1. Import dataset from the following link:
https://archive.ics.uci.edu/ml/machine-learning-databases/00360/
Perform the below written operations:

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

```
> mydata<-read_csv("C:/Users/BastianSol/Downloads/AirqualityUCI.zip")
Multiple files in zip: reading 'AirQualityUCI.csv'
Parsed with column specification:
cols(
   `Date;Time;CO(GT);PT08.S1(CO);NMHC(GT);C6H6(GT);PT08.S2(NMHC);NOx(GT);PT08.S3(NOx);NO2(GT)
;PT08.S4(NO2);PT08.S5(O3);T;RH;AH;;` = col_character()
)
Warning: 9357 parsing failures.
row col  expected    actual                                                    file
  1  -- 1 columns 6 columns 'C:/Users/BastianSol/Downloads/AirqualityUCI.zip'
  2  -- 1 columns 5 columns 'C:/Users/BastianSol/Downloads/AirqualityUCI.zip'
  3  -- 1 columns 6 columns 'C:/Users/BastianSol/Downloads/AirqualityUCI.zip'
  4  -- 1 columns 6 columns 'C:/Users/BastianSol/Downloads/AirqualityUCI.zip'
  5  -- 1 columns 6 columns 'C:/Users/BastianSol/Downloads/AirqualityUCI.zip'
... ... ......... ......... ..............................................
See problems(...) for more details.

>
> AirQualityUCI <- read_delim("C:/Users/BastianSol/Downloads/AirqualityUCI.zip", ";", escape
_double = FALSE, trim_ws = TRUE)
Multiple files in zip: reading 'AirQualityUCI.csv'
Parsed with column specification:
cols(
  Date = col_character(),
  Time = col_character(),
  `CO(GT)` = col_character(),
  `PT08.S1(CO)` = col_double(),
  `NMHC(GT)` = col_double(),
  `C6H6(GT)` = col_character(),
  `PT08.S2(NMHC)` = col_double(),
  `NOx(GT)` = col_double(),
  `PT08.S3(NOx)` = col_double(),
  `NO2(GT)` = col_double(),
  `PT08.S4(NO2)` = col_double(),
  `PT08.S5(O3)` = col_double(),
  T = col_number(),
  RH = col_number(),
  AH = col_character(),
  X16 = col_logical(),
  X17 = col_logical()
)
Warning message:
Missing column names filled in: 'X16' [16], 'X17' [17]
>
> View(AirQualityUCI)
```

b. Create Univariate for all the columns.

```
AirQualityUCI[AirQualityUCI==-200.0]<-NA
for(i in 1:ncol(AirQualityUCI))
  {AirQualityUCI[is.na(AirQualityUCI[,i]),i] <- mean(AirQualityUCI[,i], na.rm = TRUE)}

summary(AirQualityUCI)

AirQualityUCI[7:14,]

hist(AirQualityUCI$`NOx(GT)`,col="red")

dotchart(AirQualityUCI$`PT08.S2(NMHC)`,
         labels = row.names(AirQualityUCI$`PT08.S1(CO)`),cex=0.5, color = "blue")

pairs(AirQualityUCI[7:14], "Date Time CO(GT) PT08.S1(CO) NMHC(GT) C6H6(GT) PT08.S2(NMHC)")

univariateTable(~ Date + Time + CO(GT) + PT08.S1(CO) + NMHC(GT) + C6H6(GT) + PT08.S2(NMHC) + NOx(GT) +
PT08.S3(NOx) , data = AirQualityUCI)
```

c. Check for missing values in all columns.

```
> colSums(is.na(AirQualityUCI))
          Date           Time        CO(GT)   PT08.S1(CO)       NMHC(GT)       C6H6(GT) PT08.S2(
NMHC)        NOx(GT)
           114            114           1706            480           8557            114
480           1753
 PT08.S3(NOx)       NO2(GT)   PT08.S4(NO2)   PT08.S5(O3)              T             RH
AH            X16
           480           1756            480            480            480            480
480           9471
           X17
          9471
```

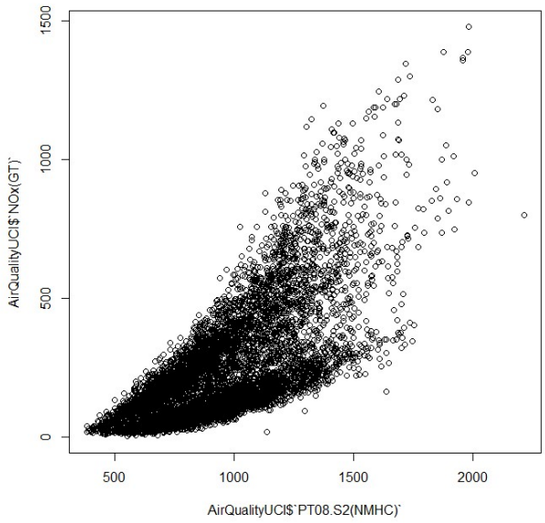
```
library(mice)
md.pattern(AirQualityUCI)
str(AirQualityUCI)

library(Amelia)
missmap(AirQualityUCI, col=c("black", "grey"), legend=FALSE)
```

d. Impute the missing values using appropriate methods

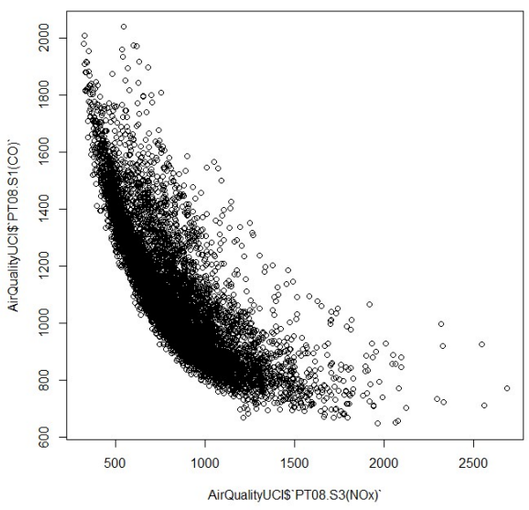
```
colSums(is.na(AirQualityUCI))

library(plyr)
AirQualityUCI[AirQualityUCI==-200.0]<-NA

for(i in 1:ncol(AirQualityUCI))
  { AirQualityUCI[is.na(AirQualityUCI[,i]),i] <- mean(AirQualityUCI[,i], na.rm = TRUE)}

summary(AirQualityUCI)
```

e. Create bi-variate analysis for all relationships

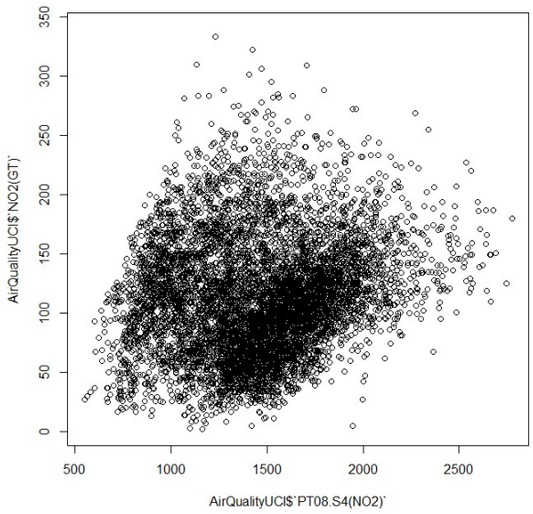
```
> plot(AirQualityUCI$`NOx(GT)`~AirQualityUCI$`PT08.S2(NMHC)`)
```



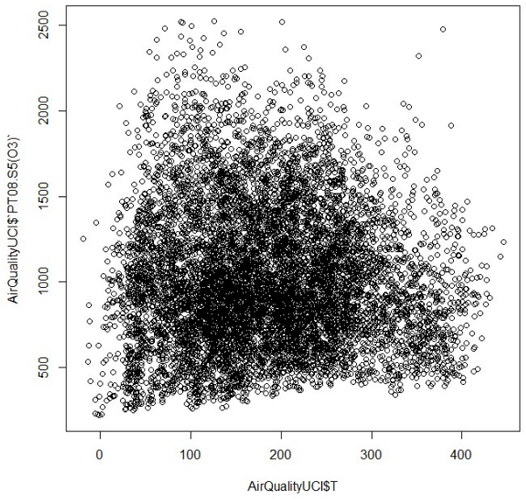

```
> plot(AirQualityUCI$`PT08.S1(CO)`~AirQualityUCI$`PT08.S3(NOx)`)
```



```
> plot(AirQualityUCI$`NO2(GT)`~AirQualityUCI$`PT08.S4(NO2)`)
```




```
> plot(AirQualityUCI$`PT08.S5(O3)`~AirQualityUCI$T)
```



f. Test relevant hypothesis for valid relations

```
plot(AirQualityUCI$`PT08.S1(CO)`,AirQualityUCI$T)
lm(formula=AirQualityUCI$`PT08.S3(NOx)`~AirQualityUCI$`NOx(GT)`)
lm(formula = AirQualityUCI$PT08.S1(CO)~AirQualityUCI$T)
lm(formula = AirQualityUCI$NMHC(GT)~AirQualityUCI$PT08.S2(NMHC))
plot(AirQualityUCI$PT08.S5(O3),AirQualityUCI$NOx(GT))
lm(formula =AirQualityUCI$PT08.S5(O3)~AirQualityUCI$NOx(GT) )

pnorm(1.49)
pnorm(1.097)
qnorm(0.9318879)
qnorm(0.8636793)
```

g. Create cross tabulations with derived variables
h. check for trends and patterns in time series
i. Find out the most polluted time of the day and the name of the chemical compound.