

Code Sleuth

This activity uses a data frame called “penguins”. Here’s what it looks like at the console.

```
> penguins
# A tibble: 344 × 8
  species island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex    year
  <fct>   <fct>         <dbl>         <dbl>         <int>         <int> <fct> <int>
1 Adelie  Torgersen         39.1           18.7           181           3750 male    2007
2 Adelie  Torgersen         39.5           17.4           186           3800 female  2007
3 Adelie  Torgersen         40.3            18           195           3250 female  2007
4 Adelie  Torgersen          NA            NA            NA           NA NA      2007
5 Adelie  Torgersen         36.7           19.3           193           3450 female  2007
6 Adelie  Torgersen         39.3           20.6           190           3650 male    2007
7 Adelie  Torgersen         38.9           17.8           181           3625 female  2007
8 Adelie  Torgersen         39.2           19.6           195           4675 male    2007
9 Adelie  Torgersen         34.1           18.1           193           3475 NA      2007
10 Adelie Torgersen         42            20.2           190           4250 NA      2007
# i 334 more rows
```

Here’s what the data frame looks like using the “glimpse()” function.

```
> glimpse(penguins)
Rows: 344
Columns: 8
$ species      <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adelie,
Adelie, Adelie, ...
$ island       <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Torgersen,
Torgersen, Tor...
$ bill_length_mm <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, 42.0, 37.8,
37.8, 41.1, 38.6...
$ bill_depth_mm <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, 20.2, 17.1,
17.3, 17.6, 21.2...
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186, 180, 182,
191, 198, 185, 1...
$ body_mass_g   <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, 4250, 3300,
3700, 3200, 3800...
$ sex          <fct> male, female, female, NA, female, male, female, male, NA, NA, NA,
NA, female, ma...
$ year         <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007,
2007, 2007, 20...
```

What does the “glimpse()” function do?

Working with rows: filter()

Note: code outputs have been truncated for clarity.

Inference

Code example	Output
<code>filter(penguins, species == "Chinstrap")</code>	<pre>species island bill_length_mm <fct> <fct> <dbl> 1 Chinstrap Dream 46.5 2 Chinstrap Dream 50 3 Chinstrap Dream 51.3 4 Chinstrap Dream 45.4 5 Chinstrap Dream 52.7</pre>
<code>filter(penguins, body_mass_g < 3000)</code>	<pre>species island body_mass_g <fct> <fct> <int> 1 Adelie Dream 2975 2 Adelie Biscoe 2900 3 Adelie Biscoe 2850 4 Adelie Biscoe 2850 5 Adelie Dream 2900</pre>
<code>filter(penguins, sex == "female" flipper_length_mm > 190)</code>	<pre>species flipper_length_mm sex <fct> <int> <fct> 1 Adelie 186 female 2 Adelie 195 female 3 Adelie 193 female 4 Adelie 181 female 5 Adelie 195 male</pre>

What does the “filter()” function do?

What do you think the difference between | and || is? Recall we used || in conditional statements last week.

Prediction

What will be the output of the following expressions? Describe your prediction in one sentence (don't write out the whole data frame).

Code example	Output
<code>filter(penguins, species == "Chinstrap")</code>	
<code>filter(penguins, flipper_length_mm > 229)</code>	
<code>filter(penguins, sex == "female" & flipper_length_mm > 190)</code>	

Working with rows: arrange()

Note: code outputs have been truncated for clarity.

Inference

Code example	Output
<code>arrange(penguins, bill_depth_mm)</code>	<pre>species island bill_depth_mm <fct> <fct> <dbl> 1 Gentoo Biscoe 13.1 2 Gentoo Biscoe 13.2 3 Gentoo Biscoe 13.3 4 Gentoo Biscoe 13.4 5 Gentoo Biscoe 13.5</pre>
<code>arrange(penguins, desc(bill_depth_mm))</code>	<pre>species island bill_depth_mm <fct> <fct> <dbl> 1 Adelie Torgersen 21.5 2 Adelie Torgersen 21.2 3 Adelie Dream 21.2 4 Adelie Torgersen 21.1 5 Adelie Dream 21.1</pre>

What does the “arrange()” function do?

Prediction

What will be the output of the following expressions? Describe your prediction in one sentence (don't write out the whole data frame).

Code example	Output
<code>arrange(penguins, body_mass_g)</code>	
<code>arrange(penguins, desc(body_mass_g))</code>	

Working with columns: mutate()

Note: code outputs have been truncated for clarity.

Inference

Code example	Output
<pre>mutate(penguins, body_mass_kg = body_mass_g / 1000)</pre>	<pre>species body_mass_g body_mass_kg <fct> <int> <dbl> 1 Adelie 3750 3.75 2 Adelie 3800 3.8 3 Adelie 3250 3.25 4 Adelie NA NA 5 Adelie 3450 3.45</pre>
<pre>mutate(penguins, body_mass_kg = body_mass_g / 1000, bill_length_cm = bill_length_mm / 10)</pre>	<pre>species body_mass_kg bill_length_cm <fct> <dbl> <dbl> 1 Adelie 3.75 3.91 2 Adelie 3.8 3.95 3 Adelie 3.25 4.03 4 Adelie NA NA 5 Adelie 3.45 3.67</pre>

What does the “mutate()” function do?

What does the *parameter name* in mutate() calls become in the result?

What does the *argument* in mutate() calls become in the result?

Prediction

What will be the output of the following expressions? Describe your prediction in one sentence (don't write out the whole data frame).

Code example	Output
<pre>mutate(penguins, bill_length_cm = bill_length_mm / 10)</pre>	
<pre>mutate(penguins, bill_ratio = bill_length_mm / bill_depth_mm)</pre>	
<pre>mutate(penguins, body_mass_kg = body_mass_g / 1000, bill_ratio = bill_length_mm / bill_depth_mm,)</pre>	

Working with columns: select()

Note: code outputs have been truncated for clarity.

Inference

Code example	Output
<code>select(penguins, island)</code>	island <fct> 1 Torgersen 2 Torgersen 3 Torgersen
<code>select(penguins, species:bill_length_mm)</code>	species island bill_length_mm <fct> <fct> <dbl> 1 Adelie Torgersen 39.1 2 Adelie Torgersen 39.5 3 Adelie Torgersen 40.3
<code>select(penguins, !island)</code>	species bill_length_mm bill_depth_mm <fct> <dbl> <dbl> 1 Adelie 39.1 18.7 2 Adelie 39.5 17.4 3 Adelie 40.3 18
<code>select(penguins, starts_with("bill"))</code>	bill_length_mm bill_depth_mm <dbl> <dbl> 1 39.1 18.7 2 39.5 17.4 3 40.3 18

What does the “select()” function do?

What does the sequence operator : do in select()?

What does the not operator ! do in select()?

What do you think the “starts_with()” function does?

Prediction

What will be the output of the following expressions? Describe your prediction in one sentence (don't write out the whole data frame).

Code example	Output
<code>select(penguins, species, body_mass_g)</code>	
<code>select(penguins, flipper_length_mm:year)</code>	
<code>select(penguins, !species)</code>	
<code>select(penguins, ends_with("g"))</code>	

The pipe

Note: code outputs have been truncated for clarity.

Inference

Code example	Output
<pre>filter(mutate(penguins, body_mass_kg = body_mass_g / 1000), body_mass_kg < 2.8)</pre>	<pre>species island body_mass_kg <fct> <fct> <dbl> 1 Chinstrap Dream 2.7</pre>
<pre>penguins > mutate(body_mass_kg = body_mass_g / 1000) > filter(body_mass_kg < 2.8)</pre>	<pre>species island body_mass_kg <fct> <fct> <dbl> 1 Chinstrap Dream 2.7</pre>
<pre>select(arrange(penguins, species, bill_length_mm), species:bill_length_mm)</pre>	<pre>species island bill_length_mm <fct> <fct> <dbl> 1 Adelie Dream 32.1 2 Adelie Dream 33.1 3 Adelie Torgersen 33.5</pre>
<pre>penguins > arrange(species, bill_length_mm) > select(species:bill_length_mm)</pre>	<pre>species island bill_length_mm <fct> <fct> <dbl> 1 Adelie Dream 32.1 2 Adelie Dream 33.1 3 Adelie Torgersen 33.5</pre>

How does the order of the function calls change between the 1st and 2nd example, and between the 3rd and 4th example?

Describe what you think `|>` does.

Prediction

What will be the output of the following expressions? Describe your prediction in one sentence (don't write out the whole data frame).

Code example	Output
<pre>penguins > select(species:island, sex:year) > mutate(year2 = year - 2000, sex = substr(sex, start = 1, stop = 1))</pre>	
<pre>penguins > filter(bill_length_mm <= 34 bill_depth_mm <= 13.5) > mutate(bill_ratio = bill_depth_mm / bill_length_mm)</pre>	

Summarizing groups

Inference

Code example	Output
<pre>penguins > summarize(avg_size = mean(body_mass_g))</pre>	<pre>avg_size <dbl> 1 NA</pre>
<pre>penguins > summarize(avg_size = mean(body_mass_g, na.rm = TRUE))</pre>	<pre>avg_size <dbl> 1 4202.</pre>
<pre>penguins > summarize(biggest_bill = max(bill_length_mm, na.rm = TRUE), .by = island)</pre>	<pre>island biggest_bill <fct> <dbl> 1 Torgersen 46 2 Biscoe 59.6 3 Dream 58</pre>
<pre>penguins > summarize(avg_size = mean(body_mass_g, na.rm = TRUE), biggest_bill = max(bill_length_mm, na.rm = TRUE), .by = species)</pre>	<pre>species avg_size biggest_bill <fct> <dbl> <dbl> 1 Adelie 3701. 46 2 Gentoo 5076. 59.6 3 Chinstrap 3733. 58</pre>

What do you think “NA” represents? What do you think the “na.rm” parameter in mean, max, and other summary functions might do?

How many rows does “summarize()” return? How does the “.by” parameter affect number of rows returned?

Prediction

What will be the output of the following expressions? Describe your prediction in one sentence (don't write out the whole data frame).

Code example	Output
<pre>penguins > summarize(flipper_mean = mean(flipper_length_mm, na.rm = TRUE), flipper_sd = sd(flipper_length_mm, na.rm = TRUE))</pre>	
<pre>penguins > summarize(flipper_mean = mean(flipper_length_mm, na.rm = TRUE), flipper_sd = sd(flipper_length_mm, na.rm = TRUE), .by = sex)</pre>	