Problem Set #2

Pedro Alberto Arroyo

10/17/2019

Computation

Questions 1-3

You fielded a survey and collected some wildly descriptive feature vectors. Use the following vectors to address questions 1-3:

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?

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?

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?

```
#Data
v1 <- c(1,2)
v2 <- c(3,4)
mx <- matrix(c(v1,v2), byrow=T, nrow=2)
mx
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    3    4
```

```
col_names_vectors <- c("x1", "x2")
colnames(mx) <- col_names_vectors
row_names_vectors <- c("p", "q")
rownames(mx) <- row_names_vectors

#Manhattan Distance by Hand
md <- abs(1-3)+abs(2-4)
md
```

```
## [1] 4
```

```
#Euclidean Distance by Hand
ed <- sqrt((1-3)^2+(2-4)^2)
ed
```

```
## [1] 2.828427
```

```
#Canberra Distance by Hand
cd <- (abs(1-3)/(abs(1)+abs(3)))+(abs(2-4)/(abs(2)+abs(4)))
cd
```

```
## [1] 0.8333333
```

```
#Manhattan w/Dist
dist(mx, method = "manhattan")
```

```
##   p
```

```
## q 4
```

```
#Euclidean w/Dist
dist(mx, method = "euclidean")
```

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

```
#Canberra w/Dist
dist(mx, method = "canberra")
```

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

The Manhattan distance is 4, the Euclidean distance is 2.82, and the Canberra distance is 0.833. Originally, my *by hand* numbers did not match the *dist()* numbers. It turned out to be due to a couple of syntax errors: an errant space between math operators, a missing paranthesis, etc.

We see that each of these approaches gives a different value for the distance between p and q, reflecting that *distance* is a general concept whereas *manhattan, euclidean, and canberra measues of distance* are operationalized deployments of that concept. They report different values because they capture different phenomena.

Euclidean distance is, for me, the easiset to interpret: its basic component is the shortest distance between two points and corresponds to our intuitive sense of proximity. When people give the distance bewtween two towns as *ten miles, as the crow flies*, they are talking about Euclidean distance.

Most of us don't fly to the next town over. Instead, we use transportation methods that are constrained by the local topography. **Manhattan distance** captures this intuition by requiring that paths between two points cover intermediate points.

This is reflected in the values we calculated in our mini-dataset: the Euclidean distance is shorter than the Manhattan distance because the Manhattan distance is constrained.

Canberra distance is a weighted version of *Manhattan.*

Old Faithful

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

5. Calculate a dissimilarity matrix of these data.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------ tidy
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

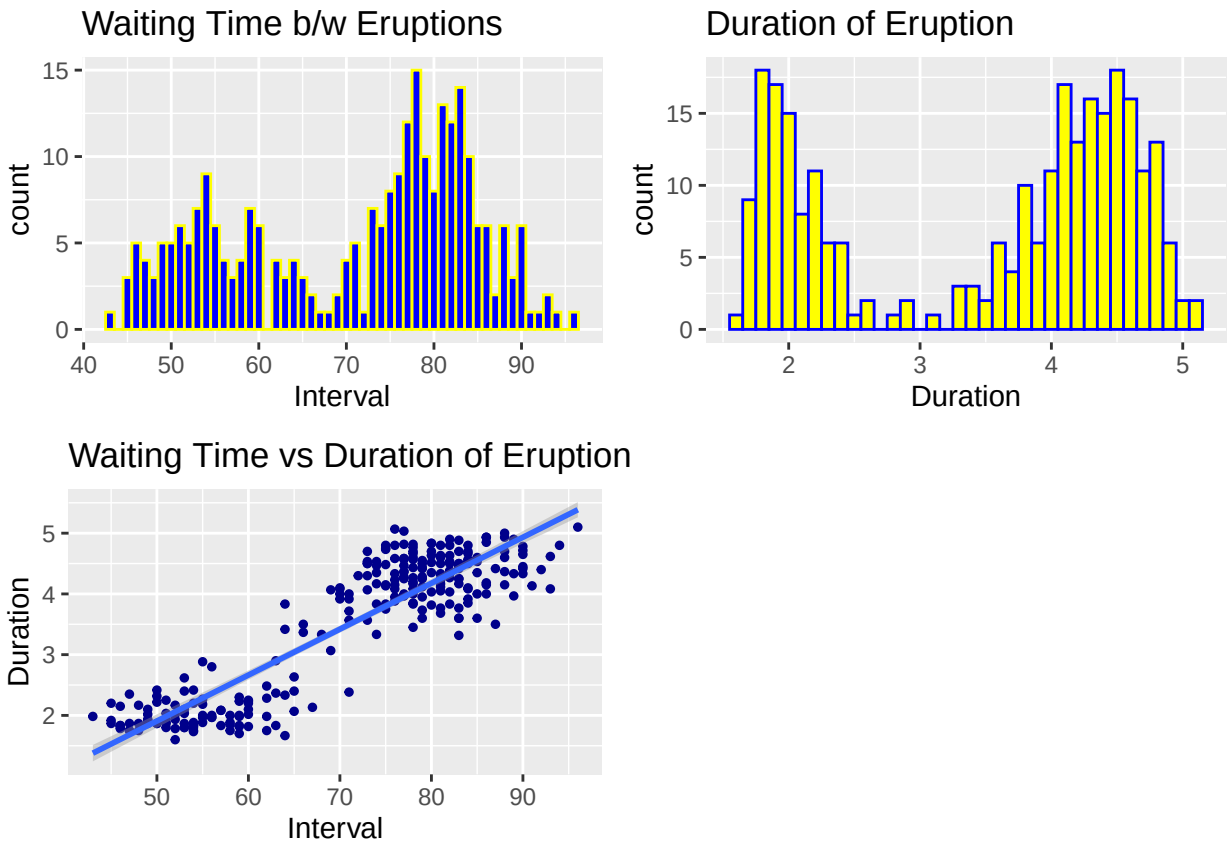
```
library(cowplot)
```

```
##
## ************************************************************
```

```
## Note: As of version 1.0.0, cowplot does not change the
```

```
##   default ggplot2 theme anymore. To recover the previous
##   behavior, execute:
##   theme_set(theme_cowplot())
## ******************************************************
```

```
data("faithful")
of <- faithful
names(of) <- c("Duration", "Interval")


p1 <- ggplot(of, aes(x=Interval, y=Duration)) +
  geom_point(color="darkblue", size = 1) +
  ggtitle("Waiting Time vs Duration of Eruption") +
  geom_smooth(method='lm')

p2 <-ggplot(of, aes(x=Interval)) +
  geom_histogram(color="yellow", fill="blue", binwidth = 1) +
  ggtitle("Waiting Time b/w Eruptions")

p3 <- ggplot(of, aes(x=Duration)) +
  geom_histogram(color="blue", fill="yellow", binwidth = .1) +
  ggtitle("Duration of Eruption")

plot_grid(p2, p3, p1)
```



All three visualizations support the impression that the distribution has two clusters: both histagrams have a clear bimodal distribution; in adition, the scatterplot shows two areas of relative density with a sparse area

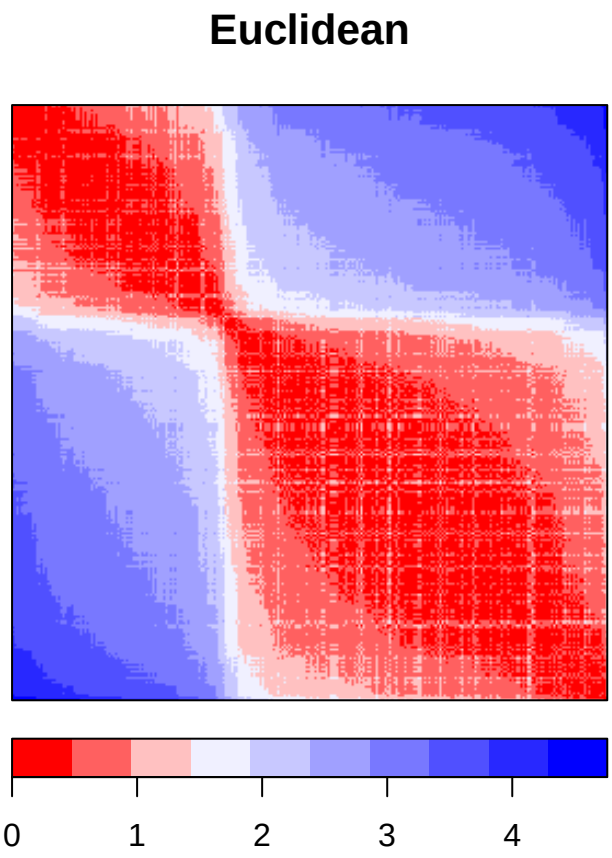
seperating them. Generally, there seems to be a roughly linear relationship between Duration and Interval; though it's possible that the observed relationship is an artifact of the relative location of the clusters.

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

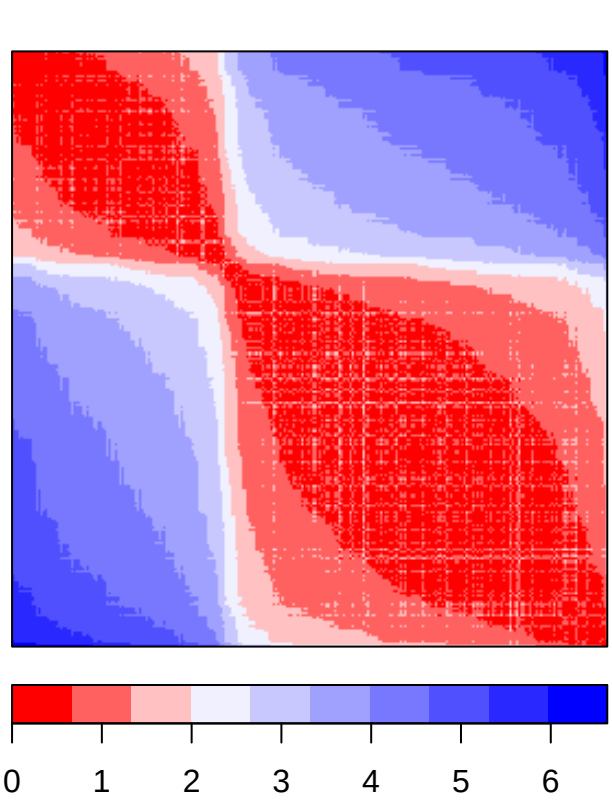
```
library(seriation)
```

```
## Registered S3 method overwritten by 'seriation':
##   method         from
##   reorder.hclust gclus
```

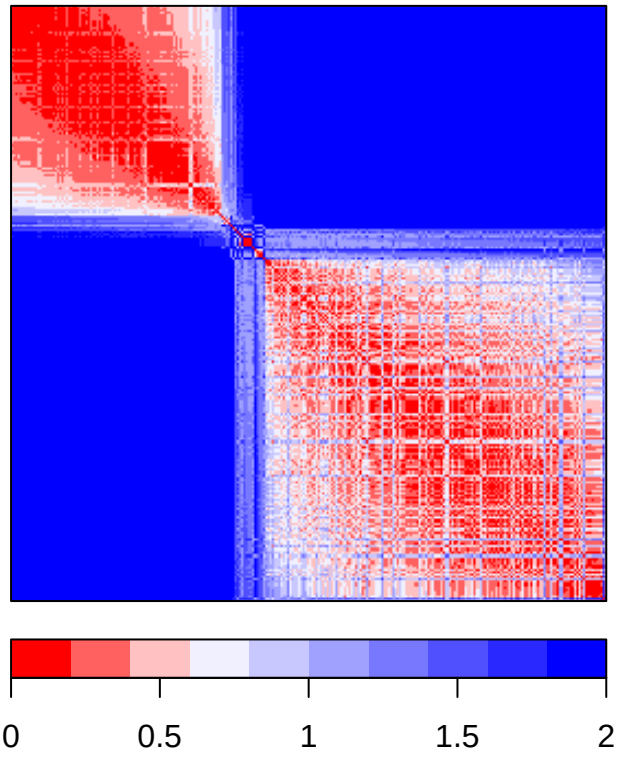
```
data("faithful")
of <- faithful
of_scaled <- scale(of)
of_dist_e <- dist(of_scaled,
                  method = "euclidean")
of_dist_m <- dist(of_scaled,
                    method = "manhattan")
of_dist_c <- dist(of_scaled,
                    method = "canberra")

p3 <- dissplot(of_dist_e, plot = FALSE, main = "Euclidean")
p4 <- dissplot(of_dist_m, plot = FALSE, main = "Manhattan")
p5 <- dissplot(of_dist_c, plot = FALSE, main = "Canberra")
plot(p3, options = list(main = "Euclidean",col= bluered(10, bias=.5), newpage = TRUE))
```



```
plot(p4, options = list(main = "Manhattan",col= bluered(10, bias=.5), newpage = TRUE))
```




```
plot(p5, options = list(main = "Canberra",col= bluered(10, bias=.5), newpage = TRUE))
```



The ODI plots confirm the impression from earlier that there are two clusters. From the ODI plots, we can more clearly see that one of the clusters seems to be larger, but less dense than the other. The Manhattan, Euclidean, and Canberra approaches all reveal a two-cluste structure. The Canberra plot is much more dramatic in its presentation; but, in all sincerity, I am not sure how to interpret that.

IRIS

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 %>% for stacking functions to do this quickly, e.g.).

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?

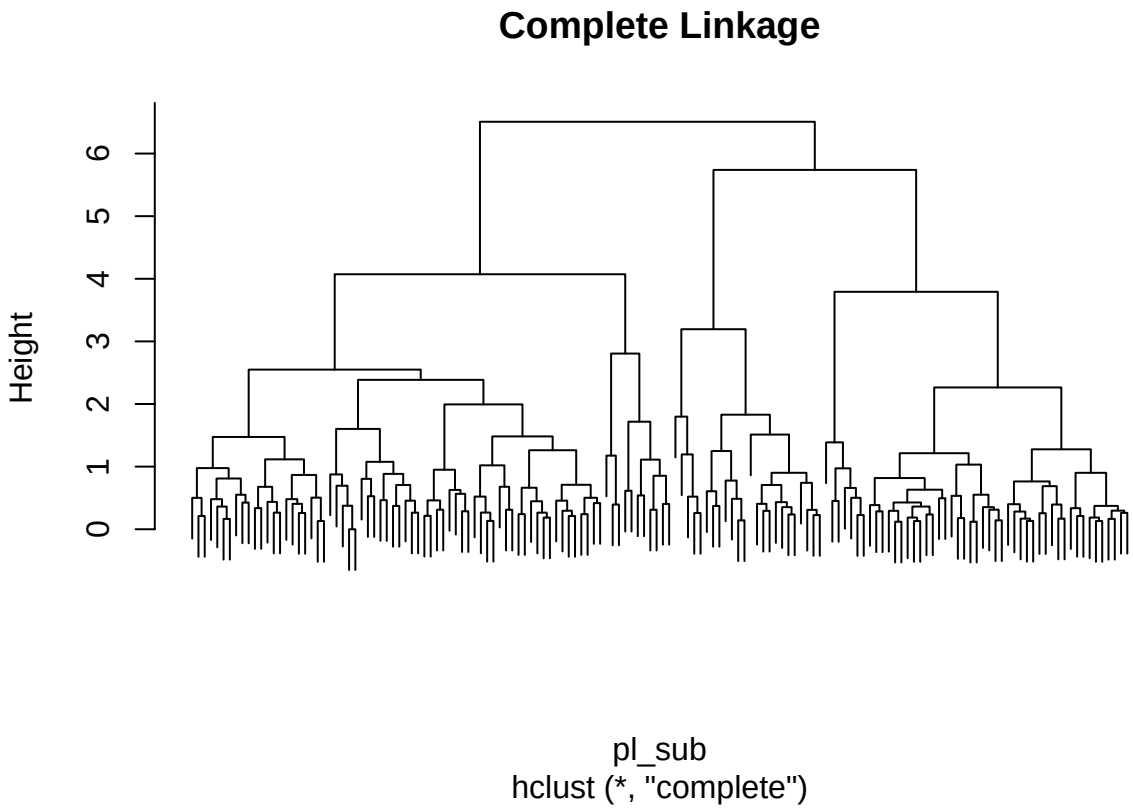
```
library(tidyverse)
library(cowplot)
library(skimr)
```

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

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

```
library(seriation)
data("iris")
pl <- iris
```

```r
pl_sub <- pl %>%
  dplyr::select(-matches("Species")) %>%
  scale() %>% #standardizes values, making values essentially unitness
  dist(method = "euclidean")

hc_complete <- hclust(pl_sub,
                      method = "complete"); plot(hc_complete, labels = FALSE,
                                                 main = 'Complete Linkage')
```

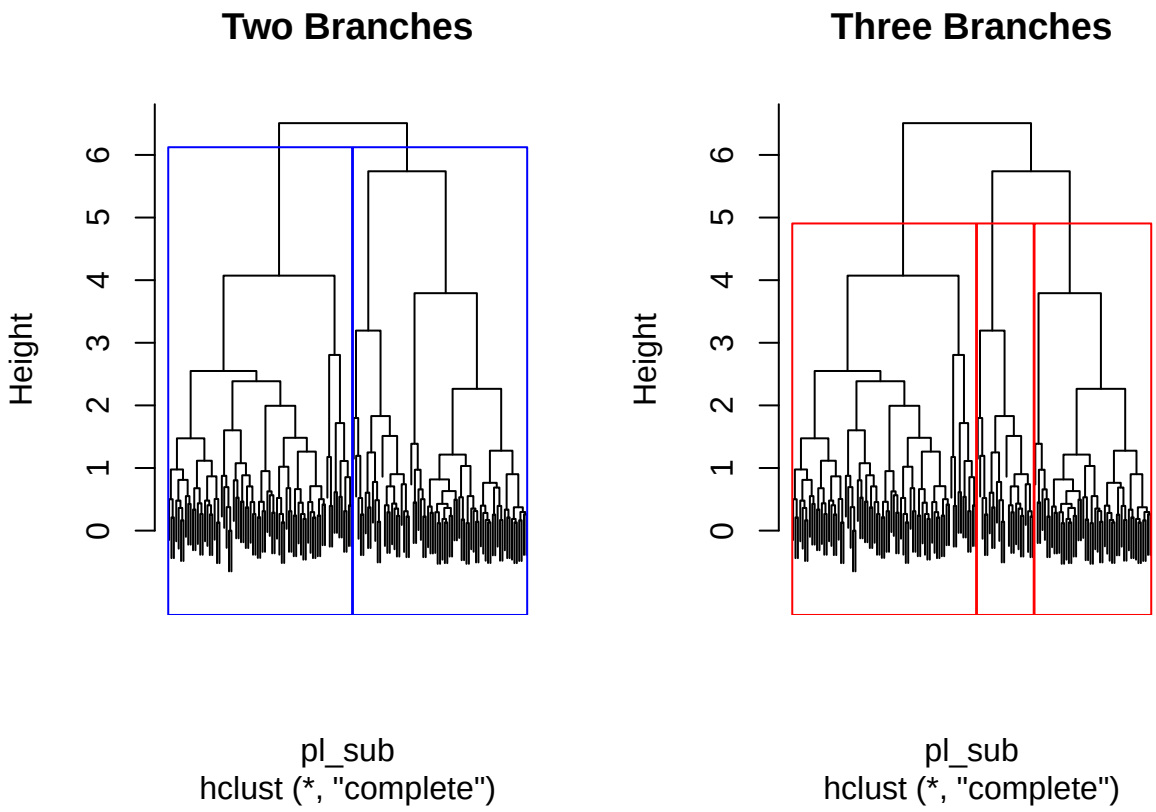


The drendogram begins with individual elements at the lowest level and agglomerates them in a stepwise fashion. Every item seems to find a cluster at early iterations of the clustering algorithm, which is interesting. In theory, we *know* that the *correct* number of clusters for this data is 3, because that's the number of species that are represented. This is probalby a good point to remember the admonition to come in *with minimal priors*. For example, it's tempting to think of specie as a *natural kind*. Of course, as any biologist will readily admit, it isn't: it's a heuristic for thinking about organism behavior and interaction as well as the distribution of traits and the genes that code for them. Let's assume, for the sake of argume, that the drendogram shows us meaningful clusters at every level. At iterations that produce more than three clusters, what are we learning? One thing we could be learning is that there are meaningful differences within the members of a specie; which, of course we know to be true.

But working just off the image, it isn't readily apperent that there is anything special about *three clusters* as the definitive analytic lens. At this point, I would think that the next step is to consider *what kind of data* are we working with? What might we want to know about it? Again, assuming that every cluster iteration is meaningful, you still have to ask, *for what purpose is it meaningful?*

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

```r
data("iris")
pl <- iris
pl_sub <- pl %>%
  dplyr::select(-matches("Species")) %>%
  scale() %>% #standardizes values, making values essentially unitness
  dist(method = "euclidean")

par(mfrow = c(1,2))

hc_cu2 <- hclust(pl_sub,
                      method = "complete"); plot(hc_cu2, labels = FALSE,
                                                 main = 'Two Branches')

rect.hclust(hc_cu2, k = 2, which = NULL, x = NULL, h = NULL,
            border = 4, cluster = NULL)

hc_cu3 <- hclust(pl_sub,
                 method = "complete"); plot(hc_cu3, labels = FALSE,
                                            main = 'Three Branches')

rect.hclust(hc_cu3, k = 3, which = NULL, x = NULL, h = NULL,
            border = 2, cluster = NULL)
```



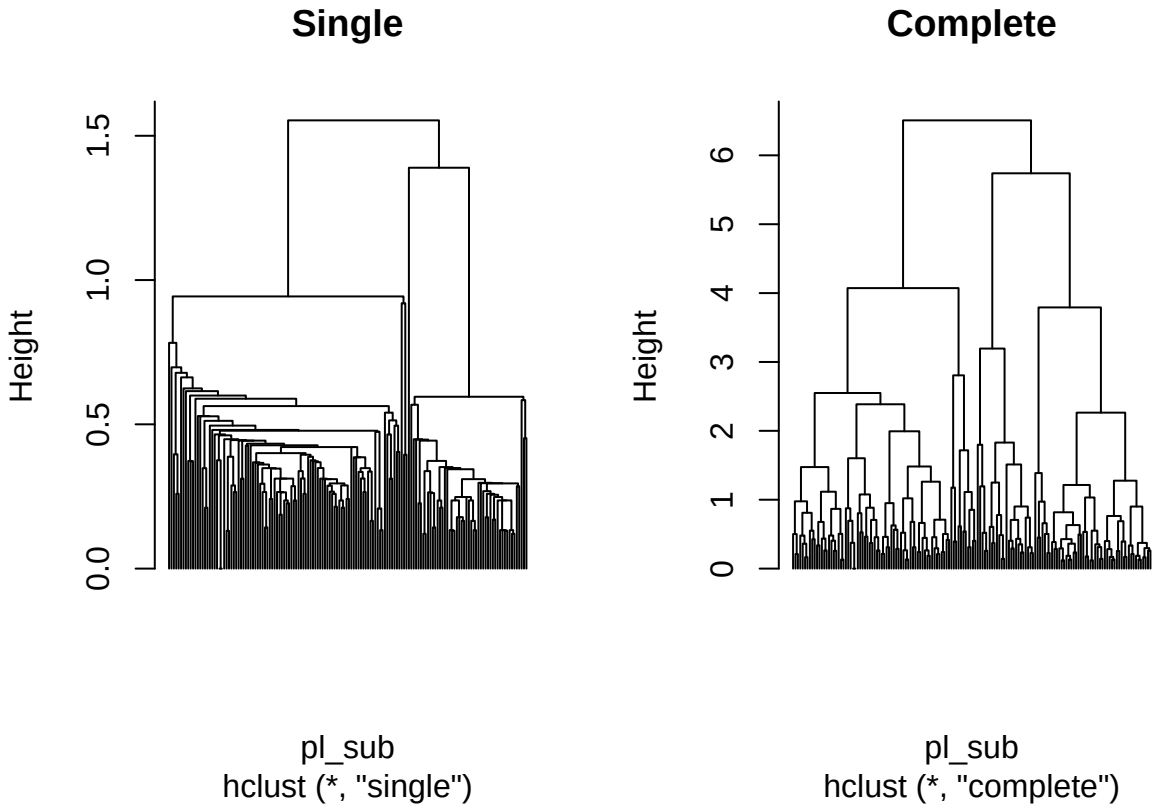

```r
par(mfrow = c(1,1))
```

Comparing the 2-branch and the 3-branch plots highlights the relative size of the clusters, roughly half of the items fall into one cluster in both cases. I expect that the 3-branch plot picks out the three species in the dataset. The 2-branch cluster might be showing that two of the species are closely related. It might not,

though: shared morphological traits don't always reflect shared descent.

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?

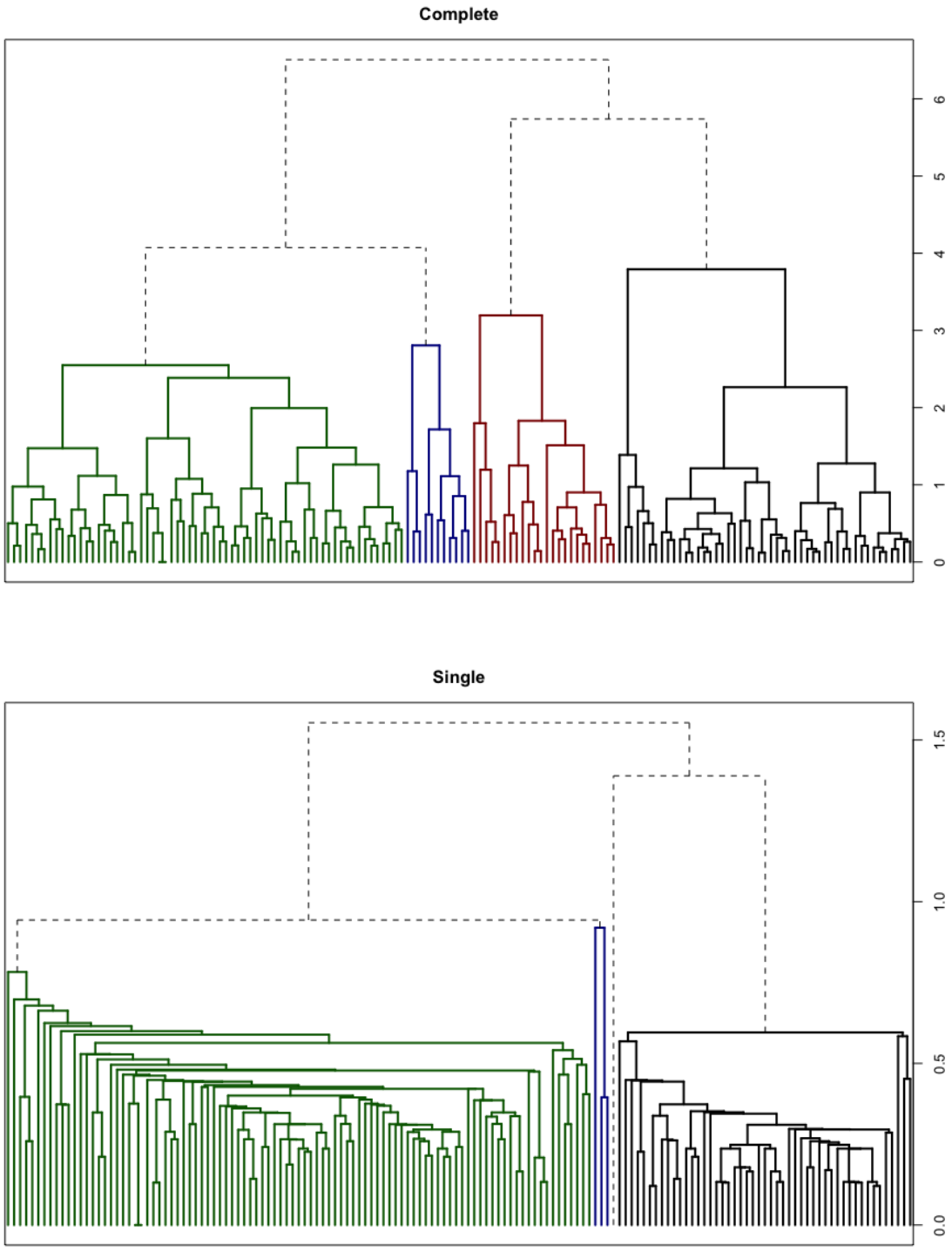
```
par(mfrow = c(1,2))

hc_single <- hclust(pl_sub,
                    method = "single"); plot(hc_single, labels = FALSE,
                                             main = "Single", hang = -1)

hc_complete <- hclust(pl_sub,
                      method = "complete"); plot(hc_complete, labels = FALSE,
                                                 main = "Complete", hang = -1)
```




```
par(mfrow = c(1,1))
```

In truth, I don't find the side-by-side comparison very helpful. I think that's mostly due to there being so many items at the bottom-level that it's hard to really get a whole-structure view of what's going on in order to get a good sense of what's going on. The **complete** linkage approach gives a more *pleasing* structure when comapred to the **single** linkage approach, but it isn't obvious what's actually happening that produces what looks like a more jumbled structure.

I tried using the AZRplot package that we used in class and was able to get it to run, but couldn't get it to run from inside the Rmarkdown file. So, instead, I'm adding the plots here as images:



The color overlay seems helpful. Particularly, it drew my attention to the elements between the green and black clusters in the **single** linkage cluster. As best as I can tell, the red 'cluster' is just one item and it doesn't join another cluster until the second-to-last iteration. In contrast, the **complete** linkage approach identified two clusters in this range, with the result that there was relatively more 'even' rate of cluter formation in under that approach.

My sense is that the **single** linkage approach is *greedy*; that is, it subsumes adjacent points rapidly and prevents them from joining other clusters. In contrast, the **complete** linkage approach seems more 'communal', allowing for a more fluid or organic structure.

I could imagine the **single** linkage approach being used to model a *scramble for resources* scenario, while the **complete** linkage approach might do a better job of explaining coalition building.

Critical Thinking

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

- **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?**

When considering the clusterability of a feature space, my first step would be to represent the data visually. To do this, I would generate an ODI and a scatterplot. The scatterplot can help me gain a sense of the underlying structure and, depending on the kind of data I'm using, might start pointing me towards some notion of what kind of cluters the feature space might exhibit. The ODI is a more constrained visual representationa and therefore is less likely to encourage seeing patterns where there aren't any. Lastly, I would compute the Hopkins statistic, looking for values greater than 0.5.

I would consider the results from these approaches collectively: if there is support for clusterability from each of these, then it is appropriate to perform a clustering analysis.

The reason for performing a prior assesment of clusterability is that clustering algorithms *will* find clusters in a feature space, that doesn't mean those clusters are meaningful. If there is no prior support for clusterability, then performing a clustering exercise is going to generate spurrious findings.

2. **Locate (and read) a paper that applies the hierarchical agglomerative clustering technique. Address the following questions:**

Stashevsky, P. S., Yakovina, I. N., Alarcon Falconi, T. M., & Naumova, E. N. (2019). Agglomerative Clustering of Enteric Infections and Weather Parameters to Identify Seasonal Outbreaks in Cold Climates. *International Journal Of Environmental Research And Public Health*, 16(12). https://doi-org.proxy.uchicago.edu/10.3390/ijerph16122083

- **Describe the author(s) process.**

This is a study examining the relationship between weather conditions and disease outbreak. They had access to a dataset with multiple measures of meteorological conditions as well as another dataset with disease outbreak among target conditions. Both datasets provided data in terms of *days*. After constructing their dataset and normalizing their values, they constructed a distance matrix using *Ward's measurement*. To select the number or clusters, they calculated the *silhouette metric* for models with cluster numbers ranging from 4-15; four selected because it would split the data into the four seasons, and 15 selected in order to ''warrant at least two major seasons over the study period".

Their *methods* sections is very clear and linear - to the extent that I would feel comfortable attempting to replicate it. They lay out their methods, the statistical packages they used, and the reasoning behind their choices. It might have put a slight bounce in my step.

- **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?

It's a bit unclear whether or not they did this before they began their analysis. They present the data visually early on (in the form of stacked line graphs), and those images display clear periodicity. They might have used that as a check on clusterability, but they don't explicitly say so. They also say that they used interquartile range, kurtosis, and the silhouette metric as tests of clustering quality. From their discussion, it looks like they did a quality test after the fact, but not a plausibility test before the fact.

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

Their results indicate a relationship between unseasonably warm weather and disease spread. That opens up a few avenues for further research. First, the findings need to be replicated. If they hold, then further work could focus on creating more precise models as well as forecasting tools. Unseasonably warm weather is likely to increase in many populated areas. The application of this sort of research is one way in which the deleterious effects of that increase can be mitigated.