Palmer Penguins Notebook

Drew Croteau

Purpose A simple *progression-based* analysis of penguin data.

Creating the R environment:

Here we'll install and load the necessary R packages.

Package installation: **Note**: If receiving **error(s)** then the packages are most likely *already* installed.

```r
install.packages('tidyverse')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
install.packages('palmerpenguins')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
install.packages('ggplot2')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r

library('tidyverse')
```

Load packages:

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library('palmerpenguins')
library('ggplot2')
```

Load and preview the dataset: **Note**: The 'penguins' dataset is pulled from the 'palmerpenguins' package.

```r
data(penguins)
head(penguins)
```

```
## # A tibble: 6 x 8
##   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
## 1 Adelie  Torgersen           39.1          18.7               181        3750
## 2 Adelie  Torgersen           39.5          17.4               186        3800
## 3 Adelie  Torgersen           40.3          18                 195        3250
## 4 Adelie  Torgersen           NA            NA                  NA          NA
## 5 Adelie  Torgersen           36.7          19.3               193        3450
## 6 Adelie  Torgersen           39.3          20.6               190        3650
## # i 2 more variables: sex <fct>, year <int>
```

Cleaning the data:

In the preview of the ‘penguins’ dataset we can see ‘NA’ values which require cleaning before manipulating the data and creating visuals.

Creating a new table: First we’ll start by creating a new table and filter the data from ‘penguins’ to only pull values that are **NOT** ‘NA’.

```r
clean_penguins = filter(penguins, penguins$body_mass_g != 'NA')
```

View the new table: Let’s take a peek at the new filtered data.

```r
head(clean_penguins)
```

```
## # A tibble: 6 x 8
##   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
## 1 Adelie  Torgersen           39.1          18.7               181        3750
## 2 Adelie  Torgersen           39.5          17.4               186        3800
## 3 Adelie  Torgersen           40.3          18                 195        3250
## 4 Adelie  Torgersen           36.7          19.3               193        3450
## 5 Adelie  Torgersen           39.3          20.6               190        3650
## 6 Adelie  Torgersen           38.9          17.8               181        3625
## # i 2 more variables: sex <fct>, year <int>
```

Data manipulation

Now that we have a clean dataset, we can manipulate the data to make visualizations easier to create.

Body mass summary Here we’ll create a subset summary table of our clean dataset with a focus on body mass by species.

```r
body_mass_summary = clean_penguins %>%
  group_by(species) %>%
  summarize(avg_body_mass = mean(body_mass_g))
```

View the summary Let's view the new summary table and make sure everything is in order.

```
print(body_mass_summary)
```

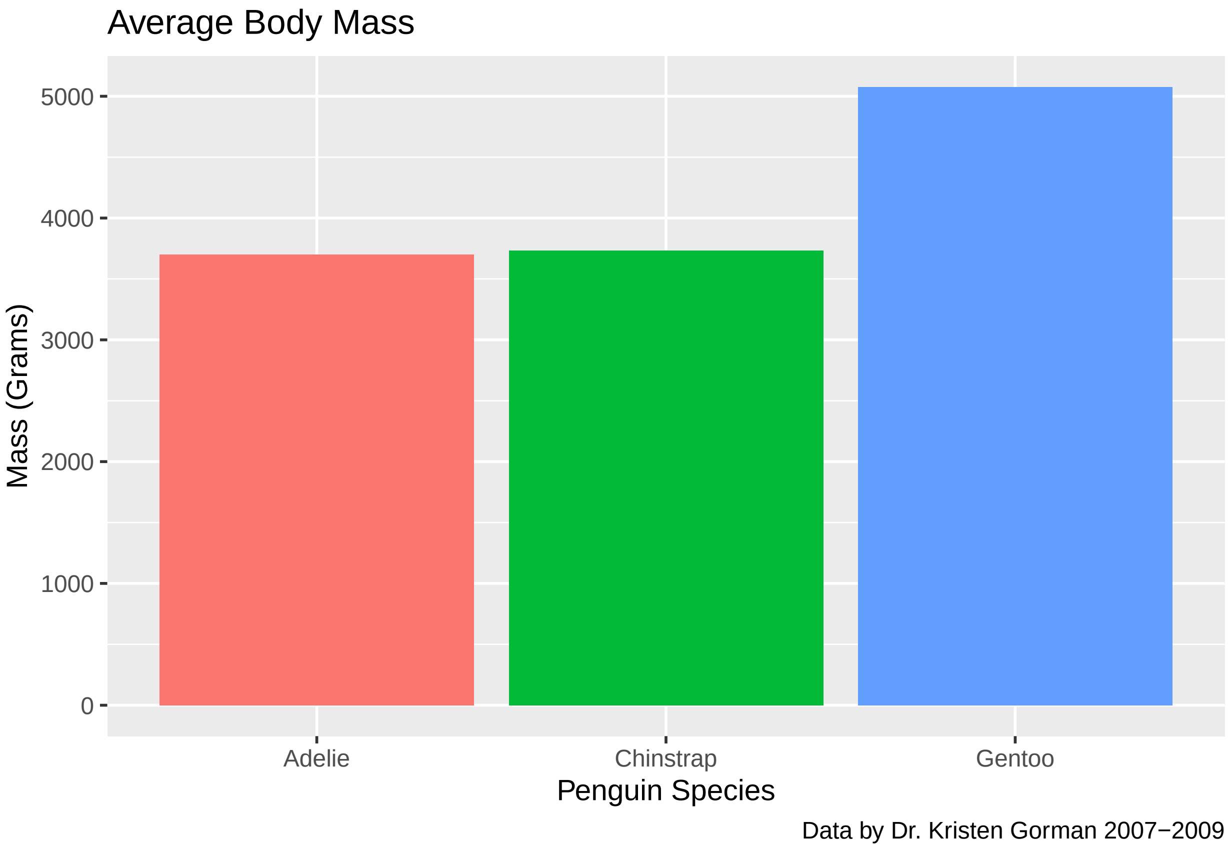
```
## # A tibble: 3 x 2
##   species   avg_body_mass
##   <fct>             <dbl>
## 1 Adelie            3701.
## 2 Chinstrap         3733.
## 3 Gentoo            5076.
```

Data visualization

Even though the summary paints a fairly clear picture already, let's make a visualization to show the differences anyway.

Average body mass by species visualization Utilizing the 'ggplot2' package, we can create an easy-to-understand visual representing the body mass by species summary. In this instance, we'll be using a simple bar chart.

```
ggplot(data = body_mass_summary) %>%
  + geom_col(mapping = aes(x = species, y = avg_body_mass, fill = species)) %>%
  ## The rest of this code is for readability and aesthetic purposes.
  + guides(fill = 'none') %>%
  + labs(title = 'Average Body Mass', x = 'Penguin Species', y = 'Mass (Grams)', caption =
         'Data by Dr. Kristen Gorman 2007-2009')
```


Average Body Mass
Mass (Grams)
Adelie
Chinstrap
Gentoo
Penguin Species
Data by Dr. Kristen Gorman 2007−2009