Project Component 1

Himanshu, MDS202327

2023-09-29

Introduction

In this report we try to discover patterns and make inferences about the pollution level in stations in and around Delhi. We also look at the relation between the pollutant parameters and try to find which parameter are related to each other and see if this is the case across all ten stations.

Data Description

The data contains 10600 rows and 9 columns, namely, siteName, siteCode, Date and six air pollution parameters i.e. PM2.5, PM10, NO_2, NH_3, SO_2, Ozone for ten stations in New Delhi, collected from CPCB website from 08-02-2018 to 02-01-2021 on daily basis. There are 1060 entries for each station, one for all the dates between 08-02-2018 and 02-01-2021 (both inclusive). The data for the parameters is average of 24 hour data collected every 15 minutes. The units for all the parameters in the data are $\frac{ug}{m^3}$ that represents micrograms(one-millionth of a gram) of a gaseous pollutant per cubic meter of air.

siteName	**siteCode**	**Date**	**PM2.5**	**PM10**	**NO_2**	**NH_3**	**SO_2**	**Ozone**
<chr>	<int>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Sonia Vihar	1432	2019-09-19	17.62	65.71	13.18	26.37	12.64	36.09
Jahangirpuri	1423	2020-03-01	51.20	120.17	72.40	36.34	2.04	12.23
Wazirpur	1434	2020-04-12	44.46	85.50	32.24	23.36	14.07	52.15
Najafgarh	1427	2018-05-19	100.06	287.78	28.60	46.65	7.63	73.52
Patparganj	1431	2018-10-27	189.89	384.89	63.65	85.26	4.39	18.85

Table 1: A glimpse of random sample of the data.

The names of all ten stations with their respective site codes are displayed in the table below.

Site Name	Ashok Vihar	Dwarka-Sector	Jahangirpuri	Najafgarh	Narela
Site Code	1420	1422	1423	1427	1426
Site Name	Patparganj	Rohini	Sonia Vihar	Vivek Vihar	Wazirpur
Site Code	1431	1430	1432	1435	1434

Table 2: Site Names and corresponding Site Codes

Exploratory Data Analysis

We use the following libraries for handling data and creating plots for this report.

```r
# Importing libraries
library(tidyverse)
library(dplyr)
library(ggplot2)
library(knitr)
```

Since our data has a date column, we would want to exploit it to our use to plot some time-series plots and analysis. As displayed in table 1, the data column has type chr, so we must first convert it to date type.

```r
#Changing Date data type from chr to date
df$Date <- as.Date(df$Date)
```

siteName	**siteCode**	**Date**	**PM2.5**	**PM10**	**NO_2**	**NH_3**	**SO_2**	**Ozone**
<chr>	<int>	<date>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Sonia Vihar	1432	2019-09-19	17.62	65.71	13.18	26.37	12.64	36.09
Jahangirpuri	1423	2020-03-01	51.20	120.17	72.40	36.34	2.04	12.23
Wazirpur	1434	2020-04-12	44.46	85.50	32.24	23.36	14.07	52.15
Najafgarh	1427	2018-05-19	100.06	287.78	28.60	46.65	7.63	73.52
Patparganj	1431	2018-10-27	189.89	384.89	63.65	85.26	4.39	18.85

Table 3: A glimpse of random sample of the data after chnaging type of Date column.

We now start with visualizing some graphs!

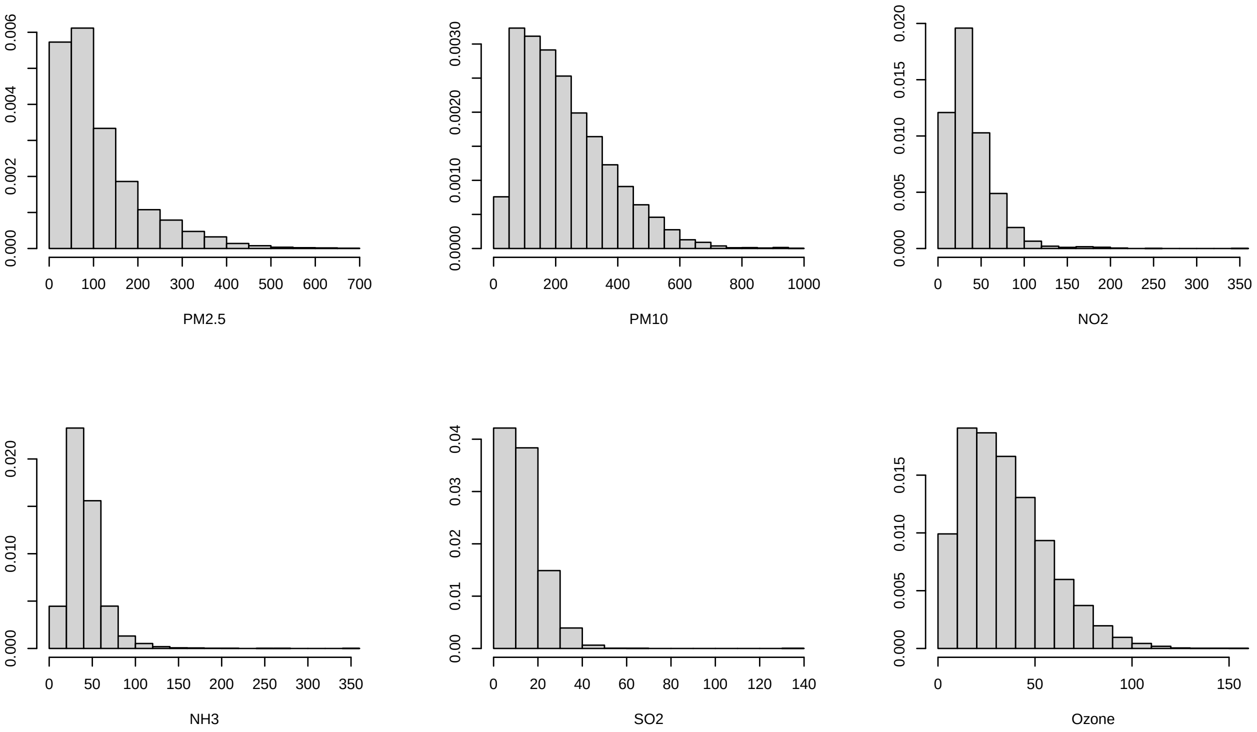

Figure: The above graph shows the distribution of values of the six pollutant parameters.

An important part of exploratory data analysis is to find out which parameters are correlated, positive or negative. This helps one decide on which parameters to include in their prediction model.

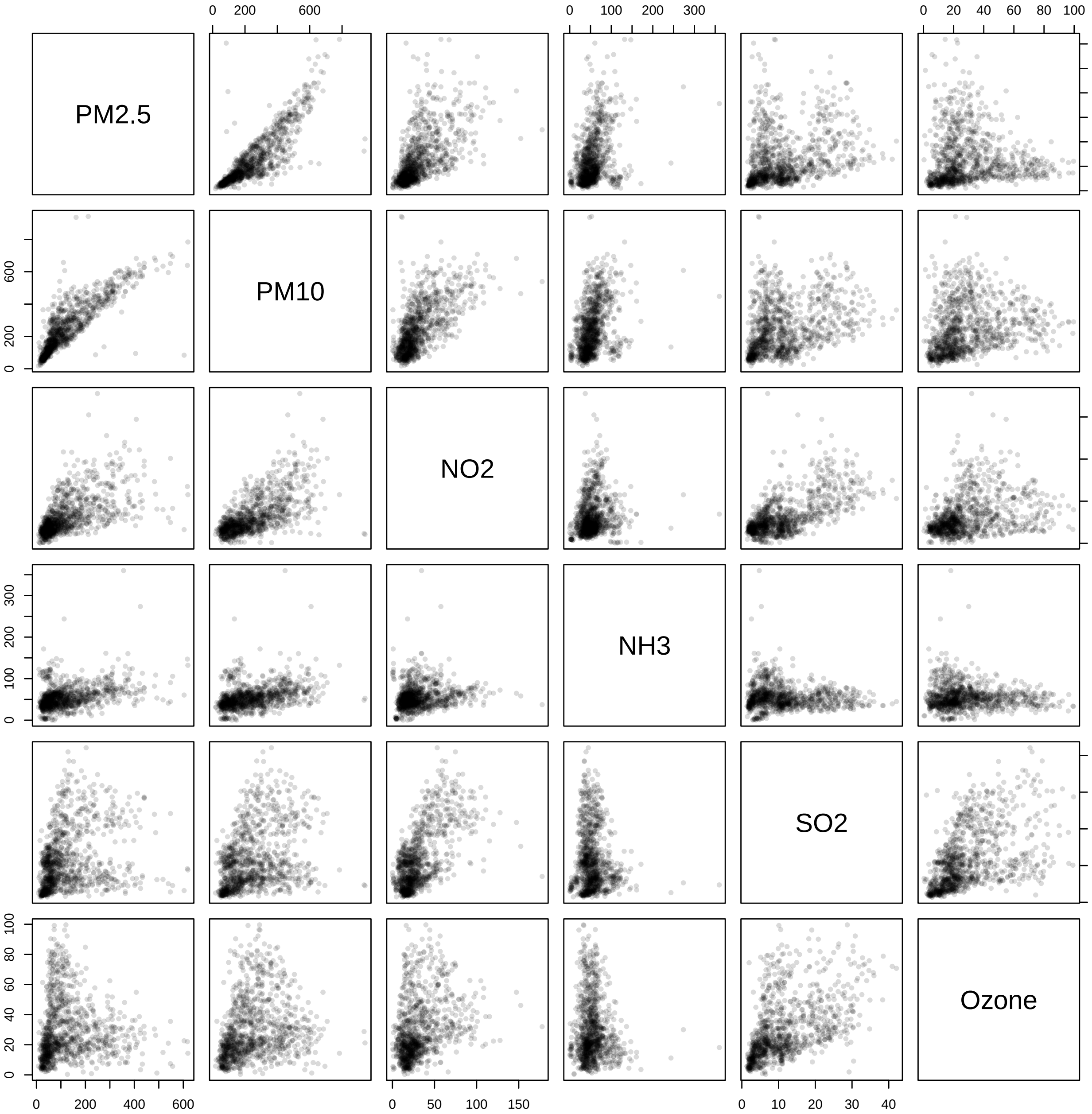

Figure: The above graph shows the correlation between all the six pollutant parameters.

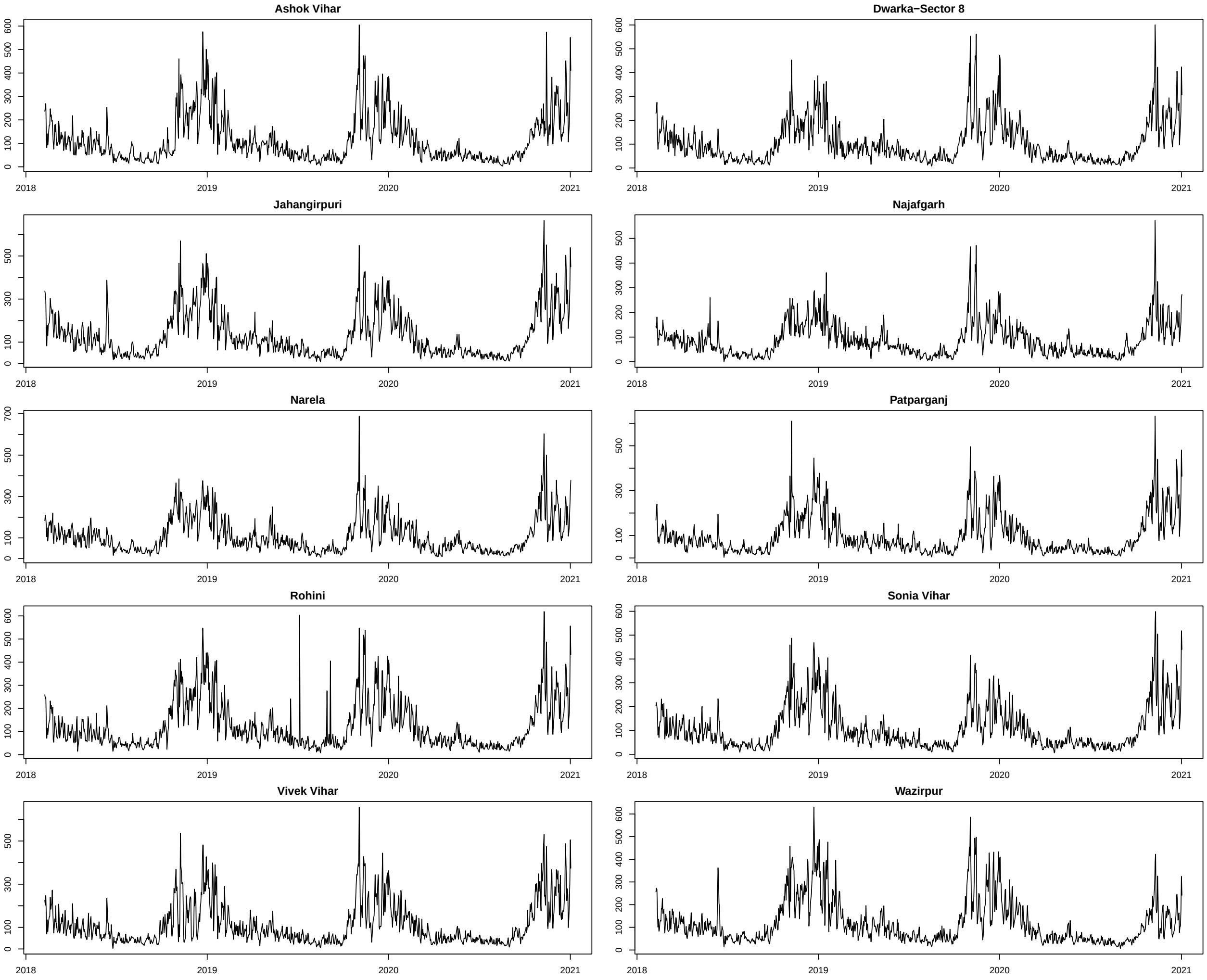

Figure: The above graph shows the time series plot of PM2.5 parameter for all 10 stations in the data.

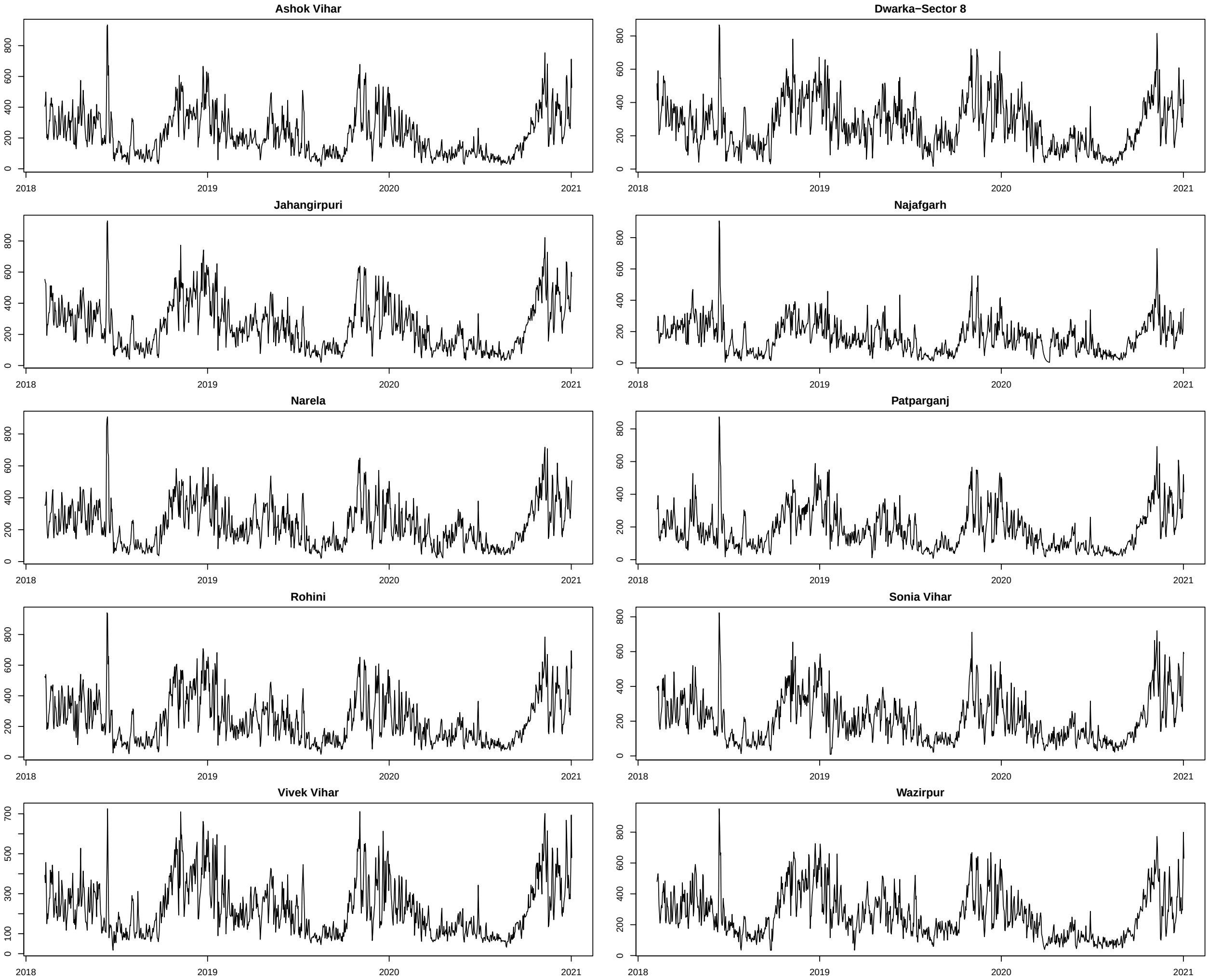

Figure: The above graph shows the time series plot of PM10 parameter for all 10 stations in the data.

Results

Conclusion