USCOTS 2025 breakout session: Download the airlines data

Nicholas Horton (nhorton@amherst.edu) and Jo Hardin (jo.hardin@pomona.edu)

2025-07-11

Table of contents

This file downloads the airline flight data from the American Statistical Association's Data Expo 2024. The data are stored in Parquet format, which is efficient for large datasets and can be read using SQL commands using the **arrow** package in R.

Once you've successfully rendered this Quarto file (`1_download_data.qmd`), open and render `2_explore-sql.qmd` to carry out an analysis of the downloaded data. (Note that both files must be saved and rendered from the same folder to access the downloaded data.)

See https://community.amstat.org/dataexpo/home for background and the source for much of the code below and https://github.com/nicholasjhorton/uscots2025-sql-data-technologies for more resources.

```
tictoc::tic()

folder_name <- "data_airlines"
dir.create(folder_name, showWarnings = TRUE)
stopifnot(file.exists(folder_name))

list.files(folder_name)
```

```
character(0)
```

```
base_url <- "https://blobs.duckdb.org/flight-data-partitioned/"

years <- 2023:2024 # could be 1987:2024, but that's a *lot* of data!
```

```
# please don't download too much data during the breakout session

files <- paste0("Year=", years, "/data_0.parquet")
my_files <- paste0(folder_name, "/Year=", years, "/data_0.parquet")

for (dir in dirname(my_files)) {
  dir.create(dir, showWarnings = FALSE)
}

out <- curl::multi_download(paste0(base_url, files), my_files, resume = TRUE)

tictoc::toc()
```

```
9.904 sec elapsed
```

The **tictoc** package will tell you how long it took to download the files into your local system. It should not take too long since you are only downloading a proper subset of what's there (to avoid overloading the network at the breakout session).

```
list.files(folder_name)
```

```
[1] "Year=2023" "Year=2024"
```

We will be accessing these files in the next Quarto file, `2_explore-sql.qmd`.