Lecture_7_Week3

Thrisha Rajkumar

2024-12-26

load packages: the Palmer penguins data set & tidyverse

First, load tidyverse

```
library(tidyverse)
```

```
## -- 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 1.9.3     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 errors
```

Second, we will be based on the Palmer penguins data set, so load the dataset

```
library(palmerpenguins)
head(penguins)
```

```
## # A tibble: 6 x 8
##   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
## 1 Adelie  Torgersen           39.1          18.7               181        3750
## 2 Adelie  Torgersen           39.5          17.4               186        3800
## 3 Adelie  Torgersen           40.3          18                 195        3250
## 4 Adelie  Torgersen           NA            NA                  NA          NA
## 5 Adelie  Torgersen           36.7          19.3               193        3450
## 6 Adelie  Torgersen           39.3          20.6               190        3650
## # i 2 more variables: sex <fct>, year <int>
```

We also need the ggplot2 package, which is contained in tidyverse. The ggplot2 package implements Leland Wilkinson's Grammar of Graphics: 1. An aesthetic is a mapping between a variable and a visual cue. 2. A glyph is a basic graphical element e.g. a mark or symbol. 3. A guide is an annotation which provides context.

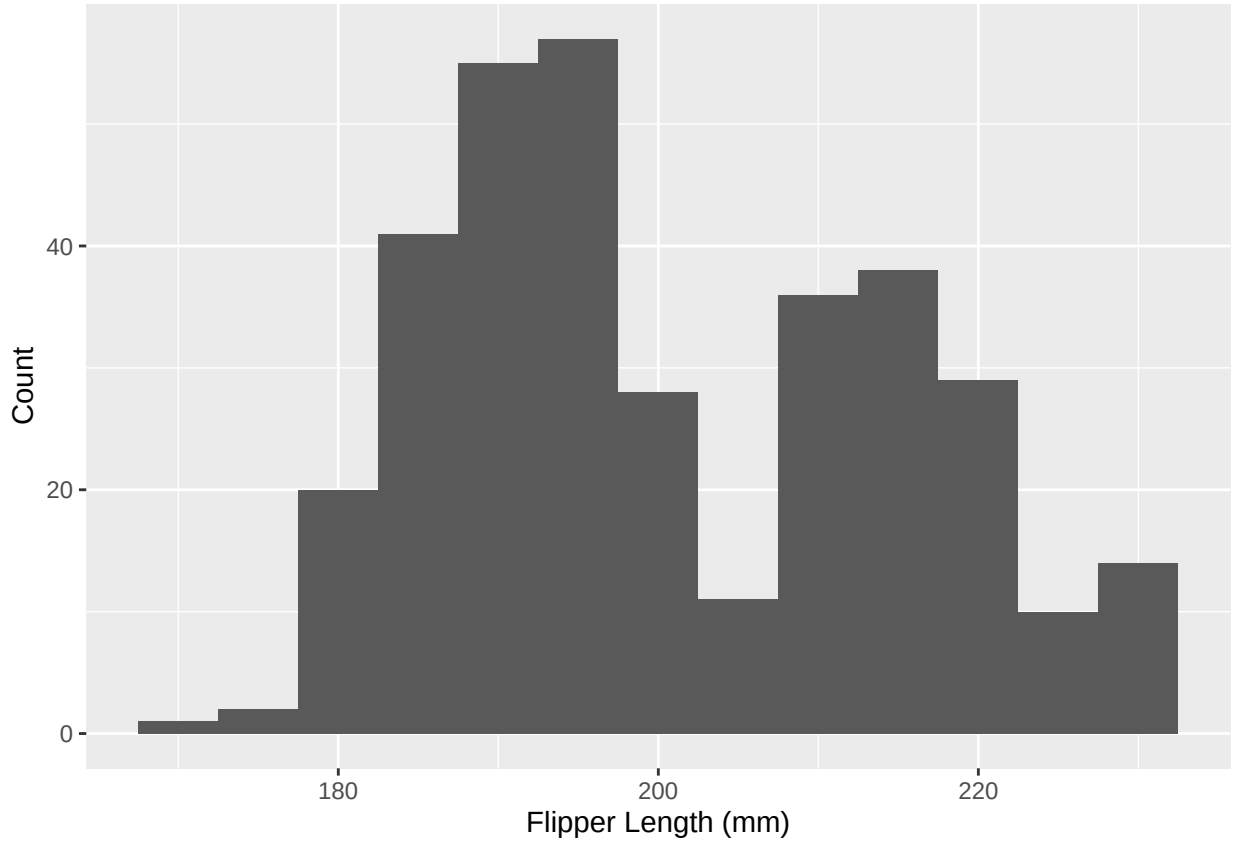
Univariate plot - histogram

First, we create a plot object using the ggplot function. The plot object specifies the aesthetic (using the aes() function).

A histogram plot with an aesthetic that maps Flipper length to horizontal position. Using the geom_histogram function to plot histogram

```
univar_plot <- ggplot(data=penguins, aes(x=flipper_length_mm)) + xlab("Flipper Length (mm)")
univar_plot+geom_histogram(binwidth = 5)+ylab("Count")
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



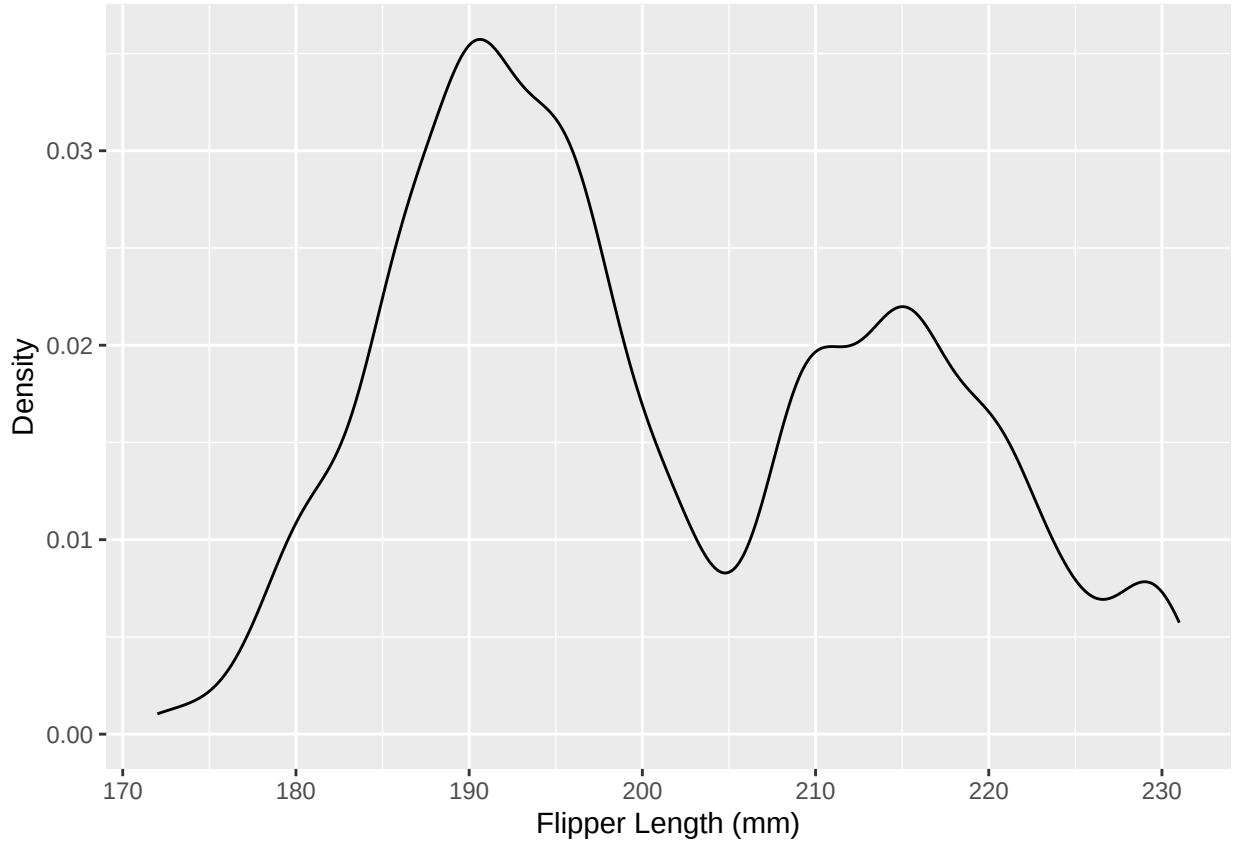
Note that we have created guides using xlab & ylabels;

Univariate plot - density plot

Replacing the above geom_histogram with geom_density to plot density:

```
univar_plot+geom_density(adjust=0.5)+ylab('Density')
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_density()`).
```

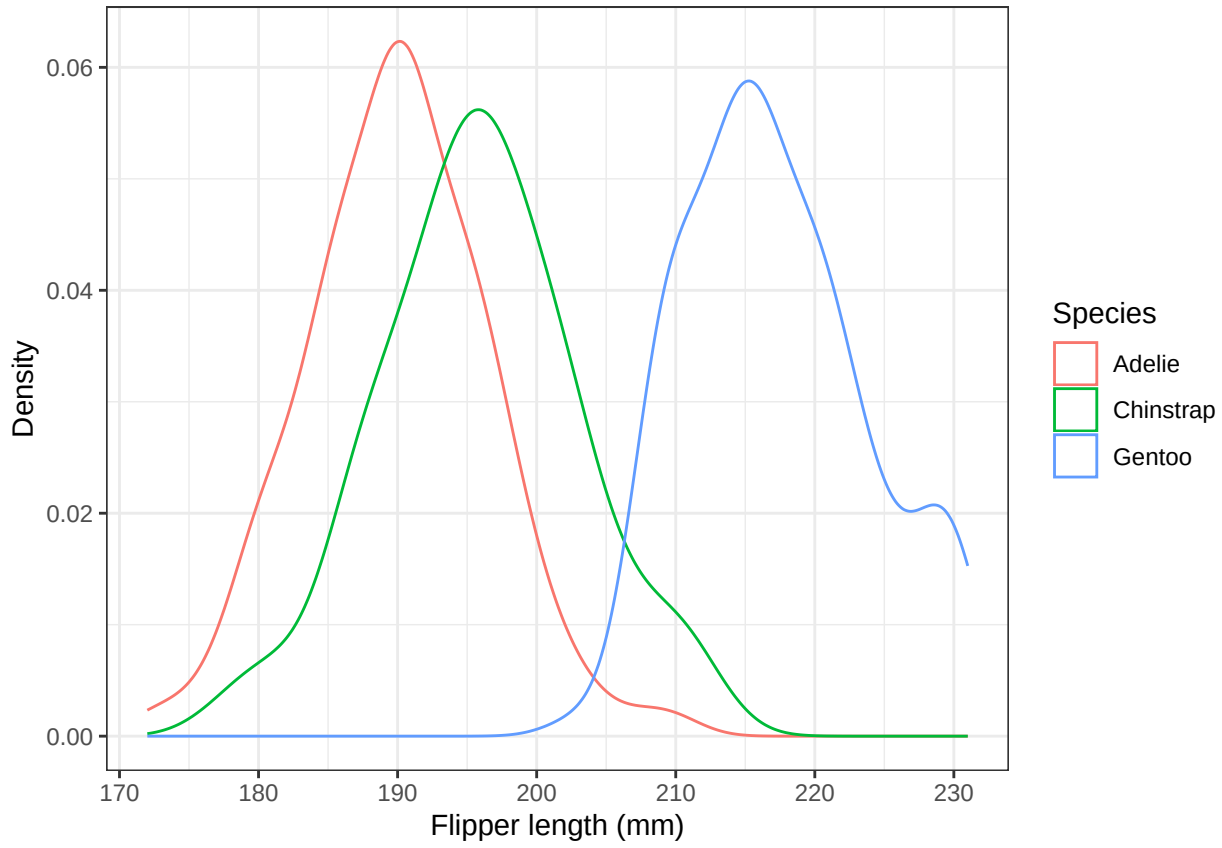


Bivariate plots - density

Adding another aesthetic that maps species to color, and use geom_density in a simlar way to the above density plot:

```
ggplot(data=rename(penguins, Species=species), aes(x=flipper_length_mm, color=Species))+
  geom_density()+theme_bw()+xlab("Flipper length (mm)")+ylab("Density")
```

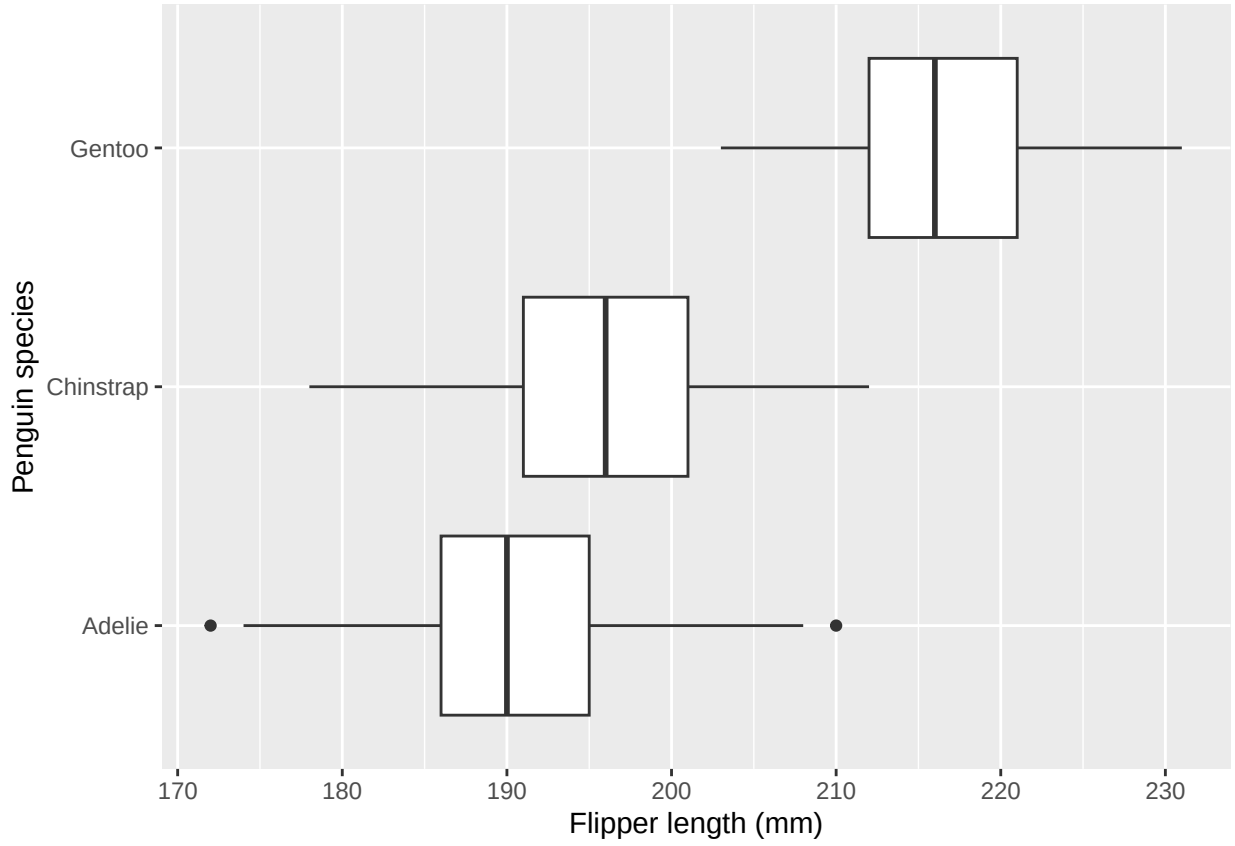
```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_density()`).
```



Bivariate plots - box

```
ggplot(data=penguins, aes(x=flipper_length_mm, y=species))+geom_boxplot()+
  xlab('Flipper length (mm)') + ylab("Penguin species")
```

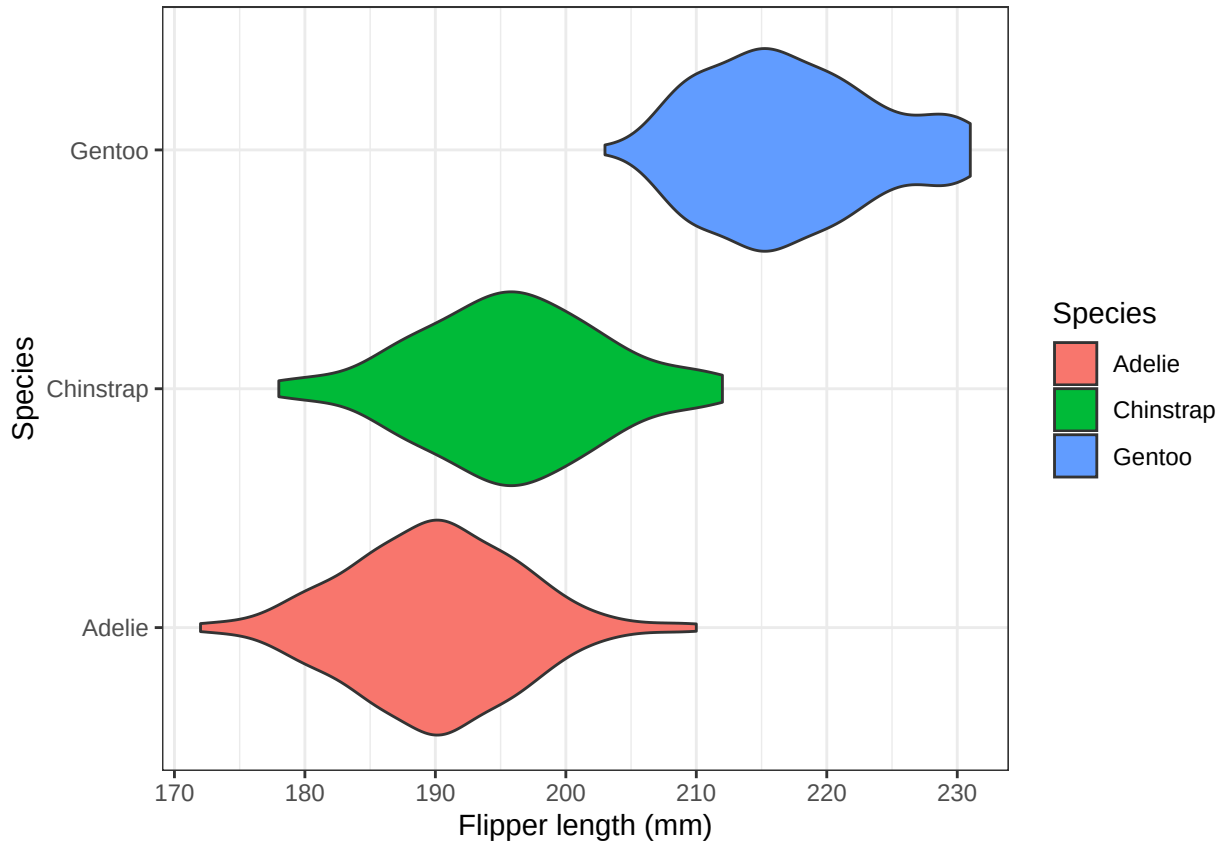
```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



Bivariate plots - violin

```
ggplot(data=rename(penguins, Species=species), aes(x=flipper_length_mm, y=Species, fill=Species))+geom_
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_ydensity()`).
```

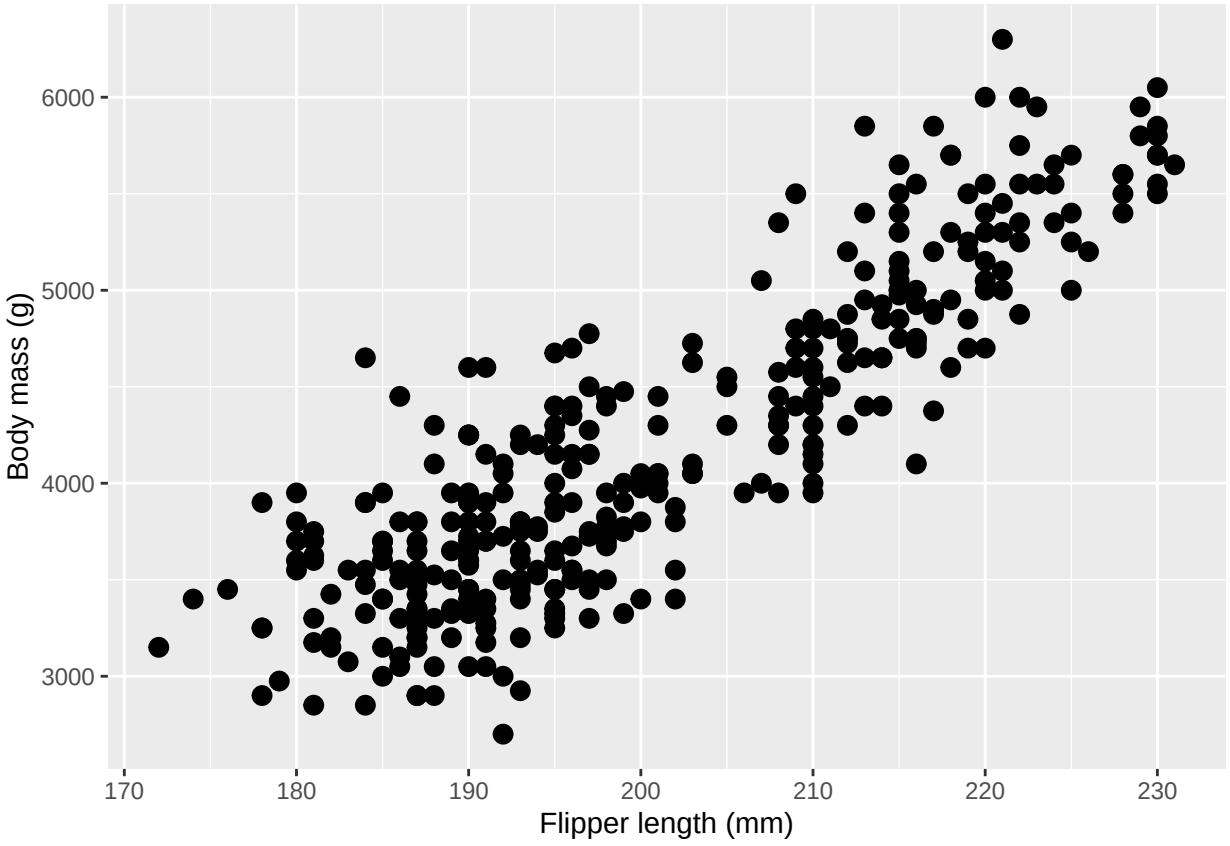


Bivariate plots - scatter

Scatter plot for flipper length vs body mass, using geom_point:

```
mass_flipper_scatter <- ggplot(data=penguins, aes(y=body_mass_g, x=flipper_length_mm))+
  xlab("Flipper length (mm)") + ylab("Body mass (g)")
mass_flipper_scatter+geom_point(size=3)
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

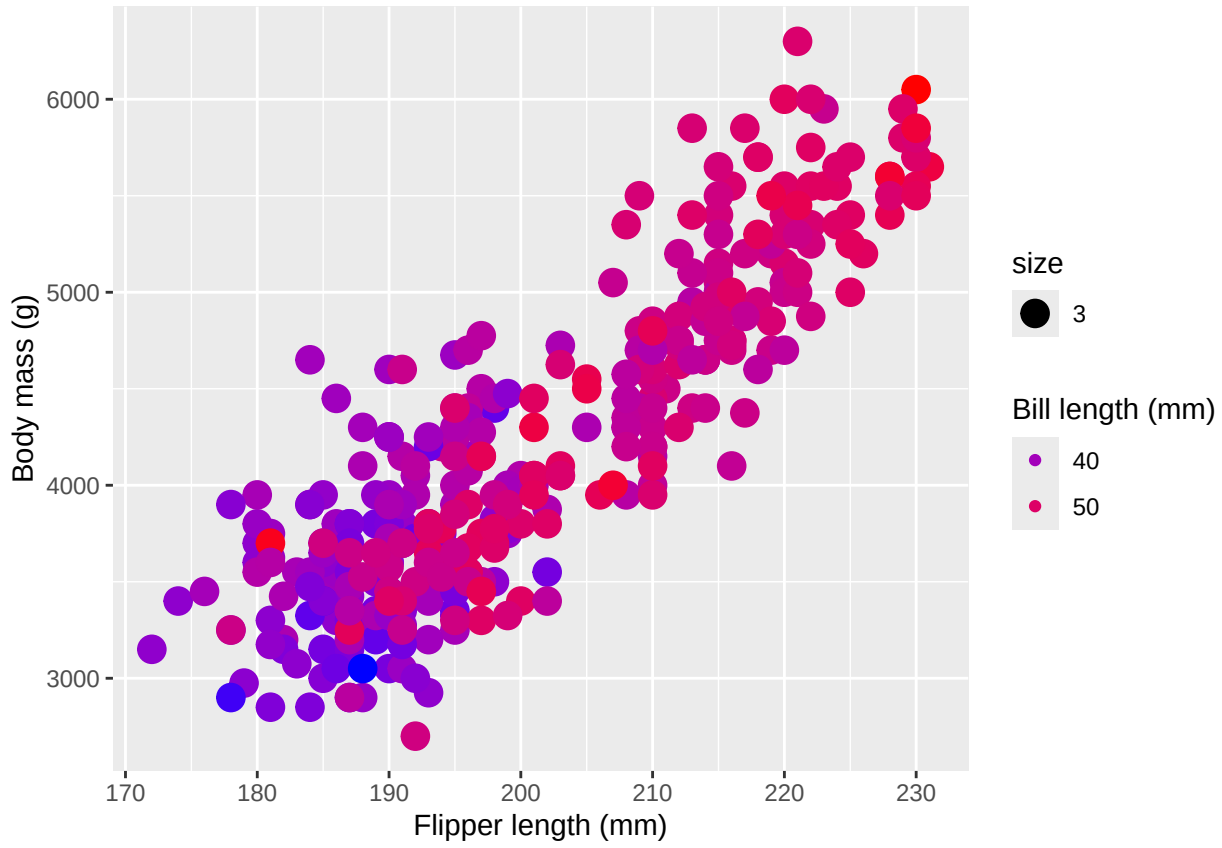


Adding more aesthetics

Adding another aesthetic using arguments in the geom_point function (Bill length to color):

```
mass_flipper_scatter+geom_point(aes(color=bill_length_mm, size=3))+
  scale_color_gradient(low="blue", high="red")+guides(color=guide_legend("Bill length (mm)"))
```

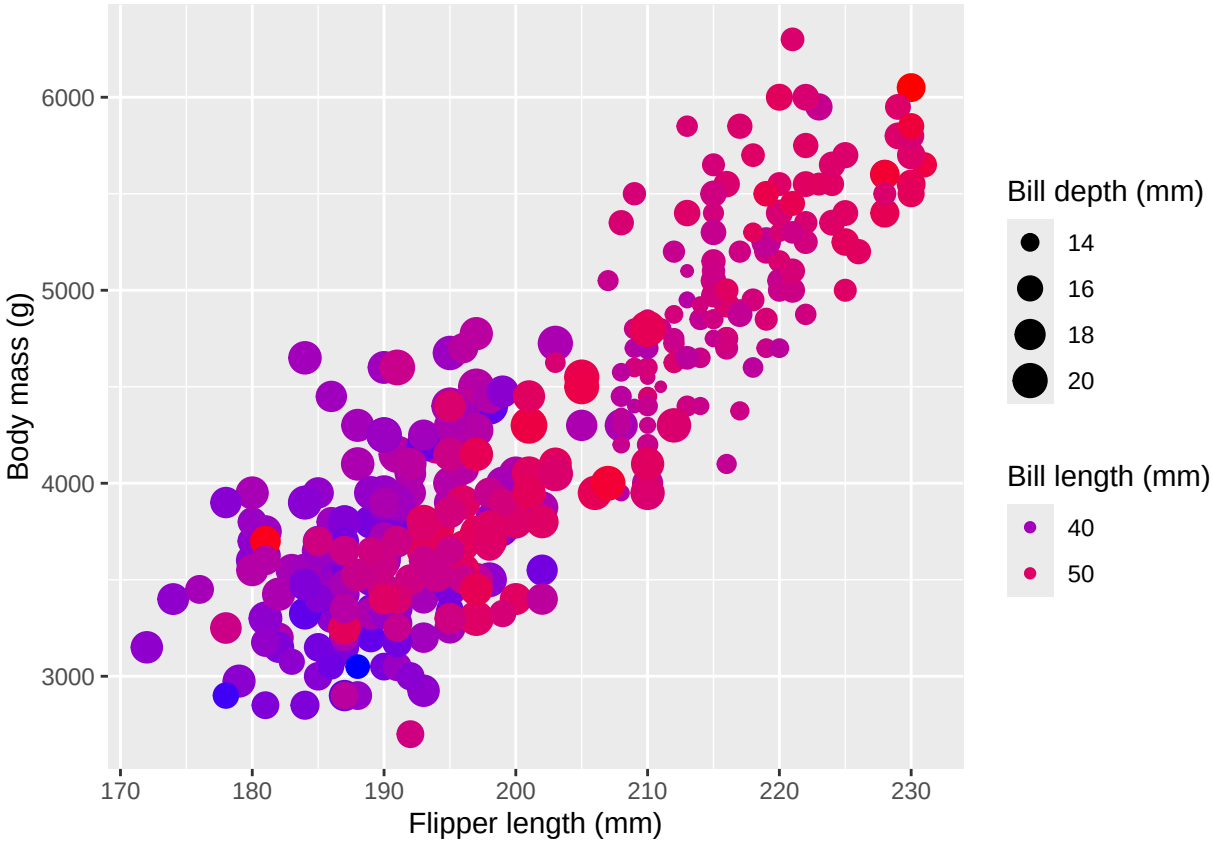
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Adding one more aesthetic (Bill depth to size):

```
mass_flipper_scatter+geom_point(aes(color=bill_length_mm, size=bill_depth_mm))+
  scale_color_gradient(low="blue", high="red")+
  guides(color=guide_legend("Bill length (mm)"), size=guide_legend("Bill depth (mm)"))
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

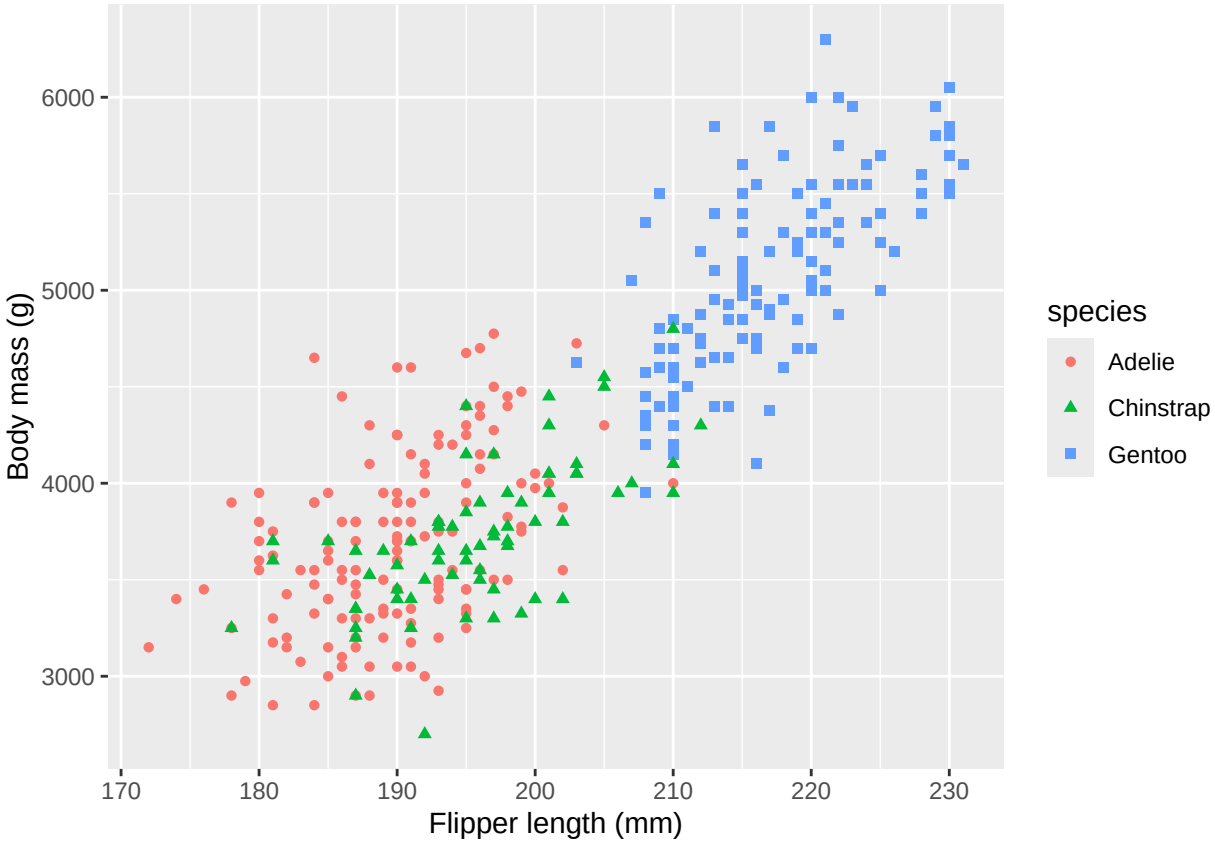


Multivariate plots

Changing two of the aesthetic, one which maps species to color, another one with map species to shape:

```
mass_flipper_scatter+geom_point(aes(color=species, shape=species))
```

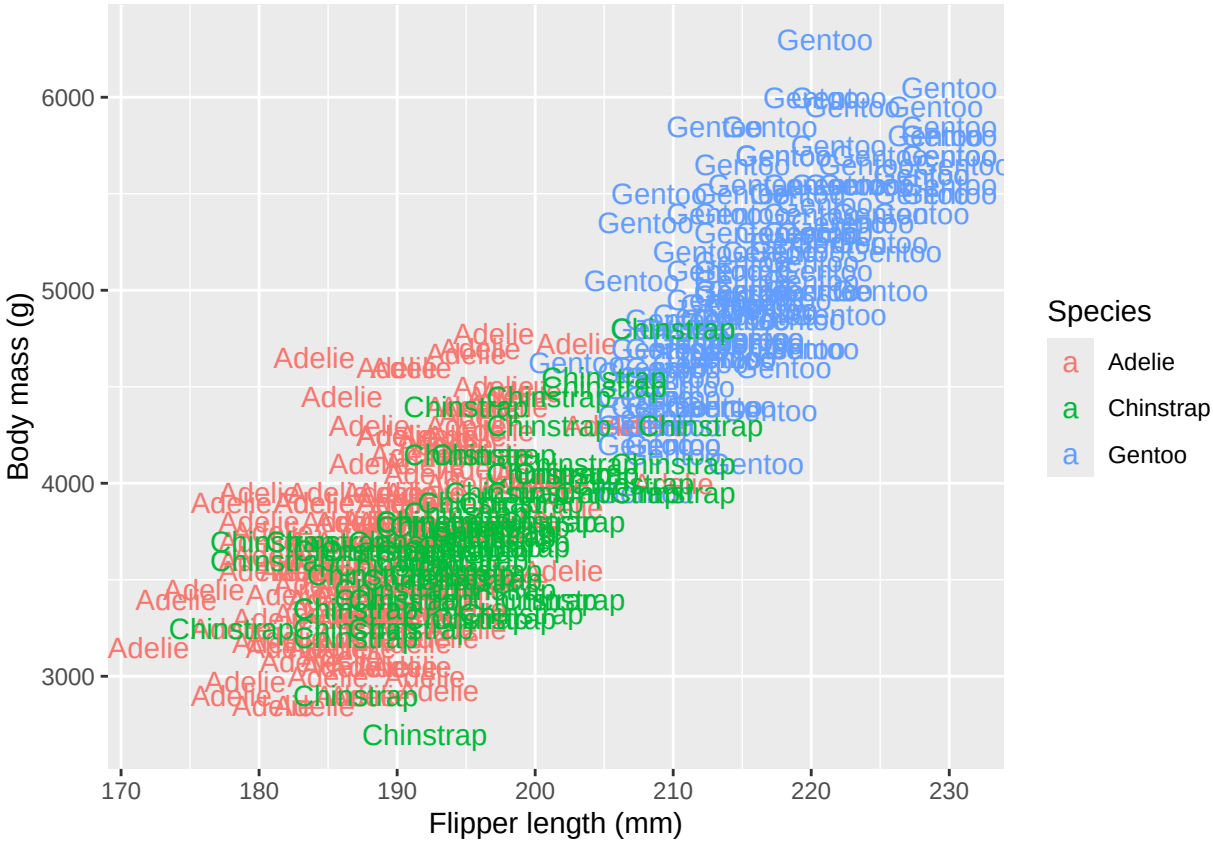
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



One can also map species to text:

```
mass_flipper_scatter + geom_text(aes(label=species, color=species)) +
  guides(color=guide_legend("Species"))
```

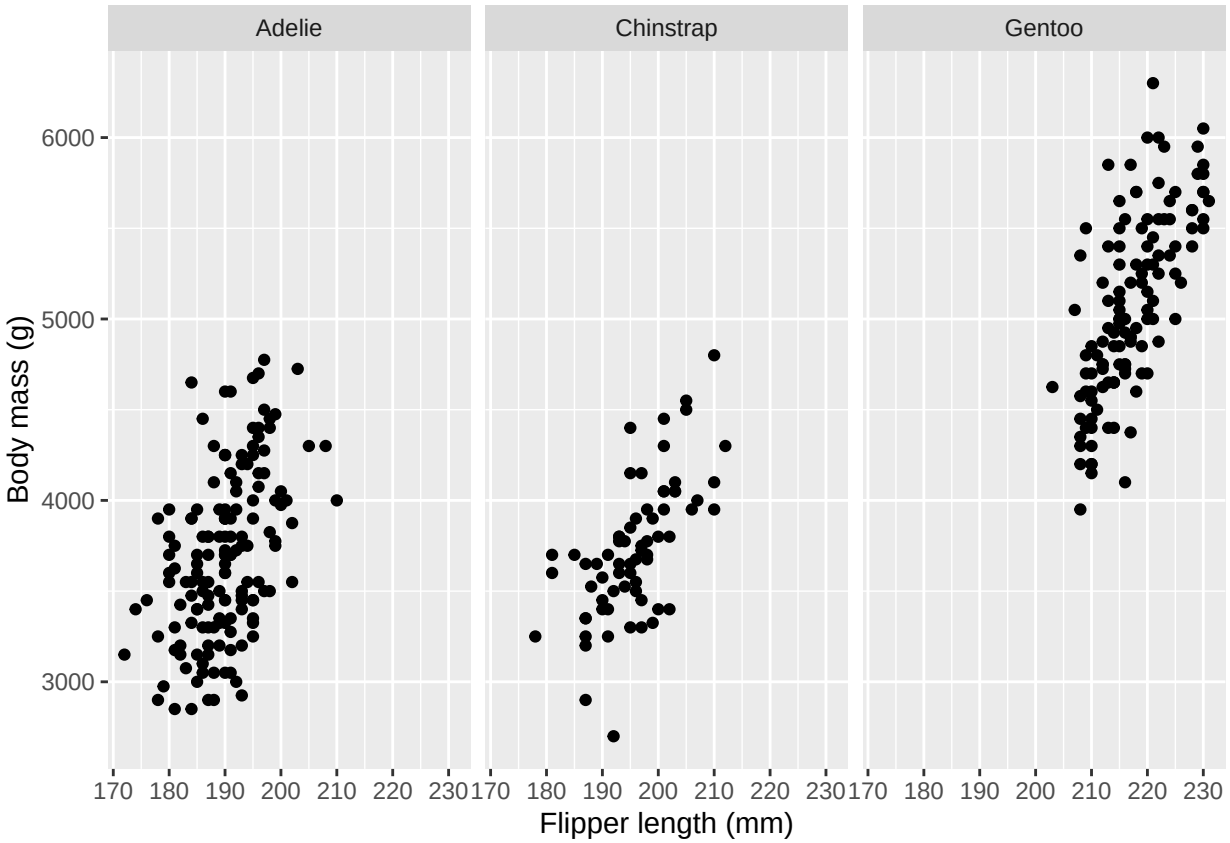
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_text()`).
```



Alternatively, one can use facets to display a categorical variable (species):

```
mass_flipper_scatter + geom_point() + facet_wrap(~species)
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Trend lines

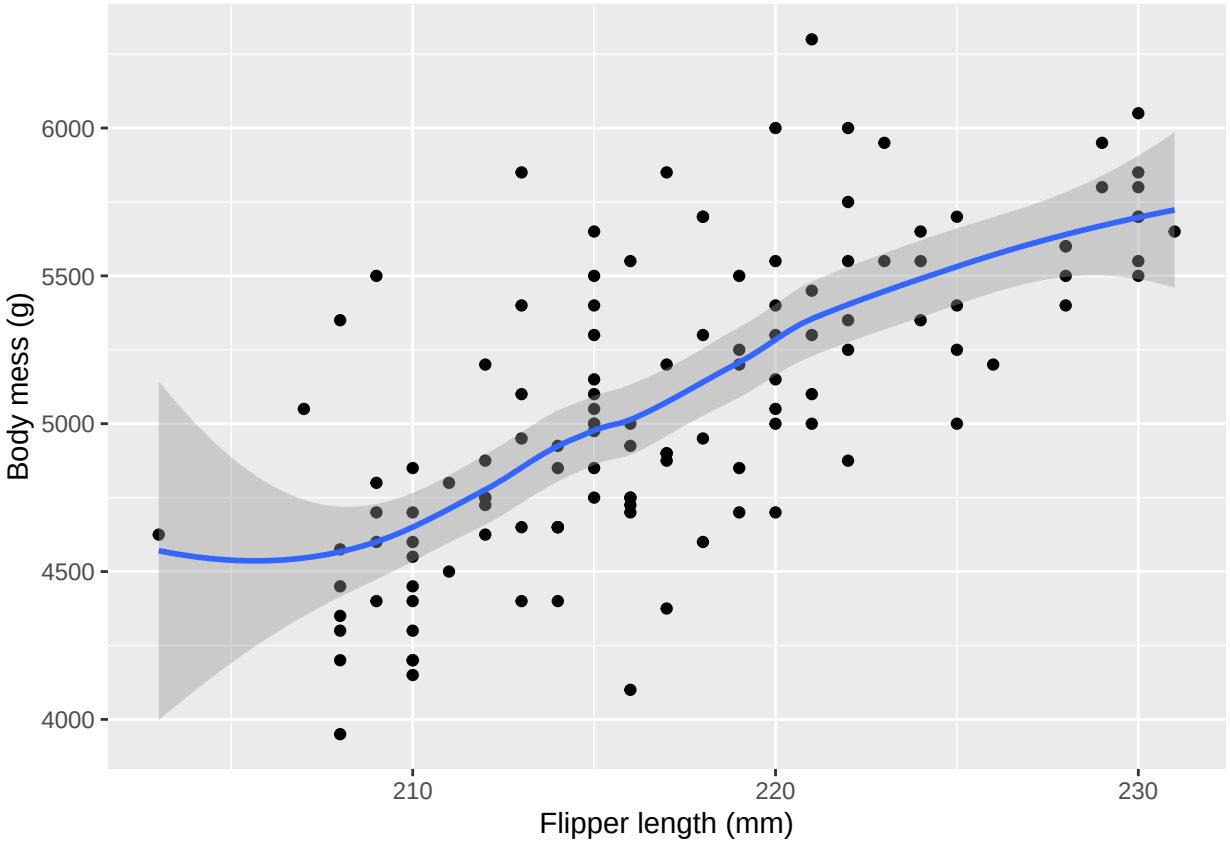
Create trend line to illustrate the relationship between two variables (e.g., flipper length and body mass), using the geom_smooth function:

```
trend_plot <- ggplot(data=filter(penguins, species=='Gentoo'), aes(y=body_mass_g, x=flipper_length_mm))
trend_plot + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```



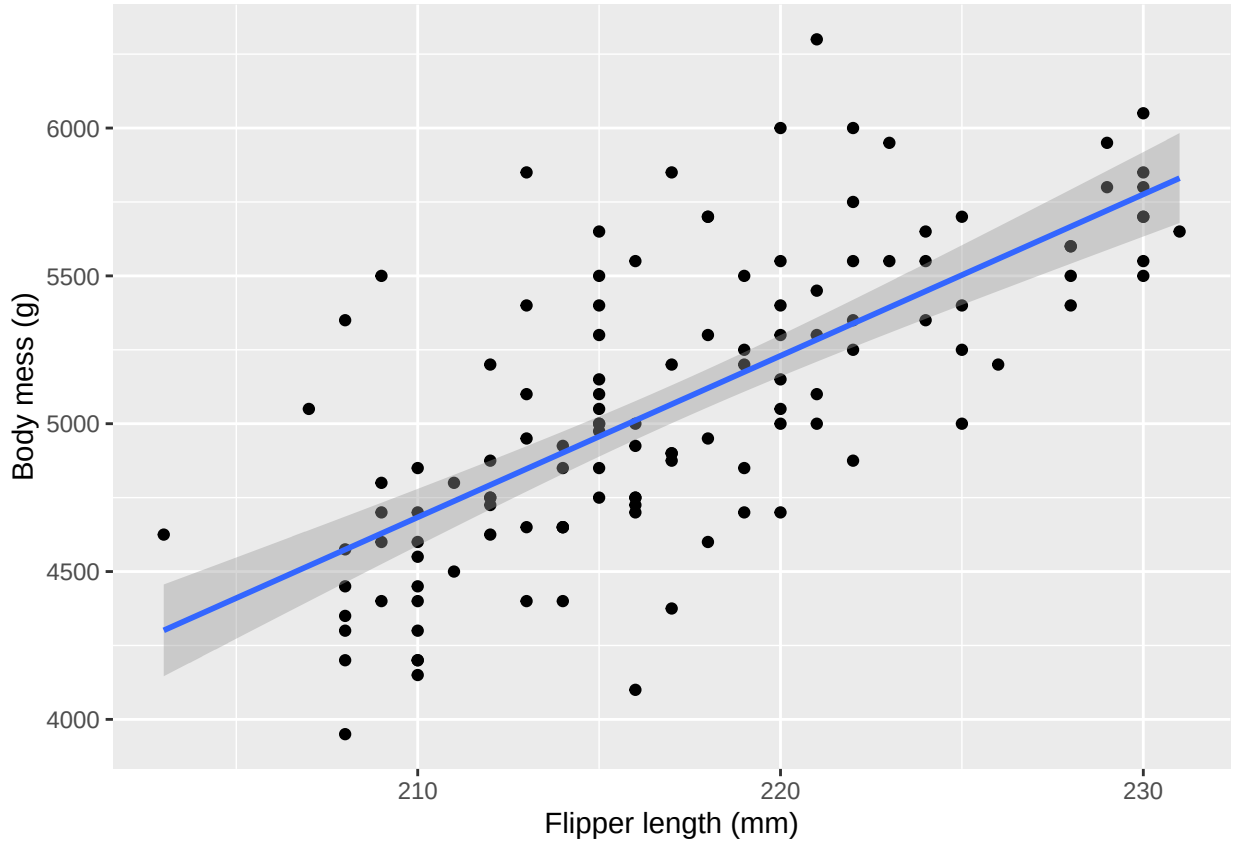
Alternatively, we can add a linear trend line, using method "lm" of geom_smooth:

```
trend_plot+geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```



Adding annotation to the plot

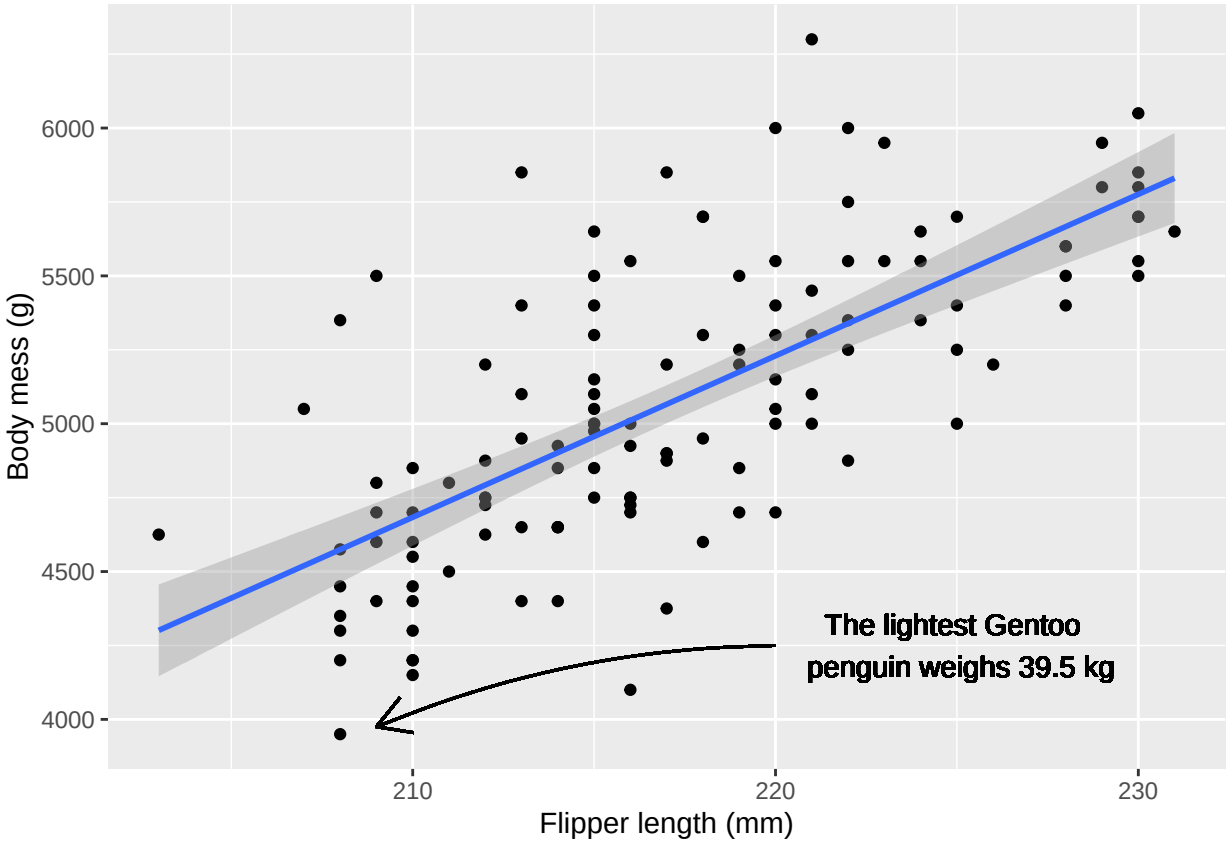
Using a combination of geom_curve and geom_text:

```
trend_plot + geom_smooth(method="lm") +
  geom_curve(x=220, xend=209, y=4250, yend=3975, arrow=arrow(length=unit(0.5, 'cm')), curvature=0.1) +
  geom_text(x=225, y=4250, label="The lightest Gentoo \n penguin weighs 39.5 kg")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```


6000
5500
5000
4500
4000
Body mess (g)
210
220
230
Flipper length (mm)
The lightest Gentoo
penguin weighs 39.5 kg