Module 1: Tabular Data

Working with larger-than-RAM data using duckdbfs

Joey Zhang

```
# Set CRAN mirror to avoid prompts
library(duckdbfs)
library(dplyr)
```

```
Attaching package: 'dplyr'

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

    filter, lag

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

    intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.5.2
```

```
# Remote S3 path to EXIOBASE 3 (Source Cooperative)

duckdbfs::duckdb_secrets(
    key = "",
    secret = "",
    endpoint = "s3.amazonaws.com",
    region = "us-west-2"
)
```

```
[1] 1
```

```
s3_url <- "s3://us-west-2.opendata.source.coop/youssef-harby/exiobase-3/4588235/parquet/**"

# Open the dataset lazily
exio <- open_dataset(s3_url)

# View the schema (column names and types) without reading data
glimpse(exio)
```

```
Rows: ??
Columns: 8
Database: DuckDB 1.4.4 [root@Darwin 24.5.0:R 4.5.1/:memory:]
$ stressor <chr> "Value Added", "Value Added", "Value Added", "Value Added", "~
$ region   <chr> "AT", "AT", "AT", "AT", "AT", "AT", "AT", "AT", "AT", "AT", "~
$ sector   <chr> "Cultivation of wheat", "Cultivation of cereal grains nec", "~
$ value    <dbl> 183.1118891, 402.2305799, 830.2127384, 101.9705426, 31.763189~
$ unit     <chr> "M.EUR", "M.EUR", "M.EUR", "M.EUR", "M.EUR", "M.EUR", "M.EUR"~
$ year     <dbl> 1995, 1995, 1995, 1995, 1995, 1995, 1995, 1995, 1995, 1995, 1~
$ format   <chr> "ixi", "ixi", "ixi", "ixi", "ixi", "ixi", "ixi", "ixi", "ixi"~
$ matrix   <chr> "F_impacts", "F_impacts", "F_impacts", "F_impacts", "F_impact~
```

```r
# Get all CO2-related stressors from F_satellite matrix
co2 <- exio |>
    filter(matrix == "F_satellite") |>
    filter(grepl("CO2|carbon dioxide", stressor, ignore.case = TRUE)) |>
    collect()  # Collect data upfront to avoid connection issues

co2
```

```
# A tibble: 973,538 x 8
   stressor               region sector          value unit   year format matrix
   <chr>                  <chr>  <chr>           <dbl> <chr> <dbl> <chr>  <chr>
 1 CO2 - combustion - air AT     Cultivation o~ 2.27e8 kg     1995 ixi    F_sat~
 2 CO2 - combustion - air AT     Cultivation o~ 2.16e8 kg     1995 ixi    F_sat~
 3 CO2 - combustion - air AT     Cultivation o~ 1.01e8 kg     1995 ixi    F_sat~
 4 CO2 - combustion - air AT     Cultivation o~ 4.98e7 kg     1995 ixi    F_sat~
 5 CO2 - combustion - air AT     Cultivation o~ 1.10e7 kg     1995 ixi    F_sat~
 6 CO2 - combustion - air AT     Cultivation o~ 1.13e4 kg     1995 ixi    F_sat~
 7 CO2 - combustion - air AT     Cultivation o~ 1.74e6 kg     1995 ixi    F_sat~
 8 CO2 - combustion - air AT     Cattle farming 6.79e7 kg     1995 ixi    F_sat~
 9 CO2 - combustion - air AT     Pigs farming   4.87e7 kg     1995 ixi    F_sat~
10 CO2 - combustion - air AT     Poultry farmi~ 4.62e7 kg     1995 ixi    F_sat~
# i 973,528 more rows
```

identify which regions are the top 5 co2 emitters

```r
co2_top5 <- co2 |>
    filter(year == 2022) |>
    group_by(region) |>
    summarize(total_co2 = sum(value, na.rm = TRUE)) |>
    arrange(desc(total_co2)) |>
    head(5)

co2_top5
```

```
# A tibble: 5 x 2
  region total_co2
```

```
  <chr>      <dbl>
1 CN       2.24e13
2 US       7.43e12
3 IN       5.31e12
4 WA       4.12e12
5 WM       4.11e12
```

filter the co2 data down to just these top countries, and plot thier total co2 emissions by year

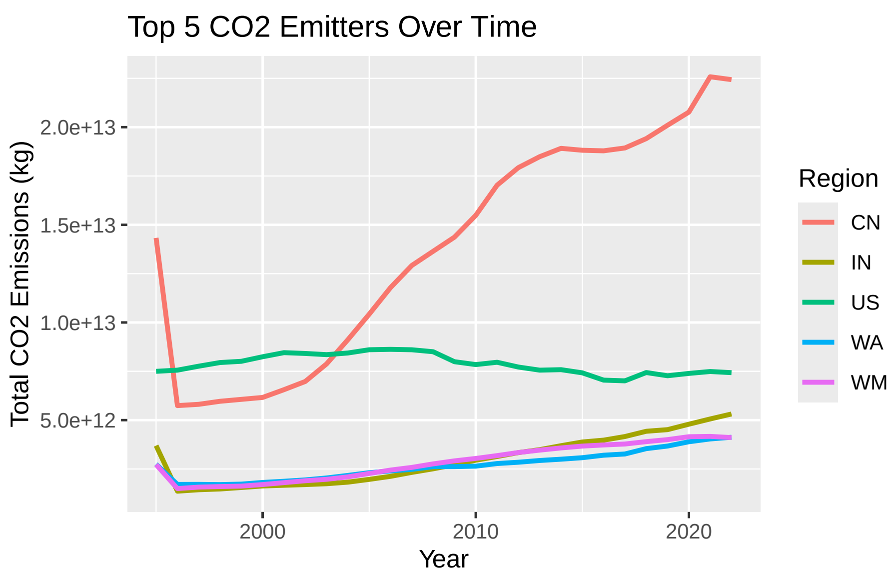
```r
library(ggplot2)

co2_top5_filtered <- co2 |>
    filter(region %in% co2_top5$region) |>
    group_by(region, year) |>
    summarize(total_co2 = sum(value, na.rm = TRUE), .groups = "drop")
co2_top5_filtered
```

```
# A tibble: 140 x 3
   region  year total_co2
   <chr>  <dbl>     <dbl>
 1 CN      1995   1.43e13
 2 CN      1996   5.74e12
 3 CN      1997   5.81e12
 4 CN      1998   5.96e12
 5 CN      1999   6.06e12
 6 CN      2000   6.16e12
 7 CN      2001   6.56e12
 8 CN      2002   6.98e12
 9 CN      2003   7.88e12
10 CN      2004   9.12e12
# i 130 more rows
```

```r
library(ggplot2)
p <- ggplot(co2_top5_filtered, aes(x = year, y = total_co2, color = region)) +
    geom_line(linewidth = 1) +
    labs(title = "Top 5 CO2 Emitters Over Time",
         x = "Year",
         y = "Total CO2 Emissions (kg)",
         color = "Region")
ggsave("co2_top5_plot.png",plot=p)
```

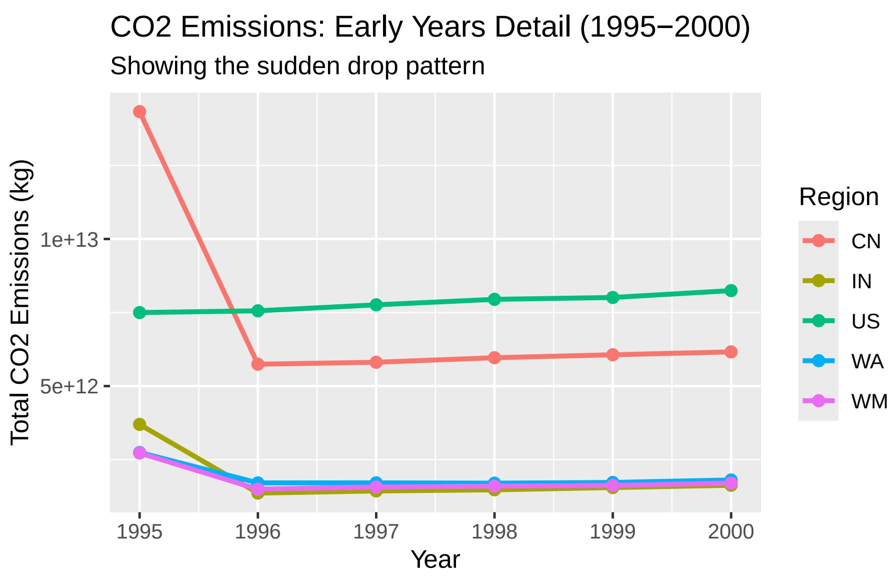
```
Saving 5.5 x 3.5 in image
```

```r
p
```



take a look at the early years of the data to see if we can identify the sudden drop pattern in the top 5 emitters

```r
# Plot early years in detail
early_years <- co2 |>
    filter(region %in% co2_top5$region, year >= 1995, year <= 2000) |>
    group_by(region, year) |>
    summarize(total_co2 = sum(value, na.rm = TRUE), .groups = "drop")
ggplot(early_years, aes(x = year, y = total_co2, color = region)) +
    geom_line(linewidth = 1) +
    geom_point(size = 2) +
    labs(
        title = "CO2 Emissions: Early Years Detail (1995-2000)",
        subtitle = "Showing the sudden drop pattern",
        x = "Year",
        y = "Total CO2 Emissions (kg)",
        color = "Region"
    )
```




```r
ggsave("co2_early_years_plot.png", plot = last_plot())
```

```
Saving 5.5 x 3.5 in image
```