**Homework5**

**20200639 Chae Woojin**

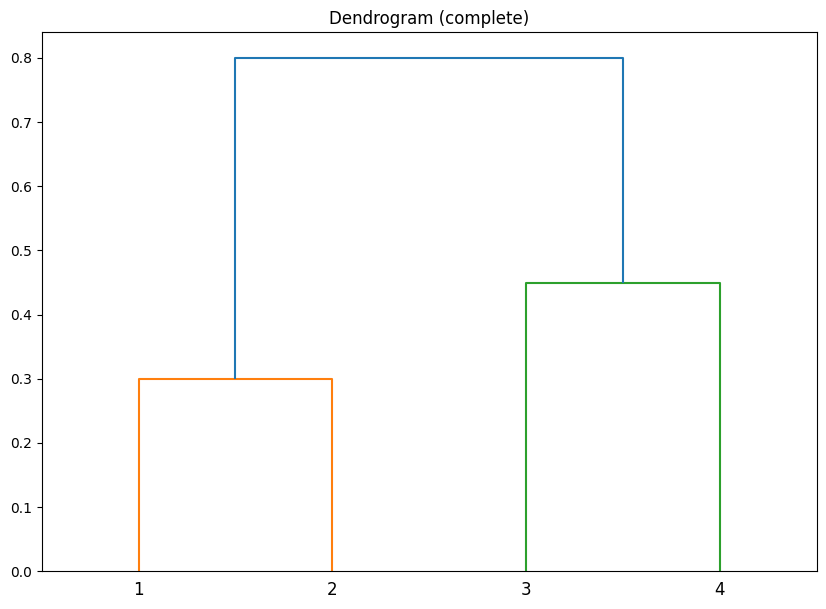
**Chapter 12:** 2, 3, 8, 9, 13

**Chapter 13:** 1, 2, 4, 6, 8

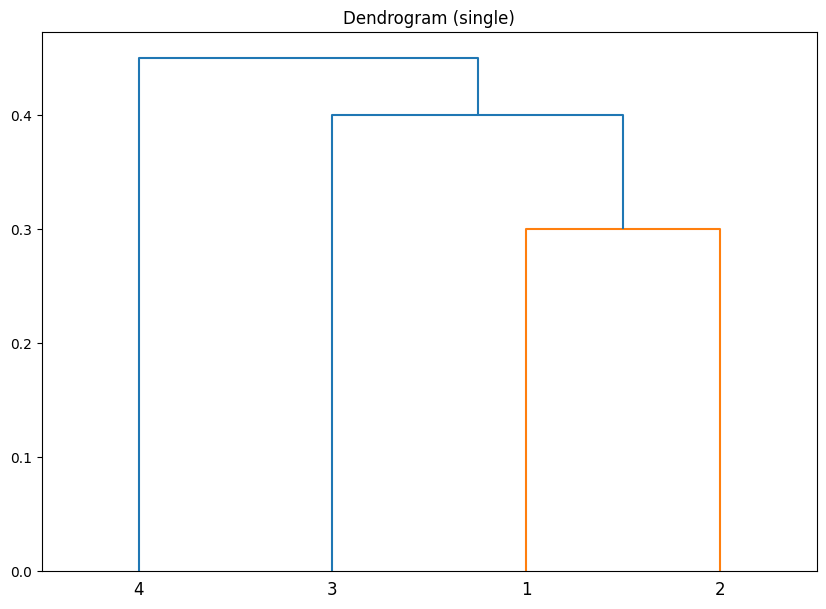
**Chapter 12**

**2.**

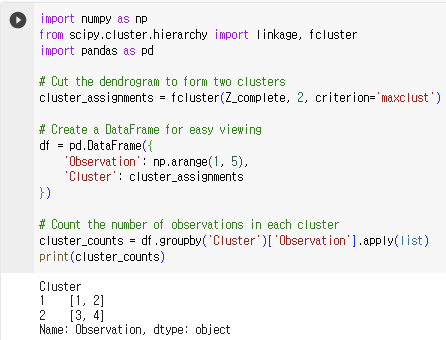
(a)



(b)

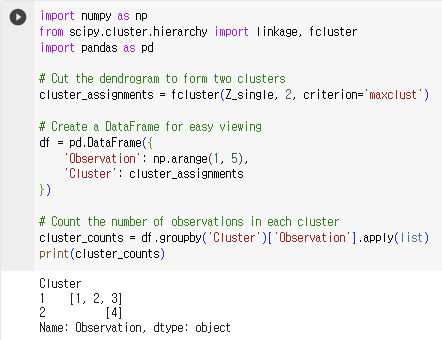


(c)



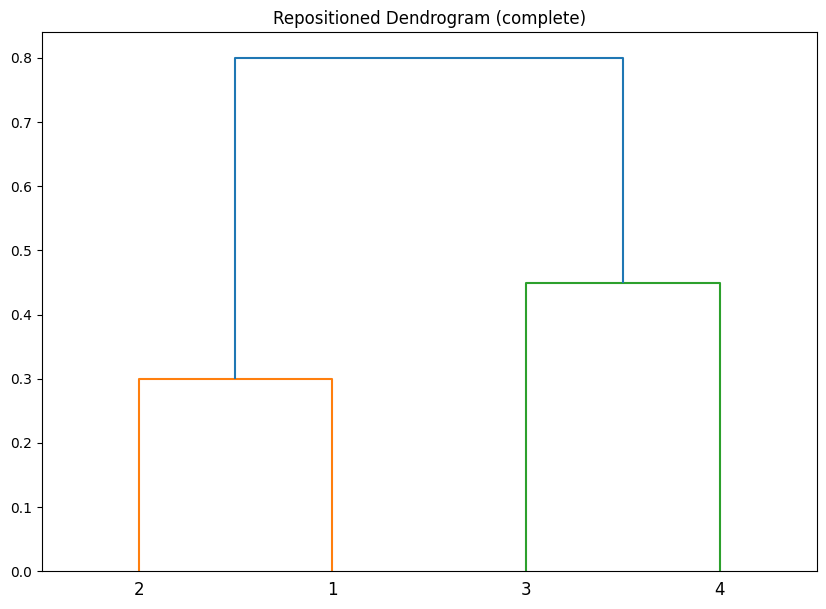
1st, 2nd observations are in the same cluster, and 3rd, 4th observations are in the same cluster.

(d)



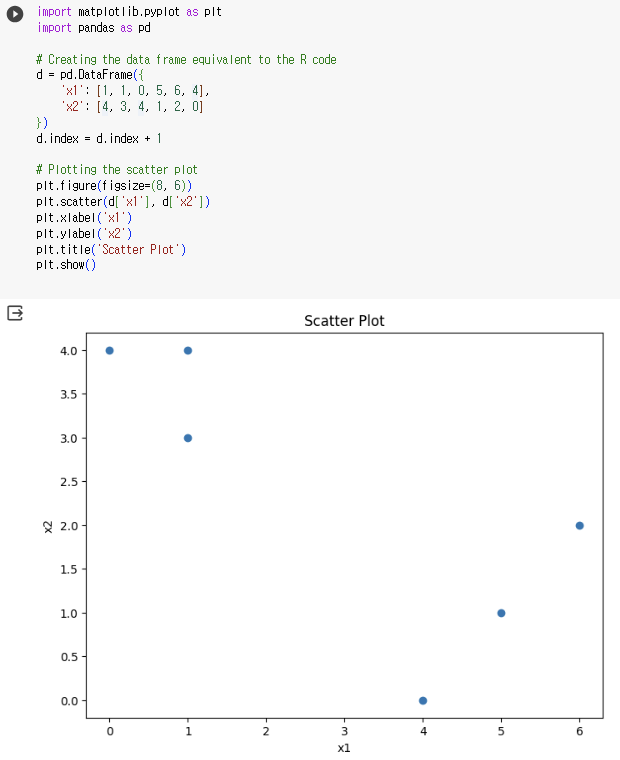
1st, 2nd, and 3rd observations are grouped into a same cluster, and the 4th observation mapped to the other one.

(e)

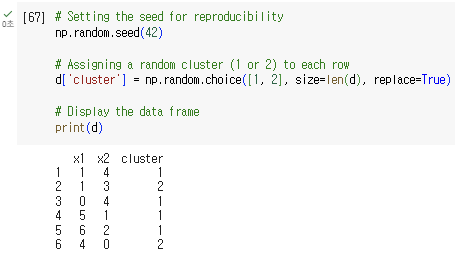


**3.**

(a)

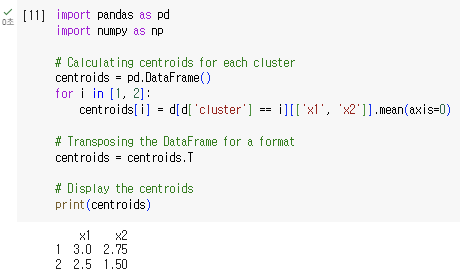


(b)



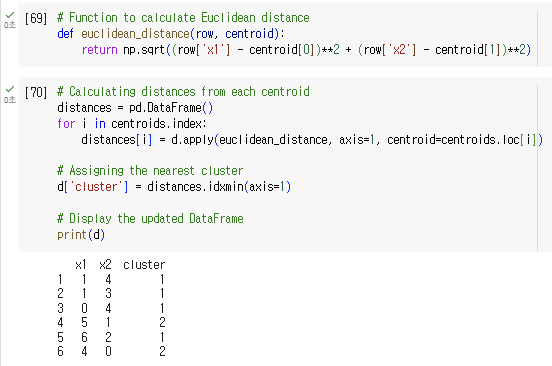
1st, 3rd, 4th, and 5th observations are grouped into the same cluster, and the other observations are grouped into the other cluster.

(c)



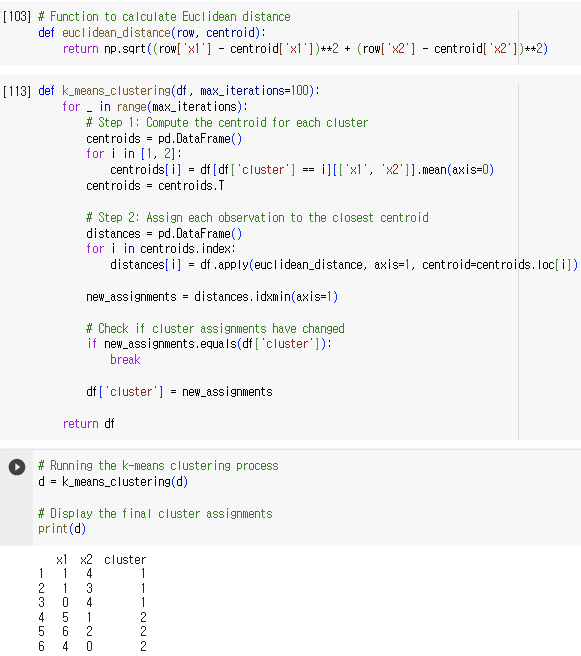
Centroids of two clusters are , respectively.

(d)



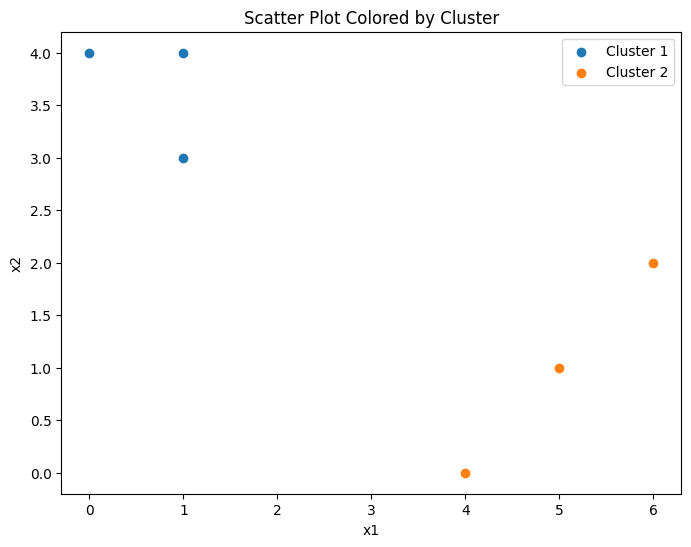
After new assignment, 1st, 2nd, 3rd, and 5th observations are grouped into same cluster, and 4th, and 6th observations are grouped into the same cluster.

(e)



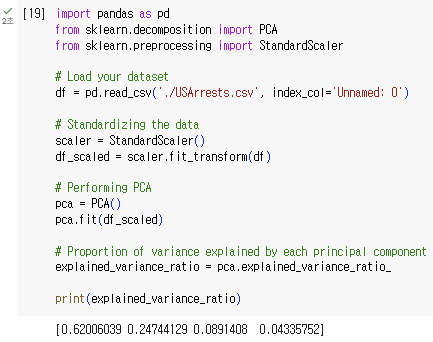
After equilibrium reached, one can see that 1st, 2nd, and 3rd observations are grouped into the same cluster, and the others are grouped into the same cluster.

(f)



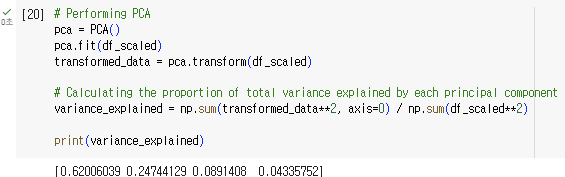
**8.**

(a)



PVE is [0.62006039, 0.24744129, 0.0891408, 0.04335752].

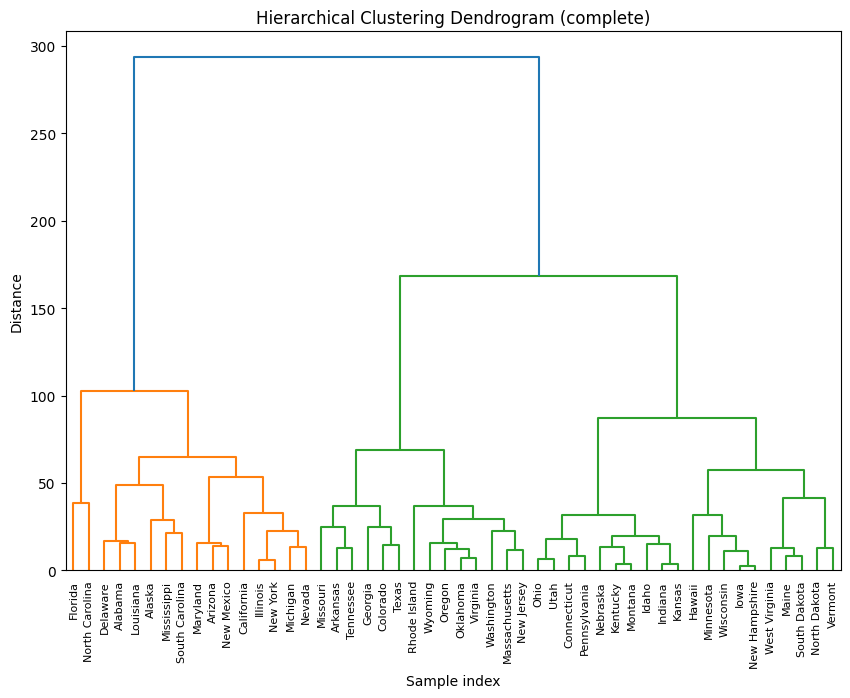
(b)



PVE is [0.62006039, 0.24744129, 0.0891408, 0.04335752] which is identical to the result of (a).

**9.**

(a)



(b)

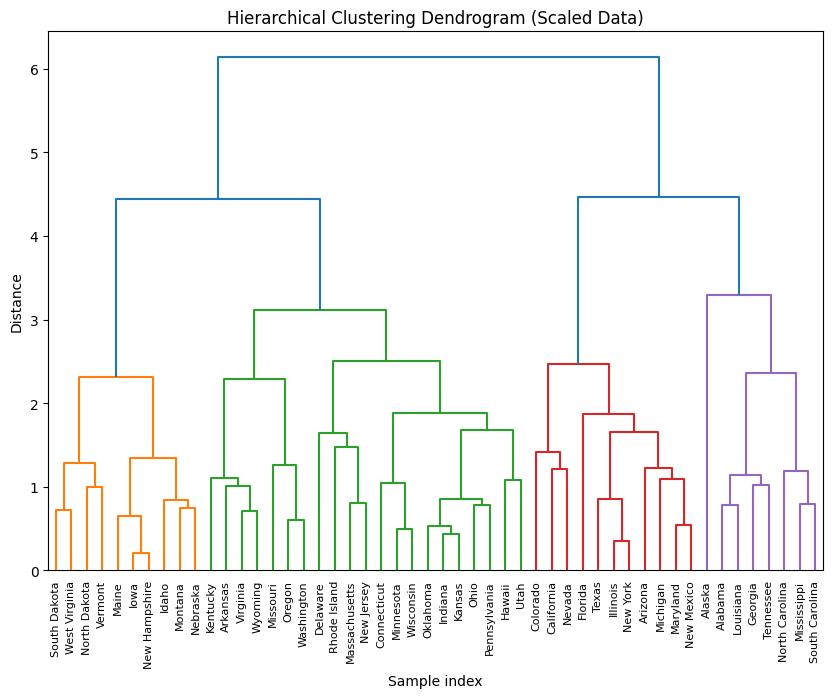


Cluster 1: ['Alabama', 'Alaska', 'Arizona', 'California', 'Delaware', 'Florida', 'Illinois', 'Louisiana', 'Maryland', 'Michigan', 'Mississippi', 'Nevada', 'New Mexico', 'New York', 'North Carolina', 'South Carolina']

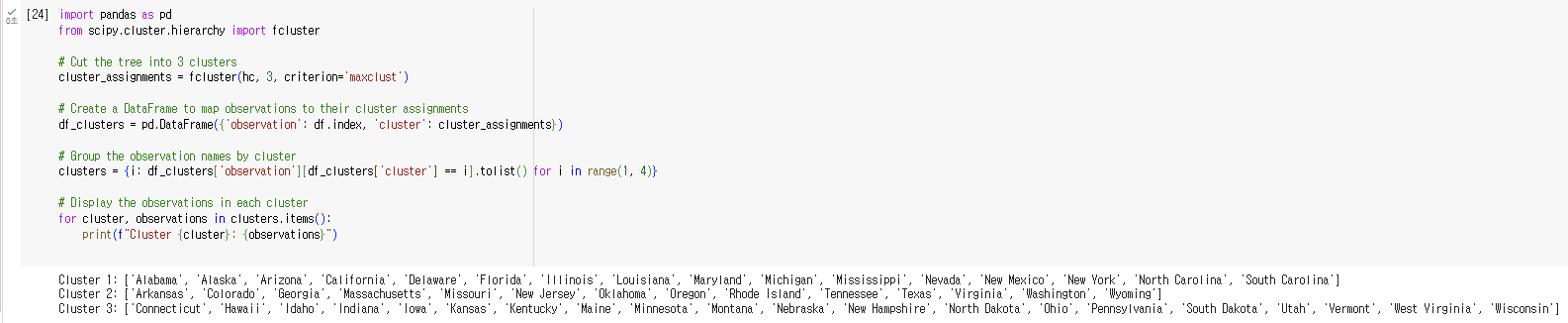
Cluster 2: ['Arkansas', 'Colorado', 'Georgia', 'Massachusetts', 'Missouri', 'New Jersey', 'Oklahoma', 'Oregon', 'Rhode Island', 'Tennessee', 'Texas', 'Virginia', 'Washington', 'Wyoming']

Cluster 3: ['Connecticut', 'Hawaii', 'Idaho', 'Indiana', 'Iowa', 'Kansas', 'Kentucky', 'Maine', 'Minnesota', 'Montana', 'Nebraska', 'New Hampshire', 'North Dakota', 'Ohio', 'Pennsylvania', 'South Dakota', 'Utah', 'Vermont', 'West Virginia', 'Wisconsin']

(c)



(d)



Cluster 1: ['Alabama', 'Alaska', 'Arizona', 'California', 'Delaware', 'Florida', 'Illinois', 'Louisiana', 'Maryland', 'Michigan', 'Mississippi', 'Nevada', 'New Mexico', 'New York', 'North Carolina', 'South Carolina']

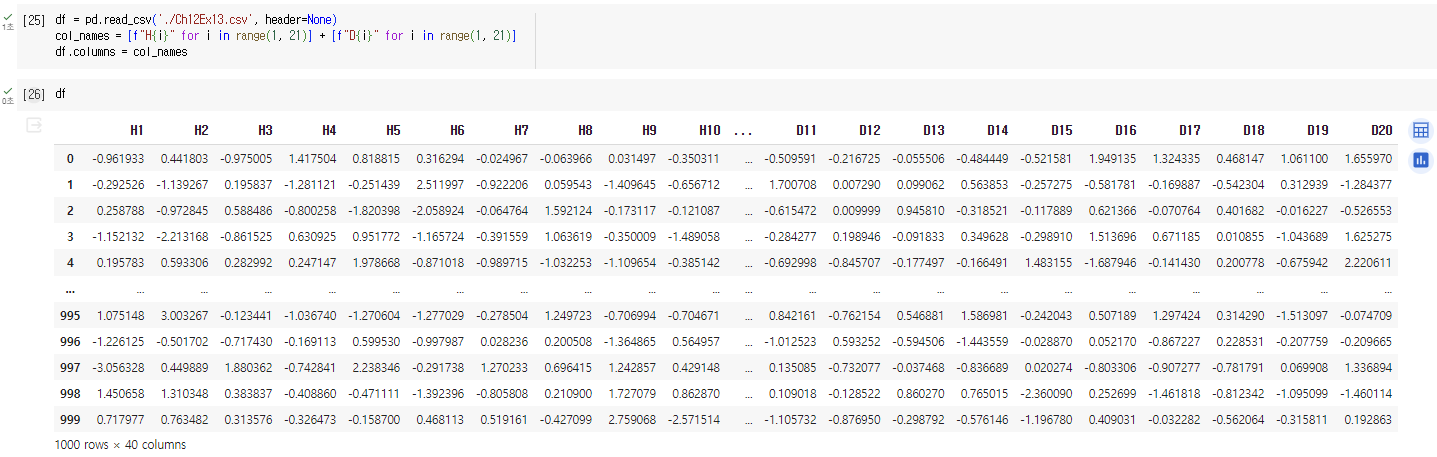
Cluster 2: ['Arkansas', 'Colorado', 'Georgia', 'Massachusetts', 'Missouri', 'New Jersey', 'Oklahoma', 'Oregon', 'Rhode Island', 'Tennessee', 'Texas', 'Virginia', 'Washington', 'Wyoming']

Cluster 3: ['Connecticut', 'Hawaii', 'Idaho', 'Indiana', 'Iowa', 'Kansas', 'Kentucky', 'Maine', 'Minnesota', 'Montana', 'Nebraska', 'New Hampshire', 'North Dakota', 'Ohio', 'Pennsylvania', 'South Dakota', 'Utah', 'Vermont', 'West Virginia', 'Wisconsin']

One can see that scaling the variables have an effect on clustering results. Therefore, scaling ought to be informed by the unique attributes of the dataset under consideration. In the current scenario, the data fields are not uniform in terms of units (**Murder (Assault, Rape)**: **numeric number** of arrests per 100,000 people pertaining to murder (assault, rape) crime, **UrbanPop:** **numeric** **percent** urban population). Given this context, scaling the data prior to clustering seems to be a more appropriate approach.

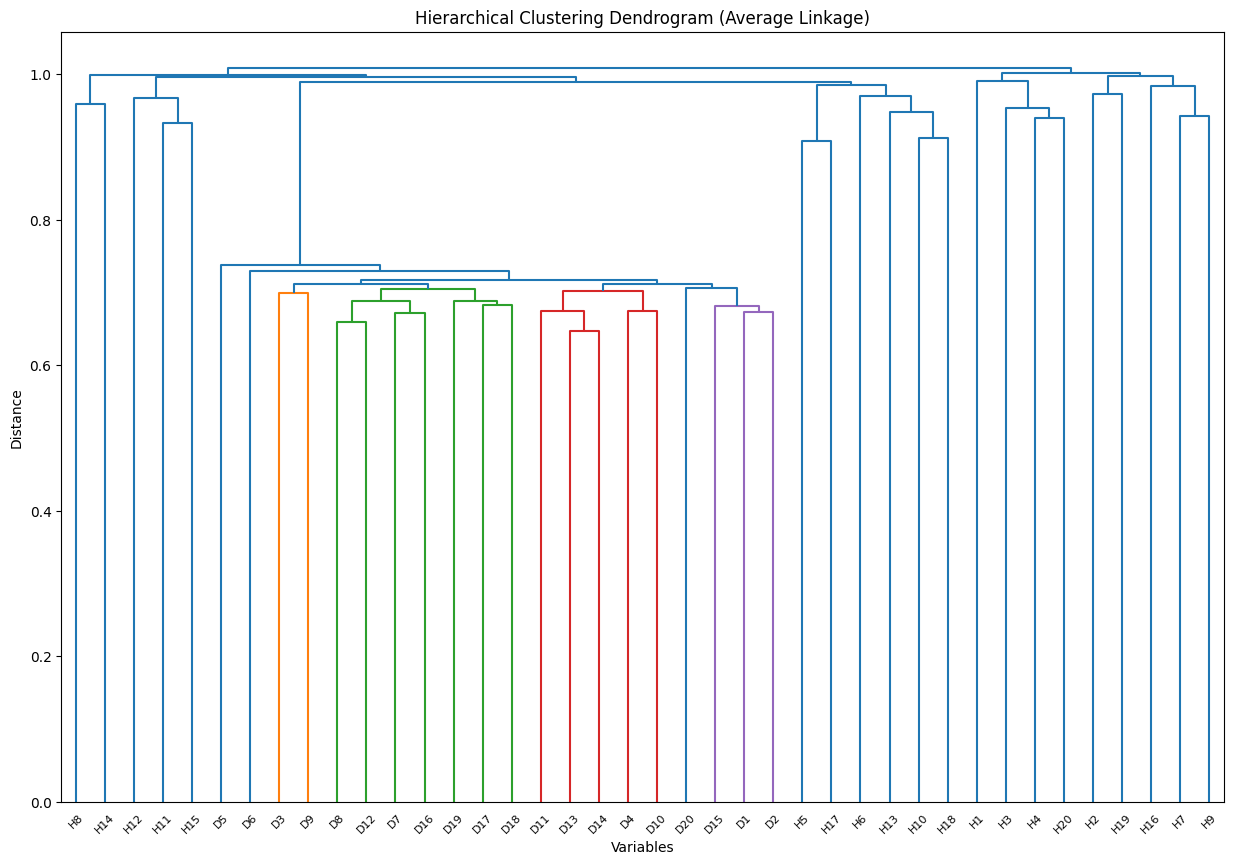
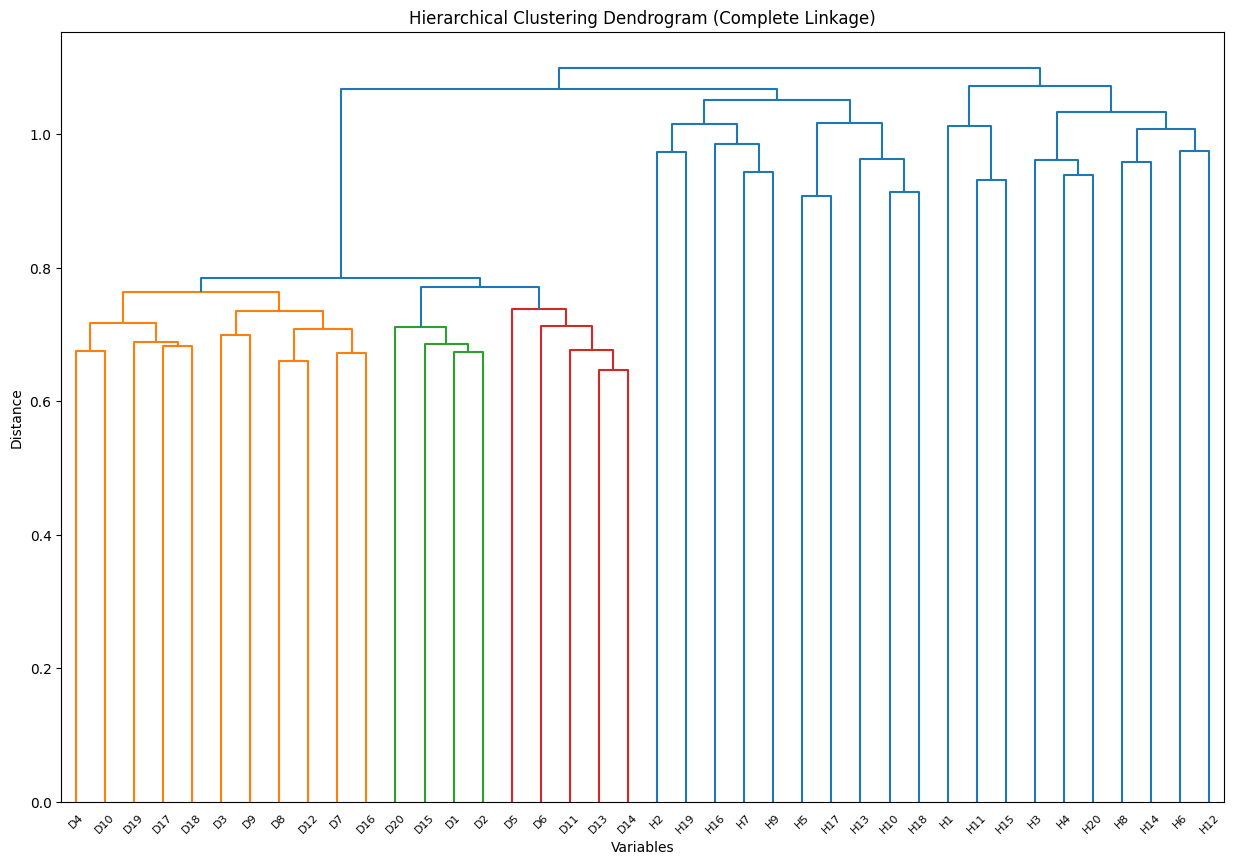
**13.**

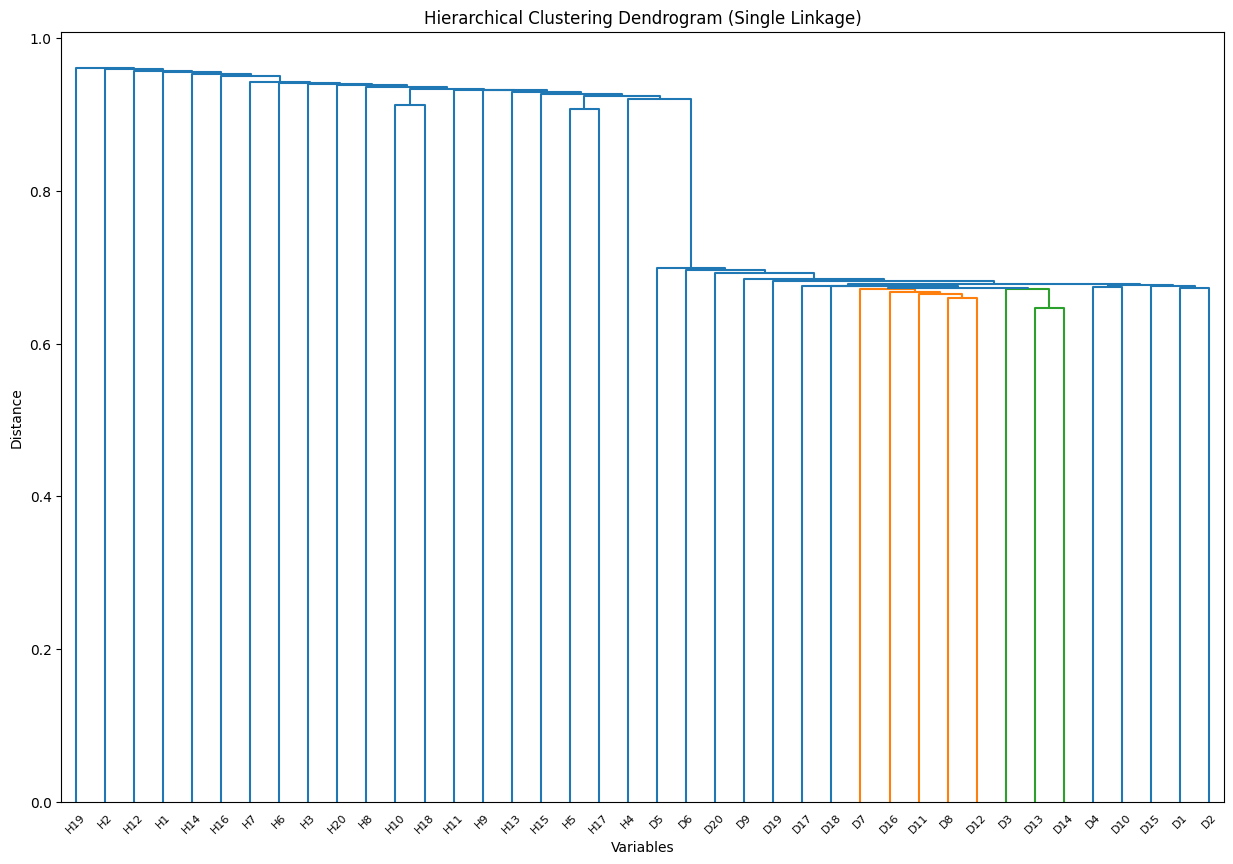
(a)



This code reads the whole raw data (1000 rows). Column H1~H20 represents samples from 20 healthy patients, and remaining columns (D1~D20) denotes the ones from 20 diseased group.

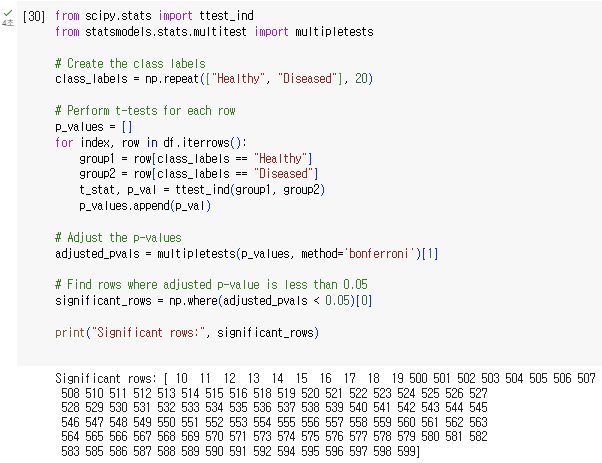
(b)





The samples are clearly divided into two distinct groups. However, it's crucial to understand that the degree of this division is affected by the selected linkage method. In particular, with average clustering, all disease samples cluster together, and they are encompassed within a portion of the healthy samples.

(c)

****

The group comparison test provides significant genes that appear to give different results between the two groups. Test results which reveal p-values under 0.05 are deployed as above. One can conclude that corresponding genes to these row indices differ the most across the two groups.

**Chapter 13**

**1.**

(a) We expect

(b) Let represents the number of TypeⅠerrors when performing m independent null hypothesis tests, all of which are true. Then,

(c) If the p-values for the two tests are perfectly correlated, we can regard them as a single test. Therefore, we can expect FWER to be less than the case when two tests are independent. This can be explained mathematically, as . In a similar manner, one can conclude that if two tests are positively correlated, FWER will be lower than . This is because positively correlated tests have the effect of reducing the effective number of tests.

(d) If the p-values for the two tests are negatively correlated, the family-wise error rate will tend to be higher compared to the case where the tests are independent. This is because it leads to an increased likelihood of making TypeⅠerror in one or both tests, which raises the overall FWER. When the p-values are negatively correlated, it becomes more likely that one of the tests will produce a small p-value even if the null hypothesis is true for both tests. This is because the negative correlation increases the chance that one test will produce a significant result when the other does not. Consequently, the FWER is more likely to be elevated compared to the independent case ().

**2.**

(a) follows a Bernoulli distribution with probability (i.e,)

(b)

(c) Using the fact that binomial distribution has standard deviation ,

the standard deviation of the number of Type 1 error(’s) we will make is .

**4.**

(a) We should reject every null hypothesis with p-value under 0.05. Therefore, should be rejected.

(b) If we want to control FWER at level 0.05, we can apply Bonferroni’s method as a conservative approach. Using this approach, every null hypothesis with p-value under 0.005 will be rejected. Therefore, should be rejected.

(c) Using Benjamini-Hochberg Procedure, we can control the FDR. As we want to control FDR at level 0.05, we have to find , where ’s are the sequence of p-values sorted in ascending order. In this case, it is {0.0009, 0.0011, 0.004, 0.006, 0.017, 0.031, 0.07, 0.11, 0.32, 0.90}. Then, simple evaluation gives . Finally, we have rejects null hypotheses having p-value Therefore, will be rejected.

(d) Similar to (c), using this procedure with different threshold 0.2 will reject .

(e) 20% of these rejections are expected to be false positives.

**6.**

For each of the three panels in Figure 13.3, there are always 8 positives(red) and 2 negatives(black).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Panel | FP | FN | TP | TN | Type 1 (FP) | Type 2 (FN) |
| 1 | 0 | 1 | 7 | 2 | 0 | 1 |
| 2 | 0 | 1 | 7 | 2 | 0 | 1 |
| 3 | 0 | 5 | 3 | 2 | 0 | 5 |

(a) Under Bonferroni procedure to control the FWER at level 0.05, classification result of each panel display can be summarized as follows.

(b) Under Holm procedure to control FWER at level 0.05, classification result of each panel display can be summarized as follows.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Panel | FP | FN | TP | TN | Type 1 (FP) | Type 2 (FN) |
| 1 | 0 | 1 | 7 | 2 | 0 | 1 |
| 2 | 0 | 0 | 8 | 2 | 0 | 0 |
| 3 | 0 | 0 | 8 | 2 | 0 | 0 |

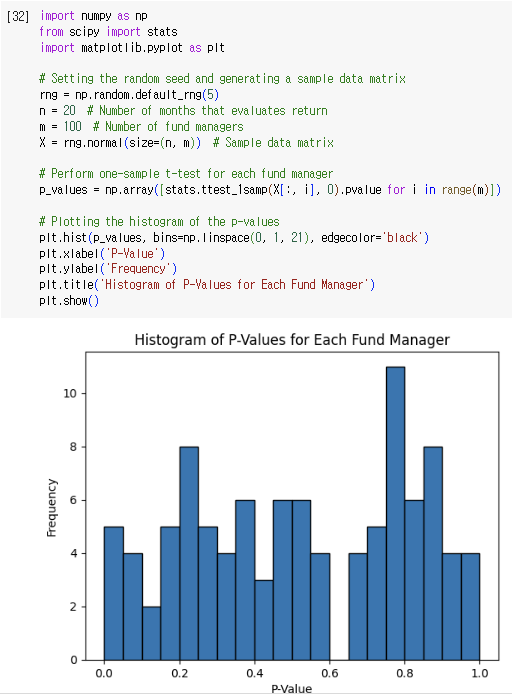
(c) False discovery rate is the expected ratio of false positives to total positives. As the number of false positives are 0 in every panel, false discovery rates are also 0.

(d) False discovery rate for panel 1, 2, and 3 are 0 with the same reason in the case of (c).

(e) Although we lower the threshold of FWER at level 0.001, the number of false positives doesn’t change. Therefore, the results would not change.

**8.**

(a)

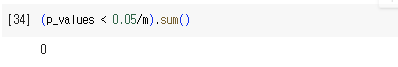


(b)



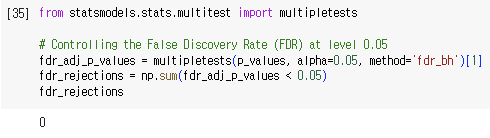
The number of rejections if we control Type 1 error at level 0.05 was 5.

(c)



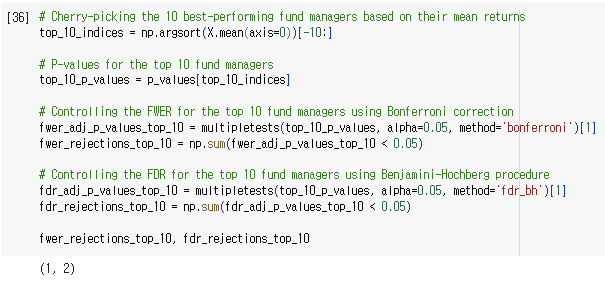
The number of rejections if we control FWER at level 0.05 is 0.

(d)



The number of rejections if we control FDR at level 0.05 is 0.

(e)



The number of rejections under cherry-picked data if we control FWER, and FDR at level 0.05 are 1, and 2 respectively.

(f)

It could be misleading, because we selectively performed the test under cherry-picked data that are likely to be rejected. Therefore, results from this test cannot represent the whole data.