ZAAWANSOWANE METODY EKSPLORACJI

INFORMACJE WSTĘPNE

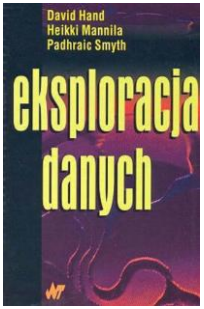
Prof. dr hab. inż. Grzegorz Dudek
Wydział Matematyki i Informatyki
Uniwersytet Łódzki

PLAN WYKŁADU

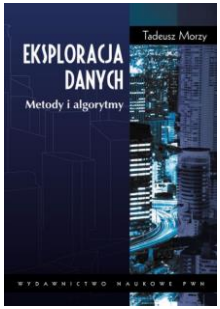
1. Informacje wstępne *
2. Podstawy uczenia maszynowego
3. Uczenie się indukcyjne
4. Przekształcanie, selekcja i ekstrakcja atrybutów
5. Wykrywanie obserwacji odstających i uzupełnianie brakujących danych
6. Grupowanie danych
7. Drzewa decyzyjne do klasyfikacji
8. Drzewa decyzyjne do regresji
9. Liniowe metody klasyfikacji
10. Sieci neuronowe do regresji
11. Sieci neuronowe do klasyfikacji
12. Klasyfikator Bayesa
13. Klasyfikatory minimalnoodległościowe

* Ten wykład oparty jest na materiałach dostępnych w internecie:
http://home.agh.edu.pl/~pszwed/wiki/doku.php?id=med:wyklady, http://wazniak.mimuw.edu.pl/index.php?title=Eksploracja_danych
Racka K: Metody eksploracji danych i ich zastosowanie. Zeszyty Naukowe PWSZ w Płocku, Nauki Ekonomiczne, t. XXI, 2015.
Mirończuk M.: Przegląd i klasyfikacja zastosowań, metod oraz technik eksploracji danych. Studia i Materiały Informatyki Stosowanej 2(2), str. 35-46, 2010

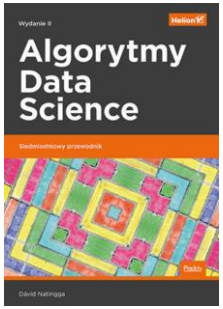
LITERATURA

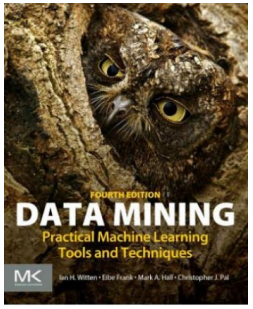

David Hand
Heikki Mannila
Padhraic Smyth
eksploracja danych

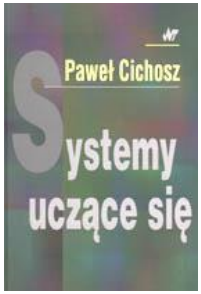

data mining
INFORMATYKA I ZASTOSOWANIA
Daniel T. Larose
Metody i modele eksploracji danych
WYDAWNICTWO NAUKOWE PWN


EKSPLORACJA DANYCH
Tadeusz Morzy
Metody i algorytmy
WYDAWNICTWO NAUKOWE PWN


Metody i narzędzia eksploracji danych
Stanisław Osowski

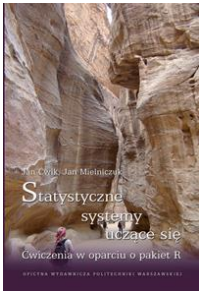

Algorytmy Data Science
David Natingga

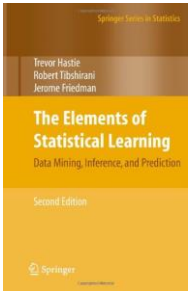

FOURTH EDITION
DATA MINING
Practical Machine Learning Tools and Techniques


Paweł Cichosz
Systemy uczące się


Jacek Koronacki
Jan Ćwik
statystyczne systemy uczące się


Systemy uczące się
rozpoznawanie wzorców
analiza skupień
i redukcja wymiarowości


Statystyczne systemy uczące się
Ćwiczenia w oparciu o pakiet R


Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer


MACHINE LEARNING
TOM M. MITCHELL

DEFINICJE

Eksploracja danych (*data mining*, odkrywanie wiedzy w bazach danych, zgłębianie danych), to proces odkrywania nowych reguł, wzorców i zależności w zbiorach danych.

Cel – zasadniczym celem eksploracji danych jest wydobycie wiedzy z danych. Wiedza ta pozwala zrozumieć różne zjawiska i procesy, wspomaga procesy podejmowania decyzji.

Narzędzia – statystyka, uczenie maszynowe

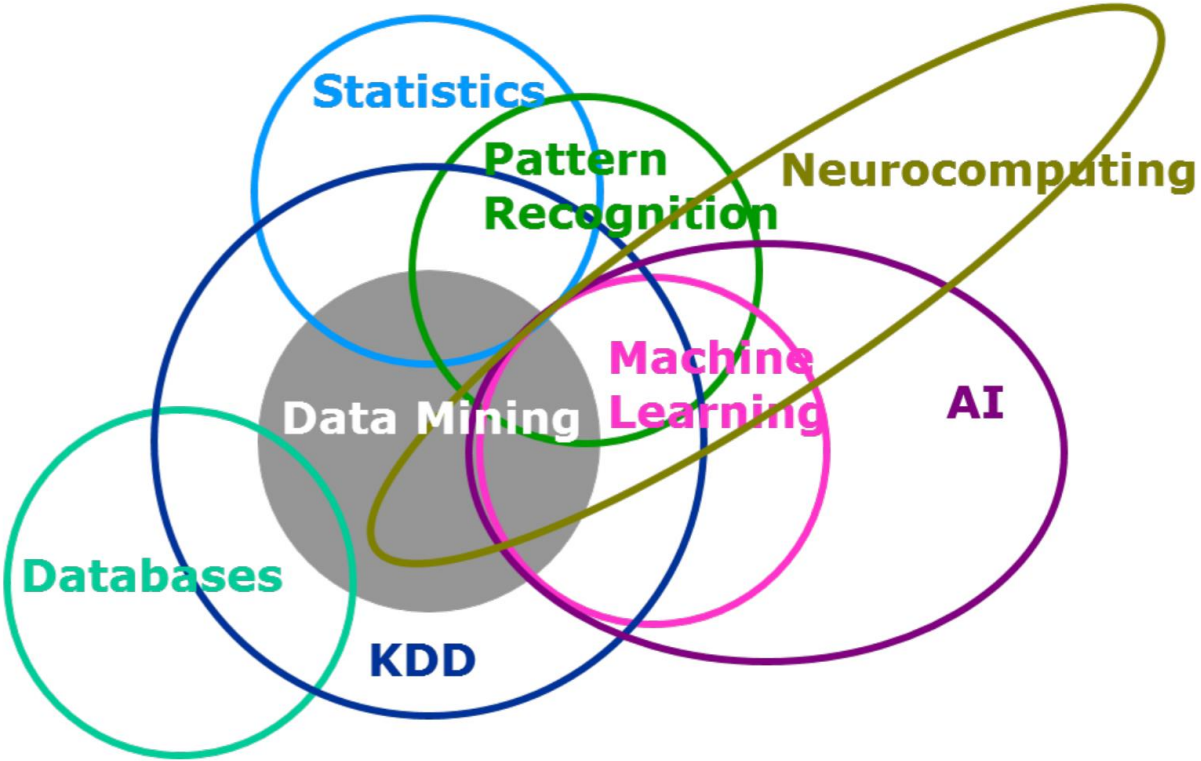
Eksploracja danych jest dziedziną multidyscyplinarną, która skupia wokół siebie wiele dziedzin związanych z przechowywaniem, przetwarzaniem i analizowaniem danych, wydobywaniem z nich wiedzy i jej wdrażaniem. W kontekście zastosowań biznesowych stosuje się też nazwę *business intelligence*.

ETAPY ODKRYWANIA WIEDZY

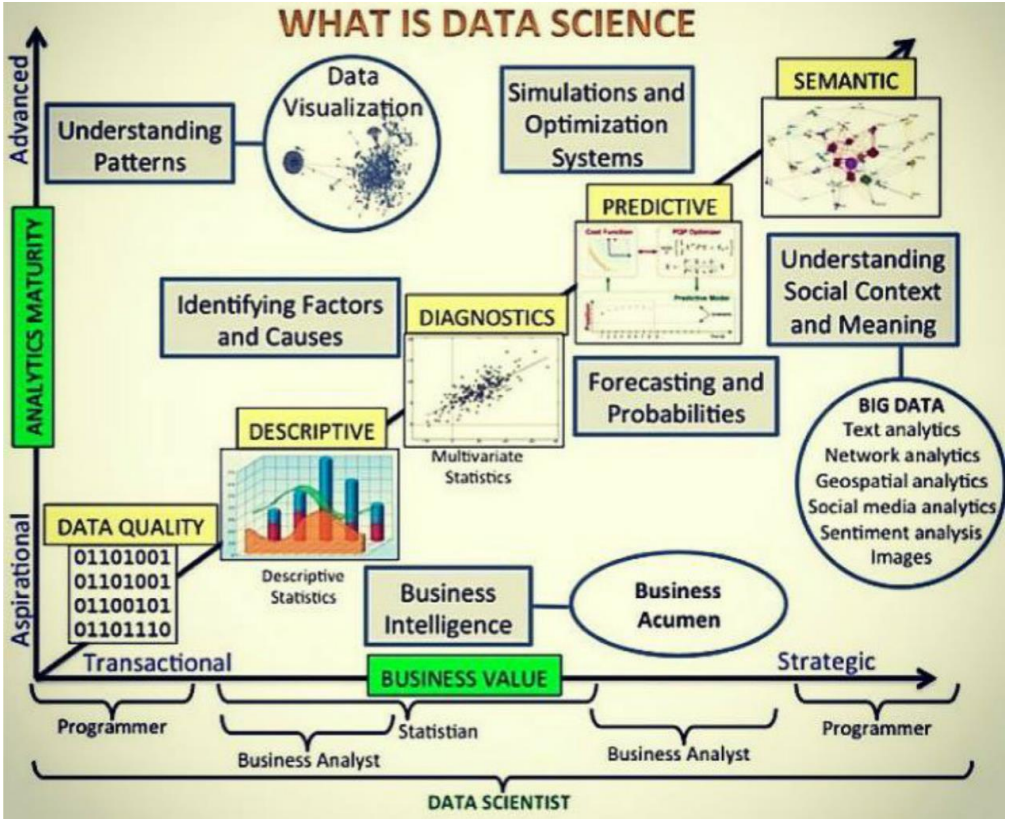
Eksploracja danych jest jednym z etapów procesu odkrywania wiedzy z baz danych (*Knowledge Discovery in Databases*, KDD), który składa się z następujących kroków:

1. Czyszczenie danych (*data cleaning*) – usuwanie błędów, danych nadmiarowych, powtórzonych
2. Integracja danych – łączenie danych pochodzących z różnych źródeł (baz danych), posiadających różną strukturę oraz różne modele danych
3. Selekcja danych – z bazy danych pobierane są dane do przeprowadzenia analiz
4. Transformacja danych – przetwarzanie lub łączenie danych w formach odpowiednich dla eksploracji
5. Eksploracja danych – stosowanie metod eksploracji danych w celu wydobycia z danych wzorców, reguł, zależności
6. Ocena odkrytych wzorców, reguł, zależności – identyfikacja najbardziej interesujących wzorców
7. Prezentacja odkrytej wiedzy – przedstawienie odkrytej wiedzy użytkownikowi za pomocą technik wizualizacji i reprezentacji danych

DATA MINING NA TLE INNYCH DYSCYPLIN



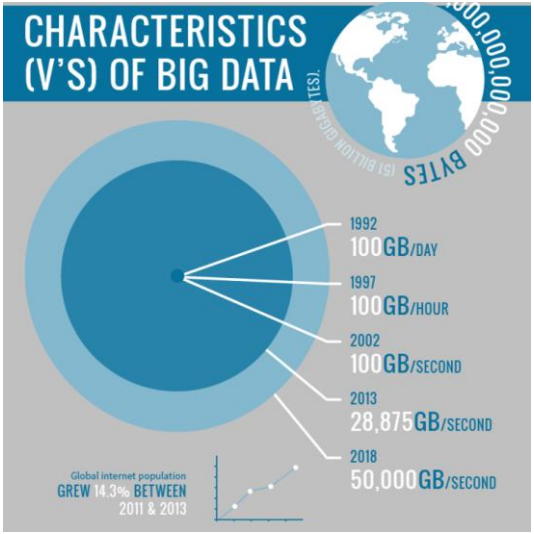
DATA SCIENCE



ŹRÓDŁA DANYCH

Codziennie wytwarzane jest około $2.5 \cdot 10^{18}$ bajtów danych

- dane operacyjne firm i instytucji (banki, ubezpieczalnie, przemysł, sieci handlowe, …)
- aktywność internetowa (e-handel,
- dane z urządzeń mobilnych
- pomiary z rozproszonych czujników
- dane eksperymentalne (fizyczne, biologiczne, astronomiczne, …)
- …



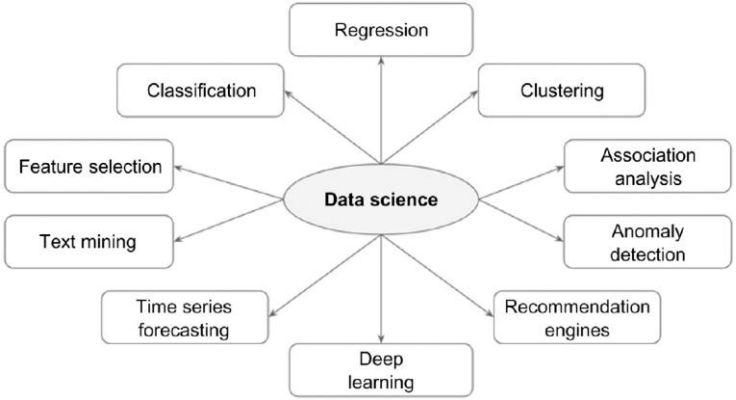
PRZYKŁAD

Przykład wykorzystania eksploracji danych: pozyskiwanie informacji na temat klientów sieci handlowej

- Jakie inne jeszcze produkty, najczęściej, kupują klienci, którzy kupują wino?
- Czym różnią się koszyki klientów kupujących wino i piwo?
- W jaki sposób można scharakteryzować klientów kupujących wino?
- W jaki sposób pogrupować klientów kupujących wino?
- Czy można dokonać predykcji, że dany klient kupi wino?

ZADANIA EKSPLORACJI DANYCH

- klasyfikacja
- regresja
- grupowanie
- odkrywanie sekwencji
- odkrywanie charakterystyk
- analiza przebiegów czasowych
- odkrywanie asocjacji
- eksploracja stron WWW, sieci społecznościowych
- eksploracja tekstów
- eksploracja obrazów i wideo
- eksploracja danych dźwiękowych
- wykrywanie anomalii, zmian i odchyleń
- rekomendowanie
- …

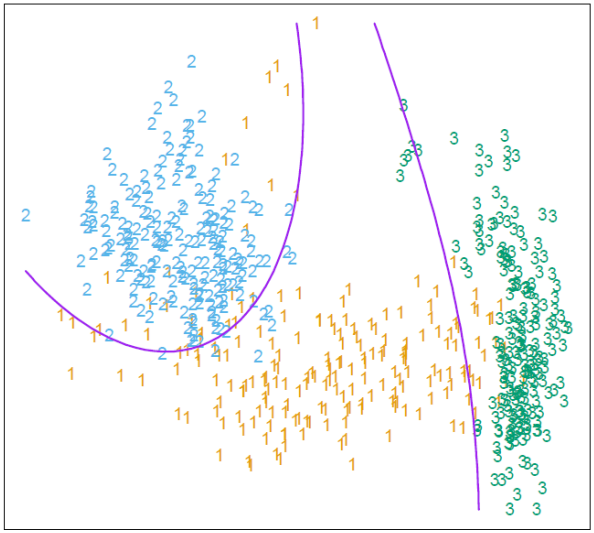


Klasyfikacja

Klasyfikacja polega na zaklasyfikowaniu obserwacji (obiektu) do pewnej klasy.

Metoda klasyfikacji (klasyfikator) przypisuje klasę obserwacji na podstawie charakterystyk klas i reguł klasyfikacji.

Charakterystyki klas powstają w procesie uczenia się systemu klasyfikującego na pewnym zbiorze danych sklasyfikowanych.

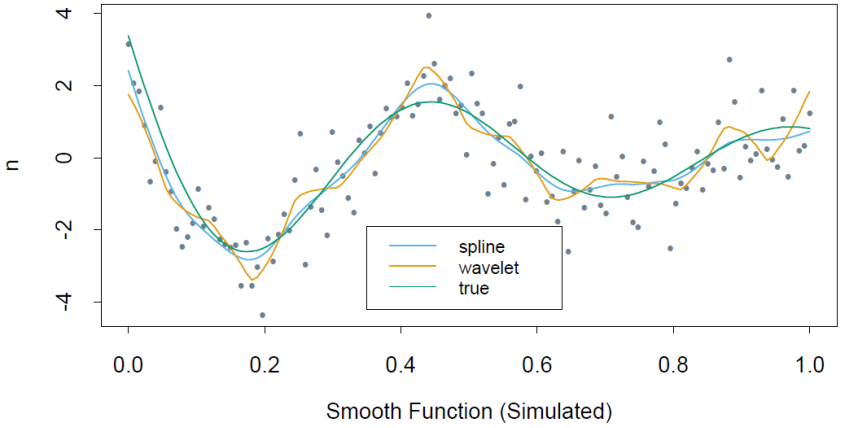


Regresja

Regresja (aproksymacja funkcji) polega na przedstawieniu pewnej funkcji f(x) w innej, zazwyczaj prostszej postaci h(x).

Metoda regresji przypisuje obserwacji x wartość funkcji h(x).

Funkcja aproksymująca powstaje w procesie uczenia się modelu na pewnym zbiorze danych reprezentujących funkcję f(x).

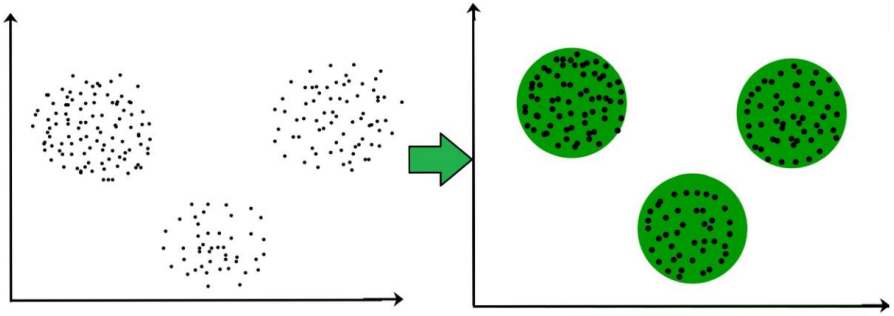


Grupowanie

Podział danych na grupy tak, aby wewnątrz każdej grupy znalazły się obserwacje podobne.

Metoda grupowania dokonuje podziału na grupy na podstawie podobieństwa pomiędzy obserwacjami.

Charakterystyki grup powstają w procesie uczenia się systemu na pewnym zbiorze danych.



Odkrywanie sekwencji (*sequential pattern mining, string mining*)

Odkrywanie wzorców w danych sekwencyjnych.
Odkrywanie wzorców zachowań, na podstawie analizy danych zmieniających się w czasie.

Przykłady zastosowania:

- Odkrywanie wzorców zachowań użytkowników korzystających z Internetu.
- Badanie notowań akcji i odkrywanie wzorców w celu ustalenia modelu decyzyjnego dla strategii inwestycyjnych.
- Odkrywanie wzorców sekwencji w DNA.

Odkrywanie charakterystyk

Odkrywanie charakterystyk polega na znajdowaniu zwięzłych opisów (charakterystyk) podanego zbioru danych.

Przykłady zastosowania:

- określanie powszechnych symptomów danej choroby
  Przykładowo, symptomy określonej choroby mogą być charakteryzowane przez zbiór reguł charakteryzujących (np. pacjenci chorujący na COVID-19 cechują się temperaturą ciała większa niż 37.5 C, suchym kaszlem, zmęczeniem, utratą smaku lub węchu)
- Określenie profilu klienta
- znajdowanie zależności funkcyjnych pomiędzy zmiennymi

Analiza szeregów czasowych

Wykrywanie trendów, wahań okresowych, właściwości szeregów, prognozowanie.

Przykłady zastosowania:

- Analiza i prognoza zapotrzebowania na dobra
- Badanie dynamiki zjawisk masowych
- Diagnostyka predykcyjna w przemyśle
- Analizy giełdowe
- Badanie rozwoju gospodarczego
- Analiza i predykcja szeregów czasowych w meteorologii

Odkrywanie asocjacji

Badanie współwystępowania wartość (wariantów) różnych zmiennych. Wyniki analizy tego typu mają postać reguł asocjacyjnych postaci jeżeli A, to B.

Pierwotnym zastosowaniem była analiza koszykowa: badanie współzależności pomiędzy kupowanymi produktami.

Przykłady zastosowania:

- Analiza usług pod katem zwiększenia sprzedaży
- Optymalizacja pakietów usług, opłat i taryf w sektorze finansowym, telekomunikacyjnym, itp.
- Planowanie kampanii promocyjnych
- Weryfikacja skuteczności kampanii marketingowych

Eksploracja danych tekstowych (*text mining*)

Metody eksploracji danych służące analizie treści dokumentów tekstowych w celu znalezienia nowych informacji, które nie są dostępne bezpośrednio.

Przykłady zastosowania:

- Wykrywanie plagiatów
- Klasyfikacja dokumentów tekstowych np. poczty internetowej - oddzielenie informacji ważnych od nieistotnych
- Grupowanie danych tekstowych np. artykułów ze względu na tematy i treść
- Analiza treści zamieszczanych na portalach społecznościowych

Eksploracja obrazów

Wydobywania wiedzy poprzez odkrywanie relacji między obrazami, czy też wzorów ukrytych (niejawnie) w obrazach (np. rozpoznawanie obiektów na obrazach.

Dziedzina ta wykorzystuje metody pochodzące z widzenia komputerowego (ang. Computer vision), przetwarzania obrazu, odzyskiwania obrazu, eksploracji danych, uczenia maszynowego, baz danych i sztucznej inteligencji.

Eksploracja danych wideo

Odkrywanie wzorów w zawartościach baz multimedialnych przechowujących dane wideo.

Cele:

- wykrywanie przyczyn zarejestrowanych zdarzeń
- wykrywanie zdarzeń niepożądanych, np. pojazdów wjeżdżających na teren chroniony, ludzi zachowujących się nietypowo
- określanie typowych i nieprawidłowych zachowań,
- klasyfikacja zachowania do wybranej kategorii np. chodzenie, bieganie, skakanie
- grupowanie i określanie interakcji pomiędzy obiektami.

Przedmiotem zainteresowania statystyki opisowej są przede wszystkim:

- miary położenia (m.in. tendencji centralnej): np. średnia, dominanta (moda), kwantyle.
- miary dyspersji (rozproszenia, zróżnicowania): np. wariancja, odchylenie standardowe, odchylenie przeciętne.
- miary asymetrii (skośności): np. klasyczny/pozycyjny współczynnik asymetrii.

Miary średnie

Średnia arytmetyczna liczb $x_1, \ldots, x_n$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Średnia arytmetyczna ważona liczb $x_1, \ldots, x_n$ z wagami $a_1, \ldots, a_n$, gdzie $a_i \geqslant 0$ dla każdego $i = 1, \ldots, n$ oraz $\sum_{i=1}^{n} a_i = 1$

$$\bar{x} = \sum_{i=1}^{n} a_i x_i.$$

Średnia geometryczna liczb dodatnich $x_1, \ldots, x_n$

$$\bar{g} = \sqrt[n]{x_1 \cdots x_n}.$$

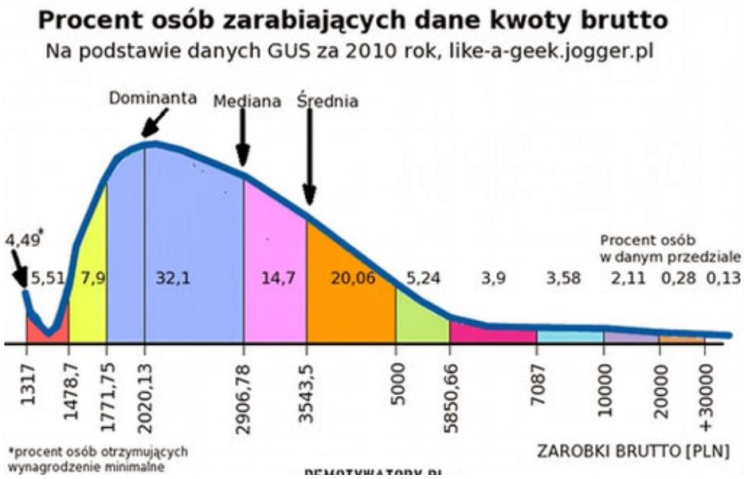
Średnia harmoniczna liczb $x_1, \ldots, x_n$ różnych od zera

$$\bar{h} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

Mediana (wartość środkowa) m_e dla danych niezgrupowanych

$$m_e = \begin{cases} x_{(\frac{n+1}{2})}, & n - \text{nieparzyste}, \\ \frac{x_{(\frac{n}{2})}+x_{(\frac{n}{2}+1)}}{2}, & n - \text{parzyste}. \end{cases}$$

Moda (dominanta) m_o dla danych niezgrupowanych – wartość, która pojawia się najczęściej.



65% Polaków zarabia poniżej średniej krajowej

Miary zmienności

Rozstęp $r = x_{\max} - x_{\min}$.

Kwantyle – wartości danej cechy, które dzielą ją na określone części pod względem liczby jednostek. Dane muszą być uporządkowane niemalejąco.

Kwartyle dla danych niezgrupowanych

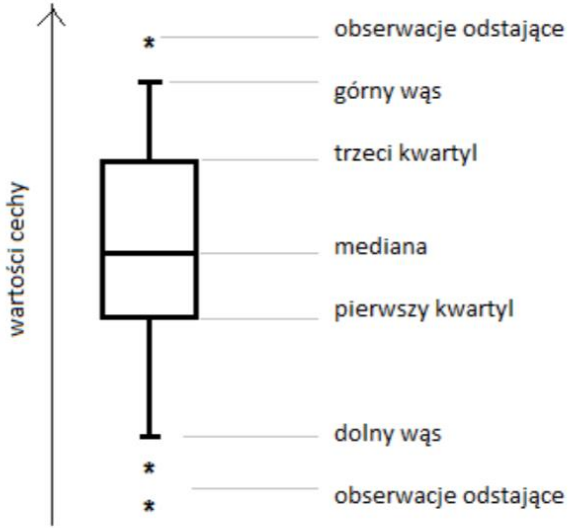
- *Kwartyl pierwszy (dolny)* Q_1 – dzieli dane tak, że $1/4$ jednostek ma wartości niższe lub równe, a $3/4$ jednostek ma wartości wyższe lub równe niż kwartyl. (Mediana „lewej połowy danych”.)
- *Kwartyl drugi (środkowy)* Q_2 to mediana. Dzieli dane tak, że $1/2$ jednostek ma wartości niższe lub równe i $1/2$ jednostek ma wartości wyższe lub równe niż kwartyl.
- *Kwartyl trzeci (górny)* Q_3 – dzieli dane tak, że $3/4$ jednostek ma wartości niższe lub równe, a $1/4$ jednostek ma wartości wyższe lub równe niż kwartyl. (Mediana „prawej połowy danych”.)

Rozstęp międzykwartylowy $Q_3 - Q_1$.

Odchylenie ćwiartkowe (rozstęp międzykwartylowy połówkowy) $Q = \frac{Q_3 - Q_1}{2}$.

Typowy obszar zmienności $(m_e - Q, m_e + Q)$.

Wykres pudełkowy



Długość każdego z wąsów jest równa $1{,}5IQR$, chyba, że:

- wartość maksymalna jest mniejsza niż $Q_3 + 1{,}5IQR$
- wartość minimalna jest większa niż $Q_1 - 1{,}5IQR$,

W takim przypadku długość wąsa jest zdeterminowana przez odpowiednio wartość maksymalną lub minimalną. Obserwacje znajdujące się poza 3 rozstępami IQR to obserwacje odstające.

Odchylenie przeciętne dla danych niezgrupowanych

$$d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

Wariancja dla danych niezgrupowanych

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Odchylenie standardowe

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Typowy obszar zmienności $(\bar{x} - s, \bar{x} + s)$.

Współczynnik zmienności

Miara pozycyjna

$$v = \frac{Q}{m_e}, \quad \text{gdzie } Q \text{ – odchylenie ćwiartkowe}$$

Miary klasyczne

$$v = \frac{d}{\bar{x}}, \quad \text{gdzie } d \text{ – odchylenie przeciętne}$$

$$v = \frac{s}{\bar{x}}$$

Rozkład normalny

Zwany także rozkładem Gaussowskim. Parametry: średnia $\mu \in \mathbb{R}$, wariancja $\sigma^2 \in \mathbb{R}_+$, z dodatnim pierwiastkiem (odchylenie standardowe).

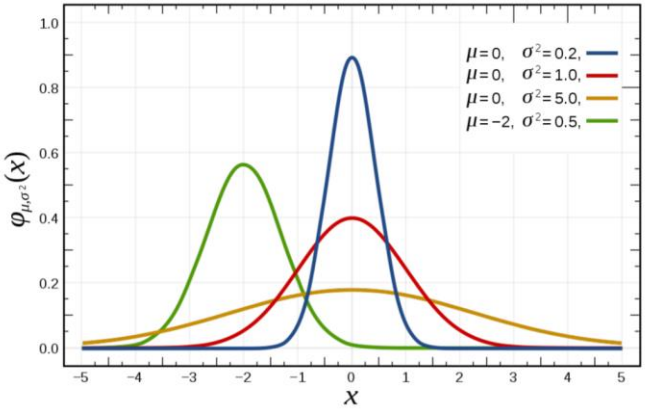
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \text{ dla } x \in \mathbb{R}$$

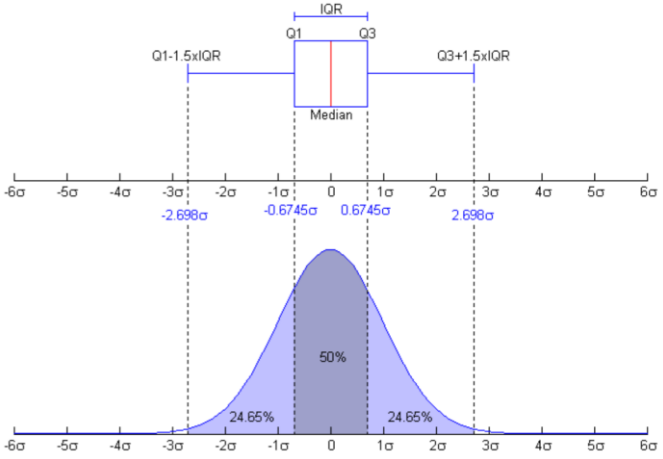
$$\mathbb{E}(X) = \mu$$

$$\mathrm{Var}(X) = \sigma^2$$

Rozkład normalny standardowy ma średnią $\mu = 0$ i wariancję $\sigma^2 = 1$. Stąd, jego funkcja gęstości to

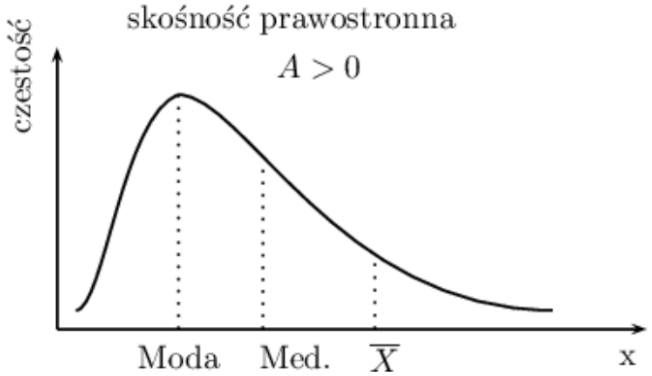
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \text{ dla } x \in \mathbb{R}$$

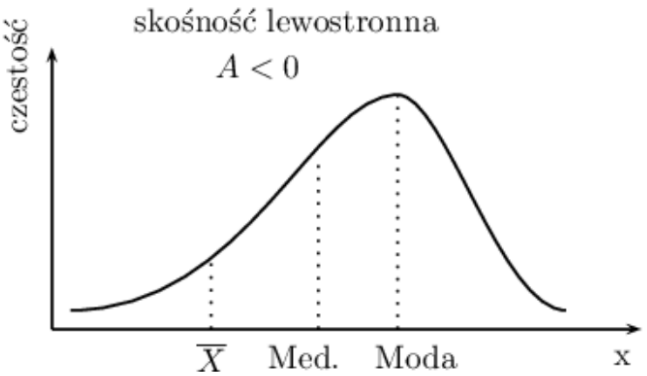

$\mu=0, \quad \sigma^2=0.2,$
$\mu=0, \quad \sigma^2=1.0,$
$\mu=0, \quad \sigma^2=5.0,$
$\mu=-2, \quad \sigma^2=0.5,$
$\varphi_{\mu,\sigma^2}(x)$
x


IQR
Q1
Q3
Q1-1.5xIQR
Q3+1.5xIQR
Median
-2.698σ
-0.6745σ
0.6745σ
2.698σ
50%
24.65%
24.65%

Współczynnik asymetrii

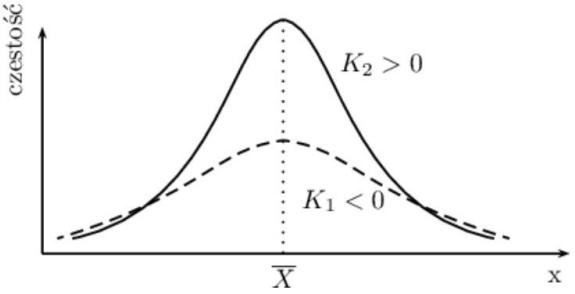
$$A_s = \frac{1}{s^3} \cdot (\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^3)$$





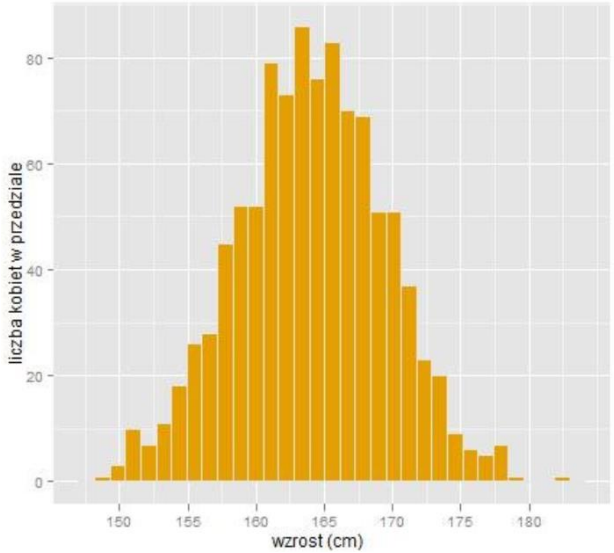
Kurtoza

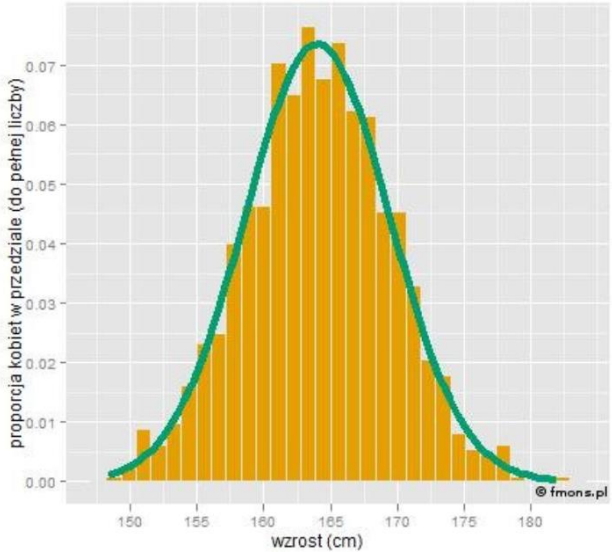
$$K = \frac{1}{s^4} \cdot \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4\right)$$



Histogram

Graficzny sposób przedstawiania rozkładu empirycznego. Przedstawia liczebności obserwacji w zadanych przedziałach badanej zmiennej.





Współzależność

Kowariancja

$$\mathrm{cov}(x, y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

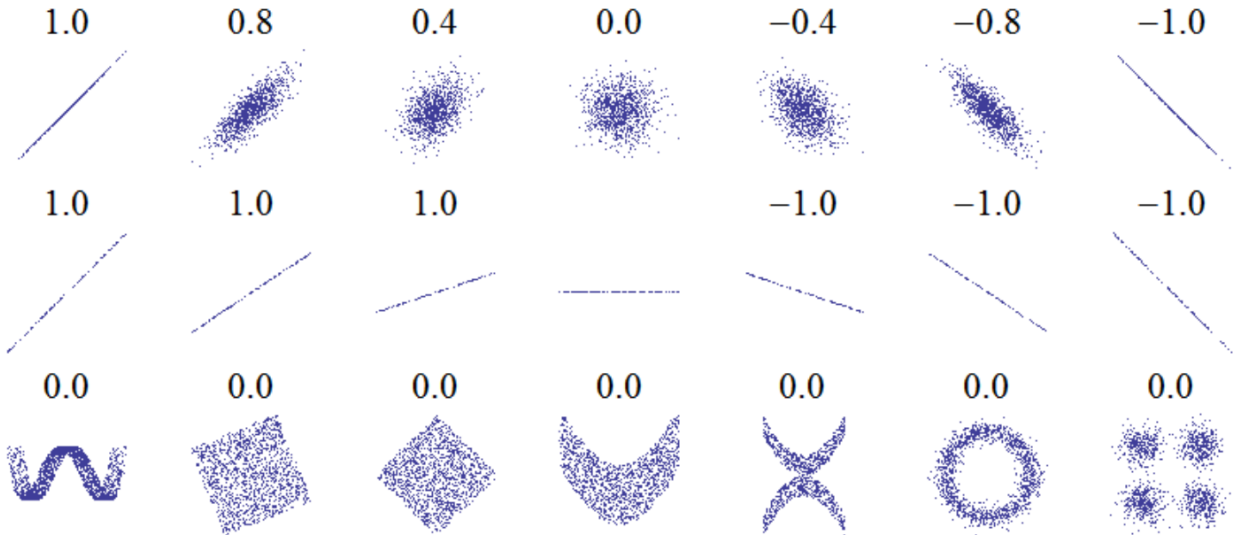
Współczynnik korelacji liniowej Pearsona

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{s(x) \cdot s(y)}$$

gdzie
$s(x)$, $s(y)$ – odchylenia standardowe zmiennych x, y.

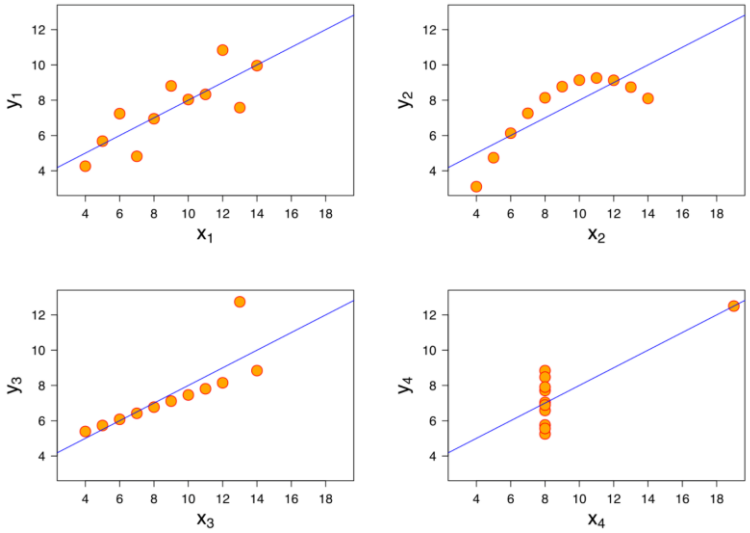
* r_{xy} bada tylko **liniową** zależność między zmiennymi,
* $r_{xy} \in \langle -1, 1 \rangle$ i pozwala określić siłę i kierunek zależności liniowej
  - jeśli $|r_{xy}|$ jest bliskie 0, to mamy słabą zależność liniową między zmiennymi,
  - jeśli $|r_{xy}|$ jest bliskie 1, to mamy silną zależność liniową między zmiennymi,
  - jeśli $r_{xy} > 0$, to zależność między zmiennymi jest dodatnia,
  - jeśli $r_{xy} < 0$, to zależność między zmiennymi jest ujemna.

r_{xy} =

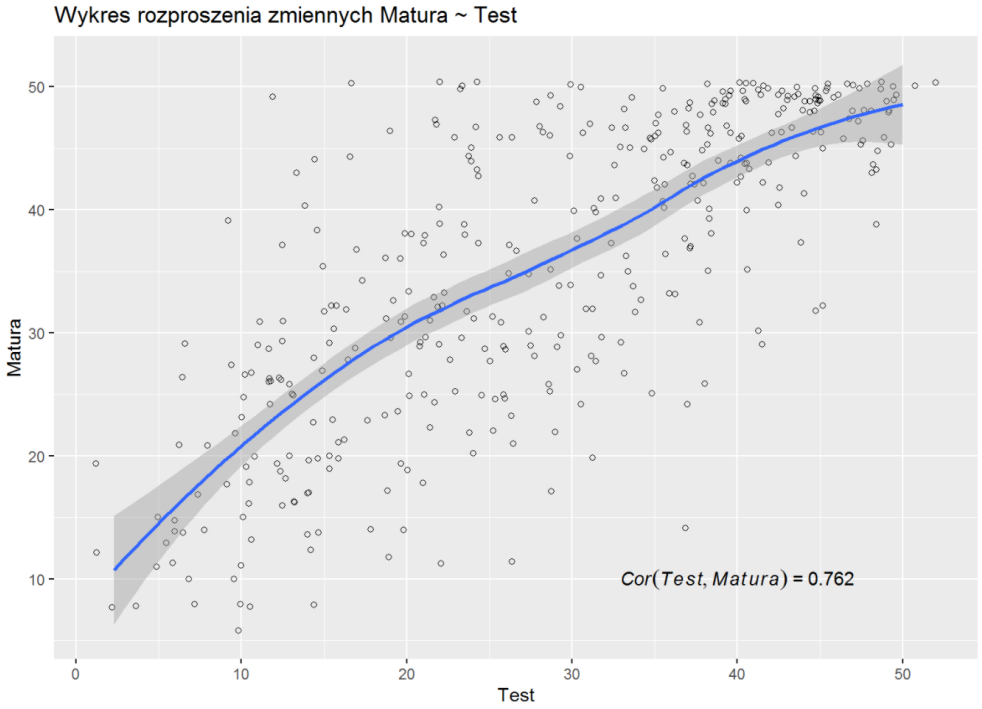


Kwartet Anscombe'a

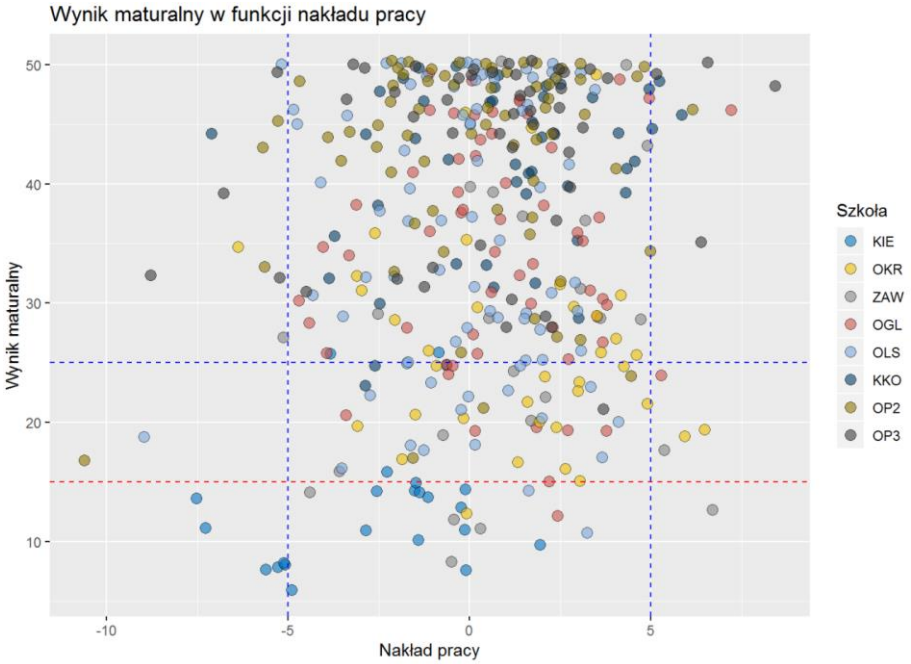
Cecha	Wartość
Średnia arytmetyczna zmiennej x	9
Wariancja zmiennej x	11
Średnia arytmetyczna zmiennej y	7.50
Wariancja zmiennej y	4.122
Współczynnik korelacji pomiędzy zmiennymi	0.816
Równanie regresji liniowej	$y = 3.00 + 0.500x$



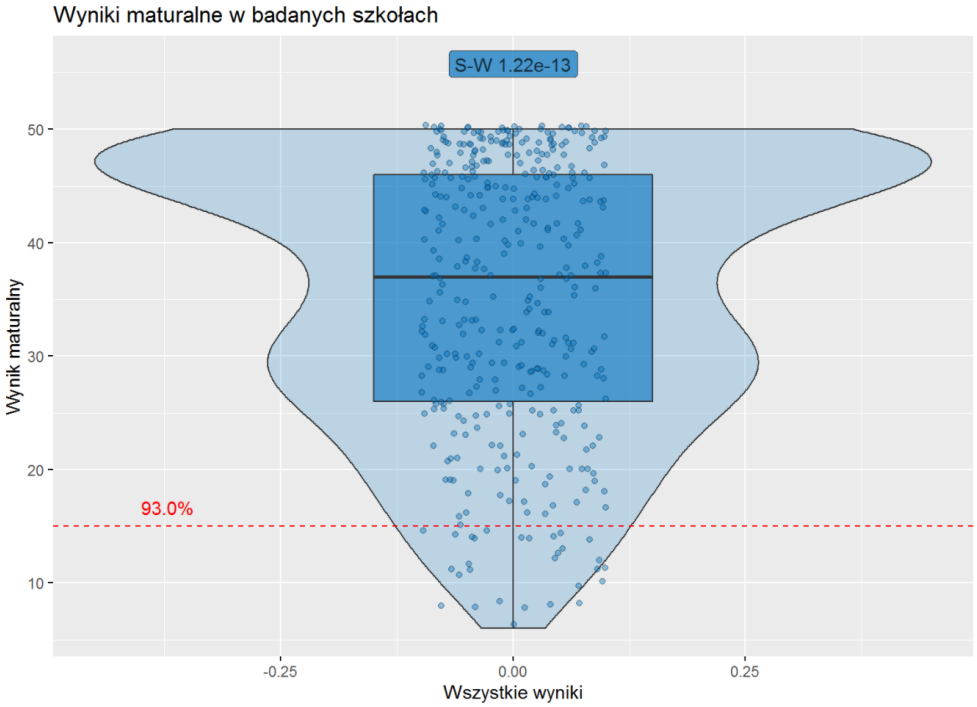
GRAFICZNA PREZENTACJA DANYCH



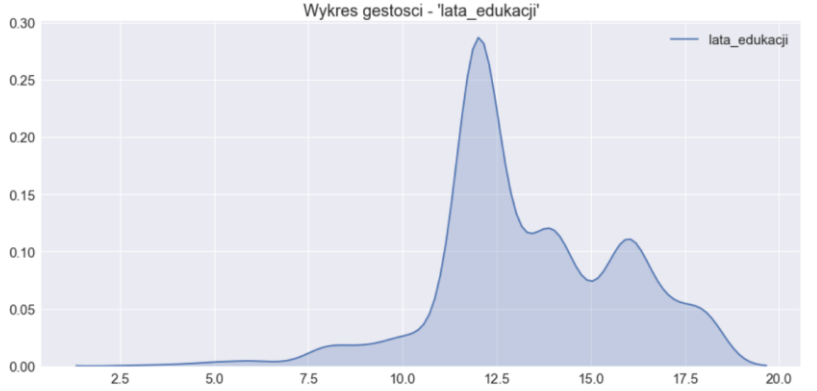
GRAFICZNA PREZENTACJA DANYCH

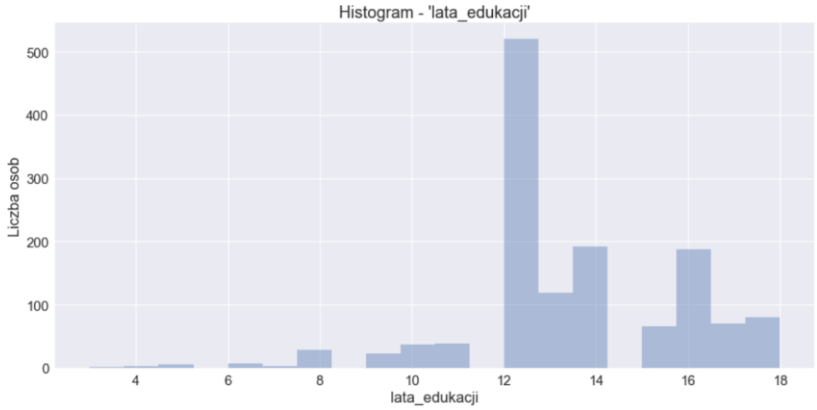


GRAFICZNA PREZENTACJA DANYCH

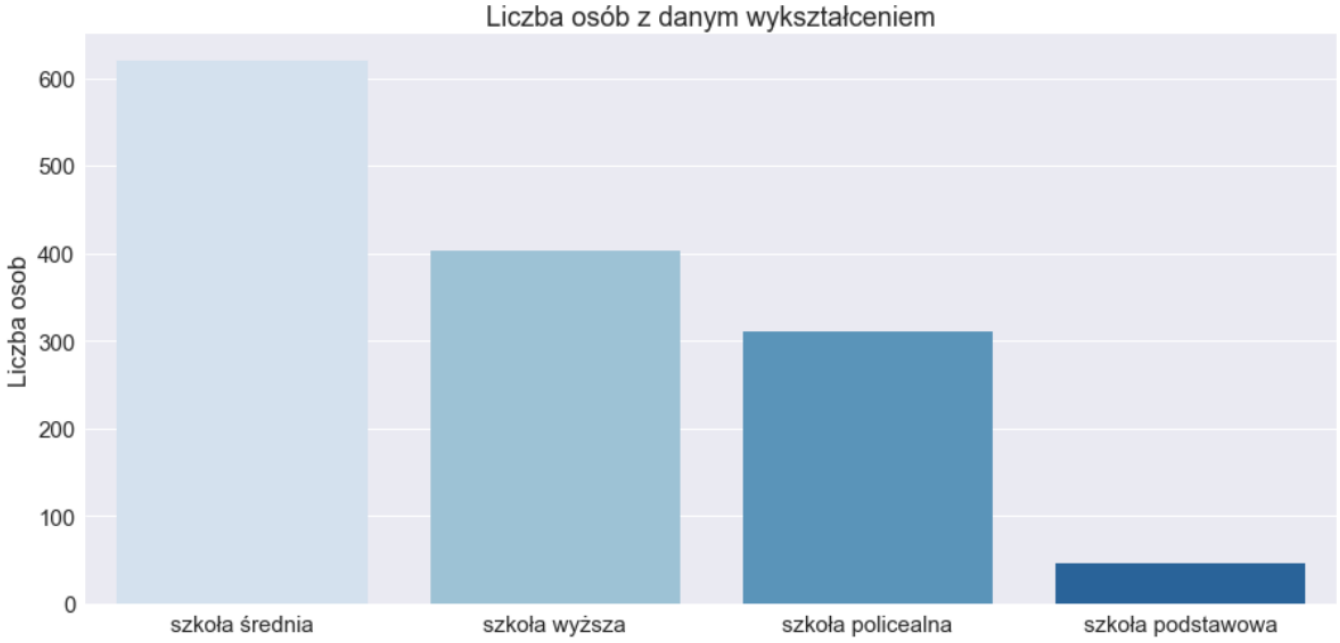


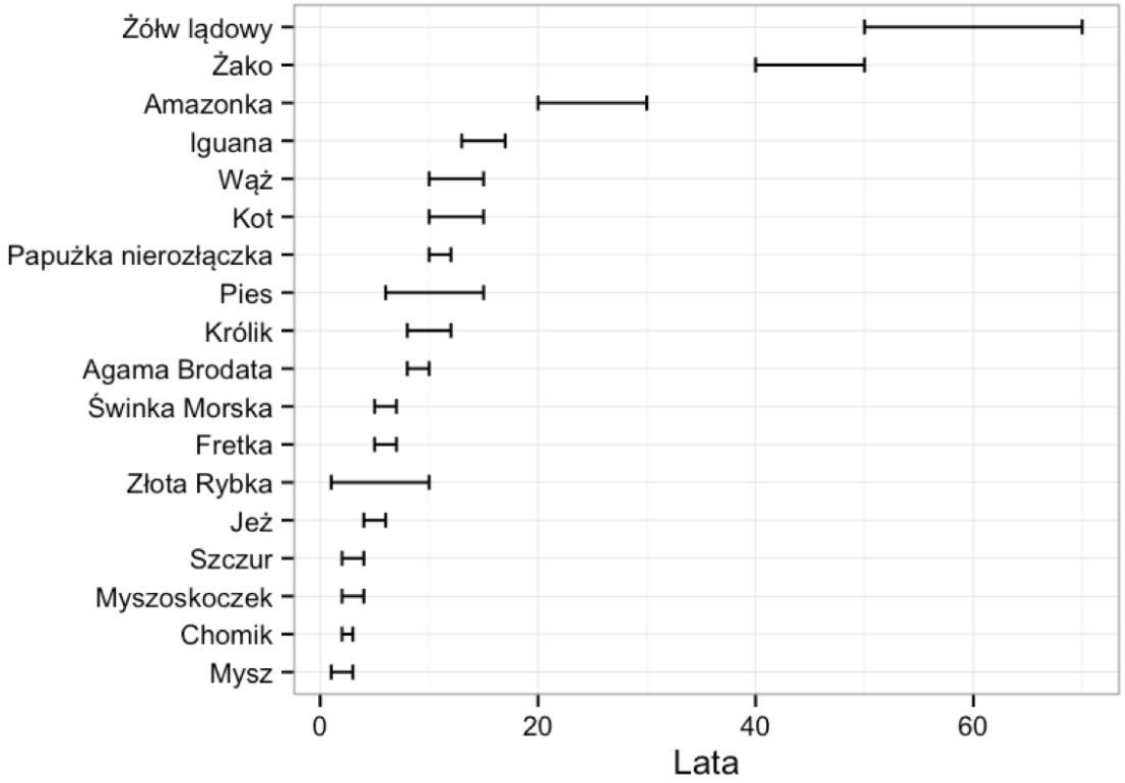
GRAFICZNA PREZENTACJA DANYCH

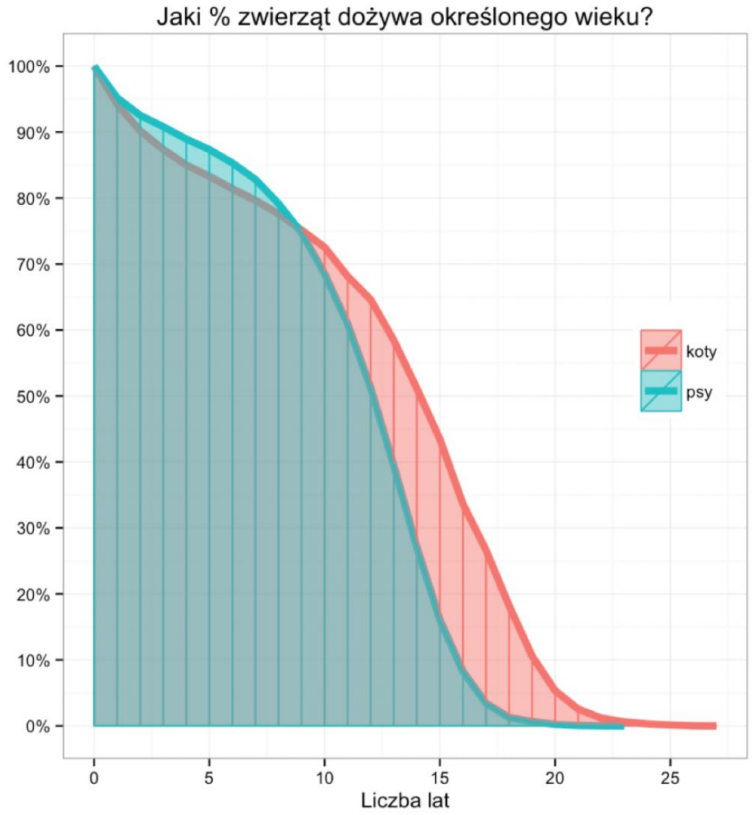




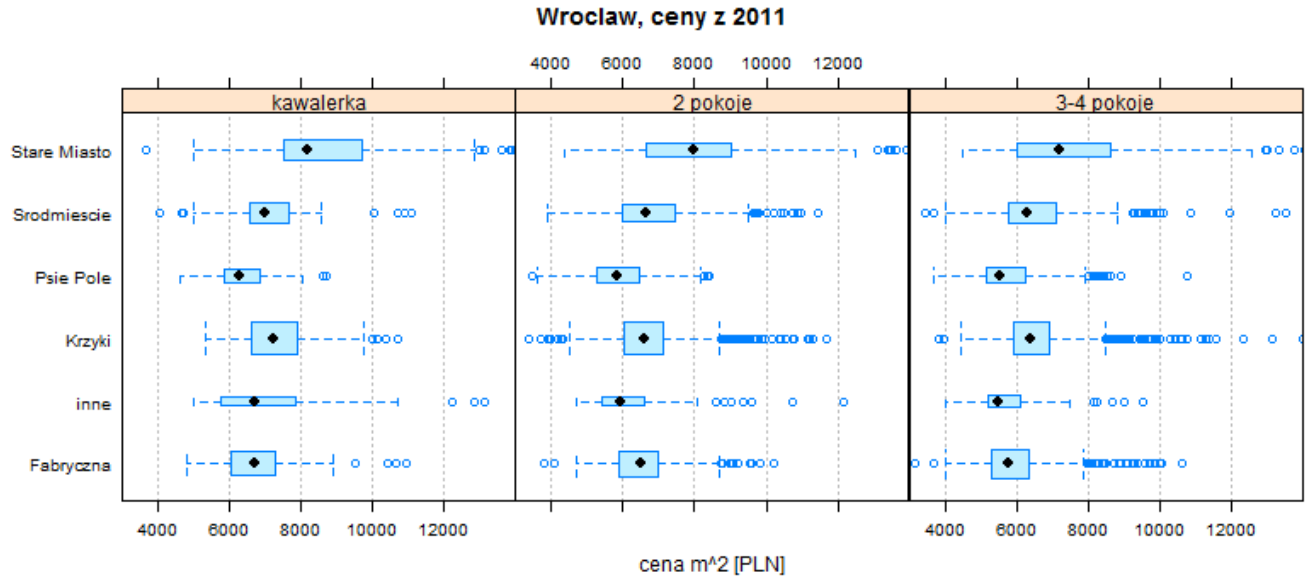
GRAFICZNA PREZENTACJA DANYCH

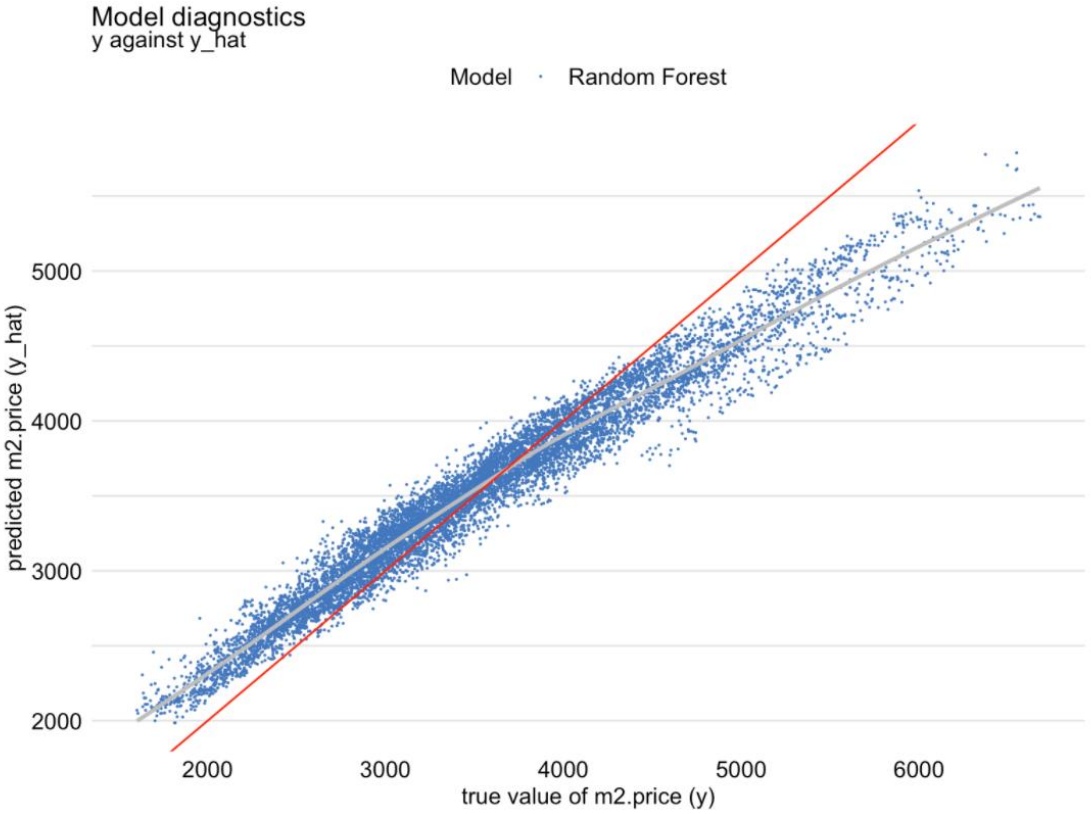



Żółw lądowy
Żako
Amazonka
Iguana
Wąż
Kot
Papużka nierozłączka
Pies
Królik
Agama Brodata
Świnka Morska
Fretka
Złota Rybka
Jeż
Szczur
Myszoskoczek
Chomik
Mysz
0
20
40
60
Lata

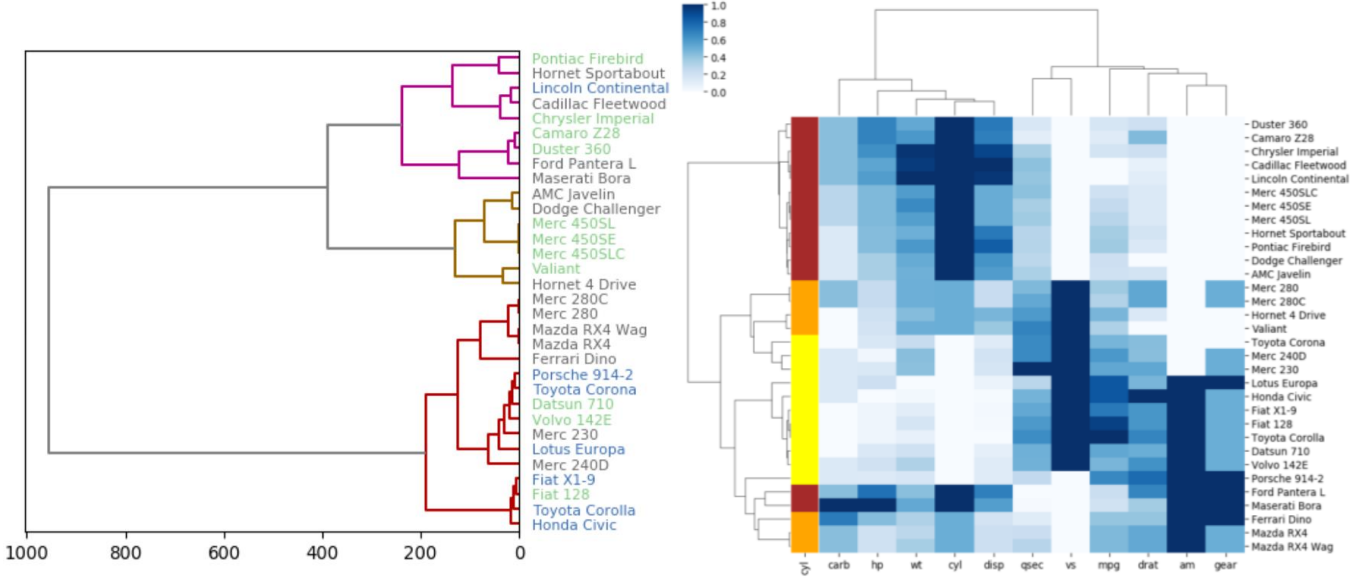

Jaki % zwierząt dożywa określonego wieku?
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0
5
10
15
20
25
Liczba lat
koty
psy

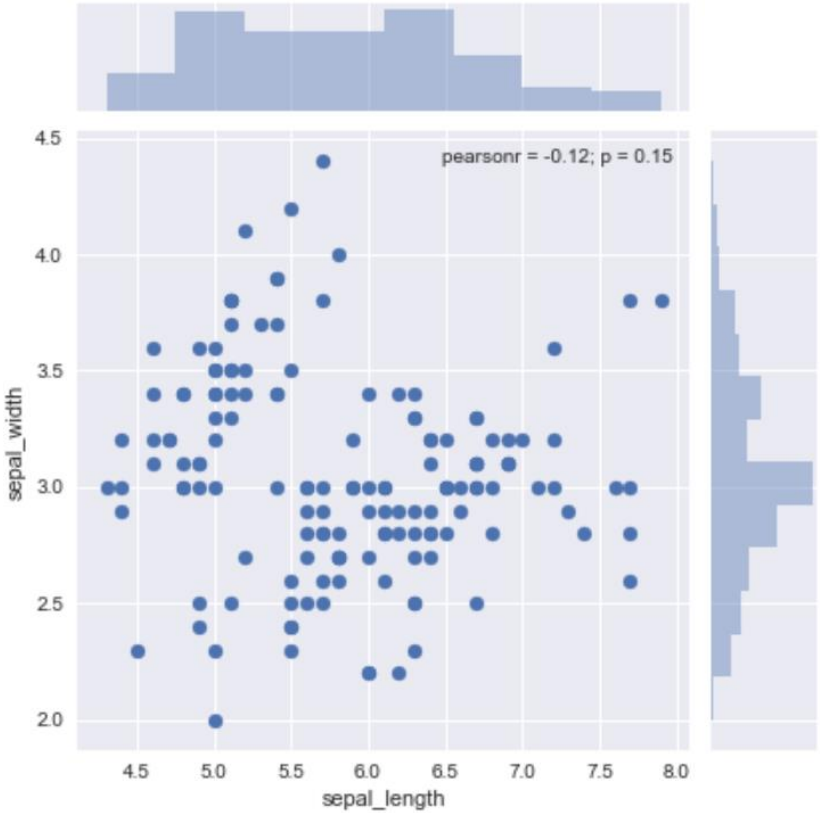
GRAFICZNA PREZENTACJA DANYCH



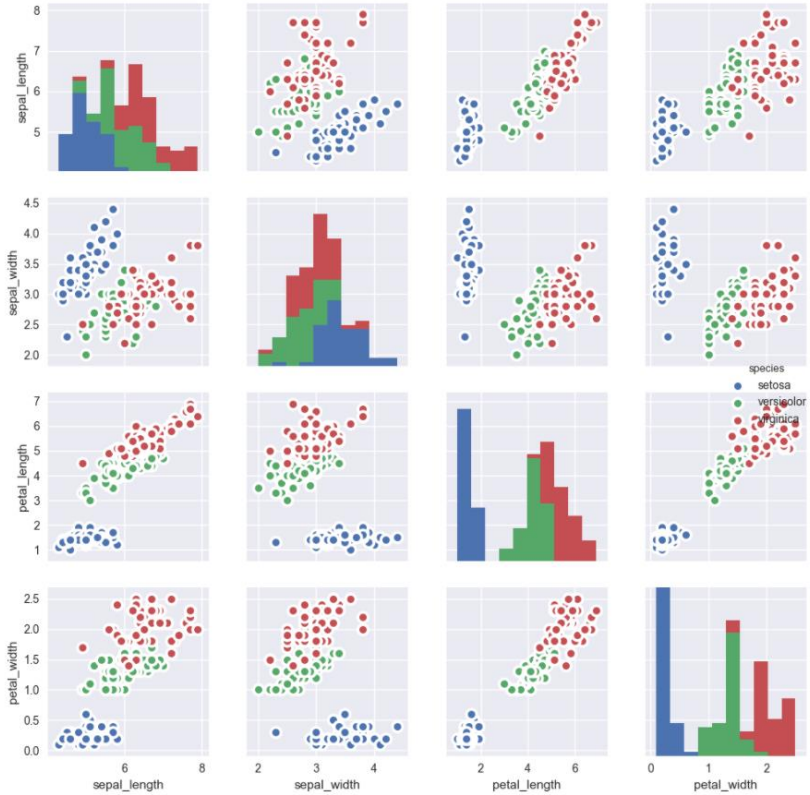

Model diagnostics
y against y_hat
Model · Random Forest
predicted m2.price (y_hat)
5000
4000
3000
2000
2000
3000
4000
5000
6000
true value of m2.price (y)

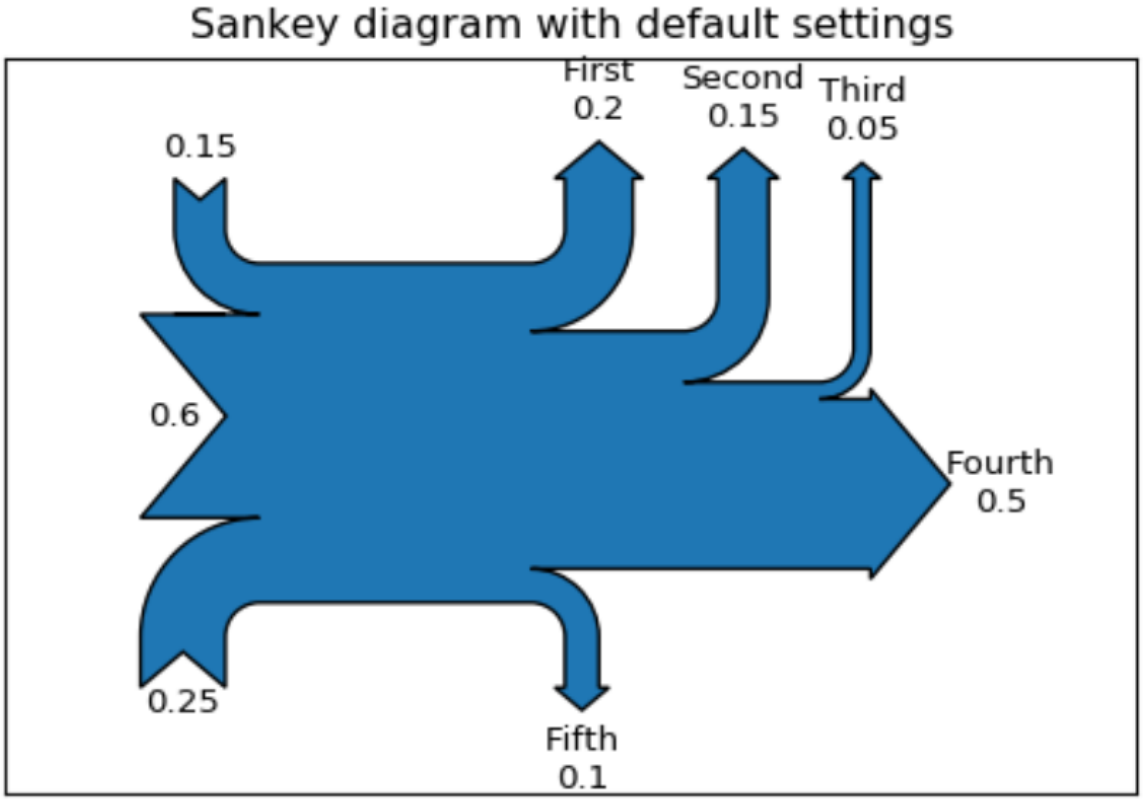
GRAFICZNA PREZENTACJA DANYCH

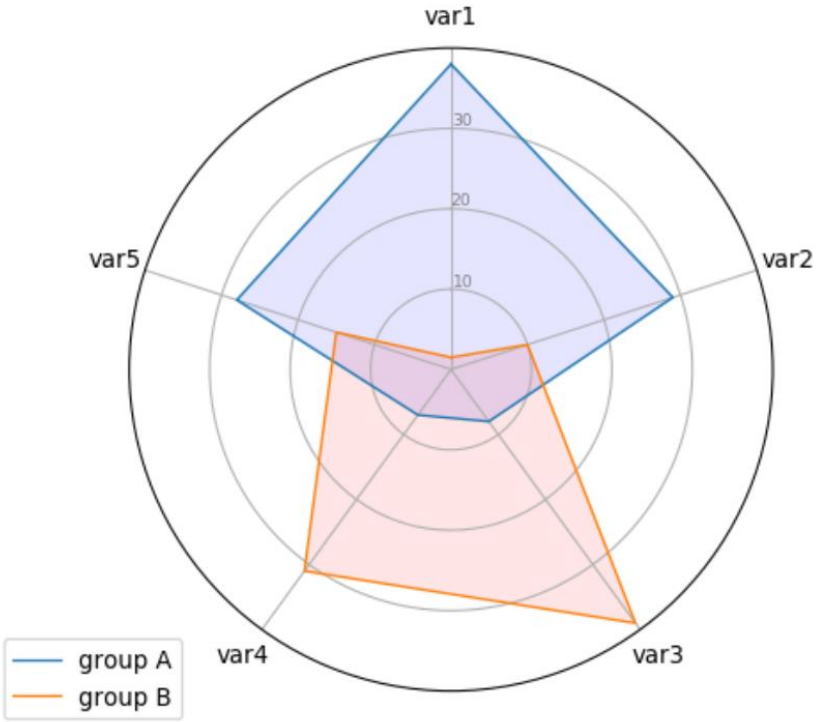



pearsonr = -0.12; p = 0.15
sepal_width
sepal_length
4.5
4.0
3.5
3.0
2.5
2.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

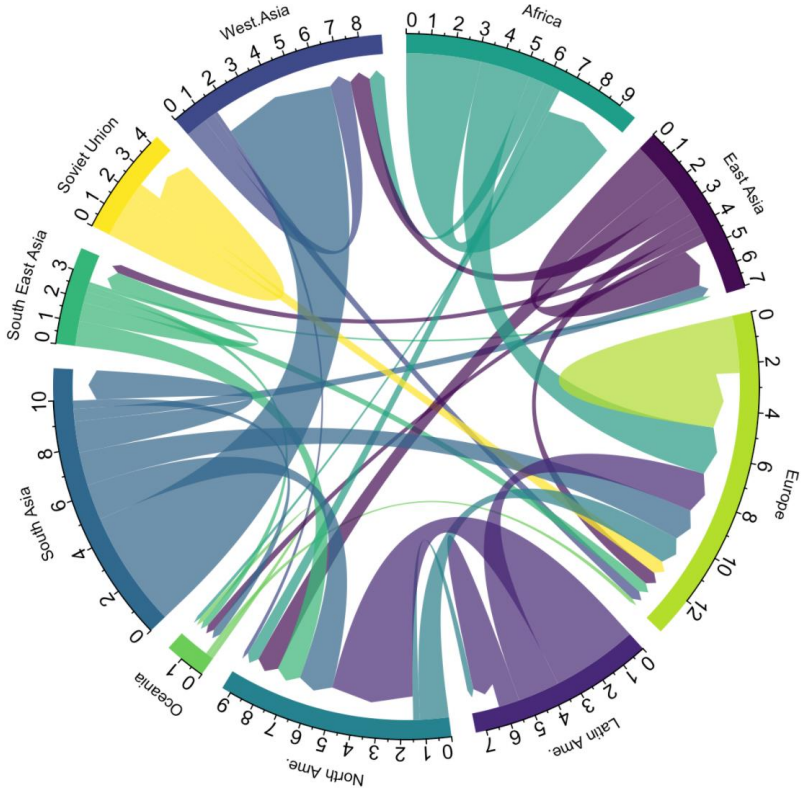
GRAFICZNA PREZENTACJA DANYCH




Sankey diagram with default settings
First
0.2
Second
0.15
Third
0.05
0.15
0.6
Fourth
0.5
0.25
Fifth
0.1



#391 Several group on the same radar chart


West.Asia
Africa
East Asia
Soviet Union
South East Asia
Europe
South Asia
Latin Ame.
Oceania
North Ame.