

K-Means, Hierarchical Clustering, & Model Based Clustering

Subash Chandra, Derrick Martin, Abigail Solomon, Aditi Chaudhari

2022-10-08

K-Means

We will attempt to cluster the Dry Bean Data Set from <http://archive.ics.uci.edu/ml/datasets/Dry+Bean+Dataset>. Firstly, the data is read into a data frame.

```
library("readxl")
df <- read_excel("Dry_Bean_Dataset.xlsx")
```

Next, we want to see what the first few rows look like in the data frame.

```
head(df)

## # A tibble: 6 x 17
##   Area Perimeter Major-1 Minor-2 Aspec-3 Eccen-4 Conve-5 Equiv-6 Extent Solid-7
##   <dbl>      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 28395      610.    208.    174.    1.20    0.550   28715    190.    0.764    0.989
## 2 28734      638.    201.    183.    1.10    0.412   29172    191.    0.784    0.985
## 3 29380      624.    213.    176.    1.21    0.563   29690    193.    0.778    0.990
## 4 30008      646.    211.    183.    1.15    0.499   30724    195.    0.783    0.977
## 5 30140      620.    202.    190.    1.06    0.334   30417    196.    0.773    0.991
## 6 30279      635.    213.    182.    1.17    0.520   30600    196.    0.776    0.990
## # ... with 7 more variables: roundness <dbl>, Compactness <dbl>,
## #   ShapeFactor1 <dbl>, ShapeFactor2 <dbl>, ShapeFactor3 <dbl>,
## #   ShapeFactor4 <dbl>, Class <chr>, and abbreviated variable names
## #   1: MajorAxisLength, 2: MinorAxisLength, 3: AspectRatio, 4: Eccentricity,
## #   5: ConvexArea, 6: EquivDiameter, 7: Solidity
```

We have decided to try to cluster the dry beans based off of area and perimeter. The data is scaled because we are not entirely sure if area and perimeter were measured in the same units.

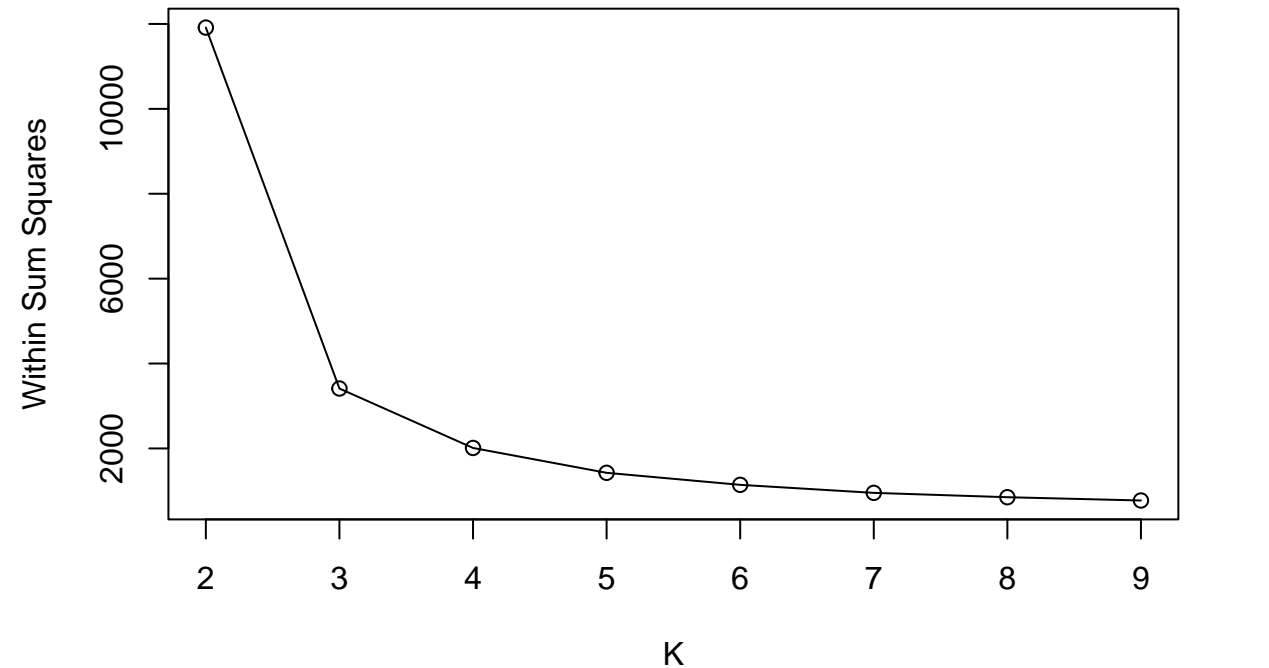
```
scaled <- df[c("Area", "Perimeter")]
scaled <- scale(scaled)
```

Next, we use this algorithm to determine what a good number of clusters is. As seen from the graph below, there is an elbow at K=3, so we will use 3 clusters.

```
plot_withinss <- function(df, max_clusters){
  withinss <- rep(0, max_clusters-1)
  for (i in 2:max_clusters){
    set.seed(1234)
    withinss[i] <- sum(kmeans(df, i)$withinss)
  }
  plot(2:max_clusters, withinss[2:max_clusters], type="o",
       xlab="K", ylab="Within Sum Squares")
}
```

```
plot_withinss(scaled, 9)
```

```
## Warning: Quick-TRANSfer stage steps exceeded maximum (= 680550)
```



Now, let's use the `kmeans()` function to build our model.

```
set.seed(1234)
beans_cluster <- kmeans(scaled, 3, nstart=25)
beans_cluster
```

```
## K-means clustering with 3 clusters of sizes 8734, 4356, 521
```

##

```
## Cluster means:
```

```
##          Area  Perimeter
```

```
## 1 -0.4993405 -0.5889355
```

```
## 2  0.5095162  0.7727822
```

```
## 3  4.1109160  3.4117532
```

##

```
## Clustering vector:
```

```
##      [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```
## [37] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```
## [73] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```
## [109] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```
## [145] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

[illegible][illegible][illegible]

[illegible]

[illegible]

[illegible]

[illegible]

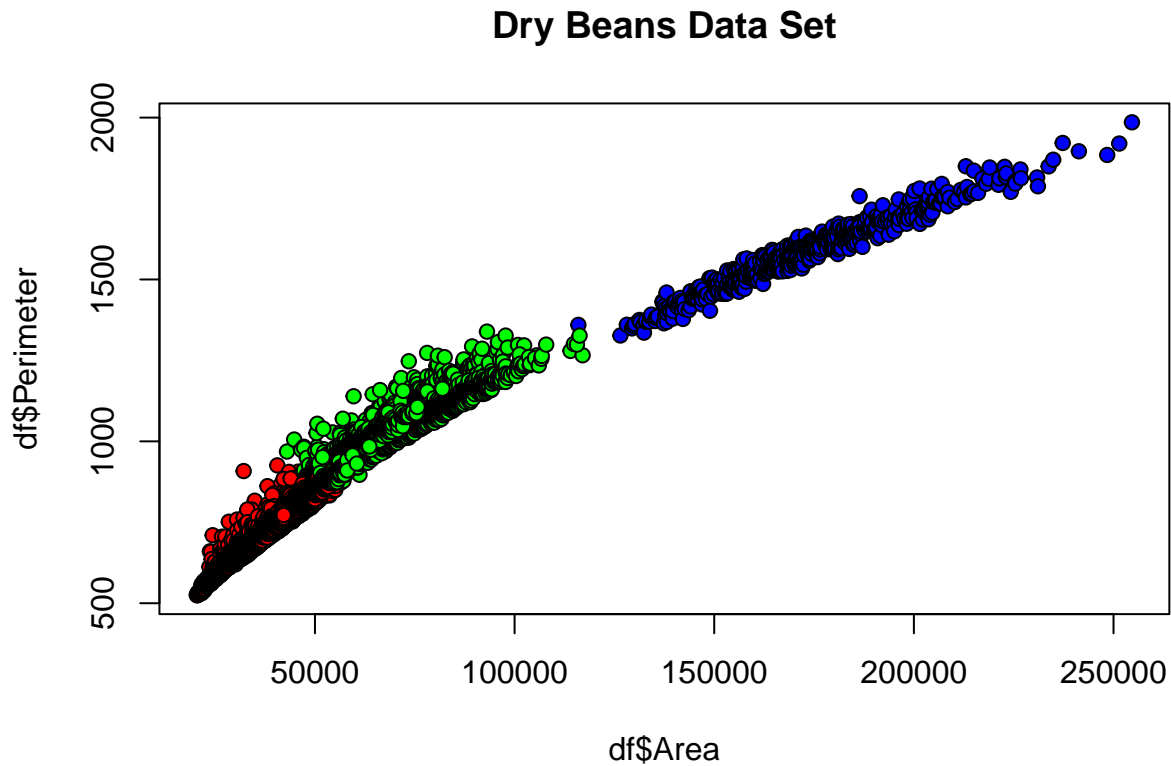
[illegible]

[illegible]


```
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"       "
```

After building our model, let's graph the clusters on a graph. Each color represents a different cluster.

```
plot(df$Area, df$Perimeter, pch=21, bg=c("red", "green", "blue")[unclass(beans_cluster$cluster)],
     main="Dry Beans Data Set")
```



Hierarchical Clustering

Next, let's attempt clustering the Dry Beans Data Set with hierarchical clustering. Using the `hclust()` function, we can create a dendrogram of the clustering (as seen below). The dendrogram is cut with $k = 3$.

```
d <- dist(scaled, method="euclidean")
fit.average <- hclust(d, method="ward")
```

```
## The "ward" method has been renamed to "ward.D"; note new "ward.D2"
```

```
plot(fit.average, hang=-1, cex=.8, main="Hierarchical Clustering")
groups <- cutree(fit.average, k=3)
rect.hclust(fit.average, k=3, border="red")
```

Hierarchical Clustering



d
hclust (*, "ward.D")

Model Based Clustering

Now, let's use model based clustering to cluster the Dry Beans Data Set. Using the mclust library, we can determine the most likely model and an ideal number of clusters. The BIC suggests VEV with 9 groups.

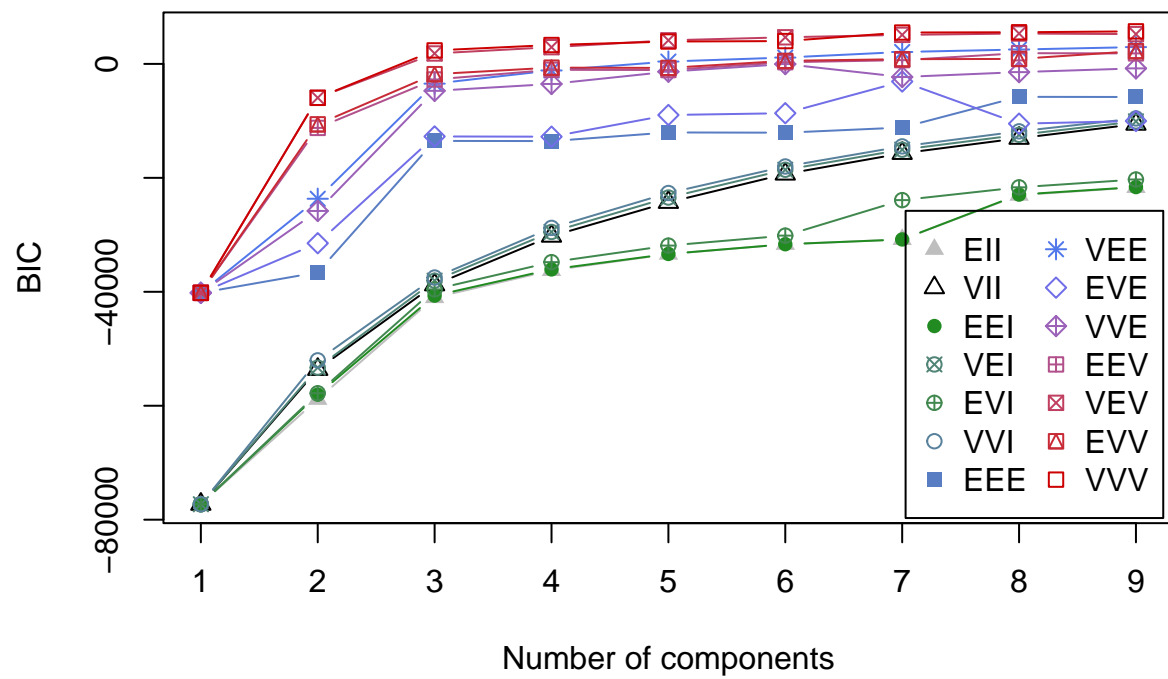
```
library(mclust)
```

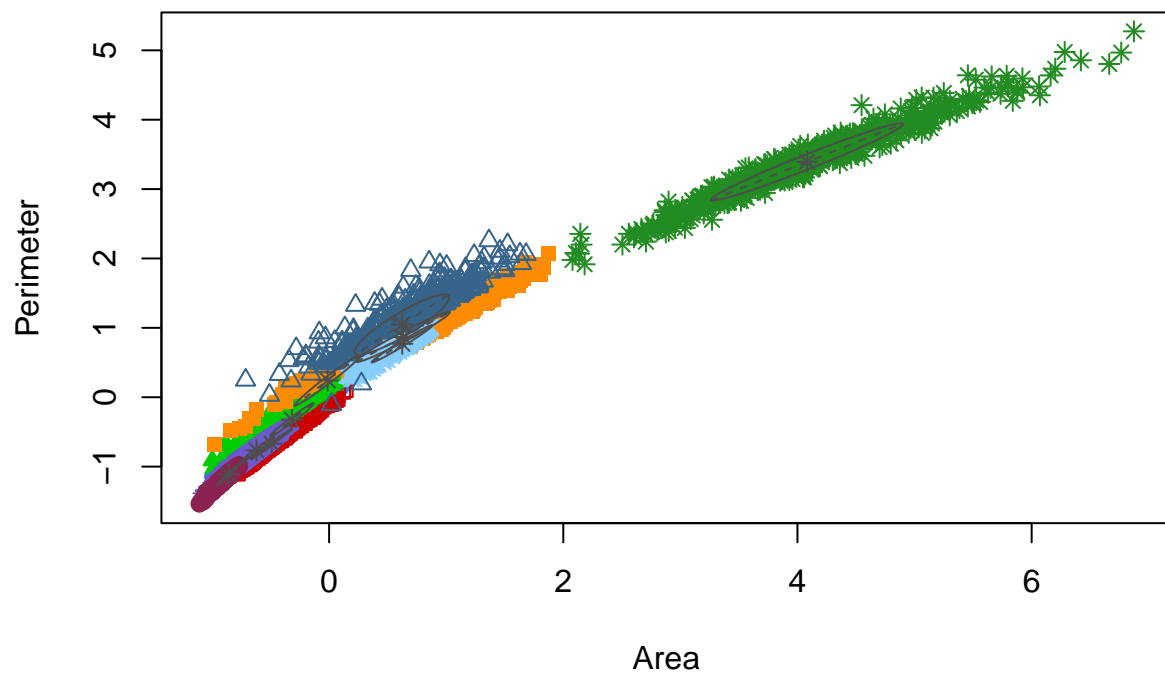
```
## Package 'mclust' version 5.4.10
```

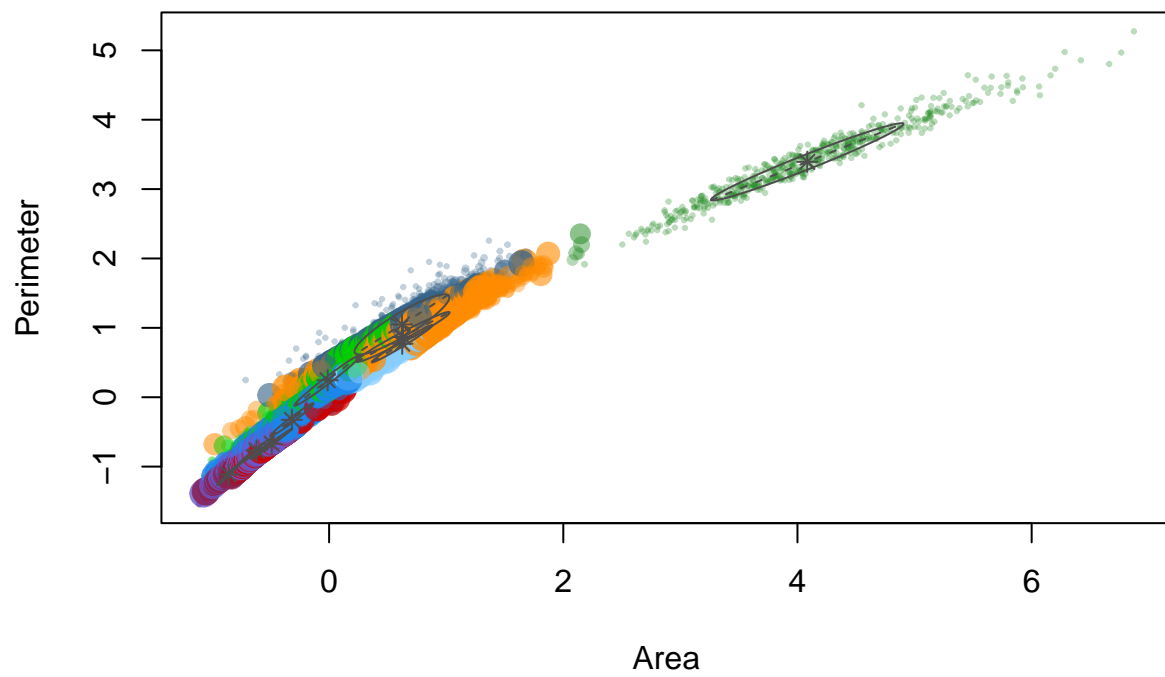
```
## Type 'citation("mclust")' for citing this R package in publications.
```

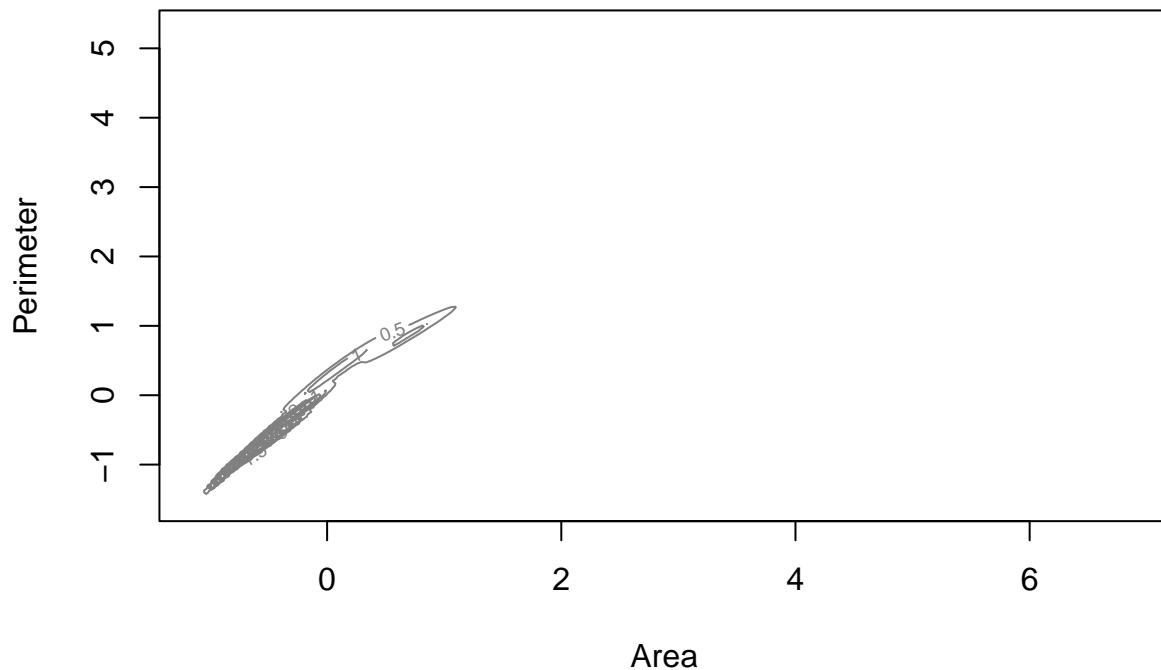
```
m_fit <- Mclust(scaled)
```

```
plot(m_fit)
```









```
summary(m_fit) # displays the best model
```

```
## -----
## Gaussian finite mixture model fitted by EM algorithm
## -----
##
## Mclust VVV (ellipsoidal, varying volume, shape, and orientation) model with 9
## components:
##
##   log-likelihood      n df      BIC      ICL
##      3108.445 13611 53 5712.402 -1362.76
##
## Clustering table:
##    1    2    3    4    5    6    7    8    9
## 3092 1484 2068 2515 1889  599 1048  526  390
```

Results

In this exercise, we used three different algorithms in order to cluster the data from the Dry Beans Data Set based on the area and perimeter of the bean. The k-means clustering algorithm identifies k centers and groups the other observations around those centers. In our notebook, we explored our data and found that $k=3$ was the optimal amount of clusters. We used the `kmeans()` function to create those clusters, but learned that the k-means algorithm may not have performed incredibly well due to large within sum of squares values. The hierarchical clustering algorithm uses a distance measure in order to create clusters that are organized within a hierarchy. In our notebook, we used the `hclust()` function to create a dendrogram, which we cut at $k=3$. Hierarchical clustering tends to get bogged down with larger amounts of data, which makes using hierarchical clustering on this data set difficult. Finally, the model based clustering algorithm utilizes a

variety of data models and Bayes criteria in order to find the model that fits the best and then determine the ideal number of clusters. In our notebook, we used the `mclust` function to graph the BIC, which suggested VEV with 9 groups.