Name: SUN RUI

Student ID: 18083229g

**Answer 1:**

a)

sort the column Age: [17, 18, 18, 22, 22, 27, 29, 34, 35, 38, 39, 39, 46, 54, 59]

sort the column Monthly Income: [3000, 4000, 7500, 7800, 7800, 7900, 8500, 14000, 18000, 21000, 24700, 30000, 31000, 31110, 40500]

sort the column Service Plan: [100, 100, 100, 200, 200, 400, 600, 600, 600, 600, 800, 1000, 1600, 1600, 1600]

sort the column Extra Usage: [0, 0, 0, 7, 25, 31, 31, 50, 54, 64, 211, 254, 290, 303, 311]

Equal-width:

Age: (59-17)/3=14 → a=[17, 31), b=[31, 45), c=[45, 59]

Monthly Income: (40500-3000)/3=12500 → x=[3000, 15500), y=[15500, 28000), z=[28000, 40500]

Service Plan: (1600-100)/3=500 → A=[100, 600), B=[600, 1100), C=[1100, 1600]

Extra Usage: (311-0)/3=103.67 →X=[0, 103.67), Y=[103.67, 207.34), Z=[207.34, 311.01]

Change the data into below table according to above ranges:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **(0)** | **(i)** | **(ii)** | **(iii)** | **(iv)** | **(v)** | **(vi)** |
| **Ref** | **Age** | **Sex** | **Monthly Income** | **Marital**  **Status** | **Service**  **Plan** | **Extra**  **Usage** |
| 1 | 54 c | FEMALE | 3000 x | YES | 100 A | 0 X |
| 2 | 59 c | FEMALE | 4000 x | NO | 600 B | 54 X |
| 3 | 38 b | MALE | 7800 x | NO | 200 A | 31 X |
| 4 | 18 a | FEMALE | 8500 x | NO | 600 B | 311 Z |
| 5 | 27 a | MALE | 14000 x | YES | 100 A | 211 Z |
| 6 | 29 a | FEMALE | 31000 z | YES | 1600 C | 25 X |
| 7 | 17 a | MALE | 7500 x | NO | 600 B | 254 Z |
| 8 | 22 a | FEMALE | 7900 x | NO | 200 A | 31 X |
| 9 | 34 b | MALE | 24700 y | NO | 100 A | 7 X |
| 10 | 46 c | FEMALE | 31110 z | YES | 600 B | 0 X |
| 11 | 39 b | FEMALE | 21000 y | YES | 800 B | 64 X |
| 12 | 35 b | FEMALE | 30000 z | NO | 1600 C | 0 X |
| 13 | 39 b | MALE | 40500 z | YES | 1600 C | 50 X |
| 14 | 18 a | MALE | 7800 x | NO | 1000 C | 290 Z |
| 15 | 22 a | MALE | 18000 y | YES | 400 A | 303 Z |

initial cluster centers：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Locus | | | | | |
| **Ref** | **Age** | **Sex** | **Monthly Income** | **Marital**  **Status** | **Service**  **Plan** | **Extra**  **Usage** |
| **1**(1) | 54 c | FEMALE | 3000 x | YES | 100 A | 0 X |
| **2**(8) | 22 a | FEMALE | 7900 x | NO | 200 A | 31 X |
| **3**(15) | 22 a | MALE | 18000 y | YES | 400 A | 303 Z |

cluster distances of first step:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **(0)** | **(i)** | **(ii)** | **(iii)** | **(iv)** | **(v)** | **(vi)** | **distance** | | |
| **Ref** | **Age** | **Sex** | **Monthly Income** | **Marital**  **Status** | **Service**  **Plan** | **Extra**  **Usage** | **1** | **2** | **3** |
| 1 | 54 c | FEMALE | 3000 x | YES | 100 A | 0 X | **0** | 2 | 4 |
| 2 | 59 c | FEMALE | 4000 x | NO | 600 B | 54 X | **2** | 2 | 6 |
| 3 | 38 b | MALE | 7800 x | NO | 200 A | 31 X | 3 | **2** | 4 |
| 4 | 18 a | FEMALE | 8500 x | NO | 600 B | 311 Z | 4 | **2** | 4 |
| 5 | 27 a | MALE | 14000 x | YES | 100 A | 211 Z | 3 | 3 | **1** |
| 6 | 29 a | FEMALE | 31000 z | YES | 1600 C | 25 X | **3** | 3 | 4 |
| 7 | 17 a | MALE | 7500 x | NO | 600 B | 254 Z | 5 | **3** | 3 |
| 8 | 22 a | FEMALE | 7900 x | NO | 200 A | 31 X | 2 | **0** | 4 |
| 9 | 34 b | MALE | 24700 y | NO | 100 A | 7 X | 4 | **3** | 3 |
| 10 | 46 c | FEMALE | 31110 z | YES | 600 B | 0 X | **2** | 4 | 5 |
| 11 | 39 b | FEMALE | 21000 y | YES | 800 B | 64 X | **3** | 4 | 4 |
| 12 | 35 b | FEMALE | 30000 z | NO | 1600 C | 0 X | 4 | **3** | 6 |
| 13 | 39 b | MALE | 40500 z | YES | 1600 C | 50 X | **4** | 5 | 4 |
| 14 | 18 a | MALE | 7800 x | NO | 1000 C | 290 Z | 5 | **3** | 3 |
| 15 | 22 a | MALE | 18000 y | YES | 400 A | 303 Z | 4 | 4 | **0** |

update cluster centers:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Locus | | | | | |
| **Ref** | **Age** | **Sex** | **Monthly Income** | **Marital**  **Status** | **Service**  **Plan** | **Extra**  **Usage** |
| **1**(1) | 54 c | FEMALE | **z** | YES | **B** | 0 X |
| **2**(8) | 22 a | **MALE** | 7900 x | NO | 200 A | 31 X |
| **3**(15) | 22 a | MALE | 18000 y | YES | 400 A | 303 Z |

Equal-depth:

Age: a={17,18,18,22,22}, b={27,29,34,35,38}, c={39,39,46,54,59}

Monthly Income: x={3000,4000,7500,7800,7800}, y={7900,8500,14000,18000,21000}, z={24700,30000,31000,31110,40500}

Service Plan: A={100,100,100,200,200}, B={400,600,600,600,600}, C={800,1000,1600,1600,1600}

Extra Usage: X={0,0,0,7,25}, Y={31,31,50,54,64}, Z={211,254,290,303,311}

Change the data into below table according to above rules:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **(0)** | **(i)** | **(ii)** | **(iii)** | **(iv)** | **(v)** | **(vi)** |
| **Ref** | **Age** | **Sex** | **Monthly Income** | **Marital**  **Status** | **Service**  **Plan** | **Extra**  **Usage** |
| 1 | 54 c | FEMALE | 3000 x | YES | 100 A | 0 X |
| 2 | 59 c | FEMALE | 4000 x | NO | 600 B | 54 Y |
| 3 | 38 b | MALE | 7800 x | NO | 200 A | 31 Y |
| 4 | 18 a | FEMALE | 8500 y | NO | 600 B | 311 Z |
| 5 | 27 b | MALE | 14000 y | YES | 100 A | 211 Z |
| 6 | 29 b | FEMALE | 31000 z | YES | 1600 C | 25 X |
| 7 | 17 a | MALE | 7500 x | NO | 600 B | 254 Z |
| 8 | 22 a | FEMALE | 7900 y | NO | 200 A | 31 Y |
| 9 | 34 b | MALE | 24700 z | NO | 100 A | 7 X |
| 10 | 46 c | FEMALE | 31110 z | YES | 600 B | 0 X |
| 11 | 39 c | FEMALE | 21000 y | YES | 800 C | 64 Y |
| 12 | 35 b | FEMALE | 30000 z | NO | 1600 C | 0 X |
| 13 | 39 c | MALE | 40500 z | YES | 1600 C | 50 Y |
| 14 | 18 a | MALE | 7800 x | NO | 1000 C | 290 Z |
| 15 | 22 a | MALE | 18000 y | YES | 400 B | 303 Z |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Locus | | | | | |
| **Ref** | **Age** | **Sex** | **Monthly Income** | **Marital**  **Status** | **Service**  **Plan** | **Extra**  **Usage** |
| **1**(1) | 54 c | FEMALE | 3000 x | YES | 100 A | 0 X |
| **2**(8) | 22 a | FEMALE | 7900 y | NO | 200 A | 31 Y |
| **3**(15) | 22 a | MALE | 18000 y | YES | 400 B | 303 Z |

initial cluster centers：

cluster distances of first step:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **(0)** | **(i)** | **(ii)** | **(iii)** | **(iv)** | **(v)** | **(vi)** | **distance** | | |
| **Ref** | **Age** | **Sex** | **Monthly Income** | **Marital**  **Status** | **Service**  **Plan** | **Extra**  **Usage** | **1** | **2** | **3** |
| 1 | 54 c | FEMALE | 3000 x | YES | 100 A | 0 X | **0** | 4 | 5 |
| 2 | 59 c | FEMALE | 4000 x | NO | 600 B | 54 Y | **3** | 3 | 5 |
| 3 | 38 b | MALE | 7800 x | NO | 200 A | 31 Y | 4 | **3** | 5 |
| 4 | 18 a | FEMALE | 8500 y | NO | 600 B | 311 Z | 5 | **2** | 2 |
| 5 | 27 b | MALE | 14000 y | YES | 100 A | 211 Z | 4 | 4 | **2** |
| 6 | 29 b | FEMALE | 31000 z | YES | 1600 C | 25 X | **3** | 5 | 5 |
| 7 | 17 a | MALE | 7500 x | NO | 600 B | 254 Z | 5 | 4 | **2** |
| 8 | 22 a | FEMALE | 7900 y | NO | 200 A | 31 Y | 4 | **0** | 4 |
| 9 | 34 b | MALE | 24700 z | NO | 100 A | 7 X | **4** | 4 | 5 |
| 10 | 46 c | FEMALE | 31110 z | YES | 600 B | 0 X | **2** | 5 | 4 |
| 11 | 39 c | FEMALE | 21000 y | YES | 800 C | 64 Y | **3** | 3 | 4 |
| 12 | 35 b | FEMALE | 30000 z | NO | 1600 C | 0 X | **4** | 4 | 6 |
| 13 | 39 c | MALE | 40500 z | YES | 1600 C | 50 Y | **4** | 5 | 4 |
| 14 | 18 a | MALE | 7800 x | NO | 1000 C | 290 Z | 5 | 4 | **3** |
| 15 | 22 a | MALE | 18000 y | YES | 400 B | 303 Z | 5 | 4 | **0** |

Update cluster centers:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Locus | | | | | |
| **Ref** | **Age** | **Sex** | **Monthly Income** | **Marital**  **Status** | **Service**  **Plan** | **Extra**  **Usage** |
| **1**(1) | 54 c | FEMALE | **z** | YES | **C** | 0 X |
| **2**(8) | 22 a | FEMALE | 7900 y | NO | 200 A | 31 Y |
| **3**(15) | 22 a | MALE | 18000 y | YES | 400 B | 303 Z |

b)

the clustering result of k-means by Python with package sklearn (**initial centers are 1, 8 and 15**):

samples of 1, 2, 10, 11 are cluster 1

samples of 3, 4, 6, 8, 9, 12 are cluster 2

samples of 5, 7, 13, 14, 15 are cluster 3

the program:

*# -\*- encoding:utf-8 -\*-  
  
# Name: SUN RUI ID:18083229g*import pandas as pd  
import numpy as np  
from sklearn import preprocessing  
from sklearn.cluster import KMeans  
  
  
def load\_data(file\_path):  
 df\_data = pd.read\_csv(file\_path)  
 df\_data.columns = ["Age", "Sex", "MonthlyIncome", "MaritalStatus", "ServicePlan", "ExtraUsage"]  
 scaled\_data = preprocessing.scale(df\_data)  
 return scaled\_data  
  
def do\_k\_means(data):  
 centers = np.vstack((data[0], data[7], data[14]))  
 k\_means = KMeans(n\_clusters=3, init=centers, n\_init=1, max\_iter=1000)  
 return k\_means.fit\_predict(data), k\_means.fit(data).cluster\_centers\_  
  
def print\_clusters(cluster\_index):  
 cluster\_dict = {0:[],1:[],2:[]}  
 for i in range(len(cluster\_index)):  
 cluster\_dict[cluster\_index[i]].append(i+1)  
 print("cluster result: {}".format(cluster\_dict))  
  
if \_\_name\_\_ == '\_\_main\_\_':  
 df\_scaled\_data = load\_data("data.csv")  
 cluster\_index, cluster\_center = do\_k\_means(df\_scaled\_data)  
 print("cluster index: {}".format(cluster\_index))  
 print\_clusters(cluster\_index)  
 print("cluster centers:\n{}".format(cluster\_center))

c）to compare the performances between k-means and k-modes, I plan to use scatter picture to show the differences. Because of the data has six attributes, I have to reduce dimensions from 6 to 2 or from 6 to 3 by PCA, so the results are a little different from part b). Let us see the results of k-means firstly:

The cluster result:

samples of 6, 10, 11, 12, 13 are cluster 1

samples of 1, 2, 3, 8, 9 are cluster 2

samples of 4, 5, 7, 14, 15 are cluster 3

 