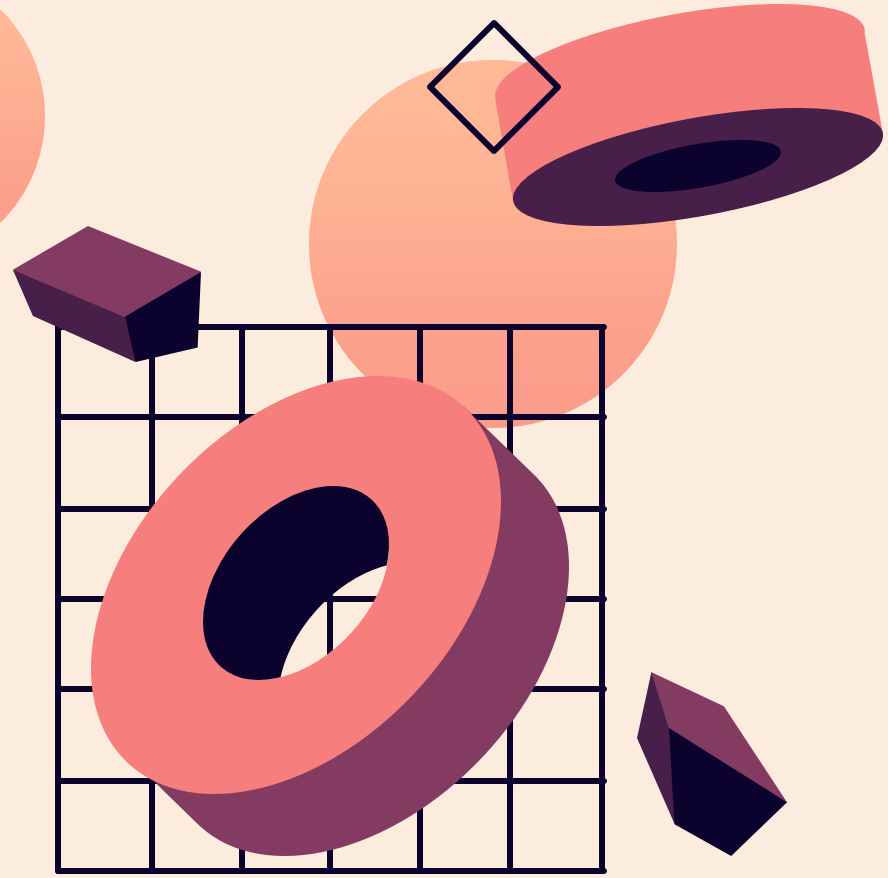
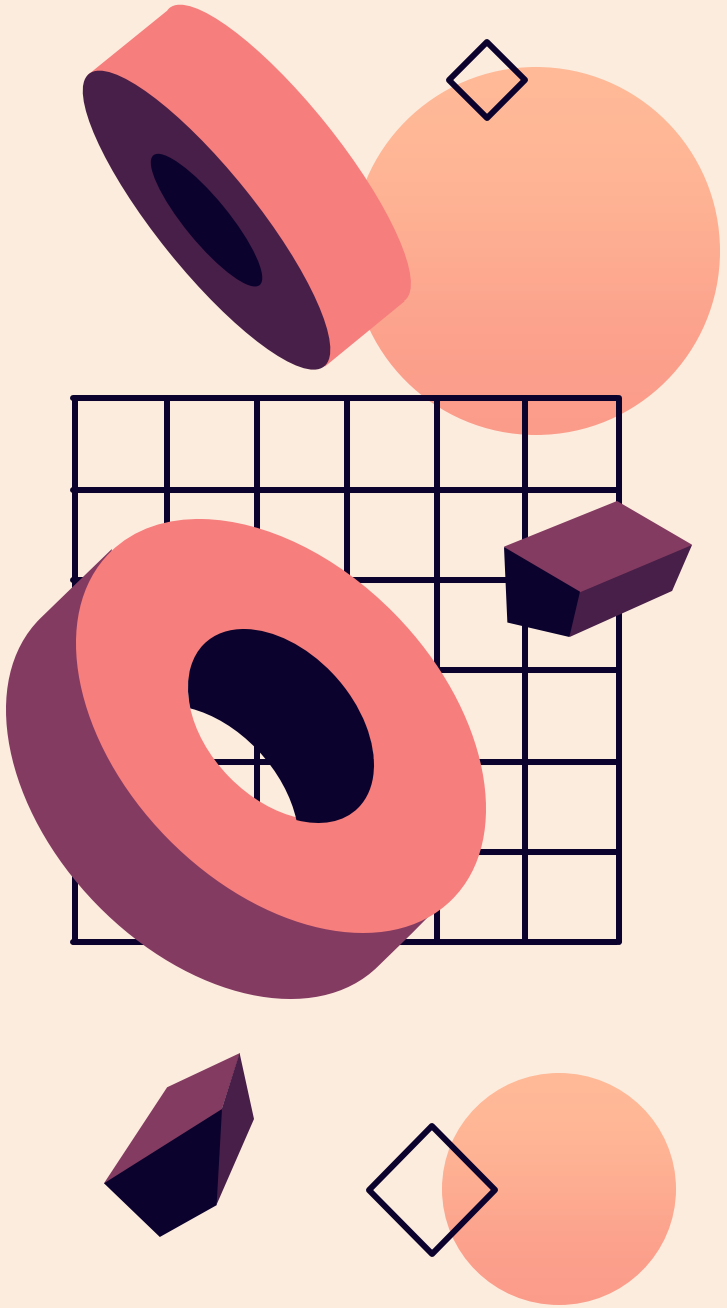
Statystyczna analiza danych



Skład grupy:
Weronika Belniak, 249048
Jakub Kudryk, 259434
Bartosz Matysiak, 252757
Mateusz Śmigielski, 260457



Spis treści

1. Nasz zbiór danych

- Dane o zanieczyszczeniach powietrza
- Pochodzą z projektu Edukacyjnej Sieci Antysmogowej (ESA)
- Sieć czujników mierzących poziomy zanieczyszczeń PM2.5 oraz PM10 jest zainstalowana na szkołach w całej Polsce
- Czujniki są zainstalowane na ponad dwóch tysiącach szkół w całym kraju

1. Nasz zbiór danych

- Plik POSE-205.xlsx
- Odczyty pomiarów z danego dnia dla danej placówki

date	rspo	city	name	dew_point_avg	humidity_avg	pm10_avg
01.01.2022	8723	WEJHEROWO	SZKOŁA PODSTAWOWA NR 8 IM. MARTYROLOGII PIAŚNICY W WEJHEROWIE	0	67.7657070707071	33.1169696969697
01.01.2022	11470	GDYNIA	SZKOŁA PODSTAWOWA NR 48 W GDYNI	0	100	9.4109435261708
01.01.2022	15250	PRZYBOROWO	SZKOŁA PODSTAWOWA IM. ARKADEGO FIEDLERA W PRZYBOROWIE		92.7208333333333	5.8875
01.01.2022	22227	POBIEDZISKA LETNISKO	ZESPÓŁ SZKÓŁ IM. KONSTYTUCJI 3 MAJA W POBIEDZISKACH LETNISKU		93.5378472222222	11.0361111111111
01.01.2022	30660	MODRZE	ZESPÓŁ SZKOLNO-PRZEDSZKOLNY W MODRZU		4.22361111111111	50.8701388888889
01.01.2022	31112	PŁOCK	SZKOŁA PODSTAWOWA NR 5 IM. WŁADYSŁAWA BRONIEWSKIEGO W PŁOCKU	0	59.7949494949495	24.2722853535354
01.01.2022	34711	JEZIORKI	SZKOŁA PODSTAWOWA IM. PRZYJACIÓŁ WIELKOPOLSKI W JEZIORKACH		94.1569444444444	19.0861111111111
01.01.2022	38778	PŁOCK	SZKOŁA PODSTAWOWA NR 20 IM. WŁADYSŁAWA BRONIEWSKIEGO W PŁOCKU	0	48.967803030303	16.4508207070707
01.01.2022	44065	BABOROWO	SZKOŁA PODSTAWOWA IM. PPŁK. MAKSYMILIANA CIĘŻKIEGO W BABOROWIE		93.5902777777778	4.20625
01.01.2022	59645	POBIEDZISKA	SZKOŁA PODSTAWOWA IM. KAZIMIERZA ODNOWICIELA W POBIEDZISKACH		95.3215277777778	10.4333333333333
01.01.2022	70077	KÓRNIK	SZKOŁA PODSTAWOWA NR 2 IM. TEOFILI Z DZIAŁYŃSKICH SZOŁDRSKIEJ-POTULICKIEJ W KÓRNIKU		95.7965277777778	23.3298611111111

['date', 'rspo', 'city', 'name', 'dew_point_avg', 'humidity_avg', 'pm10_avg', 'pm25_avg', 'pressure_avg', 'temperature_avg', 'humidity_min', 'humidity_max', 'pm10_min', 'pm10_max', 'pm25_min', 'pm25_max', 'pressure_min', 'pressure_max', 'temperature_min', 'temperature_max', 'dew_point_min', 'dew_point_max']

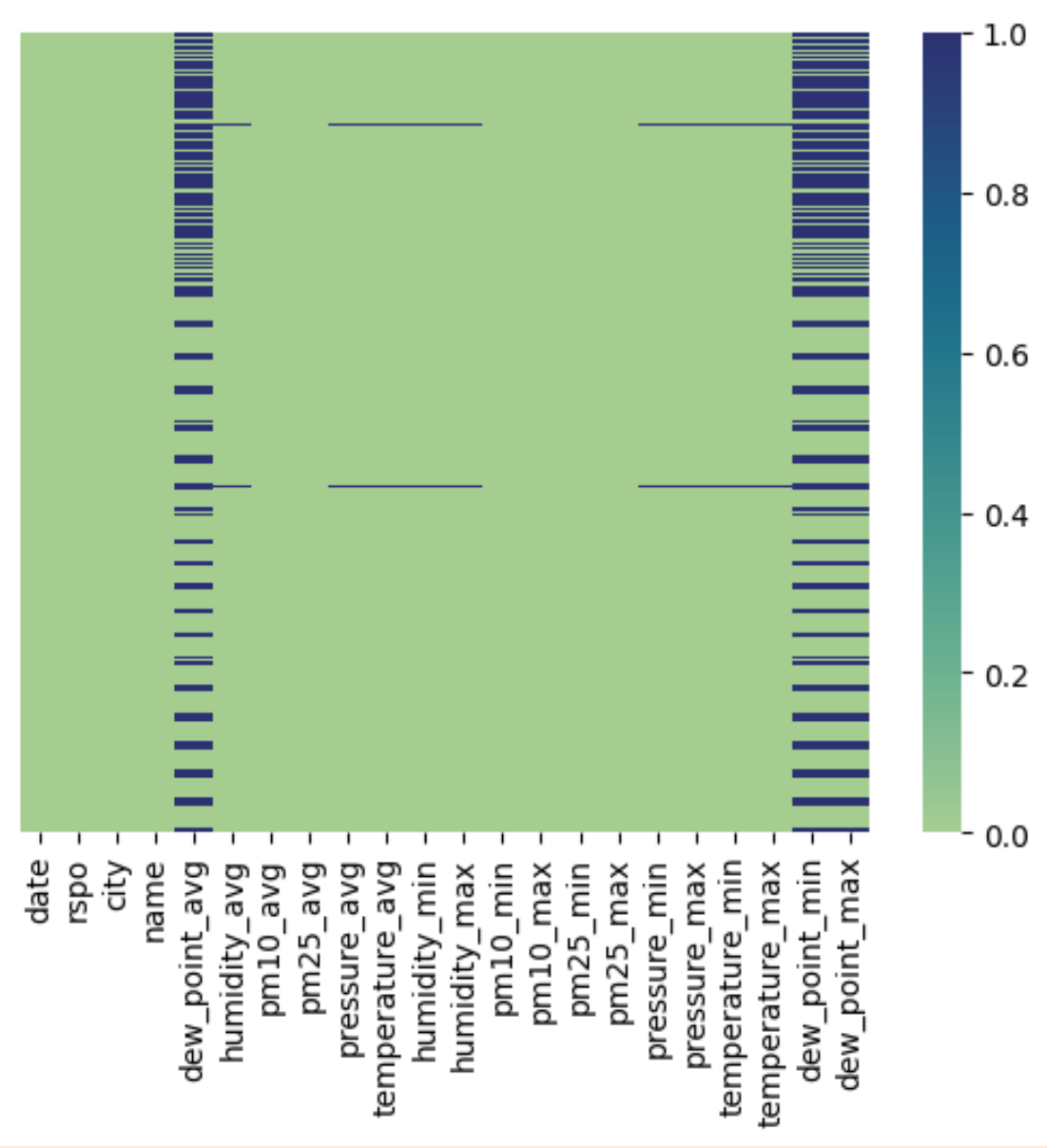
1. Nasz zbiór danych

- Plik rspo_data.json
- Dane szkół z Rejestru Szkół i Placówek Oświatowych


```
"@context": "/api/contexts/Placowka",
"@id": "/api/placowki/",
"@type": "hydra:Collection",
"hydra:member": [
    {
        "@id": "/api/placowki/2869",
        "@type": "Placowka",
        "numerRspo": 2869,
        "dataZalozenia": "1977-08-31T00:00:00+02:00",
        "dataRozpoczecia": "1977-08-31T00:00:00+02:00",
        "dataZakonczenia": "9999-12-31T00:00:00+01:00",
        "nip": "7393077918",
        "regon": "510015171",
        "nazwaSkrocona": "PM37",
        "nazwa": "PRZEDSZKOLE MIEJSKIE NR 37 W OLSZTYNIE",
```

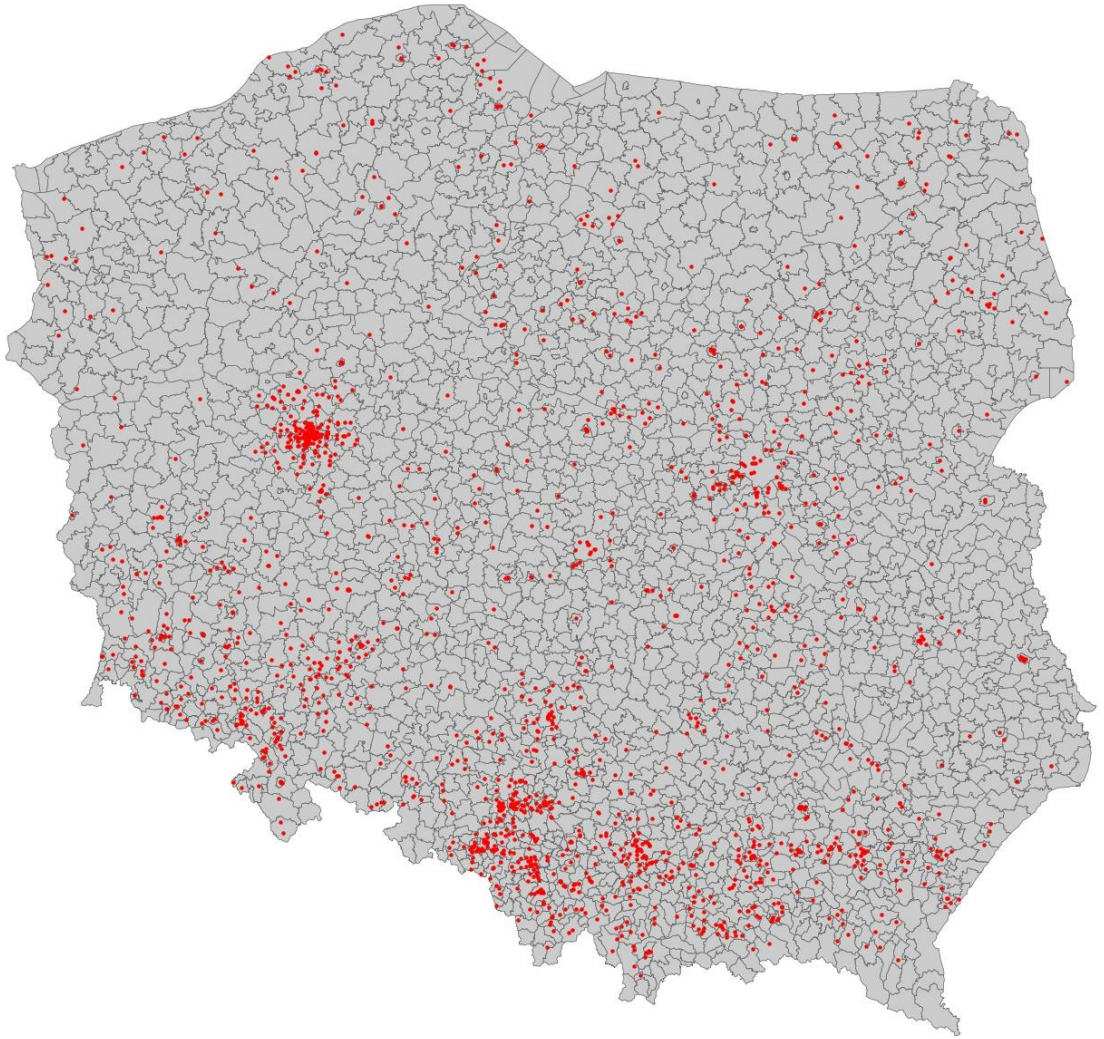

Problem z danymi: wartości "rspo" w pliku .xlsx, miały odpowiadać wartościom z pliku .json, ale jakiekolwiek próby połączenie tych danych ze sobą kończyły się porażką.

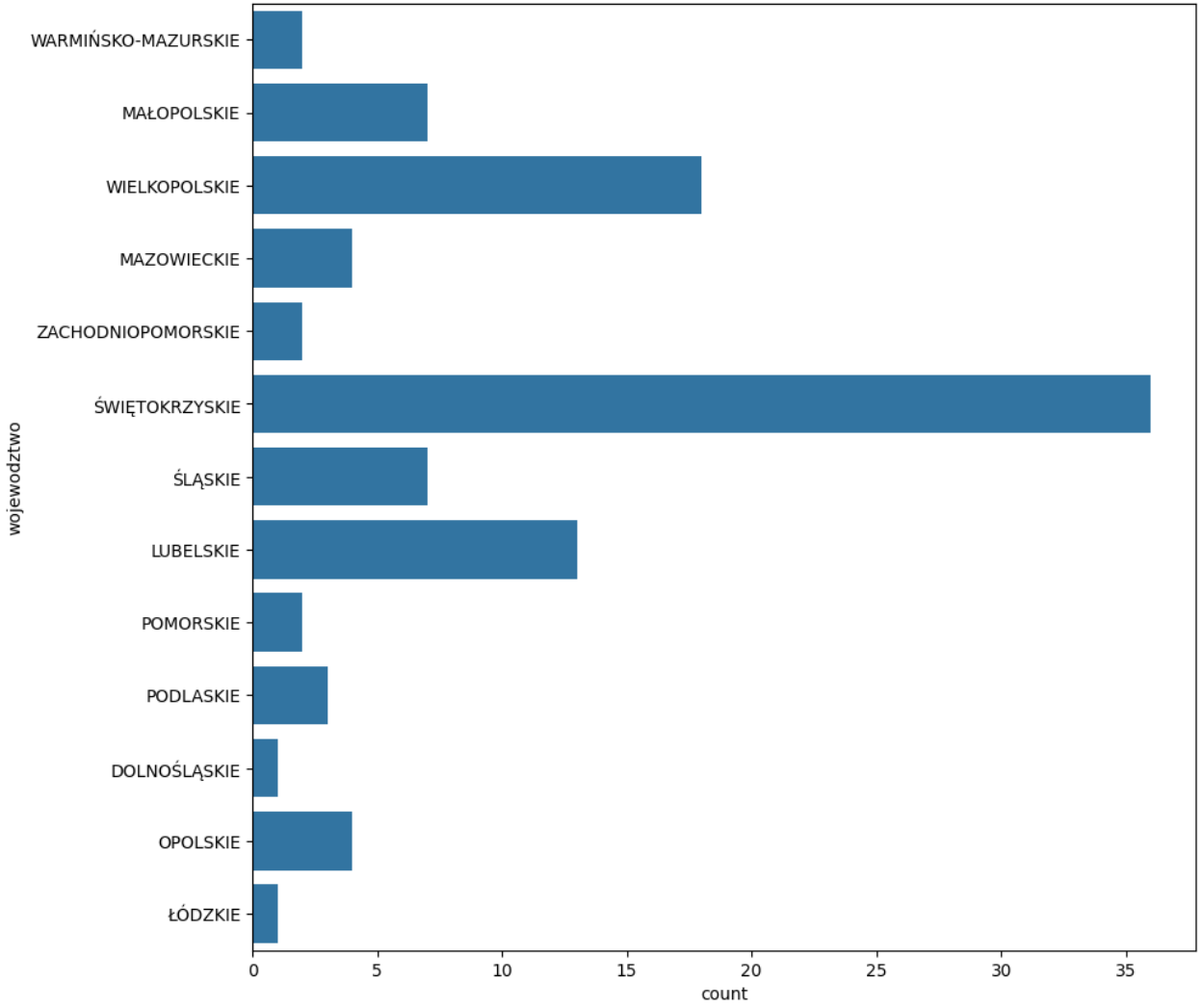
2. Analiza danych



Brakujące wartości

Rozmieszczenie placówek pomiarowych





Patroni szkół

Analiza poprzez zastosowanie NLP do danych

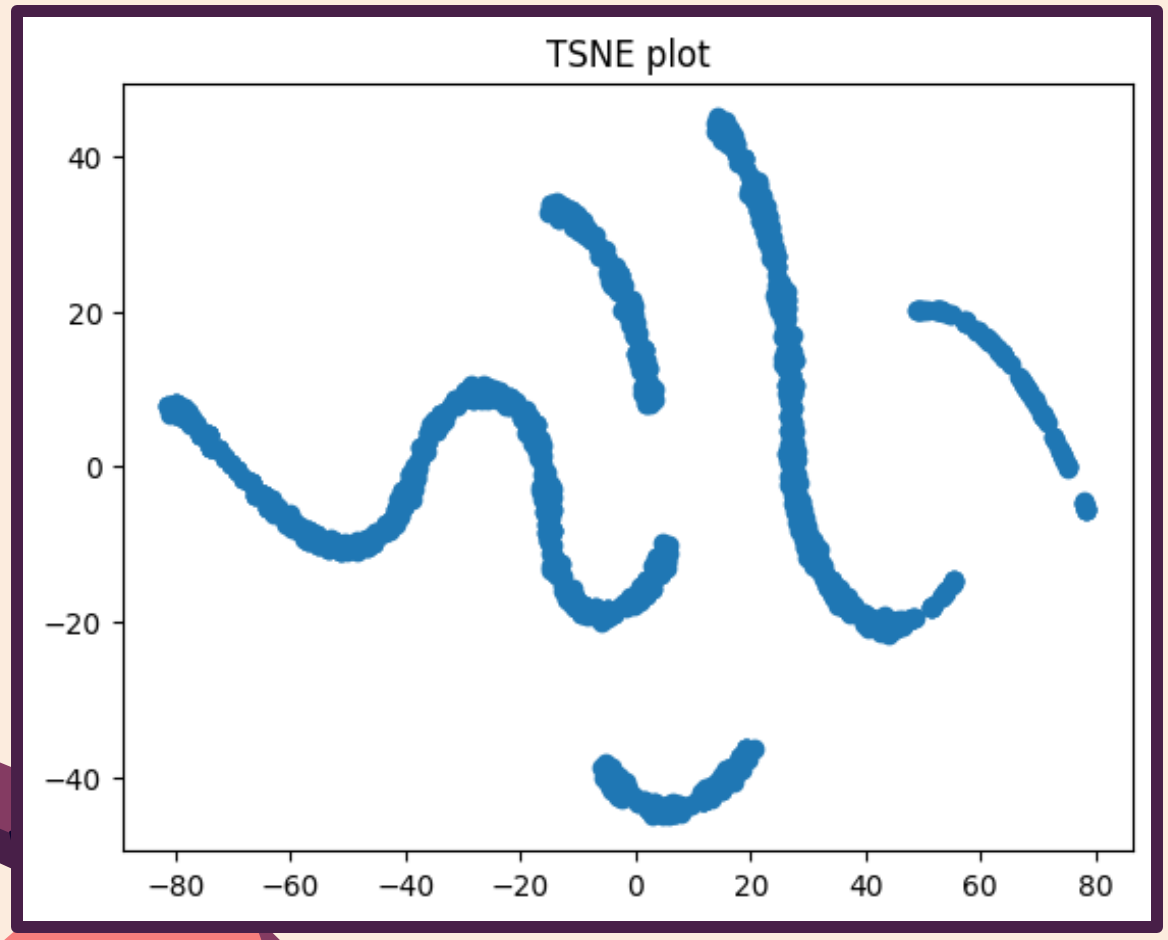
- 'Jan Paweł II': 56,
- 'Mikołaj Kopernik': 45,
- 'Maria Konopnicka': 42,
- 'Adam Mickiewicz': 19,
- 'Janusz Korczak': 16,
- 'Jan Kochanowski': 16,
- 'Jan Twardowski': 14,
- 'Kornel Makuszyński': 14,
- 'Stefan Wyszyński': 13,
- 'Henryk Sienkiewicz': 13

```
import spacy    Nie można rozpoznać importu „spacy”.
from collections import Counter

nlp = spacy.load("pl_core_news_lg")

doc = nlp(' | '.join(df2['name']).title())

person_lemmas = []

for ent in doc.ents:
    if ent.label_ == 'persName':
        person_lemmas.append(ent.lemma_)

person_lemmas = [x.title() for x in person_lemmas if len(x) > 3]

# get the most common names
most_common = dict(Counter(person_lemmas).most_common(10))
if 'Jan Paweł Ii W' in most_common:
    most_common['Jan Paweł II'] = most_common['Jan Paweł Ii W']
    most_common.pop('Jan Paweł Ii W')
most_common = dict(sorted(most_common.items(), key=lambda item: item[1], reverse=True))
for k, v in most_common.items():
    print(f'{k}: {v}')
```

✓ 4.0s

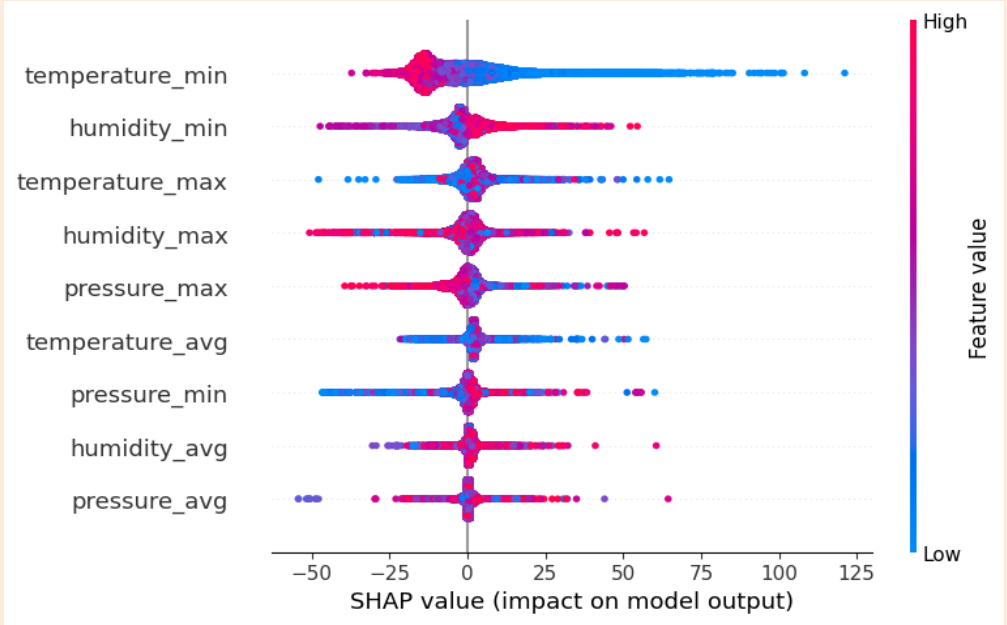
```
Jan Paweł II: 56
Mikołaj Kopernik: 45
Maria Konopnicka: 42
Adam Mickiewicz: 19
Janusz Korczak: 16
Jan Kochanowski: 16
Jan Twardowski: 14
Kornel Makuszyński: 14
Stefan Wyszyński: 13
Henryk Sienkiewicz: 13
```

Redukcja wymiarów

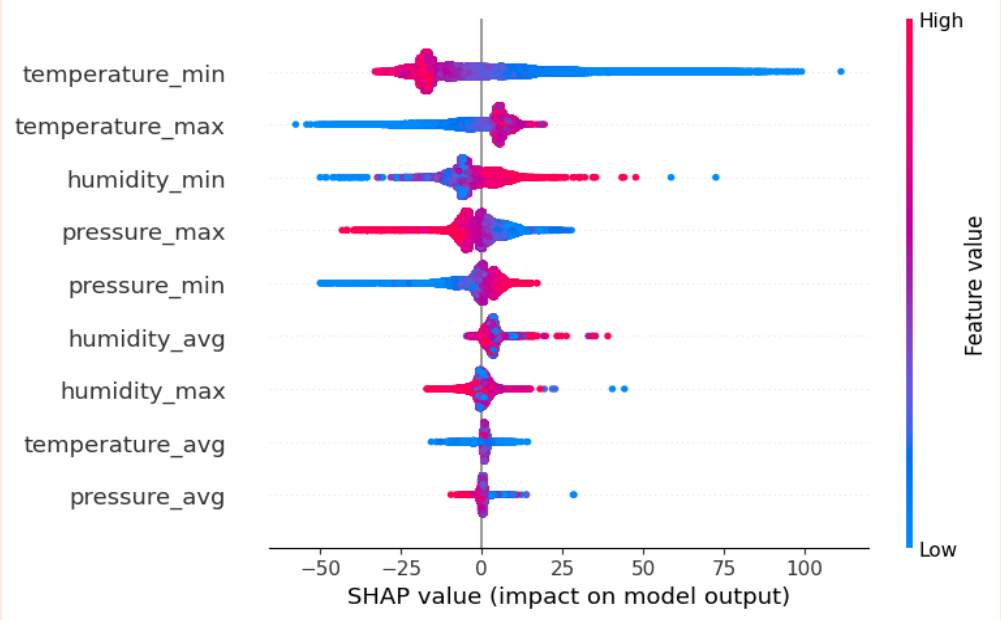


3. Ewaluacja modelu

- Wybrano modele XGBoost, DecisionTrees i ARIMA.
- Modele przewidują wartości zanieczyszczeń PM10 i PM25.

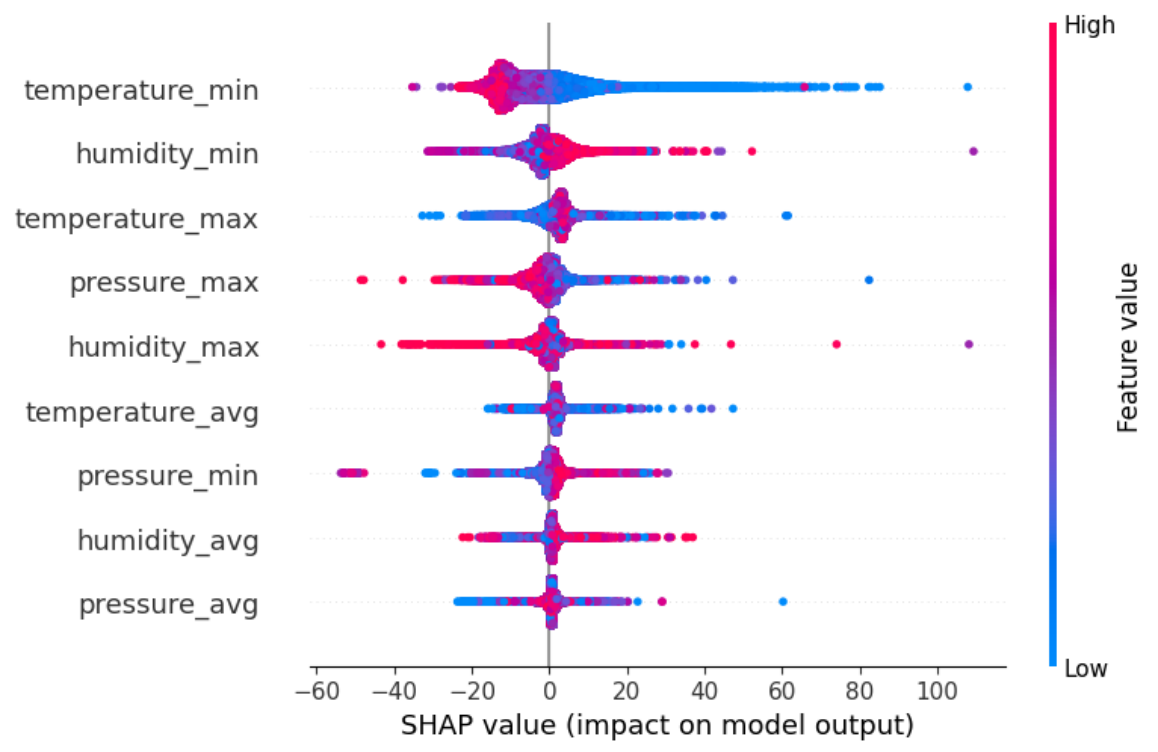


DecisionTreeRegressor(PM10)

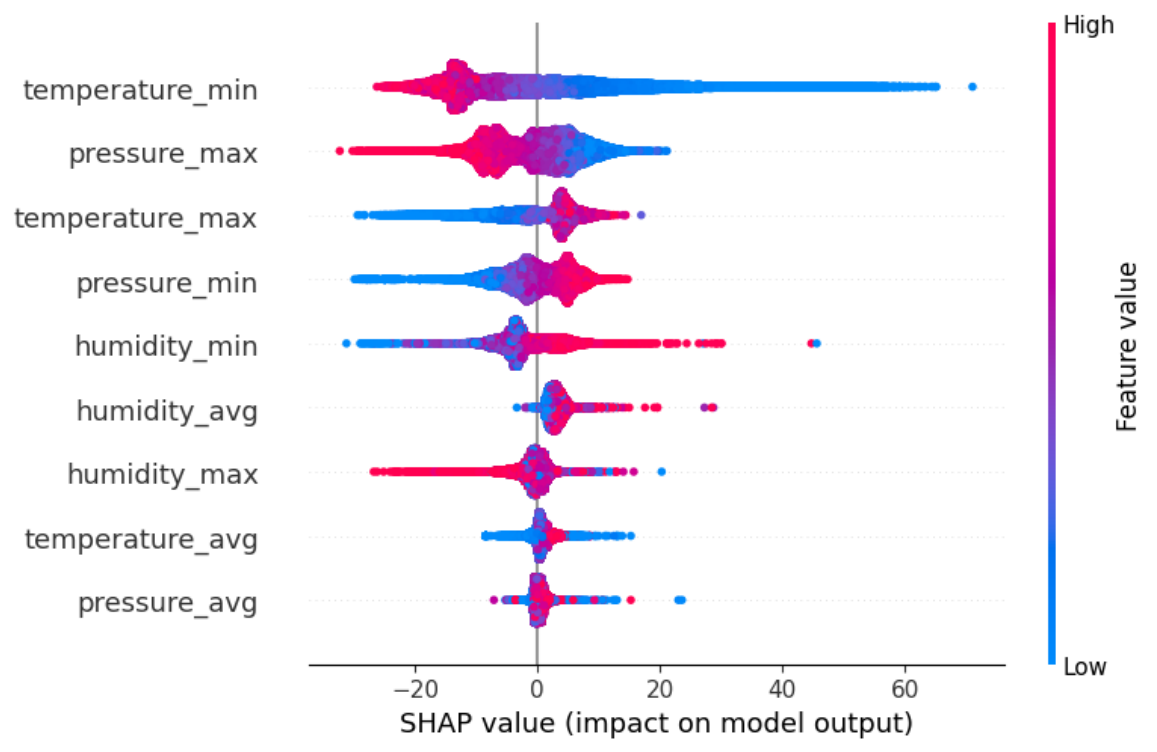


XGBRegressor(PM10)

Porównanie pm25

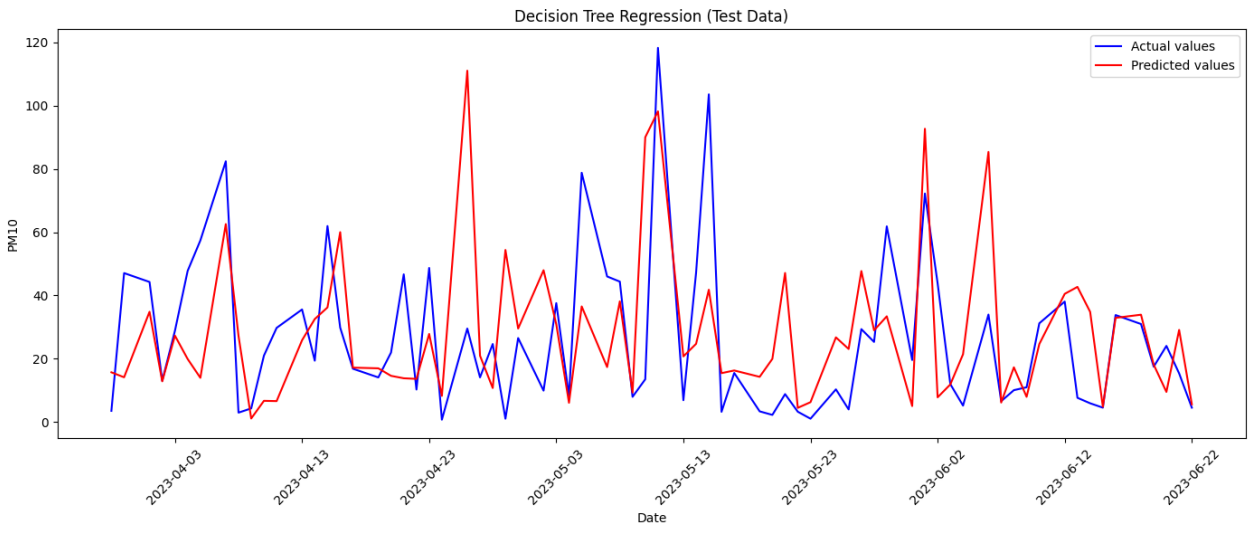


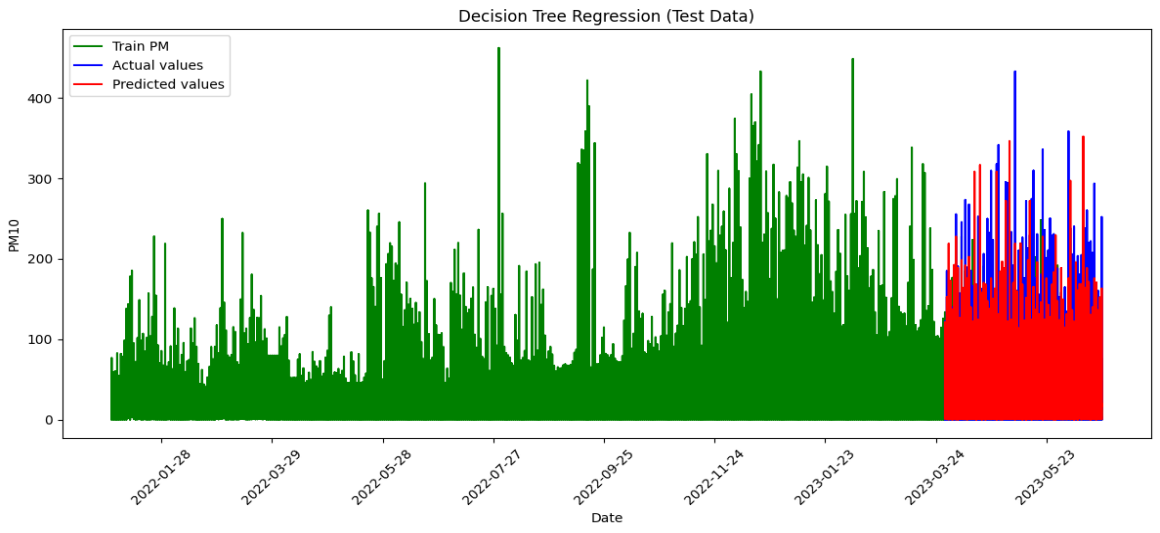
DecisionTreeRegressor

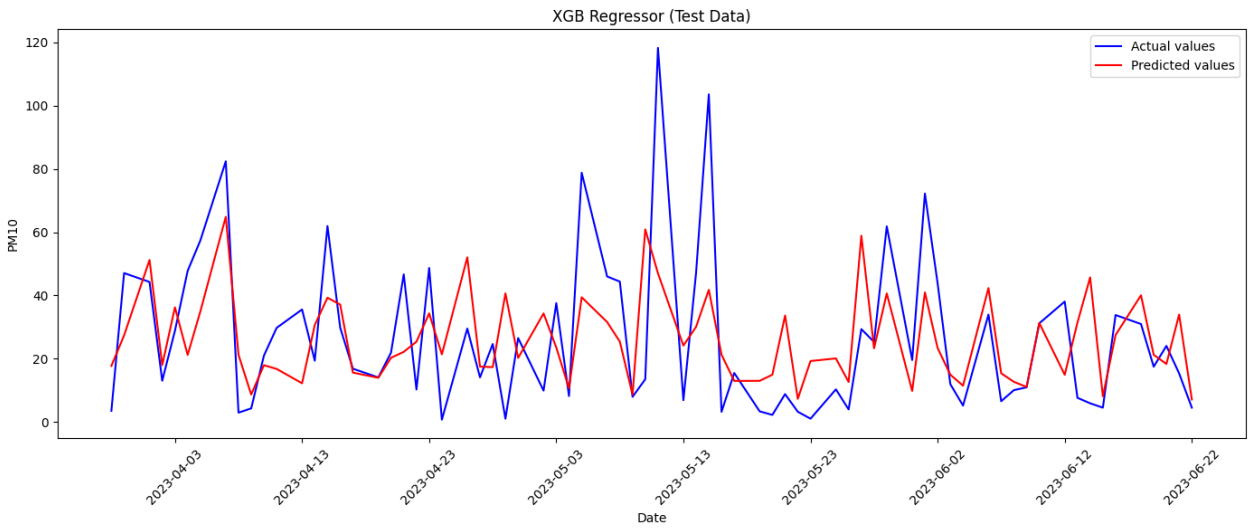


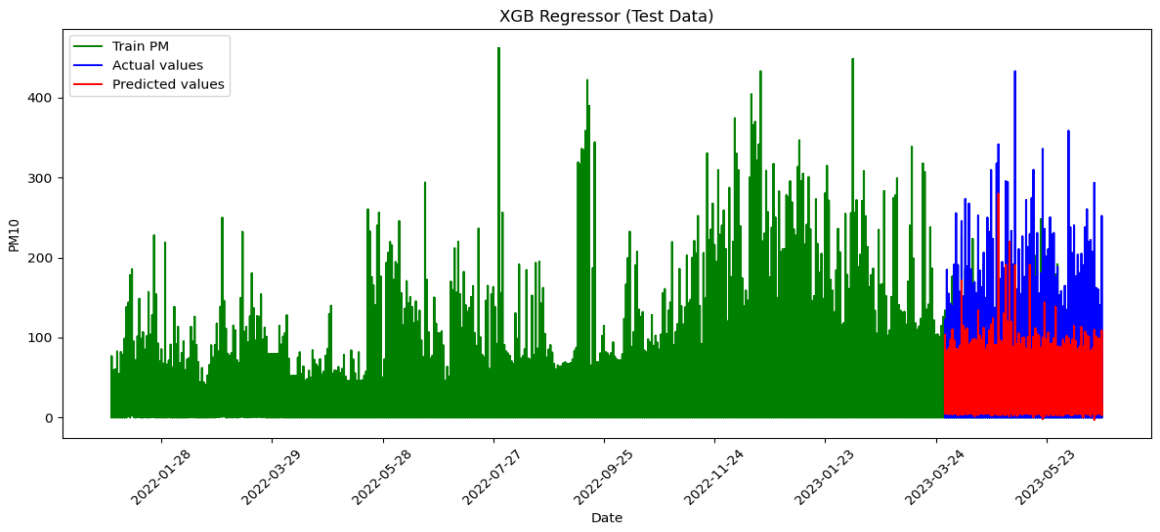
XGBRegressor

Porównanie predykcji dla atrybutu PM10.

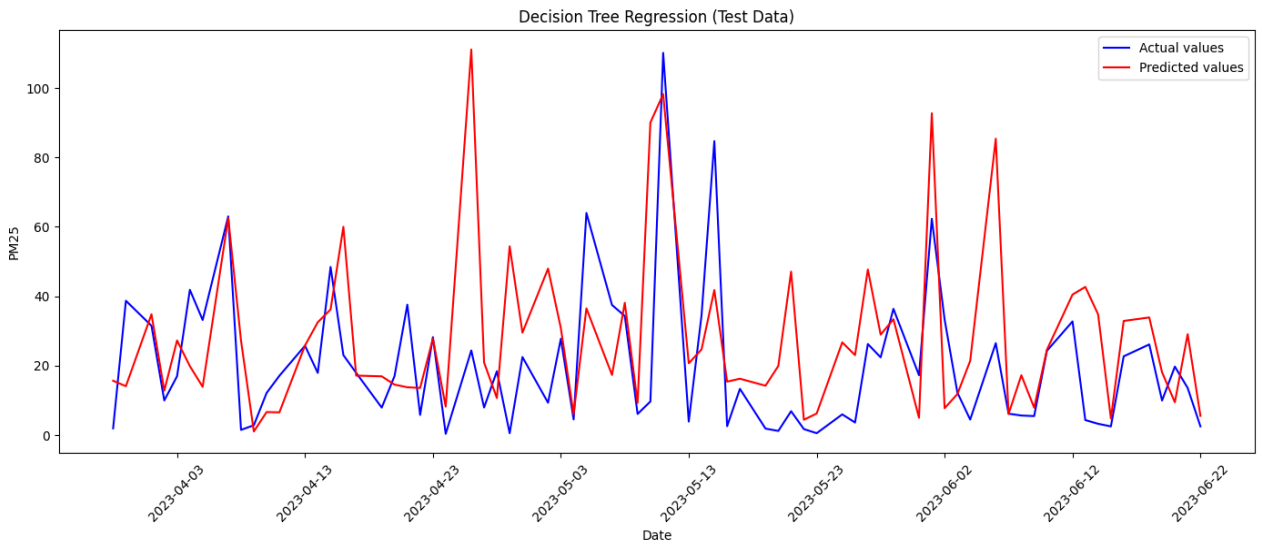


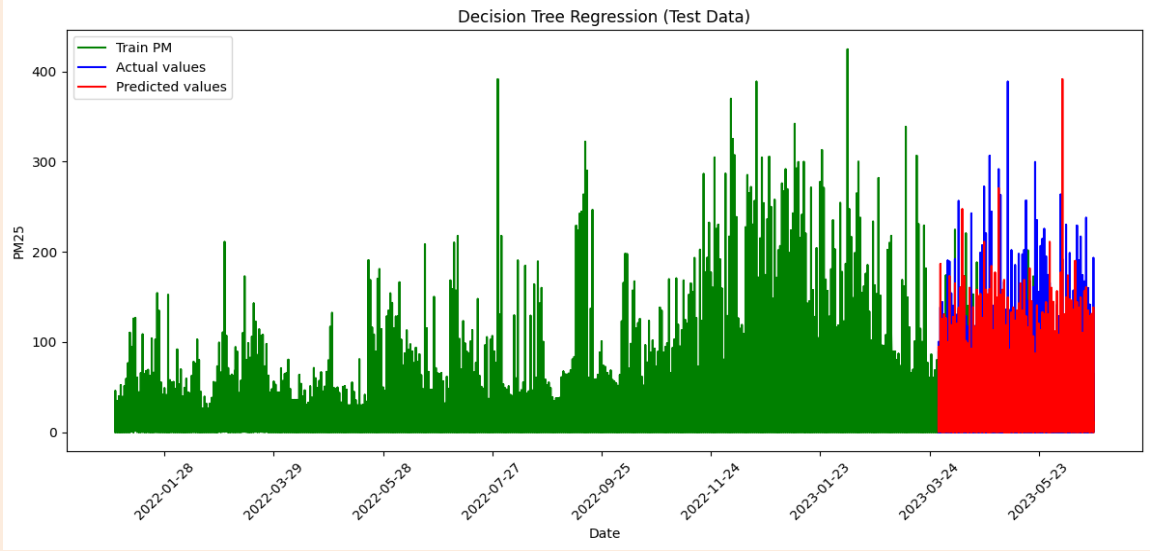


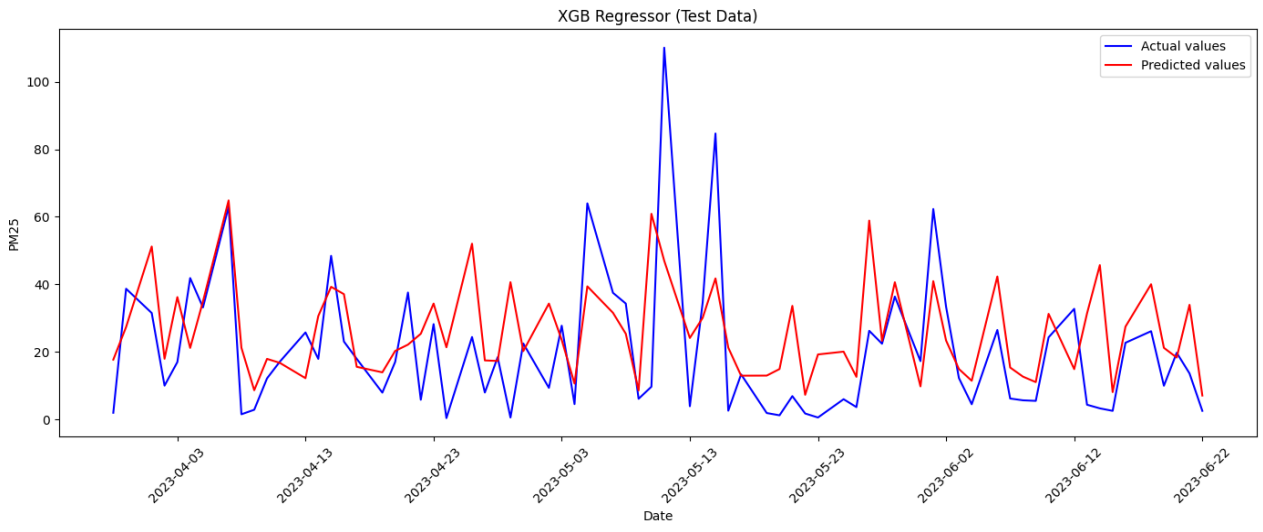


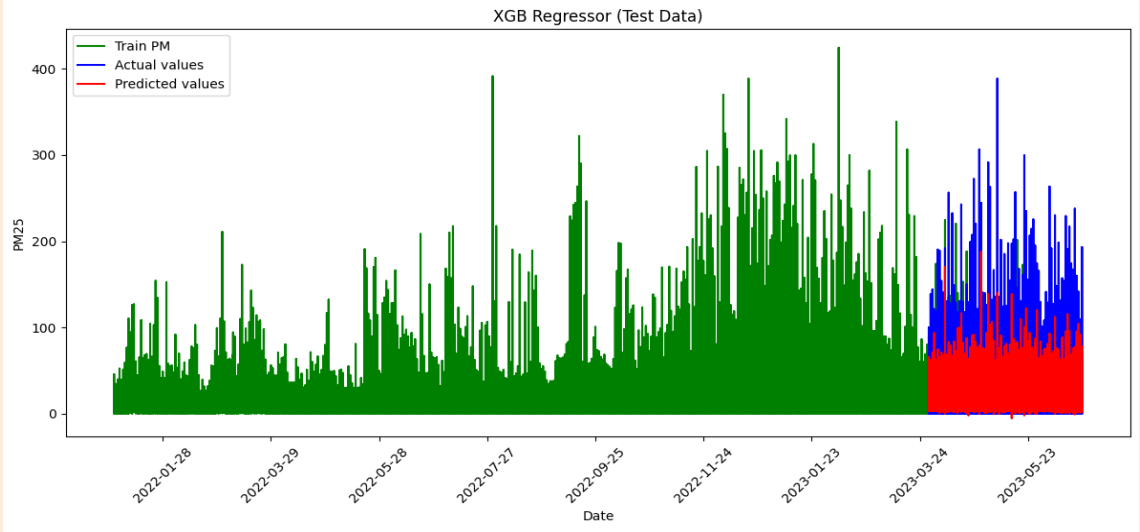


Porównanie predykcji dla atrybutu PM25.



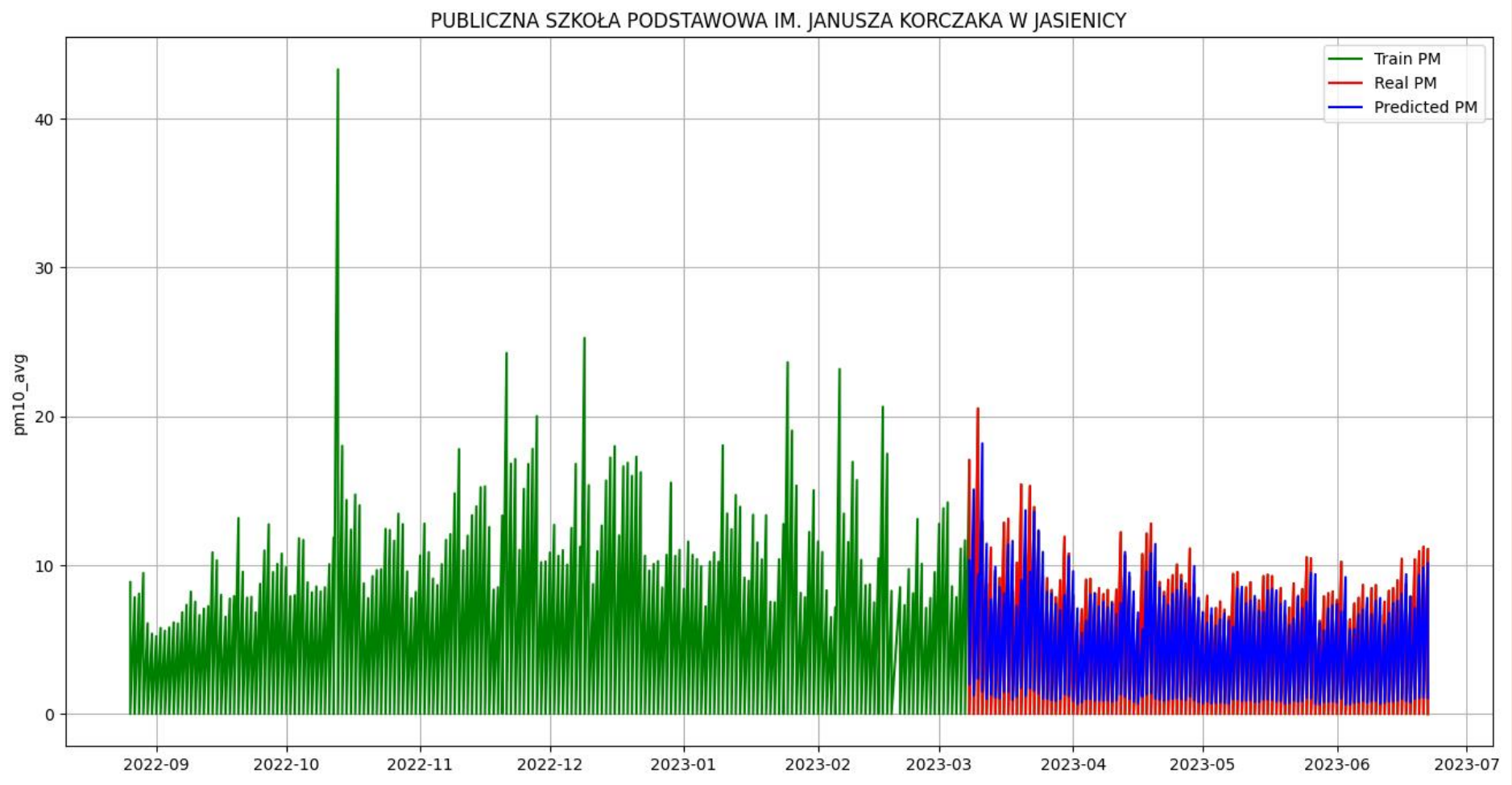






Model ARIMA

Na podstawie jednego punktu pomiarowego

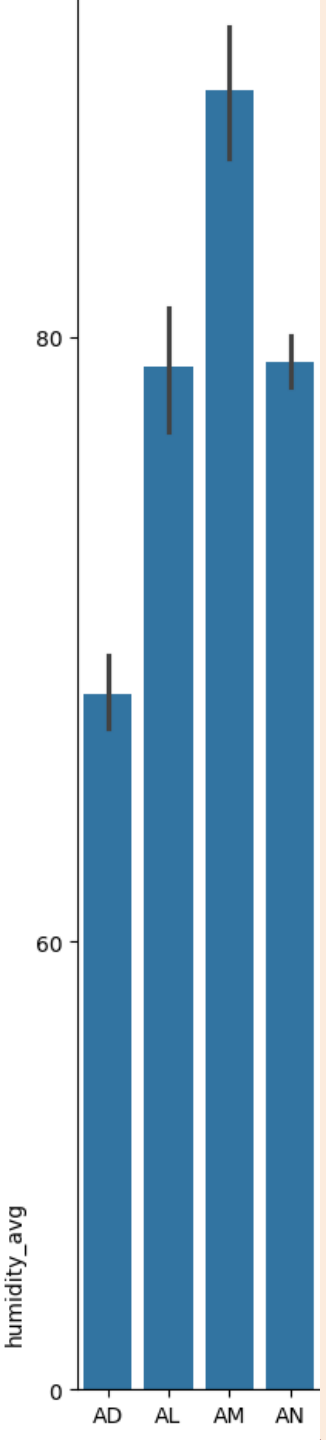


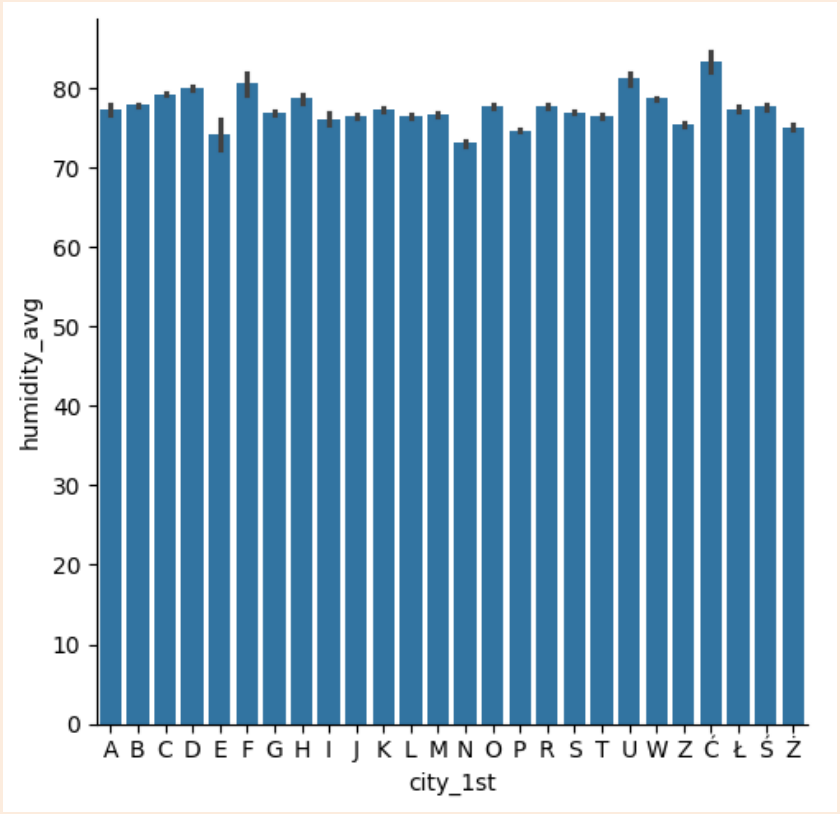
4. Najciekawsze elementy kursu

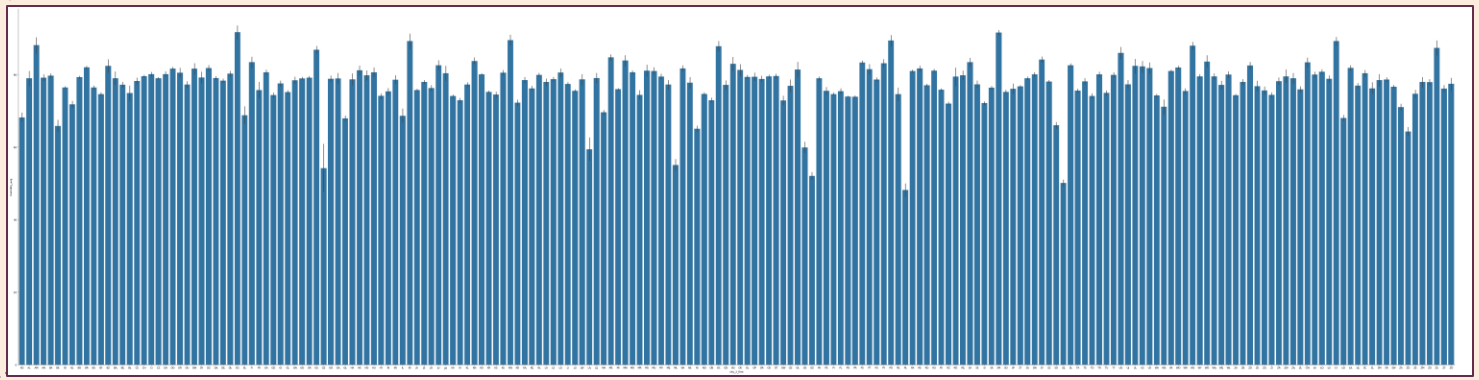
- Współpraca
- Wiele sposobów patrzenia na te same dane
- Temat wyjaśnialności modeli

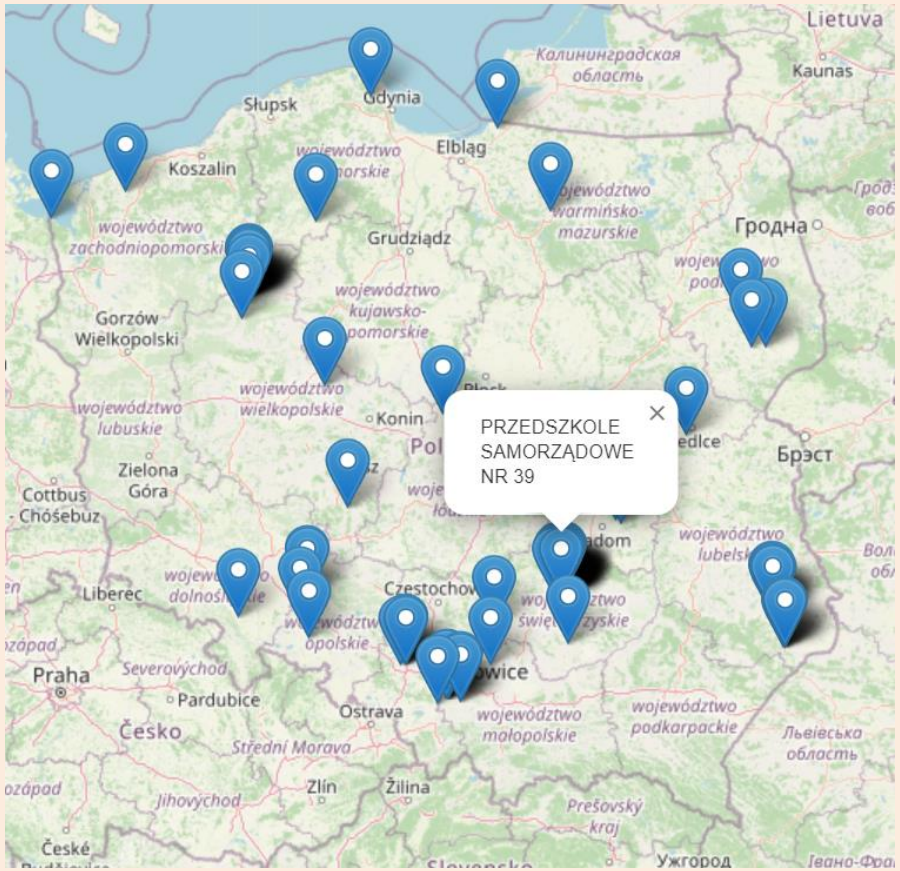
5. Czego brakowało?

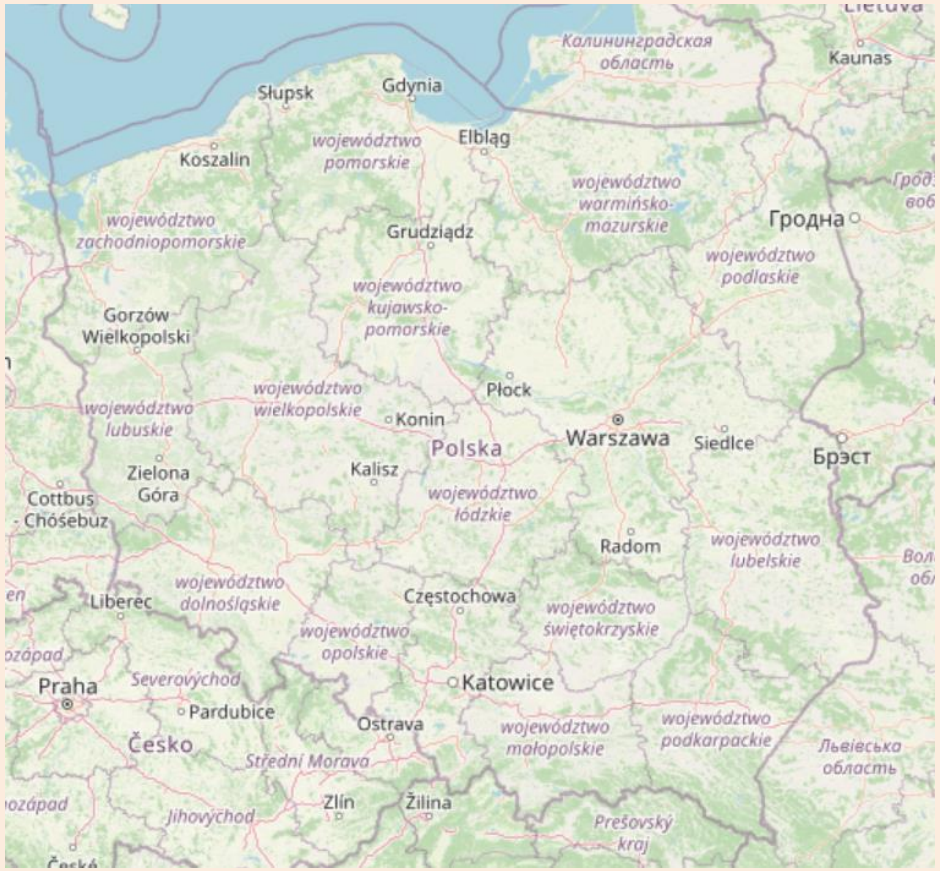
- W pewnym momencie integralność danych (a w zasadzie jej brak) okazała się przeszkodą do realizacji części naszych celów.

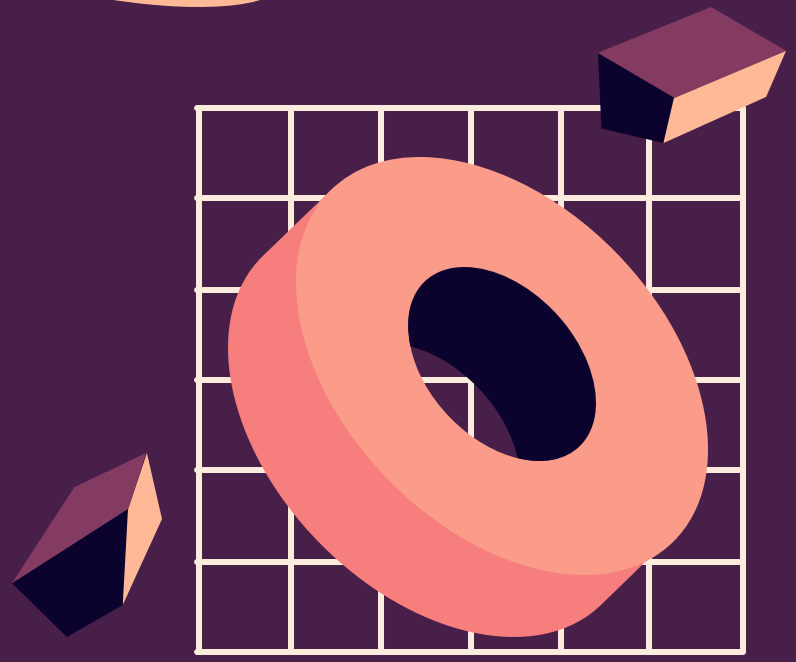

80
60
0
humidity_avg
AD
AL
AM
AN


80
70
60
50
40
30
20
10
0
humidity_avg
A B C D E F G H I J K L M N O P R S T U W Z Ć Ł S Ż
city_1st









Dziękujemy za uwagę!