# Analysis of India’s Air Quality Index using R

---

## Project Report


Pratik P. Patil (202118023)
Jeffrey James (202118031)

Course: Statistical Methods using R (SC - 612)

M.Sc. Data Science
Dhirubhai Ambani Institute of Informationa and Communiction Technology - Gandhinagar

# Index

## ❖ *Problem Definition:*

The air quality index (AQI) is an index for reporting air quality on a daily basis. It is a measure of how air pollution affects one's health within a short time period. The purpose of the AQI is to help people know how the local air quality impacts their health. The higher the AQI value, the greater the level of air pollution and the greater the health concerns.

## ❖ *Project Description:*

In this project, we have analyzed the Air Quality Index of a few selected states and their cities to gain inference on the various factors affecting the Air Quality of any given region.

Analyzing the exact Air Pollutant’s like “NO2”, “PM2.5”, “PM10”, “CO”, “CO2”, “Benzene”, “O3”, “Xylene”, “Toluene”, etc. by using various statistical methods like:

a. Normality Test
b. Summary Statistic
c. Correlation
d. Linear Regression
e. Multiple Linear Regression
   And a few Visualization techniques like Bar Plot, Pie Chart, Scatter Plot, etc.

to analyze the given dataset.

## ❖ *Methodology and Data Cleaning:*

We acquired this dataset from Kaggle.com. This data consisted of 15 columns and around 10,000 rows.

We performed data cleaning in R Studio & MS Excel and removed Null values along with unwanted rows and columns.

Further we performed various above mentioned statistical tools on this data to get reliable inferences.

We have taken the AQI Range and AQI Status by using the scale:

| | |
|---|---|
| 0 – 50 – Good | 50 – 100 – Satisfactory |
| 100 – 200 – Moderate | 200 – 300 – Poor |
| 300 – 400 – Very Poor | More than 400 -- Severe |

## ❖ *Code and Analysis:*

### ➢ *Phase - 1: Importing the csv file into R studio:*

```
getwd()

## [1] "E:/DA-IICT/Study/Sem 1/R Project/Indian_Air_Pollution_Data"

AQI_data = read.csv('AQIdatasetmain.csv')
```

**getwd() :** – To get or set Working directory.
**read.csv() :** – To read the comma seperated value files.

### ➢ *Phase - 2: To see the content that the data holds:*

```
head(AQI_data)

##        StationId      City PM2.5   PM10   NO   NO2   NOx   NH3   CO    S
O2
## 1 Andhra_Pradesh Amaravati 71.36 115.75 1.75 20.65 12.40 12.19 0.10 10.
76
## 2 Andhra_Pradesh Amaravati 81.40 124.50 1.44 20.50 12.08 10.72 0.12 15.
24
## 3 Andhra_Pradesh Amaravati 78.32 129.06 1.26 26.00 14.85 10.28 0.14 26.
96
## 4 Andhra_Pradesh Amaravati 73.96 113.56 4.58 19.29 13.97 10.95 0.10 13.
90
## 5 Andhra_Pradesh Amaravati 89.90 140.20 7.71 26.19 19.87 13.12 0.10 19.
37
## 6 Andhra_Pradesh Amaravati 87.14 130.52 0.97 21.31 12.12 14.36 0.15 11.
41
##       O3 Benzene Toluene Xylene AQI AQI_Bucket
## 1 109.26    0.17    5.92   0.10 132   Moderate
## 2 127.09    0.20    6.50   0.06 184   Moderate
## 3 117.44    0.22    7.95   0.08 197   Moderate
## 4 123.80    0.17    2.85   0.04 191   Moderate
## 5 128.73    0.25    2.79   0.07 191   Moderate
## 6 114.80    0.23    3.82   0.04 227       Poor
```

**head(data) :** – Return the first parts of the matrix,table or dataframe

- **Normality Check: To check whether the data follows normal distribution.**

For Normality check, we use Shapiro-Wilk test on all the columns in our data.

**Shapiro Test:**

```
apply(AQI_data[,c(3:14)],2,shapiro.test)
```

```
## $PM2.5
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.72028, p-value = 1.398e-11
##
##
## $PM10
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.88156, p-value = 9.438e-07
##
##
## $NO
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.7365, p-value = 3.39e-11
##
##
## $NO2
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.92744, p-value = 0.0001137
##
##
## $NOx
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.77378, p-value = 3.015e-10
##
##
## $NH3
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.92673, p-value = 0.0001045
```

```
##
## $CO
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.36576, p-value < 2.2e-16
##
##
## $SO2
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.78091, p-value = 4.704e-10
##
##
## $O3
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.75265, p-value = 8.505e-11
##
##
## $Benzene
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.87525, p-value = 5.322e-07
##
##
## $Toluene
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.74148, p-value = 4.482e-11
##
##
## $Xylene
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.83072, p-value = 1.418e-08
```

For Normality check, if P-value is > Alpha (0.05), then we can say that the data follows normal distribution.

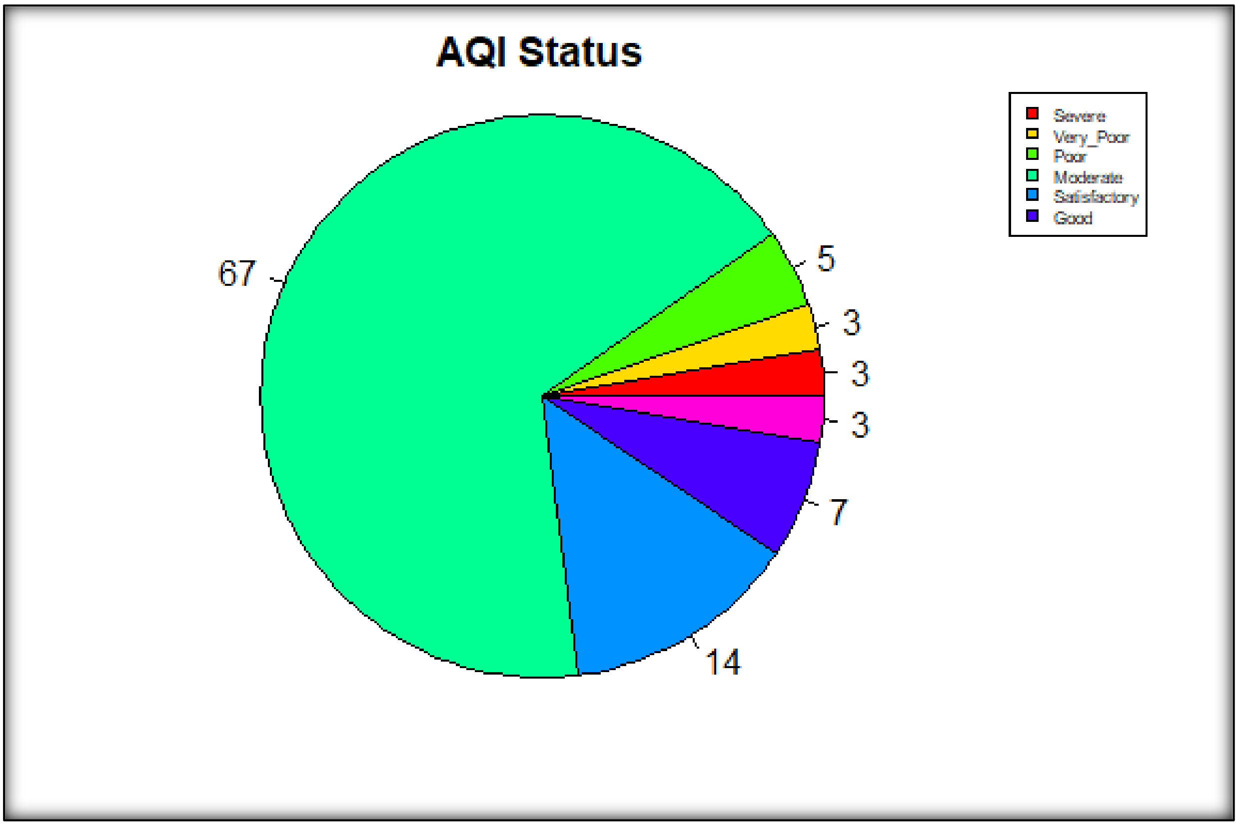
But here, for all the columns P-value < Alpha (0.05). So, we conclude that our data is not normal.

## ➢ *Phase - 3: To know the Summary Statistic of data (mean, median and Quantile):*

```
# Moderate , Poor , Satisfactory, Very Poor, Severe,  Good
summary(AQI_data)


##   StationId             City                PM2.5             PM10
##  Length:87          Length:87          Min.   : 11.04   Min.   : 18.03
##  Class :character   Class :character   1st Qu.: 26.41   1st Qu.: 70.61
##  Mode  :character   Mode  :character   Median : 64.01   Median :107.73
##                                        Mean   : 73.51   Mean   :126.63
##                                        3rd Qu.: 90.66   3rd Qu.:179.50
##                                        Max.   :412.34   Max.   :404.52




##        NO              NO2              NOx              NH3
##  Min.   : 0.570   Min.   : 3.13   Min.   :  2.18   Min.   : 2.77
##  1st Qu.: 3.615   1st Qu.:12.73   1st Qu.: 14.29   1st Qu.:12.12
##  Median : 7.250   Median :19.87   Median : 18.21   Median :14.36
##  Mean   :13.867   Mean   :24.25   Mean   : 28.75   Mean   :17.51
##  3rd Qu.:15.985   3rd Qu.:32.09   3rd Qu.: 34.78   3rd Qu.:22.60
##  Max.   :78.090   Max.   :66.23   Max.   :112.44   Max.   :41.78




##        CO              SO2               O3            Benzene
##  Min.   : 0.100   Min.   : 3.390   Min.   :  1.31   Min.   : 0.010
##  1st Qu.: 0.575   1st Qu.: 9.315   1st Qu.: 18.39   1st Qu.: 0.580
##  Median : 0.850   Median :12.500   Median : 26.34   Median : 2.940
##  Mean   : 1.785   Mean   :15.611   Mean   : 33.49   Mean   : 3.871
##  3rd Qu.: 1.130   3rd Qu.:18.995   3rd Qu.: 36.52   3rd Qu.: 6.270
##  Max.   :27.140   Max.   :79.470   Max.   :128.73   Max.   :12.450




##     Toluene           Xylene             AQI          AQI_Bucket
##  Min.   : 0.040   Min.   : 0.000   Min.   : 28.0   Length:87
##  1st Qu.: 1.520   1st Qu.: 0.575   1st Qu.: 95.0   Class :character
##  Median : 2.850   Median : 5.060   Median :136.0   Mode  :character
##  Mean   : 6.290   Mean   : 6.136   Mean   :178.6
##  3rd Qu.: 9.975   3rd Qu.: 7.680   3rd Qu.:227.0
##  Max.   :41.200   Max.   :25.660   Max.   :653.0
```

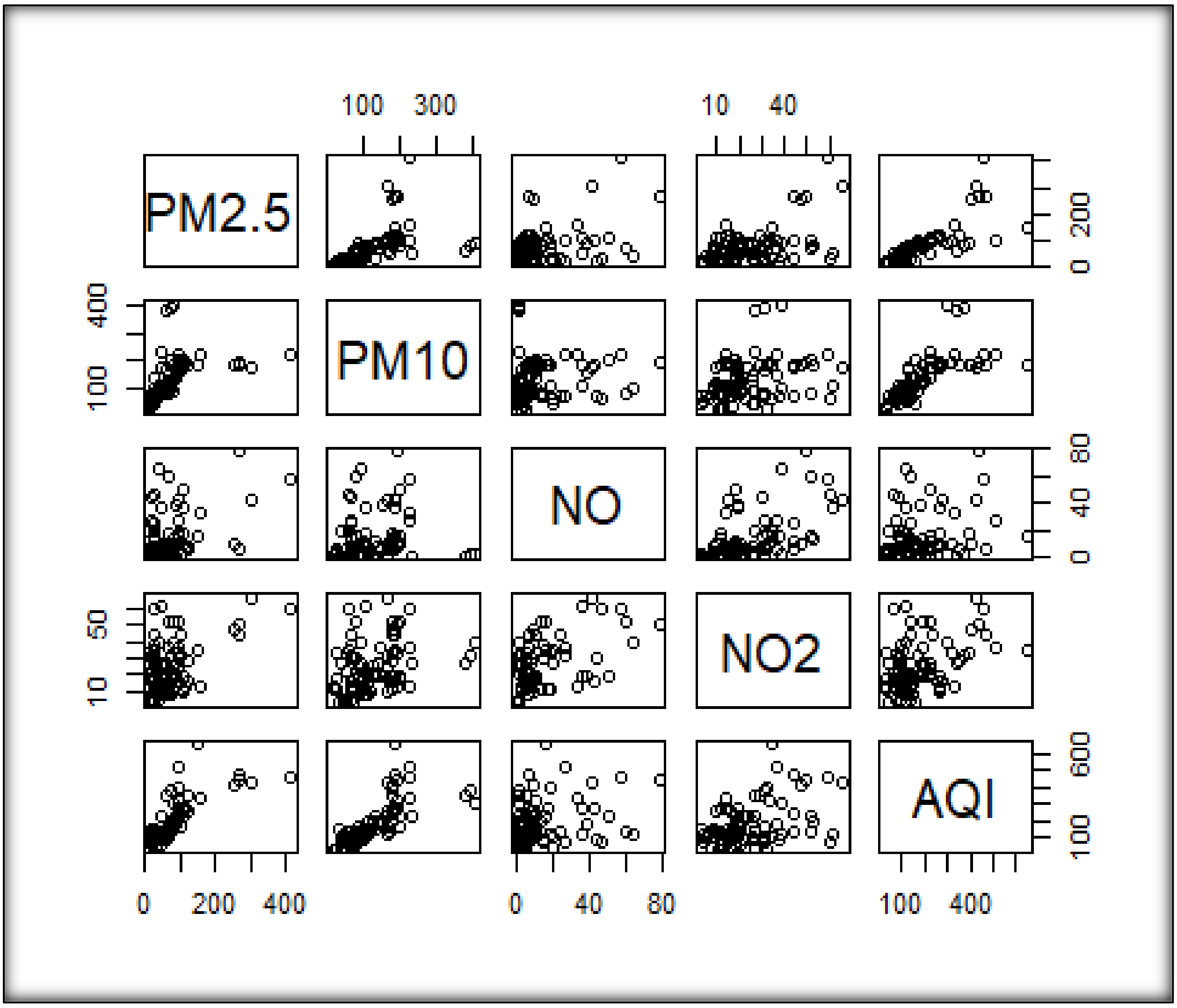
## ➢ *Phase - 4: Checking the Air Quality Status:*

```
#Moderate,Very Poor,Poor , Severe , Good, Satisfactory
#Average = mean(AQI_data$AQI)
moderate = length(which(AQI_data$AQI_Bucket == 'Moderate'))
VeryPoor = length(which(AQI_data$AQI_Bucket == 'Very_Poor'))
Poor = length(which(AQI_data$AQI_Bucket == 'Poor'))
Severe = length(which(AQI_data$AQI_Bucket == 'Severe'))
Satisfactory = length(which(AQI_data$AQI_Bucket == 'Satisfactory'))
Good = length(which(AQI_data$AQI_Bucket == 'Good'))


pie_chart = cbind(Severe,VeryPoor,Poor,moderate,Satisfactory,Good)
percentage = round(100*pie_chart/sum(pie_chart))


pie(pie_chart,labels = percentage,col = rainbow(length(pie_chart)),main =
"AQI Status",radius = 1)

legend("topright",c("Severe","Very_Poor","Poor","Average","Moderate","Sati
sfactory","Good"),cex = 0.5,fill = rainbow(length(pie_chart)))
```

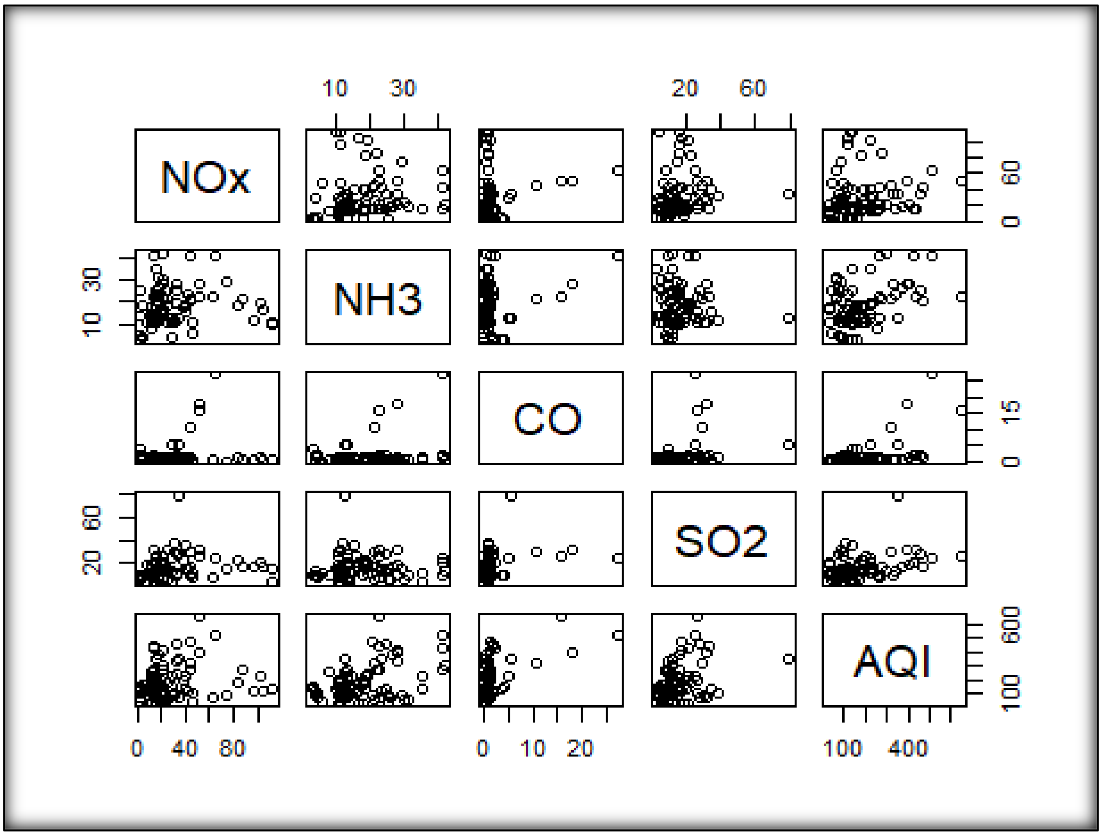


## ➢ *Phase - 5: Correlation between different affecting factors affecting AQI:*

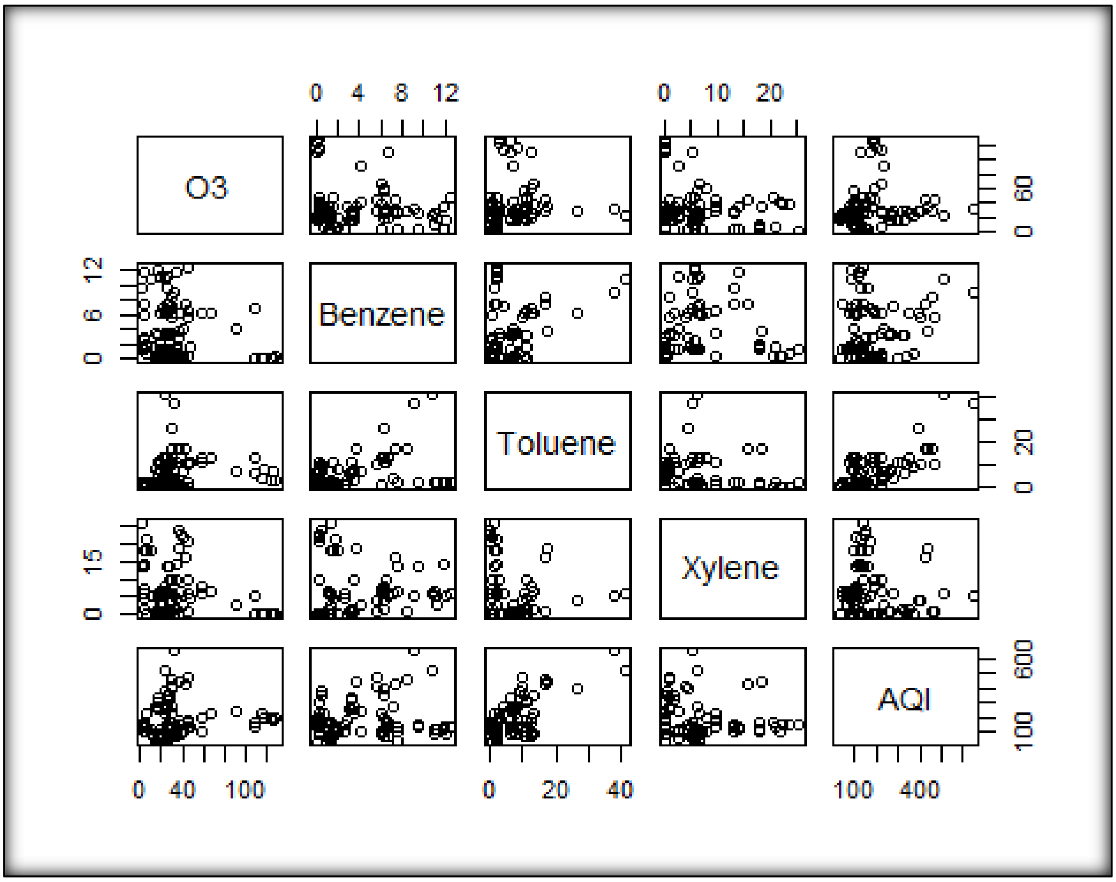
```
pairs(AQI_data[,c(3:6,15)])
```



```r
pairs(AQI_data[,c(7:10,15)])
```




```r
pairs(AQI_data[,c(11:15)])
```



# ➢ *Phase - 6: Testing of Hypothesis*

- **Correlation**

  - **1:**

H0: Correlation between Xylene and AQI is zero.

H1: Correlation between Xylene and AQI is not zero.

```
cor.test(AQI_data$Xylene,AQI_data$AQI)

##
##  Pearson's product-moment correlation
##
## data:  AQI_data$Xylene and AQI_data$AQI
## t = -0.81602, df = 85, p-value = 0.4168
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2933652  0.1248007
## sample estimates:
##         cor
## -0.08816532
```

**Inference:** Here, we can see that the correlation coefficient is -0.088. i.e., Weak Negative Correlation. But the P-vale > Alpha (0.05) and the Correlation coefficient is very close to 0, so we cannot reject H0 and the correlation is almost equal to 0.

  - **2:**

H0: Correlation between PM2.5 and AQI is zero.

H1: Correlation between PM2.5 and AQI is not zero.

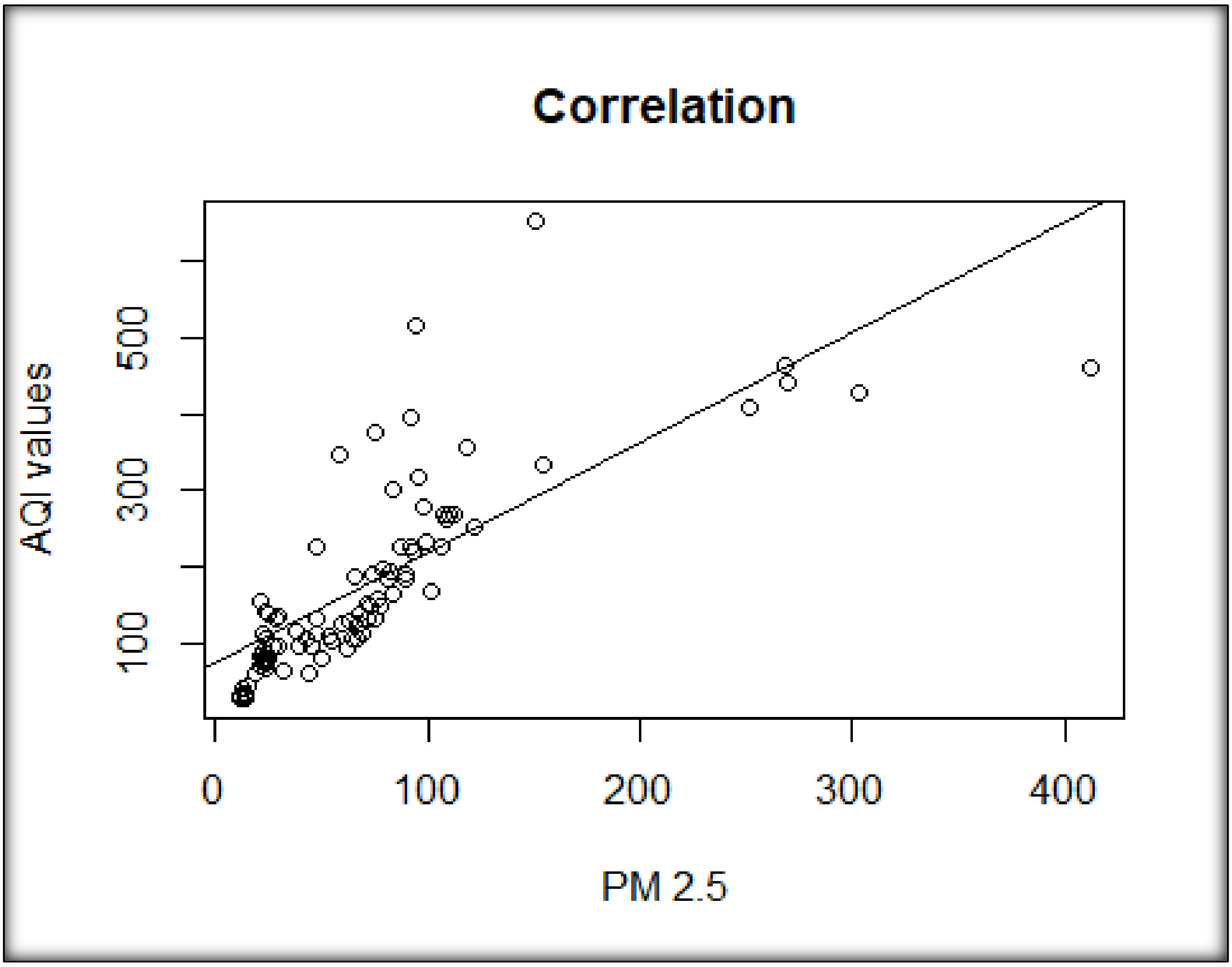
```
cor.test(AQI_data$PM2.5,AQI_data$AQI)

##
##  Pearson's product-moment correlation
##
## data:  AQI_data$PM2.5 and AQI_data$AQI
## t = 11.307, df = 85, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6744753 0.8473296
## sample estimates:
##       cor
## 0.7750121
```

**Inference:** Here, we can see that the correlation coefficient is 0.775. i.e., Strong Positive Correlation. Also, the P-vale < Alpha (0.05), so we can reject H0.

We can conclude that there is Strong Correlation between PM2.5 and AQI.

- **Graphical Representation:**

```
plot(x = AQI_data$PM2.5, y = AQI_data$AQI,xlab ="PM 2.5",ylab = "AQI value
s",main = "Correlation")
abline(lm(AQI_data$AQI~AQI_data$PM2.5))
```

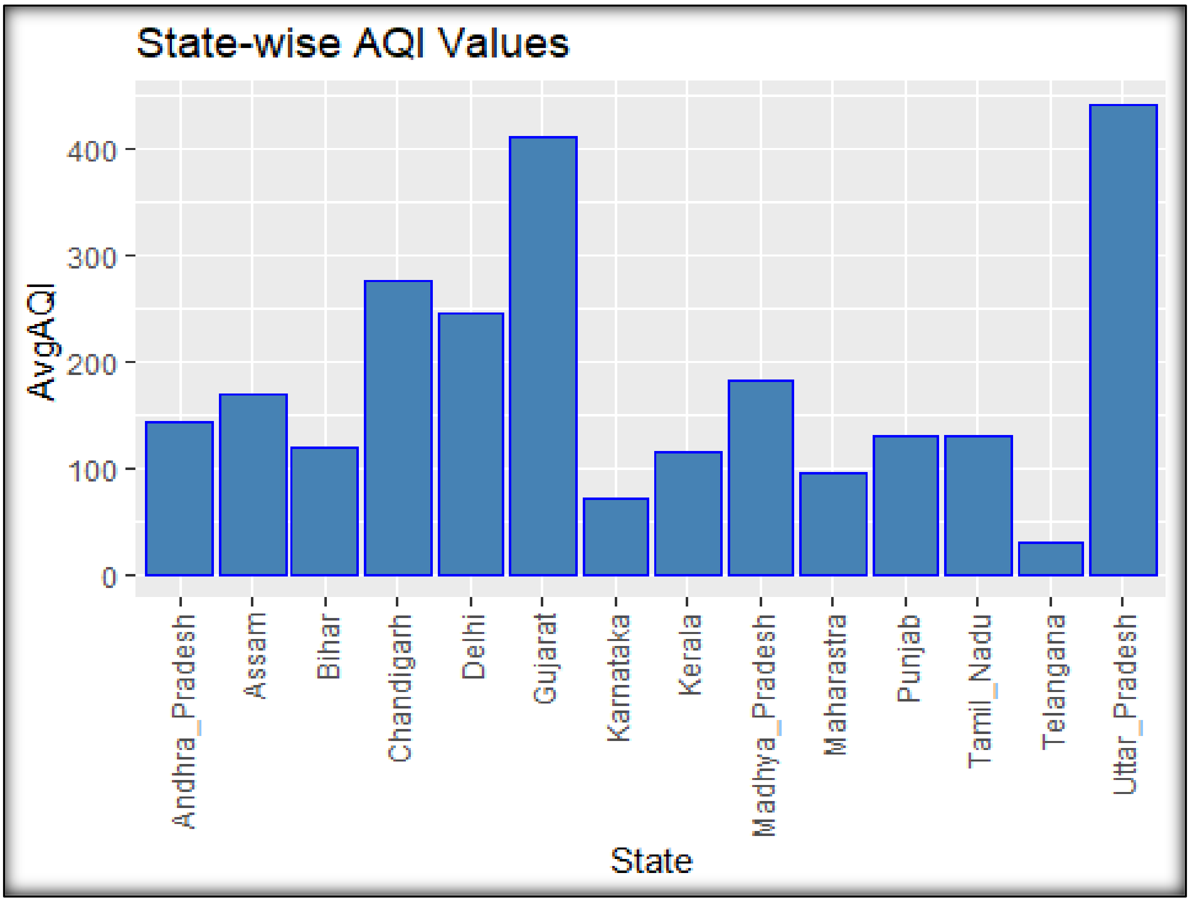



**Inference:** We can visualize that there is positive correlation between PM2.5 and AQI, which says that AQI values are dependent on PM2.5 (Higher the PM2.5 values, higher the AQI).

We can also see that in the Trend line.

- ***Phase - 7: State-wise Average AQI index:***

```
Lalu = aggregate(x = AQI_data$AQI,by = list(AQI_data$StationId),FUN = mean)
colnames(Lalu)[1] = "State"
colnames(Lalu)[2] = "AvgAQI"
print(Lalu)

##              State    AvgAQI
## 1   Andhra_Pradesh 142.25000
## 2            Assam 169.66667
## 3            Bihar 118.71429
## 4       Chandigarh 276.33333
## 5            Delhi 244.16667
## 6          Gujarat 410.83333
## 7        Karnataka  71.42857
## 8           Kerala 113.80000
## 9   Madhya_Pradesh 182.83333
## 10      Maharastra  96.16667
## 11          Punjab 130.80000
## 12      Tamil_Nadu 129.80000
## 13       Telangana  30.00000
## 14   Uttar_Pradesh 439.80000

require(ggplot2)

## Loading required package: ggplot2

ggplot(Lalu,aes(x = State,y = AvgAQI)) + ggtitle("State-wise AQI Values") + geom_b
ar(stat = "identity",color ="blue", fill = "steelblue") + theme(axis.text.x = elem
ent_text(angle = 90,hjust = 1,vjust = 0.5))
```



## ➢ *Phase - 8: Regression:*

Ho: There is no significant relationship between the variables.

H1: There is significant relationship between the variables.

**Linear Regression:**

```
linear_rega= lm(AQI ~ CO,AQI_data)
summary(linear_rega)

##
## Call:
## lm(formula = AQI ~ CO, data = AQI_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -123.50  -74.59  -30.05   48.35  292.43
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  146.460     12.327  11.881  < 2e-16 ***
## CO            18.007      2.887   6.236 1.67e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 104.4 on 85 degrees of freedom
## Multiple R-squared:  0.3139, Adjusted R-squared:  0.3058
## F-statistic: 38.89 on 1 and 85 DF,  p-value: 1.672e-08

# Predicting the Value of AQI for a given CO reading:

df1 = data.frame(CO = 1.69)
predict(linear_rega,df1)

##        1
## 176.8913
```

**Inference:** Here, P-value (1.67e-08) < Alpha (0.001) which shows that the regression model fits properly and also has a high F-statistic (38.89).

So, we reject H0.

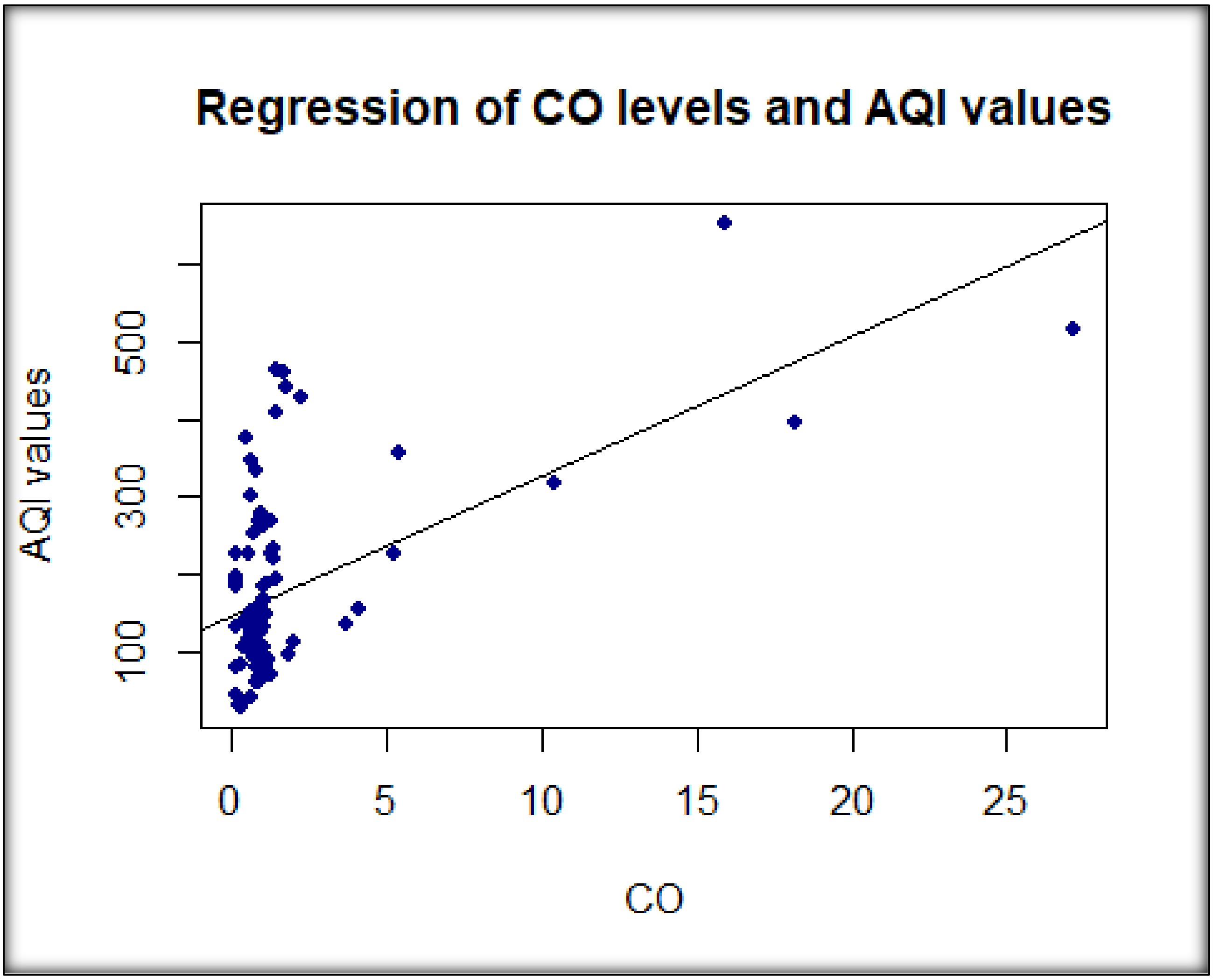
This shows that AQI and CO levels are highly related to each other.

**Regression Equation:** y = 146.460 + 18.007 * x

Thereafter, we have also predicted the value of AQI for a given CO value by using the *predict()* function.

- **Graphical Representation:**

```
# Plotting scatter plot with Regression Line:

plot(x = AQI_data$CO, y = AQI_data$AQI,xlab ="CO",ylab = "AQI values",main
= "Regression of CO levels and AQI values",abline(linear_rega),cex = 1.3,p
ch=20,col="Dark Blue")
```




**Inference:** We can visualize the regression and regression line.

**Multiple Regression:**

```
rega = lm(AQI ~ PM2.5 + PM10 + NO + NO2 + NOx + NH3 + CO + SO2 + O3 + Benz
ene + Toluene + Xylene ,AQI_data)
summary(rega)

##
## Call:
## lm(formula = AQI ~ PM2.5 + PM10 + NO + NO2 + NOx + NH3 + CO +
##     SO2 + O3 + Benzene + Toluene + Xylene, data = AQI_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -106.054  -24.209    0.804   23.109  154.998
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.85229   17.92562  -0.103   0.9180
## PM2.5        0.93116    0.09160  10.166 1.08e-15 ***
## PM10         0.63656    0.07584   8.394 2.33e-12 ***
## NO          -0.09341    0.44768  -0.209   0.8353
## NO2         -0.53375    0.42835  -1.246   0.2167
## NOx          0.10776    0.26438   0.408   0.6847
## NH3         -0.22906    0.66572  -0.344   0.7318
## CO           9.62891    1.88765   5.101 2.52e-06 ***
## SO2          0.34155    0.50460   0.677   0.5006
## O3           0.18276    0.16717   1.093   0.2778
## Benzene      1.25150    1.50435   0.832   0.4081
## Toluene      2.21884    1.23219   1.801   0.0758 .
## Xylene      -0.14488    0.69765  -0.208   0.8361
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.26 on 74 degrees of freedom
## Multiple R-squared:  0.9199, Adjusted R-squared:  0.9069
## F-statistic: 70.78 on 12 and 74 DF,  p-value: < 2.2e-16

# Predicting the value of AQI:

xyz = data.frame(PM2.5=25.2,PM10=69,NO=28.5,NO2=37,NOx=24.3,NH3=21.3,CO=0.
89,SO2=22.9,O3=108,Benzene=3.5,Toluene=6.9,Xylene=4)
predict(rega,xyz)

##        1
## 96.10458
```

Here, we have taken the Multiple Linear regresion of AQI (dependent variable) with respect to all the other Independent variables.

We can see that there is a very high F-statistic value (70.78) and also P-value (2.2e-16) < Alpha (0.05) so we can conclude that AQI is highly dependent on all these independent variables.

After that, we have also predicted the AQI for given values of all the other variables.

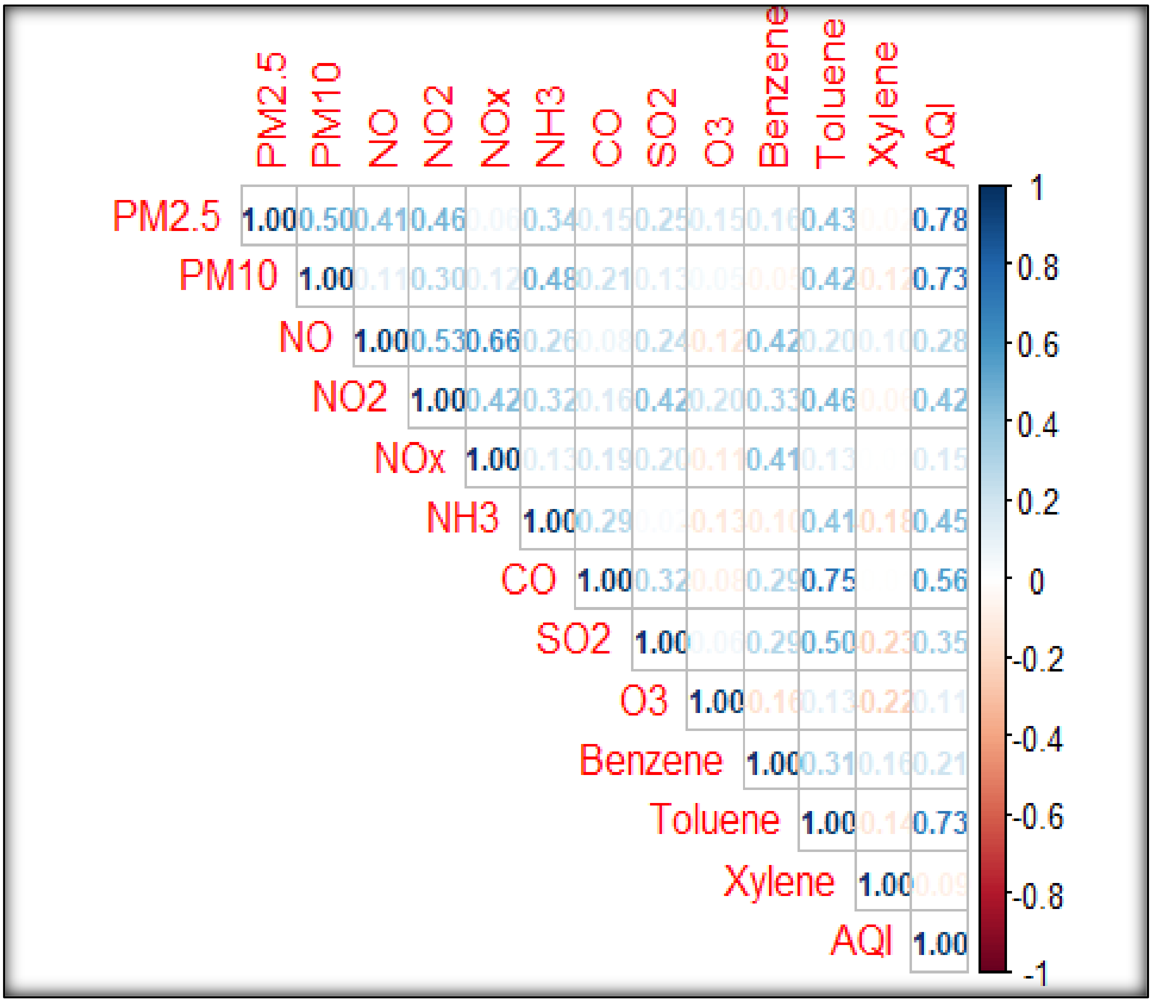
- **Correlation Matrix:**

```
cor_matrix = cor(AQI_data[,c(3:15)])
round(cor_matrix,2)
```

```
##          PM2.5  PM10    NO   NO2   NOx   NH3    CO   SO2    O3 Benzene T
oluene
## PM2.5     1.00  0.50  0.41  0.46  0.06  0.34  0.15  0.25  0.15    0.16
0.43
## PM10      0.50  1.00  0.11  0.30  0.12  0.48  0.21  0.13  0.05   -0.05
0.42
## NO        0.41  0.11  1.00  0.53  0.66  0.26  0.08  0.24 -0.12    0.42
0.20
## NO2       0.46  0.30  0.53  1.00  0.42  0.32  0.16  0.42  0.20    0.33
0.46
## NOx       0.06  0.12  0.66  0.42  1.00  0.13  0.19  0.20 -0.11    0.41
0.13
## NH3       0.34  0.48  0.26  0.32  0.13  1.00  0.29  0.02 -0.13   -0.10
0.41
## CO        0.15  0.21  0.08  0.16  0.19  0.29  1.00  0.32 -0.08    0.29
0.75
## SO2       0.25  0.13  0.24  0.42  0.20  0.02  0.32  1.00  0.06    0.29
0.50
## O3        0.15  0.05 -0.12  0.20 -0.11 -0.13 -0.08  0.06  1.00   -0.16
0.13
## Benzene   0.16 -0.05  0.42  0.33  0.41 -0.10  0.29  0.29 -0.16    1.00
0.31
## Toluene   0.43  0.42  0.20  0.46  0.13  0.41  0.75  0.50  0.13    0.31
1.00
## Xylene   -0.02 -0.12  0.10 -0.06  0.01 -0.18 -0.01 -0.23 -0.22    0.16
-0.14
## AQI       0.78  0.73  0.28  0.42  0.15  0.45  0.56  0.35  0.11    0.21
0.73
##          Xylene   AQI
## PM2.5     -0.02  0.78
## PM10      -0.12  0.73
## NO         0.10  0.28
## NO2       -0.06  0.42
## NOx        0.01  0.15
## NH3       -0.18  0.45
## CO        -0.01  0.56
## SO2       -0.23  0.35
## O3        -0.22  0.11
## Benzene    0.16  0.21
## Toluene   -0.14  0.73
## Xylene     1.00 -0.09
## AQI       -0.09  1.00
```

**Inference:** Here, we can see the correlation matrix for the data.

- **Corrplot:**
- `require(corrplot)`
- `## Loading required package: corrplot`
- `## corrplot 0.90 loaded`
- `corrplot(cor(AQI_data[,c(3:15)],method = "pearson"), method = 'number', type = 'upper',number.cex = 0.75, )`

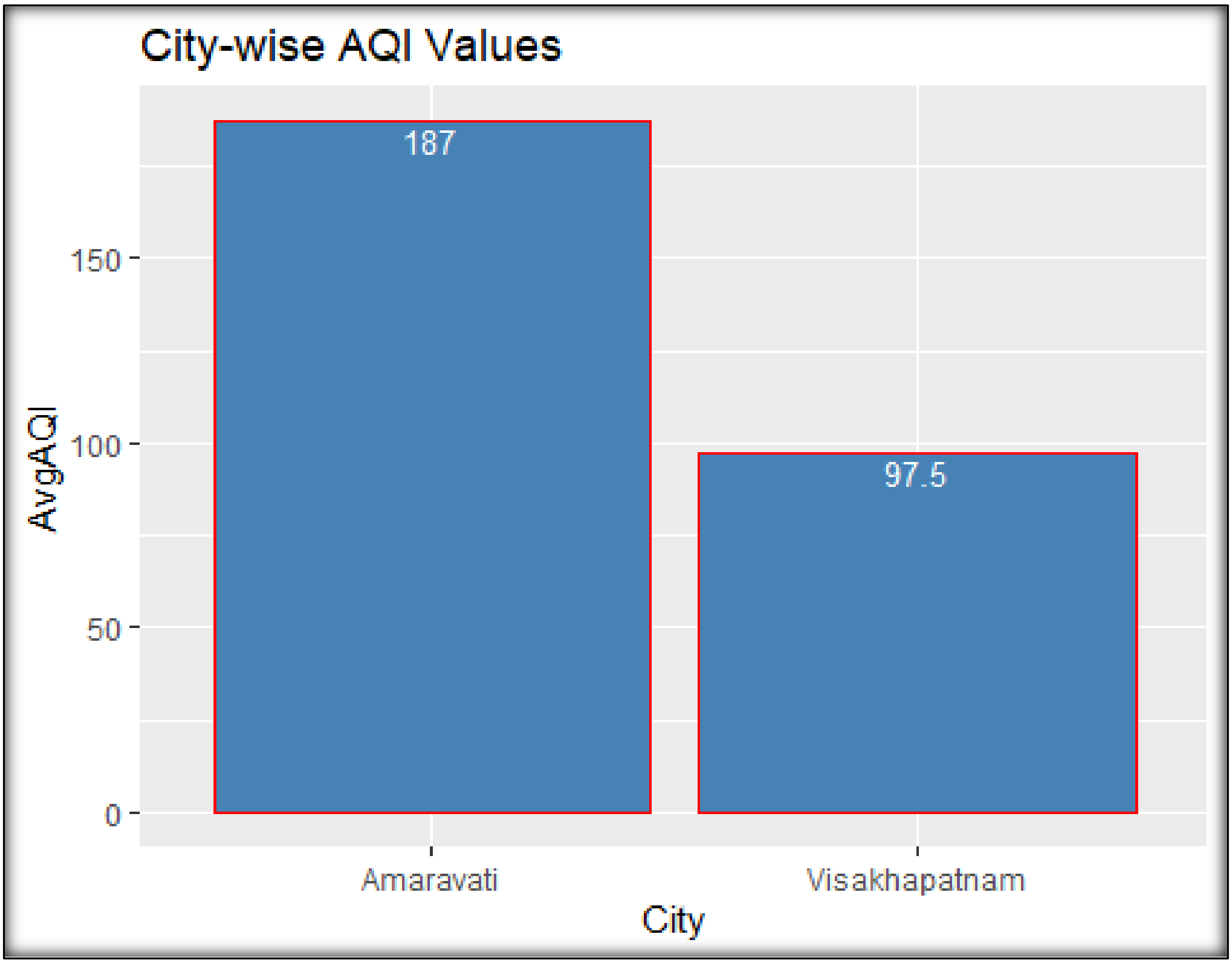



**Inference:** Here, we can see the correlation matrix using corrplot for the data.

With reference to the multiple regression in the above section, the independent variables namely “PM2.5”, “PM10”, “CO”, “Toluene” are perfectly regressed wrt AQI and we can see in the regression matrix, these variables have the highest correlation with AQI.

## ➢ *Phase - 9: Average AQI for given cities in a state:*

```
City_avg = aggregate(x=AQI_data$AQI,by=list(AQI_data$City),FUN = mean)
City_avg_1 = City_avg[c(2,15),]
colnames(City_avg_1)[1] = "City"
colnames(City_avg_1)[2] = "AvgAQI"
ggplot(City_avg_1,aes(x = City, y = AvgAQI)) + ggtitle("City-wise AQI Valu
es") + geom_bar(stat = "identity",color ="red", fill = "steelblue") + geom
_text(aes(label = AvgAQI),vjust = 1.2,color = "white", position = position
_dodge(0.9),size = 3.5 )
```




**Inference:** In the Andhra Pradesh, we have considered two cities 'Amravati' and 'Vishakhapatnam', looking at the bar plot, we can say that the mean Air Quality Index of 'Amravati' is higher than 'Vishakhapatnam', therefore Vishakhapatnam has safer air than Amravati.

## ➢ *Phase – 9: Chi Square Test:*

In this section, we took the AQI Status that was given as ‘Good’, ‘Satisfactory’, ‘Moderate’, ‘Poor’, ‘Very Poor’ and ‘Severe’ and converted them into numeric data i.e., 1, 2, 3, 4, 5 and 6 respectively.

```
getwd()

## [1] "E:/DA-IICT/Study/Sem 1/R Project/Indian_Air_Pollution_Data"

chi_sq = read.csv("Column_15_16.csv")
head(chi_sq)

##   Toluene AQI_Bucket
## 1    5.92          3
## 2    6.50          3
## 3    7.95          3
## 4    2.85          3
## 5    2.79          3
## 6    3.82          4

# Chi-Square test:
data1 = data.matrix(chi_sq)
chisq.test(data1)

## Warning in chisq.test(data1): Chi-squared approximation may be incorrec
t

##
##  Pearson's Chi-squared test
##
## data:  data1
## X-squared = 146.95, df = 86, p-value = 4.708e-05
```

**Inference:** Here, we have performed Chi-Square test between two variables CO and AQI Status.

We can see that the p-value (4.708e-05) < Alpha (0.05), and also observed Chi-squared value (146.95) > Critical Chi-squared value (101.87) for 86 degrees of freedom.

So, we reject the H0 and can conclude that there is a significant relationship between these variables CO and AQI status.

## ❖ *Conclusion:*

- We found that all the factors i.e., “NO2”, “PM2.5”, “PM10”, “CO”, “CO2”, “Benzene”, “O3”, “Xylene”, “Toluene”, etc. are the pollutants that generally affect the AQI of any place.

- From the above-mentioned factors, “PM2.5”, “PM10”, “CO” and “Toluene” are the major affecting factors for AQI and also have very high Correlation with it.

- “O3” and “Xylene” are least corelated and don’t have much effect on the AQI values.

- “Uttar-Pradesh” and “Gujarat” have the highest mean AQI which means that these states have high air pollution and lie in the “Severe” category.

- “Telangana” has a lower mean AQI and comes in the “Good” AQI category.

  And on the second number is “Karnataka” with “Satisfactory” AQI values.

## ❖ *References:*


- Dataset: Air Quality Data India.
  Link: https://www.kaggle.com/rohanrao/air-quality-data-in-india?select=station_day.csv