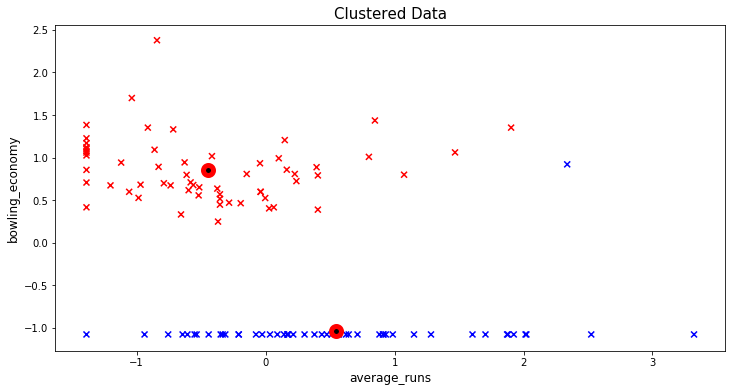
2.

This is the rationale for clustering – we do not know what related clusters or groups of data are present in the input and neither do we know what the prediction result is going to be classified into given a test point.

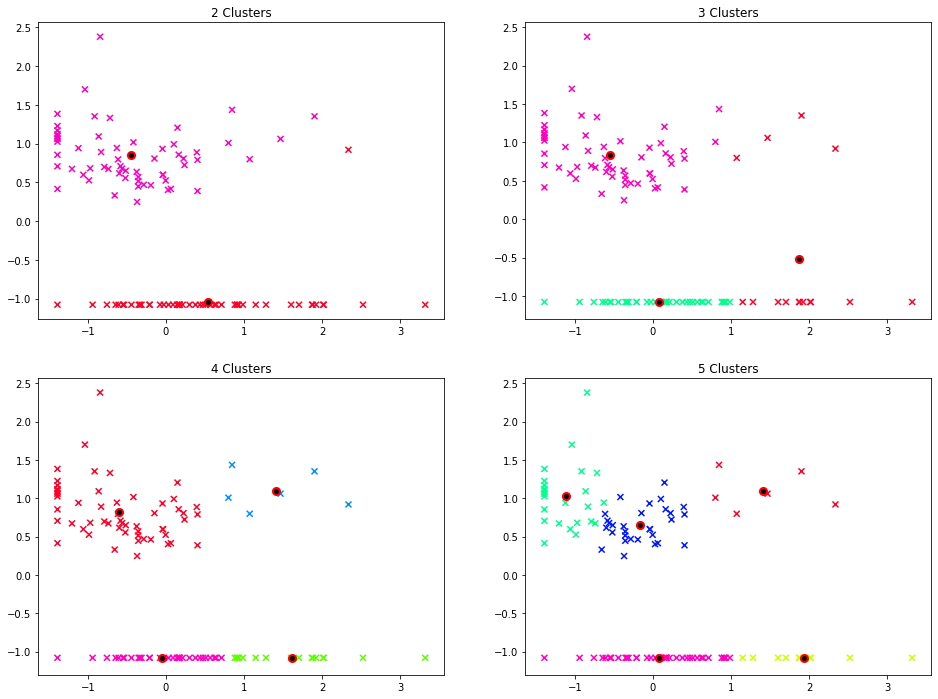


Interpreting with labels

Cluster-1 (Red) represents a cluster with all data where economy greater than the minimum economy in Cluster-2 (Blue) that tend to cluster together. There can also be an ambiguous 3rd cluster (and potentially 4th) if we halve the graph towards the right where the average runs scored can be classified/grouped again to be above some threshold (say ’1’ on the above normalized picture-graph).

Interpreting without labels or bias

* Cluster 1 and 2 are visually grouped into 2 clusters though it is not obvious and is ambiguous
* Both clusters have a lot of variability and do not indicate a very good quality clustering

3.

As explained in 2) for k=2, the results to interpret are cohesion and variation. Cohesion should be high while variation or dispersion should be low.

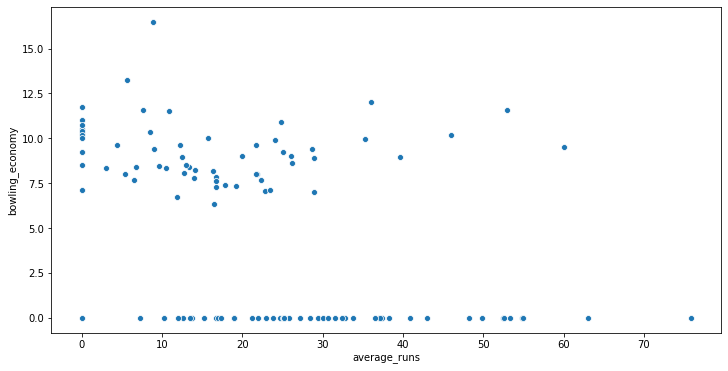
* For k=3 🡪 The third cluster (points in Red to the far right) is ambiguous and does not have good cohesion (centroid is placed badly) and has high variation (distance of each point from the centroid is large)
* For k=4 🡪 The ambiguous cluster from k=3 has been split in 2 for good and looks better than the previous clustering
* For k=5 🡪 There are 5 groups now where the groups 1 and 2 (green and blue) are split in 2 and this is again ambiguous as to whether it is a real meaningful cluster in the data or not

4.

The distortion function is a strict and monotonically decreasing function and we are minimizing it to find good clusters. Hence saying a cluster is good just based on distortion score is a fallacy.

The best cluster can be either 4/5 depending on what we are trying to do or achieve. If we are trying to segregate **all-rounders** (who can bat and bowl) based on average runs scored into LOW, MEDIUM and HIGH then 5 clusters make sense, else 4 would be a good choice for the clusters where the runs will be LOW and HIGH with average bowling economy. The other two clusters in both the cases being pure **batsmen** with LOW and HIGH average runs scored.

My choice would be 4 clusters focusing on average runs scored split into <35 and >35 slots while also segregating players into who can also bowl (>0 bowling economy).



The usefulness comes in decision making and during critical match situations –

* Identify key players (there are relatively few all-rounders with >35 average who also can bowl)
* Identify batting order (low scores and high scores on average)
* Know your team composition better to identify strengths and weaknesses and for potential recruits in future to balance the team