



Techniki Wizualizacji Danych

Politechnika Warszawska

Anna Kozak

Zanim wizualizacja to chwilę
o eksploracji danych

Dane

Mogą być generowane przez:

- ?

Dane

Mogą być generowane przez:

- banki,
- ubezpieczenia,
- portale społecznościowe,
- firmy telekomunikacyjne,
- szpitale,
- dane eksperymentalne,
- tekst,
- mapy,
- sklepy internetowe,
- ...

Eksploracja danych – czym jest?

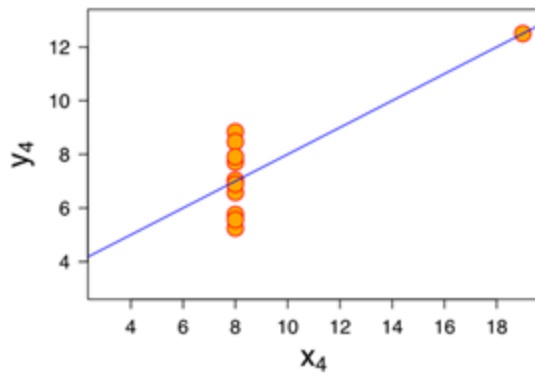
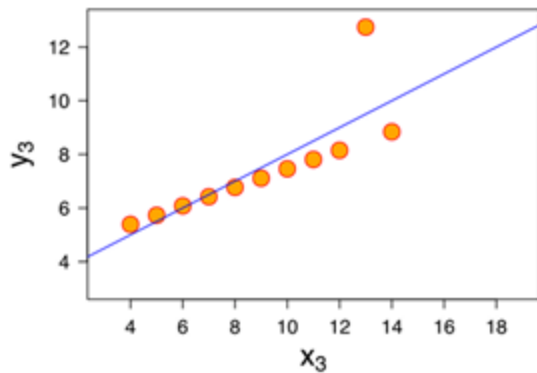
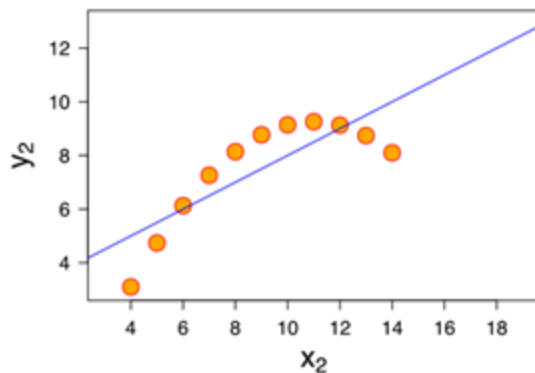
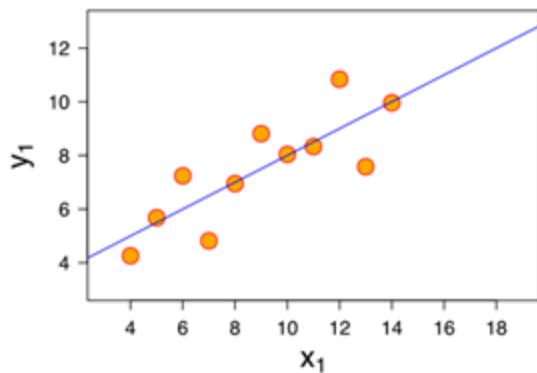
“proces odkrywania nietrywialnych, dotychczas nieznanых, potencjalnie użytecznych reguł, zależności, trendów”

Cel: analiza danych w celu lepszego ich zrozumienia

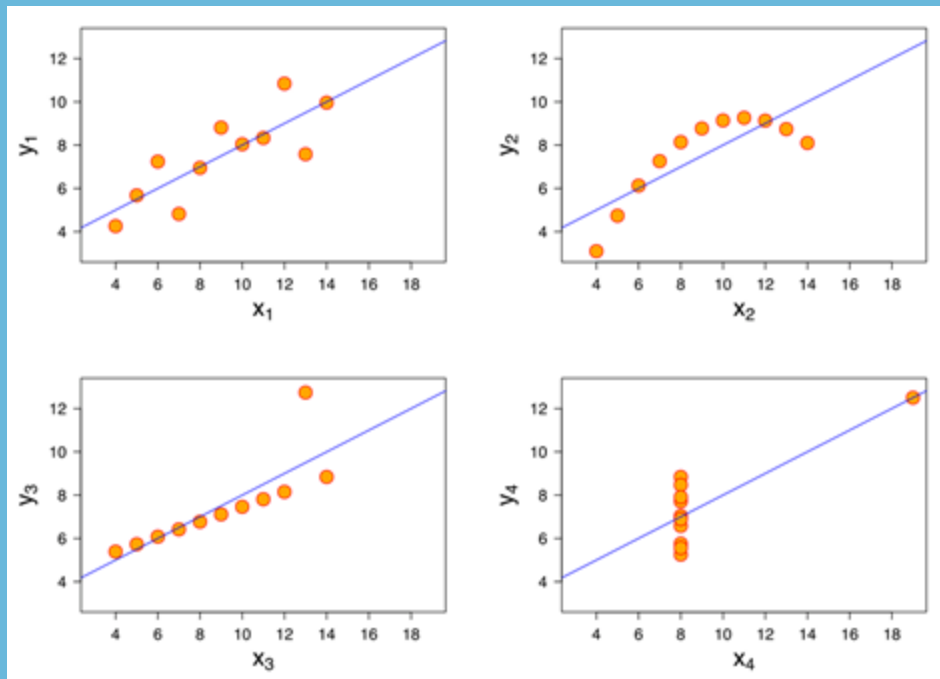
Eksploracja danych - czym jest?

Na eksplorację danych składa się wiele dyscyplin, między innymi:

- bazy danych
- statystyka
- uczenie maszynowe
- techniki wizualizacji danych
- wyszukiwanie informacji

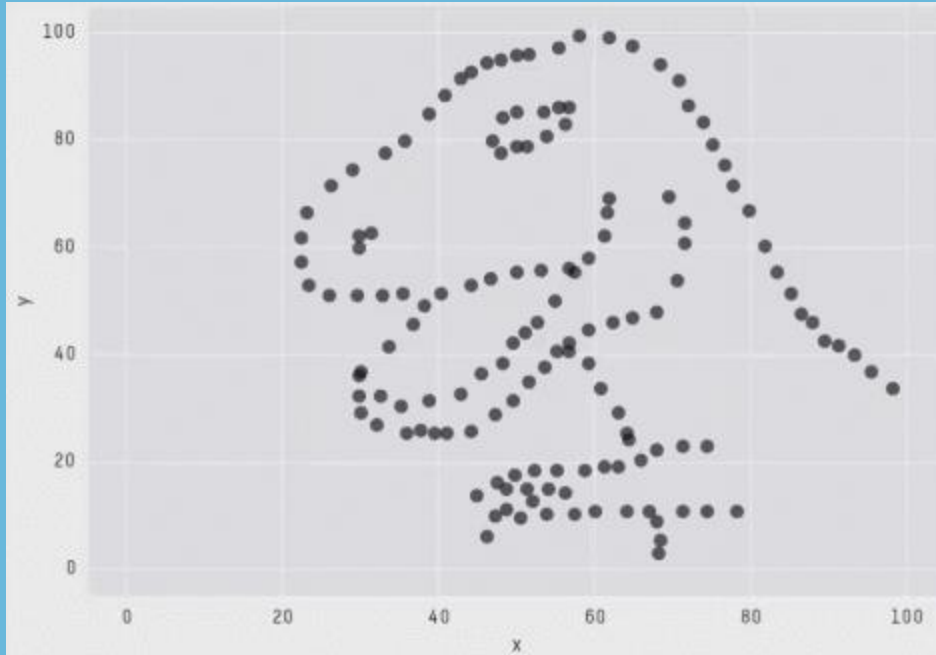


Kwartet Anscombe' a



Cecha	Wartość
Średnia arytmetyczna zmiennej x	9
Wariancja zmiennej x	11
Średnia arytmetyczna zmiennej y	7.50 (identyczna do dwóch cyfr po przecinku)
Wariancja zmiennej y	4.122 lub 4.127 (identyczna do trzech cyfr po przecinku)
Współczynnik korelacji pomiędzy zmiennymi	0.816 (identyczny do trzech cyfr po przecinku)

The Datasaurus Dozen



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Jak rozpoznać rodzaj zmiennej?

“dane liczbowe to nie tylko liczby”

Typy danych

Zmienne jakościowe (nazywane również wyliczeniowymi, czynnikowymi lub kategorycznymi), to zmienne przyjmujące określoną liczbę wartości (najczęściej nie liczbowych). Zmienne te można dalej podzielić na:

- *binarne* (nazywane również dwumianowymi, dychotomicznymi) np. płeć (poziomy: kobieta/mężczyzna),
- *nominalne* (nazywane również zmiennymi jakościowymi nieuporządkowanymi) np. marka samochodu,
- *uporządkowane*, np. wykształcenie (poziomy: podstawowe/średnie/wyższe), ocena z przedmiotu.

Typy danych

Zmienne ilościowe, z których można dodatkowo wyróżnić:

- *zliczenia* (liczba wystąpień pewnego zjawiska, opisywana liczbą całkowitą), np. liczba lat nauki, liczba wypadków,
- *ilorazowe*, czyli zmienne mierzone w skali, w której można dzielić wartości (ilorazy mają sens). Np. długość w metrach (coś jest 2 razy dłuższe, 10 razy krótsze itp.),
- *przedziałowe* (nazywane też interwałowymi), mierzone w skali, w której można odejmować wartości (wyznaczać długość przedziału).

Struktura zbioru danych

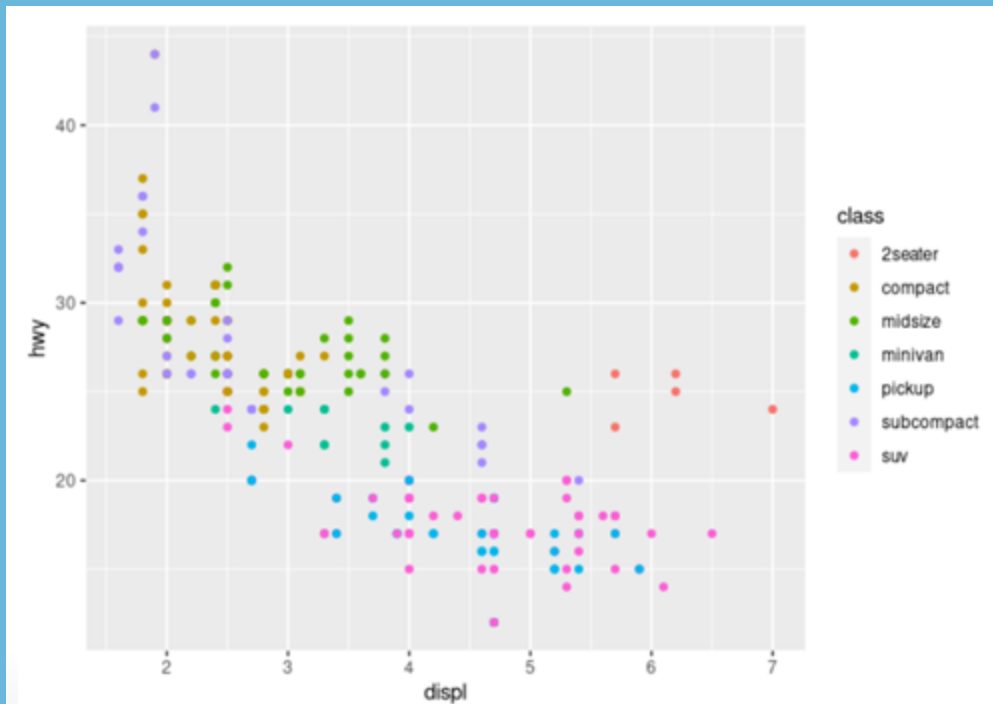
ID	PŁEĆ	ZAWÓD	WZROST	DATA URODZENIA
ID_23	K	INFORMATYK	158	1978-03-12
ID_45	K	PRAWNIK	178	1989-05-29
ID_46	M	MATEMATYK	183	1991-01-19
ID_89	M	INFORMATYK	167	1982-02-20
ID_101	K	LEKARZ	163	1973-02-23

Narzędzia do wizualizacji danych

- programistyczne (R, Python, JavaScript)
- programy graficzne (Inkscape, Gimp, Canva)
- programy dedykowane do wizualizacji danych (Tableau, Power BI)

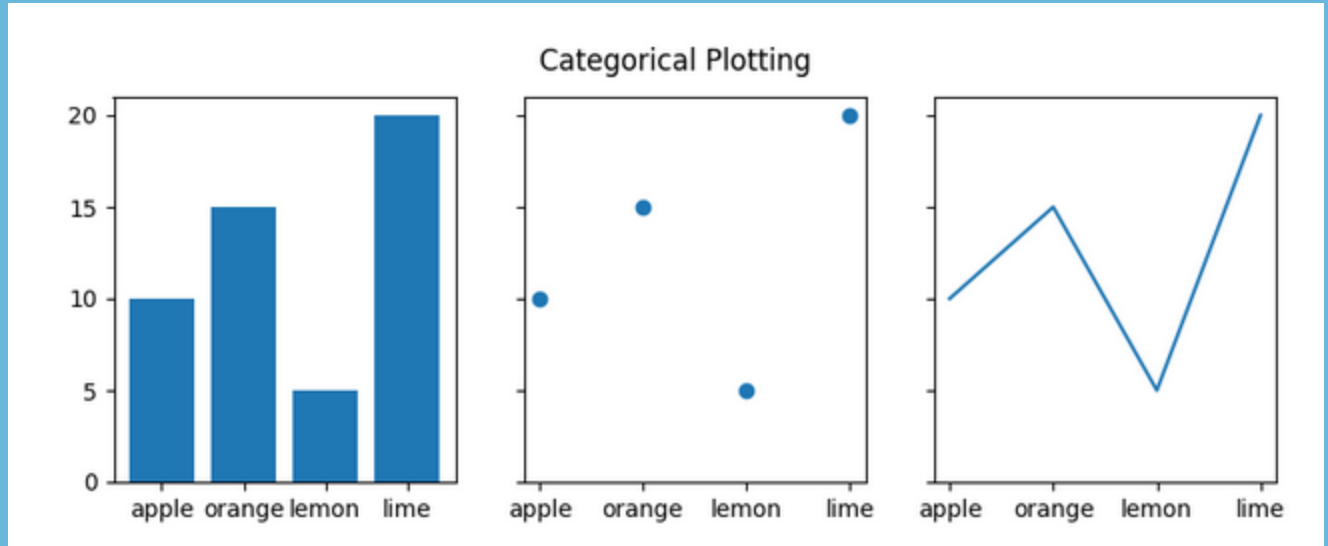
ggplot2 (R)

<https://ggplot2.tidyverse.org/>



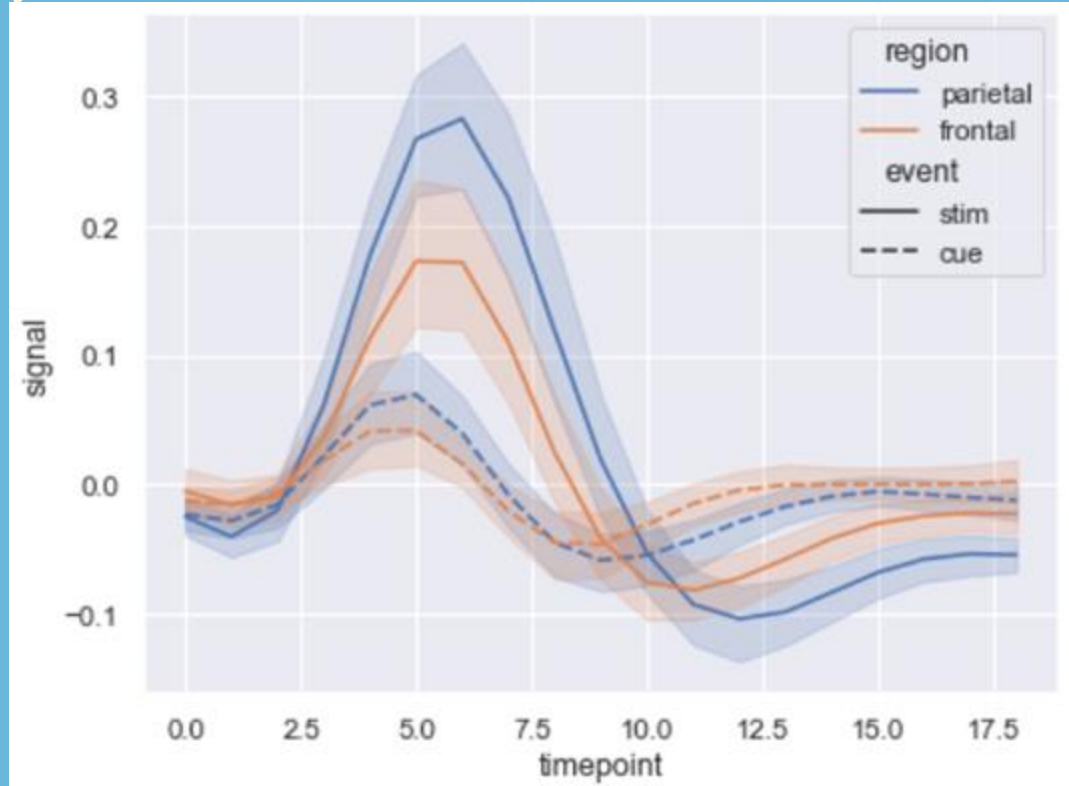
matplotlib (Python)

<https://matplotlib.org/>



seaborn (Python)

<https://seaborn.pydata.org/>



Plot.ly (R, Python)

Interaktywne wizualizacje w Javascript z interfejsem w Python i R.

<https://plotly.com/>

plotly.js: <https://github.com/plotly/plotly.js>

plotly.py: <https://github.com/plotly/plotly.py>

plotly.R: <https://github.com/plotly/plotly.R>