Unsupervised Learning: PSet 2

Felipe Alamos

10/19/2019

Computation

1

Calculate Manahattan, Canberra, and Euclidean distances "by hand" (i.e., create the data, program each line, and make the calculations). What are the values for each measure?

```
library(knitr)
library(skimr)
```

```
##
## Attaching package: 'skimr'

## The following object is masked from 'package:knitr':
##
##     kable

## The following object is masked from 'package:stats':
##
##     filter
```

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

manhattam <- abs(p[1]-q[1])+abs(p[2]-q[2])
canberra <- abs(p[1]-q[1])/(abs(p[1])+abs(q[1]))+abs(p[2]-q[2])/(abs(p[2])+abs(q[2]))
euclidean <- sqrt((p[1]-q[1])^2+(p[2]-q[2])^2)

df <- data.frame(manhattam, canberra, euclidean)

#Specify number of digits to show
options(digits=3)
kable(df,caption="Manhattan, Canberra and Euclidean distances of 2 vectors")
```

manhattam	canberra	euclidean
4	0.833	2.83

2

Use the dist() function in R to check your work. Were you right or wrong? (be honest in your reporting) If wrong, after debugging, where and why did you go wrong?

```
dist(rbind(p, q))
```

```
##      p
## q 2.83
```

We observe that the distance that we had calculated by hand was right.

3

What are the key differences between these measures, and why does it matter? How might you see these differences "in action" with these fictitious data?

The key difference is the way the measures are calculated. Manhattan and Canberra use absolute values, whereas euclidean uses exponentiation.

The Manhattan distance represents the sum of the lengths on each axis that separate two points. Using absolute values instead of exponentiation, it could be considered to be more robust to outliers.[^http://rstudio-pubs-static.s3.amazonaws.com/476168_58516a3d6685427badf52a263e690975.html] If we would like to weight each of these lengths by the value of the coordinates of each point, we could use Canberra. Canberra distance is useful in domains such as measure of disarray for ranked lists, where rank differences in top positions need to pay higher penalties than movements in the bottom part of the lists[^https://pdfs.semanticscholar.org/a298/02f53f3bcf348e2c9a95df748277ae7571c5.pdf].

On the other hand, the euclidean distance is not the addition of lengths on each axis, but rather measures a "straight line" between the points.

This differences between these measures are important because the appropriate distance measure should be used depending on the context. Example, if we are trying to model the distance between two houses, but there is no straight road between them, a Manhattan distance would be more appropriate than euclidean.

Naturally, Manhattan distances will be bigger than Canberra (because the latter each length is weighted) and euclidean (because this one is a straight line). This can be observed in our data for the two vectors example.

4

Use some basic EDA techniques to present and discuss the data (e.g., visualize, describe in multiple ways, etc.)

```r
#Tried to use skim but Knit is thorwing error
nrow(faithful)
```

```
## [1] 272
```

```r
ncol(faithful)
```

```
## [1] 2
```

```r
summary(faithful)
```

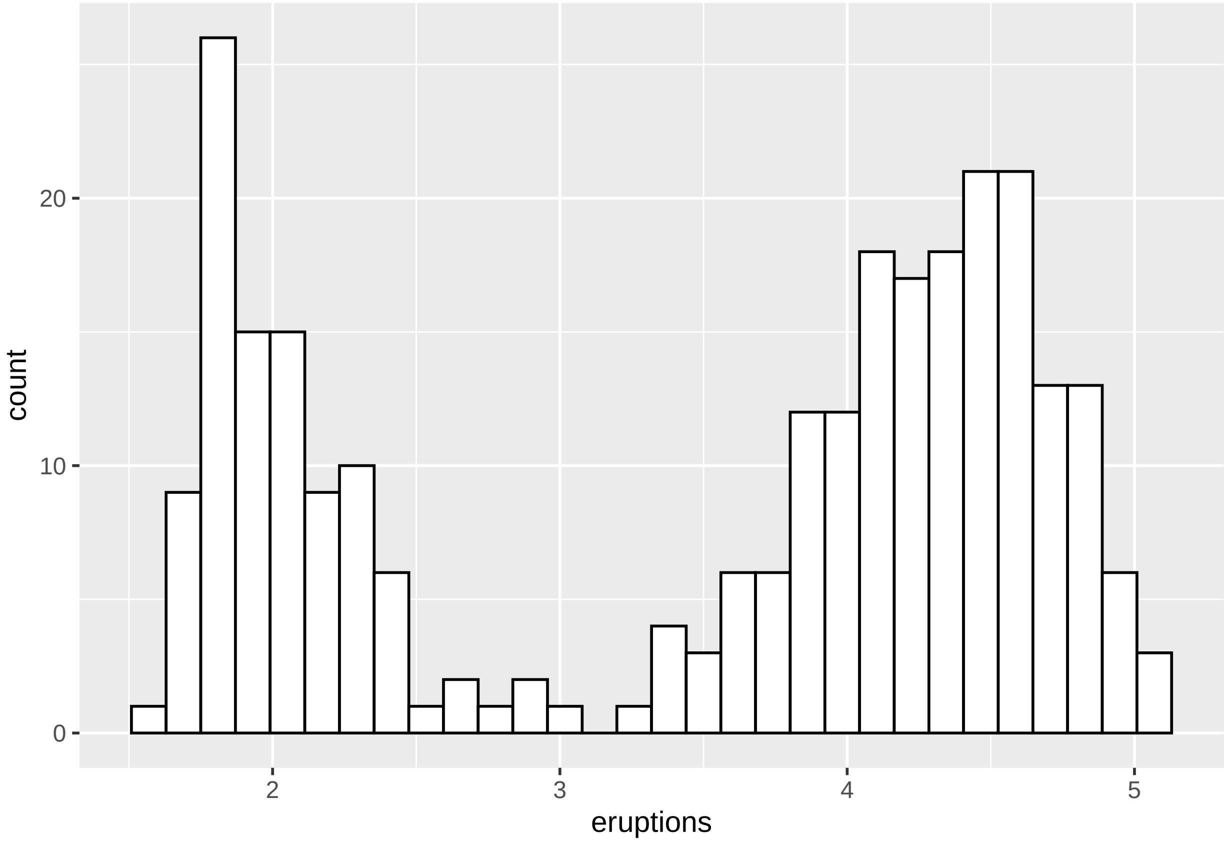
```
##    eruptions        waiting    
##  Min.   :1.60   Min.   :43.0  
##  1st Qu.:2.16   1st Qu.:58.0  
##  Median :4.00   Median :76.0  
##  Mean   :3.49   Mean   :70.9  
##  3rd Qu.:4.45   3rd Qu.:82.0  
##  Max.   :5.10   Max.   :96.0
```

With summary we can observe basic statistics of this dataset. First of all, we have 272 observations, with 2 columns each: eruptions and waitings. The table presents mean, standard deviations and quartiles for each of them.

We present basic histograms for each variable so as to have a clearer idea of their distribution. In addition, we present a scatterplot so as to understand how the two variables relate to each other.

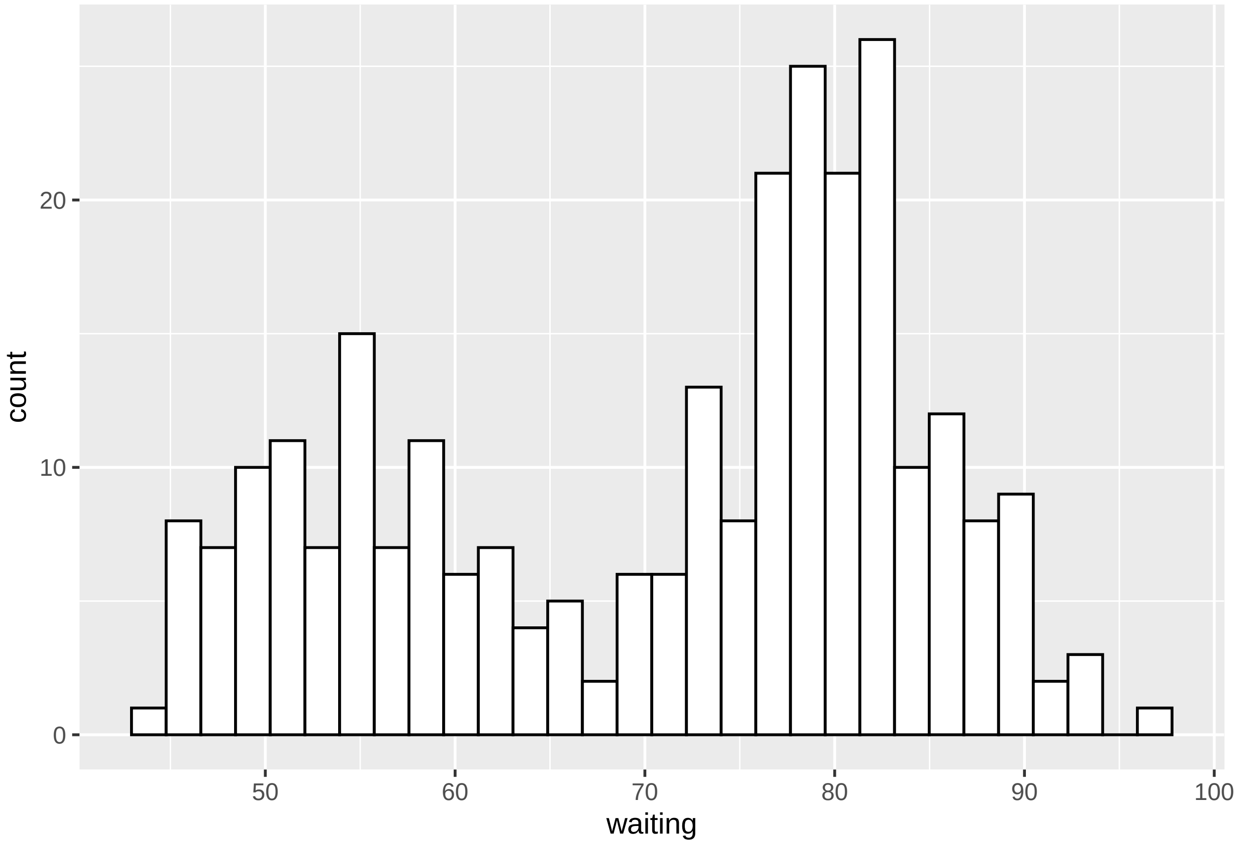
```r
library(ggplot2)

ggplot(faithful, aes(x=eruptions)) +
  geom_histogram(color="black", fill="white")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

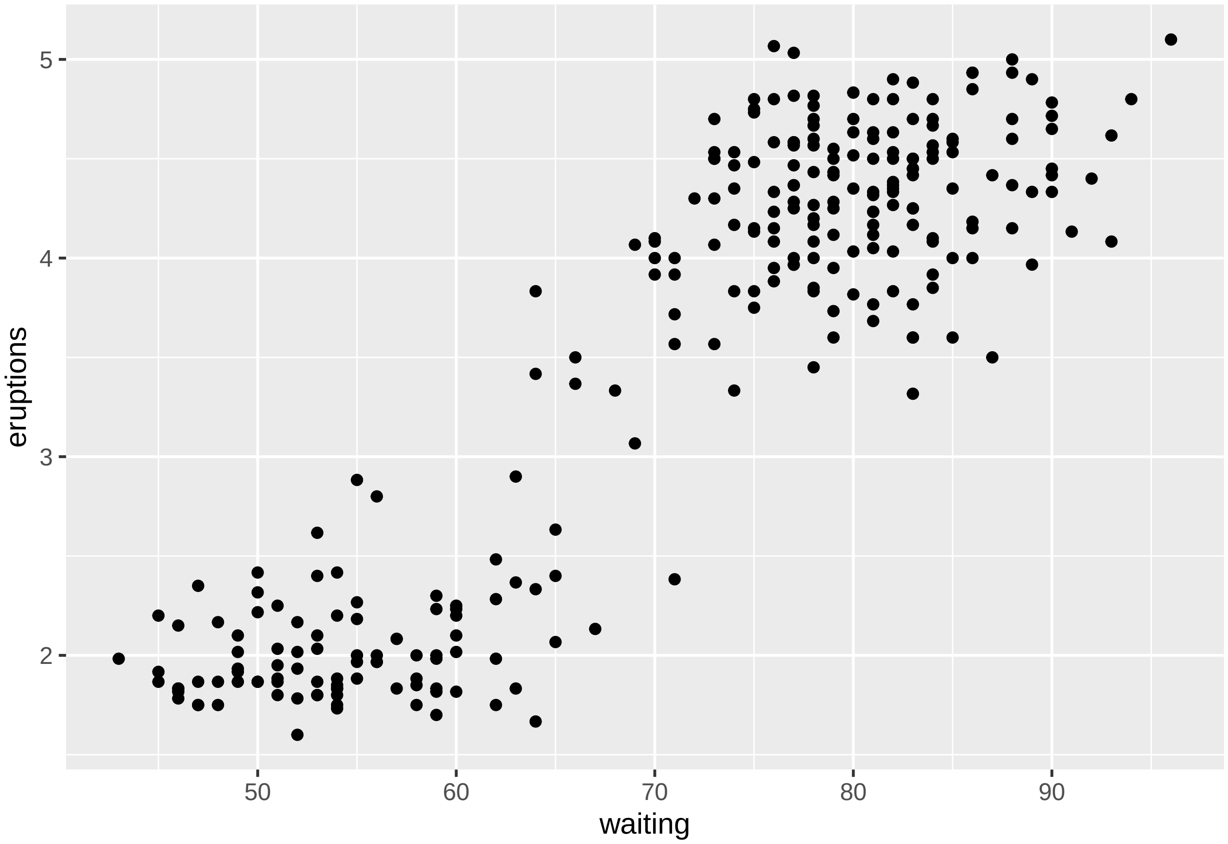


```r
ggplot(faithful, aes(x=waiting)) +
  geom_histogram(color="black", fill="white")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(faithful, aes(x=waiting, y=eruptions)) +
  geom_point()
```



5

Calculate a dissimilarity matrix of these data

Because of the dimension of the dissimilarity matrix, we will display only distances for the first 5 points.

```
library(cluster)

diss <- daisy(faithful, metric = "euclidean", stand = FALSE) #Equivalent to scaling and calculating dis
diss_matrix <- as.matrix(diss)
diss_matrix[0:5,0:5]
```

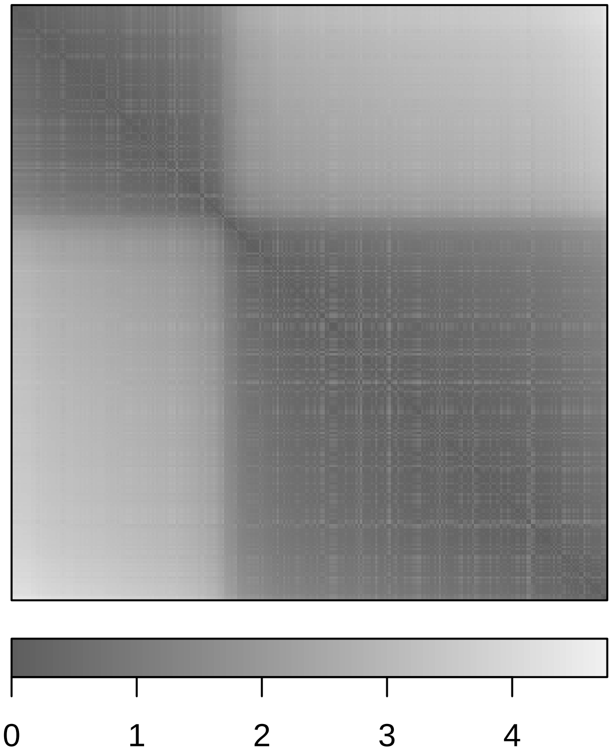
```
##       1     2     3     4     5
## 1  0.00 25.06  5.01 17.05  6.07
## 2 25.06  0.00 20.06  8.01 31.12
## 3  5.01 20.06  0.00 12.05 11.07
## 4 17.05  8.01 12.05  0.00 23.11
## 5  6.07 31.12 11.07 23.11  0.00
```

6

Generate an ODI for the Old Faithful data. What do you see?

```
library("seriation")

#We start by scaling the data using the function scale().
```

```r
df_scaled <- scale(faithful)
#Next we compute the dissimilarity matrix between observations using the function dist().
df_dist<- dist(df_scaled)
#Finally the function dissplot() [in the package seriation] is used to display an ordered dissimilarity
dissplot(df_dist)
```



[^http://www.sthda.com/english/wiki/print.php?id=238]

We observe a very clear pattern: there seem to be two groups or clusters, one bigger than the other. This can be observed from the image because of the dark squares, which represent points that are close to each other in distance.

Download the Iris data set we used in class (e.g., data(iris) in R), and use to address questions 7-10: ## 7 Using any munging tools you'd like (e.g., dplyr from the Tidyverse), create a subset of the data excluding the species feature, scaling the features, and calculating a dissimilarity matrix (think "pipe" for stacking functions to do this quickly, e.g.).

```r
library(dplyr)
```

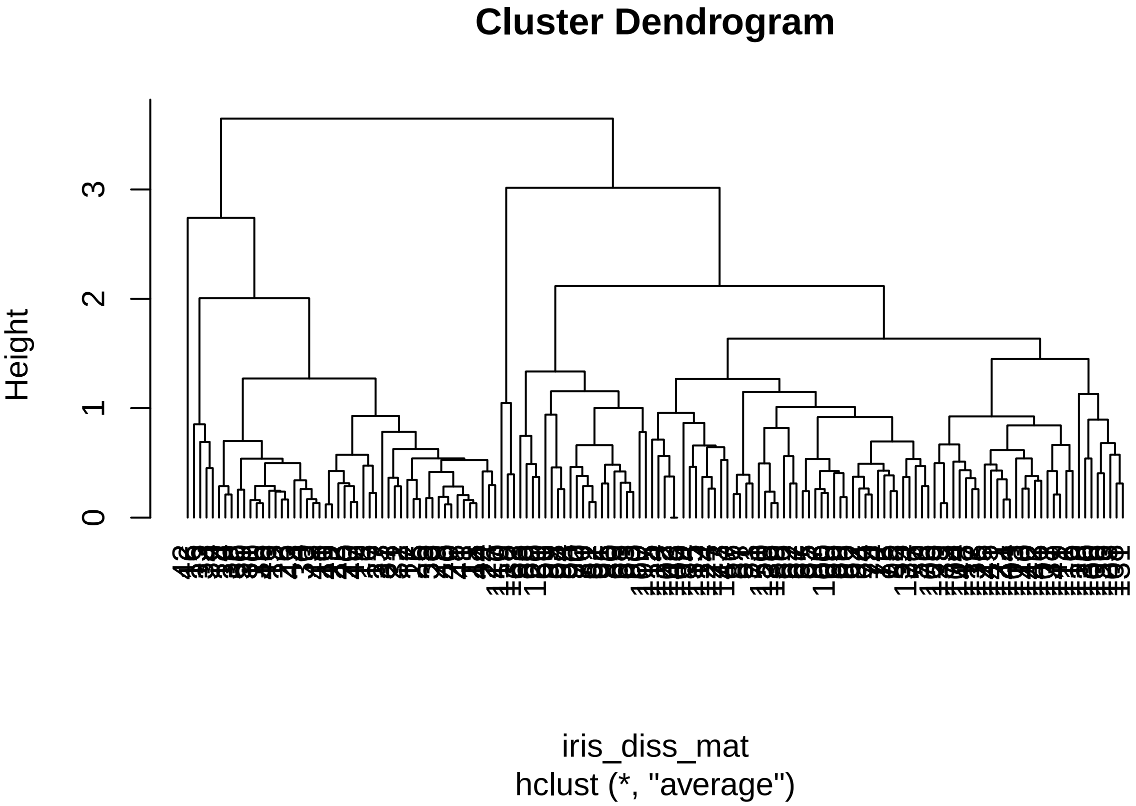
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
iris_diss_mat <- iris %>%
  select(-Species) %>%
  scale() %>%
  dist()
```

8

Fit an agglomerative hierarchical clustering algorithm using complete linkage on your subset data and render the dendrogram of clustering results. What do you see?

```r
hc_complete <- hclust(iris_diss_mat,
                      method = "average"); plot(hc_complete, hang = -1)
```



Observations: * When clustering in two groups, the right one seems to be bigger than the left one * The observations on the farthest left seem to be quite unique - distanced from other samples, and hence join clusters at higher levels. In particular, the first sample remains as a unique cluster almost until the end.

9.

Try cutting the tree at 2 and 3 branches and show these trees side-by-side. How do they differ?

```r
cuts <- cutree(hc_complete,
               k = c(2,3))

table(`2 Clusters` = cuts[,1],
      `3 Clusters` = cuts[,2])
```

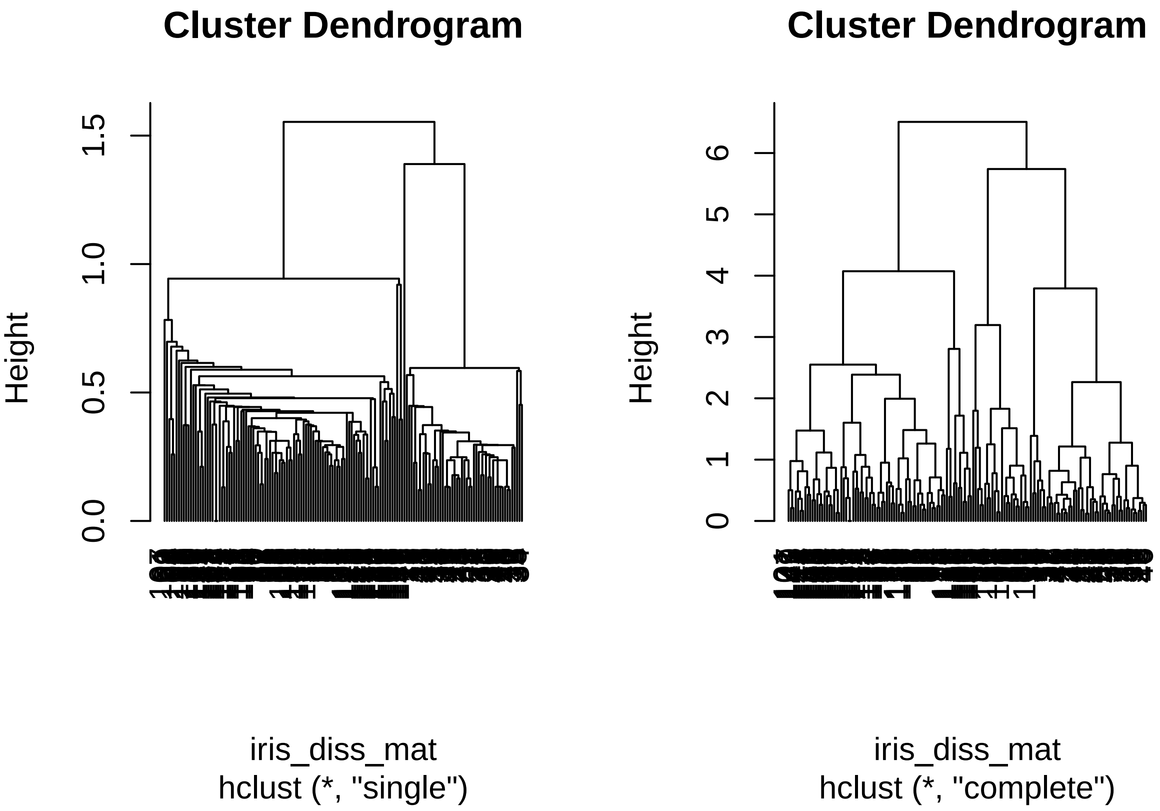
```
##             3 Clusters
## 2 Clusters   1  2  3
##          1 50  0  0
##          2  0 97  3
```

We observe that, in case of clustering in 2 groups, one group contains 50 members and the other 100. When clustering in 3 groups, the only change respect the previous clustering is that now 3 observations of the second cluster form a third cluster.

10

Now fit the algorithm using single and complete linkage and present each dendrogram side-by-side. Discuss the differences. What effects can we see in the clustering patterns when using different linkage methods?

```
par(mfrow = c(1,2))
hc_single <- hclust(iris_diss_mat,
                    method = "single"); plot(hc_single, hang = -1)

hc_complete <- hclust(iris_diss_mat,
                      method = "complete"); plot(hc_complete, hang = -1)
```



It seems that complete linkage generates clusters that are more evenly distributed bottom up (i.e. it doesnt happen that in higer levels of aggregation some groups have only a few members).

We can also perceive how the differences in distance between the groups (measured by the height) is more distinguishable on the complete linkage.

Critical Thinking

1.

You just assessed the clusterability of some feature space, R. Address the following questions:

a.

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

I would try to do some exploration of the feature space to identify if some cluster detection is feasible.

b.

What are techniques you would use, and what might you be looking for from each?

- Informal diagnosis: Simple distribution plots. I would look do some scatter plots with main features so as to see if those plots reveal certain patterns or clusters. I would use this technique if I had some intuition in advance of which features could be the most important when clustering.
- Visually (VAT/ODI plots). This technique would give us a better visual representation of distances between the elements considering all features, and not only some arbitrarly chosen. It would give us a sense of spatial similarity/dissimilarity of the observations.
- Mathematically: sparse sampling. Particularly using a Hopkins statistic test, which calculates the probability that a given dataset is generated by a uniform (with no clusters) distribution or not (non-random, with clustering likely). With this technique we try to reject the null hypothesis which is that the data is uniformly distributed (so, we want to mathematically show that it has clusters).

c.

How might these techniques work together to motivate clustering or not?

These techniques reveal different things from the data feature space, so it is a great idea to combine them. That said, it is important to start using the one that is more appropiate according to our knowledge of the data. For example, if we know in advance that some clusterability is possible, we could try some informal diagnosis. If, on the other hand, we know little about the data, ODI plots or Hopkins test can be useful as a starting point of the diagnosis.

d.

And ultimately, can/should you proceed if you find little to no support for clusterability? Why or why not?

I believe that, in general, the costs of fitting a clustering algorithm and studying the possible clusters is not high (subject to data dimension). Hence, we should proceed with this even if the clusterability analysis did not show much evidence of clusters. In particular, there might be some hidden combination of features in the data that do create clusters. In that sense, my opinion is that a positive result in clusterability diagnosis is a great proxy to start training our cluster algorithms, but a negative result in the diagnosis does not inform as that clusters do NOT exist, it just does not add information. In general, I think trying to cluster our data does not hurt, it might help explore, we might possible learn something new, and we might at least give stronger evidence of our beliefs of no clusters.

2. Locate (and read) a paper that applies the hierarchical agglomerative clustering technique.

Address the following questions:

Paper: Identifying National Types: A Cluster Analysis of Politics, Economics, and Conflict [^https://www.jstor.org/stable/4149616?seq=1#metadata_info_tab_contents]

a.

Describe the author(s) process

The authors use the Dataset on National Attributes to collect variety of data of different countries through time. Their aim is to identify clusters/groups of countries that are similar to each other in different moments of time. In other words, what countries are similar to each other considering a variety of features.

They specifically use features related to economics, politics and conflicts. Their premise is that these variables show bidirectional and interactive effects upon each other, so they would be good proxys to identify countries that are similar to each other.

b.

Do they go through similar steps as we covered this week both in setting the stage for clustering (e.g., assessing clusterability, calculating distance, etc.), as well as in fitting the algorithm? If not, what did they omit and does this omission impact their findings in your opinion?

The authors did not go through assessing clusterability: they relied on their intuition of countries being similar to each other based on economical, political and conflict variables, and wanted to prove this through the existence of the clusters.

They did do the other similar steps we discussed in class: they define a clustering technique (agglomeration by euclidean distance), they discuss the use of single and complete linkage, and they debate over the appropriate distance measure. On this latter point, they decide to use euclidean distances because they believe that grouping of countries should be based on a great deal of similarity across all variables, and that distinctions should be formed based on outliers (Euclidean is great for this in particular). Authors also fit the algorithms and showed dendograms of the clusters.

I do not think that omitting the "assessing clusterability" affected their findings, mainly because their intuitions that the selected features were good to create clusters was correct. That said, I would have recommended to use, for example, some informal diagnosis through simple distribution plots, so as to improve their intuitions and give more ground to their initial thoughts/hypothesis.

c

Describe at least one possible extension from the study that could emerge based on their findings

The study uses data from 5 different years - 1967, 1974, 1981, 1988, and 1995 - and clusters countries for each of them. They justify this approach because averaging data across time could, for example, hide countries dynamics and evolutions.

Although I agree with their point of view, I still think that having aggregated data thorugh time and clustering based on that could have been useful and valuable to the study. I stand this based on the idea that differences across countries at the same time should be generally greater than those within a country over time. Clustering countries thorugh time could give another informative insight on which countries are, in time and in average, similar to each other. Using that general clusters map and the per year clusters could also be useful to understand which general clusters remain through time and which do not.