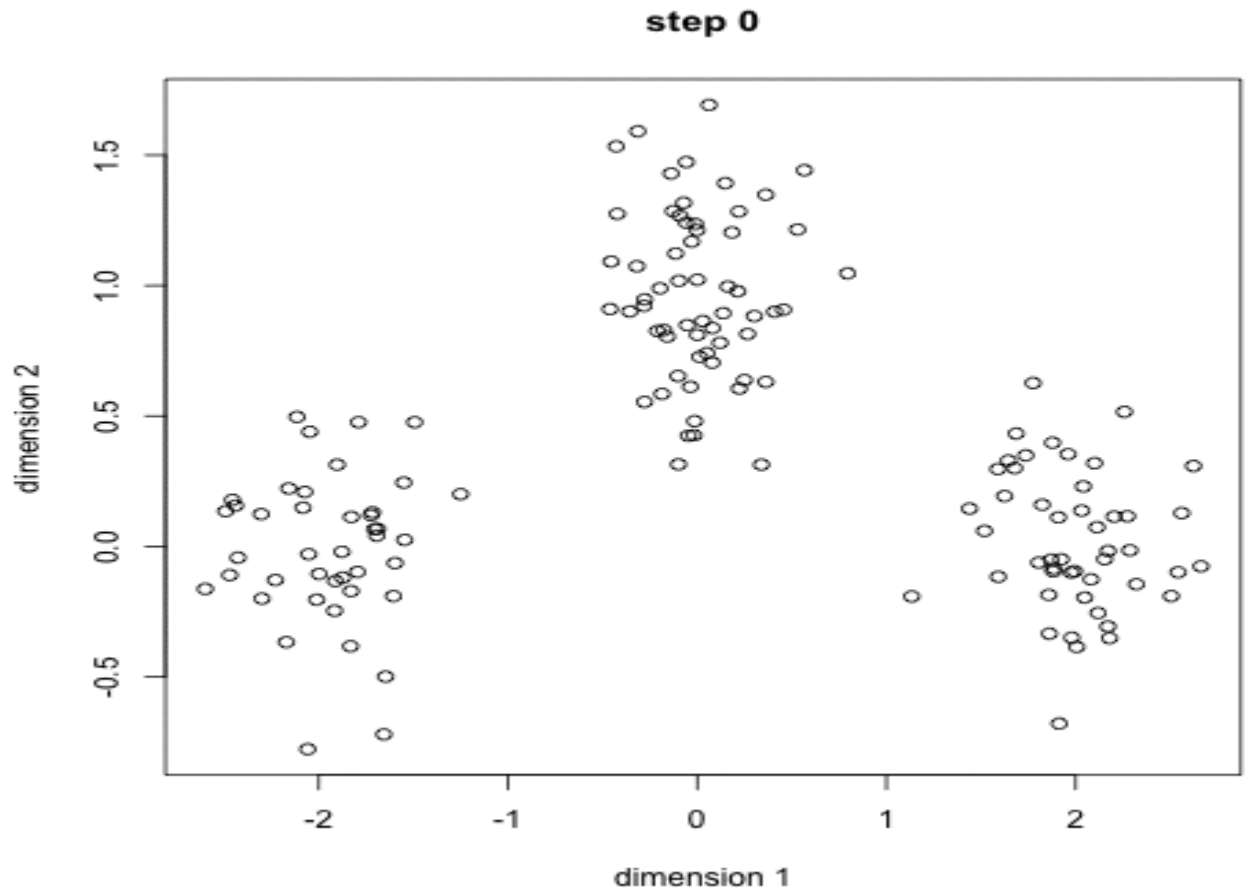
K-Means Clustering

K-Means is probably the most well-known clustering algorithm.

The K-means clustering algorithm computes centroids and repeats until the optimal centroid is found. It is presumptively known how many clusters there are. It is also known as the flat clustering algorithm. The number of clusters found from data by the method is denoted by the letter ‘K’ in K-means.



Step 1: Choose the number K of clusters

Step 2: Select random K points, the centroids (not necessarily from your dataset)

Step 3: Assign each data point to the closest centroid -> that forms K clusters

Step 4: Compute and place the new centroid of each cluster

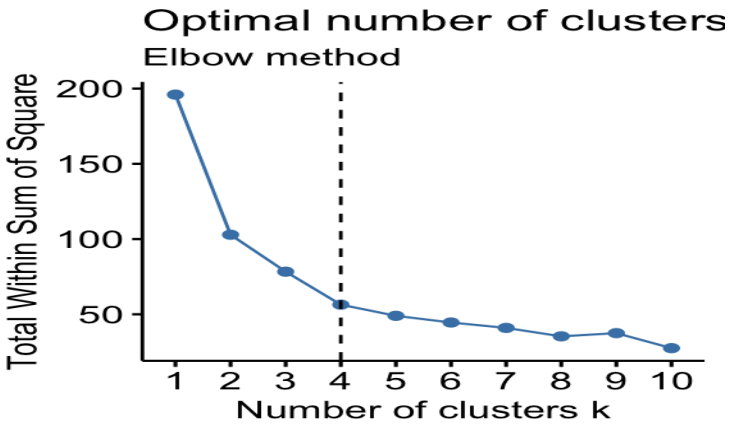
Step 5: Reassign each data point to the new closest centroid.

If any reassignment took place, go to step 4

Step 6: Your model is ready

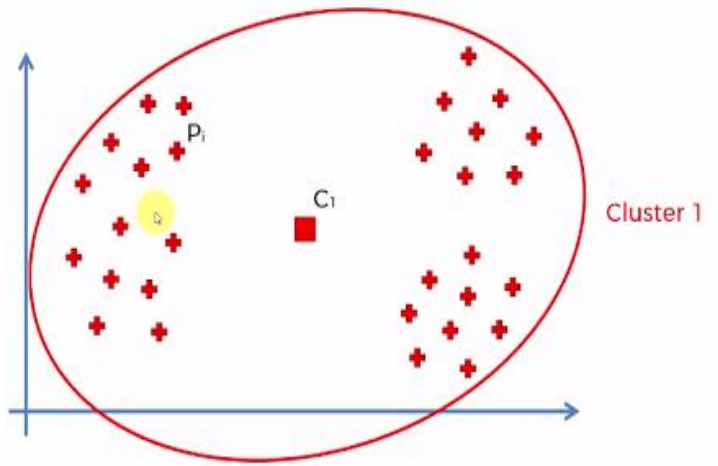
Height	Weight
185	72
170	56
168	60
179	68
182	72
188	77
180	71
180	70
183	84
180	88
180	67
177	76

Elbow Method: The Elbow method is a heuristic method of interpretation and validation of consistency within cluster analysis designed to help finding the appropriate number of clusters in a dataset.

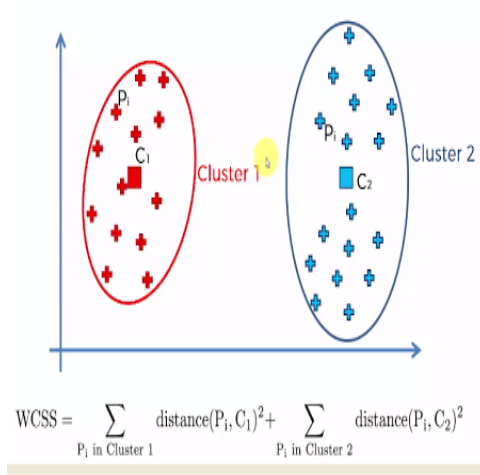


Within-Cluster-Sum-of-Squares (**WCSS**).

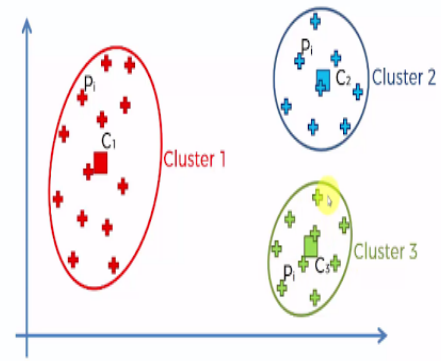
$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$



$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2$$



$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2$$



$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$