GDAŃSK UNIVERSITY OF TECHNOLOGY

Eksploracja danych

EKSPLORACYJNA ANALIZA DANYCH (EDA)

Wojciech Waloszek

wowal@eti.pg.gda.pl

Teresa Zawadzka

tegra@eti.pg.gda.pl

Katedra Inżynierii Oprogramowania
Wydział Elektroniki, Telekomunikacji i Informatyki
Politechnika Gdańska

GDAŃSK UNIVERSITY OF TECHNOLOGY

EDA — *exploratory data analysis*

EDA to zestaw interaktywnych i wizualnych technik analizy zbioru danych bez bardziej precyzyjnego wskazania celu eksploracji.

EDA to także bardzo ważny element etapu wstępnej analizy danych.

EDA pozwala na:

- zgłębienie zbioru danych,
- sprawdzenie zależności między atrybutami,
- identyfikację nietypowych przykładów bądź podzbiorów przykładów,
- opracowanie wstępnych wniosków dotyczących prawidłowości w zbiorze przykładów.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Poznaj swoje dane!

Dane nie zawsze są tak eleganckie…

S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	**800**	**32**	**wyższe**	**tak**	**tak**
S	**1200**	**35**	**średnie**	**tak**	**tak**
S	**700**	**26**	**podstawowe**	**nie**	**nie**
M	**600**	**45**	**wyższe**	**nie**	**tak**
M	**650**	**38**	**średnie**	**tak**	**tak**
S	**900**	**28**	**wyższe**	**nie**	**nie**
S	**1100**	**65**	**średnie**	**tak**	**nie**
M	**500**	**22**	**średnie**	**nie**	**nie**
S	**800**	**43**	**podstawowe**	**tak**	**nie**

GDAŃSK UNIVERSITY OF TECHNOLOGY

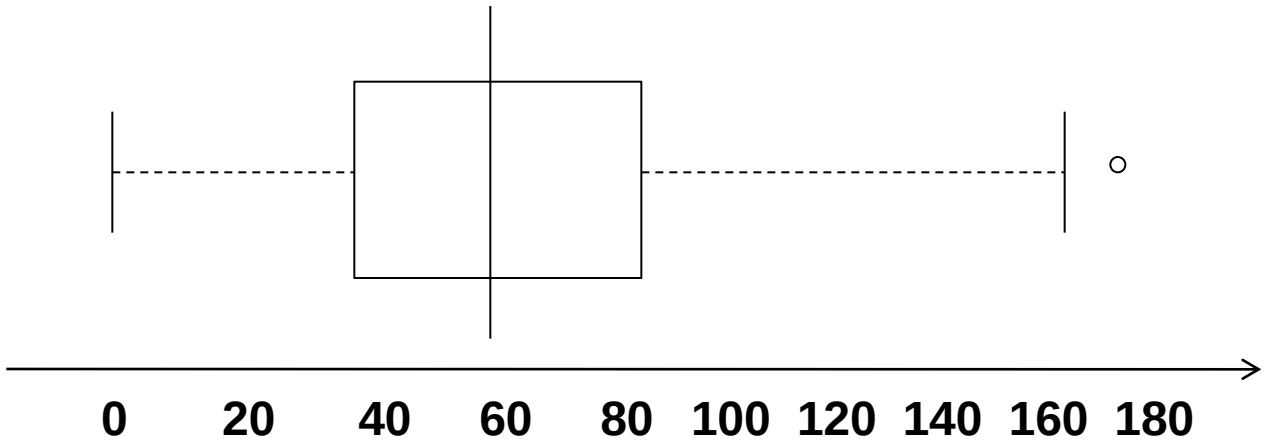
Poznaj swoje dane!

Kod pocztowy	Dochód roczny (tys. zł)	Wiek	Stan cywilny	Płeć
80249	**70**	**32**	**S**	**M**
00245	**50000**	**35**	**M**	
J2S7K7	**40**	**26**	**K**	**M**
90120		**0**	**Ż**	**K**
8230	**999**	**38**		**K**

GDAŃSK UNIVERSITY OF TECHNOLOGY

Wykres pudełkowy (*boxplot*)

- **wykres skrzynkowy**
- **wykres ramka-wąsy**



GDAŃSK UNIVERSITY OF TECHNOLOGY

Wykres pudełkowy (2)

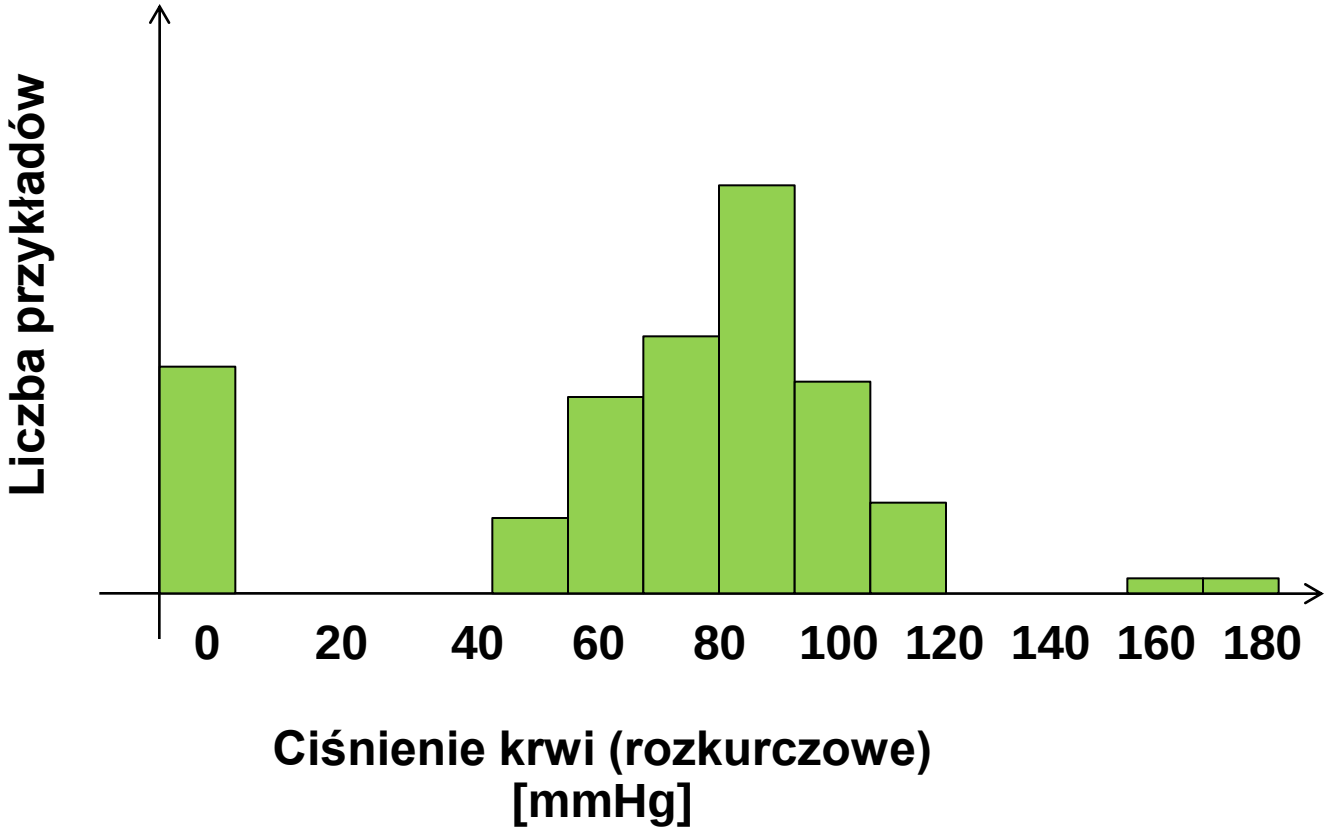
Wykres pudełkowy służy do przedstawienia podstawowych parametrów rozkładu wartości atrybutów numerycznych. Występuje w różnych odmianach i może być narysowany pionowo lub poziomo.

Składa się z *pudełka* i *wąsów* (*wibrys*):

- lewa (dolna) krawędź pudełka odpowiada pierwszemu a prawa (dolna) krawędź trzeciemu kwartylowi (Q1, Q3); pudełko przecina odcinek – mediana
- wąsy pokazują rozrzut danych „poza pudełkiem”, najczęściej jest to najbardziej skrajna wartość mieszcząca się jeszcze w odległości 1,5 * (Q3 – Q1) od odpowiednio Q1 i Q3
- punkty mieszczące się „poza wąsami” oznaczane są osobno jako *oddalone*

GDAŃSK UNIVERSITY OF TECHNOLOGY

Histogram



GDAŃSK UNIVERSITY OF TECHNOLOGY

Histogram (2)

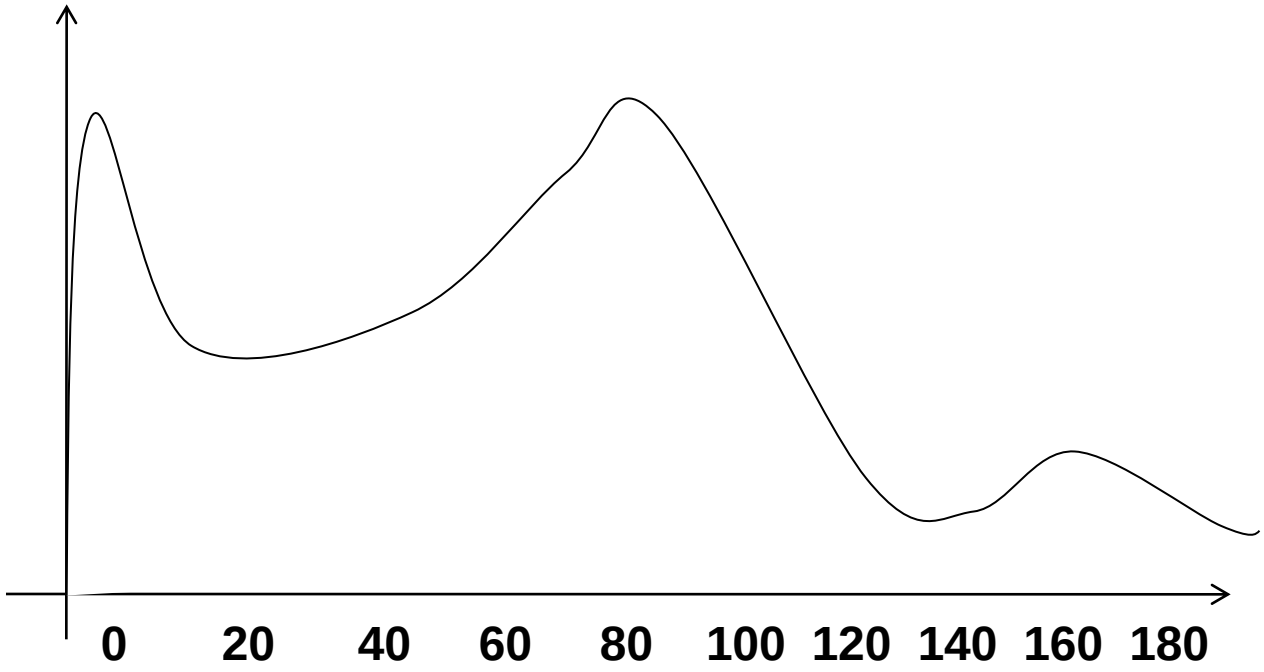
Histogram jest diagramem pokazującym liczność przykładów o wartości danego atrybutu zawartej w pewnej klasie.

Histogram pozwala na:

- przybliżenie kształtu rozkładu,
- identyfikację punktów oddalonych,
- identyfikację wartości specjalnych.

„Poszarpany” kształt histogramu utrudnia czasem interpretację rozkładu.

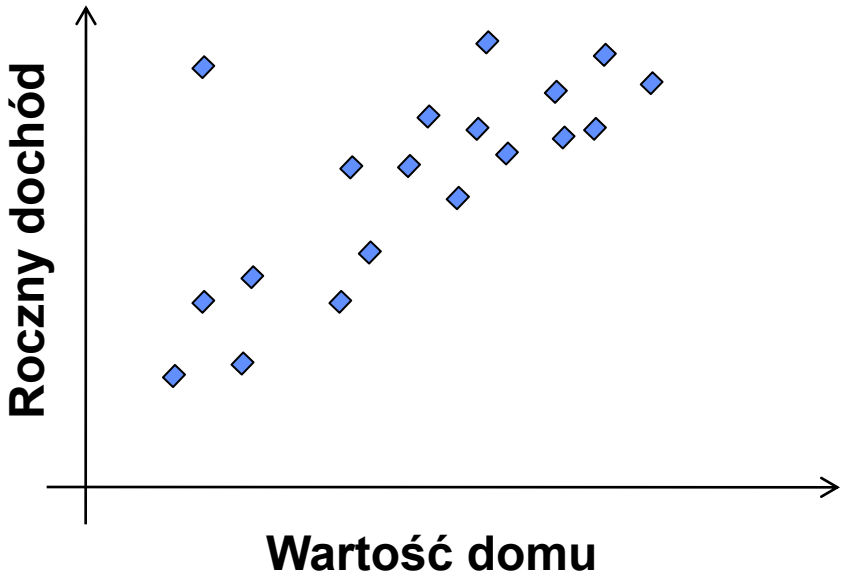
Przybliżanie rozkładu za pomocą wygładzania

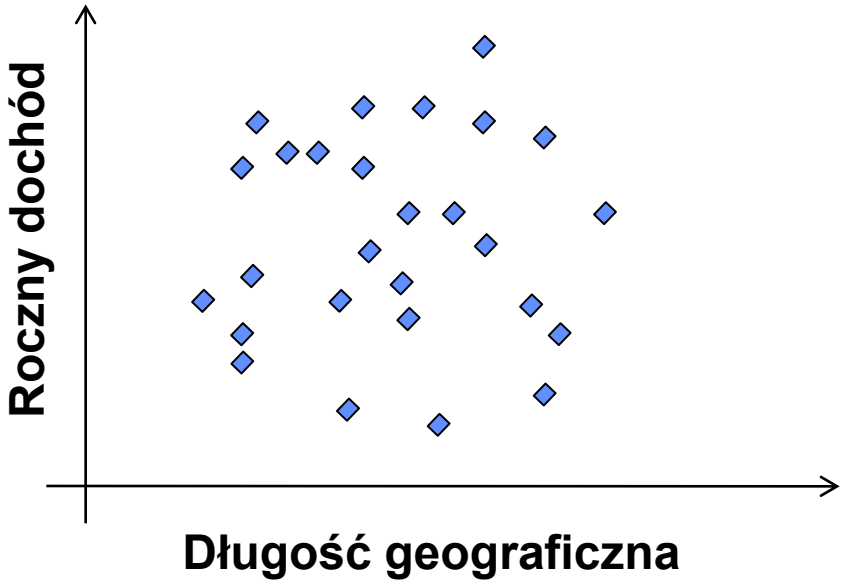


$$f(x) = \frac{1}{n \cdot h} \sum_{i=1}^{n} K \frac{x - x_i}{h}$$

GDAŃSK UNIVERSITY OF TECHNOLOGY

Wykresy punktowe





GDAŃSK UNIVERSITY OF TECHNOLOGY

Wykresy punktowe (2)

Wykresy punktowe pozwalają na wizualizację zależności pomiędzy parą zmiennych.

Wykresy punktowe pozwalają m.in. na:

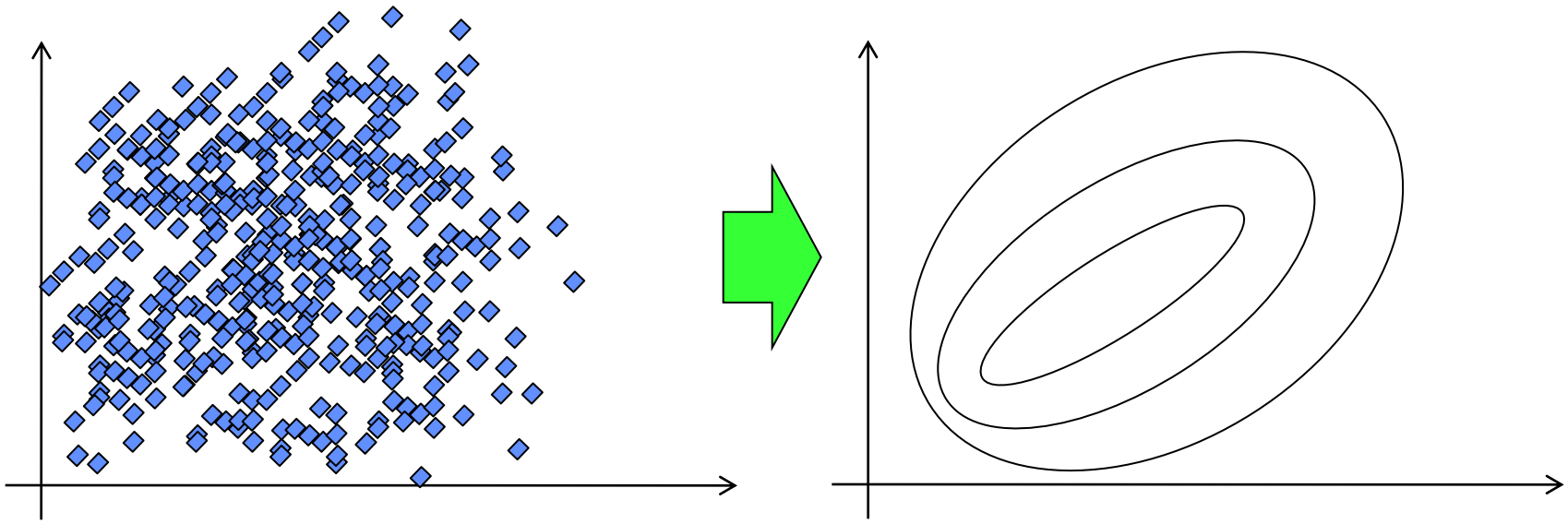
- odkrywanie korelacji między parą zmiennych,
- identyfikację punktów oddalonych od ogólnego trendu.

Macierze wykresów punktowych pozwalają na badanie większych podzbiorów atrybutów.

Wykresy punktowe podatne są na zjawisko tzw. overprintingu.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Overprinting i wykresy konturowe



GDAŃSK UNIVERSITY OF TECHNOLOGY

EDA – podsumowanie

EDA jest bardzo ważnym etapem procesu eksploracji danych, czasem nawet występującym samodzielnie. EDA jest także bardzo cennym elementem procesu wstępnej analizy danych pozwalającym na:

- lepsze poznanie zbioru przykładów, również poprzez wskazanie nietypowych wartości atrybutów, punktów tzw. oddalonych i anomalnych itp.,
- wyciągnięcie wstępnych, często bardzo interesujących z punktu widzenia klienta, wniosków na temat zależności pomiędzy wartościami poszczególnych zmiennych,
- opracowanie zbioru zaleceń co do sposobów przekształcenia i modelowania danych na dalszych etapach procesu eksploracji.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Dziękujemy za uwagę

Zapraszamy na wykład:

KLASYFIKACJA I REGRESJA cz. 1