Stat7350: Plotting with ggplot2 (lesson 3)

AC Gerstein

2019-03-12

Learning Objectives

- Visualizing Interactions
- Mosaic plots
- How to evaluate figure quality?
- TidyTuesday figure comparison

Pre-analysis workflow: packages & data prep

Packages

```
suppressMessages(library(tidyverse))
suppressMessages(library(gridExtra))
suppressMessages(library(Hmisc)) #%nin%
```

Load Data

Load our survey data again. I'm going to save it as a shorter variable name this time.

```
sc <- read_csv("data_output/surveys_complete.csv")
```

```
## Parsed with column specification:
## cols(
##   record_id = col_double(),
##   month = col_double(),
##   day = col_double(),
##   year = col_double(),
##   plot_id = col_double(),
##   species_id = col_character(),
##   sex = col_character(),
##   hindfoot_length = col_double(),
##   weight = col_double(),
##   genus = col_character(),
##   species = col_character(),
##   taxa = col_character(),
##   plot_type = col_character()
## )
```

Interactions

Interactions are really common in ecology - it is often that case that "the effect of X on Y depends on M". We have aleady looked at interactions indirectly, when we've used different colours for `fill` or `facet_wrap`

to see whether different parts of our dataset (e.g., sex or species) responded differently to a given pair of factors (e.g., weight, hindfoot_length, year).

We're going to start with a fairly simple interaction plot that is used quite often to visualize interactions in biological data.

Let's think more about how animals might change within a year. Let's make a hypothesis: do you think that weight is typically higher in a summer month (e.g., January) or a winter month (e.g., July)?

If we look across the whole dataset:

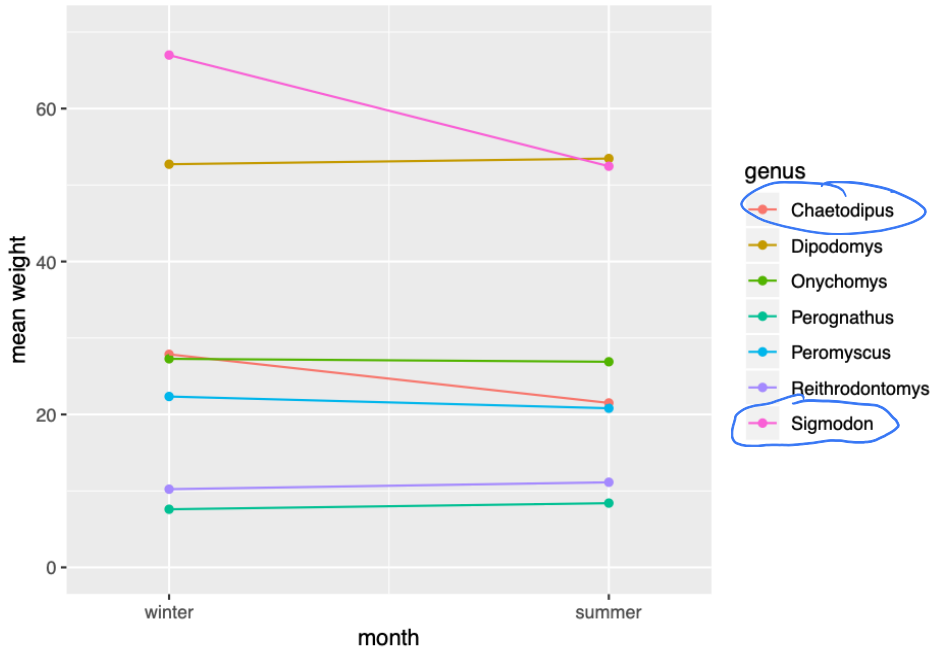
```
wt_month <- sc %>%
  filter(month %in% c(1, 7)) %>%
  group_by(month) %>%
  summarise(wt = mean(weight))

wt_month
```

```
## # A tibble: 2 x 2
##   month    wt
##   <dbl> <dbl>
## 1     1  41.5
## 2     7  38.2
```

summer mean weight is higher in this case

It looks like it's (a very little bit) higher in the summer than the winter. But maybe this depends on species?

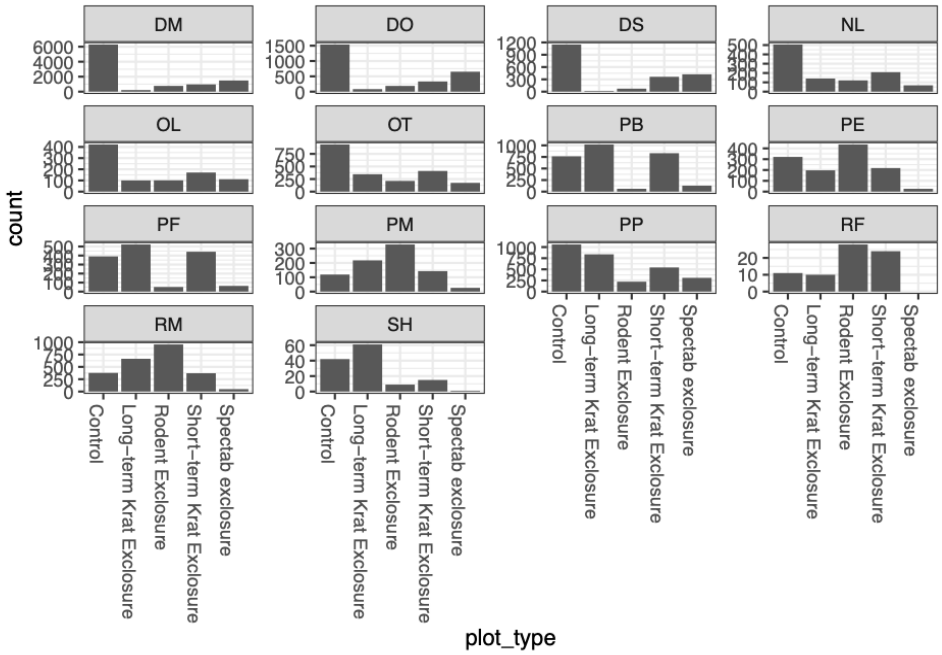
```
t <- sc %>%
  filter(genus %nin% c("Neotoma"), month %in% c(1, 7)) %>%
  group_by(month, genus) %>%
  summarise(wt = mean(weight), wt.se = sd(weight), na.rm=TRUE)

ggplot(t, aes(x = month, y =wt, colour = genus)) +
  geom_point() +
  geom_line() +
  scale_x_continuous(name = "month", breaks = c(1, 7), labels= c("winter", "summer"), limits = c(0, 8))
  scale_y_continuous(name = "mean weight", limits = c(0, 70))
```



Indeed, based on the crossing lines it looks like the effect of month on weight does depend on species.

Let's go back to our investigation of whether plot type influences the species that are caught.

```
ggplot(data = sc, mapping = aes(x = plot_type)) +
    geom_bar() +
    facet_wrap(~ species_id,  scales="free_y") +
    theme_bw() +
    theme(axis.text.x=element_text(angle = -90, hjust = 0))
```

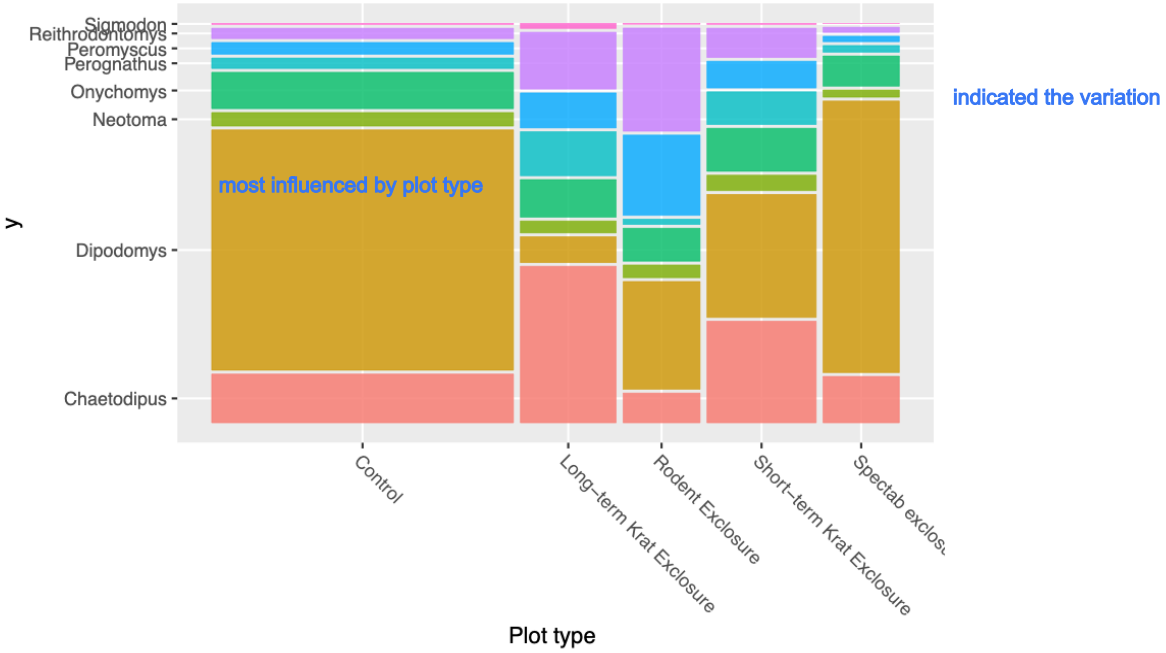


Another way we can visualize interactions is with mosaic plots using the `geom_mosaic`. This isn't built in to the tidyverse (yet?) so requires us to load a new library, `ggmosaic`. Vignette here: https://cran.r-project.org/web/packages/ggmosaic/vignettes/ggmosaic.html

```
#install.packages("ggmosaic")
library("ggmosaic")
```

```
ggplot(sc) +
  geom_mosaic(aes(x = product(plot_type), fill=genus), na.rm=TRUE) +
   labs(x="Plot type ") +
  theme(axis.text.x=element_text(angle = -45, hjust = 0)) +
  theme(legend.position = "none")
```

essentially, like stack bar plot with facet wrap

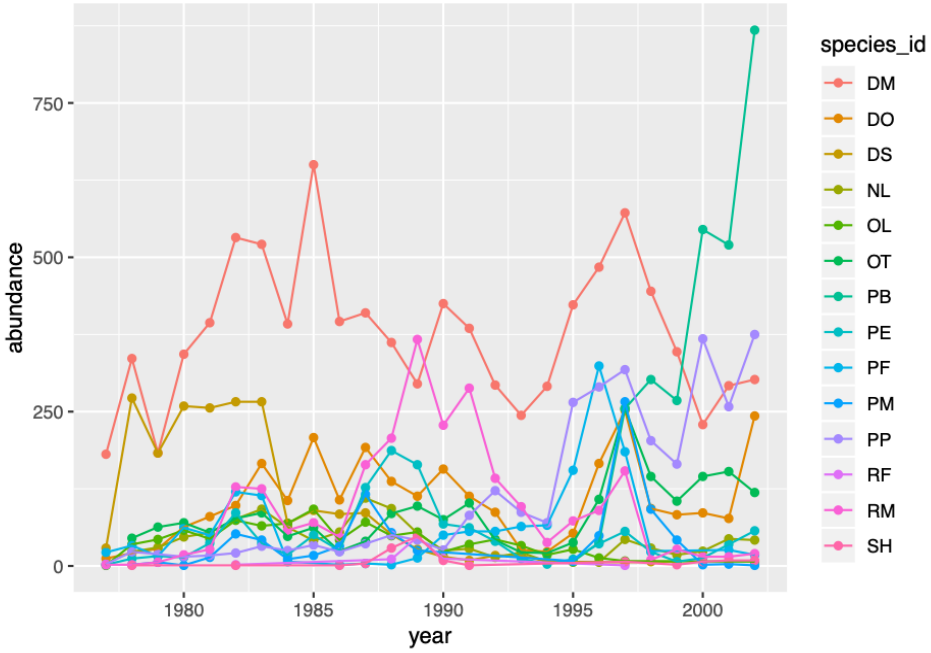


indicated the variation

What do you think? Which one do we like better?

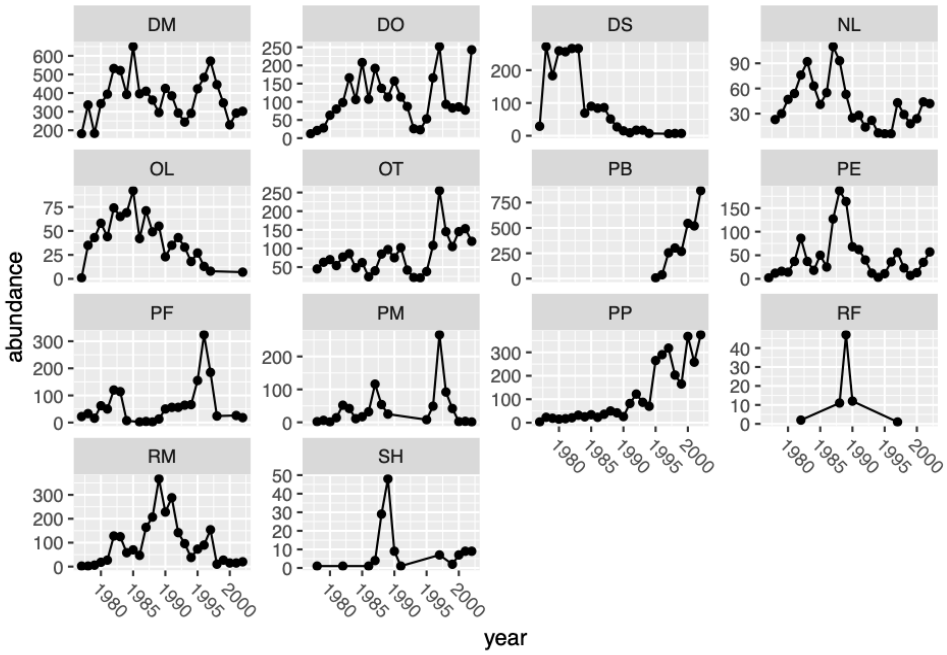
Similarly, in the first set of lecture notes at the very end there was a plot of abundance by year by species (i.e., "does the effect of year on abundance depend on species"). Here it is again:

```
sc_counts <- sc %>%
  count(year, species_id, name = "abundance")

ggplot(sc_counts, aes(year, abundance, col = species_id)) +
  geom_point() +
  geom_line()
```



This is probably enough to see that there is an iteraction (or 'moderator effect'). But it's pretty messy.

```
ggplot(sc_counts, aes(year, abundance)) +
  geom_point() +
  geom_line() +
  facet_wrap(~ species_id, scales = "free_y")  +
  theme(axis.text.x=element_text(angle = -45, hjust = 0))
```



There is definitely year-to-year variation in abundance, and it depends on the species. We can build this story up even further, by adding another layer on top to see how the type of plot is influenced by these variables.

```
sc_counts_plot <- sc %>%
  count(year, species_id, plot_type, name = "abundance")

ggplot(sc_counts_plot, aes(year, abundance, col = plot_type)) +
  geom_point() +
  geom_line() +
  facet_wrap(~ species_id, scale = "free_y") +
  theme(legend.position = "top", legend.text = element_text(size = 6)) +
  theme(axis.text.x=element_text(angle = -45, hjust = 0))
```

That is not very effective. I think it shows us that there is something interesting going on, but this particular visualization is pretty hard to tease apartt. So let's try plotting the same data in a different way.

```
ggplot(sc_counts_plot) +
  geom_col(aes(year, abundance, fill = plot_type)) +
  facet_wrap(~ species_id, scale = "free_y") +
  theme(legend.position = "top", legend.text = element_text(size = 6)) +
  theme(axis.text.x=element_text(angle = -45, hjust = 0))
```

There are four different variables explored here! Let's make a list of all the things that we think are true about this dataset (really about this ecological experiment), based only on this figure.

Overall visualization goals

(Whitlock & Schluter, The Analysis of Biological Data)

Principles of effective display
We will follow these metrics to evaluate figures:
1. Show the data
2. Make patterns in the data easy to see
3. Represent magnitudes honestly
4. Draw graphical elements clearly, minimizing clutter