Spis treści

[Czym jest eksploracja danych? 1](#_Toc164634696)

[Eksploracja i wykopywanie danych (data mining) 1](#_Toc164634697)

[Data Mining a Data Science 2](#_Toc164634698)

[Eksploracja danych i odkrywanie wiedzy (Knowledge Discovery) 2](#_Toc164634699)

[Knowledge Discovery w eksploracji danych 3](#_Toc164634700)

[Problemy w eksploracji danych 3](#_Toc164634701)

[Przykłady zastosowań eksploracji danych 4](#_Toc164634702)

[Data Exploration w GIS 4](#_Toc164634703)

[Eksploracja danych w uczeniu maszynowym 4](#_Toc164634704)

[Interaktywna eksploracja danych 5](#_Toc164634705)

[Eksploracja danych w Pythonie 5](#_Toc164634706)

[Eksploracja danych w R 6](#_Toc164634707)

[Przewodnik do eksploracji danych 6](#_Toc164634708)

[Etapy eksploracji i przygotowania danych 6](#_Toc164634709)

[Identyfikacja zmiennych 7](#_Toc164634710)

[Analiza jednej zmiennej 8](#_Toc164634711)

[Analiza dwóch zmiennych (Bi-variate analysis) 8](#_Toc164634712)

[Ciągła i ciągła: 8](#_Toc164634713)

[Kategoryczna & Kategoryczna: 9](#_Toc164634714)

[Kategoryczna & Ciągła: 10](#_Toc164634715)

[Obróbka wartości brakujących 10](#_Toc164634716)

[Dlaczego w moich danych brakuje wartości? 11](#_Toc164634717)

[Jakie są metody leczenia brakujących wartości? 11](#_Toc164634718)

[Techniki wykrywania i leczenia wartości odstających 13](#_Toc164634719)

[Co to jest wartość odstająca (outlier)? 13](#_Toc164634720)

[Jakie są typy wartości odstających? 13](#_Toc164634721)

[Co powoduje wartości odstające? 13](#_Toc164634722)

[Jaki jest wpływ wartości odstających na zestaw danych? 14](#_Toc164634723)

[Jak wykryć wartości odstające? 15](#_Toc164634724)

[Jak usunąć wartości odstające? 15](#_Toc164634725)

[Co to jest inżynieria zmiennych (cech, atrybutów)? 16](#_Toc164634726)

[Proces inżynierii cech 17](#_Toc164634727)

[Transformacja zmiennych 17](#_Toc164634728)

[Co to jest tworzenie cech / zmiennych i jego zalety? 18](#_Toc164634729)

[Ekstrakcja cech 19](#_Toc164634730)

[Wybór cechy (feature selection) 19](#_Toc164634731)

[Ważność cechy dla modelu. Feature importance 19](#_Toc164634732)

[Rodzaje cech/modeli 20](#_Toc164634733)

[Współczynniki regresji 20](#_Toc164634734)

[Korelacja z etykietą 21](#_Toc164634735)

[Funkcja Ablacja 21](#_Toc164634736)

[Znaczenie permutacji 21](#_Toc164634737)

[Znaczenie oparte na modelu drzewa 22](#_Toc164634738)

[SHAP 22](#_Toc164634739)

[Wniosek 22](#_Toc164634740)

[Wizualizacja danych 23](#_Toc164634741)

[Definicja wizualizacji danych 23](#_Toc164634742)

[Co sprawia, że wizualizacja danych jest skuteczna? 23](#_Toc164634743)

[5 typów wizualizacji danych 23](#_Toc164634744)

[Czasowy 23](#_Toc164634745)

[Hierarchiczny 23](#_Toc164634746)

[Sieć 24](#_Toc164634747)

[Wielowymiarowy 24](#_Toc164634748)

[Geo-przestrzenny 24](#_Toc164634749)

[Jak dobrać właściwą wizualizację do mojego zadania eksploracji danych? 24](#_Toc164634750)

[Co chcesz zwizualizować, pokazać? 25](#_Toc164634751)

[Porównania (ze współrzędnymi i bez) 25](#_Toc164634752)

[Proporcje 25](#_Toc164634753)

[Relacje 25](#_Toc164634754)

[Hierarchia 25](#_Toc164634755)

[Dystrybucja 25](#_Toc164634756)

[Ruch lub przepływ 25](#_Toc164634757)

[Wzory 26](#_Toc164634758)

[Zasięg (Zakres) 26](#_Toc164634759)

[Dane w czasie 26](#_Toc164634760)

[Porównania 26](#_Toc164634761)

[Porównania (ze współrzędnymi) ? 26](#_Toc164634762)

[Wykres słupkowy 26](#_Toc164634763)

[Wykres słupkowy grupowy 27](#_Toc164634764)

[Wykres Marimekko 28](#_Toc164634765)

[Wykres róży 29](#_Toc164634766)

[Wykres równoległych współrzędnych 30](#_Toc164634767)

[Wykres radarowy 31](#_Toc164634768)

[Radialny wykres kolumnowy 32](#_Toc164634769)

[Mapa termiczna 33](#_Toc164634770)

[Kartogram 34](#_Toc164634771)

[Piktogram 35](#_Toc164634772)

[Proporcje 36](#_Toc164634773)

[Skumulowany wykres słupkowy 36](#_Toc164634774)

[Wykres kołowy 37](#_Toc164634775)

[Diagram (wykres) pierścieniowy 38](#_Toc164634776)

[Diagram macierzy punktowej 39](#_Toc164634777)

[Dystrybucje 40](#_Toc164634778)

[Wykres zagęszczenia 40](#_Toc164634779)

[Histogram 41](#_Toc164634780)

[Wykres pudełkowy z wąsami 42](#_Toc164634781)

[Piramida ludności 43](#_Toc164634782)

[Dane w czasie 44](#_Toc164634783)

[Wykres liniowy 44](#_Toc164634784)

[Wykres warstwowy 45](#_Toc164634785)

[Wykres strumieniowy 46](#_Toc164634786)

[Ruch i przepływy 47](#_Toc164634787)

[Wykresy zestawów równoległych 47](#_Toc164634788)

[Diagram Sankey’a 48](#_Toc164634789)

[Relacje (wykresy) 49](#_Toc164634790)

[Wykres rozrzutu 49](#_Toc164634791)

[Mapa termiczna - macierz korelacji 50](#_Toc164634792)

[3D Wykresy rozrzutu 51](#_Toc164634793)

[3D Wykresy rozrzutu z regresją trójwymiarową 51](#_Toc164634794)

[Wykres bąbelkowy 52](#_Toc164634795)

[Relacje (sieci) 53](#_Toc164634796)

[Diagram sieci (Graf) 53](#_Toc164634797)

[Diagram cięciwy 54](#_Toc164634798)

[Diagram cięciwy bez wstążki 55](#_Toc164634799)

[Diagram drzewa 55](#_Toc164634800)

[Hierarchia 56](#_Toc164634801)

[Mapy drzew 56](#_Toc164634802)

[Pakowanie w kółko 57](#_Toc164634803)

[Diagram „rozbłysk słońca” (Sunburst) 58](#_Toc164634804)

[Dendrogram 59](#_Toc164634805)

# Czym jest eksploracja danych?

**Eksploracja danych** oraz **eksploracyjna analiza danych (EDA**) **dostarcza(ją?)** prosty zestaw **narzędzi** **eksploracyjnych**, które pozwalają na podstawowe **zrozumienie** danych w czasie rzeczywistym w analityce danych.

Rezultaty eksploracji danych mogą być potężnym czynnikiem w **zrozumieniu:** struktury danych, rozkładu wartości i wzajemnych powiązań.

Eksploracja danych może być również pomocna dla naukowców w uzyskaniu właściwego wglądu w dane biznesowe, które wcześniej nie były łatwo dostrzegalne.

**Eksploracja danych jest pierwszym krokiem w analityce danych.** **Zrozumienie** danych biznesowych jest niezbędne do podjęcia dobrze zaplanowanej decyzji, co zazwyczaj wiąże się z podsumowaniem głównych cech zbioru danych, takich jak jego rozmiar, wzór, charakterystyka, dokładność i inne.

Cały proces jest prowadzony przy użyciu narzędzi do analizy wizualnej oraz zaawansowanego oprogramowania statystycznego, takiego jak R lub Python oraz specjalistyczne oprogramowania.

**Eksploracyjna analiza danych (EDA),** podobnie jak eksploracja danych, jest techniką statystyczną służącą do analizy zbiorów danych pod kątem ich szerokiej charakterystyki. Narzędzia wizualizacyjne do analizy danych eksploracyjnych umożliwiają interakcję z surowymi zbiorami danych, dając analitykom lepszy wgląd we **wzorce** i **zależności** występujące w danych.

## Eksploracja i wykopywanie danych (data mining)

Istnieją **2 główne metodologie** (techniki)używane do pobierania odpowiednich danych z dużych, niezorganizowanych zbiorów.

Są to metody **ręczne** (tzw.manualna, eksploracja danych) i **automatyczne** (data mining).

Eksploracja i wizualizacja danych dostarcza **wskazówek** do zastosowania najbardziej efektywnych metod statystycznych i data mining.

Po ujawnieniu **zależności** pomiędzy poszczególnymi **zmiennymi**, analitycy mogą przystąpić do procesu eksploracji danych, budując i wdrażając modele danych z wykorzystaniem zdobytych informacji. **Eksploracja danych i data mining mogą być stosowane zamiennie.**

**Data Mining** to zestaw metod, które mają zastosowanie do dużych i złożonych baz danych. Ma to na celu ***wyeliminowanie przypadkowości*** *i* ***odkrycie ukrytego wzoru***. Ponieważ te metody eksploracji danych prawie zawsze wymagają dużej mocy obliczeniowej. Używamy narzędzi, metodologii i teorii eksploracji danych do ujawniania wzorców w danych.

Stosujemy techniki eksploracji danych w **długim procesie badań** i **rozwoju produktu**. Pozwala także użytkownikom poruszać się po danych w czasie rzeczywistym. Korzystamy z eksploracji danych w środowisku biznesowym, ponieważ jest ona obsługiwana przez trzy dojrzałe technologie:

* **Ogromne gromadzenie danych**
* **Potężne komputery wieloprocesorowe**
* **Algorytmy eksploracji danych**

## Data Mining a Data Science

Obraz zawierający tekst, krąg, Czcionka, diagram

Opis wygenerowany automatycznie**Dane ustrukturyzowane** pochodzą z danych transakcji, systemów zarządzania relacyjnymi bazami danych, **Dane nieustrukturyzowane** pochodzą z wiadomości e-mail, blogów, aktywności w mediach społecznościowych.

Istnieją również **dane** **częściowo** **ustrukturyzowane** pozyskiwane jako pliki tekstowe i dzienniki systemowe.

**Data Science** to pula operacji danych, która obejmuje również **Data Mining**.

**Data Scientist** jest odpowiedzialny za rozwój produktów danych dla przemysłu. Z drugiej strony **eksploracja danych** jest odpowiedzialna za wydobywanie użytecznych danych również z innych zasobów.

**Data Scientist** wykonuje wiele operacji, takich jak analiza danych, opracowywanie modeli predykcyjnych, odkrywanie ukrytych wzorców itp. Mimo, że **Data Mining** obejmuje modelowanie statystyczne w celu uzyskania przydatnych informacji.

**Data Science** ma do czynienia zarówno z danymi ustrukturyzowanymi, jak i nieustrukturyzowanymi.

Obraz zawierający tekst, zrzut ekranu, Czcionka, diagram

Opis wygenerowany automatycznie**Data Mining** zajmuje się wyłącznie informacjami strukturalnymi.

## Eksploracja danych i odkrywanie wiedzy (Knowledge Discovery)

Oto lista kroków związanych z procesem odkrywania wiedzy w eksploracji danych :

1. **Czyszczenie danych** - usuwanie szumów i niespójnych danych.

2. **Integracja danych** - Na ogół łączenie wielu źródeł danych.

3. **Wybór danych** - Zasadniczo pobieranie danych z bazy danych istotnych do analizy.

4. **Transformacja danych** – przekształcanie danych w formy odpowiednie do eksploracji, wykonując operacje sumowania lub agregacji.

5**. Data Mining** –stosowanie inteligentnych metod w celu wydobycia wzorców danych.

**6. Ocena wzorca danych**.

**7. Prezentacja i wizualizacja wiedzy** – reprezentowanie wiedzy jako wykresy, tabele, diagramy, sieci, grafy etc.

### Knowledge Discovery w eksploracji danych

Proces znajdowania i interpretowania wzorców na podstawie danych obejmuje wielokrotne stosowanie następujących kroków:

**Rozwijanie zrozumienia:**

* Domena aplikacji
* Odpowiednia wcześniejsza wiedza
* Cele użytkownika końcowego

**Tworzenie docelowego zestawu danych:**

* Wybieranie zestawu danych lub skupianie się na podzbiorze zmiennych (**cech**) lub **próbkach danych**, na których ma zostać przeprowadzone wykrywanie.

**Czyszczenie i wstępne przetwarzanie danych:**

* Usuwanie szumów lub wartości odstających.
* Zbieranie niezbędnych informacji do modelowania lub rozliczania szumów.
* Strategie postępowania z brakującymi polami danych.
* Uwzględnianie informacji o sekwencji czasowej i znanych zmianach.

**Redukcja danych i projekcja:**

* Znajdowanie przydatnych zmiennych (**cech**) do reprezentowania danych w zależności od celu zadania.
* Stosowanie metod redukcji wymiarów w celu zmniejszenia efektywnej liczby zmiennych. Jest to rozważane w celu znalezienia niezmiennych reprezentacji danych.

**Wybieranie zadania eksploracji danych:**

* Decyzja, czy celem procesu KDD jest klasyfikacja, regresja, grupowanie itp.

**Wybór algorytmu (-ów) eksploracji danych:**

* Wybór metody (metod) do wyszukiwania wzorców w danych.
* Podejmowanie decyzji, które modele i parametry mogą być odpowiednie.
* Obraz zawierający zrzut ekranu, tekst, design

  Opis wygenerowany automatycznieDopasowywanie określonej metody eksploracji danych do kryteriów procesu KDD.

**Data Mining:**

* Poszukiwanie wzorców zainteresowania w określonej formie reprezentacyjnej. Takie reprezentacje, jak reguły klasyfikacji lub drzewa, regresja, grupowanie itd.
* Interpretowanie wydobytych wzorów.

**Konsolidacja odkrytej wiedzy.**

## Problemy w eksploracji danych

Kwestie, które należy rozwiązać to:

I. Obsługa niepewności

II. Radzenie sobie z brakującymi wartościami

III. Radzenie sobie z nagłośnionymi danymi

IV. Wydajność algorytmów

V. Ograniczanie wiedzy odkrytej do użytecznej

VI. Uwzględnianie wiedzy o domenach skąd pochodzą dane

VII. Rozmiar i złożoność danych

VIII. Wybór danych

IX. Zrozumiałość odkrytej wiedzy: spójność między danymi a odkrytą wiedzą

## Przykłady zastosowań eksploracji danych

### Image shows an example of data exploration for GIS on OmniSci's Immerse data visualization platform.Data Exploration w GIS

GIS (Geographic Information Systems) to system gromadzenia i analizy danych związanych z lokalizacjami geograficznymi (dane wzbogacone o lokalizację) i ich związkiem z ludzką lub naturalną aktywnością na Ziemi.

Ułatwienie analizy przestrzenno-czasowej do istniejących przepływów danych związanych z analizą big data.

Umiejętność scharakteryzowania i zawężenia surowych danych jest niezbędnym krokiem dla analityków danych przestrzennych, którzy mogą mieć do czynienia z milionami wielokątów i miliardami odwzorowanych punktów.

### Eksploracja danych w uczeniu maszynowym

Aby dobrze działać, modele uczenia maszynowego muszą pobierać duże ilości danych, a dokładność modelu **ucierpi**, jeśli dane te nie zostaną najpierw dokładnie **zbadane**. Kroki eksploracji danych, które należy wykonać przed zbudowaniem modelu uczenia maszynowego, obejmują:

* **Identyfikację** **zmiennych**: zdefiniowanie każdej zmiennej i jej roli w zbiorze danych
* **Analiza** **jednowariantowa**:
  + **dla zmiennych ciągłych** (numerycznych?), budujemy wykresy pudełkowe lub histogramy dla każdej zmiennej niezależnie.
  + **dla zmiennych kategorycznych**, budujemy wykresy słupkowe, aby pokazać częstości.
* **Analiza dwu- i wielozmienna** - określa interakcje pomiędzy zmiennymi poprzez tworzenie narzędzi wizualizacyjnych
* Wykrywanie i obróbka **brakujących** wartości
* Wykrywanie i obróbka wartości **odstających**

Ostatecznym celem maszynowego uczenia eksploracji danych **jest dostarczenie wglądu w dane, który zainspiruje późniejszą inżynierię cech (feature engineering) i proces budowania modelu**.

**Inżynieria cech** ułatwia proces uczenia maszynowego i zwiększa moc predykcyjną algorytmów uczenia maszynowego poprzez **tworzenie cech** z surowych danych.

### Interaktywna eksploracja danych

Zaawansowane techniki wizualizacji są stosowane w wielu dziedzinach, aby umożliwić użytkownikom wizualizację wzorców i wgląd w złożone przepływy danych, a następnie podejmowanie decyzji opartych na danych.

W narzędziach do eksploracji big data, **interaktywność** jest ważnym elementem w postrzeganiu technologii wizualnych do eksploracji danych i rozpowszechnianiu spostrzeżeń.

Obraz zawierający tekst, zrzut ekranu, Oprogramowanie multimedialne, mapa

Opis wygenerowany automatycznieInteraktywna eksploracja danych podkreśla znaczenie pracy zespołowej i ułatwia interakcję międzyludzką dzięki integracji zaawansowanych technologii interakcji i wizualizacji.

Przyspieszone platformy interakcji multimodalnej wyposażone w graficzne interfejsy użytkownika, które ułatwiają eksplorację dużych danych poprzez analizę wizualną, przyspieszają dzielenie się opiniami, usuwają wąskie gardło danych w analizie indywidualnej i skracają czas odkrywania.

## Eksploracja danych w Pythonie

Ma przewagę w łatwości nauki, gotowości do produkcji, integracji z popularnymi narzędziami, bogatej bibliotece i wsparciu ogromnej społeczności. Prawie każdy zestaw narzędzi i funkcjonalności jest dostępny jako pakiet metod łatwo wywoływalnych (nazwa metody).

Eksploracja danych w Pythonie (np. na **Jupiter Notebook**) jest łatwiejsza dzięki otwartej bibliotece **Pandas** do analizy danych, która może samodzielnie sprofilować dowolną ramkę danych i wygenerować kompletny raport HTML na temat zbioru danych. Pandas pozwala na importowanie plików w różnych formatach, z których najpopularniejszym jest CSV.

**Biblioteka eksploracji danych Pandas zapewnia**:

* Wydajny obiekt **dataframe** do manipulacji danymi ze zintegrowanym indeksowaniem
* Narzędzia do odczytu i zapisu danych pomiędzy różnymi formatami
* Zintegrowaną obsługę brakujących danych i inteligentne wyrównywanie danych
* Elastyczne przestawianie i przekształcanie zbiorów danych
* Funkcjonalność szeregów czasowych
* Inteligentne, oparte na etykietach krojenie, wymyślne indeksowanie i podzbiory dużych zbiorów danych
* Możliwość wstawiania i usuwania kolumn ze struktur danych w celu zmiany rozmiaru
* Agregowanie lub przekształcanie danych za pomocą wydajnego mechanizmu grupowania pozwalającego na operacje dzielenia, stosowania i łączenia zbiorów danych
* Wysokowydajne łączenie i scalanie zbiorów danych
* Hierarchiczne indeksowanie osi

## Eksploracja danych w R

Proces eksploracji i wizualizacji danych za pomocą R wygląda następująco:

* **Wczytywanie danych**: Ze względu na dostępność predefiniowanych bibliotek i prostą składnię, ładowanie danych z różnych formatów, takich jak .XLS, TXT, CSV i JSON, jest bardzo proste
* **Konwersja zmiennych**: Proces konwersji zmiennej na inny typ danych w R obejmuje dodanie łańcucha znaków do wektora liczbowego, konwersję wszystkich elementów wektora na znak.
* **Transpozycja zbioru danych**: R dostarcza kod do transpozycji zbioru danych z szerokiej struktury do znacznie węższej struktury
* **Sortowanie ramek danych**: osiągane poprzez użycie kolejności jako indeksu
* **Tworzenie wykresów** lub **histogramów**
* **Generowanie tabel częstości**, aby jak najlepiej zrozumieć dystrybucję pomiędzy kategoriami
* Generowanie zestawu próbek z kilkoma losowymi indeksami
* **Usuwanie zduplikowanych wartości** zmiennej
* **Znajdź średnią i sumę zliczeń na poziomie klasy**: Techniki eksploracji danych w R zawierają funkcje, które pozwalają to osiągnąć
* **Rozpoznanie i leczenie brakujących wartości** i **wartości odstających** poprzez wprowadzenie **średniej** z innych liczb
* **Scalanie i łączenie zbiorów danych**: R zawiera funkcję dołączania zbiorów danych oraz funkcję bind

# Przewodnik do eksploracji danych

## Etapy eksploracji i przygotowania danych

**Jakość danych wejściowych decyduje o jakości wyjścia (modelu i wniosków z niego)**. Tak więc, gdy już przygotujesz swoją hipotezę biznesową, warto poświęcić tutaj dużo czasu i wysiłku. Eksploracja, czyszczenie i przygotowywanie danych może zająć do 70% całkowitego czasu projektu.

Poniżej przedstawiono kroki niezbędne do zrozumienia, oczyszczenia i przygotowania danych do budowania modelu predykcyjnego:

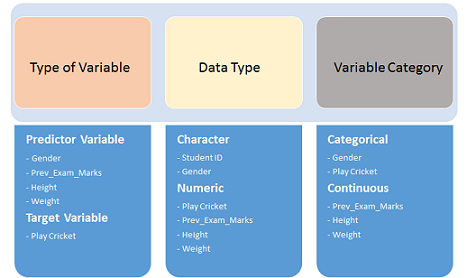
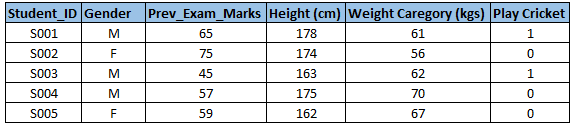
1. Identyfikacja zmiennych
2. Analiza jednowymiarowa
3. Analiza dwuwymiarowa
4. Leczenie brakujących wartości
5. Leczenie odstających wartości
6. Transformacja zmiennych
7. Tworzenie zmiennych

Wreszcie, będziemy musieli powtórzyć kroki 4 - 7 wiele razy, zanim wymyślimy nasz wyrafinowany model.

## Identyfikacja zmiennych

Najpierw zidentyfikuj zmienne **Predykator (wejście) i cel (wyjście**). Następnie określ **typ danych i kategorię zmiennych.**

Przykład: - *Załóżmy, że chcemy przewidzieć, czy uczniowie będą grać w krykieta, czy nie. Tutaj musisz zidentyfikować zmienne predykcyjne, zmienną docelową, typ danych zmiennych i kategorię zmiennych.*



Zmienne zostały zdefiniowane w różnych kategoriach:

**Typ zmiennych**

**Zmienne wejścia, predykatory**

(gender, Prev. Exam. Mark, Height, Weight)

**Zmienna wyjścia, cel**

(Play Cricket)

**Typ danych**

**Kategoryczne**

(student ID, gender)

**Numeryczne**

(Play Cricket, Prev. Exam Marks, Height, Weight)

**Kategoria zmiennych**

**Klasyfikacyjne**

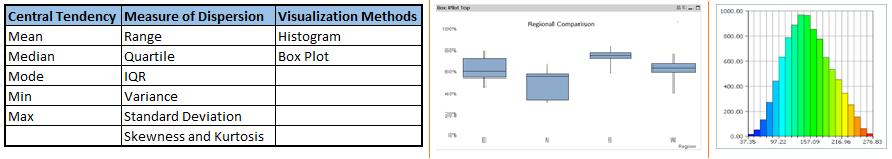
(Gender, Play Cricket)

**Ciągłe**

(Prev. Exam marks, Height, Weight)

## Analiza jednej zmiennej

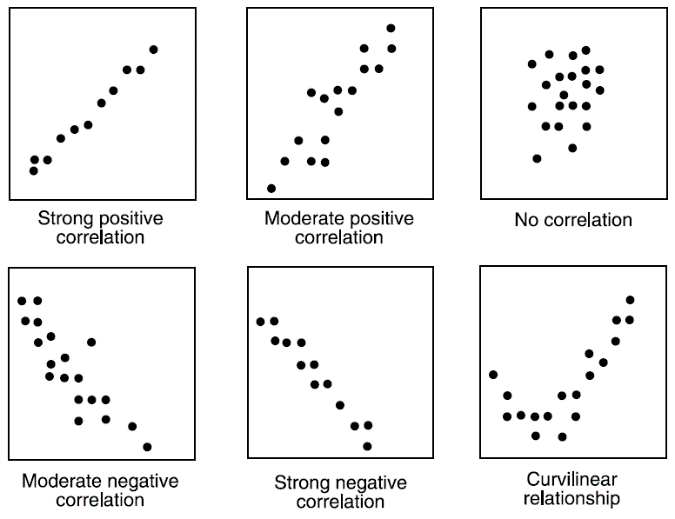
Na tym etapie badamy zmienne jedną po drugiej. Metoda przeprowadzania analizy jednowymiarowej będzie zależeć od tego, czy **typ** **zmiennej** jest **kategoryczny** czy **ciągły**(numeryczny). Przyjrzyjmy się tym metodom i miarom statystycznym dla zmiennych kategorycznych i ciągłych indywidualnie:

**Zmienne ciągłe**: W przypadku zmiennych ciągłych musimy zrozumieć tendencję centralną i rozkład zmiennej. Są one mierzone przy użyciu różnych metod wizualizacji metryk statystycznych, jak pokazano poniżej:

Wbrew stwierdzeniom obecnym w niektórych podręcznikach, **kurtoza** nie mierzy "spłaszczenia", "wysmukłości" ani "spiczastości" rozkładu. Na **kurtozę** ma wpływ intensywność występowania wartości skrajnych, mierzy więc ona, co się dzieje w "**ogonach**" rozkładu, natomiast kształt "**czubka**" rozkładu jest praktycznie bez znaczenia

**Zmienne kategoryczne**: W przypadku zmiennych kategorycznych użyjemy tabeli częstotliwości, aby zrozumieć rozkład każdej kategorii. Możemy również odczytać jako procent wartości w każdej kategorii. Można go zmierzyć za pomocą dwóch wskaźników, **liczbowego i procentowego** dla każdej kategorii. **Wykres słupkowy** może być używany jako wizualizacja.

## Analiza dwóch zmiennych (Bi-variate analysis)

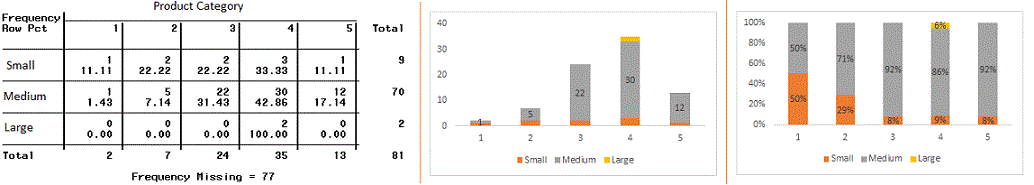
**Analiza dwuwymiarowa** odkrywa związek między dwiema zmiennymi. Tutaj szukamy skojarzenia i rozdzielenia między zmiennymi na wcześniej zdefiniowanym poziomie istotności. Możemy przeprowadzić analizę dwuwymiarową dla dowolnej kombinacji zmiennych **kategorycznych** i **ciągłych**.   
Różne metody są stosowane do radzenia sobie z tymi kombinacjami podczas procesu analizy.  
Zrozummy szczegółowo możliwe kombinacje:

Ciągła i ciągła:Wykonując analizę dwuwymiarową między dwiema zmiennymi ciągłymi, powinniśmy spojrzeć na wykres punktowy. Jest to sprytny sposób na poznanie związku między dwiema zmiennymi.

Wzorzec wykresu punktowego wskazuje relację między zmiennymi. Relacja może być **liniowa** lub **nieliniowa** (krzywoliniowa?).

### Kategoryczna & Kategoryczna:

Aby znaleźć związek między dwiema zmiennymi kategorycznymi, możemy użyć następujących metod:  
**Tabela dwukierunkowa**: Możemy rozpocząć analizę relacji, tworząc dwukierunkową tabelę liczenia i liczenia%. Wiersze reprezentują kategorię jednej zmiennej, a kolumny reprezentują kategorie drugiej zmiennej. Pokazujemy liczbę lub procent liczby obserwacji dostępnych w każdej kombinacji kategorii wierszy i kolumn.  
**Skumulowany wykres kolumnowy**: Ta metoda jest bardziej wizualną formą tabeli dwukierunkowej.



#### Chi-Kwadrat Test:

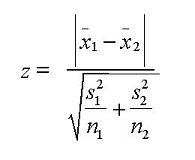
Badanie to służy do uzyskania statystycznej istotności relacji między zmiennymi. Ponadto sprawdza, czy dowody w próbie są wystarczająco silne, aby uogólnić, że związek dotyczy również większej populacji. **Chi-kwadrat opiera się na różnicy między oczekiwanymi i obserwowanymi częstościami w jednej lub kilku kategoriach w tabeli dwukierunkowej**.  
Zwraca prawdopodobieństwo dla obliczonego rozkładu chi-kwadrat ze stopniem swobody.

Data Exploration, Chi Square, Business AnalyticsStatystyka testu chi-kwadrat dla testu niezależności dwóch zmiennych kategorycznych oblicza się za pomocą:

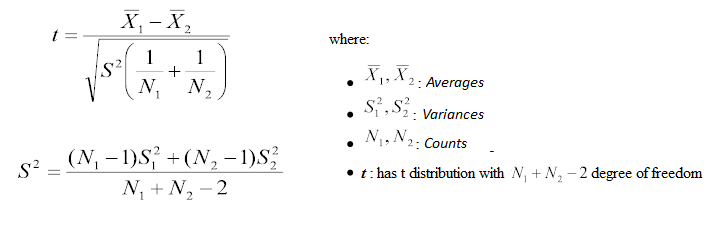
Data Exploration, Chi Square, Business AnalyticsGdzie:  
**O** reprezentuje obserwowaną częstość.   
**E** jest oczekiwaną częstością w hipotezie zerowej i obliczoną przez:

**Prawdopodobieństwo 0**: Wskazuje, że obie zmienne kategoryczne są zależne  
**Prawdopodobieństwo 1**: Pokazuje, że obie zmienne są niezależne.  
**Prawdopodobieństwo mniejsze niż 0,05**: Wskazuje, że relacja między zmiennymi jest znacząca przy 95% ufności.

### Kategoryczna & Ciągła:

Badając relacje między zmiennymi kategorycznymi i ciągłymi, możemy rysować wykresy **pudełkowe** dla każdego poziomu zmiennych **kategorycznych**. Jeśli poziomy są małe, nie pokażą istotności statystycznej. Aby obliczyć **istotność statystyczną**, możemy wykonać test Z, test T lub ANOVA.

#### Z-Test/ T-Test:

W każdym z testów ocenia się, czy średnia z dwóch grup statystycznie różnią się od siebie lub nie.

Jeśli prawdopodobieństwo **Z** jest małe, różnica dwóch średnich jest **bardziej** znacząca.   
Test T jest bardzo podobny do testu Z, ale jest używany, gdy **liczba obserwacji** dla obu kategorii jest **mniejsza niż 30**.

#### Obraz zawierający tekst, zrzut ekranu, Czcionka, numer Opis wygenerowany automatycznieAnaliza wariancji

##### ANOVA:

Ocenia, czy średnia z więcej niż dwóch grup jest statystycznie różna.  
Test ANOVA jest stosowany, gdy zakładamy **normalny** rozkład zmiennych.

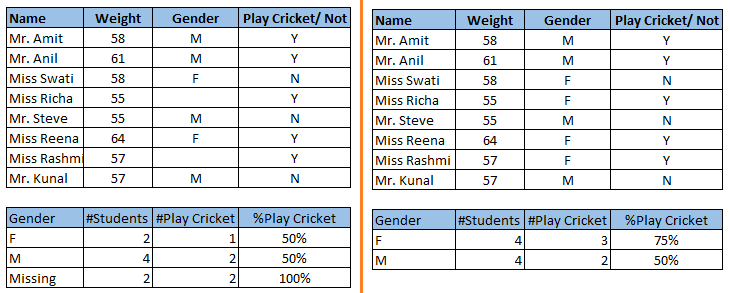
##### Kruskal-Wallis test:

Test Kruskala-Wallisa jest nieparametryczną alternatywą dla jednoczynnikowego testu ANOVA dla niezależnych pomiarów. Opiera się na uporządkowaniu danych, a nie na obliczeniach obejmujących średnie i wariancje, i pozwala ocenić różnice między trzema lub więcej niezależnymi próbkami (zabiegami).

Kruskal\_Wallis test jest preferowany, gdy zakładamy **nienormalny** rozkład zmiennych.

## Obróbka wartości brakujących

**Dlaczego leczenie brakujących wartości jest wymagane?**

Brakujące dane w zestawie danych treningowych mogą zmniejszyć moc / dopasowanie modelu lub mogą prowadzić do modelu tendencyjnego, ponieważ nie przeanalizowaliśmy poprawnie zachowania i relacji z innymi zmiennymi. Może to prowadzić do błędnego przewidywania lub klasyfikacji.

Zwróć uwagę na brakujące wartości na obrazku: W scenariuszu po lewej stronie nie usunęliśmy brakujących wartości. Wniosek z tego zestawu danych jest taki, że szanse na grę w krykieta przez mężczyzn są wyższe niż kobiet. Z drugiej strony, jeśli spojrzymy na drugą tabelę, która pokazuje dane po leczeniu brakujących wartości (w oparciu o płeć), widzimy, że kobiety mają większe szanse na grę w krykieta w porównaniu do mężczyzn.

### Dlaczego w moich danych brakuje wartości?

#### Ekstrakcja danych:

Możliwe, że występują problemy z procesem ekstrakcji. W takich przypadkach powinniśmy dokładnie sprawdzić poprawność danych u opiekunów danych. Niektóre procedury mieszania mogą być również używane do upewnienia się, że wyodrębnianie danych jest poprawne. Błędy na etapie ekstrakcji danych są zazwyczaj łatwe to find i można je również łatwo poprawić.

#### Gromadzenie danych:

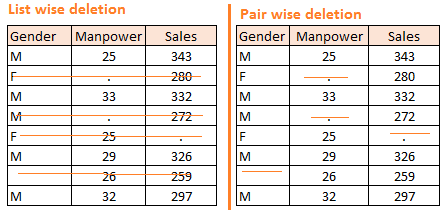
Błędy te występują w momencie zbierania danych i są trudniejsze do naprawienia. Można je podzielić na cztery typy:  
**Brak powstały całkowicie losowo**: Jest to przypadek, w którym prawdopodobieństwo brakującej zmiennej jest takie samo dla wszystkich obserwacji**. Na przykład**: *respondenci procesu zbierania danych decydują, że zadeklarują swoje zarobki po rzuceniu monetą. Jeśli wystąpi orzeł, respondent deklaruje swoje zarobki i odwrotnie.* Tutaj każda obserwacja ma równe szanse na utratę wartości.  
**Brakujące losowo**: Jest to przypadek, gdy brakuje zmiennej losowo, a brakujący stosunek zmienia się dla różnych wartości / poziomu innych zmiennych wejściowych. **Na przykład**: *Zbieramy dane dotyczące wieku, a kobieta ma wyższą brakującą wartość w porównaniu z mężczyzną.***Brak, który zależy od nieobserwowanych predykatorów**: Jest to przypadek, gdy brakujące wartości nie są losowe i są związane z nieobserwowaną zmienną wejściową. **Na przykład:** *W badaniu medycznym, jeśli konkretna diagnoza powoduje dyskomfort, istnieje większa szansa na rezygnację z badania.* Ta brakująca wartość nie jest przypadkowa, chyba że uwzględniliśmy "dyskomfort" jako zmienną wejściową dla wszystkich pacjentów.  
**Brak, który zależy od samej brakującej wartości**: Jest to przypadek, gdy prawdopodobieństwo brakującej wartości jest bezpośrednio skorelowane z samą brakującą wartością. **Na przykład:** *Osoby o wyższych lub niższych dochodach prawdopodobnie nie zareagują na swoje zarobki.*

### Jakie są metody leczenia brakujących wartości?

#### Usunięcie:

W przypadku **usuwania listy** (List Wise Deletion)usuwamy obserwacje, w których brakuje którejkolwiek ze zmiennych. Prostota jest jedną z głównych zalet tej metody, ale ta metoda zmniejsza moc modelu, ponieważ zmniejsza rozmiar próbki.

W **usunięciu parami** (Pair Wise Deletion) przeprowadzamy analizę ze wszystkimi przypadkami, w których obecne są interesujące nas zmienne. Zaletą tej metody jest to, że zachowuje ona jak najwięcej przypadków dostępnych do analizy. Jedną z wad tej metody jest to, że wykorzystuje ona różną wielkość próbki dla różnych zmiennych.

Metody usuwania są używane, gdy charakter brakujących danych to **"Brakujące całkowicie losowo**", w przeciwnym razie nielosowe brakujące wartości mogą odchylać dane wyjściowe modelu.

#### Przypisanie Średniej lub Mediany:

**Imputacja** to metoda uzupełniania brakujących wartości wartościami **szacowanymi**. Celem jest wykorzystanie znanych relacji, które można zidentyfikować w prawidłowych wartościach zestawu danych, aby pomóc w oszacowaniu brakujących wartości.

**Średnia** / **Tryb** / **Mediana** imputacji jest jedną z najczęściej stosowanych metod. Polega na zastąpieniu brakujących danych dla danego atrybutu **średnią** lub **medianą** (atrybut ilościowy) lub **trybem** (atrybut jakościowy) wszystkich znanych wartości tej zmiennej. **Rodzaje imputacji:**

**Uogólniona imputacja**: W tym przypadku obliczamy **średnią** lub **medianę** dla wszystkich nie brakujących wartości tej zmiennej, a następnie zastępujemy brakującą wartość. Podobnie jak w powyższej tabeli, brakuje zmiennej "Manpower", więc bierzemy średnią wszystkich nie brakujących wartości "Manpower" (28.33), a następnie zastępujemy nią brakującą wartość.

**Podobny przypadek**: W tym przypadku obliczamy średnią dla płci "Mężczyzna" (29,75) i "Kobieta" (25) indywidualnie brakujących wartości, a następnie zastępujemy brakującą wartość na podstawie płci. W przypadku "Mężczyzny" zastąpimy brakujące wartości siły roboczej 29,75, a "Kobiety" 25.

#### Model predykcyjny:

Model predykcyjny jest jedną z zaawansowanych metod obsługi brakujących danych. Tutaj tworzymy model predykcyjny do szacowania wartości, które zastąpią brakujące dane.  W takim przypadku dzielimy nasz zestaw danych na dwa zestawy:   
Jeden zestaw **bez brakujących** wartości dla zmiennej **i drugi z brakującymi wartościami**.

**Pierwszy zestaw danych staje się zbiorem danych treningowych modelu, podczas gdy drugi zestaw danych z brakującymi wartościami jest zestawem danych testowych, a zmienna z brakującymi wartościami jest traktowana jako zmienna docelowa.**

Następnie tworzymy **model do przewidywania zmiennej docelowej** na podstawie innych atrybutów (zmiennych) zestawu danych treningowych i wypełniamy brakujące wartości zestawu danych testowych.

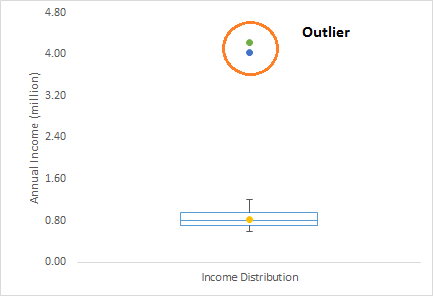
W tym celu możemy użyć **regresji**, **ANOVA**, **regresji logistycznej** **i różnych technik modelowania.**Istnieją 2 wady tego podejścia:

1.Wartości szacowane w modelu są zwykle bardziej dobrze zachowane niż wartości rzeczywiste

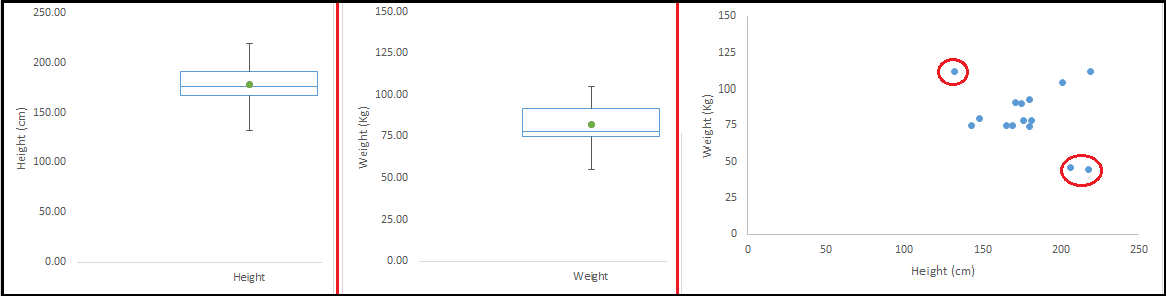
2. Jeśli w zestawie danych nie ma relacji z atrybutami i atrybutem z brakującymi wartościami, model nie będzie precyzyjny do szacowania brakujących wartości.

## Techniki wykrywania i leczenia wartości odstających

### Co to jest wartość odstająca (outlier)?

Wartość odstająca jest powszechnie używaną terminologią przez analityków, ponieważ wymaga szczególnej uwagi, w przeciwnym razie może spowodować bardzo błędne szacunki. Mówiąc najprościej, **Outlier** to **obserwacja**, która wydaje się odległa i odbiega od ogólnego wzorca w próbce.  
**Przykład**, robimy profilowanie klientów i dowiadujemy się, że średni roczny dochód klientów wynosi 0,8 miliona dolarów. Ale jest dwóch klientów o rocznym dochodzie w wysokości 4 i 4,2 miliona dolarów. Dochód tych dwóch gości jest znacznie wyższy niż reszta populacji. Te dwie obserwacje będą postrzegane jako wartości odstające.

### Jakie są typy wartości odstających?

Wartość odstająca może być dwojakiego typu: **jednowymiarowa i wielowymiarowa**. Powyżej omówiliśmy przykład jednowymiarowej wartości odstającej. Te wartości odstające można znaleźć, gdy spojrzymy na rozkład pojedynczej zmiennej. Wielowymiarowe wartości odstające są wartościami odstającymi w przestrzeni n-wymiarowej. Aby je znaleźć, musisz spojrzeć na rozkłady w wielu wymiarach.  
**Przykład**. Powiedzmy, że rozumiemy związek między wzrostem a wagą. Poniżej mamy rozkład jednowymiarowy i dwuwymiarowy dla wzrostu, wagi. Spójrz na pudełko. Nie mamy żadnych wartości odstających (powyżej i poniżej **1,5 \*** **IQR**, najczęstsza metoda). Teraz spójrz na wykres punktowy. Tutaj mamy dwie wartości poniżej i jedną powyżej średniej w określonym segmencie wagi i wzrostu.  


Co powoduje wartości odstające?Przyczyny wartości odstających można podzielić na dwie szerokie kategorie: Sztuczne (Błąd) / Nienaturalne i Naturalne.

**Błędy wprowadzania danych**: - Błędy ludzkie, takie jak błędy spowodowane podczas zbierania, rejestrowania lub wprowadzania danych, mogą powodować wartości odstające w danych. **Na** **przykład**: Roczny dochód klienta wynosi 100 000 USD. Przypadkowo operator wprowadzania danych umieszcza dodatkowe zero na rysunku. Teraz dochód wynosi 1 000 000 USD, co jest 10 razy wyższe. Oczywiście będzie to wartość odstająca w porównaniu z resztą populacji.

**Błąd pomiaru**: Jest to najczęstsze źródło wartości odstających. Dzieje się tak, gdy zastosowany przyrząd pomiarowy okazuje się wadliwy. **Na przykład**: Istnieje 10 wag. 9 z nich jest poprawnych, 1 jest wadliwy. Waga mierzona przez osoby na wadliwej maszynie będzie wyższa/niższa niż reszta osób w grupie. Ciężary mierzone na wadliwej maszynie mogą prowadzić do wartości odstających.

**Błąd eksperymentalny**: Inną przyczyną wartości odstających jest błąd eksperymentalny. **Na przykład:** W sprincie na 100 m z udziałem 7 biegaczy jeden z biegaczy nie skoncentrował się na wezwaniu "Go", co spowodowało, że zaczął późno. W związku z tym spowodowało to, że czas biegu biegacza był dłuższy niż innych biegaczy. Jego całkowity czas pracy może być odstający.

**Celowe wartości odstające**: Jest to często spotykane w środkach zgłaszanych przez siebie, które obejmują poufne dane. **Na przykład:** Nastolatki zakłamują ilość spożywanego alkoholu. Tylko ułamek z nich zgłosiłby rzeczywistą wartość. Tutaj rzeczywiste wartości mogą wyglądać jak wartości odstające, ponieważ reszta nastolatków nie zgłasza użycia.

**Błąd przetwarzania danych**: Za każdym razem, gdy przeprowadzamy eksplorację danych, wyodrębniamy dane z wielu źródeł. Możliwe, że niektóre błędy manipulacji lub wyodrębniania mogą prowadzić do wartości odstających w zestawie danych.

**Błąd próbkowania**: **Na przykład** musimy zmierzyć wzrost sportowców. Przez pomyłkę włączamy kilku koszykarzy do próbki. To włączenie prawdopodobnie spowoduje wartości odstające w zestawie danych.

**Naturalnie odstający**: Gdy wartość odstająca nie jest sztuczna (z powodu błędu), jest naturalną wartością odstającą. **Na przykład**: Podczas mojego ostatniego zadania w jednej z renomowanych firm ubezpieczeniowych zauważyłem, że wyniki 50 najlepszych doradców finansowych były znacznie wyższe niż reszty populacji. Co zaskakujące, nie było to spowodowane żadnym błędem. Dlatego za każdym razem, gdy wykonujemy jakąkolwiek działalność związaną z eksploracją danych z doradcami, traktowaliśmy ten segment osobno.

### Jaki jest wpływ wartości odstających na zestaw danych?

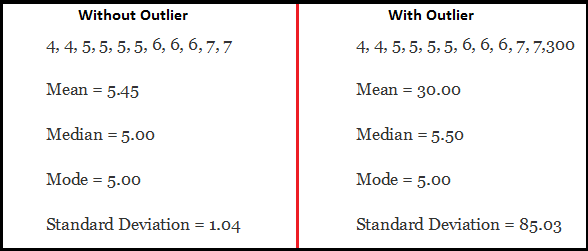
Wartości odstające mogą drastycznie zmienić wyniki analizy danych i modelowania statystycznego. Istnieje wiele niekorzystnych skutków wartości odstających w zbiorze danych:

* Zwiększa wariancję błędu i zmniejsza moc testów statystycznych
* Jeśli wartości odstające są rozmieszczone nielosowo, mogą zmniejszyć normalność
* Mogą one wpływać na szacunki, które mogą mieć istotne znaczenie.
* Mogą również wpływać na podstawowe założenie regresji, ANOVA i inne założenia modelu statystycznego.

Aby dogłębnie zrozumieć wpływ, weźmy przykład, aby sprawdzić, co dzieje się z zestawem danych z wartościami odstającymi i bez nich w zestawie danych.

**Przykład:**

Jak widać, zbiór danych z wartościami odstającymi ma znacząco różną średnią i odchylenie standardowe. W pierwszym scenariuszu powiemy, że średnia wynosi 5,45. Ale z wartością odstającą średnia wzrasta do 30. Zmieniłoby to całkowicie szacunki..



### Jak wykryć wartości odstające?

Obraz zawierający tekst, diagram, linia, Wykres

Opis wygenerowany automatycznieNajczęściej stosowaną metodą wykrywania wartości odstających jest **wizualizacja**. Używamy różnych metod wizualizacji, takich jak **Box-plot**, **Histogram**, **Scatter Plot** (na prawo **wykresu pudełkowego** i **wykresu punktowego** do wizualizacji). Niektórzy analitycy używają również „reguły kciuka” do wykrywania wartości odstających.

Niektóre z nich to:   
**IQR**. Dowolna wartość, która wykracza poza zakres od -1,5 x IQR do 1,5 x IQR

* **Rozstęp międzykwartylowy** (IQR) jest miarą rozproszenia statystycznego.

**Metoda zamykania**. Każda wartość, która poza zakresem 5. i 95. percentyla może być uznana za wartość odstającą  
**Punkty danych**, trzy lub więcej odchyleń standardowych od średniej są uważane za wartości odstające

Wykrywanie wartości odstających jest jedynie szczególnym przypadkiem badania danych pod kątem wpływowych punktów danych i zależy również od zrozumienia biznesowego.  
Dwuwymiarowe i wielowymiarowe wartości odstające są zazwyczaj mierzone za pomocą indeksu wpływu lub dźwigni lub odległości. Popularne indeksy, takie jak odległość Mahalanobisa i D Cooka, są często używane do wykrywania wartości odstających.

### Jak usunąć wartości odstające?

Większość sposobów radzenia sobie z wartościami odstającymi jest podobna do metod brakujących wartości, takich jak usuwanie obserwacji, przekształcanie ich, łączenie, traktowanie ich jako oddzielnej grupy, przypisywanie wartości i inne metody statystyczne. Tutaj omówimy typowe techniki stosowane w radzeniu sobie z wartościami odstającymi:

**Usuwanie obserwacji**: Usuwamy wartości odstające, jeśli jest to spowodowane błędem wprowadzania danych, błędem przetwarzania danych lub obserwacjami odstającymi, których liczba jest bardzo mała. Możemy również użyć przycinania na obu końcach, aby usunąć wartości odstające.

**Przekształcanie i łączenie wartości**: Przekształcanie zmiennych może również wyeliminować wartości odstające. Naturalny dziennik wartości zmniejsza zmienność spowodowaną wartościami ekstremalnymi. Binning jest również formą transformacji zmiennej. Algorytm drzewa decyzyjnego pozwala dobrze radzić sobie z wartościami odstającymi dzięki wiązaniu zmiennych. Możemy również skorzystać z procesu przypisywania wag do różnych obserwacji. 

**Przypisywanie:** Podobnie jak przypisywanie brakujących wartości, możemy również przypisywać wartości odstające. Możemy użyć metod imputacji średniej, mediany, trybu. Przed przypisaniem wartości powinniśmy przeanalizować, czy jest to naturalna wartość odstająca, czy sztuczna. Jeśli jest sztuczny, możemy iść z wartościami przypisującymi. Możemy również użyć modelu statystycznego do przewidywania wartości obserwacji odstających, a następnie możemy przypisać go przewidywanymi wartościami.

**Traktowanie osobno:** Jeśli istnieje znaczna liczba wartości odstających, powinniśmy traktować je osobno w modelu statystycznym. Jednym z podejść jest traktowanie obu grup jako dwóch różnych grup i zbudowanie indywidualnego modelu dla obu grup, a następnie połączenie wyników.

# Co to jest inżynieria zmiennych (cech, atrybutów)?

**Feature engineering** to nauka (i sztuka) wydobywania większej ilości informacji z istniejących danych. Nie dodajesz tutaj żadnych nowych danych, ale w rzeczywistości sprawiasz, że dane, które już masz, są bardziej przydatne.  
**Przykład**, *że próbujesz przewidzieć wpływ spadku stopy procentowej na ceny w centrum handlowym na podstawie dat. Jeśli spróbujesz użyć danych miesięcznych, możesz nie być w stanie wyodrębnić istotnych informacji z danych o spadku stopy. Wynika to z faktu, że spadek stopy procentowej mniejszy wpływ ma na uśredniony dzień miesiąca niż na taki dzień tygodnia. Te informacje o dniu tygodnia są ukryte w Twoich danych. Musisz je wydobyć, aby ulepszyć swój model przewidywania wpływu spadku stopy procentowej na ceny.*

Obraz zawierający tekst, zrzut ekranu, Czcionka, linia

Opis wygenerowany automatycznieObraz zawierający tekst, zrzut ekranu, Czcionka, linia

Opis wygenerowany automatycznieMiejsce inżynierii cech w Data Science 4 etapy inżynierii cech

## Proces inżynierii cech

Wykonujemy inżynierię cech po wykonaniu pierwszych 5 kroków w eksploracji danych -  
Identyfikacja zmiennych, analiza jednowymiarowa, analiza dwuwymiarowa, imputacja brakujących wartości i leczenie wartości odstających.

Samą inżynierię zmiennych można podzielić na 2 kroki:

1. **Transformacja zmiennych (cech).**

2. **Tworzenie zmiennych (cech)**

Te dwie techniki są niezbędne w eksploracji danych i mają niezwykły wpływ na moc przewidywania. Zrozummy każdy z tych kroków bardziej szczegółowo.

### Transformacja zmiennych

W modelowaniu danych transformacja odnosi się do zastąpienia zmiennej inną zmienną.

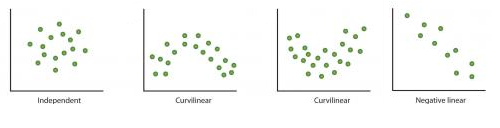
Na przykład zastąpienie zmiennej x pierwiastkiem kwadratowym / sześciennym lub logarytmem x jest transformacją. Innymi słowy, transformacja jest procesem, który zmienia rozkład lub relację zmiennej z innymi.

Przyjrzyjmy się sytuacjom, w których transformacja zmiennych jest przydatna.

#### Kiedy powinniśmy stosować transformację zmiennej?

Kiedy chcemy **zmienić skalę zmiennej (skalowanie)** lub **standaryzować wartości zmiennej** dla lepszego zrozumienia. Ta transformacja jest koniecznością, jeśli dane są w różnych skalach, ta transformacja nie zmienia kształtu rozkładu zmiennej

Kiedy możemy **przekształcić złożone relacje nieliniowe w relacje liniowe**. Istnienie liniowej relacji między zmiennymi jest łatwiejsze do zrozumienia w porównaniu z relacją nieliniową lub zakrzywioną.

Transformacja pomaga nam przekształcić relację nieliniową w relację liniową. **Wykres punktowy** może służyć do znajdowania relacji między **dwiema zmiennymi ciągłymi**. Transformacje te poprawiają również przewidywanie. 

Transformacja logarytmicznajest jedną z powszechnie stosowanych technik transformacji w takich sytuacjach.

Rozkład symetrycznyjest preferowany w stosunku do rozkładu przekrzywionego, ponieważ łatwiej jest interpretować i generować wnioski.

Niektóre techniki modelowania wymagają normalnego rozkładu zmiennych. Tak więc, ilekroć mamy wypaczony rozkład, możemy użyć **transformacji, które zmniejszają skośność**.

Dla prawego pochylonego rozkładu bierzemy pierwiastek kwadratowy / sześcienny lub **logarytm zmiennej**, a dla lewego pochylonego rozkładu bierzemy kwadrat / sześcian lub **funkcję wykładniczą zmiennych.**

**Transformacja zmiennej odbywa się również z punktu widzenia implementacji** (zaangażowanie człowieka).

W jednym z projektów dotyczących wydajności pracowników wiek ma bezpośredni związek z wydajnością pracownika, tj. wyższy wiek, lepsza wydajność. Z punktu widzenia implementacji, uruchomienie programu opartego na wieku może stanowić wyzwanie wdrożeniowe.

**Kategoryzacja** pracowników w trzech grupach wiekowych <30 lat, 30-45 lat i >45, a następnie sformułowanie trzech różnych strategii dla każdej grupy jest rozsądnym podejściem. **Ta technika kategoryzacji jest znana jako wiązanie zmiennych.**

#### Jakie są typowe metody transformacji zmiennej?

Niektóre z nich obejmują pier. kwadr., pier. sześcienny, funkcje logar,, potęgowanie i wiele innych. Przyjrzyjmy się tym metodom szczegółowo, podkreślając ich zalety i wady.

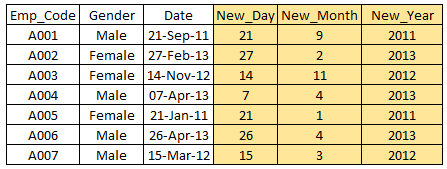
**Logarytm zmiennej**: jest powszechną metodą transformacji używaną do zmiany kształtu rozkładu zmiennej na wykresie rozkładu. Jest zwykle używany *do* ***zmniejszania prawej skośności zmiennych****.* **Nie** można go zastosować do wartości zerowych lub ujemnych.

**Pierw. Kwadr. / sześcienny** **zmiennej**: ma wpływ na rozkład zmiennych. Nie jest on jednak tak znaczący jak transformacja logarytmiczna. Można go zastosować do wartości ujemnych i zera.

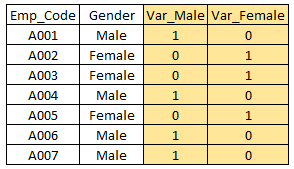
**Binning**: Służy do kategoryzacji zmiennych. Jest wykonywany na oryginalnych wartościach, percentylu lub częstości. Decyzja o technice kategoryzacji opiera się na zrozumieniu biznesowym. Na przykład możemy kategoryzować dochód w trzech kategoriach, a mianowicie: **Wysoki**, **Średni** i **Niski**. Możemy również wykonać współzmienne wiązanie, które zależy od wartości więcej niż jednej zmiennej.

### Co to jest tworzenie cech / zmiennych i jego zalety?

**Tworzenie cech (zmiennych) pochodnych** to proces generowania nowych zmiennych (cech) na podstawie istniejących zmiennych.

Załóżmy na przykład, że mamy datę (dd-mm-rr) jako zmienną wejściową w zestawie danych. Możemy generować nowe zmienne, takie jak dzień, miesiąc, rok, tydzień, dzień tygodnia, które mogą mieć lepszą relację ze zmienną docelową. Ten krok służy do wyróżnienia ukrytej relacji w zmiennej:

**Tworzenie zmiennych pozorowanych**: Jednym z najczęstszych zastosowań zmiennej pozorowanej jest konwersja zmiennej kategorycznej na zmienne liczbowe. Zmienne pozorowane są również nazywane **zmiennymi wskaźnikowymi**.

Przydatne jest przyjęcie zmiennej kategorycznej jako predyktora w modelach statystycznych.  Zmienna kategoryczna może przyjmować wartości 0 i 1. Weźmy zmienną "płeć". Możemy wytworzyć dwie zmienne, a mianowicie "Var\_Male" o wartościach 1 (Mężczyzna) i 0 (Brak mężczyzny) oraz "Var\_Female" o wartościach 1 (Kobieta) i 0 (Brak Kobiety). Możemy również tworzyć zmienne pozorowane dla więcej niż dwóch klas zmiennych kategorycznych ze zmiennymi atrapami n lub n-1.

## Ekstrakcja cech

**Ekstrakcja cech** jest ważnym krokiem w inżynierii cech, ponieważ ma na celu zminimalizowanie wymiarowości danych poprzez przełożenie ich na reprezentację o niższych wymiarach przy jednoczesnym zachowaniu użytecznych informacji.

Technika ta jest przydatna w przypadku danych wielowymiarowych lub zwartych reprezentacji, które zachowują kluczowe cechy danych. Techniki ekstrakcji cech zmniejszają złożoność i poprawiają wizualizację. Zmniejszenie wymiarów może spowodować utratę możliwości interpretacji. O tym, czy ilość cech (i które) można ograniczyć, decyduje charakter danych, problem i kompromisy.

## Wybór cechy (feature selection)

**Wybór cech** to kluczowy etap uczenia maszynowego, obejmujący identyfikację i wybór odpowiednich i znaczących zmiennych z surowych danych w celu zwiększenia mocy predykcyjnej i dokładności modelu.

Celem selekcji cech jest zwiększenie wydajności modelu poprzez minimalizację złożoności danych wejściowych przy jednoczesnym zachowaniu najbardziej znaczących i informacyjnych elementów, ograniczeniu nadmiernego dopasowania, przyspieszeniu uczenia i zwiększeniu możliwości interpretacji.

Istnieją trzy typy metod selekcji cech:

**Metody filtrowania:** Zbadaj znaczenie cech niezależnie od danego modelu, stosując środki statystyczne do uszeregowania lub oceniania cech.

**Metody opakowujące:** obejmują wielokrotne szkolenie i ocenę modelu z różnymi podzbiorami cech

**Metody wbudowane:** Włączają wybór cech do procesu budowania modelu.

Cechy w zbiorze danych nie są równie cenne. Wybranie niepotrzebnych lub nadmiarowych zmiennych może skutkować nadmiernym dopasowaniem, zwiększonym kosztem obliczeniowym i zmniejszoną interpretowalnością modelu.

## Ważność cechy dla modelu. Feature importance

**Ważność cechy** to koncepcja stosowana w uczeniu maszynowym, która odnosi się do **względnego znaczenia lub wkładu każdej zmiennej wejściowej (**znanej również jako cecha lub predyktor**) w przewidywaniu zmiennej docelowej lub wyniku**.

Zapewnia wgląd w to, **które cechy mają największy wpływ na przewidywania modelu**. Do obliczenia znaczenia cech można zastosować różne metody. W tych algorytmach szukamy listy funkcji wraz z odpowiadającymi im wartościami ważności.

**Dlaczego musisz zidentyfikować ważne zmienne**? Kluczowym powodem identyfikacji krytycznych cech jest zrozumienie odpowiedniego zestawu cech, które pomogłyby osiągnąć najlepszą wydajność. Kluczową korzyścią rozwiązania tego problemu jest to, że może pomóc zmniejszyć liczbę cech w modelu, pomagając nam skupić się na cechach o wystarczająco dużym znaczeniu.

Zmniejszenie liczby cech bezpośrednio **zmniejsza koszty i czas szkolenia modelu i wnioskowania**. Innym kluczowym powodem oceny znaczenia cech jest **zwiększenie wyjaśnialności modeli**. Ponieważ większość przypadków użycia uczenia maszynowego przechodzi do głębokiego uczenia się, nieprzezroczystość modeli znacznie wzrosła. Ważność cechy i zasięg cechy zapewniają głębsze zrozumienie, dlaczego i jak modele ML działają. Taka **inżynieria wsteczna działania mo**delu jest pomocna w debugowaniu i raportowaniu wykonawczym.

### Rodzaje cech/modeli

Cechy mogą być różnego typu. Zacznijmy od **cech** **gęstych**, które są numeryczne, tj. liczb całkowitych (integer) i zmiennoprzecinkowych (float). Liczba słów, znaków, polubień, odpowiedzi, waga, wzrost itp. to dobre przykłady gęstych cech.

Następnie istnieją **cechy kategoryczne**, zwykle przedstawiane jako małe wektory (one-hot-vectors). Cechy takie jak temat, płeć, grupa wiekowa itp.

Następnie istnieją **rzadkie cechy**, takie jak osadzanie identyfikatora języka, osadzanie identyfikatora pozycji, osadzanie tekstu itp.

Większość algorytmów ważności cech radzi sobie bardzo dobrze z cechami gęstymi i kategorycznymi. Jednak borykają się z cechami rzadkimi.

Opisujemy różne algorytmy ważności cech, szczegółowo opiszemy, jak dobrze współpracują one z różnymi typami funkcji. Podobnie, jak zobaczymy, tylko niektóre metody są niezależne od modelu. Niektóre metody działają w przypadku określonych typów modeli tj. modele liniowe i modele drzewiaste, ale niektóre podejścia dobrze się uogólniają i działają dla wszystkich modeli.

### Współczynniki regresji

Pomysł polega na wykorzystaniu współczynników modeli liniowych do **uchwycenia wartości, jaką każda cecha dodaje do prognozy**.

Obraz zawierający linia, Wykres, zrzut ekranu, diagram

Opis wygenerowany automatycznieW tym podejściu cechy można dopasować do modelu regresji liniowej, aby przewidzieć etykietę docelową. Wartości bezwzględne współczynników wygenerowanych przez model liniowy dla każdej cechy wskazują znaczenie odpowiednich cech. Większe wartości bezwzględne sugerują bardziej znaczący wpływ. Oto przykładowa implementacja regresji Ridge w zadaniu przewidywania cukrzycy z różnymi cechami liczbowymi. Ważność oparta na współczynnikach służy do obliczania ważności różnych cech.

Metoda ta jest pomocna w obliczaniu ważności cech modeli liniowych. Ma to te same zalety i wady, co korelacja. Wyjaśnialność jest zaletą tej metody, ale obsługuje ona tylko **złożone typy funkcji.**

### Korelacja z etykietą

Najprostszym sposobem oceny znaczenia cechy jest **jej korelacja ze zmienną docelową**. Im wyższa bezwzględna wielkość korelacji, tym ważniejsza jest ta cecha. Istnieje wiele sposobów obliczania korelacji. Formuła i kod, które omówiliśmy w naszych przykładach, opierają się na korelacji Pearsona.

Obraz zawierający tekst, zrzut ekranu, Wykres, linia

Opis wygenerowany automatycznieW tym wzorze obliczamy korelację dwóch zmiennych, X i Y, gdzie średnia jest średnią wartością odpowiedniej zmiennej, a n odnosi się do liczby wystąpień obu zmiennych X i Y. Oto przykładowa implementacja obliczania ważności przy użyciu korelacji z etykietą docelową.

Główną zaletą znaczenia opartego na korelacji jest prostota i możliwość wyjaśnienia metody. Ta metoda jest niezależna od typu modelu, który zbudujesz.

Głównym zastrzeżeniem dotyczącym znaczenia opartego na korelacji jest to, że nie pokrywa się ono między cechami. Kolejną wadą tej metody jest to, że nie działa ona zbyt dobrze w przypadku rzadkich obiektów.

### Funkcja Ablacja

Ablacja cech odnosi się do usunięcia zestawu cech z modelu. Prostym sposobem określenia ważności cechy jest sprawdzenie spadku wydajności modelu (mierzonego docelowymi metrykami, takimi jak auc-roc, auc-pr, precyzja i przypominanie), gdy funkcja zostanie usunięta. **Najważniejsza jest cecha, której usunięcie powoduje najbardziej znaczący spadek wydajności docelowej.** Ważność cech można posortować, sortując malejąco wzrost strat.

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznieTa metoda ma kilka zalet, bardzo dobrze działa w przypadku wszystkich typów cech i modeli. Brak tu nakładania się cech. Jest ona również łatwe do zrozumienia i wyjaśnienia.

Jednak główną wadą tego podejścia jest jego pracochłonność. Będziesz musiał przeprowadzić szkolenie w zakresie ablacji każdej cechy, a jeśli masz wiele cech, będzie to oznaczać znaczne zużycie zasobów i czasu.

### Znaczenie permutacji

Obraz zawierający tekst, zrzut ekranu, Czcionka, linia

Opis wygenerowany automatycznieTechnika ta polega na losowym mieszaniu wartości cechy i mierzeniu wynikającego z tego spadku wydajności modelu. Im bardziej znaczący spadek wydajności, tym ważniejsza jest dana cecha. To podejście jest niezależne od modelu i może być stosowane z różnymi algorytmami uczenia maszynowego. Oto przykład wdrożenia regresji grzbietu w zadaniu przewidywania cukrzycy z różnymi cechami liczbowymi. Ważność permutacji służy do obliczania ważności różnych cech.

To podejście ma te same zalety i wady, co ablacja cech. Większość bibliotek ważności funkcji online ma taką wersję. Dlatego łatwo jest go importować i używać w Pythonie.

### Znaczenie oparte na modelu drzewa

Ważność cech oparta na drzewach to technika stosowana do określania ważności cechy w modelach uczenia maszynowego opartych na drzewach, takich **jak losowe lasy** i **algorytmy wzmacniania gradientu** (np. XGBoost, LightGBM).

Obraz zawierający tekst, zrzut ekranu, Czcionka, linia

Opis wygenerowany automatycznieKrytycznym elementem w obliczaniu ważności cech opartych na drzewie jest **znaczenie** **Giniego** lub **zanieczyszczenie** (inpurity). Metoda ta mierzy znaczenie cechy na podstawie tego, jak bardzo zmniejsza ona zanieczyszczenie w węzłach drzewa. Wartość Giniego danej cechy jest obliczana jako **suma zanieczyszczeń** **zmniejszająca** się we wszystkich węzłach drzewa, które dzielą się na tę cechę. Cechy o większej wadze Giniego wskazują na większą zdolność różnicowania zmiennej docelowej.

Jak opisano w metodzie, ta technika ważności cech jest naturalnie stworzona dla modeli opartych na drzewach.

Ponadto nie działa dobrze w przypadku rzadkich obiektów, ale będzie działać dobrze w przypadku gęstych i kategorycznych obiektów.

### SHAP

Obraz zawierający tekst, zrzut ekranu, linia, Czcionka

Opis wygenerowany automatycznieSHAP (SHapley Additive exPlanations) to ujednolicony framework do interpretacji przewidywań modeli uczenia maszynowego. **Wartości SHAP umożliwiają przypisanie wkładu każdej cechy w predykcji, wskazując, w jakim stopniu każda cecha przyczynia się do ostatecznej prognozy w porównaniu do średniej prognozy.**

Różne kombinacje są tworzone przy użyciu pozostałych cech danej cechy, a działanie modelu jest oceniane dla tych kombinacji z daną cechą i bez niej. Te różnice w wydajności z rozważaną funkcją i bez niej mogą stanowić marginalny udział tej funkcji.

Wartość SHAP jest sumą ważoną takich marginalnych wkładów. Ważność cech oparta na SHAP zapewnia kompleksowe i niezależne od modelu podejście do zrozumienia wkładu każdej cechy w przewidywania modelu. Pozwala określić, które cechy są ważne i jak współdziałają z innymi cechami, zapewniając bardziej szczegółowe zrozumienie zachowania modelu.

### Wniosek

Z eksperymentów przeprowadzonych na podstawie studium przypadku dotyczącego cukrzycy wynika, że ​​BMI i s5 mają największe znaczenie w przypadku wielu metod.

Z drugiej strony badanie ablacyjne pokazuje, że usunięcie cech związanych z wiekiem i s3 prowadzi do zmniejszenia strat, co oznacza lepszą wydajność bez tych cech. Należy jednak również zauważyć, że wszystkie cechy mają mniejsze, ale są znaczące w porównaniu z innymi algorytmami, co wskazuje, że wszystkie cechy są ważne dla wydajności.

SHAP, który dokonuje szacunków opartych na ważonej ablacji w różnych ustawieniach kombinacji, wykazuje niezerowe znaczenie dla wszystkich cech.

Każda metoda wymieniona w tym artykule ma swoje zalety i wady. **Zawsze zaleca się wypróbowanie różnych strategii dotyczących ważności cech, aby dogłębnie zrozumieć model i dane**.

# Wizualizacja danych

## Definicja wizualizacji danych

**Wizualizacja danych to wizualna komunikacja danych lub przekształcanie surowych danych w spostrzeżenia, które czytelnicy mogą łatwo zinterpretować.**

lub

Wizualizacja danych odnosi się do technik używanych do przekazywania danych lub informacji przez kodowanie ich jako obiektów wizualnych (punktów, linii lub pasków) zawartych w grafice.

## Co sprawia, że wizualizacja danych jest skuteczna?

Wizualizacja danych jest skuteczna, jeśli jest wykonana poprawnie. Definiujemy dokładnie, kiedy wizualizacje danych spełniły swoje zadanie.

**Szybki test** - gdy ludzie mogą zinterpretować twoją wizualizację, zadając więcej pytań na temat wyświetlanych informacji w porównaniu do sposobu lub tego, co jest wyświetlane, wiesz, że jesteś na dobrej drodze.

Dlatego, aby być wysoce skutecznym, ważne jest zaprojektowanie odpowiednich wizualizacji danych, aby umożliwić tobie i członkom zespołu interpretację i podejmowanie decyzji na podstawie tego, co obserwują.

Jak to zrobimy? To proste. Tworzymy odpowiednie wizualizacje poprzez zrozumienie różnych rodzajów wizualizacji.

## 5 typów wizualizacji danych

### Czasowy

Przykłady czasowej wizualizacji danych obejmują:

* Wykresy rozrzutu
* Diagramy obszarów polarnych
* Sekwencje szeregów czasowych
* Osie czasu
* Wykresy liniowe

Wizualizacje danych należą do kategorii czasowej, jeśli spełniają dwa warunki: że są liniowe i że są jednowymiarowe. Wizualizacje czasowe zwykle przedstawiają linie, które albo biegną samodzielnie, albo nakładają się na siebie, z czasem rozpoczęcia i zakończenia.

### Hierarchiczny

Przykłady hierarchicznych wizualizacji danych obejmują:

* Diagramy drzew
* Wykresy pierścieniowe
* Diagramy Sunburst

Wizualizacje danych należące do kategorii hierarchicznej to te, które porządkują grupy w większych grupach. Hierarchiczne wizualizacje najlepiej nadają się, jeśli chcesz wyświetlić klastry informacji, zwłaszcza jeśli pochodzą one z jednego punktu początkowego.

Minusem tych wykresów jest to, że są one bardziej złożone i trudne do odczytania, dlatego najczęściej stosuje się diagram drzewa. Jest najprostszy do interpretacji ze względu na ścieżkę liniową.

### Sieć

Przykłady wizualizacji danych sieciowych obejmują:

* Wykresy macierzy
* Diagramy węzłów i krawędzi (Grafy)
* Chmury słów
* Diagramy przepływowe

Zestawy danych łączą się głęboko z innymi zestawami danych. Wizualizacje danych sieciowych pokazują ich wzajemne relacje w sieci. Innymi słowy, demonstrowanie związków między zestawami danych bez niepotrzebnych wyjaśnień.

### Wielowymiarowy

Przykłady wielowymiarowych wizualizacji danych obejmują:

* Wykresy rozrzutu
* Wykresy kołowe
* Diagramy Venna
* Ułożone wykresy słupkowe
* Histogramy

Podobnie jak nazwa, wielowymiarowe wizualizacje danych mają wiele wymiarów. Oznacza to, że w mieszance zawsze znajdują się 2 lub więcej zmiennych, aby utworzyć wizualizację danych 3D. Ze względu na wiele współbieżnych warstw i zestawów danych, tego rodzaju wizualizacje wydają się być najbardziej żywymi lub przyciągającymi wzrok elementami wizualnymi.

### Geo-przestrzenny

Przykłady wizualizacji danych geoprzestrzennych obejmują:

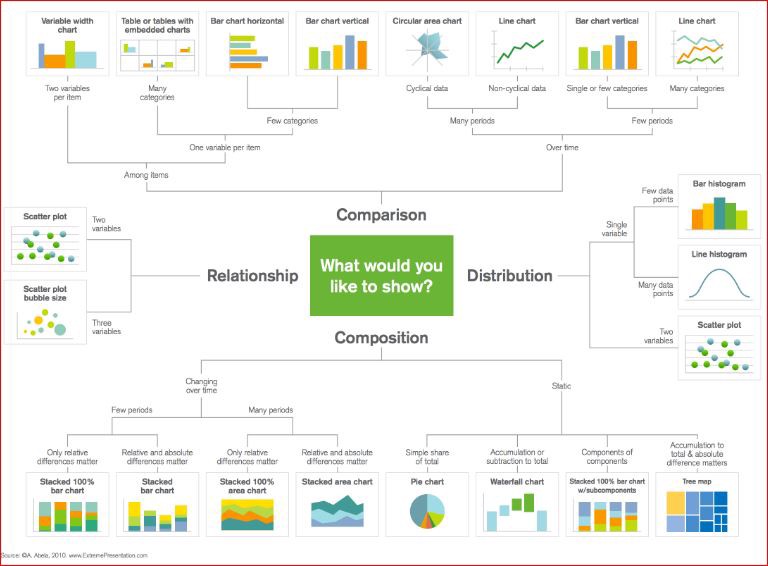
* Mapa przepływu
* Mapa zagęszczenia
* Kartogram
* Mapa ciepła

Wizualizacje danych geo-przestrzennych lub przestrzennych odnoszą się do rzeczywistych fizycznych lokalizacji, nakładając znane mapy z różnymi punktami danych. Tego rodzaju wizualizacje danych są powszechnie używane do wyświetlania sprzedaży lub przejęć w czasie i mogą być najbardziej rozpoznawalne dzięki ich zastosowaniu w kampaniach politycznych lub w celu pokazania penetracji rynku w międzynarodowych korporacjach.

# Jak dobrać właściwą wizualizację do mojego zadania eksploracji danych?

Prezentacja danych i informacji nie polega tylko na wybraniu dowolnego projektu wizualizacji danych. Dopasowanie danych do właściwej wizualizacji informacji rozpoczyna się od odpowiedzi na 5 kluczowych pytań:

1. Jaki związek próbuję zrozumieć między moimi zbiorami danych?
2. Czy chcę zrozumieć dystrybucję danych i szukać wartości odstających?
3. Czy chcę porównać wiele wartości lub przeanalizować jedną wartość w czasie?
4. Czy jestem zainteresowany analizowaniem trendów w moich zestawach danych?
5. Czy ta wizualizacja jest ważną częścią mojej nadrzędnej historii danych?



# Co chcesz zwizualizować, pokazać?

## Porównania (ze współrzędnymi i bez)

Metody wizualizacji, które pomagają pokazać **różnice lub podobieństwa między wartościami.**

## Proporcje

Metody wizualizacji wykorzystujące **rozmiar lub obszar do pokazywania różnic lub podobieństw między wartościami lub dla części w całości.**

## Relacje

Metody wizualizacji, które pokazują **relacje i połączenia między danymi lub pokazują korelacje między dwiema lub więcej zmiennymi.**

## Hierarchia

Metody wizualizacji, które pokazują, **w jaki sposób dane lub obiekty są uszeregowane i uporządkowane** razem w organizacji lub systemie.

## Dystrybucja

Metody wizualizacji, które wyświetlają **częstotliwość, rozkład danych w danym przedziale lub ich grupowanie.**

## Ruch lub przepływ

Metody wizualizacji, które są przydatne **do wyświetlania danych o ruchu lub przepływu danych**.

## Wzory

Metody wizualizacji, które mogą **ujawnić formy lub wzorce w danych, aby nadać im znaczenie.**

## Zasięg (Zakres)

Metody wizualizacji, które wyświetlają **różnice między górną a dolną granicą na skali.**

## Dane w czasie

Metody wizualizacji, które pokazują **dane w określonym przedziale czasowym w celu wyświetlenia trendów lub zmian w czasie.**

# Porównania

## Porównania (ze współrzędnymi) ?

Metody wizualizacji, które pomagają pokazać różnice lub podobieństwa między wartościami.

### Obraz zawierający tekst, zrzut ekranu, diagram, design Opis wygenerowany automatycznieWykres słupkowy

Znany jako **wykres słupkowy** lub **wykres kolumnowy**.

Klasyczny wykres słupkowy używa poziomych lub pionowych słupków (wykres kolumnowy), aby pokazać dyskretne, numeryczne porównania między kategoriami.

Jedna oś wykresu pokazuje porównywane kategorie, a druga oś reprezentuje dyskretną skalę wartości.

***Funkcje***

**Porównania**

**Wzory**

Wykresy słupkowe różnią się od histogramów, ponieważ nie wyświetlają ciągłego rozwoju w określonym przedziale czasu.

Dyskretne dane wykresu słupkowego są danymi kategorycznymi i dlatego odpowiadają na pytanie „ile?” w każdej kategorii.

Jedną z głównych wad wykresów słupkowych jest to, że etykietowanie staje się problematyczne, gdy występuje duża liczba słupków.

### Wykres słupkowy grupowy

**Opis**

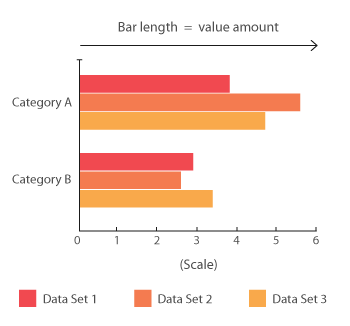
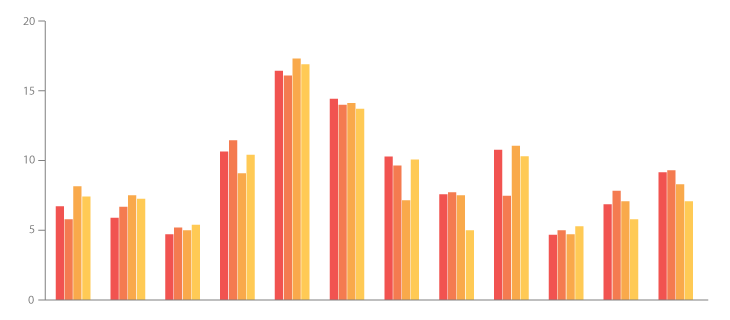
Znany również jako zgrupowany wykres słupkowy lub klastrowany wykres słupkowy.

Ta odmiana wykresu słupkowego jest używana, gdy dwie lub więcej serii danych jest kreślonych obok siebie i grupowanych razem w kategorie, wszystkie na tej samej osi.

Długość wykresu słupkowego, podobnie jak wykres słupkowy, służy do przedstawienia dyskretnych, numerycznych porównań między kategoriami. Każda seria danych ma przypisany indywidualny kolor lub inny odcień tego samego koloru, aby je rozróżnić. Każda grupa prętów jest następnie oddalona od siebie.

Stosowanie wykresów słupkowych z wieloma zestawami zwykle służy do porównywania zgrupowanych zmiennych lub kategorii z innymi grupami z tymi samymi zmiennymi lub typami kategorii. Wykresy słupkowe z wieloma zestawami mogą być również używane do porównywania ze sobą mini-histogramów, więc każdy słupek w grupie reprezentuje znaczące przedziały zmiennej.

Minusem wielu zestawów wykresów słupkowych jest to, że coraz trudniej jest odczytać więcej słupków w jednej grupie.



**Funkcje**

**Porównania**

**Dystrybucja**

**Wzory**

**Relacje**

### Wykres Marimekko

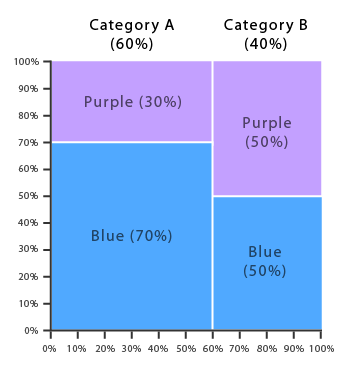
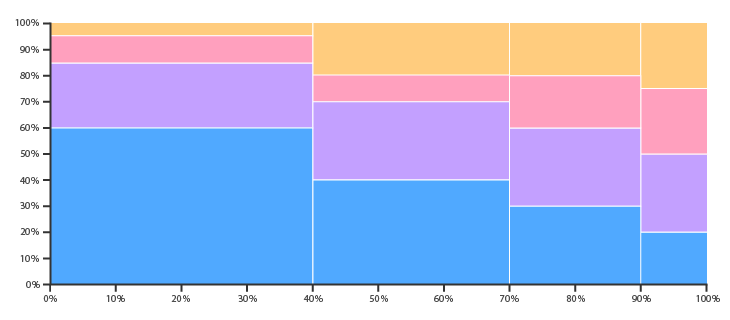
**Opis**

Znany również jako wykres mozaiki.

Wykresy Marimekko służą do wizualizacji danych kategorialnych za pomocą pary zmiennych. Na wykresie Marimekko obie osie są zmienne wraz ze skalą procentową, która określa zarówno szerokość, jak i wysokość każdego segmentu.

Wykresy Marimekko działają więc jako rodzaj dwukierunkowego 100% skumulowanego wykresu słupkowego. Umożliwia to wykrycie związków między kategoriami i ich podkategoriami za pomocą dwóch osi.

Główne wady wykresów Marimekko polegają na tym, że mogą być trudne do odczytania, szczególnie w przypadku wielu segmentów. Ponadto trudno jest dokładnie porównywać każdy segment, ponieważ nie wszystkie są ułożone obok siebie na wspólnej linii bazowej. Dlatego wykresy Marimekko lepiej nadają się do przedstawienia bardziej ogólnego przeglądu danych.



**Funkcje**

**Porównania**

**Część do całości**

**Proporcje**

**Relacje**

### Wykres róży

**Opis**

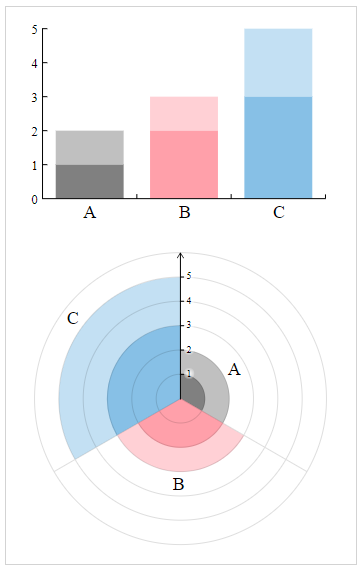
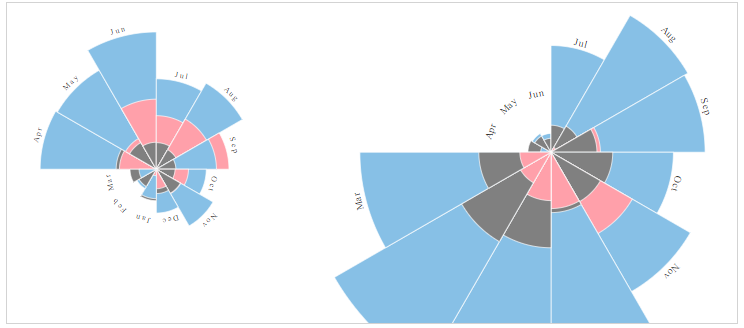
Znany również jako wykres Coxcomba, wykres obszaru polarnego.

Ta mapa była używana przez statystę i reformatora medycznego, Florence Nightingale, do informowania o możliwej do uniknięcia śmierci żołnierzy podczas wojny krymskiej.

Wykresy róż są rysowane na biegunowej siatce współrzędnych. Każda kategoria lub przedział w danych jest podzielony na równe segmenty na tym wykresie promieniowym. To, jak daleko każdy segment rozciąga się od środka osi biegunowej, zależy od wartości, którą reprezentuje.

Tak więc każdy pierścień od środka siatki biegunowej może być użyty jako skala do wykreślenia wielkości segmentu i przedstawienia wyższej wartości. Dlatego ważne jest, aby za pomocą wykresu róży zauważyć, że to obszar, a nie promień segmentu reprezentujący jego wartość.

Główną wadą wykresów róż Słowika jest to, że zewnętrzne segmenty są bardziej naciskane ze względu na ich większy rozmiar powierzchni. To nieproporcjonalnie reprezentuje wzrost wartości.



**Funkcje**

**Porównania**

**Dane w czasie**

**Proporcje**

### Wykres równoległych współrzędnych

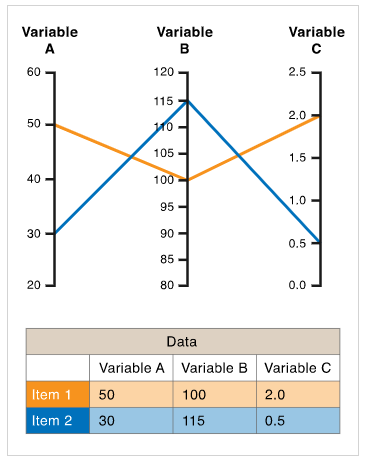
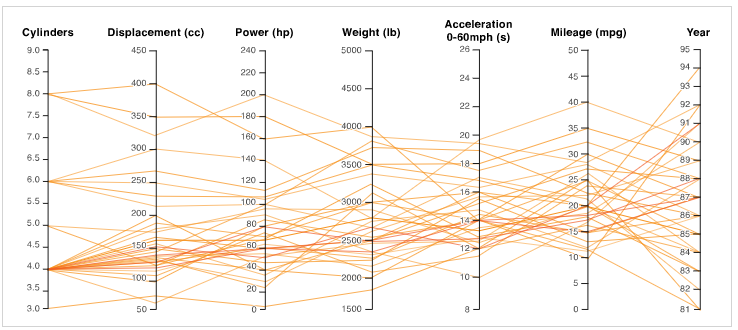
**Opis**

Ten typ wizualizacji służy do kreślenia wielowymiarowych danych liczbowych. Wykresy równoległych współrzędnych są idealne do porównywania wielu zmiennych razem i dostrzegania relacji między nimi. Na przykład, jeśli trzeba porównać szereg produktów o tych samych atrybutach (porównanie specyfikacji komputera lub samochodu w różnych modelach).

Na wykresie współrzędnych równoległych każdej zmiennej przypisano własną oś, a wszystkie osie są umieszczone równolegle względem siebie. Każda oś może mieć inną skalę, ponieważ każda zmienna działa na innej jednostce miary, lub wszystkie osie mogą być znormalizowane, aby wszystkie skale były jednolite. Wartości są wykreślane jako szereg linii, które łączą się we wszystkich osiach. Oznacza to, że każda linia jest zbiorem punktów umieszczonych na każdej osi, które zostały ze sobą połączone.

Kolejność ułożenia osi może mieć wpływ na sposób, w jaki czytnik rozumie dane. Jednym z powodów jest to, że relacje między sąsiednimi zmiennymi są łatwiejsze do zauważenia niż w przypadku zmiennych niesąsiadujących. Ponowne uporządkowanie osi może pomóc w wykryciu wzorców lub korelacji między zmiennymi.

Wadą równoległych wykresów współrzędnych jest to, że mogą one być nadmiernie zagracone, a zatem nieczytelne, gdy są bardzo zagęszczone. Najlepszym sposobem na rozwiązanie tego problemu jest interaktywność i technika znana jako „Szczotkowanie”. Pędzel podświetla wybraną linię lub kolekcję linii, jednocześnie wyciszając wszystkie pozostałe. Dzięki temu możesz wydzielić fragmenty interesującego Cię wątku, odfiltrowując hałas.



**Funkcje**

**Porównania**

**Relacje**

**Wzory**

### Wykres radarowy

**Opis**

Znany jako: wykres pająka, wykres sieci, wykres polarny, wykresy gwiezdne.

Mapy radarowe są sposobem porównywania wielu zmiennych ilościowych. To sprawia, że ​​są przydatne do sprawdzania, które zmienne mają podobne wartości lub czy są jakieś wartości odstające między każdą zmienną. Wykresy radarowe są również przydatne do sprawdzania, które zmienne mają wysoką lub niską ocenę w zbiorze danych, co czyni je idealnymi do wyświetlania wydajności.

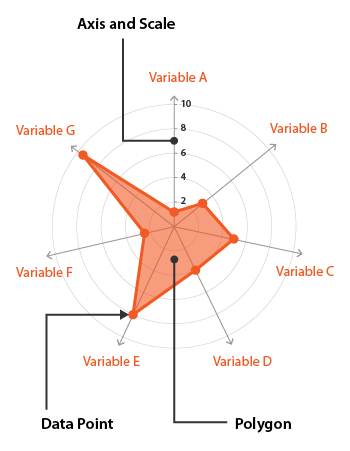
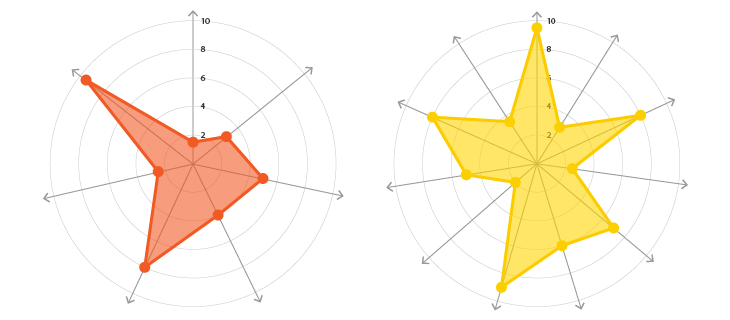
Każda zmienna jest wyposażona w oś, która zaczyna się od środka. Wszystkie osie są rozmieszczone promieniowo, w równych odległościach między sobą, przy zachowaniu tej samej skali między wszystkimi osiami. Linie siatki łączące oś z osią są często używane jako wskazówki. Każda wartość zmiennej jest wykreślana wzdłuż jej indywidualnej osi i wszystkich zmiennych w zbiorze danych i łączona razem w celu utworzenia wielokąta.

Istnieją jednak poważne wady w mapach radarowych:

Posiadanie wielu wielokątów na jednym wykresie radarowym utrudnia czytanie, jest mylące i zbyt zagracone. Zwłaszcza jeśli wielokąty są wypełnione, ponieważ górny wielokąt zakrywa wszystkie pozostałe wielokąty pod nim.

Posiadanie zbyt wielu zmiennych tworzy zbyt wiele osi, a także może utrudniać czytanie i komplikowanie wykresu. Dlatego dobrą praktyką jest upraszczanie map radarowych i ograniczanie liczby używanych zmiennych.

Kolejną wadą radarów jest to, że nie są one tak dobre do porównywania wartości dla każdej zmiennej. Nawet z pomocą prowadnicy siatki przypominającej pajęczynę. Porównywanie wartości na jednej prostej osi jest znacznie łatwiejsze.



**Funkcje**

**Porównania**

**Relacje**

**Wzory**

### Radialny wykres kolumnowy

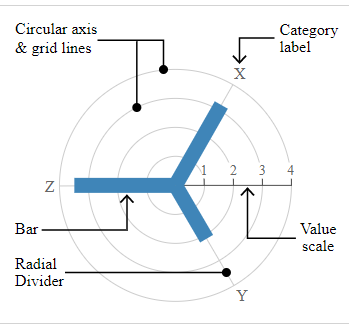
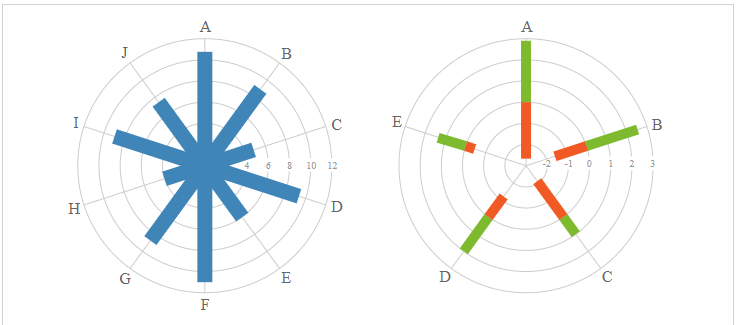
**Opis**

Znany również jako wykres kołowej kolumny lub wykres gwiazdy.

Ten typ wykresu używa siatki koncentrycznych okręgów do kreślenia słupków. Każde koło na wykresie reprezentuje wartość na skali, podczas gdy dzielniki promieniowe (linie rozciągające się od środka) są używane dla każdej kategorii lub przedziału (jeśli jest to histogram).

Zazwyczaj niższe wartości na skali zaczynają się od środka i rosną z każdym okręgiem. Jednak wartości ujemne można również wyświetlić na Radialnym Wykresie Kolumnowym, mając zero zaczynające się od dowolnego zewnętrznego koła (od środkowego) i wszystkie znajdujące się w nim koła użyte do wartości ujemnych.

Słupki zwykle zaczynają się od środka i rozciągają na zewnątrz, jednak zakresy mogą być pokazane ze zmiennymi punktami początkowymi, jak na wykresie rozpiętości. Słupki można również ustawiać w stosy w taki sam sposób jak skumulowany wykres słupkowy.



**Funkcja**

**Porównania**

### Obraz zawierający tekst, zrzut ekranu, diagram, Równolegle Opis wygenerowany automatycznieMapa termiczna

**Opis**

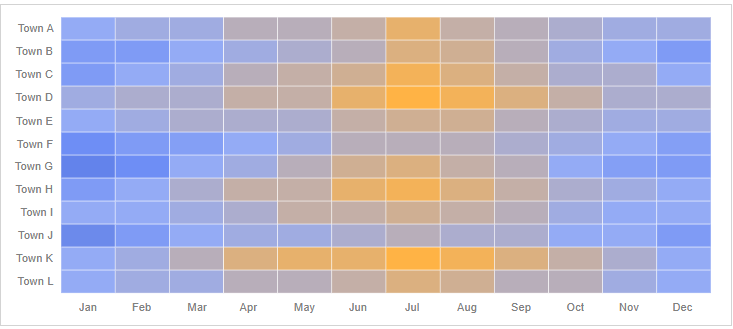
Mapy termiczne wizualizują dane poprzez różne kolory. Po zastosowaniu do formatu tabelarycznego mapy cieplne są przydatne do krzyżowego badania danych na wielu odmianach, poprzez umieszczanie zmiennych w wierszach i kolumnach oraz kolorowanie komórek w tabeli. Mapy termiczne są dobre do pokazywania wariancji między wieloma zmiennymi, ujawniania dowolnych wzorców, wyświetlania, czy jakieś zmienne są do siebie podobne oraz do wykrywania, czy istnieją między nimi jakieś korelacje.

Zazwyczaj wszystkie wiersze należą do jednej kategorii (etykiety wyświetlane po lewej lub prawej stronie), a wszystkie kolumny należą do innej kategorii (etykiety wyświetlane u góry lub u dołu). Poszczególne wiersze i kolumny są podzielone na podkategorie, które wszystkie pasują do siebie w macierzy. Komórki zawarte w tabeli zawierają dane kategoryczne oznaczone kolorami lub dane liczbowe oparte na skali kolorów. Dane zawarte w komórce są oparte na relacji między dwiema zmiennymi w łączącym rzędzie i kolumnie.

Legenda jest wymagana obok mapy termicznej, aby można ją było z powodzeniem odczytać. Dane kategorialne są kodowane kolorami, podczas gdy dane liczbowe wymagają skali kolorów, która łączy się z jednym kolorem w drugi, aby przedstawić różnicę w wysokich i niskich wartościach. Wyboru jednolitych kolorów można użyć do przedstawienia wielu zakresów wartości (0–10, 11–20, 21–30 itd.) Lub można użyć skali gradientu dla pojedynczego zakresu (na przykład 0–100) przez zmieszanie dwóch lub więcej kolorów razem.

Ze względu na poleganie na kolorze w celu przekazywania wartości, mapy termiczne są lepiej dostosowane do wyświetlania bardziej ogólnego widoku danych liczbowych, ponieważ trudniej jest dokładnie określić różnice między odcieniami kolorów i wyodrębnić określone punkty danych (chyba że oczywiście uwzględnij surowe dane w komórkach).

Mapy termiczne można również wykorzystać do wyświetlania zmian danych w czasie, jeśli jeden z wierszy lub kolumn jest ustawiony na przedziały czasowe. Przykładem może być użycie mapy cieplnej do porównania zmian temperatury w ciągu roku w wielu miastach, aby zobaczyć, gdzie są najgorętsze lub najzimniejsze miejsca. Aby wiersze zawierały listę miast do porównania, kolumny zawierają co miesiąc, a komórki zawierają wartości temperatur.



**Funkcje**

**Porównania**

**Dane w czasie**

**Wzory**

**Relacje**

### Kartogram

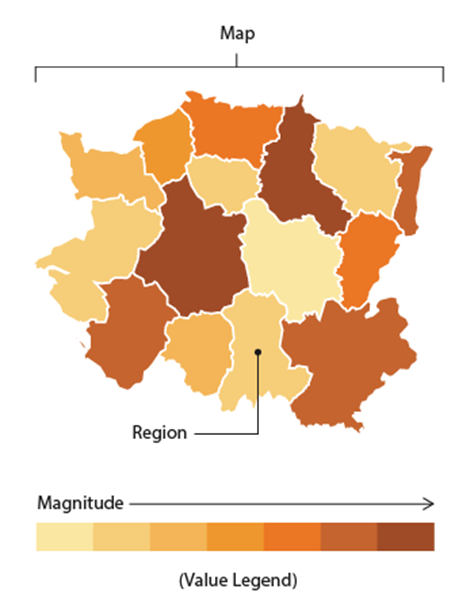
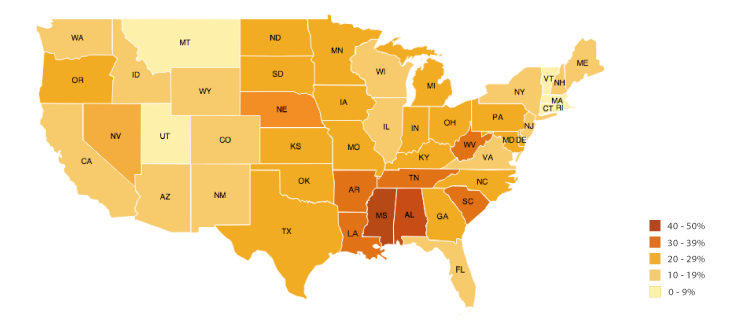
**Opis**

Kartogramy pokazują podzielone obszary geograficzne lub regiony, które są kolorowe, zacieniowane lub wzorowane względem zmiennej danych. Umożliwia to wizualizację wartości w obszarze geograficznym, który może pokazywać zmienność lub wzorce w wyświetlanej lokalizacji.

Zmienna danych używa progresji kolorów do reprezentowania siebie w każdym regionie mapy. Zwykle może to być mieszanie jednego koloru do drugiego, pojedyncza zmiana odcienia, przezroczysta do nieprzezroczystej, jasna do ciemnej lub całe spektrum kolorów.

Wadą użycia koloru jest to, że nie można dokładnie odczytać ani porównać wartości z mapy. Inną kwestią jest to, że większe regiony wydają się bardziej zaakcentowane niż mniejsze, więc wpływa to na postrzeganie cieniowanych wartości przez widza.

Częstym błędem podczas tworzenia Kartogramów jest zakodowanie surowych wartości danych (takich jak populacja) zamiast używania znormalizowanych wartości (na przykład obliczanie populacji na kilometr kwadratowy) w celu uzyskania mapy gęstości.



**Funkcje**

**Porównania**

**Lokalizacja**

**Wzory**

### Piktogram

**Opis**

Znany również jako wykres piktograficzny, wykres obrazkowy.

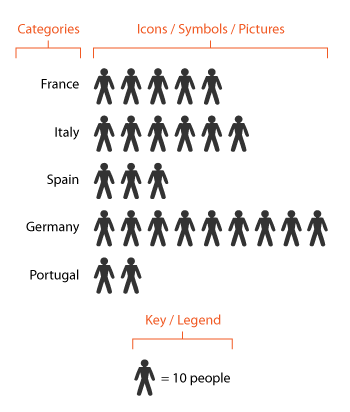
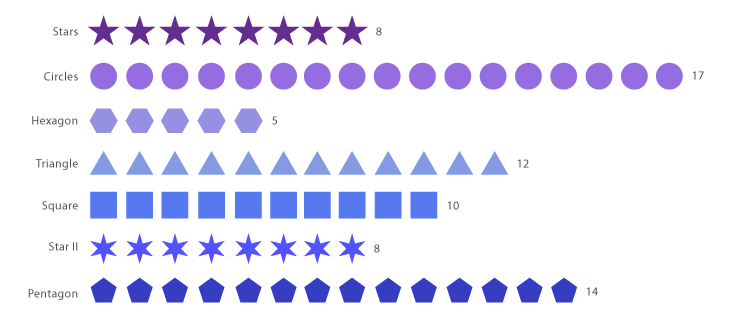
Wykresy piktogramów używają ikon, aby uzyskać bardziej wciągający ogólny widok małych zestawów dyskretnych danych. Zazwyczaj ikony reprezentują temat lub kategorię danych, na przykład dane dotyczące populacji wykorzystują ikony osób. Każda ikona może reprezentować jedną jednostkę lub dowolną liczbę jednostek (np. Każda ikona reprezentuje 10). Zestawy danych są porównywane obok siebie w kolumnach lub rzędach ikon, w celu porównania każdej kategorii ze sobą.

Korzystanie z ikon może czasem pomóc w przezwyciężeniu różnic w języku, kulturze i edukacji. Ikony mogą również dać bardziej reprezentatywny widok danych. Na przykład, jeśli dane zawierają 5 samochodów, na wykresie wyświetla się 5 ikon samochodów.

Podczas korzystania z wykresów piktogramowych należy unikać dwóch rzeczy:

Używanie ich do dużych zestawów danych, co utrudnia policzenie wartości na wykresie.

Wyświetlanie częściowych ikon, ponieważ może to wprowadzić zamieszanie w ich reprezentacji.



**Funkcje**

**Porównania**

**Dystrybucja**

## Proporcje

### Skumulowany wykres słupkowy

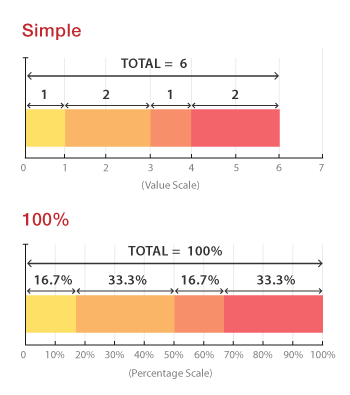
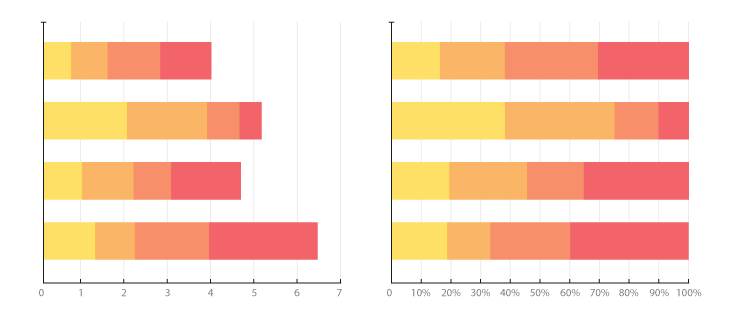
**Opis**

Skumulowane wykresy słupkowe dzielą na siebie słupki wielu zestawów danych. Służą do pokazania, w jaki sposób większa kategoria jest podzielona na mniejsze kategorie i jaki jest stosunek każdej części do całkowitej kwoty. Istnieją dwa rodzaje skumulowanych wykresów słupkowych:

**Proste skumulowane wykresy słupkowe** umieszczają każdą wartość dla segmentu po poprzedniej. Łączna wartość paska to wszystkie wartości segmentów dodane razem. Idealny do porównywania całkowitych kwot dla każdej grupy / segmentowanego paska.

**Wykresy słupkowe 100%** pokazują procent całości każdej grupy i są wykreślane jako procent każdej wartości do całkowitej kwoty w każdej grupie. Ułatwia to obserwowanie względnych różnic między wielkościami w każdej grupie.

Jedną z głównych wad stosowania wykresów słupkowych jest to, że trudniej jest odczytać więcej segmentów na każdym słupku. Porównywanie każdego segmentu ze sobą jest trudne, ponieważ nie są one wyrównane na wspólnej linii bazowej.



**Funkcje**

**Proporcje**

**Porównania**

Kiedy 100% wykres słupkowy:

**Część do całości**

### Wykres kołowy

Opis

Wykresy kołowe, szeroko stosowane w prezentacjach i biurach, pomagają pokazać proporcje i procenty między kategoriami, dzieląc okrąg na proporcjonalne segmenty. Każda długość łuku reprezentuje część każdej kategorii, podczas gdy pełny okrąg reprezentuje całkowitą sumę wszystkich danych, równą 100%.

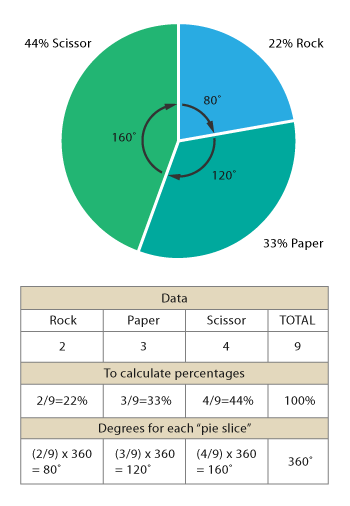
Wykresy kołowe są idealne, aby dać czytelnikowi szybki obraz proporcjonalnego rozkładu danych. Jednak głównymi wadami wykresów kołowych są:

Nie mogą pokazywać więcej niż kilku wartości, ponieważ wraz ze wzrostem liczby wyświetlanych wartości rozmiar każdego segmentu / plasterka staje się mniejszy. To sprawia, że ​​nie nadają się do dużych ilości danych.

Zajmują więcej miejsca niż ich alternatywy, na przykład 100% skumulowany wykres słupkowy. Głównie ze względu na ich rozmiar i zwykłą potrzebę legendy.

Nie są one doskonałe do dokonywania dokładnych porównań między grupami wykresów kołowych. W związku z tym trudniej jest rozróżnić rozmiar przedmiotów według powierzchni, jeśli chodzi o długość.

Mimo to, porównanie danej kategorii (jednego wycinka) w ramach jednego wykresu kołowego może być często bardziej skuteczne.



**Funkcje**

**Proporcje**

**Porównania**

**Część do całości**

### Diagram (wykres) pierścieniowy

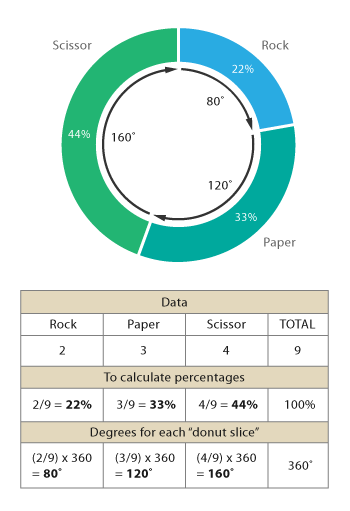
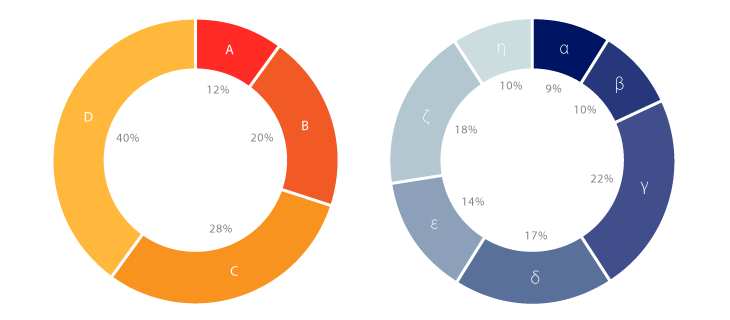
**Opis**

Wykres pączka jest zasadniczo wykresem kołowym z wyciętym obszarem środka.

Wykresy kołowe są czasami krytykowane za skupianie czytelników na proporcjonalnych obszarach wycinków względem siebie i na wykresie jako całości. To sprawia, że trudno jest zobaczyć różnice między plasterkami, zwłaszcza gdy próbujesz porównać wiele wykresów kołowych razem.

Wykres pączków w pewien sposób rozwiązuje ten problem, nie podkreślając wykorzystania obszaru. Zamiast tego czytelnicy skupiają się bardziej na czytaniu długości łuków, niż na porównywaniu proporcji między plasterkami.

Ponadto wykresy pierścieniowe są bardziej wydajne pod względem przestrzeni niż wykresy kołowe, ponieważ pusta przestrzeń wewnątrz wykresu pierścieniowego może służyć do wyświetlania informacji w nim zawartych.



**Funkcje**

**Proporcje**

**Porównania**

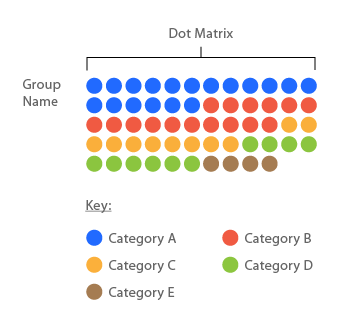
**Część do całości**

### Diagram macierzy punktowej

Opis

Diagram macierzy punktowej wyświetlają dyskretne dane w jednostkach kropek, z których każdy jest kolorowy, aby reprezentować określoną kategorię i zgrupowane razem w macierzy. Służą do szybkiego przeglądu rozkładu i proporcji każdej kategorii w zbiorze danych, a także do porównania rozkładu i proporcji w innych zestawach danych w celu wykrycia wzorców.

Gdy w zestawie danych używana jest tylko jedna zmienna / kategoria, a wszystkie kropki są tego samego koloru, można użyć wykresu macierzy kropek, aby przede wszystkim pokazać proporcje.



**Funkcje**

**Proporcje**

**Porównania**

**Dystrybucja**

**Wzory**

## Dystrybucje

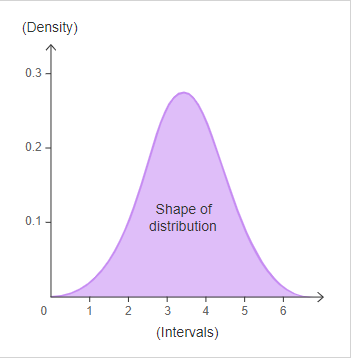
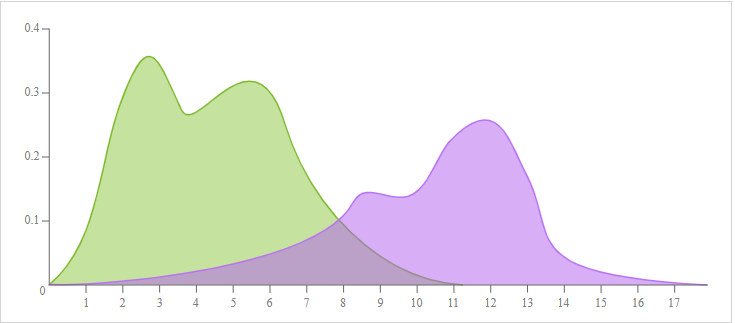
### Wykres zagęszczenia

Opis

Znany jako wykresy gęstości jądra, wykres śledzenia gęstości.

Wykres gęstości wizualizuje rozkład danych w ciągłym odstępie czasu lub okresie. Ten wykres jest odmianą histogramu, który wykorzystuje wygładzanie jądra do wykreślania wartości, umożliwiając bardziej płynne rozkłady poprzez wygładzenie szumu. Piki wykresu gęstości pomagają wyświetlać, gdzie wartości są skoncentrowane w przedziale.

Zaletą wykresów gęstości w porównaniu z histogramami jest to, że lepiej określają kształt rozkładu, ponieważ nie wpływa na nie liczba używanych pojemników (każdy słupek użyty w typowym histogramie). Histogram składający się tylko z 4 przedziałów nie dałby wystarczająco wyraźnego kształtu dystrybucji, tak jak zrobiłby to histogram 20-bin. Jednak w przypadku wykresów gęstości nie stanowi to problemu.



**Funkcje**

Dystrybucja

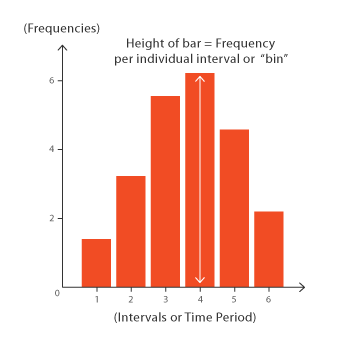
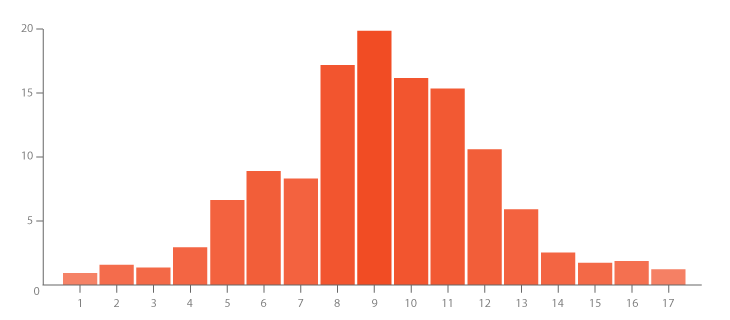
Wzory

### Histogram

**Opis**

Histogram wizualizuje rozkład danych w ciągłym przedziale lub określonym okresie czasu. Każdy słupek histogramu przedstawia częstotliwość w tabelach dla każdego interwału / przedziału.

Histogramy pomagają oszacować, gdzie koncentrują się wartości, jakie są skrajności i czy są jakieś luki lub nietypowe wartości. Są one również przydatne do przybliżenia rozkładu prawdopodobieństwa.



***Funkcje***

**Porównania**

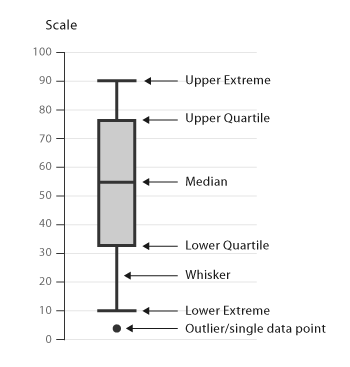
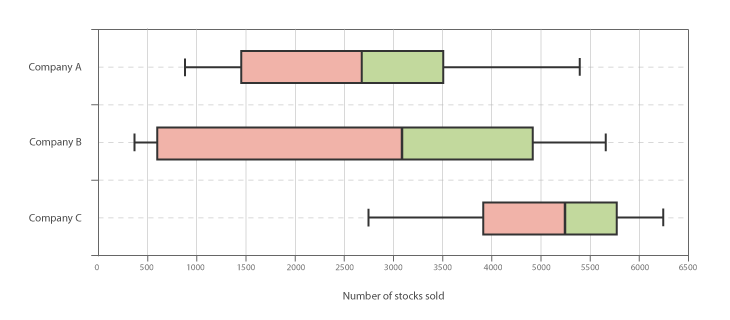
**Dane w czasie**

**Dystrybucja**

**Wzory**

**Zakres**

### Obraz zawierający czarne, ciemność Opis wygenerowany automatycznieWykres pudełkowy z wąsami



***Funkcje***

**Dystrybucja**

**Zakres**

### Piramida ludności

**Opis**

Znany jako piramida wieku i płci.

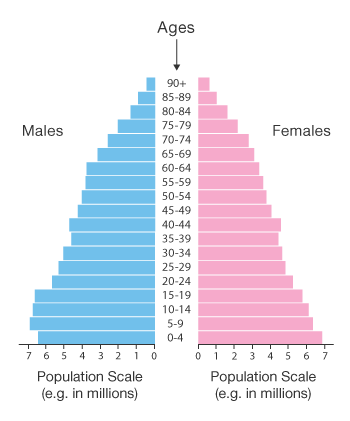
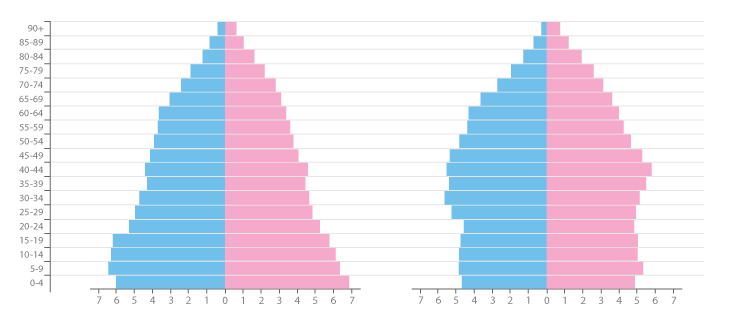
Piramida ludnościowa to para histogramów (dla każdej płci), które pokazują rozkład populacji we wszystkich grupach wiekowych i obu płciach. Oś X służy do wykreślania liczby ludności, a oś Y wymienia wszystkie grupy wiekowe.

Piramidy populacji są idealne do wykrywania zmian lub różnic we wzorcach populacji. Piramidy z wieloma populacjami mogą być używane do porównywania wzorców między narodami lub wybranymi grupami populacji.

Kształt piramidy populacji można wykorzystać do interpretacji populacji. Na przykład piramida z bardzo szeroką podstawą i wąską górną częścią sugeruje populację o wysokiej dzietności i śmiertelności. Natomiast piramida z szerszą górną połową i węższą podstawą sugerowałaby starzenie się populacji o niskim współczynniku dzietności.

Piramidy populacji można również wykorzystać do spekulacji przyszłym rozwojem populacji. Starzejąca się populacja, która się nie rozmnaża, w końcu napotkałaby takie problemy, jak (nie?) wystarczająca liczba potomstwa do opieki nad osobami starszymi. Inne teorie, takie jak „Youth Bulge”, stwierdzają, że gdy występuje szerokie wybrzuszenie w wieku 16-30 lat, szczególnie u mężczyzn, prowadzi to do niepokojów społecznych, wojny i terroryzmu.

To sprawia, że ​​piramidy populacji są przydatne w takich dziedzinach, jak ekologia, socjologia i ekonomia.



**Funkcje**

**Porównania**

**Dystrybucja**

**Wzory**

## Dane w czasie

### Wykres liniowy

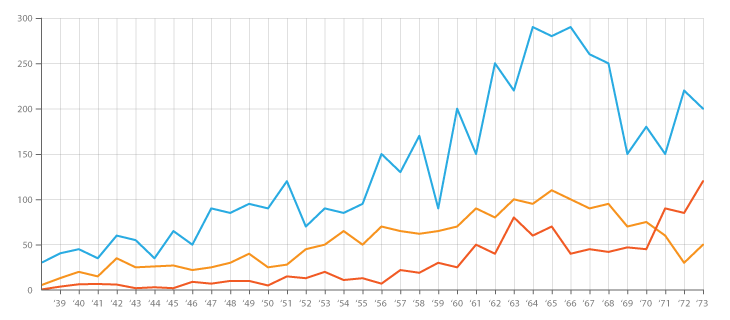
**Opis**

Wykresy liniowe służą do wyświetlania wartości ilościowych w ciągłym przedziale lub okresie. Wykres liniowy jest najczęściej używany do pokazywania trendów i analizowania zmian danych w czasie.

Wykresy liniowe są rysowane przez wykreślenie punktów danych na kartezjańskiej siatce współrzędnych, a następnie połączenie linii między wszystkimi tymi punktami. Zazwyczaj oś y ma wartość ilościową, podczas gdy oś x jest skalą czasu lub sekwencją przedziałów. Wartości ujemne mogą być wyświetlane poniżej osi x.

Kierunek linii na wykresie działa jako ładna metafora dla danych: nachylenie w górę wskazuje, gdzie wzrosły wartości, a nachylenie w dół wskazuje, gdzie spadły wartości. Podróż linii przez wykres może tworzyć wzorce, które ujawniają trendy w zbiorze danych.

Po zgrupowaniu z innymi liniami (innymi seriami danych) poszczególne linie można porównać ze sobą. Należy jednak unikać używania więcej niż 3-4 linii na wykres, ponieważ powoduje to, że wykres jest bardziej zagracony i trudniejszy do odczytania. Rozwiązaniem tego problemu jest podzielenie wykresu na mniejsze wielokrotności (mieć mały Wykres liniowy dla każdej serii danych)



***Funkcje***

**Dane w czasie**

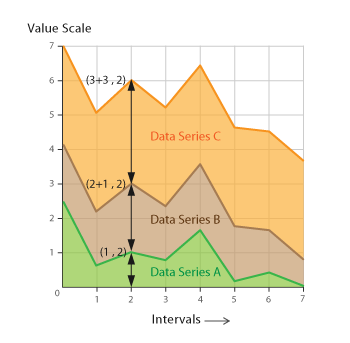
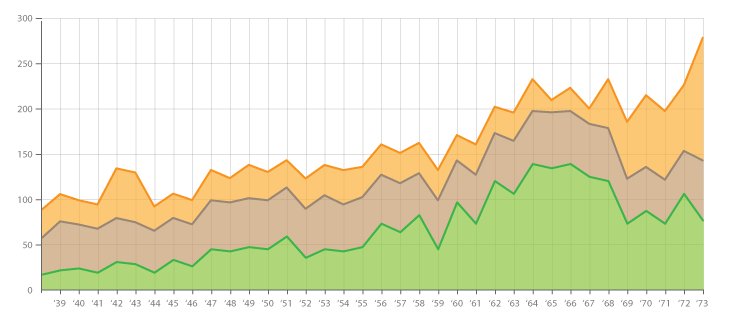
**Wzory**

### Wykres warstwowy

Opis

Wykresy warstwowe działają w taki sam sposób jak proste wykresy powierzchniowe, z wyjątkiem użycia wielu serii danych, które rozpoczynają każdy punkt od punktu pozostawionego przez poprzednią serię danych.

Cały wykres przedstawia sumę wszystkich wykreślonych danych. Wykresy warstwowe również używają obszarów do przenoszenia liczb całkowitych, więc nie działają one na wartości ujemne. Ogólnie są one przydatne do porównywania wielu zmiennych zmieniających się w danym przedziale.



**Funkcje**

**Dane w czasie**

**Porównania**

**Wzory**

### Wykres strumieniowy

Opis

Znany również jako ThemeRiver.

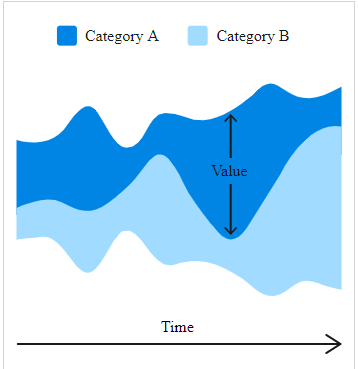
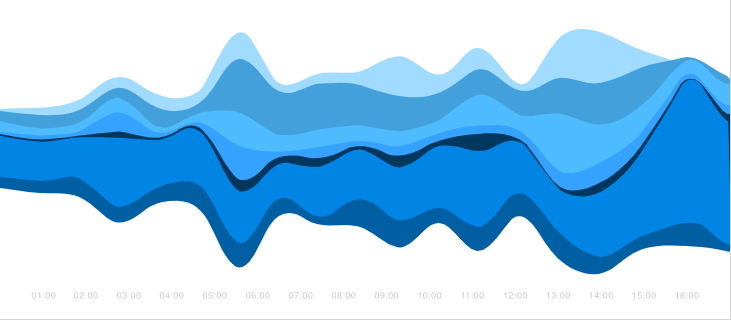
Ten rodzaj wizualizacji jest odmianą wykresu skumulowanego, ale zamiast kreślenia wartości względem stałej, prostej osi, wykres strumienia ma wartości przesunięte wokół zmieniającej się centralnej linii bazowej. Wykresy strumieniowe przedstawiają zmiany danych w czasie dla różnych kategorii dzięki zastosowaniu płynnych, organicznych kształtów, które nieco przypominają strumień podobny do rzeki. To sprawia, że ​​wykresy strumieniowe są estetyczne i bardziej interesujące.

Na wykresie strumienia rozmiar każdego indywidualnego kształtu strumienia jest proporcjonalny do wartości w każdej kategorii. Oś, do której wykres strumienia płynie równolegle, jest używana w skali czasu. Kolor można wykorzystać do rozróżnienia każdej kategorii lub do wizualizacji dodatkowych wartości ilościowych każdej kategorii poprzez zmianę odcienia koloru.

Wykresy strumieniowe są idealne do wyświetlania dużych zbiorów danych, aby odkrywać trendy i wzorce w czasie w szerokim zakresie kategorii. Na przykład sezonowe szczyty i doliny w kształcie strumienia mogą sugerować wzór okresowy. Wykres strumienia można również wykorzystać do wizualizacji zmienności dużej grupy aktywów w określonym przedziale czasu.

Wadą wykresów strumieniowych jest to, że mają problemy z czytelnością, ponieważ często są bardzo zaśmiecone dużymi zestawami danych. Kategorie o mniejszych wartościach są często zagłuszane, aby zrobić miejsce dla kategorii o znacznie większych wartościach, uniemożliwiając zobaczenie wszystkich danych. Niemożliwe jest również odczytanie dokładnych wartości wizualizowanych na wykresie strumienia, ponieważ nie ma osi, której można by użyć jako odniesienia.

Dlatego wykresy strumieniowe powinny być zastrzeżone dla odbiorców, którzy nie zamierzali spędzać dużo czasu na rozszyfrowywaniu wykresu i eksplorowaniu jego danych. Wykresy strumieniowe lepiej nadają bardziej ogólny widok danych. Mają też znacznie lepszą funkcjonalność jako element interaktywny niż grafika statyczna lub drukowana.



**Funkcje**

Dane w czasie

Wzory

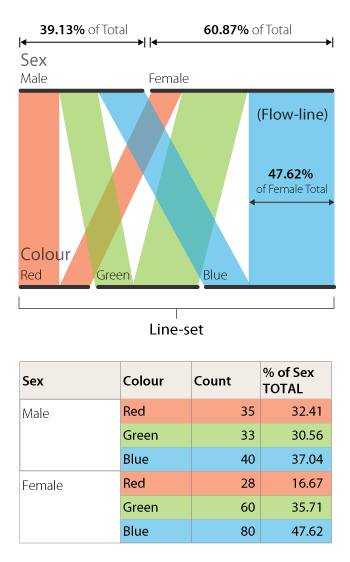
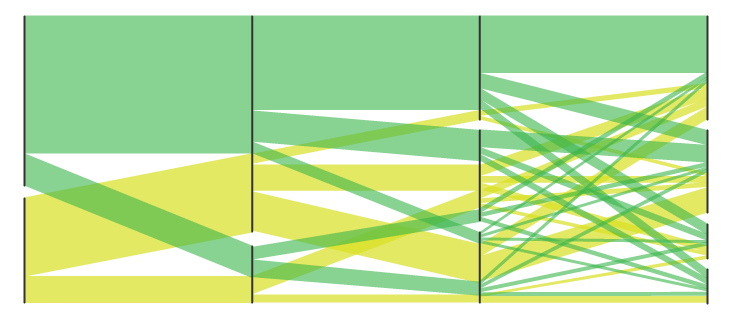
## Ruch i przepływy

### Wykresy zestawów równoległych

**Opis**

Wykresy zestawów równoległych są podobne do diagramów Sankeya, ponieważ pokazują przepływ i proporcje. Jednak zestawy równoległe nie używają strzałek i dzielą ścieżkę przepływu w każdym wyświetlonym zestawie linii.

Każdy zestaw linii odpowiada wymiarowi / zestawowi danych, których wartości / kategorie są reprezentowane w każdym podziale linii w tym zestawie linii. Szerokość każdej linii i wynikająca z niej ścieżka przepływu jest określona przez proporcjonalny ułamek sumy kategorii. Każda ścieżka przepływu może być pokolorowana, aby pokazać i porównać rozkład między różnymi kategoriami.



**Funkcje**

**Porównania**

**Dystrybucja**

**Przepływy**

**Procesy & Metody**

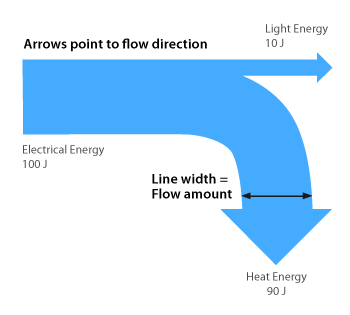
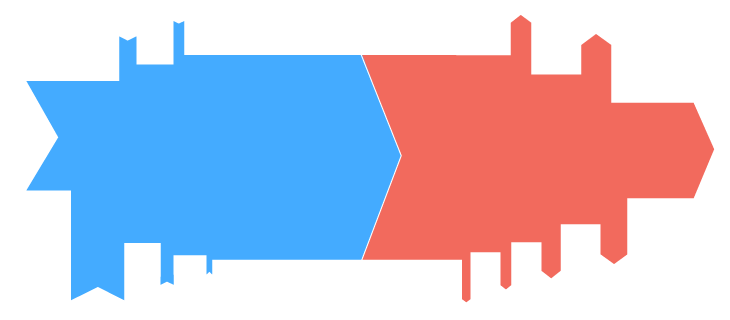
**Proportcje**

### Diagram Sankey’a

Opis

Diagramy Sankeya pokazują przepływy i ich ilości proporcjonalnie do siebie. Szerokość strzałek lub linii służy do pokazania ich wielkości, więc im większa strzałka, tym większa ilość przepływu. Strzałki lub linie przepływu mogą łączyć się ze sobą lub rozdzielać ścieżki na każdym etapie procesu. Koloru można użyć do podzielenia diagramu na różne kategorie lub do pokazania przejścia z jednego stanu procesu do drugiego.

Zazwyczaj diagramy Sankeya służą do wizualnego pokazania transferu energii, pieniędzy lub materiałów, ale można je wykorzystać do pokazania przepływu dowolnego izolowanego procesu systemowego.



**Funkcje**

Przepływy

Procesy

Proporcje

## Relacje (wykresy)

### Wykres rozrzutu

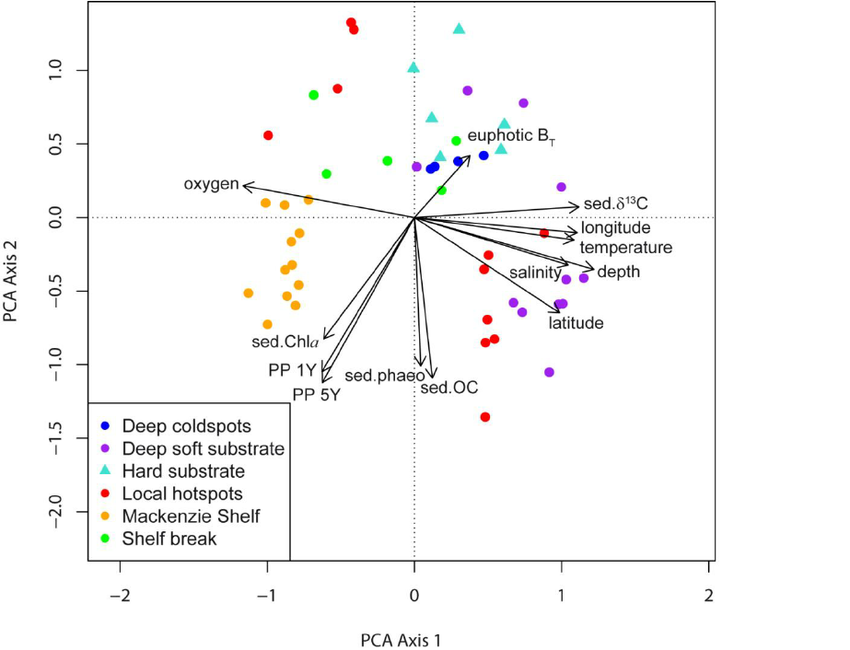
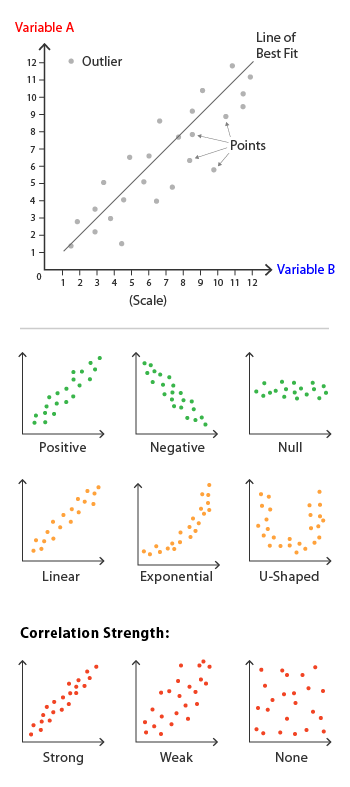
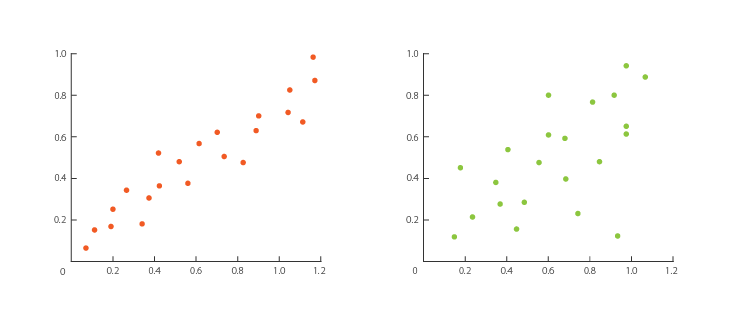
Opis: Znany również jako **wykres rozproszenia, wykres punktowy, wykres X-Y, wykres rozproszenia lub Scattergram.**

Wykresy rozrzutu używają zbioru punktów umieszczonych za pomocą współrzędnych kartezjańskich do wyświetlania wartości z dwóch zmiennych. Wyświetlając zmienną w każdej osi, możesz wykryć, czy istnieje relacja lub korelacja między dwiema zmiennymi.

Różne typy korelacji można interpretować za pomocą wzorów wyświetlanych na wykresach rozrzutu. Są to: **dodatnie** (wartości rosną razem), **ujemne** (jedna wartość maleje wraz ze wzrostem drugiej), **zero** (brak korelacji), **liniowe**, **wykładnicze** i w **kształcie litery U**. Siła korelacji może być określona przez to, jak ściśle upakowane są punkty na wykresie. Punkty, które kończą się daleko poza ogólnym skupieniem punktów, są znane jako **wartości odstające**.

Linie lub krzywe są dopasowywane do wykresu w celu ułatwienia analizy i są narysowane tak blisko wszystkich punktów, jak to możliwe i aby pokazać, jak wyglądałyby wszystkie punkty skondensowane w jedną linię. Zazwyczaj jest to znane jako Linia najlepszego dopasowania lub Linia trendu i może być użyte do oszacowania za pomocą interpolacji.

Wykresy rozrzutu są idealne, gdy sparujesz dane liczbowe i chcesz sprawdzić, czy jedna zmienna wpływa na drugą. Pamiętaj jednak, że korelacja nie jest przyczyną, a inna niezauważalna zmienna może wpływać na wyniki.



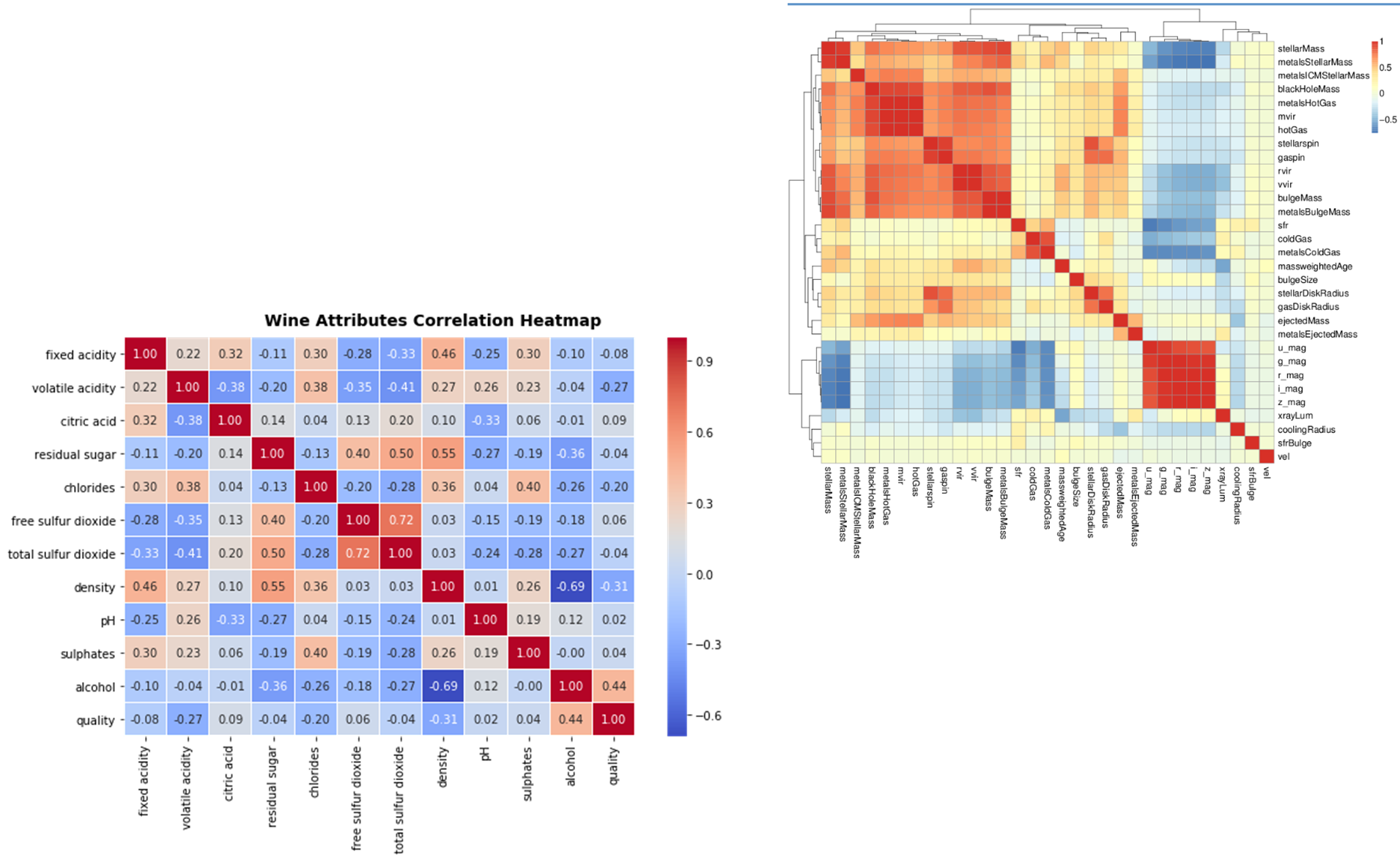
Wykres do Analizy Głównych Składowych

**Funkcje**

Wzory

Relacje

### Mapa termiczna - macierz korelacji

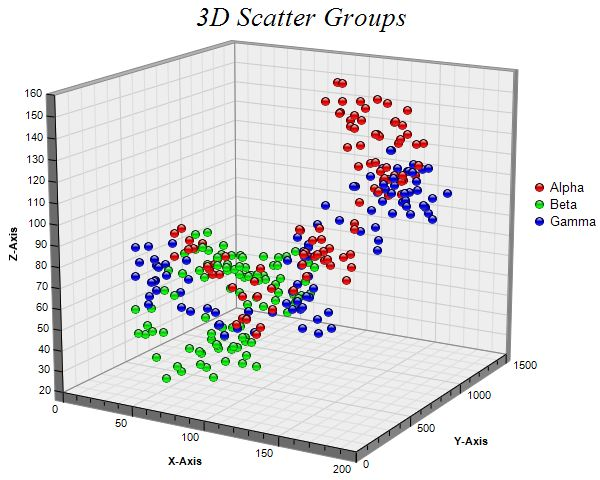
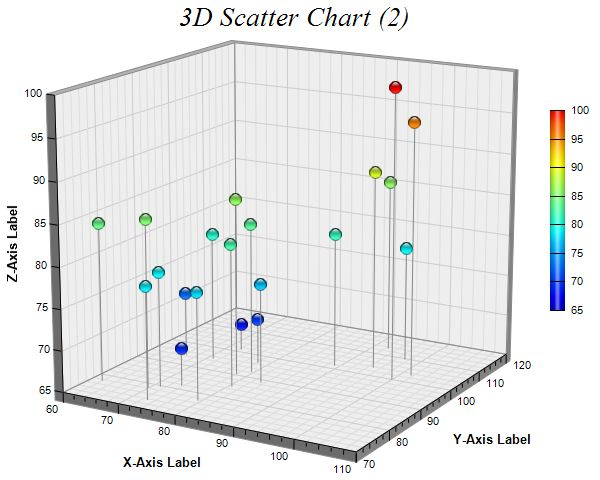
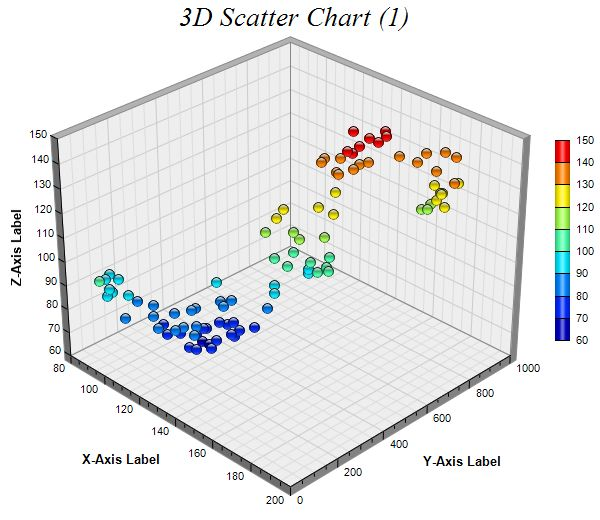


Wizualizacja korelacyjnej mapy cieplnej. Mapa termiczna korelacji wykorzystuje kolorowe komórki, zwykle w skali monochromatycznej, aby pokazać macierz korelacji 2D (tabela) między dwoma dyskretnymi wymiarami lub typami zdarzeń. ... Wartość koloru komórek jest proporcjonalna do współczynnika korelacji.

Wizualizacja mapy cieplnej macierzy korelacji (za pomocą miary korelacji Pearsona) niektórych właściwości galaktyki z próbnego katalogu galaktyk przez \ cite {Guo2011}. Kolor czerwony oznacza silną korelację dodatnią, a niebieski oznacza silną korelację ujemną. Żółci są powiązane z korelacjami bliskimi zeru.

Dendrogram właściwości galaktyki z katalogu \ citet {Guo2011}.

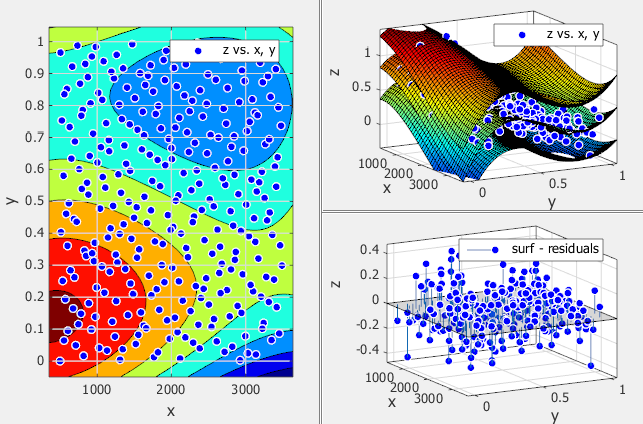
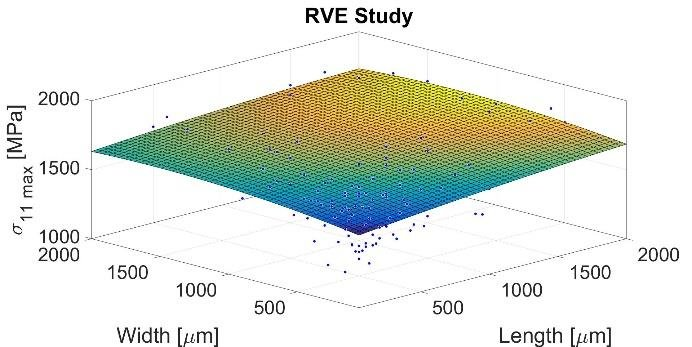
### 3D Wykresy rozrzutu



3D wykres rozrzutu z symbolami pokolorowanymi na podstawie ich wartości Z.

3D wykres rozrzutu z różnymi grupami symboli reprezentowanymi przy użyciu różnych kolorów.

### 3D Wykresy rozrzutu z regresją trójwymiarową



Redukcja wymiarów

### Wykres bąbelkowy

**Opis**

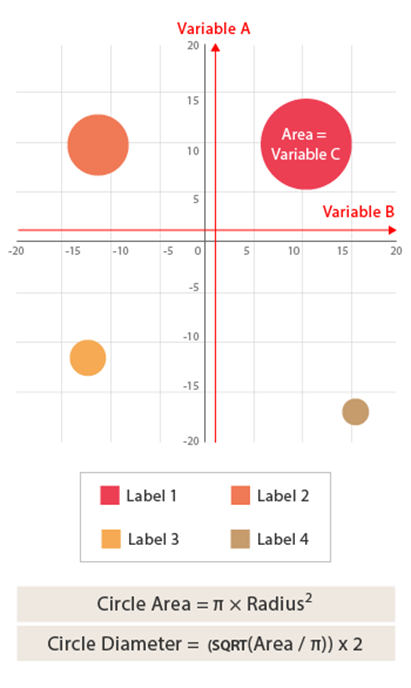
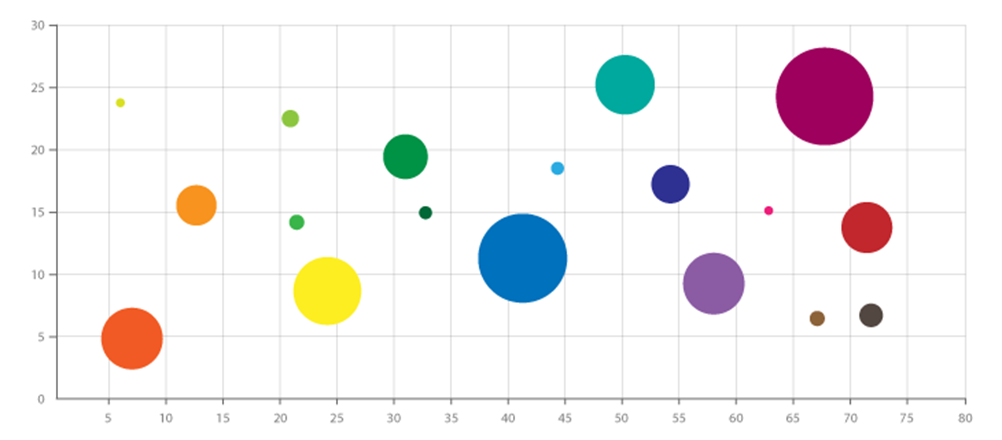
Wykres bąbelkowy to wykres o wielu zmiennych, który jest skrzyżowaniem wykresu rozrzutu z proporcjonalnym wykresem powierzchni.

Podobnie jak wykres rozrzutu, wykresy bąbelkowe używają kartezjańskiego układu współrzędnych do kreślenia punktów wzdłuż siatki, w której oś X i Y są osobnymi zmiennymi. Jednak. w przeciwieństwie do Planu rozrzutu, każdemu punktowi przypisana jest etykieta lub kategoria (wyświetlane obok lub na legendzie). **Każdy narysowany punkt reprezentuje następnie trzecią zmienną** przez obszar jego okręgu. Kolorów można również użyć do rozróżnienia kategorii lub użyć do przedstawienia dodatkowej zmiennej danych. Czas można pokazać, ustawiając go jako zmienną na jednej osi lub animując zmienne danych zmieniające się w czasie.

Wykresy bąbelkowe są zwykle używane do porównywania i pokazywania związków między skategoryzowanymi kręgami, przy użyciu pozycjonowania i proporcji. Ogólny obraz wykresów bąbelkowych można wykorzystać do analizy wzorców / korelacji.

Zbyt wiele bąbelków może utrudniać odczytanie wykresu, więc wykresy bąbelkowe mają ograniczoną pojemność. Można temu zaradzić poprzez interaktywność: kliknięcie lub najechanie myszką na bąbelki, aby wyświetlić ukryte informacje, z możliwością reorganizacji lub odfiltrowania zgrupowanych kategorii.

Podobnie jak w przypadku wykresów obszarów proporcjonalnych, rozmiary kół należy rysować na podstawie powierzchni koła, a nie jego promienia lub średnicy. Wielkość kręgów nie tylko zmieni się wykładniczo, ale doprowadzi to do błędnej interpretacji przez ludzki system wzrokowy.



***Funkcje***

**Porównania**

**Dane w czasie**

**Dystrybucja**

**Wzory**

**Proporcje**

**Relacje**

## Relacje (sieci)

### Diagram sieci (Graf)

**Opis**

Znany również jako **sieć grafu , mapa sieci, schemat połączenia z węzłami.**

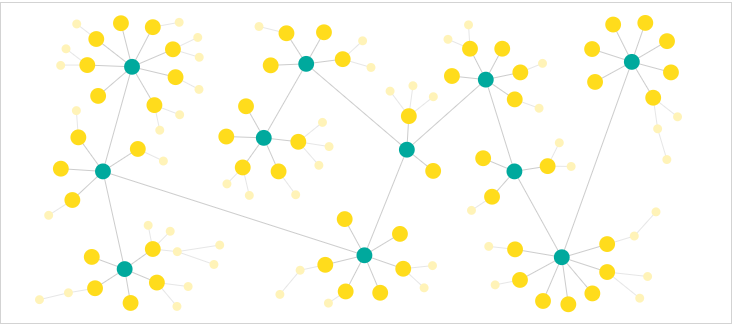
Ten rodzaj wizualizacji pokazuje, jak rzeczy są ze sobą połączone za pomocą węzłów / wierzchołków i linii łączących do reprezentowania ich połączeń i pomagają uwidocznić rodzaj relacji między grupą jednostek.

Zazwyczaj węzły są rysowane jako małe kropki lub kółka, ale można również użyć ikon. Łącza są zwykle wyświetlane jako proste linie połączone między węzłami. Jednak na niektórych diagramach sieciowych nie wszystkie węzły i łącza są tworzone w jednakowy sposób: dodatkowe zmienne mogą być wizualizowane, na przykład poprzez proporcję wielkości węzła lub wagi obrysu łącza do przypisanej wartości.

Dzięki odwzorowaniu połączonych systemów można wykorzystać diagramy sieciowe do interpretacji struktury sieci poprzez poszukiwanie klastrowania węzłów, gęstość połączeń węzłów lub układ aranżacji diagramów.

Dwa ważne typy diagramów sieciowych to „bezkierunkowy” i „ukierunkowany”. Niekierowane diagramy sieciowe wyświetlają tylko połączenia między jednostkami, natomiast ukierunkowane diagramy sieciowe pokazują, czy połączenia są jednokierunkowe czy dwukierunkowe za pomocą małych strzałek.

Diagramy sieciowe mają ograniczoną pojemność danych i zaczynają być trudne do odczytania, gdy jest zbyt wiele węzłów i przypominają „kule”.



krawędź

węzeł

**Funkcja**

Relacje

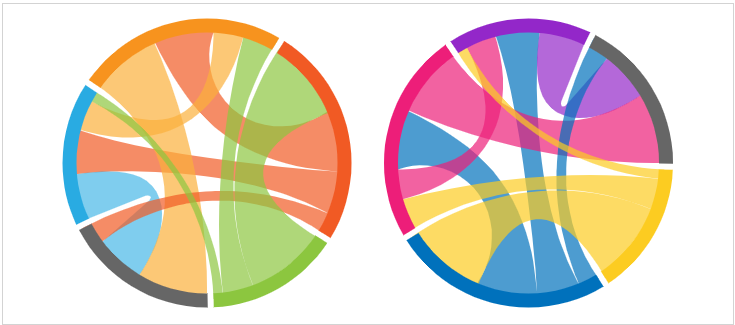
### Diagram cięciwy

Opis

Ten typ diagramu wizualizuje wzajemne relacje między jednostkami. Połączenia między jednostkami służą do wyświetlania, że mają coś wspólnego. To sprawia, że diagramy akordów są idealne do porównywania podobieństw w zbiorze danych lub między różnymi grupami danych.

Węzły są rozmieszczone wzdłuż okręgu, a relacje między punktami są ze sobą połączone za pomocą łuków lub krzywych Béziera. Wartości są przypisane do każdego połączenia, które jest reprezentowane proporcjonalnie przez rozmiar każdego łuku. Koloru można użyć do pogrupowania danych w różne kategorie, co pomaga w dokonywaniu porównań i rozróżnianiu grup.

Przepełnienie staje się problemem w diagramach akordów, gdy wyświetlanych jest zbyt wiele połączeń.



**Funkcje**

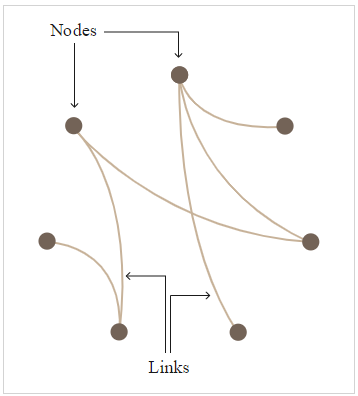
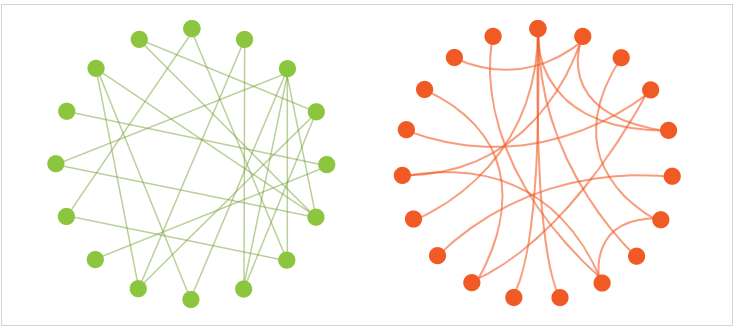
**Porównania**

**Relacje**

### Diagram cięciwy bez wstążki

**Opis**

Diagram cięciw nie wstążkowych to uproszczona wersja diagramu cięciw, w którym widoczne są tylko węzły i linie połączeń. Zapewnia to większy nacisk na połączenia w danych. Używany również do wizualizacji teorii grafów.



**Funkcja**

Relacje

### Diagram drzewa

Obraz zawierający tekst, linia, Czcionka, diagram

Opis wygenerowany automatycznie**Opis**

Znany również jako schemat organizacyjny, drzewo powiązań.

Diagram drzewa to sposób wizualnego przedstawienia hierarchii w strukturze drzewiastej. Zazwyczaj struktura diagramu drzewa składa się z takich elementów, jak **węzeł** **główny**, który nie ma elementu nadrzędnego. Następnie są węzły, które są połączone razem z połączeniami liniowymi zwanymi **gałęziami**, które reprezentują relacje i połączenia między węzłami. Wreszcie **węzły liści** (lub węzły końcowe) to węzły, które nie mają węzłów potomnych.

Diagramy drzew są często używane:

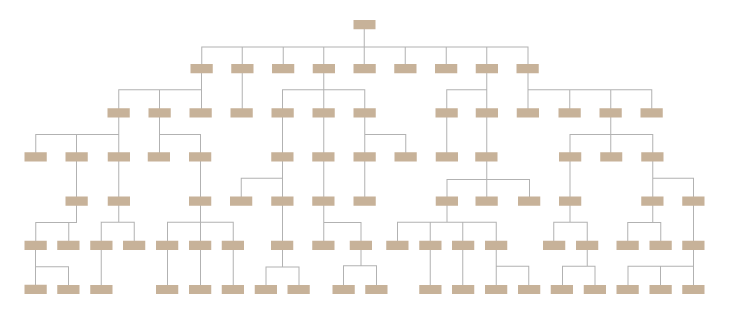
Aby pokazać relacje rodzinne i pochodzenie.

W taksonomii praktyka i nauka klasyfikacji.

W nauce ewolucyjnej, aby pokazać pochodzenie gatunków.

W informatyce i matematyce.

W przedsiębiorstwach i organizacjach do celów zarządczych.



**Funkcje**

Hierarchia

Relacje

## Hierarchia

### Mapy drzew

**Opis**

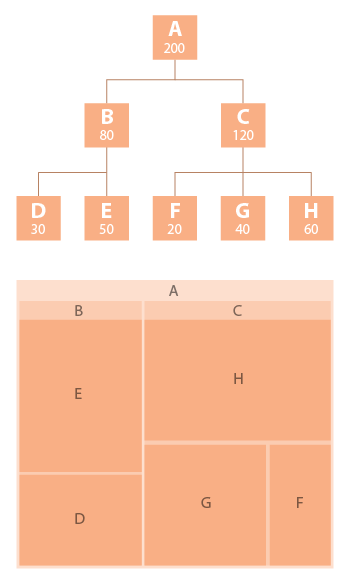
Mapy drzew są alternatywnym sposobem wizualizacji hierarchicznej struktury diagramu drzewa, a także wyświetlania ilości dla każdej kategorii według wielkości obszaru. Każda kategoria ma przypisany obszar prostokąta z zagnieżdżonymi w niej prostokątami podkategorii.

Gdy ilość jest przypisana do kategorii, jej powierzchnia jest wyświetlana proporcjonalnie do tej ilości i do innych ilości w tej samej kategorii nadrzędnej w relacji części do całości. Rozmiar obszaru kategorii nadrzędnej jest także sumą jego podkategorii. Jeśli żadna ilość nie jest przypisana do podkategorii, wówczas jej obszar jest dzielony równo między inne podkategorie w ramach kategorii nadrzędnej.

Sposób, w jaki prostokąty są podzielone i uporządkowane w pod-prostokąty, zależy od zastosowanego algorytmu kafelkowania. Opracowano wiele algorytmów ustawiania kafelków, ale „algorytm kwadratowy”, który utrzymuje każdy prostokąt tak kwadratowy, jak to możliwe, jest powszechnie stosowany.

Ben Shneiderman pierwotnie opracował mapy witryn jako sposób wizualizacji obszernego katalogu plików na komputerze, nie zajmując zbyt wiele miejsca na ekranie. To sprawia, że ​​mapy witryn są bardziej kompaktową i zajmującą mało miejsca opcją wyświetlania hierarchii, która zapewnia szybki przegląd struktury. Mapy drzew świetnie sprawdzają się również w porównywaniu proporcji między kategoriami według wielkości ich powierzchni.

Wadą tej mapy jest to, że nie pokazuje ona poziomów hierarchicznych tak wyraźnie, jak inne wykresy wizualizujące dane hierarchiczne (takie jak diagram drzewa lub diagram Sunburst).



**Funkcje**

**Porównania**

**Hierarchia**

**Część do całości**

**Proporcje**

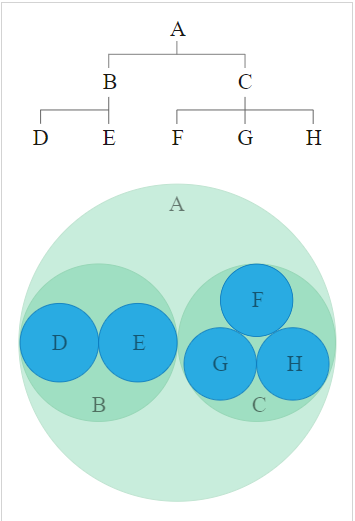
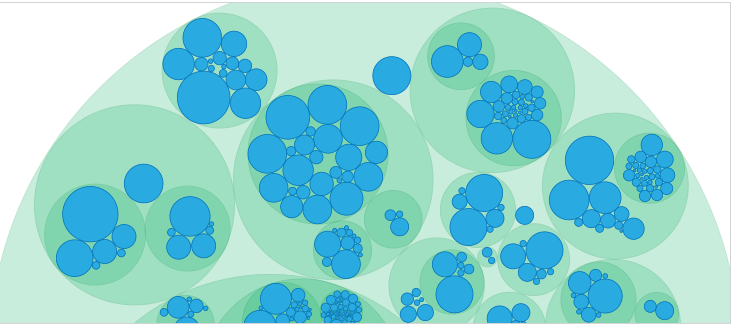
### Pakowanie w kółko

Opis

Znany jako okrągła mapa drzewa.

Pakowanie w kółko to odmiana mapy drzewa, która wykorzystuje koła zamiast prostokątów. Ograniczenie w obrębie każdego koła reprezentuje poziom w hierarchii: każda gałąź drzewa jest reprezentowana jako okrąg, a jego gałęzie są reprezentowane jako koła wewnątrz niego. Obszar każdego koła można również wykorzystać do przedstawienia dodatkowej arbitralnej wartości, takiej jak ilość lub rozmiar pliku. Kolor może być również użyty do przypisania kategorii lub przedstawienia innej zmiennej za pomocą różnych odcieni.

Tak piękne, jak się wydaje **Packing Circle**, nie jest ono tak oszczędne jak mapa Treemap, ponieważ w kręgach jest dużo pustej przestrzeni. Mimo to Pakowanie w kółko ujawnia strukturę hierarchiczną lepiej niż mapa drzewa.



**Funkcje**

Hierarchia

Proporcja

### Diagram „rozbłysk słońca” (Sunburst)

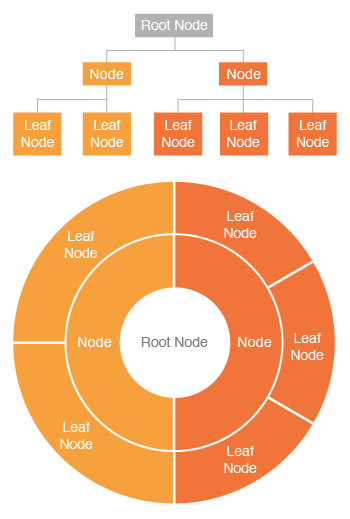
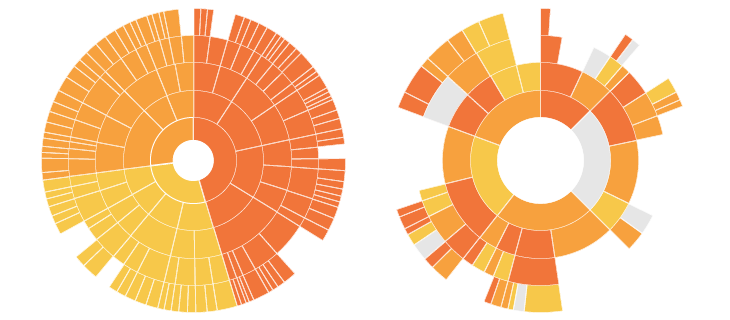
Opis

Wykres pierścieniowy, wielopoziomowy wykres kołowy, wykres pasowy, mapa promieniowa.

Ten typ wizualizacji pokazuje hierarchię poprzez serię pierścieni, które są pocięte na poszczególne węzły kategorii. Każdy pierścień odpowiada poziomowi w hierarchii, przy czym środkowy okrąg reprezentuje węzeł główny, a hierarchia odchodzi od niego na zewnątrz.

Pierścienie są krojone i dzielone na podstawie ich hierarchicznej relacji z wycinkiem nadrzędnym. Kąt każdego wycinka jest albo równo dzielony pod jego węzłem macierzystym, albo może być proporcjonalny do wartości.

Koloru można użyć do wyróżnienia grup hierarchicznych lub określonych kategorii.



**Funkcje**

Hierarchia

Część do całości

### Dendrogram

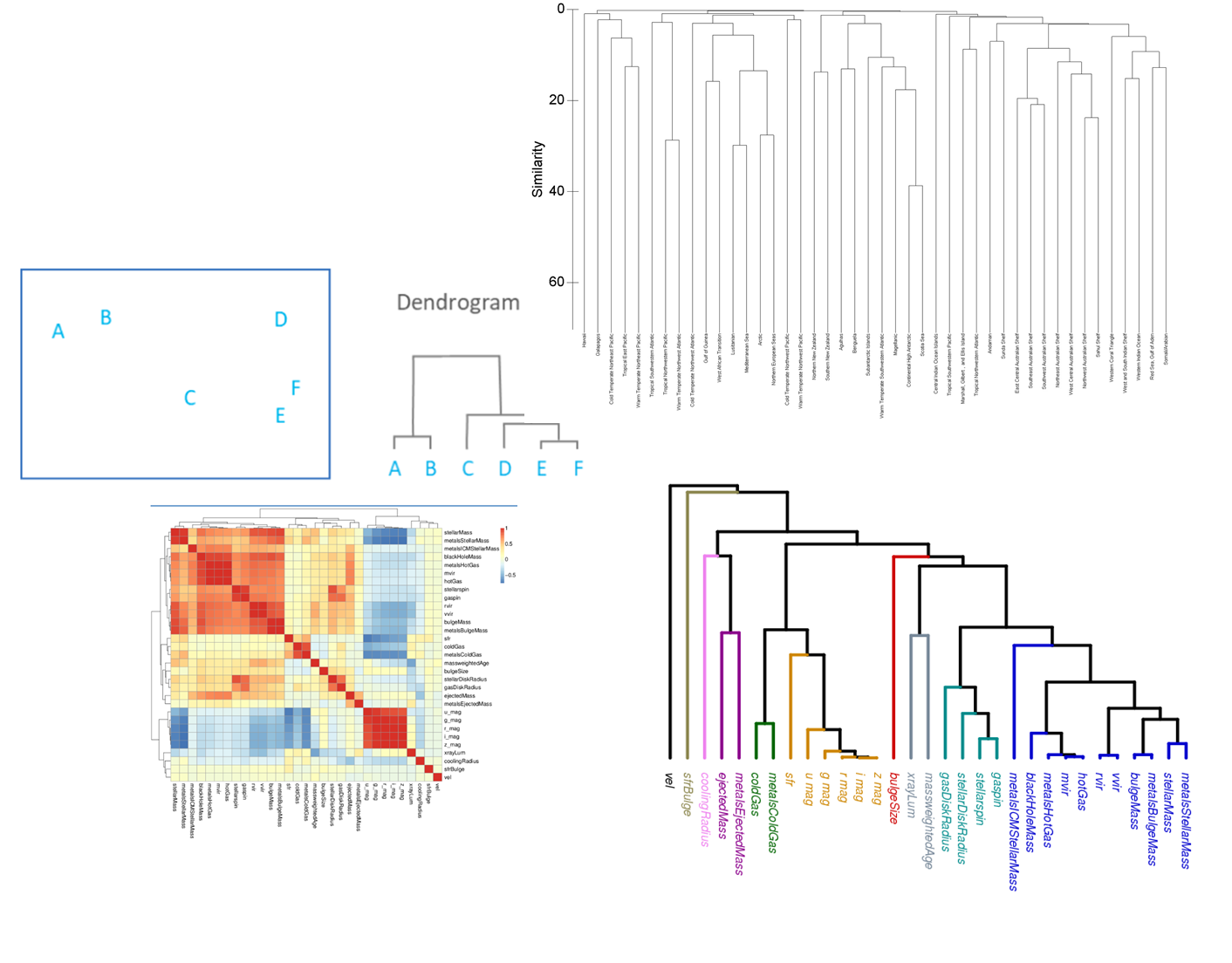
**Opis**

Dendrogram to diagram pokazujący hierarchiczną zależność między obiektami. Najczęściej jest tworzony jako wynik hierarchicznego grupowania. Głównym zastosowaniem dendrogramu jest wypracowanie najlepszego sposobu przydzielania obiektów do klastrów. Poniższy dendrogram pokazuje hierarchiczne skupienie sześciu obserwacji pokazanych na wykresie rozrzutu po lewej stronie.

**Jak czytać dendrogram**

Kluczem do interpretacji dendrogramu jest skupienie się na wysokości, na której dowolne dwa obiekty są ze sobą połączone. W powyższym przykładzie widzimy, że E i F są najbardziej podobne, ponieważ wysokość łączącego je łącza jest najmniejsza. Następne dwa najbardziej podobne obiekty to A i B.

W powyższym dendrogramie wysokość dendrogramu wskazuje kolejność łączenia klastrów. Można stworzyć bardziej pouczający dendrogram, w którym wysokości odzwierciedlają odległość między skupieniami, jak pokazano poniżej. W tym przypadku dendrogram pokazuje nam, że duża różnica między skupieniami występuje pomiędzy skupieniem A i B w porównaniu do skupienia C, D, E i F.



Dane wyjściowe programu Dendrogram dla hierarchicznego grupowania prowincji morskich z wykorzystaniem obecności / braku gatunków gąbek

Dendrogram w wyniku mapy cieplnej macierzy korelacji właściwości galaktyki z katalogu \ citet {Guo2011}. Różne podgrupy właściwości galaktyki, przypisane za pomocą indeksu \ citeauthor {Calinski74}, są pokolorowane zgodnie z przypisaniem gromady.