Assignment 2: Visualization by example

Minsu Kang and Minsung Kim

2022-06-29

Exercise 1

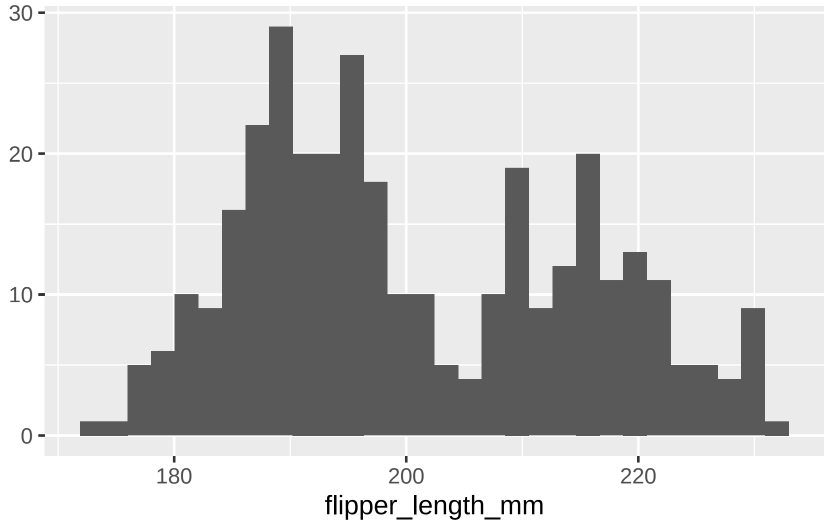
How many rows and columns? 344 rows(observations) and 8 columns(variables) * What is recorded in the observations? species, island, bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g
* What are the 3 categorical variables? spcies, island and sex * What are the 4 continuous variables? bil_length_mm, bill_depth_mm, flipper_length_mm, and body_mass_g * Which variable could be continuous or categorical? year * What are the 3 species of penguin included in the dataset? Adelie, Chinstrap, Gentoo

Exercise 2

```
qplot(x=flipper_length_mm, data=penguins)
```

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

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

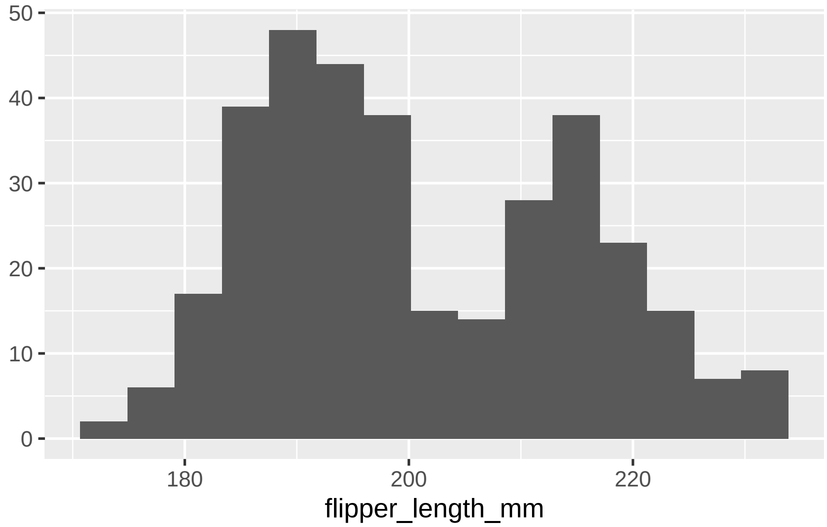


- Which axis measures flipper length? x-axis
- What do the numbers on the outer axis represent? It represents the number of penguins of each flipper lengths
- What is the modality of the flipper length variable? It has binominal modality.

Exercise 3

```
qplot(x=flipper_length_mm, data=penguins, bins=15)
```

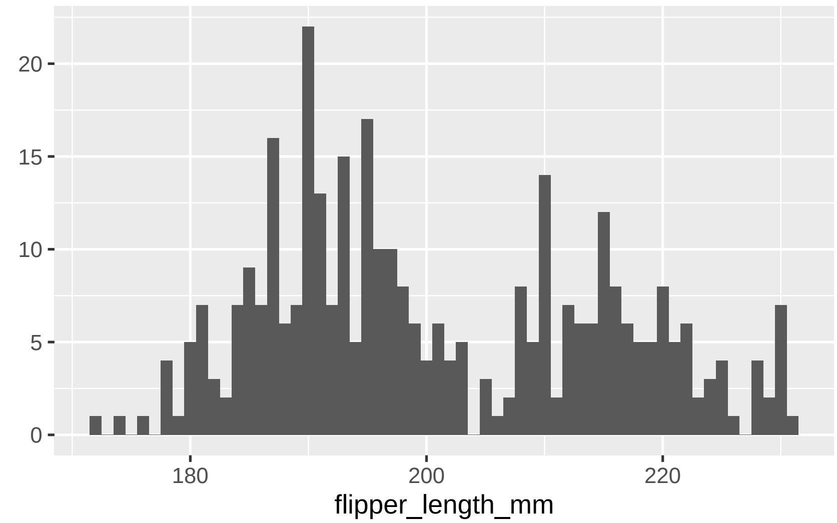
```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```



- Does the distribution in the histogram look more or less noisy than in exercise2? 15 bins histogram has less noise. More simplified and dulled by dividing the whole data sets with less bins.

```
qplot(x=flipper_length_mm, data=penguins, binwidth=1)
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

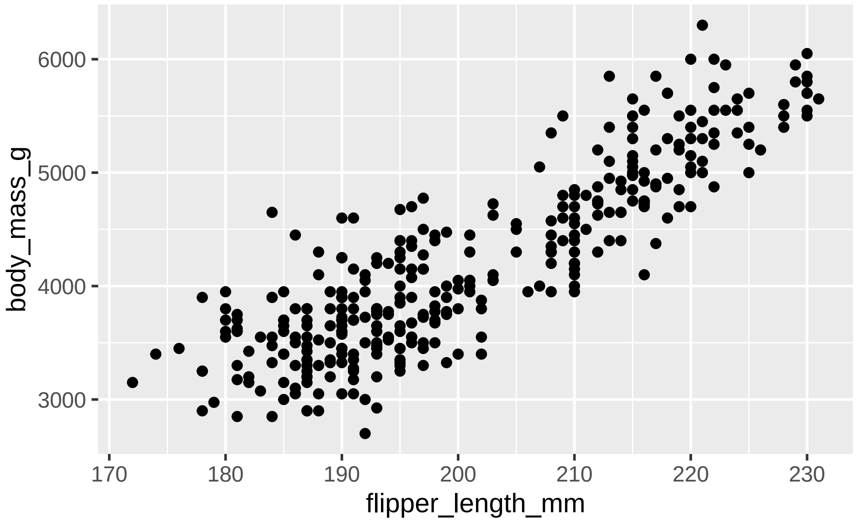


- Are there more or fewer bins in the new histogram? Yes, default was 30, and now it's 15.
- Does this increase or decrease the smotthness of the distribution? It increase smoothness of the histogram.
- Does the make the pattern of 2 peaks easier or harder to visualize? It makes to compare the patterns of two peaks easier, because it reduce other noises.

Exercise 4

```
qplot(x=flipper_length_mm, y=body_mass_g, data=penguins)
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

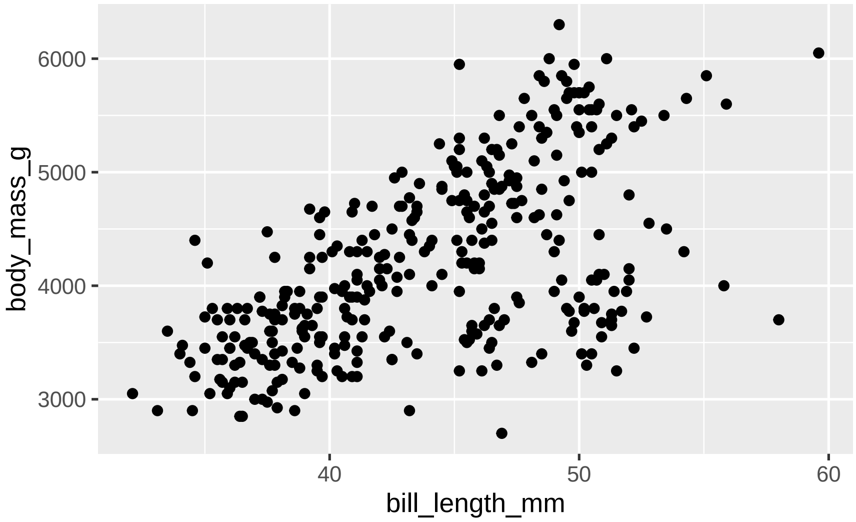


- What variable is no the y-axis? The body mass of the penguins is in the y-axis.
- Is there a relationship between these variables? Is it linear or non-linear? There is linear relationship between two variables.

Exercise 5

```
qplot(x=bill_length_mm, y=body_mass_g, data=penguins)
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

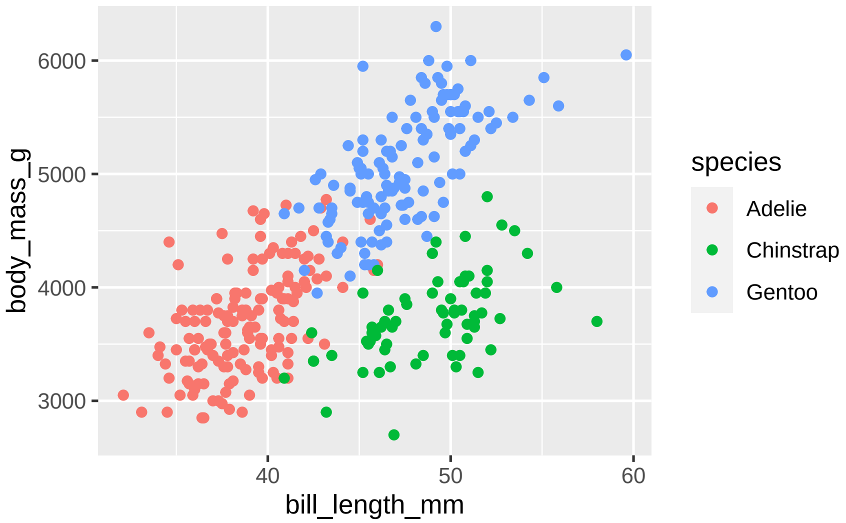


- Does the correlation between bill length and body mass look stronger or weaker than the relationship between the variables in exercise #04? It is weaker to find relationship. It is more scattered and spread than excercise 04 graph.

Exercise 6

```
qplot(x=bill_length_mm, y=body_mass_g, color=species, data=penguins)
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



- How does coloring the data points by species of penguin make the relationship between bill length and body mass easier to understand? Is any pattern more obvious than before? Compare to exercise5, it is more easy to understand cause by adding colour, we can specify the species, and each spicies have relationship between bodymass and bill lngth. Longer bill length has longer body mass.