Podstawy eksploracji danych wykład I

Agnieszka Nowak - Brzezińska

Plan wykładu

- Podstawowe pojęcia
- Możliwości eksploracji danych
- Metody eksploracji danych
- Zadania dotąd niezrealizowane w ramach eksploracji danych
- Przygotowanie danych do analizy – I etap ED

Definicja nr 1

Eksploracja danych jest procesem odkrywania znaczących nowych powiązań, wzorców i trendów przez przeszukiwanie danych zgromadzonych w skarbnicach, przy wykorzystaniu metod rozpoznawania wzorców, jak również metod statystycznych i matematycznych.

Definicja nr 2

Eksploracja danych jest analiza (często ogromnych) zbiorów danych obserwacyjnych w celu znalezienia nieoczekiwanych związków i podsumowania danych w oryginalny sposób tak, aby były zarówno zrozumiałe, jak i przydatne dla ich właściciela przez przeszukiwanie danych zgromadzonych w skarbnicach, przy wykorzystaniu metod rozpoznawania wzorców, jak również metod statystycznych i matematycznych.

Definicja nr 3

Eksploracja danych jest między dyscyplinarną dziedziną, łączącą techniki uczenia maszynowego, rozpoznawania wzorców, statystyki, baz danych i wizualizacji w celu uzyskiwania informacji z dużych baz danych.

- Wszystkie definicje mówią o tym, że celem eksploracji danych jest odkrywanie nowych zależności w zbiorach danych, które nie były wcześniej znane odbiorcy. Wyjątek stanowi tutaj chęć potwierdzenia pewnej określonej hipotezy, czyli potwierdzenia znanej wcześniej zależności w danych.
- Rezultaty procesu eksploracji są nazywane modelami lub wzorcami. Mogą to być równania liniowe, reguły, skupienia, grafy, struktury drzewiaste lub wzorce rekurencyjne w szeregach czasowych.

“Eksploracja danych polega na torturowaniu danych tak długo, aż zaczną zeznawać”

Schemat ogólny DM

1. Zdefiniować problem/zadanie i zanalizować otoczenie.
2. Wybrać zbiór danych do eksploracji i atrybuty.
3. Zdecydować jak przygotować dane do przetwarzania.
Na przykład: czy wiek reprezentować jako przedział
(np. 40-45 lat), czy jako liczbę (np. 40 lat).
4. Wybrać algorytm (lub ich kombinację) eksploracji
i wykonać program realizujący ten algorytm.
5. Zanalizować wyniki wykonania programu i wybrać te,
które uznajemy za rezultat pracy.
6. Przedłożyć wyniki kierownictwu organizacji i zasugerować
sposób ich wykorzystania.

Zastosowania KDD znajdują zastosowania przy:

- eksploracji danych o ruchu internetowym,
- rozpoznawaniu sygnałów obrazu, mowy, pisma,
- wspomaganiu diagnostyki medycznej,
- badaniach genetycznych,
- analizie operacji bankowych,
- projektowaniu hurtowni danych,
- tworzeniu reklam skierowanych (ang. Targeted ads),
- prognozowaniu sprzedaży (ang. Sales forecast),
- wdrażaniu strategii Cross-selling'owej,
- wykrywaniu nadużyć(ang. Fraud detection),
- ocenie ryzyka kredytowego,
- segmentacji klientów.

Przykładem może być odkrycie w danych z supermarketu zależności polegającej na tym że klient, który kupuje szampana i kwiaty, kupuje zwykle również czekoladki.

Problem eksplozji danych

- Narzędzia zbierania danych+rozwój systemów bazodanowych
- gwałtowny wzrost ilości danych zgromadzonych w bazach danych, hurtowniach danych i magazynach danych
- Np.:
- $N = 10^9$ rekordów w danych astronomicznych,
- $d = 10^2 \sim 10^3$ atrybutów w systemach diagnozy medycznej

- „Jesteśmy zatopieni w morzu danych, podczas gdy pragniemy wiedzę”
- PROBLEM: jak wydobyć użyteczne informacje/wiedzy z dużego zbioru danych?
- Rozwiązanie: hurtownia danych + data mining
- Zbieranie danych (w czasie rzeczywistym)
- Odkrywanie interesującej wiedzy (reguł, regularności, wzorców, modeli ...) z dużych zbiorów danych

Ewolucja w bazach danych

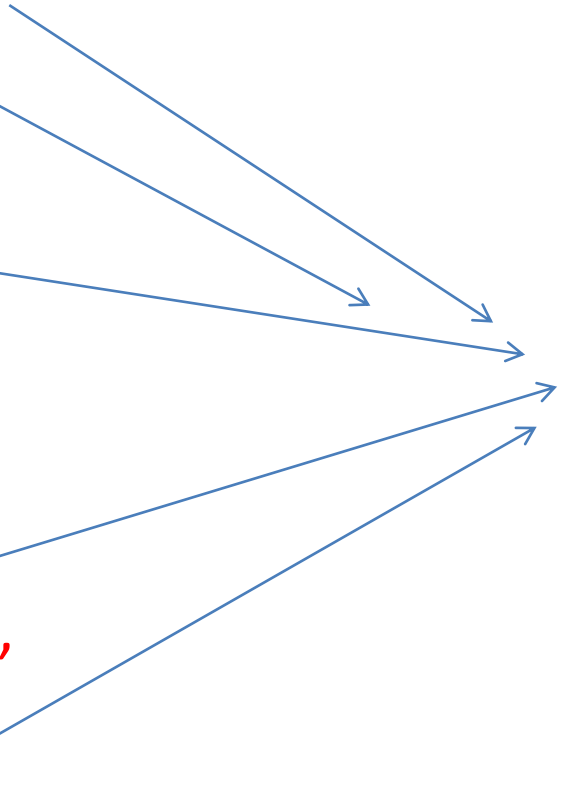
- W latach 60-tych:
- Kolekcja danych, tworzenie baz danych, IMS oraz sieciowe DBMS
- W latach 70-tych:
- Relacyjny model danych, implementacja relacyjnych DBMS
- W latach 80-tych:
- RDBMS, zaawansowane modele danych (extended-relational, OO, deductive, ...) oraz aplikacyjno-zorientowaneDBMS
- Od 90-tych —obecnie:
- Data mining, hurtownia danych, multimedialne bazy danych oraz „Web databases”

10 najważniejszych algorytmów eksploracji danych

- C4.5,
- k-Means,
- SVM,
- Apriori,
- EM,
- PageRank,
- AdaBoost,
- kNN,
- Naive Bayes,

oraz

- CART.



Poznamy je na wykładzie i zajęciach laboratoryjnych PED

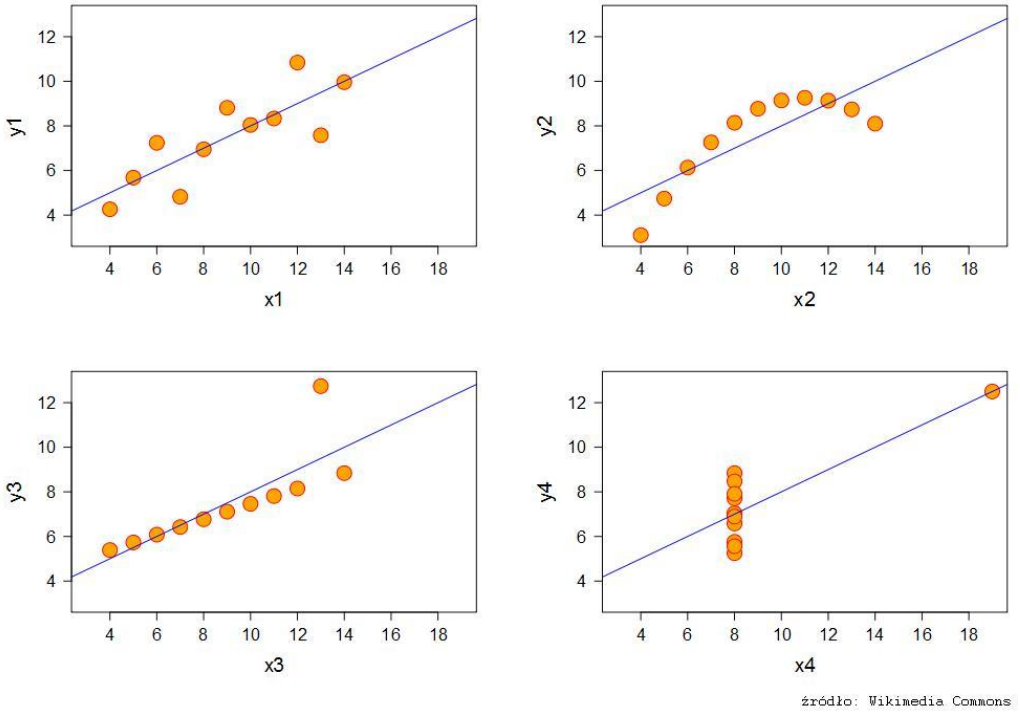
10 najważniejszych problemów eksploracji danych

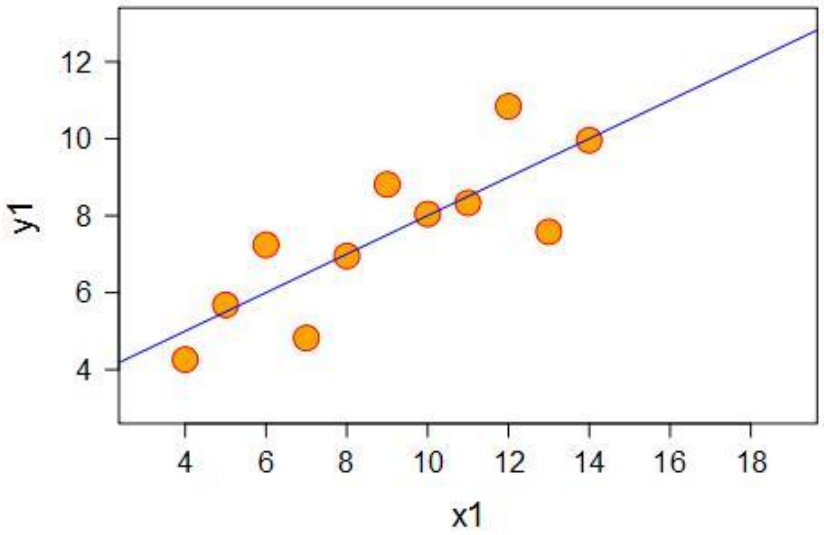
Oto lista 10 najważniejszych problemów eksploracji danych (kolejność nie odzwierciedla ważności):

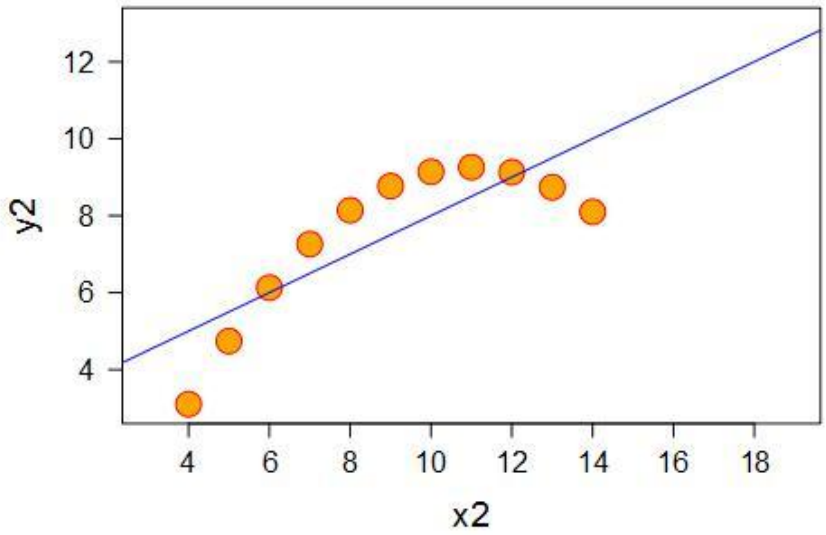
1. stworzenie Unifikującej Teorii Eksploracji Danych (UTED),
2. opracowanie skalowalnych metod dla problemów opisanych bardzo wieloma wymiarami oraz dla problemów opisanych przez dane strumieniowe,
3. praca nad eksploracją przebiegów czasowych i danych sekwencyjnych,
4. odkrywanie złożonych wzorców w złożonych typach danych,
5. eksploracja struktur sieciowych (zarówno sieci społecznych jak i sieci komputerowych),
6. opracowanie metod rozproszonej eksploracji danych oraz wykorzystanie systemów agentowych do odkrywania wiedzy,
7. eksploracja danych w dziedzinie biologii i ekologii,
8. eksploracja danych opisujących procesy (np. przepływy pracy),
9. bezpieczeństwo, poufność i spójność danych,
10. praca z danymi, które są niezrównoważone, dynamiczne, podlegające ewolucji.

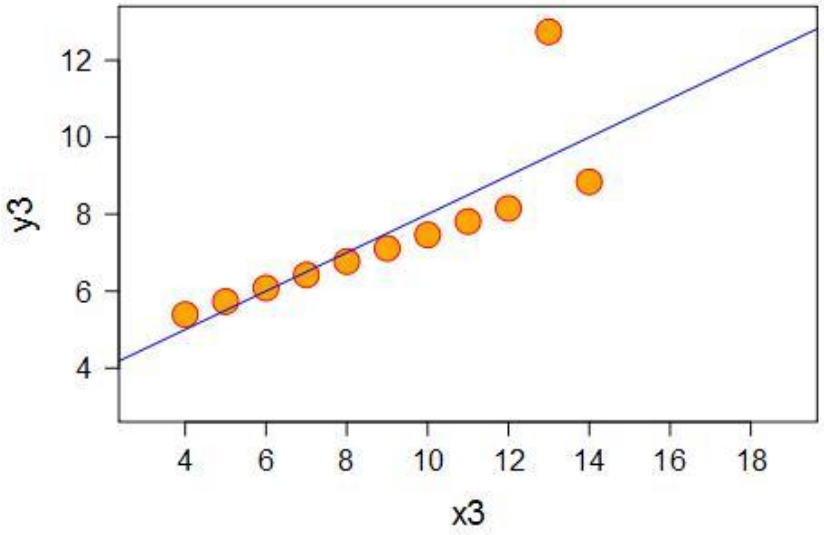
kwartet Anscombe’a

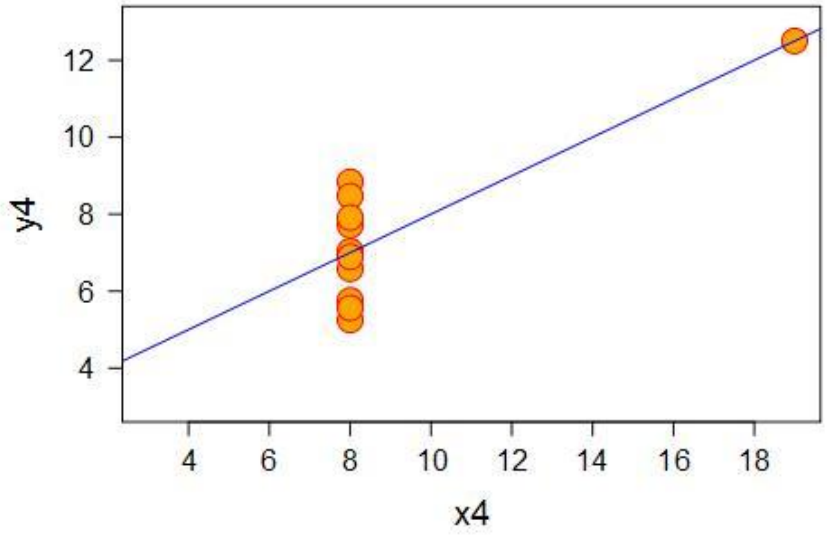
Francis Anscombe był angielskim statystykiem, który dużą część życia spędził na uniwersytetach Yale i Princeton. Anscombe był jednym z pionierów analizy wizualnej i często podkreślał istotność wizualizacji zbioru danych poddawanego analizie. Na potrzeby ilustracji stworzył cztery proste zbiory danych, nazwane kwartetem Anscombe’a.





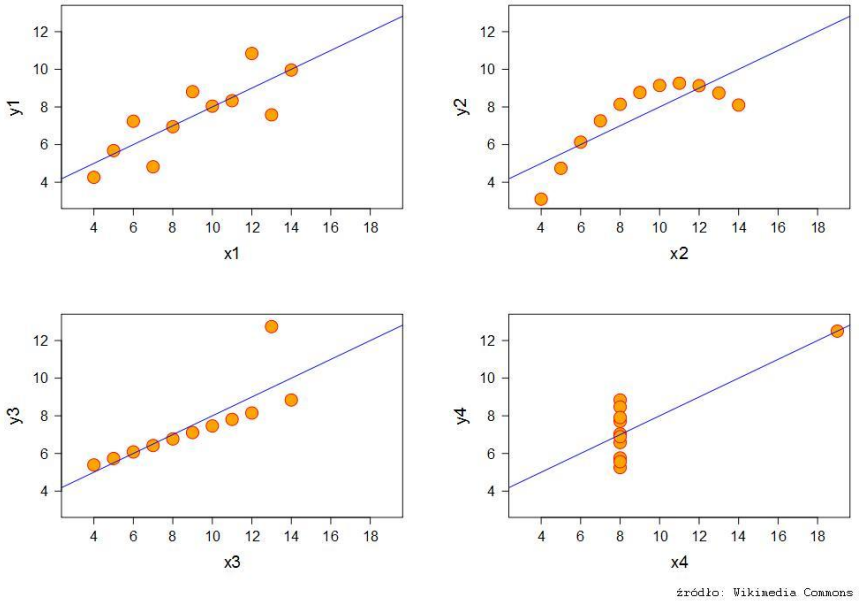






źródło: Wikimedia Commons

- Co jest takiego ciekawego w tych czterech zbiorach? Otóż wszystkie cztery mają **dokładnie te same** własności statystyczne.
- Ten przykład dobitnie pokazuje, jak istotne jest zapoznanie się i zaprzyjaźnienie z eksplorowanymi danymi.



źródło: Wikimedia Commons

średnia $\overline{x_i} = 9.0$

wariancja $\sigma_x^2 = 10.0$

średnia $\overline{y_i} = 7.5$

wariancja $\sigma_y^2 = 3.75$

korelacja $\rho_{x,y} = 0.816$

regresja liniowa: $y = 0.5x + 3$

Data mining

I. Zrozumienie dziedziny problemu i celu analizy

II. Budowa roboczego zbioru danych

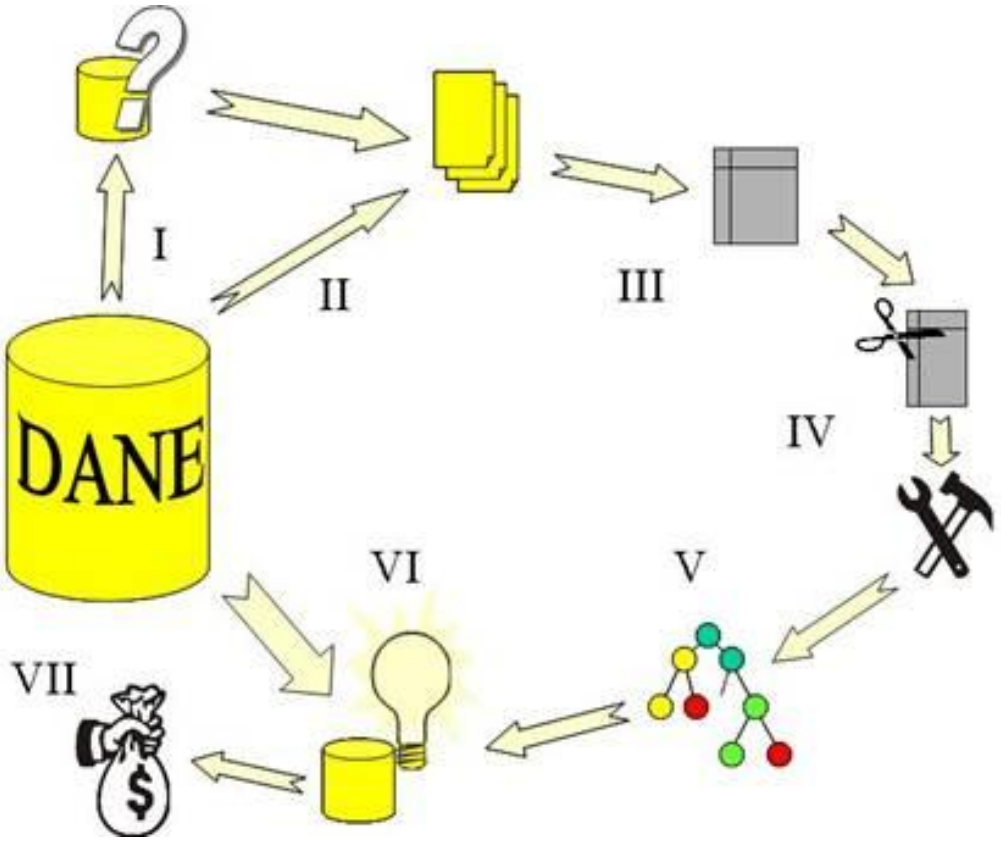
III. Przygotowanie i oczyszczenie danych

IV. Wybór metody analizy danych

V. Eksploracja danych (data mining)

VI. Interpretacja znalezionych regularności

VII. Wykorzystanie odkrytej wiedzy



- Każda technika pasuje tylko do pewnych problemów
- Trudność polega na znalezieniu właściwego sformułowania problemu (dobre pytania)
- Nie ma jeszcze żadnego kryterium, potrzebne jest wyczucie eksperta!!!

Techniki eksploracji

- Eksploracja danych posługuje się różnymi technikami, które budują specyficzne rodzaje wiedzy.
- W zależności od przeznaczenia odkrywanej wiedzy, może ona odwzorowywać klasyfikacje, regresje, klastrowanie, charakterystyki, dyskryminacje, asocjacje itp.. Poniżej dokonano krótkiej charakterystyki każdej z wymienionych technik eksploracji danych.

Klasyfikacja

- Klasyfikacja polega na znajdowaniu sposobu odwzorowania danych w zbiór predefiniowanych klas.
- Na podstawie zawartości bazy danych budowany jest model (np. drzewo decyzyjne, reguły logiczne), który służy do klasyfikowania nowych obiektów w bazie danych lub głębszego zrozumienia istniejących klas. Przykładowo, w medycznej bazie danych znalezione mogą być reguły klasyfikujące poszczególne schorzenia, a następnie przy pomocy znalezionych reguł automatycznie może być przeprowadzone diagnozowanie kolejnych pacjentów. Inne przykłady zastosowań klasyfikacji to:

1. *rozpoznawanie trendów an rynkach finansowych,*
2. *automatyczne rozpoznawanie obiektów w dużych bazach danych obrazów,*
3. *wspomaganie decyzji przyznawania kredytów bankowych*.

Regresja

- Regresja jest techniką polegającą na znajdowaniu sposobu odwzorowania danych w rzeczywistoliczbowe wartości zmiennych predykcyjnych.
- Przykłady zastosowań regresji:

1. przewidywanie zawartości biomasy obecnej w ściółce leśnej na podstawie dokonanych zdalnych pomiarów mikrofalowych
2. szacowanie prawdopodobieństwa wyzdrowienia pacjenta na podstawie przeprowadzonych testów diagnostycznych

Grupowanie (analiza skupień)

- Grupowanie polega na znajdowaniu skończonego zbioru kategorii opisujących dane.
- Kategorie mogą być rozłączne, zupełne, mogą też tworzyć struktury hierarchiczne i nakładające się.
- Przykładowo, zbiór danych o nieznanych chorobach może zostać w wyniku klastrowania podzielony na szereg grup cechujących sie najsilniejszym podobieństwem symptomów.
- Innymi przykładami zastosowań klastrowania mogą być:

1. określanie segmentów rynku dla produktu na podstawie informacji o klientach
2. znajdowanie kategorii widmowych spektrum promieniowania nieba

Odkrywanie charakterystyk

- Odkrywanie charakterystyk polega na znajdowaniu zwięzłych opisów (charakterystyk) podanego zbioru danych.
- Przykładowo, symptomy określonej choroby mogą być charakteryzowane przez zbiór reguł charakteryzujących.
- Inne przykłady odkrywania charakterystyk to:

1. znajdowanie zależności funkcyjnych pomiędzy zmiennymi
2. określanie powszechnych symptomów wskazanej choroby

Dyskryminacja

- Dyskryminacja polega na znajdowaniu cech, które odróżniają wskazaną klasę obiektów (*target class) od innych klas* (*contrasting classes).*
- *Przykładowo, zbiór reguł dyskryminujących może opisywać te cechy objawowe, które odróżniają* daną chorobę od innych.

Odkrywanie asocjacji

- Odkrywanie asocjacji polega na znajdowaniu związków pomiędzy występowaniem grup elementów w zadanych zbiorach danych.
- Najpopularniejszym przykładem odkrywania asocjacji jest tzw. analiza koszyka - przetwarzanie baz danych supermarketów i hurtowni w celu znalezienia grup towarów, które są najczęściej kupowane wspólnie.
- Przykładowo, znalezione asocjacje mogą wskazywać, że kiedy klient kupuje słone paluszki, wtedy kupuje także napoje gazowane.

Proces eksploracji danych

1. **Data cleaning (to remove noise and inconsistent data)**
2. **Data integration (where multiple data sources may be combined)**
3. **Data selection (where data relevant to the analysis task are retrieved fromthe database)**
4. **Data transformation (where data are transformed or consolidated into forms appropriate** for mining by performing summary or aggregation operations, for instance)
5. **Data mining (an essential process where intelligent methods are applied in order to** extract data patterns)
6. **Pattern evaluation (to identify the truly interesting patterns representing knowledge** based on some interestingness measures)
7. **Knowledge presentation (where visualization and knowledge representation techniques** are used to present the mined knowledge to the user)

Metody eksploracji

- **Charakterystyka danych (opis)**
- **Wzorce, asocjacje, powiązania w danych**

  buys(X; “computer”)→buys(X; “software”)

  [support = 1%; confidence = 50%]
- **Klasyfikacja i predykcja (drzewa decyzyjne, sieci neuronowe, analiza regresji)**
- **Analiza skupień (grupowanie)**
- **Wykrywanie odchyleń**

Wszystko to są reguły…

(a)

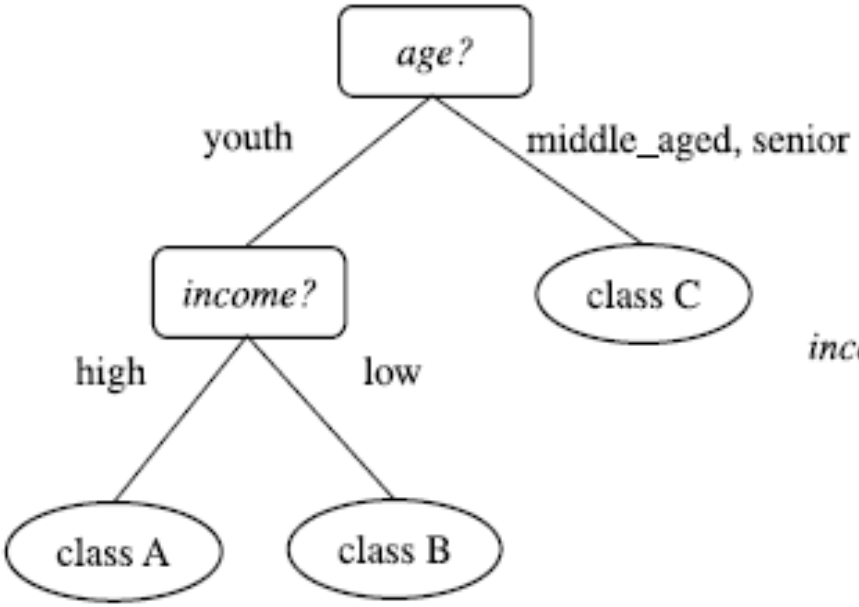
age(X, "youth") AND income(X, "high") → class(X, "A")

age(X, "youth") AND income(X, "low") → class(X, "B")

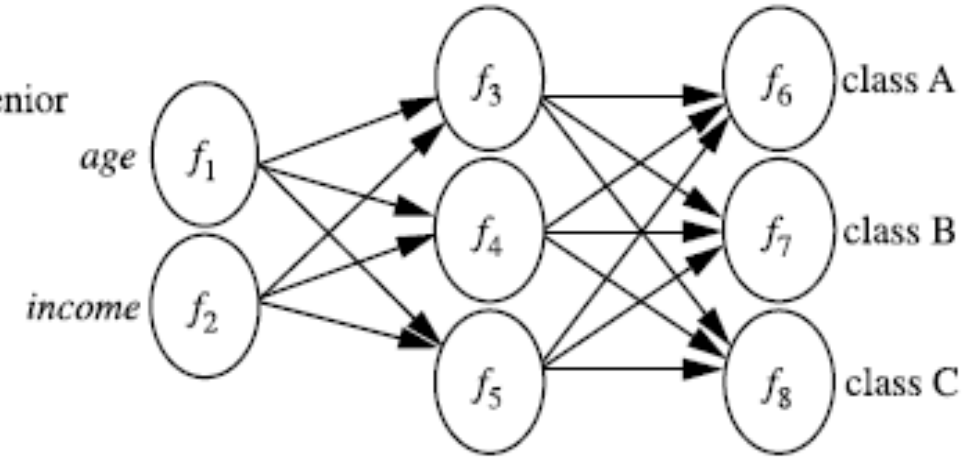
age(X, "middle_aged") → class(X, "C")

age(X, "senior") → class(X, "C")

(b)



(c)



Czy wszystkie wzorce są użyteczne?

- A data mining system has the potential to generate thousands or even millions of patterns, or rules. only a small fraction of the patterns potentially generated would actually be of interest to any given user.
- This raises some serious questions for data mining: *“What makes a pattern interesting? Can a data mining system generate all of the interesting patterns? Can a data mining system generate only interesting patterns?”*

Wzorzec jest interesujący jeśli jest …

- Łatwo zrozumiały dla odbiorcy,
- Użyteczny i poprawny na nowych danych z zadowalającym stopniem pewności,
- Niebanalny,
- Oryginalny (nowy, zaskakujący).

Wzorzec jest interesujący jeśli potwierdza (sprawdza) hipotezy stawiane przez użytkownika. Każdy taki wzorzec reprezentuje pewną wiedzę.

Przykład reguły asocjacyjnej

opis pacjentów chorujących na anginę:

- *pacjenci chorujący na anginę cechują się temperaturą ciała większą niż 37.5 C, bólem gardła, osłabieniem organizmu*

Przykład reguły asocjacyjnej

1. W sklepie *Elektronika* odkryto następującą regułę asocjacyjną:
   - wiek(x, „30...39”) $\wedge$ dochód(x, „1500...2900”) $\rightarrow$ kupuje(x, „CD-RW”)
   - [wsparcie = 2%, ufność = 60%]
   - x oznacza klienta.
2. Reguła mówi, że 2% (wsparcie, *support*) transakcji zakupów w *Elektronice* dotyczyło klientów w wieku od 30 do 39 lat o dochodach od 1500 do 2900 PLN i kupujących CD-RW. 60% (ufność, *confidence*) transakcji dokonanych przez osoby w wieku 30 do 39 lat i o dochodach z przedziału 1500 - 2900 dotyczyło zakupu CD-RW.
3. Jest to przykład wielowymiarowej reguły asocjacyjnej (występują trzy predykaty: *wiek, dochód, kupuje*).

Ocena stopnia jakości wzorca…

- Istnieje wiele metod.
- Często stosowanymi są miar z reguł asocjacyjnych: wsparcie i zaufanie.
- Wsparcie jest stosunkiem ilości wystąpień przykładów P, które zawierają w całości opisy X oraz Y do ilości wszystkich przykładów P, gdzie P jest dowolnym podzbiorem zawierającym tylko i wyłącznie przykłady Z:
- Wsparcie($X \Rightarrow Y$) = $|PX \cup Y| / |P|$
- Każda reguła może być więc wspierana, albo naruszana przez przykład. Natomiast zaufanie reguły w zbiorze P definiuje się następująco:
- Zaufanie($X \Rightarrow Y$) = $|PX \cup Y| / |PX|$
- czyli jako stosunek liczby przykładów ze zbioru P, w których opisach występują zarówno opisy X jak i Y do liczby przykładów, w których opisach występują tylko elementy zbioru X, czyli zbioru wartości warunkujących.

Przygotowanie danych do analizy – I etap eksploracji wiedzy

Low-quality data will lead to low-quality mining results.

- *Techniki preprocessingu stosujemy by poprawić jakość analizowanych danych, by wydobyć z nich potem użyteczną wiedzę.*
- *Czyszczenie danych pozwala pozbyć się szumu w danych i poprawić wszelkie nieścisłości w danych.*
- *Integracja danych pozwala połączyć dane z wielu źródeł np. hurtowni danych.*
- *Transformacja danych (normalizacja) może poprawić dokładność i efektywność algorytmów eksploracji danych te które stosują miarę odległości w danych.*
- *Redukcja danych z kolei pozwala zredukować rozmiar danych poprzez np. agregację, pewnych cech, eliminację tych nadmiarowych, poprzez grupowanie również.*
- *Oczywiście możemy stosować jednocześnie wszystkie te techniki.*

Czym mogą być dane…

- Dokumenty tekstowe
- Bazy danych ilościowych (numerycznych) i nienumerycznych
- Obrazy, multimedia
- Sekwencje DNA
- Inne.

Struktura zbioru danych

	płeć	Data_ur	Zawód	wzrost
Id_23	K	1978-03-09	Informatyk	169
Id_24	M	1977-05-20	Informatyk	190
Id_25	M	1965-05-04	Prawnik	173
Id_26	K	1982-02-02	dziennikarz	167

Cechy jakościowe

	płeć
Id_23	K
Id_24	M
Id_25	M
Id_26	K

- Płeć, grupa zawodowa… to atrybuty tzw. jakościowe (nominalne)
- Nie pozwolą nam policzyć wartości średniej, minimalnej, maksymalnej.
- Nie można więc uporządkować wartości w takim zbiorze.

Cechy ilościowe

	wzrost
Id_23	169
Id_24	190
Id_25	173
Id_26	167

- Wzrost, waga, dochód, liczba godzin…. to atrybuty tzw. ilościowe (numeryczne)
- Pozwolą nam policzyć wartości średniej, minimalnej, maksymalnej.
- Można więc uporządkować takie wartości od najmniejszej do największej.
- **Czy jest jakaś wada ?**

Typy danych

- W polskiej systematyce podręcznikowej dzielimy typy danych (zmienne) na:
- ilościowe (mierzalne) - np. wzrost, masa, wiek
  - ciągłe - np. wzrost, masa, wiek (w rozumieniu ilości dni między datą urodzin a datą badania)
  - porządkowe (quasi-ilościowe) - np. klasyfikacja wzrostu: (niski,średni,wysoki)
  - skokowe (dyskretne) - np. ilość posiadanych dzieci, ilość gospodarstw domowych, wiek (w rozumieniu ilości skończonych lat)
- jakościowe (niemierzalne) - np. kolor oczu, płeć, grupa krwi

kobieta mężczyzna

Przykład pomiaru nominalnego: **płeć**

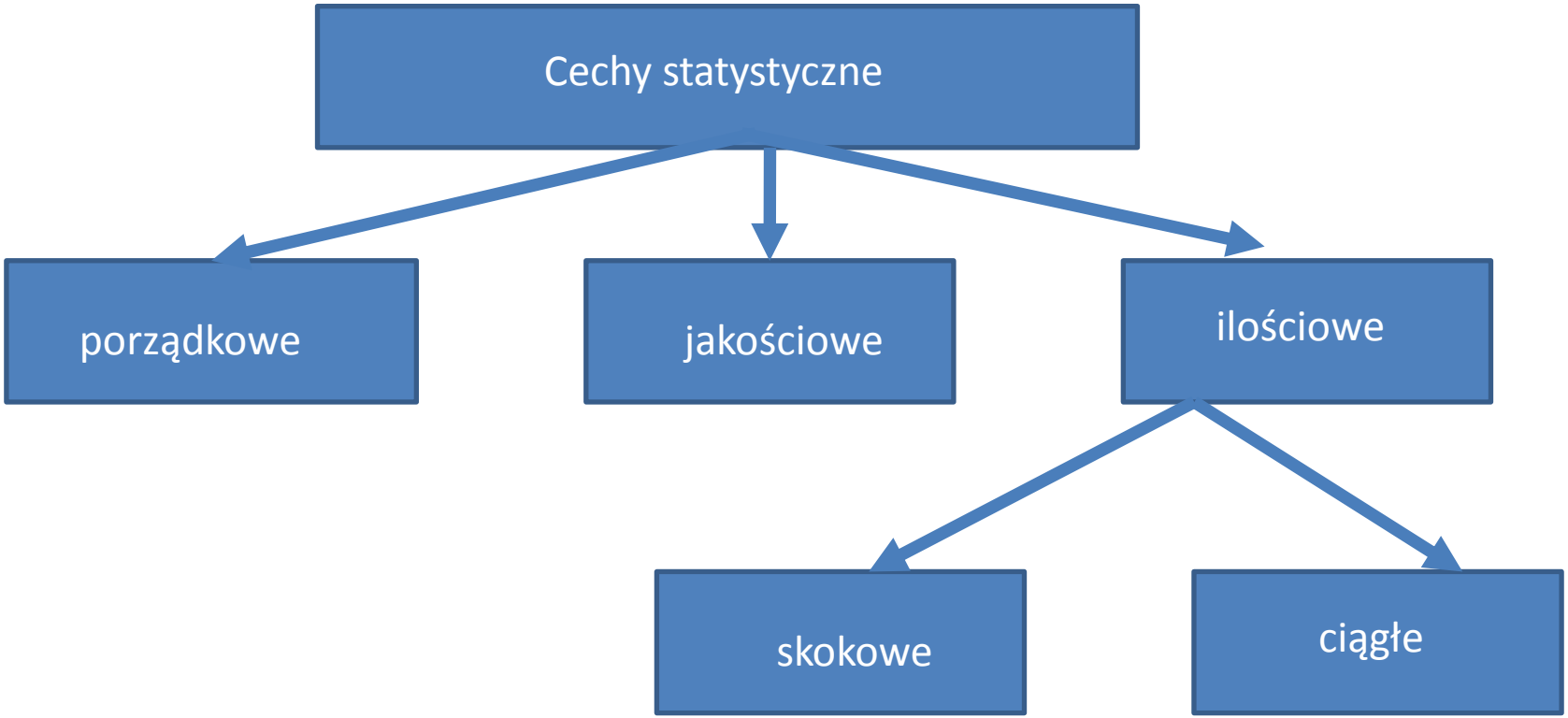
niskie średnie wysokie

Przykład pomiaru porządkowego: **stopień zaangażowania w trening sportowy**

0 25 m 50 m 75 m 100 m 125 m

Przykład pomiaru ilorazowego: **odległość w skokach narciarskich [m]**

Może więc lepszy taki podział ?



- **Cechy jakościowe** (niemierzalne) to takie, których nie można jednoznacznie scharakteryzować za pomocą liczb (czyli nie można zmierzyć). Efekt to podział zbioru na podzbiory rozłączne, np. płeć, grupę krwi, kolor włosów, zgon lub przeżycie, stan uodpornienia przeciwko ospie (zaszczepiony lub nie) itp. W przypadku grupy krwi rezultat pomiaru będzie następujący: n1 pacjentów ma grupę krwi A, n2 pacjentów - grupę krwi B, n3 pacjentów - grupę AB i n4 - grupę O.
- **Cechy porządkowe** umożliwiają porządkowanie wszystkich elementów zbioru wyników. Cechy takie najlepiej określa się przymiotnikami i ich stopniowaniem, np. dla wzrostu: "niski", "średni" lub "wysoki".
- **Cechy ilościowe** (mierzalne) to takie, które dadzą się wyrazić za pomocą jednostek miary w pewnej skali. Cechami mierzalnymi są na przykład: wzrost (w cm), waga (w kg), stężenie hemoglobiny we krwi (w g/dl), wiek (w latach) itp.
  - **Cecha ciągła** to zmienna, która może przyjmować każdą wartość z określonego skończonego przedziału liczbowego, np. wzrost, masa ciała czy temperatura. **Cechy skokowe** mogą przyjmować wartości ze zbioru skończonego lub przeliczalnego (zwykle całkowite), na przykład: liczba łóżek w szpitalu, liczba krwinek białych w 1 ml krwi.

Cecha objaśniana a objaśniająca ?

- **Cecha objaśniana** to ta, której wartość próbujemy określić na podstawie wartości cech objaśniających.
- **Cecha objaśniająca** – to cecha opisująca obserwację w zbiorze.

$$y = f(x)$$

Cecha objaśniana

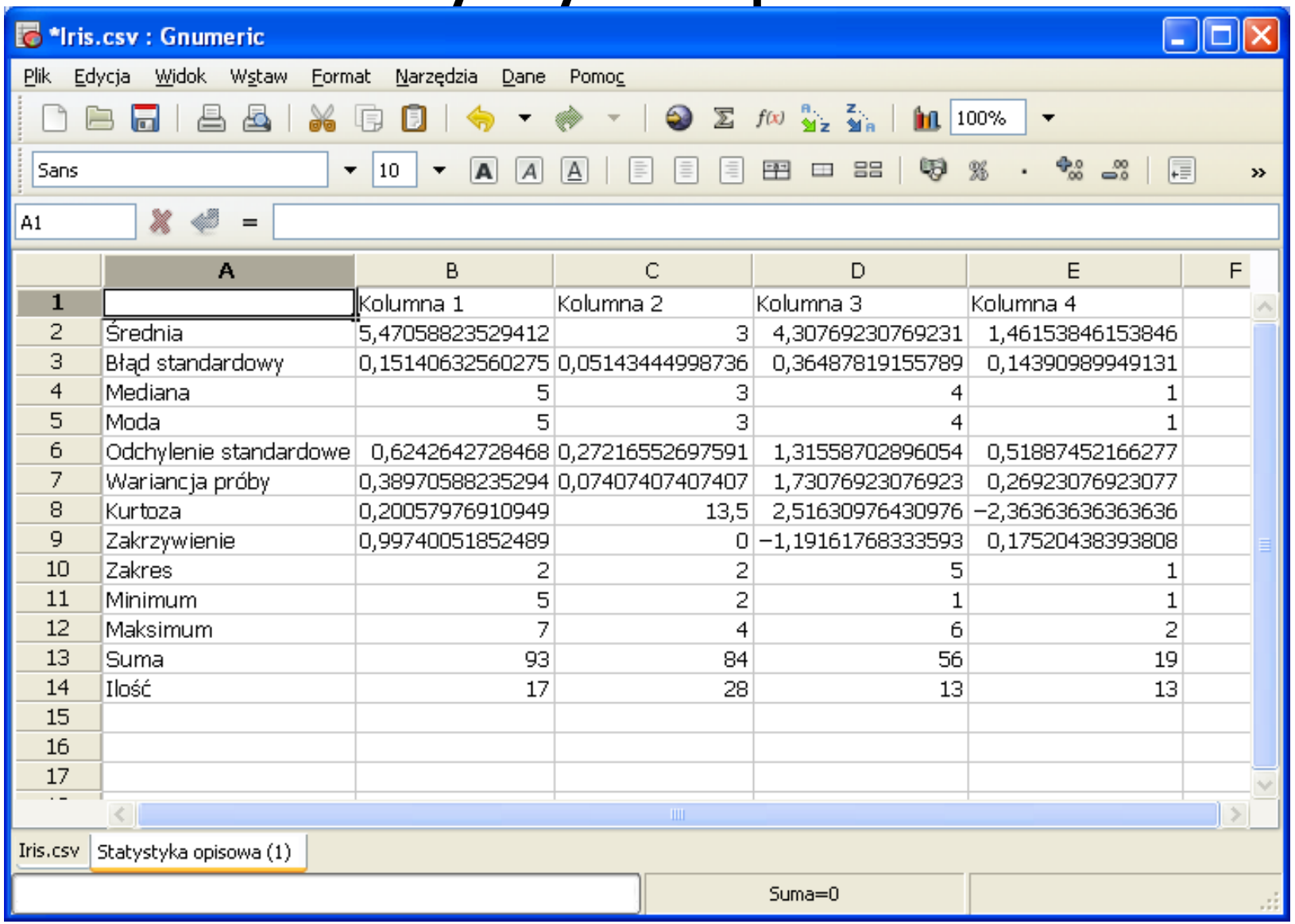
Cecha objaśniająca

Opis danych

Różne statystyczne metody:

1. Opisowe statystyki jak średnia, mediana, minimum, maksimum, moda, odchylenie standardowe, wariancja.
2. Graficzne metody (wykresy): histogram, wykres pudełkowy, wykresy rozrzutu.

Statystyka opisowa



	A	B	C	D	E	F
1		Kolumna 1	Kolumna 2	Kolumna 3	Kolumna 4	
2	Średnia	5,47058823529412	3	4,30769230769231	1,46153846153846	
3	Błąd standardowy	0,15140632560275	0,05143444998736	0,36487819155789	0,14390989949131	
4	Mediana	5	3	4	1	
5	Moda	5	3	4	1	
6	Odchylenie standardowe	0,6242642728468	0,27216552697591	1,31558702896054	0,51887452166277	
7	Wariancja próby	0,38970588235294	0,07407407407407	1,73076923076923	0,26923076923077	
8	Kurtoza	0,20057976910949	13,5	2,51630976430976	−2,36363636363636	
9	Zakrzywienie	0,99740051852489	0	−1,19161768333593	0,17520438393808	
10	Zakres	2	2	5	1	
11	Minimum	5	2	1	1	
12	Maksimum	7	4	6	2	
13	Suma	93	84	56	19	
14	Ilość	17	28	13	13	
15						
16						
17						

Co nam daje wartość średnia ?

- Ile zarabiają dyrektorzy w działach sprzedaży?

  Średnia zarobków dyrektorów sprzedaży wynosi 12161 PLN.

- Czy dyrektorzy to generalnie bogaci ludzie?
- Czy można określić ile zarabia konkretny dyrektor?
- Czy można obliczyć średnią płeć dyrektorów?

Co nam daje mediana czy moda ?

- W jakim przedziale mieszczą się zarobki większości dyrektorów? (mediana)
- Ile zarabia „przeciętny” dyrektor? (mediana nie BO z bycia w środku nie wynika bycie „przeciętnym”), (moda tak, bo przeciętny to najczęściej występujący)
- Czy prawie wszyscy dyrektorzy to bogaci ludzie? (nie BO nie wiemy NIC o całości badanej grupy)
- Czy większość dyrektorów to bogaci ludzie? (mediana)
- Czy zarobki dyrektorów różnią się mocno od siebie? (nie BO nie wiemy NIC o całości badanej grupy)
- Jakie zarobki są najczęstsze wśród dyrektorów? (nie BO to co jest w środku nie musi być najbardziej popularne), moda – tak.

Czego nam nie powie moda ?

- W jakim przedziale mieszczą się zarobki większości dyrektorów?
- Czy większość dyrektorów to bogaci ludzie? (BO najczęstsza wartość wcale nie musi dotyczyć większości)
- Czy prawie wszyscy dyrektorzy to bogaci ludzie? (BO jeśli nie wiemy nic o większości, to tym bardziej o prawie wszystkich)
- Czy zarobki dyrektorów różnią się mocno (Od siebie? BO nie wiemy nic o całości grupy)

Graficzny opis danych

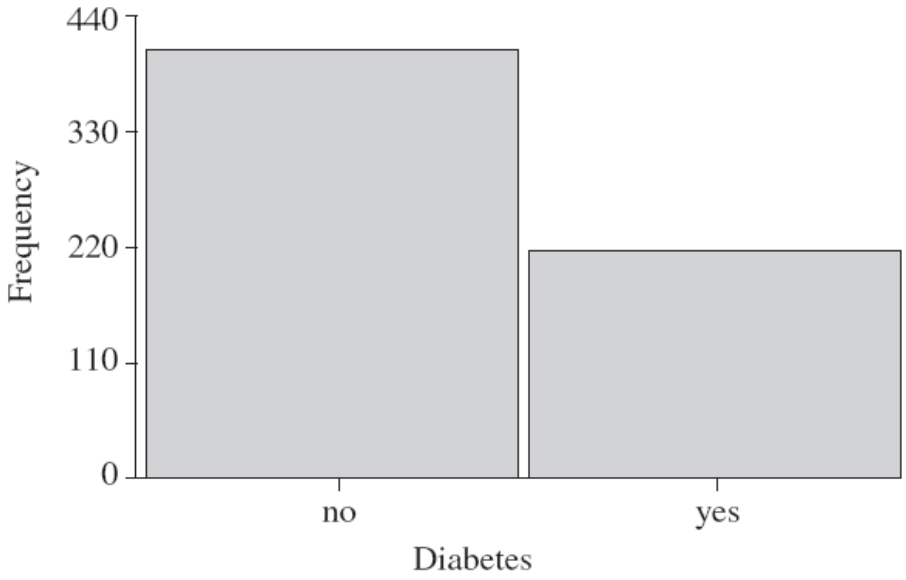
- histogramy i wykresy częstości
- wykresy rozrzutu (scatterplots)
- wykresy pudełkowe (boxplot)

Summary		Data	Uses
Graphs	**Frequency polygram**	Single variable, any type	Viewing trends, ranges, frequency distribution, and outliers
	Histogram	Single variable, any type	Viewing trends, ranges, frequency distribution, and outliers
	Scatterplot	Two ratio or interval variables	Viewing relationships between continuous variables and outliers
	Box plot	Single ratio, or interval variable	Viewing ranges, frequency distributions, and outliers
	Multiple graphs	Data dependent on individual graph	Viewing multi-dimensional relationships, multi-dimensional summaries, and comparisons

histogramy

Histogram to jeden z graficznych sposobów przedstawienia rozkładu empirycznego cechy.

Składa się z szeregu prostokątów umieszczonych na osi współrzędnych. Na osi „X” mamy przedziały klasowe wartości cechy np. dla atrybutu płeć: „K, M”, na osi „Y” liczebność tych przedziałów.

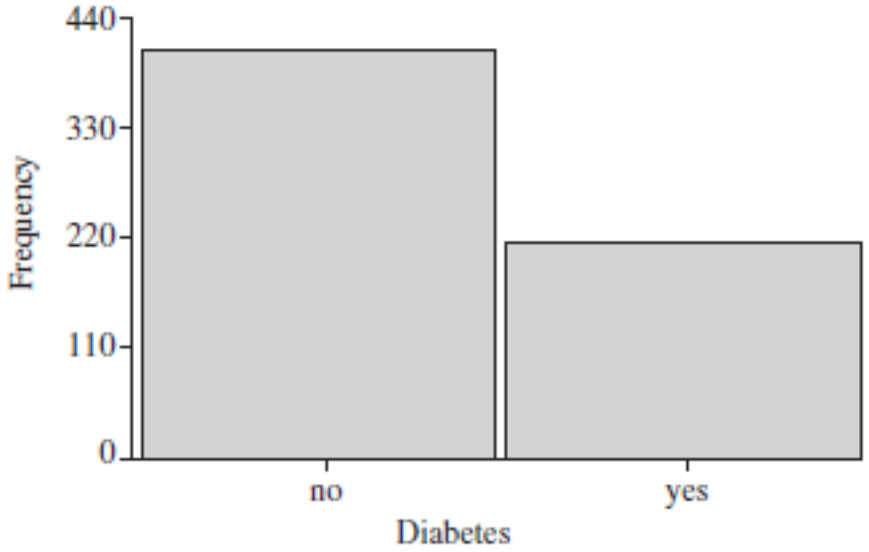


- Dla danych jakościowych
- Porządkują wiedze o danych analizowanych
- Pokazują odchylenia w danych
- Pokazują dane dominujące w zbiorze

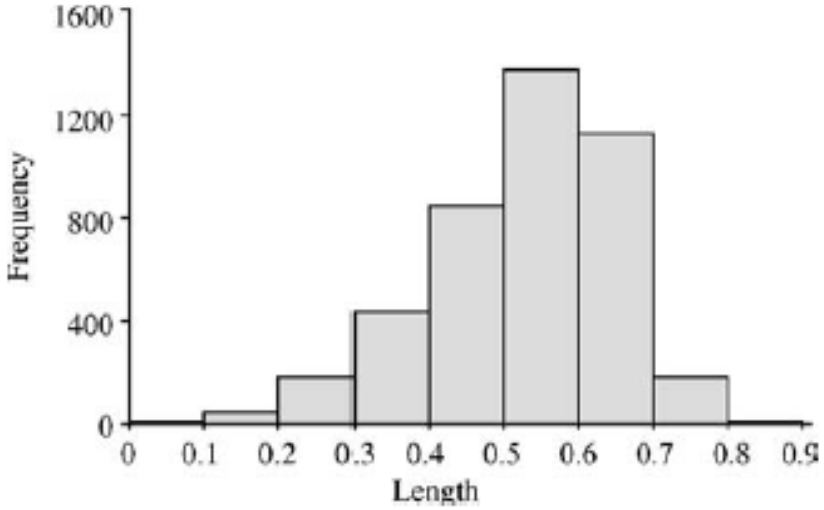
Histogram

- Najpopularniejsza statystyka graficzna. Przedstawia liczności pacjentów w poszczególnych przedziałach (nazywanych tez kubełkami) danej zmiennej.
- Domyślnie w funkcji histogram liczba kubełków dobierana jest w zależności od liczby obserwacji jak i ich zmienności.
- Możemy jednak subiektywnie wybrać interesującą nas liczbę kubełków.

Histogram a rodzaj danych



Dane jakościowe



Dane ilościowe

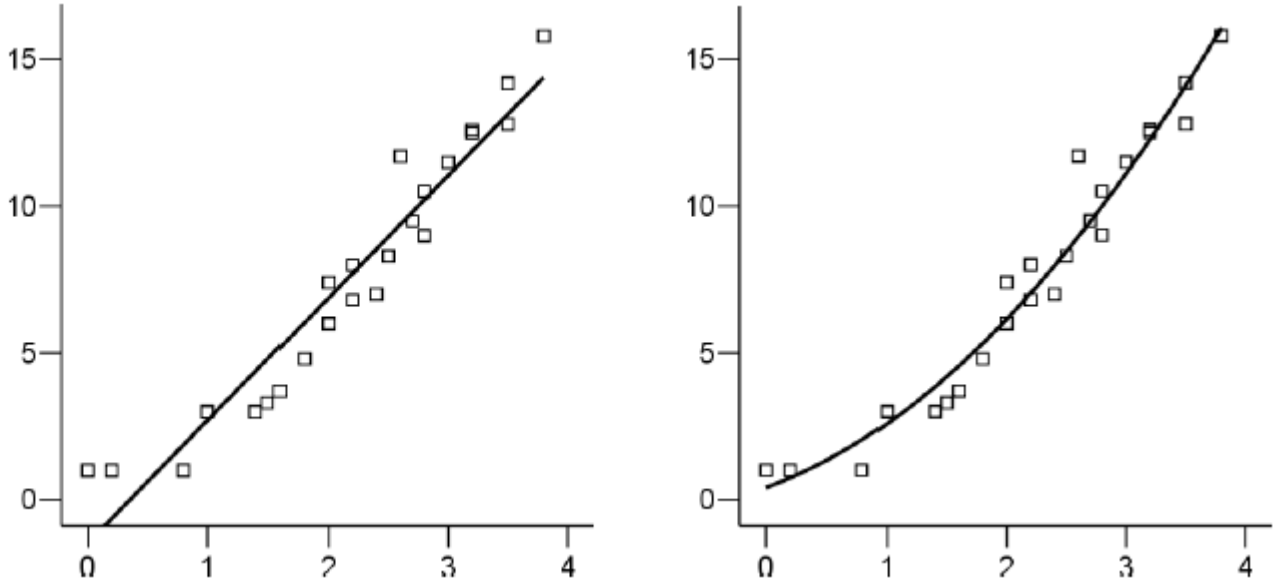
Wykres punktowy (rozrzutu)

W diagnostyce powiązań między cechami[25] we wstępnym rozpoznaniu charakteru i kierunku zależności par cech ilościowych pomocny jest najprostszy z wykresów korelacyjnych tzw. **wykres punktowy** w programie SPSS zwany **wykresem rozrzutu** (*Scatter Plots*).

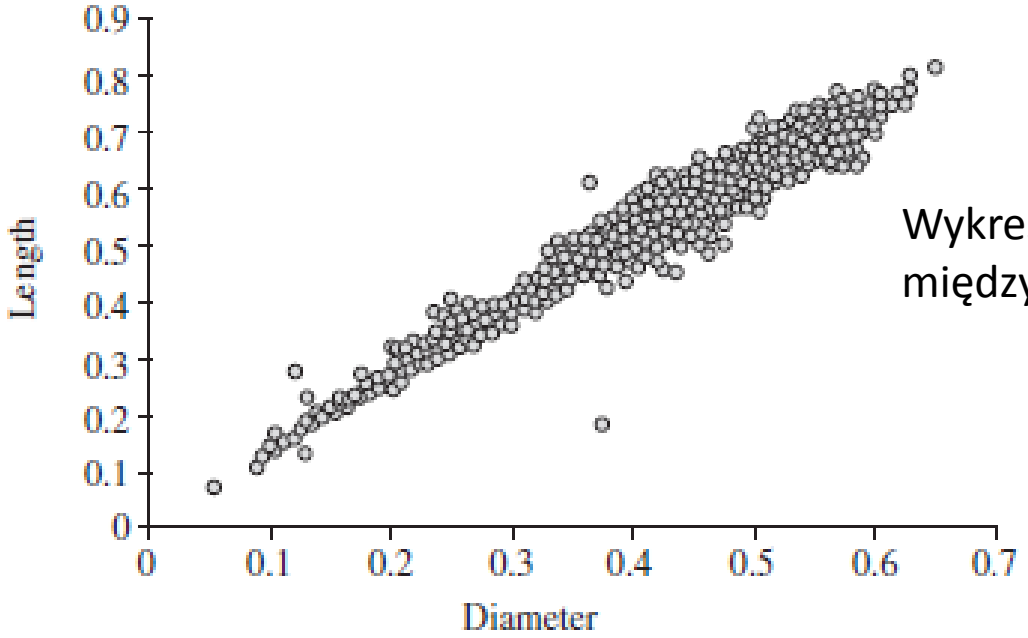
Dla tych samych danych

a *b*

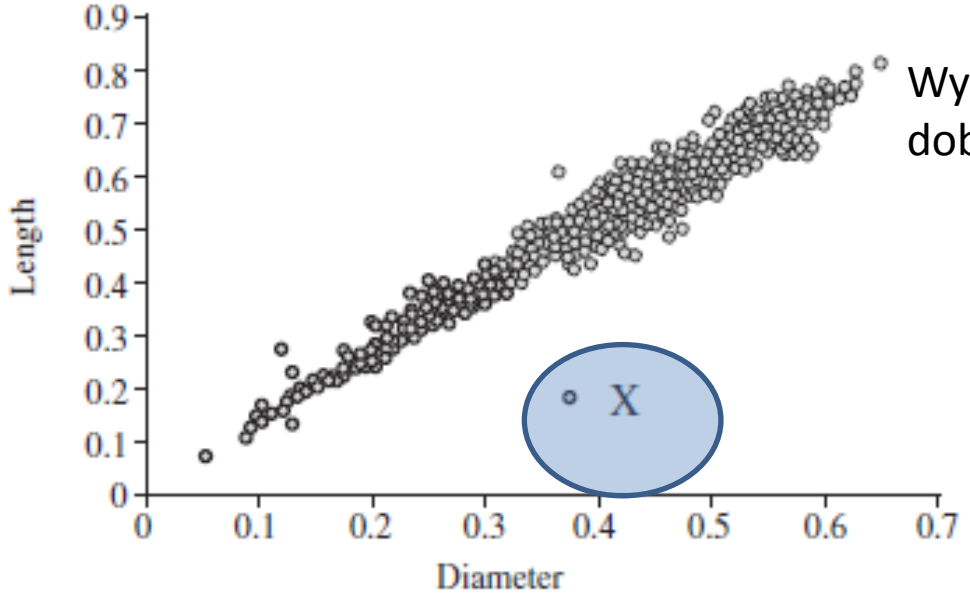
Wykres rozrzutu dodatniej zależności cech X i Y wg typu $\hat{y}_t = a + bx_t$, $\hat{y}_t = a + bx_t + cx_t^2$:
a) zależność prostoliniowa;
b) zależność kwadratowa



O tym która linia regresji lepiej odwzrowuje dane decyduje współczynnik determinacji R^2.

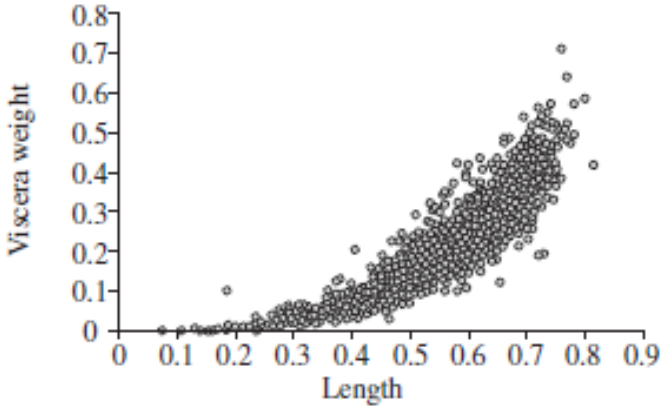


Wykresy rozrzutu pokazują relację między daną na osi X a daną na osi Y

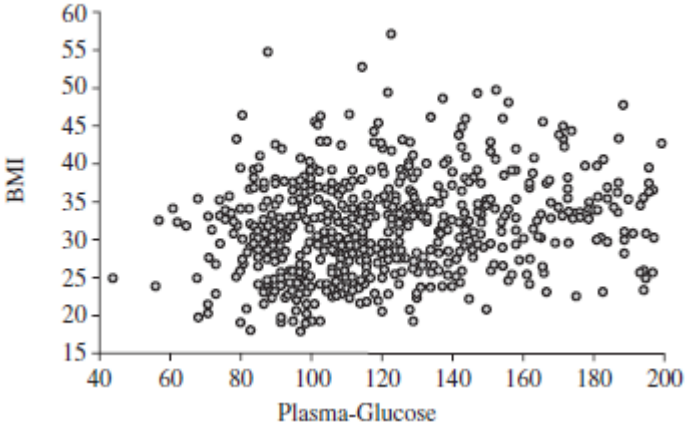


Wykresy rozrzutu też wskazują dobrze odchylenia w danych

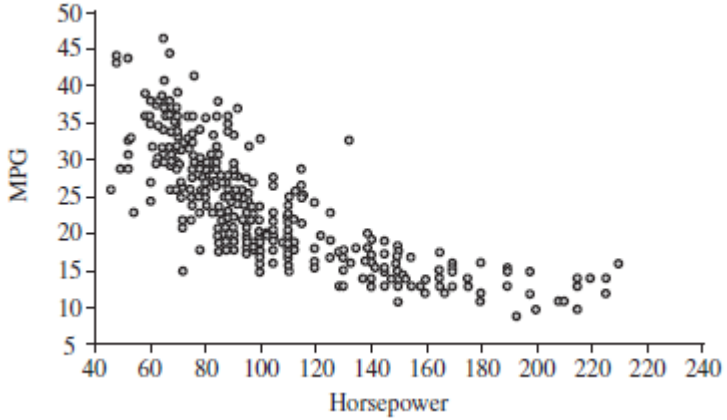
Typ korelacji



Nieliniowa zależność danych



Scatterplot showing no discernable relationship



Korelacja ujemna

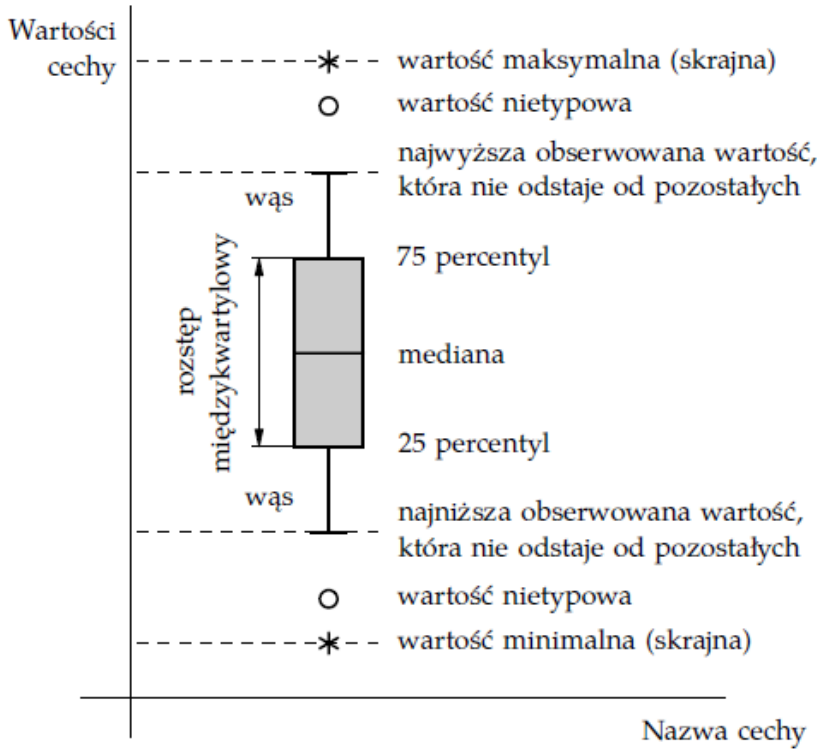
Wykres pudełkowy

- Wykres pudełkowy można wyznaczać dla pojedynczej zmiennej, dla kilku zmiennych lub dla pojedynczej zmiennej w rozbiciu na grupy.
- Wykres przedstawia medianę (środek pudełka), kwartyle (dolna i górna granica pudełka), obserwacje odstające (zaznaczane kropkami) oraz maksimum i minimum po usunięciu obserwacji odstających.
- Wykres pudełkowy jest bardzo popularną metodą prezentacji zmienności pojedynczej zmiennej.

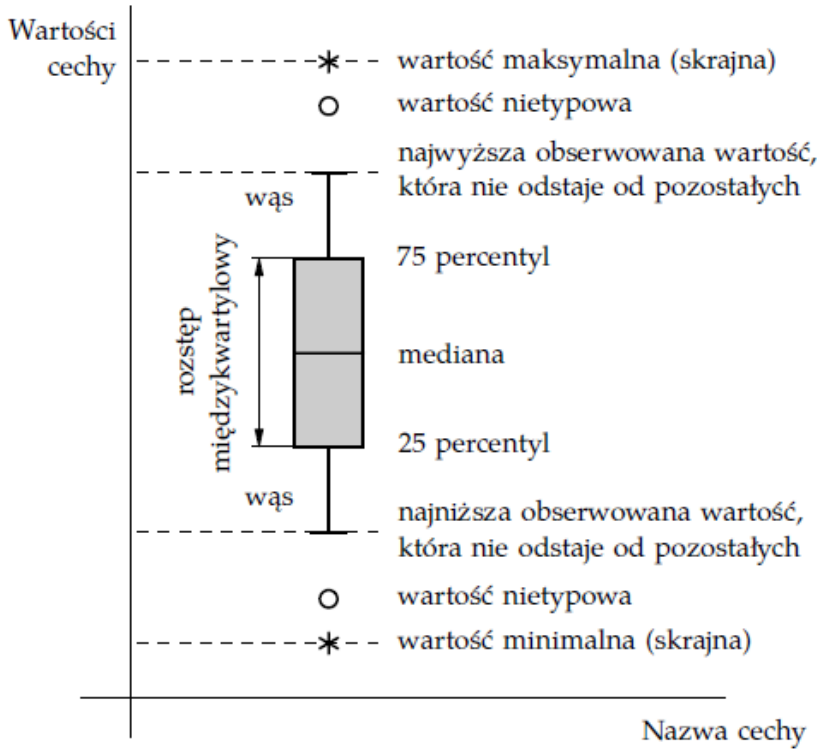
Wykres skrzynkowy (*Box-and-Whisker Plot, Boxplot*), zwany też **pudełkowym** lub **skrzynką z wąsami**, przedstawia rozkład uporządkowanych wartości cechy pod postacią wykorzystanego w nazwie prostego przedmiotu. Ułatwia diagnostykę rozproszenia wartości cechy oraz charakteru (typu) skośności rozkładu cechy.

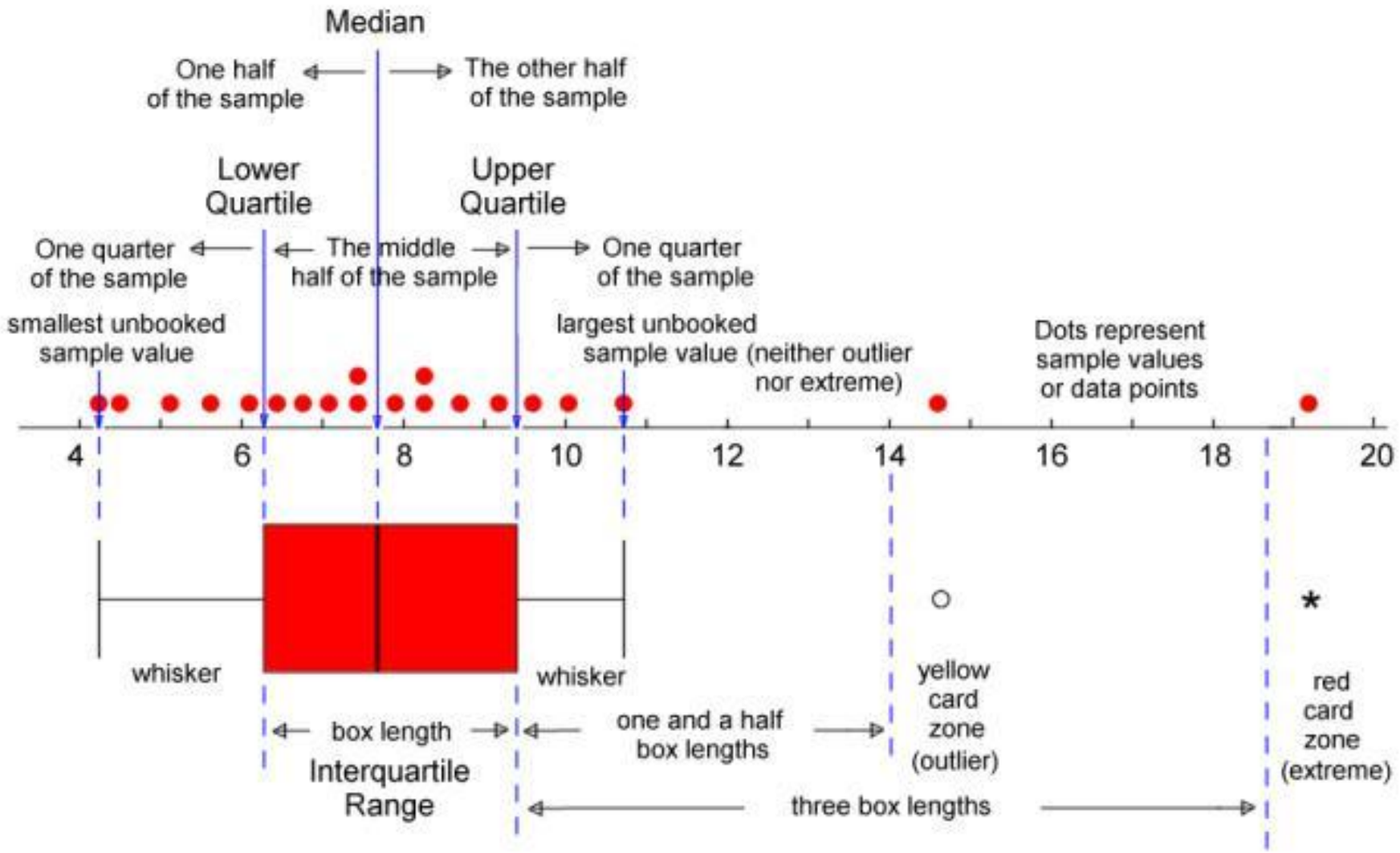
Z wykresu skrzynkowego nietrudno odczytać:

- położenie wartości środkowej (mediany);
- wartości kwartyli (pierwszego i trzeciego);
- położenie wariantów cechy, które nie odstają od tendencji centralnej;
- występowanie nietypowych wariantów cechy;
- występowanie ekstremalnych wariantów cechy.



Długość *rozstępu międzykwartylowego* (*hspread*) jest różnicą między krawędziami skrzynki (pudełka). W obszarze pudełka mieści się 50% wartości cechy. *Wąsy* skrzynki pokazują wartości cechy, jakie mieszczą się wewnątrz 1.5 długości zawiasu (choć 1.5 długości zawiasu może wykraczać poza wąsy). Symbolem „∗" oznaczone są nietypowe wartości ekstremalne oddalone od 25. (75.) percentyla dalej niż 3 długości pudełka, zaś symbolem „∘" — nietypowe wartości, oddalone od 25. (75.) percentyla dalej niż 1.5 długości pudełka.




Median
One half of the sample
The other half of the sample
Lower Quartile
Upper Quartile
One quarter of the sample
The middle half of the sample
One quarter of the sample
smallest unbooked sample value
largest unbooked sample value (neither outlier nor extreme)
Dots represent sample values or data points
4
6
8
10
12
14
16
18
20
whisker
whisker
box length
Interquartile Range
one and a half box lengths
yellow card zone (outlier)
red card zone (extreme)
three box lengths

Co można odczytać z wykresów?

	Boxplot	Histogram
Kwantyl	tak	nie
Mediana	tak	nie
Wartość min	tak	tak
Wartość max	tak	tak
Wartość cechy	tak	tak
Liczebność	nie	tak
Częstość	nie	tak
Wzajemna korelacja zmiennych	nie	tak

Przed analizą danych – obróbka danych

- Dane są przeważnie:
- Nieobrobione
- Niekompletne
- Zaszumione
- To sprawia konieczność czyszczenia i przekształcania danych.

- Np. baza danych zawiera pola przestarzałe lub zbędne, rekordy z brakującymi wartościami, punkty oddalone, dane z niewłaściwym formatem.

Czyszczenie danych i ich przekształcanie ma na celu realizację zadania nazywanego często: Minimalizacją GIGO

- GIGO – garbage in garbage out

Czyszczenie danych

	A	B	C
1	**id**	**kod**	**liczba_mieszkańców**
2	1	41400	12 tys
3	2	41310	13 tys
4	3	22543	4.5 tys
5	4	**2687**	45 tys
6	5	44789	12.5 tys
7	6	87211	33 tys

- To być może miał być kod „02687” lecz niektóre programy jeśli pole jest typem numerycznym to zero na początku nie będzie ujmowane.

Maski wprowadzania

- Maska wprowadzania to zestaw znaków literałowych i znaków masek umożliwiający sterowanie zakresem danych, które można wprowadzać w polu. Maska wprowadzania może na przykład wymagać od użytkowników wprowadzania dat czy numerów telefonów zgodnie z konwencją przyjętą w danym kraju/regionie — tak jak w poniższych przykładach:
- RRRR-MM-DD
- (___) ___-____ wew. ___

Maski takie

- ułatwiają zapobieganie wprowadzaniu przez użytkowników nieprawidłowych danych (na przykład wpisaniu numeru telefonu w polu daty).
- Dodatkowo zapewniają one spójny sposób wprowadzania danych przez użytkowników, co z kolei ułatwia wyszukiwanie danych i obsługę bazy danych.

Gdzie stosować maski:

- do pola tabeli typu Data/godzina lub formantu pola tekstowego w formularzu powiązanego z polem typu Data/godzina.
- Masek wprowadzania nie można jednak używać bezkrytycznie.
- Maski wprowadzania można domyślnie stosować do pól tabeli z typem danych ustawionym na Tekst, Liczba (oprócz identyfikatora replikacji), Waluta i Data/godzina.
- Maski wprowadzania można także stosować dla formantów formularza, takich jak pola tekstowe, powiązanych z polami tabeli, dla których ustawiono te typy danych.

http://office.microsoft.com/pl-pl/access-help/tworzenie-maski-wprowadzania-do-wprowadzania-wartosci-pol-lub-formantow-w-okreslonym-formacie-HA010096452.aspx#BM1

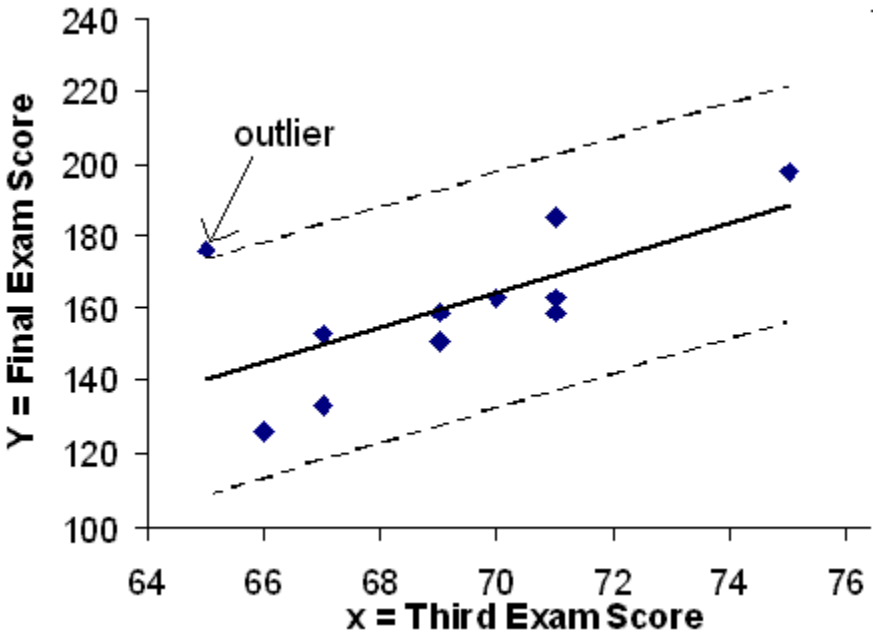
Co z danymi oddalonymi ?

	A	B	C
1	**id**	**kod**	**dochód**
2	1	41400	12 tys
3	2	41310	13 tys
4	3	22543	4.5 tys
5	4	**2687**	45 tys
6	5	44789	**-12.5 tys**
7	6	87211	33 tys

- Błędne dane typu dochód z minusem na początku: to błąd we wprowadzaniu danych, czy faktyczny ujemny dochód ?

Odchylenia w danych

- W bazie danych mogą być zawarte obiekty, które nie spełniają wymagań ogólnego modelu zachowań.
- Te obiekty nazywamy ***odchyleniami***.
- W większości przypadków obiekty takie są odrzucane jako zakłócenia, śmieci lub wyjątki.
- Niekiedy jednak identyfikacja takich odchyleń może być bardzo interesująca, na przykład w systemach wykrywania oszustw (fraud detection).
- Odchylenia mogą być wykrywane z wykorzystaniem testów statystycznych, w których przyjmowany jest określony rozkład
- prawdopodobieństwa dla danych.
- Można też stosować miary odległości, a obiekty, których odległość od utworzonych skupień jest duża traktowane są jako odchylenia



Numeryczne metody wykrywania danych oddalonych:

1. Metoda ze średniej i odchylenia standardowego
2. Metoda rozstępu międzykwartylowego

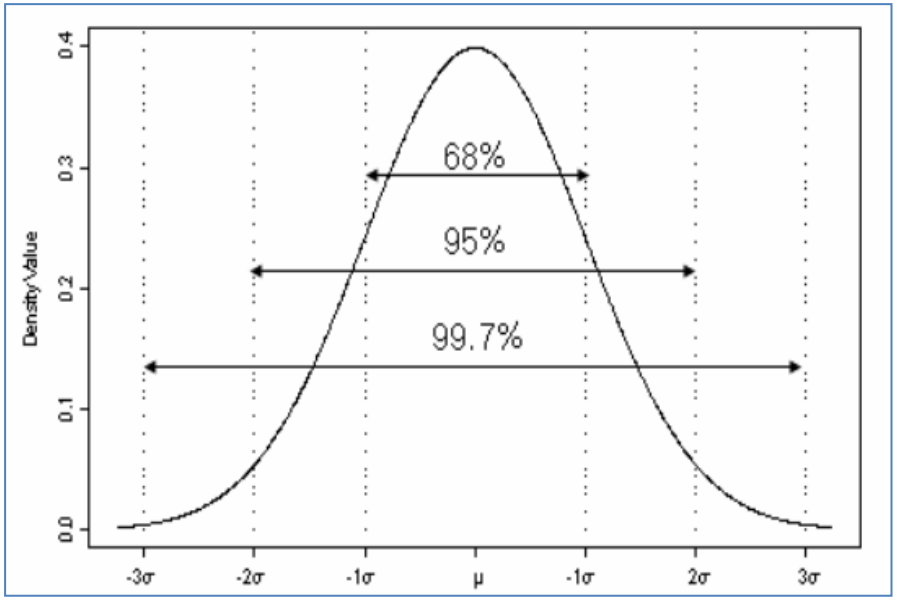
Metoda ze średniej i odchylenia standardowego

- Wyznaczamy przedział tzw. wartości typowych jako: średnia ± *k* SD gdzie *k* wynosi 2, 2.5 lub 3
- Wartości z poza tego przedziału uznajemy za odstające

Rozkład danych – metoda Tukeya

średnia ± *k* SD

- Około 68%, 95%, i 99.7% danych powinno mieścić 1,2, i 3 wielkości różnicy odchylenia standardowego i średniej.
- Więc obserwacje większe lub mniejsze niż wartość: śr_x ± 2 *SD lub śr_x ± 3 *SD uznajemy za odchylenia.



3.2, 3.4, 3.7, 3.7, 3.8, 3.9, 4, 4, 4.1, 4.2, 4.7, 4.8, 14, 15.

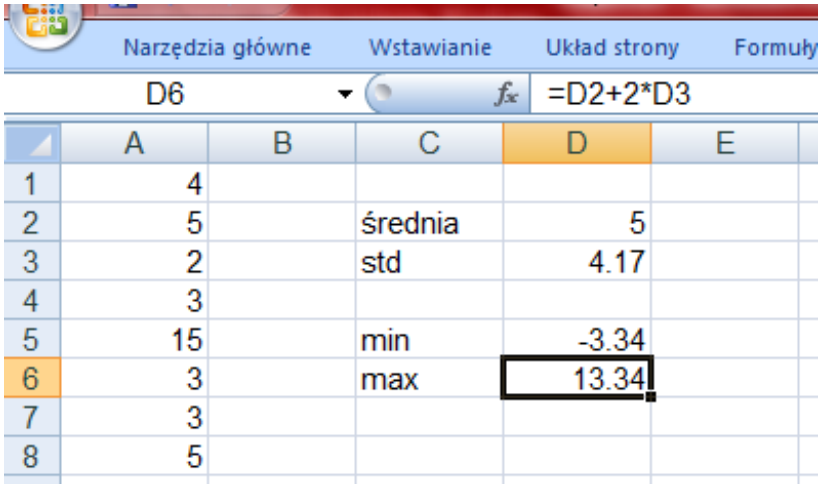
śr_x = 5.46, SD=3.86.

Dla śr_x ± 2 *SD: <-2.25, 13.18 > zaś dla śr_x ± 3*SD: <-6.11, 17.04>.

Wówczas 14 i 15 uznamy za odchylenia dla reguły: śr_x ± 2 *SD ale nie będą odchyleniami dla reguły: śr_x ± 3*SD

Metoda średniej i odchylenia standardowego

Często do wykrywania odchyleń w danych używa się wartości średniej i odchylenia standardowego. Mówi się wówczas, że jeśli jakaś wartość jest większa bądź mniejsza o wartość równą dwukrotnej wartości odchylenia standardowego od wartości średniej to należy ją uznać za odchylenie.



Rozstęp międzykwartylowy IQR

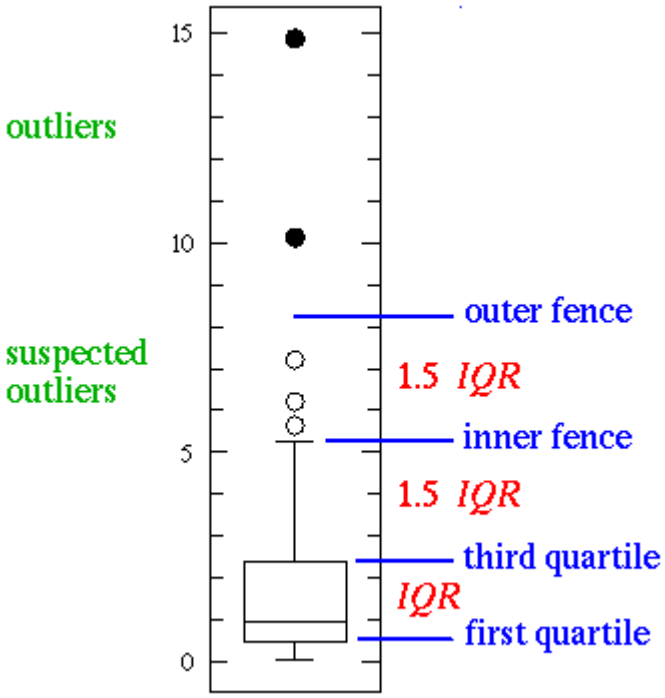
- To bardziej odporna metoda.
- Kwartyle dzielą zbiór danych na 4 części z których każda zawiera 25 % danych.
- Rozstęp międzykwartylowy to miara zmienności, która jest dużo bardziej odporna niż odchylenie standardowe
- IRQ = Q3 – Q1

Metoda rozstępu międzykwartylowego
Metoda boxplotów - Tukey (1977)

Dane wejściowe: O1, O3

Dane wyznaczane w algorytmie Tukeya:

IQR (**interquartile range**) = Q3 – Q1



<Q1 – k* IQR, Q3+ k*IQR>

Dane które nie należą do tego przedziału uznajemy za odchylenia.

Zauważmy, że wartość współczynnika „k” odpowiednio rozszerza albo zawęża ten przedział.

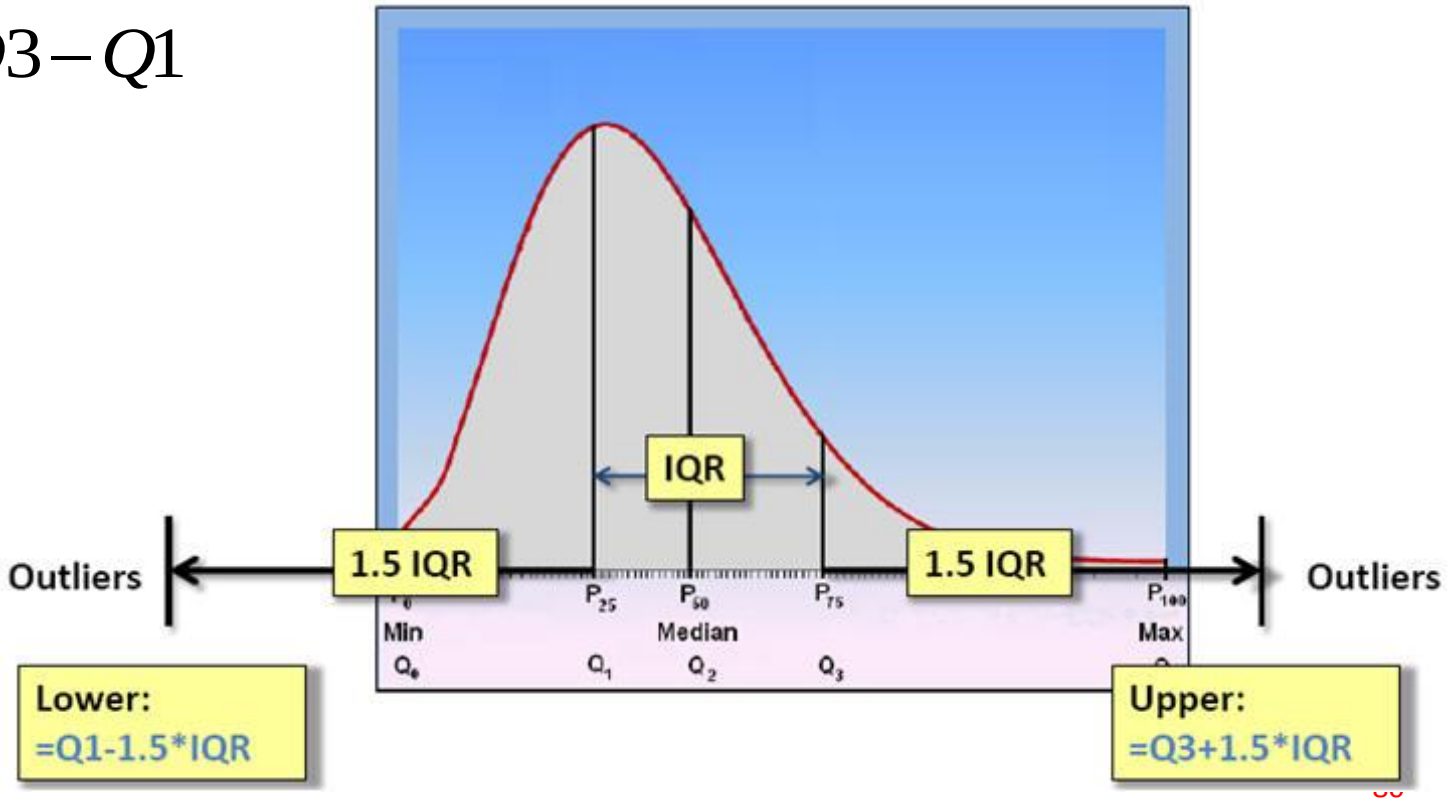
Metoda „rozstępu międzykwartylowego”

$$\min = Q1 - 1.5 * IQR$$

$$\max = Q3 + 1.5 * IQR$$

$$IQR = Q3 - Q1$$

Odchyleniem są obserwacje nie wchodzące do przedziału: $< \min, \max >$



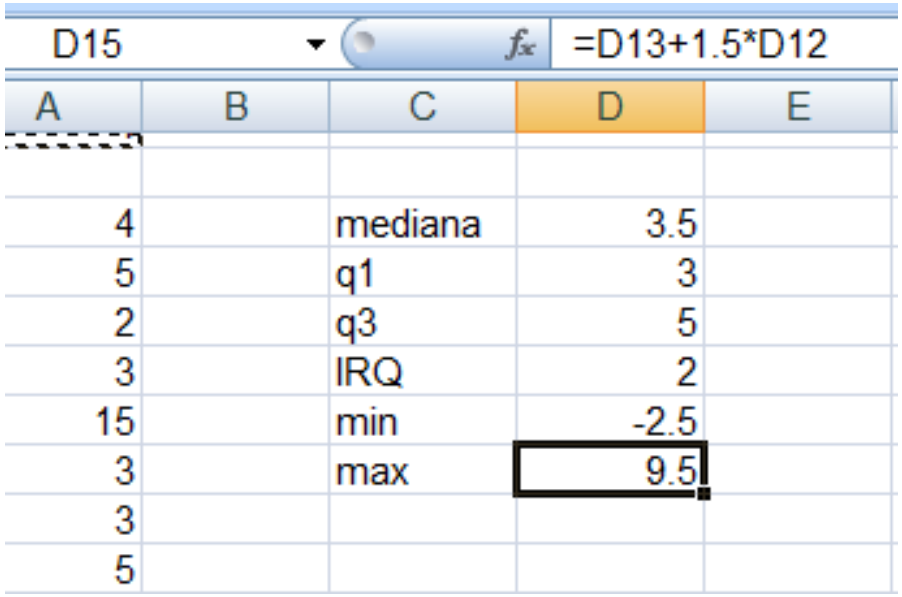
Dana jest punktem oddalonym gdy:

- Jest położona przynajmniej o 1.5 x IQR poniżej Q1

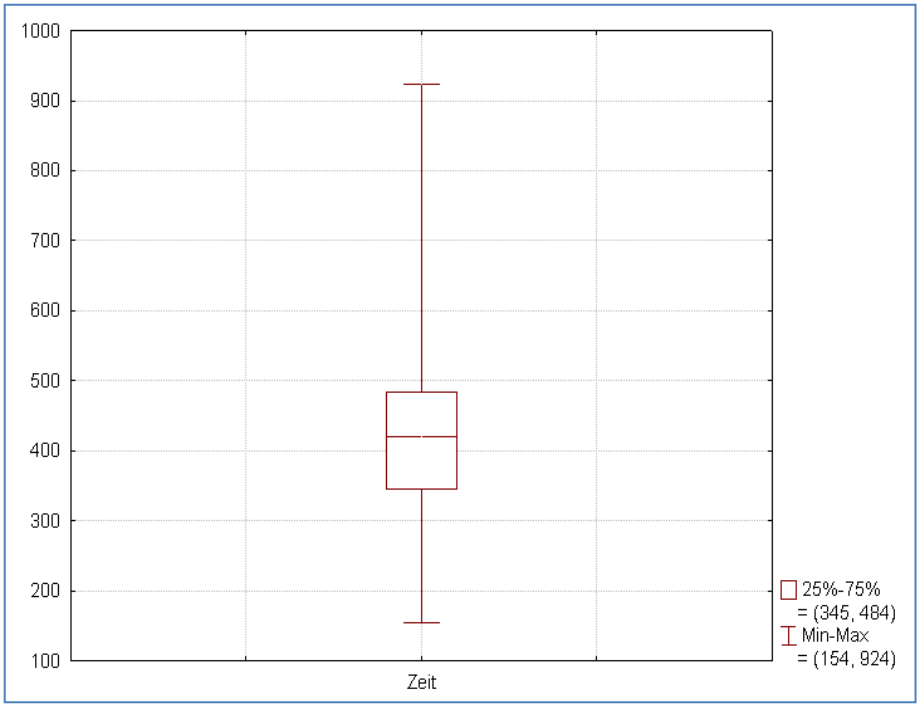
(a więc: ≤ Q1-1.5 * IQR)

- Jest położona przynajmniej o 1.5 x IQR powyżej Q3

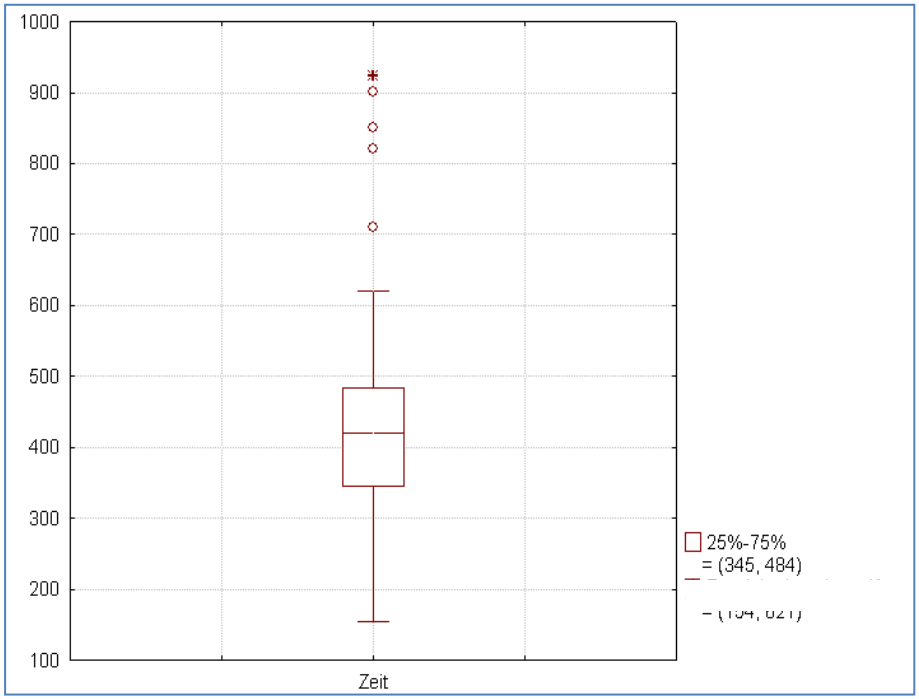
(a więc ≥ Q3+1.5 * IQR)



Boxploty



Zwykły boxplot



Boxplot Tukeya

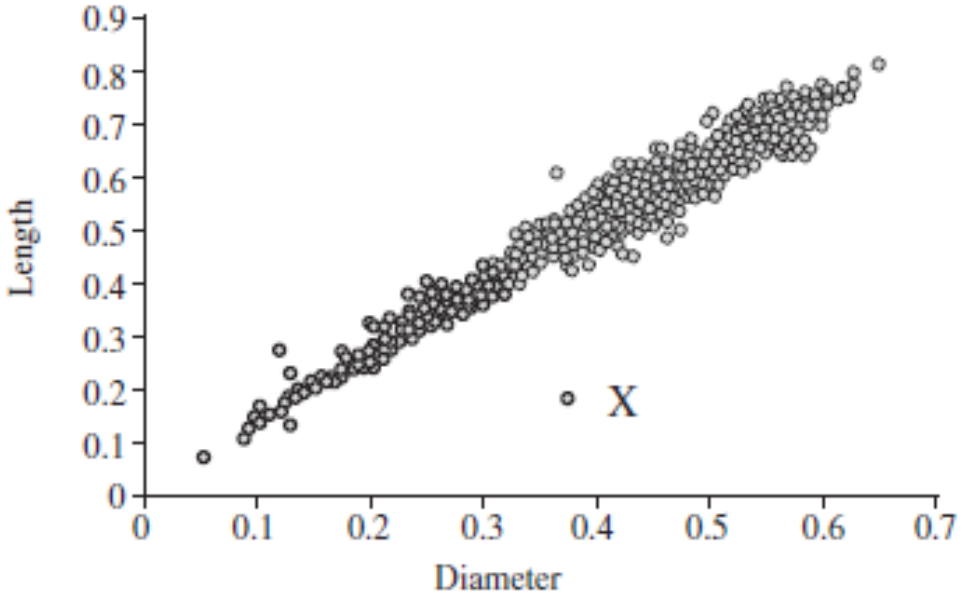
Graficzne metody wykrywania wartości oddalonych:

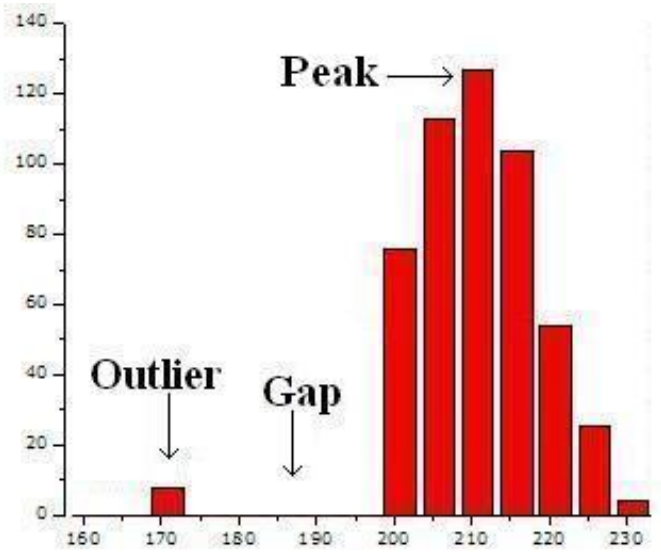
Punkty oddalone to skrajne wartości, znajdujące się blisko granic zakresu danych bądź są sprzeczne z ogólnym trendem pozostałych danych.

Metody:

Histogram lub dwuwymiarowe wykresy rozrzutu, które potrafią wskazać obserwacje oddalone dla więcej niż 1 zmiennej.

Histogram lub dwuwymiarowe wykresy rozrzutu – pozwalają wykryć odchylenia





Inne problemy z danymi?

- Np. wartość „99999” może być prawidłową daną, a może być także błędem w danych. W starszych BD pewne określone wartości oznaczały kod dla niewłaściwie wprowadzonych danych i właśnie wartość „99999” może być w tym względzie wartością oznaczającą błąd.

Złe dane

wiek	rok_urodzenia
32	1978
60	1950
34	1976
25	1985

- Np. kolumna „wiek” czy „rok_urodzenia” ?
- Czy jest jakas różnica między nimi ?
- Wiek - **źle**, rok_urodzenia - **dobrze**

Brakujące dane – bardzo poważnym problemem przy analizie danych

Nie wiadomo jaka jest przyczyna braku danych i jak z tymi brakami w danych postępować.

Powody niekompletności danych:

- atrybuty najbardziej pożądane do analizy mogą być niedostępne
- dane nie były możliwe do zdobycia w określonym czasie, co spowodowało nie zidentyfikowanie pewnych ważnych zależności
- czasami winą jest błąd pomiaru
- dane mogły być zapisane ale potem usunięte
- o prostu może brakować pewnych wartości dla atrybutów.

Metody na brakujące dane:

Są 2 możliwości:

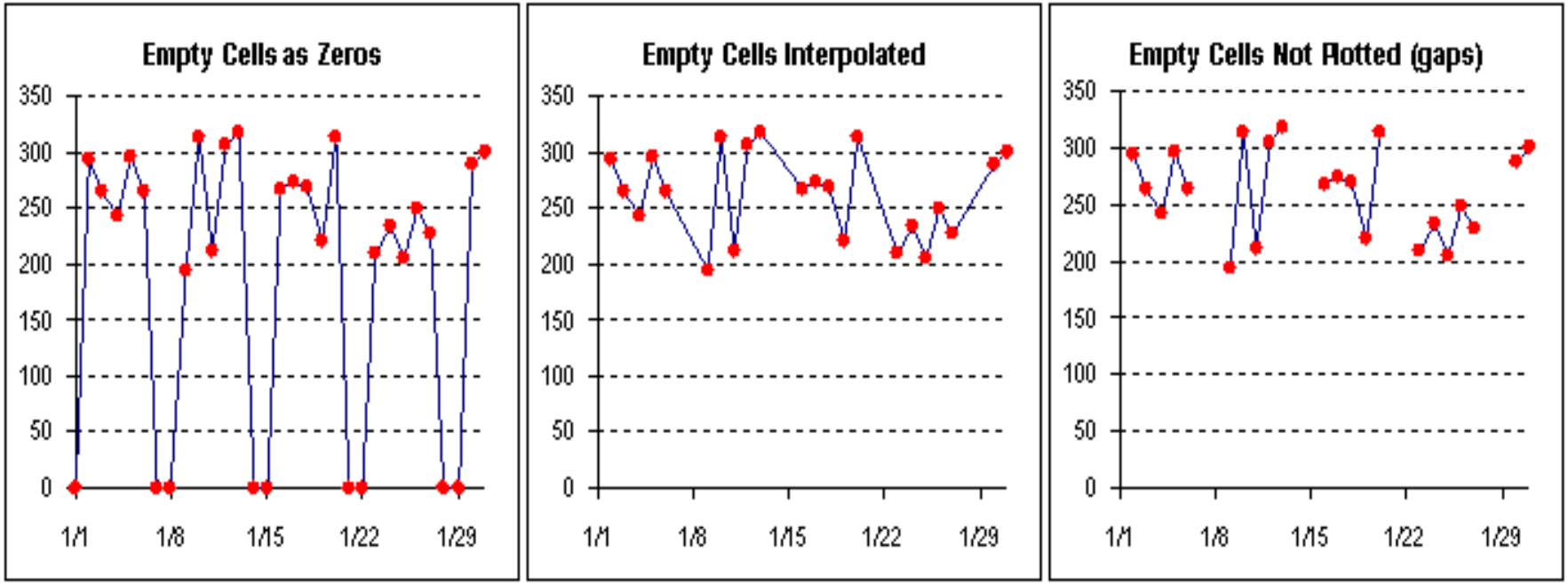
1. Pomijanie danych – niebezpieczny krok
2. Zastępowanie danych (różne metody):
    1. Zastąpienie pewną stałą podaną przez analityka
    2. Zastąpienie wartością średnią lub modalną
    3. Zastąpienie wartością losową.

Ad .1. Zastąpienie pewną stałą podaną przez analityka

- Braki w danych numerycznych zastępuje się wartością „0”
- Braki w danych tekstowych zastępuje się wartością „missing”

Ad. 2. Zastąpienie wartością średnią lub modalną

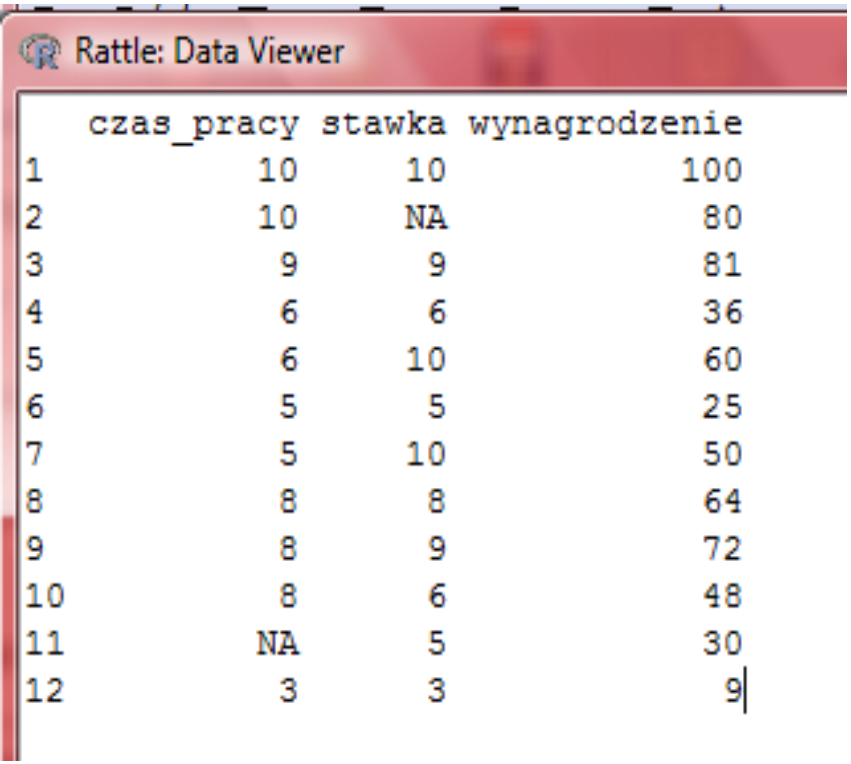
- Dane numeryczne zastępuje się wartością średnią w zbiorze danych
- Dane nienumeryczne (tekstowe) zastępuje się wartością modalną a więc wartością najczęściej występującą w zbiorze.



- w 1 przypadku dane z uwzględnieniem danych brakujących
- w 2 przypadku dane z uwzględnieniem metod interpolacji
- w 3 przypadku gdy dane brakujące są ignorowane, a więc nie są brane pod uwagę przy wykreślaniu wykresu.

R i Rattle a brakujące dane

Przypuśćmy, że mamy do czynienia ze zbiorem danych, w którym brak niektórych informacji. Konkretnie brakuje nam stawki godzinowej w wierszu 2 oraz informacji o czasie pracy w wierszu 11.



	czas_pracy	stawka	wynagrodzenie
1	10	10	100
2	10	NA	80
3	9	9	81
4	6	6	36
5	6	10	60
6	5	5	25
7	5	10	50
8	8	8	64
9	8	9	72
10	8	6	48
11	NA	5	30
12	3	3	9

W Rattle w zakładce „Transform” możemy użyć jednej z kilku metod radzenia sobie z brakami w danych:

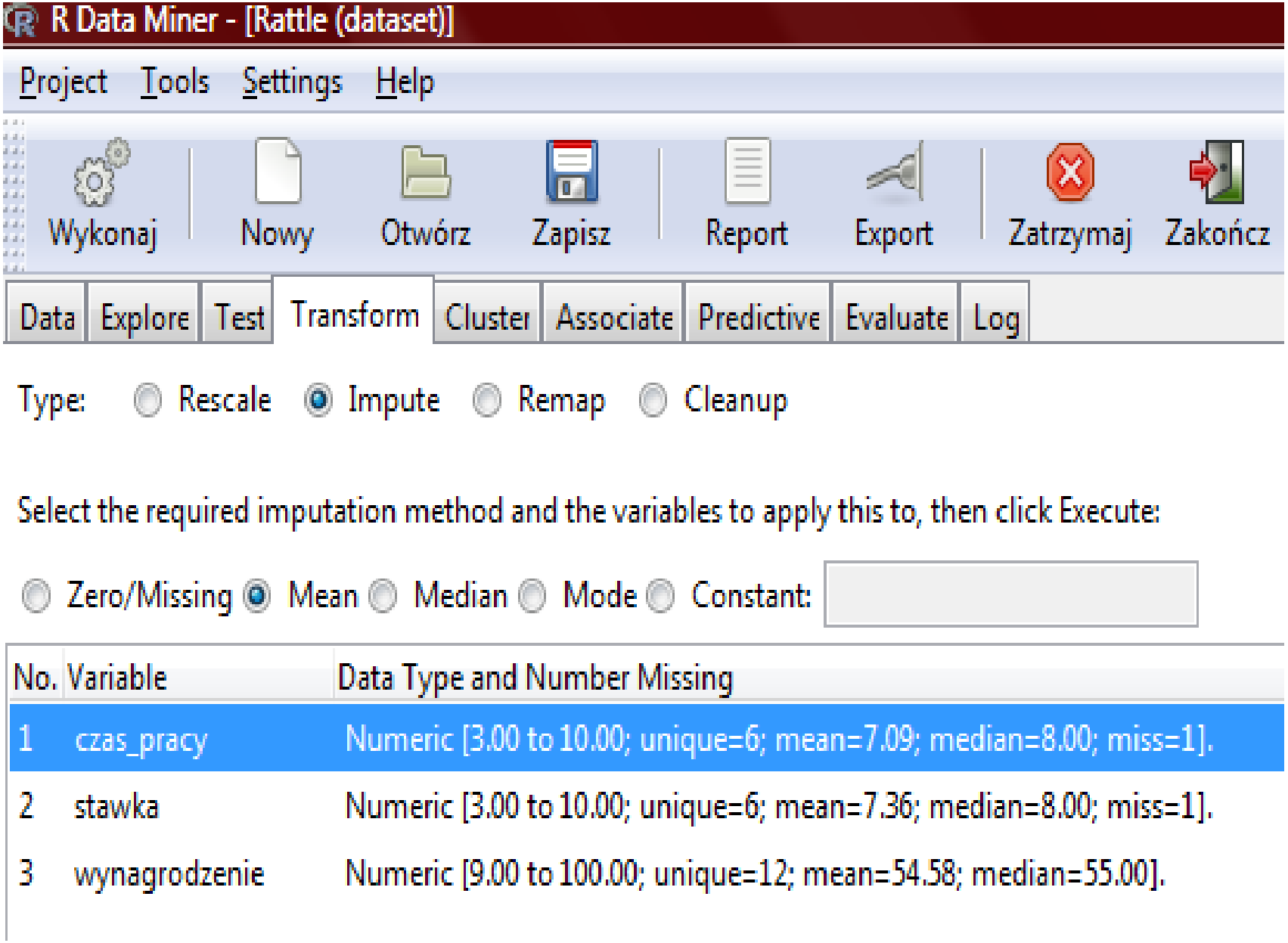
Zero/Missing – zastępowanie braków w danych wartością „0”

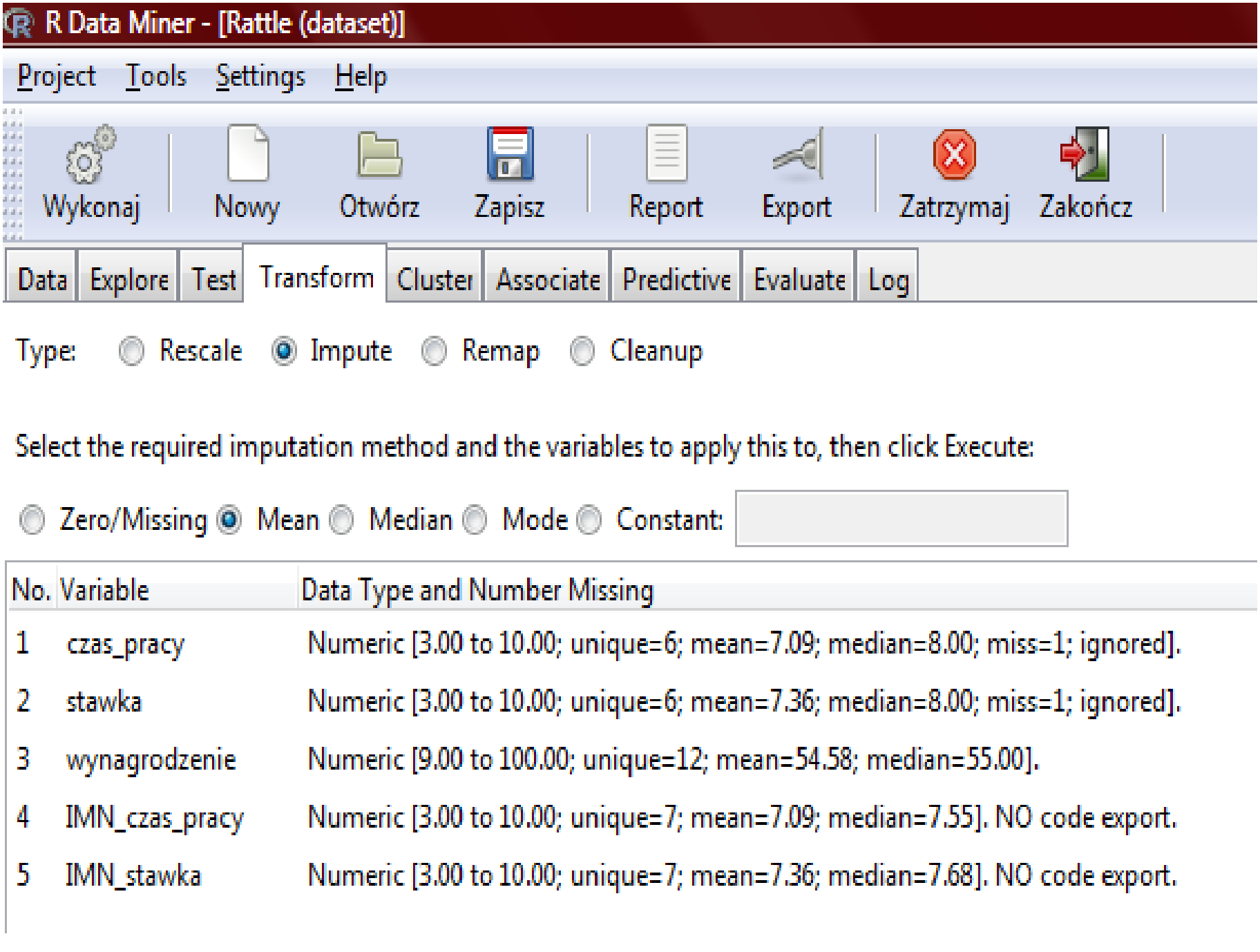
Mean – zastępowanie braków w danych wartością średnią w danym zbiorze (tutaj można rozważyć także uśrednianie w ramach danej podgrupy!!!)

Median – zastępowanie braków w danych medianą w danym zbiorze

Mode– zastępowanie braków w danych modą w danym zbiorze

Constant – stała wartość, którą będą zastępowane wszelkie braki w danych. Może to być np. wartość 0, "*unknown*", "*N/A*" lub -∞

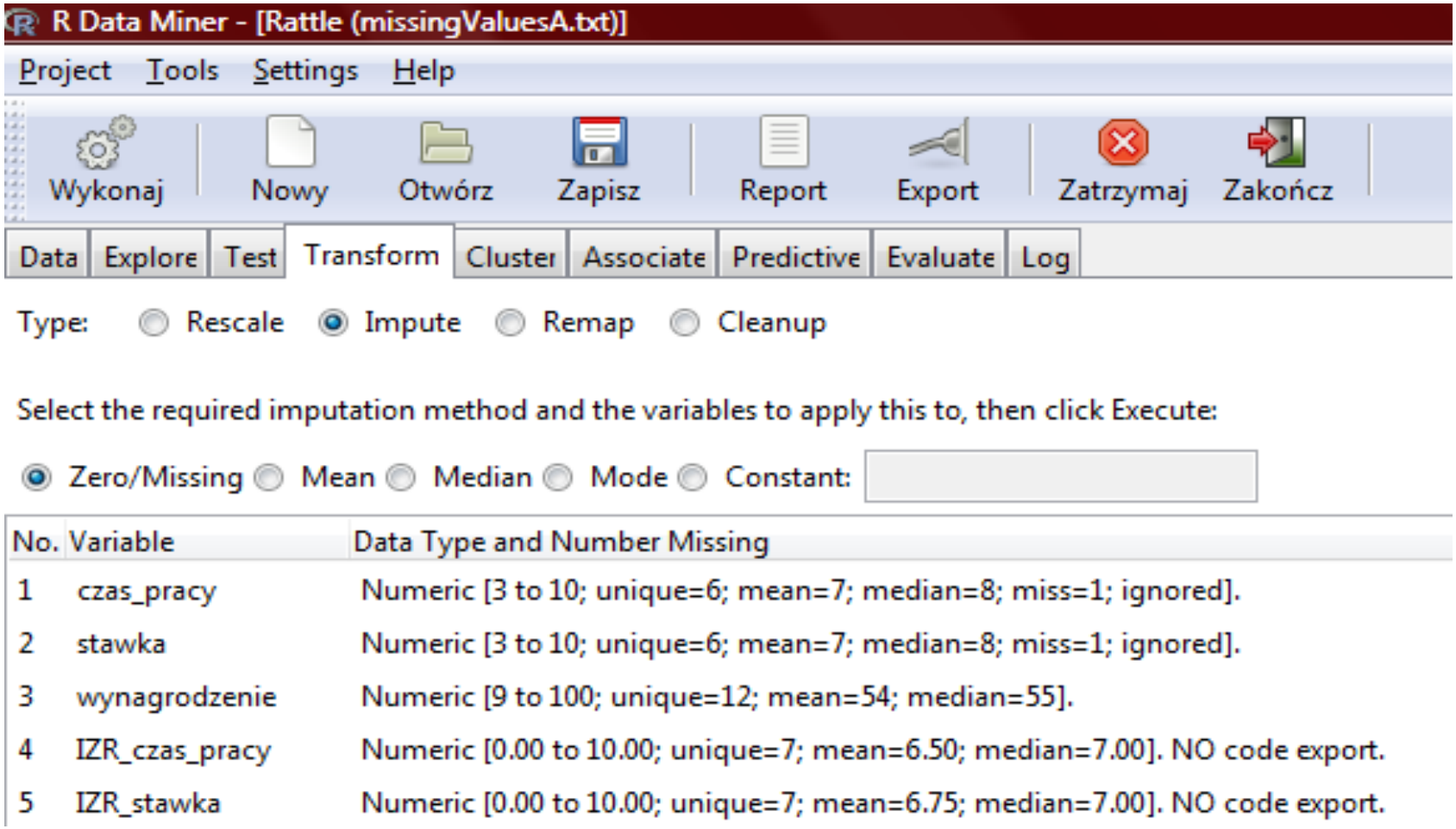

R Data Miner - [Rattle (dataset)]
Project Tools Settings Help
Wykonaj
Nowy
Otwórz
Zapisz
Report
Export
Zatrzymaj
Zakończ
Data Explore Test Transform Cluster Associate Predictive Evaluate Log
Type: Rescale Impute Remap Cleanup
Select the required imputation method and the variables to apply this to, then click Execute:
Zero/Missing Mean Median Mode Constant:
No. Variable Data Type and Number Missing
1 czas_pracy Numeric [3.00 to 10.00; unique=6; mean=7.09; median=8.00; miss=1].
2 stawka Numeric [3.00 to 10.00; unique=6; mean=7.36; median=8.00; miss=1].
3 wynagrodzenie Numeric [9.00 to 100.00; unique=12; mean=54.58; median=55.00].


R Data Miner - [Rattle (dataset)]
Project Tools Settings Help
Wykonaj
Nowy
Otwórz
Zapisz
Report
Export
Zatrzymaj
Zakończ
Data Explore Test Transform Cluster Associate Predictive Evaluate Log
Type: Rescale Impute Remap Cleanup
Select the required imputation method and the variables to apply this to, then click Execute:
Zero/Missing Mean Median Mode Constant:
No. Variable Data Type and Number Missing
1 czas_pracy Numeric [3.00 to 10.00; unique=6; mean=7.09; median=8.00; miss=1; ignored].
2 stawka Numeric [3.00 to 10.00; unique=6; mean=7.36; median=8.00; miss=1; ignored].
3 wynagrodzenie Numeric [9.00 to 100.00; unique=12; mean=54.58; median=55.00].
4 IMN_czas_pracy Numeric [3.00 to 10.00; unique=7; mean=7.09; median=7.55]. NO code export.
5 IMN_stawka Numeric [3.00 to 10.00; unique=7; mean=7.36; median=7.68]. NO code export.

Rattle: Data Viewer

	czas_pracy	stawka	wynagrodzenie	IMN_czas_pracy	IMN_stawka
1	10	10	100	10.000000	10.000000
2	10	NA	80	10.000000	7.363636
3	9	9	81	9.000000	9.000000
4	6	6	36	6.000000	6.000000
5	6	10	60	6.000000	10.000000
6	5	5	25	5.000000	5.000000
7	5	10	50	5.000000	10.000000
8	8	8	64	8.000000	8.000000
9	8	9	72	8.000000	9.000000
10	8	6	48	8.000000	6.000000
11	NA	5	30	7.090909	5.000000
12	3	3	9	3.000000	3.000000

Gdzie widzimy, że zarówno wiersz 2 jak i 11 mają teraz nowe wartości:
będące wartościami średnimi w zbiorze.


R Data Miner - [Rattle (missingValuesA.txt)]
Project Tools Settings Help
Wykonaj
Nowy
Otwórz
Zapisz
Report
Export
Zatrzymaj
Zakończ
Data Explore Test Transform Cluster Associate Predictive Evaluate Log
Type: Rescale Impute Remap Cleanup
Select the required imputation method and the variables to apply this to, then click Execute:
Zero/Missing Mean Median Mode Constant:
No. Variable Data Type and Number Missing
1 czas_pracy Numeric [3 to 10; unique=6; mean=7; median=8; miss=1; ignored].
2 stawka Numeric [3 to 10; unique=6; mean=7; median=8; miss=1; ignored].
3 wynagrodzenie Numeric [9 to 100; unique=12; mean=54; median=55].
4 IZR_czas_pracy Numeric [0.00 to 10.00; unique=7; mean=6.50; median=7.00]. NO code export.
5 IZR_stawka Numeric [0.00 to 10.00; unique=7; mean=6.75; median=7.00]. NO code export.

Rattle: Data Viewer

	czas_pracy	stawka	wynagrodzenie	IZR_czas_pracy	IZR_stawka
1	10	10	100	10	10
2	10	NA	80	10	0
3	9	9	81	9	9
4	6	6	36	6	6
5	6	10	60	6	10
6	5	5	25	5	5
7	5	10	50	5	10
8	8	8	64	8	8
9	8	9	72	8	9
10	8	6	48	8	6
11	NA	5	30	0	5
12	3	3	9	3	3

- Metoda zastępowania braków w danych w dużej mierze zależy od typu danych. Gdy brakuje danych w kolumnach z danymi numerycznymi często stosuje się uzupełnianie braków w danych wartością średnią czy medianą np. **Jednak jeśli brakuje danych w kolumnach z danymi typu nominalnego wówczas powinno się wypełniać braki wartością najczęściej występującą w zbiorze !**

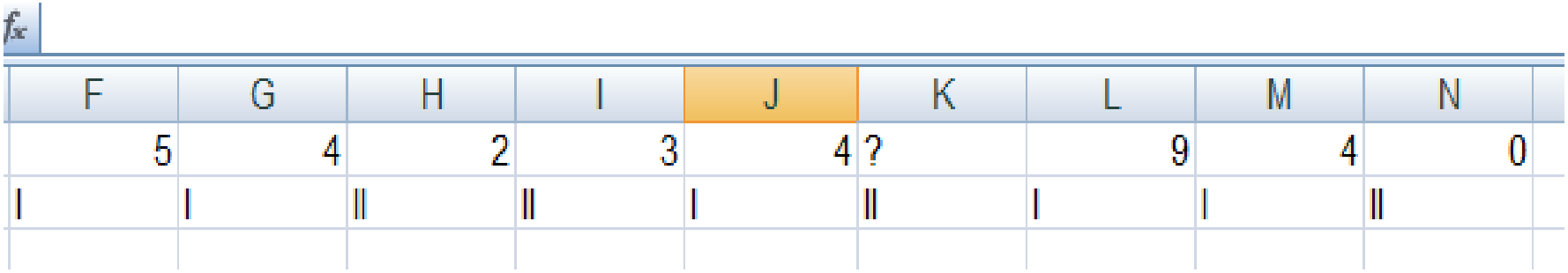
Zastosowanie metody „k-NN” do uzupełniania braków w danych

- Metoda ta polega na tym, by znaleźć K takich przykładów, które są najbardziej podobne do obiektu, dla którego mamy pewne wartości puste. Wówczas brakująca wartość jest wyznaczana jako średnia wartość tej danej (zmiennej, kolumny) wśród tych K wybranych wartości.
- Wówczas wartość brakująca jest wypełniana jako:

$$\hat{y}_{ih} = \frac{\sum_{j \in I_{Kih}} y_{jh}}{|I_{Kih}|},$$

- , gdzie I_{Kih} jest zbiorem przykładów wziętych pod uwagę jako najbardziej podobne obserwacje, y_{jh} jest wartością brakującą. Wadą tej metody jest fakt, że nie wiadomo jaka wartość liczby K jest najwłaściwsza – i dobiera się ją czysto doświadczalnie.

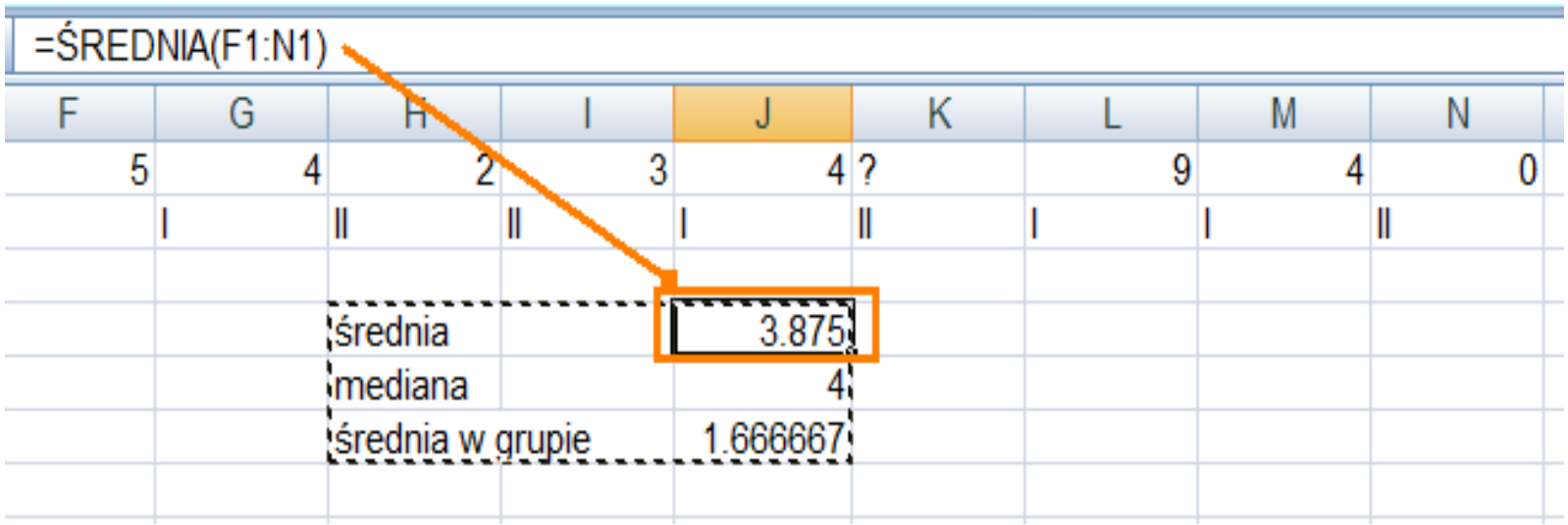
Przykład

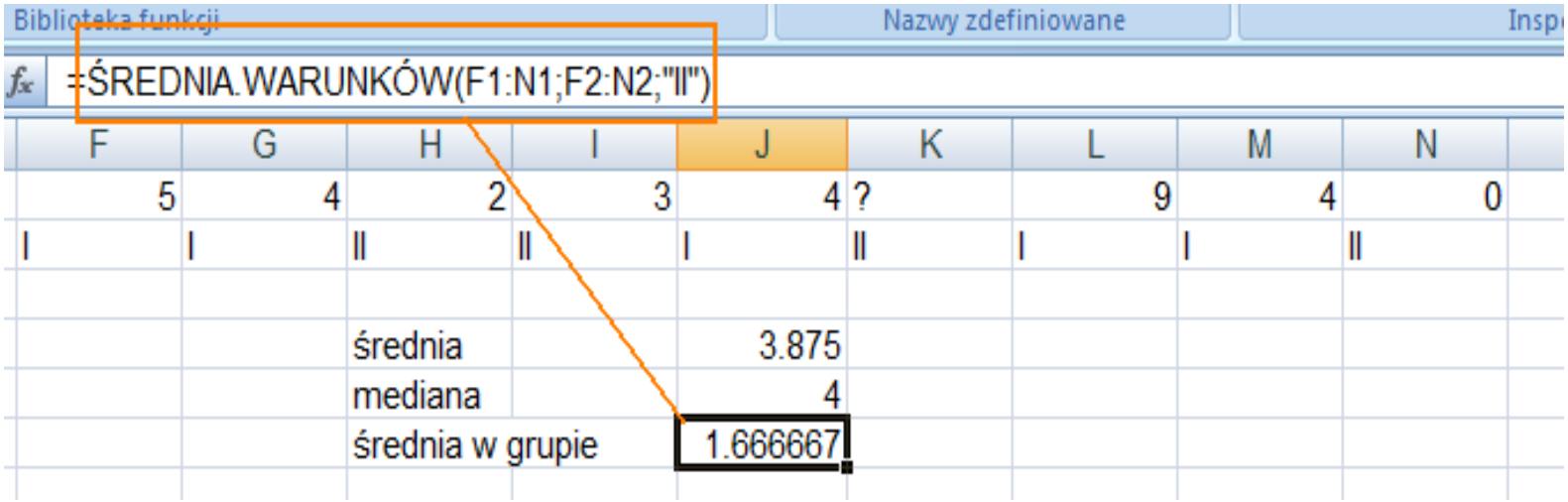


F	G	H	I	J	K	L	M	N
5	4	2	3	4	?	9	4	0
I	I	II	III	I	II	I	I	II

Widzimy, że w komórce K1 brakuje wartości. Excel rozpoznaje komórki z błędnymi wartościami – w tym przypadku będzie to zawartość tej komórki równa „?” i nie wlicza takich wartości przy podstawowych statystykach tupu średnia czy mediana.

średnia	3.875
mediana	4
średnia w grupie	1.666667


=ŚREDNIA(F1:N1)
F G H I J K L M N
5 4 2 3 4 ? 9 4 0
I I II II I II I I II
średnia 3.875
mediana 4
średnia w grupie 1.666667


Biblioteka funkcji
Nazwy zdefiniowane
=ŚREDNIA.WARUNKÓW(F1:N1;F2:N2;"II")
F G H I J K L M N
5 4 2 3 4 ? 9 4 0
I I II II I II I I II
średnia 3.875
mediana 4
średnia w grupie 1.666667

Przekształcanie danych

- Dane należy normalizować, by ujednolicić wpływ każdej zmiennej na wyniki. Jest kilka technik normalizacji:

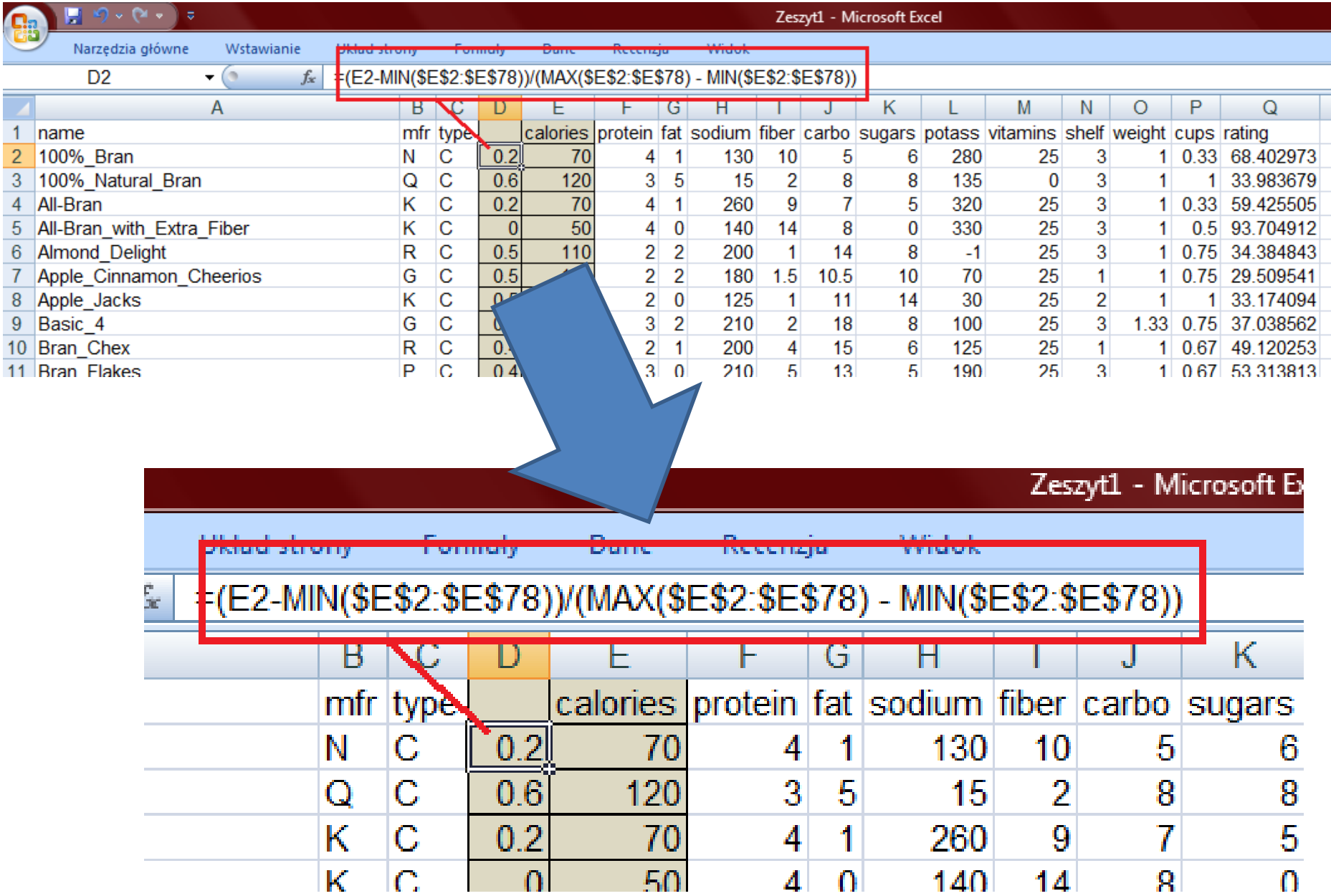
1. Min- Max
2. Standaryzacja Z-score

Normalizacja Min-Max

Metoda przeprowadza liniową transformację pierwotnych danych najczęściej do przedziału $[0, 1]$ według wzoru:

$$V' = \frac{V - min}{max - min} * (new\_max - new\_min) + new\_min$$

gdzie odpowiednio V oraz V' to wartości przed i po normalizacji, $[min, max]$ jest oryginalnym przedziałem, w którym mieszczą się dane wejściowe, natomiast $[new\_min, new\_max]$ jest nowym (docelowym) przedziałem danych. Atutem tej metody jest fakt, że zachowuje ona prawidłowe relacje pomiędzy wartościami.

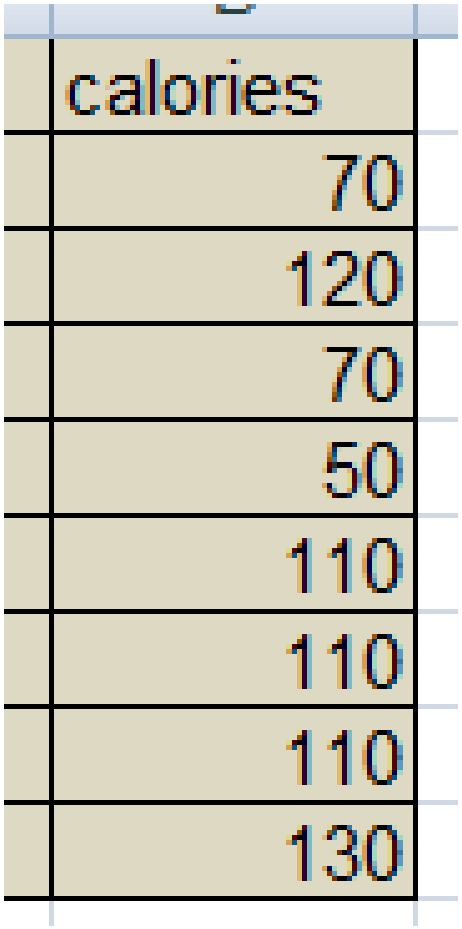

Zeszyt1 - Microsoft Excel
Narzędzia główne
Wstawianie
D2
=(E2-MIN($E$2:$E$78))/(MAX($E$2:$E$78) - MIN($E$2:$E$78))
name mfr type calories protein fat sodium fiber carbo sugars potass vitamins shelf weight cups rating
100%_Bran N C 0.2 70 4 1 130 10 5 6 280 25 3 1 0.33 68.402973
100%_Natural_Bran Q C 0.6 120 3 5 15 2 8 8 135 0 3 1 1 33.983679
All-Bran K C 0.2 70 4 1 260 9 7 5 320 25 3 1 0.33 59.425505
All-Bran_with_Extra_Fiber K C 0 50 4 0 140 14 8 0 330 25 3 1 0.5 93.704912
Almond_Delight R C 0.5 110 2 2 200 1 14 8 -1 25 3 1 0.75 34.384843
Apple_Cinnamon_Cheerios G C 0.5 2 2 180 1.5 10.5 10 70 25 1 1 0.75 29.509541
Apple_Jacks K C 2 0 125 1 11 14 30 25 2 1 1 33.174094
Basic_4 G C 3 2 210 2 18 8 100 25 3 1.33 0.75 37.038562
Bran_Chex R C 2 1 200 4 15 6 125 25 1 1 0.67 49.120253
Bran_Flakes P C 3 0 210 5 13 5 190 25 3 1 0.67 53.313813
Zeszyt1 - Microsoft Ex
=(E2-MIN($E$2:$E$78))/(MAX($E$2:$E$78) - MIN($E$2:$E$78))
mfr type calories protein fat sodium fiber carbo sugars
N C 0.2 70 4 1 130 10 5 6
Q C 0.6 120 3 5 15 2 8 8
K C 0.2 70 4 1 260 9 7 5
K C 0 50 4 0 140 14 8 0

Normalizacja „Min-Max” – wersja bardziej uniwersalna

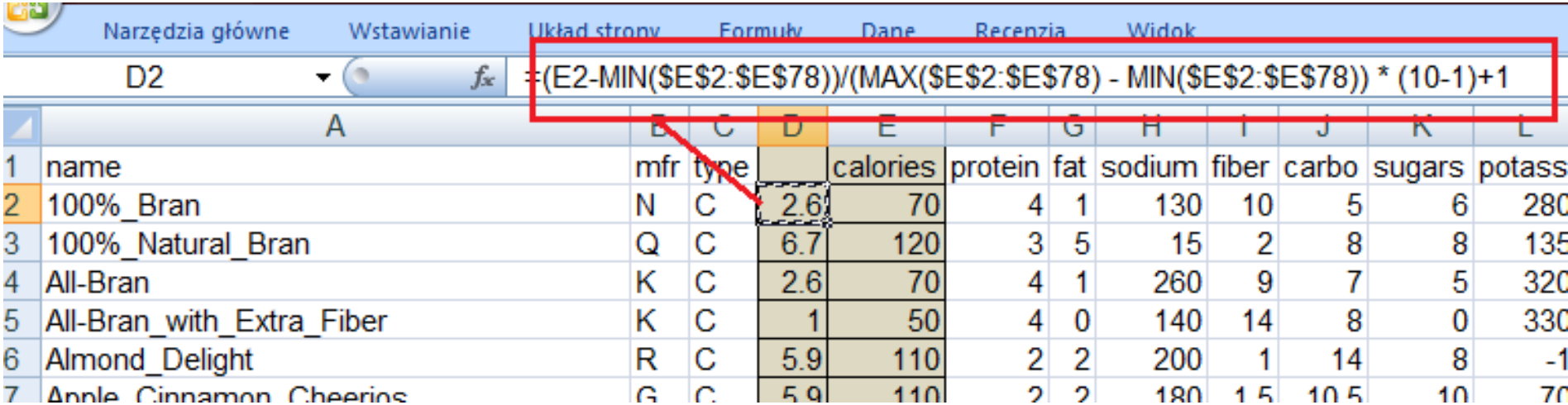
$$V' = \frac{(V - \min)}{\max - \min} * (new\_\max - new\_\min) + new\_\min$$

New_min to nowa wartość minimalna, którą chcemy uzyskać
New_max – nowa wartość maksymalna.
Min – to dotychczasowa wartość minimalna
Max – dotychczasowa wartość maksymalna

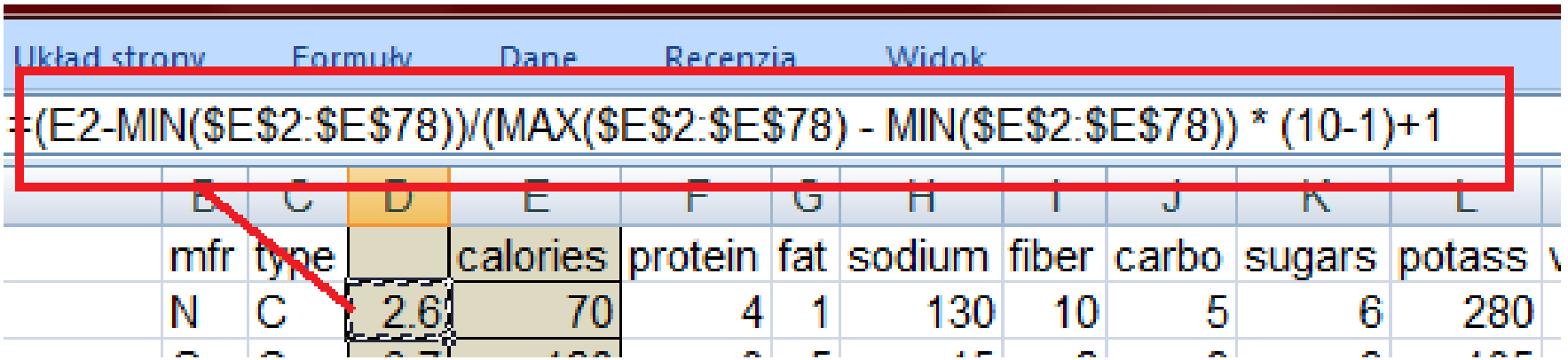
- Min = 50
- Max = 130
- New_min = 0
- New_max = 1



	calories
	70
	120
	70
	50
	110
	110
	110
	130

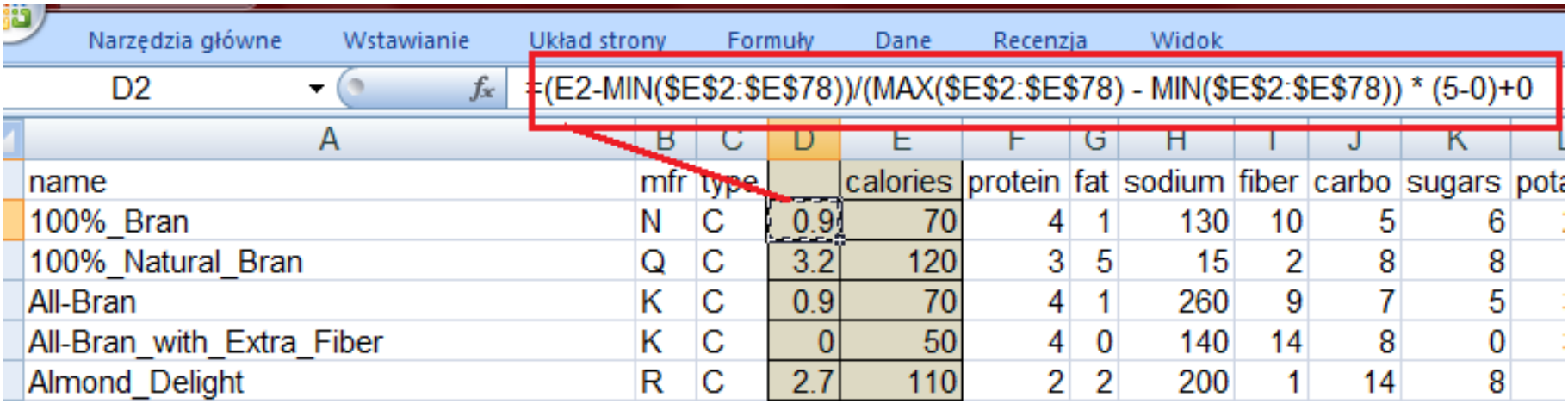


	A	B	C	D	E	F	G	H	I	J	K	L
1	name	mfr	type		calories	protein	fat	sodium	fiber	carbo	sugars	potass
2	100%_Bran	N	C	2.6	70	4	1	130	10	5	6	280
3	100%_Natural_Bran	Q	C	6.7	120	3	5	15	2	8	8	135
4	All-Bran	K	C	2.6	70	4	1	260	9	7	5	320
5	All-Bran_with_Extra_Fiber	K	C	1	50	4	0	140	14	8	0	330
6	Almond_Delight	R	C	5.9	110	2	2	200	1	14	8	-1
7	Apple_Cinnamon_Cheerios	G	C	5.9	110	2	2	180	1.5	10.5	10	70



B	C	D	E	F	G	H	I	J	K	L
mfr	type		calories	protein	fat	sodium	fiber	carbo	sugars	potass
N	C	2.6	70	4	1	130	10	5	6	280

0..5



	A	B	C	D	E	F	G	H	I	J	K	L
	name	mfr	type		calories	protein	fat	sodium	fiber	carbo	sugars	pota
	100%_Bran	N	C	0.9	70	4	1	130	10	5	6	
	100%_Natural_Bran	Q	C	3.2	120	3	5	15	2	8	8	
	All-Bran	K	C	0.9	70	4	1	260	9	7	5	
	All-Bran_with_Extra_Fiber	K	C	0	50	4	0	140	14	8	0	
	Almond_Delight	R	C	2.7	110	2	2	200	1	14	8	

Normalizacja Z-score

Nazwa metody wywodzi się z jej własności, zgodnie z którą po normalizacji wartość średnia powinna wynosić 0. Normalizacja danych polega na transformacji danych wg wzoru:

$$V' = \frac{V - \overline{x}}{\sigma},$$

gdzie odpowiednio V oraz V' to wartości przed i po normalizacji, $\overline{x}$ to wartość będąca średnią dla całego normalizowanego zbioru, zaś σ to odchylenie standardowe

Wartości większe od średniej po standaryzacji będą na pewno dodatnie !

Korelacja - definicja

Korelacja (współzależność cech) określa wzajemne powiązania pomiędzy wybranymi zmiennymi.

- Charakteryzując korelację dwóch cech podajemy dwa czynniki: kierunek oraz siłę.
- Wyrazem liczbowym korelacji jest współczynnik korelacji (r lub R), zawierający się w przedziale [-1; 1].

Korelacja

- Korelacja (współzależność cech) określa wzajemne powiązania pomiędzy wybranymi zmiennymi.
- Charakteryzując korelację dwóch cech podajemy dwa czynniki: kierunek oraz siłę.

Współczynnik korelacji Pearsona

Współczynnik ten wykorzystywany jest do badania związków prostoliniowych badanych

zmiennych, w których zwiększenie wartości jednej z cech powoduje proporcjonalne zmiany średnich wartości drugiej cechy (wzrost lub spadek).

Współczynnik ten obliczamy na podstawie wzoru:

$$r_{xy} = \frac{\operatorname{cov}(x, y)}{Sd_x \cdot Sd_y} \qquad \operatorname{cov}(x, y) = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{n}$$

Korelacja Pearsona

Korelacja jest miarą pozwalającą badać stopień zależności między analizowanymi danymi. Jedną z podstawowych miar korelacji jest współczynnik korelacji Pearsona wyrażany wzorem:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

gdzie xi i yi to poszczególne wartości rzeczywiste zmiennych x i y, $\bar{x}$ i $\bar{y}$ określają wartości średnie tych zmiennych zaś sx i sy to odpowiednio odchylenia standardowe tych zmiennych w zbiorze n elementów. Przyjmuje on wartości z przedziału [−1, 1].

Dodatnia wartość tego współczynnika oznacza, że wzrost wartości jednej zmiennej generalnie pociąga za sobą wzrost wartości drugiej zmiennej (wartość ujemna oznacza odpowiednio spadek wartości dla drugiej zmiennej). Specyficzny przypadek gdy r =0 oznacza brak związku między zmiennymi x i y.

Przykład – brak korelacji

Badaniu poddano długość kończyny dolnej (x_i) oraz moc (y_i) u siedmiu uczniów IV klasy szkoły podstawowej. Na podstawie poniższych danych oszacować współzależność obu analizowanych cech.

x_i	83,1	88,2	87,3	90,4	80,6	87,1	85,3
y_i	41	45	42	52	52	46	47

W pierwszej kolejności należy wyliczyć:

$\bar{x}_{xi}$; $\bar{y}_{yi}$; Sd_x ; Sd_y oraz wartość kowariancji

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
83,1	41	-2,9	-5,4	15,66	8,41	29,16
88,2	45	2,2	-1,4	-3,08	4,84	1,96
87,3	42	1,3	-4,4	-5,72	1,69	19,36
90,4	52	4,4	5,6	24,64	19,36	31,36
80,6	52	-5,4	5,6	-30,24	29,16	31,36
87,1	46	1,1	-0,4	-0,44	1,21	0,16
85,3	47	-0,7	0,6	-0,42	0,49	0,36
Σ=602,0	Σ=325,0			Σ=-1,24	Σ=65,16	Σ=113,72

$$\bar{x} = \frac{602}{7} = 86;\ \bar{y} = \frac{325}{7} = 46,4$$

$$Sd_x = \sqrt{\frac{\sum (x_i - \bar{x}_x)^2}{n}} = \sqrt{\frac{65,16}{7}} = 3,05$$

$$Sd_y = \sqrt{\frac{\sum (y_i - \bar{x}_y)^2}{n}} = \sqrt{\frac{113,72}{7}} = 4,03$$

$$\mathrm{cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{-1,24}{7} = -0,18$$

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{Sd_x \cdot Sd_y} = \frac{-0,18}{3,05 \cdot 4,03} = -0,015$$

Interpretacja: wyliczony współczynnik korelacji wskazuje na **brak związku** długości kończyny z mocą objętych badaniem uczniów.

Przykład – wysoka korelacja

- Sportowców poddano badaniom szybkości reakcji na bodziec wzrokowy (yi) oraz badaniom wzroku (xi).
- Oszacować współzależność obu analizowanych cech

x_i	3,5	3,4	2,1	5,4	1,1	5,1	6,9	4,0	4,5	2,5
y_i	1,6	2,9	1,5	3,5	0,6	2,5	7,1	3,5	2,1	2,6

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
3,5	1,6	-0,35	-1,19	0,42	0,12	1,42
3,4	2,9	-0,45	0,11	-0,05	0,20	0,01
2,1	1,5	-1,75	-1,29	2,26	3,06	1,66
5,4	3,5	1,55	0,71	1,10	2,40	0,50
1,1	0,6	-2,75	-2,19	6,02	7,56	4,80
5,1	2,5	1,25	-0,29	-0,36	1,56	0,08
6,9	7,1	3,05	4,31	13,15	9,30	18,58
4	3,5	0,15	0,71	0,11	0,02	0,50
4,5	2,1	0,65	-0,69	-0,45	0,42	0,48
2,5	2,6	-1,35	-0,19	0,26	1,82	0,04
38,5	**27,9**			**22,445**	**26,485**	**28,069**

Występuje bardzo wysoka dodatnia korelacja pomiędzy analizowanymi cechami. Wzrostowi jakości wzroku towarzyszy wzrost szybkości reakcji.

$r=0,8232 \in < 0,7;0,9)$

Współczynnik R Spearmana

Współczynnik korelacji rang Spearmana wykorzystywany jest do opisu siły korelacji dwóch cech, w przypadku gdy:

- **cechy mają charakter jakościowy,** pozwalający na uporządkowanie ze względu na siłę tej cechy,
- **cechy mają charakter ilościowy, ale ich liczebność jest niewielka.**

- Ranga jest to liczba odpowiadająca miejscu w uporządkowaniu każdej z cech. Jeśli w badanej zbiorowości jest więcej jednostek z identycznym natężeniem badanej cechy, to jednostkom tym przypisuje się identyczne rangi, licząc średnią arytmetyczną z rang przynależnych tym samym jednostkom.
- Współczynnik korelacji rang przyjmuje wartości z przedziału [-1; 1].
- Interpretacja jest podobna do współczynnika korelacji liniowej Pearsona.
- d_i^2 - *różnica pomiędzy rangami odpowiadających sobie wartości cech* x_i *i* y_i

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Na podstawie opinii o zdrowiu 10 pacjentów wydanych przez dwóch lekarzy chcemy ustalić współzależność między tymi opiniami, które zostały wyrażone w punktach.

Nr Pacjenta	1	2	3	4	5	6	7	8	9	10
Punkty od I lekarza	42	27	36	33	24	47	39	52	43	37
Punkty od II lekarza	39	24	35	29	26	47	44	51	39	32

Nr Pacjenta	1	2	3	4	5	6	7	8	9	10
Rangi od I lekarza	7	2	4	3	1	9	6	10	8	5
Rangi od II lekarza	6,5	1	5	3	2	9	8	10	6,5	4

x_i	y_i	Ranga x	Ranga y	d_i	d_i^2
42	39	7	6,5	0,5	0,25
27	24	2	1	1	1
36	35	4	5	-1	1
33	29	3	3	0	0
24	26	1	2	-1	1
47	47	9	9	0	0
39	44	6	8	-2	4
52	51	10	10	0	0
43	39	8	6,5	1,5	2,25
37	32	5	4	1	1
		Σ			10,5

$$r_s = 1 - \frac{6 \cdot 10,5}{10(100-1)} = 1 - \frac{63}{990} = 0,936$$

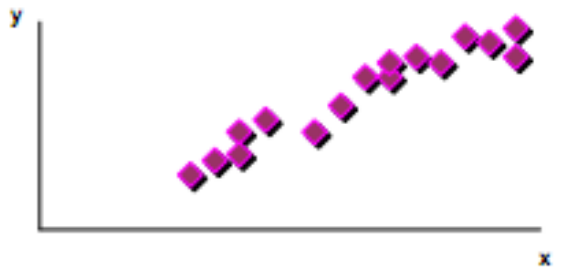
Korelacja: zalety i wady

- korelacja dodatnia (wartość współczynnika korelacji od 0 do 1) –informuje, że wzrostowi wartości jednej cechy towarzyszy wzrost średnich wartości drugiej cechy,
- korelacja ujemna (wartość współczynnika korelacji od -1 do 0) -informuje, że wzrostowi wartości jednej cechy towarzyszy spadek średnich wartości drugiej cechy.

Wartość korelacji dodatnie

- poniżej 0,2 - korelacja słaba (praktycznie brak związku)
- 0,2 –0,4 - korelacja niska (zależność wyraźna)
- 0,4 – 0,6 - korelacja umiarkowana 0,4 – 0,6 - korelacja umiarkowana (zależność istotna)
- 0,6 –0,8 - korelacja wysoka (zależność znaczna)
- 0,8 –0,9 - korelacja bardzo wysoka (zależność bardzo duża)
- 0,9 –1,0 - zależność praktycznie pełna

Korelacyjne wykresy rozrzutu



zależność liniowa dodatnia (r>0)

zależność liniowa ujemna (r<0)

brak zależności (r=0)

Wiadomości testowe ?

- Co to jest korelacja ?
- Jak wykrywać wartości oddalone w zbiorze danych ?
- Jak zastępować braki w danych ?
- Jak normalizować dane do jakiegoś przedziału ?
- Czy typ danych ma wpływ na wybór graficznej reprezentacji ?
- W czym może pomóc eksploracja danych ?

Dziękuję za uwagę !