Module 1: Tabular Data

Working with larger-than-RAM data using duckdbfs

Sarah Sarfaty Epstein

Setup

```
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)
if (!require("patchwork", quietly = TRUE)) {
  install.packages("patchwork", repos = "https://cloud.r-project.org")
}
library(patchwork)
```

Exercise 1: connecting to remote data

We can open the entire dataset without downloading it using `open_dataset()`. The data is hosted on Source Cooperative. The `**` pattern allows recursive scanning of the partitioned parquet files.

```
# 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)
```

I want to identify data on the C02 production country by country over time. Open the f_satellite matrix table only and give me unique values of the 'stressor' column.

```
# Get unique values of the stressor column from F_satellite matrix
exio |>
  filter(matrix == "F_satellite") |>
  distinct(stressor) |>
  collect()
```

```
# A tibble: 1,051 x 1
   stressor
   <chr>
 1 Domestic Extraction Used - Metal Ores - Other non-ferrous metal ores
 2 Domestic Extraction Used - Metal Ores - Uranium and thorium ores
 3 Domestic Extraction Used - Non-Metallic Minerals - Chemical and fertilizer m~
 4 Domestic Extraction Used - Non-Metallic Minerals - Other minerals
 5 Domestic Extraction Used - Primary Crops - Agave Fibres nes
 6 Domestic Extraction Used - Primary Crops - Almonds
 7 Domestic Extraction Used - Primary Crops - Cassava
 8 Domestic Extraction Used - Primary Crops - Chillies and peppers, dry
 9 Domestic Extraction Used - Primary Crops - Cotton Lint
10 Domestic Extraction Used - Primary Crops - Cottonseed
# i 1,041 more rows
```

How can I find which stressors are just about C02 (carbon dioxide emissions)?

```
# Filter stressors containing CO2 or carbon dioxide
co2 <- exio |>
  filter(matrix == "F_satellite") |>
  distinct(stressor) |>
  filter(grepl("CO2|carbon dioxide", stressor, ignore.case = TRUE)) |>
  collect()

co2
```

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

Identify which regions are the top 5 CO2 emmitters

```
# Get top 5 CO2 emitting regions using the co2 stressors we identified
exio |>
  filter(matrix == "F_satellite") |>
  filter(stressor %in% local(pull(co2, stressor))) |>
  group_by(region) |>
  summarize(total_co2 = sum(value, na.rm = TRUE)) |>
  arrange(desc(total_co2)) |>
  head(5) |>
  collect()
```

```
# A tibble: 5 x 2
  region total_co2
  <chr>      <dbl>
1 CN       3.87e14
2 US       2.20e14
3 IN       8.24e13
4 WM       7.95e13
5 RU       7.48e13
```

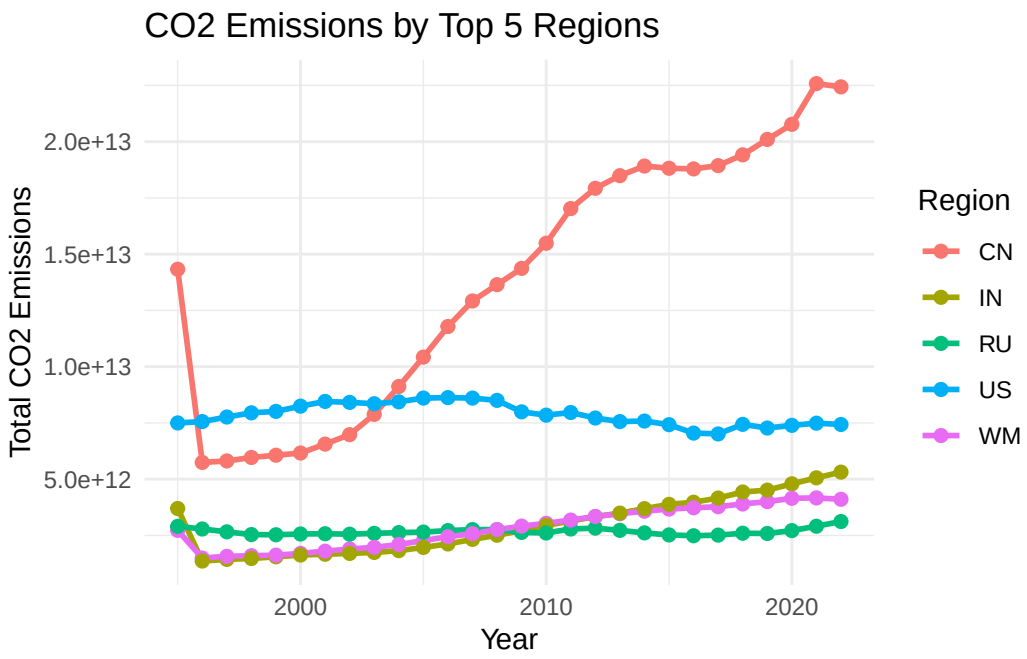
Filter the co2 data down to just these top 5 regions and plot it over the past three years

```
# Store top 5 regions
top5_regions <- exio |>
  filter(matrix == "F_satellite") |>
  filter(stressor %in% local(pull(co2, stressor))) |>
  group_by(region) |>
  summarize(total_co2 = sum(value, na.rm = TRUE)) |>
  arrange(desc(total_co2)) |>
  head(5) |>
  collect() |>
  pull(region)

# Filter for top 5 regions and plot over all years
co2_top5 <- exio |>
  filter(matrix == "F_satellite") |>
  filter(stressor %in% local(pull(co2, stressor))) |>
  filter(region %in% local(top5_regions)) |>
  group_by(region, year) |>
  summarize(total_co2 = sum(value, na.rm = TRUE)) |>
  collect()
```

```
`summarise()` has grouped output by "region". You can override using the
`.groups` argument.
```

```
# Plot
ggplot(co2_top5, aes(x = year, y = total_co2, color = region)) +
  geom_line(linewidth = 1) +
  geom_point(size = 2) +
  labs(title = "CO2 Emissions by Top 5 Regions",
       x = "Year",
       y = "Total CO2 Emissions",
       color = "Region") +
  theme_minimal()
```



going back to all regions can you identify any sectors that have had decreasing CO2 emissions over time?

```
# Calculate CO2 emissions by sector over time
sector_trends <- exio |>
  filter(matrix == "F_satellite") |>
  filter(stressor %in% local(pull(co2, stressor))) |>
  group_by(sector, year) |>
  summarize(total_co2 = sum(value, na.rm = TRUE)) |>
  collect()
```

```
`summarise()` has grouped output by "sector". You can override using the
`.groups` argument.
```

```
# Calculate the change from first to last year for each sector
sector_changes <- sector_trends |>
  group_by(sector) |>
  arrange(year) |>
  summarize(
    first_year = first(year),
    last_year = last(year),
    first_value = first(total_co2),
    last_value = last(total_co2),
```

```
    change = last_value - first_value
  ) |>
  filter(change < 0) |>
  arrange(change) |>
  head(10) |>
  select(sector, change)

# Display the top 10 sectors with largest decreases
sector_changes
```

```
# A tibble: 10 x 2
   sector                                                          change
   <chr>                                                            <dbl>
 1 Production of electricity by coal                              -
5.80e11
 2 Mining of copper ores and concentrates                         -
2.64e11
 3 Mining of coal and lignite; extraction of peat (10)            -
1.01e11
 4 Electricity by petroleum and other oil derivatives             -
9.57e10
 5 Manufacture of basic iron and steel and of ferro-alloys and first p~ -
9.27e10
 6 Manufacture of fabricated metal products, except machinery and equi~ -
7.59e10
 7 Sea and coastal water transport                                -
6.86e10
 8 Production of electricity by gas                               -
6.51e10
 9 Petroleum Refinery                                             -
6.27e10
10 Manufacture of gas; distribution of gaseous fuels through mains      -
6.10e10
```

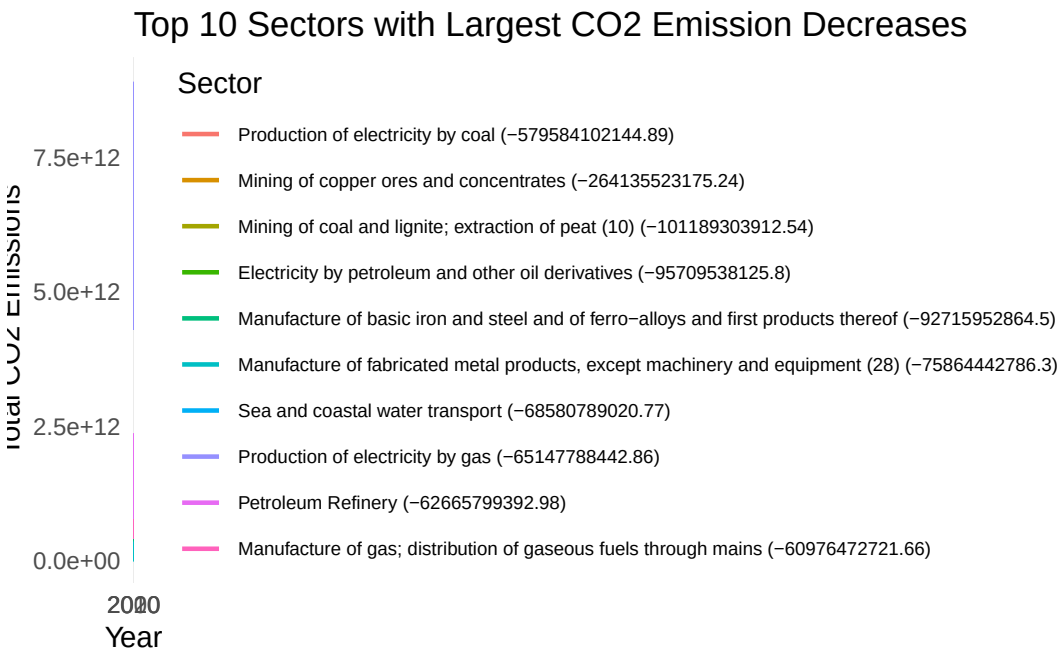
```
# Plot these top 10 sectors over time
top10_sectors <- sector_changes |> pull(sector)

sector_trends_top10 <- sector_trends |>
  filter(sector %in% top10_sectors)

# Create labels with sector name and change
```

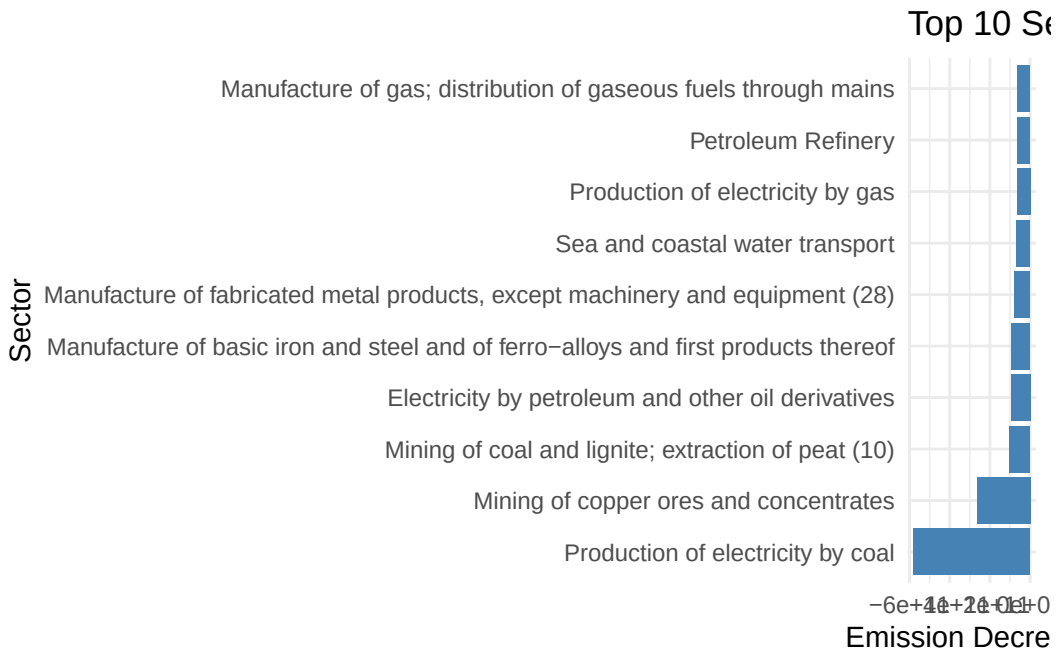
```r
labels <- paste0(sector_changes$sector, " (", round(sector_changes$change, 2), ")")

ggplot(sector_trends_top10, aes(x = year, y = total_co2, color = sector)) +
  geom_line(linewidth = 0.8) +
  labs(title = "Top 10 Sectors with Largest CO2 Emission Decreases",
       x = "Year",
       y = "Total CO2 Emissions",
       color = "Sector") +
  scale_color_discrete(labels = labels) +
  theme_minimal() +
  theme(legend.position = "right", legend.text = element_text(size = 7))
```



in a new plot, make a bar chart of the same information

```r
# Bar chart of the emission decreases
ggplot(sector_changes, aes(x = reorder(sector, change), y = change)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  labs(title = "Top 10 Sectors: CO2 Emission Decreases",
       x = "Sector",
       y = "Emission Decrease") +
  theme_minimal()
```



in a new block of code, please analyze the top 5 sectors with emissions decreases in the US ONLY

```
# Calculate CO2 emissions by sector over time in the US only
us_sector_trends <- exio |>
  filter(matrix == "F_satellite") |>
  filter(stressor %in% local(pull(co2, stressor))) |>
  filter(region == "US") |>
  group_by(sector, year) |>
  summarize(total_co2 = sum(value, na.rm = TRUE)) |>
  collect()
```

```
`summarise()` has grouped output by "sector". You can override using the
`.groups` argument.
```

```
# Calculate the change from first to last year for each sector in US
us_sector_changes <- us_sector_trends |>
  group_by(sector) |>
  arrange(year) |>
  summarize(
    first_year = first(year),
    last_year = last(year),
    first_value = first(total_co2),
```

```
    last_value = last(total_co2),
    change = last_value - first_value
  ) |>
  filter(change < 0) |>
  arrange(change) |>
  head(5) |>
  select(sector, change)

# Display the top 5 US sectors with largest decreases
us_sector_changes
```

```
# A tibble: 5 x 2
  sector                                                                change
  <chr>                                                                  <dbl>
1 Electricity by coal                                                   -
4.09e11
2 Air transport services (62)                                           -
6.56e10
3 Natural gas and services related to natural gas extraction, excludin~ -
5.68e10
4 Basic iron and steel and of ferro-alloys and first products thereof   -
4.21e10
5 Production of electricity by gas                                      -
4.21e10
```

now in a new code block, plot these top 5 sectors in the US

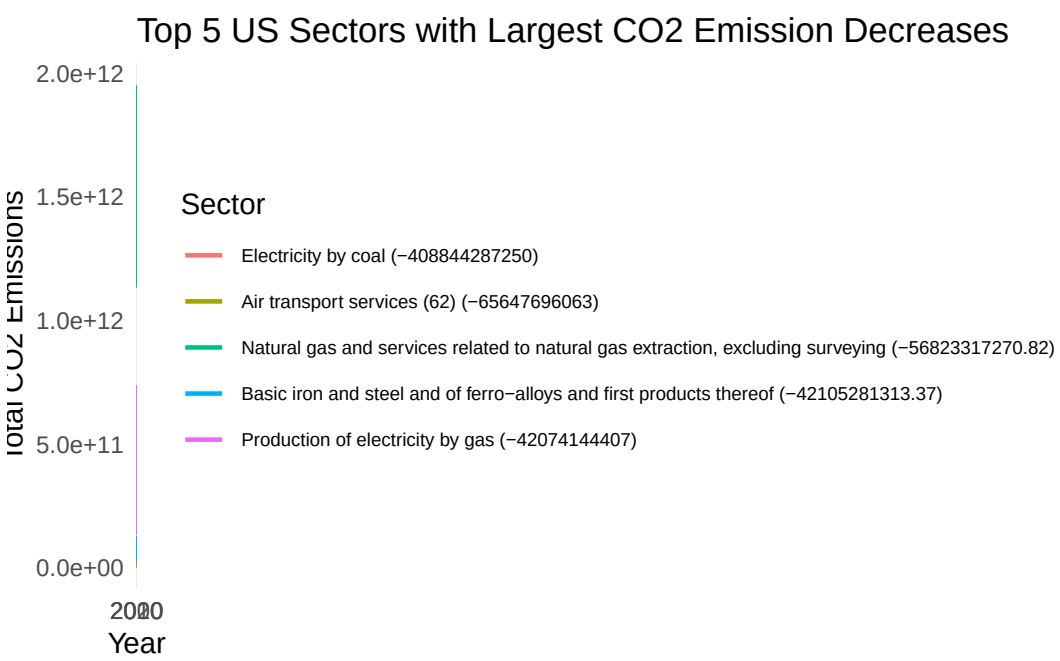
```
# Plot these top 5 US sectors over time
us_top5_sectors <- us_sector_changes |> pull(sector)

us_sector_trends_top5 <- us_sector_trends |>
  filter(sector %in% us_top5_sectors)

# Create labels with sector name and change
us_labels <- paste0(us_sector_changes$sector, " (", round(us_sector_changes$change, 2), ")")

ggplot(us_sector_trends_top5, aes(x = year, y = total_co2, color = sector)) +
  geom_line(linewidth = 0.8) +
  labs(title = "Top 5 US Sectors with Largest CO2 Emission Decreases",
       x = "Year",
       y = "Total CO2 Emissions",
```

```
       color = "Sector") +
  scale_color_discrete(labels = us_labels) +
  theme_minimal() +
  theme(legend.position = "right", legend.text = element_text(size = 7))
```



can you create a side by side plot of the 5 sectors with largest DECREASE in emissions along side of the 5 sectors with largest INCREASE (all in the US)

```
# Calculate the change from first to last year for each sector in US (including increases)
us_all_changes <- us_sector_trends |>
  group_by(sector) |>
  arrange(year) |>
  summarize(
    first_year = first(year),
    last_year = last(year),
    first_value = first(total_co2),
    last_value = last(total_co2),
    change = last_value - first_value
  )

# Top 5 decreasing sectors
us_decrease <- us_all_changes |>
```

```r
  filter(change < 0) |>
  arrange(change) |>
  head(5)

# Top 5 increasing sectors
us_increase <- us_all_changes |>
  filter(change > 0) |>
  arrange(desc(change)) |>
  head(5)

# Plot decreasing sectors
us_decrease_sectors <- us_decrease |> pull(sector)
us_trends_decrease <- us_sector_trends |> filter(sector %in% us_decrease_sectors)
decrease_labels <- paste0(us_decrease$sector, " (", round(us_decrease$change, 2), ")")

p1 <- ggplot(us_trends_decrease, aes(x = year, y = total_co2, color = sector)) +
  geom_line(linewidth = 0.8) +
  labs(title = "Top 5 US Sectors: Largest CO2 Emission Decreases",
       x = "Year",
       y = "Total CO2 Emissions") +
  scale_color_discrete(labels = decrease_labels) +
  theme_minimal() +
  theme(legend.position = "right", legend.text = element_text(size = 8), legend.title = eleme

# Plot increasing sectors
us_increase_sectors <- us_increase |> pull(sector)
us_trends_increase <- us_sector_trends |> filter(sector %in% us_increase_sectors)
increase_labels <- paste0(us_increase$sector, " (+", round(us_increase$change, 2), ")")

p2 <- ggplot(us_trends_increase, aes(x = year, y = total_co2, color = sector)) +
  geom_line(linewidth = 0.8) +
  labs(title = "Top 5 US Sectors: Largest CO2 Emission Increases",
       x = "Year",
       y = "Total CO2 Emissions") +
  scale_color_discrete(labels = increase_labels) +
  theme_minimal() +
  theme(legend.position = "right", legend.text = element_text(size = 8), legend.title = eleme

# Arrange side by side
library(patchwork)
p1 + p2
```

;t CO2 Emission Decreases

7250)

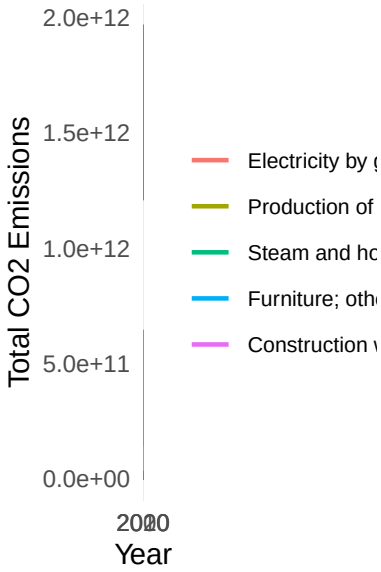
5647696063)

ed to natural gas extraction, excluding surveying (−56823317270.82)

o−alloys and first products thereof (−42105281313.37)

; (−42074144407)

Top 5 US Sectc



‘‘