HW_06 Agglomerative Clustering

1. You will need to remove one of the attributes in the CSV file.
Which one should you *always be certain to* remove?

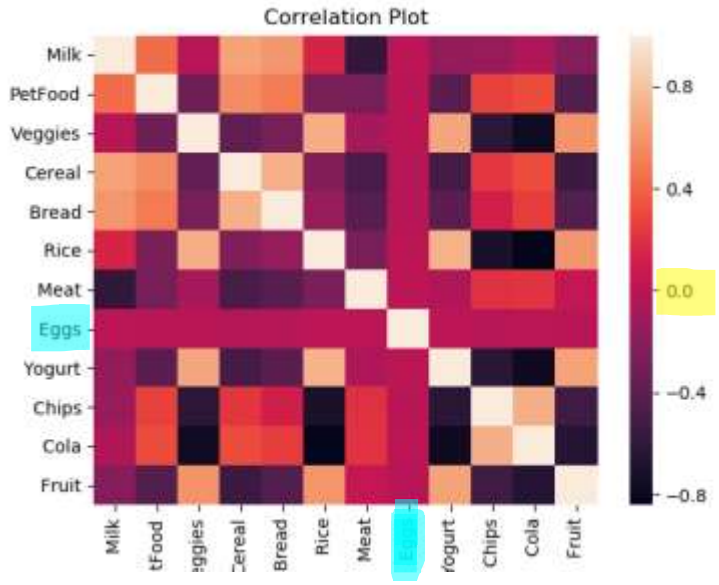
Clustering is the process to from group of datapoints with similar properties based on the attributes.
Thus, we do need the **unique identification** information for the clustering method.
Thus, one should always be certain to remove the **unique ID** from the data.

2. Remark on the cross-correlation coefficients of the attributes. What information do they reveal?

The correlation matrix is as follows:

	Milk	PetFood	Veggies	Cereal	Bread	Rice	Meat	Eggs	Yogurt	Chips	Cola	Fruit
Milk	1	0.43443	-0.00586	0.656148	0.608239	0.137587	-0.59153	0.012702	-0.15692	-0.13449	-0.03917	-0.21867
PetFood	0.43443	1	-0.30492	0.561622	0.495695	-0.26193	-0.28307	0.003835	-0.40844	0.280244	0.314515	-0.43107
Veggies	-0.00586	-0.30492	1	-0.36772	-0.28507	0.705635	-0.07942	0.007277	0.670604	-0.62211	-0.75449	0.594803
Cereal	0.656148	0.561622	-0.36772	1	0.720604	-0.23437	-0.48238	-0.02048	-0.50561	0.229975	0.314731	-0.5458
Bread	0.608239	0.495695	-0.28507	0.720604	1	-0.14409	-0.41595	-0.01208	-0.39824	0.114293	0.252806	-0.4334
Rice	0.137587	-0.26193	0.705635	-0.23437	-0.14409	1	-0.27186	0.018745	0.730972	-0.70969	-0.84061	0.611131
Meat	-0.59153	-0.28307	-0.07942	-0.48238	-0.41595	-0.27186	1	0.013568	-0.0378	0.202524	0.21182	0.050255
Eggs	0.012702	0.003835	0.007277	-0.02048	-0.01208	0.018745	0.013568	1	0.015229	-0.00116	0.016095	-0.00135
Yogurt	-0.15692	-0.40844	0.670604	-0.50561	-0.39824	0.730972	-0.0378	0.015229	1	-0.63473	-0.76533	0.655999
Chips	-0.13449	0.280244	-0.62211	0.229975	0.114293	-0.70969	0.202524	-0.00116	-0.63473	1	0.712859	-0.52774
Cola	-0.03917	0.314515	-0.75449	0.314731	0.252806	-0.84061	0.21182	0.016095	-0.76533	0.712859	1	-0.66068
Fruit	-0.21867	-0.43107	0.594803	-0.5458	-0.4334	0.611131	0.050255	-0.00135	0.655999	-0.52774	-0.66068	1

The correlation plot is as follows:



Any attribute is highly correlated with itself. Thus, in the plot we see diagonal white blocks representing the correlation=1 with itself which can also be seen in the correlation matrix. (logically correlation of a variable with itself has no meaning)
We also see that most variables have positive or negative correlation with at least one of the other attributes except for the attribute Eggs.
For Eggs we see a pink block (which stands for correlation = 0) throughout, with all other variables. If we the correlation matrix for the actual values, we see that the correlation of Eggs with any other variable is negligible.

3. You can keep all the other attributes or remove a few of them.
Which attribute(s), if any, did you remove?

In clustering, we group together the variables with similar properties.
For this, the attributes need to be related. From the explanation above, we see that the attribute Eggs is uncorrelated with any other attribute in the data. Thus, Eggs would not be a part of any of the cluster. Therefore, it is okay to remove the Attribute Eggs from the data for further Cluster Analysis.

4. At each stage of clustering, you record the size of the smaller cluster being merged in.
For the last ten merges, what was the size of smaller cluster that was merged in?
What does this indicate about the true number of clusters?

[1,1,1,1,1,1,1,1,125,62]
We see a jump at the 9^{th} and the 10^{th} merge. Thus, we see the two clusters (125,62)here clearly.
The third cluster is the remaining elements viz 145 datapoints.
Thus, we see the three clusters being formed from this list of sizes.

5. Look at the average amount of milk, etc… purchased by the third cluster of shoppers.
What typifies the third cluster? What nick-name should we give these customers? (be polite)

The ID's of the datapoints that belong to each of the three cluster are as follows:

1st cluster : Size =62

[289, 185, 169, 313, 186, 33, 101, 240, 269, 209, 319, 5, 262, 200, 268, 239, 86, 217, 284, 272, 134, 40, 251, 317, 136, 276, 197, 207, 12, 306, 220, 39, 95, 132, 36, 85, 327, 112, 232, 290, 318, 88, 287, 107, 53, 294, 10, 35, 283, 211, 135, 231, 139, 115, 128, 213, 61, 258, 75, 219, 51, 192]

2nd cluster: Size =125

[145, 203, 173, 65, 212, 109, 17, 297, 50, 233, 157, 273, 116, 215, 15, 195, 56, 104, 167, 196, 226, 81, 312, 100, 308, 52, 234, 49, 78, 286, 118, 153, 301, 229, 334, 63, 182, 54, 208, 42, 204, 179, 11, 223, 256, 205, 243, 316, 131, 333, 166, 238, 248, 127, 282, 119, 27, 23, 147, 264, 314, 193, 320, 250, 55, 154, 93, 277, 66, 37, 108, 225, 44, 74, 206, 198, 336, 41, 161, 67, 288, 16, 174, 148, 58, 241, 90, 309, 43, 164, 263, 155, 303, 30, 126, 133, 326, 247, 310, 92, 25, 71, 292, 170, 246, 190, 253, 106, 21, 87, 259, 3, 249, 315, 79, 142, 228, 124, 19, 184, 99, 72, 285, 80, 291]

3rd cluster: Size =145

[7, 2, 323, 151, 271, 143, 216, 254, 275, 102, 257, 210, 168, 202, 201, 214, 1, 4, 64, 278, 38, 187, 295, 224, 265, 20, 141, 125, 252, 73, 24, 158, 96, 9, 13, 110, 175, 113, 331, 218, 227, 121, 150, 45, 83, 29, 178, 307, 324, 89, 281, 293, 156, 160, 188, 94, 130, 242, 47, 122, 172, 322, 163, 270, 18, 77, 111, 144, 69, 222, 28, 266, 302, 162, 31, 321, 280, 149, 103, 62, 244, 76, 46, 230, 57, 120, 304, 98, 26, 329, 299, 311, 191, 84, 180, 91, 300, 332, 183, 328, 199, 138, 236, 140, 152, 34, 165, 325, 22, 8, 330, 260, 261, 279, 137, 105, 237, 114, 305, 117, 60, 70, 296, 335, 337, 97, 176, 82, 255, 14, 274, 221, 171, 181, 267, 235, 298, 32, 59, 68, 177, 146, 48, 6, 123, 245, 129, 189, 159, 194]

The centers (12) of each of the three clusters are as follows:

Cluster 1: [7.64, 4.78, 7.693333333333333, 8.026666667, 6.5266666, 8.28, 3.92666667, 5.046666666666667, 5.326666666666667, 2.9866667, 2.83333333335, 5.03333333]
Name: Family

Cluster 2: [2.161290322580645, 1.8548387096774193, 9.129032258064516, 0.8548387096774194, 1.4516129032258065, 8.903225806451612, 7.741935483870968, 5.080645161290323, 8.17741935483871, 2.435483870967742, 1.1451612903225807, 8.064516129032258]
Name: Party People

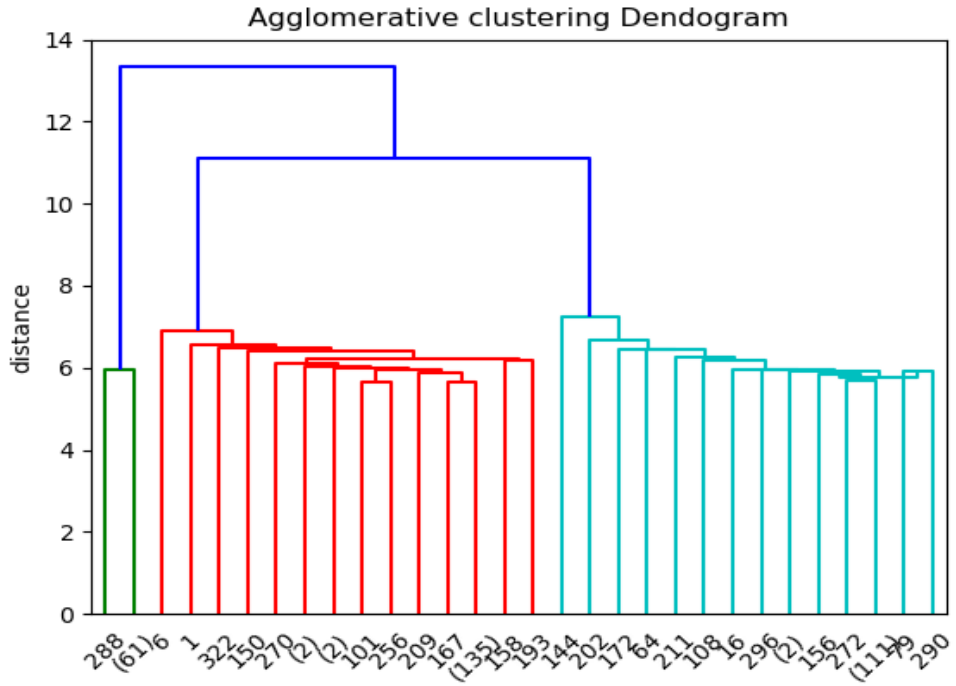
Cluster 3: [2.161290322580645, 1.8548387096774193, 9.129032258064516, 0.8548387096774194, 1.4516129032258065, 8.903225806451612, 7.741935483870968, 5.080645161290323, 8.17741935483871, 2.435483870967742, 1.1451612903225807, 8.064516129032258]
Name: Diet Cautious.

From the centers for the third cluster, we can see that the people in this cluster and very careful about their food buying habits.
We see a high average for, Veggie, Rice, Meat, Yogurt, Fruits.
Low average for, Milk, Pet food, Cereal, Bread, Chips and cola.

6. If we switched from a "central link" to a "single link" merge step, what else would you need to add to the algorithm when computing the distance between two clusters?

When computing the distance between two cluster with 'single link',
We calculate the distance between attributes of an object and all the attributes of the other object and find the minimum distance of all the attribute distances for these two objects. We do this for all the objects to find the distance metric (minimum) for each object.

7. Generate a dendrogram of the clusters as they are being merged.
Show the code that demonstrates your understanding of this. This is easy in R.
You can use a package in R, Python, Matlab or Java for this, but you cannot use a web resource.
You do not need to use the same language for everything in the entire homework.



Discussed the Homework with Utsav Patel.