Adam Dracz

Nr indeksu: 415151

# Eksploracyjna analiza zbioru danych Edukacyjnej Sieci Antysmogowej

Link do folderu z kodem:

Link

# Spis treści

# 1 Wstęp

Celem niniejszego raportu jest eksploracyjna analiza danych pochodzących z Edukacyjnej Sieci Antysmogowej (ESA).

Edukacyjna Sieć Antysmogowa to program informacyjny na rzecz czystego powietrza. W jego ramach w szkołach, które biorą w nim udział, są montowane czujniki, które mierzą parametry dotyczące warunków środowiskowych takie jak: ciśnienie atmosferyczne, stężenie PM10 i PM2.5 w powietrzu, wilgotność względna. Dane, które są zapisywane i udostępniane, są średnimi kroczącymi z ostatnich 60 minut, które zmieniają się co 5 minut.

# 2 Ogólne informacje o zbiorze danych

## 2.1 Źródło zbioru danych

Dane zostały pobrane z serwisu dane.gov.pl, na którym bezpłatnie są udostępniane różne zbiory danych. Mogą one zostać użyte w celach prywatnych, komercyjnych lub naukowych. Zbiory są regularnie aktualizowane.

Dane te dotyczą różnych dziedzin: gospodarki, środowiska, edukacji, technologii i wielu innych.

Zbiorów danych można szukać ze względu na obszar, którego dotyczą lub ze względu na dostawcę danych (np. ministerstwa, urzędy miast).

Dane mogą zostać pobrane przez API w postaci pliku JSON (jedyny rodzaj API) lub można je pobrać bezpośrednio w innych udostępnionych formatach, np. csv, xml, xlsx.

Dane pochodzące z Edukacyjnej Sieci Antysmogowej zostały udostępnione przez Naukową i Akademicką Sieć Komputerową – Państwowy Instytut Badawczy.

## 2.2 Przykłady wykorzystania danych z ESA

Dane z Edukacyjnej Sieci Antysmogowej mogą zostać wykorzystane w różnych rozwiązaniach korzystających z technik uczenia maszynowego.

Przy wykorzystaniu uczenia nadzorowanego mogą zostać użyte np. do predykcji jednej zmiennej na podstawie innych.

W przypadku uczenia nienadzorowanego można przeprowadzić klasteryzację zbioru czy też poszukiwanie wartości odstających (które np. ukażą ekstremalne i szkodliwe dla zdrowia wartości w jakimś miejscu w Polsce).

W danych są też udostępnione kolumny ze współrzędnymi geograficznymi miejsc, w których wykonywano pomiary, a więc na zbiór można też patrzeć pod kątem analiz przestrzennych.

## 2.3 Zmienne występujące w zbiorze

Na wstępnym etapie analizy można przyjrzeć się bliżej zbiorowi danych i spojrzeć na kilka pierwszych wierszy:

| Timestamp | Name | Street | Post Code | City | Longitude | Latitude | Humidity Avg | Pressure Avg | Temperature Avg | PM10 Avg | PM25 Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2025-04-28 11:21:09 | Szkoła Podstawowa im. Mariana Falskiego w Kras... | ul. Szkolna | 63-522 | Kraszewice | 18.224030 | 51.515630 | 95.120000 | 1015.110000 | 8.610000 | 32.860000 | 20.200000 |
| 2025-04-28 11:21:09 | Szkoła Podstawowa we Wrząsowicach | ul. Szkolna | 32-040 | Wrząsowice | 19.942820 | 49.961030 | 34.200000 | 988.925000 | 18.775000 | 6.175000 | 5.525000 |
| 2025-04-28 11:21:09 | Publiczna Szkoła Podstawowa nr 2 im. Kazimierz... | ul. Wawrzyńca Świerzego | 47-100 | Strzelce Opolskie | 18.314889 | 50.503431 | 60.000000 | 1043.166667 | 12.266667 | 1.000000 | 0.490000 |
| 2025-04-28 11:21:09 | Zespół Szkół nr 1 w Pszczynie | ul. Kazimierza Wielkiego | 43-200 | Pszczyna | 18.945706 | 49.965883 | 79.666667 | 1005.566667 | 11.166667 | 10.800000 | 6.586667 |
| 2025-04-28 11:21:09 | Zespół Szkół im. Powstańców Wielkopolskich w J... | Szkolna | 63-421 | Janków Przygodzki | 17.788907 | 51.596172 | 57.166667 | 1012.300000 | 14.400000 | 15.700000 | 8.893333 |
| 2025-04-28 11:21:09 | Szkoła Podstawowa nr 7 im. Żołnierzy Września ... | ul. Męczenników Oświęcimskich | 43-229 | Ćwiklice | 18.989839 | 49.971937 | 37.233333 | 996.366667 | 23.666667 | 8.066667 | 4.590000 |
| 2025-04-28 11:21:09 | Szkoła Podstawowa nr 12 w Studzionce | ul. Jordana | 43-245 | Studzionka | 18.774985 | 49.960356 | 99.333333 | 1004.266667 | 11.500000 | 14.800000 | 8.833333 |
| 2025-04-28 11:21:09 | Zespół Szkolno-Przedszkolny w Piasku | Szkolna | 43-211 | Piasek | 18.946340 | 50.009550 | 53.533333 | 1005.700000 | 11.700000 | 11.600000 | 7.260000 |
| 2025-04-28 11:21:09 | Zespół Szkolno-Przedszkolny w Łące | Fitelberga | 43-241 | Łąka | 18.906757 | 49.958244 | 51.166667 | 1003.033333 | 12.666667 | 14.066667 | 8.236667 |

*Fig. 2.1. Pierwsze 9 wierszy analizowanego zbioru danych*

Kolumny są indeksowane przez znacznik czasu (Timestamp), który informuje o dacie i godzinie wykonania pomiaru. W przypadku badanego zbioru wszystkie pomiary zostały pobrane w tym samym momencie – 28 kwietnia 2025 r. o godzinie 11:29:09.

Zmienne występujące w zbiorze:

- Name – nazwa szkoły, w której zamontowano czujnik,
- Street – ulica, przy której jest położona szkoła,
- Post Code – kod pocztowy dla danej szkoły,
- City – miasto (miejscowość), w której znajduje się szkoła,
- Longitude – długość geograficzna,
- Latitude – szerokość geograficzna,
- Humidity Avg – średnia wilgotność powietrza (w %),
- Pressure Avg – średnie ciśnienie atmosferyczne (w hPa),
- Temperature Avg – średnia temperatura (w °C),
- PM10 Avg – średnie stężenie PM10 (w $\mu g/m^3$),
- PM25 Avg – średnie stężenie PM2.5 (w $\mu g/m^3$).

Zbiór danych zawiera 1727 rekordów.

# 3 Analiza zbioru danych

## 3.1 Wartości brakujące

Dla zmiennych występujących w zbiorze procentowy udział wartości brakujących dla poszczególnych zmiennych wynosi odpowiednio:

| | |
|---|---|
| Name | 0.000000 |
| Street | 18.297626 |
| Post Code | 0.000000 |
| City | 0.000000 |
| Longitude | 0.000000 |
| Latitude | 0.000000 |
| Humidity Avg | 0.115808 |
| Pressure Avg | 0.115808 |
| Temperature Avg | 0.115808 |
| PM10 Avg | 0.000000 |
| PM25 Avg | 0.000000 |

*Fig. 3.1 Procentowy udział wartości brakujących dla zmiennych*

Dla wilgotności, ciśnienia i temperatury jest to ok. 0.12%, więc w związku z małym udziałem wartości te mogą zostać uznane za nieistotne i usunięte.

Widoczny udział mają wartości brakujące dla zmiennej Street.

| Timestamp | Name | Street | Post Code | City | Longitude | Latitude | Humidity Avg | Pressure Avg | Temperature Avg | PM10 Avg | PM25 Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2025-04-28 11:21:09 | Zespół Szkół w Piotrowie | NaN | 62-814 | Piotrów | 18.054850 | 51.815700 | 27.633333 | 1013.666667 | 23.433333 | 7.200000 | 4.070000 |
| 2025-04-28 11:21:09 | Szkoła Podstawowa w Czarnymlesie | NaN | 63-421 | Czarnylas | 17.762604 | 51.508366 | 73.800000 | 1008.866667 | 13.466667 | 16.200000 | 9.376667 |
| 2025-04-28 11:21:09 | Szkoła Podstawowa im. Bohaterów Powstań Śląski... | NaN | 48-364 | Kałków | 17.185241 | 50.404200 | 96.340000 | 1010.980000 | 8.550000 | 31.050000 | 18.640000 |
| 2025-04-28 11:21:09 | Szkoła Podstawowa w Maciejowicach | NaN | 48-385 | Maciejowice | 17.134970 | 50.499800 | 38.433333 | 991.800000 | 21.433333 | 4.466667 | 2.390000 |
| 2025-04-28 11:21:09 | Szkoła Podstawowa w Szkodnej | NaN | 39-126 | Szkodna | 21.647559 | 49.984184 | 39.955556 | 989.711111 | 12.766667 | 3.266667 | 3.255556 |

*Fig. 3.2 Pięć przykładowych wierszy, gdzie brakuje wartości dla zmiennej Street*

Po zbadaniu okazuje się, że wartości brakujące dla Street występują dla niewielkich miejscowości, w których nie ma podziału na ulice.

W zbiorze występują jednak również miejscowości, w których nie ma ulic, a jako wartość w kolumnie Street jest podana nazwa miejscowości.

| Timestamp | Name | Street | Post Code | City | Longitude | Latitude | Humidity Avg | Pressure Avg | Temperature Avg | PM10 Avg | PM25 Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2025-04-28 11:21:09 | Szkoła Podstawowa im. Arkadego Fiedlera w Rasz... | Pogrzybów | 63-440 | Pogrzybów | 17.723099 | 51.713966 | 48.70 | 1018.200000 | 22.900000 | 6.000000 | 3.500000 |
| 2025-04-28 11:21:09 | Szkoła Podstawowa im. Orła Białego w Sokołowicach | Sokołowice | 56-400 | Sokołowice | 17.446900 | 51.245000 | 77.98 | 1009.180000 | 9.680000 | 10.980000 | 6.150000 |
| 2025-04-28 11:21:09 | Szkoła Podstawowa w Ligocie Polskiej | Ligota Polska | 56-400 | Ligota Polska | 17.542356 | 51.239942 | 99.90 | 1005.966667 | 15.833333 | 9.000000 | 4.830000 |
| 2025-04-28 11:21:09 | Szkoła Podstawowa im. Unicef w Ligocie Małej | Ligota Mała | 56-400 | Ligota Mała | 17.365476 | 51.142467 | 52.70 | 984.200000 | 11.600000 | 15.766667 | 8.633333 |
| 2025-04-28 11:21:09 | Szkoła Podstawowa we Wszechświętem | Wszechświęte | 56-400 | Wszechświęte | 17.484968 | 51.191751 | 99.90 | 1008.266667 | 13.033333 | 6.833333 | 3.946667 |

*Fig. 3.3 Przykładowe wiersze, w których nazwa ulicy jest taka jak nazwa miejscowości*

Jest to niekonsekwencja przy zapisie danych w zbiorze. Wobec tego wartości NA w kolumnie mogą zostać uzupełnione przez wpisanie w to miejsce nazwy miejscowości.

W tym wypadku jest to jednak zabieg kosmetyczny, niemający wpływu na dalszą analizę.

## 3.2 Kardynalność zmiennych nominalnych

Liczebność etykiet dla zmiennych nominalnych wynosi odpowiednio:

```
Ilość etykiet dla zmiennej Name: 1722
Ilość etykiet dla zmiennej Post Code: 1139
Ilość etykiet dla zmiennej City: 1203
Ilość etykiet dla zmiennej Street: 1149
```

*Fig. 3.4 Kardynalność zmiennych nominalnych*

Są to duże wartości w stosunku do wielkości zbioru danych. Ze względu na charakter analiz i skupienia w dalszej części na cechach numerycznych, nie są to jednak istotne informacje. W związku z tym nie zachodzi potrzeba kodowania zmiennych nominalnych.

## 3.3 Miasta, w których wykonano pomiary

W celu wizualizacji przestrzennej na wykresie punktowym można przedstawić miasta, w których wykonano pomiary.

Prowadzona analiza nie ma charakteru analizy przestrzennej, a ogólną wizualizację w celu detekcji potencjalnych anomalii i powodów, dla których dane mogą nie być reprezentatywne, więc punkty nie są nanoszone na mapę.

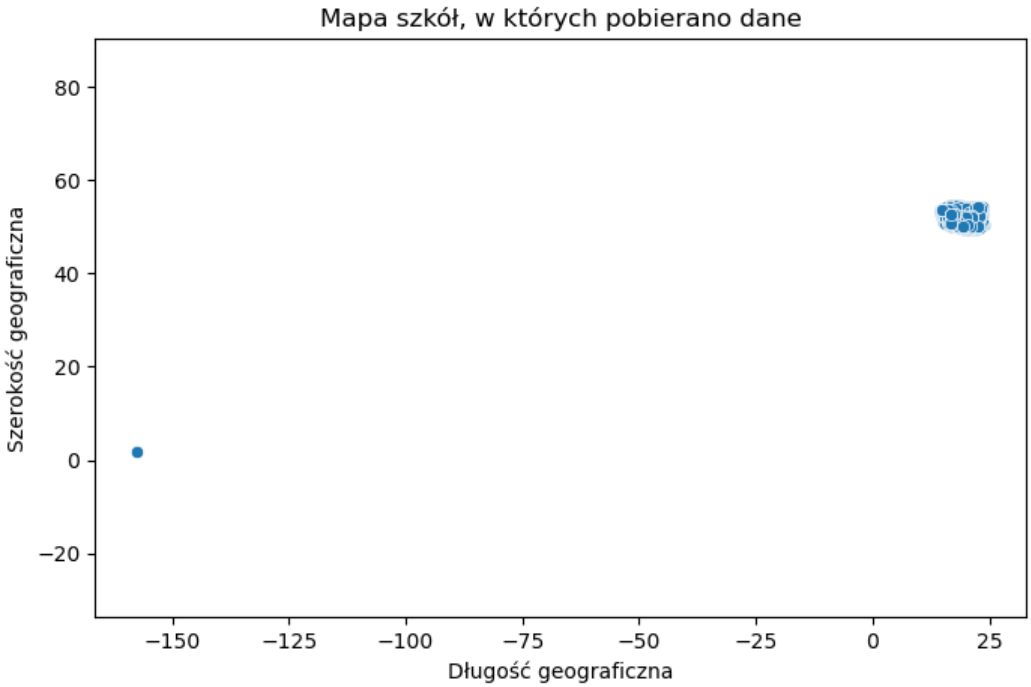


*Fig. 3.5 Mapa rozłożenia przestrzennego szkół*

Można zauważyć jeden punkt bardzo widocznie odstający.

| Timestamp | Name | Street | Post Code | City | Longitude | Latitude | Humidity Avg | Pressure Avg | Temperature Avg | PM10 Avg | PM25 Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2025-04-28 11:21:09 | Test Szkoła | ul Testowa | 00-000 | Testcity | -157.529698 | 1.858685 | 89.69 | 992.47 | 7.76 | 50.17 | 30.32 |

*Fig. 3.6 Rekord dla punktu o odstających współrzędnych geograficznych*

Po sprawdzeniu okazuje się, że jest to wartość testowa, sztuczna. Można ją zatem usunąć ze zbioru.

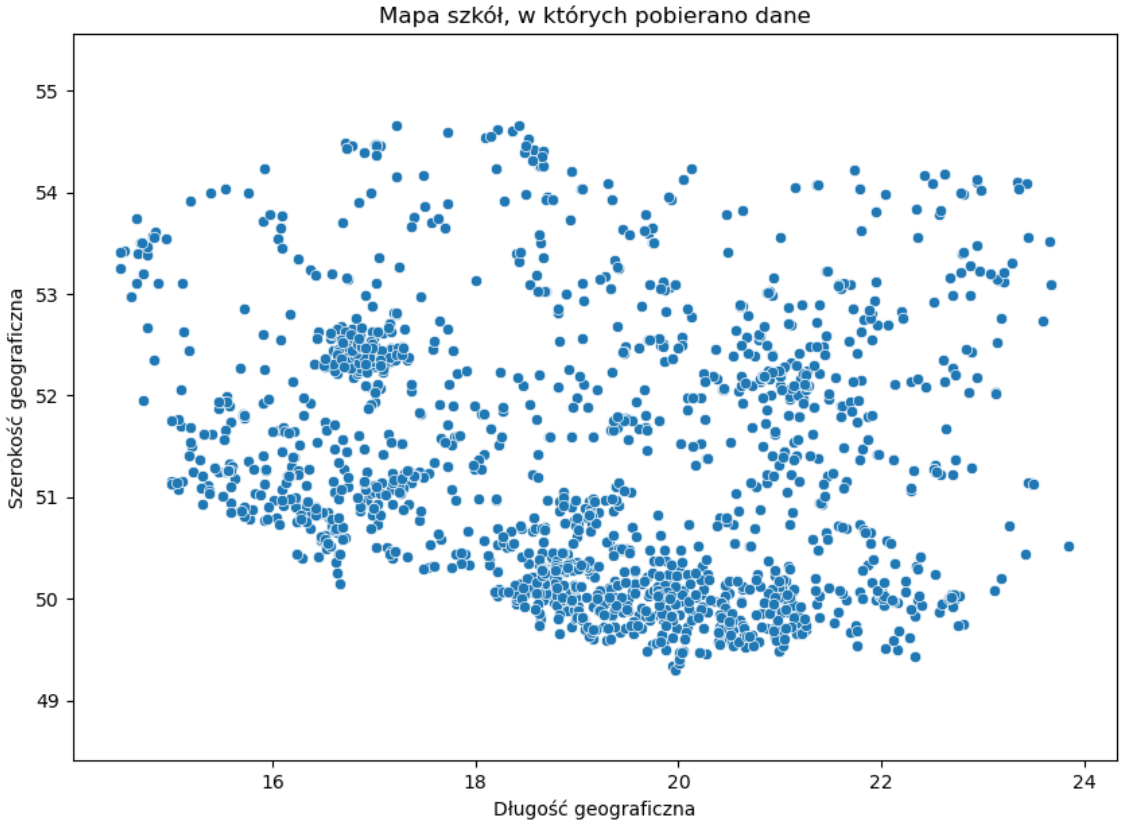


*Fig. 3.7 Mapa szkół po usunięciu wartości odstającej*

Można zauważyć, że w przybliżeniu dane pokrywają cały obszar Polski. Największe skupiska można dostrzec na południu i zachodzie kraju. Oznacza to, że dane mogą być w pewien sposób niereprezentatywne (dane ze skupisk będą do siebie podobne i prawdopodobnie inne od pochodzących z innych rejonów). Wada to może być jednak zrównoważona przez fakt, że pozostałe dane pochodzą z miejsc rozproszonych po całym kraju.

## 3.4 Analiza pozostałych zmiennych numerycznych

### 3.4.1 Badanie wartości odstających

Płynnie można przejść do analizy pozostałych zmiennych numerycznych w zbiorze: wilgotność, ciśnienie, temperatura, stężenie PM10 i PM2.5.

| | Humidity Avg | Pressure Avg | Temperature Avg | PM10 Avg | PM25 Avg |
|---|---|---|---|---|---|
| **count** | 1724.000000 | 1724.000000 | 1724.000000 | 1724.000000 | 1724.000000 |
| **mean** | 45.892248 | 1003.547592 | 14.511042 | 9.540906 | 7.893529 |
| **std** | 14.905450 | 14.198035 | 3.256907 | 10.575737 | 8.351864 |
| **min** | 0.000000 | 941.675000 | -40.000000 | 0.000000 | 0.000000 |
| **25%** | 35.947917 | 994.739583 | 12.508050 | 5.272917 | 4.566667 |
| **50%** | 42.864800 | 1005.045833 | 13.868333 | 7.430952 | 6.416667 |
| **75%** | 51.783333 | 1014.920417 | 16.495833 | 11.037500 | 9.300000 |
| **max** | 100.000000 | 1097.800000 | 31.800000 | 245.666667 | 241.000000 |

*Fig. 3.8 Statystyki opisowe dla zmiennych numerycznych*

Już na tym etapie można zauważyć kilka nietypowych wartości: temperatura minimalna równa -40, wilgotność równa 0 i 100, czy wartości stężenia PM10 i PM2.5 powyżej 240.

Pomocne jest również przedstawienie danych na wykresach pudełkowych, które m.in. wskażą wartości odstające:

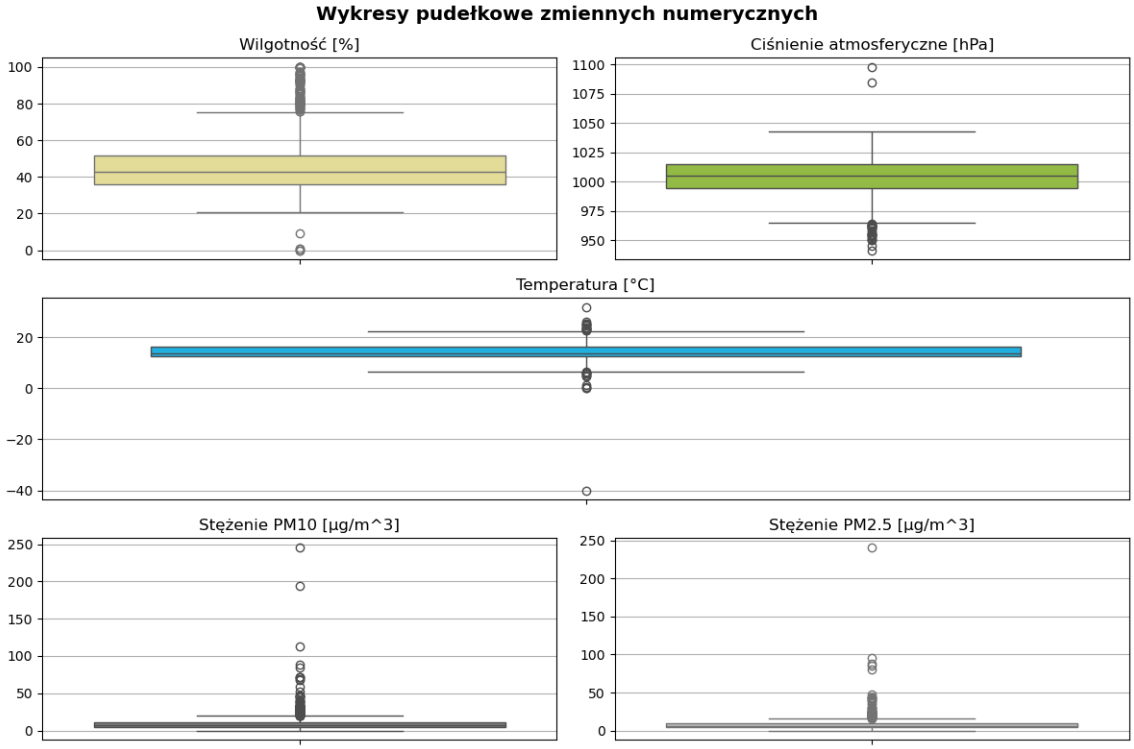


*Fig. 3.9 Wykresy pudełkowe dla zmiennych numerycznych*

Na powyższym wykresie można zauważyć wartości odstające dla danych zmiennych. Niektóre są ewidentnymi błędami i są mało realne (np. temperatura równa -40, wilgotność równa 0 lub ciśnienie równe 1100).

Pozostałe (np. PM10 do wartości 150) są jednak realne i nie ma podstaw, aby uważać je za błędne. Nie należy ich zatem odrzucać z danych, ponieważ też niosą potencjalnie ważną informację.

Z danych odrzucone zostały następujące wartości odstające (na podstawie analizy danych pogodowych i ustaleniu przybliżonych zakresów wartości typowych):

- wilgotność powyżej 90% i poniżej 7%
- ciśnienie atmosferyczne powyżej 1050 hPa i poniżej 960 hPa
- temperatura poniżej 5 °C i powyżej 26 °C
- stężenie PM10 powyżej 100 µg/m$^3$
- stężenie PM2.5 powyżej 79 µg/m$^3$

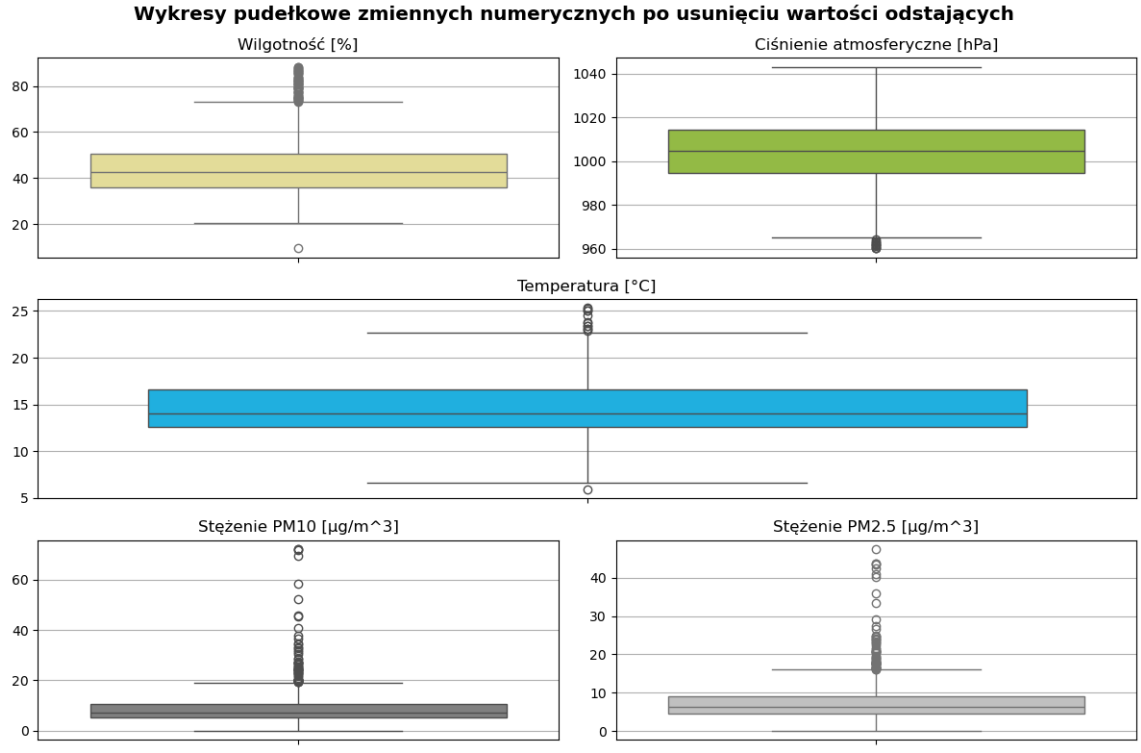


*Fig. 3.10 Wykresy pudełkowe dla zmiennych numerycznych po usunięciu wartości odstających*

### 3.4.2 Badanie rozkładu zmiennych

Dane przedstawić można również za pomocą histogramu, aby zbadać ich rozkład.

Do histogramu dodane zostały też jądrowe estymatory gęstości, które pozwolą łatwiej ocenić kształt rozkładu.

Histogram został narysowany w wersji znormalizowanej (tak, że suma pól poszczególnych słupków sumuje się do 1), aby przybliżać funkcję gęstości prawdopodobieństwa.

Zaznaczone zostały również średnia i mediana dla danej zmiennej.

Ważną uwaga: wykresy dla stężenia PM10 i PM2.5 zostały ograniczone tylko do wartości 35 na osi poziomej ze względu na niewielkie wartości występujące powyżej tej wartości, które zaburzają czytelność wykresu.

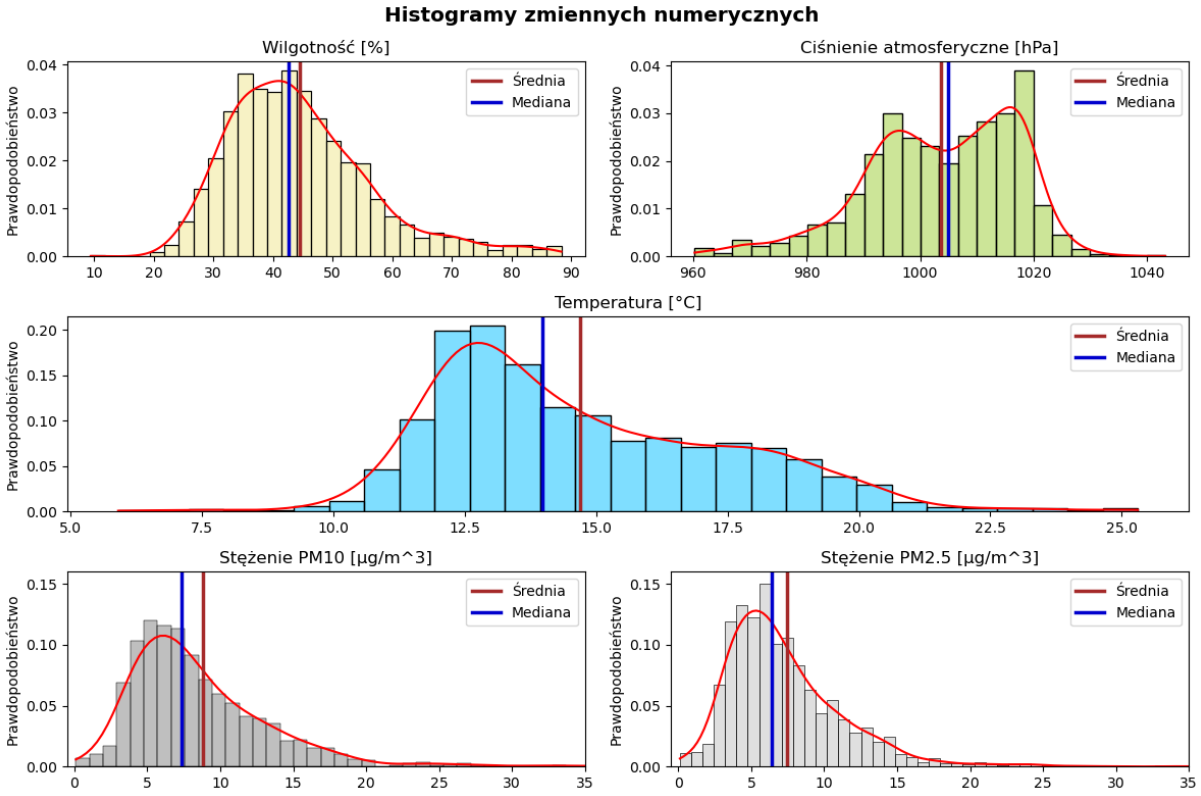


*Fig. 3.12 Histogramy z jądrowymi estymatorami gęstości dla zmiennych numerycznych*

Możemy zauważyć, że rozkłady zmiennych są lewoskośne i jednomodalne.

Wyjątkiem jest ciśnienie atmosferyczne, gdzie możemy dostrzec bimodalność (w okolicach 995 hPa i 1115 hPa).

### 3.4.3 Badanie zależności między zmiennymi

Poniżej przedstawiono również kilka wykresów punktowych, w których przedstawiono związki między niektórymi ze zmiennych.

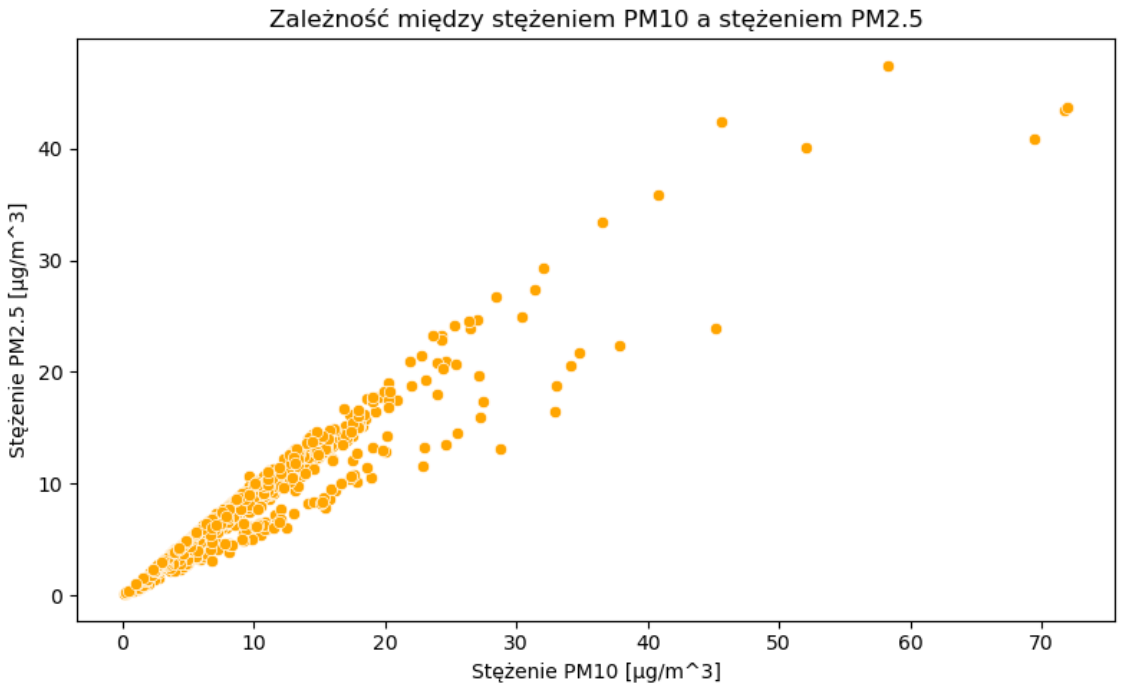


*Fig. 3.13 Zależność między stężeniem PM10 a stężeniem PM2.5*

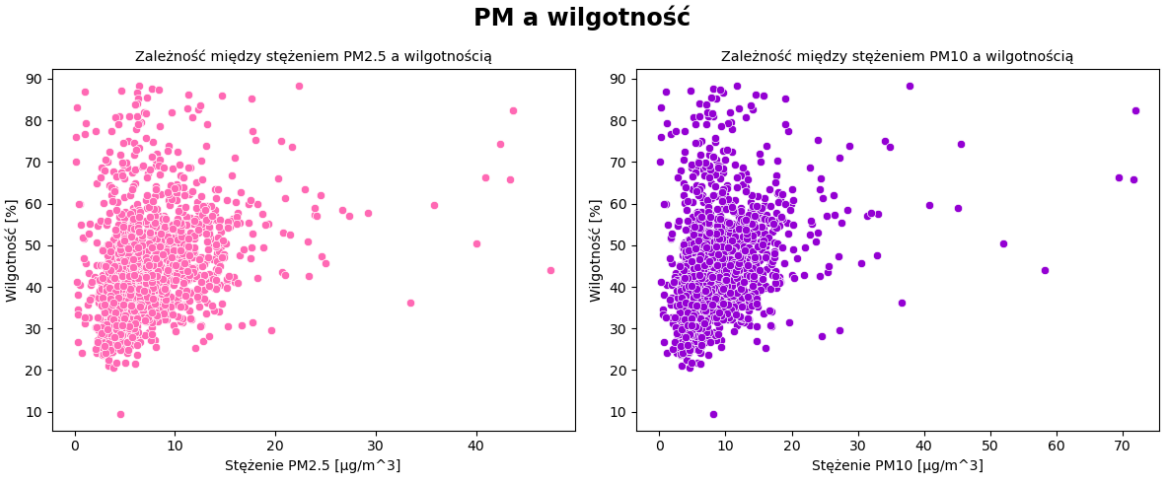


*Fig. 3.14 Zależności między stężeniem PM2.5 oraz PM10 a wilgotnością*

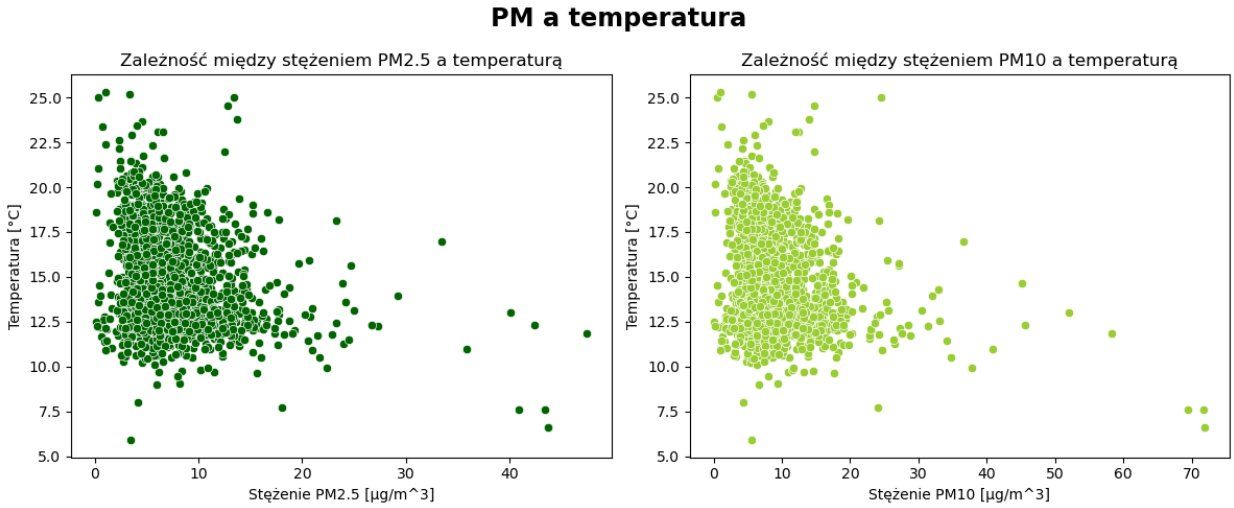


*Fig. 3.15 Zależność między stężeniem PM2.5 i PM10 a temperaturą*

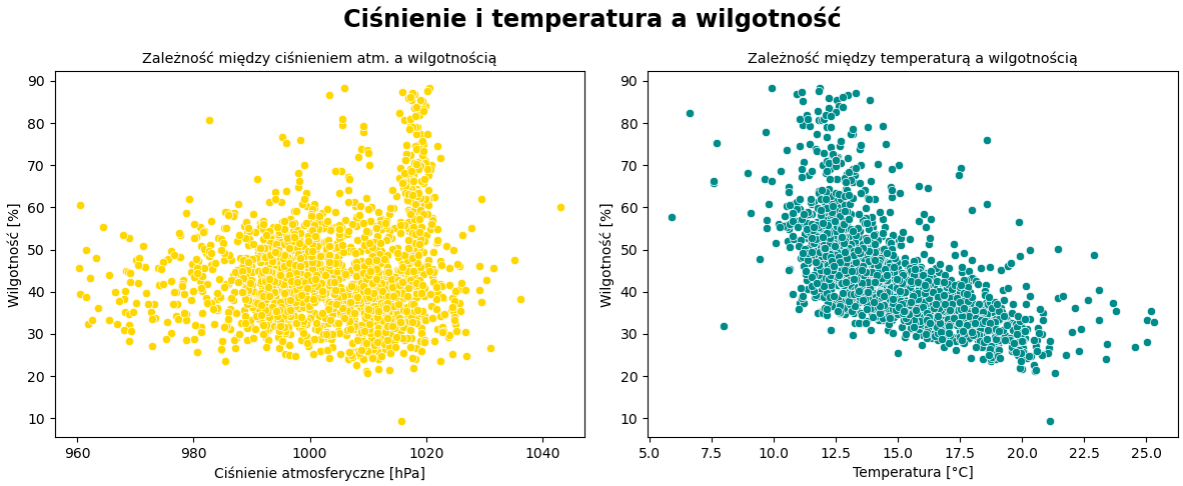


*Fig. 3.16 Zależność między ciśnieniem i temperaturą a wilgotnością*

Na podstawie powyższych wykresów można wyciągnąć wnioski:

- istnieje bardzo silna zależność między stężeniem PM2.5 i PM10 (PM2.5 zawiera się w PM10),
- istnieją też widoczne zależności między stężeniami PM2.5 i PM10 a wilgotnością i temperaturą, ale ze względu na nadreprezentację wartości niskich stężeń pyłów informacja ta jest niepełna,
- trudno zauważyć jednoznaczną zależność między ciśnieniem a wilgotnością, ale zależność między wilgotnością a temperaturą jest już silna.

# 4 Wnioski

Dane pochodzące z Edukacyjnej Sieci Antysmogowej nie są zbiorem o wielu zmiennych, ale ze względu na liczbę rekordów mogą służyć do analizy, predykcji czy też klasteryzacji.

Przykładem (przy zastosowaniu technik nadzorowanego uczenia maszynowego) może być predykcja wilgotności, dzięki której można prognozować opady. Może ona się odbyć przy użyciu zmiennych Pressure, Temperature i PM10. Jest to związane z tym, że warunki atmosferyczne mają wpływ na wilgotność powietrza. PM10 dodatkowo pomoże zbadać wpływ zanieczyszczeń na badany parametr. PM2.5 nie powinno być uwzględniane, ponieważ zawiera się w PM10, więc użycie go przy budowie modelu byłoby niepotrzebnym duplikowaniem informacji.

Badany zbiór nie jest jednak reprezentatywny. Widać to po mapie rozmieszczenia szkół, gdzie widać rejony, z których pobieranych jest więcej pomiarów, a dla warunków klimatycznych oznacza to, że zbiór będzie zawierał dużo danych podobnych do siebie, a wartości wyraźnie różniące się od większości, mające znaczącą informację, ale występujące w mniejszej ilości będą miały mniejszy udział w wymodelowaniu zjawiska.

Reprezentatywność mogłaby być lepsza, gdyby zbiór poddano analizom przestrzennym i udałoby się uśrednić dane pochodzące z danych zgrupowań (podobnych obszarów). Dodatkowo należałoby dysponować danymi pochodzącymi z różnym momentów w czasie i związanymi z różnymi warunkami atmosferycznymi, a także dotyczącymi zanieczyszczeń powietrza.

# 5 Źródła

http://meteo2.ftj.agh.edu.pl/meteo/archiwalneDaneMeteo
Dostęp: 09.05.2025 godz. 14.59

https://powietrze.gios.gov.pl/pjp/current
Dostęp: 09.05.2025 godz. 15.20

https://en.wikipedia.org/wiki/Humidity
Dostęp: 08.05.2025 godz. 13.50