Struktury danych i typy atrybutów

Jarosław Jasiewicz

Eksploracja danych i Uczenie maszynowe

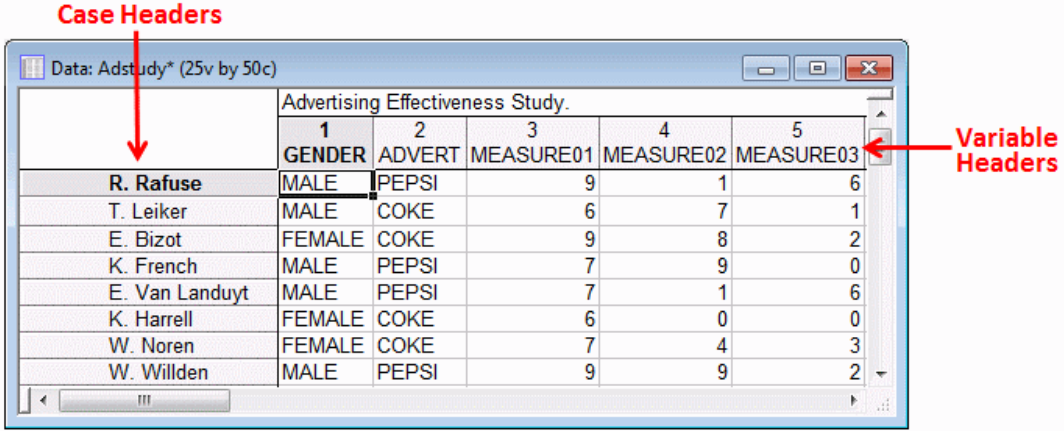
Geoinformacja program magisterski
Specjalność Geoinformatyka

Co to są dane?

- Kolekcja **obiektów** i ich **atrybutów**, gdzie każdy atrybut jest właściwością lub charakterystyką obiektu:
  - Oko:kolor
  - Osoba:kolor oczu
  - Osoba:wzrost

- Obiekt jest opisywany kolekcją atrybutów
- **Inne nazwy obiektu**: rekord (bazy danych), przypadek, encja (entity), instancja, punkt (analiza wielowymiarowa)

- **Inne nazwy atrybutu**: zmienna, pole (bazy danych), cecha



Atrybuty ciągłe i kategoryzacyjne i dyskretne

- Jakościowe
  - Muszą przybierać wartości ze skończonego zbioru
  - Są kategoriami, ale często są reprezentowane przez liczby całkowite (np. kod pocztowy lub PESEL)
  - Mogą ale nie muszą być uporządkowane
  - Wartości tekstowe są reprezentowane przez liczby całkowite (factor w R)
  - Specjalną odmianą są atrybuty binarne (Prawda/Fałsz)
- Dyskretne
  - Muszą być policzalne (mogą być nieskończone)
  - Reprezentowane są przez liczby całkowite
- Ilościowe
  - Reprezentowane przez liczby rzeczywiste, o nieskończonej ilości pomiędzy dwoma dowolnymi wartościami; w praktyce o skończonej precyzji ograniczonej precyzją zapisu liczb

Typy atrybutów

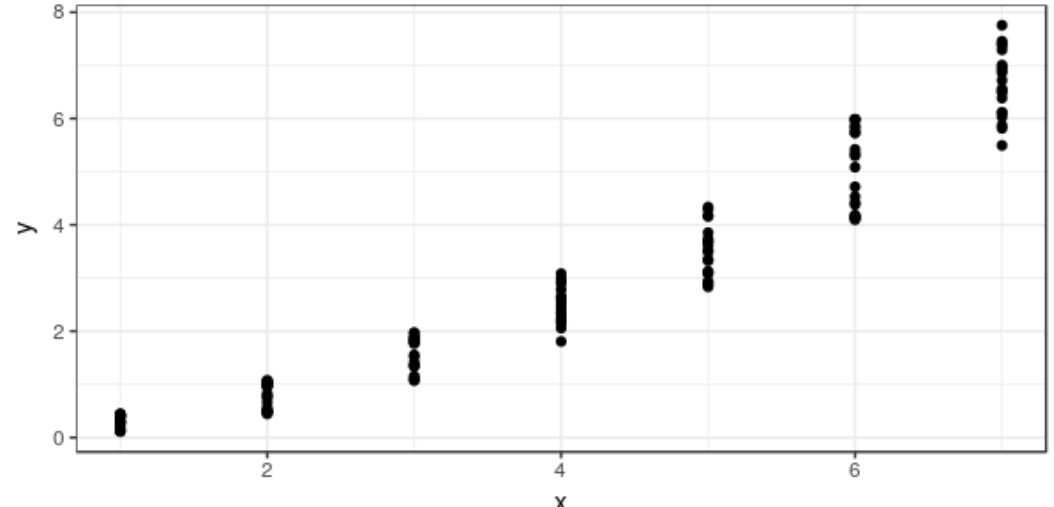
	Typ	Opis	Działania	Operacje	Przykład
Jakościowe kategoryzacyjne	Nominalne (nominal)	Wartości jako różne nazwy lub klasy, pozwalają rozróżniać obiekty między sobą	= ≠	Moda, entropia, Chi-square	Kod pocztowy, PESEL, kolor włosów
	Porządkowe (ordinal)	Wartości jako różne nazwy lub klasy, pozwalają rozróżniać obiekty między sobą, dodatkowo pozwalają uporządkować obiekty względem cechy	= ≠ < >	Mediana,korelacja rang, testy znaków	Twardość minerałów, numery domów, oceny
Ilościowe ciągłe	Interwałowe (interval)	Wartości w skalach umownych. Różnice pomiędzy wartościami są znaczące w obrębie skali	= ≠ < > + -	Mediana, Średnia, SD, korelacja Pearsona, testy F i t	Data (rok), temp. C i F
	Ilorazowe (ratio)	Wartości w skalach bezwzględnych wszystkie operacje są dozwolone i są znaczące	= ≠ < > + - * /	jw. + Średnia geometryczna i harmoniczna	Temp. K, masa, długość, natężenie, czas (S) – jednostki SI

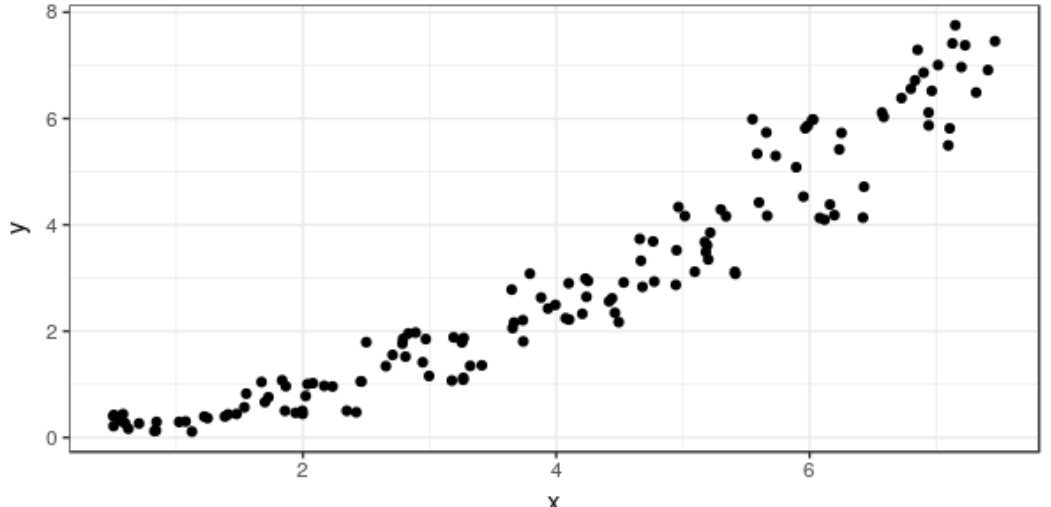
Konwersja pomiędzy atrybutami - dane → dane jakościowe

- Dyskretyzacja – to zamiana zbioru ciągłego na dane jakościowe. Najczęściej wykonuje się tę operację poprzez podział wartości ciągłej na przedziały. Stosuje się metody:
  - Naturalne (Jenks'a) – oparta na minimalizacji wariancji w każdej klasie
  - Grupowanie hierarchiczne
  - Okrągłe (pretty)
  - O stałej wielkości (dla skończonego zbioru wartości)
- Binaryzacja – zamiana wartości ciągłej lub dyskretnej na TRUE/FALSE w zależności od spełniania kryterium logicznego

Konwersja pomiędzy atrybutami - dane → dane ilościowe

- Kwantyfikacja – proces ilościowego ujmowania zjawisk ujętych opisowo. Możliwe do zastosowania dla danych porządkowych, jeżeli znane są relacje pomiędzy cechą jakościową a przedziałem wartości
- Rozrzucanie (jitter) – jest operacją stosowaną dla liczb całkowitych, jej celem jest usunięcie efektu skupiania na okrągłych wartościach. Stosuje się głównie w celach wizualnych
- Zaokrąglanie – operacje wykonywane na liczbach zmiennoprzecinkowych. Polega na usunięciu informacji zawartej w mniej znaczących części liczby. Jest odwrotnością rozrzucania





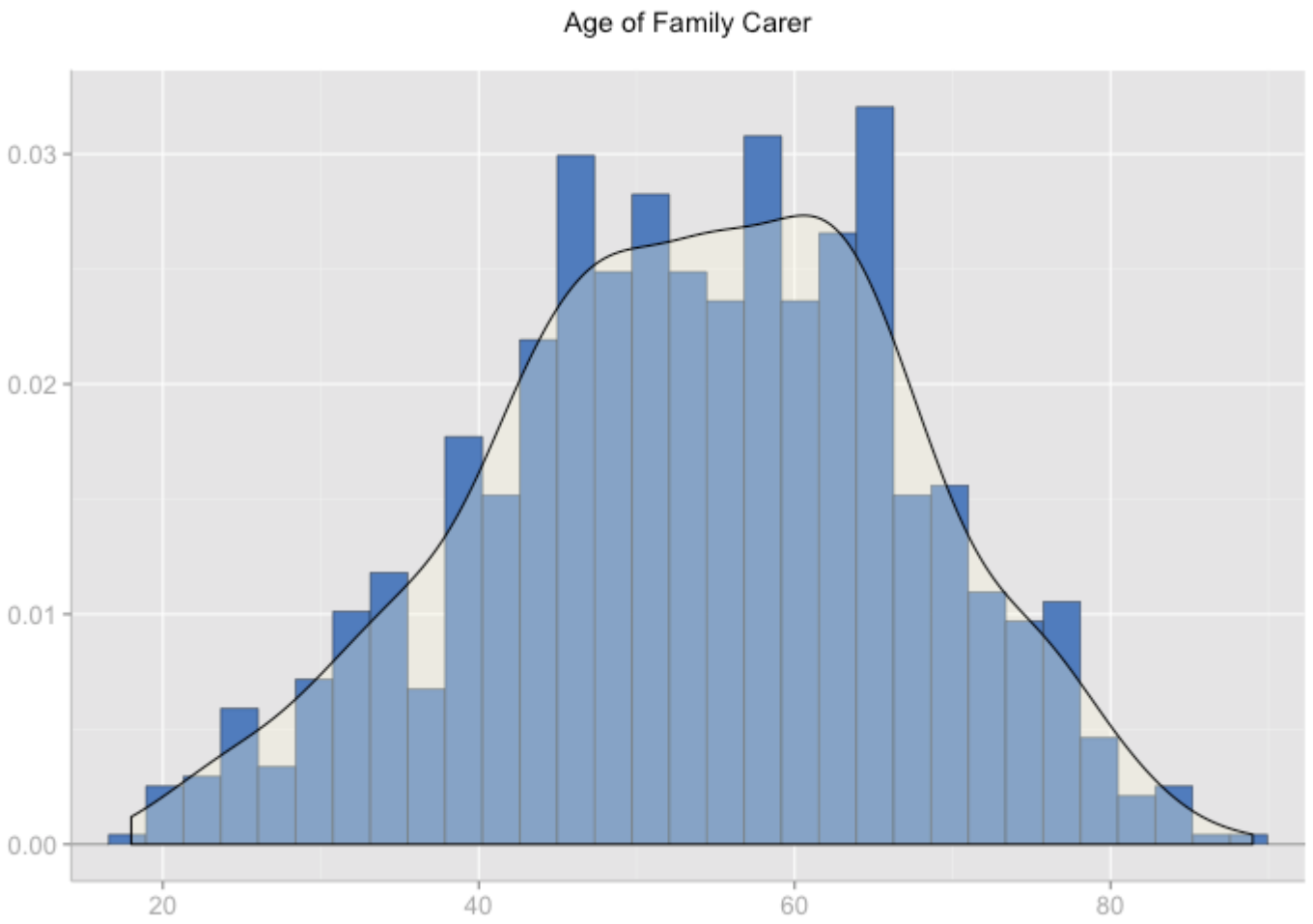
Dziedziny atrybutów

- Dziedzina to ograniczenie, które decyduje jakie wartości może przyjmować atrybut
- W przypadku atrybutów jakościowych jest to **lista klas** jakie może przyjąć atrybut: **{mały, średni, duży}**
- Dla atrybutów binarnych jest to **{Prawda - Fałsz}**
- W przypadku atrybutów ilościowych jest to zakres (przedział) jaki może przyjąć atrybut. Przedział może być jednostronny **(x,∞)** lub dwustronny **(x,y)**. Przedziały mogą być domknięte **[x,y]**, tj wartość graniczna jest częścią dziedziny lub niedomknięte, wtedy warość graniczna nie jest częścią dziedziny **(0,x]**. Na przykład wartość atrybutu musi być większa od 0

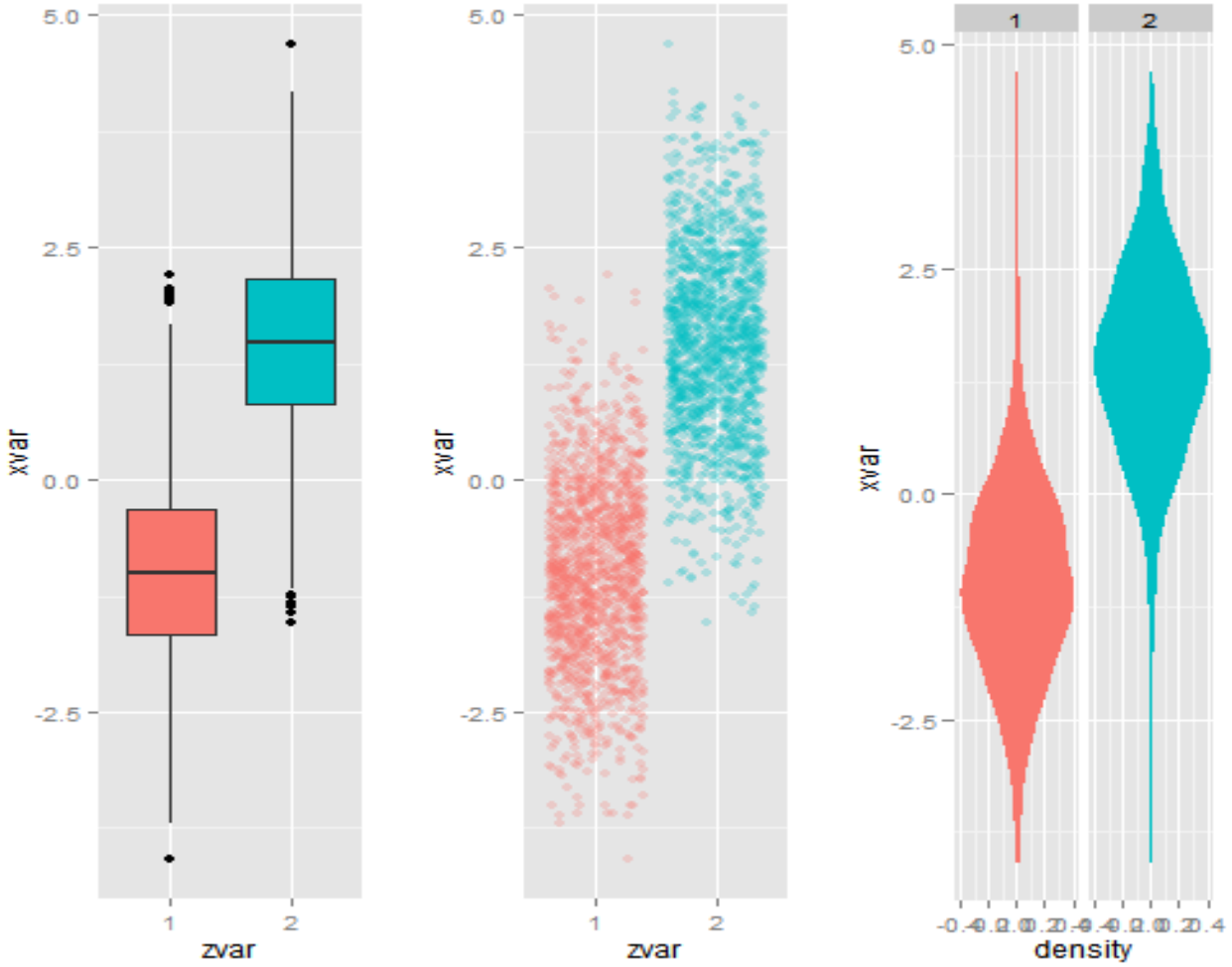
Wizualizacja atrybutów

- W zależności od typów danych stosuje się różne metody ich wizualizacji
- Dla danych ciągłych:
  - Histogramy
  - Wykresy gęstościowe
  - Wykresy pudełkowe
- Dla danych dyskretnych
  - Wykresy słupkowe
- Dla wszystkich typów:
  - Koordynaty równoległe

Dane ciągłe – histogram i krzywa gęstościowa

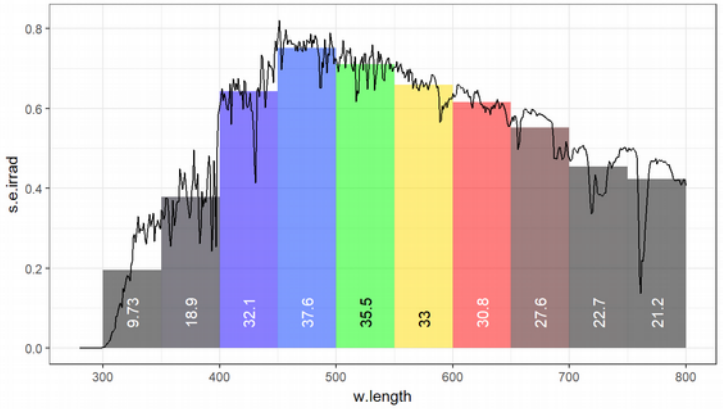


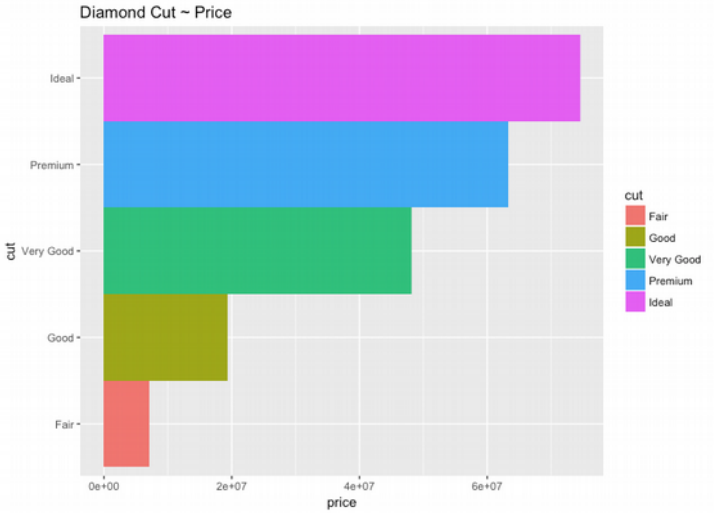
Dane ciągłe Wykres pudełkowy i jego odmiany



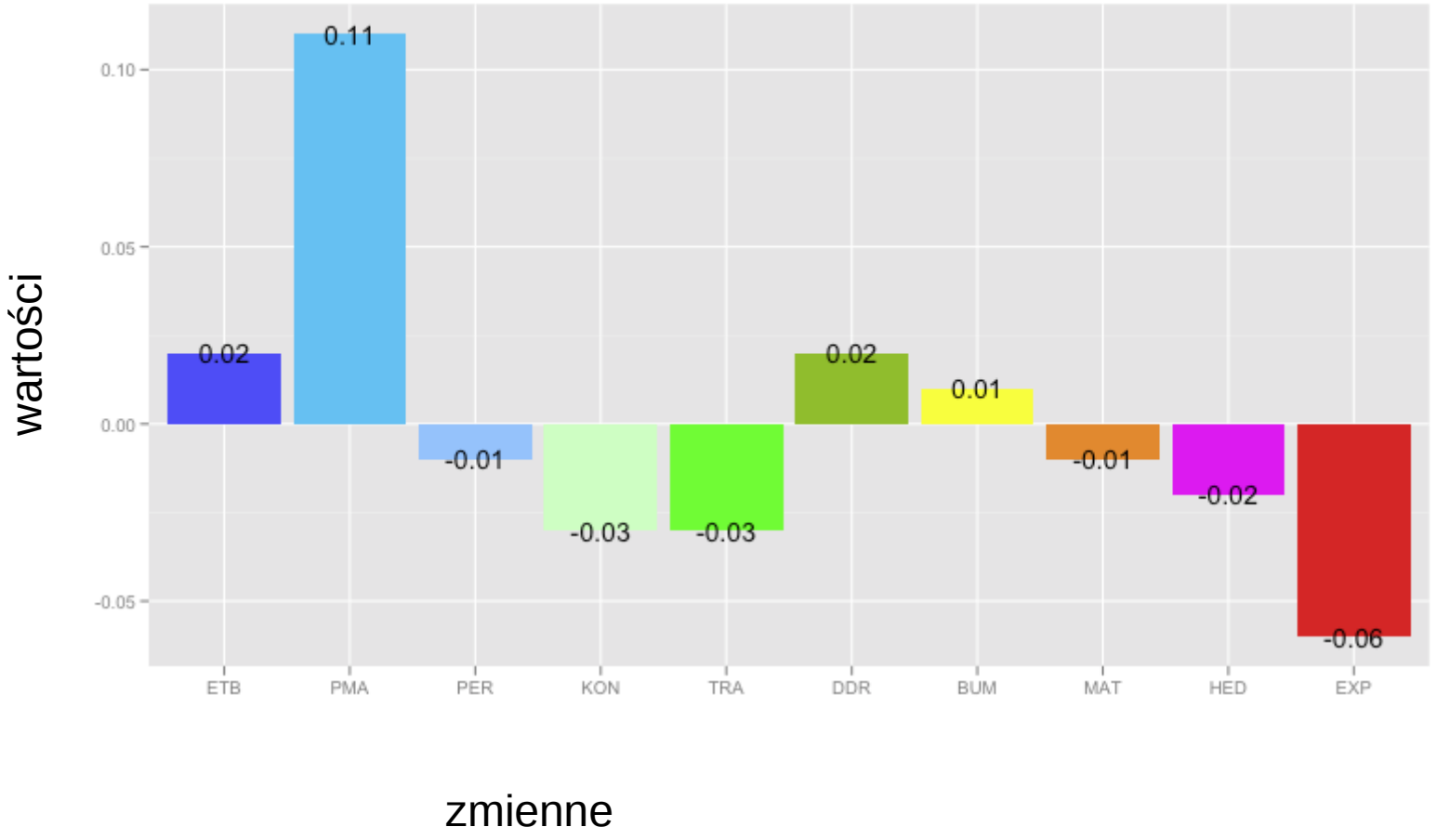
Atrybuty dyskretne

- Atrybuty dyskretne przedstawia się za pomocą wykresów słupkowych, gdzie każdy słupek pokazuje udział danej klasy w zmiennej
- Jeżeli zmienna jest porządkowana powinno się stosować stopniowaną skalę barwną



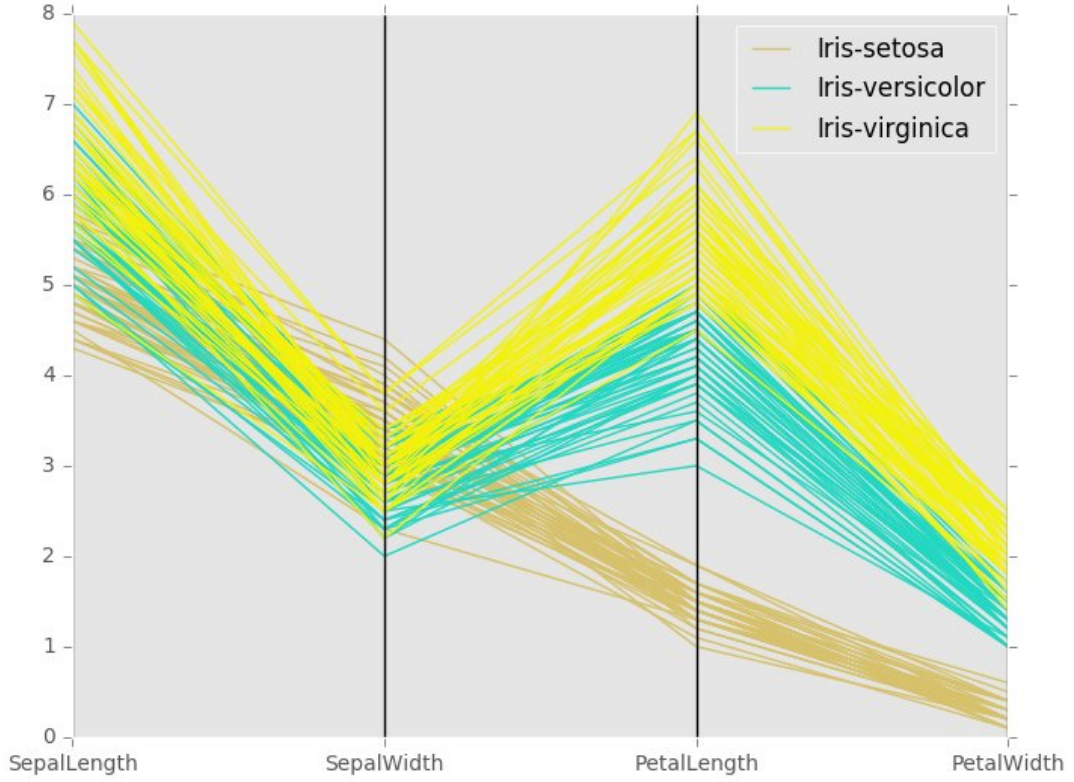


Wizualizacja wartości atrybutów jednego obiektu



Porównanie zasięgu zmiennych dla różnych obiektów

- Koordynaty równoległe
- Zmienne na osi X, każda linia reprezentuje osobny obiekt



Co wpływa na jakość danych?

- Znaczenie (relevance)
  - Zrozumiałość
  - Dostępność
  - Kompletność (ilość cech)
  - Spójność
- Problemy z danymi
  - Braki danych
  - Szumy
  - Wartości odstające
  - Dokładność
  - Błędy
  - Duplikaty
  - Integralność

- Jak wykryć problemy w danych?
- Co można z problematycznymi danymi zrobić

Przygotowanie do analizy

- Dane w postaci surowej z reguły nie nadają się do przetwarzania. **90%** czasu w Data Science zajmuje „użeranie się” z danymi (*wrangling*)
- Preprocessing obejmuje:
  - Usuwanie braków
  - Odszumienie
  - Usunięcie obserwacji odstających
  - Usunięcie błędów i duplikatów

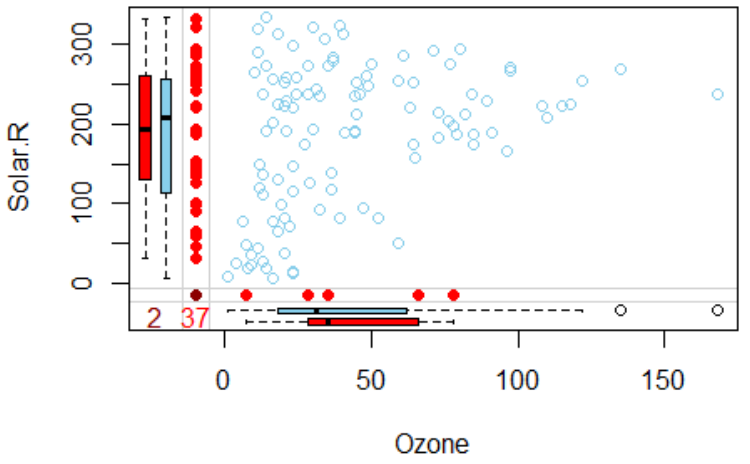
GARBAGE IN – GARBAGE OUT

Braki danych

- Źródła braków
  - Brak pomiaru (np. odmowa odpowiedzi, awaria czujnika)
  - Utrata wyniku (dane archiwalne)
  - Nieistotność cechy (np. ciąża u mężczyzny, zarobki u dziecka)
- Zarządzanie brakami danych
  - Usuwanie przypadków lub całych atrybutów
  - Szacowanie (wstawianie) brakującej wartości
  - Ignorowanie w czasie analizy

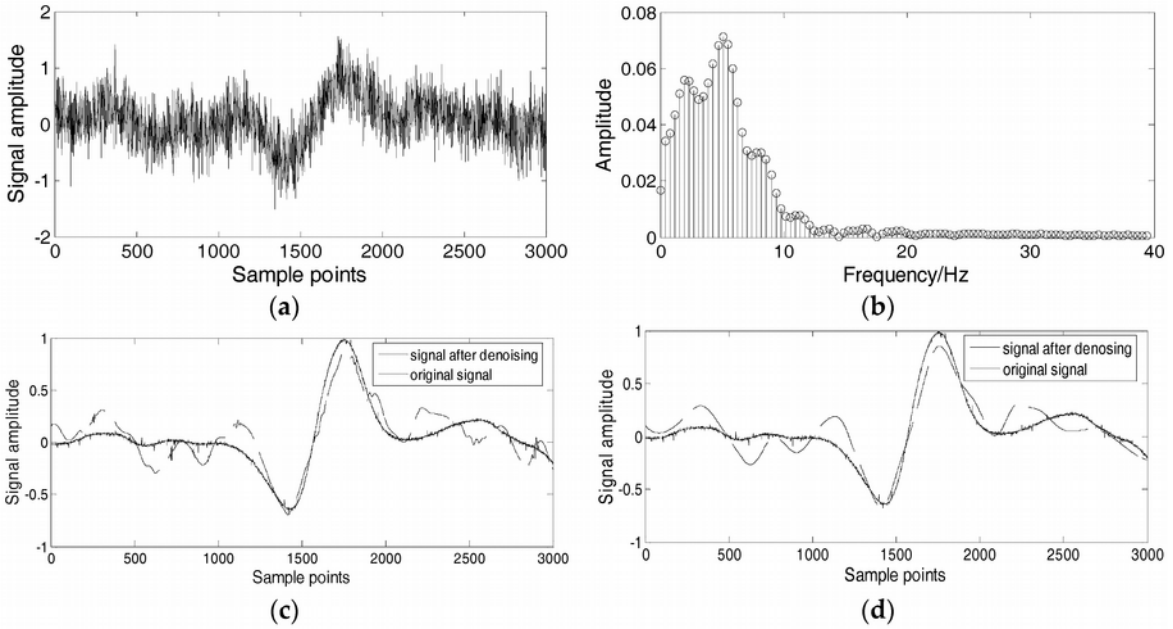
Metody uzupełniania braków

- Uzupełnienie pomiaru – jeżeli atrybut jest niezmienny w czasie (dodaje brakującą informację)
- Imputacja – uzupełnianie braków na podstawie kryteriów matematycznych
- Imputacja to podejście pragmatyczne, nigdy nie dodaje informacji
- Można je stosować jeżeli braki są **losowe**
  - Uśredniona wartość atrybutu
  - Modelowanie wartości na postawie innych wartości (modelowanie predykcyjne)
  - Zero lub ε (jeżeli brak jest spowodowany czułością urządzenia
  - Wartość losowa z dziedziny



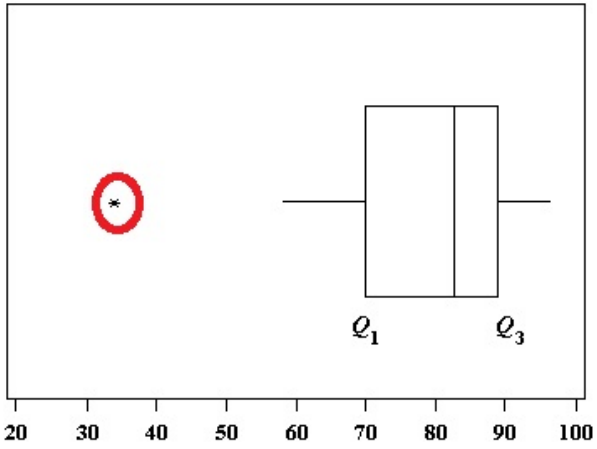
Szumy

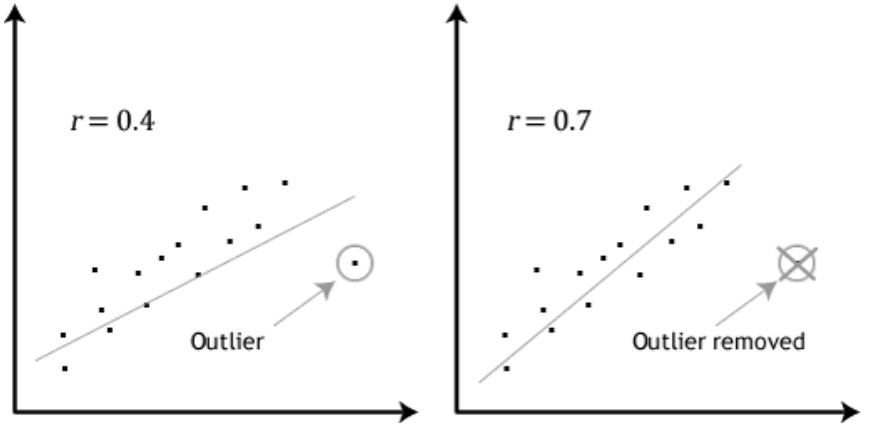
- Szum - losowe, nadmiarowe, nie interpretowane zmiany wartości atrybutu, mogące wpływać na wynik analizy. Mogą być wynikać ze zbyt dużej czułości sensora mieć charakter przypadkowy
- Szumy usuwa się poprzez operacje wygładzania wartości atrybutu (tylko dla danych uporządkowanych)



Wartości (obserwacje) odstające

- Punkty lub wartości odległe od pozostałych obserwacji/modelu
- Mogą być wynikiem błędu pomiaru lub być wynikiem zarejestrowania unikalnego obiektu
- Dla pojedynczych atrybutów są to wartości w dużej odległości od średniej (min: **2x SD**)
- Obiekty odstające usuwa się danych, wartości odstające, jeżeli uważamy że są efektem błędu można usunąć i potraktować jak braki w danych





Błędy i duplikaty

- Sytuacje, gdy obiekt został zarejestrowany wielokrotnie
- **Pełne duplikaty** – wszystkie atrybuty (w tym identyfikator) mają taką samą wartość
- **Częściowe duplikaty** – niektóre wartości się różnią pomimo że jest to ten sam obiekt (np. różne adresy e-mail tej samej osoby)
- **Pseudo-duplikaty** – różne obiekty (różne identyfikatory) posiadają ten sam zestaw wartości (np. temperatura powietrza, wilgotność i opad mogą się powtarzać)
- Pełne i częściowe duplikaty usuwa się/łączy. Pseudo-duplikaty rozdziela się dodając niewielką wartość losową.

Organizacja danych

- Dane rekordowe
  - Ramki danych/macierze
  - Macierze komplementarne (histogramy)
  - Wektory cech
  - Transakcje i zbiory
- Dane sekwencyjne/uporządkowane (w jednym lub więcej wymiarów)
  - Serie czasowe (1 wymiar)
  - Obrazy i rastry (2 wymiary)
  - Dane czasowo-przestrzenne (3 wymiary)
  - Sekwencje np. genów
- Grafy
  - WWW
  - Molekuły (cząsteczki chemiczne)

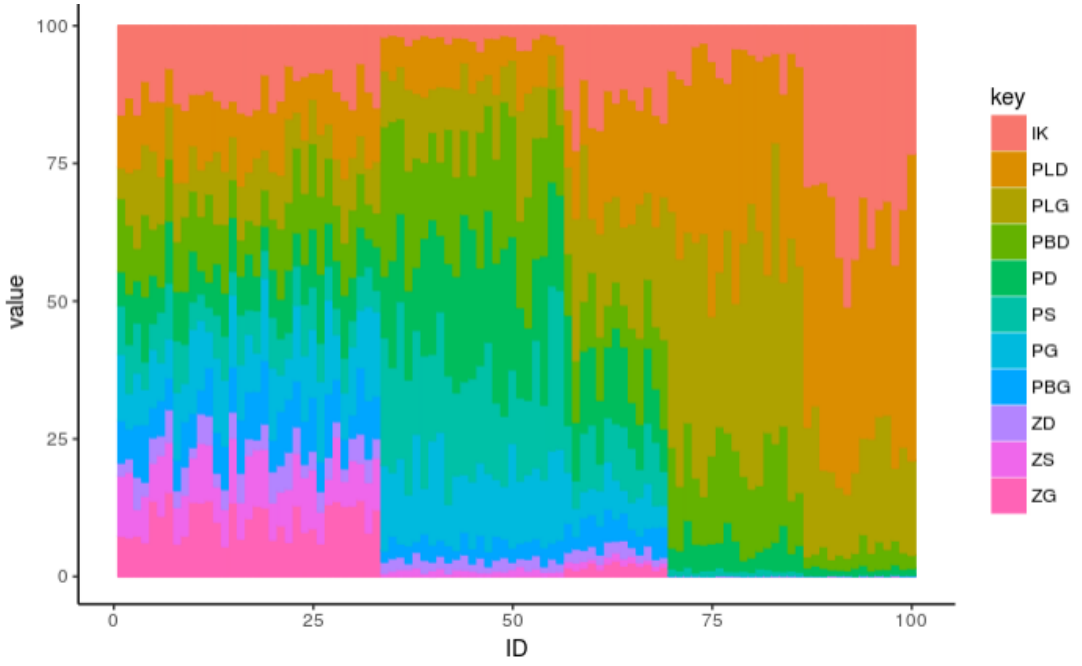
Dane rekordowe – ramki danych

- Klasyczny sposób strukturalizacji danych
- Kolejność rekordów i atrybutów jest dowolna
- Każdy rekord (wiersz) to stały zbiór atrybutów
- Atrybuty mogą być różnych typów (nominalne, ilorazowe, wskaźnikowe itp)
  - Jeżeli atrybuty są typu ilorazowego – macierze, gdzie każdy wiersz to punkt w wielowymiarowej przestrzeni
- Formalny brak związków między atrybutami
- Klasyczna tabela w I postaci normalnej, każdy zbiór danych można sprowadzić do tej postaci
- Przykład: dane socjo-ekonomiczne krajów europejskich

	AREA	POPTL	POPGR	POPDN	LIFEXP	FERTR	MRTCH	SCHCNT	URBGR	GDPGR	INFLT	GNIPK	GDPTL
ALBN	28750	2889104	-0.2	105.4	77.8	1.8	14.4	96.4	1.6	1.8	1.4	4450	1.321986e+10
AUST	83879	8541575	0.7	103.5	81.5	1.5	3.7	99.3	0.8	0.6	1.8	50150	4.383762e+11
BLRS	207600	9474511	0.1	46.7	73.0	1.7	4.7	107.0	0.6	1.7	18.1	7600	7.881305e+10
BELG	30530	11209057	0.2	370.2	81.3	1.7	4.2	164.8	0.3	1.6	0.7	46930	5.317509e+11
BOSN	51210	3566002	-1.1	69.6	76.4	1.3	5.7	92.1	-0.8	1.1	1.0	5160	1.852148e+10
BULG	111000	7223938	-0.6	66.5	74.5	1.5	10.9	100.9	-0.1	1.3	0.5	7720	5.673201e+10
CRAT	56590	4238389	-0.4	75.7	77.5	1.5	4.5	99.0	0.1	-0.5	0.0	13150	5.708037e+10
CYPR	9250	1152309	0.7	124.7	80.1	1.4	2.9	99.4	0.6	-1.5	-1.5	26500	2.330821e+10
CZCH	78870	10525347	0.1	136.3	78.8	1.5	3.5	105.1	0.0	2.7	2.5	18790	2.078183e+11
DENM	42922	5643475	0.5	133.5	80.7	1.7	3.6	129.9	0.7	1.7	0.8	63640	3.522970e+11
ESTN	45230	1314545	-0.3	31.0	77.0	1.5	3.2	110.4	-0.4	2.8	1.7	18790	2.621394e+10
FINL	338420	5461512	0.4	18.0	81.2	1.7	2.4	145.5	0.6	-0.6	1.7	48990	2.726093e+11
FRAN	549087	66331957	0.5	121.1	82.7	2.0	4.4	110.6	0.8	0.9	0.6	42750	2.849305e+12
GERM	357380	80982500	0.4	232.1	81.1	1.5	3.8	102.4	0.7	1.6	1.8	47680	3.879277e+12
GERC	131960	10892413	-0.7	84.5	81.4	1.3	4.6	106.5	-0.2	0.4	-1.8	22000	2.360798e+11
HUNG	93030	9866468	-0.3	109.0	75.8	1.4	6.1	107.0	0.4	4.0	3.4	13460	1.392946e+11
ICLD	103000	327386	1.1	3.3	82.9	1.9	2.1	108.2	1.2	1.9	4.1	48160	1.717896e+10
IRLD	70280	4617225	0.4	67.0	81.3	1.9	3.7	127.2	0.9	8.5	-1.2	47040	2.562713e+11
ITAL	301340	60789140	0.9	206.7	83.1	1.4	3.6	102.6	1.1	0.1	1.0	34760	2.151733e+12
LATV	64490	1993782	-0.9	32.1	74.1	1.6	8.2	115.4	-1.0	2.1	1.6	15330	3.135225e+10
LITH	65286	2932367	-0.9	46.8	74.5	1.6	5.2	106.8	-0.9	3.5	1.0	16030	4.854525e+10
LUXM	2590	556319	2.4	214.8	82.2	1.5	2.0	102.3	2.7	5.6	1.6	76900	6.629806e+10
MACD	25710	2077495	0.1	82.4	75.3	1.5	6.0	78.6	0.2	3.6	1.4	5200	1.136227e+10
MOLD	33850	3556397	-0.1	123.9	71.5	1.3	16.1	87.3	0.0	4.8	6.4	2560	7.983271e+09
MONT	13810	621810	0.1	46.2	76.2	1.7	5.0	93.8	0.4	1.8	1.0	7320	4.587929e+09
NETH	41540	16865008	0.4	500.6	81.7	1.7	3.9	132.3	1.1	1.4	0.1	51330	8.796351e+11
NORW	385178	5137232	1.1	14.1	82.1	1.8	2.7	112.6	1.5	1.9	0.3	104860	4.983398e+11
POLD	312680	38011735	-0.1	124.1	77.6	1.3	5.2	108.1	-0.2	3.3	0.5	13630	5.451518e+11

Dane rekordowe – macierze komplementarne

- W **macierzy komplementarnej** każdy rekord zawiera wartości typu proporcjonalnego (ratio). Suma wartości w każdym rekordzie sumuje się do jedności (100%). Przykład skład mineralny skały, granulometria



Dane rekordowe - wektory cech

- **Wektory cech:** Każdy rekord to sumy wystąpień poszczególnych encji (np. słów w tekście), może być w postaci bezwzględnej lub względnej, tj. sumuje się do jedności co pozwala porównywać obiekty różnej wielkości
- Są to bardzo długie wektory rzadkie

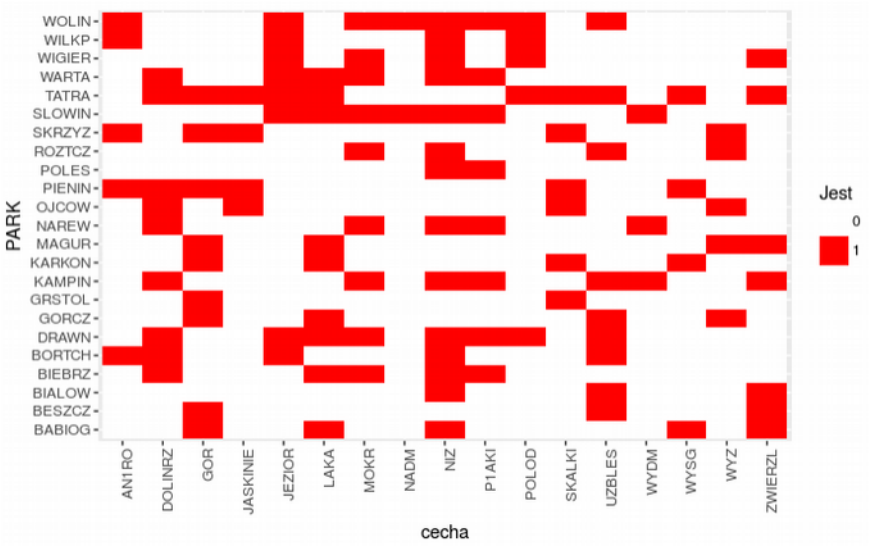
		(Var 1)	(Var 2)	(Var 3)	(Var 4)	(Var 5)	(Var 6)	...	(Var 4999)	(Var 5000)
		apple	cat	cats	dog	dogs	farm	...	White House	Senate
(Obs 1)	Doc 1	1	1	2	2	0	1	...	0	0
(Obs 2)	Doc 2	0	1	0	1	1	0	...	3	2
(Obs 3)	Doc 3	0	1	0	0	1	0	...	4	4
(Obs ...)	...	...	...	...	...	...	...	...	...	...
(Obs N)	Doc N	2	2	2	3	0	1	...	0	0

Dane rekordowe - transakcje

- Każdy rekord zawiera **zbiór** elementów: przykład kosz zakupów. W zbiorze elementy nie mogą się powtarzać, stosuje się przede wszystkim w analizie asocjacji, oraz do zbiorów cech binarnych
- Istotą zbiorów jest to że każdy element występuje tylko raz
- W praktyce jest to tzw. macierz rzadka

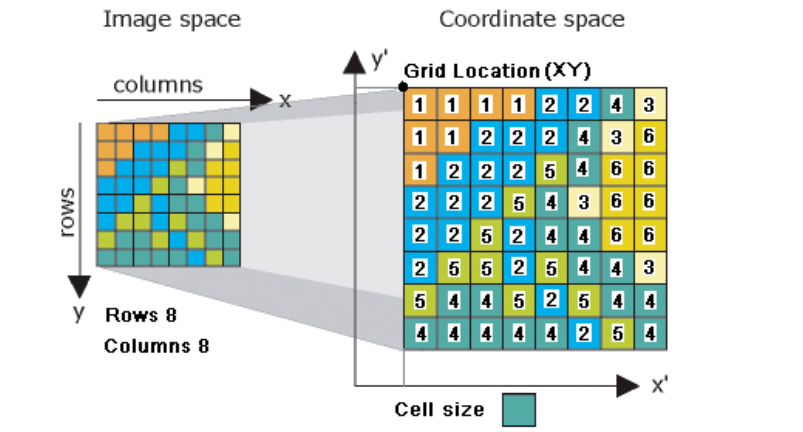
TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

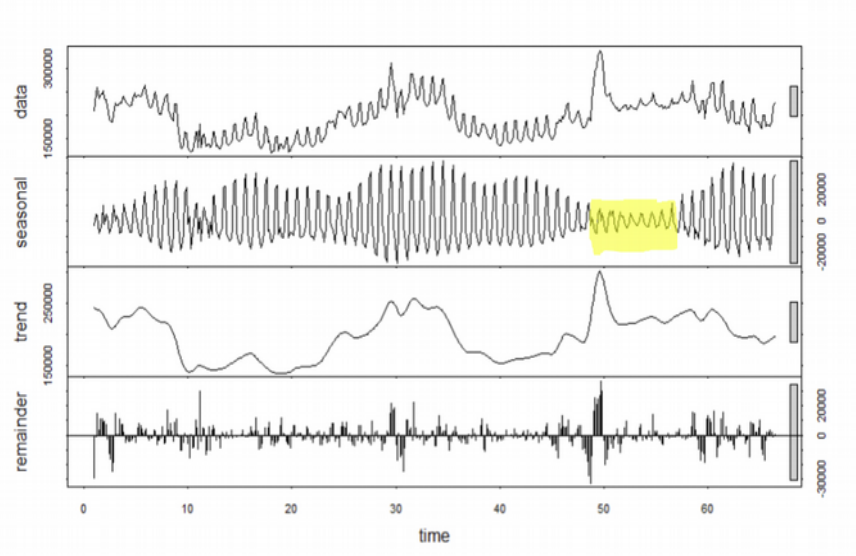
	Beer	Bread	Milk	Diaper	Eggs	Coke
T_1	0	1	1	0	0	0
T_2	1	1	0	1	1	0
T_3	1	0	1	1	0	1
T_4	1	1	1	1	0	0
T_5	0	1	1	1	0	1



Dane porządkowe

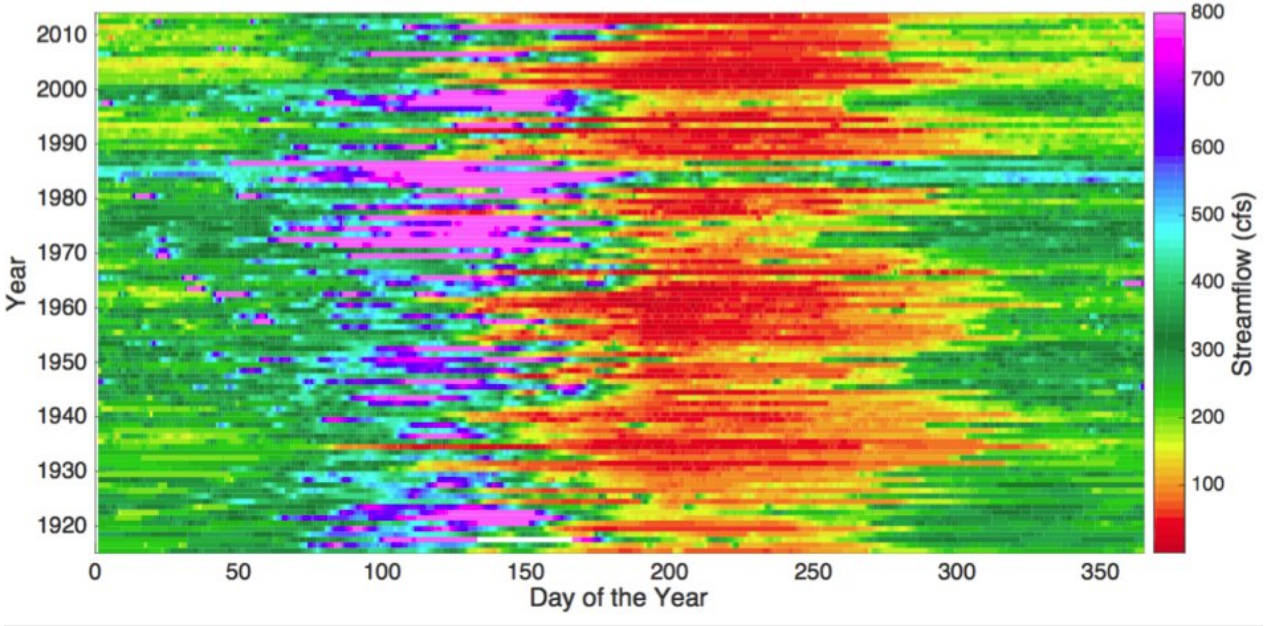
- Dane prządkowe w jednym, dwóch czy trzech wymiarach charakteryzują się tym, że każda wartość znajduje się w określonym położeniu
- Uporządkowanie dostarcza dodatkowych informacji, które mogą zostać zamienione w atrybuty (np. sąsiedztwo, autokorelacja itp.)
- Dane uporządkowane są bardziej odporne na braki danych, ich usuwanie opiera się bardziej na analizie sąsiedztwa czy trendów, co nie jest możliwe w danych nieuporządkowanych





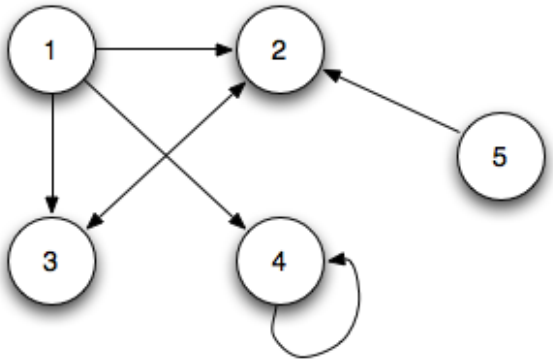
Położenie jako atrybut

- Istotą danych uporządkowanych jest to że każdy obiekt oprócz wartości ma również położenie, które również może być traktowane jako atrybut(y)
- Na przykład obrót sklepu może być opisany wielkością obrotu, rodzajem kupowanych towarów, ale również położeniem na osi czasu: kolejnego dnia, dnia tygodnia czy pory dnia, te wartości mogą być mieć charakter cykliczny



Grafy, sieci i molekuły

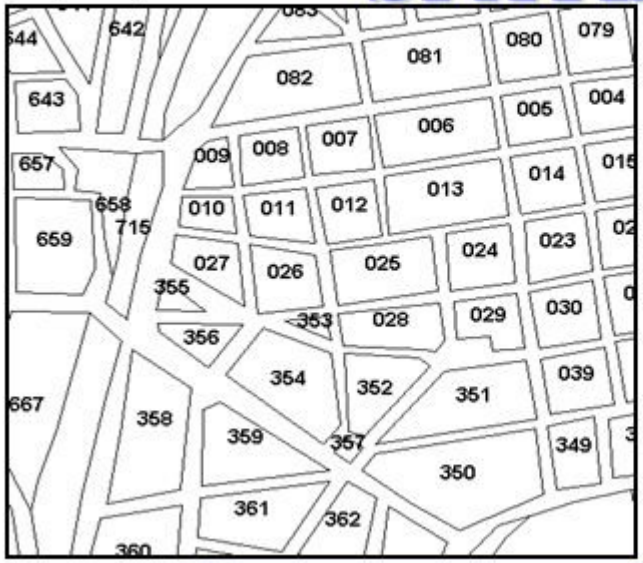
- Struktury danych składające się z węzłów (nodes, vertices) i połączeń (edges) par węzłów.
- Połączenia mogą być uporządkowane, nieuporządkowane oraz mogą mieć wartość (liczbową lub nominalną)
- Grafy można prezentować jako macierze albo listę par węzłów i ich połączeń



	1	2	3	4	5
1	0	1	1	1	0
2	0	0	1	0	0
3	0	1	0	0	0
4	0	0	0	1	0
5	0	1	0	0	0

SPATIAL AND NON-SPATIAL DATA

Map: City blocks



City blocks	Land use
001	Institutional
002	Commercial
003	Commercial
004	Residential
005	Residential
006	Residential
007	Industrial
008	Residential
009	Industrial
010	Industrial
011	Residential
012	Industrial
013	Residential
014	Residential
015	Residential

NON-SPATIAL DATA

SPATIAL DATA

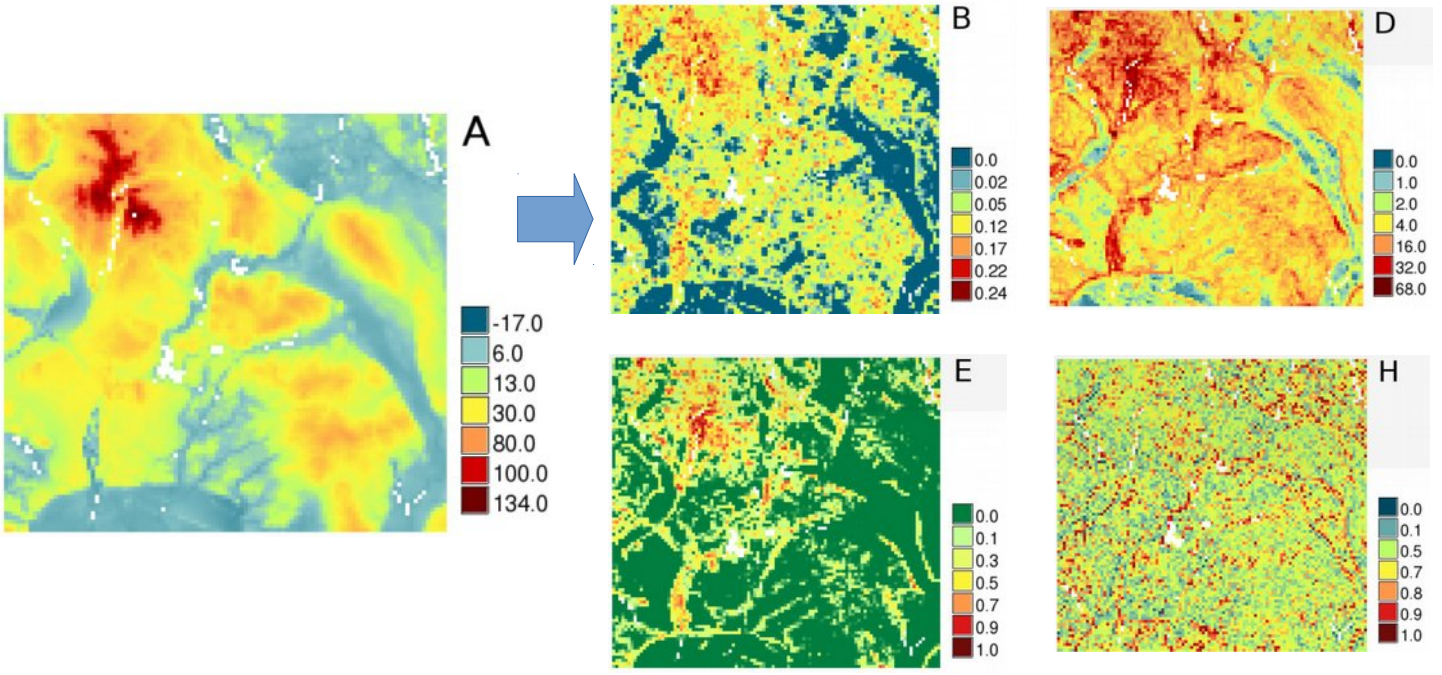
c: Commercial
n: Industrial
i: Institutional
h: Recreational
r: Residential
w: Water

Dane geoprzestrzenne

- Nieprzestrzenne – atrybuty przechowywane w modelu relacyjnym: tabelach atrybutowych lub relacyjnych bazach danych
- Przestrzenne – informacja zawarta w:
  - Lokalizacji (x,y,z)
  - Geometrii (powierzchnia, obwód, kształt, orientacja)
  - Relacjach przestrzennych: przylegania, nakładania itp.
  - Operacji przestrzenno-nieprzestrzennych: np. ilość budynków <10 m w promieniu 1000 metrów
- Informacja na temat lokalizacji i geometrii to **jawne** lub główne cechy przestrzenne, informacja wynikająca z relacji lub operacji to cechy **niejawne** albo poboczne (*secondary*)

Geomorfometria – cechy niejawne

- Przykładem cech niejawnych, które można włączyć do analizy są pochodne terenu. Na podstawie jednej zmiennej (rzędna) i operacji wyliczania pochodnych można zbudować listę dodatkowych atrybutów



	LOWLAND	RANGE	RESBASE	SLOPE
1	0.34666667	16.4405441	11.8017483	0.164444444
2	0.29777778	5.7955704	17.0003109	0.000000000
3	0.46222222	10.0126114	22.9106979	0.000000000
4	0.52444444	14.0399246	31.8086338	0.000000000
5	0.48000000	14.8845444	42.3126602	0.000000000
6	0.48000000	14.8634949	50.2092209	0.000000000
7	0.60000000	15.7373962	55.6396103	0.404444444
8	0.46222222	23.0587540	51.2119293	0.537777778
9	0.24000000	44.8523407	33.4271126	0.724444444
10	0.68444444	61.6504555	10.0911036	0.897777778
11	0.40444444	62.6353149	30.0178719	0.835555556
12	0.17777778	43.7212448	40.7687035	0.604444444
13	0.53777778	28.8774567	34.4320717	0.724444444
14	0.18222222	25.8944397	21.9936714	0.253333333
15	0.52888889	32.4404831	20.7477112	0.582222222
16	0.67555556	20.7957611	27.5051785	0.333333333
17	0.62666667	25.3899231	35.5769768	0.640000000
18	0.25777778	19.7975464	47.5967445	0.337777778
19	0.42666667	7.3471451	54.0927353	0.000000000
20	0.13777778	12.7977753	54.3353958	0.044444444
21	0.53777778	19.9280472	50.5568047	0.395555556
22	0.41333333	16.4708328	52.9926529	0.248888889

Reorganizacja danych

- Reorganizację danych przeprowadza się w celu przygotowania danych do właściwych prac nad wydobywaniem wiedzy. Część zagadnień omówiona szczegółowo w dalszej części kursu
- Szeroki i wąski format danych
- Agregacja
- Próbkowanie
- Wybór atrybutów

Format szeroki i wąski

- (lub długi i szeroki) – to dwa sposoby prezentacji danych
- Format szeroki
  - Koncepcyjnie prostszy przy małej liczbie atrybutów
  - Nie jest nadmiarowy
  - Wymaga identycznej liczby atrybutów dla każdego przypadku
- Zaletą danych wąskich jest:
  - Mniejsza liczba kolumn
  - Prosty koncepcyjnie model danych: key-value
  - Złożone analizy danych wymagają długiego formatu
  - Poszczególne przypadki mogą mieć różną liczbę atrybutów
  - Problemy z atrybutami różnego typu

Konwersja pomiędzy formatami

- Konwersja pomiędzy formatami to przestawianie (*pivoting*)
- Wąski typ szczególnie przydatny gdy dane w kolumnach są jednego typu (np. kolejne lata)

	County	LandArea	NatAmenity	College1970	College1980	College1990	College2000	Jobs1970	Jobs1980	Jobs1990	Jobs2000
1	Autauga	599	4	.064	.121	.145	.180	6853	11278	11471	16289
2	Baldwin	1578	4	.065	.121	.168	.231	19749	27861	40809	70247
3	Barbour	891	4	.073	.092	.118	.109	9448	9755	12163	15197
4	Bibb	625	3	.042	.049	.047	.071	3965	4276	5564	6098
5	Blount	639	4	.027	.053	.070	.096	7587	9490	11811	16503

	County	LandArea	NatAmenity	Year	College	Jobs
1	Autauga	599	4	1970	.064	6853
2	Autauga	599	4	1980	.121	11278
3	Autauga	599	4	1990	.145	11471
4	Autauga	599	4	2000	.180	16289
5	Baldwin	1578	4	1970	.065	19749
6	Baldwin	1578	4	1980	.121	27861
7	Baldwin	1578	4	1990	.168	40809
8	Baldwin	1578	4	2000	.231	70247
9	Barbour	891	4	1970	.073	9448
10	Barbour	891	4	1980	.092	9755
11	Barbour	891	4	1990	.118	12163
12	Barbour	891	4	2000	.109	15197
13	Bibb	625	3	1970	.042	3965
14	Bibb	625	3	1980	.049	4276
15	Bibb	625	3	1990	.047	5564
16	Bibb	625	3	2000	.071	6098
17	Blount	639	4	1970	.027	7587
18	Blount	639	4	1980	.053	9490
19	Blount	639	4	1990	.070	11811
20	Blount	639	4	2000	.096	16503

Agregacja przypadków

- Polega na łączeniu obiektów w mniejsze grupy na podstawie atrybuty grupującego: np. gminy łączone są w powiaty
- Agregację wykonuje się w celu stabilizacji wariancji (jednostki niższego rzędu są bardziej zróżnicowane między sobą niż wyższego rzędu)
- Zmiana skali analizy (czasowej, przestrzennej, tematycznej)

Agregacja i transformacja atrybutów

- Atrybuty grupuje się jeżeli są od siebie zależne, lub mogą być wyrażone relacją. Są to tzw. **atrybuty wskaźnikowe** np. zamiast czasu i odległości można podać prędkość m/s
- Atrybut grupuje się w celu zamiany wartości bezwględnych na proporcję, np. zamiast masy próby i masy składnika w próbie podaje się udział procentowy składnika w próbie
- Z danych na etapie analizy usuwa się atrybuty, które są dla analizy nieistotne lub nawet wprowadzające w błąd, np. do analizy składu mineralnego nie ma znaczenia całkowita masa próby
- Usuwanie atrybutów powinno być czasowe tj surowe dane nie powinny być modyfikowane a jedynie ich kopia (mutate)

Próbkowanie

- Jest to operacja, która ma na celu ograniczenie liczby przypadków
- Podstawą operacji próbkowania jest założenie że nie ma **istotnych różnic** pomiędzy wynikiem dla **reprezentatywnej** próby a wynikiem dla populacji
- Próbkowanie wykonuje się w sytuacji:
  - Badań wstępnych, gdy testujemy różne modele
  - Praca na całym zbiorze jest zbyt kosztowna lub niemożliwa
  - Otrzymanie całego zbioru nie jest możliwe lub jest zbyt kosztowne

Rodzaje próbkowania

- **Losowe**: każdy obiekt ma takie same szanse na wylosowanie. Domyślnie próbkowanie losowe jest próbowaniem bez zastępowania, każdy obiekt może być wylosowany raz. W tej procedurze prawdopodobieństwo wylosowania każdego kolejnego elementu wzrasta (po maleje populacja), ale przy dużych zbiorach danych i małej próbce w praktyce jest ono równoważne losowaniu z zastępowaniem (prawdopodobieństwo nie wzrasta znacząco)
- **Losowe z zastępowaniem** jest odmianą wyboru losowego, gdy każdy obiekt może być wylosowany więcej niż raz. Ten rodzaj losowania gwarantuje ze każdy element jest losowany z jednakowym prawdopodobieństwem. W praktyce stosuje się gdy populacja jest mała a próba duża a nawet w sytuacji gdy próba jest większa od populacji tzw. **nadpróbkowanie** (ang. oversampling)
- **Stratyfikowane**: Próba odzwierciedla zróżnicowanie populacji względem danej cechy, każda cecha musi być reprezentowana. Proces losowania rozpoczyna się od podziału populacji na grupy a następnie z każdej grupy losowana jest określona liczba obiektów. Małe grupy mogą być nadreprezentowane.


data
quality
management
reporting
staff
Data
processes
procedures
recording
place
reviewed
collection
responsibility
needs
covering
updates
information
clearly
arrangements
service
governance
external
taken
corporate
example
senior
body
ensure
used
basis
set
agreement
training
Systems
reported
form
delivery
necessary
guidance
action
business
Policies
provision
regular
policy
timely
periodically
analysis
systems
objectives
monitoring
assessed
day
work
policies
practice
relevant
clear
manipulation
risk
address
charged
complete
requirements
back
well
level
results
defined
associated
Internal
system
review
levels
support
working
top
changes
missing
appraisal
part
may
scrutiny
supported
Completeness
fed
tailored
guidelines
Procedures
identified
collecting
needed
evidence
commitment
relation
reinforce
understanding
Accountability
built-in
regularly
body's
Arrangements
Relevance
Information
standards
performance
although
comprehensive
valid
updated
consistently
documented
least
Act
Corrective
previous
possible
Accuracy
unauthorised
programme
areas
acted
corrective
overall
correction
descriptions
agreed
critical
rigorous
extensive
practices
document
forecasting
member
procedure
cleansing
minimum
compiling
implemented
monitor
communicated
reinforcing
reliable
actions
right
erroneous
time
role
necessitate
meet