Lab Assignment Nine

Noah Gallego

2024-09-29

Lab Assignment 9: Air Pollutant Analysis

Import Libraries

```
##
## 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
```

Read Data into R

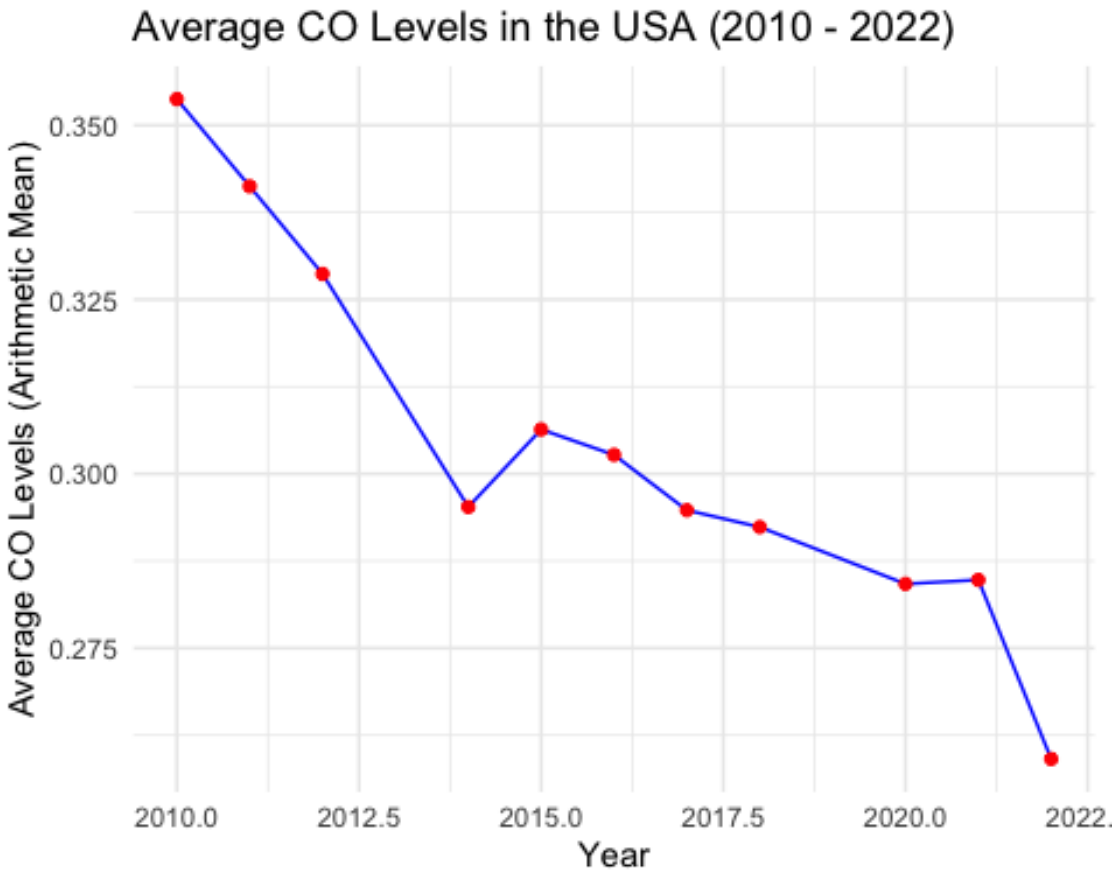
```
folder_path = "../Data/"

# List all the Excel files from 2010 to 2022
file_list = list.files(path = folder_path, pattern = "daily_42101_[0-
9]{4}.csv", full.names = TRUE)

# Use lapply to read all the Excel files into a list
all_data = lapply(file_list, function(file) read.csv(file, header = TRUE))

# Combine the list of data frames into one large data frame
combined_data = bind_rows(all_data)
df = combined_data

# View the first few rows of the combined data
head(df)

##   State.Code County.Code Site.Num Parameter.Code POC Latitude Longitude
Datum
## 1          1          73       28          42101   1 33.52944 -86.85028
WGS84
## 2          1          73       28          42101   1 33.52944 -86.85028
WGS84
## 3          1          73       28          42101   1 33.52944 -86.85028
WGS84
## 4          1          73       28          42101   1 33.52944 -86.85028
```

```
WGS84
## 5          1          73       28          42101   1 33.52944 -86.85028
WGS84
## 6          1          73       28          42101   1 33.52944 -86.85028
WGS84
##    Parameter.Name Sample.Duration Pollutant.Standard Date.Local
## 1 Carbon monoxide          1 HOUR     CO 1-hour 1971 2010-01-01
## 2 Carbon monoxide          1 HOUR     CO 1-hour 1971 2010-01-02
## 3 Carbon monoxide          1 HOUR     CO 1-hour 1971 2010-01-03
## 4 Carbon monoxide          1 HOUR     CO 1-hour 1971 2010-01-04
## 5 Carbon monoxide          1 HOUR     CO 1-hour 1971 2010-01-05
## 6 Carbon monoxide          1 HOUR     CO 1-hour 1971 2010-01-06
##    Units.of.Measure Event.Type Observation.Count Observation.Percent
## 1 Parts per million       None                24                 100
## 2 Parts per million       None                24                 100
## 3 Parts per million       None                24                 100
## 4 Parts per million       None                24                 100
## 5 Parts per million       None                23                  96
## 6 Parts per million       None                24                 100
##   Arithmetic.Mean X1st.Max.Value X1st.Max.Hour AQI Method.Code
## 1        0.470833            0.6            18  NA          54
## 2        0.479167            0.5             0  NA          54
## 3        0.462500            0.5             0  NA          54
## 4        0.579167            0.8            18  NA          54
## 5        0.582609            0.8             6  NA          54
## 6        0.612500            1.4            23  NA          54
##                            Method.Name Local.Site.Name
## 1 INSTRUMENTAL - NONDISPERSIVE INFRARED
## 2 INSTRUMENTAL - NONDISPERSIVE INFRARED
## 3 INSTRUMENTAL - NONDISPERSIVE INFRARED
## 4 INSTRUMENTAL - NONDISPERSIVE INFRARED
## 5 INSTRUMENTAL - NONDISPERSIVE INFRARED
## 6 INSTRUMENTAL - NONDISPERSIVE INFRARED
##                                    Address State.Name County.Name
City.Name
## 1 EAST THOMAS, FINLEY, 841 FINLEY AVE. BP.    Alabama   Jefferson
Birmingham
## 2 EAST THOMAS, FINLEY, 841 FINLEY AVE. BP.    Alabama   Jefferson
Birmingham
## 3 EAST THOMAS, FINLEY, 841 FINLEY AVE. BP.    Alabama   Jefferson
Birmingham
## 4 EAST THOMAS, FINLEY, 841 FINLEY AVE. BP.    Alabama   Jefferson
Birmingham
## 5 EAST THOMAS, FINLEY, 841 FINLEY AVE. BP.    Alabama   Jefferson
Birmingham
## 6 EAST THOMAS, FINLEY, 841 FINLEY AVE. BP.    Alabama   Jefferson
Birmingham
##               CBSA.Name Date.of.Last.Change
## 1 Birmingham-Hoover, AL          2021-11-08
## 2 Birmingham-Hoover, AL          2021-11-08
```

```
## 3 Birmingham-Hoover, AL          2021-11-08
## 4 Birmingham-Hoover, AL          2021-11-08
## 5 Birmingham-Hoover, AL          2021-11-08
## 6 Birmingham-Hoover, AL          2021-11-08
```

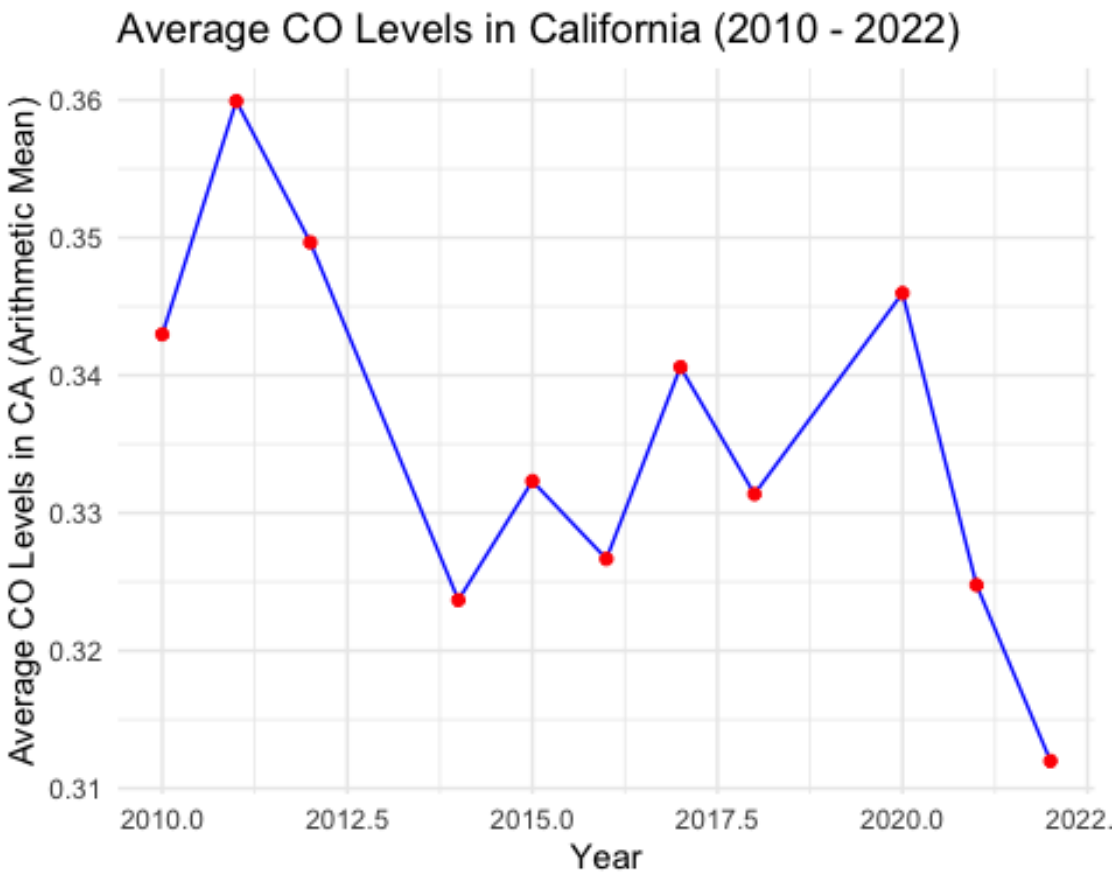
Plotting CO Levels:

```
# Map Levels Of CO vs Time

# Convert to DateTime
df$Date.Local = as.Date(df$Date.Local)
df$Year = format(df$Date.Local, "%Y")

# Get Yearly CO Levels & Group By Year
yearly_co_levels = df %>%
  group_by(Year) %>%
  summarise(mean_CO = mean(Arithmetic.Mean, na.rm = TRUE))

# Plot CO2 Over Years
yearly_co_levels %>%
  ggplot(aes(x = as.numeric(Year), y = mean_CO)) +
  geom_line(color = "blue") +
  geom_point(color = "red") +
  labs(title = "Average CO Levels in the USA (2010 - 2022)",
      x = "Year",
      y = "Average CO Levels (Arithmetic Mean)") +
      theme_minimal()
```



Plotting CO Levels (California Only)

```
# Get California-Only DF
ca_df = df %>%
  filter(State.Name == "California" & State.Code == 6)

# Get Yearly CO Levels & Group By Year
yearly_co_levels_ca = ca_df %>%
  group_by(Year) %>%
  summarise(mean_CO = mean(Arithmetic.Mean, na.rm = TRUE))

# Plot CO2 Over Years
yearly_co_levels_ca %>%
  ggplot(aes(x = as.numeric(Year), y = mean_CO)) +
  geom_line(color = "blue") +
  geom_point(color = "red") +
  labs(title = "Average CO Levels in California (2010 - 2022)",
       x = "Year",
       y = "Average CO Levels in CA (Arithmetic Mean)") +
       theme_minimal()
```



Reading SO_2 Data

```
# Fetch File List
remaining_file_list = list.files(path = folder_path, pattern =
"daily_42401_[0-9]{4}.csv", full.names = TRUE)

# Use lapply to read all the Excel files into a list
all_data_remaining = lapply(remaining_file_list, function(file)
read.csv(file, header = TRUE))

# Combine the list of data frames into one large data frame
so2_df = bind_rows(all_data_remaining)

# View the first few rows of the combined data
head(so2_df)
```

```
##   State.Code County.Code Site.Num Parameter.Code POC Latitude Longitude
Datum
## 1          1          73     1003          42401   1 33.48556   -86.915
WGS84
## 2          1          73     1003          42401   1 33.48556   -86.915
WGS84
## 3          1          73     1003          42401   1 33.48556   -86.915
WGS84
```

```
## 4          1          73     1003          42401   1 33.48556   -86.915
WGS84
## 5          1          73     1003          42401   1 33.48556   -86.915
WGS84
## 6          1          73     1003          42401   1 33.48556   -86.915
WGS84
##   Parameter.Name Sample.Duration Pollutant.Standard Date.Local
## 1 Sulfur dioxide          1 HOUR    SO2 1-hour 2010 2010-01-01
## 2 Sulfur dioxide          1 HOUR    SO2 1-hour 2010 2010-01-02
## 3 Sulfur dioxide          1 HOUR    SO2 1-hour 2010 2010-01-03
## 4 Sulfur dioxide          1 HOUR    SO2 1-hour 2010 2010-01-04
## 5 Sulfur dioxide          1 HOUR    SO2 1-hour 2010 2010-01-05
## 6 Sulfur dioxide          1 HOUR    SO2 1-hour 2010 2010-01-06
##    Units.of.Measure Event.Type Observation.Count Observation.Percent
## 1 Parts per billion       None                24                 100
## 2 Parts per billion       None                24                 100
## 3 Parts per billion       None                24                 100
## 4 Parts per billion       None                24                 100
## 5 Parts per billion       None                24                 100
## 6 Parts per billion       None                24                 100
##   Arithmetic.Mean X1st.Max.Value X1st.Max.Hour AQI Method.Code
## 1        1.291667              2             7   3          60
## 2        1.208333              3             7   4          60
## 3        2.708333              8             8  11          60
## 4        2.958333              4             8   6          60
## 5        5.833333             22            10  31          60
## 6        6.833333             30            15  43          60
##                         Method.Name Local.Site.Name
## 1 INSTRUMENTAL - PULSED FLUORESCENT       Fairfield
## 2 INSTRUMENTAL - PULSED FLUORESCENT       Fairfield
## 3 INSTRUMENTAL - PULSED FLUORESCENT       Fairfield
## 4 INSTRUMENTAL - PULSED FLUORESCENT       Fairfield
## 5 INSTRUMENTAL - PULSED FLUORESCENT       Fairfield
## 6 INSTRUMENTAL - PULSED FLUORESCENT       Fairfield
##                        Address State.Name County.Name City.Name
## 1 FAIRFIELD, PFD, 5229 COURT B    Alabama   Jefferson Fairfield
## 2 FAIRFIELD, PFD, 5229 COURT B    Alabama   Jefferson Fairfield
## 3 FAIRFIELD, PFD, 5229 COURT B    Alabama   Jefferson Fairfield
## 4 FAIRFIELD, PFD, 5229 COURT B    Alabama   Jefferson Fairfield
## 5 FAIRFIELD, PFD, 5229 COURT B    Alabama   Jefferson Fairfield
## 6 FAIRFIELD, PFD, 5229 COURT B    Alabama   Jefferson Fairfield
##               CBSA.Name Date.of.Last.Change
## 1 Birmingham-Hoover, AL          2021-11-09
## 2 Birmingham-Hoover, AL          2021-11-09
## 3 Birmingham-Hoover, AL          2021-11-09
## 4 Birmingham-Hoover, AL          2021-11-09
## 5 Birmingham-Hoover, AL          2021-11-09
## 6 Birmingham-Hoover, AL          2021-11-09
```

Merging CO & SO_2 DataFrame

```
# Clean
co_df = df
co_df$Date.Local = as.Date(co_df$Date.Local)
co_df$Year = format(co_df$Date.Local, "%Y")

so2_df$Date.Local = as.Date(so2_df$Date.Local)
so2_df$Year = format(so2_df$Date.Local, "%Y")

co_df = co_df %>% distinct(Date.Local, .keep_all = TRUE)
so2_df = so2_df %>% distinct(Date.Local, .keep_all = TRUE)

co_df = co_df %>% select(Date.Local, Arithmetic.Mean)
so2_df = so2_df %>% select(Date.Local, Arithmetic.Mean)

sum(duplicated(co_df$Date.Local))

## [1] 0

sum(duplicated(so2_df$Date.Local))

## [1] 0

# Merge using inner_join to only include matching dates
merged_data = inner_join(co_df, so2_df, by = "Date.Local")

# View the merged data
head(merged_data)

##   Date.Local Arithmetic.Mean.x Arithmetic.Mean.y
## 1 2010-01-01          0.470833          1.291667
## 2 2010-01-02          0.479167          1.208333
## 3 2010-01-03          0.462500          2.708333
## 4 2010-01-04          0.579167          2.958333
## 5 2010-01-05          0.582609          5.833333
## 6 2010-01-06          0.612500          6.833333
```

Calculate Monthly Means

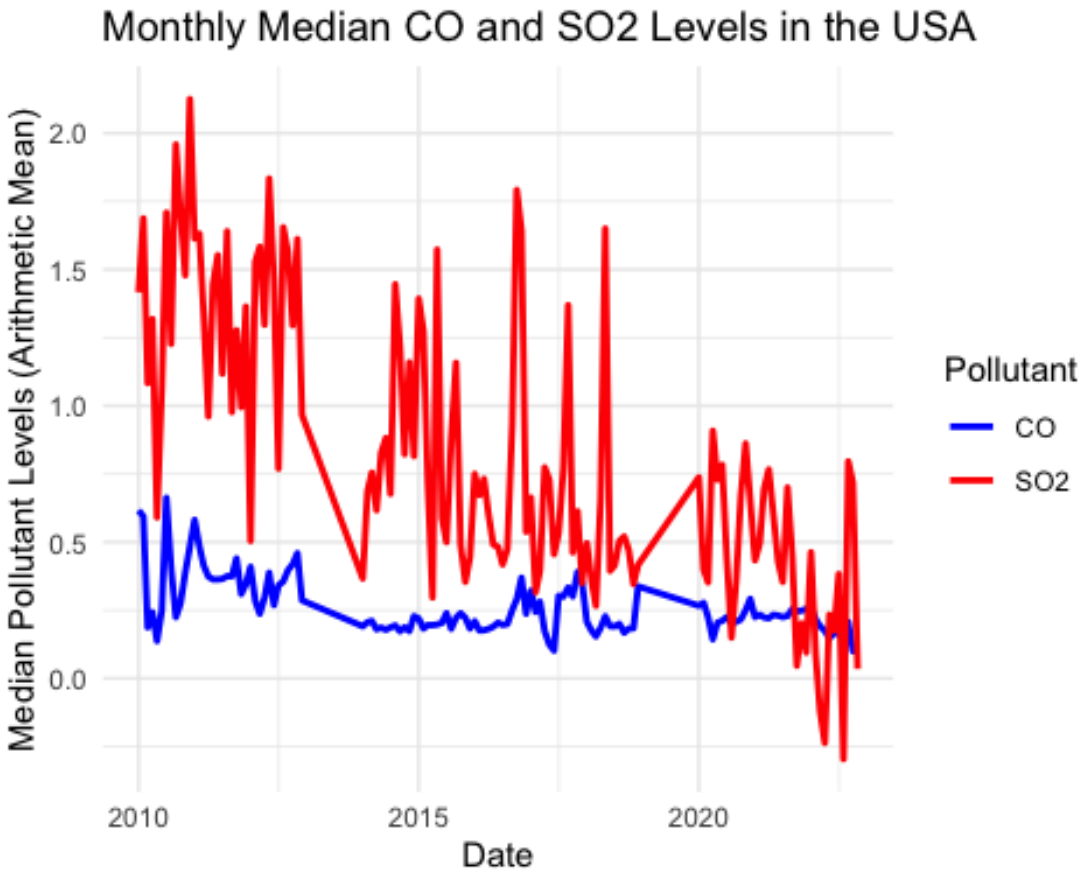
```
merged_data = merged_data %>%
  mutate(Month = format(Date.Local, "%Y-%m"))

# Calculate monthly median for both CO and SO2
monthly_medians = merged_data %>%
  group_by(Month) %>%
  summarise(monthly_median_CO = median(Arithmetic.Mean.x, na.rm = TRUE),
            monthly_median_SO2 = median(Arithmetic.Mean.y, na.rm = TRUE))

head(monthly_medians)
```

```
## # A tibble: 6 × 3
##   Month   monthly_median_CO monthly_median_SO2
##   <chr>               <dbl>              <dbl>
## 1 2010-01             0.617              1.42
## 2 2010-02             0.594              1.69
## 3 2010-03             0.188              1.08
## 4 2010-04             0.242              1.32
## 5 2010-05             0.138              0.591
## 6 2010-06             0.249              1.06
```

Visualization

```
# Plot monthly median CO and SO2 levels over time
ggplot(monthly_medians, aes(x = as.Date(paste0(Month, "-01")))) +
  geom_line(aes(y = monthly_median_CO, color = "CO"), size = 1) +
  geom_line(aes(y = monthly_median_SO2, color = "SO2"), size = 1) +
  labs(title = "Monthly Median CO and SO2 Levels in the USA",
       x = "Date",
       y = "Median Pollutant Levels (Arithmetic Mean)") +
  scale_color_manual(values = c("CO" = "blue", "SO2" = "red"), name =
"Pollutant") +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```


Monthly Median CO and SO2 Levels in the USA
Median Pollutant Levels (Arithmetic Mean)
2.0
1.5
1.0
0.5
0.0
2010
2015
2020
Date
Pollutant
CO
SO2