# Download NYC TLC Trip Data

## Table of contents



```
library(tidyverse)
library(here)
library(httr)
library(lubridate)
library(arrow)

# Create data directory if it doesn't exist
data_dir <- here::here("data")
if (!dir.exists(data_dir)) {
  dir.create(data_dir)
}

year <- params$year
```

```
# Function to get the file size from URL without downloading
get_remote_file_size <- function(url) {
  tryCatch({
    head_result <- HEAD(url)
    as.numeric(headers(head_result)$`content-length`)
  }, error = function(e) {
    NA_real_
  })
}

# Function to generate URLs for a given year
generate_urls <- function(year) {
```

```r
  months <- sprintf("%02d", 1:12)
  base_url <- "https://d37ci6vzurychx.cloudfront.net/trip-data"

  urls <- tibble(
    month = months,
    filename = str_glue("yellow_tripdata_{year}-{month}.parquet"),
    url = str_glue("{base_url}/yellow_tripdata_{year}-{month}.parquet"),
    local_path = here::here("data", filename)
  )

  return(urls)
}

# Function to download a file if it's missing or changed
download_if_needed <- function(url, local_path) {
  needs_download <- TRUE

  if (file.exists(local_path)) {
    remote_size <- get_remote_file_size(url)
    local_size <- file.size(local_path)

    if (!is.na(remote_size) && remote_size == local_size) {
      needs_download <- FALSE
      message(str_glue("File {basename(local_path)} already exists and appears unchanged. Sk
    }
  }

  if (needs_download) {
    message(str_glue("Downloading {basename(local_path)}..."))
    download.file(url, local_path, mode = "wb", method = "auto")
    message(str_glue("Successfully downloaded {basename(local_path)}"))
  }

  return(!needs_download)
}
```

### Download TLC Trip Data

We'll download yellow taxi trip data for 2024. You can modify the year parameter to download data for different years in the yaml.

```r
# Generate URLs for the specified year
urls_df <- generate_urls(year)

# Download files
results <- urls_df %>%
  rowwise() %>%
  mutate(
    skipped = download_if_needed(url, local_path),
    timestamp = Sys.time()
  ) %>%
  ungroup()
```

```
Downloading yellow_tripdata_2024-01.parquet...

Successfully downloaded yellow_tripdata_2024-01.parquet

Downloading yellow_tripdata_2024-02.parquet...

Successfully downloaded yellow_tripdata_2024-02.parquet

Downloading yellow_tripdata_2024-03.parquet...

Successfully downloaded yellow_tripdata_2024-03.parquet

Downloading yellow_tripdata_2024-04.parquet...

Successfully downloaded yellow_tripdata_2024-04.parquet

Downloading yellow_tripdata_2024-05.parquet...

Successfully downloaded yellow_tripdata_2024-05.parquet

Downloading yellow_tripdata_2024-06.parquet...

Successfully downloaded yellow_tripdata_2024-06.parquet

Downloading yellow_tripdata_2024-07.parquet...
```

```
Successfully downloaded yellow_tripdata_2024-07.parquet

Downloading yellow_tripdata_2024-08.parquet...

Successfully downloaded yellow_tripdata_2024-08.parquet

Downloading yellow_tripdata_2024-09.parquet...

Successfully downloaded yellow_tripdata_2024-09.parquet

Downloading yellow_tripdata_2024-10.parquet...

Successfully downloaded yellow_tripdata_2024-10.parquet

Downloading yellow_tripdata_2024-11.parquet...

Successfully downloaded yellow_tripdata_2024-11.parquet

Downloading yellow_tripdata_2024-12.parquet...

Successfully downloaded yellow_tripdata_2024-12.parquet
```

```r
# Display summary
results %>%
  select(filename, skipped, timestamp) %>%
  knitr::kable(
    caption = str_glue("Download summary for {year} TLC trip data"),
    col.names = c("Filename", "Skipped", "Timestamp")
  )
```

Table 1: Download summary for 2024 TLC trip data

| Filename | Skipped | Timestamp |
|---|---|---|
| yellow_tripdata_2024-01.parquet | FALSE | 2025-03-23 19:21:50 |
| yellow_tripdata_2024-02.parquet | FALSE | 2025-03-23 19:21:50 |
| yellow_tripdata_2024-03.parquet | FALSE | 2025-03-23 19:21:50 |
| yellow_tripdata_2024-04.parquet | FALSE | 2025-03-23 19:21:50 |
| yellow_tripdata_2024-05.parquet | FALSE | 2025-03-23 19:21:50 |

| Filename | Skipped | Timestamp |
|---|---|---|
| yellow_tripdata_2024-06.parquet | FALSE | 2025-03-23 19:21:50 |
| yellow_tripdata_2024-07.parquet | FALSE | 2025-03-23 19:21:50 |
| yellow_tripdata_2024-08.parquet | FALSE | 2025-03-23 19:21:50 |
| yellow_tripdata_2024-09.parquet | FALSE | 2025-03-23 19:21:50 |
| yellow_tripdata_2024-10.parquet | FALSE | 2025-03-23 19:21:50 |
| yellow_tripdata_2024-11.parquet | FALSE | 2025-03-23 19:21:50 |
| yellow_tripdata_2024-12.parquet | FALSE | 2025-03-23 19:21:50 |

### Verify Downloaded Files

Let's check the files we've downloaded and their sizes:

```
downloaded_files <- list.files(
  path = data_dir,
  pattern = str_glue("yellow_tripdata_{year}.*\\.parquet$"),
  full.names = TRUE
) %>%
  file.info() %>%
  rownames_to_column("filepath") %>%
  as_tibble() %>%
  mutate(
    filename = basename(filepath),
    size_mb = size / 1024^2
  ) %>%
  select(filename, size_mb, mtime)

downloaded_files %>%
  arrange(desc(mtime)) %>%
  knitr::kable(
    caption = "Downloaded files information",
    col.names = c("Filename", "Size (MB)", "Modified Time"),
    digits = 2
  )
```

Table 2: Downloaded files information

| Filename | Size (MB) | Modified Time |
|---|---|---|
| yellow_tripdata_2024-12.parquet | 58.67 | 2025-03-23 19:21:50 |
| yellow_tripdata_2024-11.parquet | 57.85 | 2025-03-23 19:21:48 |

| Filename | Size (MB) | Modified Time |
|---|---|---|
| yellow_tripdata_2024-10.parquet | 61.37 | 2025-03-23 19:21:46 |
| yellow_tripdata_2024-09.parquet | 58.34 | 2025-03-23 19:21:44 |
| yellow_tripdata_2024-08.parquet | 48.70 | 2025-03-23 19:21:42 |
| yellow_tripdata_2024-07.parquet | 49.88 | 2025-03-23 19:21:39 |
| yellow_tripdata_2024-06.parquet | 57.09 | 2025-03-23 19:21:37 |
| yellow_tripdata_2024-05.parquet | 59.66 | 2025-03-23 19:21:34 |
| yellow_tripdata_2024-04.parquet | 56.39 | 2025-03-23 19:21:32 |
| yellow_tripdata_2024-03.parquet | 57.30 | 2025-03-23 19:21:30 |
| yellow_tripdata_2024-02.parquet | 48.02 | 2025-03-23 19:21:28 |
| yellow_tripdata_2024-01.parquet | 47.65 | 2025-03-23 19:21:24 |