1. Data visualization

2024-07-10

```
library(ggplot2)
library(ggthemes)

cggplot <- function(...) {
    ggplot(...) + scale_color_colorblind()
}

theme_set(theme_few())
set.seed(123)
```

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

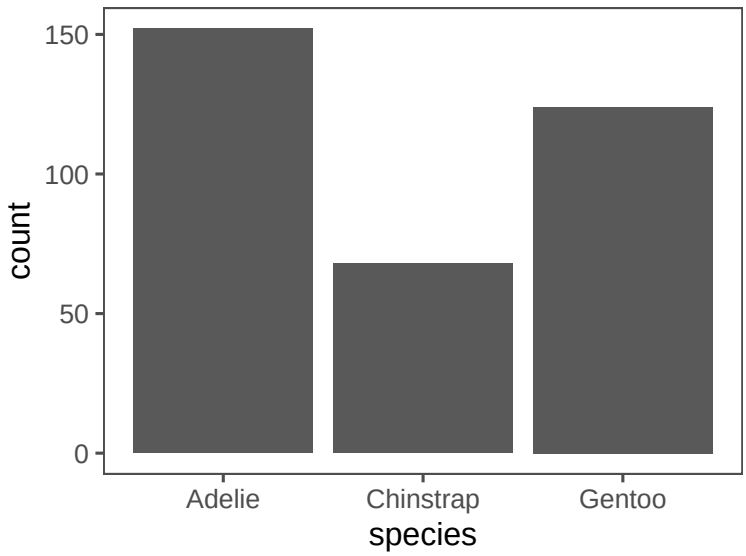
```
glimpse(penguins)
```

```
## 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~
```

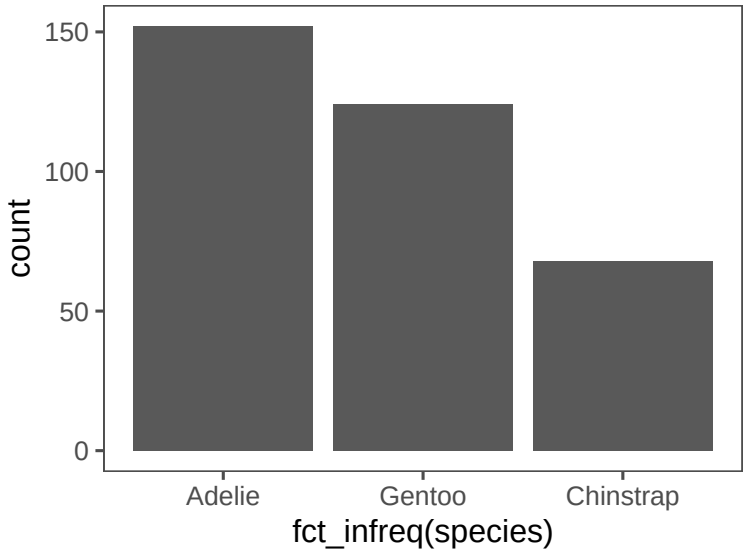
1.4 Visualizing distributions

A categorical variable

```
cggplot(penguins, aes(x = species)) + geom_bar()
```

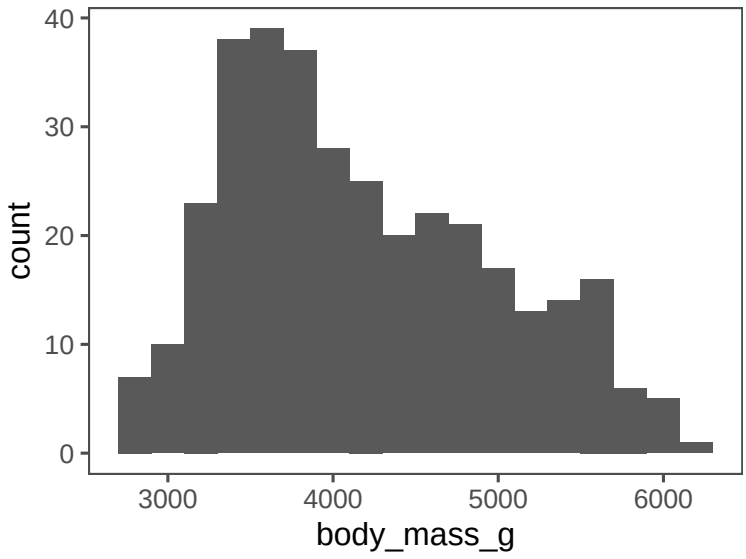



```
cggplot(penguins, aes(x = fct_infreq(species))) + geom_bar()
```



A numerical variable

```
cggplot(penguins, aes(x = body_mass_g)) + geom_histogram(binwidth = 200)
```



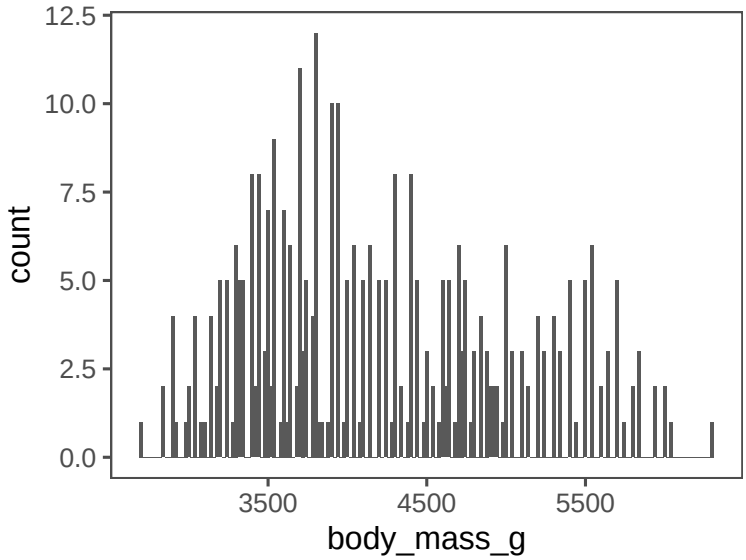
Inline code test:

```
Inline code test:
```

4

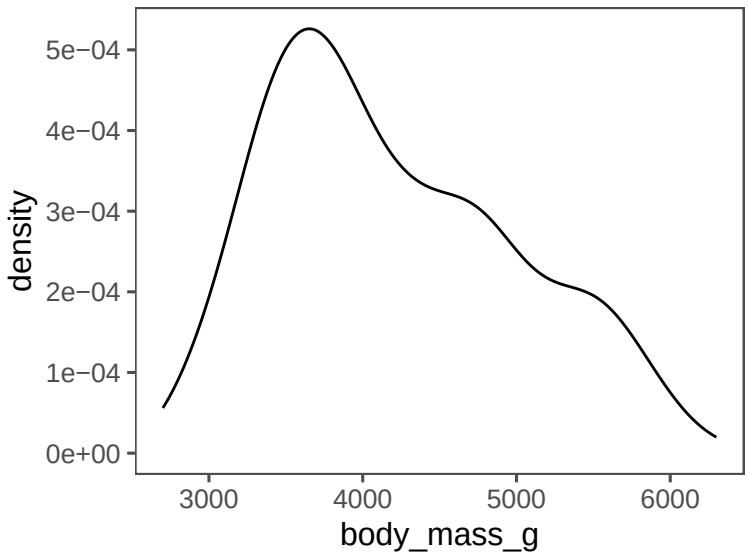
$\frac{1}{2}$

```
cggplot(penguins, aes(x = body_mass_g)) + geom_histogram(binwidth = 20)
```




```
cggplot(penguins, aes(x = body_mass_g)) + geom_density()
```

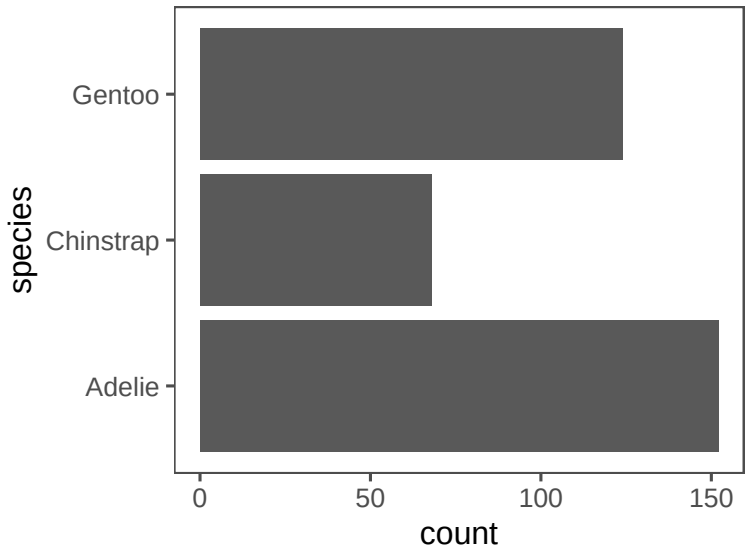
```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_density()`).
```



Exercises

1.

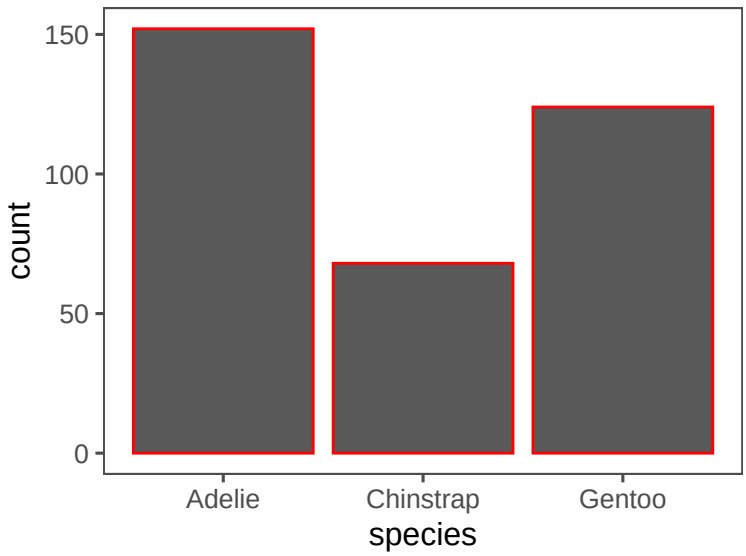
```
cggplot(penguins, aes(y = species)) + geom_bar()
```



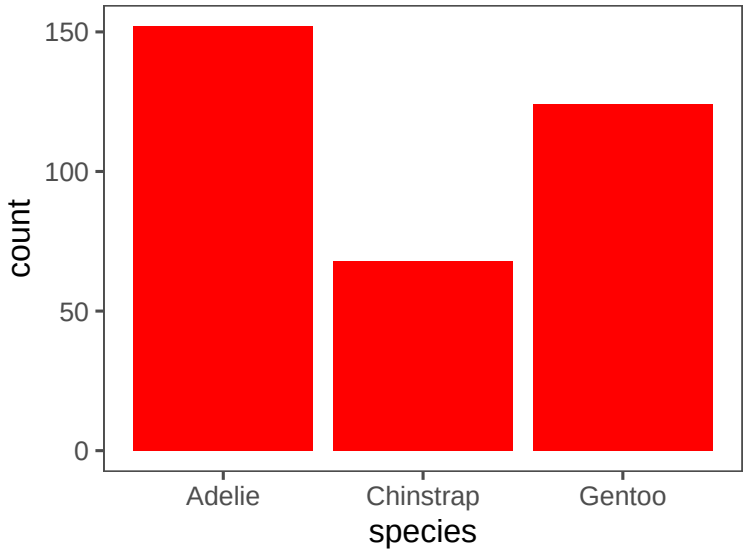
2.

```
species_plot = cggplot(penguins, aes(x = species))

species_plot + geom_bar(color = "red")
```




```
species_plot + geom_bar(fill = "red")
```

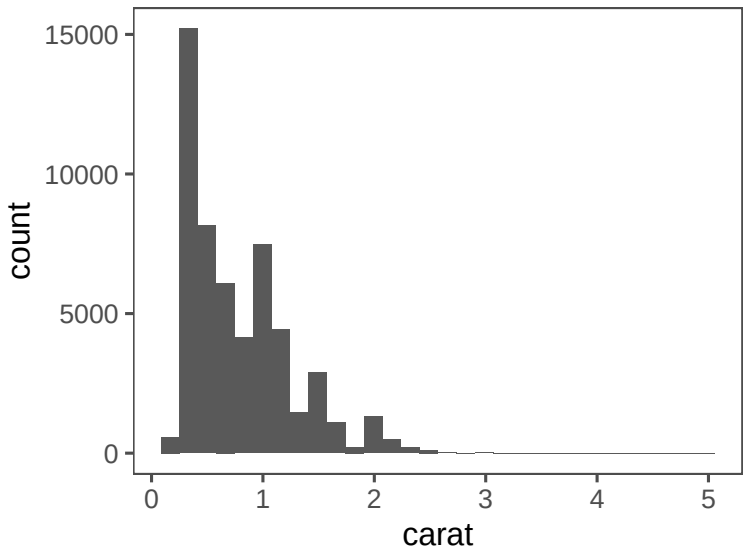


The aesthetic *fill* is more useful because it changes the whole bars; as opposed to the *color* aesthetic, which only modifies the bar borders.

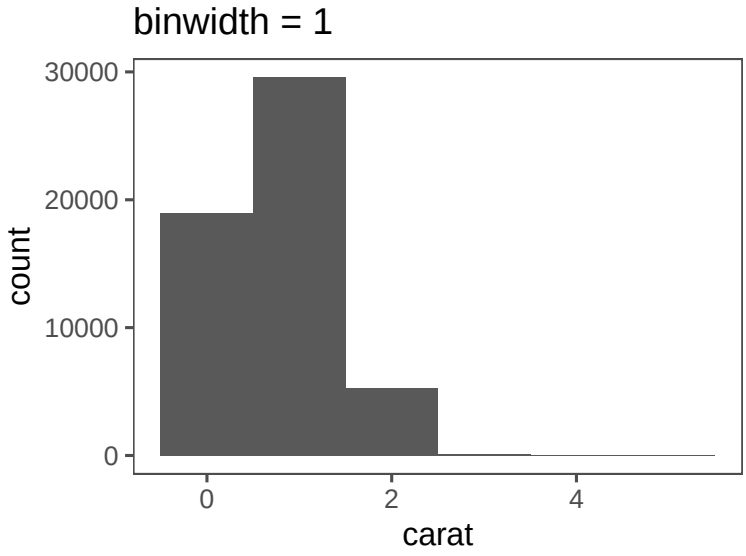
4.

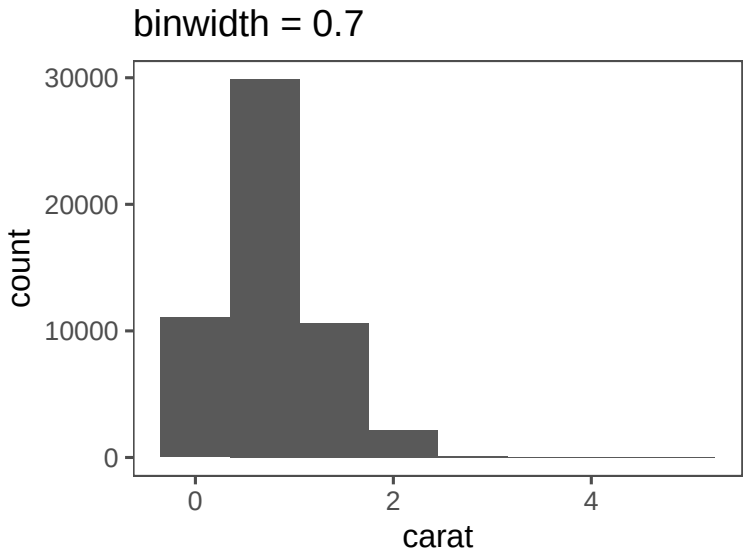
```
carat_plot <- cggplot(diamonds, aes(x = carat))
carat_plot + geom_histogram()
```

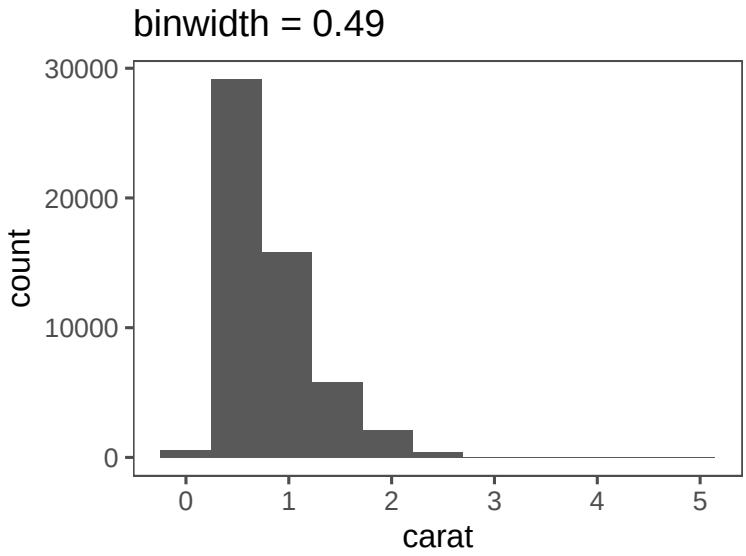
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

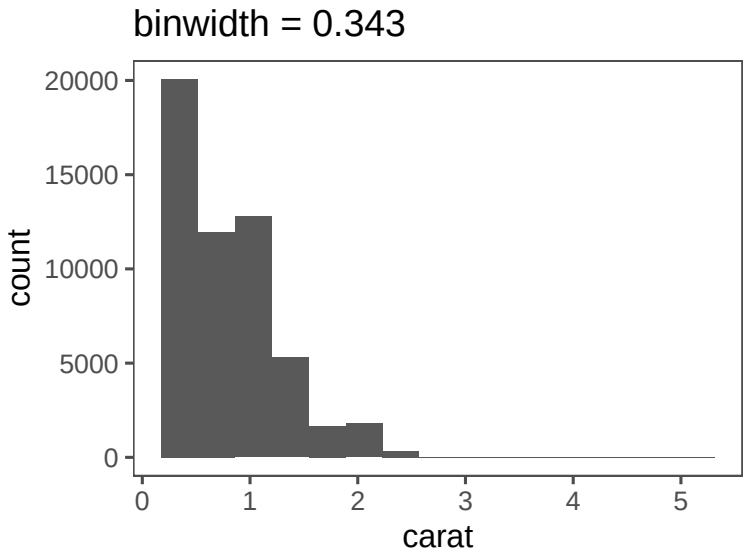


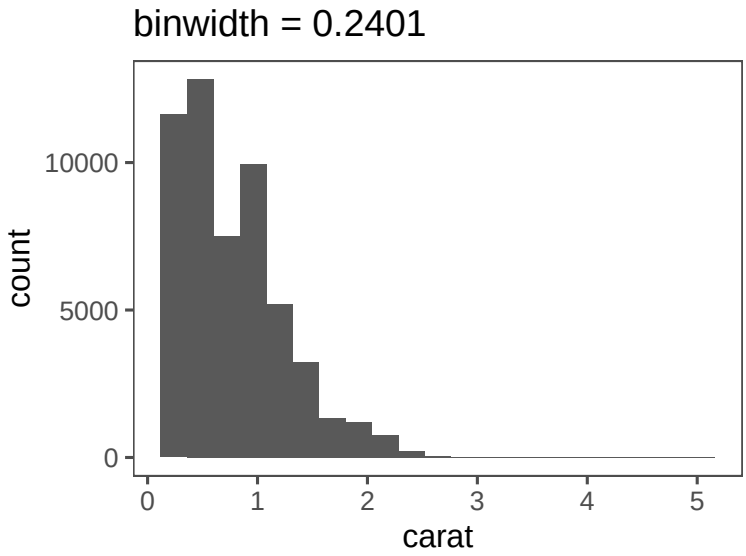

```
for (binwidth in 0.7^seq(0, 9)) {
    histogram <- carat_plot + geom_histogram(binwidth = binwidth) + labs(title = paste("binwidth =",
        binwidth))
    print(histogram)
}
```

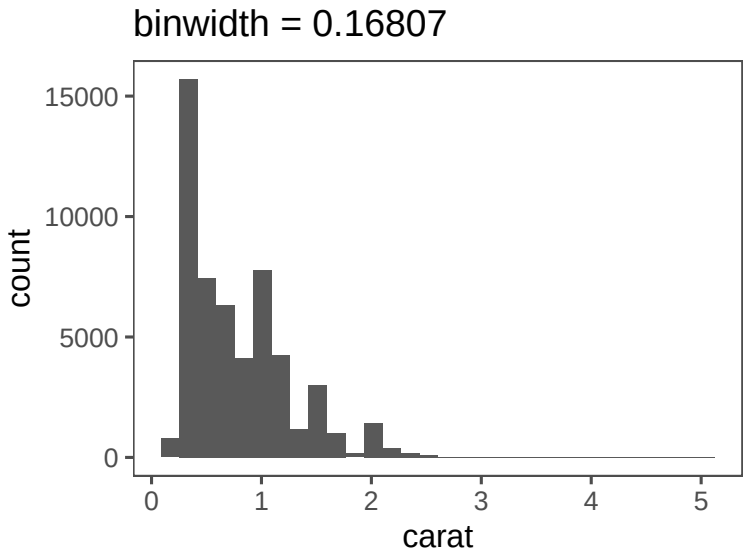


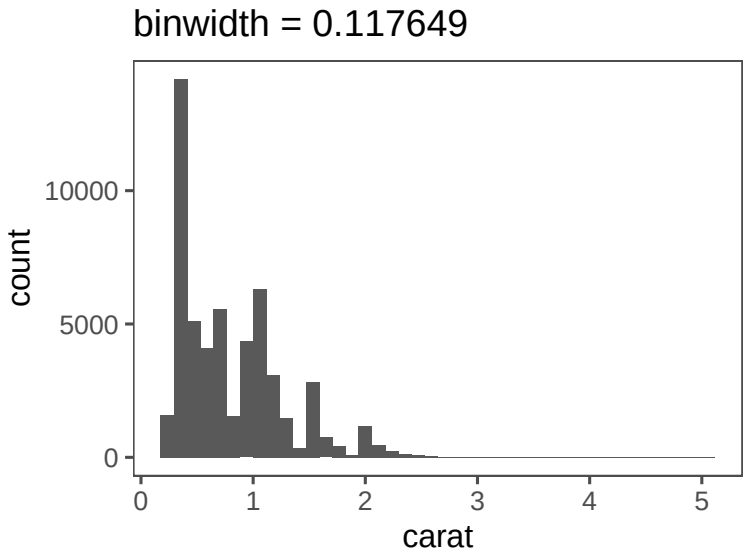

binwidth = 0.7
30000
20000
10000
0
count
0
2
4
carat

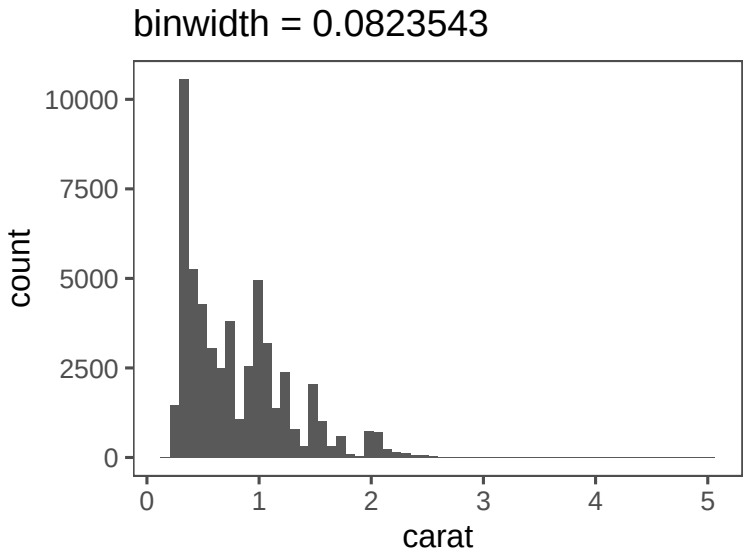

binwidth = 0.49
30000
20000
10000
0
count
0
1
2
3
4
5
carat

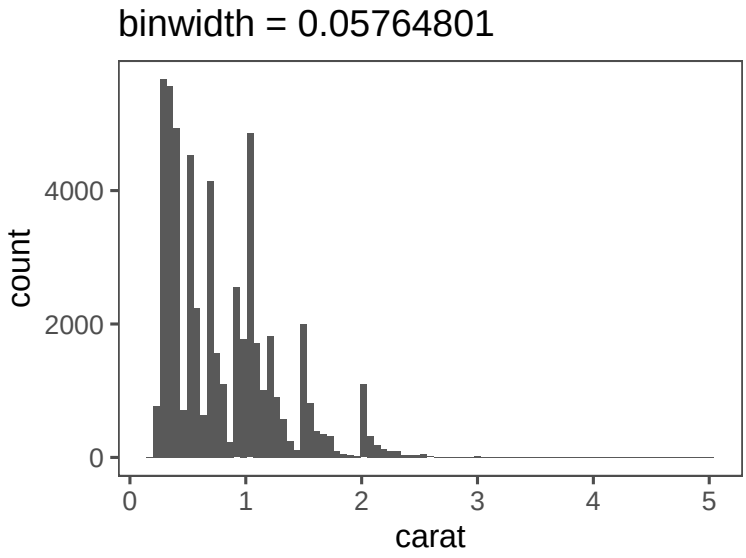

binwidth = 0.343
20000
15000
10000
5000
0
count
0
1
2
3
4
5
carat

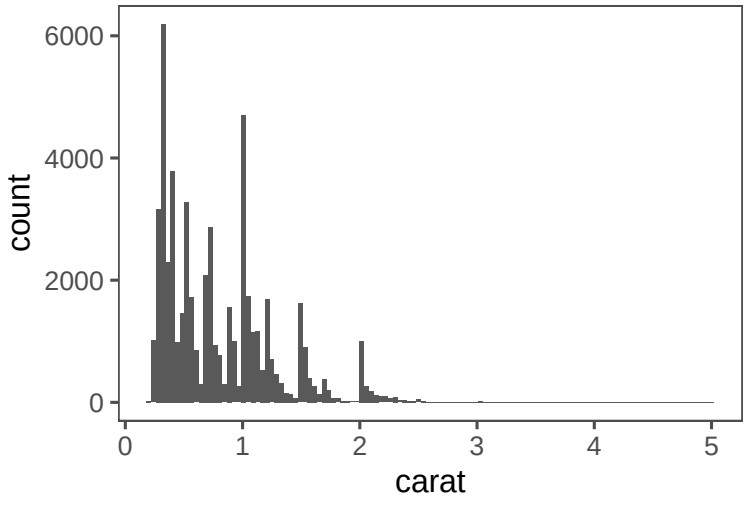

binwidth = 0.2401
count
10000
5000
0
0
1
2
3
4
5
carat


binwidth = 0.16807
count
15000
10000
5000
0
0
1
2
3
4
5
carat

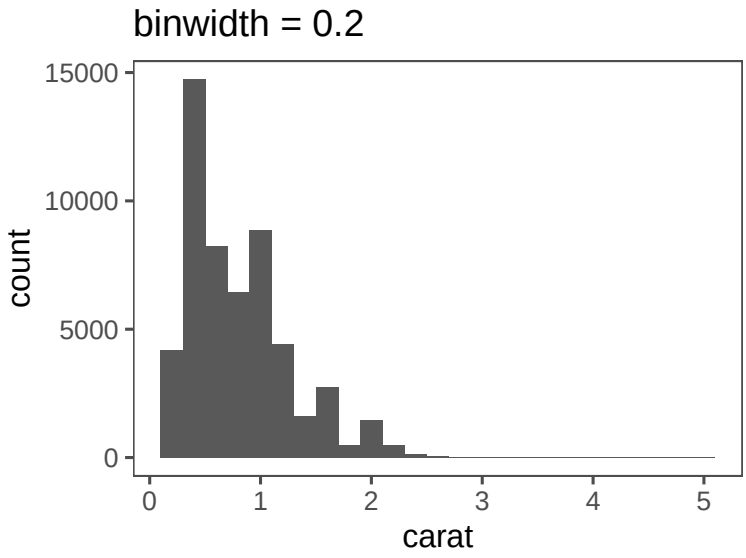

binwidth = 0.117649
count
10000
5000
0
0
1
2
3
4
5
carat


binwidth = 0.0823543
count
carat
0
2500
5000
7500
10000
0
1
2
3
4
5


binwidth = 0.05764801
count
carat
0
2000
4000
0
1
2
3
4
5


binwidth = 0.040353607
count
carat
0
2000
4000
6000
0
1
2
3
4
5

```
binwidth <- 0.2
carat_plot + geom_histogram(binwidth = binwidth) + labs(title = paste("binwidth =",
    binwidth))
```

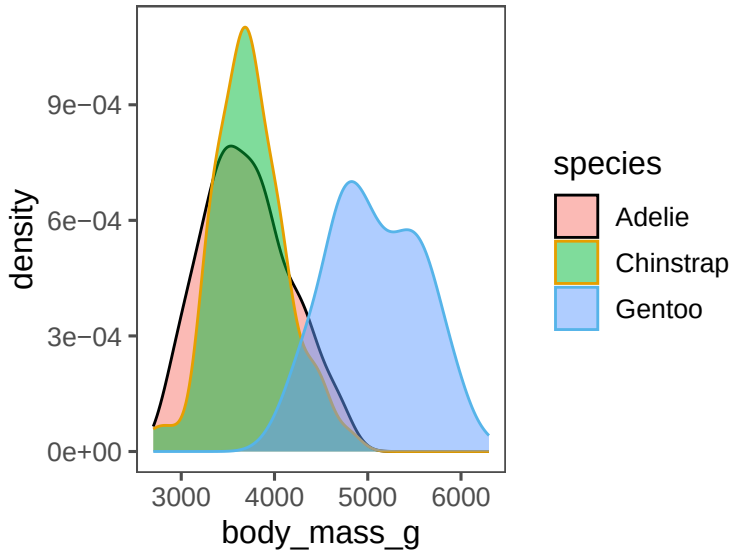


1.5 Visualizing relationships

A numerical and a categorical variable

```
cggplot(penguins, aes(x = body_mass_g, color = species, fill = species)) + geom_density(alpha = 0.5)
```

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



Exercises

1.

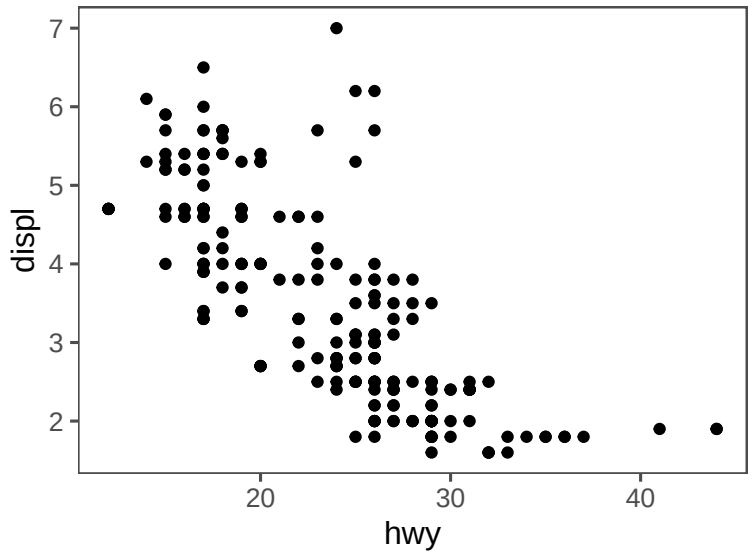
```
mpg
```

```
## # A tibble: 234 x 11
##    manufacturer model      displ  year   cyl trans drv     cty   hwy fl    class
##    <chr>        <chr>      <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>
##  1 audi         a4           1.8  1999     4 auto~ f        18    29 p     comp~
##  2 audi         a4           1.8  1999     4 manu~ f        21    29 p     comp~
##  3 audi         a4           2    2008     4 manu~ f        20    31 p     comp~
##  4 audi         a4           2    2008     4 auto~ f        21    30 p     comp~
##  5 audi         a4           2.8  1999     6 auto~ f        16    26 p     comp~
##  6 audi         a4           2.8  1999     6 manu~ f        18    26 p     comp~
##  7 audi         a4           3.1  2008     6 auto~ f        18    27 p     comp~
##  8 audi         a4 quattro   1.8  1999     4 manu~ 4        18    26 p     comp~
##  9 audi         a4 quattro   1.8  1999     4 auto~ 4        16    25 p     comp~
## 10 audi         a4 quattro   2    2008     4 manu~ 4        20    28 p     comp~
## # i 224 more rows
```

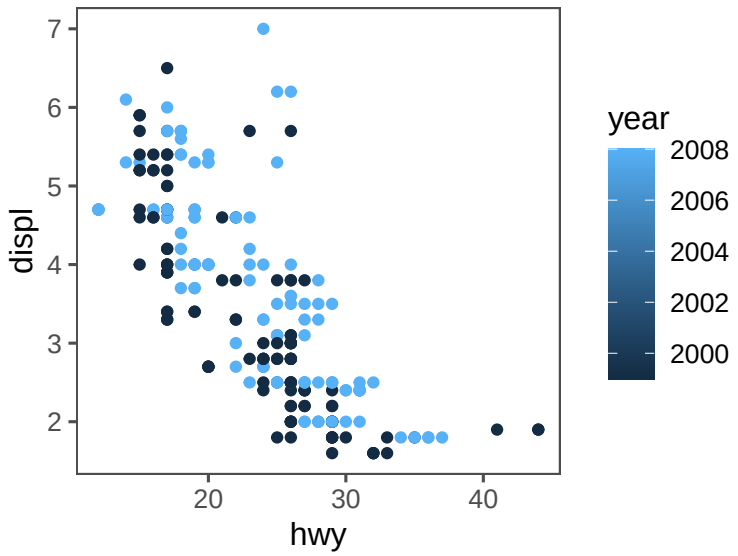
```
`?`(mpg)
```

2.

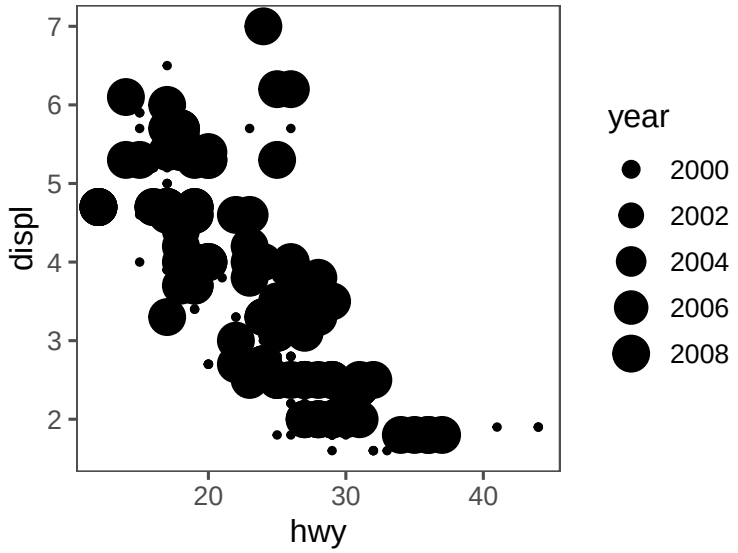
```
ggplot(mpg, aes(x = hwy, y = displ)) + geom_point()
```



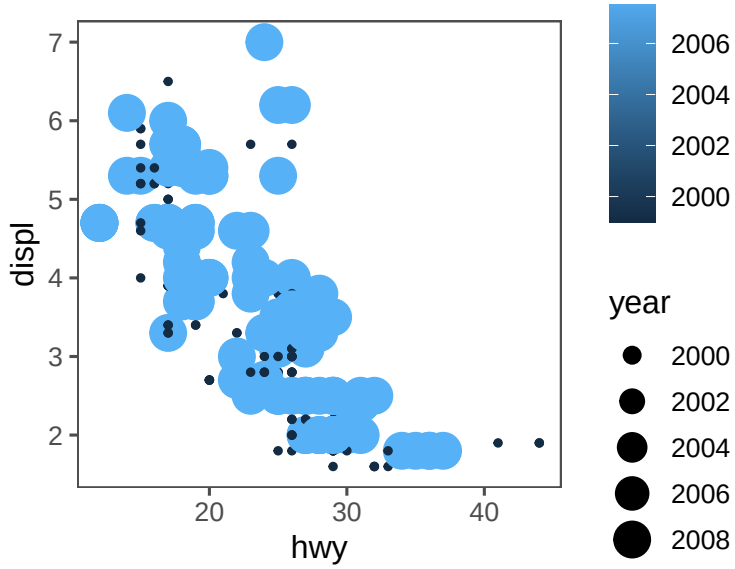

```
ggplot(mpg, aes(x = hwy, y = displ, color = year)) + geom_point()
```




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



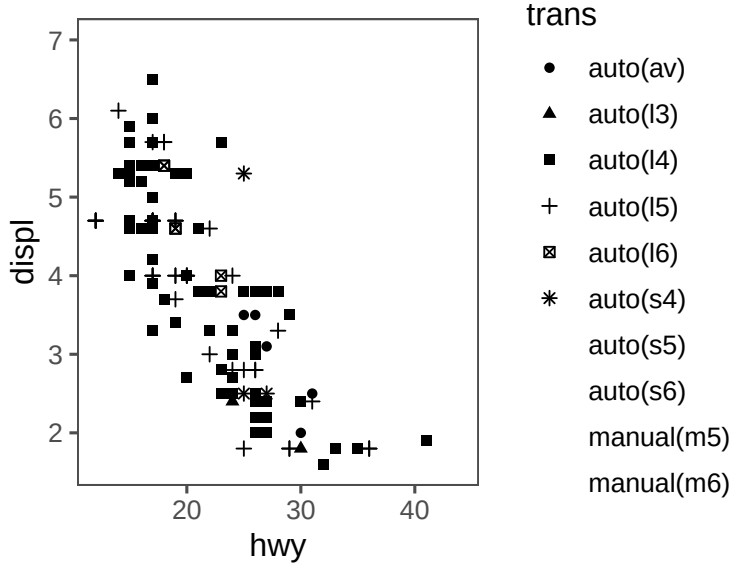

```
ggplot(mpg, aes(x = hwy, y = displ, color = year, size = year)) + geom_point()
```




```
# ggplot(mpg, aes(x = hwy, y = displ, shape = year)) + geom_point()
ggplot(mpg, aes(x = hwy, y = displ, shape = trans)) + 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 10 values. Consider specifying shapes manually if you need
##   that many have them.
```

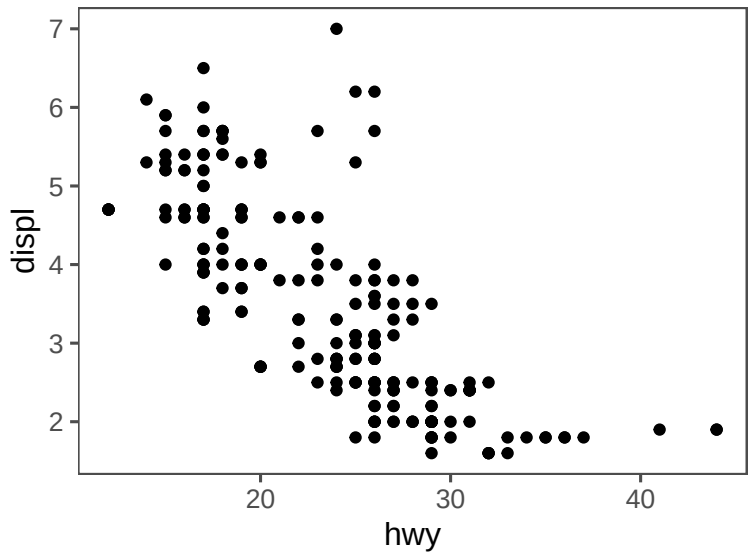
```
## Warning: Removed 96 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



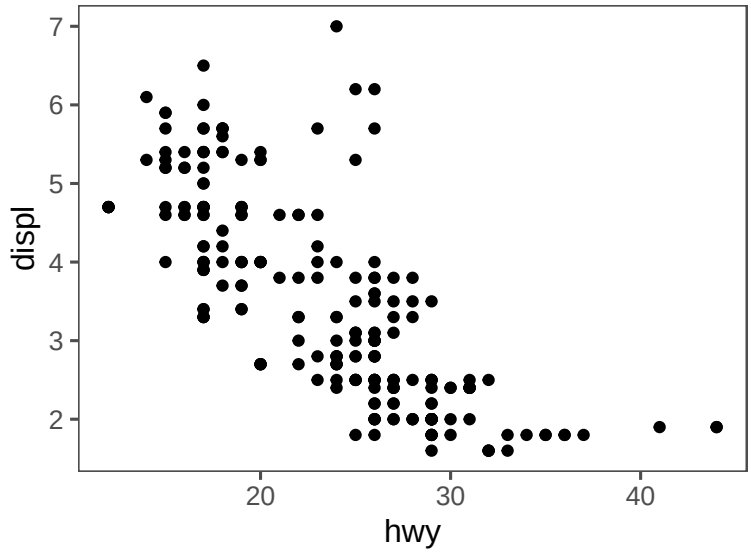
3.

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

```
## Warning in geom_point(linewidth = year): Ignoring unknown parameters:
## `linewidth`
```



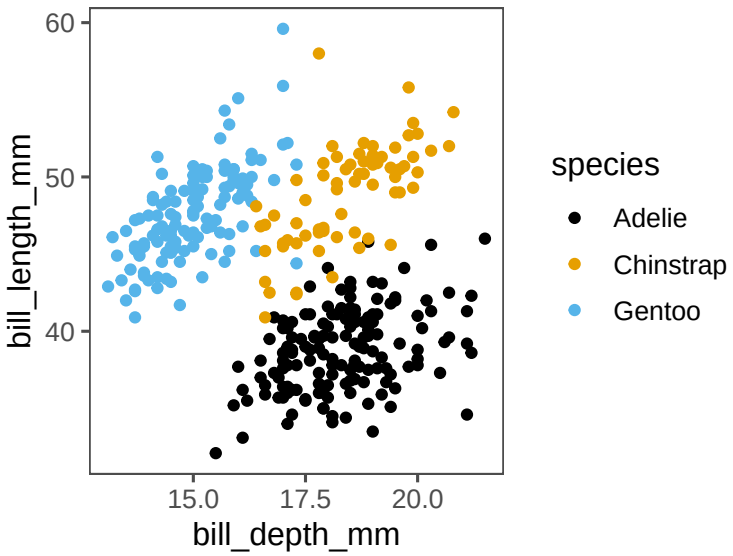

```
ggplot(mpg, aes(x = hwy, y = displ, linewidth = year)) + geom_point()
```



5.

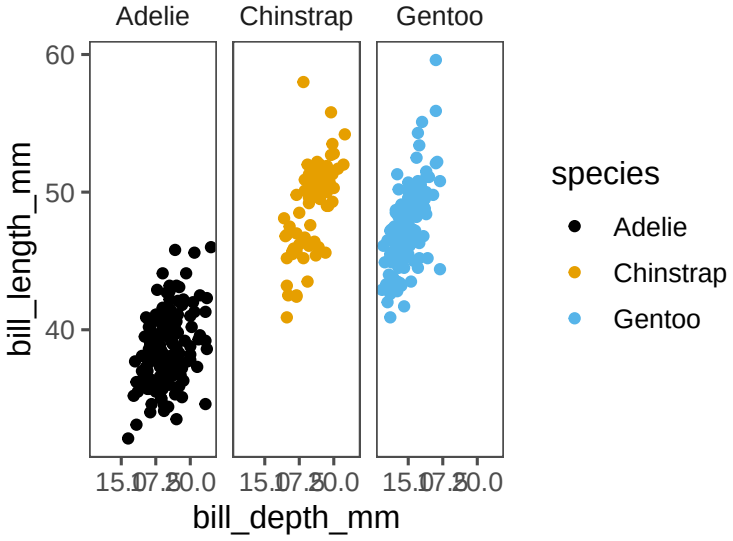
```
cggplot(penguins, aes(x = bill_depth_mm, y = bill_length_mm, color = species)) +
    geom_point()
```

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




```
cggplot(penguins, aes(x = bill_depth_mm, y = bill_length_mm, color = species)) +
    geom_point() + facet_wrap(~species)
```

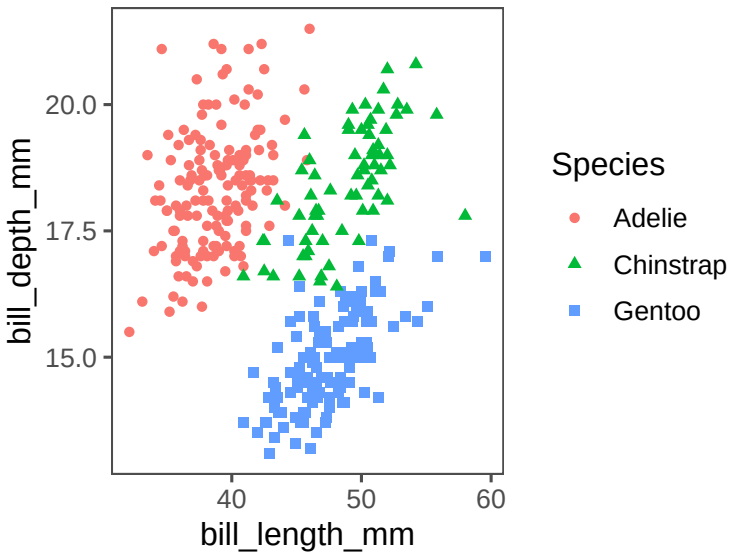
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



6.

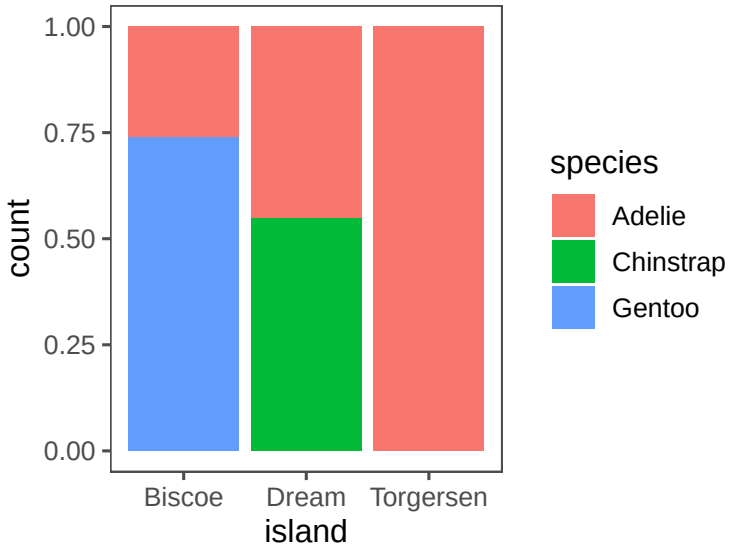
```
ggplot(data = penguins, mapping = aes(x = bill_length_mm, y = bill_depth_mm, color = species,
    shape = species)) + geom_point() + labs(color = "Species", shape = "Species")
```

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

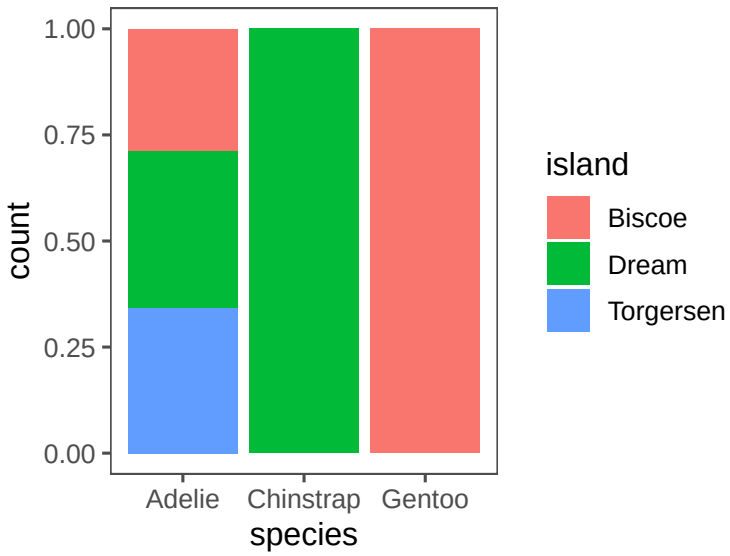


7.

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




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


1.00
0.75
0.50
0.25
0.00
count
Adelie
Chinstrap
Gentoo
species
island
Biscoe
Dream
Torgersen