

# Title - SQL demo

Author: Andrew D. Nguyen, PhD, Quantitative Biologist

Date:2023-08-27

## Contents

<b>1</b>	<b>Introduction -&gt; Goals</b>	<b>2</b>
<b>2</b>	<b>Load Libraries</b>	<b>2</b>
<b>3</b>	<b>The database - Portal Mammals Database</b>	<b>2</b>
<b>4</b>	<b>SessionInfo</b>	<b>4</b>

# 1 Introduction -> Goals

In this demo, I become more familiar with interacting with SQL databases. I'll be following a tutorial from **data carpentry**, but I'd also like to peruse open databases to become more versed in SQL and querying databases into R for data wrangling and analyses. Lastly, I'll construct my own databases from datasets I have generated.

## 2 Load Libraries

```
library(tidyverse) # data wrangling
library(RSQLite) # SQL - R package
library(dplyr) # SQL - R package, lets dplyr apply to
#SQL databases
```

## 3 The database - Portal Mammals Database

I have to first download the portal mammals database off of figshare. Then, load the mammals database.

```
#download.file(url = "https://ndownloader.figshare.com/files/2292171",
#              destfile = "data/portal_mammals.sqlite", mode = "wb")
# load data
mammals<-DBI::dbConnect(RSQLite::SQLite(), "data/portal_mammals.sqlite")
#This command does not load the data into the R session
#(as the read_csv() function did).
#Instead, it instructs R
#to connect to the SQLite database
#contained in the portal_mammals.sqlite file.

#now, lets see what the database is made of
src_dbi(mammals)
```

```
## src:  sqlite 3.41.2 [C:\Users\anbe6\Documents\GitHub\adnguyen.github.io\demos\data\portal_mammals.sqlite]
## tbls: plots, species, surveys
```

```
## querying database with tbl
tbl(mammals, sql("SELECT year, species_id, plot_id FROM surveys"))
```

```
## # Source:   SQL [?? x 3]
## # Database: sqlite 3.41.2 [C:\Users\anbe6\Documents\GitHub\adnguyen.github.io\demos\data\portal_mammals.sqlite]
##   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
```

```
surveys <- tbl(mammals, "surveys")
surveys %>%
  select(year, species_id, plot_id)
```

```
## # Source:   SQL [?? x 3]
## # Database: sqlite 3.41.2 [C:\Users\anbe6\Documents\GitHub\adnguyen.github.io\demos\data\portal_mammals]
##   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
```

```
show_query(head(surveys, n = 10))
```

```
## <SQL>
## SELECT *
## FROM 'surveys'
## LIMIT 10
```

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

```
## # Source:   SQL [?? x 3]
## # Database: sqlite 3.41.2 [C:\Users\anbe6\Documents\GitHub\adnguyen.github.io\demos\data\portal_mammals]
##   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
```

```
##R is lazy and doesn't read in data until specified
# using collect to read in the data into R
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
```

## 4 SessionInfo

```
sessionInfo()
```

```
## R version 4.3.1 (2023-06-16 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19045)
##
## Matrix products: default
##
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## time zone: America/New_York
## tzcode source: internal
##
```

```

## attached base packages:
## [1] stats      graphics  grDevices utils      datasets  methods   base
##
## other attached packages:
## [1] dbplyr_2.3.3    RSQLite_2.3.1  lubridate_1.9.2 forcats_1.0.0
## [5] stringr_1.5.0  dplyr_1.1.2    purrr_1.0.2     readr_2.1.4
## [9] tidyr_1.3.0     tibble_3.2.1   ggplot2_3.4.3   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
## [1] bit_4.0.5      gtable_0.3.4    compiler_4.3.1  tidyselect_1.2.0
## [5] blob_1.2.4     scales_1.2.1    yaml_2.3.7      fastmap_1.1.1
## [9] R6_2.5.1       generics_0.1.3  knitr_1.43      munsell_0.5.0
## [13] DBI_1.1.3      pillar_1.9.0    tzdb_0.4.0      rlang_1.1.1
## [17] utf8_1.2.3     cachem_1.0.8    stringi_1.7.12  xfun_0.40
## [21] bit64_4.0.5    memoise_2.0.1   timechange_0.2.0 cli_3.6.1
## [25] withr_2.5.0    magrittr_2.0.3  digest_0.6.33   grid_4.3.1
## [29] rstudioapi_0.15.0 hms_1.1.3       lifecycle_1.0.3 vctrs_0.6.3
## [33] evaluate_0.21  glue_1.6.2      fansi_1.0.4     colorspace_2.1-0
## [37] rmarkdown_2.24 tools_4.3.1     pkgconfig_2.0.3 htmltools_0.5.6

```