Functions and their use

dataset(filename)

* To read data from csv file.
* Dividing raw data into usable data with requested columns and header
* Cleaning of data
* Returns data and header file

create\_centroids(dataset)

* Randomly creates k number of centroids out of data points on given data set

compute\_euclidean\_distance(point, centroid)

* Returns distance between given centroid and point using standard distance formula/Euclidean distance

assign\_label\_cluster(distance, data\_point, centroids)

* Takes data point, centroids, distance which is a dictionary {index of centroid 🡪 distance of centroid from data\_point }
* Checks the dictionary distance for minimum distance
* Whichever centroid is closest to that data point, it returns the index of that centroid so that point can be clustered to that centroid.
* Returns [index of centroid with minimum distance from that data point, data point ]

Compute\_new\_centroids(cluster\_label, centroids)

* Compute new centroid function takes (data point, concerning centroid).
* It returns an array of new data point which becomes new centroid.

iterate\_k\_means(data points, centroids, total\_iteration)

* Major function that actually performs the clustering
* It performs specified number of iterations to get as close as possible to the actual centroid of a cluster
* For every iteration, it checks on each point of the dataset and calculates its distance from each centroid.
* Then it associates that point with one out of k centroids.
* Now the location of centroid is changed to the mean of that data point and centroid using compute new centroid function. Refer Compute\_new\_centroid above
* This process continues for all points for total\_iteration number of times

print\_label\_data(list)

* Accepts list of cluster label and new centroids such that cluster label itself is a 2 dimensional list with (centroid label, associated data point) on each index and centroids is the list of final centroids
* It is not necessarily to be called but provides clear picture of data points associated.

working(filename)

* This is the similar to main function where the execution takes place and clusters are assigned to different cluster variables that will be returned.
* From outside it just accepts datasets and returns 3 clusters.
* It receives the data from iterate\_k\_means function and process it in user readable format
* I have made all the print and scatterplot functions as comments. Each cluster is color coded on graph to get clear understanding. Initial and final centroids can be distinguished To get details remove the ‘#’