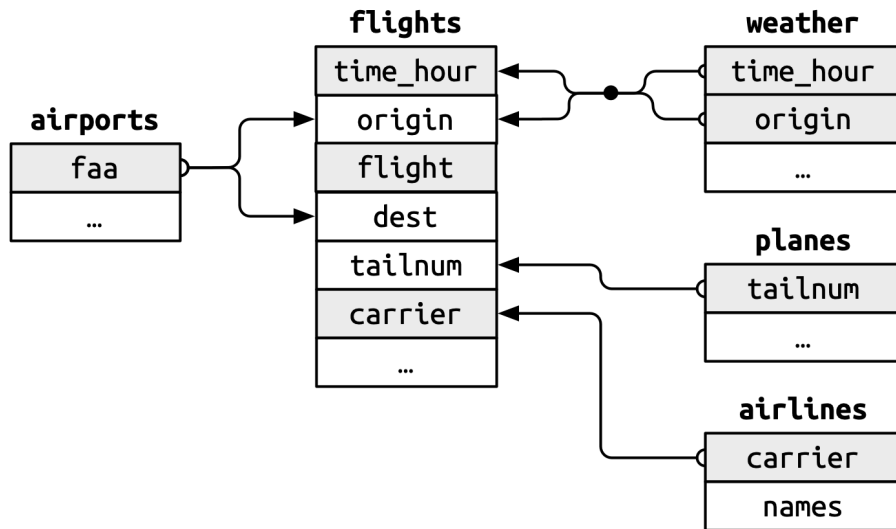


Code Sleuth

Relational data

Inference



The figure above shows a *relational schema diagram*. You'll learn about these in more detail later in EDS 213. Each name in bold is a data frame, and below it is a list of its columns. The arrows represent *keys*.

What do you think is the difference between a *primary key* and a *foreign key*?

In the diagram above, what do you think visually indicates a primary key? And what indicates a foreign key?

Topic: Joining data

Prediction

```

surveys <- read_csv("data/surveys.csv")
glimpse(surveys)
Rows: 35,549
Columns: 9
$ record_id      <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 1...
$ month          <dbl> 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7,...
$ day            <dbl> 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, ...
$ year           <dbl> 1977, 1977, 1977, 1977, 1977, 1977, 1977, 19...
$ plot_id        <dbl> 2, 3, 2, 7, 3, 1, 2, 1, 1, 6, 5, 7, 3, 8, 6,...
$ species_id     <chr> "NL", "NL", "DM", "DM", "DM", "PF", "PE", "D...
$ sex            <chr> "M", "M", "F", "M", "M", "M", "F", "M", "F", ...
$ hindfoot_length <dbl> 32, 33, 37, 36, 35, 14, NA, 37, 34, 20, 53, ...
$ weight         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...

species <- read_csv("data/species.csv")
glimpse(species)
Rows: 54
Columns: 4
$ species_id <chr> "AB", "AH", "AS", "BA", "CB", "CM", "CQ", "CS", "...
$ genus      <chr> "Amphispiza", "Ammospermophilus", "Ammodramus", "...
$ species    <chr> "bilineata", "harrisi", "savannarum", "taylori", ...
$ taxa       <chr> "Bird", "Rodent", "Bird", "Rodent", "Bird", "Bird...

plots <- read_csv("data/plots.csv")
glimpse(plots)
Rows: 24
Columns: 2
$ plot_id    <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,...
$ plot_type  <chr> "Spectab exclosure", "Control", "Long-term Krat Ex..."

```

Draw a *relational schema diagram* for the three tables above. Try to identify the primary and foreign keys first.

Keys

Inference

Code example	Output
<pre>surveys > count(record_id) > filter(n > 1)</pre>	<pre># A tibble: 0 × 2 # i 2 variables: record_id <dbl>, n <int></pre>
<pre>surveys > count(species_id) > filter(n > 1)</pre>	<pre># A tibble: 43 × 2 species_id n <chr> <int> 1 AB 303 2 AH 437 3 AS 2 4 BA 46 5 CB 50 6 CM 13 7 CQ 16 8 DM 10596 9 DO 3027 10 DS 2504 # i 33 more rows # i Use `print(n = ...)` to see more rows</pre>
<pre>species > count(species_id) > filter(n > 1)</pre>	<pre># A tibble: 0 × 2 # i 2 variables: species_id <chr>, n <int></pre>

What do the pipelines in the three code examples each do?

What does it mean if the output has no rows? What does it mean if the result has one or more rows?

How does the number of rows in the output relate to *primary* and *foreign keys*?

Prediction

Will the output for the following code examples have 0 rows or 1+ rows?

Code example	Output
<pre>surveys > count(plot_id) > filter(n > 1)</pre>	
<pre>plots > count(plot_id) > filter(n > 1)</pre>	
<pre>plots > count(plot_type) > filter(n > 1)</pre>	

Joins

Note: code outputs have been truncated for clarity.

Inference

Code example	Output
<pre>surveys > inner_join(species, by = join_by(survey_id)) > glimpse()</pre>	<pre>Error in `inner_join()`: ! Join columns in `x` must be present in the data. ✖ Problem with `survey_id`.</pre>
<pre>surveys > inner_join(species, by = join_by(species_id)) > glimpse()</pre>	<pre>Rows: 34,786 Columns: 12 \$ record_id <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, ... \$ month <dbl> 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, ... \$ day <dbl> 16, 16, 16, 16, 16, 16, 16, 16, ... \$ year <dbl> 1977, 1977, 1977, 1977, 1977, ... \$ plot_id <dbl> 2, 3, 2, 7, 3, 1, 2, 1, 1, 6, ... \$ species_id <chr> "NL", "NL", "DM", "DM", "DM", ... \$ sex <chr> "M", "M", "F", "M", "M", "M", ... \$ hindfoot_length <dbl> 32, 33, 37, 36, 35, 14, NA, 37... \$ weight <dbl> NA, NA, NA, NA, NA, NA, NA, NA... \$ genus <chr> "Neotoma", "Neotoma", "Dipodom... \$ species <chr> "albigula", "albigula", "merri... \$ taxa <chr> "Rodent", "Rodent", "Rodent", ...</pre>
<pre>birds <- species > filter(taxa == "Bird") surveys > inner_join(birds, by = join_by(species_id)) > glimpse()</pre>	<pre>Rows: 450 Columns: 12 \$ record_id <dbl> 3126, 3146, 3152, 3153, 3189, ... \$ month <dbl> 7, 7, 7, 7, 8, 8, 8, 8, 9, 9, ... \$ day <dbl> 21, 21, 21, 21, 13, 13, 13, 13... \$ year <dbl> 1980, 1980, 1980, 1980, 1980, ... \$ plot_id <dbl> 8, 24, 19, 22, 23, 9, 19, 8, 2... \$ species_id <chr> "AB", "AB", "AB", "AB", "CB", ... \$ sex <chr> NA, NA, NA, NA, NA, NA, NA, NA... \$ hindfoot_length <dbl> NA, NA, NA, NA, NA, NA, NA, NA... \$ weight <dbl> NA, NA, NA, NA, NA, NA, NA, NA... \$ genus <chr> "Amphispiza", "Amphispiza", "A... \$ species <chr> "bilineata", "bilineata", "bil... \$ taxa <chr> "Bird", "Bird", "Bird", "Bird"...</pre>

Why does the first code example throw an error, while the second one returns a valid result?

Topic: *Joining data*

What columns are in the second and third output? What columns are in the two data frames in the join?

Compare the number of rows in “surveys”, the second output, and the third output. Why do you think they have different rows?

Prediction

One of the following throws an error, the other returns a data frame. Indicate which one throws an error and list the columns in the output of the other.

Code example	Output
<pre>species > inner_join(plots, by = join_by(plot_id))</pre>	
<pre>surveys > inner_join(plots, by = join_by(plot_id))</pre>	