AIR QUALITY

1. Wstęp

   Celem analizy jest sprawdzenie czy zanieczyszczenia powietrza jednego rodzaju mają wpływ na wielkość wartości pomiarowych zanieczyszczeń innych rodzajów. A także czy wpływa na to panująca temperatura lub wilgotność powietrza.
   Innym pomysłem może być ocena zdolności pomiarowej urządzenia oraz podatności na rozkalibrowanie na przestrzeni czasu.

2. Opis danych

   Wykorzystany dataset pochodzi ze strony UCI (Machine Learning Repository): http://archive.ics.uci.edu/ml/datasets/Air+Quality

   Informacje o datasecie:
   Dataset zawiera 9358 obserwacji pomiarowych. Prezentowane wartości odczytów są średnimi pomiarami godzinowymi podawanymi przez 5 czujników znajdujących się w urządzeniu (Air Quality Chemical Multisensor Device) służącym do pomiaru stężeń związków znajdujących się w powietrzu.
   Urządzenie zostało zamontowane na wolnej przestrzeni, na wysokości jezdni, we włoskim mieście o dość dużym stopniu zanieczyszczenia powietrza. Dane pomiarowe pochodzą z okresu 1 roku (marzec 2004 – luty 2015).
   Lotne związki, których stężenia są mierzone to: CO, NMHC (Non Metanic Hydrocarbons), Benzen (C6H6), NOx (Total Nitrogen Oxides), NO2 (Nitrogen Dioxide). Ich wartości są podawane przez czujniki (analizatory) referencyjne.
   Brakujące wartości pomiarowe zostały w zbiorze danych oznaczone wartością -200.

   Opis zmiennych / struktura zbiorów:
   0) Date (DD/MM/YYYY)
   1) Time (HH.MM.SS)
   2) CO(GT) - rzeczywista średnio-godzinna koncentracja (wartość) CO w mg/m^3 (analizator referencyjny)
   3) PT08.S1 (tin oxide) - średnio-godzinna odpowiedź z sensora (pomiar CO)
   4) NMHC(GT) - rzeczywista średnio-godzinna koncentracja dla Non Metanic HydroCarbons w microg/m^3 (analizator referencyjny)
   5) C6H6(GT) - rzeczywista średnio-godzinna koncentracja dla Benzenu w microg/m^3 (analizator referencyjny)
   6) PT08.S2 (titania) - średnio-godzinna odpowiedź z sensora (pomiar NMHC)
   7) NOx(GT) - rzeczywista średnio-godzinna koncentracja dla NOx w ppb (analizator referencyjny)
   8) PT08.S3 (tungsten oxide) - średnio-godzinna odpowiedź z sensora (pomiar NOx)
   9) NO2(GT) - rzeczywista średnio-godzinna koncentracja NO2 w microg/m^3 (analizator referencyjny)
   10) PT08.S4 (tungsten oxide) - średnio-godzinna odpowiedź z sensora (pomiar NO2)
   11) PT08.S5 (indium oxide) - średnio-godzinna odpowiedź z sensora (pomiar O3)
   12) T - Temperatura w °C
   13) RH - Względna wilogotność powietrza w (%)
   14) AH – Absolutna wilgotność

Próbka danych:

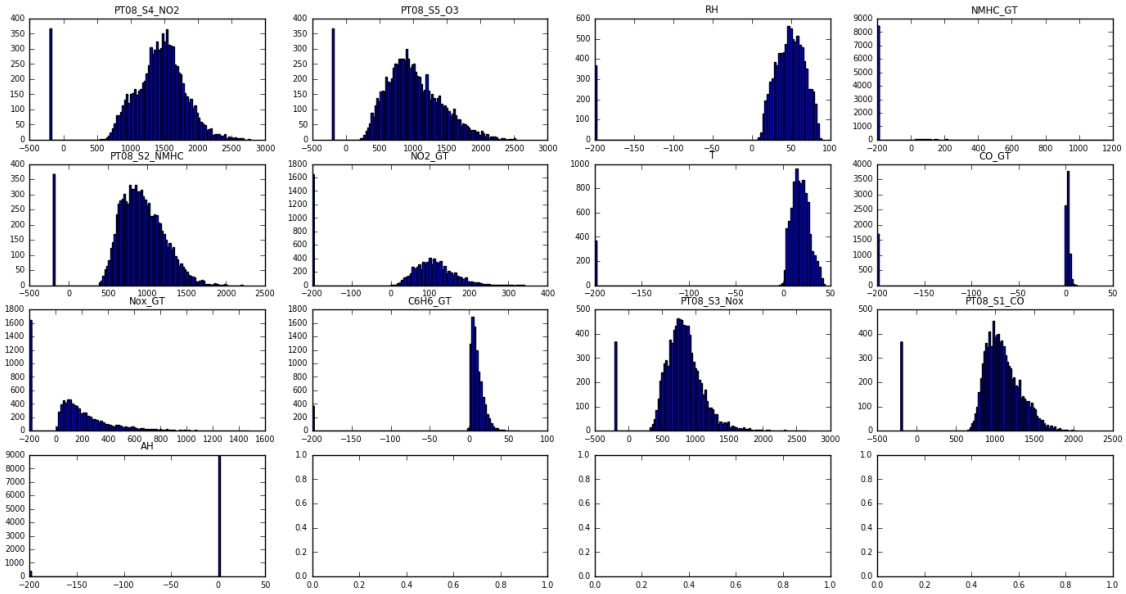
```
In [1]: air_quality=read.csv("AirQualityUCI.csv", header=TRUE, sep=";")

In [2]: head(air_quality)
```

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

3. Opis procesu przygotowywania danych do analizy - kolejne kroki

- Pierwszym krokiem było zapoznanie się z danymi, przejrzenie ich, orientacja jak dużej ilości danych w każdym typie pomiaru (dane kolumnowe) brakuje i jaki może mieć to wpływ na analizę. W jaki sposób brak danych wpływa na inne wartości oraz jakość analizy.

- Zwizualizowałem to ilościowo na histogramach:



Wartości -200 symbolizują brakujące wartości. Po przyjrzeniu się histogramom mogłem ocenić jak duże są braki. (Histogramy oraz inne wykresy zamieściłem w katalogu /charts na githubie).

- Usunięcie danych nie wchodziło w grę, ponieważ brakujące wartości znajdowały się w różnych (niezależnych) miejscach dla zmiennych, których wzajemne oddziaływanie chciałem mierzyć. Przykładowo ubytek danych dla wielkości CO_GT występował w zakresie dat 08/05/2004 – 10/05/2004, podczas gdy dla PT08.S1(CO) były dane, za to nie było ich w innym przedziale, w którym za to istniały dla CO_GT. Usunięcie danych w przypadku kilku zmiennych skutkowałoby wyłączeniem ich z analizy.

- Zdecydowałem się na zastąpienie brakujących danych medianą.

```
def corrections(variables):
    print("Wstępne poprawianie danych:")
    for name, variable in variables.items():
        for index, value in enumerate(variable):
            if (value == p.missing_value):
                #konsola
                print ("Zmienna:", name, "indeks:", index, "anomalia o wartości:", value)
                median = np.median(variable)
                print("Naprawiam. Stara wartość:", variable[index], ", nowa wartość:", median)
                variable[index] = median
        #wyliczamy podstawowe statystyki
        basic_stats(name, variable)
```

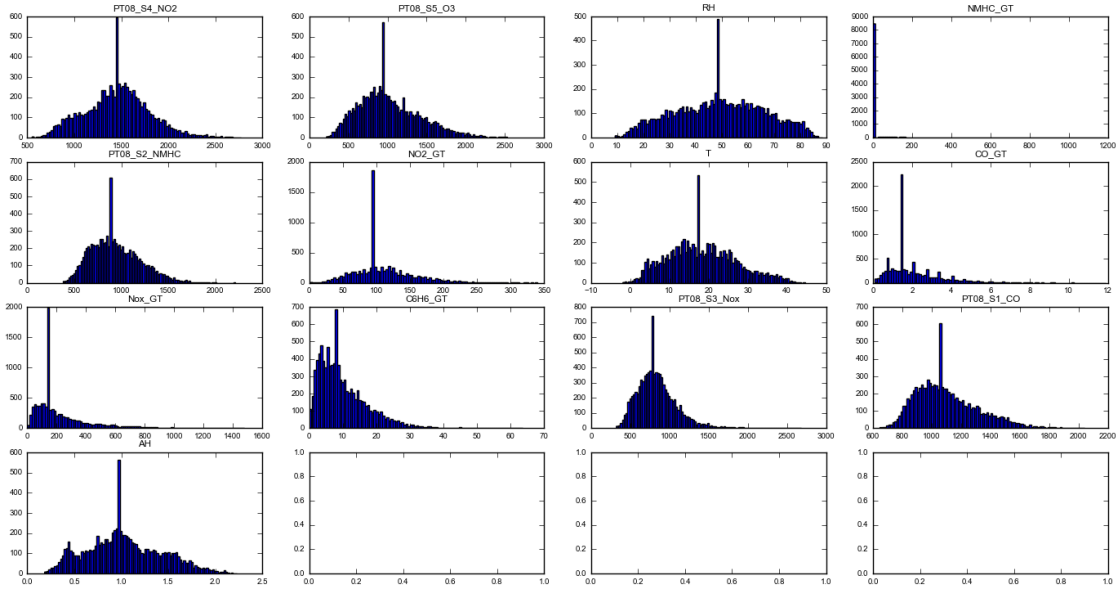
- Jadnak po wyliczeniu podstawowych statystyk danych okazało się, że dla zmiennej NMHC(GT) brakujących wartości jest na tyle dużo, że mediana wychodzi -200, czyli jest ich ponad połowę.
Oto ta statystyka:

Podstawowa statystyka dla: NMHC_GT
MIN: -200.0
MAX: 1189.0
ŚREDNIA: -159.090092979
MEDIANA: -200.0
ODCH. STD: 139.781622937
WARIANCJA: 19538.9021109
HISTOGRAM: (array([8443, 264, 322, 151, 73, 46, 33, 19, 3, 3], dtype=int64), array([-200. , -61.1, 77.8, 216.7, 355.6, 494.5, 633.4,
772.3, 911.2, 1050.1, 1189.]))

- Zrozumiałem, że liczenie mediany, uwzględniając wartość oznaczoną jako brakującą (-200) nie jest w tym przypadku właściwe i może przesuwać charakterystykę w stronę niższych wartości.
Dlatego zastąpiłem brakujące wartości medianą, która jeśli liczona bez wartości -200 (missing_value).

```
def corrections_final(variables):
    print("Ostateczne poprawienie danych:")
    for name, variable in variables.items():
        list = []; i=0
        for index, value in enumerate(variable):
            if (value != p.missing_value):
                i += 1
                list.append(value)
        sorted_list = sorted(list)
        i = int(i/2)
        print("Mediana:", name, "po wykluczeniu wartości nieprawdłowych:", sorted_list[i])
```

- Następnie policzyłem nowe statystyki oraz wylistowałem do pliku wszystkie wartości, które zastąpiłem podczas tej procedury (rezultat zapisałem w pliku verbose_full.txt)
- Wygenerowałem nowe histogramy dla zmiennych, aby upewnić się, że w zbiorze danych nie ma zarurzeń. Wartości brakujące zostały zastąpione nową wartością mediany, co nie przesuwa wartości w lewą stronę.



4. Analiza danych - przyjęte założenia, krótki opis metod i obranej metodologii analizy.

- Aby zrozumiem w jaki sposób zmienne zależą od siebie, policzyłem wzajemne korelacje. Skupiłem się na silnych korelacjach, tzn. Przyjąłem, że wartość bezwzględna dla współczynnika korelacji ma być większa niż 0.7 (abs(correlation_ratio) > 0,7).

- Okazało się, że najsilniej są skorelowane następujące zmienne:

Silna korelacja między PT08_S2_NMHC a C6H6_GT : 0.981673245516

Silna korelacja między PT08_S2_NMHC a PT08_S1_CO : 0.893069642991

Silna korelacja między PT08_S2_NMHC a PT08_S4_NO2 : 0.777016978588

Silna korelacja między PT08_S2_NMHC a CO_GT : 0.790321338488

Silna korelacja między PT08_S2_NMHC a PT08_S5_O3 : 0.880710257499

Silna korelacja między C6H6_GT a PT08_S1_CO : 0.88397896334

Silna korelacja między C6H6_GT a PT08_S4_NO2 : 0.764774586697

Silna korelacja między C6H6_GT a CO_GT : 0.802251581903

Silna korelacja między C6H6_GT a PT08_S5_O3 : 0.865863834946

Silna korelacja między PT08_S1_CO a CO_GT : 0.775029012384

Silna korelacja między PT08_S1_CO a PT08_S5_O3 : 0.89949290942

Silna korelacja między Nox_GT a NO2_GT : 0.768881291528

Silna korelacja między Nox_GT a CO_GT : 0.791624000154

Silna korelacja między CO_GT a PT08_S5_O3 : 0.762061522303

- Są to korelacje wzajemne, wyeliminowałem więc duplikaty, czyli jeśli mamy, że PT08_S2_NMHC silnie koreluje z C6H6_GT, to w 2 stronę już nie przedstawiam.

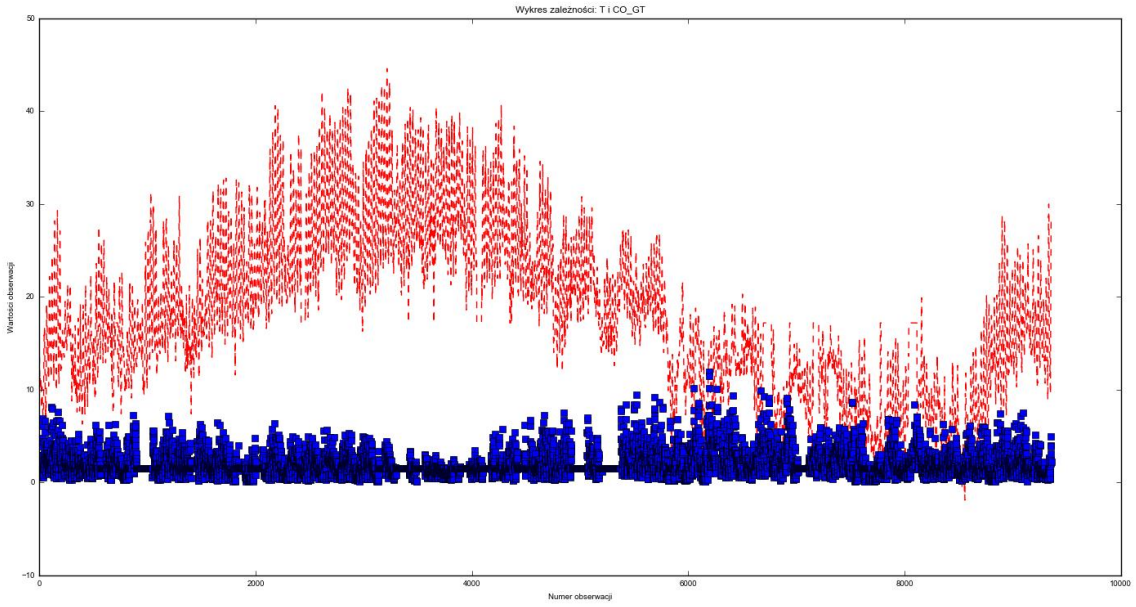
Z wartości korelacji wyciągnąłem następujące wnioski:
+ Odczyty 5 czujników PT08_S2_NMHC, PT08_S1_CO, PT08_S4_NO2, PT08_S5_O3, PT08_S4_NO2, czyli wszystkich, są skorelowane z innymi zmiennymi.

+ W przypadku wszystkich czujników są to zarówno korelacje z wartościami pomiarów stężeń związków lotnych, jak i z odczytami innych innych czujników.

+ Nie obserwuje się silnych korelacji pomiędzy wartościami temperatury (T), względnej wilgotności powietrza (RH), bezwzględnej wilgotności (AH), a odczytami pomiarów stężeń, czy odczytami czujników.

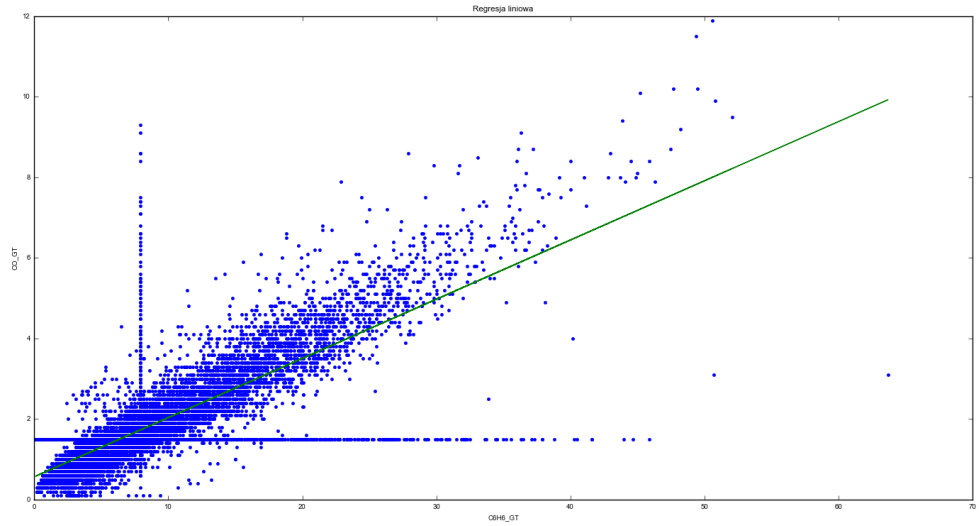
+ Wielkości te nie są też skorelowane między sobą, co biorąc pod uwagę naturę tych wielkości – ma sens.

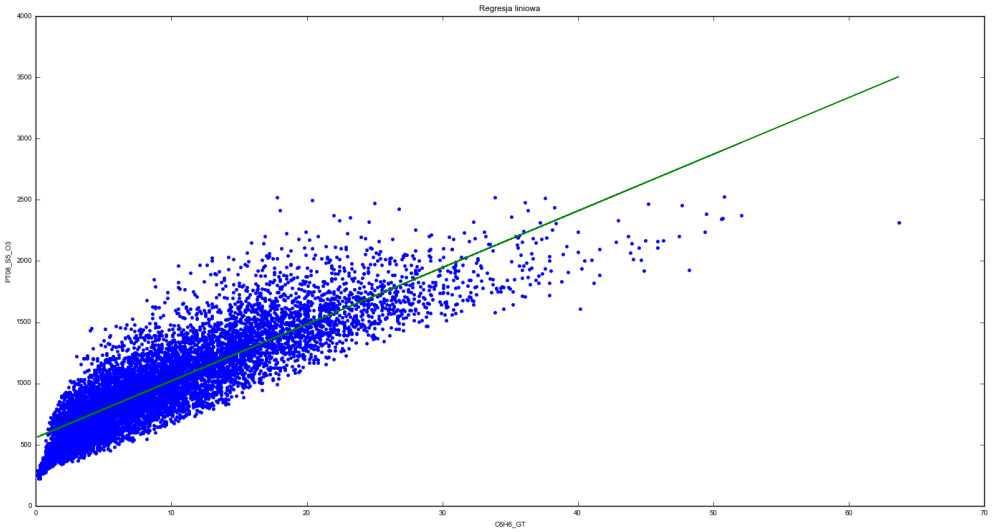
- Dla potwierdzenia tej prawidłowości, zwizualizowałem na jednym wykresie wartości temperatury (T) oraz pomiaru CO (CO_GT):

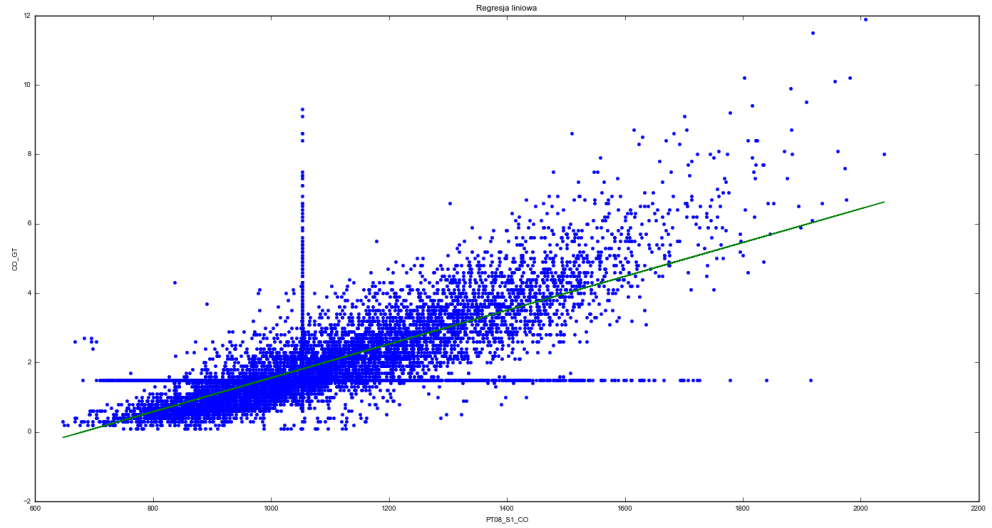


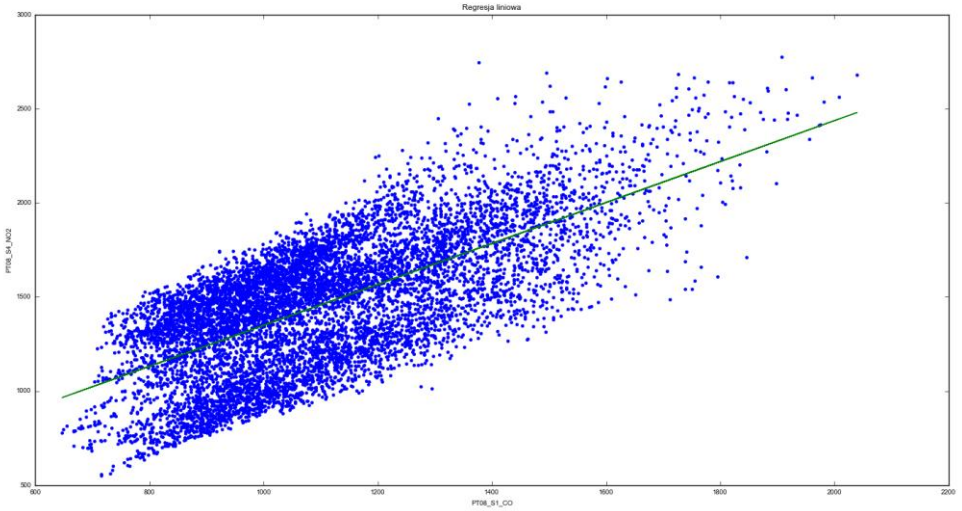
- Wykonałem regresję liniową dla zmiennych skorelowanych. Wykresy poniżej:

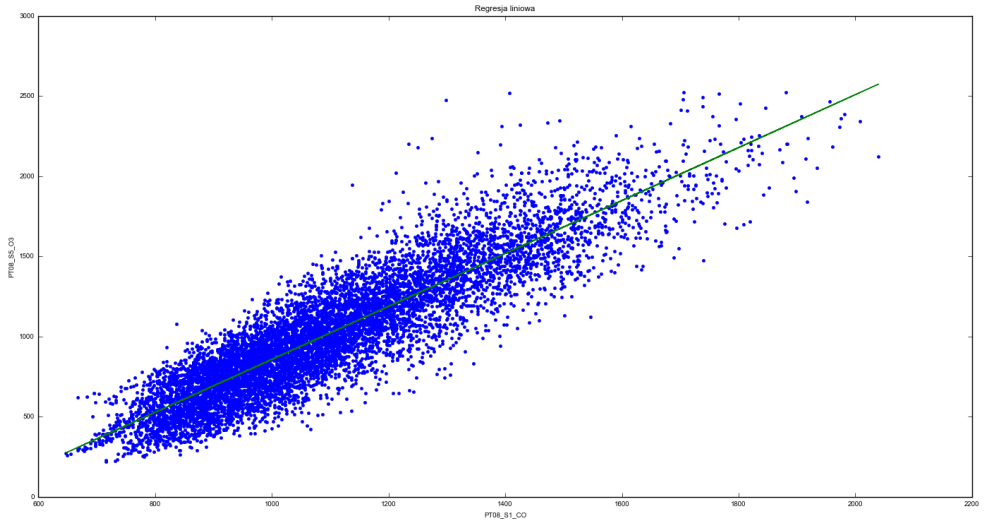
```
def linear_regression_simple(list1, name1, list2, name2):
    #Maksymalizacja okna
    wm = plt.get_current_fig_manager()
    wm.window.state('zoomed')
    # Regresja liniowa dla zmiennych z dużą korelacją
    a, b = np.polyfit(list1, list2, 1)  # Wielomian 1 rzędu - prosta
    yreg = [a * i + b for i in list1]
    #Wykresy
    plt.plot(list1, list2, ".")
    plt.plot(list1, yreg)
    plt.title("Regresja liniowa")
    plt.xlabel("%s" %(name1))
    plt.ylabel("%s" %(name2))
    plt.show()
```

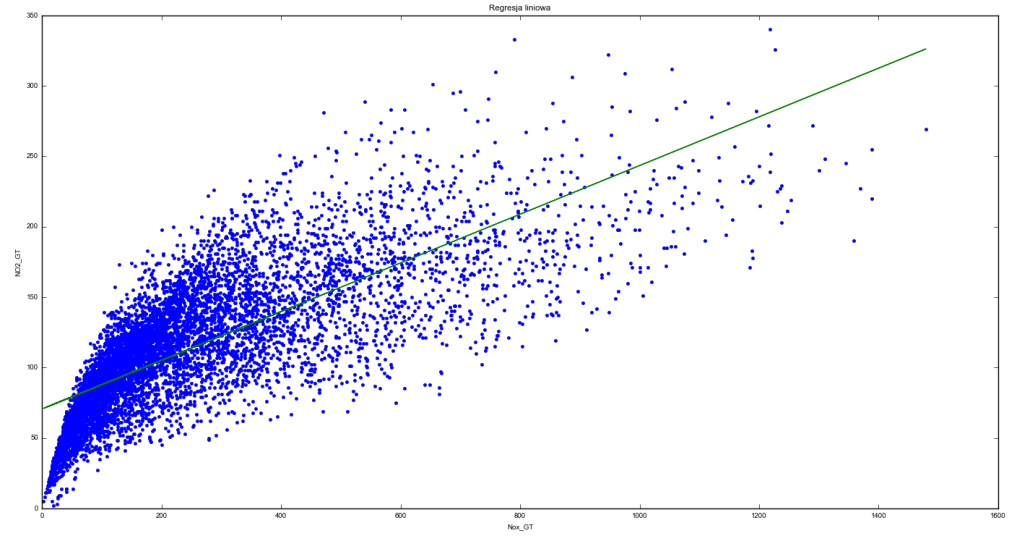

Regresja liniowa
CO_GT
C6H6_GT

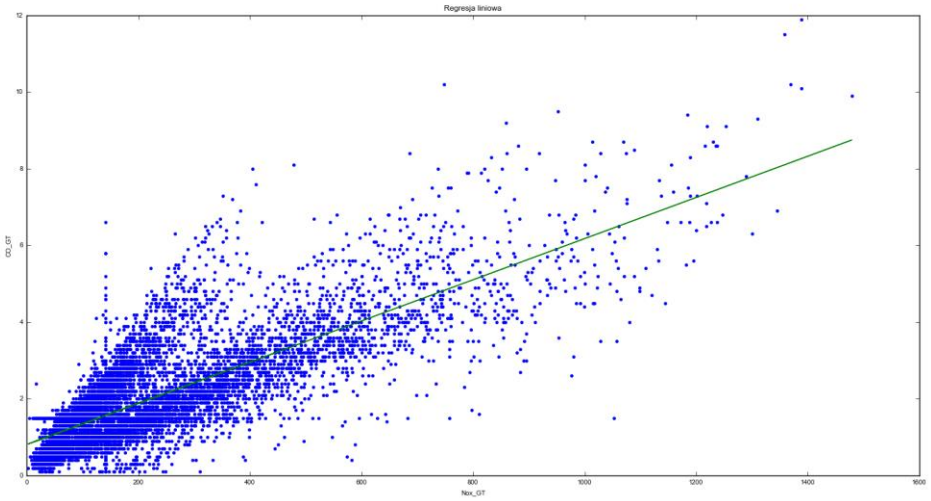

Regresja liniowa
PT08_S5_O3
C6H6_GT

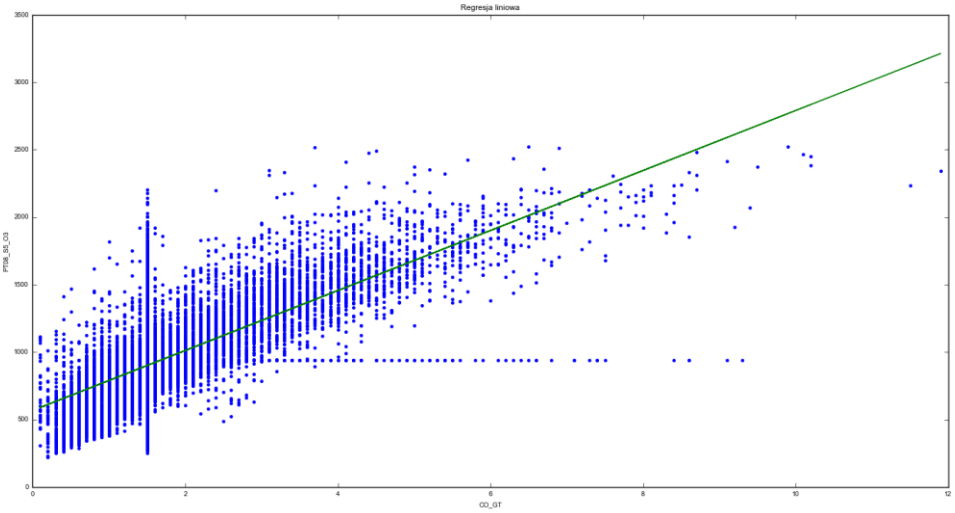

Regresja liniowa
CO_GT
PT08_S1_CO

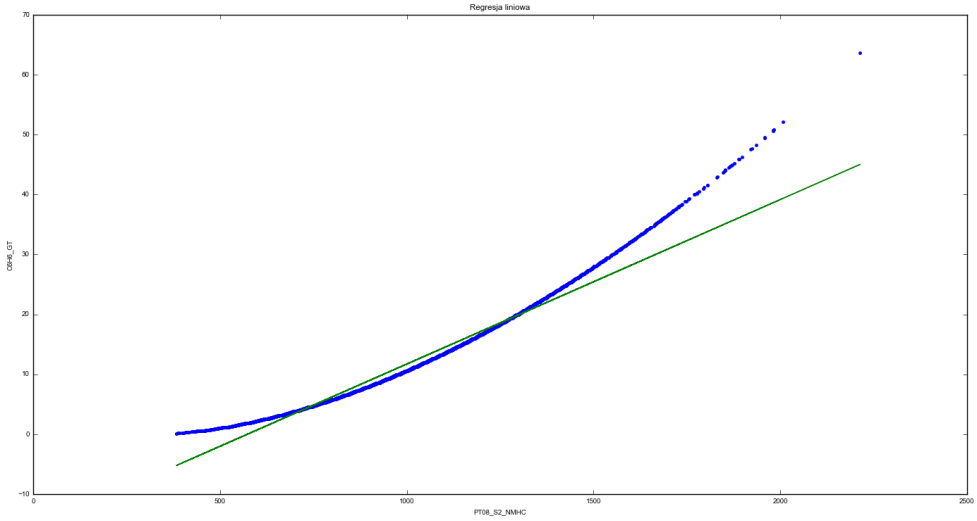

Regresja liniowa
PT08_S4_NO2
PT08_S1_CO

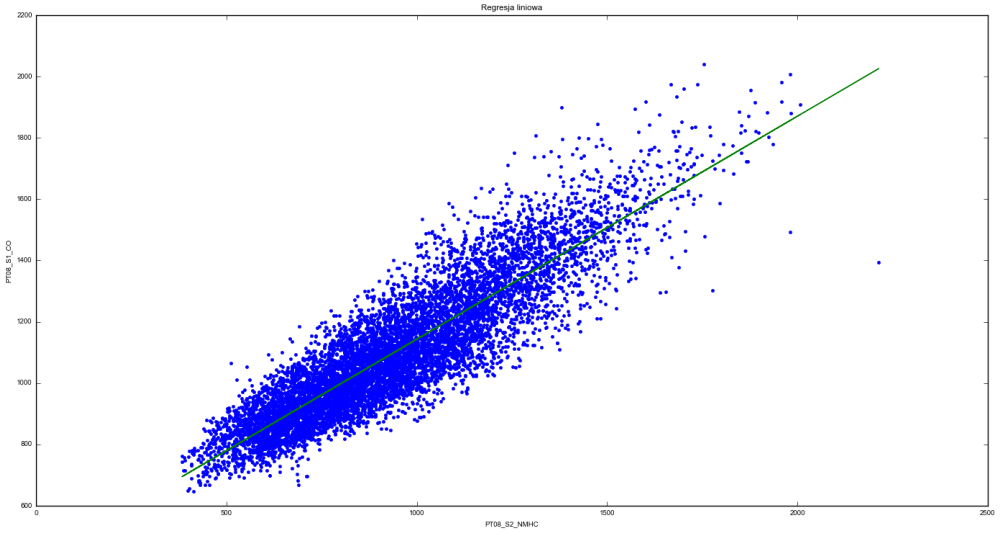

Regresja liniowa
PT08_S5_O3
PT08_S1_CO

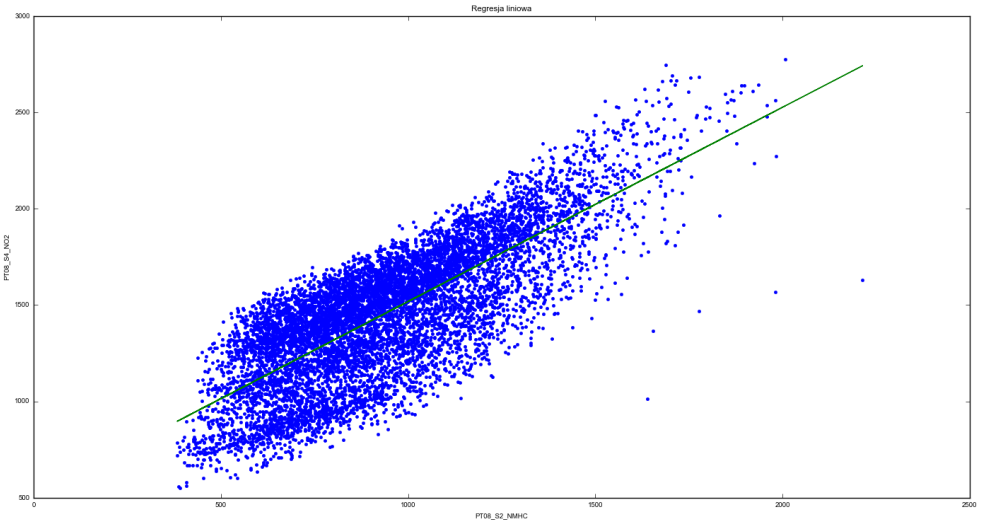

Regresja liniowa
NO2_GT
Nox_GT

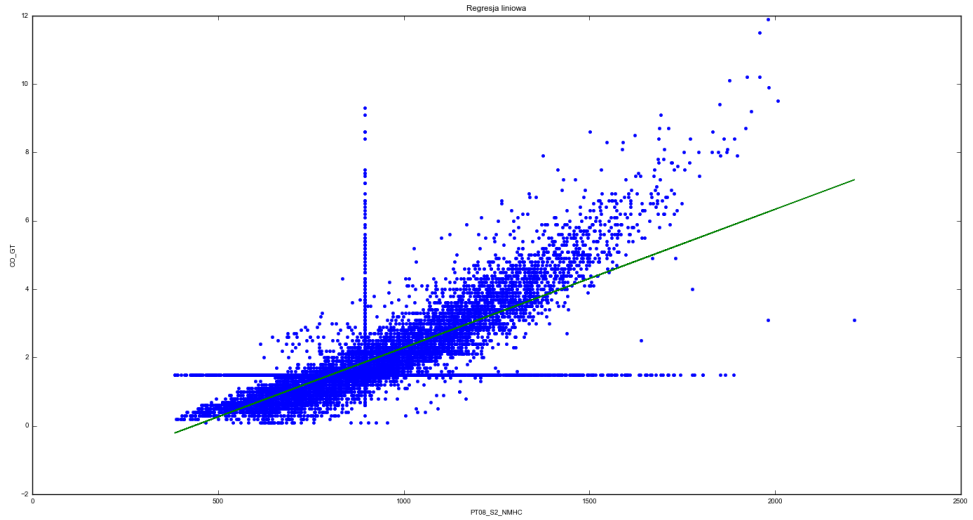

Regresja liniowa
CO_GT
Nox_GT

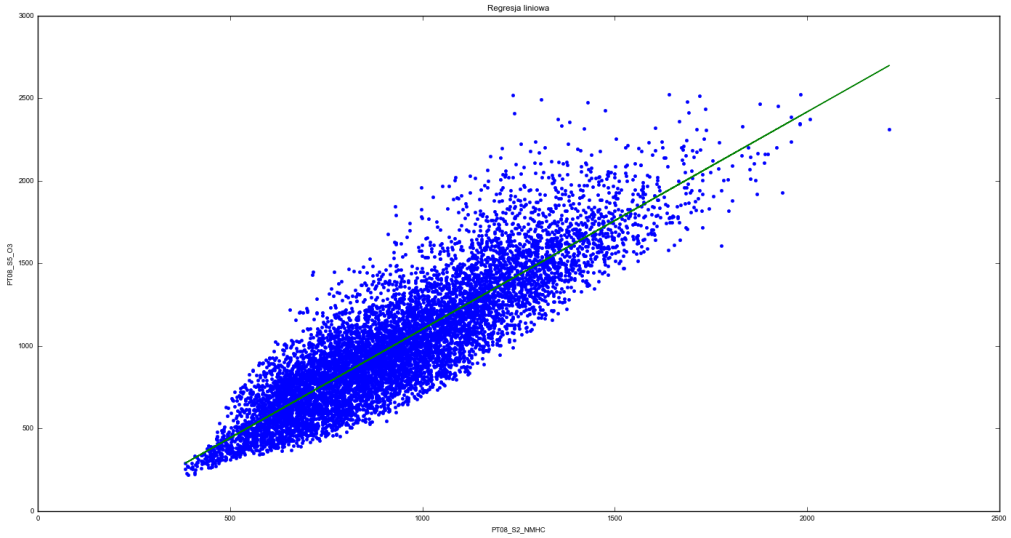

Regresja liniowa
PT08_S5_O3
CO_GT

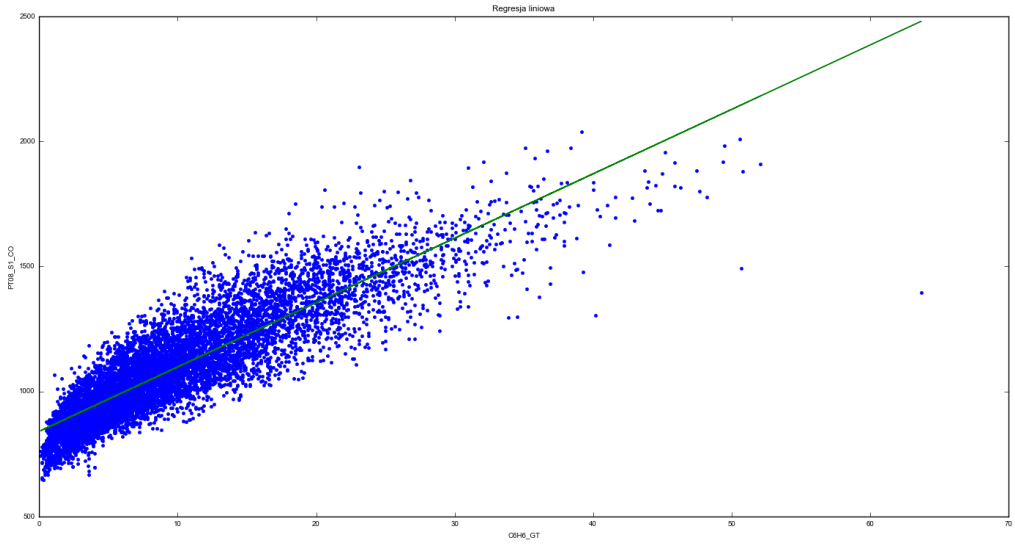

Regresja liniowa
C6H6_GT
PT08_S2_NMHC

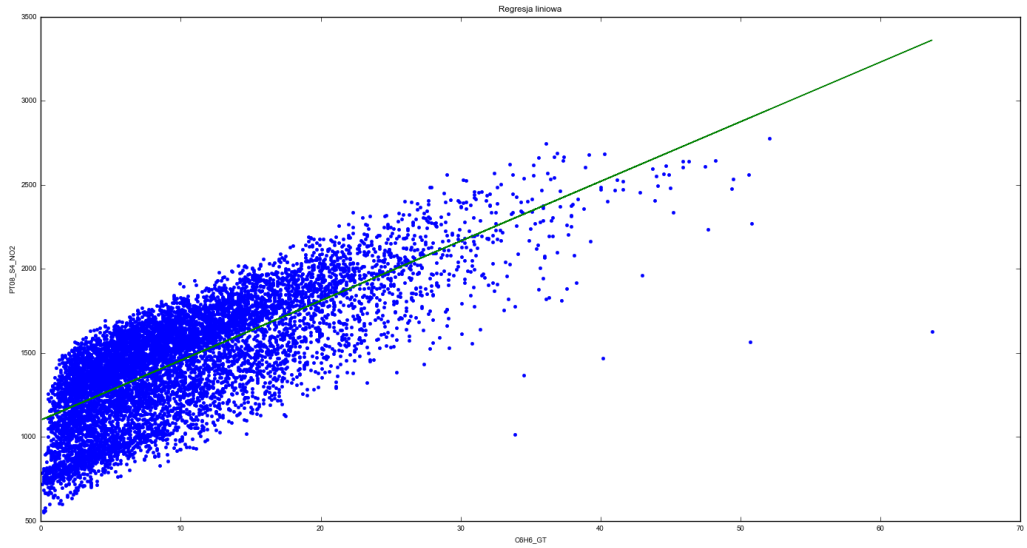

Regresja liniowa
PT08_S1_CO
PT08_S2_NMHC
2200
2000
1800
1600
1400
1200
1000
800
600
0
500
1000
1500
2000
2500


Regresja liniowa
PT08_S4_NO2
PT08_S2_NMHC
3000
2500
2000
1500
1000
500
0
500
1000
1500
2000
2500


Regresja liniowa
CO_GT
PT08_S2_NMHC
12
10
8
6
4
2
0
−2
0
500
1000
1500
2000
2500


Regresja liniowa
PT08_S5_O3
PT08_S2_NMHC
0
500
1000
1500
2000
2500
3000


Regresja liniowa
PT08_S1_CO
C6H6_GT
500
1000
1500
2000
2500
0
10
20
30
40
50
60
70

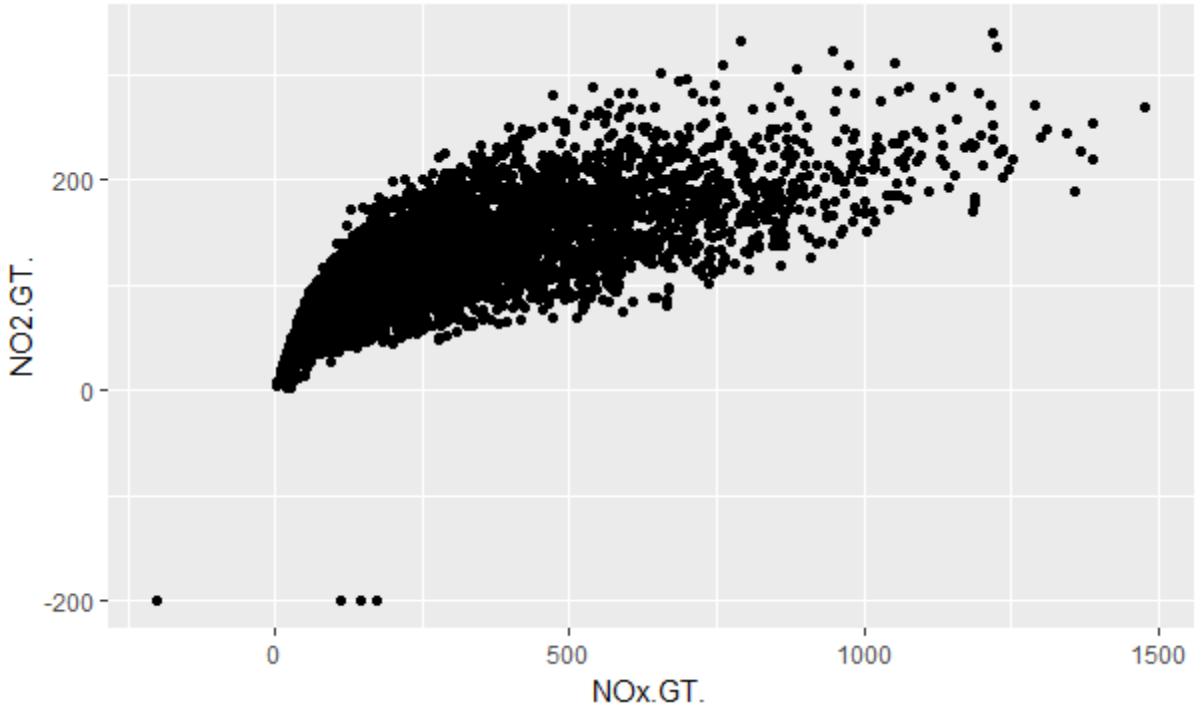

Regresja liniowa
PT08_S4_NO2
C6H6_GT
500
1000
1500
2000
2500
3000
3500
0
10
20
30
40
50
60
70

- Jeśli chodzi o zależności pomiędzy wynikami pomiarów rzeczywistych wielkości, to ujawniły się jedynie 3 korelacje, które możemy uznać za silne:

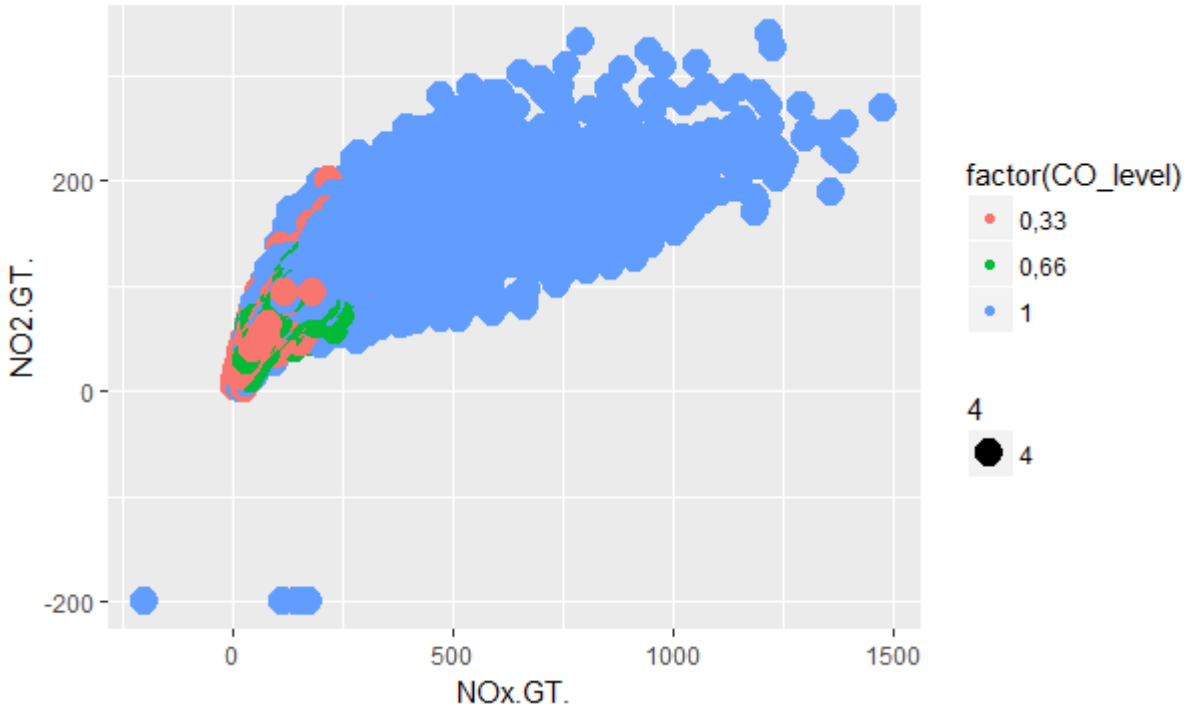
C6H6_GT a CO_GT : 0.802251581903 / Nox_GT a CO_GT : 0.791624000154 /

Nox_GT a NO2_GT : 0.768881291528.

- Aby sprawdzić jak w czasie kształtowały się obie wartości w czasie wykonałem kilka wykresów w R:



- Wprowadziłem także dodatkową zmienną CO_level. CO_level jest to wielkość stężenia CO w powietrzu – w zależności od stosunku danego pomiaru stężenia do mediany stężenia CO, przyjmuje ona 3 poziomy 0,33; 0,66; 1 (intuicyjnie małe, średnie i duże).
Poniższy wykres obrazuje jak do relacji zobrazowanej na wykresie powyżej, ma się poziom CO.



5. Rezultaty, wnioski i ich dyskusja

Struktura projektu:
- Na potrzeby analizy danych został wykonany program w pythonie, skałdający się z 3 plików:
    + config.py – zmienne konfiguracyjne
    + operations.py – metody do operowania na danych
    + data_preparation.py – logika, wywołania metod, wykresy oraz obliczenia pomocnicze
- skrypt w R – ciekawe wykresy w oparciu o biblioteke ggplot2
- plik csv z danymi (AirQualityUCI.csv) znajduje się w katalogu /data_science/proj
- spakowany projekt (data_science.zip) znajduje się w katalogu /data_science/proj
- wygenerowane wykresy zamieściłem w katalogu /data_science/charts
- raport zostanie umieszczony w pliku /data_science/report

Odnosząc się go celów ze wstępu:
+ Na podstawie przeprowadzonej analizy można stwierdzić, że nie obserwuje się wyraźnego wpływu jednych rzeczywistych wartości pomierzonych stężeń substancji na inne (nie można jednak zupełnie ich wykluczyć. Może to sugerować ostatni wykres, jeśliby zbadać pozostałe 2 skorelowane zmienne lub założyć bardziej skomplikowane wpływy.)
+ Wielkości stężeń rzeczywistych i rezultatów uzyskanych z sensorów nie zależą od T, RH i AH.
+ Rezultaty z czujników są dość mocno skorelowane z innymi wielkościami – może być to związane z klasą urządzenia i/lub użytych sensorów.
+ Z ostatniego wykresu (R) można wyciągnąć wniosek, że w otoczeniu zaszła zmiana, jesli chodzi o powiomy stężeń. Obserwuje się spory wzrost CO.
+ Rozkłady wartości podawanych przez sensory prezentowane na histogramach są równomierne, nie wykazują też tendencji do zawyżania/zmniejszania odczytów wraz z czasem. Nie świadczy to o ich zużyciu/rozkaliborowaniu. Korelacje z odczytami innych sensorów i stężaniami, na pomiar których nie są nastawione, może świadczyć o niskiej jakości lub innych niedoskonałościach, np. montażu.
+ Brakujące wartości mogą oznaczać akcje serwisowe, lub awarie.