DATA ANALYSIS OF AIR QUALITY IN INDIA

Parmpreet Singh[#1], Sonia Bhalla[*2], Shikha Tuteja[*3]

Master of Data Science[#1], Assistant Professor[*2], Assistant Professor[*3], Chandigarh University, Gharuan, Mohali, Punjab-140413, India

parmsohi1@gmail.com[#1], soniamaths5@gmail.com[*2], shikhatuteja85@gmail.com[*3]

Abstract:

In this paper, the data analysis, and its visualization related to air quality in India is studied. For data visualization and analysis, multifarious languages are available such as Python, R, etc. Most of the researchers utilized the R programming language as it has better-inbuilt tools as compared to the python language. So, here we have also considered the R language for the data analysis and data visualization process. Further, the Dataset "Air Quality Data in India (2015 - 2020)" has been used for the analysis and visualization of air quality. Mainly four parameters are focused for the discussion of the air quality of cities in India. The first parameter is Particulate Matter 2.5(PM 2.5). It has a diameter less than 2.5 micrometers. The second parameter is AQI (Air Quality Index). The third parameter is Particulate Matter-10(PM 10) having diameter less than ten micrometers. The last one is O3(Ozone gas). Some major causes and its constructive measures are also discussed.

Keywords: Data visualization; R language; Particulate Matter 2.5; AQI; Particulate Matter-10; ozone gas

DOI Number: 10.14704/nq.2022.20.9.NQ44377 **Neuro Quantology 2022; 20(9):3271-3277**

Background

In the contemporary era, with the advancement in technology, air pollution is emerging as the most grave problem in the world. There are some papers on the internet resonating air pollution in India. These papers mainly focus on air pollution in the capital of India that is Delhi and also some cities in India have very good air quality that is not cover in that papers.

Motivation

Nowadays, data is very cardinal for us. We can anticipate the future by analyzing the past data. It succors the organizations to ameliorate their product. It also helps in marketing the most apposite product to the right consumer. With this method, companies are able to save exorbitant amount of money and invest that money in their product that escalates the company's economic growth.

Data visualization is desideratum because of transformation of huge amounts of complex data sets into the human brain is very easy visually with graphs and charts than reading the spreadsheet and reports. Trends in the data can be easily recognized by the data visualization tools.

Objective

In this paper, air quality of India has been discussed with the air quality of cities of India with the help of graphics. For the analysis of air quality, we need a large dataset that is easily available on internet. The aim of this paper is to study the air quality of cities of India and find out which cities of India is good for living and which are bad with data visualization and analysis using R programing.

In R language for any type of data visualization we have some standard graphics like Scatterplots, Pie charts, Boxplots, Bar plots, etc. but in this paper, bar plot and box plot graph have been utilized because we need to figure out about the most polluted city in India and less polluted. Over and above, to calculate the average value of air quality, a parameter has been plotted by bar plot and to embellish more details about data box plot has been employed.

Contribution/Methods

The Dataset "Air Quality Data in India (2015 - 2020)" is raw data. To get any information from that data without working on it is impossible. To find desired result first understanding of data is crucial. To find some useful information from raw data is very complex task. In this paper firstly get cities names. In next step select some columns according to the result and then match the data according to anticipated outcome after that other analysis work done on dataset.

1.Data Visualization of air quality of India:

To check the air quality of India, different cities of India have been compared which are namely Ahmedabad, Aizawl, Amaravati, Amritsar, Bengaluru, Bhopal,

Brajrajnagar, Chandigarh, Chennai, Coimbatore, Delhi, Ernakulam, Gurugram, Guwahati, Hyderabad, Jaipur, Jorapokhar, Kochi, Kolkata, Lucknow, Mumbai, Patna, Shillong, Talcher, Thiruvananthapuram, Visakhapatnam. In the data visualization dataset "Air quality data in India" has been used which is available on "www.kaggle.com". [6]

2.Process of data analysis and visualization:

2.1 Collection of data:

Data has been gathered from two different sources; first are from internal sources (own sources) like customer info, sales data, and external databases such as internet, any third party storage devices.

In this paper data is collected from external source "www.kaggle.com" website.

2.2 Import data in R studio:

For import data in R studio we need to move downloaded file into current working directory

```
data<-read.csv("city_day.csv")
```

and after that we have to execute this script.

```
# get name of cities into vector form
v=table(Fdata$City)
V<-as.data.frame(v)
cities<-as.vector(V$Var1)
```

	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
1	Ahmedabad	2015-01-01	NA	NA	0.92	18.22	17.15	NA	0.92	27.64	133.36	0.00	0.02	0.00	NA	
2	Ahmedabad	2015-01-02	NA	NA	0.97	15.69	16.46	NA	0.97	24.55	34.06	3.68	5.50	3.77	NA	
3	Ahmedabad	2015-01-03	NA	NA	17.40	19.30	29.70	NA	17.40	29.07	30.70	6.80	16.40	2.25	NA	
4	Ahmedabad	2015-01-04	NA	NA	1.70	18.48	17.97	NA	1.70	18.59	36.08	4.43	10.14	1.00	NA	
5	Ahmedabad	2015-01-05	NA	NA	22.10	21.42	37.76	NA	22.10	39.33	39.31	7.01	18.89	2.78	NA	
6	Ahmedabad	2015-01-06	NA	NA	45.41	38.48	81.50	NA	45.41	45.76	46.51	5.42	10.83	1.93	NA	
7	Ahmedabad	2015-01-07	NA	NA	112.16	40.62	130.77	NA	112.16	32.28	33.47	0.00	0.00	0.00	NA	
8	Ahmedabad	2015-01-08	NA	NA	80.87	36.74	96.75	NA	80.87	38.54	31.89	0.00	0.00	0.00	NA	
9	Ahmedabad	2015-01-09	NA	NA	29.16	31.00	48.00	NA	29.16	58.68	25.75	0.00	0.00	0.00	NA	
10	Ahmedabad	2015-01-10	NA	NA	NA	7.04	0.00	NA	NA	8.29	4.55	0.00	0.00	0.00	NA	
11	Ahmedabad	2015-01-11	NA	NA	132.07	55.80	24.53	NA	132.07	25.03	6.79	0.00	0.00	0.00	NA	
12	Ahmedabad	2015-01-12	NA	NA	52.04	40.67	90.24	NA	52.04	51.84	45.89	2.41	0.03	7.88	NA	
13	Ahmedabad	2015-01-13	NA	NA	48.82	44.20	87.09	NA	48.82	68.21	35.16	9.45	13.35	12.50	NA	
14	Ahmedabad	2015-01-14	NA	NA	19.20	27.86	33.05	NA	19.20	52.65	20.96	2.16	2.26	5.19	NA	
15	Ahmedabad	2015-01-15	NA	NA	0.60	16.96	16.80	NA	0.60	28.89	47.83	0.14	0.04	1.35	NA	
16	Ahmedabad	2015-01-16	NA	NA	1.63	21.72	22.86	NA	1.63	38.27	46.03	0.35	0.05	2.01	NA	
17	Ahmedabad	2015-01-17	NA	NA	11.44	24.73	34.75	NA	11.44	49.50	52.24	0.68	0.00	3.27	NA	
18	Ahmedabad	2015-01-18	NA	NA	6.10	25.77	29.57	NA	6.10	48.41	53.49	0.74	0.21	2.75	NA	

Showing 1 to 19 of 29,531 entries, 16 total columns

Fig.1 View of raw importer data

2.3 Data processing

After collecting the data, we need to process that data and organize it for further analysis with data analysis and visualization software or programing languages. In our case data is already organised.

2.4 Cleaning of data

Data should be clean for analysis and visualization. In R some functions used for data cleaning For example "cleandata=dataframe$column[!is.na(dataframe $column)]" this line of code retrieve the without NA values into "cleandata" variable or we can say remove null values from column of dataframe and store clean data into "cleandata" variable.

2.5 Manipulating data

Data manipulation means to filter or rearrange the data for desire output.

In this paper, the filtered data has been mentioned below—

```
install.packages('dplyr')  #install dplyr package
library(dplyr)  #load dplyr package
Fdata<-
select(data,c('City','Date','PM2.5','PM10','O3','AQI'))  #select useful data
```

	City	Date	PM2.5	PM10	O3	AQI
1	Ahmedabad	2015-01-01	NA	NA	133.36	NA
2	Ahmedabad	2015-01-02	NA	NA	34.06	NA
3	Ahmedabad	2015-01-03	NA	NA	30.70	NA
4	Ahmedabad	2015-01-04	NA	NA	36.08	NA
5	Ahmedabad	2015-01-05	NA	NA	39.31	NA
6	Ahmedabad	2015-01-06	NA	NA	46.51	NA
7	Ahmedabad	2015-01-07	NA	NA	33.47	NA
8	Ahmedabad	2015-01-08	NA	NA	31.89	NA
9	Ahmedabad	2015-01-09	NA	NA	25.75	NA
10	Ahmedabad	2015-01-10	NA	NA	4.55	NA
11	Ahmedabad	2015-01-11	NA	NA	6.79	NA
12	Ahmedabad	2015-01-12	NA	NA	45.89	NA
13	Ahmedabad	2015-01-13	NA	NA	35.16	NA
14	Ahmedabad	2015-01-14	NA	NA	20.96	NA
15	Ahmedabad	2015-01-15	NA	NA	47.63	NA
16	Ahmedabad	2015-01-16	NA	NA	46.03	NA
17	Ahmedabad	2015-01-17	NA	NA	52.24	NA
18	Ahmedabad	2015-01-18	NA	NA	51.49	NA

Showing 1 to 19 of 29,531 entries, 6 total columns

Fig 2. View of data after process

2.6 Analysing data

Extracting useful information from data is known as data analysis.

If we want to analyse Pm 10 parameter by mean and compare all cities which we have then first, we need to categorize the names of Cities from abundant data. After that we have to find mean of parameter. Process has been described in below code.

```
# create new vector for cities mean data frame
cities_v<-c()
Average_PM10<-c()
Average_PM2.5<-c()
Average_AQI<-c()
Average_o3<-c()

# calculate mean by cities
for (c in cities){
  m_PM10<-mean(Fdata$PM10[Fdata$City==c &
!is.na(Fdata$PM10)])
  m_PM2.5<-mean(Fdata$PM2.5[Fdata$City==c &
!is.na(Fdata$PM2.5)])
  m_AQI<-mean(Fdata$AQI[Fdata$City==c &
!is.na(Fdata$AQI)])
  m_o3<-mean(Fdata$O3[Fdata$City==c &
!is.na(Fdata$O3)])


  cities_v<-append(cities_v,c)
  Average_PM10<-append(Average_PM10,m_PM10)
  Average_PM2.5<-append(Average_PM2.5,m_PM2.5)
  Average_AQI<-append(Average_AQI,m_AQI)
  Average_o3<-append(Average_o3,m_o3)
}

# create data frame for store mean by cities in ascending
order
city_PM10<-data.frame(
  city=cities_v[order(Average_PM10)],

Average_PM10=Average_PM10[order(Average_PM10)
]
)
```

```
city_PM2.5<-data.frame(
  city=cities_v[order(Average_PM2.5)],

Average_PM2.5=Average_PM2.5[order(Average_PM2.
5)]
)

city_AQI<-data.frame(
  city=cities_v[order(Average_AQI)],
  Average_AQI=Average_AQI[order(Average_AQI)]
)

city_o3<-data.frame(
  city=cities_v[order(Average_o3)],
  Average_o3=Average_o3[order(Average_o3)]
)

View(city_PM10)
```

	city	Average_PM10
1	Aizawl	23.35286
2	Coimbatore	37.98214
3	Shillong	42.19917
4	Ernakulam	48.33563
5	Thiruvananthapuram	52.86366
6	Chennai	62.83440
7	Kochi	67.33543
8	Amaravati	76.36202
9	Bengaluru	83.24329
10	Chandigarh	85.65655
11	Hyderabad	91.93153
12	Mumbai	96.74506
13	Visakhapatnam	105.95135
14	Ahmedabad	114.58403
15	Amritsar	115.20260
16	Kolkata	115.62950
17	Guwahati	116.60490
18	Bhopal	119.32432

Fig 3. Pm 10 levels in ascending order

2.7 Data Visualization

Boxplot and bar graph have been used to visualize air quality data of India. Because we need to plot mean value and compare according to city that's why we use barplot. Moreover, we have sorted data of all parameters and as a result, understanding the bar plot graph seems not to be a demanding task.. Box plot has been used for further explaining the details.

2.7.1 Visualization of data with barplot:

```
# plot graph of pm 10
barplot(city_PM10$Average_PM10[!is.na(city_PM10$Average_PM10)],

        names.arg = citiy_PM10$city[!is.na(city_PM10$Average_PM10)],
        col = "red",
        main = "Average PM 10 index value",
        las =2,
        cex.names = 0.7)
```

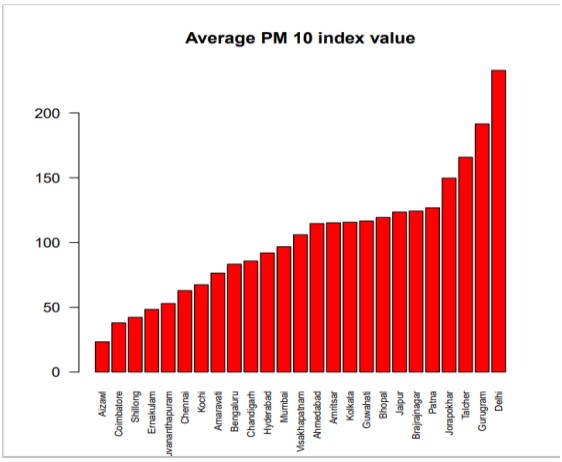

Fig 4. Average pm10 index value

2.7.2 Visualize data with box plot:

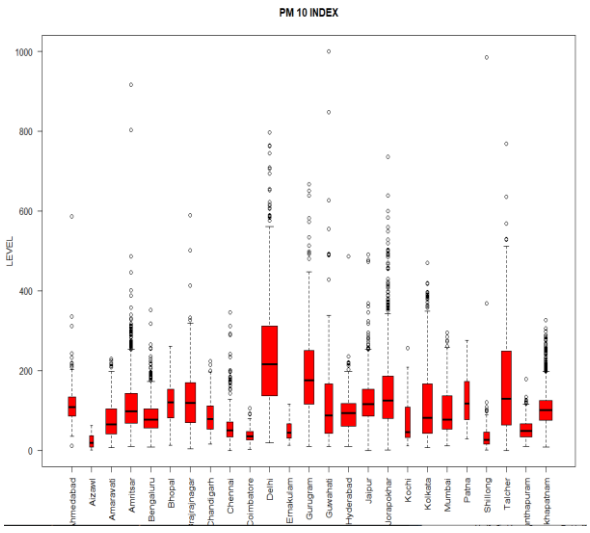

Fig 5. Pm10 index value

```
#PM10 INDEX
plt10<-boxplot(
  Fdata$PM10~Fdata$City,    #we need to replace only pm10 to other index name for
            #further visualization
  varwidth = T,
  main="PM 10 INDEX",  #we also need to change main name according to above
            #index value selected
  notch=F,
  col="Red",
  xlab = "",
  ylab = "LEVEL",
  las=2,
)
text(plt10,
     par("usr")[3],
     labels = row.names(data$City),
     srt = 45,
     adj = c(1.1,1.1),
     xpd = TRUE,
     cex=0.5
)
```

AQI Category	AQI	PM 10	PM 2.5	O_3
GOOD	0 -50	0-50	0-30	0-50
SATISFACTORY	51 - 100	51 - 100	31 – 60	51-100
MODERATELY POLLUTED	101 - 200	101 – 250	61 - 90	101-168
POOR	201 – 300	251 - 350	91 – 120	169-208
VERY POOR	301 – 400	351 - 430	121 – 250	209-748
SEVERE	401+	430+	250+	748+

Fig. 6. Health concern with index values [8]

3. Particulate Matter 2.5 (PM 2.5):

Effects: "Particles in the PM2.5 size range easily travel into lung. It leads to some deleterious effects on health affecting body parts such as nose, eye, throat and breeding ailments like lung irritation, coughing, sneezing, runny nose and shortness of breath." [7]

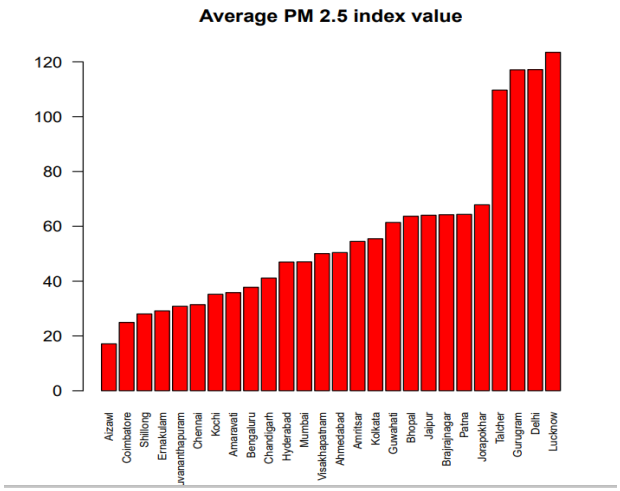

Fig 7. Pm2.5 index value

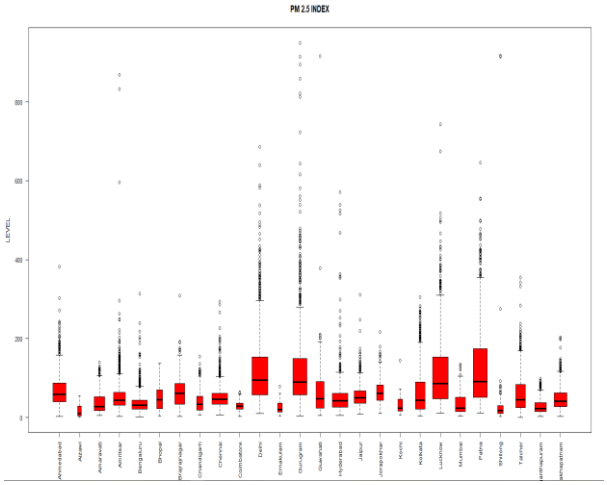

Fig. 8. This graph describes air quality with PM 2.5 index

Analysis based on PM 2.5 index:

In fig. 7 and fig. 8 Lucknow , Delhi, Gurugram, Talcher . These are those 4 places that have a high pm 2.5 level. These cities are in a very poor stage.

Aizawl, Ernakulam, Thiruvananthapuram, Coimbatore, Shillong, Kochi, Mumbai, Bengaluru, Amaravati, Chandigarh are in Moderate stage and are good for those people who are suffering from lungs related issues.

4. Air quality index(AQI):

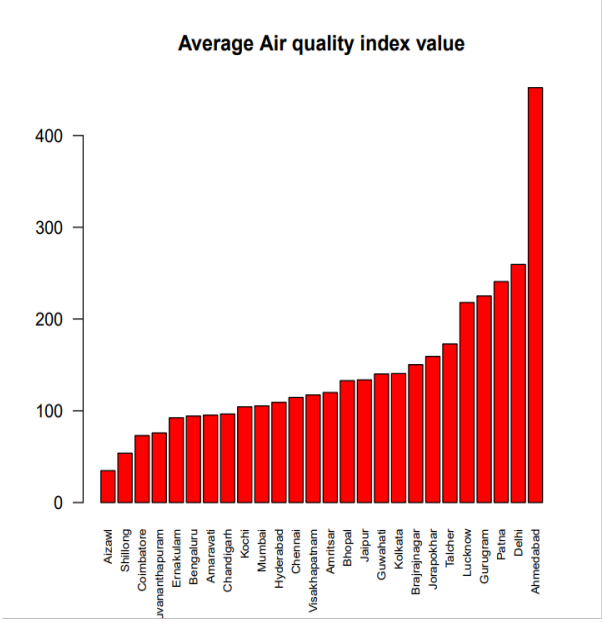

Fig. 9 Air quality index value

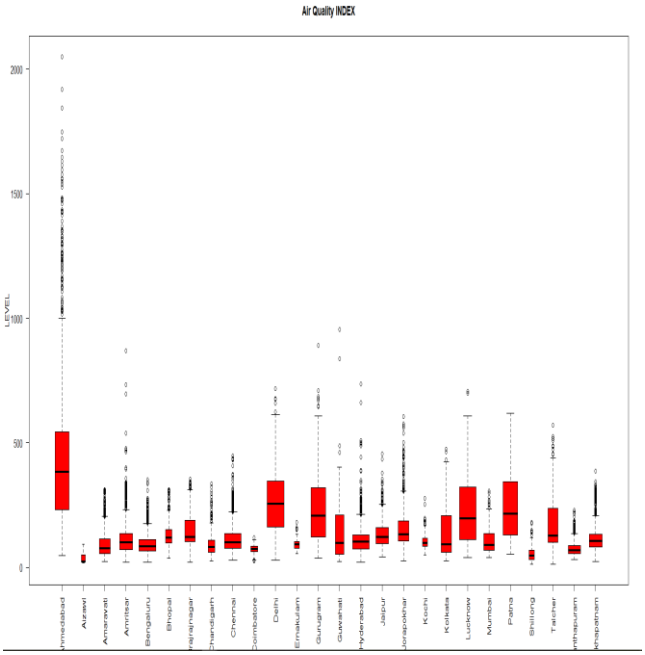

Fig. 10. This graph describes the overall air quality

Analysis of air quality based on Air quality index(AQI):

It is conspicuous from fig. 9 and fig. 10 that Ahmedabad is at a severe level although such time AQI level rich up to 2000 it is hazardous for the health of people who reside in that locality. Delhi, Patna, Gurugram, Lucknow, etc. are in

very poor level and pathetic condition of air inclines the number of health issues.

Aizawl, Shillong, Coimbatore, Thiruvananthapuram, Ernakulam, Bengaluru, Amaravati, Chandigarh, Kochi, Mumbai are in moderate category so these cities are considerable for living.

5. Particulate Matter 10 (PM 10):

Effect: "Ultra-fine dust is the most damaging variant of fine particles because the particles penetrate directly through the lungs into the bloodstream and,thereafter, spread into the organs." [7]

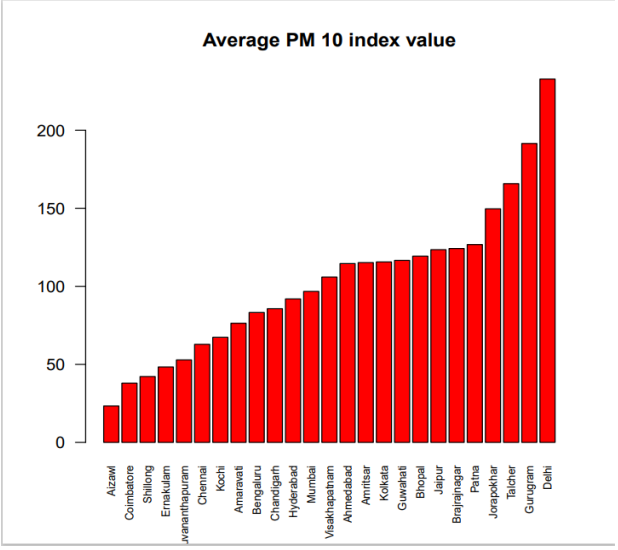

Fig 11. Pm10 index value

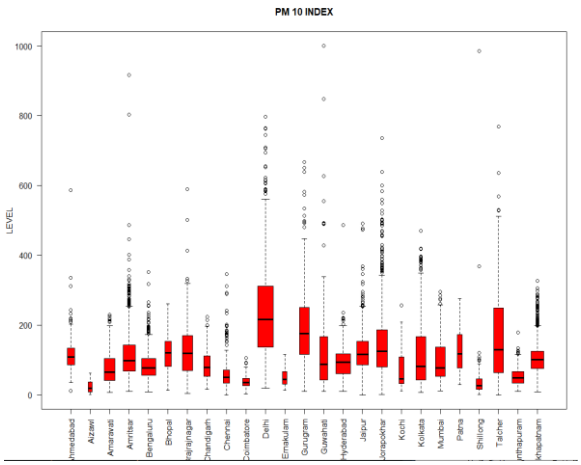

Fig. 12. This graph describes air quality based on PM 10

Analysis of air quality based on Pm10 index:

It is pellucid from fig. 11 and fig. 12 that Delhi, Gurugram, Talcher, Jorapokhar, Patna etc. have higher level of pm10 and all these cities are in a Moderate level. Thiruvananthapuram, Chennai, Kochi, Amaravati, Bengaluru, Chandigarh, Hyderabad, Mumbai come under satisfactory stage whilst Aizawl, Coimbatore, Shillong, Ernakulam are in good stage.

6. Ozone (O3):

"Ozone, O3, is a pale blue gas that is unstable and obnoxious. It is a powerful allotropic of the element oxygen; high concentration of ozone around the earth absorbs heat. Ozone is a molecule containing three oxygen atoms; and, it is unstable and harmful on ground level." [7]

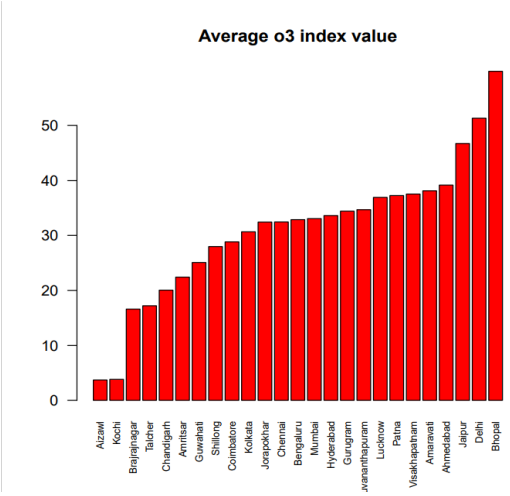

Fig. 13 ozone gas lavels

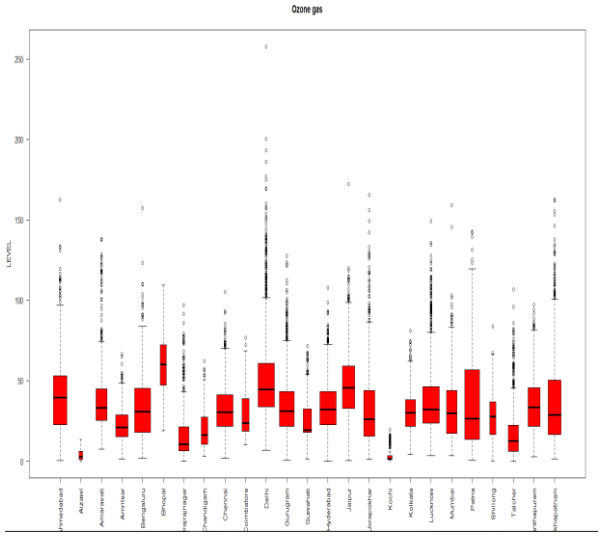

Fig. 14. This graph describes ozone levels at ground

Analysis of air quality based on ozone on the ground:

It is observed from fig.13 and fig. 14 that Bhopal, Delhi, Jaipur, Ahmedabad, Amaravati have higher ozone levels on the ground which is bad for these cities.

Aizawl, Kochi, Brajrajnagar, Talcher, Chandigarh, Amritsar, Guwahati, Shillong, Coimbatore, Kolkata have less ozone ground level which is beneficial for these cities.

7. Major Causes of air pollution:

There are three main causes of air pollution. First, is the Combustion of fossil fuels and second is Emissions from industries and factories and at last come agricultural activities, use of insecticides, pesticides, and fertilizers that produce harmful chemicals which are fatal for human body.

8. Solution

Mankind ought to give preferably to renewable energy sources such as solar energy, wind energy, Hydropower, Geothermal, and biomass over fossil fuels. In future, vehicle manufacturers should establish more and more electric car charging stations so thaf people can think about purchasing electric cars and government should also give subsidies on electric cars because diesel vehicles produce very harmful gases which can claim human lives. Additionally, industries should use natural gas instead of coal or oil and farmers should practice organic farming and lastly, government should provide information or support to farmers for organic farming.

Results

In this paper, we have discussed India's air quality data from the year 2015 to 2020 and have observed that Ahmedabad and Delhi is the most polluted cities in India. On other hand, Aizawl, Shillong, Coimbatore, Thiruva-nanthapuram, Ernakulam, Bengaluru, Amaravati, Chandigarh, etc. are having better air quality than other remaining cities which have been included in data visualization and all these places are good for living.

Discussions

Data always develops at an exponential rate. We can make lots of modifications in data and make comparisons to obtain valuable outcomes. Data visualization provides us a better understanding of data. Based on that, we can make the best decision for business or individual goals.

Conclusions

To curtail the air pollution, we should prefer electric vehicles over diesel or petrol vehicles because diesel is the major cause of air pollution in India. Similarly, car manufacturers ought to emphasize on decreasing manufacturing cost of electric cars and also reduce the production cost of the lithium-ion battery which will make electric cars affordable.

REFRENCES:

[1] Garrett Grolemund , Hands-On programming with R. 2014
[2] Rob Kabacoff , Data Visualization with R, 2018
[3] Ghosh, Debosree and Parida, Pratap, Air Pollution and India: Current Scenario, 2015
[4] Terry Gordon,a Kalpana Balakrishnan,b Sagnik Dey,c Sanjay Rajagopalan , Air pollution health research priorities for India: Perspectives of the Indo-U.S. Communities of Researchers, 2018
[5] Girish Agrawal, Dinesh Mohan, Hifzur Rahman, Ambient air pollution in selected small cities in India: Observed trends and future challenges, 2021
[6] https://www.kaggle.com/rohanrao/air-quality-data-in-india
[7] https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health
[8] https://pib.gov.in/newsite/printrelease.aspx?relid=110654
[9] Ravi Yadav and Pushpendra Vyas, Particulate Matter Pollution in Urban Cities of India During Unusually Restricted Anthropogenic Activities, 2022
[10] Sunil Gulia and Nidhi Shukla, Evolution of air pollution management policies and related research in India, 2022