3: avg. dist = 0.639





k = 1: avg. dist = 1.261

k = 2: avg. dist = 0.923

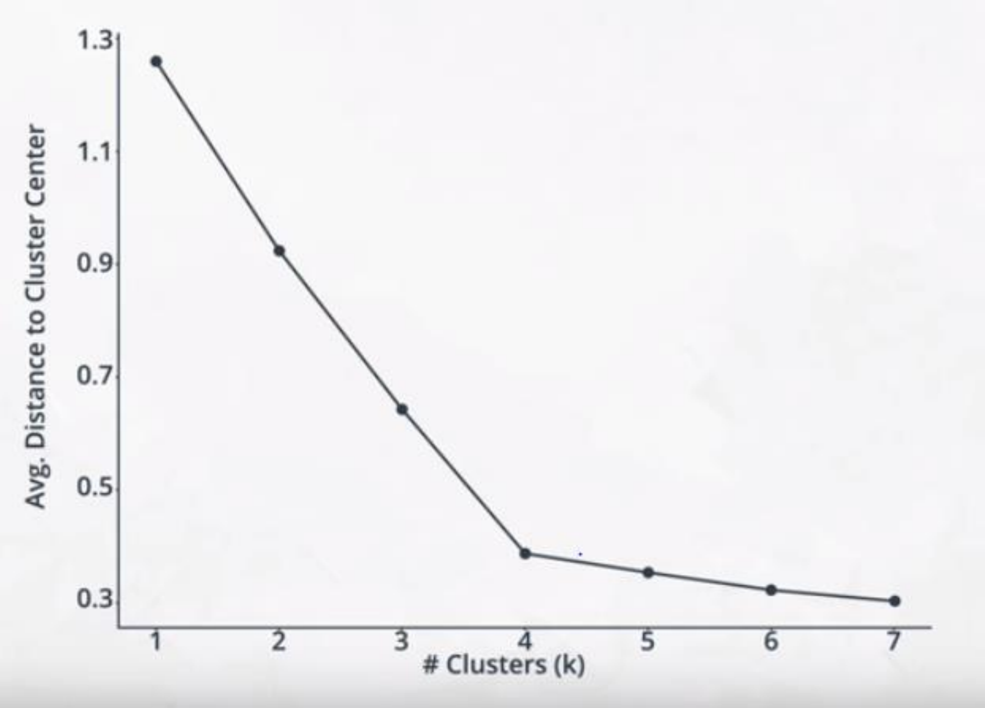
k = 3: avg. dist = 0.639

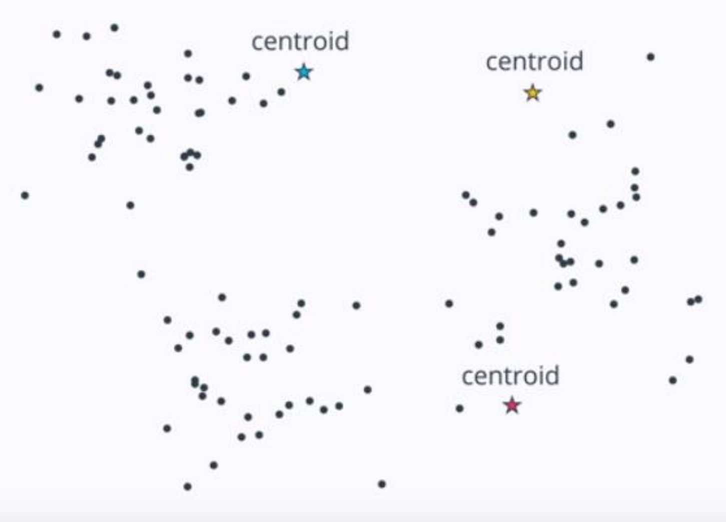
k = 4: avg. dist = 0.382

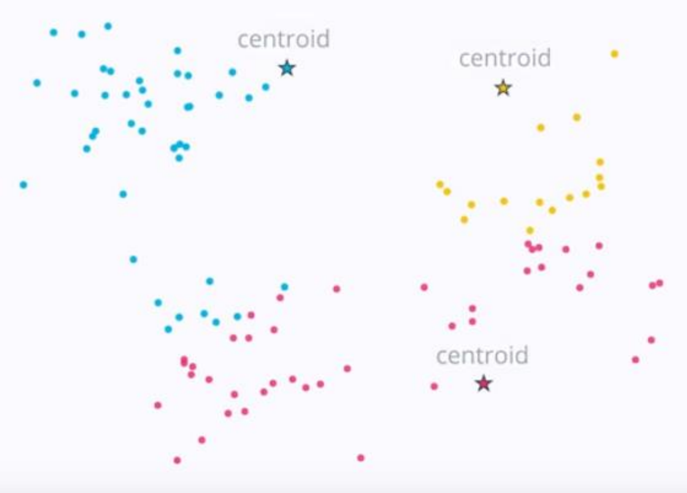
k = 5: avg. dist = 0.348

k = 6: avg. dist = 0.318

k = 7: avg. dist = 0.298



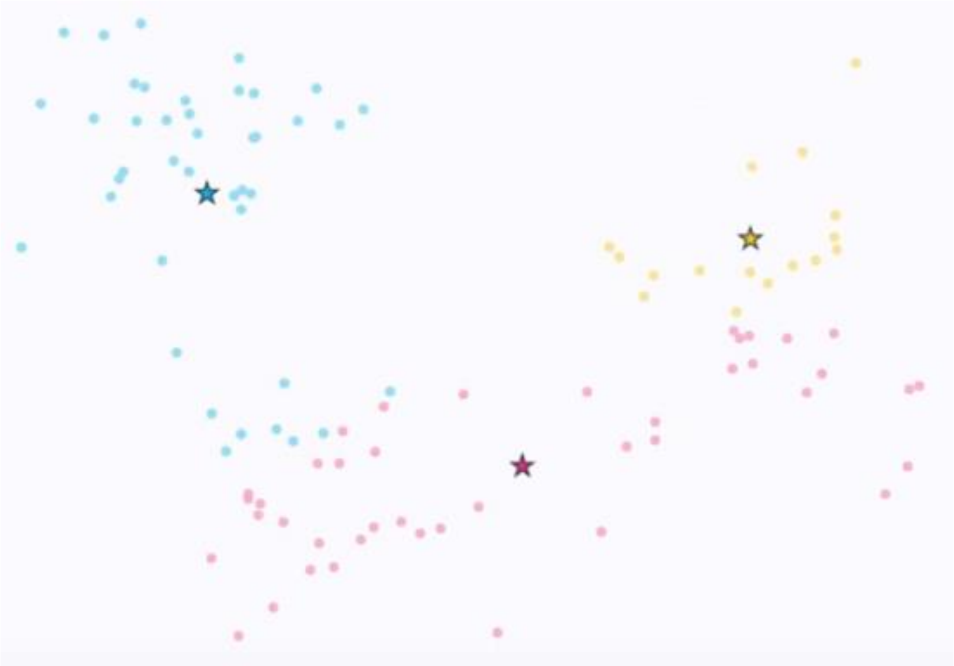




1. Initially assign 3 random centroids

2. Find the distance of the points closest to the centroid and assign it to it

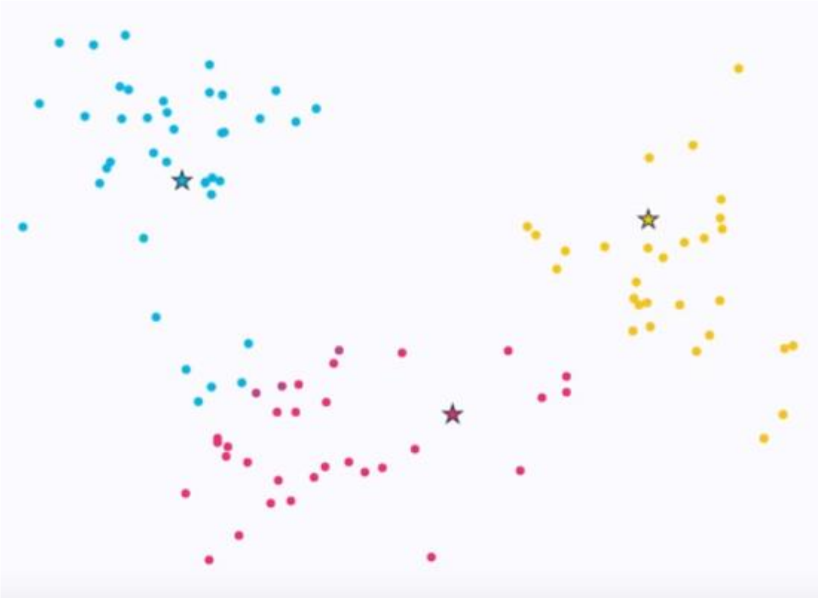
Step 1



Step 2

3. Re compute the new Centroid from the average of all the points in that group.

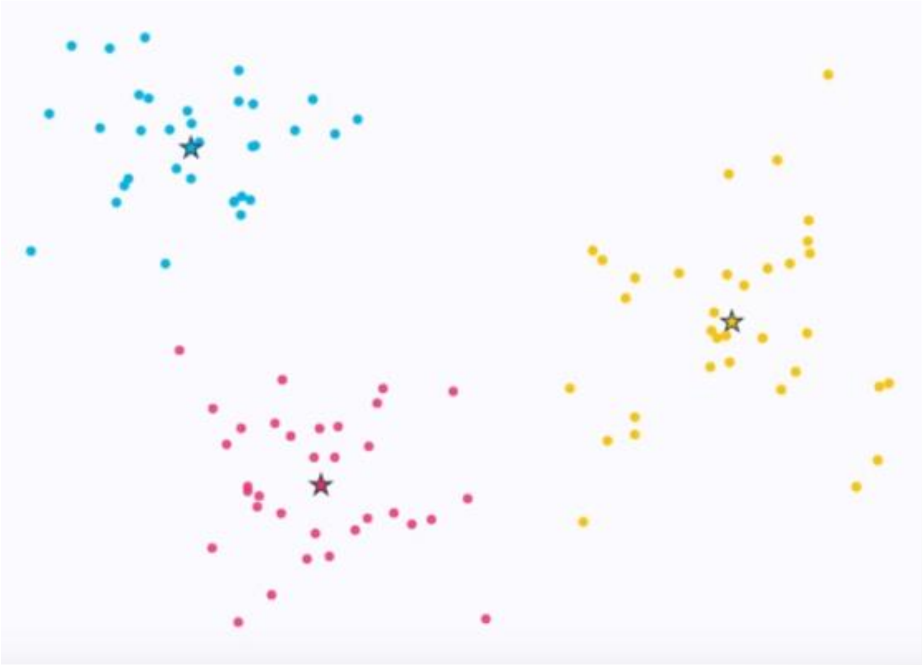
4. Iterate the process again



NOTE: See the change in Points assigned

Step 3

5. Iterate the process untill no points change the group

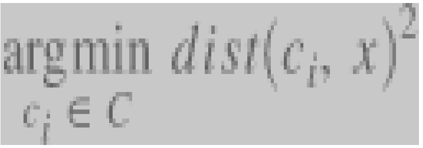


Pseudocode

- Input the algorithm with the number of clusters **K** and the data set.
- Randomly generate or randomly select K centroids from the data set.

The algorithm then iterates between two steps:

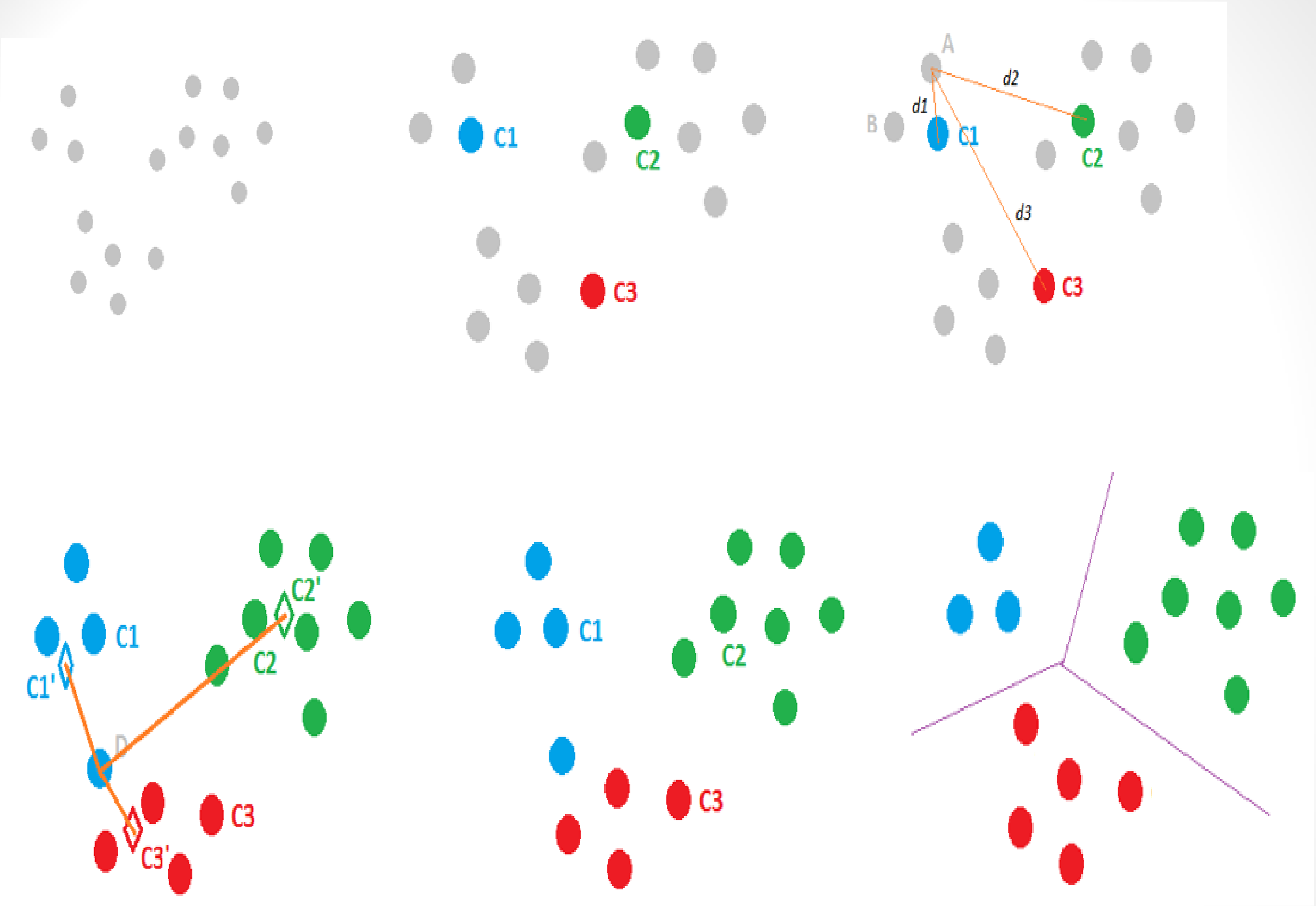
1. Data assignment step



$$\underset{c_i \in C}{\operatorname{argmin}}\ dist(c_i, x)^2$$

where *dist*(·) is the standard (L_2)

Euclidean distance


C1
C2
C3
A
d2
d1
B
C1
C2
d3
C3
C2'
C1
C2
C1'
D
C3
C3'
C1
C2
C3

2. Centroid update step:

In this step, the centroids are recomputed. This is done by taking the mean of all data points assigned to that centroid's cluster.

$$c_i = \frac{1}{|S_i|}\sum\nolimits_{x_i \in S_i} x_i$$

The algorithm iterates between steps one and two

1. No data points change clusters

2. The sum of the distances is minimized or some maximum number of iterations is reached

K Means ++

From

K Means

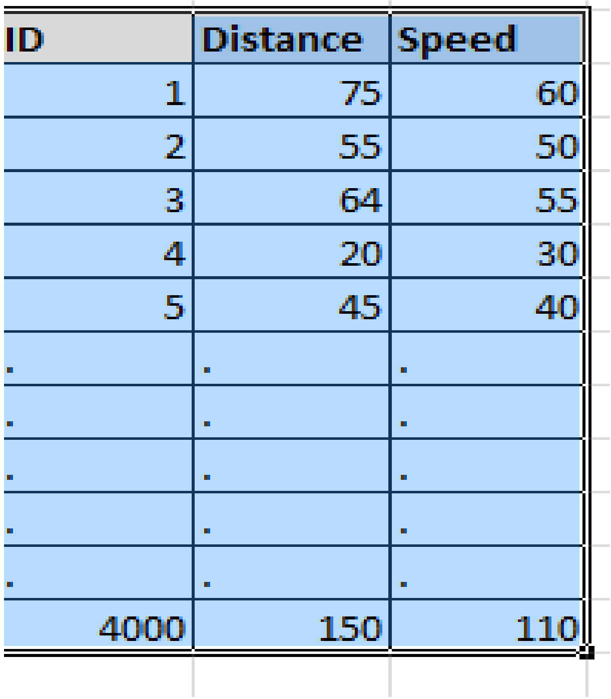
K Means ++

- In K means the initial centroids are randomly placed. This makes the points to locate different clusters based on how the centroids are placed initially

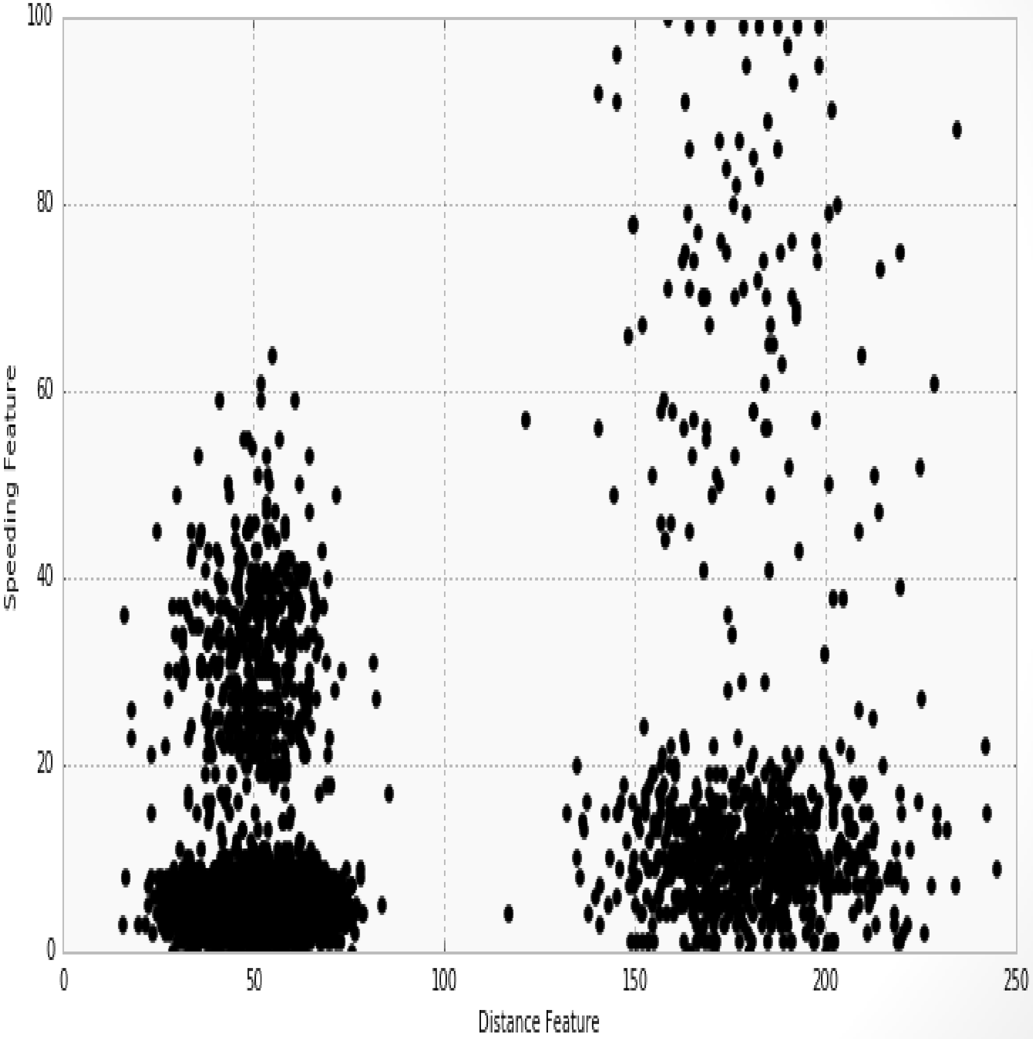
- To overcome this K-means ++ uses the farthest distant placement of centroids initially when assigned

Simulation: https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

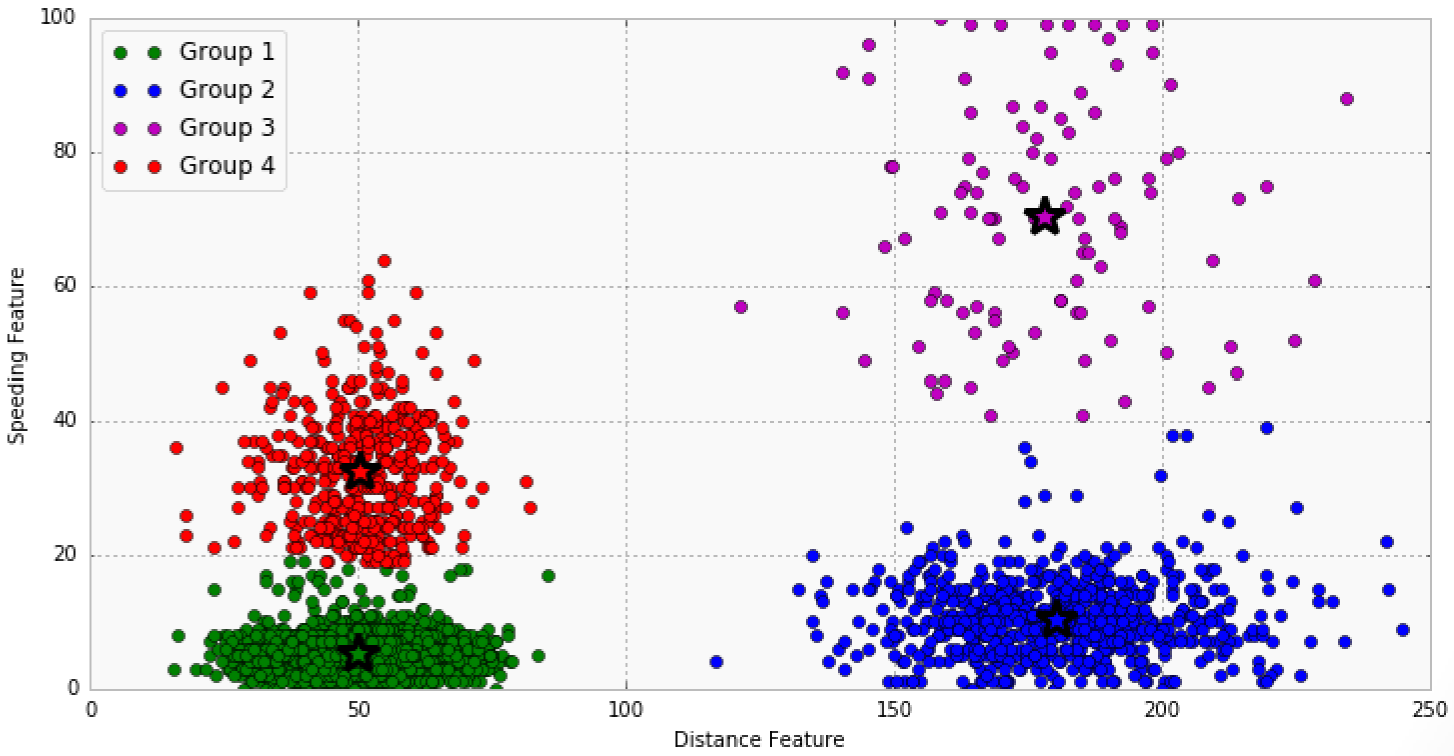
Distance and Speed



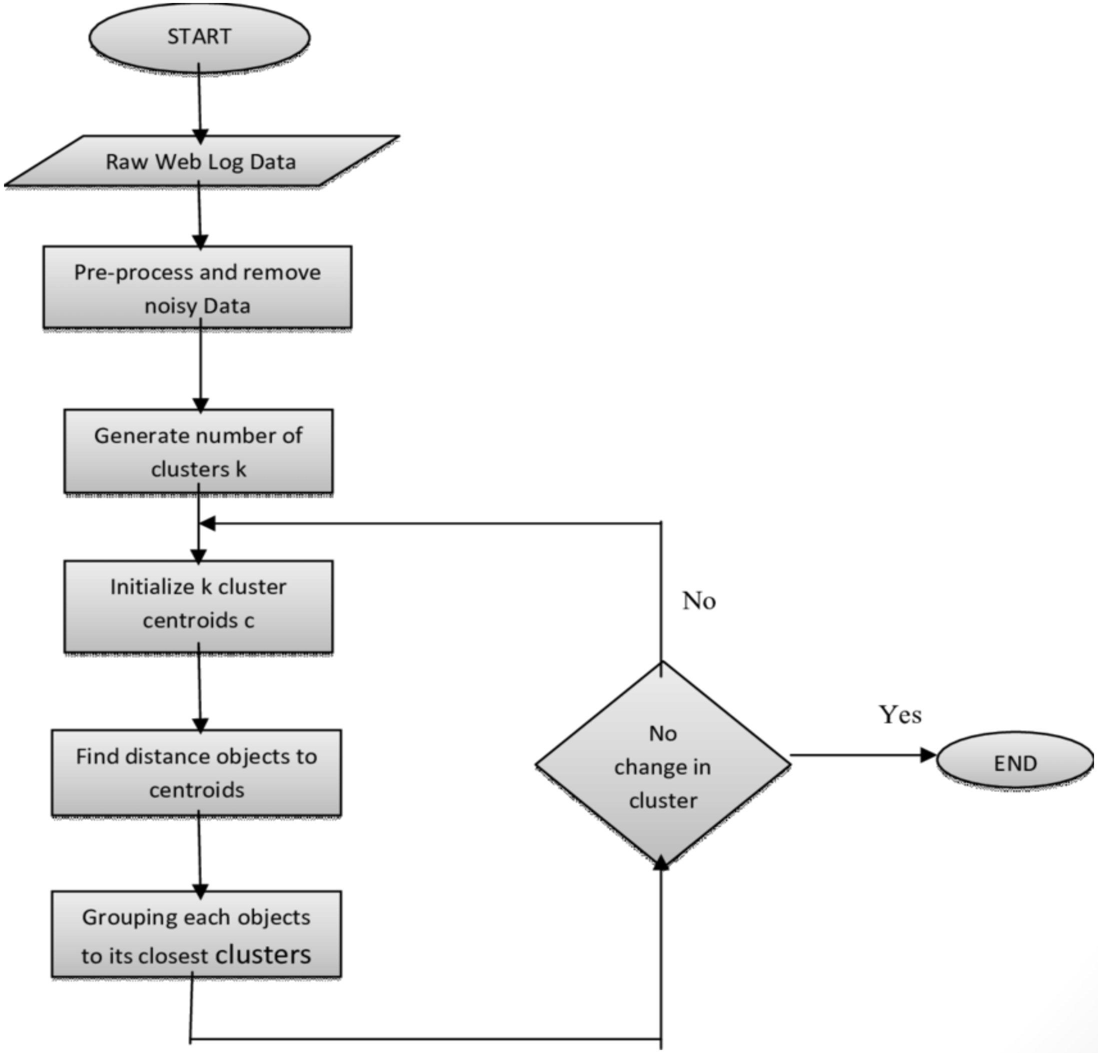
ID	Distance	Speed
1	75	60
2	55	50
3	64	55
4	20	30
5	45	40
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
4000	150	110



Graph


Group 1
Group 2
Group 3
Group 4
100
80
60
40
20
0
0
50
100
150
200
250
Speeding Feature
Distance Feature

Flow Chart


START
Raw Web Log Data
Pre-process and remove noisy Data
Generate number of clusters k
Initialize k cluster centroids c
Find distance objects to centroids
Grouping each objects to its closest clusters
No change in cluster
No
Yes
END

Key Points

- No prediction – The interest is group to similar kind of attributes to a common class

Example -

- Same language documents are one group.
- While categorising the news articles (Same news category(Sport) articles are one group)

Result of K- means

1. The centroids of the K clusters, which can be used to label new data

2. Labels for the training data (each data point is assigned to a single cluster)