

# Tammy Glazer Homework 2

10/19/2019

## Computation

Question 1: Calculate Manhattan, Canberra, and Euclidean distances “by hand” (ie. create the data, program each line, and make the calculations).

```
library(skimr)
library(tidyverse)
library(gridExtra)
library(seriation)
library(dendextend)
```

```
# Create the dataframe
x <- data.frame("Variable" = c("p", "q"),
               "Feature_1" = c(1, 3),
               "Feature_2" = c(2, 4))
x
```

```
##   Variable Feature_1 Feature_2
## 1      p          1          2
## 2      q          3          4
```

```
# Manhattan distance = |x1 - x2| + |y1 - y2|
abs(x[1,]$Feature_1 - x[2,]$Feature_1) + abs(x[1,]$Feature_2 - x[2,]$Feature_2)
```

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

```
# Canberra distance = [|x1 - x2| / (|x1| + |x2|)] + [|y1 - y2| / (|y1| + |y2|)]
a <- ((abs(x[1,]$Feature_1 - x[2,]$Feature_1)
      / (abs(x[1,]$Feature_1) + abs(x[2,]$Feature_1)))
      + (abs(x[1,]$Feature_2 - x[2,]$Feature_2)
      / (abs(x[1,]$Feature_2) + abs(x[2,]$Feature_2))))
a
```

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

```
# Euclidean distance = square_root[(x1 - x2)^2 + (y1 - y2)^2]
sqrt((x[1,]$Feature_1 - x[2,]$Feature_1)^2 + (x[1,]$Feature_2 - x[2,]$Feature_2)^2)
```

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

Manhattan distance: 4

Canberra distance: 0.8333333

Euclidean distance: 2.828427

**Question 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?

```
x <- subset(x, select=c("Feature_1", "Feature_2"))
dist(x, method="manhattan")
```

```
##      1
## 2 4
```

```
dist(x, method="canberra")
```

```
##              1
## 2 0.8333333
```

```
dist(x, method="euclidean")
```

```
##              1
## 2 2.828427
```

My distance measures calculated by hand were correct.

**Question 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?

Distance measures are relational, and are used to reflect dissimilarity across features. Euclidean distance can be thought of as the “straight line” distance between points or vectors, calculated as the square root of the sum of squared differences between each axis. Because it squares values, it can be very sensitive to large outliers, but is easy to interpret and explain geometrically. Manhattan distance can be thought of the “city block” distance, and is calculated as the sum of lengths on each axis. Because it uses absolute values instead of exponentiation and rooting, it is more robust to outliers. It is also computationally faster to compute than Euclidean distance, since modulus is faster to compute than exponentiation. Canberra distance is a weighted version of the Manhattan distance, and as such, is also robust to outliers but sensitive to values around 0. It is often used to detect outliers, but is slower to compute than Manhattan distance since it is more complex.

These distinctions matter because they can produce varied results based on the nature and magnitude of the variables, occurrence of outliers, and scales of measurement. Number of records is also important to consider because each method has different computational requirements and will therefore have varied performance. Put another way, distance measures have different levels of complexity, interpretability, and performance. Since we don’t have specific domain expertise or information about whether this fictitious data represents outliers, their relative magnitudes, units of measurement, or performance constraints, it is difficult to determine which measure is most appropriate.



**Question 4:** Use some basic EDA techniques to present and discuss the data (eg. visualize, describe in multiple ways, etc.)

```
data <- faithful
s <- skim(data) # Including an image of the results due to difficulty knitting
knitr::include_graphics("skim.png")
```

#### Skim summary statistics

n obs: 272  
n variables: 2

#### — Variable type:numeric —

variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100	hist
eruptions	0	272	272	3.49	1.14	1.6	2.16	4	4.45	5.1	
waiting	0	272	272	70.9	13.59	43	58	76	82	96	

The data contain 272 observations, with two features called eruptions and waiting. There are no missing values. The mean of eruptions is 3.49 (eruption time in minutes), with a standard deviation of 1.14, and a range of 1.6 to 5.1. The mean of waiting is 70.9 (waiting time to next eruption in minutes), with a standard deviation of 13.59, and a range of 43 to 96.

```
summary(data)
```

```
##      eruptions      waiting
##  Min.   :1.600   Min.   :43.0
##  1st Qu.:2.163   1st Qu.:58.0
##  Median :4.000   Median :76.0
##  Mean   :3.488   Mean   :70.9
##  3rd Qu.:4.454   3rd Qu.:82.0
##  Max.   :5.100   Max.   :96.0
```

```
# Create a function to calculate mode
```

```
mode <- function(options) {
  uniqv <- unique(options)
  uniqv[which.max(tabulate(match(options, uniqv)))]
}
```

```
e_mean <- mean(data$eruptions)
e_median <- median(data$eruptions)
e_mode <- mode(data$eruptions)
e_var <- var(data$eruptions)
e_sd <- sd(data$eruptions)
e_IQR <- IQR(data$eruptions)
```

```
w_mean <- mean(data$waiting)
w_median <- median(data$waiting)
w_mode <- mode(data$waiting)
w_var <- var(data$waiting)
w_sd <- sd(data$waiting)
w_IQR <- IQR(data$waiting)
```

```
kable(data.frame("Variable" = c("Eruption Time", "Waiting Time"),
  "Mean" = c(e_mean, w_mean),
  "Median" = c(e_median, w_median),
```

```

"Mode" = c(e_mode, w_mode),
"Variance" = c(e_var, w_var),
"SD" = c(e_sd, w_sd),
"IQR" = c(e_IQR, w_IQR))

```

Variable	Mean	Median	Mode	Variance	SD	IQR
Eruption Time	3.487783	4	1.867	1.302728	1.141371	2.2915
Waiting Time	70.897059	76	78.000	184.823312	13.594974	24.0000

The median for eruption time (4) is higher than its mean, indicating a potential skew towards higher values. Interquartile range tells you the difference between the 3rd and 1st quartile in a dataset. It is a measure of spread with a higher resistance to outliers, and falls at 2.29 for eruption time.

The median for waiting time (76) is also higher than its mean, again indicating a potential skew towards higher values. IQR for this variable is 24.

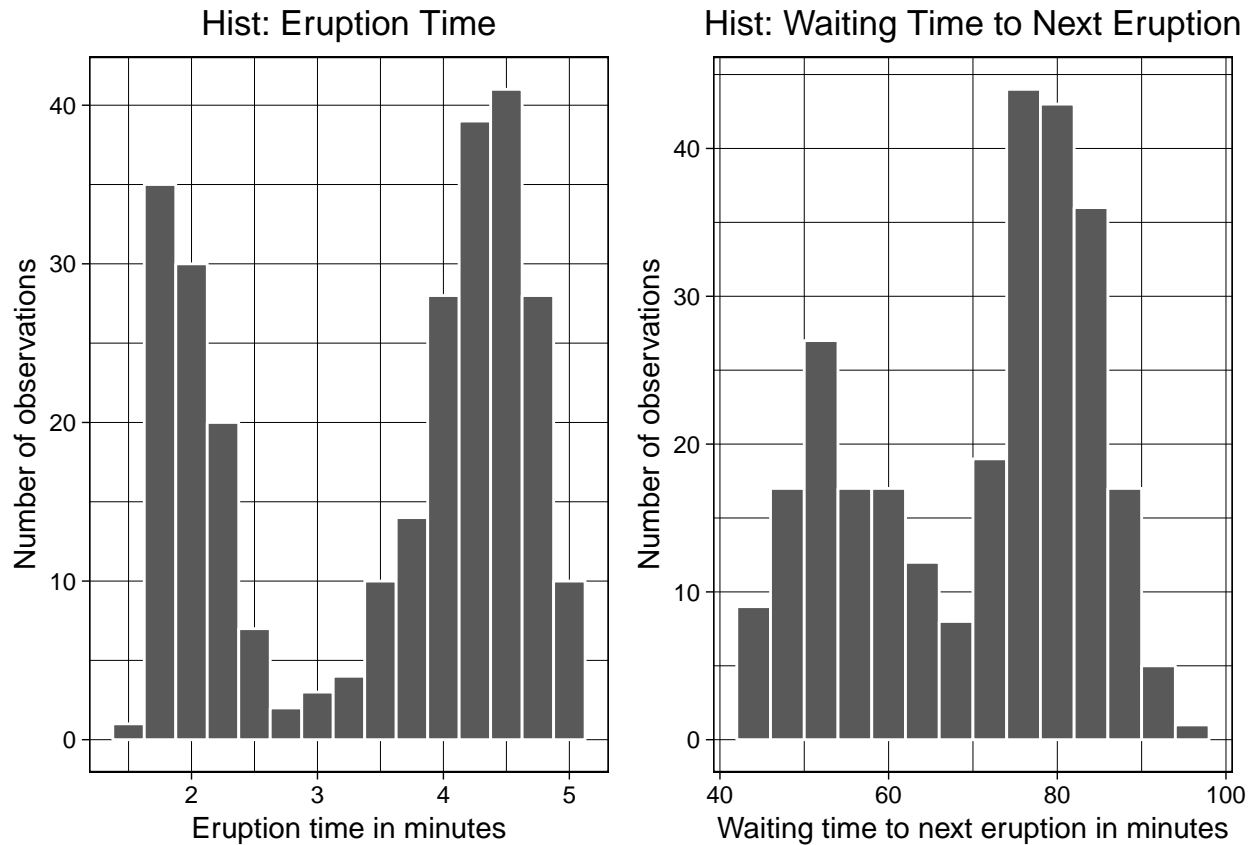
```

plot1 <- ggplot(data) +
  geom_histogram(aes(x = eruptions), binwidth = 0.25, color = "white") +
  labs(x = "Eruption time in minutes",
       y = "Number of observations") +
  ggtitle("Hist: Eruption Time") +
  theme_linedraw() +
  theme(plot.title = element_text(hjust = 0.5))

plot2 <- ggplot(data) +
  geom_histogram(aes(x = waiting), binwidth = 4, color = "white") +
  labs(x = "Waiting time to next eruption in minutes",
       y = "Number of observations") +
  ggtitle("Hist: Waiting Time to Next Eruption") +
  theme_linedraw() +
  theme(plot.title = element_text(hjust = 0.5))

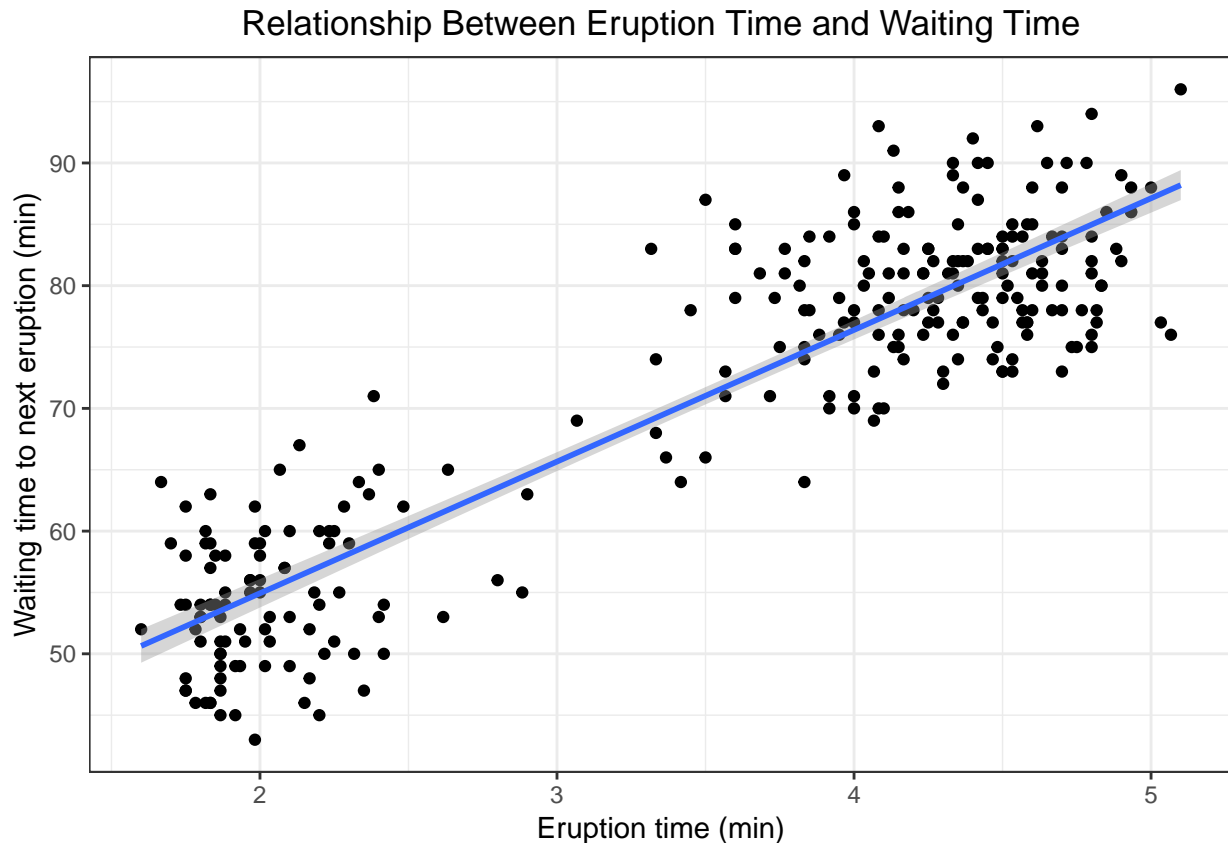
grid.arrange(plot1, plot2, ncol=2)

```



In this dataset, the most common eruption times are centered just under 2 minutes in length and around 4.5 minutes in length. There are very few observations that fall just under 3 minutes in length, while a majority of observations seem to fall between 3.5 and 5 minutes in length. Moreover, in this dataset, the most common waiting time to next eruption is centered just under 80 minutes in length. Very few observations have waiting times above 90 minutes or below 50 minutes. There is a notable dip in the number of observations just before the 70 minute mark. Both histograms demonstrate similar trends: an initial increase in observations at low values, followed by a clear decrease, and then a second clear increase at a higher value.

```
ggplot(data, aes(x = eruptions, y = waiting)) +
  geom_point() +
  geom_smooth(method = lm) +
  labs(x = "Eruption time (min)",
       y = "Waiting time to next eruption (min)",
       title = "Relationship Between Eruption Time and Waiting Time") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```



There is a noticeable, positive relationship between eruption time and waiting time to next eruption (as one increases, the other also seems to increase). The smoothed conditional means linear plot is included simply to aid the eye in seeing this pattern. There also appears to be two groupings or location-based clusters in the data, with very few points falling between them. These occur around an eruption time of 2 minutes and a waiting time of 55 minutes, and again around an eruption time of 4.5 minutes and a waiting time of 80 minutes. The fewest points appear around an eruption time of 3 minutes and a waiting time of 70 minutes.

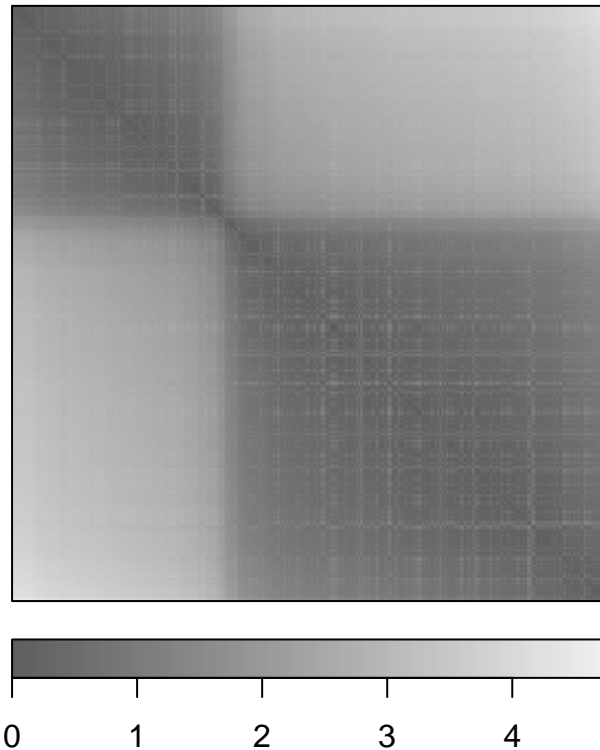
**Question 5:** Calculate a dissimilarity matrix of these data.

```
dis_matrix <- data %>%
  scale() %>% # standardize variables
  dist(method="euclidean")
```

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

```
dissplot(dis_matrix,
  main="ODI for Eruption Time and Waiting Time")
```

## ODI for Eruption Time and Waiting Time



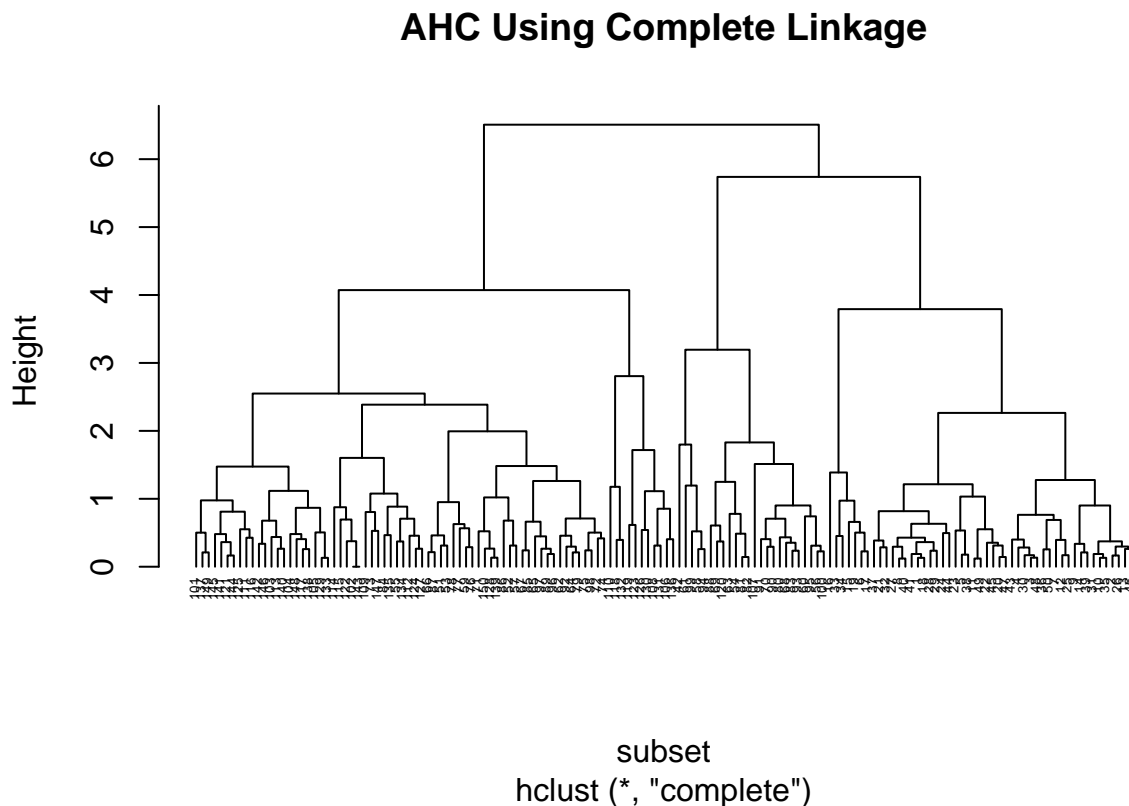
After ordering by spatial proximity, this ordered dissimilarity matrix (ODI) demonstrates a strong case for clusterability in the data and confirms that there is a cluster structure. There is a small dark square in the upper lefthand corner of the image, indicating a grouping of similar vectors (low dissimilarity), as well as a large dark square in the bottom righthand corner of the image, indicating another grouping of similar vectors (again, low dissimilarity). The lighter sections on either side demonstrate high dissimilarity between these groupings. There appears to be only a few square, dark blocks along the diagonal.

**Question 7:** Using any munging tools you'd like (eg. 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)

```
data <- iris
subset <- data %>%
  select("Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width") %>%
  scale() %>%
  dist(method="euclidean")
```

**Question 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?

```
hc_complete <- hclust(subset, method = "complete")
plot(hc_complete, hang = -1, cex = 0.4,
     main = "AHC Using Complete Linkage")
```



The height (y-axis) represents dissimilarity between clusters. The x-axis represents the objects and clusters. Each joining of two clusters is represented by the splitting of a vertical line into two vertical lines, and the height of the split represents the dissimilarity between the two clusters. On first glance, it appears that there are three main clusters in the data (if you draw a horizontal line at height 2.5, it cuts the tree in 3 locations). It also does not appear that there are any extremely visible outliers in the data, since there are no strikingly tall vertical bars connecting a cluster to a specific point on the x-axis. The low height of a majority of the connections indicates a fair amount of similarity in the dataset.

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

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

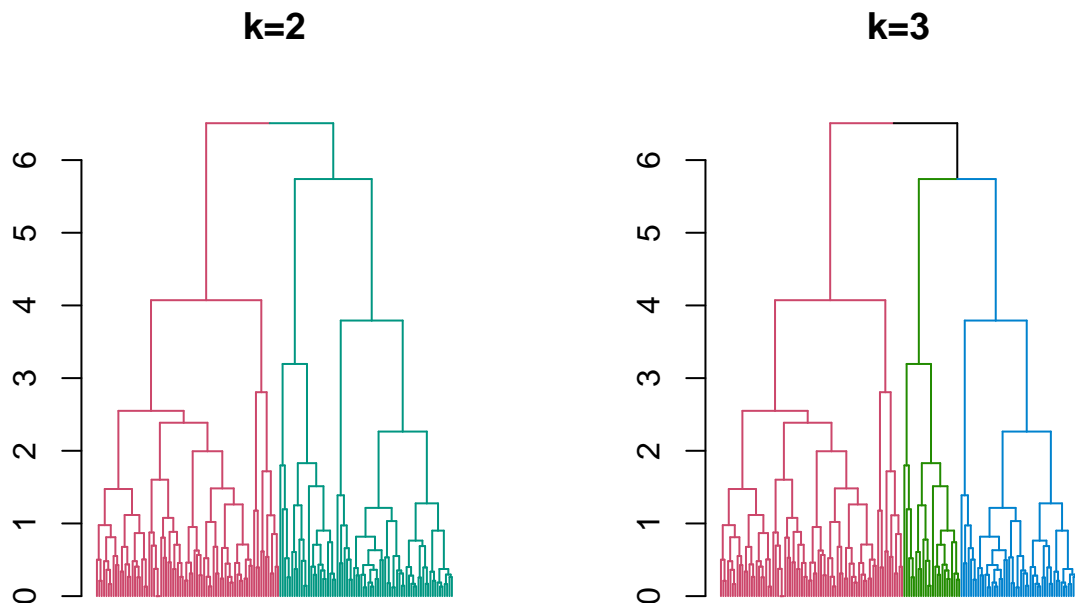
par(mfrow = c(1,2))
plot(hc_complete %>%
     as.dendrogram %>%
```



```

set("branches_k_color", k=2),
leaflab="none",
main = "k=2")
plot(hc_complete %>%
  as.dendrogram %>%
  set("branches_k_color", k=3),
  leaflab="none",
  main = "k=3")

```



The tree cut at 2 branches appears to have two large, approximately evenly sized clusters (distinguished by color), with a dissimilarity represented by a height of ~6.5. Meanwhile, the tree cut at 3 appears to have three clusters (again, distinguished by color). It appears that one of the clusters remains the same in the two plots (maroon), but that the second cluster in  $k=2$  is separated into two distinct groups in  $k=3$  (green and blue). Therefore, the tree cut at 3 branches has one large cluster, one mid-sized cluster, and one small cluster. Because the height of the split between the green and blue clusters is at a height of ~5.5, while the split between them and the third cluster is at ~6.5, it would appear that the smaller two clusters have a higher similarity than the third (maroon).

**Question 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?

```

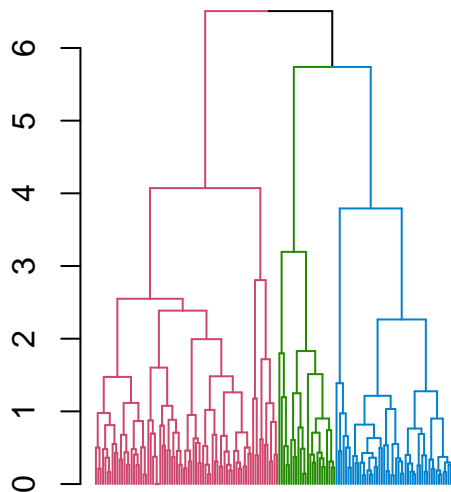
hc_single <- hclust(subset, method = "single")

par(mfrow = c(1,2))
plot(hc_complete %>%
  as.dendrogram %>%
  set("branches_k_color", k=3),
  leaflab="none",
  main = "Complete Linkage, k=3")
plot(hc_single %>%
  as.dendrogram %>%

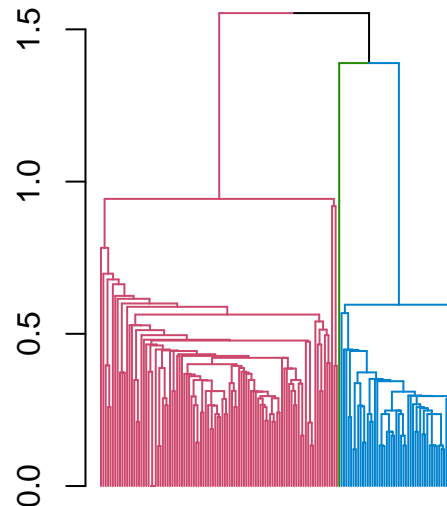
```

```
set("branches_k_color", k=3),
leaflab="none",
main = "Single Linkage, k=3")
```

### Complete Linkage, k=3



### Single Linkage, k=3



Linkage defines the dissimilarity between two groups of observations. Complete linkage uses the maximal inter-cluster dissimilarity, whereas single linkage uses the minimal inter-cluster dissimilarity. In this example, it is clear that the model fit using complete linkage is far more balanced. It appears that several individual observations fuse early on, resulting in the fusing of larger and larger clusters over time. When the tree is cut at  $k=3$ , therefore, the result is one large cluster, one mid-sized cluster, and one small cluster. Additionally, the height scale is from 0-6. Meanwhile, the model fit using single linkage is far less balanced and displays very different clustering patterns. The clusters are very elongated, and it appears that individual observations are fusing one-at-a-time or one-by-one. When the tree is cut at  $k=3$ , the result is one very large cluster, one mid-size cluster, and one extremely small cluster, which may have only a single observation. It is far less balanced, and the height scale is interestingly from 0-1.5. Therefore, different linkage methods can result in very different clustering patterns and output. Selecting the most useful method depends largely on domain expertise.

## Critical Thinking

### Question 1a) How would you go about determining whether clustering made sense to consider or not?

To determine whether clustering makes sense to consider, I use a variety of informal and mathematical strategies to decide if there may be natural groupings present in the data, such as by location, shape, or density. To make this decision, I would begin by exploring the feature space through informal EDA. I would spend time looking at measures of central tendency, measures of variation, and creating a variety of exploratory visualizations that may provide some insight into strong relationships within the data. For instance, I may plot a matrix of scatterplots to compare each pair of continuous variables (overlying color to find patterns in categorical variables), plot histograms of individual continuous variables, or plot choropleths to uncover density patterns.

I would also leverage a Visual Assessment of Tendency (VAT/ODI) plot to visualize the dissimilarity matrix for the data. If I were to see distinct, dark regions next to lighter regions, clustering may make sense. How-

ever, if I were to find seemingly random light and dark blocks, I would progress to a mathematical method, such as calculating the Hopkins statistic. This statistic tests the null hypothesis of spatial randomness in the data using a sparse sampling test, and calculates the probability that a given dataset is generated by a uniform distribution or not. If I were to reject the null hypothesis, I would certainly consider clustering. Ideally, I would also be able to consult a domain expert who has more familiarity with the dataset before making my final determination.

### **Question 1b) What are techniques you would use, and what might you be looking for from each?**

Informal Techniques (EDA):

1. Summarize measures of central tendency and variation: I would look for wide or inconsistent variation in the data, such as a high standard deviation among features. I would also look for skew in the data, such as a mean that is very different than a median for a few features.
2. Plot visualizations including a matrix of scatterplots, histograms, and choropleths: I would look for visual cues that there are natural groupings in the data, such as wide gaps creating distinct sections in scatterplots or groups of data that form shapes in scatterplots. I would also look for distinct peaks and valleys in histograms, indicating non-uniform distributions. In choropleths, I would look for darker and lighter areas, indicating varied density and potential relationships between variables.
3. Consult a domain expert: I would also discuss the data with a domain expert, who may hypothesize that there are particular groupings or insist that there are none, based on a more contextual understanding.

Visual Technique:

1. VAT/DOI plots: I would look for large, dark squares along the diagonal that can be visually parsed from light regions. This is because darker blocks along the diagonal reflect greater spatial similarity, compared to lighter shaded blocks, which inversely suggest greater dissimilarity.

Mathematical Technique:

1. Hopkins statistic using a sparse sampling test: I would look for a value of the Hopkins statistic that is close to 1, indicating that the data are highly clustered, rather than a value close to 0.5, indicating random data, or a value close to 0, indicating a uniform distribution of data. I would also look for high statistical significance. In other words, I would attempt to reject the null hypothesis of spatial randomness in the data. The Hopkins statistic calculates the probability that a given dataset is generated by a uniform distribution or not, by comparing pairwise dissimilarity with a set of simulated data drawn from a random distribution with the same standard deviation as the original data.

### **Question 1c) How might these techniques work together to motivate clustering or not?**

No technique on its own should be used to diagnose clusterability. Rather, a combination of these techniques should be used, including consultation with a domain expert when possible, to determine if there is enough evidence to support the notion that there may be “natural groupings” in the data. For instance, one could begin by consulting a domain expert who thinks that a particular dataset may contain groupings, and then could apply informal and visual techniques to find support for the initial claim. I believe that clustering should be used if one can use the Hopkins statistic to reject the null hypothesis of spatial randomness. That said, I do not believe that it is an absolutely necessary condition to meet if interesting patterns have already

emerged from informal techniques. Finally, even if I were to visualize the data, find clear separation, and verify this with ODI and the Hopkins statistic, I would calculate measures of central tendency within the clusters themselves to better understand if they are substantially different from each other. If not, clustering may not produce new or interesting insights.

**Question 1d) And ultimately, can/should you proceed if you find little to no support for clusterability? Why or why not?**

Because methods for evaluating clusterability vary radically, it may be challenging to select a set of suitable methods. It may be possible, for instance, that a method you did not select would have supported clusterability, even though the method(s) you selected did not. With unlimited time and resources, it may still be interesting to proceed without direct support for clusterability. Ambiguity underlies unsupervised machine learning, and although diagnosing clusterability represents informed guessing, it is not always going to be conclusive given unlabeled data. Moreover, clustering could be used to confirm the null hypothesis that there are no clusters, or to compare output with that of a more suitable dataset. However, if time and resources are limited, I would be less inclined to recommend proceeding based on an informed guess that the data are not clusterable.

**Identify a paper that applies the hierarchical agglomerative clustering technique:**

Title: "Identifying Subgroups of Complex Patients with Cluster Analysis"

Authors: Sophia R. Newcomer, John F. Steiner, and Elizabeth A. Bayliss

Date: August, 2011

Source: The American Journal of Managed Care

Link: [https://ajmc.s3.amazonaws.com/\\_media/\\_pdf/948a5455554eb134cf52e6662520b33.pdf](https://ajmc.s3.amazonaws.com/_media/_pdf/948a5455554eb134cf52e6662520b33.pdf)

**Question 2a) Describe the author(s) process.**

The objective of the study is to illustrate the use of cluster analysis for identifying sub-populations of complex patients with multiple interacting diseases, who may benefit from targeted care management strategies. Care management has the potential to improve health outcomes for people with multimorbidities, but most strategies focus on populations defined by a single disease. To begin, they developed a hypothesis: within a complex patient population, cluster analysis will reveal groups of patients with distinct patterns of comorbid conditions. Next, they identified an appropriate dataset for this task: 15,480 Kaiser Permanente Colorado members, age 21+, who were in the top 20% of total cost of care in 2006-2007, and had 2+ of 17 common chronic medical conditions. The data contain diagnoses, clinical conditions, demographic attributes, healthcare utilization, and a comorbidity score.

Next, the authors performed informal EDA to gather summary statistics, focusing on medians and interquartile range. Based on this analysis, they decided to leverage HAC methods to identify discrete groups of patients with specific combinations of comorbid conditions. They randomly split the data into two equal parts and converted each section into a dissimilarity matrix using Jaccard's coefficient to measure distance. They determined that Jaccard's coefficient is appropriate for clinical conditions since it considers the number of conditions that two people have in common and ignores conditions that neither person has. Next, they fit a HAC algorithm using Ward's minimum variance as a linkage method, because it required the fewest assumptions and they felt it would have the greatest explanatory power. They also compared results with the flexible beta algorithm, to assess the consistency of the clustering process. Once stability was established, they re-joined the data and ran Ward's algorithm on the entire dataset.

The authors selected a cutoff of 10 relevant clusters, based on clinical importance and pre-specified criteria. These clusters revealed distinct groups of patients including: coexisting chronic pain and mental illness, frail elderly, and specific surgical procedures. They described the 10 clusters by number of members, median age,

percentage with the most prevalent conditions, and relative cost of care ratios. They concluded that although several clusters lend themselves to existing care and disease management protocols, care management for other subgroups is less well-defined. This will help them to target care management interventions moving forward.

**Question 2b) Do they go through similar steps as we covered this week both in setting the stage for clustering (eg. assessing clusterability, calculating distance, etc.), as well as fitting the algorithm? If not, what did they omit and does this omission impact their findings in your opinion?**

In general, the authors followed similar steps as we covered in class in setting the stage for clustering and fitting the algorithm. To assess clusterability and identify outliers, they used informal techniques such as calculating frequencies and measures of central tendency (median and interquartile range). Through data exploration, they decided to exclude individuals with long-term care facility stay, chronic kidney dialysis, or inpatient visits of over 30 days, since their unique care management needs are likely already defined. They also identified 6 extremely high cost outliers and decided to remove them. They do not seem to employ any VAT/ODI plots to visualize the dissimilarity matrix, or to apply mathematical tests such as the Hopkins statistic to assess clusterability. Given their domain expertise (all authors have either an MPH or MD), however, the omission does not seem to directly impact findings. They are still able to successfully identify clinically relevant clusters, as compared to a simple tabulation of chronic conditions. A simple tabulation would be difficult to interpret, given over 100,000 different possible combinations of conditions.

Next, they decided to use Jaccard's coefficient to measure distance and build a dissimilarity matrix, because it is most appropriate for clinical conditions. Specifically, it considers the number of conditions that two people have in common and ignores conditions that neither person has. I believe that it may have been interesting to compare two or three different distance measures. Interestingly, they decided to split the full data into two parts before creating these matrices, and fit each model using Ward's minimum variance method for linkage. They examined the pseudo F, T, and R<sup>2</sup> statistics for the different clusters in each dataset, and compared cluster membership to verify consistency, before rejoining the data. Ward's method minimizes variance within clusters and is known to produce clusters of similar sizes. It may have been interesting, again, if they compared the output from several different linkage methods, in order to better assess consistency. They did compare the results from Ward's method to the results of a flexible beta algorithm, and decided that Ward's method was the most parsimonious.

Although they did not describe the process of creating a dendrogram to visualize output, this may be assumed. Dendrograms are helpful in determining where to cut the tree. They did subjectively determine that a 10-cluster solution produced the most clinically relevant clusters. In sum, the authors closely followed the steps we described in class, with less emphasis on diagnosing clusterability due to domain expertise.

**Question 2c) Describe at least one possible extension from the study that could emerge based on their findings.**

To further assess the stability of clusters over time, a researcher could conduct another analysis using cohorts from different years. He/she could also attempt the same analysis with a larger dataset or with a different patient population (such as from a different location), to determine localization and/or consistency of results. Moreover, he/she could focus on children, individuals with at least 3 chronic conditions, or could select conditions of interest based on prevalence in a particular region of the country. All of these variations might produce different, clinically relevant clusters. Finally, given sufficient resources, it would be interesting to develop a targeted care management strategy based on a few specific clusters, and then to assess the impact of interventions over time to help verify the effectiveness of HAC. For instance, they could develop a care management strategy for a less well-defined subgroup like "mental health conditions and obesity in younger adults", and then examine how this strategy impacts trends in prevalence using a difference-in-differences methodology.