Consider the dataset (attached below) comprising of 109 cricket players (filename is ‘cricketers.csv’). Each row comprises of following details of a cricketer:

1. matches\_played: number of matches played by the player,
2. innings\_batted: number of times got the chance to bat,
3. runs\_scored: total runs scores across all innings,
4. highest\_runs: highest run scored taking all innings in account,
5. ball\_faced: total balls faced across all innings,
6. average\_runs: average runs scores taking all innings in account,
7. strike\_rate: it is computed as (runs\_scored / balls\_faced) x 100,
8. innings\_bowled: number of times got the chance to bowl,
9. overs: number of overs bowled across all innings,
10. runs\_given: total run given across all innings,
11. wickets\_obtained: total wickets obtained across all innings,
12. average\_runs\_per\_wicket: it is computed as (runs\_given / wickets\_obtained)
13. bowling\_economy: average runs scored per over

Questions:

Notes: Write all code related to this assignment in single jupyter notebook. Implement K-means algorithm from scratch, do not use any predefined ML library function for K-means algorithm. Marks will be deducted if any library function is used for KMeans. Use single word document to answer questions.

1. Given that K-means depends on distance metric, it is a convention to normalize the data attributes so that attributes are on the same scale.  So, in this first task, normalize all data attributes. [2 points]

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]

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]

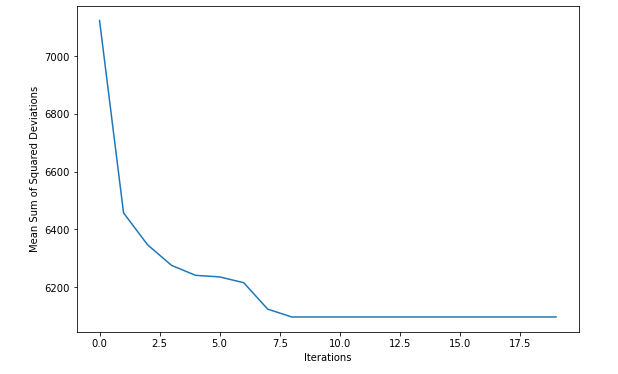
4. Write a few lines in a word document about the interpretation of the best clusters obtained. Also write a few statements about how these clusters can be useful. [2 points]

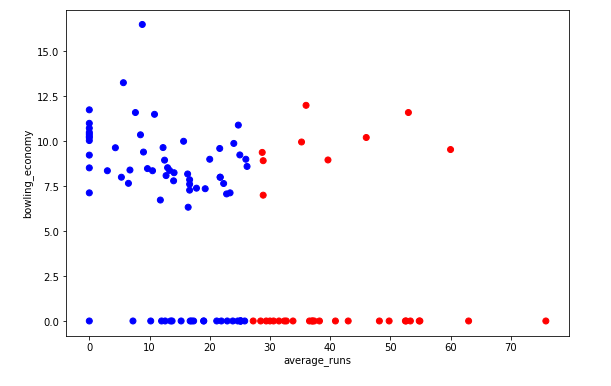
Deliverables:

One jupyter notebook for Q 1,2,3.

One word document for Q 2,3,4.

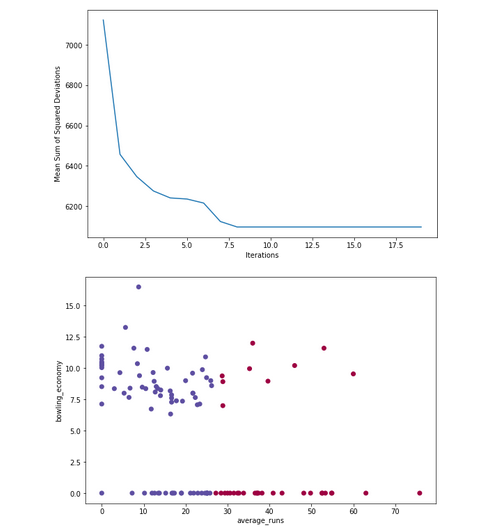
Dataset: [cricketers.csv](https://bits-pilani.instructure.com/courses/872/files/158789?wrap=1)[download](https://bits-pilani.instructure.com/courses/872/files/158789/download?download_frd=1)

2. 



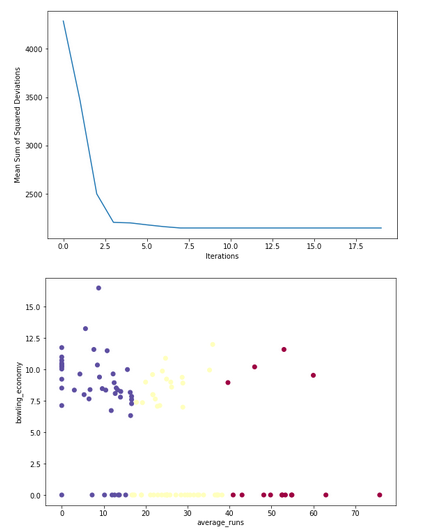
From above plot, we see kmeans has divided data into 2 clusters. With increasing average\_runs, bowling economy of more cricketers are 0, which may mean they are batsman. Not much can be interpreted overall.

3. K=2



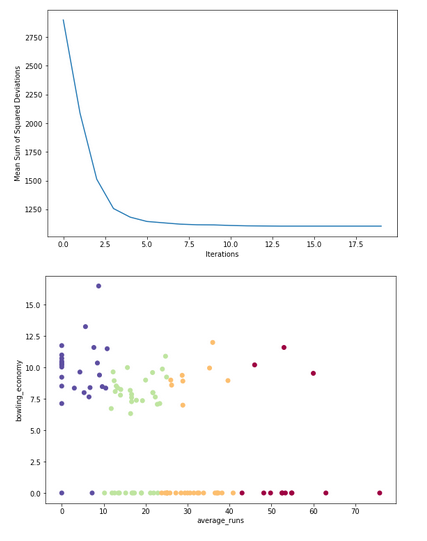
With increasing average\_runs, bowling economy of more cricketers are 0, which may mean they are batsman

K=3

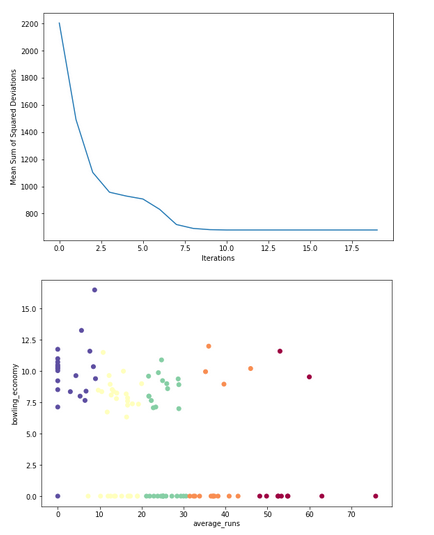


First cluster shows mostly bowlers who have low average runs and good bowling economy. Second cluster shows 20+ batting average and some good bowling economy- probably allrounders. Third cluster shows more batsman with good batting average and few allrounders with average bowling economy.

K=4



K=5



Cluster 4 and 5 are good clustering predictions. First cluster shows mostly bowlers who have very low average\_runs score and good bowling economy except one or two. Second and third clusters shows 10+/20+ batting average and some good bowling economy- probably allrounders. Fourth cluster shows more batsman with good batting average and few allrounders with average bowling economy. Fifth cluster shows great batsman.

4. Cluster 4 and 5 are good clustering predictions. First cluster shows mostly bowlers who have very low average\_runs score and good bowling economy except one or two. Second and third clusters shows 10+/20+ batting average and some good bowling economy- probably allrounders. Fourth cluster shows more batsman with good batting average and few allrounders with average bowling economy. Fifth cluster shows great batsman.

These clusters can be useful in categorizing among good batsman and good bowlers and good overall allrounders.