Module 1: Tabular Data

Working with larger-than-RAM data using duckdbfs

Mei Collins

Case Study: Global Supply Chains

We will be working with EXIOBASE 3.8.1, a global Multi-Regional Input-Output (MRIO) database. This dataset tracks economic transactions between sectors and regions, along with their environmental impacts (emissions, resource use, etc.).

Data description: - **Coverage**: 44 countries + 5 rest-of-world regions. - **Timeframe**: 1995–2022. - **Content**: Economic transactions (Z matrix), final demand (Y matrix), and environmental stressors (F matrix). - **Format**: Cloud-optimized Parquet, partitioned by year and matrix type.

Setup

```
library(duckdbfs)
```

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

```
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)
```

Connecting to data and quick exploration

```
# Open the dataset - filtering for F_satellite matrix (environmental stressors)

# Note: The F matrix is called "F_satellite" in this dataset
# Using wildcard pattern to capture all years (1995-2022)
matrix_f <- open_dataset("s3://us-west-2.opendata.source.coop/youssef-harby/exiobase-3/45882

# Check the structure
matrix_f
```

```
# Source:   table<gnuwiripurecfjz> [?? x 8]
# Database: DuckDB 1.4.3 [colli@Windows 10 x64:R 4.5.1/:memory:]
   stressor                      region sector   value unit   year format matrix
   <chr>                         <chr>  <chr>    <dbl> <chr> <dbl> <chr>  <chr>
 1 Taxes less subsidies on prod~ AT     Culti~ 3.30e+0 M.EUR  1995 ixi    F_sat~
 2 Taxes less subsidies on prod~ AT     Culti~ 6.72e+0 M.EUR  1995 ixi    F_sat~
 3 Taxes less subsidies on prod~ AT     Culti~ 5.22e+0 M.EUR  1995 ixi    F_sat~
 4 Taxes less subsidies on prod~ AT     Culti~ 6.98e-1 M.EUR  1995 ixi    F_sat~
 5 Taxes less subsidies on prod~ AT     Culti~ 5.31e-1 M.EUR  1995 ixi    F_sat~
 6 Taxes less subsidies on prod~ AT     Culti~ 2.79e-4 M.EUR  1995 ixi    F_sat~
 7 Taxes less subsidies on prod~ AT     Culti~ 4.13e+0 M.EUR  1995 ixi    F_sat~
 8 Taxes less subsidies on prod~ AT     Cattl~ 1.09e+1 M.EUR  1995 ixi    F_sat~
 9 Taxes less subsidies on prod~ AT     Pigs ~ 8.43e+0 M.EUR  1995 ixi    F_sat~
10 Taxes less subsidies on prod~ AT     Poult~ 2.48e+0 M.EUR  1995 ixi    F_sat~
# i more rows
```

```
# Preview the first few rows
matrix_f |>
  head(10) |>
  collect()
```

```
# A tibble: 10 x 8
```

```
   stressor                      region sector   value unit   year format matrix
   <chr>                         <chr>  <chr>    <dbl> <chr> <dbl> <chr>  <chr>
 1 Taxes less subsidies on prod~ AT     Culti~ 3.30e+0 M.EUR  1995 ixi    F_sat~
 2 Taxes less subsidies on prod~ AT     Culti~ 6.72e+0 M.EUR  1995 ixi    F_sat~
 3 Taxes less subsidies on prod~ AT     Culti~ 5.22e+0 M.EUR  1995 ixi    F_sat~
 4 Taxes less subsidies on prod~ AT     Culti~ 6.98e-1 M.EUR  1995 ixi    F_sat~
 5 Taxes less subsidies on prod~ AT     Culti~ 5.31e-1 M.EUR  1995 ixi    F_sat~
 6 Taxes less subsidies on prod~ AT     Culti~ 2.79e-4 M.EUR  1995 ixi    F_sat~
 7 Taxes less subsidies on prod~ AT     Culti~ 4.13e+0 M.EUR  1995 ixi    F_sat~
 8 Taxes less subsidies on prod~ AT     Cattl~ 1.09e+1 M.EUR  1995 ixi    F_sat~
 9 Taxes less subsidies on prod~ AT     Pigs ~ 8.43e+0 M.EUR  1995 ixi    F_sat~
10 Taxes less subsidies on prod~ AT     Poult~ 2.48e+0 M.EUR  1995 ixi    F_sat~
```

```r
# See what years are available
matrix_f |>
  distinct(year) |>
  arrange(year) |>
  collect()
```

```
# A tibble: 28 x 1
    year
   <dbl>
 1  1995
 2  1996
 3  1997
 4  1998
 5  1999
 6  2000
 7  2001
 8  2002
 9  2003
10  2004
# i 18 more rows
```

I want to identify data on the CO2 production country by country, over time.

Which stressors are related to CO2?

```r
# Filter stressors to find those related to CO2
co2_stressors <- matrix_f |>
  distinct(stressor) |>
  filter(grepl("CO2|carbon dioxide", stressor, ignore.case = TRUE)) |>
```

```
  collect()

co2_stressors
```

```
# A tibble: 6 x 1
  stressor
  <chr>
1 CO2 - waste - fossil - air
2 CO2 - waste - biogenic - air
3 CO2 - agriculture - peat decay - air
4 CO2 - non combustion - Lime production - air
5 CO2 - combustion - air
6 CO2 - non combustion - Cement production - air
```

Great, now we know there are 6 stressors. Let's identify which regions are the top 5 CO2 emitters based on these stressors, and then we can explore their emissions over time.

```
# Filter data for CO2 stressors and aggregate by region
co2_by_region <- matrix_f |>
  filter(stressor %in% co2_stressors$stressor) |>
  group_by(region) |>
  summarise(total_co2 = sum(value, na.rm = TRUE)) |>
  arrange(desc(total_co2)) |>
  head(5) |>
  collect()

co2_by_region
```

```
# A tibble: 5 x 2
  region total_co2
  <chr>      <dbl>
1 CN       1.96e14
2 US       1.11e14
3 IN       4.27e13
4 WM       4.02e13
5 RU       3.83e13
```

```
top5_regions <- co2_by_region$region
```

Now that we know the top 5 emitting regions, we can plot their CO2 emissions over time.

First let's pull the relevant data as a time series

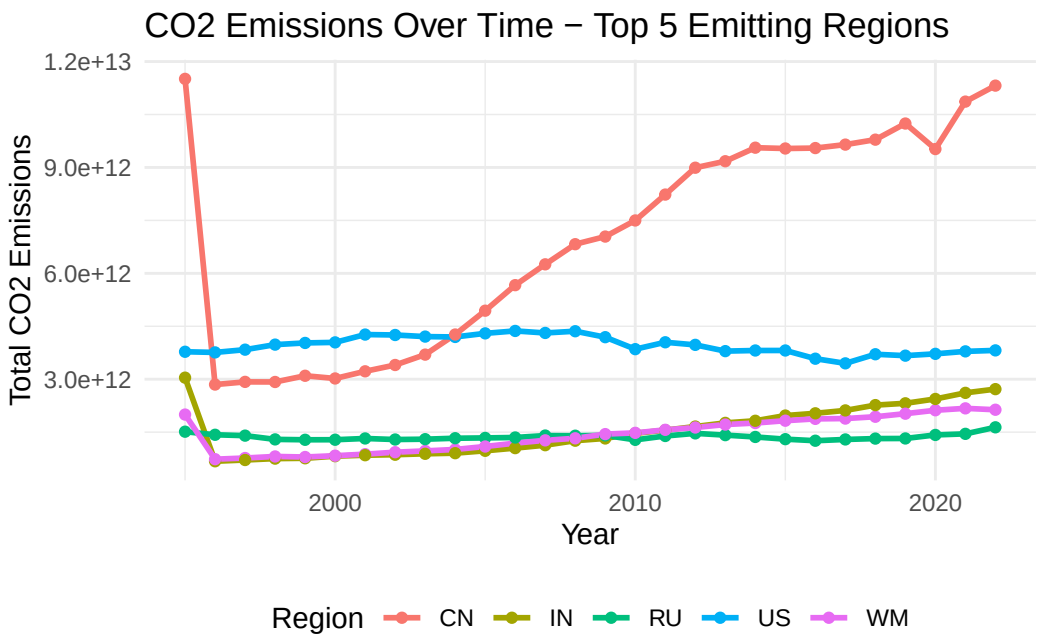
```r
# Filter and aggregate CO2 emissions by region and year for top 5
co2_time_series <- matrix_f |>
  filter(stressor %in% co2_stressors$stressor,
         region %in% top5_regions) |>
  group_by(region, year) |>
  summarise(total_co2 = sum(value, na.rm = TRUE), .groups = "drop") |>
  collect()
```

Then let's make a graph and save it

```r
# Plot emissions over time
emissions_graph <- ggplot(co2_time_series, aes(x = year, y = total_co2, color = region)) +
  geom_line(linewidth = 1) +
  geom_point() +
  labs(title = "CO2 Emissions Over Time - Top 5 Emitting Regions",
       x = "Year",
       y = "Total CO2 Emissions",
       color = "Region") +
  theme_minimal() +
  theme(legend.position = "bottom")

# Save the plot to a file
ggsave("top5emitters_timeseries.png", plot = emissions_graph, width = 10, height = 6)

# Display the plot
emissions_graph
```



Other questions we can explore with this data include: - which countries are decreasing in emissions and why? - what are the top sectors and how do they change over time?

Let's tackle the first question: which countries are decreasing in emissions and why?

```
# Calculate the change in emissions from 1995 to 2022 for each region
emissions_change <-  matrix_f|>
  filter(stressor %in% co2_stressors$stressor) |>
  group_by(region) |>
  summarise(emissions_1995 = sum(value[year == 1995]),
            emissions_2022 = sum(value[year == 2022]),
  )|>
    mutate(percent_change = (emissions_2022 - emissions_1995) / emissions_1995 * 100) |>
  arrange(percent_change) |>
  head(5) |>
  collect()
```

```
Warning: Missing values are always removed in SQL aggregation functions.
Use `na.rm = TRUE` to silence this warning
This warning is displayed once every 8 hours.
```

```
emissions_change
```

```
# A tibble: 5 x 4
  region emissions_1995 emissions_2022 percent_change
  <chr>           <dbl>          <dbl>          <dbl>
1 NL      393732291151.  135085242891.          -65.7
2 MT        2438603334.    1179562247.          -51.6
3 IE       71979157797.   51076082468.          -29.0
4 MX      643260822196.  456667256854.          -29.0
5 EE      273320649809.  196793051039.          -28.0
```

Now plot the emissions over time for these five regions

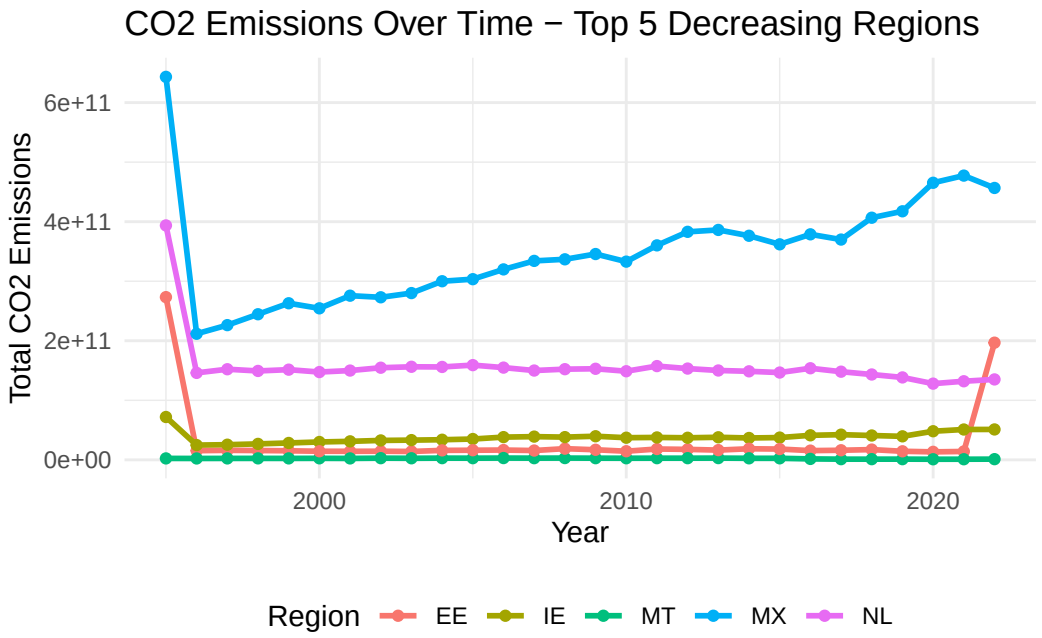
```
# Get the top 5 regions with the largest decrease in emissions
top5_decreasing_regions <- emissions_change$region


# Filter and aggregate CO2 emissions by region and year for top 5 decreasing regions
co2_decreasing_time_series <- matrix_f |>
    filter(stressor %in% co2_stressors$stressor,
             region %in% top5_decreasing_regions) |>
    group_by(region, year) |>
    summarise(total_co2 = sum(value, na.rm = TRUE), .groups = "drop") |>
    collect()

# Plot emissions over time for decreasing regions
decreasing_emissions_graph <- ggplot(co2_decreasing_time_series, aes(x = year, y = total_co2
  geom_line(linewidth = 1) +
  geom_point() +
  labs(title = "CO2 Emissions Over Time - Top 5 Decreasing Regions",
       x = "Year",
       y = "Total CO2 Emissions",
       color = "Region") +
  theme_minimal() +
  theme(legend.position = "bottom")

# Save the plot to a file
ggsave("top5decreasing_emitters_timeseries.png", plot = decreasing_emissions_graph, width =

# Display the plot
decreasing_emissions_graph
```


CO2 Emissions Over Time − Top 5 Decreasing Regions
Total CO2 Emissions
6e+11
4e+11
2e+11
0e+00
2000
2010
2020
Year
Region
EE
IE
MT
MX
NL