Task 1

Air Quality data set

```
airq_data <- read.table('/home/marina/Загрузки/AirQualityUCI.csv', sep = ';', header = TRUE, dec = ",")
```

• Deleting NA columns and rows. Averaged concentration CO, Averaged Benzene concentration, Temperature, Humidity are factor columns in data, so it will be better to convert them into numeric. Actually, I don't need date and time for futher analysis, so I will slice the data.

```
airq_data <- airq_data %>% select_if(~sum(!is.na(.)) > 0)  %>% drop_na()
airq_data[c('CO.GT.', 'C6H6.GT.', 'T', 'RH', 'AH')] <-  sapply(airq_data[c('CO.GT.', 'C6H6.GT.', 'T', 'RH', 'AH')], as.numeric)
airq_data <- airq_data[, c(3:15)]
```

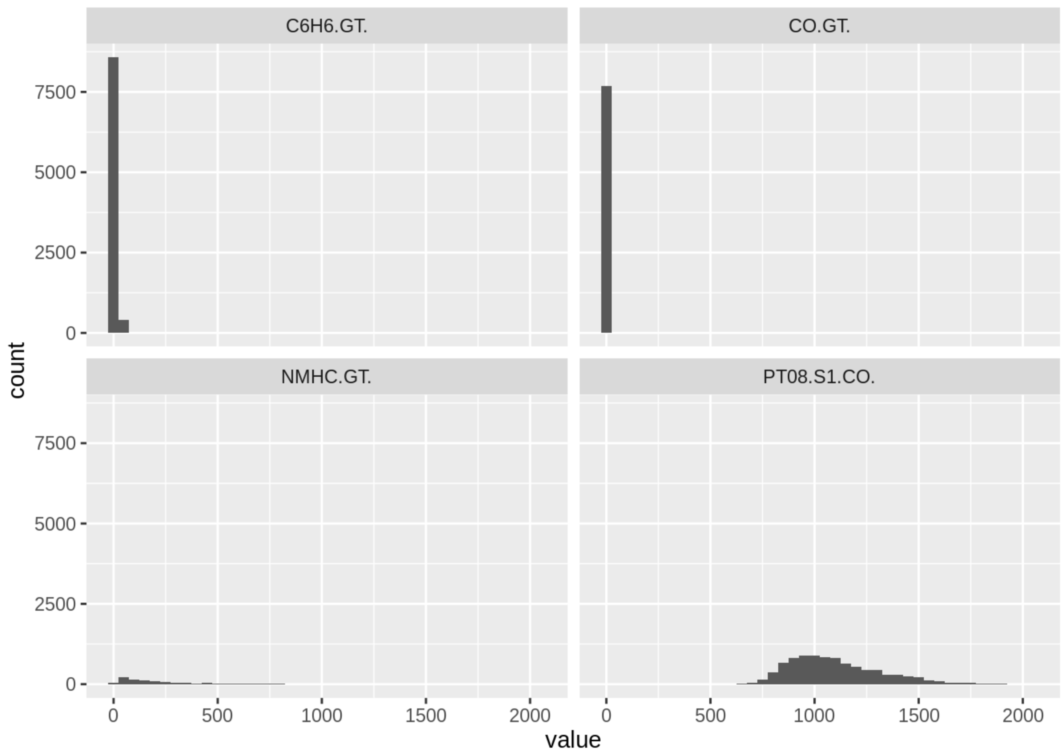
• Replacing -200 to NA.

```
airq_data <- sapply(airq_data, function(x){ifelse (x == -200, NA, x)})
airq_data <- as.data.frame(airq_data)
```

• Exploring the variables (I make it in a 3 parts to better visualizing)

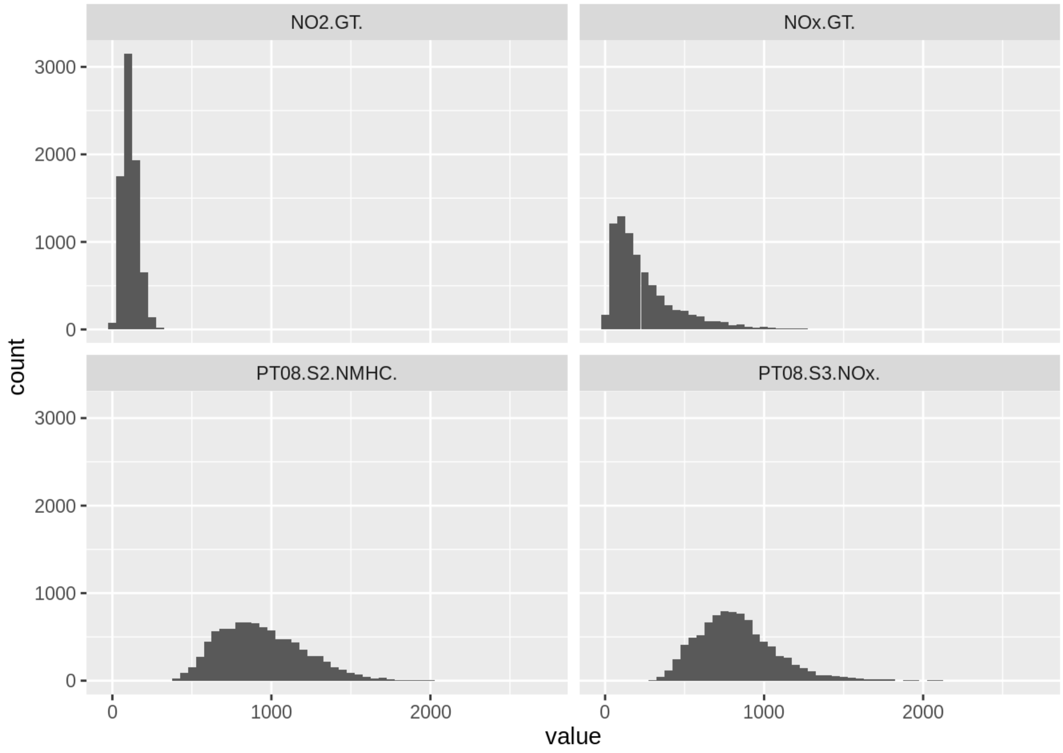
```
ggplot(gather(airq_data[, c(1:4)], cols, value), aes(x = value)) +
  geom_histogram(binwidth = 50) + facet_wrap(.~cols)
```

```
## Warning: Removed 10858 rows containing non-finite values (stat_bin).
```



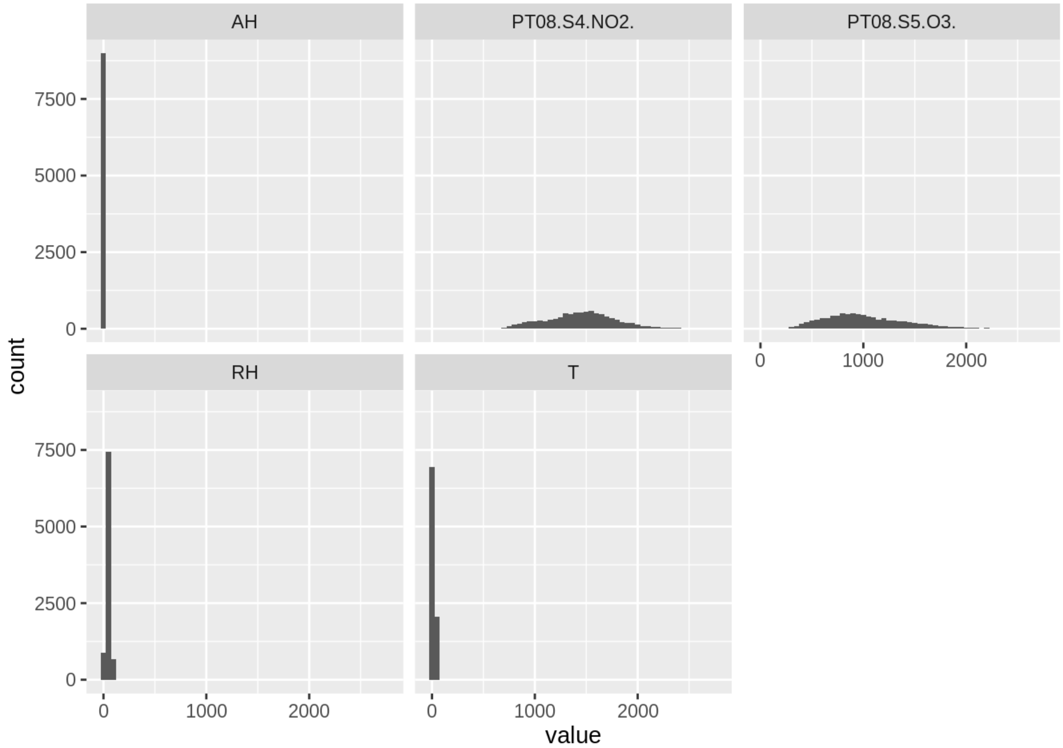

```
ggplot(gather(airq_data[, c(5:8)], cols, value), aes(x = value)) +
  geom_histogram(binwidth = 50) + facet_wrap(.~cols)
```

```
## Warning: Removed 4013 rows containing non-finite values (stat_bin).
```




```
ggplot(gather(airq_data[, c(9:13)], cols, value), aes(x = value)) +
  geom_histogram(binwidth = 50) + facet_wrap(.~cols)
```

```
## Warning: Removed 1830 rows containing non-finite values (stat_bin).
```



Ok, so it looks like a lot of variables have outliers and all of them in a different scale range So, as for me, it will be better to log-transform the data

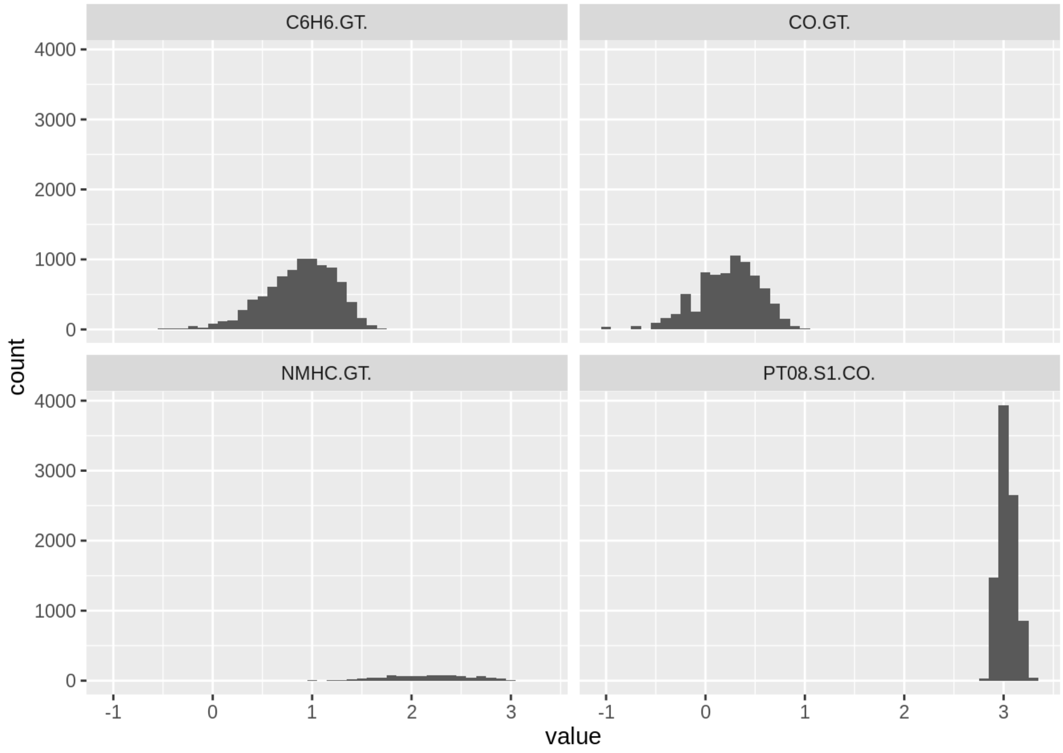
- Log-transformation

```
airq_data <- log10(airq_data)
```

So, let's look now as an example for the first 4 columns:

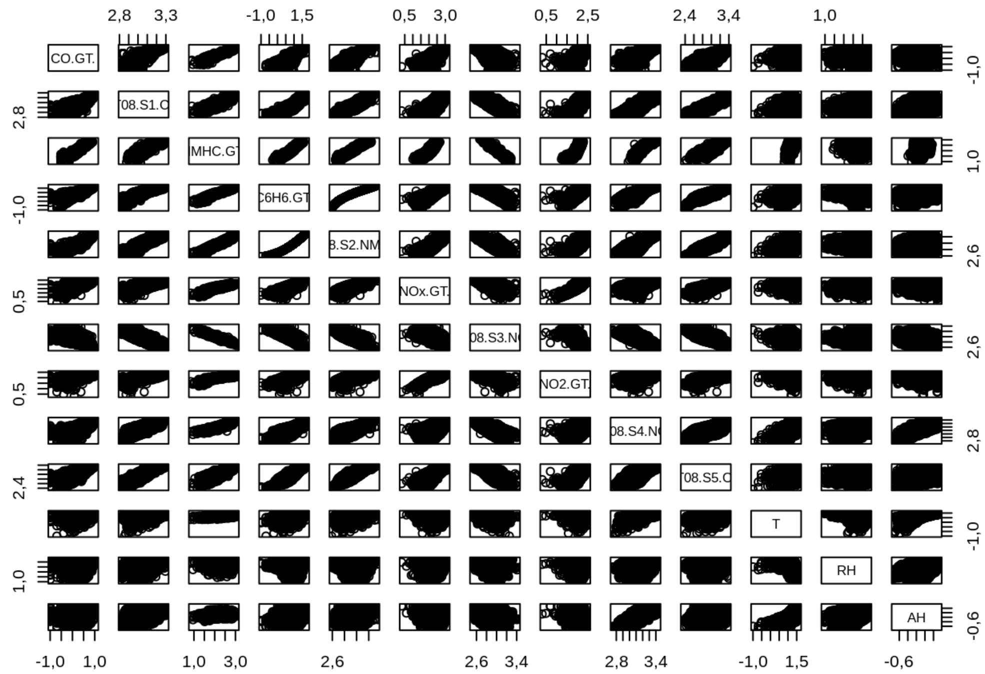
```
ggplot(gather(airq_data[, c(1:4)], cols, value), aes(x = value)) +
  geom_histogram(binwidth = 0.1) + facet_wrap(.~cols)
```

```
## Warning: Removed 10858 rows containing non-finite values (stat_bin).
```

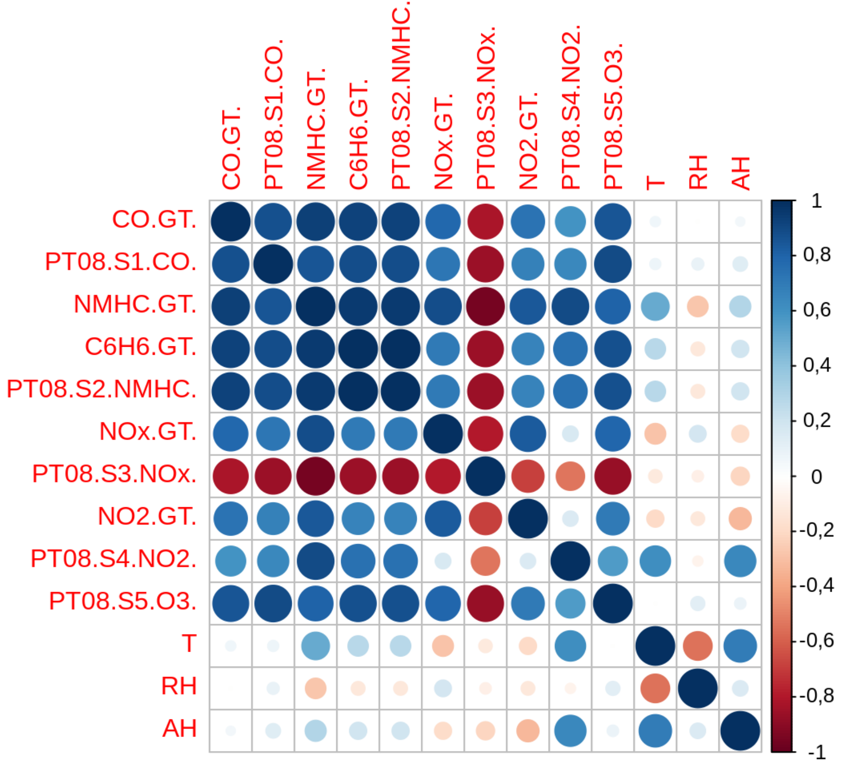


• Relationships between all variables

```
pairs(airq_data)
```

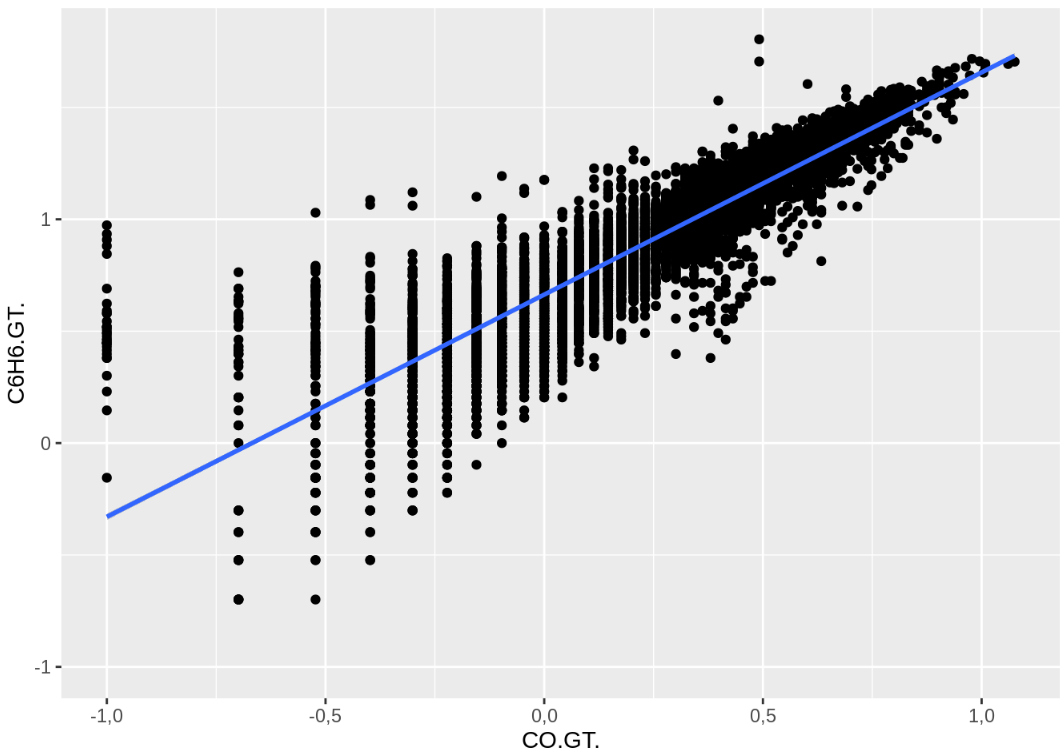



```
cor_data <- cor(airq_data, method = 'spearman', use = 'pairwise.complete.obs')
corrplot(cor_data, method = 'circle')
```

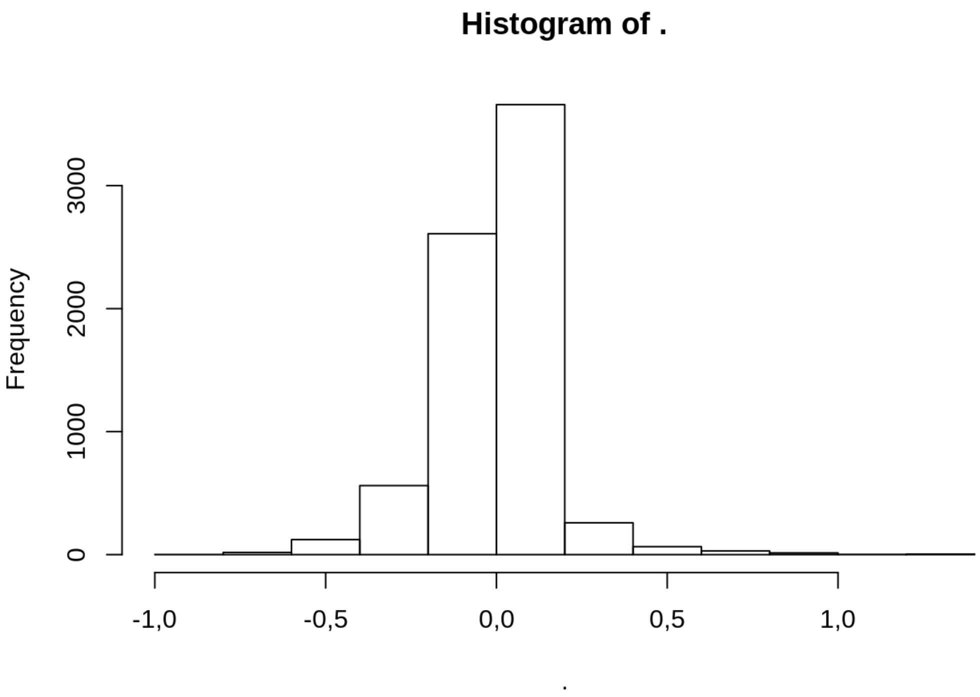


Example of C6.H6.GT.:

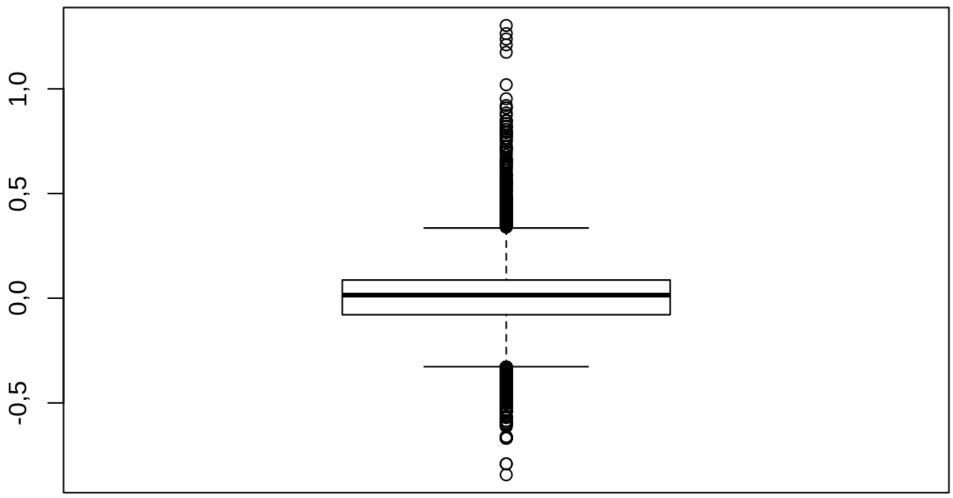
```
ggplot(airq_data, aes(x = CO.GT., y = C6H6.GT. )) +
  geom_point() +
  geom_smooth(method = 'lm')
```



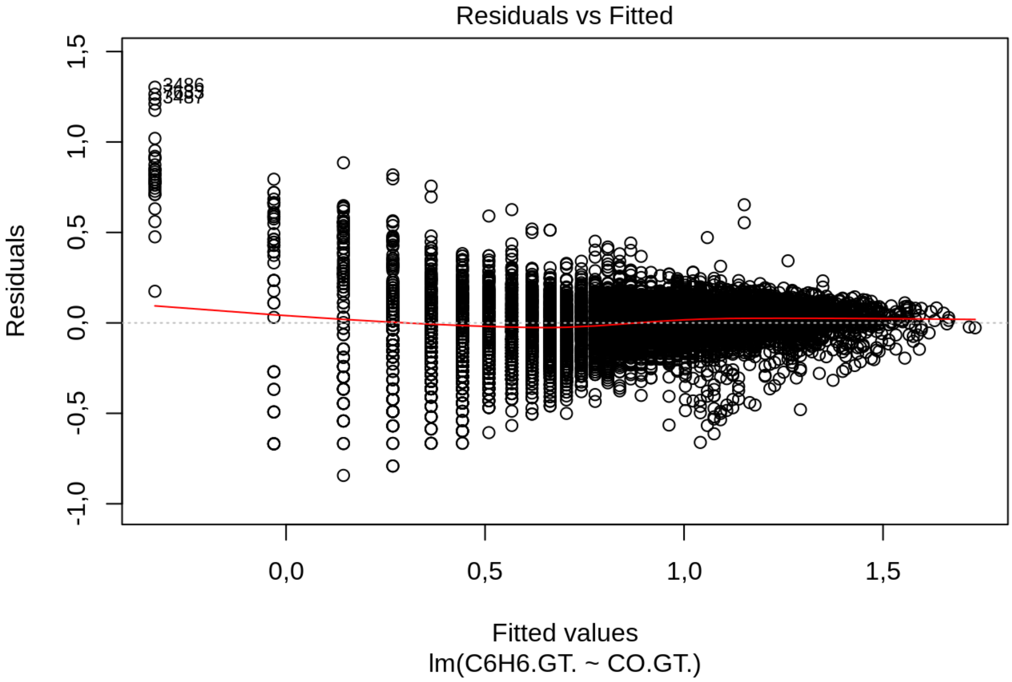

```
model_CO.GT. <- airq_data  %>%
   lm(data = .,  C6H6.GT. ~ CO.GT.)
residuals(model_CO.GT.) %>% hist()
```

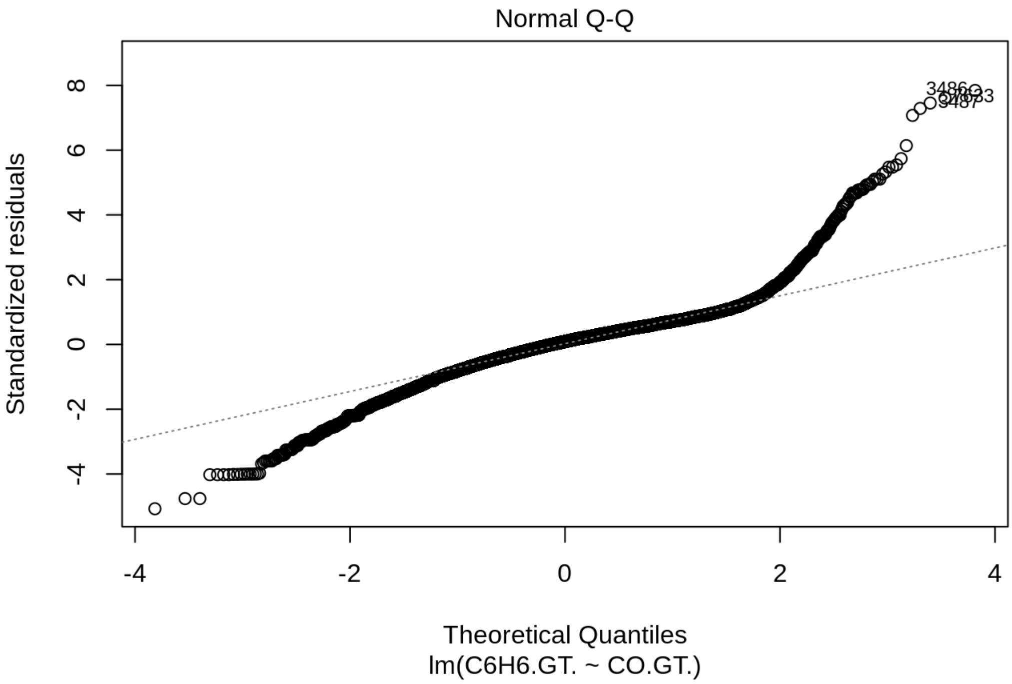



```
residuals(model_CO.GT.) %>% boxplot()
```




```
plot(model_CO.GT., which = c(1,2))
```






```
summary(model_CO.GT.)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ CO.GT., data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0,84319 -0,07876  0,01511  0,08693  1,30265
##
## Coefficients:
```

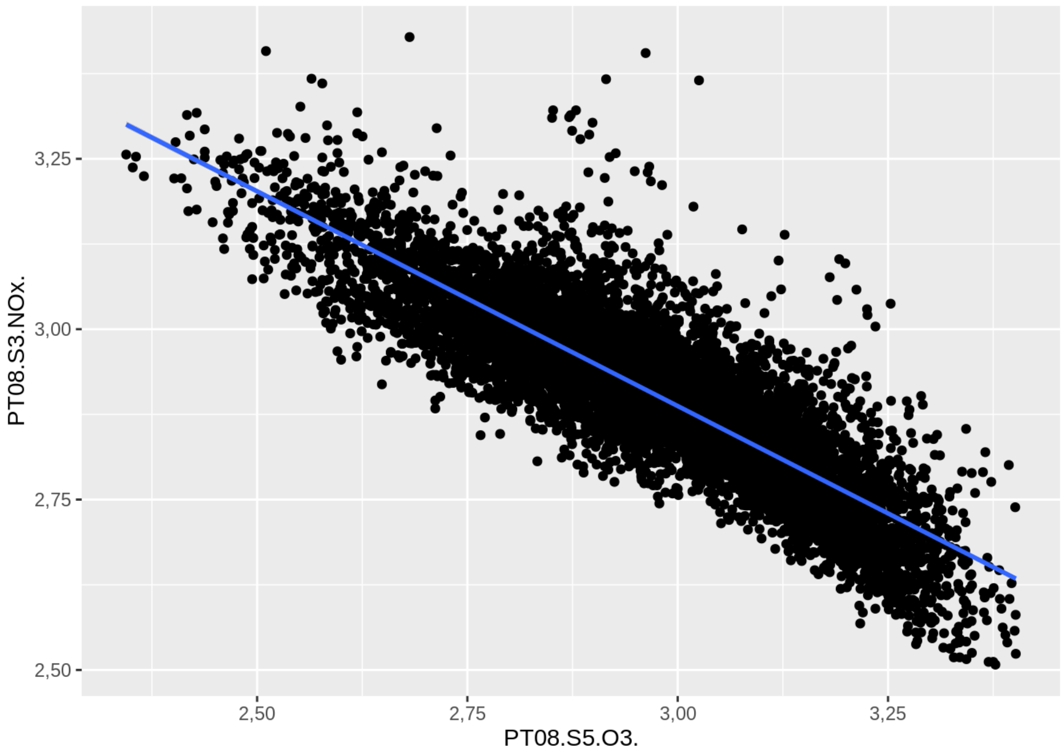
```
## Warning in printCoefmat(coefs, digits = digits, signif.stars = signif.stars, : в
## результате преобразования созданы NA
```

```
##            Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0,663401   0,002378   279,0   <2e-16 ***
## CO.GT.      0,992928   0,006098   162,8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## Residual standard error: 0,1662 on 7342 degrees of freedom
##   (2013 observations deleted due to missingness)
## Multiple R-squared:  0,7831, Adjusted R-squared:  0,7831
## F-statistic: 2,651e+04 on 1 and 7342 DF,  p-value: < 2,2e-16
```

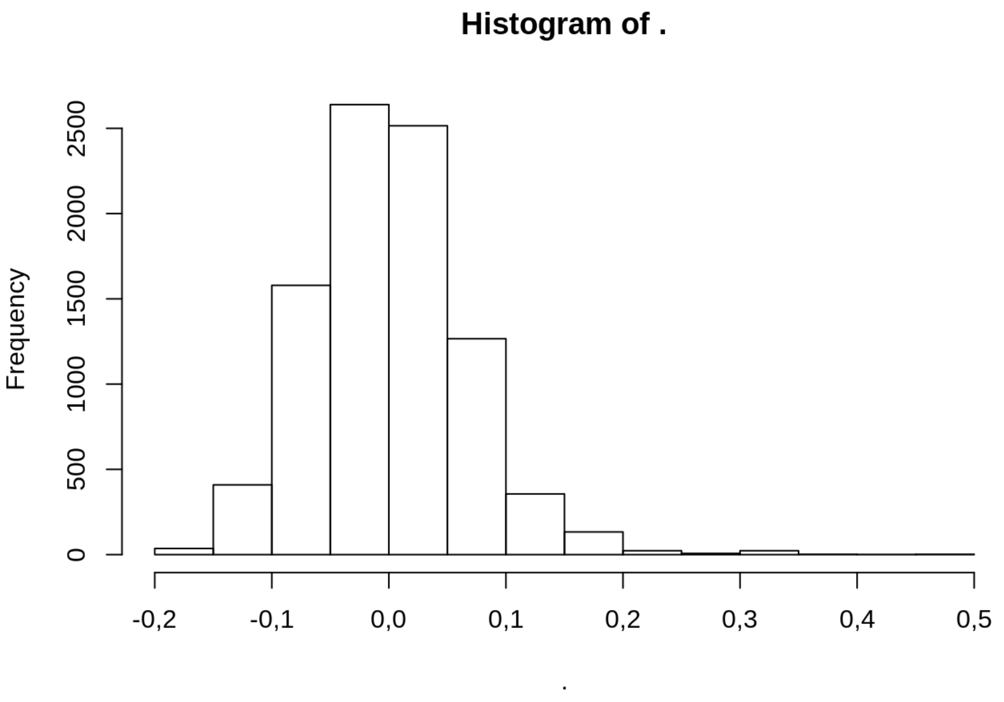
So I decided to take a PT08.S3.NOx. as a predictor, because for me it seems a little bit better. I checked the assumptions for each variable and took 4 best:

1. PT08.S5.O3.

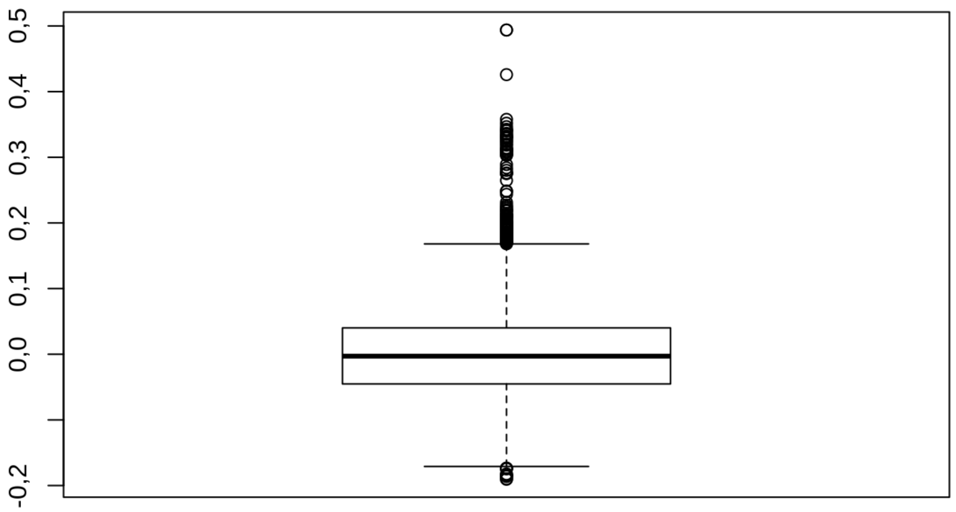
```
ggplot(airq_data, aes(x = PT08.S5.O3., y = PT08.S3.NOx. )) +
  geom_point() +
  geom_smooth(method = 'lm')
```



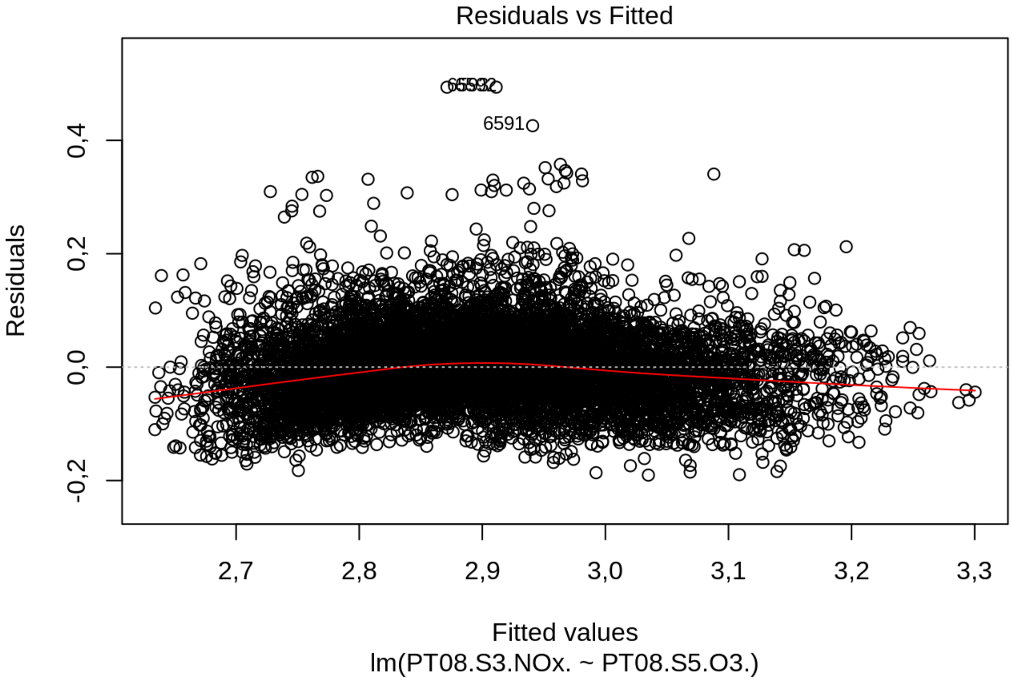

```
model_PT08.S5.O3. <- airq_data %>% lm(data = ., PT08.S3.NOx. ~ PT08.S5.O3., na.action = na.exclude)
residuals(model_PT08.S5.O3.) %>% hist()
```

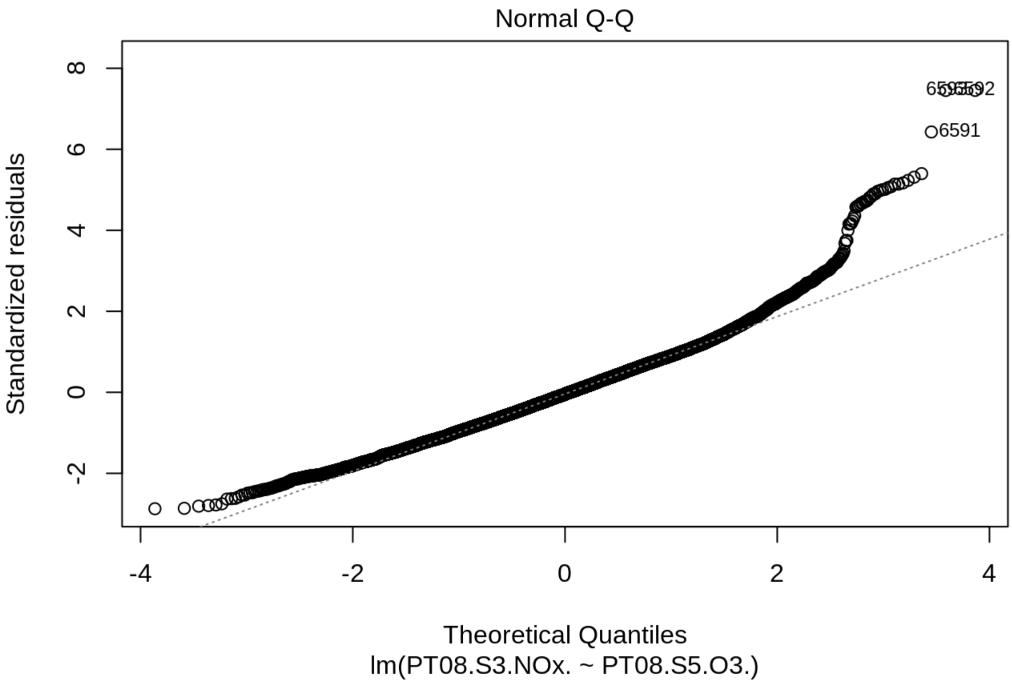



```
residuals(model_PT08.S5.O3.) %>% boxplot()
```




```
plot(model_PT08.S5.O3., which = c(1,2))
```






```
summary(model_PT08.S5.O3.)
```

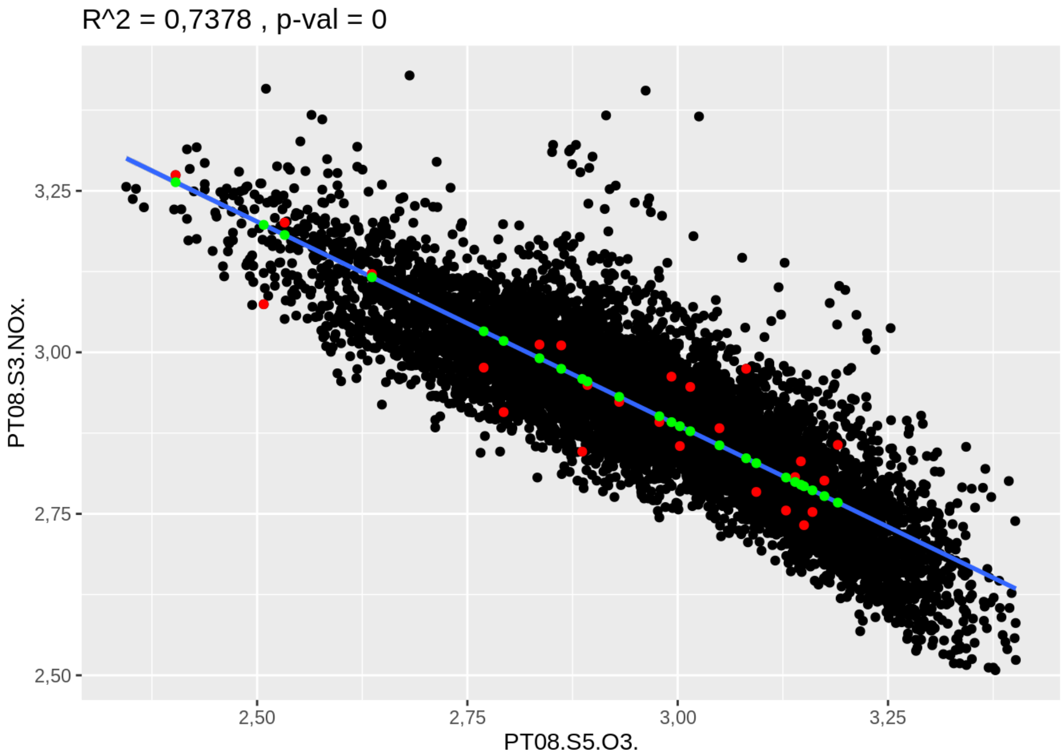
```
##
## Call:
## lm(formula = PT08.S3.NOx. ~ PT08.S5.O3., data = ., na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0,19045 -0,04522 -0,00291  0,04010  0,49389
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4,777576   0,011808   404,6   <2e-16 ***
## PT08.S5.O3. -0,630100   0,003961  -159,1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## Residual standard error: 0,06626 on 8989 degrees of freedom
##   (366 observations deleted due to missingness)
## Multiple R-squared:  0,7379, Adjusted R-squared:  0,7378
## F-statistic: 2,53e+04 on 1 and 8989 DF,  p-value: < 2,2e-16
```

Prediction:

```
test_subset_PT08.S5.O3. <-  airq_data[which(row.names(airq_data) %in% sample(row.names(airq_data), 25, replace = FALSE)), c(10,7)]
test_PT08.S5.O3. <- data.frame(PT08.S5.O3. = test_subset_PT08.S5.O3.$PT08.S5.O3.)
test_subset_PT08.S5.O3.$pred_PT08.S3.NOx. <- predict(model_PT08.S5.O3., newdata = test_PT08.S5.O3.)
colnames(test_subset_PT08.S5.O3.) <- c('real_PT08.S5.O3.', 'real_PT08.S3.NOx.', 'pred_PT08.S3.NOx.')
head(test_subset_PT08.S5.O3.)
```

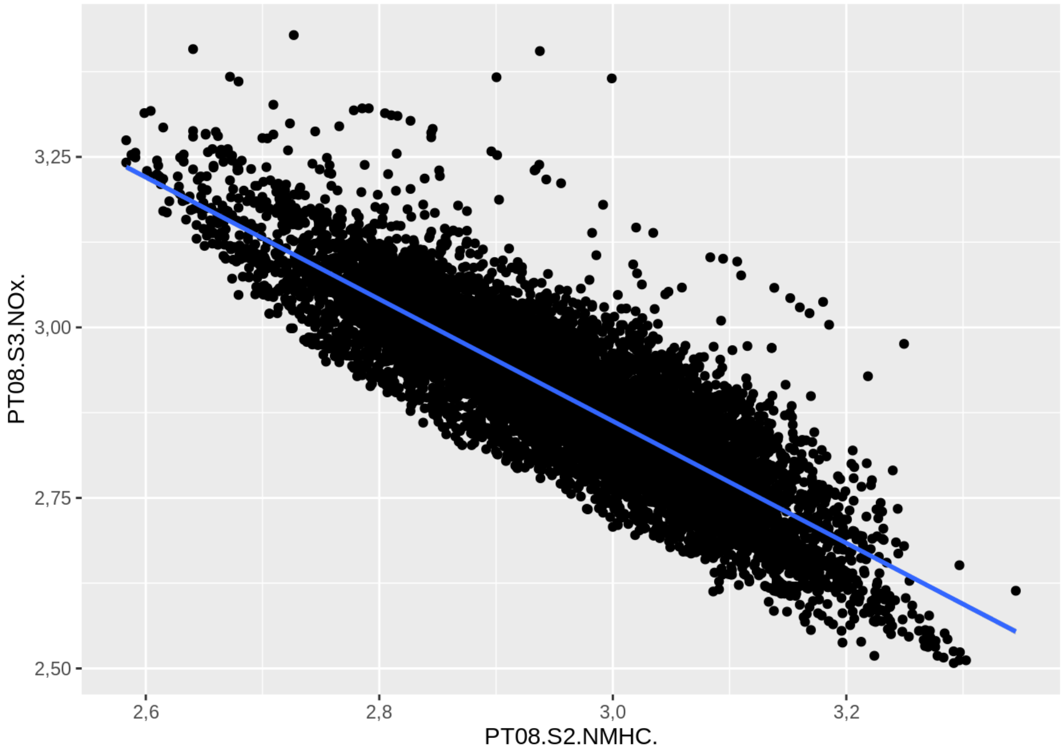
```
##      real_PT08.S5.O3. real_PT08.S3.NOx. pred_PT08.S3.NOx.
## 96          3,081347          2,974512          2,836018
## 290         3,146438          2,831230          2,795004
## 440         2,532754          3,200850          3,181687
## 483         3,014940          2,946452          2,877861
## 663         3,002598          2,854913          2,885638
## 1406        2,835691          3,011993          2,990806
```

```
R <- round(summary(model_PT08.S5.O3.)$adj.r.squared, digits = 4)
p <- round(summary(model_PT08.S5.O3.)$coefficients[2,4], digits = 3)
titl <- paste('R^2 =', as.character(R),', p-val =', as.character(p))
ggplot() +
  geom_point(data = airq_data, aes(PT08.S5.O3., PT08.S3.NOx.)) +
  geom_smooth(data = airq_data, aes(PT08.S5.O3., PT08.S3.NOx.), method = 'lm') +
  geom_point(data = test_subset_PT08.S5.O3., aes(real_PT08.S5.O3., real_PT08.S3.NOx.), color = 'red') +
  geom_point(data = test_subset_PT08.S5.O3., aes(real_PT08.S5.O3., pred_PT08.S3.NOx.), color = 'green') +
  labs(title = titl)
```

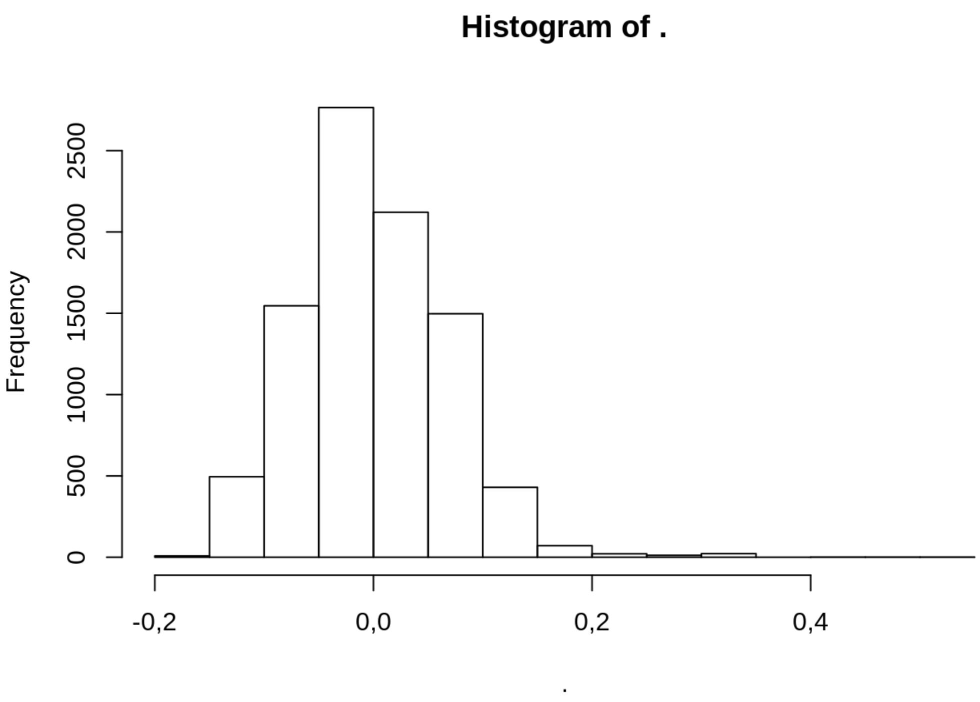


3. PT08.S2.NMHC.

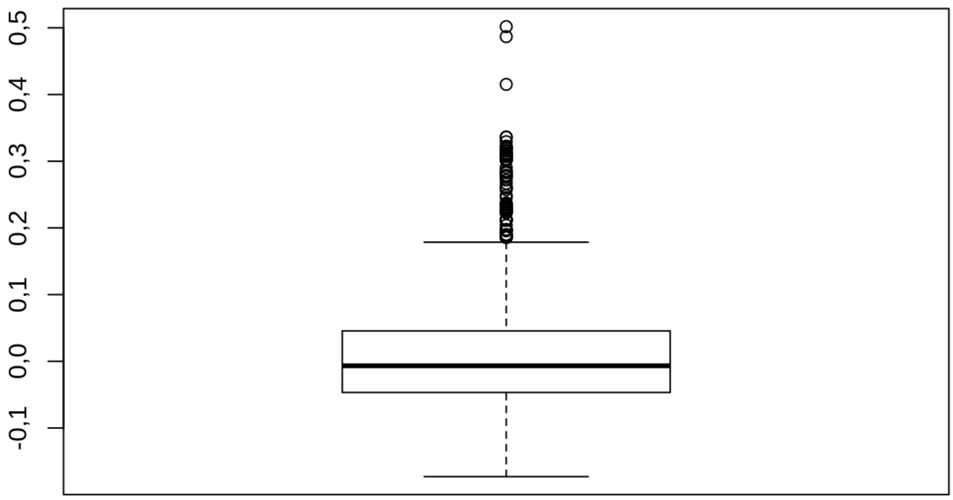
```
ggplot(airq_data, aes(x = PT08.S2.NMHC., y = PT08.S3.NOx. )) +
  geom_point() +
  geom_smooth(method = 'lm')
```



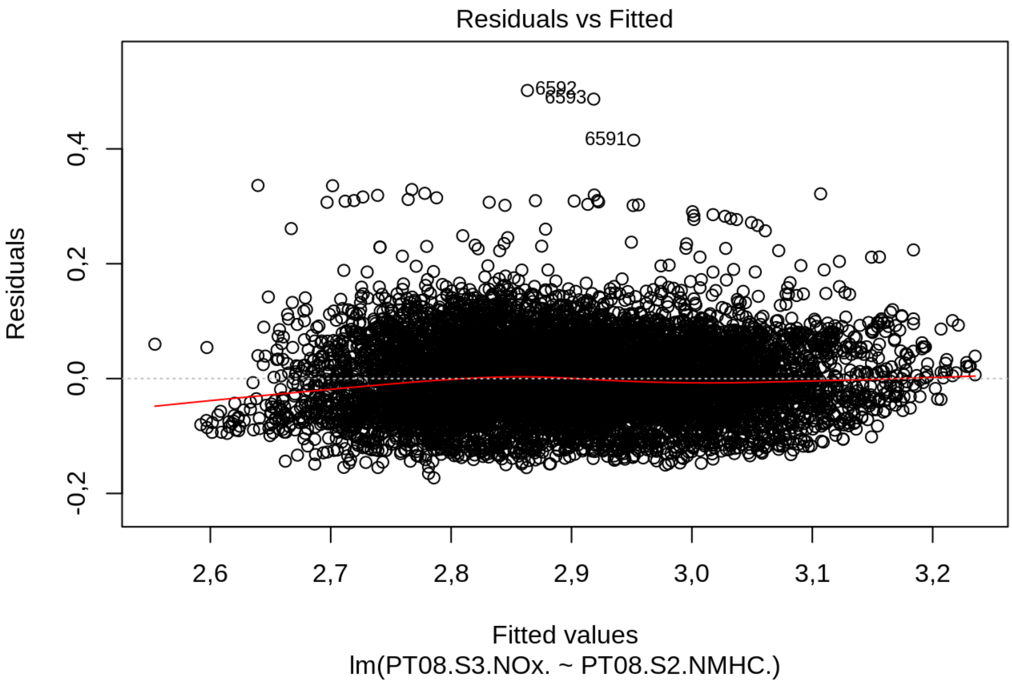

```
model_PT08.S2.NMHC. <- airq_data  %>% lm(data = .,  PT08.S3.NOx. ~ PT08.S2.NMHC.)
residuals(model_PT08.S2.NMHC.) %>% hist()
```

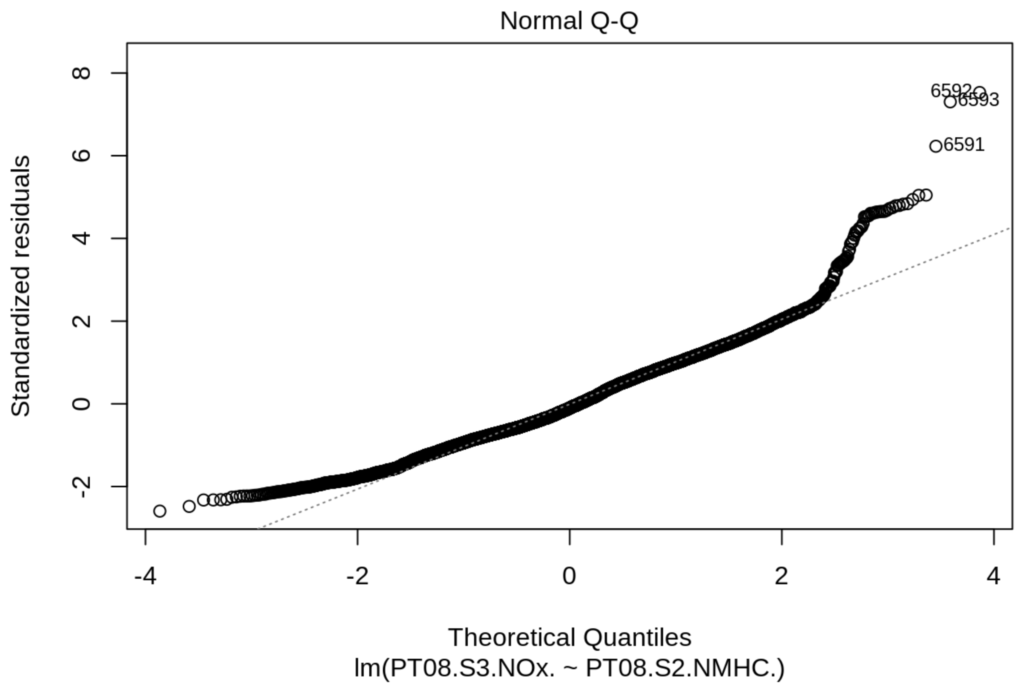



```
residuals(model_PT08.S2.NMHC.) %>% boxplot()
```




```
plot(model_PT08.S2.NMHC., which = c(1,2))
```






```
summary(model_PT08.S2.NMHC.)
```

```
##
## Call:
## lm(formula = PT08.S3.NOx. ~ PT08.S2.NMHC., data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0,17290 -0,04669 -0,00673  0,04555  0,50176
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5,544914   0,016758   330,9   <2e-16 ***
## PT08.S2.NMHC. -0,894112   0,005666  -157,8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## Residual standard error: 0,06665 on 8989 degrees of freedom
##   (366 observations deleted due to missingness)
## Multiple R-squared:  0,7348, Adjusted R-squared:  0,7348
## F-statistic: 2,491e+04 on 1 and 8989 DF,  p-value: < 2,2e-16
```

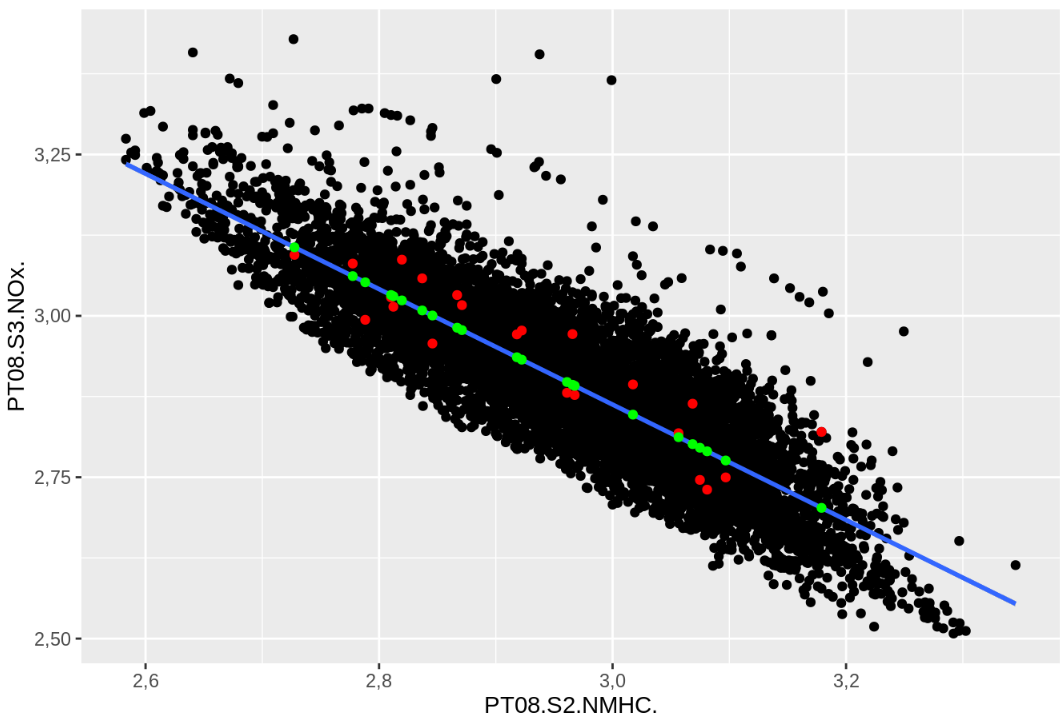
Prediction:

```
test_subset_PT08.S2.NMHC. <-  airq_data[which(row.names(airq_data) %in% sample(row.names(airq_data), 25, replace = FALSE)), c(5,7)]
test_PT08.S2.NMHC. <- data.frame(PT08.S2.NMHC. = test_subset_PT08.S2.NMHC.$PT08.S2.NMHC.)
test_subset_PT08.S2.NMHC.$pred_PT08.S3.NOx. <- predict(model_PT08.S2.NMHC., newdata = test_PT08.S2.NMHC.)
colnames(test_subset_PT08.S2.NMHC.) <- c('real_PT08.S2.NMHC.', 'real_PT08.S3.NOx.', 'pred_PT08.S3.NOx.')
head(test_subset_PT08.S2.NMHC.)
```

```
##      real_PT08.S2.NMHC. real_PT08.S3.NOx. pred_PT08.S3.NOx.
## 174            3,068557          2,863917          2,801282
## 254            2,836957          3,058046          3,008358
## 524            2,965672          2,971740          2,893272
## 791            2,819544          3,087071          3,023927
## 1033           3,178977          2,820201          2,702554
## 1509           3,056524          2,818226          2,812041
```

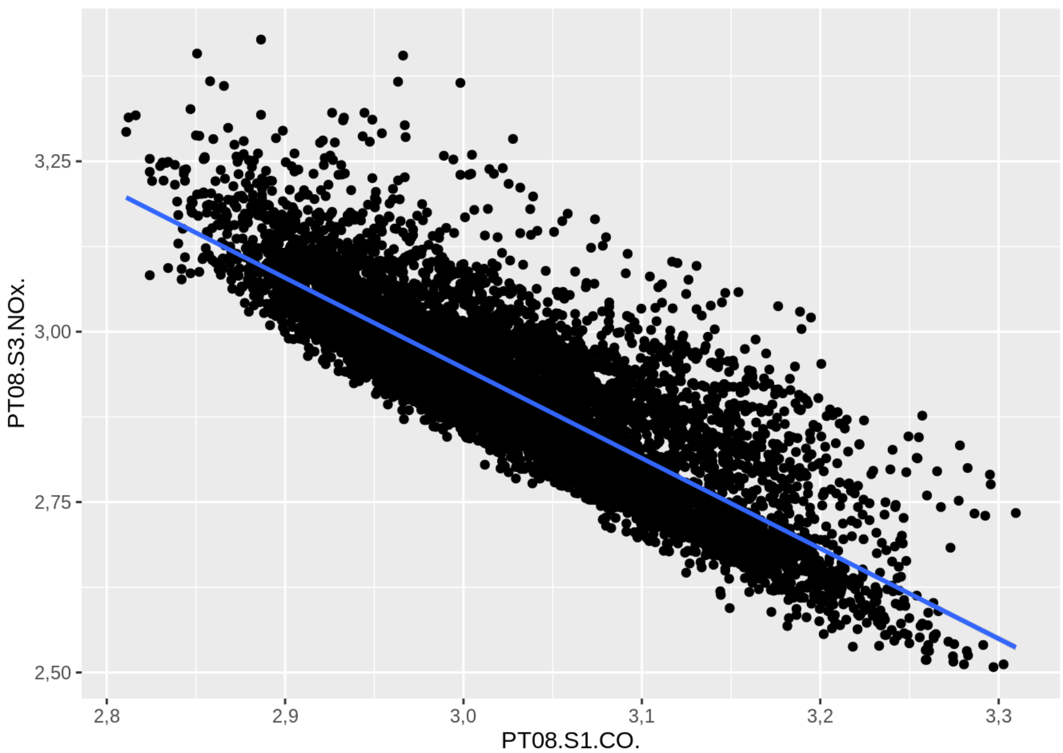
```
R <- round(summary(model_PT08.S2.NMHC.)$adj.r.squared, digits = 3)
p <- round(summary(model_PT08.S2.NMHC.)$coefficients[2,4], digits = 3)
titl <- paste('R^2 =', as.character(R),', p-val =', as.character(p))
ggplot() +
  geom_point(data = airq_data, aes(PT08.S2.NMHC., PT08.S3.NOx.)) +
  geom_smooth(data = airq_data, aes(PT08.S2.NMHC., PT08.S3.NOx.), method = 'lm') +
  geom_point(data = test_subset_PT08.S2.NMHC., aes(real_PT08.S2.NMHC., real_PT08.S3.NOx.), color = 'red') +
  geom_point(data = test_subset_PT08.S2.NMHC., aes(real_PT08.S2.NMHC., pred_PT08.S3.NOx.), color = 'green') +
  labs(title = titl)
```

R^2 = 0,735 , p-val = 0

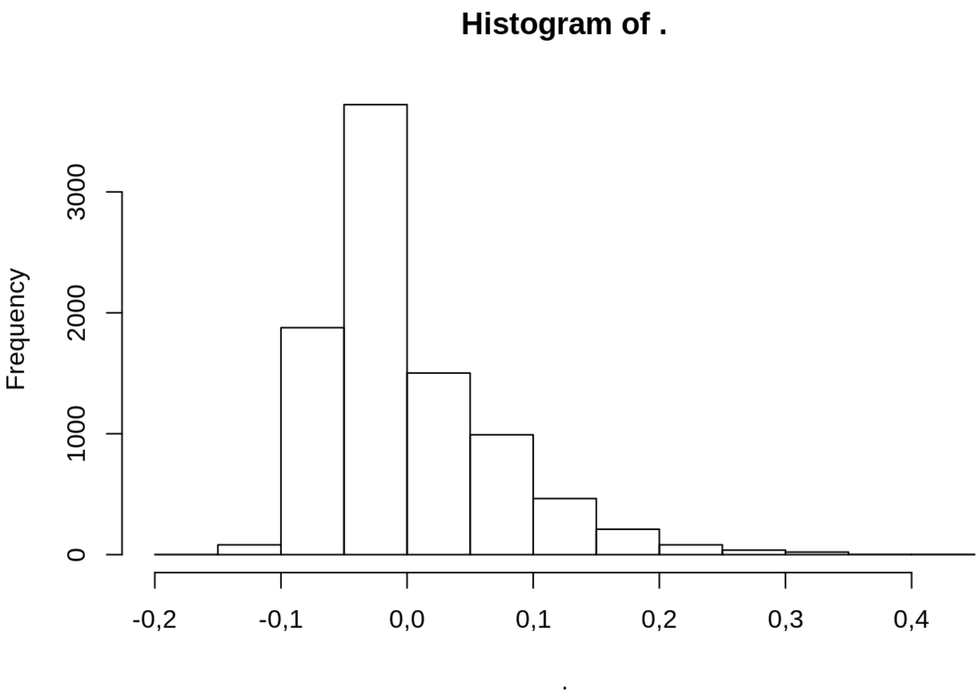


3. PT08.S1.CO.

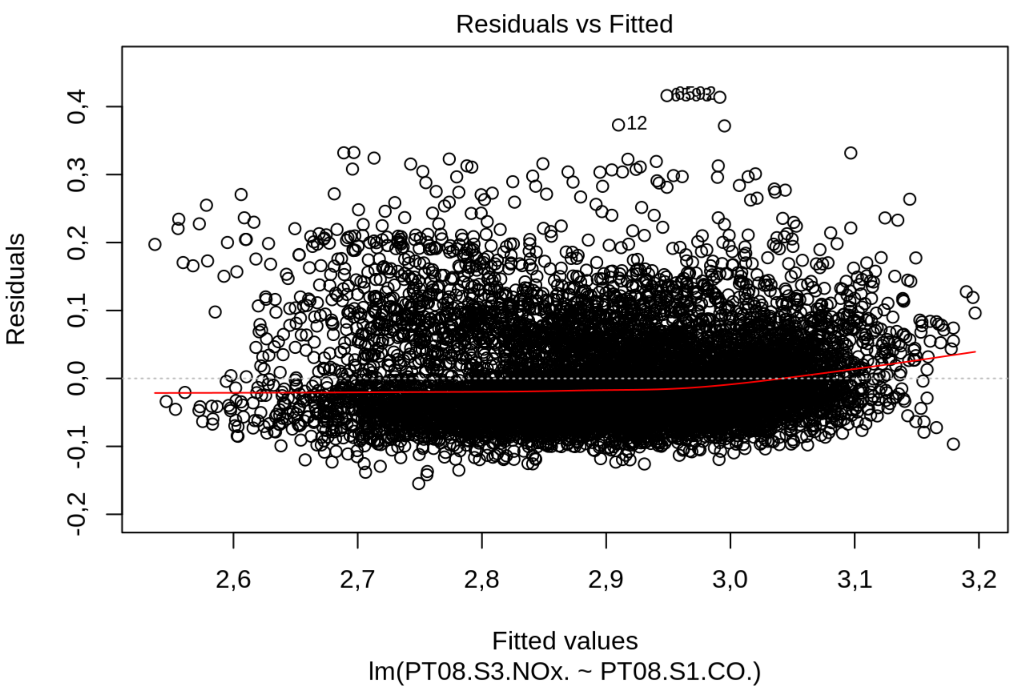
```
ggplot(airq_data, aes(x = PT08.S1.CO., y = PT08.S3.NOx. )) +
  geom_point() +
  geom_smooth(method = 'lm')
```

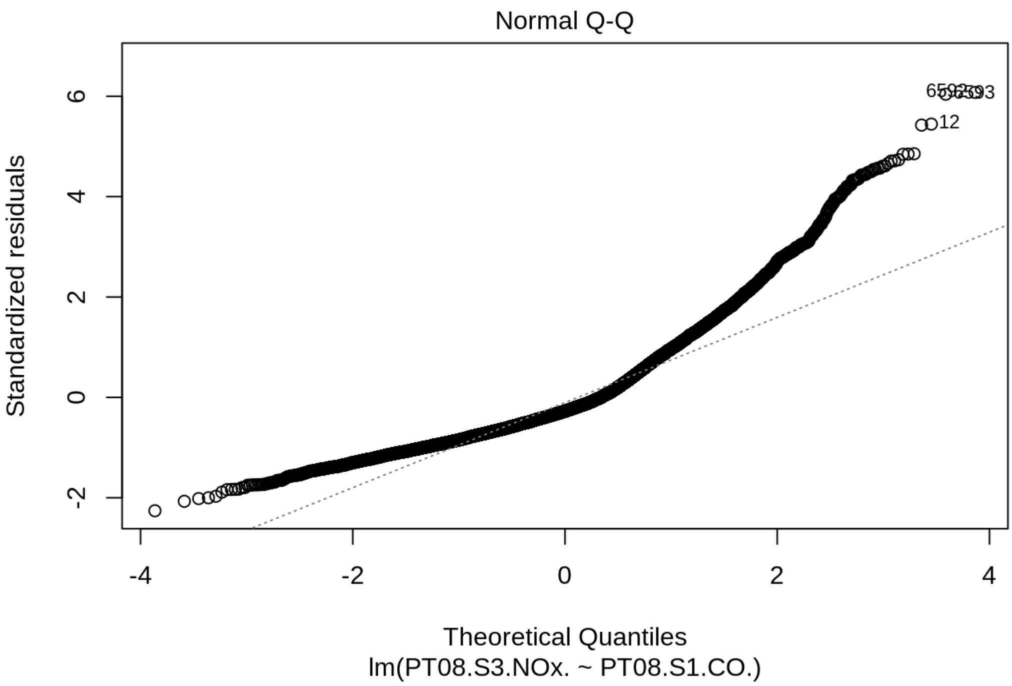



```
model_PT08.S1.CO. <- airq_data  %>% lm(data = .,  PT08.S3.NOx. ~ PT08.S1.CO.)
residuals(model_PT08.S1.CO.) %>% hist()
```




```
plot(model_PT08.S1.CO., which = c(1,2))
```






```
summary(model_PT08.S1.CO.)
```

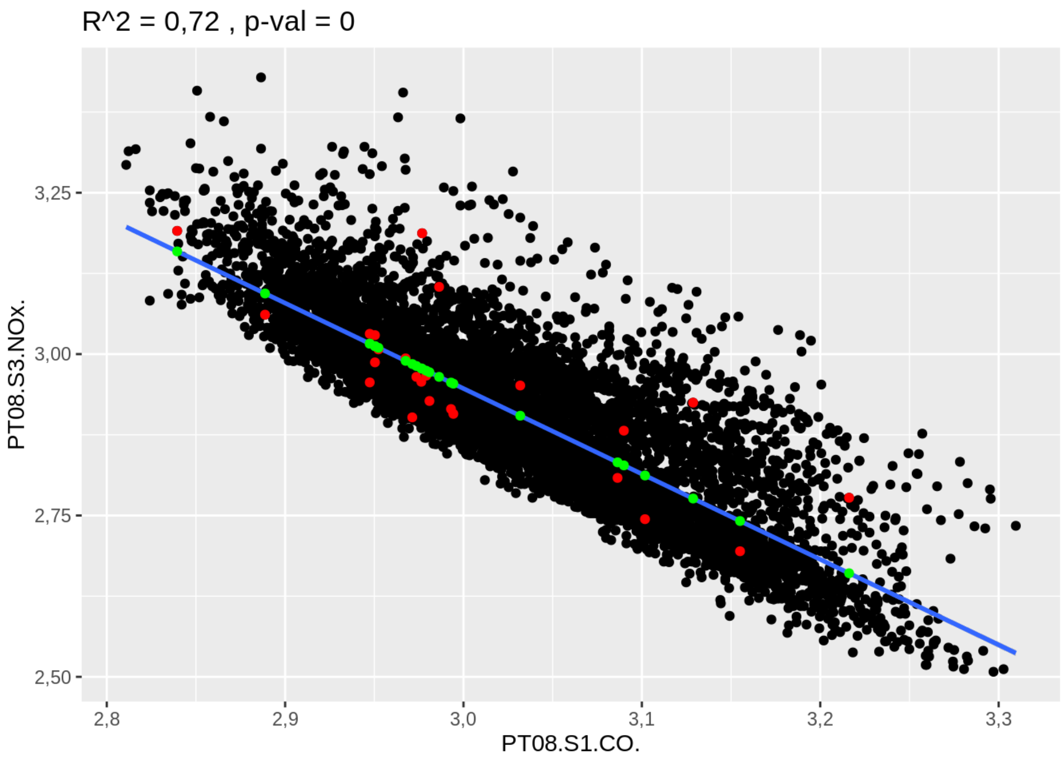
```
##
## Call:
## lm(formula = PT08.S3.NOx. ~ PT08.S1.CO., data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0,15472 -0,04621 -0,01888  0,03216  0,41618
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6,917600   0,026427   261,8   <2e-16 ***
## PT08.S1.CO. -1,323658   0,008709  -152,0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## Residual standard error: 0,0685 on 8989 degrees of freedom
##   (366 observations deleted due to missingness)
## Multiple R-squared:  0,7199, Adjusted R-squared:  0,7198
## F-statistic: 2,31e+04 on 1 and 8989 DF,  p-value: < 2,2e-16
```

Prediction:

```
test_subset_PT08.S1.CO. <-  airq_data[which(row.names(airq_data) %in% sample(row.names(airq_data), 25, replace = FALSE)), c(2,7)]
test_PT08.S1.CO. <- data.frame(PT08.S1.CO. = test_subset_PT08.S1.CO.$PT08.S1.CO.)
test_subset_PT08.S1.CO.$pred_PT08.S3.NOx. <- predict(model_PT08.S1.CO., newdata = test_PT08.S1.CO.)
colnames(test_subset_PT08.S1.CO.) <- c('real_PT08.S1.CO.', 'real_PT08.S3.NOx.', 'pred_PT08.S3.NOx.')
head(test_subset_PT08.S1.CO.)
```

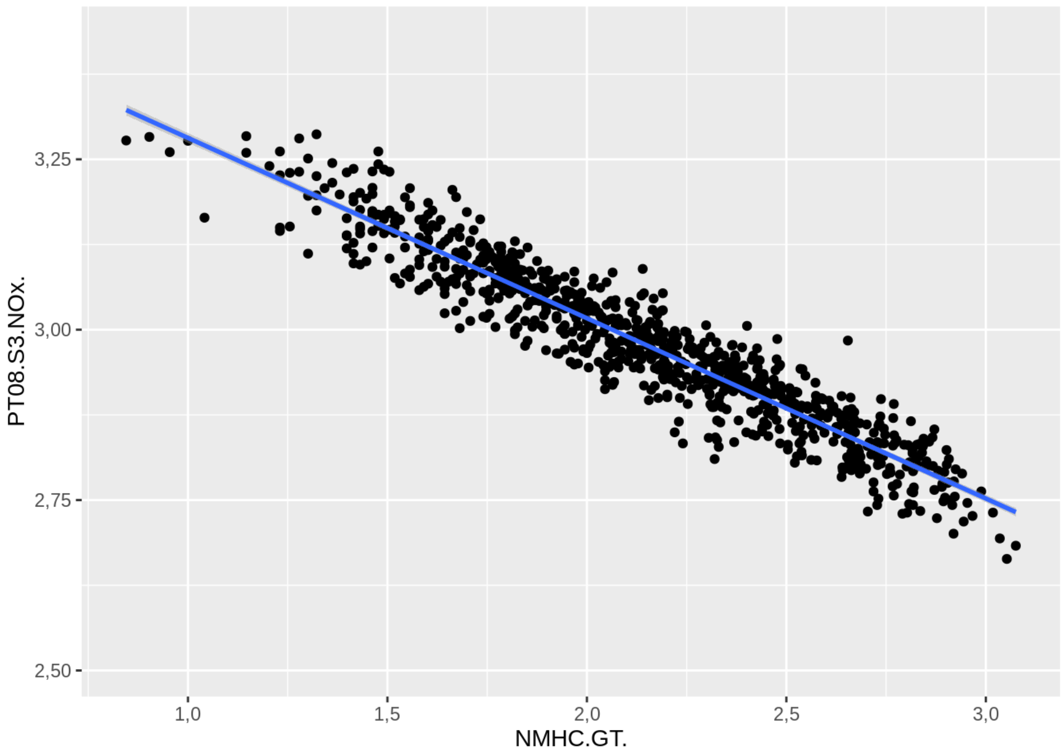
```
##      real_PT08.S1.CO. real_PT08.S3.NOx. pred_PT08.S3.NOx.
## 296          2,986324          3,104146          2,964728
## 599          3,128722          2,924796          2,776241
## 1056         3,216166          2,777427          2,660495
## 2084         2,952308          3,007321          3,009753
## 2182         3,031812          2,951338          2,904516
## 2770         2,980912          2,927370          2,971891
```

```
R <- round(summary(model_PT08.S1.CO.)$adj.r.squared, digits = 3)
p <- round(summary(model_PT08.S1.CO.)$coefficients[2,4], digits = 3)
titl <- paste('R^2 =', as.character(R),', p-val =', as.character(p))
ggplot() +
  geom_point(data = airq_data, aes(PT08.S1.CO., PT08.S3.NOx.)) +
  geom_smooth(data = airq_data, aes(PT08.S1.CO., PT08.S3.NOx.), method = 'lm') +
  geom_point(data = test_subset_PT08.S1.CO., aes(real_PT08.S1.CO., real_PT08.S3.NOx.), color = 'red') +
  geom_point(data = test_subset_PT08.S1.CO., aes(real_PT08.S1.CO., pred_PT08.S3.NOx.), color = 'green') +
  labs(title = titl)
```

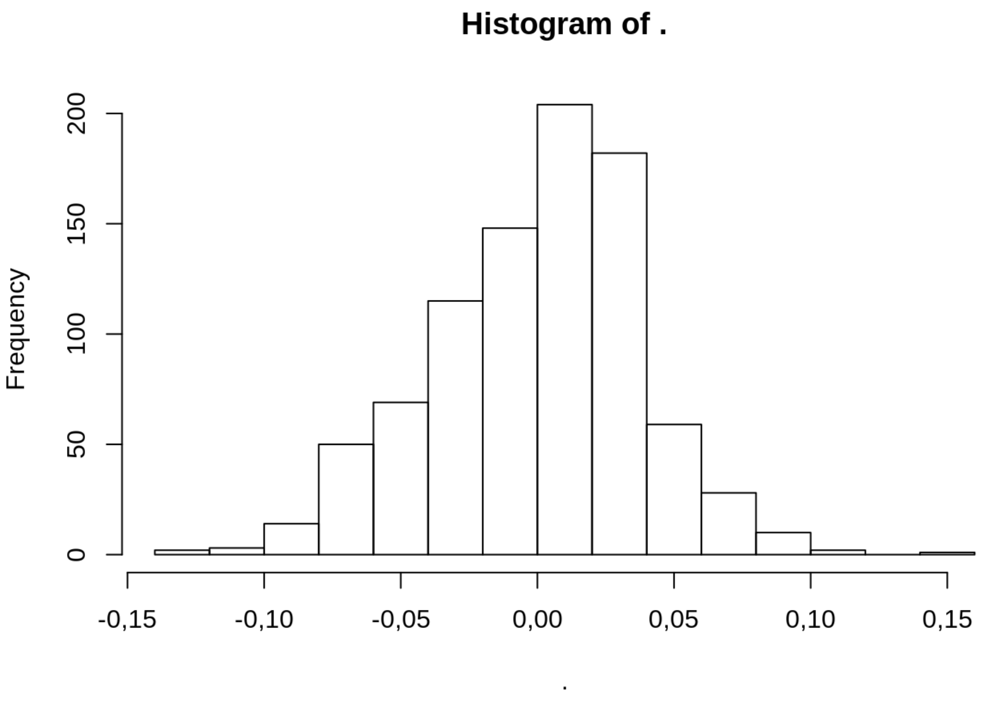


4. NMHC.GT. - there are a lot of missing data, but it was just interesting that for the rest data the model and prediction is very good. *But, of course we should not use this model, due to missing a lot of data.

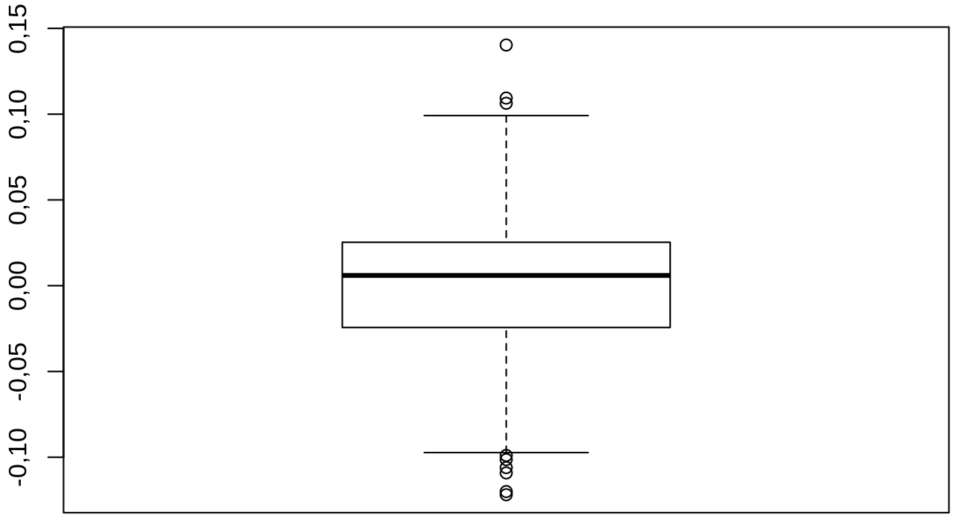
```
ggplot(airq_data, aes(x = NMHC.GT., y = PT08.S3.NOx.)) +
  geom_point() +
  geom_smooth(method = 'lm')
```



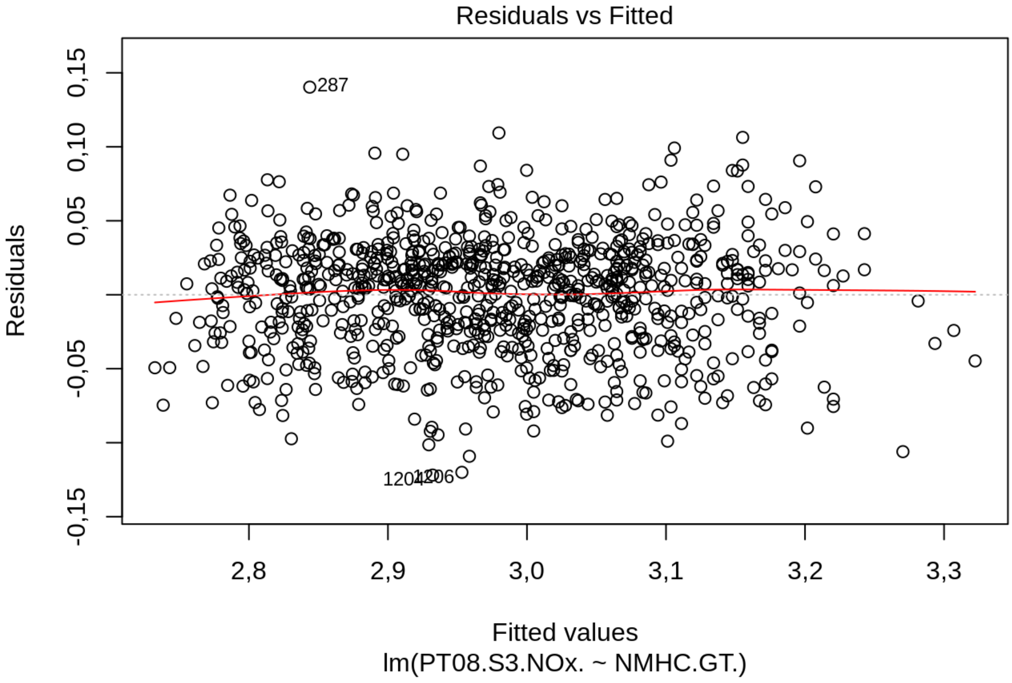

```
model_NMHC.GT. <- airq_data %>%  lm(data = ., PT08.S3.NOx.~ NMHC.GT., na.action = na.exclude)
residuals(model_NMHC.GT.) %>% hist()
```

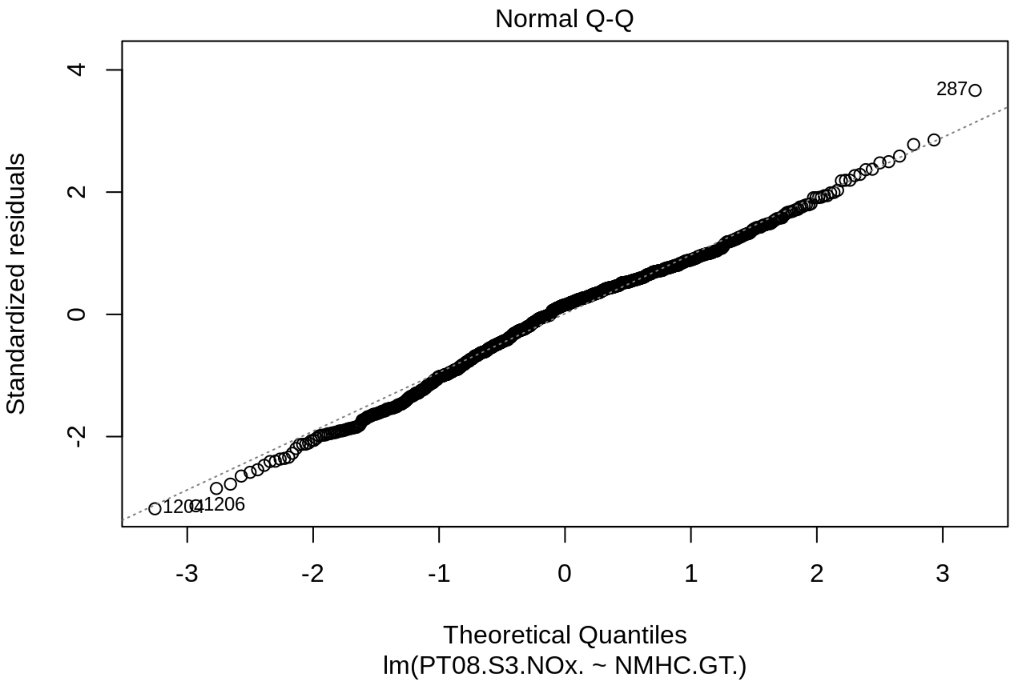



```
residuals(model_NMHC.GT.) %>% boxplot()
```




```
plot(model_NMHC.GT., which = c(1,2))
```






```
summary(model_NMHC.GT.)
```

```
##
## Call:
## lm(formula = PT08.S3.NOx. ~ NMHC.GT., data = ., na.action = na.exclude)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0,121896 -0,024351  0,005981  0,025296  0,140314
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3,545910   0,006578  539,08   <2e-16 ***
## NMHC.GT.    -0,264544   0,003004  -88,07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## Residual standard error: 0,03834 on 885 degrees of freedom
##   (8470 observations deleted due to missingness)
## Multiple R-squared:  0,8976, Adjusted R-squared:  0,8975
## F-statistic:  7757 on 1 and 885 DF,  p-value: < 2,2e-16
```

Prediction:

```
airq_data_2 <- airq_data %>% drop_na()
test_subset_NMHC.GT. <-  airq_data_2[which(row.names(airq_data_2) %in% sample(row.names(airq_data_2), 25, replace = FALSE)), c(3,7)]
test_NMHC.GT. <- data.frame(NMHC.GT. = test_subset_NMHC.GT.$NMHC.GT.)
test_subset_NMHC.GT.$pred_PT08.S3.NOx. <- predict(model_NMHC.GT., newdata = test_NMHC.GT.)
colnames(test_subset_NMHC.GT.) <- c('real_NMHC.GT.', 'real_PT08.S3.NOx.', 'pred_PT08.S3.NOx.')
head(test_subset_NMHC.GT.)
```

```
##     real_NMHC.GT. real_PT08.S3.NOx. pred_PT08.S3.NOx.
## 17       1,886491          3,085647          3,046850
## 68       2,187521          2,994317          2,967214
## 91       2,161368          2,957607          2,974133
## 103      1,819544          2,993436          3,064560
## 126      1,826075          3,003461          3,062833
## 314      2,372912          2,931458          2,918170
```

```
R <- round(summary(model_NMHC.GT.)$adj.r.squared, digits = 3)
p <- round(summary(model_NMHC.GT.)$coefficients[2,4], digits = 3)
titl <- paste('R^2 =', as.character(R),', p-val =', as.character(p))
ggplot() +
  geom_point(data = airq_data, aes(NMHC.GT., PT08.S3.NOx.)) +
  geom_smooth(data = airq_data, aes(NMHC.GT., PT08.S3.NOx.), method = 'lm') +
  geom_point(data = test_subset_NMHC.GT., aes(real_NMHC.GT., real_PT08.S3.NOx.), color = 'red') +
  geom_point(data = test_subset_NMHC.GT., aes(real_NMHC.GT., pred_PT08.S3.NOx.), color = 'green') +
  labs(title = titl)
```