Assignment 8

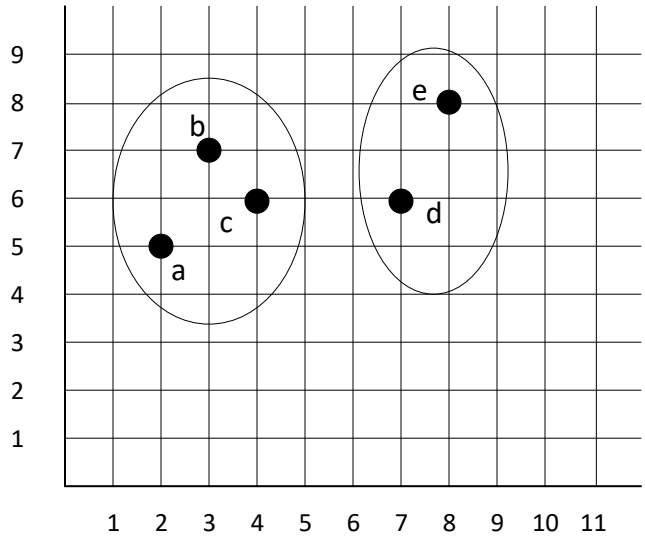
Due: 11/16

Problem 1 (10 points). The k-means algorithm, which we discussed in the class, is being run on a small two-dimensional dataset. After a certain number of iterations, you have two clusters as shown below:

ID	x	y	Cluster
a	3	6	Cluster 1
b	7	2	Cluster 1
c	5	6	Cluster 1
d	4	5	Cluster 1
e	1	5	Cluster 2
f	5	3	Cluster 2
g	6	7	Cluster 2

Run one more iteration of the k-Means clustering algorithm and show the two clusters at the end of the iteration. Use Manhattan distance when calculating the distance between objects.

Problem 2 (10 points). Consider the following two clusters:



(1). Compute the distance between the two clusters using the average method that we discussed in the class. Use the Euclidean distance measure when calculating the distance between objects.
(2). Compute the distance between the two clusters using the Ward's method that we discussed in the class. Use the Euclidean distance measure when calculating the distance between objects.

Problem 3 (10 points). *Use hw8_p3.csv file*.

(1). Standardize all columns using the *scale* function.
(2). Run the k-means algorithm using *fviz_nbclust* 10 times (this is default) with the *wss* method.
(3). Show all 10 *wss* values.
(4). Plot *wss-k* graph. Can you see an obvious *elbow*?
(5). Run the k-means algorithm using *fviz_nbclust* 10 times (this is default) with *silhouette* method and plot *silhouette-k* graph. What is the optimal number of clusters?
(6). Run the k-means algorithm once more with the following code segment:

```
set.seed(31)
k3 <- kmeans(task.df, centers = 3, nstart = 25)
```

From the clustering result, prepare the following table, which is a cluster profile:

		Cluster		
		1	2	3
Weight	mean			
	max			
	min			
Turning_Circle	mean			
	max			
	min			
Displacement	mean			
	max			
	min			
Horsepower	mean			
	max			
	min			
Gas_Tank_Size	mean			
	max			
	min			

Note that you may not get all values in the above table using the example R code we used in the class. You may have to write your own code segment to obtain all necessary values.

Problem 4 (10 points). Use *hw8_p4.csv*.

Run the *pam* (k-medoids) algorithm using the *Gower* distance 9 times with k = 2 through 10 and plot the *silhouette-k* graph. What is an optimal *k*?

Problem 5 (10 points). Use *hw8_p5.csv*

(1). Run the *agnes* hierarchical clustering algorithm using the *Gower* distance 4 times with *average*, *single*, *complete*, and *ward* methods.
(2). Collect agglomerative coefficients and plot a bar graph of the coefficients. Which method has the highest coefficient?
(5). Run the *agnes* algorithm once more with the best method you identified in the above Problem 5-(2).
(6). Cut the tree with k = 3.
(7). Show the row numbers for each cluster, which shows cluster membership. The format of your answer should look like:

> Cluster 1: rows {1, 5, 22, 30, … }
> Cluster 2: rows {2, 15, 17, 41, … }
> Cluster 3: rows {8, 52, 53, 61, … }
>
> Note: The row numbers shown above are random numbers just as an example (not actual cluster membership).

You may have to write your own code segment to obtain the above result.

Submission:

Include all answers in a single file and name it *LastName_FirstName_hw*8.EXT. Here, "EXT" is an appropriate file extension (e.g., docx or pdf). If you have multiple files, then combine all files into a single archive file and name it *LastName_FirstName_hw*8.EXT. Here, "EXT" is an appropriate archive file extension (e.g., zip or rar). Upload the file to Blackboard. You also need to submit your R code file.