2. Here I am going to explain one of the clustering technique used in regression operation for python coding.

Hierarchical clustering:-

This is one of the technique used to perform clustering which is used for plotting data on linear graph depending on raw data available at input. The major advantage is you do not need to know number of clusters in given data in advance. User can perform hierarchical clustering without knowing total numbers of clusters. There are few functions cuts hierarchical clusterings into flat clustering or provides the root of cluster by performing flat clustering on each given cluster. Those functions are,

\* fcluster:- forms flat clusters given from hierarchical clusters defined by given linkage matrix.

\* fclusterdata:- Cluster observation data using given metric

\* linkage:- Performs hierarchical/agglomerative clustering.

\* weighted:- Performs weighted/WPGMA linkage on given condensed distance matrix.

\* median:- Performs median/WPGMC linkage.

\* ward:- Performs ward's linkage on a condensed distance matrix

Syntax:- **from** **scipy.cluster.hierarchy** **import** fcluster

This is the syntax to use scipy cluster hierarchy in python file.

Example:-

*# generate the linkage matrix*

Z = linkage(X, 'ward')

here Ward's method is special case of objective function approach where criterion to choose pair of cluster to merge at each step is based on optimal value of objective function. The objective function can be anything which reflects investigator purpose. This is the method to calculate distance between newly formed clusters. Ward method causes linkage to use ward variance algorithm method.

**from** **scipy.cluster.hierarchy** **import** cophenet

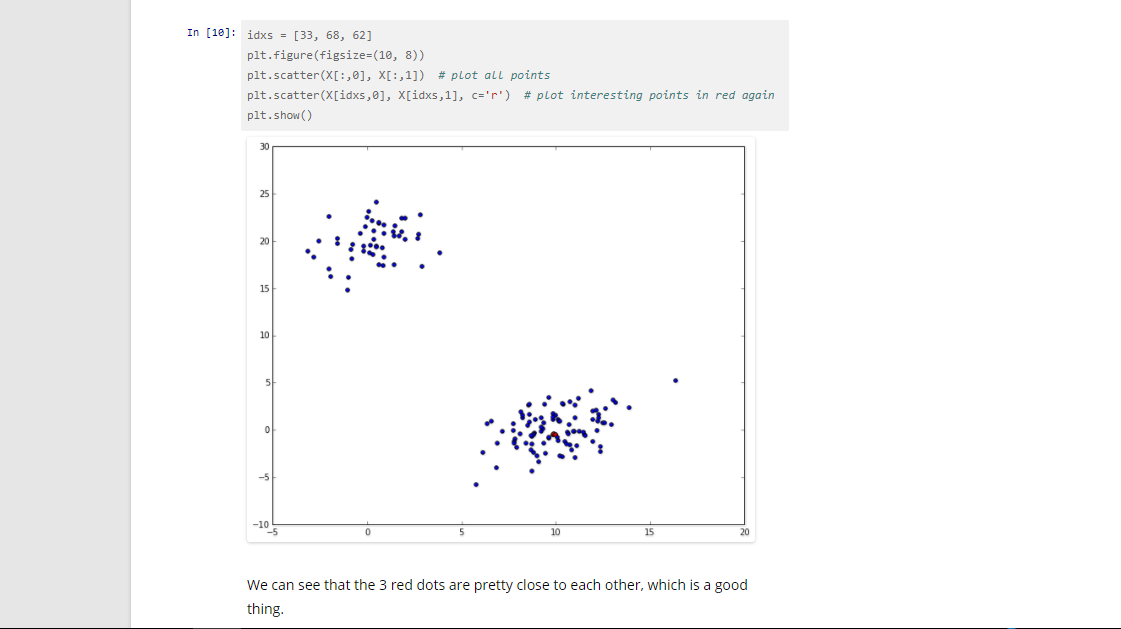
**from** **scipy.spatial.distance** **import** pdist

This function will give output of some decimal value such as 0.9864585584564576. No matter what the linkage function will use that method and metric to calculate distance of the clusters of the clusters and merge with cluster in each iteration.

array([ 14. , 79. , 0.05914, 2. ])

In the second iteration we have decided to merge two clusters, one is original and other one is new with indices 14 and 79 which had a distance of 0.04789. This again formed another cluster with a total of two samples.

Result of hierarchical clustering:-



Disadvantages of Hierarchical clustering:-

1. Cannot undo the work what done before(divisive work on sub clusters, cannot remerge)

2. Every split or merge must be redefined.

3. Methods may not scale well, checking all possible pairs, complexity goes high.