CSCI 720 - Big Data Analytics

Nihal Surendra Parchand
Rohit Kunjilikattil

np9603@rit.edu
rk4447@rit.edu

Homework 8 PCA and Agglomeration

Part 1: Using Cross-Correlation for Feature Rejection and Selection:

1. Compute the cross-correlation coefficient of all attributes. Use a package to do this. Your matrix should be n by n where n is the number of attributes.
All values computed should be in the range [-1, 1].

	Beans	Bread	Cerel	ChdBby	Chips	Corn	Eggs	Fish	Fruit	Meat	Milk	Pepper	Rice	Salza	Sauce	Soda	Tomato	Tortya	Vegges	YogChs
Beans	1.00	-0.12	-0.18	0.03	-0.06	0.43	-0.51	-0.37	-0.01	-0.63	-0.08	0.28	0.41	0.52	0.52	-0.40	0.22	0.42	0.26	0.56
Bread	-0.12	1.00	0.44	0.01	0.12	-0.13	0.00	-0.07	0.04	0.01	0.44	0.20	-0.08	0.02	-0.08	-0.06	-0.27	-0.21	-0.03	-0.07
Cerel	-0.18	0.44	1.00	0.01	0.34	-0.31	-0.01	-0.04	-0.01	-0.02	0.29	0.01	-0.36	0.08	-0.04	0.27	-0.47	-0.35	-0.40	-0.41
ChdBby	0.03	0.01	0.01	1.00	-0.04	-0.04	-0.02	0.03	-0.02	-0.04	-0.04	-0.05	0.02	-0.02	0.04	-0.03	-0.02	-0.05	0.00	-0.03
Chips	-0.06	0.12	0.34	-0.04	1.00	0.28	0.19	-0.31	-0.01	0.30	0.05	0.40	-0.51	0.49	-0.46	0.54	0.21	0.37	-0.54	-0.29
Corn	0.43	-0.13	-0.31	-0.04	0.28	1.00	0.11	-0.25	0.02	0.02	0.23	0.72	0.35	0.63	-0.29	-0.32	0.69	0.83	0.32	0.53
Eggs	-0.51	0.00	-0.01	-0.02	0.19	0.11	1.00	0.45	-0.04	0.54	0.24	0.11	-0.04	-0.04	-0.69	0.07	0.25	0.13	-0.03	-0.27
Fish	-0.37	-0.07	-0.04	0.03	-0.31	-0.25	0.45	1.00	-0.04	0.08	0.16	-0.34	0.22	-0.27	-0.14	-0.20	-0.10	-0.29	0.13	-0.35
Fruit	-0.01	0.04	-0.01	-0.02	-0.01	0.02	-0.04	-0.04	1.00	-0.00	0.00	0.00	-0.02	-0.01	0.01	-0.00	-0.04	-0.03	0.00	0.02
Meat	-0.63	0.01	-0.02	-0.04	0.30	0.02	0.54	0.08	-0.00	1.00	0.09	0.10	-0.36	-0.22	-0.73	0.39	0.20	0.06	-0.21	-0.27
Milk	-0.08	0.44	0.29	-0.04	0.05	0.23	0.24	0.16	0.00	0.09	1.00	0.49	0.25	0.22	-0.33	-0.42	0.01	0.11	0.30	0.14
Pepper	0.28	0.20	0.01	-0.05	0.40	0.72	0.11	-0.34	0.00	0.10	0.49	1.00	0.19	0.63	-0.40	-0.27	0.51	0.68	0.23	0.43
Rice	0.41	-0.08	-0.36	0.02	-0.51	0.35	-0.04	0.22	-0.02	-0.36	0.25	0.19	1.00	0.17	0.21	-0.80	0.26	0.26	0.69	0.56
Salza	0.52	0.02	0.08	-0.02	0.49	0.63	-0.04	-0.27	-0.01	-0.22	0.22	0.63	0.17	1.00	-0.10	-0.17	0.40	0.64	0.02	0.25
Sauce	0.52	-0.08	-0.04	0.04	-0.46	-0.29	-0.69	-0.14	0.01	-0.73	-0.33	-0.40	0.21	-0.10	1.00	-0.21	-0.37	-0.31	0.12	0.20
Soda	-0.40	-0.06	0.27	-0.03	0.54	-0.32	0.07	-0.20	-0.00	0.39	-0.42	-0.27	-0.80	-0.17	-0.21	1.00	-0.17	-0.20	-0.76	-0.59
Tomato	0.22	-0.27	-0.47	-0.02	0.21	0.69	0.25	-0.10	-0.04	0.20	0.01	0.51	0.26	0.40	-0.37	-0.17	1.00	0.74	0.23	0.38
Tortya	0.42	-0.21	-0.35	-0.05	0.37	0.83	0.13	-0.29	-0.03	0.06	0.11	0.68	0.26	0.64	-0.31	-0.20	0.74	1.00	0.22	0.47
Vegges	0.26	-0.03	-0.40	0.00	-0.54	0.32	-0.03	0.13	0.00	-0.21	0.30	0.23	0.69	0.02	0.12	-0.76	0.23	0.22	1.00	0.61
YogChs	0.56	-0.07	-0.41	-0.03	-0.29	0.53	-0.27	-0.35	0.02	-0.27	0.14	0.43	0.56	0.25	0.20	-0.59	0.38	0.47	0.61	1.00

2. Report:

```
The two attributes that are most strongly cross-correlated with each other: Tortya and Corn -> 0.8328401558134411
Which attribute is fish most strongly cross-correlated with?    Eggs and   Fish -> 0.44905640801135965
Which attribute is meat most strongly cross-correlated with?   Sauce and   Meat -> 0.7290675318817506
Which attribute is beans most strongly cross-correlated with?   Beans and   Meat -> 0.632335336015049
The least cross-correlated attribute is   Fruit
The second least cross-correlated attribute is  ChdBby
```

a. Which two attributes are most strongly cross-correlated with each other?
→ The two most cross-correlated are Tortya and corn.

b. Which attribute is fish most strongly cross-correlated with?
→ Fish is most strongly cross-correlated with Eggs.

c. Which attribute is meat most strongly cross-correlated with?
→ Meat is most strongly cross-correlated with Sauce.

d. Which attribute is beans most strongly cross-correlated with?
→ Beans is most strongly cross-correlated with Meat.

e. Which one attribute is least correlated with all other attributes?
→ Fruit

f. Which second attribute is least correlated with all other attributes?
→ ChdBby

g. If you were to delete two attributes, which would you guess were irrelevant?
→ We would guess that Fruit and ChdBby were irrelevant as they are least cross-correlated with all the other attributes.

h. If buying fish is strongly cross-correlated with buying cereal, and buying cereal is strongly cross correlated with buying baby products, is buying fish strongly cross-correlated with buying baby products? Can you explain this?
→ No, because cross correlation is not necessarily transitive.

Part 2: Principal Components Analysis:

c. Sort the eigenvalues from highest absolute value to lowest absolute value. Print all the eigenvalues. How many are large? Report what you observe.

```
------------EIGEN VALUES----------------
[41.94493984 32.7810728  20.34207505 13.3309452   7.3526868   2.72002896
  2.50910589  2.19527927  2.12862555  1.9966341   1.88078113  1.71518714
  1.67828302  1.57841269  1.45884985  1.4179065   1.30796831  1.20484114
  1.1475737   1.08965803]
```

→ Looking at the values above, we can see that there are 5 eigenvalues which are large. Also, we can see that all the eigenvalues are positive.

e. (1) Print out the first five eigenvectors to one significant digit – the eigenvectors associated with the five eigenvalues that have the largest five absolute values. Look at the components of each eigenvector. Tell me about them. What do they tell you about the data? What does this tell you about the attributes? Which attribute is least important? Which attribute is most important? Is there an attribute that is zero in all four eigenvectors? What is the relationship between these attributes and the cross-correlation values?

```
------------EIGEN VECTORS----------------
Eigenvector 1
[ 0.3 -0.1  0.4 -0.   0.2 -0.1 -0.  -0.  -0.1  0.3 -0.2  0.3 -0.1 -0.
 -0.2  0.3  0.6  0.1  0.  -0. ]
```

```
Eigenvector 2
[-0.   0.  -0.  -0.4 -0.2 -0.1  0.1 -0.7  0.4  0.1  0.1  0.1  0.1  0.
  0.2  0.1  0.1 -0.1  0.2  0.1]
```

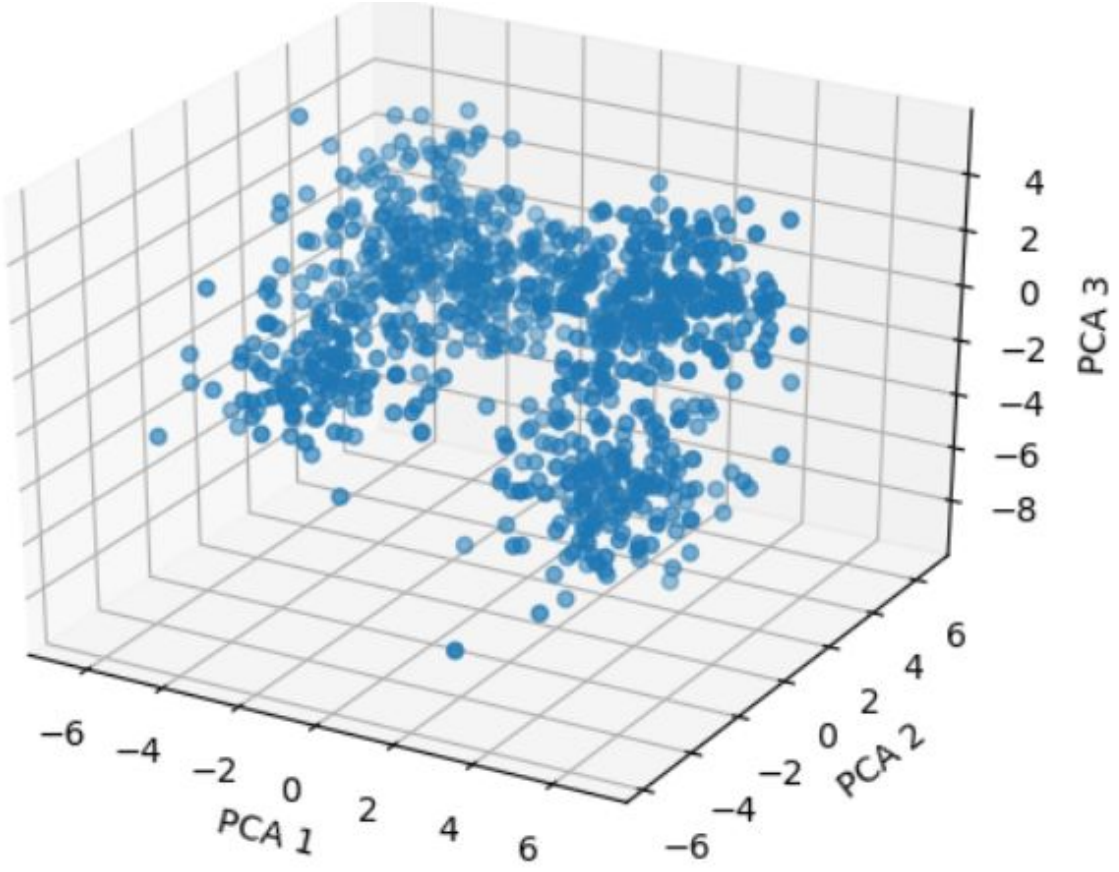
```
Eigenvector 3
[-0.2  0.1  0.1 -0.5  0.  -0.1 -0.2 -0.  -0.5 -0.2  0.  -0.   0.1  0.1
  0.2 -0.4  0.3  0.1  0.1  0.2]
```

```
Eigenvector 4
[-0.  -0.   0.  -0.   0.  -1.   0.2  0.1  0.  -0.  -0.  -0.  -0.  -0.1
  0.   0.  -0.1  0.   0.   0. ]
```

```
Eigenvector 5
[-0.1  0.5  0.3 -0.1  0.2  0.   0.  -0.1  0.1 -0.2 -0.  -0.2  0.2  0.1
 -0.4  0.3 -0.2 -0.1 -0.2  0.5]
```

Each component of the eigenvectors tells us how each attribute of the data is important. Now as we know that the first 5 eigenvalues are the highest and hence we only look at the first 5 eigenvectors. These eigenvectors tells us which attribute of the data will be important for clustering the data. Now we know that eigenvectors are descending in importance. So we look at the first eigenvector initially as it explains most of the data. As we can see, the 17th attribute, ‘tomato’ has a high value of 0.6 which is suggestive of the fact that ‘tomato’ may be an important attribute because it has values in the following eigenvectors too. Now as we can see the 14th attribute, ‘Salza’ has very low values in all the eigenvectors which leads us to believe that ‘salza’ may not be that important. No, there is no attribute that is zero in all the eigenvectors.

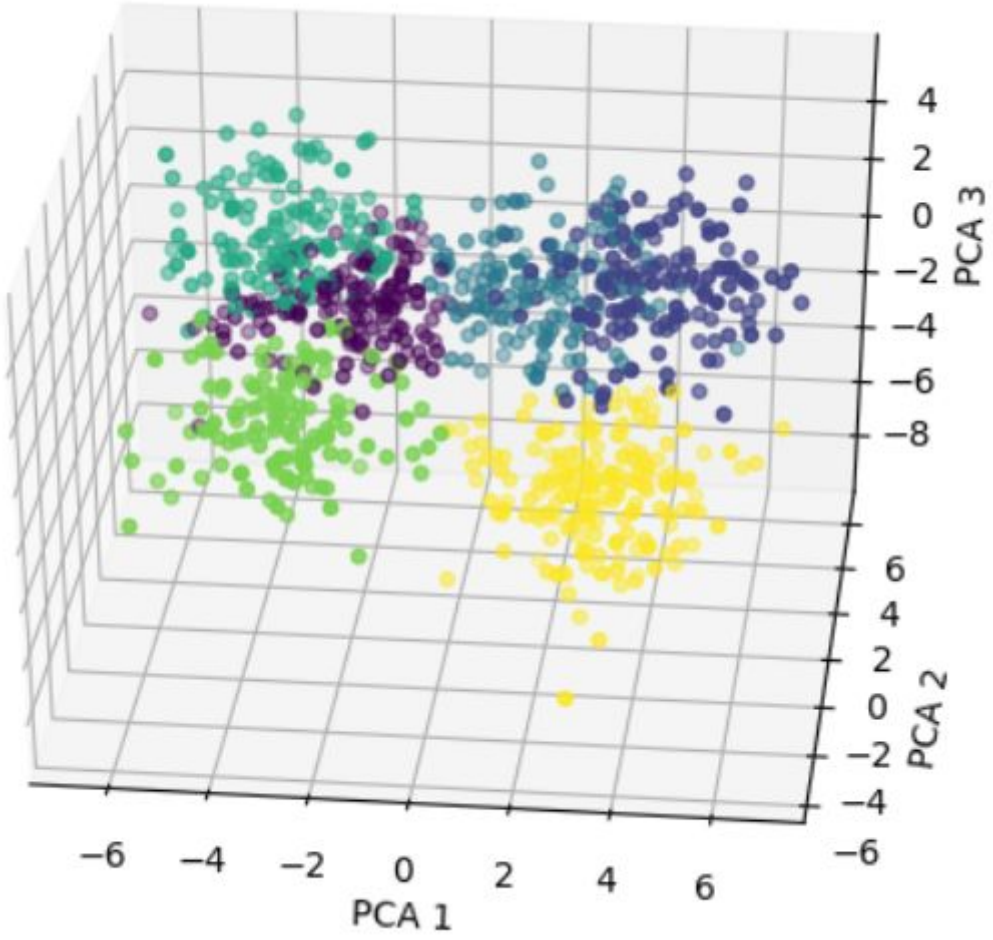
f. (2) Project the original Agglomeration data onto eigenvectors: [1, 2, 3]. PLOT: Generate a 3D plot of these newly projected points, and show a 3D scatter-gram of the data in 3D.



g. (2) In this new, projected, three dimensional space, perform k-Means clustering using a package of your choice, or the code from your previous homework. Do this for k=6. For simplicity, use the Euclidean distance.

What are the six cluster centers? Can you make sense of them?

```
The k-means cluster centers are:  [[-2.39879856  3.13873283 -0.68581544]
 [ 3.99667818  1.42982604  1.69013923]
 [ 1.44013798  3.37834578  0.08004271]
 [-3.69137198  2.99513306  1.89583829]
 [-3.0953735  -1.50393997 -1.15099783]
 [ 2.96672539  0.55307253 -4.84208476]]
```



h. (1) What evidence indicates that a k of 6 is a good choice? By inspection, of your 3D plot, how many clusters do you suspect?
As we can see from the 3D plot, all the points are clustered into 6 different clusters.

Part 3: Agglomeration:

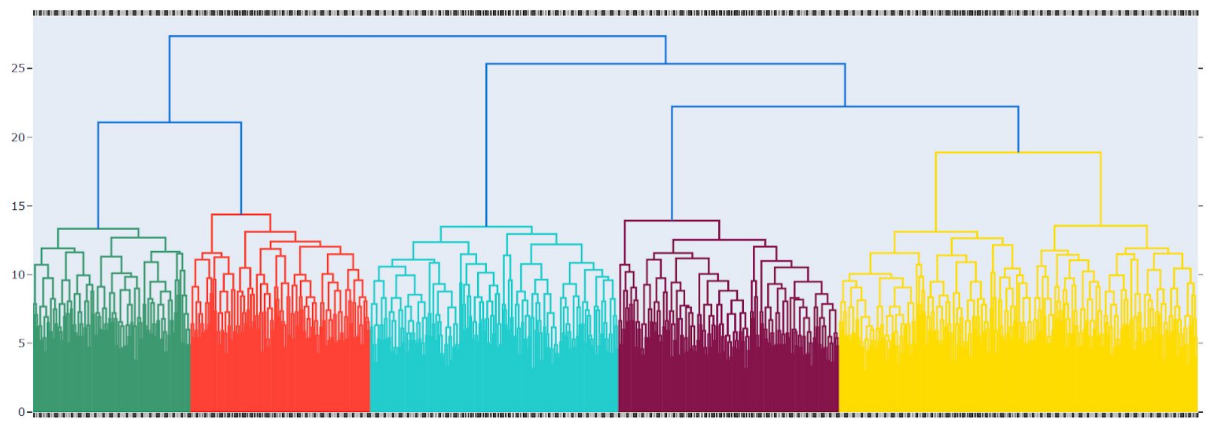
3.
d. (5) Note: At each step of clustering, two clusters are merged together. Track the size of the smallest of the two clusters that are merged together. There are questions about this later. Write down the size of the smallest cluster in the last 20 merges. For example, if we merge a cluster of size 30 with a cluster of size 10, you remember that a 10 was merged in. Cluster to completion. Record and report the size of the last 18 smallest clusters merged.
→ The following table shows the smaller cluster sizes for the last 20 merges.

Last 20 merges
347
17
50
138
129
1
149
3
69
63
3
9
82
75
34
48
4
1
29
24

e. (3) Based on agglomeration, how many clusters do you think are in the data? Why did you reach this conclusion? Support your guess. Can you support this guess with a dendrogram?

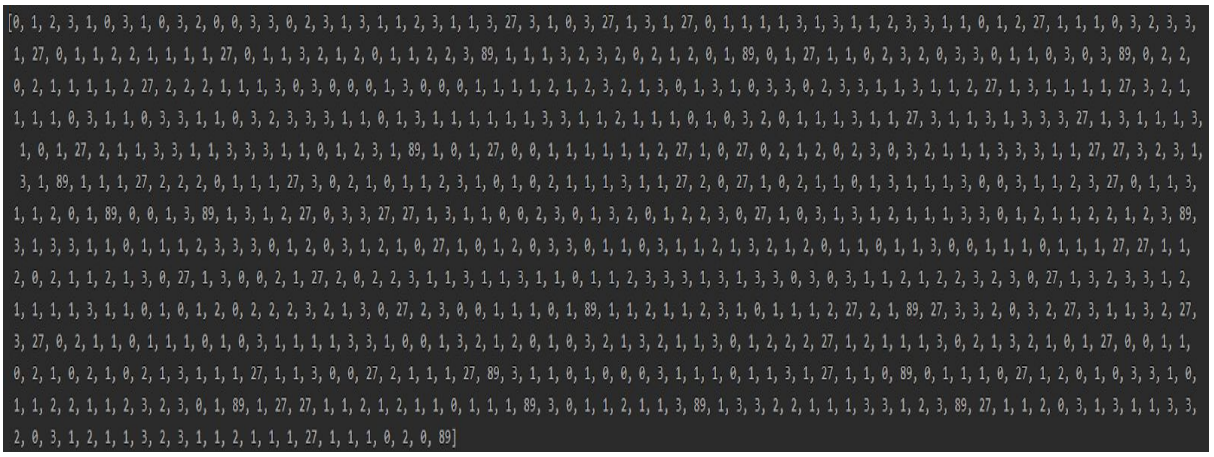
→ As we can see from the following figure, we can see that there is a jump in the cluster sizes from 3 to 149 in the last line. If we count the number of cluster lumps after that, we see that there are a total of 5 lumps. Hence according to the rule, there should be 5+1=6 clusters.

```
[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1,
1, 1, 1, 1, 1, 1, 2, 1, 2, 1, 1, 1, 2, 1, 1, 1, 2, 1, 1, 1, 2, 1, 2, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 1, 3, 2, 2, 2, 1, 1, 1, 2, 2, 4, 2, 1, 1, 1, 3, 1, 1, 3,
2, 3, 1, 2, 2, 1, 2, 1, 2, 1, 1, 1, 1, 2, 1, 1, 4, 2, 1, 1, 1, 1, 2, 2, 1, 2, 1, 2, 2, 3, 2, 1, 1, 1, 1, 2, 2, 1, 1, 2, 1, 2, 3, 1, 1, 2, 1, 1, 2, 1, 3, 1, 2, 1, 3, 1, 2, 3, 1, 1, 1, 1, 2, 1, 1,
2, 1, 2, 2, 1, 1, 1, 3, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 2, 1, 4, 1, 1, 3, 2, 2, 4, 4, 3, 5, 2, 4, 3, 3,
4, 2, 1, 2, 4, 1, 2, 1, 1, 3, 2, 2, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 3, 2, 1, 2, 1, 1, 1, 1, 3, 1, 4, 2, 5, 4, 1, 6, 4, 1, 5, 3, 2, 2, 3, 2, 2, 4, 1, 7, 2, 1, 3, 2, 3, 4, 3, 1, 2, 2, 2, 3, 3, 2,
4, 1, 1, 1, 1, 2, 1, 5, 1, 5, 3, 3, 3, 3, 1, 1, 3, 3, 6, 7, 4, 4, 5, 2, 1, 1, 2, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 4, 4, 2, 2, 2, 2, 1, 2, 1, 3, 4, 1, 3, 5, 2, 8, 2, 5, 2, 1, 2, 1, 2, 5, 7, 3, 5, 4,
6, 5, 3, 6, 2, 3, 3, 4, 1, 4, 3, 3, 5, 2, 7, 4, 4, 10, 6, 4, 2, 6, 6, 11, 3, 4, 10, 4, 5, 6, 2, 9, 4, 4, 3, 4, 3, 13, 1, 4, 4, 5, 5, 3, 8, 4, 12, 3, 4, 15, 1, 11, 4, 3, 16, 8, 8, 1, 4, 11, 9, 1,
7, 6, 5, 8, 5, 3, 1, 3, 6, 8, 3, 15, 7, 13, 11, 3, 9, 9, 11, 12, 19, 10, 18, 3, 15, 7, 1, 25, 32, 10, 22, 6, 12, 9, 25, 23, 9, 19, 14, 21, 36, 1, 14, 15, 1, 11, 36, 18, 44, 46, 23, 52, 24, 29, 1,
4, 48, 34, 75, 82, 9, 3, 63, 69, 3, 149, 1, 129, 138, 50, 17, 347]
```

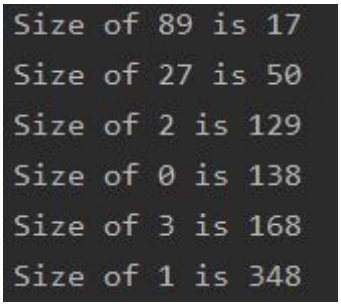


Now, according to the dendrogram of the data, we can see that the natural clustering would be 6 as the highest difference between two cluster heights cuts across 6 vertical lines.

Discussion Questions – Copy and paste these so that you can understand the context of your answers later on:



1. (1) When you have clustered to six clusters, report the size of each cluster, from lowest to highest.



2. (1) When you have clustered to six clusters, report the average prototype of these six clusters. What is their relationship to the Eigenvectors?

3. (1) What typifies each of the six clusters? What name should we give each of these prototypes?

4. (1) What advantage and/or disadvantage is there in performing k-Means clustering on data which has been projected using PCA?
→ While performing k means clustering on data which has been projected using PCA, the computation is faster as there are fewer attributes to worry about. PCA is a pre-processing technique which may or may not help us. The only way to find out is to do it. PCA works best on ratio and interval data.

5. (1) What advantage and/or disadvantage would there be to performing Agglomeration on data which has been projected onto Principal components?
→ PCA is a preprocessing technique that is mainly used for dimensionality reduction. For this data, we can see that the number of attributes has been reduced from 20 to 3. This helps a lot in agglomeration. Firstly, while calculating the linkage distance the computations are much faster. Further, we are saved from unnecessary calculations as we have already projected the data on to the eigenvectors that best explain the data. However, there is one disadvantage that information is lost.

6. (5) Write a conclusion about what you learned overall. If each of you learned different things, tell me what each of you learned.
→ Overall, the assignment was very interesting as well as challenging. But i think the Professor's decision to combine both the PCA and the agglomeration homework was a great idea because it helped us in learning both PCA and agglomeration as part of a process. A process which included cleaning the data and preprocessing it by identifying the Principal Components and then projecting the original data onto these. Then performing clustering on this projected data using either K-means or hierarchical clustering. During the course of this assignment, we learnt many things like how PCA actually performs dimensionality reduction, how agglomeration works, what happens if the matrices are not compatible to multiplied and smaller things like how to calculate the second minimum of a list using the 'nsmallest' function from the heapq package. Further, we learned that we can store tuple values as keys of a dictionary. This made it easier to store and access the key. We also learnt how to create a dendrogram for a given data. An interesting that we learned was that if we just use the plotly.FigureFactory's create_dendrogram function to create a dendrogram, it automatically performs the hierarchical clustering for us. Lastly, we learned about guessing the number of clusters from the list of sizes of smaller clusters in a merge. This was a question on the midterm which we had read about. For this assignment, we used that knowledge.