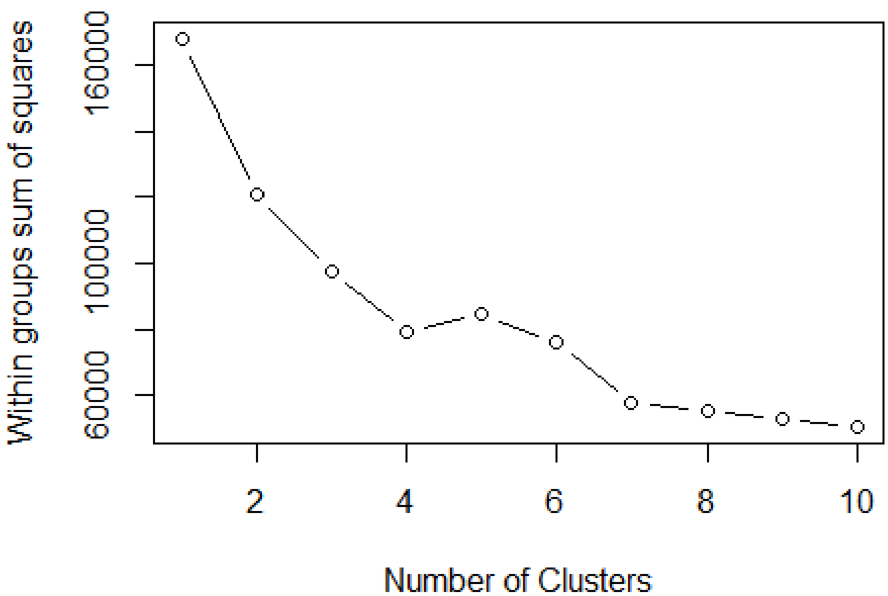
Clustering

Shiyou Yan

#Dataset from kaggle (https://www.kaggle.com/datasets/syuzai/perth-house-prices?resource=download)

1.k-Means clustering

```
mydata <- read.csv("C:/Program Files/datasets/Perth_House_Prices.csv")
mydata <- mydata[, c("PRICE","BEDROOMS", "BATHROOMS", "LAND_AREA", "FLO
OR_AREA")]
mydata <- scale(mydata)
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:10) wss[i] <- sum(kmeans(mydata,
   centers=i)$withinss)
plot(1:10, wss, type="b", xlab="Number of Clusters",
  ylab="Within groups sum of squares")
```

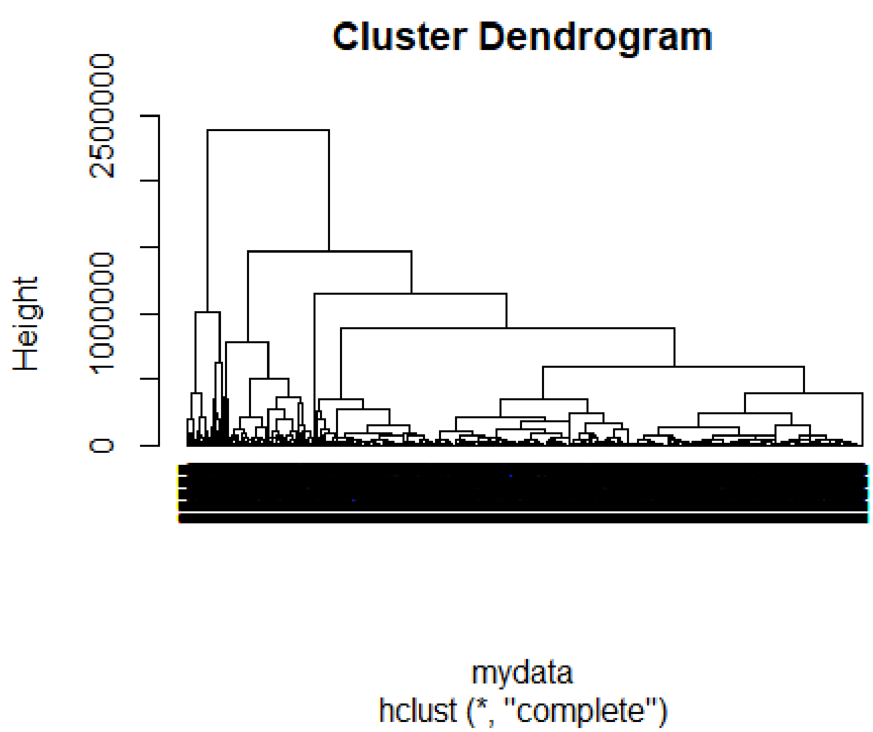



```
set.seed(1)
fit <- kmeans(mydata, centers = 5,nstart = 25)
aggregate(mydata,by=list(fit$cluster),mean)

##   Group.1      PRICE   BEDROOMS   BATHROOMS   LAND_AREA  FLOOR_AREA
## 1       1 -0.4601482 -0.9149740 -1.40113248 -0.05608916 -0.8611228
```

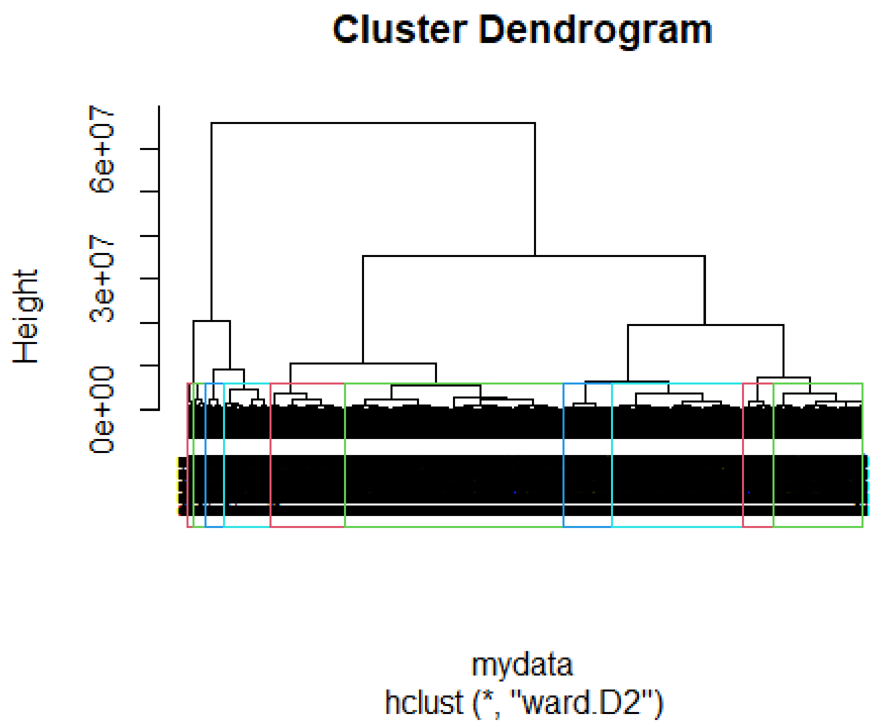
```
## 2         2  1.7752265  0.8769393  1.10783490  0.10798362  1.5599572
## 3         3 -0.1916127 -0.9369833  0.32942461 -0.07266516 -0.4109066
## 4         4  0.9399094  0.4532884  0.06287975 19.34748377  0.4909430
## 5         5 -0.2120373  0.5763961  0.31854503 -0.03885036  0.1528457

mydata <- cbind(mydata,cluster = fit$cluster)
```

2.Hierarchical clustering

```
mydata <- read.csv("C:/Program Files/datasets/Perth_House_Prices.csv")
mydata <- na.omit(mydata)
mydata <- mydata[, c("PRICE","BEDROOMS", "BATHROOMS", "LAND_AREA", "FLO
OR_AREA")]
mydata <- dist(mydata, method = "euclidean")
fit <- hclust(mydata, method = "complete")
plot(fit,cex = 0.6, hang = -1)
```

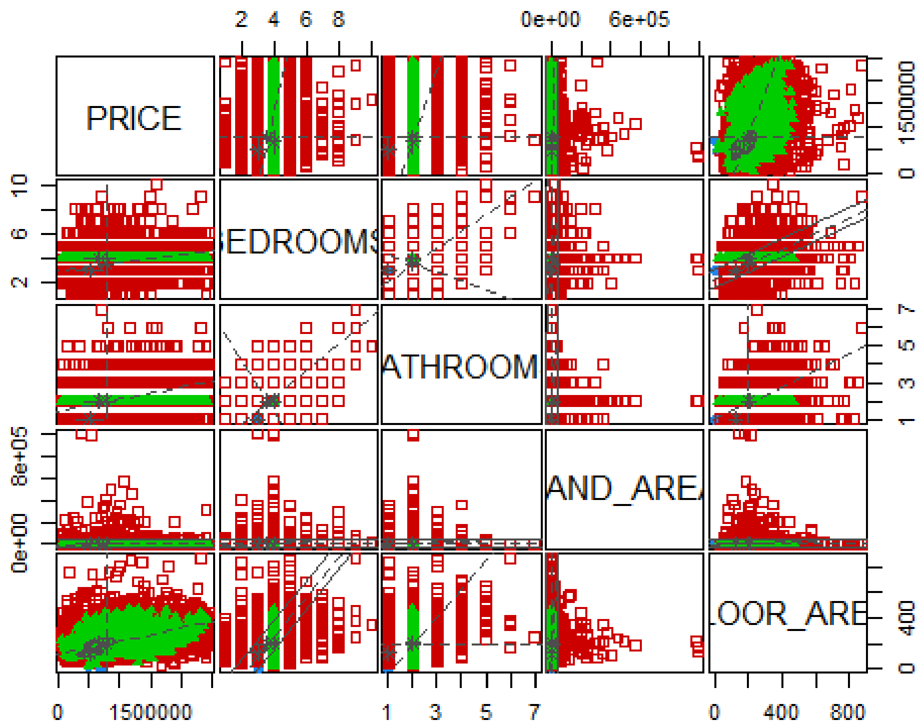



```
fit <- hclust(mydata,method ="ward.D2")
groups <- cutree(fit, k = 10)
plot(fit,cex=0.6)
rect.hclust(fit, k = 10, border = 2:5)
```

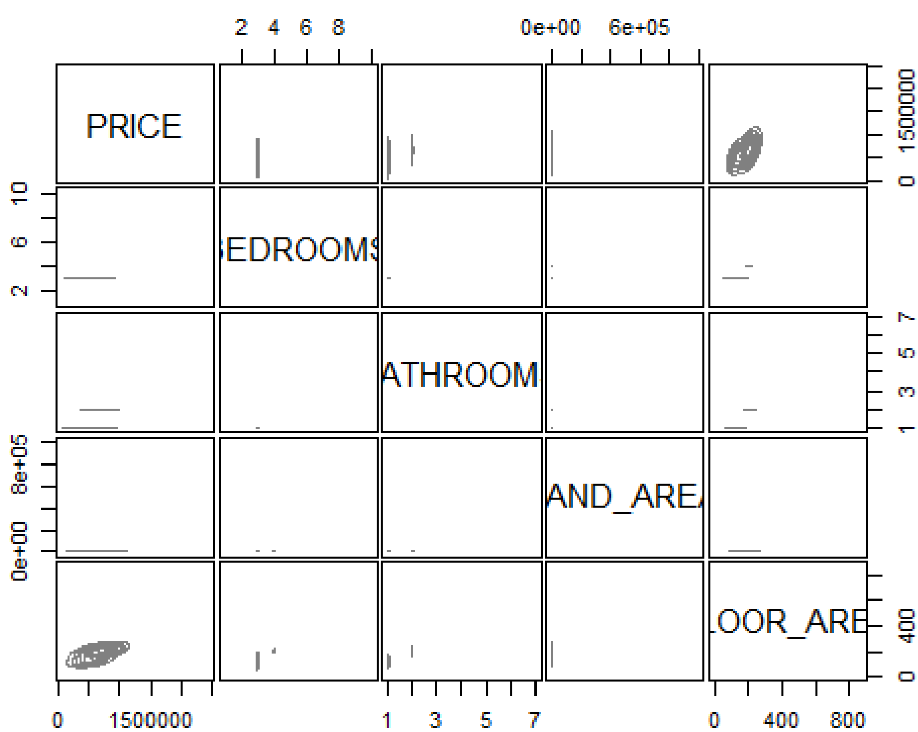


3.Model-based clustering

```
library(mclust)

## Warning: 程辑包'mclust'是用R版本4.2.3 来建造的

## Package 'mclust' version 6.0.0
## Type 'citation("mclust")' for citing this R package in publications.

mydata <- read.csv("C:/Program Files/datasets/Perth_House_Prices.csv")
mydata <- na.omit(mydata)
mydata <- mydata[, c("PRICE","BEDROOMS", "BATHROOMS", "LAND_AREA", "FLO
OR_AREA")]
fit <- Mclust(mydata)
plot(fit, what=c("classification"))
```




```r
plot(fit, what=c("density"))
```




```r
summary(fit,parameters = TRUE)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VEV (ellipsoidal, equal shape) model with 3 components:
##
##  log-likelihood     n df      BIC      ICL
##       -655084.1 22704 54 -1310710 -1310744
##
## Clustering table:
##    1    2    3
## 3644 9990 9070
##
## Mixing probabilities:
##         1         2         3
## 0.1605296 0.4401202 0.3993502
##
## Means:
##                  [,1]         [,2]        [,3]
## PRICE      509192.0109 7.789826e+05 670924.7423
## BEDROOMS        3.0000 3.583754e+00      4.0000
## BATHROOMS       1.0000 2.006571e+00      2.0000
## LAND_AREA     717.0867 5.237192e+03    672.0691
## FLOOR_AREA    122.9669 1.973740e+02    202.3620
##
## Variances:
## [,,1]
##                 PRICE      BEDROOMS     BATHROOMS     LAND_AREA    
FLOOR_AREA
## PRICE      41990821903  0.000000e+00  0.000000e+00 -2.656692e+06  2.
481576e+06
## BEDROOMS             0  1.118436e-05 -4.161360e-08 -8.320145e-22 -9.
217508e-22
## BATHROOMS            0 -4.161360e-08  4.397480e-06  3.667102e-22 -3.
223386e-22
## LAND_AREA     -2656692 -8.320145e-22  3.667102e-22  5.522088e+04  5.
974839e+02
## FLOOR_AREA     2481576 -9.217508e-22 -3.223386e-22  5.974839e+02  8.
688111e+02
## [,,2]
##                  PRICE     BEDROOMS    BATHROOMS    LAND_AREA   FLO
OR_AREA
## PRICE      1.278797e+15 8.425173e+08 8.295332e+08 2.292686e+11 1.341
389e+11
## BEDROOMS   8.425173e+08 1.451680e+03 1.021068e+03 1.547820e+05 2.277
473e+05
## BATHROOMS  8.295332e+08 1.021068e+03 7.894488e+02 1.483226e+05 1.607
970e+05
## LAND_AREA  2.292686e+11 1.547820e+05 1.483226e+05 1.718012e+09 2.431
295e+07
```

```
## FLOOR_AREA 1.341389e+11 2.277473e+05 1.607970e+05 2.431295e+07 3.574
297e+07
## [,,3]
##                  PRICE      BEDROOMS     BATHROOMS   LAND_AREA    FL
OOR_AREA
## PRICE      55293947599  0.000000e+00  0.000000e+00 5412097.211  4.43
5039e+06
## BEDROOMS             0  1.472318e-05 -2.077077e-07       0.000 -7.92
1011e-22
## BATHROOMS            0 -2.077077e-07  5.795141e-06       0.000 -5.46
9089e-22
## LAND_AREA      5412097  0.000000e+00  0.000000e+00   73023.282  1.44
6545e+03
## FLOOR_AREA     4435039 -7.921011e-22 -5.469089e-22    1446.545  1.30
7199e+03
```

4. Write a paragraph comparing the results of each algorithm and what insights they gave you to this data There are three popular clustering algorithms, K-means clustering, hierarchical clustering, and model-based clustering. They are used to be analyze the Perth house prices data set.

K-means clustering is a straightforward algorithm that works well with large data sets. It partitions the data into a predefined number of clusters based on their similarities. K-means clustering can provide insights into which neighborhoods or areas have similar house prices and can help identify outlines in the data. Cluster 1 was characterized by smaller houses with fewer bedrooms and bathrooms.Cluster 2 had larger houses with more bedrooms and bathrooms.Cluster 3 had houses with high land areas and floor areas.Cluster 4 had houses with relatively low land areas and floor areas.Cluster 5 was characterized by houses with high prices and relatively large land and floor areas.

Hierarchical clustering can be useful for identifying natural groupings in the data and can help visualize the relationships between different neighborhoods or regions in Perth. In this analysis, the Ward.D2 method was employed to group the observations into distinct clusters. Through analysis of the results, it was discovered that some clusters were dominated by larger houses with multiple bedrooms and bathrooms, while others contained smaller houses with fewer bedrooms and bathrooms. Additionally, some clusters contained houses with substantial land and floor areas, while others featured houses with relatively limited land and floor space. The results obtained from hierarchical clustering are similar to those obtained from k-means clustering.

Model-based clustering is to perform Gaussian finite mixture modeling on the Perth house prices data set. The classification plot shows the observations grouped into different clusters, while the density plot shows the density of the data within each cluster.