Data Visualization

Gabriel Montes

2025-07-15

Exercise 1.2.5

1.How many rows are in penguins? How many columns?

```
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   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- 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
```

```
glimpse(penguins)
```

```
## Rows: 344
## Columns: 8
## $ species     <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Ad~
## $ island      <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Tor~
## $ bill_len    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, 42.0, ~
## $ bill_dep    <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, 20.2, ~
## $ flipper_len <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186, 180,~
## $ body_mass   <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, 4250, ~
## $ sex         <fct> male, female, female, NA, female, male, female, male, NA, ~
## $ year        <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

344 rows and 8 columns

2. What does the bill_depth_mm variable in the penguins data frame describe? Read the help for ?penguins to find out.

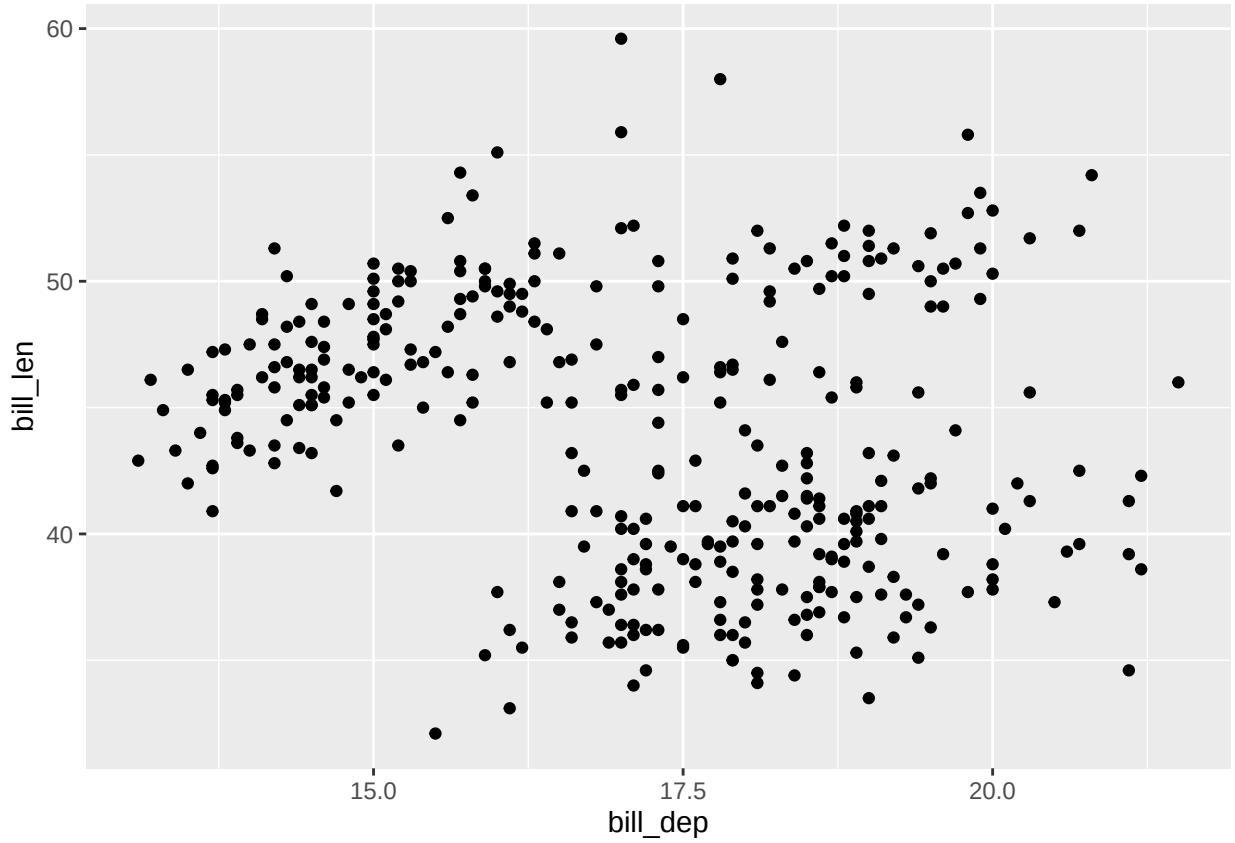
```
?penguins
```

numeric, bill depth (millimeters)

3. Make a scatterplot of bill_depth_mm vs. bill_length_mm. That is, make a scatterplot with bill_depth_mm on the y-axis and bill_length_mm on the x-axis. Describe the relationship between these two variables.

```
library(ggplot2)
ggplot(
  data = penguins,
  mapping = aes(x = bill_dep, y = bill_len)
) +
  geom_point()
```

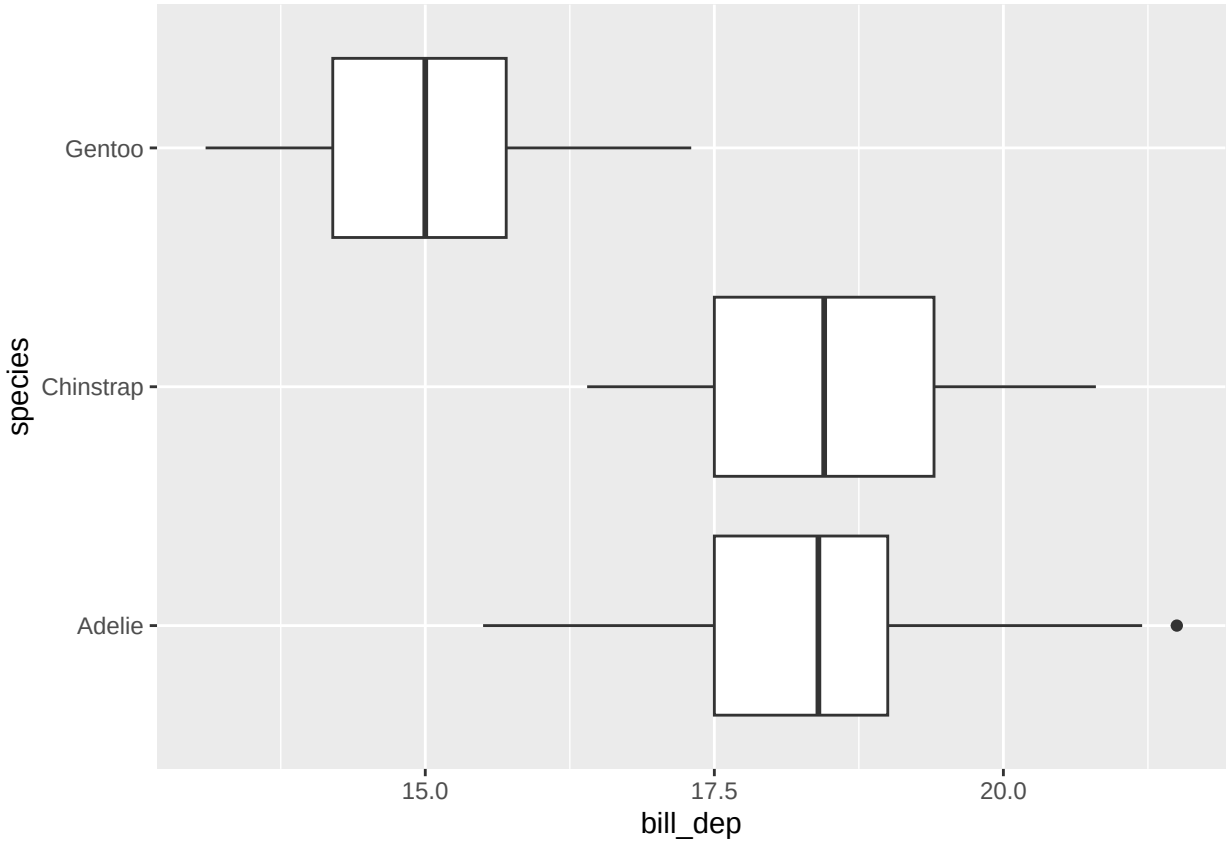
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



4. What happens if you make a scatterplot of species vs. bill_depth_mm? What might be a better choice of geom?

```
library(ggplot2)
ggplot(
  data = penguins,
  mapping = aes(x = bill_dep, y = species)
) +
  geom_boxplot()
```

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



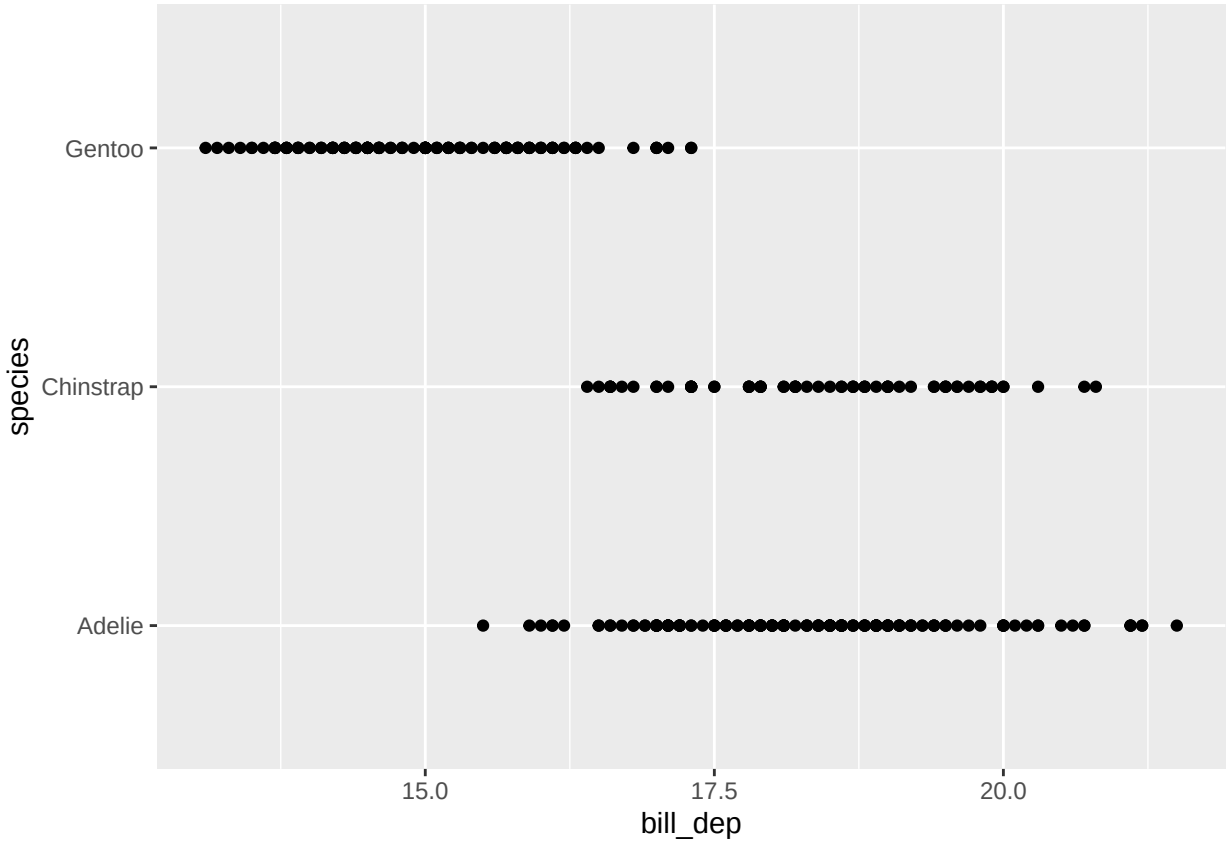
three lines appear if i scatterplot them, better would be geom_boxplot()

5. Why does the following give an error and how would you fix it?

ggplot(data = penguins) + geom_point()

```
library(ggplot2)
ggplot(
  data = penguins,
  mapping = aes(x = bill_dep, y = species)
) +
  geom_point()
```

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

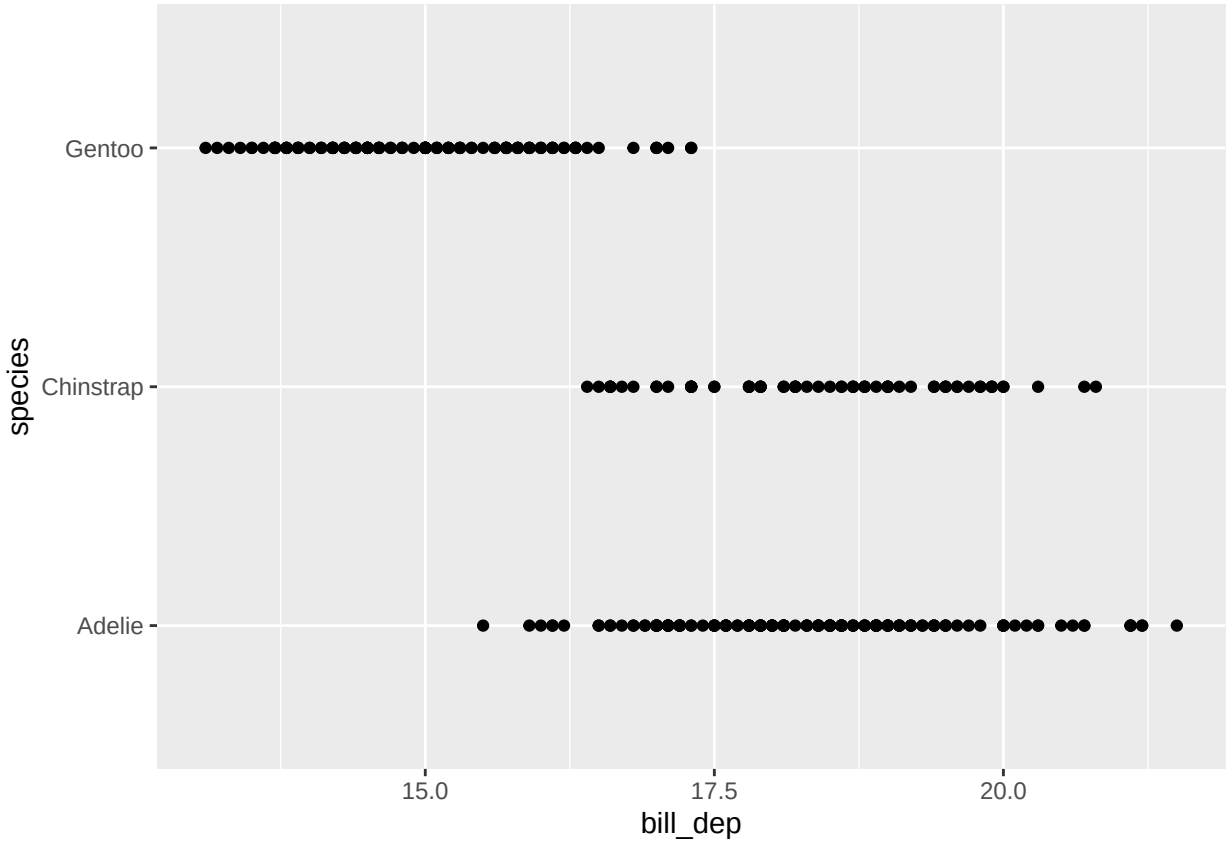


provide the x and y values to fix the code

6. What does the na.rm argument do in geom_point()? What is the default value of the argument? Create a scatterplot where you successfully use this argument set to TRUE.

The na.rm argument removes missing values if set to true

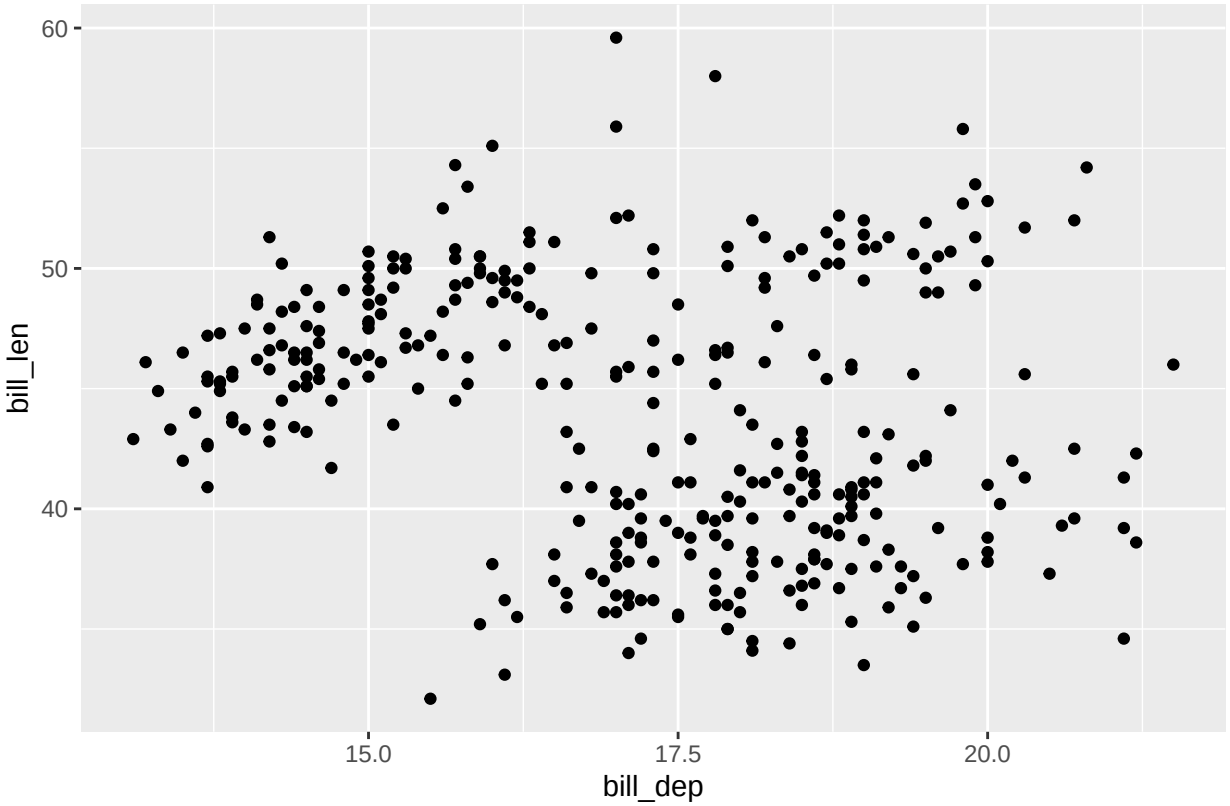
```
library(ggplot2)
ggplot(
  data = penguins,
  mapping = aes(x = bill_dep, y = species)
) +
  geom_point(na.rm = TRUE)
```



7. Add the following caption to the plot you made in the previous exercise: "Data come from the palmerpenguins package." Hint: Take a look at the documentation for labs().

```
library(ggplot2)
ggplot(
  data = penguins,
  mapping = aes(x = bill_dep, y = bill_len)
) +
  geom_point() +
  labs(caption = "Data come from the palmerpenguins package.")
```

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

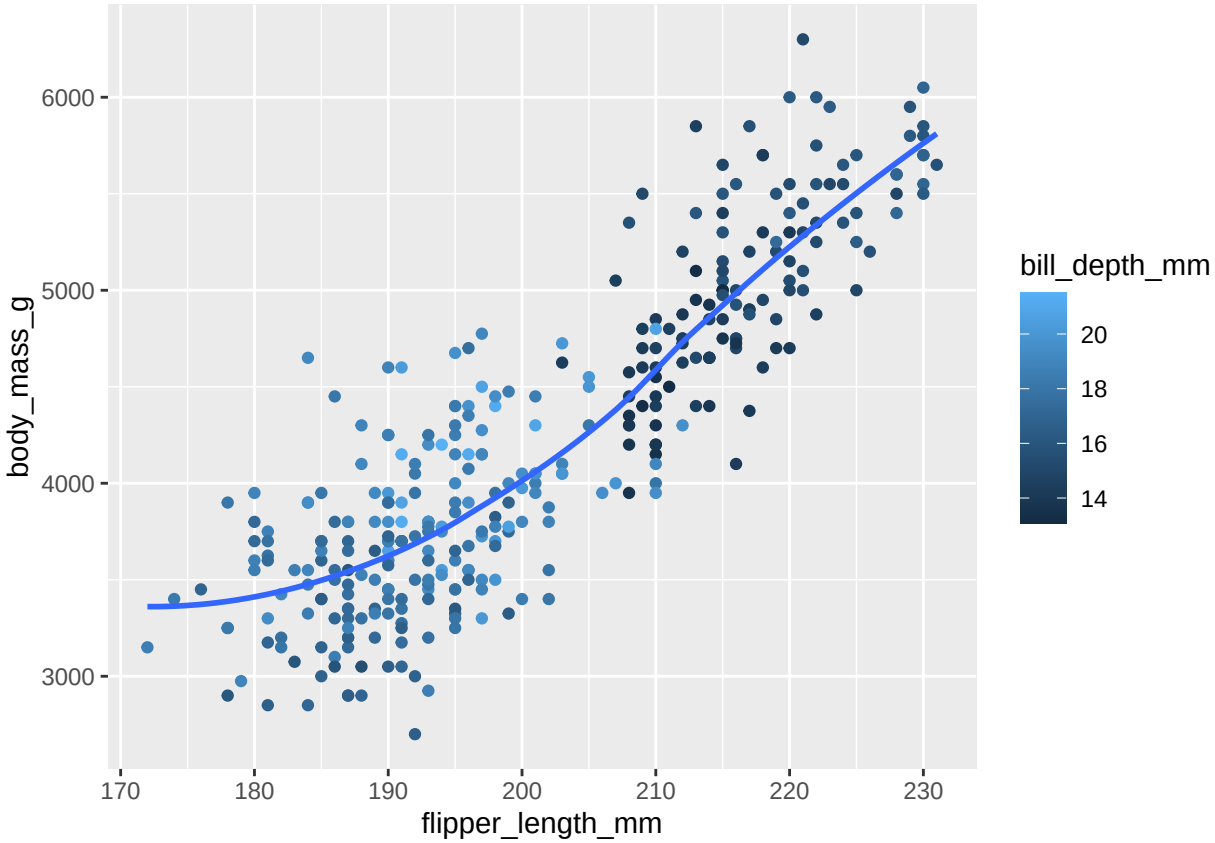


Data come from the palmerpenguins package.

8. Recreate the following visualization. What aesthetic should bill_depth_mm be mapped to? And should it be mapped at the global level or at the geom level?

```
ggplot(data = penguins, aes(x = flipper_len, y = body_mass)) +
  geom_point(aes(color = bill_dep), na.rm = TRUE) +
  geom_smooth(se = FALSE, na.rm = TRUE) +
  labs(
    x = "flipper_length_mm",
    y = "body_mass_g",
    color = "bill_depth_mm"
  )
```

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



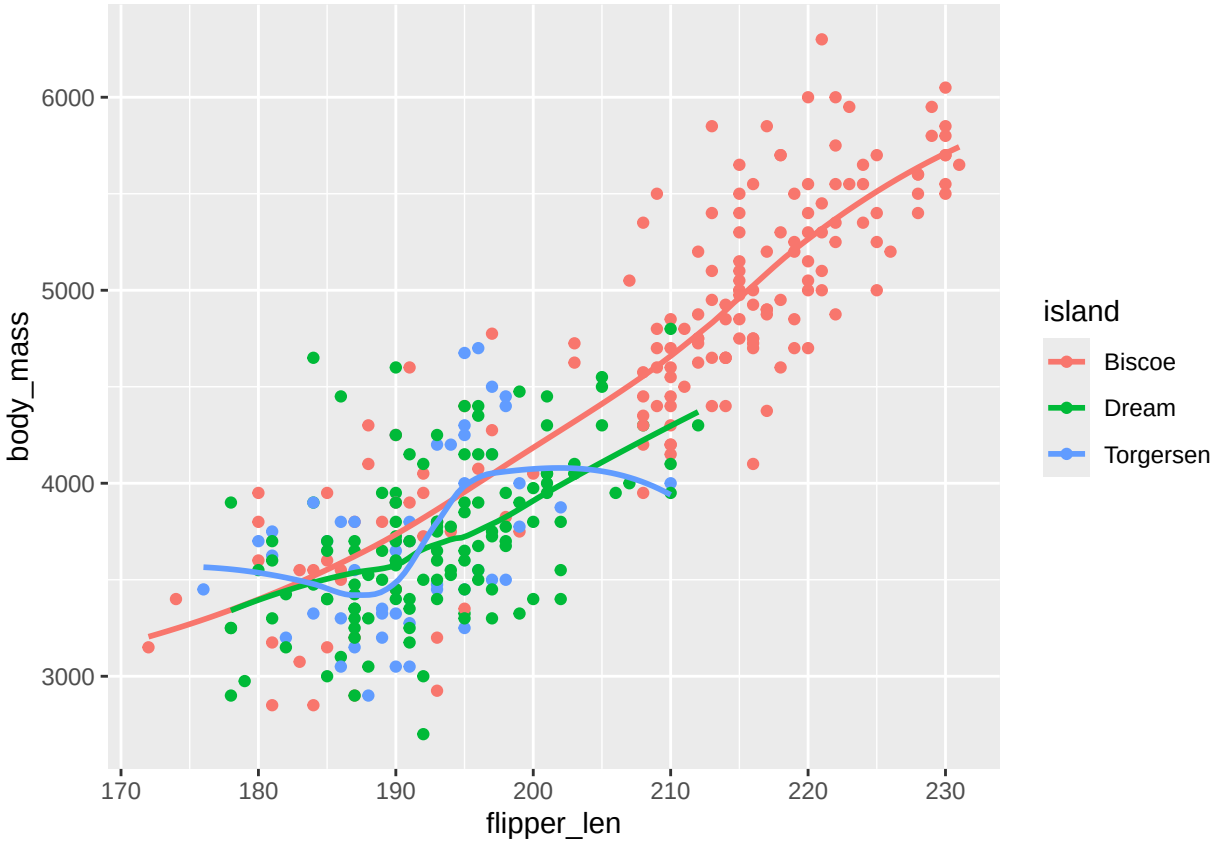
9. Run this code in your head and predict what the output will look like. Then, run the code in R and check your predictions.

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_len, y = body_mass, color = island)
) +
  geom_point() +
  geom_smooth(se = FALSE)
```

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

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

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



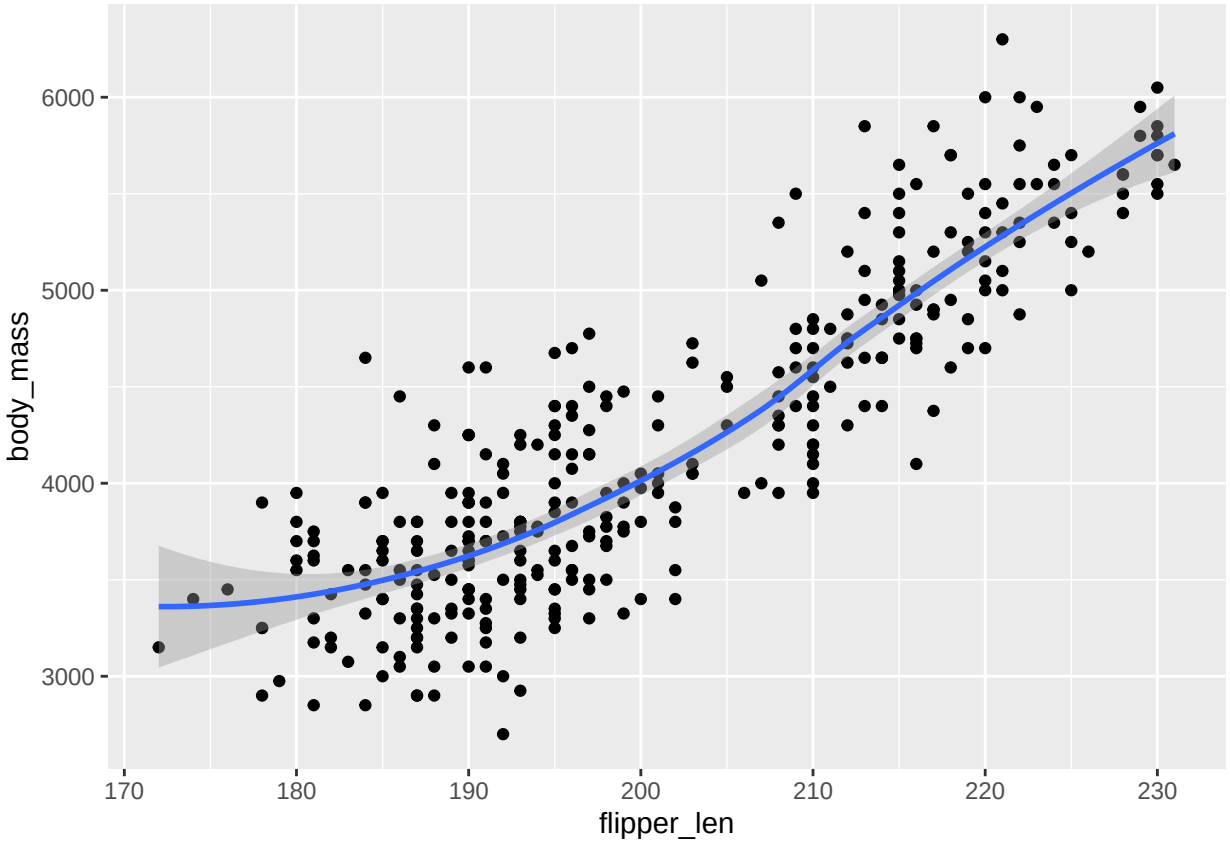
10. Will these two graphs look different? Why/why not?

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_len, y = body_mass)
) +
  geom_point() +
  geom_smooth()
```

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

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

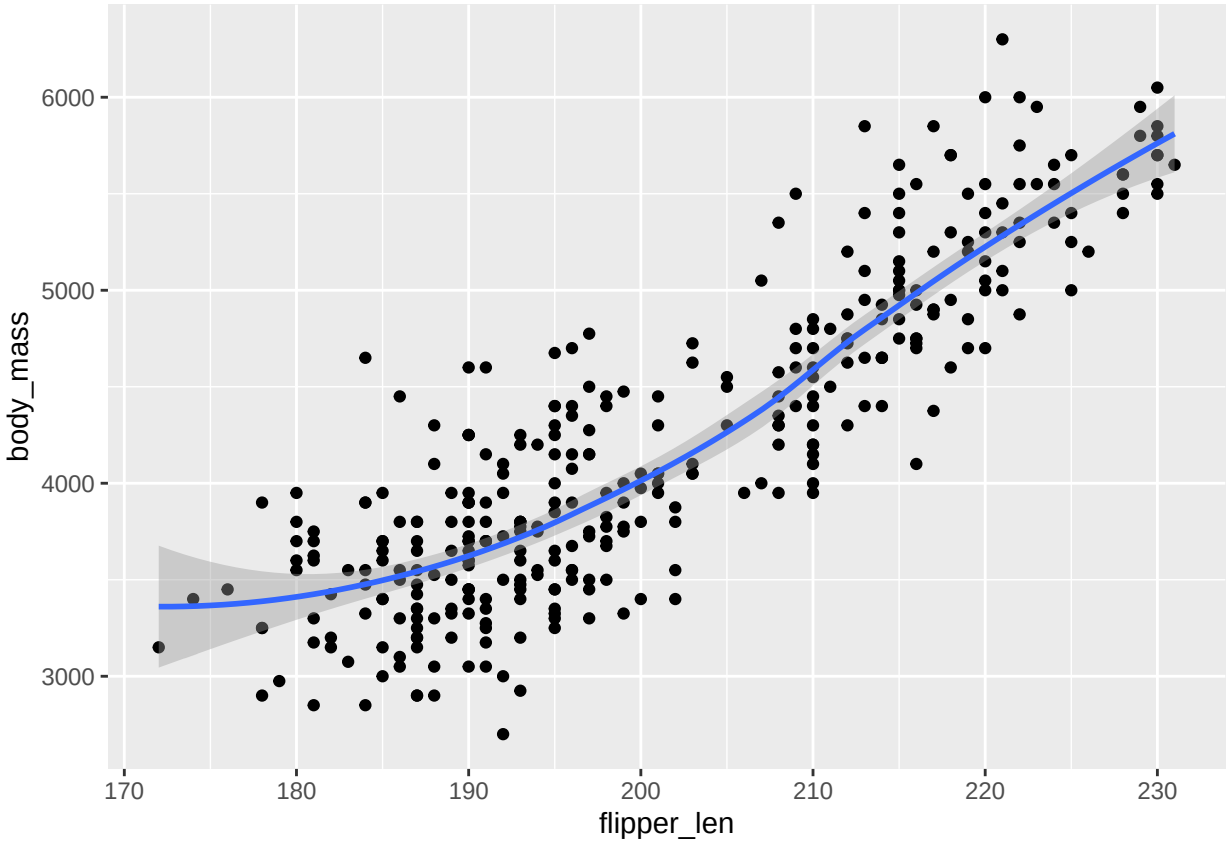
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
ggplot() +
  geom_point(
    data = penguins,
    mapping = aes(x = flipper_len, y = body_mass)
  ) +
  geom_smooth(
    data = penguins,
    mapping = aes(x = flipper_len, y = body_mass)
  )
```

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

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

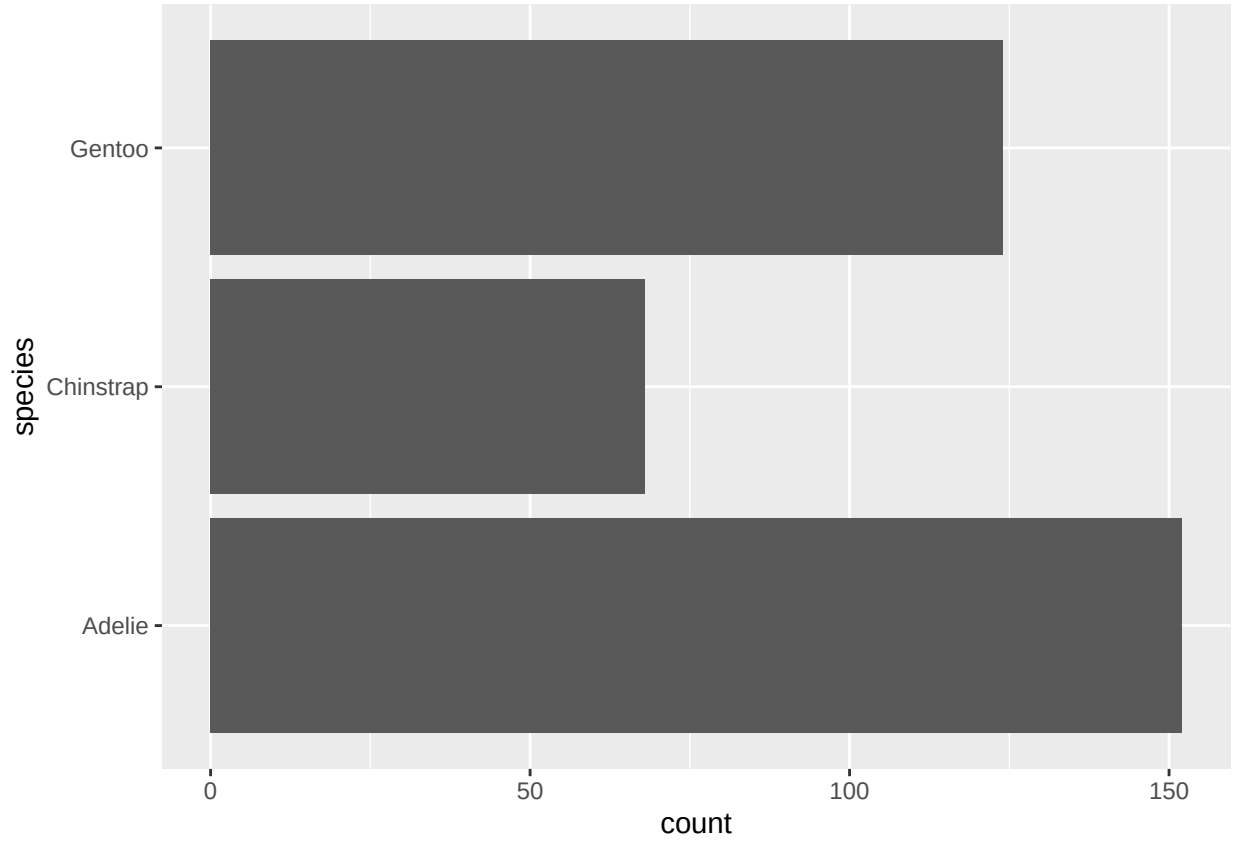


they look the same because they use the same data, in graph 2 geom_smooth/geom_pint is applied separately

Exercise 1.4.3

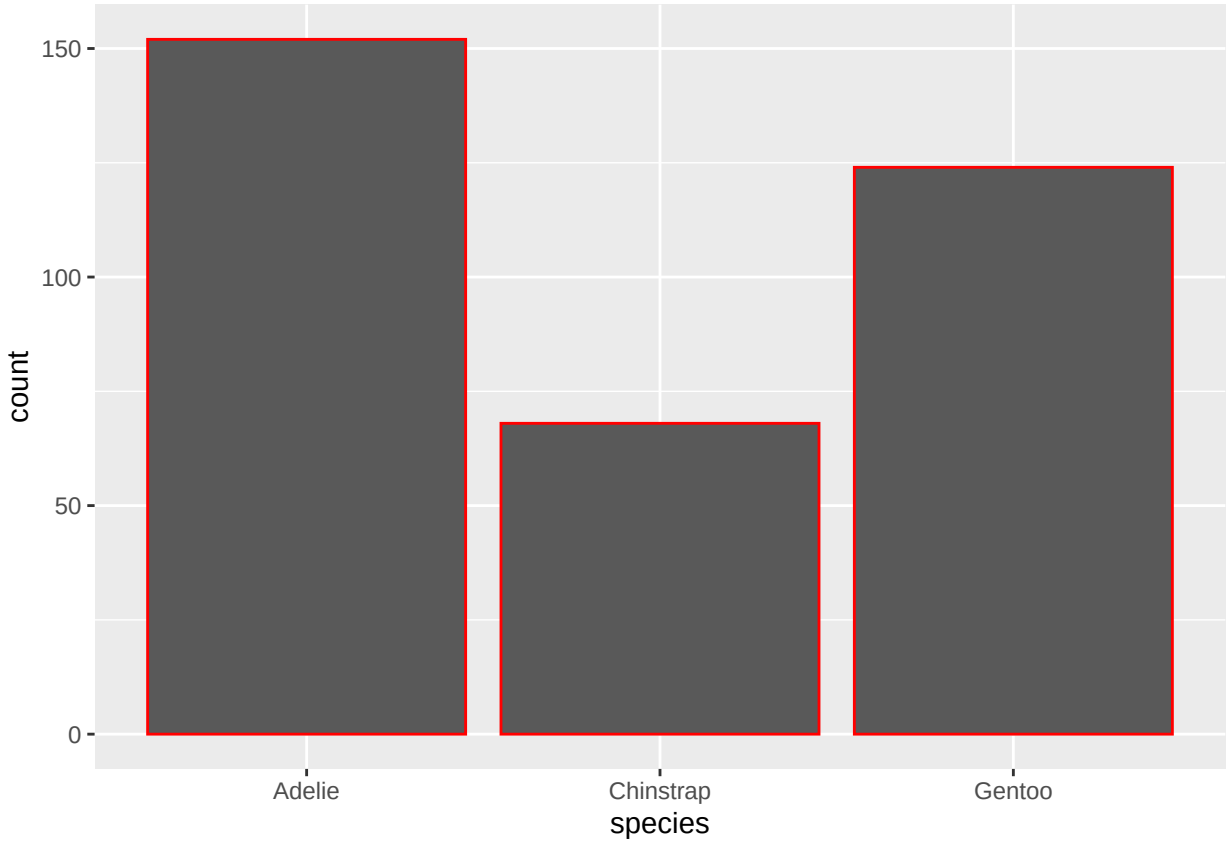
1. Make a bar plot of species of penguins, where you assign species to the y aesthetic. How is this plot different?

```
library(ggplot2)
ggplot(data = penguins, aes(y = species)) +
  geom_bar()
```

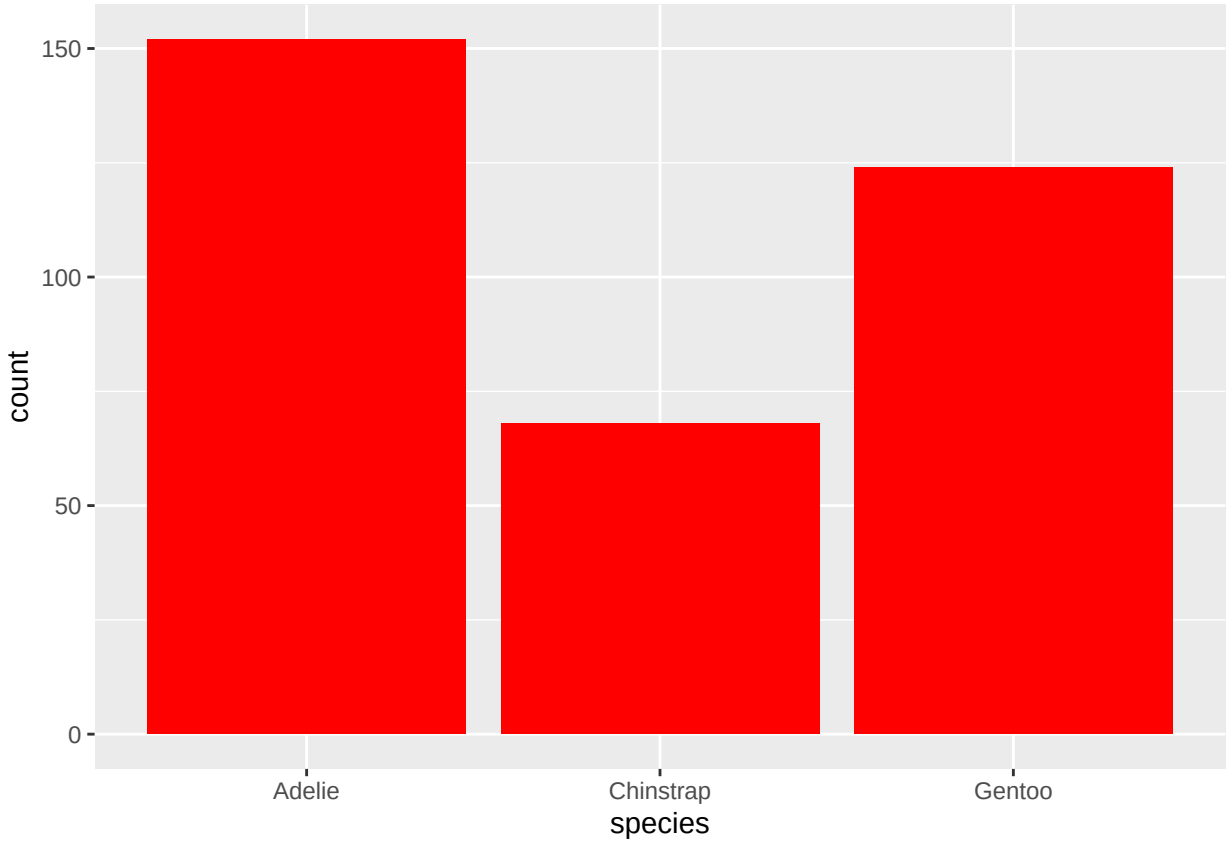


2. How are the following two plots different? Which aesthetic, color or fill, is more useful for changing the color of bars?

```
ggplot(penguins, aes(x = species)) +
  geom_bar(color = "red")
```



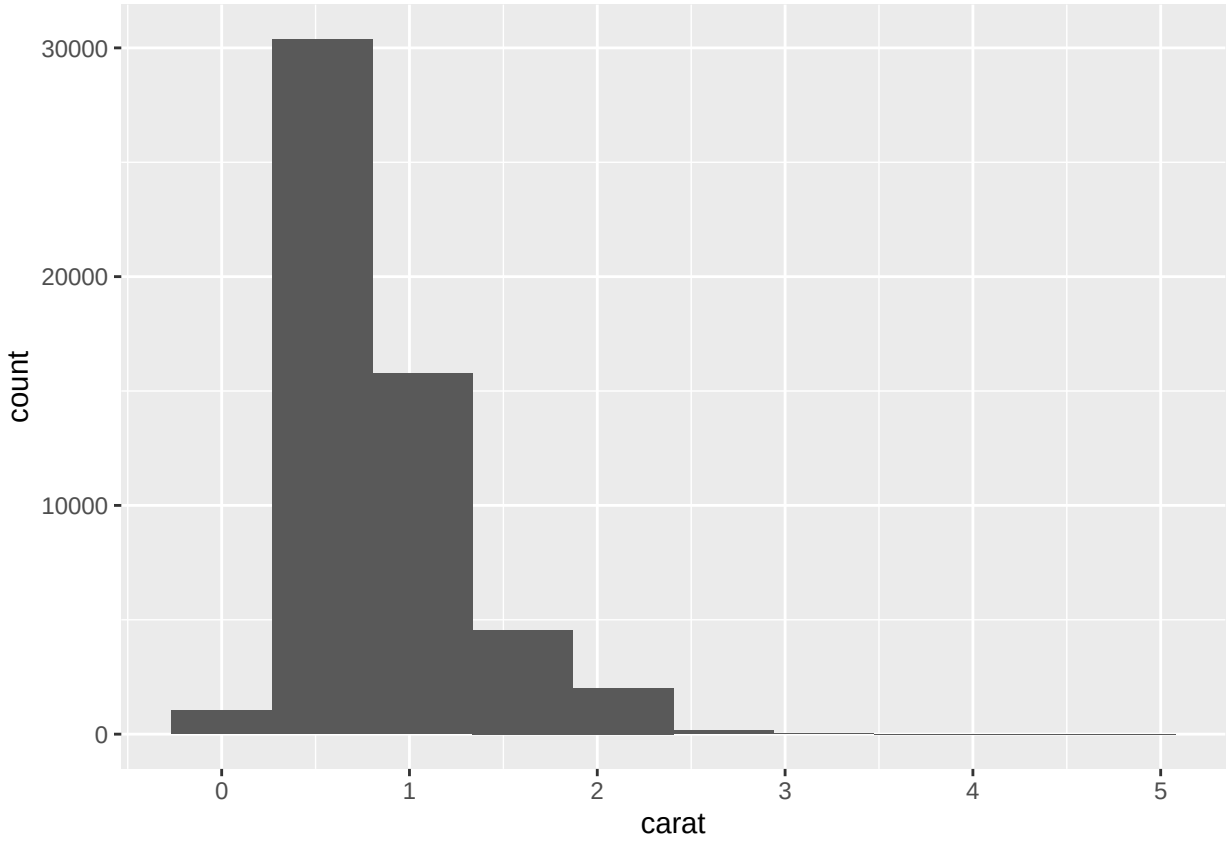

```
ggplot(penguins, aes(x = species)) +
  geom_bar(fill = "red")
```



color= changes the border and fil= fills the whole square

3. What does the bins argument in geom_histogram() do?

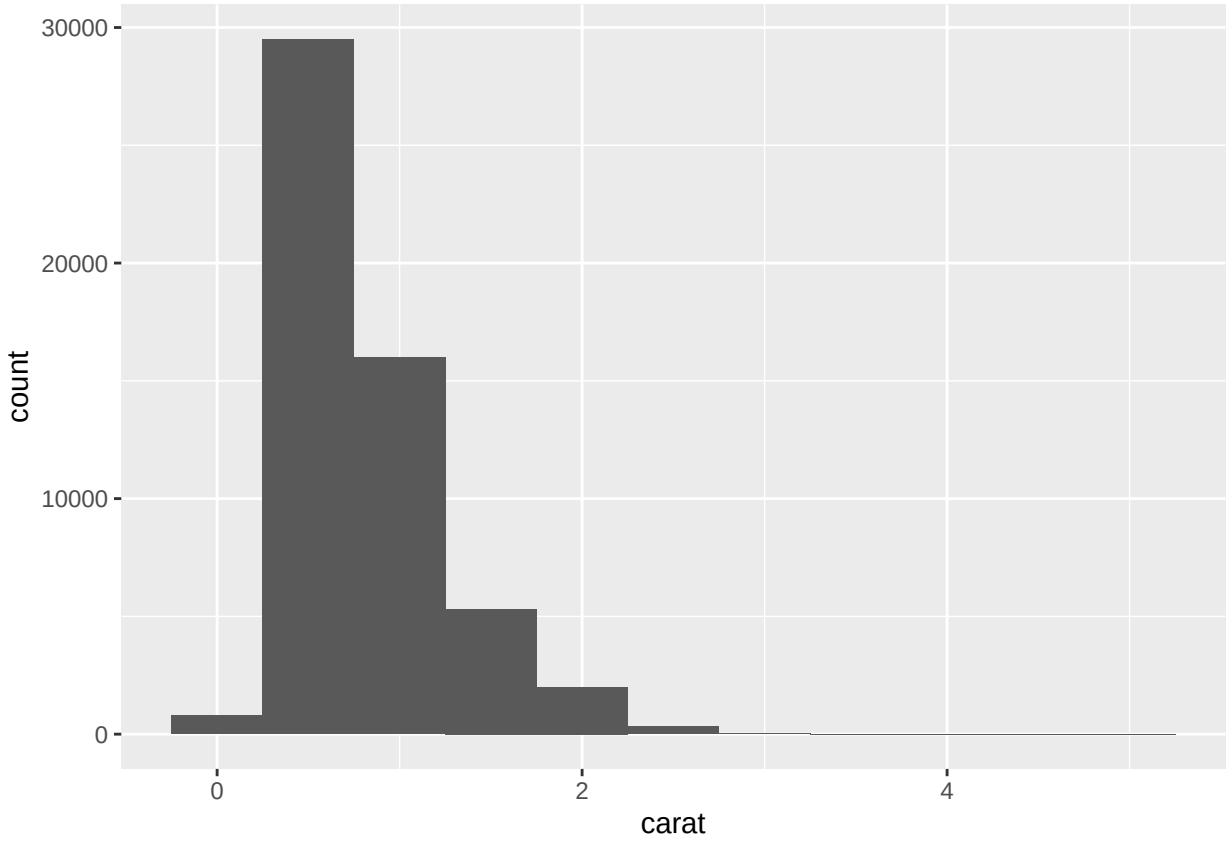
```
ggplot(data = diamonds, aes(x = carat)) +
  geom_histogram(bins = 10)
```



bins tell how many to divide the x axis

4. Make a histogram of the carat variable in the diamonds dataset that is available when you load the tidyverse package. Experiment with different binwidths. What binwidth reveals the most interesting patterns?

```
ggplot(diamonds, aes(x = carat)) +
  geom_histogram(binwidth = 0.5)
```



Exercise 1.5.5

1. The mpg data frame that is bundled with the ggplot2 package contains 234 observations collected by the US Environmental Protection Agency on 38 car models. Which variables in mpg are categorical? Which variables are numerical? (Hint: Type ?mpg to read the documentation for the dataset.) How can you see this information when you run mpg?

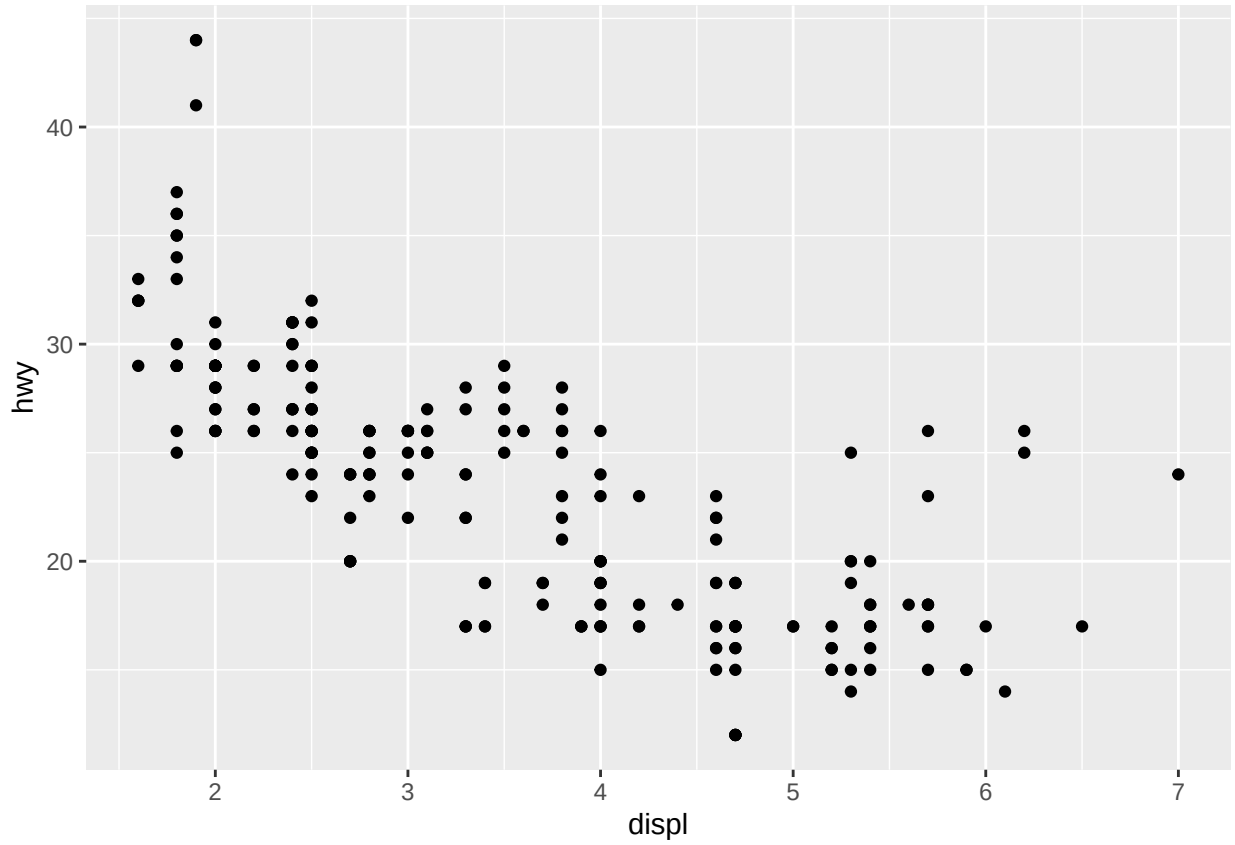
```
library(tidyverse)
str(mpg)
```

```
## tibble [234 x 11] (S3: tbl_df/tbl/data.frame)
##  $ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" ...
##  $ model       : chr [1:234] "a4" "a4" "a4" "a4" ...
##  $ displ       : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int [1:234] 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int [1:234] 4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr [1:234] "f" "f" "f" "f" ...
##  $ cty         : int [1:234] 18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int [1:234] 29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr [1:234] "p" "p" "p" "p" ...
##  $ class       : chr [1:234] "compact" "compact" "compact" "compact" ...
```

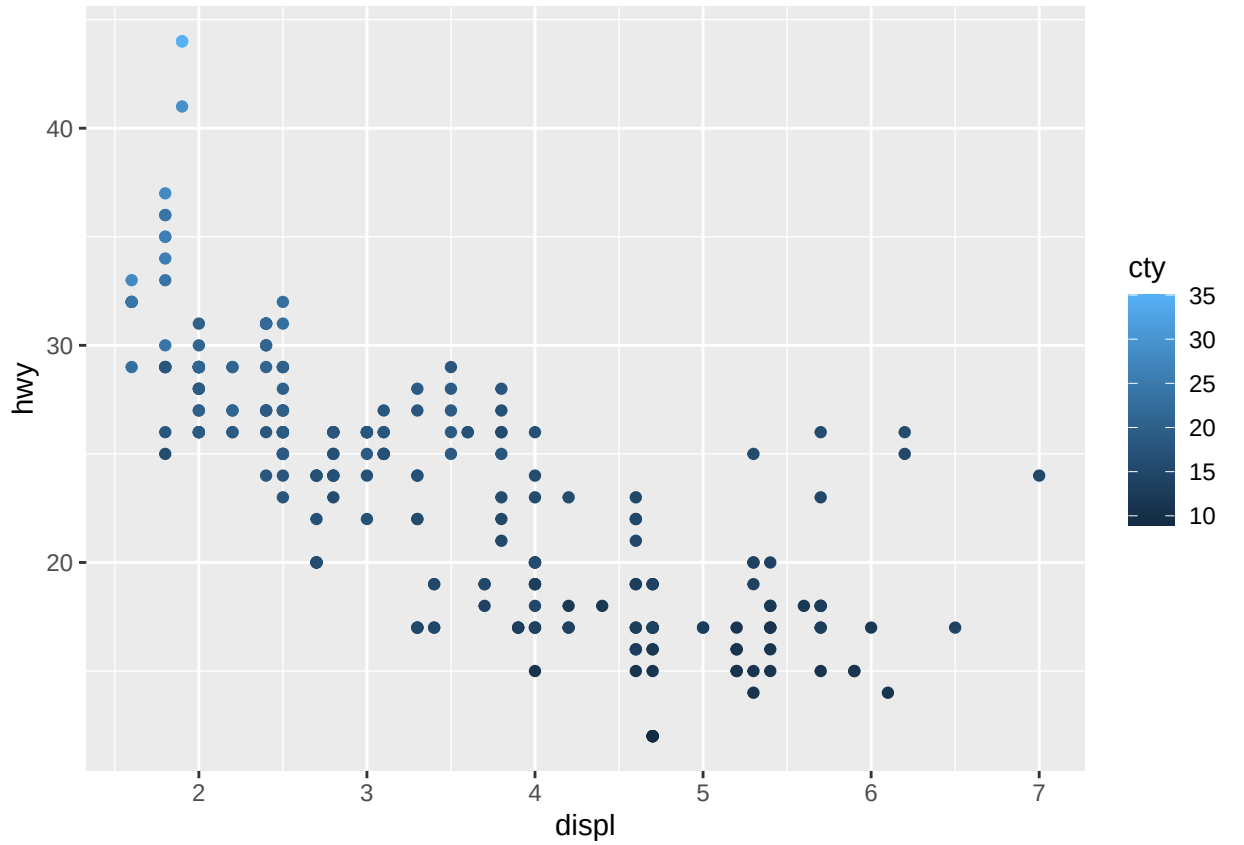
using str() we can see which one is numerical or categorical

2. Make a scatterplot of hwy vs. displ using the mpg data frame. Next, map a third, numerical variable to color, then size, then both color and size, then shape. How do these aesthetics behave differently for categorical vs. numerical variables?

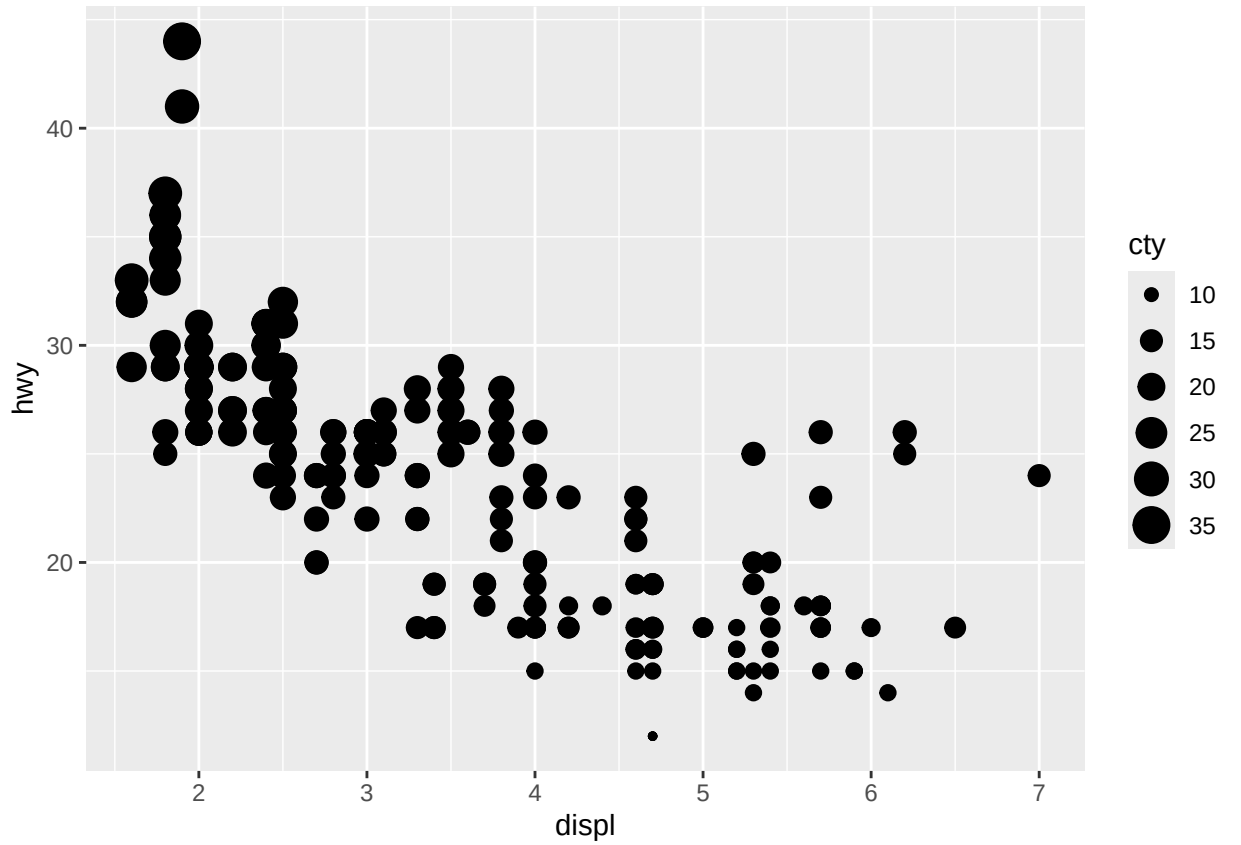
```
library(ggplot2)

ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point()
```



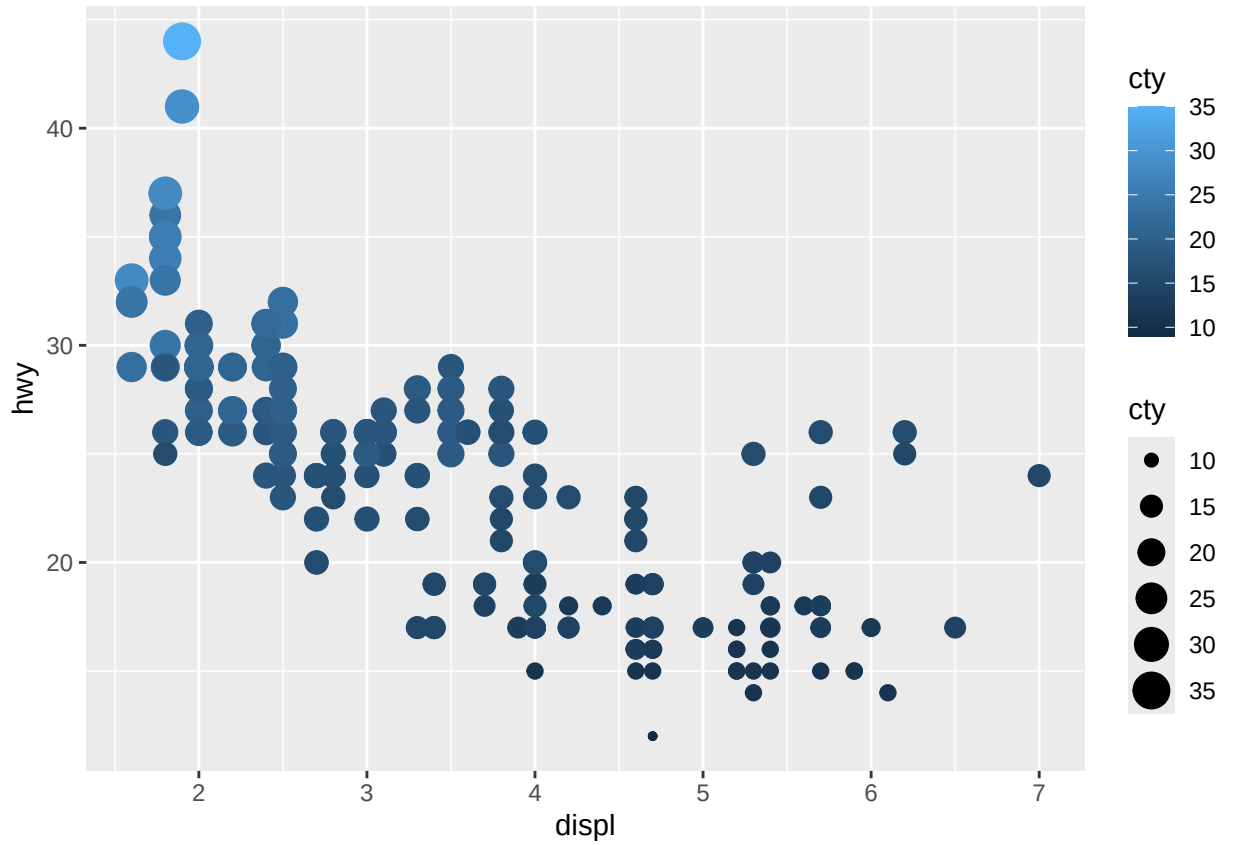

```
ggplot(mpg, aes(x = displ, y = hwy, color = cty)) +
  geom_point()
```




```
ggplot(mpg, aes(x = displ, y = hwy, size = cty)) +
  geom_point()
```



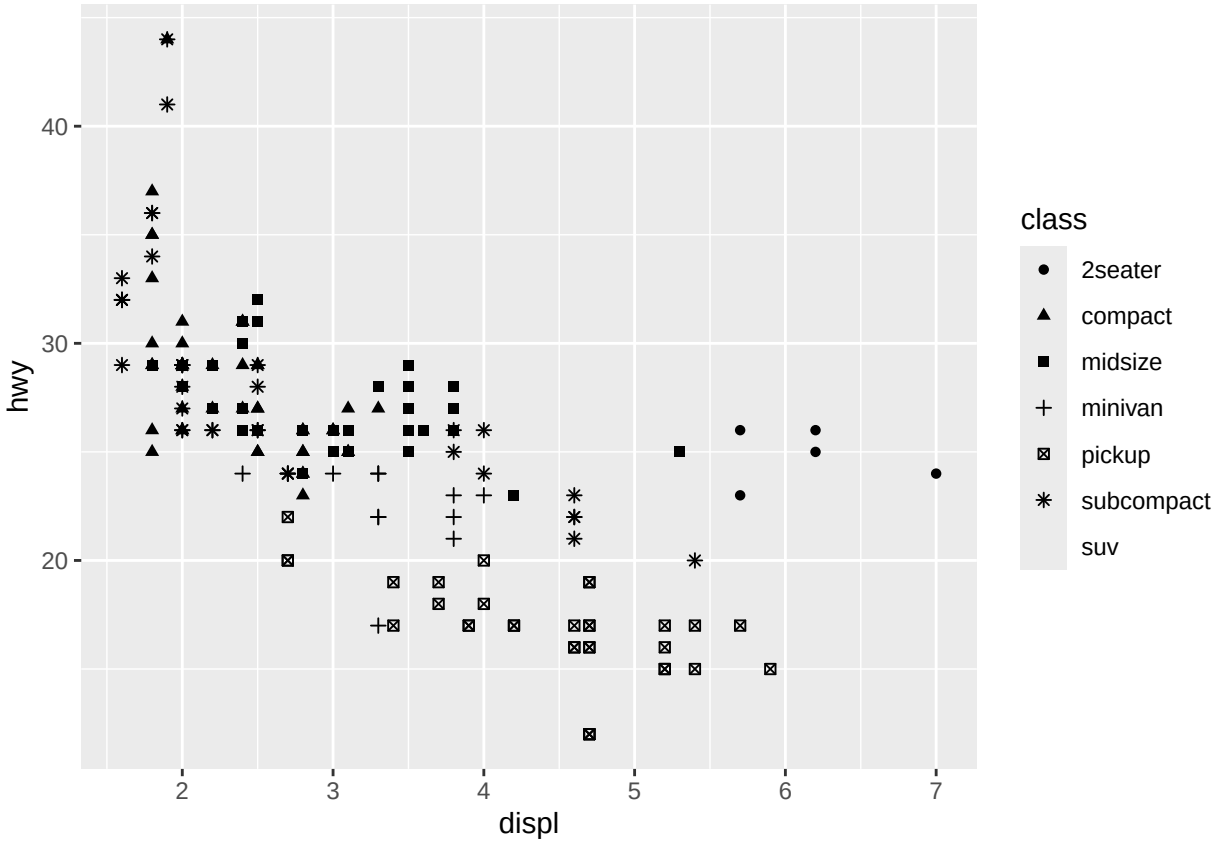

```
ggplot(mpg, aes(x = displ, y = hwy, color = cty, size = cty)) +
  geom_point()
```




```
ggplot(mpg, aes(x = displ, y = hwy, shape = class)) +
  geom_point()
```

```
## Warning: The shape palette can deal with a maximum of 6 discrete values because more
## than 6 becomes difficult to discriminate
## i you have requested 7 values. Consider specifying shapes manually if you need
##   that many of them.
```

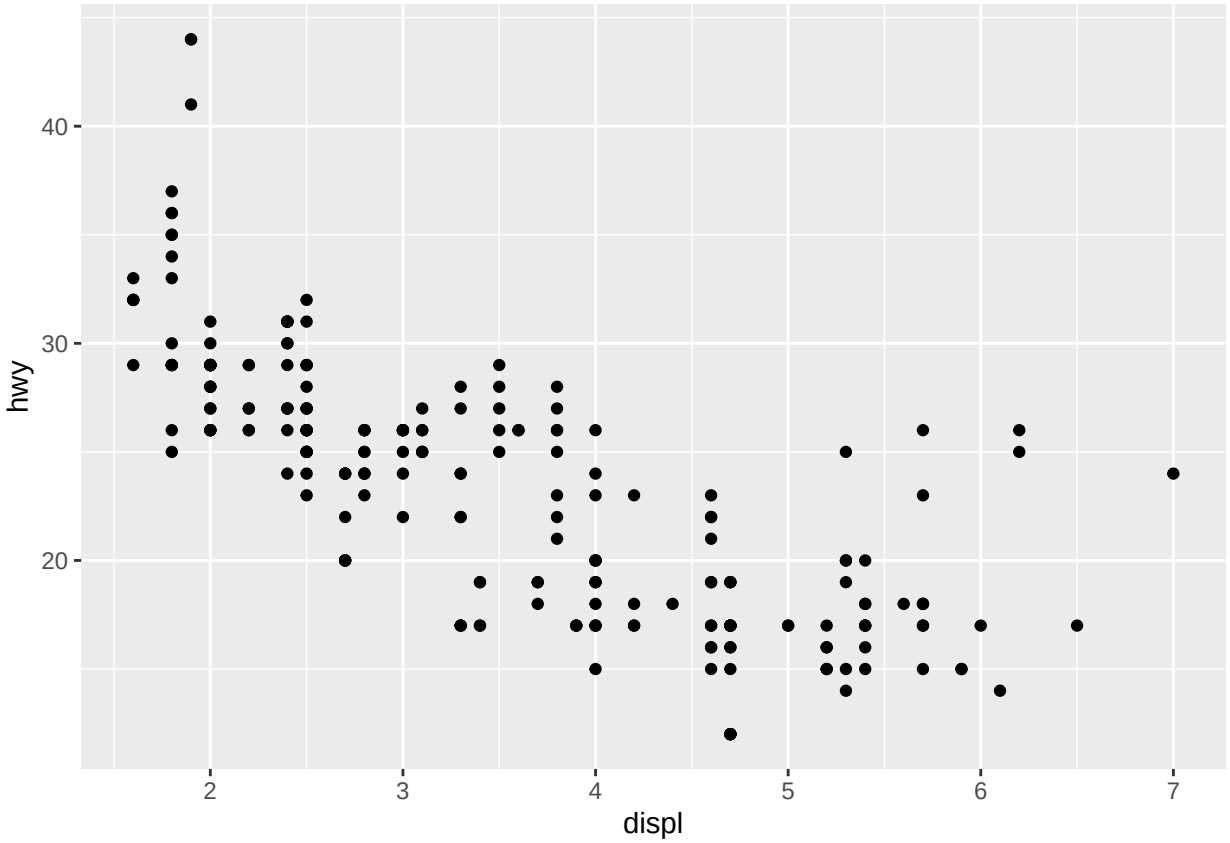
```
## Warning: Removed 62 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



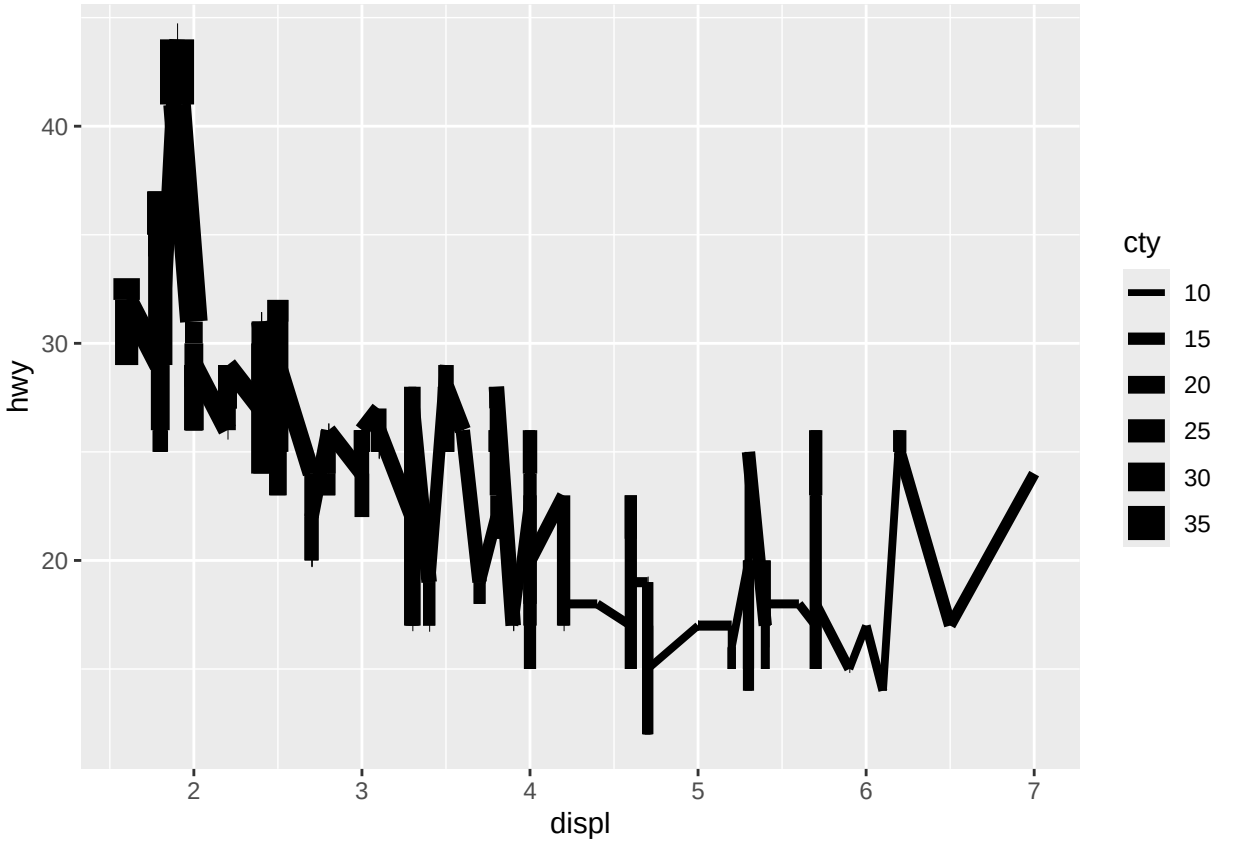
they differ by the points, their shapes, and colors

3. In the scatterplot of hwy vs. displ, what happens if you map a third variable to linewidth?

```
ggplot(mpg, aes(x = displ, y = hwy, linewidth = cty)) +
  geom_point()
```




```
ggplot(mpg, aes(x = displ, y = hwy, linewidth = cty)) +
  geom_line()
```



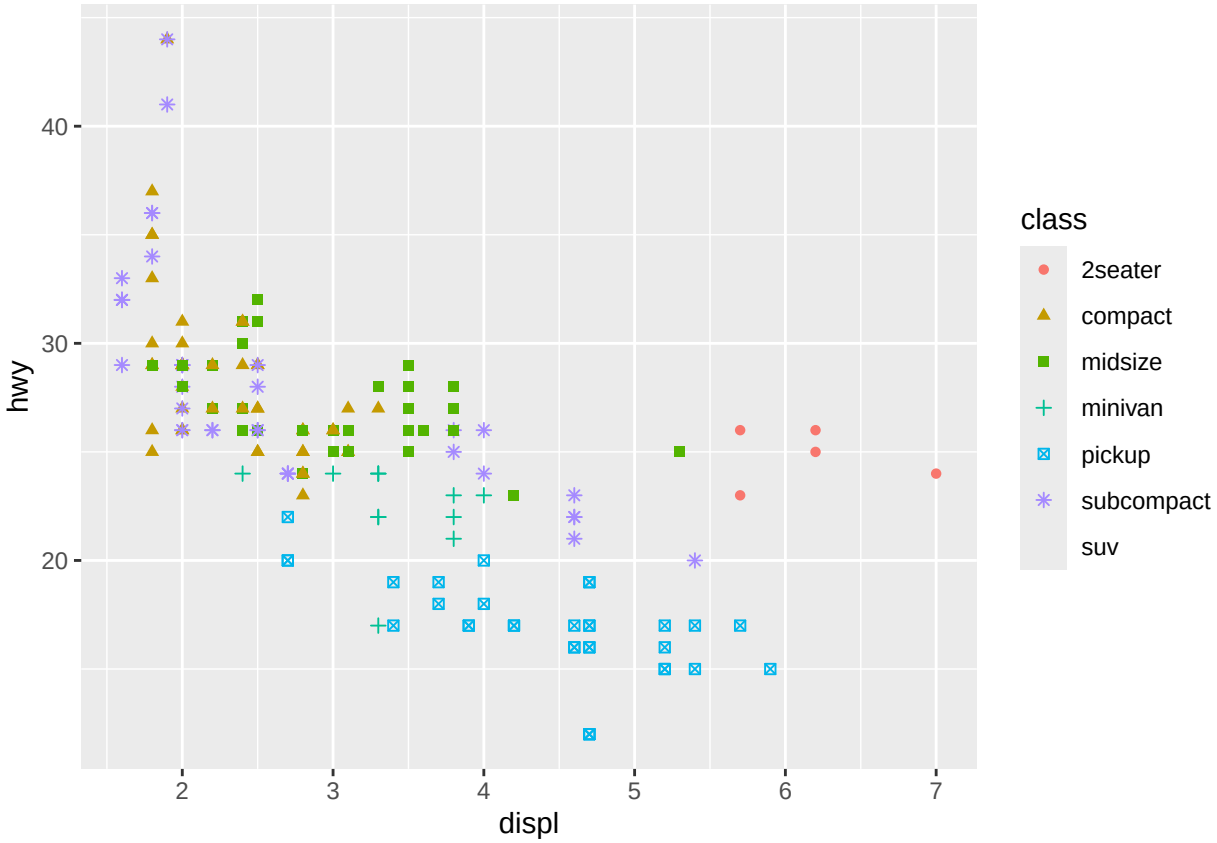
if you use geom_point it has no affect, mainly linewidth is used with geom_line

4. What happens if you map the same variable to multiple aesthetics?

```
ggplot(mpg, aes(x = displ, y = hwy, color = class, shape = class)) +
  geom_point()
```

```
## Warning: The shape palette can deal with a maximum of 6 discrete values because more
## than 6 becomes difficult to discriminate
## i you have requested 7 values. Consider specifying shapes manually if you need
##   that many of them.
```

```
## Warning: Removed 62 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

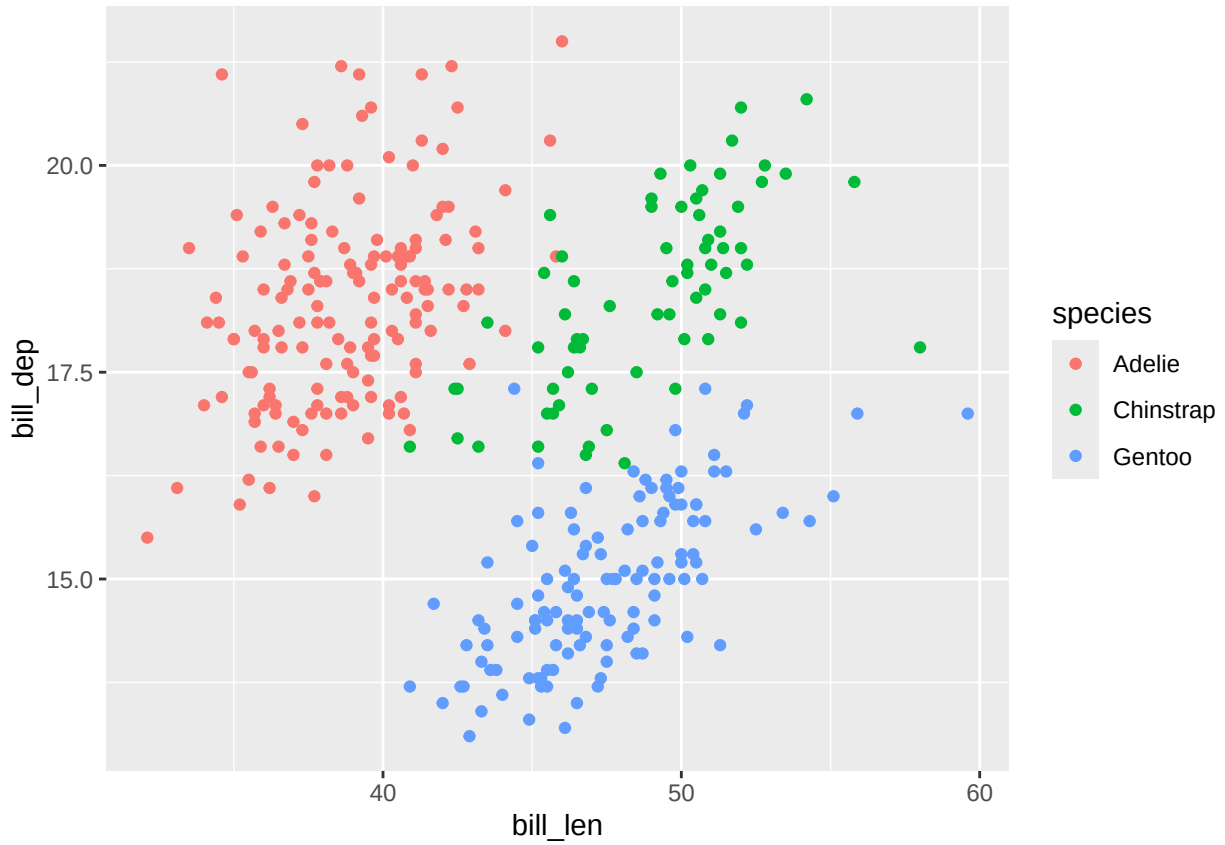


it is useful for highlighting because it highlights the different colors and shapes

5. Make a scatterplot of bill_depth_mm vs. bill_length_mm and color the points by species. What does adding coloring by species reveal about the relationship between these two variables? What about faceting by species?

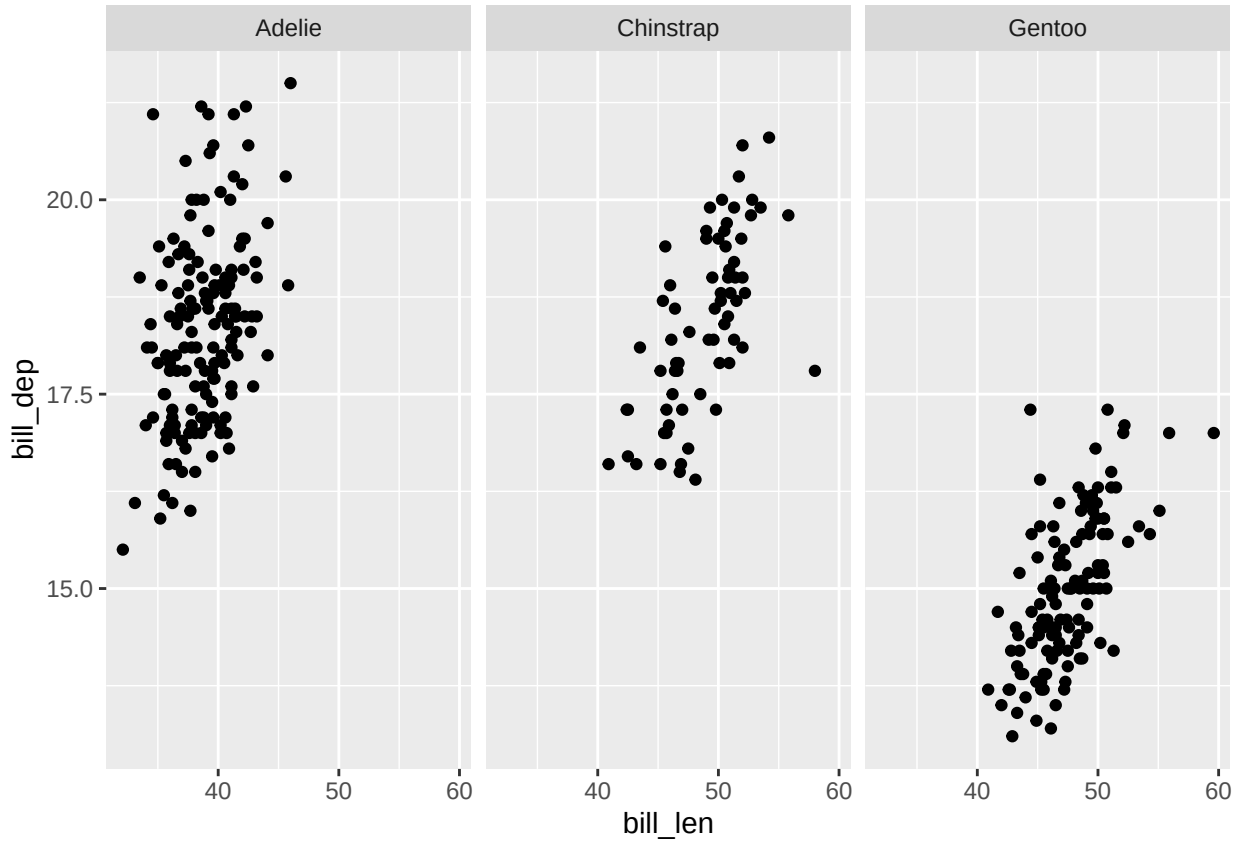
```
ggplot(penguins, aes(x = bill_len, y = bill_dep, color = species)) +
  geom_point()
```

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




```
ggplot(penguins, aes(x = bill_len, y = bill_dep)) +
  geom_point() +
  facet_wrap(~species)
```

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

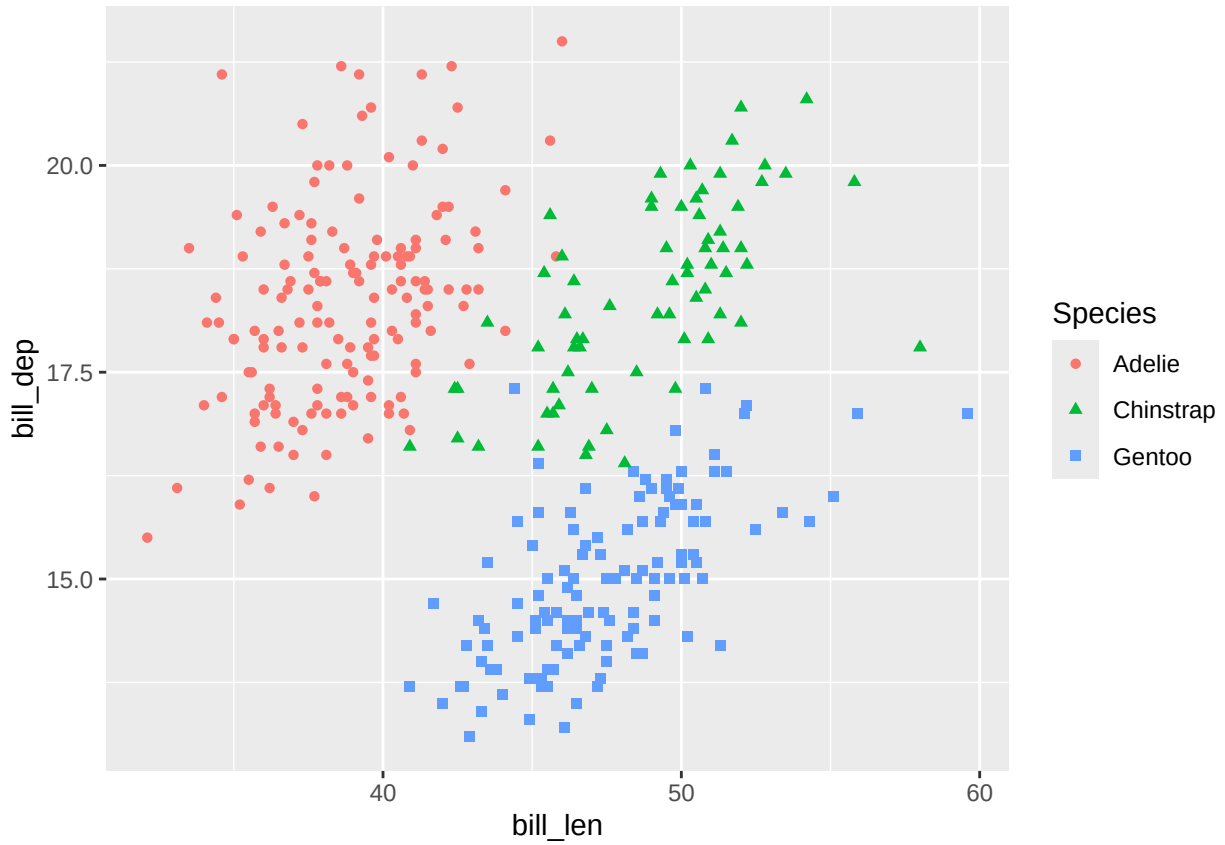


they reveal how scattered they are with color by species, and with facet_wrap, it separates them

6. Why does the following yield two separate legends? How would you fix it to combine the two legends?

```
ggplot(
  data = penguins,
  mapping = aes(
    x = bill_len, y = bill_dep,
    color = species, shape = species
  )
) +
  geom_point() +
  labs(color = "Species", shape = "Species")
```

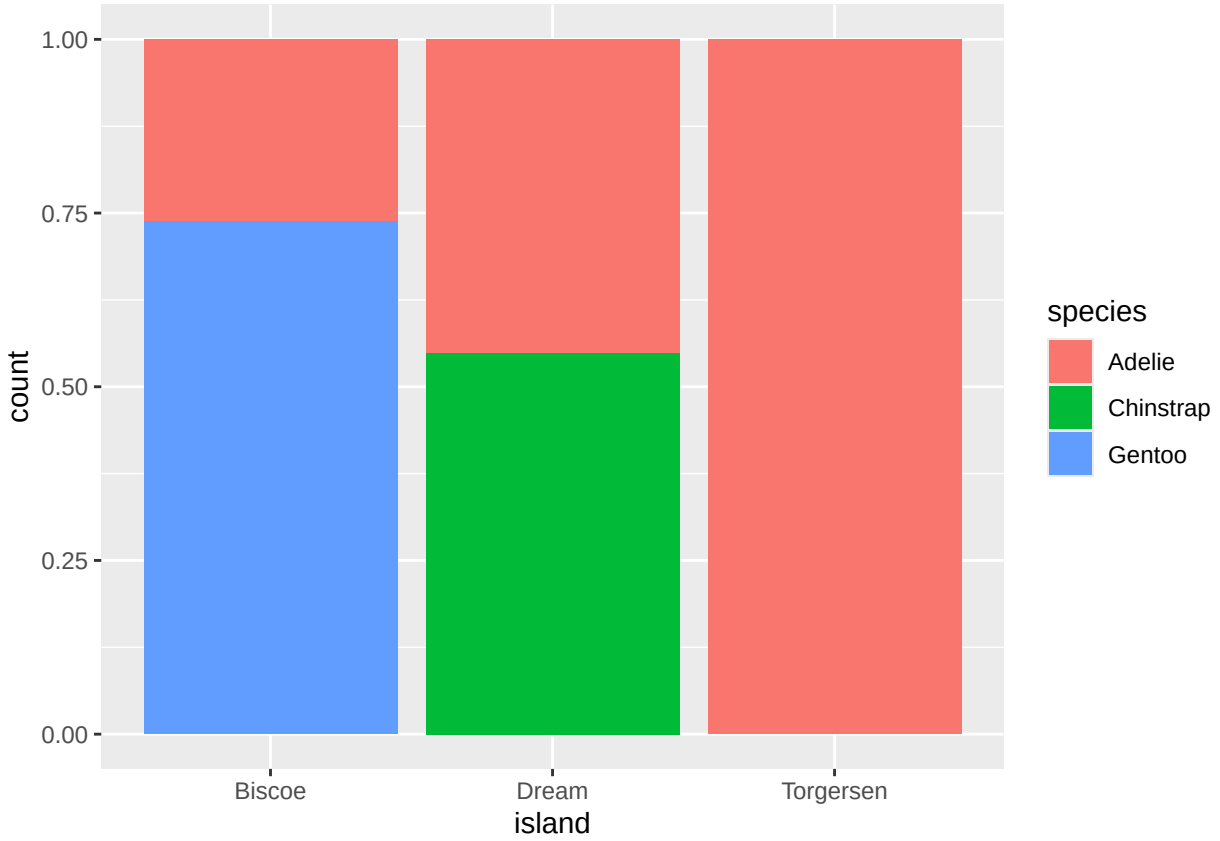
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



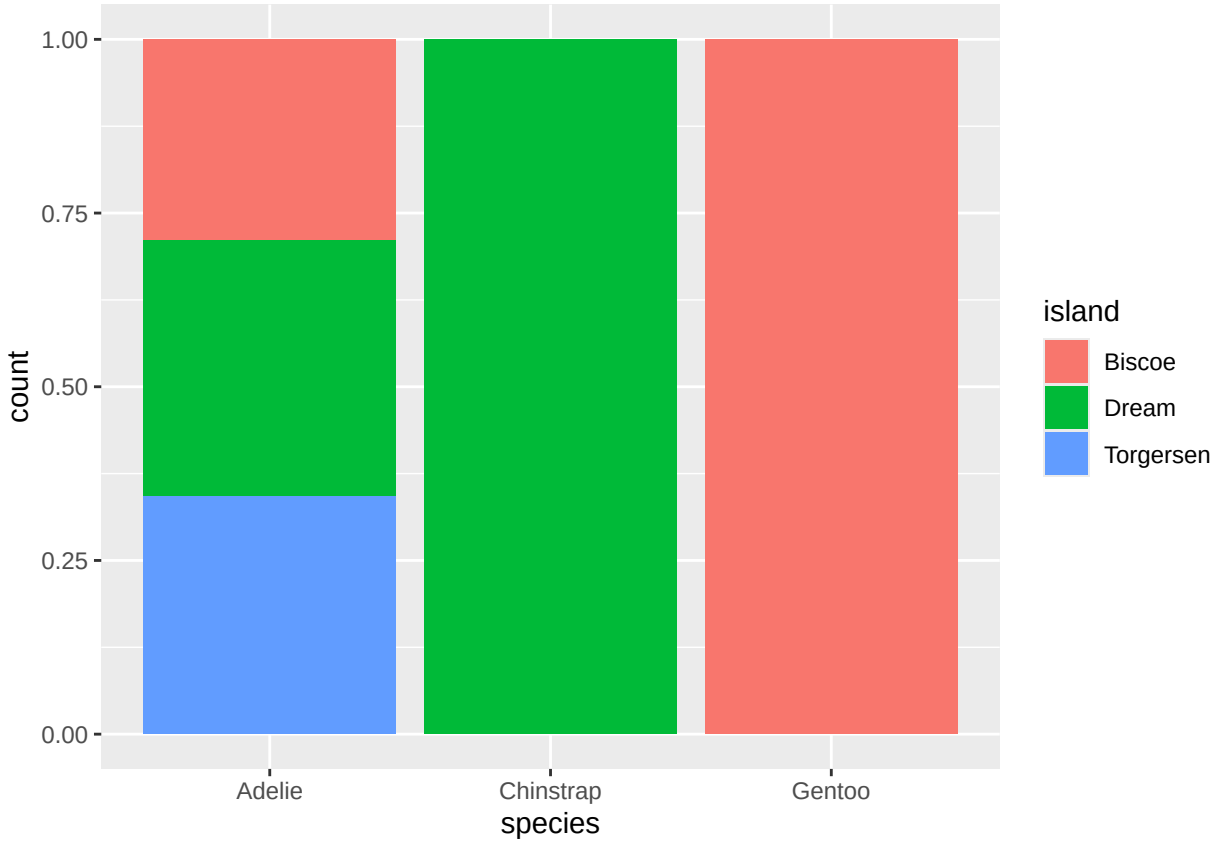
you can fix it by also adding shape = "Species in labs

7. Create the two following stacked bar plots. Which question can you answer with the first one? Which question can you answer with the second one?

```
ggplot(penguins, aes(x = island, fill = species)) +
  geom_bar(position = "fill")
```




```
ggplot(penguins, aes(x = species, fill = island)) +
  geom_bar(position = "fill")
```

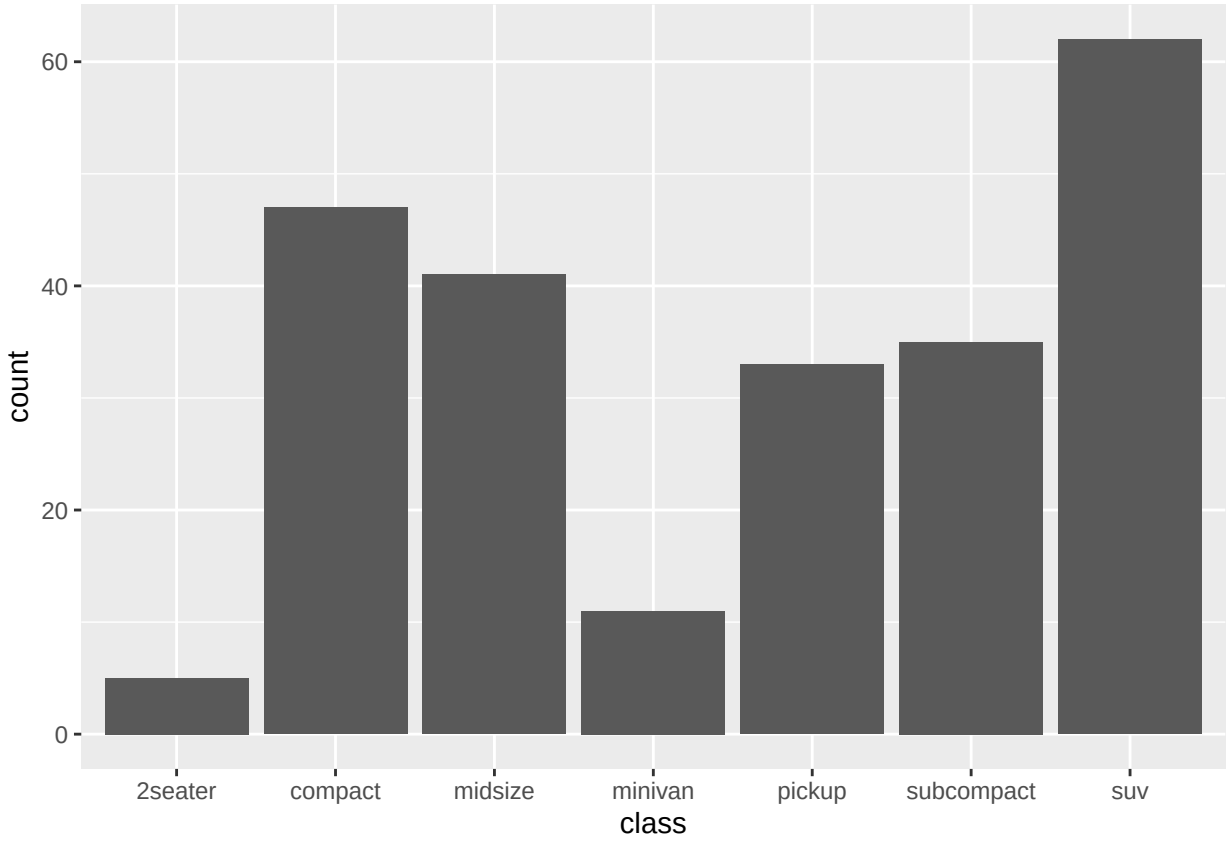


the first graph tells how many species in island, and the second tells how many species comes from what island

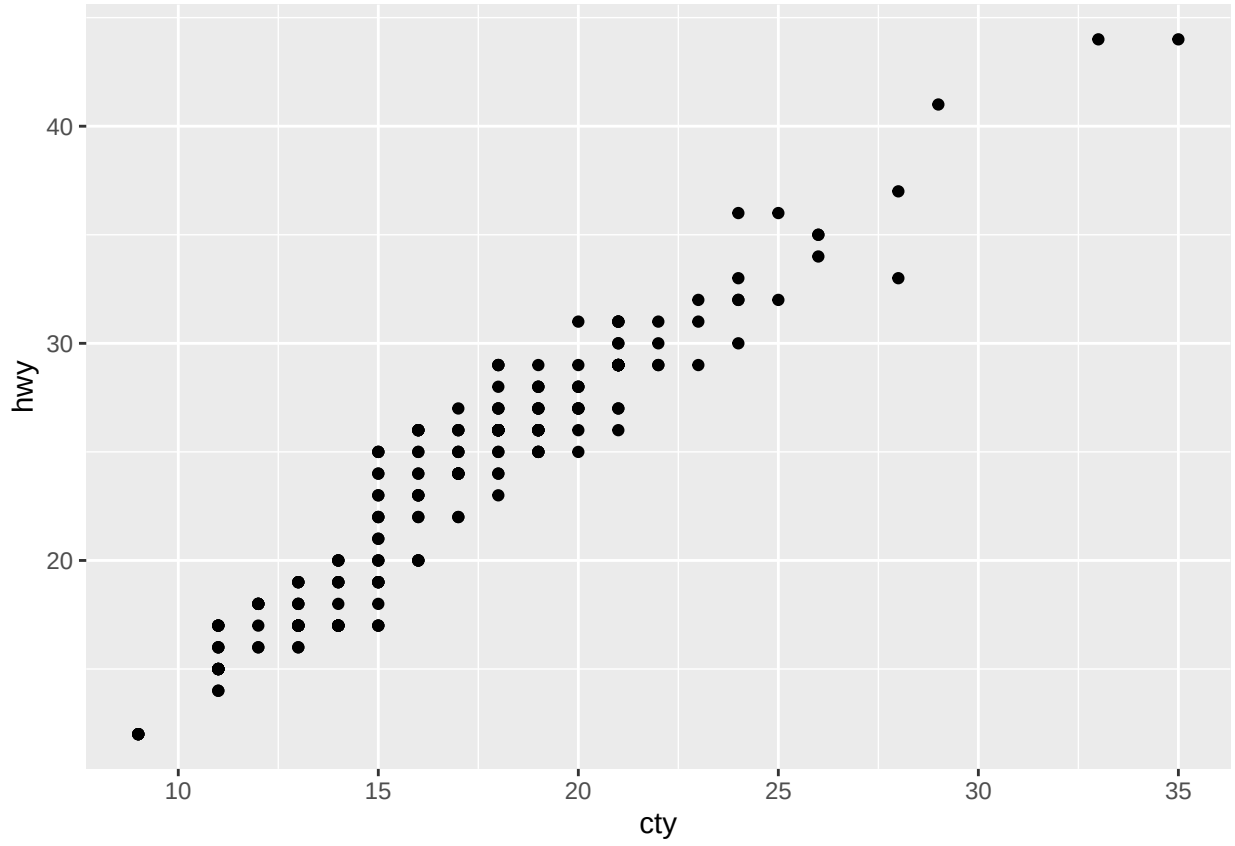
Exercise 1.6.1

1. Run the following lines of code. Which of the two plots is saved as mpg-plot.png? Why?

```
library(ggplot2)
ggplot(mpg, aes(x = class)) +
  geom_bar()
```




```
ggplot(mpg, aes(x = cty, y = hwy)) +
  geom_point()
```



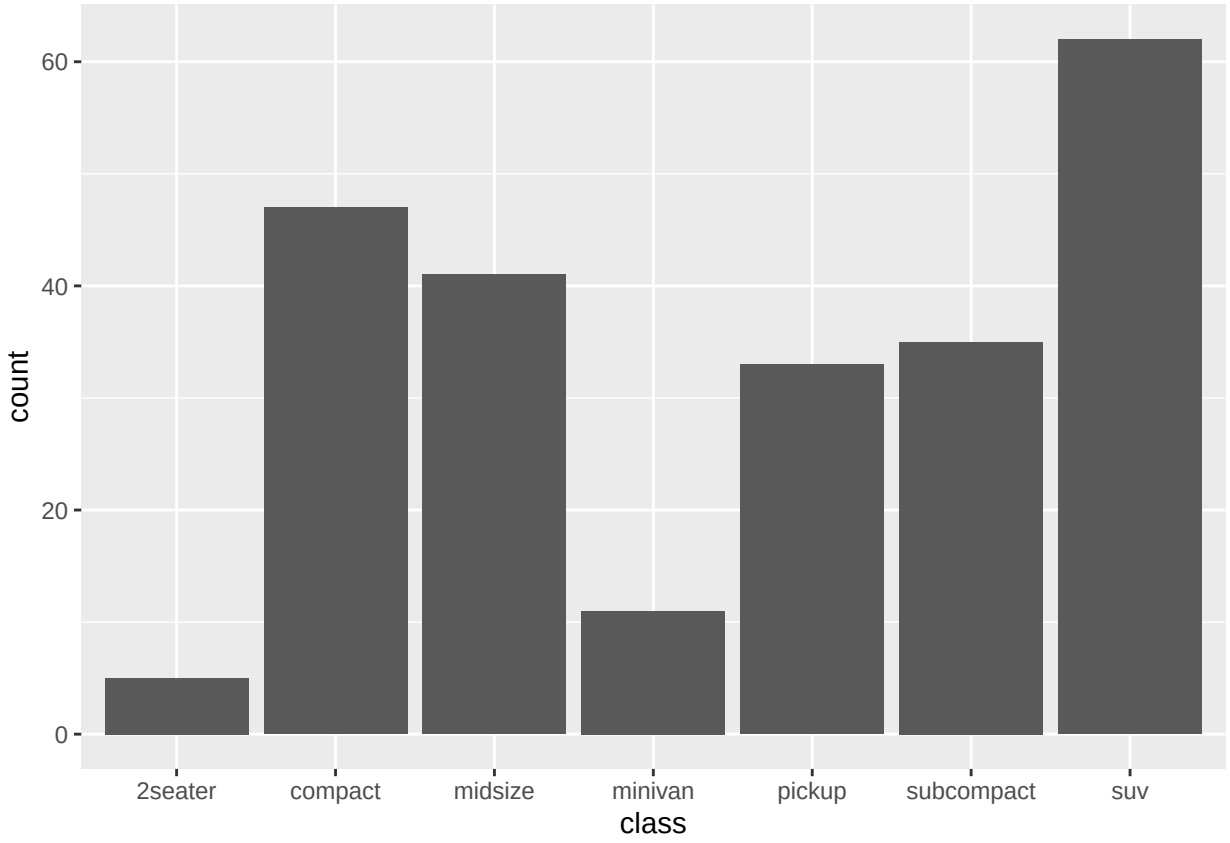

```
ggsave("mpg-plot.png")
```

```
## Saving 6.5 x 4.5 in image
```

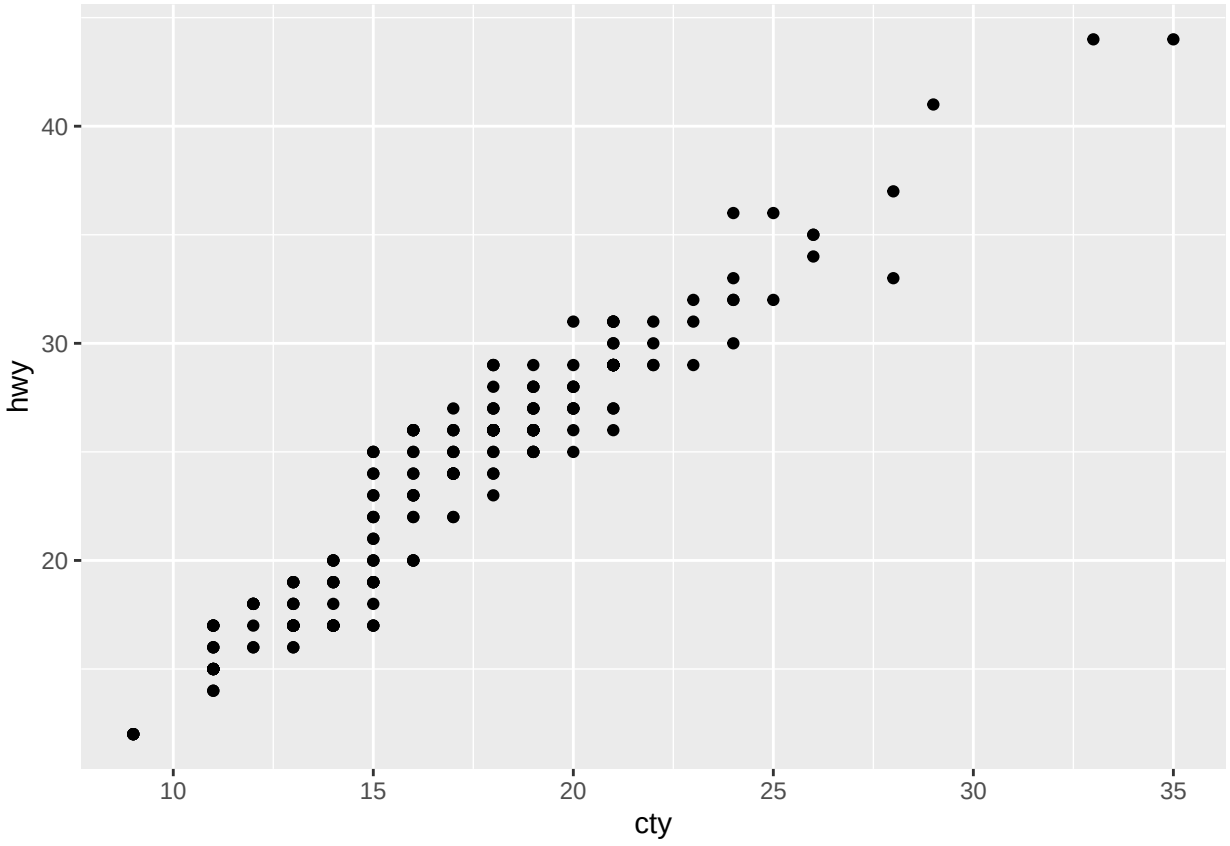
it saves the second plot, the first one got overwritten

2. What do you need to change in the code above to save the plot as a PDF instead of a PNG? How could you find out what types of image files would work in ggsave()?

```
ggplot(mpg, aes(x = class)) +
  geom_bar()
```




```
ggplot(mpg, aes(x = cty, y = hwy)) +
  geom_point()
```




```
ggsave("mpg-plot.pdf")
```

```
## Saving 6.5 x 4.5 in image
```

by running ?ggsave it reveals the formats that you can save it to, just change the .png to .pdf