Q2. 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]



For K=2

Interpretation:

* Chosen K=2, which means grouping the entire data into 2 clusters based on average\_runs and bowling\_economy.
* Purple and Red clusters are created by the KNN algorithm.
* Purple cluster is group of bowlers with all kinds of averages.
* Red cluster has both allrounders and batsmen grouped.

Q3. 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]



For K=3

Interpretation:

* Chosen K=3, which means grouping the entire data into 3 clusters based on average\_runs and bowling\_economy.
* Purple, Red and Yellow clusters are created by the KNN algorithm.
* Purple cluster is group of bowlers with all kinds of averages.
* Red and Yellow clusters has both allrounders and batsmen grouped.

For K=4

Interpretation:

* Chosen K=4, which means grouping the entire data into 4 clusters based on average\_runs and bowling\_economy.
* Green, Purple, Red and Yellow clusters are created by the KNN algorithm.
* Green cluster is group of bowlers with all kinds of averages.
* Purple, Red and Yellow clusters has both allrounders and batsmen grouped.
* **Yellow clusters are bowlers who can bat a bit and Red cluster are batsmen who can bowl.**

For

K=5

Interpretation

* Chosen K=5, which means grouping the entire data into 5 clusters based on average\_runs and bowling\_economy.
* Yellow cluster is group of bowlers with all kinds of averages.
* Compared to the clustering done when K=4, a new orange cluster got created. Do not see any significance for this cluster, apart from that those are only batsmen who doesn’t bowl.

Q4. 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]

For K=2 and K=5, there is either too less information or too much more information. K=2 only provides the classification on a high level whereas K=5 is creating too many clusters.

Ideally, when K=3 and K=4, there is clear segregation of the batsmen whether they are bowlers, batsmen or batting allrounders or bowling allrounders.