Design a distance function to assess the similarity of customers of a supermarket; each customer in a supermarket is characterized by the following attributes:

a) SSN

b) Items\_Bought (The set of items the bought last month)

c) Amount\_spend (Average amount spent per purchase in dollars and cents; it has a mean of

50.00, a standard deviation of 40, the minimum is 0.05 and the maximum is 600)

d) Age (is an ordinal attribute taking 5 values: child, young, medium, old, very\_old)

Assume that Items\_Bought and Amount\_Spend are of major importance and Age is of a minor importance when assessing the similarity of the customers.

MANY POSSIBLE ANSWERS

One possible answer: Ignore SSN as it is not important.

Find distance between items bought using Jaccard: 1 – Jaccard Index(u.Items\_Bought,v.Items\_Bought)

Normalize Amount\_spend using Z-score and find distance by L-1 norm

Assign 5 values to Age using a function : child = 0, young = 1, medium = 2, old = 3, very old = 4

Find distance by taking L-1 norm and dividing by range i.e. 4

Assign weights 0.4 to Items\_Bought, 0.4 to Amount\_spend and 0.2 to age.

Compute the distance of the following 2 customers:

c1 = (111111111, {A, B, C, D}, 40.00, ‘old’) and

c2 = (222222222, {D, E, F}, 100.00, ‘young’)

with your distance function:

d (c1, c2) = 0.4 \* (1 – 1/6) + 0.4 \*|(40-50)/40 – (100-50)/40| + 0.2 \*|3-1|/4

= 0.33 + 0.6 + 0.1

= 1.03