Code ▾

Clustering

Nikilas John

October 9th, 2022

Read in the data

Hide

```
df <- read.csv("avocado.csv")
```

K-Means Cluster Analysis

Since we have two groups we would like to cluster, we will use that to find the K-Means Analysis

Hide

```
# K-Means Cluster Analysis
fit <- kmeans(df[, c(4,8)], 3) # 2 cluster solution
```

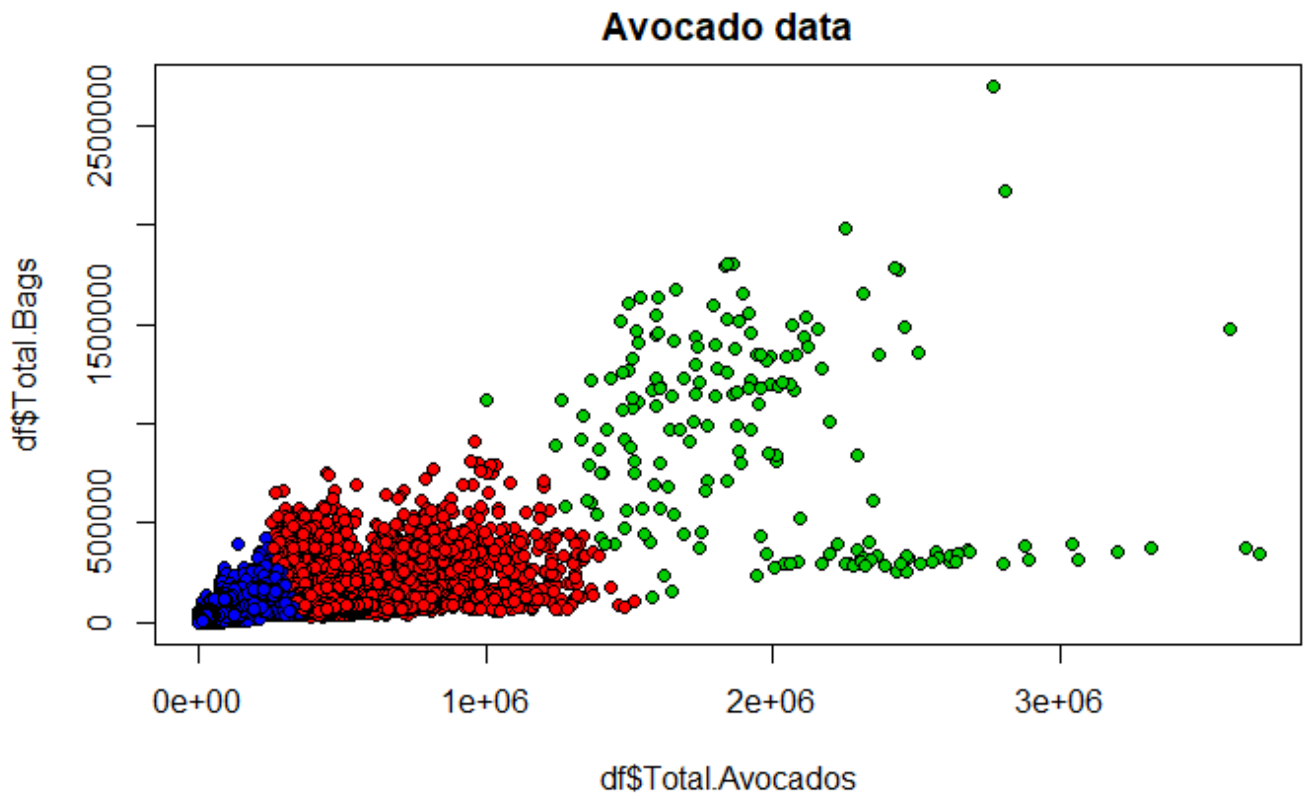
Plotting the Clusters

Time to really see the fruits of our labor and plot the clusters

As we can see, there is three sort of distinct clusters. The lines between the clusters is not as defined as I thought it would be, but oh well. The green and blue clusters are pretty close together and most likely only separated by the algorithm, but the red cluster represents almost all of the outliers (in regards to the majority of the data).

Hide

```
#plot the clusters
plot(df$Total.Avocados, df$Total.Bags, pch=21, bg=c("red","green3", "blue")
[unclass(fit$cluster)], main="Avocado data")
```



Hierarchical Clustering

Since Hierarchical clustering works best on smaller datasets, we are going to subset the data to a 20 values

Hide

```
set.seed(80)
library("dplyr")
sample <- sample_n(df, 20)
```

Plotting the Dendogram

This section will display the dendogram using average linkage (the average distance between points in clusters)

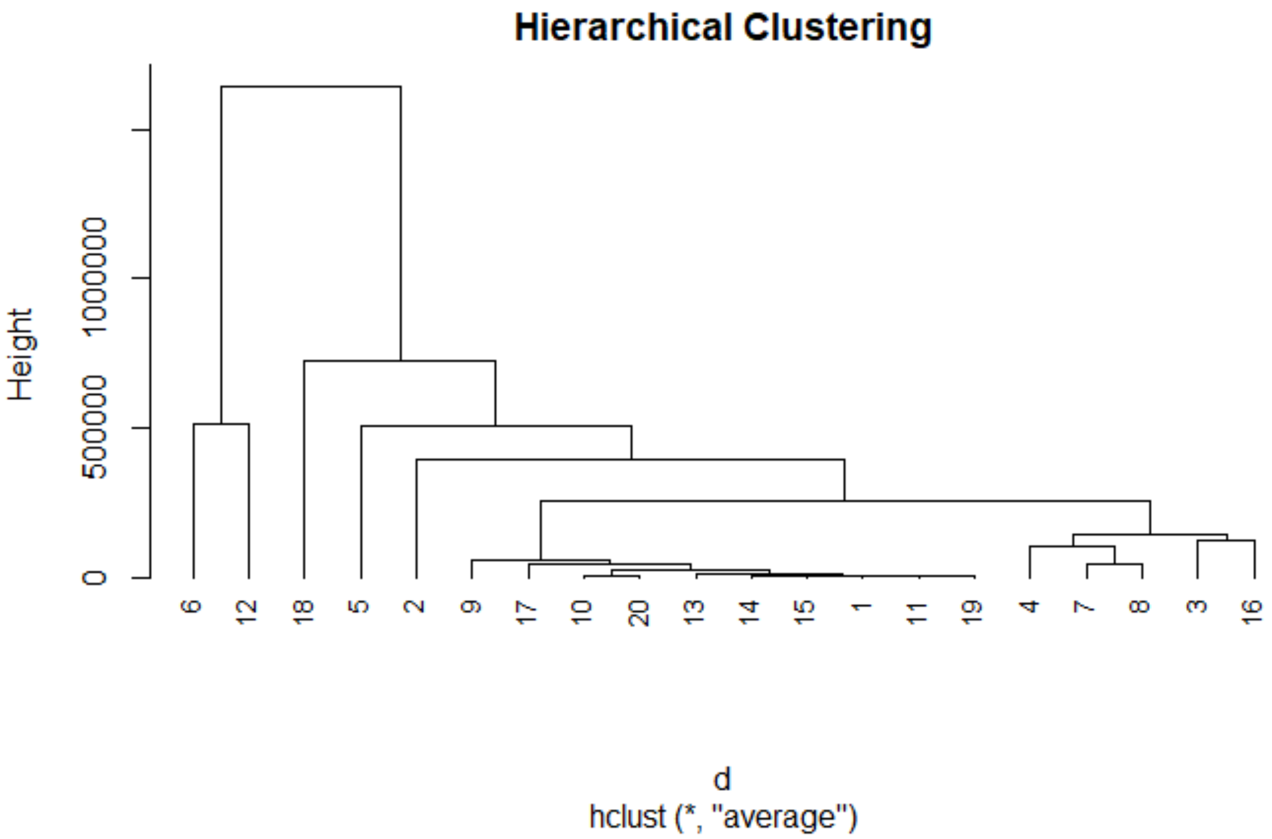
Hide

```
d <- dist(sample)
```

```
Warning: NAs introduced by coercion
```

Hide

```
fit.average <- hclust(d, method="average")
plot(fit.average, hang=-1, cex=.8,
     main="Hierarchical Clustering")
```



Model-Based Clustering

All information in this section was found through these links: https://www.statmethods.net/advstats/cluster.html (https://www.statmethods.net/advstats/cluster.html)

Model-based clustering gets its name from assuming a variety of data models and “apply maximum likelihood estimation and Bayes criteria to identify the most likely model and number of clusters” (cited from the first link)

The benefit I see to this clustering method is that there is that the programmer does not set the number of clusters, instead the algorithm does. A lot of the work is actually done by the algorithm, as we will see by how short the code is

Hide

```
# Model Based Clustering
library(mclust)
test <- Mclust(df)
```

```
  |==========================================================================================
===========        |  93%
  |
  |==========================================================================================
============       |  94%
  |
  |==========================================================================================
=============      |  94%
  |
  |==========================================================================================
==============     |  95%
  |
  |==========================================================================================
===============    |  96%
  |
  |==========================================================================================
================   |  97%
  |
  |==========================================================================================
================   |  98%
  |
  |==========================================================================================
=================  |  98%
  |
  |==========================================================================================
================== |  99%
  |
  |==========================================================================================
===================| 100%
```

Hide

```
plot(test) # plot results
```

```
Model-based clustering plots:

1: BIC
2: classification
3: uncertainty
4: density
```

Hide

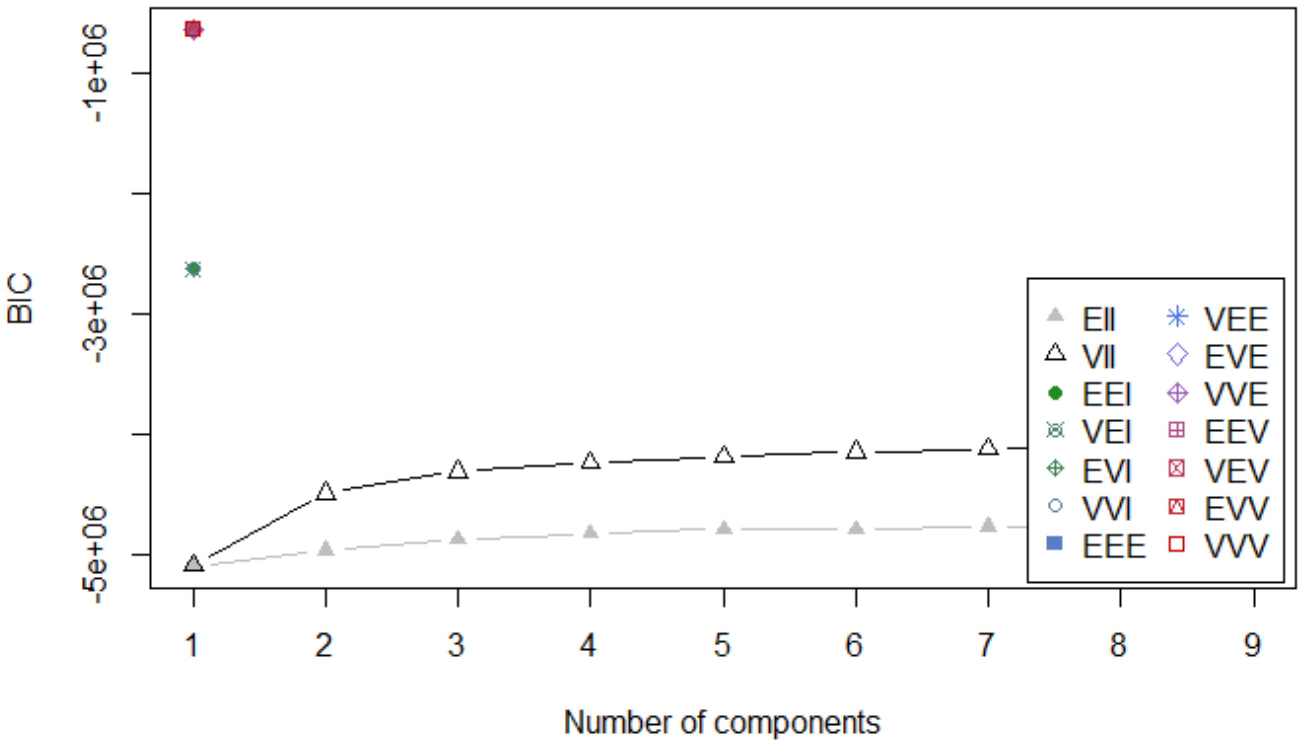
```
1
```

```
Model-based clustering plots:

1: BIC
2: classification
3: uncertainty
4: density
```

Hide

```
0
```



Hide

```
summary(test) # display the best model
```

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust XXX (ellipsoidal multivariate normal) model with 1 component:
```

log-likelihood	n	df	BIC	ICL
<dbl>	<int>	<dbl>	<dbl>	<dbl>
-320361.5	14869	104	-641722.2	-641722.2

1 row

```


Clustering table:
    1 
14869 
```

Comparison of the algorithms

All of these algorithms have their advantages and disadvantages. Being able to cluster data into different groups can allow for another method of finding trends in a dataset. The applications this can have in the field will, however, will result in groups with no label. k-Means clustering has the advantage being (personally) the easiest to read and interpret. The disadvantage of this algorithm is that it requires multiple iterations to get the best result. This can sometimes lead to long processing times. Hierarchical clustering is an algorithm that is used to cluster smaller datasets, which can be an advantage and a disadvantage at the same time. This is the best algorithm to cluster smaller datasets, but with larger datasets, interpreting the dendogram that is outputted by this algorithm is not the best. Model-based clustering allows for a non-heuristic approach to clustering and an advantage that it brings is that it sets its own optimal number of clusters. The disadvantage is that the output from the commands is very hard to interpret without extensive knowledge on it. The best algorithm out of the three in my opinion was the k-Means algorithm. It was very easy to understand and despite having to define and optimal “k”, it allowed me to draw more conclusions from the resulting plot