**Assignment No. 10.1**

**Data Set Information:**

The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level, within an Italian city. Data were recorded from March 2004 to February 2005 (one year) representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Methane Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer. Evidences of cross-sensitivities as well as both concept and sensor drifts are present as described in De Vito et al., Sens. And Act. B, Vol. 129,2,2008 (citation required) eventually affecting sensors concentration estimation capabilities. Missing values are tagged with -200 value. This dataset can be used exclusively for research purposes. Commercial purposes are fully excluded.

**Attribute Information:**

 0 Date (DD/MM/YYYY)

 1 Time (HH.MM.SS)

 2 True hourly averaged concentration CO in mg/m^3 (reference analyzer)

 3 PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)

4 True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)

 5 True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)

 6 PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)

7 True hourly averaged NOx concentration in ppb (reference analyzer)

 8 PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)

 9 True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)

10 PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)

11 PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)

 12 Temperature in Â°C

13 Relative Humidity (%)

 14 AH Absolute Humidity

The following questions are expected to answered in this assignment

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

2. Create Univariate for all the columns.

3. Check for missing values in all columns.

4. Impute the missing values using appropriate methods.

5. Create bi-variate analysis for all relationships.

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.

**READING THE ZIP FILE INTO R**

|  |
| --- |
| > zipF<- "E:\\kamagyana\\Computing\\DARET\\Assignments\\AirQualityUCI.zip"  > outDir<-"E:\\kamagyana\\Computing\\DARET\\Assignments"  > unzip(zipF,exdir=outDir) |
|  |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | > library(readxl)  > AirQualityUCI <- read\_excel("E:/kamagyana/Computing/DARET/Assignments/AirQualityUCI.xlsx")   |  | | --- | | > head(AirQualityUCI)  # A tibble: 6 x 15  Date Time `CO(GT)` `PT08.S1(CO)`  *<dttm>* *<dttm>* *<dbl>* *<dbl>*  1 2004-03-10 00:00:00 1899-12-31 18:00:00 2.6 1360  2 2004-03-10 00:00:00 1899-12-31 19:00:00 2 1292.  3 2004-03-10 00:00:00 1899-12-31 20:00:00 2.2 1402  4 2004-03-10 00:00:00 1899-12-31 21:00:00 2.2 1376.  5 2004-03-10 00:00:00 1899-12-31 22:00:00 1.6 1272.  6 2004-03-10 00:00:00 1899-12-31 23:00:00 1.2 1197  # ... with 11 more variables: `NMHC(GT)` *<dbl>*, `C6H6(GT)` *<dbl>*,  # `PT08.S2(NMHC)` *<dbl>*, `NOx(GT)` *<dbl>*, `PT08.S3(NOx)` *<dbl>*,  # `NO2(GT)` *<dbl>*, `PT08.S4(NO2)` *<dbl>*, `PT08.S5(O3)` *<dbl>*,  # T *<dbl>*, RH *<dbl>*, AH *<dbl>*  > str(AirQualityUCI)  Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 9357 obs. of 15 variables:  $ Date : POSIXct, format: "2004-03-10" ...  $ Time : POSIXct, format: "1899-12-31 18:00:00" ...  $ 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 ...  **CREATE UNIVARIATE FOR ALL COLUMNS** | |  | | |  | | --- | | > summary(AirQualityUCI)  Date Time  Min. :2004-03-10 00:00:00 Min. :1899-12-31 00:00:00  1st Qu.:2004-06-16 00:00:00 1st Qu.:1899-12-31 05:00:00  Median :2004-09-21 00:00:00 Median :1899-12-31 11:00:00  Mean :2004-09-21 04:30:05 Mean :1899-12-31 11:29:55  3rd Qu.:2004-12-28 00:00:00 3rd Qu.:1899-12-31 18:00:00  Max. :2005-04-04 00:00:00 Max. :1899-12-31 23:00:00  CO(GT) PT08.S1(CO) NMHC(GT) C6H6(GT)  Min. :-200.00 Min. :-200 Min. :-200.0 Min. :-200.000  1st Qu.: 0.60 1st Qu.: 921 1st Qu.:-200.0 1st Qu.: 4.005  Median : 1.50 Median :1052 Median :-200.0 Median : 7.887  Mean : -34.21 Mean :1049 Mean :-159.1 Mean : 1.866  3rd Qu.: 2.60 3rd Qu.:1221 3rd Qu.:-200.0 3rd Qu.: 13.636  Max. : 11.90 Max. :2040 Max. :1189.0 Max. : 63.741  PT08.S2(NMHC) NOx(GT) PT08.S3(NOx) NO2(GT)  Min. :-200.0 Min. :-200.0 Min. :-200.0 Min. :-200.00  1st Qu.: 711.0 1st Qu.: 50.0 1st Qu.: 637.0 1st Qu.: 53.00  Median : 894.5 Median : 141.0 Median : 794.2 Median : 96.00  Mean : 894.5 Mean : 168.6 Mean : 794.9 Mean : 58.14  3rd Qu.:1104.8 3rd Qu.: 284.2 3rd Qu.: 960.2 3rd Qu.: 133.00  Max. :2214.0 Max. :1479.0 Max. :2682.8 Max. : 339.70  PT08.S4(NO2) PT08.S5(O3) T RH  Min. :-200 Min. :-200.0 Min. :-200.000 Min. :-200.00  1st Qu.:1185 1st Qu.: 699.8 1st Qu.: 10.950 1st Qu.: 34.05  Median :1446 Median : 942.0 Median : 17.200 Median : 48.55  Mean :1391 Mean : 975.0 Mean : 9.777 Mean : 39.48  3rd Qu.:1662 3rd Qu.:1255.2 3rd Qu.: 24.075 3rd Qu.: 61.88  Max. :2775 Max. :2522.8 Max. : 44.600 Max. : 88.72  AH  Min. :-200.0000  1st Qu.: 0.6923  Median : 0.9768  Mean : -6.8376  3rd Qu.: 1.2962  Max. : 2.2310 | |  | |   The above summary shows that there is a particular minimum which is featuring in all the variables. It is not natural and the nature of variables suggests that such a reading is not valid for the respective variable. Further the dataset information also suggests that they have assigned a number “-200” for the missing data.  **CHECKING FOR MISSING VALUES IN THE DATA FOR ALL COLUMNS** | |

> nrow(AirQualityUCI) # shows the number of rows in the data frame

[1] 9357

> ncol(AirQualityUCI) # shows the number of columns in the data frame

[1] 15

> nrow(AirQualityUCI) \* ncol(AirQualityUCI)

[1] 140355

> sum(is.na(AirQualityUCI)) # shows that there are no missing values

[1] 0

> sum(!is.na(AirQualityUCI)) # shows that all the cells in a row column matrix are filled with some number

[1] 140355

This shows clearly that the existing dataset which is given in AirQualityUCI, does not show any missing values. This is because the missing values in thdataset are recorded with “-200”. Hence to convert it back to the original dataset, we have to globally replace all the “-200” values with NA and then proceed with the analysis.

> library(naniar) # to load a package which has global replace commands

> AQUCI <- AirQualityUCI %>% replace\_with\_na\_all(~.x == -200)

> sum(is.na(AQUCI))

[1] 16701

> sum(!is.na(AQUCI))

[1] 123654

> sum(is.na(AQUCI)) + sum(!is.na(AQUCI))

[1] 140355

Now that the missing values are recoded to NA. The summaries of all various columns with numerical values would change and hence there is a need to recalculate the summaries and undertake univariate analysis.

> summary(AQUCI)

Date Time CO(GT)

Min. :2004-03-10 05:30:00 Min. :1899-12-31 05:21:10 Min. : 0.100

1st Qu.:2004-06-16 05:30:00 1st Qu.:1899-12-31 10:21:10 1st Qu.: 1.100

Median :2004-09-21 05:30:00 Median :1899-12-31 16:21:10 Median : 1.800

Mean :2004-09-21 10:00:05 Mean :1899-12-31 16:51:05 Mean : 2.153

3rd Qu.:2004-12-28 05:30:00 3rd Qu.:1899-12-31 23:21:10 3rd Qu.: 2.900

Max. :2005-04-04 05:30:00 Max. :1900-01-01 04:21:10 Max. :11.900

NA's :1683

PT08.S1(CO) NMHC(GT) C6H6(GT) PT08.S2(NMHC)

Min. : 647.2 Min. : 7.0 Min. : 0.149 Min. : 383.2

1st Qu.: 936.8 1st Qu.: 67.0 1st Qu.: 4.437 1st Qu.: 734.4

Median :1063.0 Median : 150.0 Median : 8.240 Median : 909.0

Mean :1099.7 Mean : 218.8 Mean :10.083 Mean : 939.0

3rd Qu.:1231.2 3rd Qu.: 297.0 3rd Qu.:13.989 3rd Qu.:1116.2

Max. :2039.8 Max. :1189.0 Max. :63.742 Max. :2214.0

NA's :366 NA's :8443 NA's :366 NA's :366

NOx(GT) PT08.S3(NOx) NO2(GT) PT08.S4(NO2)

Min. : 2.0 Min. : 322.0 Min. : 2.0 Min. : 551

1st Qu.: 98.0 1st Qu.: 657.9 1st Qu.: 78.0 1st Qu.:1227

Median : 179.8 Median : 805.5 Median :109.0 Median :1463

Mean : 246.9 Mean : 835.4 Mean :113.1 Mean :1456

3rd Qu.: 326.0 3rd Qu.: 969.2 3rd Qu.:142.0 3rd Qu.:1674

Max. :1479.0 Max. :2682.8 Max. :339.7 Max. :2775

NA's :1639 NA's :366 NA's :1642 NA's :366

PT08.S5(O3) T RH AH

Min. : 221.0 Min. :-1.90 Min. : 9.175 Min. :0.1847

1st Qu.: 731.4 1st Qu.:11.79 1st Qu.:35.812 1st Qu.:0.7368

Median : 963.2 Median :17.75 Median :49.550 Median :0.9954

Mean :1022.8 Mean :18.32 Mean :49.232 Mean :1.0255

3rd Qu.:1273.4 3rd Qu.:24.40 3rd Qu.:62.500 3rd Qu.:1.3137

Max. :2522.8 Max. :44.60 Max. :88.725 Max. :2.2310

NA's :366 NA's :366 NA's :366 NA's :366

Though we understood that there are missing values, a little deeper understanding of the missing values can be had from the following analysis

> library(pastecs)

> stat.desc(AQUCI)

Date Time CO(GT) PT08.S1(CO)

nbr.val 9.357000e+03 9.357000e+03 7.674000e+03 8.991000e+03

nbr.null 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00

nbr.na 0.000000e+00 0.000000e+00 1.683000e+03 3.660000e+02

min 1.078877e+09 -2.209075e+09 1.000000e-01 6.472500e+02

max 1.112573e+09 -2.208992e+09 1.190000e+01 2.039750e+03

range 3.369600e+07 8.280000e+04 1.180000e+01 1.392500e+03

sum 1.025285e+13 -2.066993e+13 1.652020e+04 9.887473e+06

median 1.095725e+09 -2.209036e+09 1.800000e+00 1.063000e+03

mean 1.095741e+09 -2.209034e+09 2.152750e+00 1.099708e+03

SE.mean 1.005323e+05 2.576558e+02 1.658938e-02 2.289417e+00

CI.mean.0.95 1.970652e+05 5.050614e+02 3.251971e-02 4.487780e+00

var 9.456878e+13 6.211785e+08 2.111941e+00 4.712571e+04

std.dev 9.724648e+06 2.492345e+04 1.453252e+00 2.170846e+02

coef.var 8.874952e-03 -1.128251e-05 6.750679e-01 1.974020e-01

NMHC(GT) C6H6(GT) PT08.S2(NMHC) NOx(GT)

nbr.val 9.140000e+02 8.991000e+03 8.991000e+03 7.718000e+03

nbr.null 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00

nbr.na 8.443000e+03 3.660000e+02 3.660000e+02 1.639000e+03

min 7.000000e+00 1.490477e-01 3.832500e+02 2.000000e+00

max 1.189000e+03 6.374148e+01 2.214000e+03 1.479000e+03

range 1.182000e+03 6.359243e+01 1.830750e+03 1.477000e+03

sum 1.999940e+05 9.065619e+04 8.442812e+06 1.905430e+06

median 1.500000e+02 8.239851e+00 9.090000e+02 1.798000e+02

mean 2.188118e+02 1.008299e+01 9.390292e+02 2.468813e+02

SE.mean 6.762933e+00 7.856539e-02 2.814032e+00 2.424201e+00

CI.mean.0.95 1.327270e+01 1.540061e-01 5.516144e+00 4.752091e+00

var 4.180386e+04 5.549713e+01 7.119772e+04 4.535674e+04

std.dev 2.044599e+02 7.449640e+00 2.668290e+02 2.129712e+02

coef.var 9.344099e-01 7.388322e-01 2.841541e-01 8.626464e-01

PT08.S3(NOx) NO2(GT) PT08.S4(NO2) PT08.S5(O3)

nbr.val 8.991000e+03 7.715000e+03 8.991000e+03 8.991000e+03

nbr.null 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00

nbr.na 3.660000e+02 1.642000e+03 3.660000e+02 3.660000e+02

min 3.220000e+02 2.000000e+00 5.510000e+02 2.210000e+02

max 2.682750e+03 3.397000e+02 2.775000e+03 2.522750e+03

range 2.360750e+03 3.377000e+02 2.224000e+03 2.301750e+03

sum 7.510820e+06 8.723776e+05 1.309219e+07 9.195822e+06

median 8.055000e+02 1.090000e+02 1.462750e+03 9.632500e+02

mean 8.353710e+02 1.130755e+02 1.456143e+03 1.022781e+03

SE.mean 2.708423e+00 5.505688e-01 3.651142e+00 4.202459e+00

CI.mean.0.95 5.309127e+00 1.079264e+00 7.157071e+00 8.237778e+00

var 6.595400e+04 2.338617e+03 1.198576e+05 1.587870e+05

std.dev 2.568151e+02 4.835925e+01 3.462045e+02 3.984809e+02

coef.var 3.074264e-01 4.276722e-01 2.377544e-01 3.896054e-01

T RH AH

nbr.val 8.991000e+03 8.991000e+03 8.991000e+03

nbr.null 0.000000e+00 0.000000e+00 0.000000e+00

nbr.na 3.660000e+02 3.660000e+02 3.660000e+02

min -1.900000e+00 9.175000e+00 1.846790e-01

max 4.460000e+01 8.872500e+01 2.231036e+00

range 4.650000e+01 7.955000e+01 2.046357e+00

sum 1.646796e+05 4.426482e+05 9.220542e+03

median 1.775000e+01 4.955000e+01 9.953952e-01

mean 1.831605e+01 4.923236e+01 1.025530e+00

SE.mean 9.315340e-02 1.826221e-01 4.258692e-03

CI.mean.0.95 1.826019e-01 3.579809e-01 8.348007e-03

var 7.801991e+01 2.998573e+02 1.630649e-01

std.dev 8.832888e+00 1.731639e+01 4.038129e-01

coef.var 4.822484e-01 3.517278e-01 3.937602e-01

> naper <- function(x) {sum(is.na(x)/length(x)\*100)}; apply(AQUCI,2,naper); apply(AQUCI,1,naper)

Date Time CO(GT) PT08.S1(CO) NMHC(GT)

0.00000 0.00000 17.98653 3.91151 90.23191

C6H6(GT) PT08.S2(NMHC) NOx(GT) PT08.S3(NOx) NO2(GT)

3.91151 3.91151 17.51630 3.91151 17.54836

PT08.S4(NO2) PT08.S5(O3) T RH AH

3.91151 3.91151 3.91151 3.91151 3.91151

True hourly averaged concentration CO in mg/m^3 (reference analyzer)

True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)

True hourly averaged NOx concentration in ppb (reference analyzer)

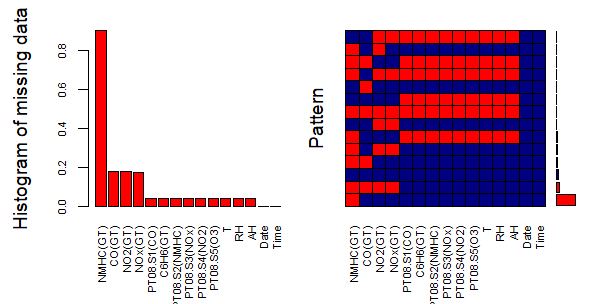
True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)

The above are the variables which display percentage of ‘NA’ values which is beyond the usual norm of 5% of the total number of observations. The Non Metanic HydroCarbons concentration is also to left out for analysis, since it shows that 90% of the data is ‘NA’

The percentage ‘NA’ in other variables shows a peculiar 3.91151 value, which is same across all the variables. This leads to a doubt, whether the observation on a particular day w.r.t. these aspects is completely absent, else technically or the domain experts may comment, whether it does not make sense to collect such an information on such days where the other related variables record a particular value. So it becomes an important point to investigate further.

> library(VIM)

> missdataplot <- aggr(AQUCI, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(data), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))



The visual representation of the missing data shows that predominantly the missing data is in the four variables NMHC(GT), CO(GT), NO2(GT),NOX(GT).

> md.pattern(AQUCI)

Date Time PT08.S1(CO) C6H6(GT) PT08.S2(NMHC) PT08.S3(NOx)

827 1 1 1 1 1 1

6114 1 1 1 1 1 1

24 1 1 1 1 1 1

428 1 1 1 1 1 1

3 1 1 1 1 1 1

36 1 1 1 1 1 1

364 1 1 1 1 1 1

1195 1 1 1 1 1 1

26 1 1 0 0 0 0

291 1 1 0 0 0 0

5 1 1 0 0 0 0

1 1 1 0 0 0 0

12 1 1 0 0 0 0

31 1 1 0 0 0 0

0 0 366 366 366 366

PT08.S4(NO2) PT08.S5(O3) T RH AH NOx(GT) NO2(GT) CO(GT)

827 1 1 1 1 1 1 1 1

6114 1 1 1 1 1 1 1 1

24 1 1 1 1 1 1 1 0

428 1 1 1 1 1 1 1 0

3 1 1 1 1 1 1 0 1

36 1 1 1 1 1 0 0 1

364 1 1 1 1 1 0 0 1

1195 1 1 1 1 1 0 0 0

26 0 0 0 0 0 1 1 1

291 0 0 0 0 0 1 1 1

5 0 0 0 0 0 1 1 0

1 0 0 0 0 0 0 0 1

12 0 0 0 0 0 0 0 1

31 0 0 0 0 0 0 0 0

366 366 366 366 366 1639 1642 1683

NMHC(GT)

827 1 0

6114 0 1

24 1 1

428 0 2

3 0 2

36 1 2

364 0 3

1195 0 4

26 1 9

291 0 10

5 0 11

1 1 11

12 0 12

31 0 13

8443 16701

The pattern shows that 827 rows with complete information, and 6114 and 24 rows with missing information ONLY in NHMC(GT) and CO(GT) columns. If we consider their aggregate 6965 rows with either 0 or 1 column values missing out of the total 9357 rows, making a percentage of 74.44%. So aroung 75% of the data is available in the data set in most of the columns and variables. Only NHMC(GT) data is not available in 6114 out of 9357 rows, meaning around 65% of the rows this data is not present. We can always take a call in dropping this variable completely from analysis and continue with the rest, but if the information is significant then we need to retain it.

**INPUTING THE MISSING VALUES**

Then the process of imputation begins. The convention for imputation of missing data is that variables which have percentage of missing data upto 5% will be chosen for imputation. The rest are not usually considered for analysis. But here for the assignment we are making imputation for all the variables in the dataset.

> colnames(AQUCI)

> colnames(AQUCI)[c(3,4,5,6,7,8,9,10,11,12)] <- c("CO","TINO","NMHC","C6H6","TITAN","NOX", "TUNGOX1","NO2","TUNGOX2","INDOX")

> colnames(AQUCI)

[1] "Date" "Time" "CO" "TINO" "NMHC" "C6H6"

[7] "TITAN" "NOX" "TUNGOX1" "NO2" "TUNGOX2" "INDOX"

[13] "T" "RH" "AH"

> tempdata <- mice(AQUCI,m=5,maxit=50,meth='pmm',seed=500)

iter imp variable

1 1 COError in solve.default(xtx + diag(pen)) :

system is computationally singular: reciprocal condition number = 2.39134e-24

> library(dplyr)

> colnames(AQUCI)[-c(1,2)]

[1] "CO" "TINO" "NMHC" "C6H6" "TITAN" "NOX"

[7] "TUNGOX1" "NO2" "TUNGOX2" "INDOX" "T" "RH"

[13] "AH"

> NUMAQUCI <- AQUCI %>% select(colnames(AQUCI)[-c(1,2)])

> str(NUMAQUCI)

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 9357 obs. of 13 variables:

$ CO : num 2.6 2 2.2 2.2 1.6 1.2 1.2 1 0.9 0.6 ...

$ TINO : num 1360 1292 1402 1376 1272 ...

$ NMHC : num 150 112 88 80 51 38 31 31 24 19 ...

$ C6H6 : num 11.88 9.4 9 9.23 6.52 ...

$ TITAN : num 1046 955 939 948 836 ...

$ NOX : num 166 103 131 172 131 89 62 62 45 NA ...

$ TUNGOX1: num 1056 1174 1140 1092 1205 ...

$ NO2 : num 113 92 114 122 116 96 77 76 60 NA ...

$ TUNGOX2: num 1692 1559 1554 1584 1490 ...

$ INDOX : 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 ...

> NUMAQUCI <- as.data.frame(NUMAQUCI)

> tempdata <- mice(NUMAQUCI,m=5,maxit=50,meth='pmm',seed=500)

> compdata <- complete(tempdata,1)

> nrow(compdata)

> ncol(compdata)

> sum(is.na(compdata))

> compdata <- cbind(compdata,AQUCI$Date,AQUCI$Time)

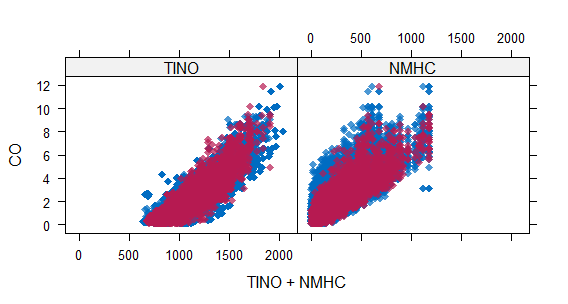
> ncol(compdata)

> sum(is.na(compdata))

> library(lattice)

> library(ggplot2)

> xyplot(tempdata,CO~TINO+NMHC,pch=18,cex=1)



**BIVARIATE ANALYSIS**

> numdata <- complete(tempdata,1)

> library(stats)

> library(car)

> library(Hmisc)

> rcorr(as.matrix(numdata))

CO TINO NMHC C6H6 TITAN NOX TUNGOX1 NO2 TUNGOX2 INDOX

CO 1.00 0.88 0.79 0.94 0.92 0.80 -0.70 0.70 0.64 0.85

TINO 0.88 1.00 0.65 0.88 0.89 0.73 -0.77 0.66 0.68 0.90

NMHC 0.79 0.65 1.00 0.82 0.79 0.40 -0.55 0.40 0.71 0.61

C6H6 0.94 0.88 0.82 1.00 0.98 0.74 -0.73 0.64 0.76 0.87

TITAN 0.92 0.89 0.79 0.98 1.00 0.72 -0.79 0.67 0.77 0.88

NOX 0.80 0.73 0.40 0.74 0.72 1.00 -0.65 0.76 0.27 0.79

TUNGOX1 -0.70 -0.77 -0.55 -0.73 -0.79 -0.65 1.00 -0.65 -0.54 -0.79

NO2 0.70 0.66 0.40 0.64 0.67 0.76 -0.65 1.00 0.19 0.72

TUNGOX2 0.64 0.68 0.71 0.76 0.77 0.27 -0.54 0.19 1.00 0.59

INDOX 0.85 0.90 0.61 0.87 0.88 0.79 -0.79 0.72 0.59 1.00

T 0.03 0.04 0.28 0.19 0.23 -0.25 -0.14 -0.17 0.56 -0.03

RH 0.02 0.12 -0.14 -0.06 -0.08 0.19 -0.07 -0.12 -0.02 0.13

AH 0.04 0.13 0.20 0.16 0.18 -0.14 -0.23 -0.33 0.63 0.07

T RH AH

CO 0.03 0.02 0.04

TINO 0.04 0.12 0.13

NMHC 0.28 -0.14 0.20

C6H6 0.19 -0.06 0.16

TITAN 0.23 -0.08 0.18

NOX -0.25 0.19 -0.14

TUNGOX1 -0.14 -0.07 -0.23

NO2 -0.17 -0.12 -0.33

TUNGOX2 0.56 -0.02 0.63

INDOX -0.03 0.13 0.07

T 1.00 -0.56 0.65

RH -0.56 1.00 0.19

AH 0.65 0.19 1.00

n= 9357

P

CO TINO NMHC C6H6 TITAN NOX TUNGOX1 NO2 TUNGOX2

CO 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

TINO 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

NMHC 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

C6H6 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

TITAN 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

NOX 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

TUNGOX1 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

NO2 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

TUNGOX2 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

INDOX 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

T 0.0014 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

RH 0.0747 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.1313

AH 0.0003 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

INDOX T RH AH

CO 0.0000 0.0014 0.0747 0.0003

TINO 0.0000 0.0000 0.0000 0.0000

NMHC 0.0000 0.0000 0.0000 0.0000

C6H6 0.0000 0.0000 0.0000 0.0000

TITAN 0.0000 0.0000 0.0000 0.0000

NOX 0.0000 0.0000 0.0000 0.0000

TUNGOX1 0.0000 0.0000 0.0000 0.0000

NO2 0.0000 0.0000 0.0000 0.0000

TUNGOX2 0.0000 0.0000 0.1313 0.0000

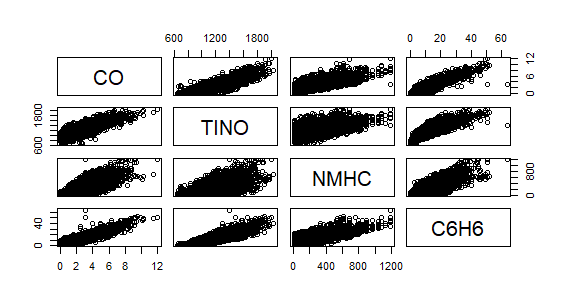
INDOX 0.0023 0.0000 0.0000

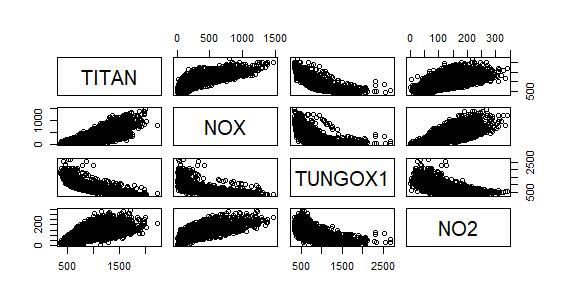
T 0.0023 0.0000 0.0000

RH 0.0000 0.0000 0.0000

AH 0.0000 0.0000 0.0000

> pairs(~CO+TINO+NMHC+C6H6, data=numdata)



> pairs(~TITAN+NOX+TUNGOX1+NO2, data=numdata)

> pairs(~TUNGOX2+INDOX+T+RH+AH,data=numdata)



> colnames(compdata)[14] <- "Date"

> colnames(compdata)[15] <- "Time"

> install.packages("lubridate")

> library(lubridate)

> head(compdata$Time)

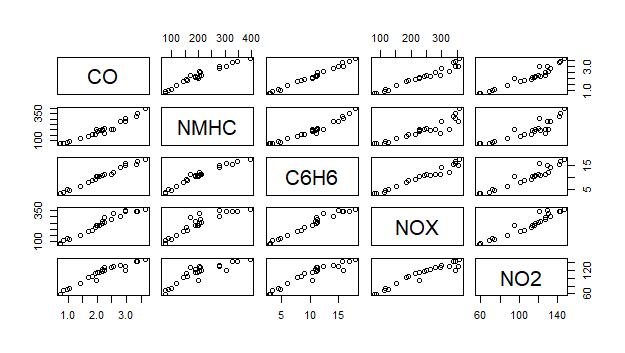
> head(compdata$Date)

>timepolu<-aggregate(cbind(CO,TINO,NMHC,C6H6,TITAN,NOX,TUNGOX1,NO2,TUNGOX2,INDOX)

~hour(Time),data=compdata,mean)

> timepolu <- as.data.frame(timepolu)

> pairs(~CO+NMHC+C6H6+NOX+NO2, data = timepolu)



> polnumdata <- numdata %>% select("CO",'NMHC',"C6H6","NOX","NO2","T","RH","AH")

> library(Hmisc)

> rcorr(as.matrix(polnumdata))

CO NMHC C6H6 NOX NO2 T RH AH

CO 1.00 0.79 0.94 0.80 0.70 0.03 0.02 0.04

NMHC 0.79 1.00 0.82 0.40 0.40 0.28 -0.14 0.20

C6H6 0.94 0.82 1.00 0.74 0.64 0.19 -0.06 0.16

NOX 0.80 0.40 0.74 1.00 0.76 -0.25 0.19 -0.14

NO2 0.70 0.40 0.64 0.76 1.00 -0.17 -0.12 -0.33

T 0.03 0.28 0.19 -0.25 -0.17 1.00 -0.56 0.65

RH 0.02 -0.14 -0.06 0.19 -0.12 -0.56 1.00 0.19

AH 0.04 0.20 0.16 -0.14 -0.33 0.65 0.19 1.00

n= 9357

P

CO NMHC C6H6 NOX NO2 T RH AH

CO 0.0000 0.0000 0.0000 0.0000 0.0014 0.0747 0.0003

NMHC 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

C6H6 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

NOX 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

NO2 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

T 0.0014 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

RH 0.0747 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

AH 0.0003 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

**TESTING RELEVANT HYPOTHESIS**

1. Null Hypothesis = The concentration of CO is independent on the day of the week.
2. Alternate Hypothesis = The concentration of CO is dependent on the day of the week.

> colnames(compdata)

[1] "CO" "TINO" "NMHC" "C6H6" "TITAN" "NOX" "TUNGOX1"

[8] "NO2" "TUNGOX2" "INDOX" "T" "RH" "AH" "Date"

[15] "Time" "Day" "Month" "Hour"

> rescoday <- aov(CO~Day,data=compdata)

> summary(rescoday)

Df Sum Sq Mean Sq F value Pr(>F)

Day 6 1089 181.57 91.35 <2e-16 \*\*\*

Residuals 9350 18585 1.99

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The above results show clearly that CO emissions are very strongly dependent on the day of the week. The large F value and the almost 0% significance level show that the NULL hypothesis receives the necessary statistical power to be rejected and hence to consider the alternate hypothesis that CO concentrations are dependent on the day of the week. The Tukey results below also show that the difference between Beginning days and the Ending days of the week are very significant.

> TukeyHSD(rescoday)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = CO ~ Day, data = compdata)

$`Day`

diff lwr upr p adj

Monday-Friday -0.41471425 -0.5753665 -0.25406198 0.0000000

Saturday-Friday -0.61979167 -0.7801739 -0.45940947 0.0000000

Sunday-Friday -1.00550595 -1.1658881 -0.84512376 0.0000000

Saturday-Monday -0.20507741 -0.3657297 -0.04442514 0.0031871

Sunday-Monday -0.59079170 -0.7514440 -0.43013943 0.0000000

Thursday-Monday 0.35571128 0.1950590 0.51636355 0.0000000

Tuesday-Monday 0.30341420 0.1420358 0.46479261 0.0000006

Wednesday-Monday 0.27090803 0.1097133 0.43210275 0.0000151

Sunday-Saturday -0.38571429 -0.5460965 -0.22533209 0.0000000

Thursday-Saturday 0.56078869 0.4004065 0.72117088 0.0000000

Tuesday-Saturday 0.50849161 0.3473821 0.66960117 0.0000000

Wednesday-Saturday 0.47598544 0.3150599 0.63691100 0.0000000

Thursday-Sunday 0.94650298 0.7861208 1.10688517 0.0000000

Tuesday-Sunday 0.89420590 0.7330963 1.05531545 0.0000000

Wednesday-Sunday 0.86169973 0.7007742 1.02262528 0.0000000

**CROSS TABULATIONS WITH DERIVED VARIABLES**

> table(compdata$CO)

0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9

43 56 139 222 309 332 337 327 310 389 333 297 304 336 333 341 322 226 283

2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3 3.1 3.2 3.3 3.4 3.5 3.6 3.7 3.8

272 257 242 225 198 176 185 163 162 132 135 136 121 122 120 104 102 68 86

3.9 4 4.1 4.2 4.3 4.4 4.5 4.6 4.7 4.8 4.9 5 5.1 5.2 5.3 5.4 5.5 5.6 5.7

85 85 72 60 69 55 52 59 38 44 41 46 34 21 16 24 26 27 21

5.8 5.9 6 6.1 6.2 6.3 6.4 6.5 6.6 6.7 6.8 6.9 7 7.1 7.2 7.3 7.4 7.5 7.6

16 19 9 13 18 14 11 14 15 10 11 5 1 2 5 7 4 9 1

7.7 7.8 7.9 8 8.1 8.3 8.4 8.5 8.6 8.7 9.1 9.2 9.3 9.4 9.5 9.9 10.1 10.2 11.5

6 2 3 6 5 2 6 2 5 3 3 1 1 1 1 1 1 2 1

11.9

1

> compdata <- mutate(compdata, COLEVEL = ifelse(CO <= 3,"LOW",ifelse(CO >3 & CO <= 6,"MODERATE",ifelse(CO > 6 & CO <= 9,"HIGH",ifelse(CO > 9,"ALARM","NA")))))

> table(compdata$COLEVEL)

ALARM HIGH LOW MODERATE

13 180 7386 1778

> head(compdata$COLEVEL)

[1] "LOW" "LOW" "LOW" "LOW" "LOW" "LOW"

> with(compdata, tapply(CO, list(day,COLEVEL),FUN=Mean))

Error in unique.default(x, nmax = nmax) :

unique() applies only to vectors

> table(compdata$Day, compdata$COLEVEL)

ALARM HIGH LOW MODERATE

Friday 1 41 958 344

Monday 2 36 1071 226

Saturday 0 1 1183 160

Sunday 0 2 1265 77

Thursday 3 39 983 319

Tuesday 6 30 947 337

Wednesday 1 31 979 315

> table(compdata$Hour, compdata$COLEVEL)

ALARM HIGH LOW MODERATE

0 2 46 168 174

1 4 25 193 168

2 0 7 291 92

3 0 0 353 37

4 0 0 361 29

5 0 0 366 24

6 0 0 373 17

7 0 0 382 8

8 0 0 389 1

9 0 0 389 1

10 0 0 390 0

11 0 0 389 1

12 0 2 311 77

13 0 20 206 164

14 0 15 212 163

15 0 3 264 123

16 0 3 300 87

17 0 0 318 72

18 0 1 319 70

19 0 3 325 62

20 0 3 329 57

21 0 4 315 70

22 0 21 250 118

23 7 27 193 163

**PATTERNS IN TIME SERIES AND TRENDS**

> daypolu <- aggregate(cbind(CO,NMHC,C6H6,NOX,NO2)~weekdays(Date),data=compdata,mean)

> daypolu

weekdays(Date) CO NMHC C6H6 NOX NO2

1 Friday 2.428497 237.5990 11.954476 266.8183 118.51979

2 Monday 2.013783 198.1993 9.989174 228.1749 106.57880

3 Saturday 1.808705 145.0037 8.475635 212.5828 104.77597

4 Sunday 1.422991 116.8118 6.087604 154.2033 89.16905

5 Thursday 2.369494 221.4271 11.695314 266.1958 117.14695

6 Tuesday 2.317197 232.8788 11.757576 261.8845 114.53500

7 Wednesday 2.284691 219.2202 11.414197 259.8796 115.50611

> timepolu1 <- aggregate(cbind(CO,NMHC,C6H6,NOX,NO2)~hour(Time),data=compdata,mean)

> timepolu1

hour(Time) CO NMHC C6H6 NOX NO2

1 0 3.6523077 394.64359 17.815002 358.35000 147.49231

2 1 3.3441026 326.60769 15.571180 339.29795 141.75821

3 2 2.4946154 207.65128 11.266639 277.88051 127.53641

4 3 1.9271795 156.58462 9.088093 214.74103 112.55821

5 4 1.8202564 158.34103 8.518392 193.15231 105.75077

6 5 1.7107692 143.53846 7.820645 186.46103 101.03949

7 6 1.4097436 119.12051 6.128055 154.02564 87.87821

8 7 1.0546154 98.43590 4.490881 116.22359 73.88615

9 8 0.8505128 77.58205 3.493694 114.53949 68.55590

10 9 0.7435897 77.43590 3.003365 81.65744 58.32333

11 10 0.7371795 74.96154 3.182142 88.41410 59.55744

12 11 0.9671795 87.55897 4.787204 125.45769 72.35487

13 12 1.9666667 201.09744 10.386174 228.95179 95.48590

14 13 2.9671795 305.91026 15.850869 345.05154 121.09923

15 14 2.9641026 277.93077 14.896083 352.99103 129.40564

16 15 2.5351282 204.63333 12.319607 324.16359 130.26667

17 16 2.2561538 168.98462 11.143009 296.60026 127.84615

18 17 2.1671795 190.63077 10.946891 263.16821 121.92026

19 18 2.1669231 204.94872 11.019043 243.05667 117.52436

20 19 2.0628205 195.21282 10.417378 227.93205 113.71231

21 20 2.0156812 188.53728 10.344801 228.29769 115.94627

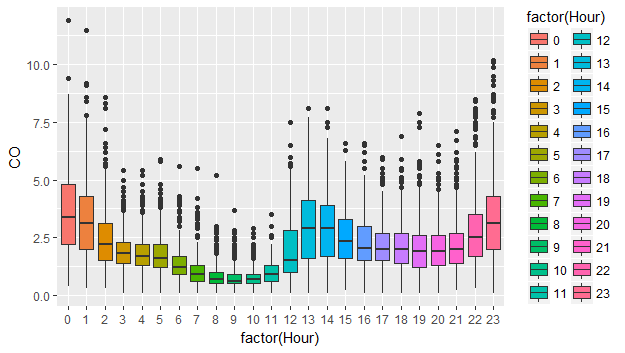
22 21 2.2305913 210.61440 11.218693 250.66427 120.73985

23 22 2.7845758 279.36761 14.064823 298.65630 133.25553

24 23 3.3646154 347.54615 16.803522 344.13897 142.76846

> compdata$Hour <- hour(compdata$Time)

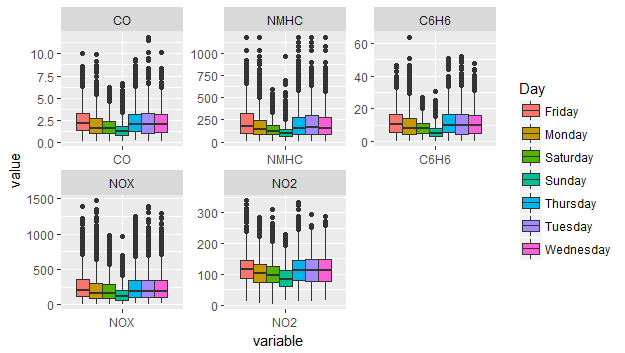
> ggplot(compdata,aes(factor(Hour),CO,fill=factor(Hour))) + geom\_boxplot()



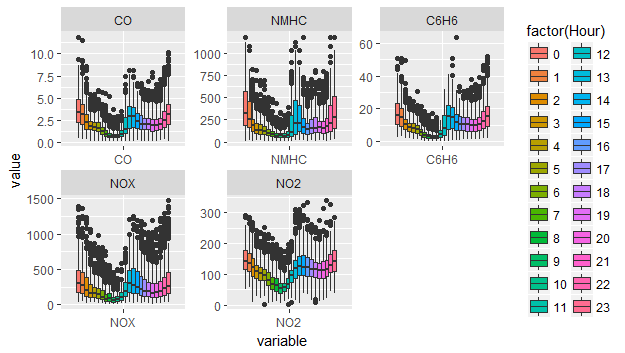
> polnumdata <- compdata %>% select("CO","NMHC","C6H6","NOX","NO2","T","RH","AH","Date","Time","Day","Hour")

> Day = polnumdata %>% select("CO","NMHC","C6H6","NOX","NO2","Day") %>% melt(id.var="Day")

> ggplot(Day,aes(variable,value))+geom\_boxplot(aes(fill=Day))+facet\_wrap(~variable,scales='free')



> ggplot(Hour,aes(variable,value))+geom\_boxplot(aes(fill=factor(Hour)))+facet\_wrap(~variable,scales='free')



**MOST POLLUTED TIME OF THE DAY**

> timepolu1

hour(Time) CO NMHC C6H6 NOX NO2

1 0 3.6523077 394.64359 17.815002 358.35000 147.49231

2 1 3.3441026 326.60769 15.571180 339.29795 141.75821

3 2 2.4946154 207.65128 11.266639 277.88051 127.53641

4 3 1.9271795 156.58462 9.088093 214.74103 112.55821

5 4 1.8202564 158.34103 8.518392 193.15231 105.75077

6 5 1.7107692 143.53846 7.820645 186.46103 101.03949

7 6 1.4097436 119.12051 6.128055 154.02564 87.87821

8 7 1.0546154 98.43590 4.490881 116.22359 73.88615

9 8 0.8505128 77.58205 3.493694 114.53949 68.55590

10 9 0.7435897 77.43590 3.003365 81.65744 58.32333

11 10 0.7371795 74.96154 3.182142 88.41410 59.55744

12 11 0.9671795 87.55897 4.787204 125.45769 72.35487

13 12 1.9666667 201.09744 10.386174 228.95179 95.48590

14 13 2.9671795 305.91026 15.850869 345.05154 121.09923

15 14 2.9641026 277.93077 14.896083 352.99103 129.40564

16 15 2.5351282 204.63333 12.319607 324.16359 130.26667

17 16 2.2561538 168.98462 11.143009 296.60026 127.84615

18 17 2.1671795 190.63077 10.946891 263.16821 121.92026

19 18 2.1669231 204.94872 11.019043 243.05667 117.52436

20 19 2.0628205 195.21282 10.417378 227.93205 113.71231

21 20 2.0156812 188.53728 10.344801 228.29769 115.94627

22 21 2.2305913 210.61440 11.218693 250.66427 120.73985

23 22 2.7845758 279.36761 14.064823 298.65630 133.25553

24 23 3.3646154 347.54615 16.803522 344.13897 142.76846

Note: The above table shows the mean hourly concentrations of various compounds, by the hour of the day for the entire 1 year of observations. Midnight 12 O clock seems to be the most polluted time of the day and Non Methane Hydro Carbons are the chemicals which are polluting the most.