Multivariate Data Analysis Assignment #2

2014150137, Statistics

박 정진

**8.2**

(a)

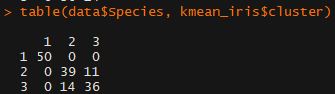
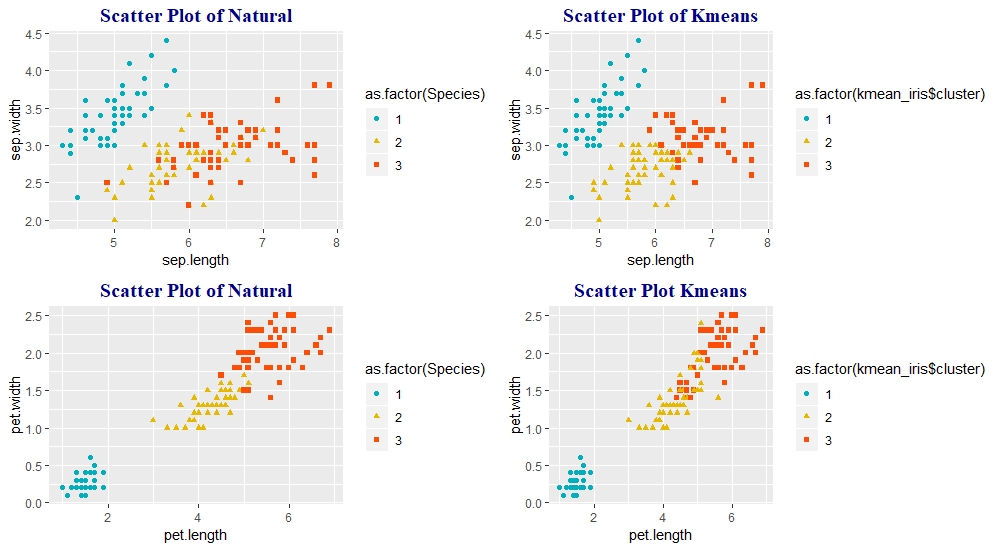
Species can be the natural modality of this data set.



You can see clear differences among Species.

Which means, there are 3 natural clusters.

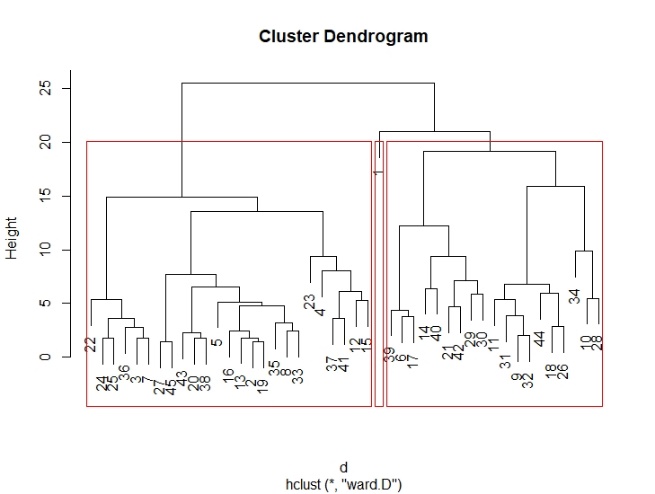
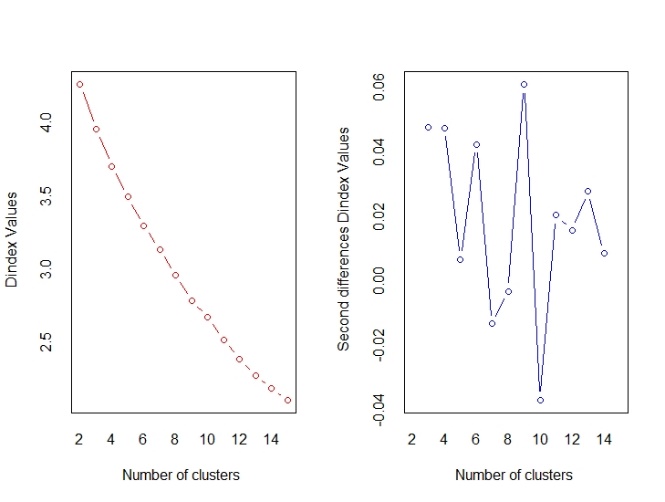
(b)-Using K-means Clustering



By Comparing Scatter plots of Natural Modality and Kmeans Clustering, you can see the similarity between two method.

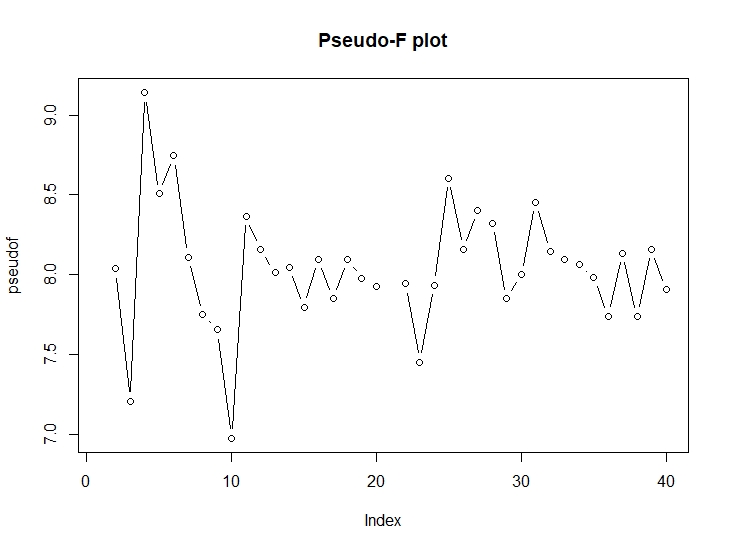
Also, Cluster Table shows the same result.

8.4

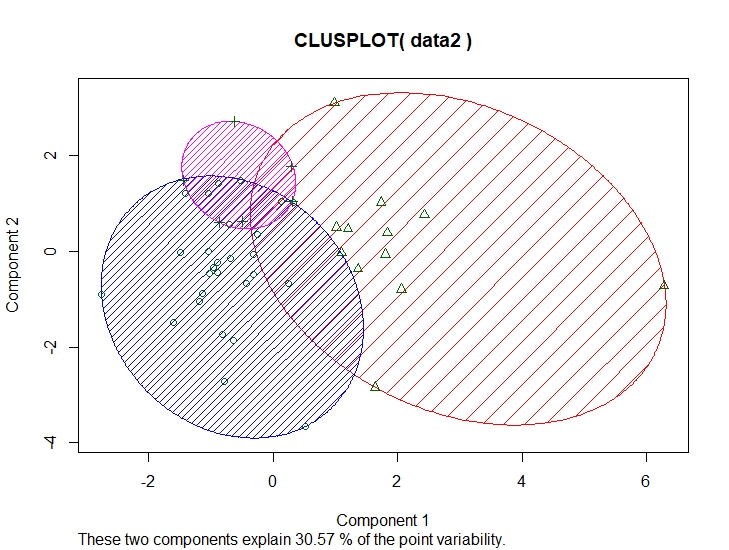


By Using Nbclust package, we could find out 3 clusters would be the best number.

Cluster Dendrogram of hierarchical clustering shows that it is appropriate.

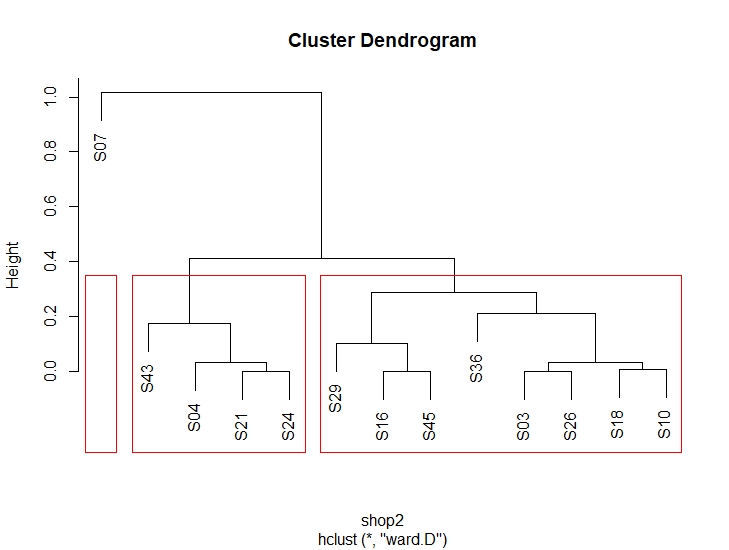


By calculating Pseudo-F statistics, we also could see that the best number of clusters is 3.



The plot above shows the data is pretty well clustered with 3 clusters.

**8.5i**



The Cluster Dendrogram clearly shows that the best number of clusters would be 3.

The clusters consist of (S07), (S43, S04, S21, S24), (S29 S16, S45, S36, S03, S26, S18, S10)

If the stores are in the same cluster, then we can say that it means they are not competitive.

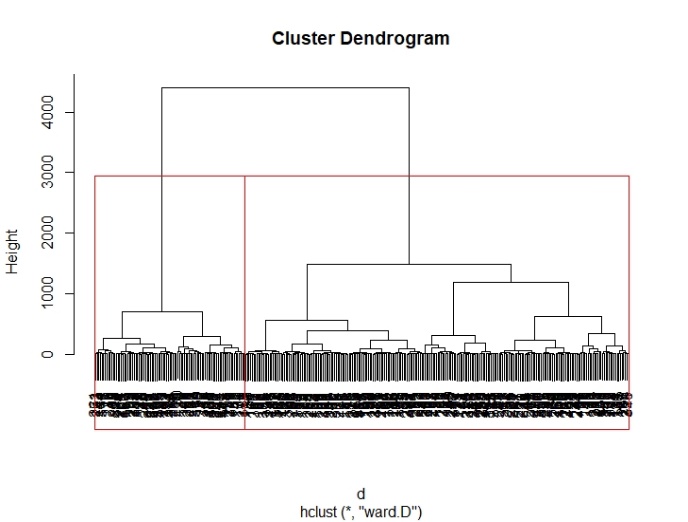
Else, they are competitive for each other.

Because the measure we are using is , and if the ith and jth store choice shares are similar, the distance between two stores would be bigger.

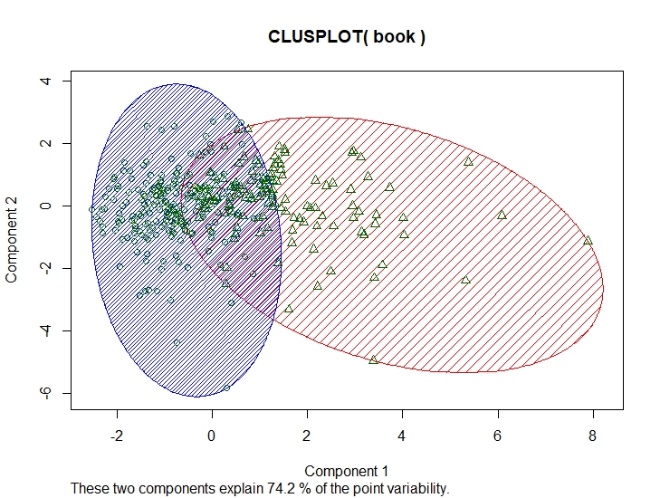
So as an example, S07 makes up its cluster by its own, which means it sells special products distinctive from other stores.

**8.8**

(a)

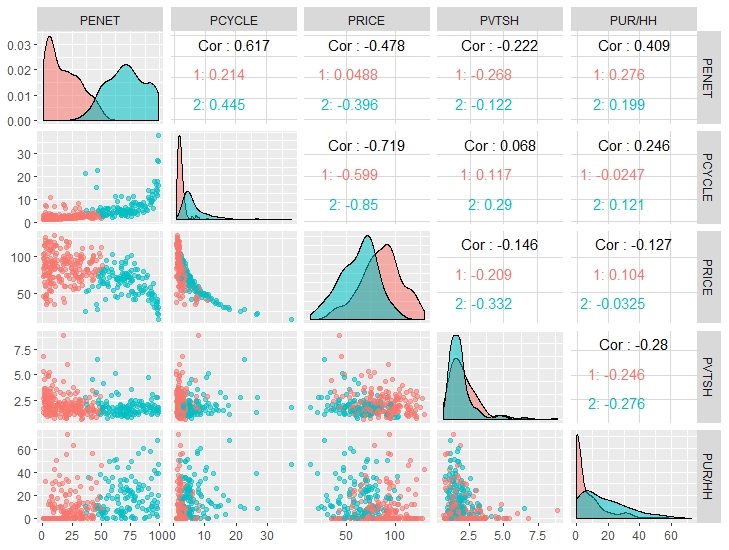


We can find out that the best number of clusters is 2 from Cluster Dendrogram and pseudo-F statistics.



Clusplot also shows that K-means clustering with K=2 is well done.

(b)



By paring plots, Cluster1 shows lower Penent, Pcycle and Pur/HH values, and Cluster2 has higher values in contrast.

(c)

Looking at the scatter plots and its distribution, Price and Pvtsh of two clusters show similarities and they are not clearly distinguished by clusters.

And it means that the argument of Fader and Lodish is true.

**8.10**

(a)

I used the measure studied by Rao and Sabavala which can be demonstrated as,

.

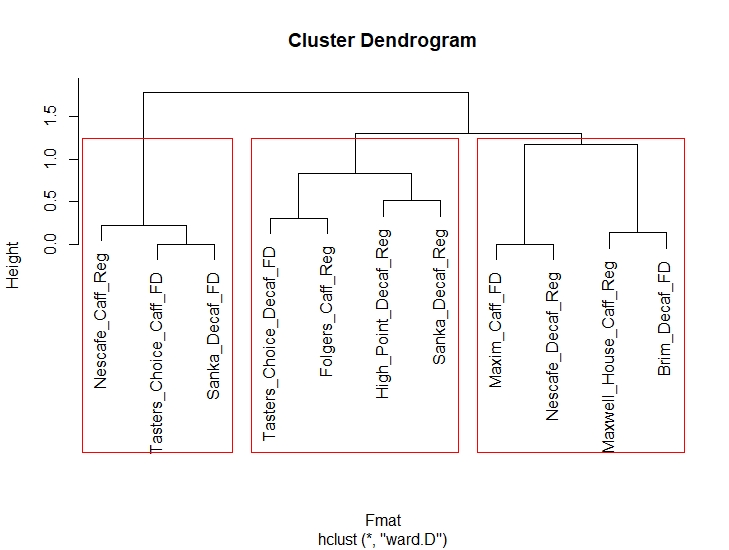
= the number of consumer who changed brand from I to j

= the number of consumer who first purchased brand i

= the number of consumer who purchased brand j in the second occasion

n = the number of sampled population

(b)



We used hierarchical clustering for this method and we could clearly find out that 3 clusters would be good for this data, there are some points we can say.

First, instant coffees from same brands are not competitive for each other. As you can see, Nescafe, Sanka, Tasters coffees are not in the same cluster.

Second, the form of coffee doesn’t have strong effect on competitiveness. You can see that clear in the dendrogram.

**10.1**

(a) 28 – 17 = 11

(b) 0.88\*0.82\*(0.43 - 0.29\*0.26) = 0.2558794

(c)

> gam

[,1] [,2]

[1,] 0.43 -0.29

> phi

[,1] [,2]

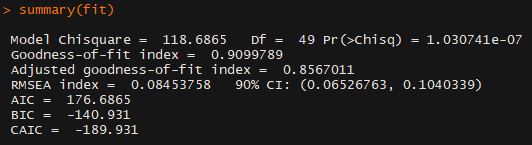
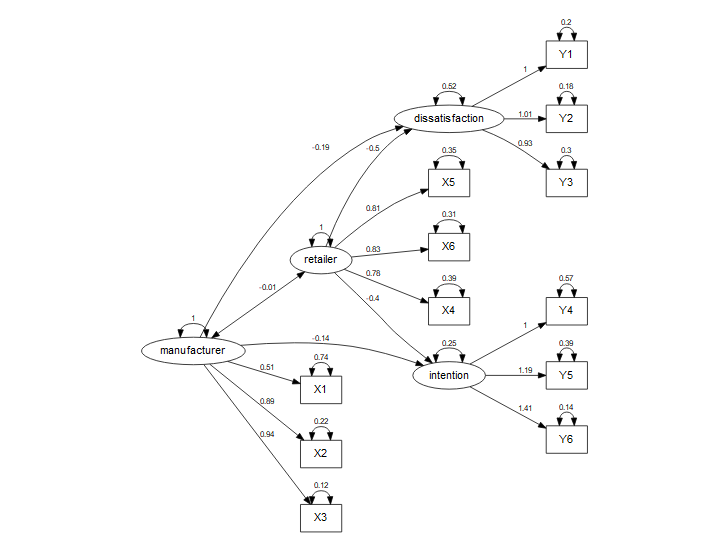
[1,] 1.00 0.26

[2,] 0.26 1.00

reliability = gam%\*%phi%\*%t(gam)

reliability = 0.204156

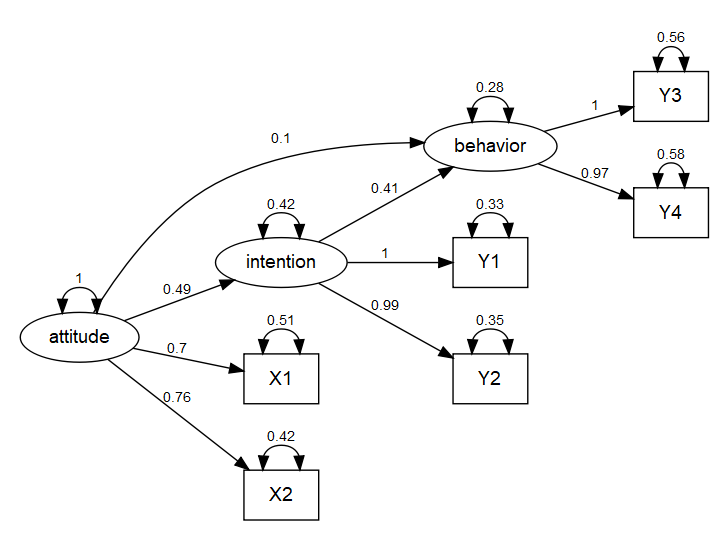
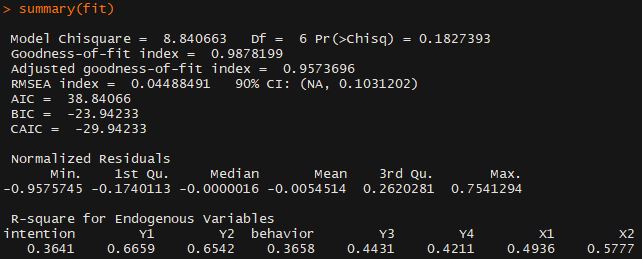
**10.3**



Estimated parameters can be shown as above.

But the model chisquare statistics is really big and p-value is almost 0, which means the model is not well-defined so we need additional parameters.

**10.5**



(a)

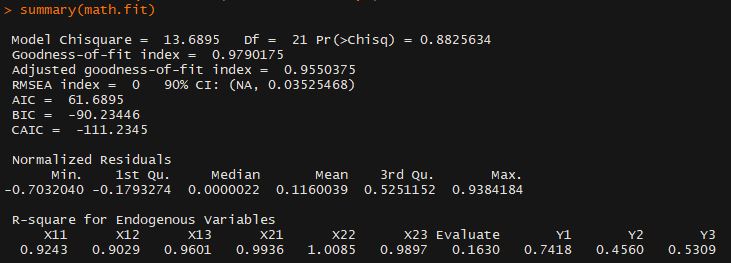
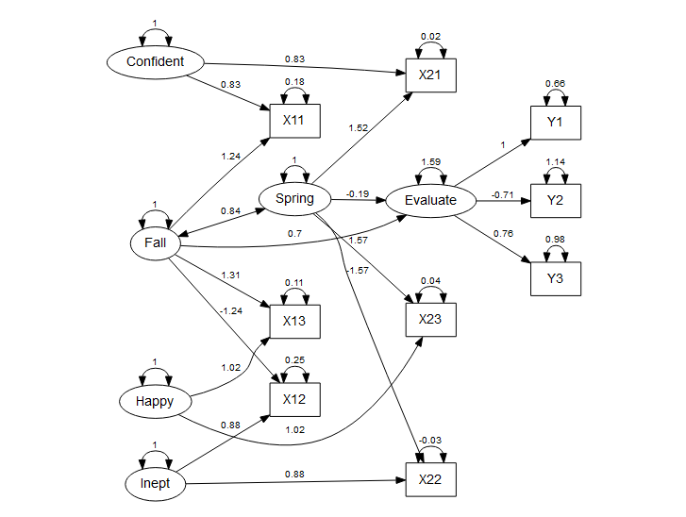
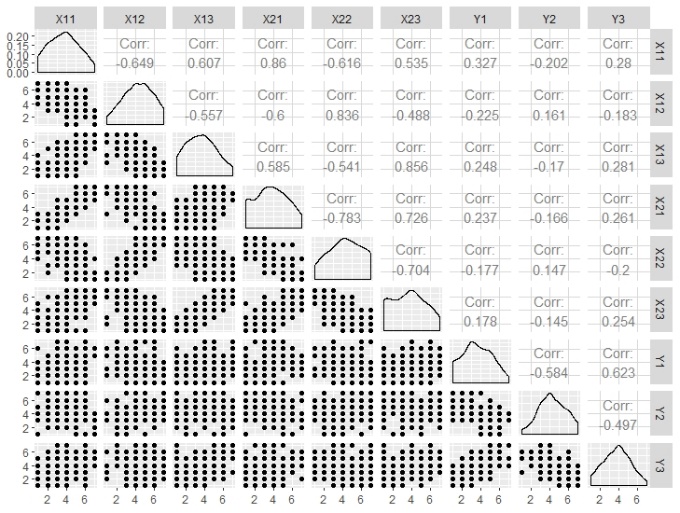
Our goodness-of-fit is 0.99 and P-value is 0.18, which shows the model is well-defined.

(b)

The effect of attitude on behavior is both direct and indirect.

But, as you can see above, direct effect is small, but indirect effect through intention is much higher than direct effect.

**10.7**



(a)

We can model SEM, and can see the difference between the effects of attitudes on evaluation in time 1 and time 2.

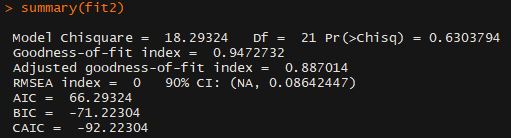
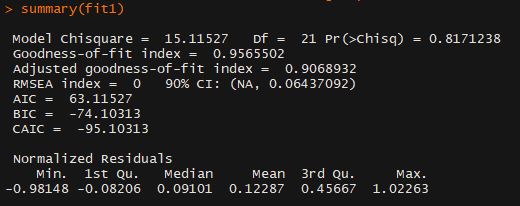
First, the p-value of our model is 0.88 and goodness of fit is over 0.95, which means the model is well-defined.

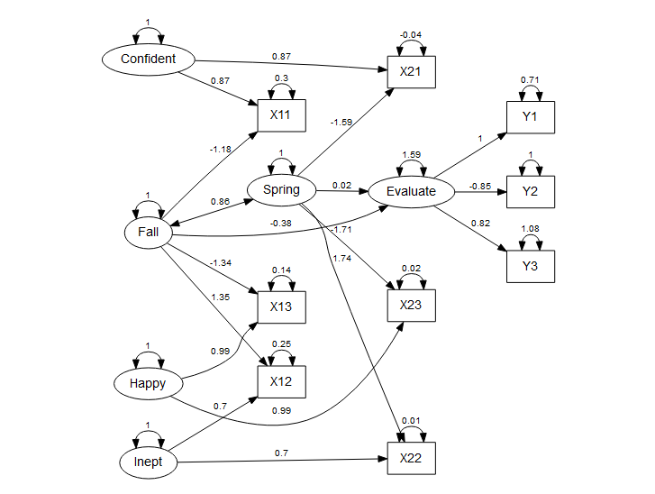
You can that the effect in time 1(Spring) is -0.19 and time 2(Fall) is 0.7. So, our conclusion is that a new quantitative methods course has impact on evaluation.

(b)



Aptitude Score histogram shows that we can divide students into 2 groups to control the effect of aptitude test score.





Left sided tables and graphs are for higher scored students, and Right sided are for lower scored students. Two models are well-defined for divided data. We can find out by their p-values and goodness-of-fit.

The effect of higher scored students is almost same as the original complete data. But, the effect of lower scored students is different from the original. Time1 effect is 0.02 and Time2 effect is -0.3. It means that the aptitude test score has great impact on evaluation. So we have to reconsider the hypothesis that there is effect of attitude toward math on student evaluations.