Data exploration using the palmerpenguins dataset

Olivier Gimenez

10/25/2020

Data exploration

Motivation

In this section, we explore the data from package palmerpenguins. A recent publication from the researcher, Dr Kristen Gorman, who shared the data is Connors et al. (2020).

Data

The data are displayed below (first 10 rows):

```r
penguins %>%
  slice(1:10) %>%
  knitr::kable()
```

<table>
<thead>
<tr>
<th>species</th>
<th>island</th>
<th>bill_length_mm</th>
<th>bill_depth_mm</th>
<th>flipper_length_mm</th>
<th>body_mass_g</th>
<th>sex</th>
<th>year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adelie</td>
<td>Torgersen</td>
<td>39.1</td>
<td>18.7</td>
<td>181</td>
<td>3750</td>
<td>male</td>
<td>2007</td>
</tr>
<tr>
<td>Adelie</td>
<td>Torgersen</td>
<td>39.5</td>
<td>17.4</td>
<td>186</td>
<td>3800</td>
<td>female</td>
<td>2007</td>
</tr>
<tr>
<td>Adelie</td>
<td>Torgersen</td>
<td>40.3</td>
<td>18.0</td>
<td>195</td>
<td>3250</td>
<td>female</td>
<td>2007</td>
</tr>
<tr>
<td>Adelie</td>
<td>Torgersen</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>2007</td>
</tr>
<tr>
<td>Adelie</td>
<td>Torgersen</td>
<td>36.7</td>
<td>19.3</td>
<td>193</td>
<td>3450</td>
<td>female</td>
<td>2007</td>
</tr>
<tr>
<td>Adelie</td>
<td>Torgersen</td>
<td>39.3</td>
<td>20.6</td>
<td>190</td>
<td>3650</td>
<td>male</td>
<td>2007</td>
</tr>
<tr>
<td>Adelie</td>
<td>Torgersen</td>
<td>38.9</td>
<td>17.8</td>
<td>181</td>
<td>3625</td>
<td>female</td>
<td>2007</td>
</tr>
<tr>
<td>Adelie</td>
<td>Torgersen</td>
<td>39.2</td>
<td>19.6</td>
<td>195</td>
<td>4675</td>
<td>male</td>
<td>2007</td>
</tr>
<tr>
<td>Adelie</td>
<td>Torgersen</td>
<td>34.1</td>
<td>18.1</td>
<td>193</td>
<td>3475</td>
<td>NA</td>
<td>2007</td>
</tr>
<tr>
<td>Adelie</td>
<td>Torgersen</td>
<td>42.0</td>
<td>20.2</td>
<td>190</td>
<td>4250</td>
<td>NA</td>
<td>2007</td>
</tr>
</tbody>
</table>

Numerical exploration

There are 344 penguins in the dataset, and 3 different species. The data were collected in 3 islands of the Palmer archipelago in Antarctica.

The mean of all traits that were measured on the penguins are:

```r
## A tibble: 3 x 6
##  species bill_length_mm bill_depth_mm flipper_length_mm body_mass_g year
##  <fct>       <dbl>        <dbl>             <dbl>          <dbl> <dbl>
## 1 Adelie      39.1         18.7              181            3750  2007
## 2 Adelie      39.5         17.4              186            3800  2007
## 3 Adelie      40.3         18.0              195            3250  2007
```
Graphical exploration

A histogram of body mass per species:

```r
penguins %>%
ggplot() +
aes(x = body_mass_g) +
geom_histogram(aes(fill = species),
               alpha = 0.5,
               position = "identity") +
scale_fill_manual(values = c("darkorange","purple","cyan4")) +
theme_minimal() +
labs(x = "Body mass (g)",
     y = "Frequency",
     title = "Penguin body mass")
```
References