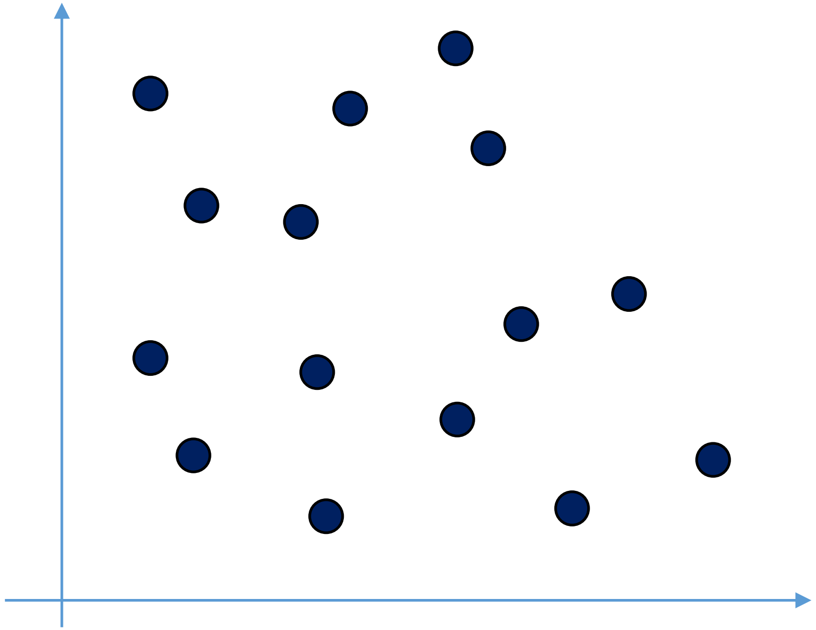
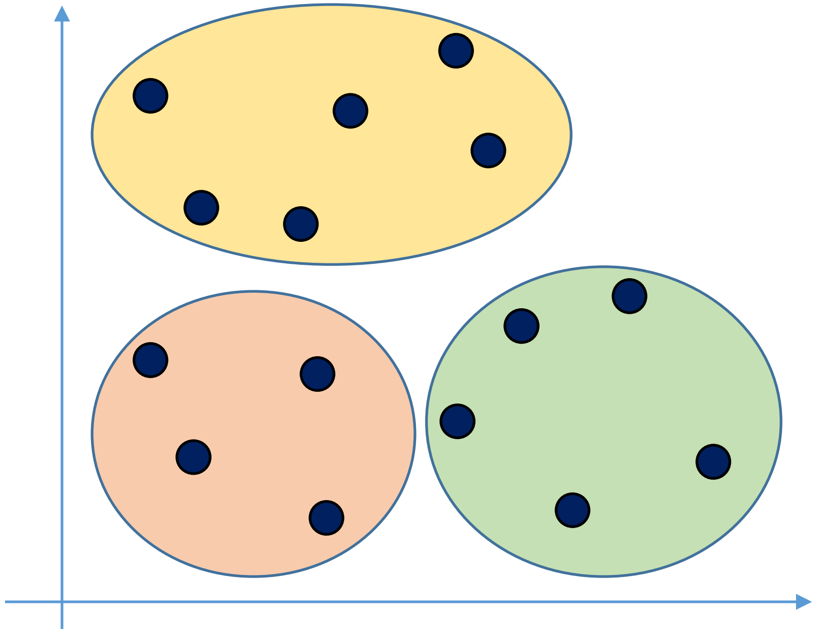
k-Means Clustering

k-Means

- Widely used in classification of data based on the Centroid-based clustering
- A black-box algorithm
- Algorithm breaks the dataset into 'k' different clusters
- Number of clusters to be broken into is specified by the user (Eg. **k=3** breaks dataset into 3 clusters)
- Number of clusters has to be known beforehand



Before clustering



After clustering

K-Means Algorithm

- Identify the number of clusters (**k**=*n*) [n >= 2] (Optional Step)
- Algorithm assigns <k> random values as Centroid values - one for each cluster
- Assign every record (observation) to the nearest centroid based on **Distance Calculation**
  - forms **k**-clusters, each having **n** observations
- Compute new centroids for each cluster
  - The means of each cluster become the new centroids
- Reassign record to the new centroid **(step 3)** and repeat process 4 till no new assignments
- Build the Model

X	C1	C2	C3	d1	d2	d3	Min	Cluster
94.8	15	55	30	79.82	39.82	64.82	39.82	2
26.7				11.71	28.29	3.29	3.29	3
32.0				17.05	22.95	2.05	2.05	3
62.6				47.6	7.60	32.60	7.60	2
30.0				15.02	24.98	0.02	0.02	3
25.5				10.55	29.45	4.45	4.45	3
31.6				16.61	23.39	1.61	1.61	3
58.2				43.2	3.20	28.20	3.20	2
46.1				31.11	8.89	16.11	8.89	2
2.7				12.26	52.26	27.26	12.26	1
94.5				79.45	39.45	64.45	39.45	2
95.0				79.97	39.97	64.97	39.97	2
9.7				5.257	45.26	20.26	5.26	1
74.6				59.56	19.56	44.56	19.56	2
21.3				6.265	33.74	8.74	6.26	1
90.2				75.16	35.16	60.16	35.16	2
15.2				0.217	39.78	14.78	0.22	1
82.2				67.21	27.21	52.21	27.21	2
64.7				49.65	9.65	34.65	9.65	2
44.7				29.73	10.27	14.73	10.27	2

	cluster 1	cluster 2	cluster 3	
	12.26	39.82	3.29	
	5.26	7.6	2.05	
	6.26	3.2	0.02	
	0.22	8.89	4.45	
		39.45	1.61	
		39.97		
		19.56		
		35.16		
		27.21		
		9.65		
		10.27		
mean	**6.000**	**21.889**	**2.284**	
var	**24.40**	**217.39**	**2.83**	**244.63**

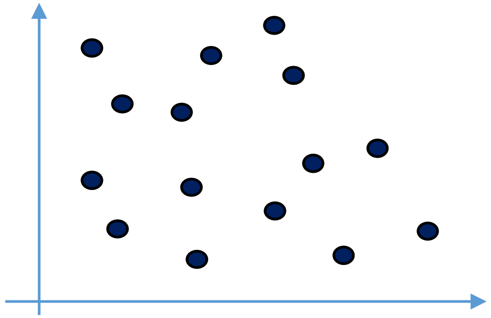
C1	C2	C3
6	21.889	2.284

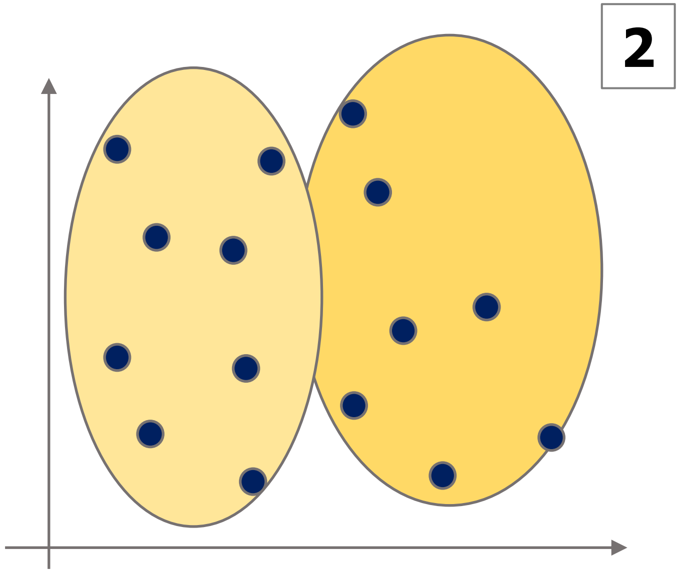
Random Initialization trap

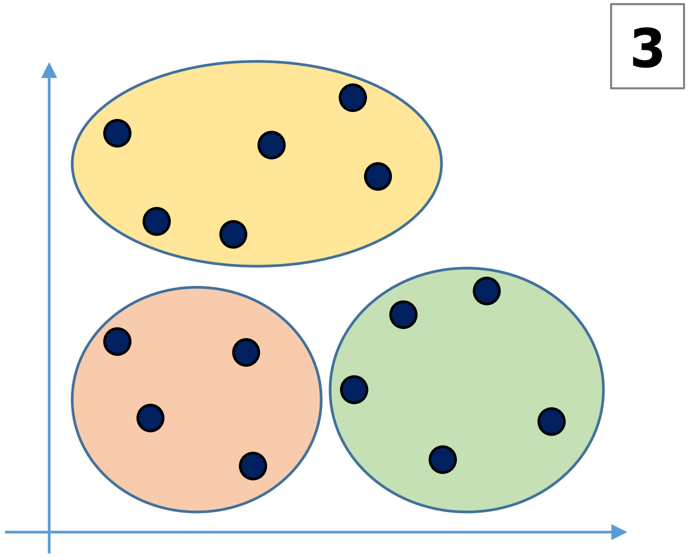
- Random values taken as weights for each 'k' cluster

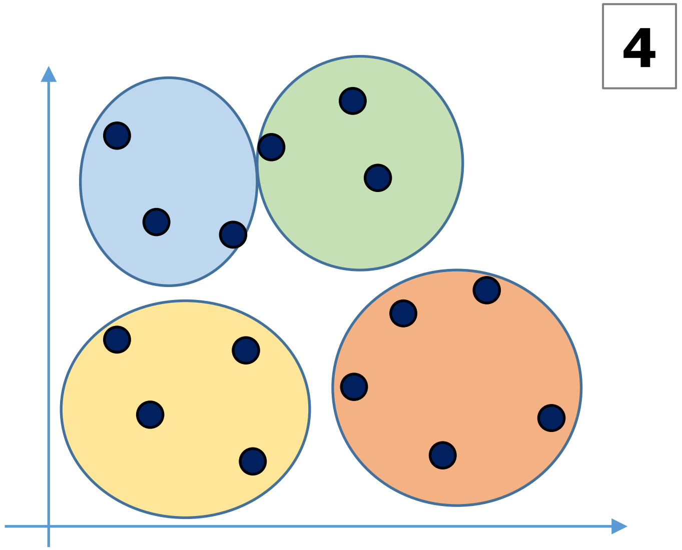
- Observations in the Clusters might change depending upon these random values
  - A cluster **k1** can have more observations
  - A cluster **k1** can have less observations
  - A cluster **k1** having an observation could have moved to another cluster **k2**

Optimum selection of Clusters

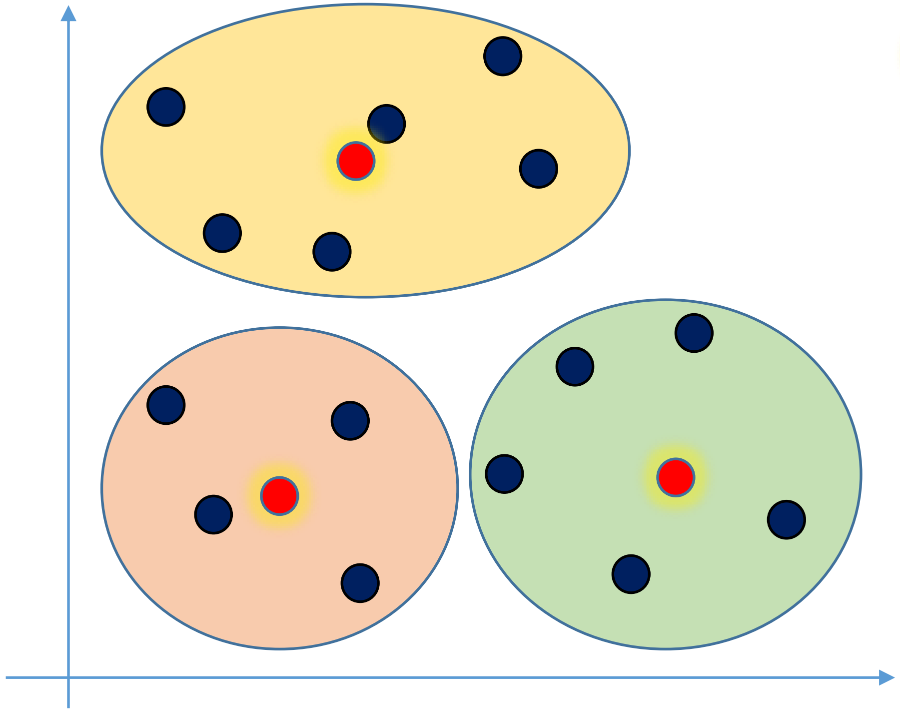








Within Cluster Sum of Squares (WCSS)



Element within a cluster (e)

Centroid of cluster (c)

Within Cluster Sum of Squares (WCSS) =

$$\sum_{c} \Sigma_{e_c}\, distance\,(e,c)^2$$

- As the number of clusters increase, Errors decrease
- Optimum cluster is the one that shows less difference in the errors with the previous error component
- Using the Elbow chart, it is easy to determine