(a) Describe the dataset in terms of rows, columns, types of data and any outliers and missing

data . . . the usual.

This dataset has 64535 observations and 99 columns, which is adequate for cluster analysis.

Half of the data are numeric, and each percent type data also combines with category type data.

However, the dataset has a lot of Incorrect measurements that will affect the analysis, and we have to remove them.

(b) Clean the data - describe what you did to clean the data

For cleaning the data, we have to detect whether the value is Incorrectly measured or not.

for (i in names(test)) {

if (is.numeric(test[[i]])) {

test[[i]][test[[i]] < 0] <- NaN

}

}

Using this for loop to detect if the value is negative or not, if so, we will change it to NaN value, then use complete.case() to remove the whole row.

test\_r\_m=test[complete.cases(test),]

test\_r\_m

check =nrow(test) - nrow(test\_r\_m)

check

(c) Create a set number of groups of “housing” observations.

1. Determine which variables will you then cluster on. Remember we are focused

mainly on the energy cost (UTILITY VARIABLE).

test = data[,-c(1,3,4,13,16,17,23,26,34,46,48,50,52,54,56,58,60,62,64,66,68,70,72,74,75:99)]

we removed columns 1,3,4,13,16,17,23,26,34,46,48,50,52,54,56,58,60,62,64,66,68,70,72,74,75:99

column “control” is meanless for this analysis, so we removed it.

After data cleaning, column “STATUS”, “VACANCY”, “TENURE” ,”ASSIST” have too many same data, so we decide to removed it.

Columns 46,48,50,52,54,56,58,60,62,64,66,68,70,72,74 and 75 to 99 are same data with different data type with others, therefore, we removed them.

(ii) Conduct cluster analyses using two agglomerative methods and a k-means cluster.

How many clusters do you settle on using each method. Why? Provide the necessary charts.

In this case, we use “centroid” and “ward distance” method for Hierarchical Cluster

to analysis.

We use agglomeration\_coefficients function to measuring how many clusters we should put in this model.

agglomeration\_coefficients=function(x,data){

ac=0

n=length(x)

for (i in 1:n) {

pac=0

m=length(x[[i]])

x0=data[x[[i]],]

x0\_mean=colMeans(x0)

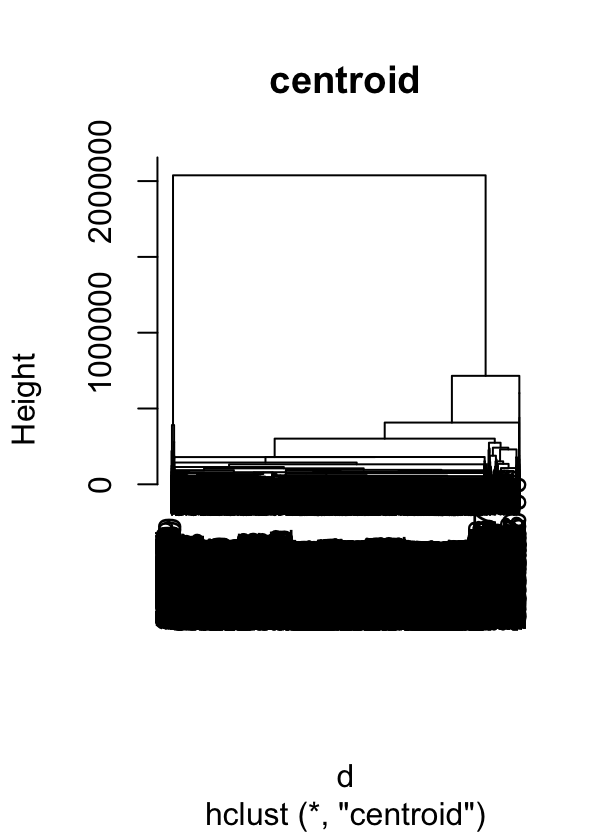
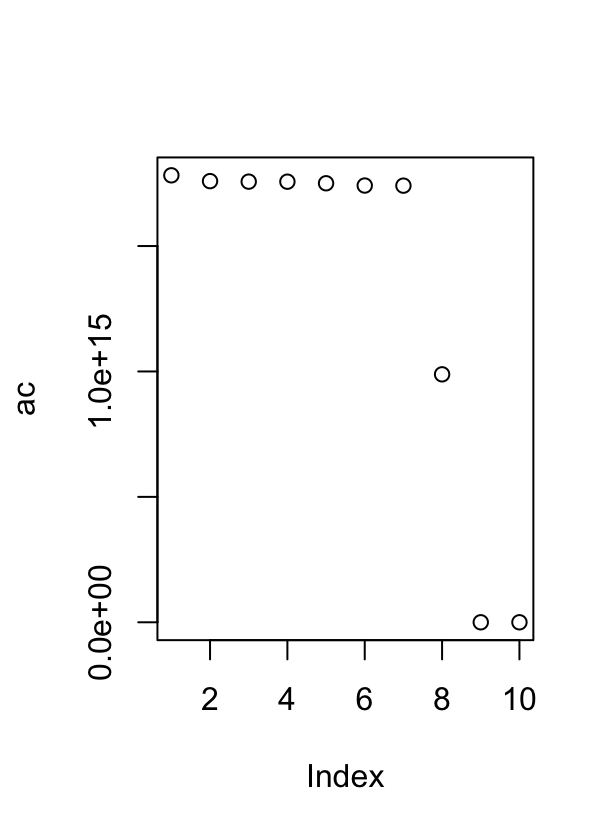
for (j in 1:m) {

pac=pac+sum((x0[j,]-x0\_mean)^2)}

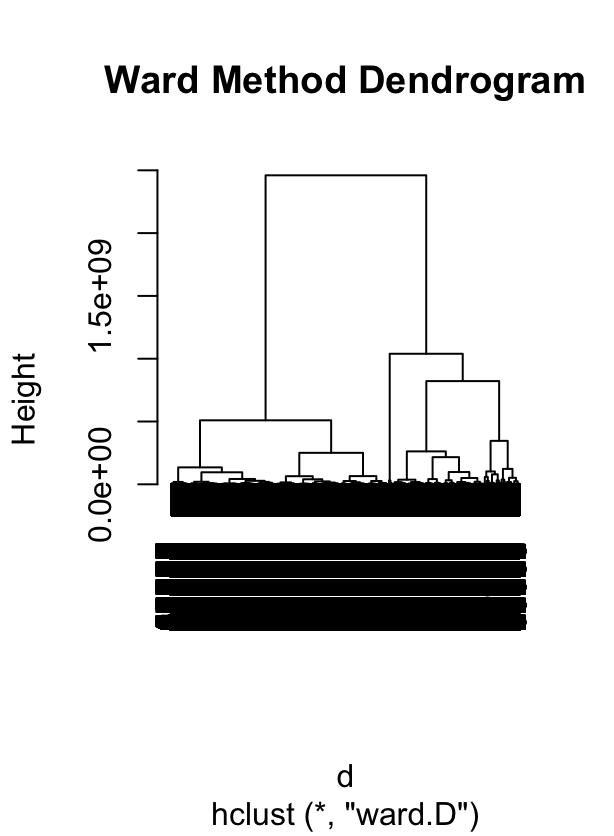
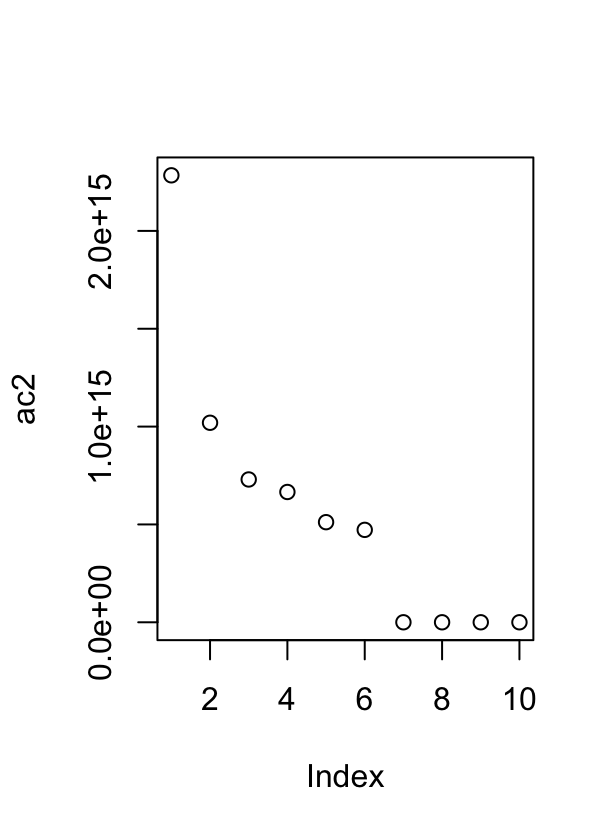
ac=ac+pac}

return(ac)}

The agglomeration coefficient is a metric used to evaluate the clustering results, which reflects the degree of closeness between each cluster in the clustering process, A smaller clustering coefficient indicates a better clustering result because it indicates a higher degree of intra-cluster similarity of clusters and a higher degree of inter-cluster variation, indicating a better classification. So it is better to find the K that has a smallest agglomeration coefficient in this case.

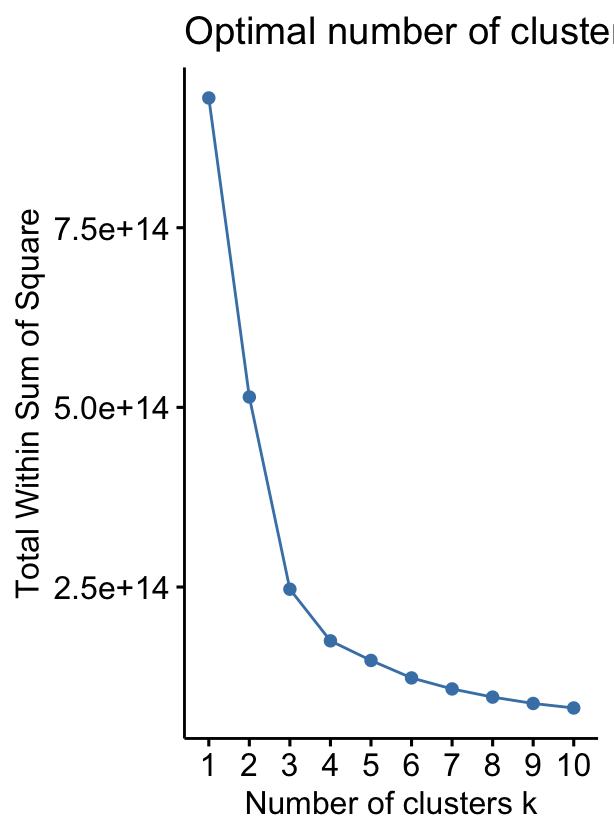
 

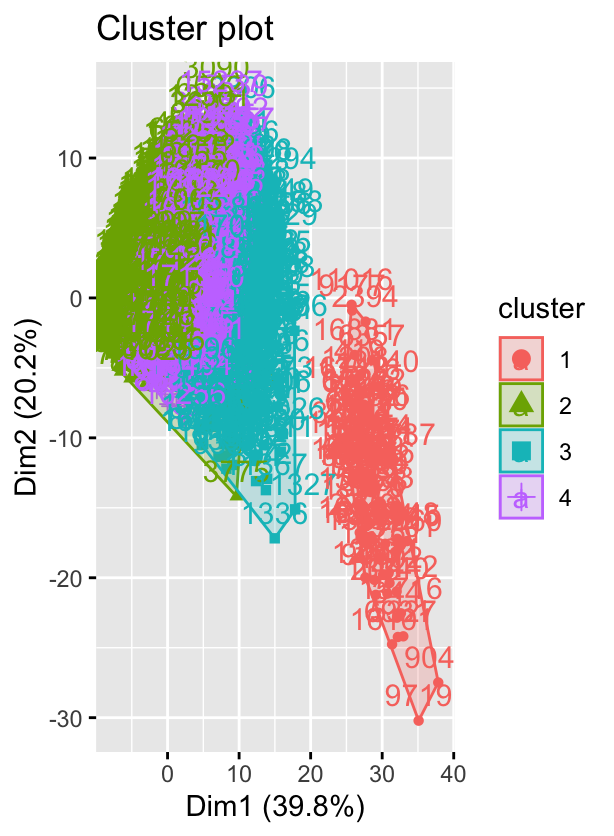
For “Centroid” method the agglomeration coefficient is 9, so we decide divided into ten categories.

For “Ward distance” method the agglomeration coefficient is 7, so we decide divided into seven categories.

For k-means, we use fviz\_nbclust function to analysis how many clusters we should put in this model. Using elbow method to pick the leveling point in Within-Cluster-Sum-of-Squares can get the best result that avoid overfitting. In this case we choose 4 for appropriate K.





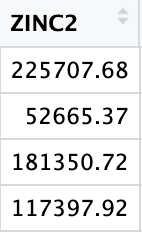
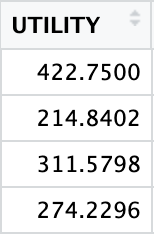
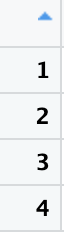
(iii) Define how you value or discern each cluster.

1. buildings that have Higher age, lager size and expensive stuff
2. buildings that have smaller size, utility and lower cost
3. buildings that have median size and for median income’s family
4. buildings that have higher Housing cost as a fraction of income

(iv) Name your clusters.

1. elder building
2. Cheaper building
3. Median income & bedroom
4. Rental building

(d) Create a new variable in your dataset which identifies which observation is within whichcluster (k-means only), then provide measures for each cluster on three variables (UTILITY,TOTALSAL, ZINC2 - know what they are for your assessment, don’t just give me thevariable names).



For cluster 1, it has the higher values in three variables, for cluster 2, it has lowest values of three variables in this model.

For cluster 3 and4, they only have slightly difference. But we still can recognize cluster 3 has higher values than cluster 4.

For cluster 1 and 3, “ToTSAL” dosen’t have significant difference.

E:

For the "zinc2" variable, each cluster group displays a significant difference. As all p-values are below 0.05, we can conclude that each group has a statistically significant difference.

> TukeyHSD(model\_zinc)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = ZINC2 ~ cluster, data = k\_m)

$cluster

diff lwr upr p adj

2-1 -173042.31 -193338.20 -152746.43 0e+00

3-1 -44356.97 -65656.91 -23057.02 5e-07

4-1 -108309.76 -128752.22 -87867.30 0e+00

3-2 128685.35 121445.28 135925.41 0e+00

4-2 64732.55 60655.17 68809.93 0e+00

4-3 -63952.80 -71594.12 -56311.48 0e+00

For the "Utility" variable, each cluster group displays a significant difference. As all p-values are below 0.05, we can conclude that each group has a statistically significant difference.

> TukeyHSD(model\_UTILITY)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = UTILITY ~ cluster, data = k\_m)

$cluster

diff lwr upr p adj

2-1 -207.90980 -242.00354 -173.81605 0

3-1 -111.17020 -146.95060 -75.38979 0

4-1 -148.52045 -182.86042 -114.18047 0

3-2 96.73960 84.57749 108.90171 0

4-2 59.38935 52.54002 66.23868 0

4-3 -37.35025 -50.18641 -24.51408 0

For the "Totsal" variable, most cluster group displays a significant difference. As all p-values are below 0.05, we can conclude that those group has a statistically significant difference.

However, When comparing Cluster 3 with Cluster 1, p-value is above 0.05, we cannot reject the null hypothesis that the means of these two groups are not significantly different.

> TukeyHSD(model\_TOTSAL)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = TOTSAL ~ cluster, data = k\_m)

$cluster

diff lwr upr p adj

2-1 -101910.471 -121105.86 -82715.08 0.0000000

3-1 -9810.662 -29955.67 10334.35 0.5940803

4-1 -51226.398 -70560.42 -31892.38 0.0000000

3-2 92099.809 85252.32 98947.30 0.0000000

4-2 50684.073 46827.78 54540.37 0.0000000

4-3 -41415.736 -48642.73 -34188.75 0.0000000