PollutionDataMadrid

Vivek

24/12/2019

Setting the root directory

```
install.packages("knitr", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/toviv/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## package 'knitr' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\toviv\AppData\Local\Temp\RtmpKsKwVq\downloaded_packages
```

```
knitr::opts_knit$set(root.dir = "D:/IE/Term 1/Programming R/Practice/data/")
```

Changing the Current working directory to the location where all datasets are present:

```
setwd("D:/IE/Term 1/Programming R/Practice/data/")
Myfiles<-list.files(path = "D:/IE/Term 1/Programming R/Practice/data/")
```

Here, 72 datasets consist information about hourly pollution in Madrid, Spain from the period 2011 to 2016. These datasets should first be combined into a single dataset before beginning our analysis:

```
Read_Madrid_Data <- function(y, m) {
  Myfiles <- paste0('hourly_data_', y, '_', m, '.csv')
  z <- read.csv(Myfiles)
  z$date <- as.Date(paste0('20', y, '-', m, '-', z$day))
  z$day <- NULL
  return(z)
}
Entire_Madrid_Data <- function(ys,ms) {
  pollutants_H <- data.frame();
  for(y in ys) {
    for(m in ms) {

      monthly <- Read_Madrid_Data(y,m)
      pollutants_H <- rbind(pollutants_H, monthly )
    }
  }
  return(pollutants_H)
}

ys <- seq(11,16,1)
ms <-seq(1,12,1)

raw_data <- Entire_Madrid_Data(ys,ms)

# Deleting irrelevant information
raw_data$station<-NULL
```

Viewing the first 6 rows of the combined data:

```
head(raw_data)
```

```
##   hour parameter value       date
## 1    1         1     6 2011-01-01
## 2    1         1    12 2011-01-01
## 3    1         1    12 2011-01-01
## 4    1         1    10 2011-01-01
## 5    1         1     7 2011-01-01
## 6    1         1    11 2011-01-01
```

Calculating the mean of the parameters based on date and the parameter number

```
rawdata2<-aggregate(raw_data[ ,'value'],mean, by= list(date1=raw_data$date,parameter2=raw_dat
a$parameter), na.rm=T)
```

And checking for NULL values within this dataset:

```
sum(is.na(rawdata2))>0
```

```
## [1] FALSE
```

In this case, the following air-pollutants are considered:

```
NO2 <- rawdata2[rawdata2$parameter2==8,]
SO2 <- rawdata2[rawdata2$parameter2==1,]
O3 <- rawdata2[rawdata2$parameter2==14,]
PM2.5 <- rawdata2[rawdata2$parameter2==9,]
PM10 <- rawdata2[rawdata2$parameter2==10,]
CO <- rawdata2[rawdata2$parameter2==6,]
NO <- rawdata2[rawdata2$parameter2==7,]
```

A simple merge function is used to combine all data together:

```
merge1 <- merge(NO2,SO2,by.x = 'date1',by.y = 'date1')
merge2 <- merge(O3,PM2.5,by.x = 'date1',by.y = 'date1')
merge3 <- merge(PM10,CO,by.x = 'date1',by.y = 'date1')
merge4 <- merge(NO,merge1,by.x = 'date1',by.y = 'date1')
merge5 <- merge(merge2,merge3,by.x = 'date1',by.y = 'date1')
FinalData <- merge(merge4,merge5,by.x = 'date1',by.y = 'date1')

colnames(FinalData) <- c('Date','NO','avg_NO','NO2','avg_NO2','SO2','avg_SO2','O3','avg_O3',
'PM2.5','avg_PM2.5','PM10','avg_PM10','CO','avg_CO')
```

The final cleaned dataset looks like this:

```
head(FinalData)
```

```
##         Date NO   avg_NO NO2  avg_NO2 SO2   avg_SO2 O3    avg_O3 PM2.5
## 1 2011-01-01  7 16.53819   8 41.51042   1 10.712500 14 20.473214     9
## 2 2011-01-02  7 28.82292   8 48.47396   1 11.933333 14 15.562500     9
## 3 2011-01-03  7 71.14236   8 63.63368   1 11.912134 14  9.446429     9
## 4 2011-01-04  7 27.82639   8 46.29514   1  8.841667 14 13.342262     9
## 5 2011-01-05  7 36.02951   8 51.51736   1  9.508403 14 10.883929     9
## 6 2011-01-06  7 11.19271   8 35.32812   1  8.633333 14 23.419643     9
##   avg_PM2.5 PM10  avg_PM10 CO    avg_CO
## 1  9.363636   10 13.831579  6 0.3725000
## 2  9.076389   10 13.656250  6 0.4529167
## 3 11.944444   10 21.555556  6 0.5325000
## 4  9.402778   10 14.357639  6 0.3670833
## 5 10.513889   10 15.597222  6 0.3950000
## 6  6.979167   10  9.753472  6 0.2854167
```

To read the excel files, since the weather data is available in .xlsx format:

```
library("xlsx")
weather <- read.xlsx("weather.xlsx", sheetIndex = 1)
colnames(weather)[1] <- c('Date')
```

Merging the Pollution data with the weather information provides the data frame shown below:

```
MadridData <-merge(FinalData,weather,by.x = 'Date',by.y = 'Date')
head(MadridData)
```

```
##         Date NO   avg_NO NO2  avg_NO2 SO2   avg_SO2 O3    avg_O3 PM2.5
## 1 2011-01-01  7 16.53819   8 41.51042   1 10.712500 14 20.473214     9
## 2 2011-01-02  7 28.82292   8 48.47396   1 11.933333 14 15.562500     9
## 3 2011-01-03  7 71.14236   8 63.63368   1 11.912134 14  9.446429     9
## 4 2011-01-04  7 27.82639   8 46.29514   1  8.841667 14 13.342262     9
## 5 2011-01-05  7 36.02951   8 51.51736   1  9.508403 14 10.883929     9
## 6 2011-01-06  7 11.19271   8 35.32812   1  8.633333 14 23.419643     9
##   avg_PM2.5 PM10  avg_PM10 CO    avg_CO temp_avg temp_max temp_min
## 1  9.363636   10 13.831579  6 0.3725000      8.3     13.0      3.0
## 2  9.076389   10 13.656250  6 0.4529167      8.6     13.0      4.0
## 3 11.944444   10 21.555556  6 0.5325000      4.2      9.4     -1.6
## 4  9.402778   10 14.357639  6 0.3670833      6.5      8.0      4.1
## 5 10.513889   10 15.597222  6 0.3950000      8.9     10.0      6.3
## 6  6.979167   10  9.753472  6 0.2854167     12.2     15.0      8.9
##   precipitation humidity wind_avg_speed
## 1          0.00       84            5.2
## 2          0.00       81            5.4
## 3          0.00       86            3.5
## 4          0.00       93            6.3
## 5          0.00       90           10.4
## 6          0.51       87           15.7
```

To obtain the Linear Regression Model of the dataset:

```
multi_model<-lm(avg_NO2~ avg_O3+avg_SO2+avg_PM2.5+humidity+precipitation+wind_avg_speed+temp_
avg, data=MadridData)
summary(multi_model)
```

```
##
## Call:
## lm(formula = avg_NO2 ~ avg_O3 + avg_SO2 + avg_PM2.5 + humidity +
##     precipitation + wind_avg_speed + temp_avg, data = MadridData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.768  -4.631   0.168   4.440  24.986
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    62.52120    1.59964  39.085  < 2e-16 ***
## avg_O3         -0.39164    0.01182 -33.133  < 2e-16 ***
## avg_SO2         1.67329    0.07577  22.082  < 2e-16 ***
## avg_PM2.5       0.95055    0.03628  26.202  < 2e-16 ***
## humidity       -0.25297    0.01329 -19.036  < 2e-16 ***
## precipitation   0.17130    0.05100   3.359 0.000795 ***
## wind_avg_speed -0.66390    0.03457 -19.206  < 2e-16 ***
## temp_avg       -0.31467    0.03568  -8.820  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.095 on 2184 degrees of freedom
## Multiple R-squared:  0.8292, Adjusted R-squared:  0.8287
## F-statistic:  1515 on 7 and 2184 DF,  p-value: < 2.2e-16
```

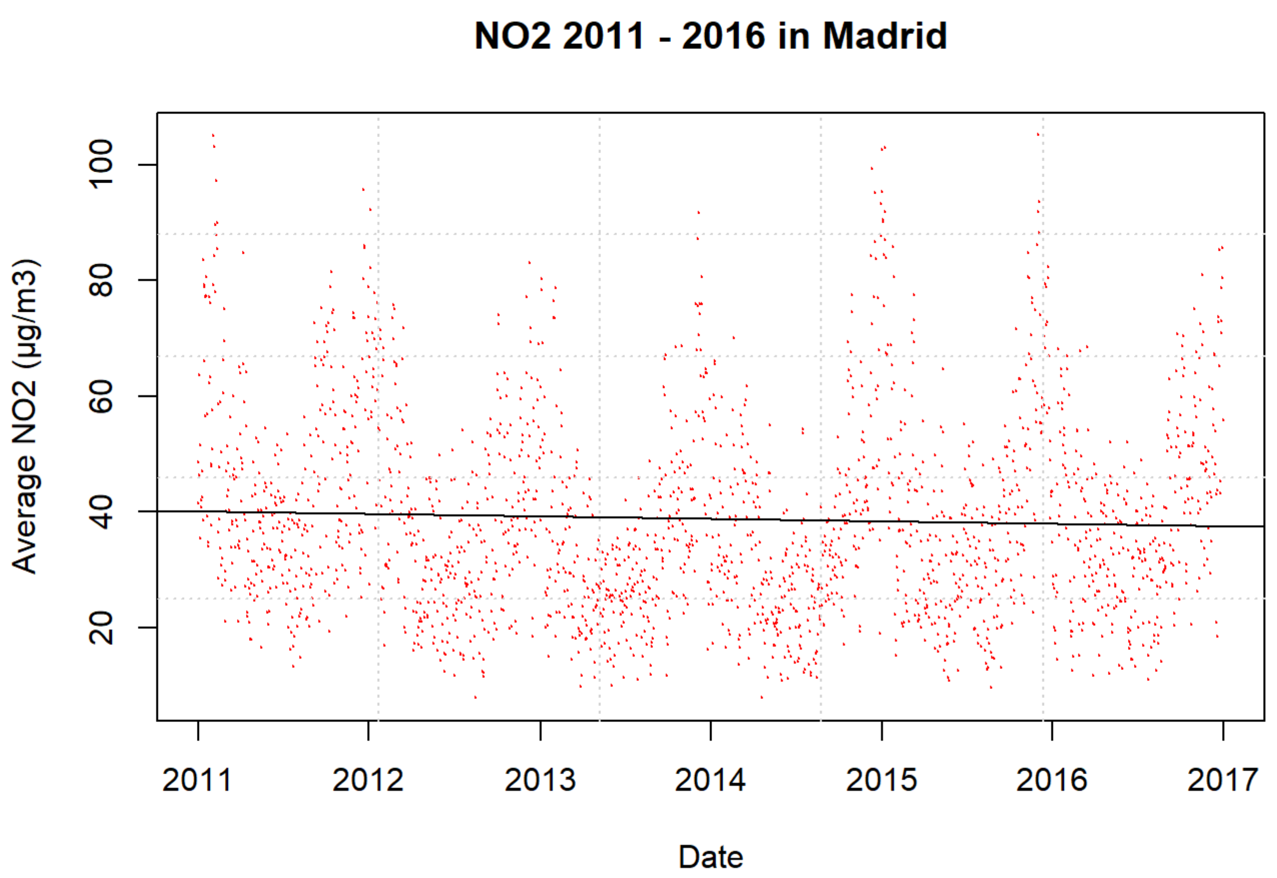
It can be seen from the summary that all the variables are statistically significant, that is, the *p-value* is within the default threshold of 0.95 confidence.

Now, to plot the distribution of air-pollutants over the years:

```
par(mar=c(5,5,5,1))
par(mfrow=c(1,1))
```

NO2 pollution is caused due to vehicular emissions, and also due to burning of fossil fuels.

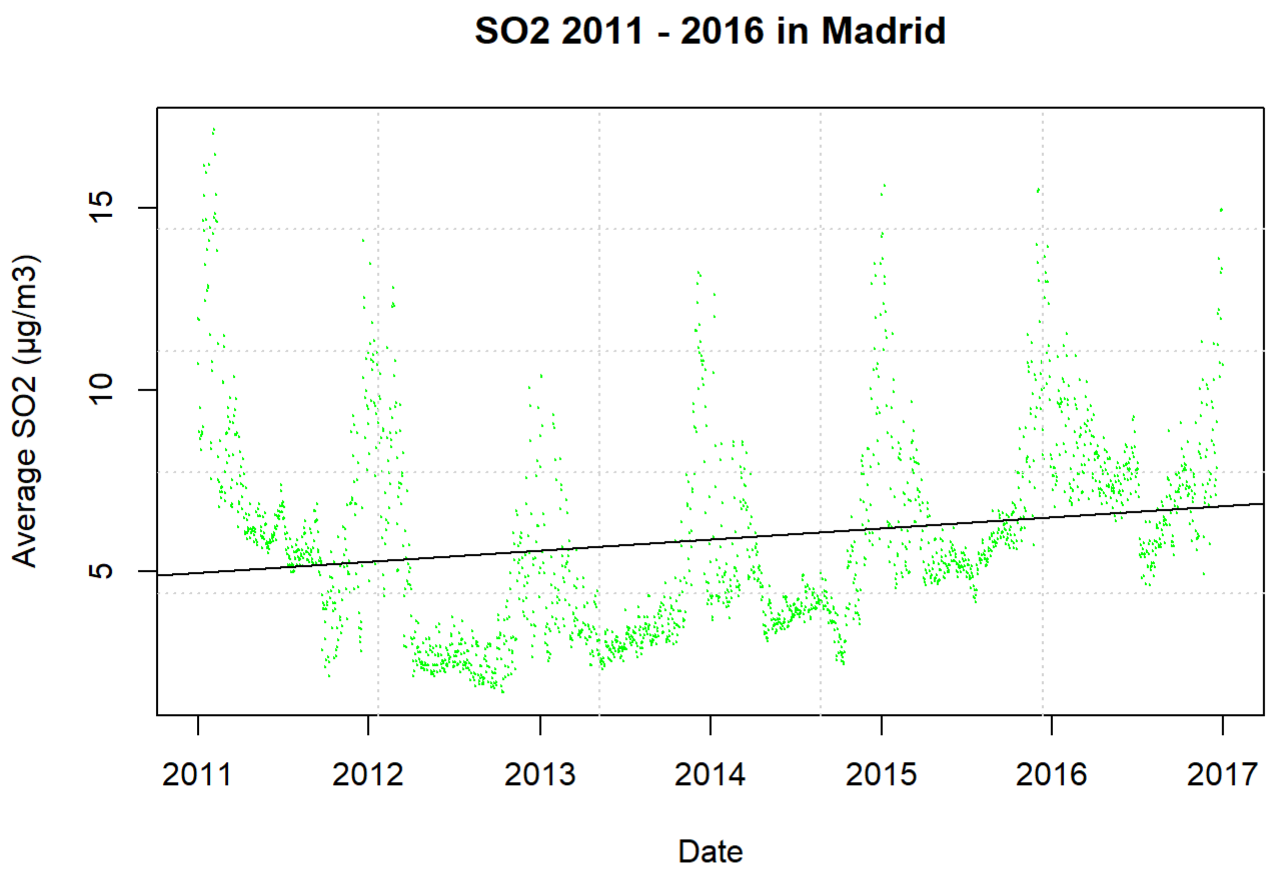
```
plot(MadridData$Date,MadridData$avg_NO2, col='red', pch=19 , cex=0.1,main='NO2 2011 - 2016 in
Madrid', xlab='Date', ylab='Average NO2 (µg/m3)'); grid(5)
abline(mC <- lm(avg_NO2 ~ Date, data = MadridData))
```




```
# As can be seen from the plot, the average emissions have reduced over the years
```

SO2 reacts with substances from the atmosphere to form acid rain.

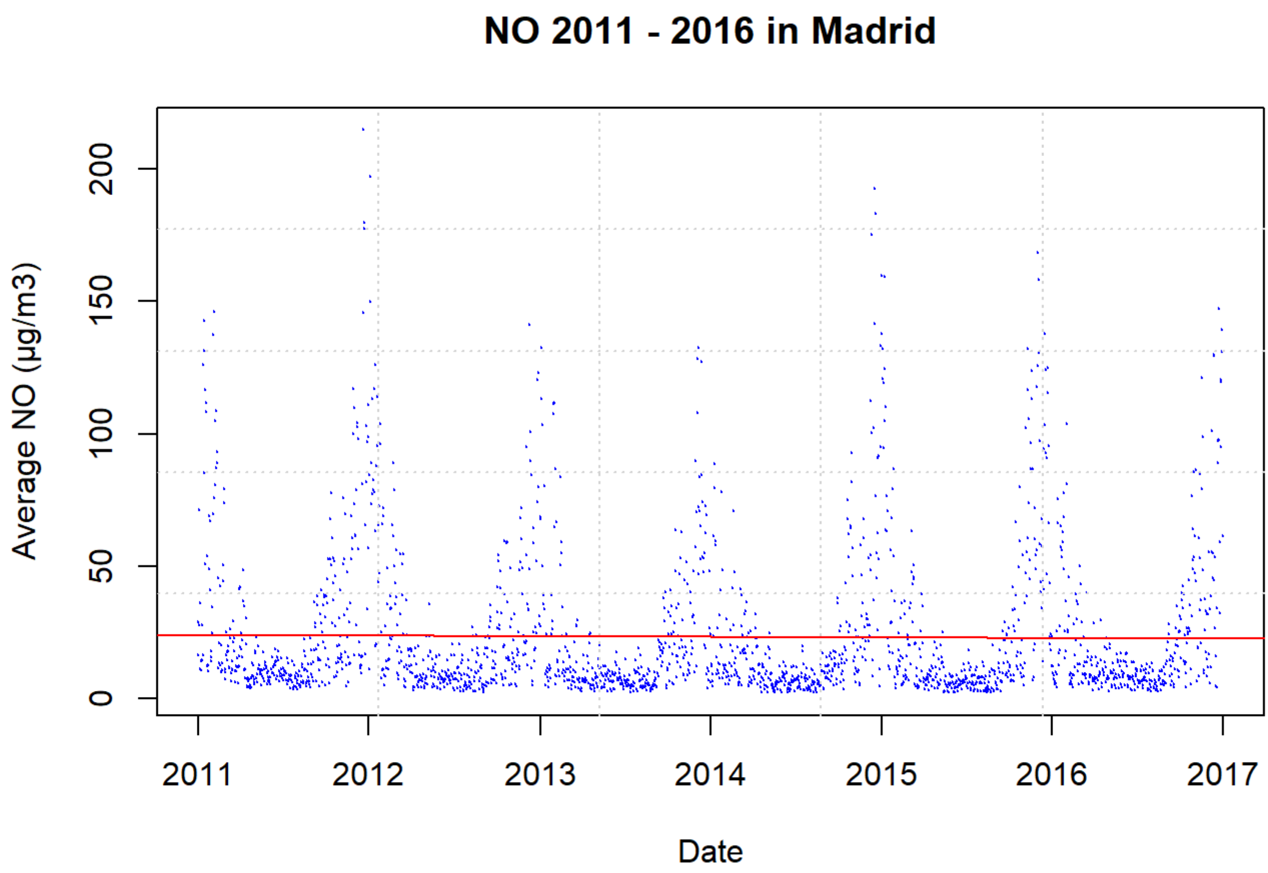
```
plot(MadridData$Date,MadridData$avg_SO2, col='green', pch=19 , cex=0.1,main='SO2 2011 - 2016
 in Madrid', xlab='Date', ylab='Average SO2 (µg/m3)'); grid(5)
abline(mC <- lm(avg_SO2 ~ Date, data = MadridData))
```




```
# As can be seen from the plot, the average concentration level of SO2 is on an increase sinc
e 2011
```

NO (Nitrous Oxide) comes from wastewaters.

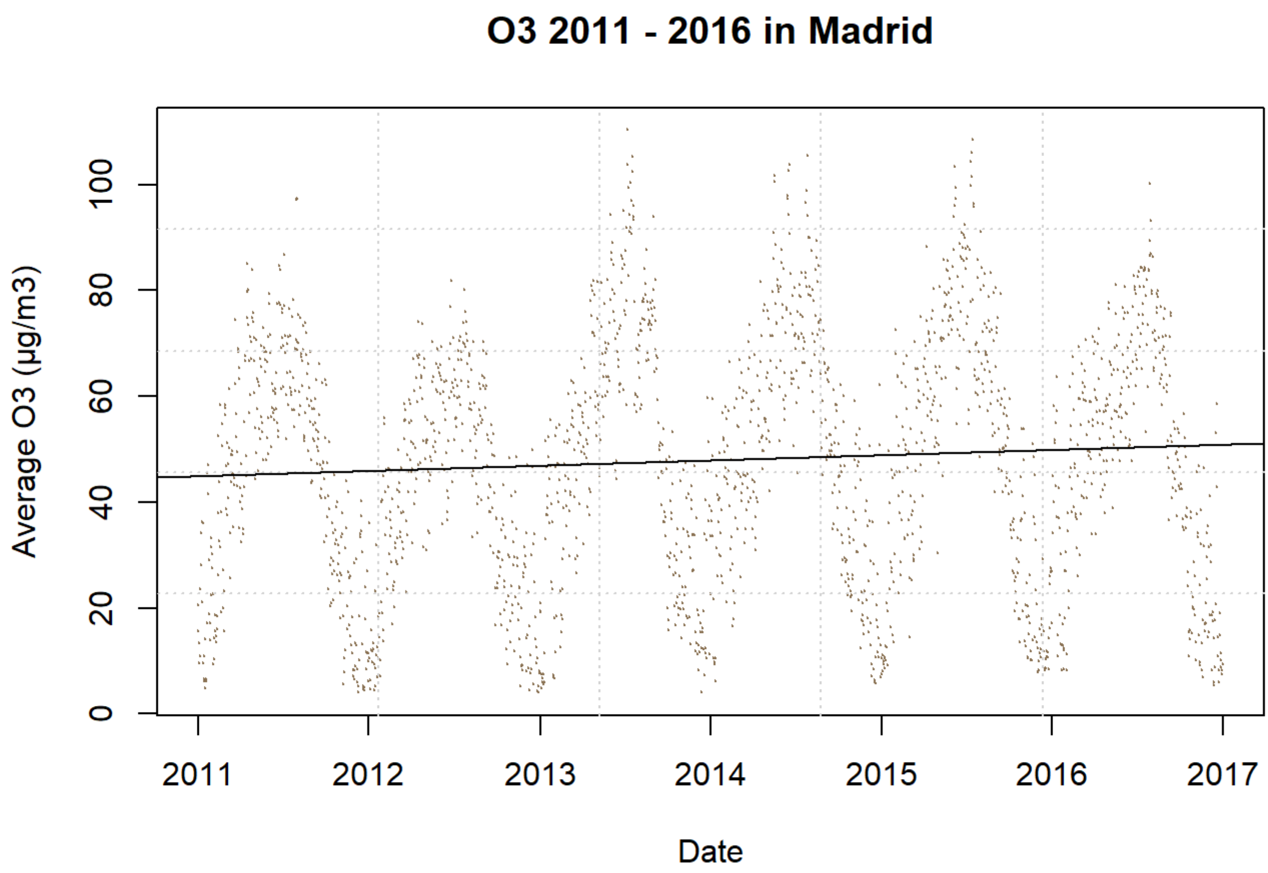
```
plot(MadridData$Date,MadridData$avg_NO, col='blue', pch=19 , cex=0.1,main='NO 2011 - 2016 in
 Madrid', xlab='Date', ylab='Average NO (µg/m3)'); grid(5)
abline(mC <- lm(avg_NO ~ Date, data = MadridData), col='red')
```




```
# As can be seen from the plot, the average concentration level of NO has remained fairly con
stant over the years
```

O3 (Ozone) is different from the ozone that protects people from the sun. It is created on the ground when volatile organic compounds chemically react with oxides of nitrogen in the presence of sunlight.

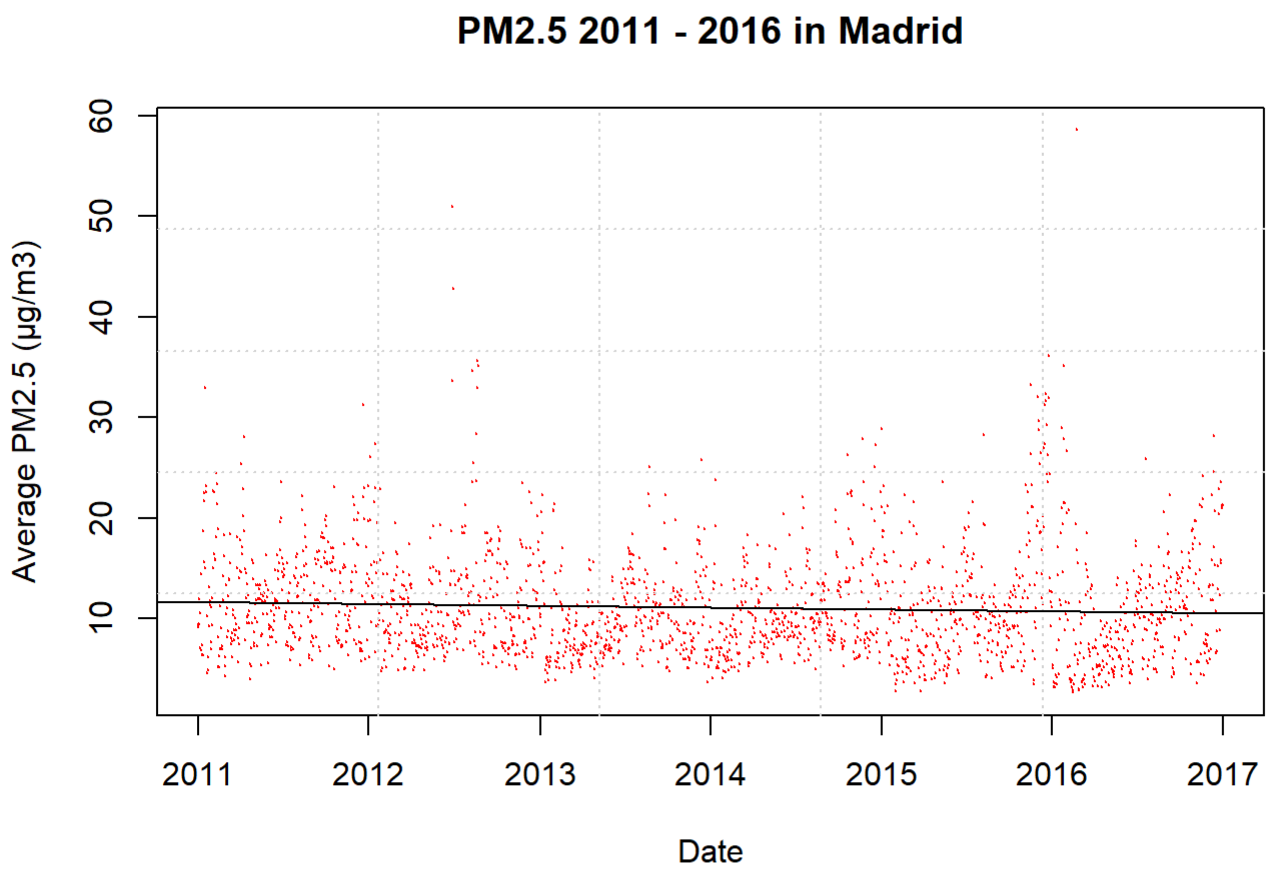
```
plot(MadridData$Date,MadridData$avg_O3, col='burlywood4', pch=19 , cex=0.1, main='O3 2011 - 2
016 in Madrid', xlab='Date', ylab='Average O3 (µg/m3)'); grid(5)
abline(mC <- lm(avg_O3 ~ Date, data = MadridData))
```




```
# As can be seen from the plot, the average concentration level of O3 has gradually increased over the years
```

PM2.5 has been plotted as shown below:

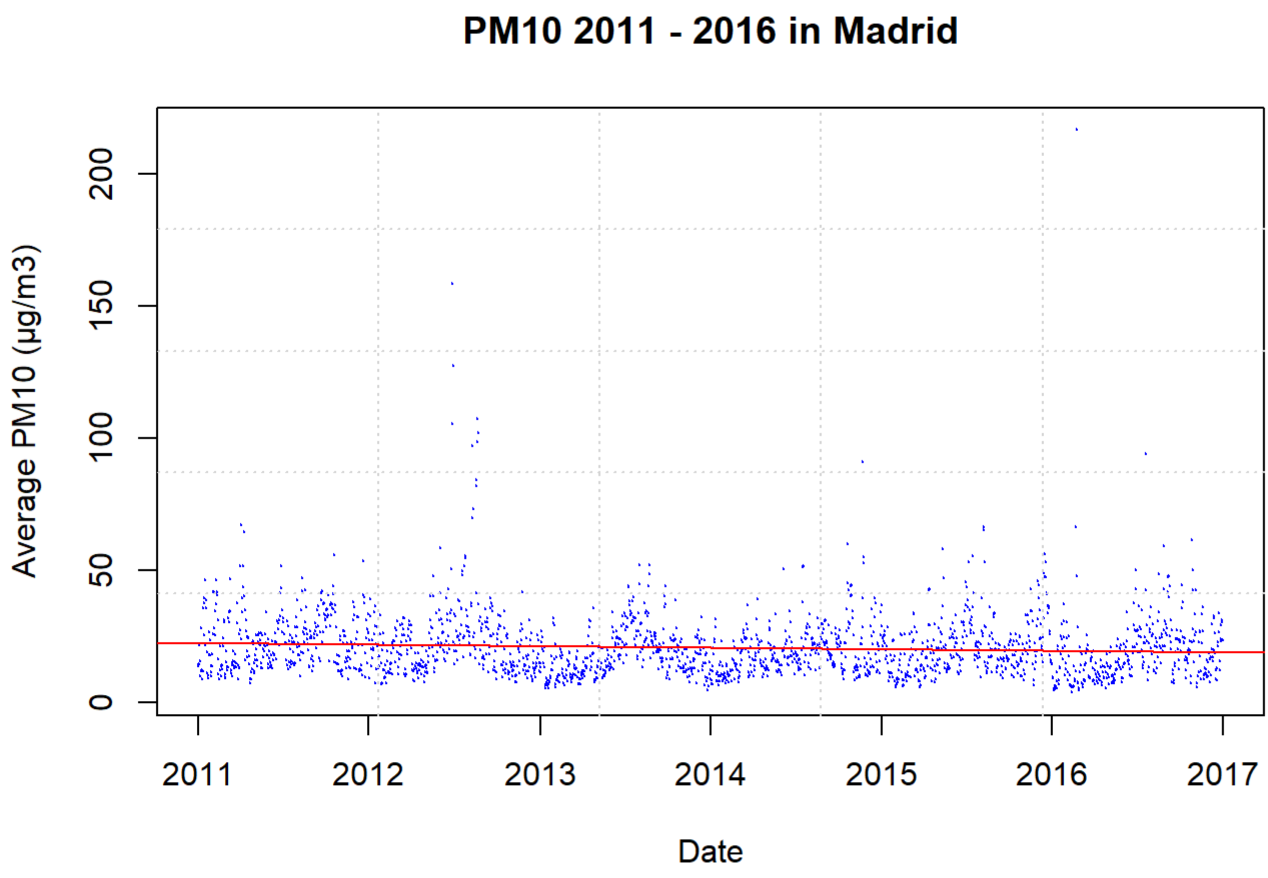
```
plot(MadridData$Date,MadridData$avg_PM2.5, col='red', pch=19 , cex=0.1, main='PM2.5 2011 - 2016 in Madrid', xlab='Date', ylab='Average PM2.5 (µg/m3)'); grid(5)
abline(mC <- lm(avg_PM2.5 ~ Date, data = MadridData))
```




```
# As can be seen from the plot, the average concentration level of PM2.5 decreased slightly o
ver the years
```

PM10 has been plotted as shown below:

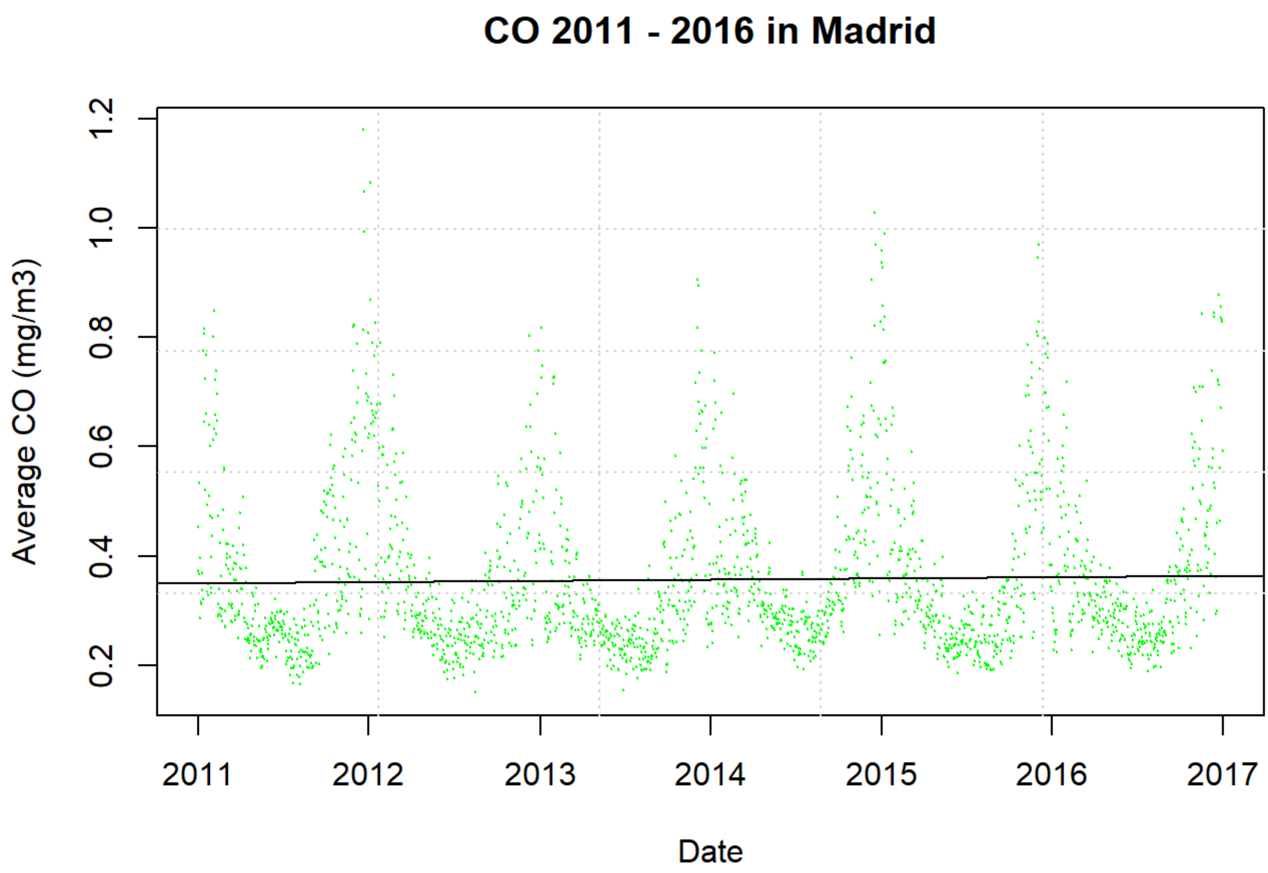
```
plot(MadridData$Date,MadridData$avg_PM10, col='blue', pch=19 , cex=0.1, main='PM10 2011 - 201
6 in Madrid', xlab='Date' ,ylab='Average PM10 (µg/m3)'); grid(5)
abline(mC <- lm(avg_PM10 ~ Date, data = MadridData), col='red')
```




```
# As can be seen from the plot, the average concentration level of PM10 has remained constant
over the years
```

The graph of CO (Carbon Mono-oxide) has been plotted as shown below:

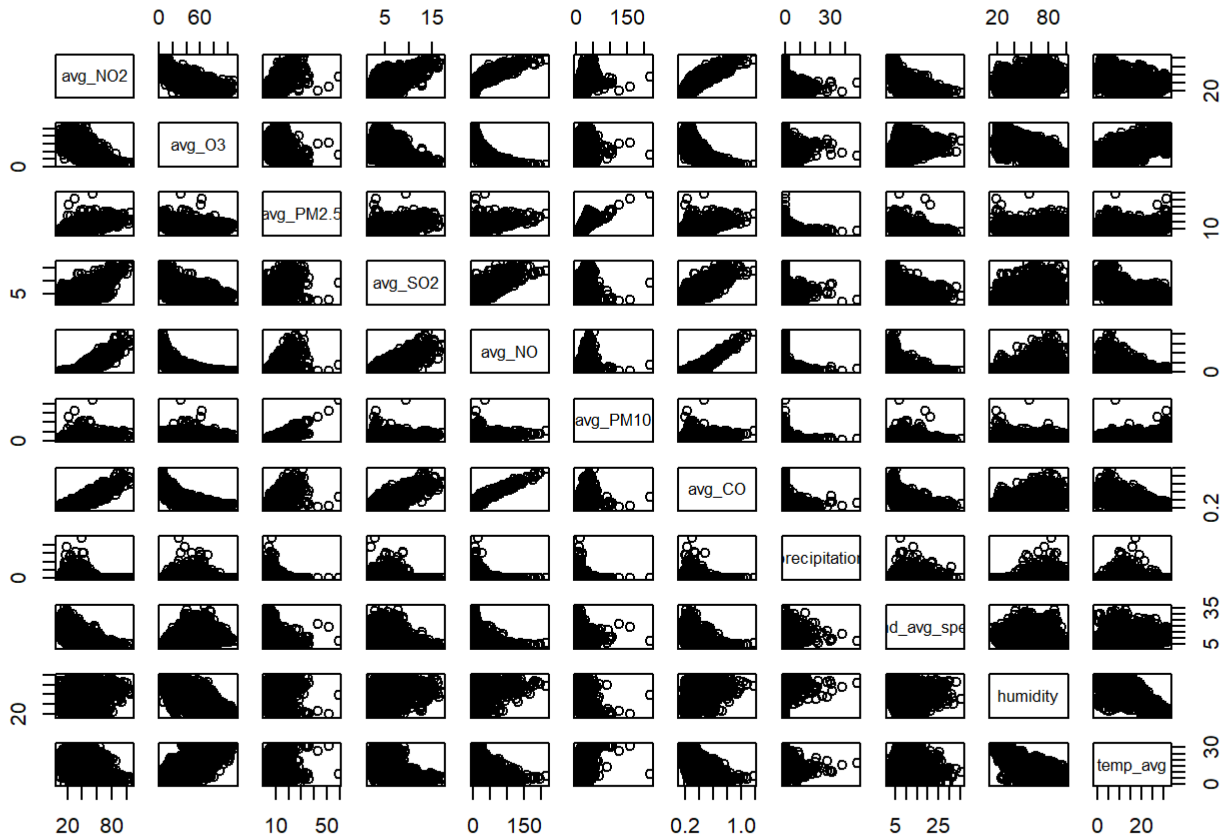
```
plot(MadridData$Date,MadridData$avg_CO, col='green', pch=19 , cex=0.1, main='CO 2011 - 2016 i
n Madrid', xlab='Date', ylab='Average CO (mg/m3)'); grid(5)
abline(mC <- lm(avg_CO ~ Date, data = MadridData))
```




```
# As can be seen from the plot, the co average ncentration level of CO has slightly increased over the years
```

Finally, the Correlation Matrix for each variable can be plotted as shown below:

```
pairs(MadridData[,c('avg_NO2','avg_O3','avg_PM2.5','avg_SO2','avg_NO','avg_PM10', 'avg_CO','precipitation', 'wind_avg_speed','humidity','temp_avg')])
```



Alternately, the correlation can also be shown using the correlation function:

```
res <-cor(MadridData[,c('avg_NO','avg_NO2','avg_O3','avg_CO','avg_PM2.5','avg_PM10','avg_SO2'
,'temp_avg','humidity','precipitation','wind_avg_speed')])
round(res,2)
```

```
##                avg_NO avg_NO2 avg_O3 avg_CO avg_PM2.5 avg_PM10 avg_SO2 temp_avg
## avg_NO           1.00    0.86  -0.72   0.95      0.58     0.28    0.69    -0.47
## avg_NO2          0.86    1.00  -0.71   0.89      0.64     0.37    0.67    -0.38
## avg_O3          -0.72   -0.71   1.00  -0.76     -0.33    -0.04   -0.51     0.67
## avg_CO           0.95    0.89  -0.76   1.00      0.57     0.24    0.72    -0.57
## avg_PM2.5        0.58    0.64  -0.33   0.57      1.00     0.85    0.34     0.07
## avg_PM10         0.28    0.37  -0.04   0.24      0.85     1.00    0.09     0.35
## avg_SO2          0.69    0.67  -0.51   0.72      0.34     0.09    1.00    -0.49
## temp_avg        -0.47   -0.38   0.67  -0.57      0.07     0.35   -0.49     1.00
## humidity         0.38    0.27  -0.69   0.44      0.00    -0.30    0.32    -0.77
## precipitation   -0.12   -0.13  -0.02  -0.10     -0.21    -0.21   -0.05    -0.11
## wind_avg_speed  -0.46   -0.60   0.35  -0.49     -0.46    -0.28   -0.35     0.01
##                humidity precipitation wind_avg_speed
## avg_NO             0.38         -0.12          -0.46
## avg_NO2            0.27         -0.13          -0.60
## avg_O3            -0.69         -0.02           0.35
## avg_CO             0.44         -0.10          -0.49
## avg_PM2.5          0.00         -0.21          -0.46
## avg_PM10          -0.30         -0.21          -0.28
## avg_SO2            0.32         -0.05          -0.35
## temp_avg          -0.77         -0.11           0.01
## humidity           1.00          0.27          -0.06
## precipitation      0.27          1.00           0.13
## wind_avg_speed    -0.06          0.13           1.00
```