

## 9.2. Wstępne przetwarzanie danych: czyszczenie, brak i wartości odstające

iBigWorld:  
Innovations for Big Data in a Real World

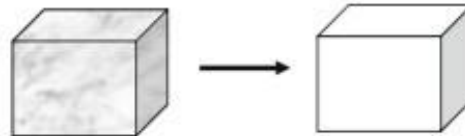
**Disclaimer:** Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the National Agency (NA). Neither the European Union nor NA can be held responsible for them.



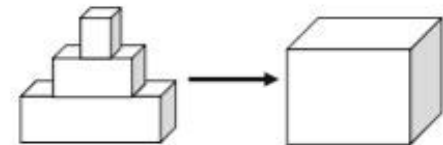
# Cel wstępnego przetwarzania danych:

uzyskanie ostatecznych zbiorów danych, które można uznać za poprawne i przydatne dla dalszych algorytmów eksploracji danych

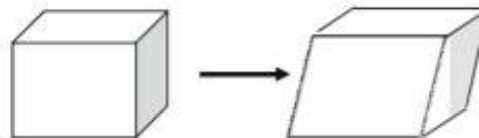
Data cleaning



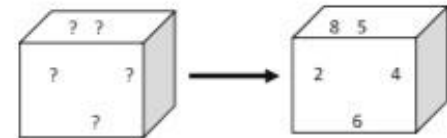
Data normalization



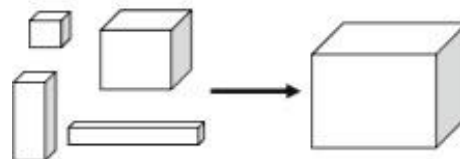
Data transformation



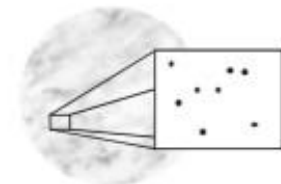
Missing values imputation



Data integration



Noise identification



# Wstępne przetwarzanie danych

W zależności od zestawu danych, samo wstępne przetwarzanie danych może zająć 10-60% całkowitego czasu i wysiłku całego procesu analizy danych



# Czyszczenie danych

Czyszczenie danych to proces poprawiania lub usuwania nieprawidłowych, uszkodzonych, zniekształconych, duplikatów lub niekompletnych danych w zbiorze danych.



## Czyszczenie danych

---

Jeśli dane są błędne, wyniki i algorytmy są niewiarygodne, nawet jeśli wydają się poprawne.

Nie ma jednego bezwzględnego sposobu na określenie dokładnych kroków w procesie czyszczenia danych, ponieważ procesy różnią się w zależności od zestawu danych.

# Czyszczenie danych

Jednak ważne jest, aby utworzyć szablon dla procesu czyszczenia danych, aby zawsze wiedzieć, że robisz to dobrze.



## Jak czyścić dane?

---

**Krok 1:** Usuń  
duplikaty lub  
nieistotne  
obserwacje

**Krok 2:** Napraw  
błędy strukturalne

**Krok 3:** Odfiltruj  
niepożądane  
wartości  
odstające

**Krok 4:** Przetwórz  
brakujące dane

**Krok 5:** Inspekcja  
i kontrola jakości



## Składniki jakości danych

---

- **Ważność.** Zakres, w jakim Twoje dane są zgodne z określonymi regułami biznesowymi lub ograniczeniami.
- **Precyzja.** Upewnij się, że Twoje dane są zbliżone do prawdziwych wartości.
- **Kompletność.** Stopień informacji wszystkich niezbędnych danych.
- **Sekwencja.** Upewnij się, że Twoje dane są spójne w jednym zestawie danych i/lub wielu zestawach danych.



## Składniki jakości danych

---

**Jednolitość.** Stopień, w jakim dane są określone przy użyciu tej samej jednostki miary.

## Korzyści z czyszczenia danych

---

- Wyeliminuj błędy, gdy zaangażowanych jest wiele źródeł danych.
- Mniej błędów oznacza szczęśliwszych klientów i mniej sfrustrowanych pracowników.
- Możliwość mapowania różnych funkcji i do czego wykorzystywane są Twoje dane.

## Korzyści z czyszczenia danych

---

- Śledzenie błędów i lepsze raportowanie, aby zobaczyć, skąd pochodzą błędy, co ułatwia naprawę błędnych lub uszkodzonych danych w przyszłych aplikacjach.
- Wykorzystanie narzędzi do czyszczenia danych przyczyni się do wydajniejszej działalności biznesowej i szybszego podejmowania decyzji

# Brakujące wartości

---

Wszystkie rzeczy są równe, więcej informacji jest prawie zawsze lepsze.



# Ogólne kryteria wyboru wartości podstawień dla brakujących danych

---

1. Zastęp brakującą wartość stałą określoną przez analityka.

2. Zastęp brakującą wartość polem średniej (dla zmiennych numerycznych) lub trybu (dla zmiennych kategoryzowanych).

# Ogólne kryteria wyboru wartości podstawień dla brakujących danych

---

3. Zastąp brakujące wartości wartością wygenerowaną losowo z obserwowanego rozkładu zmiennej.

4. Zastąp brakujące wartości wartościami imputowanymi na podstawie innych cech rekordu.

# GRAFICZNE METODY OKREŚLANIA WARTOŚCI ODSTAJĄCYCH

---

Wartości odstające to wartości ekstremalne, które są sprzeczne z trendem pozostałych danych.

Identyfikacja wartości odstających jest ważna, ponieważ mogą one reprezentować błędy we wprowadzaniu danych.

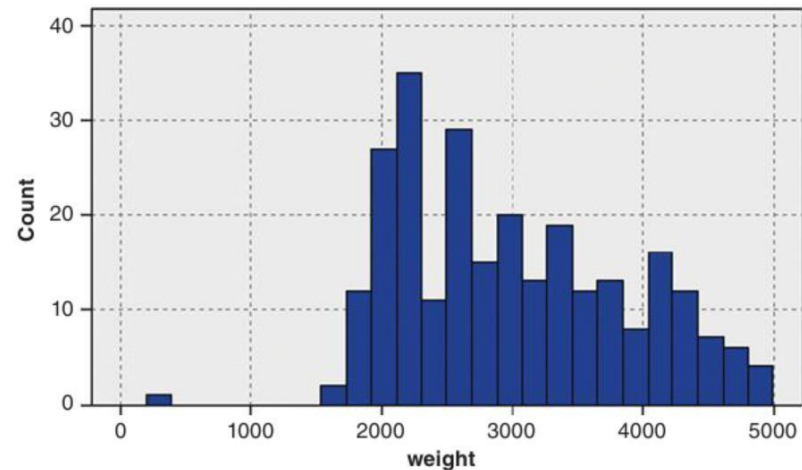
Jeśli wartość odstająca jest prawidłowym punktem danych, a nie błędem, niektóre metody statystyczne są wrażliwe na obecność wartości odstających i mogą dawać niewiarygodne wyniki.



# GRAFICZNE METODY OKREŚLANIA WARTOŚCI ODSTAJĄCYCH

1. histogram; Zwróć uwagę na coś dziwne w rozkładzie częstości?

Brand	Frequency
USA	1
France	1
US	156
Europe	46
Japan	51

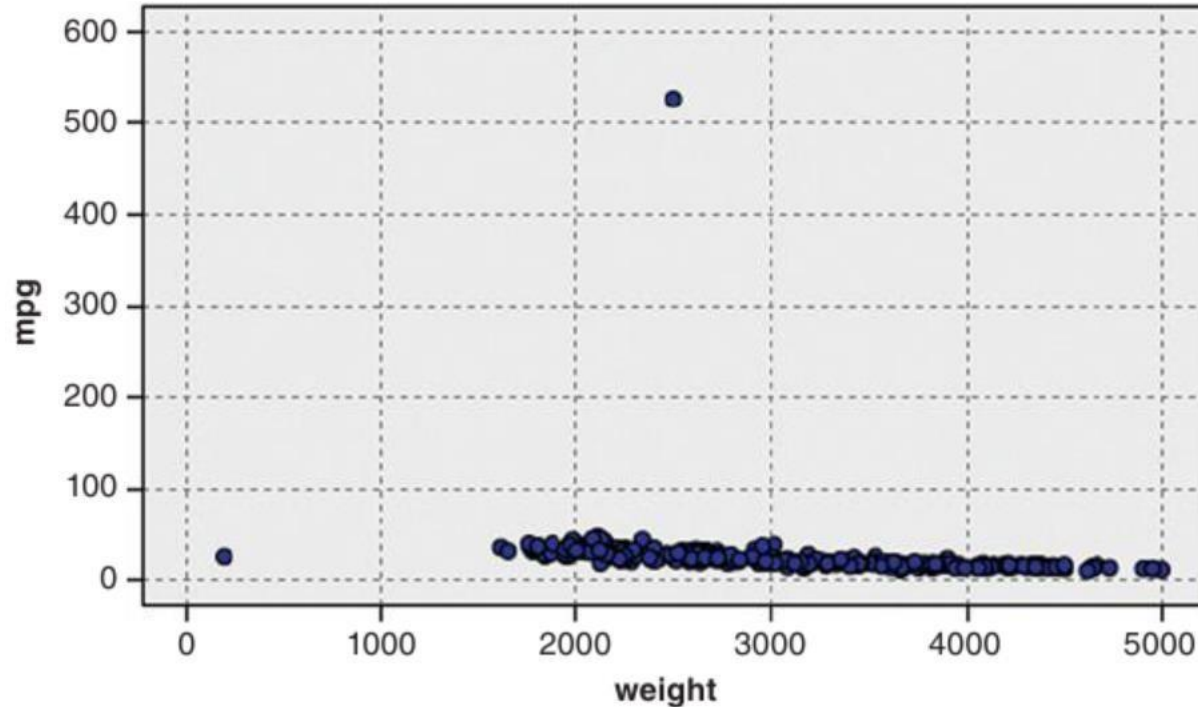


Histogram masy samochodu: czy potrafisz dostrzec wartość odstającą?



# GRAFICZNE METODY OKREŚLANIA WARTOŚCI ODSTAJĄCYCH

## 2. Wykres rozrzutu;



# Niezawodne metody OKREŚLANIA WARTOŚCI ODSTAJĄCYCH

---

użyj rozstępu międzykwartylowego (rozstępu międzykwartylowego, IQR)  
 $IQR = Q3 - Q1$  jest obliczane i może być interpretowane jako reprezentujące środkowe 50% danych.

Rozstęp międzykwartylowy (IQR) jest miarą zmiany, która jest znacznie bardziej wiarygodna niż odchylenie standardowe.



## Usuwanie szumów: używanie

---

**metody przetwarzania danych**, zwłaszcza jeśli ma to wpływ na etykietowanie instancji. Uważa się, że nawet częściowa korekcja szumu jest przydatna, ale jest to trudne zadanie, które zwykle ogranicza się do niewielkich ilości szumu

**filtry szumów**, które identyfikują i usuwają zaszumione przypadki w danych treningowych i nie wymagają modyfikacji techniki eksploracji danych

## Redukcja wymiarowości

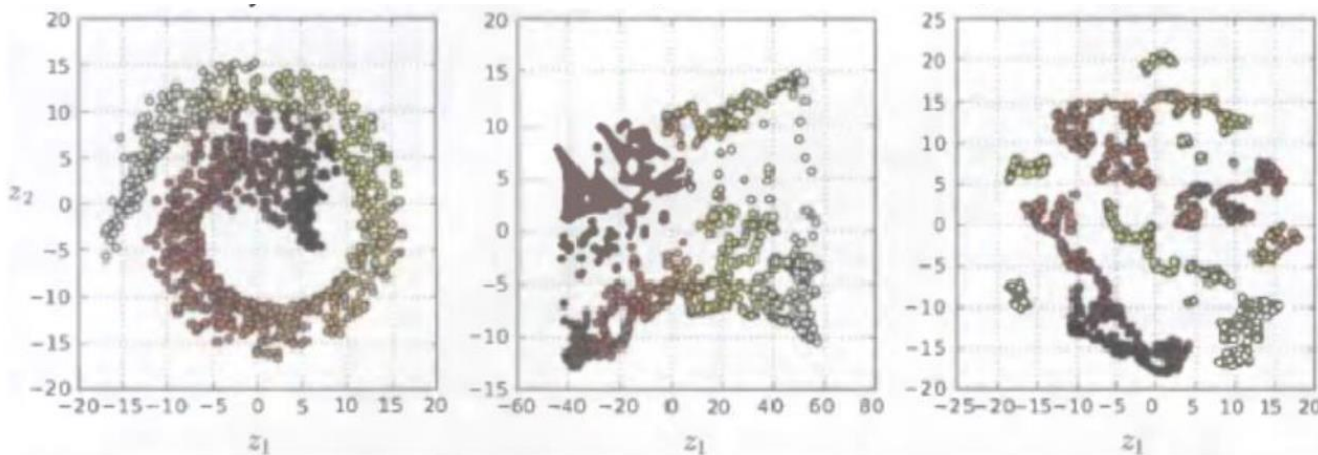
---

Gdy zbiory danych stają się duże pod względem liczby zmiennych predykcyjnych lub liczby instancji, algorytmy eksploracji danych stają przed problemem wymiarowości.

Jest to poważny problem, ponieważ utrudni wydajność większości algorytmów eksploracji danych wraz ze wzrostem kosztów obliczeniowych.

# Redukcja wymiarowości: użycie

- **Analiza czynników**
- **Metoda głównych składników** (analiza głównych składników, PCA)
- **Nieliniowe relacje** między zmiennymi (LLE, ISOMAP i pochodne)



---

**Dziękuję za uwagę!**