NYC Yellow Cab Ride Data 2023 Q1

Assignment 1.2 - Example submission

```r
library(arrow)
library(fpp3)
library(here)
library(plotly)
```

Data Source

This data ways downloaded from the NYC.gov **Taxi & Limousine Commission**, and represents Yellow Cab trip data from January through March of 2023. The data is in monthly `parquet` files, and has been assembed into a single 3 month time series. The original data is available at https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

Importing the Data

Downloaded files are in the `data` directory of this Github repo.

```r
pfiles <- here('data', list.files(here('data')))
```

Read in a small sample to explore the dataset.

```r
frac <- 1/20
set.seed(2023)
small <- pfiles[1] |>
    read_parquet() |>
    sample_frac(frac)
```

```
small_ts <- small |>
    mutate(ymd = as.Date(with_tz(tpep_pickup_datetime, "America/New_York"))) |>
    count(ymd) |>
    mutate(trips = n / 1000) |>
    select(-n) |>
    as_tsibble(index = ymd)

est_rows <- (1e-6 * length(pfiles) * nrow(small) / frac) |> round(1)
est_rows
```

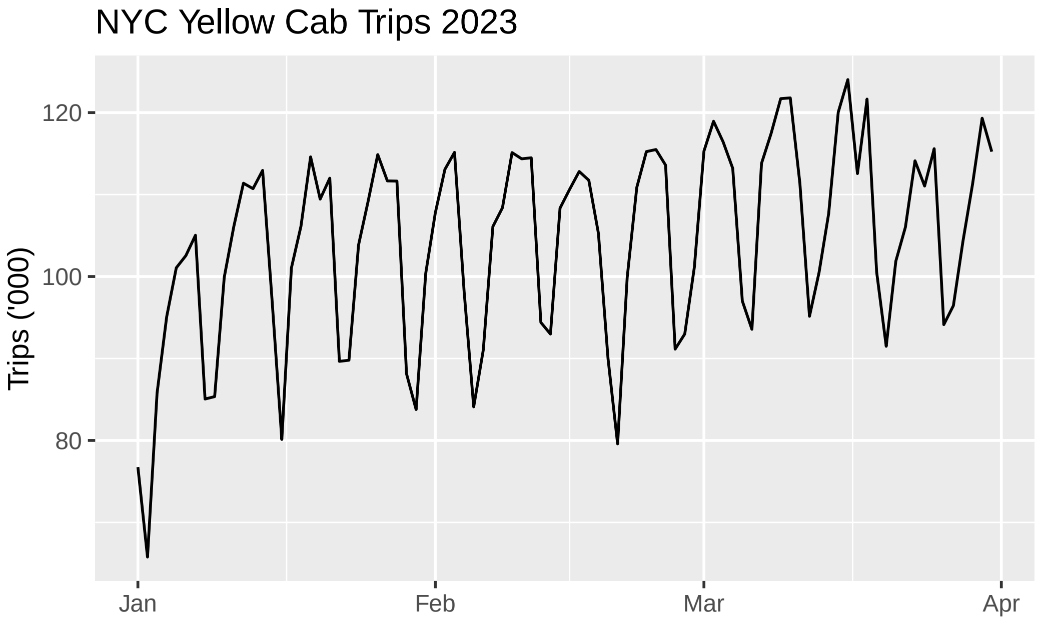
```
[1] 9.2
```

Since we have roughly 9.2M rows for the three months covered by the data, we read and aggregate the data in one step.

```
rides <- purrr::map_dfr(pfiles, read_parquet) |>
    mutate(ymd = as.Date(with_tz(tpep_pickup_datetime, "America/New_York"))) |>
    filter(year(ymd) == 2023,
           month(ymd) <= 3) |>
    count(ymd) |>
    mutate(trips = n / 1000) |>
    select(-n) |>
    as_tsibble(index = ymd)
```

Time Series Plot

```
rides_tsplt <- rides |>
    autoplot(trips) +
      labs(title = "NYC Yellow Cab Trips 2023",
       y = "Trips ('000)",
       x = "")
rides_tsplt
```



Discussion

The data shows a clear weekly seasonality, and perhaps a mild trend. We can see minimums for Sunday and Monday, but there is an irregular pattern of 2-day verus 1-day troughs. These may be difficult to model without multiple years of data.

The 3 month series may be adequate for producing a forecast of trips that would be useful for operational planning, since support staff and resources are required in proportion to the number of trips. There is other data in the dataset, not explored above that may also help forecast the trips volume.