Clustering

Bishal Neupane, Saugat Gyawali, Spencer Gray, Michael Stinnett

Data Set Link (https://archive.ics.uci.edu/ml/datasets/Dry+Bean+Dataset)

I used data describing 7 types of dry beans. Originally this data was intended to be used for classification, but I omitted the target bean class to enable clustering.

Set up Data

```
rm(list=ls())
```

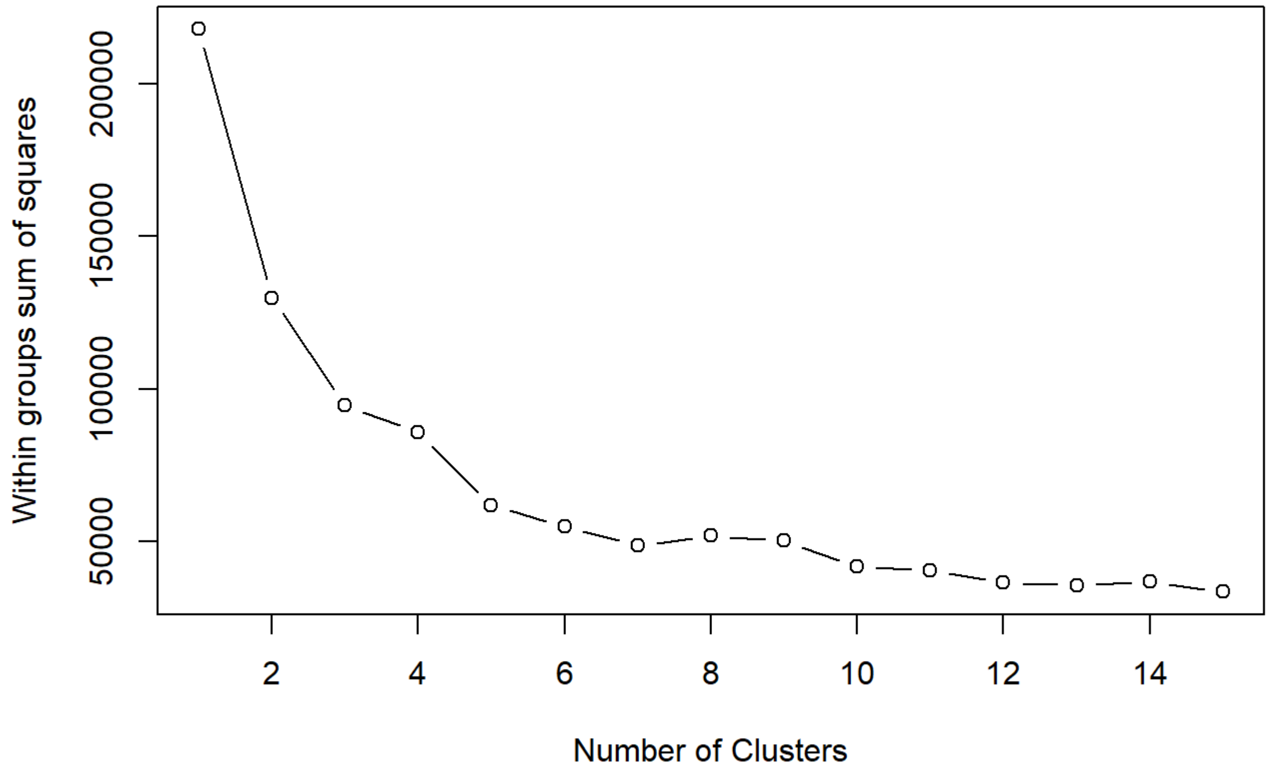
```
set.seed(1234)
data <- read.csv("Dry_Bean_Dataset.csv")
data <- na.omit(data)
```

```
data.scaled <- data[, -c(17)]
data.scaled <- scale(data.scaled)
```

Kmeans

Graphing a few groups sum of squares with different cluster size to see if 7 clusters will really work out.

```
wss <- (nrow(data.scaled)-1)*sum(apply(data.scaled,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(data.scaled,
   centers=i)$withinss)
plot(1:15, wss, type="b", xlab="Number of Clusters",
  ylab="Within groups sum of squares")
```



Performing kmeans operation and putting the cluster profiles back into the data

```
fit <- kmeans(data.scaled, 7)
aggregate(data.scaled,by=list(fit$cluster),FUN=mean)
```

Grou… <int>	Area <dbl>	Perimeter <dbl>	MajorAxisLength <dbl>	MinorAxisLength <dbl>	AspectRation <dbl>	Eccent
1	-0.32903010	-0.3619829	-0.3541018	-0.25560484	-0.25238297	-0.00772
2	-0.77058050	-0.9731079	-0.9464942	-0.89101907	-0.43432667	-0.20666
3	-0.46778917	-0.6257123	-0.8395457	-0.01951913	-1.41908510	-1.88675
4	4.11098583	3.4109128	3.1890420	3.82874389	0.01084531	0.21483
5	0.03344718	0.3224118	0.6405849	-0.40215522	1.84805407	1.29378
6	-0.30537665	-0.2687848	-0.2260681	-0.34456047	0.11939338	0.33521
7	0.71393517	0.9612035	0.8849435	0.82087086	0.30962112	0.44550

7 rows | 1-7 of 17 columns

```
data <- data.frame(data, fit$cluster)
```

This is a small example showing that it seems like the clustering generally placed the observations into the already their already defined classifications that was omitted from the clustering analysis

```
head(data[, c(17,18)])
```

	Class <chr>	fit.cluster <int>
1	SEKER	3
2	SEKER	3
3	SEKER	3
4	SEKER	3
5	SEKER	3
6	SEKER	3

6 rows

```
tail(data[, c(17,18)])
```

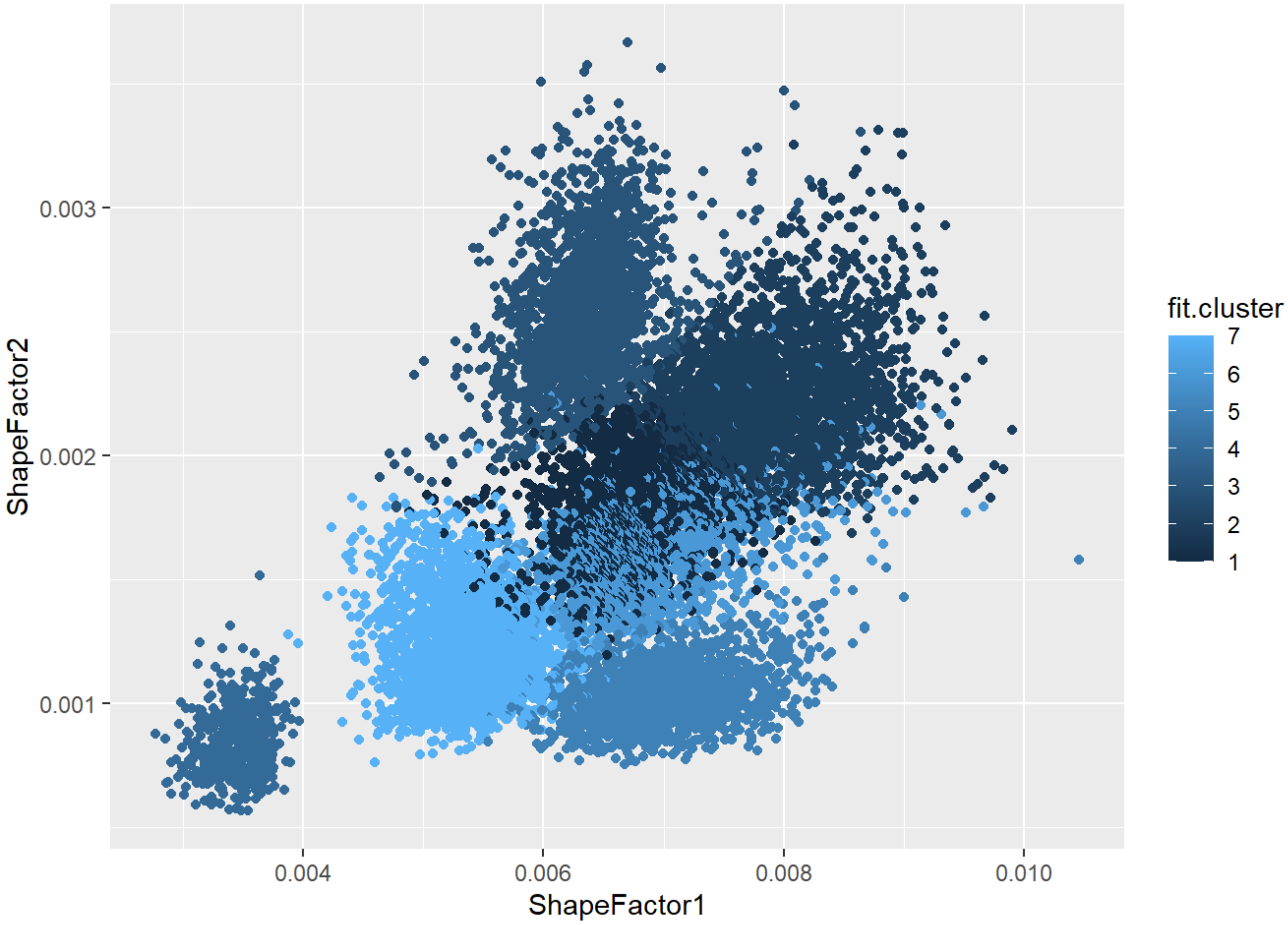
	Class <chr>	fit.cluster <int>
13606	DERMASON	1
13607	DERMASON	1
13608	DERMASON	1
13609	DERMASON	1
13610	DERMASON	6
13611	DERMASON	1

6 rows

A simple k means cluster plot

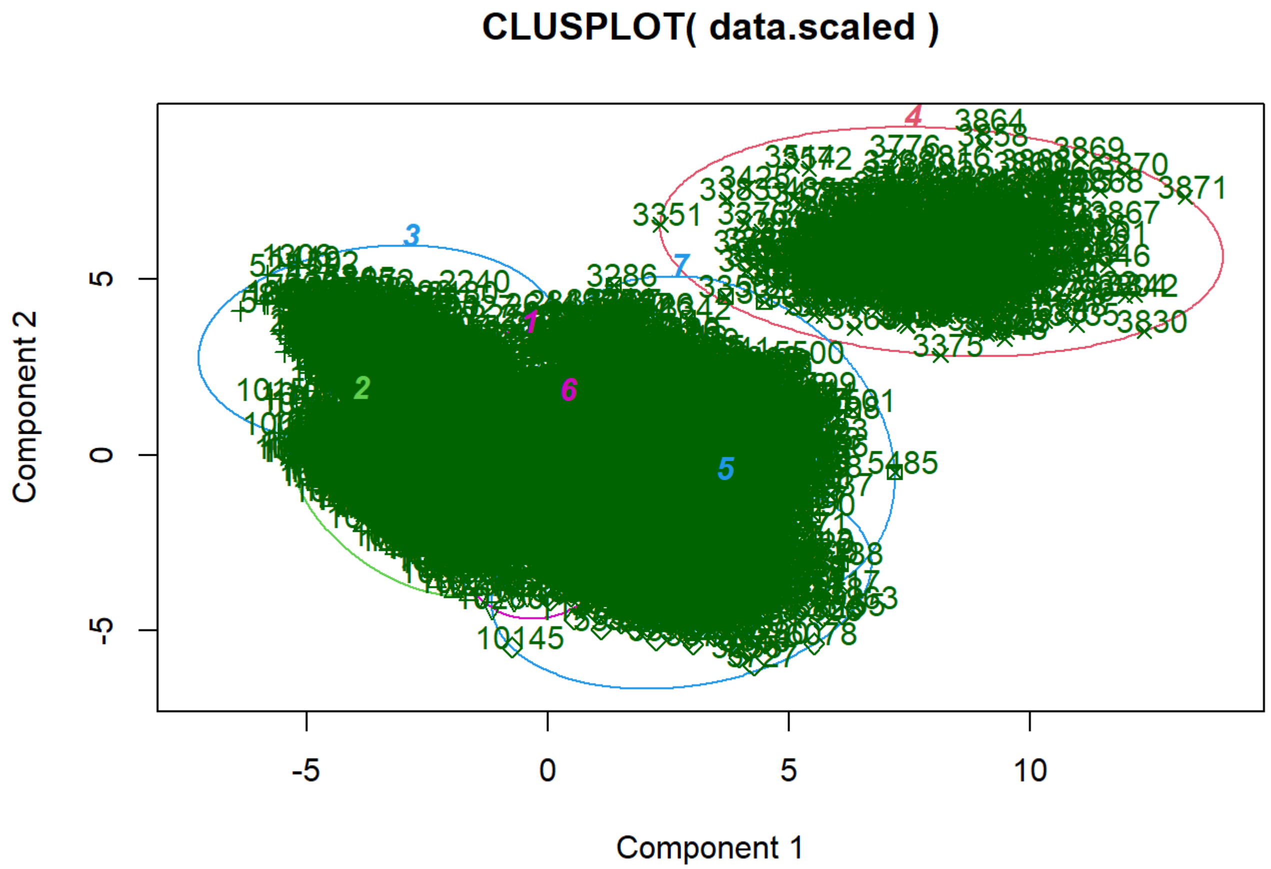
```
library("ggplot2")

ggplot(data,aes(x=ShapeFactor1,y=ShapeFactor2,group=fit.cluster)) +
  geom_point(aes(color=fit.cluster))
```



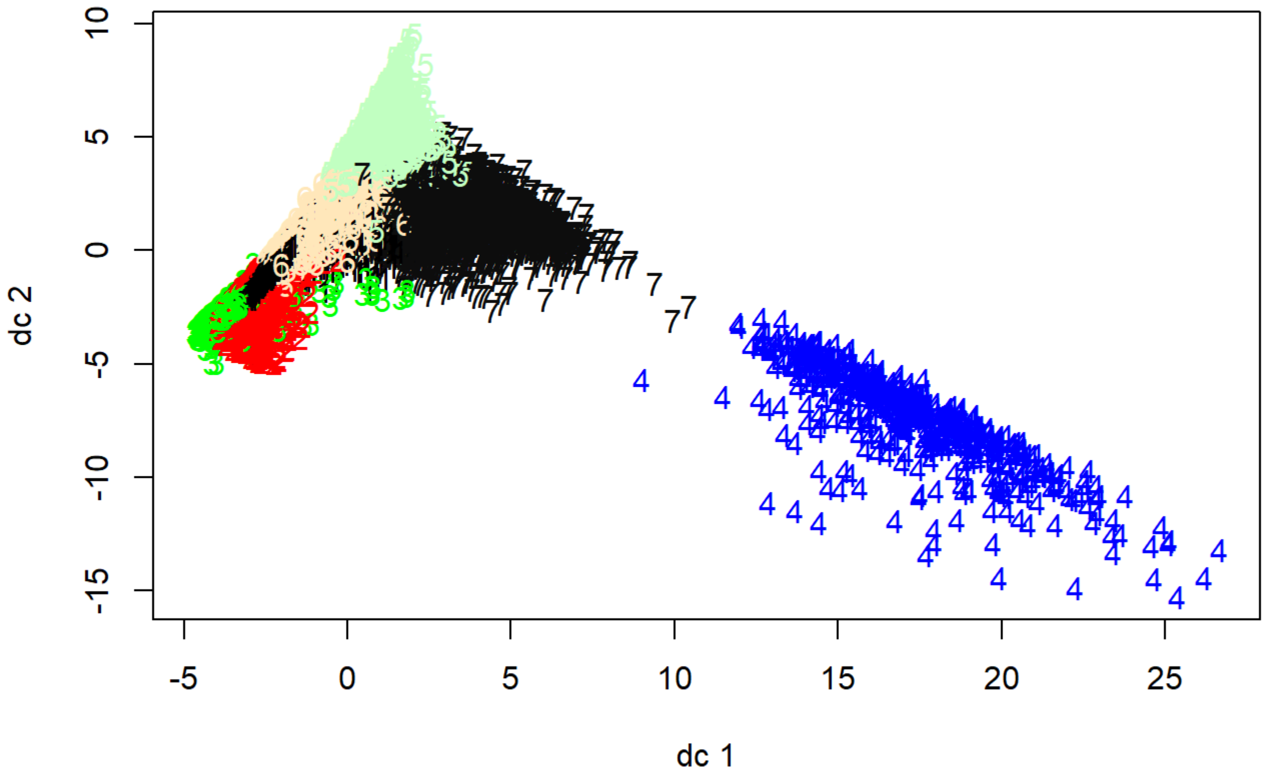
Some more complex cluster plots

```
library(cluster)
clusplot(data.scaled, fit$cluster, color=TRUE,
   labels=2, lines=0)
```



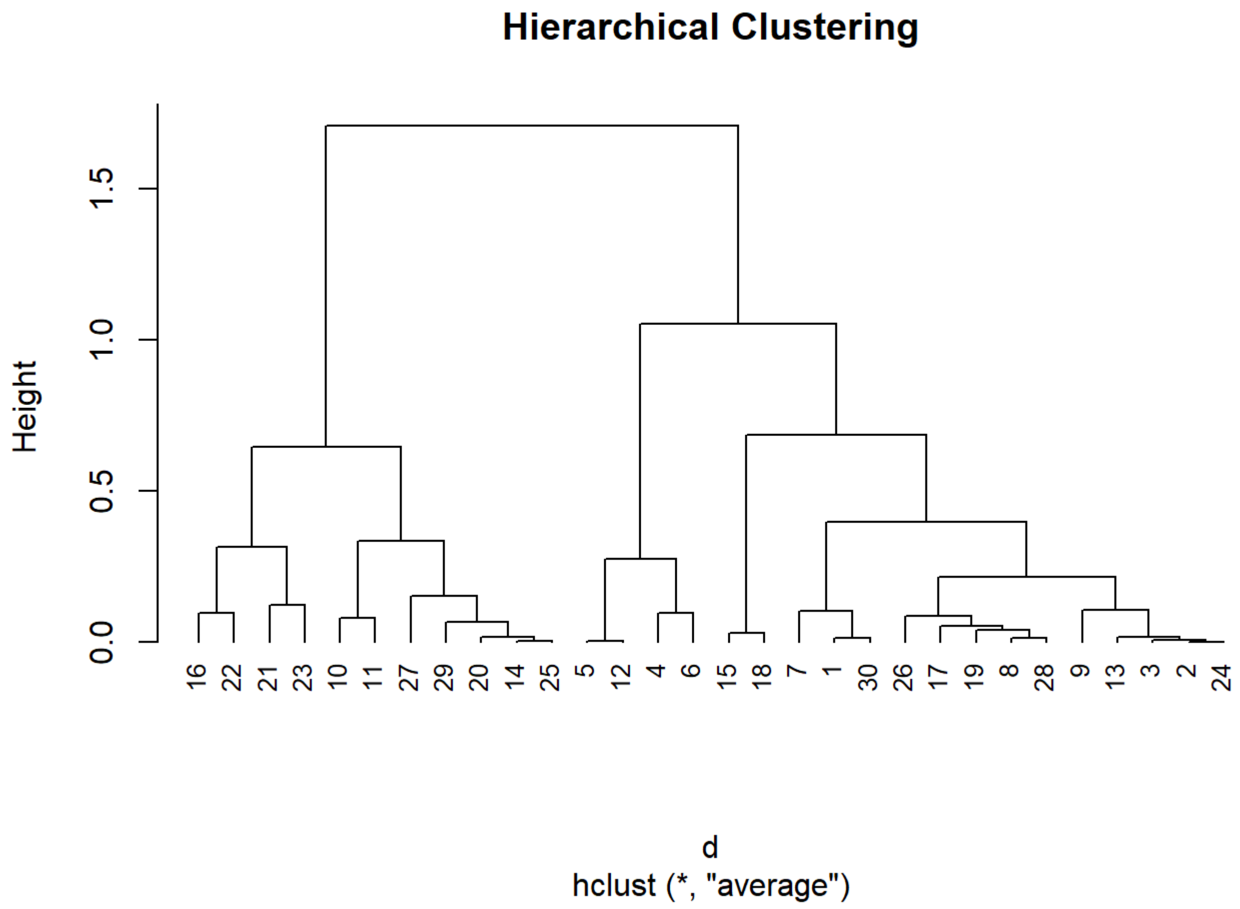
These two components explain 81.9 % of the point variability.

```
library(fpc)
plotcluster(data.scaled, fit$cluster)
```



Hierarchical

```
d <- dist(sample(data.scaled[, c(13,14,15,16)],30))
fit.average <- hclust(d, method="average")
plot(fit.average, hang=-1, cex=.8,
     main="Hierarchical Clustering")
```



Model Based

```
library(mclust)
```

```
## Package 'mclust' version 5.4.10
## Type 'citation("mclust")' for citing this R package in publications.
```

```
fit.m <- Mclust(data)
summary(fit.m)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VEV (ellipsoidal, equal shape) model with 9 components:
##
##  log-likelihood     n   df     BIC     ICL
##        745655.7 13611 1573 1476339 1476318
##
## Clustering table:
##    1    2    3    4    5    6    7    8    9
##  624 1807 1492  521 1747 1545 3036 1203 1636
```

Comparisons

These three clustering solutions set out to find different things about the data.

Kmeans tries to group observations into meaningful groups. What exactly the groups mean is not always easy to find. In our case we knew to try 7 clusters because this data has observations for 7 different kinds of beans. Kmeans was able catch on to this classification pretty closely, however, classification does not really matter to Kmeans.

Hierarchical clustering tries to find if there is some type of hierarchical taxonomy within the data. Interestingly it found that most beans belong to one of two main families. This is not something that was directly said within the data set, but both hierarchical and kmeans point towards this type of dichotomy.

Finally model-based clustering just told us the general shape of our data. It is VEV (ellipsoidal, equal shape). I could not find much on what exactly that means.

Ultimately these models reaffirmed things we already knew about the data set and gave us some hints into something that can be investigated furthered, like the bean family hierarchy.