Problem Set 2

Jesús Pacheco

10/15/2019

Computation

1. Distances

```
p <- c(1,2)
q <- c(3,4)

d_manhattan <- sum(abs(p-q))
d_canberra <- sum(abs(p-q)/(abs(p) + abs(q)))
d_euclidean <- (sum((p-q)^2))^.5
c(d_manhattan, d_canberra, d_euclidean)
```

```
## [1] 4.0000000 0.8333333 2.8284271
```

2. Using the build-up function

```
dist(rbind(p, q), method = "manhattan")
```

```
##   p
## q 4
```

```
dist(rbind(p, q), method = "canberra")
```

```
##           p
## q 0.8333333
```

```
dist(rbind(p, q), method = "euclidean")
```

```
##          p
## q 2.828427
```

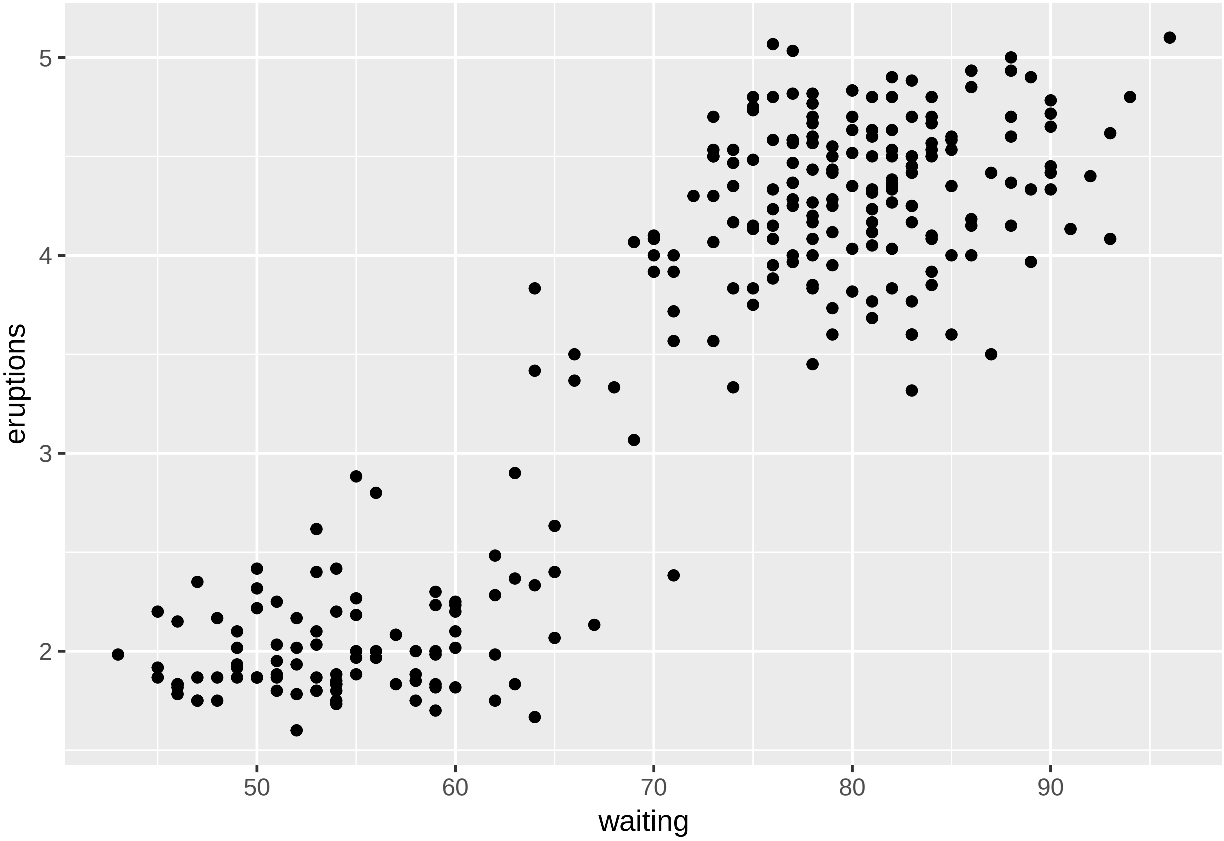
I was wrong at first at computing the Canberra distance, I was missing a parentheses to do the vector computations correctly. Using the build-up function helps with this.

3. Key differences

The three distances measure different relational attributes in the data, hence they are not comparable amongst them. The Manhattan distance sums over absolute values of all dimensions in which points are dissimilar to each other. Canberra does this in a standardized way, hence it is always smaller. The euclidean distance measures the shortest distance between these points (a straight line), hence it also produces a smaller value than the Manhattan distance.

4. Old Faithful EDA

```r
faithful %>% ggplot() + geom_point(aes(x=waiting, eruptions))
```



```r
faithful %>% summarize(Mean_waiting = mean(waiting), Mean_eruption = mean(eruptions),
                       Var_waiting = var(waiting), Var_eruption = var(eruptions),
                       corr = cor(waiting, eruptions))
```

```
##   Mean_waiting Mean_eruption Var_waiting Var_eruption      corr
## 1     70.89706      3.487783    184.8233     1.302728 0.9008112
```

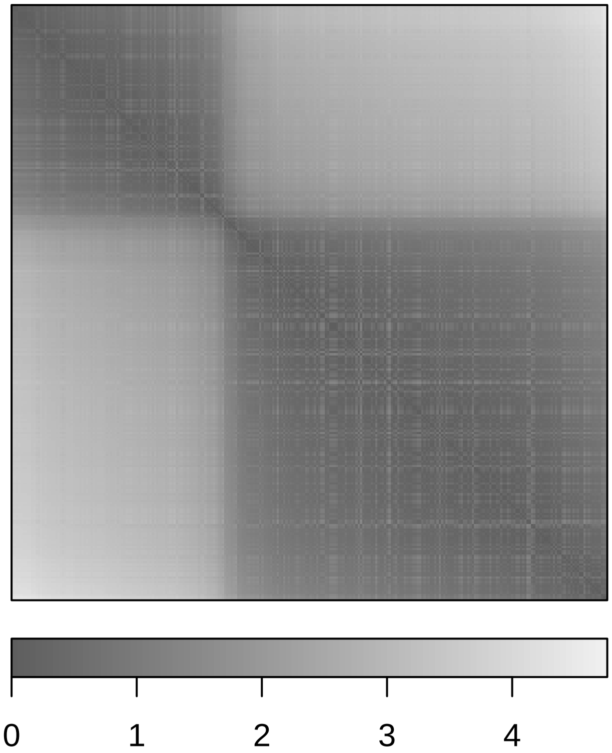
From the visualization, we see that there might be some positive relation between waiting and eruption times in the Old Faithful geyser. We can also see that in the correlation coefficient between the two variables. I also see that there might be potential clusters since we see two “groups” of density of observations: one in the lower left side (less than 70 minutes of waiting time and less than 3 minutes of eruptions) and the other one in the upper right. Furthermore, we see more variation in the data in the right-hand side of the graph. We can already start to guess that there might be two types of relations of eruption durations and waiting times.

5. Dissimilarity matrix

```r
#Scaling the data
faithful_scaled <- scale(faithful)
dis_mat <- dist(faithful_scaled, method = "euclidean")
```

6. ODI

```
dissplot(dis_mat)
```



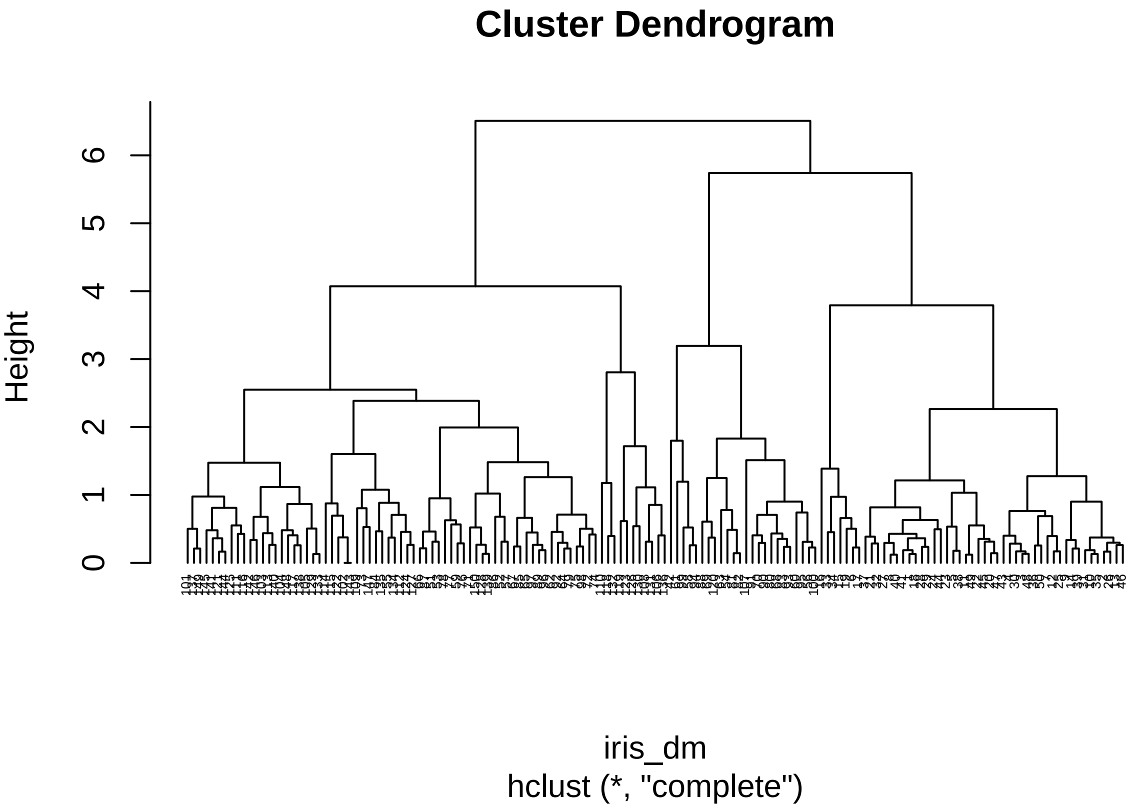
The ODI clearly displays two types of relations in the data. The darker blocks along the diagonal line suggest that there might be a good case for clusters in the data. In other words, the dissimilarities are smaller in certain parts of the distribution.

7. The iris data

```
iris_dm <- iris %>% select(-Species) %>% scale() %>% dist( ,method="euclidean")
```

8. Hierarchical clustering

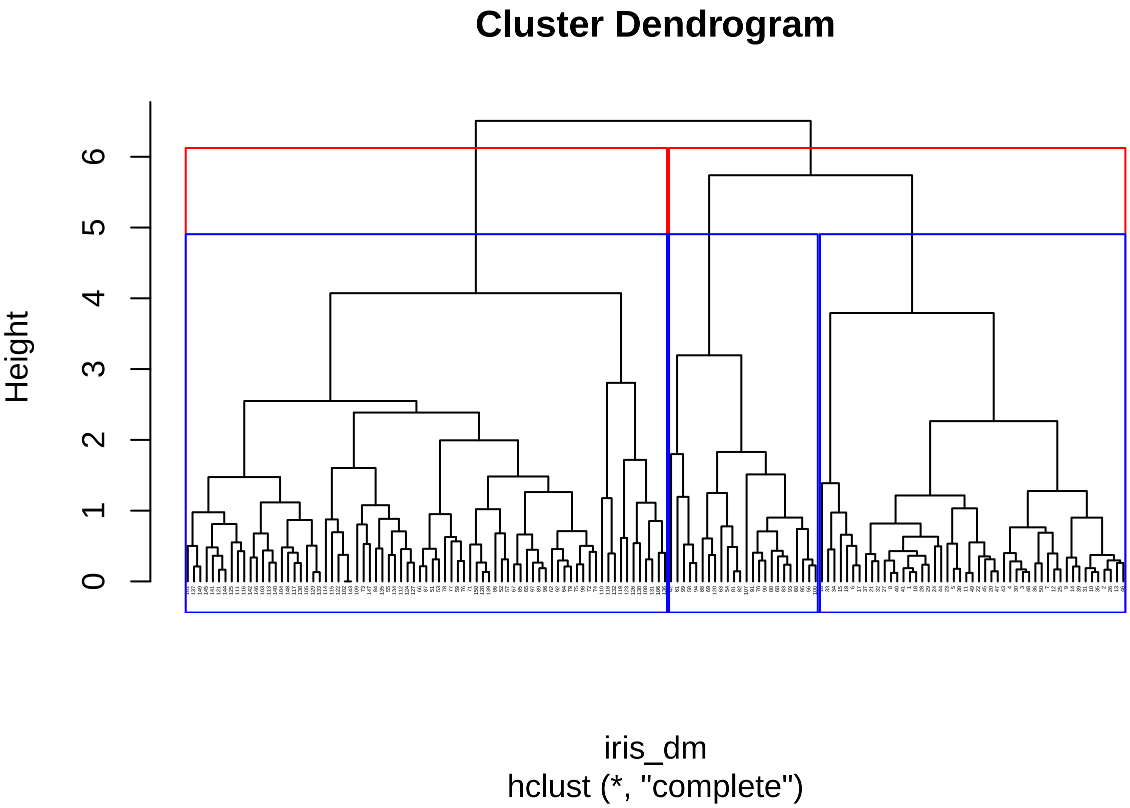
```
hc_complete <- hclust(iris_dm, method = "complete"); plot(hc_complete, hang = -1, cex=.4)
```



From this dendrogram, we can see three branches at a pretty simliar height as expected since we know there are three species. Taking a closer look, we see the number of rows in the 100s clustered in the left (virginica) and the lower numbers clustered to the right (setosa) but the second 50 numbers are not very close together in the dendrogram. Hence, we could see that under this cluster procedure, virginica and setosa species are more similar within each of them than the versicolor specie. Overall, we see also that the clustering was not done perfectly.

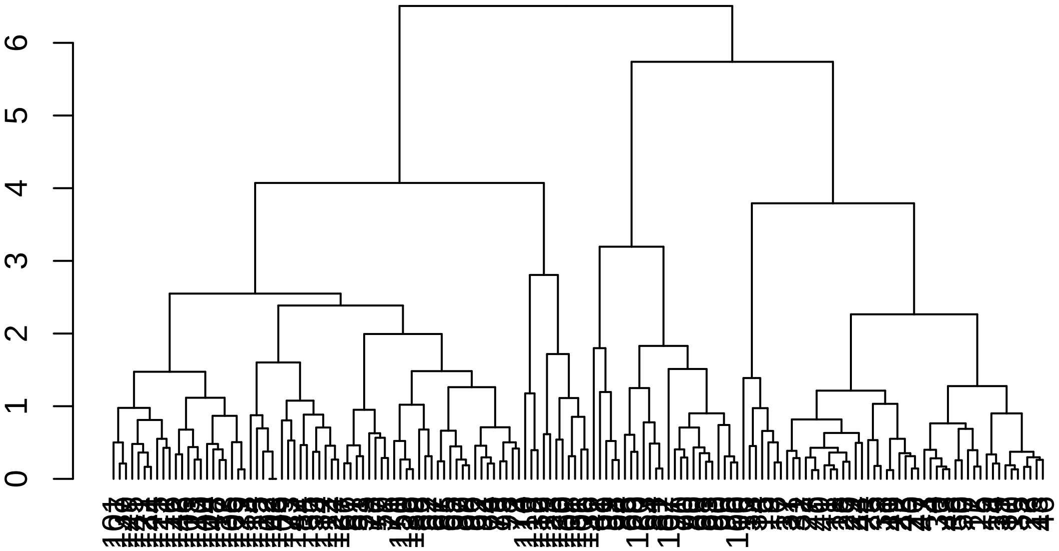
9. Cutting the tree

```
hc_complete <- hclust(iris_dm, method = "complete"); plot(hc_complete, hang = -1, cex=.2)
rect.hclust(hc_complete, k=2, border="red")
rect.hclust(hc_complete, k=3, border="blue")
```



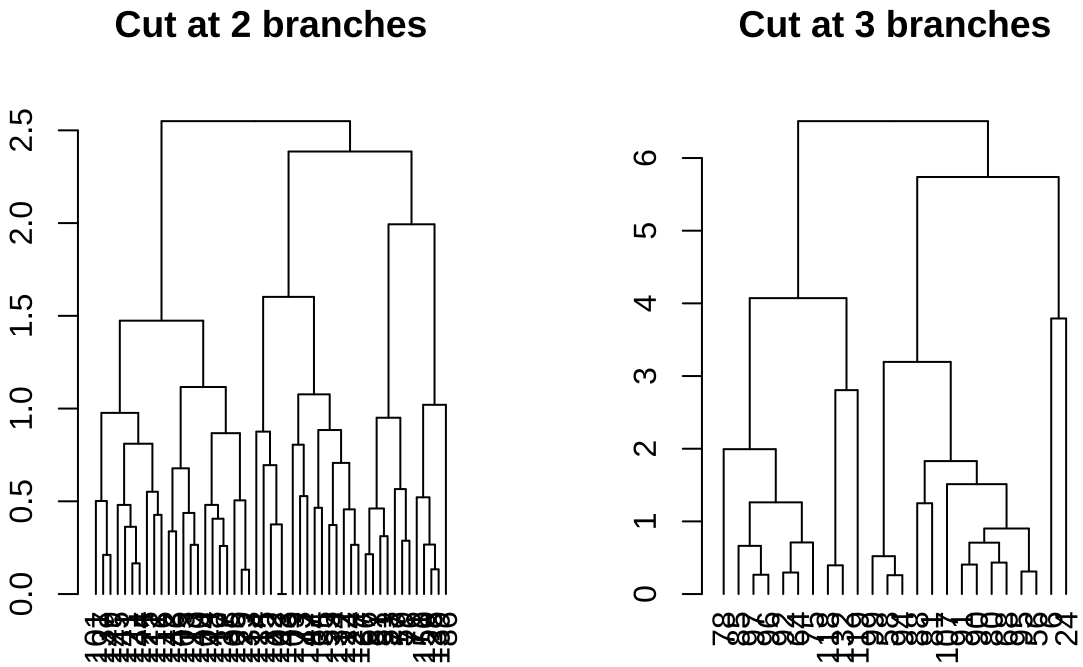
The red rectangle show the tree with two clusters and the blue rectangles show three clusters. We see that the two-cluster division basically divided the observations right in the middle, we see a mixture of species if we take only two clusters, specially the observations in the right seem quite dissimilar from looking at the whole tree. The three-clustered tree seem a little more homogenous. Here is how the individual trees look like:

```
k <- 2
dend <- as.dendrogram(hc_complete)
plot(dend)
```



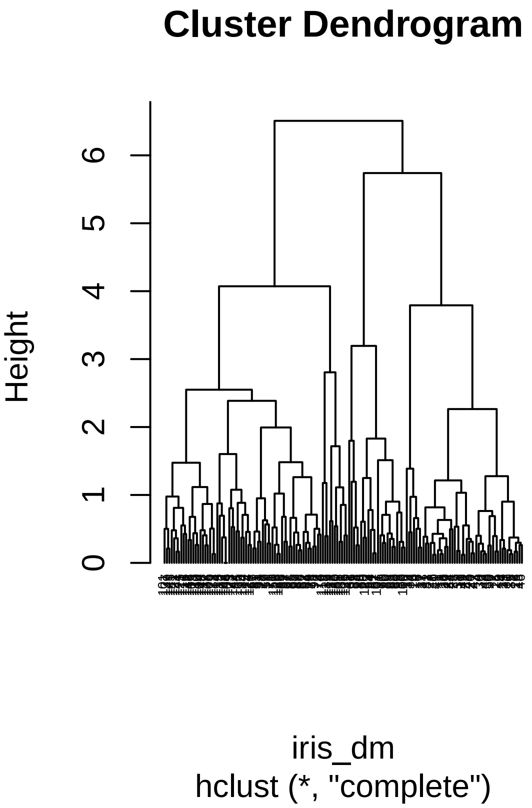
```
labels_dend <- labels(dend)
groups <- cutree(dend, k=3)
dends <- list()
```

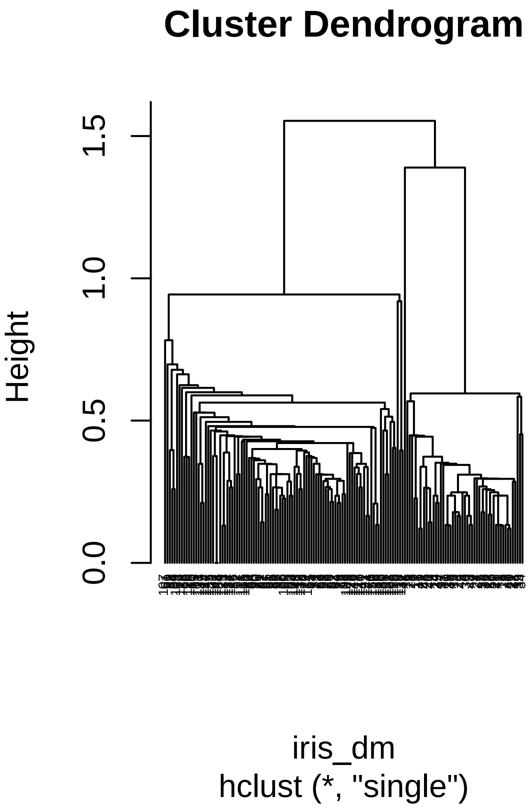
```
for(i in 1:k) {
  labels_to_keep <- labels_dend[i != groups]
  dends[[i]] <- prune(dend, labels_to_keep)
}
par(mfrow = c(1,2))
for(i in 1:k) {
  plot(dends[[i]],
       main = paste0("Cut at ", i+1, " branches"))
}
```



10. Different cluster linkages

```
par(mfrow = c(1,2))
hc_complete <- hclust(iris_dm, method = "complete"); plot(hc_complete, hang=-1, cex=.4)
hc_single <- hclust(iris_dm, method = "single"); plot(hc_single, hang=-1, cex = .4)
```





We can see that the two dendrograms are different from each other. The branches in the dendrogram using a complete linkage method display something closer to three clusters in the observations but the observations are not ordered in a way that match the actual species. The dendrogram with the single linkage method has the observations ordered in a way that's closer to the actual species classification but the branches are completely different, we don't find the three-cluster pattern with this linkage method. As discussed in class, agglomeration hierarchical clustering produces significantly diverse results depending on the linkage method. In this case we can imagine how it can be very different to match one cluster to another by looking at the minimal euclidean distance between and to match them by looking at the maximal distance; the clusters that are tied up together are differing by a lot.

Critical thinking

1. How would you go about determining whether clustering made sense to consider or not?

What are techniques you would use, and what might you be looking for from each? How might these techniques work together to motivate clustering or not? And ultimately, can/should you proceed if you find little to no support for clusterability? Why or why not?

And ultimately, can/should you proceed if you find little to no support for clusterability? Why or why not? First, I would use some informal EDA with visualizations and ODI that might help me understand the data better, I could calculate mathematically the sparsity with a Hopkins statistic to see if I can reject the hypothesis of a distribution that is completely random. After this, even if the clusterability is not clear, I can draw a dendrogram to try to find patterns.

These two techniques might work together either by confirming our "suspicions" or can complete each other (for example if the similarities on the observations are due to bigger set of variables than what we can see on a twoway plot.

Hierarchical clustering might make sense, but it's not the only way, as I understand, we can still find other types of clustering in the data. So, if we still have some motivation why the data might be clusterable, we should keep trying to find useful relations.

2. A paper that applies the hierarchical agglomerative clustering technique.

Borgen, F. H. & Barnett, D. C. (1987). *Applying cluster analysis in counselling psychology research.* Journal of Counseling Psychology, 34(4), 456-468.

The paper suggests the use of clustering techniques to group entering freshmen students who have not yet declared a major to work together. Based on their preferences, academic environment and vocational results from psychological tests, they start by looking at the data and doing some informal EDA. They discuss the choice of the clustering and distance method. They also show some summary statistics about the final results to see what are the characteristics of the groups formed. There's not much for the algorithm fitting but I think it was a clear and easy example of an application of clustering as well as a thorough process for exemplifying the process. The article shows examples of the usage of hierarchical clustering analysis in psychology research, the use of this technique can be further evaluated to see how much students might benefit from interacting with peers with similar interests and if it would be more beneficial than more diverse groups in terms of interests.