Contents

1 Introduction

Over the last few decades, inhabitants of cities worldwide have gotten more concerned about air pollution. The two most polluted cities in the world are New Delhi, the capital of India and Kolkata. According to the World Air Quality Report prepared by IQAir on pollution at particulate levels, Delhi was the fourth most polluted city in the world in 2022 out of 50 cities and Kolkata is the second most polluted city in the world, according to the State of Global Air report for 2022. The large population and rapid economic expansion are the major causes of the high levels of pollution. Typically, Delhi's air gets worse in October and November (winter) near Diwali season and gets better by March and April (summer).and Kolkata is also mostly polluted in October during the big fest of West Bengal Durga Pooja.

The present study analyzed the air pollution of two Indian megacities: Delhi, Kolkata's last two consecutive season, the winter(October, November, December) season of year 2022 and summer(April, May, June) season 2023. Six pollutants are considered PM2.5, PM10, NOx, SO2, Toluene, CO.

2 Data Description

The data set contains air pollution data from two cities for the summer and winter seasons. We have downloaded the pollutant concentrations from the Central Pollution Control Board website and meteorological data from LARC NASA

Location Information:

1. Latitude - 22.5726 Longitude - 88.3639
   Station Name - Ballygunge, Kolkata - WBPCB
   City - Kolkata

2. Latitude - 28.6791 Longitude - 77.0697
   Station Name - Anand Vihar, Delhi - DPCC
   City - Delhi

Time Information:

1. Winter: From Date: 1st October 2022
   To Date: 31st December 2022

2. Summer: From Date: 1st April 2023
   To Date: 30th June 2023

Feature Information:
Here we have taken various pollutants such as

1. Particulate Matter(PM2.5, PM10)($\mu g/m^3$)
2. Sulfur Dioxide(SO2)($\mu g/m^3$)
3. Nitrogen Oxide(NOx)($\mu g/m^3$)
4. Carbon Monoxide(CO)(mg/m^3)
5. Toluene($\mu g/m^3$)
6. Relative Humidity(RH)(%)
7. Wind Speed(WS)(m/sec)

8. Wind Direction(WD)($degrees$)
9. Temperature in 2M range(T2M)($^{\circ}C$)
10. Precipitation Corrected(PRECTOTCORR) (mm/day)

3 Data Preprocessing

If a significant portion of the data is missing, it may lead to biased or inaccurate results in subsequent analyses. So Handling of missing values is very important. There is some missing values in features like Toluene and CO. So the missing values are replaced by median of the remaining data.

Also there are some features in the data in which most of the data are missing so these columns are dropped.

Also there are so many similar features in the data like temperature in different ranges so we kept only one of them and others are dropped.

We have used the data of two cities of two seasons. So we labeled the data according to city and season which was useful for different plottings for different cities and seasons.

4 Data Visualisation:

4.1 Box Plots:

A box plot is a standardized way of displaying the distribution of the dataset based on its five-number summary of the data points: "minimum", "first quartile(Q1)", "median (second quartile(Q2))", "third quartile(Q3)" and "maximum."

Now, we have drawn box plots of the parameters of the given dataset and the box plots are given below:

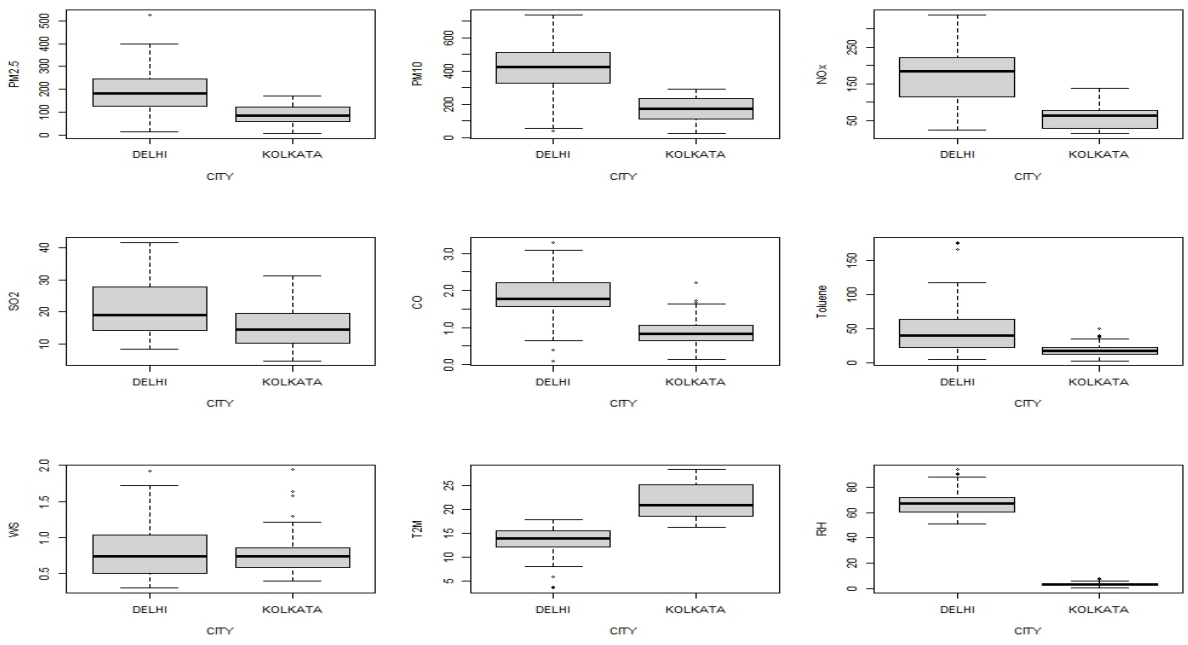

Figure 1: Comparing PM2.5, PM10, NOx, SO2, CO, Toluene, Wind Speed, Temperature, Relative Humidity between Delhi and Kolkata in Winter Season

In the above box plots, we are comparing air pollutants like PM2.5, PM10, NOx, SO2, CO, Toluene, WS, T2M, and RH with respect to the two cities between Delhi and Kolkata in the Winter Season. From the box plots, When comparing the air pollutants like PM2.5, PM10, NOx, SO2, CO, Toluene, T2M and RH, we observed that Delhi has consistently higher median air pollutants than Kolkata. From this observation, it suggests that, on average, the air quality in Delhi is worse.

Similarly,

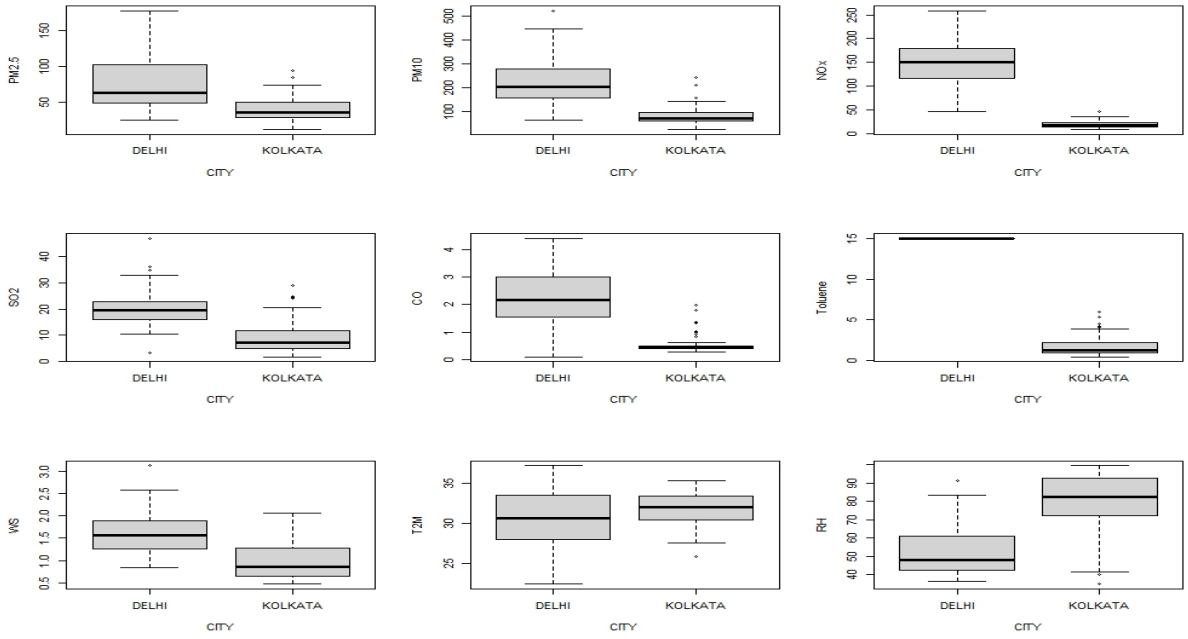

Figure 2: Comparing PM2.5, PM10, NOx, SO2, CO, Toluene, Wind Speed, Temperature, Relative Humidity between Delhi and Kolkata in Summer Season

In the above box plots, we are comparing air pollutants like as PM2.5, PM10, NOx, SO2, CO, Toluene, WS, T2M, RH with respect to the two cities between Delhi and Kolkata in Summer Season. When comparing the air pollutants like PM2.5, PM10, NOx, SO2, CO, Toluene, WS and RH, we observed that Delhi has consistently higher median air pollutants than Kolkata. From this observation, it suggests that, on average, the air quality of Delhi is worse.

But, when we compare with respect to T2M and RH, Kolkata has a consistently higher median level than Delhi.

4.2 Density

In this analysis, we utilize histogram density plots to visually depict the shape of the distribution across eight different graphs. These graphs correspond to both cities and various seasons, focusing on the levels of PM2.5 and PM10 pollutants. In each plot, we examine the distribution of PM2.5 and PM10 concentrations, considering the unique characteristics presented in different seasons within each city. The use of histogram density plots allows for a comprehensive visualization of the data distribution in these diverse contexts.

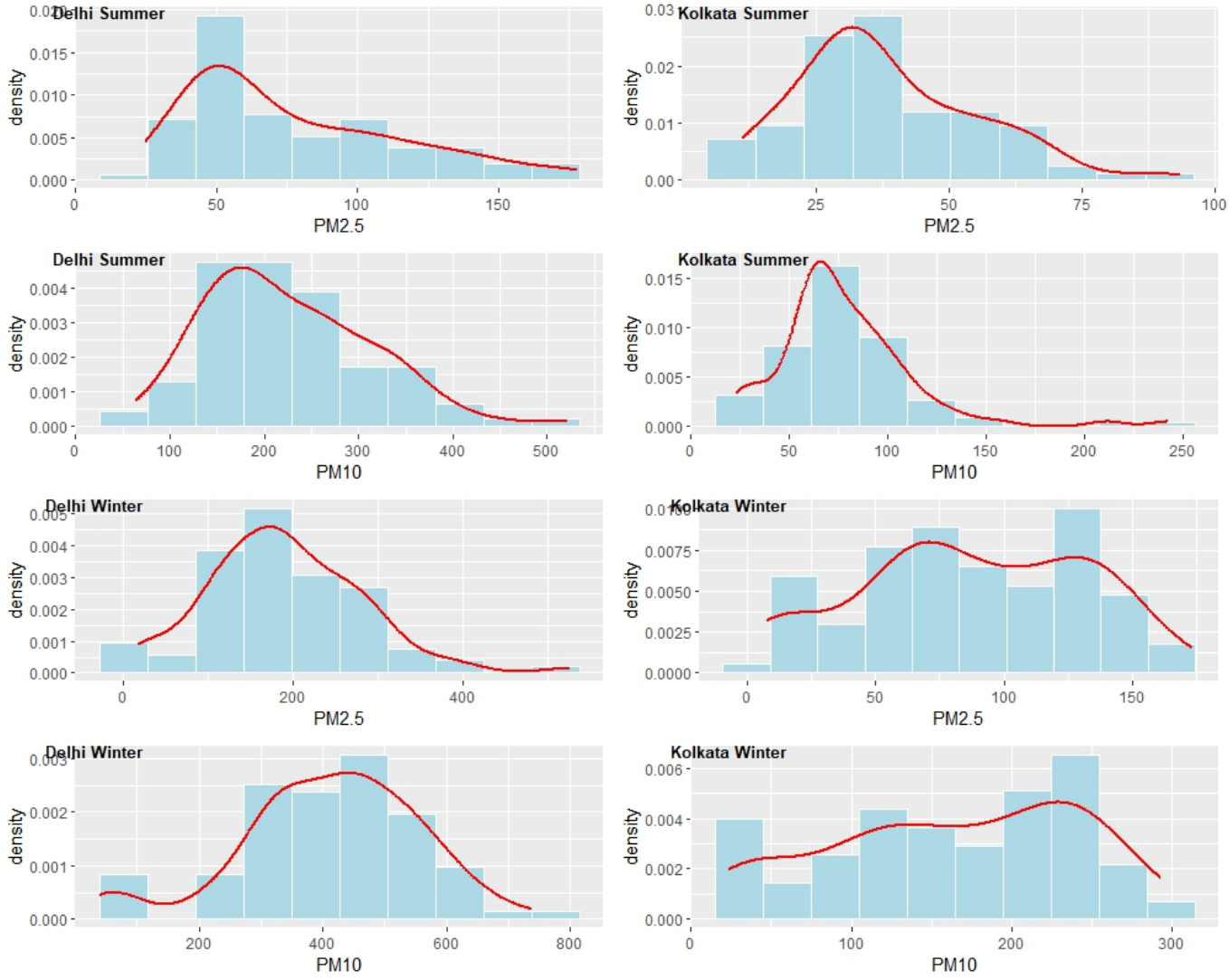

Figure 3: Histogram plots of PM2.5 and PM10 for Summer and Winter Seasons of Delhi and Kolkata

We see that the data points are symmetrically distributed around the mean, resembling the characteristic bell-shaped curve indicative of a normal distribution.

5 Hypothesis Testing

From the box plots in figure1 and figure2, we can see that Kolkata appears to have less PM2.5 and PM10 than Delhi. However, the graphical representation doesn't tell us if it is statistically significant or not. So we will do hypothesis testing to check that.

Why did we choose the t-test?

The t-test is used to check if there is a significant difference between the means of the two groups. This test is useful when two independent samples follow a normal distribution, the sample size is small and the population standard deviation is unknown. Here from the Histogram and Density plots in figure3, it is clear that PM2.5 and PM10 are following normal distribution.
We have the sample data of PM2.5 and PM10 of two cities that are Kolkata and Delhi. We will compare the means of PM2.5 and PM10 in the summer and winter of Kolkata and Delhi. There are different types of t-tests, here we will use two-sample t-tests.

5.1 Hypothesis:

Null Hypothesis(H0): Assume there is no significant difference between the mean of PM2.5 of two cities in summer i.e. $\mu_{sd} = \mu_{sk}$

Alternative Hypothesis(H1): Assume that mean of PM2.5 of Delhi is more than Kolkata i.e. $\mu_{sd} > \mu_{sk}$.

In this test, the p-value is 2.478e-15 which is very low. As the p-value is small (typically less than the chosen significance level, often 0.05) we have evidence to reject the null hypothesis. It indicates Delhi has significantly more PM2.5 than Kolkata in summer.

Null Hypothesis(H0): Assume there is no significant difference between the mean of PM2.5 of two cities in winter that is $\mu_{wd} = \mu_{wk}$.

Alternative Hypothesis(H1): Assume that mean of PM2.5 of Delhi is more than Kolkata that is $\mu_{wd} > \mu_{wk}$.

In this test, the p-value is less than 2.2e-16 which is very low. As the p-value is small (typically less than the chosen significance level, often 0.05) we have evidence to reject the null hypothesis. It indicates Delhi has significantly more PM2.5 than Kolkata in winter.

Similarly, we have done hypothesis testing of PM10 for summer and winter between Delhi and Kolkata. In both cases, the null hypothesis is rejected i.e. Delhi Air has significantly more PM10 than Kolkata.

6 Model Evaluation and Residual Analysis for Predicting PM2.5 and PM10 Levels:

From the dataset, we have calculated the correlation matrix, was given in the AppendixA.1, from there one can check that PM2.5 and PM10 both are highly correlated with the factors of air such as NOx, SO_2, CO, Toluene, Wind Speed, Relative Humidity, Temperature. After analyzing the multiple regression models, we got the results which were given in the AppendixB.2.

Model 1 PM2.5:

The linear regression model for predicting PM2.5 levels demonstrates a notable range of residuals, spanning from -125.620 to 159.006. This range reflects the variance between predicted and observed values, indicating instances where the model both overpredicts and underpredicts PM2.5 concentrations. The residuals, as calculated differences between observed and predicted values, underscore the inherent variability in the model's predictive accuracy across the dataset.

NOx, Toluene, and RH are statistically significant predictors with p-values < 0.05. T2M(Temperature at 2 Meters) and PRECTOTCORR(Corrected Total Precipitation) are also significant.

Adjusted R-squared is 0.7695 suggests that the model explains about 76.95% of the variance in the response variable.

Model 2 PM10:

In a similar vein, the residuals of the linear regression model for PM10 levels vary, ranging from -247.943 to 244.963. This range highlights the predictability gaps in the model, exposing situations where PM10 concentrations are both overestimated and underestimated. The distribution of residuals illustrates how different prediction errors might be from one another and shows how well the model can capture the subtleties of the observed data.

NOx, Toluene, WS, RH, T2M, and PRECTOTCORR are statistically significant predictors with p-values < 0.05.

Adjusted R-squared is 0.8175 suggests that the model explains about 81.75% of the variance in the

response variable.

6.1 Comparative Analysis:

In comparing the two models, the wider range of residuals in Model 2 suggests a potentially higher degree of variability in its predictive accuracy for PM10 levels. Model 1, with its narrower range, may indicate a more consistent predictive performance for PM2.5 levels. However, the interpretation of these ranges should be tempered with consideration for the specific context and variability inherent in environmental datasets.

6.2 Prediction Analysis:

Now, using test data during predicting PM2.5 Levels we have got:

R^2	RMSE	MAE
0.718475	38.30676	28.62038

Approximately 71.18% of the variability in PM2.5 levels is explained by the model. On average, the model's predictions deviate by approximately 38.30676 units from the observed PM2.5 values. On average, the absolute difference between predicted and observed PM2.5 values is approximately 28.62038 units.
And using test data during predicting PM10 Levels we have got:

R^2	RMSE	MAE
0.7293892	79.6395	56.42672

Approximately 72.93% of the variability in PM10 levels is explained by the model. On average, the model's predictions deviate by approximately 79.6395 units from the observed PM10 values. On average, the absolute difference between predicted and observed PM10 values is approximately 56.42672 units.
In summary, these metrics suggest that Model 1 performs better than Model 2 in terms of explaining variability and producing more precise forecasts for the dependent variables that are assumed (PM2.5 and PM10).

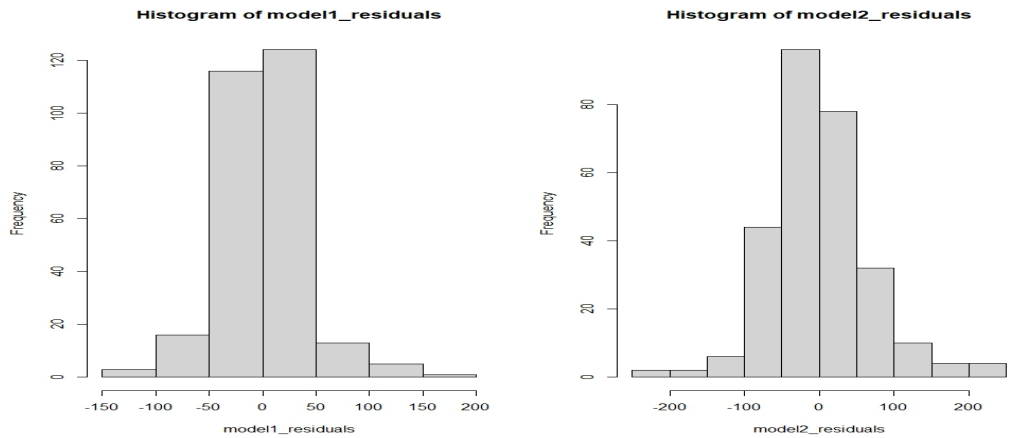

Figure 4: Histogram plot of Residuals for Model 1 and Model 2

7 Conclusion

In this study, the pollutant concentration of Delhi city was higher in comparison to Kolkata in both seasons, but season-wise, the air quality was worse in the winter season for all pollutants PM2.5, PM10, NOx, SO2, CO, and Toluene. From the box plot, the seasonal analysis shows the concentration dispersion depends upon various meteorological conditions like WS, Temp, RH, precipitation, and wind direction.

The relationship between pollutant concentration and meteorology was found to have a strong negative correlation, which shows that Delhi meteorology is responsible for the higher pollution in winter due to lower temperature and wind speed as compared to Kolkata. Kolkata is the coastal reason while Delhi is surrounded by the hills of Aravali and the Himalayas and wind speed is lower as well, and the lower mixing height in winter results improper dispersion takes place and in winter season, the most polluted city in India.

8 References

1. Mahanta, Soubhik, et al. "Urban air quality prediction using regression analysis." TENCON 2019-2019 IEEE Region 10 Conference (TENCON). IEEE, 2019.

2. Pal, Vijay, Surinder Deswal, and Mahesh Pal. "Spatial Variation of Air Quality in Delhi During Diwali: A Case Study of Covid-19 Period." International Conference on Advancements in Interdisciplinary Research. Cham: Springer Nature Switzerland, 2022.

3. Keywood, M. D., et al. "Relationships between size segregated mass concentration data and ultrafine particle number concentrations in urban areas." Atmospheric Environment 33.18 (1999): 2907-2913.

4. Rice, John A. Mathematical statistics and data analysis. Cengage Learning, 2006.

5. Power LARC NASA. https://power.larc.nasa.gov/data-access-viewer/. Accessed 10 November 2023

6. CPCB. http://www.cpcbenvis.nic.in. Accessed 10 November 2023

A Summary of all the features

Attribute	Min	1st Qu.	Median	Mean	3rd Qu.	Max
PM2.5	16.38	126.14	181.03	189.05	244.60	525.00
PM10	39.18	325.55	422.26	410.87	505.17	738.00
NOX	23.23	115.03	184.69	172.44	221.91	337.10
SO2	8.29	14.41	18.96	21.18	27.88	41.79
CO	0.100	1.570	1.780	1.845	2.190	3.300
Toluene	5.22	22.12	39.85	46.86	64.17	176.70
RH	50.91	60.06	67.00	67.42	71.66	93.83
WS	0.3000	0.5075	0.7400	0.8095	1.0350	1.9200
WD	1.75	96.03	171.19	151.87	218.34	353.00
PRECTOTCORR	0.00	0.00	0.00	1.04	0.00	56.75
T2M	3.48	12.18	13.94	13.61	15.54	17.79

Table 1: Delhi Winter

Attribute	Min	1st Qu.	Median	Mean	3rd Qu.	Max
PM2.5	24.74	48.07	62.90	77.51	102.12	177.97
PM10	63.45	157.48	201.82	224.97	278.26	521.36
NOX	46.49	115.72	150.90	147.03	179.19	258.17
SO2	3.13	15.93	19.54	20.06	22.92	47.07
CO	0.100	1.555	2.170	2.257	2.985	4.400
Toluene	15.04	15.04	15.04	15.04	15.04	15.04
RH	36.58	42.39	48.33	52.59	61.10	91.40
WS	0.830	1.250	1.560	1.598	1.890	3.120
WD	61.86	121.02	169.54	163.93	206.14	256.93
PRECTOTCORR	0.00	0.00	0.12	2.22	2.42	22.15
T2M	22.47	28.05	30.65	30.53	33.55	37.22

Table 2: Delhi Summer

Attribute	Min	1st Qu.	Median	Mean	3rd Qu.	Max
PM2.5	7.82	60.13	86.06	87.61	124.90	173.08
PM10	23.61	109.68	170.13	163.18	231.43	292.80
NOX	14.86	29.44	62.56	57.64	77.89	138.83
SO2	4.74	10.34	14.55	15.09	19.37	31.35
CO	0.1300	0.6500	0.8300	0.8472	1.0450	2.2200
Toluene	2.88	12.43	17.08	18.04	22.25	49.48
RH	0.170	2.138	3.145	3.045	3.970	8.130
WS	0.4000	0.5950	0.7400	0.7668	0.8600	1.9400
WD	42.39	63.85	76.67	75.51	85.75	99.45
PRECTOTCORR	0.0000	0.0000	0.0000	1.5059	0.0975	19.8900
T2M	16.21	18.66	20.84	21.80	25.07	28.30

Table 3: Kolkata Winter

Attribute	Min	1st Qu.	Median	Mean	3rd Qu.	Max
PM2.5	11.18	27.83	34.76	38.50	50.05	93.52
PM10	23.10	60.38	69.66	78.04	93.22	242.36
NOX	8.74	14.49	17.62	19.79	23.45	46.40
SO2	1.730	4.865	7.220	9.149	11.680	29.150
CO	0.2900	0.4000	0.4400	0.5076	0.4950	1.9700
Toluene	0.420	0.880	1.230	1.624	2.135	5.940
RH	3.96	72.05	82.54	80.54	92.72	99.46
WS	0.4700	0.6500	0.8500	0.9758	1.2850	2.0600
WD	137.2	156.2	186.5	197.5	229.3	298.2
PRECTOTCORR	0.000	0.015	0.490	4.064	4.535	64.230
T2M	25.87	30.50	32.05	31.89	33.45	35.30

Table 4: Kolkata Summer

A.1 Correlation Matrix

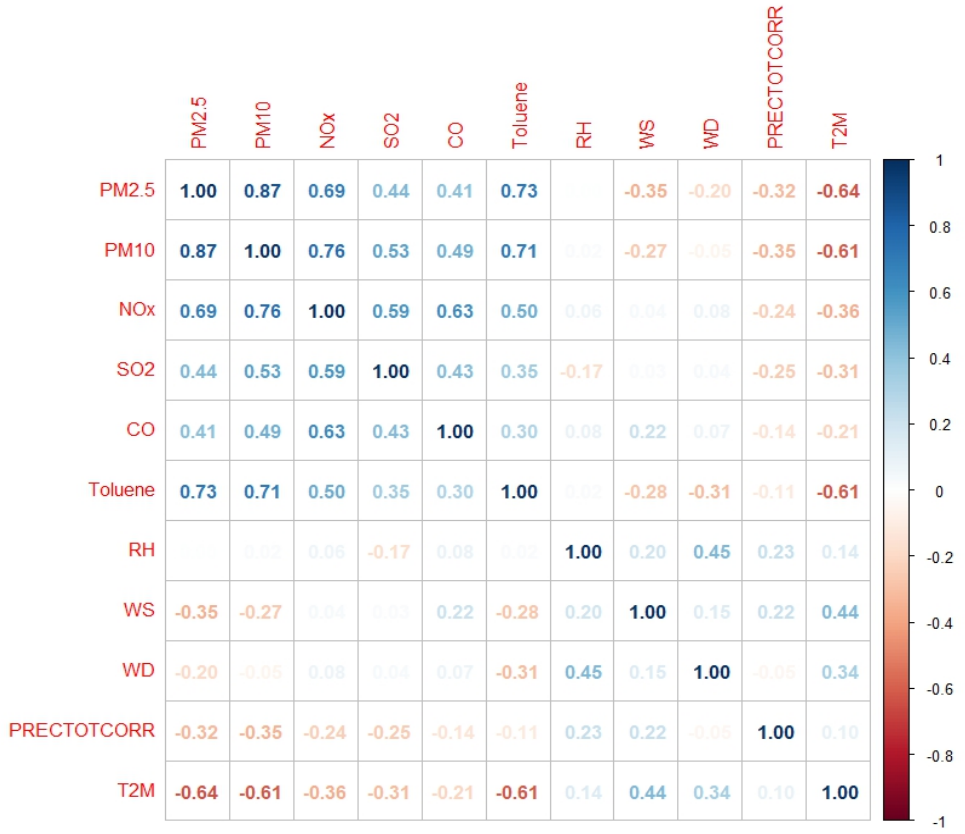

Figure 5: Correlation Matrix between all features

B R Programming Code used for the project work

```
library(tidyverse) #for transforming datas
library(corrplot) #for correlation plot
library(ggplot2) #for plotting histogram and density plots
library(cowplot) # for multiple plot in a grid format in ggplot
library(caret) # for prediction analysis
```

```r
#Reading data

air<-read.csv("FINAL_DATASET.csv")
print(air)
dim(air)
names(air)
#Checking which columns have how many missing values
mis<-colSums(is.na(air))
print(mis)
#Removing the features which are not useful for this analysis purpose
air<-subset(air,select = -c(Temp,NH3,Ozone,NO2,RF,TOT.RF,NO,RH2M,WS10M,WD10M,
            Benzene,SR,BP,AT))
#Replacing NA by median value of the corresponding feature vector
air<- air %>% mutate(across(c(SO2,CO,RH,WS,WD,PRECTOTCORR,T2M,PM2.5,PM10,NOx,Toluene),
            ~replace_na(., median(., na.rm=TRUE))))
#Printing first five and last five rows of the datasets
head(air)
tail(air)
air$CITY[air$CITY==1]<-'DELHI'
air$CITY[air$CITY==0]<-'KOLKATA'

#Making different subsets of dataset according to seasons and cities
winter<-subset(air,(SEASON=='W'))
summer<-subset(air,(SEASON=='S'))
delhi<-subset(air,(CITY=='DELHI'))
kolkata<-subset(air,(CITY=='KOLKATA'))
dwinter<-subset(air,(CITY=='DELHI' & SEASON=='W'))
dsummer<-subset(air,(CITY=='DELHI' & SEASON=='S'))
kwinter<-subset(air,(CITY=='KOLKATA' & SEASON=='W'))
ksummer<-subset(air,(CITY=='KOLKATA' & SEASON=='S'))
#Summary of different features of Kolkata and Delhi in summer and winter.
summary(dwinter)
summary(dsummer)
summary(kwinter)
summary(ksummer)

#Making box plot of all features of two cities in summer
par(mfrow=c(3,3))
boxplot(PM2.5 ~ CITY,data=summer)
boxplot(PM10 ~ CITY,data=summer)
boxplot(NOx ~ CITY,data=summer)
boxplot(SO2 ~ CITY,data=summer)
boxplot(CO ~ CITY,data=summer)
boxplot(Toluene ~ CITY,data=summer)
boxplot(WS ~ CITY,data=summer)
boxplot(T2M ~ CITY,data=summer)
boxplot(RH ~ CITY,data=summer)
#boxplot(PRECTOTCORR ~ CITY,data=summer)

#Making box plot of all features of two cities in winter
par(mfrow=c(3,3))
boxplot(PM2.5 ~ CITY,data=winter)
boxplot(PM10 ~ CITY,data=winter)
```

```r
boxplot(NOx ~ CITY,data=winter)
boxplot(SO2 ~ CITY,data=winter)
boxplot(CO ~ CITY,data=winter)
boxplot(Toluene ~ CITY,data=winter)
boxplot(WS ~ CITY,data=winter)
boxplot(T2M ~ CITY,data=winter)
boxplot(RH ~ CITY,data=winter)
#boxplot(PRECTOTCORR ~ CITY,data=winter)

# Hypothesis testing of significant difference of PM2.5 and
PM10 between Kolkata and Delhi

t.test(dwinter$PM2.5, kwinter$PM2.5, alternative = "greater", mu = 0,
            paired = FALSE, conf.level = 0.95)
t.test(dsummer$PM2.5, ksummer$PM2.5, alternative = "greater", mu = 0,
            paired = FALSE, conf.level = 0.95)
t.test(dwinter$PM10, kwinter$PM10, alternative = "greater", mu = 0,
            paired = FALSE, conf.level = 0.95)
t.test(dsummer$PM10, ksummer$PM10, alternative = "greater", mu = 0,
            paired = FALSE, conf.level = 0.95)

# Checking the distribution of dataset by using histogram density plot

p1<-ggplot(dsummer, aes(x = PM2.5, y = after_stat(density))) +
  geom_histogram(bins = 10, color = "white", fill = "lightblue") +
  geom_density(color ="red", linewidth = 1)

p2<-ggplot(ksummer, aes(x = PM2.5, y = after_stat(density))) +
  geom_histogram(bins = 10, color = "white", fill = "lightblue") +
  geom_density(color ="red", linewidth = 1)


p3<-ggplot(dsummer, aes(x = PM10, y = after_stat(density))) +
  geom_histogram(bins = 10, color = "white", fill = "lightblue") +
  geom_density(color ="red", linewidth = 1)

p4<-ggplot(ksummer, aes(x = PM10, y = after_stat(density))) +
  geom_histogram(bins = 10, color = "white", fill = "lightblue") +
  geom_density(color ="red", linewidth = 1)

p5<-ggplot(dwinter, aes(x = PM2.5, y = after_stat(density))) +
  geom_histogram(bins = 10, color = "white", fill = "lightblue") +
  geom_density(color ="red", linewidth = 1)

p6<-ggplot(kwinter, aes(x = PM2.5, y = after_stat(density))) +
  geom_histogram(bins = 10, color = "white", fill = "lightblue") +
  geom_density(color ="red", linewidth = 1)

p7<-ggplot(dwinter, aes(x = PM10, y = after_stat(density))) +
  geom_histogram(bins = 10, color = "white", fill = "lightblue") +
  geom_density(color ="red", linewidth = 1)

p8<-ggplot(kwinter, aes(x = PM10, y = after_stat(density))) +
```

```
  geom_histogram(bins = 10, color = "white", fill = "lightblue") +
  geom_density(color ="red", linewidth = 1)
plot_grid(p1, p2,p3, p4, p5,p6, p7, p8,ncol=2,
labels = c('Delhi Summer', 'Kolkata Summer','Delhi Summer', 'Kolkata Summer',
'Delhi Winter','Kolkata Winter','Delhi Winter','Kolkata Winter'), label_size = 10)

#Removing data which are not numeric for correlation matrix
par(mfrow=c(1,1))
air2<-subset(air,select=-c(DATE,CITY,SEASON))
M <- cor(air2)
corrplot(M, method = 'number')

par(mfrow=c(1,2))
#make this example reproducible
set.seed(1)

#use 75% of dataset as training set and 25% as test set
sample <- sample(c(TRUE, FALSE), nrow(air), replace=TRUE, prob=c(0.75,0.25))
train  <- air[sample, ]
test   <- air[!sample, ]

# MULTI LINEAR REGRESSION for PM2.5
model1 = lm(formula = PM2.5~ NOx+SO2+CO+Toluene+WS + RH + T2M + PRECTOTCORR, data = train)
summary(model1)
model1_residuals = model1$residuals
hist(model1_residuals)

predictions1 <- predict(model1, test)

# computing model performance metrics
data.frame( R2 = R2(predictions1, test $ PM2.5),
            RMSE = RMSE(predictions1, test $ PM2.5),
            MAE = MAE(predictions1, test $ PM2.5))

# MULTI LINEAR REGRESSION for PM10

model2 = lm(formula = PM10~ NOx+SO2+CO+Toluene+WS + RH + T2M + PRECTOTCORR, data = train)
summary(model2)
model2_residuals = model2$residuals
hist(model2_residuals)

predictions2 <- predict(model2, test)
# computing model performance metrics
data.frame( R2 = R2(predictions2, test $ PM10),
            RMSE = RMSE(predictions2, test $ PM10),
            MAE = MAE(predictions2, test $ PM10))
```

B.1 Code output for differents hypothesis testings

```
#Hypothesis testing of significant difference of PM2.5 and PM10 between Kolkata and Delhi
>
> t.test(dwinter$PM2.5, kwinter$PM2.5, alternative = "greater", mu = 0,
```

```
            paired = FALSE, conf.level = 0.95)

Welch Two Sample t-test

data:  dwinter$PM2.5 and kwinter$PM2.5
t = 9.6889, df = 132.21, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 84.09542      Inf
sample estimates:
mean of x mean of y
 189.0489   87.6112

> t.test(dsummer$PM2.5, ksummer$PM2.5, alternative = "greater", mu = 0,
            paired = FALSE, conf.level = 0.95)

Welch Two Sample t-test

data:  dsummer$PM2.5 and ksummer$PM2.5
t = 8.9199, df = 124.45, p-value = 2.478e-15
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 31.76628      Inf
sample estimates:
mean of x mean of y
 77.51484  38.50022

> t.test(dwinter$PM10, kwinter$PM10, alternative = "greater", mu = 0,
            paired = FALSE, conf.level = 0.95)

Welch Two Sample t-test

data:  dwinter$PM10 and kwinter$PM10
t = 14.389, df = 139.02, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 219.1858      Inf
sample estimates:
mean of x mean of y
 410.8664  163.1759

> t.test(dsummer$PM10, ksummer$PM10, alternative = "greater", mu = 0,
            paired = FALSE, conf.level = 0.95)

Welch Two Sample t-test

data:  dsummer$PM10 and ksummer$PM10
t = 14.768, df = 117.49, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 130.4319      Inf
sample estimates:
mean of x mean of y
```

```
224.96505  78.03857
```

B.2 Code output for Multi Linear Regression Model 1 for PM2.5

```
Call:
lm(formula = PM2.5 ~ NOx + SO2 + CO + Toluene + WS + RH + T2M +
    PRECTOTCORR, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-125.620  -19.086    0.472   17.451  159.006

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 103.53169   12.77385   8.105 1.87e-14 ***
NOx           0.42942    0.04697   9.141  < 2e-16 ***
SO2          -0.22006    0.35777  -0.615    0.539
CO            3.97789    3.27730   1.214    0.226
Toluene       1.05093    0.12374   8.493 1.39e-15 ***
WS          -30.86490    5.88360  -5.246 3.15e-07 ***
RH            0.19500    0.07962   2.449    0.015 *
T2M          -1.84741    0.41407  -4.462 1.20e-05 ***
PRECTOTCORR  -1.73238    0.38595  -4.489 1.06e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.25 on 269 degrees of freedom
Multiple R-squared:  0.7761,Adjusted R-squared:  0.7695
F-statistic: 116.6 on 8 and 269 DF,  p-value: < 2.2e-16
```

B.3 Code output for Multi Linear Regression Model 2 for PM10

```
Call:
lm(formula = PM10 ~ NOx + SO2 + CO + Toluene + WS + RH + T2M +
    PRECTOTCORR, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-247.943  -33.271   -5.568   27.081  244.963

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 173.71678   22.21988   7.818 1.22e-13 ***
NOx           0.88367    0.08171  10.814  < 2e-16 ***
SO2           1.31495    0.62234   2.113 0.035529 *
CO            7.21929    5.70080   1.266 0.206478
Toluene       1.83981    0.21525   8.547 9.59e-16 ***
WS          -38.34298   10.23442  -3.746 0.000219 ***
RH            0.44104    0.13849   3.185 0.001620 **
T2M          -3.55870    0.72027  -4.941 1.37e-06 ***
PRECTOTCORR  -3.93164    0.67135  -5.856 1.38e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 66.53 on 269 degrees of freedom
Multiple R-squared:  0.8227,Adjusted R-squared:  0.8175
F-statistic: 156.1 on 8 and 269 DF,  p-value: < 2.2e-16
```