

Krótkie wprowadzenie do wizualizacji danych i eksploracyjnej analizy danych

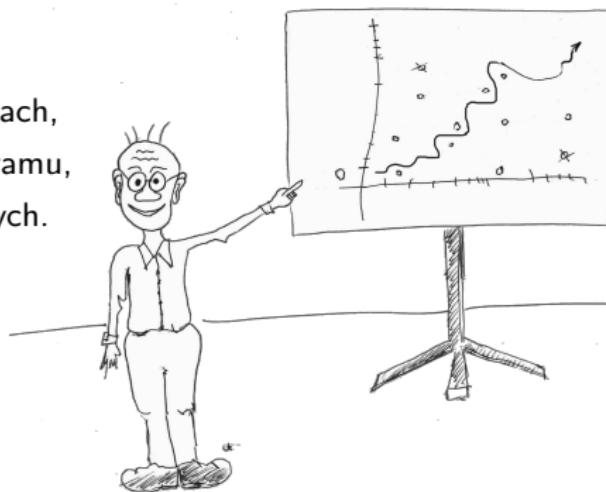
Przemysław.Biecek@gmail.com, MIMUW / SmarterPoland.pl

Plan części 1:

- ① Czym jest ten data mining?
- ② Skalowanie wielowymiarowe na przykładach,
- ③ Analiza skupisk na przykładzie dendrogramu,
- ④ Klasyfikacja na przykładzie lasów losowych.

Plan części 2:

- ① Ładniej nie znaczy lepiej,
- ② Po pierwsze „nie kłamać”,
- ③ Dobre rady wujka Edwarda Tufte'go.



Co to ten Data Mining?

To paraphrase provocatively, 'data mining is statistics minus any checking of models and assumptions'.

Brian D. Ripley (zapytany o różnice pomiędzy DM a statystyką).

Rozwój technik komputerowych, baz danych oraz galopująca miniaturyzacja umożliwiają zbieranie coraz większych zbiorów danych. W wielu badaniach analizowane są tysiące zmiennych i setki milionów przypadków. Standardowe techniki statystyczne słabo nadają się do analizy takich zbiorów danych. Np. problem kontroli błędu pierwszego rodzaju traci na znaczeniu, co podważa stosowalność całej klasycznej teorii testowania w analizie naprawdę dużych zbiorów danych.

Terry Speed w swojej kolumnie w biuletynie IMS często podkreśla, że autorami nowych technik analizy danych coraz rzadziej są matematycy czy statystycy a coraz częściej informatycy, biolodzy, fizycy itp. Nie rozwijają oni co prawda matematycznych, jednolitych teorii ale tworzą algorytmy rozwiązuje (z różnym skutkiem) konkretne problemy.

Co to ten Data Mining?

Nie ma (=nie znam) powszechnie przyjętej definicji dziedziny data mining.

To określenie pojawia się najczęściej w sytuacji gdy analizowane są duże zbiory danych. Roboczo przyjmijmy więc, że data mining to metody analizy dużych zbiorów danych.

Można wyróżnić trzy grupy metod, które często pojawiają się w podręcznikach czy szkoleniach poświęconych tematyce data mining.

Są to:

- Skalowanie wielowymiarowe, konstrukcja cech, wybór cech,
- Analiza skupisk, podejście modelowe, kombinatoryczne, hierarchiczne i mieszane,
- Klasifikacja, tak podejście parametryczne jak i nieparametryczne.

Skalowanie wielowymiarowe jest techniką pozwalającą na transformacje, najczęściej redukcje, przestrzeni w której opisywane są obiekty. Zazwyczaj oznacza to redukcje przestrzeni cech do przestrzeni \mathcal{R}^2 lub \mathcal{R}^3 .

Do tego celu może służyć wiele technik i algorytmów. Poniżej, w krótkich żołnierskich słowach, opiszemy techniki PCA i MDS.

Skalowanie wielowymiarowe ma wiele zastosowań, w tym:

- ułatwia wizualizacje, najczęściej przez redukcje danych do 2-3 wymiarów,
- etap pośredni w analizie skupisk,
- ekstrakcja cech, w sytuacji gdy oryginalnych zmiennych jest zbyt dużo by móc je analizować (np. PCR).

Przykład: Psychotyczność pacjentów z projektu EDEN

Zastosowanie skalowania wielowymiarowego przedstawimy na przykładzie danych pochodzących z projektu EDEN. W ramach tego projektu przebadano pod kątem psychiatrycznym ponad 3 tysiące pacjentów z 6 europejskich ośrodków. Dla każdego pacjenta określono ponad tysiąc cech, między innymi początkową psychotyczność mierzoną w czterech obszarach oraz początkową jakość życia.

Pytanie: jak określać podobieństwo pomiędzy pacjentami, czy zbiór danych jest jednorodny ze względu na te cechy, czy w danych można wyróżnić typowe profile pacjentów, czy są wyraźne skupiska w danych?

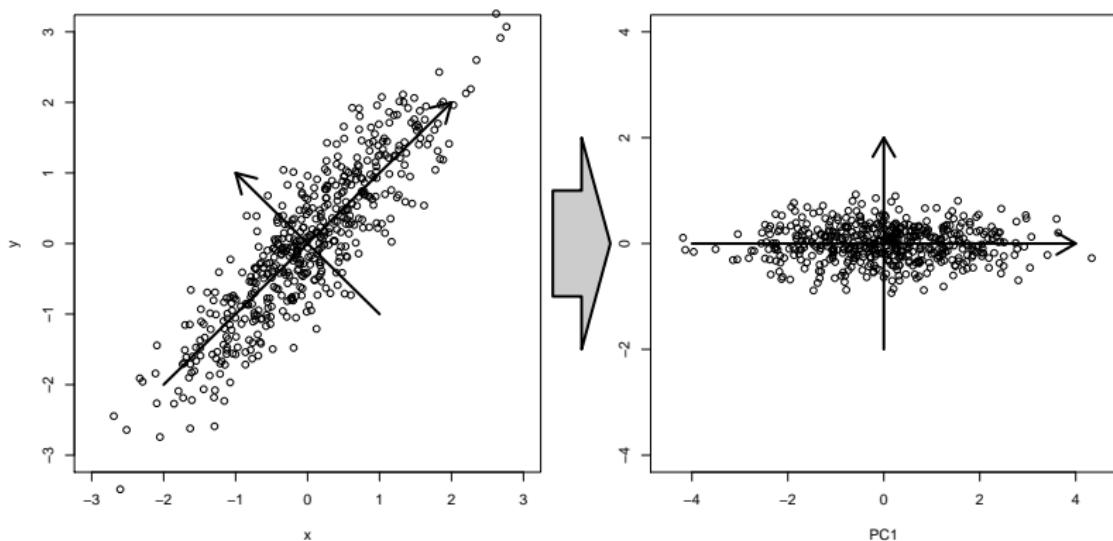
	BPRS.Maniac	BPRS.Negative	BPRS.Positive	BPRS.Depression	MANSA
1	1.8	2.3	2.1	3.4	3.1
2	1.8	1.3	1.1	2.1	3.0
3	1.0	2.0	1.3	3.0	4.1
4	1.3	1.1	1.1	1.5	4.0
5	1.2	1.9	1.4	3.4	4.1
6	1.3	1.9	1.3	4.0	4.0

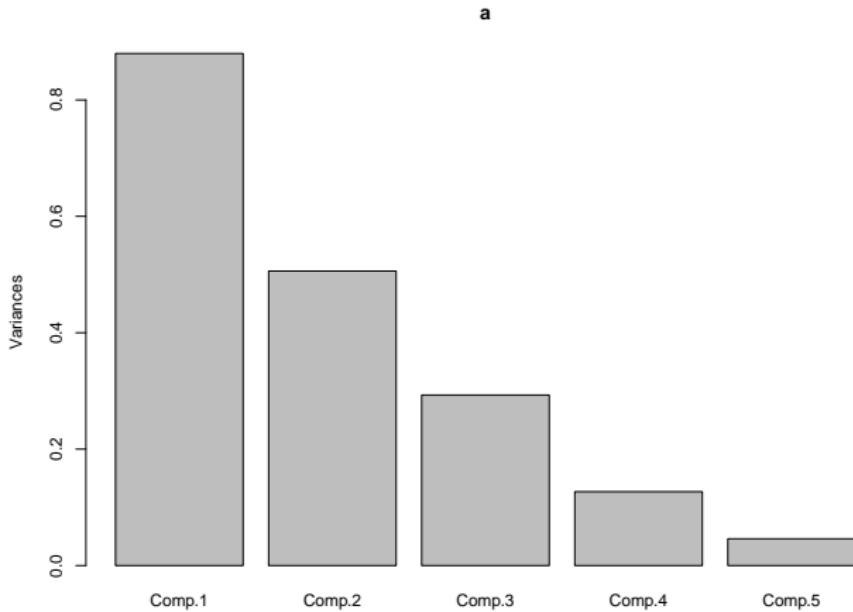
Metoda PCA pozwala na przekształcenie danych do przestrzeni k -wymiarowej. W opisanym przykładzie przekształcimy 5-wymiarową przestrzeń cech do dwuwymiarowej przestrzeni.

Wykorzystany jest następujący algorytm:

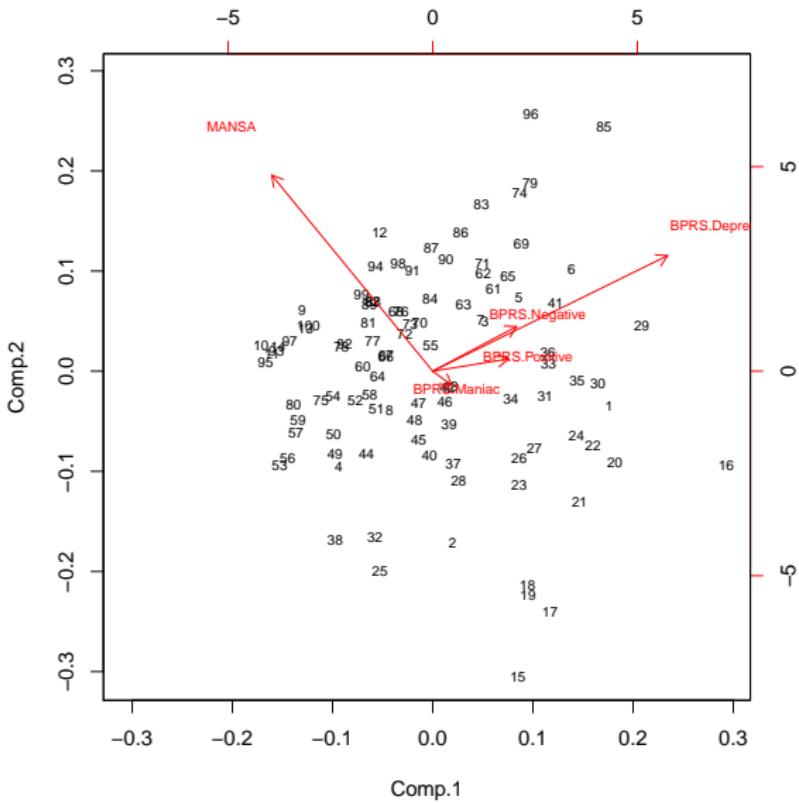
- wyznacz rozkład SVD dla analizowanej macierzy $X = UDV^T$.
Macierz U opisuje bazę ortonormalną w przestrzeni kolumn macierzy X a macierz V opisuje bazę w przestrzeni wierszy,
- jako nowe współrzędne wybierz k pierwszych wektorów z macierzy V ,
- wyznacz współrzędne obiektów w nowym (zredukowanym) układzie współrzędnych.

Analiza składowych głównych (PCA, ang. Principal Components Analysis)





Zmienna w danych wyjaśniona przez kolejne współrzędnej nowej bazy.

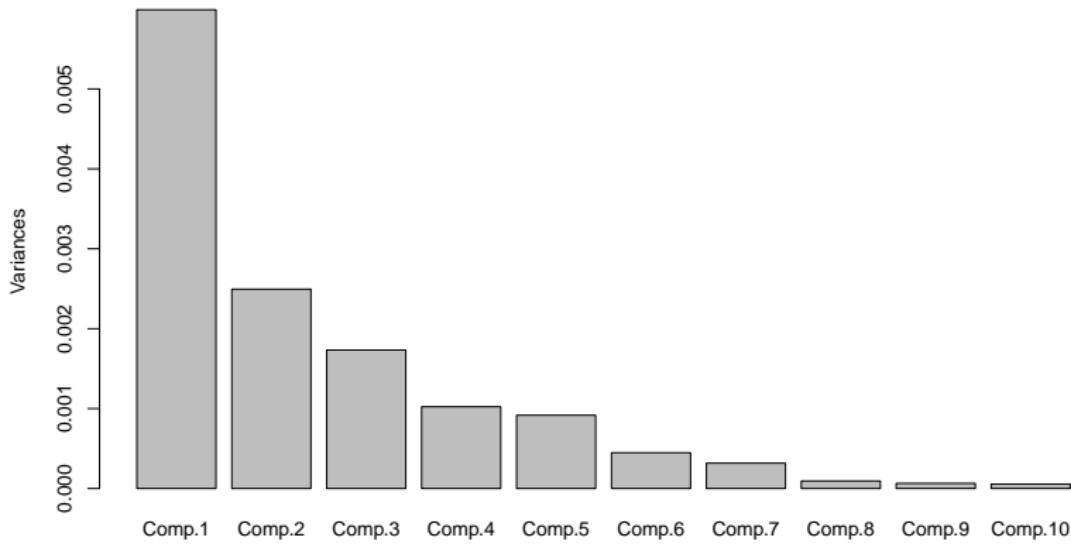


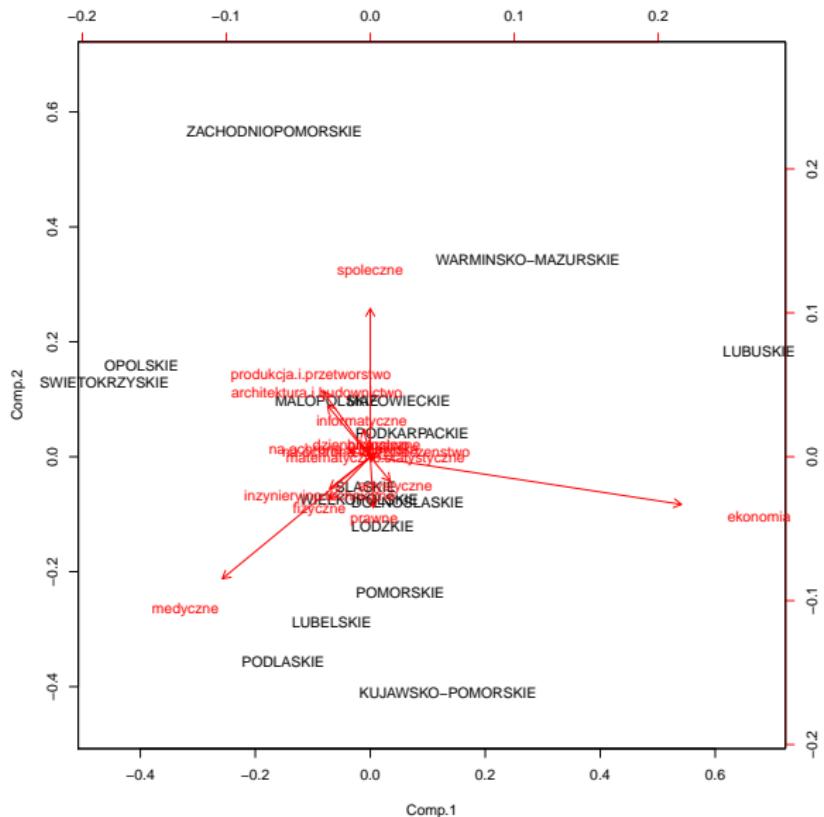
Zobaczmy teraz zbiór danych pobrany ze strony GUSu dotyczący liczby studentów studujących w określonym województwie określony kierunek studiów.

Zbiór danych dotyczy 16 województw i 15 wyróżnionych kierunków.

Pytanie: Które województwa mają podobną strukturę kierunków studiów.

	studenci.artystyczne	studenci.spoleczne	studenci.ekonomia
DOLNOSLASKIE	2179	13313	23998
KUJAWSKO-POMORSKIE	1205	5097	10098
LODZKIE	2934	10242	13990
LUBELSKIE	341	5971	9655
LUBUSKIE	498	3474	7772
MALOPOLSKIE	2741	19528	23555
MAZOWIECKIE	3231	27381	30501





Dla metody PCA skalowanie było przeprowadzane tak by w „docelowej” przestrzeni odległość pomiędzy obiektami była możliwie bliska odległości w przestrzeni oryginalnej. Odpowiada to minimalizacji tzw. stresu (kalka z ang. stress)

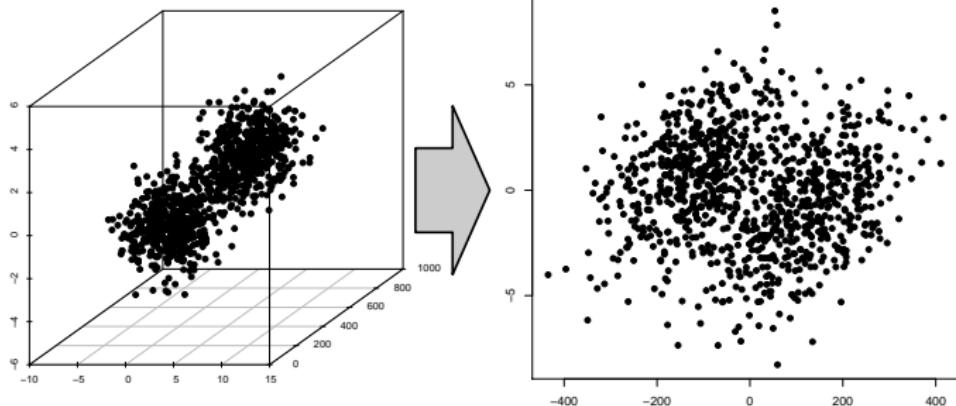
$$\text{stress} = \frac{\sum_{i,j} (d_{ij} - \tilde{d}_{ij})^2}{\sum_{i,j} d_{ij}}.$$

W przypadku metody MDS minimalizowana jest wartość

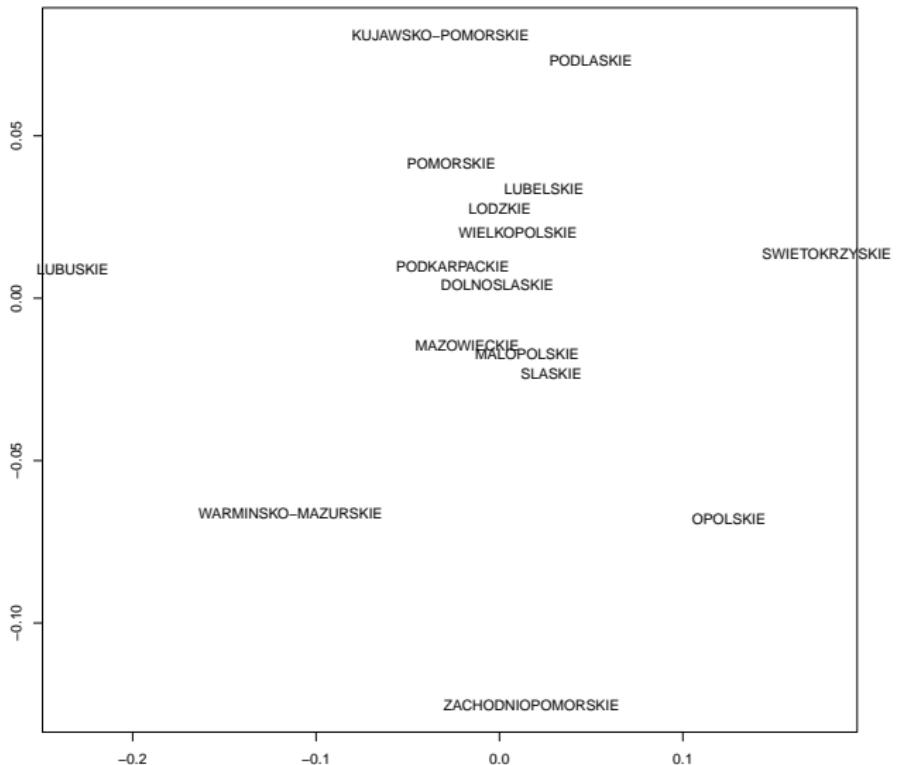
$$\text{stress} = \frac{\sum_{i,j} (f(d_{ij}) - \tilde{d}_{ij})^2}{\sum_{i,j} f(d_{ij})^2},$$

gdzie \tilde{d}_{ij} to odległość pomiędzy obiektami i i j w nowej k -wymiarowej przestrzeni a d_{ij} to oryginalne odległości pomiędzy obiektami przekształcone przez pewną monotoniczną funkcję $f()$ (więc d_{ij} i \tilde{d}_{ij} mogą być w różnych skalach!).

Skalowanie wielowymiarowe Kruskalla (MDS, ang. Multidimensional Scaling)



Skalowanie wielowymiarowe Kruskalla (MDS, ang. Multidimensional Scaling)



Analiza skupisk pozwala na wyróżnienie zbiorów obserwacji (nazywanych skupiskami lub klastrami) podobnych do siebie. Proces szukania podziału na grupy, nazywany jest czasem klastrowaniem (kalka z ang. clustering). Jest to szybko rozwijany zbiór metod obejmujący metody kombinatoryczne, hierarchiczne, modelowe i inne.

Poniżej przedstawimy przykład analizy hierarchicznej aglomeracyjnej. Popularnie używane są również metody hierarchiczne dzielące, metoda k-średnich i metoda k-medoidów.

Analiza skupisk ma wiele zastosowań, w tym:

- segmentacja klientów,
- identyfikacja wspólnie regulowanych genów,
- konstrukcja portfeli akcji lub innych instrumentów,
- filogenetyka i konstrukcja rodowodów gatunków.

AGglomerative NESting (AGNES) jest najpopularniejszą metodą aglomeracyjną.

Dendrogram opisujący skupiska tworzony jest w następujący sposób

- ① Każdą obserwację traktujemy jako osobne skupisko.
- ② Najdujemy dwa skupiska najbliższe sobie. Odległość pomiędzy skupiskami można wyznaczać na różne sposoby (tzw. szczególny implementacyjny).
- ③ Łączymy dwa najbliższe skupiska w jedno i przeliczamy odległości pomiędzy tym nowym skupiskiem a pozostałymi.
- ④ Jeżeli pozostało więcej niż jedno skupisko to wracamy do kroku 2.

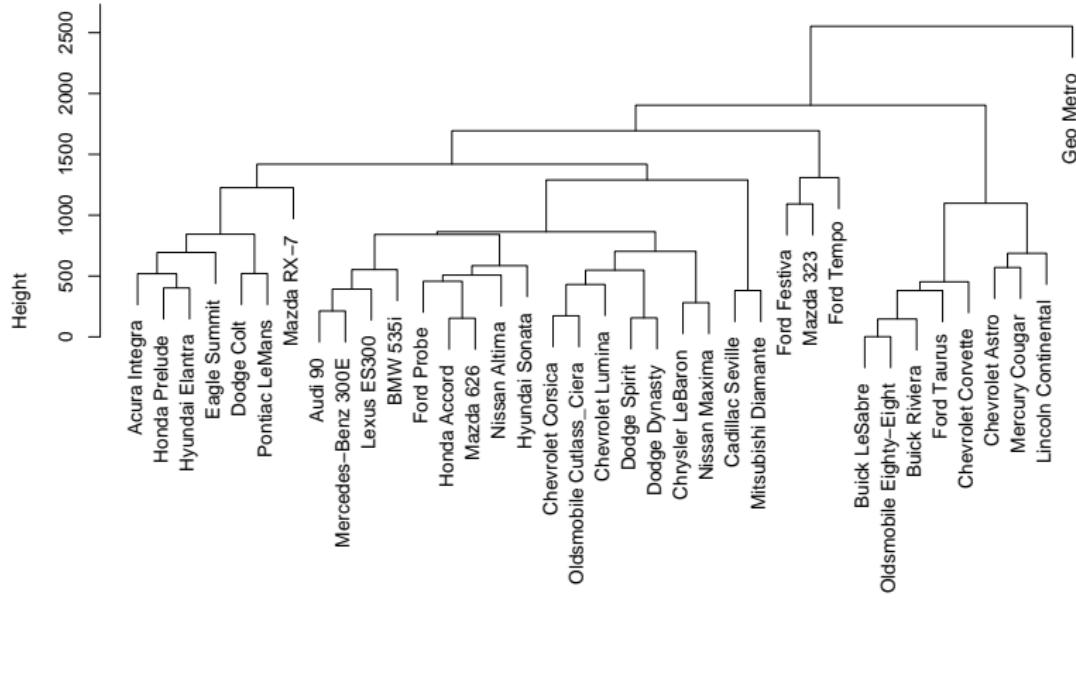
Przykład: samochody

Przedstawimy analizę skupisk na przykładzie zbioru danych dotyczącego samochodów.

Dla 93 różnych modeli samochodów zebrano informacje o 27 cechach, takich jak moc silnika, pojemność, rozmiar, liczba drzwi, liczba poduszek, liczba cylindrów, liczba koni mechanicznych, maksymalne obroty, spalanie w trasie, mieście i średnie, wielkość bagażnika, masa itp.

Na podstawie tych danych chcemy ocenić, które samochody są do siebie podobne i które z nich można ewentualnie połączyć w grupy.

Przykład: samochody



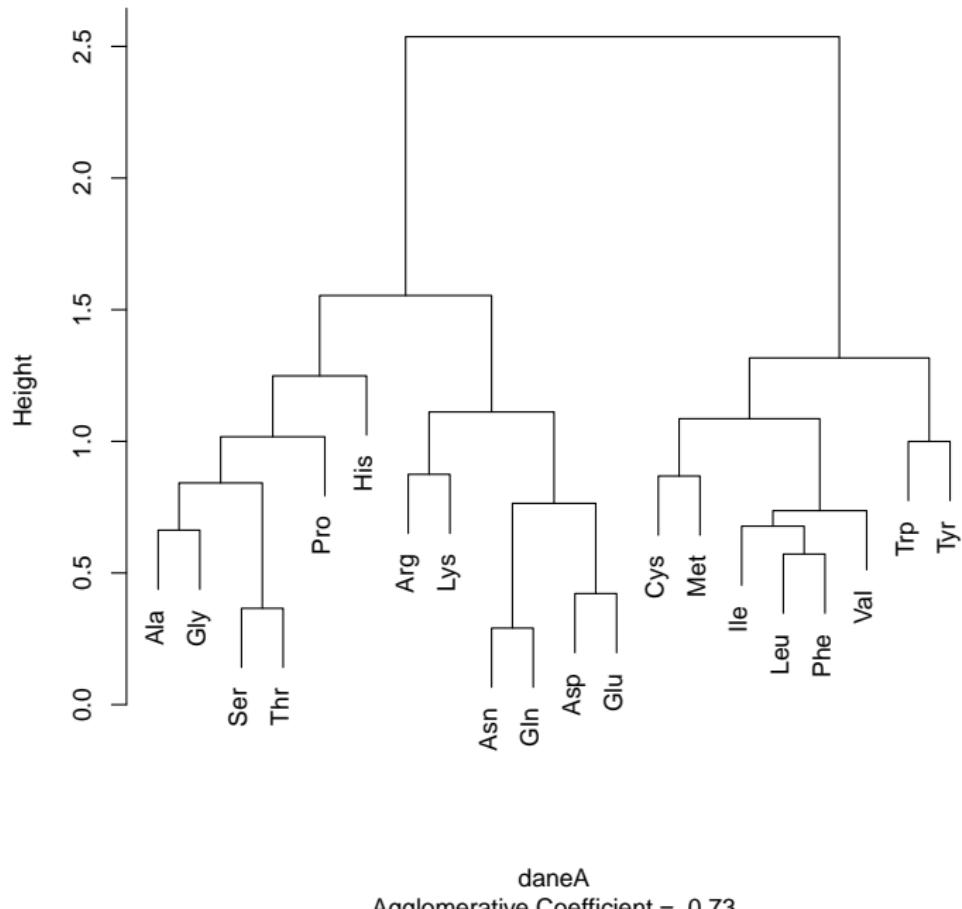
Przykład: aminokwasy

Białka wszystkich żywych organizmów składają się z aminokwasów. Różnych aminokwasów jest jedynie 20, każdy o innych właściwościach fizyko-chemicznych (zasadowość, polarność, rozmiar, homofilność, itd.).

Dysponujemy opisem każdego z 20 aminokwasów przez 23 różne współczynniki. Interesuje nas określenie podobieństwa pomiędzy aminokwasami na podstawie tych współczynników.

	Sweet	Kyte.and.Doolittle	Abraham.and.Leo	Bull.and.Breese	Guy
Ala	0.2817337	0.7000000	0.576846307	0.8625954	0.5508685
Arg	0.2229102	0.0000000	0.005988024	0.8931298	1.0000000
Asn	0.1207430	0.1111111	0.225548902	0.9694656	0.6451613
Asp	0.0000000	0.1111111	0.427145709	0.8625954	0.7196030
Cys	0.4582043	0.7777778	0.604790419	0.7671756	0.1736973
Gln	0.1238390	0.1111111	0.347305389	1.0000000	0.7617866

Dendrogram of agnes(x = daneA)



Klasyfikacja to zagadnienie w którym naszym celem jest budowa algorytmu (klasyfikatora), który na podstawie określonych cech przypisze obiekt do jednej z g kategorii.

Również ten dział szybko się rozwija. Począwszy od lat dwudziestych XX wieku (np. Fisher) budowane są nowe algorytmy budowy klasyfikatorów. Do popularniejszych należą: regresja logistyczna, drzewa decyzyjne, metoda wektorów podpierających, sieci neuronowe i wiele innych.

Klasyfikacja ma wiele zastosowań, w tym:

- klasyfikacja do grupy zdrowy-chory,
- identyfikacja z jakim rodzajem nowotworu mamy do czynienia,
- rozpoznawanie numerów rejestracyjnych ze zdjęć tablic samochodowych,
- scoring i rating dla klientów banku.

Przykład: ryzyko wznowy wśród pacjentek chorych na raka piersi

Przykładowy klasyfikator przedstawimy na przykładzie danych zebranych od pacjentek chorych na raka piersi. Dane te zostały zebrane w Dolnośląskim Centrum Onkologii i dotyczą ponad 300 pacjentek leczonych w tym centrum.

Dla każdej z pacjentek zbadano wiele cech, np. informacje o przerzutach do węzłów chłonnych, menopauzie, wieku, poziomie rozmaitych substancji, markerów genetycznych HER, BCRP, Wimentyna i szeregu innych cech. Część pacjentek w analizowanym okresie doświadczyła nawrotu choroby a część pozostała zdrowa (analizujemy 5 letni DFS).

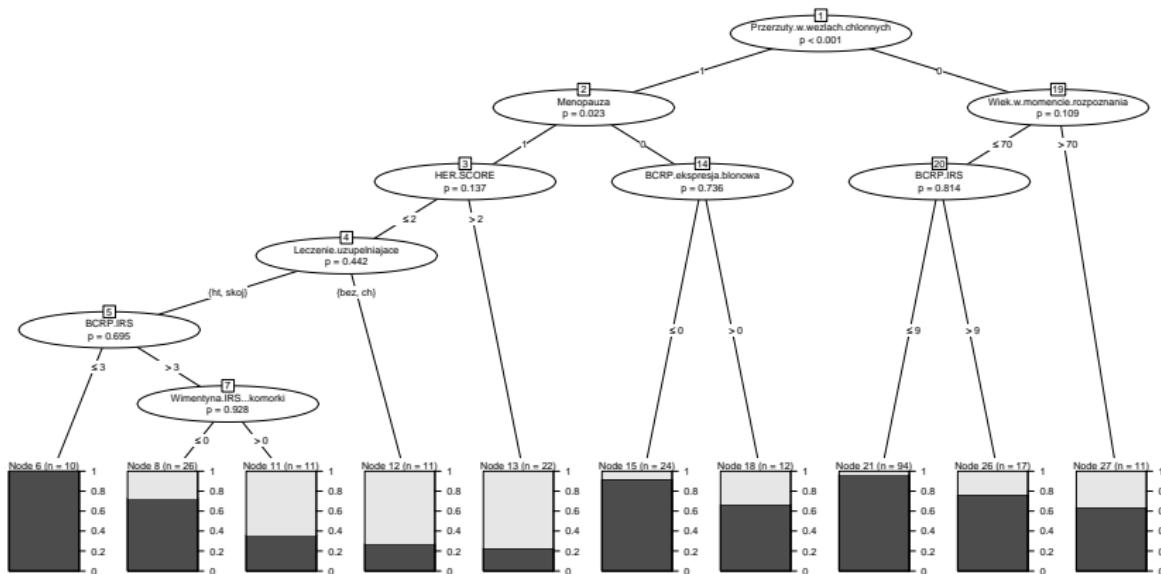
Pytanie na które chcemy odpowiedzieć to jakie czynniki mogą świadczyć o ryzyku wznowy oraz jak dla danej pacjentki ocenić ryzyko nawrotu choroby.

Drzewa decyzyjne to technika konstrukcji klasyfikatora, który można przestawić w postaci drzewa. Liście takiego drzewa odpowiadają klasom a węzły odpowiadają pewnym testom logicznym.

Drzewo konstruowane jest następująco:

- dla obserwacji znajdujących się w danym węźle identyfikujemy cechę, która najbardziej różnicuje badane klasy (używając współczynnika Giniego, entropii lub innej miary),
- jeżeli dana cecha jest ciągła to szukamy punktu podziału, który najsilniej różnicuje badane klasy,
- tworzymy dwa podwęzły, w jednym umieszczamy obserwacje o wartości wybranej cechy mniejszej niż wartość progowa, w drugim pozostałe,
- jeżeli różnorodność w danym podwęźle jest większa niż pewna zadana stała to wracamy do kroku 1.

Drzewo decyzyjne



Drzewa decyzyjne są metodą mało stabilną, której często towarzyszy problem tzw. przeuczenia. Aby poprawić właściwości tego klasyfikatora używa się tzw. metody boosting. Bazuje ona na metodzie bootstrap i pozwala na poprawę stabilności klasyfikatora.

Algorytm konstrukcji i użycia lasu losowego wygląda następująco:

- z próby danych generujemy B (np. 1000) podzbiorów danych, losując ze zwracaniem wiersze oraz bez zwracania cechy,
- na każdym z uzyskanych podzbiorów danych budujemy drzewo decyzyjne, ponieważ podzbiory zawierają średnio tylko około 35% wspólnych obserwacji, uzyskane drzewa różnią się między sobą,
- tak zbudowany las klasyfikuje nowe obserwacje do klasy najczęściej wybieranej przez drzewa z których jest zbudowany.

Po co konstruować las losowy?

Okazuje się że taki komitet klasyfikatorów ma bardzo dobre właściwości.

- Na zbiorach obserwacji, które nie znalazły się w zreplikowanych próbach (tzw. OOB, ang. out of bag) można liczyć nieobciążoną ocenę błędu predykcji.
- Badając jakość drzew w lesie można konstruować ranking ważności zmiennych.
- Można uzupełniać brakujące dane stosując tzw. imputacje.
- Można identyfikować obserwacje odstające i wpływowne.
- Otrzymuje się predyktor o małej wariancji.

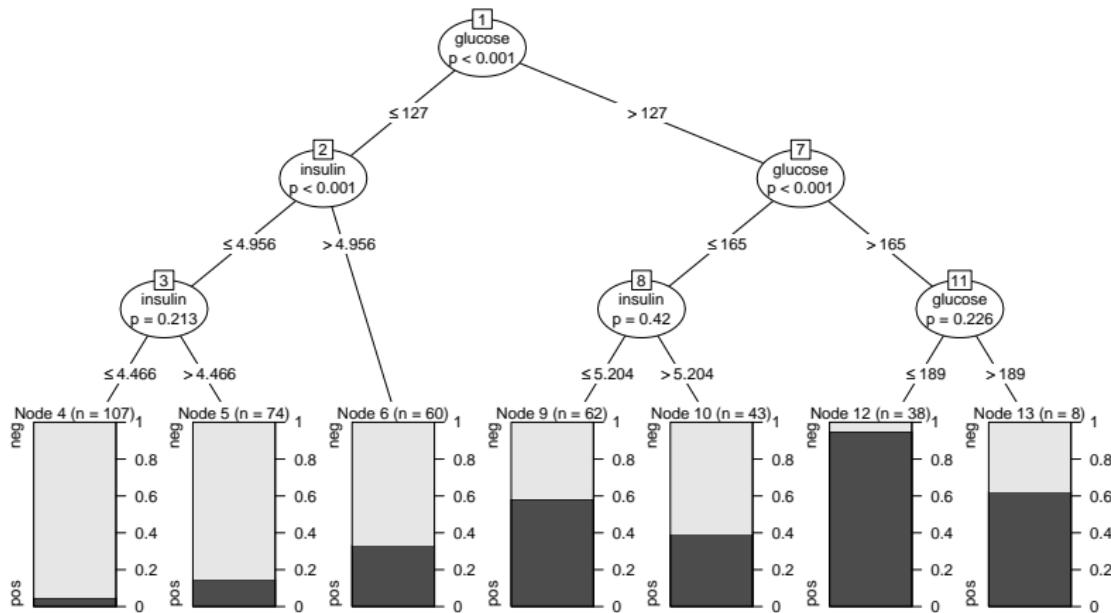
Przykład: Indianki z plemienia Pima

Kolejny przykład dotyczący klasyfikacji dotyczyć będzie badania cukrzycy u Indianek z plemienia Pima.

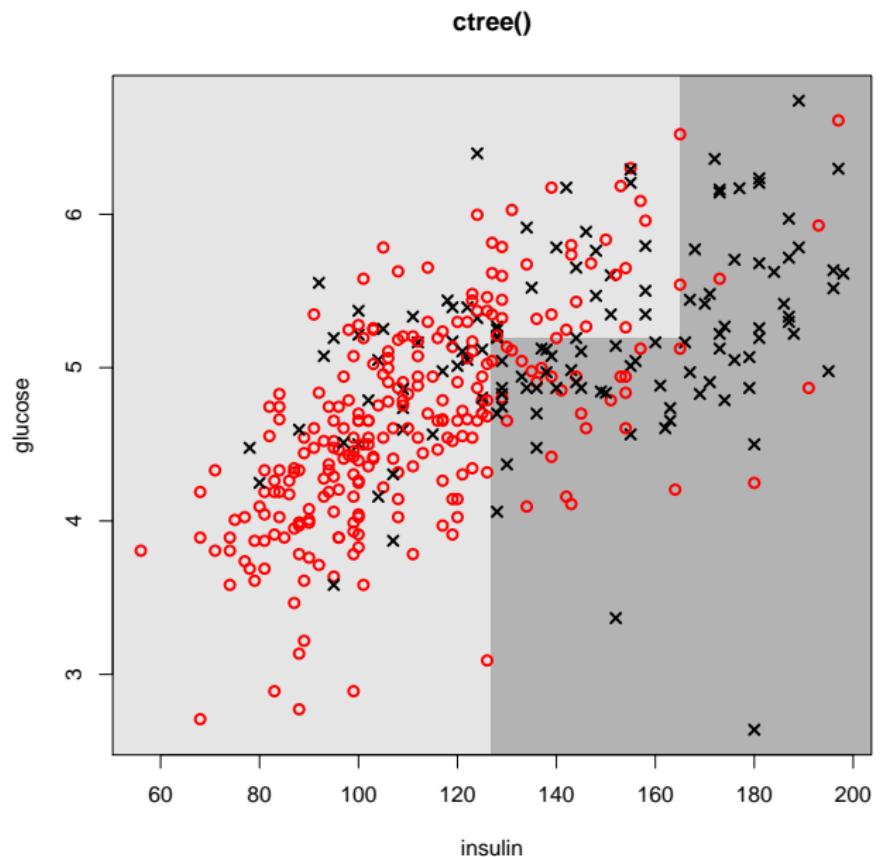
Pytanie: Która cecha lub zestaw cech pozwoli na skuteczną wstępna diagnozę cukrzycy?

	npreg	glu	bp	skin	bmi	ped	age	type
1	5	86	68	28	30.2	0.364	24	No
2	7	195	70	33	25.1	0.163	55	Yes
3	5	77	82	41	35.8	0.156	35	No
4	0	165	76	43	47.9	0.259	26	No
5	0	107	60	25	26.4	0.133	23	No
6	5	97	76	27	35.6	0.378	52	Yes

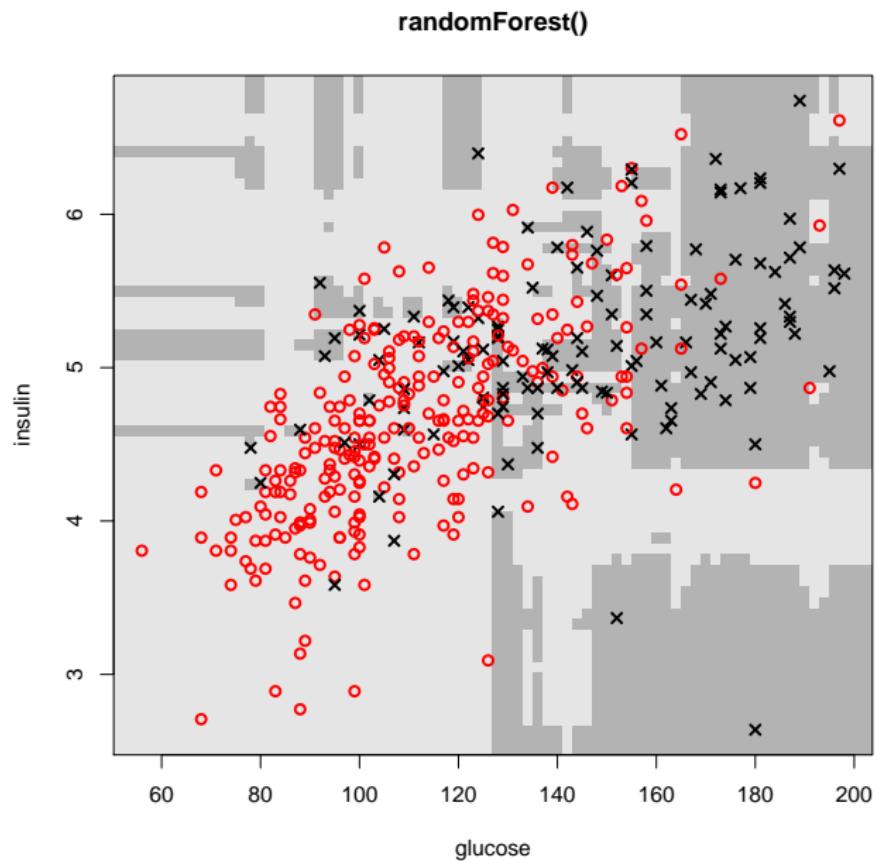
Przykład: Indianki z plemienia Pima



Przykład: Indianki z plemienia Pima, obszar decyzyjny drzewa decyzyjnego

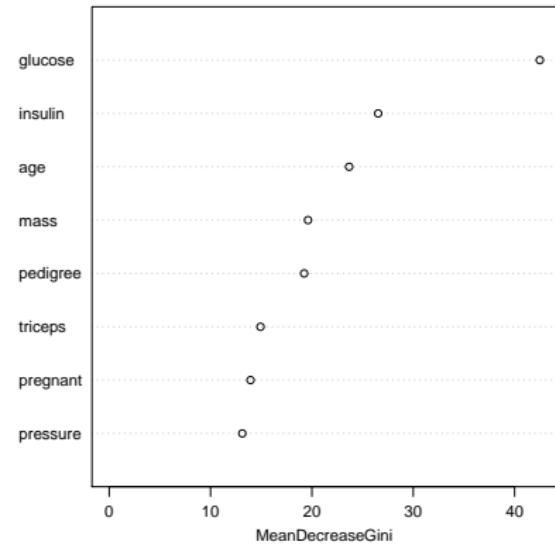
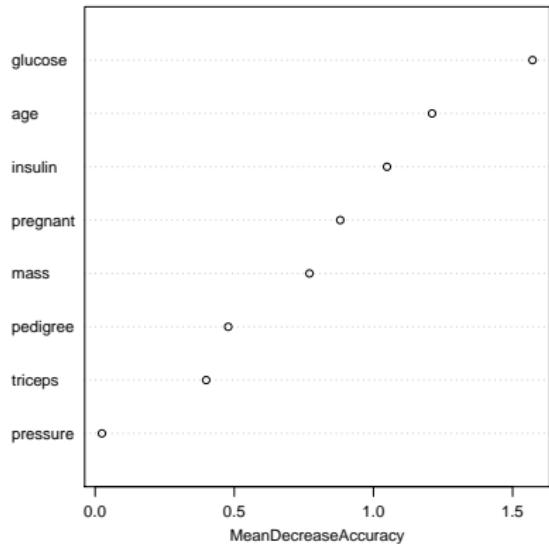


Przykład: Indianki z plemienia Pima, obszar decyzyjny lasu losowego



Przykład: Indianki z plemienia Pima

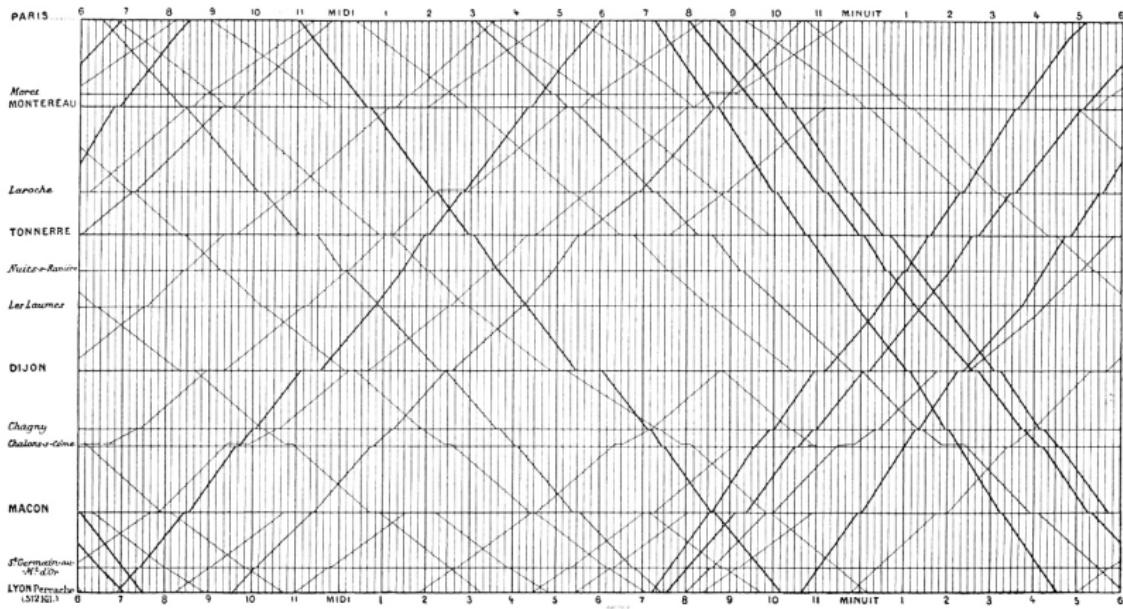
klasyfikatorRF



- Data mining = statystyka - założenia,
ale
data mining + założenia \neq statystyka.
- Przedstawione metody to algorytmy, które niczym w książce kucharskiej zostały zaprojektowane do rozwiązania określonych problemów.
- Algorytmom tym towarzyszą metody badające ich właściwości. Np. dla metod budowy klasyfikatorów potrzebne są metody do oceny błędu klasyfikacji, stabilności algorytmu, badania przeuczenia, diagnostyki modelu itp. W pewnym sensie ta diagnostyka odpowiada twierdzeniom z klasycznej statystyki matematycznej.
- Obecnie data mining jest znacznie szybciej rozwijany na wydziałach informatycznych niż przez „statystyków matematycznych”. Czy to dobrze?

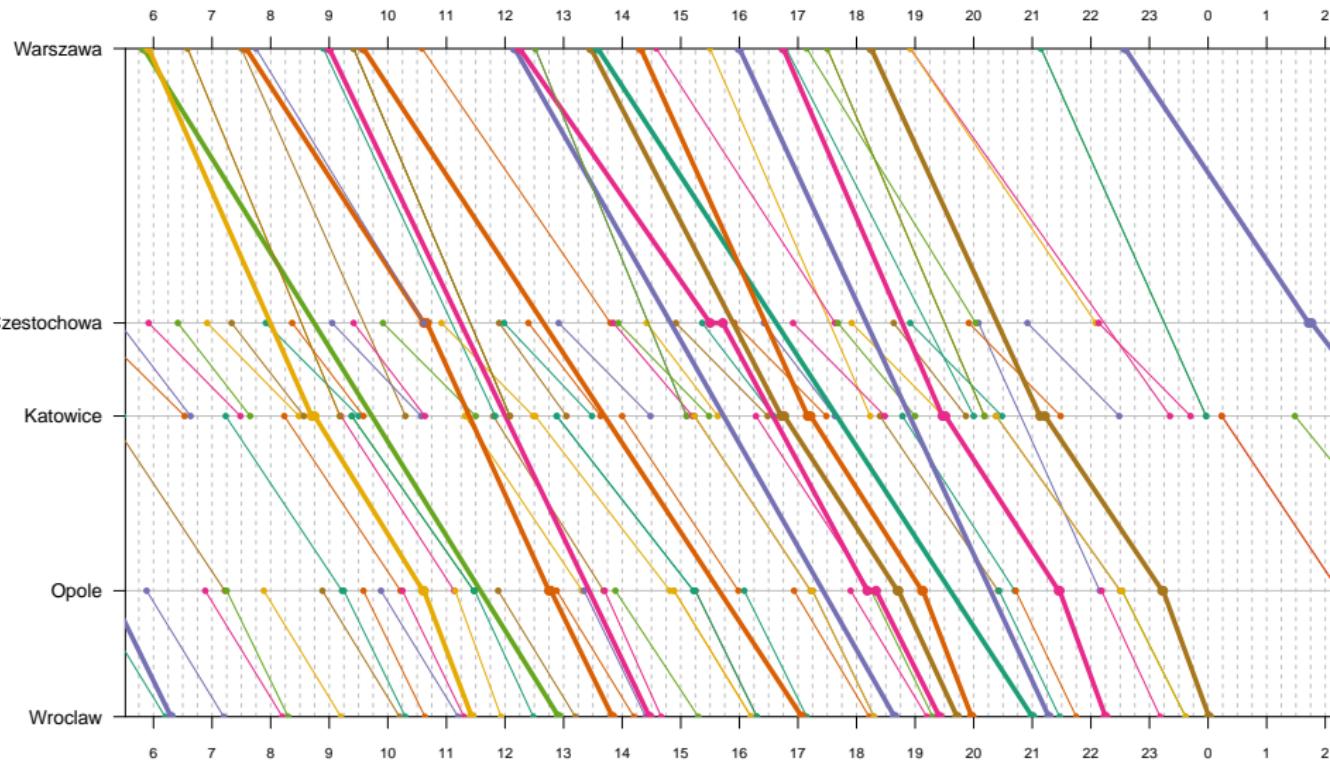
Po co nam wizualizacja? - Prezentacja informacji

Train Schedule by E.J. Marey, <http://c82.net/posts.php?id=66>



Po co nam wizualizacja? - Prezentacja informacji

Opracowanie własne na podstawie danych z serwisu pkp.pl



Prezentacja informacji

Percepcja wzorców obecnych na wykresie odbywa się w trzech krokach (zobacz też: Statistical presentation graphics, Frank Harrell)

- identyfikacja - jakie geometrie kodują prezentowane wartości (kąty, powierzchnie, długości),
- grupowanie - zestawienie listy obiektów prezentujących dane wartości,
- ocena / szacowanie różnic, porządku lub względnych proporcji pomiędzy przedstawianymi geometriami.

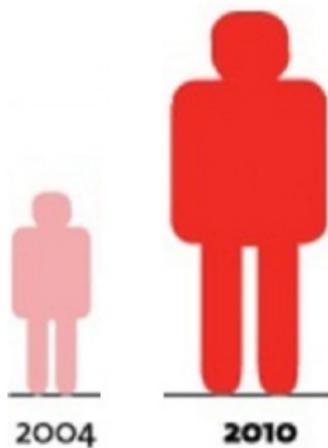
Jeżeli wizualizacja nie jest przemyślana to w każdym z tych kroków coś może pójść źle.

Identyfikacja geometrii

Przykład z <http://biznes.interia.pl/raport/emerytury/news/reforma-reformy-emerytalnej-bo-dane-sie-zdezaktualizowaly>'

Poniższy rysunek porównuje dwie liczby: liczby polaków na emigracji w 2004 i 2010 roku.

Czy można z niego odczytać o ile % liczba polaków na emigracji wyrosła?
Która właściwość „ludzika” przedstawia liczbę osób na emigracji?



Identyfikacja geometrii

Przykład z <http://biznes.interia.pl/raport/emerytury/news/reforma-reformy-emerytalnej-bo-dane-sie-zdezaktualizowaly>'

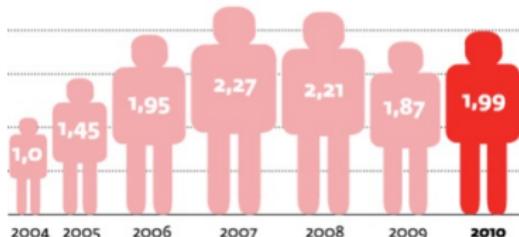
Pole czy wysokość?

Nasza percepja lepiej ocenia różnice w wysokości. Intuicyjnie jednak te obrazki różnią się polami.



Liczba emigrantów przebywających poza granicami Polski

Dane w mln



źródło: GUS

[Z serwisu interia.pl, usunięto liczby i linie poziome siatki
<http://biznes.interia.pl/news/reforma-reformy-emerytalnej-bo-dane-sie-zdezaktualizowaly,1771747,4265>]

Grupowanie obiektów

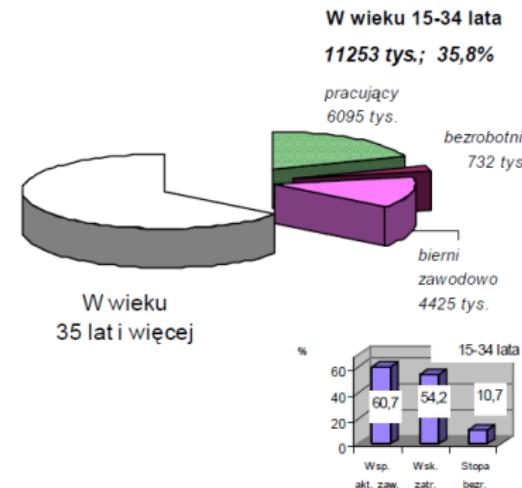
Przykład z raportu „Wejście ludzi młodych na rynek pracy”, GUS 2010

Na dobrym wykresie powinno być oczywiste, który element odpowiada jakiej grupie.

Aktywność ekonomiczna¹ ludności w wieku 15-34 lata

Zbiorowość osób młodych, zdefiniowanych tu jako ludność w wieku 15-34 lata, liczyła w II kwartale 2009 r. 11253 tys., co stanowiło nieco ponad 1/3 globalnych zasobów w pracy (w wieku 15 lat i więcej). W badanym okresie nieco częściej niż **co druga młoda osoba pracowała** – 6095 tys. (wskaźnik zatrudnienia – 54,2%), a kryteria **bezrobotnego** wg MOP spełniała częściej niż co piętnasta - 735 tys. Łącznie, trzy z pięciu młodych osób były aktywne zawodowo, tj. pracowały lub aktywnie poszukiwały pracy (współczynnik aktywności zawodowej - 60,7%).

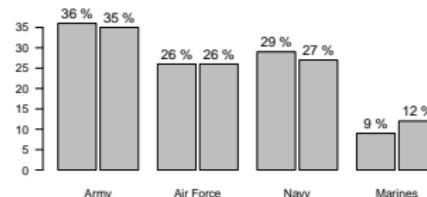
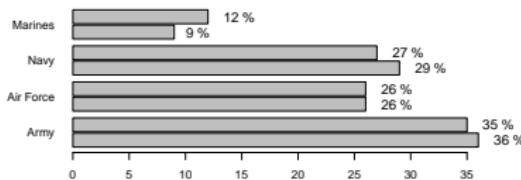
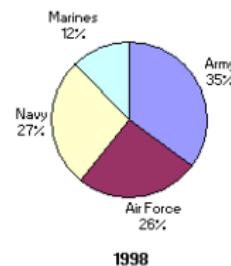
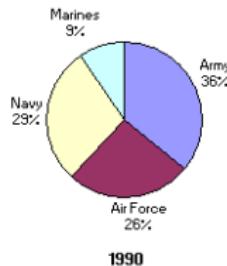
Ludność w wieku 15 lat i więcej,
w tym 15-34 lata według aktywności ekonomicznej
- II kw. 2009 r.



Ocena proporcji

Przykład z <http://lilt.ilstu.edu/gmklass/pos138/datadisplay/badchart.htm>

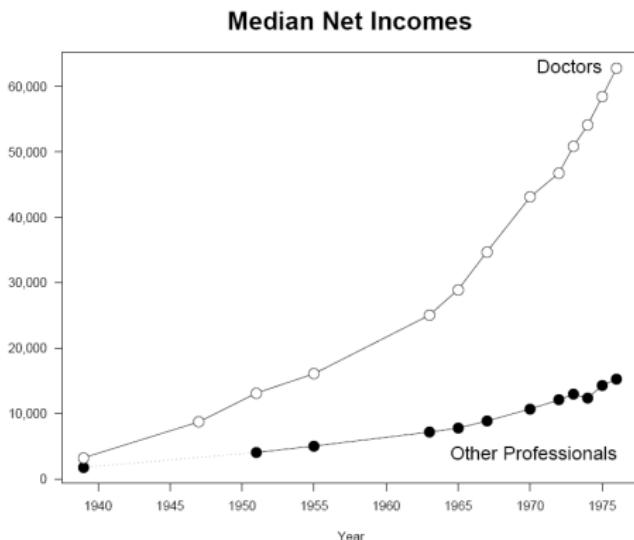
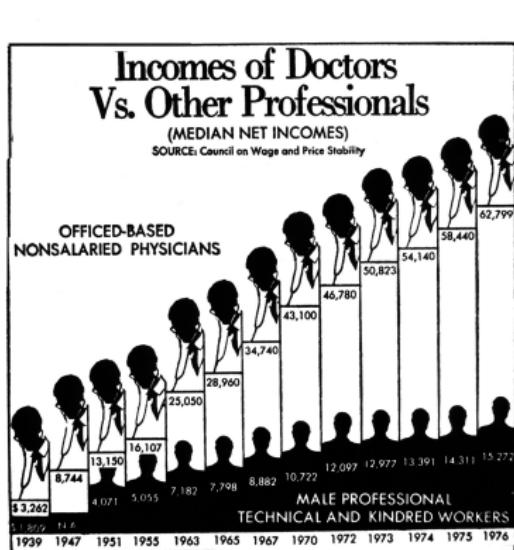
Na dobrym wykresie różnice i podobieństwa powinny być łatwe do oceny.



Zniekształcanie wykresu: gumowe osie OX

Przykład z „The Visual Display of Quantitative Information”. E. Tufte

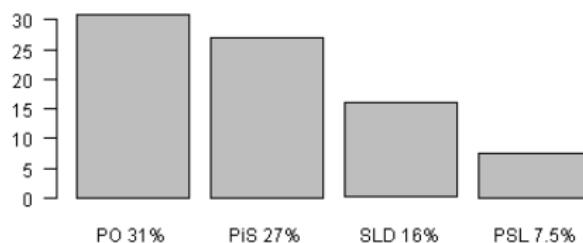
Zniekształcanie może być wynikiem zarówno nieumyślnego błędu jak i celowej manipulacji.



Zniekształcanie wykresu: gumowe osie OY

Przykład z serwisu <http://www.szczecinek.pl>

Nawet jeżeli wizualizacji towarzyszą liczby, jeżeli te same wartości przedstawione są w sposób i graficzny i liczbowy to pierwsze wrażenie dotyczące charakteru zależności oparte jest zazwyczaj o grafikę.

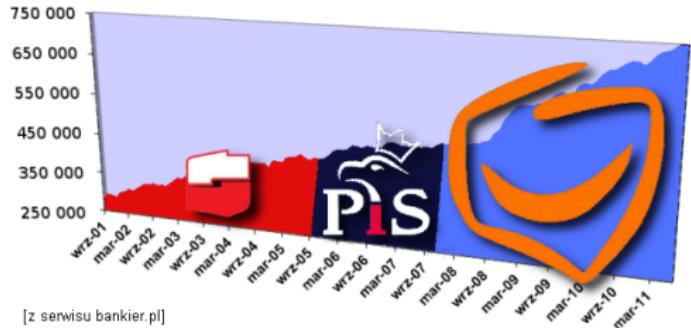
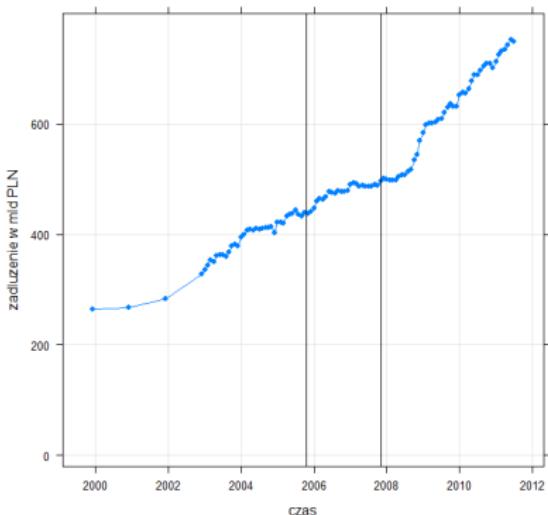


[Z serwisu szczecinek.pl]

Zniekształcanie wykresu: obroty

Przykład z serwisu <http://www.bankier.pl>

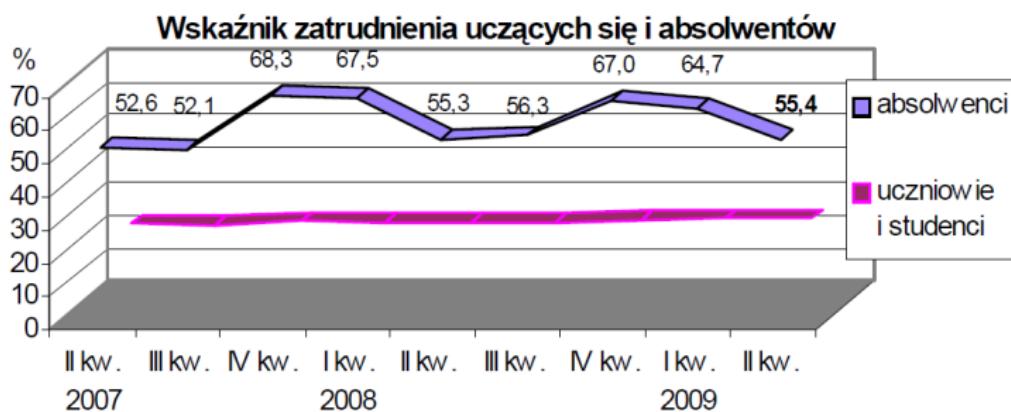
Obroty wykresu to szczególnie zły pomysł, gdy chcemy porównywać nachylenia krzywych lub wysokości punktów.



Zniekształcanie wykresu: pseudo perspektywa

Przykład z raportu „Wejście ludzi młodych na rynek pracy”, GUS 2010

Dodawanie perspektywy bardzo rzadko jest dobrym pomysłem. Czasem perspektywa uniemożliwia odczytanie informacji z wykresu.

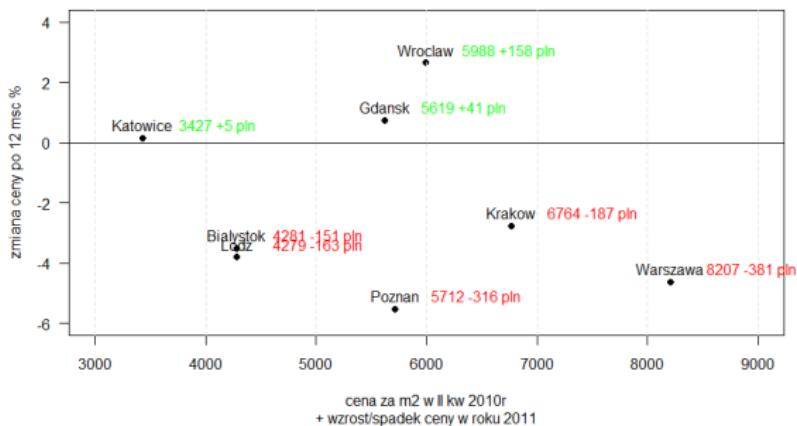
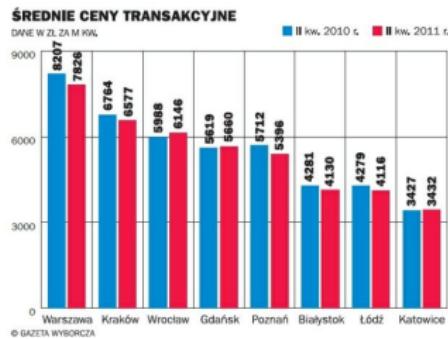


Zniekształcanie wykresu: kolejność obiektów

Przykład z serwisu <http://www.gazeta.pl>

Porządkowanie obiektów to dobry pomysł, gdy chcemy przedstawić ranking wartości.

Ale zły pomysł gdy chcemy przedstawić różnice pomiędzy wartościami.



Zniekształcanie wykresu: perspektywa zmienia kąty

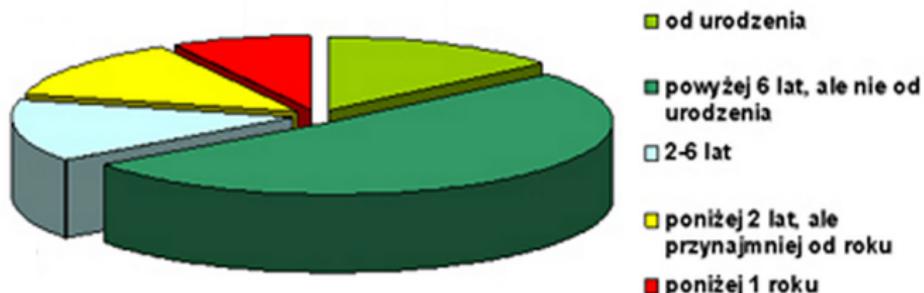
Przykład z serwisu dzielnicy warszawa Bemowo

Wykres kołowy przedstawia liczby za pomocą kątów.

Rzuty „pseudo 3D” zmieniają kąty. Więc połączenie perspektywy i wykresu kołowego to bardzo zły pomysł.

Mocne i słabe strony życia na Bemowie

OKRES ZAMIESZKIWANIA NA BEMOWIE



Zniekształcanie wykresu: perspektywa zmienia kąty

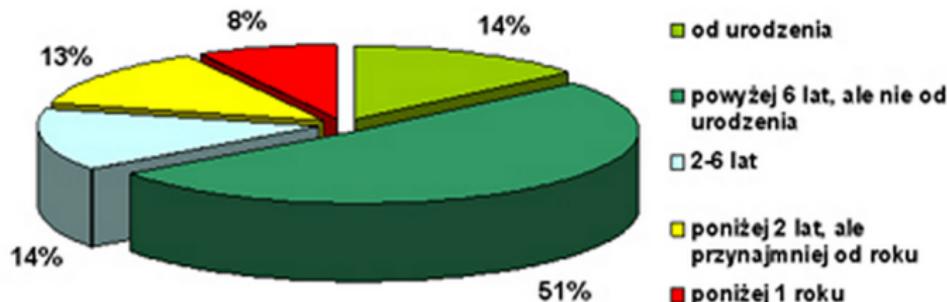
Przykład z serwisu dzielnicy Warszawa Bemowo

Wykres kołowy przedstawia liczby za pomocą kątów.

Rzuty „pseudo 3D” zmieniają kąty. Więc połączenie perspektywy i wykresu kołowego to bardzo zły pomysł.

Mocne i słabe strony życia na Bemowie

OKRES ZAMIESZKIWANIA NA BEMOWIE

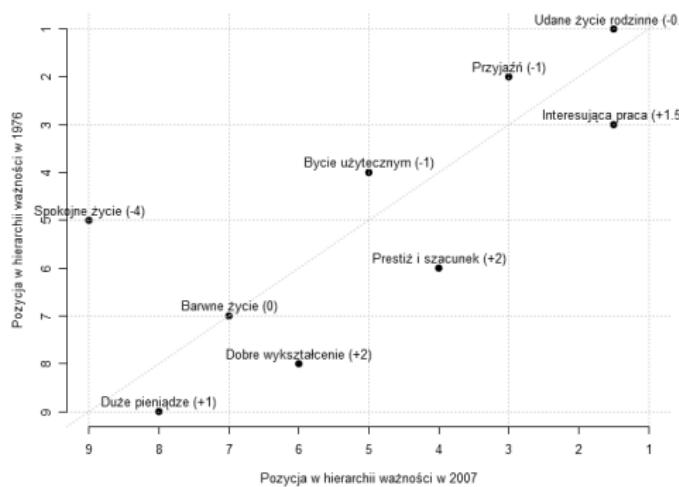


Zła identyfikacja charakterystyki do oceny

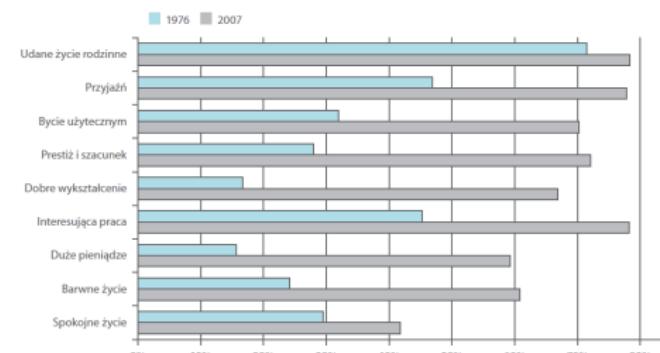
Przykład z raportu MAC „Młodzi 2011”

Nawet jeżeli wykres przedstawia dane poprawnie, wciąż pozostaje kwestia odczytania wykresu.

Z poniższych danych można mieć inne wnioski patrząc na ranking wartości a coś innego patrząc na wartości bezwzględne.



Rys.2.1. Co jest w życiu ważne? Odpowiedzi 19-letniej młodzieży w 1976 i 2007



Źródło: Badania warszawsko-kieleckie S. Nowaka (lata 70. XX w.), badania własne. „Poruszanie” – ścieżki edukacyjne i wchodzenie w dorosłość (N = 1096).

[Raport Młodzi 2011, ministerstwo MAC strona 38]

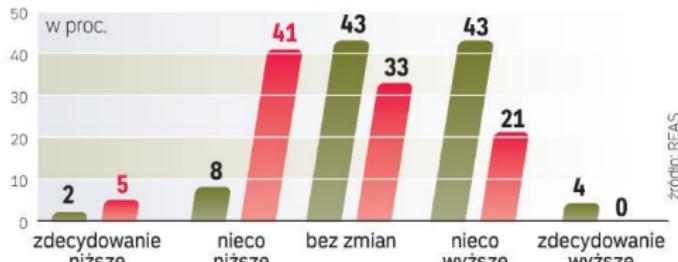
Nie wszystkie charakterystyki są sobie równe

Na przykładzie <http://www.rp.pl/galeria/8,2,641431.html>

Rodzaj wykresu powinien uwzględniać rodzaj prezentowanych zmiennych.

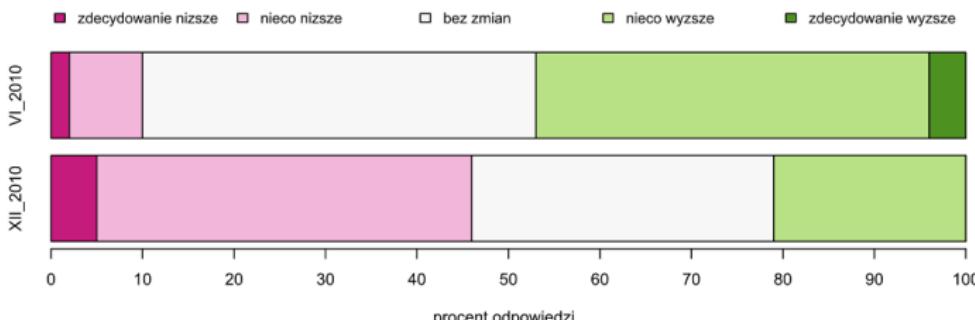
Jak zmienią się ceny mieszkań według deweloperów w ciągu 12 miesięcy

■ badanie z VI 2010 r. ■ badanie z XII 2010 r.



źródło: REAS

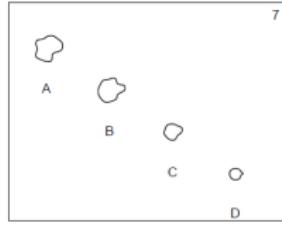
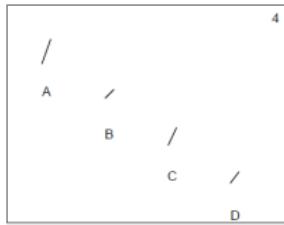
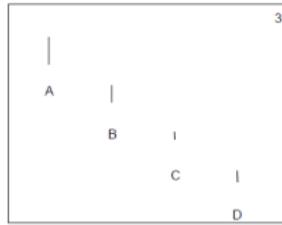
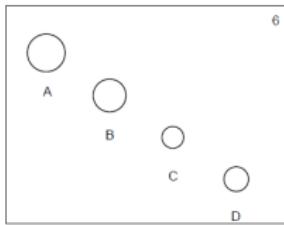
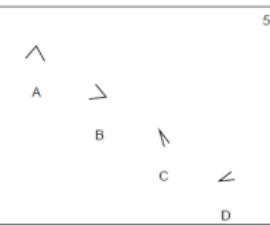
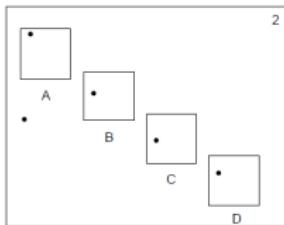
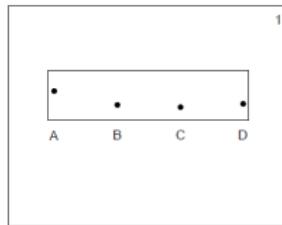
• 46 PROC. FIRM UWAŻA, ŻE CENY MIESZKAŃ SPADNAŁY



Nie wszystkie charakterystyki są sobie równe

Eksperyment Cleveland and McGill, na podstawie „Information Visualisation”, Ross Ihaka

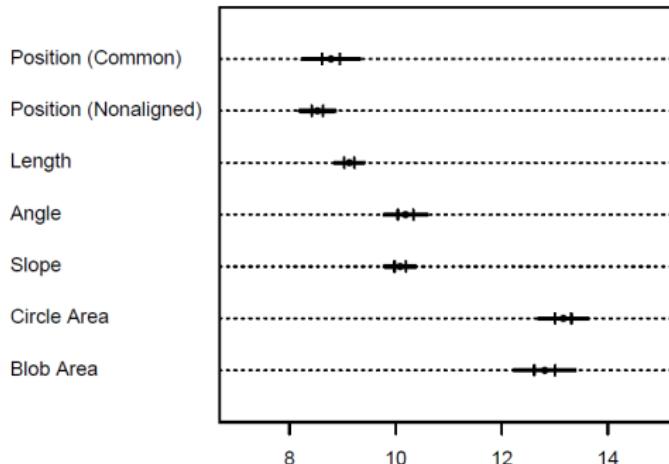
W łatwy sposób można zbadać jak dokładnie ludzie są w stanie oceniać różnice pomiędzy długościami, polami, kolorami, kątami itp.



Nie wszystkie charakterystyki są sobie równe

Eksperyment Cleveland and McGill, na podstawie „Information Visualisation”, Ross Ihaka

W łatwy sposób można zbadać jak dokładnie ludzie są w stanie oceniać różnice pomiędzy długościami, polami, kolorami, kątami itp.



Prezentacja informacji

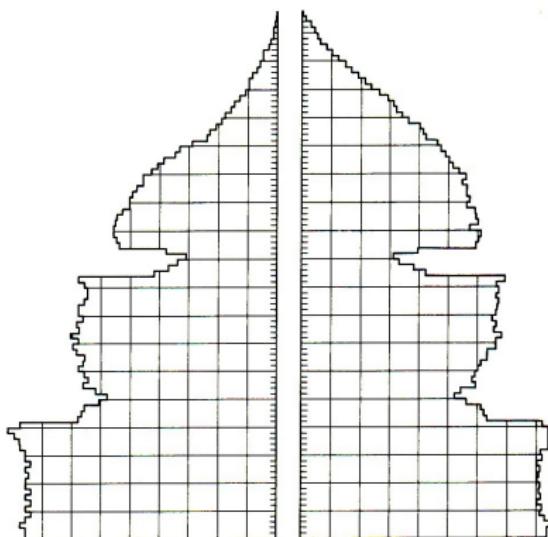
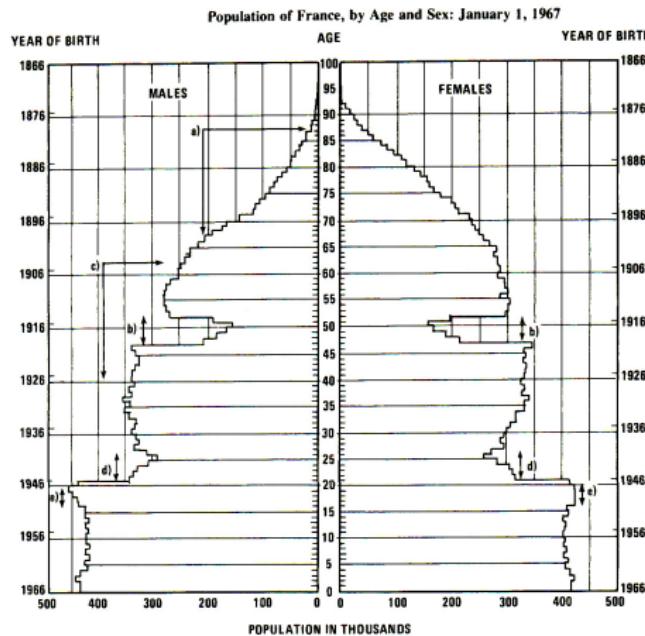
Dobre rady wujków Edwarda Tufte'go i Rosa Ihaka

- Maksymalizuj współczynnik dane / atrament (ang. data - ink ratio).
- Minimalizuj współczynnik przekłamania (ang. lie - factor).
- Proste dane przedstawiaj prosto (ang. If the „story” is simple, keep it simple).
- Złożone dane przedstawiaj w czytelny i łatwo do interpretacji sposób (ang. If the „story” is complex, make it look simple).

Współczynnik dane / atrament

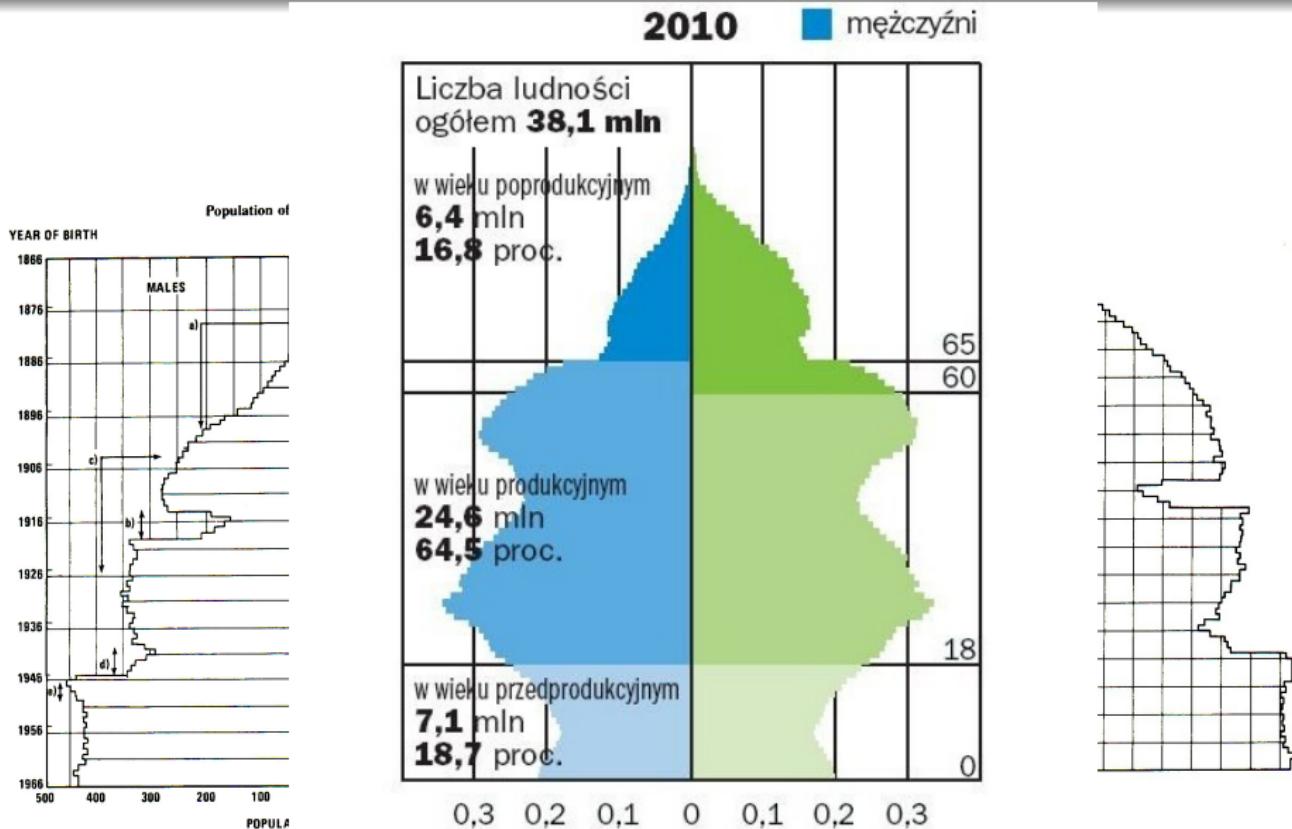
Przykład z „The Visual Display of Quantitative Information”. E. Tufte

Należy unikać rysowania elementów nie niosących dodatkowej informacji o danych. Istotne elementy nie powinny być „zalane” pomocniczymi ozdobnikami.



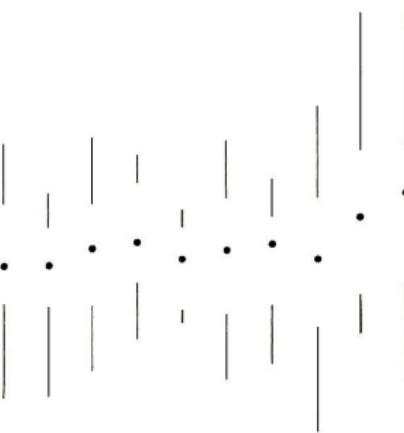
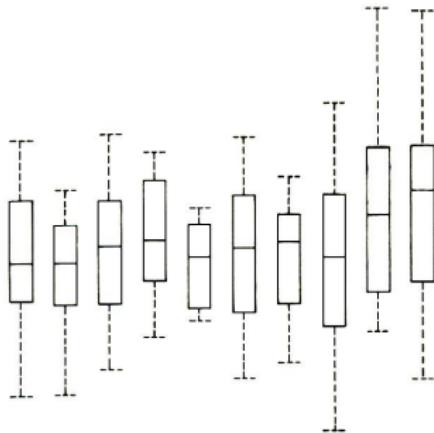
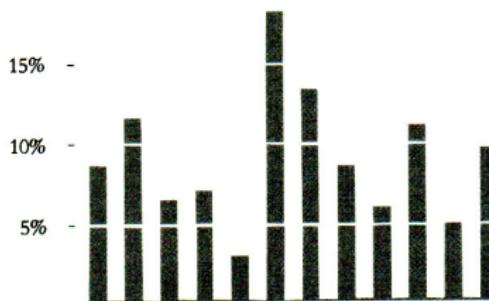
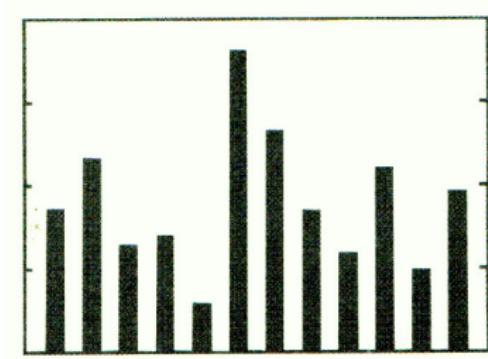
Współczynnik dane / atrament

Przykład z „The Visual Display of Quantitative Information”. E. Tufte



Współczynnik dane / atrament

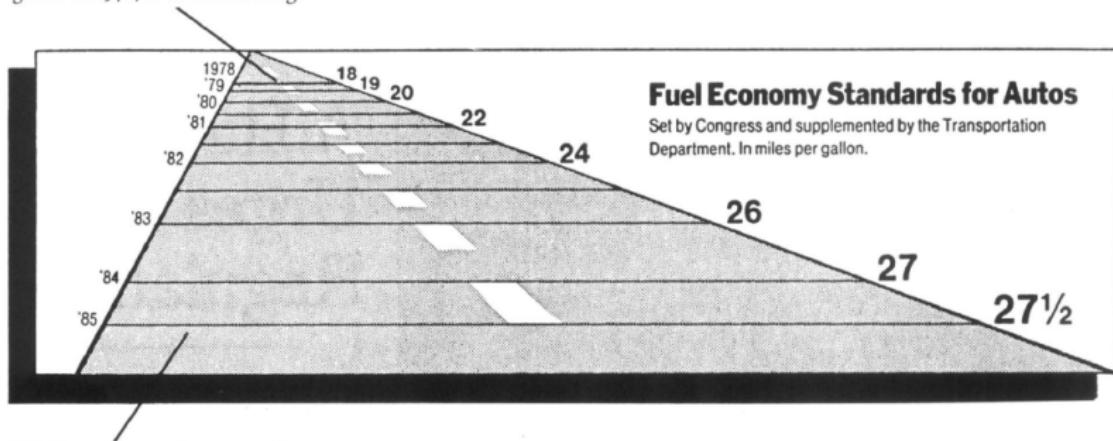
Przykład z „The Visual Display of Quantitative Information”. E. Tufte



Współczynnik przekłamania (lie - factor)

Przykład z „The Visual Display of Quantitative Information”. E. Tufte

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.



This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

$$\text{Data Effect} = \frac{27.5 - 18}{18} = 0.53,$$

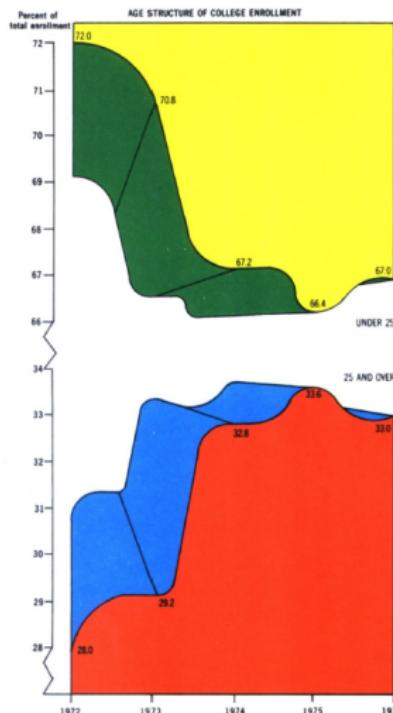
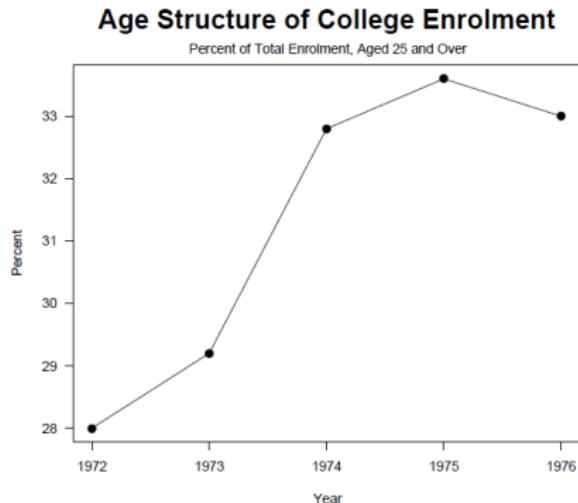
$$\text{Graph Effect} = \frac{5.3 - .6}{.6} = 7.83,$$

$$\text{Lie Factor} = 14.8$$

Proste dane przedstawiaj prosto

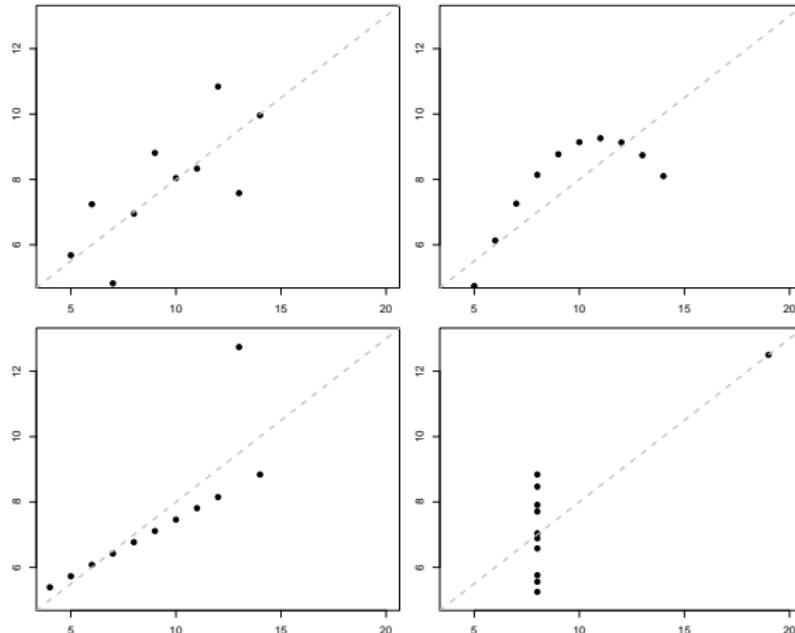
Przykład z „Information Visualisation”, Ross Ihaka

Ile kolorów i elementów potrzeba by przedstawić 5 liczb?



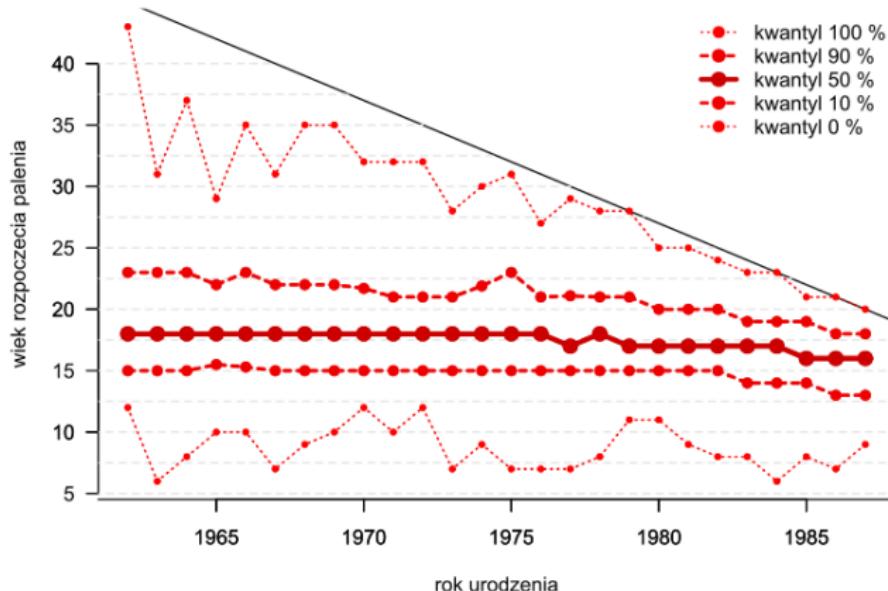
Wizualizacja - jest potrzebna

Nie wszystkie wykresy są złe, a nawet jeżeli są złe to czasem wciąż są lepsze niż ich brak. Oglądając grafiki często łatwiej nam zrozumieć charakter zależności niż patrząc na surowe liczby.



Wizualizacja pozwala na wyjaśnienie rzeczywistości

Czy to zdanie wyjaśnia coś nt. wieku palenia:
Średni wiek rozpoczęcia palenia to 17.7 ± 2.3 lat.

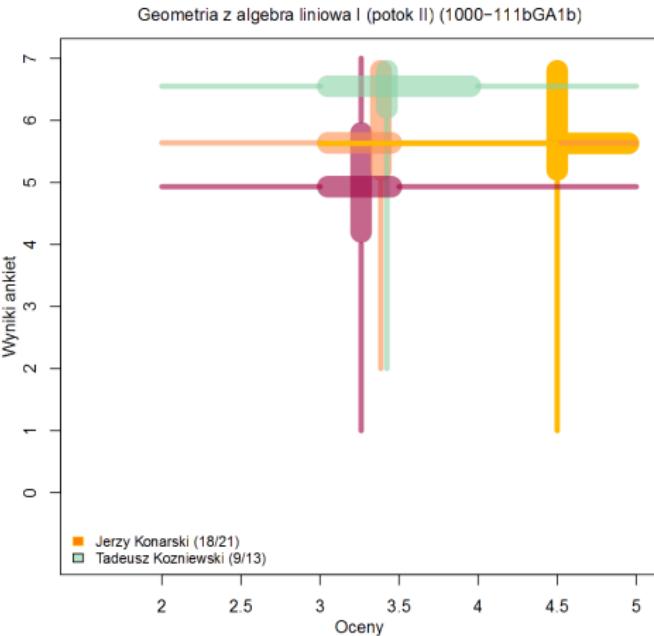
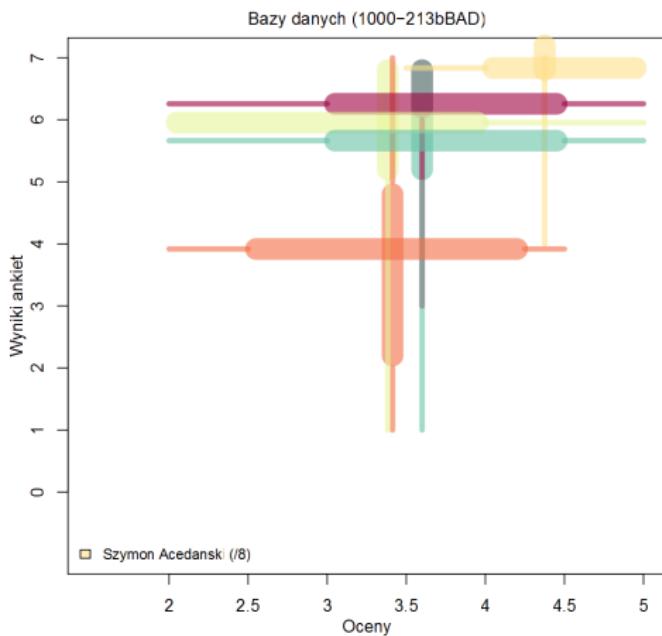


Wizualizacja bardziej przyciąga uwagę niż tabela liczb

Wizualizacja danych z systemu uniwersyteckiego USOS

Czy lepiej wybrać miłego prowadzącego czy efektywnego?

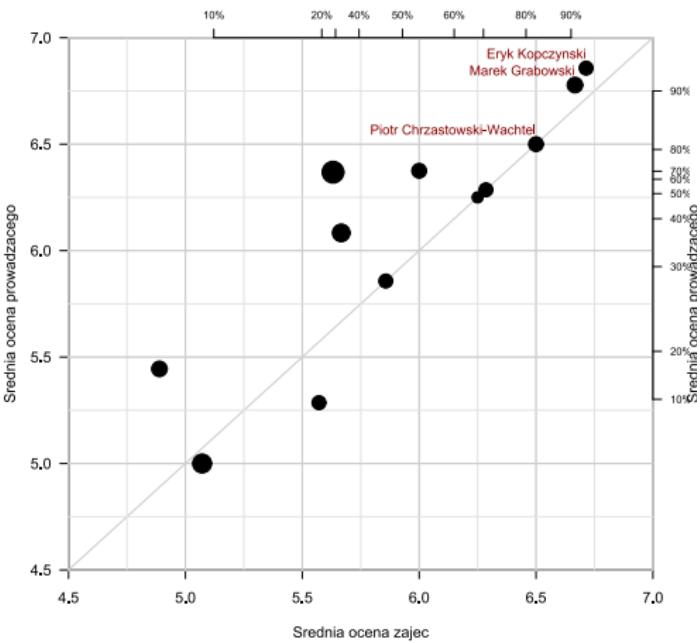
Czy mili prowadzący są bardziej efektywni od tych niemiłych?



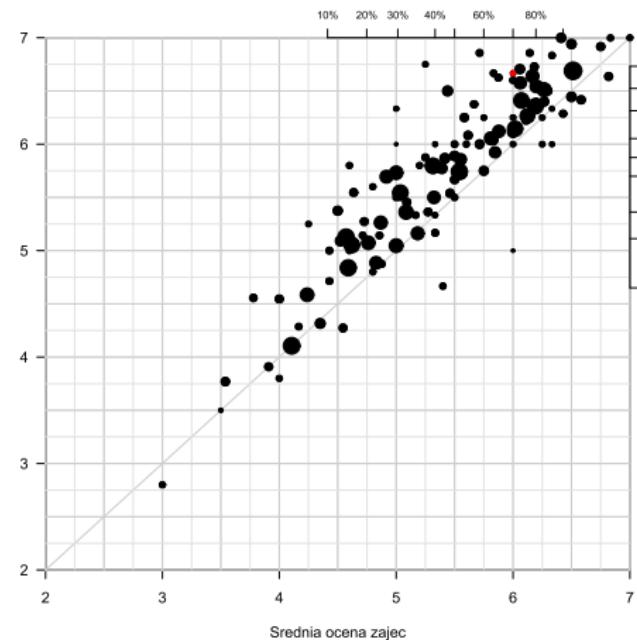
Czy wybieramy przedmiot czy prowadzącego?

Którego ćwiczeniowca wybrać do Wstępu do programowania?

Wstęp do programowania, ćwiczenia i laboratoria, semestr 2010Z



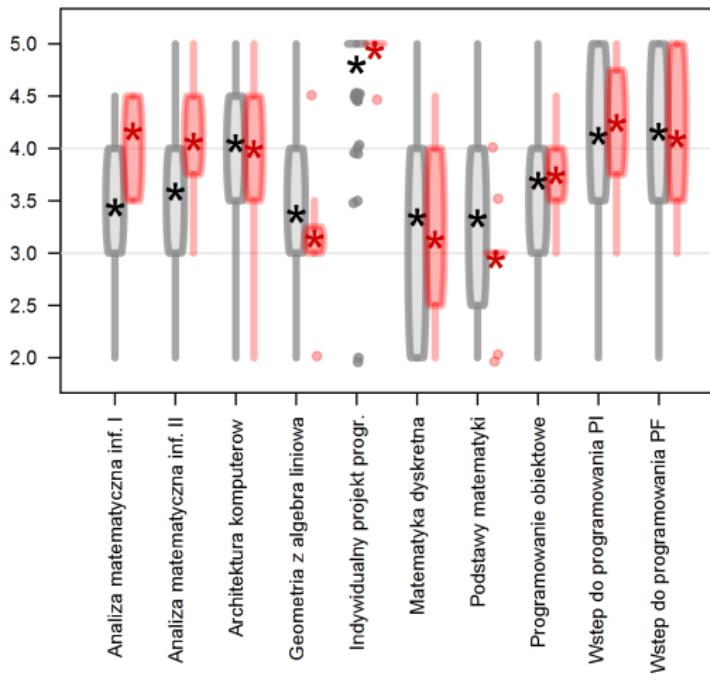
Wykłady, semestr 2010Z



Profile seminariów magisterkich

Wybór seminarium magisterskiego nie zawsze jest prosty. Opinie kolegów są często obciążone subiektywnymi doświadczeniami. Pewne informacje, które mogą pomóc w tym wyborze znajdują się w USOSie, trzeba je tylko wyciągnąć i przedstawić.

Wybrane aspekty inżynierii oprogramowania (11)



Pytania do części pierwszej

- ① Podaj przykład zastosowania skalowania wielowymiarowego.
- ② Podaj przykład zastosowania analizy dyskryminacyjnej.
- ③ Podaj przykład zastosowania analizy skupisk.
- ④ Czym różni się las losowy od drzewa klasyfikacyjnego?
- ⑤ Zaproponuj użyteczne zastosowanie jednego z przedstawionych algorytmów do danych z systemu USOS.

Pytania do części drugiej

- ① Co mierzy współczynnik „lie-factor” ?
- ② Co mierzy współczynnik „data-ink” ?
- ③ Uporządkuj te trzy właściwości: długość, kąt, objętość w kolejności od tego który można najprecyzyjniej porównać do najmniej precyzyjnego.
- ④ Wymień trzy ozdobniki/zniekształcienia często dodawane do wykresów a utrudniające precyzyjne odczytanie wartości.
- ⑤ Zaproponuj użyteczne podsumowanie danych dostępnych w systemie USOS.