Acadgild – Data Analytics – Batch 4 Assignment

SESSION: 6 To 10

Task 1:

1. Import the Titanic Dataset from the link **=>**[Titanic Data Set](#).

Perform the following:

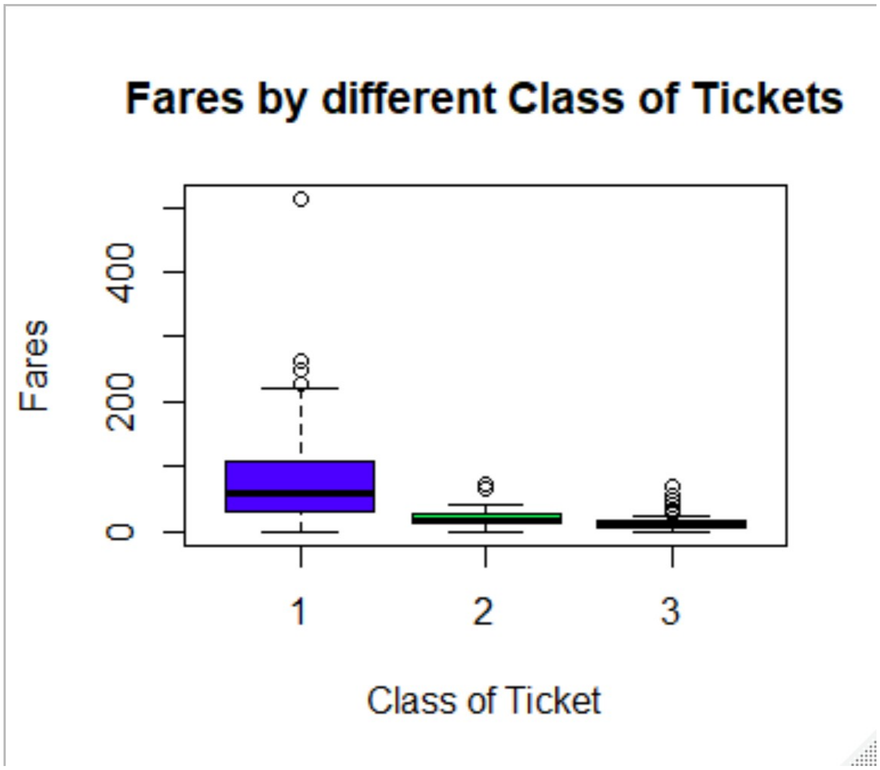
a. Is there any difference in fares by a different class of tickets?

Note - Show a boxplot displaying the distribution of fares by class

Solution:

R Script:

```
library("readr")
library(readxl)
TitanicData <- read_xls("D:/DocumentsR/R Scripts & Data- acadgild sessions/data files R
sessions/titanic3.xls")

View(TitanicData)
str(TitanicData)

colnames(TitanicData) <-
c("Pclass","Survived","Name","Sex","Age","SibSp","Parch","Ticket","Fare",
                    "Cabin","Embarked","Boat","Body","destination")

Titanic <- TitanicData %>% mutate(Pclass = as.factor(Pclass))  # Passennger class as factor
str(Titanic)
View(Titanic)

boxplot(Fare~Pclass, data = Titanic, col = topo.colors(3),
      xlab = "Class of Ticket", ylab = "Fares", main = "Fares by different Class of Tickets")
```



Yes- fares are different for different class of accommodation.

Task

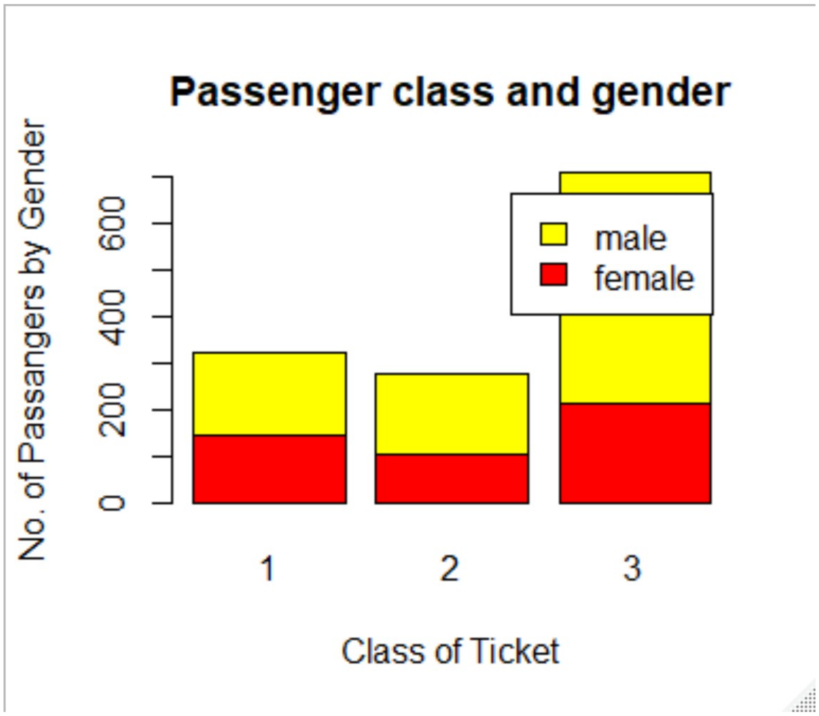
b. Is there any association with Passenger class and gender?

Note – Show a stacked bar chart

Solution:

R Script:

```
A<- table(Titanic$Sex, Titanic$Pclass)
A
str(A)
head(A)

bp <- barplot(A, col= rainbow(length(A)), legend = rownames(A),
          main = "Passenger class and gender",
          xlab = "Class of Ticket", ylab = "No. of Passangers by Gender")
```



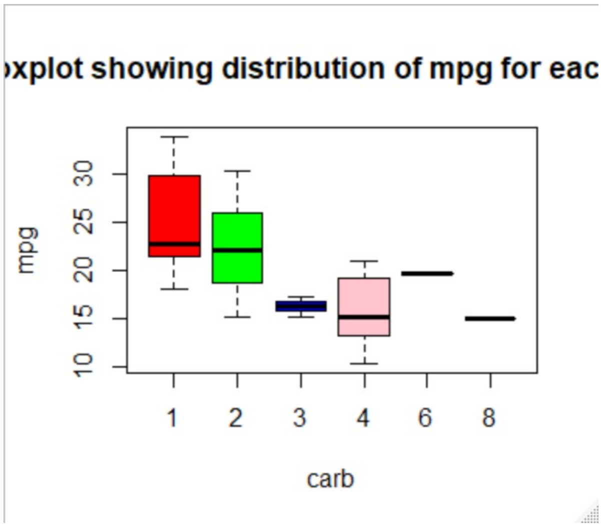
Conclusion/Interpretation:

- Male passengers are more than female in each class .
- The percentage of male passengers over Female Passengers is more in class 3 as compared to class 1 & 2 but females are higher in proportion in Class 1 than in class 2 & 3 as compared to males in each class.

Task 2:

1. Create a box and whisker plot by class using mtcars dataset.

Solution



R Script

```
### mtcars
library(readr)
library(ggplot2)
library(dplyr)
mtcars
View(mtcars)
str(mtcars)
mtcars1 <- mutate(mtcars,
                  cyl = as.factor(cyl),
                  disp = as.factor(disp),
                  vs = as.factor(vs),
                  am = as.factor(am),
                  gear = as.factor(gear),
                  carb = as.factor(carb),
                  mpg = mpg, hp = hp, drat = drat, qsec=qsec)
str(mtcars1)

boxplot(mpg~carb, data = mtcars1, col =
c("Red","Green","Blue","Pink","yellow","orange"),main="Boxplot showing distribution of mpg for
each carb")
```

Task 3:

1. A recent national study showed that approximately 44.7% of college students have used Wikipedia as a source in at least one of their term papers. Let X equal the number of students in a random sample of size n = 31 who have used Wikipedia as a source.

Perform the below functions

a. Find the probability that X is equal to 17

b. Find the probability that X is at most 13

c. Find the probability that X is bigger than 11.

d. Find the probability that X is at least 15.

e. Find the probability that X is between 16 and 19, inclusive

Solution

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

```
# a. Find the probability that X is equal to 17 dbinom(17, 31, 0.447)

# b. Find the probability that X is at most 13 pbinom(13, 31, 0.447)

# c. Find the probability that X is bigger than 11. pbinom(11, 31, 0.447, lower.tail = F)

# d. Find the probability that X is at least 15. pbinom(14, 31, 0.447, lower.tail = F)

# e. Find the probability that X is between 16 and 19, inclusive sum(dbinom(16:19, 31, 0.447))
diff(pbinom(c(19,15), 31, 0.447, lower.tail = FALSE))
```

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

```
> # a. Find the probability that X is equal to 17 > dbinom(17, 31, 0.447) [1] 0.07532248

> # b. Find the probability that X is at most 13 > pbinom(13, 31, 0.447) [1] 0.451357 > # c. Find
the probability that X is bigger than 11. > pbinom(11, 31, 0.447, lower.tail = F)

[1] 0.8020339

> # d. Find the probability that X is at least 15. > pbinom(14, 31, 0.447, lower.tail = F) [1]
0.406024
```

```
> # e. Find the probability that X is between 16 and 19, inclusive > sum(dbinom(16:19, 31,
0.447)) [1] 0.2544758 > diff(pbinom(c(19,15), 31, 0.447, lower.tail = FALSE)) [1] 0.2544758
```

Conclusion/Interpretation:

a) 0.07532248 is the probability that x is equal to 17
b) 0.451357 is the probability that x is at most 13
c) 0.8020339 is the probability that x is bigger than 11
d) 0.406024 is the probability that x is at least 15
e) 0.2544758 is the probability between 16 and 19 , inclusive

Task 4:

1. If Z is norm (mean = 0, sd = 1)

Find P(Z > 2.64)
Find P(|Z| > 1.39)

2. Suppose p = the proportion of students who are admitted to the graduate school of the University of California at Berkeley, and suppose that a public relation officer boasts that UCB has historically had a 40% acceptance rate for its graduate school. Consider the data stored in the table UCBAdmissions from 1973. Assuming these observations constituted a simple random sample, are they consistent with the officerâ..s claim, or do they provide evidence that the acceptance rate was significantly less than 40%? Use an Î± = 0.01 significance level.
3. How do you test the proportions and compare against hypothetical props?

Test Hypothesis: the proportion of automatic cars is 40%.

SOLUTION

The R-script for the given problem is as follows: # 1. If Z is norm (mean = 0, sd = 1)

```
#  Find P(Z > 2.64) pnorm(2.64, mean = 0, sd = 1, lower.tail = FALSE)

#  Find P(|Z| > 1.39) 1 - (pnorm(1.39, mean = 0, sd=1) - pnorm(-1.39, mean = 0, sd=1))
```

The output of the R-Script (from Console window) is given as follows: > pnorm(2.64, mean = 0, sd = 1, lower.tail = FALSE) [1] 0.004145301 > 1 - (pnorm(1.39, mean = 0, sd=1) - pnorm(-1.39, mean = 0, sd=1)) [1] 0.1645289

```
>pnorm(2.64, mean = 0, sd = 1, lower.tail = FALSE)
[1] 0.004145301
>#  Find P(|Z| > 1.39)
>#  = 1 - P(-1.39 < X < 1.39)
>1 - (pnorm(1.39, mean = 0, sd=1) - pnorm(-1.39, mean = 0, sd=1))
[1] 0.1645289
```

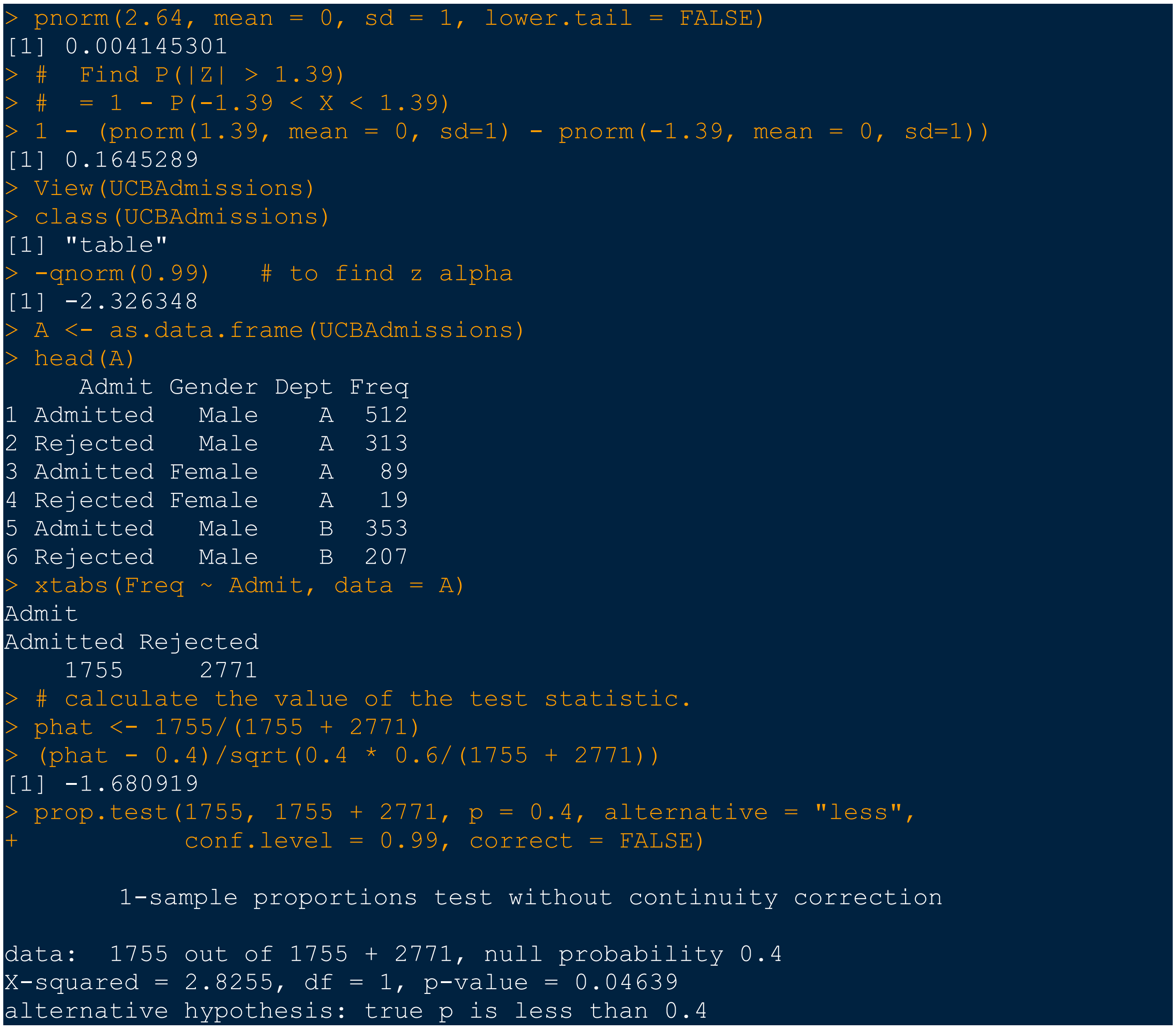
Conclusion/Interpretation:

¬⅄ P(Z > 2.64) 0.004145301
¬⅄ P(|Z| > 1.39) is 0.1645289

```
>View(UCBAdmissions)
>class(UCBAdmissions)
[1] "table"
>-qnorm(0.99)   # to find z alpha
[1] -2.326348
>A <- as.data.frame(UCBAdmissions)
>head(A)
     Admit Gender Dept Freq
1 Admitted   Male    A  512
2 Rejected   Male    A  313
3 Admitted Female    A   89
4 Rejected Female    A   19
5 Admitted   Male    B  353
6 Rejected   Male    B  207
>xtabs(Freq ~ Admit, data = A)
Admit
Admitted Rejected
    1755     2771
># calculate the value of the test statistic.
>phat <- 1755/(1755 + 2771)
>(phat - 0.4)/sqrt(0.4 * 0.6/(1755 + 2771))
[1] -1.680919
>prop.test(1755, 1755 + 2771, p = 0.4, alternative = "less",
+           conf.level = 0.99, correct = FALSE)

        1-sample proportions test without continuity correction

data:  1755 out of 1755 + 2771, null probability 0.4
X-squared = 2.8255, df = 1, p-value = 0.04639
alternative hypothesis: true p is less than 0.4
99 percent confidence interval:
 0.0000000 0.4047326
sample estimates:
        p
0.3877596
```

Conclusion/Interpretation:

- Null hypothesis, H0 is p= 0.40
- Alternative Hypothesis , Ha is p < 0.4
- z alpha = -2.326348 is found
- t-statistics is -1.680919.
- p- value i.e. 0.046 is greater than alpha i.e. 0.01
- The p value does not fall into the critical region. We fail to reject the null hypothesis that “the true proportion of students admitted to graduate school is less than 40% and say that the observed data are consistent with the officer's claim at the alpha = 0.01 significance level.



```
> pnorm(2.64, mean = 0, sd = 1, lower.tail = FALSE)
[1] 0.004145301
> #  Find P(|Z| > 1.39)
> #  = 1 - P(-1.39 < X < 1.39)
> 1 - (pnorm(1.39, mean = 0, sd=1) - pnorm(-1.39, mean = 0, sd=1))
[1] 0.1645289
> View(UCBAdmissions)
> class(UCBAdmissions)
[1] "table"
> -qnorm(0.99)   # to find z alpha
[1] -2.326348
> A <- as.data.frame(UCBAdmissions)
> head(A)
     Admit Gender Dept Freq
1 Admitted   Male    A  512
2 Rejected   Male    A  313
3 Admitted Female    A   89
4 Rejected Female    A   19
5 Admitted   Male    B  353
6 Rejected   Male    B  207
> xtabs(Freq ~ Admit, data = A)
Admit
Admitted Rejected
    1755     2771
> # calculate the value of the test statistic.
> phat <- 1755/(1755 + 2771)
>  (phat - 0.4)/sqrt(0.4 * 0.6/(1755 + 2771))
[1] -1.680919
> prop.test(1755, 1755 + 2771, p = 0.4, alternative = "less",
+            conf.level = 0.99, correct = FALSE)

	1-sample proportions test without continuity correction

data:  1755 out of 1755 + 2771, null probability 0.4
X-squared = 2.8255, df = 1, p-value = 0.04639
alternative hypothesis: true p is less than 0.4
```

```
99 percent confidence interval:
 0.0000000 0.4047326
sample estimates:
        p
0.3877596
```

Task 5:

Import dataset from the following link:AirQuality Data Set

Perform the following written operations:

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

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 bivariate 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.

1.Expected Output

Solution report with commands, explanation of commands, and screenshots of the output should be submitted in .pdf format on GitHub the same GitHub should expected to submit on student dashboard. This assignment contains 700 marks and will be evaluated within 14 days of submission.

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("C:/Users/Jagannath/Downloads/AirQualityUCI.xlsx")
>View(AirQualityUCI)
>dim(AirQualityUCI)
[1] 9357   15
>str(AirQualityUCI)
```

b) Create Univariate for all the columns. The R-script for the given problem is as follows: library(psych) describe(Air)

Conclusion/Interpretation: Univariate for all the columns is created using describe() function

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

```
AH               15 9357   -6.84  38.98    0.98    0.99   0.45 -200    2.23
202.23 -4.75     20.60
                 se
Date             NA
Time             NA
CO(GT)         0.80
PT08.S1(CO)    3.41
NMHC(GT)       1.45
C6H6(GT)       0.43
PT08.S2(NMHC)  3.54
NOx(GT)        2.66
PT08.S3(NOx)   3.33
NO2(GT)        1.31
PT08.S4(NO2)   4.83
PT08.S5(O3)    4.72
T              0.45
RH             0.53
AH             0.40
```

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)

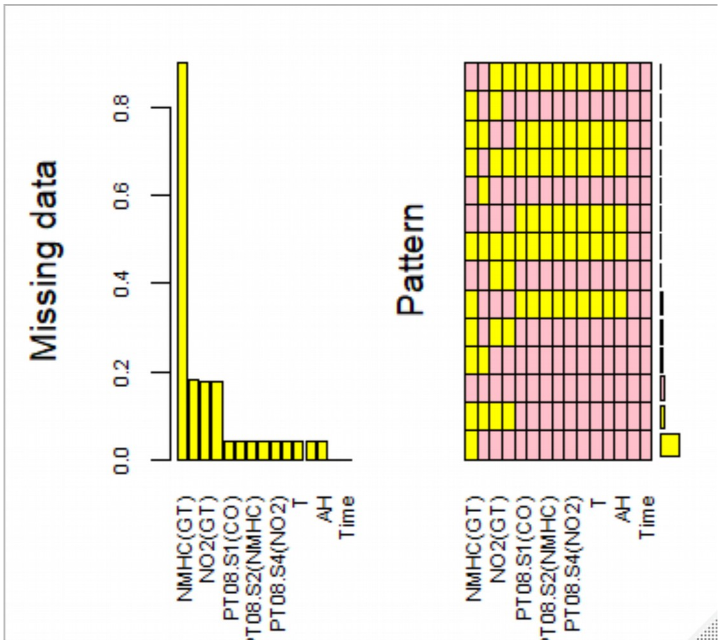
#or

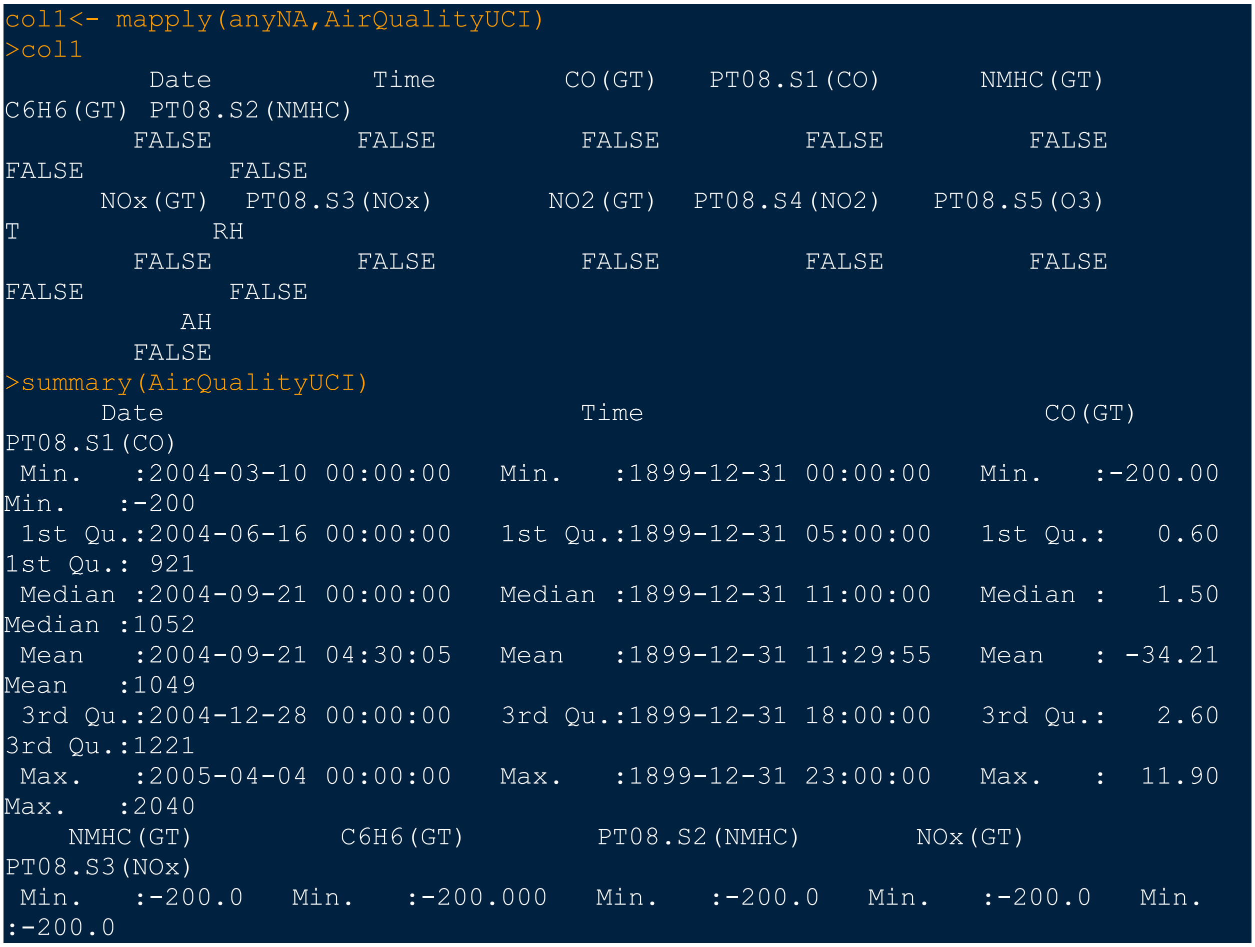
AirQualityUCI[AirQualityUCI == -200] <- NA View(AirQualityUCI) library(VIM) aggr(AirQualityUCI, col=c('pink','yellow'),   numbers=TRUE, sortVars=TRUE, labels=names(AirQualityUCI), cex.axis=.7,   gap=3, ylab=c("Missing data","Pattern"))   # graphical presentation of NAs

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

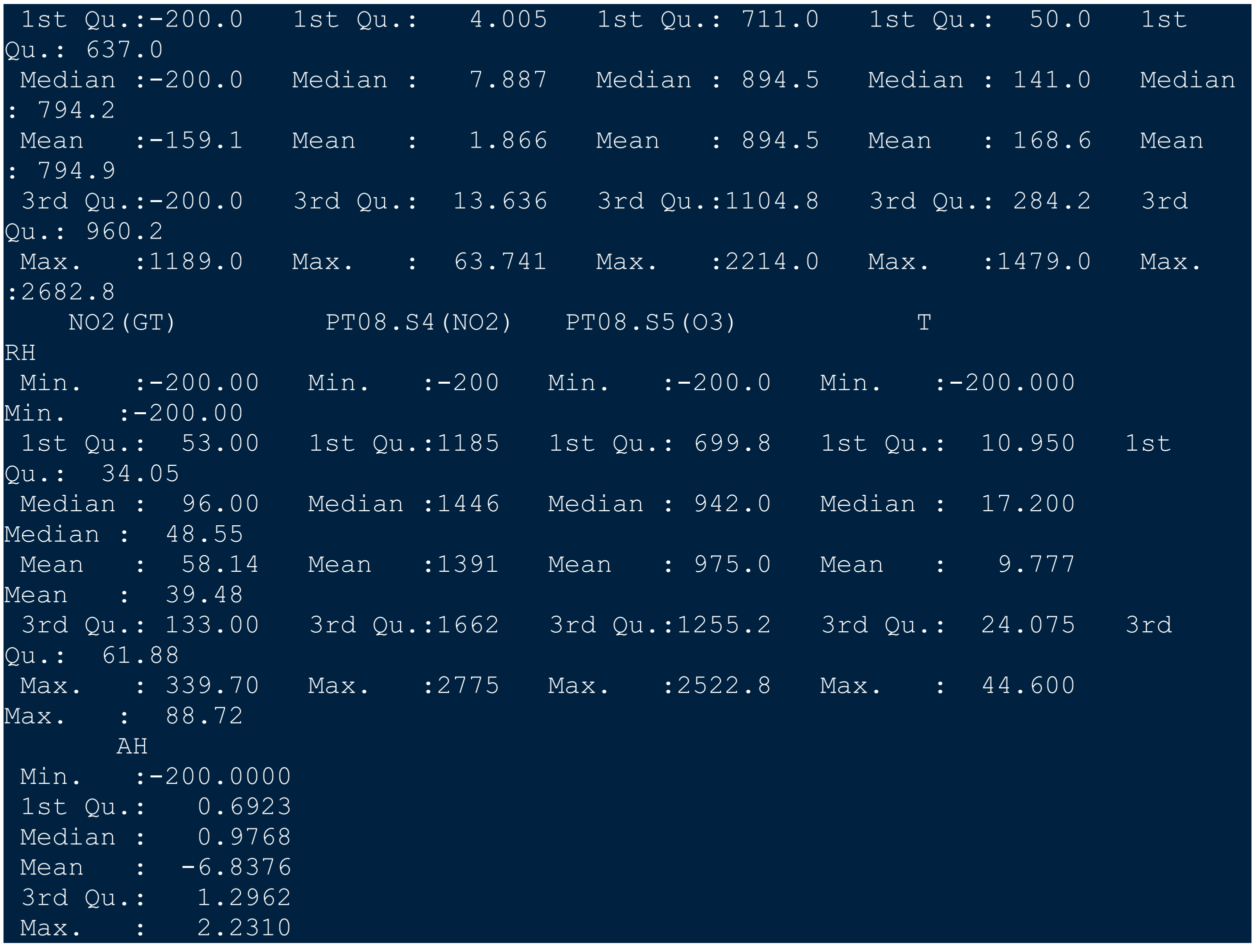
AirQualityUCI$`NMHC(GT)` <- NULL
> Air[Air == -200] <- NA > View(Air) > library(VIM) > aggr(Air, col=c('pink','yellow'), + numbers=TRUE, sortVars=TRUE, +   labels=names(Air), cex.axis=.7, +   gap=3, ylab=c("Missing data","Pattern"))   # graphical presentation of NAs

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

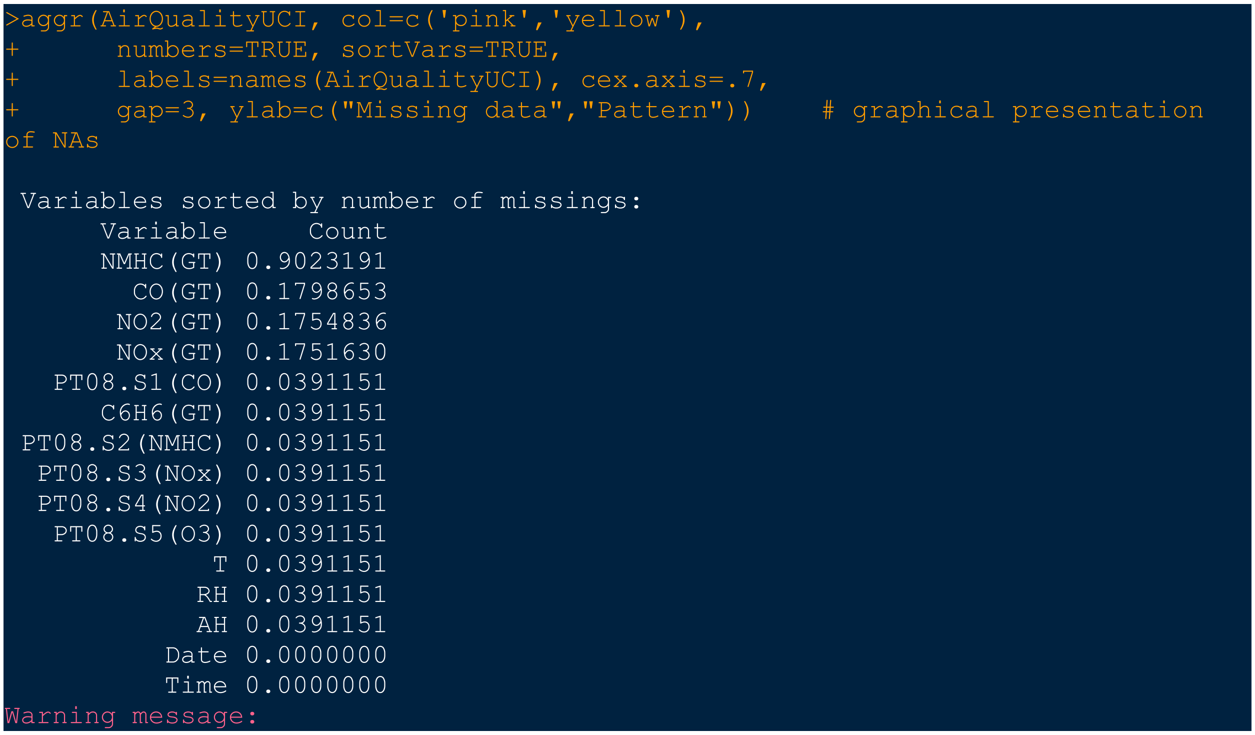




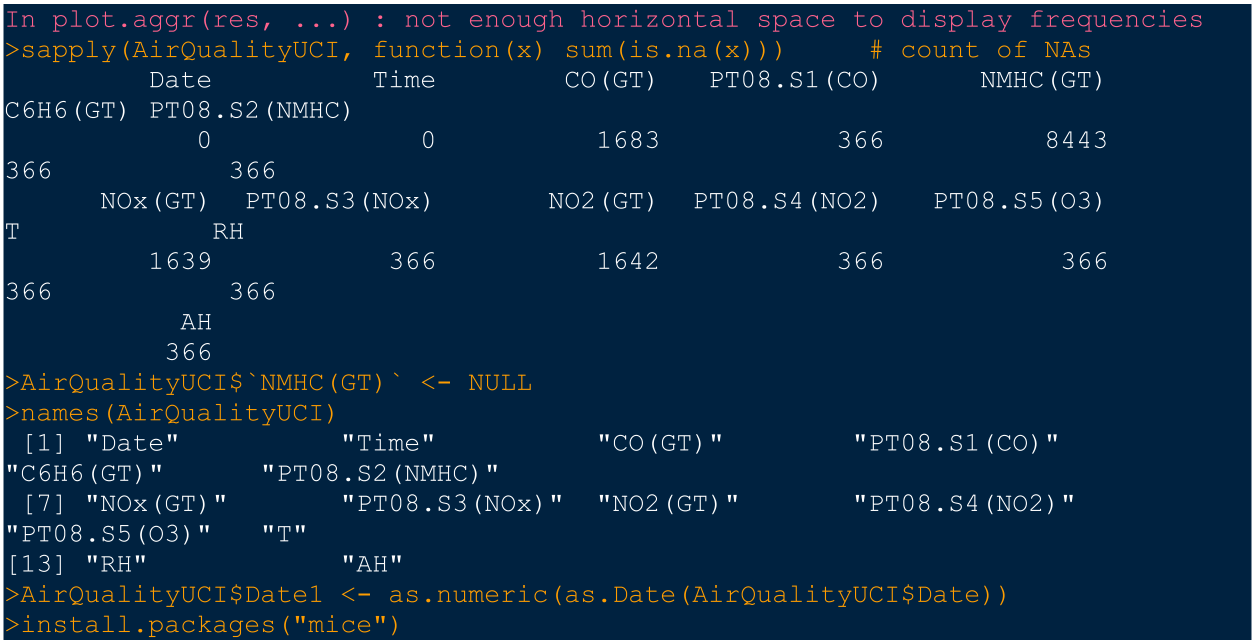
```
col1<- mapply(anyNA,AirQualityUCI)
>col1
         Date           Time         CO(GT)    PT08.S1(CO)       NMHC(GT)
C6H6(GT) PT08.S2(NMHC)
        FALSE          FALSE          FALSE          FALSE          FALSE
FALSE         FALSE
      NOx(GT)   PT08.S3(NOx)         NO2(GT)   PT08.S4(NO2)    PT08.S5(O3)
T            RH
        FALSE          FALSE          FALSE          FALSE          FALSE
FALSE         FALSE
           AH
        FALSE
>summary(AirQualityUCI)
      Date                              Time                            CO(GT)
PT08.S1(CO)
 Min.   :2004-03-10 00:00:00   Min.   :1899-12-31 00:00:00   Min.   :-200.00
Min.   :-200
 1st Qu.:2004-06-16 00:00:00   1st Qu.:1899-12-31 05:00:00   1st Qu.:   0.60
1st Qu.: 921
 Median :2004-09-21 00:00:00   Median :1899-12-31 11:00:00   Median :   1.50
Median :1052
 Mean   :2004-09-21 04:30:05   Mean   :1899-12-31 11:29:55   Mean   : -34.21
Mean   :1049
 3rd Qu.:2004-12-28 00:00:00   3rd Qu.:1899-12-31 18:00:00   3rd Qu.:   2.60
3rd Qu.:1221
 Max.   :2005-04-04 00:00:00   Max.   :1899-12-31 23:00:00   Max.   :  11.90
Max.   :2040
    NMHC(GT)          C6H6(GT)         PT08.S2(NMHC)        NOx(GT)
PT08.S3(NOx)
 Min.   :-200.0   Min.   :-200.000   Min.   :-200.0   Min.   :-200.0   Min.
:-200.0
```



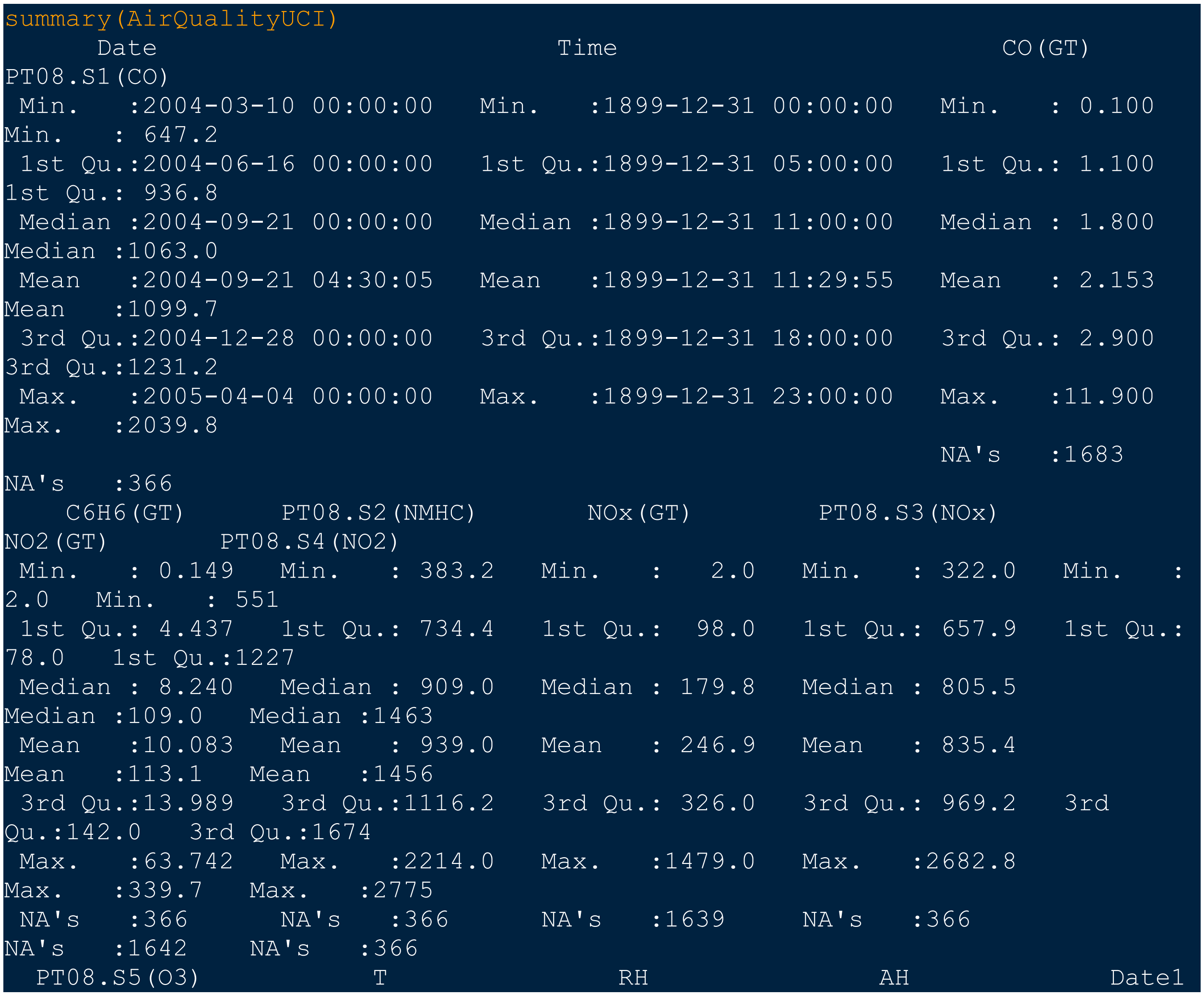
```
 1st Qu.:-200.0   1st Qu.:   4.005   1st Qu.: 711.0   1st Qu.:  50.0   1st
Qu.: 637.0
 Median :-200.0   Median :   7.887   Median : 894.5   Median : 141.0   Median
: 794.2
 Mean   :-159.1   Mean   :   1.866   Mean   : 894.5   Mean   : 168.6   Mean
: 794.9
 3rd Qu.:-200.0   3rd Qu.:  13.636   3rd Qu.:1104.8   3rd Qu.: 284.2   3rd
Qu.: 960.2
 Max.   :1189.0   Max.   :  63.741   Max.   :2214.0   Max.   :1479.0   Max.
:2682.8
    NO2(GT)          PT08.S4(NO2)    PT08.S5(O3)            T
RH
 Min.   :-200.00   Min.   :-200   Min.   :-200.0   Min.   :-200.000
Min.   :-200.00
 1st Qu.:  53.00   1st Qu.:1185   1st Qu.: 699.8   1st Qu.:  10.950   1st
Qu.:  34.05
 Median :  96.00   Median :1446   Median : 942.0   Median :  17.200
Median :  48.55
 Mean   :  58.14   Mean   :1391   Mean   : 975.0   Mean   :   9.777
Mean   :  39.48
 3rd Qu.: 133.00   3rd Qu.:1662   3rd Qu.:1255.2   3rd Qu.:  24.075   3rd
Qu.:  61.88
 Max.   : 339.70   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
```



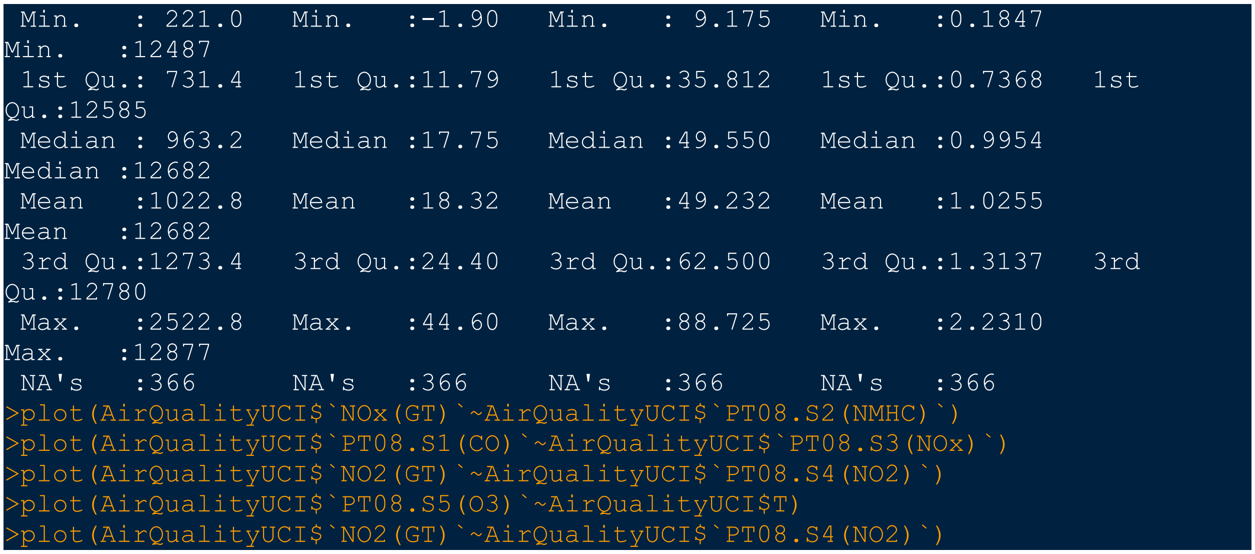
```
>aggr(AirQualityUCI, col=c('pink','yellow'),
+      numbers=TRUE, sortVars=TRUE,
+      labels=names(AirQualityUCI), cex.axis=.7,
+      gap=3, ylab=c("Missing data","Pattern"))    # graphical presentation
of NAs

 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
Warning message:
```



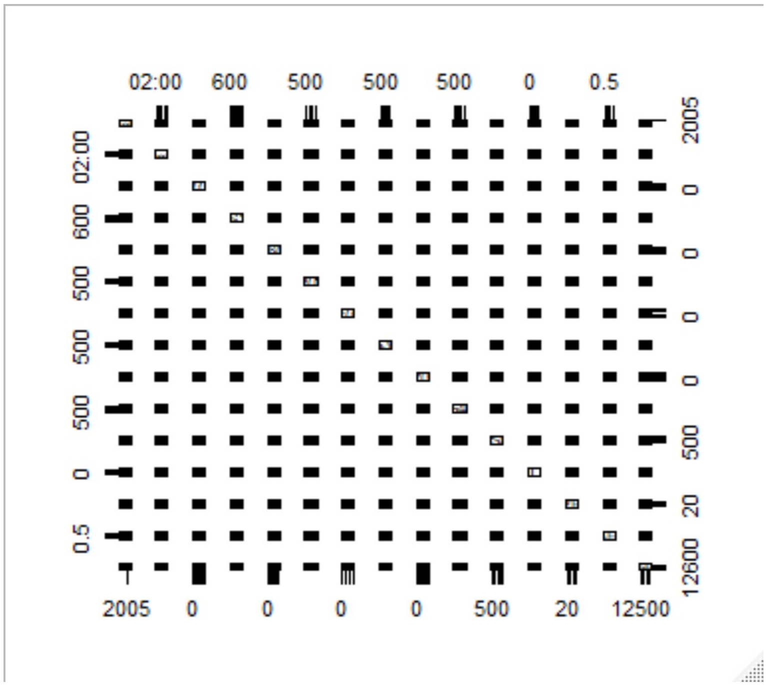
```
In plot.aggr(res, ...) : not enough horizontal space to display frequencies
>sapply(AirQualityUCI, function(x) sum(is.na(x)))     # count of NAs
         Date            Time         CO(GT)    PT08.S1(CO)       NMHC(GT)
C6H6(GT) PT08.S2(NMHC)
            0               0           1683            366           8443
366           366
      NOx(GT)   PT08.S3(NOx)        NO2(GT)   PT08.S4(NO2)    PT08.S5(O3)
T             RH
         1639            366           1642            366            366
366           366
           AH
          366
>AirQualityUCI$`NMHC(GT)` <- NULL
>names(AirQualityUCI)
 [1] "Date"           "Time"           "CO(GT)"         "PT08.S1(CO)"
"C6H6(GT)"      "PT08.S2(NMHC)"
 [7] "NOx(GT)"        "PT08.S3(NOx)"   "NO2(GT)"        "PT08.S4(NO2)"
"PT08.S5(O3)"   "T"
[13] "RH"             "AH"
>AirQualityUCI$Date1 <- as.numeric(as.Date(AirQualityUCI$Date))
>install.packages("mice")
```

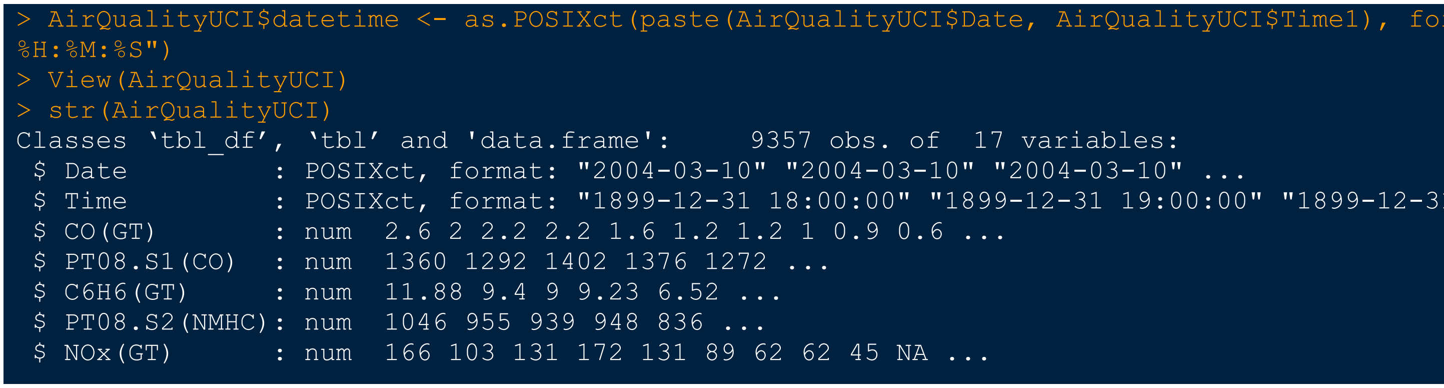


```
summary(AirQualityUCI)
      Date                              Time                              CO(GT)
PT08.S1(CO)
 Min.   :2004-03-10 00:00:00   Min.   :1899-12-31 00:00:00   Min.   : 0.100
Min.   : 647.2
 1st Qu.:2004-06-16 00:00:00   1st Qu.:1899-12-31 05:00:00   1st Qu.: 1.100
1st Qu.: 936.8
 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
    C6H6(GT)       PT08.S2(NMHC)       NOx(GT)          PT08.S3(NOx)
NO2(GT)        PT08.S4(NO2)
 Min.   : 0.149   Min.   : 383.2   Min.   :   2.0   Min.   : 322.0   Min.   :
2.0    Min.   : 551
 1st Qu.: 4.437   1st Qu.: 734.4   1st Qu.:  98.0   1st Qu.: 657.9   1st Qu.:
78.0    1st Qu.:1227
 Median : 8.240   Median : 909.0   Median : 179.8   Median : 805.5
Median :109.0    Median :1463
 Mean   :10.083   Mean   : 939.0   Mean   : 246.9   Mean   : 835.4
Mean   :113.1    Mean   :1456
 3rd Qu.:13.989   3rd Qu.:1116.2   3rd Qu.: 326.0   3rd Qu.: 969.2   3rd
Qu.:142.0    3rd Qu.:1674
 Max.   :63.742   Max.   :2214.0   Max.   :1479.0   Max.   :2682.8
Max.   :339.7    Max.   :2775
 NA's   :366      NA's   :366      NA's   :1639     NA's   :366
NA's   :1642     NA's   :366
  PT08.S5(O3)           T                RH                AH              Date1
```

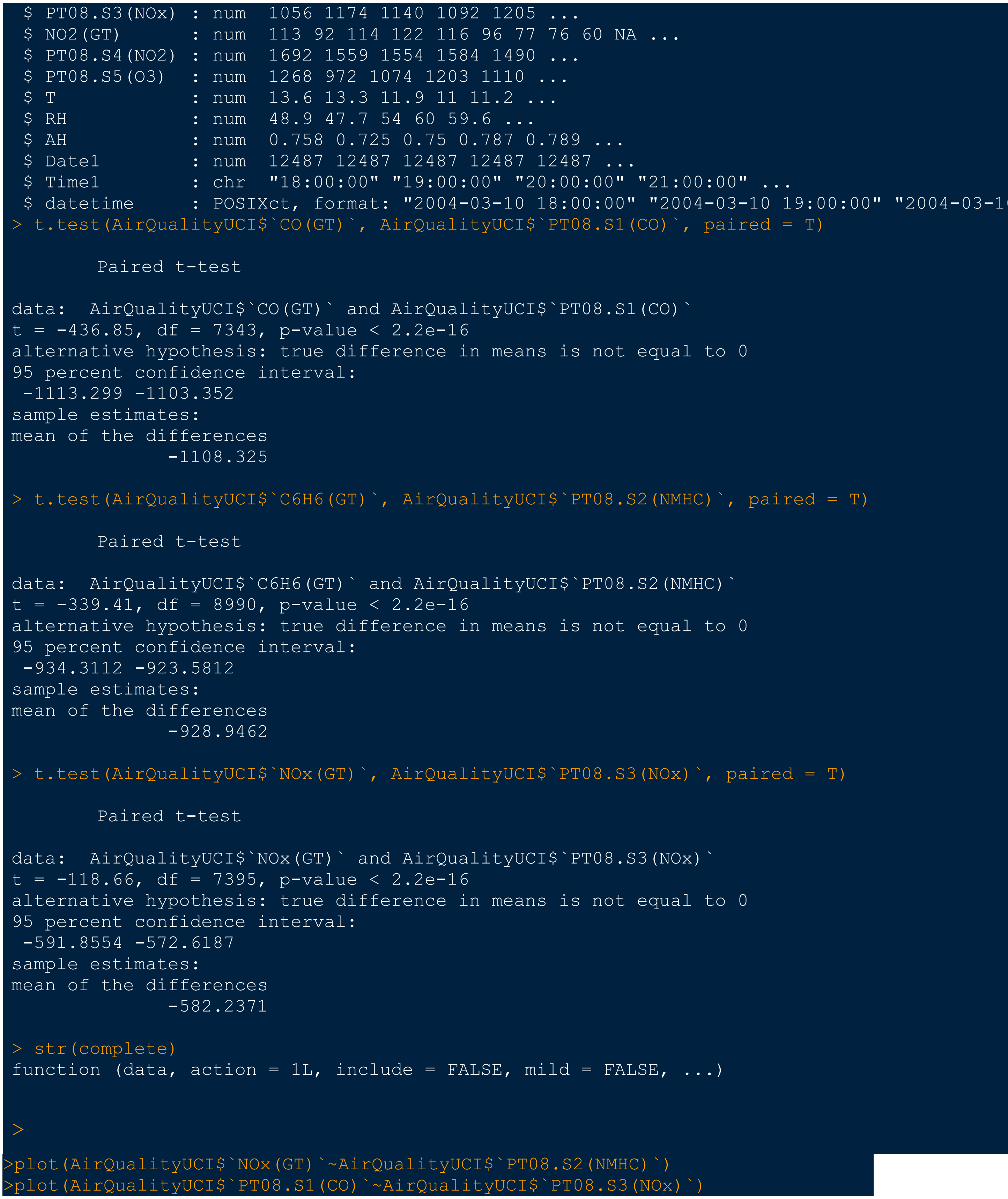


```
 Min.   : 221.0   Min.   :-1.90   Min.   : 9.175   Min.   :0.1847   
Min.   :12487  
 1st Qu.: 731.4   1st Qu.:11.79   1st Qu.:35.812   1st Qu.:0.7368   1st 
Qu.:12585  
 Median : 963.2   Median :17.75   Median :49.550   Median :0.9954   
Median :12682  
 Mean   :1022.8   Mean   :18.32   Mean   :49.232   Mean   :1.0255   
Mean   :12682  
 3rd Qu.:1273.4   3rd Qu.:24.40   3rd Qu.:62.500   3rd Qu.:1.3137   3rd 
Qu.:12780  
 Max.   :2522.8   Max.   :44.60   Max.   :88.725   Max.   :2.2310   
Max.   :12877  
 NA's   :366      NA's   :366     NA's   :366      NA's   :366      
>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)
>plot(AirQualityUCI$`NO2(GT)`~AirQualityUCI$`PT08.S4(NO2)`)
```

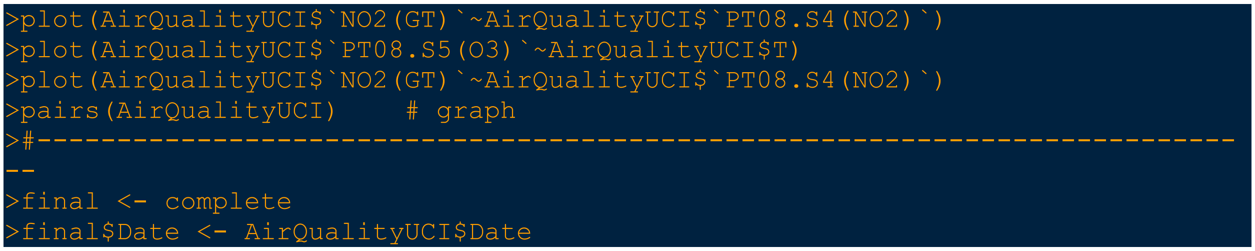




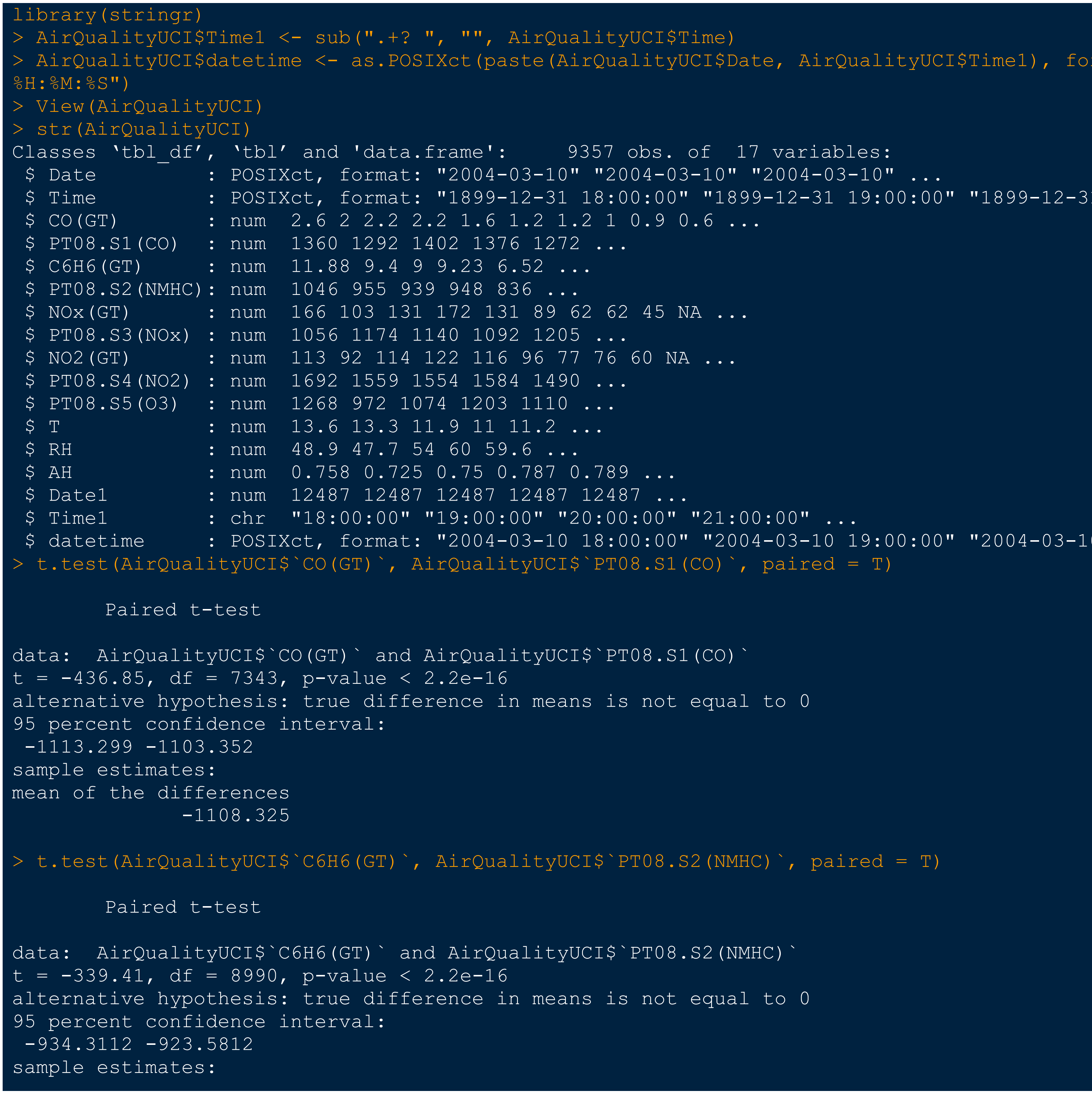
```
 > AirQualityUCI$datetime <- as.POSIXct(paste(AirQualityUCI$Date, AirQualityUCI$Time1), fo
 %H:%M:%S")
 > View(AirQualityUCI)
 > str(AirQualityUCI)
 Classes ‘tbl_df’, ‘tbl’ and 'data.frame':     9357 obs. of  17 variables:
  $ Date         : POSIXct, format: "2004-03-10" "2004-03-10" "2004-03-10" ...
  $ Time         : POSIXct, format: "1899-12-31 18:00:00" "1899-12-31 19:00:00" "1899-12-3
  $ 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 ...
```



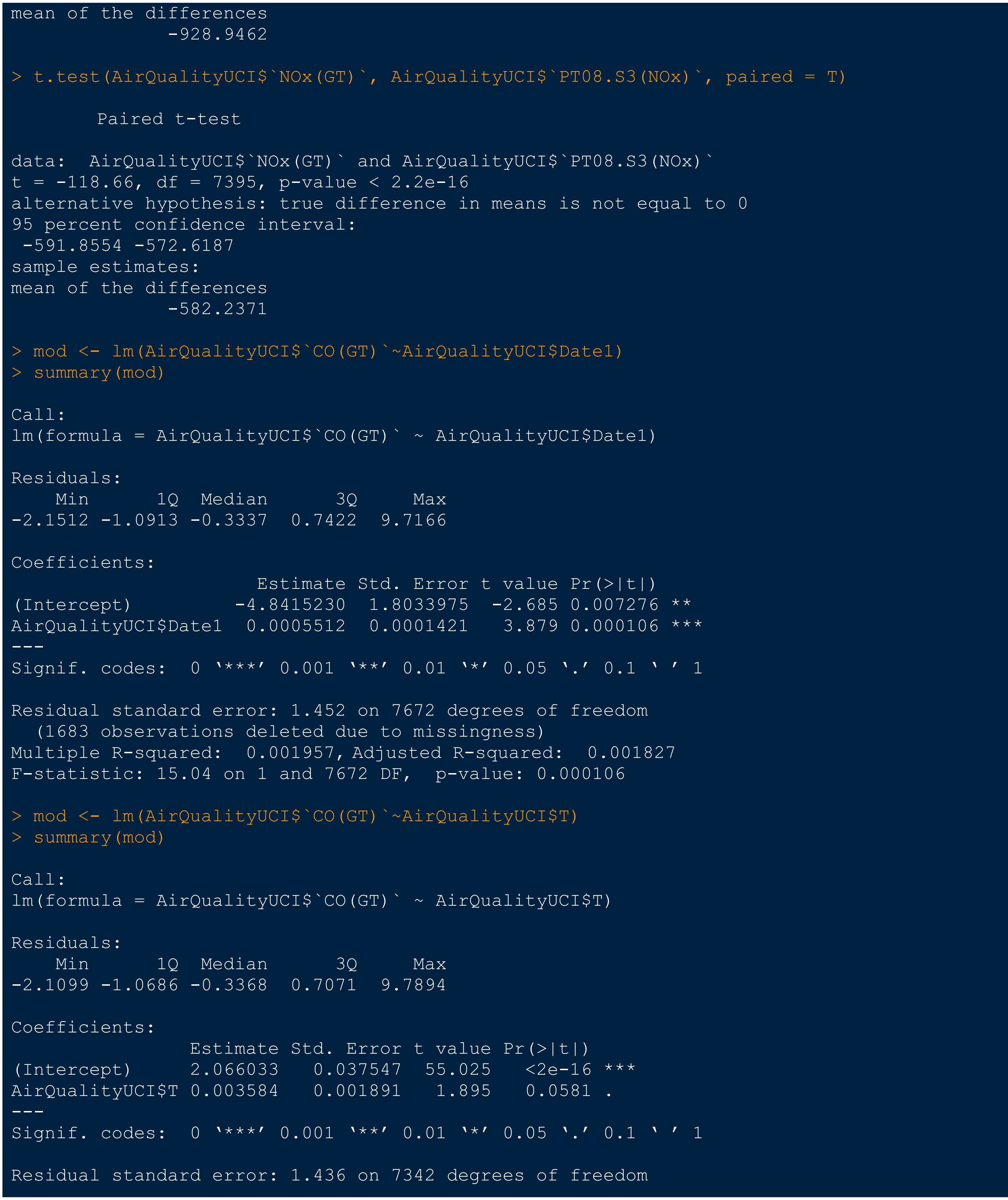
```
 $ PT08.S3(NOx) : num  1056 1174 1140 1092 1205 ...
 $ NO2(GT)      : num  113 92 114 122 116 96 77 76 60 NA ...
 $ PT08.S4(NO2) : num  1692 1559 1554 1584 1490 ...
 $ PT08.S5(O3)  : num  1268 972 1074 1203 1110 ...
 $ T            : num  13.6 13.3 11.9 11 11.2 ...
 $ RH           : num  48.9 47.7 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" "2004-03-1
> 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

> str(complete)
function (data, action = 1L, include = FALSE, mild = FALSE, ...)


>
>plot(AirQualityUCI$`NOx(GT)`~AirQualityUCI$`PT08.S2(NMHC)`)
>plot(AirQualityUCI$`PT08.S1(CO)`~AirQualityUCI$`PT08.S3(NOx)`)
```



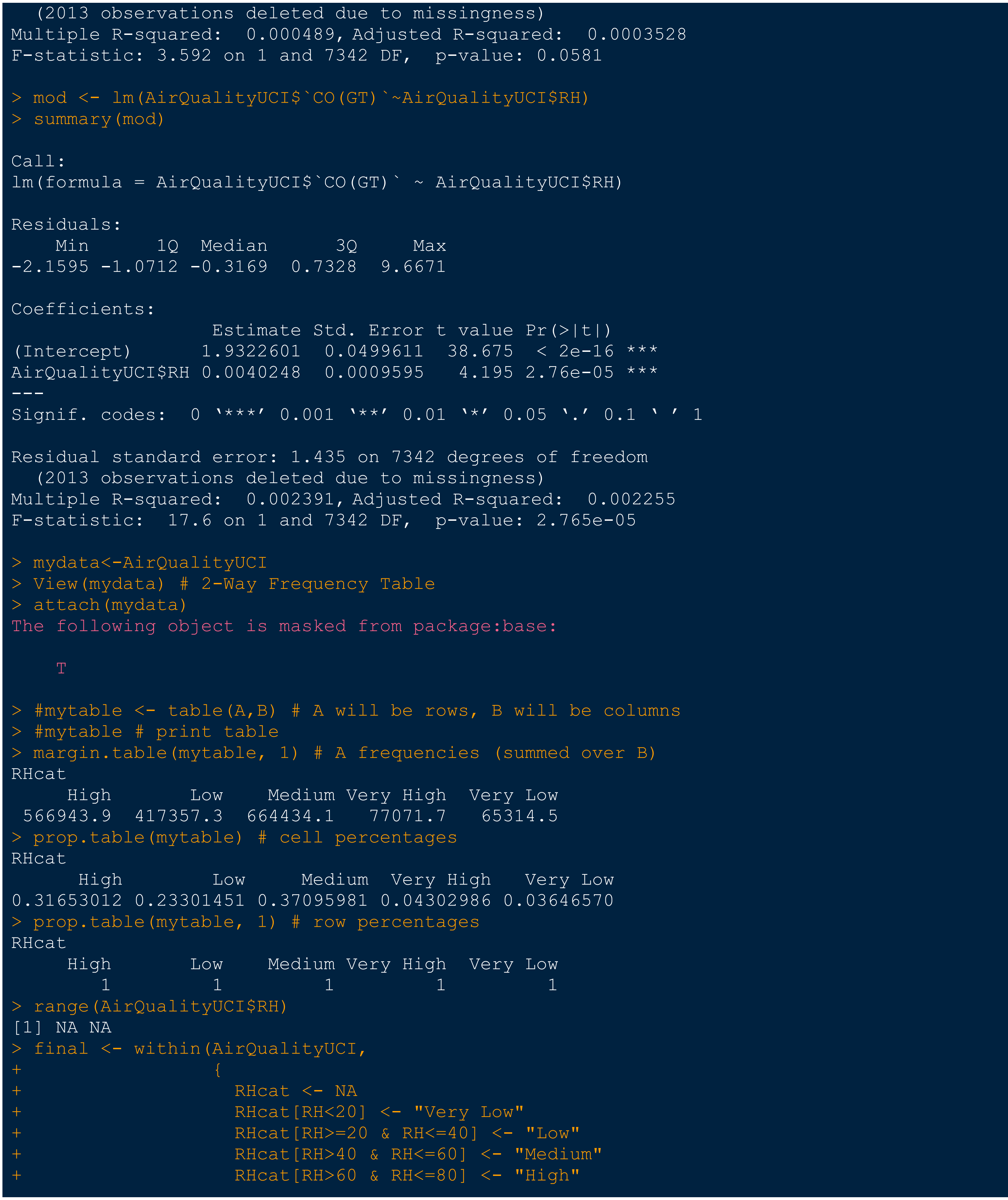
```
>plot(AirQualityUCI$`NO2(GT)`~AirQualityUCI$`PT08.S4(NO2)`)
>plot(AirQualityUCI$`PT08.S5(O3)`~AirQualityUCI$T)
>plot(AirQualityUCI$`NO2(GT)`~AirQualityUCI$`PT08.S4(NO2)`)
>pairs(AirQualityUCI)    # graph
>#----------------------------------------------------------------------------
--
>final <- complete
>final$Date <- AirQualityUCI$Date
```



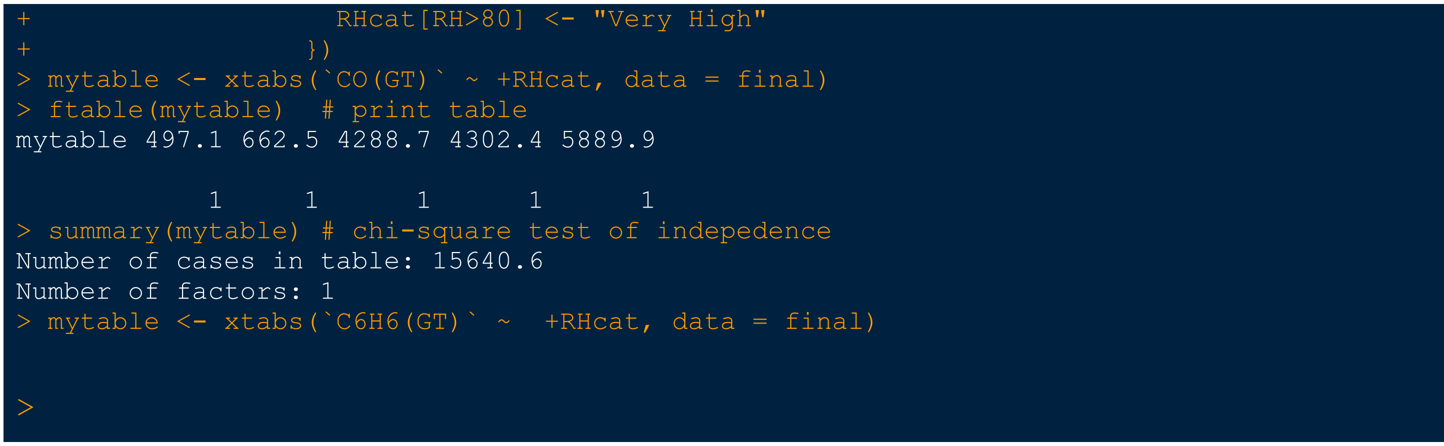
```
library(stringr)
> AirQualityUCI$Time1 <- sub(".+? ", "", AirQualityUCI$Time)
> AirQualityUCI$datetime <- as.POSIXct(paste(AirQualityUCI$Date, AirQualityUCI$Time1), fo
%H:%M:%S")
> View(AirQualityUCI)
> str(AirQualityUCI)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':     9357 obs. of  17 variables:
 $ 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" "1899-12-3
 $ 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" "2004-03-1
> 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:
```



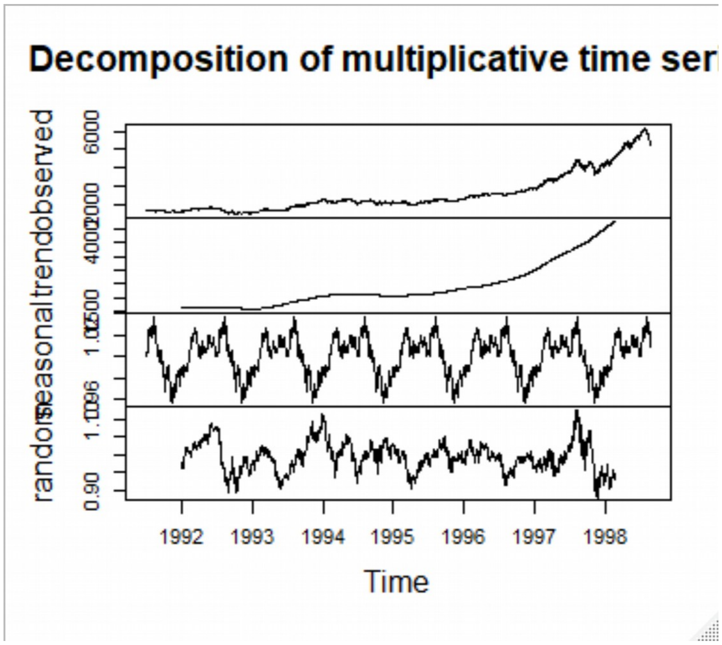
```
mean of the differences 
              -928.9462 

> 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
```

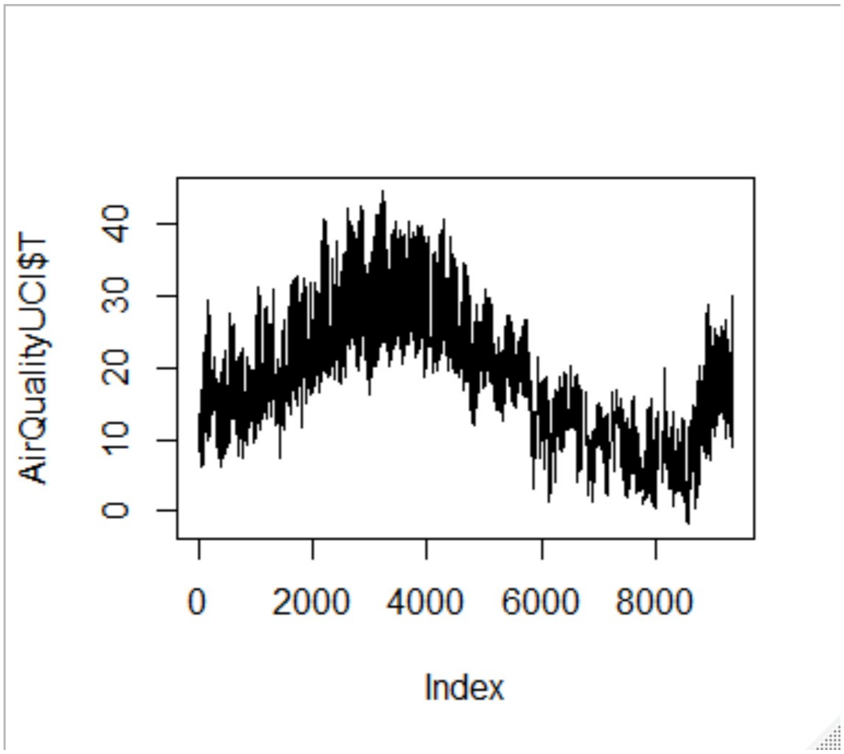


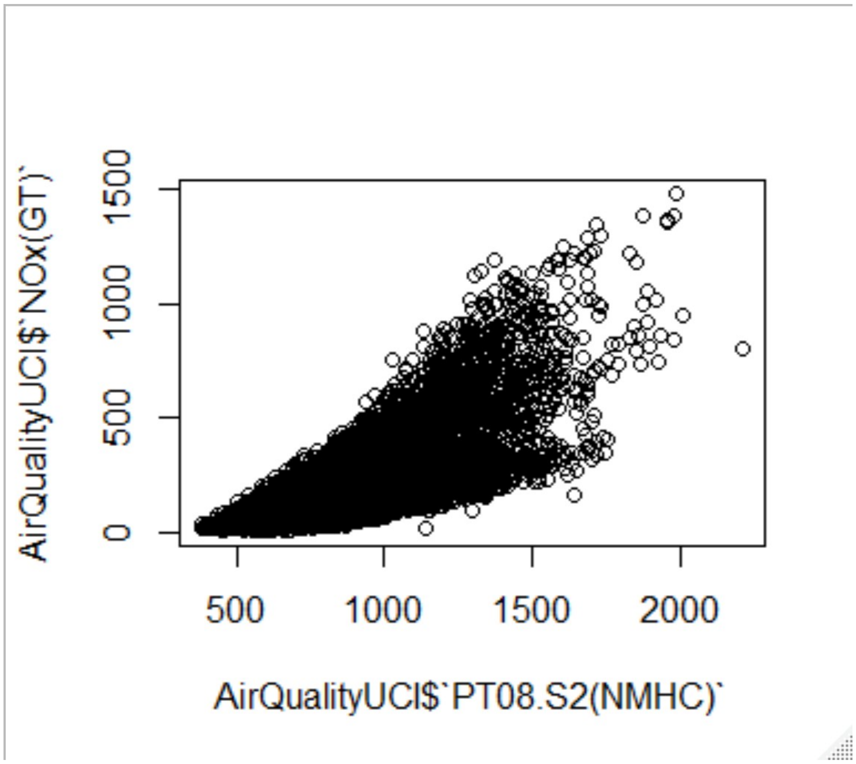
```
  (2013 observations deleted due to missingness)
Multiple R-squared:  0.000489,	Adjusted R-squared:  0.0003528 
F-statistic: 3.592 on 1 and 7342 DF,  p-value: 0.0581

> 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

> mydata<-AirQualityUCI
> View(mydata) # 2-Way Frequency Table
> attach(mydata)
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"
```

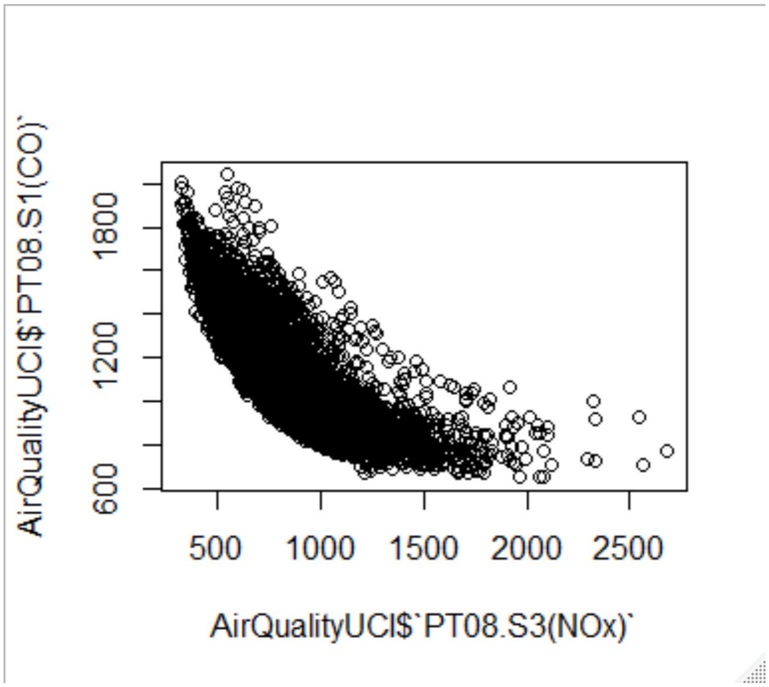


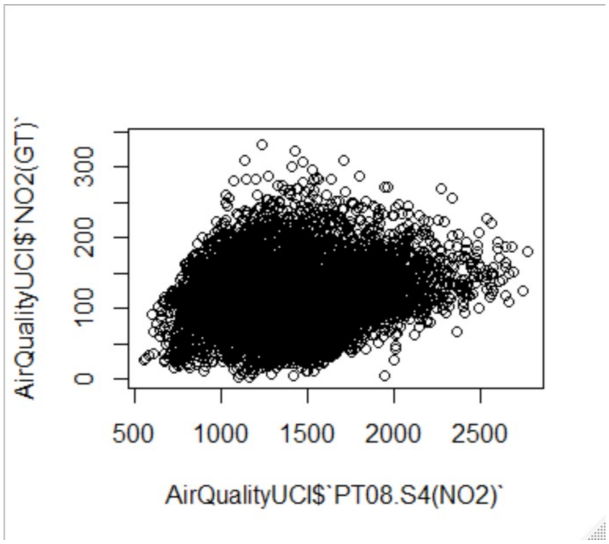
```
+                    RHcat[RH>80] <- "Very High"
+                  })
> mytable <- xtabs(`CO(GT)` ~ +RHcat, data = final)
> ftable(mytable)   # print table
mytable 497.1 662.5 4288.7 4302.4 5889.9

            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)


>
```

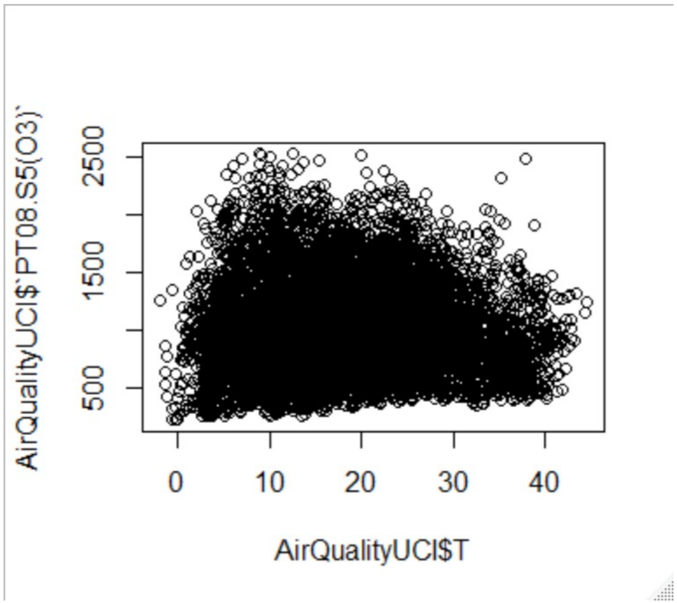



AirQualityUCI$T
0
10
20
30
40
0
2000
4000
6000
8000
Index

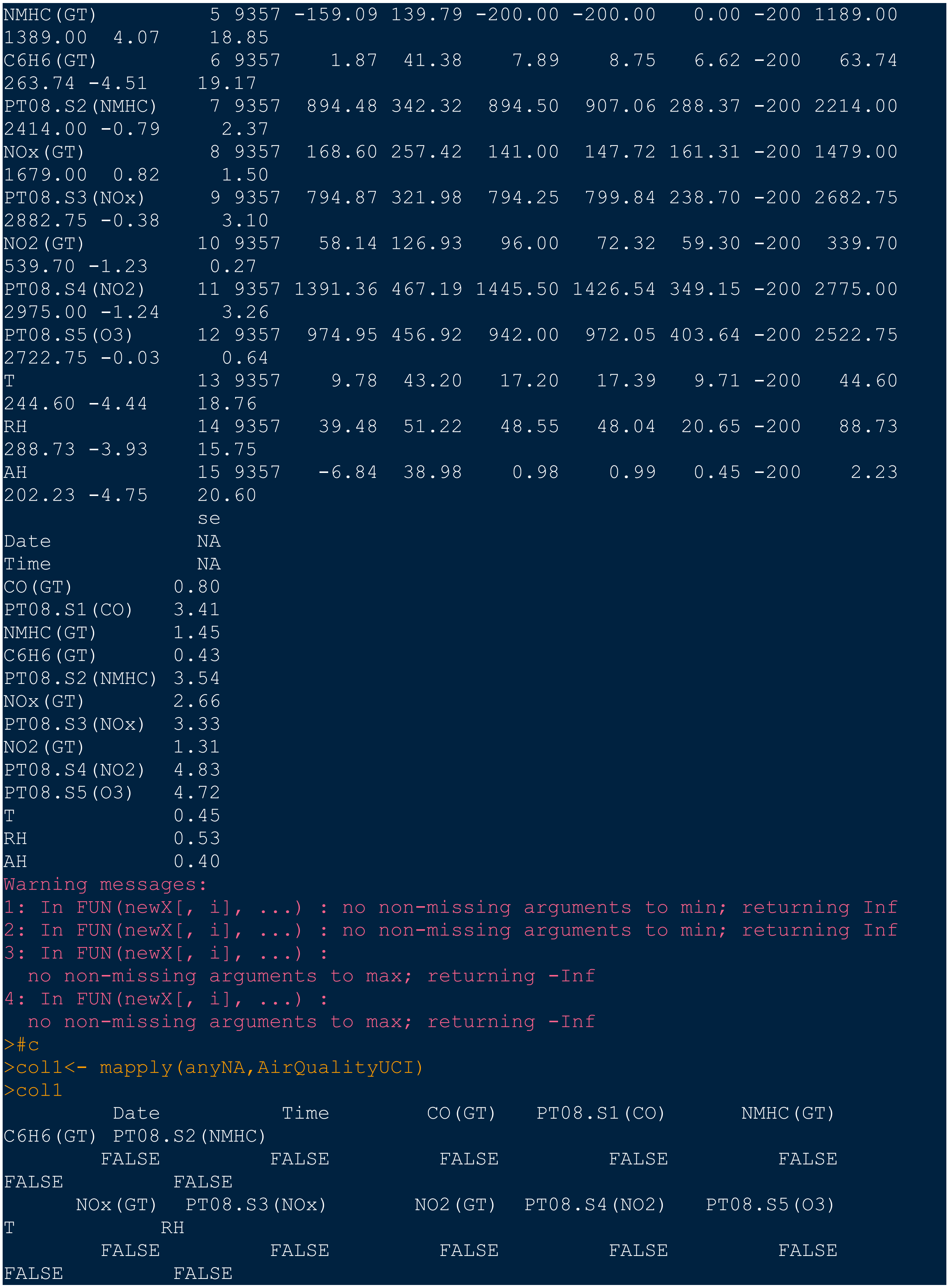

AirQualityUCI$`NOx(GT)`
0
500
1000
1500
500
1000
1500
2000
AirQualityUCI$`PT08.S2(NMHC)`


AirQualityUCI$`PT08.S1(CO)`
600
1200
1800
500
1000
1500
2000
2500
AirQualityUCI$`PT08.S3(NOx)`

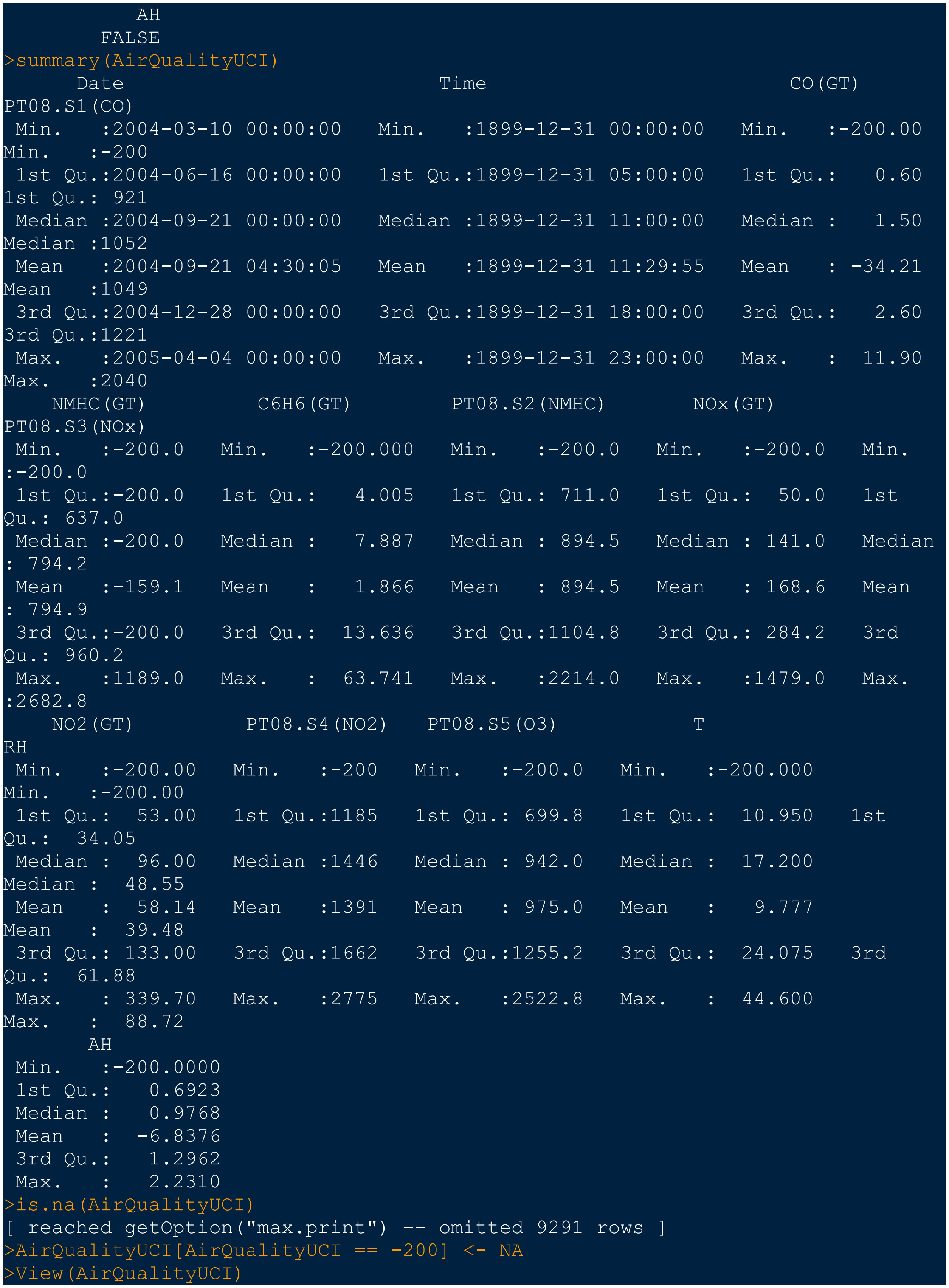

AirQualityUCI$`NO2(GT)`
0
100
200
300
500
1000
1500
2000
2500
AirQualityUCI$`PT08.S4(NO2)`



```
>library(readxl)
>AirQualityUCI <-
read_excel("C:/Users/Jagannath/Downloads/AirQualityUCI.xlsx")
>View(AirQualityUCI)
>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" ...
 $ 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 ...
 $ 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 ...
>#b
>library(psych)
>describe(AirQualityUCI)
              vars    n    mean     sd  median trimmed    mad  min     max
range  skew kurtosis
Date             1 9357     NaN     NA      NA     NaN     NA  Inf    -Inf
-Inf    NA       NA
Time             2 9357     NaN     NA      NA     NaN     NA  Inf    -Inf
-Inf    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
PT08.S1(CO)      4 9357 1048.87 329.82 1052.50 1069.72 218.19 -200 2039.75
2239.75 -1.72     5.83
```

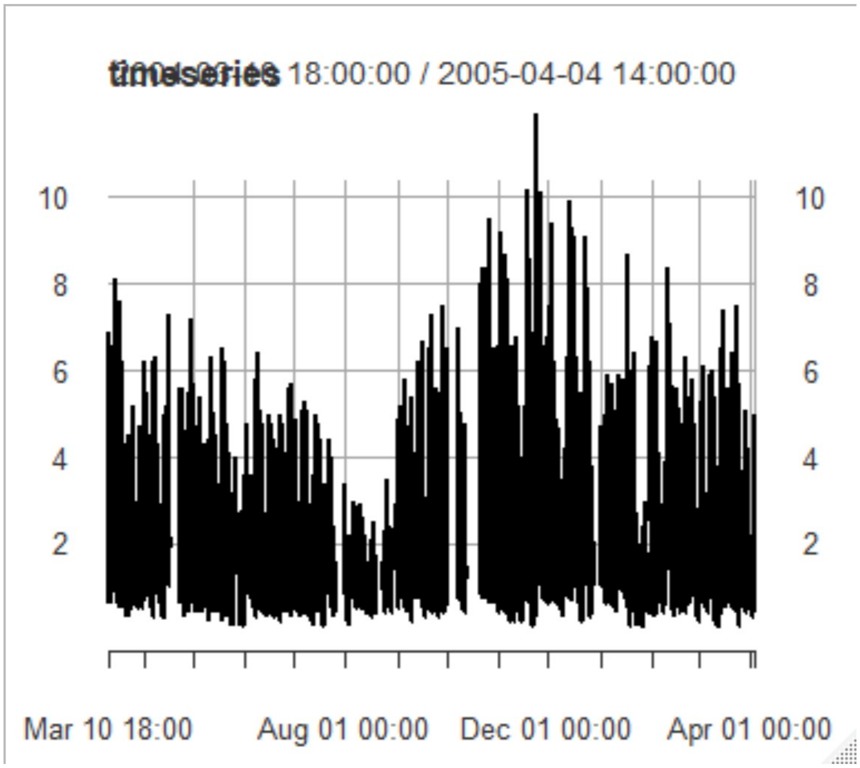


```
NMHC(GT)         5 9357 -159.09 139.79 -200.00 -200.00   0.00 -200 1189.00
1389.00  4.07    18.85
C6H6(GT)         6 9357    1.87  41.38    7.89    8.75   6.62 -200   63.74
263.74 -4.51    19.17
PT08.S2(NMHC)    7 9357  894.48 342.32  894.50  907.06 288.37 -200 2214.00
2414.00 -0.79     2.37
NOx(GT)          8 9357  168.60 257.42  141.00  147.72 161.31 -200 1479.00
1679.00  0.82     1.50
PT08.S3(NOx)     9 9357  794.87 321.98  794.25  799.84 238.70 -200 2682.75
2882.75 -0.38     3.10
NO2(GT)         10 9357   58.14 126.93   96.00   72.32  59.30 -200  339.70
539.70 -1.23      0.27
PT08.S4(NO2)    11 9357 1391.36 467.19 1445.50 1426.54 349.15 -200 2775.00
2975.00 -1.24     3.26
PT08.S5(O3)     12 9357  974.95 456.92  942.00  972.05 403.64 -200 2522.75
2722.75 -0.03     0.64
T               13 9357    9.78  43.20   17.20   17.39   9.71 -200   44.60
244.60 -4.44    18.76
RH              14 9357   39.48  51.22   48.55   48.04  20.65 -200   88.73
288.73 -3.93    15.75
AH              15 9357   -6.84  38.98    0.98    0.99   0.45 -200    2.23
202.23 -4.75    20.60
                se
Date            NA
Time            NA
CO(GT)        0.80
PT08.S1(CO)   3.41
NMHC(GT)      1.45
C6H6(GT)      0.43
PT08.S2(NMHC) 3.54
NOx(GT)       2.66
PT08.S3(NOx)  3.33
NO2(GT)       1.31
PT08.S4(NO2)  4.83
PT08.S5(O3)   4.72
T             0.45
RH            0.53
AH            0.40
Warning messages:
1: In FUN(newX[, i], ...) : no non-missing arguments to min; returning Inf
2: In FUN(newX[, i], ...) : no non-missing arguments to min; returning Inf
3: In FUN(newX[, i], ...) :
  no non-missing arguments to max; returning -Inf
4: In FUN(newX[, i], ...) :
  no non-missing arguments to max; returning -Inf
>#c
>col1<- mapply(anyNA,AirQualityUCI)
>col1
         Date          Time       CO(GT)    PT08.S1(CO)      NMHC(GT)
C6H6(GT) PT08.S2(NMHC)
        FALSE         FALSE         FALSE         FALSE         FALSE
FALSE         FALSE
      NOx(GT)  PT08.S3(NOx)       NO2(GT)  PT08.S4(NO2)   PT08.S5(O3)
T            RH
        FALSE         FALSE         FALSE         FALSE         FALSE
FALSE         FALSE
```



```
           AH
        FALSE
>summary(AirQualityUCI)
      Date                            Time                            CO(GT)
PT08.S1(CO)
 Min.   :2004-03-10 00:00:00   Min.   :1899-12-31 00:00:00   Min.   :-200.00
Min.   :-200
 1st Qu.:2004-06-16 00:00:00   1st Qu.:1899-12-31 05:00:00   1st Qu.:   0.60
1st Qu.: 921
 Median :2004-09-21 00:00:00   Median :1899-12-31 11:00:00   Median :   1.50
Median :1052
 Mean   :2004-09-21 04:30:05   Mean   :1899-12-31 11:29:55   Mean   : -34.21
Mean   :1049
 3rd Qu.:2004-12-28 00:00:00   3rd Qu.:1899-12-31 18:00:00   3rd Qu.:   2.60
3rd Qu.:1221
 Max.   :2005-04-04 00:00:00   Max.   :1899-12-31 23:00:00   Max.   :  11.90
Max.   :2040
    NMHC(GT)          C6H6(GT)         PT08.S2(NMHC)       NOx(GT)
PT08.S3(NOx)
 Min.   :-200.0   Min.   :-200.000   Min.   :-200.0   Min.   :-200.0   Min.
:-200.0
 1st Qu.:-200.0   1st Qu.:   4.005   1st Qu.: 711.0   1st Qu.:  50.0   1st
Qu.: 637.0
 Median :-200.0   Median :   7.887   Median : 894.5   Median : 141.0   Median
: 794.2
 Mean   :-159.1   Mean   :   1.866   Mean   : 894.5   Mean   : 168.6   Mean
: 794.9
 3rd Qu.:-200.0   3rd Qu.:  13.636   3rd Qu.:1104.8   3rd Qu.: 284.2   3rd
Qu.: 960.2
 Max.   :1189.0   Max.   :  63.741   Max.   :2214.0   Max.   :1479.0   Max.
:2682.8
    NO2(GT)          PT08.S4(NO2)   PT08.S5(O3)            T
RH
 Min.   :-200.00   Min.   :-200   Min.   :-200.0   Min.   :-200.000
Min.   :-200.00
 1st Qu.:  53.00   1st Qu.:1185   1st Qu.: 699.8   1st Qu.:  10.950   1st
Qu.:  34.05
 Median :  96.00   Median :1446   Median : 942.0   Median :  17.200
Median :  48.55
 Mean   :  58.14   Mean   :1391   Mean   : 975.0   Mean   :   9.777
Mean   :  39.48
 3rd Qu.: 133.00   3rd Qu.:1662   3rd Qu.:1255.2   3rd Qu.:  24.075   3rd
Qu.:  61.88
 Max.   : 339.70   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
>is.na(AirQualityUCI)
[ reached getOption("max.print") -- omitted 9291 rows ]
>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

 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
Warning message:
In plot.aggr(res, ...) : not enough horizontal space to display frequencies
>sapply(AirQualityUCI, function(x) sum(is.na(x)))     # count of NAs
         Date           Time         CO(GT)    PT08.S1(CO)       NMHC(GT)
C6H6(GT) PT08.S2(NMHC)
            0              0           1683            366           8443
366           366
      NOx(GT)   PT08.S3(NOx)        NO2(GT)   PT08.S4(NO2)    PT08.S5(O3)
T            RH
         1639            366           1642            366            366
366           366
           AH
          366
>AirQualityUCI$`NMHC(GT)` <- NULL
>names(AirQualityUCI)
 [1] "Date"          "Time"          "CO(GT)"        "PT08.S1(CO)"
"C6H6(GT)"      "PT08.S2(NMHC)"
 [7] "NOx(GT)"       "PT08.S3(NOx)"  "NO2(GT)"       "PT08.S4(NO2)"
"PT08.S5(O3)"   "T"
[13] "RH"            "AH"
>AirQualityUCI$Date1 <- as.numeric(as.Date(AirQualityUCI$Date))
>install.packages("mice")
Error in install.packages : Updating loaded packages

Restarting R session...

Loading required package: arules
Loading required package: Matrix

Attaching package: ‘arules’

The following objects are masked from ‘package:base’:
    abbreviate, write
>imputed <- mice(AirQualityUCI[,-c(1,2,4)], m=5, maxit = 5, method = 'cart',
seed = 100) # impute missing values

### time series not covered in syllabus

 iter imp variable
  1    1   CO(GT)
```



…………………………………………… END …………………………..