AirQualityUCI

Natalia

31 May 2019

```
data <- read.csv("C://Users//Natalia//Desktop//ITMO//R//CaseStudy#2models//AirQualityUCI.csv", header =
head(data)
```

```
##         Date     Time CO.GT. PT08.S1.CO. NMHC.GT. C6H6.GT. PT08.S2.NMHC.
## 1 10/03/2004 18.00.00    2,6        1360      150     11,9          1046
## 2 10/03/2004 19.00.00      2        1292      112      9,4           955
## 3 10/03/2004 20.00.00    2,2        1402       88      9,0           939
## 4 10/03/2004 21.00.00    2,2        1376       80      9,2           948
## 5 10/03/2004 22.00.00    1,6        1272       51      6,5           836
## 6 10/03/2004 23.00.00    1,2        1197       38      4,7           750
##   NOx.GT. PT08.S3.NOx. NO2.GT. PT08.S4.NO2. PT08.S5.O3.    T   RH     AH
## 1     166         1056     113         1692        1268 13,6 48,9 0,7578
## 2     103         1174      92         1559         972 13,3 47,7 0,7255
## 3     131         1140     114         1555        1074 11,9 54,0 0,7502
## 4     172         1092     122         1584        1203 11,0 60,0 0,7867
## 5     131         1205     116         1490        1110 11,2 59,6 0,7888
## 6      89         1337      96         1393         949 11,2 59,2 0,7848
##    X X.1
## 1 NA  NA
## 2 NA  NA
## 3 NA  NA
## 4 NA  NA
## 5 NA  NA
## 6 NA  NA
```

```
str(data)
```

```
## 'data.frame':    9471 obs. of  17 variables:
##  $ Date         : Factor w/ 392 levels "","01/01/2005",..: 116 116 116 116 116 116 129 129 129 129 .
##  $ Time         : Factor w/ 25 levels "","00.00.00",..: 20 21 22 23 24 25 2 3 4 5 ...
##  $ CO.GT.       : Factor w/ 105 levels "","-200","-200,0",..: 35 28 31 31 20 16 16 13 12 9 ...
##  $ PT08.S1.CO.  : int  1360 1292 1402 1376 1272 1197 1185 1136 1094 1010 ...
##  $ NMHC.GT.     : int  150 112 88 80 51 38 31 31 24 19 ...
##  $ C6H6.GT.     : Factor w/ 409 levels "","-200,0","0,1",..: 41 404 400 402 374 327 237 234 125 19 .
##  $ PT08.S2.NMHC.: int  1046 955 939 948 836 750 690 672 609 561 ...
##  $ NOx.GT.      : int  166 103 131 172 131 89 62 62 45 -200 ...
##  $ PT08.S3.NOx. : int  1056 1174 1140 1092 1205 1337 1462 1453 1579 1705 ...
##  $ NO2.GT.      : int  113 92 114 122 116 96 77 76 60 -200 ...
##  $ PT08.S4.NO2. : int  1692 1559 1555 1584 1490 1393 1333 1333 1276 1235 ...
##  $ PT08.S5.O3.  : int  1268 972 1074 1203 1110 949 733 730 620 501 ...
##  $ T            : Factor w/ 438 levels "","-0,1","-0,2",..: 68 65 51 42 44 44 45 39 39 35 ...
##  $ RH           : Factor w/ 755 levels "","-200","10,0",..: 377 365 428 488 484 480 456 488 485 490
##  $ AH           : Factor w/ 6685 levels "","-200","0,1847",..: 1898 1729 1855 2058 2068 2047 1911 19
##  $ X            : logi  NA NA NA NA NA NA ...
##  $ X.1          : logi  NA NA NA NA NA NA ...
```

```
summary(data)
```

```
##          Date            Time          CO.GT.      PT08.S1.CO.
```

```
##            : 114   00.00.00: 390   -200   :1592   Min.   :-200
##  01/01/2005:  24   01.00.00: 390   1,4    : 279   1st Qu.: 921
##  01/02/2005:  24   02.00.00: 390   1,6    : 275   Median :1053
##  01/03/2005:  24   03.00.00: 390   1,5    : 273   Mean   :1049
##  01/04/2004:  24   04.00.00: 390   1,1    : 262   3rd Qu.:1221
##  01/04/2005:  24   05.00.00: 390   0,7    : 260   Max.   :2040
##  (Other)   :9237   (Other) :7131   (Other):6530   NA's   :114
##     NMHC.GT.          C6H6.GT.    PT08.S2.NMHC.       NOx.GT.
##  Min.   :-200.0   -200,0 : 366   Min.   :-200.0   Min.   :-200.0
##  1st Qu.:-200.0          : 114   1st Qu.: 711.0   1st Qu.:  50.0
##  Median :-200.0   3,6    :  84   Median : 895.0   Median : 141.0
##  Mean   :-159.1   2,8    :  82   Mean   : 894.6   Mean   : 168.6
##  3rd Qu.:-200.0   3,8    :  79   3rd Qu.:1105.0   3rd Qu.: 284.0
##  Max.   :1189.0   4,0    :  78   Max.   :2214.0   Max.   :1479.0
##  NA's   :114      (Other):8668   NA's   :114      NA's   :114
##   PT08.S3.NOx.      NO2.GT.         PT08.S4.NO2.   PT08.S5.O3.
##  Min.   :-200   Min.   :-200.00   Min.   :-200   Min.   :-200.0
##  1st Qu.: 637   1st Qu.:  53.00   1st Qu.:1185   1st Qu.: 700.0
##  Median : 794   Median :  96.00   Median :1446   Median : 942.0
##  Mean   : 795   Mean   :  58.15   Mean   :1391   Mean   : 975.1
##  3rd Qu.: 960   3rd Qu.: 133.00   3rd Qu.:1662   3rd Qu.:1255.0
##  Max.   :2683   Max.   : 340.00   Max.   :2775   Max.   :2523.0
##  NA's   :114    NA's   :114       NA's   :114    NA's   :114
##        T              RH              AH           X
##  -200   : 366   -200   : 366   -200   : 366   Mode:logical
##         : 114          : 114          : 114   NA's:9471
##  20,8   :  57   53,1   :  31   0,7487 :   6
##  21,3   :  54   47,8   :  30   0,8394 :   6
##  13,8   :  51   57,9   :  30   0,9684 :   6
##  20,2   :  51   45,9   :  27   0,9722 :   6
##  (Other):8778   (Other):8873   (Other):8967
##    X.1
##  Mode:logical
##  NA's:9471
##
##
##
##
##
```

```
#Delete empty columns and unusful data:
data$X <- NULL
data$X.1 <- NULL
data$Date <- NULL
data$Time <- NULL
head(data)
```

```
##   CO.GT. PT08.S1.CO. NMHC.GT. C6H6.GT. PT08.S2.NMHC. NOx.GT. PT08.S3.NOx.
## 1    2,6        1360      150     11,9          1046     166         1056
## 2      2        1292      112      9,4           955     103         1174
## 3    2,2        1402       88      9,0           939     131         1140
## 4    2,2        1376       80      9,2           948     172         1092
## 5    1,6        1272       51      6,5           836     131         1205
## 6    1,2        1197       38      4,7           750      89         1337
##   NO2.GT. PT08.S4.NO2. PT08.S5.O3.    T   RH     AH
```

```
## 1     113         1692        1268 13,6 48,9 0,7578
## 2      92         1559         972 13,3 47,7 0,7255
## 3     114         1555        1074 11,9 54,0 0,7502
## 4     122         1584        1203 11,0 60,0 0,7867
## 5     116         1490        1110 11,2 59,6 0,7888
## 6      96         1393         949 11,2 59,2 0,7848
```

```
# convert all variables into numeric:

dt <- data.matrix(data)
dt <- data.frame(dt)
str(dt)
```

```
## 'data.frame':    9471 obs. of  13 variables:
##  $ CO.GT.       : int  35 28 31 31 20 16 16 13 12 9 ...
##  $ PT08.S1.CO.  : int  1360 1292 1402 1376 1272 1197 1185 1136 1094 1010 ...
##  $ NMHC.GT.     : int  150 112 88 80 51 38 31 31 24 19 ...
##  $ C6H6.GT.     : int  41 404 400 402 374 327 237 234 125 19 ...
##  $ PT08.S2.NMHC.: int  1046 955 939 948 836 750 690 672 609 561 ...
##  $ NOx.GT.      : int  166 103 131 172 131 89 62 62 45 -200 ...
##  $ PT08.S3.NOx. : int  1056 1174 1140 1092 1205 1337 1462 1453 1579 1705 ...
##  $ NO2.GT.      : int  113 92 114 122 116 96 77 76 60 -200 ...
##  $ PT08.S4.NO2. : int  1692 1559 1555 1584 1490 1393 1333 1333 1276 1235 ...
##  $ PT08.S5.O3.  : int  1268 972 1074 1203 1110 949 733 730 620 501 ...
##  $ T            : int  68 65 51 42 44 44 45 39 39 35 ...
##  $ RH           : int  377 365 428 488 484 480 456 488 485 490 ...
##  $ AH           : int  1898 1729 1855 2058 2068 2047 1911 1964 1937 1865 ...
```

```
summary(dt)
```

```
##      CO.GT.     PT08.S1.CO.      NMHC.GT.         C6H6.GT.
##  Min.   :  1   Min.   :-200   Min.   :-200.0   Min.   :  1.0
##  1st Qu.:  9   1st Qu.: 921   1st Qu.:-200.0   1st Qu.: 58.5
##  Median : 19   Median :1053   Median :-200.0   Median :154.0
##  Mean   : 23   Mean   :1049   Mean   :-159.1   Mean   :197.6
##  3rd Qu.: 35   3rd Qu.:1221   3rd Qu.:-200.0   3rd Qu.:361.0
##  Max.   :105   Max.   :2040   Max.   :1189.0   Max.   :409.0
##                NA's   :114    NA's   :114
##  PT08.S2.NMHC.       NOx.GT.        PT08.S3.NOx.     NO2.GT.
##  Min.   :-200.0   Min.   :-200.0   Min.   :-200   Min.   :-200.00
##  1st Qu.: 711.0   1st Qu.:  50.0   1st Qu.: 637   1st Qu.:  53.00
##  Median : 895.0   Median : 141.0   Median : 794   Median :  96.00
##  Mean   : 894.6   Mean   : 168.6   Mean   : 795   Mean   :  58.15
##  3rd Qu.:1105.0   3rd Qu.: 284.0   3rd Qu.: 960   3rd Qu.: 133.00
##  Max.   :2214.0   Max.   :1479.0   Max.   :2683   Max.   : 340.00
##  NA's   :114      NA's   :114      NA's   :114    NA's   :114
##   PT08.S4.NO2.   PT08.S5.O3.           T               RH
##  Min.   :-200   Min.   :-200.0   Min.   :  1.0   Min.   :  1.0
##  1st Qu.:1185   1st Qu.: 700.0   1st Qu.: 81.0   1st Qu.:223.5
##  Median :1446   Median : 942.0   Median :156.0   Median :370.0
##  Mean   :1391   Mean   : 975.1   Mean   :179.6   Mean   :361.7
##  3rd Qu.:1662   3rd Qu.:1255.0   3rd Qu.:249.0   3rd Qu.:505.0
##  Max.   :2775   Max.   :2523.0   Max.   :438.0   Max.   :755.0
##  NA's   :114    NA's   :114
##        AH
##  Min.   :   1
```

```
##  1st Qu.:1510
##  Median :3115
##  Mean   :3135
##  3rd Qu.:4751
##  Max.   :6685
##
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.1       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
map_int(dt, function(.x) sum(is.na(.x)))
```

```
##        CO.GT.   PT08.S1.CO.      NMHC.GT.      C6H6.GT. PT08.S2.NMHC.
##             0           114           114             0           114
##       NOx.GT.  PT08.S3.NOx.       NO2.GT.  PT08.S4.NO2.   PT08.S5.O3.
##           114           114           114           114           114
##             T            RH            AH
##             0             0             0
```

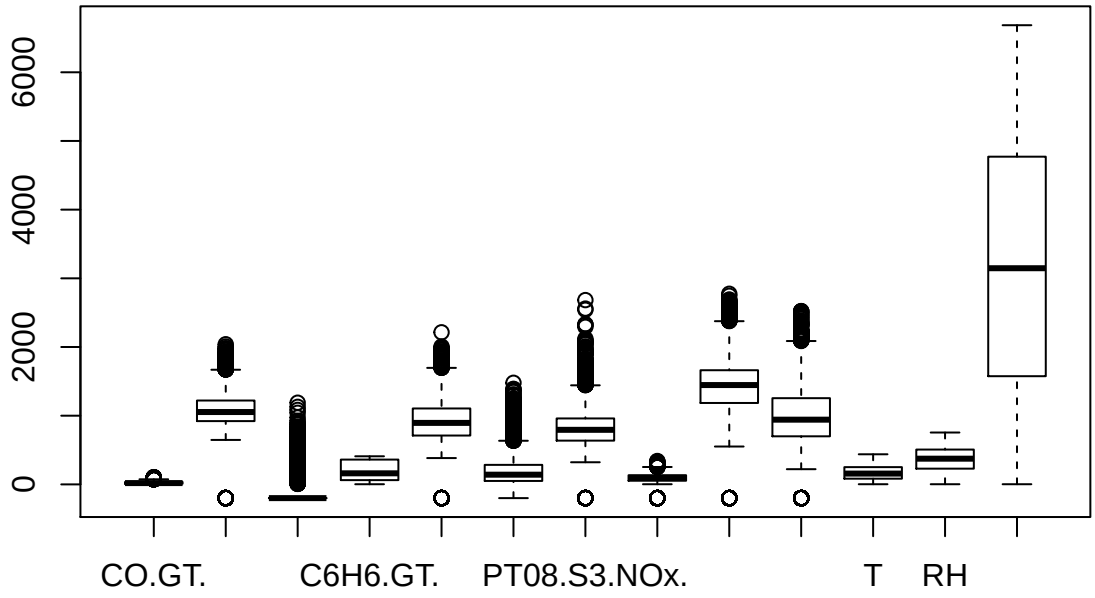
```
# Delete empty rows from the end of the dataframe:

dt <- dt[-c(9358:9471),]
tail(dt)
```

```
##      CO.GT. PT08.S1.CO. NMHC.GT. C6H6.GT. PT08.S2.NMHC. NOx.GT.
## 9352     49        1297     -200       58          1102     523
## 9353     41        1314     -200       57          1101     472
## 9354     33        1163     -200       36          1027     353
## 9355     33        1142     -200       46          1063     293
## 9356     30        1003     -200      405           961     235
## 9357     31        1071     -200       41          1047     265
##      PT08.S3.NOx. NO2.GT. PT08.S4.NO2. PT08.S5.O3.   T  RH   AH
## 9352          507     187         1375        1583 114 251 1849
## 9353          539     190         1374        1729 161 181 1891
## 9354          604     179         1264        1269 185 125 1660
## 9355          603     175         1241        1092 211  71 1360
## 9356          702     156         1041         770 225  24  899
## 9357          654     168         1129         816 227  20  861
```

```
map_int(dt, function(.x) sum(is.na(.x)))
```

```
##        CO.GT.   PT08.S1.CO.      NMHC.GT.      C6H6.GT. PT08.S2.NMHC.
##             0             0             0             0             0
##       NOx.GT.  PT08.S3.NOx.       NO2.GT.  PT08.S4.NO2.   PT08.S5.O3.
##             0             0             0             0             0
##             T            RH            AH
##             0             0             0
```

```r
#let's see each variable:
library(ggplot2)

#plot all numeric variables:
boxplot(dt)
```




```r
#There are some values = "-200" and "-200.0" so lets replace them to NAs:

dt <- na_if(dt, -200)

summary(dt)
```

```
##      CO.GT.         PT08.S1.CO.      NMHC.GT.         C6H6.GT.
##  Min.   :  2.00   Min.   : 647   Min.   :   7.0   Min.   :  2
##  1st Qu.:  9.00   1st Qu.: 937   1st Qu.:  67.0   1st Qu.: 62
##  Median : 19.00   Median :1063   Median : 150.0   Median :161
##  Mean   : 23.27   Mean   :1100   Mean   : 218.8   Mean   :200
##  3rd Qu.: 35.00   3rd Qu.:1231   3rd Qu.: 297.0   3rd Qu.:361
##  Max.   :105.00   Max.   :2040   Max.   :1189.0   Max.   :409
##                   NA's   :366    NA's   :8443
##  PT08.S2.NMHC.       NOx.GT.         PT08.S3.NOx.       NO2.GT.
##  Min.   : 383.0   Min.   :   2.0   Min.   : 322.0   Min.   :  2.0
##  1st Qu.: 734.5   1st Qu.:  98.0   1st Qu.: 658.0   1st Qu.: 78.0
##  Median : 909.0   Median : 180.0   Median : 806.0   Median :109.0
##  Mean   : 939.2   Mean   : 246.9   Mean   : 835.5   Mean   :113.1
##  3rd Qu.:1116.0   3rd Qu.: 326.0   3rd Qu.: 969.5   3rd Qu.:142.0
```

```
##  Max.   :2214.0   Max.   :1479.0   Max.   :2683.0   Max.   :340.0
##  NA's   :366      NA's   :1639     NA's   :366      NA's   :1642
##   PT08.S4.NO2.   PT08.S5.O3.           T               RH
##  Min.   : 551   Min.   : 221.0   Min.   :  2.0   Min.   :  2.0
##  1st Qu.:1227   1st Qu.: 731.5   1st Qu.: 83.0   1st Qu.:229.0
##  Median :1463   Median : 963.0   Median :157.0   Median :374.0
##  Mean   :1456   Mean   :1022.9   Mean   :181.8   Mean   :366.1
##  3rd Qu.:1674   3rd Qu.:1273.5   3rd Qu.:251.0   3rd Qu.:507.0
##  Max.   :2775   Max.   :2523.0   Max.   :438.0   Max.   :755.0
##  NA's   :366    NA's   :366
##        AH
##  Min.   :   2
##  1st Qu.:1575
##  Median :3147
##  Mean   :3173
##  3rd Qu.:4771
##  Max.   :6685
##
```

```
library(tidyverse)
map_int(dt, function(.x) sum(is.na(.x)))
```

```
##        CO.GT.   PT08.S1.CO.      NMHC.GT.      C6H6.GT. PT08.S2.NMHC.
##             0           366          8443             0           366
##       NOx.GT.  PT08.S3.NOx.       NO2.GT.  PT08.S4.NO2.   PT08.S5.O3.
##          1639           366          1642           366           366
##             T            RH            AH
##             0             0             0
```

```
#column NMHC.GT. is almost NAs values so let's get rid of it:

dt$NMHC.GT. <- NULL
```

```
library(mice)
```

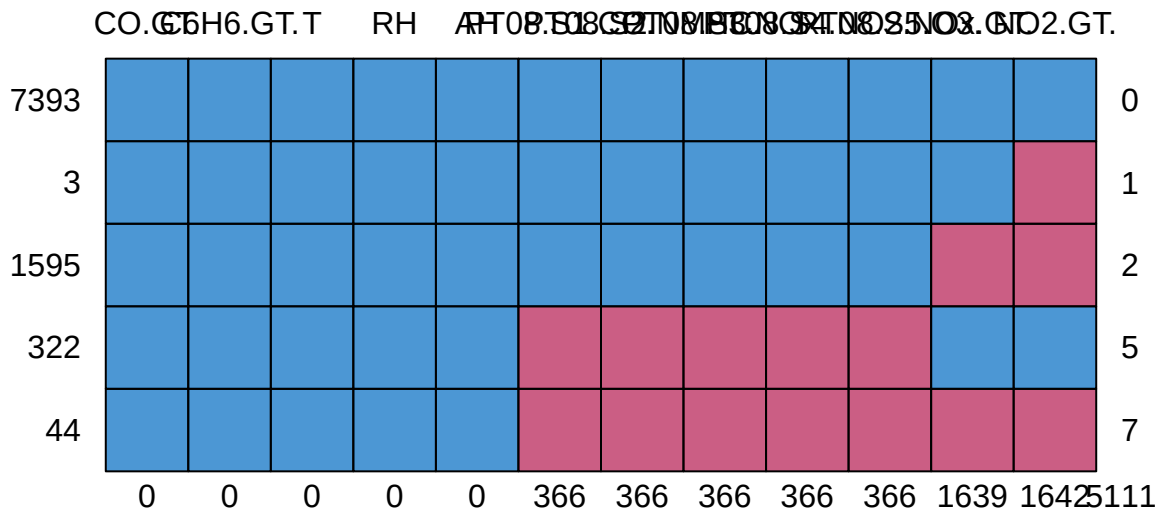
```
## Loading required package: lattice
```

```
##
## Attaching package: 'mice'
```

```
## The following object is masked from 'package:tidyr':
##
##     complete
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
md.pattern(dt)
```




```
##      CO.GT. C6H6.GT. T RH AH PT08.S1.CO. PT08.S2.NMHC. PT08.S3.NOx.
## 7393      1        1 1  1  1           1             1            1
## 3         1        1 1  1  1           1             1            1
## 1595      1        1 1  1  1           1             1            1
## 322       1        1 1  1  1           0             0            0
## 44        1        1 1  1  1           0             0            0
##           0        0 0  0  0         366           366          366
##      PT08.S4.NO2. PT08.S5.O3. NOx.GT. NO2.GT.
## 7393            1           1       1       1    0
## 3               1           1       1       0    1
## 1595            1           1       0       0    2
## 322             0           0       1       1    5
## 44              0           0       0       0    7
##               366         366    1639    1642 5111
```

```
library(VIM)
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

```
## Loading required package: data.table
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
## VIM is ready to use.
##  Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.
```

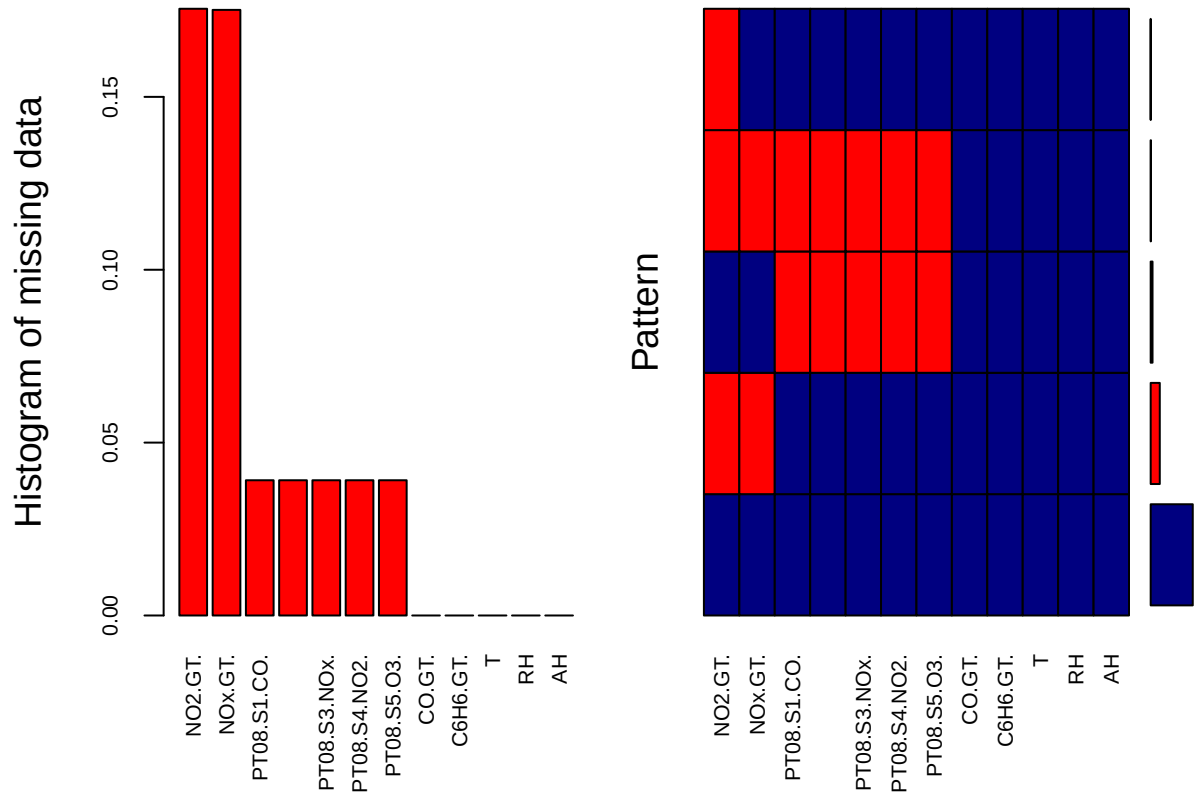
```
## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues
```

```
##
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
aggr_plot <- aggr(dt, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(dt), cex.axis=
```

```
## Warning in plot.aggr(res, ...): not enough horizontal space to display
## frequencies
```




```
##
##  Variables sorted by number of missings:
##       Variable     Count
##        NO2.GT. 0.1754836
##        NOx.GT. 0.1751630
##    PT08.S1.CO. 0.0391151
```

```
##  PT08.S2.NMHC. 0.0391151
##   PT08.S3.NOx. 0.0391151
##   PT08.S4.NO2. 0.0391151
##    PT08.S5.O3. 0.0391151
##         CO.GT. 0.0000000
##       C6H6.GT. 0.0000000
##             T 0.0000000
##            RH 0.0000000
##            AH 0.0000000
```

```
#Impute missing data:
tempData <- mice(dt,m=5,maxit=50,meth='pmm',seed=500, print = FALSE)
summary(tempData)
```

```
## Class: mids
## Number of multiple imputations:  5
## Imputation methods:
##        CO.GT.   PT08.S1.CO.      C6H6.GT. PT08.S2.NMHC.       NOx.GT.
##            ""         "pmm"            ""         "pmm"         "pmm"
##  PT08.S3.NOx.       NO2.GT.  PT08.S4.NO2.   PT08.S5.O3.             T
##         "pmm"         "pmm"         "pmm"         "pmm"            ""
##            RH            AH
##            ""            ""
## PredictorMatrix:
##               CO.GT. PT08.S1.CO. C6H6.GT. PT08.S2.NMHC. NOx.GT.
## CO.GT.             0           1        1             1       1
## PT08.S1.CO.        1           0        1             1       1
## C6H6.GT.           1           1        0             1       1
## PT08.S2.NMHC.      1           1        1             0       1
## NOx.GT.            1           1        1             1       0
## PT08.S3.NOx.       1           1        1             1       1
##               PT08.S3.NOx. NO2.GT. PT08.S4.NO2. PT08.S5.O3. T RH AH
## CO.GT.                   1       1            1           1 1  1  1
## PT08.S1.CO.              1       1            1           1 1  1  1
## C6H6.GT.                 1       1            1           1 1  1  1
## PT08.S2.NMHC.            1       1            1           1 1  1  1
## NOx.GT.                  1       1            1           1 1  1  1
## PT08.S3.NOx.             0       1            1           1 1  1  1
```

```
cdt <- complete(tempData,1)
```

```
map_int(cdt, function(.x) sum(is.na(.x)))
```

```
##        CO.GT.   PT08.S1.CO.      C6H6.GT. PT08.S2.NMHC.       NOx.GT.
##             0            0             0            0             0
##  PT08.S3.NOx.       NO2.GT.  PT08.S4.NO2.   PT08.S5.O3.             T
##             0            0             0            0             0
##            RH            AH
##             0            0
```

```
summary(cdt)
```

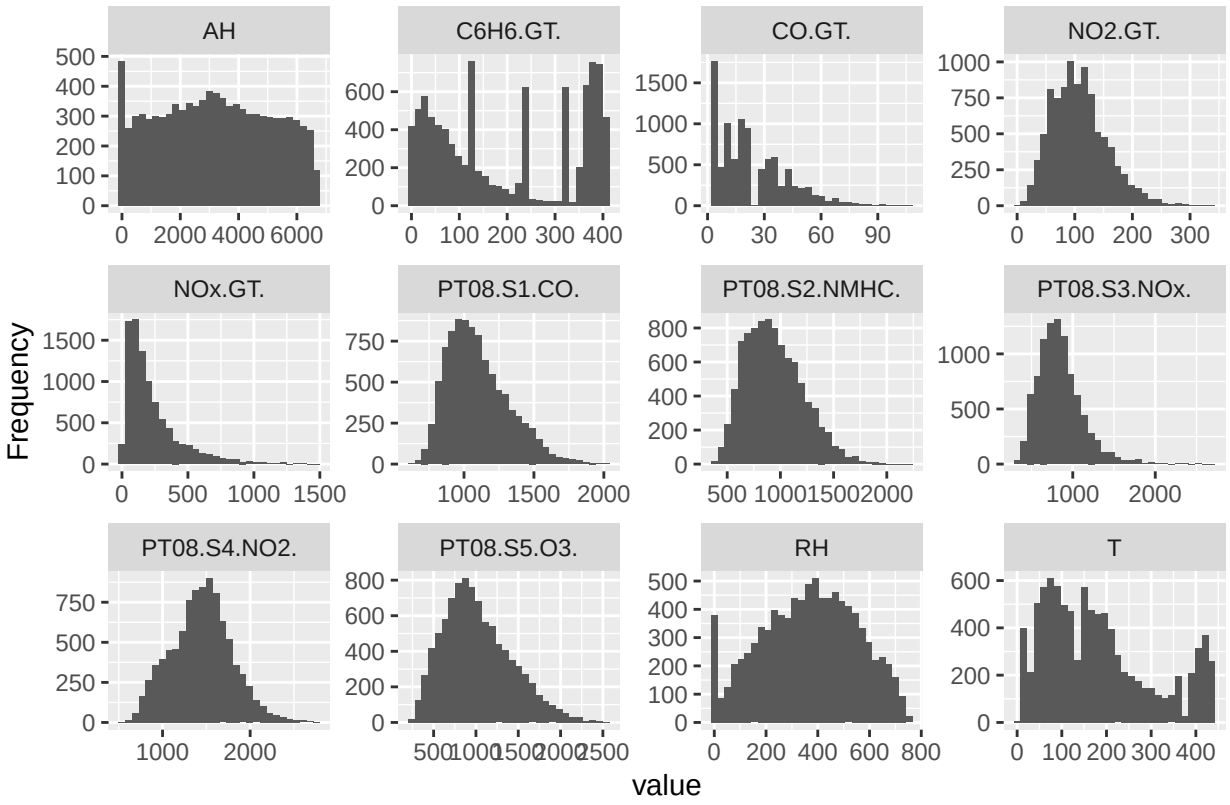
```
##      CO.GT.        PT08.S1.CO.      C6H6.GT.   PT08.S2.NMHC.
##  Min.   :  2.00   Min.   : 647   Min.   :  2   Min.   : 383.0
##  1st Qu.:  9.00   1st Qu.: 933   1st Qu.: 62   1st Qu.: 732.0
##  Median : 19.00   Median :1063   Median :161   Median : 909.0
```

```
##  Mean   : 23.27   Mean   :1099   Mean   :200   Mean   : 940.8
##  3rd Qu.: 35.00   3rd Qu.:1233   3rd Qu.:361   3rd Qu.:1120.0
##  Max.   :105.00   Max.   :2040   Max.   :409   Max.   :2214.0
##     NOx.GT.         PT08.S3.NOx.       NO2.GT.       PT08.S4.NO2.
##  Min.   :   2.0   Min.   : 322.0   Min.   :  2.0   Min.   : 551
##  1st Qu.:  88.0   1st Qu.: 660.0   1st Qu.: 72.0   1st Qu.:1213
##  Median : 163.0   Median : 810.0   Median :104.0   Median :1454
##  Mean   : 227.8   Mean   : 844.1   Mean   :108.2   Mean   :1447
##  3rd Qu.: 296.0   3rd Qu.: 981.0   3rd Qu.:136.0   3rd Qu.:1664
##  Max.   :1479.0   Max.   :2683.0   Max.   :340.0   Max.   :2775
##   PT08.S5.O3.          T               RH              AH
##  Min.   : 221   Min.   :  2.0   Min.   :  2.0   Min.   :   2
##  1st Qu.: 724   1st Qu.: 83.0   1st Qu.:229.0   1st Qu.:1575
##  Median : 961   Median :157.0   Median :374.0   Median :3147
##  Mean   :1021   Mean   :181.8   Mean   :366.1   Mean   :3173
##  3rd Qu.:1278   3rd Qu.:251.0   3rd Qu.:507.0   3rd Qu.:4771
##  Max.   :2523   Max.   :438.0   Max.   :755.0   Max.   :6685
```

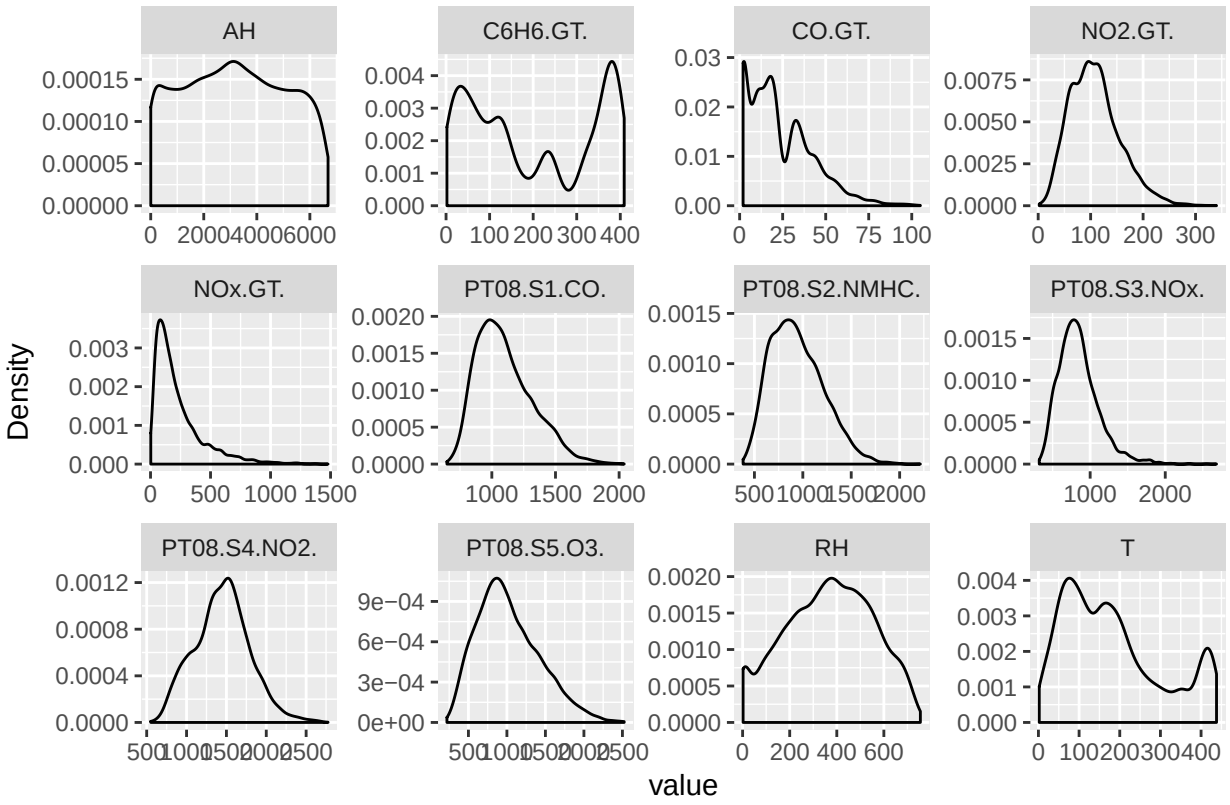
```
str(cdt)
```

```
## 'data.frame':    9357 obs. of  12 variables:
##  $ CO.GT.       : int  35 28 31 31 20 16 16 13 12 9 ...
##  $ PT08.S1.CO.  : int  1360 1292 1402 1376 1272 1197 1185 1136 1094 1010 ...
##  $ C6H6.GT.     : int  41 404 400 402 374 327 237 234 125 19 ...
##  $ PT08.S2.NMHC.: int  1046 955 939 948 836 750 690 672 609 561 ...
##  $ NOx.GT.      : int  166 103 131 172 131 89 62 62 45 63 ...
##  $ PT08.S3.NOx. : int  1056 1174 1140 1092 1205 1337 1462 1453 1579 1705 ...
##  $ NO2.GT.      : int  113 92 114 122 116 96 77 76 60 12 ...
##  $ PT08.S4.NO2. : int  1692 1559 1555 1584 1490 1393 1333 1333 1276 1235 ...
##  $ PT08.S5.O3.  : int  1268 972 1074 1203 1110 949 733 730 620 501 ...
##  $ T            : int  68 65 51 42 44 44 45 39 39 35 ...
##  $ RH           : int  377 365 428 488 484 480 456 488 485 490 ...
##  $ AH           : int  1898 1729 1855 2058 2068 2047 1911 1964 1937 1865 ...
```

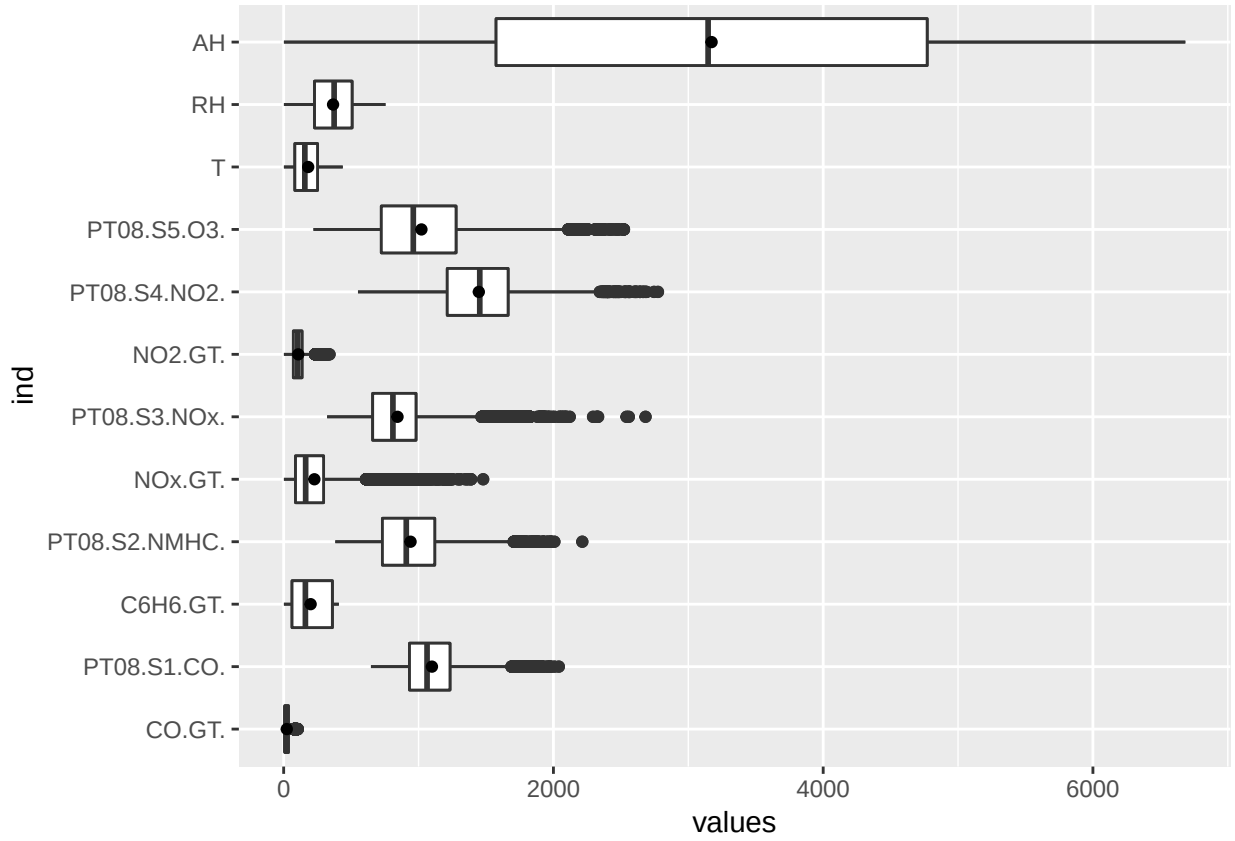
```
# To see the data distribution:
library(DataExplorer)
plot_histogram(cdt)
```




```
plot_density(cdt)
```




```
# let's make a boxplot for tidy data (cdt) to check the scales of the variables:

ggplot(stack(data.frame(cdt)), aes(x = ind, y = values)) +
geom_boxplot() +
guides(fill = FALSE) +
geom_boxplot() +
stat_summary(fun.y = mean, geom = "point") + coord_flip()
```

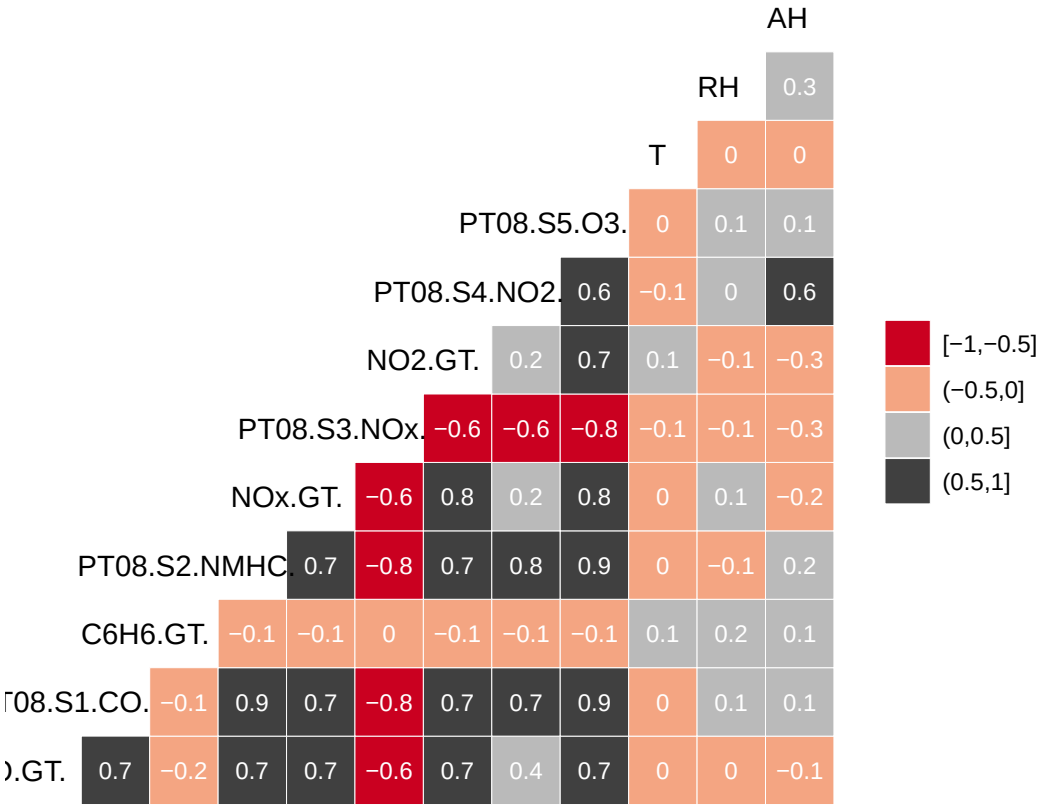



```
#To see the correlation matrix:
library(GGally)
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
ggcorr(cdt, nbreaks = 4, palette = "RdGy", label = TRUE, label_size = 3, label_color = "white", hjust =
```

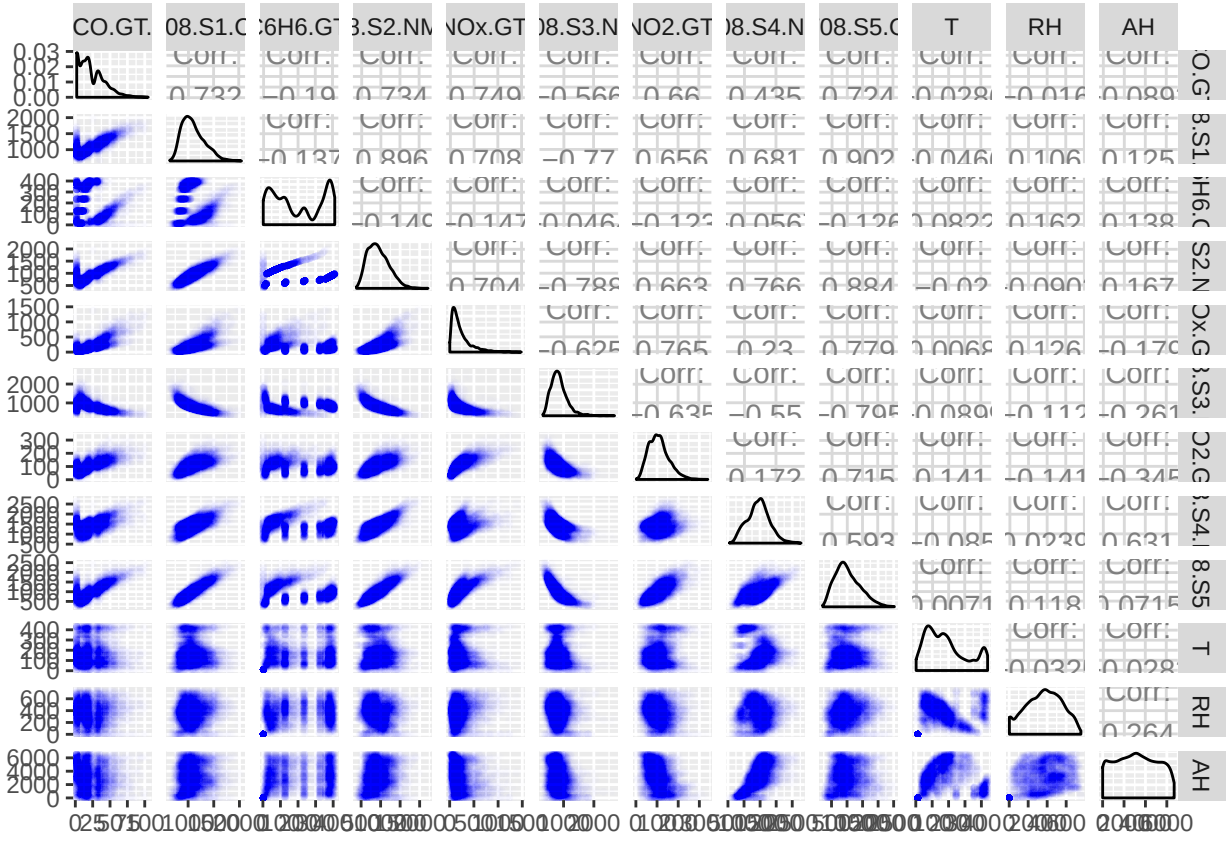



```
summary(cdt)
```

```
##      CO.GT.         PT08.S1.CO.      C6H6.GT.    PT08.S2.NMHC.
##  Min.   :  2.00   Min.   : 647   Min.   :  2   Min.   : 383.0
##  1st Qu.:  9.00   1st Qu.: 933   1st Qu.: 62   1st Qu.: 732.0
##  Median : 19.00   Median :1063   Median :161   Median : 909.0
##  Mean   : 23.27   Mean   :1099   Mean   :200   Mean   : 940.8
##  3rd Qu.: 35.00   3rd Qu.:1233   3rd Qu.:361   3rd Qu.:1120.0
##  Max.   :105.00   Max.   :2040   Max.   :409   Max.   :2214.0
##     NOx.GT.         PT08.S3.NOx.       NO2.GT.       PT08.S4.NO2.
##  Min.   :   2.0   Min.   : 322.0   Min.   :  2.0   Min.   : 551
##  1st Qu.:  88.0   1st Qu.: 660.0   1st Qu.: 72.0   1st Qu.:1213
##  Median : 163.0   Median : 810.0   Median :104.0   Median :1454
##  Mean   : 227.8   Mean   : 844.1   Mean   :108.2   Mean   :1447
##  3rd Qu.: 296.0   3rd Qu.: 981.0   3rd Qu.:136.0   3rd Qu.:1664
##  Max.   :1479.0   Max.   :2683.0   Max.   :340.0   Max.   :2775
##   PT08.S5.O3.         T               RH              AH
##  Min.   : 221   Min.   :  2.0   Min.   :  2.0   Min.   :   2
##  1st Qu.: 724   1st Qu.: 83.0   1st Qu.:229.0   1st Qu.:1575
##  Median : 961   Median :157.0   Median :374.0   Median :3147
##  Mean   :1021   Mean   :181.8   Mean   :366.1   Mean   :3173
##  3rd Qu.:1278   3rd Qu.:251.0   3rd Qu.:507.0   3rd Qu.:4771
##  Max.   :2523   Max.   :438.0   Max.   :755.0   Max.   :6685
```

```
ggpairs(cdt,
        lower = list(continuous = wrap("points",
                                       alpha = 0.004,
```

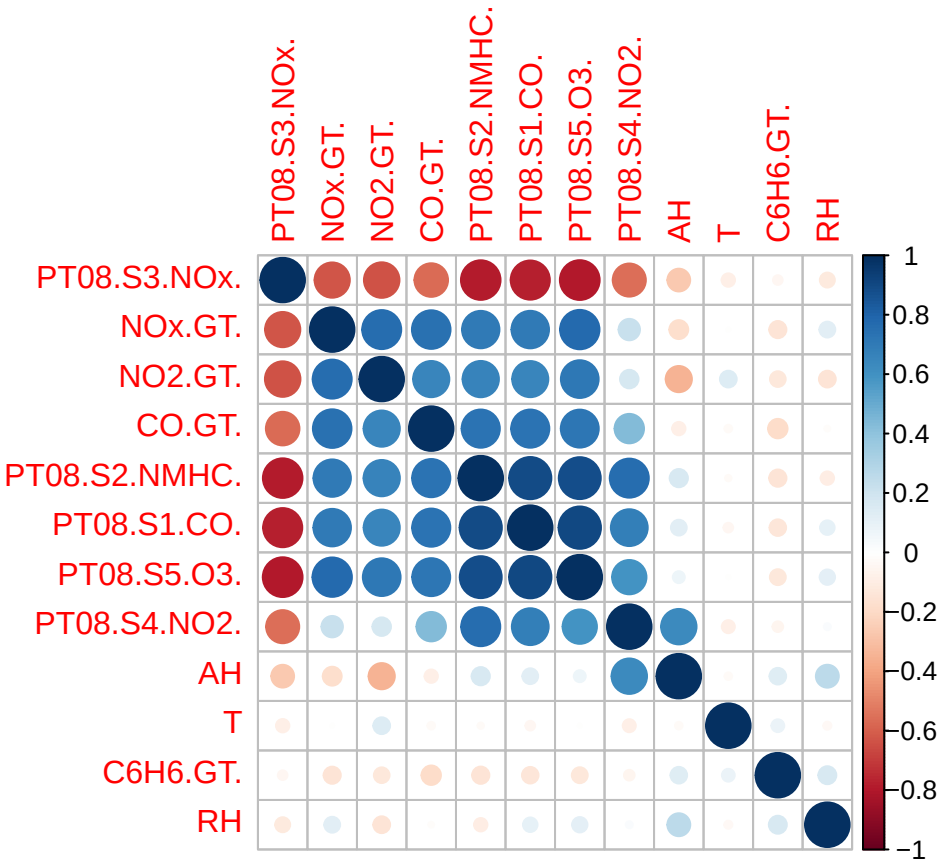
```
                                color = "blue",
                                size = .5)))
```



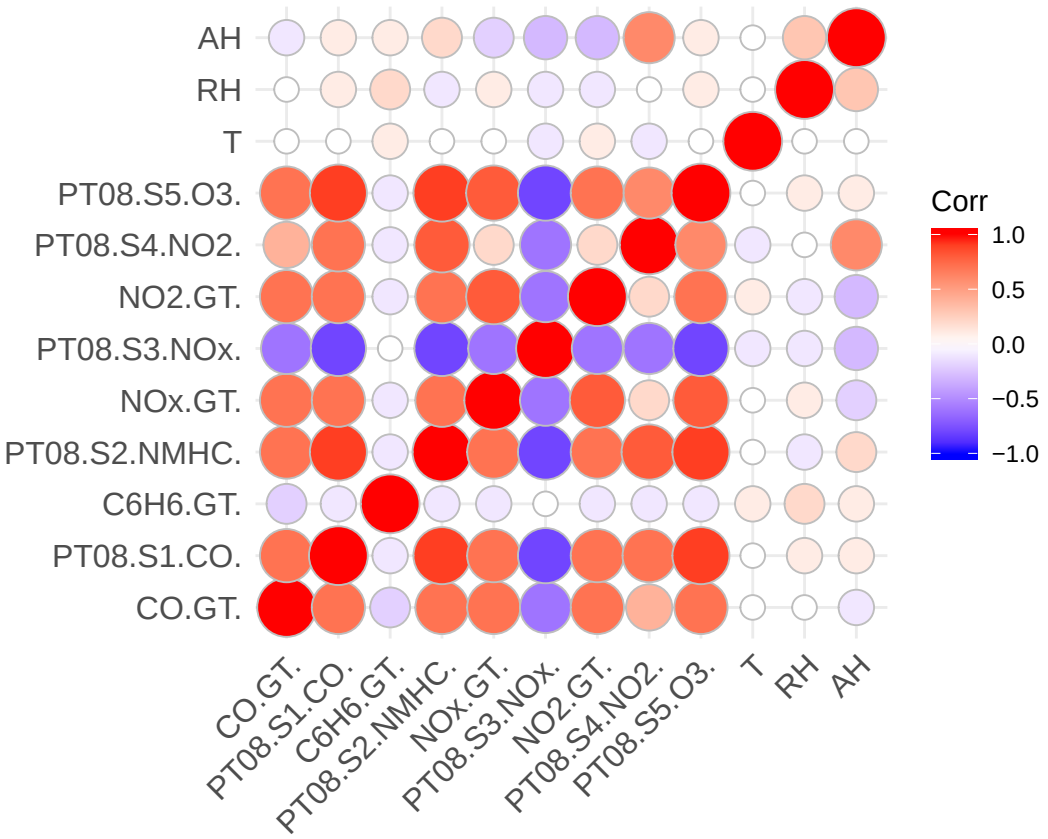

```
# Correlation matrix:
library(corrplot)
```

```
## corrplot 0.84 loaded
```

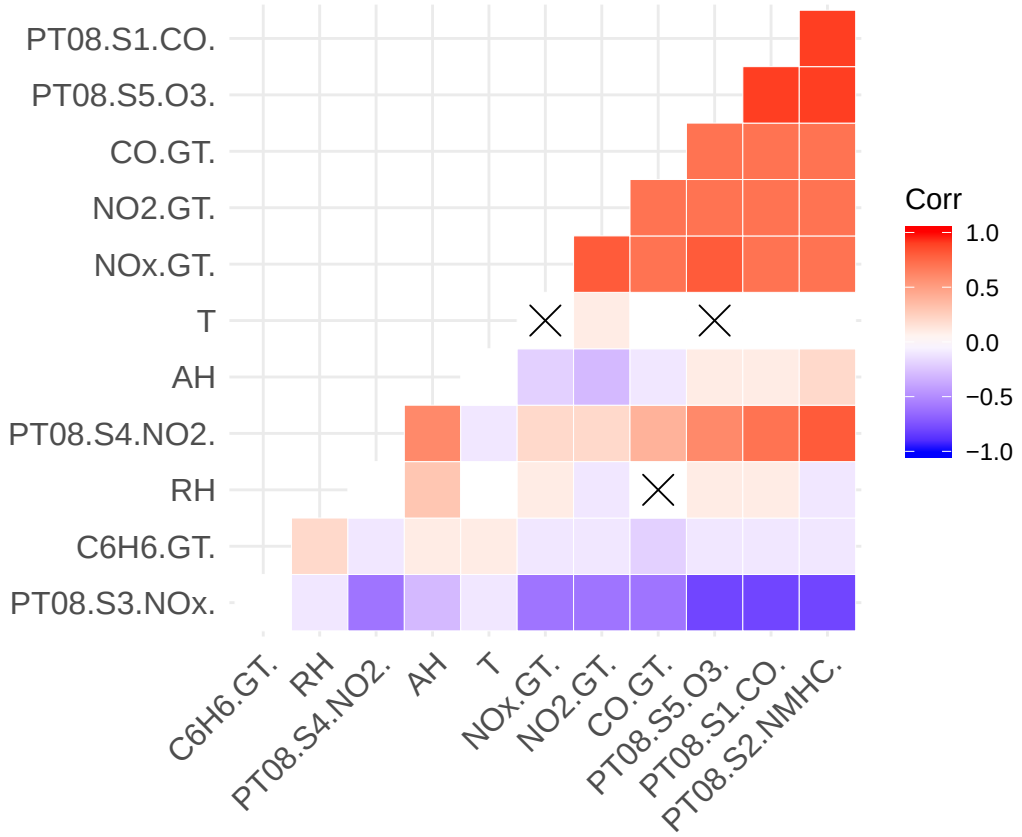
```
corMatMy <- cor(cdt)
corrplot(corMatMy, order = "hclust")
```




```
library(ggcorrplot)
corr <- round(cor(cdt), 1)
p.mat <- cor_pmat(cdt)
ggcorrplot(corr, method = "circle")
```




```
ggcorrplot(corr, hc.order = TRUE, outline.col = "white", type = "lower", p.mat = p.mat)
```

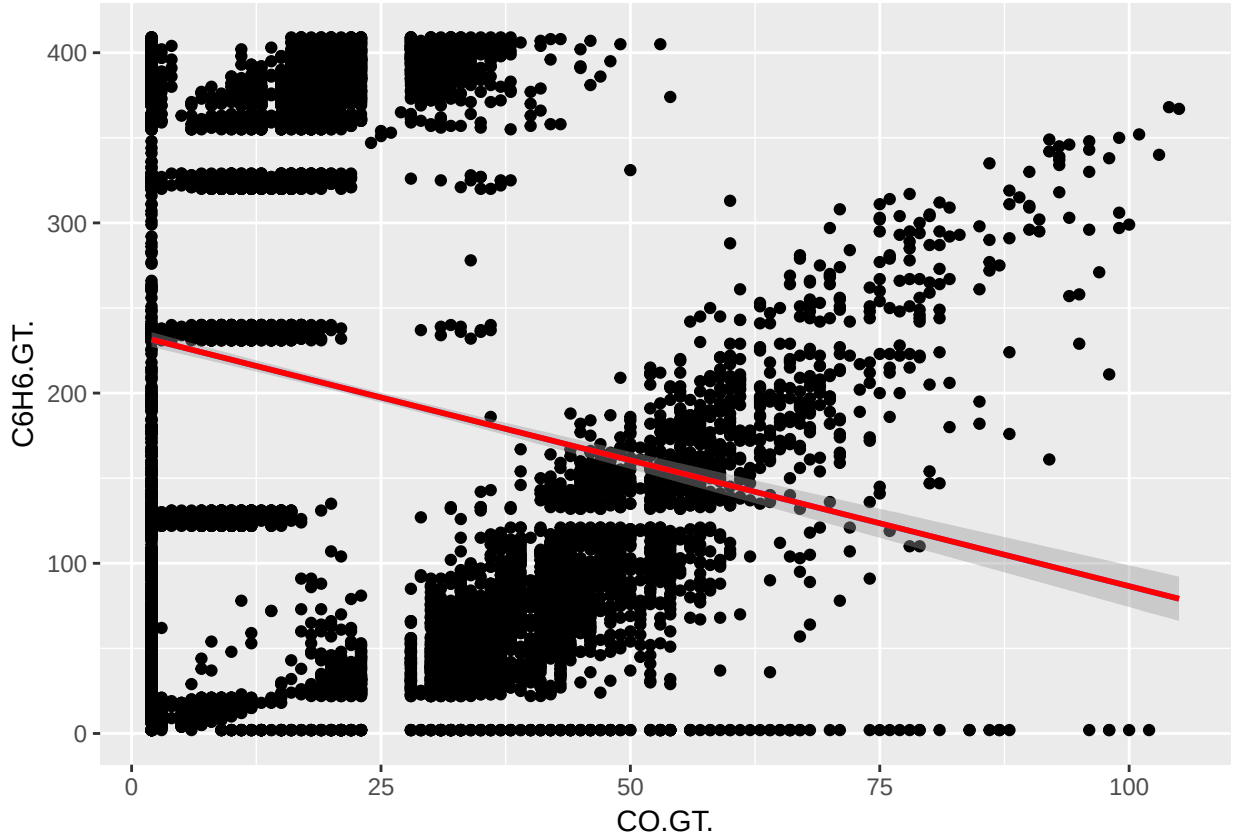



```
#Variable C6H6 has small correlation with other variables:
#more or less correlation is indicated with CO.GT.
```

```
#Try to understand degree of linearity between RH output and other input features
#plot all X-features against output variable C6H6.GT:

lr1 <- lm(C6H6.GT. ~ CO.GT., data = cdt)
summary(lr1)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ CO.GT., data = cdt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -229.434 -124.434   -3.434  149.305  287.827
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 234.39040    2.35711   99.44   <2e-16 ***
## CO.GT.       -1.47826    0.07919  -18.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 142.2 on 9355 degrees of freedom
## Multiple R-squared:  0.03591,    Adjusted R-squared:  0.03581
## F-statistic: 348.5 on 1 and 9355 DF,  p-value: < 2.2e-16
```

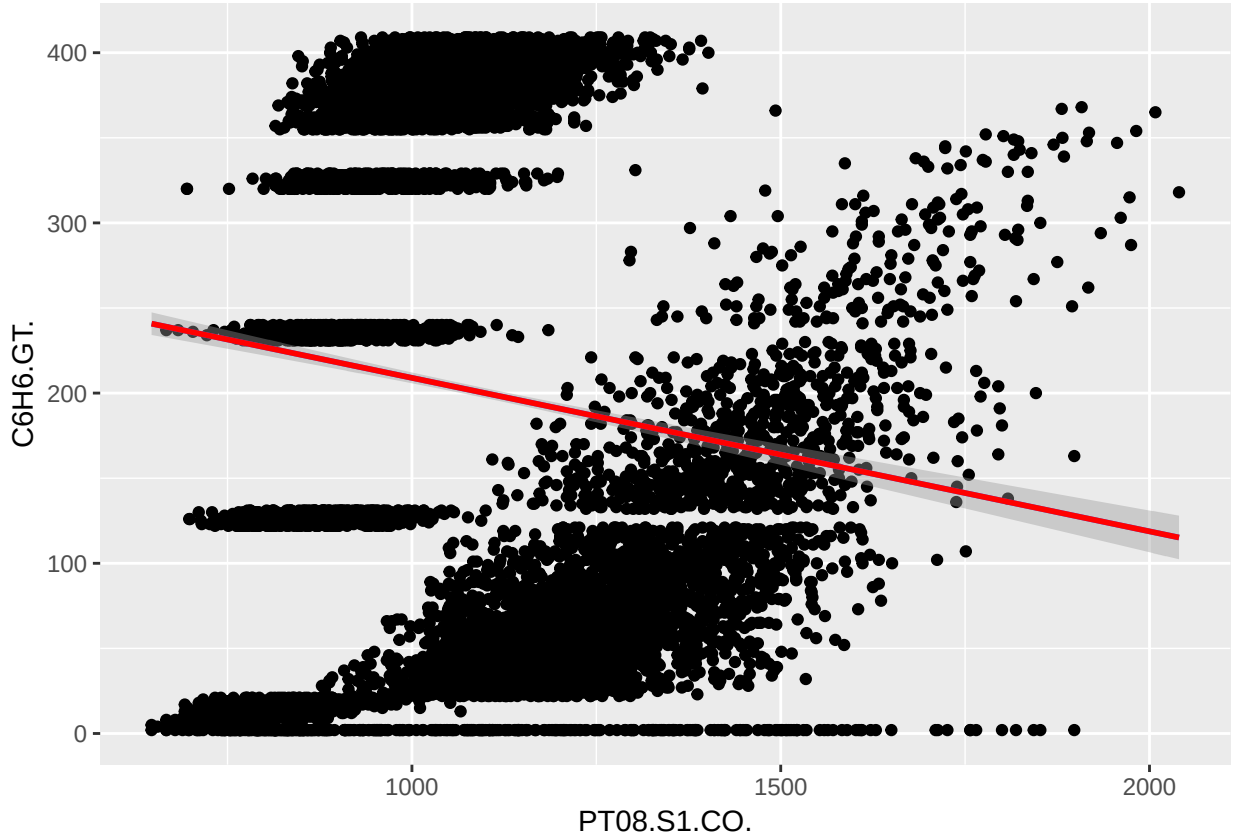
```
ggplot(lr1, aes(x = CO.GT., y = C6H6.GT.)) +
  geom_point() +
  stat_smooth(method = lm) +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```




```
lr2 <- lm(C6H6.GT. ~ PT08.S1.CO., data = cdt)
summary(lr2)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S1.CO., data = cdt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -238.794 -127.898   -9.684  153.680  246.975
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 299.156926   7.558104   39.58   <2e-16 ***
## PT08.S1.CO.  -0.090205   0.006741  -13.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143.4 on 9355 degrees of freedom
## Multiple R-squared:  0.01878,    Adjusted R-squared:  0.01867
## F-statistic:   179 on 1 and 9355 DF,  p-value: < 2.2e-16
```

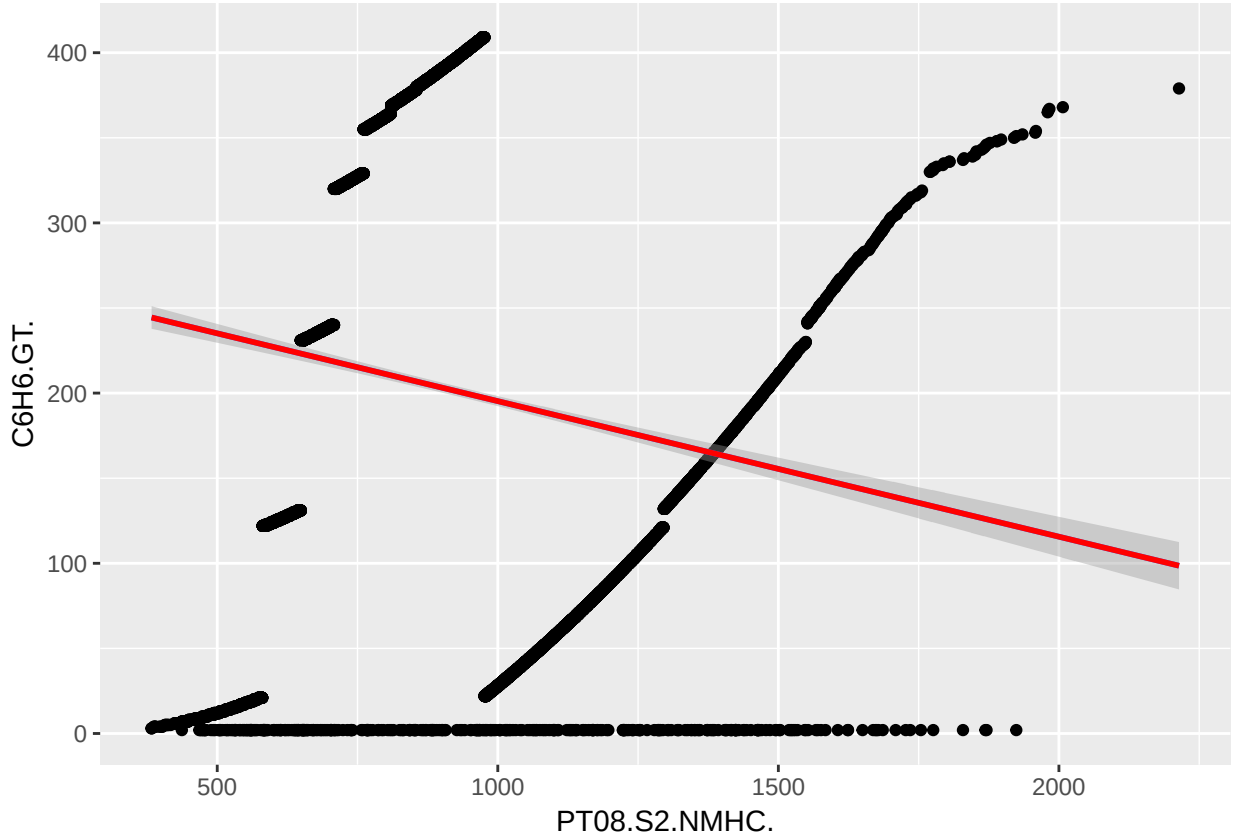
```
ggplot(lr2, aes(x = PT08.S1.CO., y = C6H6.GT.)) +
  geom_point() +
  stat_smooth(method = lm) +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```




```
lr3 <- lm(C6H6.GT. ~ PT08.S2.NMHC., data = cdt)
summary(lr3)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S2.NMHC., data = cdt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -241.426 -124.036   -4.484  150.792  280.429
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   274.934888   5.344792   51.44   <2e-16 ***
## PT08.S2.NMHC.  -0.079658   0.005459  -14.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143.2 on 9355 degrees of freedom
## Multiple R-squared:  0.02226,    Adjusted R-squared:  0.02215
## F-statistic: 212.9 on 1 and 9355 DF,  p-value: < 2.2e-16
```

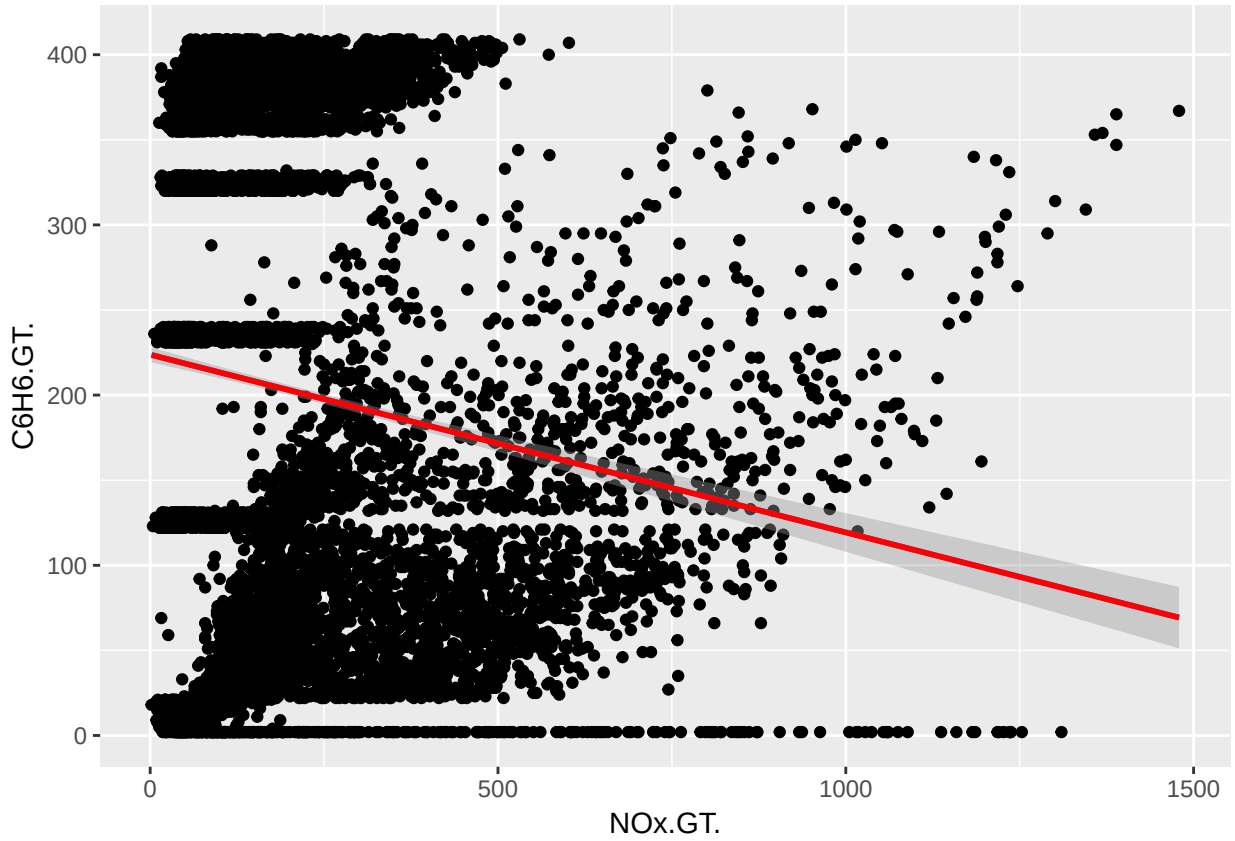
```r
ggplot(lr3, aes(x = PT08.S2.NMHC., y = C6H6.GT.)) +
  geom_point() +
  stat_smooth(method = lm) +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```




```r
lr4 <- lm(C6H6.GT. ~ NOx.GT., data = cdt)
summary(lr4)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ NOx.GT., data = cdt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -219.91 -132.74  -12.12  154.27  297.73
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 223.785934   2.219164  100.84   <2e-16 ***
## NOx.GT.      -0.104471   0.007258  -14.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143.2 on 9355 degrees of freedom
## Multiple R-squared:  0.02167,    Adjusted R-squared:  0.02156
## F-statistic: 207.2 on 1 and 9355 DF,  p-value: < 2.2e-16
```

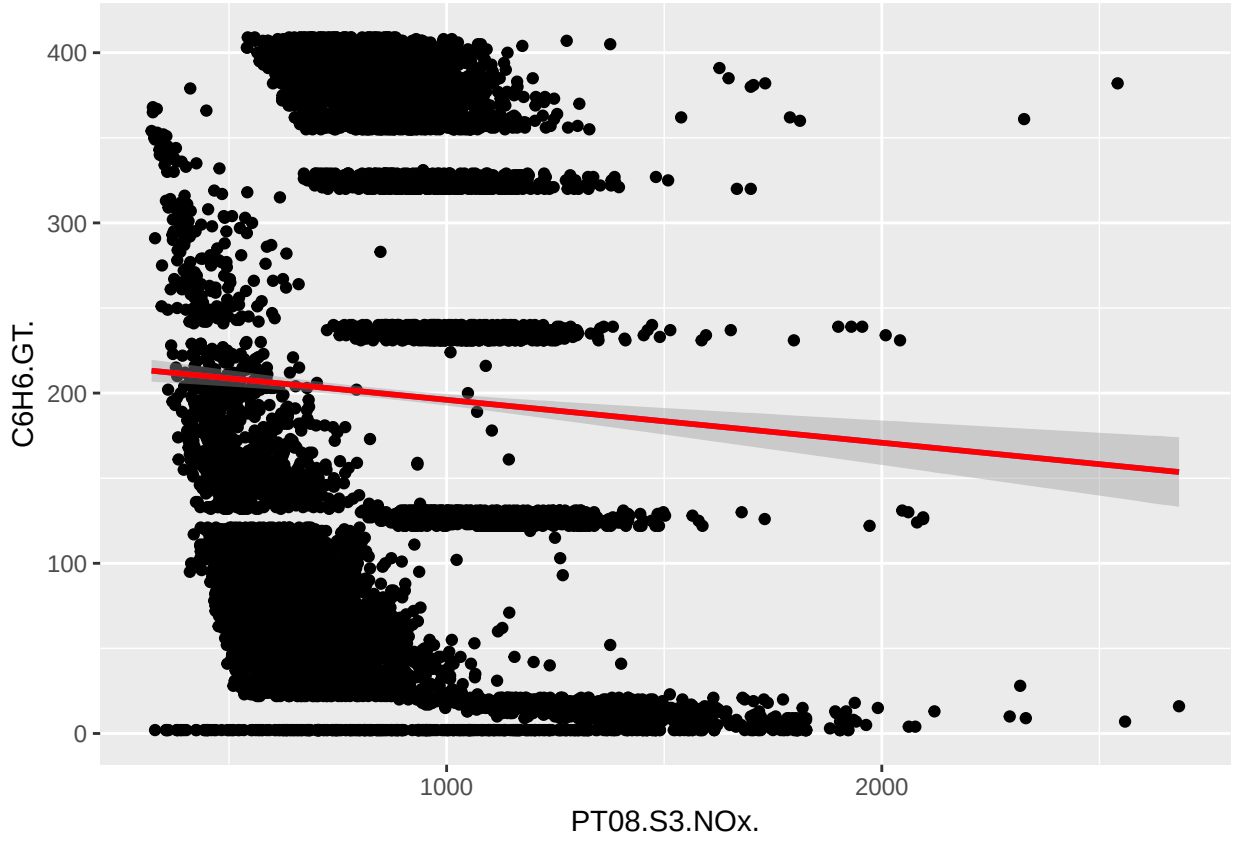
```r
ggplot(lr4, aes(x = NOx.GT., y = C6H6.GT.)) +
  geom_point() +
  stat_smooth(method = lm)+
  geom_line(aes(y = .fitted), color = "red", size = 1)
```




```r
lr5 <- lm(C6H6.GT. ~ PT08.S3.NOx., data = cdt)
summary(lr5)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S3.NOx., data = cdt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -210.95 -142.20  -46.18  162.57  224.82
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  221.271645   4.967117  44.547  < 2e-16 ***
## PT08.S3.NOx.  -0.025213   0.005612  -4.493 7.11e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 144.6 on 9355 degrees of freedom
## Multiple R-squared:  0.002153,   Adjusted R-squared:  0.002047
## F-statistic: 20.19 on 1 and 9355 DF,  p-value: 7.109e-06
```

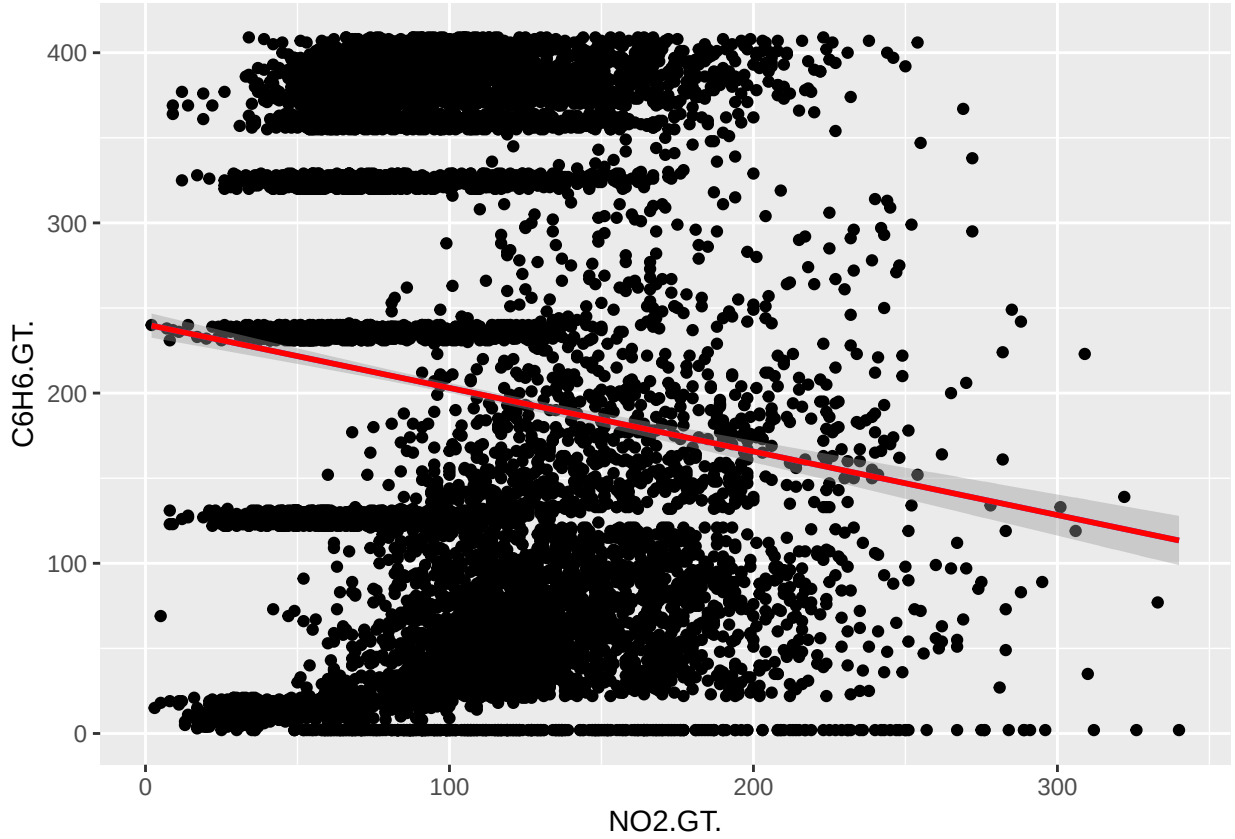
```
ggplot(lr5, aes(x = PT08.S3.NOx., y = C6H6.GT.)) +
  geom_point() +
  stat_smooth(method = lm) +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```




```
lr6 <- lm(C6H6.GT. ~ NO2.GT., data = cdt)
summary(lr6)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ NO2.GT., data = cdt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -231.0 -129.8  -21.0  154.4  260.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 240.39123    3.68230   65.28   <2e-16 ***
## NO2.GT.      -0.37350    0.03115  -11.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143.7 on 9355 degrees of freedom
## Multiple R-squared:  0.01514,    Adjusted R-squared:  0.01503
## F-statistic: 143.8 on 1 and 9355 DF,  p-value: < 2.2e-16
```

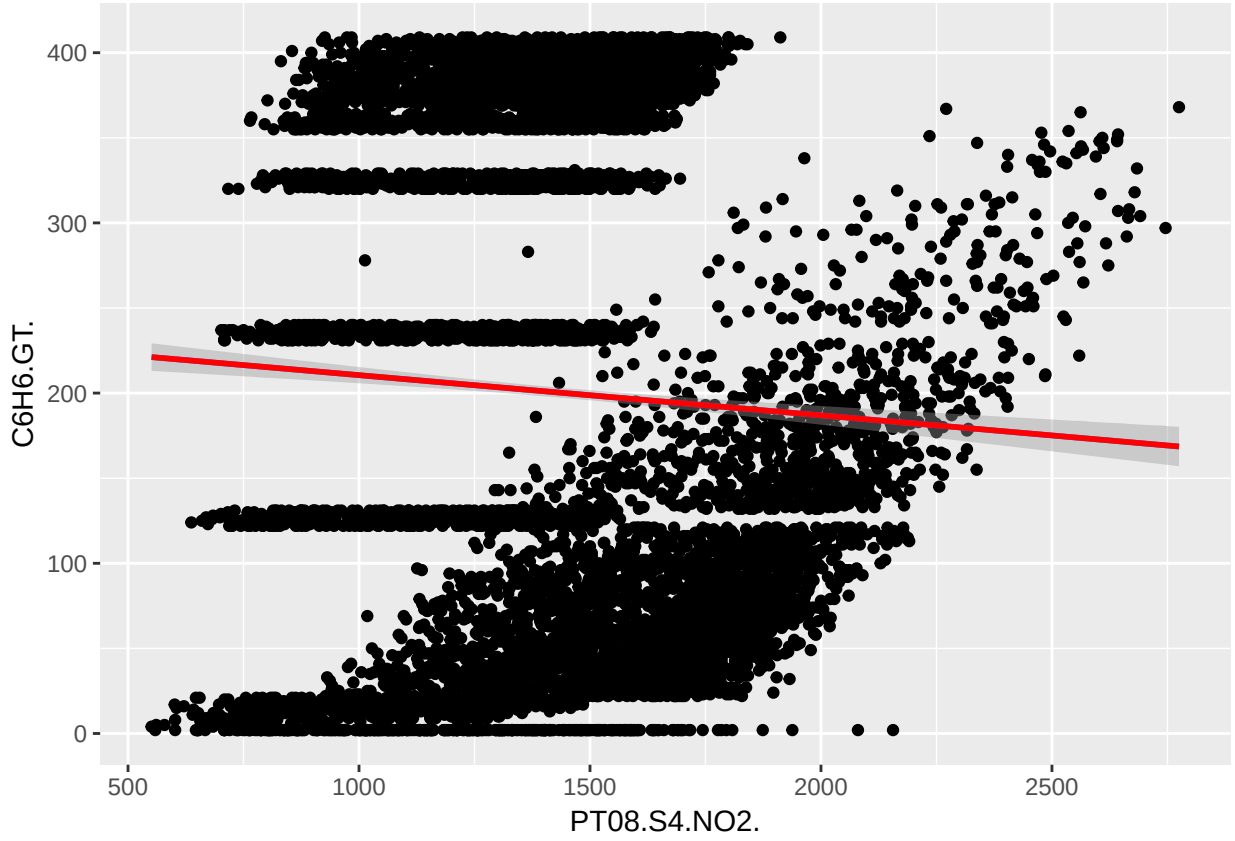
```
ggplot(lr6, aes(x = NO2.GT., y = C6H6.GT.)) +
  geom_point() +
  stat_smooth(method = lm) +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```




```
lr7 <- lm(C6H6.GT. ~ PT08.S4.NO2., data = cdt)
summary(lr7)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S4.NO2., data = cdt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -218.96 -134.34  -29.81  159.49  220.00
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  234.162980   6.400065  36.588  < 2e-16 ***
## PT08.S4.NO2.  -0.023618   0.004301  -5.491 4.09e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 144.6 on 9355 degrees of freedom
## Multiple R-squared:  0.003213,   Adjusted R-squared:  0.003106
## F-statistic: 30.15 on 1 and 9355 DF,  p-value: 4.094e-08
```

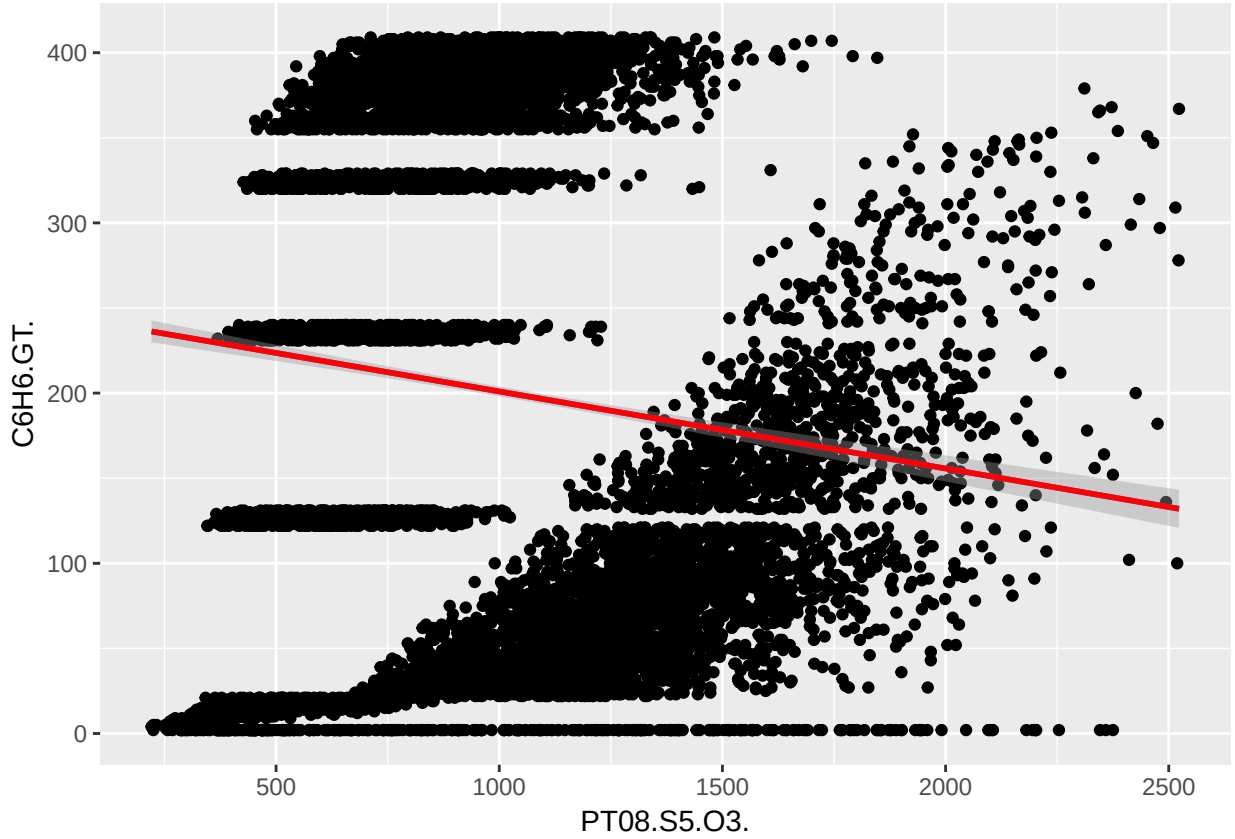
```r
ggplot(lr7, aes(x = PT08.S4.NO2., y = C6H6.GT.)) +
  geom_point() +
  stat_smooth(method = lm) +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```




```r
lr8 <- lm(C6H6.GT. ~ PT08.S5.O3., data = cdt)
summary(lr8)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S5.O3., data = cdt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -234.04 -128.23  -10.56  154.65  239.76
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 246.22508    4.03229   61.06   <2e-16 ***
## PT08.S5.O3.  -0.04526    0.00367  -12.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143.6 on 9355 degrees of freedom
## Multiple R-squared:  0.016,  Adjusted R-squared:  0.01589
## F-statistic: 152.1 on 1 and 9355 DF,  p-value: < 2.2e-16
```

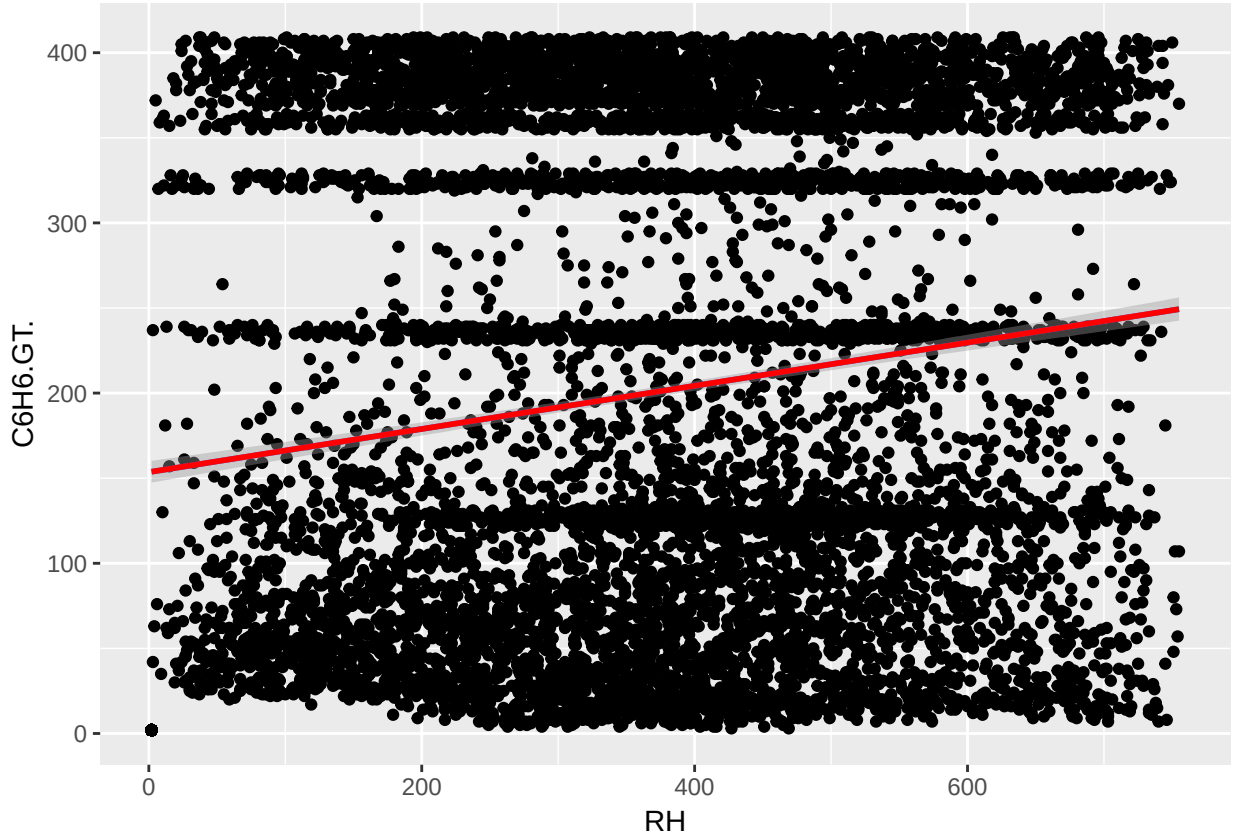
```
ggplot(lr8, aes(x = PT08.S5.O3., y = C6H6.GT.)) +
  geom_point() +
  stat_smooth(method = lm) +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```




```
lr9 <- lm(C6H6.GT. ~ RH, data = cdt)
summary(lr9)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ RH, data = cdt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -240.40 -131.14  -37.45  152.56  250.75
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 153.55972    3.27714   46.86   <2e-16 ***
## RH            0.12682    0.00799   15.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 142.9 on 9355 degrees of freedom
## Multiple R-squared:  0.02622,    Adjusted R-squared:  0.02612
## F-statistic: 251.9 on 1 and 9355 DF,  p-value: < 2.2e-16
```

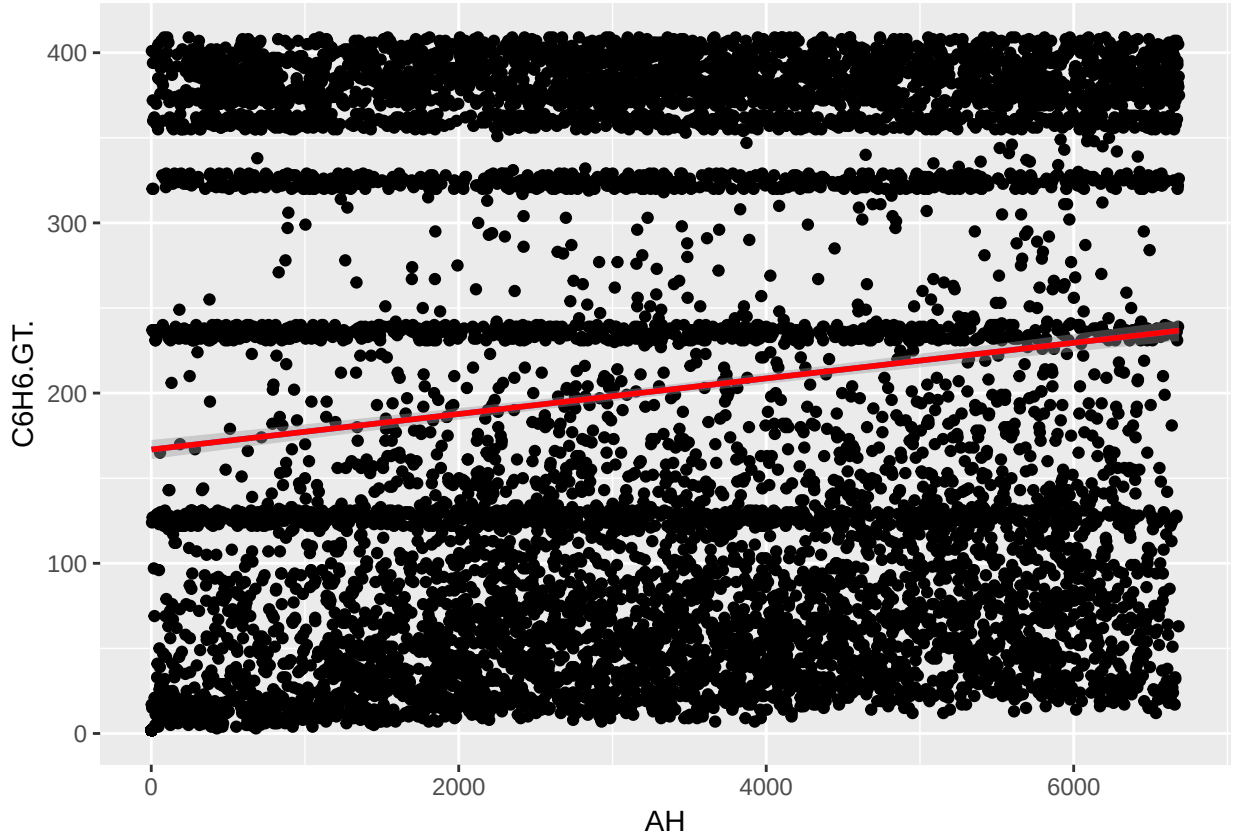
```r
ggplot(lr9, aes(x = RH, y = C6H6.GT.)) +
  geom_point() +
  stat_smooth(method = lm) +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```




```r
lr10 <- lm(C6H6.GT. ~ AH, data = cdt)
summary(lr10)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ AH, data = cdt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -223.1 -141.0  -40.4  154.1  241.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.668e+02  2.866e+00   58.21   <2e-16 ***
## AH          1.045e-02  7.730e-04   13.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143.4 on 9355 degrees of freedom
## Multiple R-squared:  0.01918,    Adjusted R-squared:  0.01907
## F-statistic: 182.9 on 1 and 9355 DF,  p-value: < 2.2e-16
```

```
ggplot(lr10, aes(x = AH, y = C6H6.GT.)) +
  geom_point() +
  stat_smooth(method = lm) +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```




```
# Find the best model for prediction:
lmMod <- lm(C6H6.GT. ~ . , data = cdt)
selectedMod <- step(lmMod)
```

```
## Start:  AIC=91818.03
## C6H6.GT. ~ CO.GT. + PT08.S1.CO. + PT08.S2.NMHC. + NOx.GT. + PT08.S3.NOx. +
##     NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. + T + RH + AH
##
##                 Df Sum of Sq       RSS   AIC
## - NOx.GT.        1      2165 170473721 91816
## <none>                       170471555 91818
## - NO2.GT.        1     61656 170533212 91819
## - AH             1    179498 170651053 91826
## - T              1    245166 170716722 91829
## - PT08.S5.O3.    1    410851 170882406 91839
## - PT08.S4.NO2.   1    577931 171049486 91848
## - CO.GT.         1    659024 171130580 91852
## - PT08.S2.NMHC.  1    755214 171226769 91857
## - PT08.S1.CO.    1    825098 171296653 91861
## - RH             1   1931514 172403070 91921
## - PT08.S3.NOx.   1   8140234 178611789 92253
```

```
##
## Step:  AIC=91816.15
## C6H6.GT. ~ CO.GT. + PT08.S1.CO. + PT08.S2.NMHC. + PT08.S3.NOx. +
##     NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. + T + RH + AH
##
##                 Df Sum of Sq       RSS   AIC
## <none>                       170473721 91816
## - NO2.GT.        1     59589 170533310 91817
## - AH             1    227188 170700909 91827
## - T              1    253853 170727574 91828
## - PT08.S5.O3.    1    432591 170906312 91838
## - CO.GT.         1    805677 171279398 91858
## - PT08.S1.CO.    1    830109 171303829 91860
## - PT08.S4.NO2.   1   1024364 171498085 91870
## - PT08.S2.NMHC.  1   1188881 171662602 91879
## - RH             1   2182951 172656672 91933
## - PT08.S3.NOx.   1   8949285 179423006 92293
```

```
summary(selectedMod)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ CO.GT. + PT08.S1.CO. + PT08.S2.NMHC. +
##     PT08.S3.NOx. + NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. + T +
##     RH + AH, data = cdt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -284.64 -119.75  -12.91  129.11  623.21
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   603.681903  22.203812  27.188  < 2e-16 ***
## CO.GT.         -0.821021   0.123535  -6.646 3.18e-11 ***
## PT08.S1.CO.    -0.125603   0.018619  -6.746 1.61e-11 ***
## PT08.S2.NMHC.  -0.174199   0.021577  -8.073 7.69e-16 ***
## PT08.S3.NOx.   -0.254127   0.011473 -22.150  < 2e-16 ***
## NO2.GT.         0.113275   0.062671   1.807 0.070723 .
## PT08.S4.NO2.    0.097121   0.012960   7.494 7.29e-14 ***
## PT08.S5.O3.    -0.049048   0.010072  -4.870 1.13e-06 ***
## T               0.044862   0.012025   3.731 0.000192 ***
## RH              0.105689   0.009661  10.940  < 2e-16 ***
## AH             -0.005559   0.001575  -3.529 0.000419 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 135.1 on 9346 degrees of freedom
## Multiple R-squared:  0.1307, Adjusted R-squared:  0.1298
## F-statistic: 140.6 on 10 and 9346 DF,  p-value: < 2.2e-16
```

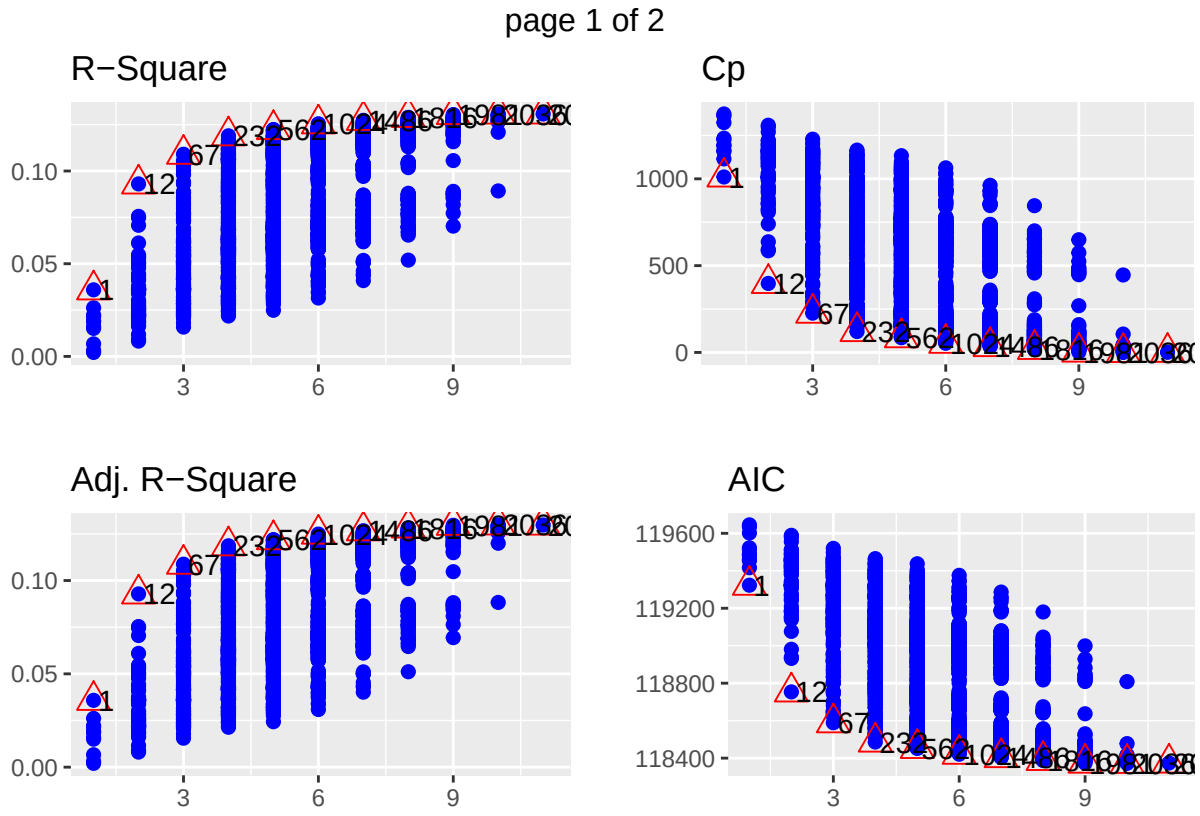
```
#Linear Regression, the Null Hypothesis is that the coefficients
#associated with the variables is equal to zero.
#The alternate hypothesis is that the coefficients are not equal to zero
#(i.e. there exists a relationship between the independent variable in
#question and the dependent variable).
```

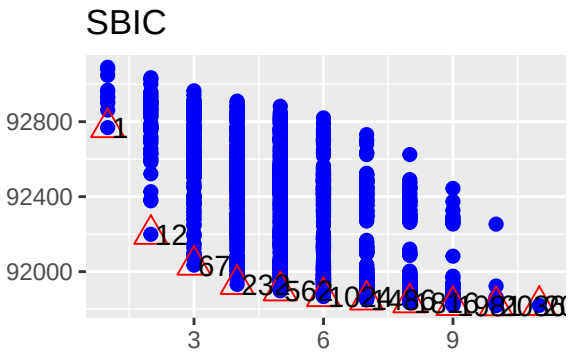
```
library(olsrr)
```

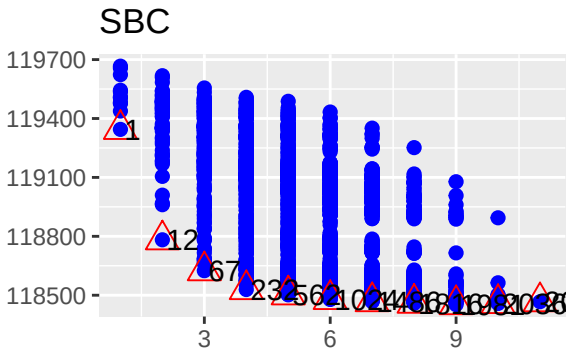
```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
#to find the best model:

model <- lm(C6H6.GT. ~ ., data = cdt)
k <- ols_step_all_possible(model)
plot(k)
```








```
ols_step_best_subset(model)
```

```
##                                          Best Subsets Regression
## ---------------------------------------------------------------------------------------------------
## Model Index    Predictors
## ---------------------------------------------------------------------------------------------------
##      1         CO.GT.
##      2         PT08.S2.NMHC. PT08.S3.NOx.
##      3         PT08.S2.NMHC. PT08.S3.NOx. PT08.S4.NO2.
##      4         PT08.S1.CO. NOx.GT. PT08.S3.NOx. RH
##      5         PT08.S1.CO. NOx.GT. PT08.S3.NOx. PT08.S5.O3. RH
##      6         CO.GT. PT08.S1.CO. PT08.S2.NMHC. PT08.S3.NOx. PT08.S4.NO2. RH
##      7         CO.GT. PT08.S1.CO. PT08.S2.NMHC. PT08.S3.NOx. PT08.S4.NO2. PT08.S5.O3. RH
##      8         CO.GT. PT08.S1.CO. PT08.S2.NMHC. PT08.S3.NOx. PT08.S4.NO2. PT08.S5.O3. T RH
##      9         CO.GT. PT08.S1.CO. PT08.S2.NMHC. PT08.S3.NOx. PT08.S4.NO2. PT08.S5.O3. T RH AH
##     10         CO.GT. PT08.S1.CO. PT08.S2.NMHC. PT08.S3.NOx. NO2.GT. PT08.S4.NO2. PT08.S5.O3. T RH A
##     11         CO.GT. PT08.S1.CO. PT08.S2.NMHC. NOx.GT. PT08.S3.NOx. NO2.GT. PT08.S4.NO2. PT08.S5.O3
## ---------------------------------------------------------------------------------------------------
##
##                                                    Subsets Regression Summary
## ---------------------------------------------------------------------------------------------------
##                        Adj.        Pred
## Model    R-Square    R-Square    R-Square      C(p)           AIC           SBIC           SBC
## ---------------------------------------------------------------------------------------------------
##   1        0.0359      0.0358      0.0355    1011.6422    119323.0473    92768.6238    119344.4790
##   2        0.0930      0.0928      0.0924     399.5767    118753.5799    92199.3187    118782.1554
```

```
##   3        0.1090      0.1088      0.1082     229.4047    118588.8814    92034.6797    118624.6008
##   4        0.1190      0.1186       0.118     124.5701    118485.9299    91931.7929    118528.7932
##   5        0.1224      0.1219      0.1212      90.0971    118451.8285    91897.7133    118501.8357
##   6        0.1254      0.1249      0.1241      59.0989    118421.0460    91866.9631    118478.1971
##   7        0.1273      0.1267      0.1258      41.0049    118403.0271    91848.9687    118467.3220
##   8        0.1289      0.1282      0.1272      25.8292    118387.8815    91833.8507    118459.3203
##   9        0.1304      0.1296      0.1284      11.3853    118373.4368    91819.4395    118452.0195
##  10        0.1307      0.1298      0.1286      10.1187    118372.1666    91818.1789    118457.8932
##  11        0.1308      0.1297      0.1284      12.0000    118374.0478    91820.0629    118466.9182
## ----------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

```
#Check all variables for their p-values:

ols_step_forward_p(model)
```

```
## Forward Selection Method
## ---------------------------
##
## Candidate Terms:
##
## 1. CO.GT.
## 2. PT08.S1.CO.
## 3. PT08.S2.NMHC.
## 4. NOx.GT.
## 5. PT08.S3.NOx.
## 6. NO2.GT.
## 7. PT08.S4.NO2.
## 8. PT08.S5.O3.
## 9. T
## 10. RH
## 11. AH
##
## We are selecting variables based on p value...
##
## Variables Entered:
##
## - CO.GT.
## - PT08.S3.NOx.
## - PT08.S2.NMHC.
## - PT08.S4.NO2.
## - RH
## - PT08.S1.CO.
## - PT08.S5.O3.
## - T
## - AH
## - NO2.GT.
##
## No more variables to be added.
```

```
##
## Final Model Output
## ------------------
##
##                            Model Summary
## ----------------------------------------------------------------
## R                       0.362       RMSE                135.057
## R-Squared               0.131       Coef. Var            67.532
## Adj. R-Squared          0.130       MSE                18240.287
## Pred R-Squared          0.129       MAE                  120.228
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                  ANOVA
## --------------------------------------------------------------------------------
##                     Sum of
##                    Squares          DF    Mean Square       F          Sig.
## --------------------------------------------------------------------------------
## Regression     25640983.366          10    2564098.337    140.573    0.0000
## Residual      170473720.670        9346      18240.287
## Total         196114704.035        9356
## --------------------------------------------------------------------------------
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##         model       Beta    Std. Error    Std. Beta       t        Sig       lower       upper
## ----------------------------------------------------------------------------------------------------
##   (Intercept)    603.682        22.204                27.188    0.000    560.158    647.206
##        CO.GT.     -0.821         0.124       -0.105    -6.646    0.000     -1.063     -0.579
##  PT08.S3.NOx.     -0.254         0.011       -0.468   -22.150    0.000     -0.277     -0.232
## PT08.S2.NMHC.     -0.174         0.022       -0.326    -8.073    0.000     -0.216     -0.132
##  PT08.S4.NO2.      0.097         0.013        0.233     7.494    0.000      0.072      0.123
##            RH      0.106         0.010        0.135    10.940    0.000      0.087      0.125
##   PT08.S1.CO.     -0.126         0.019       -0.191    -6.746    0.000     -0.162     -0.089
##   PT08.S5.O3.     -0.049         0.010       -0.137    -4.870    0.000     -0.069     -0.029
##             T      0.045         0.012        0.038     3.731    0.000      0.021      0.068
##            AH     -0.006         0.002       -0.074    -3.529    0.000     -0.009     -0.002
##       NO2.GT.      0.113         0.063        0.037     1.807    0.071     -0.010      0.236
## ----------------------------------------------------------------------------------------------------

##
##                                  Selection Summary
## ------------------------------------------------------------------------------
##         Variable                    Adj.
## Step       Entered       R-Square    R-Square      C(p)           AIC          RMSE
## ------------------------------------------------------------------------------
##    1    CO.GT.             0.0359      0.0358    1011.6422    119323.0473    142.1646
##    2    PT08.S3.NOx.       0.0707      0.0705     639.3955    118980.9339    139.5819
##    3    PT08.S2.NMHC.      0.1052      0.1049     270.8633    118629.2943    136.9763
##    4    PT08.S4.NO2.       0.1148      0.1144     169.4002    118530.1132    136.2449
##    5    RH                 0.1186      0.1181     130.7099    118492.0186    135.9606
##    6    PT08.S1.CO.        0.1254      0.1249      59.0989    118421.0460    135.4387
```

```
##    7    PT08.S5.O3.        0.1273      0.1267      41.0049    118403.0271    135.3012
##    8    T                  0.1289      0.1282      25.8292    118387.8815    135.1845
##    9    AH                 0.1304      0.1296      11.3853    118373.4368    135.0730
##   10    NO2.GT.            0.1307      0.1298      10.1187    118372.1666    135.0566
## ------------------------------------------------------------------------------------
```

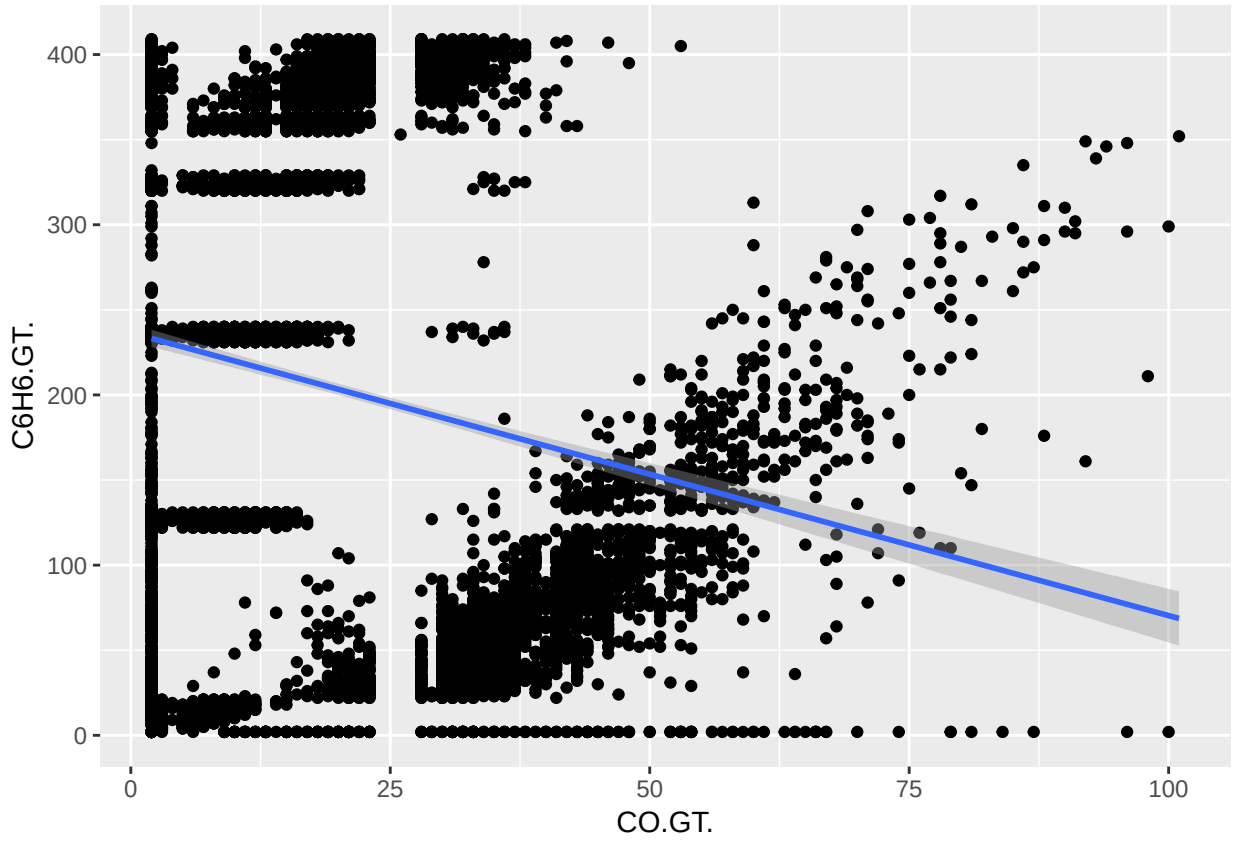
```
# from all statistic metrics let's choose ~CO.GT. model as more appropriate for further analysis:
#TASK: create train-test sets, plot the model, for the test
#set color real and predicted points differently; R^2 and p-value to title
```

```
# Prepare data for prediction and model training (75%):

library(caTools)
set.seed(42)
sample <- sample.split(cdt, SplitRatio = 0.75)
train <- subset(cdt, sample == TRUE)
test  <- subset(cdt, sample == FALSE)
```

```
new_mod <- lm(data = train, C6H6.GT. ~ CO.GT.)
summary(new_mod)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ CO.GT., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -231.232 -124.400   -1.232  149.704  283.340
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 236.5562     2.7647   85.56   <2e-16 ***
## CO.GT.       -1.6623     0.1002  -16.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143.4 on 7015 degrees of freedom
## Multiple R-squared:  0.03778,    Adjusted R-squared:  0.03765
## F-statistic: 275.5 on 1 and 7015 DF,  p-value: < 2.2e-16
```

```
ggplot(data = train, aes(x = CO.GT., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = "lm")
```




```
pred <- predict(new_mod, newdata = test)
head(pred)
```

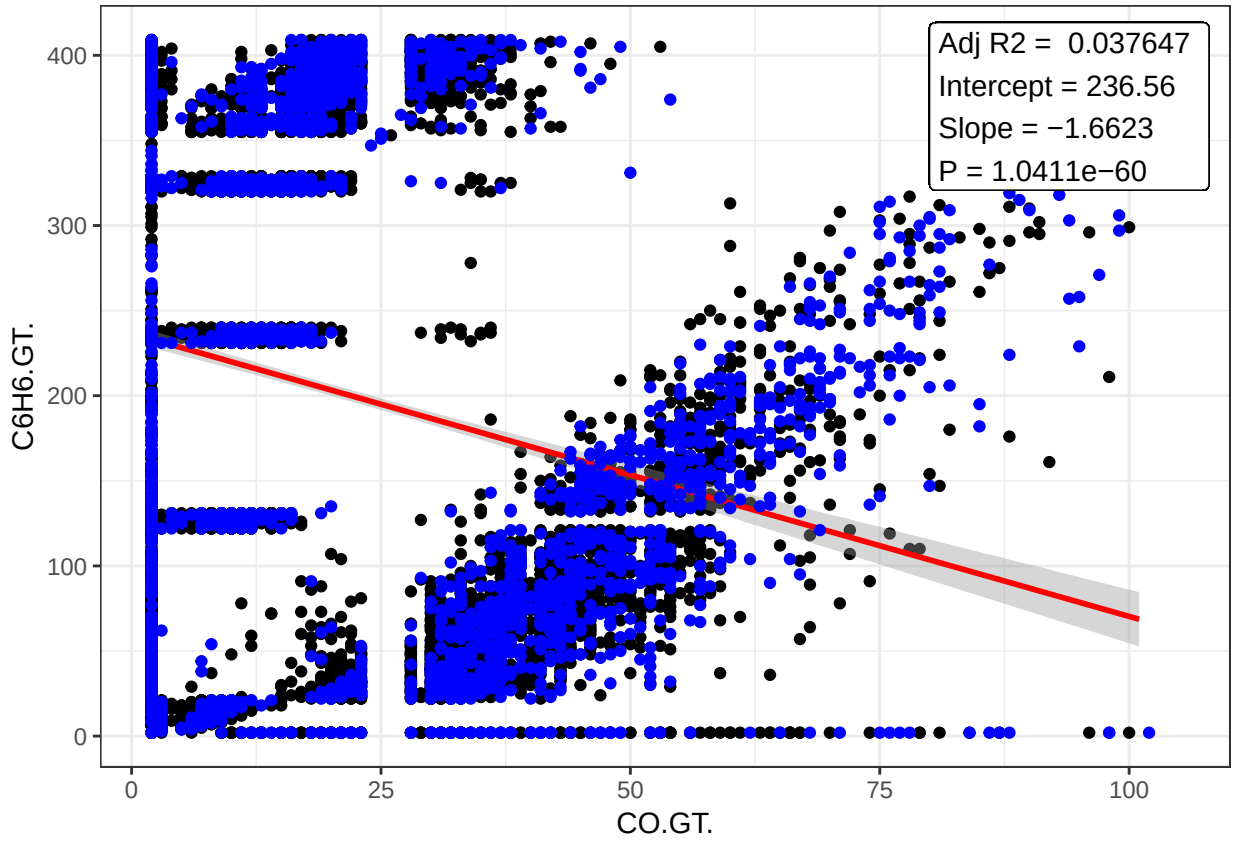
```
##        1        2        4       13       14       16
## 178.3743 190.0107 185.0237 219.9328 211.6211 185.0237
```

```
test$C6H6.GT.pred <- pred
head(test)
```

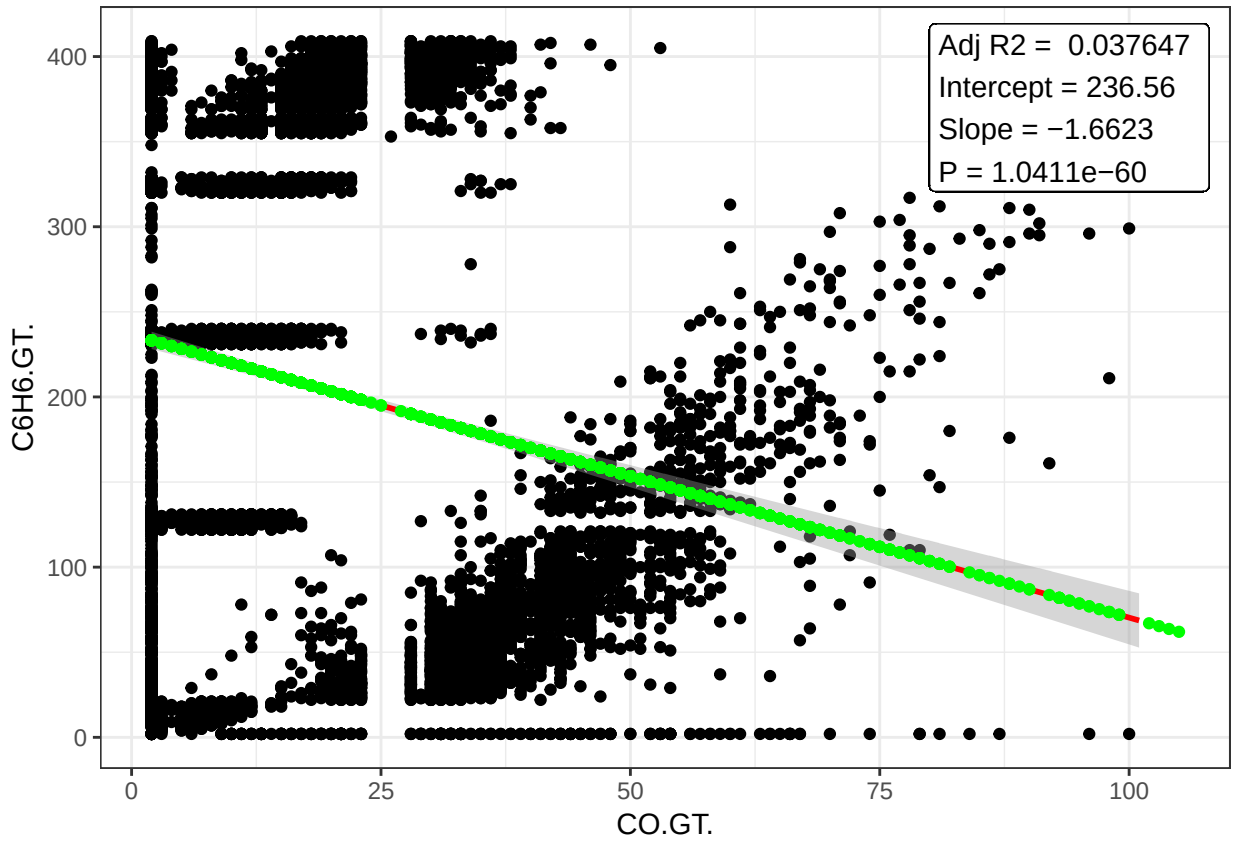
```
##    CO.GT. PT08.S1.CO. C6H6.GT. PT08.S2.NMHC. NOx.GT. PT08.S3.NOx. NO2.GT.
## 1      35        1360       41          1046     166         1056     113
## 2      28        1292      404           955     103         1174      92
## 4      31        1376      402           948     172         1092     122
## 13     10        1052       18           553      34         1738      48
## 14     15        1144      233           667      98         1490      82
## 16     31        1351      405           960     129         1079     101
##    PT08.S4.NO2. PT08.S5.O3.  T  RH   AH C6H6.GT.pred
## 1          1692        1268 68 377 1898     178.3743
## 2          1559         972 65 365 1729     190.0107
## 4          1584        1203 42 488 2058     185.0237
## 13         1221         472 37 469 1771     219.9328
## 14         1339         730 34 484 1812     211.6211
## 16         1583        1028 37 494 1958     185.0237
```

```
ggplot(train, aes(x = CO.GT., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red" ) +
  geom_point(data = test, aes(y = C6H6.GT.), color = "blue") +
```

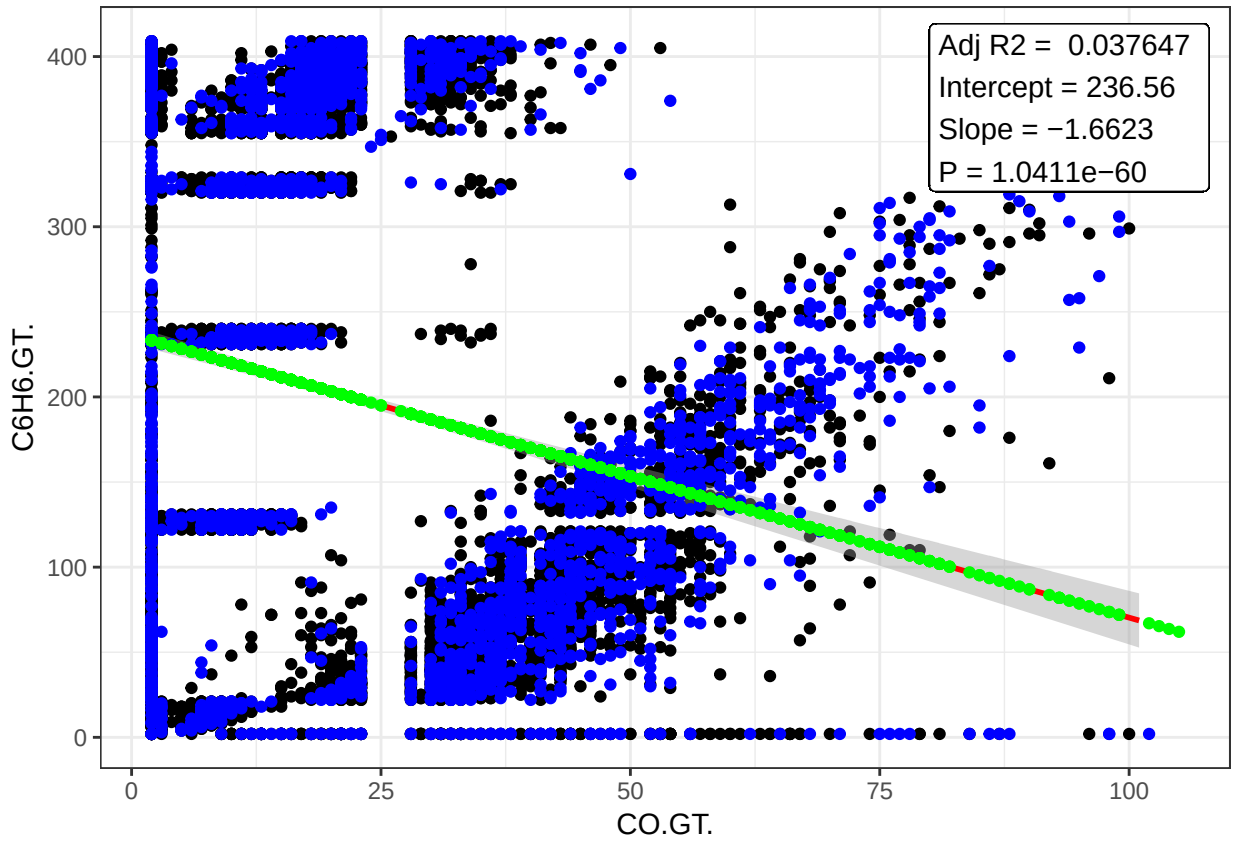
```
  theme_bw() +
  geom_label(aes(x = 80, y = 370), hjust = 0,
             label = paste("Adj R2 = ",signif(summary(new_mod)$adj.r.squared, 5),
                                                "\nIntercept =",signif(new_mod$coef[[1]],5 ),
                                                " \nSlope =",signif(new_mod$coef[[2]], 5),
                                                " \nP =",signif(summary(new_mod)$coef[2,4], 5)))
```



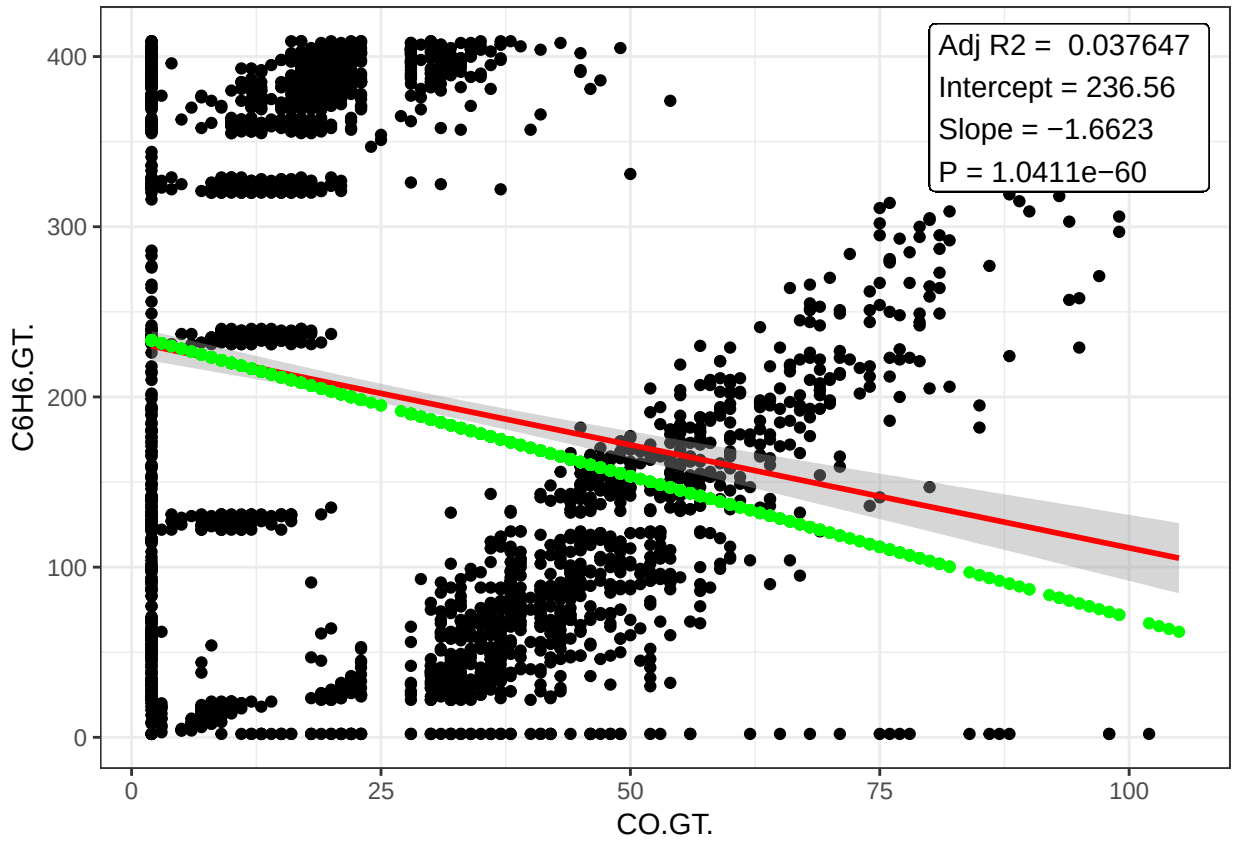

```
ggplot(train, aes(x = CO.GT., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red" ) +
  geom_point(data = test, aes(y = C6H6.GT.pred), color = "green") +
  theme_bw() +
  geom_label(aes(x = 80, y = 370), hjust = 0,
             label = paste("Adj R2 = ",signif(summary(new_mod)$adj.r.squared, 5),
                                                "\nIntercept =",signif(new_mod$coef[[1]],5 ),
                                                " \nSlope =",signif(new_mod$coef[[2]], 5),
                                                " \nP =",signif(summary(new_mod)$coef[2,4], 5)))
```




```
ggplot(train, aes(x = CO.GT., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red" ) +
  geom_point(data = test, aes(y = C6H6.GT.), color = "blue") +
  geom_point(data = test, aes(y = C6H6.GT.pred), color = "green") +
  theme_bw() +
  geom_label(aes(x = 80, y = 370), hjust = 0,
             label = paste("Adj R2 = ",signif(summary(new_mod)$adj.r.squared, 5),
                                                "\nIntercept =",signif(new_mod$coef[[1]],5 ),
                                                " \nSlope =",signif(new_mod$coef[[2]], 5),
                                                " \nP =",signif(summary(new_mod)$coef[2,4], 5)))
```




```
ggplot(test, aes(x = CO.GT., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red" ) +
  geom_point(data = test, aes(y = C6H6.GT.pred), color = "green") +
  theme_bw() +
  geom_label(aes(x = 80, y = 370), hjust = 0,
             label = paste("Adj R2 = ",signif(summary(new_mod)$adj.r.squared, 5),
                                                "\nIntercept =",signif(new_mod$coef[[1]],5 ),
                                                " \nSlope =",signif(new_mod$coef[[2]], 5),
                                                " \nP =",signif(summary(new_mod)$coef[2,4], 5)))
```




```
#Detecting multicollinearity in a regression model:
mod <- lm(C6H6.GT. ~ ., data = train)
car::vif(mod)
```

```
##        CO.GT.   PT08.S1.CO. PT08.S2.NMHC.       NOx.GT.  PT08.S3.NOx.
##      3.112173      8.297274     23.821199      7.047253      5.249414
##       NO2.GT.  PT08.S4.NO2.   PT08.S5.O3.             T            RH
##      4.974353     15.805796      8.233587      1.108072      1.929914
##            AH
##      5.676388
```

```
#Remove variables with VIF>10:

model <- lm(C6H6.GT. ~.-PT08.S2.NMHC.,-PT08.S4.NO2., data = train)
```

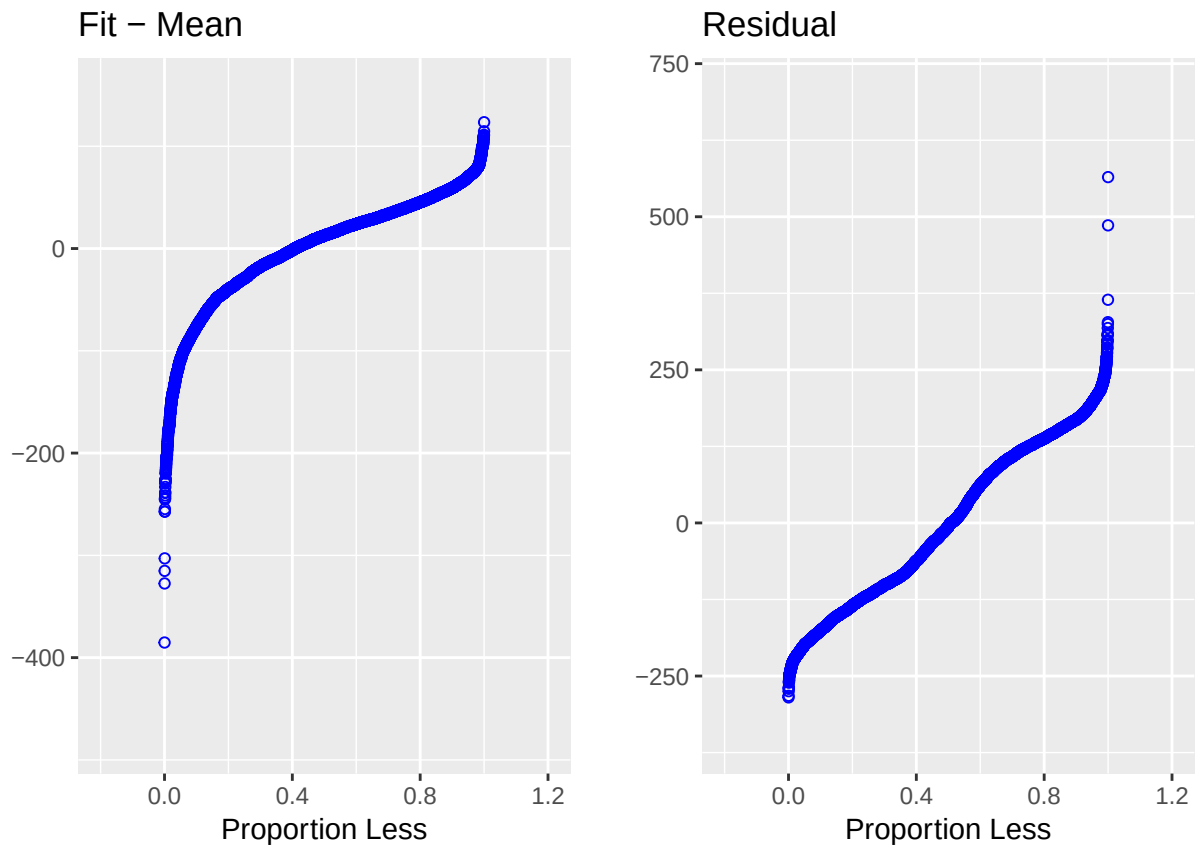
```
#variance inflation factor (VIF) is the ratio of variance in a model
#with multiple terms, divided by the variance of a model with one term alone.
#It quantifies the severity of multicollinearity in an ordinary least squares
#regression analysis. It provides an index that measures how much
#the variance (the square of the estimate's standard deviation) of
#an estimated regression coefficient is increased because of collinearity.
```

```
library(olsrr)
ols_coll_diag(model)
```

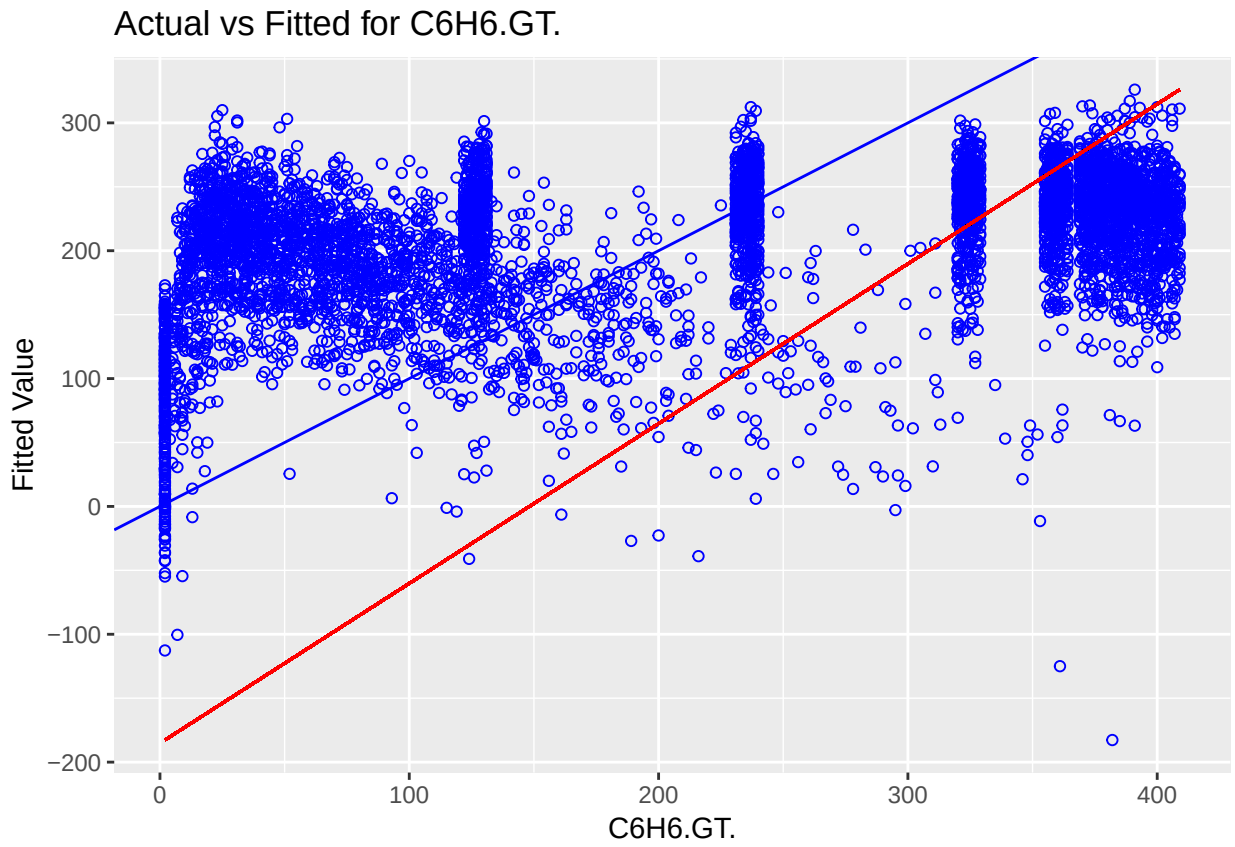
```
## Tolerance and Variance Inflation Factor
## ----------------------------------------
## # A tibble: 10 x 3
```

```
##    Variables    Tolerance   VIF
##    <chr>            <dbl> <dbl>
##  1 CO.GT.           0.304  3.29
##  2 PT08.S1.CO.      0.107  9.38
##  3 NOx.GT.          0.191  5.24
##  4 PT08.S3.NOx.     0.260  3.85
##  5 NO2.GT.          0.202  4.96
##  6 PT08.S4.NO2.     0.130  7.71
##  7 PT08.S5.O3.      0.126  7.95
##  8 T                0.918  1.09
##  9 RH               0.679  1.47
## 10 AH               0.187  5.34
##
##
## Eigenvalue and Condition Index
## ------------------------------
##     Eigenvalue Condition Index    intercept       CO.GT.  PT08.S1.CO.
## 1  9.236465226        1.000000 5.169254e-05 0.0010321475 4.578474e-05
## 2  0.788320957        3.422956 1.459173e-04 0.0335095714 1.446629e-08
## 3  0.374765201        4.964474 4.164558e-05 0.0018198727 1.387676e-05
## 4  0.234559289        6.275186 1.185163e-03 0.0007209252 1.631347e-04
## 5  0.170821991        7.353278 3.429670e-04 0.0141841103 2.899139e-04
## 6  0.098252168        9.695759 1.806065e-04 0.7426270722 4.326546e-04
## 7  0.053508593       13.138359 5.795175e-05 0.0400773583 2.349863e-03
## 8  0.026680242       18.606214 5.016115e-04 0.0725769306 5.573422e-03
## 9  0.008935166       32.151526 5.380651e-03 0.0539368442 7.360126e-02
## 10 0.005256764       41.917335 4.099577e-01 0.0289074753 8.121841e-02
## 11 0.002434404       61.596574 5.821541e-01 0.0106076924 8.363117e-01
##         NOx.GT. PT08.S3.NOx.      NO2.GT. PT08.S4.NO2.  PT08.S5.O3.
## 1  0.0006889253 0.0002179991 0.0003467213 7.849940e-05 0.0001947262
## 2  0.0301927614 0.0035096688 0.0023069710 9.750392e-05 0.0006583855
## 3  0.0001930295 0.0022689549 0.0035583134 5.424629e-04 0.0003495703
## 4  0.0026752844 0.0363938017 0.0014747695 5.056972e-06 0.0002557892
## 5  0.0250193015 0.0020555675 0.0022743617 2.700227e-03 0.0001304737
## 6  0.0525512329 0.0085017377 0.0572859853 1.015122e-04 0.0107570975
## 7  0.7331925587 0.0664713180 0.0621696226 1.925320e-03 0.0318718205
## 8  0.0146878599 0.0004658703 0.5629820858 1.416184e-02 0.2186662799
## 9  0.1217874883 0.2225405333 0.0459219193 2.306342e-01 0.7031036574
## 10 0.0028748492 0.3365001762 0.2447602518 3.955475e-01 0.0011851553
## 11 0.0161367090 0.3210743726 0.0169189983 3.542059e-01 0.0328270445
##              T          RH           AH
## 1  0.002405078 0.001299791 0.0004704972
## 2  0.022100318 0.007222149 0.0102354419
## 3  0.417116193 0.010363490 0.0497932382
## 4  0.417737496 0.002722909 0.0517013953
## 5  0.002536213 0.603928890 0.0140616441
## 6  0.042785623 0.019958857 0.0001076085
## 7  0.019544403 0.107603400 0.0574546902
## 8  0.049754604 0.032555794 0.1734107998
## 9  0.003162021 0.041552200 0.3220233860
## 10 0.009569063 0.131315849 0.0324064838
## 11 0.013288986 0.041476670 0.2883348151
```

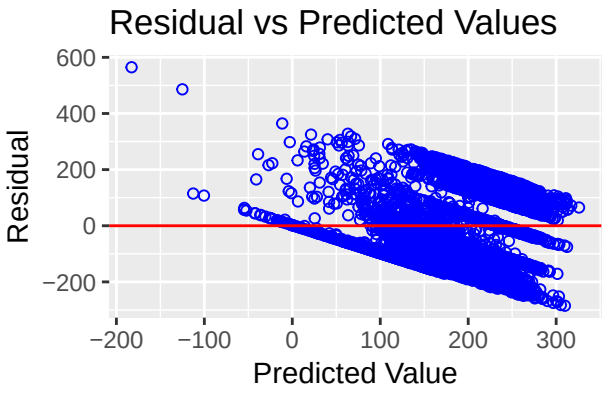
```
#Plot to detect non-linearity, influential observations and outliers.
#Consists of side-by-side quantile plots of the centered fit and the residuals.
#It shows how much variation in the data is explained by the fit and
#how much remains in the residuals. For inappropriate models,
#the spread of the residuals in such a plot is often greater than
#the spread of the centered fit.

ols_plot_resid_fit_spread(model)
```

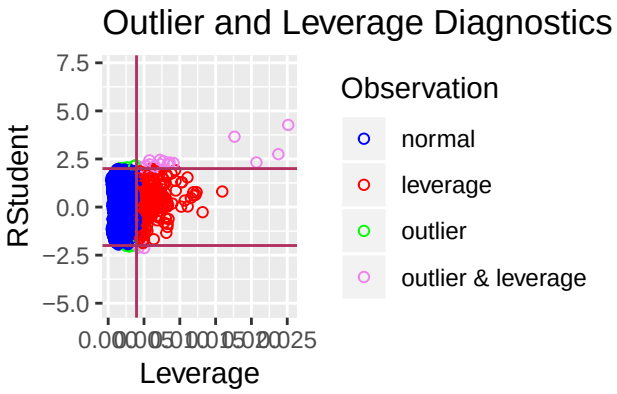


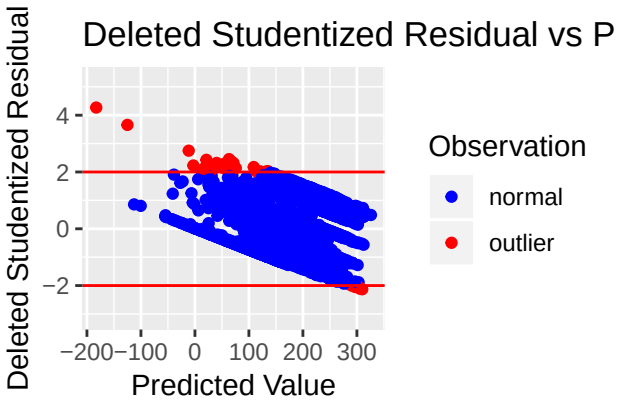

```
#Plot of observed vs fitted values to assess the fit of the model.
#Ideally, all your points should be close to a regressed diagonal line.
#Draw such a diagonal line within your graph and check out where the points lie.
#If your model had a high R Square, all the points would be close to this diagonal line.
#The lower the R Square, the weaker the Goodness of fit of your model,
#the more foggy or dispersed your points are from this diagonal line.

ols_plot_obs_fit(model)
```

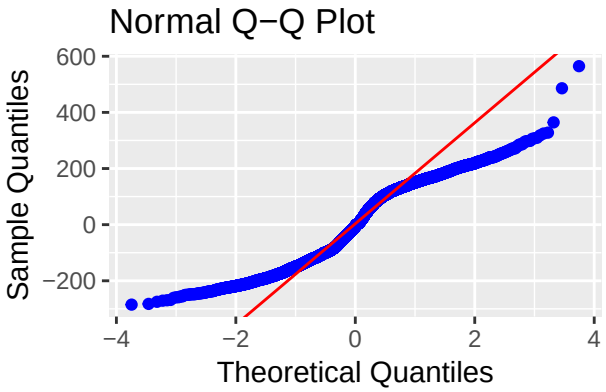



```
ols_plot_diagnostics(model)
```

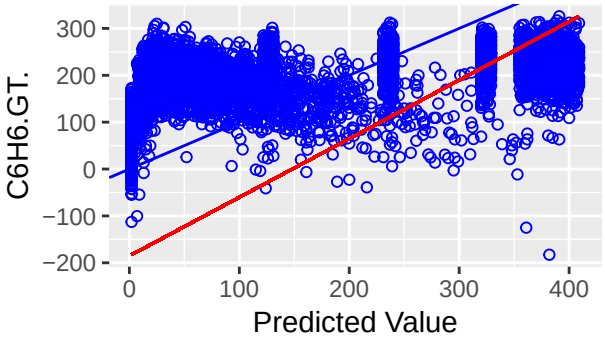

Residual vs Predicted Values
Residual
Predicted Value
600
400
200
0
−200
−200
−100
0
100
200
300


Outlier and Leverage Diagnostics
RStudent
Leverage
Observation
normal
leverage
outlier
outlier & leverage
7.5
5.0
2.5
0.0
−2.5
−5.0

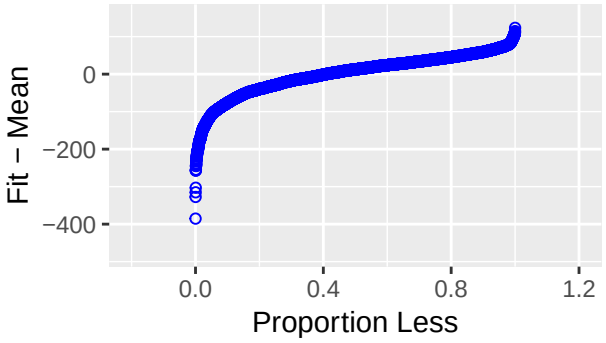

Deleted Studentized Residual vs P
Deleted Studentized Residual
Predicted Value
Observation
normal
outlier
4
2
0
−2
−200
−100
0
100
200
300


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
600
400
200
0
−200
−4
−2
0
2
4

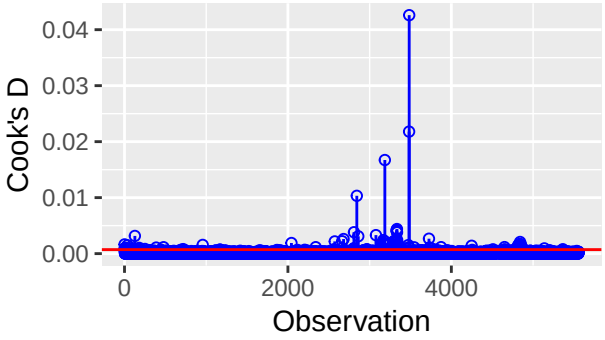
Observed by Predicted for C6H6.



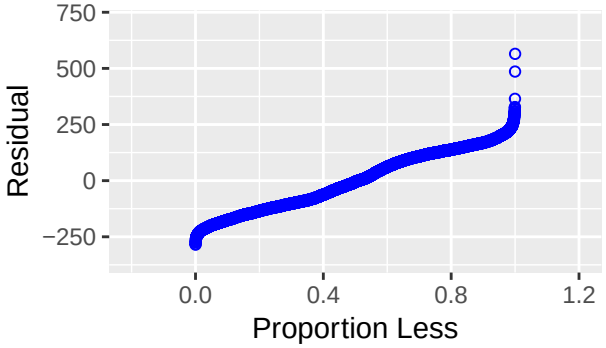
Residual Fit Spread Plot

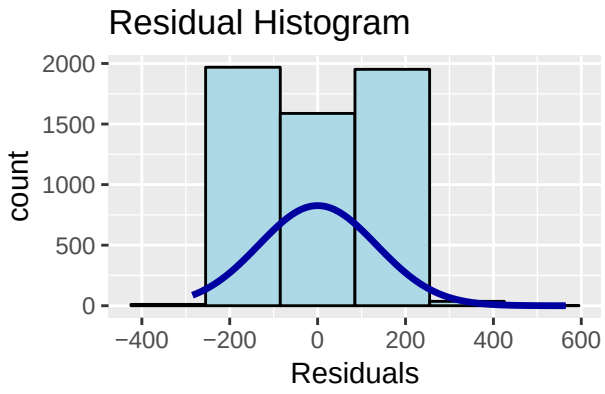


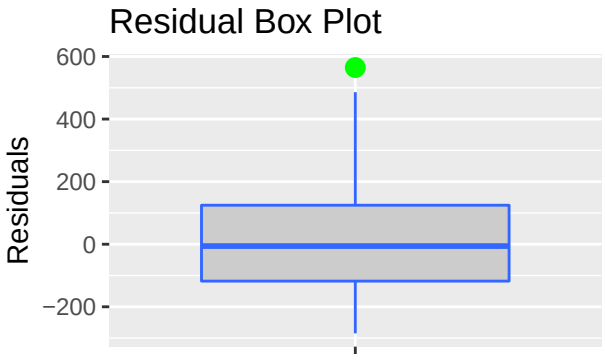
Cook's D Chart



Residual Fit Spread Plot








```
library(plotly)
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

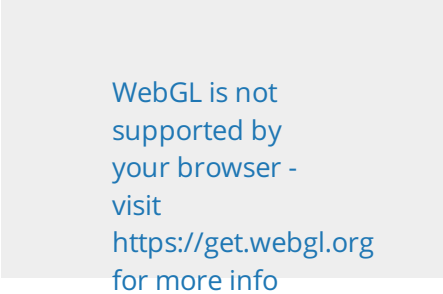
```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
plot_ly(data = test,
        z = ~C6H6.GT.,
        y = ~CO.GT.,
        x = ~PT08.S4.NO2.,
        opacity = 0.7)
```

```
## No trace type specified:
##   Based on info supplied, a 'scatter3d' trace seems appropriate.
##   Read more about this trace type -> https://plot.ly/r/reference/#scatter3d
```

```
## No scatter3d mode specifed:
##   Setting the mode to markers
```

```
##   Read more about this attribute -> https://plot.ly/r/reference/#scatter-mode
```

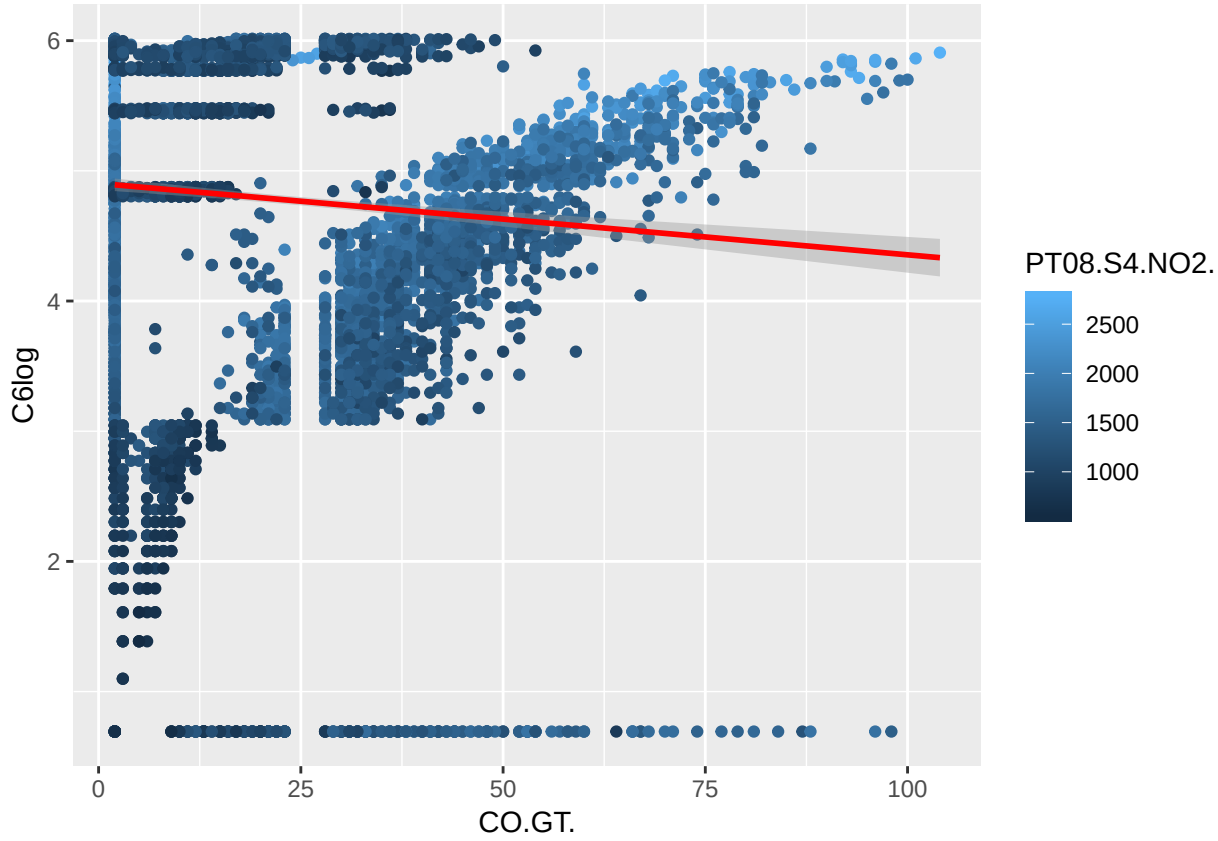



```
#let's try to use non-linear transformation for the chosen variable C6H6.GT.
# It seems that log transformations might help:

cdt$C6log <- log(cdt$C6H6.GT.)

library(caTools)
set.seed(42)
sample.log <- sample.split(cdt, SplitRatio = 0.75)
train.log <- subset(cdt, sample.log == TRUE)
test.log  <- subset(cdt, sample.log == FALSE)


ln <- lm(C6log ~ ., data = train.log)
summary(ln)
```

```
##
## Call:
## lm(formula = C6log ~ ., data = train.log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.37610 -0.28837  0.04081  0.35270  2.03414
##
```

```
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.249e+00  1.195e-01  27.192  < 2e-16 ***
## CO.GT.         2.238e-03  6.636e-04   3.372 0.000751 ***
## PT08.S1.CO.   -8.728e-04  9.278e-05  -9.408  < 2e-16 ***
## C6H6.GT.       7.300e-03  5.116e-05 142.695  < 2e-16 ***
## PT08.S2.NMHC. -4.236e-04  1.310e-04  -3.233 0.001231 **
## NOx.GT.       -3.683e-04  9.105e-05  -4.045 5.29e-05 ***
## PT08.S3.NOx.  -7.436e-04  6.148e-05 -12.094  < 2e-16 ***
## NO2.GT.        2.046e-03  3.185e-04   6.423 1.43e-10 ***
## PT08.S4.NO2.   7.248e-04  8.271e-05   8.764  < 2e-16 ***
## PT08.S5.O3.    1.620e-04  5.099e-05   3.177 0.001494 **
## T              1.004e-03  5.953e-05  16.873  < 2e-16 ***
## RH             1.141e-03  5.230e-05  21.808  < 2e-16 ***
## AH             1.678e-05  8.473e-06   1.980 0.047738 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5562 on 6465 degrees of freedom
## Multiple R-squared:  0.8206, Adjusted R-squared:  0.8203
## F-statistic:  2465 on 12 and 6465 DF,  p-value: < 2.2e-16
```

```
ggplot(data = train.log, aes(x = CO.GT., y = C6log, color = PT08.S4.NO2.)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red")
```



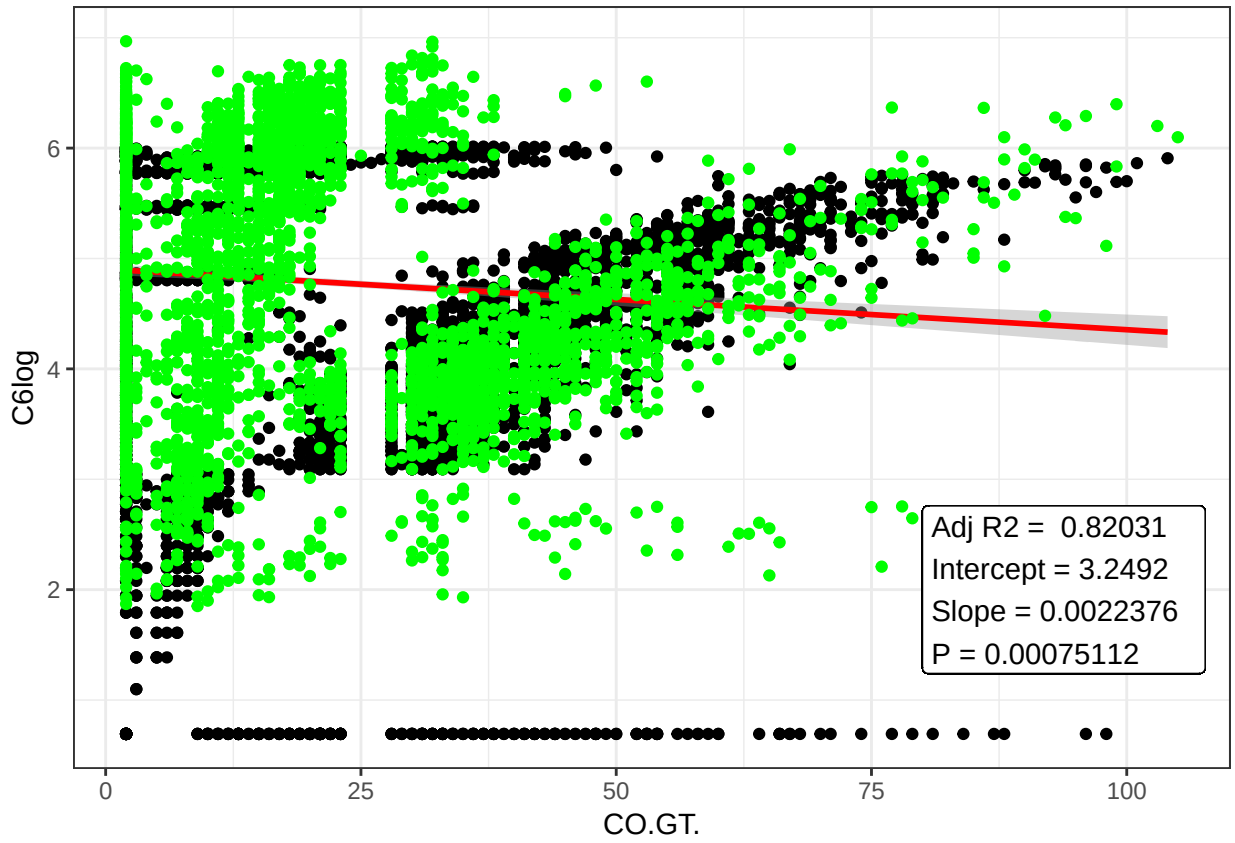
```
predlog <- predict(ln, newdata = test.log)
head(predlog)
```

```
##        1        2        4       13       14       15
## 3.343592 5.804319 6.001285 2.607217 4.431830 5.810066
```

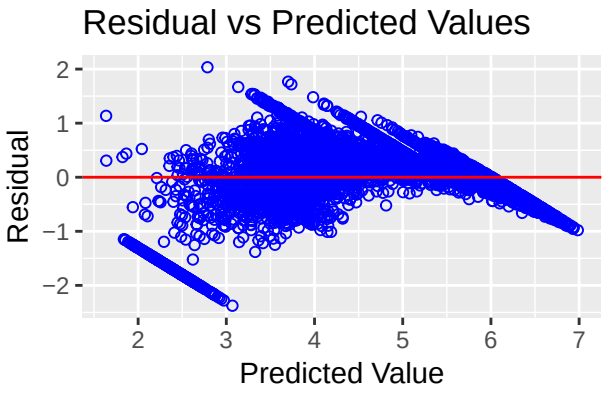
```
test.log$C6H6.GT.pred <- predlog
head(test.log)
```

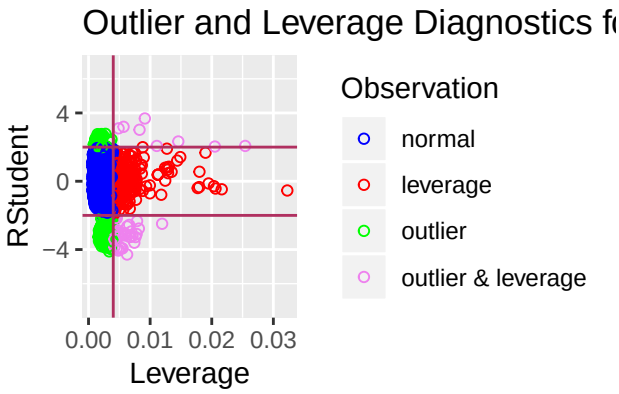
```
##    CO.GT. PT08.S1.CO. C6H6.GT. PT08.S2.NMHC. NOx.GT. PT08.S3.NOx. NO2.GT.
## 1      35        1360       41          1046     166         1056     113
## 2      28        1292      404           955     103         1174      92
## 4      31        1376      402           948     172         1092     122
## 13     10        1052       18           553      34         1738      48
## 14     15        1144      233           667      98         1490      82
## 15     28        1333      390           900     174         1136     112
##    PT08.S4.NO2. PT08.S5.O3.  T  RH   AH    C6log C6H6.GT.pred
## 1          1692        1268 68 377 1898 3.713572     3.343592
## 2          1559         972 65 365 1729 6.001415     5.804319
## 4          1584        1203 42 488 2058 5.996452     6.001285
## 13         1221         472 37 469 1771 2.890372     2.607217
## 14         1339         730 34 484 1812 5.451038     4.431830
## 15         1517        1102 40 462 1804 5.966147     5.810066
```

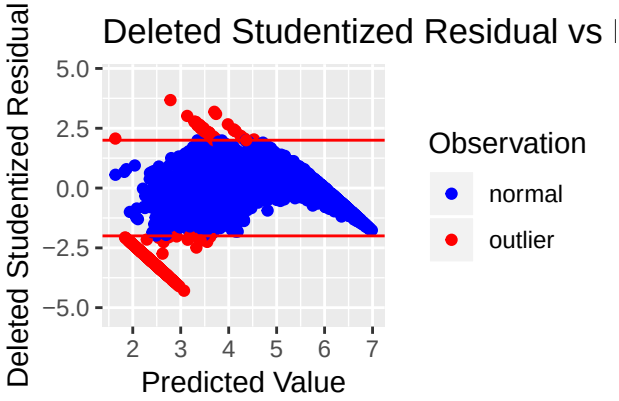
```
ggplot(train.log, aes(x = CO.GT., y = C6log)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red" ) +
  geom_point(data = test.log, aes(y = C6H6.GT.pred), color = "green") +
  theme_bw() +
  geom_label(aes(x = 80, y = 2), hjust = 0,
             label = paste("Adj R2 = ",signif(summary(ln)$adj.r.squared, 5),
                                              "\nIntercept =",signif(ln$coef[[1]],5 ),
                                              " \nSlope =",signif(ln$coef[[2]], 5),
                                              " \nP =",signif(summary(ln)$coef[2,4], 5)))
```

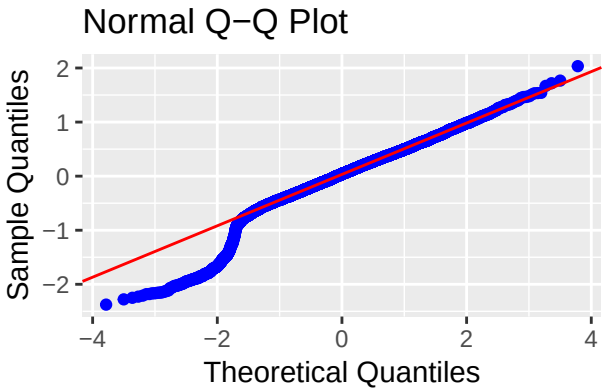


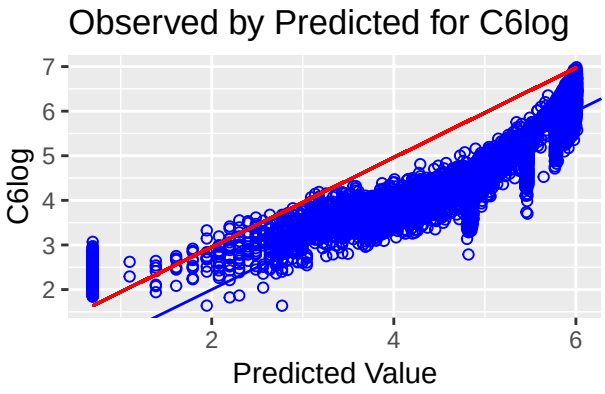

```
ols_plot_diagnostics(ln)
```

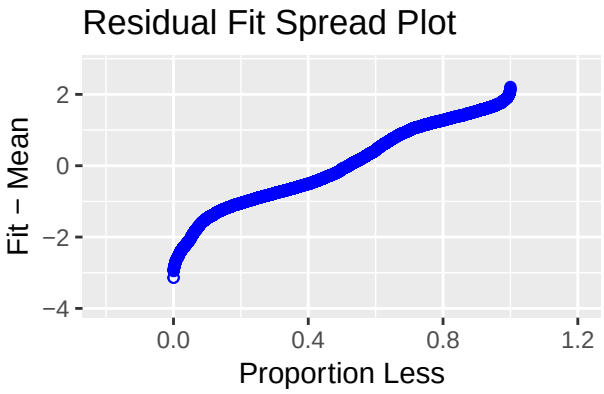

Residual vs Predicted Values
Residual
Predicted Value

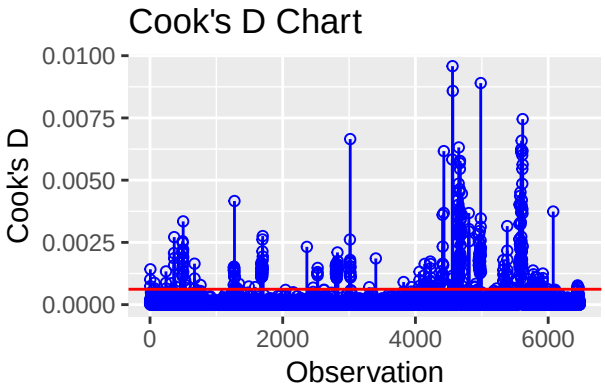

Outlier and Leverage Diagnostics f
RStudent
Leverage
Observation
normal
leverage
outlier
outlier & leverage

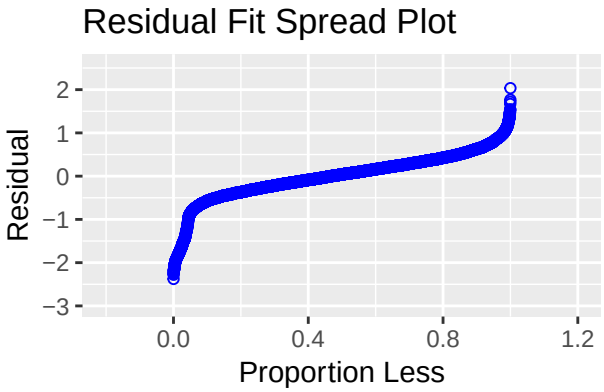

Deleted Studentized Residual vs
Deleted Studentized Residual
Predicted Value
Observation
normal
outlier

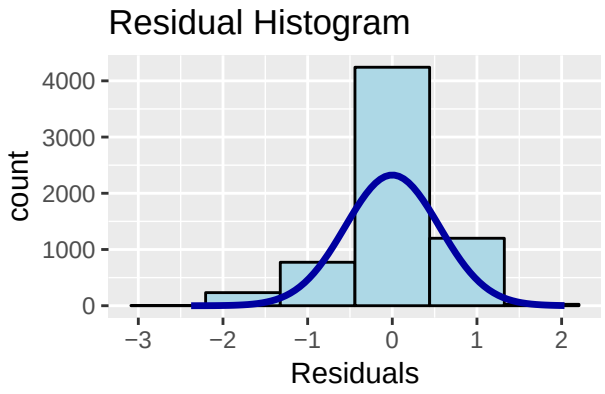

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles

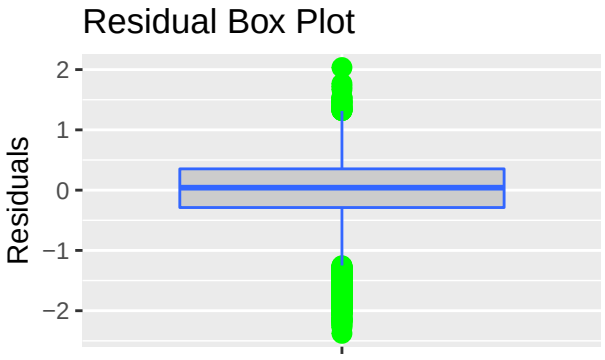

Observed by Predicted for C6log
C6log
Predicted Value
7
6
5
4
3
2
2
4
6


Residual Fit Spread Plot
Fit – Mean
Proportion Less
2
0
−2
−4
0.0
0.4
0.8
1.2


Cook's D Chart
Cook's D
Observation
0.0100
0.0075
0.0050
0.0025
0.0000
0
2000
4000
6000


Residual Fit Spread Plot
Residual
Proportion Less
2
1
0
−1
−2
−3
0.0
0.4
0.8
1.2


Residual Histogram
count
4000
3000
2000
1000
0
−3
−2
−1
0
1
2
Residuals


Residual Box Plot
Residuals
2
1
0
−1
−2