Statistics for Toastmasters

Ken Butler

September 7, 2021

... with extra penguins!

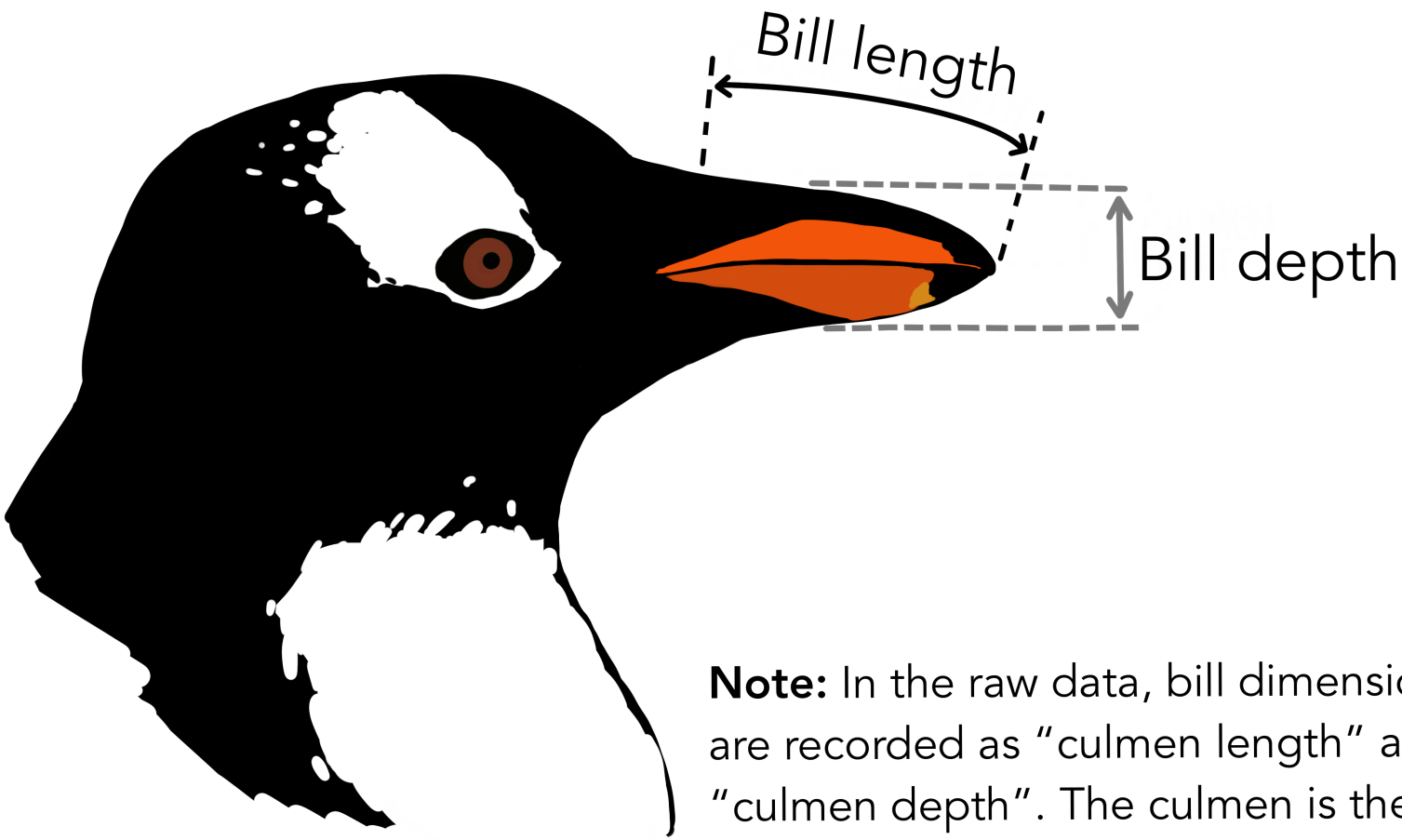
Figure 1: Gentoo penguins

Source

What we'll be doing

- Have a dataset on three different species of penguins (Adelie, Chinstrap, Gentoo)
- Contains species, measurements on bills and flippers
- Want to learn something about the data
- For example, how the penguins compare
- Or, can we tell the species apart from the measurements we have?

The bill of a penguin



Note: In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal ridge atop the bill.

Figure 2: Bill or culmen depth

Our data (of 344 penguins total)

species	bill_length_mm	bill_depth_mm	flipper_length_mm
Adelie	37.6	19.1	194
Gentoo	46.5	13.5	210
Gentoo	51.1	16.5	225
Gentoo	45.5	13.7	214
Adelie	37.6	19.3	181
Gentoo	46.4	15.0	216
Adelie	34.6	21.1	198
Gentoo	52.5	15.6	221
Adelie	37.6	17.0	185
Chinstrap	47.6	18.3	195
Gentoo	49.0	16.1	216
Chinstrap	45.6	19.4	194

There are 344 penguins altogether.

The species

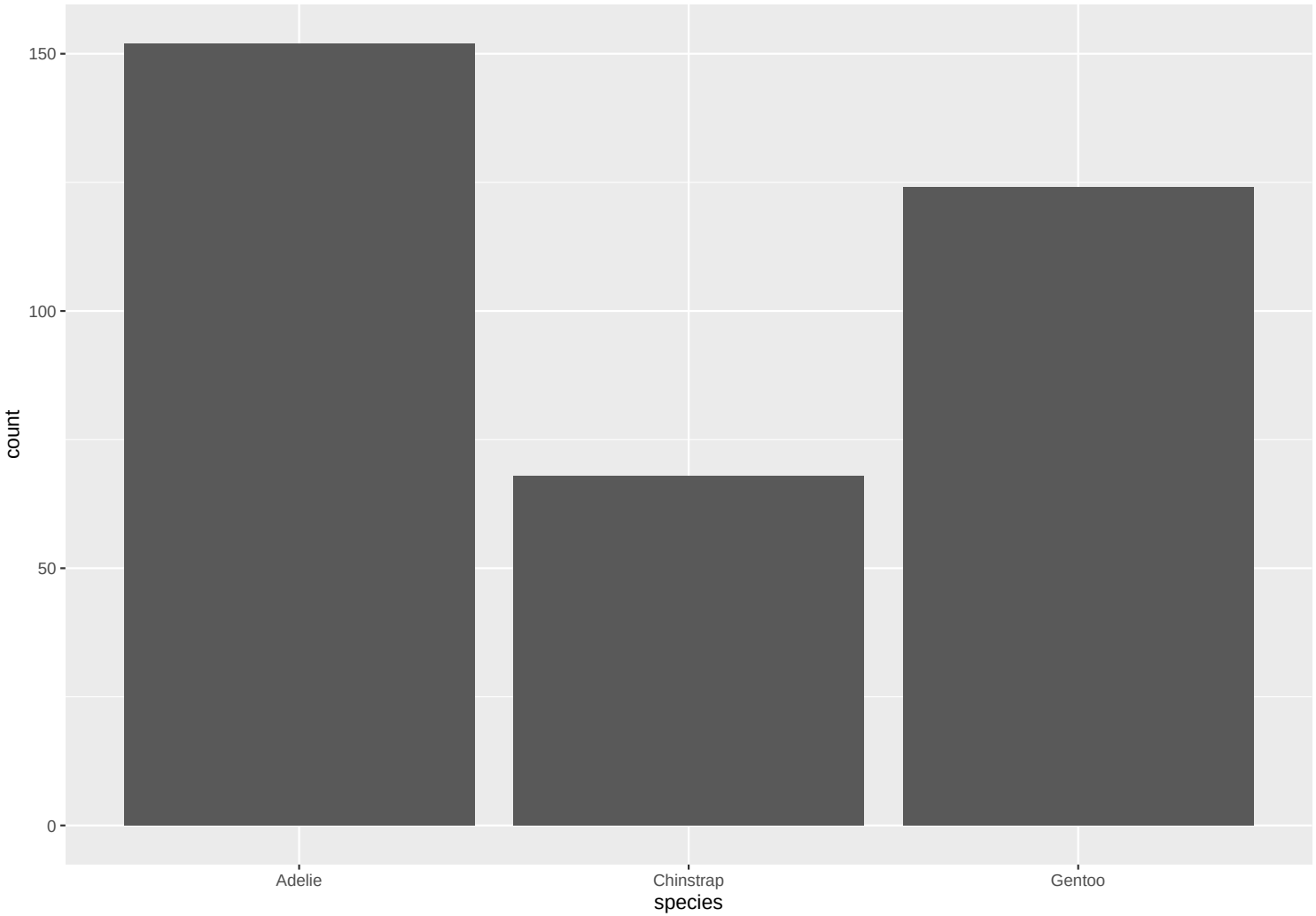
- a so-called "categorical variable" (classified)
- could count how many of each:

species	n
Adelie	152
Chinstrap	68
Gentoo	124

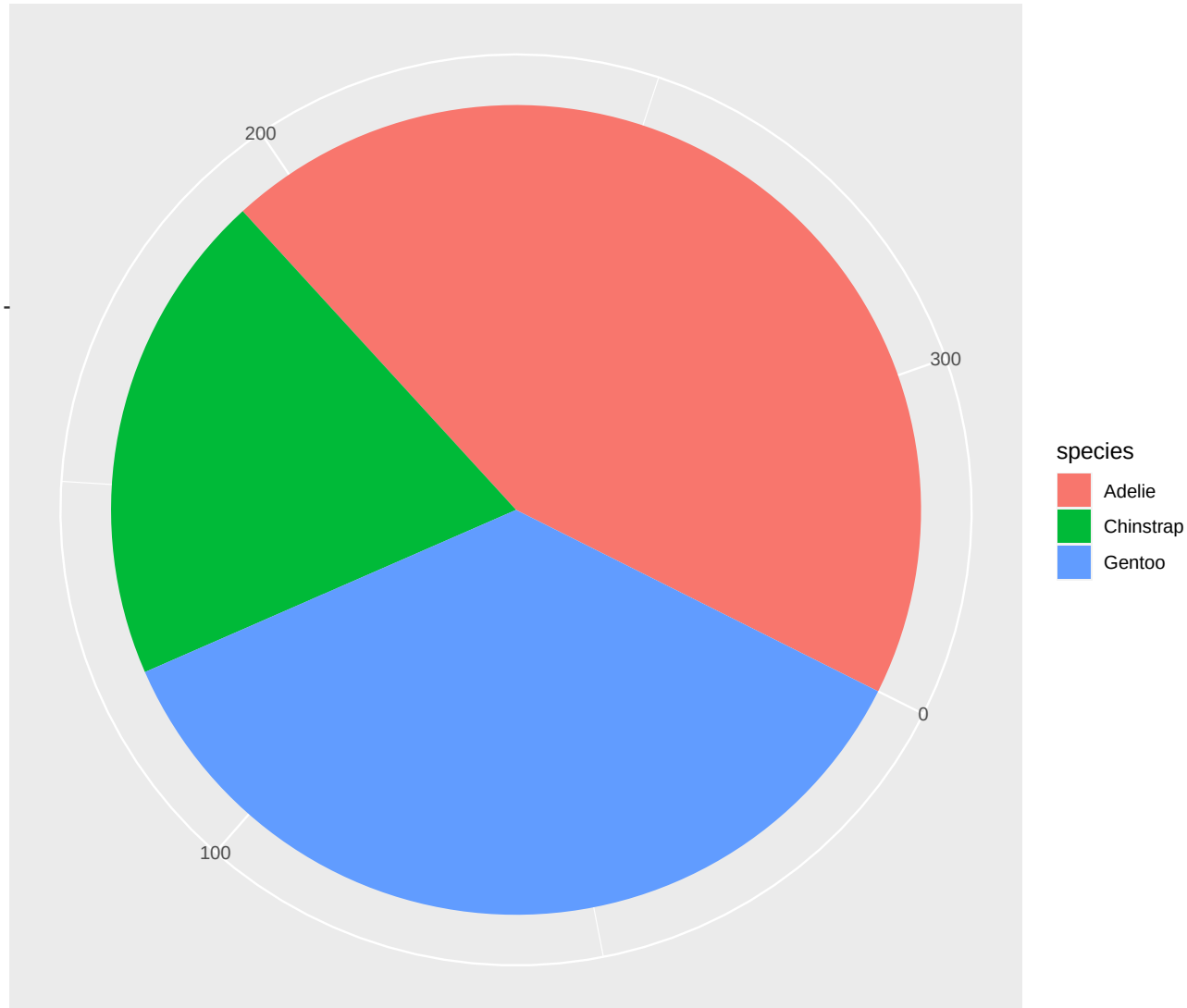
Or make a graph.

Bar chart

Best chart for a categorical variable is a bar chart:



Pie chart, if you must



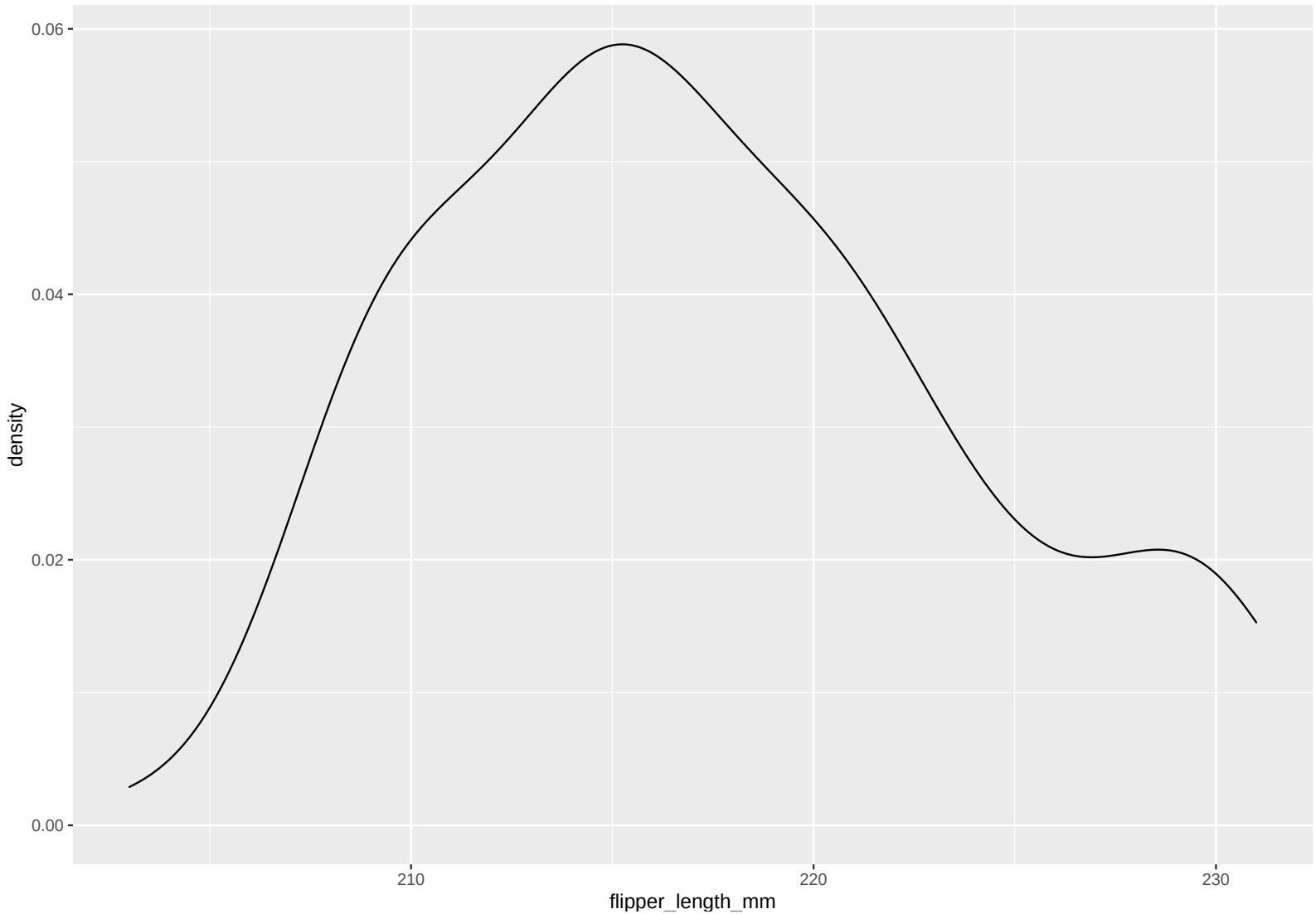
Flipper length

... is measured, not categorized ("quantitative"):

flipper_length_mm
196
202
230
197
198
190
188
215

Density plot: Gentoo flipper length

```
## Warning: Removed 1 rows containing non-finite values (stat_
```

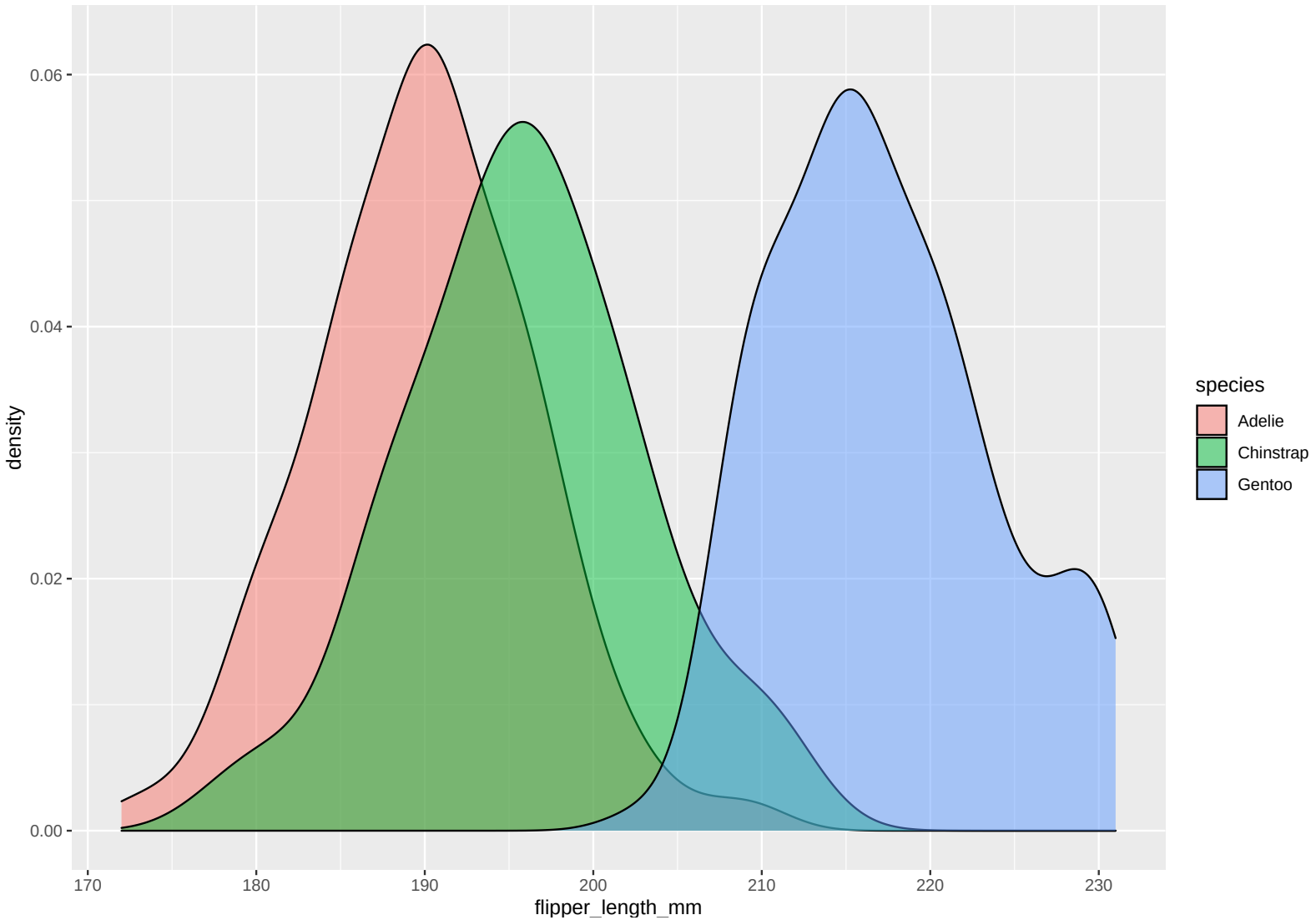


Comments

- most of the Gentoo penguins have flipper length near 215 mm
- there is variability: the flipper lengths vary between about 200 and 230 mm
- how do the other species compare?
- idea: do density plots `overlaid` (in different colours).

Density plot for all three species

```
## Warning: Removed 2 rows containing non-finite values (stat_
```



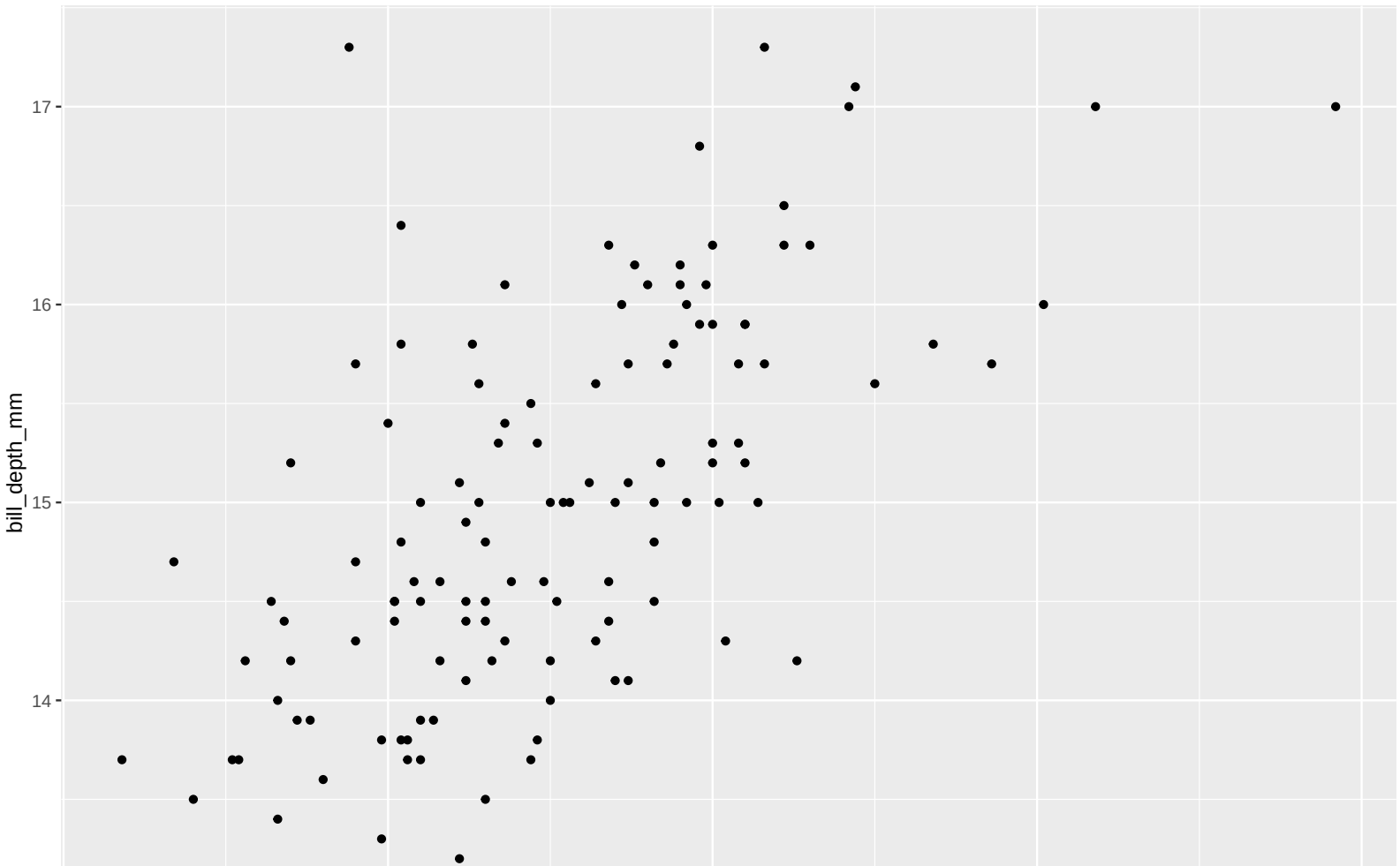
Comments

- Gentoo penguins have longer flippers than others
- Chinstrap slightly longer than Adelie, but a lot of overlap
- if all you knew was flipper length, you would do reasonably well at distinguishing Gentoo from others.
- but even then, would not be perfect (eg. what if flipper length was 205 mm?)

Two measured variables: scatterplot

Bill length vs depth for Gentoo:

```
## Warning: Removed 1 rows containing missing values (geom_poi
```



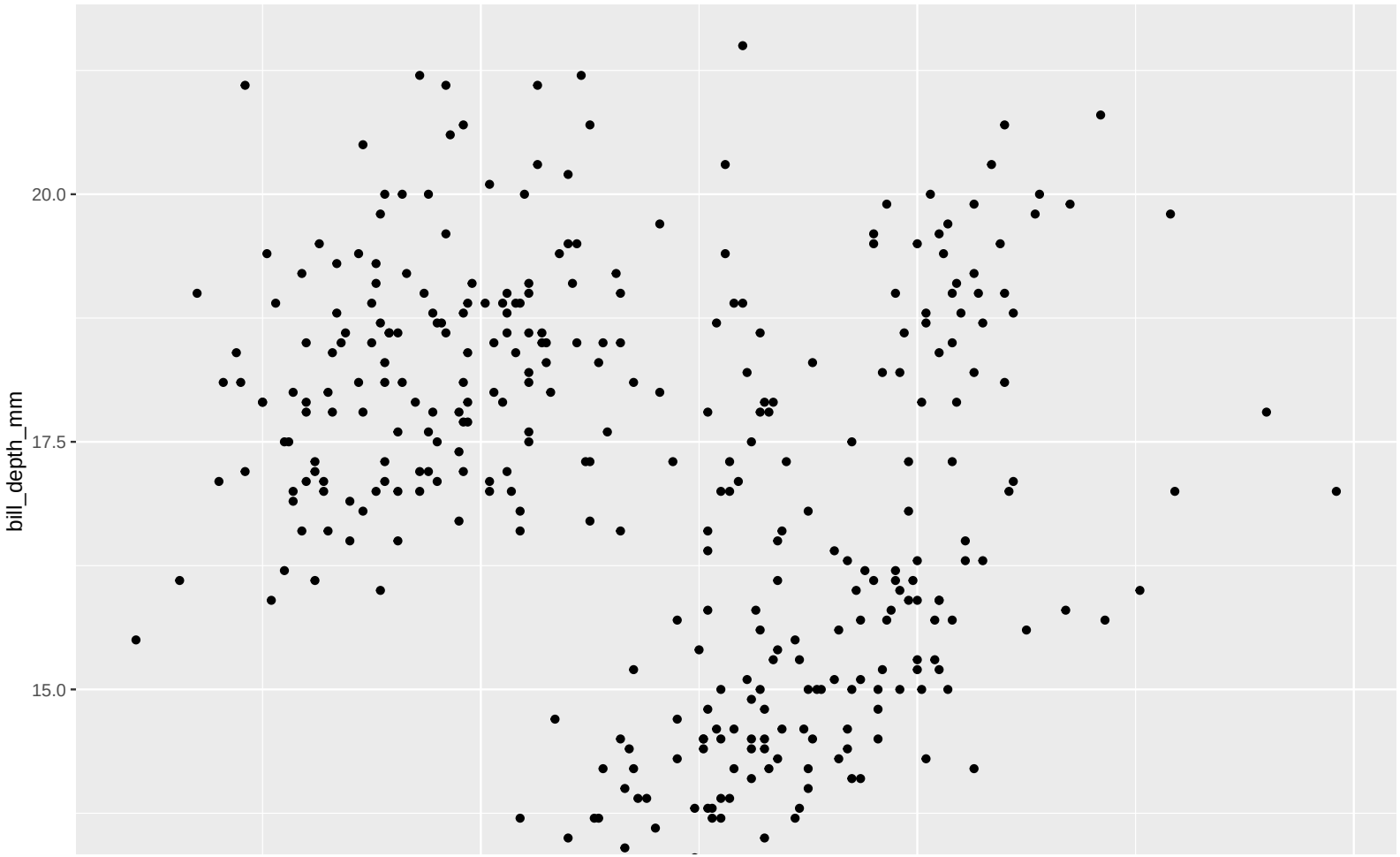
Comments

- Each dot is one (Gentoo) penguin
- for example, at the bottom left is a penguin with bill length 41 mm and bill depth 13.7 mm
- and at top left, bill length 44 and bill depth 17.3
- usually, a Gentoo penguin with greater bill length also has greater bill depth, but not always

Bill length vs. depth for all the penguins: a big mess!

```
ggplot(penguins, aes(x = bill_length_mm, y = bill_depth_mm)) +
```

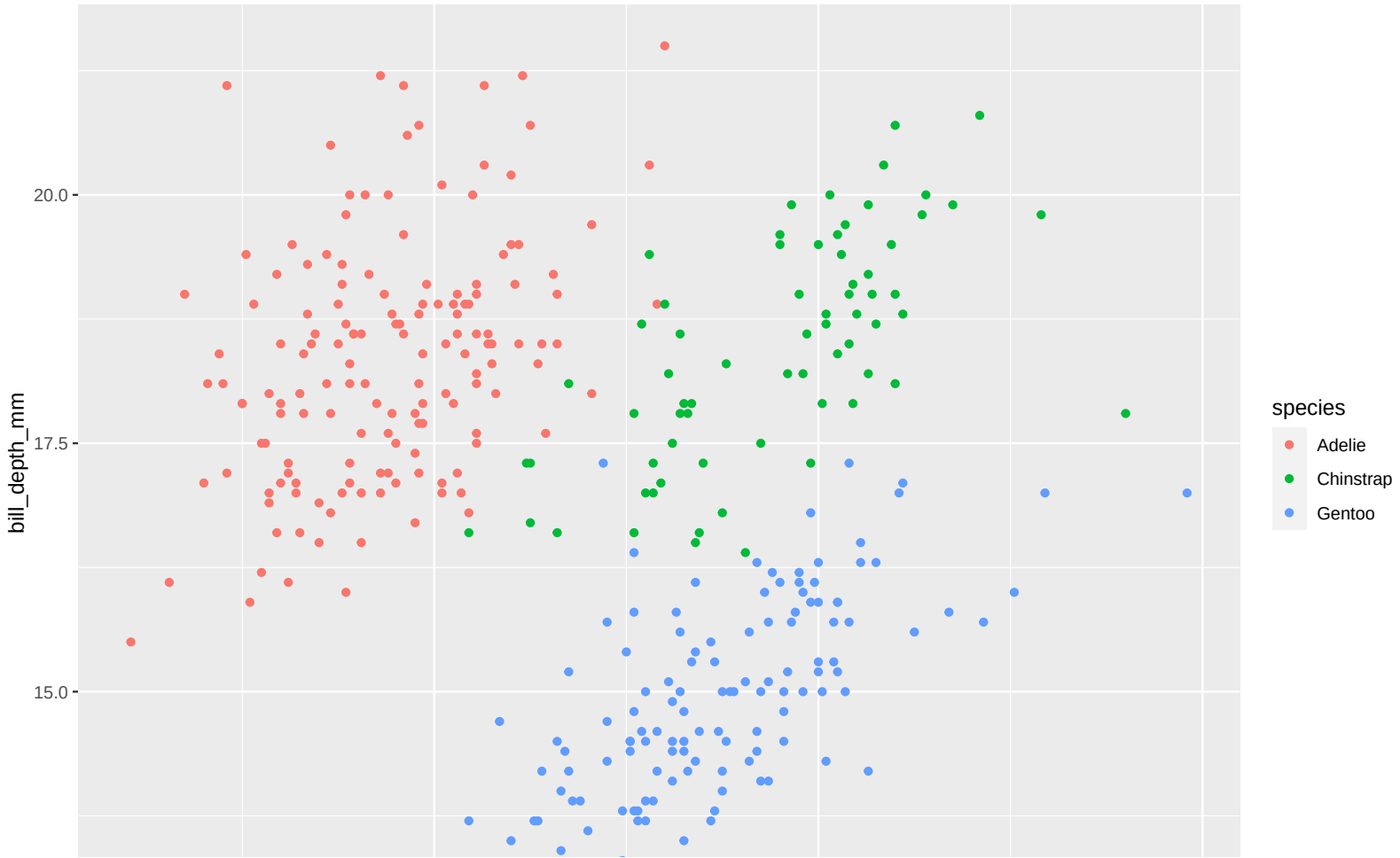
```
## Warning: Removed 2 rows containing missing values (geom_poi
```



Bill length vs. depth with species shown

```
ggplot(penguins, aes(x = bill_length_mm, y = bill_depth_mm, co
```

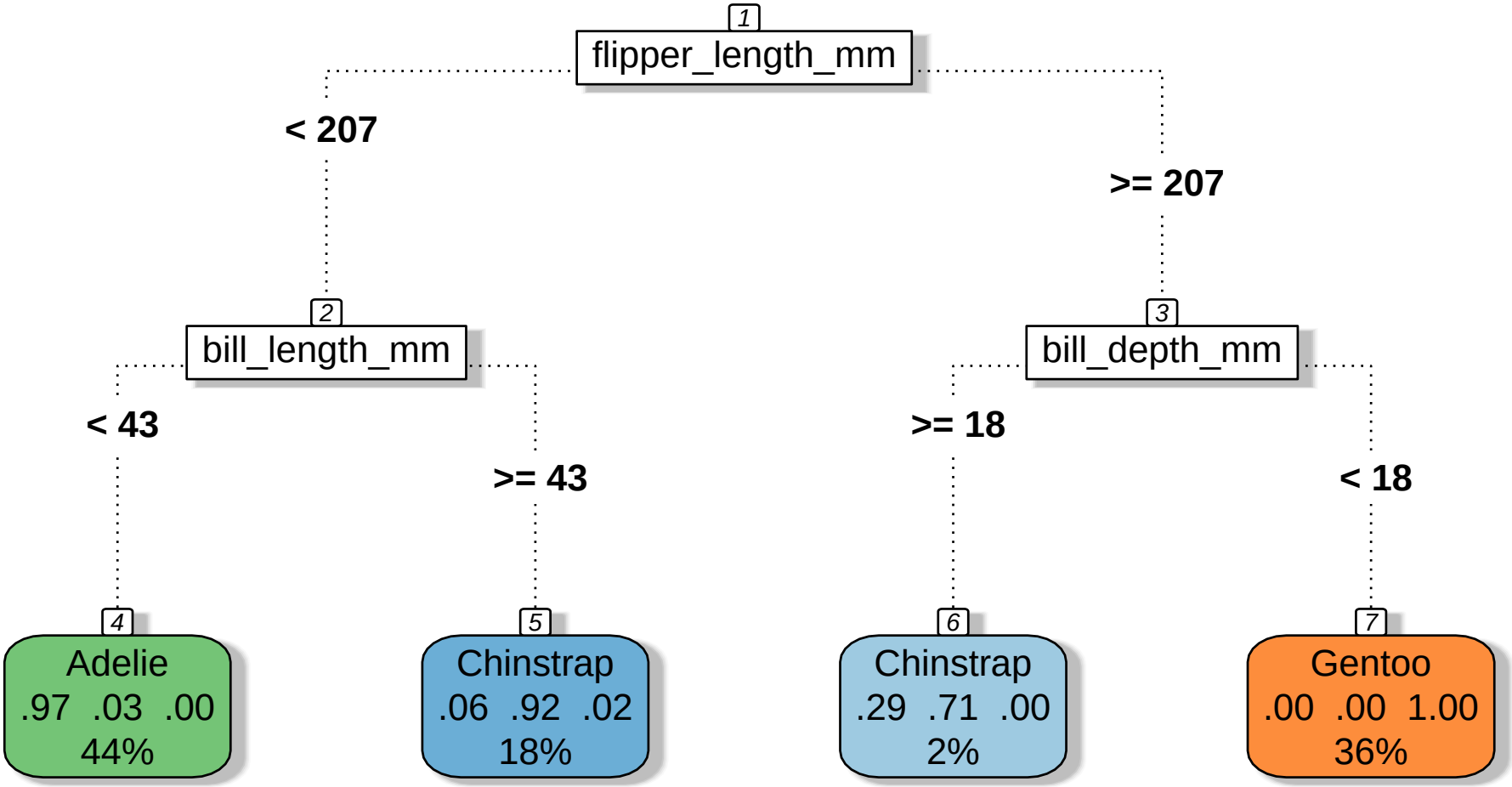
```
## Warning: Removed 2 rows containing missing values (geom_poi
```



comments

- the species are fairly distinct, eg Adelie have small bill length and large bill depth
- we know that Gentoo have large flipper length, so that plus a small bill depth identifies them
- can we do better? I ran a “classification tree”.

Classification tree



Examples:

penguin	bill_length_mm	bill_depth_mm	flipper_length_mm
50	42.3	21.2	191
202	45.2	15.8	215

- Penguin 50:
  - compare flipper length with 207. Less.
  - so compare bill length with 43. Less again.
  - so it's Adelie.
- Penguin 202:
  - compare flipper length with 207. Greater.
  - compare bill depth with 18. Less.
  - so it's Gentoo.

Were we right?

penguin	species
50	Adelie
202	Gentoo

we were!

to summarize

- variables
  - categorical: something you classify
  - quantitative: something you measure
- graphs
  - bar chart (*not* pie chart) for 1 categorical
  - density plot for 1 quantitative
  - density plots layered to compare such by groups
  - scatterplot for 2 quantitative
  - scatterplot with colours to include groups
- classification tree to understand how groups differ on quantitative variables

these also work for things other than penguins!

Extra 1: Flipper length vs bill depth

```
## Warning: Removed 2 rows containing missing values (geom_poi
```

Comments

- Look to right of solid line.
- There are a few Chinstraps with long flippers (top right).
- These are distinguished from Gentoos by bill depth: above dashed line is Chinstrap, below is Gentoo.
- A few Adelie (red) will be guessed wrong.

Extra 2: Flipper length vs. bill length

```
## Warning: Removed 2 rows containing missing values (geom_poi
```

Comments

- Look to left of solid line.
- The Adelie and Chinstrap have short flippers (left of solid line).
- To tell them apart, look at bill length: Chinstrap longer, Adelie shorter.
- A few Adelie and Chinstrap will get mistaken for each other.