Assignment 4 Chunlei Zhou

Q1
(a)
(i)

If we do not specify any prior probability, prior = c(0.6,0.4) will be used, because it is the original proportion.

(ii)

$$When\ p_{max} \geq 0.75, the\ correct\ classification\ rate\ is\ 0.8974359$$
$$When\ p_{max} < 0.75, the\ correct\ classification\ rate\ is\ 0.6363636$$

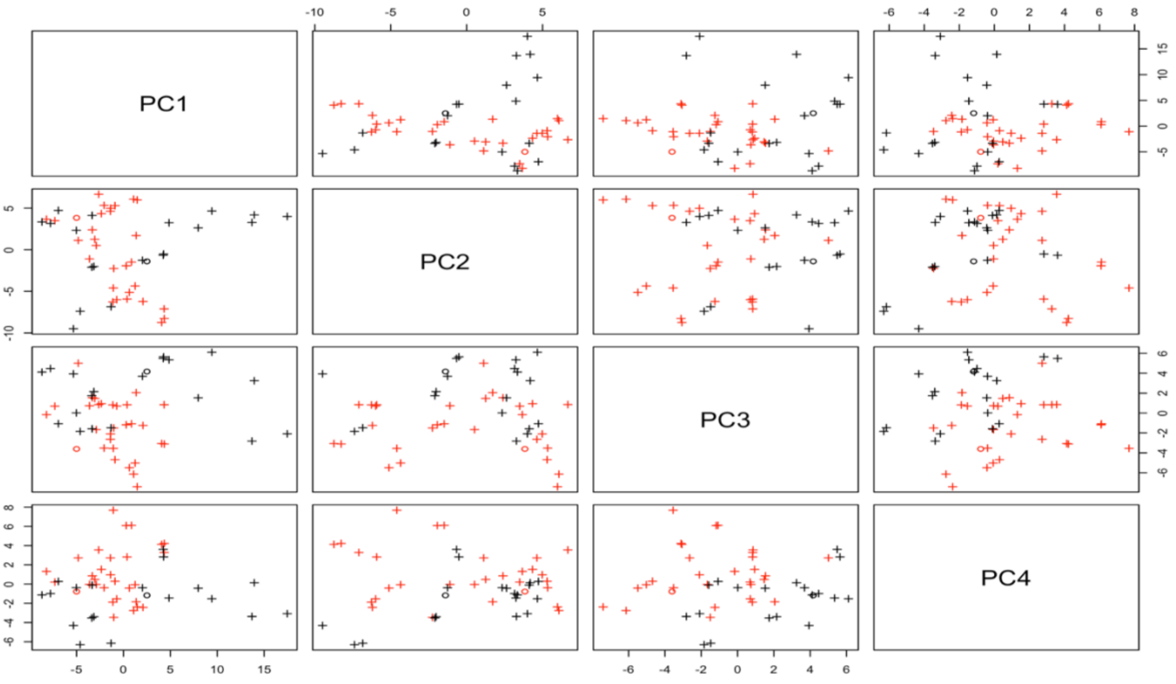
When apply Wilcoxon rank-sum test, the result is as follow:

```
Wilcoxon rank sum test

data:  pc and pnc
W = 260, p-value = 0.01336
alternative hypothesis: true location shift is not equal to 0
```

We can conclude that there is a difference in the distribution of p_{max} between observations that were correctly classified and those that weren't.
P-value is 0.01336.

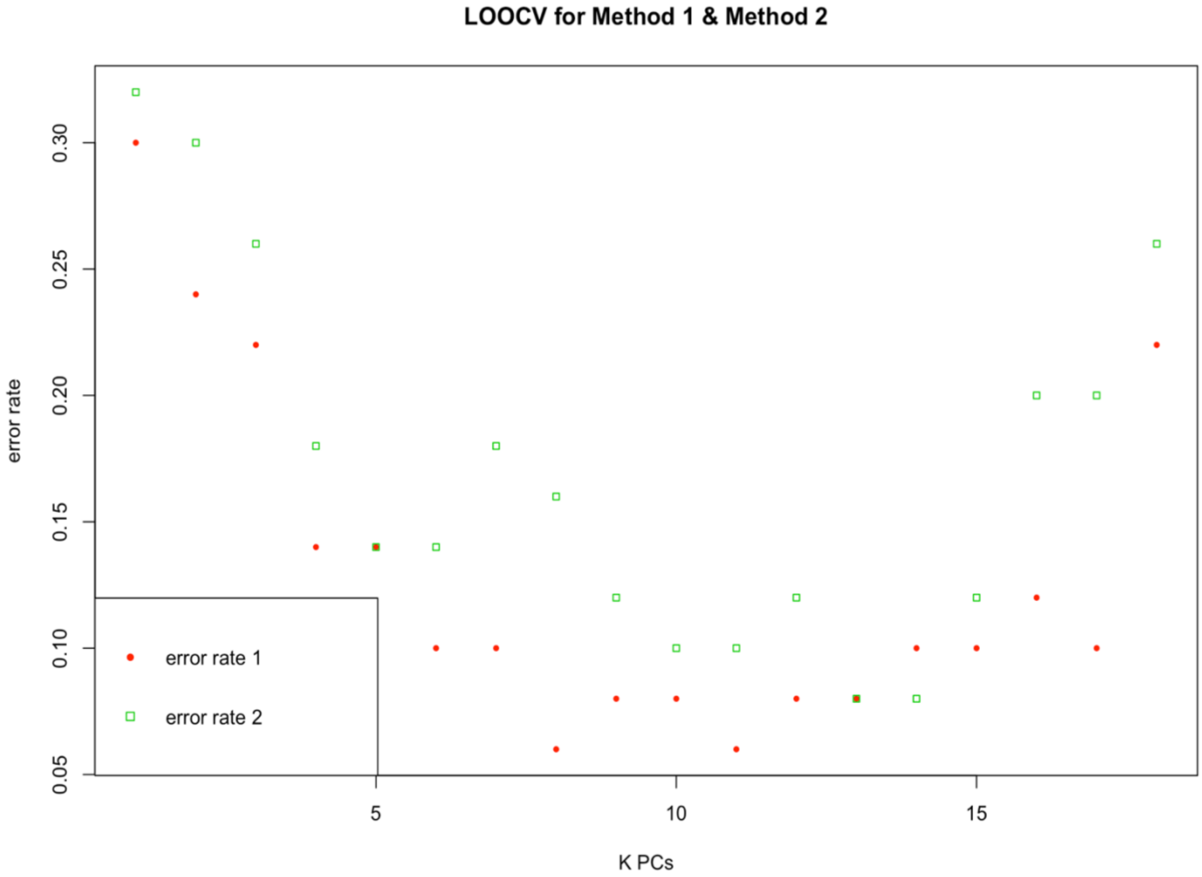
(iii)



Based on the plot, the values of the correctly classified observations are quite different from each other. The values of incorrectly classified observations are quite similar to each other.
Most values of the incorrectly classified observation on PC1 are less than 0, while their counterpart on PC2 are larger than 0. Values of the incorrectly classified observation on PC3 and PC4 are quite similar with each other.

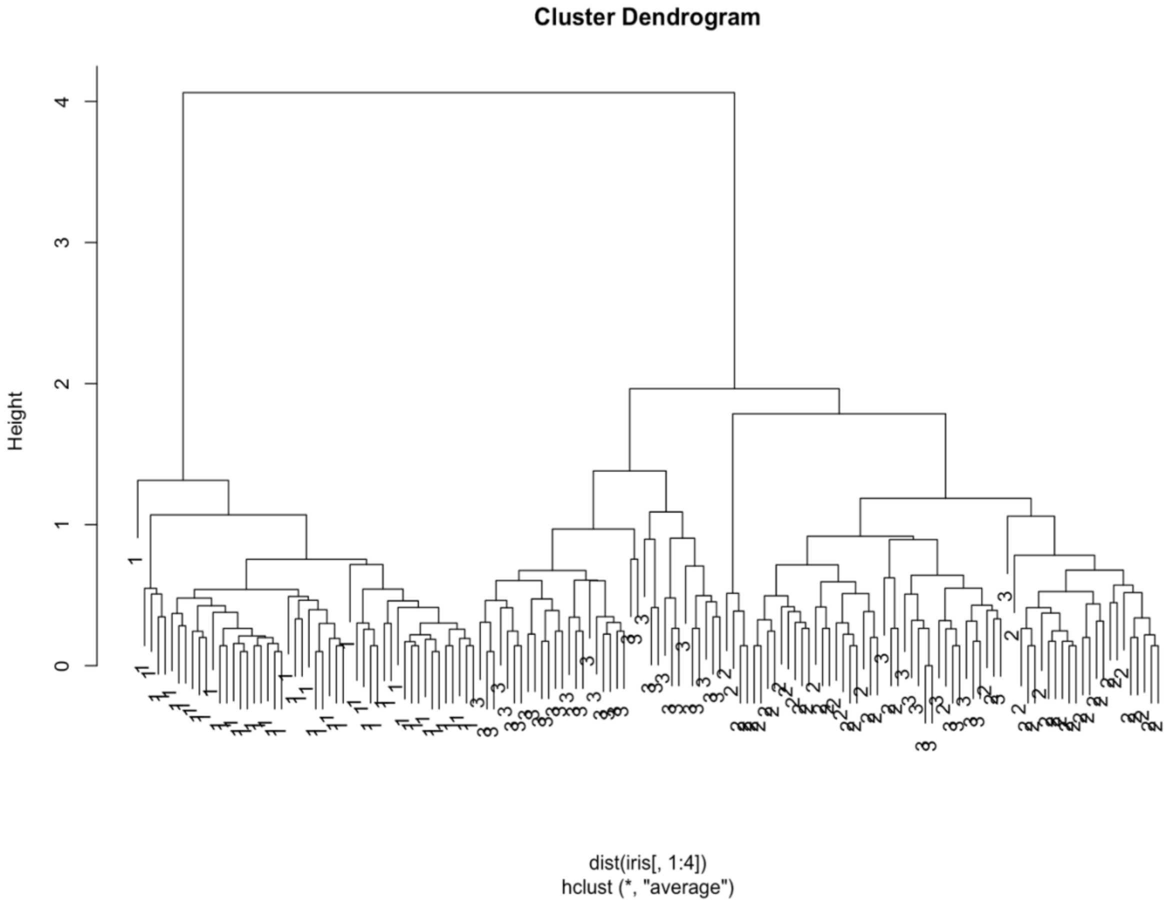
(b)
Error rate for method 1:
0.30, 0.24, 0.22, 0.14, 0.14, 0.10, 0.10, 0.06, 0.08, 0.08, 0.06, 0.08, 0.08, 0.10, 0.10, 0.12, 0.10, 0.22
Error rate for method 2:
0.32, 0.30, 0.26, 0.18, 0.14, 0.14, 0.18, 0.16, 0.12, 0.10, 0.10, 0.12, 0.08, 0.08, 0.12, 0.20, 0.20, 0.26

Plot:



The error rate of the first method is lower than that of the second method.
Based on different method, the recommended number of principal components will be different. If we use method 1, then the recommended number of principal components K will be 8. If we use method 2, then the recommended number of principal components K will be 13.

Q2
(b)

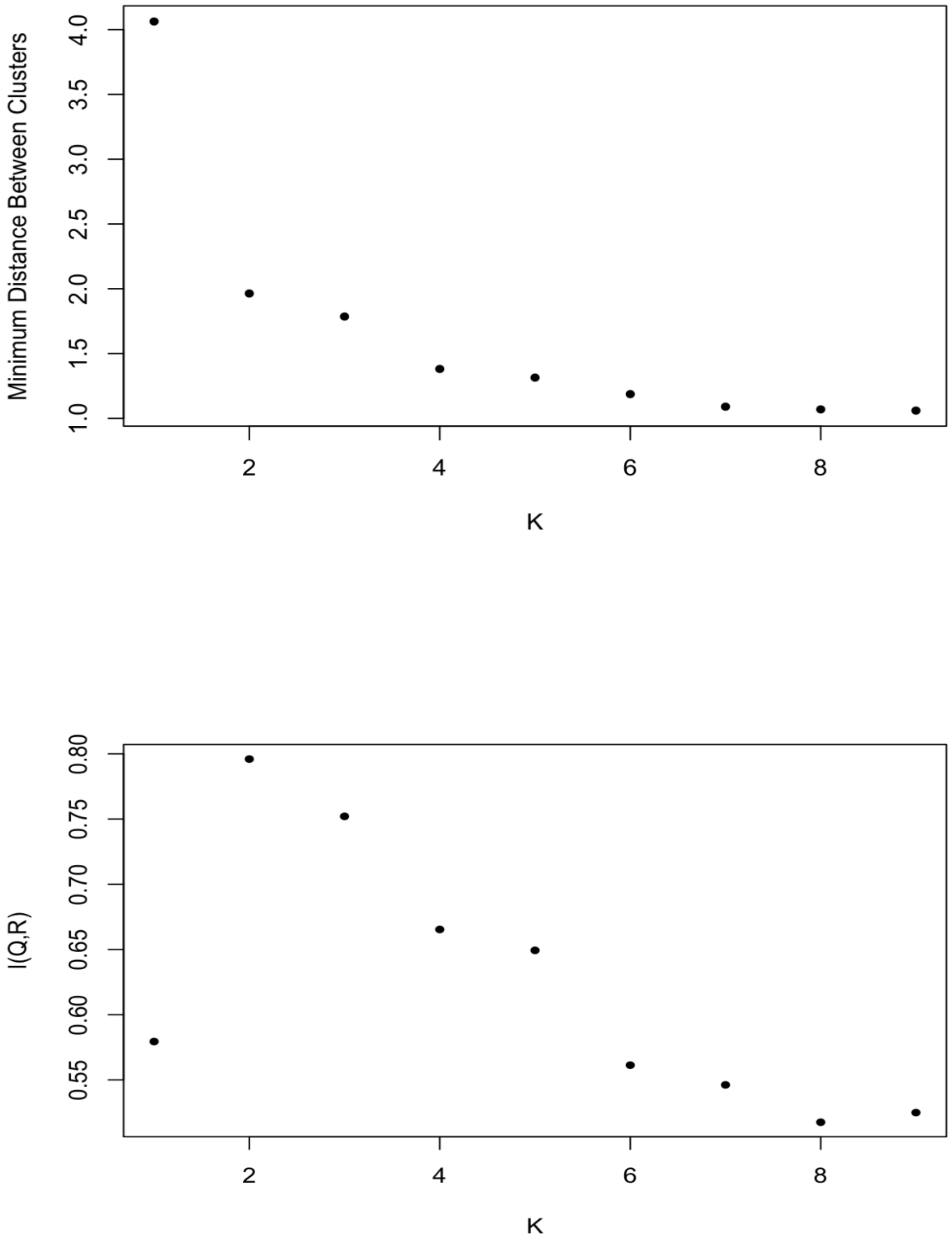


(c)

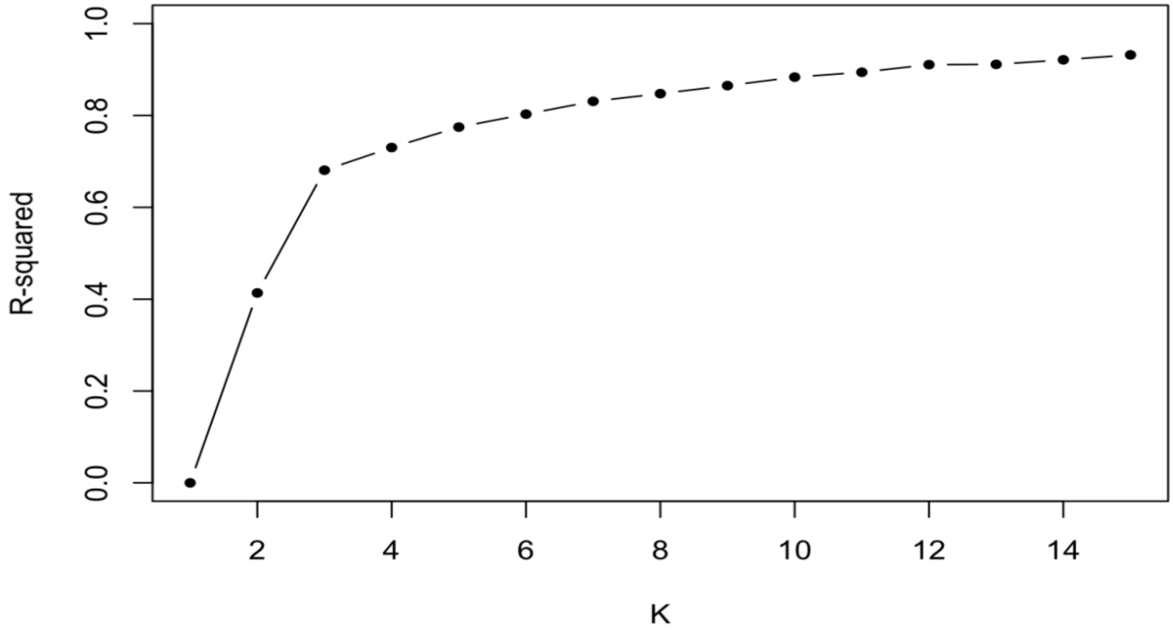
```
[1] "The smallest distance for K = 2 is 4.06268268611804"
[1] "The smallest distance for K = 3 is 1.96361408627465"
[1] "The smallest distance for K = 4 is 1.78556648202279"
[1] "The smallest distance for K = 5 is 1.38099373932928"
[1] "The smallest distance for K = 6 is 1.3141878740166"
[1] "The smallest distance for K = 7 is 1.18677850034338"
[1] "The smallest distance for K = 8 is 1.09005268813436"
[1] "The smallest distance for K = 9 is 1.06920843946017"
[1] "The smallest distance for K = 10 is 1.05978781496254"
```

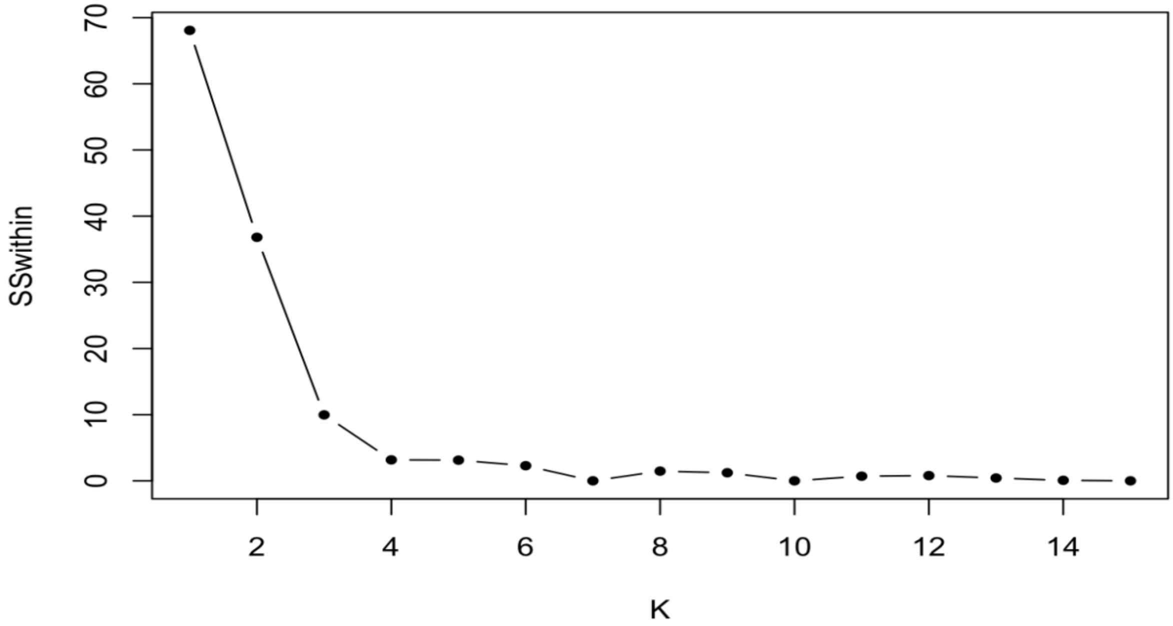
(d)
I(Q,R):
0.5793802, 0.7959816, 0.7520438, 0.6653148, 0.6492963, 0.5612752, 0.5461426, 0.5174973, 0.5248961

(e)


Minimum Distance Between Clusters
1.0
1.5
2.0
2.5
3.0
3.5
4.0
2
4
6
8
K
I(Q,R)
0.55
0.60
0.65
0.70
0.75
0.80
2
4
6
8
K

Q3
(a)





Based on the plot, the actual number of clusters might be 3 or 4.

(b)

```
[1] "K= 3 , R-squared= 0.680814667302004"
```

Brand names list 1 for brands belong to cluster 1:

```
bn1:
[[1]]
[1] "Frosted Mini-Wheats"       "Puffed Rice"
[3] "Raisin Squares"            "Shredded Wheat 'n'Bran"
[5] "Shredded Wheat spoon size"
```

Brand names list 2 for brands belong to cluster 2:

```
bn2:
[[1]]
 [1] "Apple Cinnamon Cheerios" "Apple Jacks"
 [3] "Cap'n'Crunch"            "Cheerios"
 [5] "Cinnamon Toast Crunch"   "Cocoa Puffs"
 [7] "Corn Chex"               "Corn Flakes"
 [9] "Corn Pops"               "Count Chocula"
[11] "Crispix"                 "Double Chex"
[13] "Froot Loops"             "Frosted Flakes"
[15] "Fruity Pebbles"          "Golden Crisp"
[17] "Golden Grahams"          "Grape Nuts Flakes"
[19] "Honey Graham Ohs"        "Honey-comb"
[21] "Kix"                     "Lucky Charms"
[23] "Multi-Grain Cheerios"    "Nut&Honey Crunch"
[25] "Product 19"              "Rice Chex"
[27] "Rice Krispies"           "Smacks"
[29] "Special K"               "Total Corn Flakes"
[31] "Total Whole Grain"       "Triples"
[33] "Trix"                    "Wheaties"
[35] "Wheaties Honey Gold"
```

Brand names list 3 for brands belong to cluster 3:

```
bn3:
[[1]]
 [1] "100% Bran"
 [2] "All-Bran"
 [3] "All-Bran with Extra Fiber"
 [4] "Basic 4"
 [5] "Bran Chex"
 [6] "Bran Flakes"
 [7] "Clusters"
 [8] "Cracklin' Oat Bran"
 [9] "Crispy Wheat & Raisins"
[10] "Fruit & Fibre: Dates Walnuts and Oats"
[11] "Fruitful Bran"
```

```
[12] "Grape-Nuts"
[13] "Great Grains Pecan"
[14] "Honey Nut Cheerios"
[15] "Just Right Fruit & Nut"
[16] "Life"
[17] "Mueslix Crispy Blend"
[18] "Nutri-Grain Almond-Raisin"
[19] "Oatmeal Raisin Crisp"
[20] "Post Nat. Raisin Bran"
[21] "Quaker Oat Squares"
[22] "Raisin Bran"
[23] "Raisin Nut Bran"
[24] "Total Raisin Bran"
[25] "Wheat Chex"
```

These lists suggest that the clustering might be informative. Because it seems that each cluster is mainly composed by one kind of food.

Q4

(a)

$$\alpha_i = \sum_{j=1}^{m} p_{ij} \ , i = 1, 2, \dots n = p_i$$

$$\beta_j = \sum_{i=1}^{n} p_{ij} \ , j = 1, 2, \dots m = p_j$$

$$MI(A, B) = H(P_A) + H(P_B) - H(P_{AB})$$

$$= -\sum_{i=1}^{n} p_i \times \log(p_i) - \sum_{j=1}^{m} p_j \times \log(p_j) + \sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \times \log(p_{ij})$$

$$= -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \times \log(\alpha_i) - \sum_{j=1}^{m}\sum_{i=1}^{n} p_{ij} \times \log(\beta_j) + \sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \times \log(p_{ij})$$

$$= -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \times (\log(\alpha_i) + \log(\beta_j) - \log(p_{ij}))$$

$$= -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \times \log\left(\frac{\alpha_i \beta_j}{p_{ij}}\right)$$

■

(b)
Since

$$E\ [f(x)]\ \geq f(E(x))$$

$$MI(A,B) = E\left[-log_b\big(g(I,J)\big)\right]$$

$$f(x) = \ -log_b\big(g(I,J)\big)$$

$$MI(A,B) \geq f\big(E(x)\big)$$

$$f\big(E(x)\big) = \ -log_b\big(g(I,J)\big) = 0$$

So we can have

$$MI(A,B) \geq 0$$

■

(c)

$$H(P_A) - H(P_{AB}) = \ -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \times (\log(\alpha_i) - \ \log(p_{ij}))$$

$$= \ -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \times \log\left(\frac{\alpha_i}{p_{ij}}\right)$$

Since we have

$$\alpha_i, \beta_j \ \geq \ p_{ij}$$

So we can get

$$\frac{\alpha_i}{p_{ij}}, \frac{\beta_j}{p_{ij}} \ \geq 1$$

$$\log\left(\frac{\alpha_i}{p_{ij}}\right), \log\left(\frac{\beta_j}{p_{ij}}\right) \ \geq \log(1) = 0$$

We can denote that

$$H(P_A) - H(P_{AB})\ \leq 0$$

$$H(P_B) - H(P_{AB}) \leq 0$$

■

(d)
If

$$\alpha_i \beta_j = p_{ij}$$

Then

$$\log\left(\frac{\alpha_i \beta_j}{p_{ij}}\right) = 0$$

So we can have

$$MI(A, B) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \times \log\left(\frac{\alpha_i \beta_j}{p_{ij}}\right) = 0$$

■

(e)
(i)
If for each row we have

$$\alpha_i = p_{ij}$$

And other $p_{ij} = 0$
Then we can denote that

$$MI(A, B) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \times \log\left(\frac{\alpha_i \beta_j}{p_{ij}}\right)$$

$$= -\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \times \log(\alpha_i)$$

$$= -\sum_{i=1}^{n} \alpha_i \times \log(\alpha_i)$$

$$= H(P_A)$$

So,

$$H(P_A) \leq H(P_B)$$

The upper bound is $H(P_B)$.

(ii)
If for each column we have

$$\beta_j = p_{ij}$$

And other $p_{ij} = 0$
Then we can denote that

$$MI(A,B) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \times \log\left(\frac{\alpha_i \beta_j}{p_{ij}}\right)$$

$$= -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \times \log(\beta_j)$$

$$= -\sum_{j=1}^{m} \beta_j \times \log(\beta_j)$$

$$= H(P_B)$$

So,

$$H(P_B) \leq H(P_A)$$

The upper bound is $H(P_A)$.

(f)

Compare with using $\{min\{H(P_A), H(P_B)\}$, in which we get I(Q,R) = 1, if we use $\{max\{H(P_A), H(P_B)\}$, we will have I(Q,R) $<$ 1.

We know that Q is different from R, so using $\{min\{H(P_A), H(P_B)\}$ and have I(Q,R) = 1 is not the best solution.

Thus, $\{max\{H(P_A), H(P_B)\}$ is a better method compared with $\{min\{H(P_A), H(P_B)\}$.