A PROJECT REPORT ON

EXPLORATORY DATA ANALYSIS

ON

AIR QUALITY

SUBMITTED TO

Asst. Prof. Dr. Mausam Kumari

Emp. ID: E17645

CHANDIGARH UNIVERSITY

By

Sweta Dey
24MCI10247

In partial fulfillment for the award of the degree of

MASTER OF COMPUTER APPLICATION

IN
ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

UNIVERSITY INSTITUTE OF COMPUTING,

CHANDIGARH UNIVERSITY

OCTOBER-2024

Tasks

Task to be done:

1. **Load the Dataset**

   1.1 Load the Dataset

   1.2 View the dimension of the dataset

   1.3 Check the structure of the Data

   1.4 View the first few rows.

2. **Summarize the Dataset**

   2.1 Generate Summary Statistics

   2.2 Generate Structure Statistics

3. **Visualizing the Distribution of Variables**

   3.1 Create Histograms from the data of the dataset.

4. **Identifying Outliers**

   4.1 Create Boxplot from the data of the dataset.

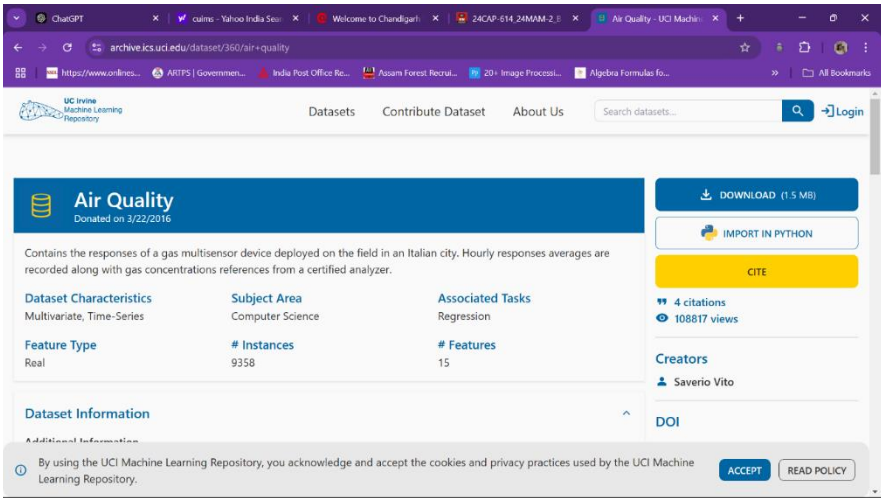
5. **Analyzing Relationships Between Variables**

   5.1 Create Scatter Plots from the data of the dataset.

Steps/Commands

Performing Exploratory Data Analysis

1. Download Air Quality dataset from UIC repository.



Air Quality

Donated on 3/22/2016

Contains the responses of a gas multisensor device deployed on the field in an Italian city. Hourly responses averages are recorded along with gas concentrations references from a certified analyzer.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate, Time-Series	Computer Science	Regression
Feature Type	**# Instances**	**# Features**
Real	9358	15

4 citations
108817 views

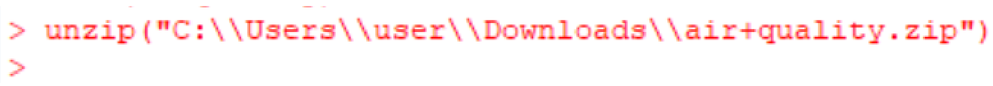
Creators
Saverio Vito

Dataset Information

DOI

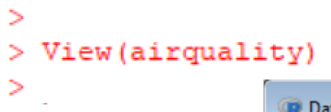
By using the UCI Machine Learning Repository, you acknowledge and accept the cookies and privacy practices used by the UCI Machine Learning Repository.

2. Open R and unzip the downloaded file.



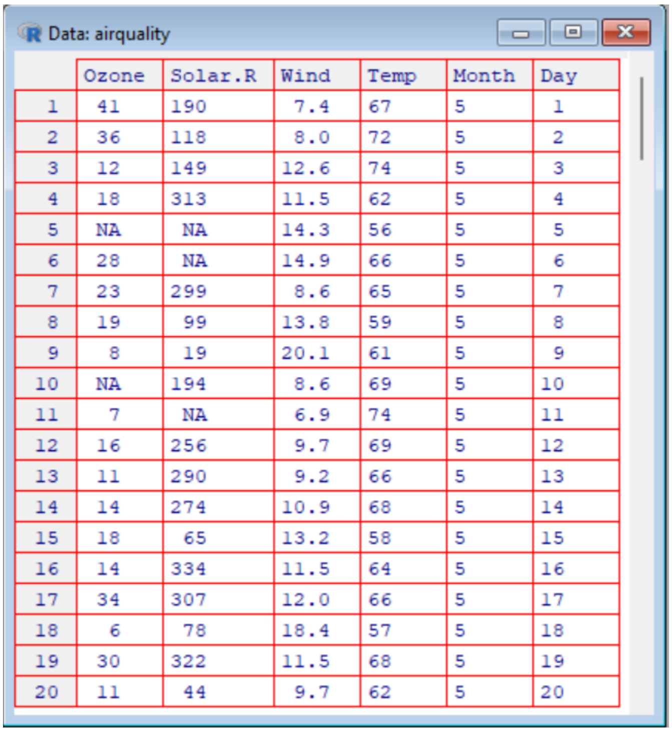
```
> unzip("C:\\Users\\user\\Downloads\\air+quality.zip")
>
```

3. View the Air Quality Dataset with the R Code.



```
>
> View(airquality)
>
```

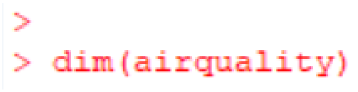
OUTPUT:



Data: airquality

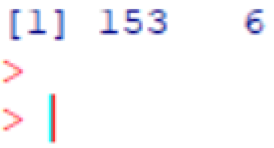
	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6
7	23	299	8.6	65	5	7
8	19	99	13.8	59	5	8
9	8	19	20.1	61	5	9
10	NA	194	8.6	69	5	10
11	7	NA	6.9	74	5	11
12	16	256	9.7	69	5	12
13	11	290	9.2	66	5	13
14	14	274	10.9	68	5	14
15	18	65	13.2	58	5	15
16	14	334	11.5	64	5	16
17	34	307	12.0	66	5	17
18	6	78	18.4	57	5	18
19	30	322	11.5	68	5	19
20	11	44	9.7	62	5	20

4. Dimension of the dataset.



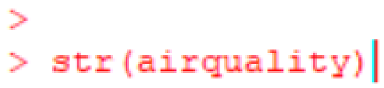
```
>
> dim(airquality)
```

OUTPUT:



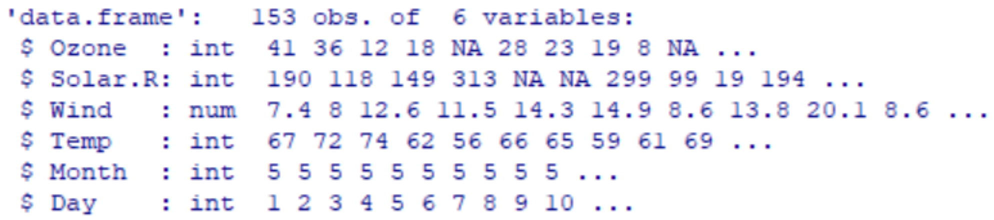
```
[1] 153   6
>
> 
```

5. Structure of the dataset.



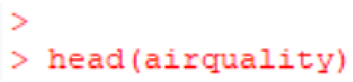
```
>
> str(airquality)
```

OUTPUT:



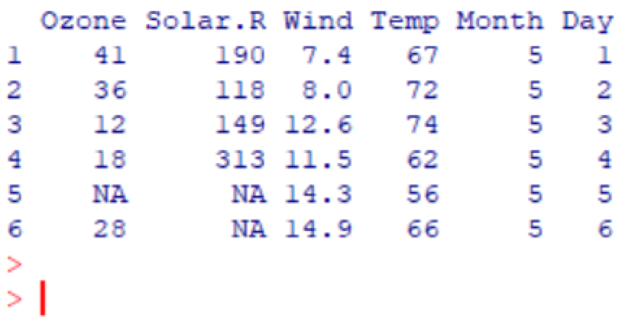
```
'data.frame':   153 obs. of  6 variables:
 $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
 $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
 $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
 $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
 $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
 $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

6. Getting first 6 rows of the dataset.



```
>
> head(airquality)
```

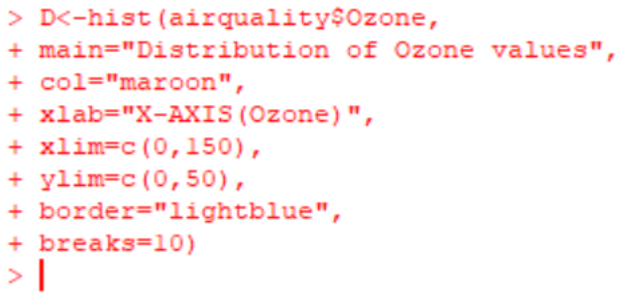
OUTPUT:



```
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
>
> 
```

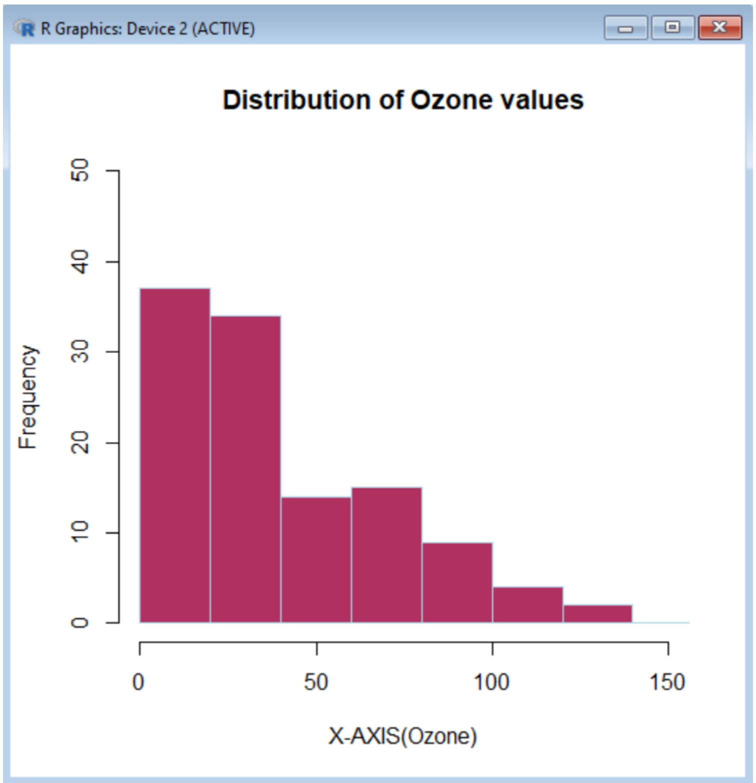
Visualizing the distribution of variables using HISTOGRAM

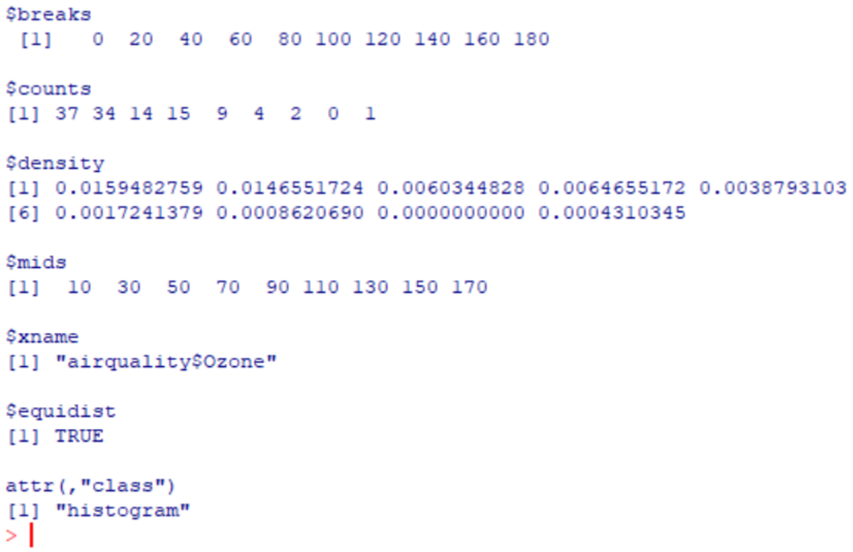
1. On Ozone values.



```
> D<-hist(airquality$Ozone,
+ main="Distribution of Ozone values",
+ col="maroon",
+ xlab="X-AXIS(Ozone)",
+ xlim=c(0,150),
+ ylim=c(0,50),
+ border="lightblue",
+ breaks=10)
> 
```

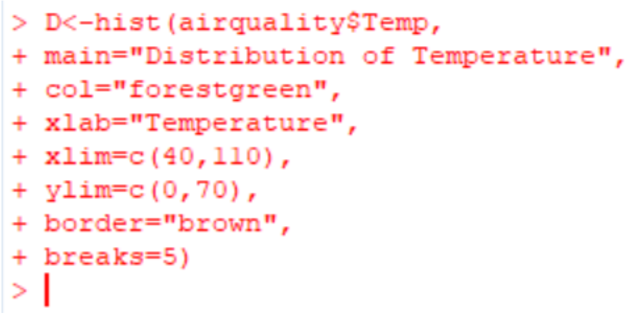
OUTPUT:





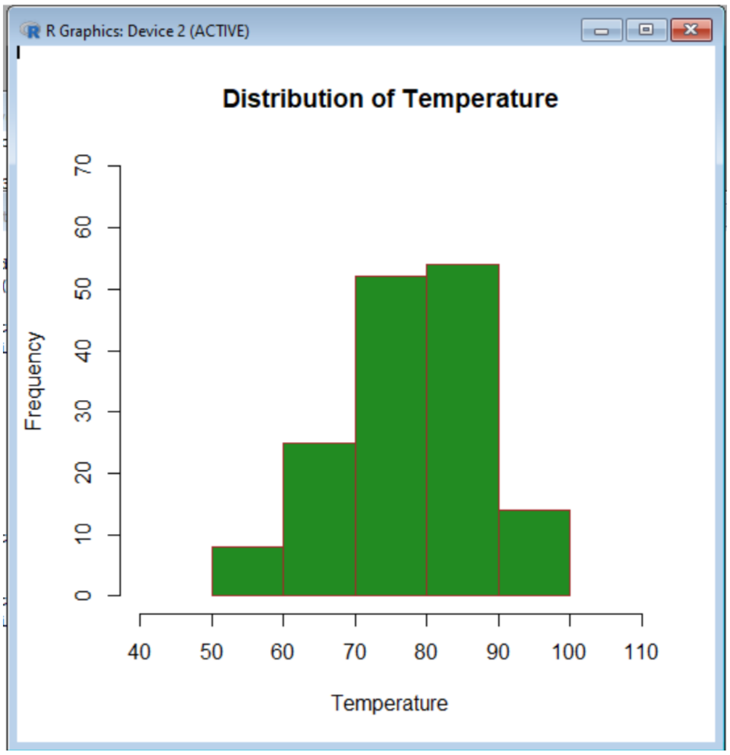
```
$breaks
 [1]   0  20  40  60  80 100 120 140 160 180

$counts
[1] 37 34 14 15  9  4  2  0  1

$density
[1] 0.0159482759 0.0146551724 0.0060344828 0.0064655172 0.0038793103
[6] 0.0017241379 0.0008620690 0.0000000000 0.0004310345

$mids
[1]  10  30  50  70  90 110 130 150 170

$xname
[1] "airquality$Ozone"

$equidist
[1] TRUE

attr(,"class")
[1] "histogram"
> 
```

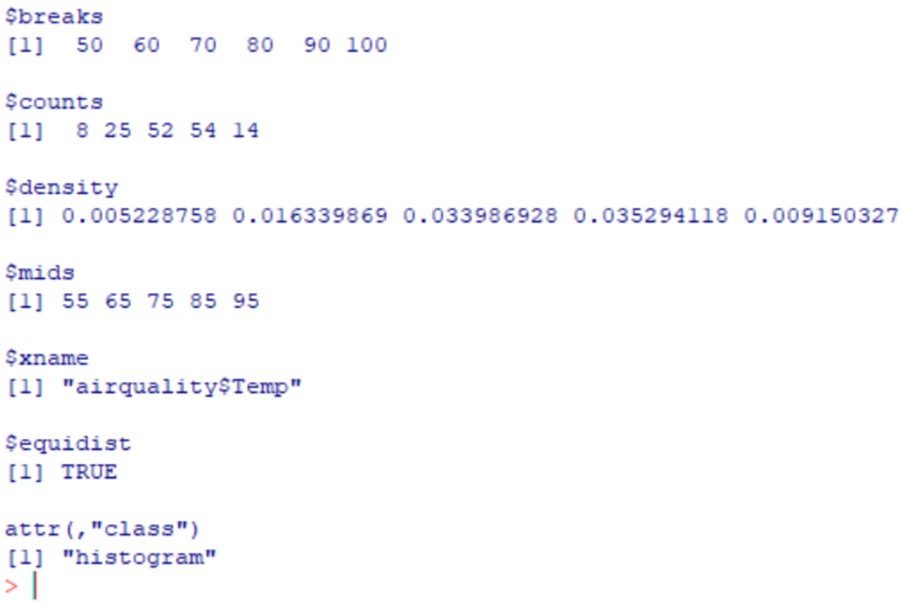
2. On Temperature values.



```
> D<-hist(airquality$Temp,
+ main="Distribution of Temperature",
+ col="forestgreen",
+ xlab="Temperature",
+ xlim=c(40,110),
+ ylim=c(0,70),
+ border="brown",
+ breaks=5)
> 
```

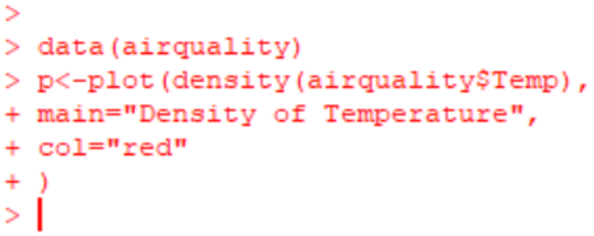
OUTPUT:





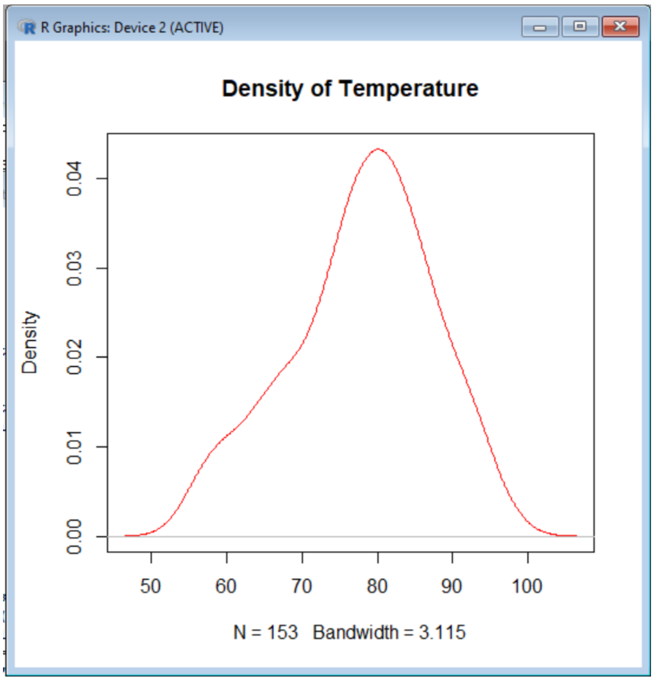
```
$breaks
[1]  50  60  70  80  90 100

$counts
[1]  8 25 52 54 14

$density
[1] 0.005228758 0.016339869 0.033986928 0.035294118 0.009150327

$mids
[1] 55 65 75 85 95

$xname
[1] "airquality$Temp"

$equidist
[1] TRUE

attr(,"class")
[1] "histogram"
> 
```

3. Density of temperature.



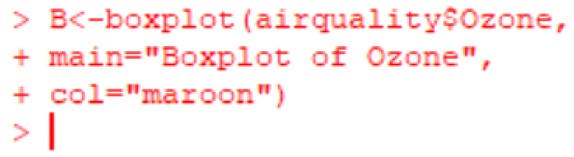
```
> 
> data(airquality)
> p<-plot(density(airquality$Temp),
+ main="Density of Temperature",
+ col="red"
+ )
> 
```

OUTPUT:



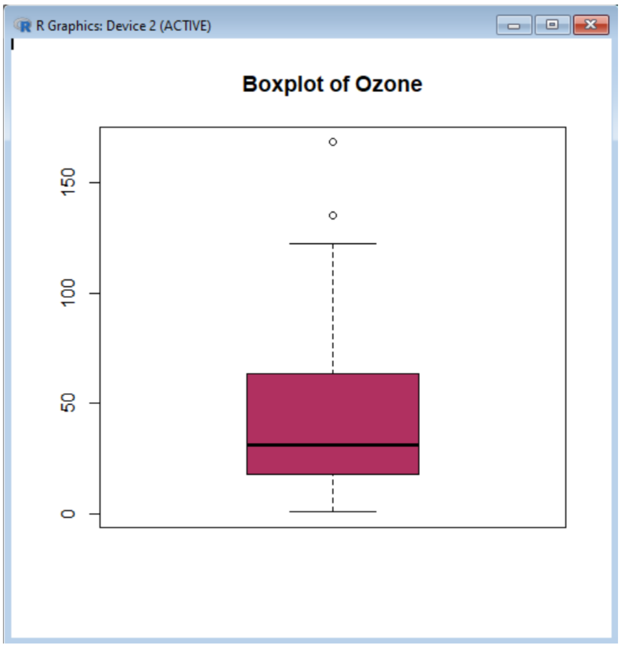
Identify the Outliers using BOX PLOT

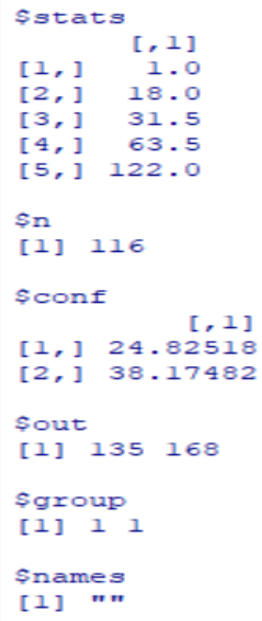
1. On Ozone values.



```
> B<-boxplot(airquality$Ozone,
+ main="Boxplot of Ozone",
+ col="maroon")
> |
```

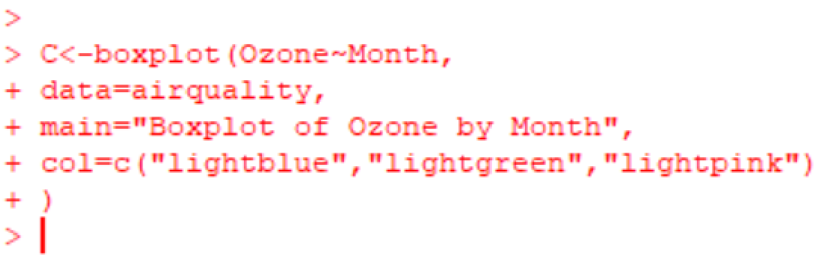
OUTPUT:





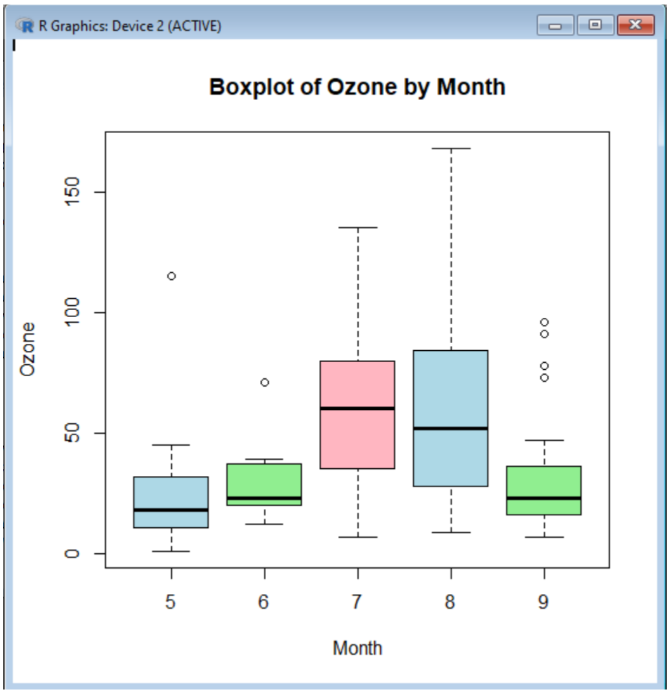
```
$stats
      [,1]
[1,]   1.0
[2,]  18.0
[3,]  31.5
[4,]  63.5
[5,] 122.0

$n
[1] 116

$conf
         [,1]
[1,] 24.82518
[2,] 38.17482

$out
[1] 135 168

$group
[1] 1 1

$names
[1] ""
```

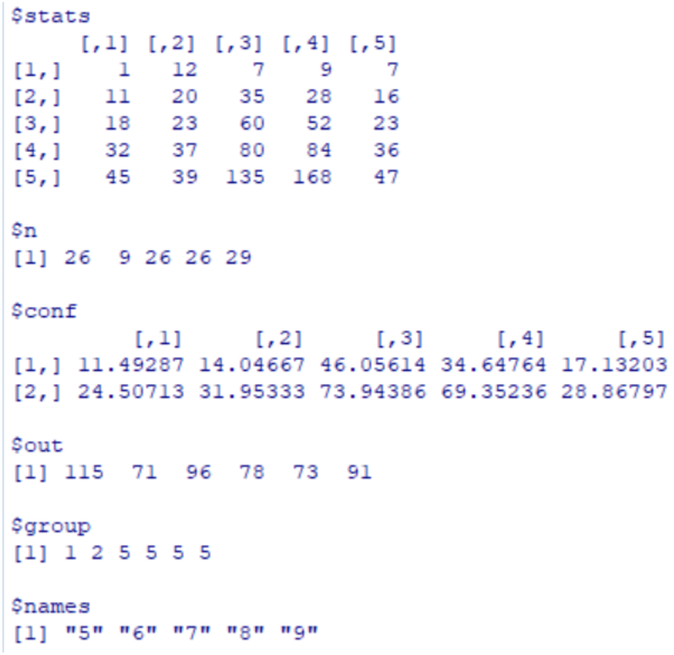
2. Comparing values using BOX PLOT.



```
>
> C<-boxplot(Ozone~Month,
+ data=airquality,
+ main="Boxplot of Ozone by Month",
+ col=c("lightblue","lightgreen","lightpink")
+ )
> 
```

OUTPUT:

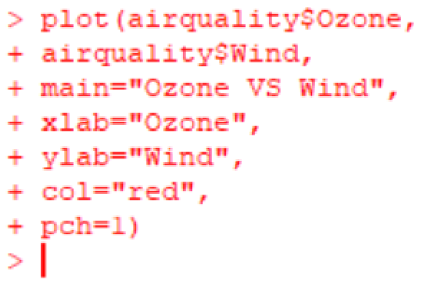




```
$stats
     [,1] [,2] [,3] [,4] [,5]
[1,]    1   12    7    9    7
[2,]   11   20   35   28   16
[3,]   18   23   60   52   23
[4,]   32   37   80   84   36
[5,]   45   39  135  168   47

$n
[1] 26  9 26 26 29

$conf
         [,1]     [,2]     [,3]     [,4]     [,5]
[1,] 11.49287 14.04667 46.05614 34.64764 17.13203
[2,] 24.50713 31.95333 73.94386 69.35236 28.86797

$out
[1] 115  71  96  78  73  91

$group
[1] 1 2 5 5 5 5

$names
[1] "5" "6" "7" "8" "9"
```

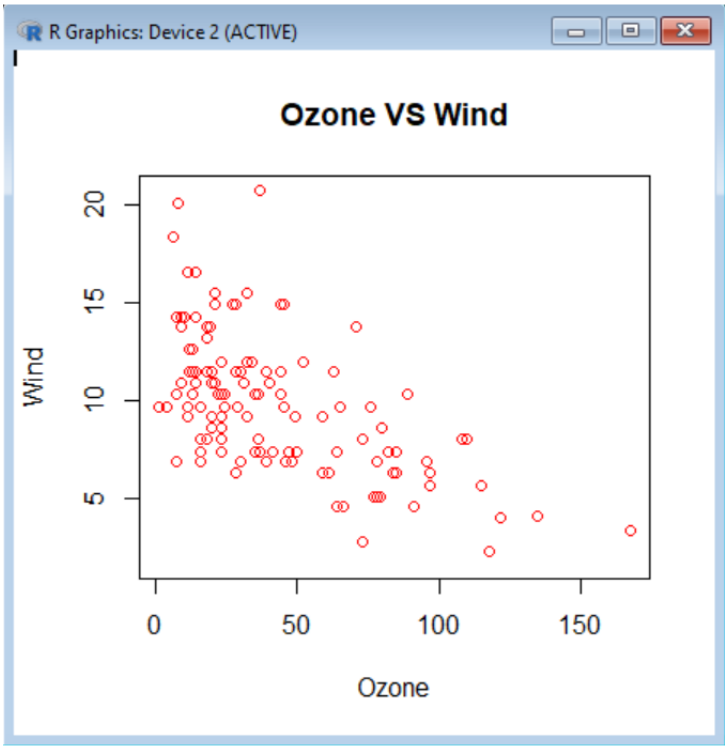
Analyzing relationship between variables using SCATTER PLOT

1. On Ozone values.



```
> plot(airquality$Ozone,
+ airquality$Wind,
+ main="Ozone VS Wind",
+ xlab="Ozone",
+ ylab="Wind",
+ col="red",
+ pch=1)
> 
```

OUTPUT:



6. Conclusion

In conclusion, the exploratory data analysis of air quality has revealed important insights into pollutant patterns and trends. By identifying geographic disparities and correlations, we have enhanced our understanding of air quality variations and improved data quality through anomaly detection.

These findings support public health and environmental policy, guiding informed decision-making and targeted interventions. This analysis highlights the importance of ongoing monitoring and lays the foundation for future research on air pollution and its health impacts.