SQL databases and R

```
#install.packages("tidyverse")
#install.packages("dbplyr")
#install.packages("RSQLite")
library(RSQLite)
library(tidyverse)
```

```
-- 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   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- 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 becor
```

The portal_mammals database

```
dir.create("data_raw", showWarnings = FALSE)
download.file(url = "https://ndownloader.figshare.com/files/2292171",
              destfile = "data_raw/portal_mammals.sqlite", mode = "wb")
```

Connecting to databases

```
library(dplyr)
library(dbplyr)
```

```
Attaching package: 'dbplyr'

The following objects are masked from 'package:dplyr':

    ident, sql
```

```
mammals <- DBI::dbConnect(RSQLite::SQLite(), "data_raw/portal_mammals.sqlite")
```

```
## A closer look at how connected mammals database looks like
src_dbi(mammals)
```

```
src:  sqlite 3.46.0 [/Users/namigabbasov/Desktop/R-Data-Carpentry/data_raw/portal_mammals.sq
tbls: plots, species, surveys
```

```
#three tables: plots, species, surveys
```

```
## Querying the database with the SQL syntax
tbl(mammals, sql("SELECT year, species_id, plot_id FROM surveys"))
```

```
# Source:   SQL [?? x 3]
# Database: sqlite 3.46.0 [/Users/namigabbasov/Desktop/R-Data-Carpentry/data_raw/portal_mamma
    year species_id plot_id
   <int> <chr>        <int>
 1  1977 NL               2
 2  1977 NL               3
 3  1977 DM               2
 4  1977 DM               7
 5  1977 DM               3
 6  1977 PF               1
 7  1977 PE               2
 8  1977 DM               1
 9  1977 DM               1
10  1977 PF               6
# i more rows
```

```
## Querying the database with the dplyr syntax
surveys <- tbl(mammals, "surveys")
surveys %>%
    select(year, species_id, plot_id)
```

```
# Source:   SQL [?? x 3]
# Database: sqlite 3.46.0 [/Users/namigabbasov/Desktop/R-Data-Carpentry/data_raw/portal_mamma
    year species_id plot_id
   <int> <chr>        <int>
 1  1977 NL               2
 2  1977 NL               3
 3  1977 DM               2
 4  1977 DM               7
 5  1977 DM               3
 6  1977 PF               1
 7  1977 PE               2
 8  1977 DM               1
 9  1977 DM               1
10  1977 PF               6
# i more rows
```

```
## look first ten observations
head(surveys, n = 10)
```

```
# Source:   SQL [10 x 9]
# Database: sqlite 3.46.0 [/Users/namigabbasov/Desktop/R-Data-Carpentry/data_raw/portal_mamma
   record_id month   day  year plot_id species_id sex   hindfoot_length weight
       <int> <int> <int> <int>   <int> <chr>      <chr>           <int>  <int>
 1         1     7    16  1977       2 NL         M                  32     NA
 2         2     7    16  1977       3 NL         M                  33     NA
 3         3     7    16  1977       2 DM         F                  37     NA
 4         4     7    16  1977       7 DM         M                  36     NA
 5         5     7    16  1977       3 DM         M                  35     NA
 6         6     7    16  1977       1 PF         M                  14     NA
 7         7     7    16  1977       2 PE         F                  NA     NA
 8         8     7    16  1977       1 DM         M                  37     NA
 9         9     7    16  1977       1 DM         F                  34     NA
10        10     7    16  1977       6 PF         F                  20     NA
```

```
# difference between read_csv and database
nrow(surveys)
```

```
[1] NA
```

```r
#SQL translation
show_query(head(surveys, n = 10))
```

```
<SQL>
SELECT `surveys`.*
FROM `surveys`
LIMIT 10
```

Simple database queries

```r
surveys %>%
  filter(weight < 5) %>%
  select(species_id, sex, weight)
```

```
# Source:   SQL [?? x 3]
# Database: sqlite 3.46.0 [/Users/namigabbasov/Desktop/R-Data-Carpentry/data_raw/portal_mamma
   species_id sex   weight
   <chr>      <chr>  <int>
 1 PF         M          4
 2 PF         F          4
 3 PF         <NA>       4
 4 PF         F          4
 5 PF         F          4
 6 RM         M          4
 7 RM         F          4
 8 RM         M          4
 9 RM         M          4
10 RM         M          4
# i more rows
```

Laziness

```r
## dplyr never pulls data into R unless you explicitly ask for it
data_subset <- surveys %>%
  filter(weight < 5) %>%
  select(species_id, sex, weight)
```

```
data_subset %>%
  select(-sex)
```

```
# Source:   SQL [?? x 2]
# Database: sqlite 3.46.0 [/Users/namigabbasov/Desktop/R-Data-Carpentry/data_raw/portal_mamma
   species_id weight
   <chr>       <int>
 1 PF              4
 2 PF              4
 3 PF              4
 4 PF              4
 5 PF              4
 6 RM              4
 7 RM              4
 8 RM              4
 9 RM              4
10 RM              4
# i more rows
```

```
# To retrieve all of the query results from the database, we add the collect() command to ou
data_subset <- surveys %>%
  filter(weight < 5) %>%
  select(species_id, sex, weight) %>%
  collect()
data_subset
```

```
# A tibble: 17 x 3
   species_id sex   weight
   <chr>      <chr>  <int>
 1 PF         M          4
 2 PF         F          4
 3 PF         <NA>       4
 4 PF         F          4
 5 PF         F          4
 6 RM         M          4
 7 RM         F          4
 8 RM         M          4
 9 RM         M          4
10 RM         M          4
11 RM         M          4
12 RM         F          4
```

```
13 RM         M           4
14 RM         M           4
15 RM         M           4
16 PF         M           4
17 PP         M           4
```

Complex database queries

```
plots <- tbl(mammals, "plots")
plots
```

```
# Source:   table<`plots`> [?? x 2]
# Database: sqlite 3.46.0 [/Users/namigabbasov/Desktop/R-Data-Carpentry/data_raw/portal_mamma
   plot_id plot_type
     <int> <chr>
 1       1 Spectab exclosure
 2       2 Control
 3       3 Long-term Krat Exclosure
 4       4 Control
 5       5 Rodent Exclosure
 6       6 Short-term Krat Exclosure
 7       7 Rodent Exclosure
 8       8 Control
 9       9 Spectab exclosure
10      10 Rodent Exclosure
# i more rows
```

```
## The plot_id column also features in the surveys table
surveys
```

```
# Source:   table<`surveys`> [?? x 9]
# Database: sqlite 3.46.0 [/Users/namigabbasov/Desktop/R-Data-Carpentry/data_raw/portal_mamma
   record_id month   day  year plot_id species_id sex   hindfoot_length weight
       <int> <int> <int> <int>   <int> <chr>      <chr>           <int>  <int>
 1         1     7    16  1977       2 NL         M                  32     NA
 2         2     7    16  1977       3 NL         M                  33     NA
 3         3     7    16  1977       2 DM         F                  37     NA
 4         4     7    16  1977       7 DM         M                  36     NA
 5         5     7    16  1977       3 DM         M                  35     NA
 6         6     7    16  1977       1 PF         M                  14     NA
```

```
 7         7     7    16  1977       2 PE         F                    NA     NA
 8         8     7    16  1977       1 DM         M                    37     NA
 9         9     7    16  1977       1 DM         F                    34     NA
10        10     7    16  1977       6 PF         F                    20     NA
# i more rows
```

```
## Inner join
plots %>%
  filter(plot_id == 1) %>%
  inner_join(surveys) %>%
  collect()
```

```
Joining with `by = join_by(plot_id)`
```

```
# A tibble: 1,995 x 10
   plot_id plot_type         record_id month   day  year species_id sex
     <int> <chr>                 <int> <int> <int> <int> <chr>      <chr>
 1       1 Spectab exclosure         6     7    16  1977 PF         M
 2       1 Spectab exclosure         8     7    16  1977 DM         M
 3       1 Spectab exclosure         9     7    16  1977 DM         F
 4       1 Spectab exclosure        78     8    19  1977 PF         M
 5       1 Spectab exclosure        80     8    19  1977 DS         M
 6       1 Spectab exclosure       218     9    13  1977 PF         M
 7       1 Spectab exclosure       222     9    13  1977 DS         M
 8       1 Spectab exclosure       239     9    13  1977 DS         M
 9       1 Spectab exclosure       263    10    16  1977 DM         M
10       1 Spectab exclosure       270    10    16  1977 DM         F
# i 1,985 more rows
# i 2 more variables: hindfoot_length <int>, weight <int>
```

```
## Left Join
species <- tbl(mammals, "species")
unique_genera <- left_join(surveys, plots) %>%
    left_join(species) %>%
    group_by(plot_type) %>%
    summarize(
        n_genera = n_distinct(genus)
    ) %>%
    collect()
```

```
Joining with `by = join_by(plot_id)`
Joining with `by = join_by(species_id)`
```

Creating a new SQLite database

```
## download files
download.file("https://ndownloader.figshare.com/files/3299483",
              "data_raw/species.csv")
download.file("https://ndownloader.figshare.com/files/10717177",
              "data_raw/surveys.csv")
download.file("https://ndownloader.figshare.com/files/3299474",
              "data_raw/plots.csv")
```

```
## import files
species <- read_csv("data_raw/species.csv")
```

```
Rows: 54 Columns: 4
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (4): species_id, genus, species, taxa

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
surveys <- read_csv("data_raw/surveys.csv")
```

```
Rows: 35549 Columns: 9
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (2): species_id, sex
dbl (7): record_id, month, day, year, plot_id, hindfoot_length, weight

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
plots <- read_csv("data_raw/plots.csv")
```

```
Rows: 24 Columns: 2
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): plot_type
dbl (1): plot_id
```

```
i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```