Clusters and Heatmaps

Jeff Oliver

15 August, 2017

Want to combine the visualization of quantitative data with clustering algorithms? Unsatisfied with the options provided by the base R packages? In this hands-on workshop, we'll use the `ggendro` package for R to make publication-quality graphics.

- Consider the `ggdendro` package? See: https://stackoverflow.com/questions/6673162/reproducing-lattice-dendrogram-graph
- ggdendro page

0. What are the data? (the otter data?)
1. Clustering approach (e.g. `hclust`)
2. Drawing just a cluster
3. Drawing just a heatmap (`geom_tile`); individuals on Y, measurement on X

Learning objectives

1. one
2. two
3. three

[DESCRIPTION OR MOTIVATION; 2-4 sentences that would be used for an announcement]

Getting started

Start by creating a new project in RStudio and creating two folders we'll use to organize our efforts. The two folders should be `data` and `output` and will store. . . data and output.

```
dir.create("data")
dir.create("output")
```

We will also need a few R packages that are not included in the standard distribution of R. Those packages are `ggplot2`, `ggdendro`, `reshape2`, and `grid` and can be installed with the `install.packages()` function. Note these packages need only be installed once on your machine.

```
install.packages("ggplot2")
install.packages("ggdendro")
install.packages("reshape2")
install.packages("ggplot2")
```

Finally, download data file from https://jcoliver.github.io/learn-r/data/otter-mandible-data.csv or http://tinyurl.com/otter-data-csv (the latter just re-directs to the former). These data are a subset of those used in a study on skull morphology and diet specialization in otters doi: 10.1371/journal.pone.0143236. **Save this file in the `data` folder we just created**.

Data preparation

For any work that we do, we want to record all the steps we take, so instead of typing commands directly into the R console, we keep all our work in an R script. These scripts are just text files with R commands; by convention, we start the script with a brief bit of information about what the script does.

```
# Cluster & heatmap on otter data
# Jeff Oliver
# jcoliver@email.arizona.edu
# 2017-08-15

rm(list = ls()) # clean up the environment

################################################################################

# Read in data
otter <- read.csv(file = "data/otter-mandible-data.csv", header = TRUE)
```

If we look at the first few rows of data, we see the first three columns have information about the specimen from which the measurements were taken, while columns 4-9 have data for six mandible measurements (`m1` - `m6`):

```
head(otter)
```

```
##        species museum accession     m1     m2     m3     m4     m5     m6
## 1 A. cinerea    AMNH    101458 15.100 27.790 21.885 13.010 10.500 61.635
## 2 A. cinerea    AMNH    101461 12.740 26.750 20.265 13.255  8.340 59.370
## 3 A. cinerea    AMNH    101466 12.425 25.915 20.735 12.300  9.430 56.270
## 4 A. cinerea    AMNH    101635 13.400 28.030 22.075 10.580 10.455 58.080
## 5 A. cinerea    AMNH    101459 14.400 26.160 21.385 12.100  9.600 58.635
## 6 A. cinerea    AMNH    101462 14.525 29.020 22.305 11.905 11.070 60.655
```

For our purposes, we only want to look at the data for two species, so we subset the data, including only those rows that have either "A. cinerea" or "L. canadensis" in the `species` column. Note, this *does not* alter the data in the original file we read into R; it only alters the data object `otter` currently in R's memory.

```
two.species <- c("A. cinerea", "L. canadensis")
otter <- otter[otter$species %in% two.species, ]
```

We also want to use the specimen accession number for categorizing specimens, so we convert this column to be of type `factor`.

```
otter$accession <- factor(otter$accession)
```

Missing data can cause problems in downstream analyses, so we will just remove any rows that have missing data. Here we replace the original data object `otter` with one in which there are no missing values.

```
otter <- na.omit(otter)
```

Because R *does not* automatically re-number the rows when we drop those with `NA` values, we can force re-numbering via:

```
rownames(otter) <- NULL
```

The last thing we need to do is scale the data variables (columns 4-9) so measurements have a mean value of 0 and unit variance.

```
otter.scaled <- otter
otter.scaled[, c(4:9)] <- scale(otter.scaled[, 4:9])
```

Clustering

To make our figure, we will build the two plots (the cluster diagram and the heatmap) separately, then use the `grid` framework to put them together. We start by making the dendrogram (or cluster).

```r
# Run clustering
library("ggdendro")
otter.matrix <- as.matrix(otter.scaled[, -c(1:3)])
rownames(otter.matrix) <- otter.scaled$accession
otter.dendro <- as.dendrogram(hclust(d = dist(x = otter.matrix)))
otter.dendro.data <- dendro_data(otter.dendro)
# otter.order <- order.dendrogram(otter.dendro)

# Create dendro
library("ggplot2")
# dendro.plot <- ggplot(data = segment(otter.dendro.data)) +
#   geom_segment(aes(x = x, y = y, xend = xend, yend=yend)) +
#   coord_flip()

dendro.plot <- ggdendrogram(data = otter.dendro, rotate = TRUE)

# Preview the plot
print(dendro.plot)
```
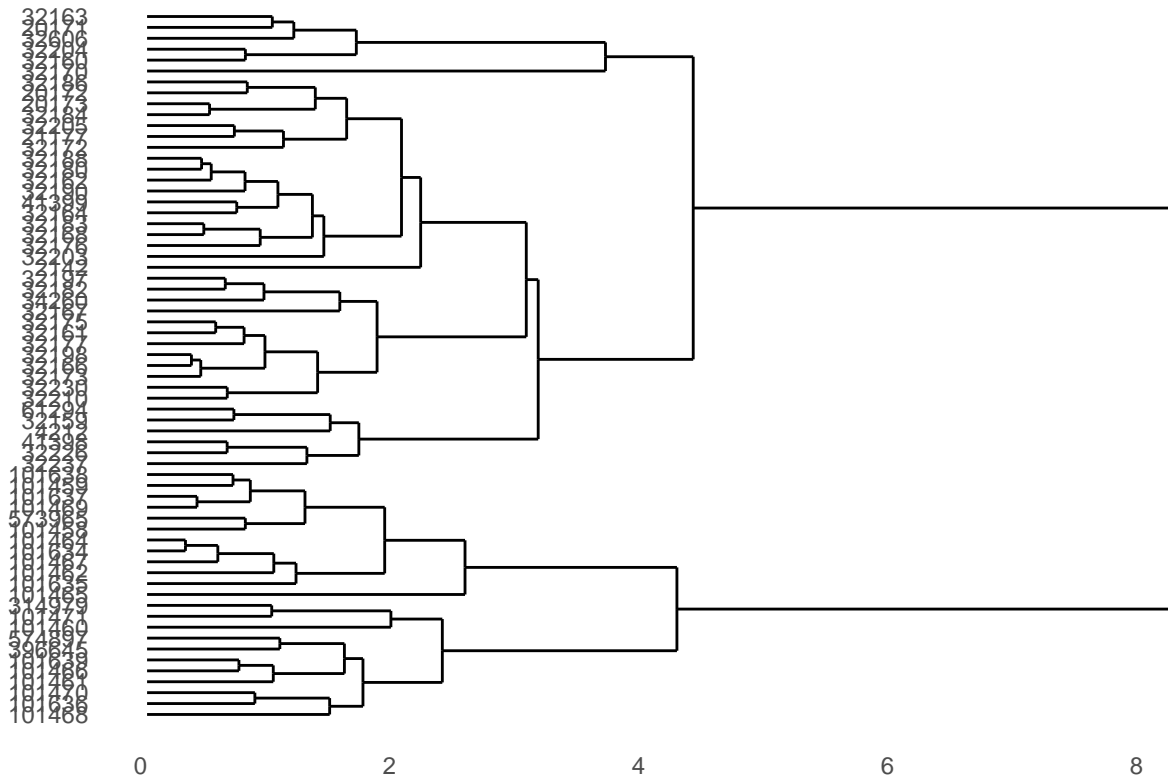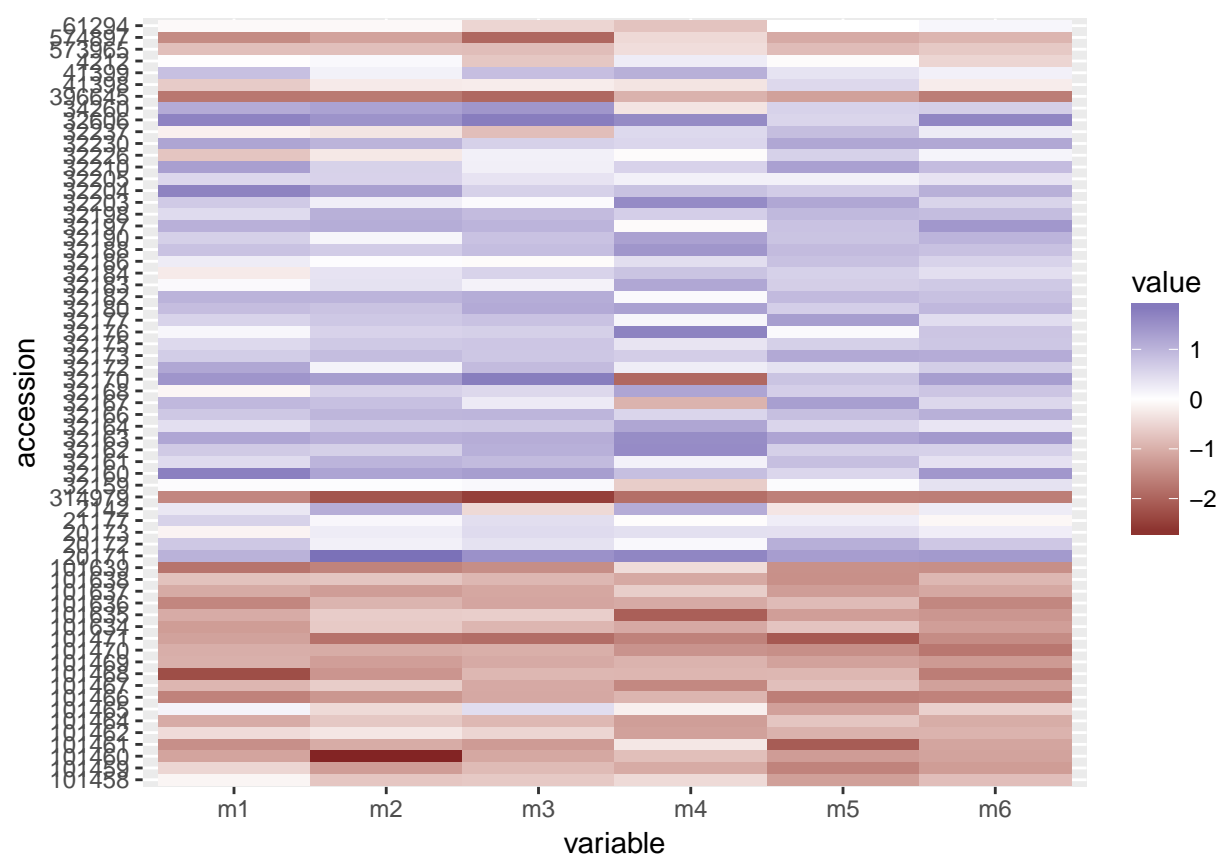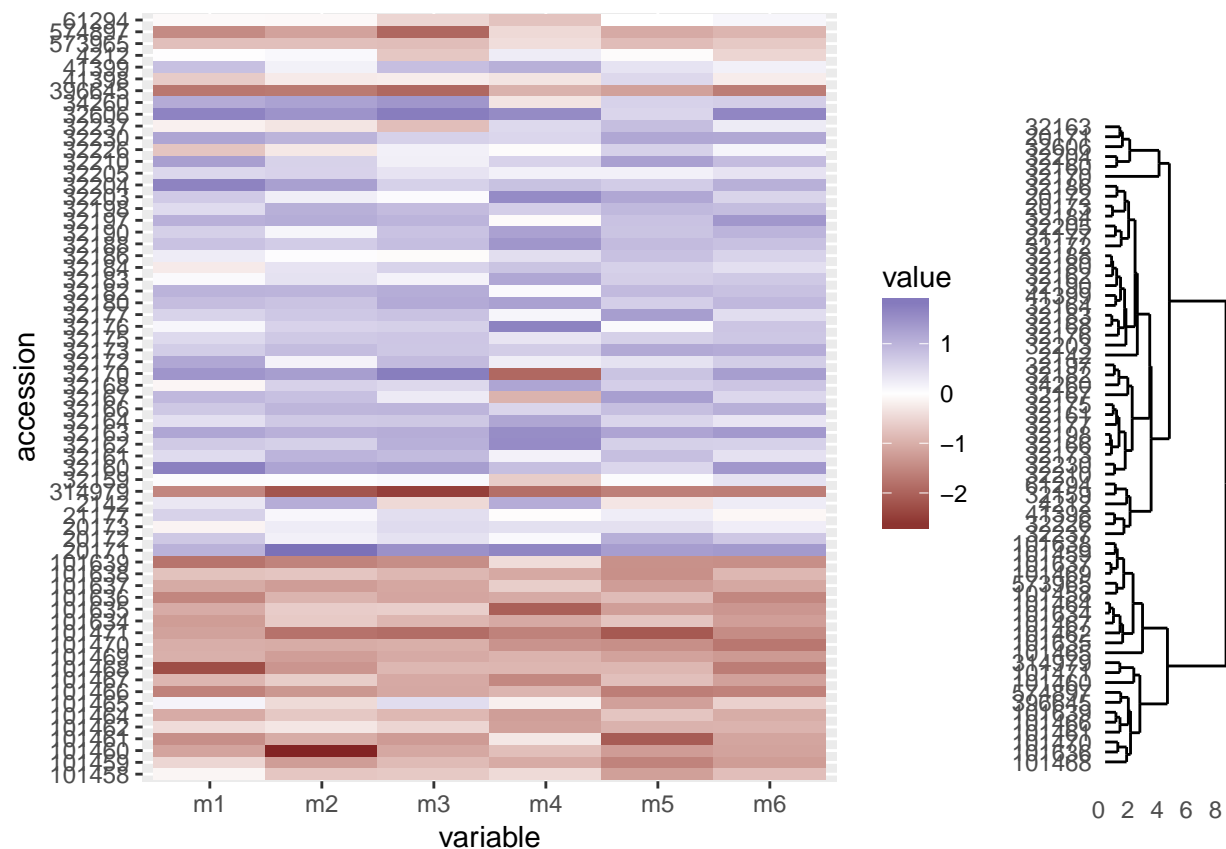
Heatmap

```r
# Heatmap
# Data wrangling
library("reshape2")
otter.long <- melt(otter.scaled, id = c("species", "museum", "accession"))
# Order the levels according to their position in the cluster
# otter.long$accession <- factor(x = otter.long$accession, levels = otter.scaled$accession[otter.order]
heatmap.plot <- ggplot(data = otter.long, aes(x = variable, y = accession)) +
  geom_tile(aes(fill = value)) +
  scale_fill_gradient2()

# Preview the heatmap
print(heatmap.plot)
```


Putting it all together

```r
library("grid")
grid.newpage()
print(heatmap.plot, vp = viewport(x = 0.4, y = 0.5, width = 0.8, height = 1.0))
print(dendro.plot, vp = viewport(x = 0.90, y = 0.445, width = 0.2, height = 0.92))
```

OK, that's a start, but there are a few things we need to address:

1.

Final-ish

```
otter <- read.csv(file = "data/otter-mandible-data.csv", header = TRUE)
two.species <- c("A. cinerea", "L. canadensis")
otter <- otter[otter$species %in% two.species, ]
otter <- na.omit(otter)
otter.scaled <- otter
otter.scaled[, c(4:9)] <- scale(otter.scaled[, 4:9])
otter.scaled$accession <- factor(otter.scaled$accession)

# Renumber rows
rownames(otter) <- NULL

# Run clustering
library("ggdendro")
otter.matrix <- as.matrix(otter.scaled[, -c(1:3)])
otter.dendro <- as.dendrogram(hclust(d = dist(x = otter.matrix)))
otter.dendro.data <- dendro_data(otter.dendro)
otter.order <- order.dendrogram(otter.dendro)
```

```r
# Create dendro
library("ggplot2")
dendro.plot <- ggplot(data = segment(otter.dendro.data)) +
  geom_segment(aes(x = x, y = y, xend = xend, yend=yend)) +
  coord_flip() +
  theme_dendro()

# Heatmap
# Data wrangling
library("reshape2")
otter.long <- melt(otter.scaled, id = c("species", "museum", "accession"))
# Order the levels according to their position in the cluster
otter.long$accession <- factor(x = otter.long$accession, levels = otter.scaled$accession[otter.order],
heatmap.plot <- ggplot(data = otter.long, aes(x = variable, y = accession)) +
  geom_tile(aes(fill = value)) +
  scale_fill_gradient2() +
  theme(legend.position = "top")

# All together
library("grid")
grid.newpage()
print(heatmap.plot, vp = viewport(x = 0.4, y = 0.5, width = 0.8, height = 1.0))
print(dendro.plot, vp = viewport(x = 0.90, y = 0.445, width = 0.2, height = 0.92))
```
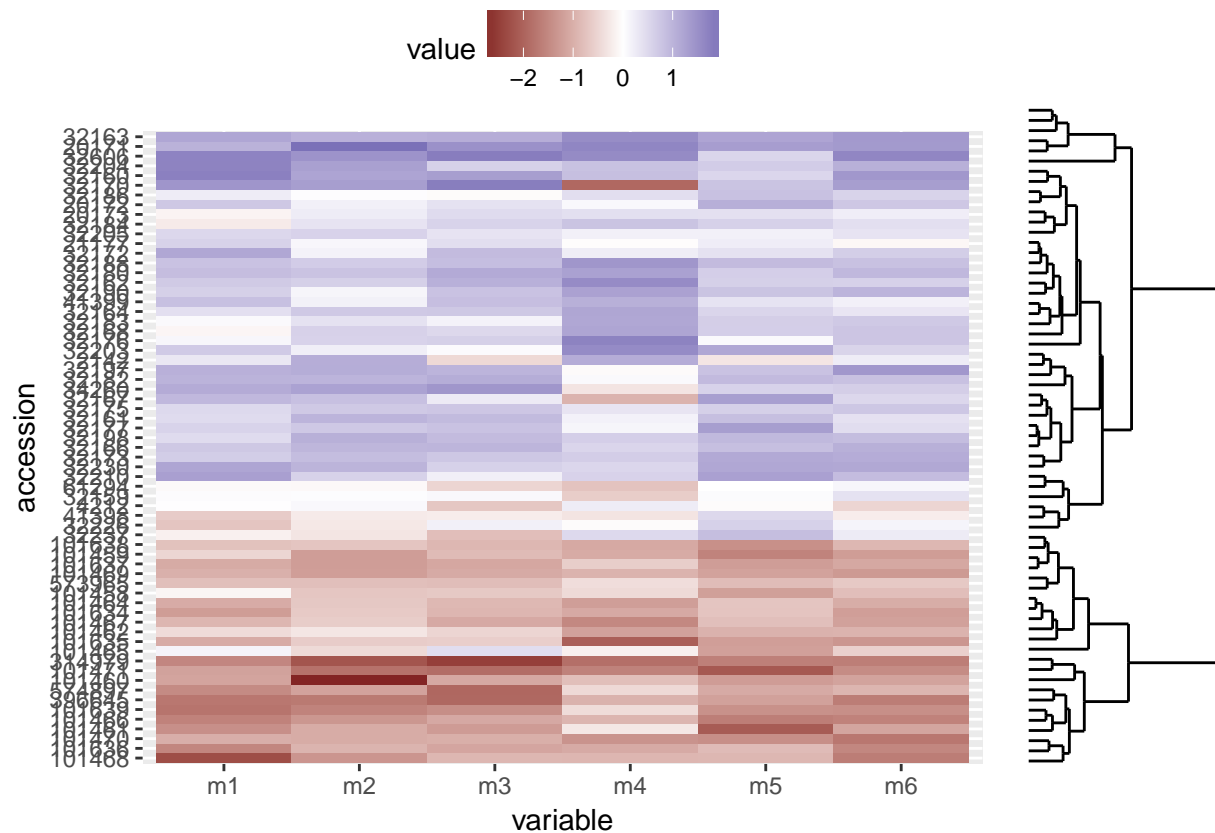
Additional resources

- resource one
- resource two

Back to learn-r main page

Questions? e-mail me at jcoliver@email.arizona.edu.