Pollution before and after lockdown in India

M.G.Manjusha

2024-07-14

```
#importing libraries
library(ggplot2)
library(readr)
library(dplyr)
library(tidyr)
library(reprex)
library(stringr)
library(maps)
```

Information about Dataset

The dataset is taken from the following source and downloaded to the system.
https://hub.tumidata.org/dataset/airpollutionbeforeandafterlockdownindelhi_delhi
The above dataset gives values of the pollutants in India from 1st January 2020 to 19th May 2020.Which is the timeline of lockdown when the covid 19 virus was rapidly spreading. It includes names of the states as well as cities. It consists of 15 columns and 687205 rows of data.
The data was being recorded every one hour from 12 AM to 11PM each day.

Columns:

- datetime: Gives the date and the time of when the value was recorded.Each day has recordings from 12 AM to 11 PM
- id : Gives the unique id for each of the cities and states.
- name : Name of the sector
- longitude
- latitude
- live: A boolean of True or False. If the data was recorded live it has a True value else it has False value.
- cityid : name of the city
- stateid : name of the state
- PM2.5 : Pollutant:Particulate Matter with a diameter of 2.5 micrometers or less, which can penetrate deep into the respiratory system.
- PM10:Pollutant: Particulate Matter with a diameter of 10 micrometers or less, which can be inhaled into the respiratory system and cause health issues.
- NO2:Pollutant: Nitrogen Dioxide, a harmful gas primarily emitted from vehicle exhaust and combustion processes, contributing to air pollution and respiratory problems.
- NH3:Pollutant:Ammonia, a compound released from agricultural activities, livestock waste, and industrial processes, contributing to air pollution and environmental concerns.
- SO2:Pollutant:Sulfur Dioxide, a gas released from burning fossil fuels containing sulfur, contributing to air pollution and acid rain formation.
- CO: Pollutant:Carbon Monoxide, which can cause harmful health effects by reducing oxygen transport in the bloodstream.
- OZONE: Pollutant: Ground-level Ozone, causing respiratory issues and environmental damage.

Reading Dataset

```
#Reading the dataset
pollution <- read_csv("/Users/mgmanjusha/Documents/NEU/Sem-1/IDMP/Hw-pratice/HW2/combined.csv"
                      ,show_col_types = FALSE)
```

Printing Summary of data

```
summary(pollution)
```

```
##    datetime              id                 name              longitude
##  Length:687205      Length:687205      Length:687205      Min.   :72.52
##  Class :character   Class :character   Class :character   1st Qu.:76.30
##  Mode  :character   Mode  :character   Mode  :character   Median :77.23
##                                                           Mean   :78.09
##                                                           3rd Qu.:78.28
##                                                           Max.   :92.72
##
##     latitude         live             cityid            stateid
##  Min.   : 8.515   Mode :logical   Length:687205      Length:687205
##  1st Qu.:19.253   FALSE:104649    Class :character   Class :character
##  Median :26.121   TRUE :582556    Mode  :character   Mode  :character
##  Mean   :23.878
##  3rd Qu.:28.625
##  Max.   :31.620
##
##      PM2.5            PM10             NO2              NH3
##  Min.   :  0.0   Min.   :  0.0    Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 42.0   1st Qu.: 59.0    1st Qu.: 13.00   1st Qu.:  3.00
##  Median : 72.0   Median : 94.0    Median : 24.00   Median :  5.00
##  Mean   :112.9   Mean   :108.4    Mean   : 34.49   Mean   :  6.35
##  3rd Qu.:149.0   3rd Qu.:134.0    3rd Qu.: 46.00   3rd Qu.:  8.00
##  Max.   :500.0   Max.   :500.0    Max.   :436.00   Max.   :125.00
##  NA's   :82781   NA's   :144575   NA's   :82427    NA's   :209625
##       SO2               CO             OZONE
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0
##  1st Qu.:  7.00   1st Qu.: 20.00   1st Qu.: 12
##  Median : 11.00   Median : 34.00   Median : 27
##  Mean   : 16.14   Mean   : 42.69   Mean   : 37
##  3rd Qu.: 20.00   3rd Qu.: 56.00   3rd Qu.: 51
##  Max.   :140.00   Max.   :200.00   Max.   :364
##  NA's   :100178   NA's   :89629    NA's   :120661
```

Tidying the data

The dataset has 15 columns, out of which 2 columns can be discarded. The columns “live” and “name” are discarded.

```
#removing unused columns
pollution_updated <- subset(pollution, select = -c(live,name))
```

In the below code, I am changing the existing datetime column to represent only dates and not times. Also, I am piping the updated data set into group_by and summarise to take the mean of all the pollutant values from 12 AM to 11 PM and make one row per day. Then I omitted the NA values.

```
pollution_updated <- mutate(pollution_updated,datetime=as.Date(datetime,format="%A, %d
                                                               %b %Y, %I:%M %p"))
pollution_updated <- pollution_updated %>%
  group_by(datetime,cityid,id,stateid) %>%
  summarise(across(where(is.numeric),~mean(.,na.rm=TRUE)))
```

Number of rows before omitting NA values

```
nrow(pollution_updated)
```

```
## [1] 28658
```

Omitting NA values

Number of rows after omitting NA values

```
pollution_updated <- na.omit(pollution_updated)
nrow(pollution_updated)
```

```
## [1] 19106
```

In the below code, I made a new column named lockdown_phase which tells which phase of lockdown was going on according to the given dates.
The lockdown in India was phase wise as below:

- 1st Jan 2020 - 24th March 2020 - Before lockdown
- 25th March 2020 - 14th April 2020 - Lockdown Phase 1
- 15th April 2020 - 3rd May 2020 - Lockdown Phase 2
- 4th May 2020 - 17th May 2020 - Lockdown Phase 3
- 18th May 2020 - 31st May 2020 - Lockdown Phase 4

```
pollution_updated$lockdown_phase <- cut(
  pollution_updated$datetime,
  breaks = as.Date(c("2020-01-01", "2020-03-24", "2020-04-14", "2020-05-03",
                     "2020-05-17", "2020-05-31")),
  labels = c("Before Lockdown", "Phase 1", "Phase 2", "Phase 3", "Phase 4"),
  include.lowest = TRUE )
```

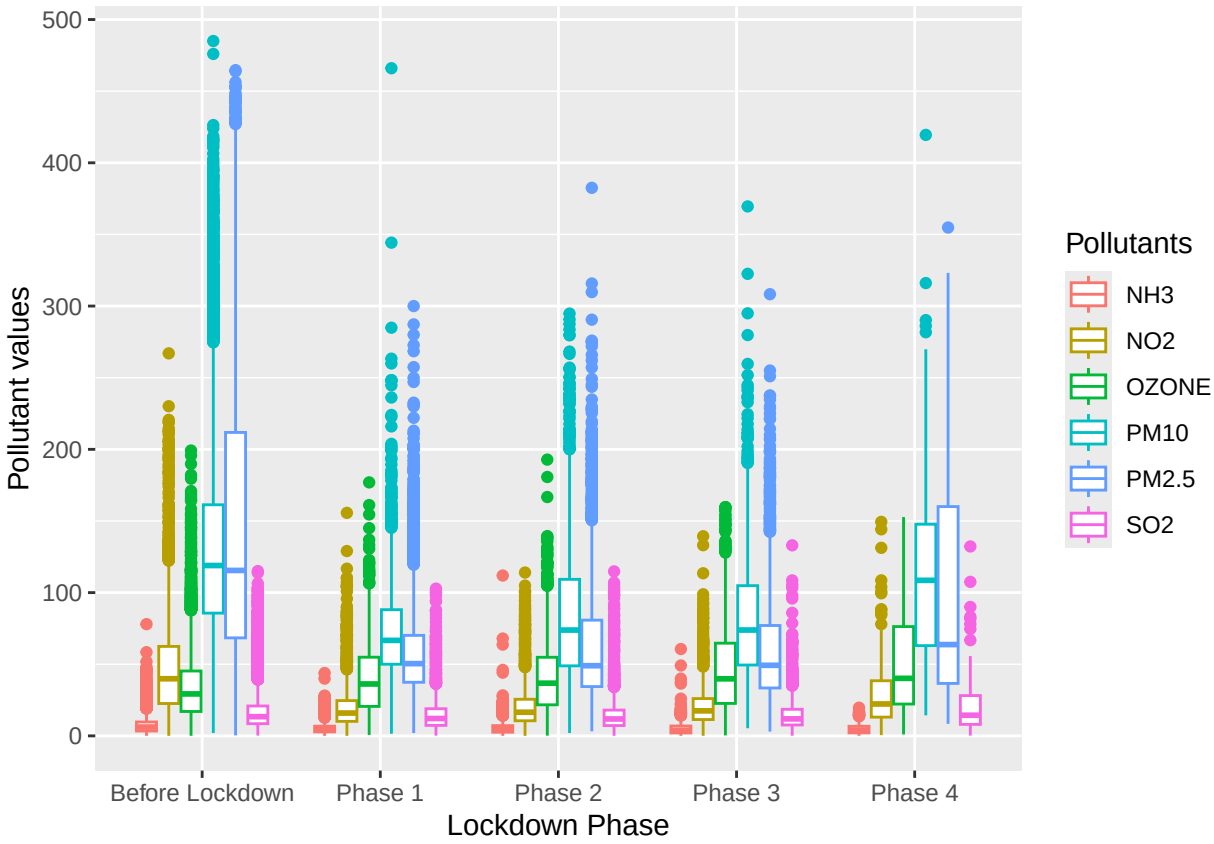
Finally, I have created a data frame called pollution_long where i selected all the columns necessary for visualization and used pivot_longer to get all the pollutant values in one column.

```
pollution_long <- pollution_updated %>%
  select(cityid, datetime, stateid, PM2.5, PM10, NO2, NH3, SO2, OZONE,lockdown_phase,longitude,latitude
  pivot_longer(cols = c("PM2.5", "PM10", "NO2", "NH3", "SO2", "OZONE"),
               names_to = "Pollutants",
               values_to = "Pollutant_values")
```

Visualizing plots

Plot 1

```
ggplot(pollution_long,aes(x=lockdown_phase,y=Pollutant_values))+
  geom_boxplot(aes(color=Pollutants))+
  xlab("Lockdown Phase")+
  ylab("Pollutant values")
```

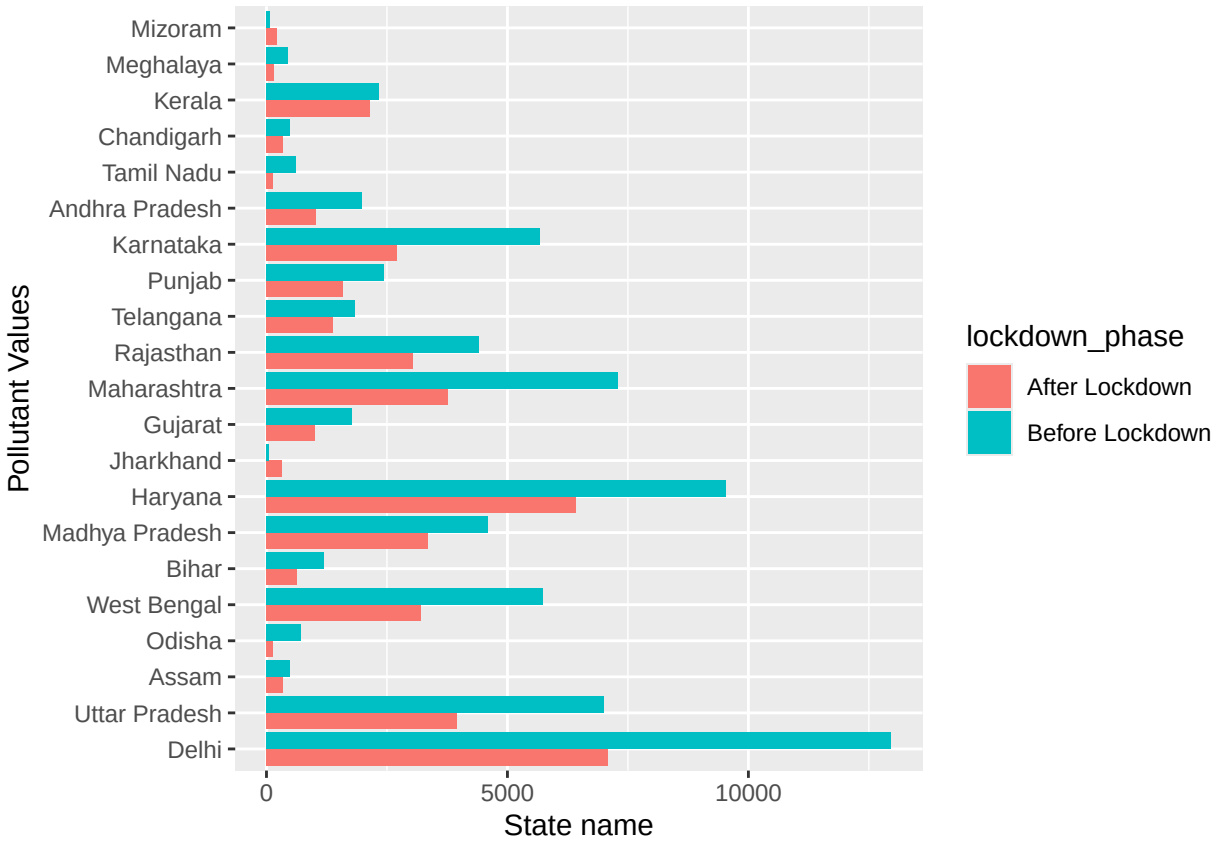


In the above plot, we can see the pollutant's minimum,maximum,and median values. All the pollutants are color coded.
Observations:

- NH3 pollutant is less compared to other pollutants before and after lockdown(Phase 1,2,3,and 4)
- The pollutants which are a lot are: PM10 and PM2.5
- All polutants except NH3 and SO2 had significant reduction after lockdown compared to before lockdown.

Plot 2

```
pollution_temp<- pollution_long
pollution_temp$lockdown_phase <- as.character(pollution_temp$lockdown_phase)
pollution_temp<- mutate(pollution_temp,lockdown_phase = ifelse(lockdown_phase %in% c("Phase 1", "Phase 1

ggplot(pollution_temp,aes(y=reorder(stateid,-Pollutant_values)))+
  geom_bar(aes(fill=lockdown_phase),position = "dodge")+
  ylab("Pollutant Values")+
  xlab("State name")
```

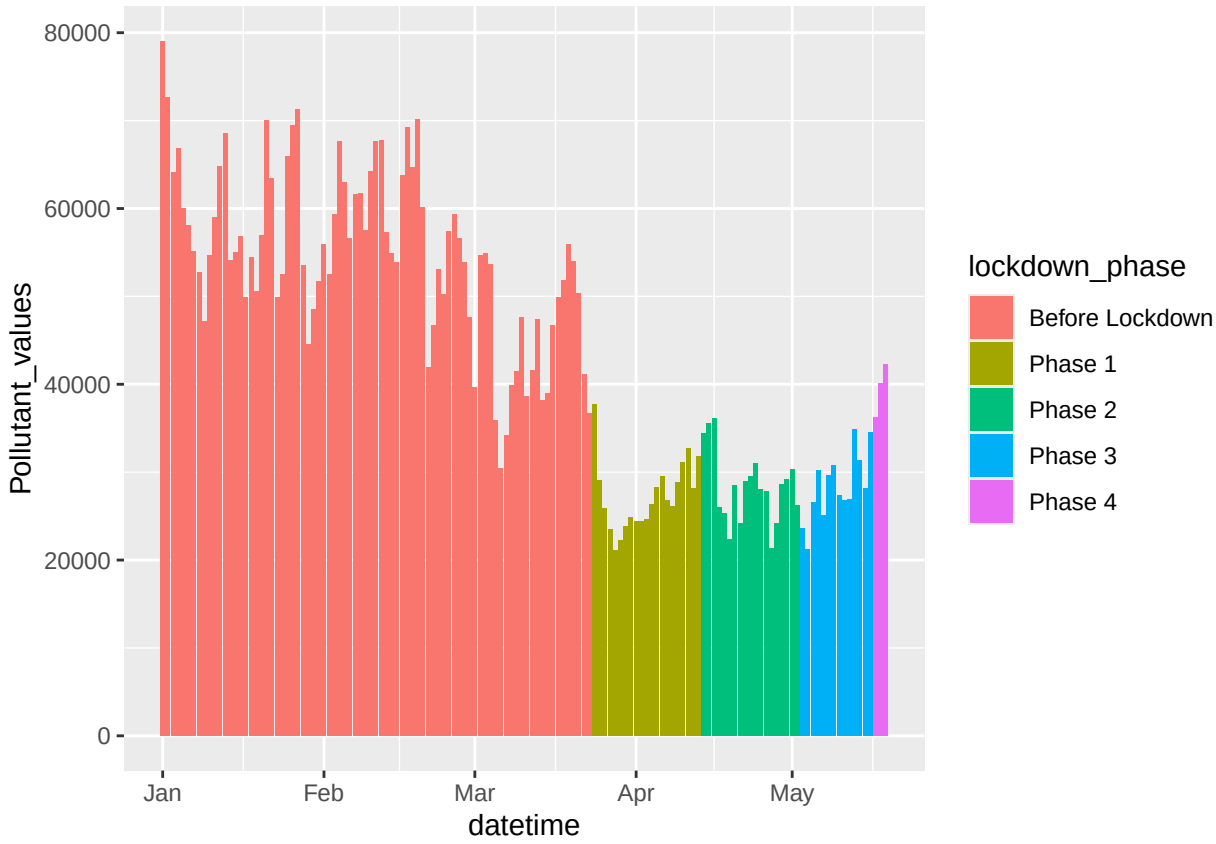


From the above plot, notable points are:

- Almost every state's pollution has reduced after lockdown.
- Jharkand and Mizoram are the only states that are not inclining towards the trend of decrease in pollution after lockdown.
- Delhi is the most polluted state before and after lockdown.
- Harayana is the second most polluted state before and after lockdown.
- Jarkhand is the least polluted state before lockdown.
- Odisha is the least polluted state after lockdown.

Plot 3

```
ggplot(pollution_long,aes(x=datetime,y=Pollutant_values))+
  geom_col(aes(fill=lockdown_phase))
```

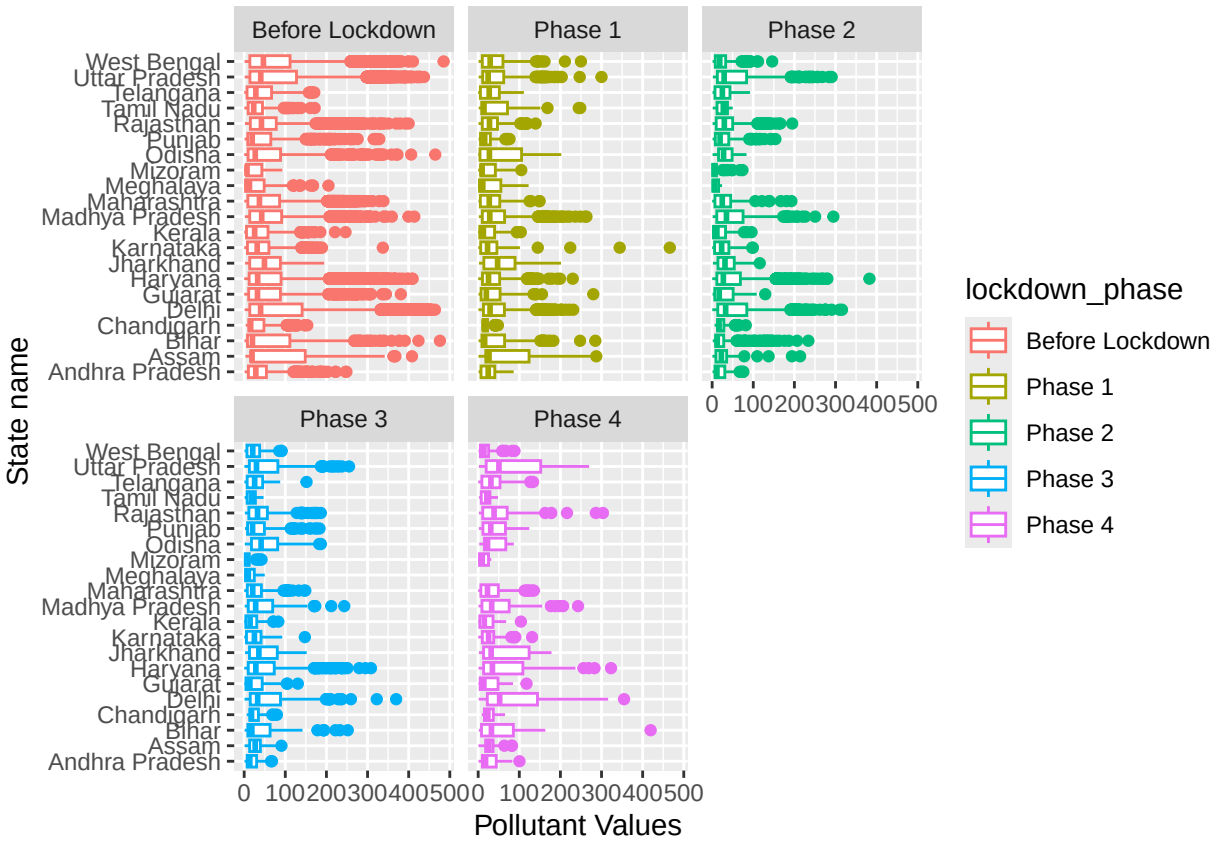


Observations:

- As we can see from the above plot, the pollution decreased drastically after lockdown. In this plot we can see the difference as each month goes by.
- A point to be noted is that, At end of each phase, we see an increase of pollutants compared to the least value in that phase of lockdown.
- We can infer that the trend of pollutant values is not continuously decreasing.
- At the beginning of each phase of lockdown including before lockdown, we see a drastic fall in the pollutant values. However, this trend does not hold true for Phase 4.

Plot 4

```
ggplot(pollution_long,aes(x=stateid,y=Pollutant_values))+
  geom_boxplot(aes(color=lockdown_phase))+
  facet_wrap(~lockdown_phase,)+
  ylab("Pollutant Values")+
  xlab("State name")+
  coord_flip()
```



From the graph above, We can see how the range of pollutant values changes from before lockdown to Phase 4

Observations: - Pollutant values are high before lockdown and have a drastic drop in Phase 1 - Pollutant values keep changing as phases increase

Here, I am creating a new column which tells if the pollutant value is low, medium or high according to the pollutant value with respective to the pollutant. I have done research for the ranges of pollutant values, the links will be provided in the end.

```
pollution_range <- function(pollutant, value) {
  if (pollutant == "PM2.5") {
    if (value <= 12) return("Low")
    else if (value <= 12 & value <= 35) return("Medium")
    else return("High")
  } else if (pollutant == "PM10") {
    if (value < 100) return("Low")
    else if (value >= 100 & value <= 250) return("Medium")
    else return("High")
  } else if (pollutant == "NO2") {
    if (value < 50) return("Low")
    else if (value >= 50 & value <= 200) return("Medium")
    else return("High")
  } else if (pollutant == "NH3") {
    if (value < 100) return("Low")
    else if (value >= 100 & value <= 300) return("Medium")
    else return("High")
  } else if (pollutant == "SO2") {
    if (value < 80) return("Low")
    else if (value >= 80 & value <= 200) return("Medium")
```

```
    else return("High")
  } else if (pollutant == "OZONE") {
    if (value < 35) return("Low")
    else if (value >= 30 & value <= 180) return("Medium")
    else return("High")
  } else {
    return("Unknown")
  }
}

pollution_long <- pollution_long %>%
  mutate(Pollution_Range = mapply(pollution_range, Pollutants, Pollutant_values))
```

Here, I am making a new dataframe which summarizes the pollutant range for each city

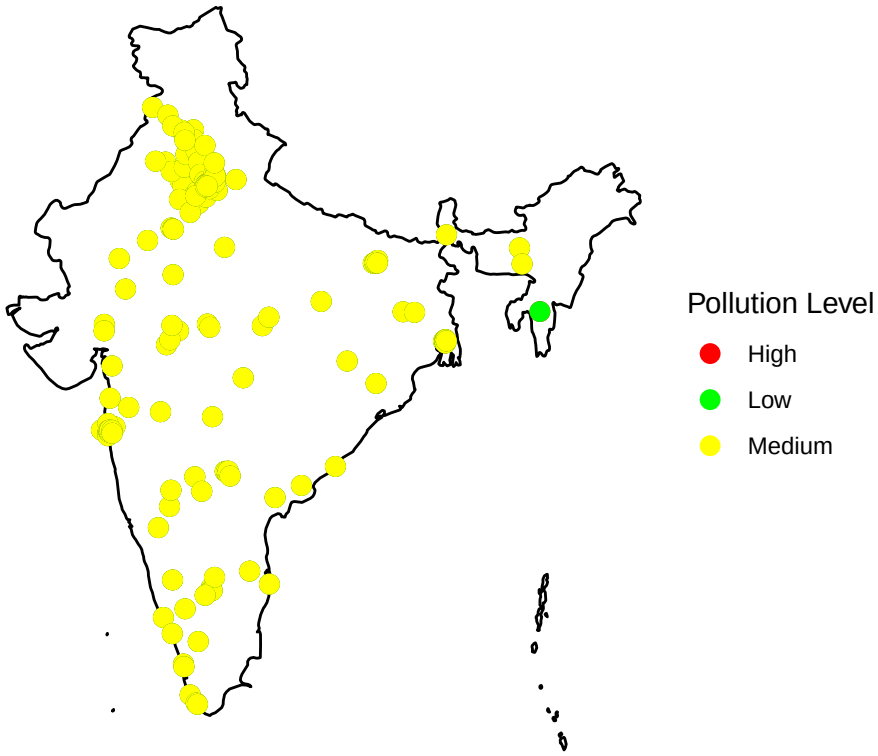
```
#Making a new dataframe
predominant_range <- function(high, medium, low) {
  if (max(c(high, medium, low)) == high) {
    return("High")
  } else if (max(c(high, medium, low)) == medium) {
    return("Medium")
  } else {
    return("Low")
  }
}

city_pollution_summary <- pollution_long %>%
  group_by(stateid, cityid, lockdown_phase,longitude,latitude,Pollution_Range) %>%
  summarise(
    High = sum(Pollution_Range == "High"),
    Medium = sum(Pollution_Range == "Medium"),
    Low = sum(Pollution_Range == "Low"),
    .groups = "drop"
  ) %>%
  mutate(Predominant_Range = mapply(predominant_range, High, Medium, Low))
```

Plot 5

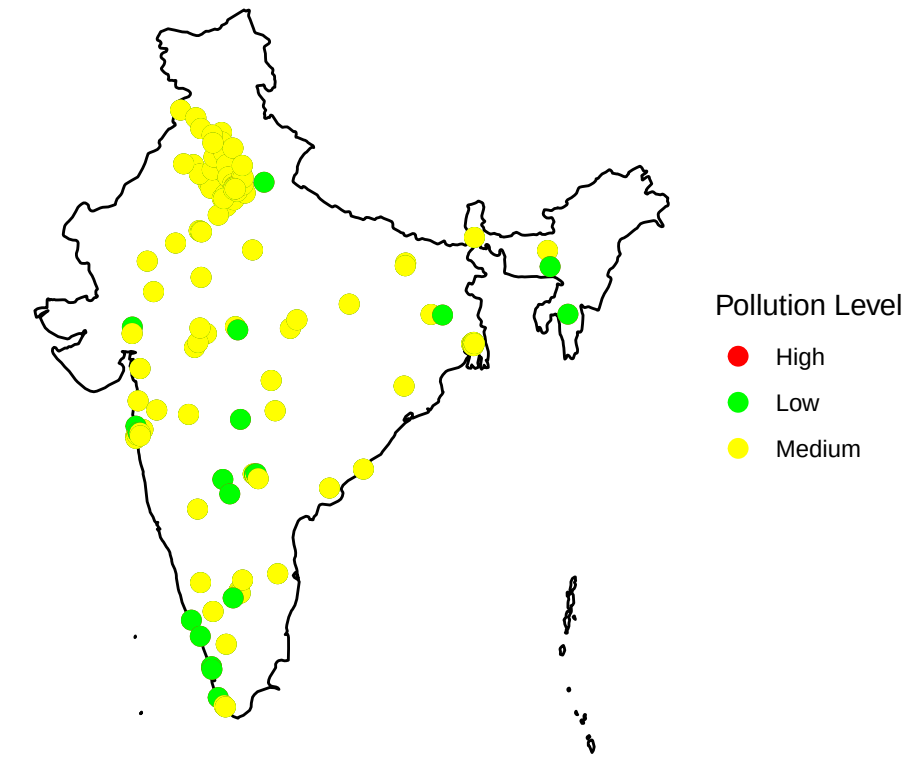
```
india_map <- map_data("world", region = "India")
india_map_df <- fortify(india_map)

pollution_long_before_lockdown <- city_pollution_summary |>
  filter(lockdown_phase=="Before Lockdown")

pollution_long_after_lockdown <- city_pollution_summary |>
  filter(lockdown_phase=="Phase 2")

ggplot(india_map_df) +
  geom_polygon(aes(x = long, y = lat, group = group), fill = "White", color = "black") +
  geom_point(data = pollution_long_before_lockdown,
             aes(x = longitude, y = latitude, color = Pollution_Range), size = 3) +
  scale_color_manual(name = "Pollution Level",
                     values = c("Low" = "Green", "Medium" = "Yellow", "High" = "Red")) +
  theme_void() +
  coord_fixed(ratio = 1.19)+
```

```
  ggtitle("Pollution Levels Across India Before Lockdown")
```

Pollution Levels Across India Before Lockdown




```
ggplot(india_map_df) +
  geom_polygon(aes(x = long, y = lat, group = group), fill = "White", color = "black") +
  geom_point(data = pollution_long_after_lockdown,
             aes(x = longitude, y = latitude, color = Pollution_Range), size = 3) +
  scale_color_manual(name = "Pollution Level",
                     values = c("Low" = "Green", "Medium" = "Yellow", "High" = "Red")) +
  theme_void() +
  coord_fixed(ratio = 1.19)+
  ggtitle("Pollution Levels Across India After Lockdown")
```



We can see from the graph that, the pollution level in many cities have reduced after lockdown

Reference for Pollutant values:

- https://www.iqair.com/th-en/india
- https://pib.gov.in/newsite/printrelease.aspx?relid=110654