ACADGILD

SESSION 10: Correlations

Assignment 1

Submitted by: Uday Kumar
Udaykumarr019@gmail.com
(M):+91-8123431864

Data Analytics

Table of Contents

1. Problem Statement

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
b) Create Univariate for all the columns.
c) Check for missing values in all columns.
d) Impute the missing values using appropriate methods
e) Create bi-variate analysis for all relationships
f) Test relevant hypothesis for valid relations
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

2. Solution

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

The R-script for the given problem is as follows:

```
library(readxl)
#AirQualityUCI <-read_excel(unzip("E:/uday/acadgild data analytics/supporting files/AirQualityUCI.zip"))
AirQualityUCI <- read_excel("E:/uday/acadgild data analytics/supporting files/AirQualityUCI.xlsx")
View(AirQualityUCI)
dim(AirQualityUCI)
str(AirQualityUCI)
```

The output of the R-Script (from Console window) is given as follows:

```
> library(readxl)
> AirQualityUCI <-read_excel(unzip("E:/uday/acadgild data analytics/supporting files/AirQualityUCI.zip"))
> View(AirQualityUCI)
```

Assignment 10_1.R* | AirQualityUCI

Filter

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)
1	2004-03-10	1899-12-31 18:00:00	2.6	1360.000	150	11.881723	1045.500	166	1056.2500	113	1692.000
2	2004-03-10	1899-12-31 19:00:00	2.0	1292.250	112	9.397165	954.750	103	1173.7500	92	1558.750
3	2004-03-10	1899-12-31 20:00:00	2.2	1402.000	88	8.997817	939.250	131	1140.0000	114	1554.500
4	2004-03-10	1899-12-31 21:00:00	2.2	1375.500	80	9.228796	948.250	172	1092.0000	122	1583.750
5	2004-03-10	1899-12-31 22:00:00	1.6	1272.250	51	6.518224	835.500	131	1205.0000	116	1490.000
6	2004-03-10	1899-12-31 23:00:00	1.2	1197.000	38	4.741012	750.250	89	1336.5000	96	1393.000
7	2004-03-11	1899-12-31 00:00:00	1.2	1185.000	31	3.624399	689.500	62	1461.7500	77	1332.750
8	2004-03-11	1899-12-31 01:00:00	1.0	1136.250	31	3.326677	672.000	62	1453.2500	76	1332.750
9	2004-03-11	1899-12-31 02:00:00	0.9	1094.000	24	2.339416	608.500	45	1579.0000	60	1276.000
10	2004-03-11	1899-12-31 03:00:00	0.6	1009.750	19	1.696658	560.750	-200	1705.0000	-200	1234.750
11	2004-03-11	1899-12-31 04:00:00	-200.0	1011.000	14	1.293620	526.750	21	1817.5000	34	1196.750
12	2004-03-11	1899-12-31 05:00:00	0.7	1066.000	8	1.133431	512.000	16	1918.0000	28	1182.000
13	2004-03-11	1899-12-31 06:00:00	0.7	1051.750	16	1.603768	553.250	34	1738.2500	48	1221.250
14	2004-03-11	1899-12-31 07:00:00	1.1	1144.000	29	3.243618	667.000	98	1489.7500	82	1339.000
15	2004-03-11	1899-12-31 08:00:00	2.0	1333.250	64	8.013773	899.750	174	1136.0000	112	1517.000
16	2004-03-11	1899-12-31 09:00:00	2.2	1351.000	87	9.540643	960.250	129	1079.0000	101	1582.750
17	2004-03-11	1899-12-31 10:00:00	1.7	1233.250	77	6.335782	827.250	112	1218.0000	98	1445.750
18	2004-03-11	1899-12-31 11:00:00	1.5	1178.750	43	4.971584	762.000	95	1327.5000	92	1361.750
19	2004-03-11	1899-12-31 12:00:00	1.6	1236.000	61	5.216919	774.250	104	1301.2500	95	1401.250
20	2004-03-11	1899-12-31 13:00:00	1.9	1285.500	63	7.269933	868.500	146	1162.2500	112	1536.750
21	2004-03-11	1899-12-31 14:00:00	2.9	1371.000	164	11.539007	1033.500	207	983.2500	128	1730.250
22	2004-03-11	1899-12-31 15:00:00	2.2	1310.000	79	8.826223	932.500	184	1081.7500	126	1646.500

Showing 1 to 23 of 9,357 entries

```
> dim(AirQualityUCI)
[1] 9357   15
> str(AirQualityUCI)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':     9357 obs. of  15 variables:
 $ Date          : POSIXct, format: "2004-03-10" "2004-03-10" "2004-03-10" "2004-03-
10" ...
 $ Time          : POSIXct, format: "1899-12-31 18:00:00" "1899-12-31 19:00:00"
"1899-12-31 20:00:00" "1899-12-31 21:00:00" ...
 $ CO(GT)        : num  2.6 2 2.2 2.2 1.6 1.2 1.2 1 0.9 0.6 ...
 $ PT08.S1(CO)   : num  1360 1292 1402 1376 1272 ...
 $ NMHC(GT)      : num  150 112 88 80 51 38 31 31 24 19 ...
 $ C6H6(GT)      : num  11.88 9.4 9 9.23 6.52 ...
 $ PT08.S2(NMHC): num  1046 955 939 948 836 ...
 $ NOx(GT)       : num  166 103 131 172 131 89 62 62 45 -200 ...
 $ PT08.S3(NOx)  : num  1056 1174 1140 1092 1205 ...
 $ NO2(GT)       : num  113 92 114 122 116 96 77 76 60 -200 ...
 $ PT08.S4(NO2)  : num  1692 1559 1554 1584 1490 ...
 $ PT08.S5(O3)   : num  1268 972 1074 1203 1110 ...
 $ T             : num  13.6 13.3 11.9 11 11.2 ...
 $ RH            : num  48.9 47.7 54 60 59.6 ...
 $ AH            : num  0.758 0.725 0.75 0.787 0.789 ...
```

Conclusion/Interpretation:

The file is read from Zip format and is viewed with name AirQualityUCI.

b) Create Univariate for all the columns.

The R-script for the given problem is as follows:

```
library(psych)
describe(Air)
```

The output of the R-Script (from Console window) is given as follows:

```
> library(psych)
> describe(Air)
              vars    n    mean     sd   median  trimmed    mad  min     max
range  skew kurtosis   se
Date             1 9357     NaN     NA       NA      NaN     NA  Inf    -Inf
-Inf    NA       NA   NA
Time             2 9357     NaN     NA       NA      NaN     NA  Inf    -Inf
-Inf    NA       NA   NA
CO(GT)           3 9357  -34.21  77.66     1.50   -18.41   1.48 -200   11.90
211.90 -1.67     0.78 0.80
PT08.S1(CO)      4 9357 1048.87 329.82  1052.50  1069.72 218.19 -200 2039.75
2239.75 -1.72     5.83 3.41
NMHC(GT)         5 9357 -159.09 139.79 -200.00 -200.00   0.00 -200 1189.00
1389.00  4.07    18.85 1.45
C6H6(GT)         6 9357    1.87  41.38     7.89     8.75   6.62 -200   63.74
263.74 -4.51    19.17 0.43
PT08.S2(NMHC)    7 9357  894.48 342.32   894.50   907.06 288.37 -200 2214.00
2414.00 -0.79     2.37 3.54
NOx(GT)          8 9357  168.60 257.42   141.00   147.72 161.31 -200 1479.00
1679.00  0.82     1.50 2.66
PT08.S3(NOx)     9 9357  794.87 321.98   794.25   799.84 238.70 -200 2682.75
2882.75 -0.38     3.10 3.33
NO2(GT)         10 9357   58.14 126.93    96.00    72.32  59.30 -200  339.70
539.70 -1.23     0.27 1.31
PT08.S4(NO2)    11 9357 1391.36 467.19  1445.50  1426.54 349.15 -200 2775.00
2975.00 -1.24     3.26 4.83
PT08.S5(O3)     12 9357  974.95 456.92   942.00   972.05 403.64 -200 2522.75
2722.75 -0.03     0.64 4.72
T               13 9357    9.78  43.20    17.20    17.39   9.71 -200   44.60
244.60 -4.44    18.76 0.45
RH              14 9357   39.48  51.22    48.55    48.04  20.65 -200   88.73
288.73 -3.93    15.75 0.53
AH              15 9357   -6.84  38.98     0.98     0.99   0.45 -200    2.23
202.23 -4.75    20.60 0.40
```

Conclusion/Interpretation:

Univariate for all the columns is created using describe() function

c) Check for missing values in all columns.

The R-script for the given problem is as follows:

```
col1<- mapply(anyNA,AirQualityUCI)
col1
summary(AirQualityUCI)
is.na(AirQualityUCI)
```

```
#or

AirQualityUCI[AirQualityUCI == -200] <- NA
View(AirQualityUCI)
library(VIM)
aggr(AirQualityUCI, col=c('pink','yellow'),
    numbers=TRUE, sortVars=TRUE,
    labels=names(AirQualityUCI), cex.axis=.7,
    gap=3, ylab=c("Missing data","Pattern"))    # graphical presentation of NAs

sapply(AirQualityUCI, function(x) sum(is.na(x)))     # count of NAs

AirQualityUCI$`NMHC(GT)` <- NULL
```

The output of the R-Script (from Console window) is given as follows:

```
> col1<- mapply(anyNA,AirQualityUCI)
> col1
         Date          Time         CO(GT)    PT08.S1(CO)       NMHC(GT)
C6H6(GT)
        FALSE         FALSE           TRUE           TRUE           TRUE
TRUE
PT08.S2(NMHC)       NOx(GT)   PT08.S3(NOx)        NO2(GT)   PT08.S4(NO2)
PT08.S5(O3)
         TRUE          TRUE           TRUE           TRUE           TRUE
TRUE
            T            RH            AH
         TRUE          TRUE           TRUE
> summary(AirQualityUCI)
      Date                         Time                          CO(GT)
PT08.S1(CO)
 Min.   :2004-03-10 00:00:00   Min.   :1899-12-31 00:00:00   Min.   : 0.100
Min.   : 647.2
 1st Qu.:2004-06-16 00:00:00   1st Qu.:1899-12-31 05:00:00   1st Qu.: 1.100
1st Qu.: 936.8
 Median :2004-09-21 00:00:00   Median :1899-12-31 11:00:00   Median : 1.800
Median :1063.0
 Mean   :2004-09-21 04:30:05   Mean   :1899-12-31 11:29:55   Mean   : 2.153
Mean   :1099.7
 3rd Qu.:2004-12-28 00:00:00   3rd Qu.:1899-12-31 18:00:00   3rd Qu.: 2.900
3rd Qu.:1231.2
 Max.   :2005-04-04 00:00:00   Max.   :1899-12-31 23:00:00   Max.   :11.900
Max.   :2039.8
                                                            NA's   :1683
NA's   :366
    NMHC(GT)          C6H6(GT)       PT08.S2(NMHC)        NOx(GT)
PT08.S3(NOx)
 Min.   :   7.0   Min.   : 0.149   Min.   : 383.2   Min.   :   2.0   Min.   :
322.0
 1st Qu.:  67.0   1st Qu.: 4.437   1st Qu.: 734.4   1st Qu.:  98.0   1st Qu.:
657.9
```

```
 Median : 150.0   Median : 8.240   Median : 909.0   Median : 179.8   Median :
805.5
 Mean   : 218.8   Mean   :10.083   Mean   : 939.0   Mean   : 246.9   Mean   :
835.4
 3rd Qu.: 297.0   3rd Qu.:13.989   3rd Qu.:1116.2   3rd Qu.: 326.0   3rd Qu.:
969.2
 Max.   :1189.0   Max.   :63.742   Max.   :2214.0   Max.   :1479.0   Max.
:2682.8
 NA's   :8443     NA's   :366      NA's   :366      NA's   :1639     NA's
:366
    NO2(GT)        PT08.S4(NO2)   PT08.S5(O3)           T                RH
 Min.   :  2.0   Min.   : 551   Min.   : 221.0   Min.   :-1.90   Min.   :
9.175
 1st Qu.: 78.0   1st Qu.:1227   1st Qu.: 731.4   1st Qu.:11.79   1st
Qu.:35.812
 Median :109.0   Median :1463   Median : 963.2   Median :17.75   Median
:49.550
 Mean   :113.1   Mean   :1456   Mean   :1022.8   Mean   :18.32   Mean
:49.232
 3rd Qu.:142.0   3rd Qu.:1674   3rd Qu.:1273.4   3rd Qu.:24.40   3rd
Qu.:62.500
 Max.   :339.7   Max.   :2775   Max.   :2522.8   Max.   :44.60   Max.
:88.725
 NA's   :1642    NA's   :366    NA's   :366      NA's   :366     NA's   :366
       AH
 Min.   :0.1847
 1st Qu.:0.7368
 Median :0.9954
 Mean   :1.0255
 3rd Qu.:1.3137
 Max.   :2.2310
 NA's   :366
> is.na(AirQualityUCI)
         Date  Time CO(GT) PT08.S1(CO) NMHC(GT) C6H6(GT) PT08.S2(NMHC)
NOx(GT) PT08.S3(NOx)
   [1,] FALSE FALSE  FALSE       FALSE    FALSE    FALSE          FALSE
FALSE         FALSE
   [2,] FALSE FALSE  FALSE       FALSE    FALSE    FALSE          FALSE
FALSE         FALSE
   [3,] FALSE FALSE  FALSE       FALSE    FALSE    FALSE          FALSE
FALSE         FALSE
   [4,] FALSE FALSE  FALSE       FALSE    FALSE    FALSE          FALSE
FALSE         FALSE
   [5,] FALSE FALSE  FALSE       FALSE    FALSE    FALSE          FALSE
FALSE         FALSE
   [6,] FALSE FALSE  FALSE       FALSE    FALSE    FALSE          FALSE
FALSE         FALSE
   [7,] FALSE FALSE  FALSE       FALSE    FALSE    FALSE          FALSE
FALSE         FALSE
   [8,] FALSE FALSE  FALSE       FALSE    FALSE    FALSE          FALSE
FALSE         FALSE
   [9,] FALSE FALSE  FALSE       FALSE    FALSE    FALSE          FALSE
FALSE         FALSE
  [10,] FALSE FALSE  FALSE       FALSE    FALSE    FALSE          FALSE
TRUE         FALSE
  [11,] FALSE FALSE   TRUE       FALSE    FALSE    FALSE          FALSE
FALSE         FALSE
```

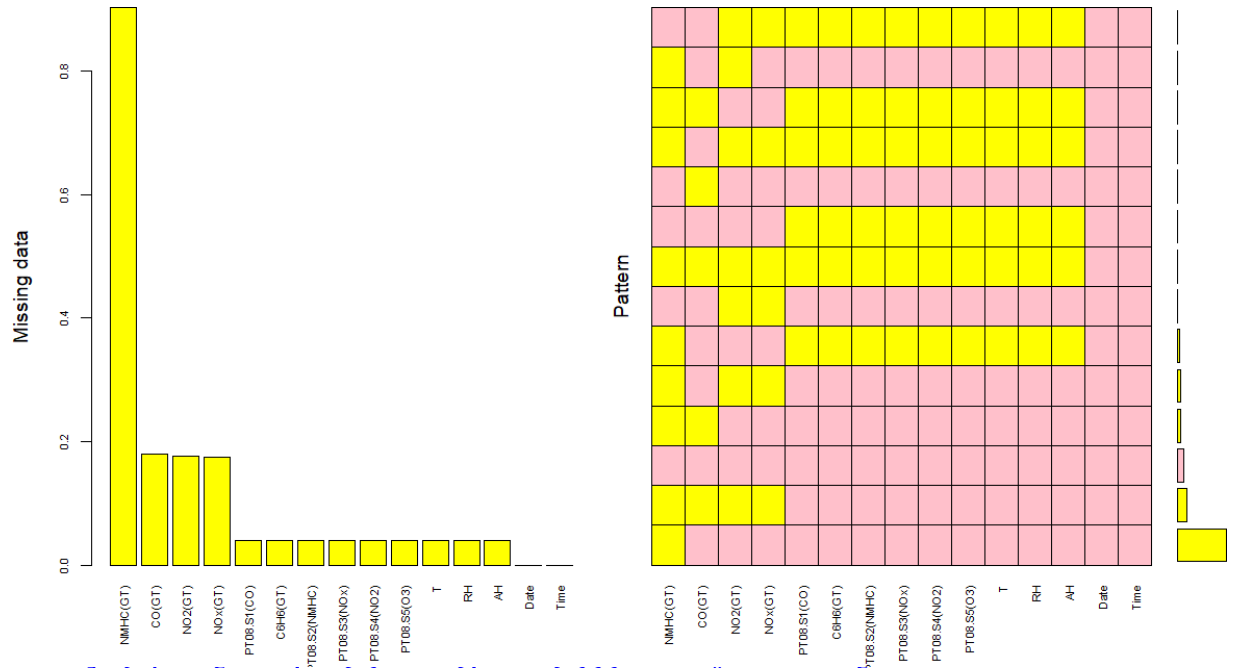
```
  [12,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [13,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [14,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [15,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [16,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [17,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [18,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [19,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [20,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [21,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [22,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [23,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [24,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [25,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [26,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [27,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [28,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [29,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [30,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [31,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [32,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [33,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [34,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
TRUE         FALSE
  [35,] FALSE FALSE    TRUE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [36,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [37,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [38,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
  [39,] FALSE FALSE   FALSE       FALSE    FALSE    FALSE         FALSE
FALSE        FALSE
```

```
  [40,] FALSE FALSE   TRUE       FALSE    TRUE    FALSE         FALSE
TRUE        FALSE
  [41,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [42,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [43,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [44,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [45,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [46,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [47,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [48,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [49,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [50,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [51,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [52,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [53,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [54,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [55,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [56,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [57,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [58,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
TRUE        FALSE
  [59,] FALSE FALSE   TRUE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [60,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [61,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [62,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [63,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [64,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [65,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
  [66,] FALSE FALSE  FALSE       FALSE   FALSE   FALSE         FALSE
FALSE       FALSE
       NO2(GT) PT08.S4(NO2) PT08.S5(O3)    T    RH    AH
  [1,]   FALSE       FALSE        FALSE FALSE FALSE FALSE
  [2,] FALSE FALSE FALSE FALSE FALSE FALSE
```

```
 [3,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
 [4,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
 [5,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
 [6,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
 [7,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
 [8,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
 [9,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[10,]    TRUE         FALSE        FALSE FALSE FALSE FALSE
[11,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[12,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[13,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[14,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[15,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[16,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[17,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[18,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[19,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[20,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[21,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[22,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[23,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[24,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[25,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[26,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[27,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[28,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[29,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[30,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[31,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[32,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[33,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[34,]    TRUE         FALSE        FALSE FALSE FALSE FALSE
[35,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[36,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[37,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[38,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[39,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[40,]    TRUE         FALSE        FALSE FALSE FALSE FALSE
[41,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[42,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[43,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[44,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[45,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[46,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[47,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[48,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[49,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[50,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[51,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[52,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[53,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[54,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[55,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[56,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[57,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
[58,]    TRUE         FALSE        FALSE FALSE FALSE FALSE
[59,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
```

```
  [60,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
  [61,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
  [62,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
  [63,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
  [64,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
  [65,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
  [66,]   FALSE         FALSE        FALSE FALSE FALSE FALSE
 [ reached getOption("max.print") -- omitted 9291 rows ]
```

```
> Air[Air == -200] <- NA
> View(Air)
> library(VIM)
> aggr(Air, col=c('pink','yellow'),
+      numbers=TRUE, sortVars=TRUE,
+      labels=names(Air), cex.axis=.7,
+      gap=3, ylab=c("Missing data","Pattern"))    # graphical presentation of NAs

 Variables sorted by number of missings:
      Variable      Count
      NMHC(GT)  0.9023191
        CO(GT)  0.1798653
       NO2(GT)  0.1754836
       NOx(GT)  0.1751630
    PT08.S1(CO)  0.0391151
       C6H6(GT)  0.0391151
  PT08.S2(NMHC)  0.0391151
   PT08.S3(NOx)  0.0391151
   PT08.S4(NO2)  0.0391151
    PT08.S5(O3)  0.0391151
              T  0.0391151
             RH  0.0391151
             AH  0.0391151
           Date  0.0000000
           Time  0.0000000
```




```
> sapply(Air, function(x) sum(is.na(x)))      # count of NAs
          Date           Time        CO(GT)    PT08.S1(CO)       NMHC(GT)
             0              0          1683           366          8443
     C6H6(GT)  PT08.S2(NMHC)       NOx(GT)    PT08.S3(NOx)       NO2(GT)
           366            366          1639           366          1642
  PT08.S4(NO2)   PT08.S5(O3)             T             RH             AH
           366            366           366           366           366
```

```
>Air$`NMHC(GT)` <- NULL
```

Conclusion/Interpretation:

Variable NMHC(GT) is having 90% of missing values. Hence, NMHC(GT) is not considered and omitted from the data frame

d) Impute the missing values using appropriate methods

The R-script for the given problem is as follows:

```
colSums(is.na(AirQualityUCI))
library(plyr)
AirQualityUCI[AirQualityUCI==-200.0]<-NA
for(i in 1:n
col(AirQualityUCI)){ AirQualityUCI[is.na(AirQualityUCI[,i]),i] <- mean(AirQualityUCI[,i], na.rm = TRUE)}
summary(AirQualityUCI)
```

The output of the R-Script (from Console window) is given as follows:

```
> 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)
      Date                 Time                          CO(GT)
 Min.   :2004-03-10 00:00:00   Min.   :1899-12-31 00:00:00   Min.   : 0.100
 1st Qu.:2004-06-16 00:00:00   1st Qu.:1899-12-31 05:00:00   1st Qu.: 1.200
 Median :2004-09-21 00:00:00   Median :1899-12-31 11:00:00   Median : 2.153
 Mean   :2004-09-21 04:30:05   Mean   :1899-12-31 11:29:55   Mean   : 2.153
 3rd Qu.:2004-12-28 00:00:00   3rd Qu.:1899-12-31 18:00:00   3rd Qu.: 2.600
 Max.   :2005-04-04 00:00:00   Max.   :1899-12-31 23:00:00   Max.   :11.900
  PT08.S1(CO)        NMHC(GT)        C6H6(GT)       PT08.S2(NMHC)
 Min.   : 647.2   Min.   :   7.0   Min.   : 0.149   Min.   : 383.2
 1st Qu.: 941.2   1st Qu.: 218.8   1st Qu.: 4.591   1st Qu.: 742.5
 Median :1074.5   Median : 218.8   Median : 8.593   Median : 923.2
 Mean   :1099.7   Mean   : 218.8   Mean   :10.083   Mean   : 939.0
 3rd Qu.:1221.2   3rd Qu.: 218.8   3rd Qu.:13.636   3rd Qu.:1104.8
 Max.   :2039.8   Max.   :1189.0   Max.   :63.741   Max.   :2214.0

    NOx(GT)        PT08.S3(NOx)      NO2(GT)       PT08.S4(NO2)   PT08.S5(O3)
 Min.   :   2.0   Min.   : 322.0   Min.   :  2.0   Min.   : 551   Min.   : 221.0
 1st Qu.: 112.0   1st Qu.: 665.5   1st Qu.: 85.9   1st Qu.:1242   1st Qu.: 741.8
 Median : 229.0   Median : 817.5   Median :113.1   Median :1456   Median : 982.5
 Mean   : 246.9   Mean   : 835.4   Mean   :113.1   Mean   :1456   Mean   :1022.8
 3rd Qu.: 284.2   3rd Qu.: 960.2   3rd Qu.:133.0   3rd Qu.:1662   3rd Qu.:1255.2
 Max.   :1479.0   Max.   :2682.8   Max.   :339.7   Max.   :2775   Max.   :2522.8
       T                RH               AH
 Min.   :-1.90   Min.   : 9.175   Min.   :0.1847
 1st Qu.:12.03   1st Qu.:36.550   1st Qu.:0.7461
 Median :18.27   Median :49.232   Median :1.0154
 Mean   :18.32   Mean   :49.232   Mean   :1.0255
 3rd Qu.:24.07   3rd Qu.:61.875   3rd Qu.:1.2962
 Max.   :44.60   Max.   :88.725   Max.   :2.2310
```

Conclusion/Interpretation:

Missing values are hence imputed

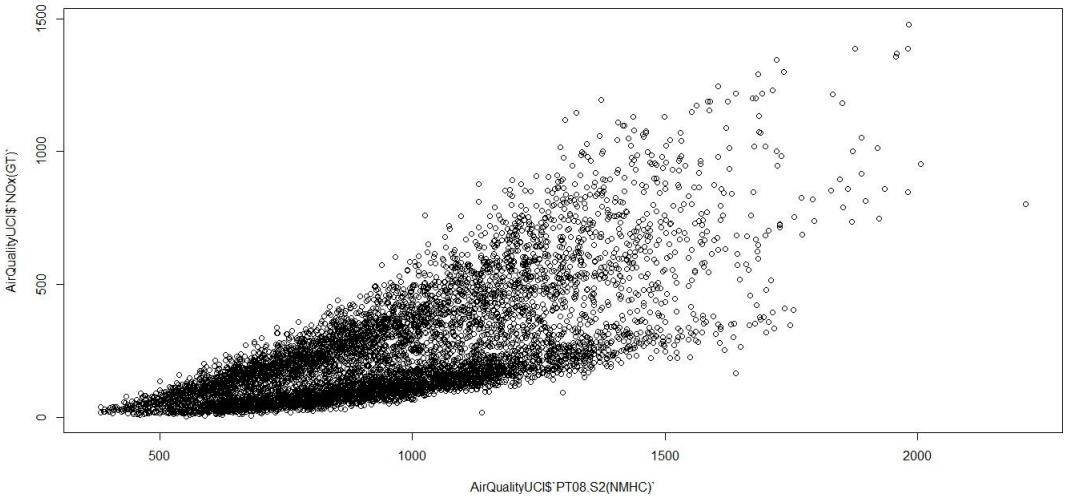
e) Create bi-variate analysis for all relationships

The R-script for the given problem is as follows:

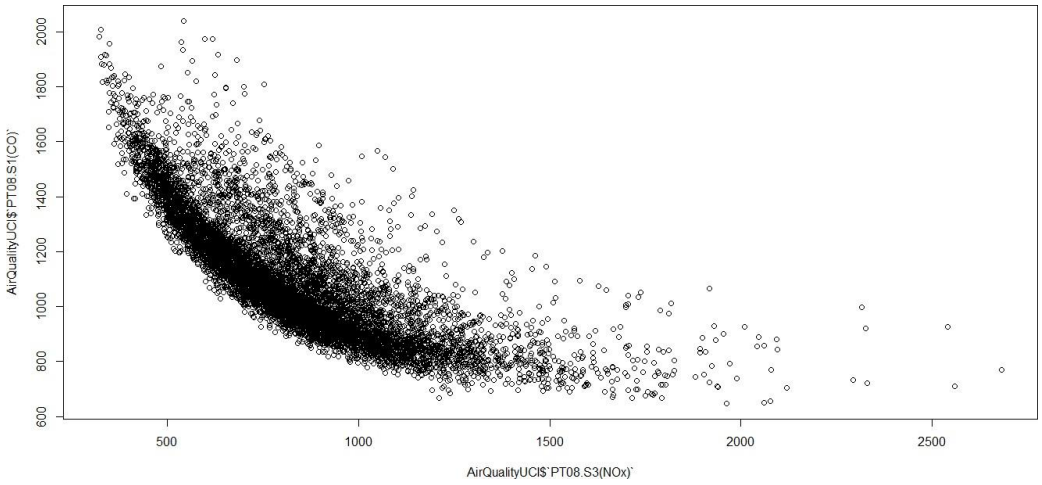
```
summary(AirQualityUCI)
plot(AirQualityUCI$`NOx(GT)`~AirQualityUCI$`PT08.S2(NMHC)`)
plot(AirQualityUCI$`PT08.S1(CO)`~AirQualityUCI$`PT08.S3(NOx)`)
plot(AirQualityUCI$`NO2(GT)`~AirQualityUCI$`PT08.S4(NO2)`)
plot(AirQualityUCI$`PT08.S5(O3)`~AirQualityUCI$T)
#or
pairs(AirQualityUCI)    # graph

final <- complete
final$Date <- AirQualityUCI$Date
final$Time <- AirQualityUCI$Time
library(stringr)
AirQualityUCI$Time1 <- sub(".+? ", "", AirQualityUCI$Time)
AirQualityUCI$datetime <- as.POSIXct(paste(AirQualityUCI$Date,
AirQualityUCI$Time1), format="%Y-%m-%d %H:%M:%S")
View(AirQualityUCI)
str(AirQualityUCI)
```

The output of the R-Script (from Console window) is given as follows:

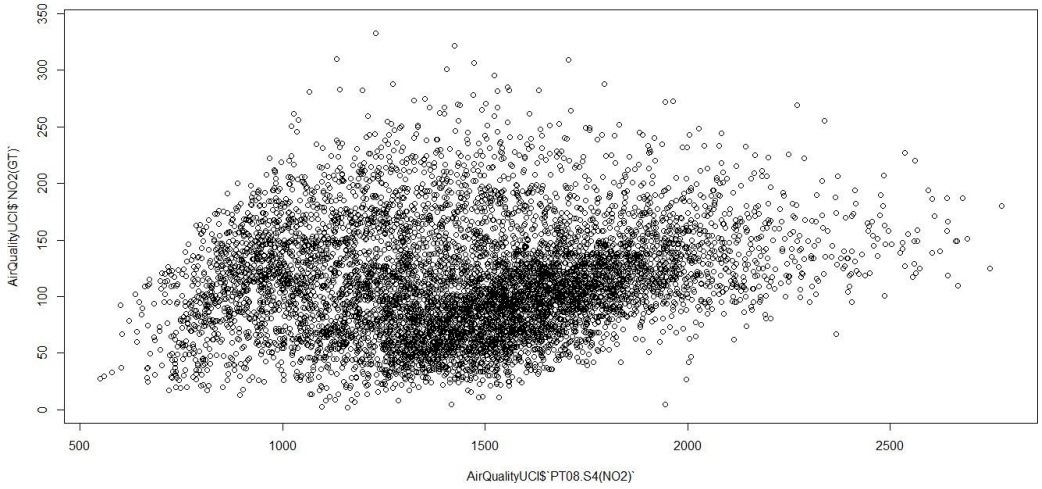
```
> plot(AirQualityUCI$`NOx(GT)`~AirQualityUCI$`PT08.S2(NMHC)`)
```



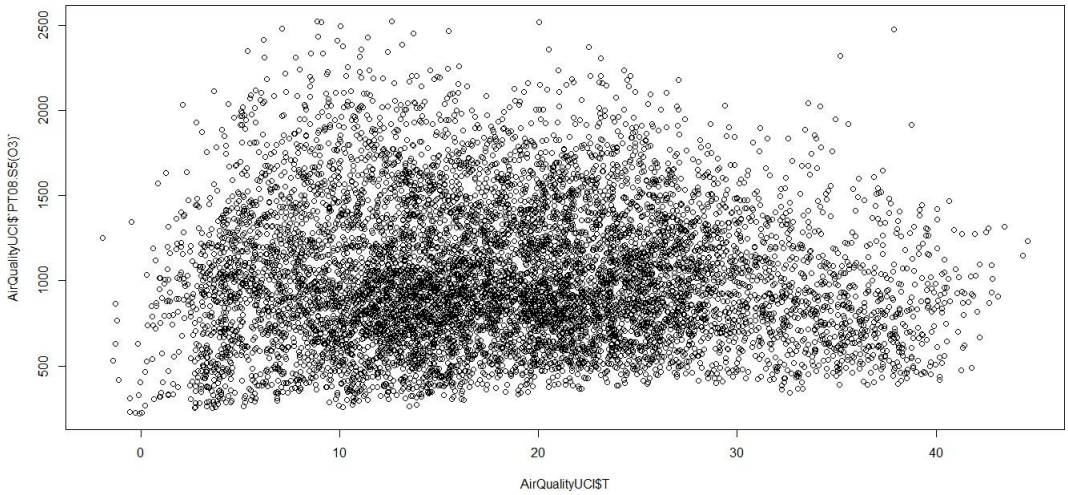

```
> plot(AirQualityUCI$`PT08.S1(CO)`~AirQualityUCI$`PT08.S3(NOx)`)
```



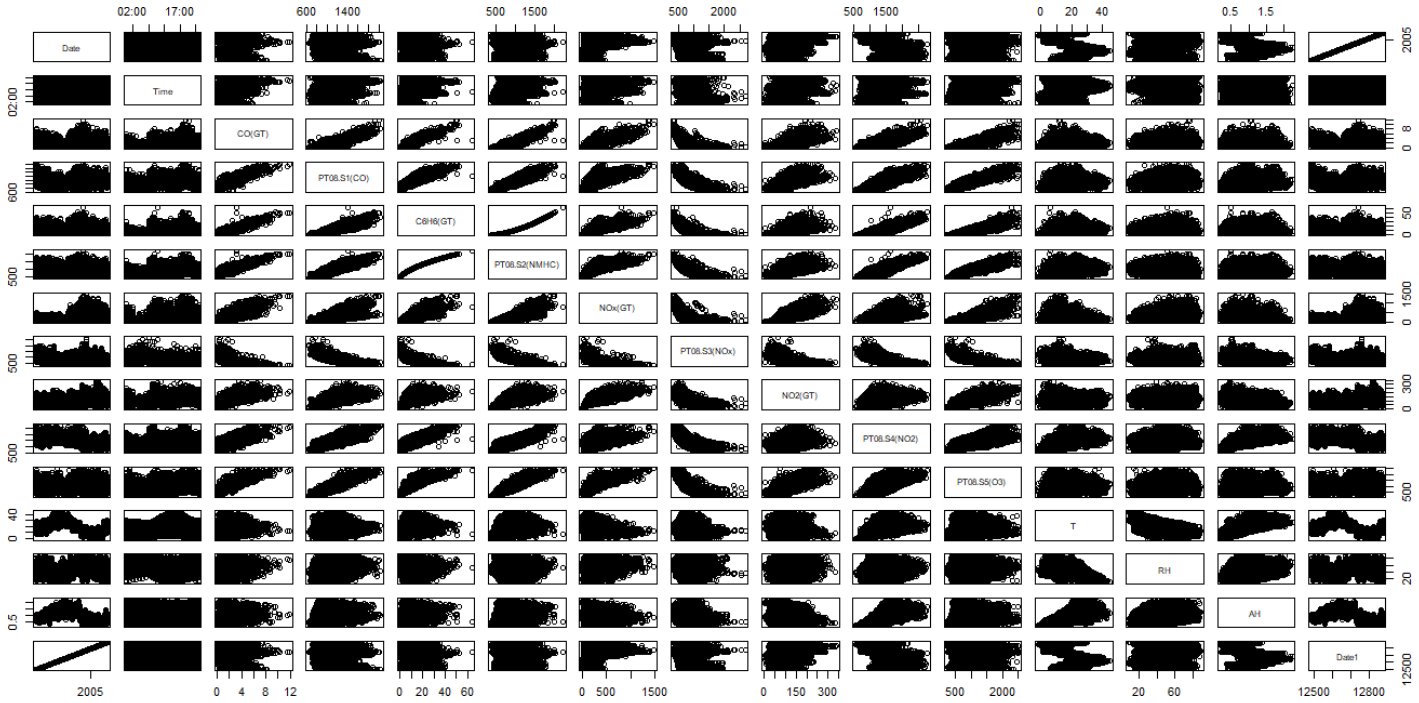

```
> plot(AirQualityUCI$`NO2(GT)`~AirQualityUCI$`PT08.S4(NO2)`)
```




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



```
> pairs(AirQualityUCI)    # graph
```




```
> final <- complete
> final$Date <- AirQualityUCI$Date
> final$Time <- AirQualityUCI$Time
> library(stringr)
> AirQualityUCI$Time1 <- sub(".+? ", "", AirQualityUCI$Time)
> AirQualityUCI$datetime <- as.POSIXct(paste(AirQualityUCI$Date,
AirQualityUCI$Time1), format="%Y-%m-%d %H:%M:%S")
> View(AirQualityUCI)
> str(AirQualityUCI)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':     9357 obs. of  17 variables:
 $ Date         : POSIXct, format: "2004-03-10" "2004-03-10" "2004-03-10" ...
 $ Time         : POSIXct, format: "1899-12-31 18:00:00" "1899-12-31
19:00:00" "1899-12-31 20:00:00" ...
 $ CO(GT)       : num  2.6 2 2.2 2.2 1.6 1.2 1.2 1 0.9 0.6 ...
 $ PT08.S1(CO)  : num  1360 1292 1402 1376 1272 ...
 $ C6H6(GT)     : num  11.88 9.4 9 9.23 6.52 ...
 $ PT08.S2(NMHC): num  1046 955 939 948 836 ...
 $ NOx(GT)      : num  166 103 131 172 131 89 62 62 45 NA ...
 $ PT08.S3(NOx) : num  1056 1174 1140 1092 1205 ...
 $ NO2(GT)      : num  113 92 114 122 116 96 77 76 60 NA ...
 $ PT08.S4(NO2) : num  1692 1559 1554 1584 1490 ...
 $ PT08.S5(O3)  : num  1268 972 1074 1203 1110 ...
 $ T            : num  13.6 13.3 11.9 11 11.2 ...
 $ RH           : num  48.9 47.7 54 60 59.6 ...
 $ AH           : num  0.758 0.725 0.75 0.787 0.789 ...
 $ Date1        : num  12487 12487 12487 12487 12487 ...
 $ Time1        : chr  "18:00:00" "19:00:00" "20:00:00" "21:00:00" ...
 $ datetime     : POSIXct, format: "2004-03-10 18:00:00" "2004-03-10
19:00:00" "2004-03-10 20:00:00" ...
```

Conclusion/Interpretation:

Bi-variate analysis for all relationships are done and plotted.

f) Test relevant hypothesis for valid relations

The R-script for the given problem is as follows:

```
t.test(AirQualityUCI$`CO(GT)`, AirQualityUCI$`PT08.S1(CO)`, paired = T)
t.test(AirQualityUCI$`C6H6(GT)`, AirQualityUCI$`PT08.S2(NMHC)`, paired = T)
t.test(AirQualityUCI$`NOx(GT)`, AirQualityUCI$`PT08.S3(NOx)`, paired = T)

mod <- lm(AirQualityUCI$`CO(GT)`~AirQualityUCI$Date1)
summary(mod)

mod <- lm(AirQualityUCI$`CO(GT)`~AirQualityUCI$T)
summary(mod)

mod <- lm(AirQualityUCI$`CO(GT)`~AirQualityUCI$RH)
summary(mod)
```

The output of the R-Script (from Console window) is given as follows:

```
> t.test(AirQualityUCI$`CO(GT)`, AirQualityUCI$`PT08.S1(CO)`, paired = T)

        Paired t-test

data:  AirQualityUCI$`CO(GT)` and AirQualityUCI$`PT08.S1(CO)`
t = -436.85, df = 7343, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1113.299 -1103.352
sample estimates:
mean of the differences
              -1108.325

> t.test(AirQualityUCI$`C6H6(GT)`, AirQualityUCI$`PT08.S2(NMHC)`, paired = T)

        Paired t-test

data:  AirQualityUCI$`C6H6(GT)` and AirQualityUCI$`PT08.S2(NMHC)`
t = -339.41, df = 8990, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -934.3112 -923.5812
sample estimates:
mean of the differences
              -928.9462

> t.test(AirQualityUCI$`NOx(GT)`, AirQualityUCI$`PT08.S3(NOx)`, paired = T)

        Paired t-test

data:  AirQualityUCI$`NOx(GT)` and AirQualityUCI$`PT08.S3(NOx)`
t = -118.66, df = 7395, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
```

```
 -591.8554 -572.6187
sample estimates:
mean of the differences
              -582.2371

> mod <- lm(AirQualityUCI$`CO(GT)`~AirQualityUCI$Date1)
> summary(mod)

Call:
lm(formula = AirQualityUCI$`CO(GT)` ~ AirQualityUCI$Date1)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1512 -1.0913 -0.3337  0.7422  9.7166

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -4.8415230  1.8033975  -2.685 0.007276 **
AirQualityUCI$Date1  0.0005512  0.0001421   3.879 0.000106 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.452 on 7672 degrees of freedom
  (1683 observations deleted due to missingness)
Multiple R-squared:  0.001957, Adjusted R-squared:  0.001827
F-statistic: 15.04 on 1 and 7672 DF,  p-value: 0.000106

>
> mod <- lm(AirQualityUCI$`CO(GT)`~AirQualityUCI$T)
> summary(mod)

Call:
lm(formula = AirQualityUCI$`CO(GT)` ~ AirQualityUCI$T)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1099 -1.0686 -0.3368  0.7071  9.7894

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.066033   0.037547  55.025   <2e-16 ***
AirQualityUCI$T 0.003584   0.001891   1.895   0.0581 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.436 on 7342 degrees of freedom
  (2013 observations deleted due to missingness)
Multiple R-squared:  0.000489, Adjusted R-squared:  0.0003528
F-statistic: 3.592 on 1 and 7342 DF,  p-value: 0.0581

>
> mod <- lm(AirQualityUCI$`CO(GT)`~AirQualityUCI$RH)
> summary(mod)

Call:
lm(formula = AirQualityUCI$`CO(GT)` ~ AirQualityUCI$RH)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1595 -1.0712 -0.3169  0.7328  9.6671
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.9322601  0.0499611  38.675  < 2e-16 ***
AirQualityUCI$RH 0.0040248  0.0009595   4.195 2.76e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.435 on 7342 degrees of freedom
  (2013 observations deleted due to missingness)
Multiple R-squared:  0.002391, Adjusted R-squared:  0.002255
F-statistic:  17.6 on 1 and 7342 DF,  p-value: 2.765e-05
```

g) Create cross tabulations with derived variables

The R-script for the given problem is as follows:

```
mydata<-AirQualityUCI
View(mydata) # 2-Way Frequency Table
attach(mydata)
#mytable <- table(A,B) # A will be rows, B will be columns
#mytable # print table
margin.table(mytable, 1) # A frequencies (summed over B)
prop.table(mytable) # cell percentages
prop.table(mytable, 1) # row percentages


range(AirQualityUCI$RH)

final <- within(AirQualityUCI,
            {
              RHcat <- NA
              RHcat[RH<20] <- "Very Low"
              RHcat[RH>=20 & RH<=40] <- "Low"
              RHcat[RH>40 & RH<=60] <- "Medium"
              RHcat[RH>60 & RH<=80] <- "High"
              RHcat[RH>80] <- "Very High"
            })

mytable <- xtabs(`CO(GT)` ~ +RHcat, data = final)
ftable(mytable) # print table
summary(mytable) # chi-square test of indepedence

mytable <- xtabs(`C6H6(GT)` ~ +RHcat, data = final)
ftable(mytable) # print table
summary(mytable) # chi-square test of indepedence

mytable <- xtabs(`NOx(GT)` ~ +RHcat, data = final)
```

ftable(mytable) # print table
summary(mytable) # chi-square test of indepedence

with(final, tapply(`NO2(GT)`, list(RHcat=RHcat), sd)) # using with()
with(final, tapply(`NO2(GT)`, list(RHcat=RHcat), mean))

The output of the R-Script (from Console window) is given as follows:

```
> mydata<-AirQualityUCI
> View(mydata) # 2-Way Frequency Table
> attach(mydata)
The following objects are masked from mydata (pos = 5):

    AH, C6H6(GT), CO(GT), Date, datetime, NO2(GT), NOx(GT), PT08.S1(CO),
    PT08.S2(NMHC), PT08.S3(NOx), PT08.S4(NO2), PT08.S5(O3), RH, T, Time, Time1

The following objects are masked from mydata (pos = 6):

    AH, C6H6(GT), CO(GT), Date, datetime, NO2(GT), NOx(GT), PT08.S1(CO),
    PT08.S2(NMHC), PT08.S3(NOx), PT08.S4(NO2), PT08.S5(O3), RH, T, Time, Time1

The following object is masked from package:base:

    T

> #mytable <- table(A,B) # A will be rows, B will be columns
> #mytable # print table
> margin.table(mytable, 1) # A frequencies (summed over B)
RHcat
     High       Low    Medium Very High  Very Low
 566943.9  417357.3  664434.1   77071.7   65314.5
> prop.table(mytable) # cell percentages
RHcat
      High        Low     Medium  Very High   Very Low
0.31653012 0.23301451 0.37095981 0.04302986 0.03646570
> prop.table(mytable, 1) # row percentages
RHcat
     High       Low    Medium Very High  Very Low
        1         1         1         1         1
>
>
> range(AirQualityUCI$RH)
[1] NA NA
>
> final <- within(AirQualityUCI,
+                 {
+                   RHcat <- NA
+                   RHcat[RH<20] <- "Very Low"
+                   RHcat[RH>=20 & RH<=40] <- "Low"
+                   RHcat[RH>40 & RH<=60] <- "Medium"
+                   RHcat[RH>60 & RH<=80] <- "High"
+                   RHcat[RH>80] <- "Very High"
+                 })
>
> mytable <- xtabs(`CO(GT)` ~ +RHcat, data = final)
> ftable(mytable)  # print table
mytable 497.1 662.5 4288.7 4302.4 5889.9

            1     1      1      1      1
> summary(mytable) # chi-square test of indepedence
Number of cases in table: 15640.6
Number of factors: 1
>
> mytable <- xtabs(`C6H6(GT)` ~  +RHcat, data = final)
> ftable(mytable)  # print table
mytable 2206.4370307221 4537.99826996217 23277.0380810769 25828.1012760302
```

```
34806.619496821

                     1                1                1                1
1
> summary(mytable) # chi-square test of indepedence
Number of cases in table: 90656.19
Number of factors: 1
>
> mytable <- xtabs(`NOx(GT)` ~  +RHcat, data = final)
> ftable(mytable)  # print table
mytable 65314.5 77071.7 417357.3 566943.9 664434.1

               1        1        1        1        1
> summary(mytable) # chi-square test of indepedence
Number of cases in table: 1791122
Number of factors: 1
>
> with(final, tapply(`NO2(GT)`, list(RHcat=RHcat), sd)) # using with()
RHcat
     High       Low    Medium Very High  Very Low
       NA        NA        NA        NA        NA
> with(final, tapply(`NO2(GT)`, list(RHcat=RHcat), mean))
RHcat
     High       Low    Medium Very High  Very Low
       NA        NA        NA        NA        NA
```

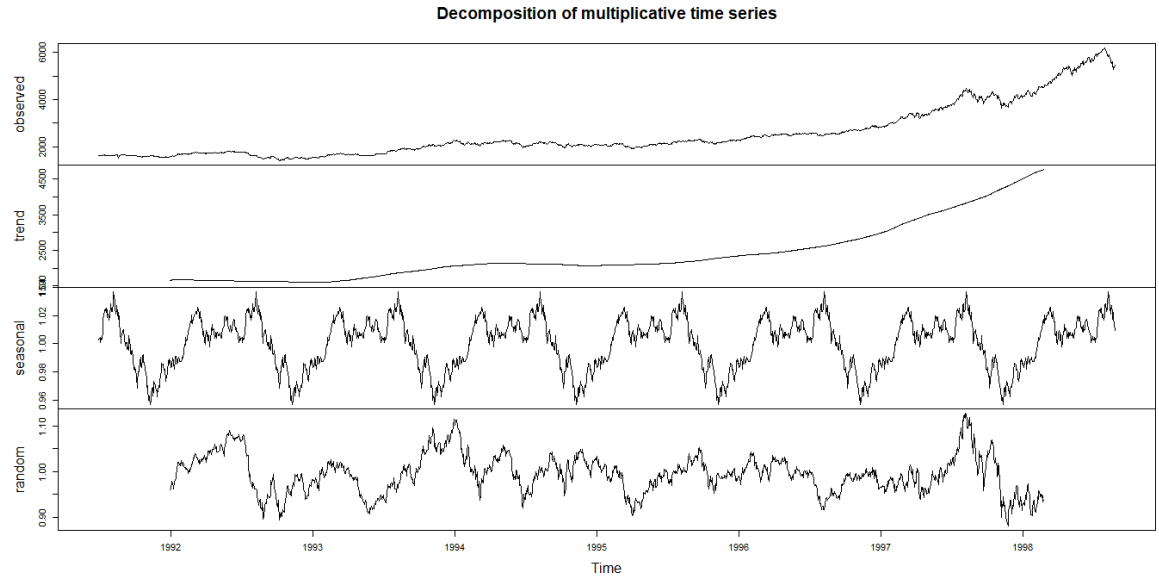
h) check for trends and patterns in time series

The R-script for the given problem is as follows:

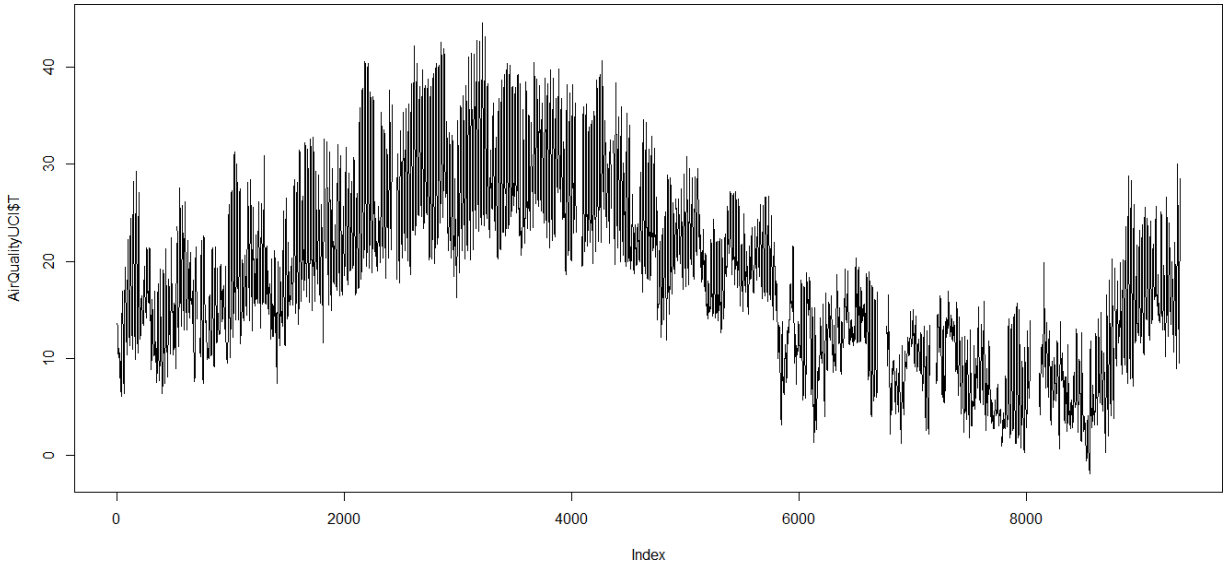
```
#plot time series
tsAirqualityUCI <- EuStockMarkets[, 1] # ts data
decomposedRes <- decompose(tsAirqualityUCI, type="mult") # use type = "additive" for
additive components
plot (decomposedRes) # see plot below
stlRes <- stl(tsAirqualityUCI, s.window = "periodic")
plot(AirQualityUCI$T, type = "l")

#or
library(xts)
timeseries <- xts(final$`CO(GT)`, final$datetime)
plot(timeseries)
summary(timeseries)
ts (AirQualityUCI, frequency = 4, start = c(1959, 2))# frequency 4 =>Quarterly Data
ts (1:10, frequency = 12, start = 1990) # freq 12 => Monthly data.
ts (AirQualityUCI, start=c(2009), end=c(2014), frequency=1) # Yearly Data
ts (1:1000, frequency = 365, start = 1990) # freq 365 => daily data.
```

The output of the R-Script (from Console window) is given as follows:

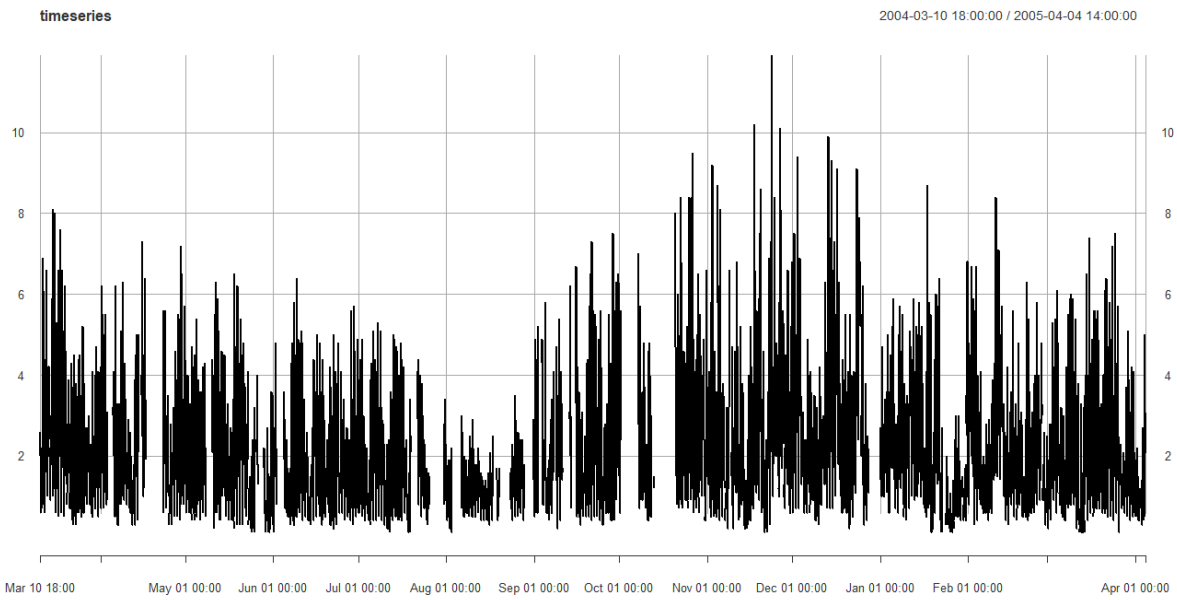
```
> #plot time series
> tsAirqualityUCI <- EuStockMarkets[, 1] # ts data
> decomposedRes <- decompose(tsAirqualityUCI, type="mult") # use type =
"additive" for additive components
> plot (decomposedRes) # see plot below
```




```
> stlRes <- stl(tsAirqualityUCI, s.window = "periodic")
> plot(AirQualityUCI$T, type = "l")
```




```
> library(xts)
>
> timeseries <- xts(final$`CO(GT)`, final$datetime)
> plot(timeseries)
> summary(timeseries)
     Index                       timeseries
 Min.   :2004-03-10 18:00:00   Min.   : 0.100
 1st Qu.:2004-06-16 05:00:00   1st Qu.: 1.100
 Median :2004-09-21 16:00:00   Median : 1.800
 Mean   :2004-09-21 16:00:00   Mean   : 2.153
 3rd Qu.:2004-12-28 03:00:00   3rd Qu.: 2.900
 Max.   :2005-04-04 14:00:00   Max.   :11.900
                               NA's   :1683
```




```
> ts (AirQualityUCI, frequency = 4, start = c(1959, 2))# frequency 4
=>Quarterly Data
             Date        Time CO(GT) PT08.S1(CO)   C6H6(GT) PT08.S2(NMHC)
NOx(GT)
1959 Q2 1078876800 -2209010400     2.6   1360.0000 11.8817235      1045.5000
166.0
1959 Q3 1078876800 -2209006800     2.0    1292.2500  9.3971649       954.7500
103.0
1959 Q4 1078876800 -2209003200     2.2    1402.0000  8.9978169       939.2500
131.0
1960 Q1 1078876800 -2208999600     2.2    1375.5000  9.2287964       948.2500
172.0
1960 Q2 1078876800 -2208996000     1.6    1272.2500  6.5182237       835.5000
131.0
1960 Q3 1078876800 -2208992400     1.2    1197.0000  4.7410124       750.2500
89.0
1960 Q4 1078963200 -2209075200     1.2    1185.0000  3.6243992       689.5000
62.0
1961 Q1 1078963200 -2209071600     1.0    1136.2500  3.3266770       672.0000
62.0
1961 Q2 1078963200 -2209068000     0.9    1094.0000  2.3394162       608.5000
45.0
1961 Q3 1078963200 -2209064400     0.6    1009.7500  1.6966583       560.7500
NA
1961 Q4 1078963200 -2209060800      NA    1011.0000  1.2936198       526.7500
21.0
1962 Q1 1078963200 -2209057200     0.7    1066.0000  1.1334306       512.0000
16.0
1962 Q2 1078963200 -2209053600     0.7    1051.7500  1.6037679       553.2500
34.0
1962 Q3 1078963200 -2209050000     1.1    1144.0000  3.2436181       667.0000
98.0
1962 Q4 1078963200 -2209046400     2.0    1333.2500  8.0137730       899.7500
174.0
```

```
1963 Q1 1078963200 -2209042800    2.2    1351.0000   9.5406429       960.2500
129.0
1963 Q2 1078963200 -2209039200    1.7    1233.2500   6.3357824       827.2500
112.0
1963 Q3 1078963200 -2209035600    1.5    1178.7500   4.9715838       762.0000
95.0
1963 Q4 1078963200 -2209032000    1.6    1236.0000   5.2169190       774.2500
104.0
1964 Q1 1078963200 -2209028400    1.9    1285.5000   7.2699334       868.5000
146.0
1964 Q2 1078963200 -2209024800    2.9    1371.0000  11.5390072      1033.5000
207.0
1964 Q3 1078963200 -2209021200    2.2    1310.0000   8.8262227       932.5000
184.0
1964 Q4 1078963200 -2209017600    2.2    1291.7500   8.3014134       911.5000
193.0
1965 Q1 1078963200 -2209014000    2.9    1383.0000  11.1515812      1019.7500
243.0
1965 Q2 1078963200 -2209010400    4.8    1580.7500  20.7992169      1318.5000
281.0
1965 Q3 1078963200 -2209006800    6.9    1775.5000  27.3598075      1487.7500
383.0
1965 Q4 1078963200 -2209003200    6.1    1640.0000  24.0177569      1404.0000
351.0
1966 Q1 1078963200 -2208999600    3.9    1312.7500  12.7793682      1076.2500
240.0
1966 Q2 1078963200 -2208996000    1.5     964.5000   4.7070719       748.5000
94.0
1966 Q3 1078963200 -2208992400    1.0     912.7500   2.6457215       629.2500
47.0
1966 Q4 1079049600 -2209075200    1.7    1080.2500   5.8548015       805.0000
122.0
1967 Q1 1079049600 -2209071600    1.9    1043.7500   6.3742975       829.0000
133.0
1967 Q2 1079049600 -2209068000    1.4     987.7500   4.1323418       718.0000
82.0
1967 Q3 1079049600 -2209064400    0.8     888.7500   1.8694446       574.2500
NA
1967 Q4 1079049600 -2209060800     NA     831.0000   1.0682926       505.7500
21.0
1968 Q1 1079049600 -2209057200    0.6     847.2500   1.0224146       501.2500
30.0
1968 Q2 1079049600 -2209053600    0.8     927.0000   1.8304312       571.2500
56.0
1968 Q3 1079049600 -2209050000    1.4    1090.5000   4.3593410       730.2500
109.0
1968 Q4 1079049600 -2209046400    4.4    1587.0000  17.8655867      1235.5000
307.0
1969 Q1 1079049600 -2209042800     NA    1544.5000  22.0741621      1353.0000
NA
1969 Q2 1079049600 -2209039200    3.1    1350.2500  14.0270114      1117.5000
187.0
1969 Q3 1079049600 -2209035600    2.7    1262.7500  11.6456466      1037.2500
216.0
1969 Q4 1079049600 -2209032000    2.1    1206.2500  10.2246621       986.0000
143.0
```

```
1970 Q1 1079049600 -2209028400    2.5    1251.5000  11.0399360     1015.7500
160.0
1970 Q2 1079049600 -2209024800    2.7    1287.0000  12.8164462     1077.5000
163.0
1970 Q3 1079049600 -2209021200    2.9    1352.7500  14.1738512     1122.2500
190.0
1970 Q4 1079049600 -2209017600    2.8    1309.0000  12.6905681     1073.2500
178.0
1971 Q1 1079049600 -2209014000    2.4    1274.0000  11.7384054     1040.5000
150.0
1971 Q2 1079049600 -2209010400    3.9    1509.5000  19.2909749     1276.5000
206.0
1971 Q3 1079049600 -2209006800    3.7    1525.2500  18.2261783     1246.0000
202.0
1971 Q4 1079049600 -2209003200    6.6    1843.0000  32.5562783     1609.7500
340.0
1972 Q1 1079049600 -2208999600    4.4    1597.7500  20.0929436     1299.0000
274.0
1972 Q2 1079049600 -2208996000    3.5    1483.5000  14.3213424     1127.0000
253.0
1972 Q3 1079049600 -2208992400    5.4    1677.2500  21.8128651     1346.0000
300.0
1972 Q4 1079136000 -2209075200    2.7    1279.5000   9.6389998      964.0000
193.0
1973 Q1 1079136000 -2209071600    1.9    1196.2500   7.3751395      873.0000
139.0
1973 Q2 1079136000 -2209068000    1.6    1183.7500   5.3696042      781.7500
83.0
1973 Q3 1079136000 -2209064400    1.7    1171.7500   5.3901039      782.7500
NA
        PT08.S3(NOx) NO2(GT) PT08.S4(NO2) PT08.S5(O3)         T        RH
AH Date1 Time1
1959 Q2    1056.2500   113.0    1692.0000   1267.5000 13.600000 48.87500
0.7577538 12487    NA
1959 Q3    1173.7500    92.0    1558.7500    972.2500 13.300000 47.70000
0.7254874 12487    NA
1959 Q4    1140.0000   114.0    1554.5000   1074.0000 11.900000 53.97500
0.7502391 12487    NA
1960 Q1    1092.0000   122.0    1583.7500   1203.2500 11.000000 60.00000
0.7867125 12487    NA
1960 Q2    1205.0000   116.0    1490.0000   1110.0000 11.150000 59.57500
0.7887942 12487    NA
1960 Q3    1336.5000    96.0    1393.0000    949.2500 11.175000 59.17500
0.7847717 12487    NA
1960 Q4    1461.7500    77.0    1332.7500    732.5000 11.325000 56.77500
0.7603119 12488    NA
1961 Q1    1453.2500    76.0    1332.7500    729.5000 10.675000 60.00000
0.7702385 12488    NA
1961 Q2    1579.0000    60.0    1276.0000    619.5000 10.650000 59.67500
0.7648187 12488    NA
1961 Q3    1705.0000      NA    1234.7500    501.2500 10.250000 60.20000
0.7516572 12488    NA
1961 Q4    1817.5000    34.0    1196.7500    445.2500 10.075000 60.47500
0.7464945 12488    NA
1962 Q1    1918.0000    28.0    1182.0000    421.7500 11.000000 56.17500
0.7365596 12488    NA
```

```
1962 Q2    1738.2500    48.0    1221.2500    471.5000 10.450000 58.12500
0.7352951 12488    NA
1962 Q3    1489.7500    82.0    1339.0000    729.7500 10.200000 59.60000
0.7417362 12488    NA
1962 Q4    1136.0000   112.0    1517.0000   1101.5000 10.750000 57.42500
0.7407946 12488    NA
1963 Q1    1079.0000   101.0    1582.7500   1027.7500 10.500000 60.60000
0.7691108 12488    NA
1963 Q2    1218.0000    98.0    1445.7500    859.7500 10.800000 58.35000
0.7551831 12488    NA
1963 Q3    1327.5000    92.0    1361.7500    670.5000 10.500000 57.92500
0.7351608 12488    NA
1963 Q4    1301.2500    95.0    1401.2500    664.0000  9.525000 66.77500
0.7950538 12488    NA
1964 Q1    1162.2500   112.0    1536.7500    799.0000  8.300000 76.42500
0.8392681 12488    NA
1964 Q2     983.2500   128.0    1730.2500   1036.5000  8.000000 81.15000
0.8735885 12488    NA
1964 Q3    1081.7500   126.0    1646.5000    946.2500  8.325000 79.80000
0.8777844 12488    NA
1964 Q4    1102.5000   131.0    1590.7500    956.7500  9.700000 71.15000
0.8569381 12488    NA
1965 Q1    1008.0000   135.0    1718.7500   1104.0000  9.775000 67.62500
0.8185012 12488    NA
1965 Q2     798.5000   151.0    2083.0000   1408.5000 10.350000 64.17500
0.8065436 12488    NA
1965 Q3     702.2500   172.0    2332.5000   1704.0000  9.650000 69.30000
0.8319211 12488    NA
1965 Q4     742.7500   165.0    2191.2500   1653.7500  9.650000 67.75000
0.8133139 12488    NA
1966 Q1     957.2500   136.0    1706.5000   1284.7500  9.125000 63.97500
0.7419242 12488    NA
1966 Q2    1325.2500    85.0    1332.5000    821.0000  8.175000 63.40000
0.6904844 12488    NA
1966 Q3    1564.5000    53.0    1252.2500    551.7500  8.250000 60.82500
0.6657444 12488    NA
1966 Q4    1253.5000    97.0    1375.0000    815.5000  8.325000 58.52500
0.6437636 12489    NA
1967 Q1    1247.2500   110.0    1378.2500    831.5000  7.725000 59.67500
0.6307661 12489    NA
1967 Q2    1395.5000    91.0    1303.5000    691.5000  7.125000 61.80000
0.6275974 12489    NA
1967 Q3    1680.2500      NA    1187.0000    512.0000  6.975000 62.27500
0.6261075 12489    NA
1967 Q4    1892.7500    32.0    1133.7500    384.0000  6.100000 65.90000
0.6247536 12489    NA
1968 Q1    1894.5000    44.0    1154.7500    394.0000  6.275000 64.97500
0.6232823 12489    NA
1968 Q2    1684.7500    71.0    1222.7500    486.5000  6.750000 62.95000
0.6234275 12489    NA
1968 Q3    1387.0000   104.0    1360.7500    748.2500  6.450000 65.07500
0.6316281 12489    NA
1968 Q4     896.5000   141.0    1900.2500   1400.2500  7.325000 63.15000
0.6499331 12489    NA
1969 Q1     767.2500      NA    2058.0000   1587.7500  9.225000 56.20000
0.6560651 12489    NA
```

```
1969 Q2     912.0000   122.0     1711.7500   1237.0000  13.225000 41.75000
0.6319501 12489    NA
1969 Q3     969.0000   143.0     1598.2500   1166.5000  14.325000 38.45000
0.6243043 12489    NA
1969 Q4    1034.5000   113.0     1537.0000    959.0000  15.025000 36.50000
0.6195323 12489    NA
1970 Q1    1007.5000   116.0     1592.7500    983.0000  16.100000 34.47500
0.6261647 12489    NA
1970 Q2     948.7500   123.0     1660.2500   1060.7500  16.275001 35.72500
0.6560306 12489    NA
1970 Q3     921.7500   126.0     1740.0000   1139.2500  15.825000 37.02500
0.6609611 12489    NA
1970 Q4     954.0000   120.0     1657.2500   1112.2500  15.875000 37.17500
0.6657285 12489    NA
1971 Q1    1005.7500   119.0     1609.7500    993.7500  16.875000 34.35000
0.6549085 12489    NA
1971 Q2     812.2500   149.0     1909.7500   1409.5000  15.150000 39.55000
0.6766265 12489    NA
1971 Q3     821.0000   145.0     1846.7500   1447.7500  14.400000 43.42500
0.7084498 12489    NA
1971 Q4     624.0000   170.0     2390.2500   1886.5000  12.875000 50.52500
0.7478032 12489    NA
1972 Q1     752.0000   149.0     1940.5000   1626.7500  12.150000 53.35000
0.7536202 12489    NA
1972 Q2     839.0000   139.0     1723.0000   1491.0000  10.975000 59.12500
0.7739800 12489    NA
1972 Q3     740.5000   134.0     2062.0000   1657.0000   9.675000 64.62500
0.7770739 12489    NA
1972 Q4     962.5000   113.0     1543.5000   1285.2500   9.450000 64.12500
0.7597465 12490    NA
1973 Q1    1071.2500    97.0     1463.2500   1144.2500   9.150000 63.90000
0.7422764 12490    NA
1973 Q2    1176.2500    82.0     1364.5000   1042.7500   8.800000 63.92500
0.7256154 12490    NA
1973 Q3    1178.5000      NA     1379.7500    995.5000   7.800000 67.52500
0.7173121 12490    NA
          datetime
1959 Q2 1078921800
1959 Q3 1078925400
1959 Q4 1078929000
1960 Q1 1078932600
1960 Q2 1078936200
1960 Q3 1078939800
1960 Q4 1078943400
1961 Q1 1078947000
1961 Q2 1078950600
1961 Q3 1078954200
1961 Q4 1078957800
1962 Q1 1078961400
1962 Q2 1078965000
1962 Q3 1078968600
1962 Q4 1078972200
1963 Q1 1078975800
1963 Q2 1078979400
1963 Q3 1078983000
1963 Q4 1078986600
1964 Q1 1078990200
```

```
1964 Q2 1078993800
1964 Q3 1078997400
1964 Q4 1079001000
1965 Q1 1079004600
1965 Q2 1079008200
1965 Q3 1079011800
1965 Q4 1079015400
1966 Q1 1079019000
1966 Q2 1079022600
1966 Q3 1079026200
1966 Q4 1079029800
1967 Q1 1079033400
1967 Q2 1079037000
1967 Q3 1079040600
1967 Q4 1079044200
1968 Q1 1079047800
1968 Q2 1079051400
1968 Q3 1079055000
1968 Q4 1079058600
1969 Q1 1079062200
1969 Q2 1079065800
1969 Q3 1079069400
1969 Q4 1079073000
1970 Q1 1079076600
1970 Q2 1079080200
1970 Q3 1079083800
1970 Q4 1079087400
1971 Q1 1079091000
1971 Q2 1079094600
1971 Q3 1079098200
1971 Q4 1079101800
1972 Q1 1079105400
1972 Q2 1079109000
1972 Q3 1079112600
1972 Q4 1079116200
1973 Q1 1079119800
1973 Q2 1079123400
1973 Q3 1079127000
 [ reached getOption("max.print") -- omitted 9299 rows ]
> ts (1:10, frequency = 12, start = 1990) # freq 12 => Monthly data.
     Jan Feb Mar Apr May Jun Jul Aug Sep Oct
1990   1   2   3   4   5   6   7   8   9  10
> ts (AirQualityUCI, start=c(2009), end=c(2014), frequency=1) # Yearly Data
Time Series:
Start = 2009
End = 2014
Frequency = 1
           Date        Time CO(GT) PT08.S1(CO)  C6H6(GT) PT08.S2(NMHC)
NOx(GT) PT08.S3(NOx)
2009 1078876800 -2209010400    2.6     1360.00 11.881723        1045.50
166      1056.25
2010 1078876800 -2209006800    2.0     1292.25  9.397165         954.75
103      1173.75
2011 1078876800 -2209003200    2.2     1402.00  8.997817         939.25
131      1140.00
2012 1078876800 -2208999600    2.2     1375.50  9.228796         948.25
172      1092.00
```

```
2013 1078876800 -2208996000    1.6     1272.25  6.518224        835.50
131      1205.00
2014 1078876800 -2208992400    1.2     1197.00  4.741012        750.25
89      1336.50
     NO2(GT) PT08.S4(NO2) PT08.S5(O3)      T     RH        AH Date1 Time1
datetime
2009     113      1692.00     1267.50 13.600 48.875 0.7577538 12487    NA
1078921800
2010      92      1558.75      972.25 13.300 47.700 0.7254874 12487    NA
1078925400
2011     114      1554.50     1074.00 11.900 53.975 0.7502391 12487    NA
1078929000
2012     122      1583.75     1203.25 11.000 60.000 0.7867125 12487    NA
1078932600
2013     116      1490.00     1110.00 11.150 59.575 0.7887942 12487    NA
1078936200
2014      96      1393.00      949.25 11.175 59.175 0.7847717 12487    NA
1078939800
> ts (1:1000, frequency = 365, start = 1990) # freq 365 => daily data.
Time Series:
Start = c(1990, 1)
End = c(1992, 270)
Frequency = 365
   [1]    1    2    3    4    5    6    7    8    9   10   11   12   13   14
15   16   17
  [18]   18   19   20   21   22   23   24   25   26   27   28   29   30   31
32   33   34
  [35]   35   36   37   38   39   40   41   42   43   44   45   46   47   48
49   50   51
  [52]   52   53   54   55   56   57   58   59   60   61   62   63   64   65
66   67   68
  [69]   69   70   71   72   73   74   75   76   77   78   79   80   81   82
83   84   85
  [86]   86   87   88   89   90   91   92   93   94   95   96   97   98   99
100  101  102
 [103]  103  104  105  106  107  108  109  110  111  112  113  114  115  116
117  118  119
 [120]  120  121  122  123  124  125  126  127  128  129  130  131  132  133
134  135  136
 [137]  137  138  139  140  141  142  143  144  145  146  147  148  149  150
151  152  153
 [154]  154  155  156  157  158  159  160  161  162  163  164  165  166  167
168  169  170
 [171]  171  172  173  174  175  176  177  178  179  180  181  182  183  184
185  186  187
 [188]  188  189  190  191  192  193  194  195  196  197  198  199  200  201
202  203  204
 [205]  205  206  207  208  209  210  211  212  213  214  215  216  217  218
219  220  221
 [222]  222  223  224  225  226  227  228  229  230  231  232  233  234  235
236  237  238
 [239]  239  240  241  242  243  244  245  246  247  248  249  250  251  252
253  254  255
 [256]  256  257  258  259  260  261  262  263  264  265  266  267  268  269
270  271  272
 [273]  273  274  275  276  277  278  279  280  281  282  283  284  285  286
287  288  289
```

```
 [290]  290  291  292  293  294  295  296  297  298  299  300  301  302  303
304  305  306
 [307]  307  308  309  310  311  312  313  314  315  316  317  318  319  320
321  322  323
 [324]  324  325  326  327  328  329  330  331  332  333  334  335  336  337
338  339  340
 [341]  341  342  343  344  345  346  347  348  349  350  351  352  353  354
355  356  357
 [358]  358  359  360  361  362  363  364  365  366  367  368  369  370  371
372  373  374
 [375]  375  376  377  378  379  380  381  382  383  384  385  386  387  388
389  390  391
 [392]  392  393  394  395  396  397  398  399  400  401  402  403  404  405
406  407  408
 [409]  409  410  411  412  413  414  415  416  417  418  419  420  421  422
423  424  425
 [426]  426  427  428  429  430  431  432  433  434  435  436  437  438  439
440  441  442
 [443]  443  444  445  446  447  448  449  450  451  452  453  454  455  456
457  458  459
 [460]  460  461  462  463  464  465  466  467  468  469  470  471  472  473
474  475  476
 [477]  477  478  479  480  481  482  483  484  485  486  487  488  489  490
491  492  493
 [494]  494  495  496  497  498  499  500  501  502  503  504  505  506  507
508  509  510
 [511]  511  512  513  514  515  516  517  518  519  520  521  522  523  524
525  526  527
 [528]  528  529  530  531  532  533  534  535  536  537  538  539  540  541
542  543  544
 [545]  545  546  547  548  549  550  551  552  553  554  555  556  557  558
559  560  561
 [562]  562  563  564  565  566  567  568  569  570  571  572  573  574  575
576  577  578
 [579]  579  580  581  582  583  584  585  586  587  588  589  590  591  592
593  594  595
 [596]  596  597  598  599  600  601  602  603  604  605  606  607  608  609
610  611  612
 [613]  613  614  615  616  617  618  619  620  621  622  623  624  625  626
627  628  629
 [630]  630  631  632  633  634  635  636  637  638  639  640  641  642  643
644  645  646
 [647]  647  648  649  650  651  652  653  654  655  656  657  658  659  660
661  662  663
 [664]  664  665  666  667  668  669  670  671  672  673  674  675  676  677
678  679  680
 [681]  681  682  683  684  685  686  687  688  689  690  691  692  693  694
695  696  697
 [698]  698  699  700  701  702  703  704  705  706  707  708  709  710  711
712  713  714
 [715]  715  716  717  718  719  720  721  722  723  724  725  726  727  728
729  730  731
 [732]  732  733  734  735  736  737  738  739  740  741  742  743  744  745
746  747  748
 [749]  749  750  751  752  753  754  755  756  757  758  759  760  761  762
763  764  765
```

```
 [766]  766  767  768  769  770  771  772  773  774  775  776  777  778  779
780  781  782
 [783]  783  784  785  786  787  788  789  790  791  792  793  794  795  796
797  798  799
 [800]  800  801  802  803  804  805  806  807  808  809  810  811  812  813
814  815  816
 [817]  817  818  819  820  821  822  823  824  825  826  827  828  829  830
831  832  833
 [834]  834  835  836  837  838  839  840  841  842  843  844  845  846  847
848  849  850
 [851]  851  852  853  854  855  856  857  858  859  860  861  862  863  864
865  866  867
 [868]  868  869  870  871  872  873  874  875  876  877  878  879  880  881
882  883  884
 [885]  885  886  887  888  889  890  891  892  893  894  895  896  897  898
899  900  901
 [902]  902  903  904  905  906  907  908  909  910  911  912  913  914  915
916  917  918
 [919]  919  920  921  922  923  924  925  926  927  928  929  930  931  932
933  934  935
 [936]  936  937  938  939  940  941  942  943  944  945  946  947  948  949
950  951  952
 [953]  953  954  955  956  957  958  959  960  961  962  963  964  965  966
967  968  969
 [970]  970  971  972  973  974  975  976  977  978  979  980  981  982  983
984  985  986
 [987]  987  988  989  990  991  992  993  994  995  996  997  998  999 1000
```

Conclusion/Interpretation:

Trends and patterns in time series are hence checked.

i) Find out the most polluted time of the day and the name of the chemical compound

The R-script for the given problem is as follows:

```
names(AirQualityUCI)
library(dplyr)

polluted <- AirQualityUCI%>%group_by(Time)%>%
  select(Time, `CO(GT)`, `C6H6(GT)`, `NO2(GT)`, `NOx(GT)` )%>%
summarise(CO = mean(`CO(GT)`), C6H6 = mean(`C6H6(GT)`), NO2 =
mean(`NO2(GT)`), NOX =mean(`NOx(GT)`))%>%


polluted[c(which.max(polluted$CO),which.max(polluted$C6H6),which.max(poll
uted$NO2),which.max(polluted$NOX)),]
```

The output of the R-Script (from Console window) is given as follows:

```
> names(AirQualityUCI)
 [1] "Date"           "Time"           "CO(GT)"         "PT08.S1(CO)"
"C6H6(GT)"
 [6] "PT08.S2(NMHC)" "NOx(GT)"        "PT08.S3(NOx)"   "NO2(GT)"
"PT08.S4(NO2)"
[11] "PT08.S5(O3)"    "T"              "RH"             "AH"             "Date1"
[16] "Time1"          "datetime"
> library(dplyr)
>
> polluted <- AirQualityUCI%>%group_by(Time)%>%
+   select(Time, `CO(GT)`, `C6H6(GT)`, `NO2(GT)`, `NOx(GT)` )%>%
+   summarise(CO = mean(`CO(GT)`), C6H6 = mean(`C6H6(GT)`), NO2 =
mean(`NO2(GT)`), NOX =mean(`NOx(GT)`))%>%
+
+
polluted[c(which.max(polluted$CO),which.max(polluted$C6H6),which.max(polluted
$NO2),which.max(polluted$NOX)),]
```

Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)
6/8/2004	8:00:00	5.8	1377	-200	36.1	1688
6/9/2004	8:00:00	6.4	1496	-200	36.9	1705
10/26/2004	18:00:00	9.5	1908	-200	52.1	2007
max		11.9	2039.8	1189.0	63.7	2214.0

Date	Time	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)
6/8/2004	8:00:00	376	525	125	2746	1708
6/9/2004	8:00:00	357	507	151	2691	2147
10/26/2004	18:00:00	952	325	180	2775	2372
max		1479.0	2682.8	339.7	2775.0	2522.8

Conclusion/Interpretation:

PT08.S4(NO2) is the highest pollution at 18.00 hrs