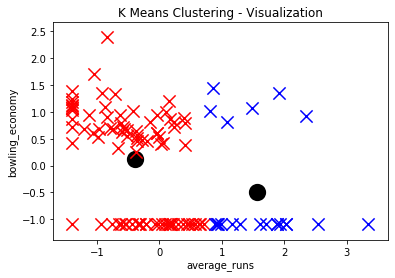
2. Write your own code for K-means algorithm using two attributes namely average\_runs and bowling\_economy. Take K=2. Plot clusters on a scatter plot with X and Y being the two attributes namely average\_runs and bowling\_economy, respectively. Color data points belonging to the first cluster with red and the second cluster with blue. Copy the plot diagram in the word document and interpret the output. [3 points]

Solution:

The following is the visualization of K Means Clustering for K=2 clusters.

It splits the normalized data into two clusters.



We can interpret the following from the above clusters

1. When the bowling economy is high the average runs made by the batsman is low. That means that most likely these cricketers are good bowlers and they bowl efficiently so their bowling economy is high.

Cluster Red – Low Average Runs and High Bowling Economy

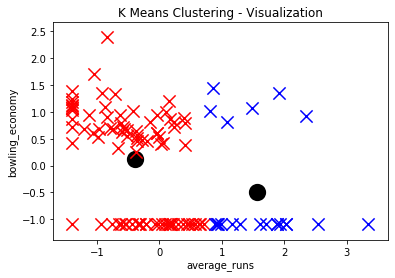
1. When the bowling economy is low that is greater than 0.5 and up till 1.5 the average runs made by the batsman is high.

Cluster Blue – High Average Runs and Low Bowling Economy

3. Redo question-2 on different values of K = 2,3,4,5. For each case, draw the plot of clusters as stated above. Visualize these plots, copy the plot diagrams in the word document,  and comment on which is better clustering (and reasons) based on visualization only. [3 points]

Solution:

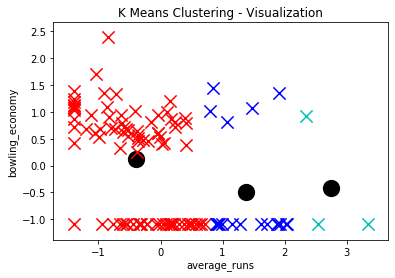
1. K=2



Cluster Red – Low Average Runs and High Bowling Economy

Cluster Blue – High Average Runs and Low Bowling Economy

1. K=3

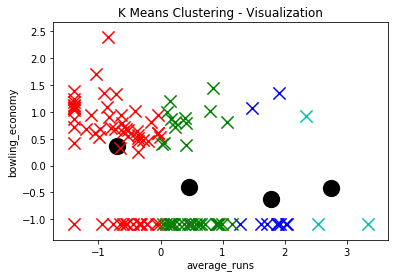


Cluster Red – Low Average Runs and High Bowling Economy

Cluster Blue – High Average Runs and Low Bowling Economy

Light Blue – Higher Average Runs and Low Bowling Economy

1. K=4



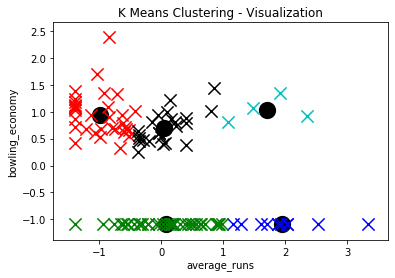
Red – Low Average Runs and High Bowling Economy

Cluster Green – High Average Runs and Low Bowling Economy

Blue – Higher Average Runs and Low Bowling Economy

Light Blue – Higher Average Runs when the Bowling Economy is low

1. K=5



Red – Low Average Runs and High Bowling Economy

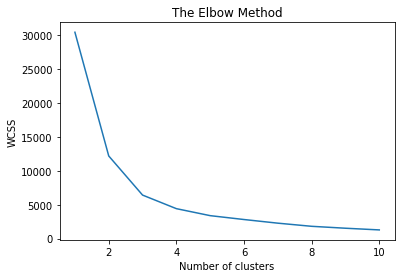
Cluster Green –Low Bowling Economy and High Average Runs (Higher Runs Batsmen)

Blue – Higher Average Runs and Low Bowling Economy (Highest Average Runs Batsmen who have low bowling economy)

Light Blue – Higher Average Runs when the Bowling Economy is high

Black – High Bowling Economy and Average of Average Runs

4. Write a few lines in a word document about the interpretation of the best clusters obtained. Also write a few statements about how these clusters can be useful. [2 points]



From the elbow method we can see that the ideal number of clusters is at 5. The Elbow method helps us determine the ideal number of clusters.

Each cluster is formed by calculating and comparing the distances of data points within a cluster to its centroid. An ideal way to figure out the right number of clusters would be to calculate the Within-Cluster-Sum-of-Squares (WCSS).

The clusters formed above can be useful in determining the following for K = 5

Cluster Red – Low Average Runs and High Bowling Economy

Cluster Green –Low Bowling Economy and High Average Runs

Cluster Blue – Higher Average Runs and Low Bowling Economy

Cluster Light Blue – Higher Average Runs when the Bowling Economy is high

Cluster Black – High Bowling Economy and Average of Average Runs

With the above deductions we can according to the cluster classify them into

1. Batsmen b) Bowlers c) All rounders who can bat and bowl d) All rounders who have high runs average and low bowling economy – (Batsmen category 1) e) All rounders who have low runs average and high bowling economy (Batsmen category 2)

So the Team can divide their cricketers into 5 classes and design different strategies for a different type of cricketers to increase their bowling economy, average runs and other factors.

The clusters as we increase the K value we see the following

1. For K=2 there is a wide separation of two clusters and they broadly classify the average runs and bowling economy statistics
2. As we move with K=3 the clusters evolve and the centroid also starts moving into the direction of where the data resides
3. When K = 4 the clusters clearly show that there are different types of interpretation as previously made available that can be interpreted
4. K=5 is the optimum cluster number that shows different kinds of clusters. The Elbow method also confirms that the cluster = 5 is the ideal number for the clusters to be segregated into different categories.