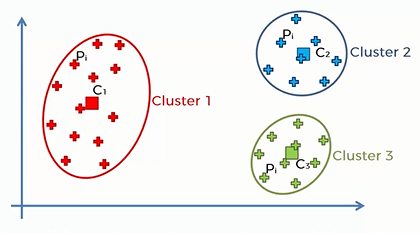
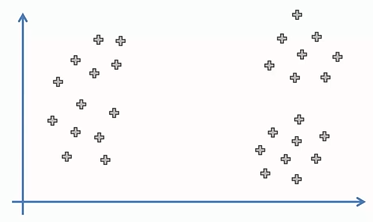
**CHOOSING THE RIGHT NUMBER OF CLUSTERS**

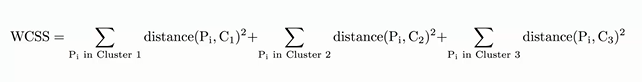
When we have a scattered plot like this then we may usually find three clusters for it.



But in order to understand if for instance two clusters would have been better or maybe 10 clusters would have been better, we need a certain metric or a way to understand and evaluate how certain number of clusters performs compared to a different number of clusters.

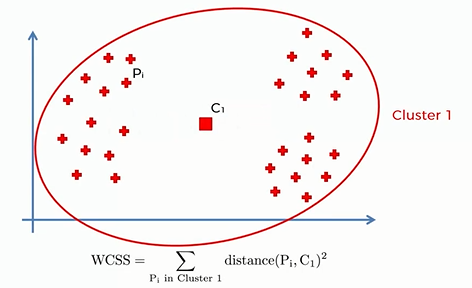
And such a **metric** is ‘within cluster sum of squares’, i.e. **wcss**.

This is the formula(for 3 clusters):

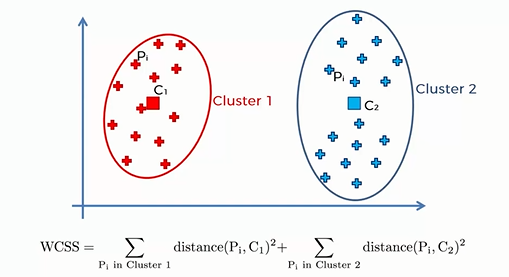


Now, let’s start from the beginning.

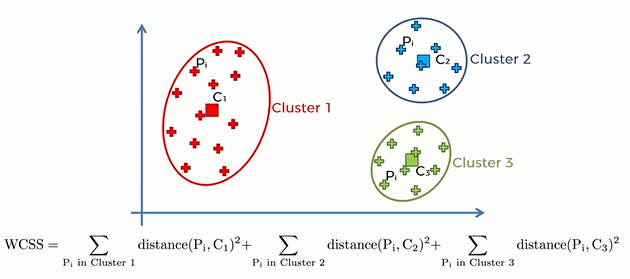
Suppose we take 1 cluster, then the graph and wcss is given by:

[distance between each point and centroid is calculated, squared and then added together.]

When we take 2 clusters:



When we take 3 clusters:



By calculations we will see that wcss is decreasing as we are increasing the number of clusters.

If we have reached the number of clusters equivalent to the number of points then

-> wcss will equate to zero

(because every single point has it’s cluster therefore has it’s own centroid and that centroid will be exactly where the point is. Therefore distance between point and centroid is zero, therefore wcss is zero.)

It is a good metric as it is constantly decreasing so it is contantly getting better.

Less the wcss or higher the number of clusters the better the goodness of fit.

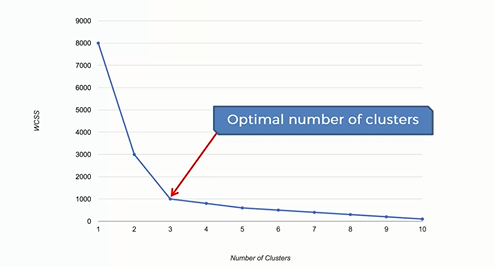
But how to find the optimal fitness of good or the right number of clusters?

-> For this ‘The Elbow Method ’ is used.

**The Elbow Method**

Here is a chart indicating different values of wcss for differnt number of clusters.

In the chart we look for that change or we look for that elbow where the drop goes from being quite substantial to being not as substantial. And that point in the chart is going to be the optimal number of clusters.



[This method is arbitrary]