hw1

AUTHOR
Andre Gala-Garza

Disclaimer: Generative AI was used to assist with templating and writing code in this assignment; however, this code was checked manually and edited by hand to ensure accuracy.

Source: OpenAI. (2026). *ChatGPT (GPT-5.2 Thinking)* [Large language model]. https://chatgpt.com/.

1. Exploratory data analysis

Define question

We will work with air pollution data from the U.S. Environmental Protection Agency (EPA). The EPA has a national monitoring network of air pollution sites that measure particulate matter (PM) concentrations. The primary question we will answer is whether daily concentrations of PM (particulate matter air pollution with aerodynamic diameter less than 2.5 µm) decreased in Michigan over the 20 years spanning from 2005 to 2025.

EDA Step 1: Download and read data

Please note that the data download link as given in the assignment description is currently not loading properly, so I have instead accessed the data directly.

Here is the data repository from the EPA Air Quality Data website:
https://aqs.epa.gov/aqsweb/airdata/download_files.html

The 2005 and 2025 PM2.5 data for all sites in Michigan were downloaded from these links:

https://aqs.epa.gov/aqsweb/airdata/daily_88101_2005.zip

https://aqs.epa.gov/aqsweb/airdata/daily_88101_2025.zip

```
pm_2005 <- read.csv("daily_88101_2005.csv")
pm_2025 <- read.csv("daily_88101_2025.csv")
```

EDA Step 2: Check data size

```
dim(pm_2005)
```

```
[1] 145913     29
```

```r
dim(pm_2025)
```

```
[1] 466760     29
```

EDA Step 3: Examine variables and their types

```r
names(pm_2005)
```

```
 [1] "State.Code"          "County.Code"         "Site.Num"
 [4] "Parameter.Code"      "POC"                 "Latitude"
 [7] "Longitude"           "Datum"               "Parameter.Name"
[10] "Sample.Duration"     "Pollutant.Standard"  "Date.Local"
[13] "Units.of.Measure"    "Event.Type"          "Observation.Count"
[16] "Observation.Percent" "Arithmetic.Mean"     "X1st.Max.Value"
[19] "X1st.Max.Hour"       "AQI"                 "Method.Code"
[22] "Method.Name"         "Local.Site.Name"     "Address"
[25] "State.Name"          "County.Name"         "City.Name"
[28] "CBSA.Name"           "Date.of.Last.Change"
```

```r
names(pm_2025)
```

```
 [1] "State.Code"          "County.Code"         "Site.Num"
 [4] "Parameter.Code"      "POC"                 "Latitude"
 [7] "Longitude"           "Datum"               "Parameter.Name"
[10] "Sample.Duration"     "Pollutant.Standard"  "Date.Local"
[13] "Units.of.Measure"    "Event.Type"          "Observation.Count"
[16] "Observation.Percent" "Arithmetic.Mean"     "X1st.Max.Value"
[19] "X1st.Max.Hour"       "AQI"                 "Method.Code"
[22] "Method.Name"         "Local.Site.Name"     "Address"
[25] "State.Name"          "County.Name"         "City.Name"
[28] "CBSA.Name"           "Date.of.Last.Change"
```

```r
str(pm_2005)
```

```
'data.frame':   145913 obs. of  29 variables:
 $ State.Code         : chr  "01" "01" "01" "01" ...
 $ County.Code        : int  3 3 3 3 3 3 3 3 3 3 ...
 $ Site.Num           : int  10 10 10 10 10 10 10 10 10 10 ...
 $ Parameter.Code     : int  88101 88101 88101 88101 88101 88101 88101 88101 88101 88101 ...
 $ POC                : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Latitude           : num  30.5 30.5 30.5 30.5 30.5 ...
 $ Longitude          : num  -87.9 -87.9 -87.9 -87.9 -87.9 ...
 $ Datum              : chr  "NAD83" "NAD83" "NAD83" "NAD83" ...
 $ Parameter.Name     : chr  "PM2.5 - Local Conditions" "PM2.5 - Local Conditions" "PM2.5 - Local
Conditions" "PM2.5 - Local Conditions" ...
 $ Sample.Duration    : chr  "24 HOUR" "24 HOUR" "24 HOUR" "24 HOUR" ...
 $ Pollutant.Standard : chr  "PM25 24-hour 2012" "PM25 24-hour 2012" "PM25 24-hour 2012" "PM25
```

```
24-hour 2012" ...
 $ Date.Local         : chr  "2005-01-01" "2005-01-04" "2005-01-07" "2005-01-10" ...
 $ Units.of.Measure   : chr  "Micrograms/cubic meter (LC)" "Micrograms/cubic meter (LC)"
"Micrograms/cubic meter (LC)" "Micrograms/cubic meter (LC)" ...
 $ Event.Type         : chr  "None" "None" "None" "None" ...
 $ Observation.Count  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Observation.Percent: num  100 100 100 100 100 100 100 100 100 100 ...
 $ Arithmetic.Mean    : num  9 7.7 7.3 8.7 3.3 7.7 16.6 7.9 15.1 23.4 ...
 $ X1st.Max.Value     : num  9 7.7 7.3 8.7 3.3 7.7 16.6 7.9 15.1 23.4 ...
 $ X1st.Max.Hour      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AQI                : int  50 43 41 48 18 43 65 44 62 78 ...
 $ Method.Code        : int  118 118 118 118 118 118 118 118 118 118 ...
 $ Method.Name        : chr  "R & P Model 2025 PM2.5 Sequential w/WINS - GRAVIMETRIC" "R & P
Model 2025 PM2.5 Sequential w/WINS - GRAVIMETRIC" "R & P Model 2025 PM2.5 Sequential w/WINS -
GRAVIMETRIC" "R & P Model 2025 PM2.5 Sequential w/WINS - GRAVIMETRIC" ...
 $ Local.Site.Name    : chr  "FAIRHOPE, Alabama" "FAIRHOPE, Alabama" "FAIRHOPE, Alabama"
"FAIRHOPE, Alabama" ...
 $ Address            : chr  "FAIRHOPE HIGH SCHOOL, 1 PIRATE DRIVE, FAIRHOPE,  ALABAMA" "FAIRHOPE
HIGH SCHOOL, 1 PIRATE DRIVE, FAIRHOPE,  ALABAMA" "FAIRHOPE HIGH SCHOOL, 1 PIRATE DRIVE, FAIRHOPE,
ALABAMA" "FAIRHOPE HIGH SCHOOL, 1 PIRATE DRIVE, FAIRHOPE,  ALABAMA" ...
 $ State.Name         : chr  "Alabama" "Alabama" "Alabama" "Alabama" ...
 $ County.Name        : chr  "Baldwin" "Baldwin" "Baldwin" "Baldwin" ...
 $ City.Name          : chr  "Fairhope" "Fairhope" "Fairhope" "Fairhope" ...
 $ CBSA.Name          : chr  "Daphne-Fairhope-Foley, AL" "Daphne-Fairhope-Foley, AL" "Daphne-
Fairhope-Foley, AL" "Daphne-Fairhope-Foley, AL" ...
 $ Date.of.Last.Change: chr  "2024-09-05" "2024-09-05" "2024-09-05" "2024-09-05" ...
```

```r
str(pm_2025)
```

```
'data.frame':	466760 obs. of  29 variables:
 $ State.Code         : int  1 1 1 1 1 1 1 1 1 1 ...
 $ County.Code        : int  3 3 3 3 3 3 3 3 3 3 ...
 $ Site.Num           : int  10 10 10 10 10 10 10 10 10 10 ...
 $ Parameter.Code     : int  88101 88101 88101 88101 88101 88101 88101 88101 88101 88101 ...
 $ POC                : int  3 3 3 3 3 3 3 3 3 3 ...
 $ Latitude           : num  30.5 30.5 30.5 30.5 30.5 ...
 $ Longitude          : num  -87.9 -87.9 -87.9 -87.9 -87.9 ...
 $ Datum              : chr  "NAD83" "NAD83" "NAD83" "NAD83" ...
 $ Parameter.Name     : chr  "PM2.5 - Local Conditions" "PM2.5 - Local Conditions" "PM2.5 - Local
Conditions" "PM2.5 - Local Conditions" ...
 $ Sample.Duration    : chr  "1 HOUR" "1 HOUR" "1 HOUR" "1 HOUR" ...
 $ Pollutant.Standard : chr  "" "" "" "" ...
 $ Date.Local         : chr  "2025-01-01" "2025-01-02" "2025-01-03" "2025-01-04" ...
 $ Units.of.Measure   : chr  "Micrograms/cubic meter (LC)" "Micrograms/cubic meter (LC)"
"Micrograms/cubic meter (LC)" "Micrograms/cubic meter (LC)" ...
 $ Event.Type         : chr  "None" "None" "None" "None" ...
 $ Observation.Count  : int  24 24 24 24 24 23 24 24 24 24 ...
 $ Observation.Percent: num  100 100 100 100 100 96 100 100 100 100 ...
 $ Arithmetic.Mean    : num  3.62 6.79 9.96 5.5 5.5 ...
 $ X1st.Max.Value     : num  19 26 25 11 12 16 5 14 35 16 ...
```

```
 $ X1st.Max.Hour      : int  21 23 1 19 14 2 9 19 22 0 ...
 $ AQI                : int  NA NA NA NA NA NA NA NA NA NA ...
 $ Method.Code        : int  209 209 209 209 209 209 209 209 209 209 ...
 $ Method.Name        : chr  "Met One BAM-1022 Mass Monitor w/ VSCC or TE-PM2.5C - Beta
Attenuation" "Met One BAM-1022 Mass Monitor w/ VSCC or TE-PM2.5C - Beta Attenuation" "Met One
BAM-1022 Mass Monitor w/ VSCC or TE-PM2.5C - Beta Attenuation" "Met One BAM-1022 Mass Monitor w/
VSCC or TE-PM2.5C - Beta Attenuation" ...
 $ Local.Site.Name    : chr  "FAIRHOPE, Alabama" "FAIRHOPE, Alabama" "FAIRHOPE, Alabama"
"FAIRHOPE, Alabama" ...
 $ Address            : chr  "FAIRHOPE HIGH SCHOOL, 1 PIRATE DRIVE, FAIRHOPE,  ALABAMA" "FAIRHOPE
HIGH SCHOOL, 1 PIRATE DRIVE, FAIRHOPE,  ALABAMA" "FAIRHOPE HIGH SCHOOL, 1 PIRATE DRIVE, FAIRHOPE,
ALABAMA" "FAIRHOPE HIGH SCHOOL, 1 PIRATE DRIVE, FAIRHOPE,  ALABAMA" ...
 $ State.Name         : chr  "Alabama" "Alabama" "Alabama" "Alabama" ...
 $ County.Name        : chr  "Baldwin" "Baldwin" "Baldwin" "Baldwin" ...
 $ City.Name          : chr  "Fairhope" "Fairhope" "Fairhope" "Fairhope" ...
 $ CBSA.Name          : chr  "Daphne-Fairhope-Foley, AL" "Daphne-Fairhope-Foley, AL" "Daphne-
Fairhope-Foley, AL" "Daphne-Fairhope-Foley, AL" ...
 $ Date.of.Last.Change: chr  "2025-10-29" "2025-10-29" "2025-10-29" "2025-10-29" ...
```

EDA Step 4: Look at top/bottom of data

```
head(pm_2005)
```

```
  State.Code County.Code Site.Num Parameter.Code POC Latitude Longitude Datum
1         01           3       10          88101   1 30.49748 -87.88026 NAD83
2         01           3       10          88101   1 30.49748 -87.88026 NAD83
3         01           3       10          88101   1 30.49748 -87.88026 NAD83
4         01           3       10          88101   1 30.49748 -87.88026 NAD83
5         01           3       10          88101   1 30.49748 -87.88026 NAD83
6         01           3       10          88101   1 30.49748 -87.88026 NAD83
            Parameter.Name Sample.Duration Pollutant.Standard Date.Local
1 PM2.5 - Local Conditions         24 HOUR  PM25 24-hour 2012 2005-01-01
2 PM2.5 - Local Conditions         24 HOUR  PM25 24-hour 2012 2005-01-04
3 PM2.5 - Local Conditions         24 HOUR  PM25 24-hour 2012 2005-01-07
4 PM2.5 - Local Conditions         24 HOUR  PM25 24-hour 2012 2005-01-10
5 PM2.5 - Local Conditions         24 HOUR  PM25 24-hour 2012 2005-01-13
6 PM2.5 - Local Conditions         24 HOUR  PM25 24-hour 2012 2005-01-16
             Units.of.Measure Event.Type Observation.Count Observation.Percent
1 Micrograms/cubic meter (LC)       None                 1                 100
2 Micrograms/cubic meter (LC)       None                 1                 100
3 Micrograms/cubic meter (LC)       None                 1                 100
4 Micrograms/cubic meter (LC)       None                 1                 100
5 Micrograms/cubic meter (LC)       None                 1                 100
6 Micrograms/cubic meter (LC)       None                 1                 100
  Arithmetic.Mean X1st.Max.Value X1st.Max.Hour AQI Method.Code
1             9.0            9.0             0  50         118
2             7.7            7.7             0  43         118
3             7.3            7.3             0  41         118
4             8.7            8.7             0  48         118
```

```
5             3.3            3.3            0  18         118
6             7.7            7.7            0  43         118
                                            Method.Name   Local.Site.Name
1 R & P Model 2025 PM2.5 Sequential w/WINS - GRAVIMETRIC FAIRHOPE, Alabama
2 R & P Model 2025 PM2.5 Sequential w/WINS - GRAVIMETRIC FAIRHOPE, Alabama
3 R & P Model 2025 PM2.5 Sequential w/WINS - GRAVIMETRIC FAIRHOPE, Alabama
4 R & P Model 2025 PM2.5 Sequential w/WINS - GRAVIMETRIC FAIRHOPE, Alabama
5 R & P Model 2025 PM2.5 Sequential w/WINS - GRAVIMETRIC FAIRHOPE, Alabama
6 R & P Model 2025 PM2.5 Sequential w/WINS - GRAVIMETRIC FAIRHOPE, Alabama
                                                  Address State.Name
1 FAIRHOPE HIGH SCHOOL, 1 PIRATE DRIVE, FAIRHOPE,  ALABAMA    Alabama
2 FAIRHOPE HIGH SCHOOL, 1 PIRATE DRIVE, FAIRHOPE,  ALABAMA    Alabama
3 FAIRHOPE HIGH SCHOOL, 1 PIRATE DRIVE, FAIRHOPE,  ALABAMA    Alabama
4 FAIRHOPE HIGH SCHOOL, 1 PIRATE DRIVE, FAIRHOPE,  ALABAMA    Alabama
5 FAIRHOPE HIGH SCHOOL, 1 PIRATE DRIVE, FAIRHOPE,  ALABAMA    Alabama
6 FAIRHOPE HIGH SCHOOL, 1 PIRATE DRIVE, FAIRHOPE,  ALABAMA    Alabama
  County.Name City.Name                 CBSA.Name Date.of.Last.Change
1     Baldwin  Fairhope Daphne-Fairhope-Foley, AL          2024-09-05
2     Baldwin  Fairhope Daphne-Fairhope-Foley, AL          2024-09-05
3     Baldwin  Fairhope Daphne-Fairhope-Foley, AL          2024-09-05
4     Baldwin  Fairhope Daphne-Fairhope-Foley, AL          2024-09-05
5     Baldwin  Fairhope Daphne-Fairhope-Foley, AL          2024-09-05
6     Baldwin  Fairhope Daphne-Fairhope-Foley, AL          2024-09-05
```

```r
tail(pm_2005)
```

```
       State.Code County.Code Site.Num Parameter.Code POC Latitude Longitude
145908         CC          11        2          88101   1 49.13514 -102.9849
145909         CC          11        2          88101   1 49.13514 -102.9849
145910         CC          11        2          88101   1 49.13514 -102.9849
145911         CC          11        2          88101   1 49.13514 -102.9849
145912         CC          11        2          88101   1 49.13514 -102.9849
145913         CC          11        2          88101   1 49.13514 -102.9849
       Datum            Parameter.Name Sample.Duration Pollutant.Standard
145908 WGS84 PM2.5 - Local Conditions         24 HOUR  PM25 24-hour 2012
145909 WGS84 PM2.5 - Local Conditions         24 HOUR  PM25 24-hour 2012
145910 WGS84 PM2.5 - Local Conditions         24 HOUR  PM25 24-hour 2012
145911 WGS84 PM2.5 - Local Conditions         24 HOUR  PM25 24-hour 2012
145912 WGS84 PM2.5 - Local Conditions         24 HOUR  PM25 24-hour 2012
145913 WGS84 PM2.5 - Local Conditions         24 HOUR  PM25 24-hour 2012
       Date.Local            Units.of.Measure Event.Type Observation.Count
145908 2005-11-24 Micrograms/cubic meter (LC)       None                 1
145909 2005-11-30 Micrograms/cubic meter (LC)       None                 1
145910 2005-12-07 Micrograms/cubic meter (LC)       None                 1
145911 2005-12-12 Micrograms/cubic meter (LC)       None                 1
145912 2005-12-18 Micrograms/cubic meter (LC)       None                 1
145913 2005-12-30 Micrograms/cubic meter (LC)       None                 1
       Observation.Percent Arithmetic.Mean X1st.Max.Value X1st.Max.Hour AQI
145908                 100             3.8            3.8             0  21
145909                 100             4.0            4.0             0  22
```

```
145910                 100            6.6            6.6             0  37
145911                 100            4.0            4.0             0  22
145912                 100            5.1            5.1             0  28
145913                 100            9.1            9.1            15  51
       Method.Code                                              Method.Name
145908         143 R & P Model 2000 PM-2.5 Air Sampler w/VSCC - Gravimetric
145909         143 R & P Model 2000 PM-2.5 Air Sampler w/VSCC - Gravimetric
145910         143 R & P Model 2000 PM-2.5 Air Sampler w/VSCC - Gravimetric
145911         143 R & P Model 2000 PM-2.5 Air Sampler w/VSCC - Gravimetric
145912         143 R & P Model 2000 PM-2.5 Air Sampler w/VSCC - Gravimetric
145913         143 R & P Model 2000 PM-2.5 Air Sampler w/VSCC - Gravimetric
       Local.Site.Name     Address State.Name  County.Name     City.Name
145908                 ESTEVAN, SK     Canada Saskatchewan Not in a city
145909                 ESTEVAN, SK     Canada Saskatchewan Not in a city
145910                 ESTEVAN, SK     Canada Saskatchewan Not in a city
145911                 ESTEVAN, SK     Canada Saskatchewan Not in a city
145912                 ESTEVAN, SK     Canada Saskatchewan Not in a city
145913                 ESTEVAN, SK     Canada Saskatchewan Not in a city
       CBSA.Name Date.of.Last.Change
145908                    2024-05-19
145909                    2024-05-19
145910                    2024-05-19
145911                    2024-05-19
145912                    2024-05-19
145913                    2024-05-19
```

```r
head(pm_2025)
```

```
  State.Code County.Code Site.Num Parameter.Code POC Latitude Longitude Datum
1          1           3       10          88101   3 30.49748 -87.88026 NAD83
2          1           3       10          88101   3 30.49748 -87.88026 NAD83
3          1           3       10          88101   3 30.49748 -87.88026 NAD83
4          1           3       10          88101   3 30.49748 -87.88026 NAD83
5          1           3       10          88101   3 30.49748 -87.88026 NAD83
6          1           3       10          88101   3 30.49748 -87.88026 NAD83
              Parameter.Name Sample.Duration Pollutant.Standard Date.Local
1 PM2.5 - Local Conditions          1 HOUR                    2025-01-01
2 PM2.5 - Local Conditions          1 HOUR                    2025-01-02
3 PM2.5 - Local Conditions          1 HOUR                    2025-01-03
4 PM2.5 - Local Conditions          1 HOUR                    2025-01-04
5 PM2.5 - Local Conditions          1 HOUR                    2025-01-05
6 PM2.5 - Local Conditions          1 HOUR                    2025-01-06
             Units.of.Measure Event.Type Observation.Count Observation.Percent
1 Micrograms/cubic meter (LC)       None                24                 100
2 Micrograms/cubic meter (LC)       None                24                 100
3 Micrograms/cubic meter (LC)       None                24                 100
4 Micrograms/cubic meter (LC)       None                24                 100
5 Micrograms/cubic meter (LC)       None                24                 100
6 Micrograms/cubic meter (LC)       None                23                  96
  Arithmetic.Mean X1st.Max.Value X1st.Max.Hour AQI Method.Code
```

```
1        3.625000             19             21  NA          209
2        6.791667             26             23  NA          209
3        9.958333             25              1  NA          209
4        5.500000             11             19  NA          209
5        5.500000             12             14  NA          209
6        2.565217             16              2  NA          209
                                                           Method.Name
1 Met One BAM-1022 Mass Monitor w/ VSCC or TE-PM2.5C - Beta Attenuation
2 Met One BAM-1022 Mass Monitor w/ VSCC or TE-PM2.5C - Beta Attenuation
3 Met One BAM-1022 Mass Monitor w/ VSCC or TE-PM2.5C - Beta Attenuation
4 Met One BAM-1022 Mass Monitor w/ VSCC or TE-PM2.5C - Beta Attenuation
5 Met One BAM-1022 Mass Monitor w/ VSCC or TE-PM2.5C - Beta Attenuation
6 Met One BAM-1022 Mass Monitor w/ VSCC or TE-PM2.5C - Beta Attenuation
    Local.Site.Name                                                Address
1 FAIRHOPE, Alabama FAIRHOPE HIGH SCHOOL, 1 PIRATE DRIVE, FAIRHOPE,  ALABAMA
2 FAIRHOPE, Alabama FAIRHOPE HIGH SCHOOL, 1 PIRATE DRIVE, FAIRHOPE,  ALABAMA
3 FAIRHOPE, Alabama FAIRHOPE HIGH SCHOOL, 1 PIRATE DRIVE, FAIRHOPE,  ALABAMA
4 FAIRHOPE, Alabama FAIRHOPE HIGH SCHOOL, 1 PIRATE DRIVE, FAIRHOPE,  ALABAMA
5 FAIRHOPE, Alabama FAIRHOPE HIGH SCHOOL, 1 PIRATE DRIVE, FAIRHOPE,  ALABAMA
6 FAIRHOPE, Alabama FAIRHOPE HIGH SCHOOL, 1 PIRATE DRIVE, FAIRHOPE,  ALABAMA
  State.Name County.Name City.Name                 CBSA.Name
1    Alabama     Baldwin  Fairhope Daphne-Fairhope-Foley, AL
2    Alabama     Baldwin  Fairhope Daphne-Fairhope-Foley, AL
3    Alabama     Baldwin  Fairhope Daphne-Fairhope-Foley, AL
4    Alabama     Baldwin  Fairhope Daphne-Fairhope-Foley, AL
5    Alabama     Baldwin  Fairhope Daphne-Fairhope-Foley, AL
6    Alabama     Baldwin  Fairhope Daphne-Fairhope-Foley, AL
  Date.of.Last.Change
1          2025-10-29
2          2025-10-29
3          2025-10-29
4          2025-10-29
5          2025-10-29
6          2025-10-29
```

```r
tail(pm_2025)
```

```
       State.Code County.Code Site.Num Parameter.Code POC Latitude Longitude
466755         80          26        6          88101   1 31.29129 -110.9515
466756         80          26        6          88101   1 31.29129 -110.9515
466757         80          26        6          88101   1 31.29129 -110.9515
466758         80          26        6          88101   1 31.29129 -110.9515
466759         80          26        6          88101   1 31.29129 -110.9515
466760         80          26        6          88101   1 31.29129 -110.9515
       Datum            Parameter.Name Sample.Duration Pollutant.Standard
466755 WGS84 PM2.5 - Local Conditions   24-HR BLK AVG  PM25 24-hour 2012
466756 WGS84 PM2.5 - Local Conditions   24-HR BLK AVG  PM25 24-hour 2012
466757 WGS84 PM2.5 - Local Conditions   24-HR BLK AVG  PM25 24-hour 2012
466758 WGS84 PM2.5 - Local Conditions   24-HR BLK AVG  PM25 24-hour 2012
466759 WGS84 PM2.5 - Local Conditions   24-HR BLK AVG  PM25 24-hour 2012
```

```
466760 WGS84 PM2.5 - Local Conditions   24-HR BLK AVG  PM25 24-hour 2012
       Date.Local           Units.of.Measure Event.Type Observation.Count
466755 2025-09-25 Micrograms/cubic meter (LC)       None                 1
466756 2025-09-26 Micrograms/cubic meter (LC)       None                 1
466757 2025-09-27 Micrograms/cubic meter (LC)       None                 1
466758 2025-09-28 Micrograms/cubic meter (LC)       None                 1
466759 2025-09-29 Micrograms/cubic meter (LC)       None                 1
466760 2025-09-30 Micrograms/cubic meter (LC)       None                 1
       Observation.Percent Arithmetic.Mean X1st.Max.Value X1st.Max.Hour AQI
466755                 100             9.9            9.9             0  52
466756                 100             3.5            3.5             0  19
466757                 100             7.9            7.9             0  44
466758                 100            11.2           11.2             0  55
466759                 100            15.3           15.3             0  63
466760                 100            17.6           17.6             0  67
       Method.Code
466755         638
466756         638
466757         638
466758         638
466759         638
466760         638
                                                                              Method.Name
466755 Teledyne T640X at 16.67 LPM w/Network Data Alignment enabled - Broadband spectroscopy
466756 Teledyne T640X at 16.67 LPM w/Network Data Alignment enabled - Broadband spectroscopy
466757 Teledyne T640X at 16.67 LPM w/Network Data Alignment enabled - Broadband spectroscopy
466758 Teledyne T640X at 16.67 LPM w/Network Data Alignment enabled - Broadband spectroscopy
466759 Teledyne T640X at 16.67 LPM w/Network Data Alignment enabled - Broadband spectroscopy
466760 Teledyne T640X at 16.67 LPM w/Network Data Alignment enabled - Broadband spectroscopy
                    Local.Site.Name
466755 Nogales Sonora Institute ITN
466756 Nogales Sonora Institute ITN
466757 Nogales Sonora Institute ITN
466758 Nogales Sonora Institute ITN
466759 Nogales Sonora Institute ITN
466760 Nogales Sonora Institute ITN
                                                                 Address
466755 Avenida Instituto Tecnologico #911, Granja, 84065 Nogales, Son., Mexico
466756 Avenida Instituto Tecnologico #911, Granja, 84065 Nogales, Son., Mexico
466757 Avenida Instituto Tecnologico #911, Granja, 84065 Nogales, Son., Mexico
466758 Avenida Instituto Tecnologico #911, Granja, 84065 Nogales, Son., Mexico
466759 Avenida Instituto Tecnologico #911, Granja, 84065 Nogales, Son., Mexico
466760 Avenida Instituto Tecnologico #911, Granja, 84065 Nogales, Son., Mexico
              State.Name County.Name     City.Name CBSA.Name
466755 Country Of Mexico      SONORA Not in a city
466756 Country Of Mexico      SONORA Not in a city
466757 Country Of Mexico      SONORA Not in a city
466758 Country Of Mexico      SONORA Not in a city
466759 Country Of Mexico      SONORA Not in a city
466760 Country Of Mexico      SONORA Not in a city
       Date.of.Last.Change
```

```
466755          2025-10-22
466756          2025-10-22
466757          2025-10-22
466758          2025-10-22
466759          2025-10-22
466760          2025-10-22
```

EDA Step 5: Visualize the distribution of PM_2.5

```r
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
```

```r
library(ggplot2)

df <- pm_2005

df2 <- df %>%
  mutate(
    Date = as.Date(Date.Local),
    PM_2_5 = as.numeric(Arithmetic.Mean)
  )

# Michigan-only (state code 26)
mi <- df2 %>%
  filter(State.Code == 26)
```

```r
# Numeric summary (all)
summary(df2$PM_2_5)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    6.80   10.70   12.73   16.40  132.60 
```

```r
quantile(df2$PM_2_5, probs = c(0, .01, .05, .10, .25, .50, .75, .90, .95, .99, 1), na.rm = TRUE)
```

```
   0%    1%    5%   10%   25%   50%   75%   90%   95%   99%  100% 
  0.0   1.9   3.3   4.4   6.8  10.7  16.4  23.8  29.2  41.6 132.6 
```

```r
mean(df2$PM_2_5, na.rm = TRUE)
```

```
[1] 12.73335
```

```r
sd(df2$PM_2_5, na.rm = TRUE)
```

```
[1] 8.487454
```

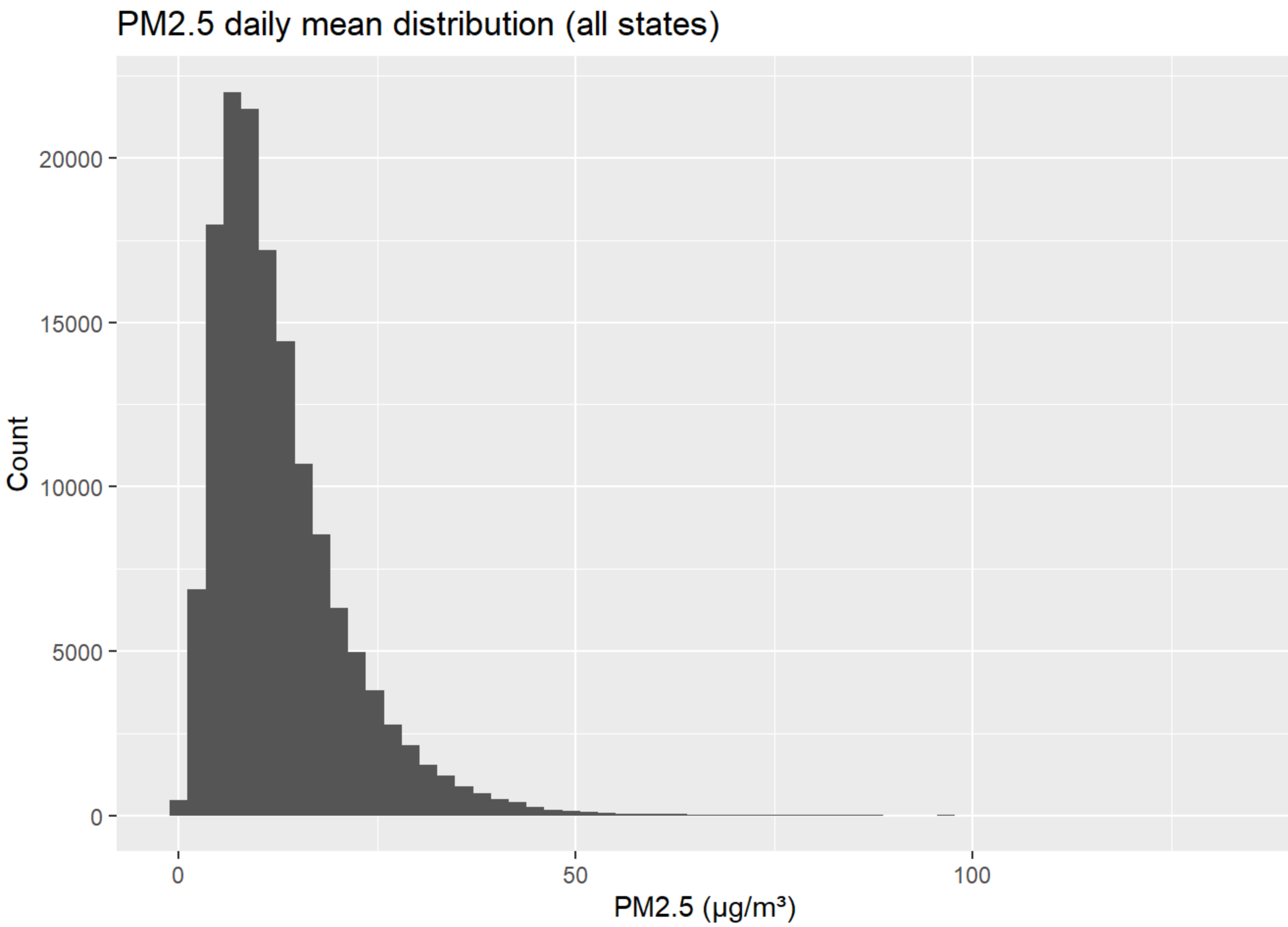
```r
# Numeric summary (Michigan)
summary(mi$PM_2_5)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.70    6.50   11.20   13.99   19.00   81.60 
```

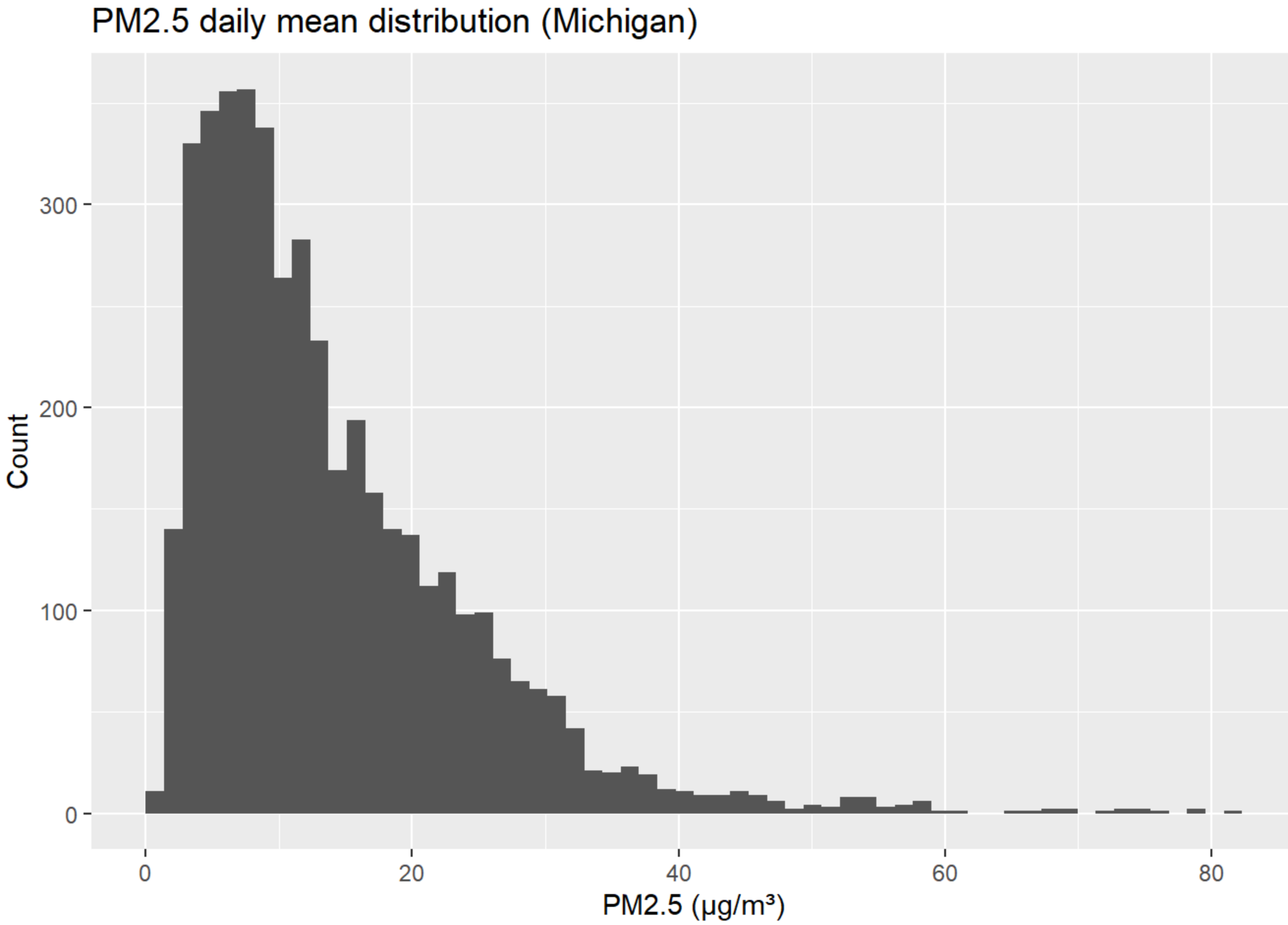
```r
quantile(mi$PM_2_5, probs = c(0, .01, .05, .10, .25, .50, .75, .90, .95, .99, 1), na.rm = TRUE)
```

```
   0%    1%    5%   10%   25%   50%   75%   90%   95%   99%  100% 
 0.70  2.00  3.20  4.00  6.50 11.20 19.00 27.20 32.35 52.30 81.60 
```

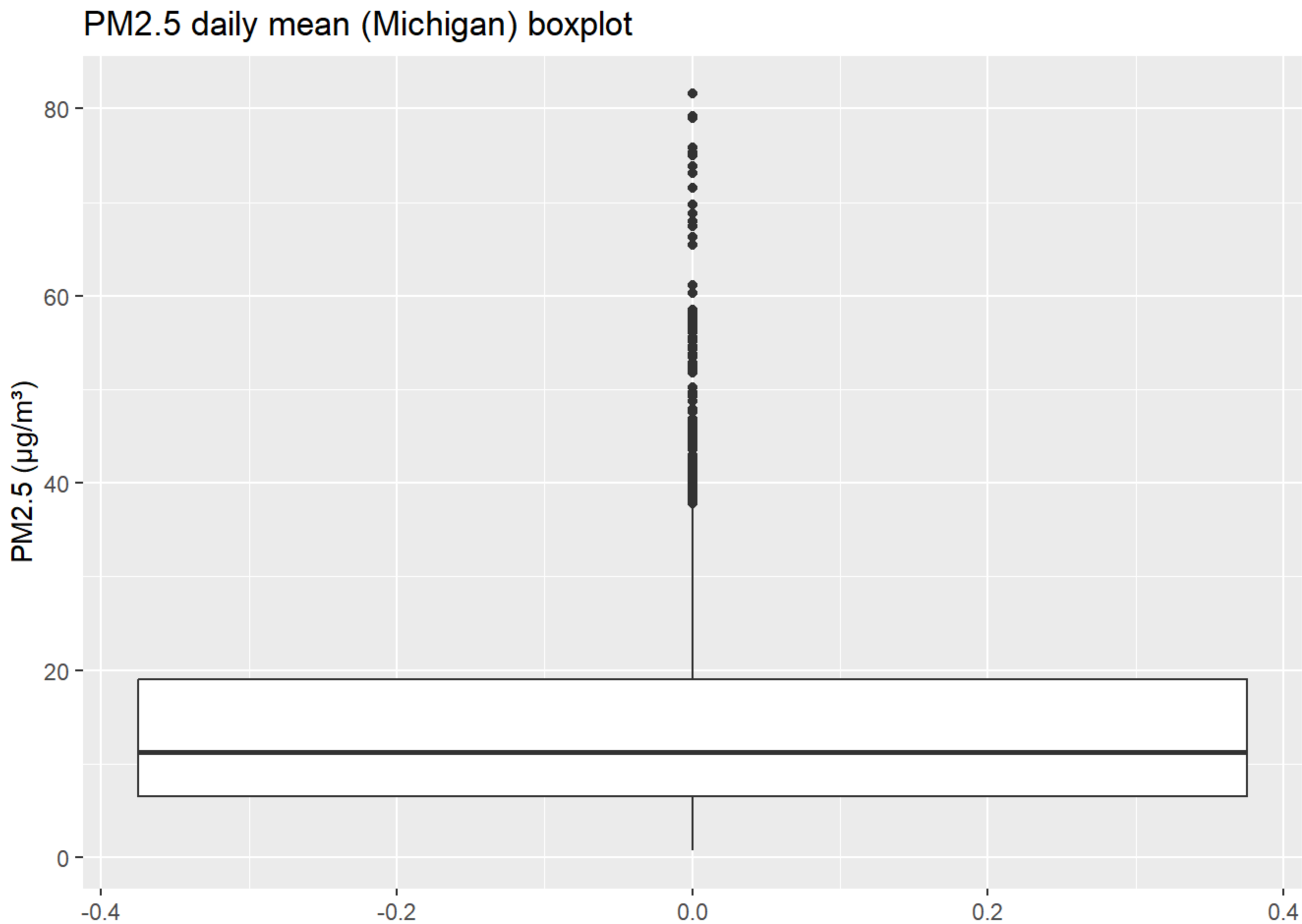
```r
ggplot(df2, aes(x = PM_2_5)) +
  geom_histogram(bins = 60) +
  labs(title = "PM2.5 daily mean distribution (all states)", x = "PM2.5 (µg/m³)", y = "Count")
```




```
ggplot(mi, aes(x = PM_2_5)) +
  geom_histogram(bins = 60) +
  labs(title = "PM2.5 daily mean distribution (Michigan)", x = "PM2.5 (µg/m³)", y = "Count")
```




```
# Boxplot (Michigan) - good for outliers / skew
ggplot(mi, aes(y = PM_2_5)) +
  geom_boxplot() +
  labs(title = "PM2.5 daily mean (Michigan) boxplot", y = "PM2.5 (µg/m³)")
```



EDA Summary

After reading the data into R, we find that both the 2005 and 2025 PM2.5 data have 29 variables (columns), which have the same names between the datasets. The 2005 data have 145,913 records (rows), while the 2025 data have 466,760 records.

The variables in the data include important information such as state, county, latitude, longitude, date, units of measurement, and the most important variable: the PM2.5 measurement, recorded as “Arithmetic.Mean”. Some variables encode the same element in different ways, such as “State.Code” and “State.Name”, although they are both the “chr” data type in this case. Other variables like latitude, longitude, and Arithmetic.Mean have the “num” datatype to store floating-point numbers.

From examining the top and bottom of each dataset, many of the values are grouped into similar sets, so the values at the top and bottom are frequently the same. The exception is Arithmetic.Mean, which always sees significant changes between records.

From observing the histograms of the distribution of PM2.5, we find that in both Michigan and overall, the distribution is right-skewed, with the mean located at roughly 8 µg/m^3. The frequency of PM2.5 concentrations decreases sharply after this point, with a very small number of records surpassing 50 µg/m^3. The boxplot of Michigan records confirms that any records above 40 µg/m^3 are considered outliers.

2. Combine data frames into one

```
library(dplyr)

clean_pm <- function(df) {
  df %>%
    mutate(
      Date = as.Date(Date.Local),
      year = as.integer(format(Date, "%Y")),
      pm25 = as.numeric(Arithmetic.Mean),
      state = State.Name,
      county = County.Name,
      site = Site.Num
    ) %>%
    select(state, county, site, year, Date, pm25,
           Latitude, Longitude, Method.Name, Units.of.Measure) %>%
    arrange(state, county, site, Date)
}

pm_all <- bind_rows(
  clean_pm(pm_2005),
  clean_pm(pm_2025)
)

# Michigan only
pm_mi <- pm_all %>% filter(state == 26)
```

```
dplyr::count(pm_all, year)
```

```
  year      n
1 2005 145913
2 2025 466760
```

```
summary(pm_all$pm25)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -5.733   4.500   6.900   8.656  10.662 288.500 
```

```
str(pm_all)
```

```
'data.frame':	612673 obs. of  10 variables:
 $ state           : chr  "Alabama" "Alabama" "Alabama" "Alabama" ...
 $ county          : chr  "Baldwin" "Baldwin" "Baldwin" "Baldwin" ...
 $ site            : int  10 10 10 10 10 10 10 10 10 10 ...
 $ year            : int  2005 2005 2005 2005 2005 2005 2005 2005 2005 2005 ...
 $ Date            : Date, format: "2005-01-01" "2005-01-04" ...
 $ pm25            : num  9 7.7 7.3 8.7 3.3 7.7 16.6 7.9 15.1 23.4 ...
 $ Latitude        : num  30.5 30.5 30.5 30.5 30.5 ...
```

```
 $ Longitude       : num  -87.9 -87.9 -87.9 -87.9 -87.9 ...
 $ Method.Name     : chr  "R & P Model 2025 PM2.5 Sequential w/WINS - GRAVIMETRIC" "R & P Model 2025 PM2.5 Sequential w/WINS - GRAVIMETRIC" "R & P Model 2025 PM2.5 Sequential w/WINS - GRAVIMETRIC" "R & P Model 2025 PM2.5 Sequential w/WINS - GRAVIMETRIC" ...
 $ Units.of.Measure: chr  "Micrograms/cubic meter (LC)" "Micrograms/cubic meter (LC)" "Micrograms/cubic meter (LC)" "Micrograms/cubic meter (LC)" ...
```

```r
head(pm_all)
```

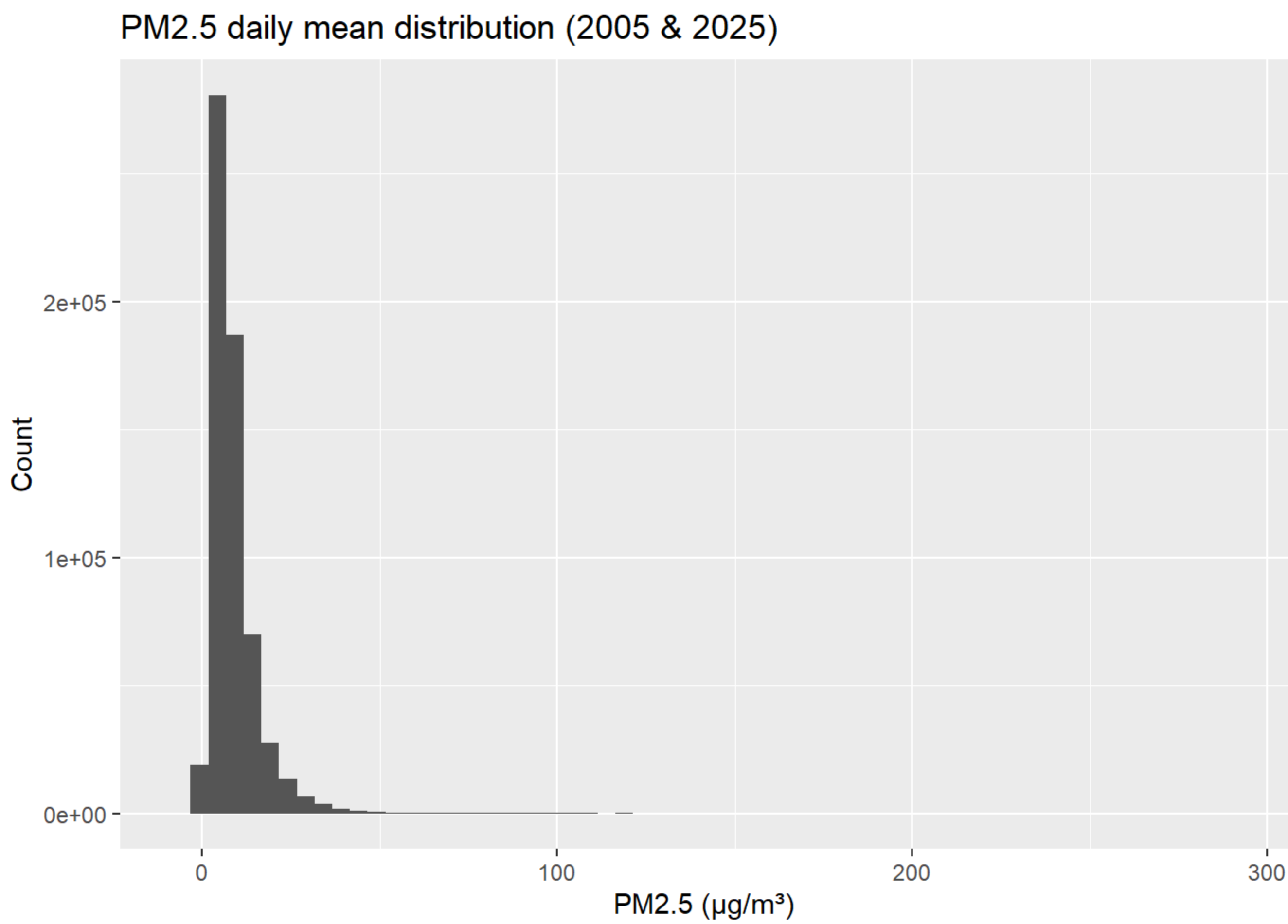
```
    state  county site year       Date pm25 Latitude Longitude
1 Alabama Baldwin   10 2005 2005-01-01  9.0 30.49748 -87.88026
2 Alabama Baldwin   10 2005 2005-01-04  7.7 30.49748 -87.88026
3 Alabama Baldwin   10 2005 2005-01-07  7.3 30.49748 -87.88026
4 Alabama Baldwin   10 2005 2005-01-10  8.7 30.49748 -87.88026
5 Alabama Baldwin   10 2005 2005-01-13  3.3 30.49748 -87.88026
6 Alabama Baldwin   10 2005 2005-01-16  7.7 30.49748 -87.88026
                                              Method.Name
1 R & P Model 2025 PM2.5 Sequential w/WINS - GRAVIMETRIC
2 R & P Model 2025 PM2.5 Sequential w/WINS - GRAVIMETRIC
3 R & P Model 2025 PM2.5 Sequential w/WINS - GRAVIMETRIC
4 R & P Model 2025 PM2.5 Sequential w/WINS - GRAVIMETRIC
5 R & P Model 2025 PM2.5 Sequential w/WINS - GRAVIMETRIC
6 R & P Model 2025 PM2.5 Sequential w/WINS - GRAVIMETRIC
             Units.of.Measure
1 Micrograms/cubic meter (LC)
2 Micrograms/cubic meter (LC)
3 Micrograms/cubic meter (LC)
4 Micrograms/cubic meter (LC)
5 Micrograms/cubic meter (LC)
6 Micrograms/cubic meter (LC)
```

```r
tail(pm_all)
```

```
         state   county site year       Date     pm25 Latitude Longitude
612668 Wyoming Washakie    2 2025 2025-07-29 4.141667 44.00792  -107.957
612669 Wyoming Washakie    2 2025 2025-07-29 4.100000 44.00792  -107.957
612670 Wyoming Washakie    2 2025 2025-07-30 4.687500 44.00792  -107.957
612671 Wyoming Washakie    2 2025 2025-07-30 4.600000 44.00792  -107.957
612672 Wyoming Washakie    2 2025 2025-07-31 4.825000 44.00792  -107.957
612673 Wyoming Washakie    2 2025 2025-07-31 4.800000 44.00792  -107.957
                                                             Method.Name
612668 Met One BAM-1022 Mass Monitor w/ VSCC or TE-PM2.5C - Beta Attenuation
612669 Met One BAM-1022 Mass Monitor w/ VSCC or TE-PM2.5C - Beta Attenuation
612670 Met One BAM-1022 Mass Monitor w/ VSCC or TE-PM2.5C - Beta Attenuation
612671 Met One BAM-1022 Mass Monitor w/ VSCC or TE-PM2.5C - Beta Attenuation
612672 Met One BAM-1022 Mass Monitor w/ VSCC or TE-PM2.5C - Beta Attenuation
612673 Met One BAM-1022 Mass Monitor w/ VSCC or TE-PM2.5C - Beta Attenuation
                  Units.of.Measure
612668 Micrograms/cubic meter (LC)
```

```
612669 Micrograms/cubic meter (LC)
612670 Micrograms/cubic meter (LC)
612671 Micrograms/cubic meter (LC)
612672 Micrograms/cubic meter (LC)
612673 Micrograms/cubic meter (LC)
```

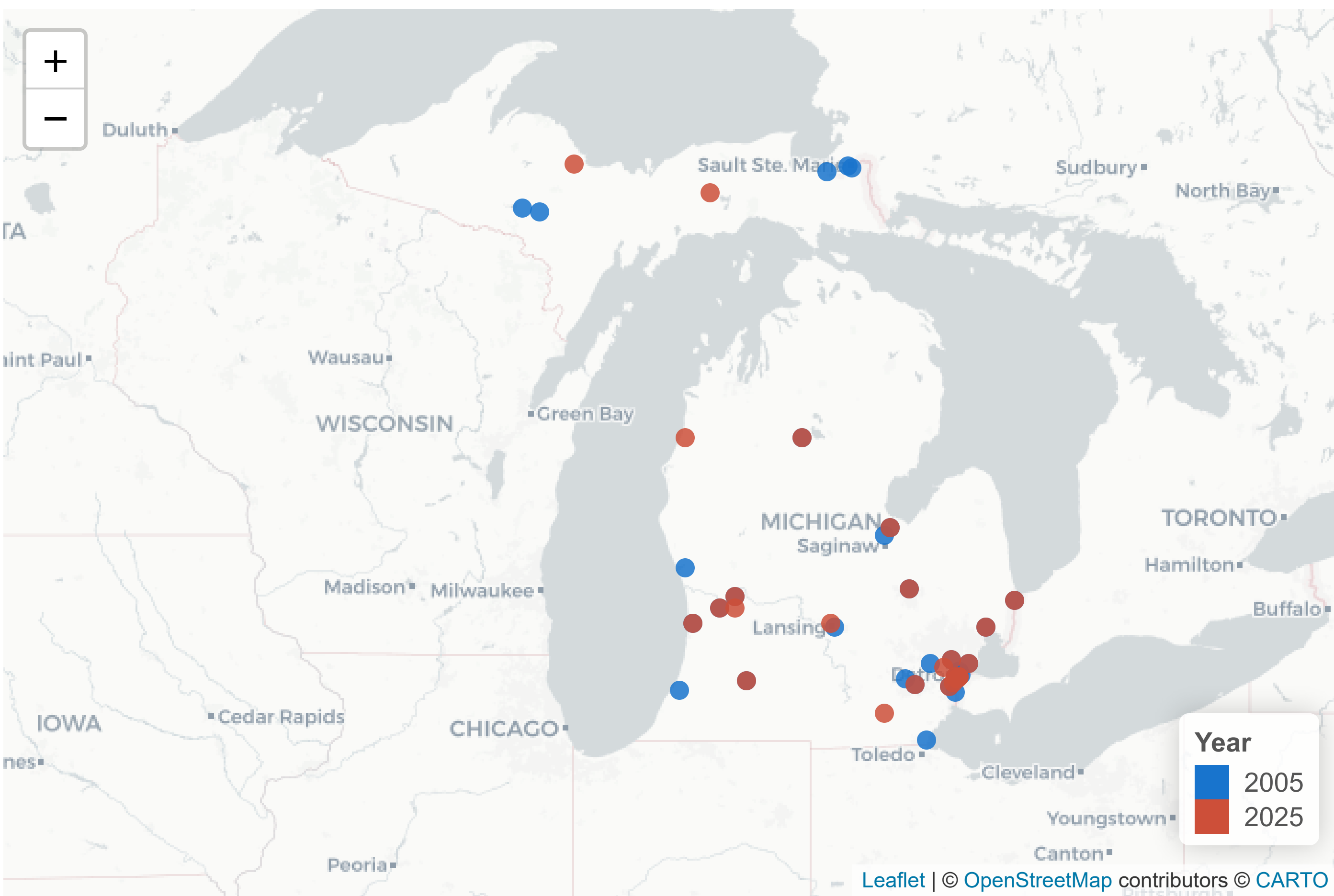
```r
ggplot(pm_all, aes(x = pm25)) +
  geom_histogram(bins = 60) +
  labs(title = "PM2.5 daily mean distribution (2005 & 2025)", x = "PM2.5 (µg/m³)", y = "Count")
```



3. Create leaflet map to show monitoring site locations

```r
library(dplyr)
library(leaflet)

# Filter to Michigan data
df <- pm_all %>% filter(state == "Michigan")

# One row per site per year
sites_year <- df %>%
  filter(!is.na(Latitude), !is.na(Longitude)) %>%
  distinct(year, state, county, site, Latitude, Longitude)
```

```
pal <- colorFactor(palette = c("dodgerblue3", "tomato3"), domain = sort(unique(sites_year$year)))

m <- leaflet(sites_year) %>%
  addProviderTiles(providers$CartoDB.Positron) %>%
  addCircleMarkers(
    lng = ~Longitude, lat = ~Latitude,
    radius = 5, stroke = FALSE, fillOpacity = 0.85,
    color = ~pal(year),
    popup = ~paste0(
      "<b>Year:</b> ", year, "<br/>",
      "<b>Site:</b> ", state, "-", county, "-", site, "<br/>",
      "<b>Lat/Lon:</b> ", round(Latitude, 4), ", ", round(Longitude, 4)
    )
  ) %>%
  addLegend(
    position = "bottomright",
    pal = pal, values = ~year,
    title = "Year", opacity = 1
  )

m
```




```
# Counts by year
sites_year %>%
```

```r
  count(year, name = "n_sites")
```

```
  year n_sites
1 2005      31
2 2025      26
```

```r
# Bounding box by year (rough geographic footprint)
sites_year %>%
  group_by(year) %>%
  summarize(
    n_sites = n(),
    lat_min = min(Latitude), lat_max = max(Latitude),
    lon_min = min(Longitude), lon_max = max(Longitude),
    .groups = "drop"
  )
```

```
# A tibble: 2 × 6
   year n_sites lat_min lat_max lon_min lon_max
  <int>   <int>   <dbl>   <dbl>   <dbl>   <dbl>
1  2005      31    41.8    46.5   -88.1   -82.5
2  2025      26    42.0    46.5   -87.5   -82.5
```

```r
# Sites present in both years vs. only one year
site_id <- sites_year %>%
  mutate(site_id = paste(state, county, site, sep = "-")) %>%
  distinct(year, site_id)

site_id %>%
  count(site_id) %>%
  count(n, name = "n_sites") %>%
  mutate(present_in_years = n) %>%
  arrange(desc(present_in_years))
```

```
  n n_sites present_in_years
1 2      15                2
2 1      27                1
```

From the map, we find the the monitoring sites are not evenly spread across Michigan. Instead, they have the highest concentration in Southeeast Michigan (the Detroit - Ann Arbor - Downriver area) for both 2005 and 2025. This makes sense, as this would be the area in Michigan where the population - and thus the emission source - is the highest. There are additional clusters of sites in central/east-central Michigan, such as around Lansing and the Saginaw / Bay City / Flint region. Meanwhile, there are very few sites in western and northern Michigan (which were only introduced as of 2025), and a small amount in the Upper Peninsula. Overall, while the number of monitoring sites increased from 2005 to 2025, the spatial distribution of the sites across Michigan remained mostly unchanged.

4. Check for data issues

Identify missing or implausible values

```
library(dplyr)

df <- pm_all %>%
  mutate(
    Date = as.Date(Date),
    year = as.integer(year),
    pm25 = as.numeric(pm25)
  )
```

```
# 1) Missingness
df %>%
  summarize(
    n = n(),
    pm25_missing = sum(is.na(pm25)),
    pm25_missing_pct = mean(is.na(pm25)) * 100
  )
```

```
       n pm25_missing pm25_missing_pct
1 612673            0                0
```

```
df %>%
  group_by(year) %>%
  summarize(
    n = n(),
    pm25_missing = sum(is.na(pm25)),
    pm25_missing_pct = mean(is.na(pm25)) * 100,
    .groups = "drop"
  )
```

```
# A tibble: 2 × 4
   year      n pm25_missing pm25_missing_pct
  <int>  <int>        <int>            <dbl>
1  2005 145913            0                0
2  2025 466760            0                0
```

We find that there are no values in the dataset that are completely missing.

```
# 2) Implausible values (negative, extremely large)
df %>%
  summarize(
    n_neg = sum(pm25 < 0, na.rm = TRUE),
    n_zero = sum(pm25 == 0, na.rm = TRUE),
    n_gt_500 = sum(pm25 > 500, na.rm = TRUE),
    n_gt_1000 = sum(pm25 > 1000, na.rm = TRUE)
  )
```

```
  n_neg n_zero n_gt_500 n_gt_1000
1  1734    338        0         0
```

```
# Table of implausible PM25 measurements
df %>%
  filter(pm25 < 0 | pm25 > 500) %>%
  select(state, county, site, year, Date, pm25) %>%
  arrange(desc(pm25)) %>%
  head(50)
```

```
            state            county site year       Date      pm25
1         Wyoming          Washakie    2 2025 2025-03-08 -0.004167
2      California              Kern   20 2025 2025-02-07 -0.008333
3           Texas            Potter 1025 2025 2025-01-02 -0.008333
4      California         Riverside 1016 2025 2025-01-02 -0.008696
5      New Jersey          Atlantic    6 2025 2025-01-09 -0.011765
6          Oregon            Marion   41 2025 2025-04-28 -0.012500
7      New Mexico              Taos    5 2025 2025-03-15 -0.020833
8          Oregon              Lane 2013 2025 2025-04-07 -0.020833
9          Oregon              Lane 2013 2025 2025-05-17 -0.020833
10     California          Imperial    7 2025 2025-01-27 -0.025000
11        Montana    Lewis and Clark    4 2025 2025-03-23 -0.025000
12         Oregon              Lane   60 2025 2025-03-21 -0.025000
13          Texas            Potter 1025 2025 2025-03-30 -0.025000
14       Nebraska              Gage    5 2025 2025-06-19 -0.029167
15   South Dakota         Minnehaha    9 2025 2025-02-28 -0.029167
16     California         Riverside 1016 2025 2025-02-15 -0.033333
17         Oregon         Deschutes  120 2025 2025-04-22 -0.033333
18         Oregon              Lane   60 2025 2025-01-03 -0.033333
19          Texas            Potter 1025 2025 2025-02-08 -0.033333
20     California            Nevada 1001 2025 2025-01-02 -0.034783
21        Montana    Lewis and Clark    4 2025 2025-05-20 -0.037500
22           Utah             Grand    7 2025 2025-03-14 -0.037500
23        Wyoming            Teton 1006 2025 2025-05-06 -0.037500
24     California            Colusa 1002 2025 2025-01-13 -0.039130
25         Alaska Matanuska-Susitna   10 2025 2025-01-15 -0.041667
26         Alaska Matanuska-Susitna   10 2025 2025-05-15 -0.041667
27     California            Colusa    7 2025 2025-02-17 -0.041667
28     California            Colusa    7 2025 2025-02-18 -0.041667
29     California       Los Angeles 4009 2025 2025-02-13 -0.041667
30     California    San Luis Obispo 8002 2025 2025-07-28 -0.041667
31     California           Ventura    9 2025 2025-03-31 -0.041667
32     California           Ventura 2002 2025 2025-01-23 -0.041667
33         Hawaii            Hawaii    5 2025 2025-06-17 -0.041667
34          Maine        Cumberland   29 2025 2025-05-19 -0.041667
35        Montana        Beaverhead    3 2025 2025-01-22 -0.041667
36 North Carolina            Durham   15 2025 2025-08-22 -0.041667
37         Oregon              Lane 1013 2025 2025-03-02 -0.041667
38     Washington           Stevens    5 2025 2025-06-21 -0.041667
39     Washington           Stevens    5 2025 2025-06-30 -0.041667
```

```
40        Wyoming           Albany   12 2025 2025-01-05 -0.041667
41        Wyoming           Albany   12 2025 2025-02-15 -0.041667
42        Wyoming          Fremont   99 2025 2025-02-17 -0.041667
43        Wyoming          Fremont  232 2025 2025-02-07 -0.041667
44        Wyoming          Fremont  232 2025 2025-03-04 -0.041667
45        Wyoming             Park    1 2025 2025-01-26 -0.041667
46        Wyoming       Sweetwater    7 2025 2025-02-13 -0.041667
47        Montana         Missoula   24 2025 2025-05-13 -0.042857
48     California           Fresno  500 2025 2025-10-14 -0.043478
49     California      Los Angeles 9035 2025 2025-03-11 -0.043478
50     California  San Luis Obispo 2004 2025 2025-07-25 -0.043478
```

We find that some of the PM2.5 measurements are negative, and a small number are exactly 0. Since these are implausible and irrelevant, respectively, we filter these records out of the dataset:

```
df <- df %>% filter(pm25 > 0)

# Table of plausible PM25 measurements
df %>%
  select(state, county, site, year, Date, pm25) %>%
  arrange(desc(pm25)) %>%
  head(50)
```

```
          state   county site year       Date     pm25
1         Texas  El Paso   44 2025 2025-04-01 288.5000
2         Texas  El Paso   55 2025 2025-03-18 285.8235
3         Texas  El Paso   55 2025 2025-03-18 285.8000
4    New Mexico Dona Ana   21 2025 2025-03-18 259.3913
5    New Mexico Dona Ana   21 2025 2025-03-18 259.3000
6       Arizona Maricopa   19 2025 2025-01-01 258.1917
7       Arizona Maricopa   19 2025 2025-01-01 258.1000
8    New Mexico Dona Ana   21 2025 2025-03-06 246.2792
9    New Mexico Dona Ana   21 2025 2025-03-06 246.2000
10   New Mexico Dona Ana   21 2025 2025-03-06 242.4208
11   New Mexico Dona Ana   21 2025 2025-03-06 242.4000
12        Texas  El Paso   55 2025 2025-03-03 238.6667
13        Texas  El Paso   55 2025 2025-03-03 238.6000
14   New Mexico Dona Ana   22 2025 2025-03-06 233.4542
15   New Mexico Dona Ana   22 2025 2025-03-06 233.4000
16        Texas  El Paso   44 2025 2025-03-18 228.6500
17        Texas  El Paso   44 2025 2025-03-18 228.6000
18   New Mexico Dona Ana   21 2025 2025-03-18 226.6870
19   New Mexico Dona Ana   21 2025 2025-03-18 226.6000
20   New Mexico Dona Ana   22 2025 2025-03-18 221.0250
21   New Mexico Dona Ana   22 2025 2025-03-18 221.0000
22      Arizona Maricopa 4003 2025 2025-01-01 210.3292
23      Arizona Maricopa 4003 2025 2025-01-01 210.3000
24 North Dakota    Burke    4 2025 2025-05-31 210.3000
25 North Dakota    Burke    4 2025 2025-05-31 210.3000
26 North Dakota     Ward    3 2025 2025-05-31 208.4292
27 North Dakota     Ward    3 2025 2025-05-31 208.4000
```

```
28 North Dakota   Mercer    4 2025 2025-05-31 208.2708
29 North Dakota   Mercer    4 2025 2025-05-31 208.2000
30 North Dakota   Mercer    4 2025 2025-05-31 207.6500
31 North Dakota   Mercer    4 2025 2025-05-31 207.6000
32   New Mexico Dona Ana   21 2025 2025-04-01 193.2083
33   New Mexico Dona Ana   21 2025 2025-04-01 193.2000
34   New Mexico Dona Ana   21 2025 2025-04-01 180.8083
35   New Mexico Dona Ana   21 2025 2025-04-01 180.8000
36      Arizona Maricopa 9812 2025 2025-01-01 167.4375
37      Arizona Maricopa 9812 2025 2025-01-01 167.4000
38        Texas  El Paso   44 2025 2025-04-01 167.0091
39        Texas  El Paso   44 2025 2025-04-01 167.0000
40 North Dakota   Oliver    2 2025 2025-05-31 166.7708
41 North Dakota   Oliver    2 2025 2025-05-31 166.7000
42        Texas  El Paso   55 2025 2025-04-27 164.3750
43        Texas  El Paso   55 2025 2025-04-27 164.3000
44        Texas  El Paso   55 2025 2025-04-01 163.1364
45        Texas  El Paso   55 2025 2025-04-01 163.1000
46   New Mexico Dona Ana   16 2025 2025-03-06 162.9875
47   New Mexico Dona Ana   16 2025 2025-03-06 162.9000
48   New Mexico Dona Ana   21 2025 2025-04-27 162.2792
49   New Mexico Dona Ana   21 2025 2025-04-27 162.2000
50   New Mexico Dona Ana   22 2025 2025-04-01 153.2875
```

```r
# 3) Distribution checks by year
df %>%
  group_by(year) %>%
  summarize(
    n = sum(!is.na(pm25)),
    min = min(pm25, na.rm = TRUE),
    p1 = quantile(pm25, 0.01, na.rm = TRUE),
    p5 = quantile(pm25, 0.05, na.rm = TRUE),
    median = median(pm25, na.rm = TRUE),
    p95 = quantile(pm25, 0.95, na.rm = TRUE),
    p99 = quantile(pm25, 0.99, na.rm = TRUE),
    max = max(pm25, na.rm = TRUE),
    mean = mean(pm25, na.rm = TRUE),
    sd = sd(pm25, na.rm = TRUE),
    .groups = "drop"
  )
```

```
# A tibble: 2 × 11
   year      n     min    p1    p5 median   p95   p99   max  mean    sd
  <int>  <int>   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1  2005 145856 0.1     1.9     3.3  10.7   29.2  41.6  133. 12.7   8.49
2  2025 464745 0.00417 0.989   2.1   6.19  16.2  27.1  288.  7.42  6.00
```

From this table, we see an overall level shift since the mean and median significantly decreased from 2005 to 2025; the mean decreased from 12.74 to 7.42, and the median decreased from 10.70 to 6.19. The highest percentiles are also lower in 2025 compared to 2005, which suggests that the overall PM2.5 concentrations in the air were less in 2025. There are also outliers of very high maximum concentrations in both years (132.6 and 288.5 respectively), which could be caused by an extreme event such as a wildfire.

```
# 4) Missing days per site-year
site_days <- df %>%
  filter(!is.na(pm25)) %>%
  group_by(state, county, site, year) %>%
  summarize(
    n_days = n_distinct(Date),
    .groups = "drop"
  )

site_days %>%
  group_by(year) %>%
  summarize(
    n_site_years = n(),
    min_days = min(n_days),
    p25_days = quantile(n_days, 0.25),
    median_days = median(n_days),
    p75_days = quantile(n_days, 0.75),
    max_days = max(n_days),
    .groups = "drop"
  )
```

```
# A tibble: 2 × 7
   year n_site_years min_days p25_days median_days p75_days max_days
  <int>        <int>    <int>    <dbl>       <int>    <dbl>    <int>
1  2005         1067        3       90         114      120      364
2  2025         1015        1      175         212      271      316
```

This shows that 2025 has roughly the same number of site-year pairs as 2005, with more frequent coverage per day in 2025 (higher median and percentiles), but some of the calendar year is not covered in 2025 (the maximum day is only 316 in 2025 compared to 364 in 2005). This makes sense because the 2025 data is only recorded as of November 24, 2025 according to the EPA website.

Check methods used for data collection

```
unique(pm_all$"Method.Name")
```

```
 [1] "R & P Model 2025 PM2.5 Sequential w/WINS - GRAVIMETRIC"
 [2] "BGI Models PQ200-VSCC or PQ200A-VSCC - Gravimetric"
 [3] "BGI Model PQ200 PM2.5 Sampler w/WINS - GRAVIMETRIC"
 [4] "Andersen RAAS2.5-300 PM2.5 SEQ w/WINS - GRAVIMETRIC"
 [5] "R & P Model 2000 PM2.5 Sampler w/WINS - GRAVIMETRIC"
 [6] "R & P Model 2000 PM-2.5 Air Sampler w/VSCC - Gravimetric"
 [7] "R & P Model 2025 PM-2.5 Sequential Air Sampler w/VSCC - Gravimetric"
 [8] "Andersen RAAS2.5-100 PM2.5 SAM w/WINS - GRAVIMETRIC"
 [9] "Thermo Electron Model RAAS2.5-300 Sequential w/VSCC - Gravimetric"
```

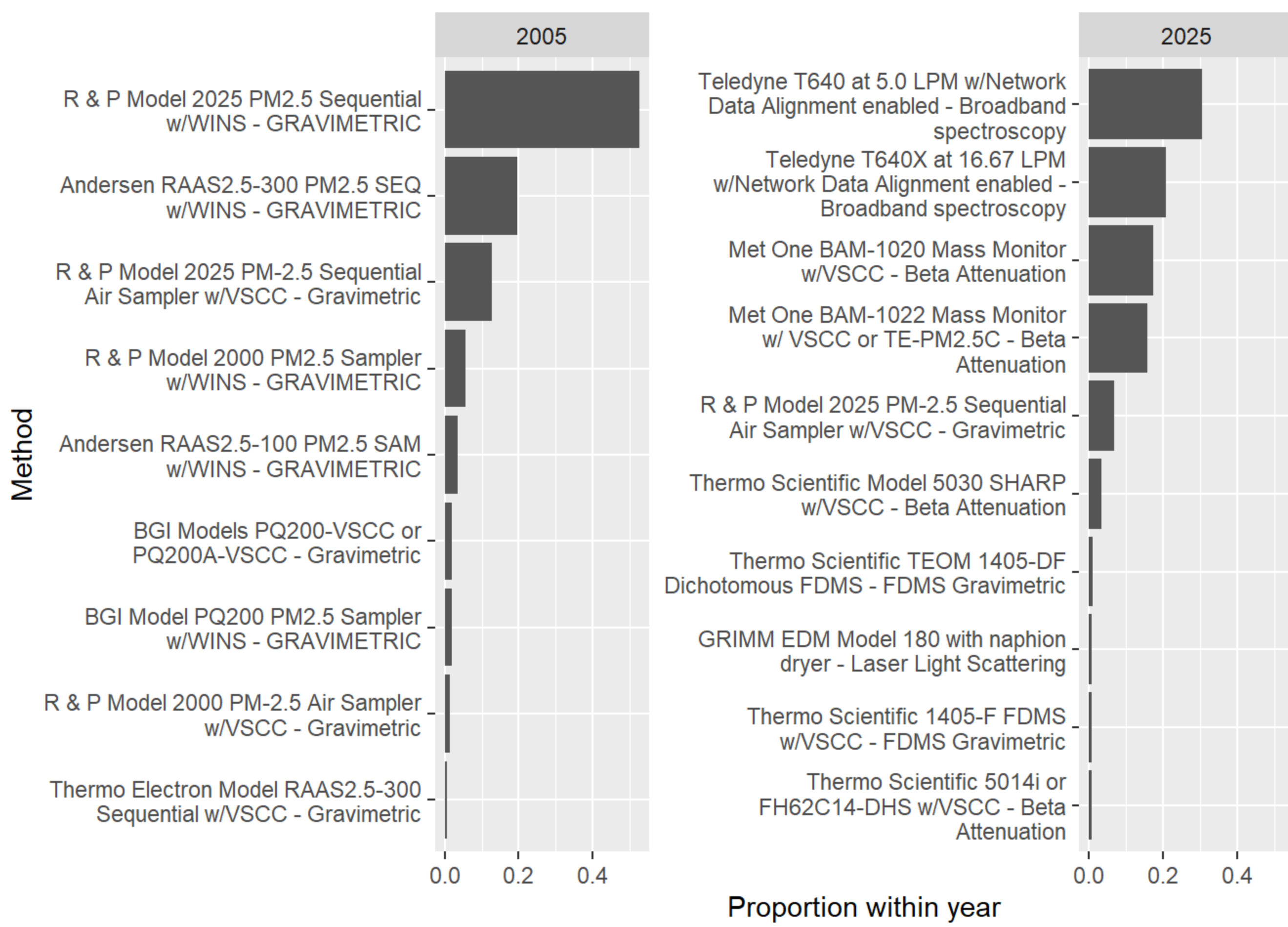
```
[10] "Met One BAM-1022 Mass Monitor w/ VSCC or TE-PM2.5C - Beta Attenuation"
[11] "Teledyne T640X at 16.67 LPM w/Network Data Alignment enabled - Broadband spectroscopy"
[12] "Met One BAM-1020 Mass Monitor w/VSCC - Beta Attenuation"
[13] "Thermo Scientific TEOM 1405-DF Dichotomous FDMS - FDMS Gravimetric"
[14] "Met One E-FRM PM2.5 with VSCC - Gravimetric"
[15] "Teledyne T640X at 16.67 LPM - Broadband spectroscopy"
[16] "Teledyne T640X at 16.67 LPM (Corrected) - Broadband spectroscopy"
[17] "Thermo Scientific TEOM 1400 FDMS or 1405 8500C FDMS w/VSCC - FDMS Gravimetric"
[18] "Teledyne T640 at 5.0 LPM - Broadband spectroscopy"
[19] "Teledyne T640 at 5.0 LPM w/Network Data Alignment enabled - Broadband spectroscopy"
[20] "Met One E-SEQ-FRM PM2.5 with VSCC - Gravimetric"
[21] "GRIMM EDM Model 180 with naphion dryer - Laser Light Scattering"
[22] "Thermo Scientific 5014i or FH62C14-DHS w/VSCC - Beta Attenuation"
[23] "Thermo Scientific 1405-F FDMS w/VSCC - FDMS Gravimetric"
[24] "Tisch Model TE-Wilbur2.5 Low-Volume Sampler - Gravimetric"
[25] "Thermo Scientific Model 5030 SHARP w/VSCC - Beta Attenuation"
```

```r
library(dplyr)

df <- pm_all %>%
  mutate(
    year = as.integer(year),
    method_name = Method.Name
  )

# 1) Proportion of missing Method.Name by year
miss_by_year <- df %>%
  group_by(year) %>%
  summarize(
    n = n(),
    n_missing = sum(is.na(method_name)),
    prop_missing = mean(is.na(method_name)),
    .groups = "drop"
  )

miss_by_year
```

```
# A tibble: 2 × 4
   year      n n_missing prop_missing
  <int>  <int>     <int>        <dbl>
1  2005 145913         0            0
2  2025 466760         0            0
```

```r
library(stringr)

# 2) Distribution of Method.Name within each year (proportions among non-missing)
method_dist <- df %>%
  filter(!is.na(method_name)) %>%
  count(year, method_name, name = "n") %>%
  group_by(year) %>%
```

```r
  mutate(prop = n / sum(n)) %>%
  arrange(year, desc(n)) %>%
  ungroup()

# 3) Top 10 method codes per year
top10 <- method_dist %>%
  group_by(year) %>%
  slice_max(n, n = 10, with_ties = FALSE) %>%
  ungroup() %>%
  mutate(method_wrap = str_wrap(method_name, width = 35))

ggplot(top10, aes(x = prop, y = reorder(method_wrap, prop))) +
  geom_col() +
  facet_wrap(~ year, scales = "free_y") +
  labs(x = "Proportion within year", y = "Method")
```



From the [EPA PM2.5 Codetable](), we find that the most common method names in 2005 correspond to codes 117 and 120, respectively, while the most common codes in 2025 are 636 and 638, respectively. However, the more helpful information is the proportion with respect to method names.

From 2005 to 2025, there is a big shift in the dominant measurement technology - instead of the overwhelming presence of gravimetric, filter-based samplers, there are more continuous and automated instrumenets (such as Teledyne for broadband spectroscopy and Met One for beta attenuation). Overall, there is a more diverse set of instruments being used in 2025 compared to in 2005, with the most frequently used method in 2025 being a smaller proportion (about 30% of the time) compared to in 2005 (over 50% of the time). These temporal patterns are important to analyze because they reflect how methodologies have been modernized over the past 20 years to include new types of instruments. In this case, the new instruments can measure PM2.5 concentrations in real time instead of only collecting samples at scheduled times.

5. Visualize and summarize daily concentrations of PM2.5

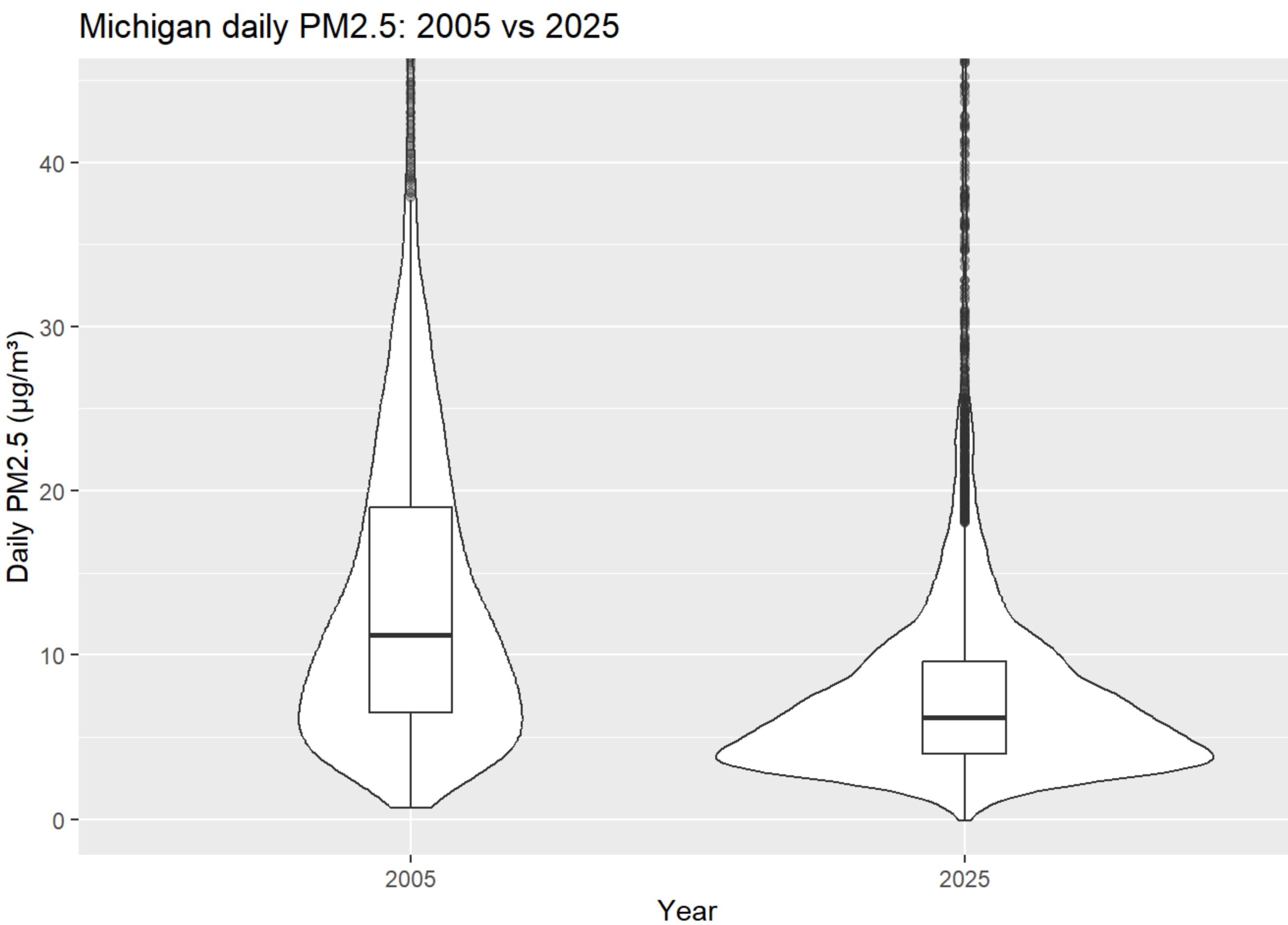
```r
# Standardize, restrict to Michigan, and keep only years of interest
library(dplyr)
library(ggplot2)
library(stringr)

pm_mi <- pm_all %>%
  mutate(
    state  = if ("state"  %in% names(.)) state  else State.Code,
    county = if ("county" %in% names(.)) county else County.Code,
    site   = if ("site"   %in% names(.)) site   else Site.Num,
    pm25   = if ("pm25"   %in% names(.)) pm25   else as.numeric(Arithmetic.Mean),
    Date   = if ("Date"   %in% names(.)) as.Date(Date) else as.Date(Date.Local),
    year   = if ("year"   %in% names(.)) year   else as.integer(format(Date, "%Y")),
    county_name = if ("County.Name" %in% names(.)) County.Name else NA_character_,
    city_name   = if ("City.Name"   %in% names(.)) City.Name   else NA_character_,
    lat = as.numeric(if ("Latitude"  %in% names(.)) Latitude  else NA),
    lon = as.numeric(if ("Longitude" %in% names(.)) Longitude else NA)
  ) %>%
  filter(state == "Michigan", year %in% c(2005, 2025))
```

Level 1: PM2 concentrations in Michigan overall

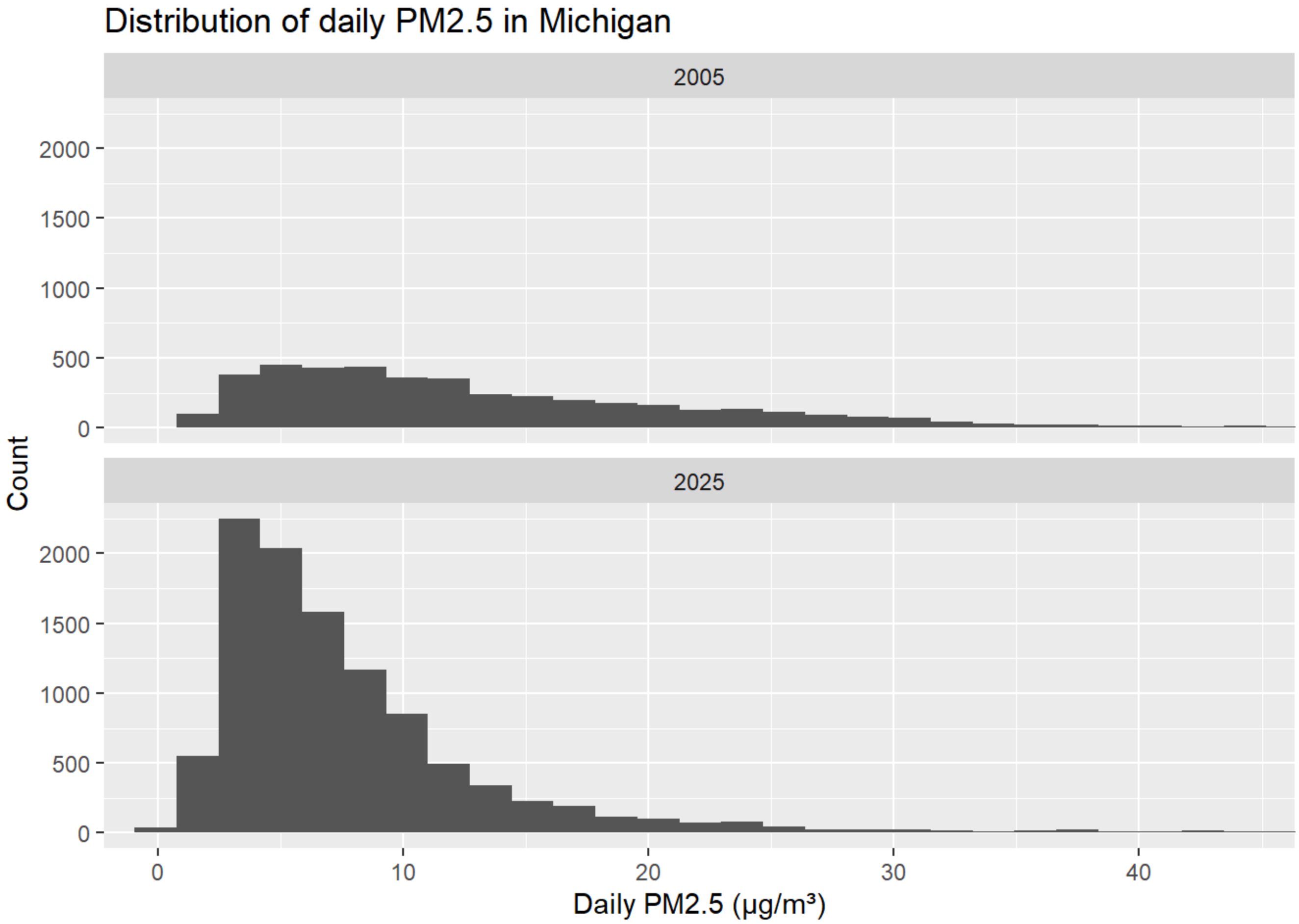
```r
state_summary <- pm_mi %>%
  group_by(year) %>%
  summarize(
    n_days = sum(!is.na(pm25)),
    mean = mean(pm25, na.rm = TRUE),
    median = median(pm25, na.rm = TRUE),
    p25 = quantile(pm25, 0.25, na.rm = TRUE),
    p75 = quantile(pm25, 0.75, na.rm = TRUE),
    p95 = quantile(pm25, 0.95, na.rm = TRUE),
    max = max(pm25, na.rm = TRUE),
    .groups = "drop"
  )

state_summary
```

```
# A tibble: 2 × 8
   year n_days  mean median   p25   p75   p95   max
  <int>  <int> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
1  2005   4391 14.0    11.2   6.5  19    32.3  81.6
2  2025  10382  8.03    6.2   4     9.6  19.9 101.
```

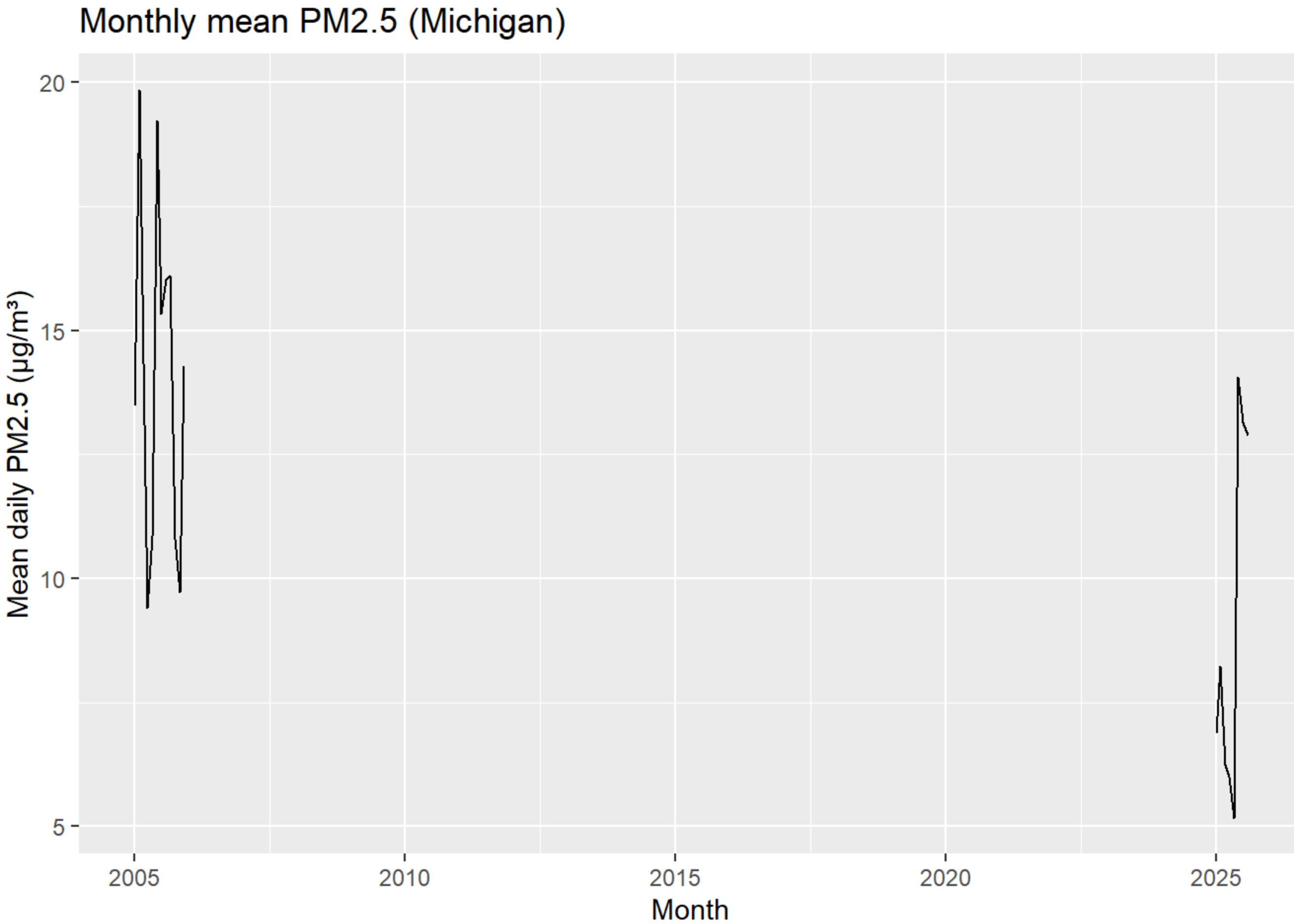
```r
ggplot(pm_mi, aes(x = factor(year), y = pm25)) +
  geom_violin(trim = TRUE) +
  geom_boxplot(width = 0.15, outlier.alpha = 0.2) +
  labs(x = "Year", y = "Daily PM2.5 (µg/m³)", title = "Michigan daily PM2.5: 2005 vs 2025") +
  coord_cartesian(ylim = c(0, quantile(pm_mi$pm25, 0.99, na.rm = TRUE)))
```




```r
ggplot(pm_mi, aes(x = pm25)) +
  geom_histogram(bins = 60) +
  facet_wrap(~ year, ncol = 1) +
  labs(x = "Daily PM2.5 (µg/m³)", y = "Count", title = "Distribution of daily PM2.5 in Michigan")
  coord_cartesian(xlim = c(0, quantile(pm_mi$pm25, 0.99, na.rm = TRUE)))
```




```
pm_mi_month <- pm_mi %>%
  mutate(month = as.Date(format(Date, "%Y-%m-01"))) %>%
  group_by(year, month) %>%
  summarize(pm25_mean = mean(pm25, na.rm = TRUE), .groups = "drop")

ggplot(pm_mi_month, aes(x = month, y = pm25_mean, group = factor(year))) +
  geom_line() +
  labs(x = "Month", y = "Mean daily PM2.5 (µg/m³)", title = "Monthly mean PM2.5 (Michigan)")
```



The summary statistics show that the mean and all of the percentiles (except for the max) of the daily PM2.5 concentration per month have decreased from 2005 to 2025; for example, the mean concentration decreased from 13.99 to 8.03 µg/m^3. The side-by-side box/violin plots and histograms also illustrate this well by showing a high frequency of values around 5 µg/m^3 in 2025, while concentrations of 20 µg/m^3 or higher (which were considered evenly distrbuted in 2005) are considered outliers in 2025. In general, the violin plot and line plot show that the distribution of concentrations has shifted down.

Level 2: PM2 concentrations by Michigan county

```
county_summary <- pm_mi %>%
  filter(!is.na(county)) %>%
  group_by(year, county) %>%
  summarize(
    n = sum(!is.na(pm25)),
    mean = mean(pm25, na.rm = TRUE),
    median = median(pm25, na.rm = TRUE),
    p95 = quantile(pm25, 0.95, na.rm = TRUE),
    .groups = "drop"
  )

county_summary
```

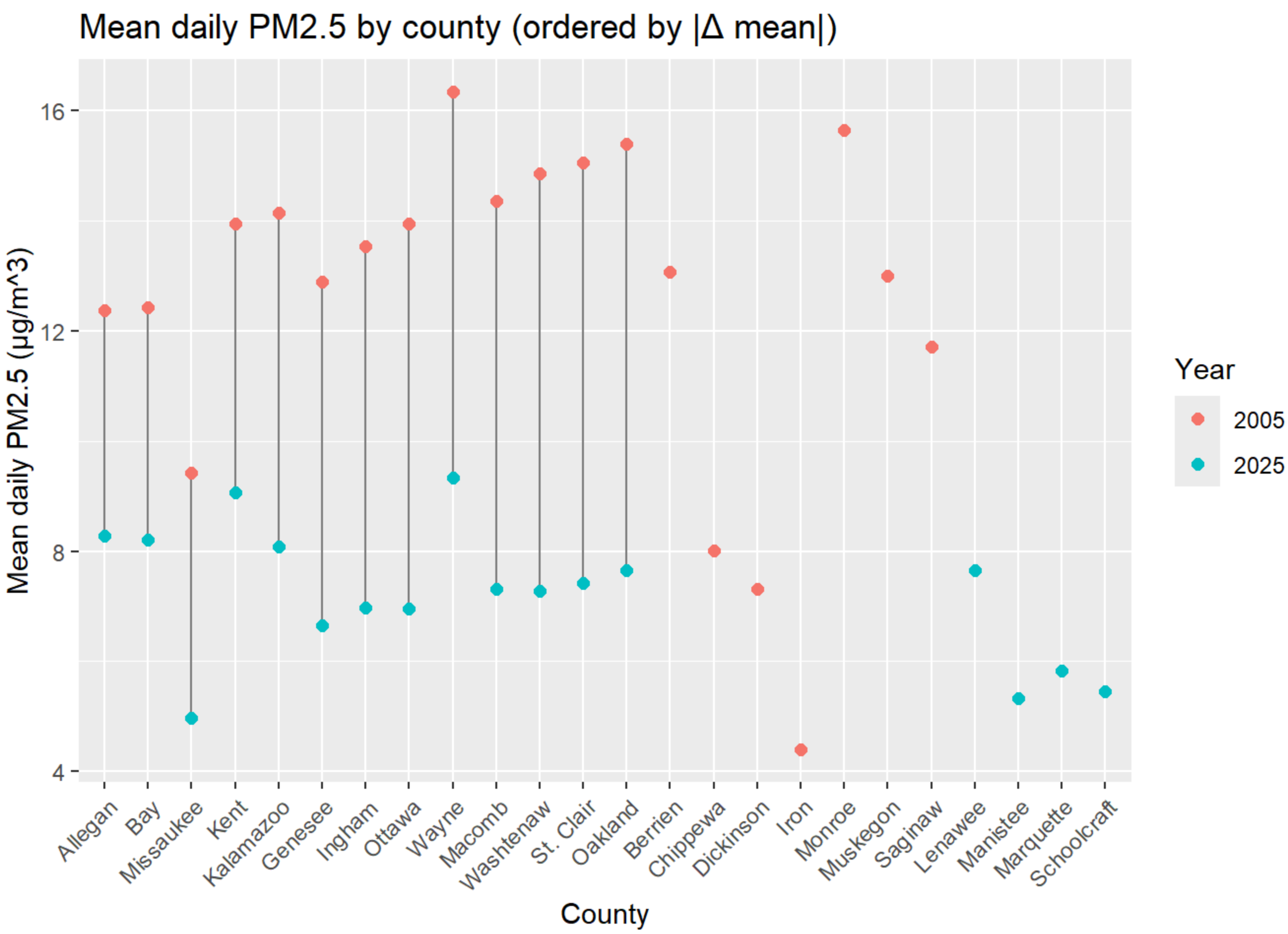
```
# A tibble: 37 × 6
    year county        n  mean median   p95
   <int> <chr>     <int> <dbl>  <dbl> <dbl>
 1  2005 Allegan     361 12.4   10     31
 2  2005 Bay         117 12.4    9.7   28.2
 3  2005 Berrien     118 13.1   10.8   31.3
 4  2005 Chippewa    294  8.01   5.85 20.8
 5  2005 Dickinson    25  7.32   6.9  13.2
 6  2005 Genesee     119 12.9   10.7  27.4
 7  2005 Ingham      120 13.5   11.2  30.0
 8  2005 Iron         25  4.40   3.5   9.36
 9  2005 Kalamazoo   167 14.1   12.6  30.6
10  2005 Kent        416 13.9   11    32.2
# ℹ 27 more rows
```

```
county_change <- county_summary %>%
  select(year, county, mean, median, p95) %>%
  tidyr::pivot_wider(names_from = year, values_from = c(mean, median, p95)) %>%
  mutate(
    d_mean   = mean_2025   - mean_2005,
    d_median = median_2025 - median_2005,
    d_p95    = p95_2025    - p95_2005
  )

county_change %>% arrange(d_mean)
```

```
# A tibble: 24 × 10
   county   mean_2005 mean_2025 median_2005 median_2025 p95_2005 p95_2025 d_mean
   <chr>        <dbl>     <dbl>       <dbl>       <dbl>    <dbl>    <dbl>  <dbl>
 1 Oakland       15.4      7.66        12.3        5.76     36.1     19.7  -7.74
 2 St. Cla…      15.0      7.43        12          5.72     36.6     17.8  -7.62
 3 Washten…      14.9      7.28        11.8        5.57     33.3     17.5  -7.58
 4 Macomb        14.4      7.32        10.8        5.43     32.6     17.5  -7.04
 5 Wayne         16.3      9.33        13.5        7.2      35.6     22.6  -7.01
 6 Ottawa        13.9      6.95        11.9        5.15     32.1     18.5  -6.99
 7 Ingham        13.5      6.96        11.2        5.4      30.0     16.0  -6.57
 8 Genesee       12.9      6.65        10.7        5.4      27.4     16.3  -6.23
 9 Kalamaz…      14.1      8.07        12.6        6.22     30.6     19.6  -6.07
10 Kent          13.9      9.06        11          7.8      32.2     20.1  -4.88
# ℹ 14 more rows
# ℹ 2 more variables: d_median <dbl>, d_p95 <dbl>
```

```
county_levels <- county_change %>%
  arrange(abs(d_mean)) %>%
  pull(county)

plot_df <- county_summary %>%
  mutate(county = factor(county, levels = county_levels))

ggplot(plot_df, aes(x = county, y = mean, color = factor(year), group = county)) +
```

```
  geom_line(color = "grey50") +
  geom_point(size = 2) +
  labs(x = "County", y = "Mean daily PM2.5 (µg/m^3)", color = "Year",
       title = "Mean daily PM2.5 by county (ordered by |Δ mean|)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

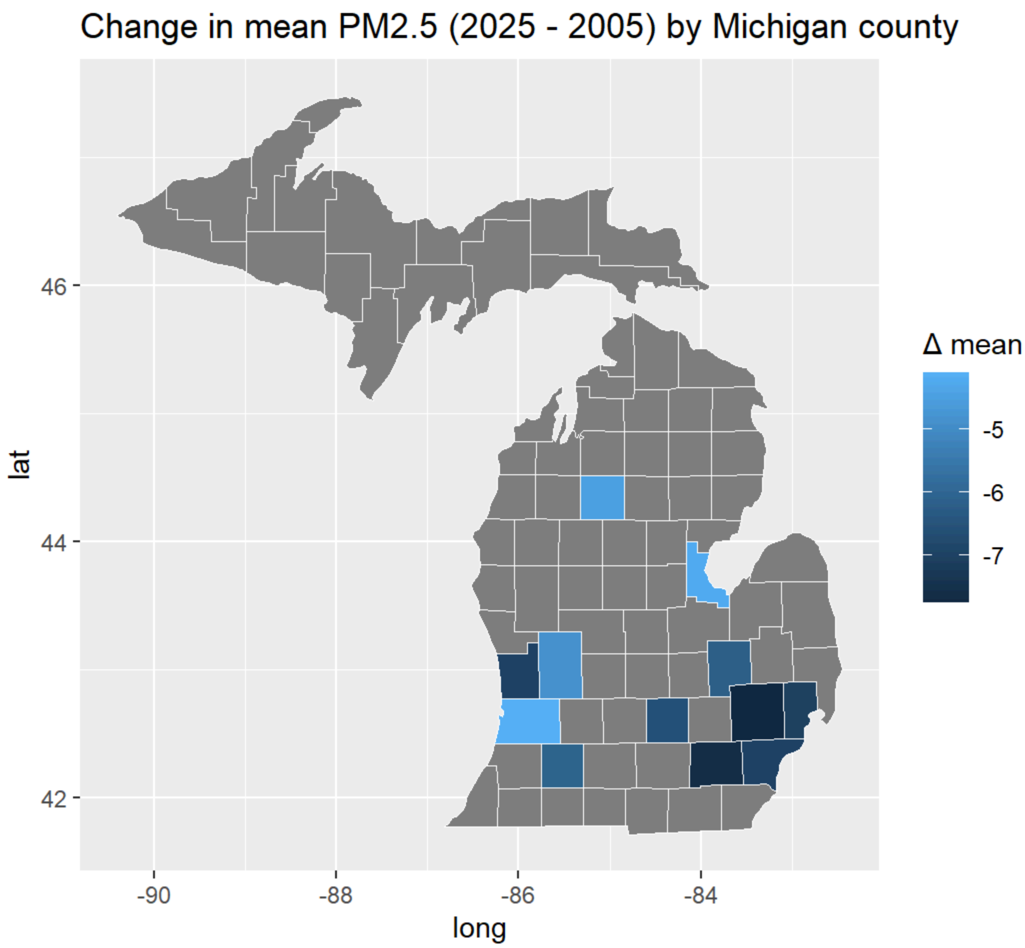


From this plot, we can see that out of every county in Michigan where measurements were recorded in both 2005 and 2025, all of the measurements from 2025 are significantly smaller than those in 2005. According to the summary statistics, the median and 95th percentile also uniformly decrease from 2005 to 2025. The differences in concentrations range from -4.09 (in Allegan) to -7.74 µ/m^3 (in Oakland), meaning the concentrations decreased by at least 4 µ/m^3 and up to almost 8 µ/m^3 by county from 2005 to 2025.

We also see an overall decrease when comparing the 7 measurements for counties only taken in 2005 compared to the 4 measurements for counties only taken in 2025, since none of the measurements from counties only recorded in 2025 are higher than 8 µg/m^3.

```
library(maps)

mi_map <- map_data("county") %>%
  filter(region == "michigan") %>%
  mutate(county = str_to_title(subregion))

county_map_df <- mi_map %>%
```

```
  left_join(county_change, by = "county")

ggplot(county_map_df, aes(long, lat, group = group, fill = d_mean)) +
  geom_polygon(color = "white", linewidth = 0.2) +
  coord_quickmap() +
  labs(title = "Change in mean PM2.5 (2025 - 2005) by Michigan county",
       fill = "Δ mean")
```



This choropleth shows the changes in mean PM2.5 concentrations per Michigan county from 2005 to 2025. We can see that the largest decreases (the deepest blue colors) are in counties in the southeast region of Michigan, which (as discussed previously) is where Metro Detroit is located. This suggests that the decreases are concentrated in metro/industrial areas, and the decreases are less pronounced in more rural counties closer to the north side of Michigan.

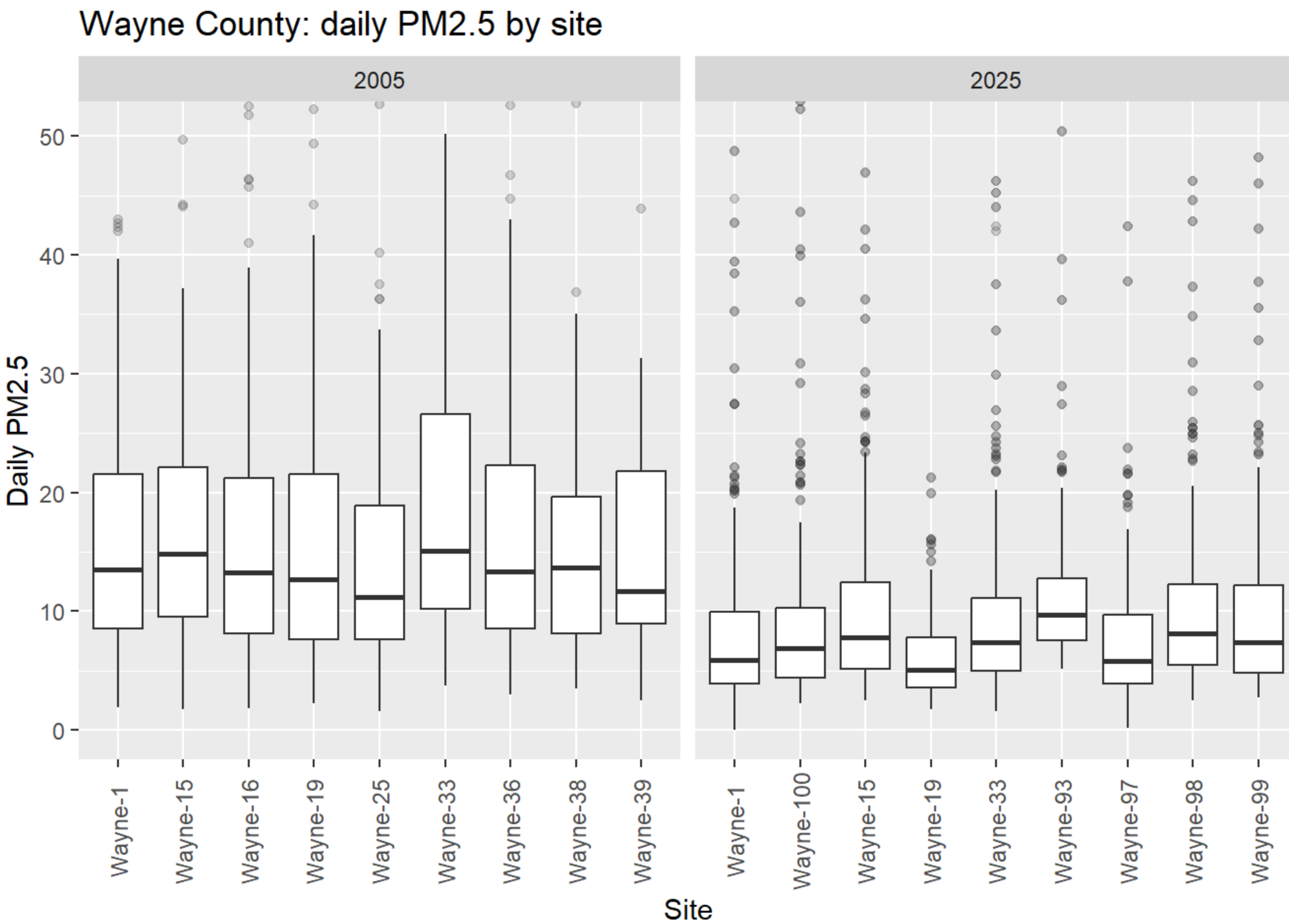
Level 3: PM2 concentrations by site in Wayne County

```
pm_wayne <- pm_mi %>%
  filter(county == "Wayne") %>%
  mutate(site_id = paste0(county, "-", site))

site_summary <- pm_wayne %>%
  group_by(year, site_id, city_name) %>%
```

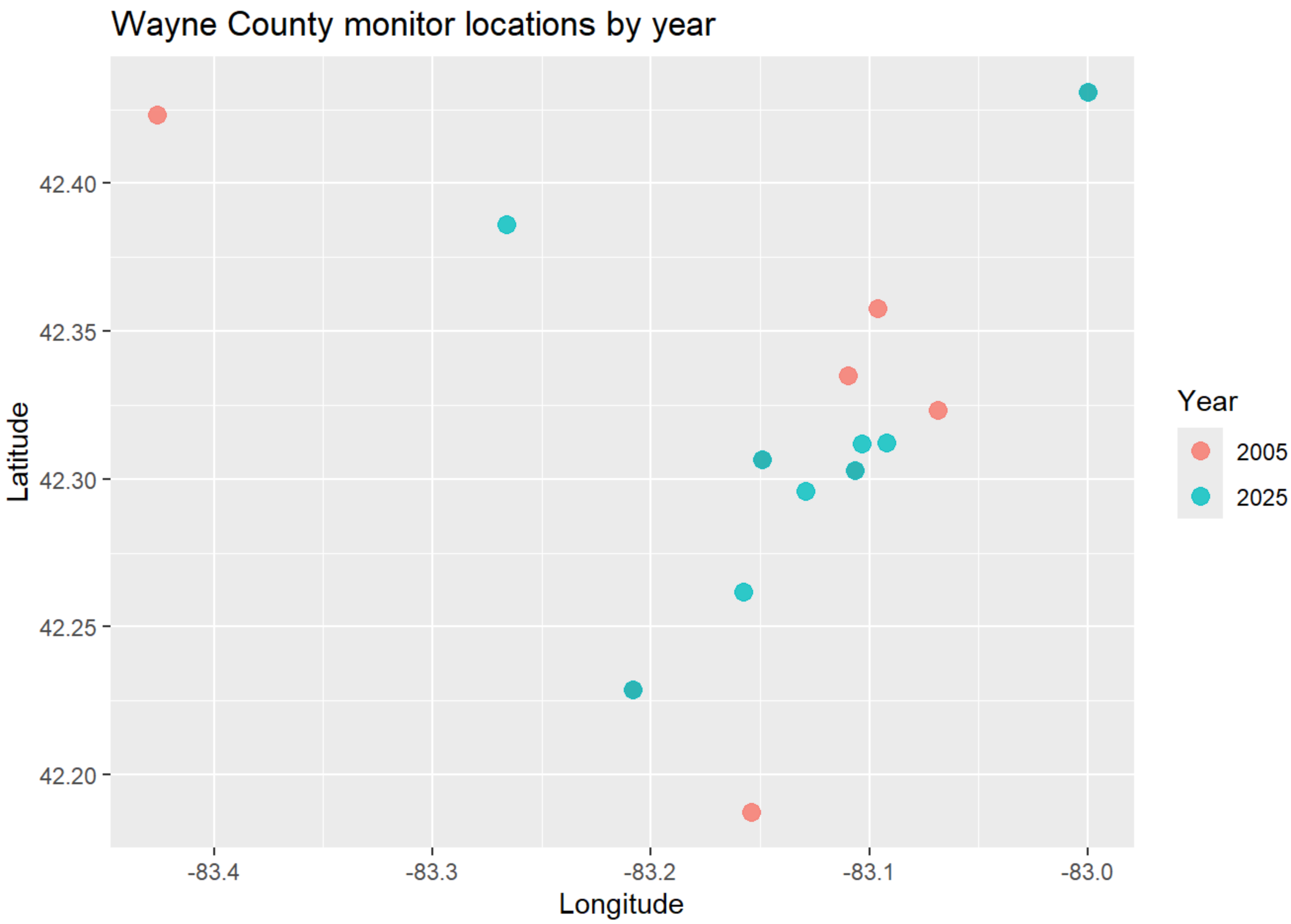
```r
  summarize(
    n = sum(!is.na(pm25)),
    mean = mean(pm25, na.rm = TRUE),
    median = median(pm25, na.rm = TRUE),
    p95 = quantile(pm25, 0.95, na.rm = TRUE),
    .groups = "drop"
  )

site_summary
```

```
# A tibble: 18 × 7
    year site_id   city_name     n  mean median   p95
   <int> <chr>     <chr>     <int> <dbl>  <dbl> <dbl>
 1  2005 Wayne-1   <NA>        408 16.2   13.5   33.7
 2  2005 Wayne-15  <NA>        114 17.2   14.8   34.2
 3  2005 Wayne-16  <NA>        338 16.0   13.2   35.4
 4  2005 Wayne-19  <NA>        117 16.4   12.7   41.5
 5  2005 Wayne-25  <NA>        114 14.9   11.2   34.6
 6  2005 Wayne-33  <NA>        115 18.5   15.1   36.2
 7  2005 Wayne-36  <NA>        113 16.4   13.3   34.2
 8  2005 Wayne-38  <NA>         75 16.4   13.7   35.6
 9  2005 Wayne-39  <NA>         35 15.0   11.7   28.8
10  2025 Wayne-1   <NA>        533  8.32   5.91  20.5
11  2025 Wayne-100 <NA>        486  8.97   6.86  22.4
12  2025 Wayne-15  <NA>        486 10.3    7.8   24.7
13  2025 Wayne-19  <NA>        282  6.27   5.08  13.5
14  2025 Wayne-33  <NA>        555  9.71   7.34  23.4
15  2025 Wayne-93  <NA>        330 11.3    9.7   21.8
16  2025 Wayne-97  <NA>        362  7.67   5.81  19.2
17  2025 Wayne-98  <NA>        486 10.4    8.1   24.9
18  2025 Wayne-99  <NA>        428 10.1    7.34  25.0
```

```r
ggplot(pm_wayne, aes(x = site_id, y = pm25)) +
  geom_boxplot(outlier.alpha = 0.2) +
  facet_wrap(~ year, scales = "free_x") +
  labs(x = "Site", y = "Daily PM2.5", title = "Wayne County: daily PM2.5 by site") +
  coord_cartesian(ylim = c(0, quantile(pm_wayne$pm25, 0.99, na.rm = TRUE))) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))
```




```r
wayne_sites <- pm_wayne %>%
  filter(!is.na(lat), !is.na(lon)) %>%
  distinct(year, site_id, city_name, lat, lon)

ggplot(wayne_sites, aes(x = lon, y = lat, color = factor(year))) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "Longitude", y = "Latitude", color = "Year",
       title = "Wayne County monitor locations by year") +
  coord_quickmap()
```



From the series of boxplots that show the distribution of daily PM2.5 concentrations in all of the Wayne County sites, we see that the medians have uniformly decreased from 2005 to 2025; in fact, according to the summary statistics, the lowest mean and median concentrations in 2005 are still higher than the highest mean and median concentrations in 2025. The boxplots also show (much like the broad, state-wide analysis from before) that the concentrations above 20 µg/m^3 are within the 75th percentile in 2005, but outliers in 2025. Finally, the spatial plot of site locations colored by year shows that the sites in 2005 (red points) are more spread out across Wayne County, but in 2025 (blue points), they are more clustered in one area between latitude 42.30 and longitude -83.15, with only three sites located away from that cluster.

Therefore, we can finally conclude that on the level of the entire state of Michigan, on the level of each county within Michigan, and the level of each observation site within Wayne County, the daily concentrations of PM2.5 have decreased in Michigan over the 20 years spanning from 2005 to 2025.