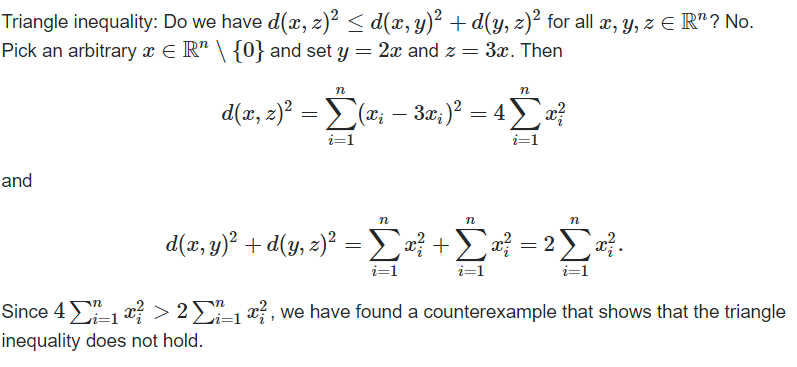
HW2-Clustering

0859605 陳冠景

1. Since Euclidean distance & Square Euclidean distance(we assume) are metric, then a metric satisfies 3 properties, which is positive reflexivity, symmetry, triangular inequality.

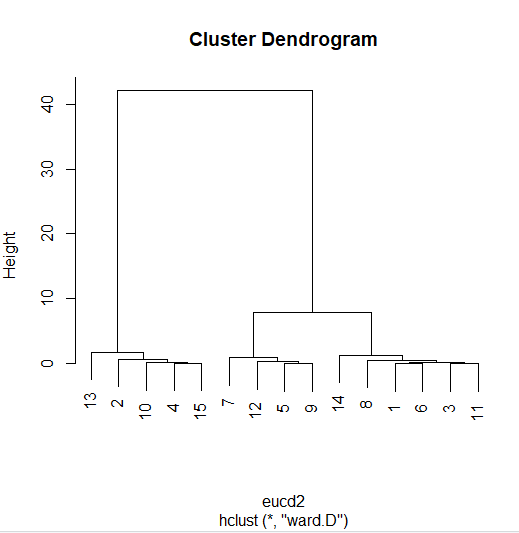
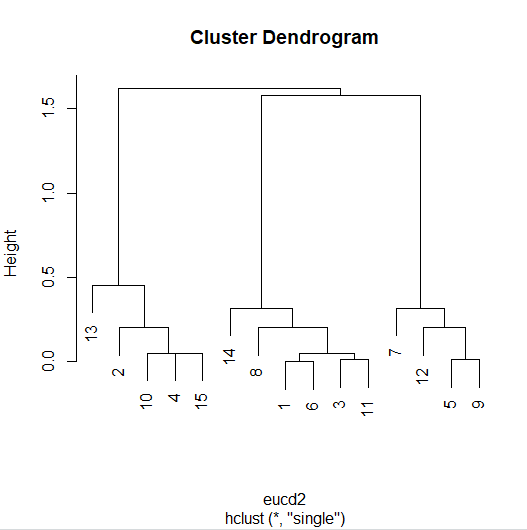


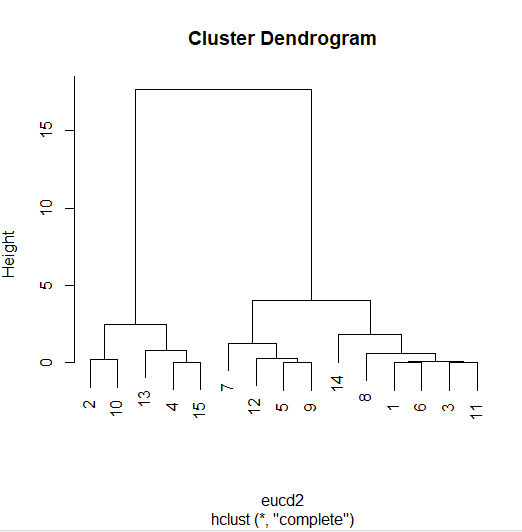
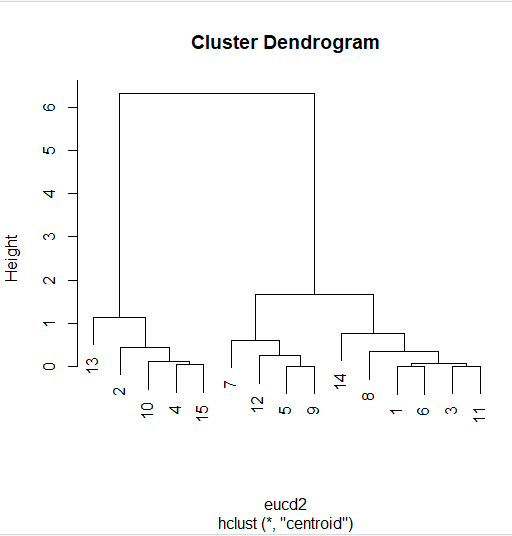
As a result, Square Euclidean distance doesn’t meet the triangle inequality, so it is not metric. When calculating , Square Euclidean distance is different from Euclidean distance.

If you use Square Euclidean distance replace Euclidean distance when doing hierarchical clustering , the output is likely to change.

(cite from: <https://reurl.cc/odYlN3>)

1. Draw diagram

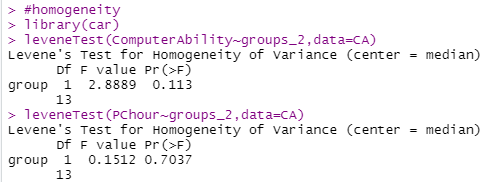
 

C.

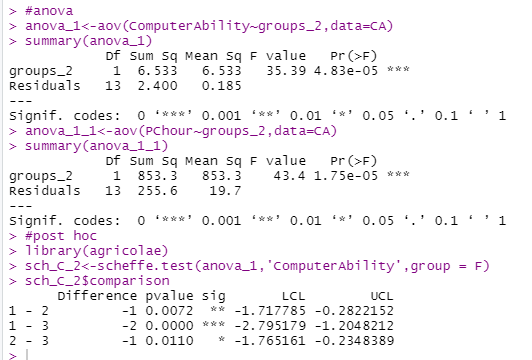
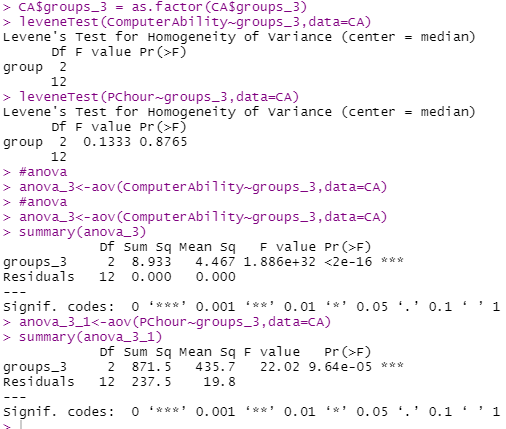
Conduct ANOVA, check significant difference in clusters to determine number of clusters.

With “ward.D”

#groups2

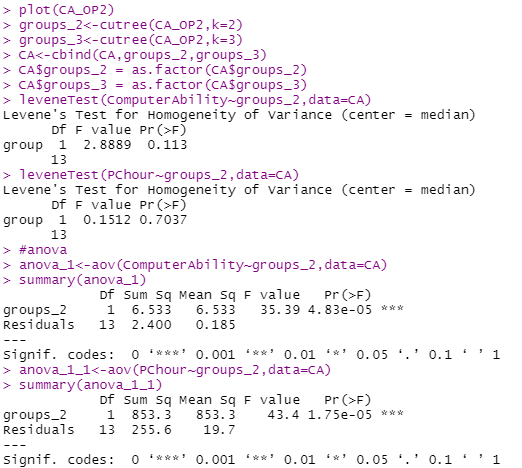
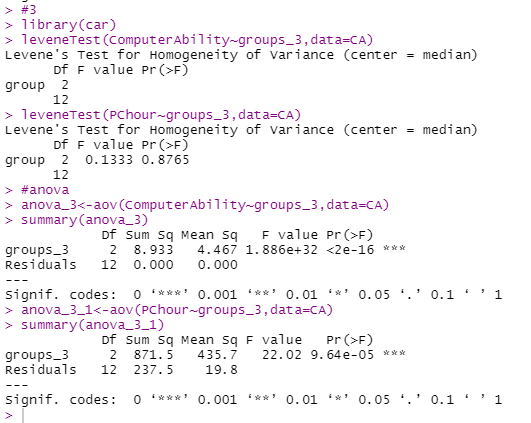


The assumption of anova isn’t violated.

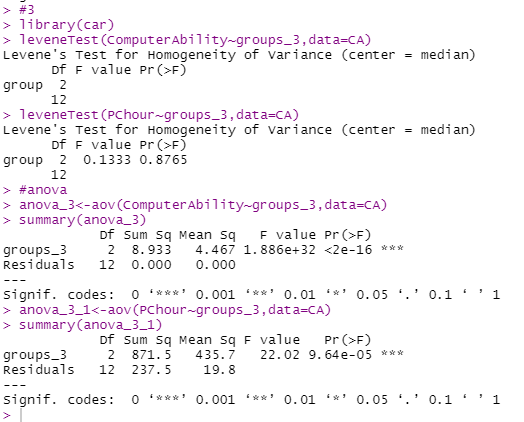
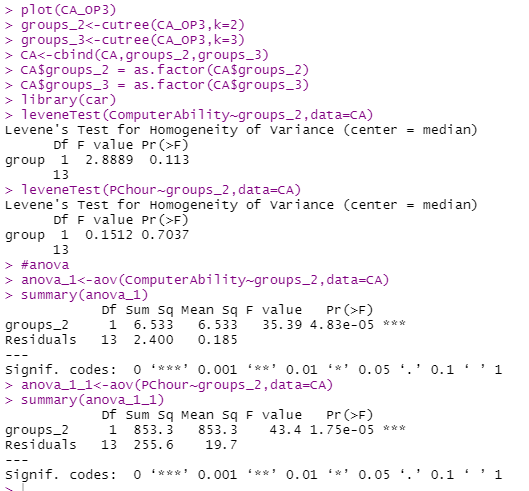
The result of cluster2 , there are significant difference about ComputerAbility and PChour in 2 groups ,as well as groups3.(right picture above)

With “nearest” method

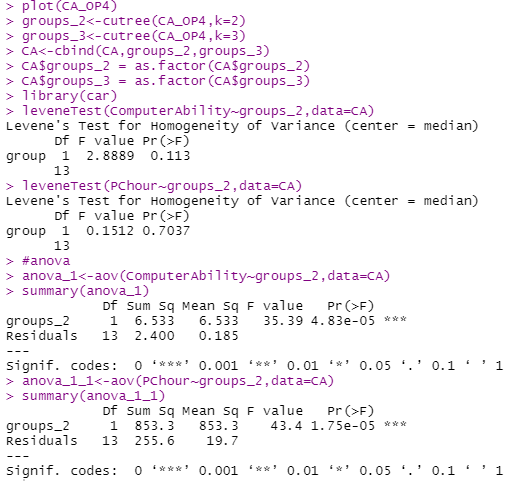
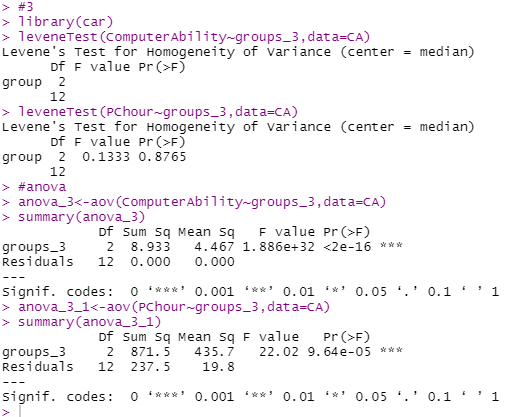
The result of cluster2 with the nearest method , there are significant difference about ComputerAbility and PChour in 2 groups, as well as groups3. (right picture above )

With ‘furthest’ method



The result of cluster2 with the furthest method , there are significant difference about ComputerAbility and PChour in 2 groups, as well as groups3. (right picture above )

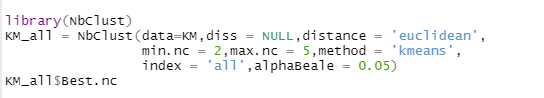
With ‘centroid’ method

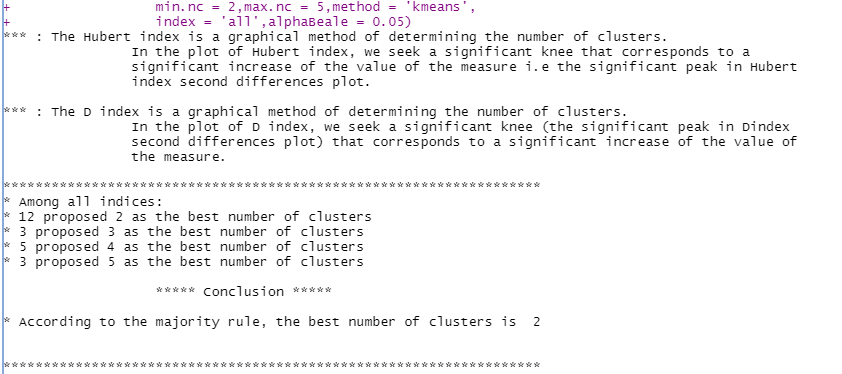
 

The result of cluster2 with the centriod method , there are significant difference about ComputerAbility and PChour in 2 groups, as well as groups3. (right picture above )

K means

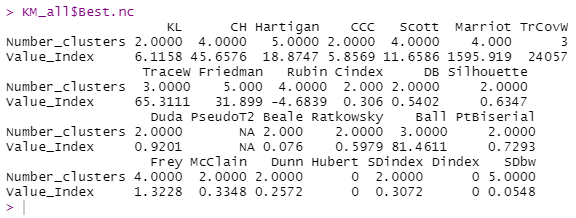
D. use NbClust() to decide the group number

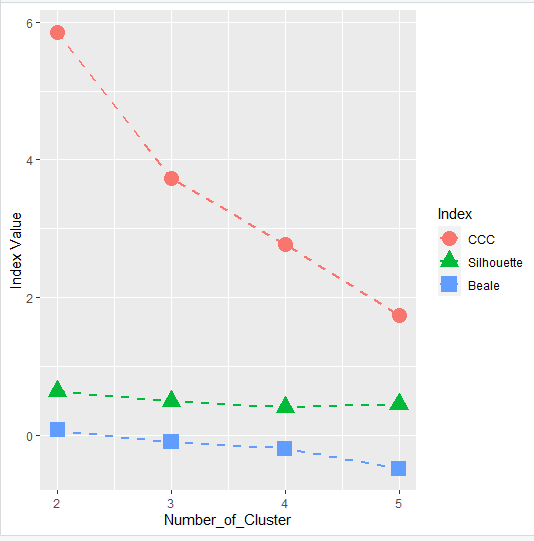




The result of the best number of clusters is 2.

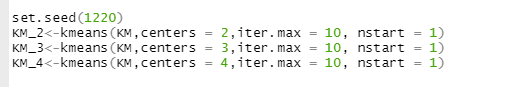
We can visualize the several common indices with cluster 2~5.





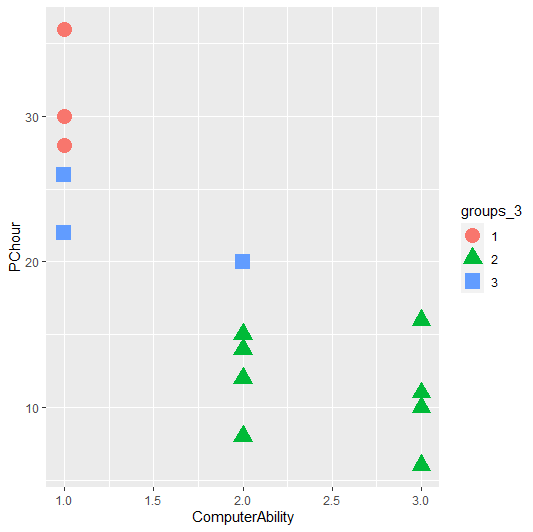
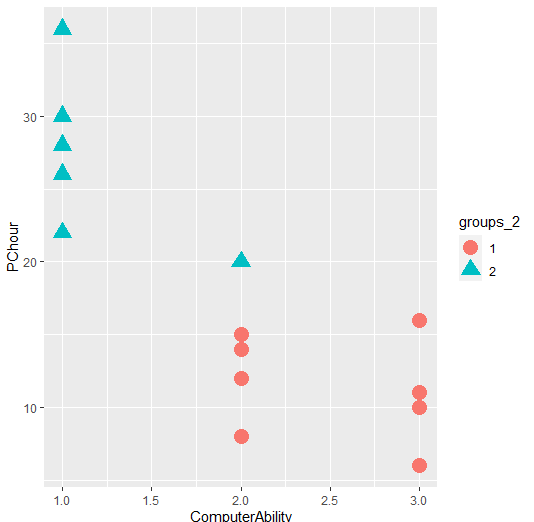
Considering the peak value of CCC index is the best, I think maybe 2 groups clustering is a good choice.

Then, before the k means analysis, set initial seed first.

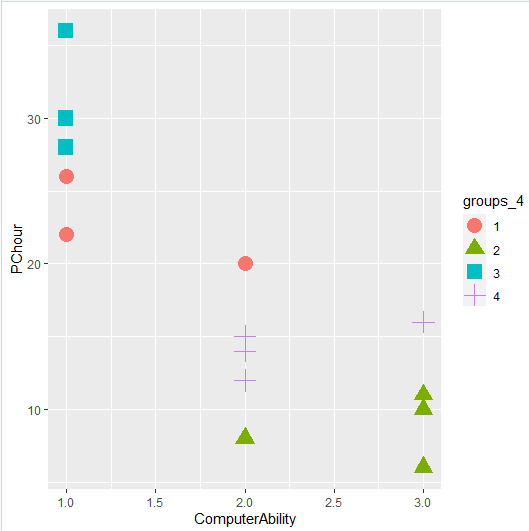


Select k=2,3,4 as the number of clusters.

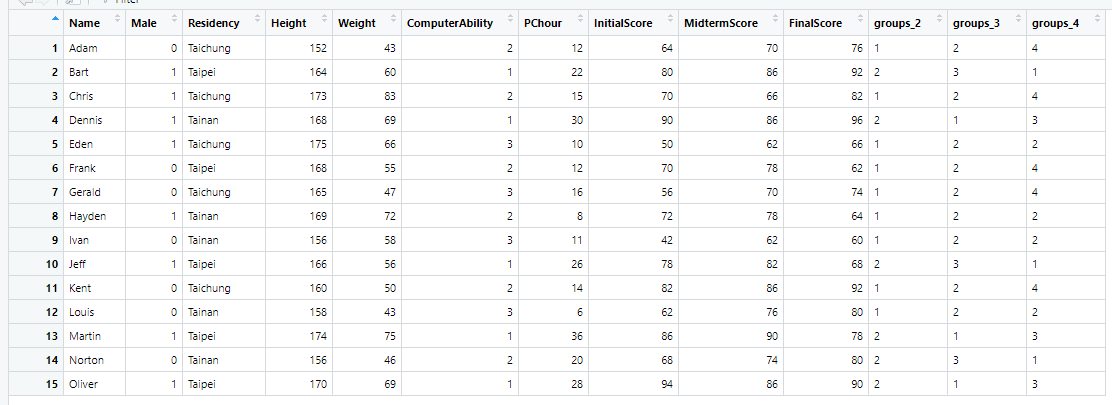
Visualize the result.



K =2 K = 3



K = 4



The result of groups 2~4.

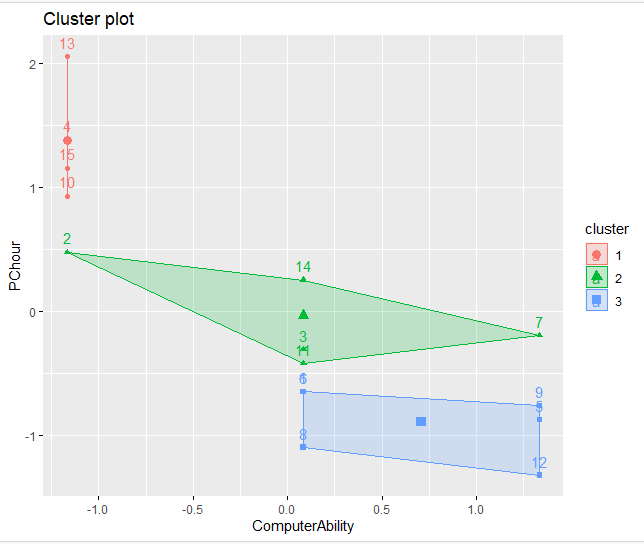
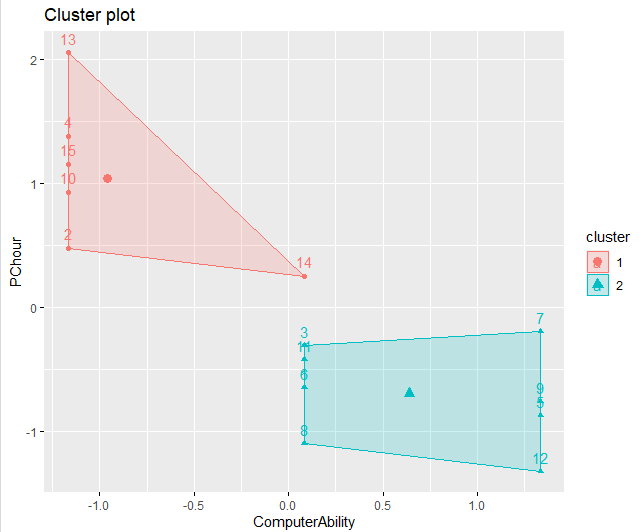
E.

Compare within cluster distance , between cluster distance in cluster 2~4

Use str() to get the distance information about 3 kinds of cluster .

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. dis. | withinss | tot.withinss | betweenss | size |
| Cluster 2 | 166.8 86.4 | 253 | 865 | 6 ,9 |
| Cluster 3 | 19.3 34.7 86.4 | 140 | 977 | 3,3,9 |
| Cluster 4 | 30.33 2.67 19.33 34.67 | 87 | 1031 | 6,3,3,3 |

Smaller within distance and larger between distance are better indices for determine better number of clusters. According to these 2 indices , cluster 3 may be a good choice.



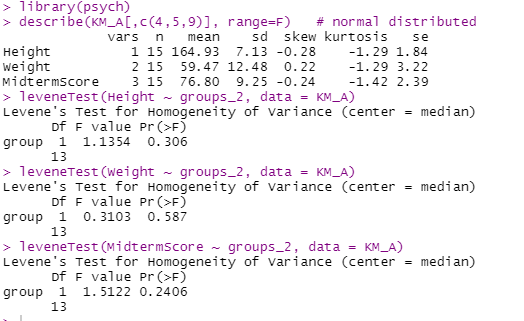
By seeing scatter plot, I think 2 group cluster the data more evenly. In 3 groups, the number of data in each group is varied.

Due to CCC index and scatter plot, we determine 2 groups is the best choice.

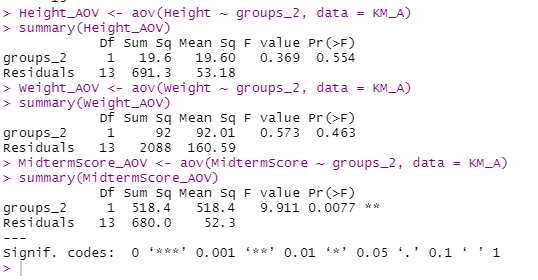
F.

Conduct anova test to find out if there is any difference between groups in Height, Weight, MidtermScore.

# check the distribution assumption, homogeneity of variance

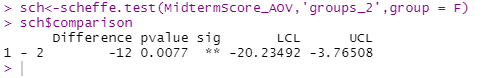


# ANOVA



Only MidtermScore has significant difference in 2 groups(*F*(1,13)=9.91,*p*=.007).

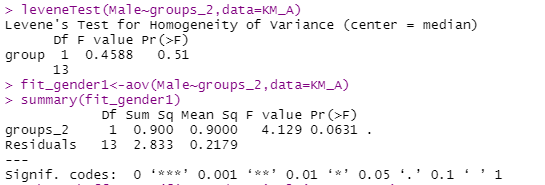
# scheffe test as post hoc



In group 1 , MidtermScore is significantly higher than group 2( ΔM1-2=-12 , *p*=.007).

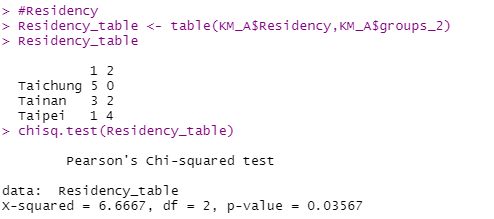
#Male(gender)

Conduct ANOVA test to find out if there is any difference in gender between 2 groups.



There is no significant difference in gender between 2 groups (*F*(1,13)=4.129,*p*=.06).

#residency



There is difference between students’ residency between 2 groups (*X*(2)=6.67,*p*=.03).

Bonus

I would like to try computer ability and student initial score as clustering variables ,since theses 2 variables could be seem as their initial ability profile before teaching .The clustering result can represent their different learning ability between groups. If there are no significant difference in their final score between groups , we may summarize teacher’s teaching is suitable to students with different abilities.