

# Analiza wyników z EDA oraz raportu ydata\_profiling

## Podstawowe statystyki zmiennych numerycznych

### Średnie wartości dla zmiennych:

- **Temperatura:** Średnia wynosi 25.46°C, a rozkład temperatur mieści się między 3.5°C a 46.2°C.
- **Wilgotność:** Średnia 60.07%, przy maksymalnej wilgotności 100%.
- **PM2.5 i PM10:** Średnie wynoszą odpowiednio 29.91  $\mu\text{g}/\text{m}^3$  i 40.00  $\mu\text{g}/\text{m}^3$ , z dużym zakresem zmienności.
- **Dwutlenek azotu (NO2):** Występują wartości negatywne, co może wskazywać na błędy w danych.
- **Gęstość zaludnienia:** Średnia wynosi 299.95 osób/km<sup>2</sup>, co sugeruje, że dane obejmują różne obszary o zróżnicowanym zaludnieniu.

## Rozkład zmiennej kategorycznej (Air Quality)

### Najczęściej występują klasy:

- **Good** (2000 próbek, 40% danych).
- **Moderate** (1500 próbek, 30% danych).
- **Poor** (1000 próbek, 20% danych).
- **Hazardous** (500 próbek, 10% danych).

Klasy są wystarczająco zróżnicowane, co umożliwia sensowną klasyfikację.

## Brakujące wartości

Brakujące wartości nie występują w żadnej kolumnie, co eliminuje potrzebę uzupełniania braków.

## Macierz korelacji

- Silna dodatnia korelacja między **PM2.5** i **PM10** (0.987): Te zmienne mogą być wzajemnie redundantne.
- Niska lub brak korelacji między zmiennymi numerycznymi a gęstością zaludnienia czy odległością od stref przemysłowych:
  - Może sugerować, że inne czynniki wpływają na poziomy zanieczyszczeń.
- **Temperatura** i **Wilgotność** wykazują bardzo niską korelację z zanieczyszczeniami, co wskazuje na słaby wpływ tych czynników.

## Wnioski z analizy EDA i raportu ydata\_profiling

### Dane odstające:

- Obserwacje w zmiennych takich jak NO2 zawierają wartości negatywne, które są niefizyczne i mogą być błędami w danych.
- PM2.5 i PM10 zawierają skrajne wartości, które mogą wymagać analizy.

### Zależności między zmiennymi:

- PM2.5 i PM10 są silnie skorelowane, co sugeruje, że w analizie można uwzględnić tylko jedną z tych zmiennych.

### Potencjalne problemy z danymi:

- Wartości negatywne w zmiennych takich jak NO2 mogą wymagać usunięcia lub korekty.
- Konieczność uwzględnienia niezależnych czynników w analizie jakości powietrza, ponieważ odległość od stref przemysłowych i gęstość zaludnienia wykazują niewielką korelację z zanieczyszczeniami.