Zachary Rutherford
HW 06 Writeup

Part A:

a. Which two attributes are most strongly positively cross-correlated with each other? (1)
The two attributes with the most positive cross-correlation are Vegges and YogChs with a coefficient of 0.76

b. Which two attributes are most strongly negatively cross-correlated with each other? (1)
The two attributes with the most negative cross-correlation are Vegges and Soda with a coefficient of -0.83

c. What is the cross-correlation coefficient of Chips with cereal? (1)
0.19

d. Which attribute is fish most strongly cross-correlated with? (1)
Chips at 0.28

e. Which attribute is Veggies most strongly cross-correlated with? (1)
YogChs at 0.76

f. According to this data, do people usually buy milk and cereal together? (1)
The data shows a coefficient of 0.01, which means people do not buy them together

g. Which two attributes are not strongly cross-correlated with anything? (1)
Salt and Fruit are the two attributes that are not strongly cross-correlated with anything

h. If you were to delete several attributes, which attributes would you believe were irrelevant? (1)
(The maximum of the absolute value of the CC is <= 0.1.)
based on the maximum given Eggs, Fruit, Beans, and Salt may be irrelevant

i. What other attribute is "child-baby" products most positively associated with? (1)
Milk

j. If buying fish is strongly cross-correlated with another item, and buying that item is strongly highly cross-correlated with a third item, is buying fish strongly cross-correlated with the third item?Explain your answer... (1)
Yes, with an asterisk. If A strongly cross-correlated B and B is strongly cross-correlated with C, then A and B will be strongly cross-correlated, depending on what the threshold is. cross-correlation is linear, so when A increases, B increases with it. If A is strongly cross-correlated with B, then their coefficient is close to 1, and if B is strongly cross-correlated with C, it is also close to 1. Therefore, the coeffecient between A and C will similarly be close, but they could be farther in the worst-case, and outside what one might consider to be "strongly" cross-correlated. Like maybe A and B are 0.95, same with B and C, but in the worst case A and C could be 0.9, which may or may not be within the category for strongly cross-correlated.

Part B:

(The full csv file was taking too long, I changed it to the first 200 values so I could get the write-up done in time, the dendrogram looks similar, so I think the final results will be similar as well)

7. Implement agglomerative clustering by hand. Do not use a package.

d. Report the size of the last 20 smallest clusters merged. (1)

From left to right by merge order

[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 42, 51, 61]

e. Based on the previous answer, how many clusters do you think are in your data? (1)

I believe there are about 4 different clusters in my data. There is a steep increase in the size of the clusters from 1 to 42.

8. Report the size of each cluster, from lowest to highest. (1)

The clusters seem to be 42, 51, 61, and our missing member being 46

9. Report the average prototype of each of the clusters. (1)

In order:

{' Milk': 4.595238095238095, 'ChildBby': 3.0714285714285716, 'Vegges': 6.523809523809524, 'Cereal': 7.238095238095238, ' Bread': 6.0, ' Rice': 6.166666666666667, ' Meat': 6.428571428571429, ' Eggs': 5.214285714285714, 'YogChs': 5.285714285714286, ' Chips': 8.238095238095237, ' Soda': 2.0, ' Fruit': 5.333333333333333, ' Corn': 3.9047619047619047, ' Fish': 5.333333333333333, ' Sauce': 2.5714285714285716, ' Beans': 5.190476190476191, 'Tortya': 7.023809523809524, ' Salt': 4.5476190476190474, 'Scented': 5.095238095238095, ' Salza': 2.0}

{' Milk': 9.470588235294118, 'ChildBby': 7.784313725490196, 'Vegges': 7.764705882352941, 'Cereal': 8.117647058823529, ' Bread': 8.098039215686274, ' Rice': 6.352941176470588, ' Meat': 7.862745098039215, ' Eggs': 5.392156862745098, 'YogChs': 4.647058823529412, ' Chips': 1.8627450980392157, ' Soda': 1.9019607843137254, ' Fruit': 5.313725490196078, ' Corn': 4.254901960784314, ' Fish': 2.411764705882353, ' Sauce': 2.0392156862745097, ' Beans': 5.196078431372549, 'Tortya': 1.2549019607843137, ' Salt': 4.568627450980392, 'Scented': 5.313725490196078, ' Salza': 2.0784313725490198}

{' Milk': 2.0, 'ChildBby': 1.9344262295081966, 'Vegges': 1.6229508196721312, 'Cereal': 8.163934426229508, ' Bread': 2.3934426229508197, ' Rice': 6.065573770491803, ' Meat': 8.147540983606557, ' Eggs': 4.80327868852459, 'YogChs': 0.9836065573770492, ' Chips': 8.754098360655737, ' Soda': 8.065573770491802, ' Fruit': 5.377049180327869, ' Corn': 6.540983606557377, ' Fish': 2.5245901639344264, ' Sauce': 5.639344262295082, ' Beans': 5.147540983606557, 'Tortya': 7.770491803278689, ' Salt': 4.524590163934426, 'Scented': 5.065573770491803, ' Salza': 5.721311475409836}

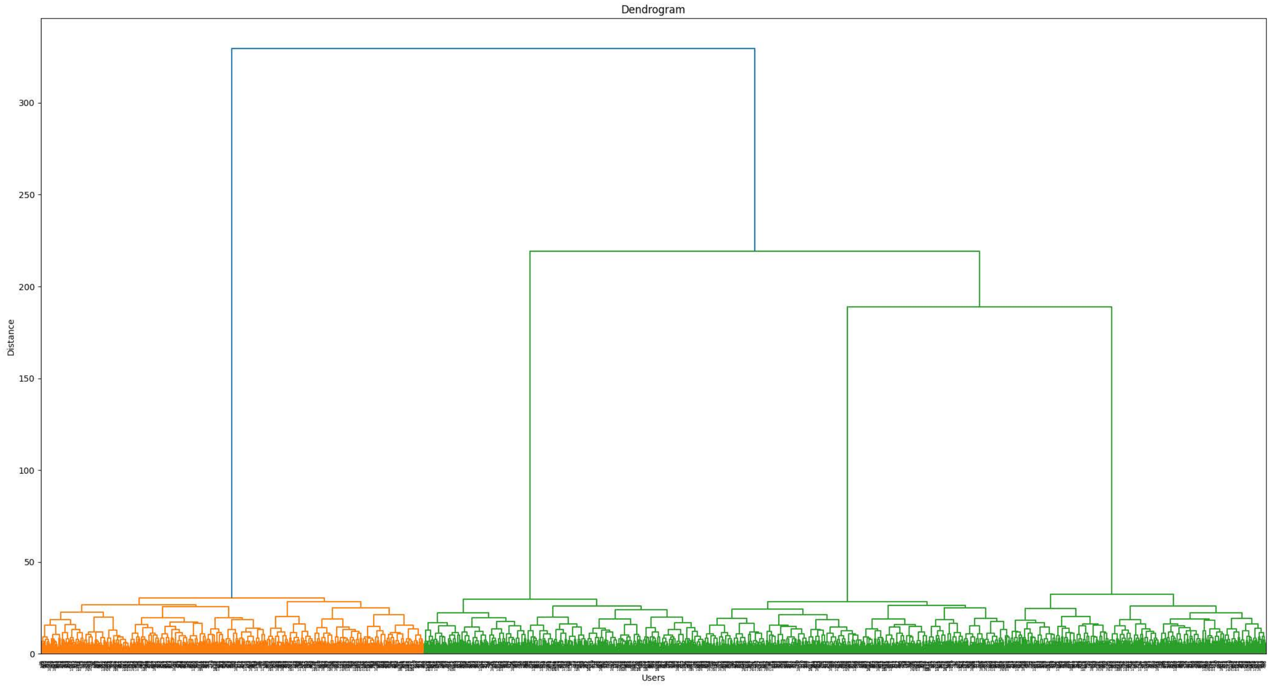
10. What typifies each of the clusters? What typical names should we give each of these prototypes? Is there a gluten-free group? Is there a family group? Is there a group of party animals? Are there vegans? Are there healthy eaters? What typifies each group you found? (1)

There seems to be 1 group buying lot's of milk and child products. They also buy cereal, eggs, meat. I would say that this group is for families do to the high number of these items bought. Next is healthy eaters, which buy manmy items in a wide variety, but especially veggies. Then there are unhealthy eaters, which buy chips, lot's of meat, soda, etc. Those might just be college students though.

Part C:

This assignment was tough. I got the cross-correlation stuff pretty easily, but when I got to the actual agglomeration, it stumped me for a while. I had to come at this from some strange angles as I'm not really used to working in the nth dimension, I'm not a huge math guy. It took a long time just to make the distance matrix, but once you have that, it becomes much easier. The only really terrible problem I had was the computational inefficiency of the algorithm. When I was researching I did see references that this is not great for large datasets, and boy are they not kidding. Like suggested, I developed most of the code in a test suite, and that worked well, with the whole process only taking a few minutes. However, when it was time to do the write-up and I set the program to compute the entire csv file, it took forever. I literally wen out to eat, finished other homework, and came back to it not even 20% done. It was a bit ridiculous. I think if I could redo this project, I would do it in java or c# so that I could parallelize the creation of the matrix, which would make things much better. Python is great, but it's just not up to the task of dealing with computationally expensive algorithms.

I included the dendrogram created from the full csv file at the end of this page.


Dendrogram
Distance
Users