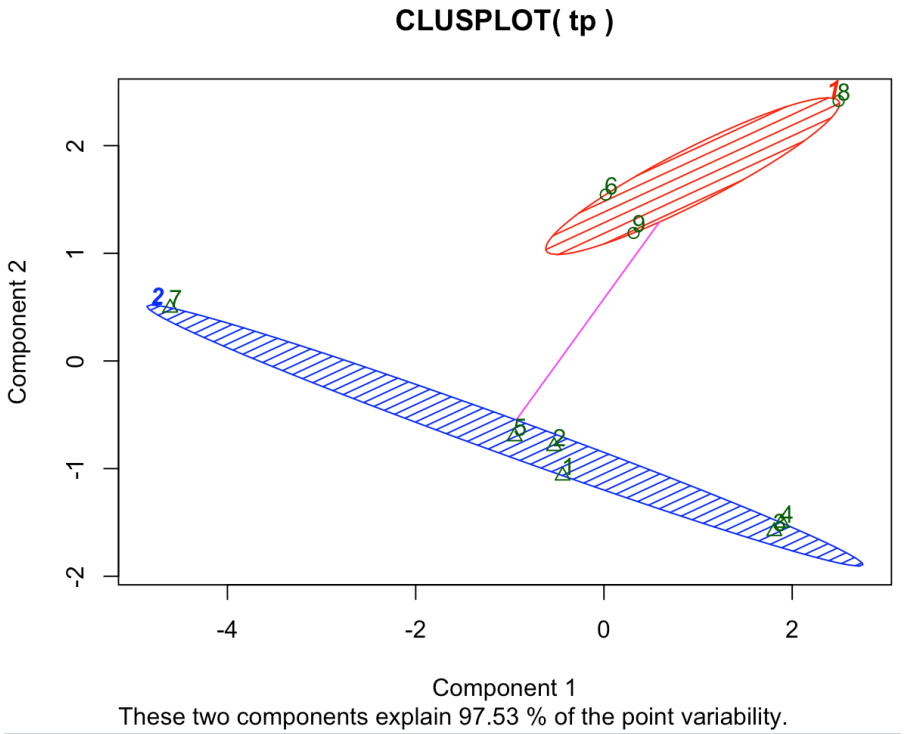
Hw12

20150056 국윤범

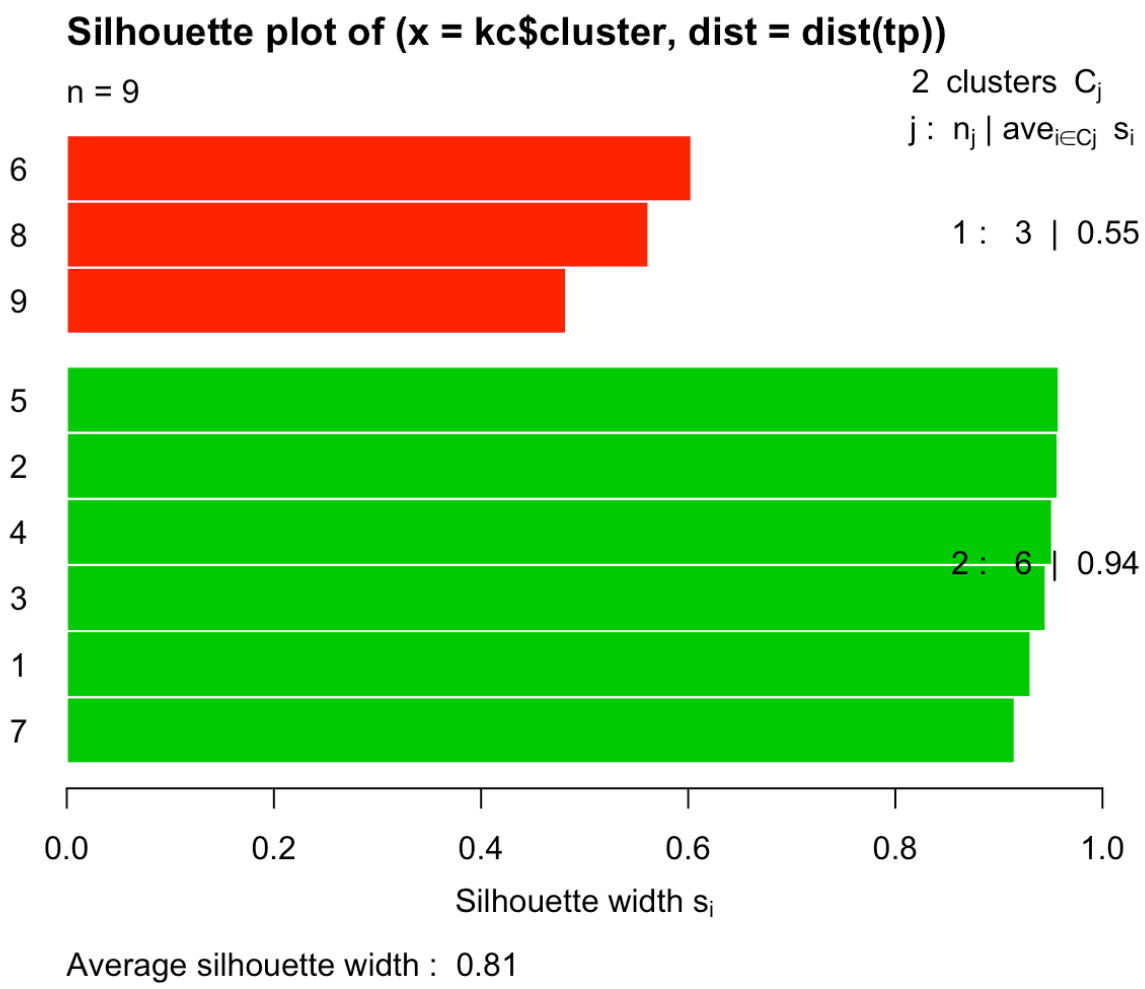
Problem 1

(a) As the professor mentioned in the KLMS, I assume that the number of clusters is 2.

```
> (kc = kmeans(tp, centers=2, nstart=3))
K-means clustering with 2 clusters of sizes 3, 6

Cluster means:
        nA    meanA      sdA       nB    meanB      sdB
1 861.3333 3.806667 4.410000 843.6667 4.046667 4.660000
2 168.8333 5.271667 4.458333 160.1667 5.368333 4.803333

Clustering vector:
1 2 3 4 5 6 7 8 9
2 2 2 2 2 1 2 1 1

Within cluster sum of squares by cluster:
[1] 246870.421   9793.245
 (between_SS / total_SS =  88.1 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betwe
enss"    "size"
[8] "iter"         "ifault"
```



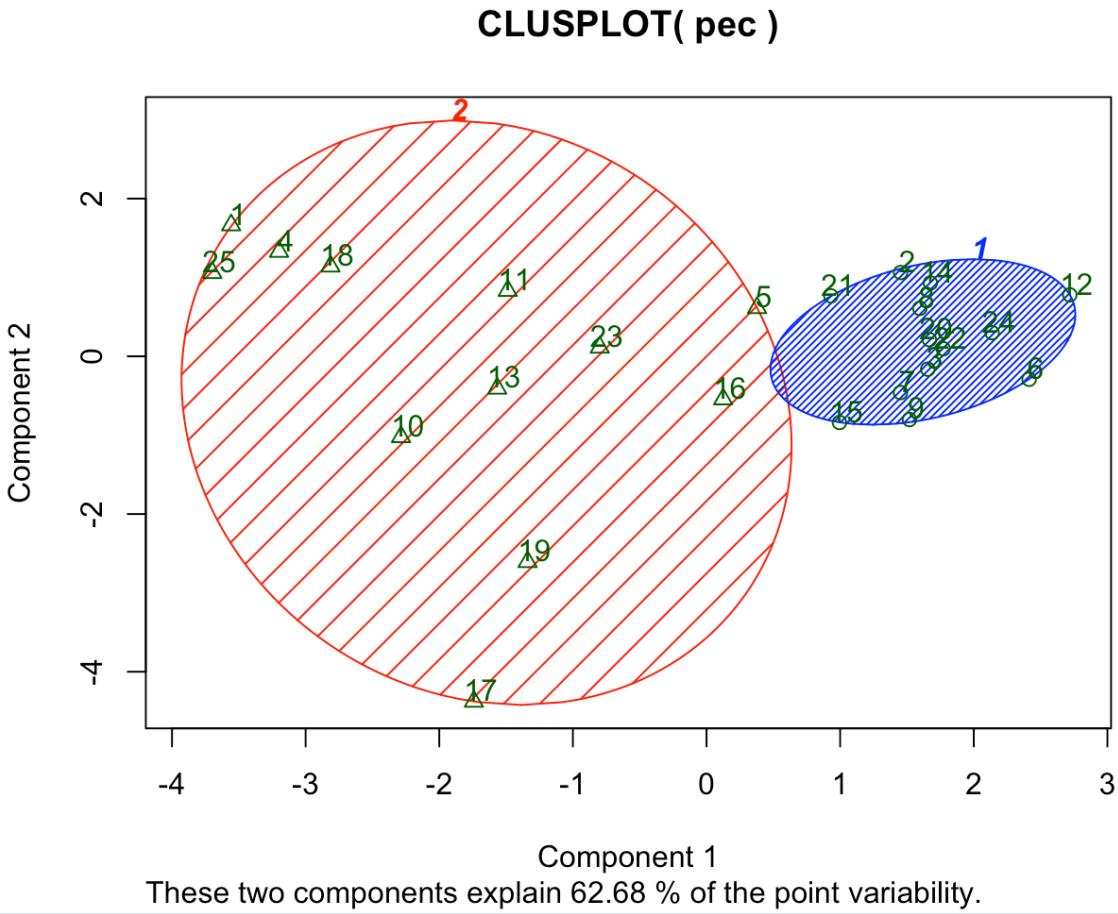
(b)


Silhouette plot of (x = kc$cluster, dist = dist(tp))
n = 9
2 clusters C_j
j : n_j | $ave_{i∈Cj}$ s_i
1 : 3 | 0.55
2 : 6 | 0.94
6
8
9
5
2
4
3
1
7
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.81

Problem 2

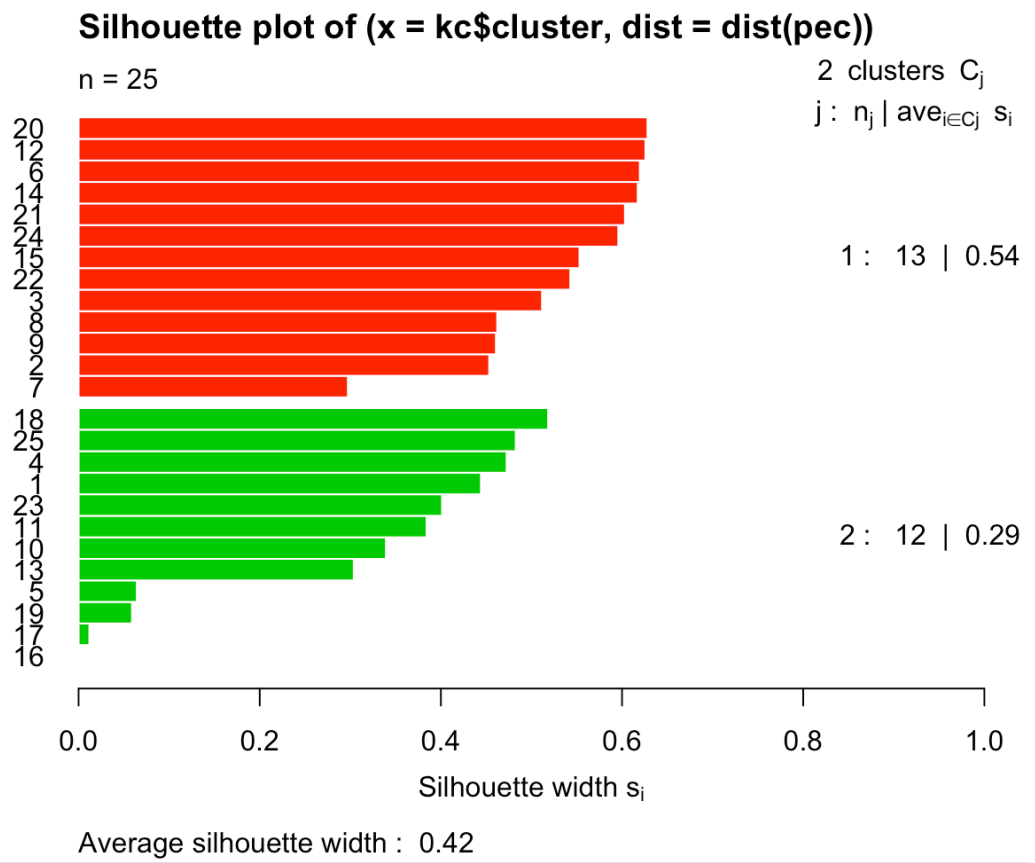
(a)

```
> (kc = kmeans(pec, centers=2, nstart=3))
K-means clustering with 2 clusters of sizes 13, 12

Cluster means:
    RedMeat WhiteMeat     Eggs     Milk     Fish  Cereals   Starch     Nuts
1 11.807692  9.607692 3.707692 22.08462 5.023077 24.06923 4.761538 1.692308
2  7.683333  6.041667 2.100000 11.72500 3.483333 41.10833 3.750000 4.566667
  FruitVeg
1    3.500
2    4.825

Clustering vector:
 [1] 2 1 1 2 2 1 1 1 1 2 2 1 2 1 1 2 2 2 2 1 1 1 2 1 2

Within cluster sum of squares by cluster:
[1]  808.6554 1668.0933
 (between_SS / total_SS =  52.8 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```



These two components explain 62.68 % of the point variability.

(b)

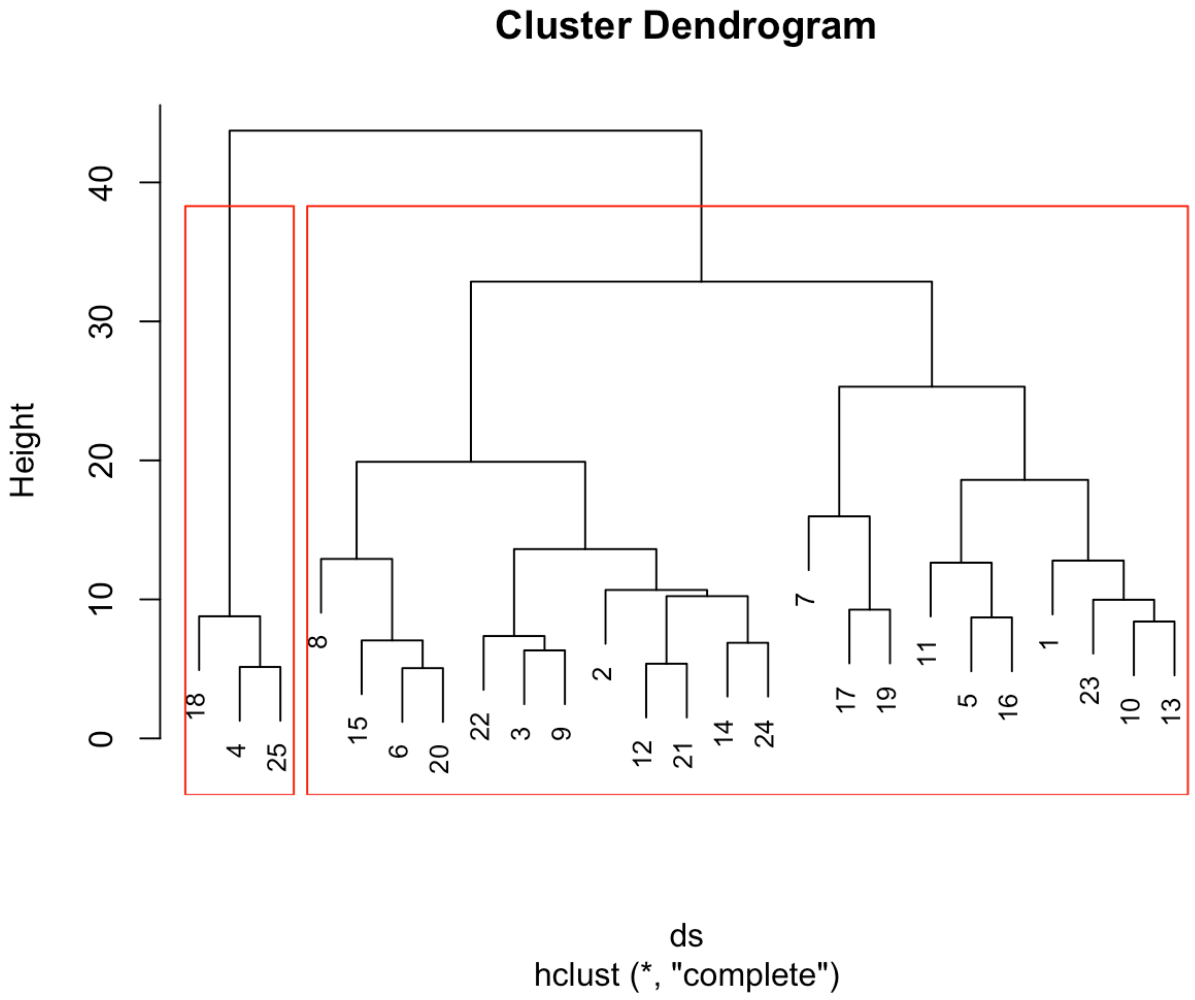


(C)

```
> cl1 = subset(pec, kc$cluster=="1")
> cl2 = subset(pec, kc$cluster=="2")
> t.test(cl1[,1],cl2[,1])$p.value #
[1] 0.0007830653
> t.test(cl1[,2],cl2[,2])$p.value #
[1] 0.01310718
> t.test(cl1[,3],cl2[,3])$p.value #
[1] 5.03499e-05
> t.test(cl1[,4],cl2[,4])$p.value #
[1] 1.930472e-05
> t.test(cl1[,5],cl2[,5])$p.value
[1] 0.2768672
> t.test(cl1[,6],cl2[,6])$p.value #
[1] 4.407262e-05
> t.test(cl1[,7],cl2[,7])$p.value
[1] 0.1388205
> t.test(cl1[,8],cl2[,8])$p.value #
[1] 0.0001519818
> t.test(cl1[,9],cl2[,9])$p.value
[1] 0.07255526
```

Hence, except for “Fish”, “Starch”, and “FruitVeg”, all categories between two clusters show significant difference.

(d)


Cluster Dendrogram
Height
0
10
20
30
40
18
4
25
8
15
6
20
22
3
9
2
12
21
14
24
7
17
19
11
5
16
1
23
10
13
ds
hclust (*, "complete")