

# Data visualization with ggplot2

*Data Carpentry contributors*

```
## Warning: package 'tidyverse' was built under R version 3.4.2
## Warning: package 'tidyverse' was built under R version 3.4.2
## Warning: package 'purrr' was built under R version 3.4.2
## Warning: package 'dplyr' was built under R version 3.4.2
```

---

## Learning Objectives

- Produce scatter plots, boxplots, and time series plots using ggplot.
- Set universal plot settings.
- Describe what facetting is and apply facetting in ggplot.
- Modify the aesthetics of an existing ggplot plot (including axis labels and color).
- Build complex and customized plots from data in a data frame.

We start by loading the required packages. `ggplot2` is included in the `tidyverse` package.

```
library(tidyverse)
```

If not still in the workspace, load the data we saved in the previous lesson.

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

## Plotting with ggplot2

`ggplot2` is a plotting package that makes it simple to create complex plots from data in a data frame. It provides a more programmatic interface for specifying what variables to plot, how they are displayed, and general visual properties. Therefore, we only need minimal changes if the underlying data change or if we decide to change from a bar plot to a scatterplot. This helps in creating publication quality plots with minimal amounts of adjustments and tweaking.

ggplot likes data in the ‘long’ format: i.e., a column for every dimension, and a row for every observation. Well-structured data will save you lots of time when making figures with ggplot.

ggplot graphics are built step by step by adding new elements. Adding layers in this fashion allows for extensive flexibility and customization of plots.

To build a ggplot, we need to:

- use the `ggplot()` function and bind the plot to a specific data frame using the `data` argument

```
ggplot(data = surveys_complete)
```

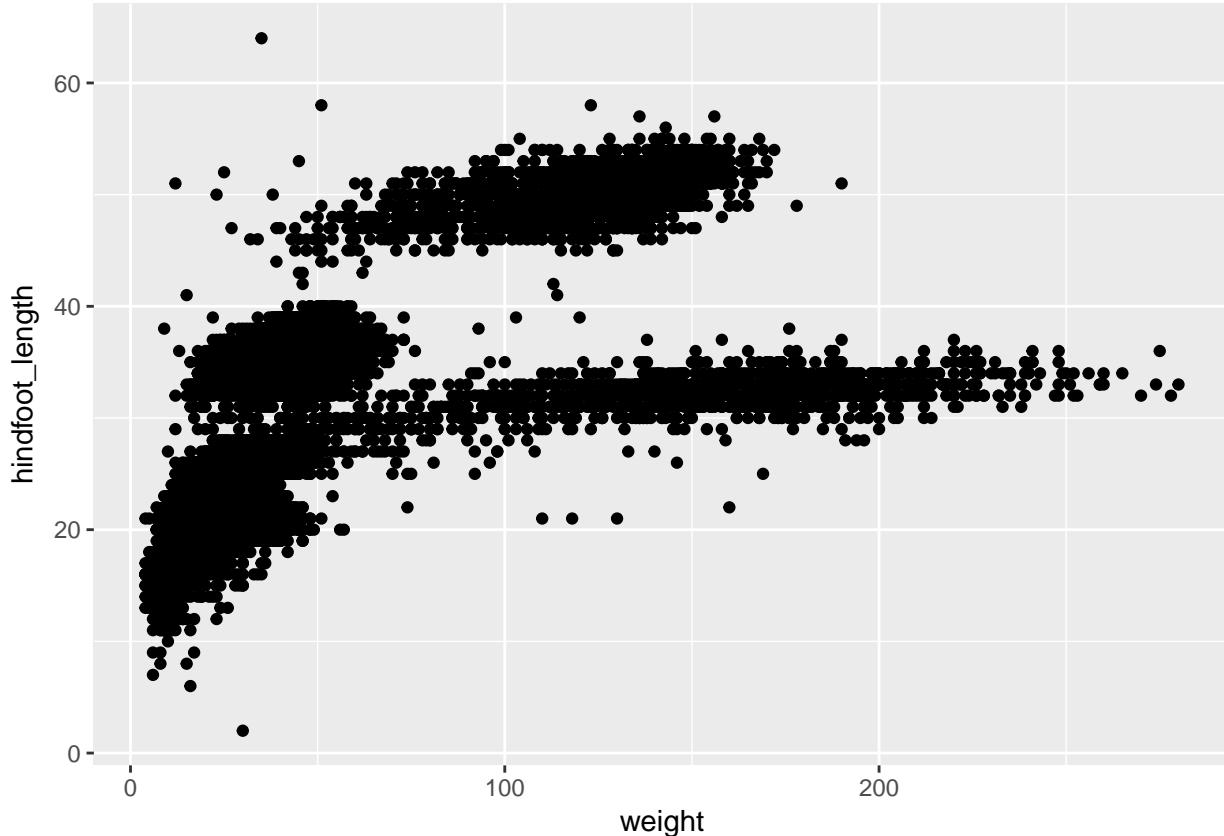
- define a mapping (using the aesthetic (`aes`) function), by selecting the variables to be plotted and specifying how to present them in the graph, e.g. as x/y positions or characteristics such as size, shape, color, etc.

```
ggplot(data = surveys_complete, aes(x = weight, y = hindfoot_length))
```

- add `geoms` – graphical representation of the data in the plot (points, lines, bars). `ggplot2` offers many different geoms; we will use some common ones today, including:
  - `geom_point()` for scatter plots, dot plots, etc.
  - `geom_boxplot()` for, well, boxplots!
  - `geom_line()` for trend lines, time series, etc.

To add a geom to the plot use the `+` operator. Because we have two continuous variables, let's use `geom_point()` first:

```
ggplot(data = surveys_complete, aes(x = weight, y = hindfoot_length)) +
  geom_point()
```



The `+` in the `ggplot2` package is particularly useful because it allows you to modify existing `ggplot` objects. This means you can easily set up plot templates and conveniently explore different types of plots, so the above plot can also be generated with code like this:

```
# Assign plot to a variable
surveys_plot <- ggplot(data = surveys_complete, aes(x = weight, y = hindfoot_length))

# Draw the plot
surveys_plot +
  geom_point()
```

\*\*Notes:\*\*

- Anything you put in the `ggplot()` function can be seen by any geom layers that you add (i.e., these are universal plot settings). This includes the x- and y-axis mapping you set up in `aes()`.
- You can also specify mappings for a given geom independently of the mappings defined globally in the `ggplot()` function.
- The `+` sign used to add layers must be placed at the end of each line containing a layer. If, instead, the

+ sign is added in the line before the other layer, `ggplot2` will not add the new layer and will return an error message.

```
# this is the correct syntax for adding layers
surveys_plot +
  geom_point()

# this will not add the new layer and will return an error message
surveys_plot
  + geom_point()
```

### Challenge (optional)

Scatter plots can be useful exploratory tools for small datasets. For data sets with large numbers of observations, such as the `surveys_complete` data set, overplotting of points can be a limitation of scatter plots. One strategy for handling such settings is to use hexagonal binning of observations. The plot space is tessellated into hexagons. Each hexagon is assigned a color based on the number of observations that fall within its boundaries. To use hexagonal binning with `ggplot2`, first install the R package `hexbin` from CRAN:

```
install.packages("hexbin")
library(hexbin)
```

Then use the `geom_hex()` function:

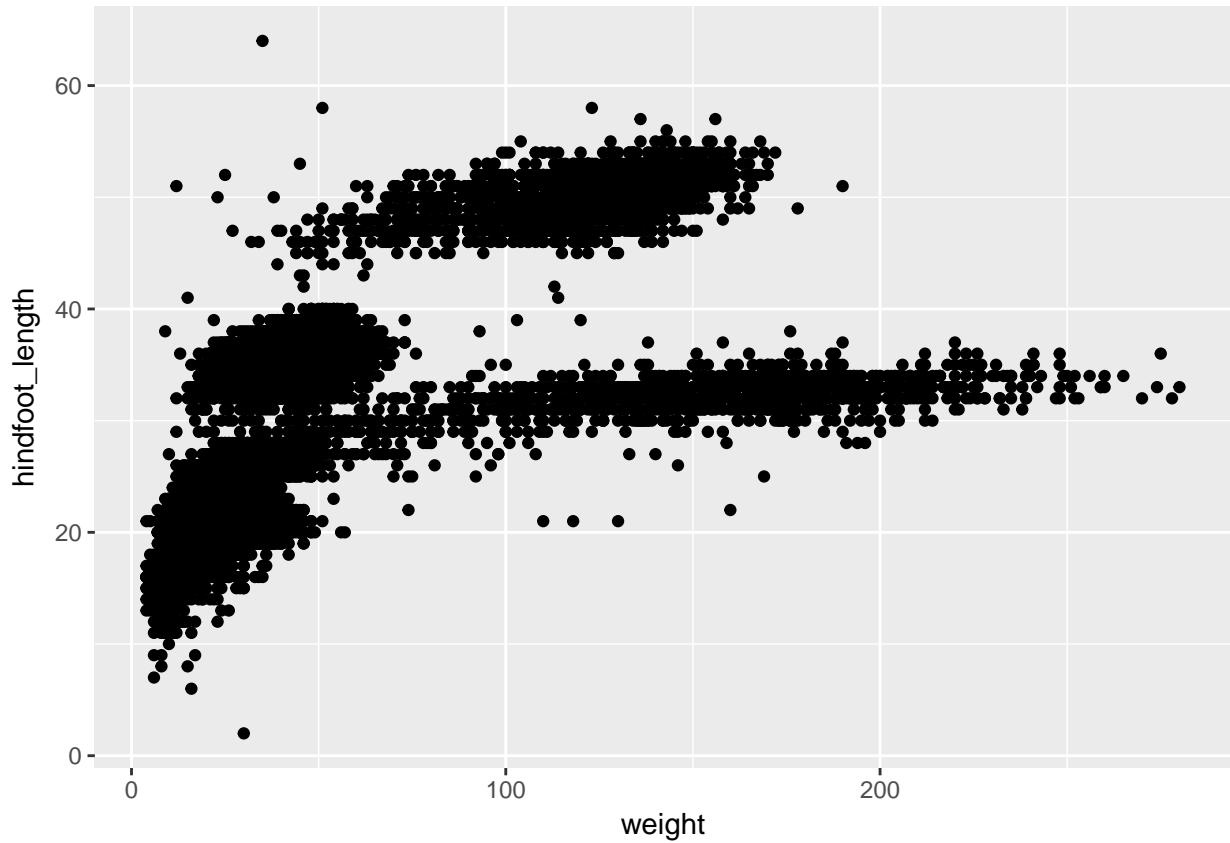
```
surveys_plot +
  geom_hex()
```

- What are the relative strengths and weaknesses of a hexagonal bin plot compared to a scatter plot? Examine the above scatter plot and compare it with the hexagonal bin plot that you created.

## Building your plots iteratively

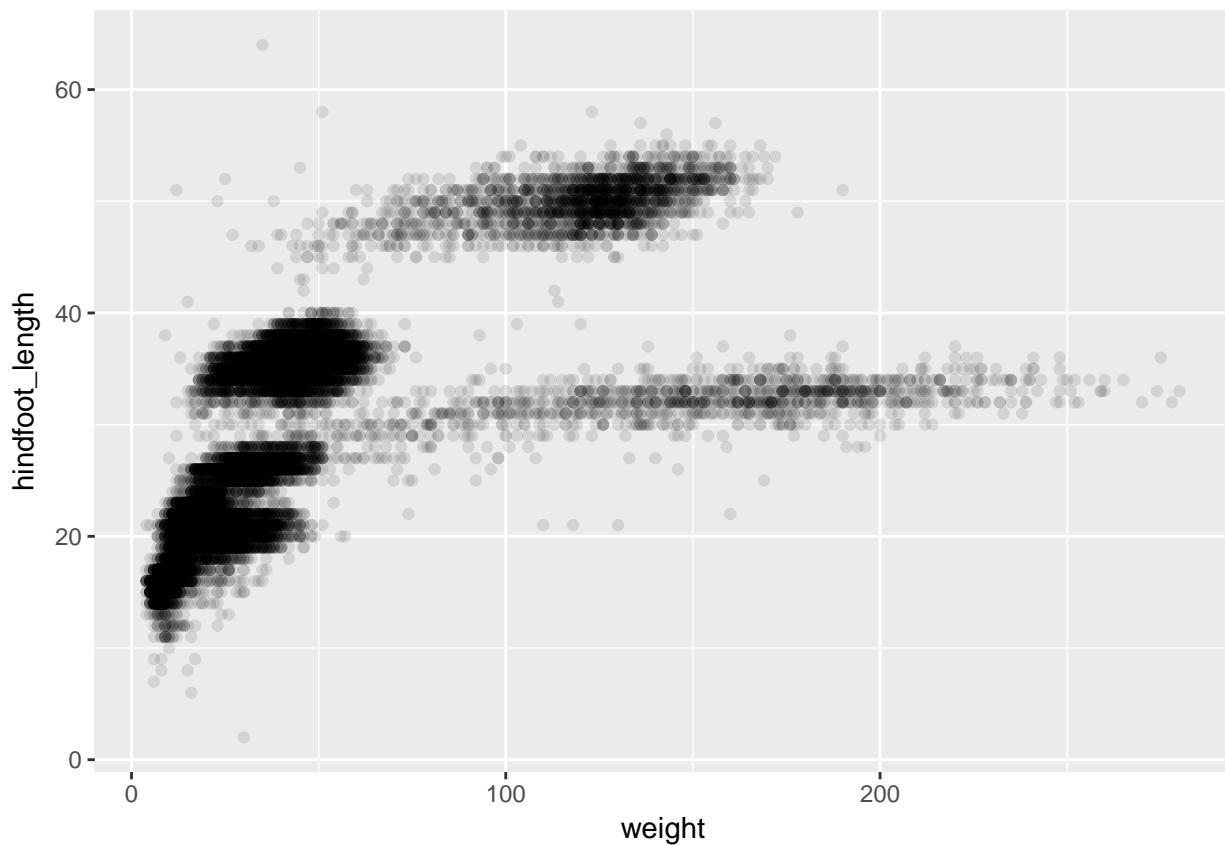
Building plots with `ggplot` is typically an iterative process. We start by defining the dataset we'll use, lay out the axes, and choose a geom:

```
ggplot(data = surveys_complete, aes(x = weight, y = hindfoot_length)) +
  geom_point()
```



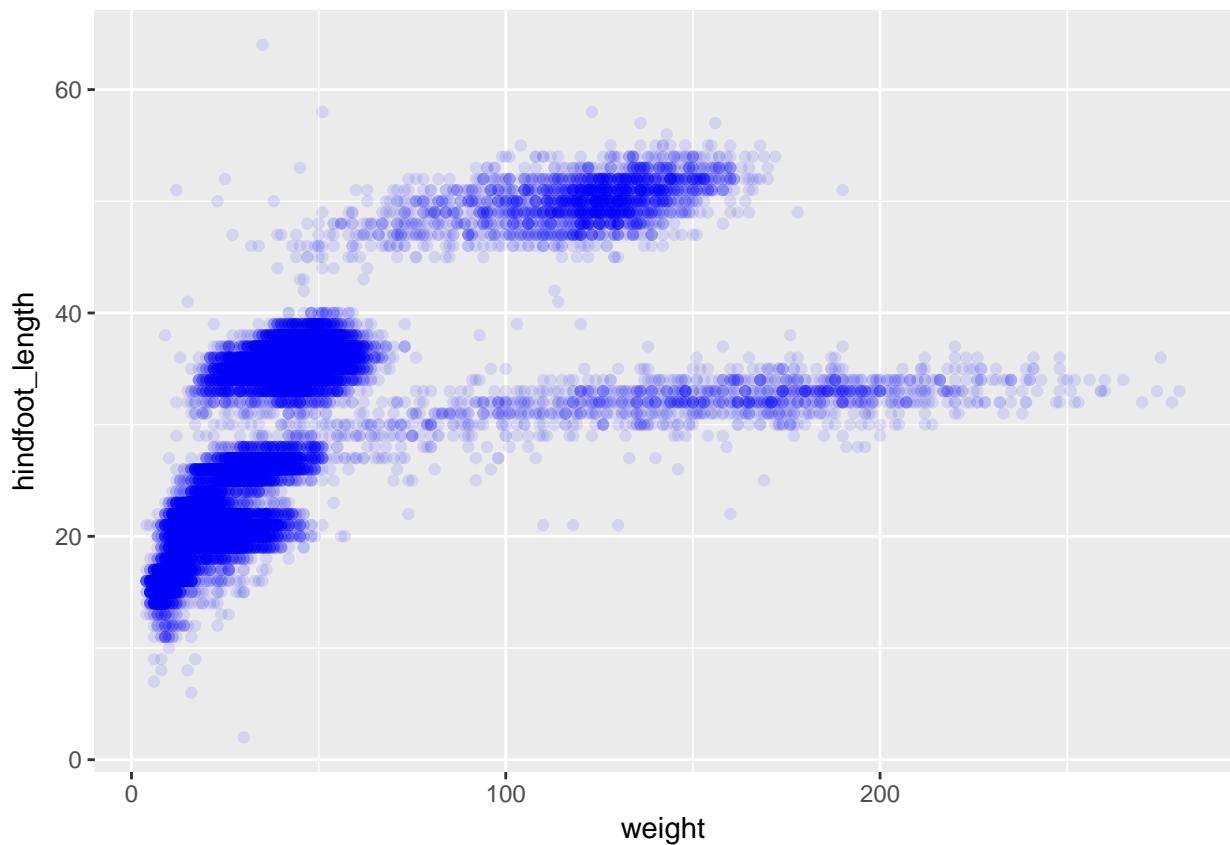
Then, we start modifying this plot to extract more information from it. For instance, we can add transparency (`alpha`) to avoid overplotting:

```
ggplot(data = surveys_complete, aes(x = weight, y = hindfoot_length)) +  
  geom_point(alpha = 0.1)
```



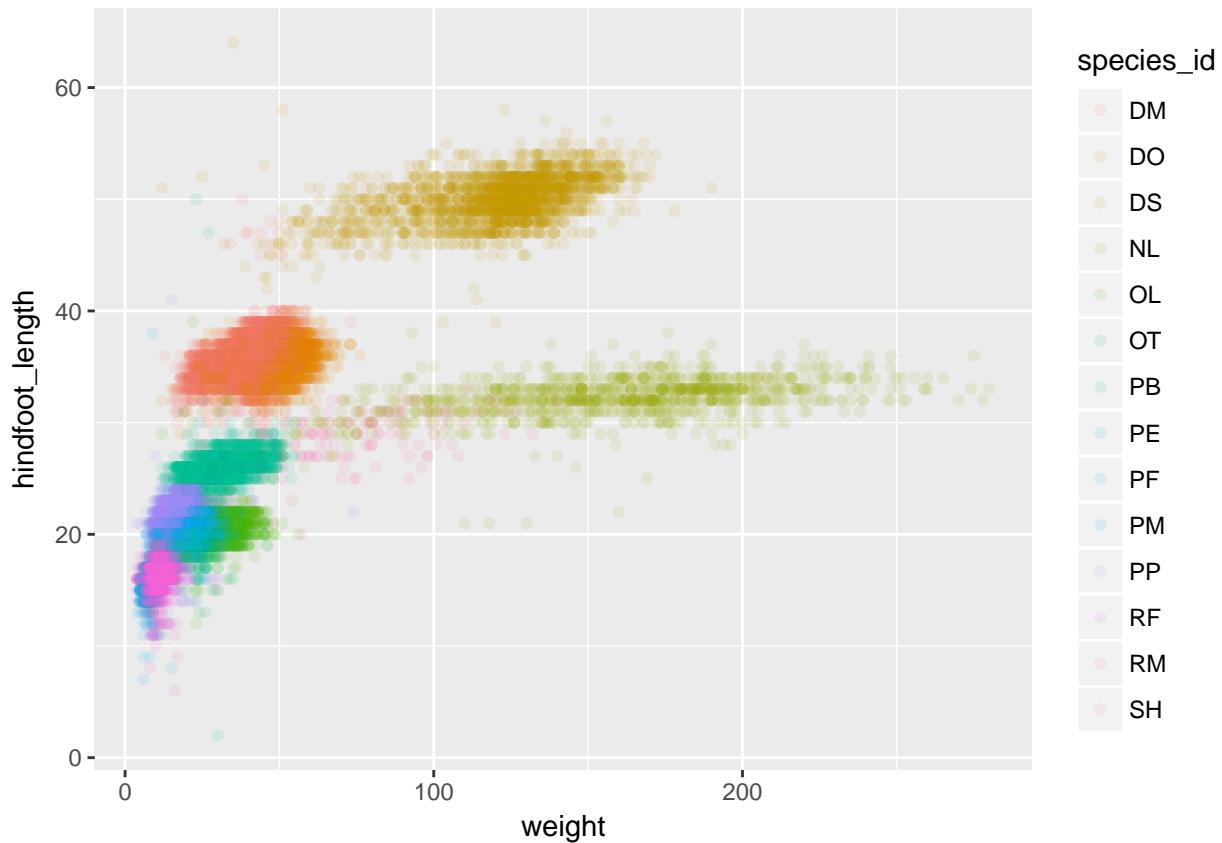
We can also add colors for all the points:

```
ggplot(data = surveys_complete, aes(x = weight, y = hindfoot_length)) +  
  geom_point(alpha = 0.1, color = "blue")
```



Or to color each species in the plot differently:

```
ggplot(data = surveys_complete, aes(x = weight, y = hindfoot_length)) +  
  geom_point(alpha = 0.1, aes(color = species_id))
```



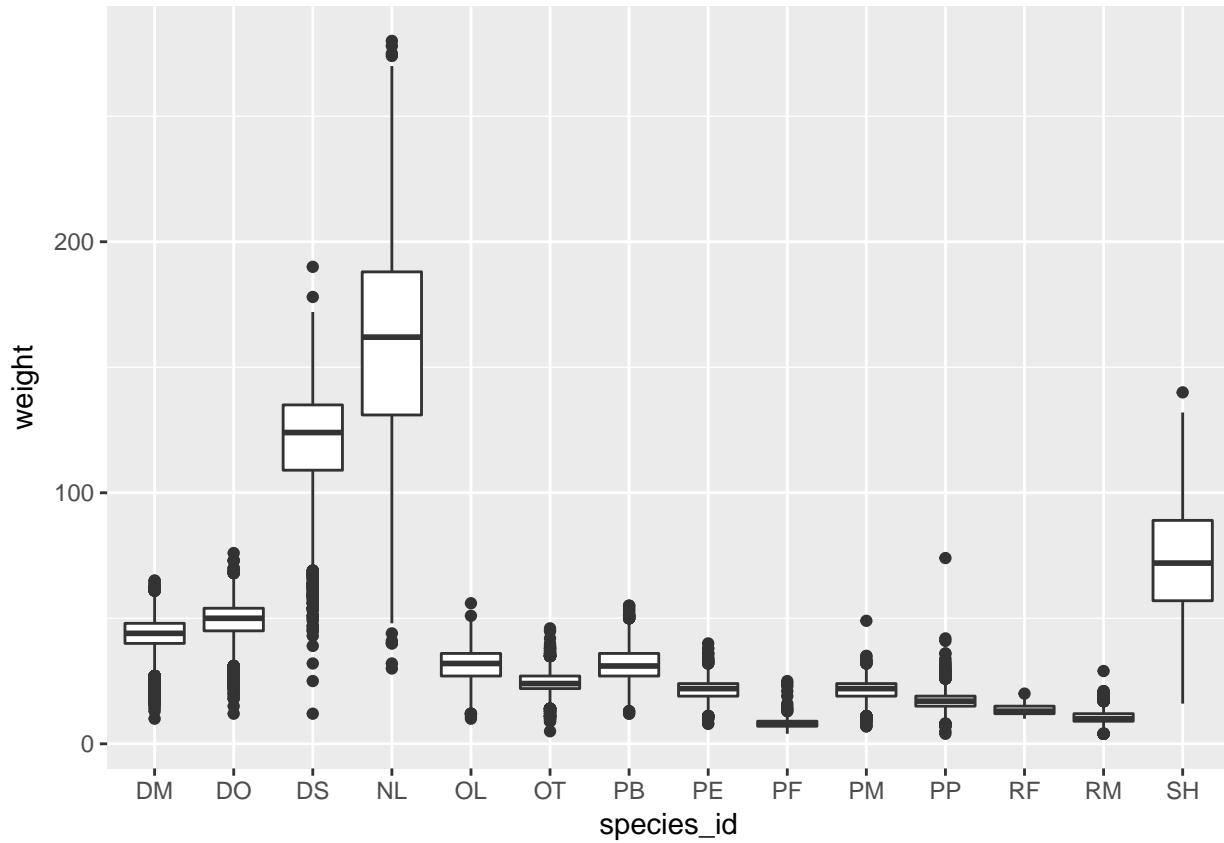
## Challenge

Use what you just learned to create a scatter plot of `weight` over `species_id` with the plot types showing in different colors. Is this a good way to show this type of data?

## Boxplot

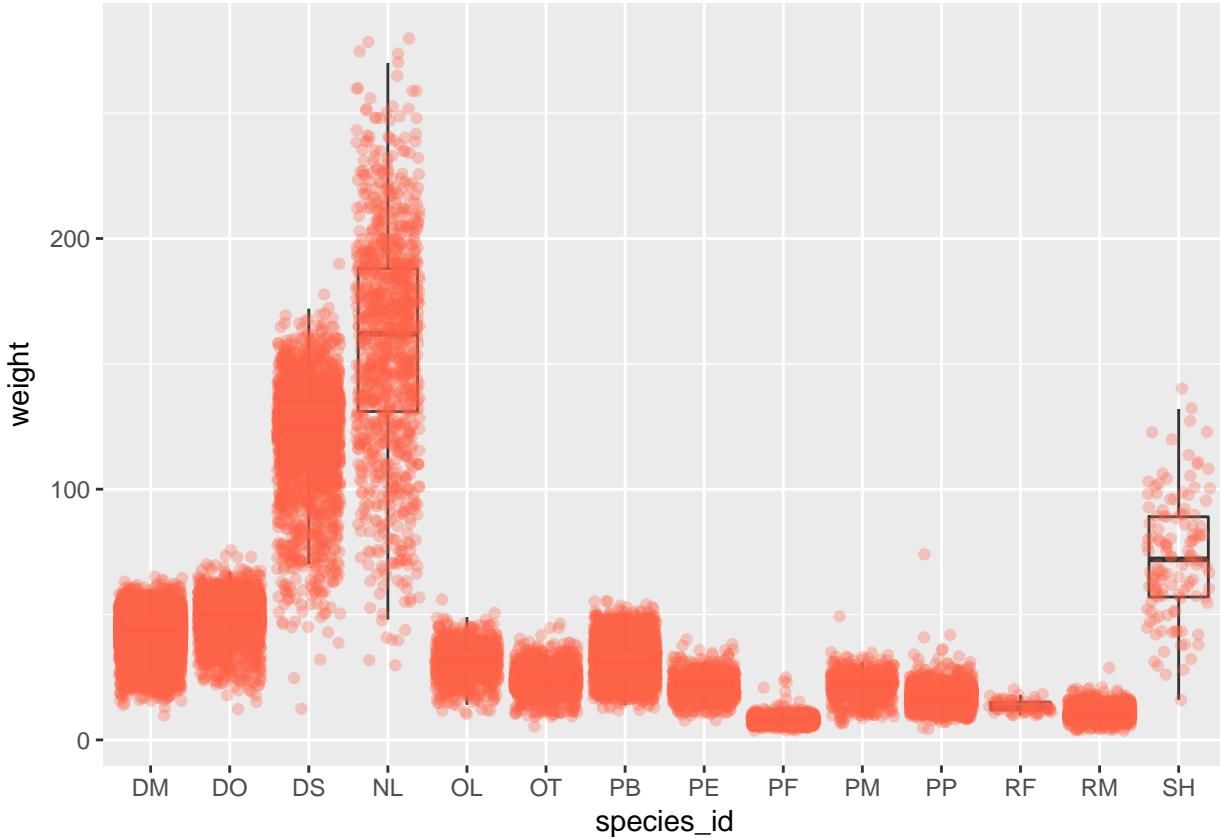
We can use boxplots to visualize the distribution of weight within each species:

```
ggplot(data = surveys_complete, aes(x = species_id, y = weight)) +
  geom_boxplot()
```



By adding points to boxplot, we can have a better idea of the number of measurements and of their distribution:

```
ggplot(data = surveys_complete, aes(x = species_id, y = weight)) +
  geom_boxplot(alpha = 0) +
  geom_jitter(alpha = 0.3, color = "tomato")
```



Notice how the boxplot layer is behind the jitter layer? What do you need to change in the code to put the boxplot in front of the points such that it's not hidden?

### Challenges

Boxplots are useful summaries, but hide the *shape* of the distribution. For example, if the distribution is bimodal, we would not see it in a boxplot. An alternative to the boxplot is the violin plot where the shape (of the density of points) is drawn.

- Replace the box plot with a violin plot; see `geom_violin()`.

In many types of data, it is important to consider the *scale* of the observations. For example, it may be worth changing the scale of the axis to better distribute the observations in the space of the plot. Changing the scale of the axes is done similarly to adding/modifying other components (i.e., by incrementally adding commands). Try making these modifications:

- Represent weight on the  $\log_{10}$  scale; see `scale_y_log10()`.

So far, we've looked at the distribution of weight within species. Try making a new plot to explore the distribution of another variable within each species.

- Create a boxplot for `hindfoot_length`. Overlay the boxplot layer on a jitter layer to show actual measurements.
- Add color to the data points on your boxplot according to the plot from which the sample was taken (`plot_id`).

*Hint:* Check the class for `plot_id`. Consider changing the class of `plot_id` from integer to factor. Why does this change how R makes the graph?

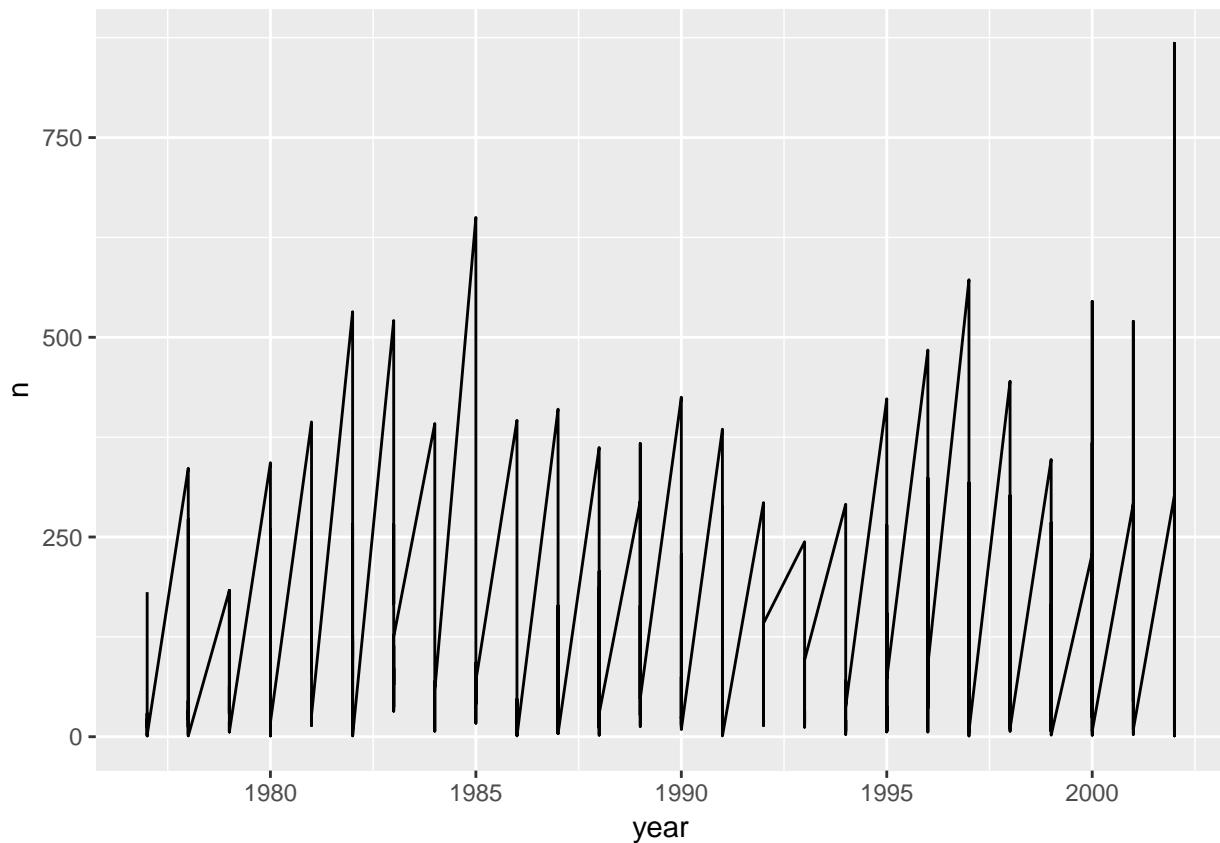
## Plotting time series data

Let's calculate number of counts per year for each species. First we need to group the data and count records within each group:

```
yearly_counts <- surveys_complete %>%
  group_by(year, species_id) %>%
  tally()
```

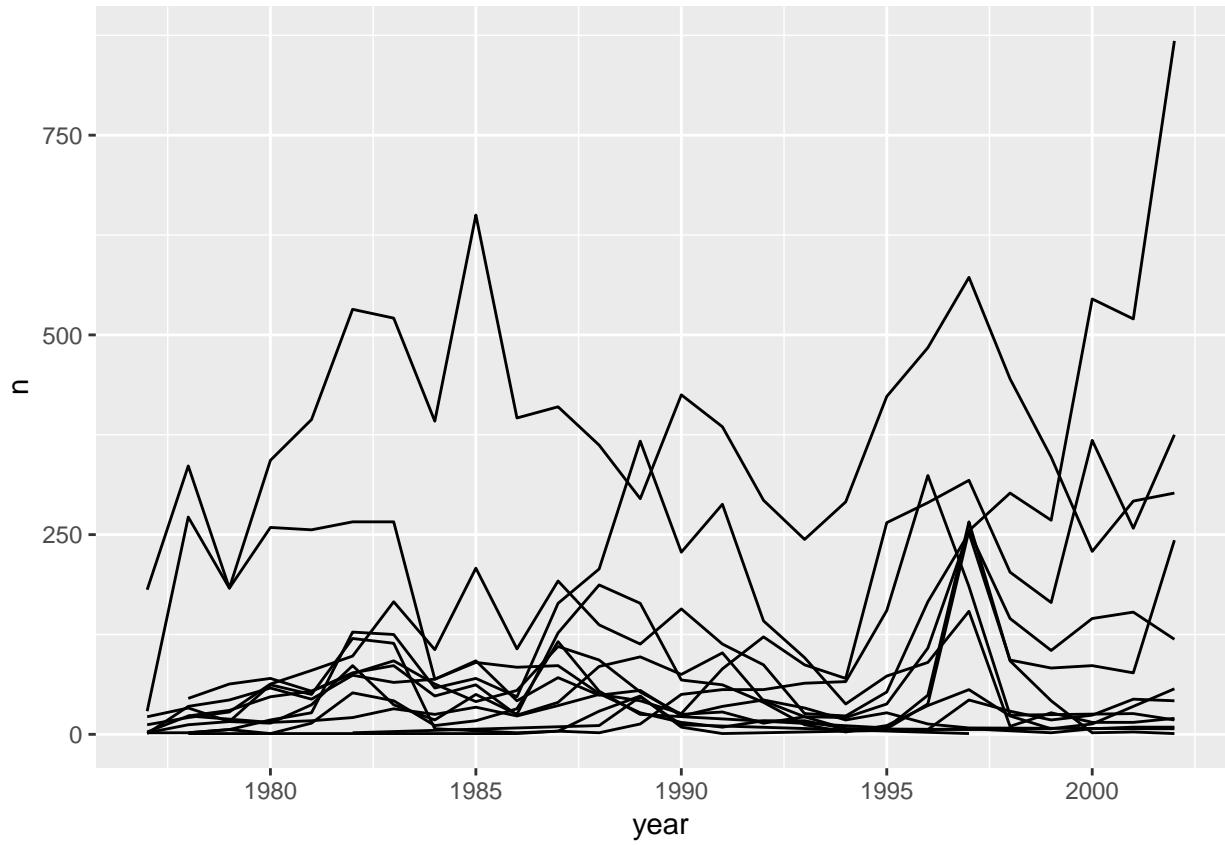
Time series data can be visualized as a line plot with years on the x axis and counts on the y axis:

```
ggplot(data = yearly_counts, aes(x = year, y = n)) +
  geom_line()
```



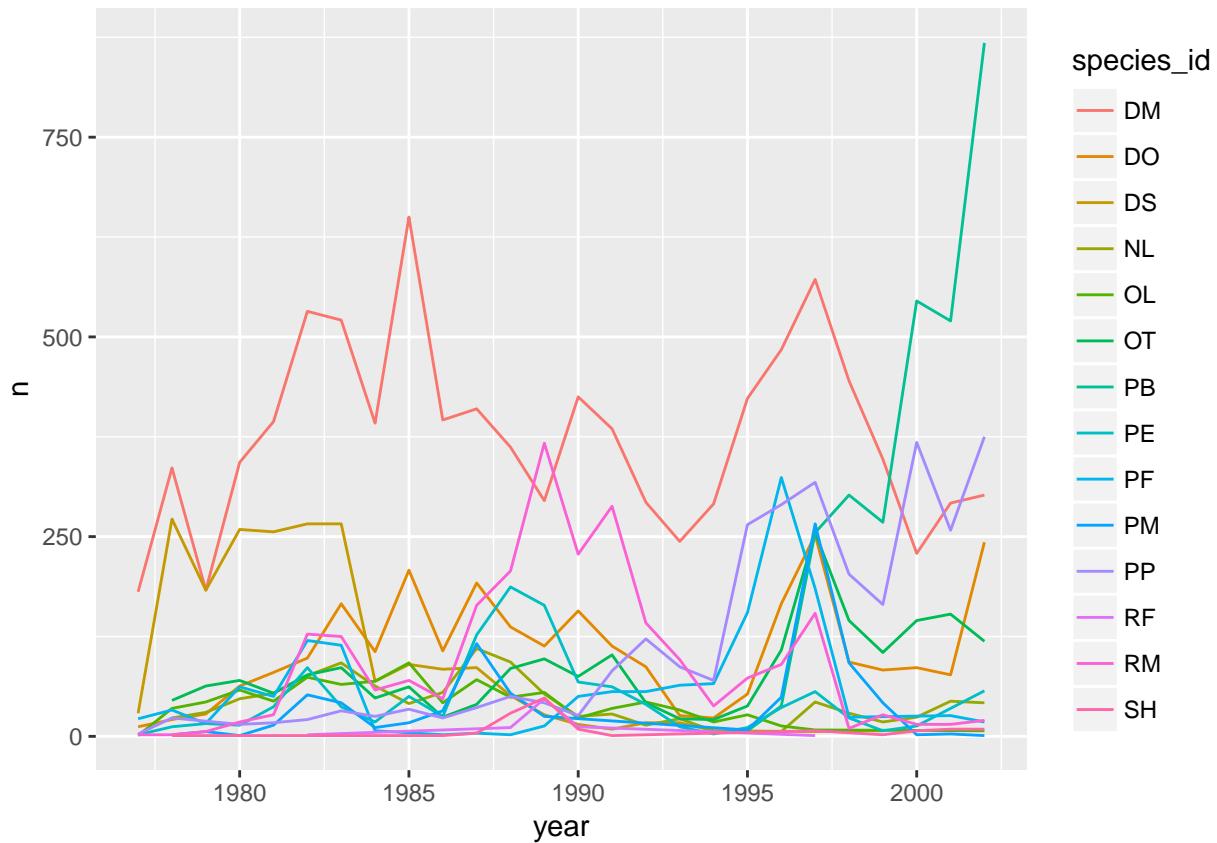
Unfortunately, this does not work because we plotted data for all the species together. We need to tell ggplot to draw a line for each species by modifying the aesthetic function to include `group = species_id`:

```
ggplot(data = yearly_counts, aes(x = year, y = n, group = species_id)) +
  geom_line()
```



We will be able to distinguish species in the plot if we add colors (using `color` also automatically groups the data):

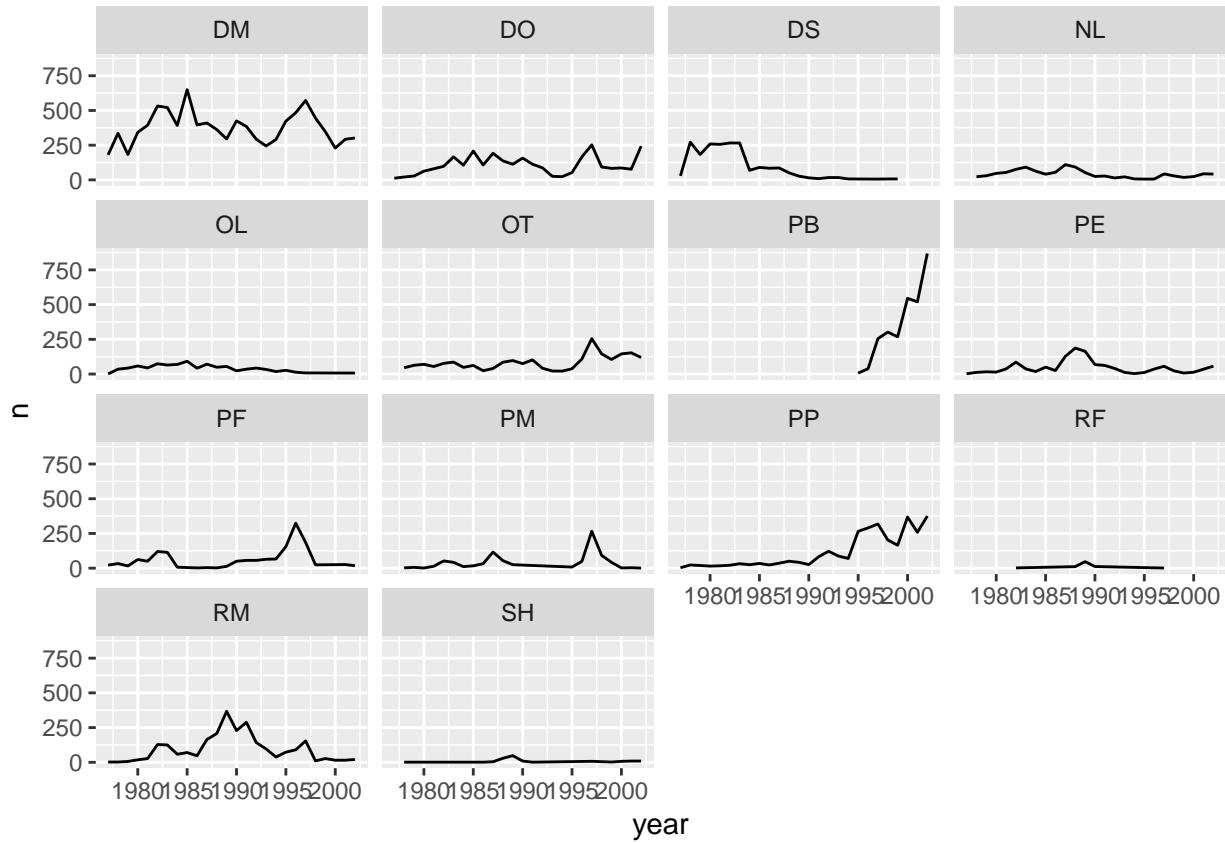
```
ggplot(data = yearly_counts, aes(x = year, y = n, color = species_id)) +  
  geom_line()
```



## Faceting

ggplot has a special technique called *faceting* that allows the user to split one plot into multiple plots based on a factor included in the dataset. We will use it to make a time series plot for each species:

```
ggplot(data = yearly_counts, aes(x = year, y = n)) +
  geom_line() +
  facet_wrap(~ species_id)
```

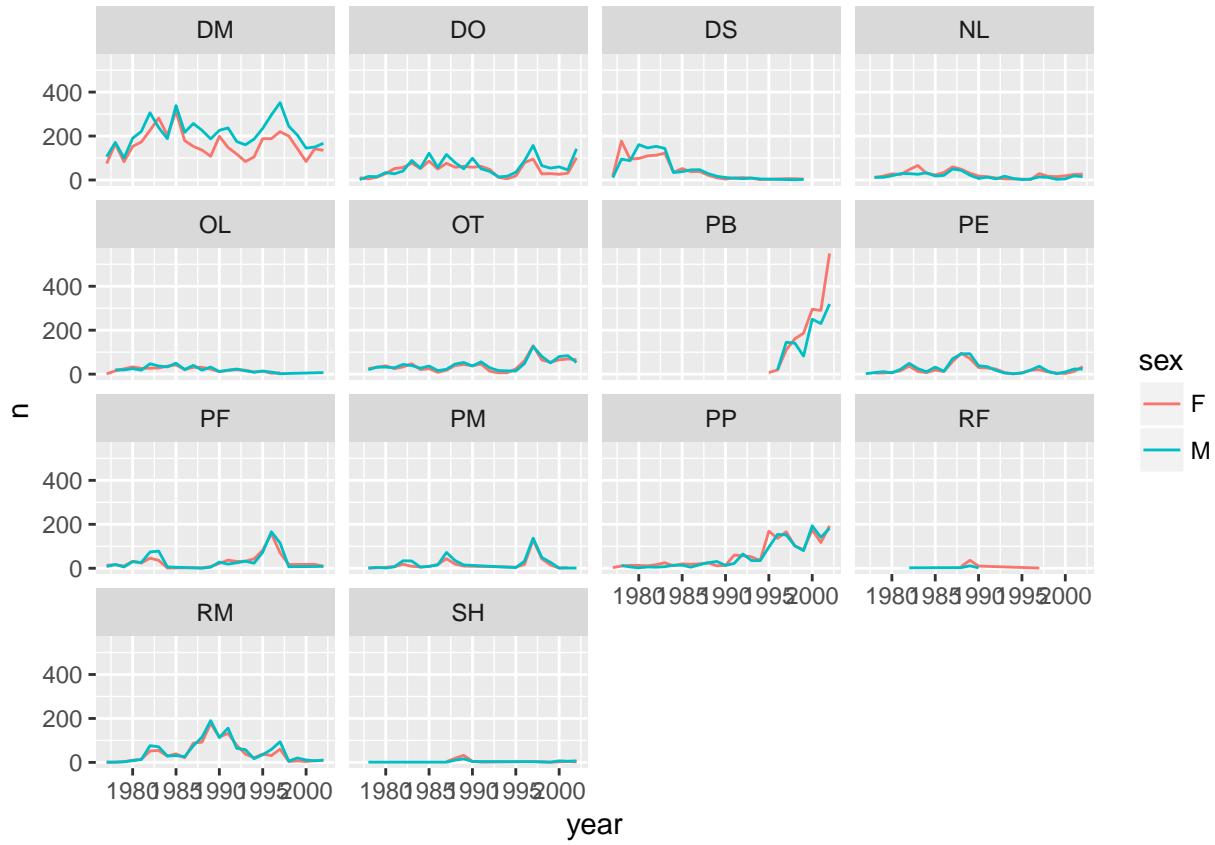


Now we would like to split the line in each plot by the sex of each individual measured. To do that we need to make counts in the data frame grouped by `year`, `species_id`, and `sex`:

```
yearly_sex_counts <- surveys_complete %>%
  group_by(year, species_id, sex) %>%
  tally()
```

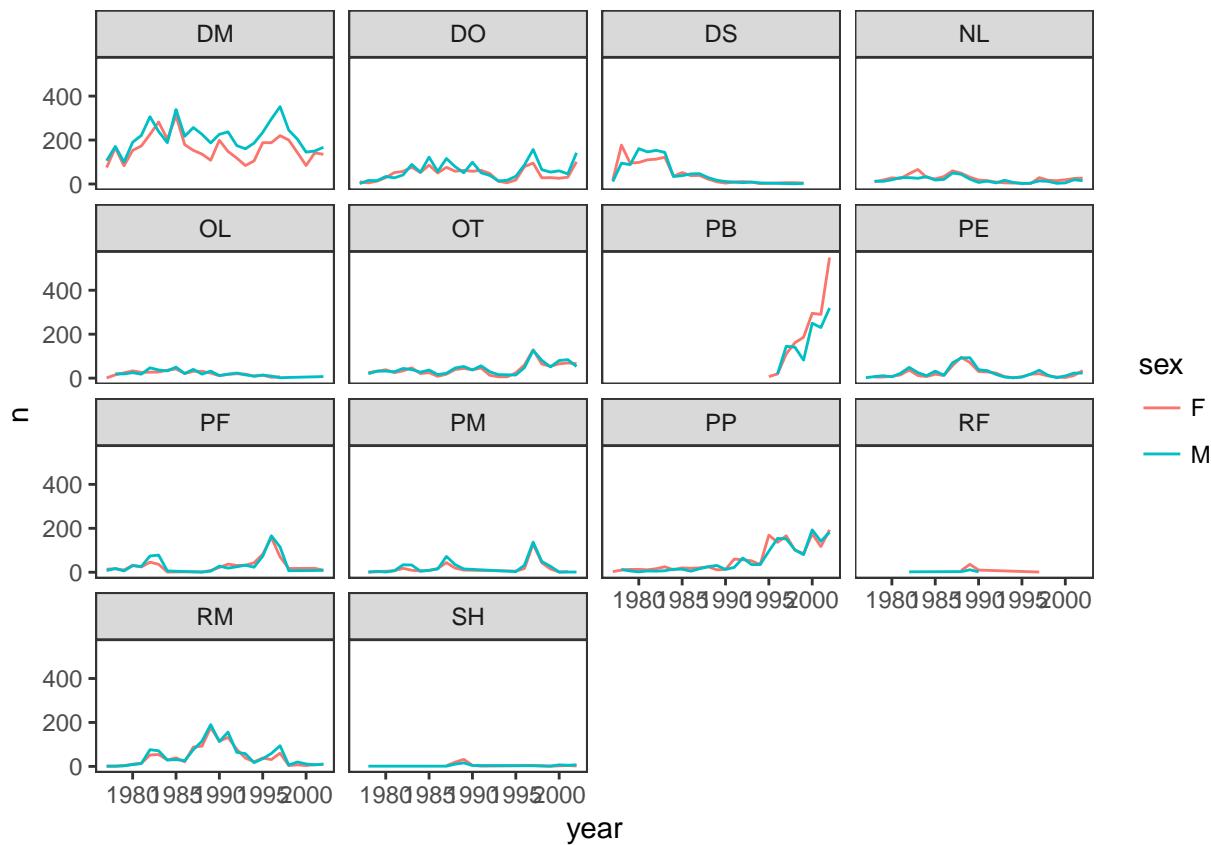
We can now make the faceted plot by splitting further by sex using `color` (within a single plot):

```
ggplot(data = yearly_sex_counts, aes(x = year, y = n, color = sex)) +
  geom_line() +
  facet_wrap(~ species_id)
```



Usually plots with white background look more readable when printed. We can set the background to white using the function `theme_bw()`. Additionally, you can remove the grid:

```
ggplot(data = yearly_sex_counts, aes(x = year, y = n, color = sex)) +
  geom_line() +
  facet_wrap(~ species_id) +
  theme_bw() +
  theme(panel.grid = element_blank())
```



## ggplot2 themes

In addition to `theme_bw()`, which changes the plot background to white, `ggplot2` comes with several other themes which can be useful to quickly change the look of your visualization. The complete list of themes is available at <http://docs.ggplot2.org/current/ggtheme.html>. `theme_minimal()` and `theme_light()` are popular, and `theme_void()` can be useful as a starting point to create a new hand-crafted theme.

The `ggthemes` package provides a wide variety of options (including an Excel 2003 theme). The `ggplot2` extensions website provides a list of packages that extend the capabilities of `ggplot2`, including additional themes.

## Challenge

Use what you just learned to create a plot that depicts how the average weight of each species changes through the years.

->

The `facet_wrap` geometry extracts plots into an arbitrary number of dimensions to allow them to cleanly fit on one page. On the other hand, the `facet_grid` geometry allows you to explicitly specify how you want your plots to be arranged via formula notation (`rows ~ columns`; `a .` can be used as a placeholder that indicates only one row or column).

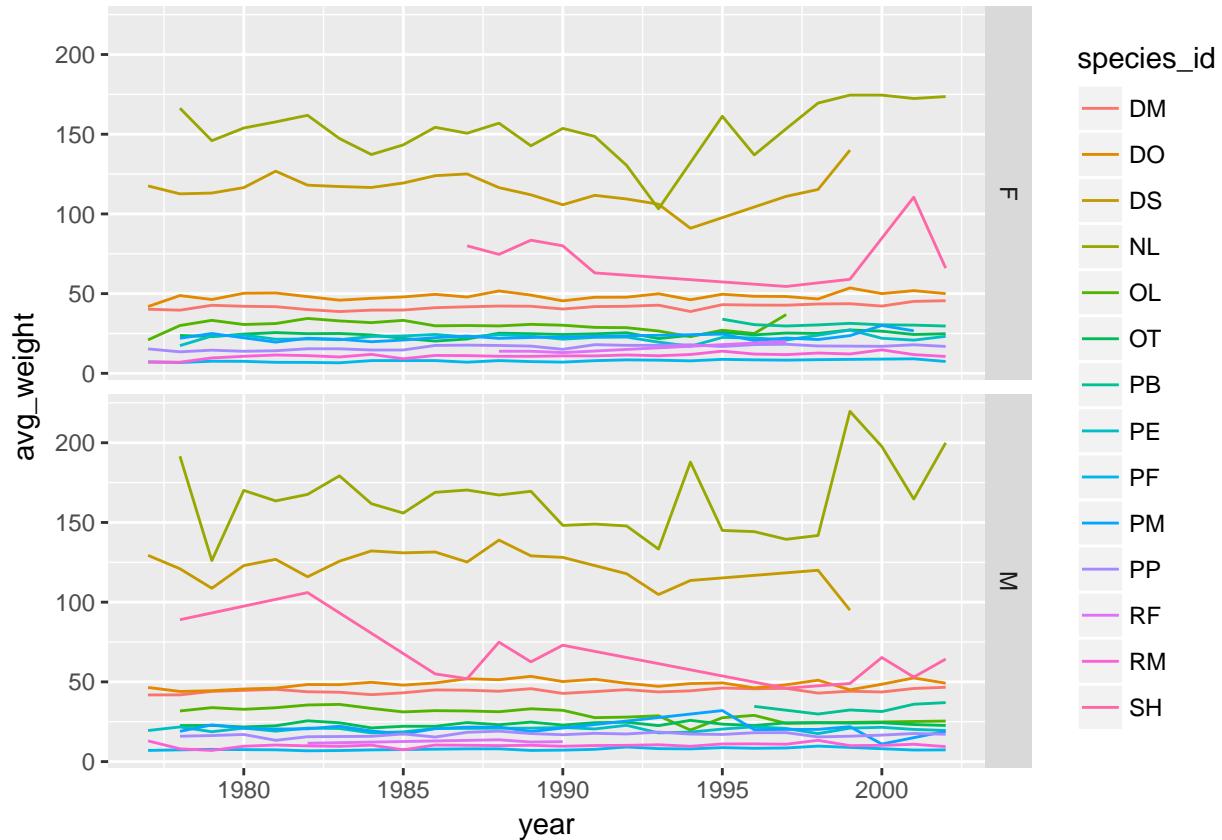
Let's modify the previous plot to compare how the weights of males and females has changed through time:

```
# One column, facet by rows
yearly_sex_weight <- surveys_complete %>%
```

```

group_by(year, sex, species_id) %>%
  summarize(avg_weight = mean(weight))
ggplot(data = yearly_sex_weight, aes(x = year, y = avg_weight, color = species_id)) +
  geom_line() +
  facet_grid(sex ~ .)

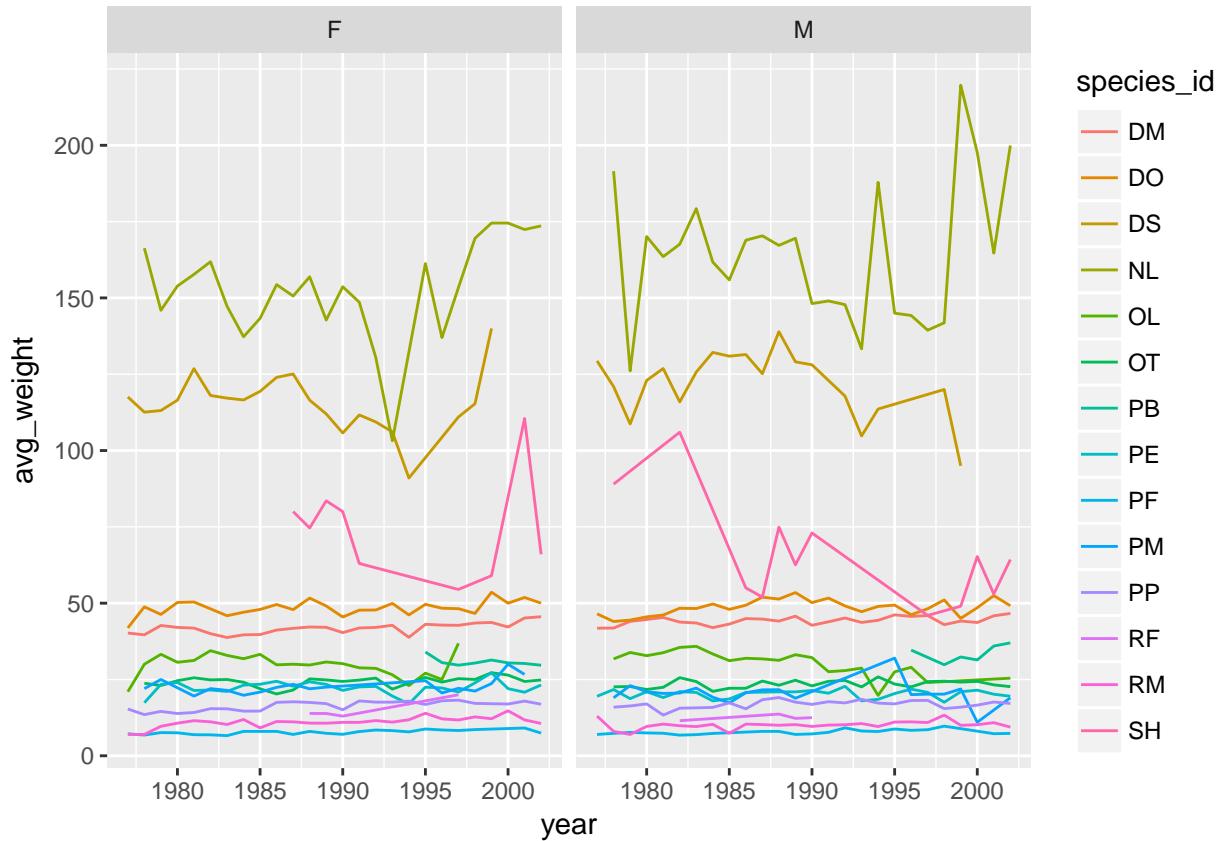
```



```

# One row, facet by column
ggplot(data = yearly_sex_weight, aes(x = year, y = avg_weight, color = species_id)) +
  geom_line() +
  facet_grid(. ~ sex)

```

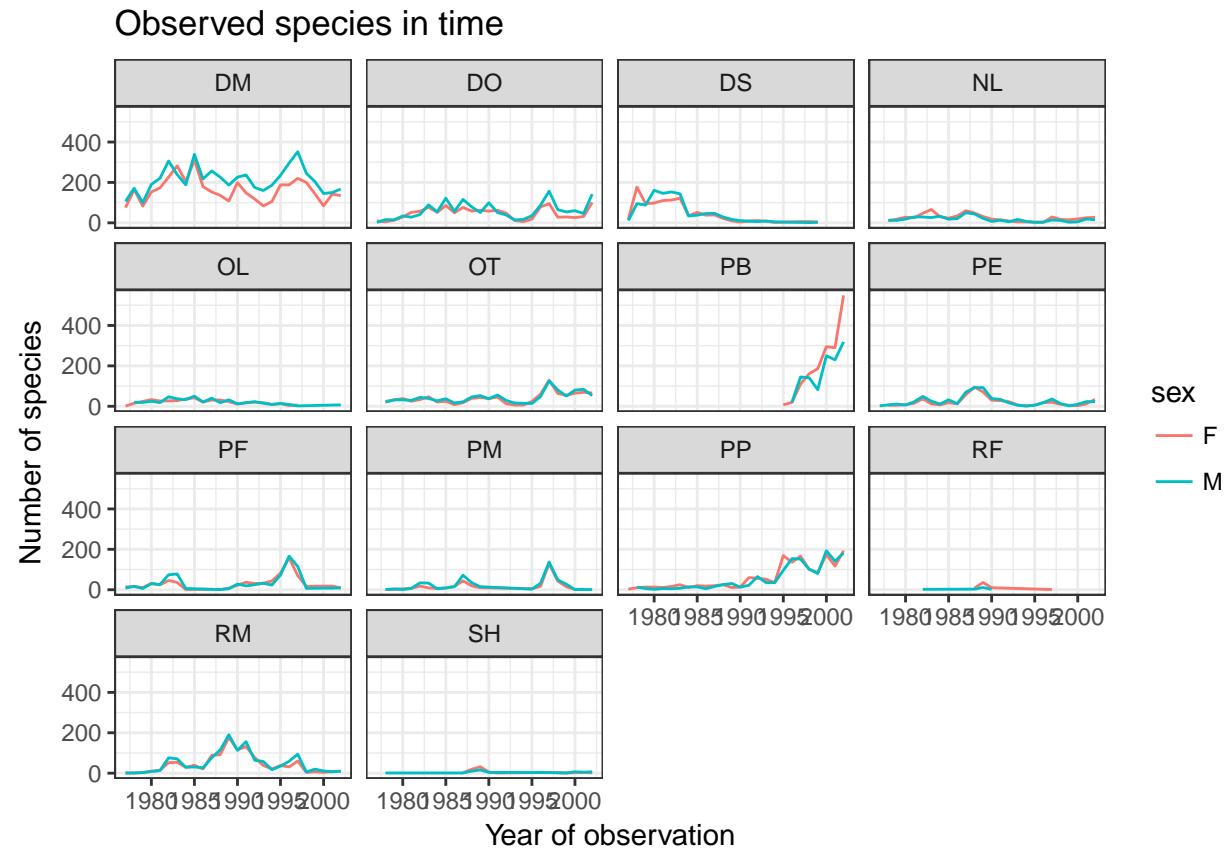


## Customization

Take a look at the `ggplot2` cheat sheet, and think of ways you could improve the plot.

Now, let's change names of axes to something more informative than 'year' and 'n' and add a title to the figure:

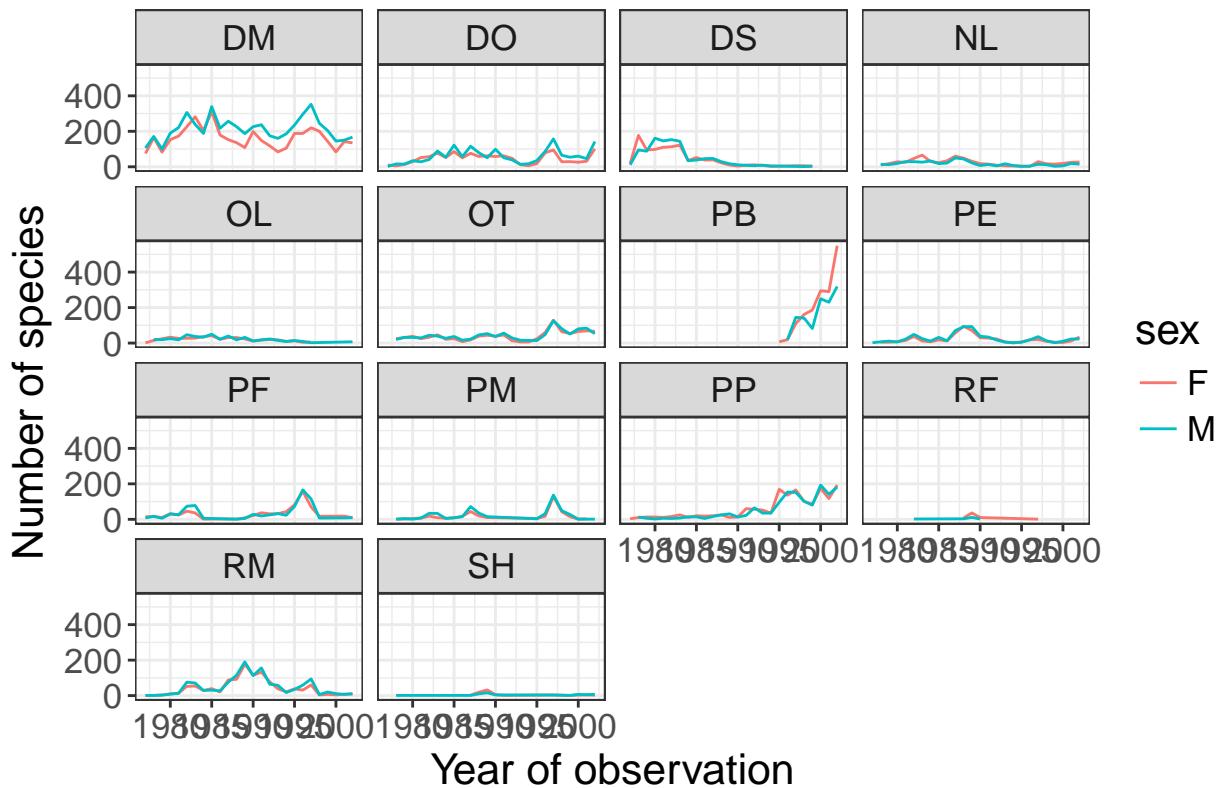
```
ggplot(data = yearly_sex_counts, aes(x = year, y = n, color = sex)) +
  geom_line() +
  facet_wrap(~ species_id) +
  labs(title = "Observed species in time",
       x = "Year of observation",
       y = "Number of species") +
  theme_bw()
```



The axes have more informative names, but their readability can be improved by increasing the font size:

```
ggplot(data = yearly_sex_counts, aes(x = year, y = n, color = sex)) +
  geom_line() +
  facet_wrap(~ species_id) +
  labs(title = "Observed species in time",
       x = "Year of observation",
       y = "Number of species") +
  theme_bw() +
  theme(text=element_text(size = 16))
```

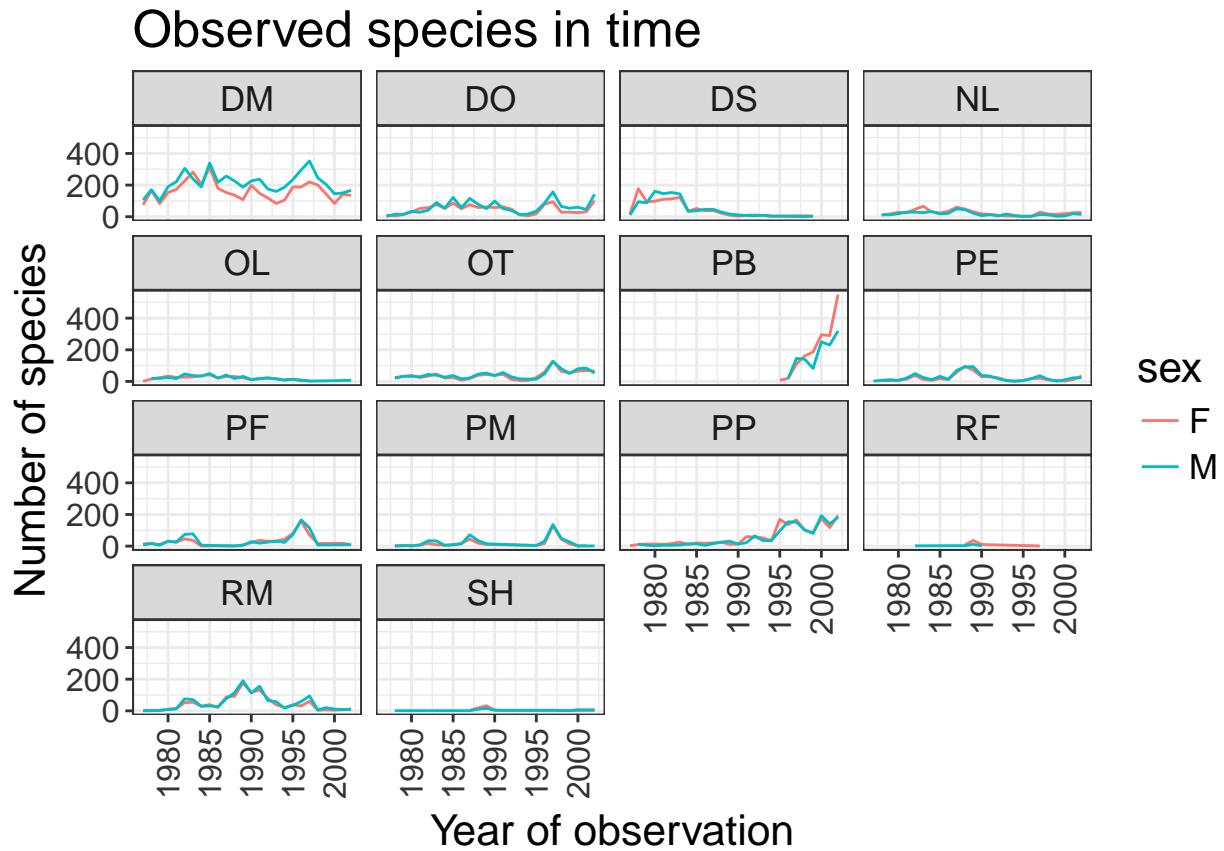
## Observed species in time



Note that it is also possible to change the fonts of your plots. If you are on Windows, you may have to install the **extrafont** package, and follow the instructions included in the README for this package.

After our manipulations, you may notice that the values on the x-axis are still not properly readable. Let's change the orientation of the labels and adjust them vertically and horizontally so they don't overlap. You can use a 90 degree angle, or experiment to find the appropriate angle for diagonally oriented labels:

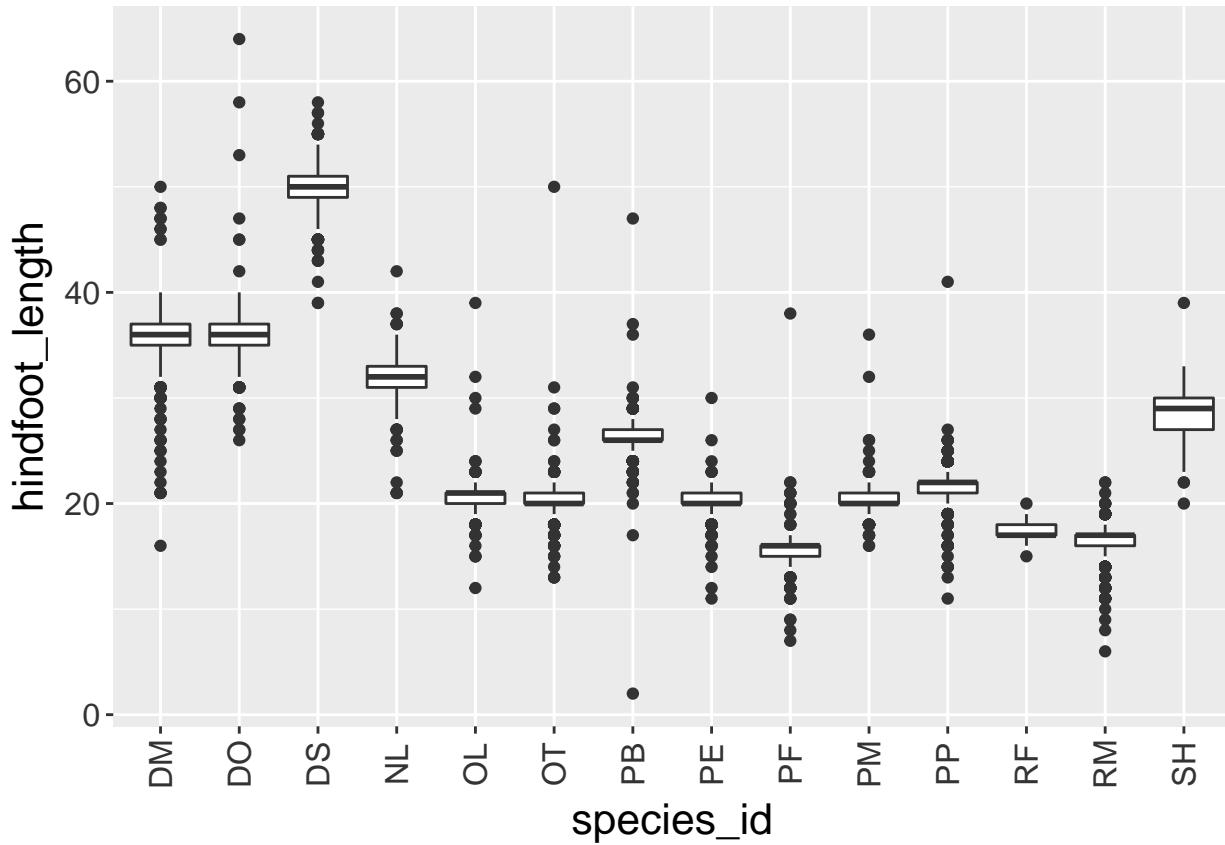
```
ggplot(data = yearly_sex_counts, aes(x = year, y = n, color = sex)) +
  geom_line() +
  facet_wrap(~ species_id) +
  labs(title = "Observed species in time",
       x = "Year of observation",
       y = "Number of species") +
  theme_bw() +
  theme(axis.text.x = element_text(colour = "grey20", size = 12, angle = 90, hjust = 0.5, vjust = 0.5),
        axis.text.y = element_text(colour = "grey20", size = 12),
        text = element_text(size = 16))
```



If you like the changes you created better than the default theme, you can save them as an object to be able to easily apply them to other plots you may create:

```
grey_theme <- theme(axis.text.x = element_text(colour = "grey20", size = 12, angle = 90, hjust = 0.5, v
```

```
axis.text.y = element_text(colour = "grey20", size = 12),
text = element_text(size = 16))
ggplot(surveys_complete, aes(x = species_id, y = hindfoot_length)) +
  geom_boxplot() +
  grey_theme
```



### Challenge

With all of this information in hand, please take another five minutes to either improve one of the plots generated in this exercise or create a beautiful graph of your own. Use the RStudio `ggplot2` cheat sheet for inspiration.

Here are some ideas:

- See if you can change the thickness of the lines.
- Can you find a way to change the name of the legend? What about its labels?
- Try using a different color palette (see [http://www.cookbook-r.com/Graphs/Colors\\_\(ggplot2\)/](http://www.cookbook-r.com/Graphs/Colors_(ggplot2)/)).

### Arranging and exporting plots

Faceting is a great tool for splitting one plot into multiple plots, but sometimes you may want to produce a single figure that contains multiple plots using different variables or even different data frames. The `gridExtra` package allows us to combine separate ggplots into a single figure using `grid.arrange()`:

```
library(gridExtra)

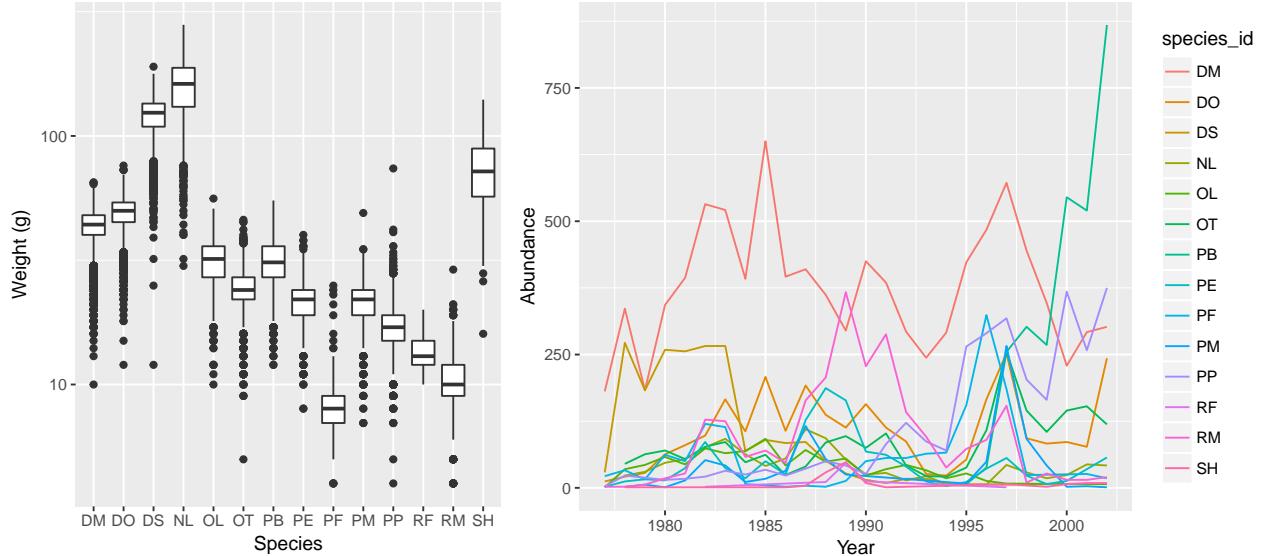
spp_weight_boxplot <- ggplot(data = surveys_complete, aes(x = species_id, y = weight)) +
  geom_boxplot() +
  xlab("Species") + ylab("Weight (g)") +
  scale_y_log10()
```

```

spp_count_plot <- ggplot(data = yearly_counts, aes(x = year, y = n, color = species_id)) +
  geom_line() +
  xlab("Year") + ylab("Abundance")

grid.arrange(spp_weight_boxplot, spp_count_plot, ncol = 2, widths = c(4, 6))

```



In addition to the `ncol` and `nrow` arguments, used to make simple arrangements, there are tools for constructing more complex layouts.

After creating your plot, you can save it to a file in your favorite format. The Export tab in the **Plot** pane in RStudio will save your plots at low resolution, which will not be accepted by many journals and will not scale well for posters.

Instead, use the `ggsave()` function, which allows you easily change the dimension and resolution of your plot by adjusting the appropriate arguments (`width`, `height` and `dpi`):

```

my_plot <- ggplot(data = yearly_sex_counts, aes(x = year, y = n, color = sex)) +
  geom_line() +
  facet_wrap(~ species_id) +
  labs(title = "Observed species in time",
       x = "Year of observation",
       y = "Number of species") +
  theme_bw() +
  theme(axis.text.x = element_text(colour = "grey20", size = 12, angle = 90, hjust = 0.5, vjust = 0.5),
        axis.text.y = element_text(colour = "grey20", size = 12),
        text=element_text(size = 16))
  ggsave("name_of_file.png", my_plot, width = 15, height = 10)

## This also works for grid.arrange() plots
combo_plot <- grid.arrange(spp_weight_boxplot, spp_count_plot, ncol = 2, widths = c(4, 6))
ggsave("combo_plot_abun_weight.png", combo_plot, width = 10, dpi = 300)

```

Note: The parameters `width` and `height` also determine the font size in the saved plot.