

Techniki Wizualizacji Danych

Politechnika Warszawska

Anna Kozak | Mateusz Krzyziński |
Hubert Ruczyński | Mikołaj Spytek

Prowadzący przedmiot



Anna Kozak

Data Scientist | Data Visualization | Responsible Machine Learning

Data Scientist (5 years)

Research and teaching assistant - MI² (3 years)

Research Assistant - WUT (2 years)

✉ anna.kozak@pw.edu.pl

/github kozaka93

/in kozakanna

/twitter kozaka93

/linked kozaka



Mateusz Krzyziński

Explainable Artificial Intelligence | Survival Analysis | Data Visualization

Data Science Student - WUT
Researcher - MI²

✉ mateusz.krzyzinski.stud@pw.edu.pl

 krzyzinskim

 krzyzinskim

 krzyzinskim

 krzyzinskim



Hubert Ruczyński

Data Scientist | AutoML | Natural Language Processing

Data Science Student - WUT
Researcher - MI²

✉ hruczynski21@interia.pl

 HubertR21

 hubert-ruczyński

 HubertRuczynski

 hruczynski21



Mikołaj Spytek

Data Science | Data Visualization | Responsible ML

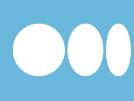
Data Science Student - WUT
Researcher - MI²

✉ mikolaj.spytek.stud@pw.edu.pl

 mikolajsp

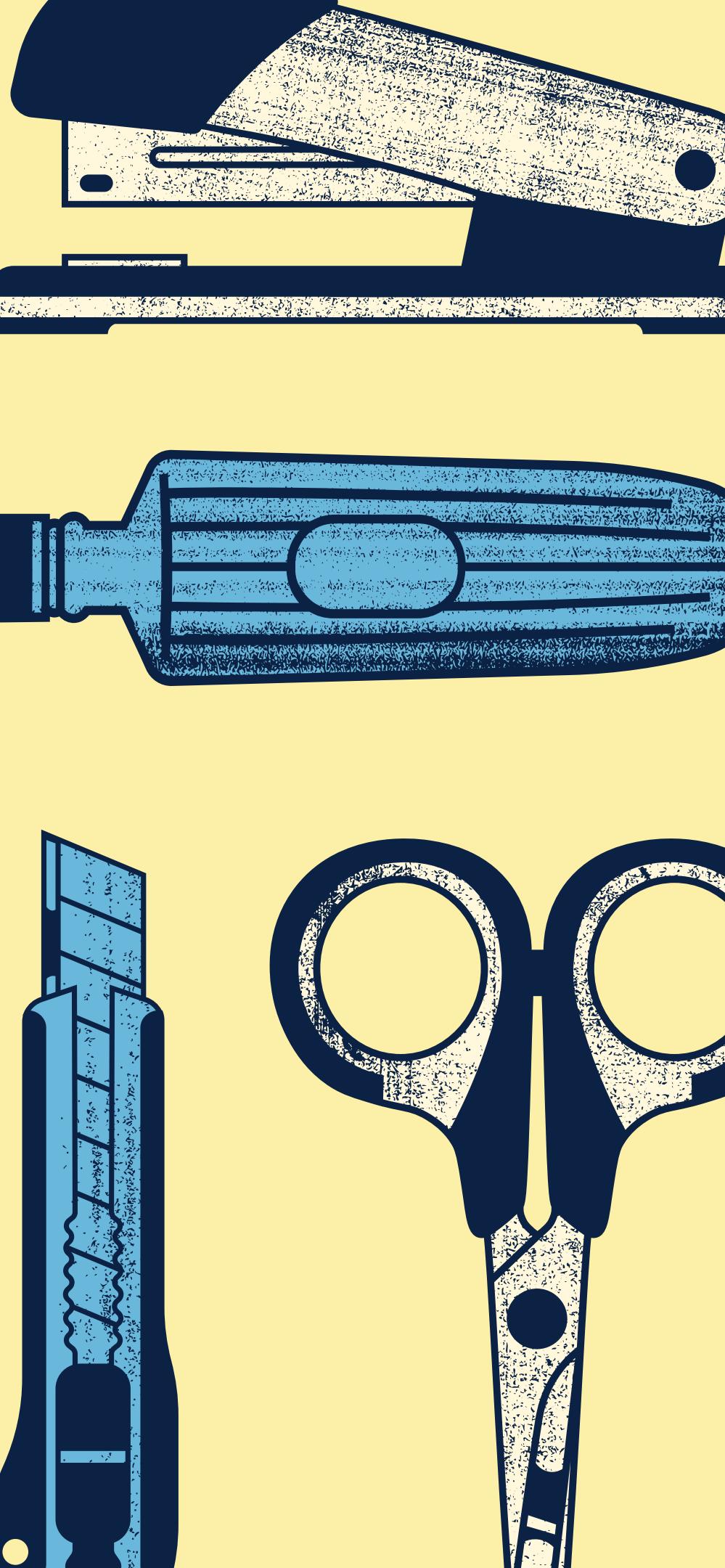
 mikolajsp

 spytekm

 spytekm

Strona przedmiotu

[https://github.com/MI2-Education/
2023Z-DataVisualizationTechniques](https://github.com/MI2-Education/2023Z-DataVisualizationTechniques)



Zajęcia

8:15 Wykład, s. 107

9:15 Zajęcia projektowe, s. 301, 302, 303, 218

10:15 Laboratoria, s. 301, 302, 303, 218

Wykład

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

15 wykładów = 12 x wykład + 3 x prezentacje projektów

Zajęcia projektowe

- 2 projekty w ciągu semestru
- zespoły 3 osobowe, różne podczas 1 i 2 projektu
- projekt trwa 7-8 tygodni
- 25 pkt za projekt (w tym 5 pkt 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
- 8 x praca domowa (6 x 5 pkt + 2 x 10 pkt)

Ocena końcowa

Suma punktów z prac domowych i projektów:

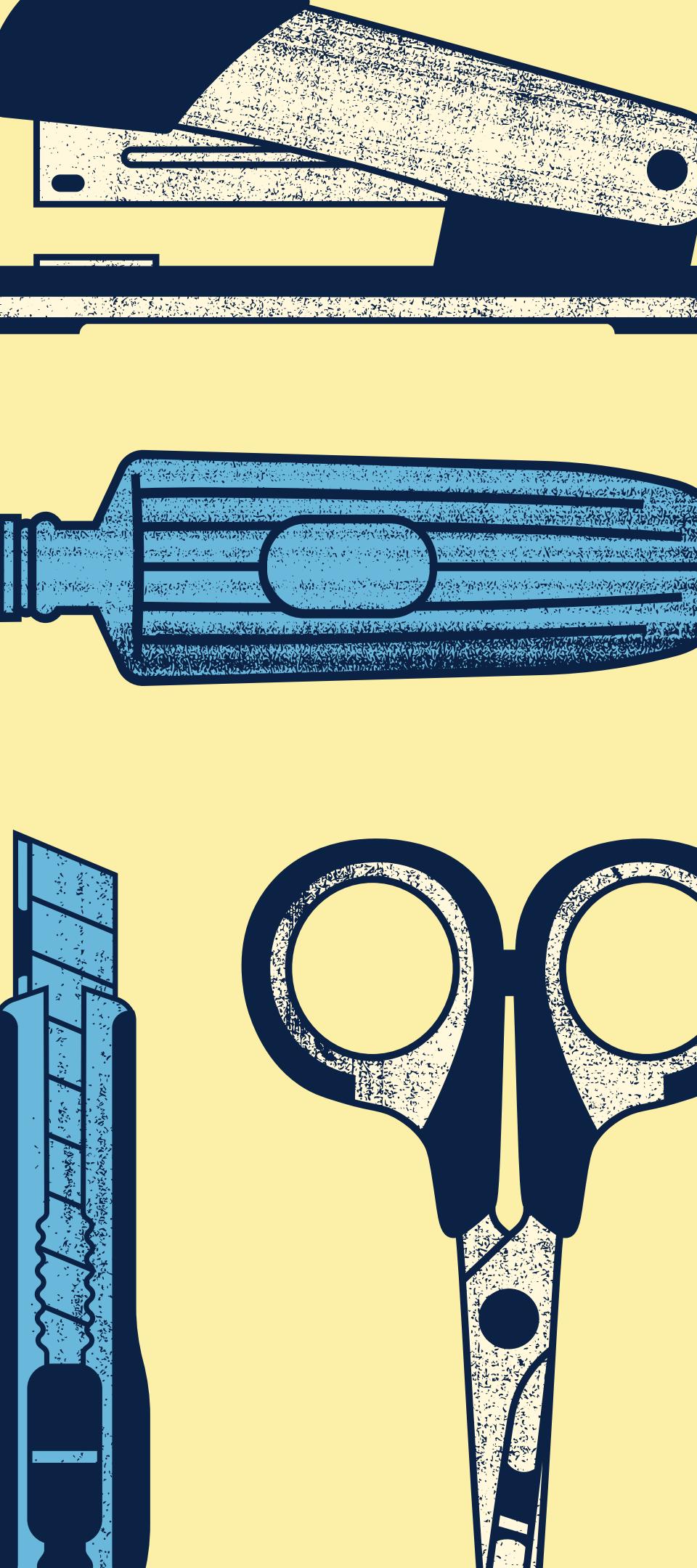
$$2 \times 25 + 2 \times 10 + 6 \times 5 = 100$$

Aby zaliczyć kurs należy uzyskać co najmniej 51 punktów,
w tym co najmniej 13 punktów z każdego z projektów.

Oceny będą wystawiane zgodnie z tabelą:

Ocena	3	3.5	4	4.5	5
Punkty	(50, 60]	(60, 70]	(70, 80]	(80, 90]	(90, ∞)

Pytania?





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?

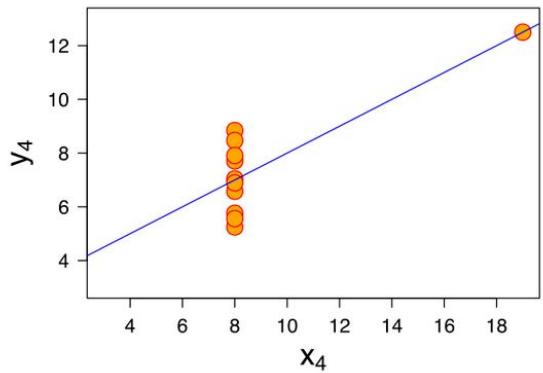
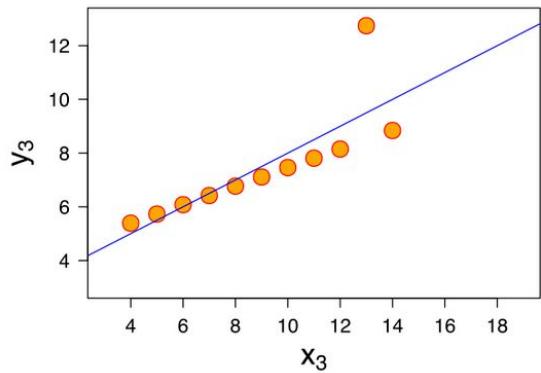
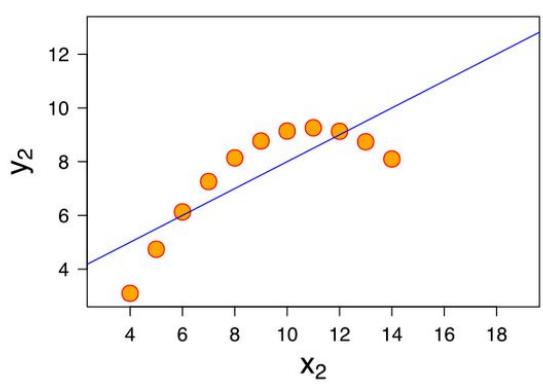
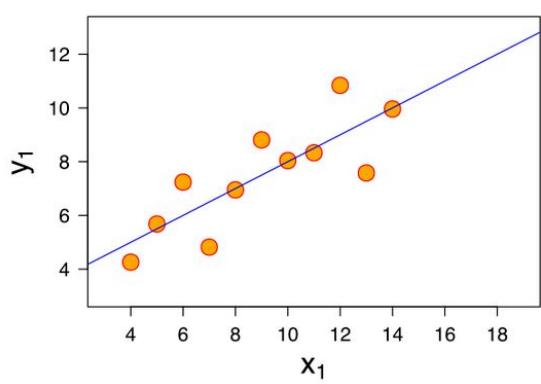
“proces odkrywania nietrywialnych, dotychczas nieznanych, potencjalnie użytecznych reguł, zależności, trendów”

Cel: analiza danych w celu lepszego ich zrozumienia

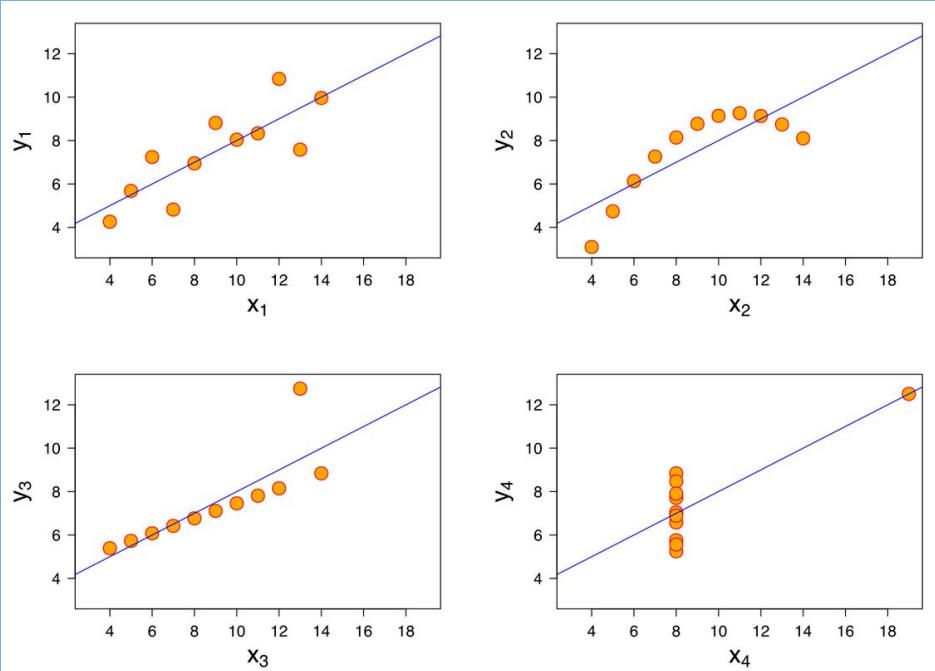
Eksploracja danych - czym jest?

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, dydaktycznymi) 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

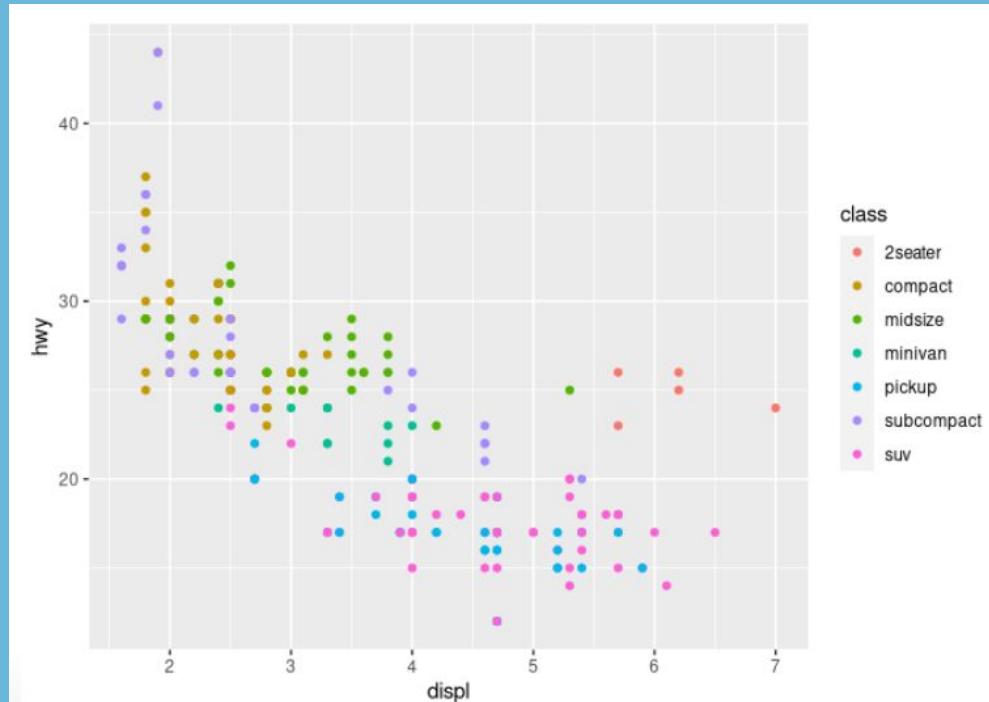
ID	PŁEĆ	ZAWÓD	WZROST	DATA URODZENIA
ID_23	K	INFORMATYK	158	1978-03-12
ID_45	K	PRAWNIK	178	1989-05-29
ID_46	M	MATEMATYK	183	1991-01-19
ID_89	M	INFORMATYK	167	1982-02-20
ID_101	K	LEKARZ	163	1973-02-23

Narzędzia do wizualizacji danych

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

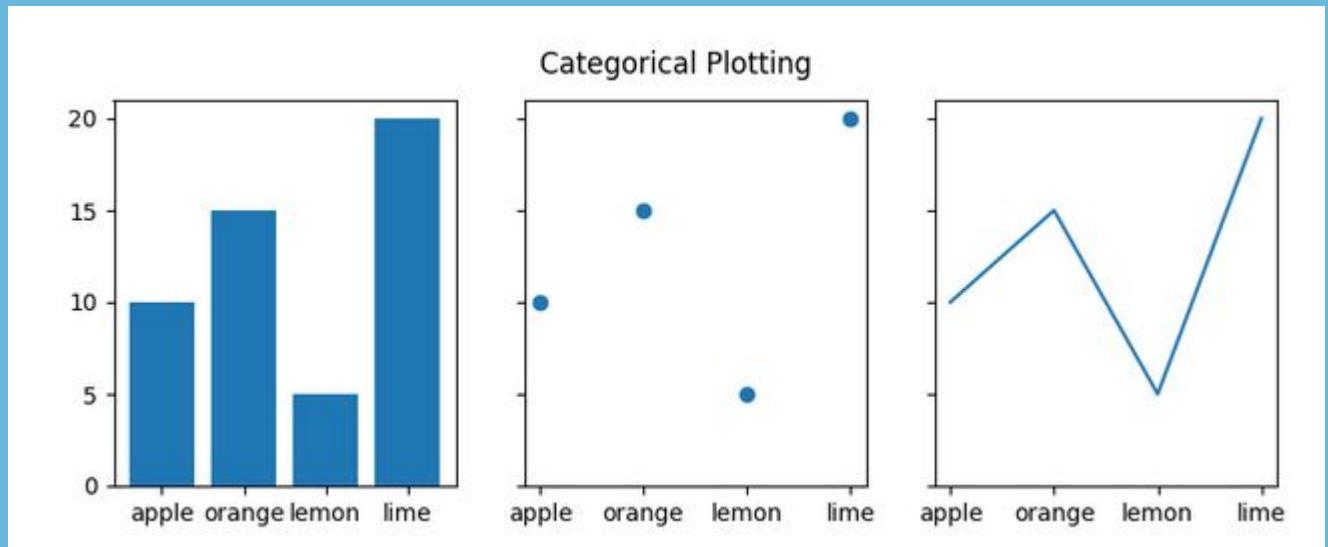
ggplot2 (R)

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



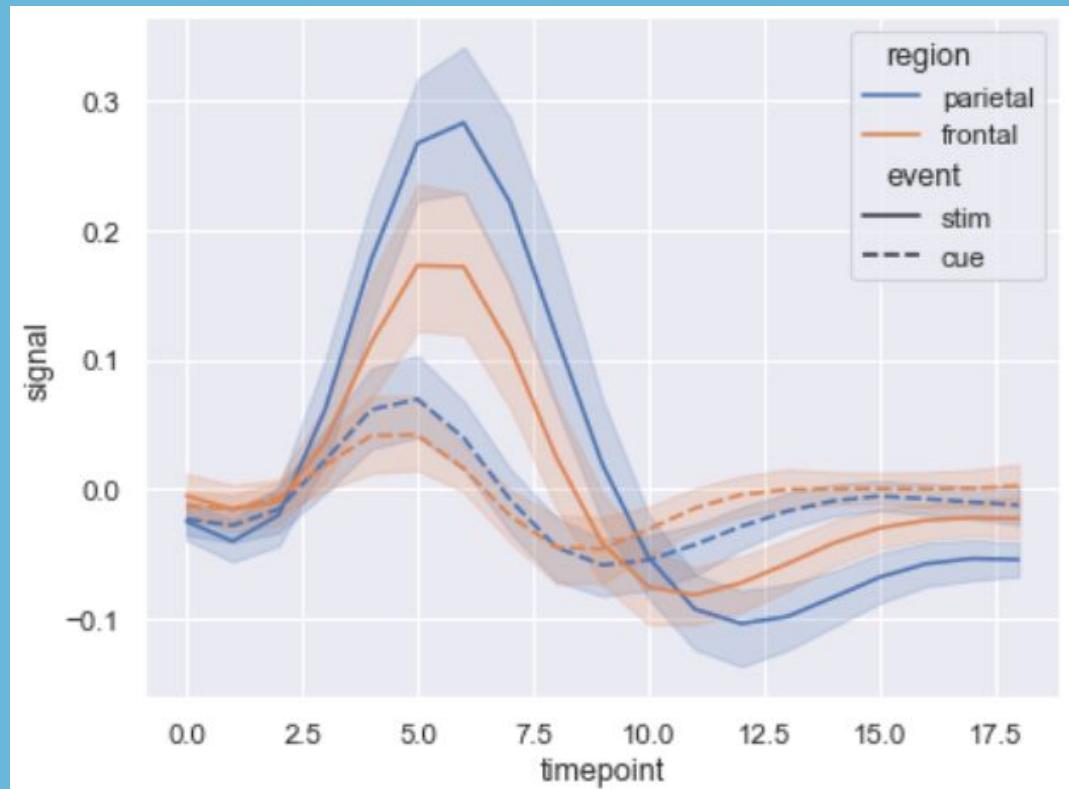
matplotlib (Python)

<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>