Wstęp do eksploracji danych

Matematyka i Analiza Danych, II rok

Prowadzący

Anna Kozak - wykład, laboratoria

Kontakt:

- MS Teams
- 99999381@pw.edu.pl

Krzysztof Spaliński - laboratoria

Strona przedmiotu: https://github.com/mini-pw/2021L-ExploratoryDataAnalysis

Wstęp do eksploracji danych składa się z:

- wykładu
- zajęć laboratoryjnych

Wykłady - wtorki 12:15

Laboratoria - wtorki i środy o 8:30

Wykład

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

8 wykładów = wykład (45 minut) + projekt (45 minut)

Dlaczego projekt?

- pozwala na ćwiczenie praktyk analizy danych, wizualizacji, sprawdzania postawionych pytań badawczych
- wykorzystanie wiedzy z wykładu i laboratorium

Projekty

- 2 projekty w ciągu semestru
- zespoły 3 osobowe, różne podczas 1 i 2 projektu
- projekt trwa 3 - 4 tygodnie
- efektem końcowym jest przygotowanie plakatu na zadany temat/pytanie badawcze (więcej na pierwszym spotkaniu projektowym)
- osoby w zespole mogą być z różnych grup laboratoryjnych
- 20 pkt za projekt

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
- pierwsza połowa w R (A. Kozak), druga połowa w Python (K. Spaliński)
- 8 x praca domowa za 5 pkt

Ocena końcowa

Suma punktów z prac domowych i projektów.

Data oddania ostatniej składowej do oceny z przedmiotu jest na dwa tygodnie przed sesją.

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?

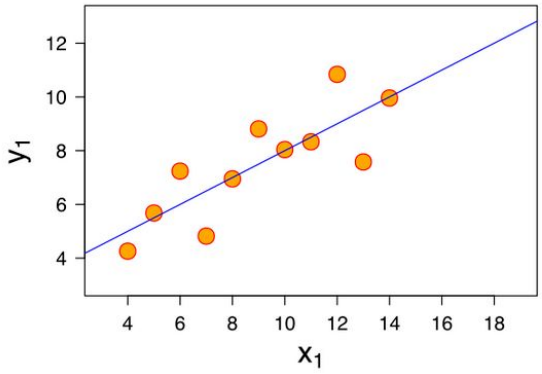
"proces odkrywania nietrywialnych, dotychczas nieznanych, potencjalnie użytecznych reguł, zależności, trendów"

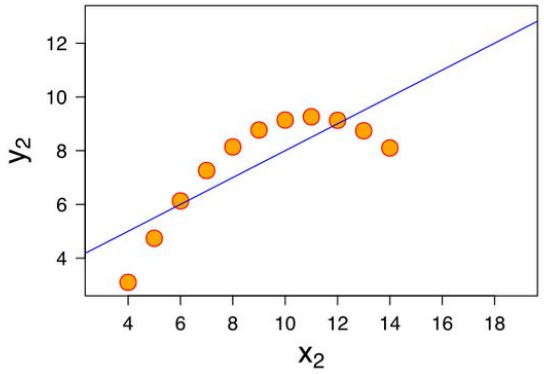
Cel: analiza danych w celu lepszego ich zrozumienia

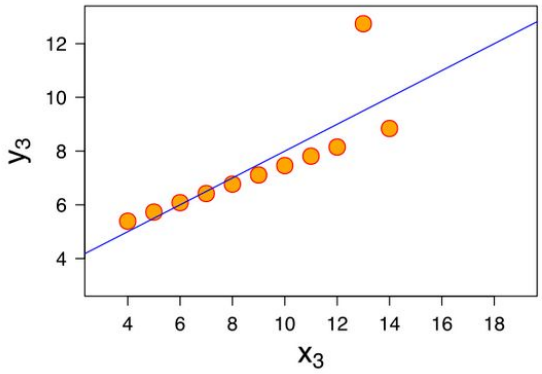
Eksploracja danych - czym jest?

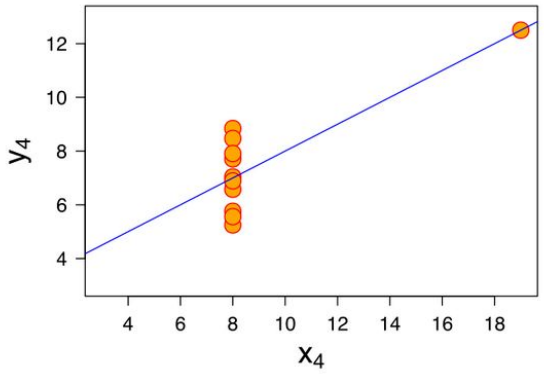
Na eksplorację danych składa się wiele dyscyplin, między innymi:

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

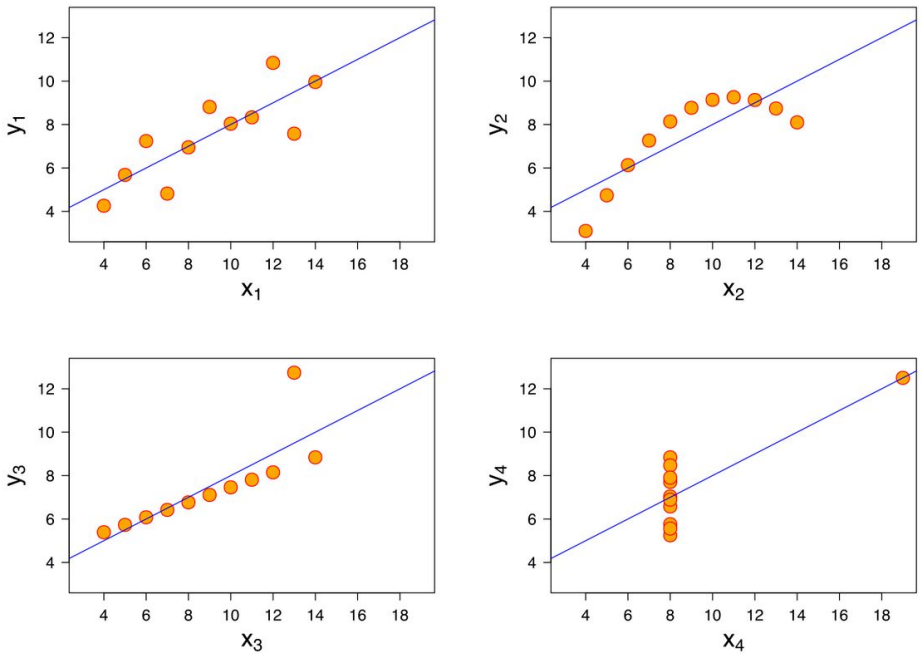








Kwartet Anscombe'a



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)

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

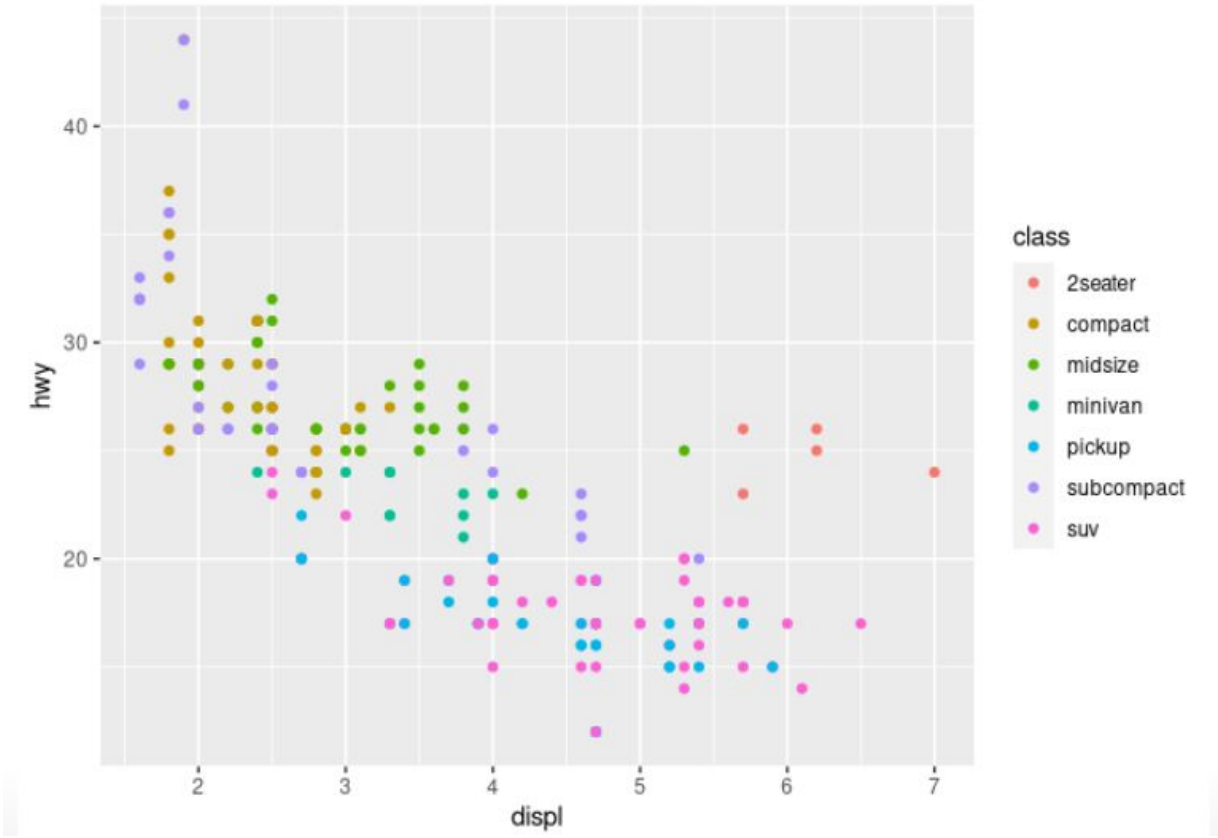
ID	PŁEĆ	ZAWÓD	WZROST	DATA URODZENIA
ID_23	K	INFORMATYK	158	1978-03-12
ID_45	K	PRAWNIK	178	1989-05-29
ID_46	M	MATEMATYK	183	1991-01-19
ID_89	M	INFORMATYK	167	1982-02-20
ID_101	K	LEKARZ	163	1973-02-23

Narzędzia do wizualizacji danych

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

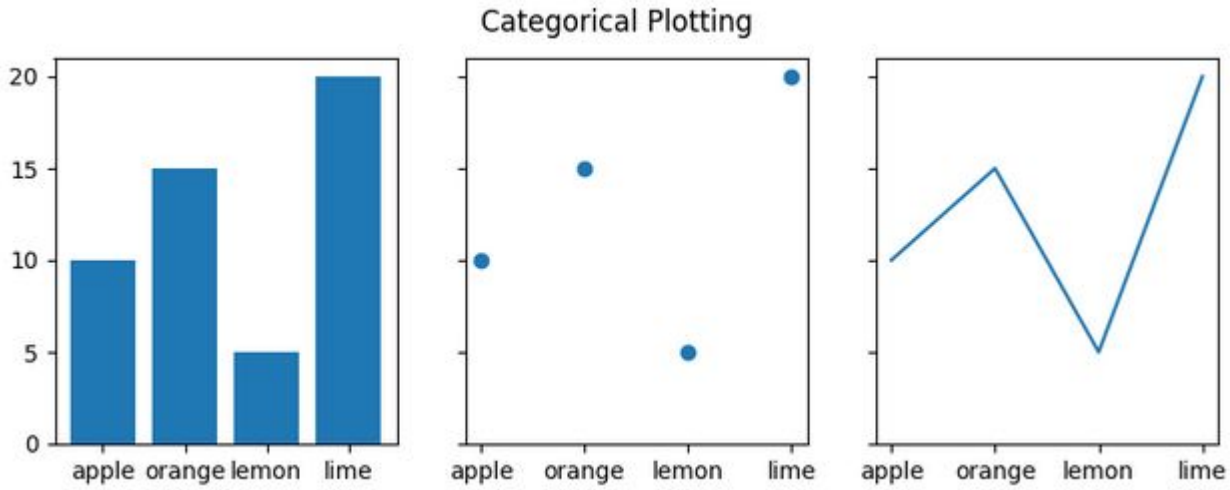
ggplot2 (R)

https://ggplot2.tidyverse.org/



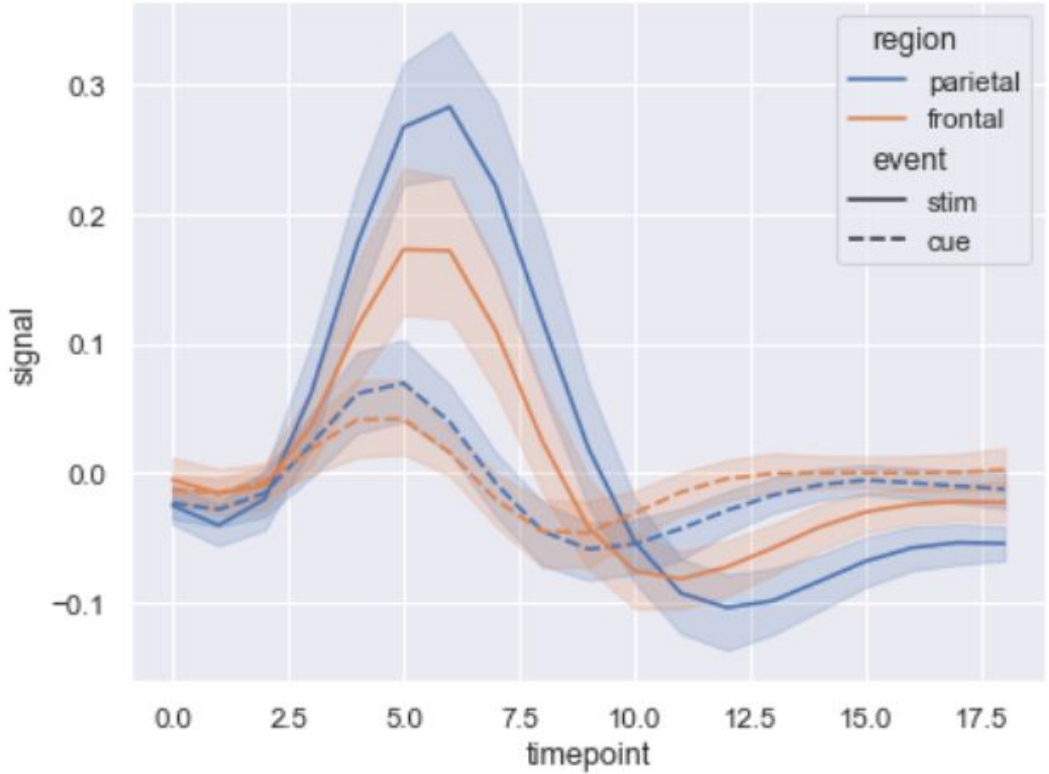
matplotlib (Python)

https://matplotlib.org/



seaborn (Python)

https://seaborn.pydata.org/



plot.ly

Interaktywne wizualizacje w Javascript z interfejsem w Python i R.

https://plotly.com/python/line-and-scatter/

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

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

plotly.R: https://github.com/ropensci/plotly