AE 02: Visualizing penguins

> **!** Important
>
> Go to the course GitHub organization and locate the repo titled `ae-02-YOUR_GITHUB_USERNAME` to get started.
> This AE is due Sunday, Sep 11 at 11:59pm.

For all analyses, we'll use the **tidyverse** and **palmerpenguins** packages.

```
library(tidyverse)
library(palmerpenguins)
```

The dataset we will visualize is called `penguins`. Let's `glimpse()` at it.

```
glimpse(penguins)
```

```
Rows: 344
Columns: 8
$ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel~
$ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse~
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, ~
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, ~
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186~
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, ~
$ sex               <fct> male, female, female, NA, female, male, female, male~
$ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

Visualizing penguin weights - Demo

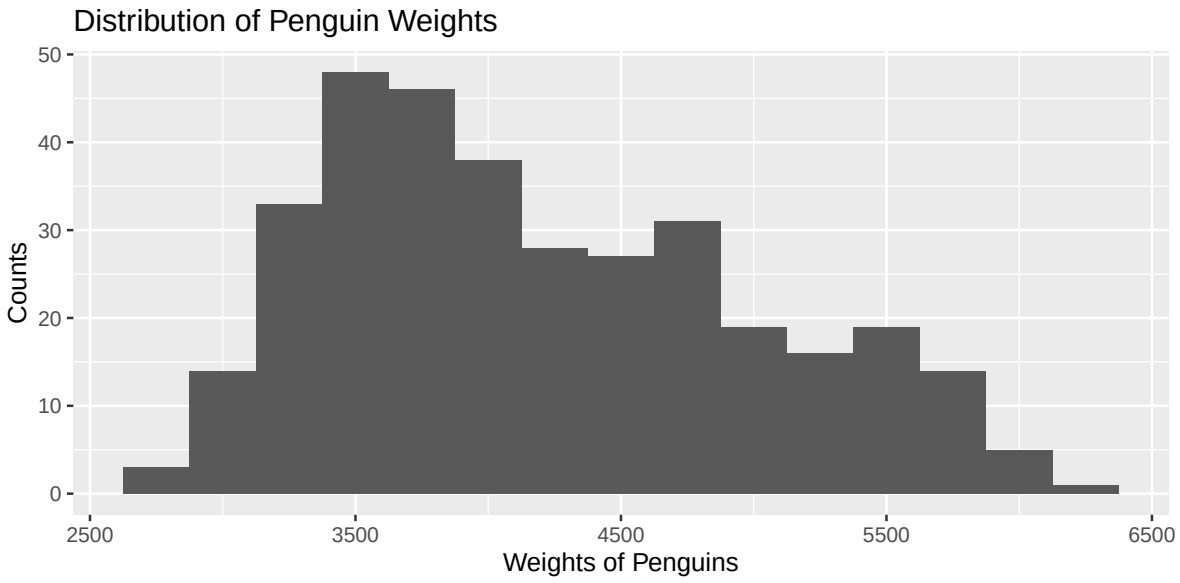
Single variable

> **Note**
>
> Analyzing the a single variable is called **univariate** analysis.

Create visualizations of the distribution of `weights` of penguins.

1. Make a histogram. Set an appropriate binwidth.

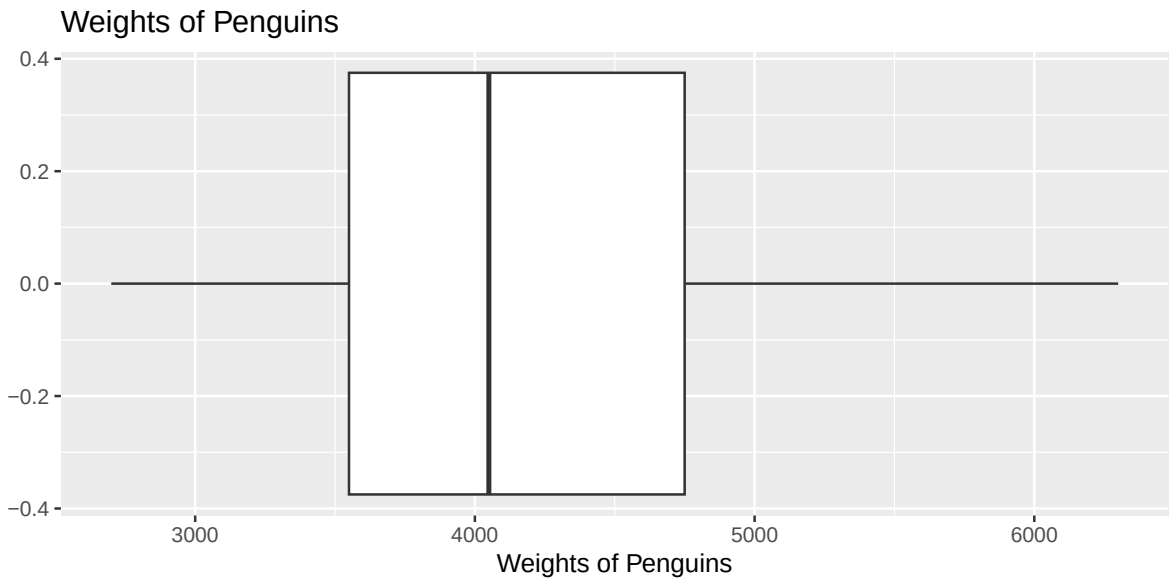
```
ggplot(penguins, aes(x = body_mass_g)) +
  geom_histogram(binwidth = 250) +
  labs(
    x = "Weights of Penguins",
    y = "Counts",
    title = "Distribution of Penguin Weights"
  )
```



2. Make a boxplot.

```
ggplot(penguins, aes(x = body_mass_g)) +
  geom_boxplot() +
```

```
  labs(
    x = "Weights of Penguins",
    title = "Weights of Penguins")
```



3. Based on these, determine if each of the following statements about the shape of the distribution is true or false.
    - The distribution of penguin weights in this sample is left skewed. **FALSE**
    - The distribution of penguin weights in this sample is unimodal. **TRUE**

Two variables

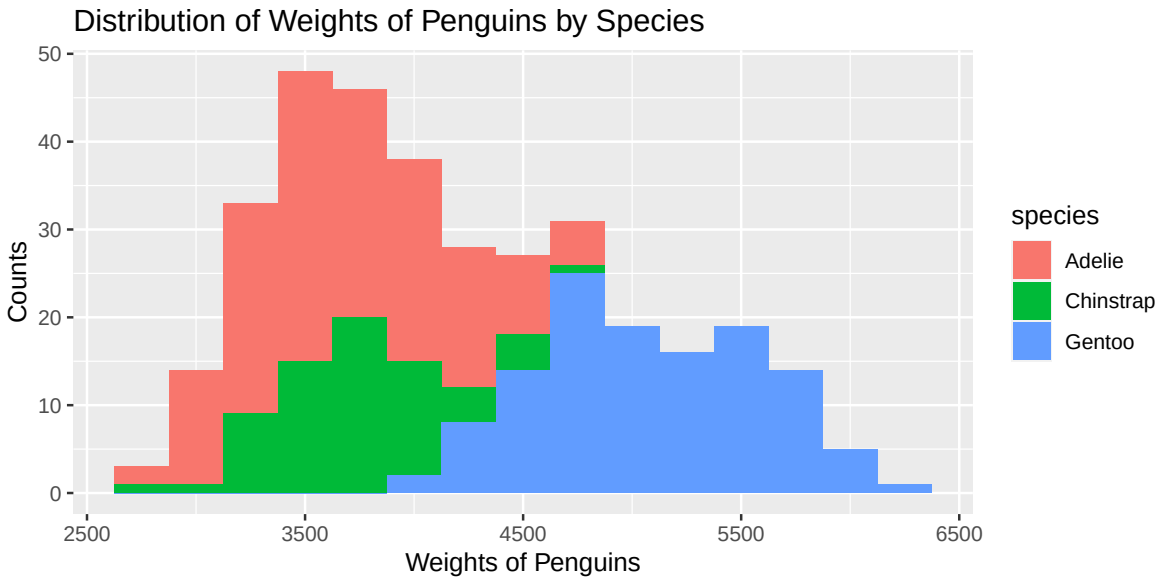
> **Note**
>
> Analyzing the relationship between two variables is called **bivariate** analysis.

Create visualizations of the distribution of `weights` of penguins by `species`.

4. Make a single histogram. Set an appropriate binwidth.

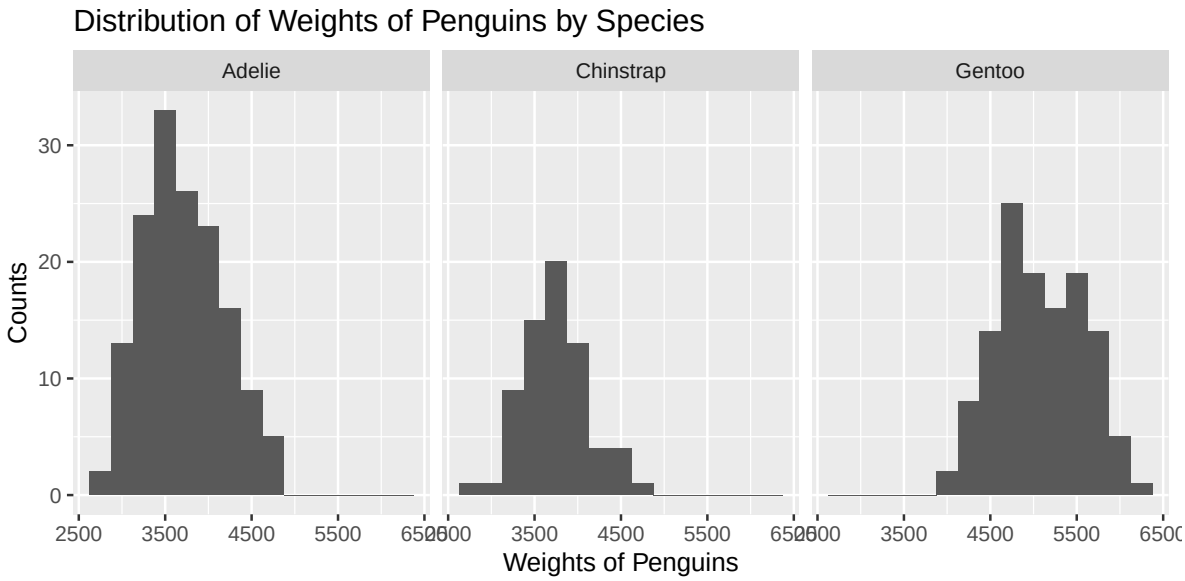
```
ggplot(penguins, aes(x = body_mass_g, fill = species)) +
  geom_histogram(binwidth = 250) +
  labs(
    x = "Weights of Penguins",
```

```
    y = "Counts",
    title = "Distribution of Weights of Penguins by Species")
```



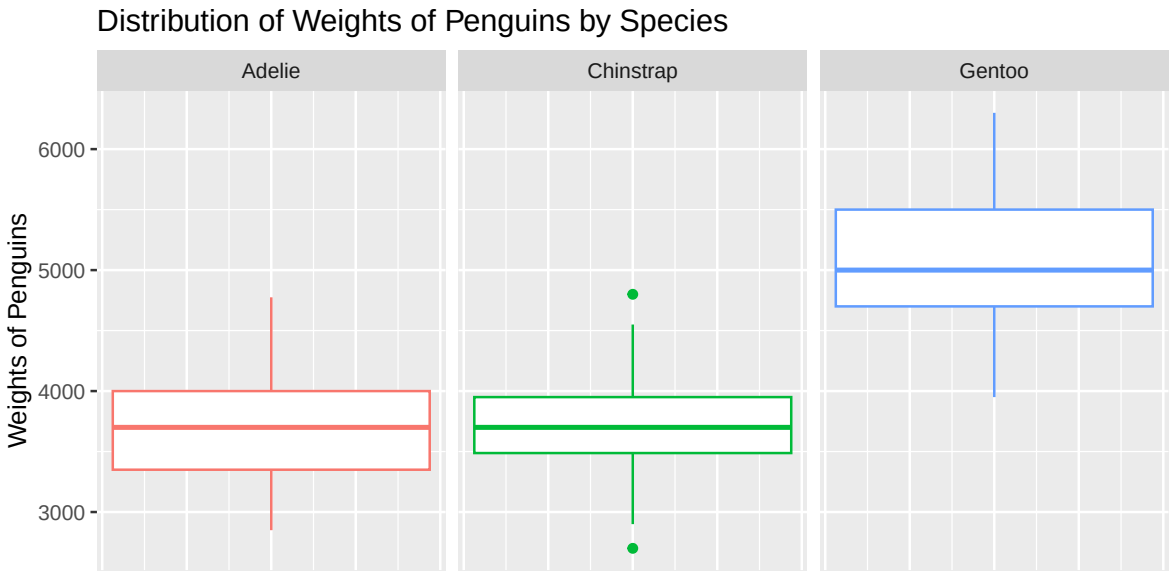
5. Use multiple histograms via faceting, one for each species. Set an appropriate binwidth, add color as you see fit, and turn off legends if not needed.

```
ggplot(penguins, aes(x = body_mass_g)) +
  geom_histogram(binwidth = 250) +
  facet_wrap(~species) +
  labs(
    x = "Weights of Penguins",
    y = "Counts",
    title = "Distribution of Weights of Penguins by Species")
```



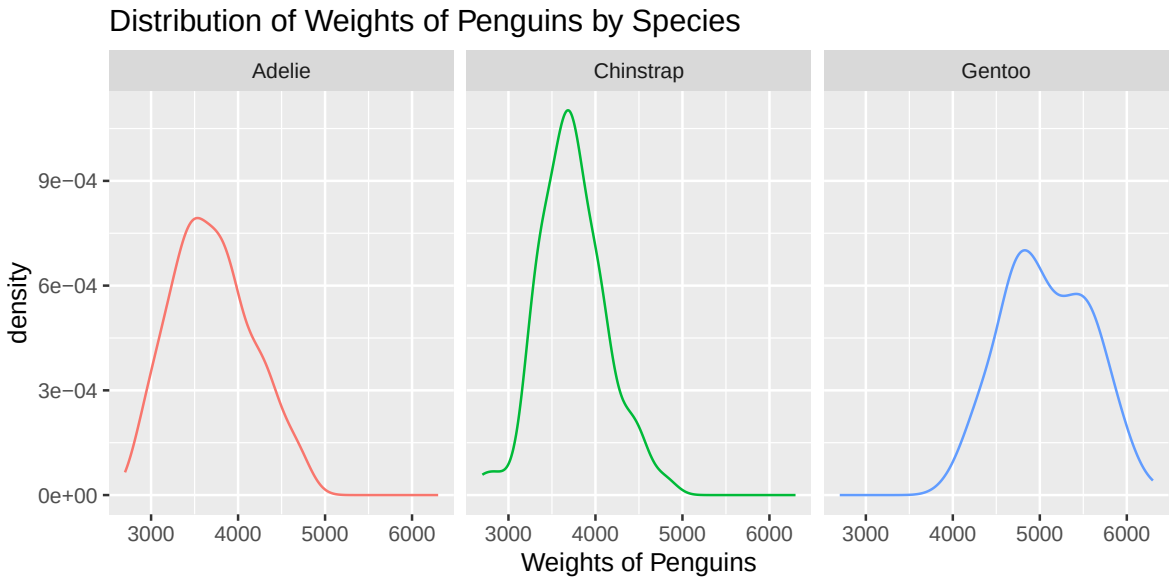
6. Use side-by-side box plots. Add color as you see fit and turn off legends if not needed.

```
ggplot(penguins, aes(y = body_mass_g, color = species)) +
  geom_boxplot() +
  facet_wrap(~species) +
  labs(
    y = "Weights of Penguins",
    title = "Distribution of Weights of Penguins by Species") +
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank(),
        legend.position = "none")
```



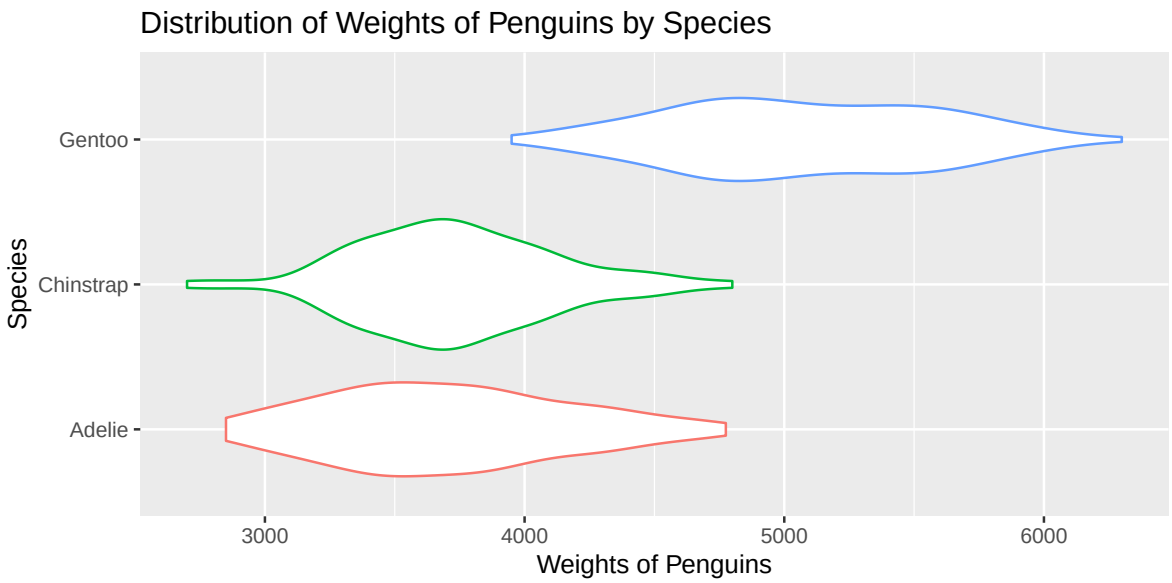
7. Use density plots. Add color as you see fit.

```
ggplot(penguins, aes(x = body_mass_g, color = species)) +
  geom_density() +
  facet_wrap(~species) +
  labs(
    x = "Weights of Penguins",
    title = "Distribution of Weights of Penguins by Species") +
  theme(legend.position = "none")
```



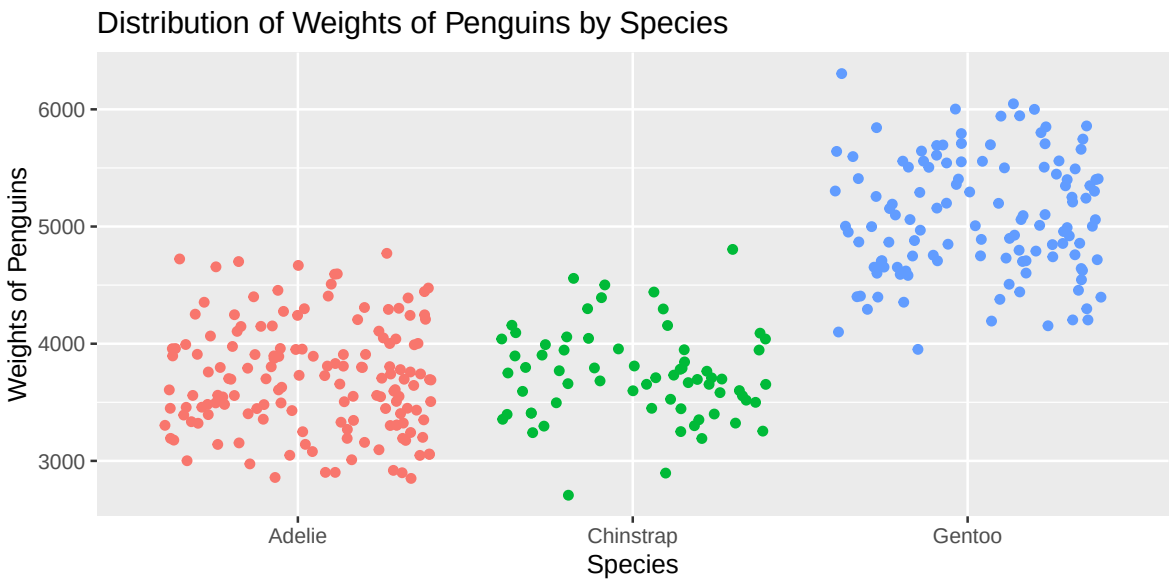
8. Use violin plots. Add color as you see fit and turn off legends if not needed.

```
ggplot(penguins, aes(x = body_mass_g, y = species, color = species)) +
  geom_violin() +
  labs(
    x = "Weights of Penguins", y = "Species",
    title = "Distribution of Weights of Penguins by Species") +
  theme(legend.position = "none")
```



9. Make a jittered scatter plot. Add color as you see fit and turn off legends if not needed.

```
ggplot(penguins, aes(x = species, y = body_mass_g, color = species)) +
  geom_jitter() +
  labs(
    x = "Species", y = "Weights of Penguins",
    title = "Distribution of Weights of Penguins by Species") +
  theme(legend.position = "none")
```

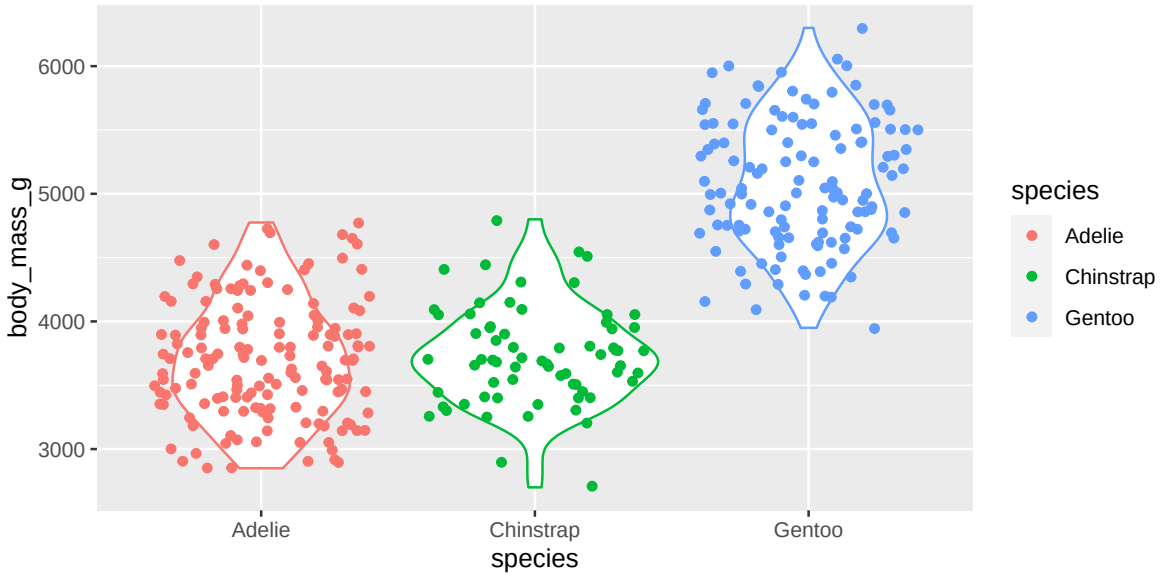


10. Use beeswarm plots. Add color as you see fit and turn off legends if not needed.

```
# ggplot(penguins, aes(x = species, y = body_mass_g, color = species)) +
#   geom_beeswarm() +
#   labs(
#     x = "Species", y = "Weights of Penguins",
#     title = "Distribution of Weights of Penguins by Species") +
#   theme(legend.position = "none")
```

11. Use multiple geoms on a single plot. Be deliberate about the order of plotting. Change the theme and the color scale of the plot. Finally, add informative labels.

```
ggplot(penguins, aes(x = species, y = body_mass_g, color = species)) +
  geom_violin(show.legend = FALSE) + geom_jitter()
```




```
  labs(
    x = "Species", y = "Weights of Penguins",
    title = "Distribution of Weights of Penguins by Species")
```

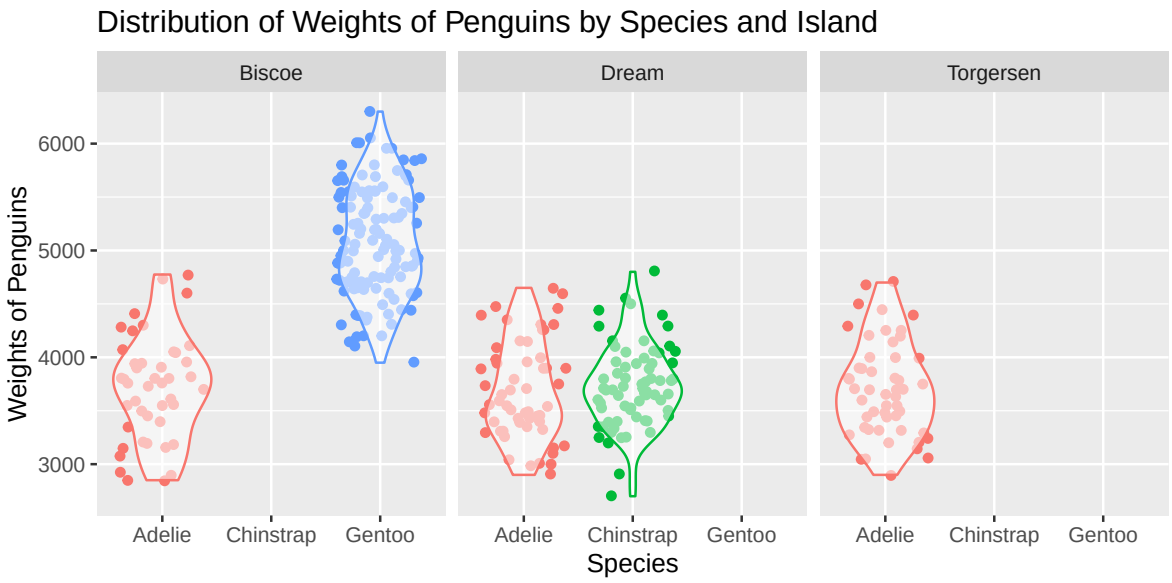
```
$x
[1] "Species"

$y
[1] "Weights of Penguins"

$title
[1] "Distribution of Weights of Penguins by Species"

attr(,"class")
[1] "labels"
```

Multiple variables

> **Note**
>
> Analyzing the relationship between three or more variables is called **multivariate** analysis.

12. Facet the plot you created in the previous exercise by `island`. Adjust labels accordingly.

```
ggplot(penguins, aes(x = species, y = body_mass_g, color = species)) +
  geom_jitter() + geom_violin(aes(alpha = 0.5)) +
  labs(
    x = "Species", y = "Weights of Penguins",
    title = "Distribution of Weights of Penguins by Species and Island") +
  theme(legend.position = "none") +
  facet_wrap(~island)
```

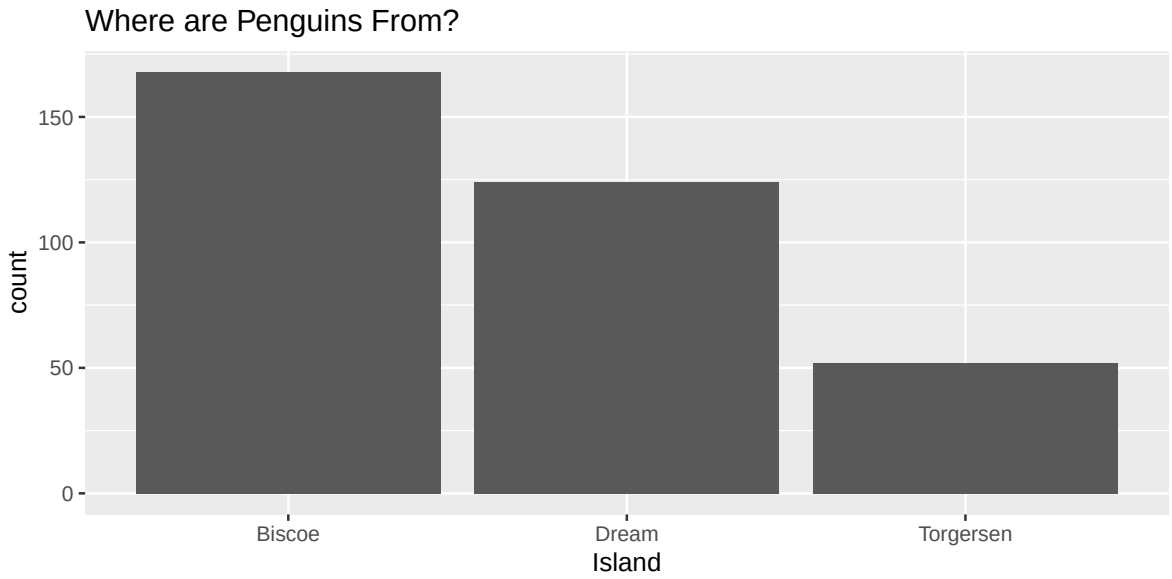


Before you continue, let's turn off all warnings the code chunks generate and resize all figures. We'll do this by editing the YAML.

Visualizing other variables - Your turn!

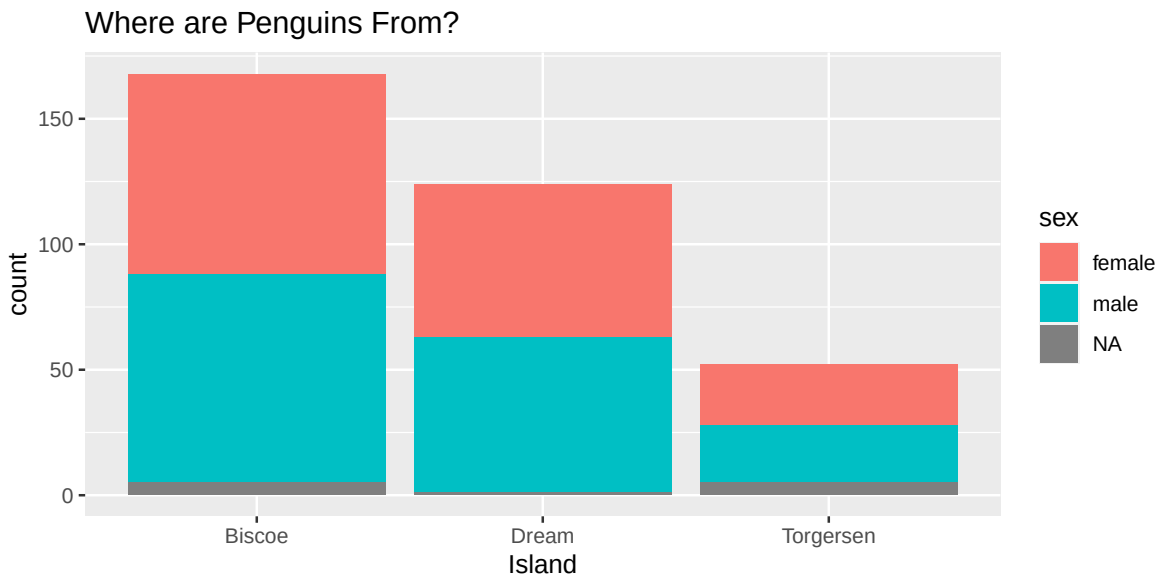
13. Pick a single categorical variable from the data set and make a bar plot of its distribution.

```
ggplot(penguins, aes(x = island)) +
  geom_bar(show.legend = FALSE) +
  labs(
    x = "Island",
    title = "Where are Penguins From?"
  )
```



14. Pick two categorical variables and make a visualization to visualize the relationship between the two variables. Along with your code and output, provide an interpretation of the visualization.

```
ggplot(penguins, aes(x = island, fill = sex)) +
  geom_bar() +
  labs(
    x = "Island",
    title = "Where are Penguins From?"
  )
```



Biscoe has the highest penguin population, followed by Dream then Torgersen. On each island, roughly 50% of the penguins are male and 50% of the penguins are female (a few penguins on each island are not labeled with a sex).

15. Make another plot that uses at least three variables. At least one should be numeric and at least one categorical. In 1-2 sentences, describe what the plot shows about the relationships between the variables you plotted. Don't forget to label your code chunk.

```
# add code here

ggplot(penguins, aes(x = body_mass_g, y = flipper_length_mm, color = sex)) +
  geom_smooth(show.legend = FALSE) + geom_point(show.legend = FALSE) +
  labs(
    x = "Weight (g)", y = "Flipper Length (mm)",
    title = "Flipper Length vs Weight, By Sex"
  ) + facet_wrap(~sex)
```

Male penguins tend to be heavier and have longer fins than female penguins.