Name: CHAN King Yeung
SID: 1155119394
SEEM2460 Assignment 3

Question 1

a) We constructed data matrix as follow,

$$\begin{pmatrix} A & B & C & D & E & F & G & H & I \\ 1 & 2 & 5 & 4 & 4 & 2 & 5 & 6 & 3 \\ 2 & 2 & 4 & 4 & 5 & 3 & 3 & 4 & 5 \end{pmatrix} \begin{matrix} Customer \\ Income \\ Age \end{matrix}$$

Then we calculated the distances based on the formula of Euclidean distance

$$d(C_{centroid}, cust_i) = \sqrt{\left(C_{income} - cust_{i,income}\right)^2 + \left(C_{age} - cust_{i,age}\right)^2} \text{ for } i \in [A, I]$$

We used K-means algorithm and obtained the following distance matrices

$$D^0 = \begin{pmatrix} 0 & 1 & 4.47 & 3.61 & 4.24 & 1.41 & 4.12 & 5.39 & 3.61 \\ 3.61 & 2.83 & 1 & 0 & 1 & 2.24 & 1.41 & 2 & 1.41 \\ 5.39 & 4.47 & 1 & 2 & 2.24 & 4.12 & 1.41 & 0 & 3.16 \end{pmatrix} \begin{matrix} C_1 = (1, 2) \\ C_2 = (4, 4) \text{ with } SSE = 6.93 \\ C_3 = (6, 4) \end{matrix}$$

Customer A, B, and F are assigned into cluster 1; Customer C, D, E, G, and I are assigned into cluster 2; Customer H is assigned into cluster 3;

 # note here: customer C and G have same distance between cluster 2 and 3, but we assigned them into cluster 2

We computed new centroids and repeated the above steps.

$$D^1 = \begin{pmatrix} 0.75 & 0.47 & 3.73 & 2.87 & 3.54 & 0.75 & 3.4 & 4.64 & 2.98 \\ 3.88 & 3.11 & 0.82 & 0.28 & 0.82 & 2.51 & 1.44 & 1.81 & 1.44 \\ 5.39 & 4.47 & 1 & 2 & 2.24 & 4.12 & 1.41 & 0 & 3.16 \end{pmatrix} \begin{matrix} C_1 = (1.67, 2.33) \\ C_2 = (4.20, 4.20) \text{ with } SSE = 4.33 \\ C_3 = (6.00, 4.00) \end{matrix}$$

Customer G no longer belongs to cluster 2 and being assigned into cluster 3

$$D^2 = \begin{pmatrix} 0.75 & 0.47 & 3.73 & 2.87 & 3.54 & 0.75 & 3.4 & 4.64 & 2.98 \\ 3.91 & 3.2 & 1.12 & 0.5 & 0.5 & 2.5 & 1.8 & 2.06 & 1.12 \\ 4.74 & 3.81 & 0.71 & 1.58 & 2.12 & 3.54 & 0.71 & 0.71 & 2.92 \end{pmatrix} \begin{matrix} C_1 = (1.67, 2.33) \\ C_2 = (4.00, 4.50) \text{ with } SSE = 2.67 \\ C_3 = (5.50, 3.50) \end{matrix}$$

Customer C no longer belongs to cluster 2 and being assigned into cluster 3

$$D^3 = \begin{pmatrix} 0.75 & 0.47 & 3.73 & 2.87 & 3.54 & 0.75 & 3.4 & 4.64 & 2.98 \\ 3.77 & 3.14 & 1.49 & 0.75 & 0.47 & 2.36 & 2.13 & 2.43 & 0.75 \\ 4.64 & 3.73 & 0.47 & 1.37 & 1.89 & 3.4 & 0.75 & 0.75 & 2.69 \end{pmatrix} \begin{matrix} C_1 = (1.67, 2.33) \\ C_2 = (3.67, 4.67) \text{ with } SSE = 2.67 \\ C_3 = (5.33, 3.67) \end{matrix}$$

We stopped here since there is no more new assignment to cluster.

Cluster 1: customer A, B, and F; Cluster 2: customer D, E, and I; Cluster 3: customer H, G, and C

b) Partial source codes are pasted here for your convenience

```
# Step-1: Assign 3 initial centroids.
centroids = {
    # Please specify three centroids below
    1: [1, 2], 2: [4, 4], 3: [6, 4],
}

# Step-2: Continue until all assigned categories don't change any more
df = assignment(df, centroids)
while True:
    closest_centroids = df['closest'].copy(deep=True)
    # Please determine the order of update() and assignment() below, Hint: 2
lines of codes
    df = assignment(df, centroids)
    centroids = update(centroids)
    if closest_centroids.equals(df['closest']):
        break
```

My source code for your reference

https://colab.research.google.com/drive/1Wr3nHm4VQkbkvjWil3yIPEL-8v5yJIhx

---