HW_2_1

```
library(data.table)
library(dplyr)
library(ggplot2)
library(ggpubr)
library(tidyr)
library(reshape2)
library(plyr)
```

Task1. Q1. Scatter plot facetted by set

```
anscombe <- as.data.frame(anscombe)
head(anscombe)
```

```
##   x1 x2 x3 x4   y1   y2    y3   y4
## 1 10 10 10  8 8.04 9.14  7.46 6.58
## 2  8  8  8  8 6.95 8.14  6.77 5.76
## 3 13 13 13  8 7.58 8.74 12.74 7.71
## 4  9  9  9  8 8.81 8.77  7.11 8.84
## 5 11 11 11  8 8.33 9.26  7.81 8.47
## 6 14 14 14  8 9.96 8.10  8.84 7.04
```

```
nrow(anscombe)
```

```
## [1] 11
```

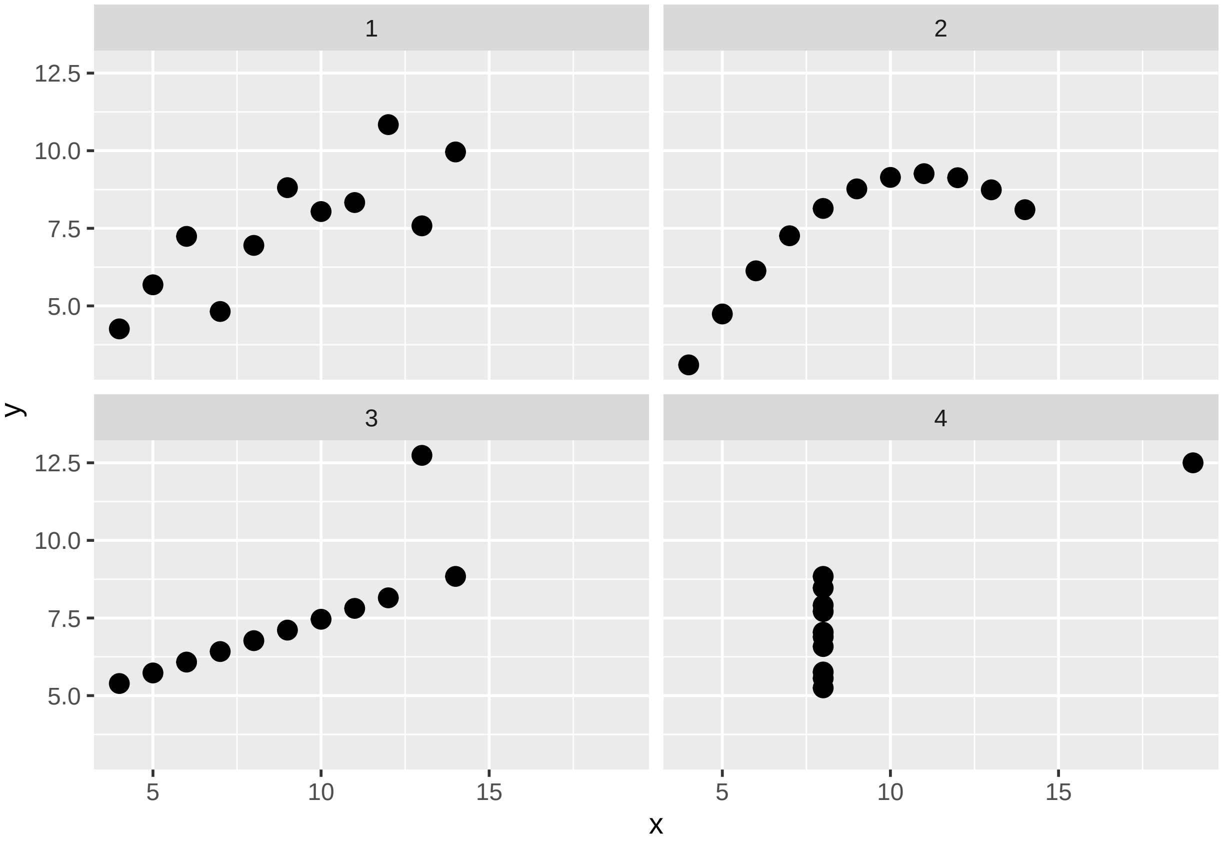
```
#generate levels to indicate which group each data point belong to
levels <- gl(4, nrow(anscombe))
levels
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4
## [39] 4 4 4 4 4 4
## Levels: 1 2 3 4
```

```
#Group it in dataframe
anscombe_grouped <- with(anscombe, data.frame(x=c(x1,x2,x3,x4),y=c(y1,y2,y3,y4), set=levels))
anscombe_grouped
```

```
##     x     y set
## 1  10  8.04   1
## 2   8  6.95   1
## 3  13  7.58   1
## 4   9  8.81   1
## 5  11  8.33   1
## 6  14  9.96   1
## 7   6  7.24   1
## 8   4  4.26   1
## 9  12 10.84   1
## 10  7  4.82   1
## 11  5  5.68   1
## 12 10  9.14   2
## 13  8  8.14   2
## 14 13  8.74   2
## 15  9  8.77   2
## 16 11  9.26   2
## 17 14  8.10   2
## 18  6  6.13   2
## 19  4  3.10   2
## 20 12  9.13   2
## 21  7  7.26   2
## 22  5  4.74   2
## 23 10  7.46   3
## 24  8  6.77   3
## 25 13 12.74   3
## 26  9  7.11   3
## 27 11  7.81   3
## 28 14  8.84   3
## 29  6  6.08   3
## 30  4  5.39   3
## 31 12  8.15   3
## 32  7  6.42   3
## 33  5  5.73   3
## 34  8  6.58   4
## 35  8  5.76   4
## 36  8  7.71   4
## 37  8  8.84   4
## 38  8  8.47   4
## 39  8  7.04   4
## 40  8  5.25   4
## 41 19 12.50   4
## 42  8  5.56   4
## 43  8  7.91   4
## 44  8  6.89   4
```

```r
#Make scattterplots
ggplot(anscombe_grouped, aes(x,y))+
  geom_point(size=3) +
  facet_wrap(~set)
```



Q2. Summary calculation(mean, sd) grouped by set

```
#Mean
aggregate(cbind(x, y) ~ set, anscombe_grouped, mean)
```

```
##   set x        y
## 1   1 9 7.500909
## 2   2 9 7.500909
## 3   3 9 7.500000
## 4   4 9 7.500909
```

```
#SD
aggregate(cbind(x, y) ~ set, anscombe_grouped, sd)
```

```
##   set        x        y
## 1   1 3.316625 2.031568
## 2   2 3.316625 2.031657
## 3   3 3.316625 2.030424
## 4   4 3.316625 2.030579
```

Q3. Pearson’s correlation by set and non-parametric, and p-value

```
anscombe_grouped%>% group_by(set) %>% summarise(cor_pearson = cor.test(x,y, method = 'pearson')$estimat
                                      cor_kendall = cor.test(x,y, method = 'kendall')$estimate,
                                      cor_spearmen = cor.test(x,y, method = 'spearman')$estimate)
```

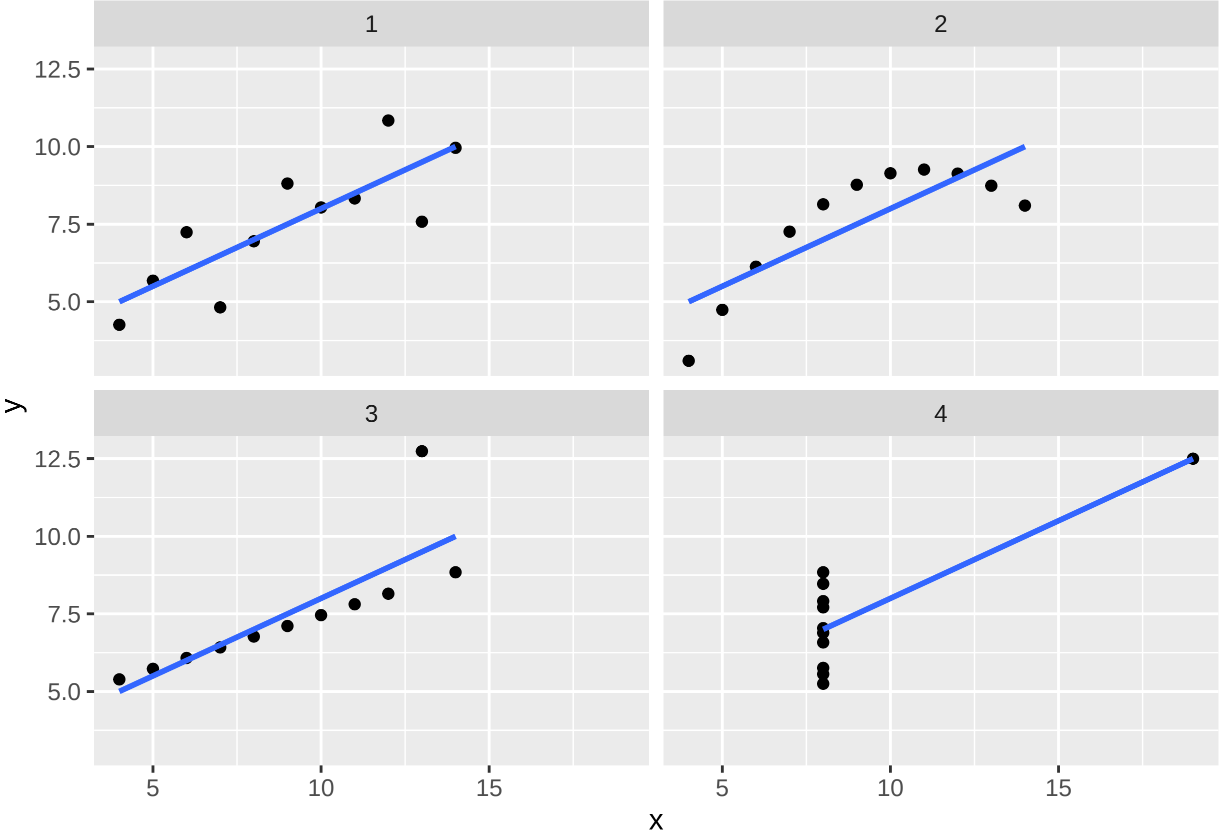
```
##   cor_pearson cor_kendall cor_spearmen
## 1   0.8163662   0.6618771    0.8168855
```

```
anscombe_grouped %>% group_by(set) %>% summarise(p_pearson = cor.test(x,y, method = 'pearson')$p.value,
                                     p_kendall = cor.test(x,y, method = 'kendall')$p.value,
                                     p_spearmen = cor.test(x,y, method = 'spearman')$p.value)
```

```
##       p_pearson    p_kendall   p_spearmen
## 1 1.436505e-11 1.422967e-09 1.360916e-11
```

Q4. Add geom smooth() to the plot

```
ggplot( anscombe_grouped, aes(x, y)) +
  geom_point() +
  geom_smooth(method = lm, se = F) +
  facet_wrap(~ set)
```



Task2. Q1. Explore data set, clean if needed

```
aq <- read.csv2("/Users/Lisa/Downloads/AirQualityUCI/AirQualityUCI.csv", header =T)
str(aq)
```

```
## '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 ...
```

```
#We see some NA. Let's remove them
colSums(is.na(aq))
```

```
##          Date          Time        CO.GT.   PT08.S1.CO.      NMHC.GT.
##             0             0           114           114           114
##      C6H6.GT. PT08.S2.NMHC.       NOx.GT.  PT08.S3.NOx.       NO2.GT.
##           114           114           114           114           114
##  PT08.S4.NO2.   PT08.S5.O3.             T            RH            AH
##           114           114           114           114           114
##             X           X.1
##          9471          9471
```

```
aq_cleaned <-  aq  %>% select_if(~sum(!is.na(.)) > 0)  %>% drop_na()
head(aq_cleaned)
```

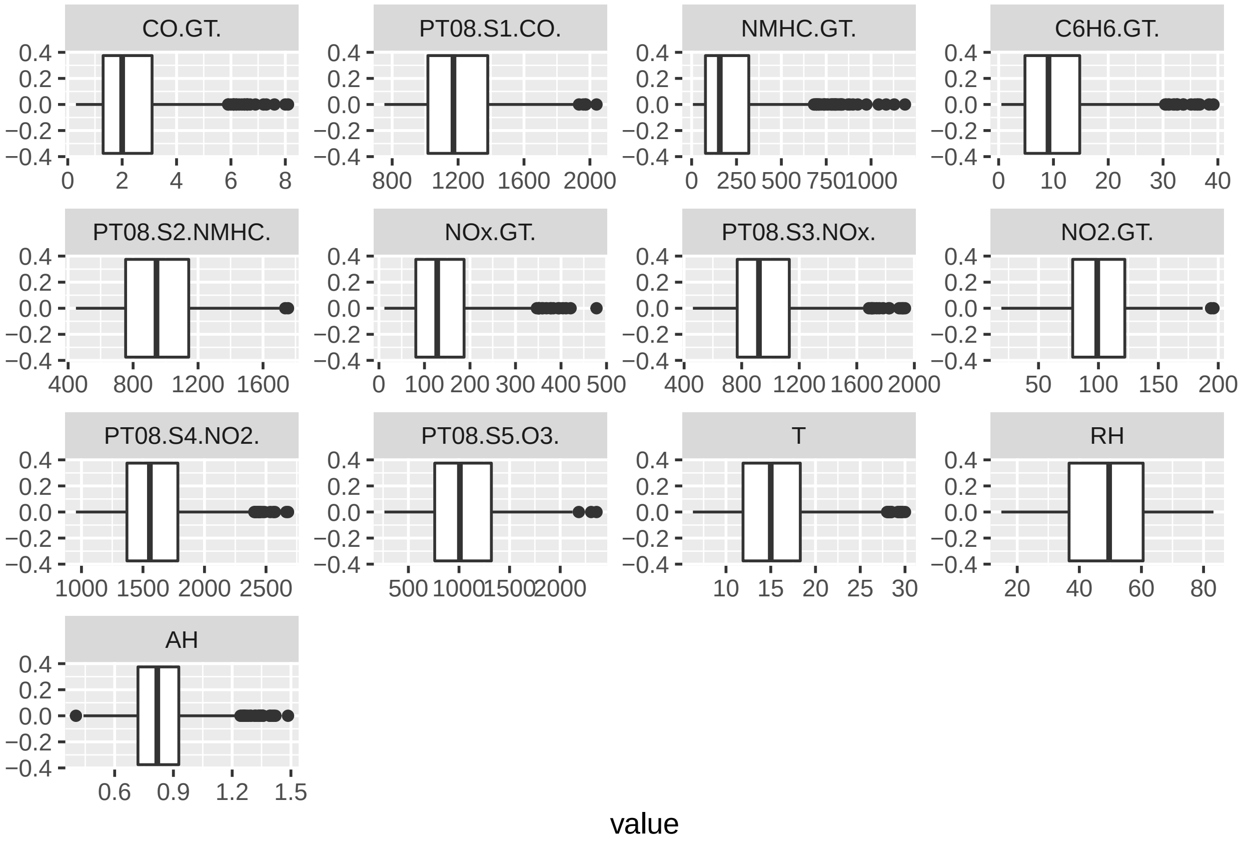
```
##         Date     Time CO.GT. PT08.S1.CO. NMHC.GT. C6H6.GT. PT08.S2.NMHC.
## 1 10/03/2004 18.00.00    2.6        1360      150     11.9          1046
## 2 10/03/2004 19.00.00    2.0        1292      112      9.4           955
## 3 10/03/2004 20.00.00    2.2        1402       88      9.0       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

```
#Delete strange variables that contain value 200
airq_fil <- aq_cleaned %>% filter_all(all_vars(. != -200))
#Convert to long dataset
air_long <- melt(airq_fil)
head(air_long)
```

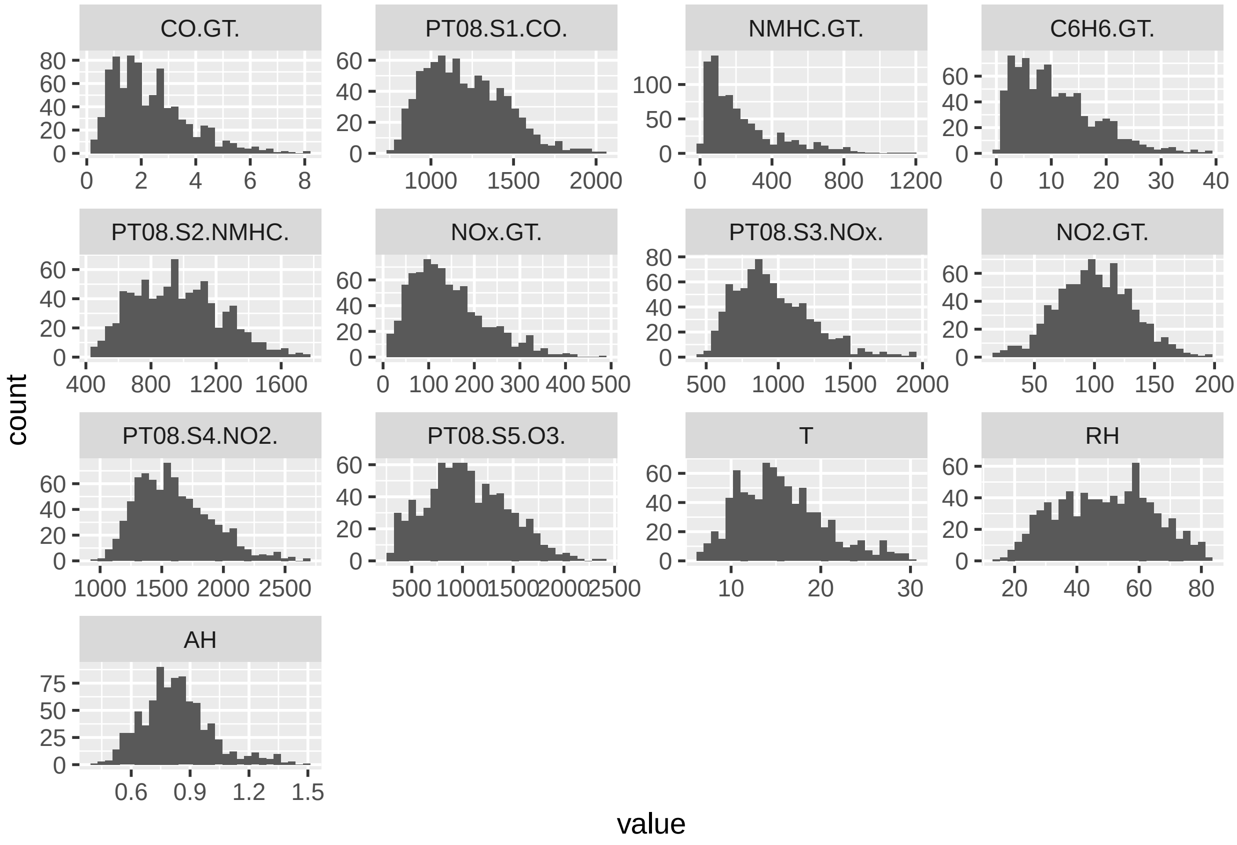
```
##         Date     Time variable value
## 1 10/03/2004 18.00.00   CO.GT.   2.6
## 2 10/03/2004 19.00.00   CO.GT.   2.0
## 3 10/03/2004 20.00.00   CO.GT.   2.2
## 4 10/03/2004 21.00.00   CO.GT.   2.2
## 5 10/03/2004 22.00.00   CO.GT.   1.6
## 6 10/03/2004 23.00.00   CO.GT.   1.2
```

Q2.Explore each variable independently

```
  ggplot(air_long, aes(value)) +
  geom_boxplot() +
  facet_wrap(~variable, scales="free")
```



```
ggplot(air_long, aes(value)) +
 geom_histogram() +
 facet_wrap(~variable, scales="free")
```

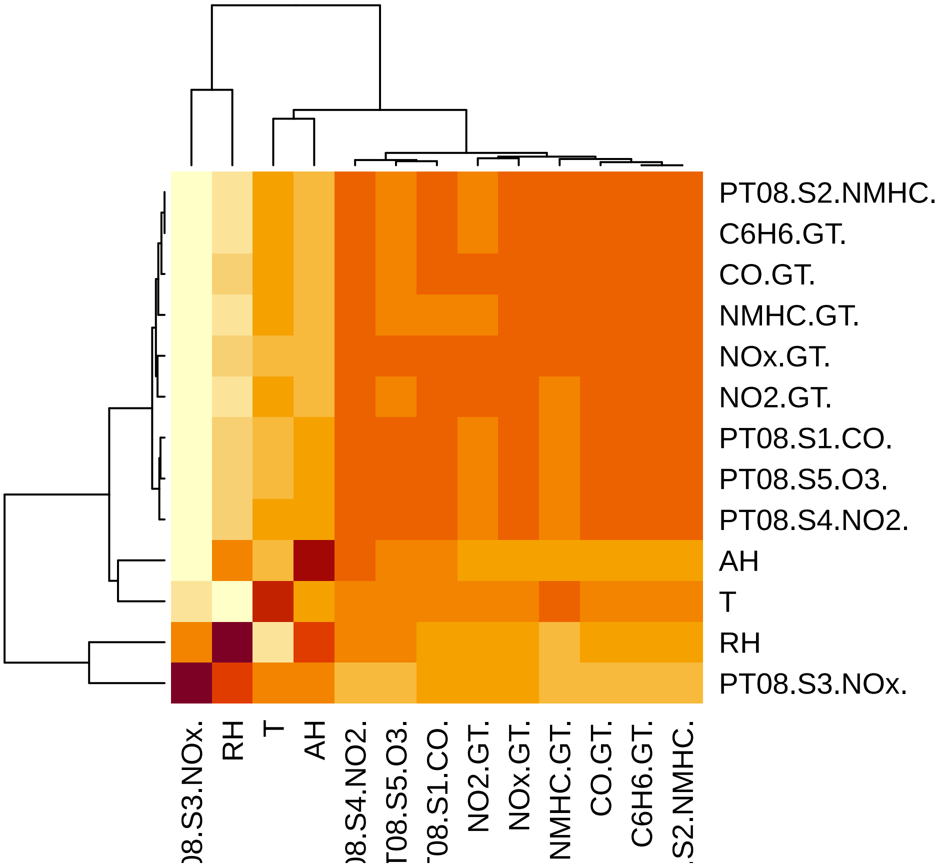


Q3.Cross correlation

```
correl <- as.matrix(airq_fil[,c(3:15)])
cor(correl, use="complete.obs", method="spearman")
```

```
##                   CO.GT. PT08.S1.CO.   NMHC.GT.   C6H6.GT. PT08.S2.NMHC.
## CO.GT.         1.0000000  0.93978597  0.9342883  0.9766091     0.9766404
## PT08.S1.CO.    0.9397860  1.00000000  0.8386699  0.9348855     0.9349883
## NMHC.GT.       0.9342883  0.83866992  1.0000000  0.9454609     0.9454232
## C6H6.GT.       0.9766091  0.93488550  0.9454609  1.0000000     0.9999832
## PT08.S2.NMHC.  0.9766404  0.93498826  0.9454232  0.9999832     1.0000000
## NOx.GT.        0.9615010  0.92622189  0.8805828  0.9431609     0.9432691
## PT08.S3.NOx.  -0.9235819 -0.89118253 -0.9461075 -0.9526035    -0.9526213
## NO2.GT.        0.9172671  0.88351864  0.8354841  0.8989692     0.8990827
## PT08.S4.NO2.   0.9330848  0.94639571  0.8924281  0.9576110     0.9576513
## PT08.S5.O3.    0.8828508  0.93386561  0.8017011  0.9002137     0.9002616
## T              0.3910139  0.34464297  0.4784004  0.4674760     0.4671404
## RH            -0.1632868 -0.05029757 -0.2410037 -0.2154793    -0.2151061
## AH             0.2947413  0.43470088  0.3163970  0.3321598     0.3323398
##                   NOx.GT. PT08.S3.NOx.    NO2.GT. PT08.S4.NO2.  PT08.S5.O3.
## CO.GT.         0.96150101  -0.92358193  0.9172671   0.93308480  0.882850756
## PT08.S1.CO.    0.92622189  -0.89118253  0.8835186   0.94639571  0.933865605
## NMHC.GT.       0.88058277  -0.94610750  0.8354841   0.89242814  0.801701058
## C6H6.GT.       0.94316091  -0.95260345  0.8989692   0.95761101  0.900213719
## PT08.S2.NMHC.  0.94326915  -0.95262132  0.8990827   0.95765129  0.900261575
## NOx.GT.        1.00000000  -0.89048355  0.9088018   0.91767242  0.891077771
## PT08.S3.NOx.  -0.89048355   1.00000000 -0.8205646  -0.95473807 -0.897821592
## NO2.GT.        0.90880183  -0.82056461  1.0000000   0.84413499  0.835499186
## PT08.S4.NO2.    0.91767242  -0.95473807  0.8441350   1.00000000  0.926023708
## PT08.S5.O3.     0.89107777  -0.89782159  0.8354992   0.92602371  1.000000000
## T               0.30603518  -0.45075496  0.3991110   0.38412086  0.321566434
## RH             -0.08554217   0.09876639 -0.2250338  -0.01714565 -0.004638448
## AH              0.29736842  -0.50185884  0.2386725   0.53597752  0.479249063
##                         T            RH          AH
## CO.GT.          0.3910139 -0.163286796  0.2947413
## PT08.S1.CO.     0.3446430 -0.050297575  0.4347009
## NMHC.GT.        0.4784004 -0.241003737  0.3163970
## C6H6.GT.        0.4674760 -0.215479340  0.3321598
## PT08.S2.NMHC.   0.4671404 -0.215106064  0.3323398
## NOx.GT.         0.3060352 -0.085542171  0.2973684
## PT08.S3.NOx.   -0.4507550  0.098766391 -0.5018588
## NO2.GT.         0.3991110 -0.225033805  0.2386725
## PT08.S4.NO2.    0.3841209 -0.017145651  0.5359775
## PT08.S5.O3.     0.3215664 -0.004638448  0.4792491
## T               1.0000000 -0.778027606  0.1490804
## RH             -0.7780276  1.000000000  0.4512274
## AH              0.1490804  0.451227370  1.0000000
```

```r
heatmap(cor(correl, use="complete.obs", method="spearman"))
```



Q4. Response C6H6(GT)

```r
airq_fil %>%
  lm(data = ., airq_fil$C6H6.GT. ~ airq_fil$CO.GT.) %>%
  summary()
```

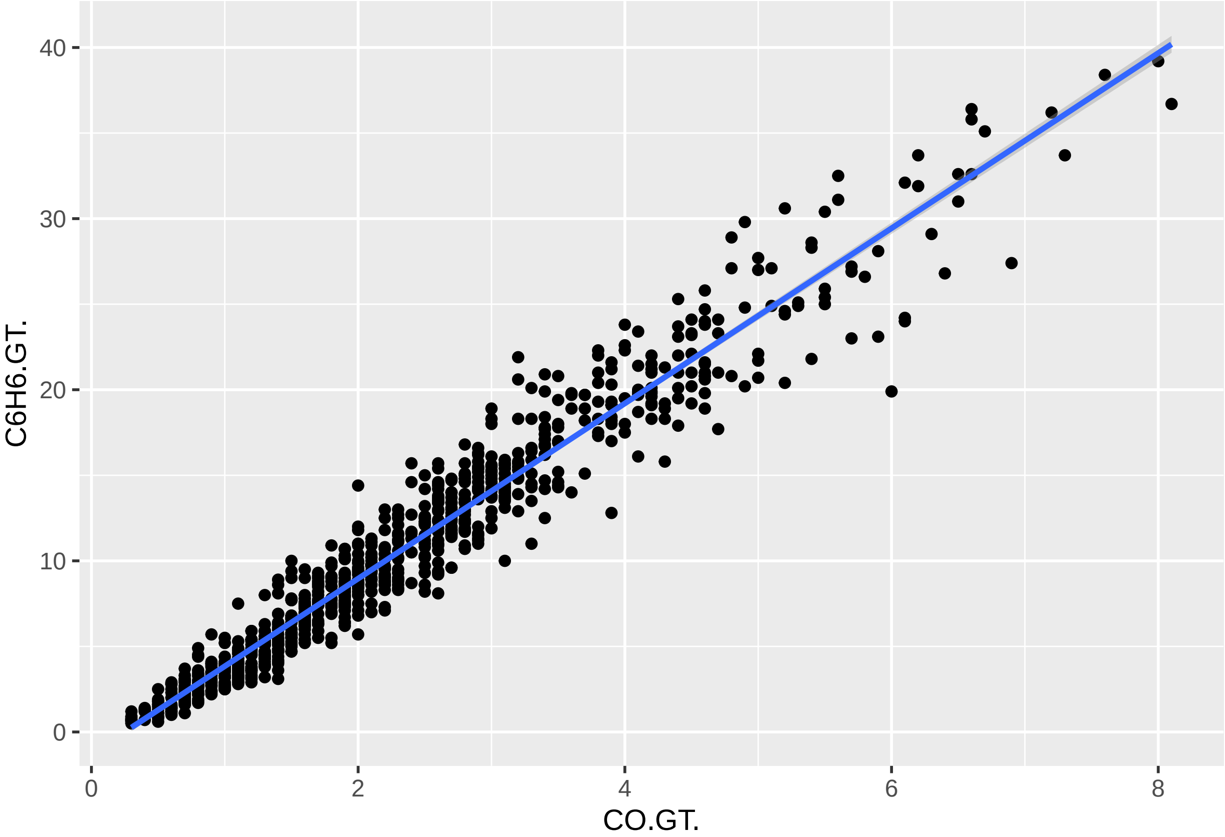
```
##
```

```
## Call:
## lm(formula = airq_fil$C6H6.GT. ~ airq_fil$CO.GT., data = .)
##
## 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 ***
## airq_fil$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
```

Q5. Build simple linear models with each predictor, check assumptions, response C6H6

```
airq_fil %>%
  ggplot(aes(x= CO.GT., y= C6H6.GT.)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
airq_fil  %>%
  lm(data= .,C6H6.GT. ~ C6H6.GT.)%>%
  summary()
```

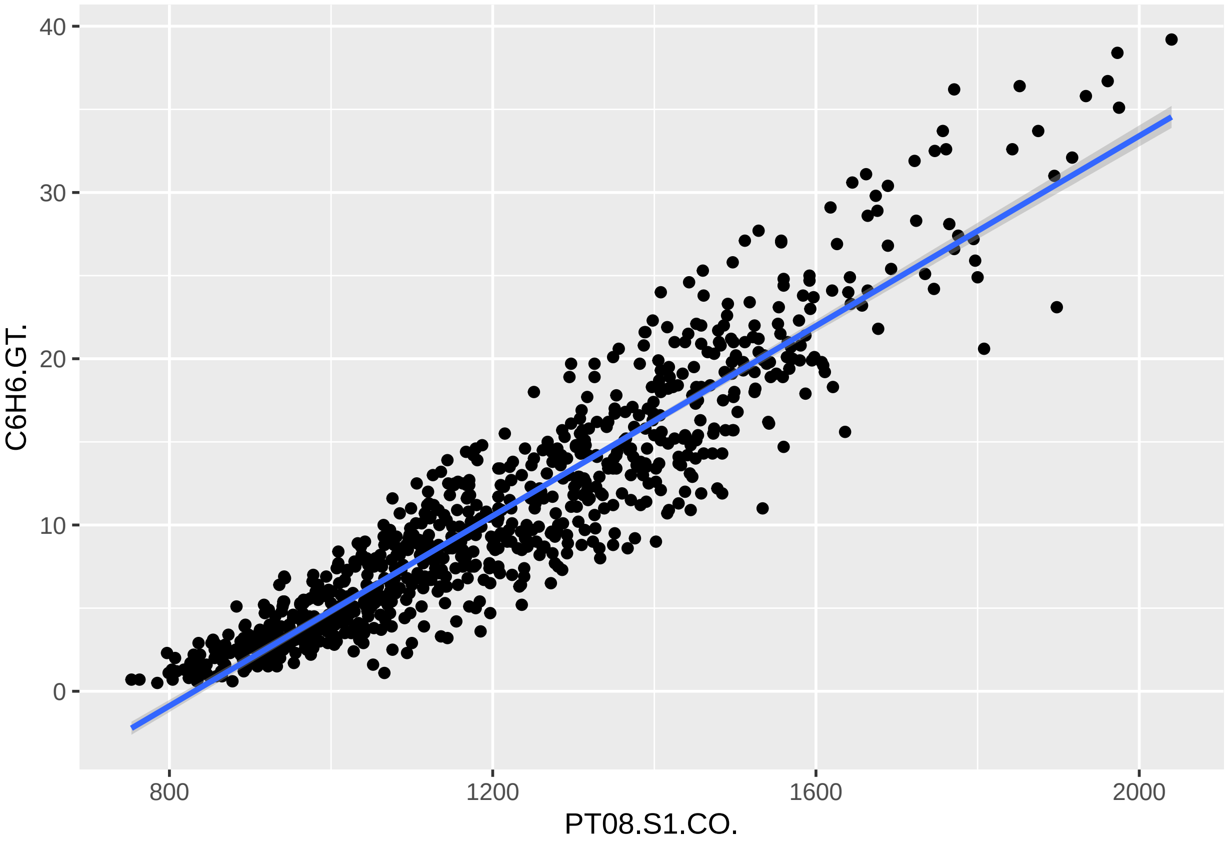
```
## Warning in model.matrix.default(mt, mf, contrasts):
##
```

```
## Warning in model.matrix.default(mt, mf, contrasts):            1
## 'model.matrix':
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ C6H6.GT., data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.271  -5.971  -1.671   4.029  28.429
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.771      0.258   41.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.418 on 826 degrees of freedom
```

```
#PT08.S1.CO
airq_fil %>%
  ggplot(aes(x= PT08.S1.CO., y= C6H6.GT.)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

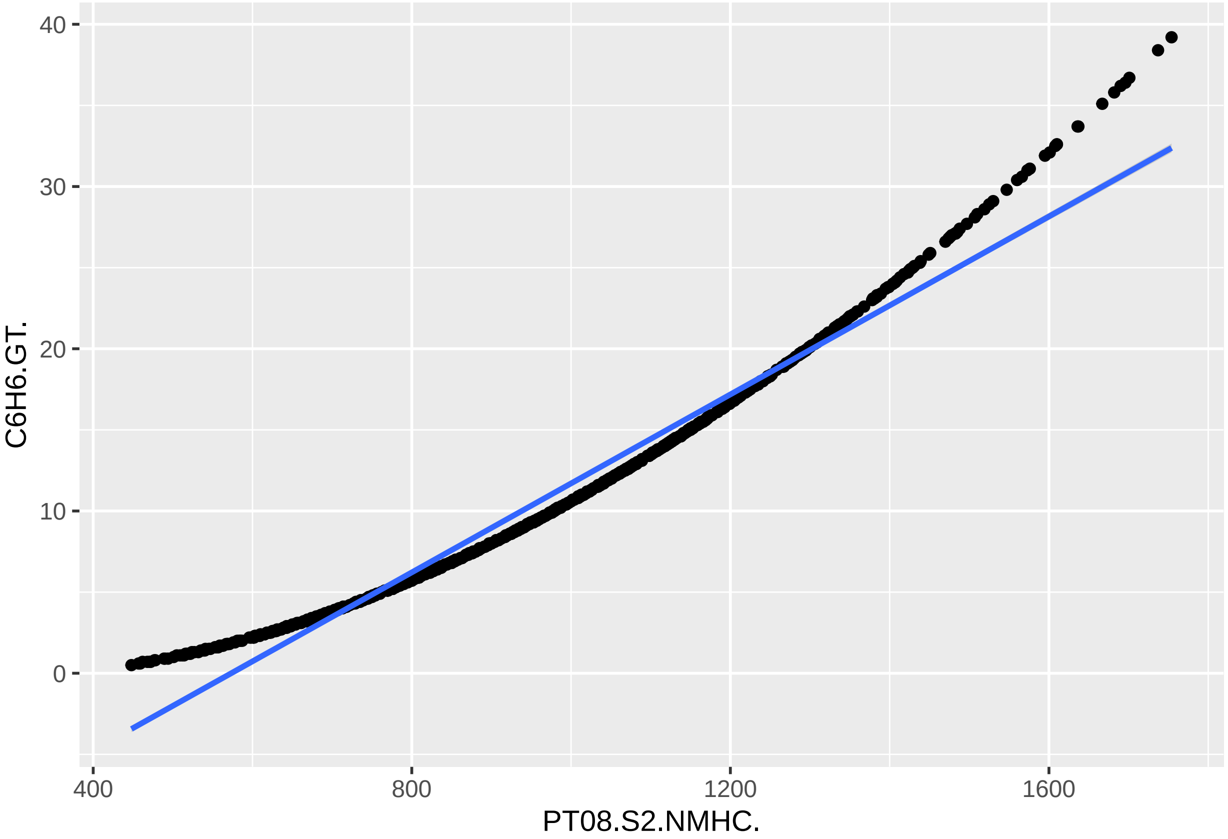


```
airq_fil  %>%
  lm(data= .,C6H6.GT. ~ PT08.S1.CO.)%>%
  summary()
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S1.CO., data = .)
##
## 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
```

```
#PT08.S2.NMHC
airq_fil %>%
  ggplot(aes(x= PT08.S2.NMHC., y= C6H6.GT.)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

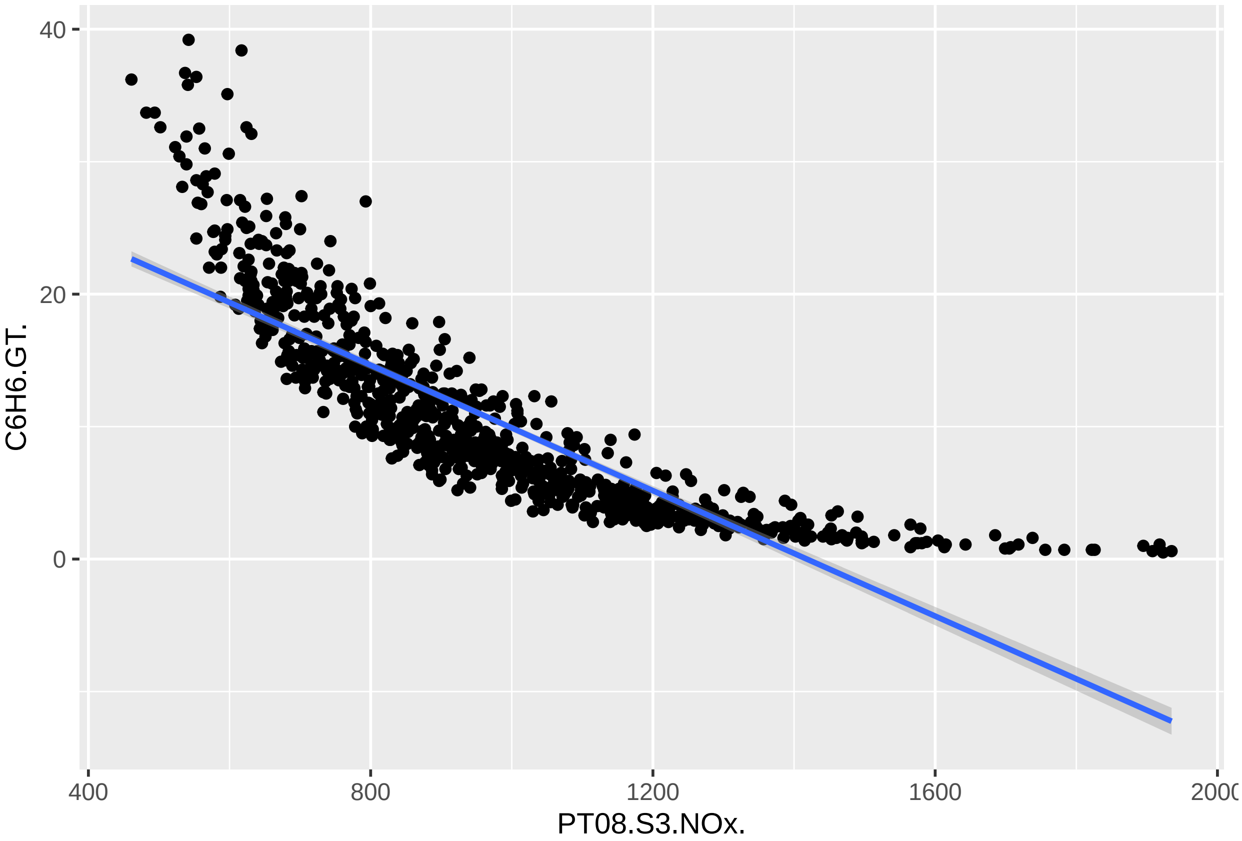


```
airq_fil  %>%
  lm(data= .,C6H6.GT. ~ PT08.S2.NMHC.)%>%
  summary()
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S2.NMHC., data = .)
##
## 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
```

```
#PT08.S3.NOx
airq_fil %>%
  ggplot(aes(x= PT08.S3.NOx., y= C6H6.GT.)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

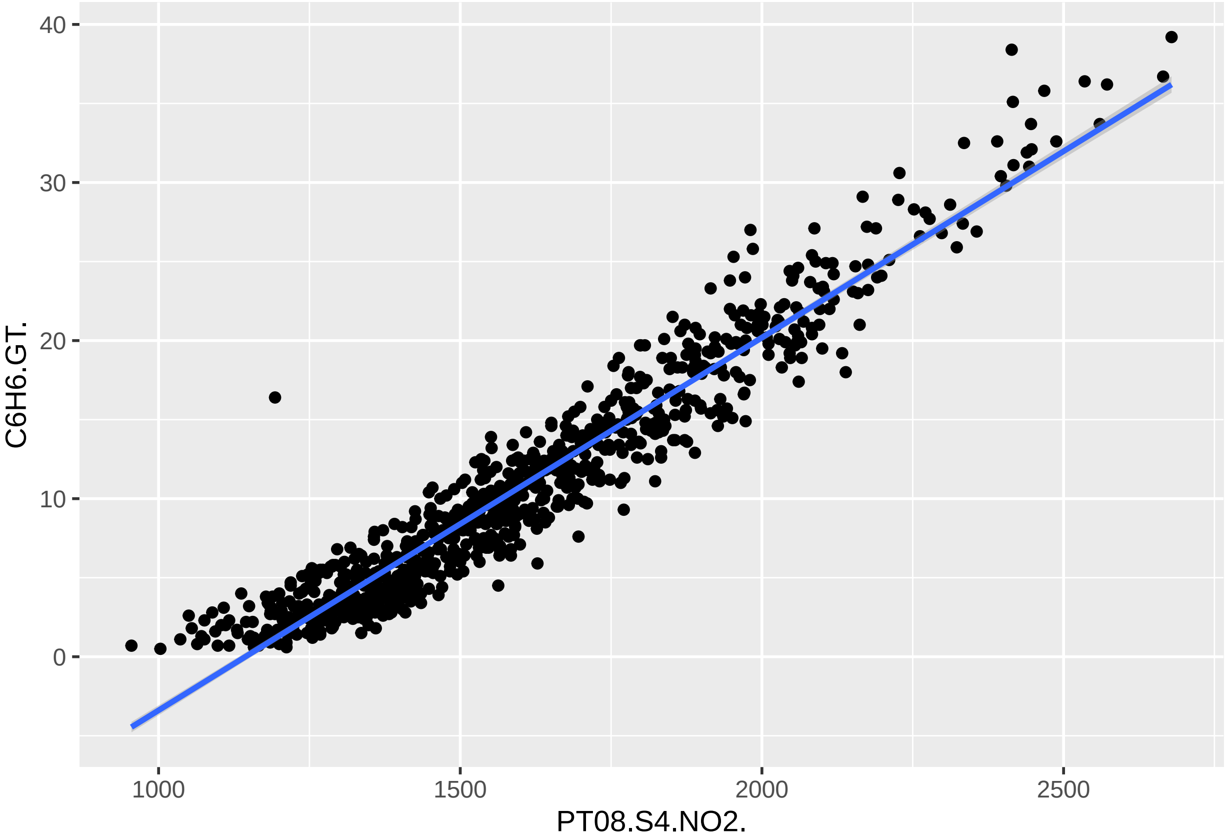


```
airq_fil  %>%
  lm(data= .,C6H6.GT. ~ PT08.S3.NOx.)%>%
  summary()
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S3.NOx., data = .)
##
## 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
```

```
#PT08.S4.NO2
airq_fil %>%
  ggplot(aes(x= PT08.S4.NO2., y= C6H6.GT.)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

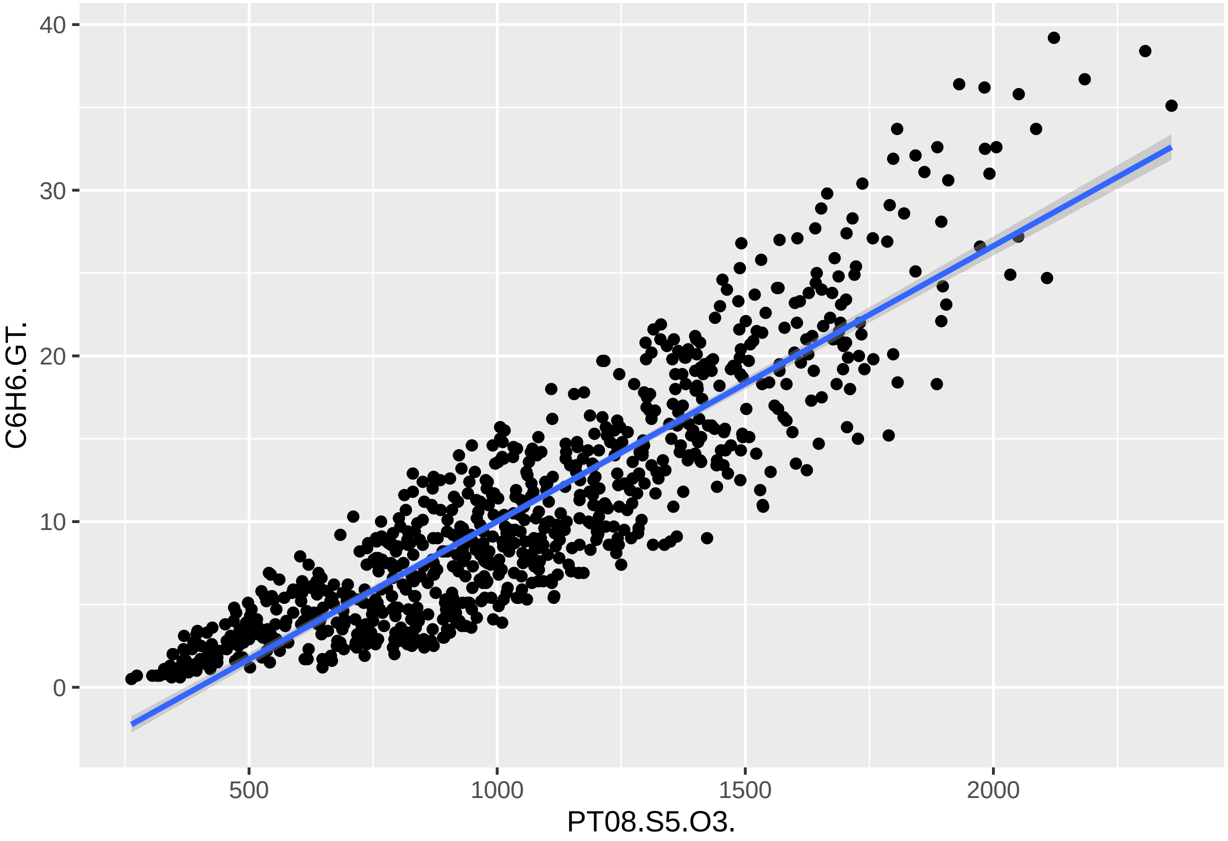


```
airq_fil  %>%
  lm(data= .,C6H6.GT. ~ PT08.S4.NO2.)%>%
  summary()
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S4.NO2., data = .)
##
## 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
```

```
#PT08.S5.O3
airq_fil %>%
  ggplot(aes(x= PT08.S5.O3., y= C6H6.GT.)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



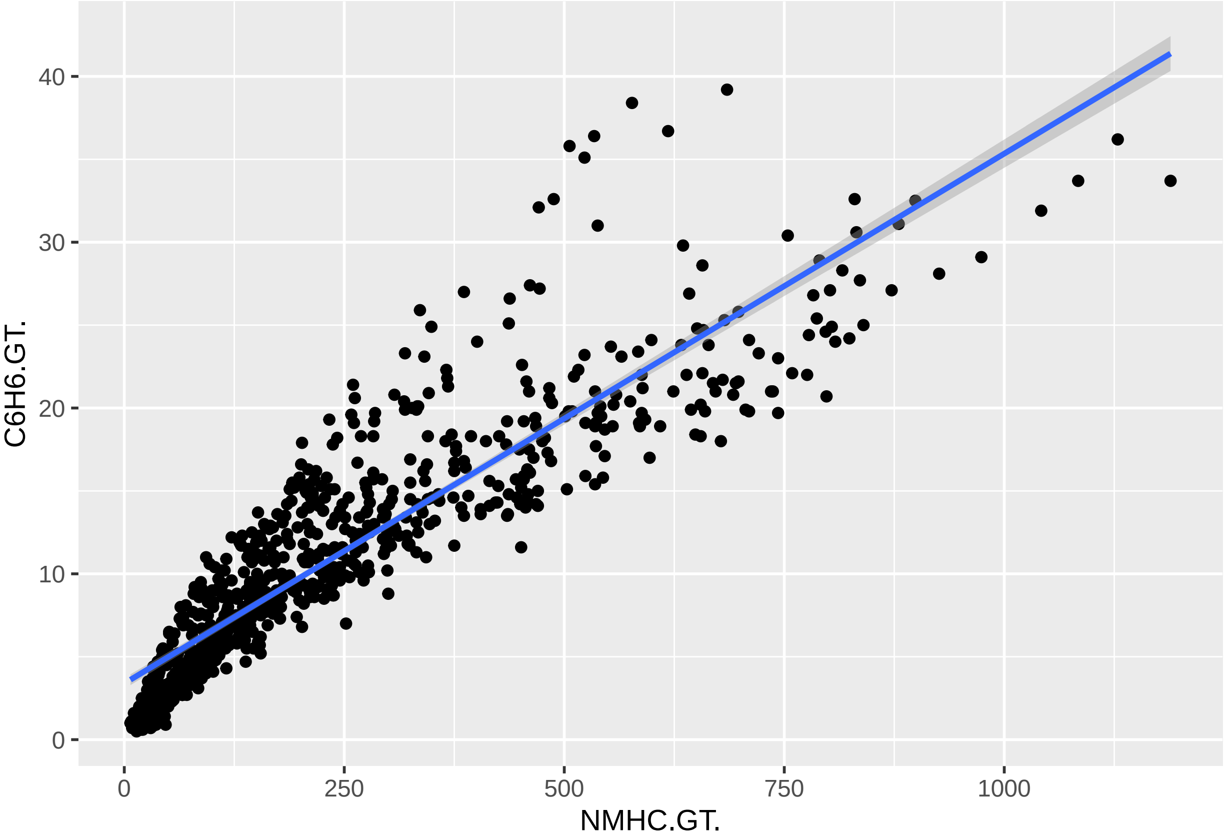
```
airq_fil  %>%
  lm(data= .,C6H6.GT. ~ PT08.S5.O3.)%>%
  summary()
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S5.O3., data = .)
##
## 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
```

```
# NMHC.GT
airq_fil %>%
  ggplot(aes(x= NMHC.GT., y= C6H6.GT.)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

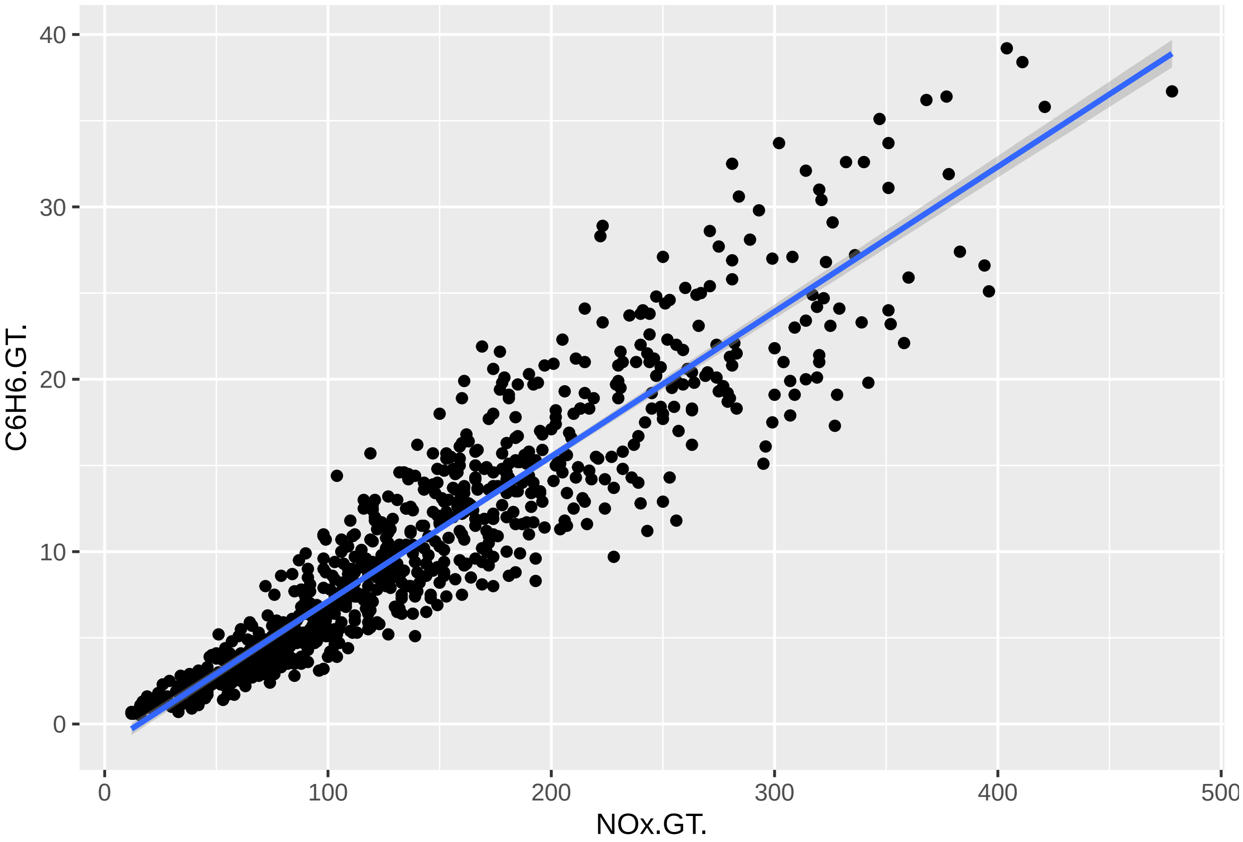


```
airq_fil  %>%
  lm(data= .,C6H6.GT. ~ NMHC.GT.)%>%
  summary()
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ NMHC.GT., data = .)
##
## 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
```

```
#NOx.GT
airq_fil %>%
  ggplot(aes(x= NOx.GT., y= C6H6.GT.)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

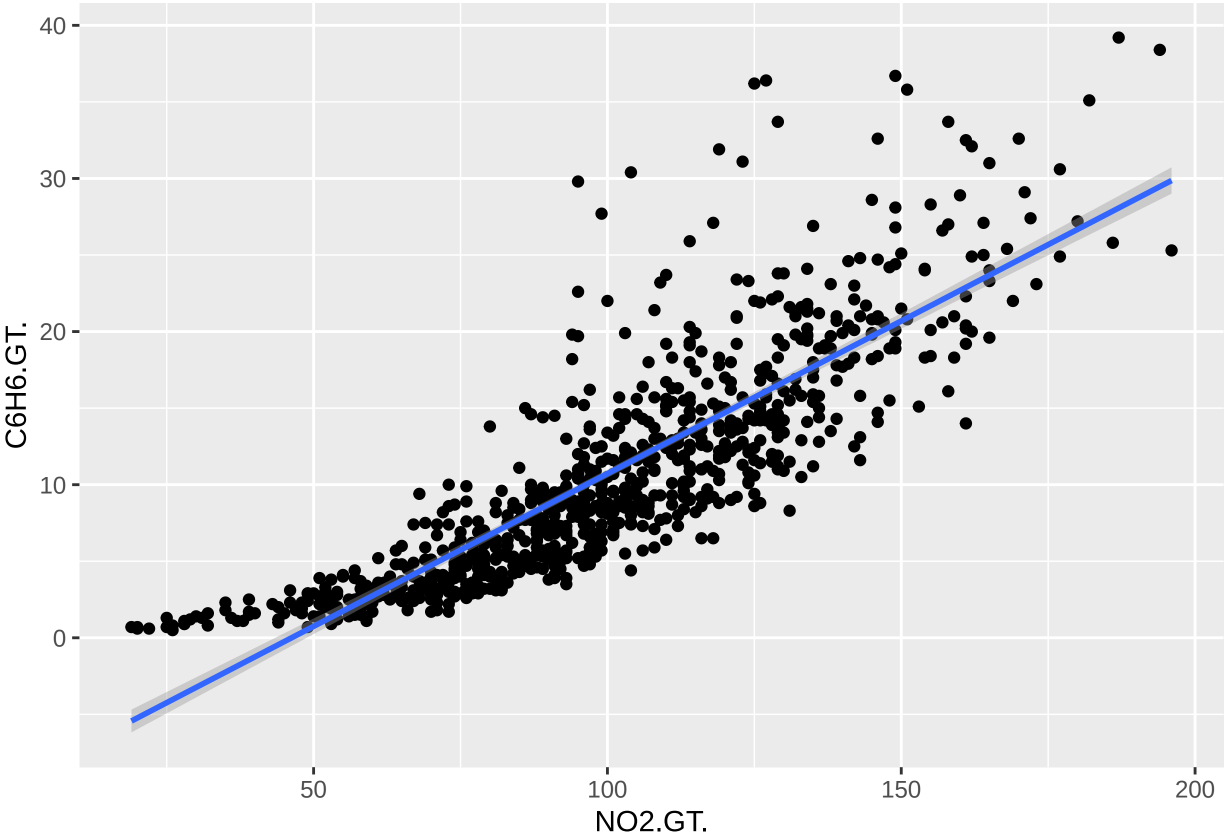


```
airq_fil  %>%
  lm(data= .,C6H6.GT. ~ NOx.GT.)%>%
  summary()
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ NOx.GT., data = .)
##
## 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
```

```
#NO2.GT
airq_fil %>%
  ggplot(aes(x= NO2.GT., y= C6H6.GT.)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



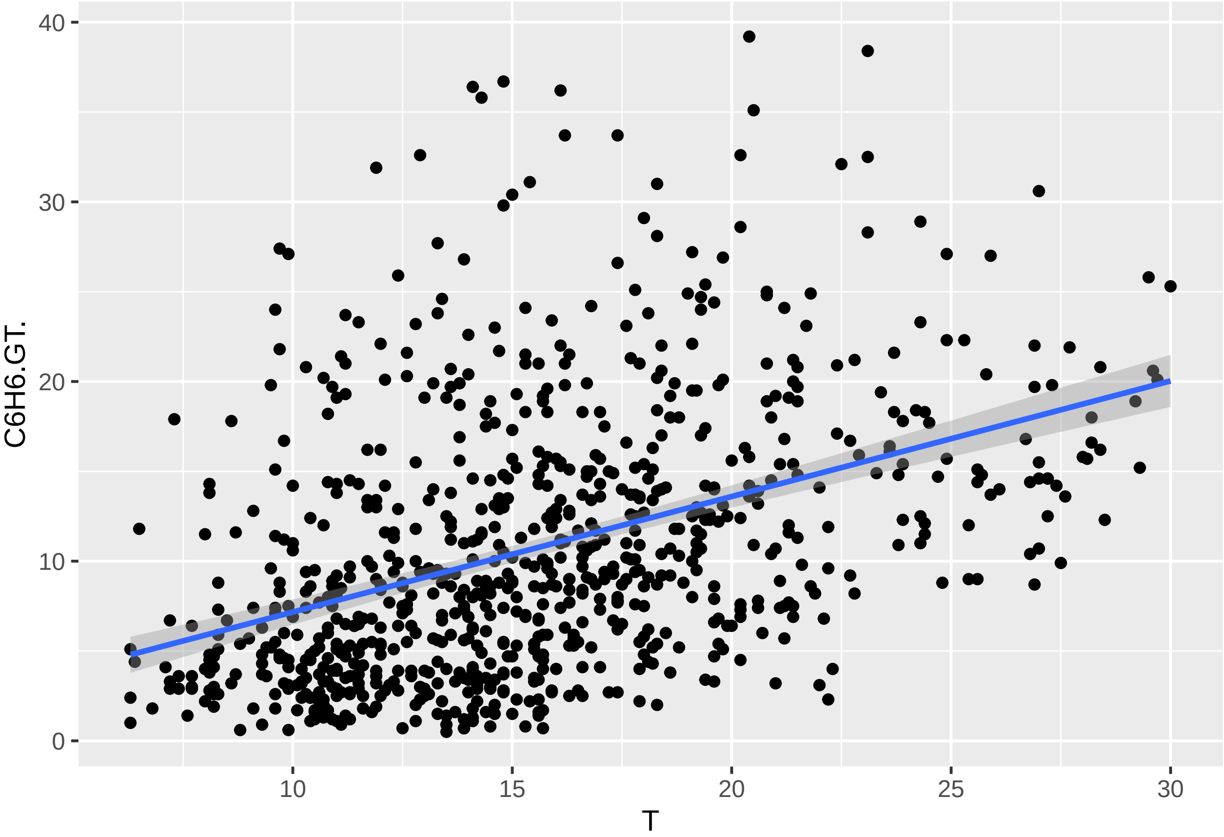
```
airq_fil  %>%
  lm(data= .,C6H6.GT. ~ NO2.GT.)%>%
  summary()
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ NO2.GT., data = .)
##
## 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
```

```
#T
airq_fil %>%
  ggplot(aes(x= T, y= C6H6.GT.)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

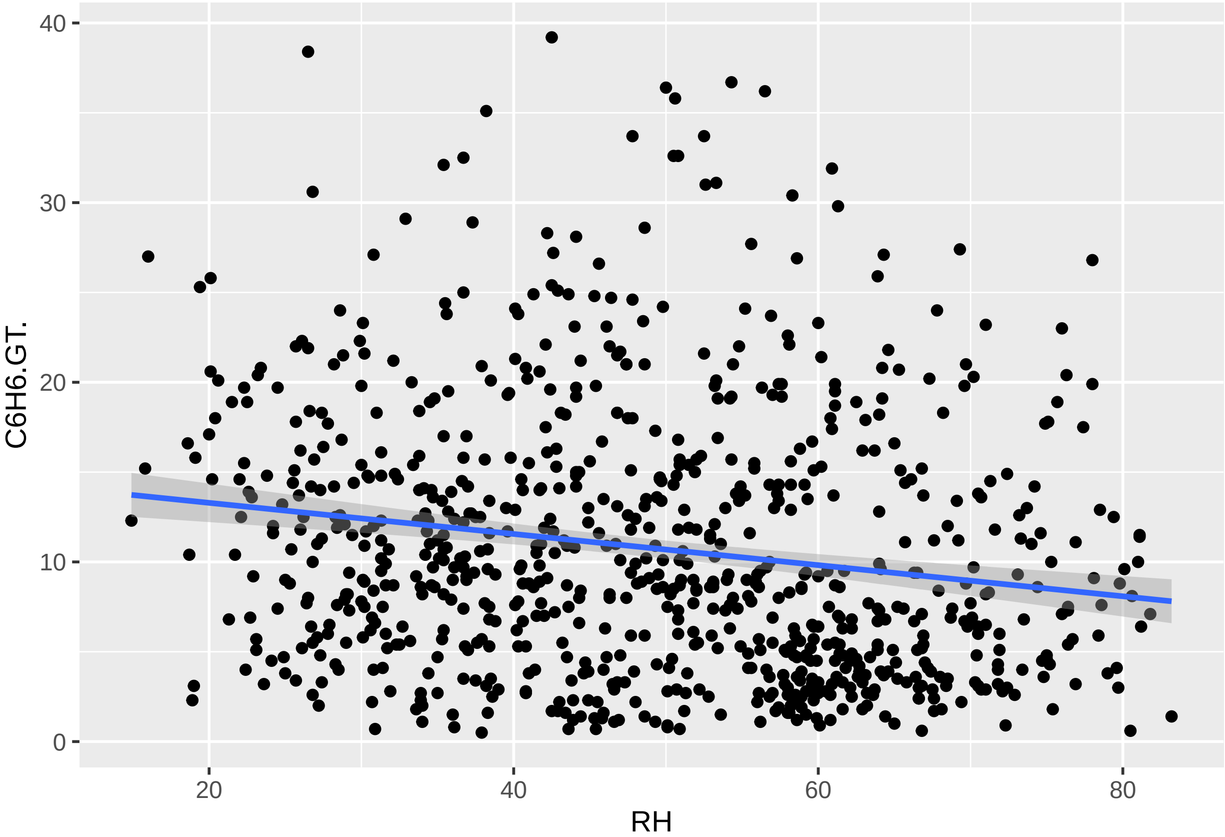


```
airq_fil  %>%
  lm(data= .,C6H6.GT. ~ T)%>%
  summary()
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ T, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.716  -4.813  -1.526   3.075  26.595
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.73568    0.79384   0.927    0.354
## T            0.64324    0.04861  13.232   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.742 on 825 degrees of freedom
## Multiple R-squared:  0.1751, Adjusted R-squared:  0.1741
## F-statistic: 175.1 on 1 and 825 DF,  p-value: < 2.2e-16
```

```
#RH
airq_fil %>%
  ggplot(aes(x= RH, y= C6H6.GT.)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



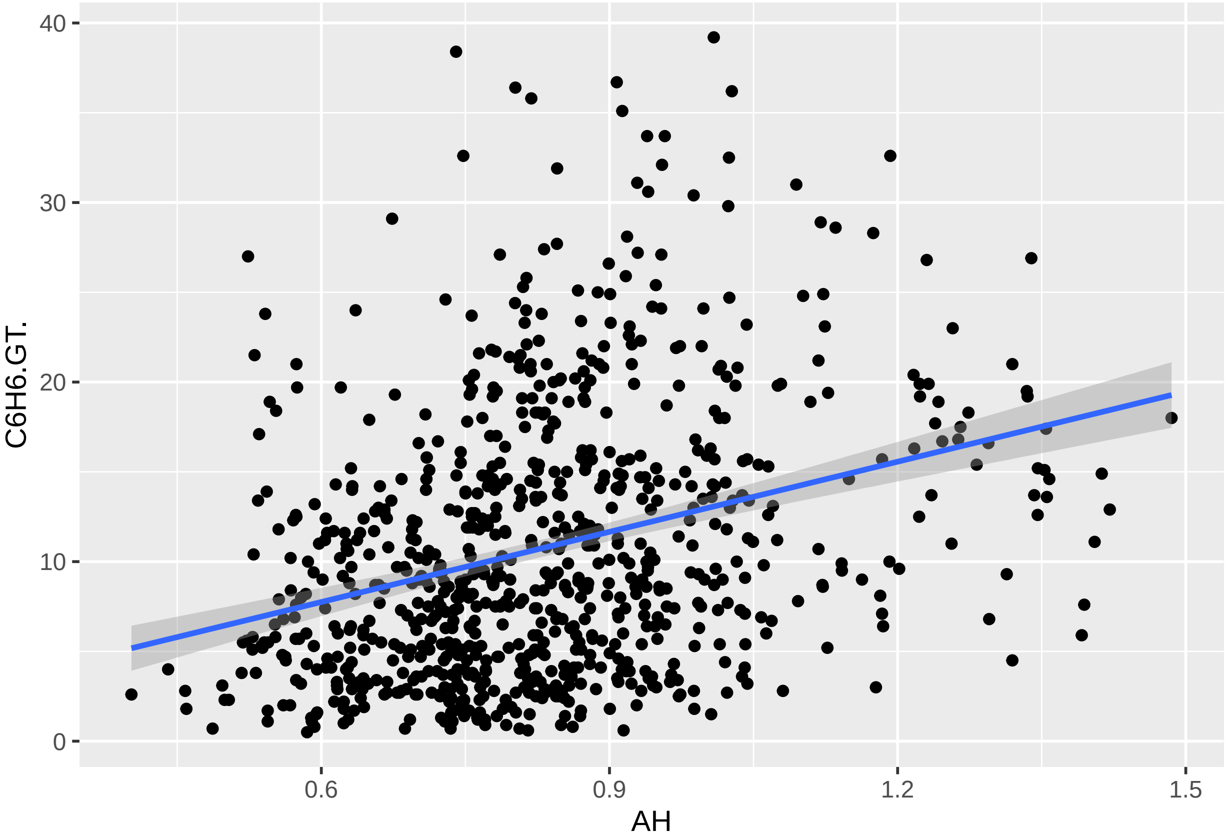
```
airq_fil  %>%
  lm(data= .,C6H6.GT. ~ RH)%>%
  summary()
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ RH, data = .)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.645  -5.566  -1.584   3.962  27.861
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.02325    0.85505  17.570  < 2e-16 ***
## RH          -0.08669    0.01665  -5.208 2.41e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.304 on 825 degrees of freedom
## Multiple R-squared:  0.03183,    Adjusted R-squared:  0.03066
## F-statistic: 27.12 on 1 and 825 DF,  p-value: 2.412e-07
```

```
airq_fil %>%
  ggplot(aes(x= AH, y= C6H6.GT.)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
airq_fil  %>%
  lm(data= .,C6H6.GT. ~ AH)%>%
  summary()
```

```
##
## Call:
```

```
## lm(formula = C6H6.GT. ~ AH, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.621  -5.372  -1.530   3.850  28.821
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.06339    1.16889  -0.054    0.957
## AH          13.02454    1.37393   9.480   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.049 on 825 degrees of freedom
## Multiple R-squared:  0.09823,    Adjusted R-squared:  0.09714
## F-statistic: 89.87 on 1 and 825 DF,  p-value: < 2.2e-16
```

Q6. For 2-3 of the models create train-test sets, plot the model for the test set color real and predicted points differently; R^2 and p-value to title

```
#PT08.S5.O3
set.seed(42)
#Separated data in 75:25%
sample <- sample.int(n = nrow(airq_fil), size = floor(.75*nrow(airq_fil)))
training <- airq_fil[sample, ]
test <- airq_fil[-sample, ]
train_PT <- training[,c("C6H6.GT.", "PT08.S5.O3." )]
test_PT <- test[,c("C6H6.GT.", "PT08.S5.O3." )]
model_PT <- lm(data = train_PT , C6H6.GT. ~ PT08.S5.O3.)
#summary
summary <- summary(model_PT)
print(summary)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S5.O3., data = train_PT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8775 -2.4821  0.1225  2.1835  8.9541
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.4573199  0.3595021  -17.96   <2e-16 ***
## PT08.S5.O3.  0.0163984  0.0003198   51.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.188 on 618 degrees of freedom
## Multiple R-squared:  0.8097, Adjusted R-squared:  0.8094
## F-statistic:  2630 on 1 and 618 DF,  p-value: < 2.2e-16
```

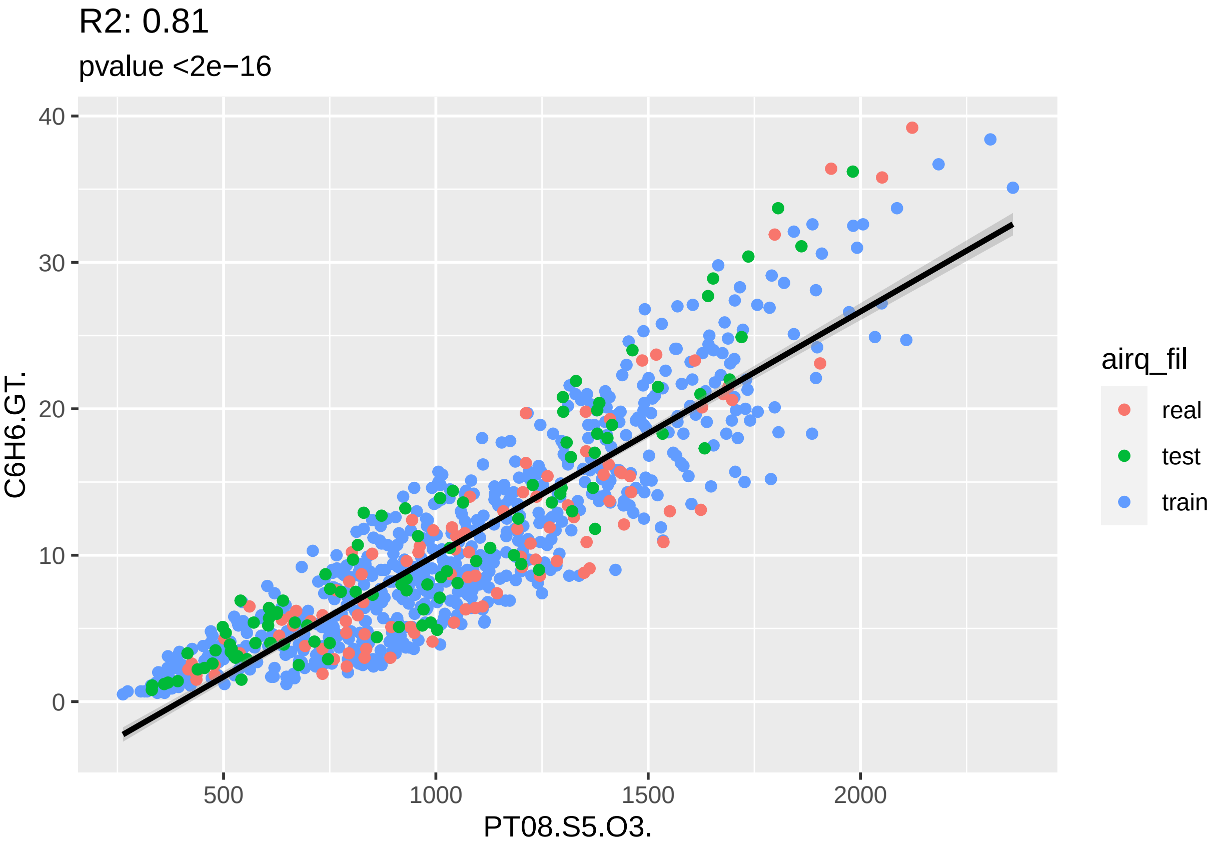
```
#Predicition
pred_PT <- predict(model_PT, newdata = test_PT)
test$PT08.S5.O3._pred <- pred_PT
#make combined dataset
```

```
train_PT$airq_fil <- "train"
test_PT$airq_fil <- "test"
test_PT[1:(nrow(test_PT)/2),3] <- "real"

all_PT <- rbind(train_PT, test_PT)
```

```
ggplot(data = all_PT, aes(x = PT08.S5.O3.,
y = C6H6.GT.,
color=airq_fil)) +
  geom_point() +
geom_smooth(method = "lm", color = "black") +
ggtitle(paste("R2", round(summary$r.squared, 3), sep = ": "),
paste("pvalue <2e-16" ))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
# NMHC.GT
set.seed(42)
#Separated data in 75:25%
sample <- sample.int(n = nrow(airq_fil), size = floor(.75*nrow(airq_fil)))
training <- airq_fil[sample, ]
test <- airq_fil[-sample, ]
train_MH <- training[,c("C6H6.GT.", "NMHC.GT." )]
test_MH <- test[,c("C6H6.GT.", "NMHC.GT." )]
model_MH <- lm(data = train_MH , C6H6.GT. ~ NMHC.GT.)
#summary
summary <- summary(model_MH)
print(summary)
```

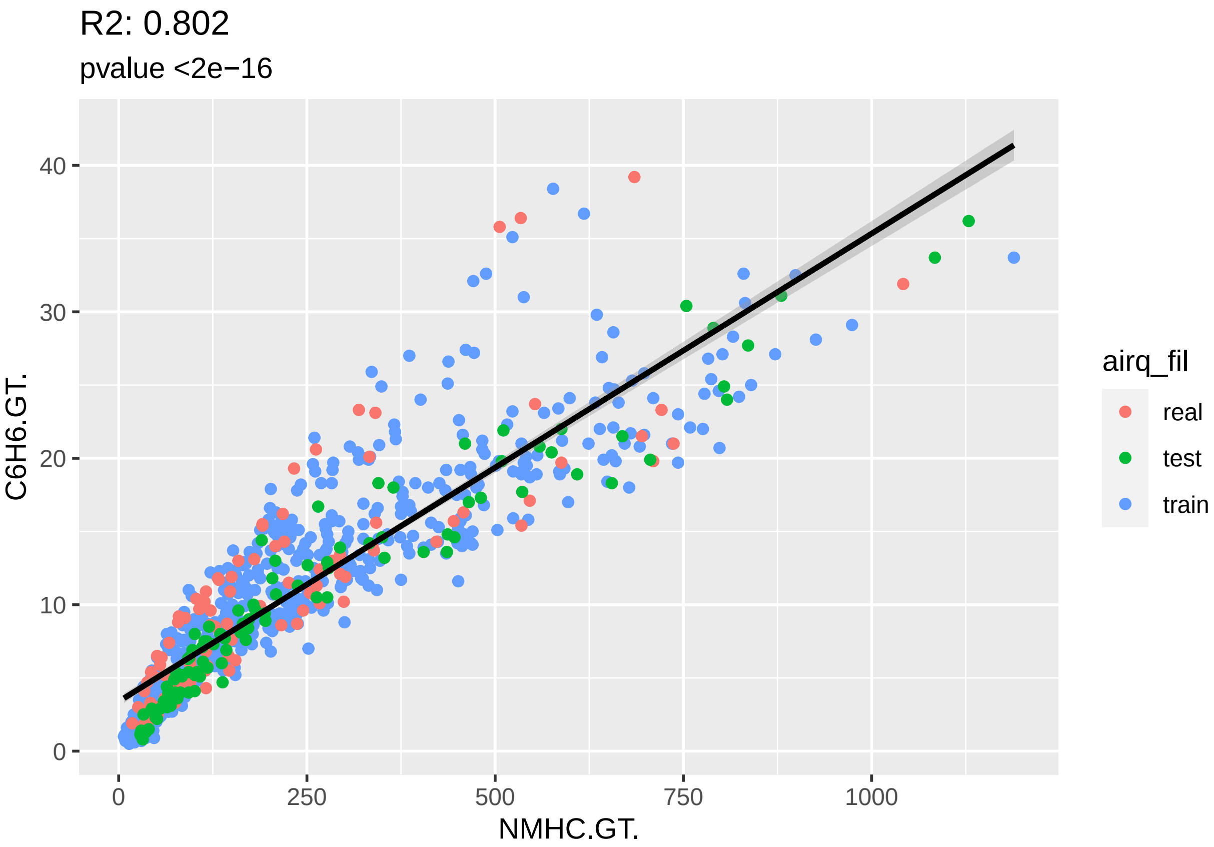
```
##
## Call:
## lm(formula = C6H6.GT. ~ NMHC.GT., data = train_MH)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4400 -2.0843 -0.6558  1.4617 16.3969
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.3695602  0.1971641   17.09   <2e-16 ***
## NMHC.GT.    0.0322938  0.0006449   50.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.25 on 618 degrees of freedom
## Multiple R-squared:  0.8023, Adjusted R-squared:  0.802
## F-statistic:  2508 on 1 and 618 DF,  p-value: < 2.2e-16
```

```
#Predicition
pred_MH <- predict(model_MH, newdata = test_MH)
test$NMHC.GT._pred <- pred_MH
#make combined dataset
train_MH$airq_fil <- "train"
test_MH$airq_fil <- "test"
test_MH[1:(nrow(test_MH)/2),3] <- "real"

all_PT <- rbind(train_MH, test_MH)
#trained and test should be similar shape
```

```
ggplot(data = all_PT, aes(x = NMHC.GT.,
y = C6H6.GT.,
color=airq_fil)) +
  geom_point() +
geom_smooth(method = "lm", color = "black") +
ggtitle(paste("R2", round(summary$r.squared, 3), sep = ": "),
paste("pvalue <2e-16" ))
```

```
## `geom_smooth()` using formula 'y ~ x'
```


R2: 0.802
pvalue <2e−16
C6H6.GT.
0
10
20
30
40
0
250
500
750
1000
NMHC.GT.
airq_fil
real
test
train