7. Worksheet: Diversity Synthesis

Jaeyoung Yoo; Z620: Quantitative Biodiversity, Indiana University

18 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.

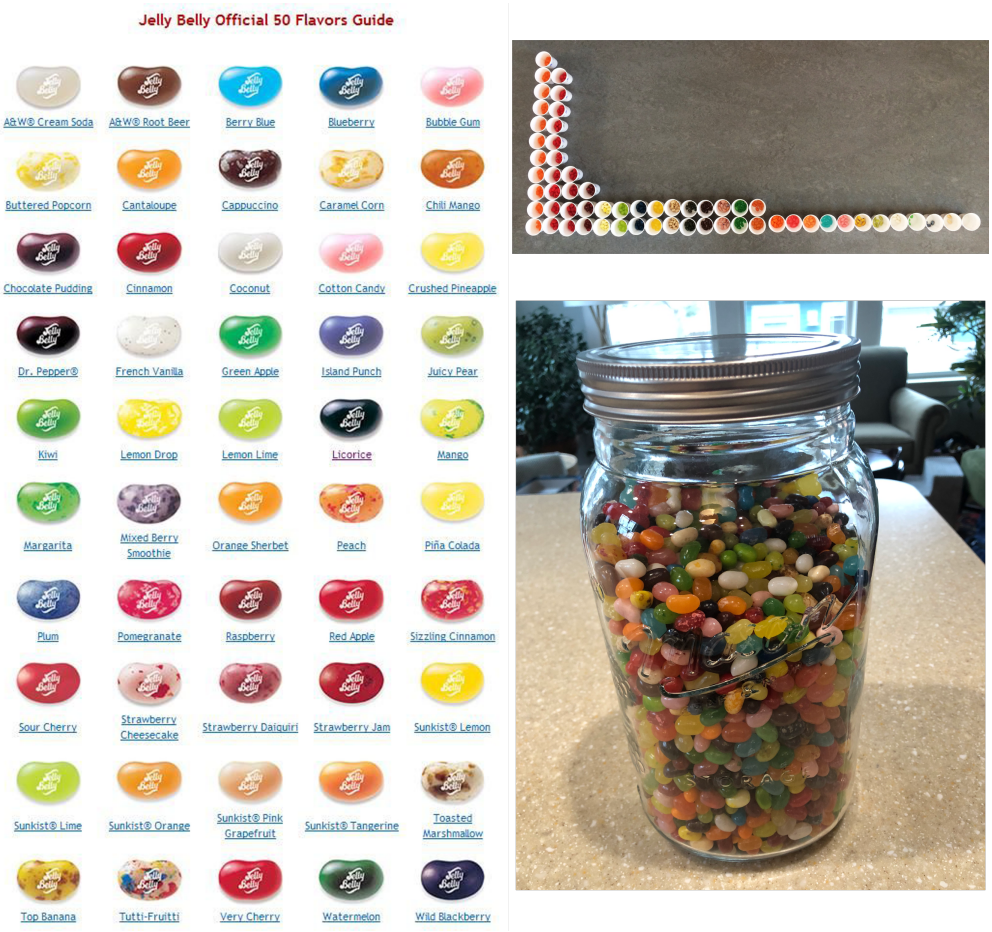


Figure 1: **Left**: taxonomic key, **Top right**: rank abundance distribution, **Bottom right**: source community

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_Yoo/Week5-Confection"
```

```
setwd("/cloud/project/QB2025_Yoo/Week5-Confection")

package.list <- c("vegan", "tidyverse", "ggplot2", "dplyr", "broom", "vegan", "ade4", "viridis", "gplots
for (package in package.list) {
  if (!require(package, character.only = TRUE, quietly = TRUE)) {
    install.packages(package)
  }
  library(c(package), character.only = TRUE)
}
```

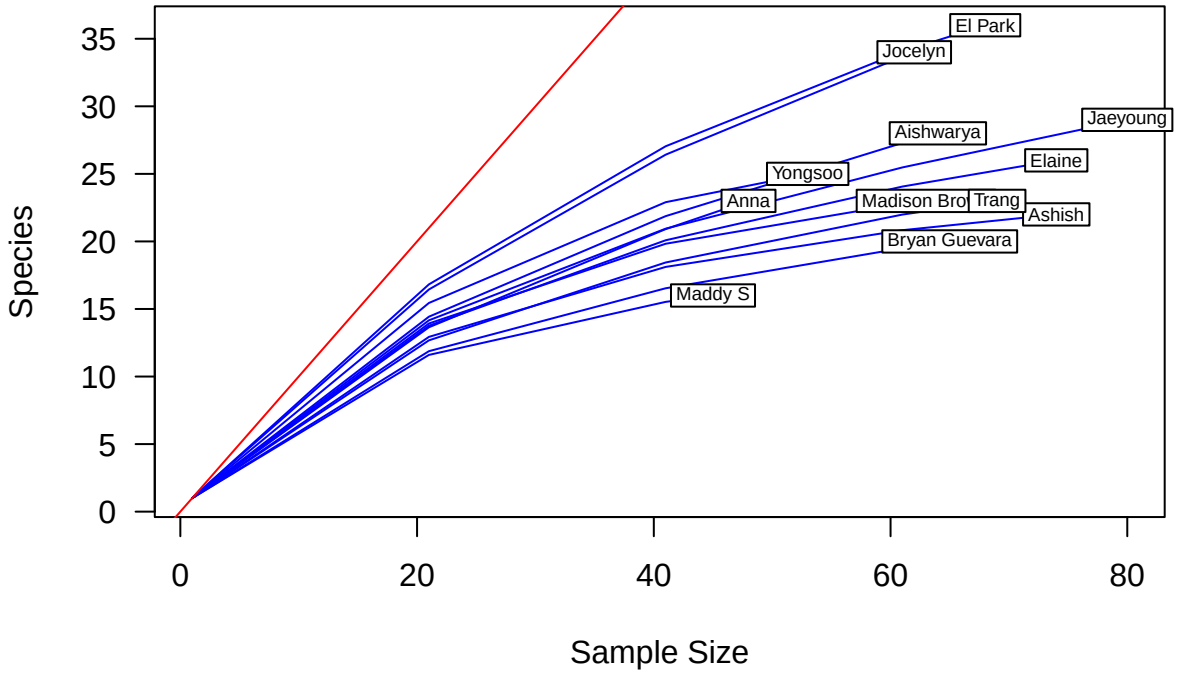
```
## This is vegan 2.6-8
```

```
## -- 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 error
##
## Attaching package: 'gplots'
##
##
## The following object is masked from 'package:stats':
##
##     lowess
```

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.

```r
dat <- read.csv(file = "./jelly_SbyS.csv", header = TRUE, row.names = 1)
prefer <- read.csv(file = "./jelly_preference.csv", header = TRUE, row.names = 1)

jelly <- specnumber(dat)
min.N <- min(rowSums(dat))
S.rarefy <- rarefy(x = dat, sample = min.N, se =TRUE)
rarecurve(x = dat, step = 20, col = "blue", cex = 0.6, las = 1)
abline(0, 1, col = "red")
text(1500, 1500, "1:1", pos = 2, col = "red")
```



Answer 1: The individual sampling efforts (Blue lines) capture less species than collective efforts (Red line). Increasing number of individuals may capture more species. Also, there are variations in the number of species captured in each individuals, Maddy captured lowest number of species while El captured the highest number of species.

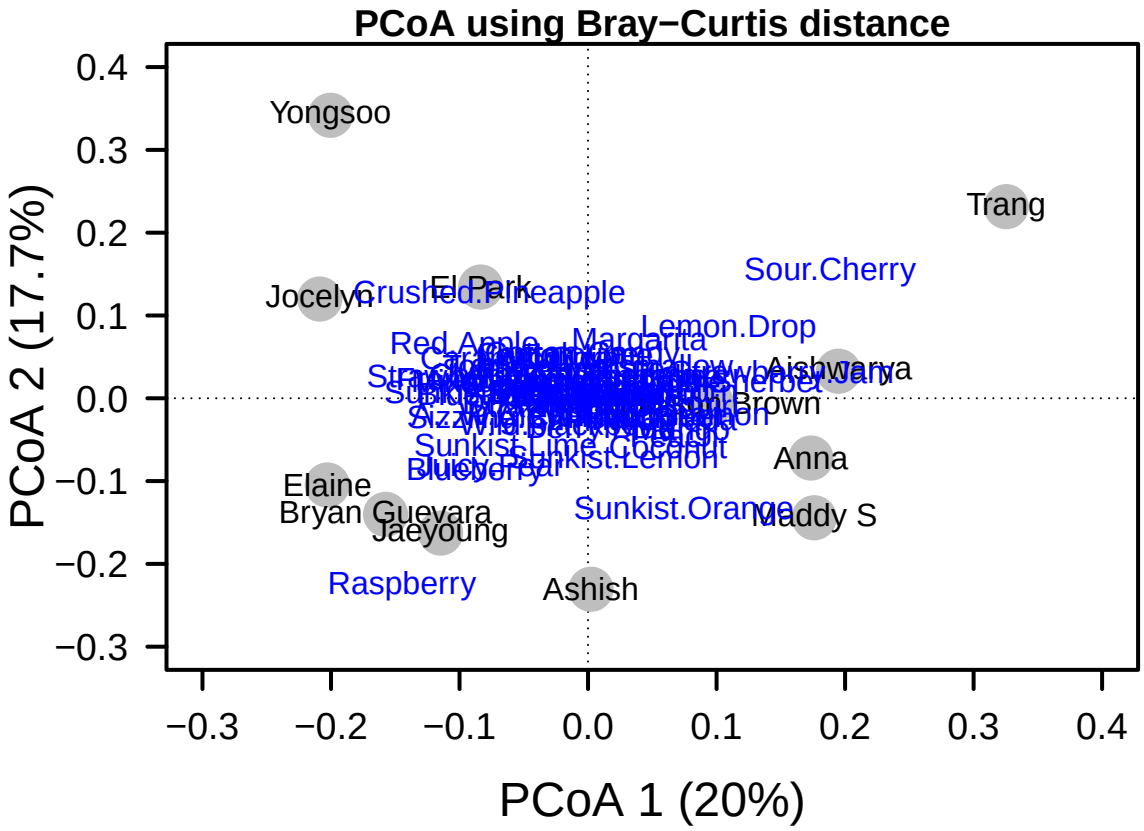
Question 2: Starting with the site-by-species matrix, visualize beta diversity. In the code chunk below, conduct principal coordinates analyses (PCoA) using both an abundance- and incidence-based resemblance matrix. Plot the sample scores in species space using different colors, symbols, or labels. Which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

```r
# Calculate Jaccard (Incidence-based)
jelly.dj <- vegdist(dat, method = "jaccard", binary = TRUE)

# Calculate Bray-Curtis (Abundance-based)
jelly.db <- vegdist(dat, method = "bray", diag = TRUE)

## PCoA using Bray-Curtis
jelly.pcoa1 <- cmdscale(jelly.db, eig = TRUE, k = 3)

explainvar1 <- round(jelly.pcoa1$eig[1] / sum(jelly.pcoa1$eig), 3) * 100
explainvar2 <- round(jelly.pcoa1$eig[2] / sum(jelly.pcoa1$eig), 3) * 100
explainvar3 <- round(jelly.pcoa1$eig[3] / sum(jelly.pcoa1$eig), 3) * 100
```

```r
sum.eig <- sum(explainvar1, explainvar2, explainvar3)

# Define plot parameters
par(mar = c(5, 5, 1, 2) + 0.1)

# Initiate plot
plot(jelly.pcoa1$points[, 1], jelly.pcoa1$points[, 2],
     xlab = paste("PCoA 1 (", explainvar1, "%)", sep = ""),
     xlim = c(-0.3, 0.4),
     ylim = c(-0.3, 0.4),
     ylab = paste("PCoA 2 (", explainvar2, "%)", sep = ""),
     pch = 16, cex = 2.0, type = "n", cex.lab = 1.5, cex.axis = 1.2, axes = FALSE)

# Add axes
axis(side = 1, labels = TRUE, lwd.ticks = 2, cex.axis = 1.2, las = 1)
axis(side = 2, labels = TRUE, lwd.ticks = 2, cex.axis = 1.2, las = 1)
abline(h = 0, v = 0, lty = 3)
box(lwd = 2)
title(main = "PCoA using Bray-Curtis distance")

# Add points & labels
points(jelly.pcoa1$points[, 1], jelly.pcoa1$points[, 2], pch = 19, cex = 3, bg = "gray", col = "gray")
text(jelly.pcoa1$points[, 1], jelly.pcoa1$points[, 2], labels = row.names(jelly.pcoa1$points))

# First we calculate the relative abundances of each species at each site
jellyREL <- dat
for (i in 1:nrow(dat)) {
    jellyREL[i, ] = dat[i, ] / sum(dat[i, ])
    }

# Now, we use this information to calculate and add species scores
jelly.pcoa1 <- add.spec.scores(jelly.pcoa1, jellyREL, method = "pcoa.scores")
```

```
## 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
text(jelly.pcoa1$cproj[, 1], jelly.pcoa1$cproj[, 2],
    labels = row.names(jelly.pcoa1$cproj), col = "blue")
```

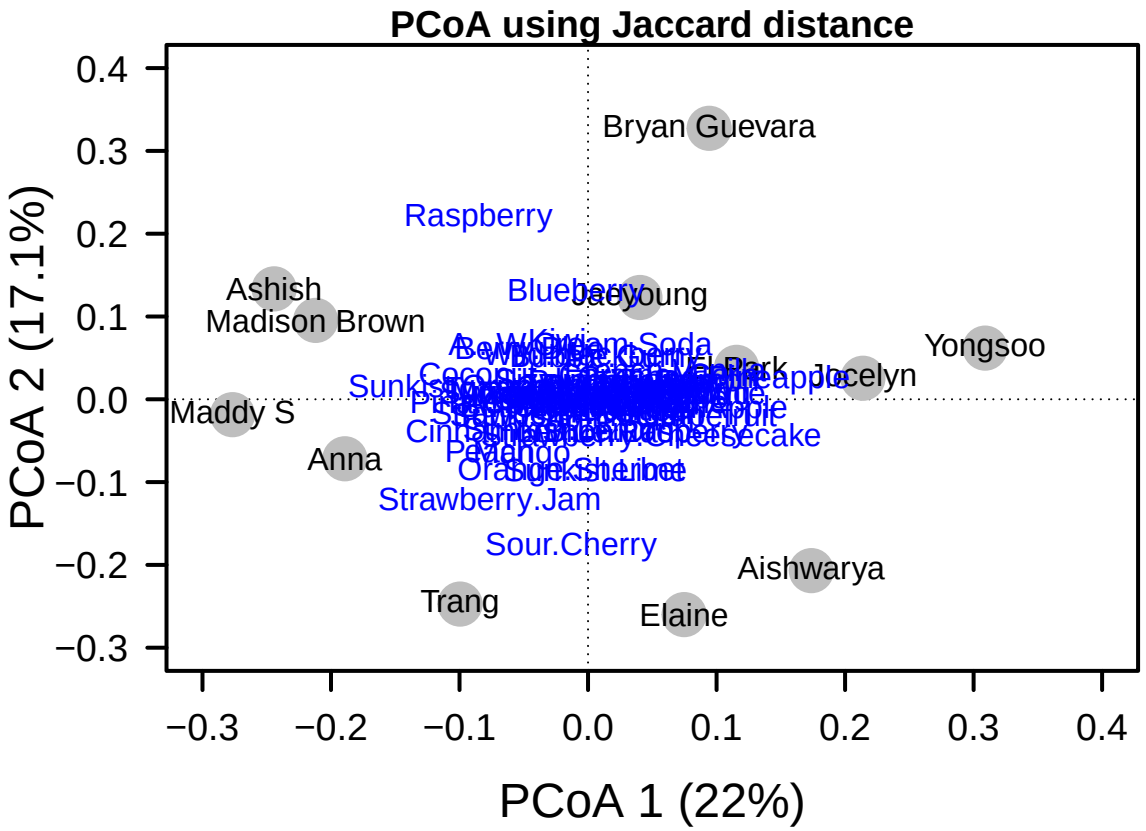



```
## PCoA using Bray-Curtis
jelly.pcoa1 <- cmdscale(jelly.dj, eig = TRUE, k = 3)

explainvar1 <- round(jelly.pcoa1$eig[1] / sum(jelly.pcoa1$eig), 3) * 100
explainvar2 <- round(jelly.pcoa1$eig[2] / sum(jelly.pcoa1$eig), 3) * 100
explainvar3 <- round(jelly.pcoa1$eig[3] / sum(jelly.pcoa1$eig), 3) * 100
sum.eig <- sum(explainvar1, explainvar2, explainvar3)

# Define plot parameters
par(mar = c(5, 5, 1, 2) + 0.1)

# Initiate plot
plot(jelly.pcoa1$points[, 1], jelly.pcoa1$points[, 2],
     xlab = paste("PCoA 1 (", explainvar1, "%)", sep = ""),
     xlim = c(-0.3, 0.4),
     ylim = c(-0.3, 0.4),
     ylab = paste("PCoA 2 (", explainvar2, "%)", sep = ""),
     pch = 16, cex = 2.0, type = "n", cex.lab = 1.5, cex.axis = 1.2, axes = FALSE)

# Add axes
axis(side = 1, labels = TRUE, lwd.ticks = 2, cex.axis = 1.2, las = 1)
axis(side = 2, labels = TRUE, lwd.ticks = 2, cex.axis = 1.2, las = 1)
abline(h = 0, v = 0, lty = 3)
box(lwd = 2)
title(main = "PCoA using Jaccard distance")

# Add points & labels
```

```r
points(jelly.pcoa1$points[, 1], jelly.pcoa1$points[, 2], pch = 19, cex = 3, bg = "gray", col = "gray")
text(jelly.pcoa1$points[, 1], jelly.pcoa1$points[, 2], labels = row.names(jelly.pcoa1$points))

# First we calculate the relative abundances of each species at each site
jellyREL <- dat
for (i in 1:nrow(dat)) {
    jellyREL[i, ] = dat[i, ] / sum(dat[i, ])
    }

# Now, we use this information to calculate and add species scores
jelly.pcoa1 <- add.spec.scores(jelly.pcoa1, jellyREL, method = "pcoa.scores")
```

```
## 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
text(jelly.pcoa1$cproj[, 1], jelly.pcoa1$cproj[, 2],
    labels = row.names(jelly.pcoa1$cproj), col = "blue")
```



Answer 2: PCoA plot based on Bray-Curtis distance (abundance-based matrix) shows that species such as Sour cherry, Raspberry and Crushed Pineapple are dispersed from main cluster, which means that they are contributing to the patterns. On the other hand, PCoA plot based on Jaccard distance (incidence-based matrix) shows that species such as Sour cherry and Raspberry are contributing to the patterns. However, the PCoA plots based on different distances show different results that pattern of sites clustered are quite different. For example, in PCoA plot based on Bray-Curtis distance, Elaine, Bryan and Jaeyoung are clustered in third quadrant, while they are dispersed in first and fourth quadrant in PCoA plot based on Jaccard distance.

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).

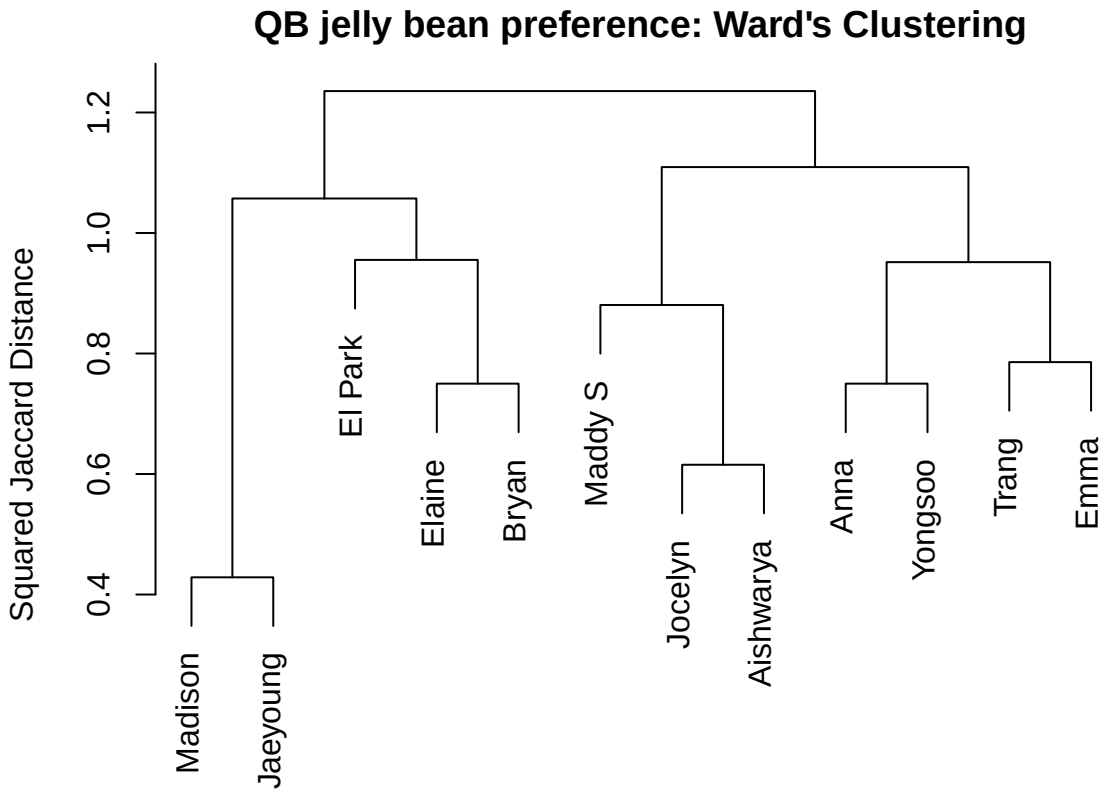
```
# Variables
max_count <- 0
best_flavor <- ""

# Iterate through each flavor
for (flavor in colnames(prefer)) {
  count_ones <- sum(prefer[[flavor]] == 1, na.rm = TRUE)
  # Check if this flavor has the most 1
  if (count_ones > max_count) {
    max_count <- count_ones
    best_flavor <- flavor
  }
}

# Print the flavor with the most 1s
cat("The most popular jelly bean flavor:", best_flavor, "with", max_count, "votes.")
```

```
## The most popular jelly bean flavor: Berry.Blue with 7 votes.
```

Answer 3: The most popular jelly bean flavor is Berry Blue with 7 votes.

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.

```
# Replace NA values with 0 in 'prefer' dataset
prefer[is.na(prefer)] <- 0

# Calculate Jaccard
prefer.dj <- vegdist(prefer, method = "jaccard", binary = TRUE)

# Perform cluster analysis
prefer.ward <- hclust(prefer.dj, method = "ward.D2")

# Plot cluster
par(mar = c(1, 5, 2, 2) + 0.1)
plot(prefer.ward, main = "QB jelly bean preference: Ward's Clustering",
     ylab = "Squared Jaccard Distance")
```



Answer 4: Madison is my jelly buddy, and our preference is quite similar because our preferences are clustered in lowest threshold of about 0.4 of squared Jaccard distance. Elain and Bryan, Jocelyn and Aishwarya, Anna and Yongsoo, Trang and Emma are clustered together as jelly buddy.

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 19th, 2025 at 12:00 PM (noon)**.