Indian Air Pollution Analysis (1987-2015)

BY

Karthikeyan K

22CSEG15

I Msc Data Analytics

Data source: data.gov.in

ASSUMPTIONS:

- The grouping of certain types of areas (e.g. "Industrial" and "Industrial Areas") into a single category is appropriate and does not significantly impact the overall analysis.
- The analysis is focused on the state of Delhi, assuming that it is a high-pollution area.
- The code assumes that the data is representative of the entire population and can be used to draw conclusions about the state of pollution in different states

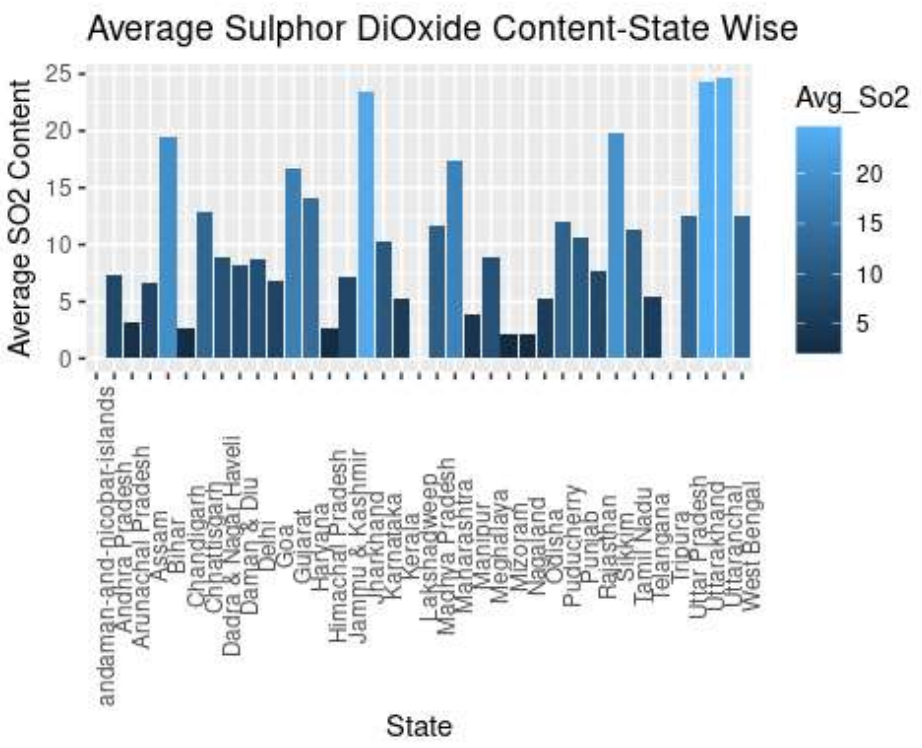
SOURCE CODE:

```
library(readr)
library(ggplot2)
library(dplyr)library(tidyr)
library(lubridate)

air <- read_csv("data.csv")

## Rows: 435742 Columns: 13

## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): stn_code, sampling_date, state, location, agency, type, location_m...
## dbl  (5): so2, no2, rspm, spm, pm2_5
## date (1): date
##
## ℹ Use `spec()` to retrieve the full column specification for this data.

colSums(is.na(air))

##                     stn_code                 sampling_date
##                       144077                             3
##                        state                      location
##                            0                             3
##                       agency                          type
##                       149481                          5393
##                          so2                           no2
##                        34646                         16233
##                         rspm                           spm
##                        40222                        237387
## location_monitoring_station                          pm2_5
##                        27491                        426428
##                         date
##                            7

#dropping unwanted columns
air <- air[, -c(1,2,5,11,12)]

air$date <-as.Date(air$date,'%Y-%m-%d')
summary(air)

##     state              location            type                so2
##  Length:435742      Length:435742      Length:435742      Min.   :  0.00
##  Class :character   Class :character   Class :character   1st Qu.:  5.00
##  Mode  :character   Mode  :character   Mode  :character   Median :  8.00
##                                                           Mean   : 10.83
##                                                           3rd Qu.: 13.70
##                                                           Max.   :909.00
##                                                           NA's   :34646
##       no2               rspm              spm               date
##  Min.   :  0.00   Min.   :   0.0   Min.   :   0.0   Min.   :1987-01-01
##  1st Qu.: 14.00   1st Qu.:  56.0   1st Qu.: 111.0   1st Qu.:2007-07-03
##  Median : 22.00   Median :  90.0   Median : 187.0   Median :2010-11-12
##  Mean   : 25.81   Mean   : 108.8   Mean   : 220.8   Mean   :2010-01-11
##  3rd Qu.: 32.20   3rd Qu.: 142.0   3rd Qu.: 296.0   3rd Qu.:2013-09-07
```

```
##  Max.   :876.00   Max.   :6307.0   Max.   :3380.0   Max.   :2015-12-31
##  NA's   :16233    NA's   :40222    NA's   :237387   NA's   :7
```

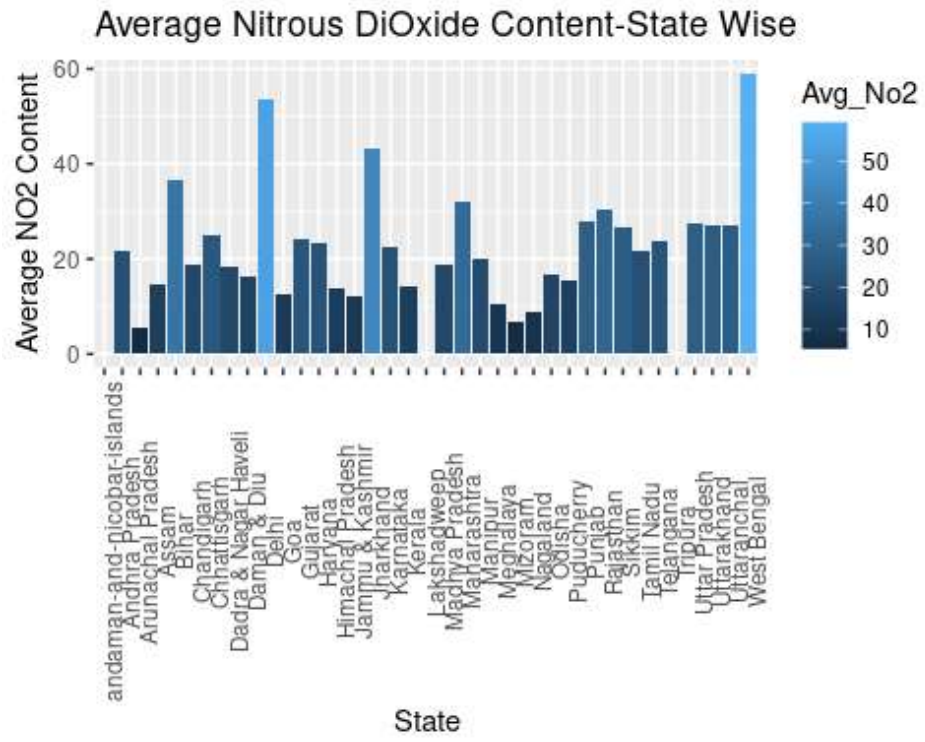
```
#replace null value by mean
#air["so2"][is.na(air['so2'])] = mean(air$so2, na.rm = TRUE)
#air["no2"][is.na(air['no2'])] = mean(air$no2, na.rm = TRUE)
#air["rspm"][is.na(air['rspm'])] = mean(air$rspm, na.rm = TRUE)
#air["spm"][is.na(air['spm'])] = mean(air$spm, na.rm = TRUE)

#some Data cleanup
air$type[air$type=="Sensitive Areas"] <-"Sensitive Area"
air$type[air$type %in% c("Industrial","Industrial Areas")] <-"Industrial Area"
air$type[air$type %in% c("Residential")] <-"Residential and others"


#Due to High Pollution in Delhi we have to analyze them
by_state_wise <-air%>%group_by(state)%>%summarise(Avg_So2=mean(so2,na.rm=TRUE),
                                                  Avg_No2=mean(no2,na.rm=TRUE),
                                                  Avg_Rspm=mean(rspm,na.rm=TRUE),
                                                  Avg_Spm= mean(spm,na.rm=TRUE))


ggplot(by_state_wise,aes(x=state,y=Avg_So2,fill=Avg_So2)) +
  geom_bar(stat="identity") +
  theme(axis.text.x =element_text(angle=90)) +
  ggtitle("Average Sulphor DiOxide Content-State Wise") +
  xlab(label="State") +
  ylab(label="Average SO2 Content")
```

```
## Warning: Removed 3 rows containing missing values (`position_stack()`).
```



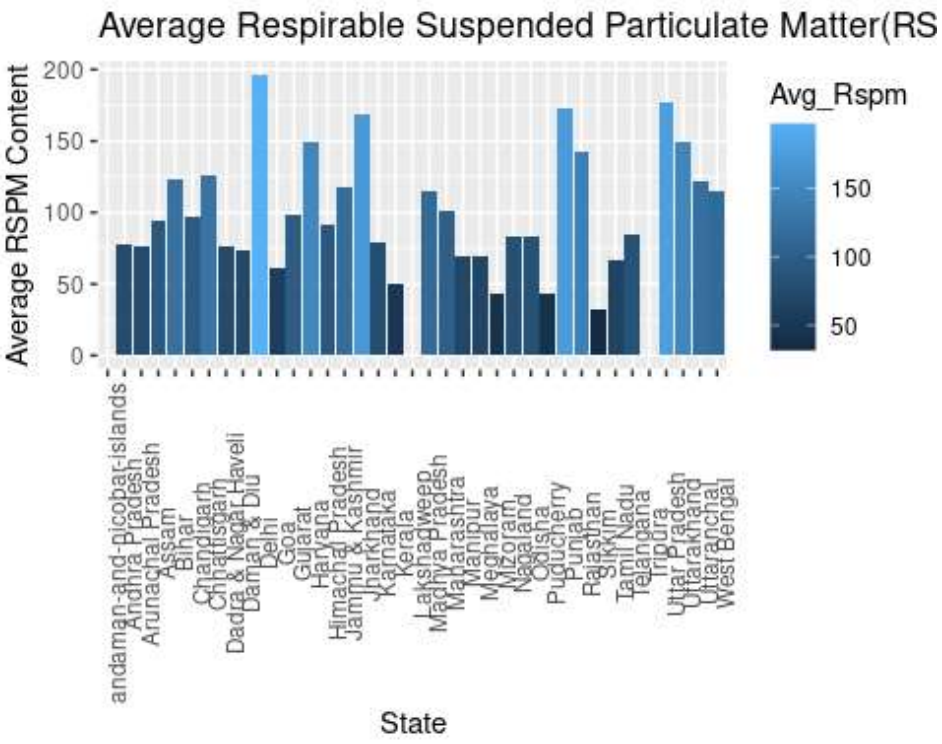
```
ggplot(by_state_wise,aes(x=state,y=Avg_No2,fill=Avg_No2)) +
  geom_bar(stat="identity") +
  theme(axis.text.x =element_text(angle=90)) +
  ggtitle("Average Nitrous DiOxide Content-State Wise") +
  xlab(label="State") +
  ylab(label="Average NO2 Content")
```

```
## Warning: Removed 3 rows containing missing values (`position_stack()`).
```



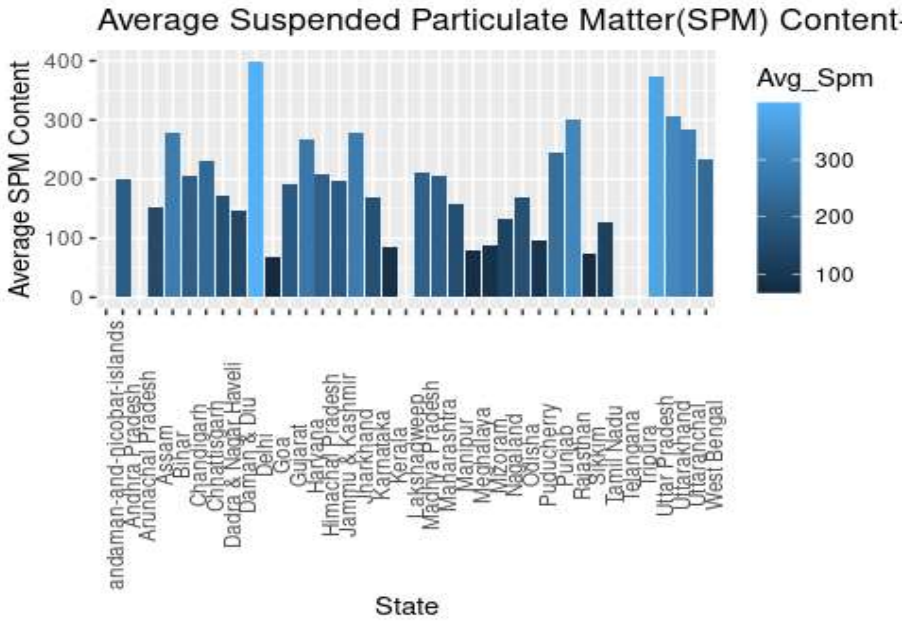

```
ggplot(by_state_wise,aes(x=state,y=Avg_Rspm,fill=Avg_Rspm)) +
  geom_bar(stat="identity") +
  theme(axis.text.x =element_text(angle=90)) +
  ggtitle("Average Respirable Suspended Particulate Matter(RSPM) Content-State Wise") +
  xlab(label="State") +
  ylab(label="Average RSPM Content")
```

```
## Warning: Removed 3 rows containing missing values (`position_stack()`).
```




```
ggplot(by_state_wise,aes(x=state,y=Avg_Spm,fill=Avg_Spm)) +
  geom_bar(stat="identity") +
  theme(axis.text.x =element_text(angle=90)) +
  ggtitle("Average Suspended Particulate Matter(SPM) Content-State Wise") +
  xlab(label="State") +
  ylab(label="Average SPM Content")
```

```
## Warning: Removed 5 rows containing missing values (`position_stack()`).
```




```
#Lets investigate more on Delhi Trend w.r.t pollution
air$date <-as.POSIXct(air$date)
air$year <-year(air$date)
```

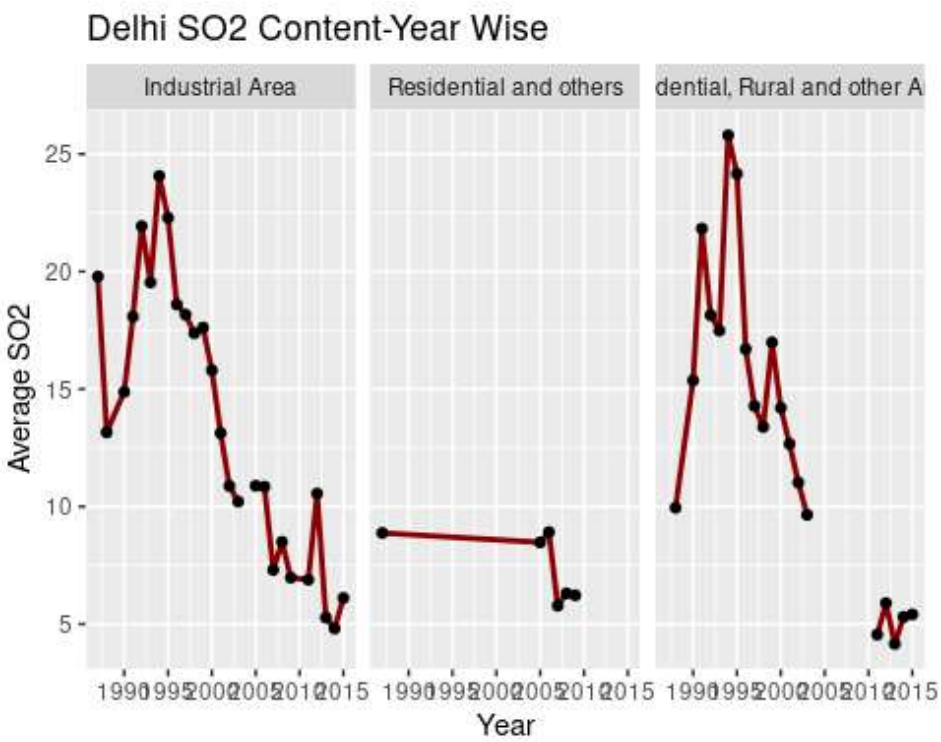
```
Delhi <-
air%>%filter(state=="Delhi")%>%group_by(year,type)%>%summarise(Avg_So2=mean(so2,na.rm=TRU
E),

Avg_No2=mean(no2,na.rm=TRUE),

Avg_Rspm=mean(rspm,na.rm=TRUE),
                                                                    Avg_Spm
=mean(spm,na.rm=TRUE))
```

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

```
ggplot(Delhi,aes(x=year,y=Avg_So2)) +
  geom_line(size=1,color="darkred") +
  geom_point()+
  facet_wrap(~type) +
  ggtitle("Delhi SO2 Content-Year Wise")+
  xlab("Year") +
  ylab("Average SO2")
```

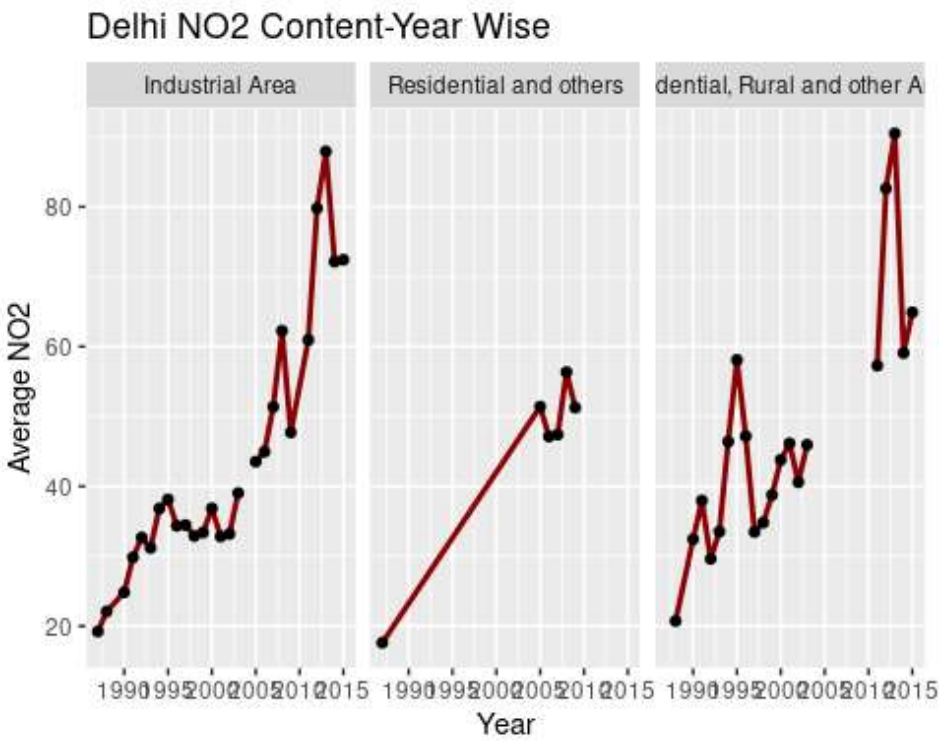
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```




```
ggplot(Delhi,aes(x=year,y=Avg_No2)) +
  geom_line(size=1,color="darkred") +
  geom_point()+
  facet_wrap(~type) +
  ggtitle("Delhi NO2 Content-Year Wise")+
  xlab("Year") +
  ylab("Average NO2")
```

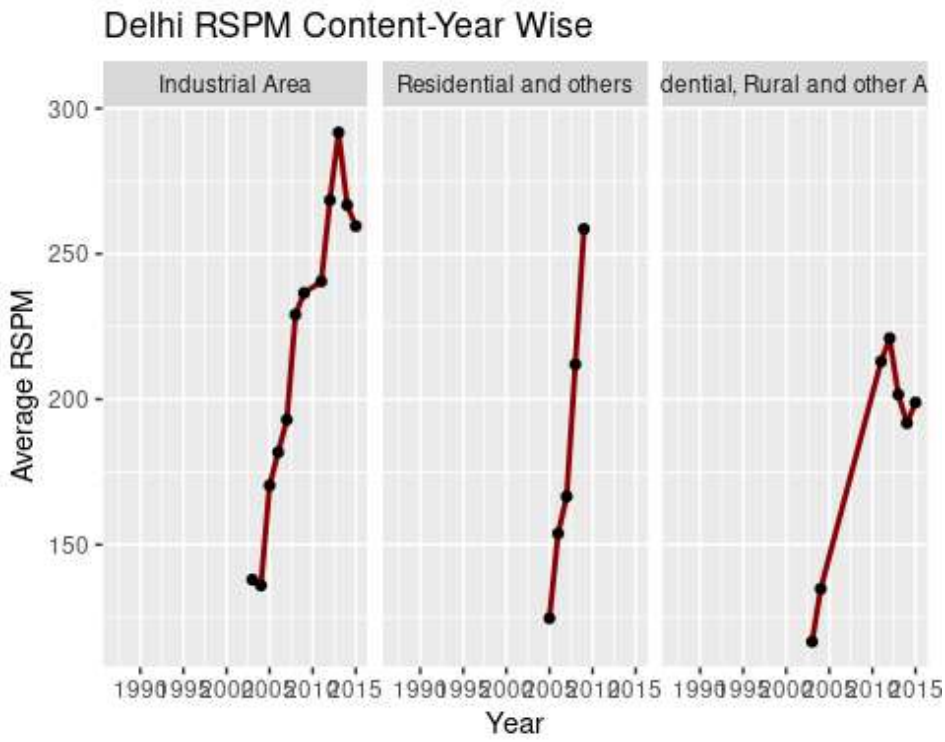
```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```




```
ggplot(Delhi,aes(x=year,y=Avg_Rspm)) +
  geom_line(size=1,color="darkred") +
  geom_point()+
  facet_wrap(~type) +
  ggtitle("Delhi RSPM Content-Year Wise")+
  xlab("Year") +
  ylab("Average RSPM")
```

```
## Warning: Removed 15 rows containing missing values (`geom_line()`).
```

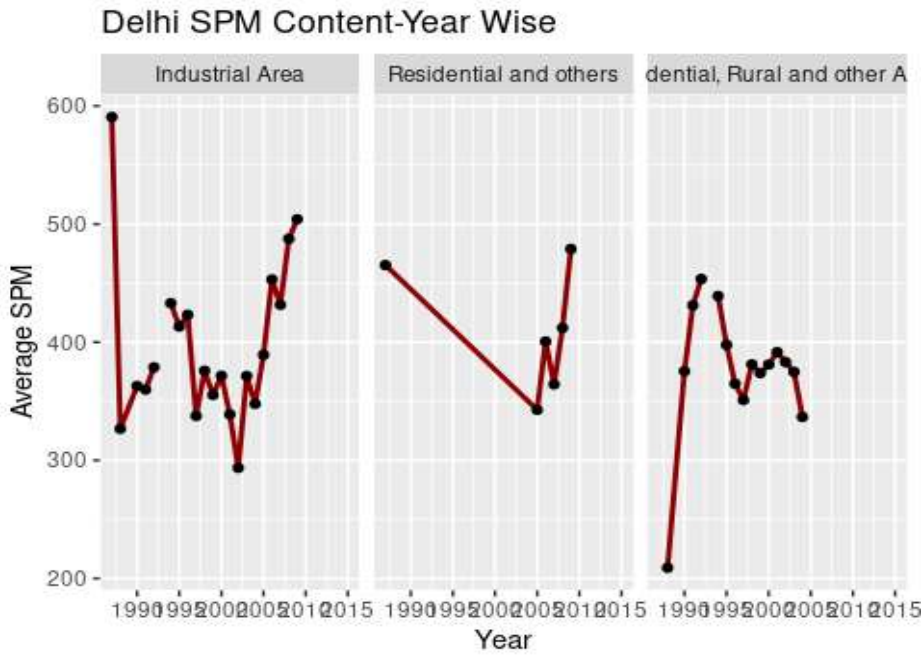
```
## Warning: Removed 30 rows containing missing values (`geom_point()`).
```




```
ggplot(Delhi,aes(x=year,y=Avg_Spm)) +
  geom_line(size=1,color="darkred") +
  geom_point()+
  facet_wrap(~type) +
  ggtitle("Delhi SPM Content-Year Wise")+
  xlab("Year") +
  ylab("Average SPM")
```

```
## Warning: Removed 5 rows containing missing values (`geom_line()`).
```

```
## Warning: Removed 12 rows containing missing values (`geom_point()`).
```



INSIGHTS:

- Delhi is the most polluted state in India with respect to Respirable Suspended Particulate Matter (RSPM) and Suspended Particulate Matter (SPM). Uttar Pradesh ranks second.
- Meghalaya and Mizoram are the least polluted states in India with respect to RSPM and SPM.
- Uttarakhand (now known as Uttarakhand) and Uttaranchal have the highest Sulphur content.
- West Bengal and Delhi rank first and second in Nitrous Oxide content.
- The analysis also shows that there is a significant variation in the pollution level of Delhi over the years, with some years showing a higher pollution level than others.
- The analysis shows that the average levels of SO2 and NO2 are highest in industrial areas, whereas the average levels of RSPM and SPM are highest in residential and sensitive areas.

INFERENCE:

- The analysis shows that Delhi has higher pollution levels than other states in India. The average content of SO2, NO2, RSPM, and SPM in Delhi is higher than the national average.
- Among the different types of areas, industrial areas have the highest pollution levels followed by sensitive areas and residential areas.
- The trend analysis of Delhi shows that there has been a gradual reduction in the pollution levels of SO2, NO2, and RSPM since 2010. However, the levels of SPM have been fluctuating with no clear trend.
- It is important for the government and other stakeholders to take action to reduce pollution levels, especially in industrial areas and sensitive areas to ensure the health and well-being of the citizens.