Assignment 5.3

Import dataset from the following link: AirQuality Data Set

Perform the following written operations:

1. Read the file in Zip format and get it into R.

Answer:

temp<-tempfile()

urlziparchive="https://archive.ics.uci.edu/ml/datasets/Air+quality"

download.file(urlziparchive,temp)

csvfile=unz(temp,"AirQualityUCI.csv")

AQ<-read.csv(csvfile)

AQ

2. Create Univariate for all the columns.

3. Check for missing values in all columns.

Answer:

|  |
| --- |
| # question 3  > for(var in names(AQ)){  + missing<-sum(is.na(AQ[,var]))  + if(missing>0){  + print(c(var,missing))  + }  + }  [1] "PT08.S1.CO." "114"  [1] "NMHC.GT." "114"  [1] "PT08.S2.NMHC." "114"  [1] "NOx.GT." "114"  [1] "PT08.S3.NOx." "114"  [1] "NO2.GT." "114"  [1] "PT08.S4.NO2." "114"  [1] "PT08.S5.O3." "114"  [1] "X" "9471"  [1] "X.1" "9471" |
|  |
| |  | | --- | | > | |

4. Impute the missing values using appropriate methods.

> library(mice)

Loading required package: lattice

Attaching package: ‘mice’

The following objects are masked from ‘package:base’:

cbind, rbind

> md.pattern(AQ)

Date Time CO.GT. C6H6.GT. T RH AH PT08.S1.CO. NMHC.GT. PT08.S2.NMHC. NOx.GT.

9357 1 1 1 1 1 1 1 1 1 1 1

114 1 1 1 1 1 1 1 0 0 0 0

0 0 0 0 0 0 0 114 114 114 114

PT08.S3.NOx. NO2.GT. PT08.S4.NO2. PT08.S5.O3. X X.1

9357 1 1 1 1 0 0 2

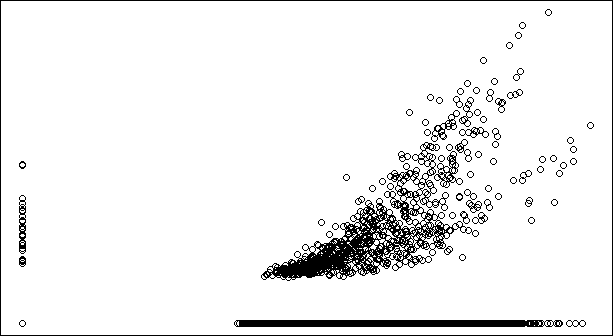
114 0 0 0 0 0 0 10

114 114 114 114 9471 9471 19854



5. Create bi-variate analysis for all relationships.

plot(AQ$PT08.S1.CO.,AQ$NMHC.GT.)



6. Test relevant hypothesis for valid relations.

7. Create cross tabulations with derived variables.

8. Check for trends and patterns in time series.

9. Find out the most polluted time of the day and the name of the chemical compound.