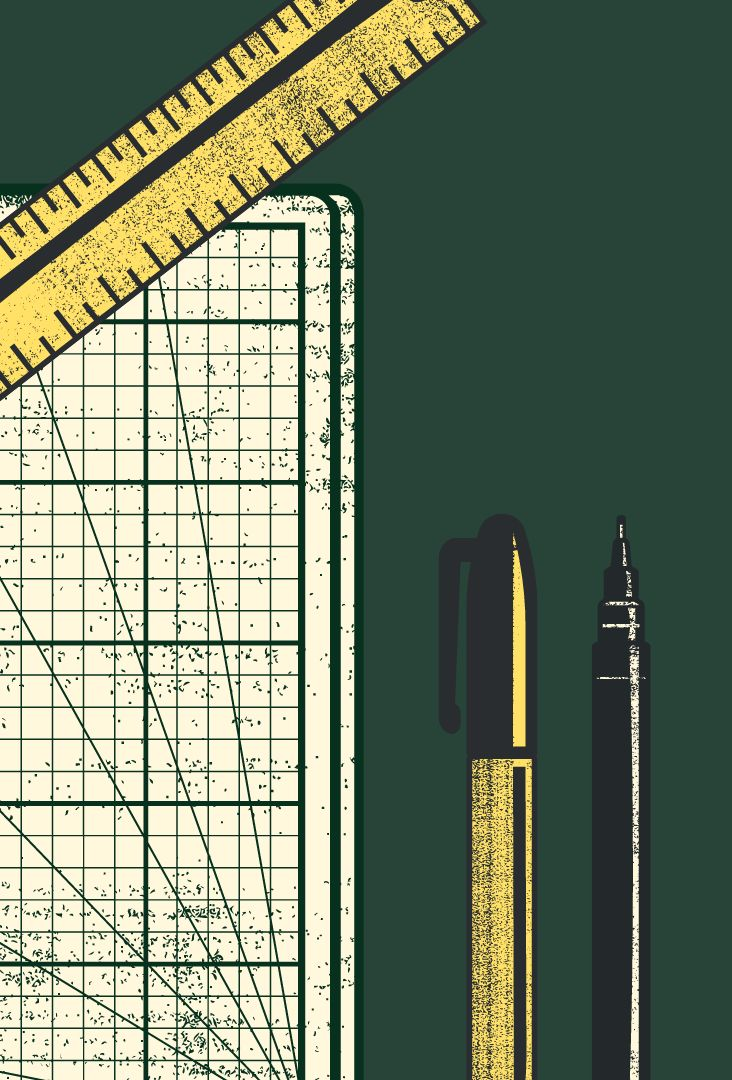


Wstęp do Eksploracji Danych

Politechnika Warszawska

Anna Kozak

Anna Kozak

anna.kozak@pw.edu.pl
MS Teams

@kozaka93

Hubert Ruczyński

hubert.ruczynski.stud@pw.edu.pl
MS Teams

@HubertR21

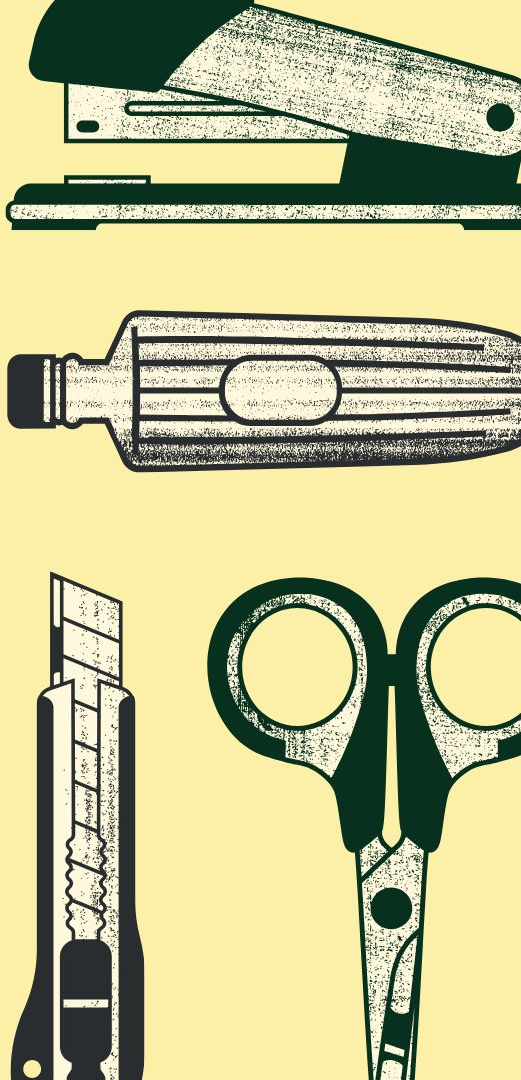
Maciej Chrabąszcz

maciej.chrabaszcz.dokt@pw.edu.pl
MS Teams

@maciejchrabaszcz

Strona przedmiotu

https://github.com/kozaka93/2024L-ExploratoryDataAnalysis



Wykład

Na wykładzie będą przedstawione zarówno teoretyczne aspekty pracy z danymi, jak i praktyczne.

15 wykładów = 13 x wykład + 2 x prezentacje projektów

Projekty

- 2 projekty w ciągu semestru
- zespoły 3 osobowe, różne podczas 1 i 2 projektu
- projekt trwa 7-8 tygodni
- 24p (P1) i 20p (P2) za projekt

*(w tym do 5p za pracę na zajęciach projektowych)

Laboratorium

- praca w R i Python
- powtórzenie operacji na danych (R: dplyr, tidyr; Python: pandas)
- wstęp do narzędzi pozwalających na estetyczne prezentowanie danych
- różne sposoby oceny zmiennych, danych, wizualizacji
- 6 x praca domowa (4 x 7p + 2 x 6p)
- 3 x wejściówka (3 x 2p)

Ocena końcowa

Suma punktów z prac domowych i projektów:

$4 \times 7 + 2 \times 6 + 24 + 20 + 3 \times 2 = 90$

(PD) + (PD) + (P1) + (P2) + (W) = (O)

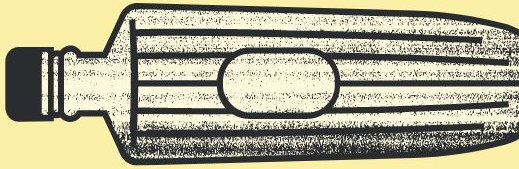
Aby zaliczyć kurs należy uzyskać ponad 45 punktów,
w tym co najmniej 50% punktów z każdego z projektów.

Zajęcia laboratoryjne są obowiązkowe, w ciągu semestru dopuszczalne są co najwyżej dwie nieusprawiedliwione nieobecności.

Oceny będą wystawiane zgodnie z tabelą:

Ocena	3	3.5	4	4.5	5
Punkty	(45, 54]	(54, 63]	(63, 72]	(72, 81]	(81, ∞)

Pytania?



Eksploracja 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?

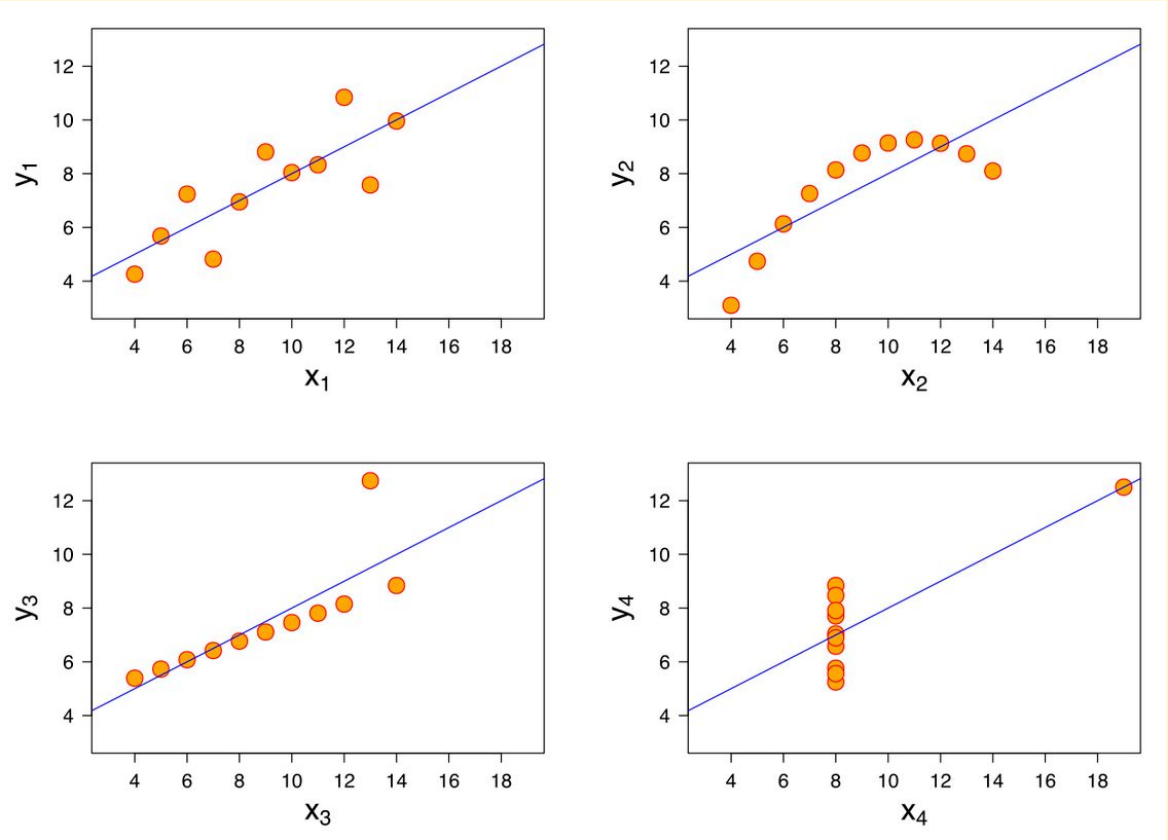
“proces odkrywania nietrywialnych, dotychczas nieznanych, potencjalnie użytecznych reguł, zależności, trendów”

Cel: analiza danych w celu lepszego ich zrozumienia

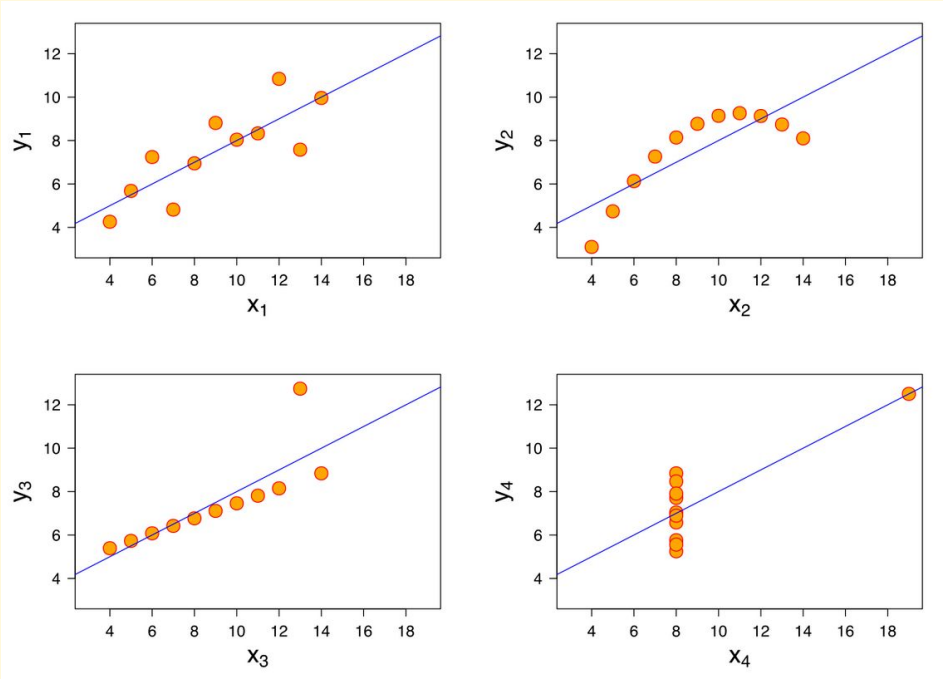
Eksploracja danych - czym jest?

Na eksplorację danych składa się wiele dyscyplin, między innymi:

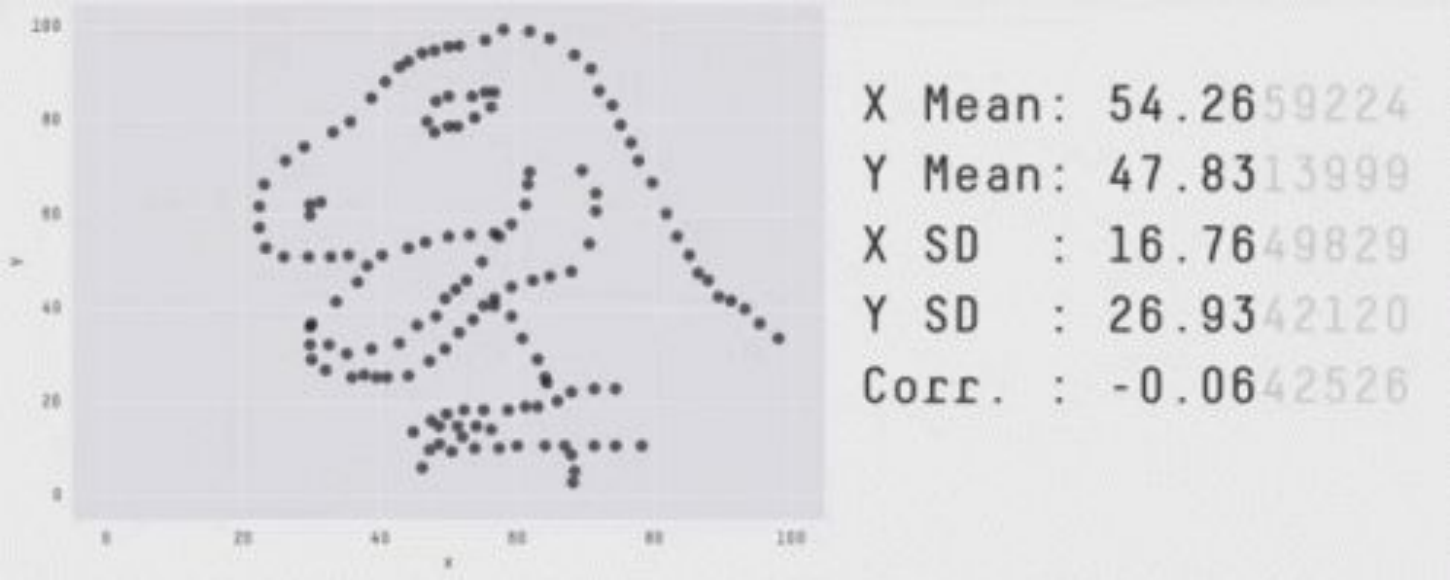
- bazy danych
- statystyka
- uczenie maszynowe
- wizualizacja danych
- wyszukiwanie informacji



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

13 zestawów danych ma te same statystyki zbiorcze (średnia x/y, odchylenie standardowe x/y i korelacja Pearsona) z dokładnością do dwóch miejsc po przecinku, a jednocześnie drastycznie różni się wyglądem.

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)
- programy dedykowane do wizualizacji danych (Tableau, Power BI)