AirQualityIndex

Abir Patnaik

12/02/2020

Air Quality Data Analysis

INTRODUCTION

According to latest stats, air pollution has been the third biggest cause of the death of humans.A few weeks back only Delhi had the worst AQI for which Supreme Court had to direct the Punjab and Haryana government to shut down the stubble burning that is adding on to the exisiting problems.Even the odd-even scheme didn't provide much solution.

For this, a statistical analysis on the air quality in india would help to understand the situation a little better.The Database is collected from Ministry of Environment and Forests and Central Pollution Control Board of India.

This data contained the sulphur dioxide(So2),Nitrous Dioxide(No2),Rspm , Spm and 2.5 MM particle has been collected for more than 15 yrs from 1987 to 2015 for all the regions of India.

Because the database contains a lot of data points a thorough analysis needed to be done by breaking into parts and analysing the data(divide and conquer approach).

Since the data is high, their are chances the data might have NA values which may not help in our analysis.For this, the dataset was checked for NA values and it was added with median values which will not affect the values as such.This approach was used because of the following reasons:

1.By checking the data, a lot of data points were missing.If data points are deleted, a lot of statistical analysis may result in false results.If the case had been minimal data points are there, then deleting the records wouldn't have been such a issue.

2.Replacing the NA values with mean is also not suggested option because take the following example: suppose the records are 1,2,2,NA,NA,100 Now we replace the NA with mean it would come approx 50 but median would be at most 2 since in here 100 is outlier and may not be representative of the whole data for which median is a suitable option.

Below code replaces the NA values with the median values.colSums first calculates the no. of NA values.

```
##                  ï..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
```

```
##   ï..stn_code        sampling_date               state
##  193    :  1428   19-03-2015:   253   Maharashtra   : 60384
##  519    :  1280   12-02-2015:   237   Uttar Pradesh : 42816
##  708    :  1273   19-02-2015:   236   Andhra Pradesh: 26368
##  541    :  1270   05-11-2015:   235   Punjab        : 25634
##  710    :  1269   11-11-2015:   234   Rajasthan     : 25589
##  (Other):285145   (Other)   :434544   Kerala        : 24728
##  NA's   :144077   NA's      :     3   (Other)       :230223
##        location
##  Guwahati  :  9984
##  Hyderabad :  9667
##  Delhi     :  8551
##  Chandigarh:  8520
##  Jaipur    :  7850
##  (Other)   :391167
##  NA's      :     3
##                                                                    agency
##  Maharashtra State Pollution Control Board                            : 27857
##  Uttar Pradesh State Pollution Control Board                          : 22686
##  Andhra Pradesh State Pollution Control Board                         : 19139
##  Himachal Pradesh State Environment Proection & Pollution Control Board: 15287
##  Punjab State Pollution Control Board                                 : 15232
##  (Other)                                                              :186060
##  NA's                                                                 :149481
##                                  type             so2              no2
##  Industrial Area                   :148071   Min.   :  0.00   Min.   :  0.00
##  Residential                       :   158   1st Qu.:  5.00   1st Qu.: 14.00
##  Residential and others            : 86791   Median :  8.00   Median : 22.00
##  Residential, Rural and other Areas:179014   Mean   : 10.83   Mean   : 25.81
##  RIRUO                             :  1304   3rd Qu.: 13.70   3rd Qu.: 32.20
##  Sensitive Area                    : 15011   Max.   :909.00   Max.   :876.00
##  NA's                              :  5393   NA's   :34646    NA's   :16233
##       rspm             spm
##  Min.   :   0.0   Min.   :   0.0
##  1st Qu.:  56.0   1st Qu.: 111.0
##  Median :  90.0   Median : 187.0
##  Mean   : 108.8   Mean   : 220.8
##  3rd Qu.: 142.0   3rd Qu.: 296.0
##  Max.   :6307.0   Max.   :3380.0
##  NA's   :40222    NA's   :237387
##                                                location_monitoring_station
##  Regional Office                                                    :  6261
##  Paonta Sahib                                                       :  1599
##  Head Office, Bamunimaidan, Guwahati                                :  1327
##  ITI Building, Gopinath Nagar, Guwahati                             :  1280
##  Bank of Baroda Building, Near Pimpri-Chinchwad M.C. Building:  1273
##  (Other)                                                            :396511
##  NA's                                                               : 27491
##      pm2_5                date
##  Min.   :  3.0    19-03-2015:   253
##  1st Qu.: 24.0    12-02-2015:   237
##  Median : 32.0    19-02-2015:   236
##  Mean   : 40.8    05-11-2015:   235
##  3rd Qu.: 46.0    11-11-2015:   234
##  Max.   :504.0    (Other)   :434540
##  NA's   :426428   NA's      :     7
```

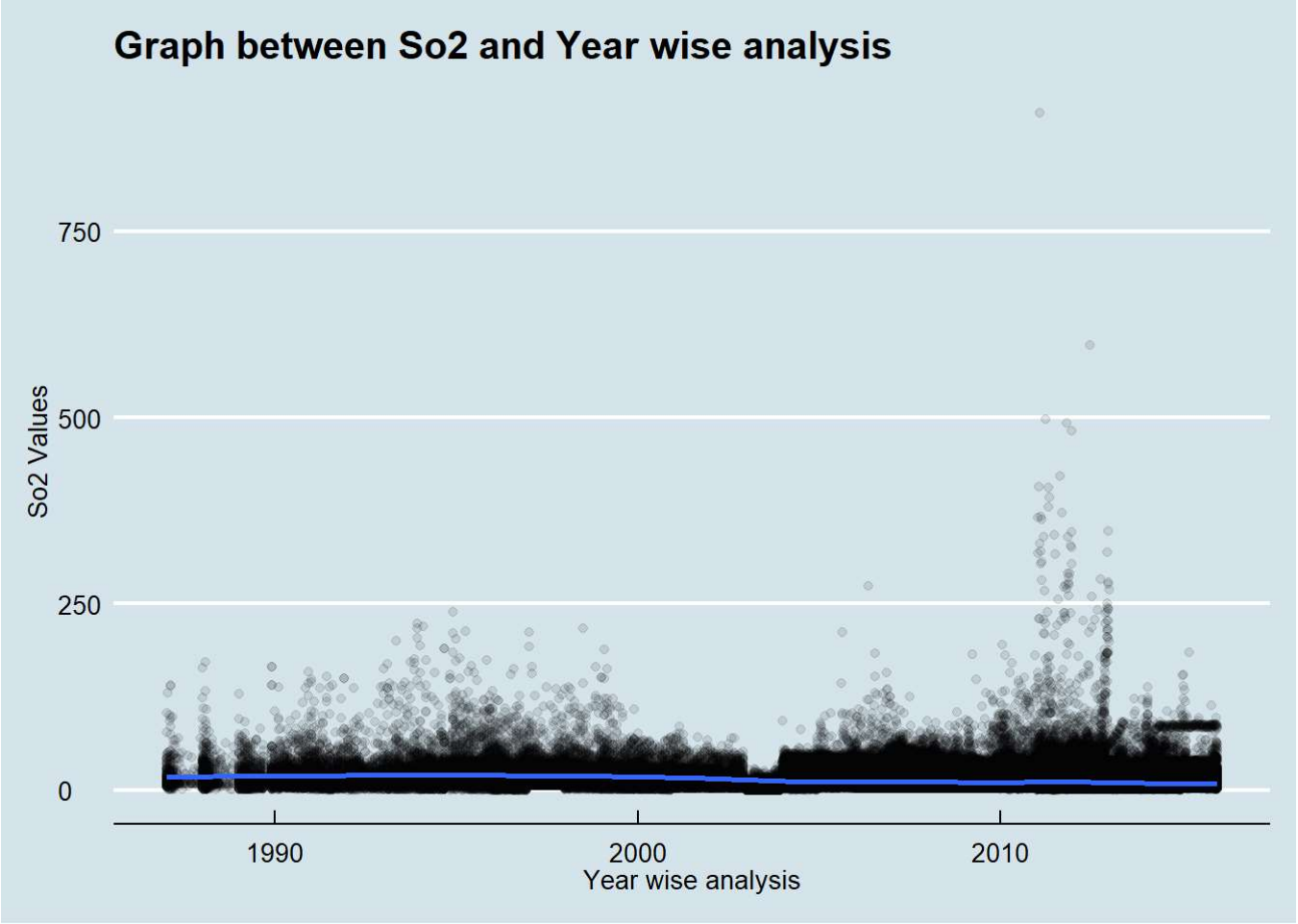
Bivariate Analysis

In the initial analysis, since the data points is high with all the states in India added in the dataset with more than 15 years of data a starter analysis is required for which all the numerical variables i.e. So2,No2,Rspm ,Spm and 2.5 mm particles analysis was done over the years.

Below is the graph plotted for changes in so2 levels over the years.

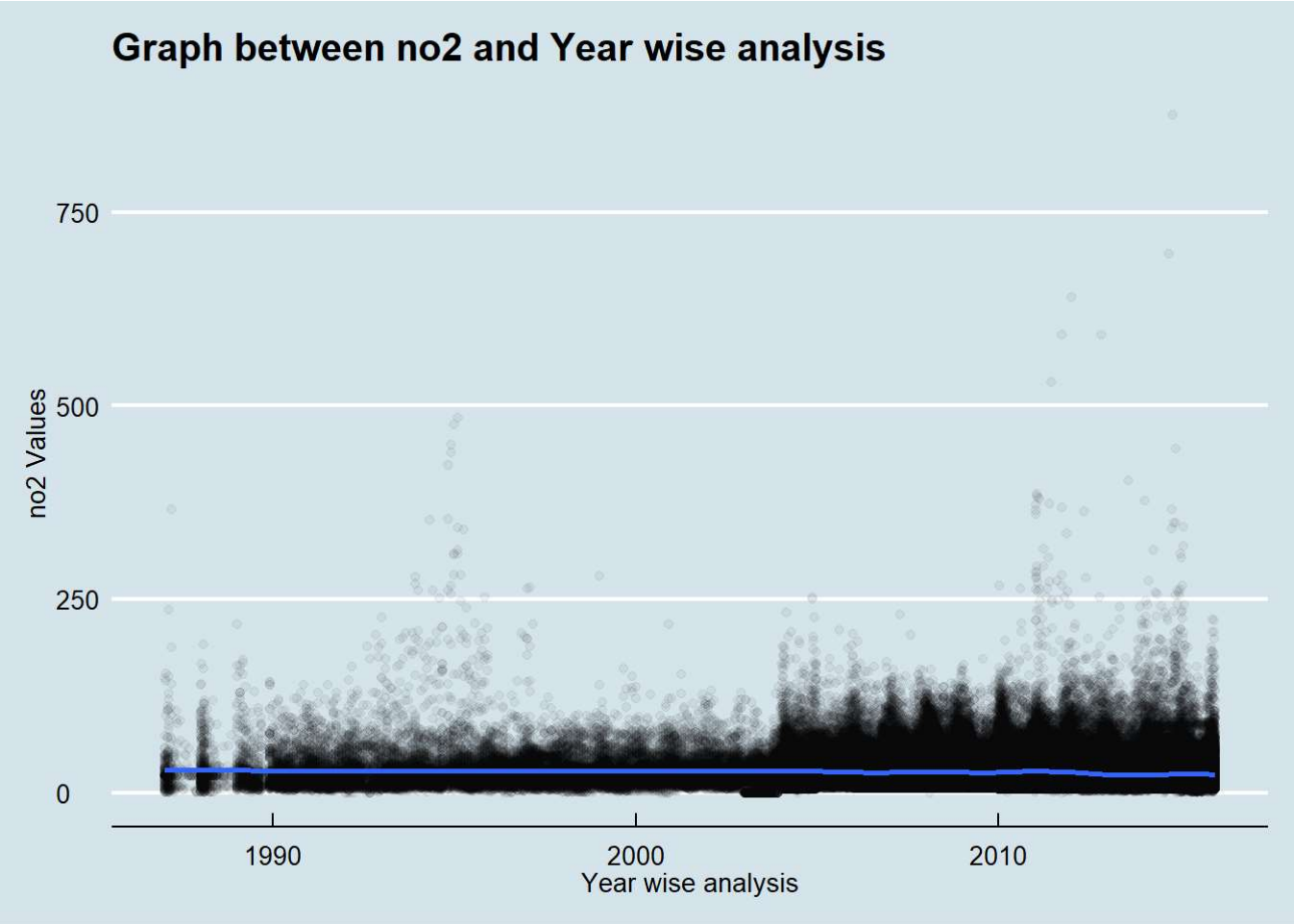
As in the graph plotted,most of the values are between 0 to 250 apart from few outliers present mainly after the year 2010.

```
ggplot(air_quality_data,aes(x=air_quality_data$date,y=air_quality_data$so2))+geom_point(alpha
=1/10)+
  geom_smooth()+
  theme_economist()+
xlab('Year wise analysis')+
  ylab('So2 Values')+
  labs(title='Graph between So2 and Year wise analysis')
```



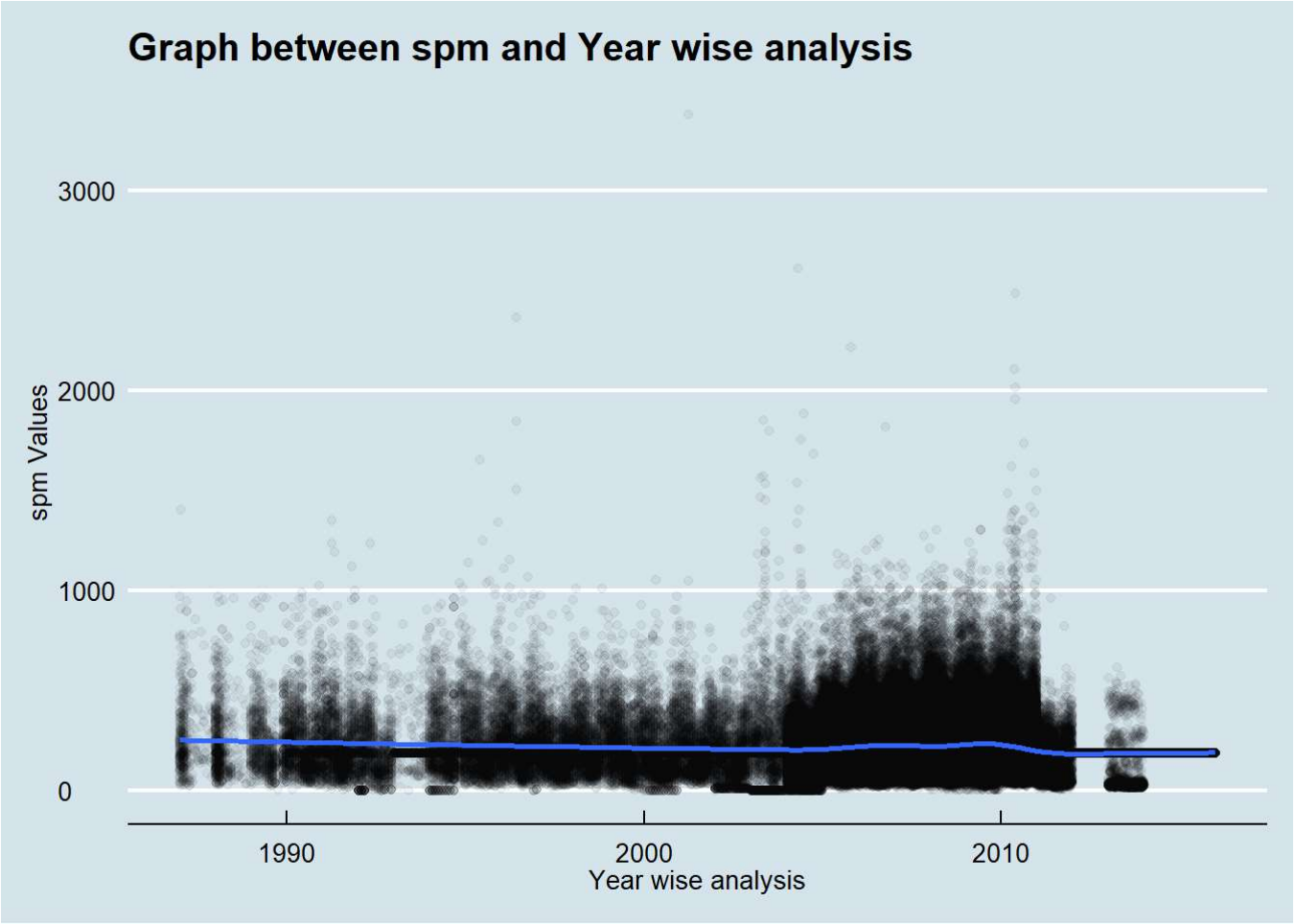
Below is the scatter plot for No2 values rise over the years.The same situation is for the no2 level rises.The years 2013-2015 had a slight rise with the amount of data points and between the 1995 and 1996 rise of levels of No2 were high.Since this includes all the states the results may vary because of that.

```
ggplot(air_quality_data,aes(x=air_quality_data$date,y=air_quality_data$no2))+geom_point(alpha
=1/20)+
  theme_economist()+
  geom_smooth()+
xlab('Year wise analysis')+
  ylab('no2 Values')+
  labs(title='Graph between no2 and Year wise analysis')
```



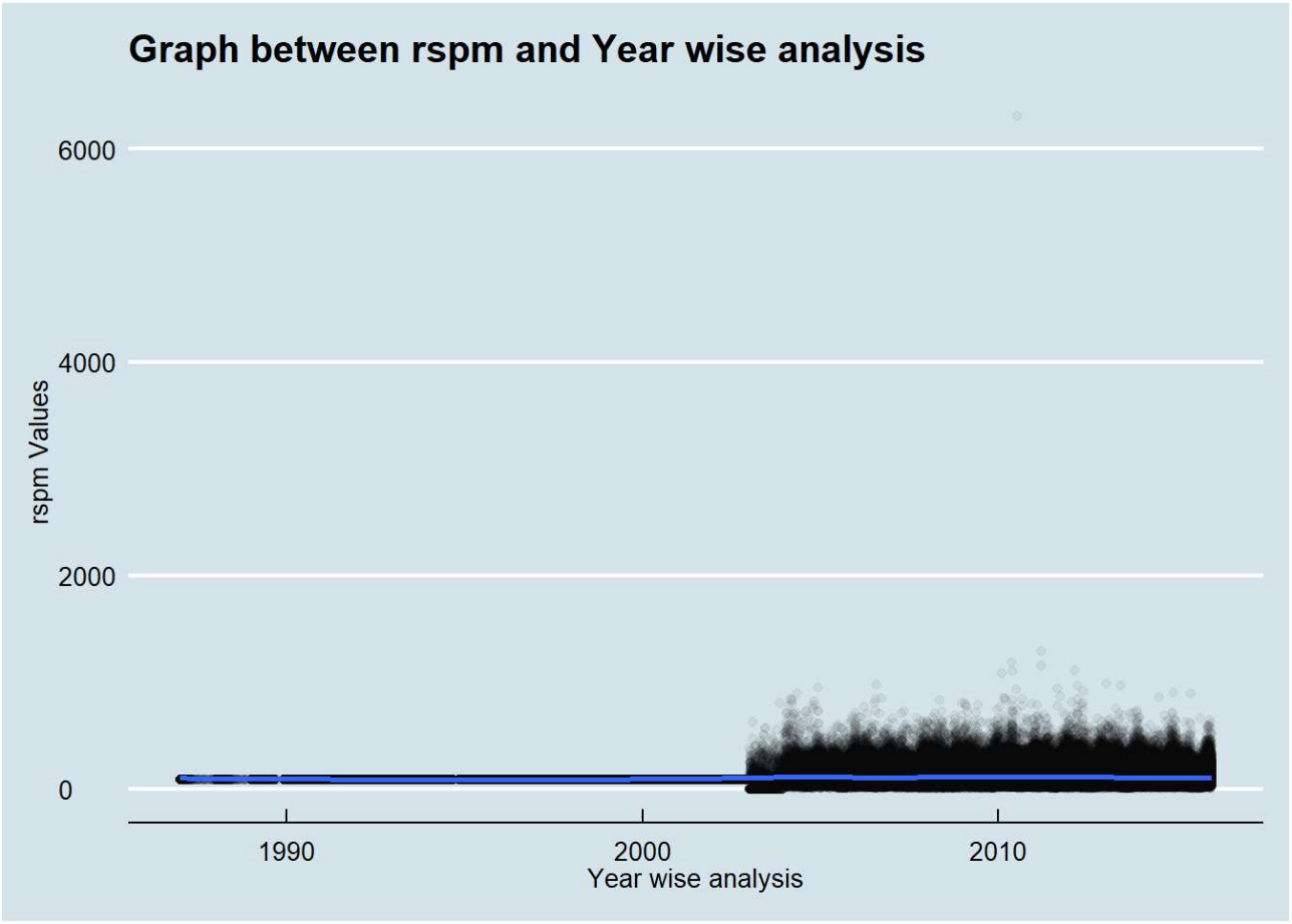
There is a gradual rise in the spm levels between the years of 2007 and 2010 .

```
ggplot(air_quality_data,aes(x=air_quality_data$date,y=air_quality_data$spm))+geom_point(alpha
=1/20)+
  geom_smooth()+theme_economist()+
xlab('Year wise analysis')+
  ylab('spm Values')+
  labs(title='Graph between spm and Year wise analysis')
```



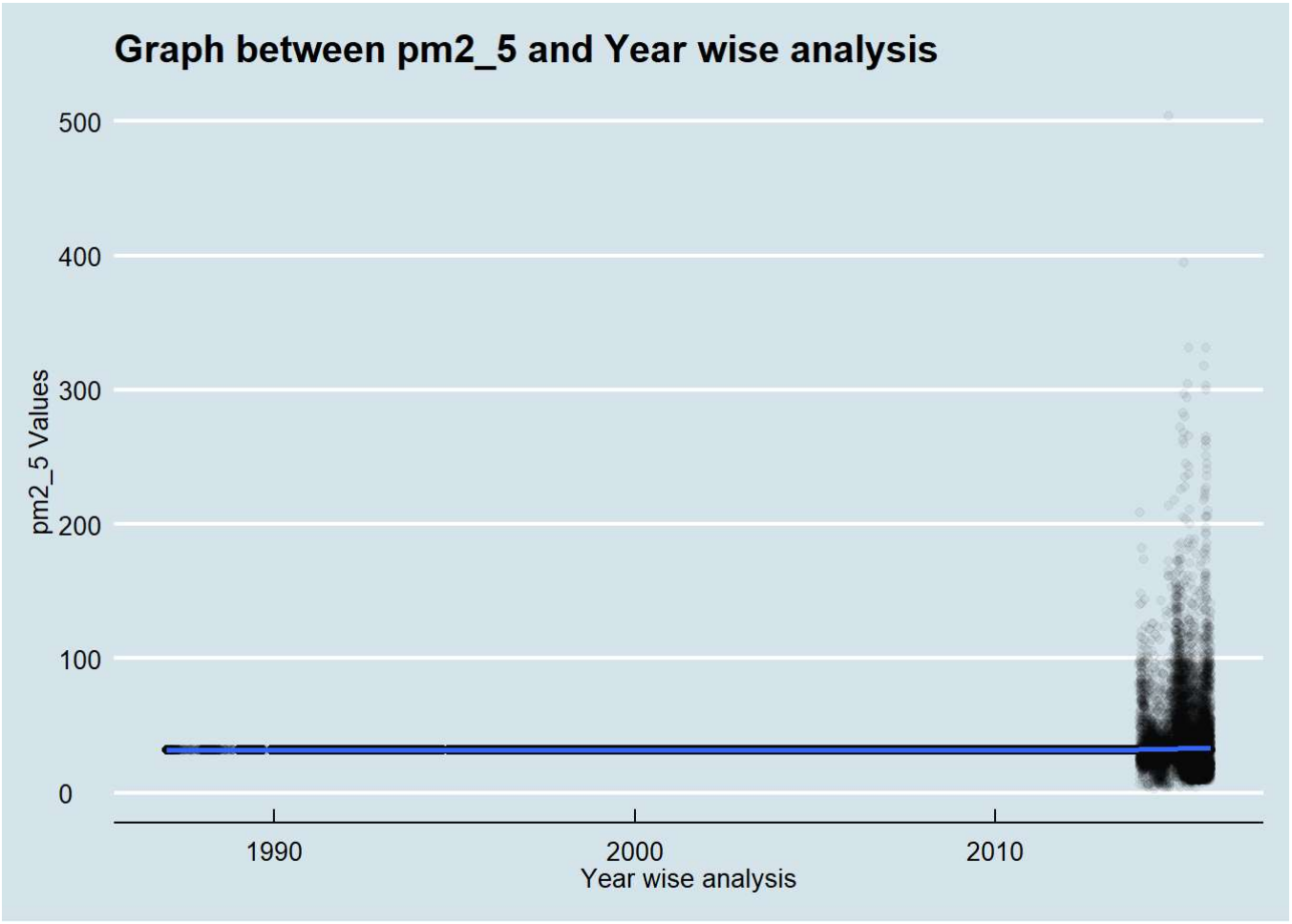
Since data was replaced from NA values to median values, the missing data is quite evident in the graph and the results can be seen only from 2003.DUe to lesser amount of data as compared to others, analysis is difficult for this.

```
ggplot(air_quality_data,aes(x=air_quality_data$date,y=air_quality_data$rspm))+geom_point(alph
a=1/20)+
  theme_economist()+geom_smooth()+xlab('Year wise analysis')+
  ylab('rspm Values')+
  labs(title='Graph between rspm and Year wise analysis')
```



Similar situation was found for this.Data arrived after 2012 only which may not help the trend but can give the overview.Optimum values need to be known for a more deeper analysis i.e. at what level the pm levels are bad.

```
ggplot(air_quality_data,aes(x=air_quality_data$date,y=air_quality_data$pm2_5))+geom_point(alp
ha=1/20)+geom_smooth()+theme_economist()+
  xlab('Year wise analysis')+
  ylab('pm2_5 Values')+
  labs(title='Graph between pm2_5 and Year wise analysis')
```



Since the data points were high and it was difficult to comprehend the results as depicted in the previous graphs, what one can do is divide the dataset into subsets based on the years and then do the analysis.

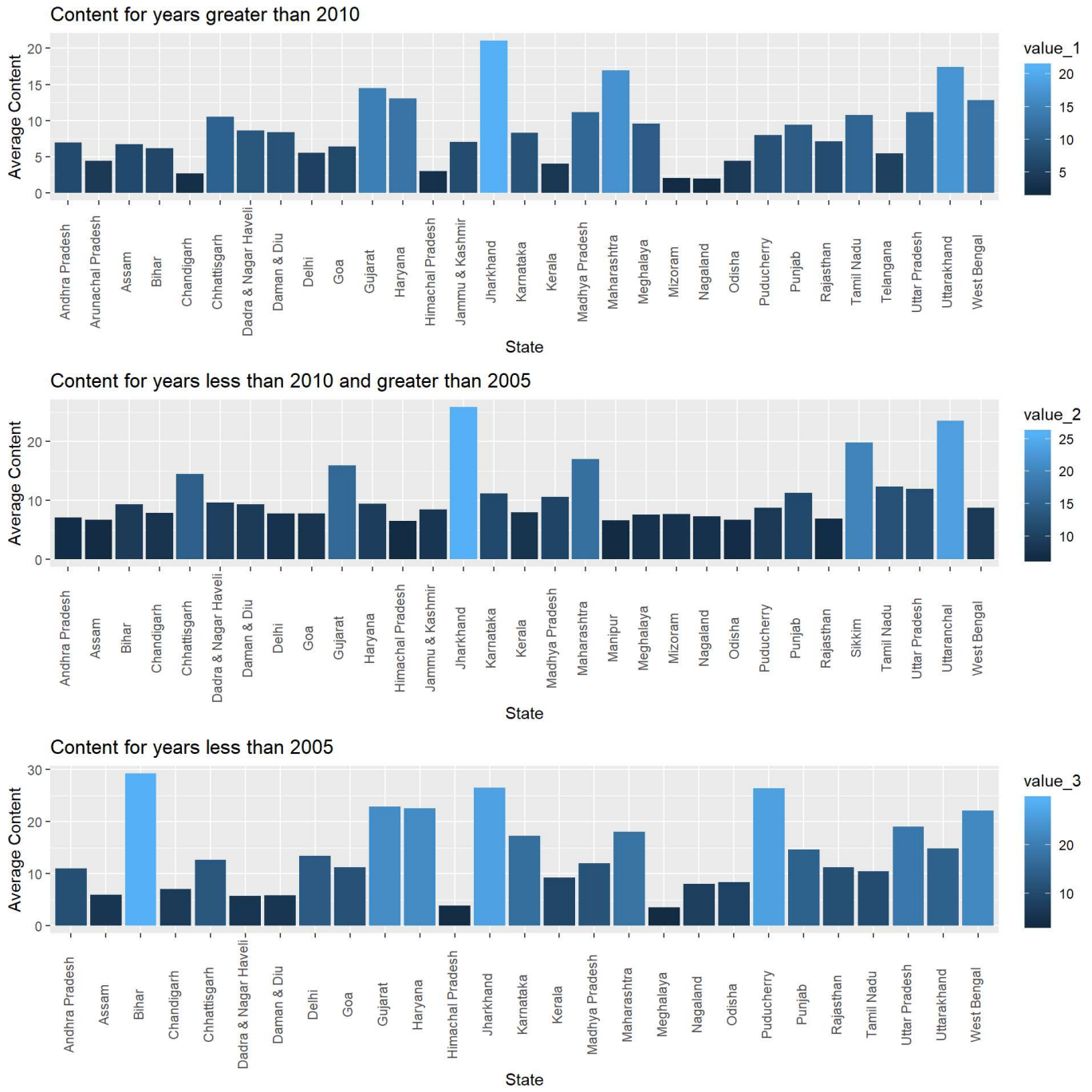
```
air_quality_data_last_5yrs <- subset(air_quality_data, air_quality_data$date > as.Date("2010-
01-01"))
air_quality_data_mid_5yrs <- subset(air_quality_data, air_quality_data$date > as.Date("2005-0
1-01") & air_quality_data$date < as.Date("2010-01-01"))
air_quality_data_first_5yrs <- subset(air_quality_data, air_quality_data$date < as.Date("2005
-01-01") )
```

Trend can be checked by grouping the data based on state and analysing it by checking average data over the years

```
summarise_data<-function(data_set){
  data_set%>%group_by(state) %>%summarise(Avg_So2=mean(so2),
                                          Avg_No2=mean(no2),
                                          Avg_Rspm=mean(rspm),
                                          Avg_Spm= mean(spm),
                                          AVG_pm2_5=mean(pm2_5))
}

Analysis_5_yrs_gap<-function(value_1,value_2,value_3){
  last_5_yrs_plot<-  ggplot(last_5_yrs,aes(x=state,y=value_1,fill=value_1)) +
    geom_bar(stat="identity") +
    theme(axis.text.x =element_text(angle=90,vjust=0.1)) +
    ggtitle("Content for years greater than 2010") +
    xlab(label="State") +
    ylab(label="Average Content")

  mid_5_yrs_plot<- ggplot(mid_5_yrs,aes(x=state,y=value_2,fill=value_2)) +
    geom_bar(stat="identity") +
    theme(axis.text.x =element_text(angle=90,vjust=0.1)) +
    ggtitle("Content for years less than 2010 and greater than 2005") +
    xlab(label="State") +
    ylab(label="Average Content")

  first_5_yrs_plot<-  ggplot(first_5_yrs,aes(x=state,y=value_3,fill=value_3)) +
    geom_bar(stat="identity") +
    theme(axis.text.x =element_text(angle=90,vjust=0.2)) +
    ggtitle("Content for years less than 2005") +
    xlab(label="State") +
    ylab(label="Average Content")
grid.arrange(last_5_yrs_plot
             ,mid_5_yrs_plot,
             first_5_yrs_plot,
             nrow=3)
}
last_5_yrs<-data.frame(summarise_data(air_quality_data_last_5yrs))
mid_5_yrs<-data.frame(summarise_data(air_quality_data_mid_5yrs))
first_5_yrs<-data.frame(summarise_data(air_quality_data_first_5yrs))
```

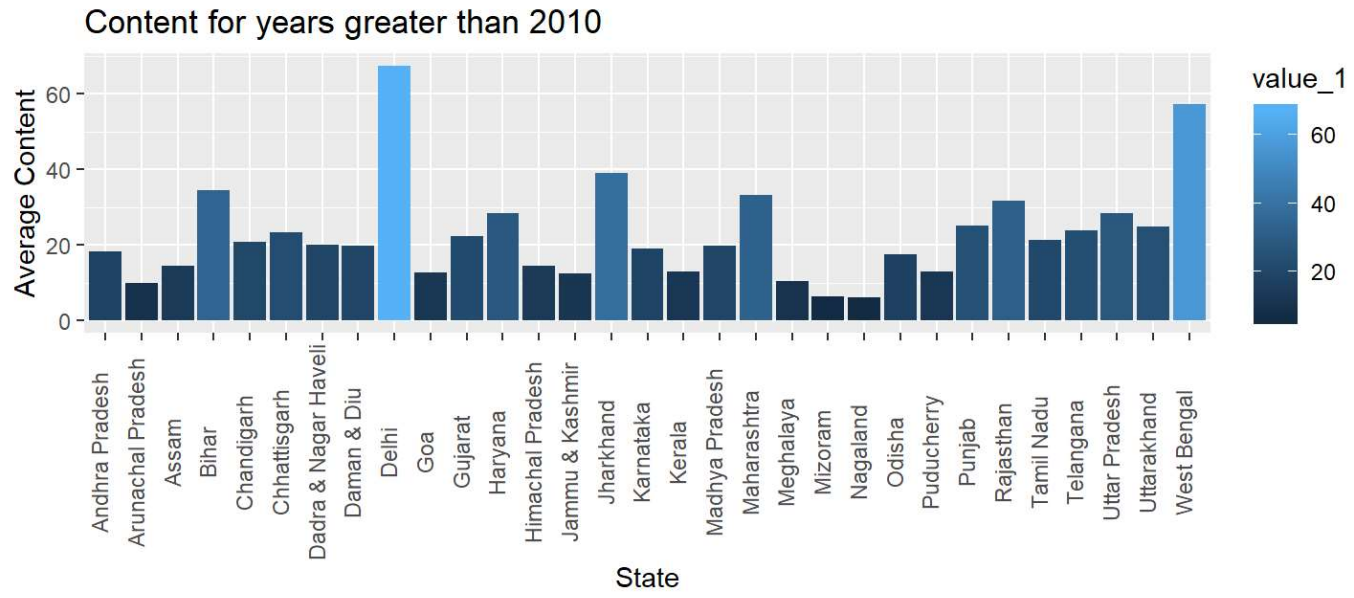
Below graph depicts changes with respect to So2 levels.Some key takeaways: 1.Jharkhand already has high amounts of SO2 mainly due to the coal mines for which rise is high as compared to other states. link (http://www.urbanemissions.info/india-apna/dhanbad-india/) 2.Haryana also has high level of So2 3.Bihar had high levels but it has decreased a lot. 4.West Bengal shows decrease in years 2005 to 2010 but it increased in 2010 to 2015

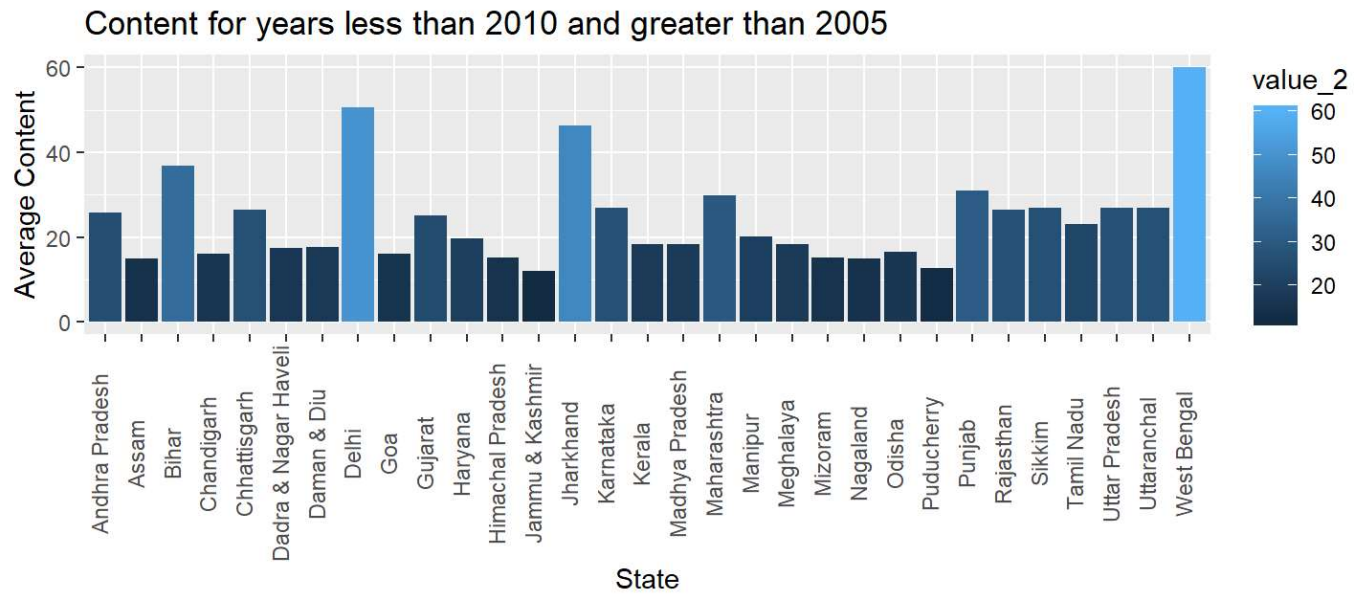
```
Analysis_5_yrs_gap(last_5_yrs$Avg_So2,mid_5_yrs$Avg_So2,first_5_yrs$Avg_So2)
```

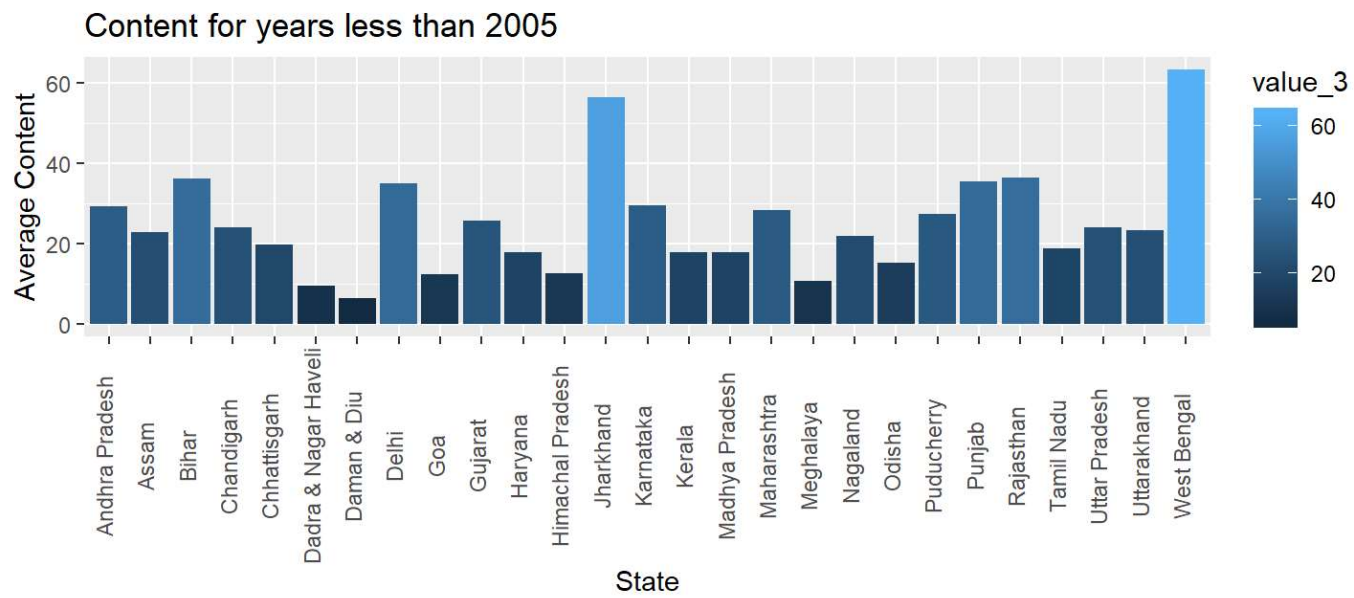


Below graph depicts changes with respect to No2 levels.Some key takeaways: 1.Delhi shows a high level of No2 over the years. 2.West Bengal already has high level of No2. link (https://timesofindia.indiatimes.com/city/kolkata/kolkata-gasps-as-no2-pollutants-rise-from-car-fumes-dust-coal-fuel/articleshow/58409054.cms) 3.Rest of the values do not have much change.

```
Analysis_5_yrs_gap(last_5_yrs$Avg_No2,mid_5_yrs$Avg_No2,first_5_yrs$Avg_No2)
```

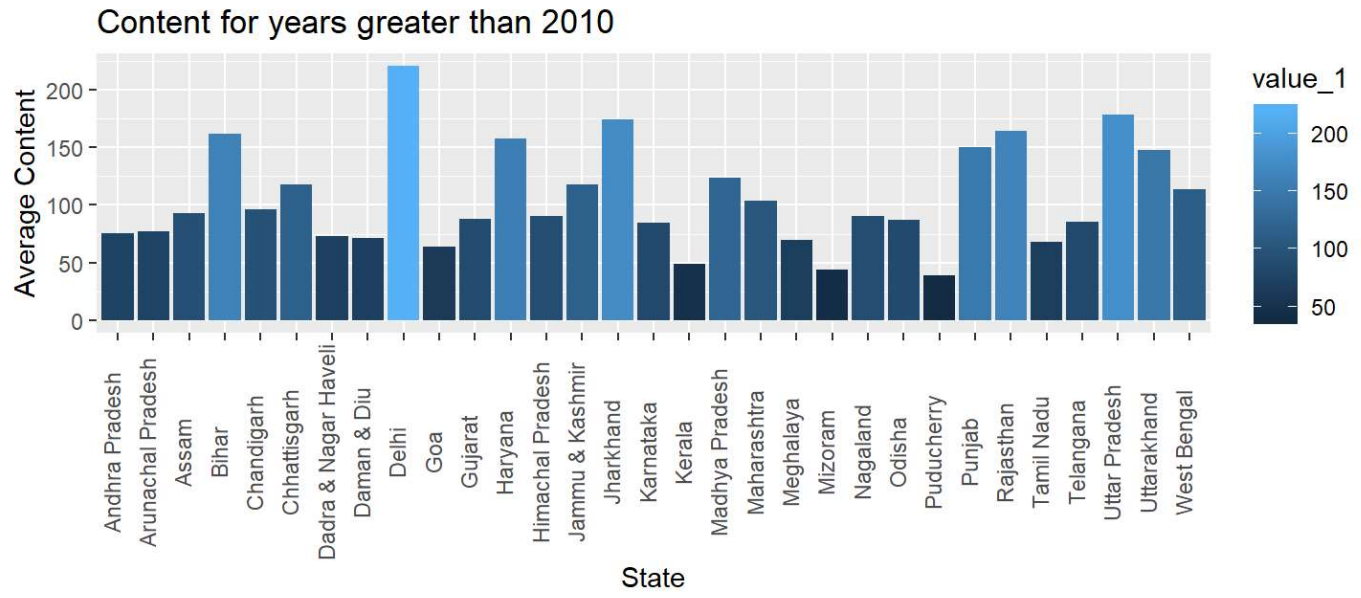


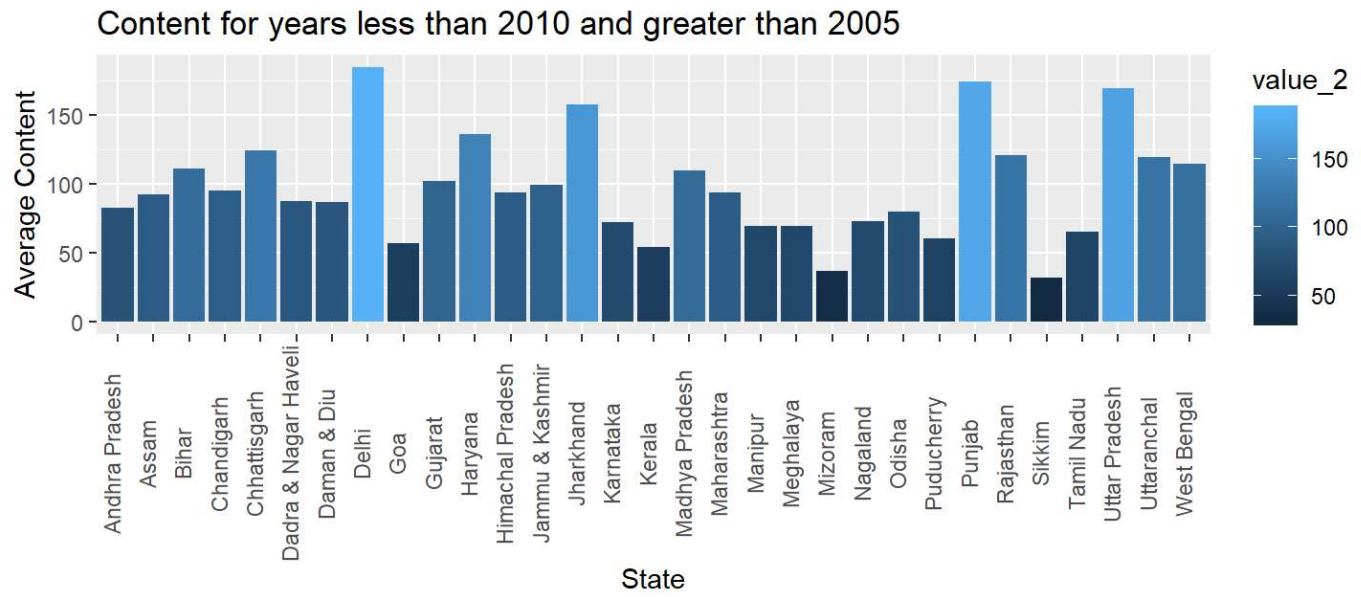


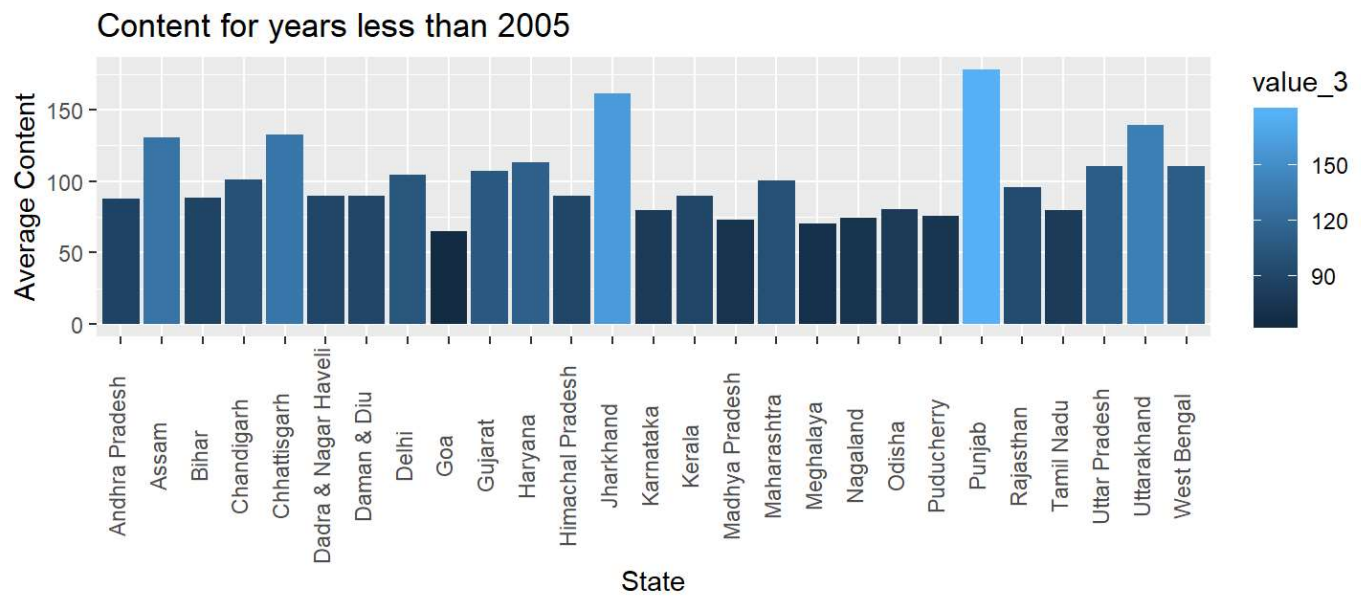


Below graph depicts changes with respect to Rspm levels.Some key takeaways: 1.Before 2010,Delhi has risen up with rspm levels 2.It can also be seen that overall levels of each of the states has decreased (e.g.Puducherry)

```
Analysis_5_yrs_gap(last_5_yrs$Avg_Rspm,mid_5_yrs$Avg_Rspm,first_5_yrs$Avg_Rspm)
```

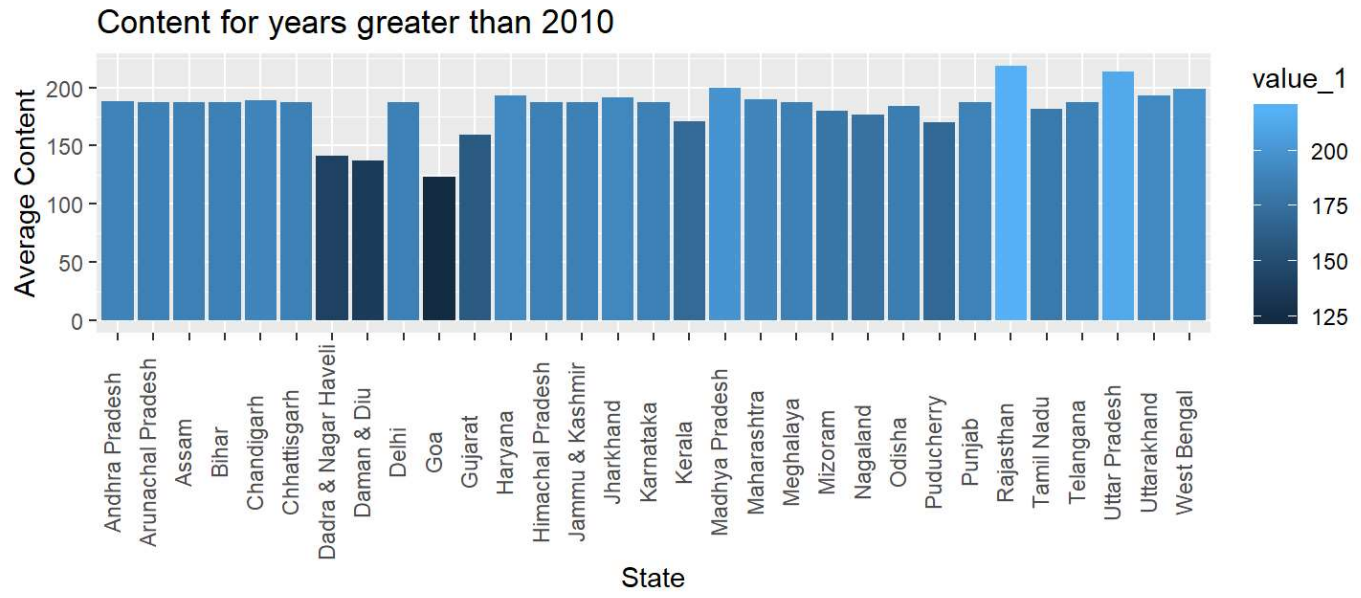


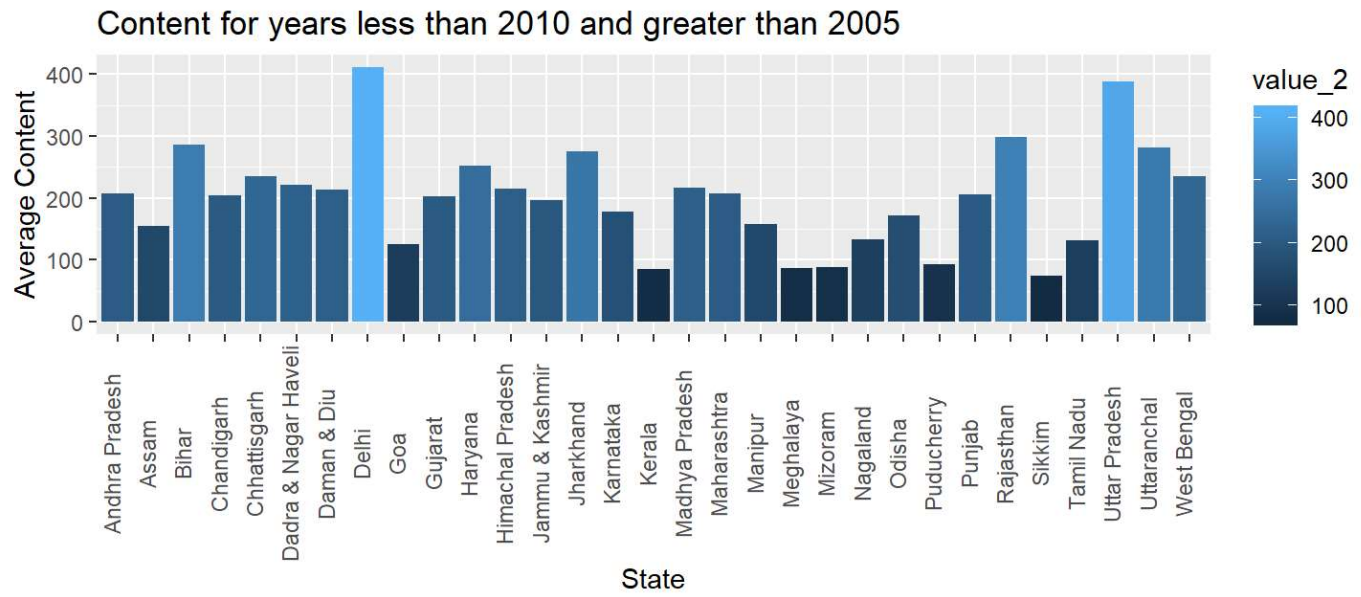


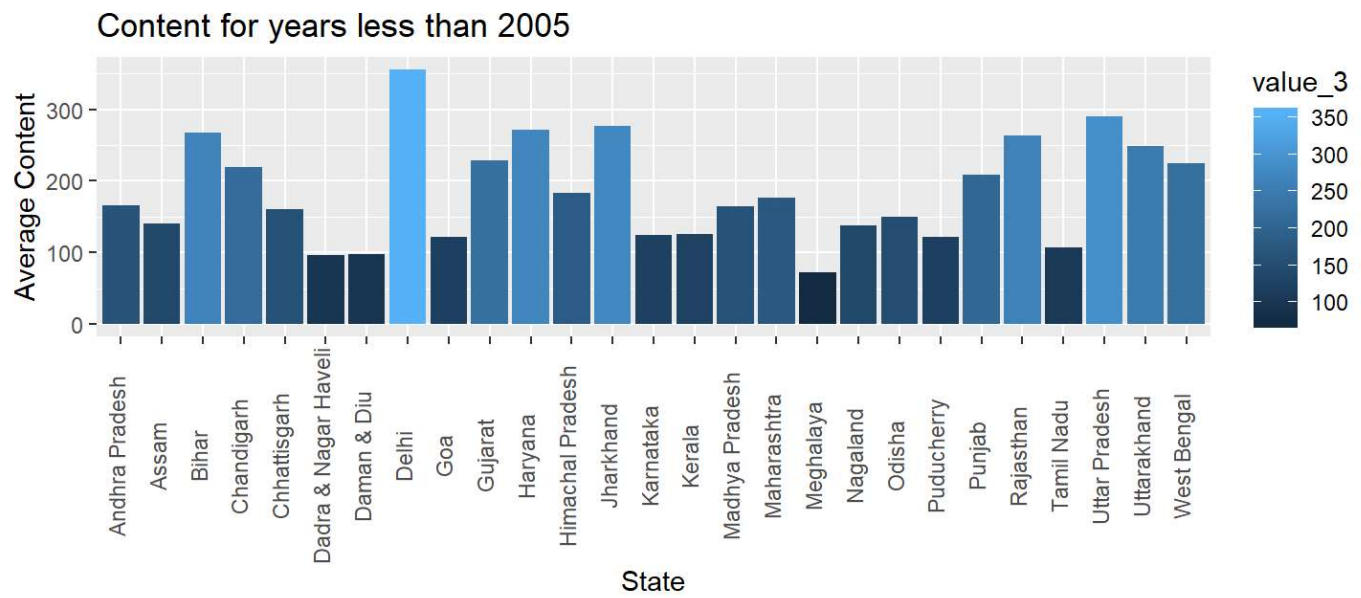


Below graph depicts changes with respect to Rspm levels.Some key takeaways: 1.This data looks farfetched as in values greater than 2010 have similar values but Delhi levels are high . 2.We can compare only the 2nd and 3rd graph as the first graph is not providing much info as the rest of the two.

```
Analysis_5_yrs_gap(last_5_yrs$Avg_Spm,mid_5_yrs$Avg_Spm,first_5_yrs$Avg_Spm)
```

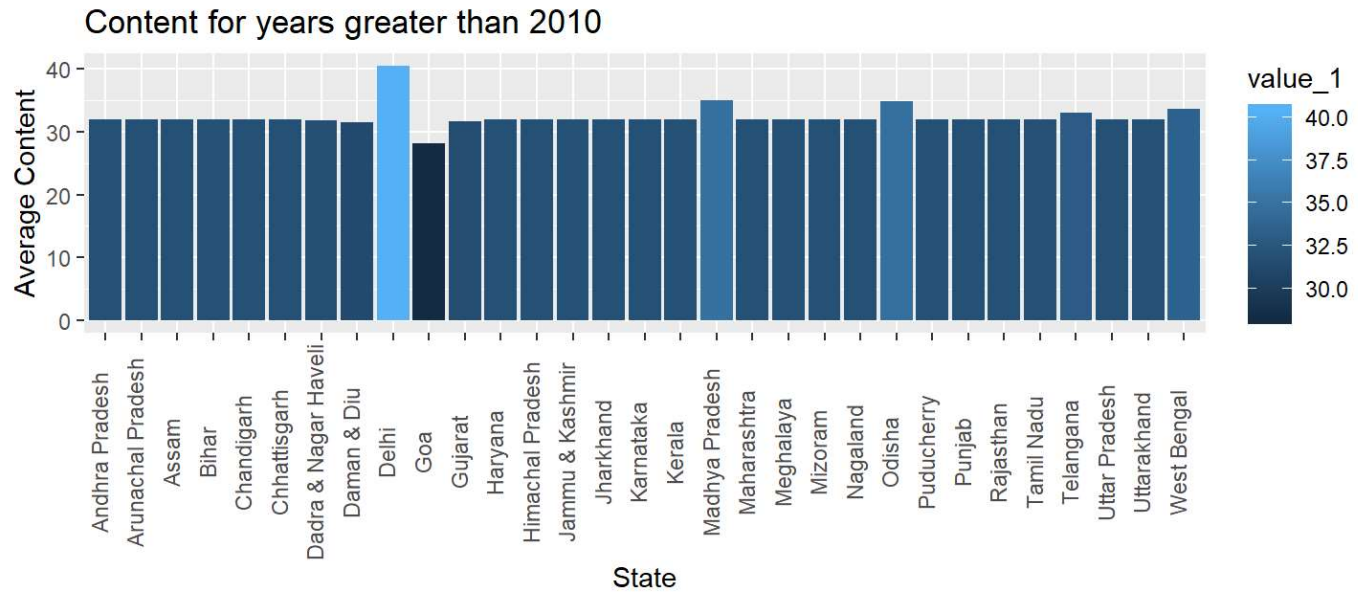


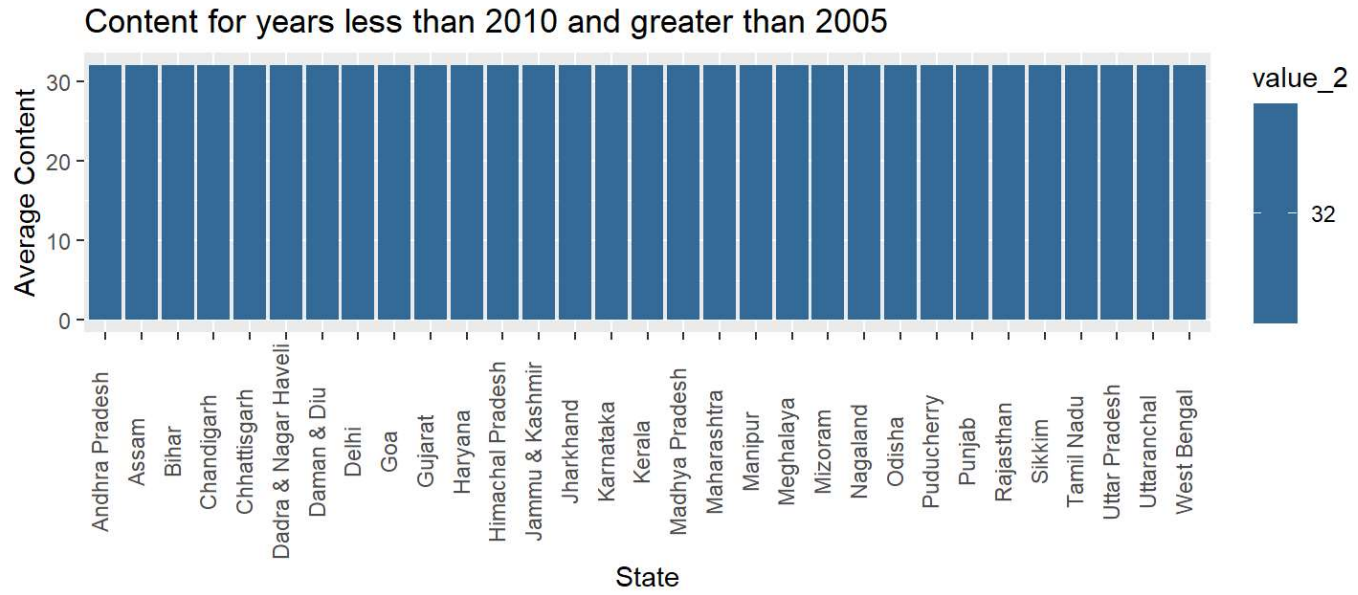


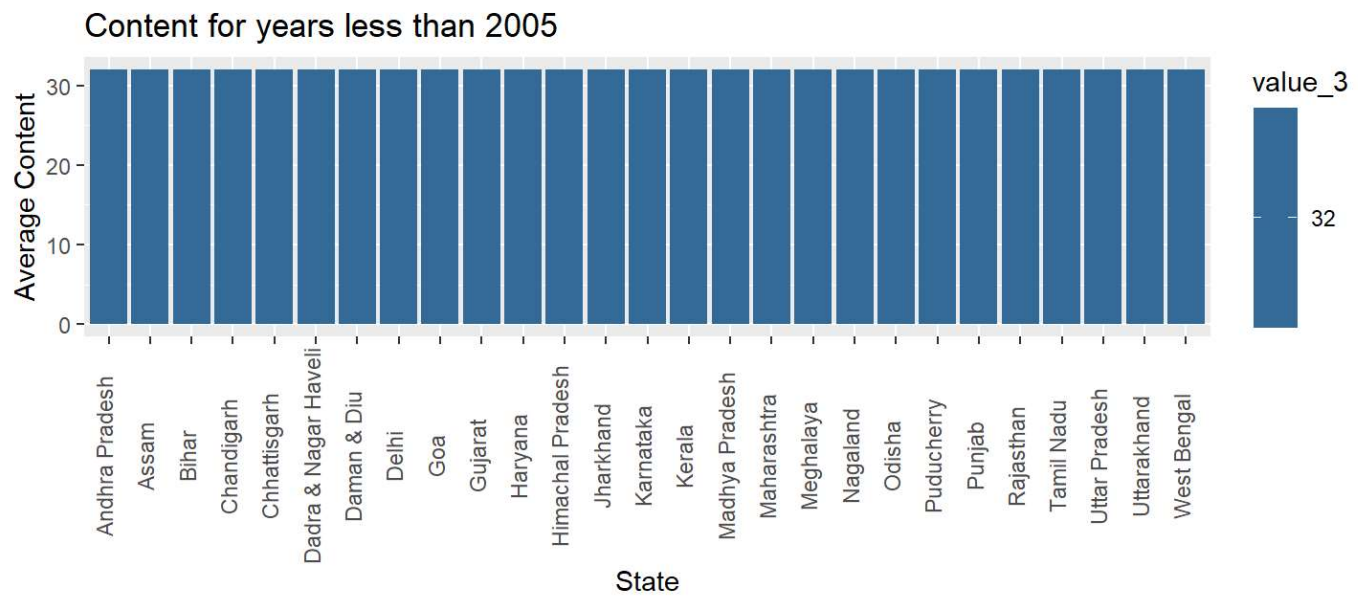


Below graph depicts changes with respect to 2.5 pm levels.Some key takeaways: 1.Because of replacement of NA values the 2nd and 3rd graph is not providing any info. 2.Only info that is able to make properly is of Delhi and that is high level.

```
Analysis_5_yrs_gap(last_5_yrs$AVG_pm2_5,mid_5_yrs$AVG_pm2_5,first_5_yrs$AVG_pm2_5)
```





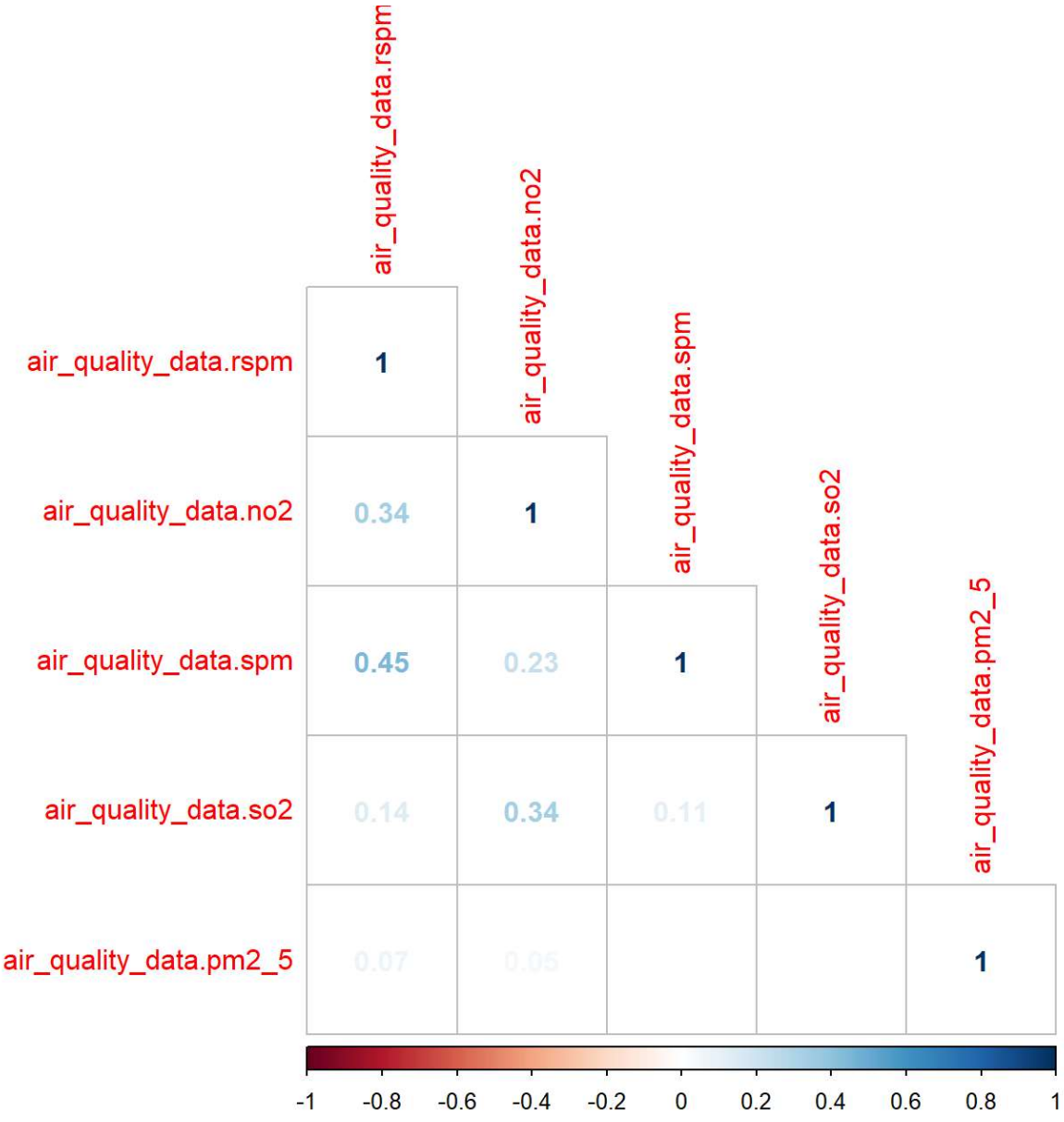


Corrplot packages gives correlation between the variables and it does give interesting insights on how to variables give us features of interest.However it should be noted that **CORRELATION IS NOT CAUSATION**. This means that even though variables may be correlated but it isn't necessary that it might be the cause.More understanding can be found by clicking this link (https://goo.gl/Trxp3W)

But there is a need of exploration on variables where they are correlated. Therefore plots were created of the following graphs on the basis of correlation. The **Rule of Thumb** says if two variables are correlated with >0.3 and <-0.3 then it is meaningful. A correlation of |0.5| is moderate. A correlation of |0.7| is large.

Inferring from the below plot there is not much correlation in the values.

```
numeric_values<-data.frame(air_quality_data$so2,air_quality_data$no2,air_quality_data$rspm,air_quality_data$spm,air_quality_data$pm2_5)
corrplot(cor(numeric_values),method="number"
         ,type = "lower"
         ,order = "FPC"
         ,number.cex=1)
```



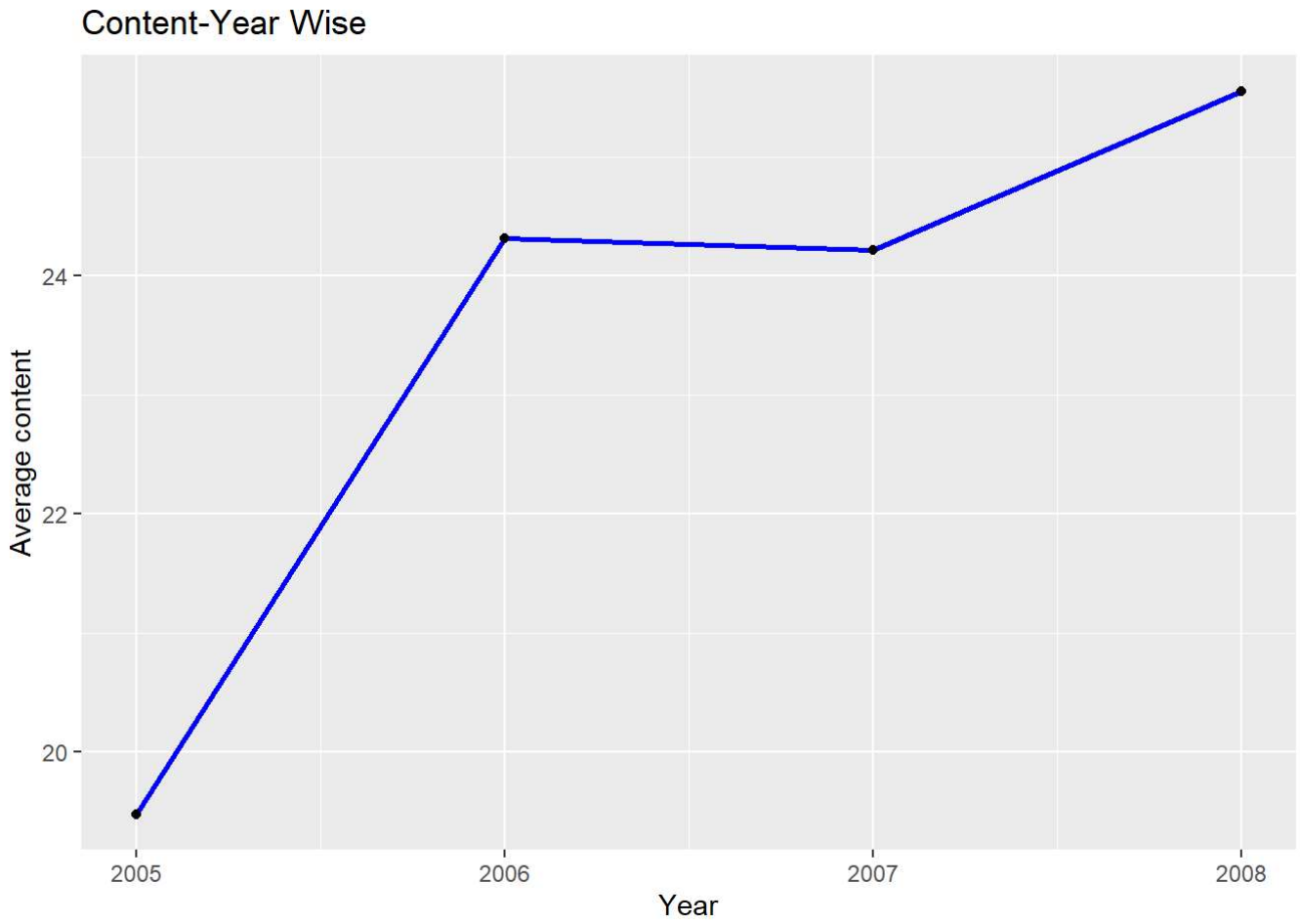
From the above analysis, one needs to find out analysis based on states and how it trends based over the years.

```
air_quality_data$year <- as.numeric(format(air_quality_data$date, "%Y"))
summarise_data_yearwise<-function(data_set,Region){
  data_set%>%filter(state==Region)%>%group_by(year) %>%summarise(Avg_So2=mean(so2),
                                        Avg_No2=mean(no2),
                                        Avg_Rspm=mean(rspm),
                                        Avg_Spm= mean(spm),
                                        AVG_pm2_5=mean(pm2_5))
}

Uttaranchal_data<-summarise_data_yearwise(air_quality_data,"Uttaranchal")
Bihar_data<-summarise_data_yearwise(air_quality_data,"Bihar")
WB_Data<-summarise_data_yearwise(air_quality_data,"West Bengal")
Delhi_Data<-summarise_data_yearwise(air_quality_data,"Delhi")

overall_data_statewise<-function(data_set,value){
  ggplot(data_set,aes(x=year,y=value)) +
    geom_line(size=1,color="blue") +
    geom_point()+
    ggtitle("Content-Year Wise")+
    xlab("Year") +
    ylab("Average content")
}
```

City-Uttaranchal

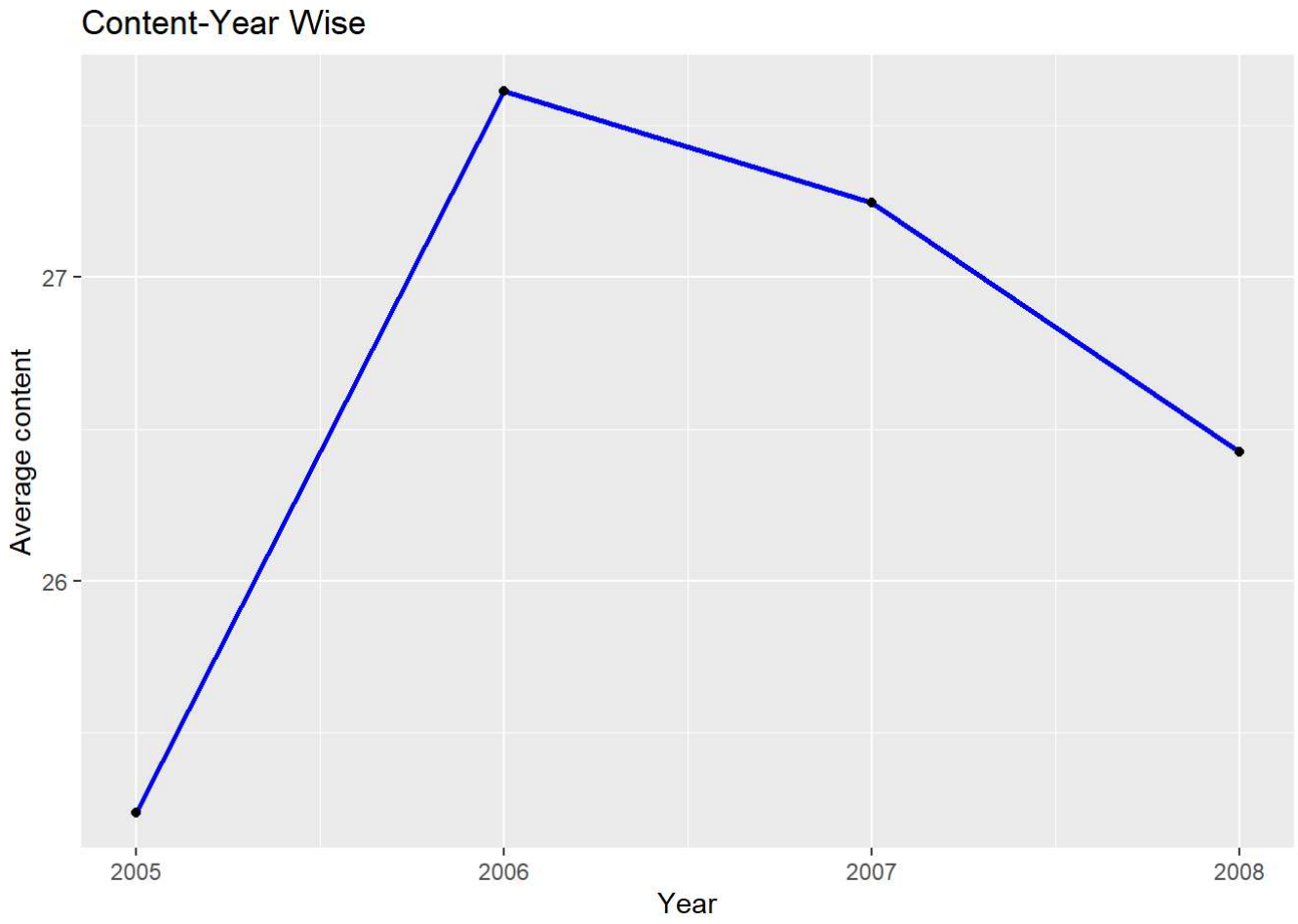
Below is the So2 Levels of Uttaranchal, there is a rise in 2005-2006 and a gradual increase in 2007-2008.As one can see, there is data missing for other areas.

```
overall_data_statewise(Uttaranchal_data,Uttaranchal_data$Avg_So2)
```



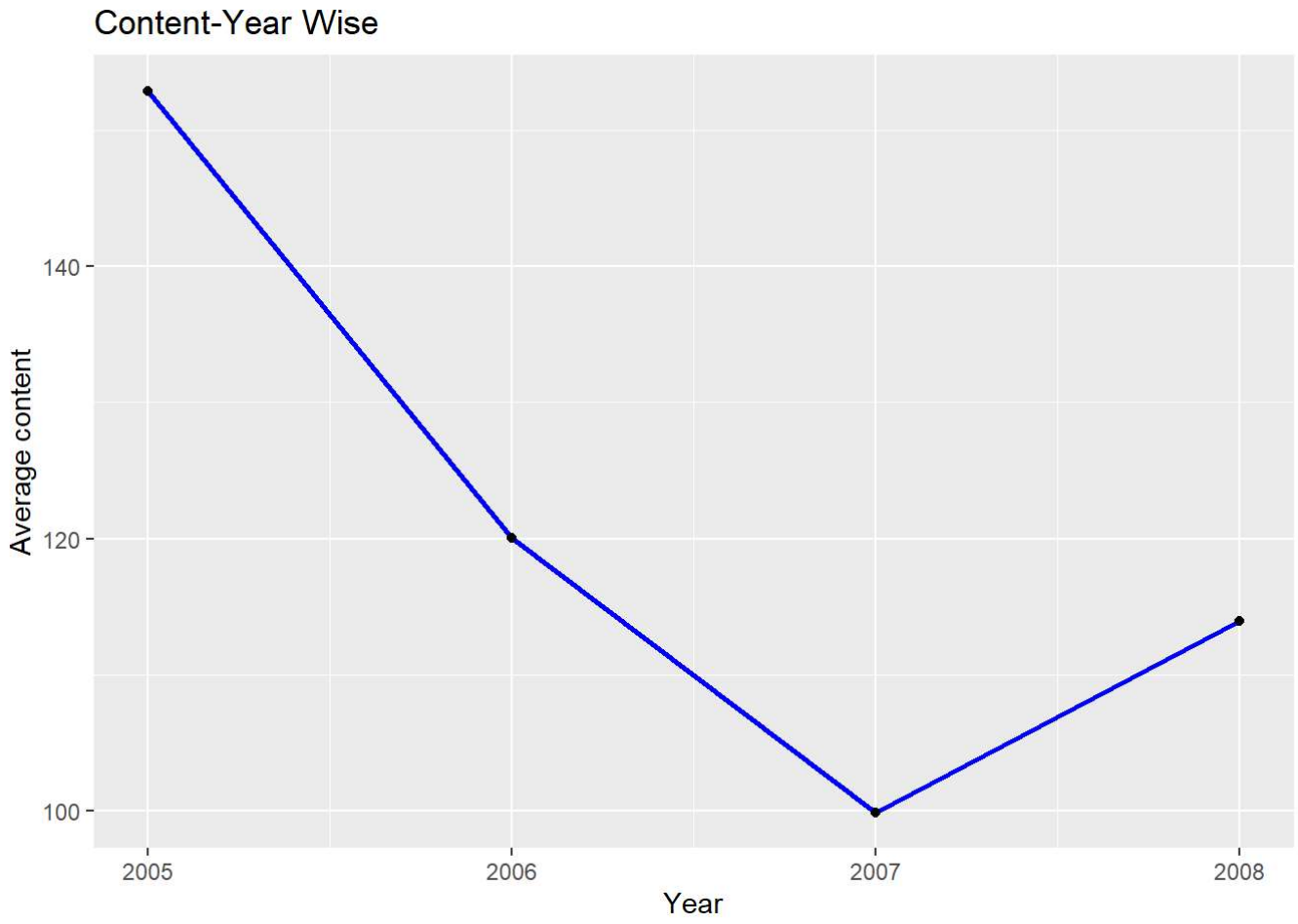
As compared to SO2, there is a decrease in No2 value over the areas after it sees a steep increase in the first year.

```
overall_data_statewise(Uttaranchal_data,Uttaranchal_data$Avg_No2)
```



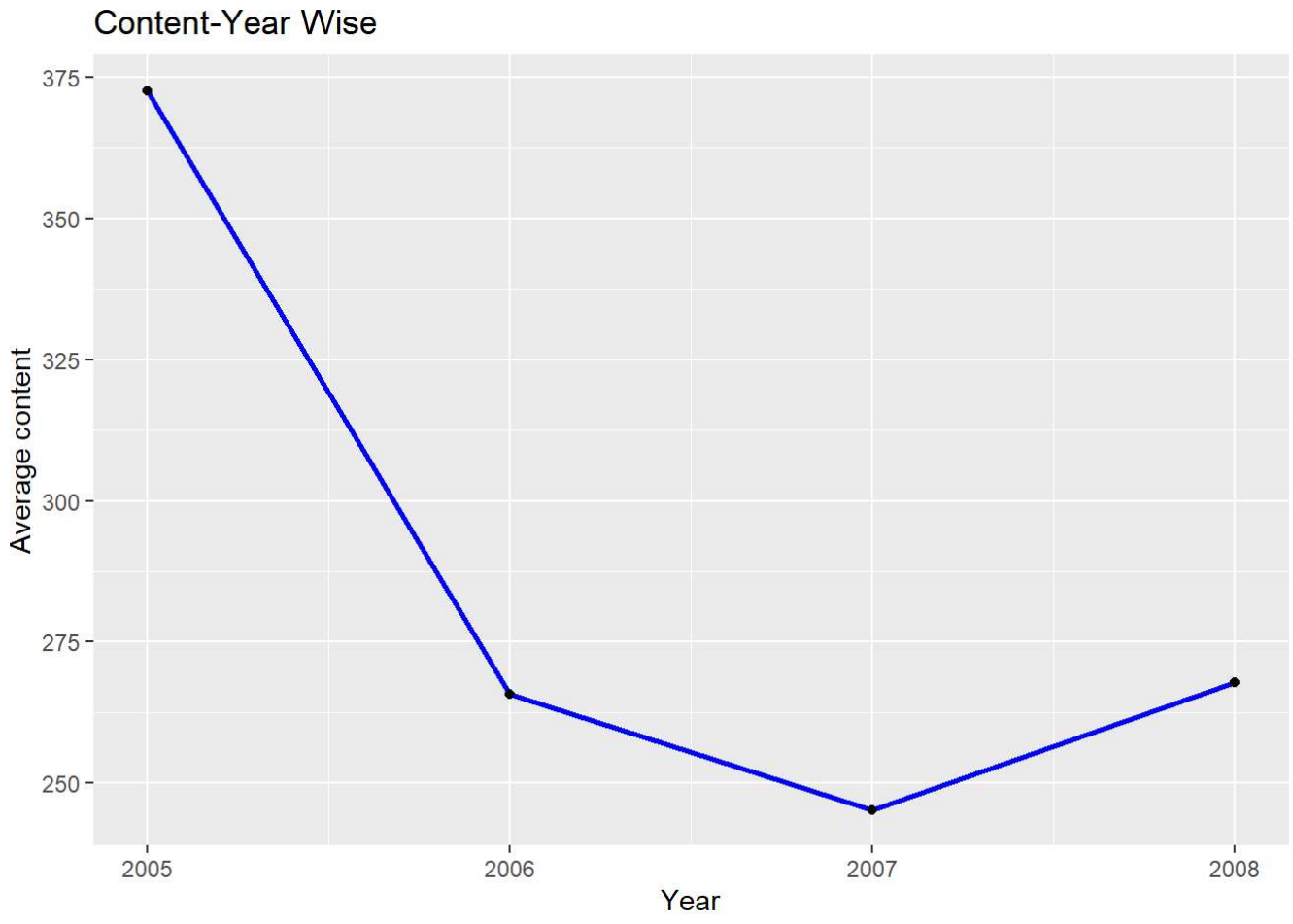
There is a decrease in rspm levels until 2007.

```
overall_data_statewise(Uttaranchal_data,Uttaranchal_data$Avg_Rspm)
```



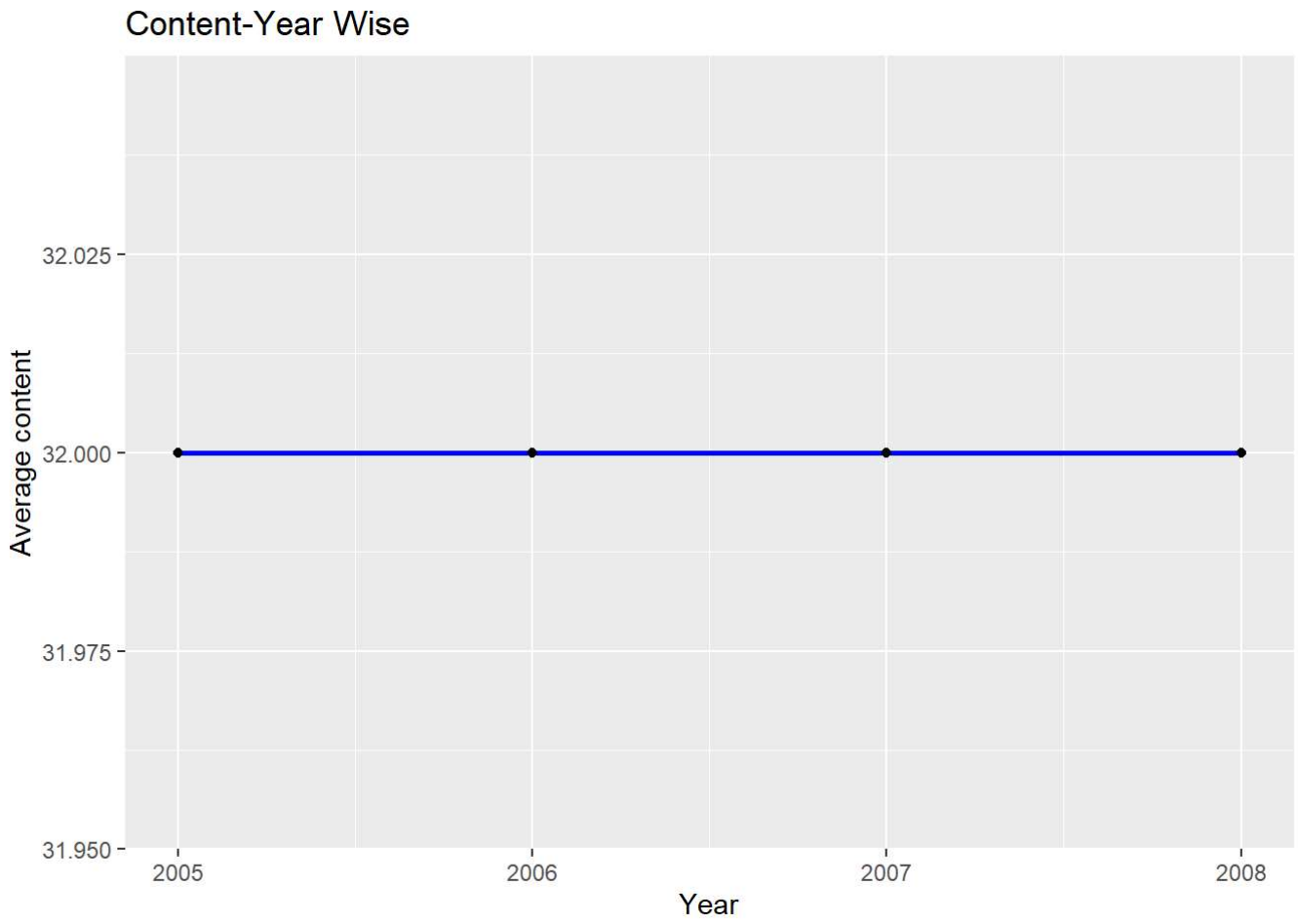
The data is similar to rspm and has seen a decrease till 2007

```
overall_data_statewise(Uttaranchal_data,Uttaranchal_data$Avg_Spm)
```



Data is not available for this one.

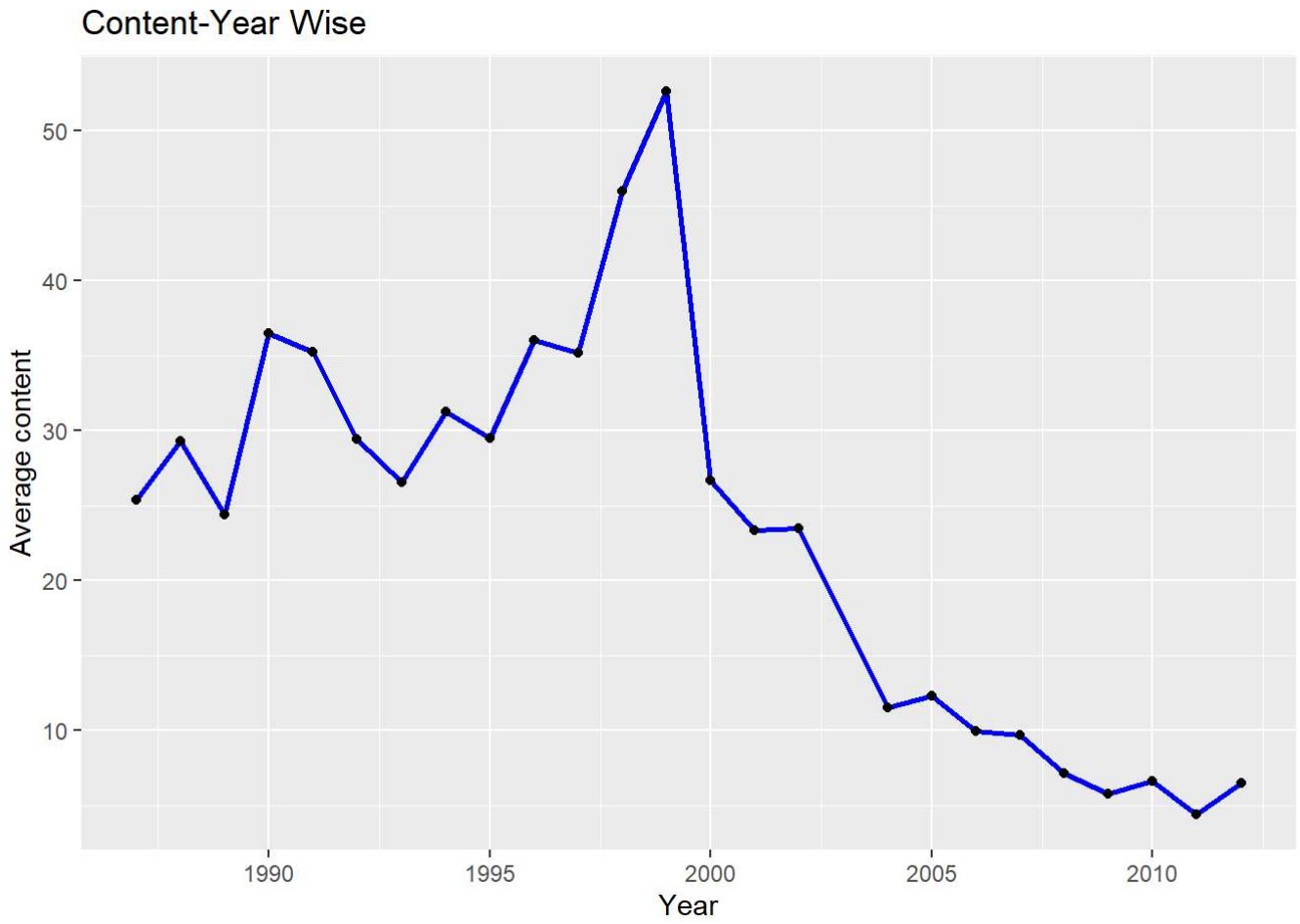
```
overall_data_statewise(Uttaranchal_data,Uttaranchal_data$AVG_pm2_5)
```



City-Bihar

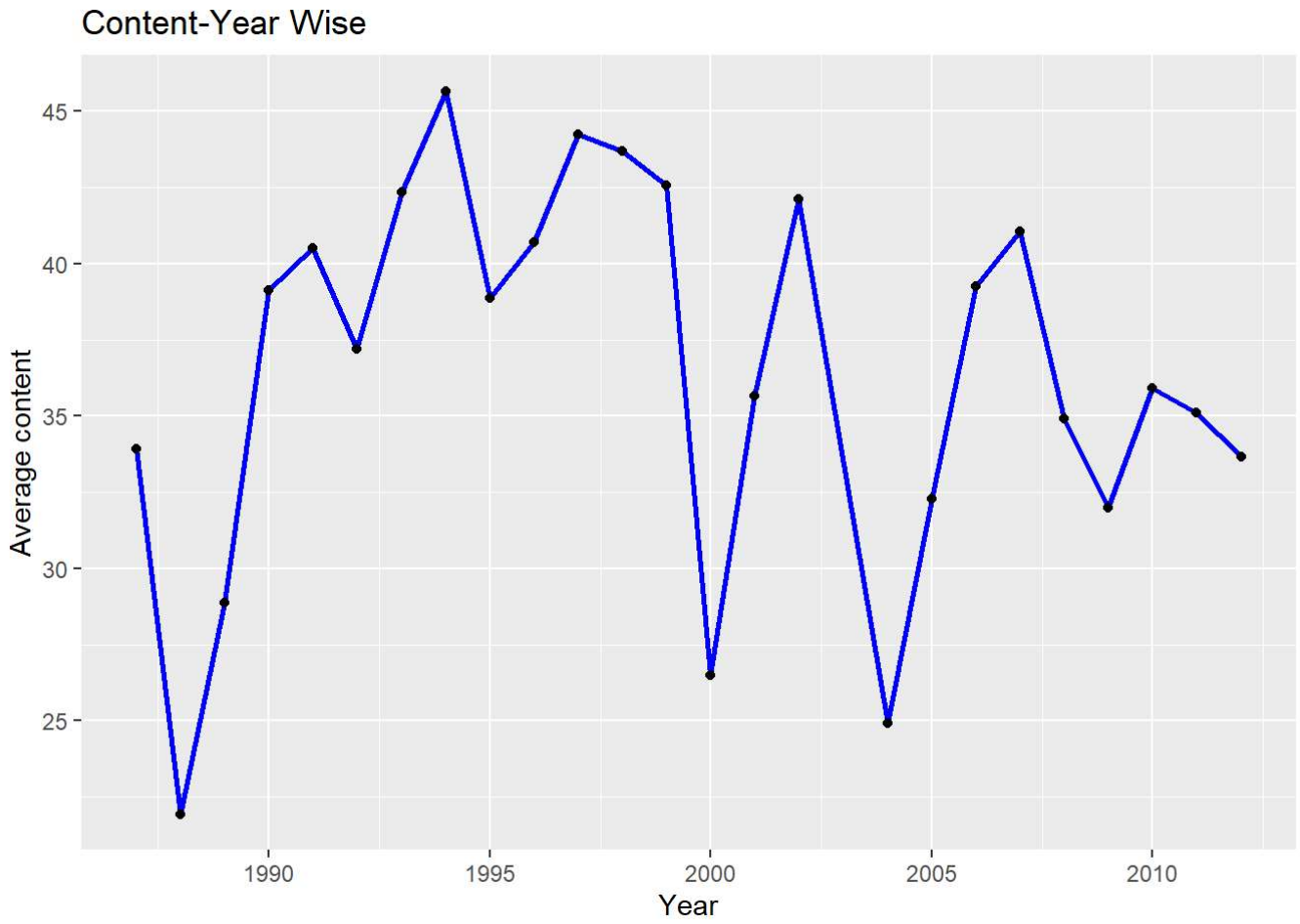
Interestingly so2 levels has decreased over the years from 1996 till 1997 after which 2000 showed a decrease.

```
overall_data_statewise(Bihar_data,Bihar_data$Avg_So2)
```



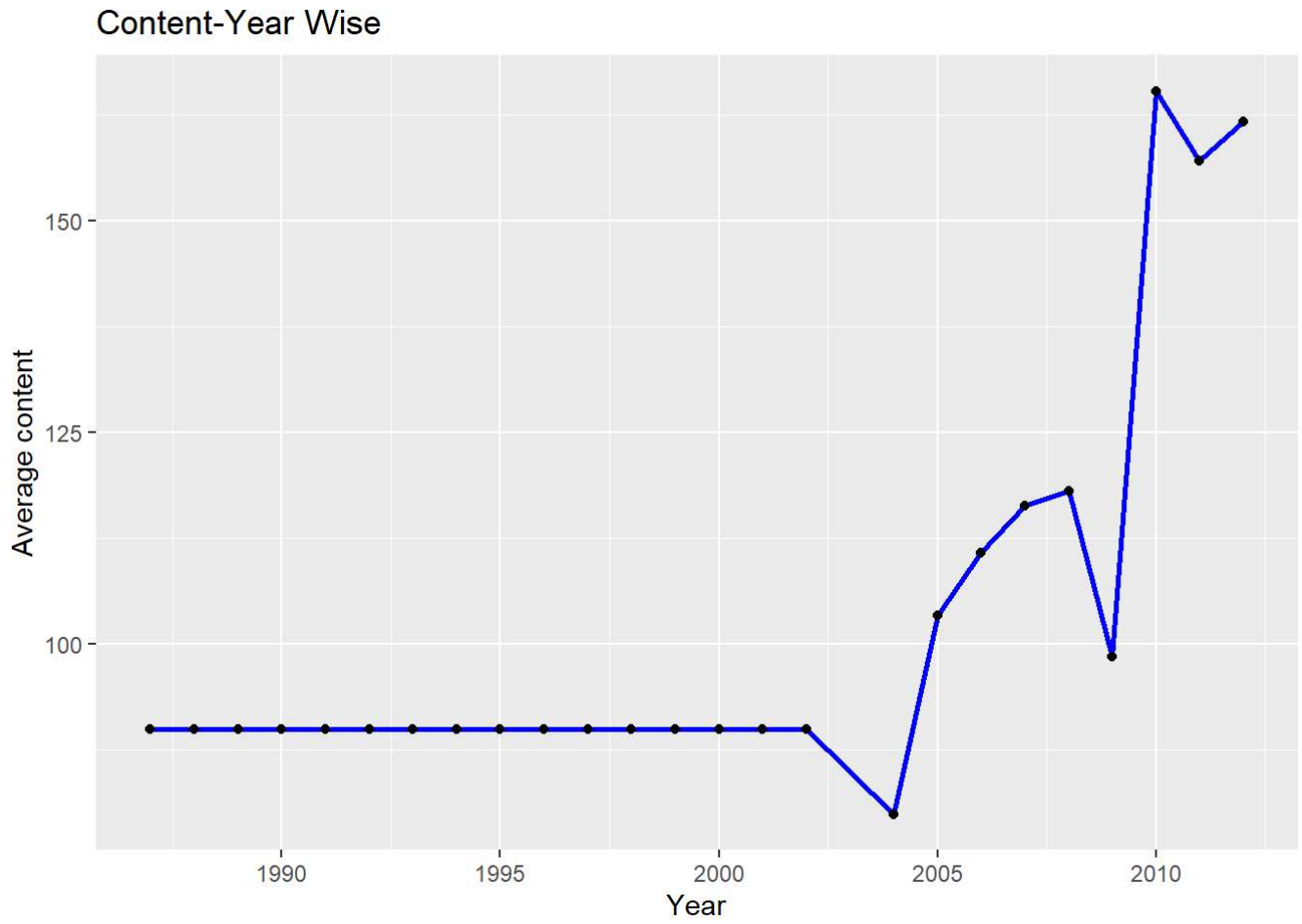
After few up and downs, it reaches a average value .

```
overall_data_statewise(Bihar_data,Bihar_data$Avg_No2)
```



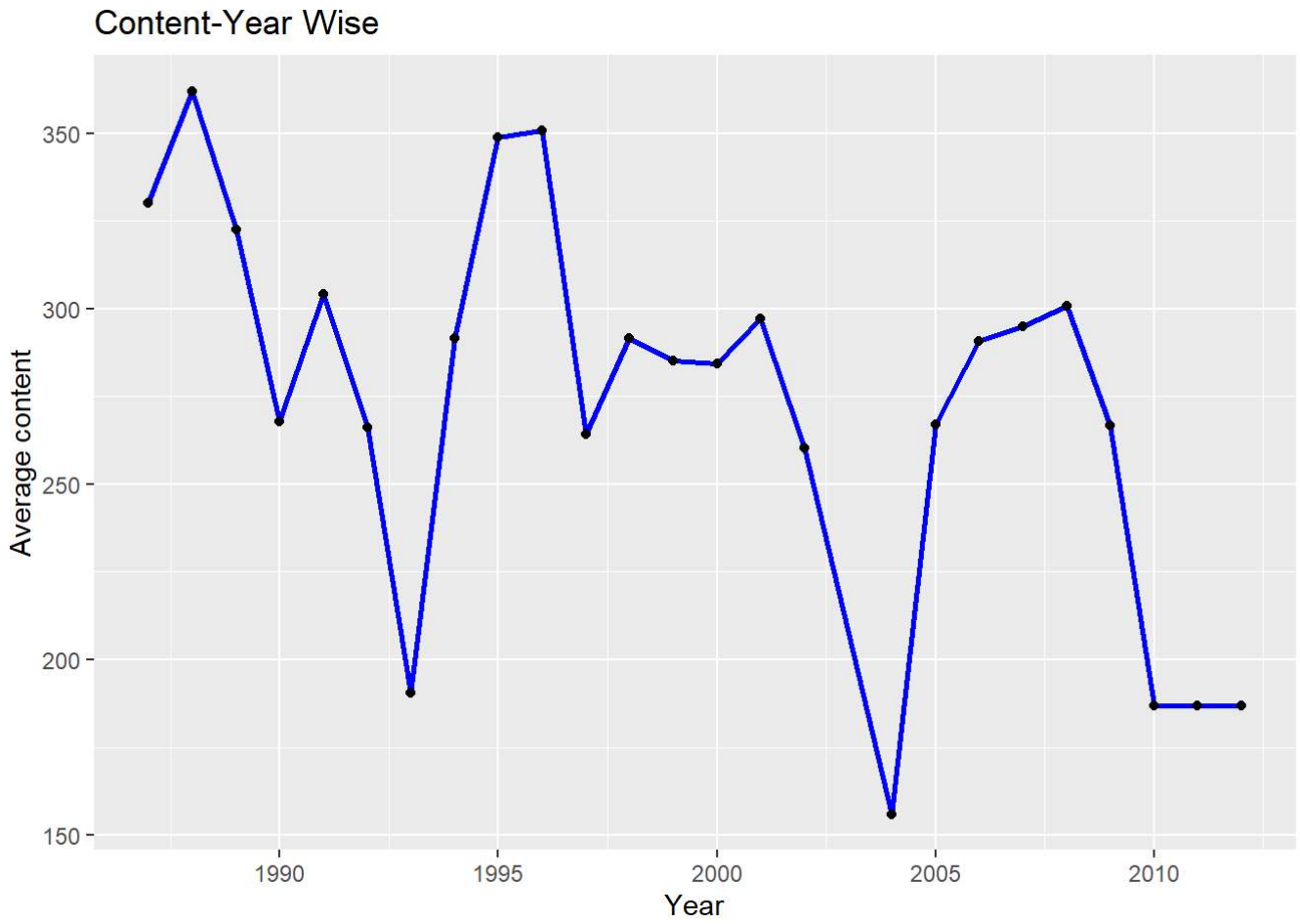
Data is available only after 2002 after which it shows a rise with a steep rise from 2009 to 2010.

```
overall_data_statewise(Bihar_data,Bihar_data$Avg_Rspm)
```



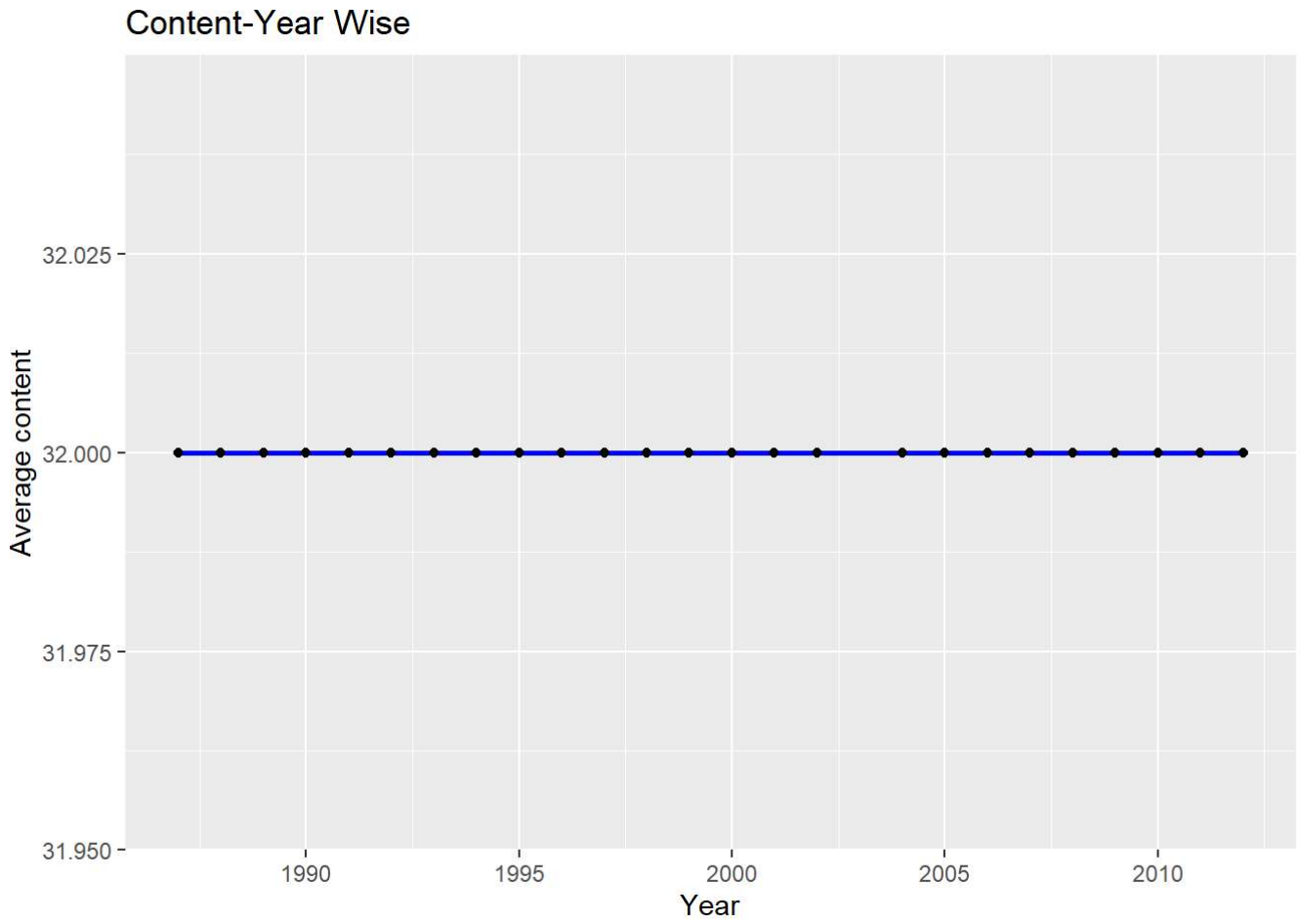
2004 has seen decrease in levels for both so2 and no2.Maybe below can be the reason link (https://en.wikipedia.org/wiki/2004_Bihar_flood)

```
overall_data_statewise(Bihar_data,Bihar_data$Avg_Spm)
```



No content available

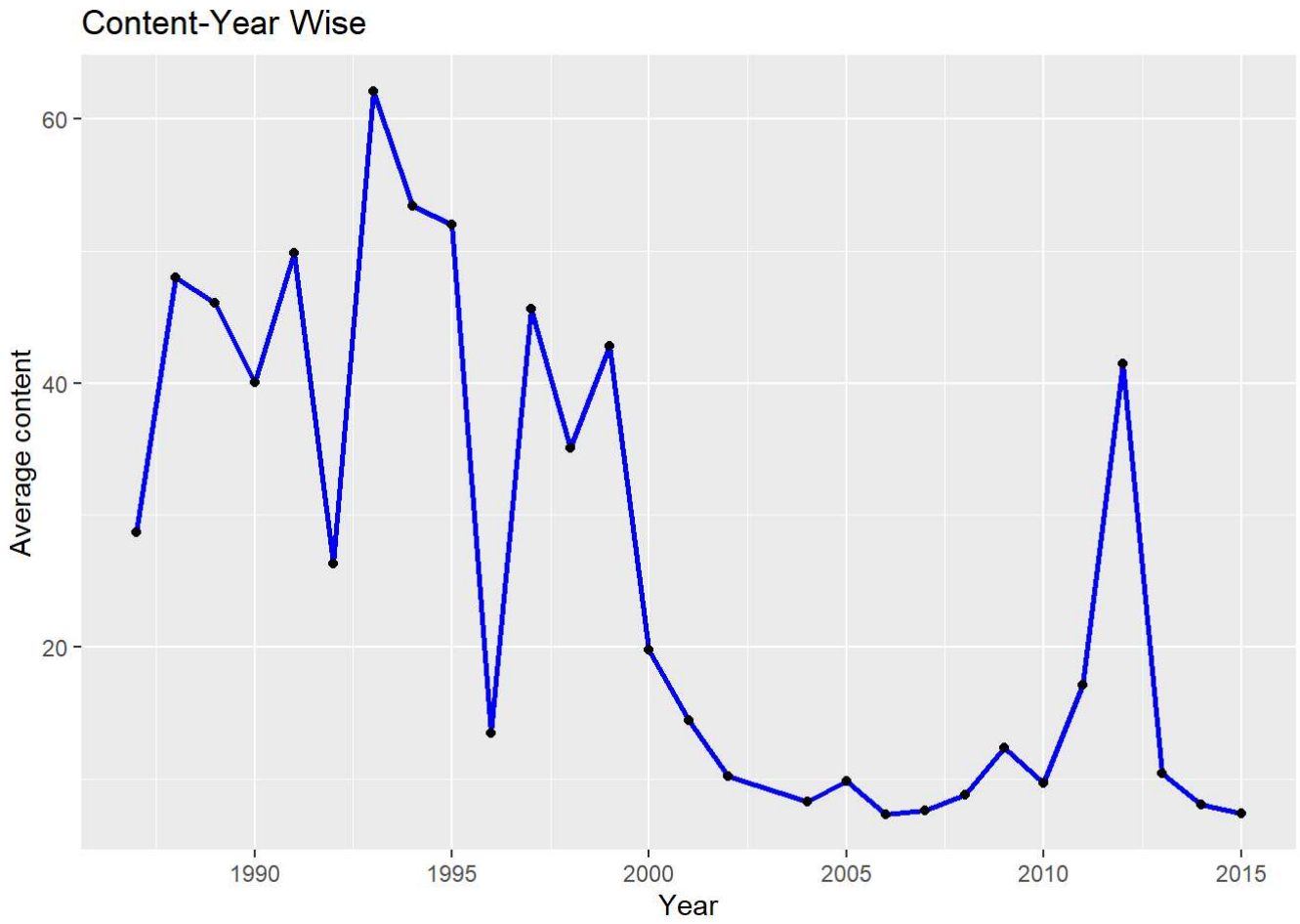
```
overall_data_statewise(Bihar_data,Bihar_data$AVG_pm2_5)
```



City-West Bengal

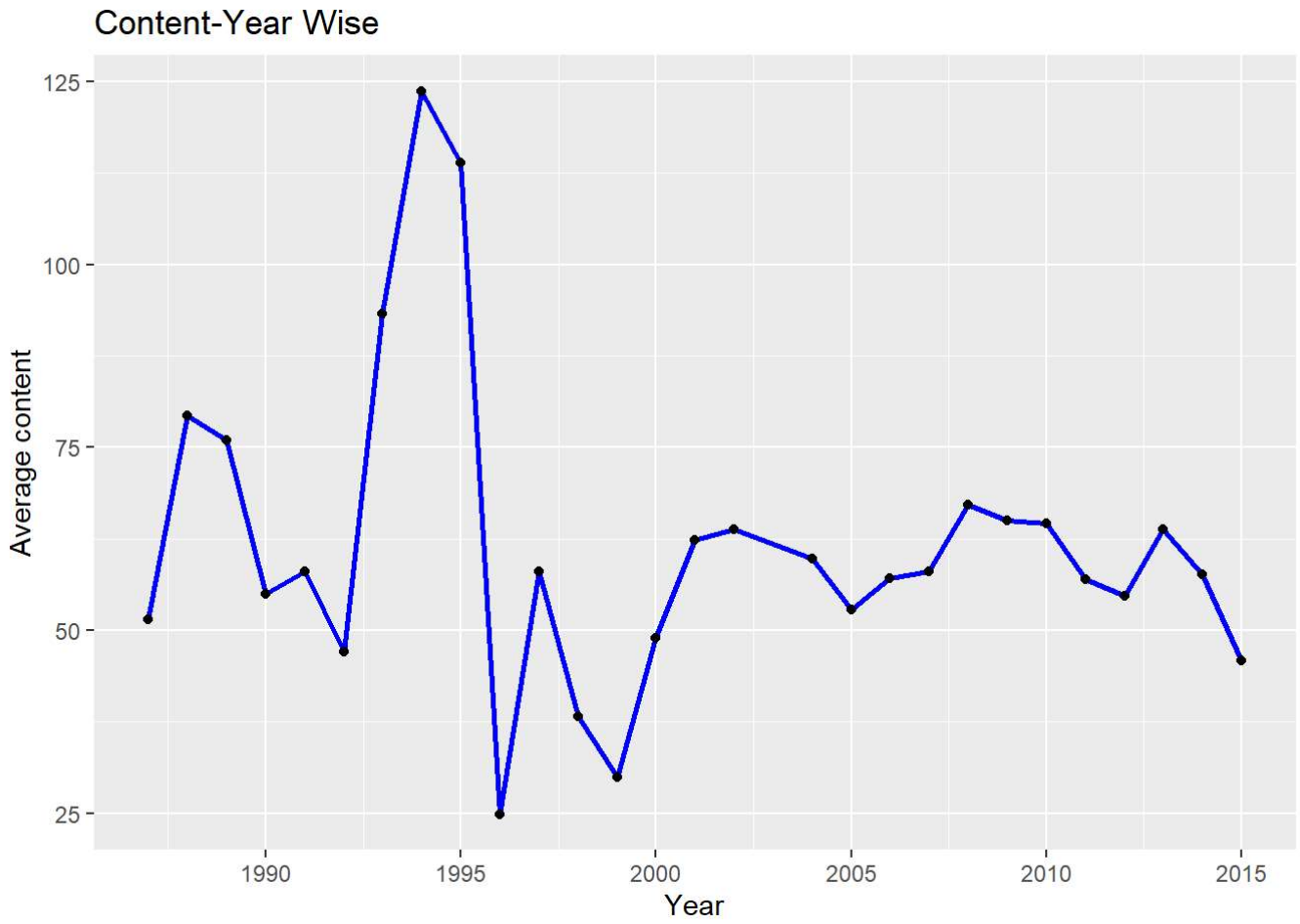
West bengal has seen rise in so2 levels from 2000 to 2010 until a rise in 2012.

```
overall_data_statewise(WB_Data,WB_Data$Avg_So2)
```



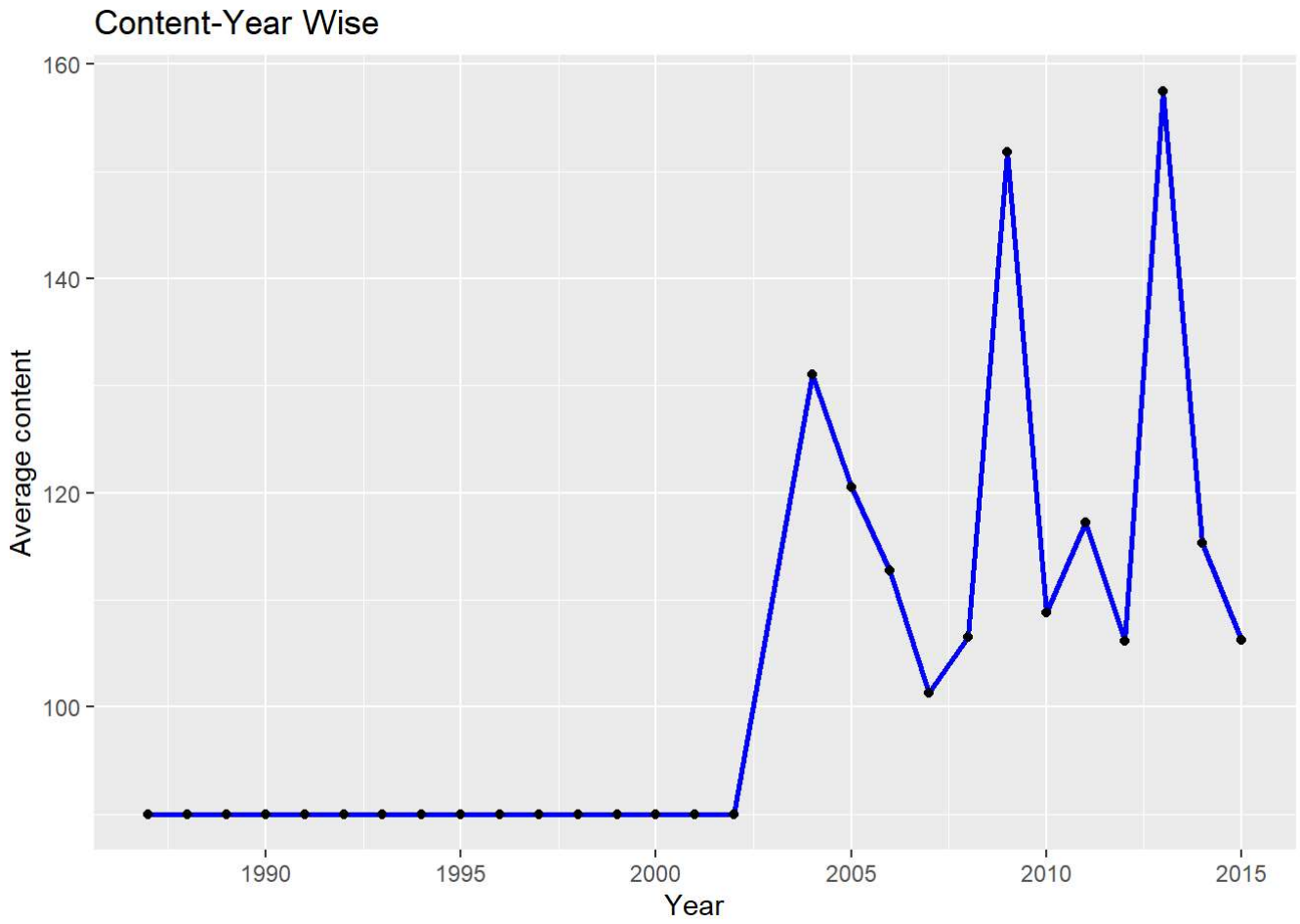
There is no2 levels only risen during 1994 and then it has got a average value for no2

```
overall_data_statewise(WB_Data,WB_Data$Avg_No2)
```



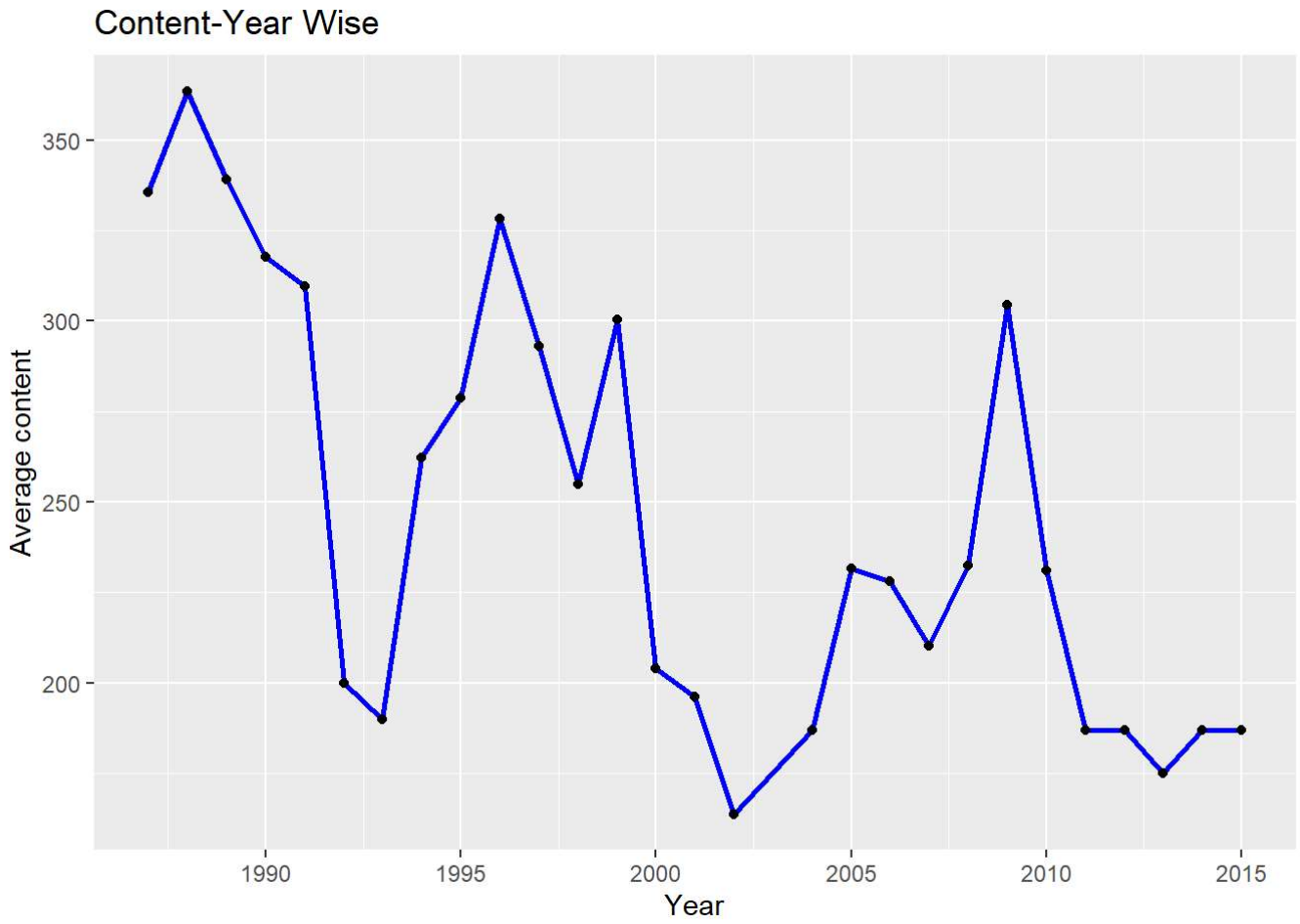
2009 and 2013 values had high level of rspm until 2015 when there is a decline.

```
overall_data_statewise(WB_Data,WB_Data$Avg_Rspm)
```



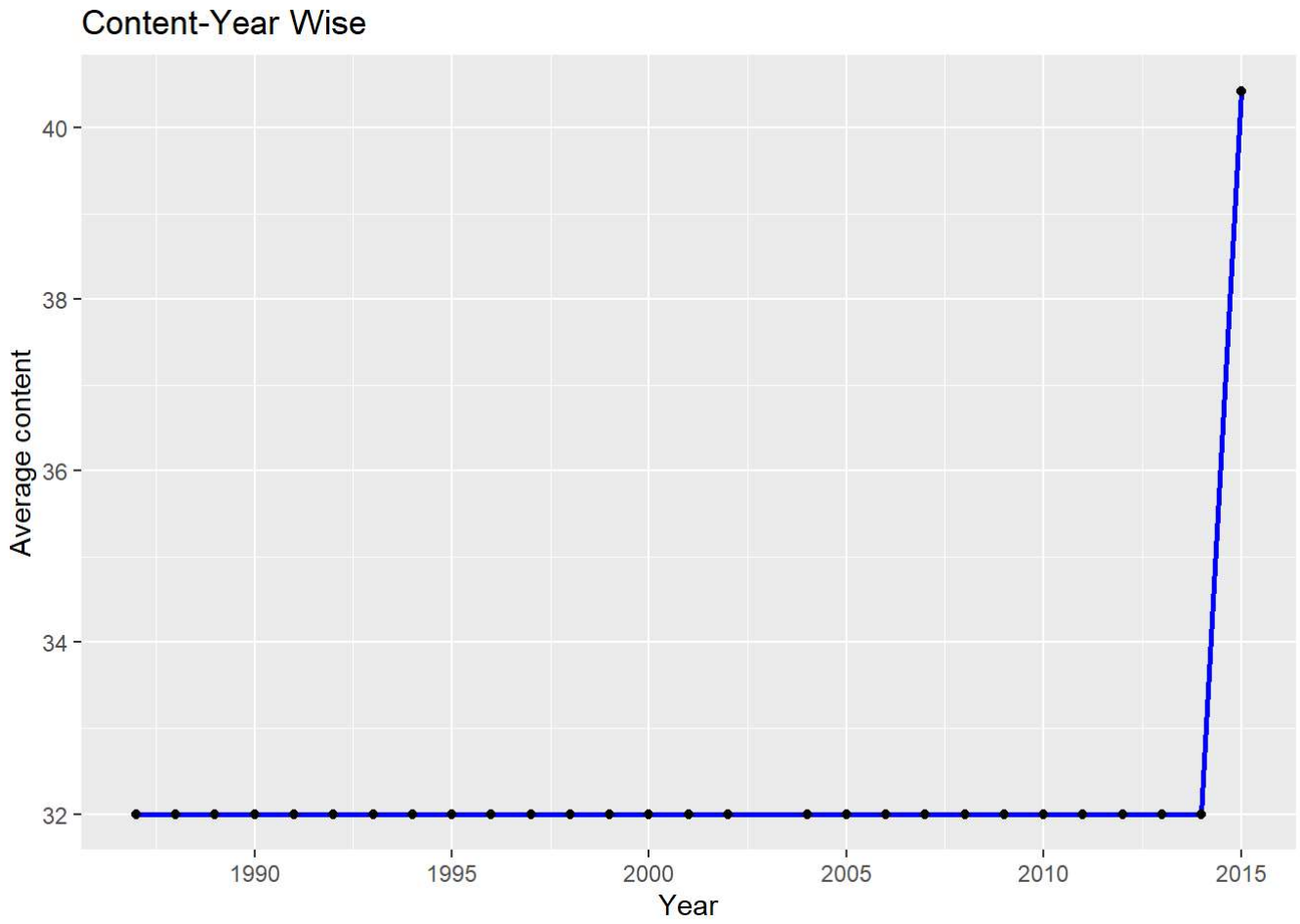
Overall there is decrease in the levels of Spm

```
overall_data_statewise(WB_Data,WB_Data$Avg_Spm)
```



There is less data available to analyse further.

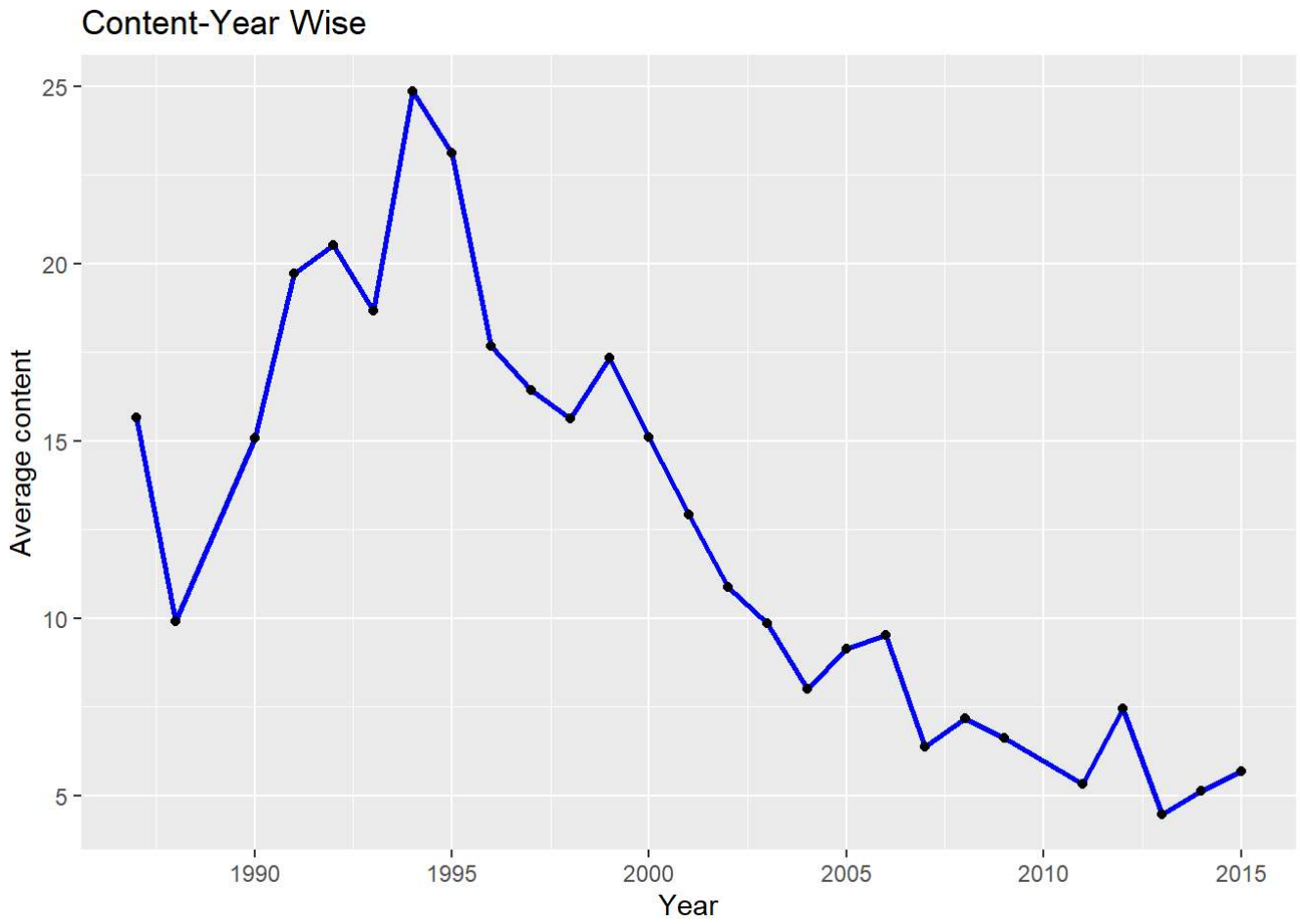
```
overall_data_statewise(WB_Data,WB_Data$AVG_pm2_5)
```



City-Delhi

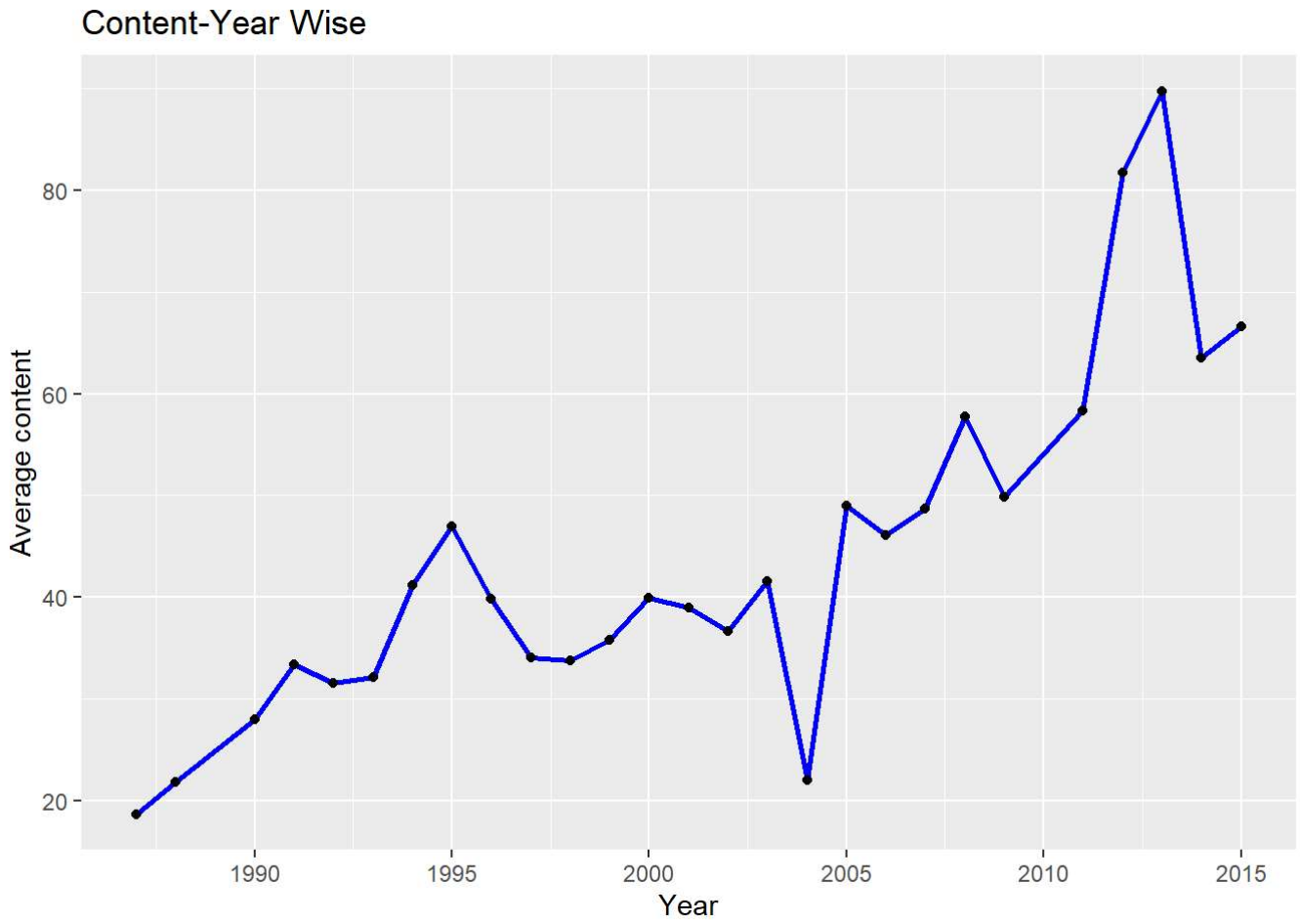
In this this has shown a decrease in so2 levels in Delhi as compared to graphs seen before

```
overall_data_statewise(Delhi_Data,Delhi_Data$Avg_So2)
```



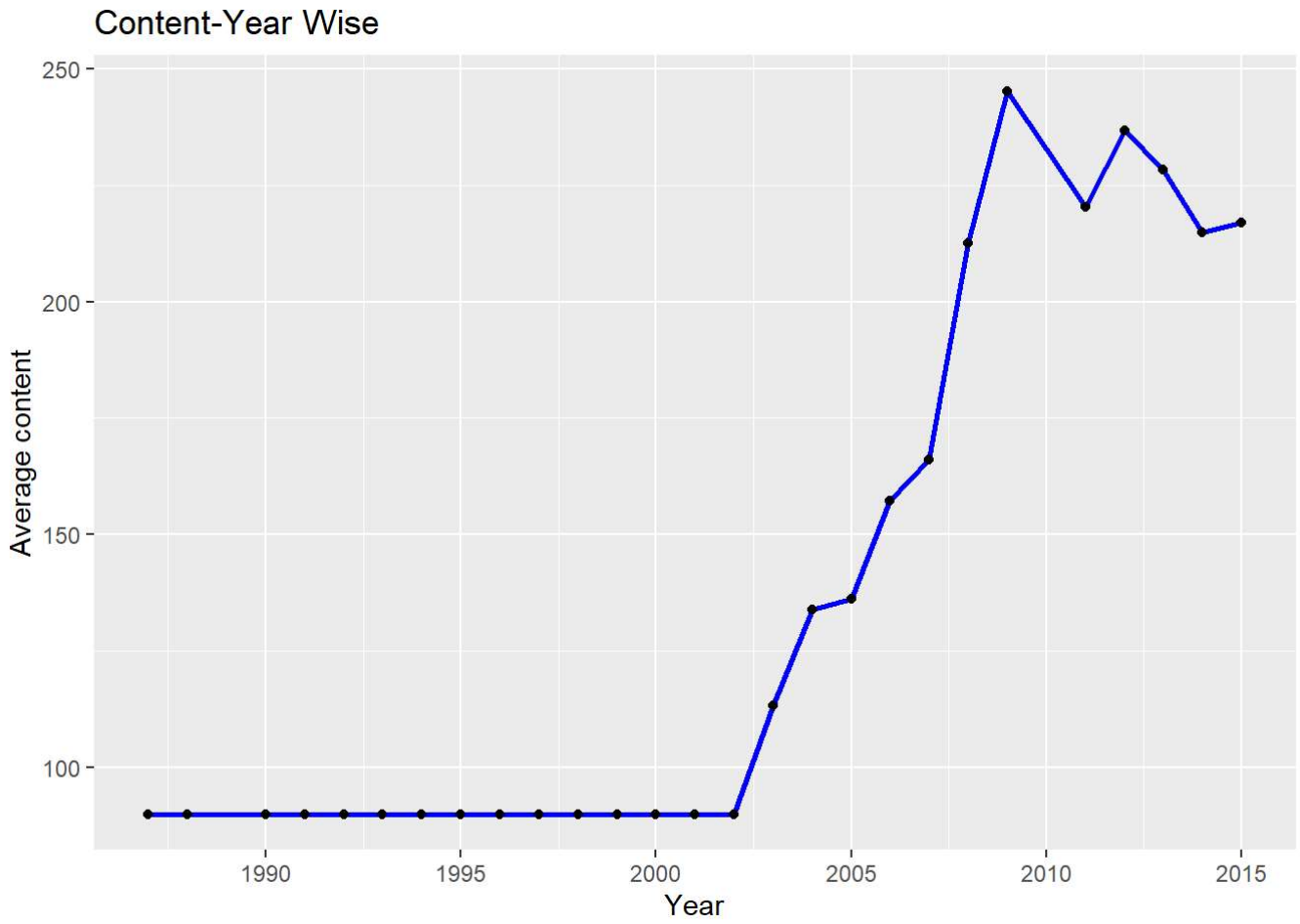
The No2 levels has seen quite a rise overall.

```
overall_data_statewise(Delhi_Data,Delhi_Data$Avg_No2)
```



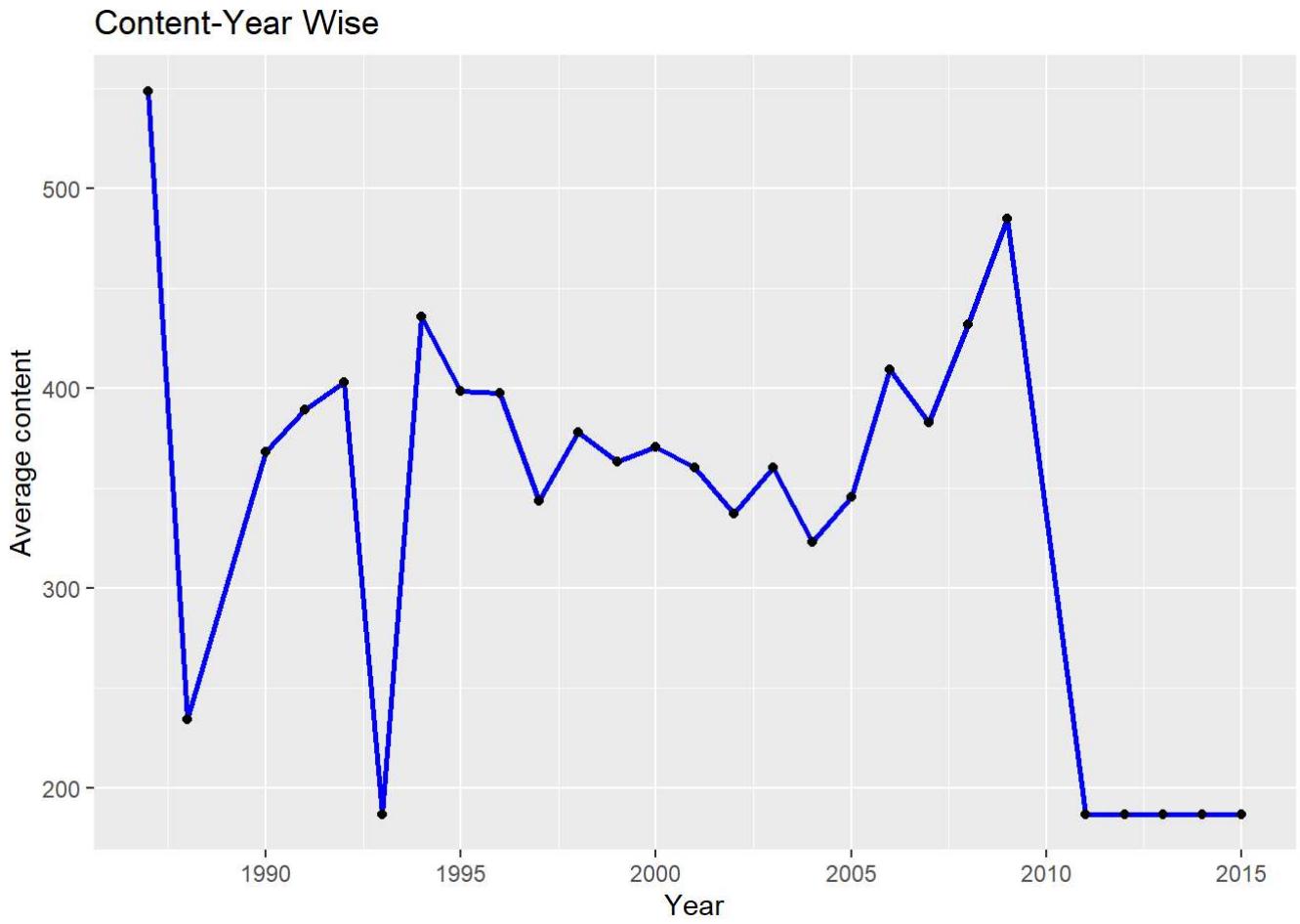
The data is available from 2002 and it has got a steep rise as compared to states we have measured before.

```
overall_data_statewise(Delhi_Data,Delhi_Data$Avg_Rspm)
```



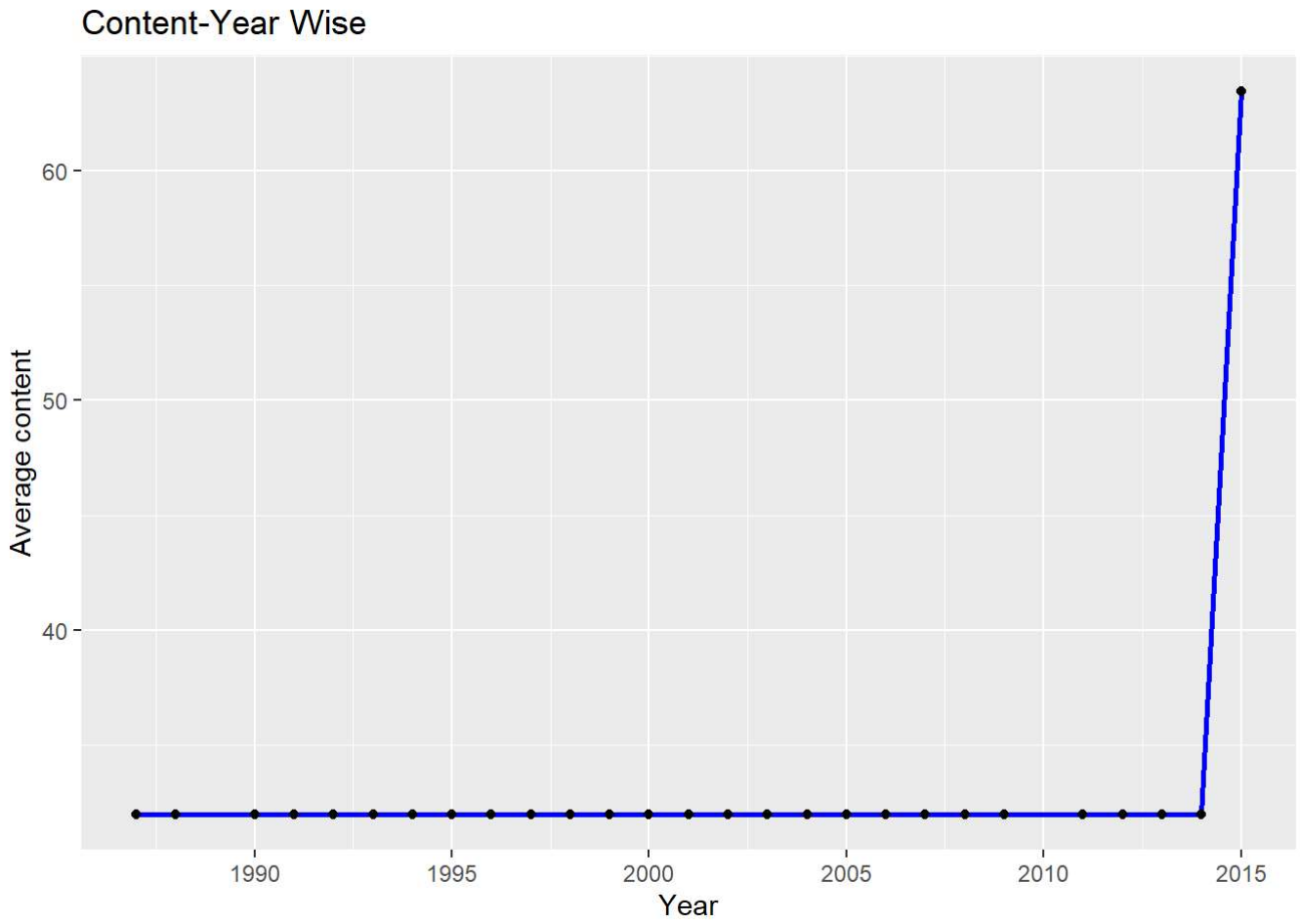
Average Spm hasn't seen much change

```
overall_data_statewise(Delhi_Data,Delhi_Data$Avg_Spm)
```



Not much can be inferred from the data

```
overall_data_statewise(Delhi_Data,Delhi_Data$AVG_pm2_5)
```



Values has seen an increas in no. of residential areas as compared to industrial areas .

```
data_based_on_type<-data.frame(unclass(table(air_quality_data$type,air_quality_data$year)))


data_based_on_type<-data.frame(aggregate(data_based_on_type, list(Group=replace(rownames(data
_based_on_type),rownames(data_based_on_type) %in% c("Residential","Residential and others","R
esidential, Rural and other Areas","RIRUO"), "Residential")), sum))
data_based_on_type
```

```
##             Group X1987 X1988 X1989 X1990 X1991 X1992 X1993 X1994 X1995 X1996
## 1 Industrial Area   233   343   430   740   733   470   539   585   707   746
## 2     Residential   158   253   343   706   700   403   538   513   650   716
## 3  Sensitive Area    26    33    30    35    36    22    22    24    31    36
##   X1997 X1998 X1999 X2000 X2001 X2002 X2003 X2004 X2005 X2006 X2007 X2008 X2009
## 1   689   652   667   715   767   784  1057  6605  7615 10537 12156 11569  9870
## 2   725   720   684   717   788   834  1298  9261 11280 17140 20895 19878 17598
## 3    36    36    36    33    35    24     0   253   441  1185  1325  1304  1281
##   X2010 X2011 X2012 X2013 X2014 X2015
## 1 12509 13394 10692 13545 13968 14752
## 2 21755 22785 23189 30606 28332 33801
## 3   711  1462  1220  1652  1915  1766
```