Project Component 1

Himanshu, MDS202327

2023-09-18

Introduction

TBA

Data Description

The data contains six air pollution parameters like PM2.5, PM10, NO2, NH3, SO2, Ozone for 15 stations in New Delhi, collected from CPCB website from 08-02-2018 to 02-01-2021 on daily basis.

```
# Importing libraries
library(tidyverse)
library(dplyr)
library(ggplot2)
library(TSstudio)
library(plotly)
```

```
# Reading the data into data frame
df <- read.csv("delhi.csv", header = TRUE)
set.seed(5)
df[sample(nrow(df), 5), ]
```

```
##          Id        siteName siteCode       Date  PM2.5   PM10   NO2    NH3
## 13122 13121     Sonia Vihar     1432 2019-03-16  76.24 128.73 26.19  26.79
## 12139 12138          Rohini     1430 2019-06-01 109.17 308.58 33.36 107.17
## 10937 10936      Patparganj     1431 2019-07-09  48.39  90.19 10.98  48.72
## 2255   2254 Dwarka-Sector 8     1422 2018-06-22  59.46 269.33 26.90   4.00
## 6859   6858       Najafgarh     1427 2019-08-30  33.23  67.69 26.37  17.75
##         SO2 Ozone
## 13122 10.62 49.66
## 12139 10.02 48.77
## 10937  2.71 15.49
## 2255   5.28  3.83
## 6859   8.56 54.52
```

```
# Variables in the data
names(df)
```

```
##  [1] "Id"       "siteName" "siteCode" "Date"     "PM2.5"    "PM10"
##  [7] "NO2"      "NH3"      "SO2"      "Ozone"
```

```r
# Dimension of the data
dim(df)
```

```
## [1] 15900    10
```

```r
# Variable types
# Note the Date column has type chr which must be converted to date type.
str(df)
```

```
## 'data.frame':    15900 obs. of  10 variables:
##  $ Id      : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ siteName: chr  "Ashok Vihar" "Ashok Vihar" "Ashok Vihar" "Ashok Vihar" ...
##  $ siteCode: int  1420 1420 1420 1420 1420 1420 1420 1420 1420 1420 ...
##  $ Date    : chr  "2018-02-08" "2018-02-09" "2018-02-10" "2018-02-11" ...
##  $ PM2.5   : num  237 250.5 269.7 146.4 82.1 ...
##  $ PM10    : num  406 423 499 315 200 ...
##  $ NO2     : num  110 79.4 183.9 41.8 23.2 ...
##  $ NH3     : num  31.4 33.5 22.7 36.7 34.8 ...
##  $ SO2     : num  11.2 13.24 7.16 8.38 4.43 ...
##  $ Ozone   : num  33.4 39.3 44.5 43 37.9 ...
```

```r
df$Date <- as.Date(df$Date)
df[sample(nrow(df), 5), ]
```

```
##          Id                      siteName siteCode       Date  PM2.5   PM10
## 13177 13176                   Sonia Vihar     1432 2019-05-10  87.55 319.71
## 1833   1832 Dr. Karni Singh Shooting Range     1421 2020-03-21  51.39 122.20
## 3797   3796                  Jahangirpuri     1423 2019-10-17 148.79 288.54
## 13534 13533                   Sonia Vihar     1432 2020-05-01  60.58 153.08
## 7239   7238                     Najafgarh     1427 2020-09-13 116.02 167.22
##             NO2   NH3       SO2  Ozone
## 13177 29.910000 27.54 12.990000  56.83
## 1833  45.230000 28.76 17.140000  76.75
## 3797  80.240000 56.53 26.090000  88.46
## 13534  3.375843 30.27  3.270609 111.14
## 7239   9.050000 14.45  9.240000  16.02
```

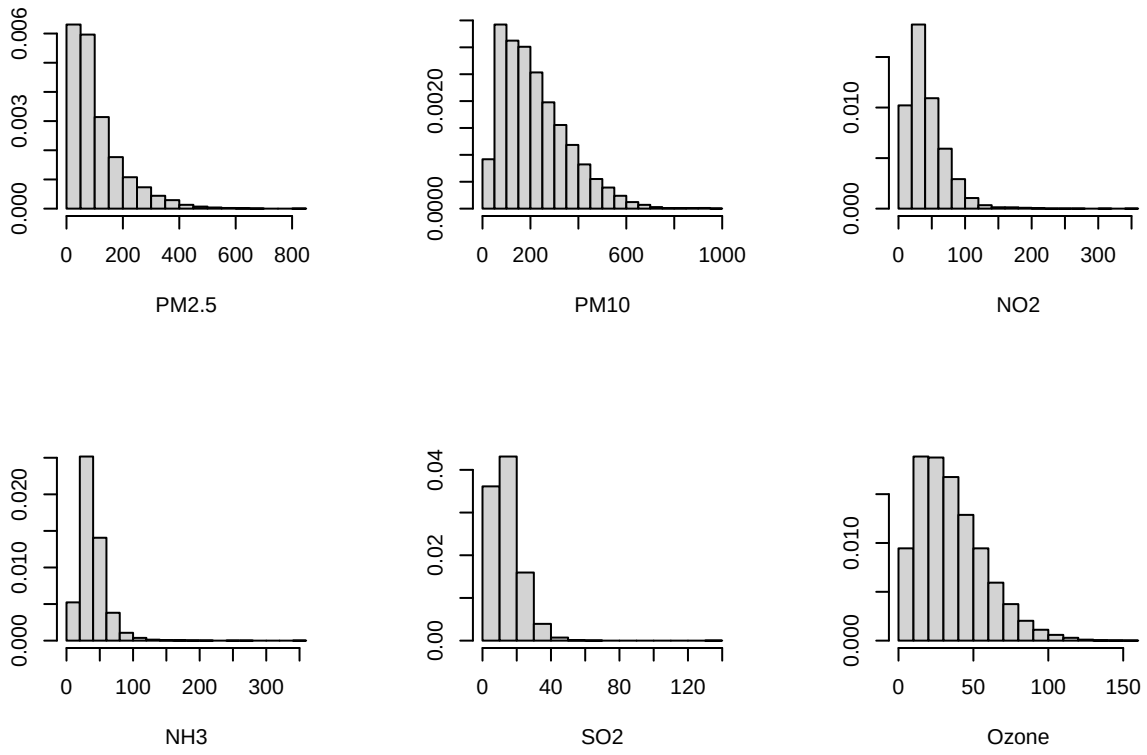
```r
str(df)
```

```
## 'data.frame':    15900 obs. of  10 variables:
##  $ Id      : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ siteName: chr  "Ashok Vihar" "Ashok Vihar" "Ashok Vihar" "Ashok Vihar" ...
##  $ siteCode: int  1420 1420 1420 1420 1420 1420 1420 1420 1420 1420 ...
##  $ Date    : Date, format: "2018-02-08" "2018-02-09" ...
##  $ PM2.5   : num  237 250.5 269.7 146.4 82.1 ...
##  $ PM10    : num  406 423 499 315 200 ...
##  $ NO2     : num  110 79.4 183.9 41.8 23.2 ...
##  $ NH3     : num  31.4 33.5 22.7 36.7 34.8 ...
##  $ SO2     : num  11.2 13.24 7.16 8.38 4.43 ...
##  $ Ozone   : num  33.4 39.3 44.5 43 37.9 ...
```

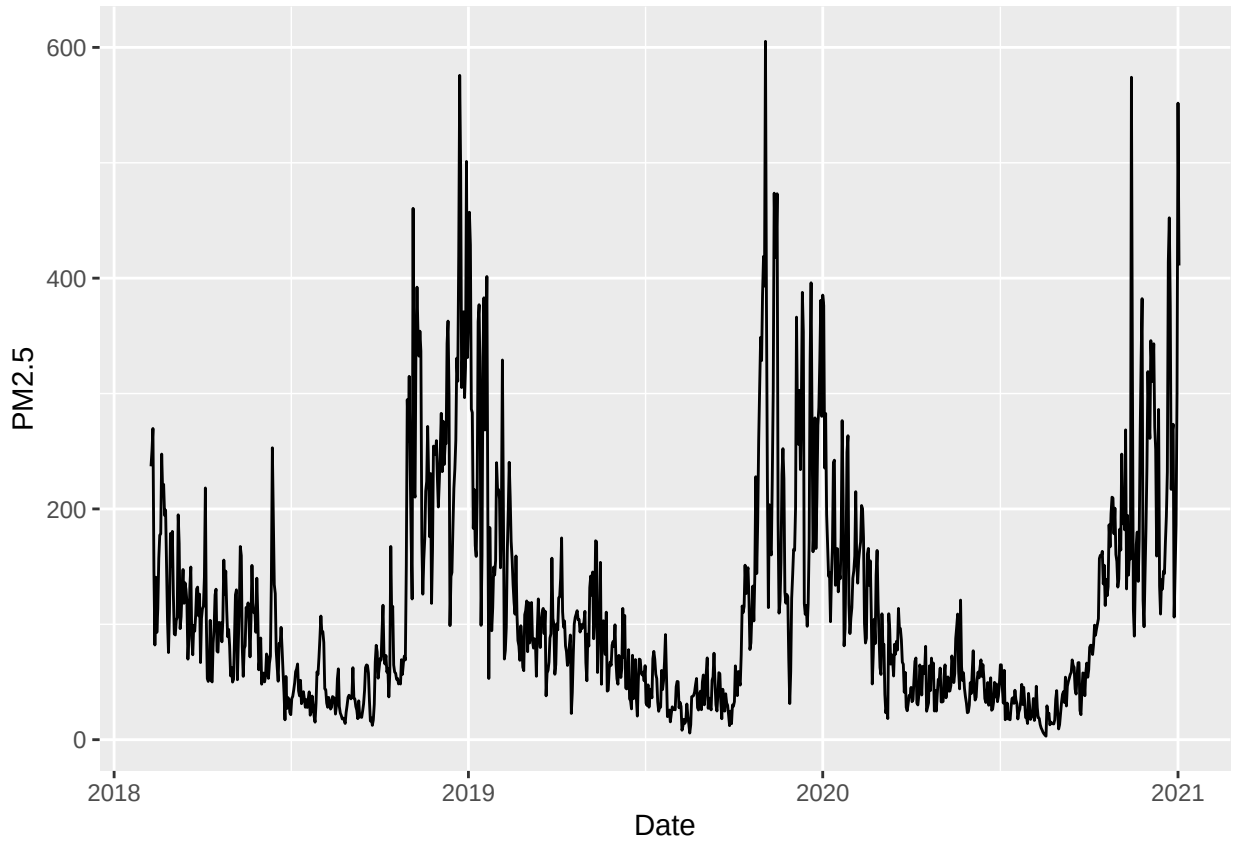
```
unique(df$siteName)
```

```
##  [1] "Ashok Vihar"                       "Dr. Karni Singh Shooting Range"
##  [3] "Dwarka-Sector 8"                   "Jahangirpuri"
##  [5] "Jawaharlal Nehru Stadium"          "Major Dhyan Chand National Stadium"
##  [7] "Najafgarh"                         "Narela"
##  [9] "Nehru Nagar"                       "Okhla Phase-2"
## [11] "Patparganj"                        "Rohini"
## [13] "Sonia Vihar"                       "Vivek Vihar"
## [15] "Wazirpur"
```

```
par(mfrow = c(2,3))
hist(df$PM2.5, probability = TRUE, main = "", xlab = "PM2.5", ylab = "")
hist(df$PM10, probability = TRUE, main = "", xlab = "PM10", ylab = "")
hist(df$NO2, probability = TRUE, main = "", xlab = "NO2", ylab = "")
hist(df$NH3, probability = TRUE, main = "", xlab = "NH3", ylab = "")
hist(df$SO2, probability = TRUE, main = "", xlab = "SO2", ylab = "")
hist(df$Ozone, probability = TRUE, main = "", xlab = "Ozone", ylab = "")
```




```
par(mfrow=c(1,2))
p <- ggplot(df[df$siteCode==1420,], aes(x=Date, y=PM2.5)) +
       geom_line()
p + scale_x_date(date_labels = "%Y")
```




```
#qplot(Date, PM2.5, data = df, geom = "line")
```

```
#ts_plot(subset(df,siteCode==1420,select=c('Date','PM2.5')), title = "",
#        Xtitle = "Time", Ytitle = "PM2.5 Levels", color = "black", width=1.5, slider = TRUE)
```

Exploratory Data Analysis

Results

Conclusion