HW11

Oksana Ivanova

```
data <- read.csv("~/Documents/ITMO/2sem/R_stat/R_ITMO/AirQualityUCI.csv", header = TRUE, sep = ";", dec=
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    2.0        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.       : num  2.6 2 2.2 2.2 1.6 1.2 1.2 1 0.9 0.6 ...
##  $ 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.     : num  11.9 9.4 9 9.2 6.5 4.7 3.6 3.3 2.3 1.7 ...
##  $ 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            : num  13.6 13.3 11.9 11 11.2 11.2 11.3 10.7 10.7 10.3 ...
##  $ RH           : num  48.9 47.7 54 60 59.6 59.2 56.8 60 59.7 60.2 ...
##  $ AH           : num  0.758 0.726 0.75 0.787 0.789 ...
##  $ 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   Min.   :-200.00   Min.   :-200
```

```
##  01/01/2005:  24   01.00.00: 390   1st Qu.:   0.60   1st Qu.: 921
##  01/02/2005:  24   02.00.00: 390   Median :   1.50   Median :1053
##  01/03/2005:  24   03.00.00: 390   Mean   : -34.21   Mean   :1049
##  01/04/2004:  24   04.00.00: 390   3rd Qu.:   2.60   3rd Qu.:1221
##  01/04/2005:  24   05.00.00: 390   Max.   :  11.90   Max.   :2040
##  (Other)   :9237   (Other) :7131   NA's   :114       NA's   :114
##     NMHC.GT.          C6H6.GT.         PT08.S2.NMHC.       NOx.GT.
##  Min.   :-200.0   Min.   :-200.000   Min.   :-200.0   Min.   :-200.0
##  1st Qu.:-200.0   1st Qu.:   4.000   1st Qu.: 711.0   1st Qu.:  50.0
##  Median :-200.0   Median :   7.900   Median : 895.0   Median : 141.0
##  Mean   :-159.1   Mean   :   1.866   Mean   : 894.6   Mean   : 168.6
##  3rd Qu.:-200.0   3rd Qu.:  13.600   3rd Qu.:1105.0   3rd Qu.: 284.0
##  Max.   :1189.0   Max.   :  63.700   Max.   :2214.0   Max.   :1479.0
##  NA's   :114      NA's   :114        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
##  Min.   :-200.000   Min.   :-200.00   Min.   :-200.0000   Mode:logical
##  1st Qu.:  10.900   1st Qu.:  34.10   1st Qu.:   0.6923   NA's:9471
##  Median :  17.200   Median :  48.60   Median :   0.9768
##  Mean   :   9.778   Mean   :  39.49   Mean   :  -6.8376
##  3rd Qu.:  24.100   3rd Qu.:  61.90   3rd Qu.:   1.2962
##  Max.   :  44.600   Max.   :  88.70   Max.   :   2.2310
##  NA's   :114        NA's   :114       NA's   :114
##    X.1
##  Mode:logical
##  NA's:9471
##
##
##
##
##
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
dt <- data %>%
  select(-c(X, X.1)) %>%
  na.omit()
```

```
head(dt)
```

```
##         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    2.0        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
```

```
summary(dt)
```

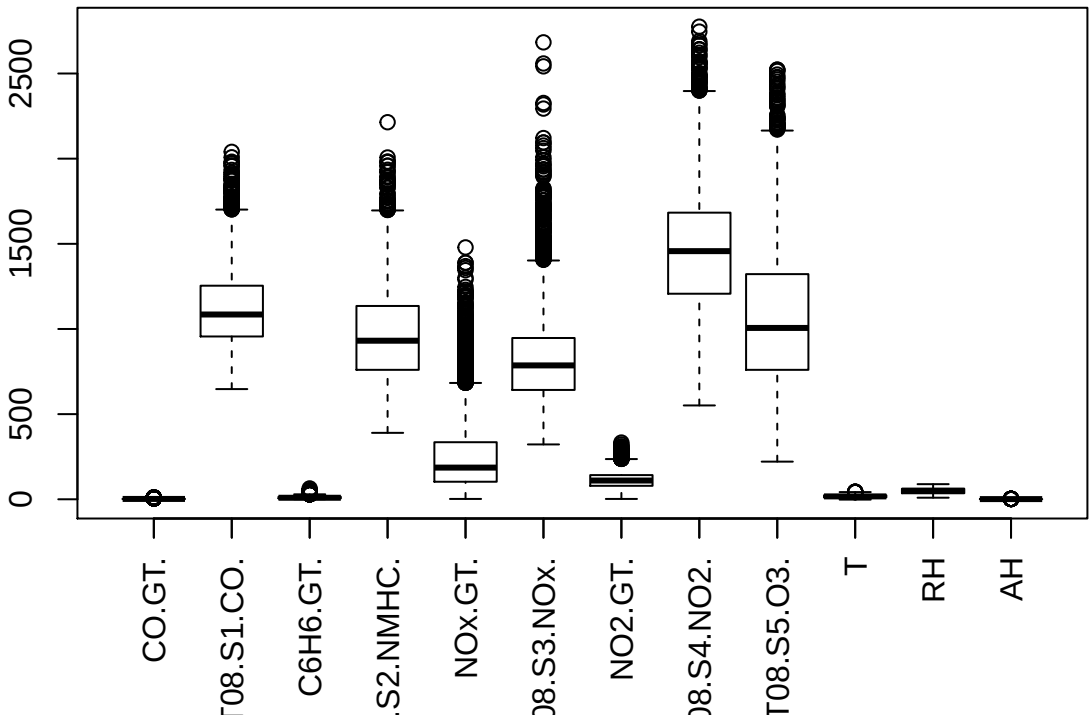
```
##          Date             Time          CO.GT.          PT08.S1.CO.
##  01/01/2005:  24   00.00.00: 390   Min.   :-200.00   Min.   :-200
##  01/02/2005:  24   01.00.00: 390   1st Qu.:   0.60   1st Qu.: 921
##  01/03/2005:  24   02.00.00: 390   Median :   1.50   Median :1053
##  01/04/2004:  24   03.00.00: 390   Mean   : -34.21   Mean   :1049
##  01/04/2005:  24   04.00.00: 390   3rd Qu.:   2.60   3rd Qu.:1221
##  01/05/2004:  24   05.00.00: 390   Max.   :  11.90   Max.   :2040
##  (Other)   :9213   (Other) :7017
##     NMHC.GT.         C6H6.GT.        PT08.S2.NMHC.       NOx.GT.
##  Min.   :-200.0   Min.   :-200.000   Min.   :-200.0   Min.   :-200.0
##  1st Qu.:-200.0   1st Qu.:   4.000   1st Qu.: 711.0   1st Qu.:  50.0
##  Median :-200.0   Median :   7.900   Median : 895.0   Median : 141.0
##  Mean   :-159.1   Mean   :   1.866   Mean   : 894.6   Mean   : 168.6
##  3rd Qu.:-200.0   3rd Qu.:  13.600   3rd Qu.:1105.0   3rd Qu.: 284.0
##  Max.   :1189.0   Max.   :  63.700   Max.   :2214.0   Max.   :1479.0
##
##   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
##
##        T                RH               AH
##  Min.   :-200.000   Min.   :-200.00   Min.   :-200.0000
##  1st Qu.:  10.900   1st Qu.:  34.10   1st Qu.:   0.6923
##  Median :  17.200   Median :  48.60   Median :   0.9768
##  Mean   :   9.778   Mean   :  39.49   Mean   :  -6.8376
##  3rd Qu.:  24.100   3rd Qu.:  61.90   3rd Qu.:   1.2962
##  Max.   :  44.600   Max.   :  88.70   Max.   :   2.2310
##
```

Here is a lot of -200 and -200.0 values that we replace with NA's

```
#column NMHC.GT. is almost NAs values so let's get rid of it:
dt$NMHC.GT. <- NULL
dt <- na_if(dt, -200) %>%
  na.omit()
summary(dt)
```

```
##          Date             Time          CO.GT.          PT08.S1.CO.
##  02/04/2005:  24   10.00.00: 312   Min.   : 0.100   Min.   : 647
##  03/04/2005:  24   20.00.00: 310   1st Qu.: 1.100   1st Qu.: 956
##  15/03/2005:  24   09.00.00: 309   Median : 1.900   Median :1085
##  16/03/2005:  24   12.00.00: 309   Mean   : 2.182   Mean   :1120
##  18/03/2005:  24   18.00.00: 309   3rd Qu.: 2.900   3rd Qu.:1254
##  19/03/2005:  24   21.00.00: 309   Max.   :11.900   Max.   :2040
##  (Other)   :6797   (Other) :5083
##     C6H6.GT.      PT08.S2.NMHC.       NOx.GT.         PT08.S3.NOx.
##  Min.   : 0.20   Min.   : 390.0   Min.   :   2.0   Min.   : 322.0
##  1st Qu.: 4.90   1st Qu.: 760.0   1st Qu.: 103.0   1st Qu.: 642.0
##  Median : 8.80   Median : 931.0   Median : 186.0   Median : 786.0
##  Mean   :10.55   Mean   : 958.5   Mean   : 250.7   Mean   : 816.9
##  3rd Qu.:14.60   3rd Qu.:1135.0   3rd Qu.: 335.0   3rd Qu.: 947.0
##  Max.   :63.70   Max.   :2214.0   Max.   :1479.0   Max.   :2683.0
##
##     NO2.GT.        PT08.S4.NO2.    PT08.S5.O3.          T
##  Min.   :  2.0   Min.   : 551   Min.   : 221   Min.   :-1.90
##  1st Qu.: 79.0   1st Qu.:1207   1st Qu.: 760   1st Qu.:11.20
##  Median :110.0   Median :1457   Median :1006   Median :16.80
##  Mean   :113.9   Mean   :1453   Mean   :1058   Mean   :17.76
##  3rd Qu.:142.0   3rd Qu.:1683   3rd Qu.:1322   3rd Qu.:23.70
##  Max.   :333.0   Max.   :2775   Max.   :2523   Max.   :44.60
##
##        RH              AH
##  Min.   : 9.20   Min.   :0.1847
##  1st Qu.:35.30   1st Qu.:0.6941
##  Median :49.20   Median :0.9539
##  Mean   :48.88   Mean   :0.9856
##  3rd Qu.:62.20   3rd Qu.:1.2516
##  Max.   :88.70   Max.   :2.1806
##
```

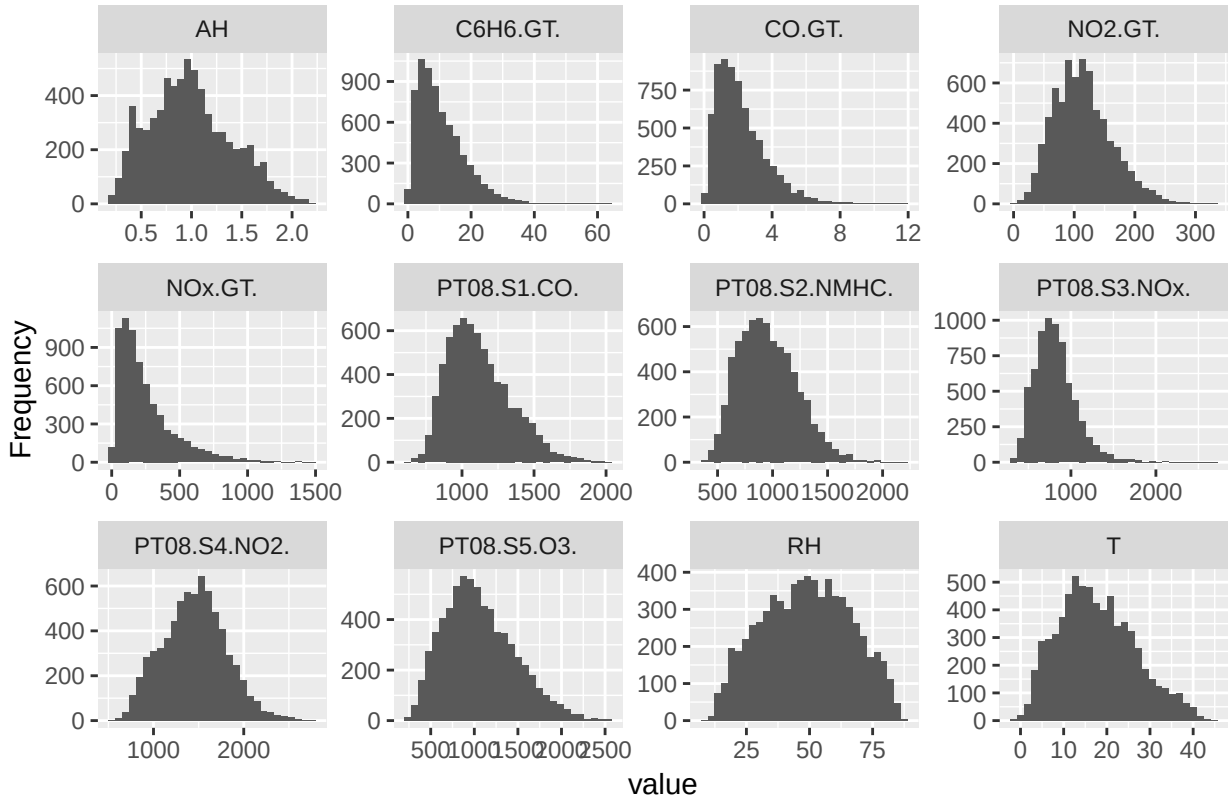
```
library(ggplot2)
dt[,-c(1:2)] %>%  boxplot(names = FALSE)
axis(1, labels=names(dt[,-c(1:2)]), at=1:12, las=2)
```



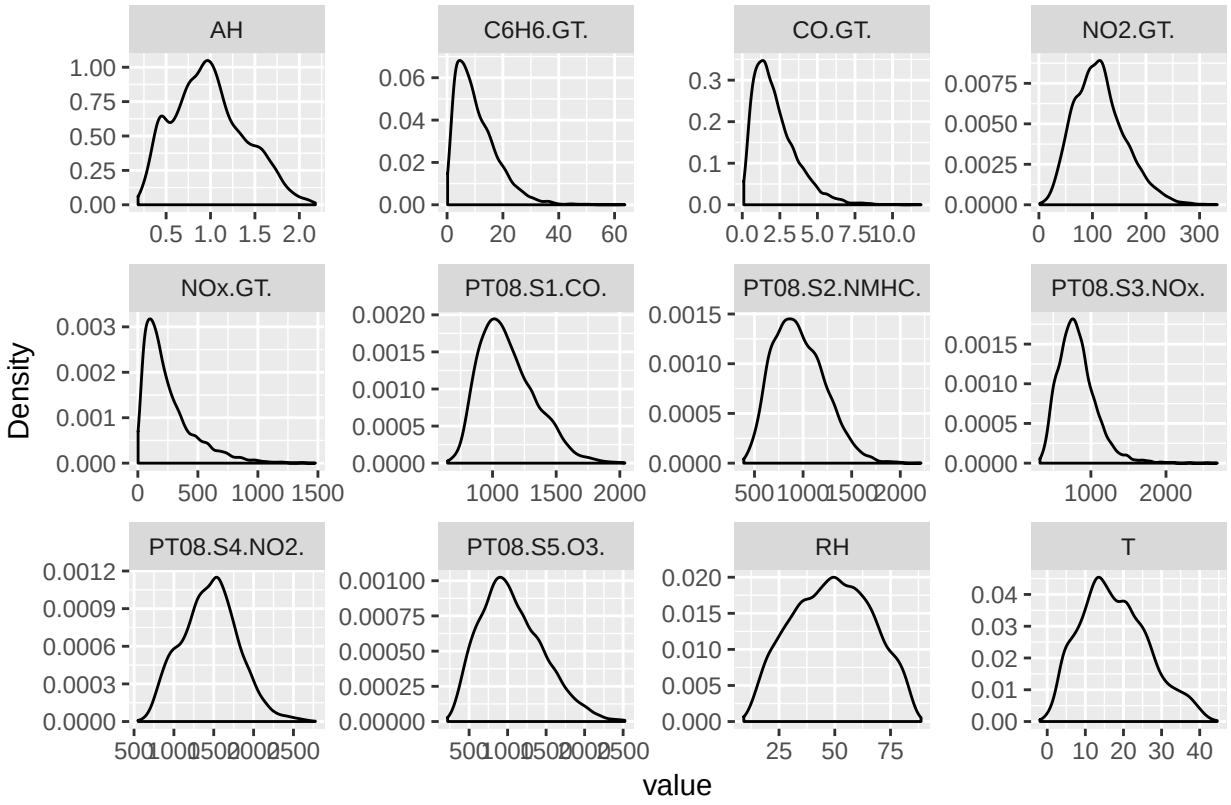

```
# To see the data distribution:
library(DataExplorer)
```

```
## Warning: package 'DataExplorer' was built under R version 3.5.2
```

```
plot_histogram(dt)
```



```
plot_density(dt)
```




```
#Looks OK, without outliers
```

```
dt[, c(3:14)] <- lapply(dt[,c(3:14)], as.numeric)
str(dt)
```

```
## 'data.frame':    6941 obs. of  14 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 7 ...
##  $ CO.GT.       : num  2.6 2 2.2 2.2 1.6 1.2 1.2 1 0.9 0.7 ...
##  $ PT08.S1.CO.  : num  1360 1292 1402 1376 1272 ...
##  $ C6H6.GT.     : num  11.9 9.4 9 9.2 6.5 4.7 3.6 3.3 2.3 1.1 ...
##  $ PT08.S2.NMHC.: num  1046 955 939 948 836 ...
##  $ NOx.GT.      : num  166 103 131 172 131 89 62 62 45 16 ...
##  $ PT08.S3.NOx. : num  1056 1174 1140 1092 1205 ...
##  $ NO2.GT.      : num  113 92 114 122 116 96 77 76 60 28 ...
##  $ PT08.S4.NO2. : num  1692 1559 1555 1584 1490 ...
##  $ PT08.S5.O3.  : num  1268 972 1074 1203 1110 ...
##  $ T            : num  13.6 13.3 11.9 11 11.2 11.2 11.3 10.7 10.7 11 ...
##  $ RH           : num  48.9 47.7 54 60 59.6 59.2 56.8 60 59.7 56.2 ...
##  $ AH           : num  0.758 0.726 0.75 0.787 0.789 ...
##  - attr(*, "na.action")= 'omit' Named int  10 11 34 35 40 58 59 82 83 106 ...
##   ..- attr(*, "names")= chr  "10" "11" "34" "35" ...
```

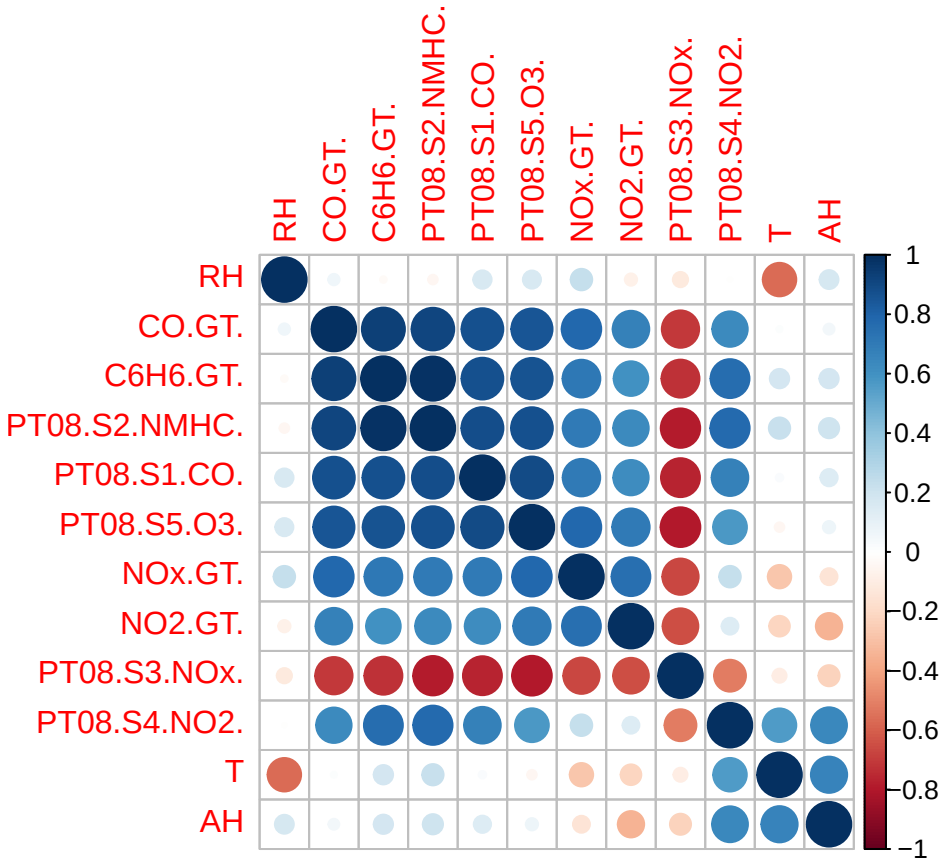
```
#also create data.frame with only numeric data (exclude Date and Time)
dt_new <- dt[,-c(1:2)]
```

OK, this is tidy data. Lets investigate dependences/correlations.

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

```
## corrplot 0.84 loaded
```

```
dt_new %>%  cor(use="pairwise.complete.obs") %>% corrplot(order = "hclust")
```

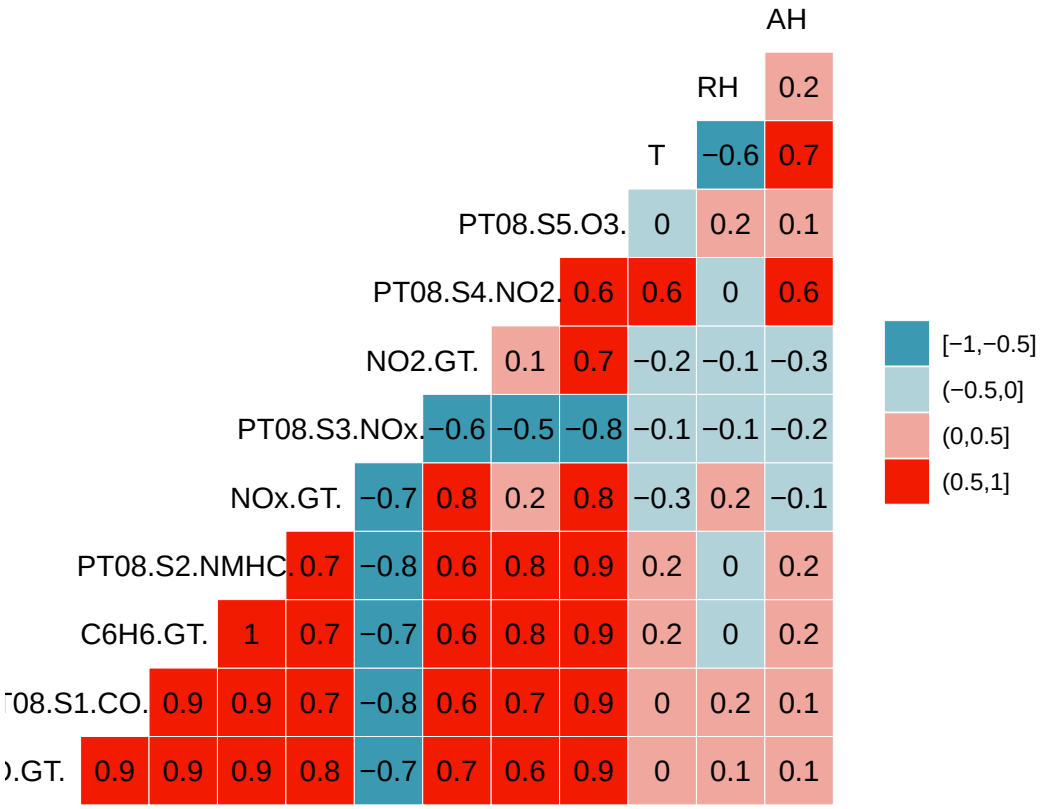



```
#Enother way
library(GGally)
```

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

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

```
ggcorr(dt_new, nbreaks = 4,
       label = TRUE,
       hjust = 0.8)
```




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

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

```
## Start:  AIC=1195.49
## 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
## <none>                       8217.2 1195.5
## - PT08.S1.CO.    1       3.0  8220.1 1196.0
## - PT08.S5.O3.    1       9.7  8226.8 1201.7
## - PT08.S4.NO2.   1      66.9  8284.1 1249.8
## - AH             1      69.1  8286.2 1251.6
## - RH             1     201.8  8418.9 1361.9
## - T              1     238.2  8455.4 1391.8
## - NOx.GT.        1     401.7  8618.8 1524.7
## - NO2.GT.        1     427.6  8644.7 1545.6
## - PT08.S3.NOx.   1     612.1  8829.3 1692.2
## - CO.GT.         1     805.2  9022.3 1842.3
## - PT08.S2.NMHC.  1    9830.6 18047.7 6654.6
```

```
summary(selectedMod)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ CO.GT. + PT08.S1.CO. + PT08.S2.NMHC. +
```

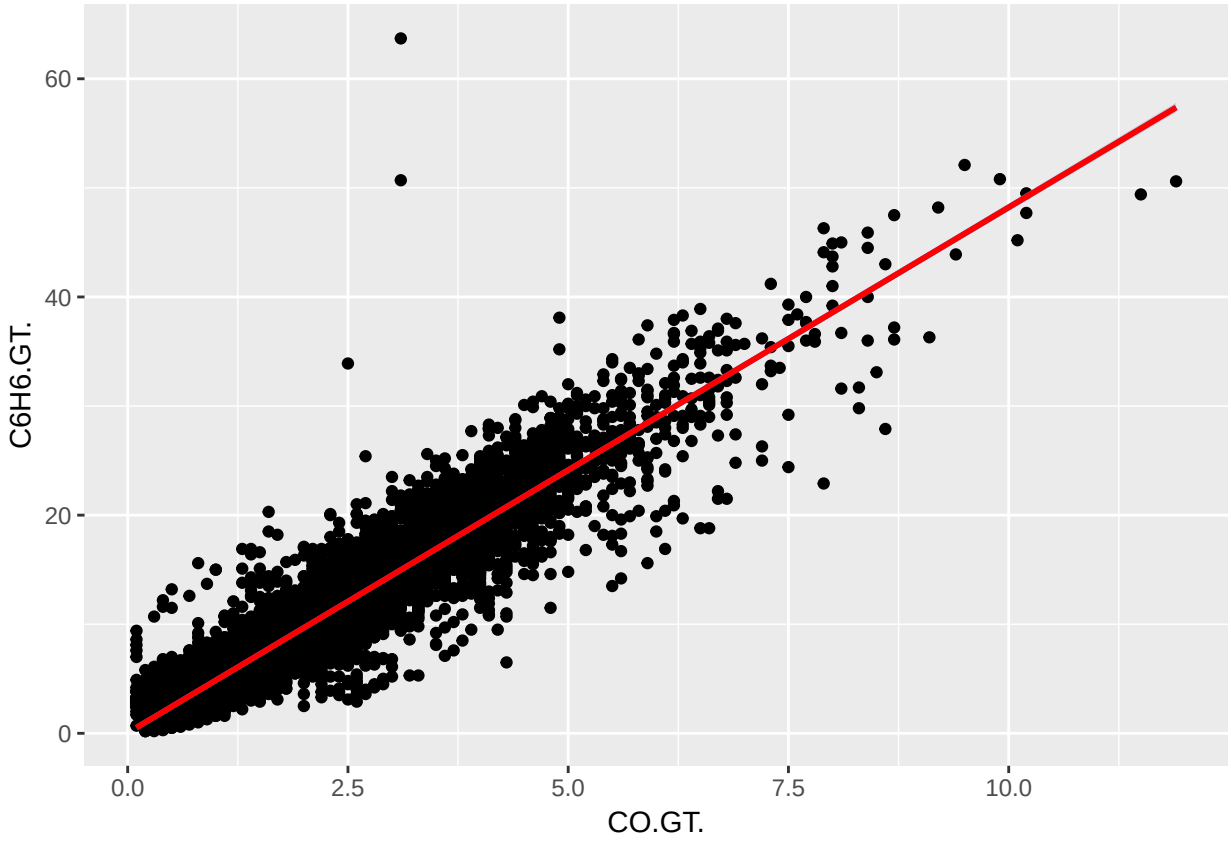
```
##     NOx.GT. + PT08.S3.NOx. + NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. +
##     T + RH + AH, data = dt_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3188 -0.6809 -0.1593  0.5259 21.6256
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.613e+01  2.599e-01 -62.068  < 2e-16 ***
## CO.GT.         7.534e-01  2.891e-02  26.057  < 2e-16 ***
## PT08.S1.CO.    2.791e-04  1.759e-04   1.587  0.11257
## PT08.S2.NMHC.  2.461e-02  2.703e-04  91.047  < 2e-16 ***
## NOx.GT.        3.192e-03  1.734e-04  18.404  < 2e-16 ***
## PT08.S3.NOx.   2.579e-03  1.135e-04  22.719  < 2e-16 ***
## NO2.GT.       -1.101e-02  5.797e-04 -18.988  < 2e-16 ***
## PT08.S4.NO2.   1.188e-03  1.582e-04   7.511 6.62e-14 ***
## PT08.S5.O3.   -2.712e-04  9.489e-05  -2.858  0.00427 **
## T             -7.836e-02  5.529e-03 -14.172  < 2e-16 ***
## RH            -2.802e-02  2.148e-03 -13.044  < 2e-16 ***
## AH             8.174e-01  1.071e-01   7.632 2.62e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.089 on 6929 degrees of freedom
## Multiple R-squared:  0.9788, Adjusted R-squared:  0.9787
## F-statistic: 2.902e+04 on 11 and 6929 DF,  p-value: < 2.2e-16
```

```
#Try to understand degree of linearity between RH output and other input features
#plot all X-features against output variable C6H6.GT:
lm1 <- lm(C6H6.GT. ~ CO.GT., data = dt_new)
summary(lm1)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ CO.GT., data = dt_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.198  -1.585  -0.158   1.497  48.725
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.04054    0.05977   0.678    0.498
## CO.GT.       4.81746    0.02286 210.782   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.744 on 6939 degrees of freedom
## Multiple R-squared:  0.8649, Adjusted R-squared:  0.8649
## F-statistic: 4.443e+04 on 1 and 6939 DF,  p-value: < 2.2e-16
```

```
ggplot(lm1, aes(x = CO.GT., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = 'lm') +
```

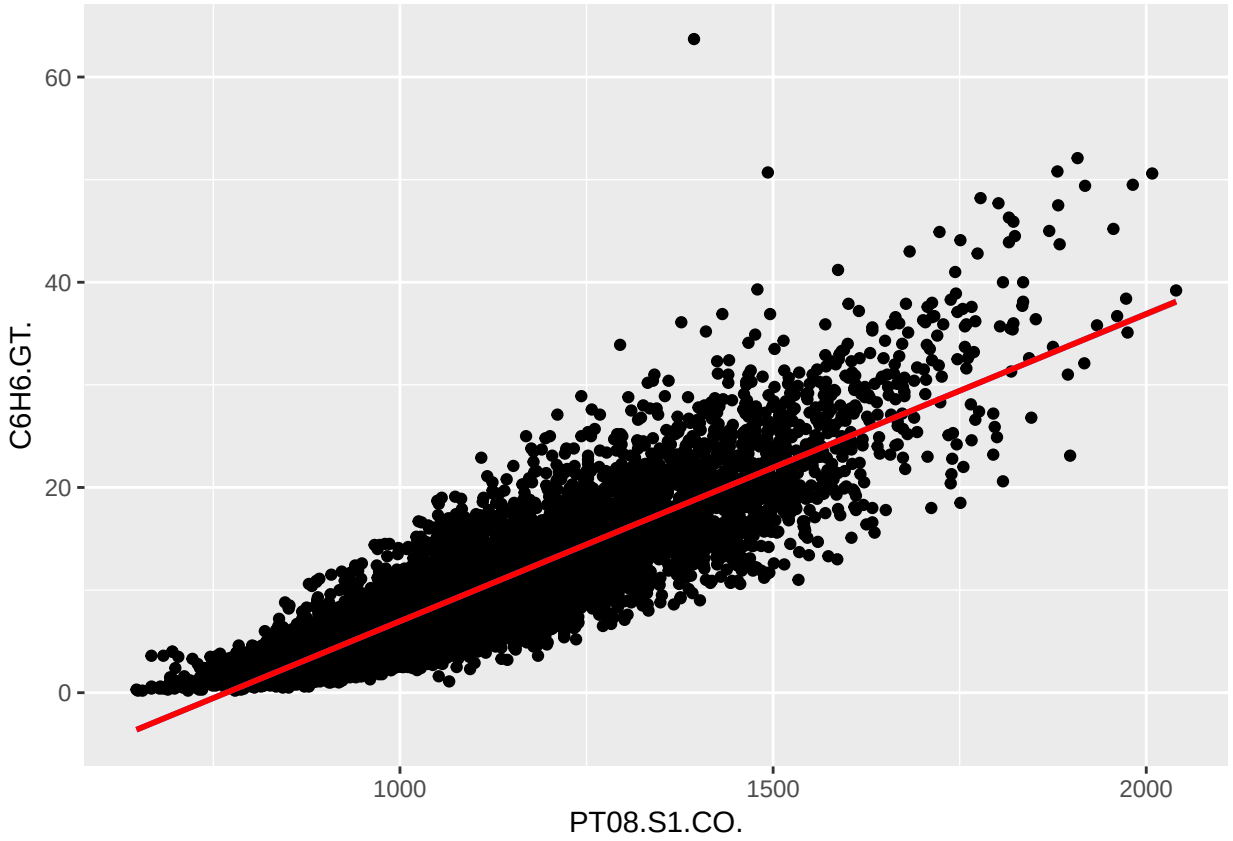
```
  geom_line(aes(y = .fitted), color = "red", size = 1)
```




```
lm2 <- lm(C6H6.GT. ~ PT08.S1.CO., data = dt_new)
summary(lm2)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S1.CO., data = dt_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.955  -2.260  -0.183   1.955  44.938
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.298e+01  2.243e-01  -102.5   <2e-16 ***
## PT08.S1.CO.  2.995e-02  1.965e-04   152.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.581 on 6939 degrees of freedom
## Multiple R-squared:  0.7699, Adjusted R-squared:  0.7699
## F-statistic: 2.322e+04 on 1 and 6939 DF,  p-value: < 2.2e-16
```

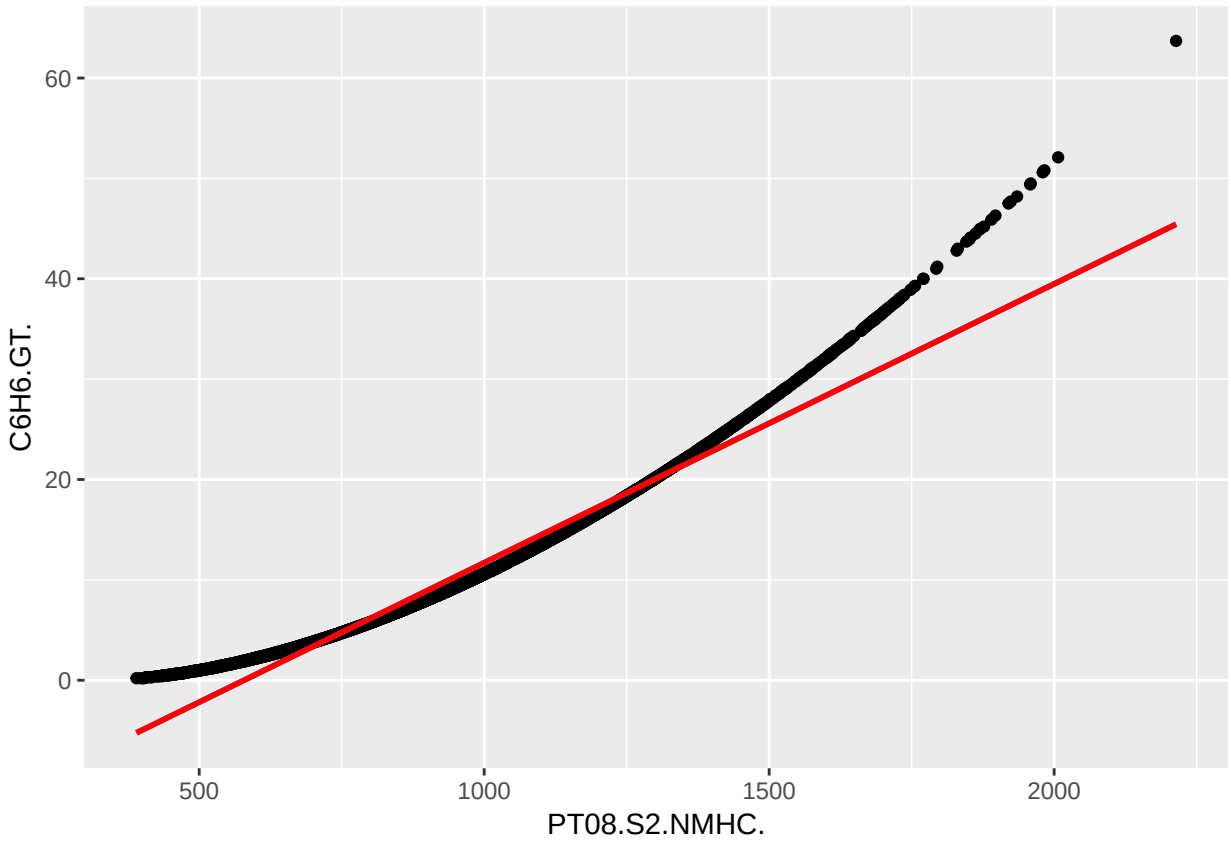
```
ggplot(lm2, aes(x = PT08.S1.CO., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = lm) +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```




```
lm3 <- lm(C6H6.GT. ~ PT08.S2.NMHC., data = dt_new)
summary(lm3)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S2.NMHC., data = dt_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1674 -0.9614 -0.4998  0.5063 18.2658
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.608e+01  6.249e-02  -257.3   <2e-16 ***
## PT08.S2.NMHC.  2.778e-02  6.285e-05   442.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.383 on 6939 degrees of freedom
## Multiple R-squared:  0.9657, Adjusted R-squared:  0.9657
## F-statistic: 1.954e+05 on 1 and 6939 DF,  p-value: < 2.2e-16
```

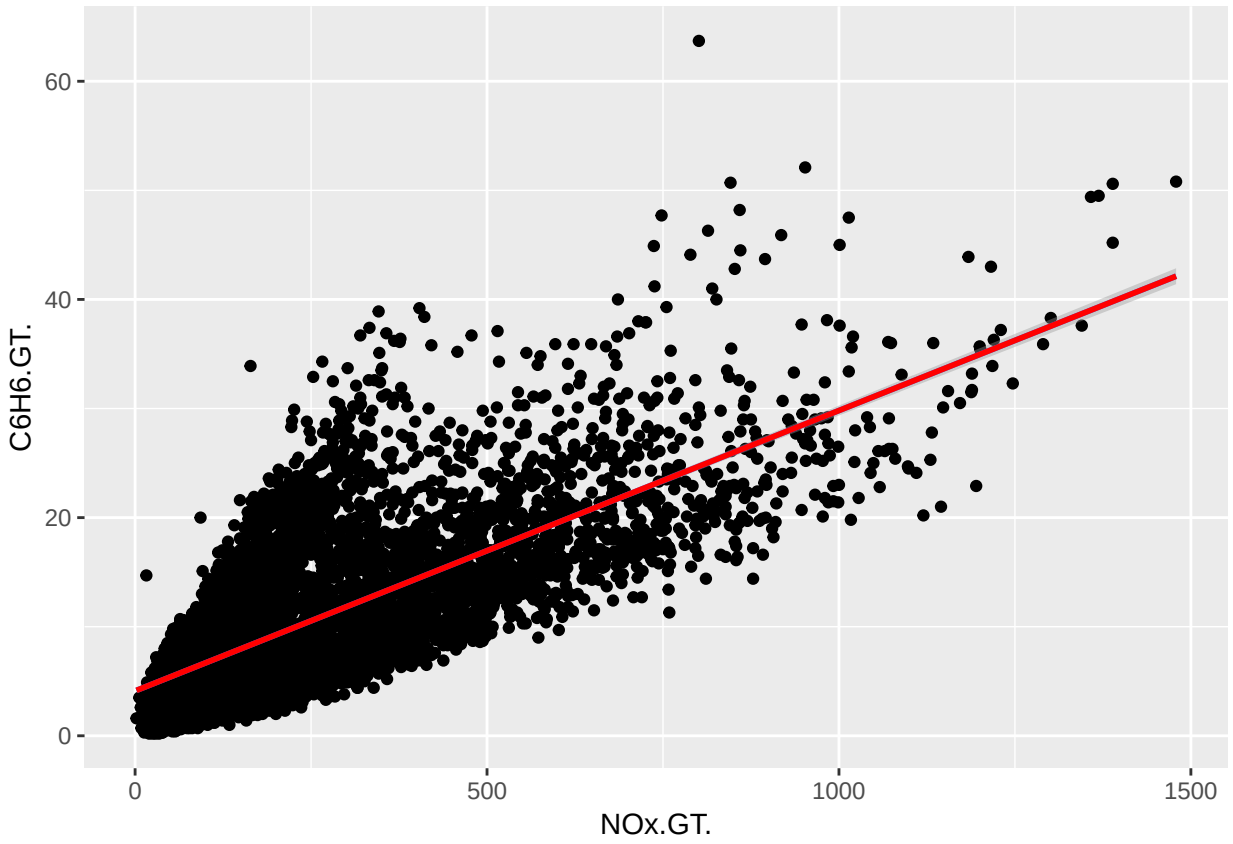
```
ggplot(lm3, aes(x = PT08.S2.NMHC., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = lm) +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```




```
lm4 <- lm(C6H6.GT. ~ NOx.GT., data = dt_new)
summary(lm4)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ NOx.GT., data = dt_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.702  -3.973  -1.113   2.790  38.999
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.1106722  0.0974625   42.18   <2e-16 ***
## NOx.GT.     0.0257062  0.0002989   86.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.194 on 6939 degrees of freedom
## Multiple R-squared:  0.516,  Adjusted R-squared:  0.5159
## F-statistic:  7398 on 1 and 6939 DF,  p-value: < 2.2e-16
```

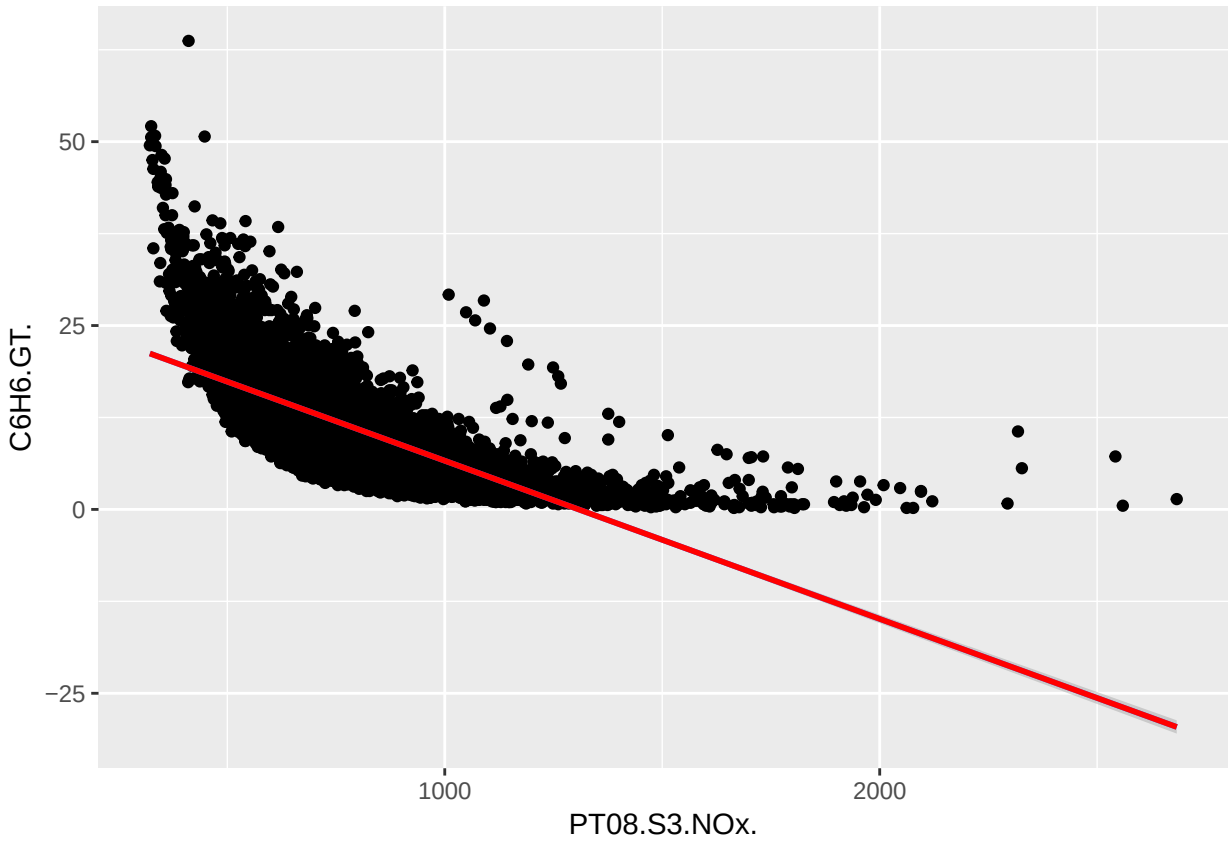
```
ggplot(lm4, aes(x = NOx.GT., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = lm) +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```




```
lm5 <- lm(C6H6.GT. ~ PT08.S3.NOx., data = dt_new)
summary(lm5)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S3.NOx., data = dt_new)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.071 -3.698 -0.810  2.306 44.416
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.1238172  0.2092372  134.41   <2e-16 ***
## PT08.S3.NOx. -0.0215075  0.0002448  -87.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.136 on 6939 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.5266
## F-statistic:  7721 on 1 and 6939 DF,  p-value: < 2.2e-16
```

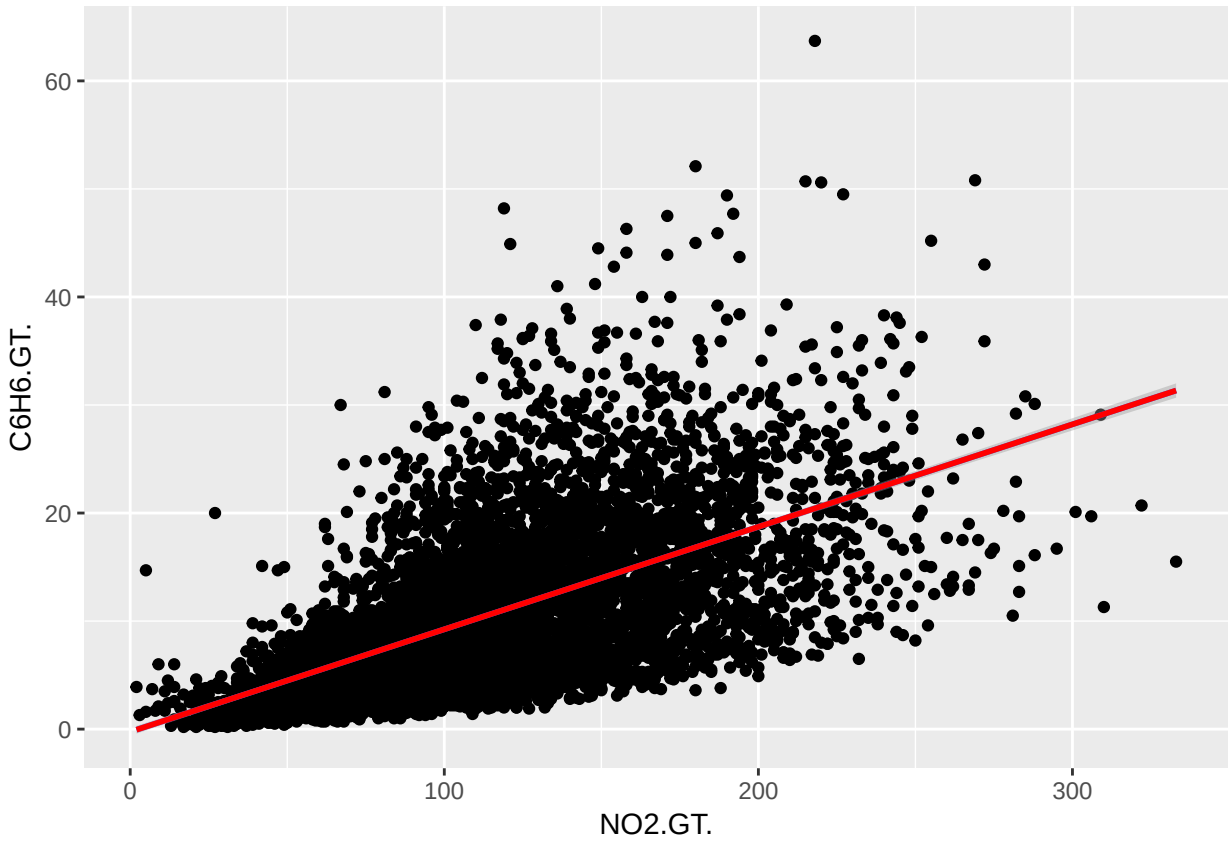
```
ggplot(lm5, aes(x = PT08.S3.NOx., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = lm) +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```




```
lm6 <- lm(C6H6.GT. ~ NO2.GT., data = dt_new)
summary(lm6)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ NO2.GT., data = dt_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.858  -3.572  -0.613   2.585  43.268
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.247248   0.185743  -1.331    0.183
## NO2.GT.      0.094857   0.001506  63.005   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.954 on 6939 degrees of freedom
## Multiple R-squared:  0.3639, Adjusted R-squared:  0.3638
## F-statistic:  3970 on 1 and 6939 DF,  p-value: < 2.2e-16
```

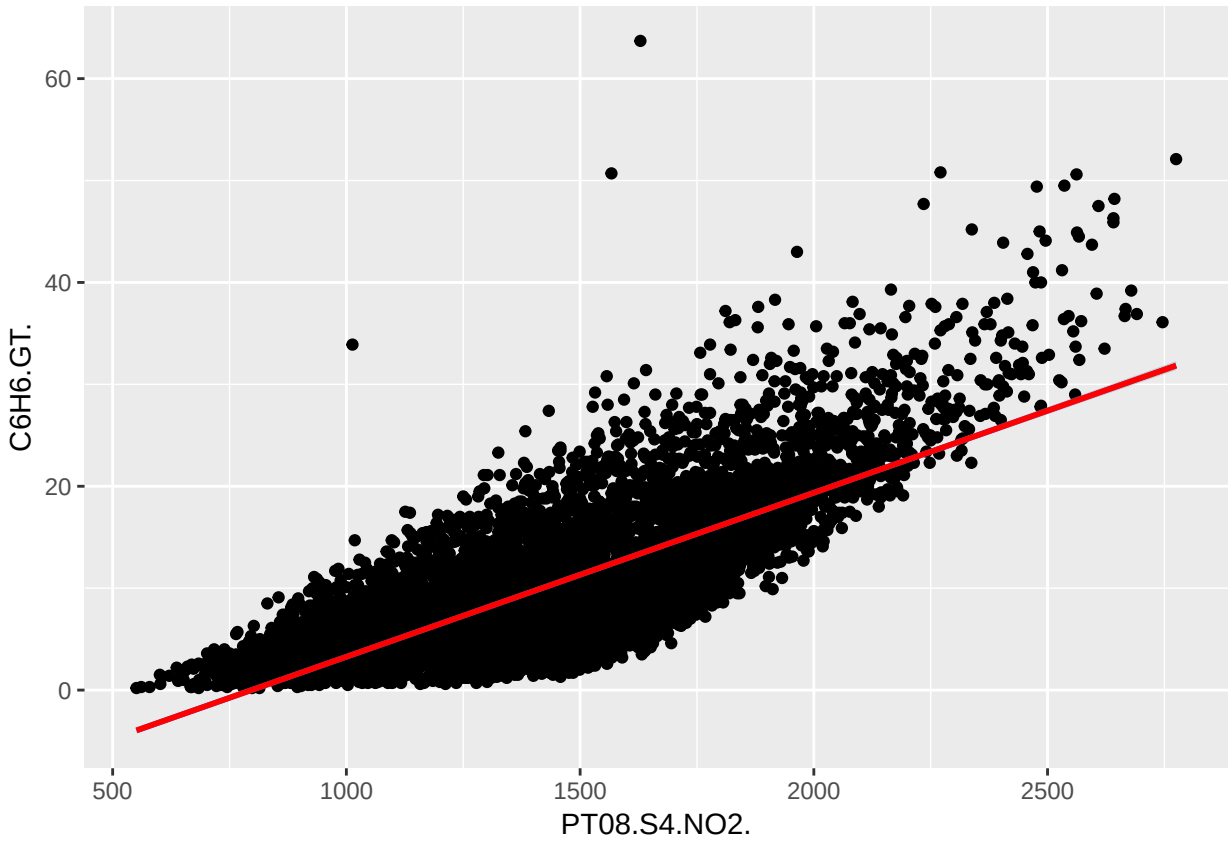
```
ggplot(lm6, aes(x = NO2.GT., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = lm) +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```




```
lm7 <- lm(C6H6.GT. ~ PT08.S4.NO2., data = dt_new)
summary(lm7)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S4.NO2., data = dt_new)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.941 -3.579 -0.460  2.919 50.307
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.283e+01  2.457e-01  -52.22   <2e-16 ***
## PT08.S4.NO2.  1.610e-02  1.643e-04   97.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.836 on 6939 degrees of freedom
## Multiple R-squared:  0.5803, Adjusted R-squared:  0.5803
## F-statistic:  9596 on 1 and 6939 DF,  p-value: < 2.2e-16
```

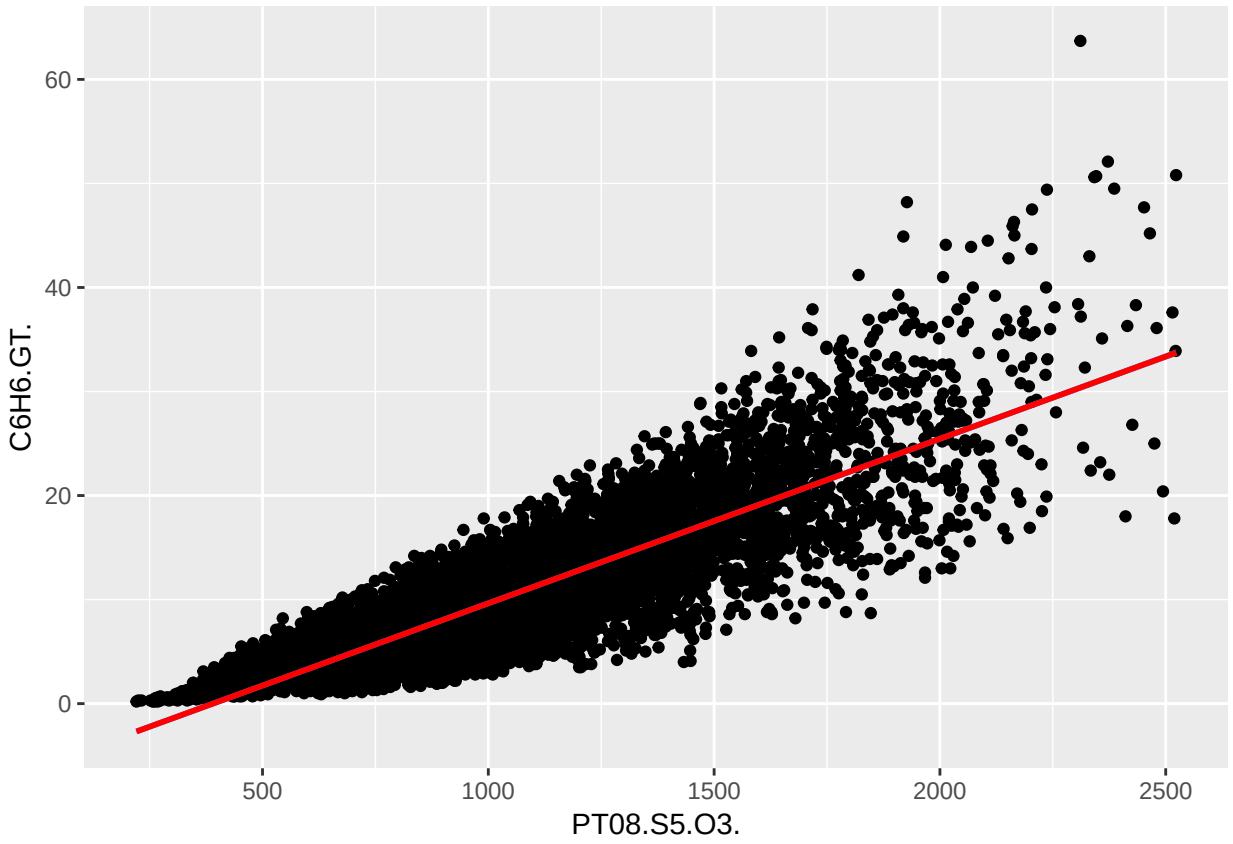
```
ggplot(lm7, aes(x = PT08.S4.NO2., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = lm) +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```




```
lm8 <- lm(C6H6.GT. ~ PT08.S5.O3., data = dt_new)
summary(lm8)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S5.O3., data = dt_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.863  -2.333  -0.038   2.112  33.326
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.1732848  0.1269922  -48.61   <2e-16 ***
## PT08.S5.O3.  0.0158144  0.0001121  141.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.795 on 6939 degrees of freedom
## Multiple R-squared:  0.7416, Adjusted R-squared:  0.7415
## F-statistic: 1.991e+04 on 1 and 6939 DF,  p-value: < 2.2e-16
```

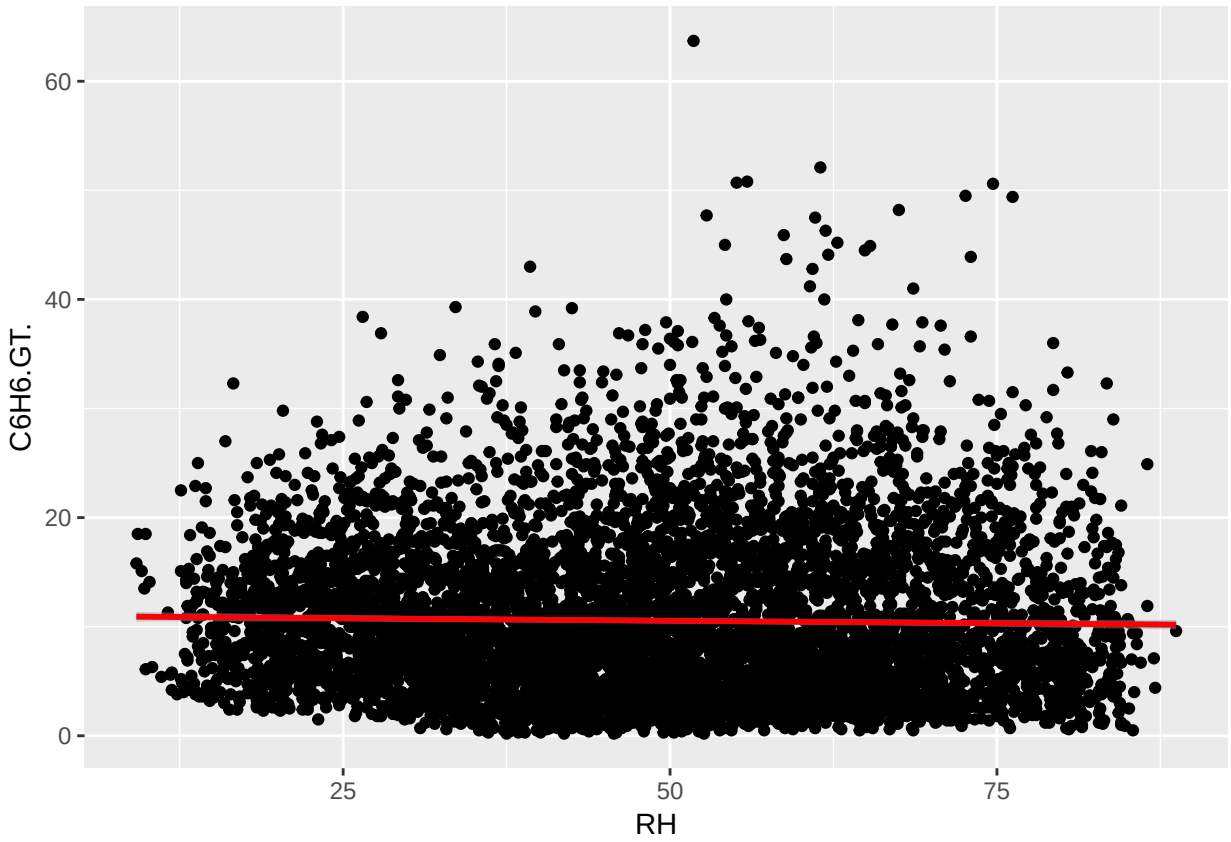
```
ggplot(lm8, aes(x = PT08.S5.O3., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = lm) +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```




```
lm9 <- lm(C6H6.GT. ~ RH, data = dt_new)
summary(lm9)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ RH, data = dt_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.460  -5.624  -1.763   3.973  53.172
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.006445   0.266723  41.266   <2e-16 ***
## RH          -0.009246   0.005139  -1.799   0.0721 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.464 on 6939 degrees of freedom
## Multiple R-squared:  0.0004662,  Adjusted R-squared:  0.0003221
## F-statistic: 3.236 on 1 and 6939 DF,  p-value: 0.07206
```

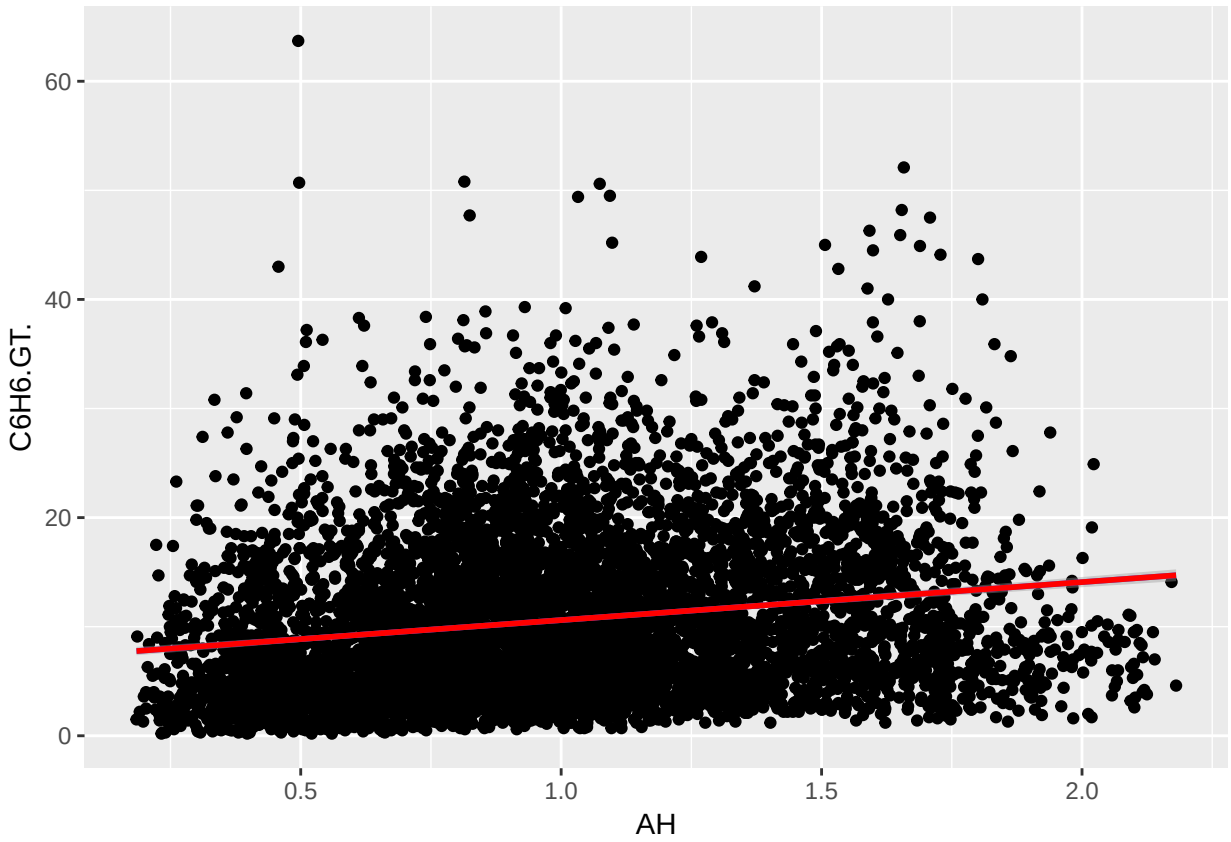
```
ggplot(lm9, aes(x = RH, y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = lm) +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```



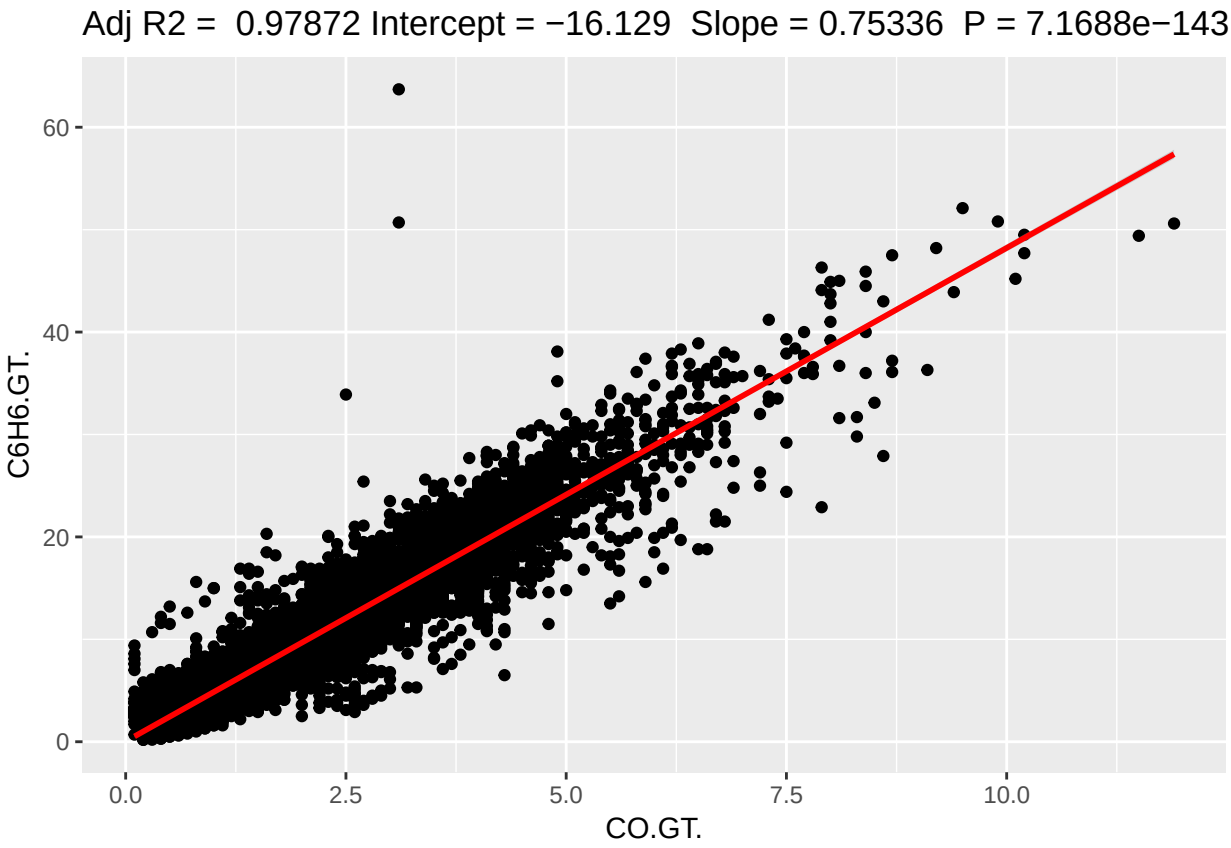

```
lm10 <- lm(C6H6.GT. ~ AH, data = dt_new)
summary(lm10)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ AH, data = dt_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.451  -5.553  -1.702   3.861  54.853
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.1229     0.2335   30.50   <2e-16 ***
## AH            3.4818     0.2195   15.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.334 on 6939 degrees of freedom
## Multiple R-squared:  0.035,  Adjusted R-squared:  0.03486
## F-statistic: 251.6 on 1 and 6939 DF,  p-value: < 2.2e-16
```

```
ggplot(lm10, aes(x = AH, y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = lm) +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```




```
ggplot(lmMod$model, aes_string(x = names(lmMod$model)[2], y = names(lmMod$model)[1])) +
  geom_point() +
  geom_smooth(method = "lm", col = "red") +
  labs(title = paste("Adj R2 = ",signif(summary(lmMod)$adj.r.squared, 5),
                     "Intercept =",signif(lmMod$coef[[1]],5 ),
                     " Slope =",signif(lmMod$coef[[2]], 5),
                     " P =",signif(summary(lmMod)$coef[2,4], 5)))
```



Now we choose ~CO.GT. variable as good predictor for our model

Lets create train-test for it and plot

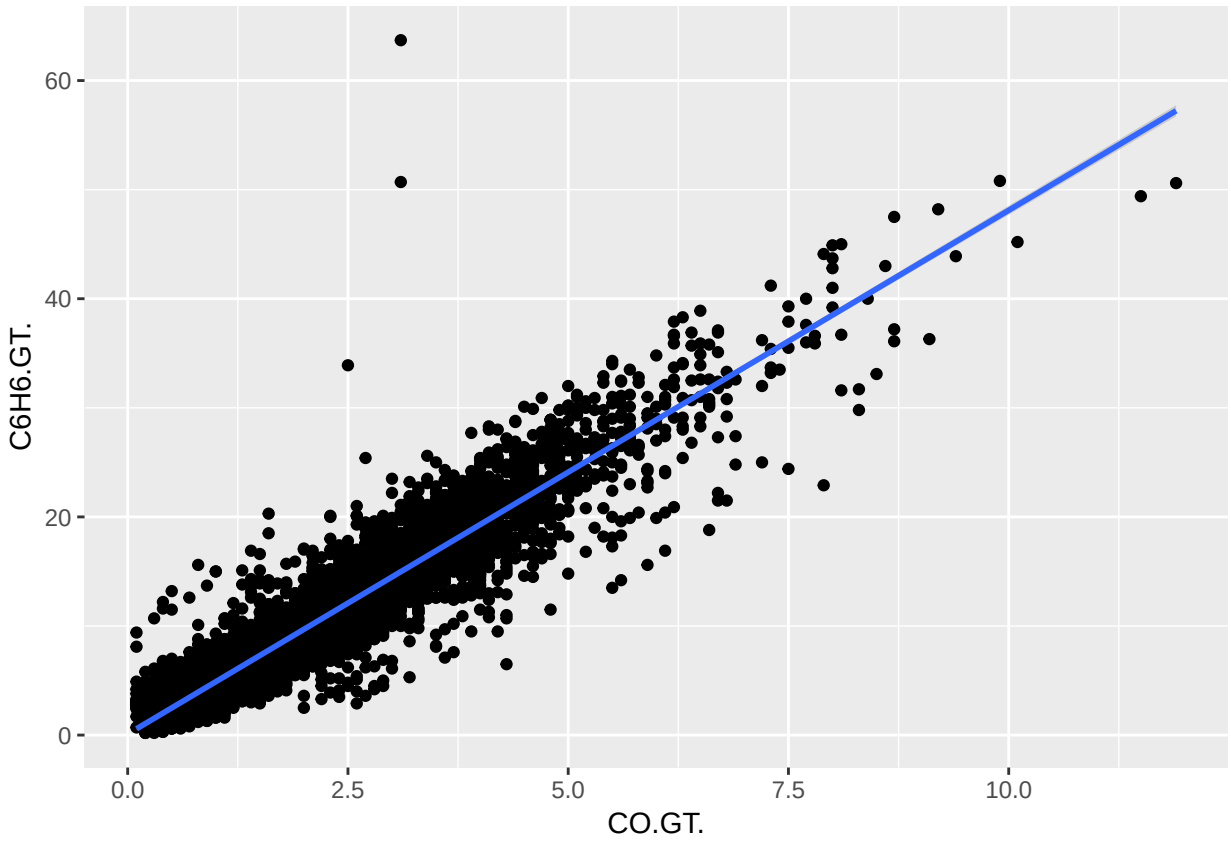
```
# Prepare data for prediction and model training (75%):
set.seed(42)
sample <- sample.int(n = nrow(dt_new), size = floor(.75*nrow(dt_new)))
train <- dt_new[sample,]
test  <- dt_new[-sample,]
```

```
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
## -15.125  -1.592  -0.172   1.485  48.723
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.09148    0.06975   1.311     0.19
```

```
## CO.GT.       4.80167    0.02665 180.175   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.768 on 5203 degrees of freedom
## Multiple R-squared:  0.8619, Adjusted R-squared:  0.8618
## F-statistic: 3.246e+04 on 1 and 5203 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)
```

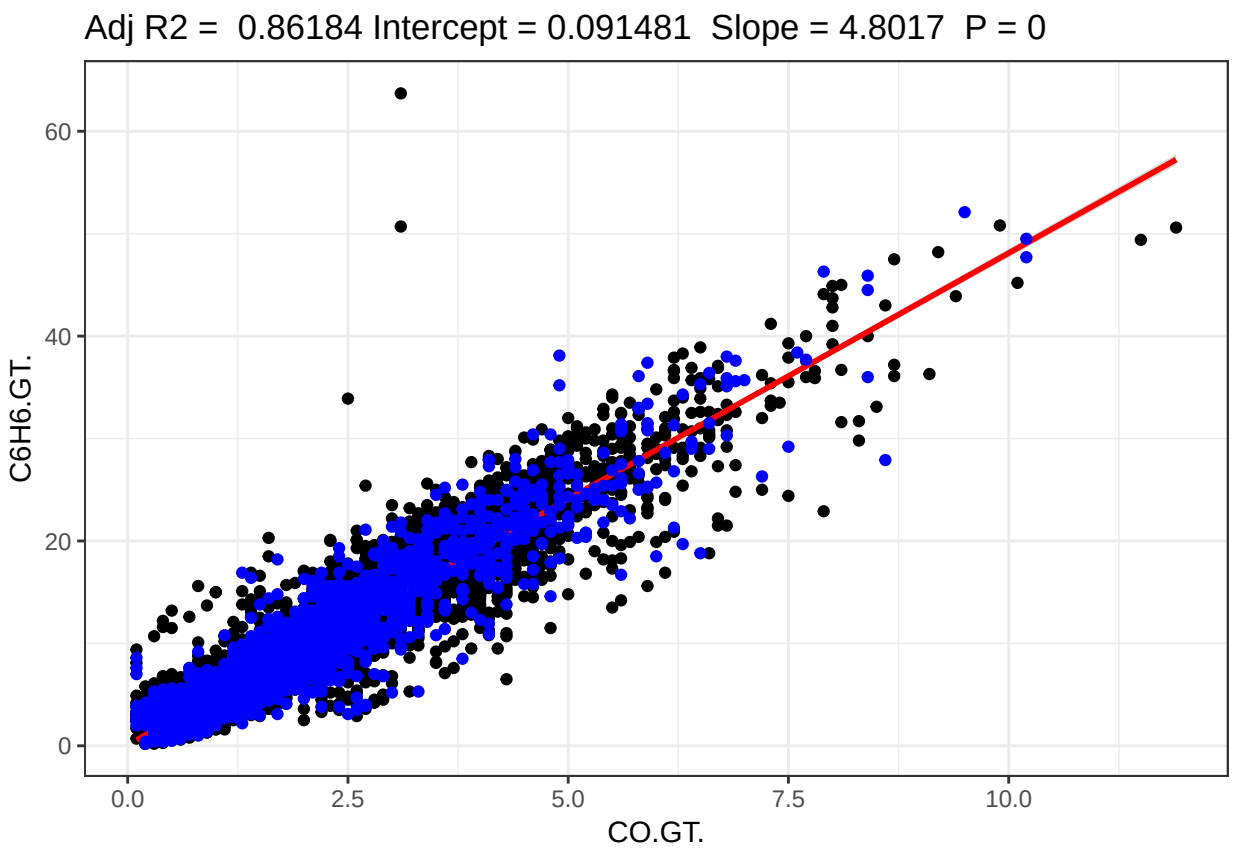
```
##         1         4         6        14        16        17
## 12.575832 10.655163  5.853489  5.373322 10.655163  8.254326
```

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

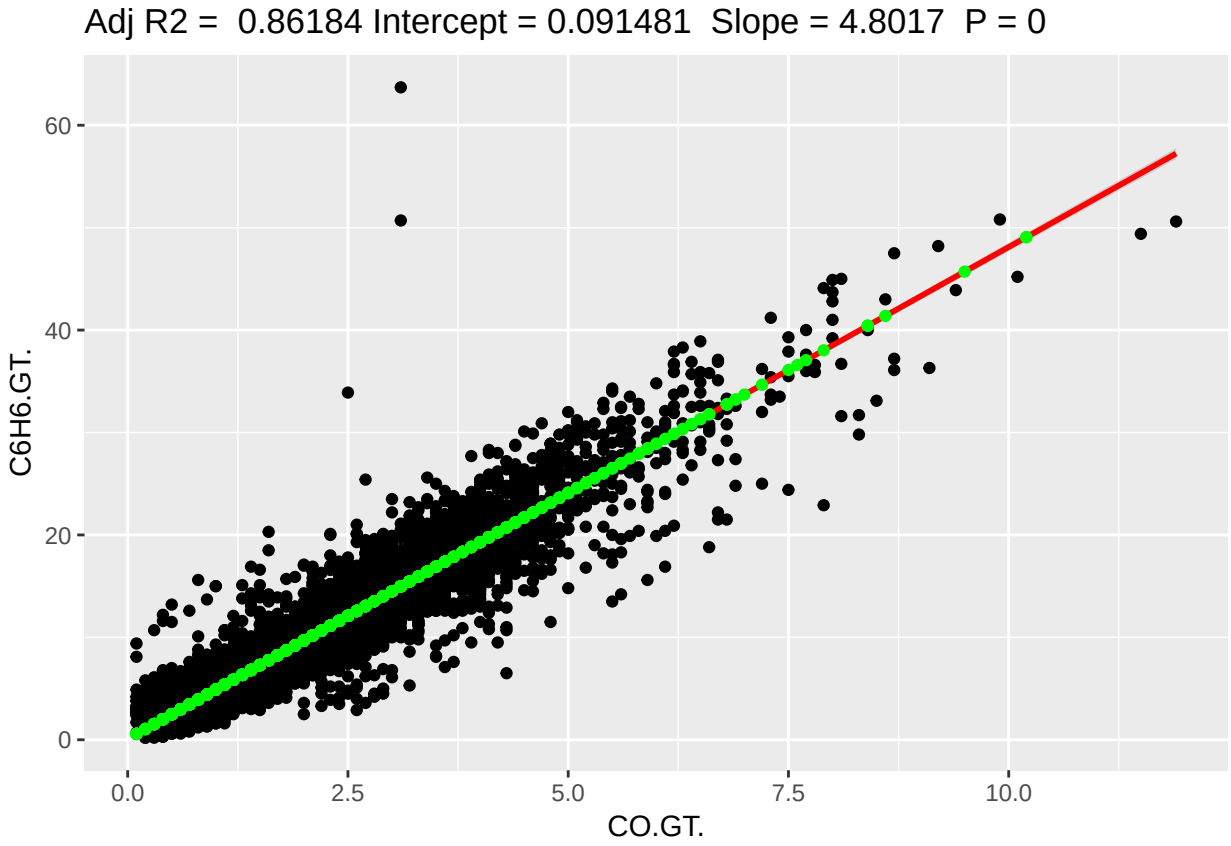
```
##    CO.GT. PT08.S1.CO. C6H6.GT. PT08.S2.NMHC. NOx.GT. PT08.S3.NOx. NO2.GT.
## 1     2.6        1360     11.9          1046     166         1056     113
## 4     2.2        1376      9.2           948     172         1092     122
## 6     1.2        1197      4.7           750      89         1337      96
## 14    1.1        1144      3.2           667      98         1490      82
## 16    2.2        1351      9.5           960     129         1079     101
## 17    1.7        1233      6.3           827     112         1218      98
##    PT08.S4.NO2. PT08.S5.O3.    T   RH     AH C6H6.GT.pred
```

```
## 1          1692        1268 13.6 48.9 0.7578    12.575832
## 4          1584        1203 11.0 60.0 0.7867    10.655163
## 6          1393         949 11.2 59.2 0.7848     5.853489
## 14         1339         730 10.2 59.6 0.7417     5.373322
## 16         1583        1028 10.5 60.6 0.7691    10.655163
## 17         1446         860 10.8 58.4 0.7552     8.254326
```

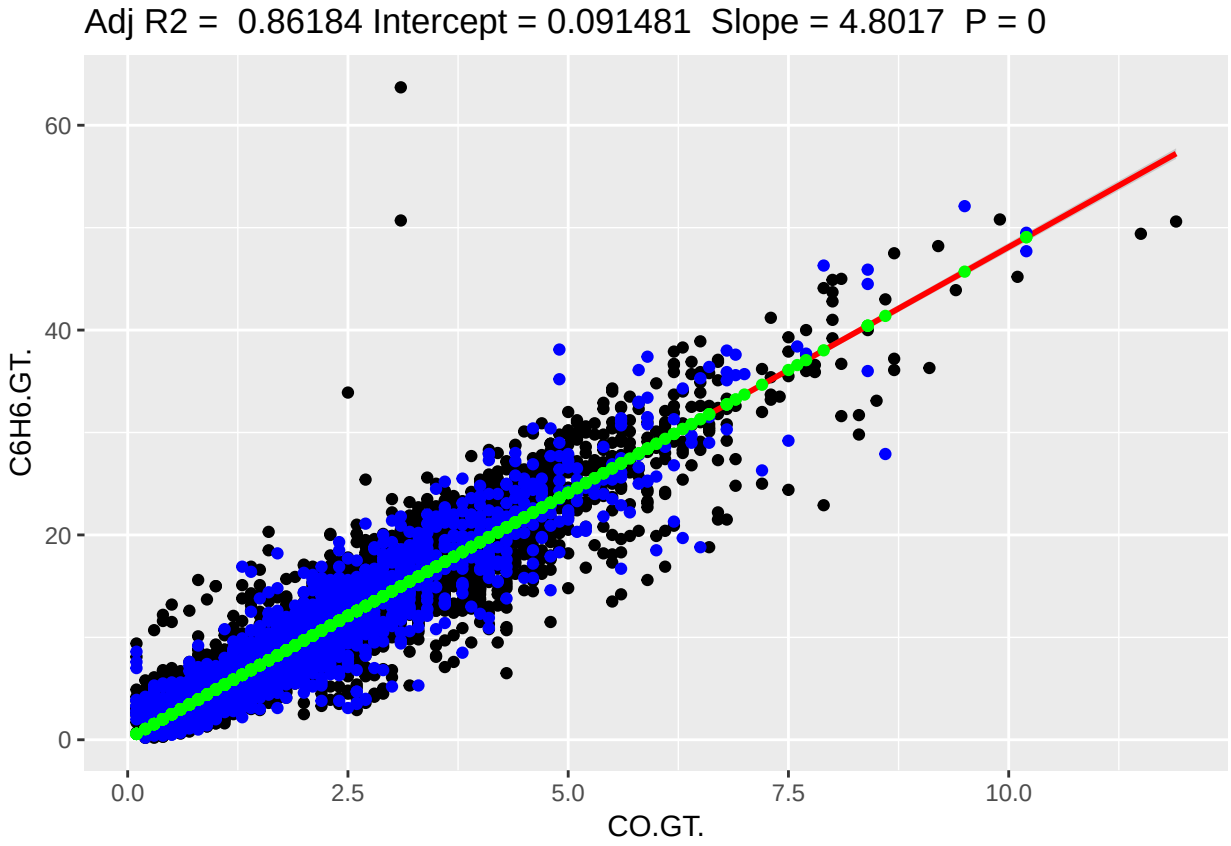
```r
#train plot
ggplot(train, aes(x = CO.GT., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red" ) +
  geom_point(data = test, aes(y = C6H6.GT.), color = "blue") +
  theme_bw() +
  labs(title = paste("Adj R2 = ",signif(summary(new_mod)$adj.r.squared, 5),
                     "Intercept =",signif(new_mod$coef[[1]],5 ),
                     " Slope =",signif(new_mod$coef[[2]], 5),
                     " P =",signif(summary(new_mod)$coef[2,4], 5)))
```



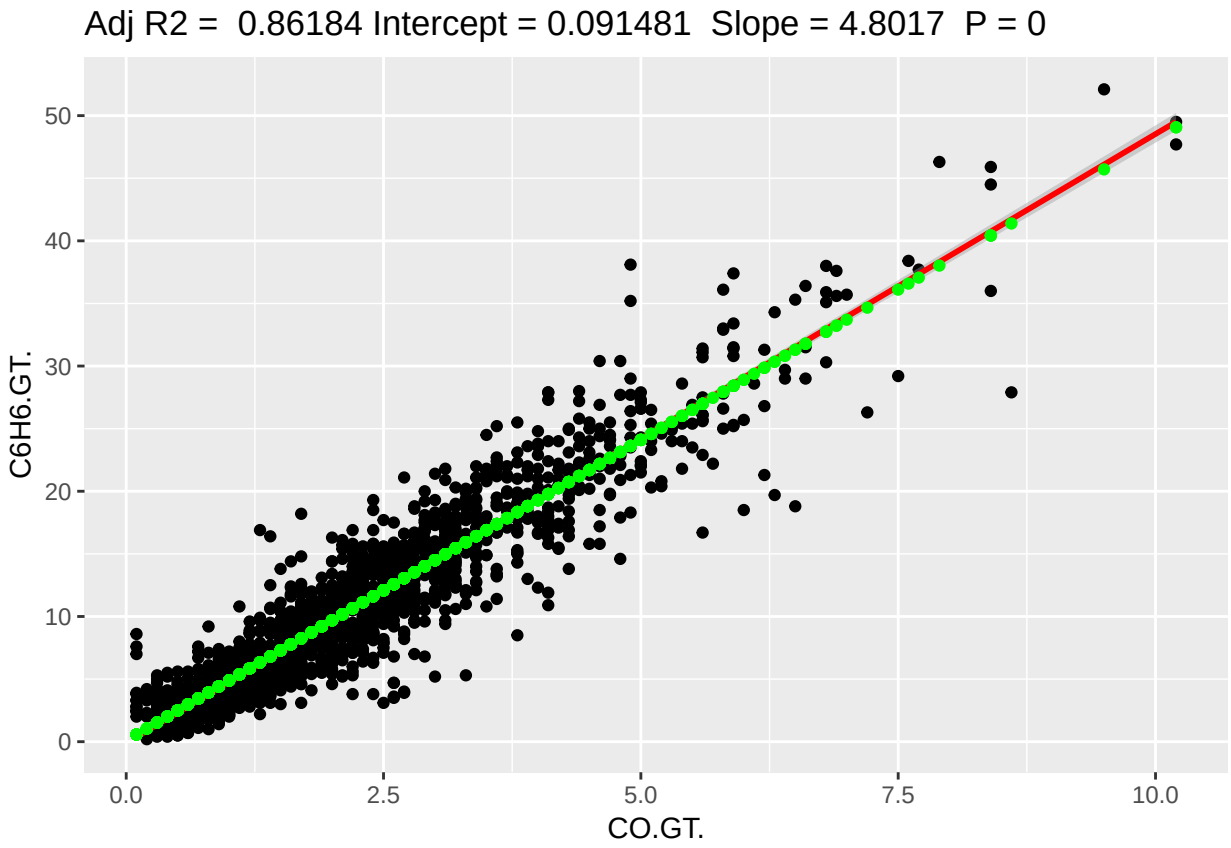

```r
# predicted plot
ggplot(train, aes(x = CO.GT., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red" ) +
  geom_point(data = test, aes(y = C6H6.GT.pred), color = "green") +
  labs(title = paste("Adj R2 = ",signif(summary(new_mod)$adj.r.squared, 5),
                     "Intercept =",signif(new_mod$coef[[1]],5 ),
                     " Slope =",signif(new_mod$coef[[2]], 5),
                     " P =",signif(summary(new_mod)$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") +
  labs(title = paste("Adj R2 = ",signif(summary(new_mod)$adj.r.squared, 5),
                     "Intercept =",signif(new_mod$coef[[1]],5 ),
                     " Slope =",signif(new_mod$coef[[2]], 5),
                     " P =",signif(summary(new_mod)$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") +
  labs(title = paste("Adj R2 = ",signif(summary(new_mod)$adj.r.squared, 5),
                     "Intercept =",signif(new_mod$coef[[1]],5 ),
                     " Slope =",signif(new_mod$coef[[2]], 5),
                     " P =",signif(summary(new_mod)$coef[2,4], 5)))
```


Adj R2 = 0.86184 Intercept = 0.091481 Slope = 4.8017 P = 0
C6H6.GT.
CO.GT.
0
10
20
30
40
50
0.0
2.5
5.0
7.5
10.0