Assignment 2: Visualization by example

Jordan Hutcherson

2024-01-24

Exercise 1

- The data set has 344 rows and 8 columns.
- The unit of observation is a penguin.
- Three categorical variables in the data set are species, island and sex.
- Four continuous variables are bill length, bill depth, flipper length and body mass.
- The variable in the dataset that could be treated as either continuous or categorical would be year.
- Three species of the penguin in the data set are Adelie, Gentoo, Chinstrap.

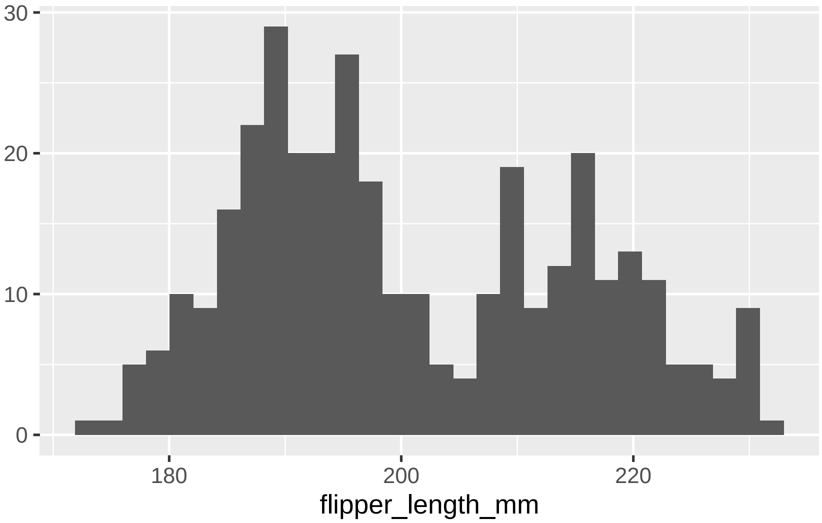
Exercise 2

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

```
## Warning: ‘qplot()‘ was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

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

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

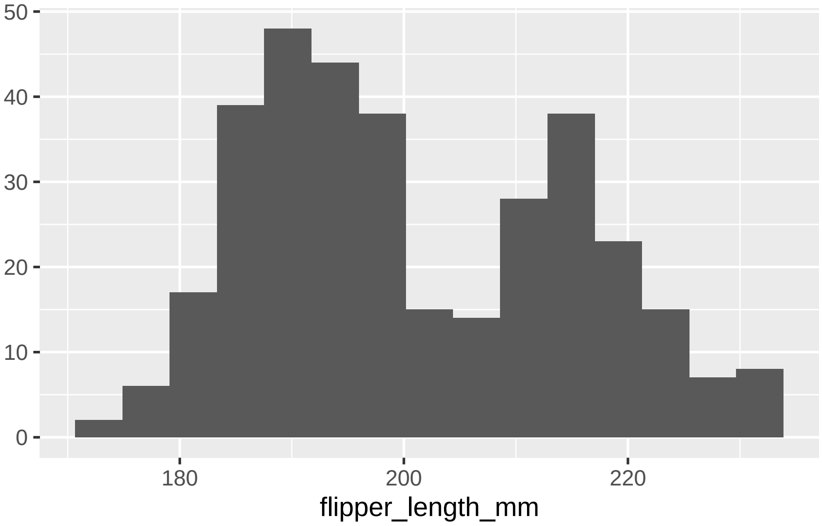


- The flipper length variable been plotted on the horizontal axis.
- The numbers on the y-axis of the graph represent the frequency count of the penguins. Each bar's height corresponds to the count of penguins whose flipper length falls within the bin's range.
- The modality of the distribution of flipper lengths refers to the number of peaks in the histogram.

Exercise 3

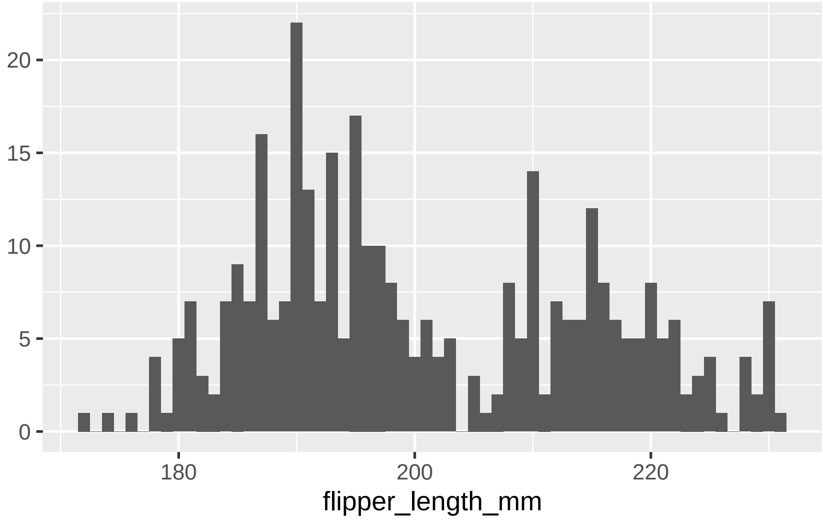
```
qplot(x = flipper_length_mm, bins = 15, data = penguins)
```

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



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

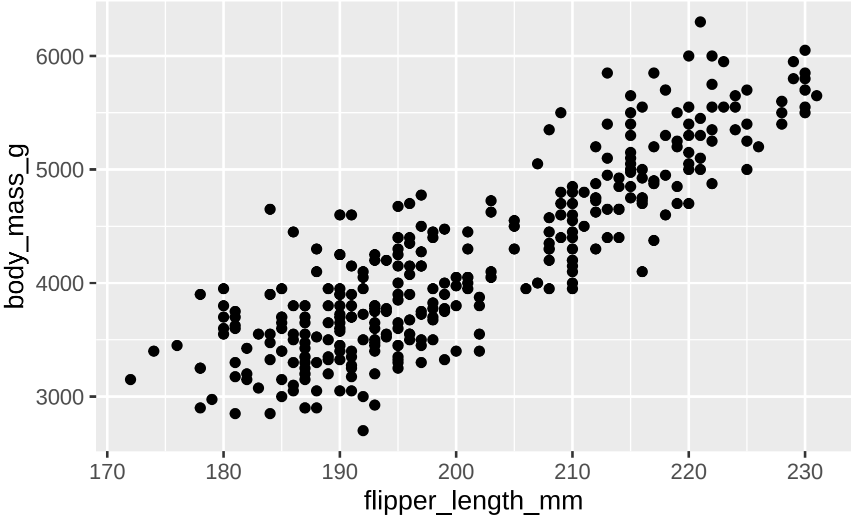
```
## Warning: Removed 2 rows containing non-finite values (`stat_bin()`).
```



Exercise 4

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

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

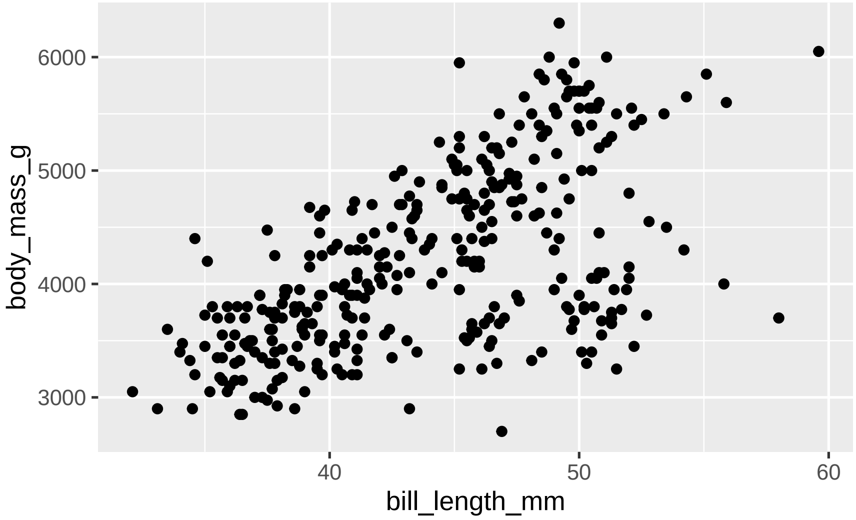


- The variable on the y-axis is body mass
- There is a relationship between the two variables and the relationship looks to be linear because the points seem to form a pattern that could be approximated by a straight line.

Exercise 5

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

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



- The correlation between bill length and body mass in this scatter plot appears weaker.

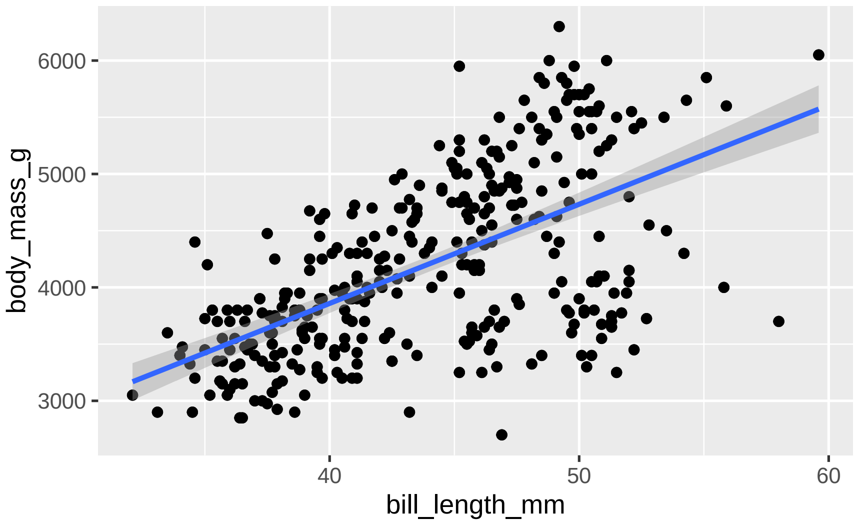
Exercise 6

```
qplot(x = bill_length_mm, y = body_mass_g, data = penguins) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```

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

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



- I think showing the patterns in the graph as this one provides a complete picture of the data. including the variability and any potential outliers. It allows the viewer to see the distribution and density of the data points.

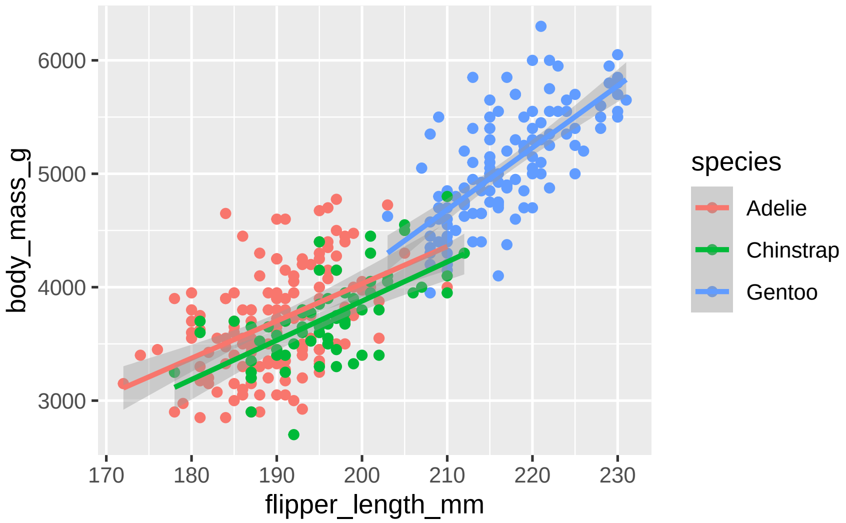
Exercise 7

```
ggplot(data = penguins, aes(x = flipper_length_mm, y = body_mass_g, color = species)) +
  geom_point() +
  geom_smooth(method = "lm")
```

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

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

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



- coloring the data points helps us understand the relationship which gives us a clear view of the data along with distinguishing patterns that might not be obvious when all points are the same color. It will allow us to see if there are distinct trends or clusters associated with each species.