ISE 5103 Intelligent Data Analytics Homework #8

Instructor: Charles Nicholson

By

Akhila Podupuganti

---

1. Data taken from UCI website.

Filter

	Sex	Length	Diameter	Height	Whole_weight	Shucked_weight	Viscera_weight	Shell_weight	Rings
1	M	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.1500	15
2	M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.0700	7
3	F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.2100	9
4	M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.1550	10
5	I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.0550	7
6	I	0.425	0.300	0.095	0.3515	0.1410	0.0775	0.1200	8
7	F	0.530	0.415	0.150	0.7775	0.2370	0.1415	0.3300	20
8	F	0.545	0.425	0.125	0.7680	0.2940	0.1495	0.2600	16
9	M	0.475	0.370	0.125	0.5095	0.2165	0.1125	0.1650	9
10	F	0.550	0.440	0.150	0.8945	0.3145	0.1510	0.3200	19
11	F	0.525	0.380	0.140	0.6065	0.1940	0.1475	0.2100	14
12	M	0.430	0.350	0.110	0.4060	0.1675	0.0810	0.1350	10

Data is about abalone from physical measurements.

Attribute Information:

Sex / nominal / -- / M, F, and I (infant)

Length / continuous / mm / Longest shell measurement

Diameter / continuous / mm / perpendicular to length

Height / continuous / mm / with meat in shell

Whole weight / continuous / grams / whole abalone

Shucked weight / continuous / grams / weight of meat

Viscera weight / continuous / grams / gut weight (after bleeding)

Shell weight / continuous / grams / after being dried

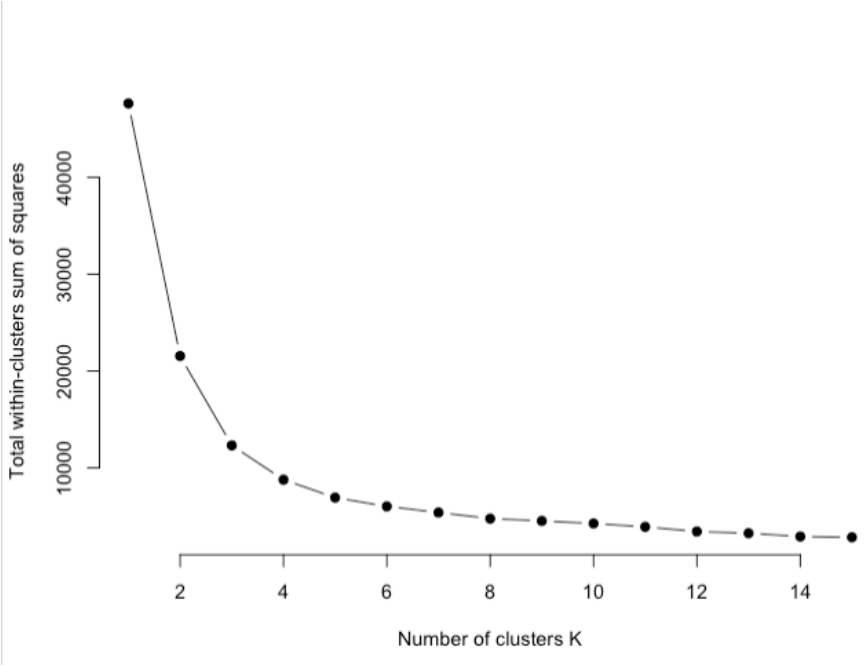
Rings / integer / -- / +1.5 gives the age in years

2. Here we don't know the age of abalone. Using clustering methods, I'm trying to group the data. So that we can see if it giving age related clusters.

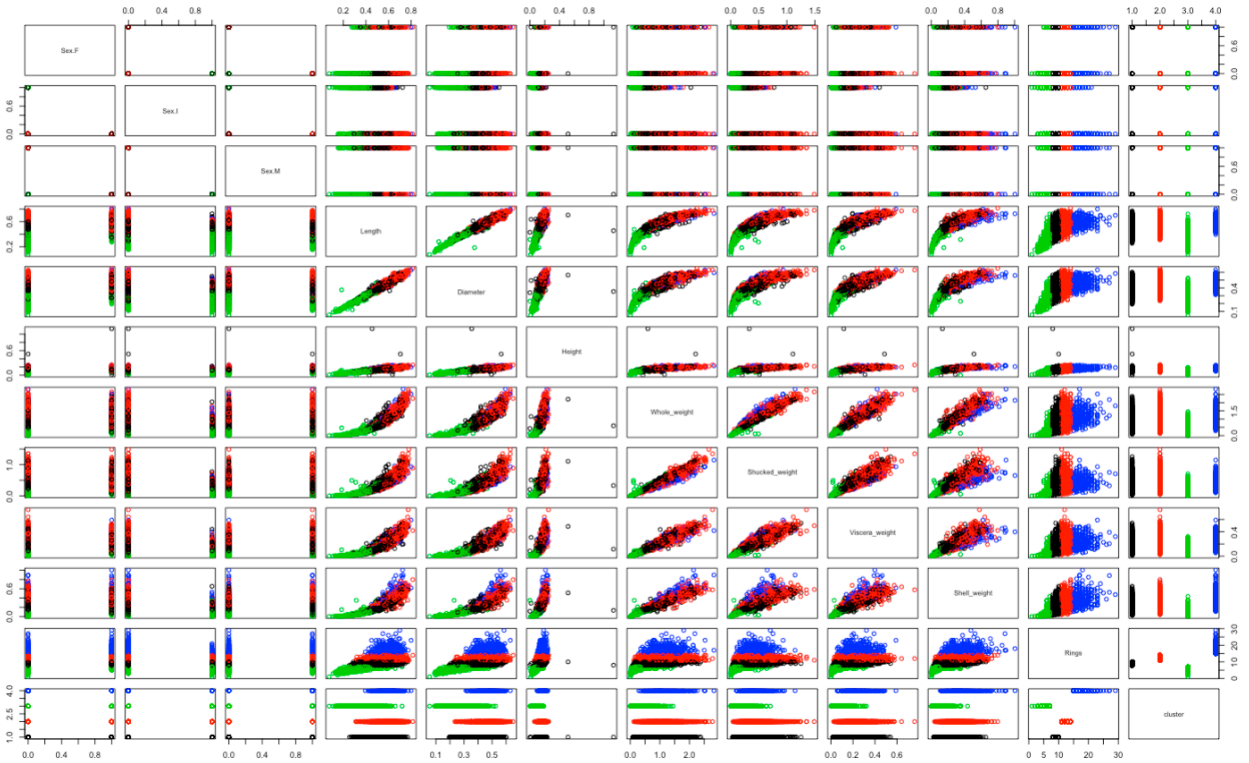
I'm performing clustering on this data using **these 3 methods** *k-means, k-medoids, hierarchical clustering.*

<u>K – means:</u>

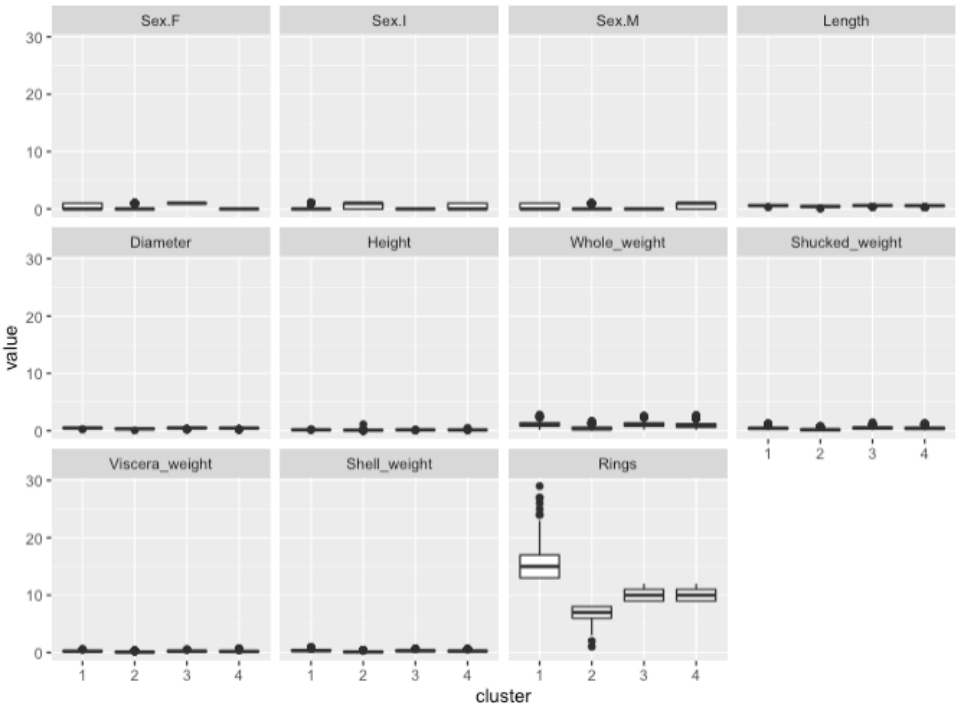
My data has factors column too. So, I'm applying one hot encoding to convert them.
From elbow method we can get to know how many clusters to make. From below fig we can do 4 or 5 clusters.



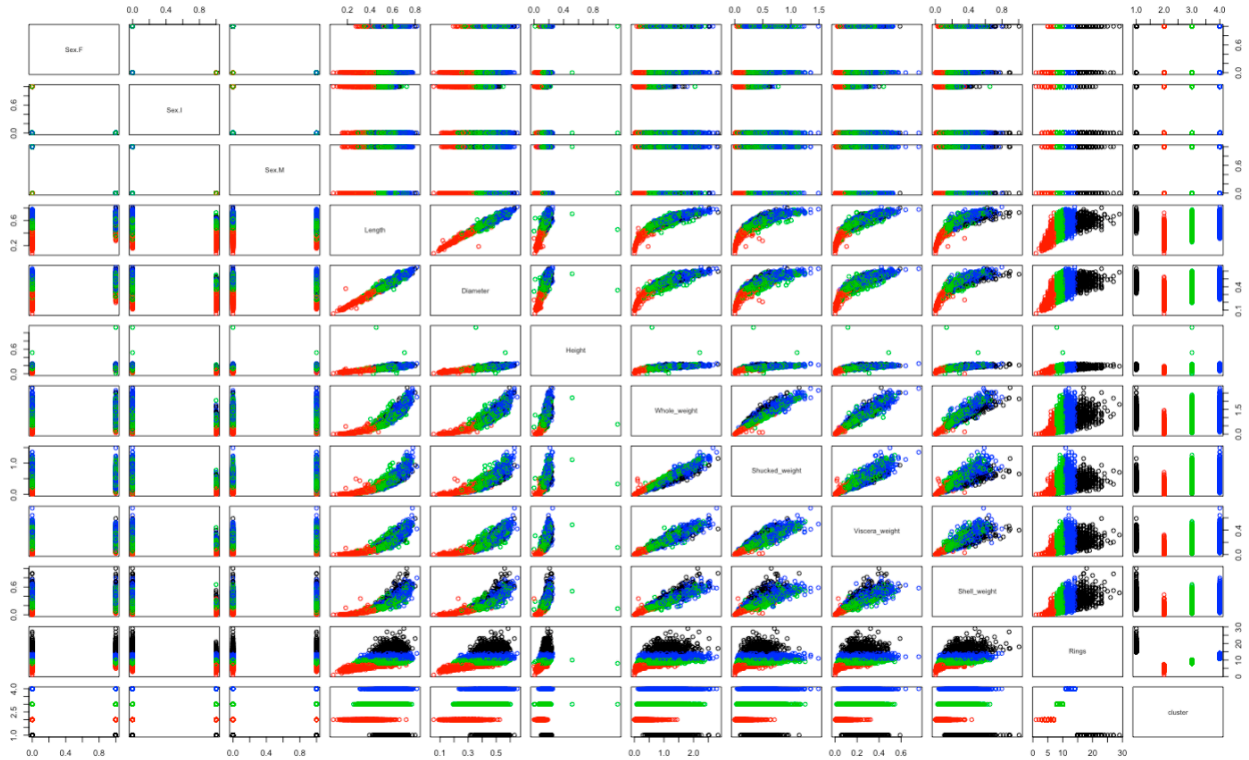

```
kmeansObj <- kmeans(x,4)
```



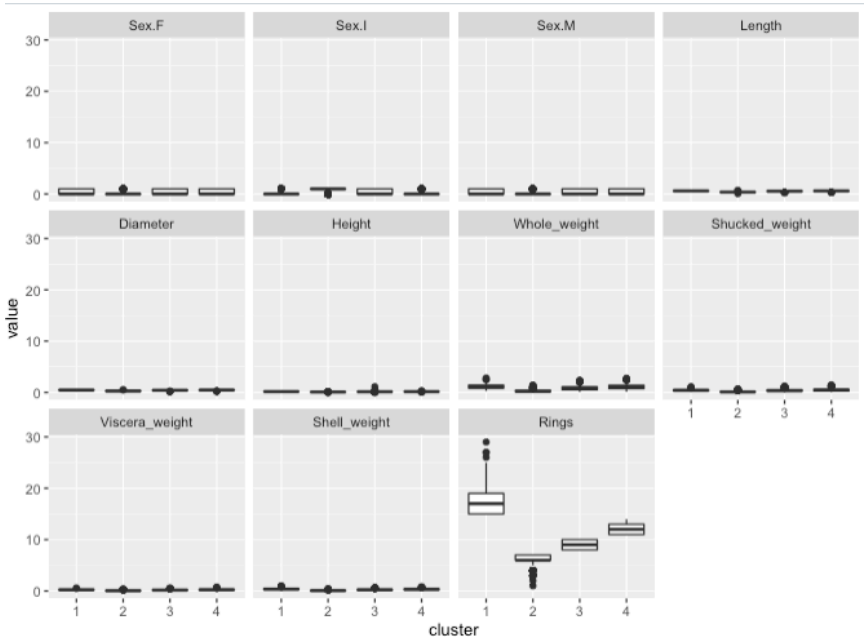
As we can see clusters are not so clear in other attributes except *Rings.* But still there is a lot of overlapping. We can see that clearly in below boxplots.



k-medoids:



We can see clusters, there is still a lot of overlapping even with k-medoids. But better than kmeans. As we can see from below plot we can see it is showing different for infant.

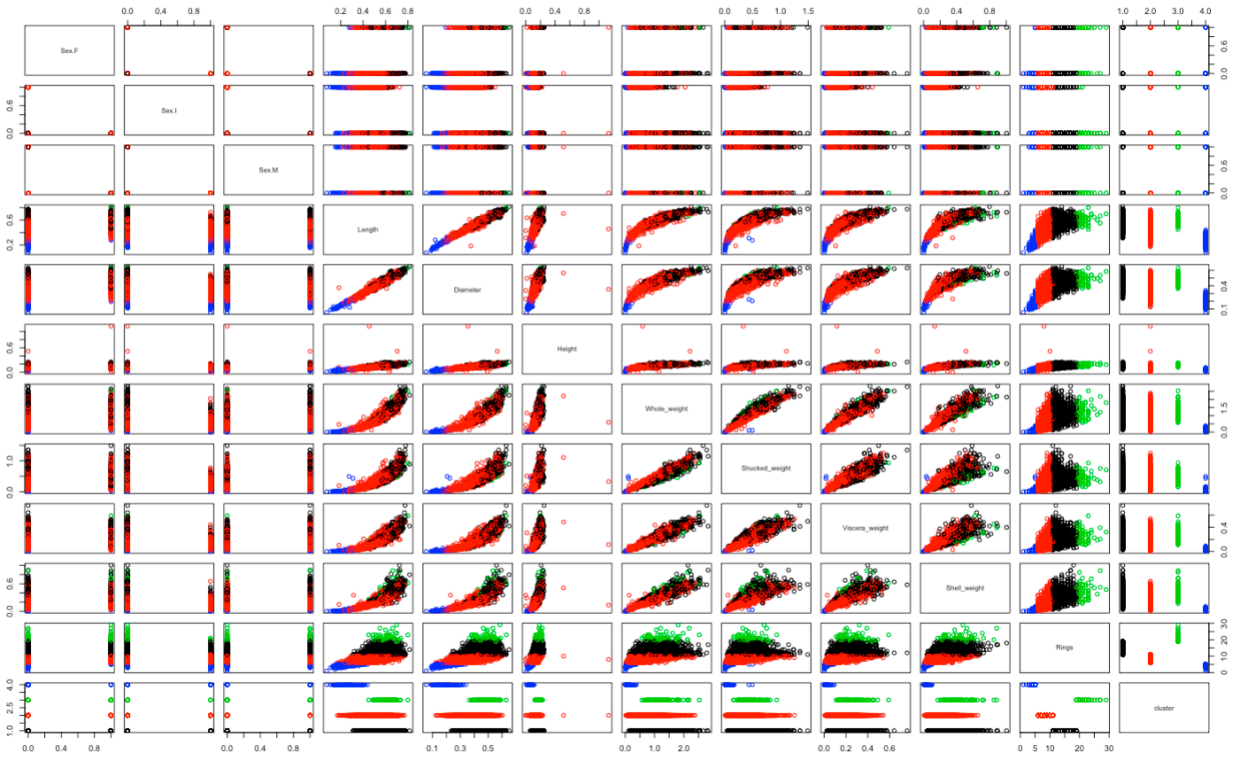


Hierarchical clustering:

Hierarchical clustering can be divided into two main types: agglomerative and divisive.

Note that agglomerative clustering is good at identifying small clusters. Divisive hierarchical clustering is good at identifying large clusters. I'm building agglomerative clustering.

```
HAgloclusters <- hclust(dist(x))

clusterCut <- cutree(HAgloclusters, 4)
```

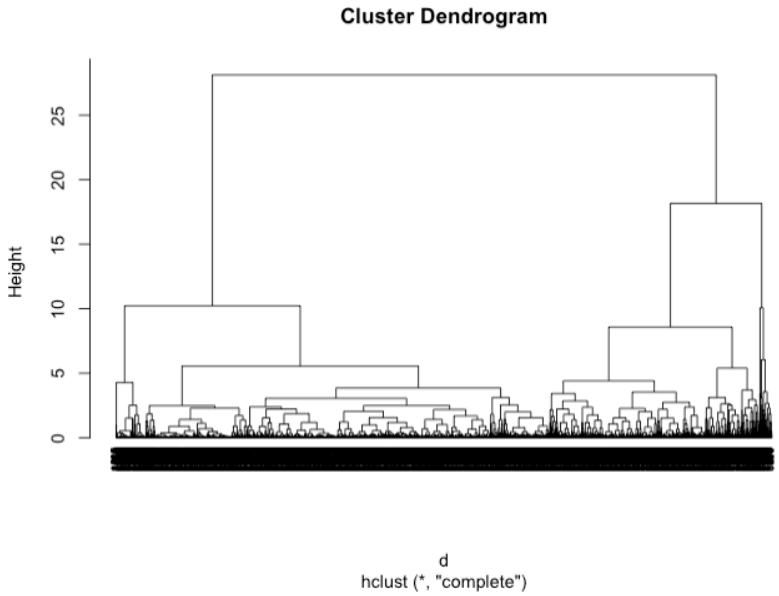


It is also giving the same results as k-medoids. As from above plot we could tell.

3. From above summarizing kmeans is not giving better results as our aim is to clusters the data to represent age. Only k-medoids and agglomerative giving better results by separating infant data along with rings.

4. More in detail about Hierarchical clustering. As I mentioned above I performed Agglomerative clustering: It's also known as AGNES (Agglomerative Nesting). It works in a bottom-up manner. That is, each object is initially considered as a single-element cluster (leaf). At each step of the algorithm, the two clusters that are the most similar are combined into a new bigger cluster (nodes). This procedure is iterated until all points are member of just one single big cluster (root) (see figure below). The result is a tree which can be plotted as a dendrogram.

```
# Dissimilarity matrix

d <- dist(x, method = "euclidean")

# Hierarchical clustering using Complete Linkage

hc1 <- hclust(d, method = "complete" )

clusterCut <- cutree(hc1, 3)

ggData$cluster <- as.factor(clusterCut)

pairs(ggData, col=ggData$cluster)

# Plot the obtained dendrogram

plot(hc1, cex = 0.6, hang = -1)
```



From the above section 2, we can see scatter plot of Hierarchical clustering where we choose cut tree at 4. Where we getting better result though it is not a great. From this we could tell they are differing more on rings and sex(adult and infant).