AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE
WYDZIAŁ GEOLOGII, GEOFIZYKI I OCHRONY ŚRODOWISKA

KATEDRA GEOINFORMATYKI I INFORMATYKI STOSOWANEJ

Raport

Inżynieria cech i eksploracyjna analiza danych środowiskowych na podstawie pomiarów z systemu ESA (Edukacyjna Sieć Antysmogowa)

Autor: *Dominik Sakłaski*
Kierunek studiów: Inżynieria i Analiza Danych

Kraków, 2025

Spis treści

1. Zapoznanie z biblioteką SEBORN

Seaborn to wysokopoziomowa biblioteka Pythona przeznaczona do tworzenia statystycznych wizualizacji danych. Została oparta na Matplotlib i zintegrowana z pandas oraz NumPy. Umożliwia szybkie tworzenie estetycznych wykresów, automatycznie mapując dane na atrybuty wizualne, takie jak kolor czy rozmiar.

Dzięki prostemu API i gotowym funkcjom biblioteka wspiera eksploracyjną analizę danych, umożliwiając badanie rozkładów, relacji oraz różnic pomiędzy grupami. Pozwala na tworzenie wielopanelowych układów wykresów, a dzięki ścisłej współpracy z Matplotlib umożliwia również ich zaawansowane dostosowywanie. Seaborn jest użyteczny na wszystkich etapach projektów analitycznych: od eksploracji danych po ocenę i prezentację wyników.

2. Charakterystyka portalu Dane.gov.pl

2.1. Ogólne informacje

Portal Dane.gov.pl stanowi centralne repozytorium danych publicznych w Polsce, zarządzane przez Ministerstwo Cyfryzacji. Umożliwia dostęp do różnorodnych zbiorów danych udostępnianych przez instytucje publiczne, takich jak Główny Urząd Statystyczny, Ministerstwo Finansów, Ministerstwo Edukacji Narodowej czy Centralna Ewidencja i Informacja o Działalności Gospodarczej.

Dane mogą obejmować m.in. informacje statystyczne, geolokalizacyjne, środowiskowe oraz rejestrowe. Portal zawiera zbiory danych z wielu dziedzin, takich jak: energetyka, transport, rolnictwo, środowisko, ludność i społeczeństwo oraz zdrowie. Wymienione dziedziny nie wyczerpują pełnego zakresu tematycznego dostępnych zasobów.

2.2. Sposoby udostępniania danych

Dane na portalu udostępniane są w różnych formatach plików, takich jak CSV (Comma-Separated Values), XLS, XLSX, XML (Extensible Markup Language) oraz JSON (JavaScript Object Notation). Dodatkowo istnieje możliwość dostępu do danych za pośrednictwem API, w tym REST API oraz SOAP API. Interfejsy API umożliwiają automatyczne pobieranie danych i ich integrację z systemami informatycznymi.

2.3. Interfejsy API dostępne na portalu

Dostęp do danych na portalu dane.gov.pl jest możliwy nie tylko poprzez bezpośrednie pobieranie plików, lecz także za pośrednictwem różnych typów API (Application Programming Interface). API umożliwia programową komunikację z bazą danych, automatyczne pobieranie, aktualizowanie i przetwarzanie danych. Na portalu dostępne są następujące typy API:

- **RESTful API** – oparty na protokole HTTP, umożliwia pobieranie i aktualizowanie danych w formacie JSON lub XML.
- **SOAP API** – standard bazujący na XML, wykorzystywany głównie w usługach takich jak CEIDG.
- **GraphQL API** – nowoczesny interfejs umożliwiający precyzyjne zapytania o wybrane dane.

Szczegóły dotyczące wykorzystania interfejsów API oraz informacje o sposobie autoryzacji i formatach danych dostępne są bezpośrednio w opisach poszczególnych zbiorów na portalu dane.gov.pl.

2.4. Rodzaje dostępnych danych

Na portalu dostępne są dane historyczne, jak i bieżąco aktualizowane obejmujące szeroki zakres tematyczny, m.in.:

- dane demograficzne i społeczne,
- dane gospodarcze i finansowe,
- informacje o środowisku i klimacie,

- dane o transporcie i infrastrukturze,
- dane dotyczące edukacji, nauki i zdrowia,
- informacje o bezpieczeństwie publicznym i administracji.

2.5. Sposoby wykorzystania danych

Dane dostępne na portalu mogą być wykorzystywane do:

- podnoszenia jakości usług publicznych poprzez analizę procesów i wydatków,
- wspomagania podejmowania decyzji administracyjnych i inwestycyjnych,
- oceny skuteczności programów rządowych i społecznych,
- analizy trendów społecznych, gospodarczych i środowiskowych,
- tworzenia modeli predykcyjnych w projektach uczenia maszynowego,
- rozwoju aplikacji opartych na danych otwartych (open data).

2.6. Szybka ocena jakości danych

W celu szybkiej weryfikacji jakości danych dostępnych na portalu należy zwrócić uwagę na:

- aktualność i datę ostatniej aktualizacji zbioru,
- częstotliwość aktualizacji danych,
- kompletność danych i sposób oznaczania brakujących wartości,
- zgodność danych z przyjętymi standardami formatowania (np. poprawne separatory w plikach CSV, spójne typy danych).

Na stronie portalu dostępne są również podstawowe informacje o zbiorach, takie jak opis danych, ich struktura oraz informacje o źródle i odpowiedzialnej instytucji, co ułatwia ocenę przydatności i jakości zbioru przed jego wykorzystaniem.

3. Wybór zestawu danych

3.1. Opis wybranego zbioru danych

Do analizy wybrano zbiór pt. **„Dane pomiarowe ESA (Edukacyjna Sieć Antysmogowa)”**, opublikowany na portalu dane.gov.pl. Zbiór zawiera dane środowiskowe rejestrowane przy placówkach edukacyjnych w całej Polsce i jest aktualizowany co 15 minut.

3.2. Powody wyboru zbioru danych

Zbiór został wybrany ze względu na swoją strukturę, która umożliwia realizację kompleksowych analiz z wykorzystaniem narzędzi i technik Data Science. Dane posiadają charakter dynamiczny oraz oznaczenia czasowe i przestrzenne, co pozwala na prowadzenie analiz w układzie czasowo-przestrzennym. Struktura zbioru wpisuje się w typowy pipeline analityczny, umożliwiając zastosowanie metod eksploracyjnych, wizualizacyjnych oraz predykcyjnych.

3.3. Znaczenie tematyki zbioru danych

Zanieczyszczenie powietrza i występowanie smogu to istotne problemy z perspektywy zdrowia publicznego. Analiza danych z systemu ESA może przyczynić się do głębszego zrozumienia warunków środowiskowych, identyfikacji czynników wpływających na jakość powietrza oraz oceny zmian w czasie i przestrzeni. Zgromadzone dane mogą znaleźć zastosowanie zarówno w badaniach naukowych, jak i w projektowaniu systemów wspomagania decyzji dla administracji publicznej i jednostek samorządowych.

4. Etapy pipeline'u ML

4.1. Pobranie danych i zapoznanie się ze strukturą zbioru danych

Dane zostały pobrane ze strony dane.gov.pl w formie plików CSV, aktualizowanych z częstotliwością co 15 minut. Na potrzeby niniejszej analizy ograniczono zakres czasowy do jednej doby – **28 kwietnia 2025 roku** – i wybrano próbkę danych z godzinnym interwałem. W tym celu pobrano 24 pliki odpowiadające kolejnym godzinom doby (od 00:00 do 23:00), a następnie połączono je w jeden zbiorczy plik zawierający dane z całego dnia.

W ramach zapoznania się ze strukturą zbioru danych przeprowadzono wstępną inspekcję zawartości pliku. Każdy wiersz odpowiada pojedynczemu pomiarowi pochodzącemu z konkretnej lokalizacji (najczęściej placówki edukacyjnej) w określonym momencie czasowym. Dane zawierają kolumny opisujące zarówno lokalizację geograficzną (miasto, ulica, kod pocztowy, współrzędne GPS), jak i parametry środowiskowe (średnia wilgotność powietrza, ciśnienie atmosferyczne, temperatura, stężenia pyłów PM10 i PM2.5). Ostatnia kolumna zawiera znacznik czasowy, umożliwiający analizę zmian w czasie.

Dane są częściowo ustrukturyzowane – występują zarówno zmienne ciągłe, jak i kategoryczne. Na podstawie wstępnego przeglądu można stwierdzić, że struktura zbioru jest spójna, a kolumny mają intuicyjne i jednoznaczne nazwy. Przykładowe rekordy danych przedstawiono na rysunku Fig. 4.1, który obrazuje fragment końcowego zbioru po scaleniu danych z całej doby.

Tak przygotowany zbiór stanowi punkt wyjścia do dalszych etapów procesu analitycznego, w tym inżynierii cech i eksploracyjnej analizy danych.

	NAME	STREET	POST_CODE	CITY	LONGITUDE	LATITUDE	HUMIDITY_AVG	PRESSURE_AVG	TEMPERATURE_AVG	PM10_AVG	PM25_AVG	TIMESTAMP
0	SZKO?A PODSTAWOWA IM. MARIANA FALSKIEGO W KRASZEWICACH	UL. SZKOLNA	63-522	KRASZEWICE	18.224030	51.515630	95.120000	1015.110000	8.610000	32.860000	20.200000	2025-04-28 00:00:00.0
1	SZKO?A PODSTAWOWA WE WRZ? SOWICACH	UL. SZKOLNA	32-040	WRZ? SOWICE	19.942820	49.961030	49.040000	988.780000	9.060000	5.780000	5.120000	2025-04-28 00:00:00.0
2	PUBLICZNA SZKO?A PODSTAWOWA NR 2 IM. KAZIMIERZA MAŁCZEWSKIEGO W STRZELCACH OPOLSKICH	UL. WAWRZY?CA ?WIERZEGO	47-100	STRZELCE OPOLSKIE	18.314889	50.503431	67.333333	1044.133333	6.133333	1.000000	0.546667	2025-04-28 00:00:00.0
40566	ZESP?? SZK?? ZAKONU PIJAR? W IM. ?W. J?ZEFA KALASANCJUSZA W POZNANIU	OSIEDLE JANA III SOBIESKIEGO	60-688	POZNA?	16.913836	52.465106	51.740000	1013.749600	12.846400	12.400000	9.160000	2025-04-28 23:00:00.0
40567	SZKO?A PODSTAWOWA IM. JANA PAW?A II W TUSZYNIE	nan	58-207	TUSZYN	16.646380	50.798453	65.841667	995.508333	11.933333	31.066667	28.425000	2025-04-28 23:00:00.0
40568	ZESP?? SZK?? IM. O. MARIANA ? ELAZKA W CHLUDOWIE	SZKOLNA	62-001	CHLUDOWO	16.845527	52.555605	85.823600	1014.150800	10.337600	37.040000	29.480000	2025-04-28 23:00:00.0

Fig. 4.1. Przykładowa zawartość końcowego zbioru danych po połączeniu 24 plików CSV z dnia 28.04.2025 r.

4.2. Zastosowanie danych w analizach

Zbiór danych z projektu ESA (Edukacyjna Sieć Antysmogowa), zawierający bieżące i regularnie aktualizowane pomiary jakości powietrza w otoczeniu placówek eduka-

cyjnych, może być szeroko wykorzystywany w analizach środowiskowych, edukacyjnych i zarządczych. Dane mają bogatą strukturę – obejmującą zarówno dane lokalizacyjne, jak i środowiskowe, z przypisanym znacznikiem czasu.

Potencjalne zastosowania obejmują m.in.:

- **Monitoring jakości powietrza w otoczeniu placówek edukacyjnych** – dane pozwalają śledzić poziom zanieczyszczeń powietrza w konkretnych szkołach i przedszkolach
- **Porównywanie rejonów pod względem zanieczyszczeń** – zróżnicowanie geograficzne placówek umożliwia analizę przestrzenną, porównującą np. poziomy PM2.5 w mieście a w gminie wiejskiej w analogicznych godzinach.
- **Analiza wpływu warunków meteorologicznych na zanieczyszczenia** – możliwe jest badanie zależności pomiędzy np. temperaturą, wilgotnością czy ciśnieniem a stężeniem cząstek pyłu, co może pomóc w identyfikacji warunków sprzyjających ich kumulacji.
- **Identyfikacja najbardziej zanieczyszczonych obszarów** – na podstawie agregacji danych czasowych można wyłonić obszary, w których regularnie przekraczane są dopuszczalne poziomy pyłów zawieszonych.
- **Wsparcie dla działań edukacyjnych i informacyjnych** – dane zebrane w placówkach edukacyjnych mogą być bezpośrednio wykorzystywane w działaniach zwiększających świadomość uczniów, rodziców i społeczności lokalnych na temat zagrożeń związanych ze smogiem i jakości powietrza.

4.3. Zastosowanie danych w uczeniu maszynowym

W kontekście uczenia maszynowego dane te oferują szerokie możliwości analityczne. Ich struktura, częstotliwość aktualizacji oraz różnorodność zmiennych pozwalają na zastosowanie zarówno uczenia nadzorowanego, jak i nienadzorowanego.

Uczenie nadzorowane:

- **Prognozowania poziomu zanieczyszczeń** – budowa modeli regresyjnych przewidujących wartości pyłów na podstawie cech takich jak temperatura, wilgotność, lokalizacja czy pora dnia.
- **Klasyfikacji jakości powietrza** – przypisywanie obserwacji do kategorii na potrzeby ostrzeżeń lub decyzji administracyjnych.
- **Tworzenia systemów wczesnego ostrzegania** – modele predykcyjne mogą być zintegrowane z automatycznym informowaniem o pogorszeniu jakości powietrza.
- **Identyfikacji kluczowych czynników wpływających na zanieczyszczenia** – analiza cech wejściowych (np. temperatura, wilgotność, pora dnia), które mają największy wpływ na wartość przewidywanej zmiennej docelowej.

Uczenie nienadzorowane:

- **Klasteryzacji lokalizacji** – grupowanie punktów pomiarowych o podobnym profilu środowiskowym (np. wysoki poziom wilgotności i niskie PM2.5).
- **Wykrywania anomalii** – identyfikacja nietypowych pomiarów wskazujących na np. awarie sensorów lub nagłe pogorszenie jakości powietrza.

- **Redukcji wymiarowości** – uproszczenie zbioru danych przy pomocy technik takich jak PCA, co pozwala na lepszą wizualizację zależności.
- **Eksploracji zależności** – odkrywanie niewidocznych na pierwszy rzut oka powiązań między zmiennymi pogodowymi a jakością powietrza.

4.4. Inżynieria cech (FE) oraz eksploracyjna analiza danych (EDA)

4.4.1. Wczytanie danych

W pierwszym kroku wykonano wczytanie danych z uprzednio połączonego zbioru obejmującego wszystkie godziny doby **28 kwietnia 2025 roku**. Dane zostały wczytane przy użyciu biblioteki pandas, co umożliwiło szybki przegląd ich struktury, typów zmiennych oraz pierwszych obserwacji. Po połączeniu 24 godzinnych plików CSV powstał finalny zbiór danych zawierający 40569 rekordów, z których każdy reprezentuje pojedynczy pomiar środowiskowy z określonej lokalizacji i godziny.

Zbiór charakteryzuje się różnorodnością cech – obejmuje dane lokalizacyjne (np. miasto, kod pocztowy, współrzędne GPS), pomiary warunków atmosferycznych (wilgotność, temperatura, ciśnienie) oraz stężenia zanieczyszczeń (PM10, PM2.5). Dodatkowo kolumna TIMESTAMP przechowuje dokładny znacznik czasu dla każdego pomiaru.

Taka struktura sprzyja zastosowaniom analitycznym i predykcyjnym, umożliwiając modelowanie zjawisk w ujęciu czasowo-przestrzennym. Przykładowy wycinek wczytanego zbioru danych przedstawiono na rysunku Fig. 4.2., ilustrując sposób prezentacji danych środowiskowych w odniesieniu do lokalizacji i czasu pomiaru.

	NAME	STREET	POST_CODE	CITY	LONGITUDE	LATITUDE	HUMIDITY_AVG	PRESSURE_AVG	TEMPERATURE_AVG	PM10_AVG	PM25_AVG	TIMESTAMP
0	SZKO?A PODSTAWOWA IM. MARIANA FALSKIEGO W KRASZEWICACH	UL. SZKOLNA	63-522	KRASZEWICE	18.224030	51.515630	95.120000	1015.110000	8.610000	32.860000	20.200000	2025-04-28 00:00:00.0
1	SZKO?A PODSTAWOWA WE WRZ?SOWICACH	UL. SZKOLNA	32-040	WRZ? SOWICE	19.942820	49.961030	49.040000	988.780000	9.060000	5.780000	5.120000	2025-04-28 00:00:00.0
2	PUBLICZNA SZKO?A PODSTAWOWA NR 2 IM. KAZIMIERZA MALCZEWSKIEGO W STRZELCACH OPOLSKICH	UL. WAWRZYC?CA ?WIERZEGO	47-100	STRZELCE OPOLSKIE	18.314889	50.503431	67.333333	1044.133333	6.133333	1.000000	0.546667	2025-04-28 00:00:00.0
3	ZESP?? SZK?? NR 1 W PSZCZYNIE	UL. KAZIMIERZA WIELKIEGO	43-200	PSZCZYNA	18.945706	49.965883	86.466667	1005.600000	6.333333	9.666667	5.790000	2025-04-28 00:00:00.0
4	ZESP?? SZK?? IM. POWSTA?C?W WIELKOPOLSKICH W JANKOWIE PRZYGODZKIM	SZKOLNA	63-421	JANK?W PRZYGODZKI	17.788907	51.596172	63.133333	1013.033333	8.266667	15.666667	9.246667	2025-04-28 00:00:00.0
5	SZKO?A PODSTAWOWA NR 7 IM. ?O?NIERZY WRZE?NIA W ? WIKLICACH	UL. M? CZENNIK?W O?WI? CIMSKICH	43-229	?WIKLICE	18.989839	49.971937	70.166667	997.166667	7.800000	14.766667	8.420000	2025-04-28 00:00:00.0
6	SZKO?A PODSTAWOWA NR 12 W STUDZIONCE	UL. JORDANA	43-245	STUDZIONKA	18.774985	49.960356	66.433333	1004.233333	8.933333	13.400000	7.540000	2025-04-28 00:00:00.0
7	ZESP?? SZKOLNO-PRZEDSZKOLNY W PIASKU	SZKOLNA	43-211	PIASEK	18.946340	50.009550	55.900000	1004.966667	6.333333	17.033333	10.010000	2025-04-28 00:00:00.0
8	ZESP?? SZKOLNO-PRZEDSZKOLNY W ??CE	FITELBERGA	43-241	??KA	18.906757	49.958244	62.166667	1002.966667	5.833333	13.666667	7.526667	2025-04-28 00:00:00.0
9	SZKO?A PODSTAWOWA NR 3 W LUBONIU	ARMII POZNA?	62-030	LUBO?	16.896800	52.348100	87.300000	1013.233333	7.733333	24.566667	15.030000	2025-04-28 00:00:00.0

Fig. 4.2 Przykładowy wycinek danych środowiskowych po wczytaniu i połączeniu wszystkich godzin doby (28.04.2025)

4.4.2. Analiza brakujących danych

W ramach eksploracyjnego przetwarzania danych przeprowadzono kontrolę brakujących wartości w każdej z kolumn zbioru. Największy udział braków dotyczy kolumny

STREET – brak danych wystąpił w ponad 23% przypadków. Nieliczne braki (ok. 0,12%) zidentyfikowano również w zmiennych środowiskowych: PRESSURE_AVG, HUMIDITY_AVG oraz TEMPERATURE_AVG. Pozostałe kolumny, w tym współrzędne geograficzne, wartości stężeń pyłów oraz znaczniki czasowe, nie zawierają braków.

Dodatkowo, na podstawie charakteru danych i kontekstu pomiarów, dokonano klasyfikacji typów braków zgodnie z typologią stosowaną w analizie danych:

- **MNAR (Missing Not At Random)** – brakujące wartości w kolumnie STREET wynikają najprawdopodobniej z nieuzupełnienia danych adresowych dla niektórych placówek, zwłaszcza szkół w mniejszych miejscowościach. Nie są one losowe i mogą być zależne od cech konkretnej instytucji.
- **MAR (Missing At Random)** – braki w kolumnach PRESSURE_AVG, HUMIDITY_AVG, TEMPERATURE_AVG pojawiają się tylko w niewielkiej liczbie rekordów, które – jak ustalono – dotyczą jedynie wybranych lokalizacji. Wartości te są prawdopodobnie pomijane ze względu na problemy z transmisją lub błędy czujników, ale ich brak jest zależny od lokalizacji (CITY), więc przypisano im typ MAR.
- **MCAR (Missing Completely At Random)** – nie stwierdzono braków, które można by uznać za całkowicie losowe (MCAR), ponieważ wszystkie zidentyfikowane przypadki można przypisać konkretnym lokalizacjom lub uwarunkowaniom technicznym.

Wyniki przedstawiono graficznie na rysunku Fig. 4.3.2., który obrazuje zarówno liczby bezwzględne, jak i procentowy udział brakujących danych w każdej kolumnie

	Ilość brakujących wartości	Stosunek brakujących do wszystkich	Udział procentowy
STREET	9724	0.2397	23.97 %
PRESSURE_AVG	51	0.0013	0.13 %
HUMIDITY_AVG	50	0.0012	0.12 %
TEMPERATURE_AVG	50	0.0012	0.12 %
NAME	0	0.0000	0.00 %
POST_CODE	0	0.0000	0.00 %
CITY	0	0.0000	0.00 %
LONGITUDE	0	0.0000	0.00 %
LATITUDE	0	0.0000	0.00 %
PM10_AVG	0	0.0000	0.00 %
PM25_AVG	0	0.0000	0.00 %
TIMESTAMP	0	0.0000	0.00 %

Fig. 4.3. Rozkład brakujących danych w kolumnach zbioru

4.4.3. Struktura typów danych

W kolejnym etapie przeanalizowano typy danych przypisane do poszczególnych kolumn zbioru. Zmiennymi tekstowymi (object) są: NAME, STREET, POST_CODE, CITY

oraz TIMESTAMP. Pozostałe kolumny – zawierające dane liczbowe dotyczące lokalizacji geograficznej oraz parametrów środowiskowych – zostały poprawnie rozpoznane jako zmienne typu float64. Przedstawienie aktualnych typów danych w zbiorze zawiera rysunek Fig. 4.4.

	Nazwa kolumny	Typ danych
0	NAME	object
1	STREET	object
2	POST_CODE	object
3	CITY	object
4	LONGITUDE	float64
5	LATITUDE	float64
6	HUMIDITY_AVG	float64
7	PRESSURE_AVG	float64
8	TEMPERATURE_AVG	float64
9	PM10_AVG	float64
10	PM25_AVG	float64
11	TIMESTAMP	object

Fig. 4.4. Typy danych przypisane kolumnom w zbiorze

4.4.4. Rozdzielenie znacznika czasowego na komponenty daty i godziny

W kolejnym kroku przetworzono kolumnę TIMESTAMP, zawierającą datę i godzinę pomiaru, poprzez jej konwersję do formatu datetime oraz rozdzielenie na cztery nowe zmienne: year, month, day i hour. Zabieg ten umożliwia bardziej precyzyjną analizę zmienności warunków środowiskowych w czasie, ułatwiając np. wykrywanie sezonowości lub identyfikację godzin szczytowego zanieczyszczenia. Przekształcone dane przedstawiono na rysunku Fig. 4.5., gdzie widoczne są nowe kolumny czasowe zastępujące pierwotny znacznik TIMESTAMP.

	NAME	STREET	POST_CODE	CITY	LONGITUDE	LATITUDE	HUMIDITY_AVG	PRESSURE_AVG	TEMPERATURE_AVG	PM10_AVG	PM25_AVG	year	month	day	hour
0	SZKO?A PODSTAWOWA IM. MARIANA FALKOWSKIEGO W KRASZEWICACH	UL. SZKOLNA	63-522	KRASZEWICE	18.224090	51.515630	95.120000	1015.110000	8.610000	52.860000	20.200000	2025	4	28	0
1	SZKO?A PODSTAWOWA WE WRZ? SOWICACH	UL. SZKOLNA	32-040	WRZ? SOWICE	19.942820	49.961030	49.040000	988.780000	9.060000	5.780000	5.120000	2025	4	28	0
2	PUBLICZNA SZKO?A PODSTAWOWA NR 2 IM. KAZIMIERZA MA?CZEWSKIEGO W STRZELCACH OPOLSKICH	UL. WAWRZYTCA ?WIERZEGO	47-100	STRZELCE OPOLSKIE	18.314889	50.509431	67.333333	1044.133333	6.133333	1.000000	0.546667	2025	4	28	0
3	ZESP?? SZK?? NR 1 W PSZCZYNIE	UL. KAZIMIERZA WIELKIEGO	43-200	PSZCZYNA	18.945706	49.965883	86.466667	1005.600000	6.333333	9.666667	5.790000	2025	4	28	0
4	ZESP?? SZK?? IM. POWSTA?CW WIELKOPOLSKICH W JANKOWIE PRZYGODZKIM	SZKOLNA	63-421	JANKÓW PRZYGODZKI	17.788907	51.596172	68.133333	1013.033333	8.266667	15.666667	9.246667	2025	4	28	0
5	SZKO?A PODSTAWOWA NR 7 IM. ?O?NIERZY WRZE?NIA W 7 WIKU?ACH	UL. MY?CZENNIK? W OTMIT CIMSKICH	43-229	?WIKLICE	18.989839	49.971987	70.166667	997.166667	7.800000	14.766667	8.420000	2025	4	28	0
6	SZKO?A PODSTAWOWA NR 12 W STUDZIONCE	UL. JORDANA	43-245	STUDZIONKA	18.774985	49.960356	66.433333	1004.233333	8.933333	13.400000	7.540000	2025	4	28	0
7	ZESP?? SZKOLNO-PRZEDSZKOLNY W PIASKU	SZKOLNA	43-211	PIASEK	18.946340	50.009550	55.900000	1004.966667	6.333333	17.033333	10.010000	2025	4	28	0
8	ZESP?? SZKOLNO-PRZEDSZKOLNY W ?YCE	FITELBERGA	43-241	?YKA	18.906757	49.958244	62.166667	1002.966667	5.833333	13.666667	7.526667	2025	4	28	0
9	SZKO?A PODSTAWOWA NR 3 W LUBONIU	ARMII POZNA?	62-030	LUBO?	16.896800	52.348100	87.300000	1013.233333	7.733333	24.566667	15.030000	2025	4	28	0

Fig. 4.5. Przykładowe rekordy po rozdzieleniu kolumny TIMESTAMP na year, month, day i hour

4.4.5. Zarządzanie brakującymi wartościami

W ramach inżynierii cech wykonano imputację brakujących danych w kolumnach zawierających informacje środowiskowe i lokalizacyjne. Uzupełniono następujące kolumny:

- **Wilgotność (HUMIDITY_AVG), ciśnienie (PRESSURE_AVG) oraz temperaturę (TEMPERATURE_AVG)** – poprzez średnią wyliczoną w obrębie tej samej miejscowości (CITY). Takie podejście pozwala zachować charakterystyki lokalnych warunków pogodowych, jednocześnie uzupełniając luki w danych w sposób kontekstowy.
- **Nazwy ulic (STREET)** – zostały zastąpione wartością "nieznana ulica" w przypadkach, gdzie brakowało informacji o adresie. Ponieważ dane te mają charakter opisowy i nie są kluczowe w kontekście modelowania numerycznego, przypisanie takiej wartości pozwala zachować spójność rekordu bez utraty innych informacji.

Po przeprowadzeniu imputacji brakujące wartości zostały całkowicie wyeliminowane ze zmiennych, co potwierdza tabela na rysunku Fig. 4.6. Zawiera ona podsumowanie braków po przeprowadzeniu uzupełnień – dla wszystkich kolumn liczba brakujących rekordów oraz udział procentowy wynosi 0.

	Ilość brakujących wartości	Stosunek brakujących do wszystkich	Udział procentowy
NAME	0	0.0000	0.00 %
STREET	0	0.0000	0.00 %
POST_CODE	0	0.0000	0.00 %
CITY	0	0.0000	0.00 %
LONGITUDE	0	0.0000	0.00 %
LATITUDE	0	0.0000	0.00 %
HUMIDITY_AVG	0	0.0000	0.00 %
PRESSURE_AVG	0	0.0000	0.00 %
TEMPERATURE_AVG	0	0.0000	0.00 %
PM10_AVG	0	0.0000	0.00 %
PM25_AVG	0	0.0000	0.00 %
year	0	0.0000	0.00 %
month	0	0.0000	0.00 %
day	0	0.0000	0.00 %
hour	0	0.0000	0.00 %

Fig. 4.6. Podsumowanie braków danych po przeprowadzonej imputacji

4.4.6. Liczba unikalnych etykiet w kolumnach i podział na kardynalność

W celu oceny różnorodności danych zawartych w zbiorze, przeanalizowano liczbę unikalnych etykiet (wartości) występujących w każdej kolumnie. Takie zestawienie pozwala zidentyfikować zmienne zawierające dużą liczbę kategorii (wysoką kardynalność) oraz te, które cechują się niewielką liczbą wartości unikalnych (niską kardynalność). Dla potrzeb interpretacyjnych przyjęto podział zmiennych na: niską kardynalność (≤ 24 unikalne wartości), wysoką kardynalność (> 24 unikalne wartości).

Zmienne takie jak hour, month, day, year cechują się niską kardynalnością, co wynika z ich charakteru czasowego. Zmienna hour zawiera 24 wartości (pełna doba), pozostałe trzy jedną wartość, ponieważ dane dotyczą jednego konkretnego dnia - 28.04.2025.

Z kolei zmienne lokalizacyjne (NAME, CITY, POST_CODE, LONGITUDE, LATITUDE) oraz środowiskowe (PRESSURE_AVG, HUMIDITY_AVG, TEMPERATURE_AVG, PM10_AVG, PM25_AVG) zawierają dużą liczbę unikatów, co świadczy o wysokiej granularności pomiarów oraz zróżnicowaniu geograficznym i środowiskowym.

```
Liczba etykiet zmiennej NAME: 1787
Liczba etykiet zmiennej STREET: 877
Liczba etykiet zmiennej POST_CODE: 1160
Liczba etykiet zmiennej CITY: 1243
Liczba etykiet zmiennej LONGITUDE: 1787
Liczba etykiet zmiennej LATITUDE: 1781
Liczba etykiet zmiennej HUMIDITY_AVG: 16008
Liczba etykiet zmiennej PRESSURE_AVG: 14683
Liczba etykiet zmiennej TEMPERATURE_AVG: 11210
Liczba etykiet zmiennej PM10_AVG: 10586
Liczba etykiet zmiennej PM25_AVG: 10769
Liczba etykiet zmiennej year: 1
Liczba etykiet zmiennej month: 1
Liczba etykiet zmiennej day: 1
Liczba etykiet zmiennej hour: 24
```

Fig. 4.6. Liczba unikalnych etykiet w poszczególnych kolumnach zbioru danych

4.4.7. Klasyfikacja zmiennych ze względu na typ danych: jakościowe vs ilościowe

W zbiorze danych wyróżniono dwie grupy zmiennych: jakościowe (kategoryczne) oraz ilościowe (numeryczne). Do zmiennych jakościowych zaliczają się m.in. nazwy szkół, miasta, ulice oraz dane czasowe (hour, day, month, year). Z kolei zmienne ilościowe to dane pomiarowe (temperatura, ciśnienie, wilgotność, pyły) oraz współrzędne geograficzne (szerokość i długość geograficzna). Podział ten ułatwia wybór odpowiednich metod analizy i wizualizacji danych. Zestawienie zmiennych według tego kryterium przedstawiono na rysunku Fig. 4.7.

	Zmienne jakościowe (kategoryczne)	Zmienne ilościowe (numeryczne)
0	NAME	LONGITUDE
1	STREET	LATITUDE
2	POST_CODE	HUMIDITY_AVG
3	CITY	PRESSURE_AVG
4	hour	TEMPERATURE_AVG
5	day	PM10_AVG
6	month	PM25_AVG
7	year	

Fig. 4.7. Zestawienie zmiennych jakościowych i ilościowych w zbiorze danych

4.4.8. Usunięcie rekordów z danymi testowymi

W trakcie przeglądu danych zidentyfikowano rekordy zawierające fikcyjne informacje – m.in. nazwę szkoły „TEST SZKOŁA”, adres „ul. Testowa”, miasto „TESTCITY” oraz współrzędne geograficzne spoza terytorium Polski (szerokość geograficzna ok. 1.85°, długość –157.52°, co wskazuje na lokalizację w rejonie Oceanu Spokojnego). Takie dane miały charakter testowy i nie reprezentowały rzeczywistych lokalizacji pomiarowych.

Z uwagi na fakt, że obecność takich rekordów mogłaby zaburzyć analizy statystyczne i geograficzne, zdecydowano się na ich całkowite usunięcie w ramach etapu czyszczenia danych w procesie inżynierii cech. Rysunek Fig. 4.8. przedstawia przykładowe wiersze, które zostały usunięte.

	NAME	STREET	POST_CODE	CITY	LONGITUDE	LATITUDE
22	TEST SZKO? A	ul Testowa	00-000	TESTCITY	-157.529698	1.858685
1758	TEST SZKO? A	ul Testowa	00-000	TESTCITY	-157.529698	1.858685
3489	TEST SZKO? A	ul Testowa	00-000	TESTCITY	-157.529698	1.858685
5186	TEST SZKO? A	ul Testowa	00-000	TESTCITY	-157.529698	1.858685
5211	TEST SZKO? A	ul Testowa	00-000	TESTCITY	-157.529698	1.858685

Fig. 4.8. Przykładowe rekordy z danymi testowymi usunięte ze zbioru danych

4.4.9. Podstawowe statystyki opisowe dla zmiennych

Na początku eksploracyjnej analizy danych (EDA) obliczono podstawowe statystyki opisowe dla wybranych zmiennych numerycznych. Uwzględniono wyłącznie kolumny zawierające dane ciągłe lub potencjalnie istotne dla dalszej analizy, pomijając zmienne typu year, month, day i hour, które mają charakter czasowy lub przyjmują ograniczony zakres wartości.

Użycie funkcji describe() pozwala na szybkie uzyskanie kluczowych informacji statystycznych, takich jak: liczność (count), średnia (mean), odchylenie standardowe (std), wartości skrajne (min, max) oraz kwartyle (25%, 50%, 75%). Technika ta umożliwia wstępną ocenę jakości danych, wykrycie potencjalnych błędów pomiarowych oraz wartości odstających.

Interpretacja wyników:

- **LONGITUDE i LATITUDE** - Średnie wartości odpowiednio 18.97°E i 51.32°N potwierdzają, że dane pochodzą z lokalizacji w Polsce. Zakresy są zgodne z położeniem geograficznym krajowych szkół i przedszkoli.
- **HUMIDITY_AVG (wilgotność względna)** - Średnia wilgotność wynosi 53.59%, co mieści się w typowym zakresie dla warunków atmosferycznych w Polsce. Rozrzut wyników od 0 do 100% wskazuje na dużą zmienność w zależności od lokalizacji i godziny pomiaru.
- **PRESSURE_AVG (ciśnienie atmosferyczne)** - Średnia wartość ciśnienia wynosi 1002.3 hPa, co odpowiada typowemu ciśnieniu na poziomie morza.
- **TEMPERATURE_AVG** - Średnia temperatura to 12.49°C. Minimalna wartość -40°C jest nietypowa i może sugerować błąd pomiarowy lub wartość testową – warto rozważyć jej dalszą weryfikację lub usunięcie.
- **PM10_AVG i PM25_AVG (stężenia pyłów zawieszonych)** - Obie zmienne charakteryzują się znaczną zmiennością. Maksymalne wartości przekraczające odpowiednio 425 µg/m³ (PM10) i 352 µg/m³ (PM2.5) mogą wskazywać na

lokalne źródła zanieczyszczeń, epizody smogowe lub wartości odstające. Konieczna jest dalsza analiza pod kątem ich zgodności z normami jakości powietrza.

Zastosowanie tej techniki było uzasadnione chęcią wczesnego rozpoznania rozkładów danych, ich poprawności oraz identyfikacji nietypowych obserwacji, które mogą zaburzać dalsze analizy. Zestawienie statystyk opisowych przedstawiono na rysunku Fig. 4.9.

	count	mean	std	min	25%	50%	75%	max
LONGITUDE	40545.0	18.986526	2.062086	14.499552	17.007729	19.007124	20.631842	23.844610
LATITUDE	40545.0	51.323641	1.270484	49.302185	50.141099	51.121530	52.384350	54.661981
HUMIDITY_AVG	40545.0	53.586517	22.098541	0.000000	33.200000	53.325000	72.060000	100.000000
PRESSURE_AVG	40545.0	1002.299593	14.319211	920.100000	993.454545	1003.533333	1013.416667	1098.275000
TEMPERATURE_AVG	40545.0	12.493824	7.552295	-40.000000	5.441667	12.633333	19.083333	38.686000
PM10_AVG	40545.0	13.316094	15.665481	0.000000	4.941667	8.790909	15.933333	425.485714
PM25_AVG	40545.0	11.062724	13.000153	0.000000	4.225000	7.500000	13.441667	352.514286

Fig. 4.9. Statystyki opisowe dla zmiennych ilościowych w zbiorze danych

4.4.10. Rozkład zmiennych ilościowych

W ramach eksploracyjnej analizy danych (EDA) przeanalizowano rozkłady wybranych zmiennych środowiskowych – HUMIDITY_AVG, PRESSURE_AVG, TEMPERATURE_AVG, PM10_AVG oraz PM25_AVG. W tym celu wykorzystano histogramy z nałożoną funkcją gęstości jądrowej (KDE), które pozwalają zarówno na ocenę kształtu rozkładu danych, jak i potencjalnych odchyleń od rozkładu normalnego.

Zastosowanie histogramów umożliwia wizualną identyfikację skupisk, rozproszenia, asymetrii, a także wartości odstających. Technika ta jest przydatna w ocenie, czy dane wymagają dalszych przekształceń – np. normalizacji lub standaryzacji – przed ich użyciem w modelowaniu.

Interpretacja wykresów:

- **HUMIDITY_AVG** - rozkład wilgotności jest wyraźnie bimodalny – wiele pomiarów skupia się wokół wartości 40% i 70%, co może świadczyć o różnicach między regionami lub porami dnia. Obecny jest również szczyt przy 100%, który może odpowiadać pełnej wilgotności powietrza.
- **PRESSURE_AVG** - rozkład ciśnienia jest dość symetryczny, skupiony wokół 1000 hPa. Taki rozkład sugeruje brak ekstremalnych warunków pogodowych w dniu pomiaru.
- **TEMPERATURE_AVG** - rozkład temperatury ma kilka pików, co może wskazywać na różnice regionalne lub zmiany temperatury w ciągu dnia. Minimalna wartość -40°C może być wartością odstającą lub błędną.
- **PM10_AVG i PM25_AVG** - rozkłady obu zmiennych są silnie skośne, z dużą liczbą małych wartości i nielicznymi bardzo wysokimi wartościami. Sugeruje to obecność lokalnych źródeł zanieczyszczeń, np. ruch drogowy, przemysł, a także możliwość występowania epizodów smogowych.

Wykresy rozkładu przedstawiono na rysunku Fig. 4.10. Analiza ta stanowi ważny etap przygotowania danych.

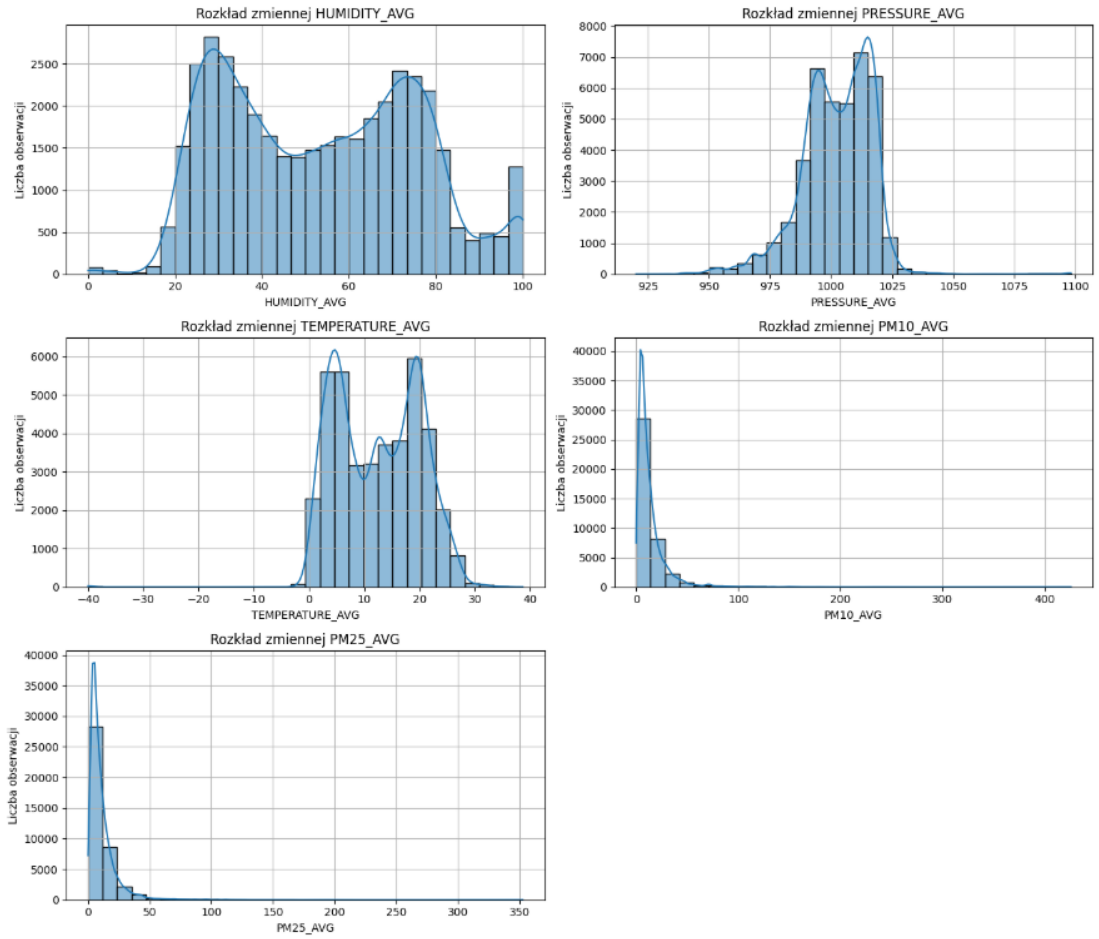

Fig. 4.10. Histogramy rozkładu zmiennych środowiskowych: wilgotności, ciśnienia, temperatury oraz stężeń pyłów PM10 i PM2.5

4.4.11. Analiza wartości odstających – wykresy pudełkowe (boxploty)

W celu identyfikacji potencjalnych wartości odstających przeprowadzono wizualizację pięciu wybranych zmiennych ilościowych za pomocą wykresów pudełkowych (*boxplotów*). Wykresy te umożliwiają szybkie wykrycie obserwacji znajdujących się poza zakresem typowych wartości (tzw. outliers), jak przedstawiono na rysunku Fig. 4.11.

Zastosowanie boxplotów jest szczególnie pomocne w eksploracyjnej analizie danych, ponieważ umożliwia szybką lokalizację potencjalnych anomalii, ocenę asymetrii rozkładów oraz identyfikację zmiennych wymagających oczyszczenia lub transformacji.

Obserwacje:

- **HUMIDITY_AVG** – rozkład stosunkowo symetryczny, bez znacznych wartości odstających, wskazuje na jednolity poziom wilgotności w większości lokalizacji.
- **PRESSURE_AVG** – kilka punktów odstających, ale ogólny rozkład ciśnienia mieści się w typowym zakresie atmosferycznym.
- **TEMPERATURE_AVG** – zauważalne są odstające wartości, szczególnie te bliskie -40°C, co sugeruje możliwy błąd pomiaru lub dane testowe.

- **PM10_AVG i PM25_AVG** – bardzo duża liczba wartości odstających. Wysokie wartości stężeń mogą wskazywać na lokalne zanieczyszczenia, ale również mogą zawierać błędy – warto rozważyć ich dalszą analizę lub oczyszczenie.

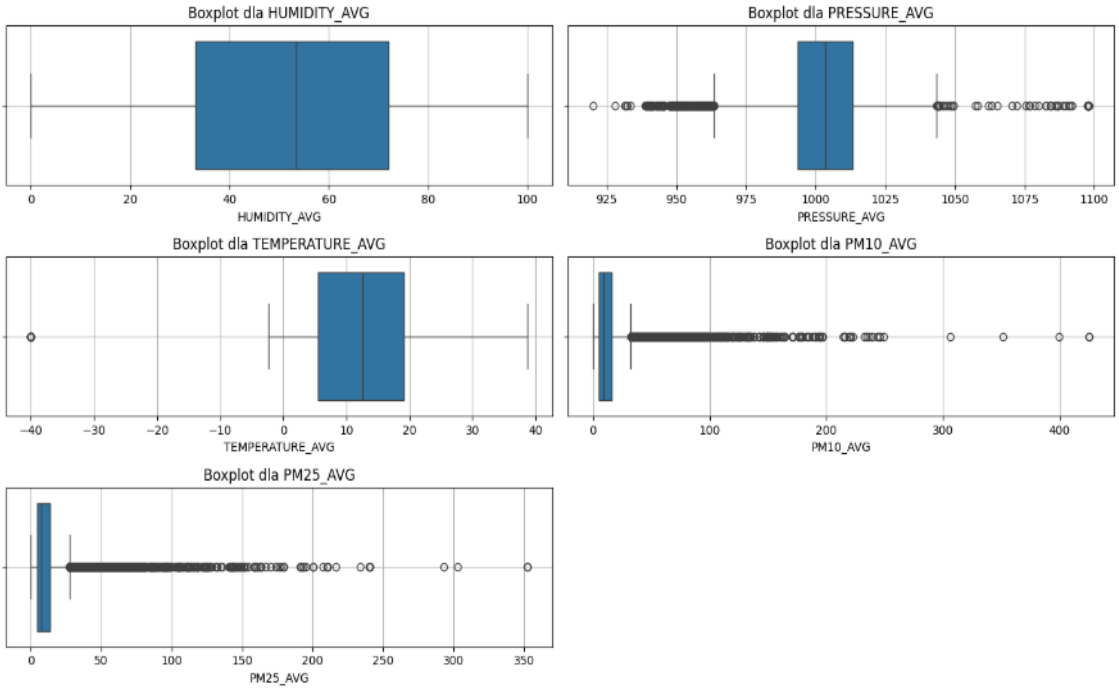

Fig. 4.11. Wykresy pudełkowe (boxploty) dla wybranych zmiennych numerycznych

4.4.12. Analiza współzależności między zmiennymi środowiskowymi

Aby zidentyfikować potencjalne zależności liniowe pomiędzy zmiennymi środowiskowymi, obliczono macierz korelacji współczynnikiem korelacji Pearsona. Wyniki przedstawiono w postaci mapy ciepła (heatmapy), na rysunku Fig. 4.12.

Heatmapa zawiera wartości korelacji między wszystkimi parami zmiennych ilościowych, takich jak: HUMIDITY_AVG, PRESSURE_AVG, TEMPERATURE_AVG, PM10_AVG, PM25_AVG. Skala kolorów pozwala wizualnie ocenić siłę i kierunek zależności – kolor czerwony oznacza silną dodatnią korelację, niebieski – silną ujemną, a biały – brak korelacji.

Interpretacja wyników:

- **PM10_AVG i PM25_AVG** - bardzo silna dodatnia korelacja (0.98), co jest zgodne z oczekiwaniami – oba wskaźniki odnoszą się do stężenia pyłów zawieszonych i często występują razem w podobnych warunkach środowiskowych.
- **HUMIDITY_AVG i TEMPERATURE_AVG** - silna ujemna korelacja (-0.84), sugerująca, że wyższa wilgotność występuje przy niższych temperaturach, co może być efektem warunków nocnych lub specyfiki klimatu.
- **PM10_AVG i TEMPERATURE_AVG / PM25_AVG i TEMPERATURE_AVG** - umiarkowanie ujemne korelacje (około -0.33), które mogą wskazywać na to, że niższe temperatury sprzyjają wyższemu stężeniu pyłów (np. przez większe emisje z ogrzewania budynków).

- **HUMIDITY_AVG i PM10_AVG / PM25_AVG** - dodatnia, ale słaba korelacja (około 0.32–0.34), może sugerować, że wilgotność powietrza ma pewien wpływ na koncentrację zanieczyszczeń, ale nie jest to dominujący czynnik.
- **PRESSURE_AVG** - nie wykazuje istotnych korelacji z innymi zmiennymi – wszystkie współczynniki są bliskie 0.

Dodatkowo warto zauważyć, że zmienne PM10_AVG i PM25_AVG są ze sobą niemal doskonale skorelowane (0.98), co wskazuje na ich bardzo silną współzależność. Może to wynikać z faktu, że cząstki PM2.5 są fizycznie częścią składową frakcji PM10 – dlatego ich poziomy zwykle zmieniają się razem. Taka silna korelacja oznacza, że w modelowaniu predykcyjnym niekoniecznie trzeba uwzględniać obie zmienne jednocześnie, ponieważ mogą wnosić zbliżoną informację (redundancja cech). Wybór jednej z nich – np. PM10_AVG – może być wystarczający przy zachowaniu pełnej reprezentatywności informacji o zanieczyszczeniu powietrza.

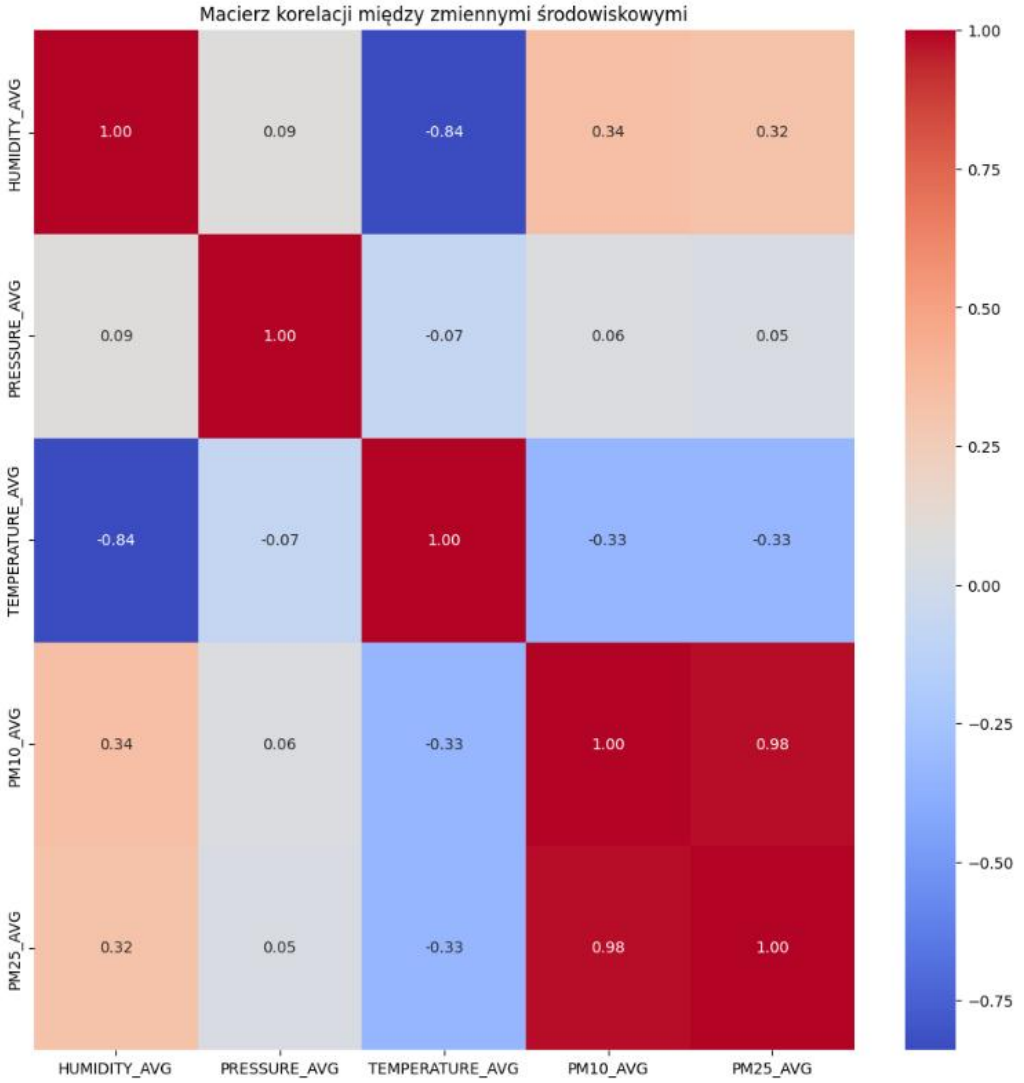

Fig. 4.12. Macierz korelacji między zmiennymi środowiskowymi

4.4.13. Rozkład godzinowy średnich wartości zmiennych środowiskowych

W celu zbadania zmienności parametrów środowiskowych w ciągu doby, przeanalizowano średnie wartości pięciu zmiennych w podziale godzinowym. Na rysunku Fig. 4.13 przedstawiono przebieg dobowy temperatury, ciśnienia, wilgotności oraz stężeń

pyłów PM10 i PM2.5. Wykresy umożliwiają identyfikację regularnych wzorców dziennych oraz potencjalnych zależności pomiędzy zmiennymi.

Wnioski z analizy:

- **HUMIDITY_AVG** - Wilgotność jest najwyższa nocą i we wczesnych godzinach porannych, a najniższa w południe i popołudniu. To typowy dzienny przebieg wilgotności wynikający z ogrzewania powietrza przez Słońce.
- **PRESSURE_AVG** - Ciśnienie atmosferyczne wykazuje lekkie wahania, ze spadkiem od godzin porannych do wieczora – zgodne z dobowym rytmem cyrkulacji atmosferycznej.
- **TEMPERATURE_AVG** - Wzrost temperatury od godzin porannych do popołudnia i spadek wieczorem to typowy dzienny przebieg promieniowania słonecznego.
- **PM10_AVG i PM25_AVG** - Największe stężenia pyłów obserwowane są w nocy i późnym wieczorem, co może wynikać z ograniczonej cyrkulacji powietrza i emisji ze źródeł grzewczych. Popołudniowy spadek może być efektem dyspersji i wyższej temperatury.

Tego typu analiza pozwala lepiej zrozumieć rytmy dzienne w danych środowiskowych, co może mieć znaczenie przy modelowaniu predykcyjnym oraz w ocenie ryzyka ekspozycji na zanieczyszczenia.

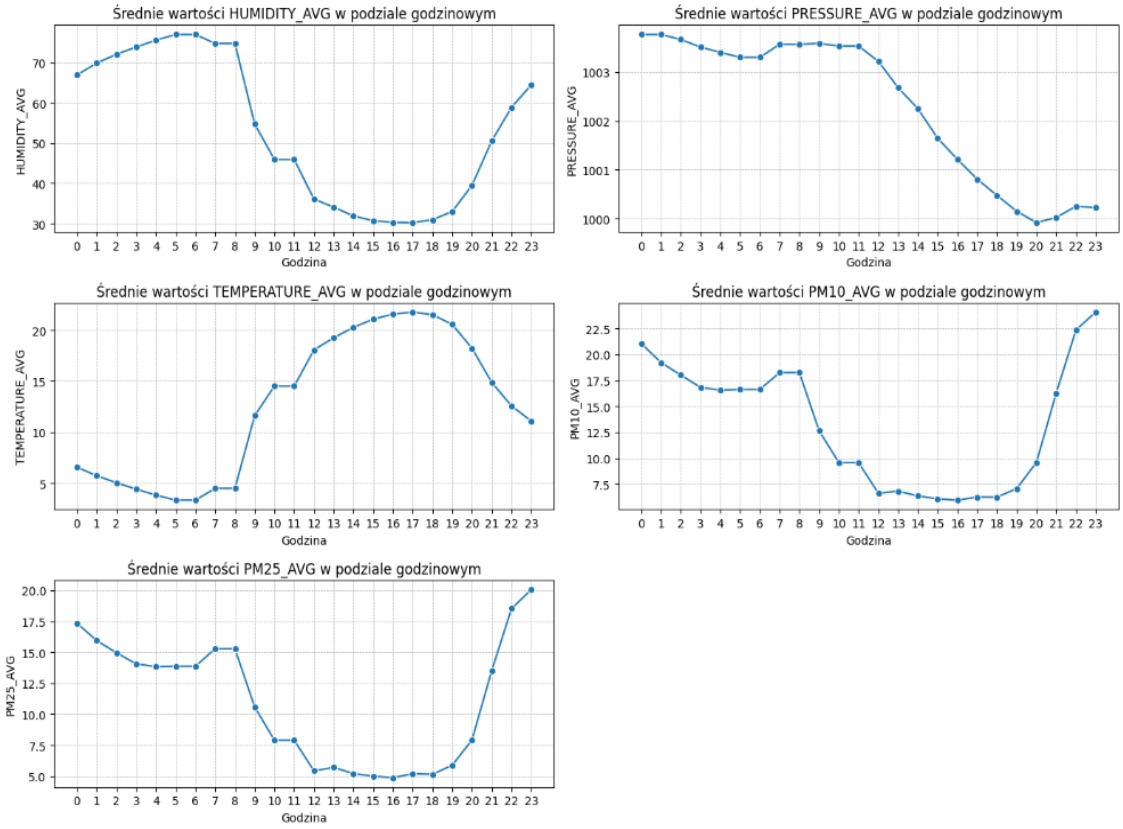

Fig. 4.13. Średnie wartości wybranych zmiennych środowiskowych w podziale godzinowym

4.4.14. Znormalizowane średnie wartości zmiennych środowiskowych w ciągu doby

W celu jednoczesnego przedstawienia trendów dobowych dla różnych zmiennych środowiskowych o odmiennych jednostkach miary (np. hPa, °C, µg/m³) przeprowadzono

ich normalizację (skalowanie Min-Max). Dzięki temu możliwe było umieszczenie wszystkich przebiegów na wspólnej skali (od 0 do 1) i ocena współzależności zmian w czasie. Wszystko zamieszczono na wykresie Fig. 4.14.

Interpretacja zależności:

- **Temperatura vs zanieczyszczenia (PM10 i PM2.5)** - Wraz ze wzrostem temperatury (TEMPERATURE_AVG) – szczególnie od godzin porannych do popołudnia (8:00–17:00) – stężenia pyłów (PM10_AVG i PM25_AVG) spadają. Oznacza to odwrotną korelację, co może wynikać z:
  - lepszej wentylacji atmosferycznej w ciągu dnia (cieplejsze powietrze sprzyja unoszeniu się zanieczyszczeń),
  - zmniejszonej emisji domowych źródeł ogrzewania w cieplejszych godzinach.
- **Wilgotność vs temperatura** - Widoczna jest silna odwrotna zależność – gdy temperatura rośnie, wilgotność (HUMIDITY_AVG) spada.
- **PM10 i PM2.5 vs wilgotność** - Lekka dodatnia zależność – wyższe poziomy zanieczyszczeń występują przy wyższej wilgotności (rano i wieczorem). Może to wskazywać na pogorszone warunki dyspersji zanieczyszczeń przy bardziej wilgotnym, stabilnym powietrzu.
- **PM10 i PM2.5** - zgodność trendów: Obie zmienne mają niemal identyczny przebieg, co potwierdza ich bardzo wysoką korelację i wspólne źródła (np. ruch drogowy, spalanie).

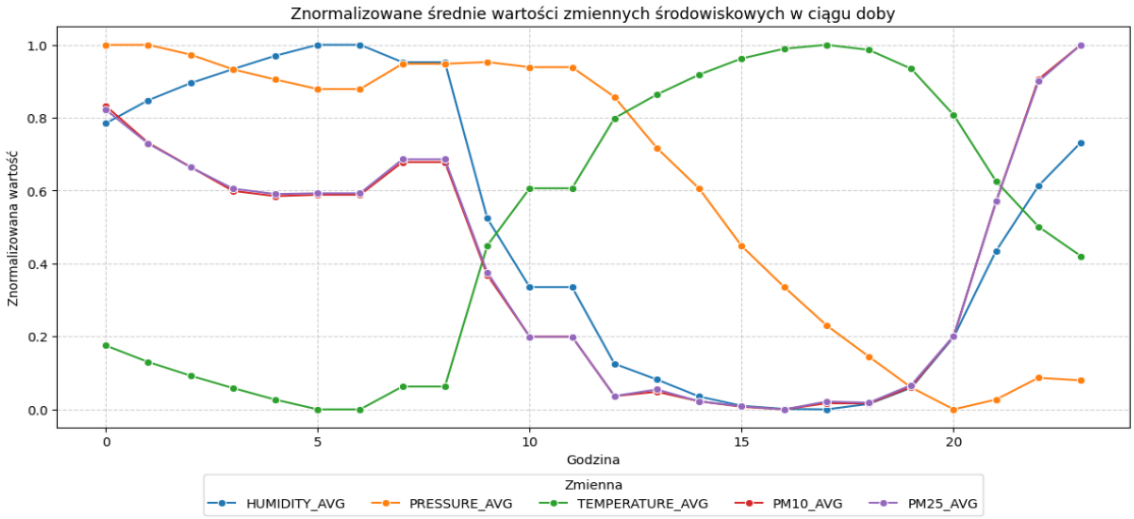

Fig. 4.14. Znormalizowane średnie wartości zmiennych środowiskowych w ciągu doby

4.4.15. Mapa średniego poziomu PM10 w miastach

W celu wizualizacji przestrzennego zróżnicowania stężeń pyłów zawieszonych PM10 przeprowadzono analizę średnich wartości dla każdego miasta na podstawie lokalizacji geograficznych. Wyniki przedstawiono na mapie punktowej (rysunek Fig. 4.16), wygenerowanej z wykorzystaniem biblioteki plotly.express, w szczególności funkcji scatter_map, która umożliwia prezentację danych geolokalizacyjnych na interaktywnej mapie.

Dla każdej miejscowości w zbiorze danych obliczono:

- średnie stężenie pyłu PM10 (PM10_AVG),
- średnią temperaturę (TEMPERATURE_AVG),

- średnią wilgotność (HUMIDITY_AVG), a następnie przypisano pierwsze (reprezentatywne) współrzędne geograficzne (LATITUDE, LONGITUDE). Ostateczna mapa przedstawia zatem uśrednione warunki środowiskowe w skali miejskiej.

Interpretacja mapy:

- Najwyższe stężenia PM10 (zaznaczone kolorem żółtym i zielonym) koncentrują się głównie w centralnej Polsce, w tym na obszarach wokół Warszawy, a także lokalnie w rejonie Poznania, Dolnego Śląska i Śląska. Może to wskazywać na duży udział emisji komunikacyjnej i przemysłowej w tych obszarach.
- Niższe poziomy PM10 (ciemnofioletowe punkty) dominują na terenach wschodniej i północnej Polski, co może wynikać z mniejszego zagęszczenia miast i mniejszej liczby źródeł emisji.
- Widoczna jest korelacja przestrzenna – największe stężenia występują głównie w obszarach zurbanizowanych i przemysłowych, co może świadczyć o wpływie emisji antropogenicznych (ruch drogowy, ogrzewanie, przemysł).
- Rzadsze występowanie skrajnych wartości w rejonach południowo-wschodnich może sugerować lepszą cyrkulację powietrza lub mniejsze natężenie ruchu.

Mapa dostarcza istotnych informacji na temat przestrzennego rozkładu zanieczyszczeń powietrza i może być punktem wyjścia do dalszych analiz – np. identyfikacji obszarów ryzyka zdrowotnego, planowania działań naprawczych lub porównań sezonowych.

Jedną z kluczowych zalet wykorzystania biblioteki plotly.express jest możliwość interaktywnej eksploracji danych. Po najechaniu kursorem na dowolny punkt na mapie użytkownik otrzymuje szczegółowe informacje o danym mieście – jak pokazano na ilustracji poniżej (rys. Fig. 4.15). Taka prezentacja pozwala szybko zlokalizować obszary o wysokim stężeniu zanieczyszczeń i ocenić jednocześnie inne zmienne środowiskowe dla danego miejsca.

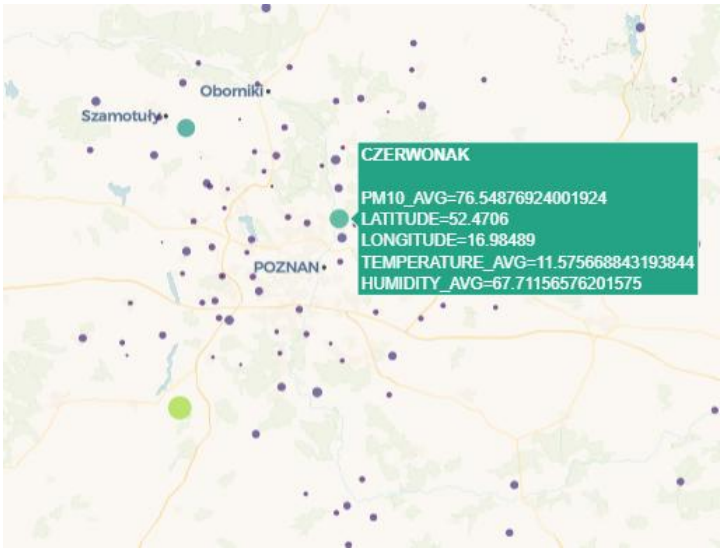

Fig. 4.15. Interaktywny podgląd danych dla miasta Czerwonak – możliwość odczytania szczegółów po najechaniu kursorem na punkt na mapie

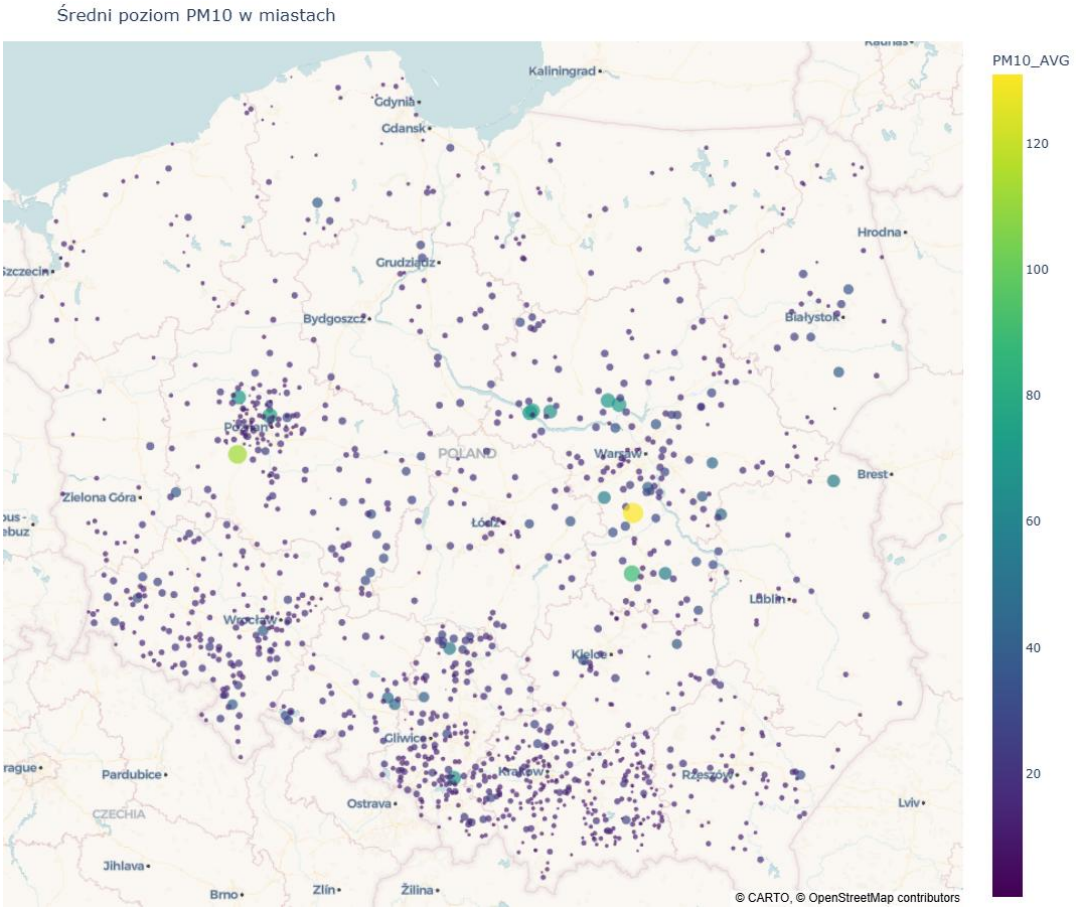

Fig. 4.16. Średni poziom stężenia pyłu PM10 w miastach – wizualizacja punktowa z użyciem plotly.express.scatter_map

4.4.16. Analiza współzależności miasto–godzina–PM10 (heatmapa)

W celu zidentyfikowania lokalnych wzorców emisji zanieczyszczeń wykonano analizę współzależności pomiędzy miastem (CITY)**,** godziną dnia (hour) a średnim poziomem stężenia pyłu zawieszonego PM10. Dane przedstawiono w postaci heatmapy (rysunek Fig. 4.17), gdzie kolor reprezentuje przeciętny poziom PM10 w danym mieście o konkretnej godzinie.

Do analizy wybrano 20 miast o najwyższej średniej dobowej wartości PM10 w celu lepszego uwidocznienia punktów problematycznych – takie podejście pozwala skupić się na lokalizacjach o największym znaczeniu środowiskowym i zdrowotnym.

Interpretacja:

- W wielu przypadkach (np. Bieniędzice, Jasieniec, Chodkowo-Działki, Lutkówka) najwyższe stężenia występują w godzinach nocnych i porannych (0:00–7:00), co może wskazywać na lokalne źródła emisji, takie jak domowe piece grzewcze, szczególnie przy braku wiatru i inwersji temperatury.
- W innych lokalizacjach, np. Czerwoniak, Baborowo czy Modrze, podwyższone poziomy występują także w ciągu dnia, co może być związane z ruchem drogowym lub przemysłem.
- Widoczna jest znaczna zmienność godzinowa w ramach jednego miasta, co wskazuje na silne działanie czynników lokalnych, zarówno czasowych (np. godziny szczytu), jak i przestrzennych (lokalne emisje).

- Częściowe braki danych (puste pola) mogą wynikać z braku pomiarów w danej godzinie lub zerowy poziom zanieczyszczeń.

Wizualizacja została wykonana przy użyciu biblioteki seaborn (sns.heatmap) po agregacji danych do poziomu CITY–hour. Technika ta jest bardzo użyteczna w kontekście monitorowania zanieczyszczeń, identyfikacji epizodów smogowych oraz planowania działań zaradczych w najbardziej narażonych lokalizacjach.

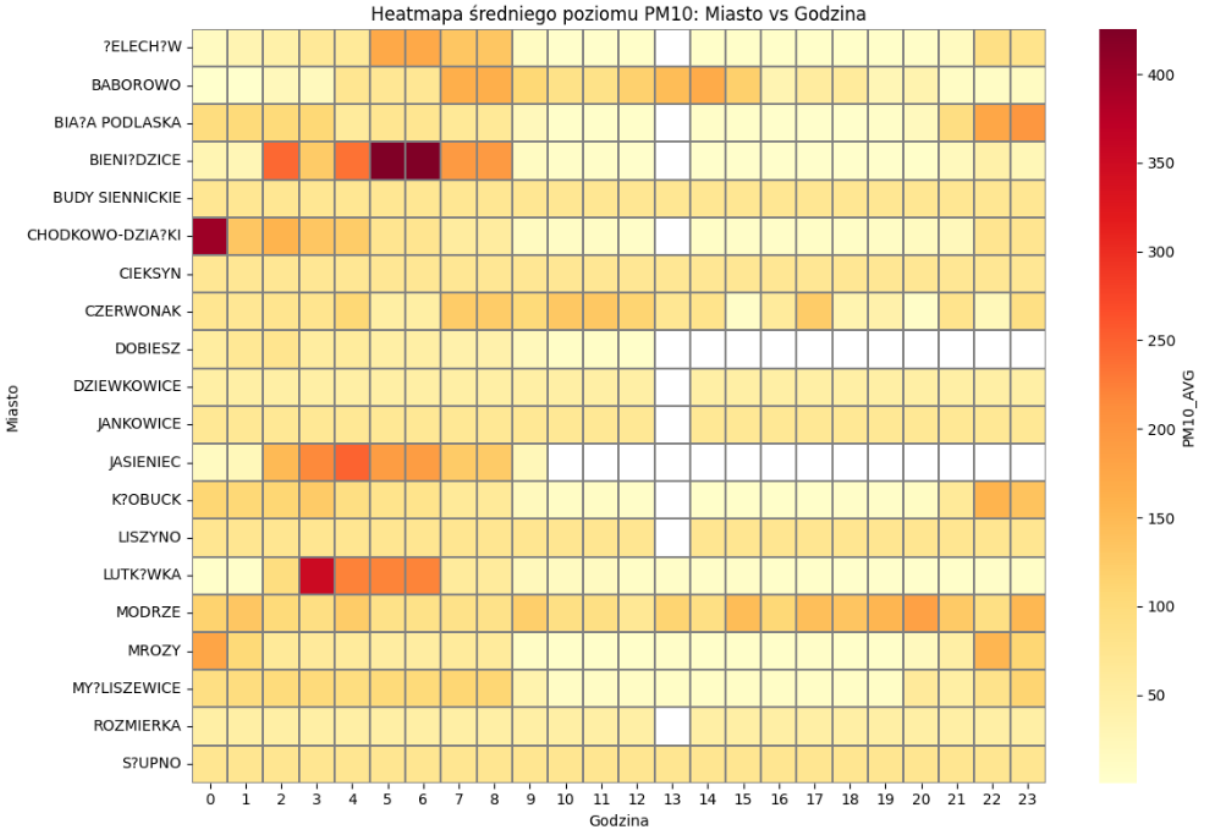

Fig. 4.17. Heatmapa średniego poziomu PM10 w zależności od miasta i godziny – analiza dla 20 miast z najwyższym średnim stężeniem

4.5. Wybór zmiennej docelowej (TARGET) do modelu uczenia nadzorowanego

Na potrzeby budowy modelu uczenia nadzorowanego zaproponowano wykorzystanie zmiennej **PM10_AVG** jako zmiennej docelowej (TARGET). Wartość ta reprezentuje średnie stężenie pyłu zawieszonego PM10 – jednego z najważniejszych parametrów opisujących jakość powietrza. Wybór tej zmiennej uzasadniony jest zarówno ze względu na jej znaczenie środowiskowe i zdrowotne, jak i dostępność oraz ciągłość danych w zbiorze.

Stężenie PM10 jest powszechnie uznawane za istotny wskaźnik jakości powietrza w analizach prowadzonych przez instytucje publiczne oraz organizacje zdrowotne. Jego wartość zmienia się dynamicznie w zależności od szeregu czynników meteorologicznych, lokalizacji geograficznej i pory dnia, co czyni ją odpowiednim kandydatem do modelowania z użyciem regresji (dla przewidywania wartości ciągłych) lub klasyfikacji (po wcześniejszym binaryzowaniu lub podziale na klasy jakości powietrza).

Dodatkowo, zmienna PM10_AVG była wielokrotnie analizowana w eksploracyjnej części projektu (EDA), a jej rozkład, wartości odstające i współzależności z innymi zmiennymi zostały dobrze poznane. Dzięki temu możliwe jest przygotowanie spójnego, dobrze uzasadnionego modelu predykcyjnego.

4.6. Wybór zmiennych wejściowych (FEATURES) do predykcji zmiennej TARGET

Na podstawie wiedzy domenowej oraz wyników eksploracyjnej analizy danych (EDA), do wyznaczenia zmiennej docelowej PM10_AVG zaproponowano następujący zestaw zmiennych wejściowych (FEATURES):

- **TEMPERATURE_AVG** – temperatura powietrza istotnie wpływa na koncentrację pyłów. Wyższe temperatury sprzyjają dyspersji zanieczyszczeń, natomiast niskie wartości mogą być związane z emisją z domowych źródeł grzewczych.
- **HUMIDITY_AVG** – wilgotność powietrza koreluje ze stężeniami PM10. Zwiększona wilgotność może ograniczać rozprzestrzenianie się pyłów w atmosferze i wskazywać na stabilne warunki sprzyjające ich kumulacji.
- **PRESSURE_AVG** – ciśnienie atmosferyczne, choć nie wykazało silnych korelacji w analizie EDA, może wpływać pośrednio na pionową cyrkulację powietrza i być pomocne w budowie modelu.
- **LATITUDE i LONGITUDE** – współrzędne geograficzne umożliwiają ujęcie przestrzennych różnic w jakości powietrza, które mogą być determinowane m.in. stopniem urbanizacji, obecnością przemysłu czy strukturą zabudowy.
- **HOUR** – pora dnia ma silny wpływ na stężenie PM10. W analizie wykazano, że największe wartości notowane są w godzinach nocnych i porannych, co może być związane z intensyfikacją emisji lokalnych.
- **CITY** – zmienna kategoryczna opisująca lokalizację. Może pełnić rolę wskaźnika warunków lokalnych (np. obecności przemysłu, stopnia urbanizacji). W przypadku modelowania regresyjnego wymaga zakodowania (np. one-hot encoding).

Powyższy zestaw zmiennych uwzględnia zarówno parametry meteorologiczne, jak i zmienne lokalizacyjne i czasowe, co zapewnia modelowi odpowiedni kontekst do dokładnego przewidywania poziomu zanieczyszczeń. Wybór tych cech oparty został na interpretacji fizycznej i środowiskowej, a nie na technikach obliczeniowych (np. selekcji cech), co było zgodne z wymaganiami zadania.

5. Podsumowanie

Przeprowadzona analiza danych środowiskowych z systemu ESA umożliwiła kompleksową ocenę jakości powietrza w otoczeniu placówek edukacyjnych. Wykorzystanie eksploracyjnej analizy danych (EDA) oraz inżynierii cech (FE) pozwoliło zidentyfikować zależności czasowe i przestrzenne, a także wskazać czynniki wpływające na poziom stężenia pyłów PM10.

Dzięki zastosowaniu bibliotek Python takich jak Seaborn czy Plotly możliwa była efektywna wizualizacja danych oraz ocena ich struktury i jakości. W oparciu o analizę korelacji i obserwacje dobowych trendów wytypowano zmienne istotne dla dalszego modelowania predykcyjnego.

Dane z ESA, ze względu na wysoką częstotliwość pomiarów, dużą liczbę lokalizacji i oznaczenia czasowe, stanowią wartościowe źródło informacji w kontekście analiz środowiskowych, prognozowania jakości powietrza oraz podejmowania decyzji wspierających ochronę zdrowia publicznego.