# 7. Worksheet: Diversity Synthesis

Jocelyn Huang; Z620: Quantitative Biodiversity, Indiana University

16 February, 2025

## OVERVIEW

In this worksheet, you will conduct exercises that reinforce fundamental concepts of biodiversity. First, you will construct a site-by-species matrix by sampling confectionery taxa from a source community. Second, you will make a preference-profile matrix, reflecting each student's favorite confectionery taxa. With this primary data structure, you will then answer questions and generate figures using tools from previous weeks, along with wrangling techniques that we learned about in class.

### Directions:

1. In the Markdown version of this document in your cloned repo, change "Student Name" on line 3 (above) to your name.
2. Complete as much of the worksheet as possible during class.
3. Refer to previous handouts to help with developing of questions and writing of code.
4. Answer questions in the worksheet. Space for your answer is provided in this document and indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">". You should notice that the answer is highlighted in green by RStudio (color may vary if you changed the editor theme).
5. Before you leave the classroom, **push** this file to your GitHub repo.
6. For the assignment portion of the worksheet, follow the directions at the bottom of this file.
7. When you are done, **Knit** the text and code into a PDF file.
8. After Knitting, submit the completed exercise by creating a **pull request** via GitHub. Your pull request should include this file `7.DiversitySynthesis_Worskheet.Rmd` and the PDF output of `Knitr` (`DiversitySynthesis_Worskheet.pdf`).

## QUANTITATIVE CONFECTIONOLOGY

We will construct a site-by-species matrix using confectionery taxa (i.e, jelly beans). The instructors have created a **source community** with known abundance ($N$) and richness ($S$). Like a real biological community, the species abundances are unevenly distributed such that a few jelly bean types are common while most are rare. Each student will sample the source community and bin their jelly beans into operational taxonomic units (OTUs).

## SAMPLING PROTOCOL: SITE-BY-SPECIES MATRIX

1. From the well-mixed source community, each student should take one Dixie Cup full of individuals.
2. At your desk, sort the jelly beans into different types (i.e., OTUs), and quantify the abundance of each OTU.
3. Working with other students, merge data into a site-by-species matrix with dimensions equal to the number of students (rows) and taxa (columns)
4. Create a worksheet (e.g., Google sheet) and share the site-by-species matrix with the class.

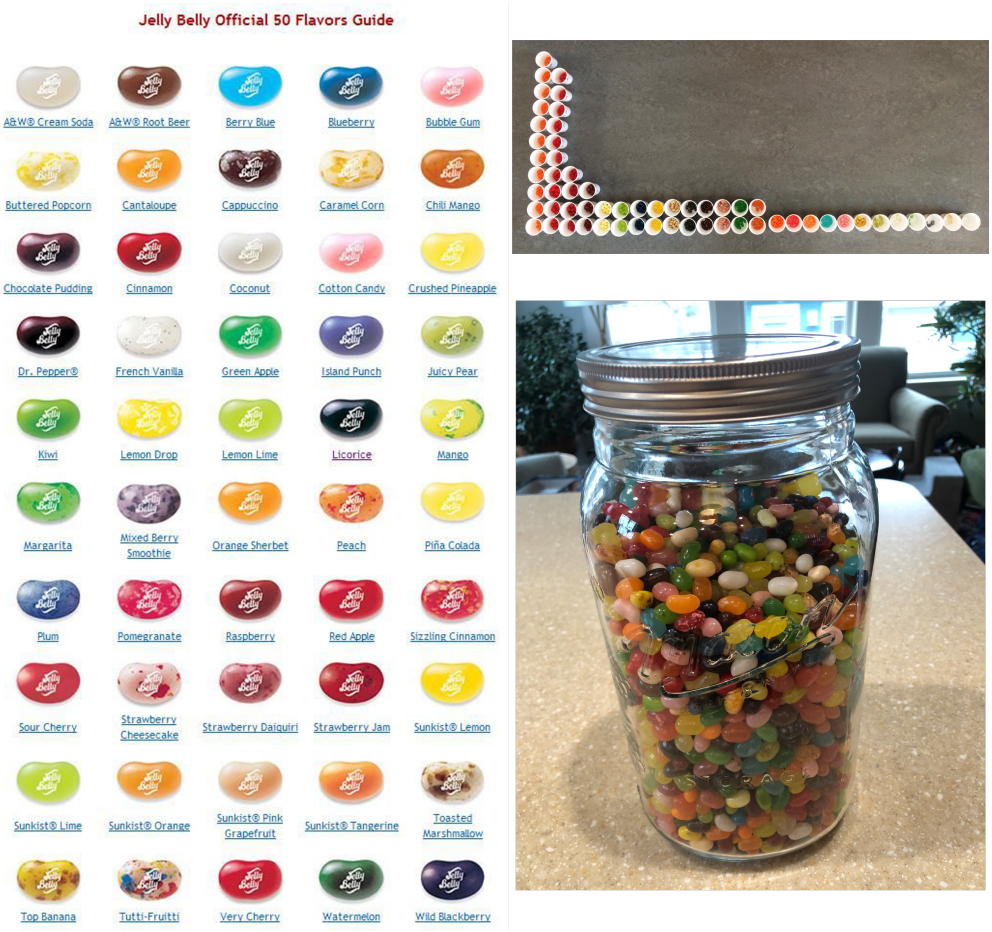



Figure 1: **Left**: taxonomic key, **Top right**: rank abundance distribution, **Bottom right**: source community

## SAMPLING PROTOCOL: PREFERENCE-PROFILE MATRIX

1. With your individual sample only, each student should choose their top 5-10 preferred taxa based on flavor, color, sheen, etc.
2. Working with other students, merge data into preference-profile incidence matrix where 1 = preferred and 0 = non-preferred taxa.
3. Create a worksheet (e.g., Google sheet) and share the preference-profile matrix with the class.

### 1) R SETUP

In the R code chunk below, please provide the code to: 1) Clear your R environment, 2) Print your current working directory, 3) Set your working directory to your `Week5-Confection/` folder, and 4) Load the `vegan` R package (be sure to install first if you have not already).

```
rm(list = ls())
getwd()
```

```
## [1] "/cloud/project/QB2025_Huang/Week5-Confection"
```

```
setwd("/cloud/project/QB2025_Huang/Week5-Confection")
library(vegan)
```

```
## Loading required package: permute
```

```
## Loading required package: lattice
```

```
## This is vegan 2.6-8
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

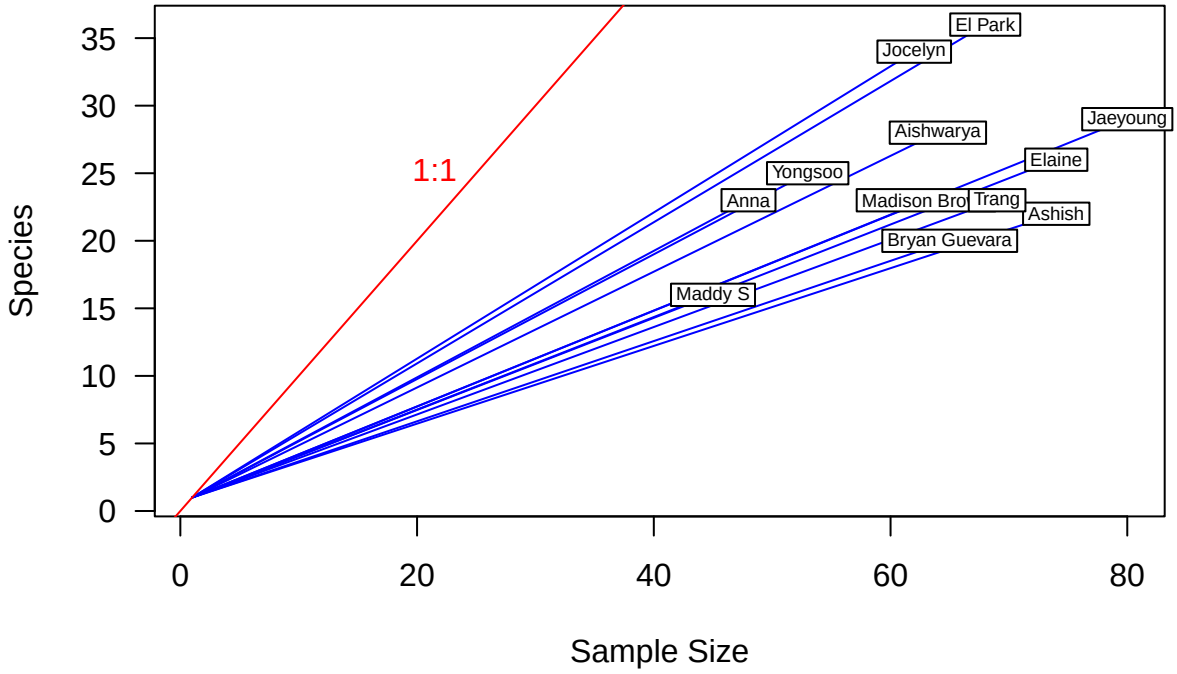
```
library(ggplot2)
library(dplyr)
library(broom)
```

## DATA ANALYSIS

**Question 1:** In the space below, generate a rarefaction plot for all samples of the source community. Based on these results, discuss how individual vs. collective sampling efforts capture the diversity of the source community.

```
#source <- read.csv(file = "./data/jelly.source.comm.csv", header = TRUE, row.names = 1)
#source.t <- as.data.frame(t(source))
#specnumber(source.t)

# load SbyS matrix:
jelly <- read.csv(file = "./data/Jelly belly data! - SbyS.csv", header = TRUE, row.names = 1)
```

```r
jelly.S <- specnumber(jelly)
min.N <- min(rowSums(jelly))
S.rarefy <- rarefy(x = jelly, sample = min.N, se = TRUE)
rarecurve(x = jelly, step = 200, col = "blue", cex = 0.6, las = 1)
abline(0, 1, col = 'red')
text(25, 25, "1:1", pos = 2, col = 'red')
```




***Answer 1***: Everyone has slightly different sample size and differnt total number of observed species in each samples. Collective sampling can reduce the individual error in sampling and identifying species, and can increase the total number of sample sizes. However, with individual sampling, the identification scheme is probably going to be more consistent. From the rarefaction plot, we can also tell that the differences in observed richness among each sample are not entirely because of the differences in the size of each sample, but might be influenced by other factors instead.

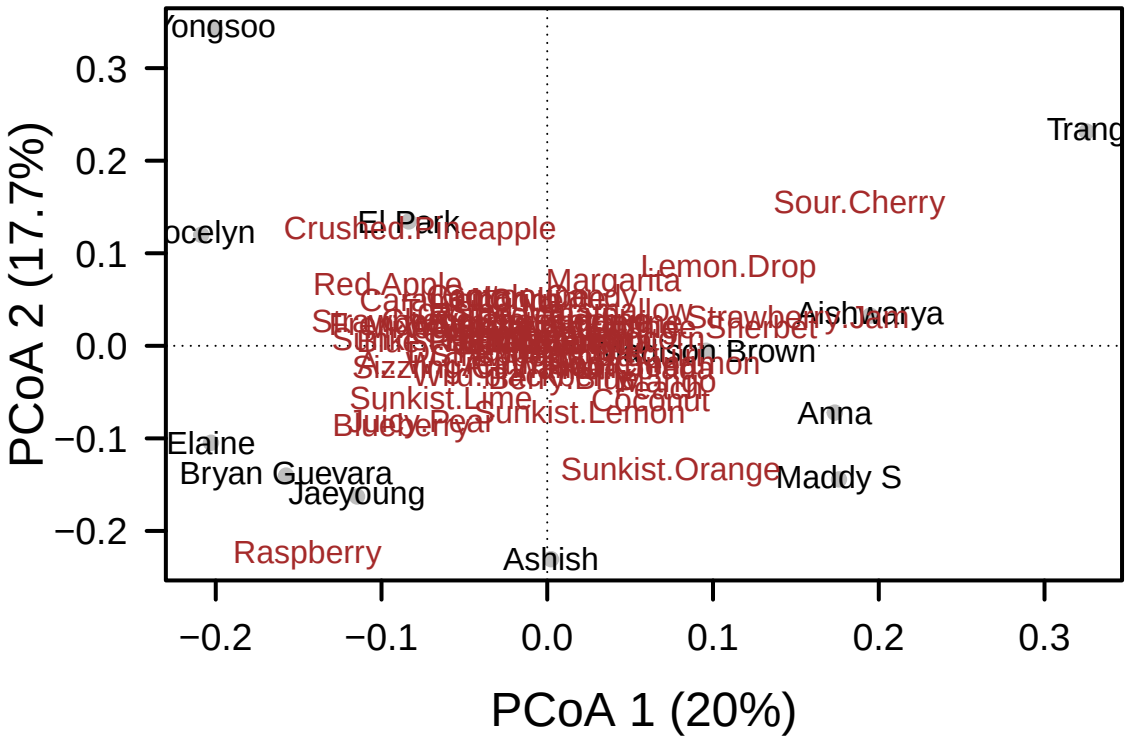
**Question 2:** Starting with the site-by-species matrix, visualize beta diversity. In the code chunk below, conduct principal coordinates analyses (PCoA) using both an abundance- and incidence-based resemblance matrix. Plot the sample scores in species space using different colors, symbols, or labels. Which "species" are contributing the patterns in the ordinations? How does the choice of resemblance matrix affect your interpretation?

```r
# Creating a distance-based resemblance matrix
jelly.dj <- vegdist(jelly, method = "jaccard", binary = TRUE) #jaccard
jelly.db <- vegdist(jelly, method = "bray") #bray-curtis

# Load the function for identifying influential species
source("/cloud/project/QB2025_Huang/Week3-Beta/bin/spec.scores.function.R")
jellyREL <- jelly
  for(i in 1:nrow(jelly)){
    jellyREL[i, ] = jelly[i, ]/sum(jelly[i, ])
  }

# Abundance-based PCoA ----
```

```r
jelly.db.pcoa <- cmdscale(jelly.db, eig = TRUE, k = 3)
db.explainvar1 <- round(jelly.db.pcoa$eig[1]/sum(jelly.db.pcoa$eig), 3)*100
db.explainvar2 <- round(jelly.db.pcoa$eig[2]/sum(jelly.db.pcoa$eig), 3)*100
db.explainvar3 <- round(jelly.db.pcoa$eig[3]/sum(jelly.db.pcoa$eig), 3)*100
db.sum.eig <- sum(db.explainvar1, db.explainvar2, db.explainvar3)
jelly.db.pcoa <- add.spec.scores.class(jelly.db.pcoa, jellyREL, method = "pcoa.scores") # calculate the
```

```
## Warning in cor(comm[, i], ordiscores[, j]): the standard deviation is zero
## Warning in cor(comm[, i], ordiscores[, j]): the standard deviation is zero
## Warning in cor(comm[, i], ordiscores[, j]): the standard deviation is zero
```

```r
# Plot the Abundance-based PCoA plot
par(mar = c(5,5,2,2) +0.3)
plot(jelly.db.pcoa$points[, 1], jelly.db.pcoa$points[, 2], xlim = range(jelly.db.pcoa$points[, 1], jelly
     xlab = paste("PCoA 1 (", db.explainvar1, "%)", sep = ""),
     ylab = paste("PCoA 2 (", db.explainvar2, "%)", sep = ""),
     pch = 16, cex = 2.0, type = "n", cex.lab = 1.5,
     cex.axis = 1.2, axes = FALSE)
axis(side = 1, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
axis(side = 2, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
abline(h = 0, v = 0, lty = 3)
box(lwd = 2)
points(jelly.db.pcoa$points[, 1], jelly.db.pcoa$points[, 2],
       pch = 19, cex = 1, bg = "gray", col = "gray")
text(jelly.db.pcoa$points[, 1], jelly.db.pcoa$points[, 2],
     labels = row.names(jelly.db.pcoa$points))
text(jelly.db.pcoa$cproj[, 1], jelly.db.pcoa$cproj[, 2],
     labels = row.names(jelly.db.pcoa$cproj), col = "brown")   #add species scores
```

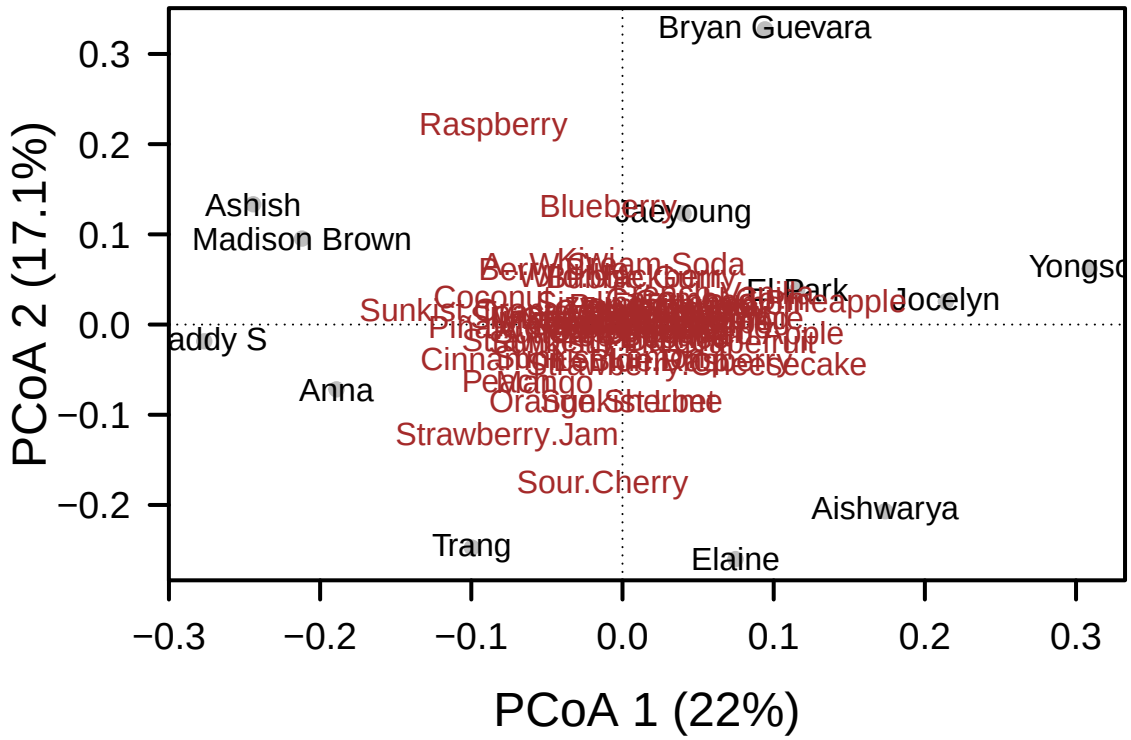



```r
# Incidence-based PCoA ----
jelly.dj.pcoa <- cmdscale(jelly.dj, eig = TRUE, k = 3)
```

```r
dj.explainvar1 <- round(jelly.dj.pcoa$eig[1]/sum(jelly.dj.pcoa$eig), 3)*100
dj.explainvar2 <- round(jelly.dj.pcoa$eig[2]/sum(jelly.dj.pcoa$eig), 3)*100
dj.explainvar3 <- round(jelly.dj.pcoa$eig[3]/sum(jelly.dj.pcoa$eig), 3)*100
dj.sum.eig <- sum(dj.explainvar1, dj.explainvar2, dj.explainvar3)
jelly.dj.pcoa <- add.spec.scores.class(jelly.dj.pcoa, jellyREL, method = "pcoa.scores") # calculate the
```

```
## Warning in cor(comm[, i], ordiscores[, j]): the standard deviation is zero
## Warning in cor(comm[, i], ordiscores[, j]): the standard deviation is zero
## Warning in cor(comm[, i], ordiscores[, j]): the standard deviation is zero
```

```r
# Plot the Abundance-based PCoA plot
par(mar = c(5,5,2,2) +0.3)
plot(jelly.dj.pcoa$points[, 1], jelly.dj.pcoa$points[, 2],
     xlim = range(jelly.dj.pcoa$points[, 1], jelly.dj.pcoa$cproj[, 1], na.rm = TRUE),
     ylim = range(jelly.dj.pcoa$points[, 2], jelly.dj.pcoa$cproj[, 2], na.rm = TRUE),
     xlab = paste("PCoA 1 (", dj.explainvar1, "%)", sep = ""),
     ylab = paste("PCoA 2 (", dj.explainvar2, "%)", sep = ""),
     pch = 16, cex = 2.0, type = "n", cex.lab = 1.5,
     cex.axis = 1.2, axes = FALSE)
axis(side = 1, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
axis(side = 2, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
abline(h = 0, v = 0, lty = 3)
box(lwd = 2)
points(jelly.dj.pcoa$points[, 1], jelly.dj.pcoa$points[, 2],
       pch = 19, cex = 1, bg = "gray", col = "gray")
text(jelly.dj.pcoa$points[, 1], jelly.dj.pcoa$points[, 2],
     labels = row.names(jelly.dj.pcoa$points))
text(jelly.dj.pcoa$cproj[, 1], jelly.dj.pcoa$cproj[, 2],
     labels = row.names(jelly.dj.pcoa$cproj), col = "brown")   #add species scores
```




***Answer 2***: Comparing to the incidence-based resemblence PCoA, abundance-based one seemed

to have more samples that share similarities in community composition (e.g. Elaine, Bryan, and Jaeyoung's samples seem to be similar to each other's, and Aishwarya, Madison, and Anna's seem to be clustered together in terms of similarity). Incidence-based PCoA shows that the community species composition of the jelly beans are more spread out and diverse among the samples, except for Ashish and Madison's samples which seem to be similar. This may suggest that most samples have many rare species, and the species composition is also different among samples. We could also infer that the source community may be composed of a lot of rare species, which means the distribution may be skewed to the right.

**Question 3** Using the preference-profile matrix, determine the most popular jelly bean in the class using a control structure (e.g., for loop, if statement, function, etc).

```
jelly.pref <- read.csv(file = "./data/Jelly belly data! - Preference.csv", header = TRUE, row.names = 1)
jelly.pref[is.na(jelly.pref)] <- 0

max.sum <- 0
max.name <- ""
for(i in 1:ncol(jelly.pref)){
  col.sum <- sum(jelly.pref[,i])
  if(col.sum > max.sum){
    max.sum <- col.sum
    max.name <- colnames(jelly.pref)[i]
  }
}
print(max.name)
```

```
## [1] "Berry.Blue"
```

```
print(max.sum)
```

```
## [1] 7
```

***Answer 3***: The most popular jelly bean is Berry Blue, which recieved 7 preference among samples.

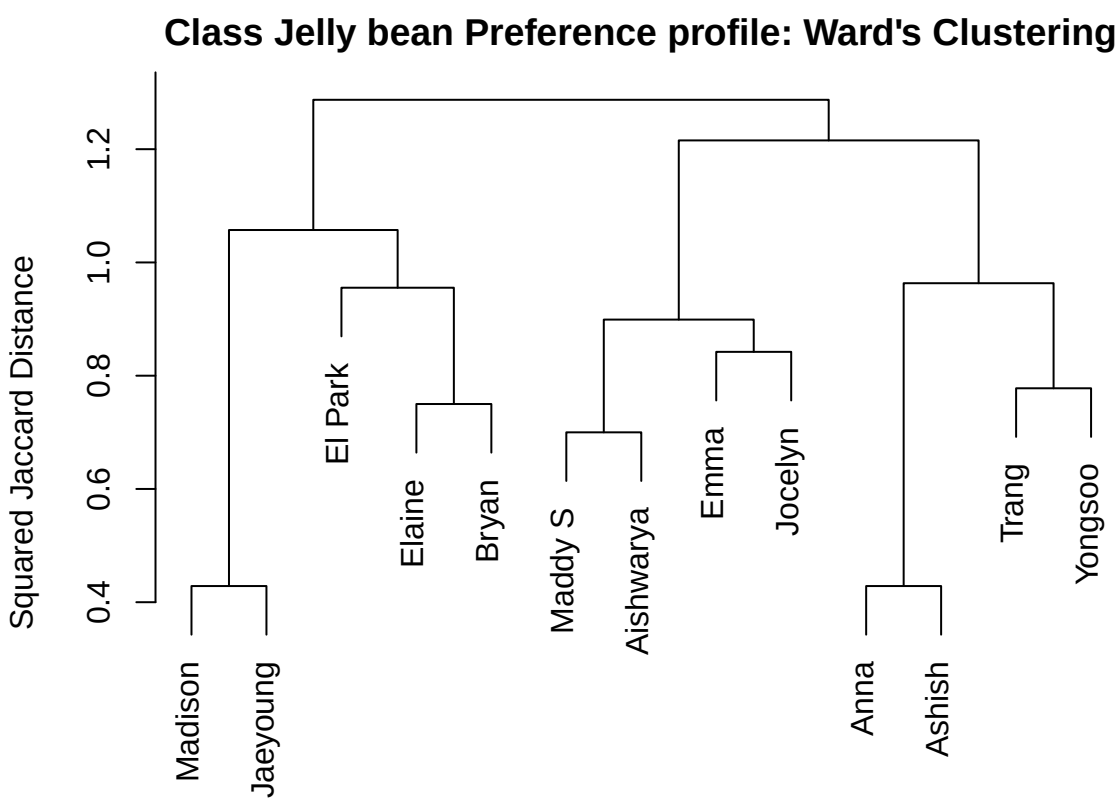
**Question 4** In the code chunk below, identify the student in QB who has a preference-profile that is most like yours. Quantitatively, how similar are you to your "jelly buddy"? Visualize the preference profiles of the class by creating a cluster dendogram. Label each terminal node (a.k.a., tip or "leaf") with the student's name or initials. Make some observations about the preference-profiles of the class.

```
pref.dj <- vegdist(jelly.pref, method = "jaccard", binary = TRUE, upper = TRUE, diag = TRUE)
pref.ward <- hclust(pref.dj, method = "ward.D2")
print(pref.dj)
```

```
##             El Park     Trang   Madison      Emma   Maddy S      Anna  Jaeyoung
## El Park   0.0000000 0.9090909 0.9090909 0.9444444 1.0000000 1.0000000 0.8181818
## Trang     0.9090909 0.0000000 0.8888889 0.7857143 1.0000000 0.8750000 0.9000000
## Madison   0.9090909 0.8888889 0.0000000 0.8666667 1.0000000 1.0000000 0.4285714
## Emma      0.9444444 0.7857143 0.8666667 0.0000000 0.8666667 0.8571429 0.8000000
## Maddy S   1.0000000 1.0000000 1.0000000 0.8666667 0.0000000 0.8750000 1.0000000
## Anna      1.0000000 0.8750000 1.0000000 0.8571429 0.8750000 0.0000000 0.8888889
## Jaeyoung  0.8181818 0.9000000 0.4285714 0.8000000 1.0000000 0.8888889 0.0000000
## Elaine    0.9166667 0.7777778 0.7777778 0.8750000 1.0000000 1.0000000 0.6666667
## Jocelyn   0.9375000 0.9285714 0.8461538 0.8421053 0.9285714 1.0000000 0.8571429
## Bryan     0.9000000 1.0000000 1.0000000 0.9333333 1.0000000 1.0000000 1.0000000
## Aishwarya 0.9285714 0.9166667 0.9166667 0.7500000 0.7000000 0.9090909 0.9230769
## Yongsoo   1.0000000 0.7777778 0.9000000 0.9411765 0.9000000 0.7500000 0.9090909
## Ashish    0.9230769 0.9090909 0.9090909 0.8823529 0.8000000 0.4285714 0.8181818
```

```
##              Elaine   Jocelyn     Bryan Aishwarya   Yongsoo    Ashish
## El Park   0.9166667 0.9375000 0.9000000 0.9285714 1.0000000 0.9230769
## Trang     0.7777778 0.9285714 1.0000000 0.9166667 0.7777778 0.9090909
## Madison   0.7777778 0.8461538 1.0000000 0.9166667 0.9000000 0.9090909
## Emma      0.8750000 0.8421053 0.9333333 0.7500000 0.9411765 0.8823529
## Maddy S   1.0000000 0.9285714 1.0000000 0.7000000 0.9000000 0.8000000
## Anna      1.0000000 1.0000000 1.0000000 0.9090909 0.7500000 0.4285714
## Jaeyoung  0.6666667 0.8571429 1.0000000 0.9230769 0.9090909 0.8181818
## Elaine    0.0000000 0.7692308 0.7500000 1.0000000 0.9090909 1.0000000
## Jocelyn   0.7692308 0.0000000 0.9230769 0.8000000 0.8571429 1.0000000
## Bryan     0.7500000 0.9230769 0.0000000 1.0000000 1.0000000 1.0000000
## Aishwarya 1.0000000 0.8000000 1.0000000 0.0000000 1.0000000 0.9285714
## Yongsoo   0.9090909 0.8571429 1.0000000 1.0000000 0.0000000 0.7000000
## Ashish    1.0000000 1.0000000 1.0000000 0.9285714 0.7000000 0.0000000
```

```
par(mar = c(1,5,2,2) +0.1)
plot(pref.ward, main = "Class Jelly bean Preference profile: Ward's Clustering", ylab = "Squared Jaccar
```




***Answer 4***: From the dissimilarity matrix generated from jaccard index, my "jelly buddy" is Elaine, who I shared the least pairwise dissimilarity (0.769) with. However, the cluster dendogram shows that my preference profile is clustered with Emma's because of similarity. The class is seperated into half based on their clustered similarities: Madison, Jaeyoung, El, Elaine and Bryan are grouped into one larger cluster, and the rest of the class into another.

## SUBMITTING YOUR ASSIGNMENT

Use Knitr to create a PDF of your completed `7.DiversitySynthesis_Worksheet.Rmd` document, push it to GitHub, and create a pull request. Please make sure your updated repo includes both the pdf and RMarkdown files.

Unless otherwise noted, this assignment is due on **Wednesday, February 19$^{\text{th}}$, 2025 at 12:00 PM (noon)**.