Анализ на данни

Методология

Съдържание

- Приложения

Анализ на данни

- Организация и обработка на данни, с цел извличане на полезна информация
- Данните се различават по форма: измервания, наблюдения, анкети и др.
- При организацята на изходните данни се наблюдават зависимости, които се подлагат на анализ с математически и други средства
- Методите за анализ могат да включват текстови и графични описания на данните и резултатите
- Качеството на резултатите зависи от качеството на данните: източници, регистриране, организиране и др.

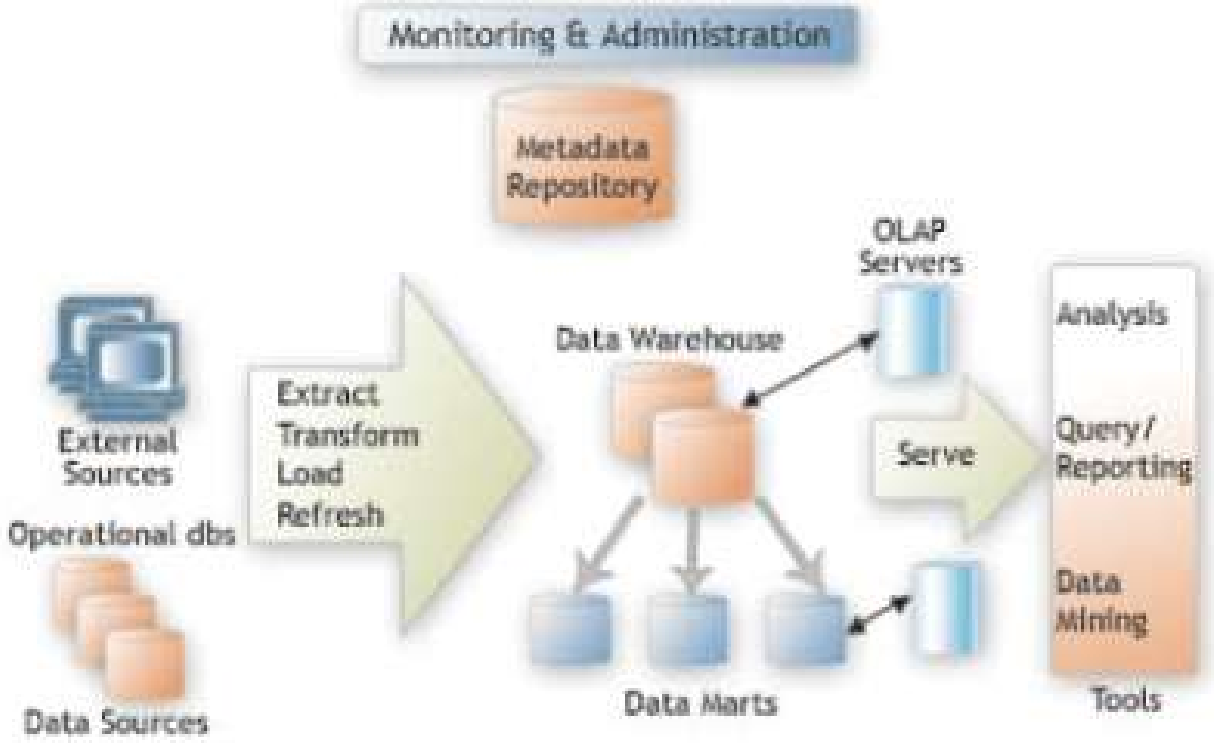
Бизнес интелигентност и анализ

- Съвокупност от програми и технологии за събиране съхраняване, анализиране и разпространение на информация, подпомагаща бизнес потребителите при вземане на по-добри решения.
- Видове приложения
  - статистически анализ
  - генериране на отчети
  - OLAP
  - системи за вземане на решение
  - прогнозиране
  - откриване на информация и знания
- Задачи на приложенията
  - аналитични – обощена информация за бизнеса (top-down)
  - оперативни – детайлна информация относно бизнес транзакциите (down-up)

Етапи на анализа

- Определяне на въпрос, който би получил отговор посредством на анализа
- Организиране на данните и определяне на обекти за изследване
- Дефиниране на данните
  - сумиране, описание и изследване на данните от извадката
  - изследване на взаимоотношенията и природата на данните в генералната съвокупност, от която произхождат
- Генериране на отчет – визуализиране и разпространение на резултатите от анализа

Технологии за анализ



Представяне на данните

- Множество – извадка репрезентативни данни, с които се провежда анализът
  - напр. таблица за продажби на стоки
- Обекти данни – екземпляри, елементи на изследваното множество
  - напр. ред от таблицата за една стока
- Атрибути – свойства на обектите на данните
  - напр. категория на стоката

Примери

Tid	Refund	Marital Status	Taxable Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

(a) Record data.

TID	ITEMS
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

(b) Transaction data.

Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

(c) Data matrix.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

(d) Document-term matrix.

Типове атрибути

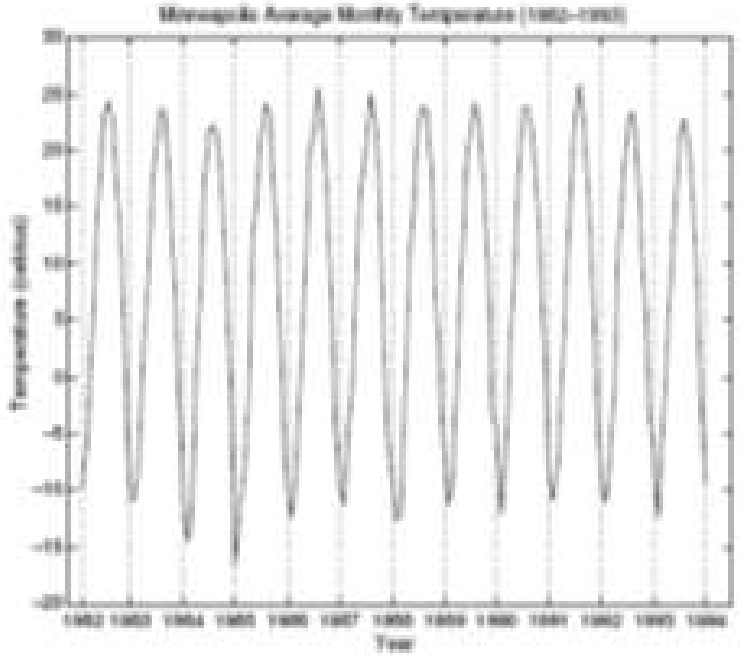
- Множество на допустимите стойности на атрибута
- Типове атрибути, според измервателната скала
  - качествени (неметрични), при които стойността се изразява без мерна единица.
    - номинални (категорийни) - код
    - рангови – система от кодове
  - количествени (метрични), при които стойността се изразява с определена мерна единица
    - интервални – метрична оценка за количество с условна нула
    - пропорционални – метрична оценка за количество с абсолютна нула
  - Типове атрибути, според измерваните величини
    - дискретни
    - непрекъснати

Дискретни и непрекъснати величини

- Дискретни величини
- Непрекъснати величини

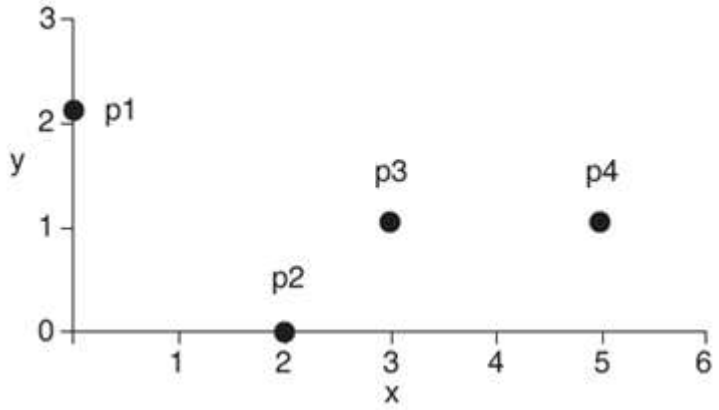
Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)



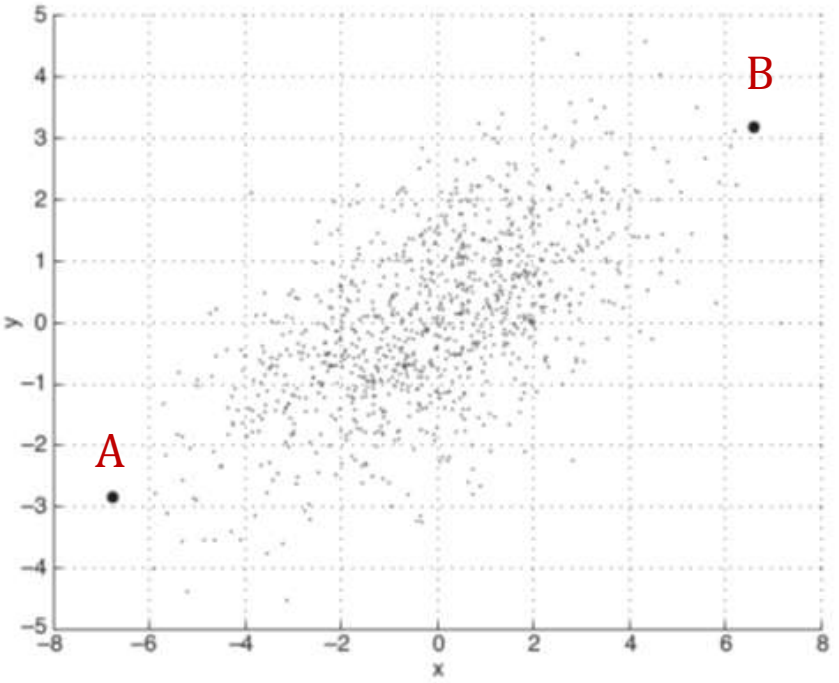
Визуализация на стойности

- Представяне на четири стойности в двумерно пространство



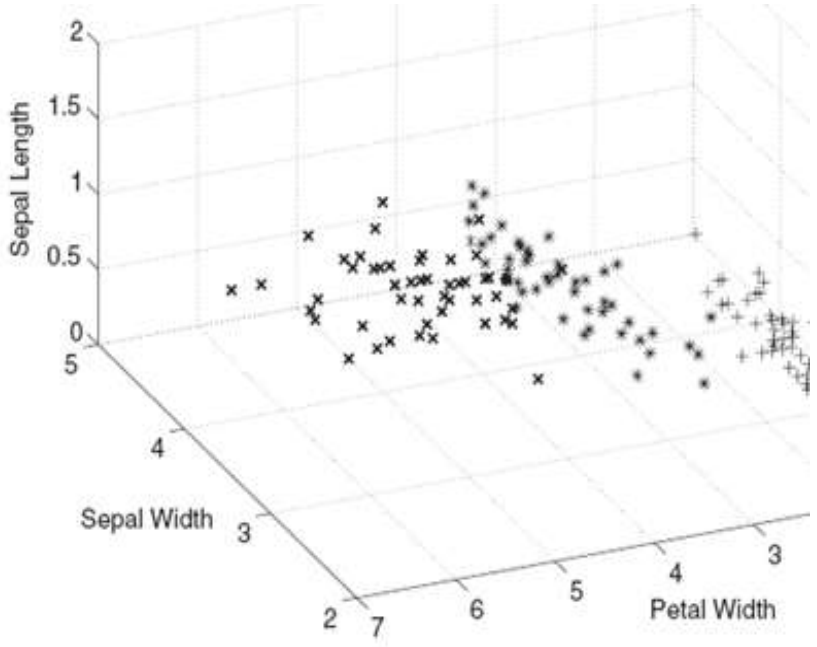
Визуализация на стойности

- Представяне на множество стойности в двумерно пространство
- Разстоянието от т. A до т. B е 14.7



Визуализация на стойности

- Представяне на множество стойности в тримерно пространство



Статистически анализ на данните

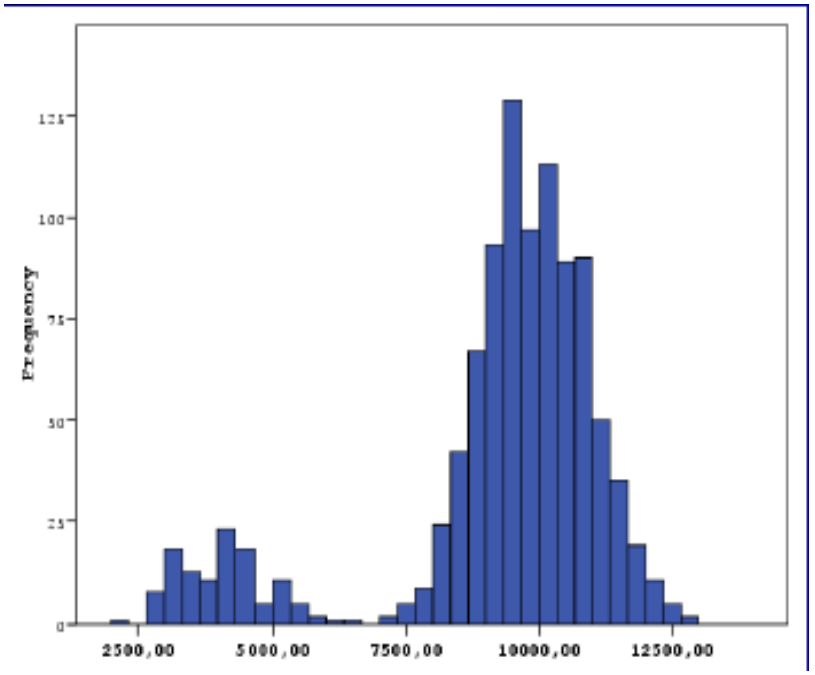
- Според типа на атрибутите
  - количествен
  - качествен
- Според количеството на наблюдаваните параметри
  - една променлива - **univariate statistics**
  - две променливи - **bivariate statistics**
  - много променливи - **multivariate statistics**
- Според използваните методи
  - описателна статистика
  - проверяваща статистика

Описателна статистика

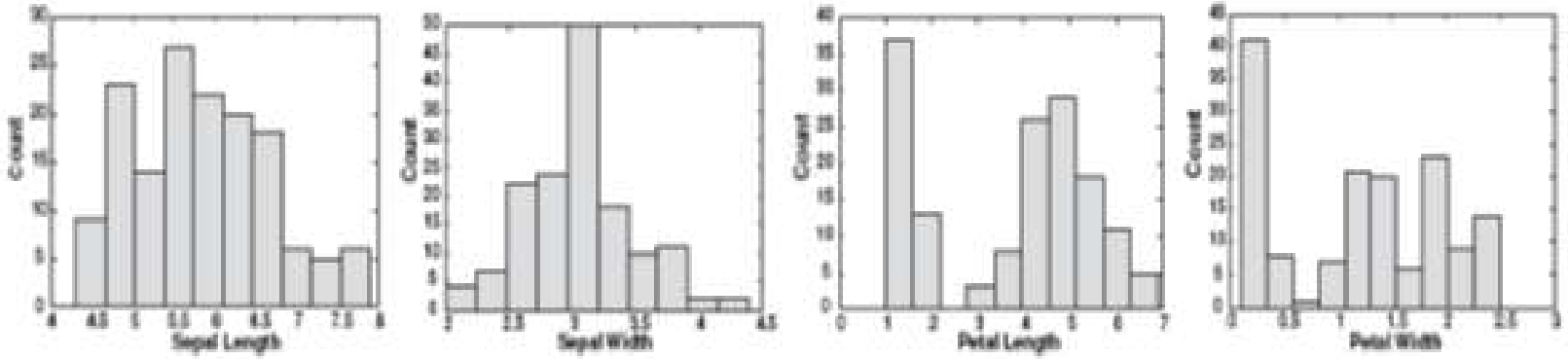
- Дефинира характеристики на извадката
  - разпределение на срещаните стойности на атрибутите
  - основни статистически величини
- Честотно разпределение
  - *normal probability distribution*
  - разделяне на измерените стойности в извадката на групи в равни интервали
  - пребройяване на срещаните стойности във всеки интервал, пропорционално разпределение
  - повечето стойности са близо до средната, някои са далече от нея

Честотно разпределение

- Хистограма
  - на абсцисата се нанасят естествените граници на интервалите
  - по ординатата - абсолютните или относителните честоти
  - в полето на диаграмата се построяват правоъгълници с основа - ширината на интервала и височина - съответната абсолютна или относителна честота

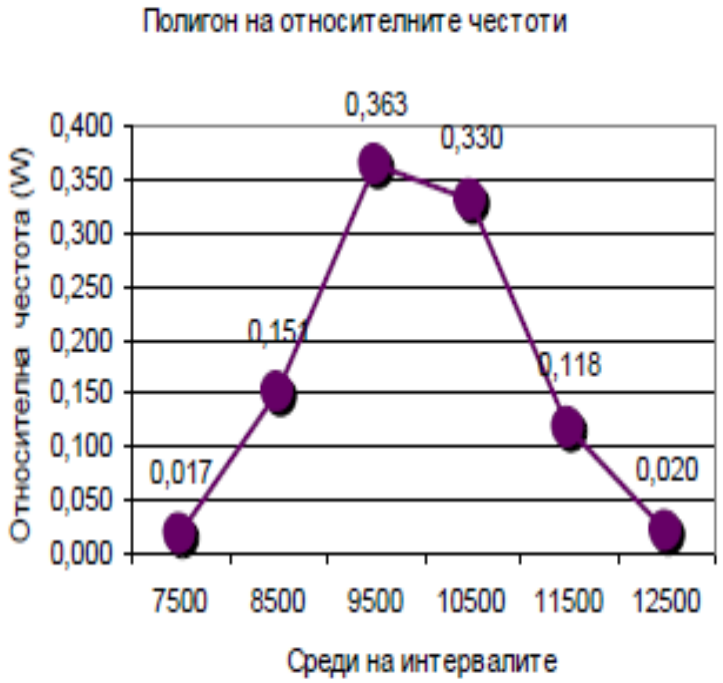


Хистограми


Count
Sepal Length
Count
Sepal Width
Count
Petal Length
Count
Petal Width

Честотно разпределение

- Полигон
  - на абсцисата се нанасят средите на интервалите,
  - на ординатата – абсолютните или относителните честоти
  - в полето на графиката се построяват точки с координати средата на интервала и съответната честота, които се свързват с начупена права линия



Приложения на анализа

- OLTP On-Line Transaction Processing
  - не е специализирано приложение за анализ на данни
  - задача на традиционните релационни бази от данни
- OLAP On-Line Analytical Processing
  - Позволява проверка на хипотези
    - IR показва какво се съдържа в базата данни
    - OLAP показва защо съществуват определени зависимости
  - Прилага се на предварителните етапи на разкриване на информация и знания

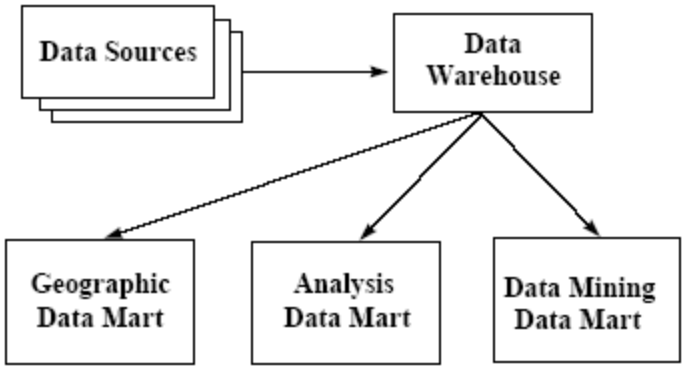
Разлика между OLTP и OLAP

	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

www.cs.uiuc.edu/~hanj

Хранилища на данни за анализ

- Data Warehouse
  - контейнер за данни от различни източници и в различни формати
- Data Mart
  - извадка от данните, подходяща за конкретно изследване; read-only database



Data Warehouse

- Специализирана база от данни
  - за подпомагане на вземане на решения, отделна от операционната база от данни на организацията
  - съдржаща консолидирани данни, полезни за анализа
- Независима във времето
  - не се нуждае от актуално състояние на данните
  - данните съдържат информация за времето
- Информационни операции
  - зареждане
  - четене
  - не се поддържа в стандартния смисъл

Приложения на Data Warehouse

- Обработка на информацията
  - статистичен анализ
  - генериране на отчети
  - визуализация чрез диаграми и др.
- Анализ
  - многомерни анализи на данните
  - OLAP операции- slice, dice, drill, pivot
- Data Mining
  - откриване на знания, чрез разкриване на скрити схеми
  - конструиране на аналитични модели, ассоциране, класифициране
  - описание и прогнозиране