Final Project Exploratory Data Analysis

Aidan Frederick, Ben Walter, Kyle Maher

2025-11-12

```
library(fpp3)
set.seed(280)  # Reprodicibility
```

```
data <- readRDS("data/loaded/delhi.rds")

delhi_city <- data %>%
  filter(file_name == "DL008.csv")

observation_counts <- delhi_city %>%
  summarise(across(everything(), ~sum(!is.na(.)))) %>%
  pivot_longer(cols = everything()) %>%
  arrange(desc(value)) %>%
  mutate(percent = (value / nrow(delhi_city)) * 100)

observation_counts
```

```
# A tibble: 60 x 3
   name             value percent
   <chr>            <int>   <dbl>
 1 From Date       114635   100
 2 To Date         114635   100
 3 file_name       114635   100
 4 CO (mg/m3)       48443    42.3
 5 Benzene (ug/m3)  48136    42.0
 6 NOx (ppb)        48007    41.9
 7 Ozone (ppb)      47576    41.5
 8 Toluene (ug/m3)  46987    41.0
 9 PM2.5 (ug/m3)    46610    40.7
10 Xylene (ug/m3)   46564    40.6
# i 50 more rows
```

Select Desired Parameters, Rename Parameters, Filter Start Date

```r
rare_parameters <- observation_counts %>%
  filter(percent < 2) %>%
  pull(name)

df <- delhi_city %>%
  select(-all_of(rare_parameters), -file_name, -`From Date`, -`Ozone (ppb)`) %>%
  rename(
    "datetime" = "To Date",
    "PM2.5" = "PM2.5 (ug/m3)",
    "PM10" = "PM10 (ug/m3)",
    "NO" = "NO (ug/m3)",
    "NO2" = "NO2 (ug/m3)",
    "NOx" = "NOx (ppb)",
    "CO" = "CO (mg/m3)",
    "Benzene" = "Benzene (ug/m3)",
    "Toluene" = "Toluene (ug/m3)",
    "Xylene" = "Xylene (ug/m3)"
  ) %>%
  as_tsibble(index = datetime) %>%
  filter(datetime >= as_datetime("2017-08-31 01:00:00"))
```

Show Hourly Data Availability

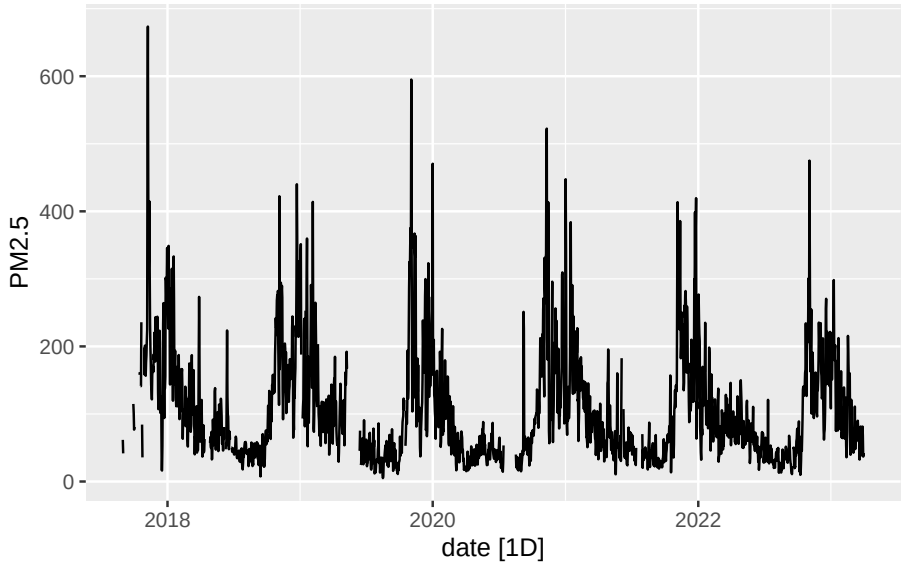
```r
df %>%
  as_tibble() %>%
  summarise(across(everything(), ~sum(!is.na(.)))) %>%
  pivot_longer(cols = everything()) %>%
  arrange(desc(value)) %>%
  mutate(percent = (value / nrow(df)) * 100)
```

```
# A tibble: 10 x 3
   name     value percent
   <chr>    <int>   <dbl>
 1 datetime 48936   100
 2 CO       44884    91.7
 3 Toluene  44635    91.2
 4 Benzene  44626    91.2
 5 NOx      44583    91.1
 6 Xylene   44212    90.3
 7 PM2.5    43227    88.3
 8 PM10     42937    87.7
 9 NO2      42161    86.2
10 NO       41407    84.6
```

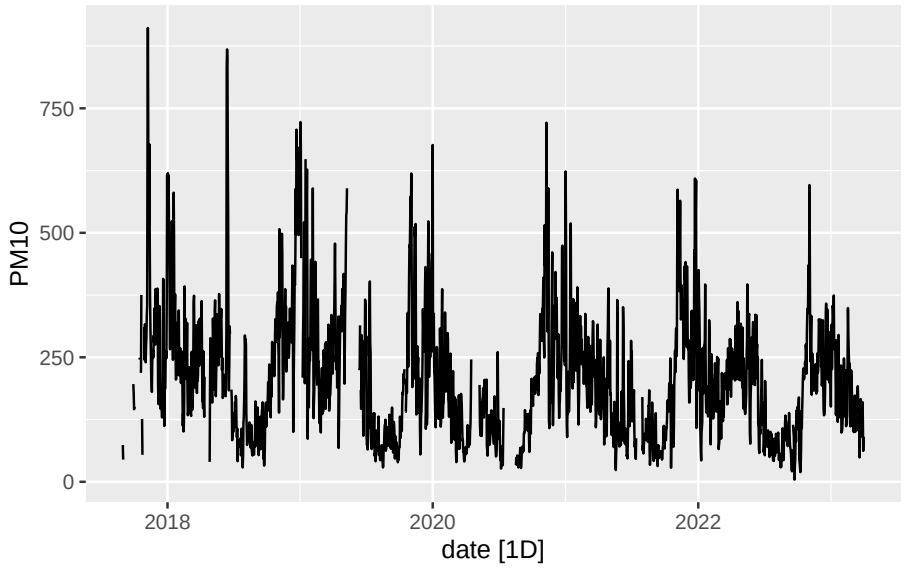
Aggregate to Daily Series

```
daily_df <- df %>%
  index_by(date = as.Date(datetime)) %>%
  summarise(
    PM2.5 = mean(PM2.5, na.rm = TRUE),
    PM10 = mean(PM10, na.rm = TRUE),
    NO = mean(NO, na.rm = TRUE),
    NO2 = mean(NO2, na.rm = TRUE),
    NOx = mean(NOx, na.rm = TRUE),
    CO = mean(CO, na.rm = TRUE),
    Benzene = mean(Benzene, na.rm = TRUE),
    Toluene = mean(Toluene, na.rm = TRUE),
    Xylene = mean(Xylene, na.rm = TRUE)
  )
```

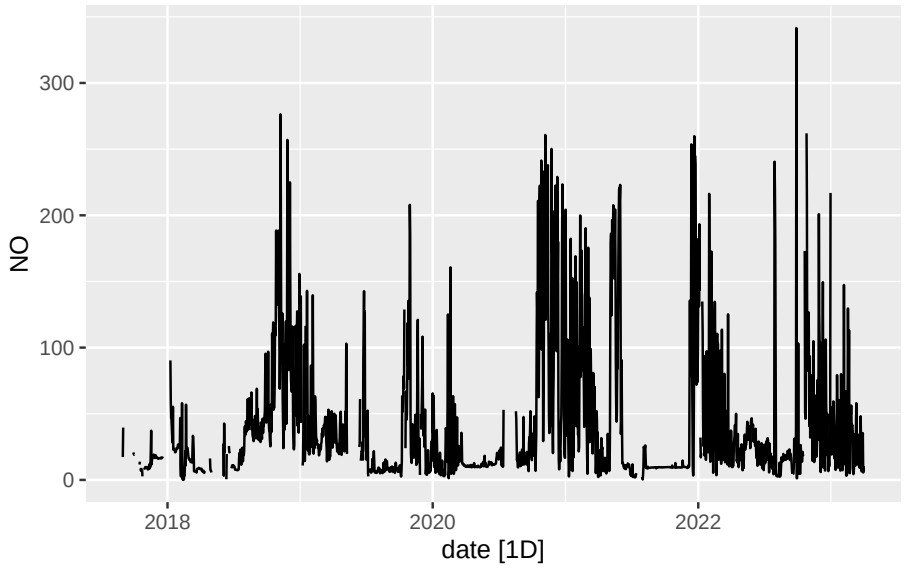
```
daily_df %>%
  autoplot(PM2.5)
```



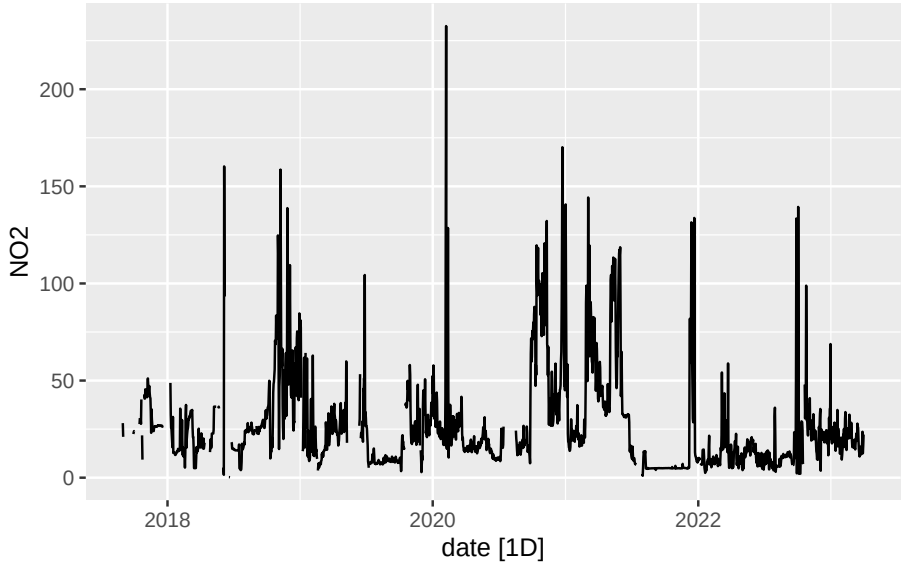

```
daily_df %>%
  autoplot(PM10)
```



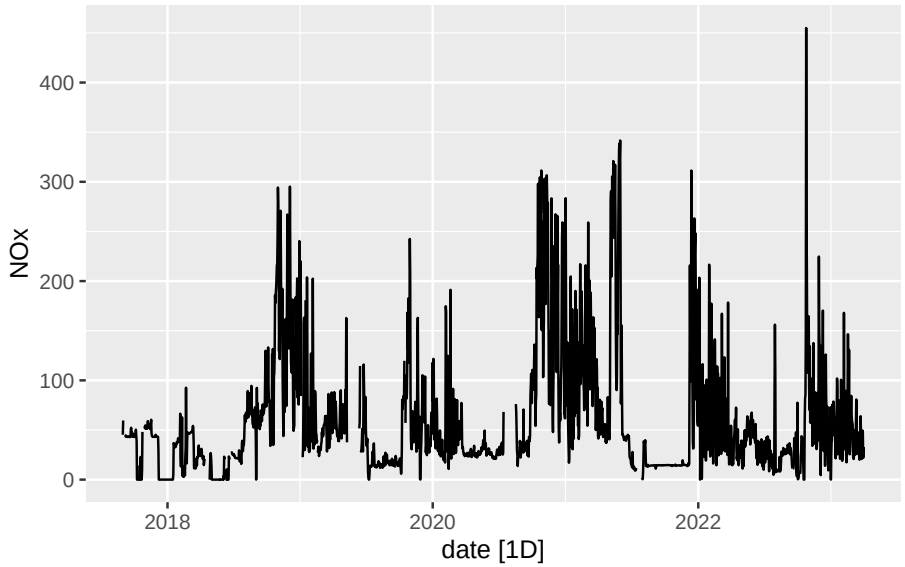

```
daily_df %>%
  autoplot(NO)
```



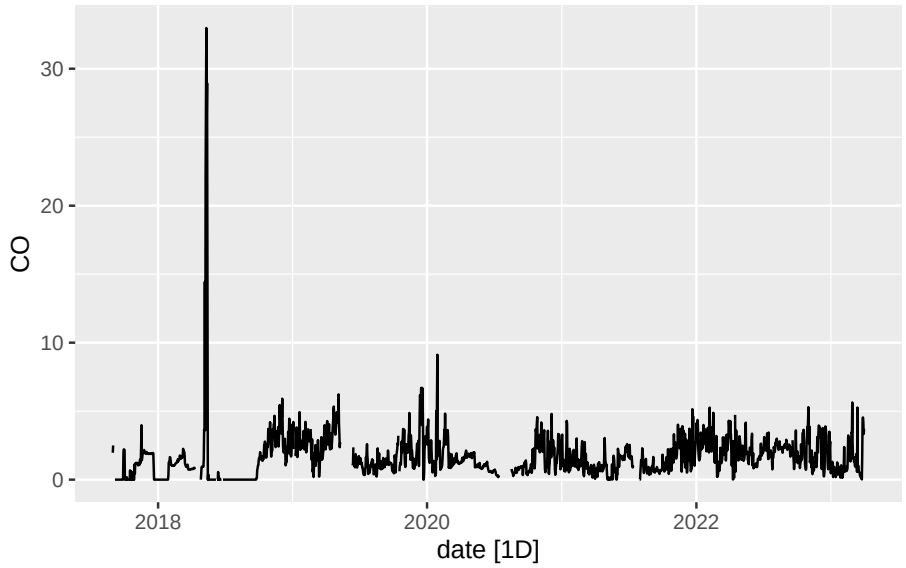

```
daily_df %>%
  autoplot(NO2)
```



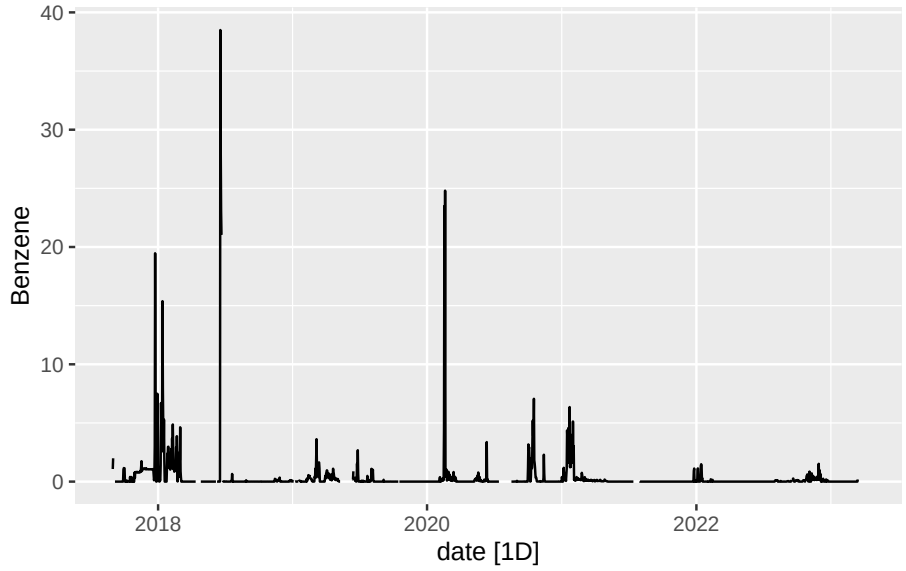

```
daily_df %>%
  autoplot(NOx)
```



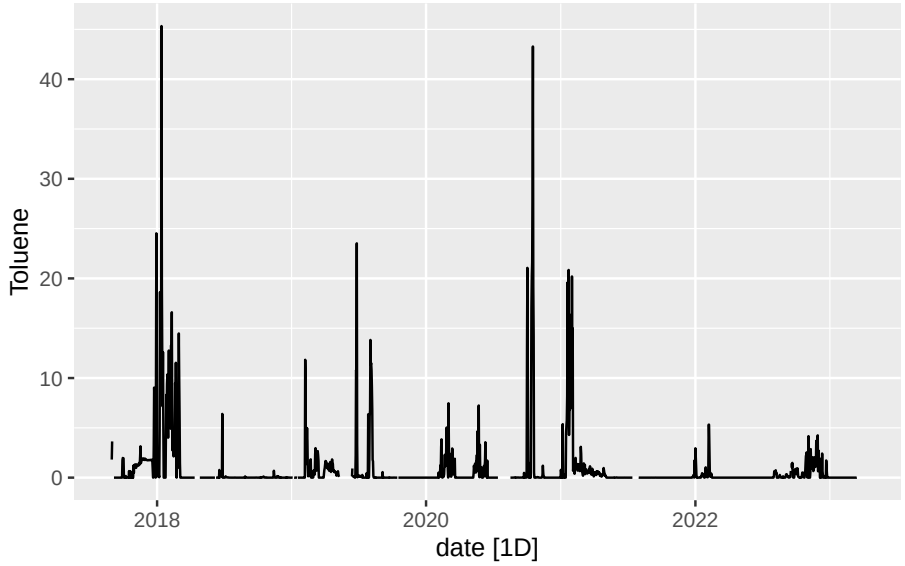

```
daily_df %>%
  autoplot(CO)
```



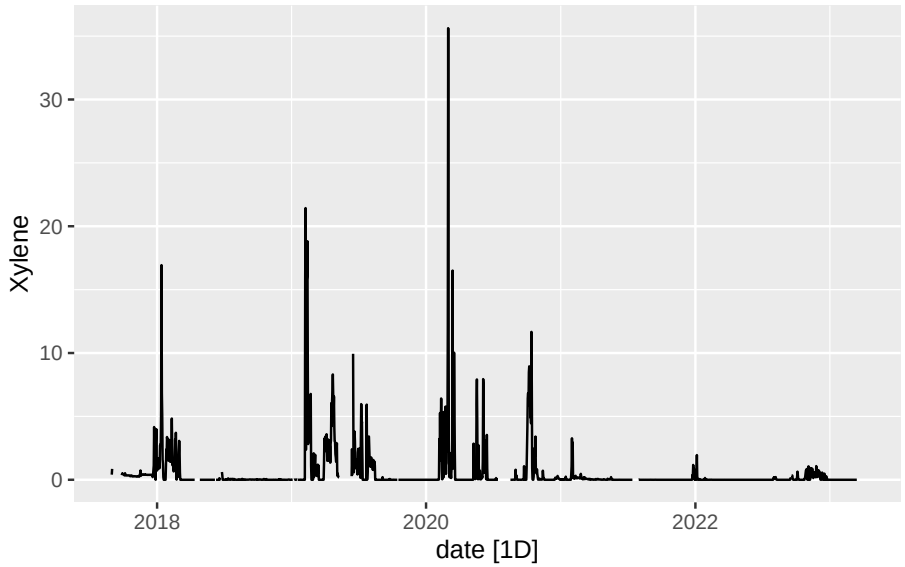

```
daily_df %>%
  autoplot(Benzene)
```




```
daily_df %>%
  autoplot(Toluene)
```

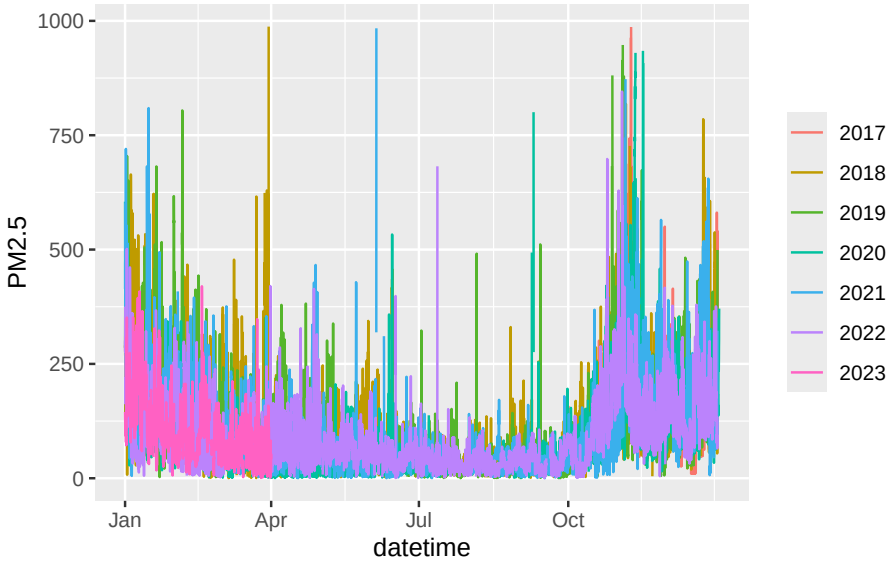



```
daily_df %>%
  autoplot(Xylene)
```



Yearly Seasonality for Particulates

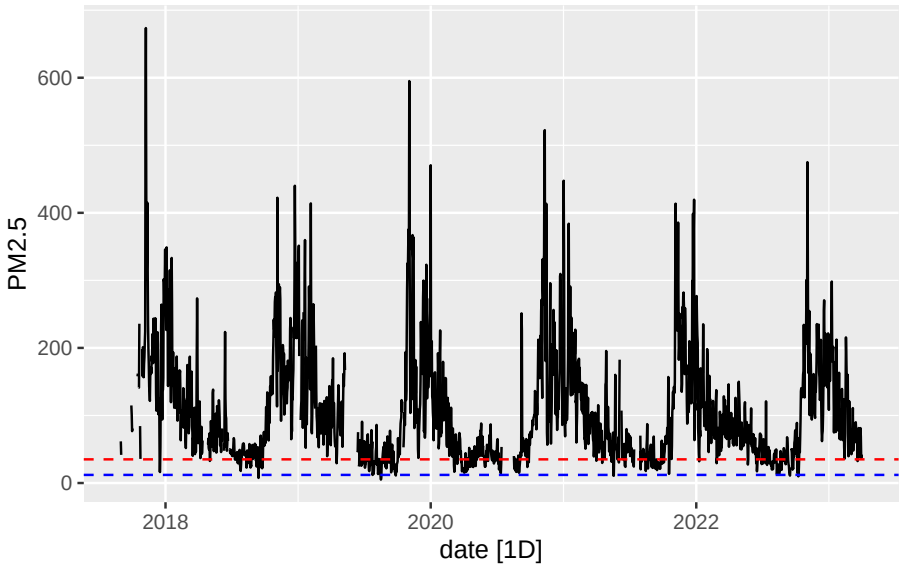
```
df %>%
  gg_season(PM2.5, period = "year")
```



Winter months have more fine particulates.

Particulates With World Health Organization Recomendation Limit

```
df %>%
  index_by(date = as.Date(datetime)) %>%
  summarise(PM2.5 = mean(PM2.5, na.rm = TRUE)) %>%
  autoplot(PM2.5) +
  geom_hline(yintercept = 35, color = "red", linetype = "dashed") +
  geom_hline(yintercept = 12, color = "blue", linetype = "dashed")
```



Fine particulate ammounts are significantly above recomended levels