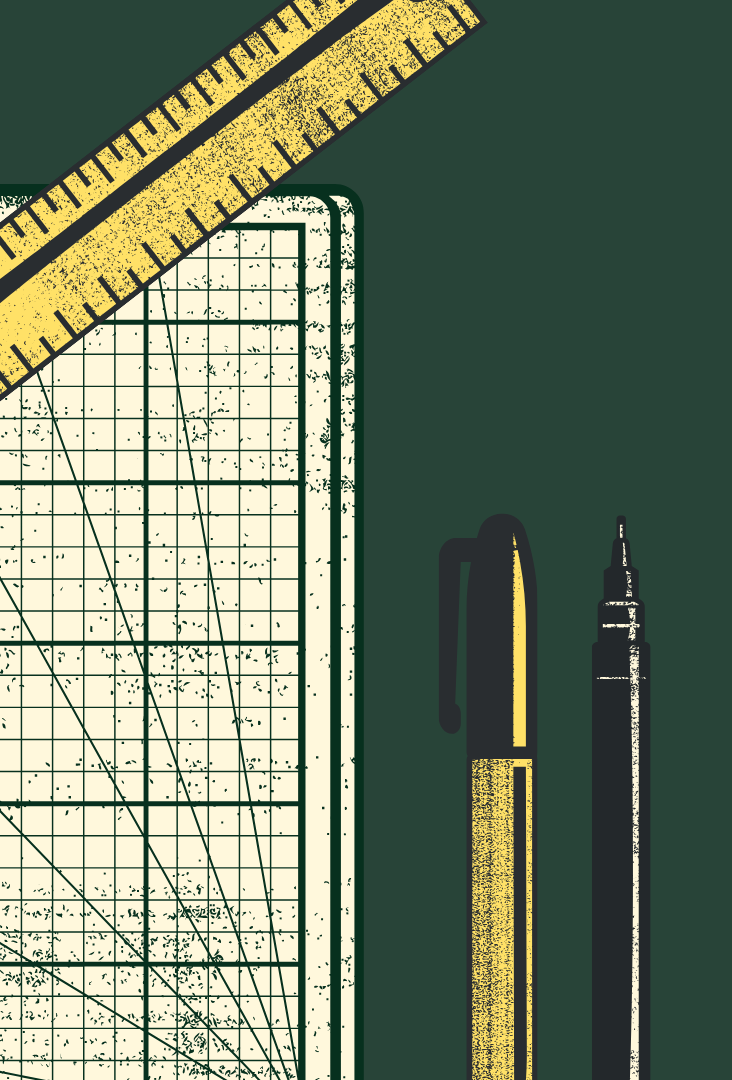


Wstęp do Eksploracji Danych

Politechnika Warszawska

Anna Kozak

Anna Kozak

koordynator przedmiotu, wykład

anna.kozak@pw.edu.pl
MS Teams

Zajęcia laboratoryjne

Antoni Chudy

Karolina Dunal

Anna Kozak

Zajęcia projektowe

Iza Danielewska

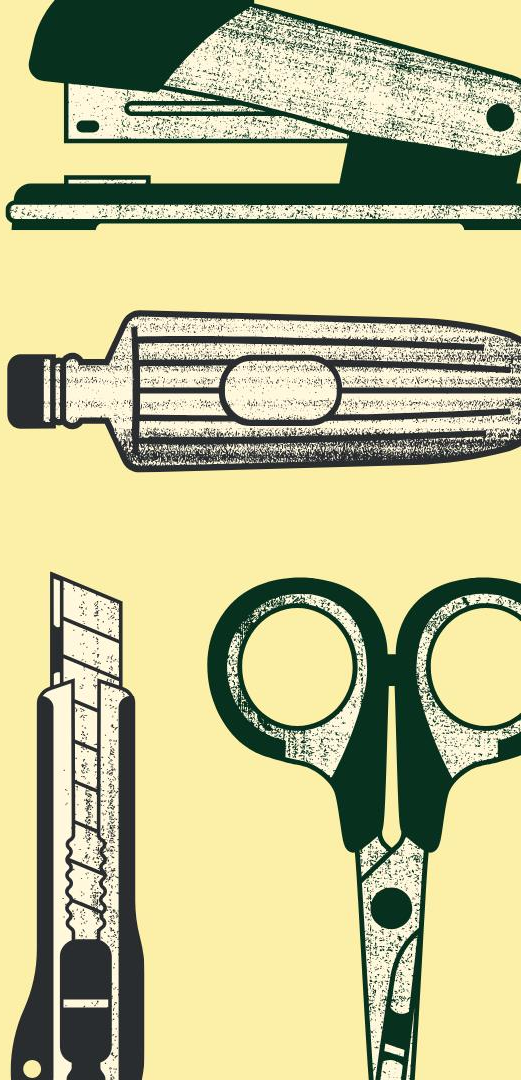
Anna Kozak

Dawid Płudowski

Katarzyna Woźnica

Strona przedmiotu

https://github.com/kozaka93/2026L-ExploratoryDataAnalysis



Wykład

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

15h wykładów = 6 x 2h wykład + 3 x 1h wykład (w tym 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 (R: ggplot, plotly; Python: matplotlib, seaborn)
- różne sposoby oceny zmiennych, danych, wizualizacji
- 6 x praca domowa (6 x 6p)

Ocena końcowa

Suma punktów z prac domowych i projektów:

$6 \times 6 + 24 + 20 + 10 = 90$

$(PD) + (P1) + (P2) + (T) = (O)$

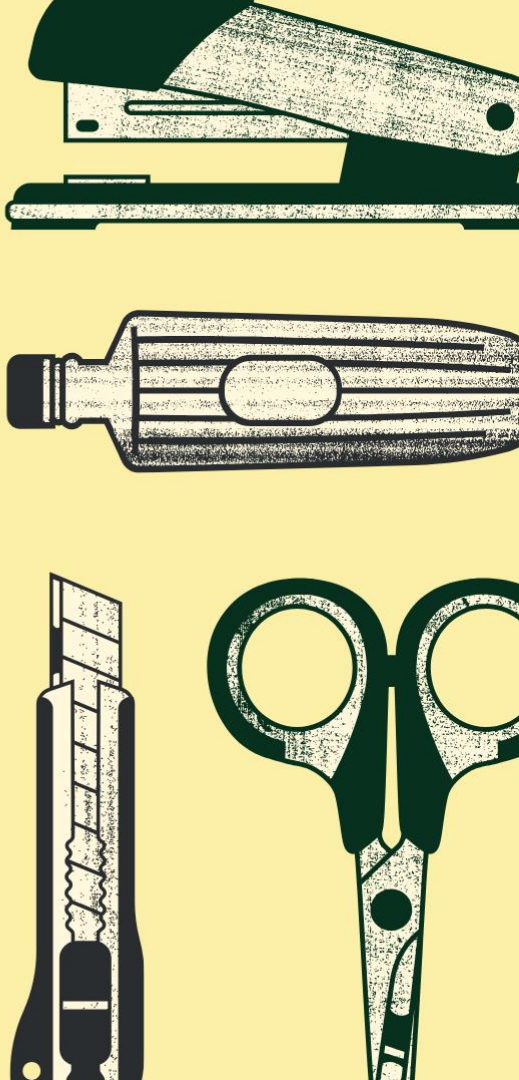
Aby zaliczyć kurs należy uzyskać ponad 45 punktów,
w tym co najmniej 50% punktów z każdego z projektów.

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?

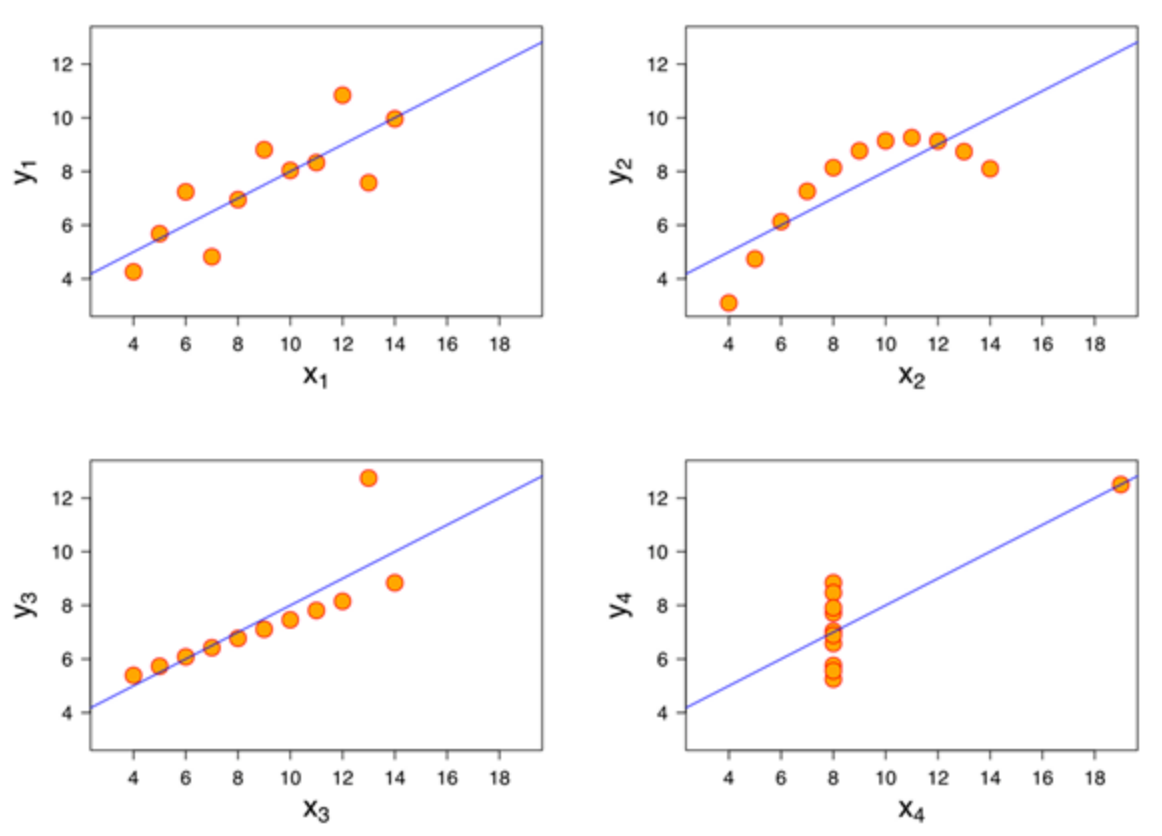
“proces odkrywania nietrywialnych, dotychczas nieznanych, potencjalnie użytecznych reguł, zależności, trendów”

Cel: analiza danych w celu lepszego ich zrozumienia

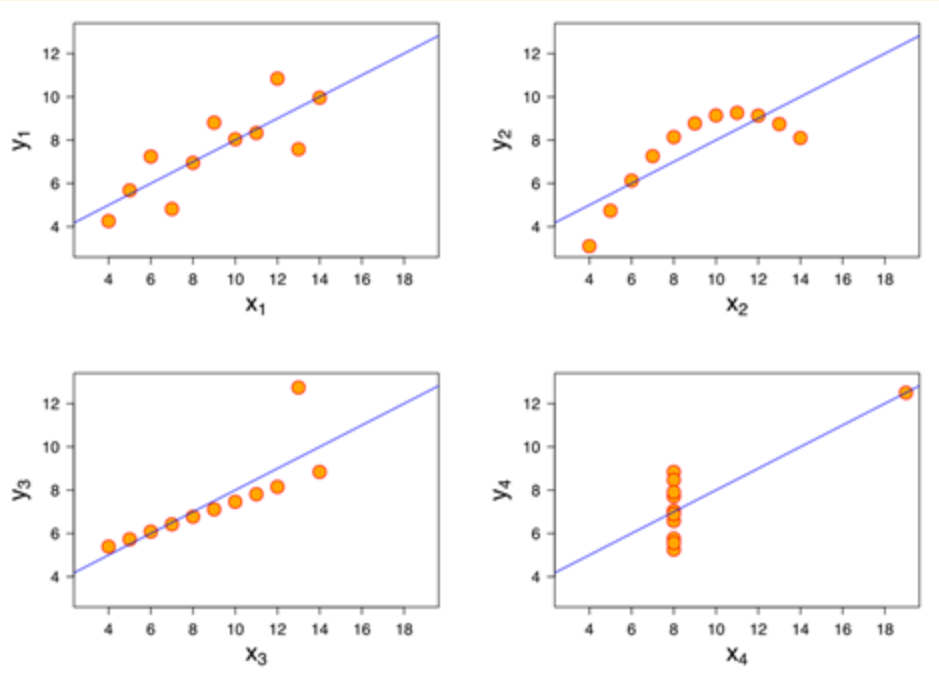
Eksploracja danych - czym jest?

Na eksplorację danych składa się wiele dyscyplin, między innymi:

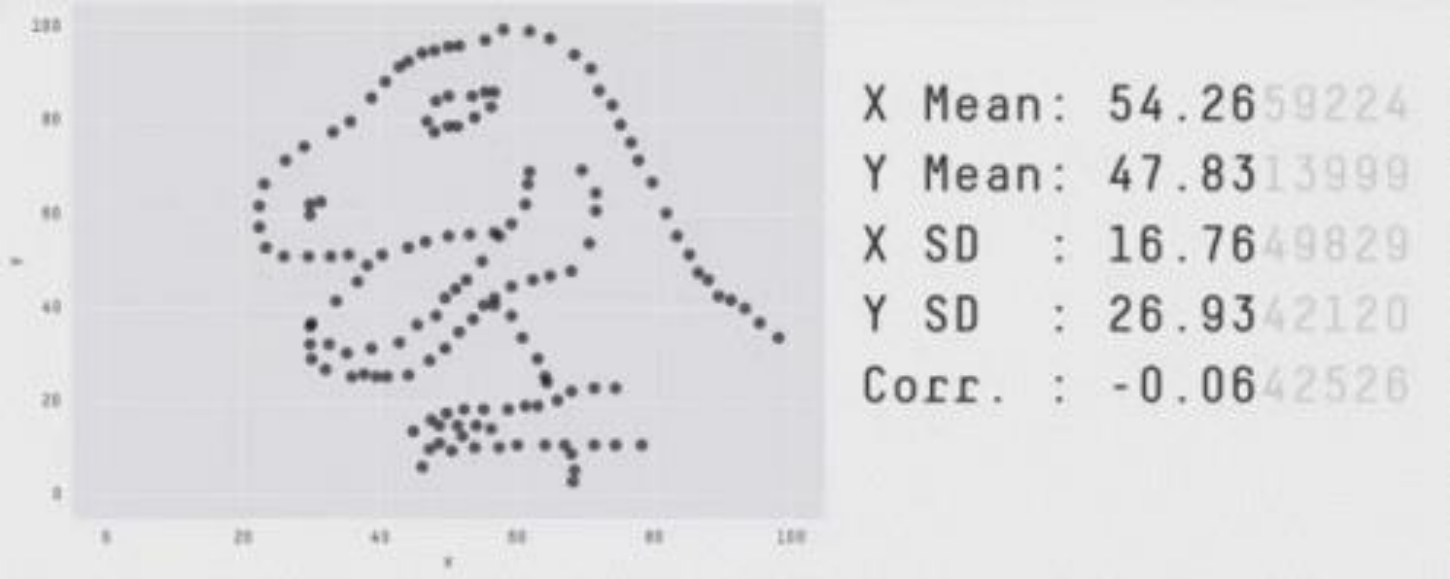
- bazy danych
- statystyka
- uczenie maszynowe
- wizualizacja danych
- wyszukiwanie informacji



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

Gramatyka języka wizualizacji

Dlaczego projektujemy wykresy?

Dlaczego projektujemy wykresy?

“aby pokazać historie ukryte w danych”

Trzy sposoby przedstawienia historii

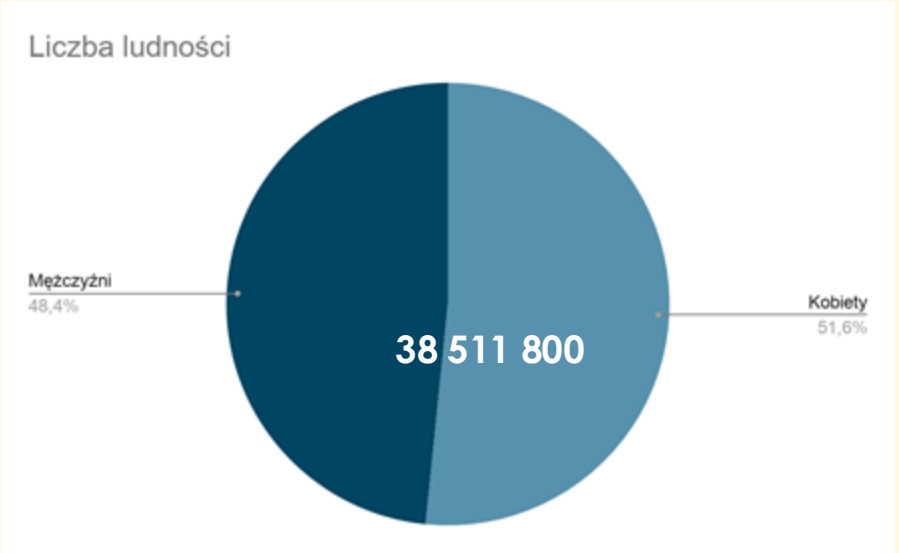
1) Opis słowny

“W wyniku przeprowadzenia Narodowego Spisu Powszechnego w roku 2011 ustalono, że w Polsce mieszka 38 511 800 osób, z czego 48,4% to mężczyźni, a 51,6% to kobiety.”

2) Tabela

Liczba ludności Polski	W tym kobiet	W tym mężczyzn
38 511 800	51,6%	48,4%

3) Wykres



http://www.biecek.pl/Eseje/indexGramatyka.html

Co w przypadku dużego zbioru danych?

Co w przypadku dużego zbioru danych?

Historia wyników polskich matur z lat 2010-2015.

1) Opis słowny

“Wyniki matury z języka polskiego mają rozkład zbliżony do normalnego. W poszczególnych latach średnie tego rozkładu nieznacznie się różnią. Rozkład ten jest zaburzony w okolicy 21-22 punktów, czyli w pobliżu wartości stanowiących granicę zaliczenia (30% możliwych do uzyskania punktów). Praktycznie nie ma uczniów, którzy uzyskaliby jeden punkt poniżej progu zaliczenia, jest za to bardzo dużo osób, które zdały egzamin, otrzymując punkt więcej. Sugeruje to, że dosyć często osoby oceniające maturę, widząc, że do zaliczenia brakuje jednego–dwóch punktów, brakujące punkty “znajdowały”. W przypadku egzaminu z matematyki rozkłady są różne w różnych rocznikach i zdecydowanie nie przypominają rozkładu normalnego. W pobliżu progu zaliczenia również widzimy pewną nieregularność, największą w roku 2014. Jest ona jednak mniejsza niż w przypadku egzaminu z języka polskiego.”

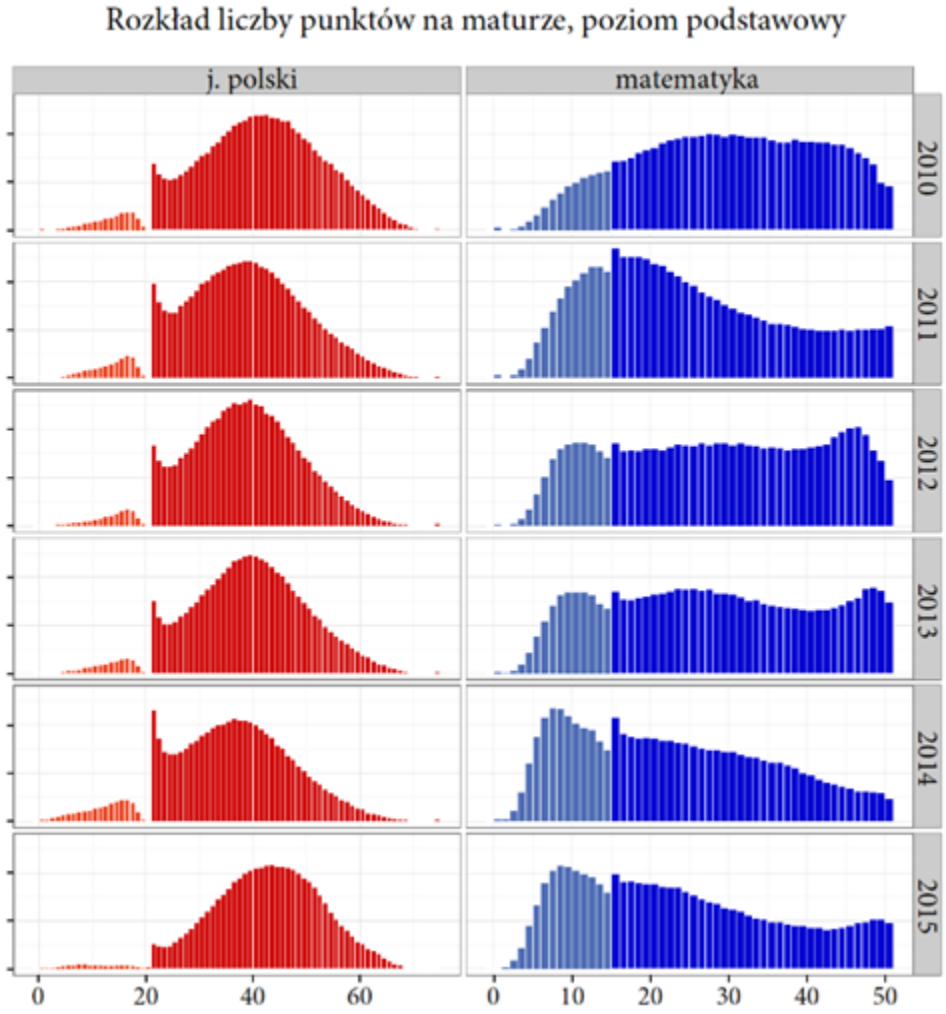
http://www.biecek.pl/Eseje/indexGramatyka.html

2) Tabela

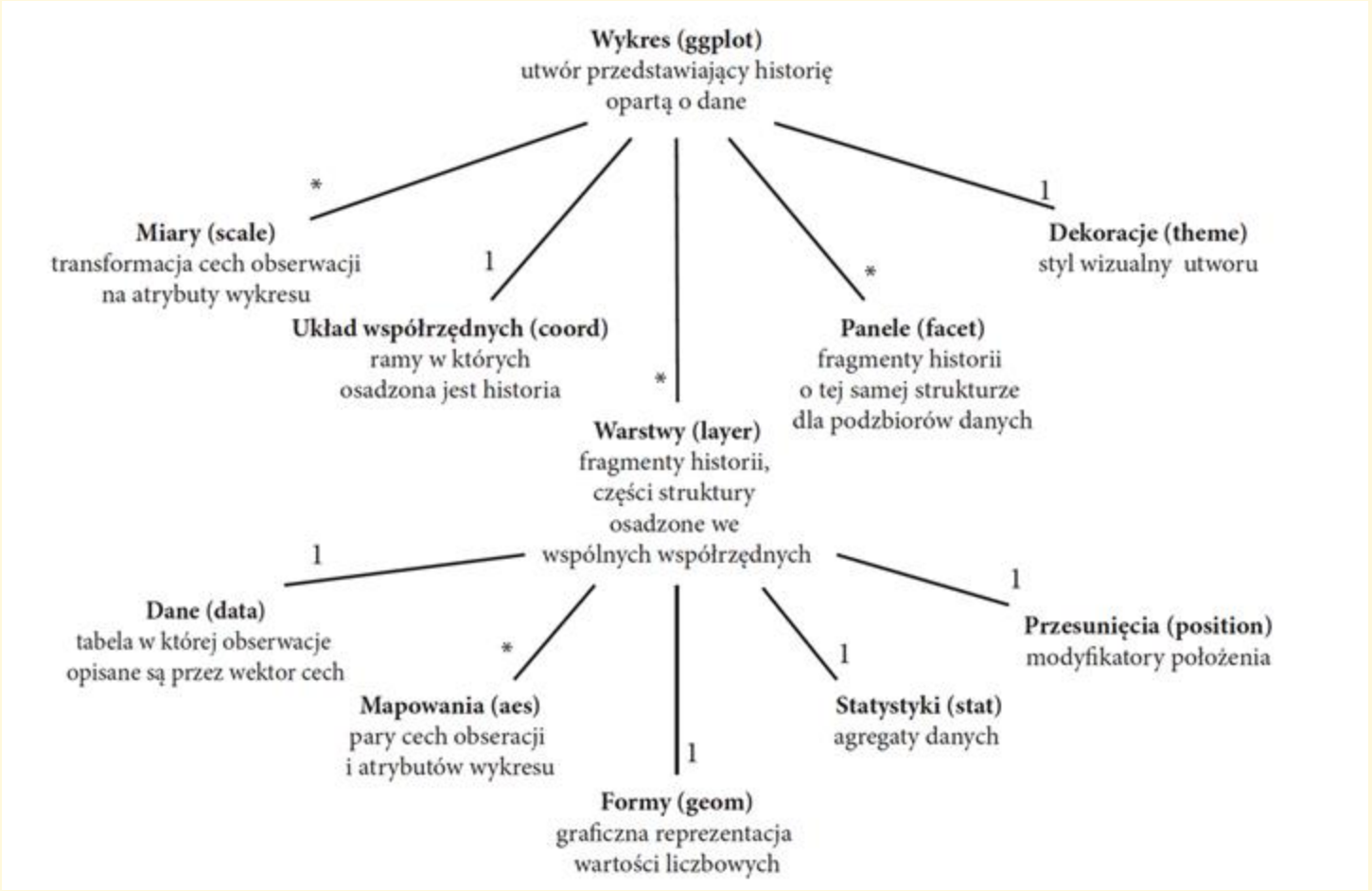
punkty	przedmiot	2010	2011	2012	2013	2014	2015
...	...	...	...	...	...	...	...
6	j. polski	0,09	0,09	0,09	0,09	0,25	0,16
7	j. polski	0,12	0,14	0,11	0,12	0,28	0,16
8	j. polski	0,16	0,18	0,12	0,14	0,34	0,19
9	j. polski	0,19	0,22	0,14	0,19	0,36	0,19
10	j. polski	0,23	0,27	0,18	0,21	0,40	0,17
11	j. polski	0,25	0,29	0,20	0,25	0,45	0,16
12	j. polski	0,28	0,31	0,23	0,28	0,47	0,15
13	j. polski	0,34	0,36	0,27	0,31	0,50	0,13
14	j. polski	0,37	0,41	0,32	0,37	0,61	0,13
15	j. polski	0,42	0,47	0,37	0,41	0,68	0,16
16	j. polski	0,49	0,57	0,45	0,45	0,73	0,16
17	j. polski	0,54	0,67	0,50	0,50	0,74	0,17
18	j. polski	0,54	0,62	0,46	0,44	0,63	0,14
19	j. polski	0,34	0,34	0,26	0,27	0,31	0,10
20	j. polski	0,13	0,09	0,09	0,09	0,07	0,06
21	j. polski	0,02	0,01	0,01	0,01	0,01	0,10
22	j. polski	1,90	2,72	2,43	2,28	3,76	0,90
23	j. polski	1,60	2,20	1,96	1,78	2,80	0,82
24	j. polski	1,46	1,95	1,80	1,56	2,36	0,81
25	j. polski	1,44	1,91	1,80	1,59	2,28	0,85
...	...	...	...	...	...	...	...

http://www.biecek.pl/Eseje/indexGramatyka.html

3) Wykres



http://www.biecek.pl/Eseje/indexGramatyka.html



http://www.biecek.pl/Eseje/indexGramatyka.html

Reprezentacja szeroka danych

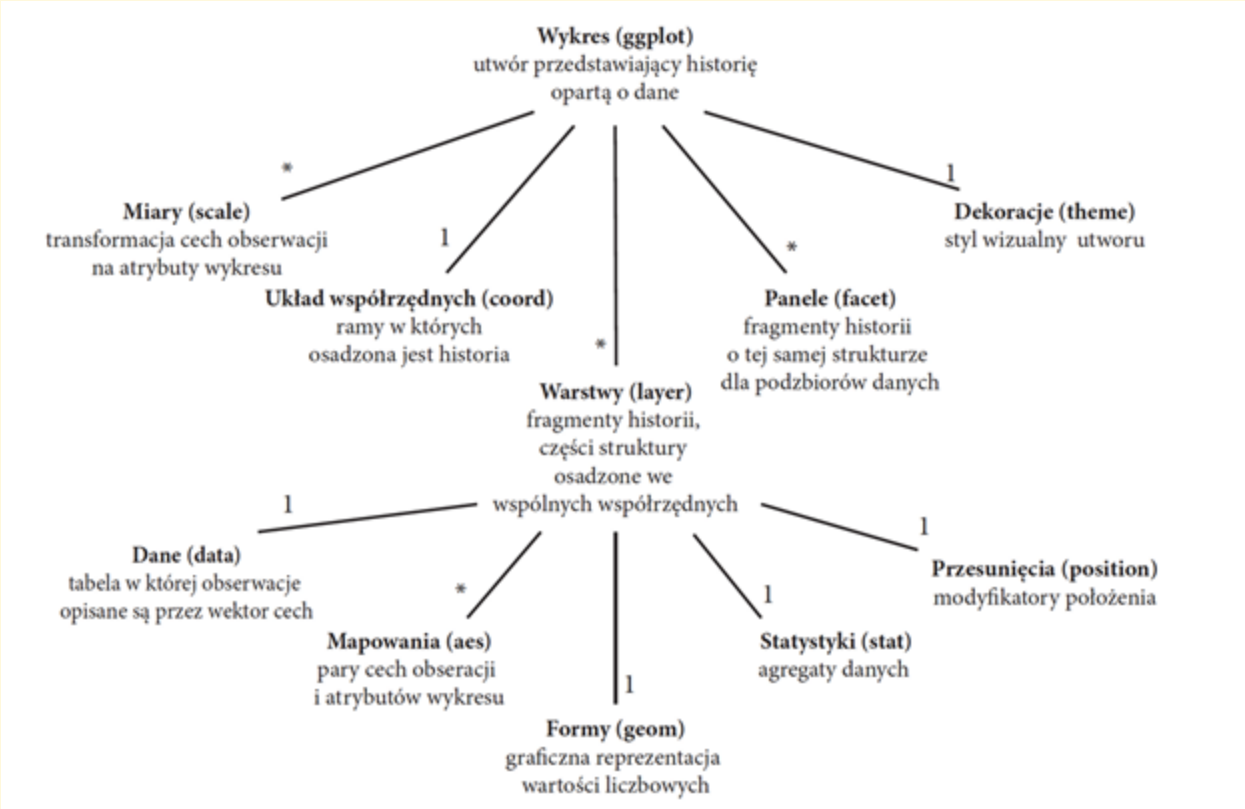
przedmiot	rok_2010	rok_2011	rok_2012	rok_2013	rok_2014	rok_2015
j. polski	40.1	37.5	37.5	38.3	35.2	41.5
matematyka	29.2	24.1	27.9	27.3	22.3	

Reprezentacja wąska danych

rok	przedmiot	srednia
rok_2010	j. polski	40.1
rok_2011	j. polski	37.5
rok_2012	j. polski	37.5
rok_2013	j. polski	38.3
rok_2014	j. polski	35.2
rok_2015	j. polski	41.5
rok_2010	matematyka	29.2
rok_2011	matematyka	24.1
rok_2012	matematyka	27.9
rok_2013	matematyka	27.3
rok_2014	matematyka	22.3

http://www.biecek.pl/Eseje/indexGramatyka.html

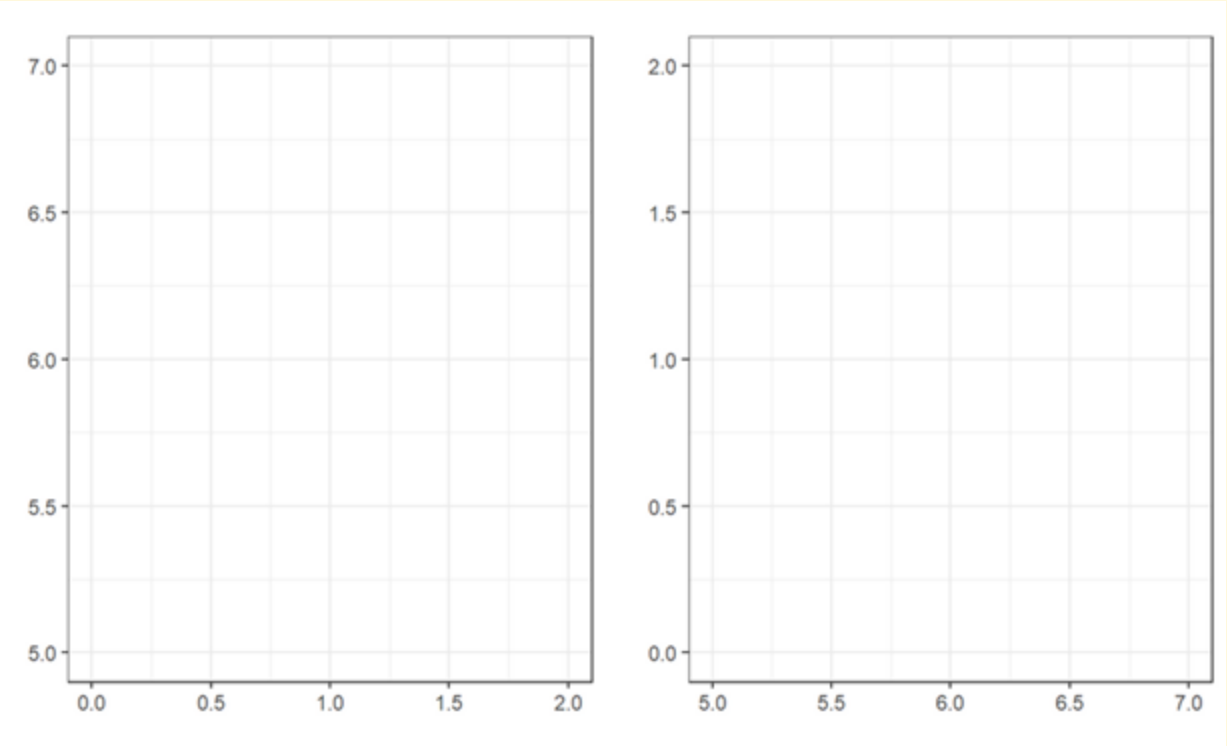
Układ współrzędnych



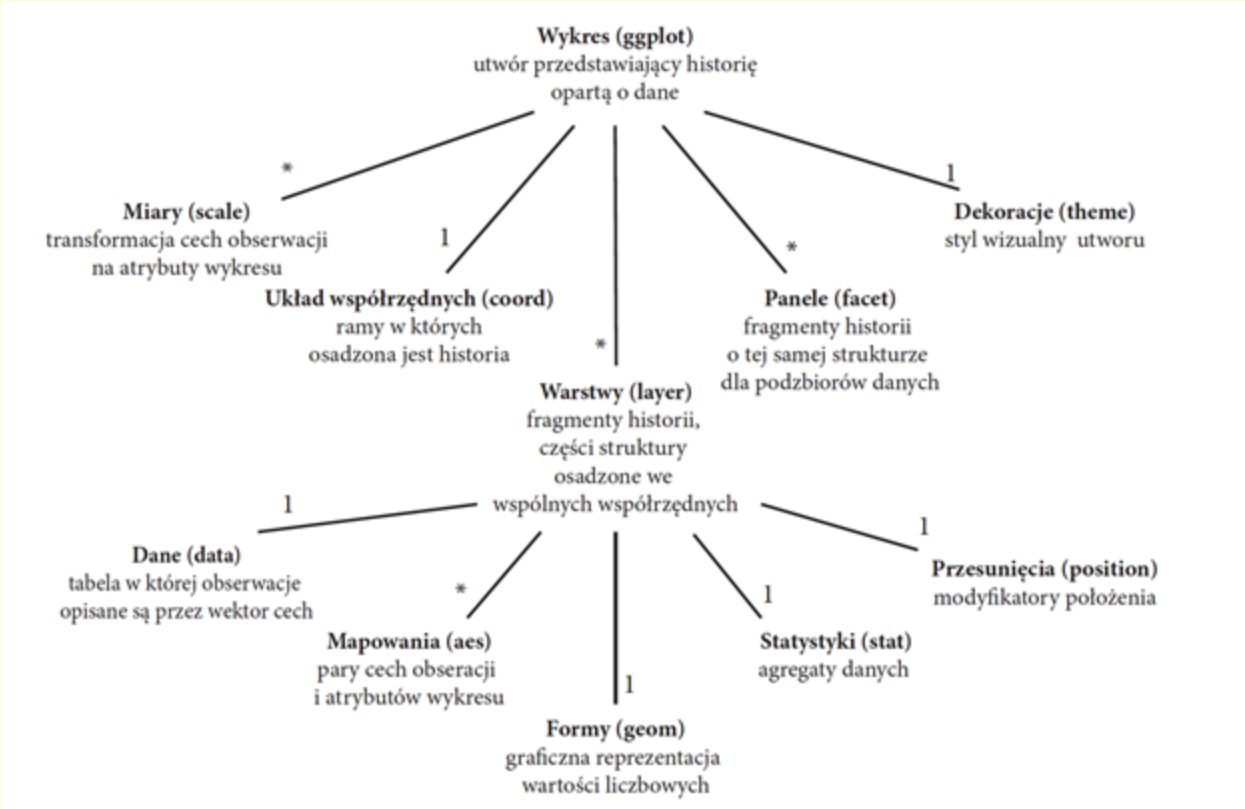
http://www.biecek.pl/Eseje/indexGramatyka.html

Układ współrzędnych

Układ współrzędnych (coords): ramy, w których osadzona jest historia.



Warstwy - Dane



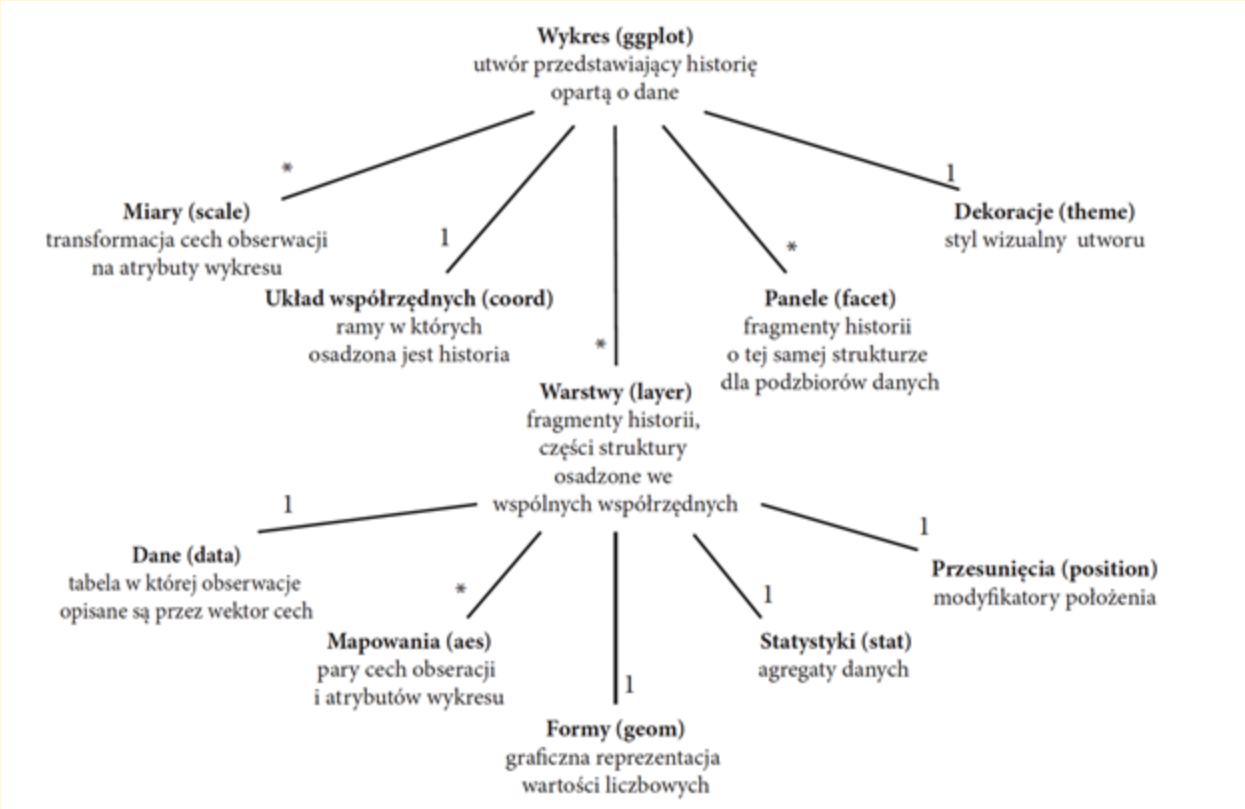
http://www.biecek.pl/Eseje/indexGramatyka.html

Warstwy - Dane

Dane (data): tabela, w której obserwacje opisane są przez wektory cech.

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

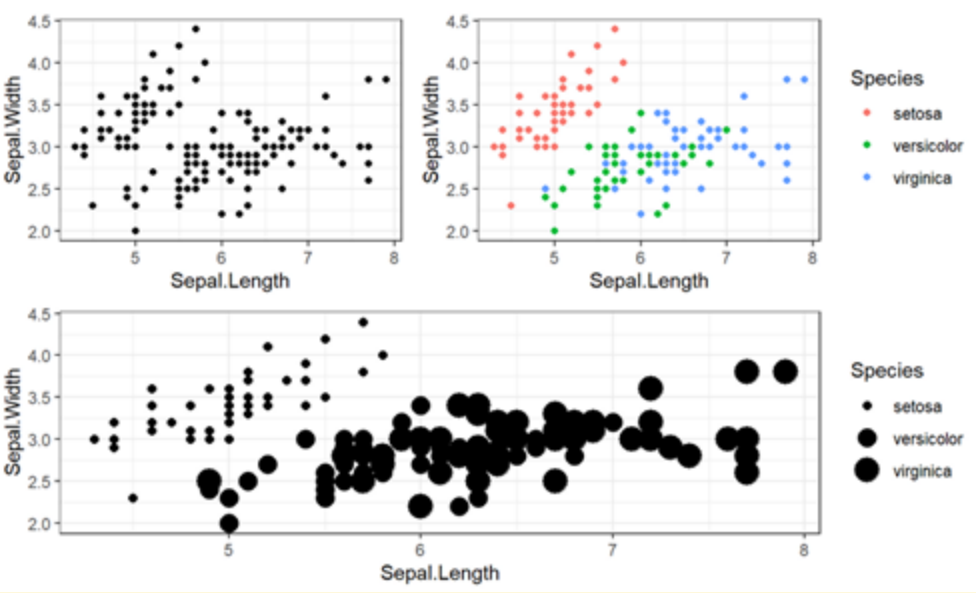
Warstwy - Mapowania



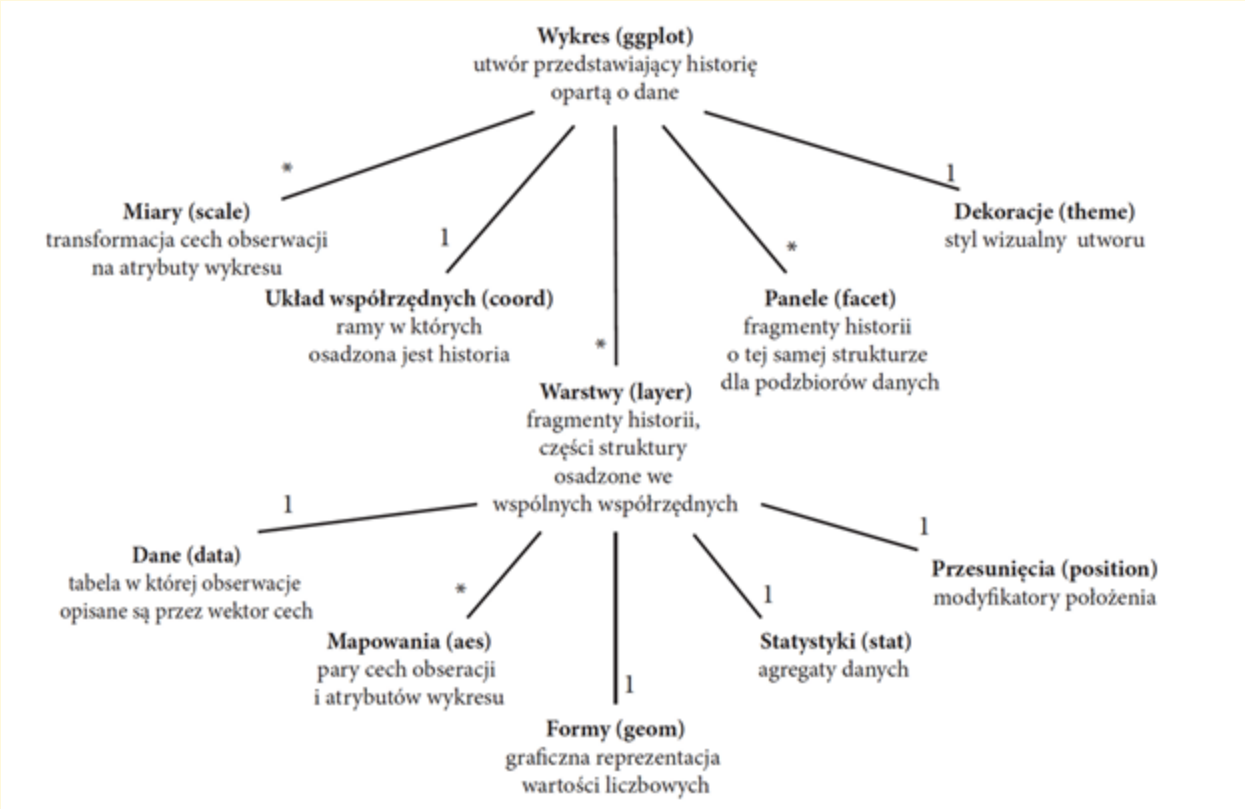
http://www.biecek.pl/Eseje/indexGramatyka.html

Warstwy - Mapowania

Mapowania (aes): pary cech obserwacji i atrybutów wykresu.



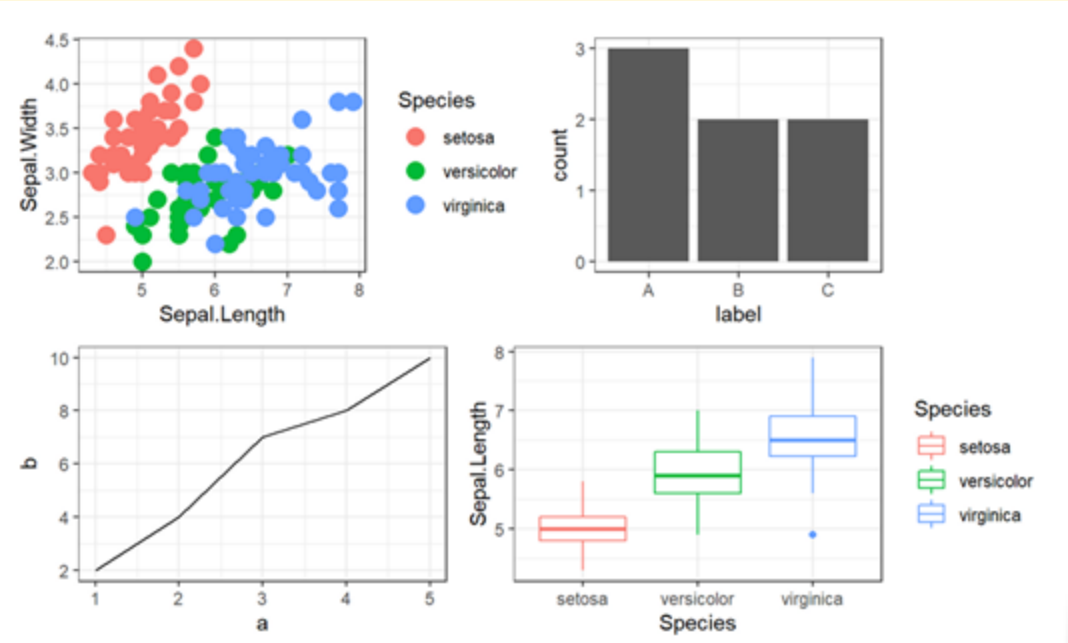
Warstwy - Formy



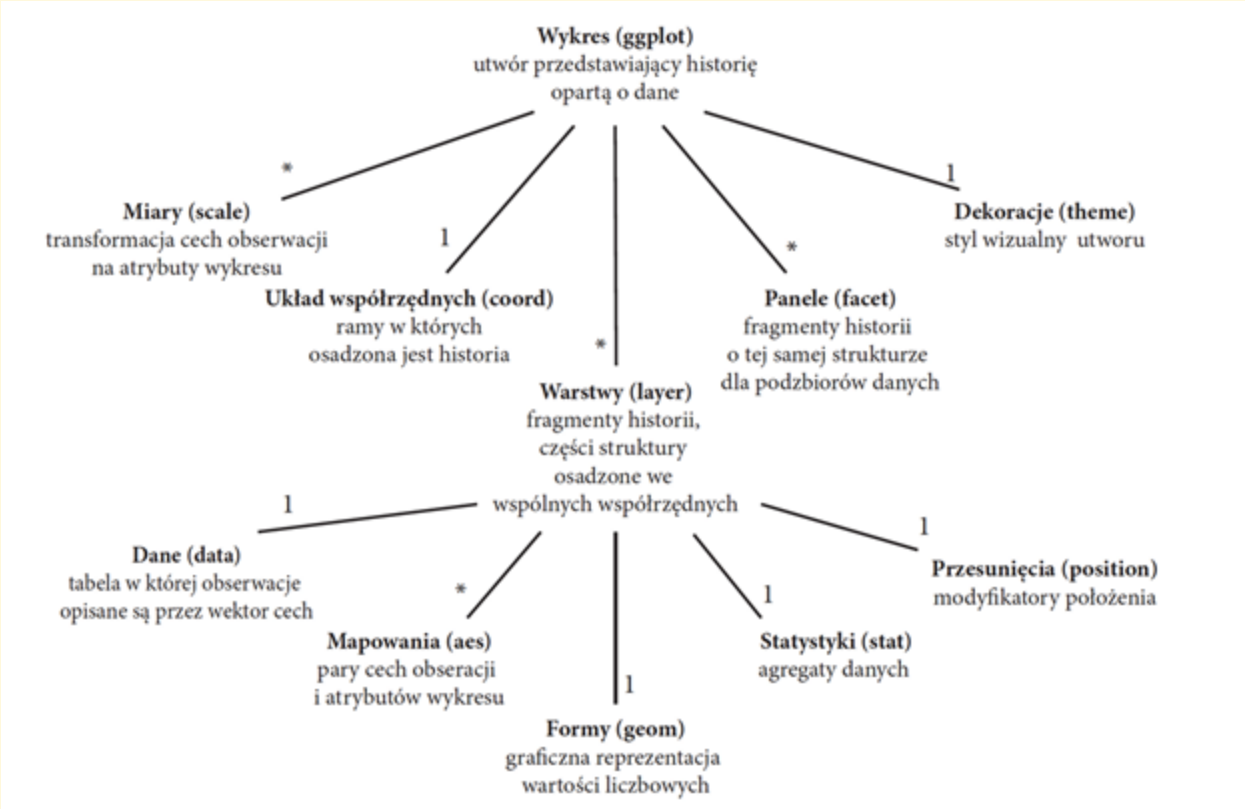
http://www.biecek.pl/Eseje/indexGramatyka.html

Warstwy - Formy

Formy (geom): graficzna reprezentacja wartości liczbowych



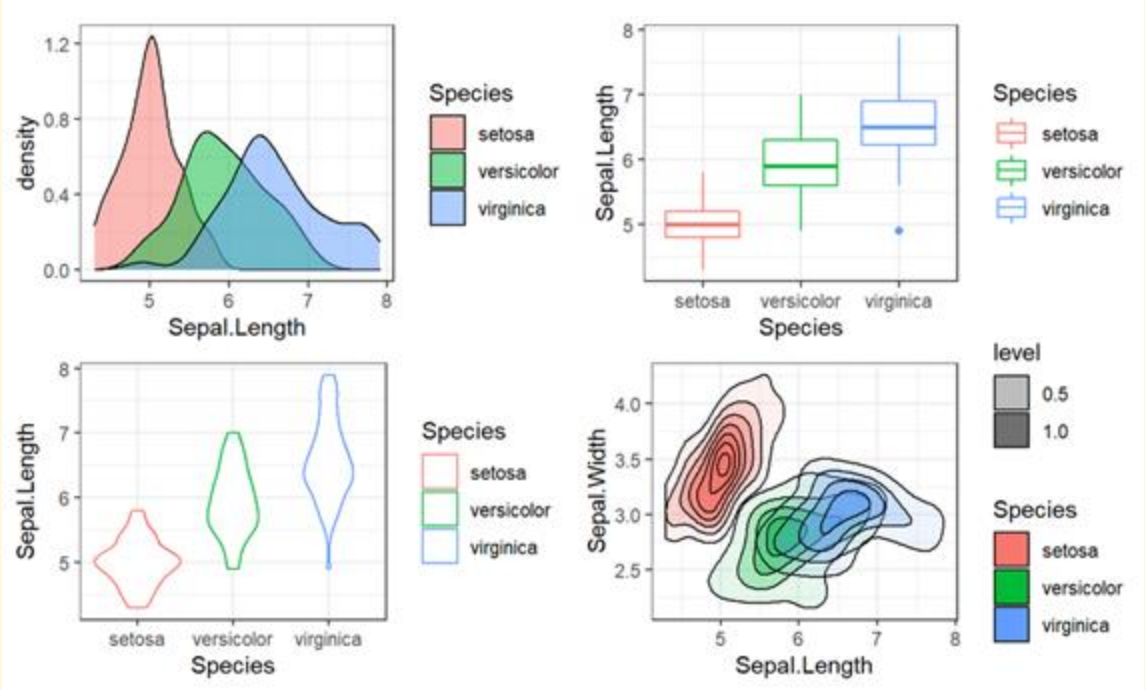
Warstwy - Statystyki



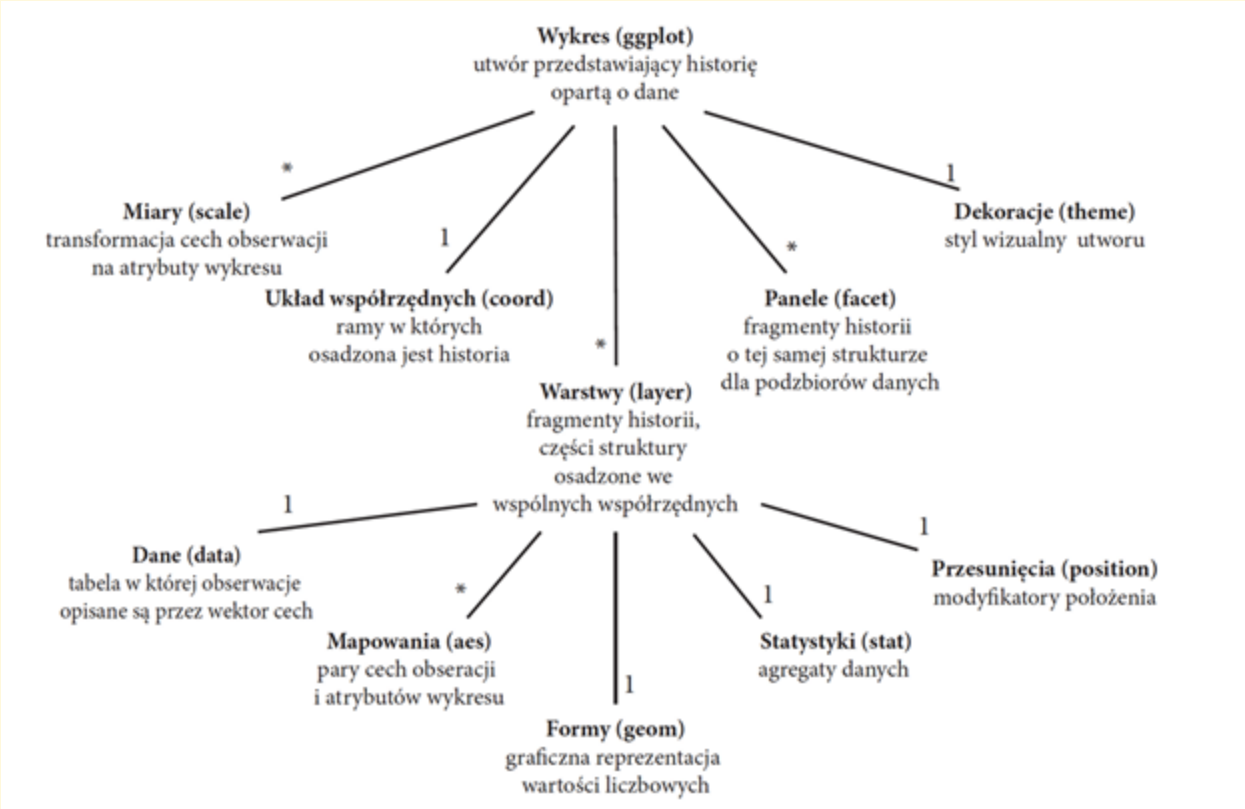
http://www.biecek.pl/Eseje/indexGramatyka.html

Warstwy - Statystyki

Statystyki (stat): agregaty danych



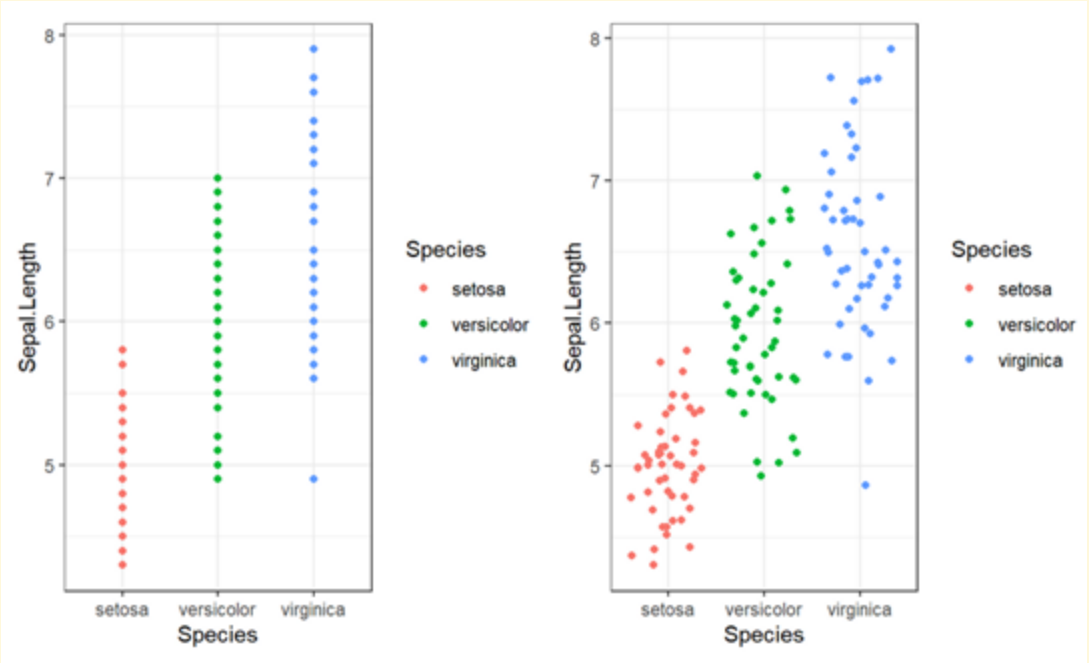
Warstwy - Przesunięcia



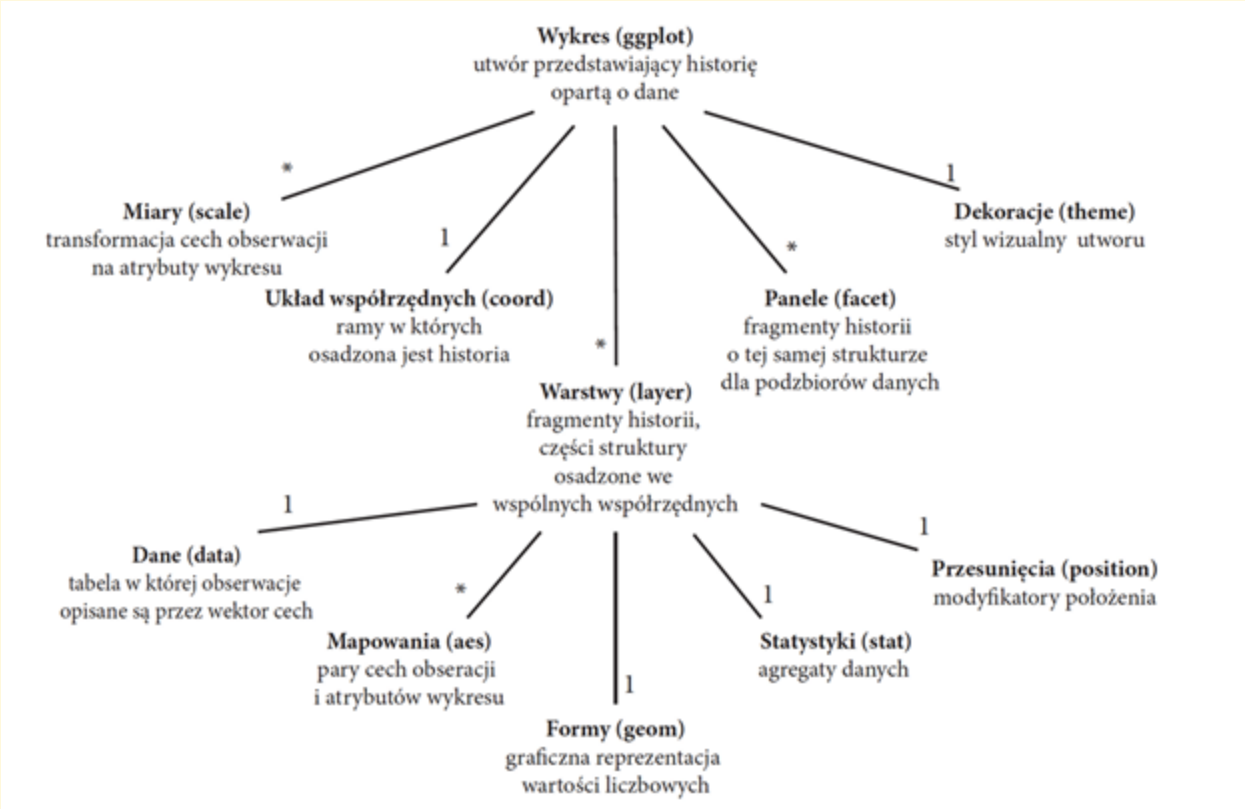
http://www.biecek.pl/Eseje/indexGramatyka.html

Warstwy - Przesunięcia

Przesunięcia (position): modyfikatory położenia



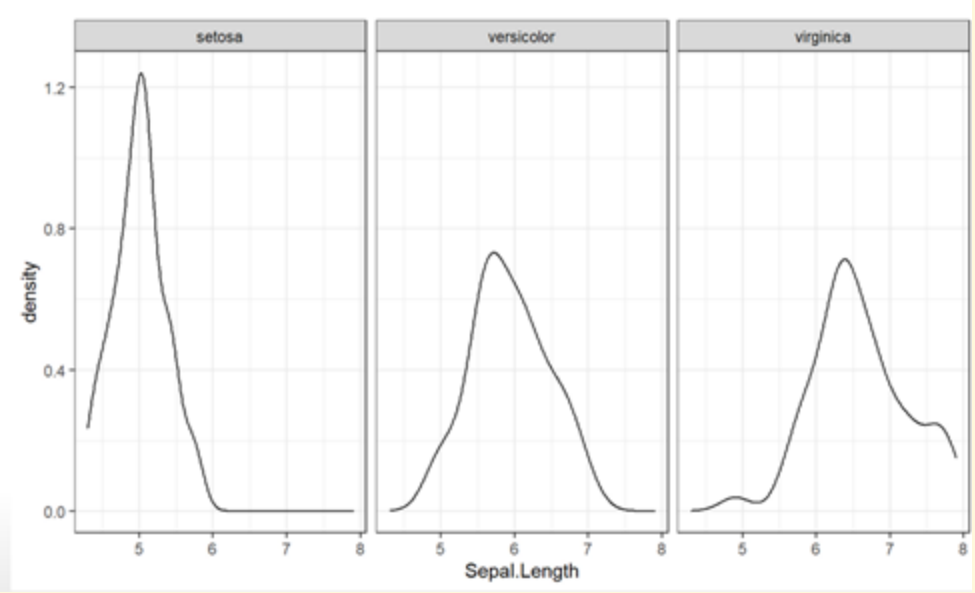
Warstwy - Panele



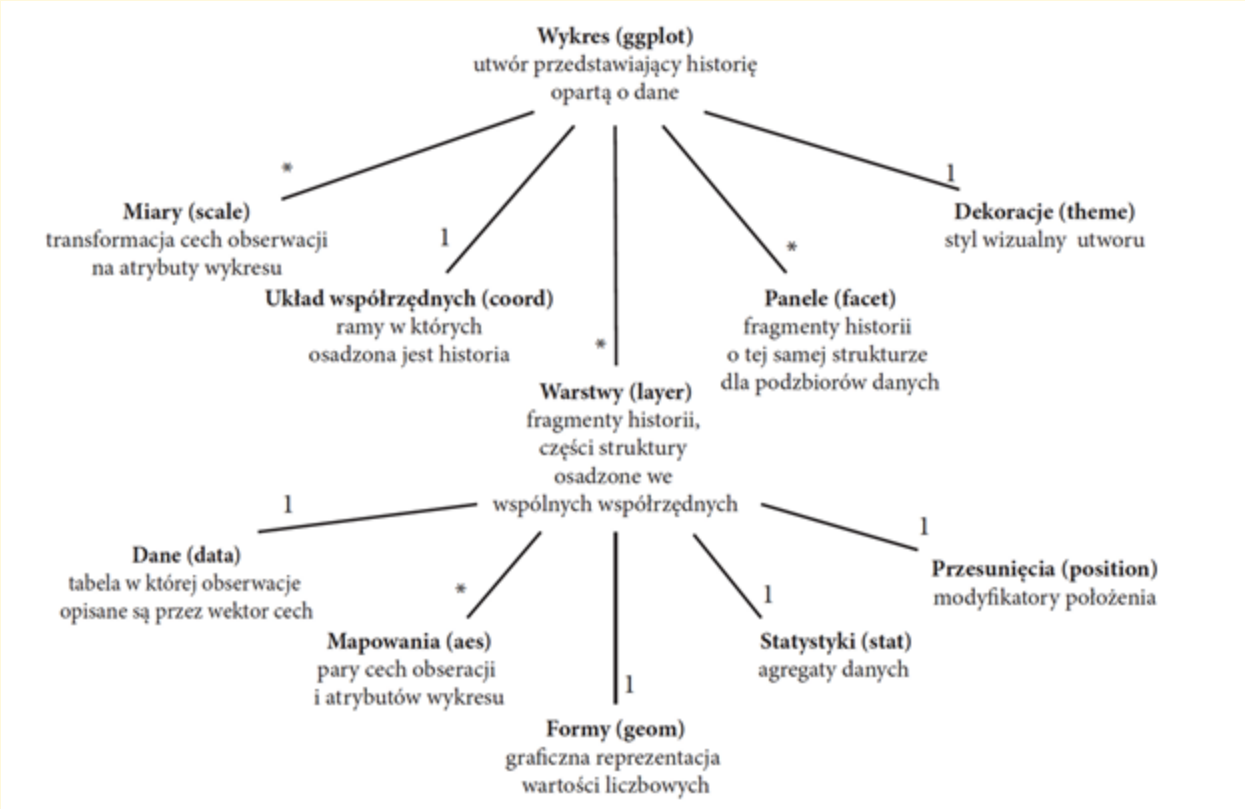
http://www.biecek.pl/Eseje/indexGramatyka.html

Warstwy - Panele

Panele (facets): fragmenty historii o tej samej strukturze dla podzbiorów danych.



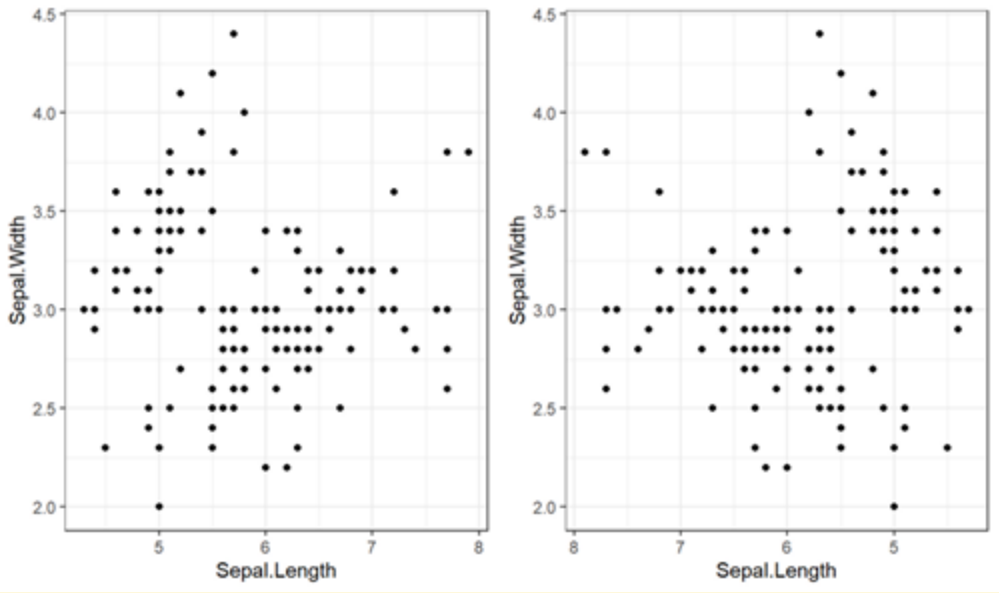
Skale



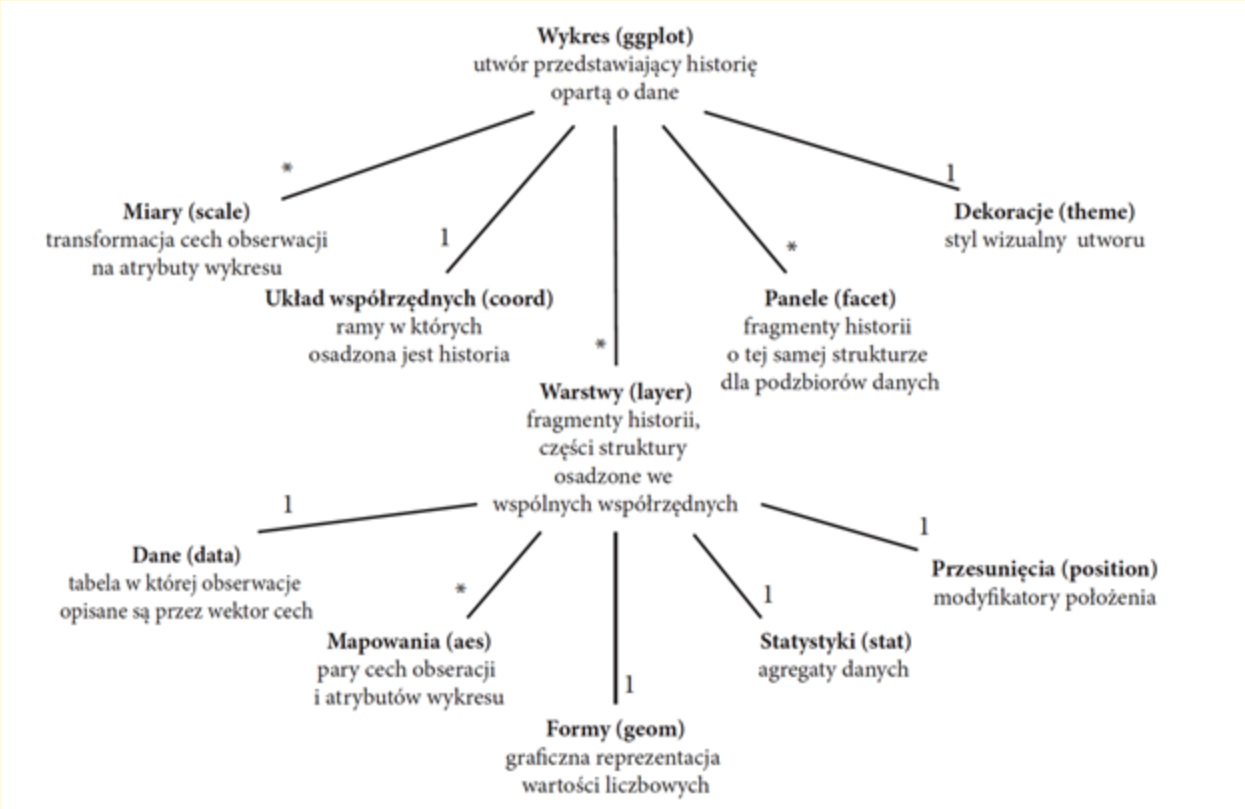
http://www.biecek.pl/Eseje/indexGramatyka.html

Skale

Skale (scale): transformacja cech obserwacji na atrybuty wykresu.



Dekoracje



http://www.biecek.pl/Eseje/indexGramatyka.html

Dekoracje

Dekoracje (theme): styl wizualny.