Project Component 1

Himanshu, MDS202327

2023-09-19

Introduction

TBA

Data Description

The data contains six air pollution parameters like PM2.5, PM10, NO2, NH3, SO2, Ozone for 12 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), ]
```

```
##              siteName siteCode       Date  PM2.5   PM10   NO2    NH3   SO2
## 13122     Sonia Vihar     1432 2019-03-16  76.24 128.73 26.19  26.79 10.62
## 12139          Rohini     1430 2019-06-01 109.17 308.58 33.36 107.17 10.02
## 10937      Patparganj     1431 2019-01-10 171.10 249.96 34.45  71.82  5.30
## 2255  Dwarka-Sector 8     1422 2018-06-22  59.46 269.33 26.90   4.00  5.28
## 6859        Najafgarh     1427 2019-06-21  71.17 145.08 39.99  36.66 20.66
##        Ozone
## 13122  49.66
## 12139  48.77
## 10937  18.40
## 2255    3.83
## 6859  103.91
```

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

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

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

```
## [1] 15900     9
```

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

```
## 'data.frame':    15900 obs. of  9 variables:
##  $ 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), ]
```

```
##                            siteName siteCode       Date  PM2.5   PM10
## 13177                    Sonia Vihar     1432 2019-05-10  87.55 319.71
## 1833  Dr. Karni Singh Shooting Range     1421 2020-03-21  51.39 122.20
## 3797                     Jahangirpuri     1423 2019-10-17 148.79 288.54
## 13534                    Sonia Vihar     1432 2020-05-01  60.58 153.08
## 7239                       Najafgarh     1427 2020-07-05  28.64  50.76
##             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   9.210000 31.50  9.430000  44.39
```

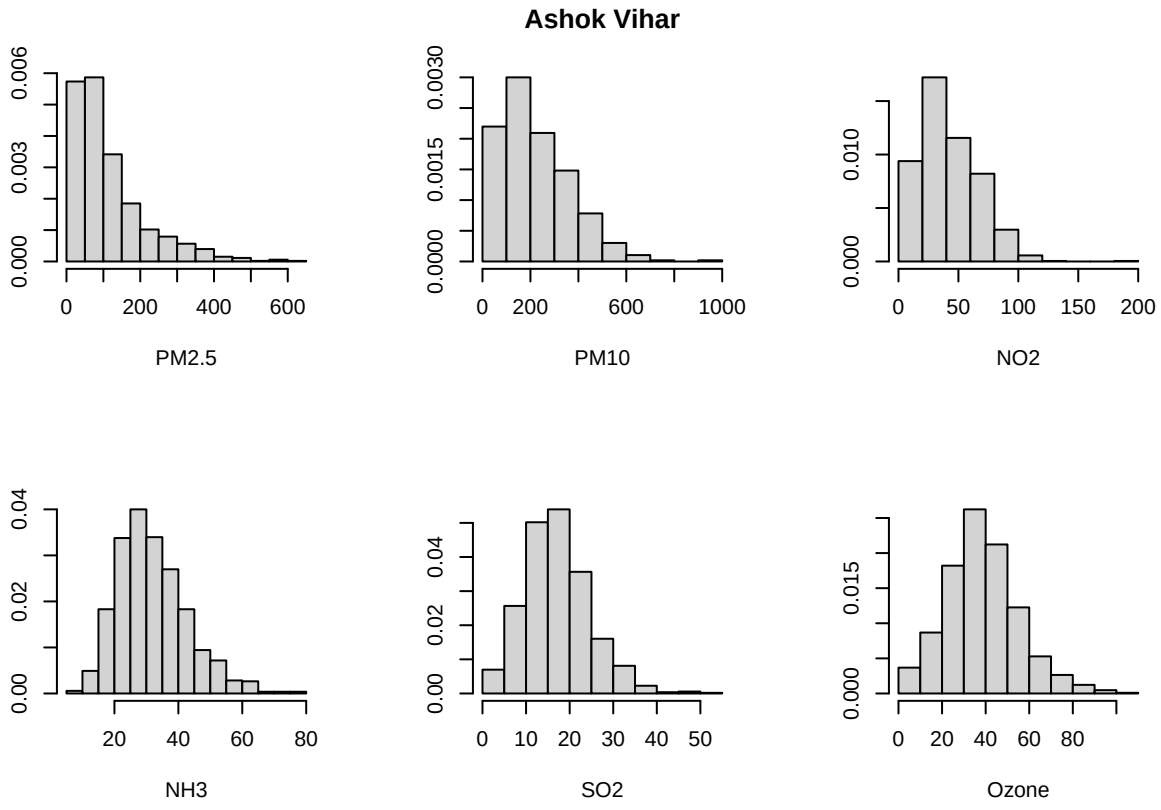
```r
str(df)
```

```
## 'data.frame':    15900 obs. of  9 variables:
##  $ 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[df$siteCode==1420,]$PM2.5, probability = TRUE, main = "", xlab = "PM2.5", ylab = "")
hist(df[df$siteCode==1420,]$PM10, probability = TRUE, main = "Ashok Vihar", xlab = "PM10", ylab = "")
hist(df[df$siteCode==1420,]$NO2, probability = TRUE, main = "", xlab = "NO2", ylab = "")
hist(df[df$siteCode==1420,]$NH3, probability = TRUE, main = "", xlab = "NH3", ylab = "")
hist(df[df$siteCode==1420,]$SO2, probability = TRUE, main = "", xlab = "SO2", ylab = "")
hist(df[df$siteCode==1420,]$Ozone, probability = TRUE, main = "", xlab = "Ozone", ylab = "")
```




```
par(mar=c(2,2,2,2))
par(mfrow=c(5,3))
for (i in unique(df$siteName)) {
  plot(df[df$siteName==i,]$Date, df[df$siteName==i,]$PM2.5, type = "l",
  main = i, xlab = "", ylab = "")
}
```

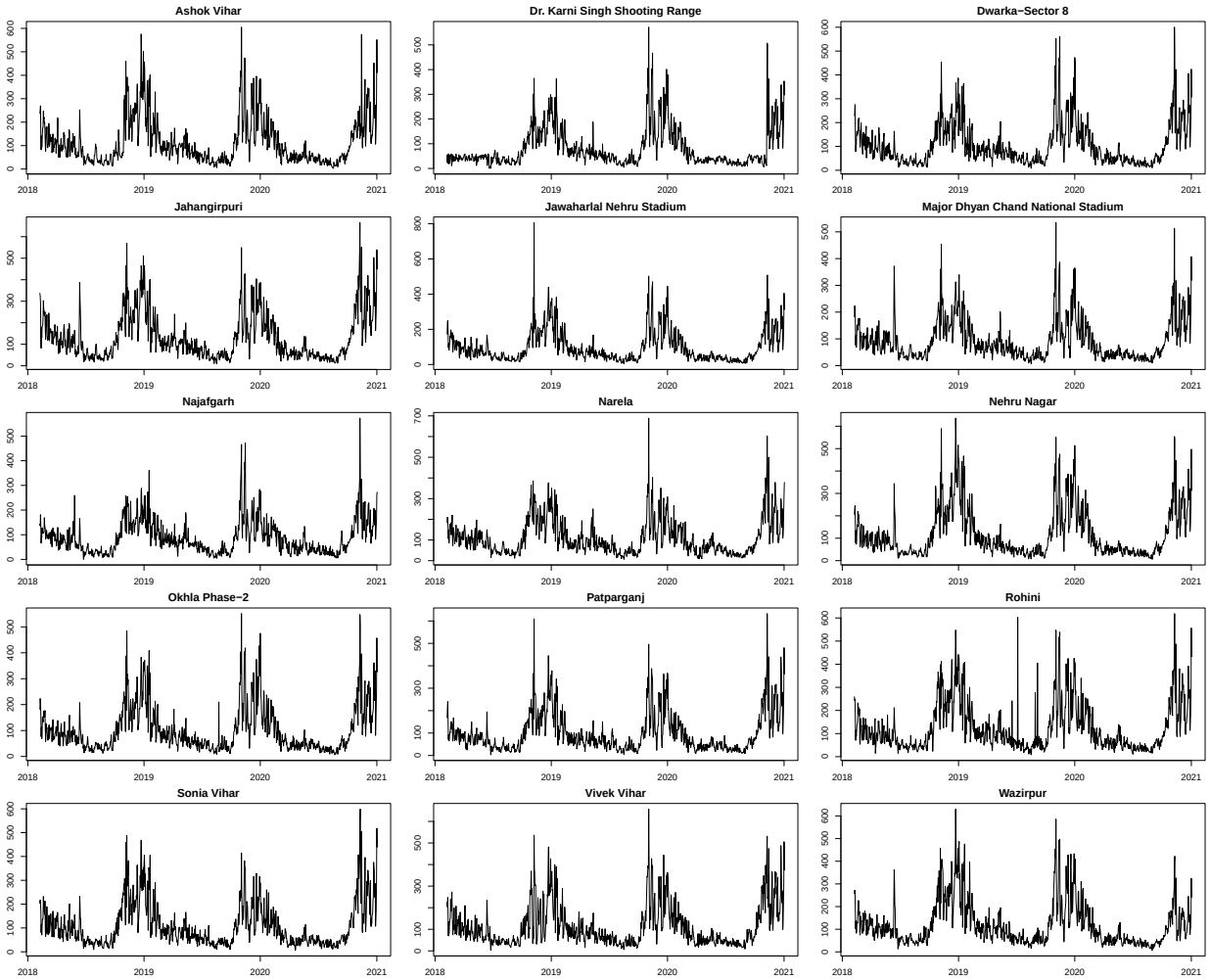
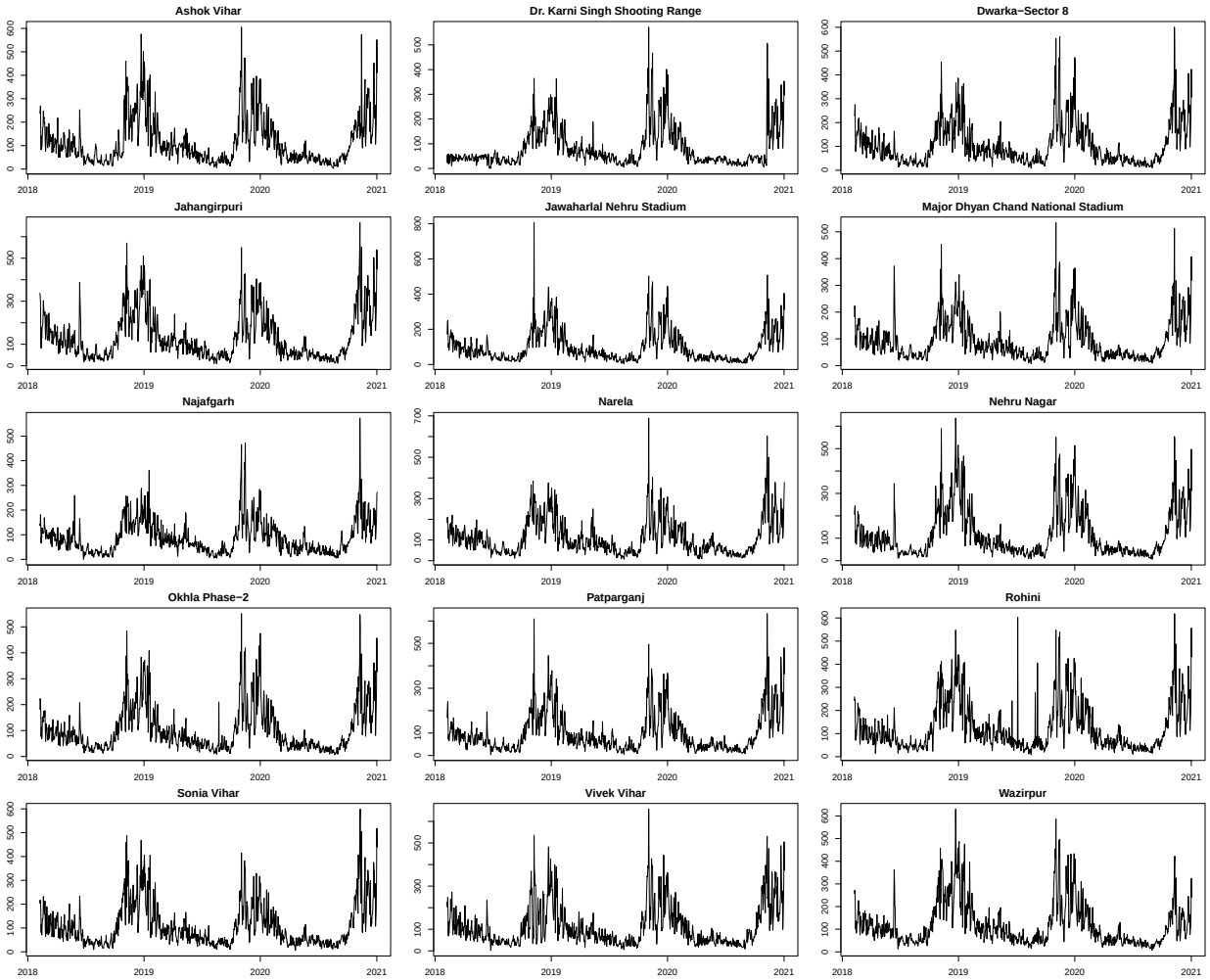
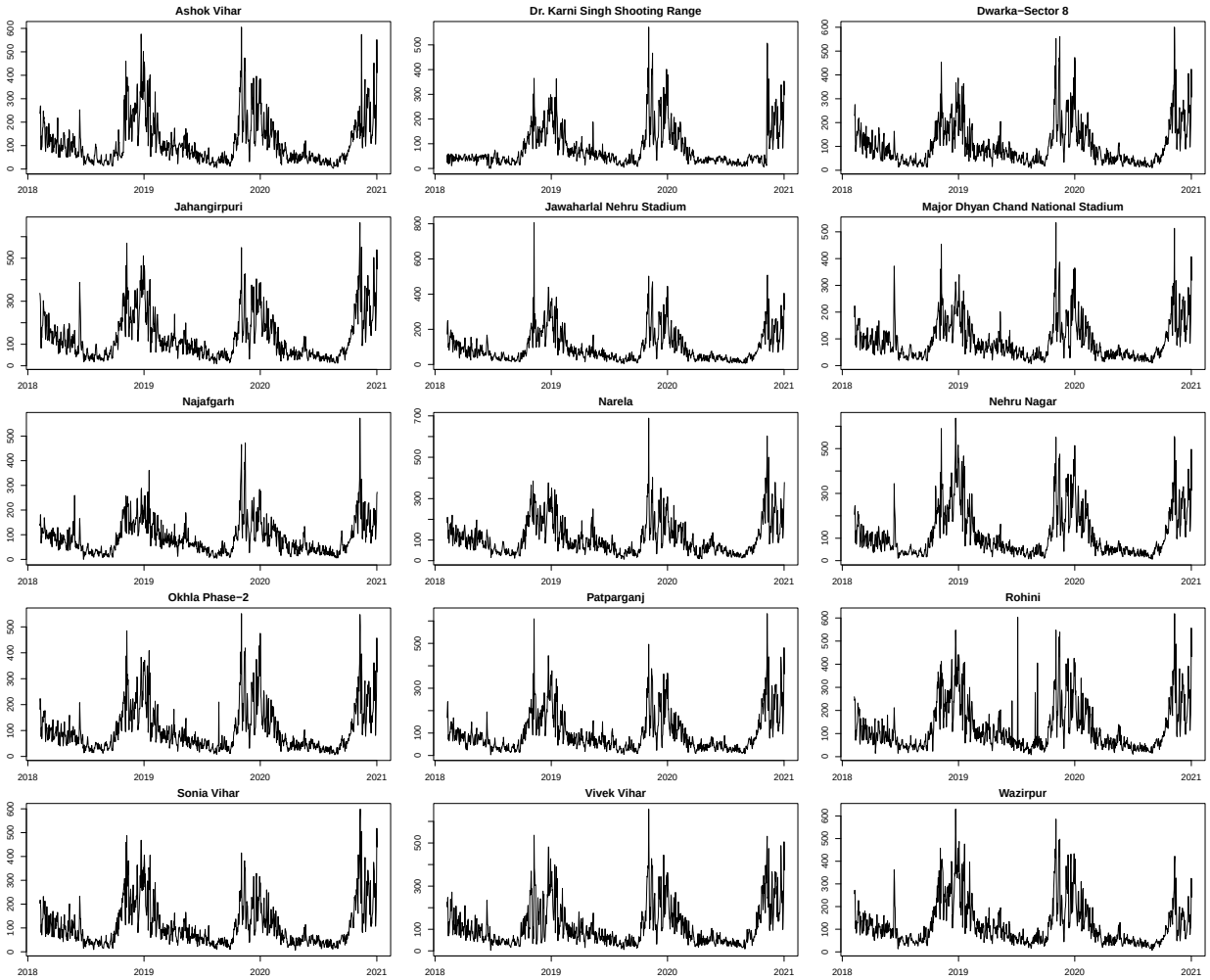


Figure: The above graph shows the time series plot of PM2.5 parameter for all 15 stations in the data.

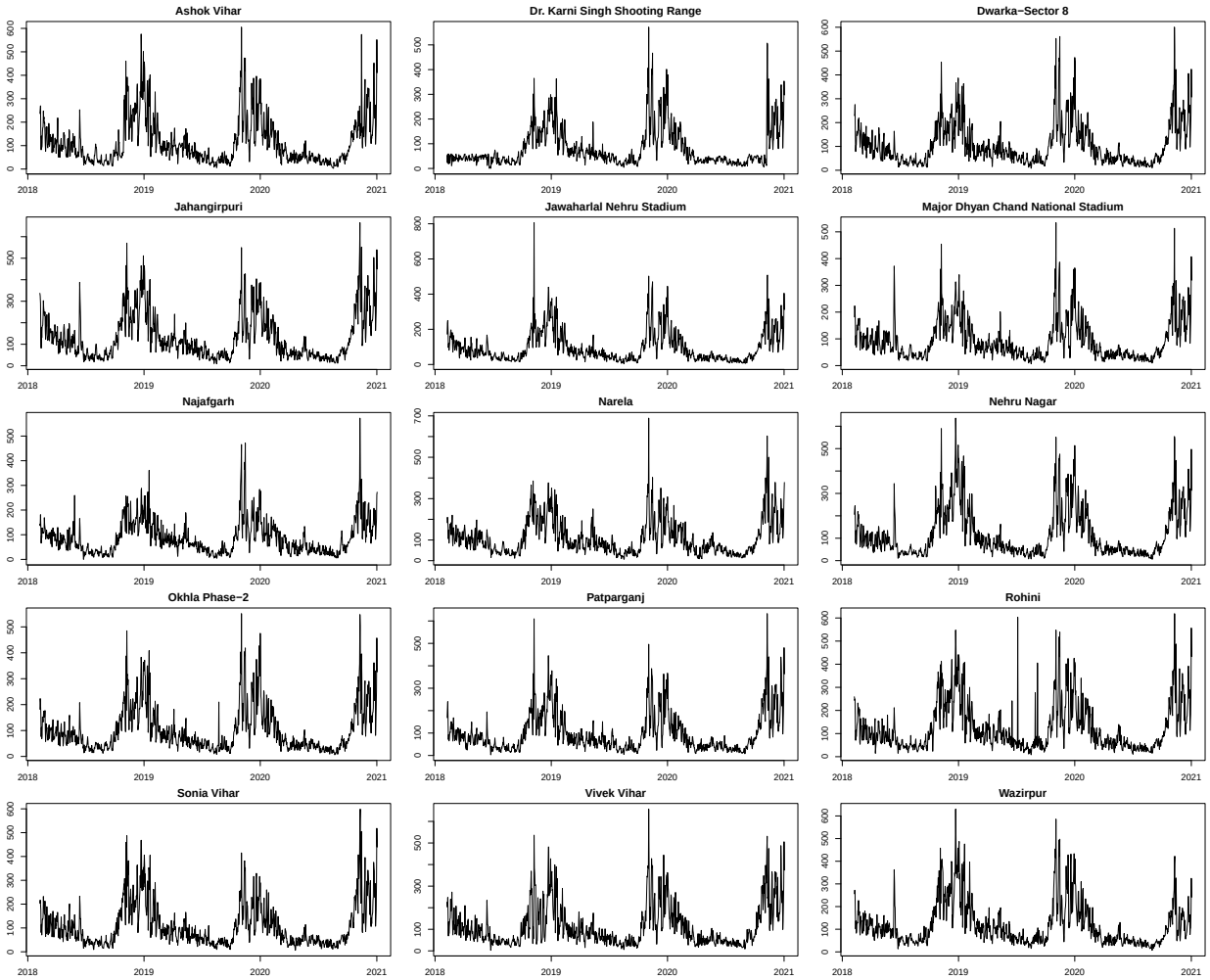
```
par(mar=c(2,2,2,2))
par(mfrow=c(5,3))
for (i in unique(df$siteName)) {
  plot(df[df$siteName==i,]$Date, df[df$siteName==i,]$PM2.5, type = "l",
  main = i, xlab = "", ylab = "PM2.5")
}
```



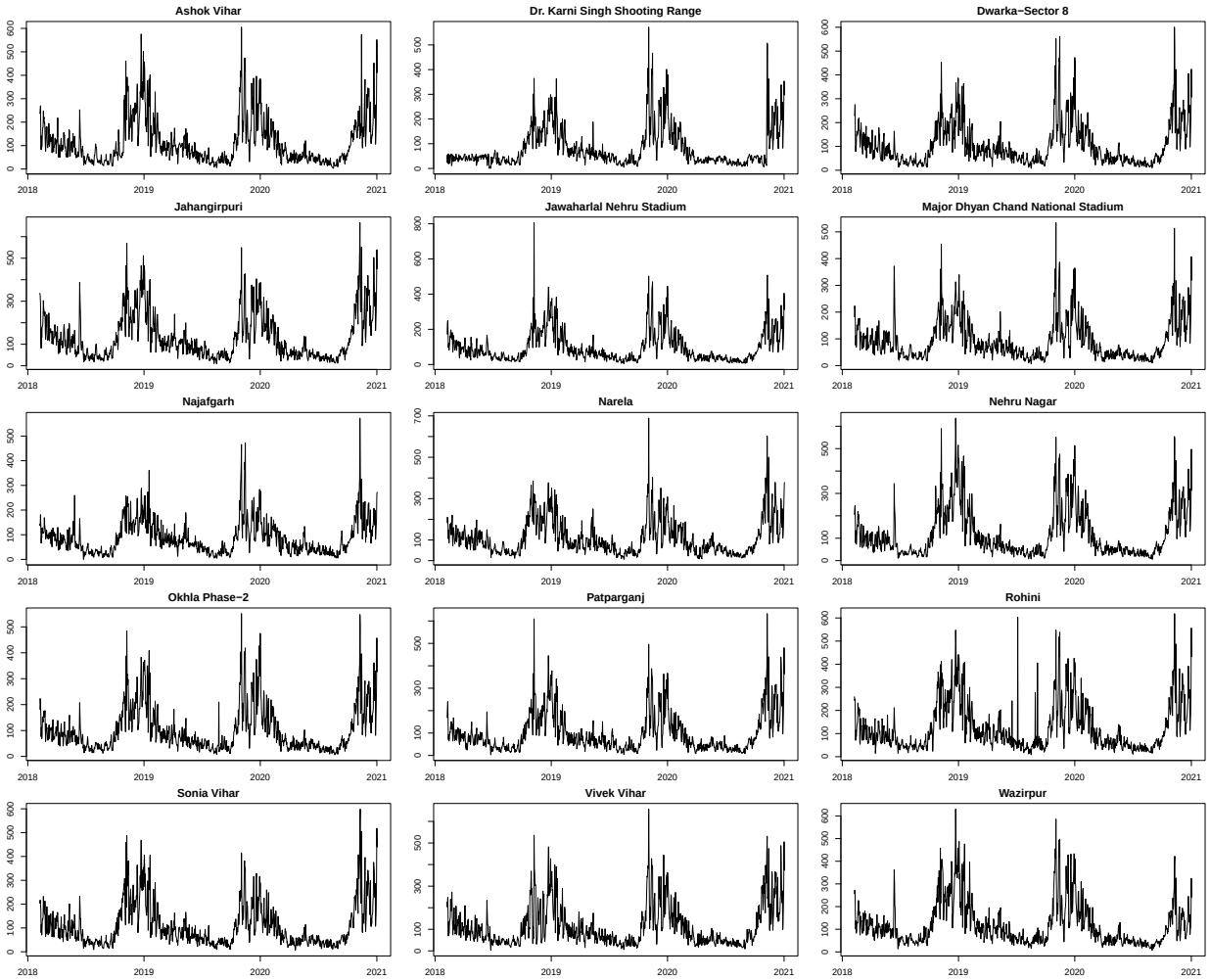

```
par(mar=c(2,2,2,2))
par(mfrow=c(5,3))
for (i in unique(df$siteName)) {
  plot(df[df$siteName==i,]$Date, df[df$siteName==i,]$PM2.5, type = "l",
  main = i, xlab = "", ylab = "PM2.5")
}
```



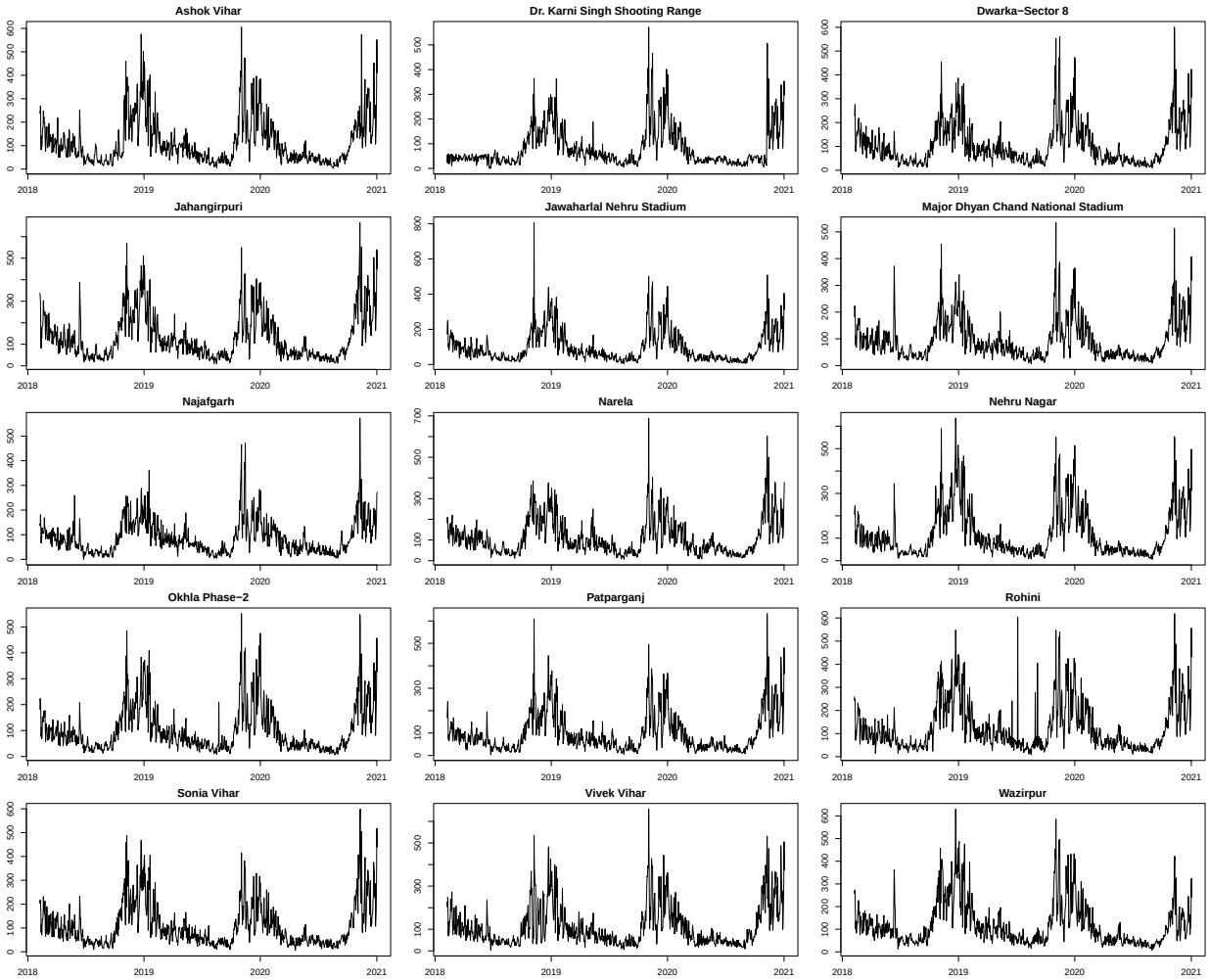

```
par(mar=c(2,2,2,2))
par(mfrow=c(5,3))
for (i in unique(df$siteName)) {
  plot(df[df$siteName==i,]$Date, df[df$siteName==i,]$PM2.5, type = "l",
  main = i, xlab = "", ylab = "PM2.5")
}
```



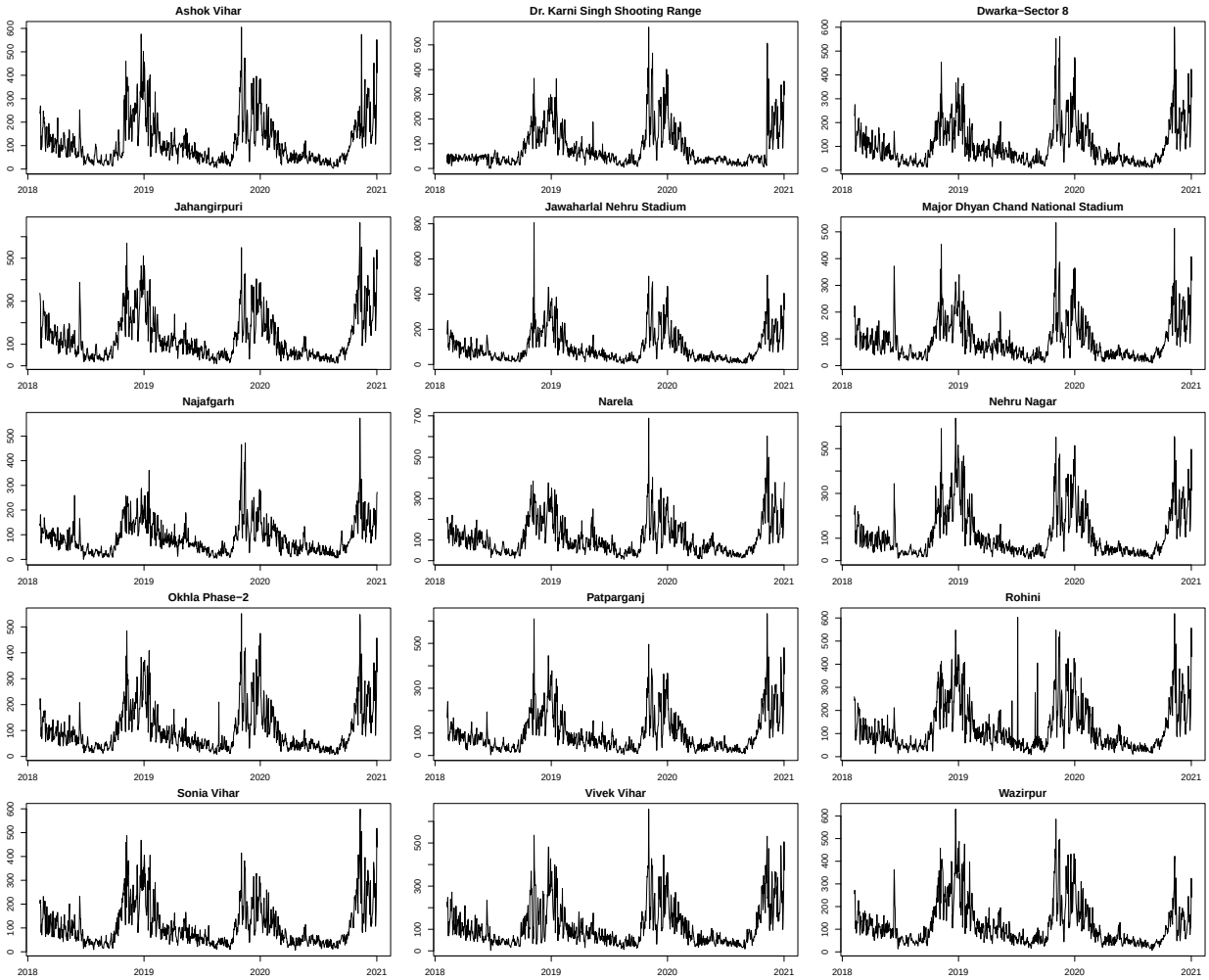

```
par(mar=c(2,2,2,2))
par(mfrow=c(5,3))
for (i in unique(df$siteName)) {
  plot(df[df$siteName==i,]$Date, df[df$siteName==i,]$PM2.5, type = "l",
  main = i, xlab = "", ylab = "PM2.5")
}
```




```
par(mar=c(2,2,2,2))
par(mfrow=c(5,3))
for (i in unique(df$siteName)) {
  plot(df[df$siteName==i,]$Date, df[df$siteName==i,]$PM2.5, type = "l",
  main = i, xlab = "", ylab = "PM2.5")
}
```




```
par(mar=c(2,2,2,2))
par(mfrow=c(5,3))
for (i in unique(df$siteName)) {
  plot(df[df$siteName==i,]$Date, df[df$siteName==i,]$PM2.5, type = "l",
  main = i, xlab = "", ylab = "PM2.5")
}
```



Exploratory Data Analysis

Results

Conclusion