POLITECHNIKA WROCŁAWSKA
WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI

METODY ANALIZY I EKSPLORACJI DANYCH

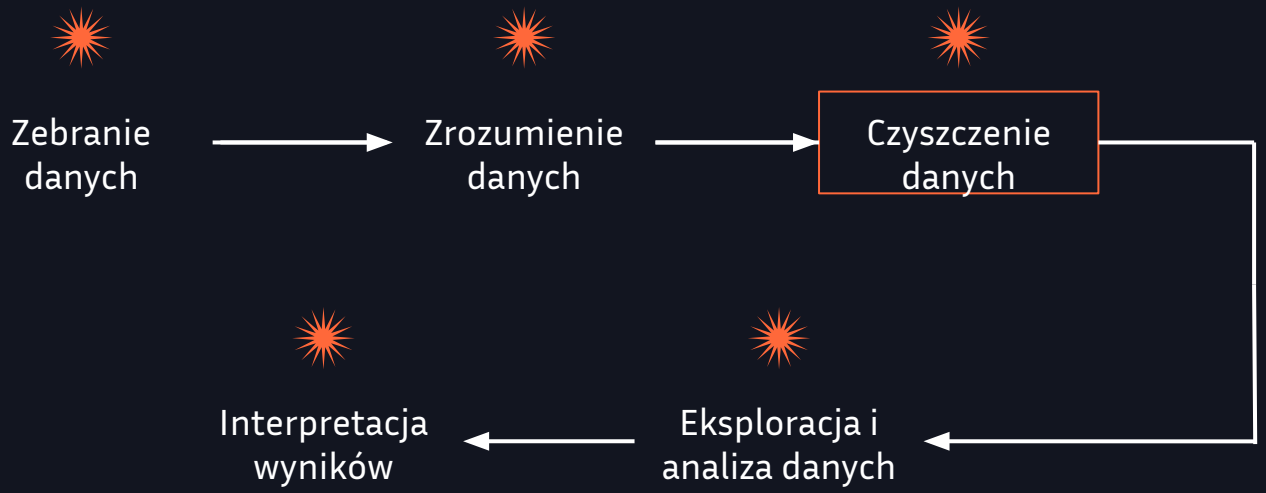
Wykład 4 - Analiza jakości danych

DR INŻ. AGATA MIGALSKA

00

CEL I MOTYWACJA

ETAPY PROJEKTU EKSPLORACJI DANYCH



OCENA JAKOŚCI DANYCH

1. **Trafność**: dane powinny spełniać wymagania dotyczące zamierzonego zastosowania.
2. **Dostępność:** dane muszą być dostępne do wykorzystania.
3. **Dokładność**: wszelkie opisane dane muszą być zbliżone do prawdziwych wartości.
4. **Kompletność**: dane nie powinny zawierać brakujących wartości ani brakujących rekordów.
5. **Aktualność**: dane powinny być aktualne.
6. **Spójność**: dane powinny mieć format zgodny z oczekiwaniami w jednym i/lub wielu zestawach danych.

PROBLEMY W RZECZYWISTYCH DANYCH

- Dane są niepełne
  - brakujące wartości
- Dane są zaszumione
  - Błędy pomiaru
  - Błędy systemowe
  - Obserwacje odstające
  - np. wiek = -10
- Dane są niespójne
  - wiek vs data urodzenia
  - skale ratingowe np. 5 i 7 punktowa

PRZYCZYNY PROBLEMÓW

- Problemy mogą nastąpić w trakcie zbierania danych, przesyłania i procesowania
- Błędy ludzkie, problemy sprzętowe, niepoprawne lub niedziałające oprogramowanie
- Zmiany w czasie
  - zaktualizowane sensory o innych zakresach i możliwościach
  - zaktualizowana ankieta
  - etc.

PRZYGOTOWANIE DANYCH

- Potencjalne problemy z danymi
  - Brakujące dane
  - Błędy systemowe
  - Niespójności
- Przygotowanie danych do procesu eksploracji danych
  - integracja wielu źródeł danych
  - czyszczenie
  - transformacja
  - redukcja
- Bez dobrze przygotowanych danych nie da się uzyskać sensownych rezultatów z eksploracyjnej analizy danych

01

BRAKUJĄCE DANE

02

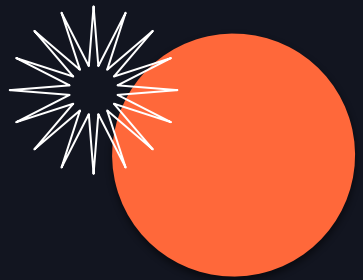
WARTOŚCI ODSTAJĄCE

03

SZUM W DANYCH

04

NIESPÓJNOŚCI W DANYCH



01

BRAKUJĄCE DANE

DLACZEGO DANYCH BRAKUJE?

- Zrozumienie przyczyn braku danych jest ważne dla prawidłowej obsługi pozostałych danych.
- **Dane brakujące całkowicie losowo (MCAR)** - zdarzenia, które prowadzą do braku określonego elementu danych, są niezależne zarówno od obserwowalnych zmiennych, jak i nieobserwowalnych parametrów będących przedmiotem zainteresowania i występują całkowicie losowo. Gdy dane są MCAR, analiza przeprowadzona na danych jest nieobciążona.
- **Dane brakujące losowo (MAR)** - brak nie jest losowy, ale może zostać w pełni wyjaśniony przez zmienne, w których istnieje pełna informacja.
- **Dane brakujące nielosowo (MNAR)** - brak nie jest losowy, wartość brakującej zmiennej jest związana z przyczyną jej braku.

SPOSOBY NA BRAKUJĄCE DANE

- Usunięcie kolumn / wierszy
- Uzupełnienie danych tzw. imputacja danych
  - Stała wartość
  - Średnia dla atrybutu np. wiek populacji
  - Średnia dla danej klasy np. w przedziałach wiekowych [18, 25), [25, 35), etc.
  - Oszacowana wartość: regresja, kNN, metody probabilistyczne

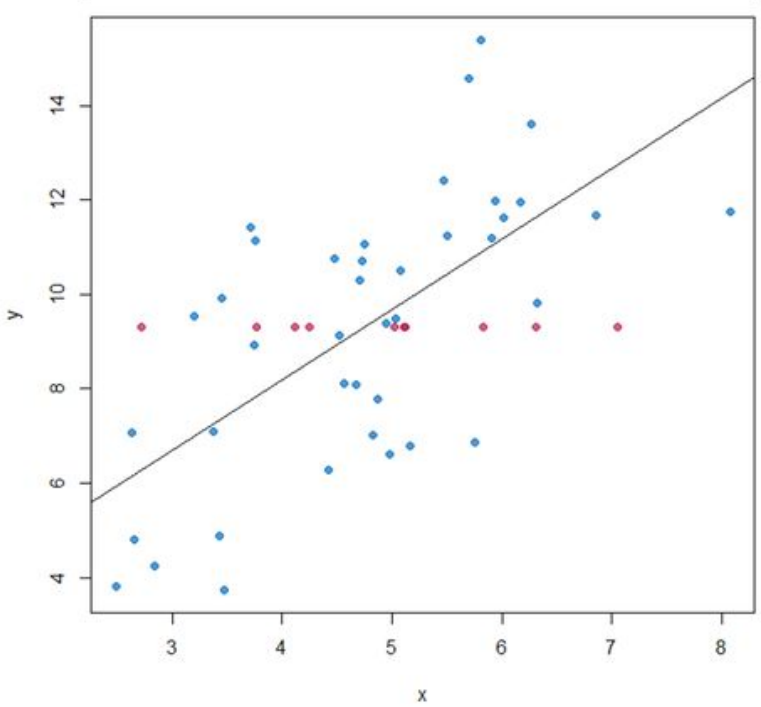
USUNIĘCIE OBSERWACJI

- Stosowane jeżeli liczba wierszy z brakującymi danymi stanowi niewielką frakcję danych.
- Nie wpływa na poprawność analizy tylko jeżeli dane są MCAR.

IMPUTACJA ŚREDNIĄ

- Polega na zastąpieniu brakujących wartości średnią z danej zmiennej obliczoną ze wszystkich innych przypadków.
- ➕ Nie zmienia to średniej z próby dla tej zmiennej.
- ➖ Osłabia wszelkie korelacje dotyczące zmiennych, które są imputowane.
  - W przypadkach z imputacją nie ma żadnego związku między zmienną przypisaną a innymi mierzonymi zmiennymi.
- Zatem imputacja średnia ma pewne atrakcyjne właściwości dla analizy jednowymiarowej, ale staje się problematyczna dla analizy wielowymiarowej.
- Średnia imputacja może być przeprowadzona w ramach klas (tj. kategorii takich jak płeć)

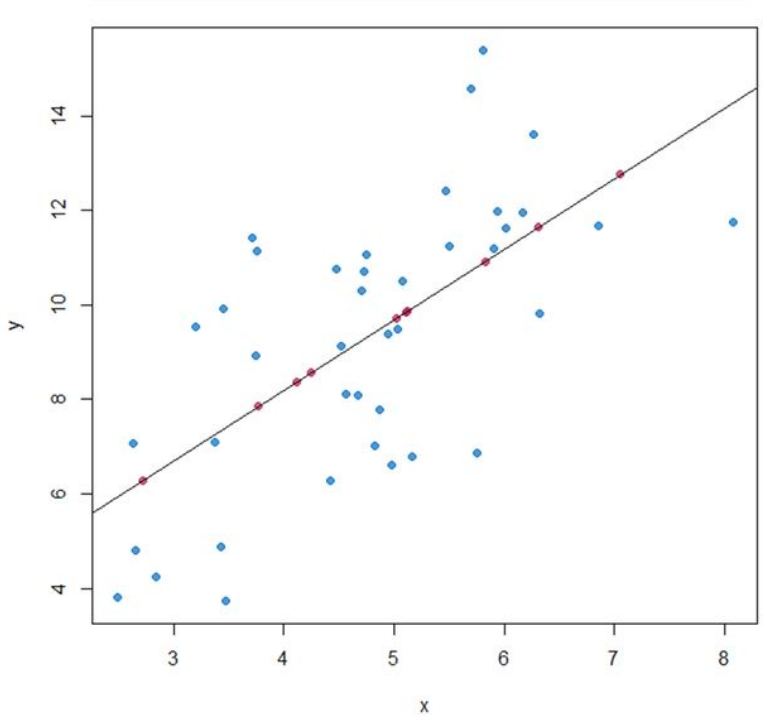
IMPUTACJA ŚREDNIĄ



REGRESJA

- Model regresji jest tworzony w celu przewidywania obserwowanych wartości zmiennej na podstawie innych zmiennych.
- Następnie model ten jest używany do przypisywania wartości w przypadkach, gdy brakuje wartości tej zmiennej.
- Dostępne informacje dla kompletnych przypadków (obserwacje o pełnych danych) są wykorzystywane do przewidywania wartości określonej zmiennej.
- Można wykorzystać jedną zmienną, najbardziej skorelowaną ze zmienną imputowaną, lub więcej zmiennych.
- Dane imputowane nie mają składnika błędu zawartego w ich estymacji, dlatego oszacowania idealnie pasują do linii regresji bez żadnej wariancji resztowej.
- ➕ Uwzględnione są relacje pomiędzy zmiennymi.
- ➖ Model regresji przewiduje najbardziej prawdopodobną wartość brakujących danych, ale nie dostarcza niepewności co do tej wartości.

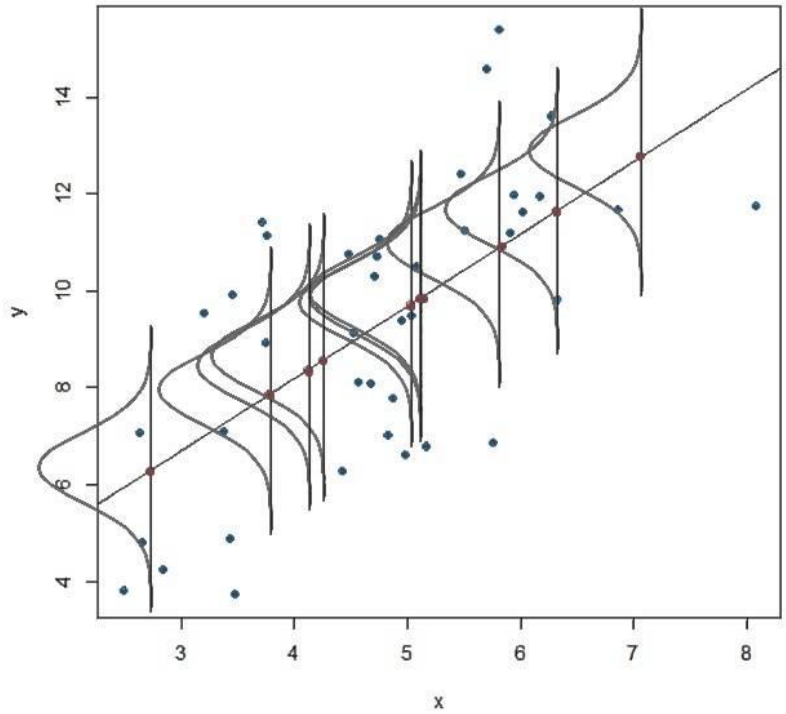
REGRESJA



REGRESJA STOCHASTYCZNA

- Rozwinięcie imputacji regresyjnej o dodatkowy krok polegający na dodaniu losowego błędu do przewidywanej wartości imputowanej zmiennej.
- Wartość błędu losowana jest z rozkładu normalnego o średniej 0 i wariancji błędu oszacowań modelu regresji.
- ➕ Regresja stochastyczna jest jedną z najlepszych klasycznych metod radzenia sobie z brakami danych.
- ➖ Wyzwanie dla testowania hipotez - ile mamy obserwacji w zbiorze? Imputowane dane mają większą wartość niż brak danych, jednak czy możemy takiej obserwacji zaufać równie mocno jak pełnej obserwacji?

REGRESJA STOCHASTYCZNA

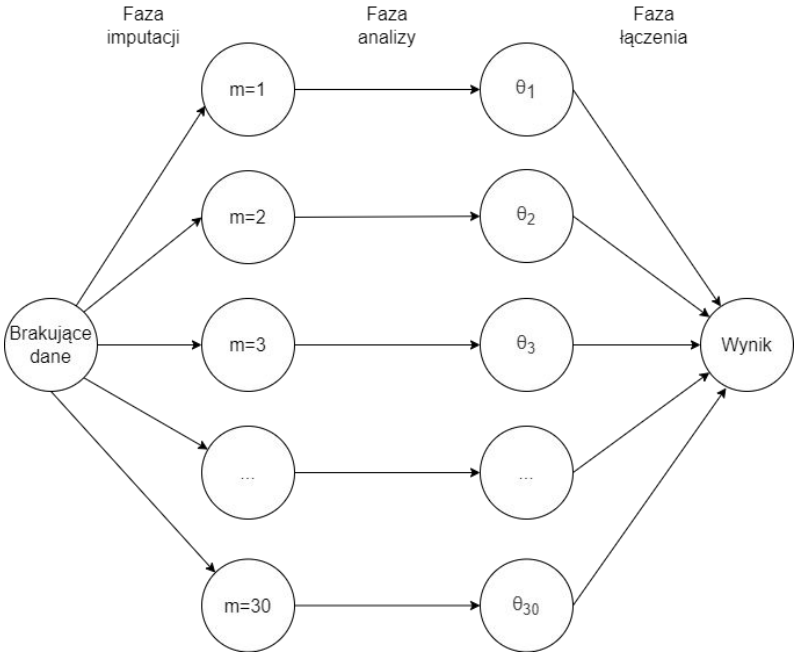


REGRESJA WIELOKROTNA

Proces wielokrotnej imputacji składa się z trzech faz:

- fazy imputacji - generowanie N imputacji na podstawie N zestawów danych,
- fazy analizy - szacowanie interesujących parametrów,
- fazy łączenia - agregacja wyników, np. za pomocą średniej.

➕ Można stosować do zmiennych o różnych rozkładach i typach danych.
➖Metoda jest wrażliwa na błędną lub niekompletną specyfikację modelu -> model imputacyjny powinien zawierać tak dużo zmiennych jak to tylko możliwe.

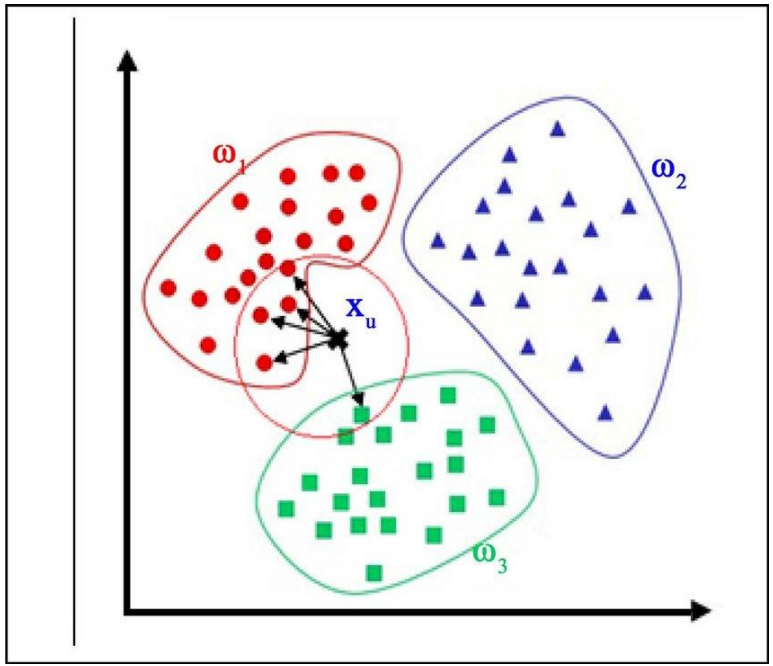


k NAJBLIŻSZYCH SĄSIADÓW

- Brakujące wartości każdej próbki są imputowane przy użyciu średniej wartości z k najbliższych sąsiadów znalezionych w zbiorze uczącym.
- Dwie próbki są zbliżone, jeśli cechy, których nie brakuje, są zbliżone.
- Do znajdowania najbliższych sąsiadów można używać dowolnej metryki, np. euklidesowej.

➕ Prosta i skuteczna strategia imputacji danych.

➖ Wybór k może być kością niezgody.



PODSUMOWANIE

- Problem braku danych jest powszechny.
- Może wynikać z czynników ludzkich, sprzętowych lub oprogramowania.
- Problem braków danych każdorazowo powinien zostać przemyślany, a sposób rozwiązania tej kwestii – opisany i uargumentowany.

02

WARTOŚCI ODSTAJĄCE

WARTOŚCI ODSTAJĄCE

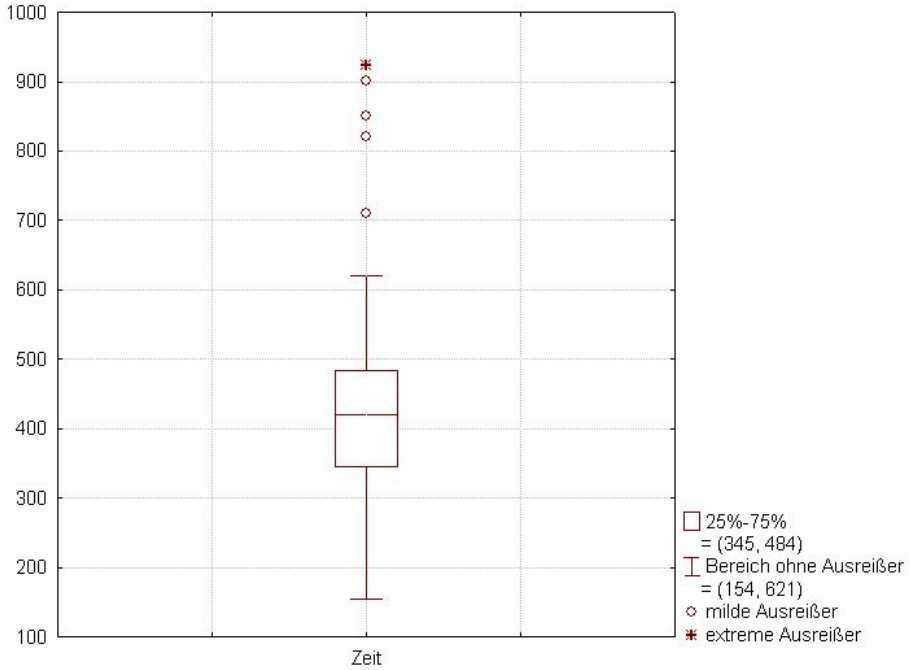
- Wartość odstająca — punkt danych, który znacznie różni się od innych obserwacji.
- Mogą powodować poważne problemy w analizach statystycznych.
- Wartości odstające mogą:
  - Pojawić się przypadkowo w dowolnym rozkładzie,
  - Wynikać z błędu pomiaru → odrzucamy takie obserwacje albo korzystamy ze statystyk odpornych na wartości odstające.
  - Wskazywać, że populacja ma rozkład z ciężkim ogonem (dużą skośność) → należy być bardzo ostrożnym w korzystaniu z narzędzi lub intuicji, które zakładają rozkład normalny.
- Nie ma ścisłej matematycznej definicji co to jest obserwacja odstająca.

KRYTERIUM TUKEYA

- Obserwacje nie są odstające powinny leżeć w przedziale:

$$[Q_1 - k \cdot IQR,\ Q_3 + k \cdot IQR]$$
$$IQR = Q_3 - Q_1$$

- John Tukey zaproponował:
  - k=1.5 dla obserwacji odstających
  - k=3 dla obserwacji ekstremalnie odstających
- Obserwacje odstające wg tego kryterium łatwo dostrzec na wykresie pudełka z wąsami.



ODLEGŁOŚĆ COOKA

Odległość Cooka określa, jaki silny wpływ ma dana obserwacja na model regresji liniowej.

Odległość Cooka D_i dla i = 1, ..., n) jest definiowany jako suma wszystkich zmian w modelu regresji po usunięciu z niego i-tej obserwacji

$$D_i = \frac{\sum_{j=1}^{n} \left(\widehat{y_j} - \widehat{y_{j(i)}}\right)^2}{ps^2}$$

gdzie:
p - liczba zmiennych wyjaśniających,
$\widehat{y_j}$ - wartość odpowiedzi modelu regresji ze wszystkimi obserwacjami,
$\widehat{y_{j(i)}}$ - wartość odpowiedzi modelu regresji z wykluczoną i-tą obserwacją,
s^2 - błąd średniokwadratowy modelu regresji.

ODLEGŁOŚĆ COOKA

Istnieje kilka interpretacji dystansu Cooka:

- Obserwacje z odległością Cooka większą niż 3-krotność średniej odległości Cooka są obserwacjami odstającymi.
- Obserwacje z odległością Cooka większą niż 4/n, gdzie n jest liczbą obserwacji.
- Należy zbadać każdą „dużą” odległość Cooka. Jak duża jest “zbyt duża”? Najczęściej podaje się 1 jako wartość graniczną.

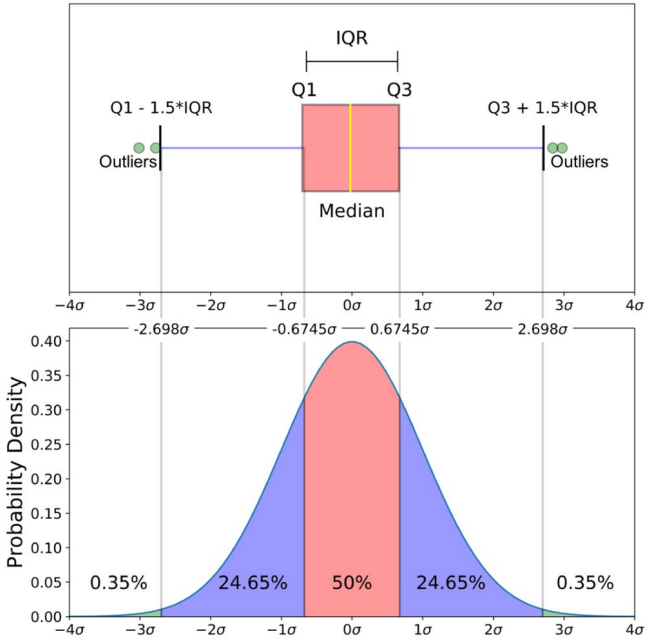
Z-SCORE

$$Z_i = \frac{X_i - \bar{X}}{s}$$

Zazwyczaj za obserwacje odstające uznajemy te, dla których $Z_i \geq 3$.

Wadą stosowania tej metody jest fakt, że jest ona bardzo wrażliwa na średnią i odchylenie standardowe, które same w sobie są wrażliwe na wartości odstające.

Metoda powinna być stosowana do zmiennych z rozkładu normalnego.



ODCHYLENIE BEZWZGLĘDNE

Mediana jest, podobnie jak średnia, miarą tendencji centralnej, ale ma tę zaletę, że jest bardzo niewrażliwa na obecność obserwacji odstających.

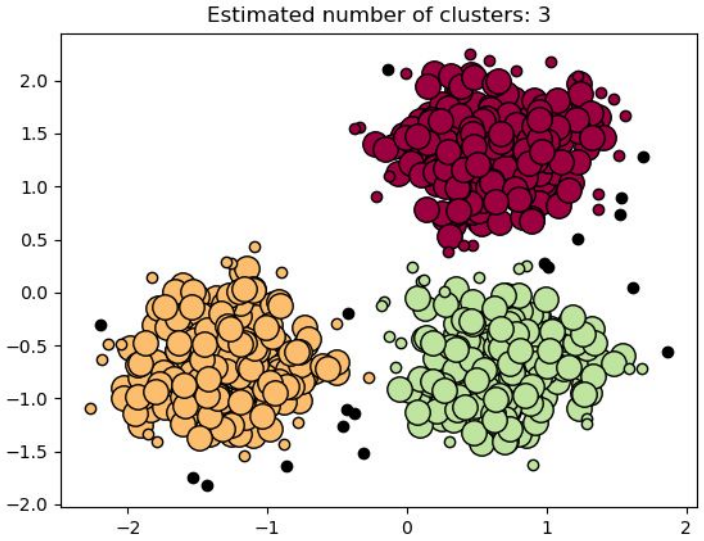
$$MAD_e = m_e(|x_i - m_e(x)|)$$

Zazwyczaj za obserwacje odstające uznajemy te, dla których $MAD_3 \geq 3$.

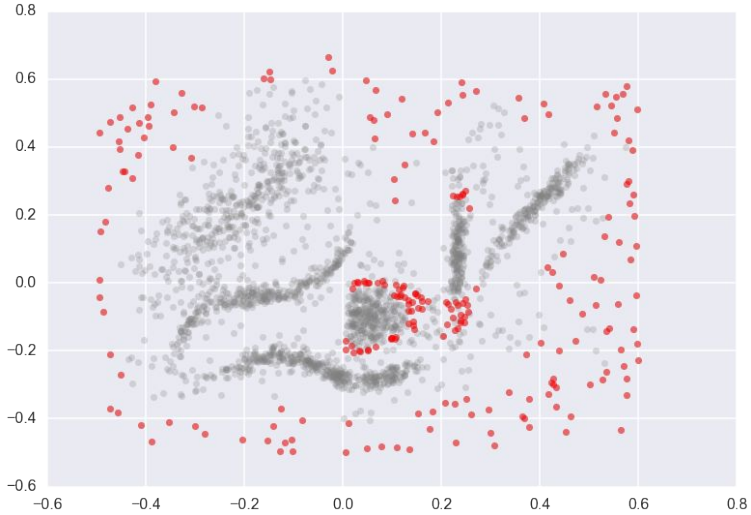
Jeśli więcej niż 50% Twoich danych ma identyczne wartości, MAD będzie równy zero. Wszystkie punkty w zestawie danych z wyjątkiem tych, które są równe medianie, zostaną następnie oznaczone jako wartości odstające, niezależnie od poziomu, na którym ustawiono wartość graniczną wartości odstających.

ALGORYTMY KLASTRUJĄCE

Obserwacje, które nie zostaną przydzielone do żadnego klastra są uznawane za obserwacje odstające.



DBSCAN



HDBSCAN

ZOSTAWIĆ CZY USUNĄĆ?

Usuwanie lub utrzymywanie wartości odstających zależy głównie od trzech czynników:

1. Dziedzina/kontekst Twoich analiz i pytanie badawcze. W niektórych dziedzinach często usuwa się wartości odstające, ponieważ często występują one z powodu nieprawidłowego działania procesu. W innych dziedzinach trzymane są wartości odstające, ponieważ zawierają cenne informacje.
2. Czasami analizy wykonuje się dwukrotnie, raz z wartościami odstającymi i raz bez wartości odstających, aby ocenić ich wpływ na wnioski. Jeśli wyniki zmienią się drastycznie z powodu pewnych wpływowych wartości, powinno to ostrzec badacza przed wysuwaniem zbyt śmiałych twierdzeń.
3. Niektóre testy i metody są odporne na obecność wartości odstających lub nie. Na przykład nachylenie prostej regresji liniowej może się znacznie różnić przy zaledwie jednej wartości odstającej, podczas gdy testy nieparametryczne są zwykle odporne na wartości odstające.

Jeżeli punkt(y) danych jest wykluczony z analizy danych, należy to wyraźnie zaznaczyć w raporcie.

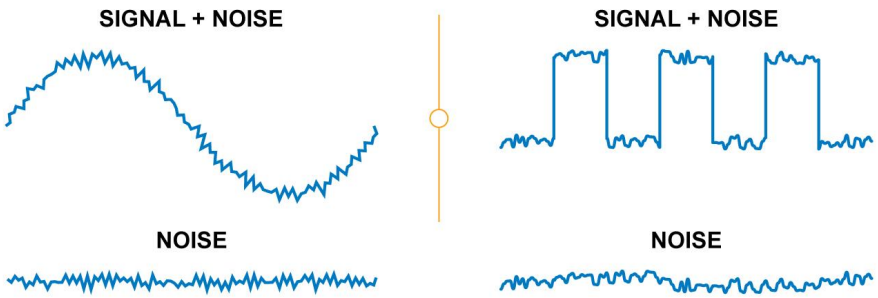
03

SZUM W DANYCH

ZASZUMIONY SYGNAŁ

Zaszumione dane – dane, które są uszkodzone, zniekształcone lub mają niski stosunek sygnału do szumu.

Dane = prawdziwy sygnał + szum



Źródło grafiki: https://www.predig.com/whitepaper/reducing-signal-noise-practice

ŹRÓDŁA SZUMU

- Szum losowy
  - Przypadkowy szum w sygnale jest mierzony jako stosunek sygnału do szumu.
  - Szum losowy zawiera prawie równe ilości szerokiego zakresu częstotliwości i jest nazywany szumem białym (ponieważ kolory światła łącząc się, tworzą biel).
- Niewłaściwe filtrowanie może dodać szum, jeśli filtrowany sygnał jest traktowany jako sygnał bezpośrednio mierzony.
  - Skutki uboczne filtrów takich jak średnia ruchoma lub wzmocnienie szumu przez filtry różnicujące
- Dane odstające
- Oszustwo
  - Celowa modyfikacja danych, aby wpłynąć na wyniki i doprowadzić do pożądanego wniosku.

BINNING

Binning lub kubełkowanie - technika, w której dzielimy dane na małe przedziały, a zaszumione dane zastępujemy wartością reprezentatywną dla przedziału, do którego wpadają.
Przedziały mogą być:

- równoliczne,
- o stałej szerokości.

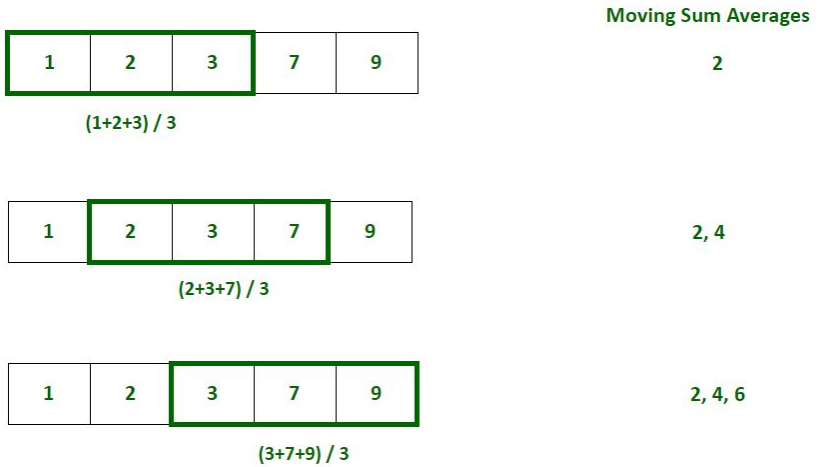
Wartością reprezentatywną dla przedziału może być:

- średnia,
- mediana,
- najbliższa wartość graniczna przedziału (dolna lub górna).

➕ Pozwala zredukować szum
➖ Utrata części informacji
❔ Zamiast zmiennej ciągłej otrzymujemy zmienną dyskretną

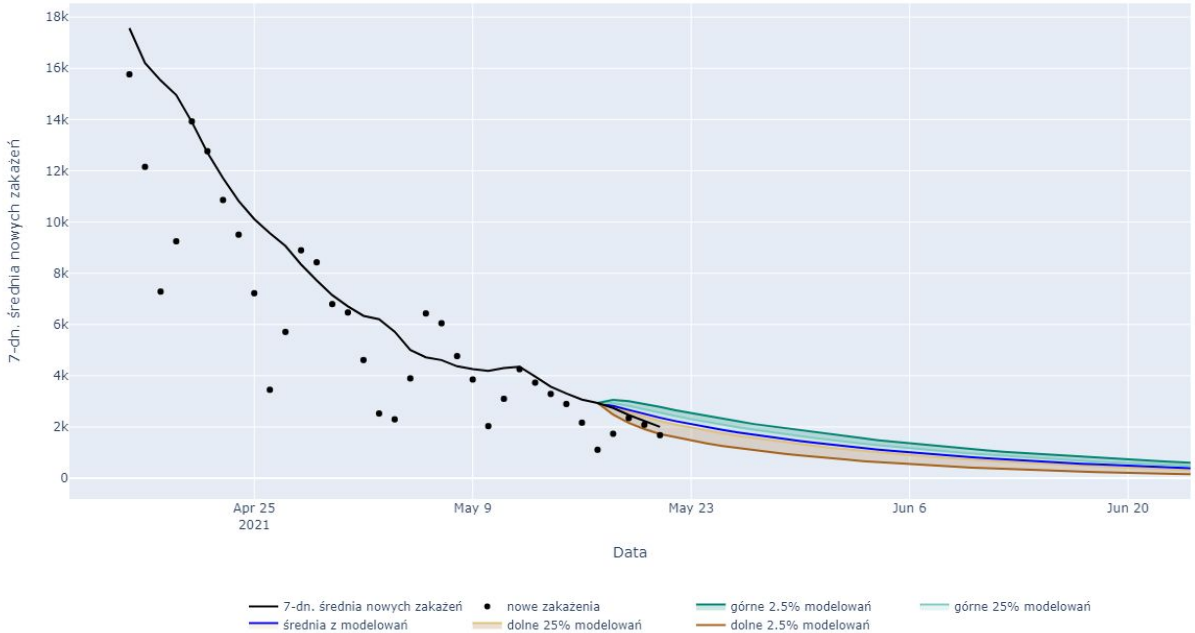
ŚREDNIA RUCHOMA (KROCZĄCA)

- Prosta i powszechna metoda stosowana w analizie szeregów czasowych.
- Polega na utworzeniu nowej serii, w której wartości składają się ze średniej surowych obserwacji w oryginalnych szeregach czasowych.
- Wymaga określenia rozmiaru okna, czyli liczby surowych obserwacji używanych do obliczania wartości średniej ruchomej.
- Okno jest przesuwane wzdłuż szeregu czasowego w celu obliczenia średnich wartości w nowym szeregu.



Źródło grafiki: https://www.geeksforgeeks.org/how-to-calculate-moving-averages-in-python/

ŚREDNIA RUCHOMA (KROCZĄCA)



Źródło grafiki: https://mocos.pl/pl/prognosis/20210517

ŚREDNIA RUCHOMA (KROCZĄCA)

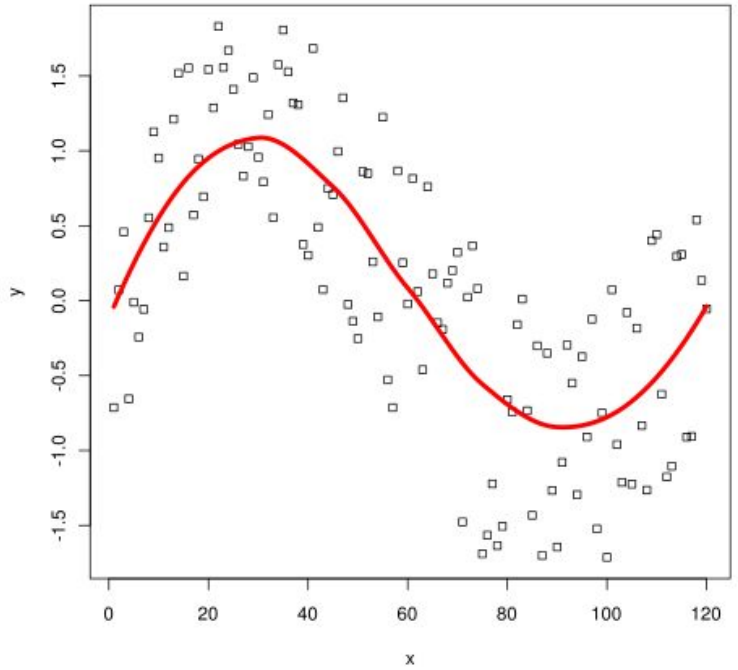
- Prosta średnia ruchoma (SMA)
  - Nadaje równą wagę wszystkim obserwacjom w oknie.
- Ważona średnia ruchoma (WMA)
  - Przypisuje różne wagi danym z poszczególnych okresów, np. w n-okresowej WMA ostatni okres ma wagę n, przedostatni n−1 i tak dalej aż do 1.
- Wykładnicza średnia ruchoma (EMA)
  - Odmiana średniej ważonej, w której znaczenie coraz bardziej odległych w czasie okresów maleje w sposób wykładniczy.

➕Prosta metoda pozwalająca na identyfikację trendów i potencjalnych obszarów wsparcia lub oporu.

➖Wprowadza własne artefakty do danych - np. opóźnienia pików, może nie odzwierciedlać dokładnie najnowszych trendów.

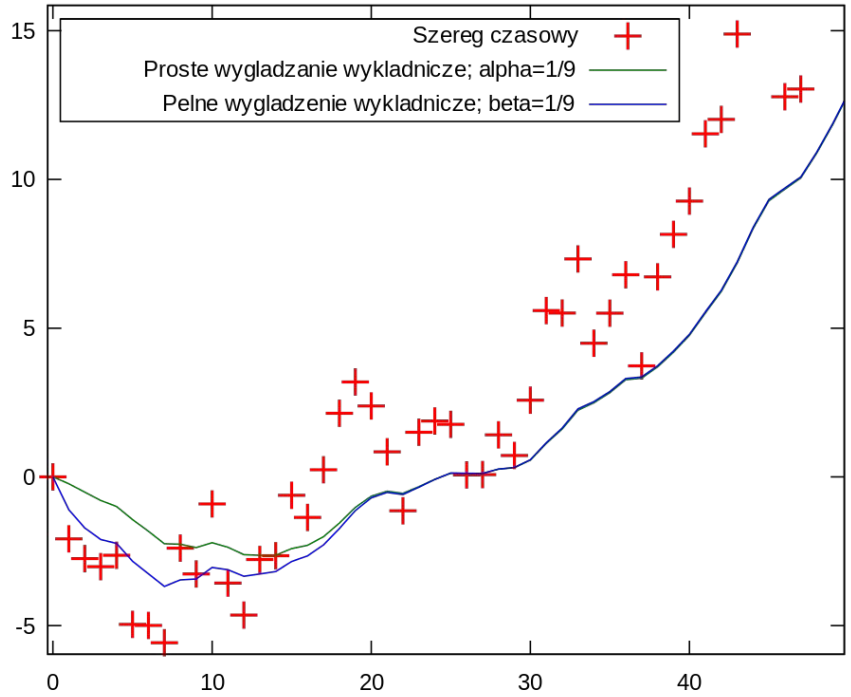
LOWESS / LOESS / filtr Savitzky'ego-Golaya

Po prawej: Krzywa LOESS dopasowana do populacji pobranej z fali sinusoidalnej z dodanym jednolitym szumem.



Źródło grafiki: https://en.wikipedia.org/wiki/Local_regression

WYGŁADZANIE WYKŁADNICZE



Źródło grafiki: https://pl.wikipedia.org/wiki/Wygładzanie_wykładnicze

04

NIESPÓJNOŚCI W DANYCH

ANOMALIE SKŁADNIOWE

- Dotyczą formatu i wartości używanych do reprezentacji encji.
  - Błędy leksykalne: błędy ortograficzne, błędy literowe.
  - Niezgodność w typie danych (wartość logiczna, numeryczna, data lub obiekt).
  - Wyjście poza zakres danych: wartości z punktacją powyżej 100 procent lub wiek ujemny.
  - Niespójny format danych łańcuchowych: np. Buntine, Wray Lindsay lub Wray L. Buntine.
  - Niejednolite użycie wartości, jednostek lub skrótów: np. wynagrodzenie w różnych walutach.
- Czasami te błędy są łatwe do znalezienia (np. ujemny wiek), ale trudne do naprawienia.
- Anomalie składniowe można wykryć za pomocą metod takich jak:
  - analiza częstości wystąpień wartości w kolumnie (pandas valuecounts)
  - weryfikacja typu danych każdej kolumn (pandas info)
  - analiza statystyczna każdej kolumny za pomocą metod statystyki opisowej (pandas describe)

ANOMALIE SEMANTYCZNE

- Wartość zawarta w samej kolumnie jest poprawna, ale razem z wartościami różnych kolumn jest niepoprawna.
  - Naruszenia ograniczeń integralności: niespełnione są warunki klucza głównego, obcego lub unikatowego.
  - Sprzeczności: Naruszenie zależności między atrybutami, np.: wiek i data urodzenia.
  - Duplikaty: obserwacje reprezentujące ten sam podmiot.
- Anomalie semantyczne można wykryć za pomocą metod takich jak:
  - tabele krzyżowe (pandas crosstab)
  - grupowanie (pandas groupby)
  - weryfikacja zduplikowanych wierszy (pandas duplicated)
- Tego rodzaju błędy są trudne do wykrycia, ponieważ najpierw musimy ustalić związek lub korelacje z innymi kolumnami.

THANKS!

DZIĘKUJĘ ZA UWAGĘ