# Particulate Matter (PM2.5)

Ricky Truong

June 2025

## Load libraries and data

```r
# Delete everything in environment
rm(list = ls())

# Load libraries
library(tidyverse)
library(readxl)
library(readr)

# Load EIA Total CO2 Emission, 1970-2022
co2_1970_2022 <- read_excel("CO2.xlsx")

# Load EPA Total PM2.5 Emission, 1990-2024
pm2.5_1990_2024 <- read.csv("national_tier1_caps_21feb2025.xlsx - PM25Primary.csv")

# Load EPA Average PM2.5 Concentration, 2000-2023
pm2.5_2000_2023 <- read.csv("PM25National.csv")

# Load NASA (Monthly) Average PM2.5 Concentration, 1980-2022 (Unweighted)
pm2.5_1980_2022 <- read.csv("MERRA2.avgM_2d_pm25_admin0x.v01.19800101-20221231.csv",
                            skip = 13, header = FALSE)
```

## EIA total CO2 emission, 1970-2022

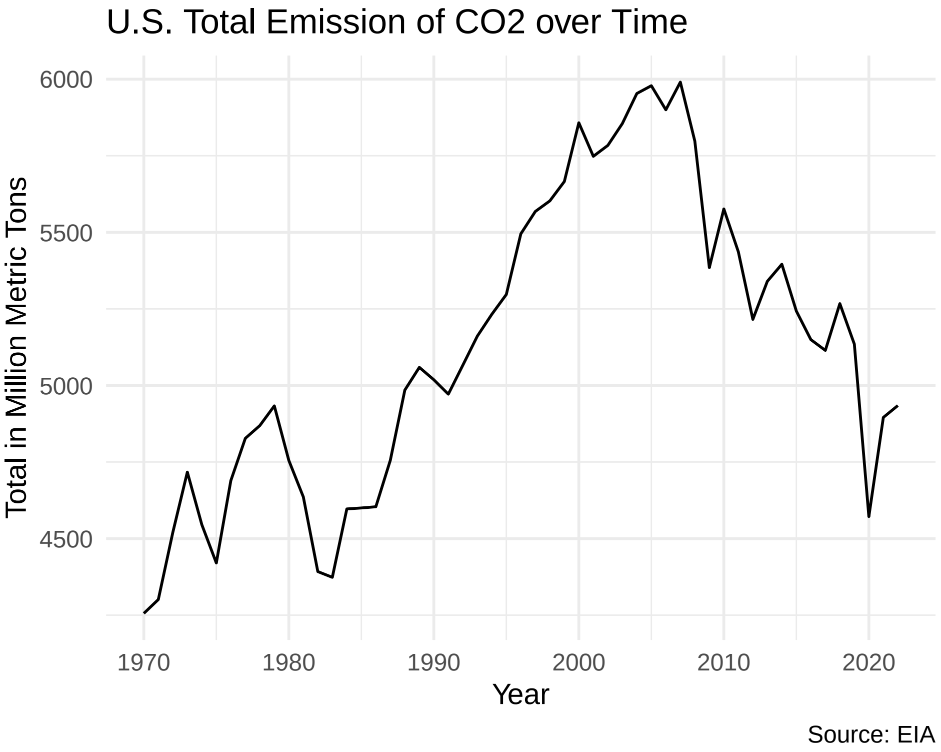
### Wrangle data

```r
# Delete (unnecessary) last four columns
co2_1970_2022 <- subset(co2_1970_2022, select = -c(...55 : ...58))

# Rename columns of data set
years <- seq(1970, 2022)
cols <- append(years, "variable", 0)
names(co2_1970_2022) <- cols
```

```
# Convert data to be in long/tidy format with correct values
co2_1970_2022 <- co2_1970_2022 %>%
  filter(variable == "Total of states") %>%
  pivot_longer(cols = -variable,
               names_to = "Year",
               values_to = "CO2") %>%
  mutate(Year = as.integer(Year),
         CO2 = as.numeric(gsub(",", "", CO2))) %>%
  select(Year, CO2)
```

### Visualize data

```
# Graph a line plot
ggplot(co2_1970_2022, aes(x = Year,
                          y = CO2)) +
  geom_line() +
  labs(x = "Year",
       y = "Total in Million Metric Tons",
       title = "U.S. Total Emission of CO2 over Time",
       caption = "Source: EIA") +
  theme_minimal()
```



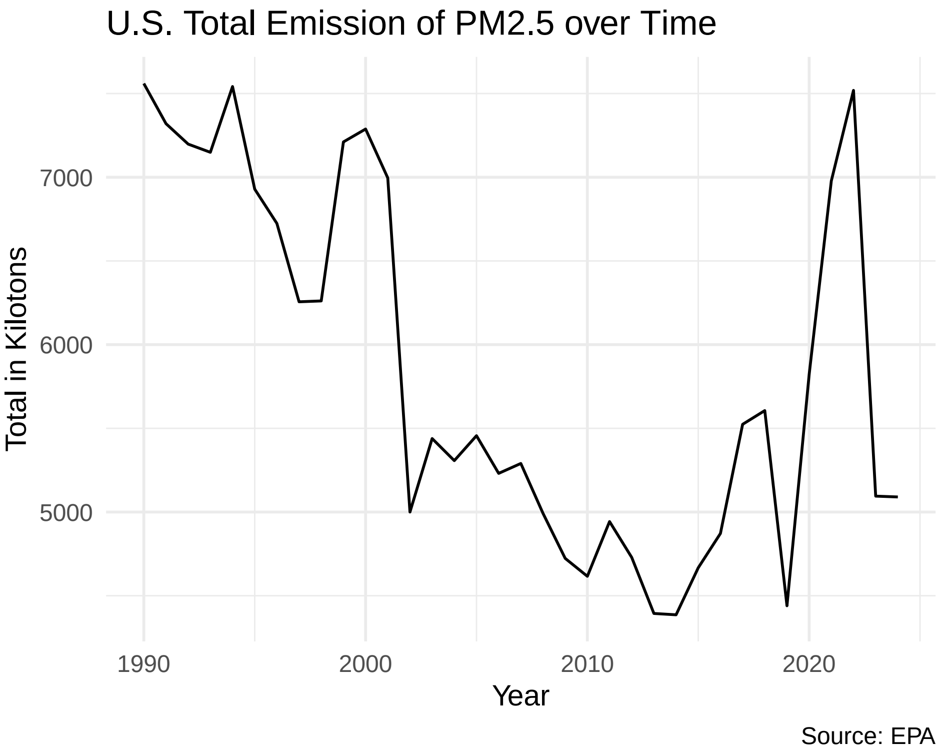
## EPA total PM2.5 emission, 1990-2024

### Wrangle data

```r
# Rename columns of data set
years <- seq(1990, 2024)
cols <- append(years, "variable", 0)
names(pm2.5_1990_2024) <- cols

# Convert data to be in long/tidy format with correct values
pm2.5_1990_2024 <- pm2.5_1990_2024 %>%
  filter(variable == "Total") %>%
  pivot_longer(cols = -variable,
               names_to = "Year",
               values_to = "PM2.5") %>%
  mutate(Year = as.integer(Year),
         PM2.5 = as.numeric(gsub(",", "", PM2.5))) %>%
  select(Year, PM2.5)
```

### Visualize data

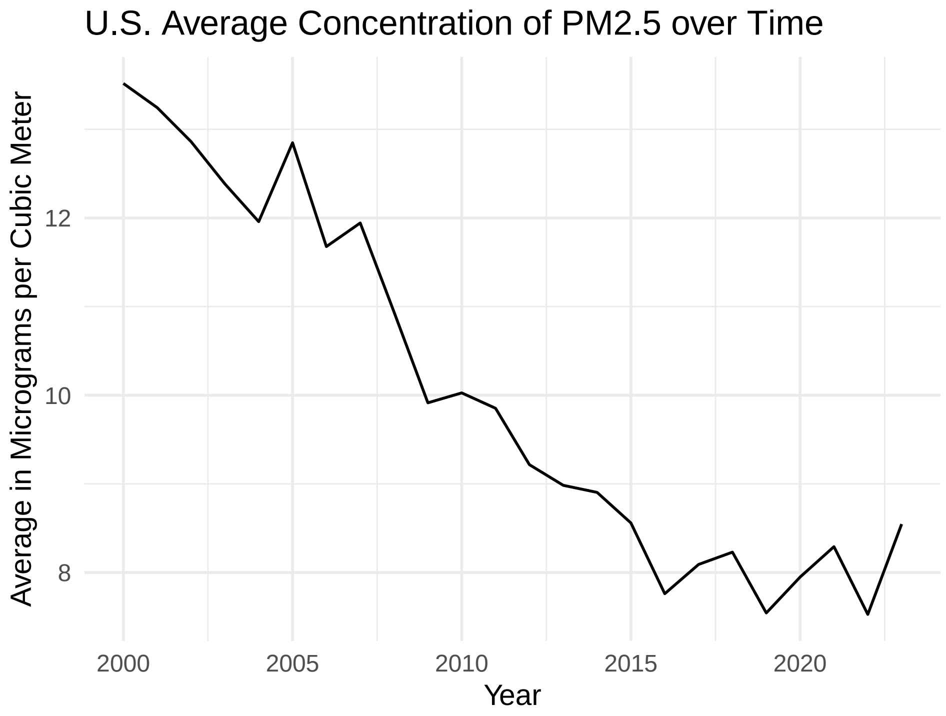
```r
# Graph a line plot
ggplot(pm2.5_1990_2024, aes(x = Year,
                            y = PM2.5)) +
  geom_line() +
  labs(x = "Year",
       y = "Total in Kilotons",
       title = "U.S. Total Emission of PM2.5 over Time",
       caption = "Source: EPA") +
  theme_minimal()
```



## EPA average PM2.5 concentration, 2000-2023

### Visualize data

```
# Graph a line plot
ggplot(pm2.5_2000_2023, aes(x = Year,
                            y = Mean)) +
  geom_line() +
  labs(x = "Year",
       y = "Average in Micrograms per Cubic Meter",
       title = "U.S. Average Concentration of PM2.5 over Time",
       caption = "Source: EPA") +
  theme_minimal()
```



Source: EPA

## NASA (monthly) average PM2.5 concentration, 1980-2022 (unweighted)

### Wrangle data

```
# Subset data for United States only
col_index <- which(pm2.5_1980_2022[1, ] == "United_States")
colname <- names(pm2.5_1980_2022)[col_index]
pm2.5_1980_2022 <- pm2.5_1980_2022 %>%
  select(V1, colname)
```

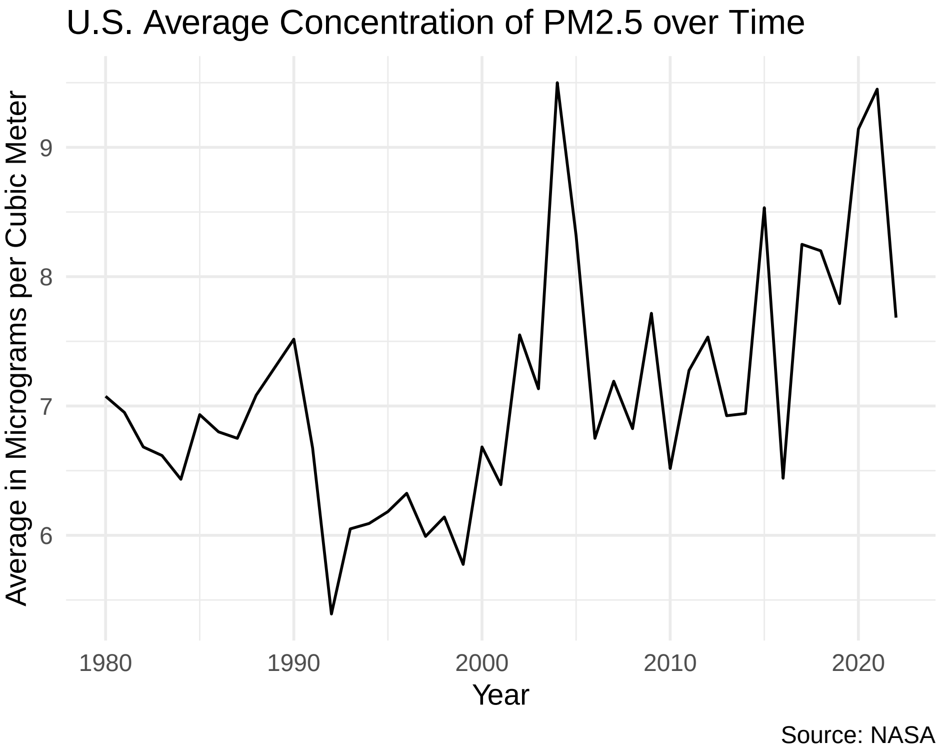
```
## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use `all_of()` or `any_of()` instead.
##   # Was:
##   data %>% select(colname)
##
##   # Now:
##   data %>% select(all_of(colname))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
# Rename columns appropriately and delete unnecessary row
cols <- c("Date", "Mean")
```

```
names(pm2.5_1980_2022) <- cols
pm2.5_1980_2022 <- pm2.5_1980_2022[-c(1),]

# Calculate annual mean
pm2.5_1980_2022_annual <- pm2.5_1980_2022 %>%
  mutate(Year = substr(Date, 1, 4),
         Mean = as.numeric(Mean)) %>%
  group_by(Year) %>%
  summarise(Annual_Mean = mean(Mean, na.rm = TRUE))
```

### Visualize data

```
# Graph a line plot (from annual data)
ggplot(pm2.5_1980_2022_annual, aes(x = as.numeric(Year),
                                   y = Annual_Mean)) +
  geom_line() +
  labs(x = "Year",
       y = "Average in Micrograms per Cubic Meter",
       title = "U.S. Average Concentration of PM2.5 over Time",
       caption = "Source: NASA") +
  theme_minimal()
```



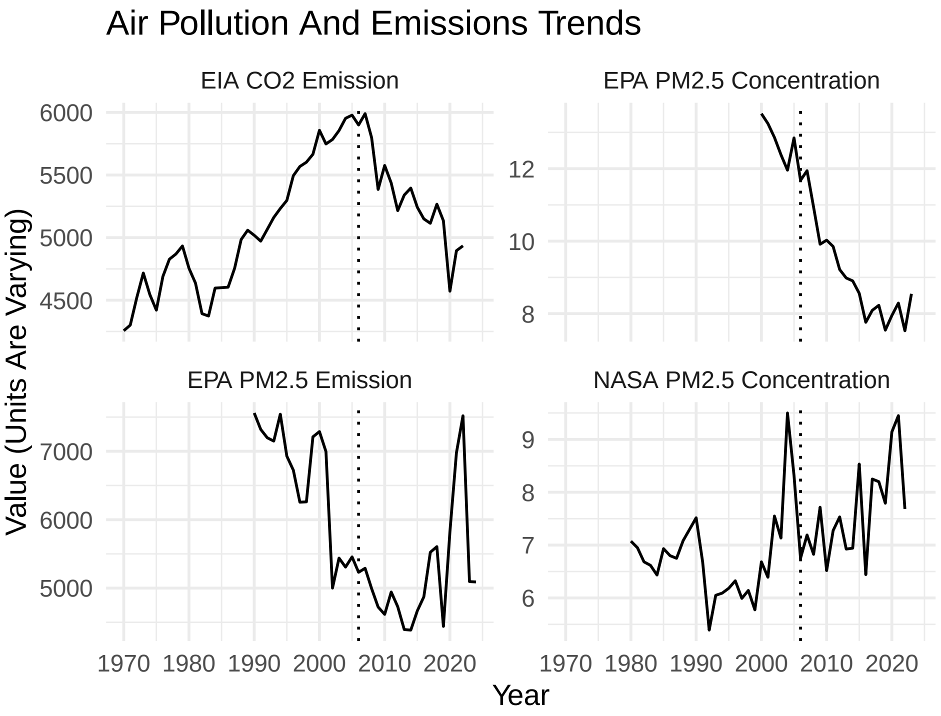
## Graph everything

### Wrangle data

```r
# Standardize each of the four data sets
eia_co2 <- co2_1970_2022 %>%
  mutate(Year = as.numeric(Year),
         Source = "EIA CO2 Emission",
         Value = CO2) %>%
  select(Year, Value, Source)

epa_emission <- pm2.5_1990_2024 %>%
  mutate(Year = as.numeric(Year),
         Source = "EPA PM2.5 Emission",
         Value = PM2.5) %>%
  select(Year, Value, Source)

epa_concentration <- pm2.5_2000_2023 %>%
  mutate(Year = as.numeric(Year),
         Source = "EPA PM2.5 Concentration",
         Value = Mean) %>%
  select(Year, Value, Source)

nasa_concentration <- pm2.5_1980_2022_annual %>%
  mutate(Year = as.numeric(Year),
         Source = "NASA PM2.5 Concentration",
         Value = Annual_Mean) %>%
  select(Year, Value, Source)

# Combine data sets into one
combined_pm_data <- bind_rows(eia_co2, epa_emission, epa_concentration, nasa_concentration)
```

### Visualize data

```r
# Graph line plots faceted
ggplot(combined_pm_data, aes(x = Year,
                             y = Value)) +
  geom_line() +
  geom_vline(xintercept = 2006, linetype = "dotted") +
  facet_wrap(~ Source, scales = "free_y") +
  labs(x = "Year",
       y = "Value (Units Are Varying)",
       title = "Air Pollution And Emissions Trends",
       caption = "Sources: EIA, EPA, NASA") +
  theme_minimal()
```



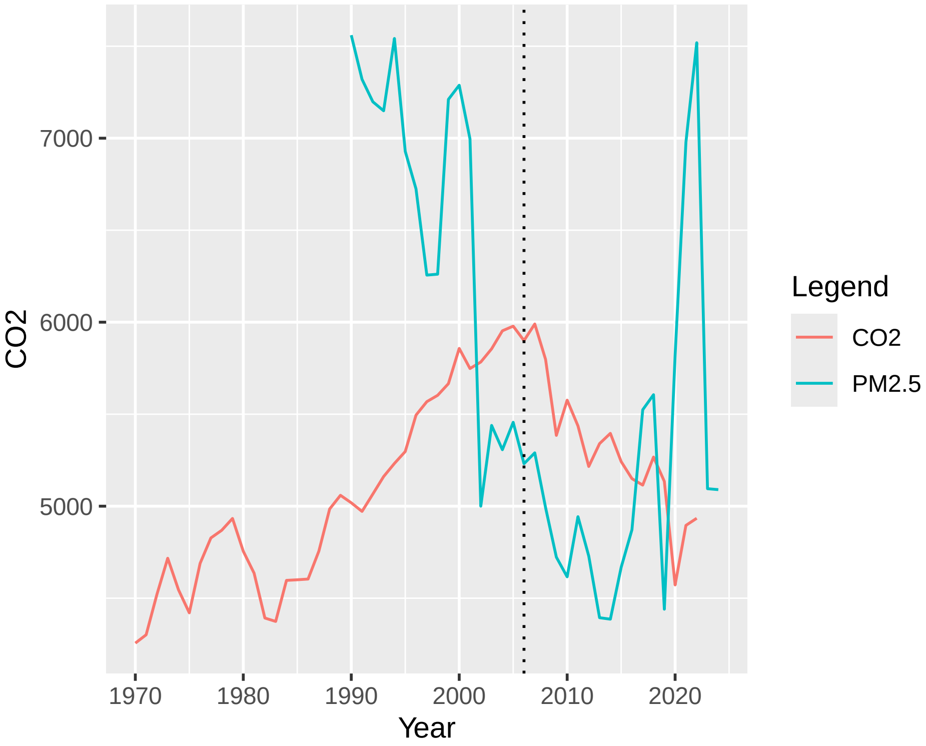
Sources: EIA, EPA, NASA

### Visualize EIA CO2 Emission and EPA PM2.5 Emission

```r
new <- full_join(co2_1970_2022, pm2.5_1990_2024, join_by("Year" == "Year"))

ggplot(new, aes(x = Year)) +
  geom_line(aes(y = CO2, color = "CO2")) +
  geom_line(aes(y = PM2.5, color = "PM2.5")) +
  geom_vline(xintercept = 2006, linetype = "dotted") +
  labs(color = "Legend")
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_line()`).
```

```
## Warning: Removed 20 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



7000
6000
5000
CO2
1970
1980
1990
2000
2010
2020
Year
Legend
CO2
PM2.5