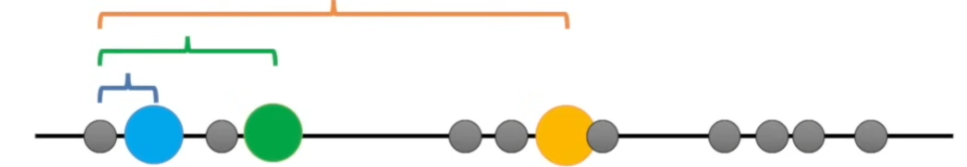
K-Means clustering

Steps:

1. Select the number of clusters you want to identify in the data. This is called K in k-Means clustering.
2. Randomly select data points (if k=3 we will select 3 data points randomly)
3. Now measure the distance between data points and each cluster (random data point selected earlier)



1. Assign the 1st data point to the nearest data point based on the cluster distance, whichever is closed then assign that data point to that. Do the same for the rest of the data points and assign each data points to a cluster based on minimum distance.



1. Now that all the points are in cluster, we have to calculate the mean between the clusters

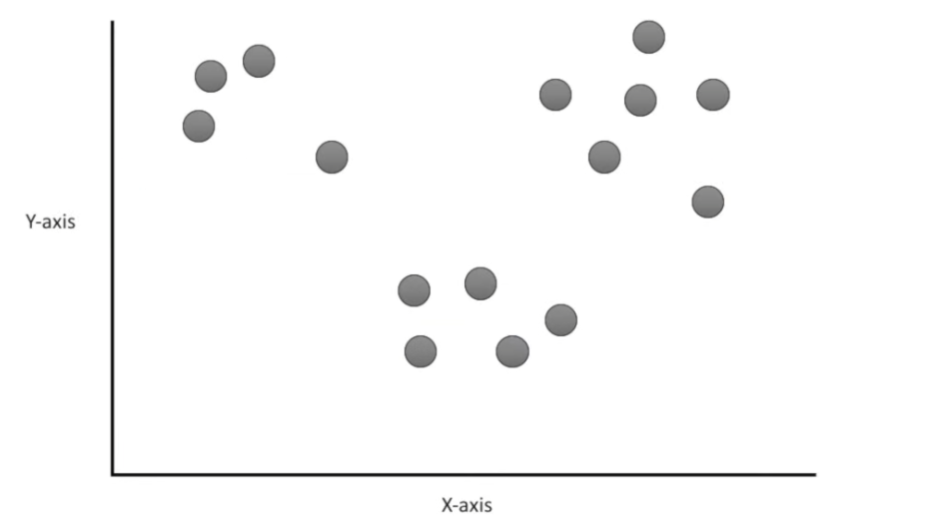


1. Now repeat the step 4 but the now the distance is between data points and the mean of the cluster and then assign the value to the closest cluster mean value.
2. This process is repeated until the cluster no longer changes.

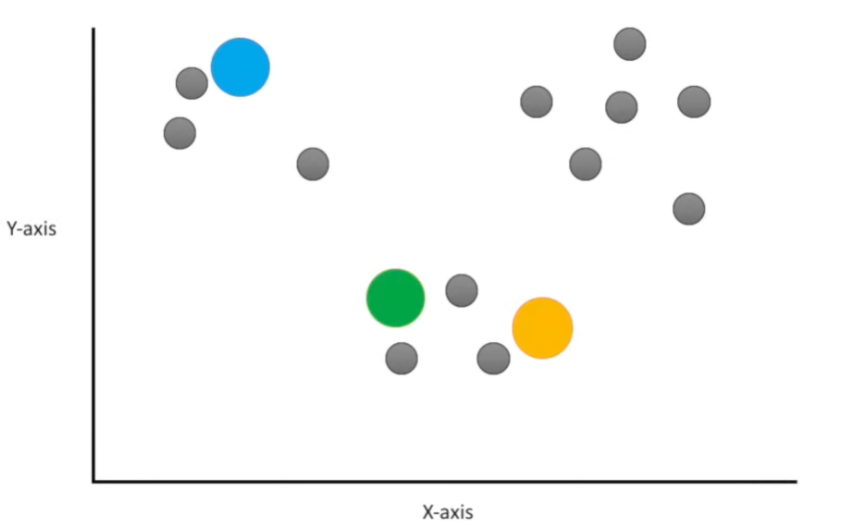
How to derive K value?

* The derive a good k value, we have to test different k value (let’s say 1,2,3,4, etc.). For each k-value we will get WSS (within cluster sum of square). Now we have sum each WSS and get total WSS per K and then find for which k we are getting less Total WSS.
* This can be found using a elbow plot.

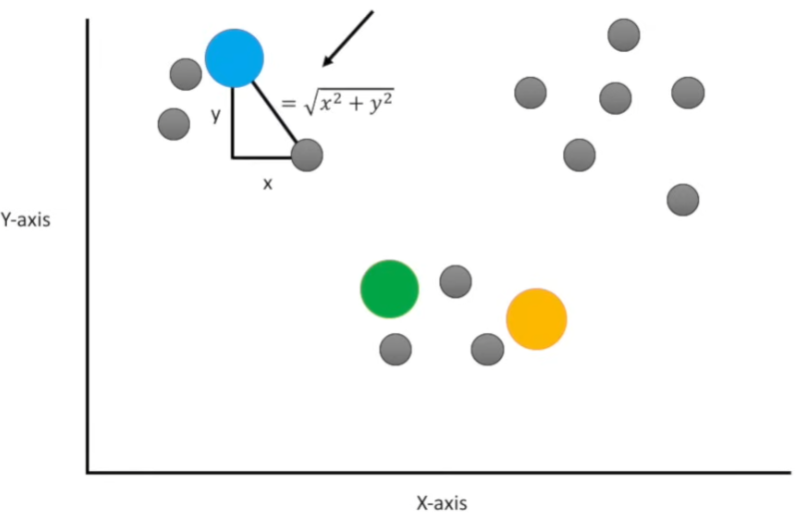
What if the data in 2 dimensional, how can we do k-means in this?



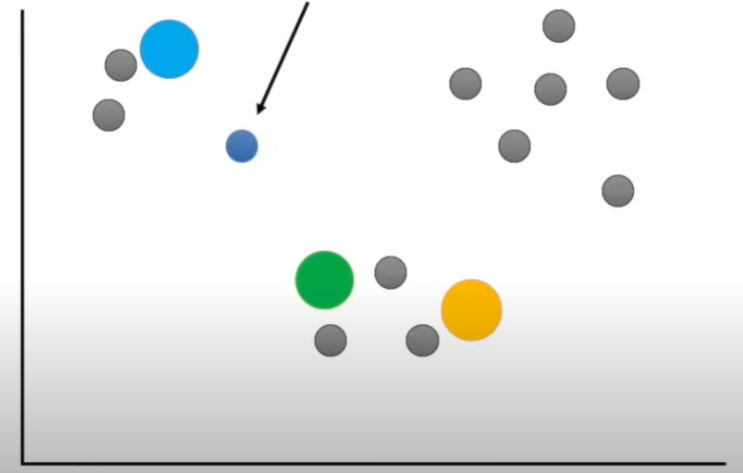
* Just like before pick data points as per k random points.



* Calculate the Euclidean distance



* Again, we assign points to the nearest cluster



* Just like before we calculate the centre of the cluster and then recalculate the mean until the cluster has no change.
* Iterate for different k values and find the k which gives minimum Total WSS value. This can be found using elbow method.

