Rohit Ravishankar
rr9105@rit.edu

Home Work 6

CSCI - 720 Big Data Analytics

Collaborators: None

1.

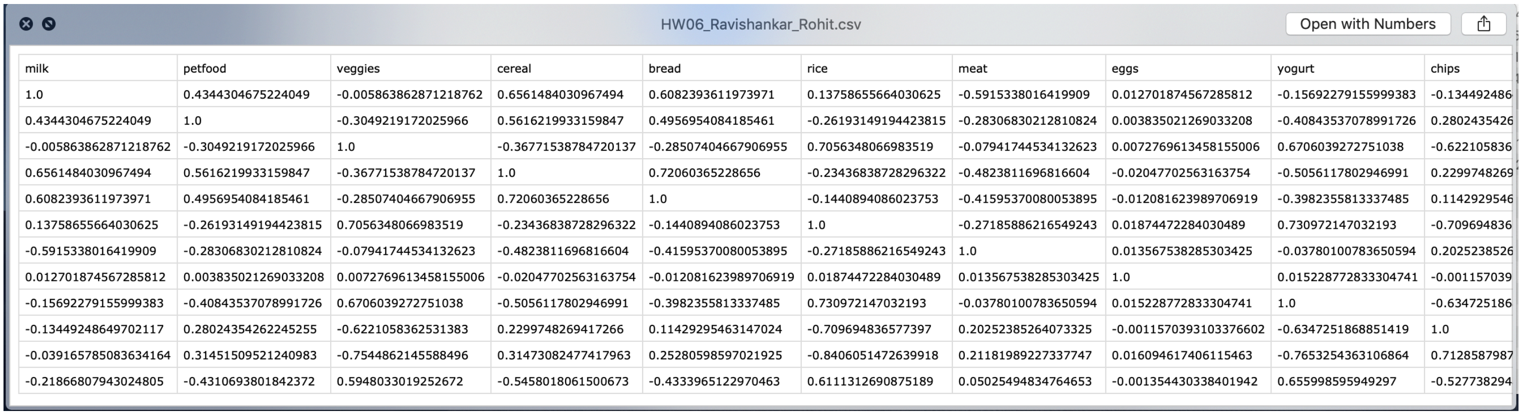
You will need to remove one of the attributes in the CSV file. Which one should you *always be certain* to remove?

Ans. The attribute we must remove in the CSV file is *'ID'*. This is because the *'ID'* field has no bearing on the correlation matrix or generating the dendrogram or the agglomerative clustering.

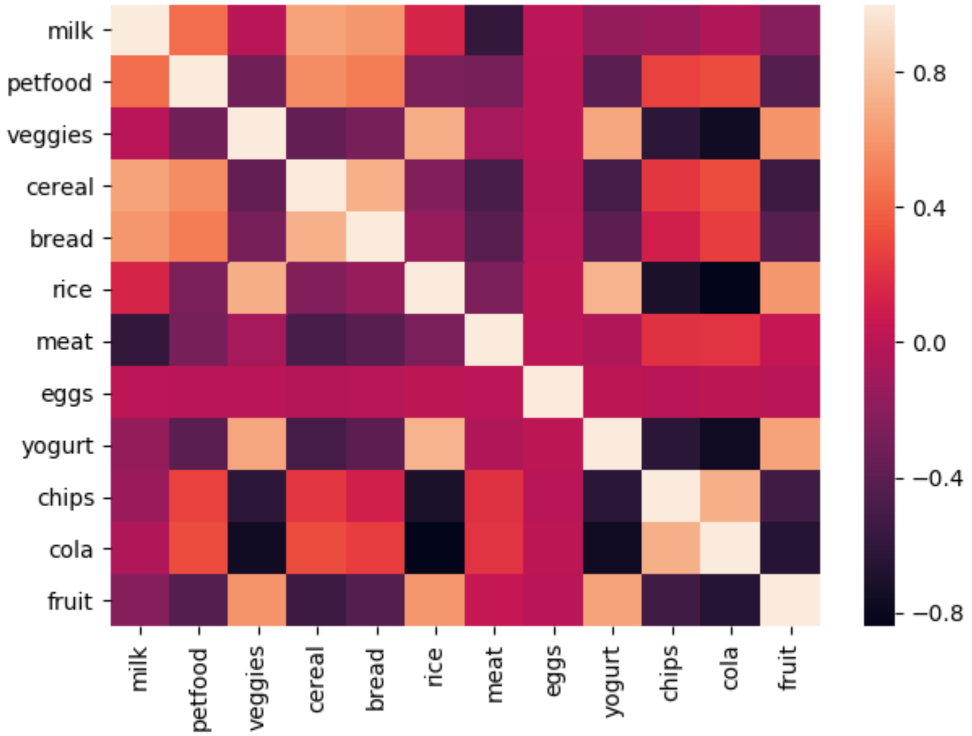
2.

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

Ans.



milk	petfood	veggies	cereal	bread	rice	meat	eggs	yogurt	chips
1.0	0.4344304675224049	-0.005863862871218762	0.6561484030967494	0.6082393611973971	0.13758655664030625	-0.5915338016419909	0.012701874567285812	-0.15692279155999383	-0.134492486
0.4344304675224049	1.0	-0.3049219172025966	0.5616219933159847	0.4956954084185461	-0.26193149194423815	-0.28306830212810824	0.003835021269033208	-0.40843537078991726	0.2802435426
-0.005863862871218762	-0.3049219172025966	1.0	-0.36771538784720137	-0.28507404667906955	0.7056348066983519	-0.07941744534132623	0.0072769613458155006	0.6706039272751038	-0.622105836
0.6561484030967494	0.5616219933159847	-0.36771538784720137	1.0	0.72060365228656	-0.23436838728296322	-0.4823811696816604	-0.02047702563163754	-0.5056117802946991	0.2299748269
0.6082393611973971	0.4956954084185461	-0.28507404667906955	0.72060365228656	1.0	-0.1440894086023753	-0.41595370080053895	-0.012081623989706919	-0.3982355813337485	0.1142929546
0.13758655664030625	-0.26193149194423815	0.7056348066983519	-0.23436838728296322	-0.1440894086023753	1.0	-0.27185886216549243	0.01874472284030489	0.730972147032193	-0.709694836
-0.5915338016419909	-0.28306830212810824	-0.07941744534132623	-0.4823811696816604	-0.41595370080053895	-0.27185886216549243	1.0	0.013567538285303425	-0.03780100783650594	0.2025238526
0.012701874567285812	0.003835021269033208	0.0072769613458155006	-0.02047702563163754	-0.012081623989706919	0.01874472284030489	0.013567538285303425	1.0	0.015228772833304741	-0.001157039
-0.15692279155999383	-0.40843537078991726	0.6706039272751038	-0.5056117802946991	-0.3982355813337485	0.730972147032193	-0.03780100783650594	0.015228772833304741	1.0	-0.634725186
-0.13449248649702117	0.28024354262245255	-0.6221058362531383	0.2299748269417266	0.11429295463147024	-0.709694836577397	0.20252385264073325	-0.0011570393103376602	-0.6347251868851419	1.0
-0.039165785083634164	0.31451509521240983	-0.7544862145588496	0.31473082477417963	0.25280598597021925	-0.8406051472639918	0.21181989227337747	0.016094617406115463	-0.7653254363106864	0.7128587987
-0.21866807943024805	-0.4310693801842372	0.5948033019252672	-0.5458018061500673	-0.4333965122970463	0.6111312690875189	0.05025494834764653	-0.001354430338401942	0.655998595949297	-0.527738294



Cross-correlation coefficients reveal information about how much is one attribute dependent on another attribute. The values for correlation coefficient lie between [-1, +1]

- Positive correlation - If one attribute increases and the other attribute increases correspondingly at a similar rate. The correlation coefficient would be positive.
- Negative correlation - If one attribute increases and the other attribute decreases correspondingly at a similar rate. The correlation coefficient would be negative

By plotting the above heat map, we can visualize all the attributes which are strongly correlated and all the attributes which are not correlated. For example,

- Cereal & bread and rice & veggies can be said to have a **strong positive** correlation
- Cola & fruit and veggies & cola can be said to have a **strong negative** correlation
- Milk & petfood and fruit & veggies can be said to have a **weak positive** correlation
- Petfood & fruit and bread & fruit can be said to have a **weak negative** correlation

An interesting point to note is that eggs have a very small correlation with all the other attributes and is very close being uncorrelated.

3.

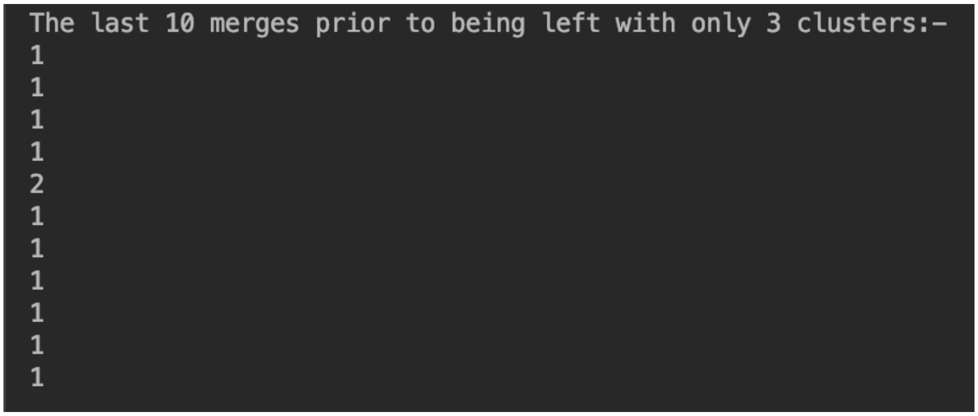
You can keep all the other attributes, or remove a few of them. What attribute(s), if any, did you remove?

Ans. I did remove the attribute, *'eggs'* apart from *'ID'*, because it had a correlation of the order of 10^{-2} which is very small in comparison to the cross correlation coefficients of the other products.

4.

At each stage of clustering, you record the size of the smaller cluster being merged. 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?

Ans. The picture below depicts the smaller cluster value being merged in for the last 10 merges. This indicates that there are small cluster which are distant and haven't been merged so far. They are being merged together to give 3 bigger clusters in the end.



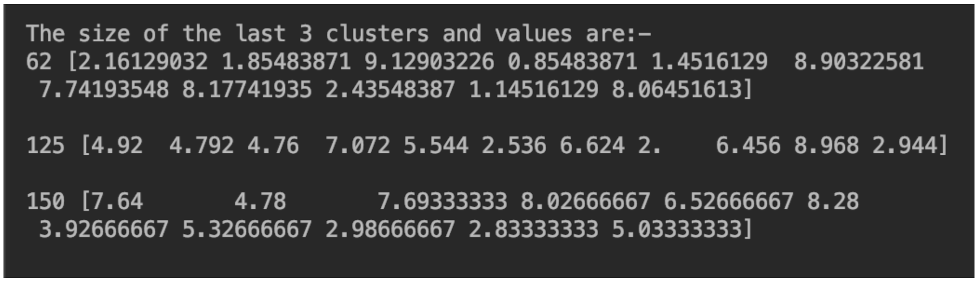
```
The last 10 merges prior to being left with only 3 clusters:-
1
1
1
1
2
1
1
1
1
1
1
```

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)

Ans.

[Milk, PetFood, Veggies, Cereal, Bread, Rice, Meat, Eggs,Yogurt, Chips, Cola, Fruit] is the order of the list of the values in the lists below.



```
The size of the last 3 clusters and values are:-
62 [2.16129032 1.85483871 9.12903226 0.85483871 1.4516129  8.90322581
 7.74193548 8.17741935 2.43548387 1.14516129 8.06451613]

125 [4.92  4.792 4.76  7.072 5.544 2.536 6.624 2.    6.456 8.968 2.944]

150 [7.64       4.78       7.69333333 8.02666667 6.52666667 8.28
 3.92666667 5.32666667 2.98666667 2.83333333 5.03333333]
```

The third cluster can be thought to be a “health conscious” set of people. If you observe the third cluster of size **62,** the quantity of veggies, rice, eggs, yogurt and fruit consumption is high and the consumption of processed food such as cereal, bread, cola and chips is low which would imply that they are possibly living healthy lifestyles. We also see that they aren’t consuming high amounts milk as milk is consumed with cereals and since the consumption of cereal is low, so is the consumption of milk.

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?

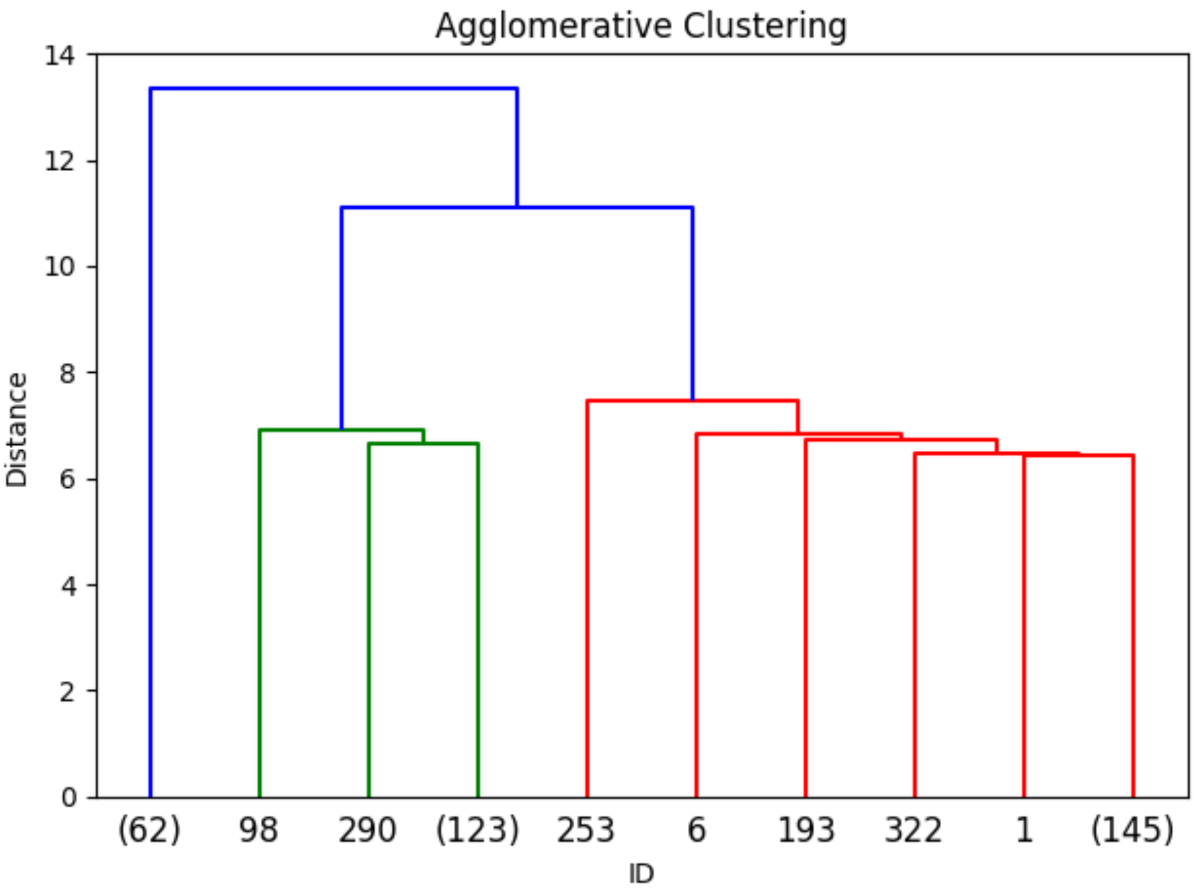
Ans. If you switched from a “central link” to a “single link” merge step, at each step we would combine two clusters that contain the closest pair of elements rather than the the distance between the centers of the clusters.

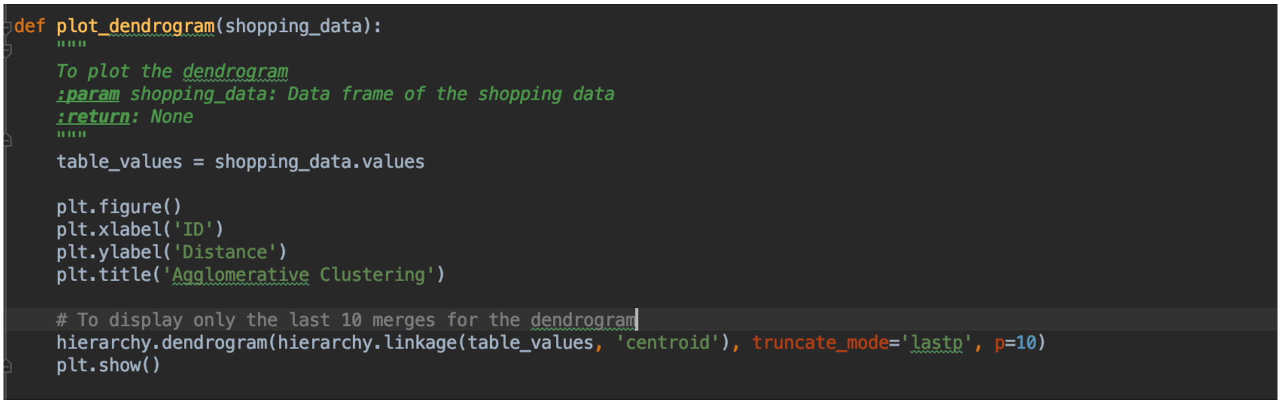
7.

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

Ans.

The dendrogram depicts the last **10** merges





```
def plot_dendrogram(shopping_data):
    """
    To plot the dendrogram
    :param shopping_data: Data frame of the shopping data
    :return: None
    """
    table_values = shopping_data.values

    plt.figure()
    plt.xlabel('ID')
    plt.ylabel('Distance')
    plt.title('Agglomerative Clustering')

    # To display only the last 10 merges for the dendrogram
    hierarchy.dendrogram(hierarchy.linkage(table_values, 'centroid'), truncate_mode='lastp', p=10)
    plt.show()
```