# Assignment 4: Final Project

## Dataset for Visualizations

*2010-2020 Specified GHG Activities*: Greenhouse gas emissions collected by the Ontario Ministry of the Environment, Conservation and Parks, as reported by businesses and industrial facilities from 2010 to 2020: <https://files.ontario.ca/moe_mapping/downloads/1Air/GHG_by_year/GHG_Data_2010_2020_data_Dec162021.csv>

## Data Visualization with R

Chart

Description automatically generated

I used R Studio to create the above data visualization. The program is free and open source, which is important for accessibility (Zobheib 2023c). The visualization is also reproducible since people can access both the dataset on the Ontario government’s website and the R script code, which can be used to recreate the graph. I also included the source, which is important for transparency and reproducibility (Zogheib, 2023a).

The data visualization is geared toward researchers. It may also be of interest to climate change activists. I wanted to show an upward trend in methane levels measured as carbon dioxide equivalent (CO2e) tonnes over time, particularly in the distribution and transport of natural gas, which has often been marketed as a clean fuel.

There were many businesses, industrial facilities and cities in the dataset, so I narrowed the focus on one greenhouse gas in the report, and sorted the dataset in the CSV file to find the top six of over 100 industry types with the highest recorded levels. I created faceted graphs using R to show methane levels over time for each of those industries. These may not be the worst polluters based on other types of greenhouse gas emissions, but this seemed the simplest approach. Also, I selected the data based on the first six industries with the highest methane levels, whether the industry had a number of high-level records or just one record. Hence, a few of the plots did not show a clear upward trend over time. I included a trend line to present any existing upward trends more substantively.

I tried to make the visualization accessible – while also reducing the cognitive load on readers to improve its perceptual quality (Zogheib, 2023b) – by using the industry names as labels for each graph rather than the NAICS numbers assigned to each industry type (e.g., “Natural gas distribution” rather than “221210”). To further improve accessibility as well as the visualization’s aesthetic quality, I added spacing between individual graphs and increased the number of year units without overcrowding the x-axis to make the graphs easier to read. I also increased the base size of the graph, which increased the font size and the readability of the text for the visually impaired.

The visualization may influence researchers and climate change activists to look deeper into different industries’ roles in greenhouse gas emissions, especially natural gas transport and distribution, which showed an upward trend in methane CO2e levels over the last five years. It may influence people’s view and support of companies in this industry.

The ‘underwater labour’ that contributed to this final visualization product includes the people or organizations collecting the information in the dataset; the people compiling all of the data in a file; the web developers uploading data files to the Ontario government website and maintaining the site; the people who developed the R libraries and functions used to create the visualization; and the people who created online resources or references about R coding or who posed and answered questions regarding coding issues on Stack Overflow, which were extremely helpful in creating the visualization.

## Data Visualization with Python

Chart, bar chart

Description automatically generated

I chose Python to create the above data visualization because the program is also free and open source, and therefore accessible, and creating the visualization using Python code makes it reproducible. I also included the source for transparency, so that readers could search for the dataset and reproduce the visualization if they wished.

The intended audience is the general public; hence, I wanted to keep the message and information of the visualization as clear and accessible as possible. I chose to create a visualization of the total CO2e tonnes from all sources produced from 2010 to 2020 for 10 industry types. In addition to showing how much CO2e tonnes can be produced in 10 years, I wanted the visualization to show which industry types were the highest producers of greenhouse gases, presented in descending order to show the worst polluters at the top. Using Python, I summed the total CO2e amounts for each industry type and sorted the resulting dataset in descending order to find the top 10 industry polluters. The visualization does not show whether the total CO2e amounts produced increased or decreased over time, or whether companies consistently reported amounts each year, but I wanted to keep things simple to be more effective.

To make the visualization more accessible and increase its perceptual and substantive qualities, I again used the industry names as y-axis labels rather than the NAICS numbers. I increased the font size of the titles and labels, and selected a text colour for the titles that had good contrast with the light background colour, making it easier to distinguish between the graph title and the axes titles and labels (Zogheib, 2023c) and improving the visualization’s accessibility and aesthetic quality. I also used a familiar chart type (i.e., a bar graph), minimized the number of visual elements, and included the CO2e values for each industry as bar labels to help decrease the cognitive load on readers and increase its perceptual quality (Zogheib, 2023b).

The people who might be impacted by my visualization is the general public, who might be moved to learn more about climate change issues and different industries’ contribution to greenhouse gas emissions. Climate change activists might also want to look further into these industries and target the companies within them in their initiatives.

As with the visualization created in R, the ‘underwater labour’ that contributed to this final visualization product include the collectors of the information in the dataset; the people compiling all of the data in a file; the web developers uploading data files to the Ontario government website and maintaining the site; the developers of Python libraries and functions used to create the visualization; people who created online resources or references about Python coding or who posed and answered questions regarding coding issues online; and the instructors for the Data Science Institute courses, whose class slides I made much use of in creating both visualizations.

## Citations

Zogheib, C. (2023a). *First steps: reproducible data visualization*. Data Visualization [PowerPoint presentation]. Toronto: Data Sciences Institute, University of Toronto.

Zogheib, C. (2023b). *Graphing our data: choosing the right visualization*. Data Visualization [PowerPoint presentation]. Toronto: Data Sciences Institute, University of Toronto.

Zogheib, C. (2023c). *Visualization with purpose: accessible data visualization*. Data Visualization [PowerPoint presentation]. Toronto: Data Sciences Institute, University of Toronto.

## Appendix A: R Script

#load libraries

library(ggplot2)

library(tidyverse)

#import greenhouse gas emissions .csv file

ghg.data <- read.csv("C:\\Users\\smych\\Documents\\Continuing Studies\\DSI Data Science Certificate Program\\Data Visualization\\Assignments\\GHG\_Data\_2010\_2020\_data\_Dec162021.csv", header = TRUE)

str(ghg.data)

#look at dataset

str(ghg.data)

head(ghg.data)

#convert Year variable to date-time type

library(lubridate)

Year <- as.Date(as.character(ghg.data$Year), format = "%Y")

#filter data by five facility NAICS code associated with highest methane levels

ghg.top5naics <-

ghg.data |>

filter(Facility.Primary.NAICS.Code %in% c(221210, 562210, 486210, 322121,

331110, 322112))

#change labels for NAICS codes:

#Source: https://ggplot2.tidyverse.org/reference/labeller.html

naics\_names <- c(

'221210' = "Natural gas distribution",

'562210' = "Waste treatment and disposal",

'486210' = "Pipeline transportation of natural gas",

'322121' = "Paper (except newsprint) mills",

'331110' = "Iron and steel mills manufacturing",

'322112' = "Chemical pulp mills"

)

#plot methane levels by years

#sources: https://stackoverflow.com/questions/38722202/how-do-i-change-the-number-of-decimal-places-on-axis-labels-in-ggplot2

#https://www.tutorialspoint.com/increase-the-space-between-facets-in-a-facetted-plot-created-using-ggplot2-in-r

#https://r-graph-gallery.com/279-plotting-time-series-with-ggplot2.html

#https://stackoverflow.com/questions/14942681/change-size-of-axes-title-and-labels-in-ggplot2

p <- ggplot(data = ghg.top5naics,

mapping = aes(x = Year,

y = Methane..CH4..in.CO2e..t.,

na.rm = TRUE))

p + geom\_line(alpha=0.3, size = 0.5) +

geom\_smooth(size = 0.75, method="loess", se=FALSE) +

facet\_wrap(~ Facility.Primary.NAICS.Code, ncol = 3, labeller = labeller(Facility.Primary.NAICS.Code = naics\_names)) +

scale\_x\_date(date\_breaks = "2 years", date\_labels = "%Y") +

scale\_y\_continuous(labels = scales::comma) +

theme(panel.spacing=unit(1,"lines")) + #add space between facet grid panels

theme\_grey(base\_size = 15) +

labs(x = "Year",

y = "Methane in CO2 equivalent (tonnes)",

title = "Methane in CO2 equivalent tonnes reported by Ontario companies, 2010-2020",

caption = "Source: Ministry of the Environment, Conservation and Parks. (2020). 2010-2020 Specified GHG Activities.")

### Appendix B: Python Code

*#load libraries*

**import** numpy **as** np

**import** pandas **as** pd

**import** matplotlib.pyplot **as** plt

**import** matplotlib.ticker **as** tick

In [68]:

*#import Ontario greenhouse gas emissions report csv file as single dataframe*

ghg\_data **=** pd**.**read\_csv('~/Documents/Continuing Studies/DSI Data Science Certificate Program/Data Visualization/Assignments/GHG\_Data\_2010\_2020\_data\_Dec162021\_UTF8.csv')

In [69]:

*#view dataset*

ghg\_data**.**columns

Out[69]:

Index(['Year', 'Ontario GHG ID', 'Facility Owner', 'Facility Name',

'Facility City', 'Facility Primary NAICS Code',

'Carbon dioxide (CO2) from non-biomass in CO2e (t)',

'Carbon dioxide (CO2) from biomass in CO2e (t)',

'Methane (CH4) in CO2e (t)', 'Nitrous oxide (N2O) in CO2e (t)',

'Sulphur hexafluoride (SF6) in CO2e (t)',

'Hydrofluorocarbons (HFCs) in CO2e (t)',

'Perfluorocarbons (PFCs) in CO2e (t)',

'Nitrogen Trifluoride (NF3) in CO2e (t)',

'Total CO2e from all sources in CO2e (t)',

'Reporting Amount in CO2e (t)', 'Verification Amount in CO2e (t)',

'Accredited Verification Body'],

dtype='object')

In [70]:

ghg\_data**.**dtypes

Out[70]:

Year int64

Ontario GHG ID int64

Facility Owner object

Facility Name object

Facility City object

Facility Primary NAICS Code int64

Carbon dioxide (CO2) from non-biomass in CO2e (t) float64

Carbon dioxide (CO2) from biomass in CO2e (t) float64

Methane (CH4) in CO2e (t) float64

Nitrous oxide (N2O) in CO2e (t) float64

Sulphur hexafluoride (SF6) in CO2e (t) float64

Hydrofluorocarbons (HFCs) in CO2e (t) float64

Perfluorocarbons (PFCs) in CO2e (t) float64

Nitrogen Trifluoride (NF3) in CO2e (t) float64

Total CO2e from all sources in CO2e (t) int64

Reporting Amount in CO2e (t) int64

Verification Amount in CO2e (t) float64

Accredited Verification Body object

dtype: object

In [71]:

*#create pivot table to get sum of total CO2e levels for each type of facility*

*#source: https://www.statology.org/pandas-pivot-table-sum/*

ghg\_data\_pivot **=** pd**.**pivot\_table(ghg\_data, values**=**'Total CO2e from all sources in CO2e (t)', index**=**'Facility Primary NAICS Code', aggfunc**=**'sum')

*#view first 5 rows of pivot table*

ghg\_data\_pivot**.**head

Out[71]:

<bound method NDFrame.head of Total CO2e from all sources in CO2e (t)

Facility Primary NAICS Code

111419 3798099

111422 130579

211110 347226

211113 1284168

212220 1848598

... ...

622112 363331

721120 70513

812320 65539

911110 177575

911910 146459

[109 rows x 1 columns]>

In [72]:

*#sort by descending order*

ghg\_data\_pivot **=** ghg\_data\_pivot**.**sort\_values(by**=**'Total CO2e from all sources in CO2e (t)', ascending**=False**)

ghg\_data\_pivot**.**head(10)

Out[72]:

|  | **Total CO2e from all sources in CO2e (t)** |
| --- | --- |
| **Facility Primary NAICS Code** |  |
| **331110** | 132141209 |
| **221112** | 101900902 |
| **327310** | 50758765 |
| **324110** | 45771289 |
| **322112** | 33820412 |
| **325110** | 17915504 |
| **322121** | 15185543 |
| **221119** | 13894399 |
| **562210** | 10469491 |
| **327410** | 10446303 |

In [188]:

*#sources: https://www.python-graph-gallery.com/matplotlib/*

*#https://www.python-graph-gallery.com/191-custom-axis-on-matplotlib-chart*

*#https://stackoverflow.com/questions/14088687/how-to-change-plot-background-color*

*#https://www.tutorialspoint.com/how-do-i-change-the-font-size-of-the-scale-in-matplotlib-plots*

*#https://pythonguides.com/matplotlib-invert-y-axis/*

*#https://stackoverflow.com/questions/70515542/adding-comma-to-bar-labels?rq=1*

*#data to plot*

totalCO2e **=** ghg\_data\_pivot['Total CO2e from all sources in CO2e (t)'][:10]

naics **=** ("Iron and steel mills manufacturing", "Fossil fuel electric power generation", "Cement manufacturing", "Petroleum refineries", "Chemical pulp mills", "Petrochemical manufacturing", "Paper (except newsprint) mills", "Other electric power generation", "Waste treatment and disposal", "Lime manufacturing")

*#create horizontal bar plot*

bar\_fig, bar\_ax **=** plt**.**subplots()

bar\_fig**.**set\_size\_inches(20,12)

bar\_fig**.**patch**.**set\_facecolor('#ffffe4')

bar\_ax**.**barh(naics, totalCO2e)

bar\_ax**.**invert\_yaxis() *#put in descending order*

bar\_ax**.**xaxis**.**set\_major\_formatter(tick**.**StrMethodFormatter('{x:,.0f}'))

bar\_ax**.**set\_axisbelow(**True**)

bar\_ax**.**grid(alpha**=**0.3)

bar\_ax**.**tick\_params(axis**=**'y', labelsize**=**15)

bar\_ax**.**tick\_params(axis**=**'x', labelsize**=**14)

**for** c **in** ax**.**containers:

bar\_ax**.**bar\_label(c, labels**=**[f'{x:,.0f}' **for** x **in** c**.**datavalues], fontsize**=**'12') *#adds bar labels and commas*

bar\_ax**.**set\_title('Top 10 industries reporting the highest total carbon dioxide equivalent \nemissions from all sources in Ontario from 2010 to 2020 \n', fontweight**=**'bold', color **=** 'brown', fontsize**=**'24')

bar\_ax**.**set\_xlabel('\n Total CO2 equivalent from all sources in CO2e (tonnes)', fontweight**=**'bold', color **=** 'brown', fontsize**=**'18')

plt**.**annotate('Source: Ministry of the Environment, Conservation and Parks. (2020). 2010-2020 Specified GHG Activities.', (0,0), (450,**-**100), fontsize**=**13,

xycoords**=**'axes fraction', textcoords**=**'offset points', va**=**'bottom')

plt**.**tight\_layout() *#prevents trimming of graph*

plt**.**savefig('TotalCO2eBarGraph.png')

Chart, bar chart

Description automatically generated