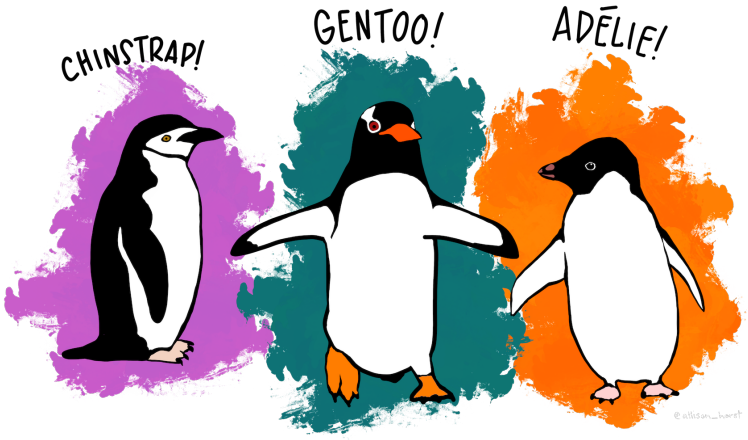
AE-01: Meet the Penguins

In this application exercise, we will meet some penguins, start thinking about data and variables, and see some R code in action.



The `penguins` data from the **palmerpenguins** package contains size measurements for three species of penguins observed on three islands in the Palmer Archipelago, Antarctica.

Data were collected and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network.

Task 1

Let's take a peek at the data with the following R code:

```
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>
```

This tables shows us the first six rows of **data frame** containing this data. When working with data, we typically want each row to be an individual observation (or case), each column to be a variable and each entry (cell) to be a single value. Data in this format is called **tidy**.

Question: What observations can you make about this table?

Task 2

Here are many other ways to view data frames.

```
print(penguins)
```

```
# A tibble: 344 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
 7 Adelie  Torgersen           38.9          17.8               181        3625
 8 Adelie  Torgersen           39.2          19.6               195        4675
 9 Adelie  Torgersen           34.1          18.1               193        3475
10 Adelie  Torgersen           42            20.2               190        4250
# i 334 more rows
# i 2 more variables: sex <fct>, year <int>
```

```
glimpse(penguins)
```

```
Rows: 344
Columns: 8
$ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel~
```

```
$ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse~
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, ~
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, ~
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186~
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, ~
$ sex               <fct> male, female, female, NA, female, male, female, male~
$ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

Question: what are the differences between these R commands?

Questions: How many penguins are included in this dataset? How many variables?

Task 3

Let's take a look at other ways to analyize our data using R. What is the code chunk below doing?

```
penguins %>%
  count(species)
```

```
# A tibble: 3 x 2
  species       n
  <fct>     <int>
1 Adelie      152
2 Chinstrap    68
3 Gentoo      124
```

Exercise:

Add a code chunk that does something similar to determine how many penguins are on each island.

Task 4

```
penguins %>%
  group_by(species) %>%
  summarize(across(where(is.numeric), ~ mean(.x, na.rm = TRUE)))
```

```
# 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              38.8          18.3               190.       3701. 2008.
2 Chinstrap           48.8          18.4               196.       3733. 2008.
3 Gentoo              47.5          15.0               217.       5076. 2008.
```

The plot below shows the relationship between flipper and bill lengths of these penguins.

```
ggplot(penguins,
       aes(x = flipper_length_mm, y = bill_length_mm)) +
  geom_point(aes(color = species, shape = species)) +
  scale_color_manual(values = c("darkorange","purple","cyan4")) +
  labs(
    title = "Flipper and bill length",
    subtitle = "Dimensions for penguins at Palmer Station LTER",
    x = "Flipper length (mm)", y = "Bill length (mm)",
    color = "Penguin species", shape = "Penguin species"
  ) +
  theme_minimal()
```