Tutorial 10

1. Suppose we have data for five objects on two features:

object	x_1	x_2
A	1	1
B	1.5	2
C	3	4
D	3.5	5
E	4.5	5

   We set $k = 2$ to cluster the six data points into two clusters, $\mathcal{P}$ and $\mathcal{Q}$, and initialize the algorithm with the centroids $(x_{1,\mathcal{P}}, x_{2,\mathcal{P}}) = (2, 2)$ and $(x_{1,\mathcal{Q}}, x_{2,\mathcal{Q}}) = (4, 4)$.

   (a) Fill up the following table to identify the objects in each cluster during the first iteration of the k-means algorithm:

cluster	**object(s)**
$\mathcal{P}$	
$\mathcal{Q}$	

   (b) Compute the new centroids for the two clusters based on cluster assignment in (a).

   (c) Based on the centroids computed in (b), identify the objects in each cluster during the second iteration of the k-means algorithm.

   (d) Calculate the Within Sum of Squares (WSS) for the clustering assignment in (c).

2. (K-Means) Consider data set `hdb-2012-to-2014.csv` which was extracted from the published data [1]. The file has information on the HDB resale flats from Jan 2012 to Dec 2014.

   (a) Load data into R. Use k means algorithm to pick an optimal value for k in term of WSS, based on two variables, `resale_price` and `floor_area_sqm`.

   (b) With the optimal k in part (a), plot the data points in the k clusters determined.

[1] `https://data.gov.sg/dataset/resale-flat-prices`