داده کاوی

فصل دوم

تهیه شده توسط
محمد پورزعفرانی
Pourzaferani.ir
دانشگاه اصفهان

دانشگاه اصفهان

تشکر ویژه

- خانم جزینی
- خانم دهقانی
- خانم سلطانی
- آقای تیموری
- خانم امین الرعایایی
- خانم بخشیانی

فصل دوم: داده‌های خود را بشناسید

1) اشیاء داده‌ای و انواع ویژگی‌ها
2) توصیفات آماری داده‌ها
3) اندازه‌گیری میزان شباهت (Similarity) و عدم شباهت (Dissimilarity) داده‌ها
4) مصورسازی داده‌ها (Data Visualization)
5) خلاصه فصل

انواع مجموعه داده‌ها (Dataset)

1. رکوردها (Record)
   a) رکوردهای رابطه‌ای(Relational Records)
   b) ماتریسی از داده‌ها مانند ماتریس عددی یا Crosstabها
   c) اسناد داده‌ای مانند متن‌ها یا term-frequency vector
   d) دا
2. گراف
   a) و
   b) ش
   c) ساختارهای مولکولی

فصل بازی	مهلت	بازنده	برنده	بازی	امتیاز	توپ
۲	۰	۲	۰	۶	۲	۰
۰	۰	۳	۰	۰	۱	۲
۰	۳	۰	۲	۲	۱	۰

TID	Item
۱	نان، زغال، شیر
۲	آبجو، نان
۳	آبجو، زغال، پارچه، شیر
۴	آبجو، نان، پارچه، شیر
۵	زغال، پارچه، شیر

انواع مجموعه داده‌ها (Dataset) (ادامه)

3. داده‌های مرتب (Ordered)
   a) داده های ویدئویی: دنباله ای از تصاویر
   b) داده های وابسته به زمان (Temporal)
   c) داده های ترتیبی: دنباله‌ای از تراکنش‌ها

4. داده‌های فضایی(Spatial)، تصویری و چند رسانه‌ای
   a) داده های فضایی: نقشه
   b) داده های تصویری
   c) داده های ویدئویی

ویژگی‌های مهم داده‌های ساخت‌یافته

1. ابعاد (Dimensionality)
   - Curse of dimensionality
2. پراکندگی یا تنک‌بودن (Sparsity)
   - تنها داده‌های موجود به کار می‌آیند
3. تجزیه و تحلیل، رفع ابهامات (Resolution)
   - داده‌ها وابسته به مقیاس هستند
4. توزیع (Distribution)
   - مرکزیت و پراکندگی

اشیاء داده‌ای(Data Objects)

- مجموعه داده‌ها از اشیاء داده‌ای تشکیل می‌شوند
- **یک شیء داده‌ای نشان‌دهنده یک موجودیت(entity) است**
- **مثال‌ها:**
  - پایگاه داده فروش: مشتریان، اقلام فروشگاه،فروش
  - پایگاده داده پزشکی:بیماران، درمان
  - پایگاه داده دانشگاه: دانشجویان، اساتید، ترم‌ها
- نام‌های دیگر اشیا داده‌ای : *samples , examples, instances, data points, objects, tuples.*
- اشیاء داده ها توسط **ویژگی‌ها(Attribute)** توصیف می‌شوند
- سطرهای پایگاه داده -> اشیاء داده ای
- ستون ها -> ویژگی‌ها

خصوصیات یا ویژگی‌ها(Attributes)

- نام‌های دیگر: **dimensions, features, variables**
- تعریف: ویژگی داده‌ای است که خصوصیات یک شی داده‌ای را توصیف می‌کند
- مثال: شناسه مشتری، نام، آدرس

- **انواع ویژگی‌ها**

1) Nominal
2) Binary
3) Ordered
4) quantitative : Numeric
   a) Interval-scaled
   b) Ratio-scaled

انواع ویژگی‌ها

- **اسمی (Nominal):** categories, states, or “names of things”
- چنانچه دسته‌هایی که متغیر در آنها قرار می گیرد، دارای هیچگونه ترتیب طبیعی نباشد، هر یک از آن متغیرها را متغیر صوری یا اسمی کیفی می‌نامند
  - مانند: رنگ مو ={ قهوه ای مایل به قرمز، سیاه، بور، قهوه ای، خاکستری، قرمز، سفید}
  - وضعیت تاهل، وضعیت اشتغال، تعداد شناسه ID، کد پستی

انواع ویژگی‌ها (ادامه)

- **دودویی(Binary):** ویژگی‌هایی که تنها دو وضعیت دارند
- (۰ یا ۱)

– دودویی متقارن (Symmetric binary): هر دو اهمیت یکسان داشته باشند مانند جنسیت (فرقی در اینکه کدام صفر یا یک باشد، نباشد)

– دودویی نامتقارن (Asymmetric binary): به یک نسبت اهمیت ندارند مانند نتیجه آزمایش HIV (با اهمیت را یک میدهیم)

انواع ویژگی‌ها (ادامه)

- **داده‌های ترتیبی (Ordinal)**
  - در یک جهت معنادار ارزش دارند (رتبه بندی) اما مقدار بین مقادیر متوالی شناخته شده نیست.
  - مثل اندازه = {کوچک، متوسط، بزرگ}
    - نمرات
    - رتبه بندی ارتش

انواع ویژگی‌های عددی

- **عددی (Numeric):** این نوع ویژگی‌ها کمیتی هستند (عدد صحیح یا حقیقی)
- فاصله **(Interval)**
  - اندازه‌گیری دریک مقیاس از واحدهای هم اندازه
  - مقادیر ترتیب دارند
    - برای مثال، درجه حرارت در سانتیگراد یا فارنهایت، تاریخ های تقویم
  - عدم وجود نقطه صفر واقعی
- نسبت**(Ratio)**
  - نقطه صفر ذاتی دارند Inherent **zero-point**
  - میتوانیم بین (مثلا ۱۰ درجه کلوین دو برابر ۵ درجه کلوین است)
  - مثالها: درجه حرارت کلوین، طول، تعداد، مقدار پول (کمیت مالی)

تقسیم ویژگی‌ها از منظری دیگر

- ویژگی‌های گسسته (Discrete Attribute)
  - فقط یک مجموعه متناهی یا نامتناهی از مقادیر قابل شمارش دارد
  - به عنوان مثال، کدهای فشرده، حرفه ای، یا مجموعه‌ای از کلمات در مجموعه‌ای از اسناد
  - گاهی اوقات به صورت عدد صحیح نمایش داده می‌شود

  - **توجه:** ویژگی‌های دودویی یک حالت خاص از ویژگی‌های گسسته هستند.

تقسیم ویژگی‌ها از منظری دیگر

- ویژگی‌های پیوسته (Continuous Attribute)
  - دارای عددهای حقیقی به عنوان مقادیر ویژگی هستند
  - برای مثال درجه حرارت، قد، وزن
  - در عمل، فقط مقادیر حقیقی می توانند اندازه گیری شوند و با استفاده از تعدادی ارقام متناهی نمایش داده می شوند.
  - ویژگی های پیوسته معمولا به صورت متغیرهای ممیز شناور نمایش داده می شوند.

فصل دوم: داده‌های خود را بشناسید

1) اشیاء داده‌ای و انواع ویژگی‌ها
2) توصیفات آماری داده‌ها
3) اندازه‌گیری میزان شباهت (Similarity) و عدم شباهت (Dissimilarity) داده‌ها
4) مصورسازی داده‌ها (Data Visualization)
5) خلاصه فصل

توصیفات آماری داده‌ها

- انگیزه: برای درک بهتر داده ها: شاخص مرکزی، تنوع(variation) و گستردگی داده‌ها (spread)
- شاخصه‌های پراکندگی داده‌ها: میانه، ماکزیمم، مینیمم، چارک، داده های پرت، واریانس، و غیره
- ابعاد عددی
  - پراکندگی داده ها: این شاخص با دقت‌های مختلفی (granularities) قابل اندازه‌گیری است
  - استفاده از Boxplot و تحلیل چارک برای بازه‌ها

شاخص‌های اندازه‌گیری مرکزیت داده‌ها

- میانگین (Mean) (اندازه گیری جبری)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- میانگین وزنی

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- میانگین هرس شده (Trimmed mean) : جدا کردن مقادیر افراطی(مثلا ۲٪ بزرگترین و کوچکترین مقادیر را حذف کنیم)

شاخص‌های اندازه‌گیری مرکزیت داده‌ها

- میانه (Median)
- در مجموعه ای از داده های مرتب شده، اگر تعداد این داده ها فرد باشد، میانه این مجموعه داده ای است که در وسط این مجموعه قرار دارد و اگر تعداد این مجموعه زوج باشد، میانه، میانگین دو داده ای است که در وسط این مجموعه قرار دارند.

تخمین میانه برای داده‌های گروه‌بندی شده

$$median = L_1 + \left(\frac{n/2 - (\sum freq)l}{freq_{median}}\right) width$$

که در آن L1 کران پایین دسته ای است که در آن میانه قرار دارد، n تعداد داده ها ، $(\sum freq)_l$ فراوانی تجمعی دستهٔ قبل از دستهٔ میانه ، $freq_{median}$ فراوانی دستهٔ میانه و w اندازه یا طول بازه میانی.

شاخص‌های اندازه‌گیری مرکزیت داده‌ها

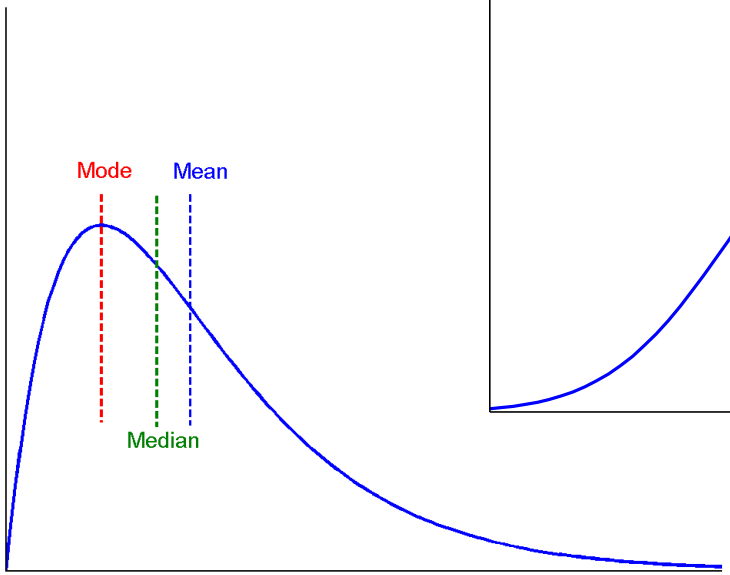
- نما یا مد (Mode)
- داده ای که اغلب در داده ها رخ می‌دهد(فراوانی آن از سایر داده ها بیشتر است)
- اصطلاحات Unimodal, bimodal, trimodal
  - مجموعه داده با یک، دو یا سه مد
  - در حالتی مانند Customer_ID که تمام مقادیر یکبار تکرار میشوند به آن no mode میگوییم
- فرمول تجربی:

$$mean - mode = 3 \times (mean - median)$$

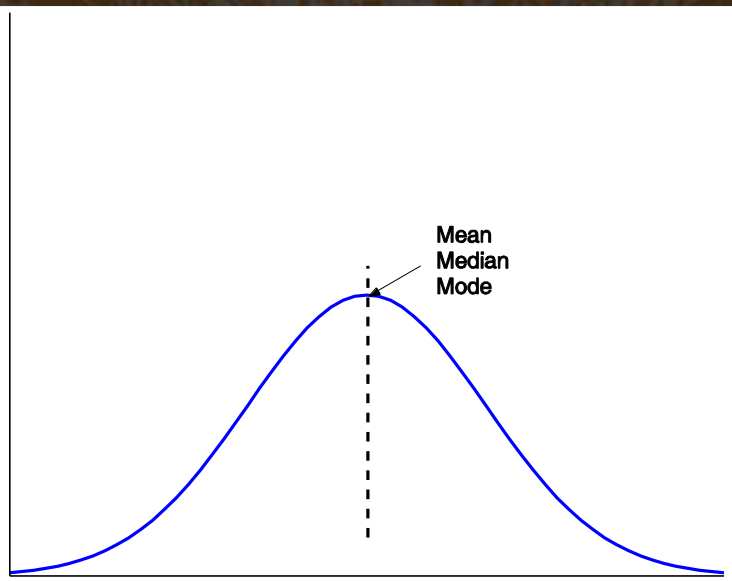
- وسط بازه (Midrange) :
  - میانگین بزرگترین و کوچکترین مقدار در مجموعه داده

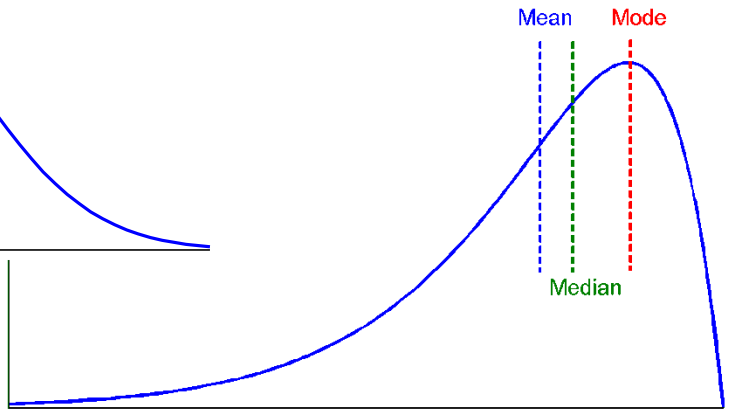
داده‌های متقارن یا نامتوازن(Symmetric vs. Skewed)

- میانه، میانگین و مد درحالت‌های متقارن و یا چوله به چپ یا راست



اریب مثبت(چولگی به راست)





اریب منفی(چولگی به چپ)

اندازه‌گیری پراکندگی داده‌ها

Quartiles, outliers and boxplots

- **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
- **Inter-quartile range**: IQR = $Q_3 - Q_1$
  - **a simple measure of spread that gives the range covered by the middle half of the data**
- **Five number summary**: min, Q_1, median(Q_2), Q_3, max
- **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
- **Outlier**: usually, a value higher/lower than 1.5 x IQR

$30 + 36 + 47 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110$

*Q*1 D \$47,000 and *Q*3 is \$63,000 , IQR=\$16,000

Variance and standard deviation (*sample: s, population: σ)*

- **Variance**: (algebraic, scalable computation)

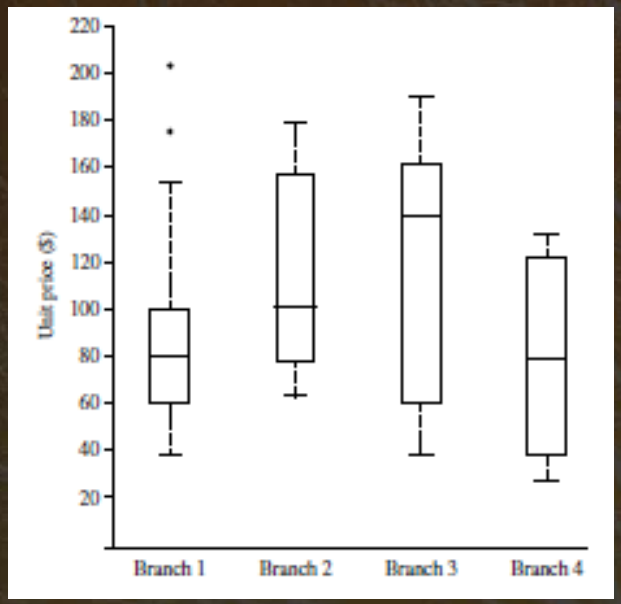
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$$

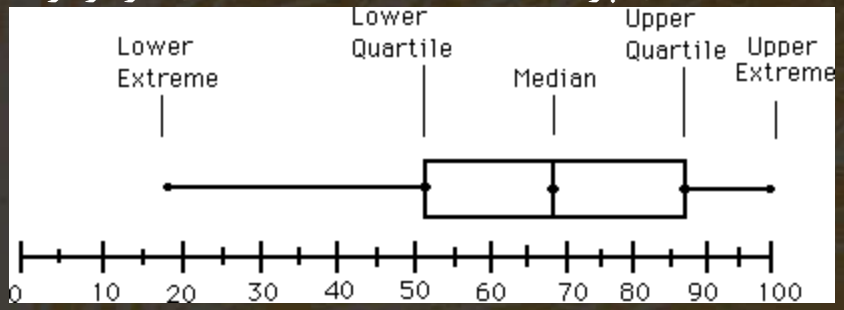
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \mu^2$$

تحلیل Boxplot

- **Five-number summary** :Minimum, Q1, Median, Q3, Maximum

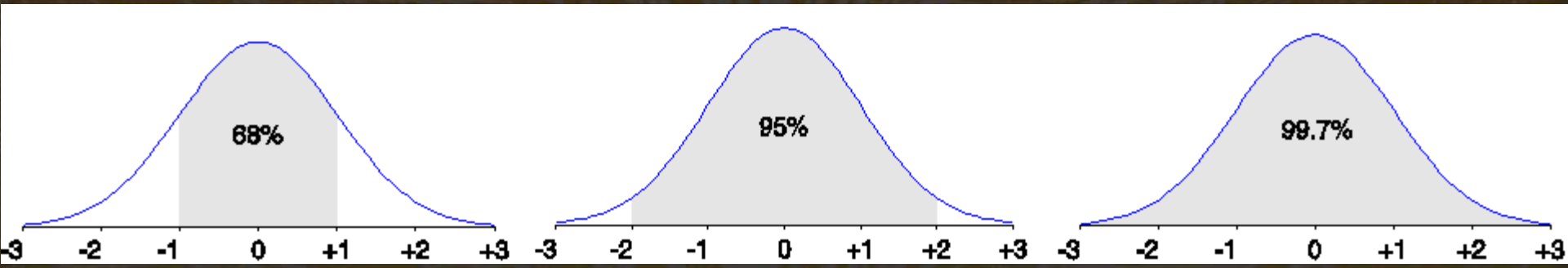
- **Boxplot**
  - داده ها با یک جعبه نشان داده شده است.
  - انتهای جعبه در چارک اول و سوم است، بنابراین ارتفاع جعبه IRQاست.
  - میانه توسط یک خط ، در جعبه مشخص میشود
  - Whisker : دو خط در خارج از جعبه که به حداقل(min) و حداکثر(max) کشیده میشوند
  - داده های پرت (Outliers): نقطه‌های فراتر از آستانه





ویژگیهای نمودار توزیع نرمال

- نمودار توزیع نرمال
  - از $\mu-\sigma$ تا $\mu+\sigma$: تقریبا شامل ۶۸٪ از اندازه گیری ها(μ: میانگین, σ:انحراف معیار)
  - از $\mu-2\sigma$ تا $\mu+2\sigma$: تقریبا شامل۹۵٪ از آن
  - از $\mu-3\sigma$ تا $\mu+3\sigma$: تقریبا شامل ۹۹.۷٪ از ان

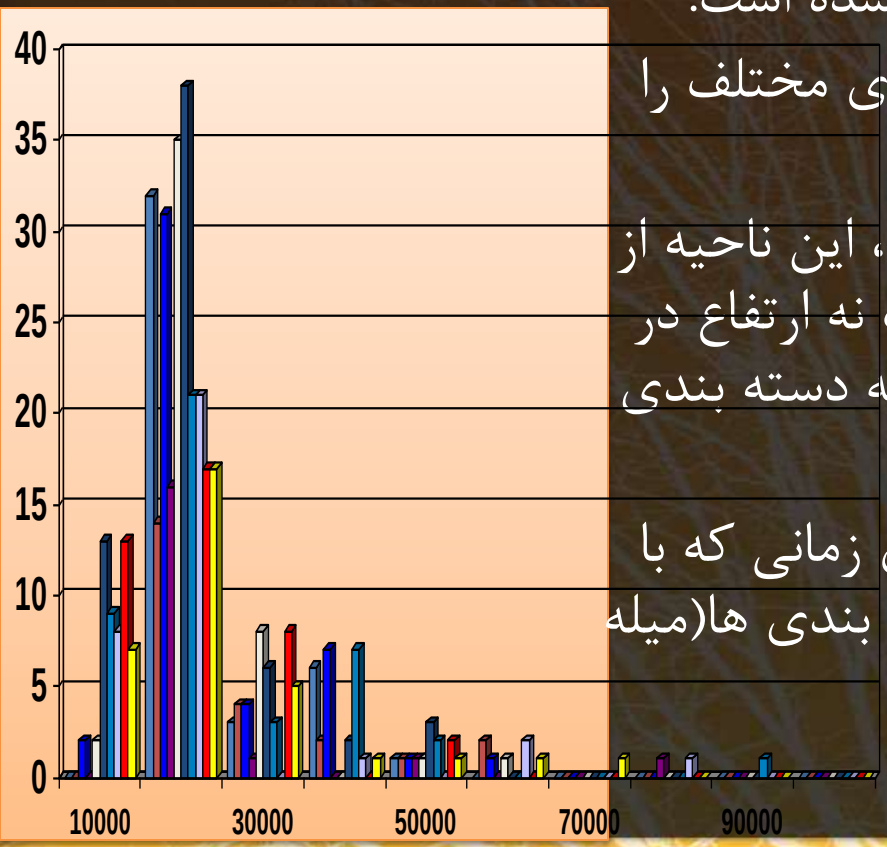


نمایش گرافیکی از توصیفات آماری

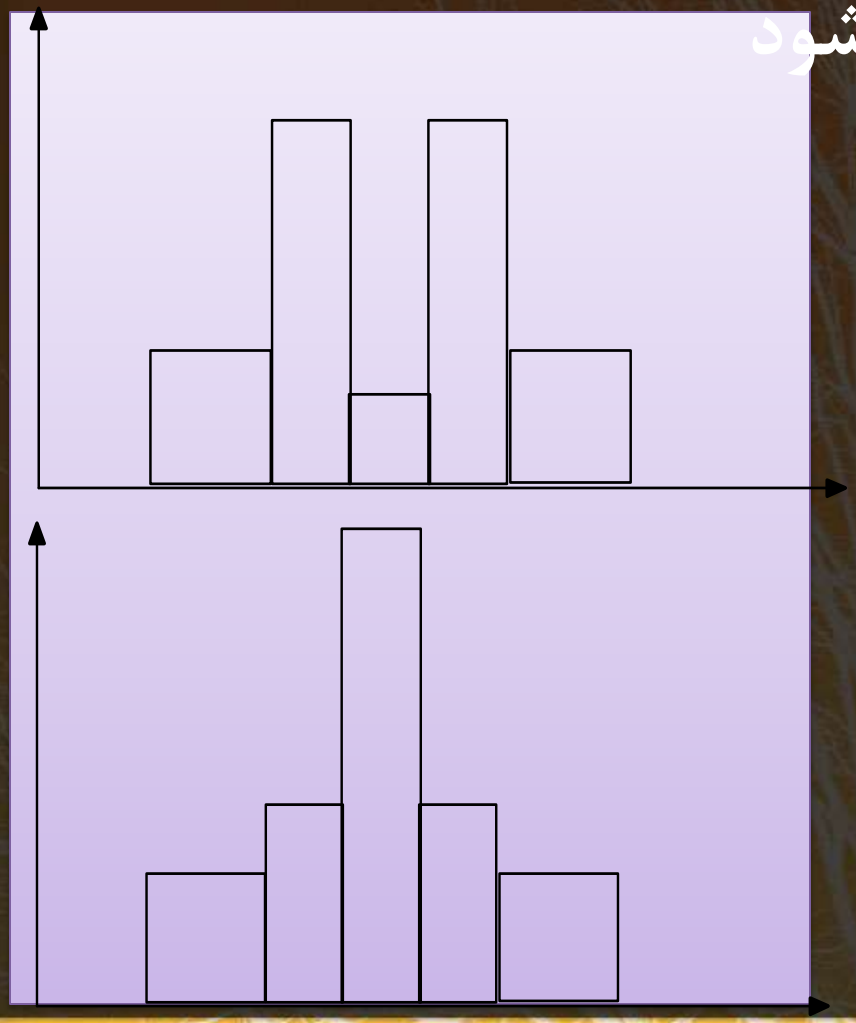
- **Boxplot**: نمایش گرافیکی از خلاصه ۵ عدد
- **Histogram**: محور xها مقادیر(value) هستند و محور yها نشان دهنده فرکانس تکرار
- **Quantile plot: هر مقدار x_i با زوج f_i** *نشان می دهد که حدود ٪۱۰۰ f_i* **داده ها از x_i کوچکتر هستند**
- **Quantile-quantile (q-q) plot: نمودار quantiles از یک توزیع univariant در مقابل quantiles مربوطه از دیگری**
- **Scatter plot: هر جفت از مقادیر یک جفت از مختصات هستند که به صورت یک نقطه در صفحه ترسیم می شوند.**

تحلیل هیستوگرام

- Histogram: نمایش نمودار فرکانس‌های جدول‌بندی شده، به صورت نمودار میله ای نشان داده شده است.
- این نسبت موارد سقوط هر یک از دسته های مختلف را نشان می دهد.
- متفاوت از یک نمودار میله ای این است که، این ناحیه از نمودار است که نشان دهنده ی ارزش است نه ارتفاع در نمودار میله ای، تمایز حیاتی زمانی است که دسته بندی ها از عرض یکسان نیستند.
- دسته بندی ها معمولا به عنوان متغییرهای زمانی که با هم تداخل ندارند مشخص می شوند. دسته بندی ها(میله ها) باید در مجاورت یکدیگر باشند.



هیستوگرام‌ها اغلب بیشتر از Boxplotsها اطلاعات دارند گفته می شود



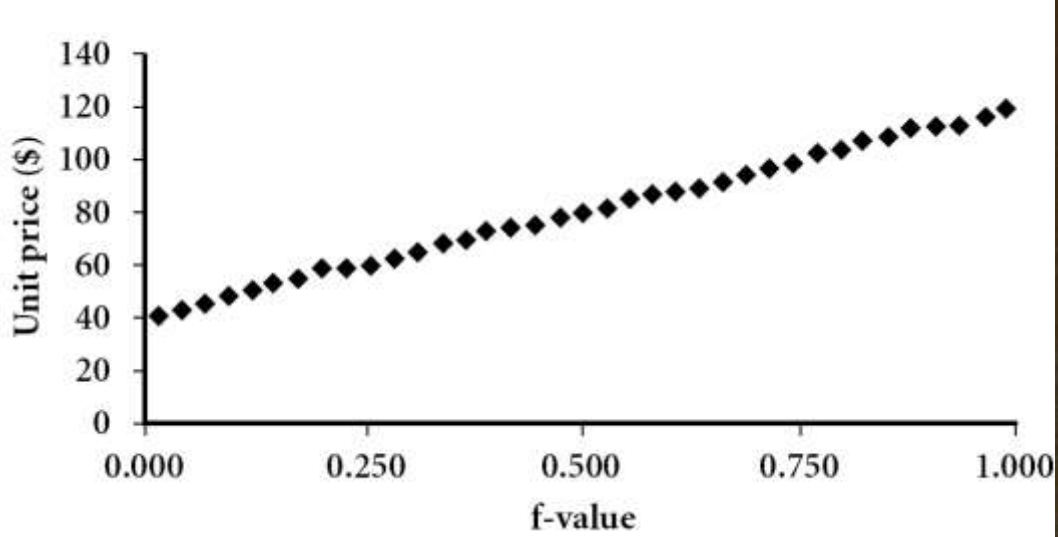
- دو نمودار هیستوگرام نشان داده شده در سمت چپ ممکن است همان boxplot باشد.
  - مقادیر یکسان برای: min, Q1, median, Q3, max
- اما آنها تقریبا متفاوت از توزیع داده ها هستند.

Quantile Plot

- نمایش تمام داده ها(اجازه دادن به کاربر برای ارزیابی هر دو رفتار کلی و رخدادهای غیر معقول)
- نمودار اطلاعات quantile
  - برای داده های x_i، داده ها به صورت صعودی مرتب می شوند، f_i نشان می دهد که تقریبا f_i% از داده ها کمتر و یا برابر x_i هستند.

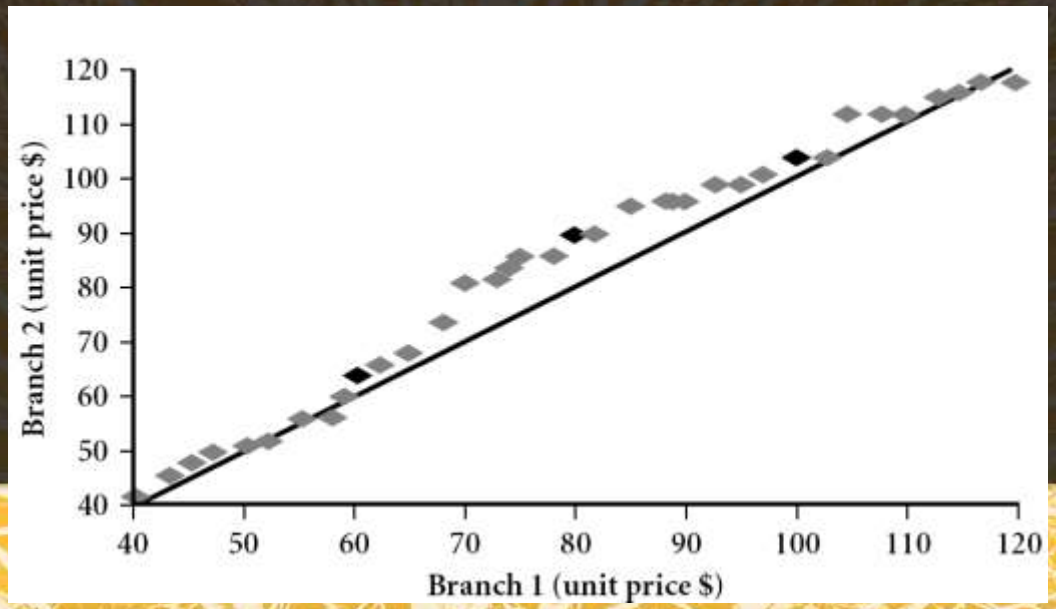
Table 2.1 A Set of Unit Price Data for Items Sold at a Branch of *AllElectronics*

Unit price ($)	*Count of items sold*
40	275
43	300
47	250
–	–
74	360
75	515
78	540
–	–
115	320
117	270
120	350



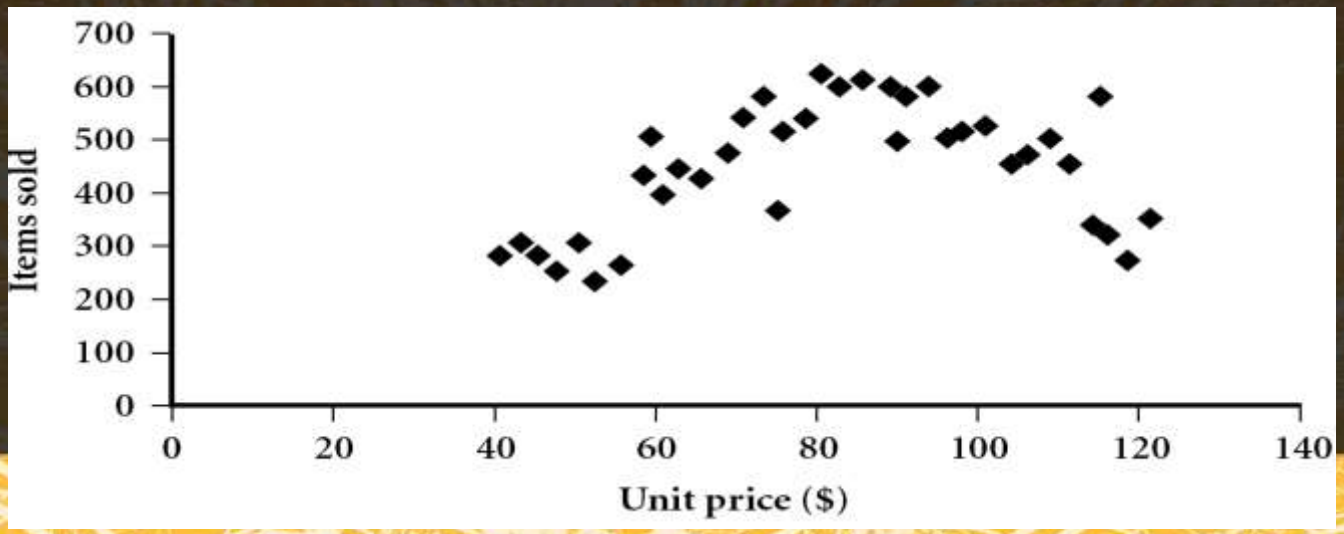
Quantile-Quantile (Q-Q) Plot

- گراف های quantiles از یک توزیع تک متغییری در برابر quantiles مربوطه از دیگری
- مشخصات: آیا یک shift برای رفتن از یک توزیع به دیگری وجود دارد؟
- مثال زیر، قیمت واحد از اقلام به فروش رسیده در شعبه ۱ را در برابر شعبه ۲ برای هر quantiles نشان می دهد. تمایل افراد برای قیمت های واحد به فروش رسیده در شعبه ۱ کمتر از شعبه۲ می باشد.

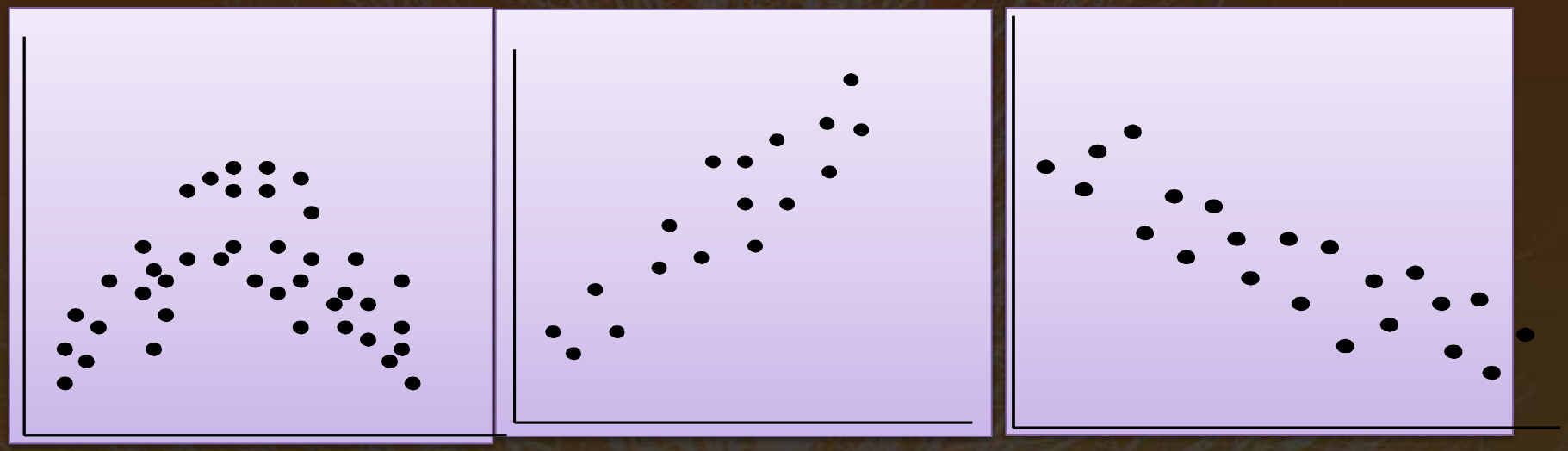


Scatter plot

- یک نگاه اولیه از داده های دو متغییره برای دیدن خوشه ای از نقاط، داده های پرت(outliers)، یا بررسی امکان روبط همبستگی را فراهم می کند.
- هر جفت از مقادیر به عنوان یک نقطه از مختصات در نمودار ترسیم می شود.



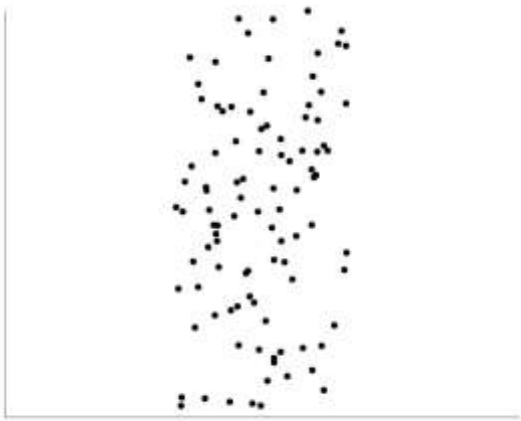
وابستگی مثبت و منفی داده‌ها



نیم قطعه ی سمت چپ وابستگی مثبت

نیمه ی سمت راست وابستگی منفی

Uncorrelated Data داده های غیر وابسته



فصل دوم: داده‌های خود را بشناسید

1) اشیاء داده‌ای و انواع ویژگی‌ها
2) توصیفات آماری داده‌ها
3) اندازه‌گیری میزان شباهت (Similarity) و عدم شباهت (Dissimilarity) داده‌ها
4) مصورسازی داده‌ها (Data Visualization)
5) خلاصه فصل

شباهت و عدم شباهت

- **شباهت:**
- اندازه گیری عددی از چگونگی یکسان بودن دو شی داده
- زمانی که اشیا شباهت بیشتری به هم دارند ارزش بیشتری دارد.
- اغلب در دامنه ی[0,1] بیان می‌شود
- **عدم شباهت:** (به عنوان مثال : فاصله)
- اندازه گیری عددی اختلاف دو شی داده
- زمانی که اشیا شباهت بیشتری به هم دارند ارزش کمتری دارد.
- حداقل عدم شباهت زمانی است که مقدار صفر باشد
- **ماتریس مجاورت (Proximity):** به شباهت و عدم شباهت اشاره می کند.

ماتریس داده و ماتریس عدم شباهت

- **ماتریس داده**
  - N نقطه داده با ابعاد p
  - دارای دو حالت است:

سطرها برای اشیا وستون‌ها برای ویژگی‌ها

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- **ماتریس عدم شباهت**
  - N نقطه داده، اما تنها فاصله درج میشود
  - یک ماتریس مثلثی می باشد
  - تنها حالت: یک نوع موجودیت

(عدم شباهت)

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

معیار شباهت برای ویژگی‌های Nominal

- دو یا چند حالت می تواند به خود بگیرد مثلا: قرمز، زرد، آبی و سبز (تعمیمی از ویژگی دودویی است)
- روش اول: تطابق ساده (Simple matching)
- m: تعداد تطابق‌ها و p: تعداد کل ویژگی‌ها

$$d(i,j)=\frac{p-m}{p}$$

$$sim(i,j)=1-d(i,j)=\frac{m}{p}.$$

- روش دوم: از تعداد زیادی از ویژگی‌های دودویی استفاده میکند
- برای هریک از M حالت اسمی یک ویژگی دودویی جدید ایجاد میکنیم

معیار شباهت برای ویژگی‌های Nominal

- از آنجایی که ما یک خصوصیت اسمی داریم، test-1 ، ما در معادله p=1 قرار می دهیم. اگر اشیا i , j شبیه باشند d(i,j) با 0 و در غیر این صورت با 1 ارزیابی خواهند شد.

A Sample Data Table Containing Attributes of Mixed Type

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

معیار شباهت برای ویژگی‌های Binary

- یک جدول تصادفی از خصوصیات دودویی.

Object *j*

Object *i*	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

- اندازه گیری فاصله برای متغیرهای دودویی متقارن

$$d(i, j) = \frac{r+s}{q+r+s+t}$$

- اندازه گیری فاصله برای متغیرهای دودویی نامتقارن

$$d(i, j) = \frac{r+s}{q+r+s}$$

- زمانی که هر دو یک باشند مهتر است از زمانی که هر دو صفر باشند

معیار شباهت برای ویژگی‌های Binary

- ضریبJaccard(میزان شباهت برای متغیرهای دودویی نامتقارن):

$$sim_{Jaccard}(i, j) = \frac{q}{q + r + s}$$

	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

- توجه داشته باشید: ضریب Jaccard مشابه coherenceاست

$$coherence(i, j) = \frac{sup(i, j)}{sup(i) + sup(j) - sup(i, j)} = \frac{q}{(q + r) + (q + s) - q}$$

عدم شباهت بین متغیرهای باینری

- مثال:

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

- جنسیت یه ویژگی متقارن است.
- مابقی ویژگی‌ها، دودویی نامتقارن هستند.
- مقادیر p و y را با 1 و N را با 0 مقدار دهی کنید.

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

$$d(i, j) = \frac{r+s}{q+r+s}$$

استاندارد سازی داده‌های عددی (Numeric)

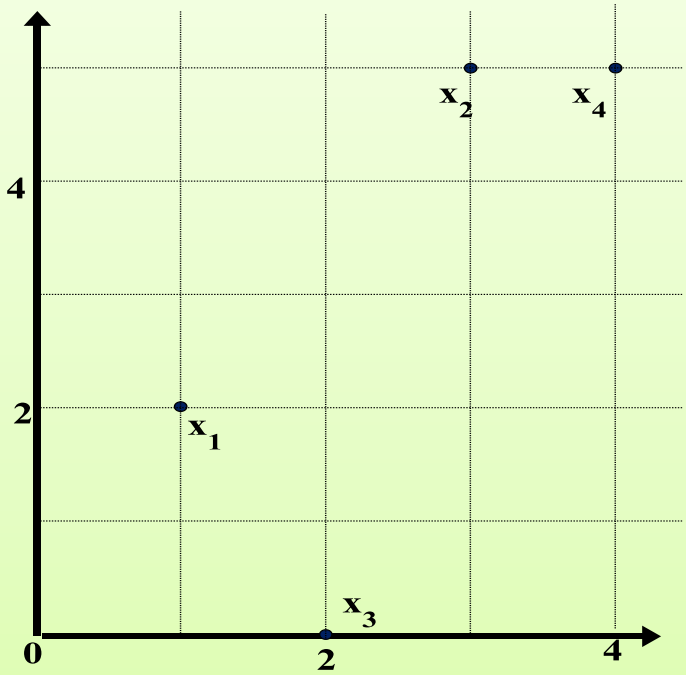
- Z-Score : $$z = \frac{x-\mu}{\sigma}$$
- X: عدد خام اولیه . μ: میانگین اعداد .σ : انحراف معیار
- فاصله بین مقدار خام و میانگین در واحد انحراف معیار
- زمانی که مقدار خام زیر میانگین باشد منفی و زمانی که بالای آن باشد مثبت است.
- یک روش جایگزین: محاسبه میانگین انحراف مطلق

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|)$$

که در آن داریم: $$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf}).$$

- میزان استاندارد شده(مقدارz): $$z_{if} = \frac{x_{if} - m_f}{s_f}$$
- استفاده از میانگین انحراف مطلق بهتر از انحراف استاندارد است

مثال ماتریس عددی و ماتریس عدم شباهت



Data Matrix

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Dissimilarity Matrix

(with Euclidean Distance)

	x1	*x2*	*x3*	*x4*
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

فاصله در داده‌های عددی:فاصله Minkowski

- فاصله Minkowski یک میزان فاصله مشهور است.

$$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

در اینجا $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ و $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ دو شی داده p بعدی هستند

خصوصیات:

- $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$(مولفه ی مثبت)
- $d(i, j) = d(j, i)$
- $d(i, j) \leq d(i, k) + d(k, j)$(نابرابری مثلث)
- فاصله‌ای که این ویژگی‌ها را داشته باشد **معیار** مینامیم

موارد خاص فاصله Minkowski

- اگر h=1 باشد فاصله منهتن نامیده میشود (Manhattan)

$$d(i,j)=|x_{i_1}-x_{j_1}|+|x_{i_2}-x_{j_2}|+\ldots+|x_{i_p}-x_{j_p}|$$

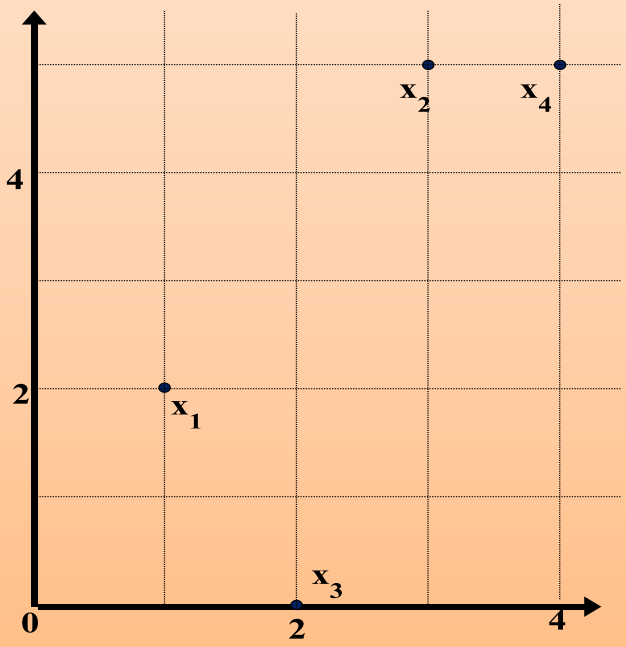
- اگر h=2 به آن فاصله اقلیدسی میگویند (Euclidean)

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+\ldots+|x_{i_p}-x_{j_p}|^2)}$$

- اگر $h \to \infty$ باشد به آن فاصله Supremum
- برای محاسبه کافی است بزرگترین اختلاف در ویژگی‌ها را در نظر بگیریم

مثال: فاصله Minkowski

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

متغیرهای ترتیبی(Ordinal)

- متغیرهای ترتیبی میتوانند پیوسته یا گسسته باشند
- روش محاسبه
  - f نشان دهنده ویژگی، Mf تعداد حالات ممکن f بنابراین داریم 1,....,Mf
  - پس از مرتب کردن به هر مقدار یک رتبه میدهیم rf
  - با استفاده از فرمول زیر مقدار نرمال شده جایگزین رتبه عدد میشود

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - پس از آن میتوان با استفاده از معیارهای پیشین فاصله بین دو شی را سنجید

مثال: محاسبه فاصله برای ویژگی‌های ترتیبی

- جدول روبه‌رو داده شده

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

محاسبه عدم شباهت با درنظرگیری انواع ویژگی‌ها

یک پایگاه داده ممکن است شامل همه‌ی انواع خصیصه ها باشد

– اسمی، باینری متقارن، باینری متقارن، عددی، ترتیبی

هنگامی که ممکن است یک فرمول وزن به منظور ترکیب اثرات آنها استفاده شود.

$$d(i,j)=\frac{\sum_{f=1}^{p}\delta_{ij}^{(f)}d_{ij}^{(f)}}{\sum_{f=1}^{p}\delta_{ij}^{(f)}}$$

– *اگر* δ_{ij} *در صورتی که مقادیر miss باشند برابر صفر و در غیر اینصورت ویژگی حساب میشود*

– *اگر F* نومینال یا باینری باشد:

$d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise

– *F* عددی باشد: استفاده از فاصله نرمال شده

$$d_{ij}^{(f)} = \frac{|x_{if}-x_{jf}|}{max_h x_{hf}-min_h x_{hf}},$$

– *F* ترتیبی باشد

- محاسبه ی رتبه های r_{if} و
- رفتار z_{if} به عنوان فاصله-کوچک

$$z_{if}=\frac{r_{if}-1}{M_f-1}$$

مثال برای انواع ویژگی‌ها

- مجموعه داده

A Sample Data Table Containing Attributes of Mixed Type

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

$$d(i,j)=\frac{\sum_{f=1}^{p}\delta_{ij}^{(f)}d_{ij}^{(f)}}{\sum_{f=1}^{p}\delta_{ij}^{(f)}}$$

Test1

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

Test2

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

Test3

$$\begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1.00 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}.$$

$$\begin{bmatrix} 0 & & & \\ 0.85 & 0 & & \\ 0.65 & 0.83 & 0 & \\ 0.13 & 0.71 & 0.79 & 0 \end{bmatrix}.$$

شباهت کسینوسی

- یک سند (Document) میتواند به وسیله ی هزاران ویژگی نمایش داده شود، که هر یک تعداد تکرار یک واژه در سند را نشان میدهند

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- برنامه های کاربردی: بازیابی اطلاعات، طبقه بندی بیولوژیک، نگاشت ویژگی ژنتیک،....
- اندازه کسیونسی: اگر d_1 , d_2 دو بردارباشند (به عنوان مثال، بردارهای نرخ تکرار)، آنگاه
- $\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2||$,

که در آن نقطه ضرب نقطه ای بردار را نشان می دهد و $||d_1||$ طول بردار را

مثال برای شباهت کسینوسی

- مثال: یافتن**شباهت** بین داکیومنت ۱ و ۲.

$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$
$d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$

$d_1 \bullet d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25$

$||d_1|| = (5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$||d_2|| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5} = (17)^{0.5} = 4.12$

$\cos(d_1, d_2) = 0.94$

فصل دوم: داده‌های خود را بشناسید

1) اشیاء داده‌ای و انواع ویژگی‌ها
2) توصیفات آماری داده‌ها
3) اندازه‌گیری میزان شباهت (Similarity) و عدم شباهت (Dissimilarity) داده‌ها
4) مصورسازی داده‌ها (Data Visualization)
5) خلاصه فصل

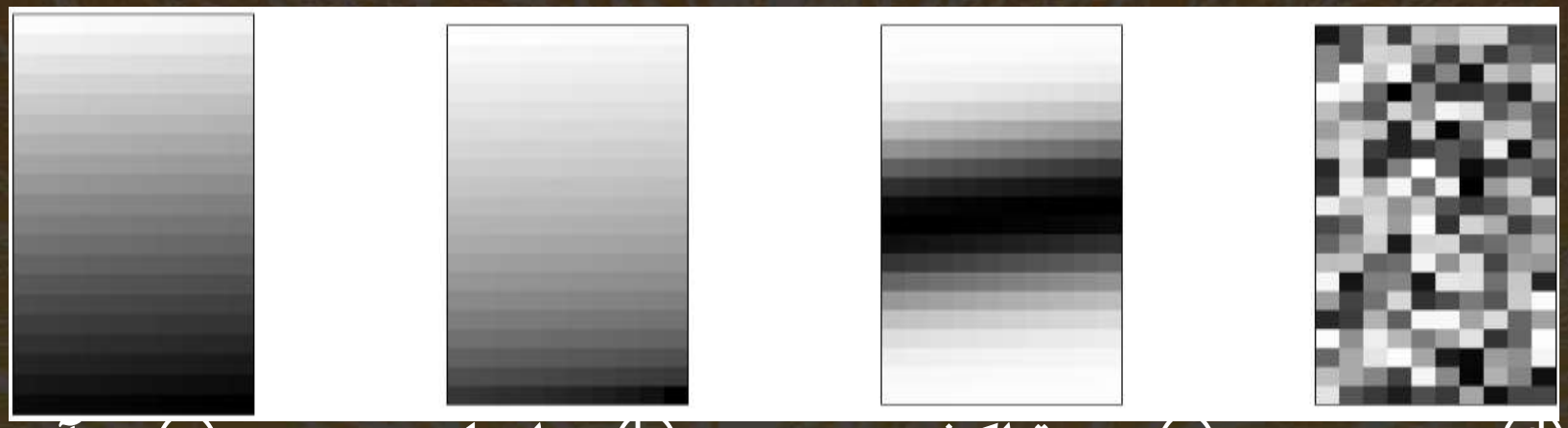
مصورسازی داده‌ها

- چرا مصورسازی داده؟
  - دستیابی به نگرشی از فضای اطلاعاتی به کمک نگاشت داده‌ها به اشکال گرافیکی
  - ارائه کلی کیفی مجموعه داده های بزرگ
  - جستجو برای الگوها، روندها، ساختار، بی نظمی ها، روابط بین داده ها
  - کمک به یافتن بخش های جالب و پارامترهای مناسب جهت تجزیه و تحلیل های بیشتر

- دسته بندی روشهای مصورسازی:
  - تکنیک های تصویر سازی پیکسل گرا
  - تکنیک های تصویر سازی طرح هندسی
  - تکنیک های تصویر سازی مبتنی بر آیکون
  - تکنیک های تصویر سازی سلسله مراتبی
  - تصویرسازی داده های پیچیده و روابط

تکنیک های تصویرسازی پیکسل گرا

- برای یک مجموعه داده m بعدی، m پنجره روی صفحه نمایش ایجاد می شود، به ازای هر بُعد یک پنجره.
- مقادیر m بعدی از یک رکورد برای m پیکسل در موقعیت متناظر در پنجره ترسیم شده اند .
- رنگ پیکسل نمایانگر مقادیر مربوطه است.
- به ظور مثال: در یک مجموعه داده رکوردها بر اساس درآمد مرتب شده‌اند
- با استفاده از این نمایش میتوان ارتباط بین ویژگی‌ها را می‌توان به سادگی فهمید

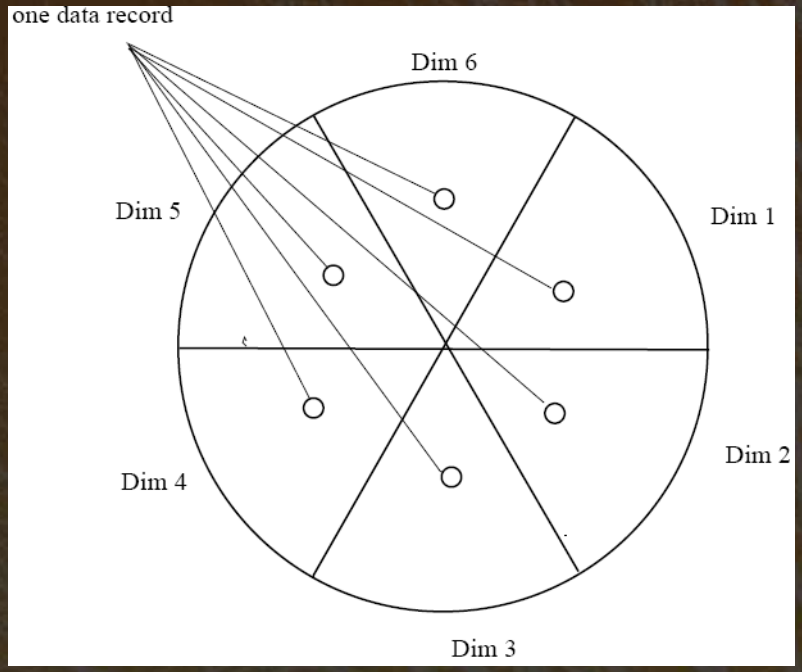


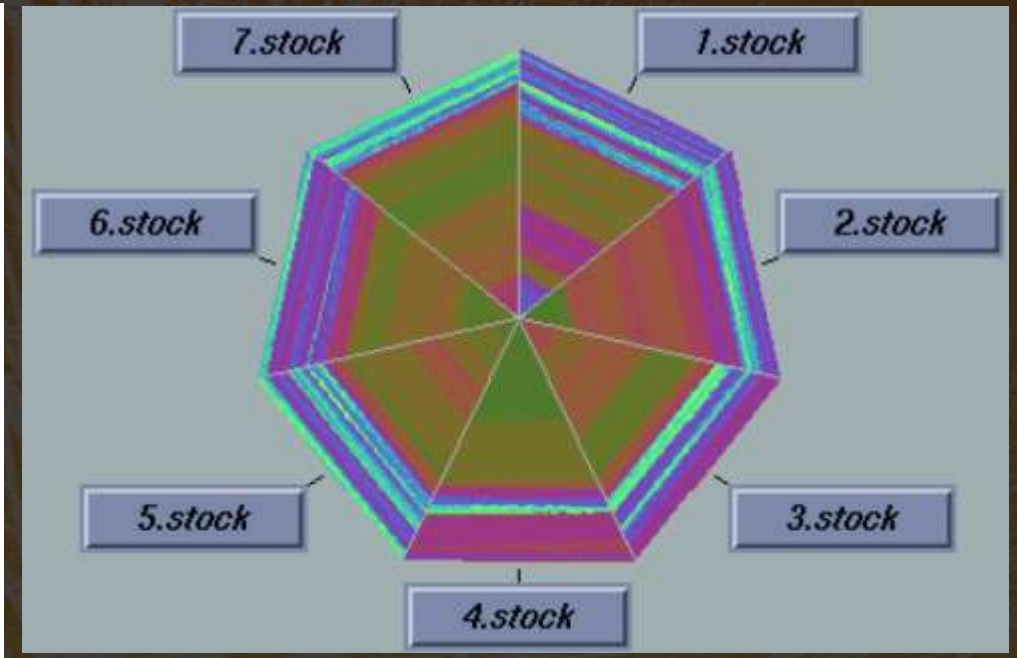
(a) درآمد (b) حد اعتبار (c) حجم تراکنش (d) سن

نمایش پیکسل ها در بخش های دایره ای

- برای صرفه جویی در فضا و نشان دادن ارتباط میان ابعاد مختلف، پر کردن فضا اغلب در بخش های دایره ای انجام می شود.
- مقاله:
- ‘Circle Segments’: A Technique for Visually Exploring Large Multidimensional Data Sets

نمایش پیکسل ها در بخش های دایره ای





تصویرسازی هندسی (Geometric)

- انواع روش‌ها

Direct visualization –

Scatterplot and scatterplot matrices –

Landscapes –

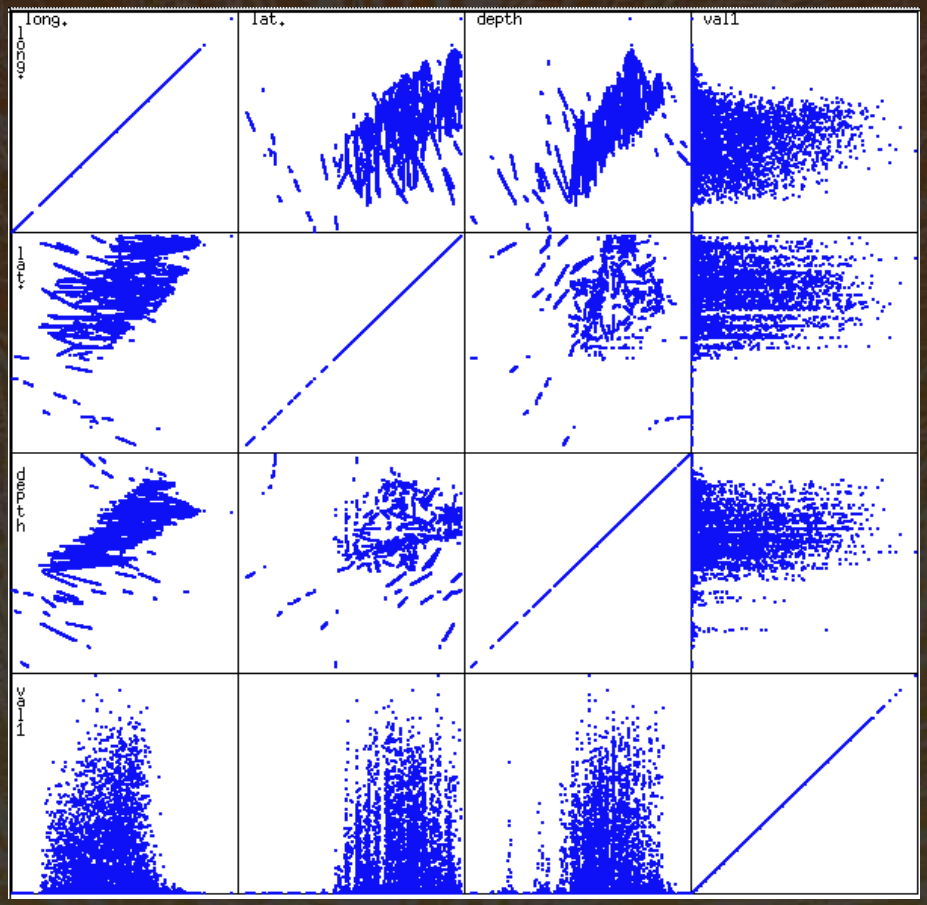
Projection pursuit technique: Help users find – meaningful projections of multidimensional data

Prosection views –

Hyperslice –

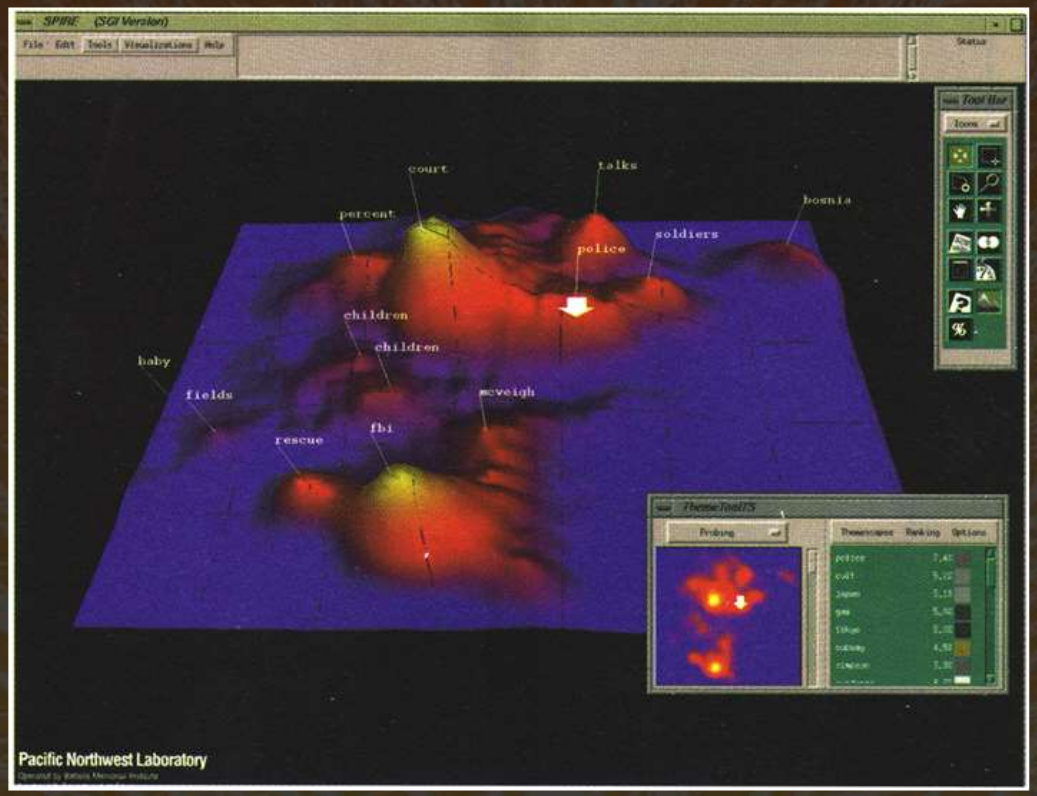
Parallel coordinates –

Scatterplot Matrices



Landscapes

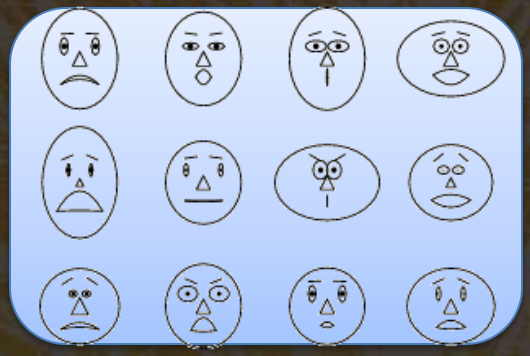
Used by permission of B. Wright, Visible Decisions Inc.



news articles visualized as a landscape

تکنیک های تصویرسازی مبتنی بر آیکون

- چهره های چرنوف
- راهی برای نشان دادن متغیر در یک سطح دو بعدی است ، به عنوان مثال، فرض کنید x کجی ابرو، y اندازه چشم و z طول بینی، و ...
- شکل ها، چهره های تولید شده با استفاده از ۱۰ ویژگی (خروج از مرکز سر، اندازه چشم، فاصله چشم، خروج از مرکز چشم، اندازه مردمک، کجی ابرو، اندازه بینی، شکل دهان، اندازه دهان و باز و بسته بودن دهان) را نشان می دهد که هر کدام به یکی از ۱۰ مقدار ممکن ، با استفاده از ریاضیات اختصاص یافته است. (S. Dickson)



تصویرسازی داده های پیچیده و روابط

- برای مصورسازی داده‌هایی مثل متن و شبکه های اجتماعی
- ابر برچسب‌ها (Tag cloud)
  - اهمیت لغات با فوت و رنگ متمایز میشود

فصل دوم: داده‌های خود را بشناسید

1) اشیاء داده‌ای و انواع ویژگی‌ها
2) توصیفات آماری داده‌ها
3) اندازه‌گیری میزان شباهت (Similarity) و عدم شباهت (Dissimilarity) داده‌ها
4) مصورسازی داده‌ها (Data Visualization)
5) خلاصه فصل

خلاصه فصل

- انواع ویژگی داده : اسمی، باینری، ترتیبی، بازه - مقیاس، نسبت - مقیاس
- انواع داده ها : به عنوان مثال، عددی، متن، نمودار، وب، تصویر.
- دستیابی به نگرشی در اطلاعات با استفاده از:
  - توصیف داده های آماری پایه : گرایش مرکزی، پراکندگی، نمایشهای گرافیکی
  - تصویر سازی داده ها: ترسیم داده ها بر روی شکلهای هندسی گرافیکی
  - اندازه گیری تشابه داده ها
- مراحل بالا آغازی بر پیش پردازش داده ها هستند .
- بسیاری از این روش ها توسعه یافته اند اما هنوز هم بستری فعال برای پژوهش محسوب میشوند.