Introduction to ggplot2

Visualizing data in R

Download the section 4 .Rmd handout to `STAT240/lecture/sect04-ggplot`.

Download the file `penguins.csv` to `STAT240/data`

Material in this section is covered by Chapter 6 on the notes website.

The Palmer penguins dataset records physical measurements of penguins taken at Palmer Research Station.

Each row is a different penguin individual.

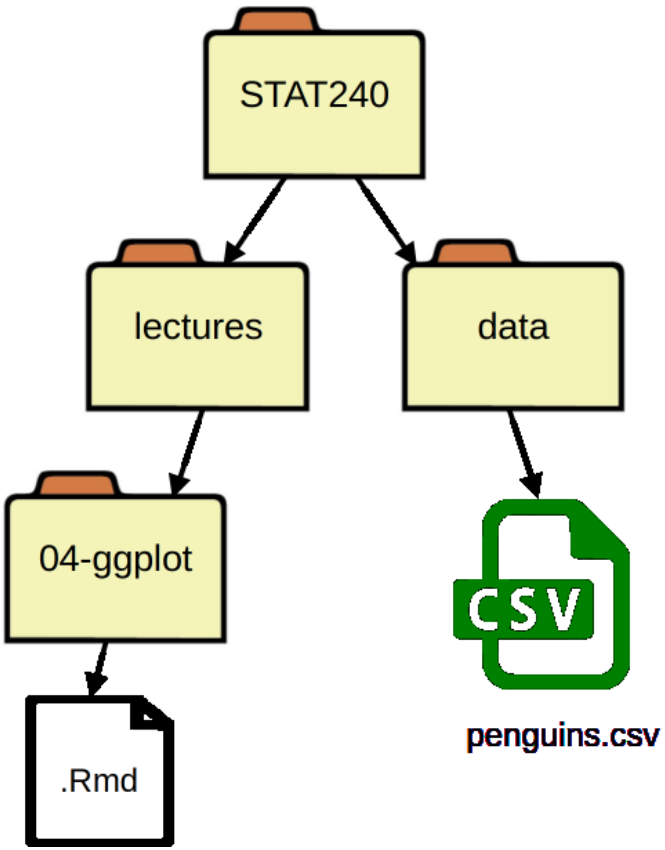
- Three species
- Measured at different locations/times
- Several physical characteristics + sex

Load the data with the `read_csv()` command.

Explore the data with `View` and `glimpse`.

Note the variable types of each column.

```
read_csv("../../data/penguins.csv")
```



`ggplot2` stands for "grammar of graphics".

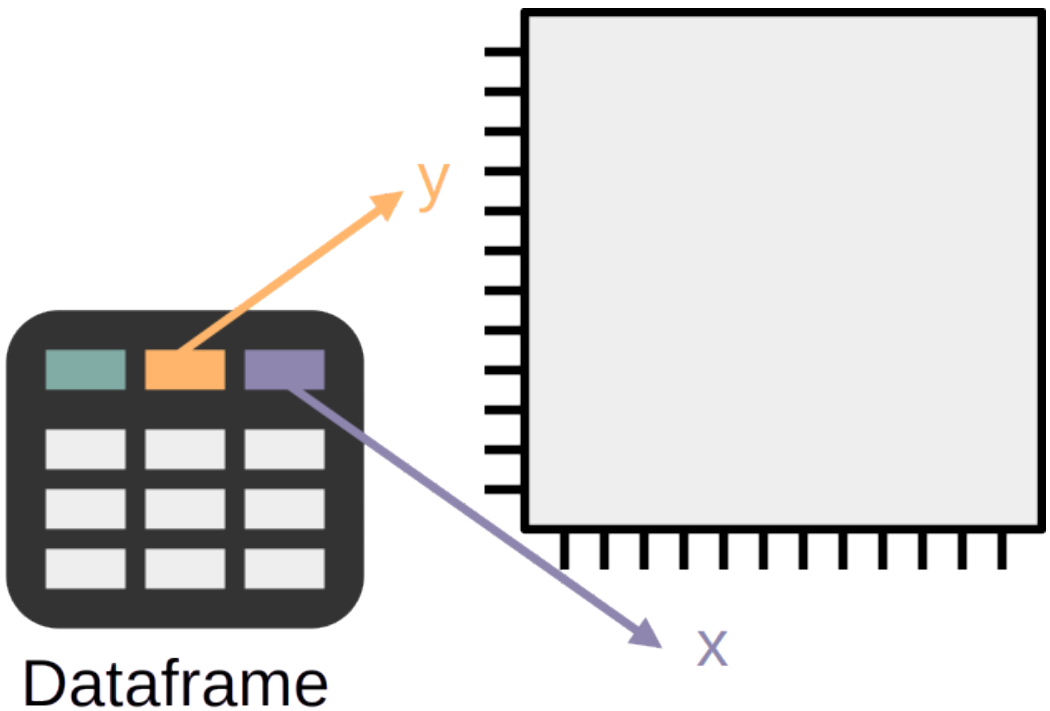
- Create different graph types with similar code
- Rich customization tools

Code will have have a specific structure.

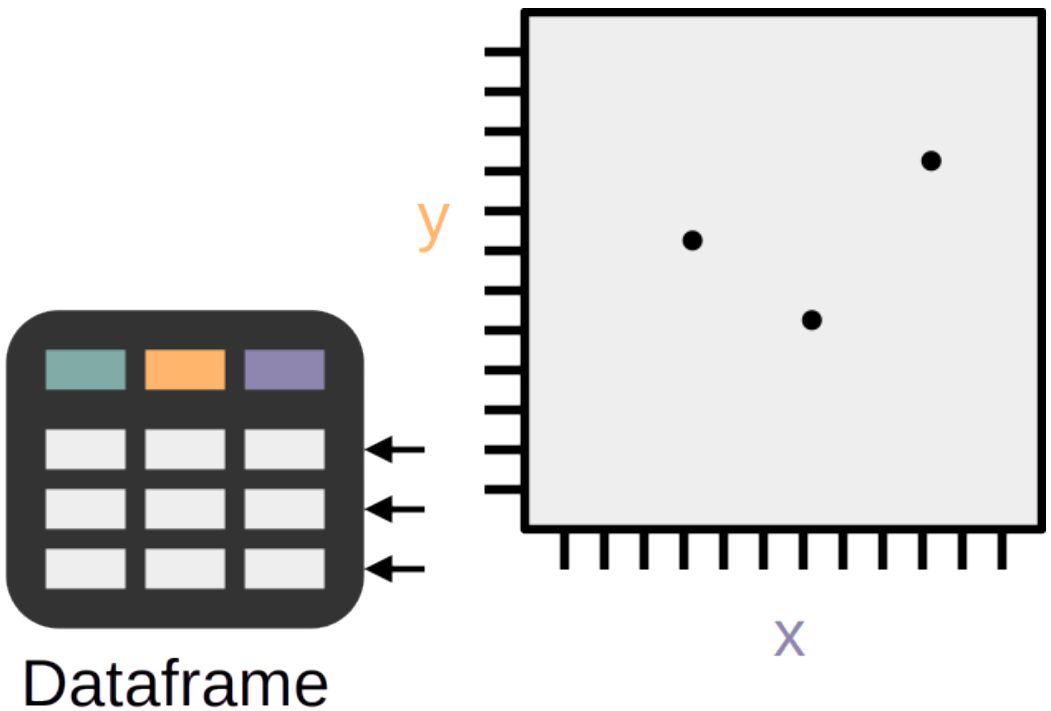
Let's build a plot to explore the relationship between body mass and flipper length.

What types of graphs could we use to answer this question?

`ggplot()` builds a canvas based on a **mapping**:



Use a `geom` to add markings:



Customization options go in the chosen `geom` function. For example:

- Color
- Shape
- Size
- Transparency

There are dozens of geometries!

- `geom_line()`
- `geom_point()`
- `geom_text()`
- `geom_smooth()`
- `geom_boxplot()`
- `geom_histogram()`
- `geom_density()`
- `geom_bar()`

And more...

Let's study the flipper length variable on its own.

Histograms, density plots, and boxplots visualize a single numeric variable.

`geom_histogram()` divides the data into bins.

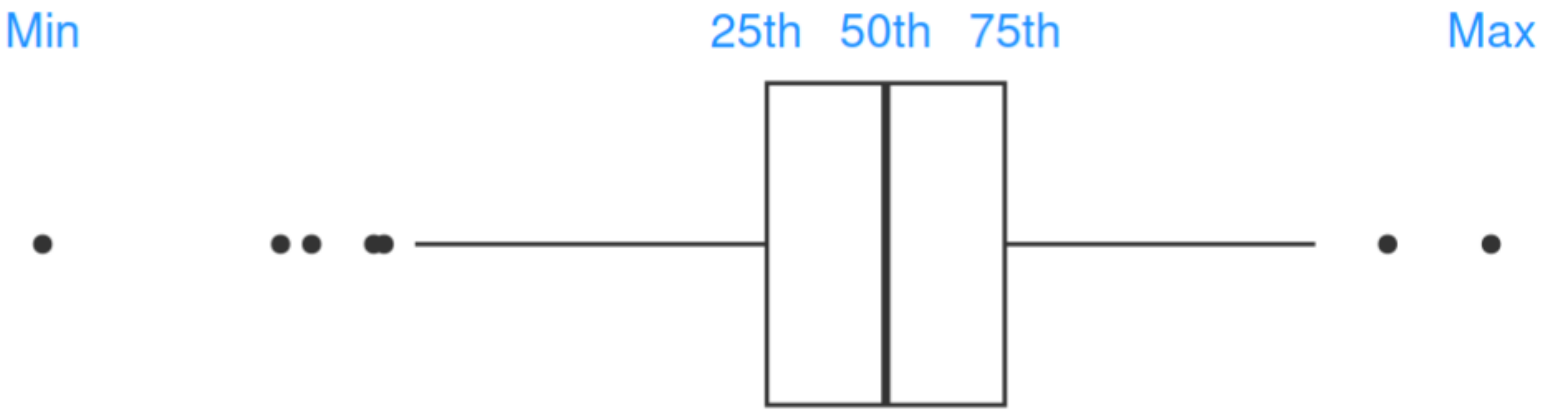
- `binwidth`: how wide the bins are
- `bins`: the number of bins

- `center`: midpoint of a bin
- `boundary`: a specific breakpoint

Use only one of (`binwidth`, `bins`) and only one of (`center`, `boundary`).

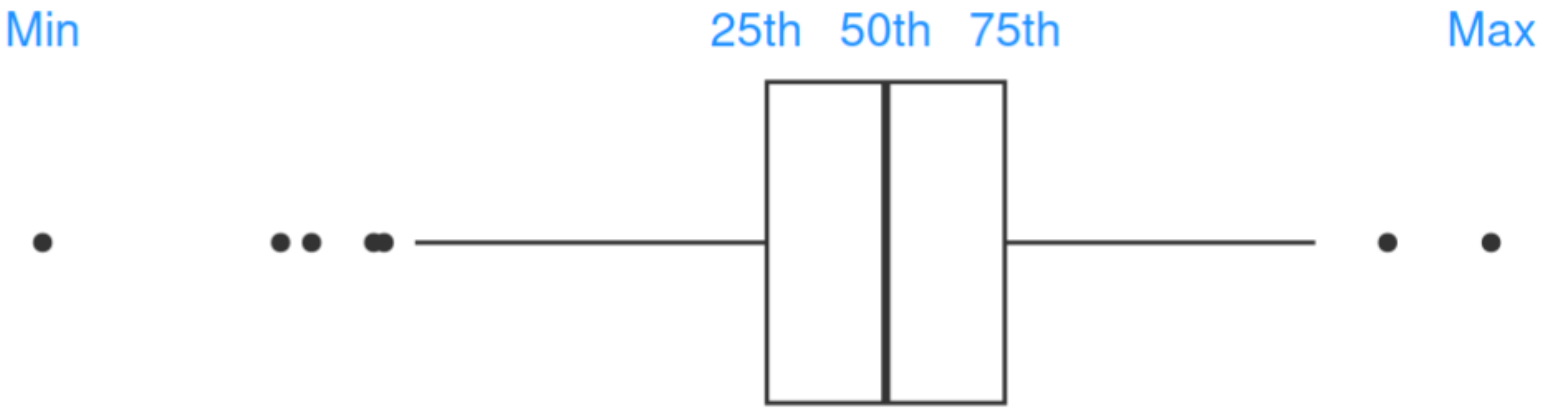
A `geom_density()` plot is similar to a histogram, but with a smooth curve.

- Shows "general trend"
- Related to integration



`geom_boxplot()` shows the **quartiles**.

- Outliers are drawn as dots



The box is the **interquartile range** (IQR).

- The "threshold" for outliers is 1.5× IQR
- Anything 1.5 "box lengths" away is a dot

(The lines only go out to data that exists.)

Now, let's compare flipper length across species.

- Add `fill = species` to color-code the plots.
- What if we use `col = species` instead?
- Make a change to the density plot to make the overlapping plots more readable.

A bar graph visualizes a single categorical variable.

Draw bars (similar to a histogram) based on the number of items in each category.

Two options: `geom_bar()` and `geom_col()`.

`geom_bar()`	`geom_col()`
Only x or y	Both x and y
Always gives counts	More flexible
Less manual calculation	Provide the bar height

Using local aesthetics, we can apply a mapping to one layer at a time.

The `color` aesthetic in `geom_point()` did not affect `geom_smooth()`.

Both geoms follow `x` and `y` from the original `ggplot()`.

Variable aesthetics are either:

- Global: apply to all layers
- Local: affect one layer

Constant aesthetics are always local.

Here are the most common geometries we'll use:

- `geom_point()`
- `geom_line()`
- `geom_smooth()`
- `geom_bar()`
- `geom_col()`
- `geom_boxplot()`
- `geom_histogram()`
- `geom_density()`

The ggplot2 package offers rich customization options.

The lectures won't include maximum detail.
Instead, I'll link to some references to use as needed.

We can annotate plots with lines and text.

- How to add reference lines
- How to add text annotations

Use the `labs` addition to customize labels.

- Title, subtitle, and caption
- Edit labels for any mapping in the graph

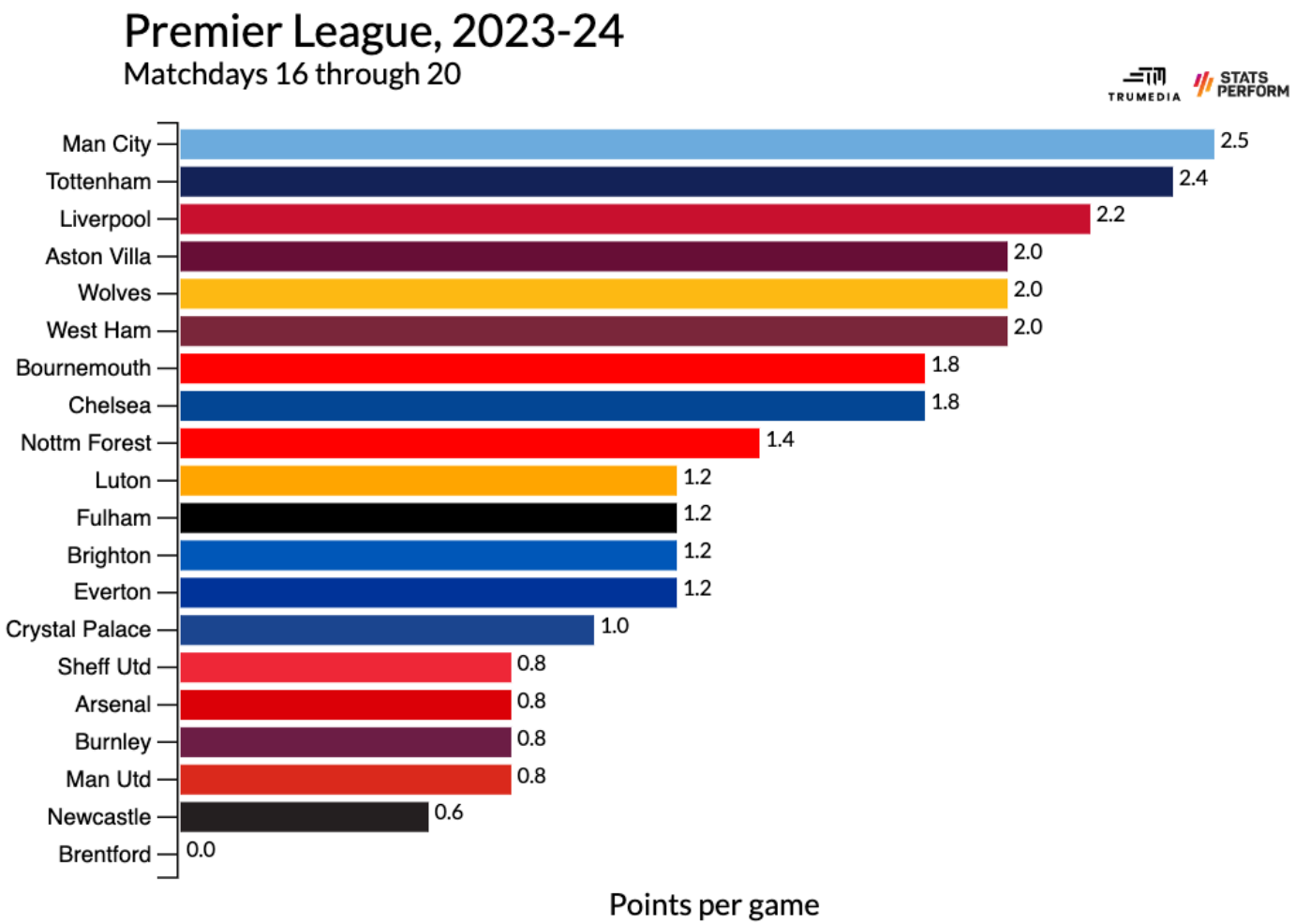
We can change the axes to be more informative.

- Use `scale_x` and `scale_y` to specify the axis
- Can be `continuous` or `discrete`
- Here is more detailed documentation

The most fun part is choosing a color scheme.

- Colorblind friendly built-in scales in `viridis`
- Can make your own custom scale with `manual`
- Specify `d` or `c` for discrete and continuous

Here are the viridis options.
Here is a list of predefined R colors.



Recreate this graphic using the partial dataframe.

Bonus topics:

- Faceting
  - `facet_grid()` and `facet_wrap()`
- Mathematical functions
  - `geom_function()`