

Matematyczne aspekty analizy danych semestr zimowy 2025/2026

Dr Anna Muranova
UWM w Olsztynie

Wykład 1

Rozdział 0. Wprowadzenie

Analiza danych

Analiza danych – proces polegający na badaniu, oczyszczaniu, przekształcaniu i modelowaniu danych w celu odkrywania przydatnych informacji, wyciągania wniosków i wspomagania podejmowania decyzji.

W zależności od rodzaju danych i stawianych problemów, może to oznaczać użycie metod statystycznych, eksploracyjnych, wizualizacji danych i innych.

- ▶ Statystyka – nauka, której przedmiotem zainteresowania są metody pozyskiwania i prezentacji, a przede wszystkim analizy danych opisujących zjawiska, w tym masowe.
- ▶ Eksploracja danych – jeden z etapów procesu uzyskiwania wiedzy z baz danych. Idea eksploracji danych polega na wykorzystaniu szybkości komputera do znajdowania ukrytych dla człowieka (właśnie z uwagi na ograniczone możliwości czasowe) prawidłowości w danych zgromadzonych w hurtowniach danych. To jest proces używany do wydobywania cennych informacji z dużych zbiorów danych.
- ▶ Wizualizacja danych – proces graficznego przekształcania surowych informacji w czytelne formaty, takie jak wykresy, diagramy, mapy czy infografik.

Analiza danych zastanych

Dane zastane – to takie dane, które już powstały i analityk nie ma wpływu na ich formę

Korzystanie z danych zastanych jest przykładem badań niereaktywnych - metod badań zachowań społecznych, które nie wpływają na te zachowania. Dane takie to: dokumenty, archiwa, sprawozdania, kroniki, spisy ludności, księgi parafialne, dzienniki, pamiętniki, blogi internetowe, audio-pamiętniki, archiwa historii mówionej i inne. (Wikipedia)

Dane zastane możemy podzielić ze względu na (Makowska red. 2013):

- ▶ Charakter: Ilościowe, Jakościowe
- ▶ Formę: Dane opracowane, Dane surowe
- ▶ Sposób powstania: Pierwotne, Wtórne
- ▶ Dynamikę: Ciągła rejestracja zdarzeń, Rejestracja w interwałach czasowych, Rejestracja jednorazowa
- ▶ Poziom obiektywizmu: Obiektywne, Subiektywne
- ▶ Źródła pochodzenia: Dane publiczne, Dane prywatne

Analiza danych to proces polegający na sprawdzaniu, porządkowaniu, przekształcaniu i modelowaniu danych w celu zdobycia użytecznych informacji, wypracowania wniosków i wspierania procesu decyzyjnego. Analiza danych ma wiele aspektów i podejść, obejmujących różne techniki pod różnymi nazwami, w różnych obszarach biznesowych, naukowych i społecznych. Praktyczne podejście do definiowania danych polega na tym, że dane to liczby, znaki, obrazy lub inne metody zapisu, w formie, którą można ocenić w celu określenia lub podjęcia decyzji o konkretnym działaniu.

Wiele osób uważa, że dane same w sobie nie mają znaczenia – dopiero dane przetworzone i zinterpretowane stają się informacją.

Proces analizy danych

Analiza odnosi się do rozbicia całości posiadanych informacji na jej odrębne komponenty w celu indywidualnego badania. Analiza danych to proces uzyskiwania nieprzetworzonych danych i przekształcania ich w informacje przydatne do podejmowania decyzji przez użytkowników. Dane są zbierane i analizowane, aby odpowiadać na pytania, testować hipotezy lub obalać teorie. Istnieje kilka faz, które można wyszczególnić w procesie analizy danych. Fazy są iteracyjne, ponieważ informacje zwrotne z faz kolejnych mogą spowodować dodatkową pracę w fazach wcześniejszych.

- ▶ Zdefiniowanie wymagań
- ▶ Gromadzenie danych
- ▶ Przetwarzanie danych
- ▶ Właściwa analiza danych
- ▶ Raportowanie i dystrybucja wyników

Zdefiniowanie wymagań

Przed przystąpieniem do analizy danych, należy dokładnie określić wymagania jakościowe dotyczące danych. Dane wejściowe, które mają być przedmiotem analizy, są określone na podstawie wymagań osób kierujących analizą lub klientów (którzy będą używać finalnego produktu analizy). Ogólny typ jednostki, na podstawie której dane będą zbierane, jest określany jako jednostka eksperymentalna (np. osoba lub populacja ludzi). Dane mogą być liczbowe lub kategoriowe (tj. Etykiety tekstowe). Faza definiowania wymagań powinna dać odpowiedź na 2 zasadnicze pytania:

- ▶ co chcemy zmierzyć?
- ▶ w jaki sposób chcemy to zmierzyć?

Gromadzenie danych

Dane są gromadzone z różnych źródeł. Wymogi, co do rodzaju i jakości danych mogą być przekazywane przez analityków do “opiekunów danych”, takich jak personel technologii informacyjnych w organizacji. Dane ponadto mogą być również gromadzone automatycznie z różnego rodzaju czujników znajdujących się w otoczeniu - takich jak kamery drogowe, satelity, urządzenia rejestrujące obraz, dźwięk oraz parametry fizyczne. Kolejną metodą jest również pozyskiwanie danych w drodze wywiadów, gromadzenie ze źródeł internetowych lub bezpośrednio z dokumentacji.

Przetwarzanie danych

Zgromadzone dane muszą zostać przetworzone lub zorganizowane w sposób logiczny do analizy.

- ▶ Dane muszą być przetworzone i uporządkowane. Na przykład, mogą one zostać umieszczone w tabelach w celu dalszej analizy - w arkuszu kalkulacyjnym lub innym oprogramowaniu.
- ▶ Oczyszczanie danych. Po fazie przetworzenia i uporządkowania, dane mogą być niekompletne, zawierać duplikaty lub zawierać błędy. Konieczność czyszczenia danych wynika z problemów związanych z wprowadzaniem i przechowywaniem danych. Czyszczenie danych to proces zapobiegania powstawaniu i korygowania wykrytych błędów. Typowe zadania obejmują dopasowywanie rekordów, identyfikowanie nieścisłości, ogólny przegląd jakości istniejących danych, usuwanie duplikatów i segmentację kolumn. Niezwykle istotne jest też zwracanie uwagi na dane których wartości są powyżej lub poniżej ustalonych wcześniej progów (ekstrema).

Właściwa analiza danych

Istnieje kilka metod, które można wykorzystać do tego celu, np.:

- ▶ Statystyki opisowe to kolejna metoda analizy zebranych informacji. Dane są badane, aby znaleźć najważniejsze ich cechy. W statystykach opisowych analitycy używają kilku podstawowych narzędzi - można użyć średniej lub średniej z zestawu liczb. Pomaga to określić ogólny trend aczkolwiek nie zapewnia to dużej dokładności przy ocenie ogólnego obrazu zebranych danych.
- ▶ Eksploracja danych (data mining). Techniki i metody służące eksploracji danych wywodzą się głównie z dziedziny badań nad sztuczną inteligencją. Główne przykłady stosowanych rozwiązań należą do następujących obszarów:
 - ▶ sieci neuronowe,
 - ▶ metody uczenia maszynowego,
 - ▶ metody ewolucyjne,
 - ▶ logika rozmyta,
 - ▶ zbiory przybliżone.
- ▶ Wizualizacja danych

W tej fazie ma miejsce również modelowanie i tworzenie formuł matematycznych - stosowane są w celu identyfikacji zależności między zmiennymi, takich jak korelacja lub przyczynowość.

Różnica pomiędzy eksploracją danych a statystyką, cz. 1

Eksploracja danych	Statystyka
Eksploracja danych to proces wydobywania przydatnych informacji, wzorców i trendów z ogromnych zbiorów danych i wykorzystywania ich do podejmowania decyzji w oparciu o dane.	Statystyka odnosi się do analizy i prezentacji danych liczbowych i stanowi główną część wszystkich algorytmów eksploracji danych.
Rodzaje eksploracji danych to grupowanie, klasyfikacja, asocjacja, sieć neuronowa, analiza oparta na sekwencjach, wizualizacja itp.	Rodzaje statystyki to statystyka opisowa i statystyka wnioskowania.
Nadaje się do dużych zbiorów danych.	Nadaje się do mniejszych zestawów danych.

Różnica pomiędzy eksploracją danych a statystyką, cz. 2

Eksploracja danych	Statystyka
Eksploracja danych jest procesem indukcyjnym. Oznacza to generowanie nowej teorii na podstawie danych.	Statystyka jest procesem dedukcyjnym. Nie pozwala sobie na jakiegokolwiek przewidywania.
Czyszczenie danych jest częścią eksploracji danych.	W statystyce czyste dane służą do wdrożenia metody statystycznej.
Walidacja modelu wymaga mniej interakcji z użytkownikiem, dzięki czemu można go łatwo zautomatyzować.	Walidacja modelu wymaga interakcji użytkownika, dlatego jest to złożone i automatyczne.
Zastosowania do eksploracji danych obejmują analizę danych finansowych, przemysł detaliczny, przemysł telekomunikacyjny, analizę danych biologicznych, niektóre zastosowania naukowe itp.	Zastosowanie statystyki obejmuje biostatystykę, kontrolę jakości, demografię, badania operacyjne itp.

Raportowanie i dystrybucja wyników

Ta faza polega na ustalaniu w jakiej formie przekazywać wyniki. Analityk może rozważyć różne techniki wizualizacji danych, aby w sposób wyraźnym i skuteczny przekazać wnioski z analizy odbiorcom. Wizualizacja danych wykorzystuje formy graficzne jak wykresy i tabele. Tabele są przydatne dla użytkownika, który może wyszukiwać konkretne rekordy, podczas gdy wykresy (np. wykresy słupkowe lub liniowe) dają spojrzenie ilościowych na zbiór analizowanych danych.

Skąd brać dane?

Darmowa repozytoria danych:

- ▶ Bank danych lokalnych GUS:
<https://bdl.stat.gov.pl/BDL/start>
- ▶ Otwarte dane:
<https://dane.gov.pl/pl>
- ▶ Bank Światowy:
<https://data.worldbank.org/>

Koncepcja “Tidy data”

Koncepcja czyszczenia danych (ang. tidy data):

- WICKHAM, Hadley . Tidy Data. Journal of Statistical Software, [S.l.], v. 59, Issue 10, p. 1 - 23, sep. 2014. ISSN 1548-7660. Available at: <https://www.jstatsoft.org/v059/i10>. Date accessed: 12 sep. 2024. doi:<http://dx.doi.org/10.18637/jss.v059.i10>.

Koncepcja “Tidy data”

Idealne dane są zaprezentowane w tabeli:

Imię	Wiek	Wzrost	Kolor oczu
Adam	26	167	Brązowe
Sylwia	34	164	Piwnie
Tomasz	42	183	Niebieskie

Na co powinniśmy zwrócić uwagę?

- ▶ jedna obserwacja (jednostka statystyczna) = jeden wiersz w tabeli/macierzy/ramce danych
- ▶ wartości danej cechy znajdują się w kolumnach
- ▶ jeden typ/rodzaj obserwacji w jednej tabeli/macierzy/ramce danych

Przykład nieuporządkowanych danych

Imię	Wiek	Wzrost	Brązowe	Niebieskie	Piwne
Adam	26	167	1	0	0
Sylwia	34	164	0	0	1
Tomasz	42	183	0	1	0

Nagłówki kolumn muszą odpowiadać cechom, a nie wartościom zmiennych.

Zbiór danych

Zawsze badamy zbiór danych.

Zbiór danych – kolekcja danych statystycznych zwykle ujętych w formie tablicy. Najczęściej kolumny odpowiadają obserwowanym, a każdy wiersz opisuje jedną obserwację z próby. Wartości komórek macierzy natomiast opisują realizacje danych zmiennych w kolejnych obserwacjach.

Zbiorowość/populacja

Zbiorowość statystyczna, populacja statystyczna: zbiór obiektów podlegających badaniu statystycznemu (wszystkie obiekty w zbiorze danych). Tworzą je jednostki podobne do siebie, logicznie powiązane, lecz nie identyczne. Mają pewne cechy wspólne oraz pewne właściwości pozwalające je różnicować.

- ▶ przykłady:
 - ▶ badanie wzrostu Polaków – mieszkańcy Polski
 - ▶ poziom nauczania w szkołach woj. warmińsko-mazurskiego – szkoły woj. warmińsko-mazurskiego.
- ▶ podział:
 - ▶ zbiorowość/populacja generalna – obejmuje całość,
 - ▶ zbiorowość/populacja próbna (próba) – obejmuje część populacji.

Jednostka statyczna

Jednostka statystyczna: każdy z elementów zbiorowości statystycznej.

▶ przykłady:

- ▶ studenci UWM – student UWM
- ▶ mieszkańcy Polski – każda osoba mieszkająca w Polsce
- ▶ maszyny produkowane w fabryce – każda maszyna

Cechy statystyczne

Cechy statystyczne – właściwości charakteryzujące jednostki statystyczne w danej zbiorowości statystycznej.

Dzielimy je na stałe i zmienne.

- ▶ Cechy stałe – takie właściwości, które są wspólne wszystkim jednostkom danej zbiorowości statystycznej.
podział:
 - ▶ rzeczowe – kto lub co jest przedmiotem badania statystycznego,
 - ▶ czasowe – kiedy zostało przeprowadzone badanie lub jakiego okresu czasu dotyczy badanie,
 - ▶ przestrzenne – jakiego terytorium (miejsce lub obszar) dotyczy badanie.

przykład: studenci WMil UWM w Olsztynie w roku akad. 2017/2018:

- ▶ cecha rzeczowa: posiadanie legitymacji studenckiej,
 - ▶ cecha czasowa: studenci studiujący w roku akad. 2017/2018
 - ▶ cecha przestrzenna: miejsce – WMil UWM w Olsztynie.
- ▶ Cechy zmienne – właściwości różnicujące jednostki statystyczne w danej zbiorowości.

przykład: studenci UWM – cechy zmienne: wiek, płeć, rodzaj ukończonej szkoły średniej, kolor oczu, wzrost.

Ważne:

- ▶ obserwacji podlegają tylko cechy zmienne,
- ▶ cecha stała w jednej zbiorowości może być cechą zmienną w innej zbiorowości.

Przykład: studenci UWM mają legitymację wydaną przez UWM. Studenci wszystkich uczelni w Polsce mają legitymacje wydane przez różne szkoły.

Klasyfikacja cech

Dane dotyczą próby z pewnej populacji, badanych pod kątem cechy (lub cech).
W polskiej systematyce podręcznikowej dzielimy cechy zmienne na:

- ▶ jakościowe (niemierzalne) – np. kolor oczu, płeć, grupa krwi.
- ▶ ilościowe (mieralne) – np. wzrost, masa, wiek:
 - ▶ ciągłe – np. wzrost, masa, wiek (w rozumieniu liczby dni między datą urodzin a datą badania),
 - ▶ porządkowe (quasi-ilościowe) – np. klasyfikacja wzrostu: (niski, średni, wysoki),
 - ▶ skokowe (dyskretne) – np. liczba posiadanych dzieci, liczba gospodarstw domowych, wiek (w rozumieniu ilości skończonych lat).

Skale

Z praktycznego punktu widzenia istotniejszy jest podział zaproponowany przez Stevensa w 1946 roku, który zaproponował czterostopniową klasyfikację pozwalającą określić zbiór dopuszczalnych operacji w tym przekształceń statystycznych. Ten system pozwala w pewien sposób usystematyzować wyniki pomiarów statystycznych.

Podstawowy podział wyróżnia 4 cechy:

- ▶ jakościowe (ang. qualitative) – np. kolor oczu, płeć:
 - ▶ nominalne (ang. nominal)
 - ▶ porządkowe (ang. ordinal)
- ▶ ilościowe (ang. quantitative) – np. wzrost, masa, wiek:
 - ▶ przedziałowe/interwałowe (ang. interval)
 - ▶ ilorazowa/stosunkowa (ang. ratio).

Podział cech statystycznych

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
NOMINAL	Determination of equality	<i>Permutation group</i> $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
ORDINAL	Determination of greater or less	<i>Isotonic group</i> $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
INTERVAL	Determination of equality of intervals or differences	<i>General linear group</i> $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
RATIO	Determination of equality of ratios	<i>Similarity group</i> $x' = ax$	Coefficient of variation

Skala nominalna

- ▶ skala, w której klasyfikujemy jednostkę statystyczną do określonej kategorii,
- ▶ wartość w tej skali nie ma żadnego uporządkowania,
- ▶ możemy tylko powiedzieć czy wartości są takie same czy różnią się,
- ▶ np. kolor oczu, płeć, grupa krwi.

Uwaga! Wartości liczbowe są tylko kodami, nprz:

Religia	Kod
Chrześcijaństwo	1
Islam	2
Buddyzm	3

Skala porządkowa

- ▶ wartości mają jasno określony porządek, ale nie są dane odległości między nimi,
- ▶ pozwala na uszeregowanie elementów.
- ▶ możemy powiedzieć czy wartości są takie same czy różnią się, i która jest większa (lepiej i t.d.)
- ▶ np. wzrost (niski, średni, wysoki), wiek (dziecko, młodość, dorośli, emeryt)

Uwaga! Kody muszą zachowywać porządek, np.:

Wykształcenie	Kod
Podstawowe	1
Średnie	2
Wyższe	3

Dochód	Kod
Niski	1
Średni	2
Wysoki	3

Skala przedziałowa/interwałowa

- ▶ wartości cechy wyrażone są poprzez konkretne wartości liczbowe,
- ▶ pozwala na porównywanie jednostek (coś jest większe lub mniejsze i na ile),
- ▶ nie możliwe jest badanie ilorazów (określenie ile razy dana wartość jest większa lub mniejsza od drugiej).
- ▶ np. temperatura w stopniach Celsjusza, data kalendarzowa.

Miasto	Temperatura w stopniach Celsjusza
Warszawa	15
Olsztyn	10
Gdańsk	5
Szczecin	20

Skala ilorazowa/stosunkowa

- ▶ wartości wyrażone są przez wartości liczbowe,
- ▶ możliwe określenie jest relacji mniejsza lub większa między wartościami,
- ▶ możliwe jest określenie stosunku (ilorazu) między wartościami,
- ▶ występuje zero absolutne.
- ▶ np. długość, masa, czas

Produkt	Cena w zł
Chleb	3
Masło	8
Gruszki	5

Podział cech statystycznych z dopuszczalnymi operacjami.

Skala	Własności miary	Operacje matematyczne	Operacje zaawansowane	Przykłady
nominalne (ang. nominal)	Klasyfikacja, członkostwo	$=, \neq$	Grupowanie	kolor oczu, płeć
porządkowe (ang. ordinal)	Porównanie, poziom	$=, \neq, >, <$	Sortowanie	wzrost, wiek
przedziałowe/interwałowe (ang. interval)	Różnica, bliskość	$=, \neq, >, <, +, -$	Porównanie ze standardem	data, temp. w $^{\circ}C$
proporcjonalne (ang. ratio)	Wielkość, ilość	$=, \neq, >, <, +, -, *, /$	Iloraz	czas, masa

Cechy binarne

- ▶ Uporządkowane cechy binarne (cechy **dwuwartościowe z wyróżnionym zerem**, naprz. {obecny, nieobecny}), tylko pozornie są cechami porządkowymi, lecz w praktyce (ze względu na zbiór dopuszczalnych przekształceń) spełniają wszystkie założenia cech proporcjonalnych.
- ▶ Nieuporządkowane cechy binarne (cechy **dwuwartościowe bez wyróżnionego zera**, naprz. płeć), tylko pozornie są cechami nominalnymi, lecz w praktyce (ze względu na zbiór dopuszczalnych przekształceń) spełniają wszystkie założenia cech interwałowych.

Ćwiczenie

Klasyfikować następujące cechy statystyczne: części mowy, energia, gęstość, gatunek, nazwiska, numer zawodnika, obecność, ocena, położenie we współrzędnych kartezjańskich, rezystancja, stadium choroby, stan cywilny, wykształcenie, inflacja, bezrobocie, temperatura w Kelvinach, temperatura w stopniach Fahrenheita.

Rozdział 1. Działania na liczbach

Liczby

- ▶ *Liczby naturalne* $\mathbb{N}$: $0, 1, 2, 3, 4, \dots$
- ▶ *Liczby całkowite* $\mathbb{Z}$: $0, 1, 2, 3, 4, \dots, -1, -2, -3, \dots$
- ▶ *Liczby wymierne* $\mathbb{Q}$: każda liczba, która można wyrazić za pomocą ułamka zwykłego
- ▶ *Liczby niewymierne* $\mathbb{R} \setminus \mathbb{Q}$: liczby na prostej liczbowej, których nie można wyrazić za pomocą ułamka zwykłego ($\pi, \sqrt{2}$).
- ▶ *Liczby rzeczywiste* $\mathbb{R}$: wymierne i niewymierne
- ▶ *Liczby zespolone* $\mathbb{C}$: $a + bi$, $a, b \in \mathbb{R}$, $i = \sqrt{-1}$.

Większość pracy będziemy wykonywać przy użyciu $\mathbb{N}$, $\mathbb{Z}$ oraz $\mathbb{R}$. Liczby urojone mogą pojawić się później w bardziej zaawansowanych zadaniach.

Uwaga: kody – liczby naturalne dodatnie.

Kolejność działań

(w matematyce i w Pythonie)

- ▶ składniki w nawiasach
- ▶ potęgi, pierwiastki
- ▶ mnożenie, dzielenie
- ▶ dodawanie, odejmowanie

$$2 \cdot (3 + 2)^2 : 5 - 4$$

$$= 6$$

$$42 : 7 * 3$$

$$= 18$$

Kolejność działań

(w matematyce i w Pythonie)

- ▶ składniki w nawiasach
- ▶ potęgi, pierwiastki
- ▶ mnożenie, dzielenie
- ▶ dodawanie, odejmowanie

$$2 \cdot (3 + 2)^2 : 5 - 4$$

$$= 6$$

$$42 : 7 * 3$$

$$= 18$$

Kolejność działań

(w matematyce i w Pythonie)

- ▶ składniki w nawiasach
- ▶ potęgi, pierwiastki
- ▶ mnożenie, dzielenie
- ▶ dodawanie, odejmowanie

$$2 \cdot (3 + 2)^2 : 5 - 4$$

$$= 6$$

$$42 : 7 * 3$$

$$= 18$$

Działanie w Pythonie

Podstawowe operacje arytmetyczne:

+	dodawanie
-	odejmowanie
*	mnożenie
/	dzielenie
**	potęgowanie

```
print(2*(3+2)**2/5-4)
print(42/7*3)
```

Uwaga z ułamkami:

$$\frac{3+1}{2}$$

```
print(3+1/2)
print((3+1)/2)
```

W dużych wyrażeniach używaj nawiasów dla "czytelności".

Uwaga:

```
print(3*0.1-0.3)
print(round(3*0.1-0.3))
```

Zmienne

Przykład: mnożenie zmiennych w Pythonie

```
x = int(input("Wprowadź liczbę : "))  
iloczyn = 3* x  
print(iloczyn)
```

Nazwy greckich zmiennych w Pythonie

```
beta = 1.75  
theta = 30.0
```

Sumowanie

Sumowanie, oznaczane grecką literą sigma σ , oznacza dodawanie wielu elementów.

$$\sum_{i=1}^5 2i = 2(1) + 2(2) + 2(3) + 2(4) + 2(5) = 30.$$

W Pythonie robić się zwykle przy pomocy pętli

```
suma = sum(2*i for i in range(1,6))  
print(suma)
```

Uwaga: funkcja range jest prawostronnie otwarta.

Sumowanie

Często używa się też zmiennej n na oznaczenie liczby elementów w kolekcji, na przykład rekordów w zbiorze danych:

$$\sum_{i=1}^n 10x_i.$$

W Pythonie robić się zwykle przy pomocy pętli

```
x = [1, 4, 6, 2]
n = len(x)
```

```
suma = sum (10*x[i] for i in range(0,n))
print(suma)
```

Lub przy pomocy List Comprehensions

```
x = [1, 4, 6, 2]
```

```
suma = sum (10*xi for xi in x)
print(suma)
```