

Introdução a Ciências de Dados

Aula 1 parte 2: Ciência de Dados e suas etapas

Francisco A. Rodrigues
ICMC/USP
francisco@icmc.usp.br



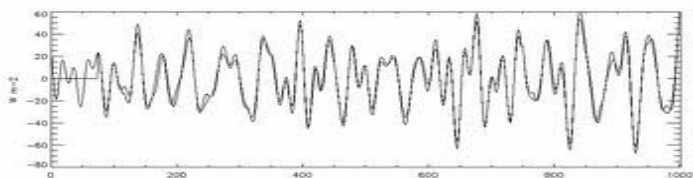
Ciências de Dados e suas Etapas

- Tipos de dados.
- Estatística descritiva.
- Visualização

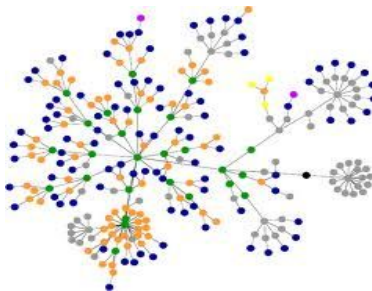
Tipos de Dados

Tipos de Dados

- Dados podem ter diferentes formatos:



Séries temporais



Grafos



Páginas web

Textos

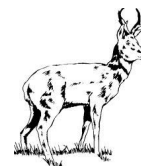
Die auch lebhafteste Partnerstrategie ziemlich auffällige Art ist besonders durch die Neigung der keltischen Bienenstockung zur Auflösung in eine ungeschickte Fleckentzue, sowie das Auftreten einer keltischen Lebensform- und -verhaltens ausgedehnten inneren Bienenstocke sehr ausgezeichnet. Ähnlich wie bei den keltischen D. Gaudin Gitta, stellt sich eine nur starker erdende Lungenzelle, die etwas innerhalb der Schilber beginnt, bis über die Mitte. Im Gegensatz zu dieser Art ist Kopf und Halsbildung in viel grosserer Ausdehnung dicht, teilsweise, da die schwebende Mittellinie und besonders die viel weniger ausgedehnten Seiten-schwebende unter Bienen für die Bildung ihrer Linsen. Diese ist sehr dicht, auf den Halsbildung zwischen den keltischen Schilber und auf den Seiten lebhaft, schwebend oder vorgeht, der stemp Teil des Kopfes, sowie eine schwebende Bewegung der glänzenden Mittellinie des Thorax und der Bienen aussenhalb der Seiten-schwebenden vordringt. Das Grundelement der Flügelbildung ist heller oder dunkler Kaffeebraun, die schwebt abgehoben weissen Bienen der O. G. sehr ungeschickte beginnt, teils in eine Reihe von Fleckchen aufsteigt, wodurch der Gesamtstempel von dem der Bienen, optischen. Daraus mit ihren schwebt beginnenden, hellen Bienen wesentlich abhebt. Von D. Uspoi Perez und Meritoni Perez unterscheidet sich wesentlich, abgesehen von der Zeichnung, durch die bei ihnen beiden Arten ganz fehlenden oder nur ungeschickten Schwebelücken des Halsbildung. Bei D. alternativen Uspoi ist die keltische Halsbildungsteile tief gefurcht und die Marginalbildung der Flügelbildung sehr schwebt, ausserdem fehlt bei dieser Art die Halsbildung.

Sigheer 7 in Cerevis (Korb, 17, 8, 87).

Áudios



Vídeos



Imagens

Geralmente transformados
para o formato
atributo-valor

Tipos de Dados

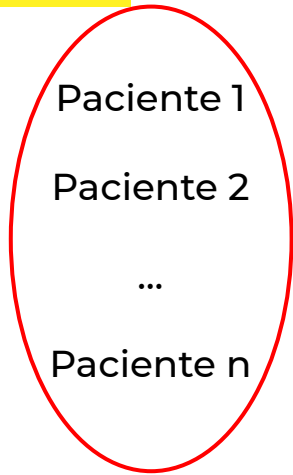
- Representação de conjunto de dados
 - Formados por objetos
 - Cada objeto corresponde a uma ocorrência dos dados

	temperatura (°C)	dor	...	pressão	doente
Paciente 1	38	Sim	...	12.7	Sim
Paciente 2	36	Não	...	12.7	Não
...
Paciente n	40	Não	...	14	Sim

Tipos de Dados

- Representação de conjunto de dados
 - Formados por objetos
 - Cada objeto corresponde a uma ocorrência dos dados

Objetos



	temperatura (°C)	dor	...	pressão	doente
Paciente 1	38	Sim	...	12.7	Sim
Paciente 2	36	Não	...	12.7	Não
...
Paciente n	40	Não	...	14	Sim

Tipos de Dados

- Representação de conjunto de dados
 - Formados por objetos
 - Cada objeto corresponde a uma ocorrência dos dados

Objetos

Atributos

temperatura (°C) dor ... pressão doente

Paciente 1

38

Sim

...

12.7

Sim

Paciente 2

36

Não

...

12.7

Não

...

...

...

...

...

...

Paciente n

40

Não

...

14

Sim

Tipos de Dados

- Representação de conjunto de dados
 - Formados por objetos
 - Cada objeto corresponde a uma ocorrência dos dados

Objetos

Atributos

**Classe
atributo meta**

temperatura (°C) dor ... pressão doente

Paciente 1

38

Sim

...

12.7

Sim

Paciente 2

36

Não

...

12.7

Não

...

...

...

...

...

...

Paciente n

40

Não

...

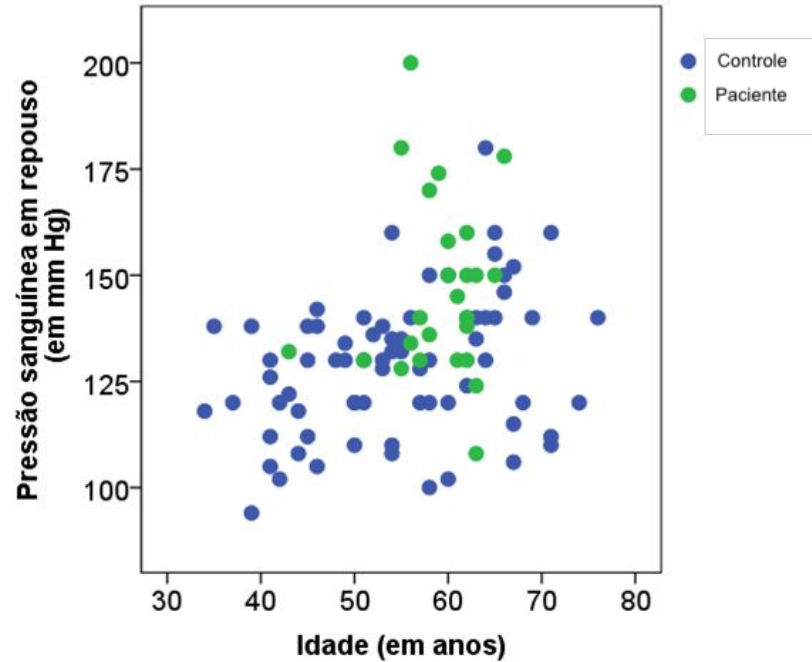
14

Sim

Conjunto de Dados

- Pode ser representado por uma matriz de objetos $\mathbf{X}_{n \times d}$
 - n = número de objetos
 - d = número de atributos (excluindo atributo-meta)
 - Dimensionalidade dos objetos
 - Elemento x_i^j (ou x_{ij}) \Rightarrow valor da j -ésima característica para o objeto i

Conjunto de Dados: Visualização



Análise de dados

- Análise das características de um conjunto de dados
 - Muitas podem ser obtidas por fórmulas estatísticas simples.
 - Estatística descritiva
 - Análise visual também é importante.

Análise de dados

- **Caracterização de dados**
 - Instâncias e Atributos
 - Tipos de Dados
- **Exploração de dados**
 - Dados univariados
 - Medidas de localidade, espalhamento e distribuição
 - Dados multivariados
 - Visualização

Análise dos dados

- Valores de atributos podem ser definidos por:
 - **Tipo**
 - Grau de quantização nos dados
 - **Escala**
 - Significância relativa dos valores

Conhecer o tipo/escala dos atributos auxilia a identificar a forma adequada de preparar os dados e posteriormente modelá-los

Tipos de Atributos

Quantitativo (numérico)

Representa quantidades.

Valores podem ser ordenados e usados em operações aritméticas.

Podem ser **contínuos** ou **discretos**.

Possuem unidade associada.

Qualitativo (simbólico ou categórico)

Representa qualidades.

Valores podem ser associados a categorias.

Alguns podem ser ordenados, mas operações aritméticas não são aplicáveis.

Ex. {pequeno, médio, grande}

Tipos de Atributos

Atributos Quantitativos

Contínuos

- Podem assumir um número infinito de valores.
- Geralmente resultados de medidas.
- Frequentemente representados por números reais
- *Ex. peso, distância.*

Discretos

- Número finito ou infinito contável de valores.
- Caso especial: atributos binários (booleanos).
- *Ex. {12, 23, 45}, {0, 1}*

Tipos de atributos

Identificação	Nome	Estado	Idade	Sexo	Nível de Glicose	Febre	Dores no corpo	Diagnóstico
100	João	SP	22	M	90,5	Baixa	Não	Saudável
102	Maria	SP	39	F	86,5	Baixa	Não	Saudável
230	Rubens	RJ	23	M	120,2	Média	Sim	Doente
40	João	RS	21	M	85,9	Alta	Não	Doente
543	Marta	SP	40	F	88,7	Alta	Sim	Doente
12	Ana	AM	32	F	78,8	Média	Não	Saudável
130	Carlos	SP	29	M	110	Alta	Sim	Doente
123	José	RJ	30	M	80,1	Baixa	Sim	Saudável

Qualitativo

Quantitativo discreto

Quantitativo contínuo

Tipos de atributos

Identificação	Nome	Estado	Idade	Sexo	Nível de Glicose	Febre	Dores no corpo	Diagnóstico
100	João	SP	22	M	90,5	Baixa	Não	Saudável
102	Maria	SP	39	F	86,5	Baixa	Não	Saudável
230	Rubens	RJ	23	M	120,2	Média	Sim	Doente
40	João	RS	21	M	85,9	Alta	Não	Doente
543	Marta	SP	40	F	88,7	Alta	Sim	Doente
12	Ana	AM	32	F	78,8	Média	Não	Saudável
130	Carlos	SP	29	M	110	Alta	Sim	Doente
123	José	RJ	30	M	80,1	Baixa	Sim	Saudável

Alguns atributos qualitativos são representados por números, mas não faz sentido a utilização de operadores aritméticos sobre seus valores.

Qualitativo

Quantitativo discreto

Quantitativo contínuo

Escala de atributos

- Define operações que podem ser realizadas sobre os valores dos atributos

– Nominiais

– Ordinais

Qualitativos

– Intervalares

– Racionais

Quantitativos

Escala de atributos

Escala nominal

- Valores são nomes diferentes e carregam a menor quantidade de informação possível
- Não existe relação de ordem entre os valores
- **Operações aplicáveis:** $=$, \neq
- *Ex.: número de conta em banco, cores, sexo*

Escala ordinal

- Valores refletem ordem das categorias representadas
- **Operações aplicáveis:** $=$, \neq , $<$, $>$, \leq , \geq
- *Ex.: hierarquia militar, avaliações qualitativas de temperatura*

Escala de atributos

Escala intervalar

- Números que variam em um intervalo
- É possível definir ordem e diferença em magnitude entre dois valores
- Origem da escala definida de maneira arbitrária
- **Operações aplicáveis:** $=, \neq, <, >, \leq, \geq, +, -$
- *Ex.: temperatura em $^{\circ}\text{C}$ ou $^{\circ}\text{F}$, datas*

Escala racional

- Carregam mais informações
- Têm significado absoluto (existe 0 absoluto)
- Razão tem significado
- **Operações aplicáveis:** $=, \neq, <, >, \leq, \geq, +, -, *, /$
- *Ex.: tamanho, distância, salário, saldo em conta*

Tipos de atributos

Identificação	Nome	Estado	Idade	Sexo	Nível de Glicose	Febre	Dores no corpo	Diagnóstico
100	João	SP	22	M	90,5	Baixa	Não	Saudável
102	Maria	SP	39	F	86,5	Baixa	Não	Saudável
230	Rubens	RJ	23	M	120,2	Baixa	Sim	Doente
40	João	RS	21	M	85,9	Alta	Não	Doente
543	Marta	SP	40	F	88,7	Alta	Sim	Doente
12	Ana	AM	32	F	78,8	Média	Não	Saudável
130	Carlos	SP	29	M	110	Média	Sim	Doente
123	José	RJ	30	M	80,1	Baixa	Sim	Saudável

Nominal

Ordinal

Intervalar

Racional

Estatística Descritiva

Exploração dos dados

- **Estatística descritiva:** resumo quantitativo das principais características de um conjunto de dados
 - Muitas medidas podem ser calculadas rapidamente
 - Captura de informações como:
 - Frequência
 - Localização ou tendência central
 - Dispersão ou espalhamento
 - Distribuição ou formato

Frequência

- Proporção de vezes que um atributo assume um dado valor.
- Aplicável a valores numéricos e simbólicos.

Identificação	Nome	Estado	Idade	Sexo	Nível de Glicose	Febre	Dores no corpo	Diagnóstico
100	João	SP	22	M	90,5	Baixa	Não	Saudável
102	Maria	SP	39	F	86,5	Baixa	Não	Saudável
230	Rubens	RJ	23	M	120,2	Baixa	Sim	Doente
40	João	RS	21	M	85,9	Alta	Não	Doente
543	Marta	SP	40	F	88,7	Alta	Sim	Doente
12	Ana	AM	32	F	78,8	Média	Não	Saudável
130	Carlos	SP	29	M	110	Média	Sim	Doente
123	José	RJ	30	M	80,1	Baixa	Sim	Saudável

62,5% dos pacientes são homens

Moda

- Retorna o valor mais comum. Geralmente usada com dados nominais.

Identificação	Nome	Estado	Idade	Sexo	Nível de Glicose	Febre	Dores no corpo	Diagnóstico
100	João	SP	22	M	90,5	Baixa	Não	Saudável
102	Maria	SP	39	F	86,5	Baixa	Não	Saudável
230	Rubens	RJ	23	M	120,2	Baixa	Sim	Doente
40	João	RS	21	M	85,9	Alta	Não	Doente
543	Marta	SP	40	F	88,7	Alta	Sim	Doente
12	Ana	AM	32	F	78,8	Média	Não	Saudável
130	Carlos	SP	29	M	110	Média	Sim	Doente
123	José	RJ	30	M	80,1	Baixa	Sim	Saudável

Moda: SP

Moda: M

Moda: Baixa

Média

$$\bar{X} = \sum_{i=1}^N \frac{X_i}{N}$$

Problema: sensível a outliers

Bom indicador apenas se valores são distribuídos simetricamente

Identificação	Nome	Estado	Idade	Sexo	Nível de Glicose	Febre	Dores no corpo	Diagnóstico
100	João	SP	22	M	90,5	Baixa	Não	Saudável
102	Maria	SP	39	F	86,5	Baixa	Não	Saudável
230	Rubens	RJ	23	M	120,2	Baixa	Sim	Doente
40	João	RS	21	M	85,9	Alta	Não	Doente
543	Marta	SP	40	F	88,7	Alta	Sim	Doente
12	Ana	AM	32	F	78,8	Média	Não	Saudável
130	Carlos	SP	29	M	110	Média	Sim	Doente
123	José	RJ	30	M	80,1	Baixa	Sim	Saudável

Média: 29,5

Média: 92,6

Mediana

Passos:

- Ordenar os valores de forma crescente
- Calcular a equação:

$$\text{mediana}(\mathbf{x}) = \begin{cases} \frac{1}{2} (x_r + x_{r+1}) & \text{se } n \text{ for par } (n = 2r) \\ x_{r+1} & \text{se } n \text{ for ímpar } (n = 2r + 1) \end{cases}$$

Média: 29,5
Mediana: 29,5

Média: 92,6
Mediana: 87,6

Identificação	Nome	Estado	Idade	Sexo	Nível de Glicose	Febre	Dores no corpo	Diagnóstico
100	João	SP	22	M	90,5	Baixa	Não	Saudável
102	Maria	SP	39	F	86,5	Baixa	Não	Saudável
230	Rubens	RJ	23	M	120,2	Baixa	Sim	Doente
40	João	RS	21	M	85,9	Alta	Não	Doente
543	Marta	SP	40	F	88,7	Alta	Sim	Doente
12	Ana	AM	32	F	78,8	Média	Não	Saudável
130	Carlos	SP	29	M	110	Média	Sim	Doente
123	José	RJ	30	M	80,1	Baixa	Sim	Saudável

Mediana

- Exemplos:
 - {17, 4, 8, 21, 4}
 - **Ordenando:** 4, 4, 8, 17, 21
 - Número ímpar de elementos \Rightarrow **mediana** = 8
 - Valor do meio na ordenação
 - {17, 4, 8, 21, 4, 15, 13, 9}
 - **Ordenando:** 4, 4, 8, 9, 13, 15, 17, 21
 - Número par de elementos \Rightarrow **mediana** = $(9+13)/2 = 11$
 - Média dos dois valores do meio na ordenação

Quartil e percentil

Quartis

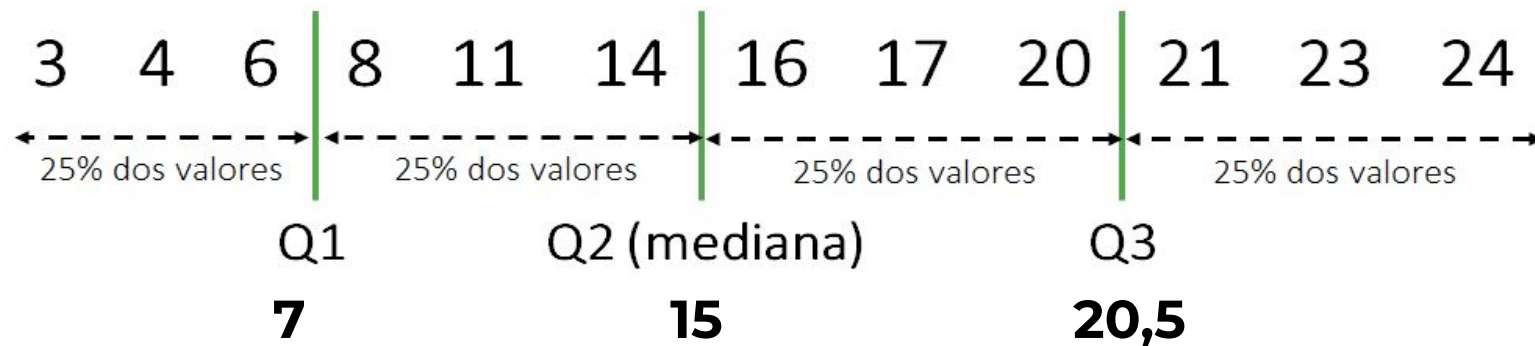
- Divide em quartos
- 1º quartil (Q1) ⇨ valor que tem 25% dos demais valores abaixo dele
- 2º quartil = mediana

Percentil

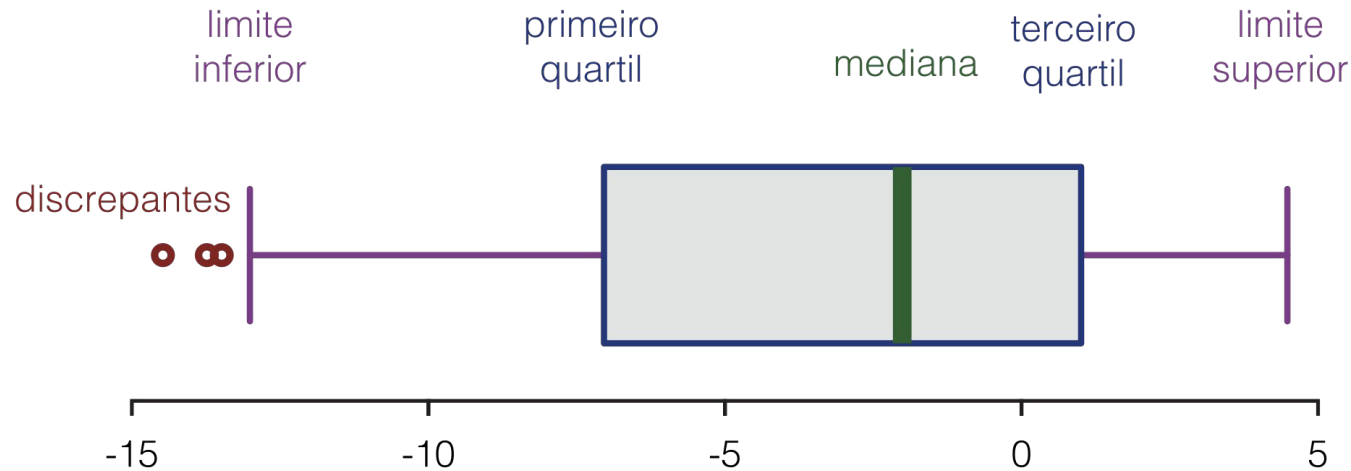
- Para p entre 0 e 100
- p percentil = Pp ⇨ x_i tal que $p\%$ dos valores observados são menores do que x_i
- $P25 = Q1$
- $P50 = Q2 = \text{mediana}$

Quartil e percentil

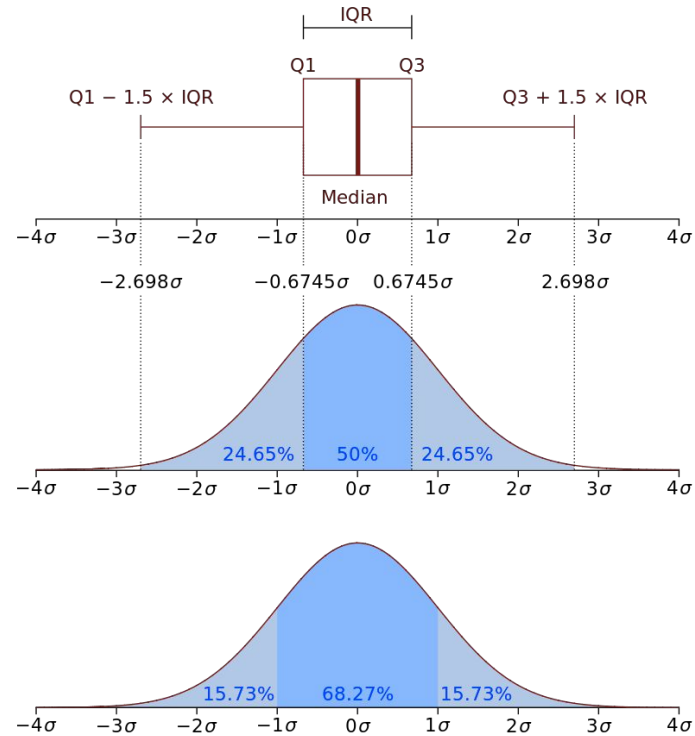
Dados: {8,3,11,14,20,4,17,6,17,24,21,23}



Boxplots



Boxplots



Medidas de dispersão

- Medem **dispersão** ou **espalhamento** de um conjunto de valores
 - Permitem observar se valores estão:
 - Espalhados
 - Concentrados em torno de um valor (ex. da média)
 - Medidas mais comuns:
 - Intervalo
 - Variância
 - Desvio padrão

Medidas de dispersão

Variância

$$\sigma^2 = E[(X - E(X))^2] = E[X^2] - E[X]^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

Desvio padrão

População

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

Amostra

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

Estimador não-viesado

Medidas de dispersão

Exemplo: Seja uma amostra do número de erros em um servidor na última hora: $X = \{6, 2, 3, 1\}$

$$\mu = \frac{6 + 2 + 3 + 1}{4} = 3$$

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

$$\sigma = \sqrt{\frac{1}{3} [(6-3)^2 + (2-3)^2 + (3-3)^2 + (1-3)^2]} = \sqrt{\frac{9+1+0+4}{3}} = 2,16$$

Medidas de correlação

Coeficiente de Pearson:

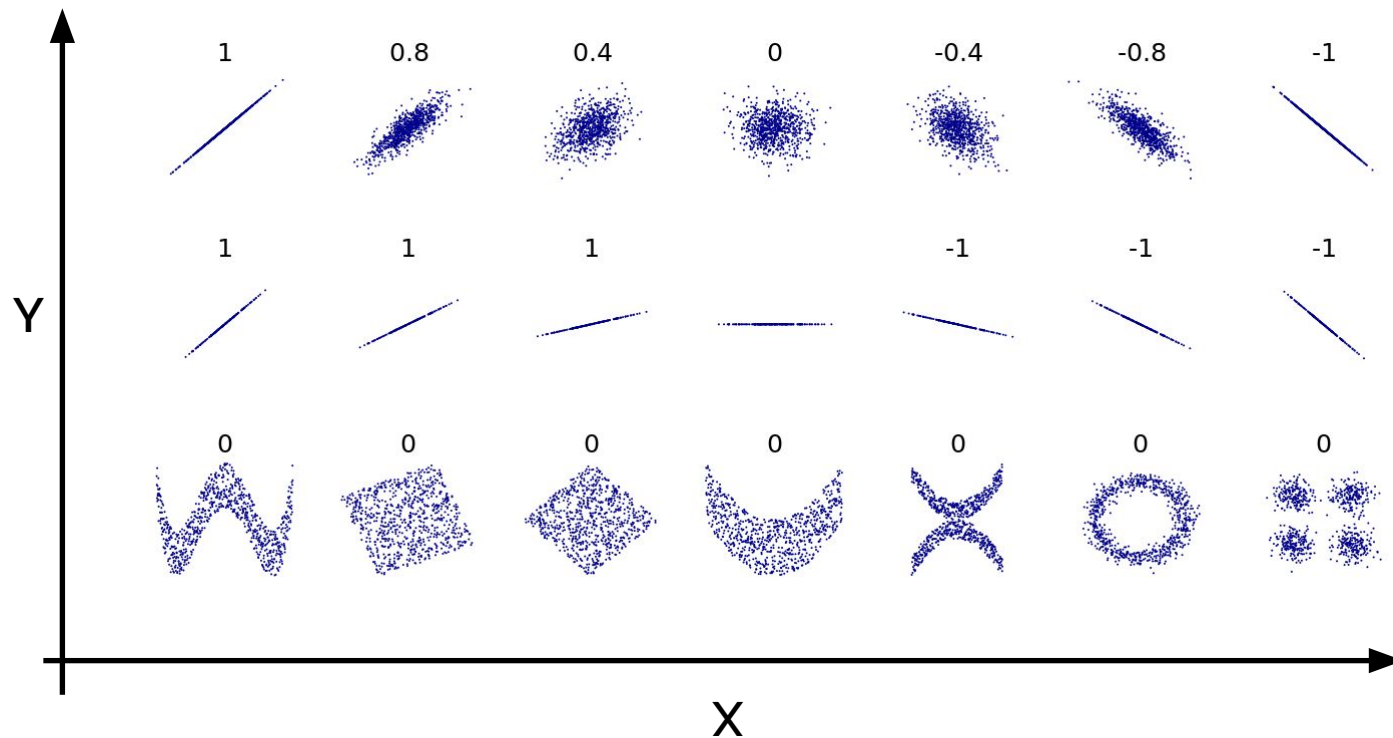
- Para duas variáveis aleatórias X e Y:

$$\rho_{X,Y} = \frac{E[XY] - E[X]E[Y]}{\sqrt{E[X^2] - [E[X]]^2} \sqrt{E[Y^2] - [E[Y]]^2}} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

- Para uma amostra:

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}}$$

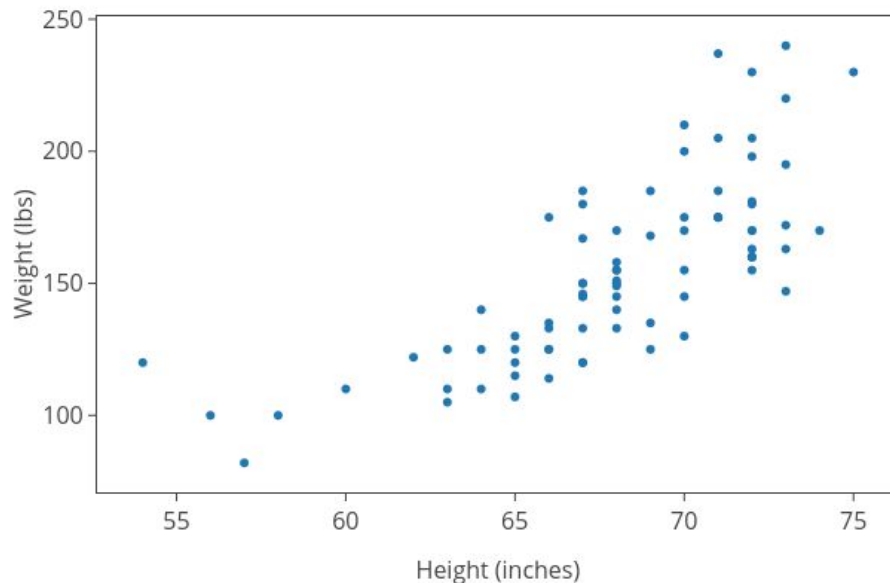
Correlação de Pearson



Correlação de Pearson

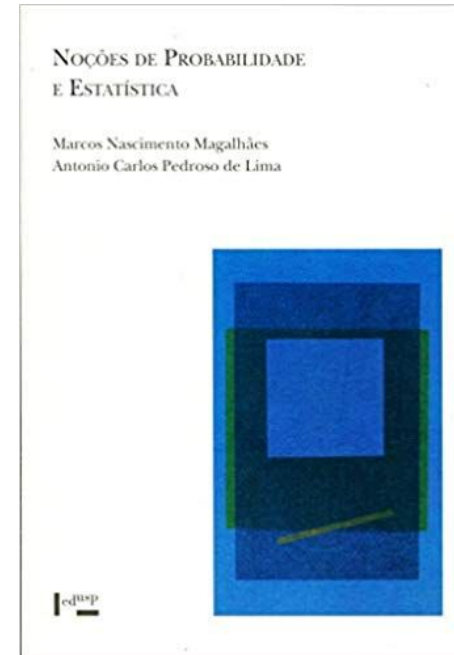
Interpretação:

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$



Sugestão de leitura

Noções de Probabilidade e Estatística,
Marcos N. Magalhães e Antonio C. P. De
Lima, Edusp.



Visualização

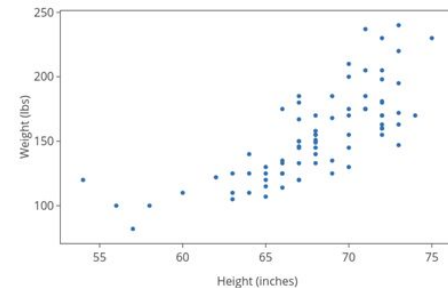
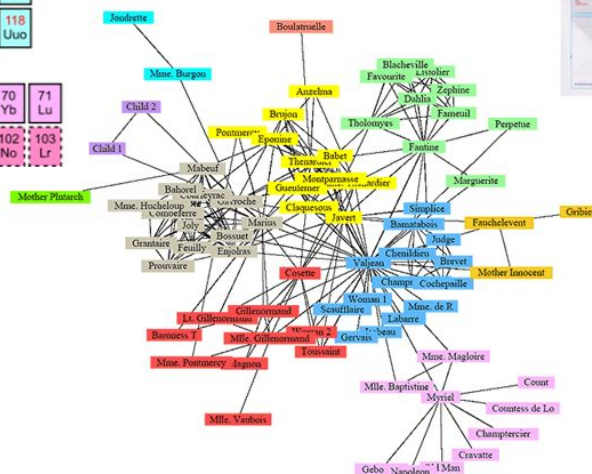
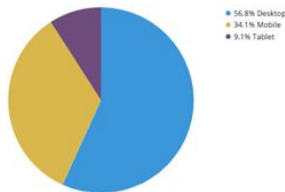
Visualização

Periodic Table visualization showing elements grouped by Period (I to VIII) and Group (I to VIII).

Period	I	II	III	IV	V	VI	VII	VIII
1	1 H							2 He
2	3 Li	4 Be						10 Ne
3	11 Na	12 Mg						18 Ar
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru
6	55 Cs	56 Ba	*	72 Hf	73 Ta	74 W	75 Re	76 Os
7	87 Fr	88 Ra	**	104 Rf	105 Db	106 Sg	107 Bh	108 Hs
8	119 Uun							

* Lanthanides
** Actinides

Number of Visits by Device



Visualização

Depende do tipo de dados:

Quantitativo (numérico)

Representa quantidades

Valores podem ser ordenados e usados em operações aritméticas

Podem ser **contínuos** ou **discretos**

Possuem unidade associada

Qualitativo (simbólico ou categórico)

Representa qualidades

Valores podem ser associados a categorias

Alguns podem ser ordenados, mas operações aritméticas não são aplicáveis

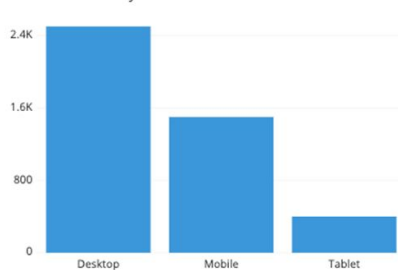
Ex. [pequeno, médio, grande]

Visualização

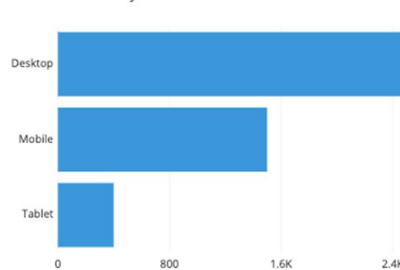
Dados qualitativos

Gráficos de barra

Number of Visits by Device

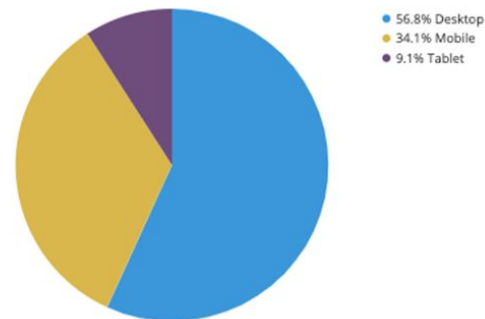


Number of Visits by Device



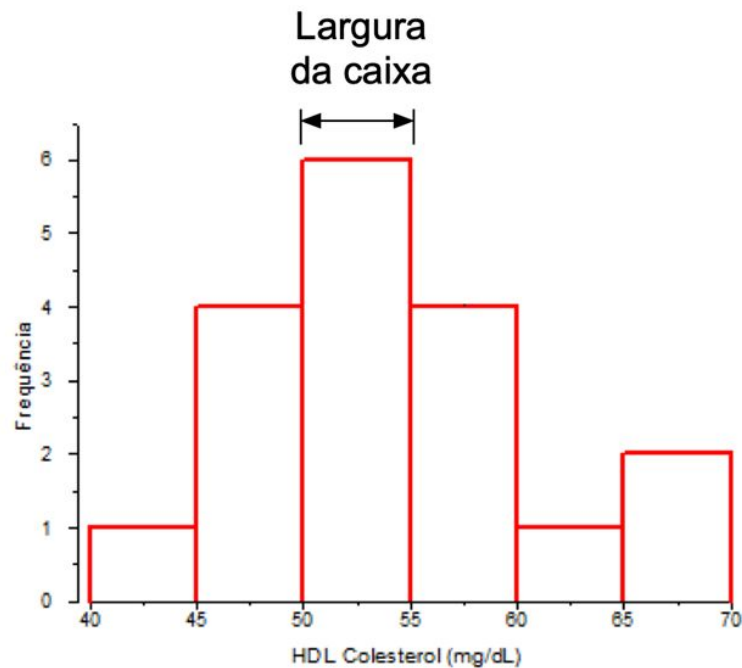
Gráficos de setores

Number of Visits by Device



Visualização

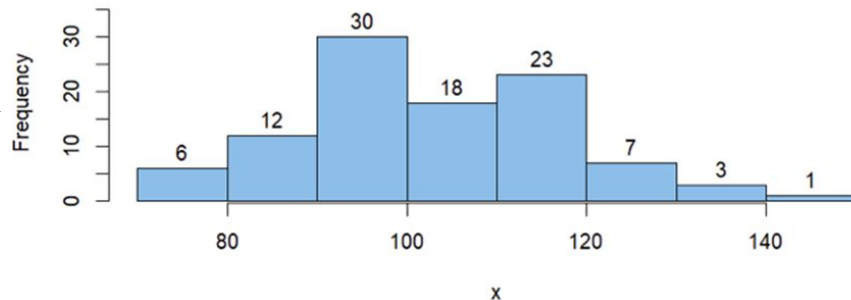
Dados quantitativos: histograma



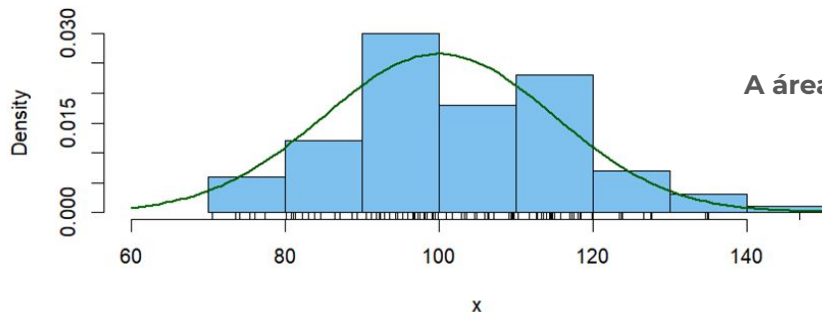
Visualização

Histograma de frequência e densidade

100 observações de uma distribuição Normal(100, 15)



Notem a
diferença no eixo y!

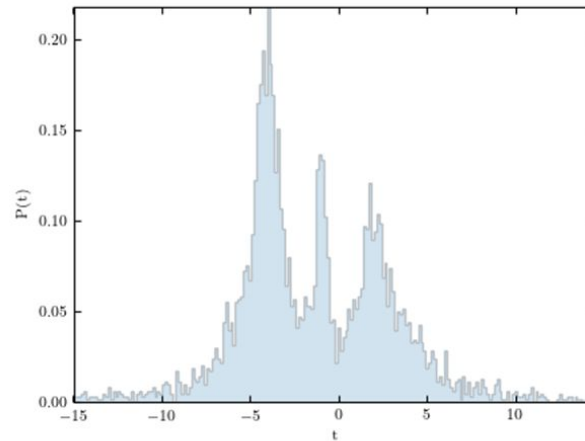
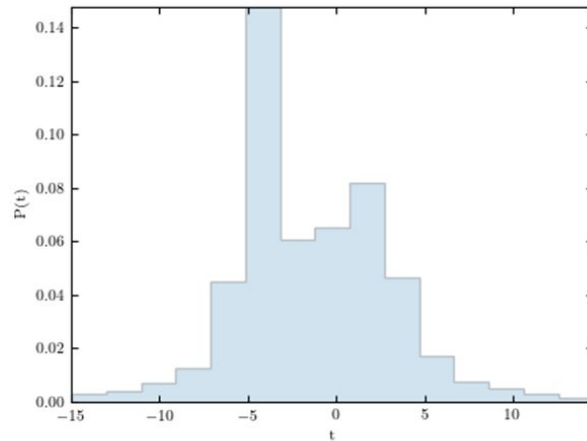


A área do histograma é
igual a 1!

Visualização

Histograma:

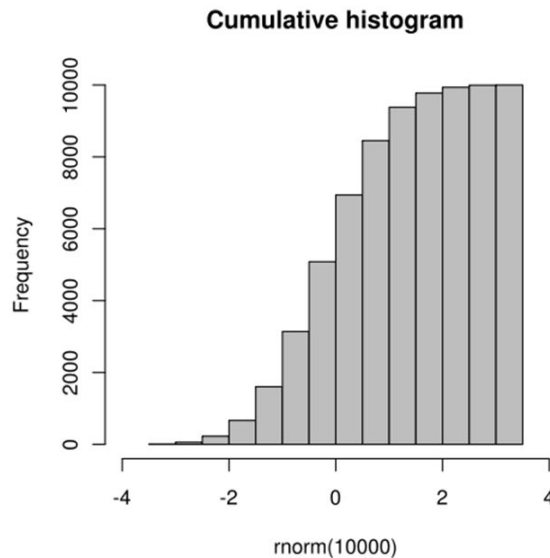
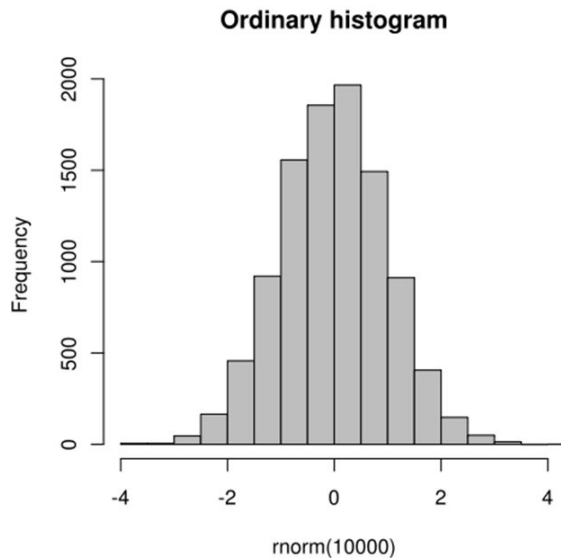
- Não há um método geral para definir o tamanho da caixa.
- Deve-se escolher de modo a não perder informação.



Visualização

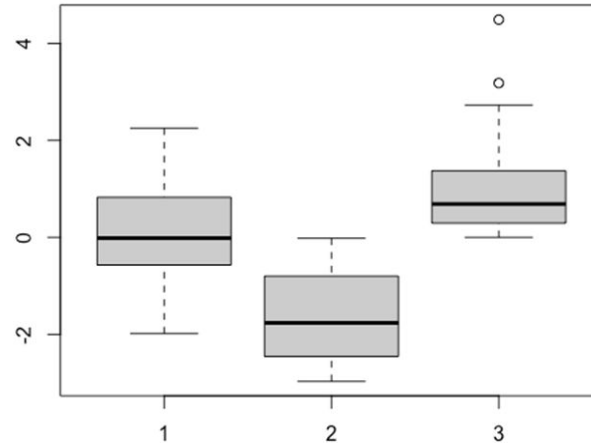
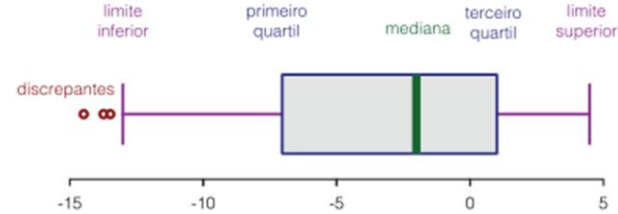
Histograma acumulado

Histograma acumulado: soma-se as frequências até um dado valor.



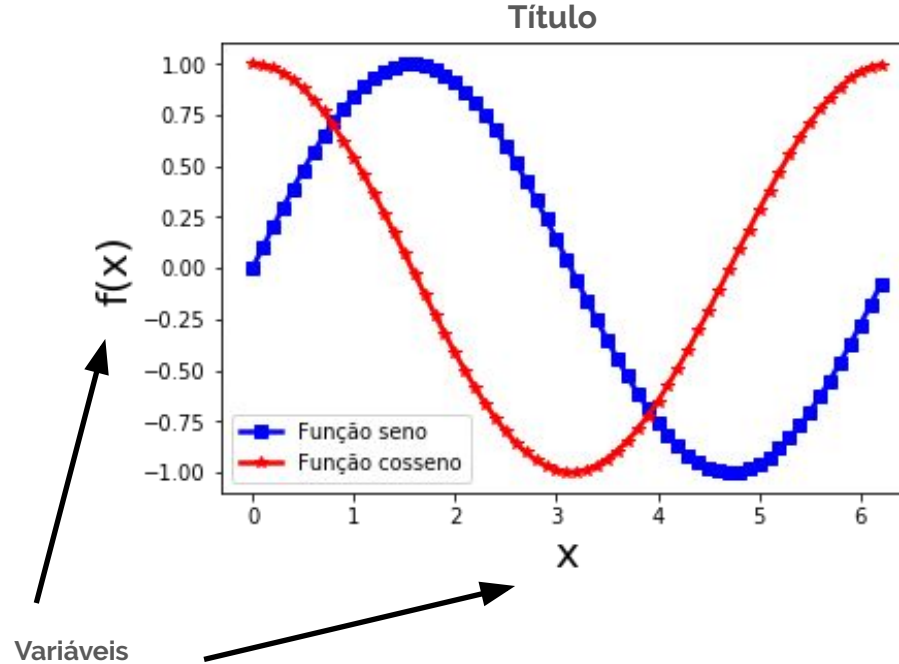
Visualização

Box plot



Visualização

Gráficos

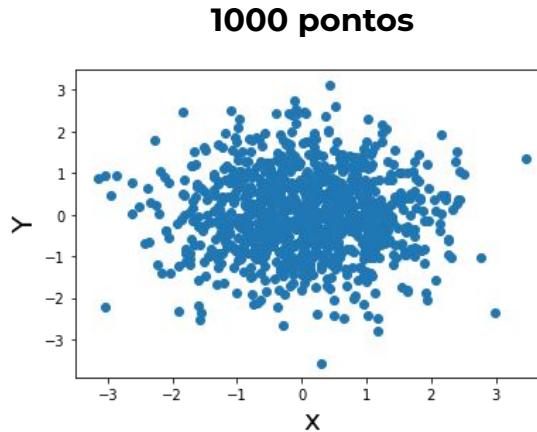
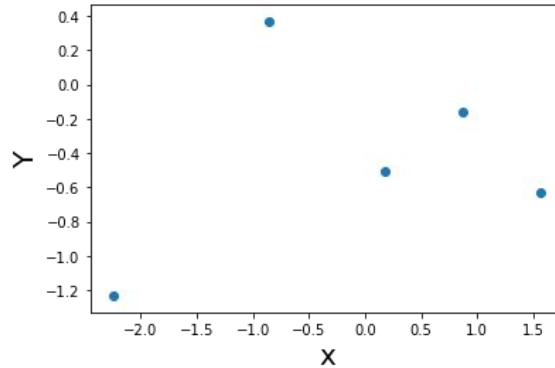


Observação: Sempre tente criar um gráfico que seja claro, com legendas, variáveis e eixos visíveis. Cuidado com as cores!

Visualização

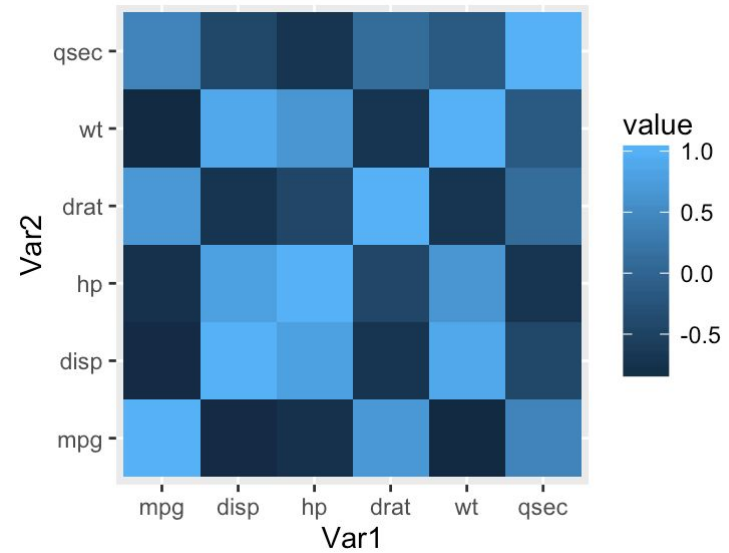
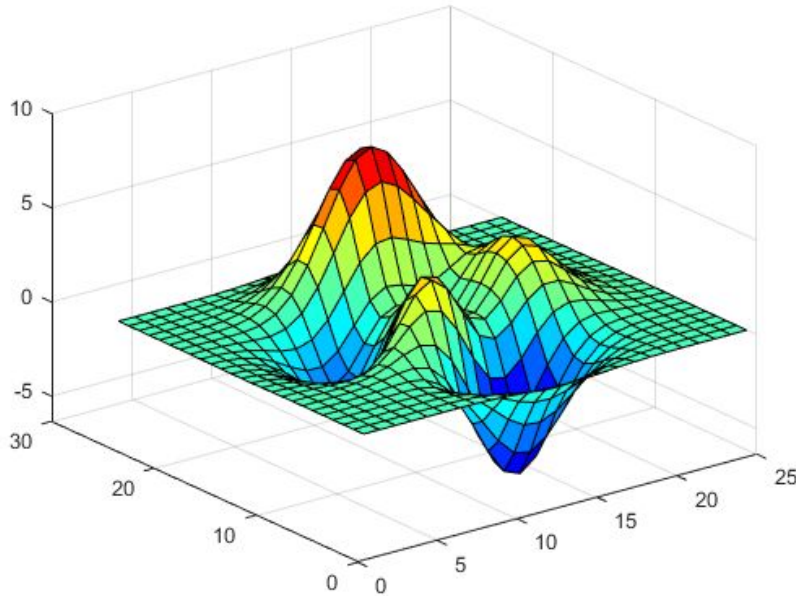
Gráfico de dispersão (scatterplot)

```
X = [-2.25230219  0.17488531  0.86807664  1.57008988 -0.8529788 ]  
Y = [-1.23335724 -0.50329834 -0.16286028 -0.63108853  0.36720953]
```



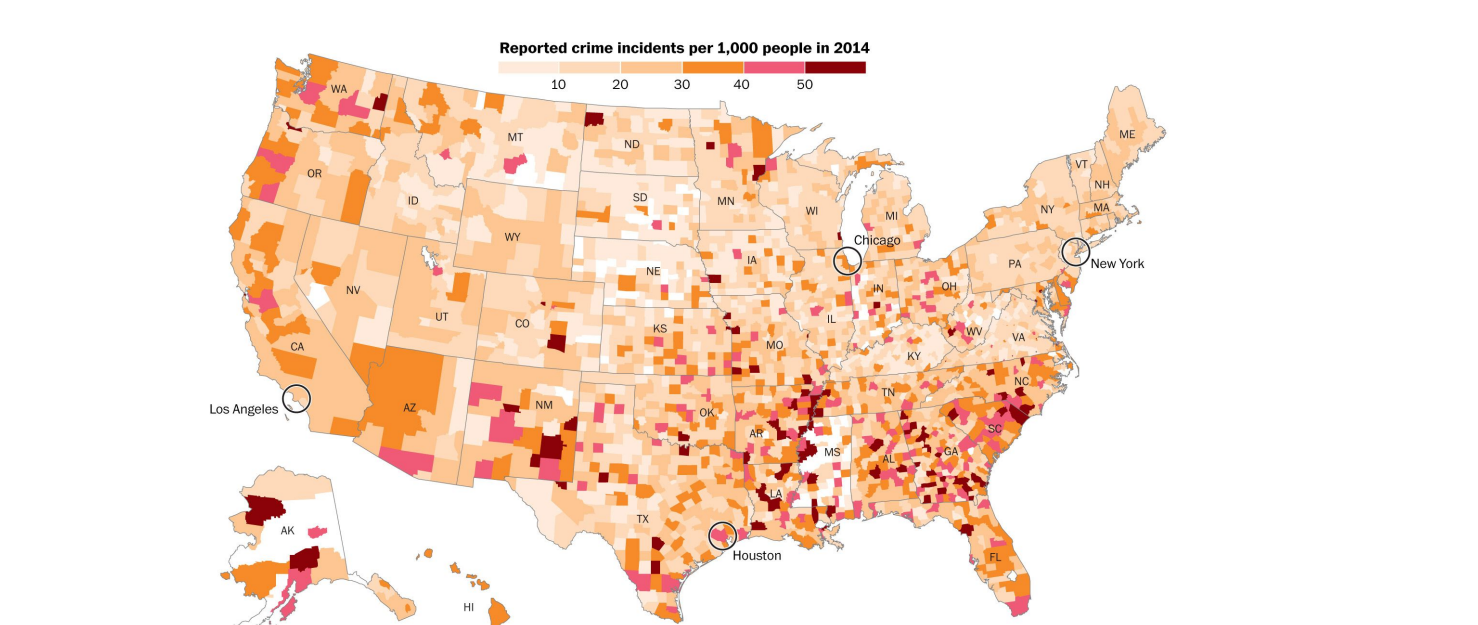
Visualização

Três variáveis



Visualização

Gráficos espaciais:



Reported crime incidents per 1,000 people in 2014

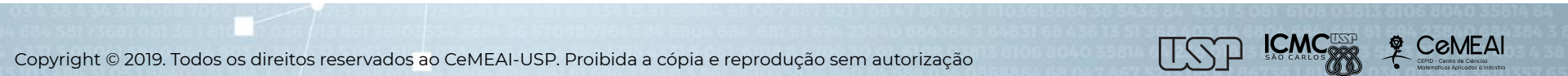
10 20 30 40 50

Los Angeles

Chicago

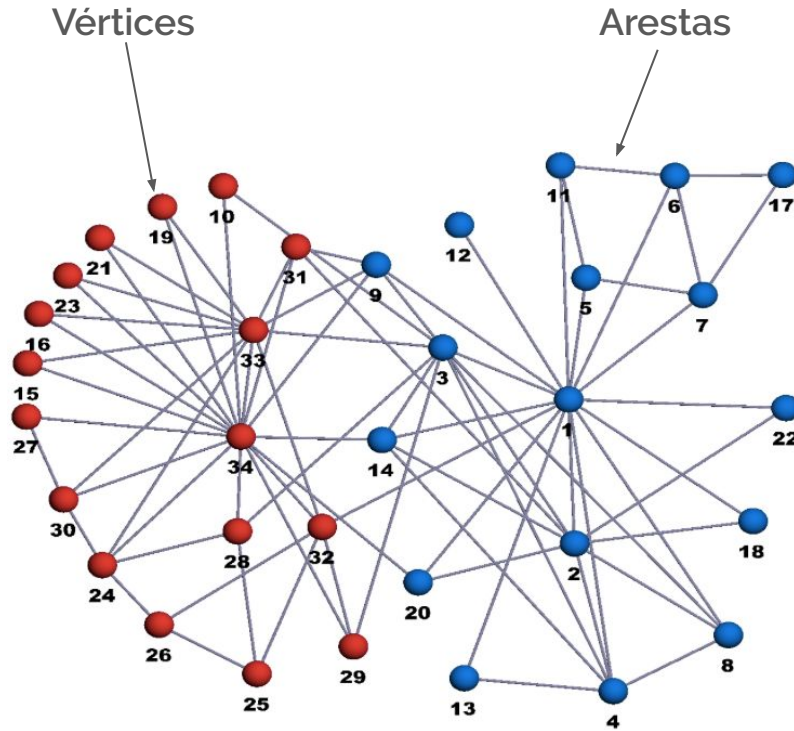
New York

Houston



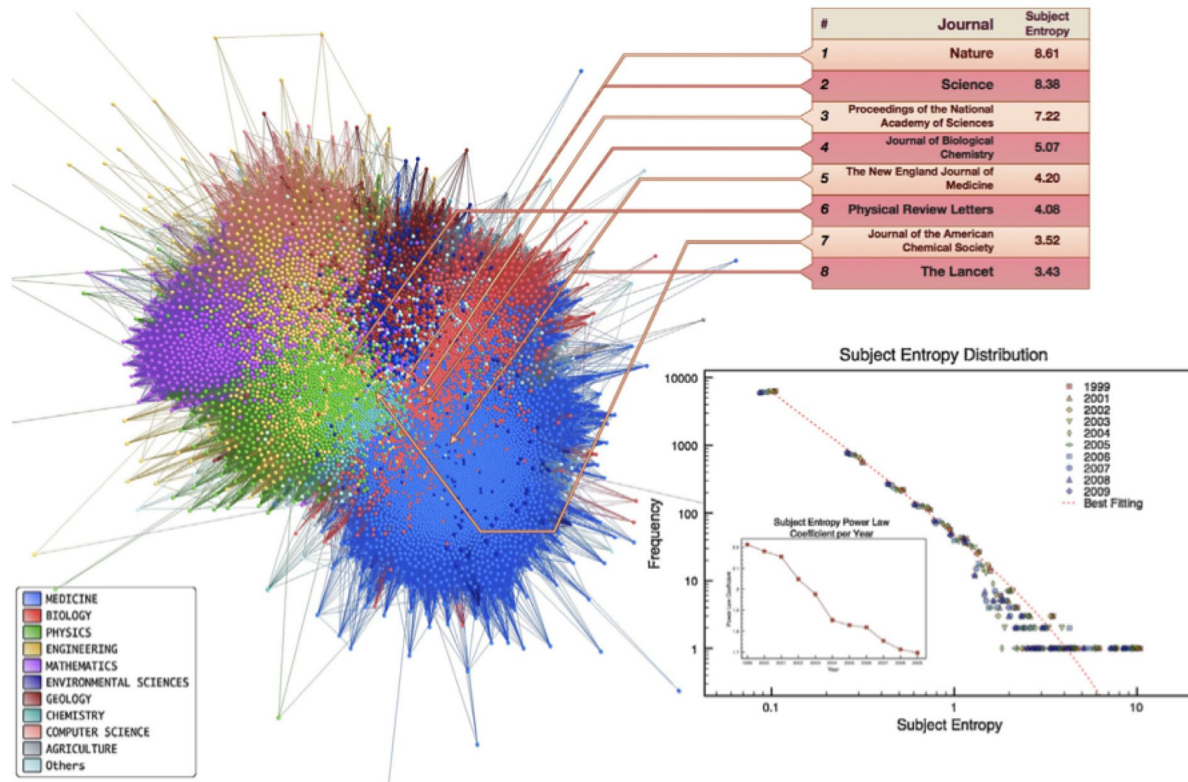
Visualização

Grafos:



Visualização

Grafos:



Sumário

Ciência de dados e suas etapas

- Tipos de dados.
- Estatística descritiva.
- Visualização

Leitura Complementar

- Chen, Härdle, Unwin, **Handbook of Data Visualization**, Springer.