This assignment covers topics belonging to clustering. All the questions are mandatory.

**Question 1**. **(30 Points)** Suppose that the data mining task is to cluster points (with (*x*, *y)* representing location) into three clusters, where the points are

*A*1 (2,10), *A*2 (2,5), *A*3 (8,4), *B*1 (5,8), *B*2 (7,5), *B*3 (6,4), *C*1 (1,2), *C*2 (4,9).

The distance function is Euclidean distance. Suppose initially we assign *A*1, *B*1, and *C*1 as the center of each cluster, respectively. Use the *k-means* algorithm to show *only*

(a) The three cluster centers after the first round of execution.

(b) The final three clusters. (up to 3 iterations only)

**Question 2**. **(20 points)** Both *k-means* and *k-medoids* algorithms can perform effective clustering.

(a) Illustrate the strength and weakness of *k-means* in comparison with *k-medoids*.

(b) Illustrate the strength and weakness of these schemes in comparison with a hierarchical clustering scheme (e.g., AGNES).

**Question 3**. **(20 points)** This question may require you to do some research on contrasting different clustering algorithms.

Describe each of the following clustering algorithms in terms of the following criteria:

(1) Shapes of clusters that can be determined; (2) input parameters that must be specified; and (3) limitations.

(a) *k*-means

(b) *k*-medoids

(c) CHAMELEON

(d) DBSCAN

**Question 4**. **(30 points)** Implement K-Means clustering algorithm for the dataset (**buddymove.csv**) provided with this assignment. You can run the algorithm three times for the values of K =3, K=4, and K=5 for each run (here K refers to # of clusters). A sample code file (**KMeans Clustering.py**) has been attached with this assignment. The description of data set is given below:

1. **Data Set Information:**

This dataset was populated from destination reviews published by 249 reviewers of holidayiq.com till October 2014. Reviews falling in 6 categories among destinations across South India were considered and the count of reviews in each category for every reviewer (traveler) is captured. More information about this dataset is available at the following link:

<https://archive.ics.uci.edu/ml/datasets/BuddyMove+Data+Set>

1. **Attribute Information:**

Attribute 1 : Unique user id  
Attribute 2 : Number of reviews on stadiums, sports complex, etc.  
Attribute 3 : Number of reviews on religious institutions  
Attribute 4 : Number of reviews on beach, lake, river, etc.  
Attribute 5 : Number of reviews on theatres, exhibitions, etc.  
Attribute 6 : Number of reviews on malls, shopping places, etc.  
Attribute 7 : Number of reviews on parks, picnic spots, etc.

You might only need to consider Attributes 2 through 7, because User\_id can be dropped. The idea is to cluster the Users that have similar characteristics of posting a # of reviews. You can implement the algorithm in any programming language you want. However, I highly recommend you use Python, R, or Matlab. Your program for question 4 should have the following:

1. Attach a snapshot of each run of your program (for different values of K) that prints clustering labels for each run.
2. Your code should also print visualization of each cluster. Research on how to create a visualization for clusters.
3. Submit the code file along with your submission.