Task10

Darya Nemirich

29 May 2019

Part 1

```
anscombe <- readRDS("D:/Bioinformatics and System Biology/2nd term/R/R_classwork/Task10_Case/anscombe.R
```

```
library(ggplot2)
library(dplyr)
```

```
##
## 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
```

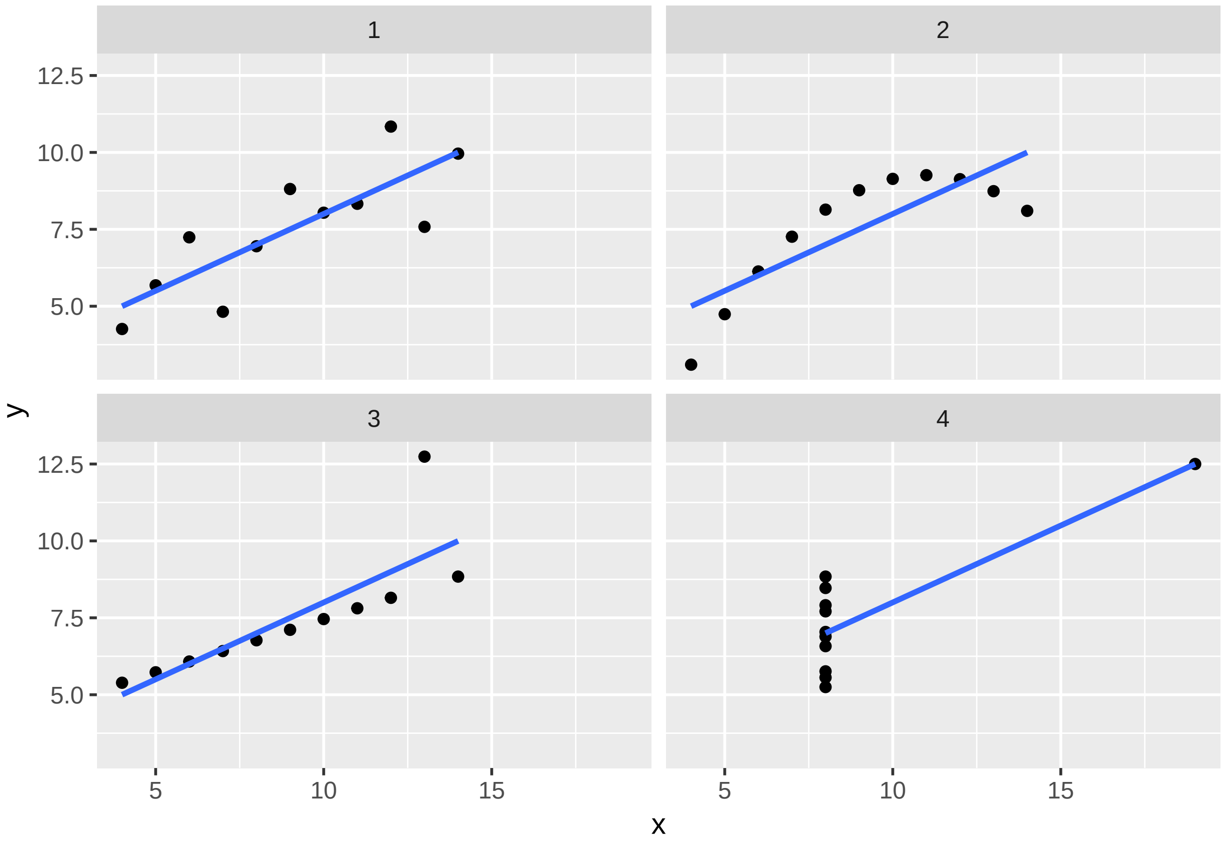
```
str(anscombe)
```

```
## 'data.frame':    44 obs. of  3 variables:
##  $ x  : num  10 8 13 9 11 14 6 4 12 7 ...
##  $ y  : num  8.04 6.95 7.58 8.81 8.33 ...
##  $ set: num  1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(anscombe)
```

```
##        x             y               set
##  Min.   : 4   Min.   : 3.100   Min.   :1.00
##  1st Qu.: 7   1st Qu.: 6.117   1st Qu.:1.75
##  Median : 8   Median : 7.520   Median :2.50
##  Mean   : 9   Mean   : 7.501   Mean   :2.50
##  3rd Qu.:11   3rd Qu.: 8.748   3rd Qu.:3.25
##  Max.   :19   Max.   :12.740   Max.   :4.00
```

```
ggplot(data = anscombe, aes(x = x,
                            y = y)) +
  geom_point() +
  facet_wrap(set ~ .) +
  geom_smooth(method = "lm", se = F)
```



```r
anscombe %>%
  group_by(set) %>%
  summarise(
    x_mean = mean(x),
    y_mean = mean(y),
    x_sd = sd(x),
    y_sd = sd(y)
  )
```

```
## # A tibble: 4 x 5
##     set x_mean y_mean  x_sd  y_sd
##   <dbl>  <dbl>  <dbl> <dbl> <dbl>
## 1     1      9   7.50  3.32  2.03
## 2     2      9   7.50  3.32  2.03
## 3     3      9   7.5   3.32  2.03
## 4     4      9   7.50  3.32  2.03
```

```r
anscombe %>%
  group_by(set) %>%
  summarise(
    correlation = cor(x, y),
    p_value = cor.test(x, y)$p.value)
```

```
## # A tibble: 4 x 3
##     set correlation p_value
##   <dbl>       <dbl>   <dbl>
## 1     1       0.816 0.00217
## 2     2       0.816 0.00218
## 3     3       0.816 0.00218
## 4     4       0.817 0.00216
```

```
anscombe %>%
  group_by(set) %>%
  summarise(
    cor_pearson = cor(x, y, method = "pearson"),
    cor_kendall = cor(x, y, method = "kendall"),
    cor_spearman = cor(x, y, method = "spearman")
  )
```

```
## # A tibble: 4 x 4
##     set cor_pearson cor_kendall cor_spearman
##   <dbl>       <dbl>       <dbl>        <dbl>
## 1     1       0.816       0.636        0.818
## 2     2       0.816       0.564        0.691
## 3     3       0.816       0.964        0.991
## 4     4       0.817       0.426        0.5
```

Part 2

```
airq <- read.csv2("D:/Bioinformatics and System Biology/2nd term/R/R_classwork/Task10_Case/AirQualityUC
```

```
head(airq)
```

```
##         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(airq)
```

```
## '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(airq)
```

```
##         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
##
##
##
##
##
```

```
airq_new <- airq %>%
  select(-c(X, X.1)) %>%
  na.omit()


head(airq_new)
```

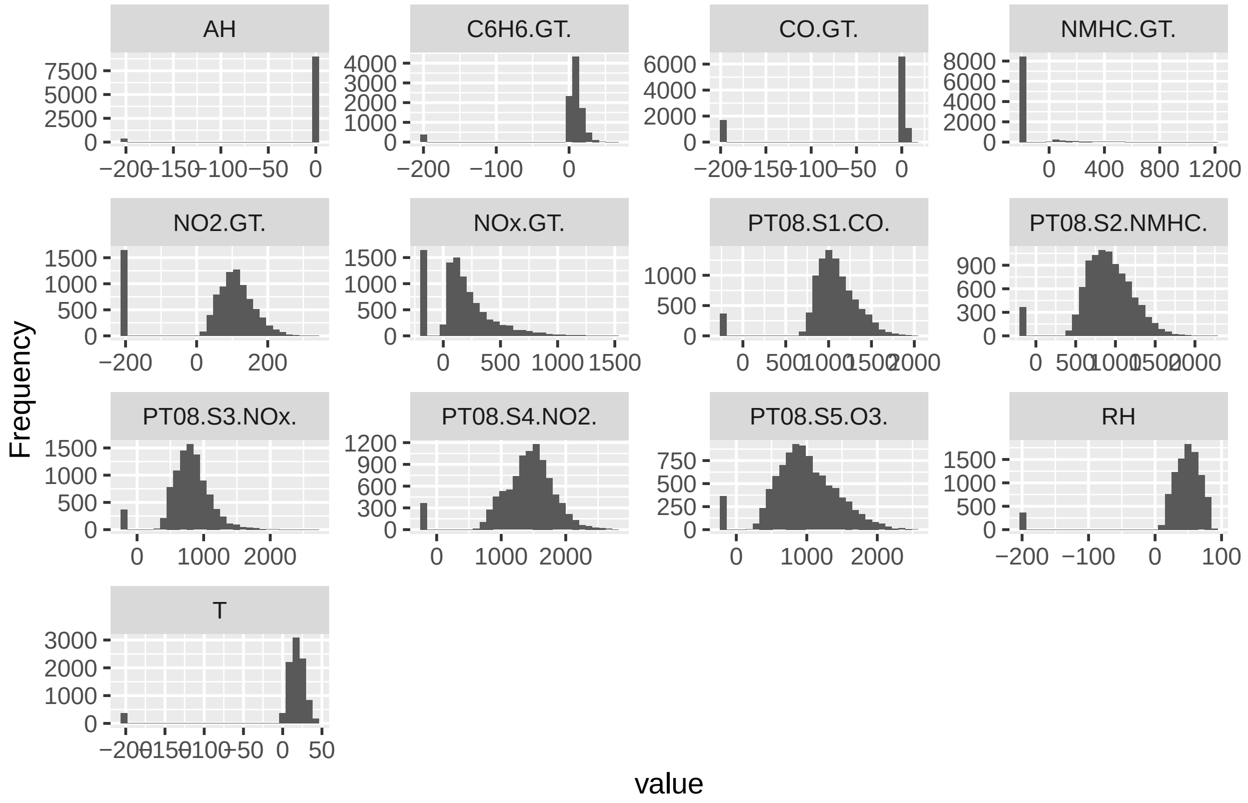
```
##         Date     Time CO.GT. PT08.S1.CO. NMHC.GT. C6H6.GT. PT08.S2.NMHC.
## 1 10/03/2004 18.00.00    2.6        1360      150     11.9          1046
## 2 10/03/2004 19.00.00    2.0        1292      112      9.4           955
## 3 10/03/2004 20.00.00    2.2        1402       88      9.0           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(airq_new)
```

```
##          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
##
```

```
library(DataExplorer)
```

```
plot_histogram(airq_new)
```

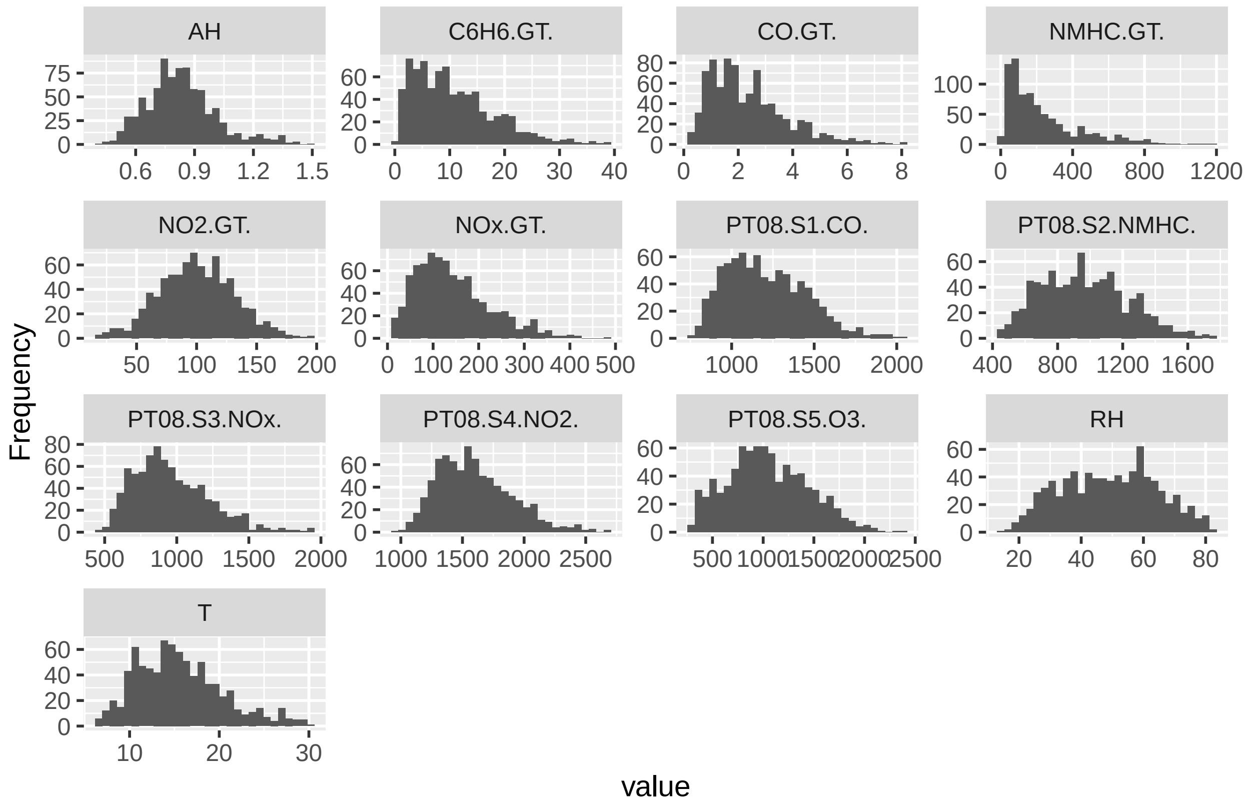


```
airq_new <- airq_new %>%
  na_if(-200) %>%
  na_if(-200.0) %>%
  na.omit()

summary(airq_new)
```

```
##          Date            Time          CO.GT.       PT08.S1.CO.
##  06/04/2004: 23   00.00.00: 38   Min.   :0.300   Min.   : 753
##  07/04/2004: 23   18.00.00: 38   1st Qu.:1.300   1st Qu.:1017
##  10/04/2004: 23   19.00.00: 38   Median :2.000   Median :1172
##  12/04/2004: 23   20.00.00: 38   Mean   :2.354   Mean   :1208
##  13/04/2004: 23   21.00.00: 38   3rd Qu.:3.100   3rd Qu.:1380
##  16/04/2004: 23   22.00.00: 38   Max.   :8.100   Max.   :2040
##  (Other)    :689   (Other) :599
##     NMHC.GT.         C6H6.GT.      PT08.S2.NMHC.      NOx.GT.
##  Min.   :   7.0   Min.   : 0.50   Min.   : 448.0   Min.   : 12.0
##  1st Qu.:  77.0   1st Qu.: 4.80   1st Qu.: 754.0   1st Qu.: 81.0
##  Median : 157.0   Median : 9.10   Median : 944.0   Median :128.0
##  Mean   : 231.0   Mean   :10.77   Mean   : 966.1   Mean   :143.5
##  3rd Qu.: 318.5   3rd Qu.:14.80   3rd Qu.:1142.5   3rd Qu.:187.0
##  Max.   :1189.0   Max.   :39.20   Max.   :1754.0   Max.   :478.0
##
##   PT08.S3.NOx.       NO2.GT.       PT08.S4.NO2.   PT08.S5.O3.
##  Min.   : 461.0   Min.   : 19.0   Min.   : 955   Min.   : 263
##  1st Qu.: 769.0   1st Qu.: 78.5   1st Qu.:1370   1st Qu.: 760
##  Median : 920.0   Median : 99.0   Median :1556   Median :1009
##  Mean   : 963.3   Mean   :100.3   Mean   :1601   Mean   :1046
##  3rd Qu.:1131.0   3rd Qu.:122.0   3rd Qu.:1784   3rd Qu.:1320
##  Max.   :1935.0   Max.   :196.0   Max.   :2679   Max.   :2359
##
##        T              RH              AH
##  Min.   : 6.3   Min.   :14.90   Min.   :0.4023
##  1st Qu.:11.9   1st Qu.:36.70   1st Qu.:0.7189
##  Median :15.0   Median :49.60   Median :0.8177
##  Mean   :15.6   Mean   :49.05   Mean   :0.8319
##  3rd Qu.:18.3   3rd Qu.:60.55   3rd Qu.:0.9275
##  Max.   :30.0   Max.   :83.20   Max.   :1.4852
##
```

```
plot_histogram(airq_new)
```



```
airq_new[, c(3:15)] <- lapply(airq_new[,c(3:15)], as.numeric)
str(airq_new)
```

```
## 'data.frame':    827 obs. of  15 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 ...
##  $ NMHC.GT.     : num  150 112 88 80 51 38 31 31 24 8 ...
##  $ 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" ...
```

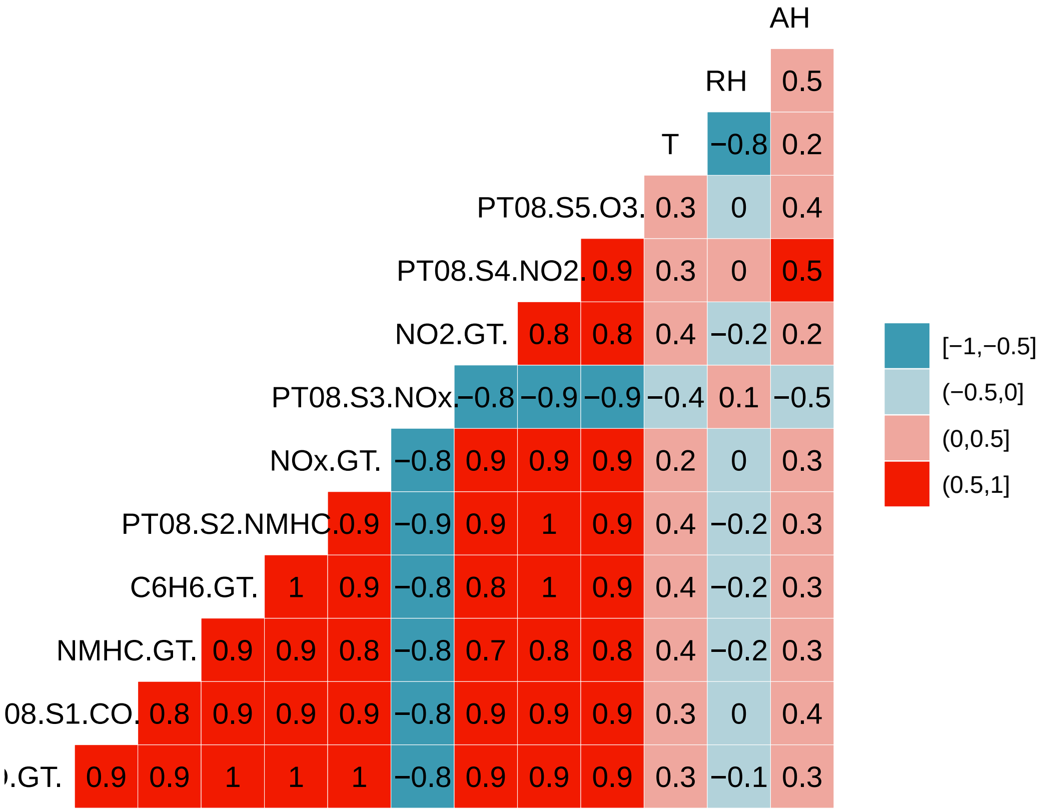
```r
library(GGally)
```

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

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

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

```
## Warning in ggcorr(airq_new, nbreaks = 4, label = TRUE, hjust = 0.8): data
## in column(s) 'Date', 'Time' are not numeric and were ignored
```



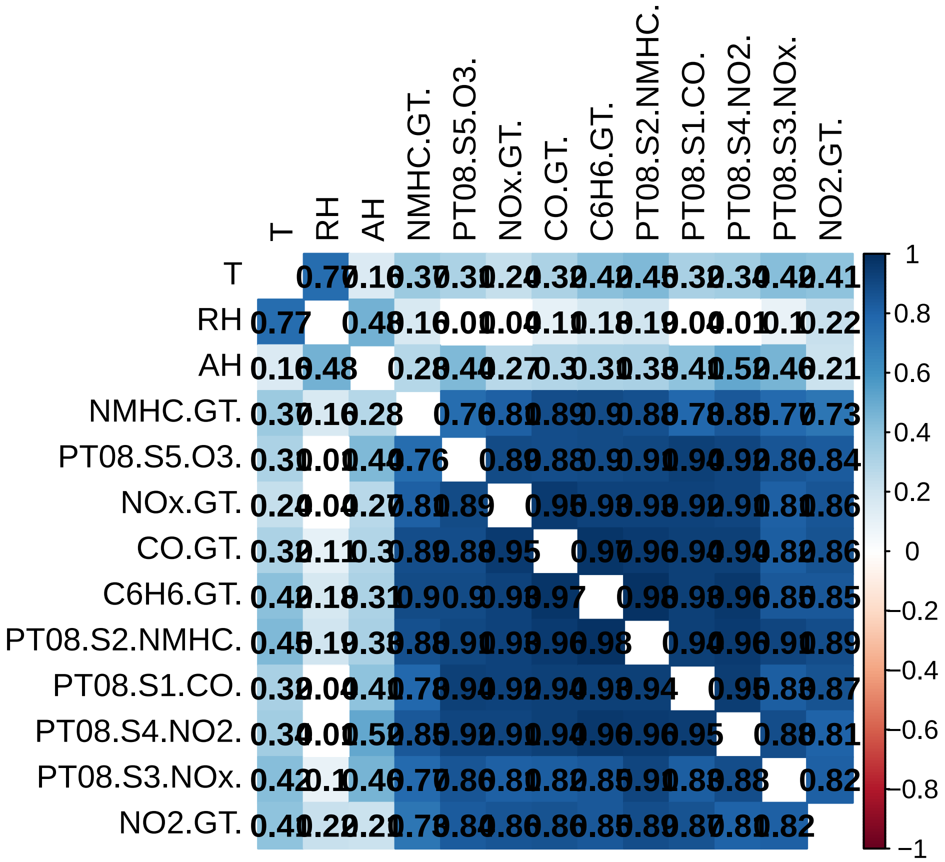
```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
visual_cor <- function(d){
  cormat <- cor(d, use = "pairwise.complete.obs")
  pvalmat <- cor.mtest(d)$p

  corrplot(abs(cormat),
           method = "color",
           order = "hclust",
           addCoef.col = "black",
           tl.col = "black", tl.srt = 90,
           p.mat = pvalmat, sig.level = 0.05,
           insig = "blank", diag = FALSE)
}
```

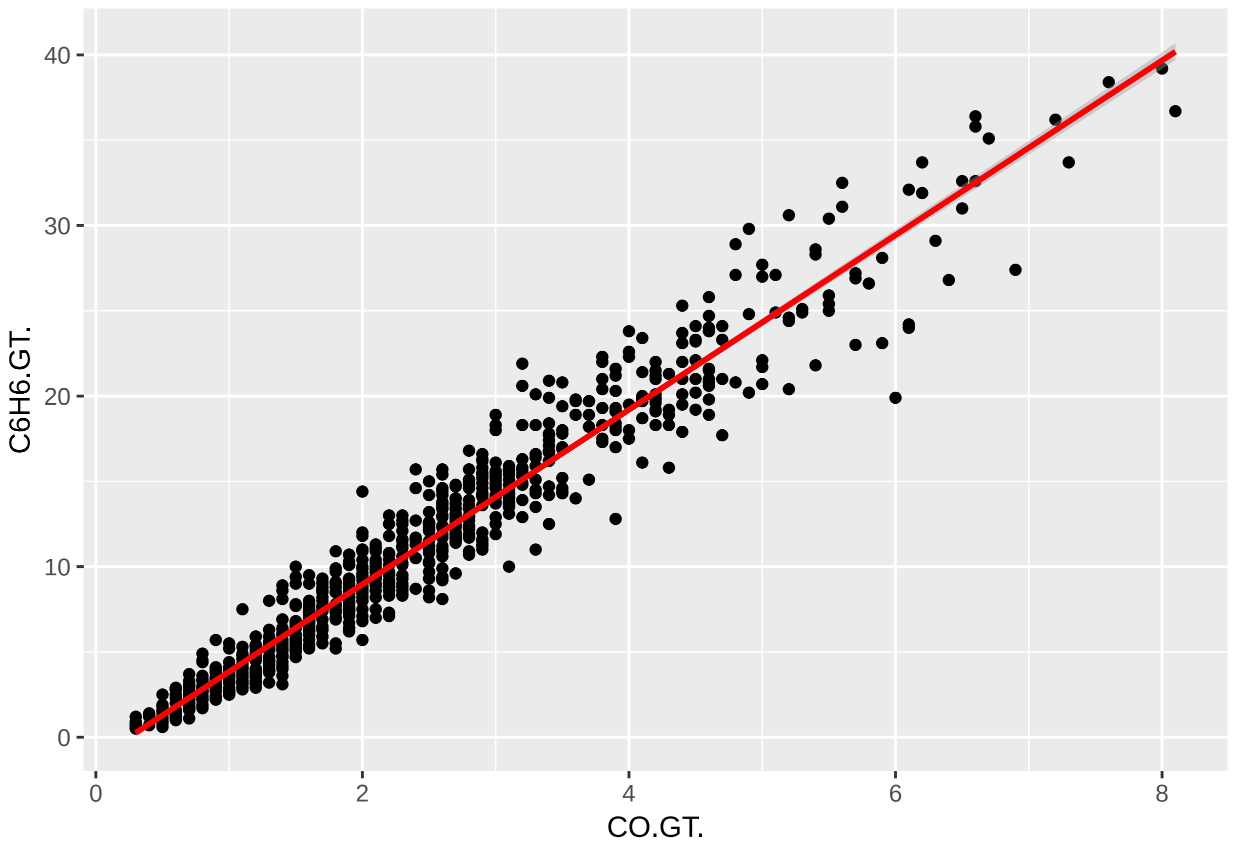
```
visual_cor(airq_new[, -c(1,2)])
```



```r
lm_1 <- lm(C6H6.GT. ~ CO.GT., data = airq_new)
summary(lm_1)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ CO.GT., data = airq_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5375 -0.9541 -0.1064  0.8293  6.7959
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.27699    0.11672  -10.94   <2e-16 ***
## CO.GT.       5.11908    0.04255  120.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.724 on 825 degrees of freedom
## Multiple R-squared:  0.9461, Adjusted R-squared:  0.946
## F-statistic: 1.447e+04 on 1 and 825 DF,  p-value: < 2.2e-16
```

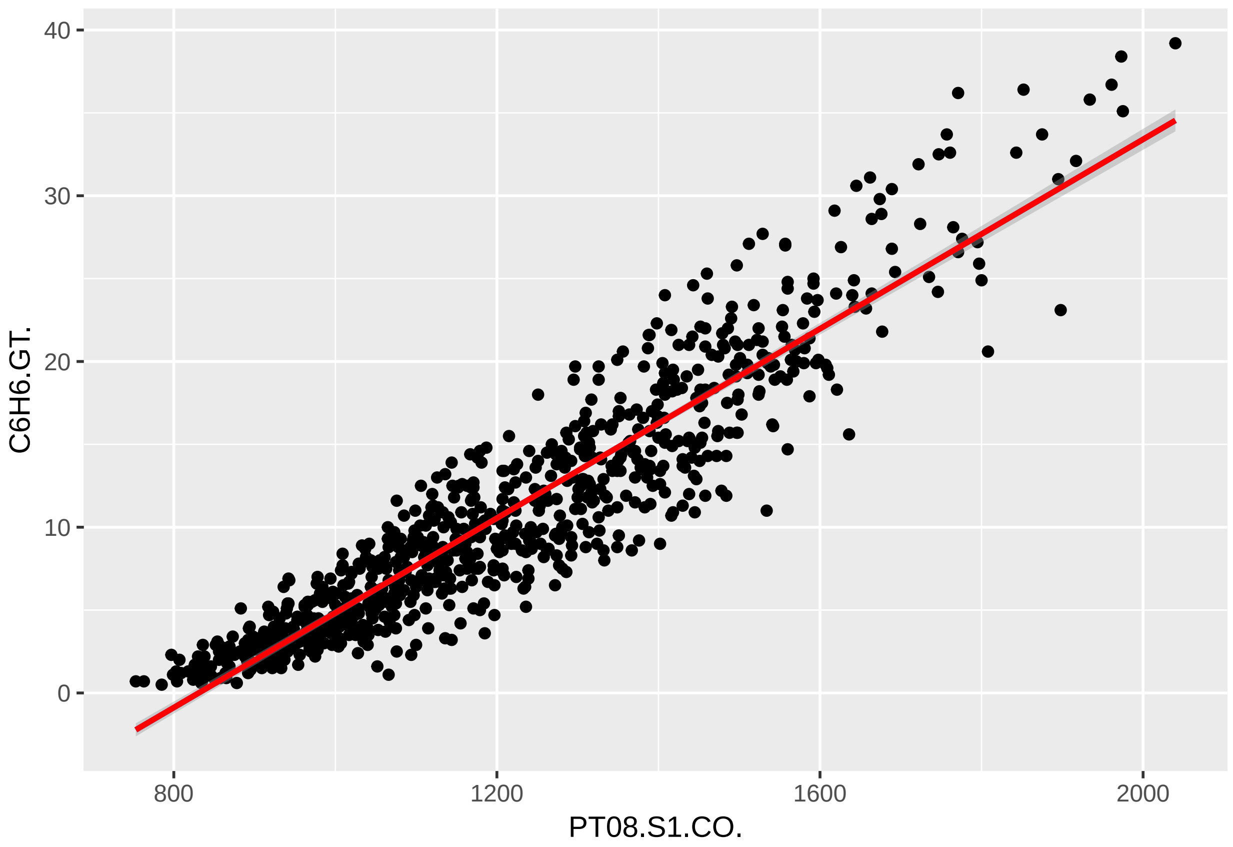
```r
ggplot(lm_1, aes(x = CO.GT., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```



```
lm_2 <- lm(C6H6.GT. ~ PT08.S1.CO., data = airq_new)
summary(lm_2)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S1.CO., data = airq_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0888 -1.6245  0.0254  1.6468  9.3398
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.374e+01  4.790e-01  -49.56   <2e-16 ***
## PT08.S1.CO.  2.857e-02  3.888e-04   73.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.702 on 825 degrees of freedom
## Multiple R-squared:  0.8674, Adjusted R-squared:  0.8673
## F-statistic:  5399 on 1 and 825 DF,  p-value: < 2.2e-16
```

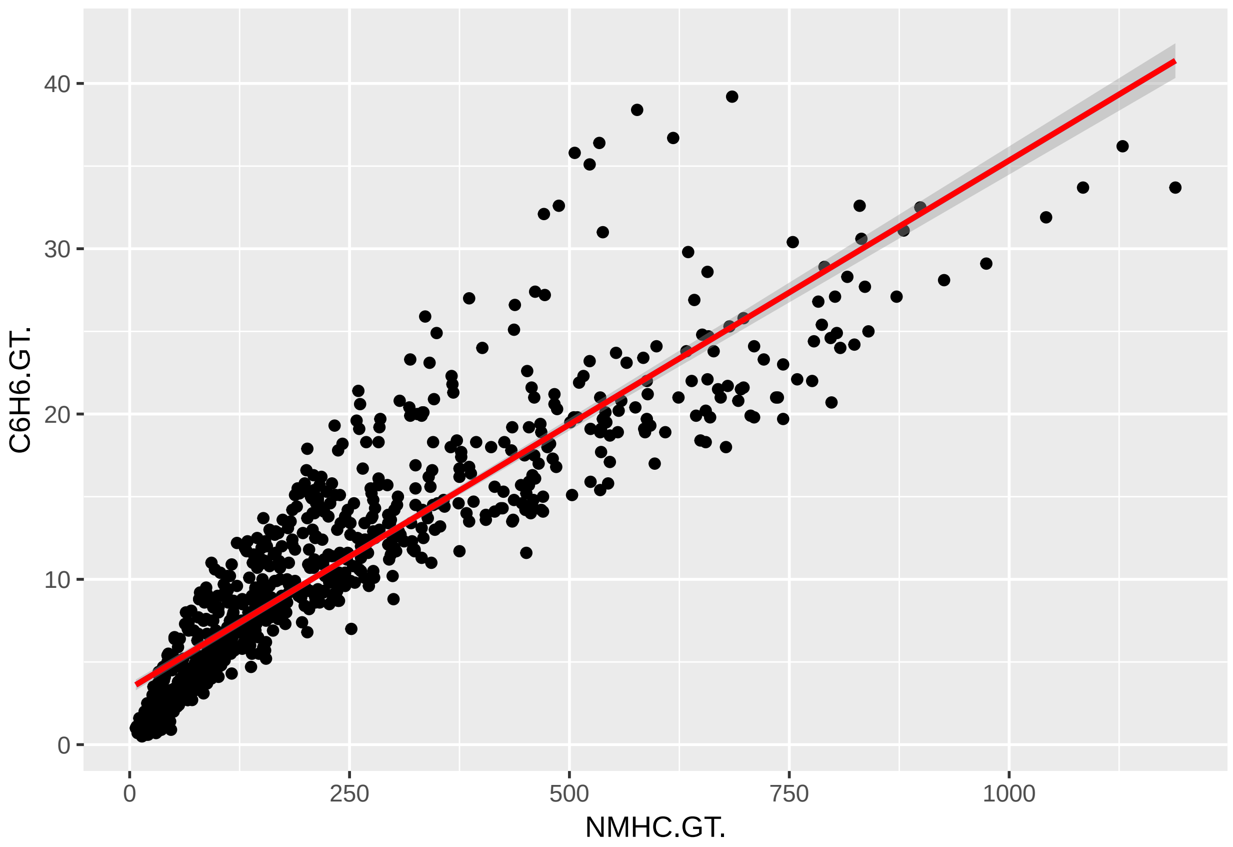
```
ggplot(lm_2, aes(x = PT08.S1.CO., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```



```r
lm_3 <- lm(C6H6.GT. ~ NMHC.GT., data = airq_new)
summary(lm_3)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ NMHC.GT., data = airq_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1876 -2.0558 -0.6626  1.3815 16.5740
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.3891818  0.1696364   19.98   <2e-16 ***
## NMHC.GT.    0.0319528  0.0005453   58.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.267 on 825 degrees of freedom
## Multiple R-squared:  0.8063, Adjusted R-squared:  0.806
## F-statistic:  3434 on 1 and 825 DF,  p-value: < 2.2e-16
```

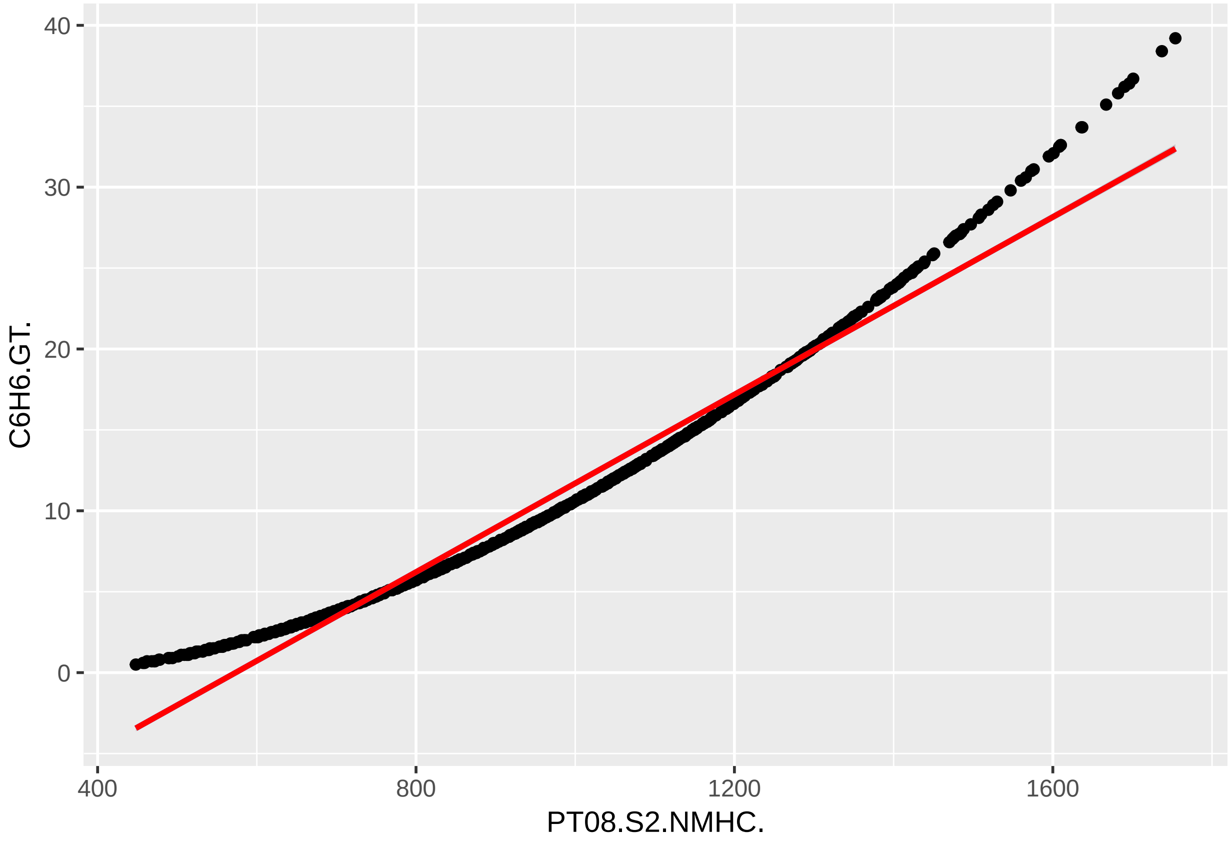
```r
ggplot(lm_3, aes(x = NMHC.GT., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```



```
lm_4 <- lm(C6H6.GT. ~ PT08.S2.NMHC., data = airq_new)
summary(lm_4)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S2.NMHC., data = airq_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1470 -0.9581 -0.4612  0.5492  6.8243
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.572e+01  1.685e-01  -93.27   <2e-16 ***
## PT08.S2.NMHC.  2.742e-02  1.682e-04  163.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.288 on 825 degrees of freedom
## Multiple R-squared:  0.9699, Adjusted R-squared:  0.9699
## F-statistic: 2.658e+04 on 1 and 825 DF,  p-value: < 2.2e-16
```

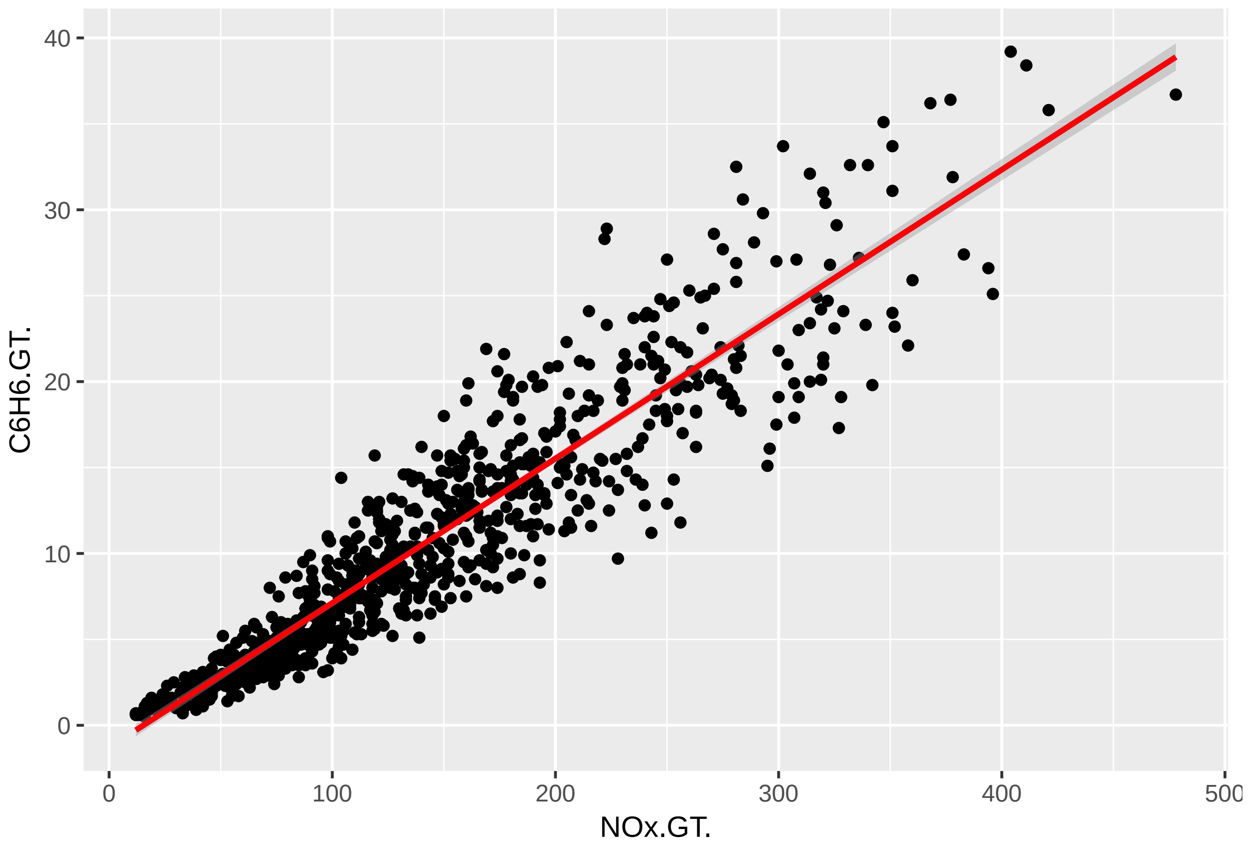
```
ggplot(lm_4, aes(x = PT08.S2.NMHC., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```



```
lm_5 <- lm(C6H6.GT. ~ NOx.GT., data = airq_new)
summary(lm_5)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ NOx.GT., data = airq_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8965 -1.5222 -0.1907  1.2497 11.4460
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.292115   0.195126  -6.622 6.39e-11 ***
## NOx.GT.      0.084063   0.001181  71.157  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.778 on 825 degrees of freedom
## Multiple R-squared:  0.8599, Adjusted R-squared:  0.8597
## F-statistic:  5063 on 1 and 825 DF,  p-value: < 2.2e-16
```

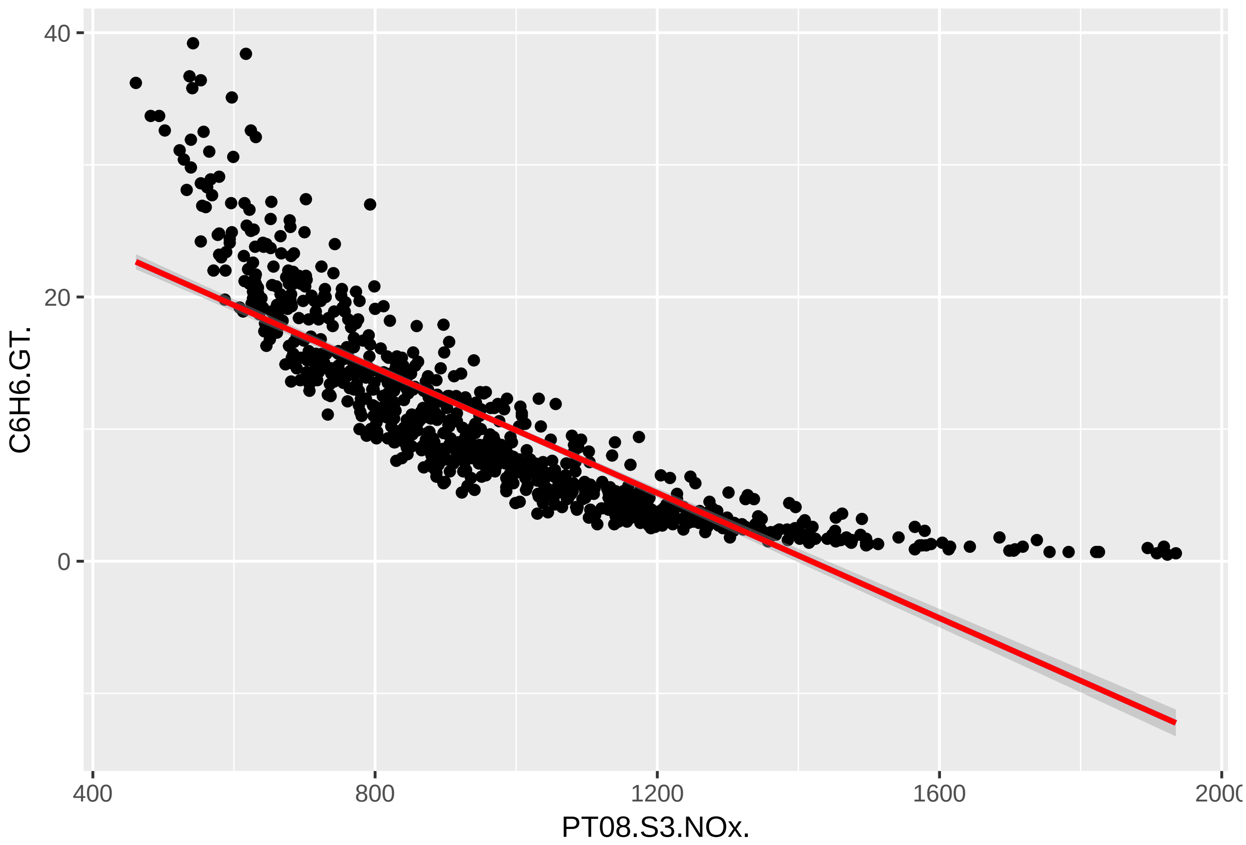
```
ggplot(lm_5, aes(x = NOx.GT., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```



```r
lm_6 <- lm(C6H6.GT. ~ PT08.S3.NOx., data = airq_new)
summary(lm_6)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S3.NOx., data = airq_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5253 -2.6883 -0.9271  1.7269 19.4285
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.5821174  0.5130616   65.45   <2e-16 ***
## PT08.S3.NOx. -0.0236801  0.0005134  -46.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.924 on 825 degrees of freedom
## Multiple R-squared:  0.7205, Adjusted R-squared:  0.7202
## F-statistic:  2127 on 1 and 825 DF,  p-value: < 2.2e-16
```

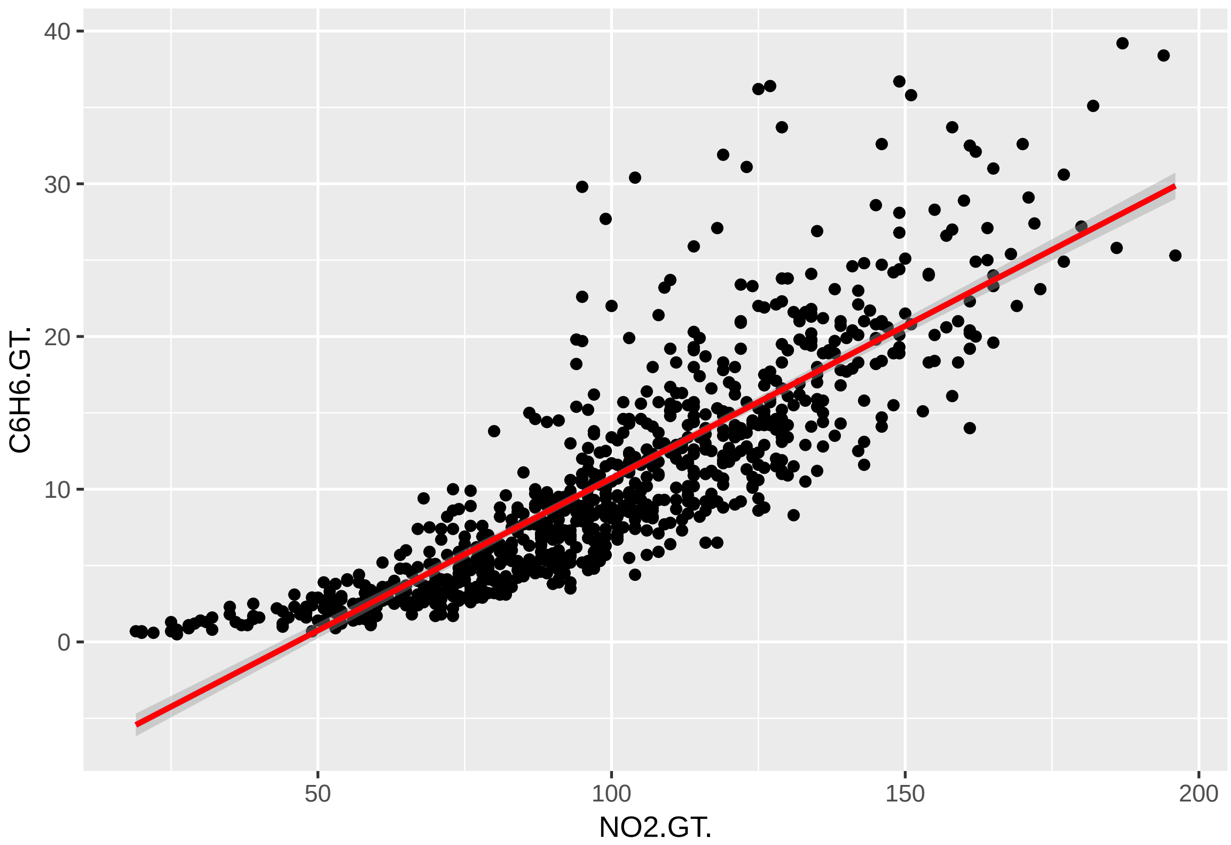
```r
ggplot(lm_6, aes(x = PT08.S3.NOx., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```



```
lm_7 <- lm(C6H6.GT. ~ NO2.GT., data = airq_new)
summary(lm_7)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ NO2.GT., data = airq_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8853 -2.5243 -0.5853  1.7552 20.4947
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.225139   0.458452  -20.12   <2e-16 ***
## NO2.GT.      0.199444   0.004363   45.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.949 on 825 degrees of freedom
## Multiple R-squared:  0.717,  Adjusted R-squared:  0.7166
## F-statistic:  2090 on 1 and 825 DF,  p-value: < 2.2e-16
```

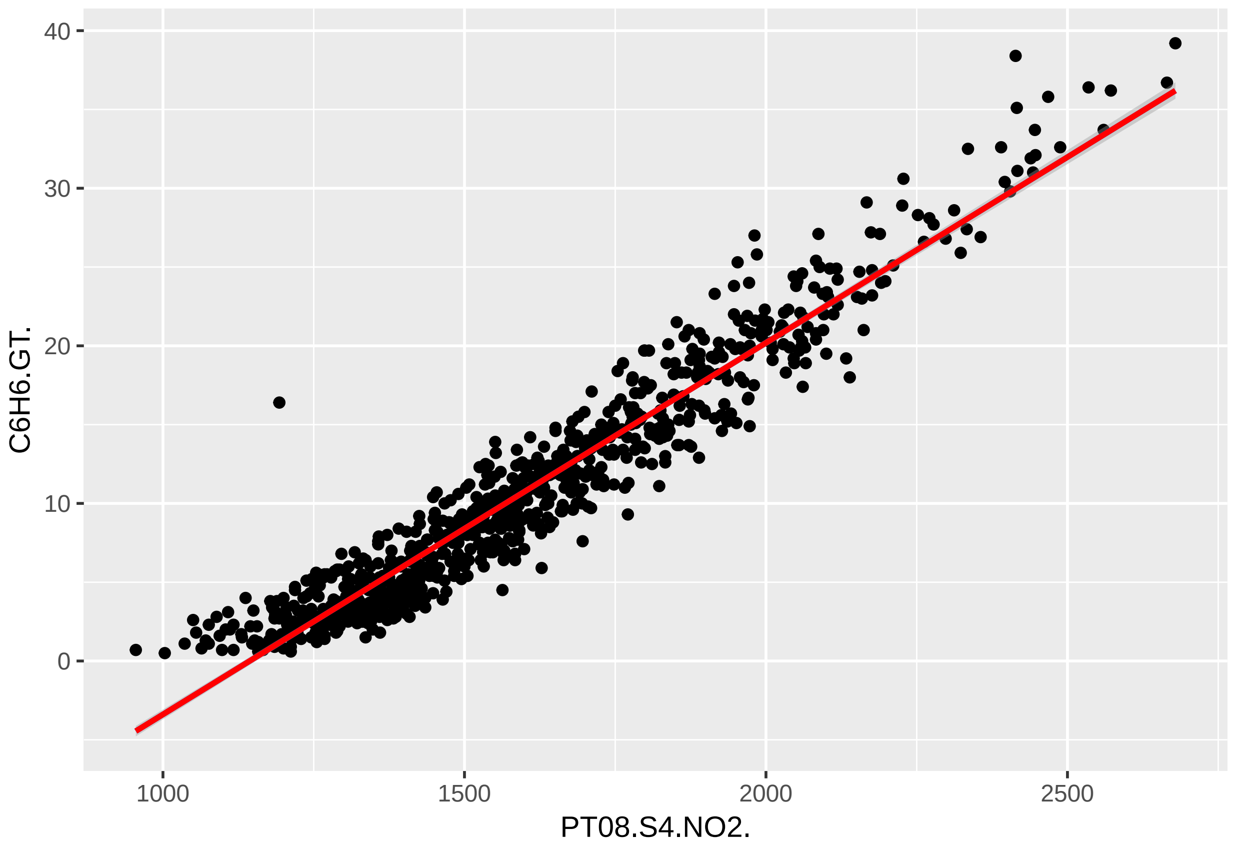
```
ggplot(lm_7, aes(x = NO2.GT., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```



```
lm_8 <- lm(C6H6.GT. ~ PT08.S4.NO2., data = airq_new)
summary(lm_8)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S4.NO2., data = airq_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5167 -1.4177  0.0103  1.1915 15.2398
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.697e+01  3.858e-01  -69.91   <2e-16 ***
## PT08.S4.NO2.  2.358e-02  2.368e-04   99.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.058 on 825 degrees of freedom
## Multiple R-squared:  0.9232, Adjusted R-squared:  0.9231
## F-statistic:  9911 on 1 and 825 DF,  p-value: < 2.2e-16
```

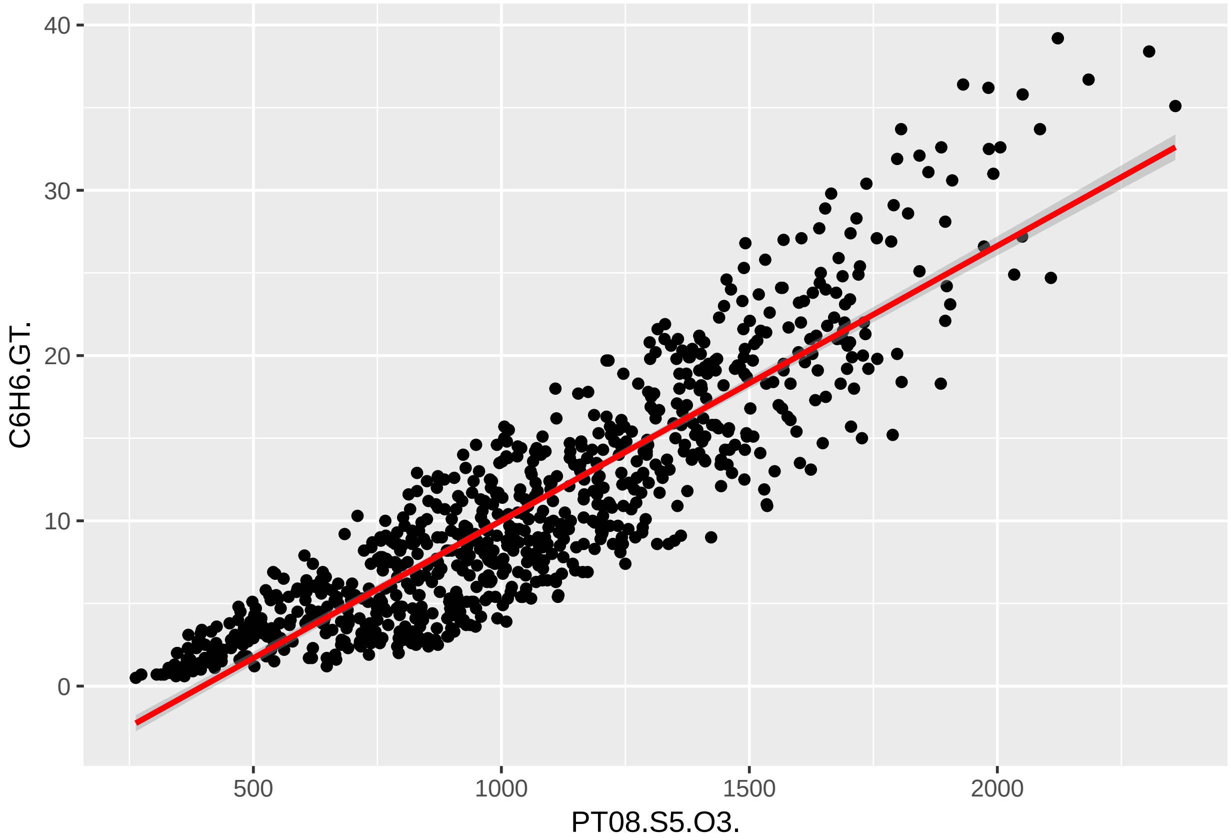
```
ggplot(lm_8, aes(x = PT08.S4.NO2., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```



```
lm_9 <- lm(C6H6.GT. ~ PT08.S5.O3., data = airq_new)
summary(lm_9)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S5.O3., data = airq_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0434 -2.5352  0.2444  2.1773 10.9090
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.6198822  0.3194802  -20.72   <2e-16 ***
## PT08.S5.O3.  0.0166292  0.0002853   58.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.281 on 825 degrees of freedom
## Multiple R-squared:  0.8046, Adjusted R-squared:  0.8043
## F-statistic:  3396 on 1 and 825 DF,  p-value: < 2.2e-16
```

```
ggplot(lm_9, aes(x = PT08.S5.O3., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  geom_line(aes(y = .fitted), color = "red", size = 1)
```

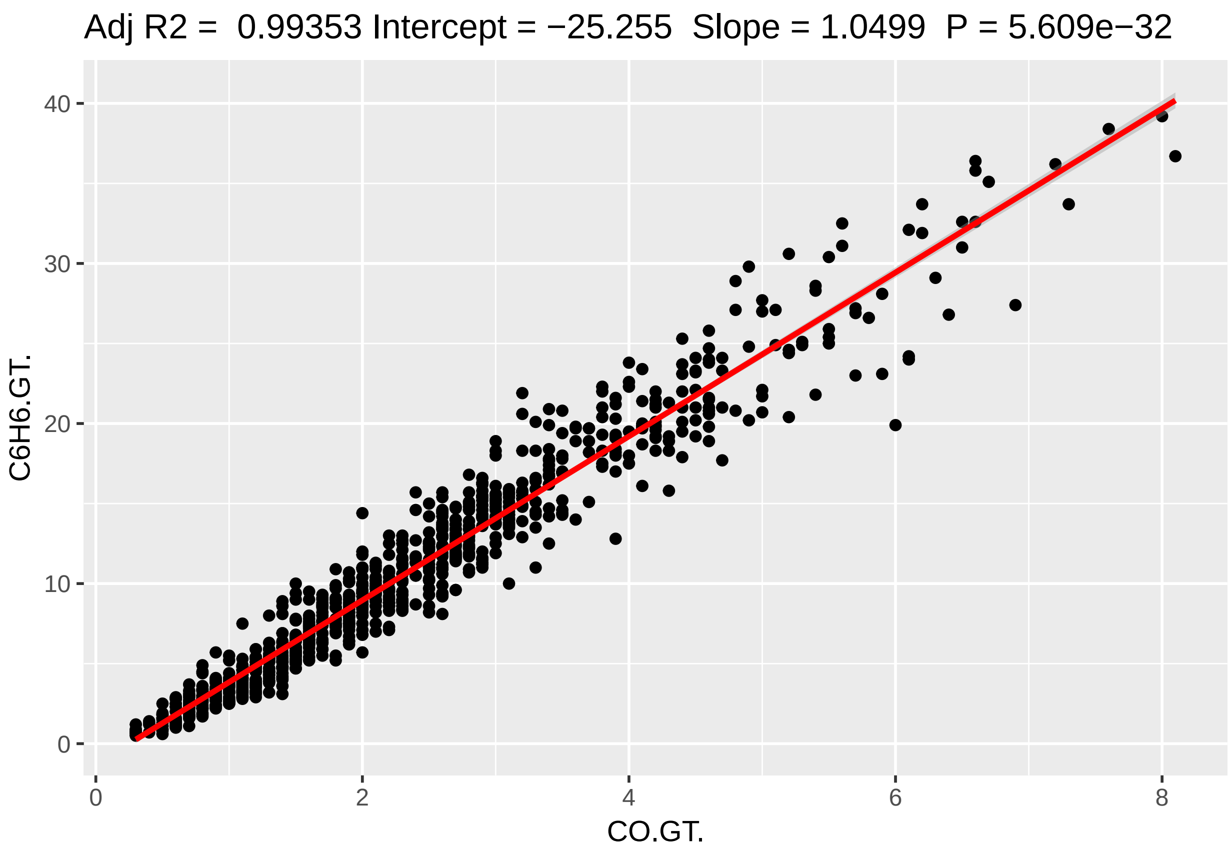


```
l_model <- lm(C6H6.GT. ~ ., data = airq_new[,-c(1,2)])

summary(l_model)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ ., data = airq_new[, -c(1, 2)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6771 -0.3765 -0.0190  0.3200  3.4535
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.525e+01  6.323e-01 -39.943  < 2e-16 ***
## CO.GT.         1.050e+00  8.541e-02  12.292  < 2e-16 ***
## PT08.S1.CO.   -2.765e-03  3.917e-04  -7.058 3.63e-12 ***
## NMHC.GT.       2.351e-03  2.525e-04   9.310  < 2e-16 ***
## PT08.S2.NMHC.  2.166e-02  7.740e-04  27.977  < 2e-16 ***
## NOx.GT.       -2.297e-03  1.009e-03  -2.277   0.0231 *
## PT08.S3.NOx.   6.408e-03  2.710e-04  23.645  < 2e-16 ***
## NO2.GT.       -1.367e-02  1.780e-03  -7.680 4.56e-14 ***
## PT08.S4.NO2.   6.454e-03  5.341e-04  12.084  < 2e-16 ***
## PT08.S5.O3.    1.436e-03  1.701e-04   8.446  < 2e-16 ***
## T              1.329e-02  2.096e-02   0.634   0.5263
## RH            -1.221e-02  7.195e-03  -1.697   0.0900 .
## AH            -5.805e-01  4.901e-01  -1.184   0.2366
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5966 on 814 degrees of freedom
## Multiple R-squared:  0.9936, Adjusted R-squared:  0.9935
## F-statistic: 1.058e+04 on 12 and 814 DF,  p-value: < 2.2e-16
```

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



```
set.seed(42)

new_dataset <- airq_new[,3:15]

sample <- sample.int(n = nrow(new_dataset),
                     size = floor(.75*nrow(new_dataset)))

training_set  <- new_dataset[sample,]

test_set  <- new_dataset[-sample,]
```

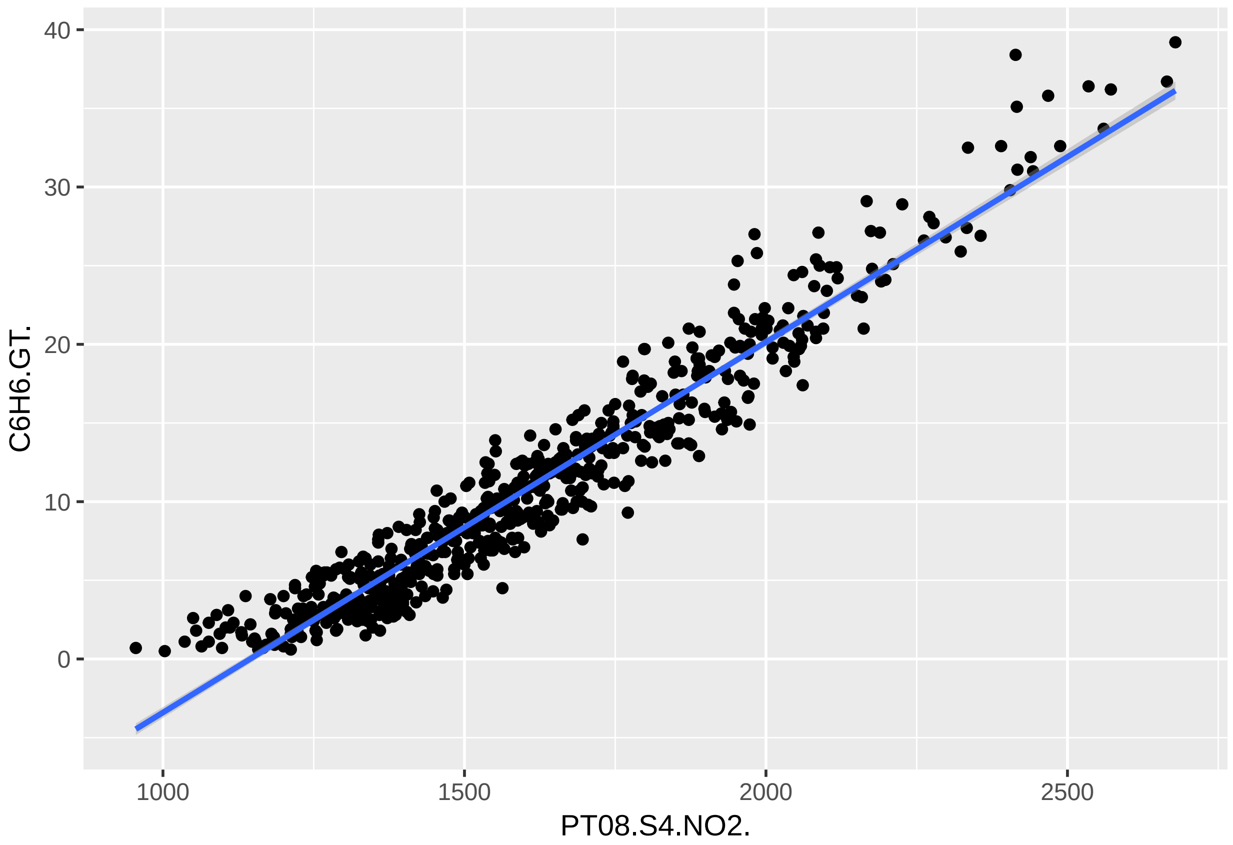
```r
new_fit<- lm(C6H6.GT. ~ PT08.S4.NO2., data=training_set)
summary(new_fit)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S4.NO2., data = training_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4510 -1.4160  0.0404  1.1696  8.5100
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.695e+01  4.270e-01  -63.10   <2e-16 ***
## PT08.S4.NO2.  2.354e-02  2.621e-04   89.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.006 on 618 degrees of freedom
## Multiple R-squared:  0.9289, Adjusted R-squared:  0.9288
## F-statistic:  8071 on 1 and 618 DF,  p-value: < 2.2e-16
```

```r
str(summary(new_fit))
```

```
## List of 11
##  $ call         : language lm(formula = C6H6.GT. ~ PT08.S4.NO2., data = training_set)
##  $ terms        :Classes 'terms', 'formula'  language C6H6.GT. ~ PT08.S4.NO2.
##   .. ..- attr(*, "variables")= language list(C6H6.GT., PT08.S4.NO2.)
##   .. ..- attr(*, "factors")= int [1:2, 1] 0 1
##   .. .. ..- attr(*, "dimnames")=List of 2
##   .. .. .. ..$ : chr [1:2] "C6H6.GT." "PT08.S4.NO2."
##   .. .. .. ..$ : chr "PT08.S4.NO2."
##   .. ..- attr(*, "term.labels")= chr "PT08.S4.NO2."
##   .. ..- attr(*, "order")= int 1
##   .. ..- attr(*, "intercept")= int 1
##   .. ..- attr(*, "response")= int 1
##   .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##   .. ..- attr(*, "predvars")= language list(C6H6.GT., PT08.S4.NO2.)
##   .. ..- attr(*, "dataClasses")= Named chr [1:2] "numeric" "numeric"
##   .. .. ..- attr(*, "names")= chr [1:2] "C6H6.GT." "PT08.S4.NO2."
##  $ residuals    : Named num [1:620] 2.31 1.94 3.04 -3.63 2.03 ...
##   ..- attr(*, "names")= chr [1:620] "1155" "1175" "362" "1077" ...
##  $ coefficients : num [1:2, 1:4] -2.69e+01 2.35e-02 4.27e-01 2.62e-04 -6.31e+01 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "(Intercept)" "PT08.S4.NO2."
##   .. ..$ : chr [1:4] "Estimate" "Std. Error" "t value" "Pr(>|t|)"
##  $ aliased      : Named logi [1:2] FALSE FALSE
##   ..- attr(*, "names")= chr [1:2] "(Intercept)" "PT08.S4.NO2."
##  $ sigma        : num 2.01
##  $ df           : int [1:3] 2 618 2
##  $ r.squared    : num 0.929
##  $ adj.r.squared: num 0.929
##  $ fstatistic   : Named num [1:3] 8071 1 618
##   ..- attr(*, "names")= chr [1:3] "value" "numdf" "dendf"
##  $ cov.unscaled : num [1:2, 1:2] 4.53e-02 -2.73e-05 -2.73e-05 1.71e-08
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "(Intercept)" "PT08.S4.NO2."
##   .. ..$ : chr [1:2] "(Intercept)" "PT08.S4.NO2."
##  - attr(*, "class")= chr "summary.lm"
```

```
ggplot(data = training_set, aes(x = PT08.S4.NO2., y = C6H6.GT.))+
  geom_point() +
  geom_smooth(method = "lm")
```



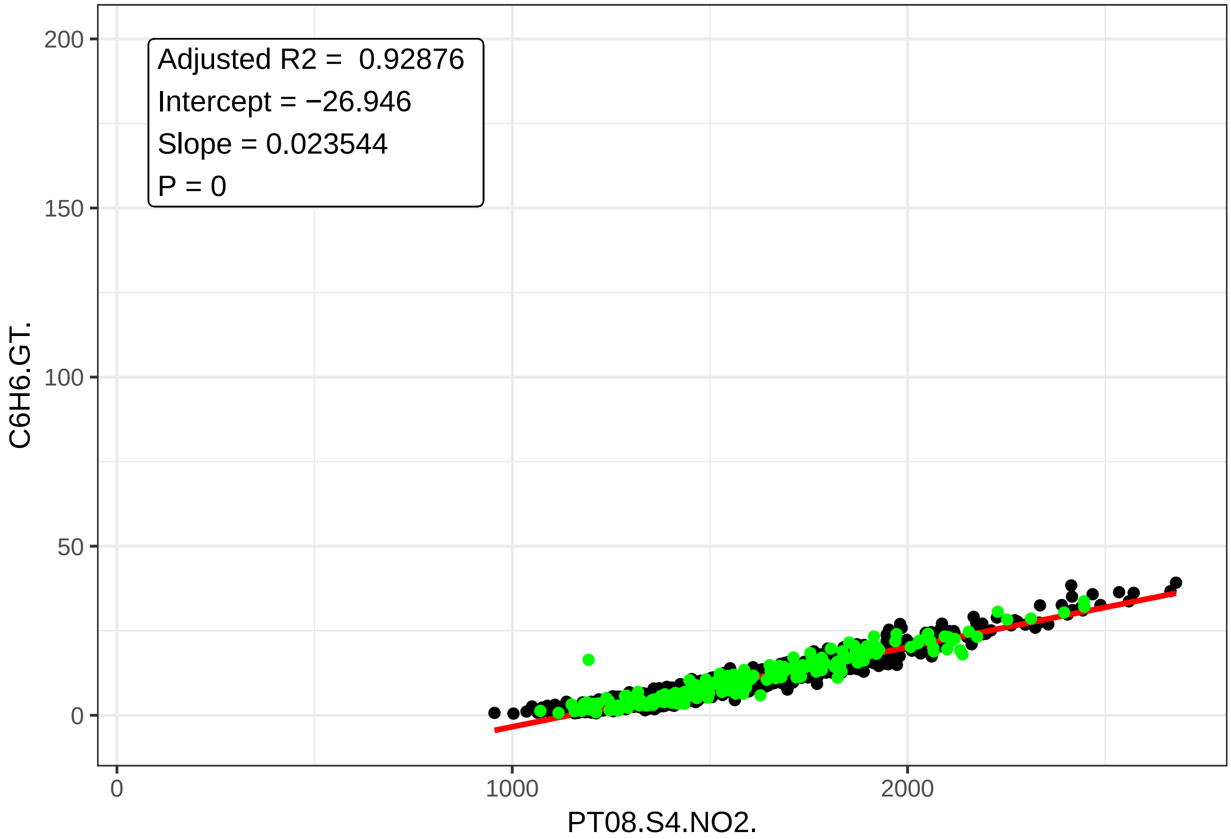
```
pred <- predict(new_fit, newdata = test_set)
head(pred)
```

```
##         3         8         9        16        17        21
##  9.665456  4.438621  3.096595 10.324697  7.099127 13.785709
```

```
test_set$C6H6.GT._pred <- pred
head(test_set)
```

```
##    CO.GT. PT08.S1.CO. NMHC.GT. C6H6.GT. PT08.S2.NMHC. NOx.GT. PT08.S3.NOx.
## 3     2.2        1402       88      9.0           939     131         1140
## 8     1.0        1136       31      3.3           672      62         1453
## 9     0.9        1094       24      2.3           609      45         1579
## 16    2.2        1351       87      9.5           960     129         1079
## 17    1.7        1233       77      6.3           827     112         1218
## 21    2.9        1371      164     11.5          1034     207          983
##    NO2.GT. PT08.S4.NO2. PT08.S5.O3.    T   RH     AH C6H6.GT._pred
## 3      114         1555        1074 11.9 54.0 0.7502      9.665456
## 8       76          1333        730 10.7 60.0 0.7702      4.438621
## 9       60          1276        620 10.7 59.7 0.7648      3.096595
## 16     101          1583       1028 10.5 60.6 0.7691     10.324697
## 17      98          1446        860 10.8 58.4 0.7552      7.099127
## 21     128          1730       1037  8.0 81.1 0.8736     13.785709
```

```
ggplot(training_set, aes(x = PT08.S4.NO2., y = C6H6.GT.)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red" ) +
  geom_point(data = test_set, aes(y = C6H6.GT.), color = "green") +
  theme_bw() +
  geom_label(aes(x = 80, y = 200), hjust = 0, vjust = 1,
             label = paste("Adjusted R2 = ",signif(summary(new_fit)$adj.r.squared, 5),
                                               "\nIntercept =",signif(new_fit$coef[[1]],5 ),
                                               " \nSlope =",signif(new_fit$coef[[2]], 5),
                                               " \nP =",signif(summary(new_fit)$coef[2,4], 5)))
```


Adjusted R2 = 0.92876
Intercept = −26.946
Slope = 0.023544
P = 0
C6H6.GT.
PT08.S4.NO2.