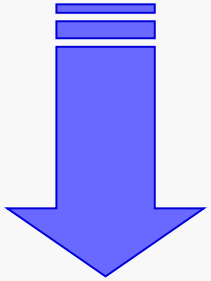
Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina

Análise de Dados

Dados

- Avanços recentes nas tecnologias de aquisição, transmissão e armazenamento de dados



Bases de dados cada vez maiores

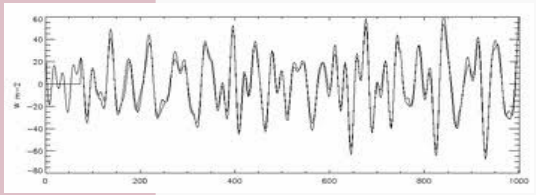
Dados

- Estima-se que a quantidade de dados em Bases de Dados mundiais dobra a cada 20 anos

- Crescimento tem ocorrido em várias áreas
  - *Transações bancárias*
  - *Utilização de cartões de crédito*
  - *Dados governamentais*
  - *Medições ambientais*
  - *Dados clínicos*
  - *Projetos genoma*
  - *Informações disponíveis na web*
  - *etc.*

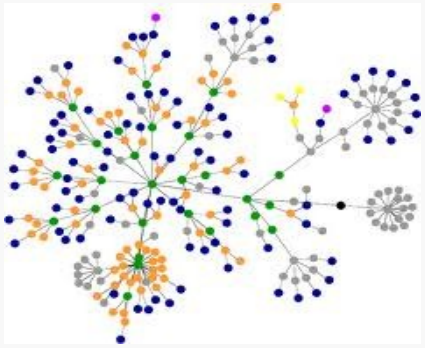
Dados

- Podem ter diferentes formatos



Séries temporais

Textos

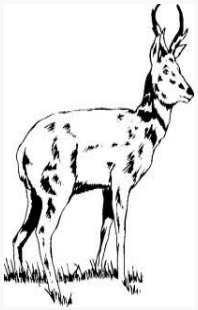


Grafos

Vídeos

Páginas web

Áudios

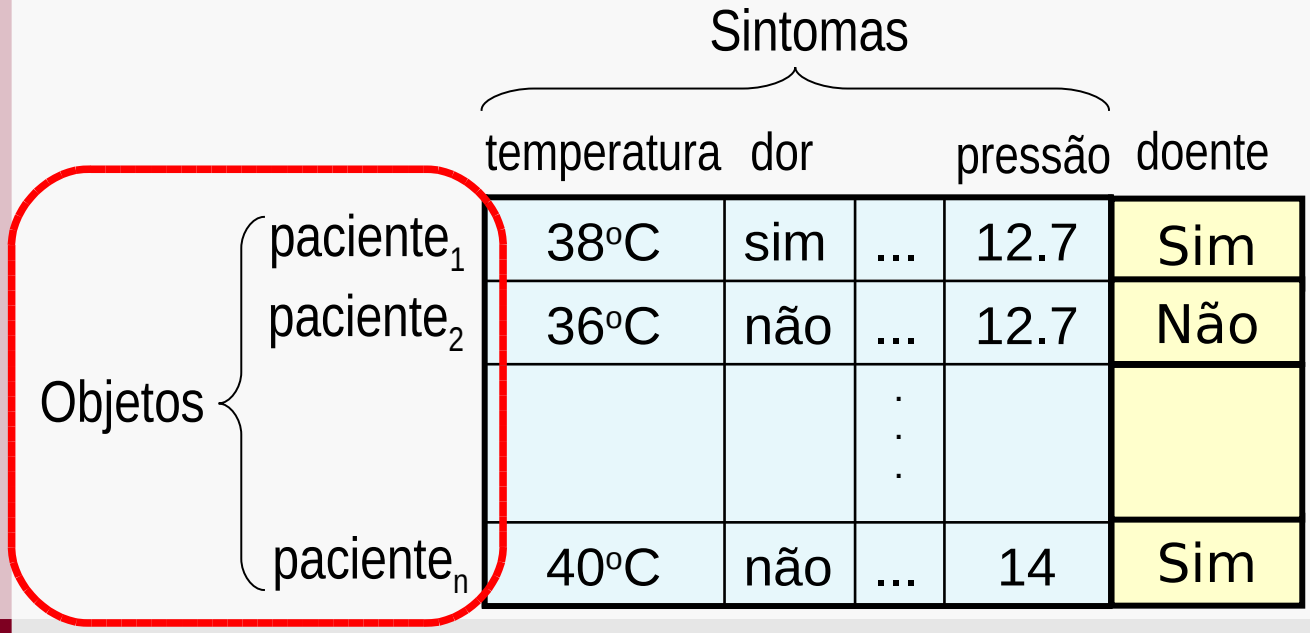


Imagens

Muitas vezes transformados para o formato atributo-valor

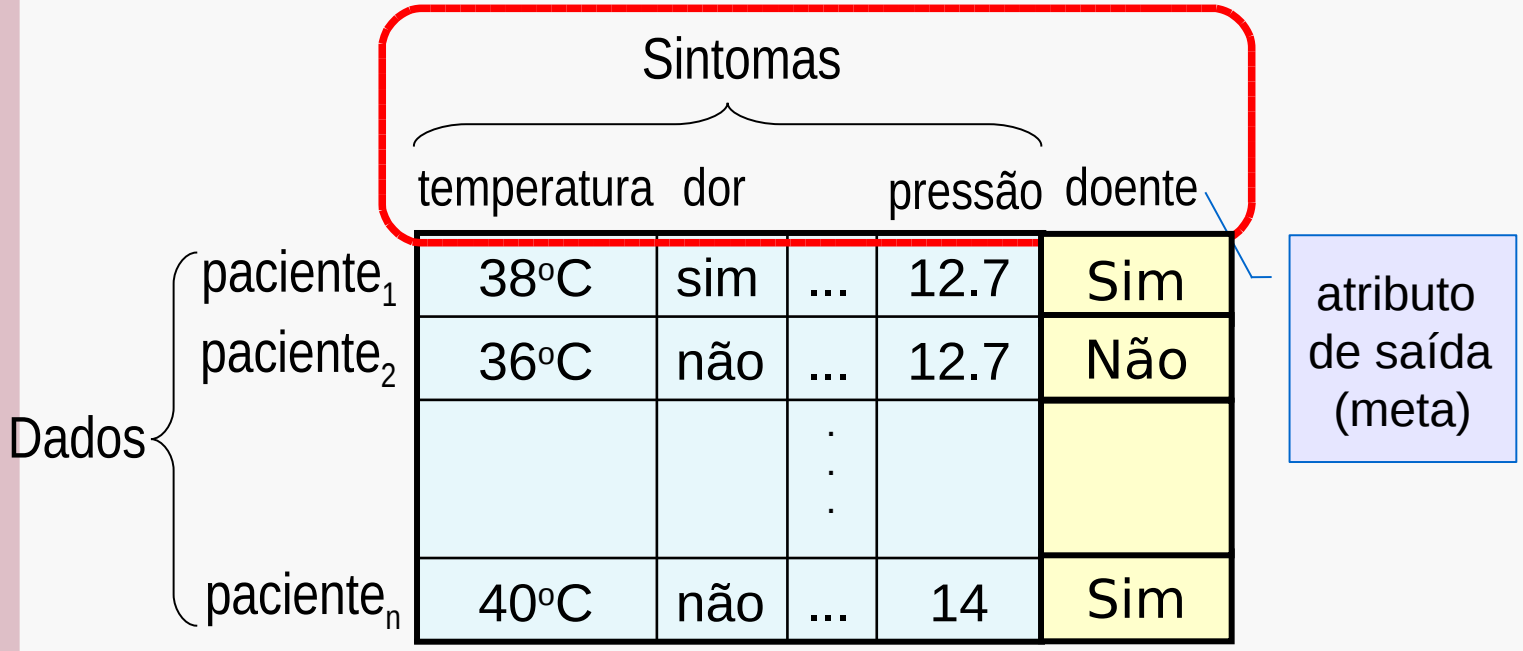
Formato atributo-valor

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



Formato atributo-valor

- Cada objeto é descrito por um conjunto de atributos de entrada (Vetor de características)
  - *Cada atributo está associado a uma propriedade do objeto*

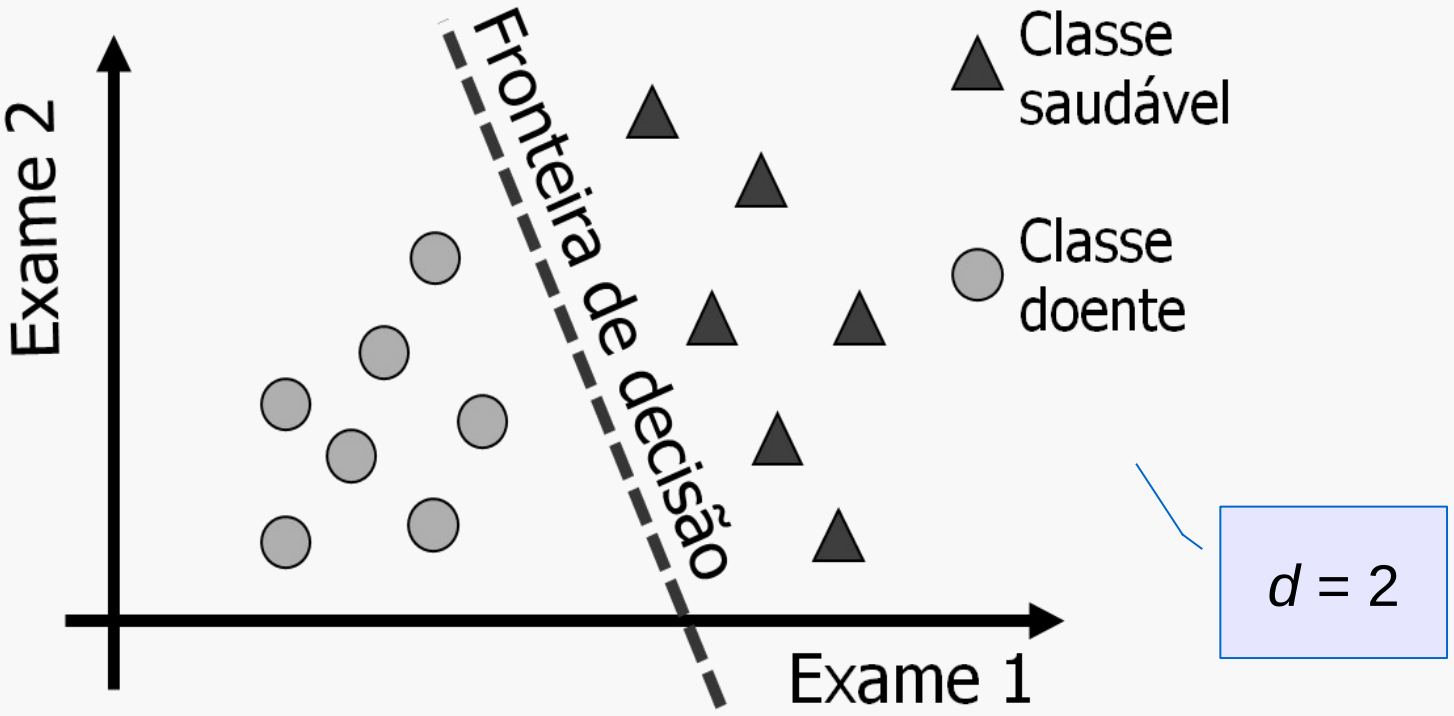


Conjunto de dados

- Pode ser representado por uma matriz de objetos $X_{n \times d}$
  - *n = número de objetos*
  - *d = número de atributos (excluindo atributo-meta)*
    - Dimensionalidade dos objetos
      - Do espaço de objetos (de entradas/de atributos)
  - *Elemento x_j (ou x_{ij}) ⇨ valor da j-ésima característica para o objeto i*

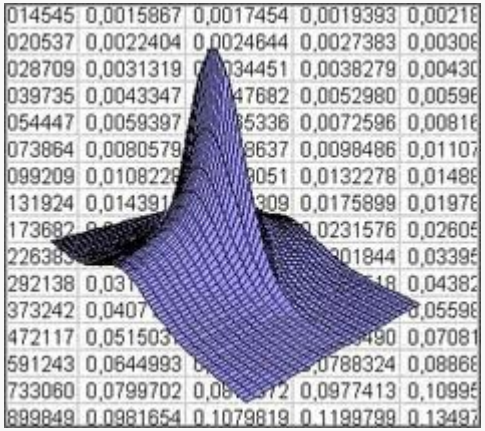
Conjunto de dados: visualização gráfica

- Supor conjunto de pacientes com dois exames



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 de 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

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Qualitativo **Quantitativo discreto** **Quantitativo contínuo**

Tipos de atributos

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

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

Escala de atributos

- Define operações que podem ser realizadas sobre os valores dos atributos
  - *Nominais*
  - *Ordinais*
  - *Intervalares*
  - *Racionais*

Escala de atributos

- Define operações que podem ser realizadas sobre os valores dos atributos
  - *Nominais*
  - *Ordinais*
  - *Intervalares*
  - *Racionais*

Qualitativos

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: =, ≠
- *Ex.: número de conta em banco, cores, sexo*

Escala ordinal

- Valores refletem ordem das categorias representadas
- Operações aplicáveis: =, ≠, <, >, ≤, ≥
- *Ex.: hierarquia militar, avaliações qualitativas de temperatura*

Escala de atributos

- Define operações que podem ser realizadas sobre os valores dos atributos
  - *Nominais*
  - *Ordinais*
  - *Intervalares*
  - *Racionais*

Quantitativos

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: =, ≠, <, >, ≤, ≥, +, -
- *Ex.: temperatura em ^{o}C ou ^{o}F, datas*

Escala racional

- Carregam mais informações
- Têm significado absoluto (existe 0 absoluto)
- Razão tem significado
- Operações aplicáveis: =, ≠, <, >, ≤, ≥, +, -, ∗, /
- *Ex.: tamanho, distância, salário, saldo em conta*

Escalas de atributos

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Nominal **Ordinal** **Intervalar** **Racional**

Exercício

- Definir o tipo e escala dos seguintes atributos:
  - *Renda mensal: ?*
  - *Número de palavras de um texto: ?*
  - *Número de matrícula: ?*
  - *Data de nascimento: ?*
  - *Código postal: ?*
  - *Posição em uma corrida: ?*

Exercício

- Definir o tipo e escala dos seguintes atributos:
  - *Renda mensal: quantitativo racional*
  - *Número de palavras de um texto: ?*
  - *Número de matrícula: ?*
  - *Data de nascimento: ?*
  - *Código postal: ?*
  - *Posição em uma corrida: ?*

Exercício

- Definir o tipo e escala dos seguintes atributos:
  - *Renda mensal: quantitativo racional*
  - *Número de palavras de um texto: quantitativo racional*
  - *Número de matrícula: ?*
  - *Data de nascimento: ?*
  - *Código postal: ?*
  - *Posição em uma corrida: ?*

Exercício

- Definir o tipo e escala dos seguintes atributos:
  - *Renda mensal: quantitativo racional*
  - *Número de palavras de um texto: quantitativo racional*
  - *Número de matrícula: qualitativo nominal*
  - *Data de nascimento: ?*
  - *Código postal: ?*
  - *Posição em uma corrida: ?*

Exercício

- Definir o tipo e escala dos seguintes atributos:
  - *Renda mensal: quantitativo racional*
  - *Número de palavras de um texto: quantitativo racional*
  - *Número de matrícula: qualitativo nominal*
  - *Data de nascimento: quantitativo intervalar*
  - *Código postal: ?*
  - *Posição em uma corrida: ?*

Exercício

- Definir o tipo e escala dos seguintes atributos:
  - *Renda mensal: quantitativo racional*
  - *Número de palavras de um texto: quantitativo racional*
  - *Número de matrícula: qualitativo nominal*
  - *Data de nascimento: quantitativo intervalar*
  - *Código postal: qualitativo nominal*
  - *Posição em uma corrida: ?*

Exercício

- Definir o tipo e escala dos seguintes atributos:
  - *Renda mensal: quantitativo racional*
  - *Número de palavras de um texto: quantitativo racional*
  - *Número de matrícula: qualitativo nominal*
  - *Data de nascimento: quantitativo intervalar*
  - *Código postal: qualitativo nominal*
  - *Posição em uma corrida: qualitativo ordinal*

Exploração de 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

Informações obtidas podem ajudar na seleção de técnicas apropriadas de pré-processamento e aprendizado

Exploração de dados

Frequência

- Proporção de vezes que um atributo assume um dado valor

- Aplicável a valores numéricos e simbólicos

- *Ex.: 40% dos pacientes têm febre*
- *Ex.: digite `summary(iris)` em R, ver o atributo classe*

Localização, dispersão e distribuição

- Diferem para dados univariados e multivariados
  - *Maioria dos dados em AM é multivariado, mas análises em cada atributo podem fornecer informações valiosas*

- Geralmente aplicados a valores numéricos

Frequência

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Frequência: 25% das manchas são médias

Dados univariados

- Objetos com apenas um atributo
  - *Conjunto com n objetos* $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$

Observação: termo conjunto não tem o mesmo significado do usado em teoria dos conjuntos
Em um conjunto de dados, o mesmo valor pode aparecer mais de uma vez em um atributo

Dados univariados: medidas de localidade

- Definem pontos de referência nos dados
  - *Valor "típico", que resume os dados*

Valores numéricos	Valores simbólicos
• Média • Mediana • Percentil	• Moda: valor mais frequente

Moda

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Moda: Grandes

Média

- Equação:

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

Problema: sensível a *outliers*

Bom indicador apenas se valores são distribuídos simetricamente

Mediana

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

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

Facilita observar se ditribuição é assimétrica ou se existem *outliers*

Mediana

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

Média e mediana

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Média: 26,1
Mediana: 21,5

`mean()` e `median()` no R

Média e mediana

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Média: 5
Mediana: 2,5

Quartis e percentis

- Mediana divide dados ordenados ao meio
  - *Quartis e percentis usam pontos de divisão diferentes*

Quartis

- Divide em quartos
- 1º quartil (*Q*1) ⇨ valor que tem 25% dos demais valores abaixo dele
- 2º quartil (Q2) = mediana
- 3º quartil (Q3) ⇨ tem 75% dos demais valores abaixo

Percentil

- Para p entre 0 e 100
- *p*º percentil = *Pp* ⇨ x_i tal que *p*% dos valores observados são menores do que x_i
- *P25* = *Q1*
- *P50* = *Q2* = mediana

Percentil

Algoritmo para cálculo do percentil

Entrada: n valores e percentil p

Saída: valor do percentil

- *Ordenar os n valores de maneira crescente*
- *Calcular k = n * p*
- *Se k não for inteiro então*
  - Arredondar para o próximo inteiro
  - Retornar o valor dessa posição na sequência
- *Senão*
  - Retornar média entre os valores nas posições k e k+1

Quartil e percentil

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Média:26,1
Mediana: 21,5
Q1: 18,7; *Q2*: 21,5; *Q3*: 29,5
*P*40: 21

`summary(x)` no R fornece várias dessas estatísticas (ver também `quantile()`)

Quartil e percentil

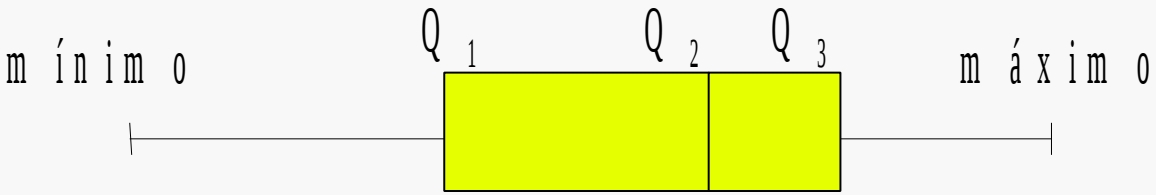
- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Média:5
Mediana: 2,5
Q1: 2; *Q2*: 2,5; *Q3*: 4,5
*P*40: 2

Boxplots

- Também chamados diagramas de Box e Whisker
- Forma gráfica de visualizar quartis
  - *Usa quartis e valores máximo e mínimo*

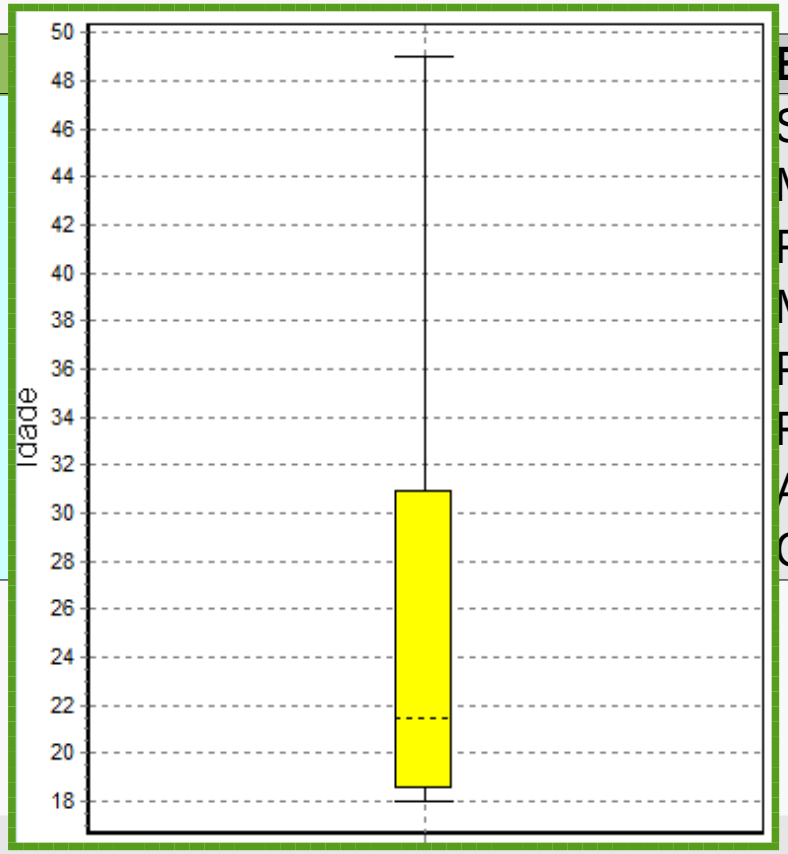


Boxplot modificado: limite superior/inferior vai até maior/menor valor apenas se esse valor não for muito distante do 3º/1º quartil (até 1,5 * intervalo entre quartis Q3 e Q1)
Valores acima/abaixo são considerados *outliers*

Boxplot

- Ex. conjunto de dados `hospital`

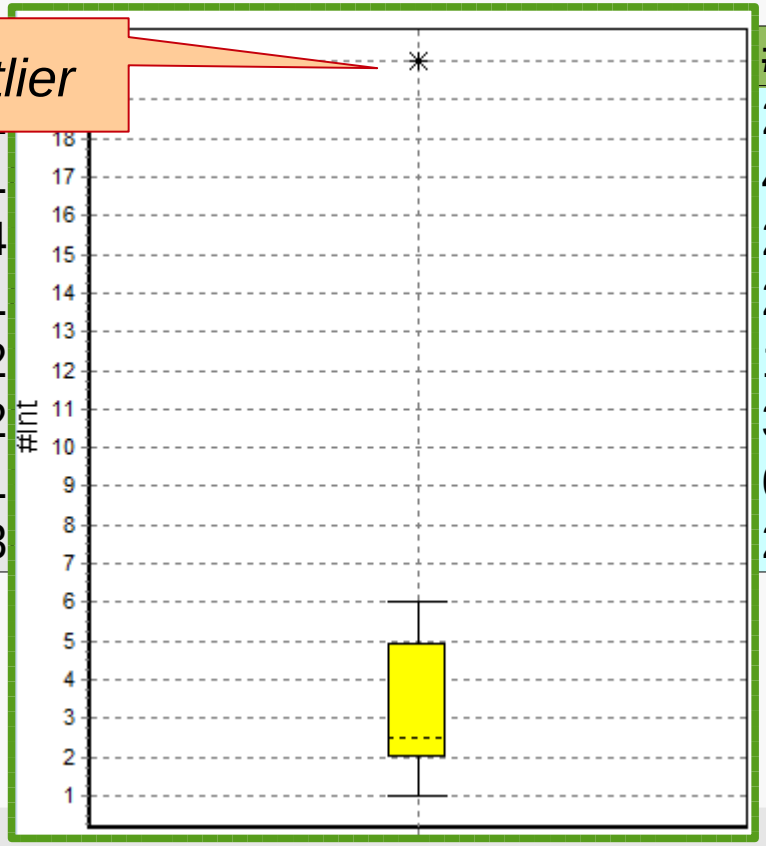
Id.	Nome	Idade	Est.	Diagnóstico
4201	João	28	SP	Doente
3217	Maria	18	MG	Doente
4039	Luiz	49	RS	Saudável
1920	José	18	MG	Doente
4340	Cláudia	21	PE	Saudável
2301	Ana	22	RJ	Doente
1322	Marta	19	AM	Doente
3027	Paulo	34	GO	Saudável



Boxplot

- Ex. conjunto de dados `hospital`

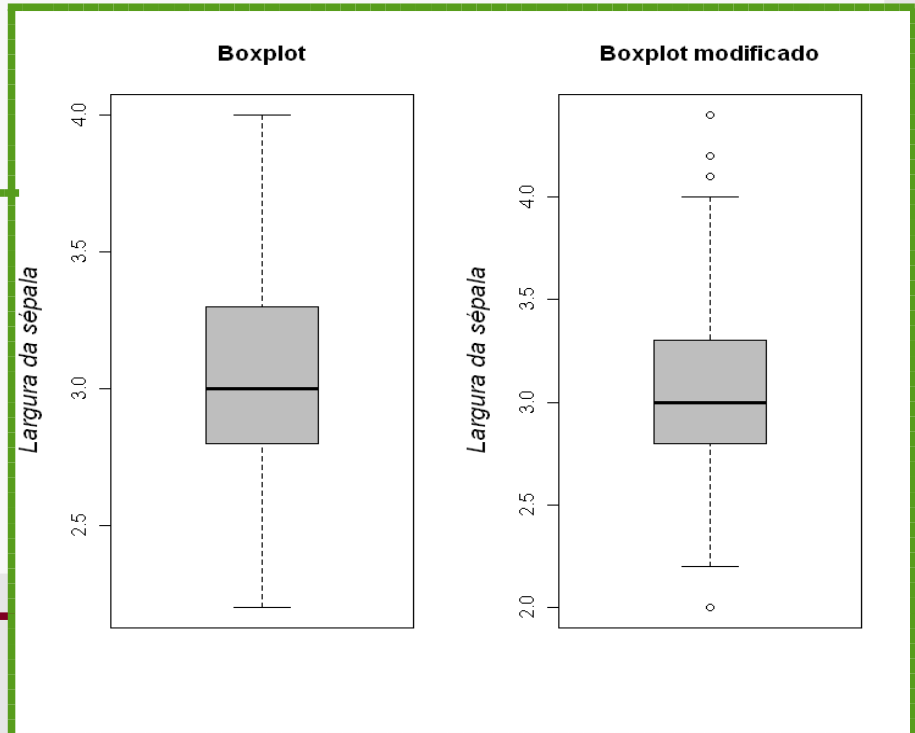
Id.	No		# Int.	Est.	Diagnóstico
4201	Joã		2	SP	Doente
3217	Maria	1	4	MG	Doente
4039	Luiz	4	2	RS	Saudável
1920	José	1	20	MG	Doente
4340	Cláudia	2	1	PE	Saudável
2301	Ana	2	3	RJ	Doente
1322	Marta	1	6	AM	Doente
3027	Paulo	3	2	GO	Saudável



Boxplot



- Ex. conjunto de dados `iris`
  - *150 objetos*
  - *4 atributos de entrada (contínuos):*
    - Tamanho pétala
    - Tamanho sépala
    - Largura pétala
    - Largura sépala
  - *3 classes (espécies de íris):*
    - Íris vírginica
    - Íris setosa
    - Íris versicolor



Dados univariados: medidas de espalhamento

- 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

Intervalo

- Mostra espalhamento máximo entre valores
  - *Medida mais simples*

$$\text{intervalo}(\mathbf{x}) = \max_{i=1,\ldots,n}(x_i) - \min_{i=1,\ldots,n}(x_i)$$

Problema: não é boa medida se maioria dos valores está próxima de um ponto, com um pequeno número de valores extremos

Intervalo

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Intervalo: 31

Em R:
`max() - min()`

Intervalo

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Intervalo: 19

Variância e desvio padrão

- Mais utilizadas para avaliar espalhamento

$$\text{variância}(\mathbf{x}) = \frac{1}{n\text{-}1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\text{desvio padrão}(\mathbf{x}) = \sqrt{\text{variância}(\mathbf{x})}$$

Problema: também são distorcidas pela presença de *outliers*

Desvio padrão

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Intervalo: 31
Desvio padrão: 10,8

Em R:
`var()` e `sd()`

Desvio padrão

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Intervalo: 19
Desvio padrão: 6,3

Outras medidas de espalhamento

- Desvio médio absoluto

$$DMA(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$$

- Desvio mediano absoluto

$$DMedA(\mathbf{x}) = mediana(\{|x_1 - \bar{x}|, \ldots, |x_n - \bar{x}|\})$$

- Intervalo interquartil

$$IQ(\mathbf{x}) = P75 - P25$$

Outras medidas de espalhamento

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Intervalo: 31
Desvio padrão: 10,8
DMA: 8,2
DmedA: 7,5
IQ: 10,8

Em R:
DMA:
`sum(abs(x-mean(x)))/length(x)`
DmedA: `median(abs(x-mean(x)))`
IQ: `IQR(x)`

Outras medidas de espalhamento

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Intervalo: 19
Desvio padrão: 6,3
DMA: 4
DmedA: 3
IQ: 2,5

Momento

- Medidas em torno da média de um conjunto de valores, em geral, são instanciações de medida de momento:

$$\text{momento}_k(\mathbf{x}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^k}{(n\text{-}1)}$$

- $k = 1$ ⇨ momento central (primeiro momento em torno da origem)
- $k = 2$ ⇨ variância (segundo momento central)
- $k = 3$ ⇨ obliquidade (terceiro momento central)
- $k = 4$ ⇨ curtose (quarto momento central)

Dados univariados: medidas de distribuição

- Obliquidade e curtose são medidas de distribuição
  - *Mostram como valores estão distribuídos*

Obliquidade	Curtose
• *Skweness* • Mede simetria da distribuição em torno da média	• *Kurtosis* • Captura achatamento da função de distribuição

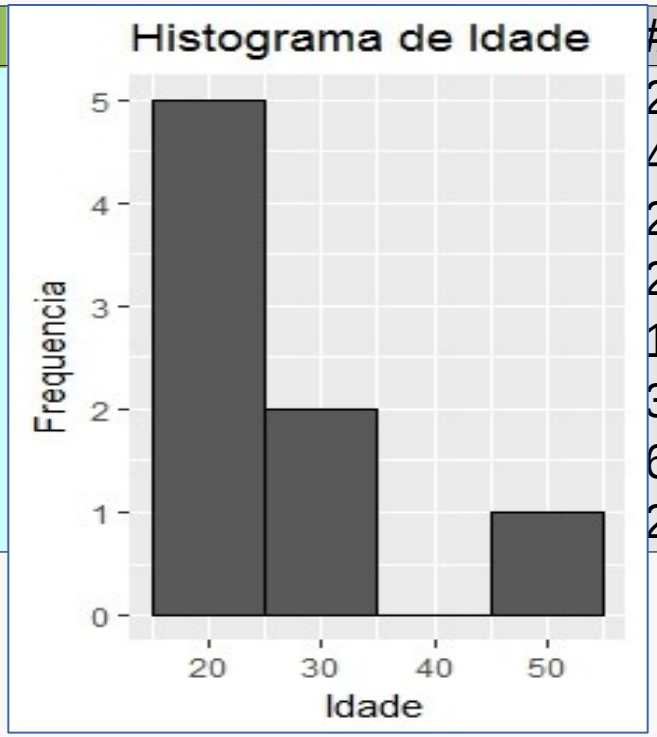
Histograma

- Forma gráfica para visualizar distribuição: histograma
  - *Divide valores em cestas*
    - Valores categóricos: cada valor é uma cesta
    - Valores numéricos: divisão em intervalos contíguos de mesmo tamanho e cada intervalo é uma cesta
  - *Para cada cesta, desenha uma barra com altura proporcional ao número de elementos na cesta*

Histograma

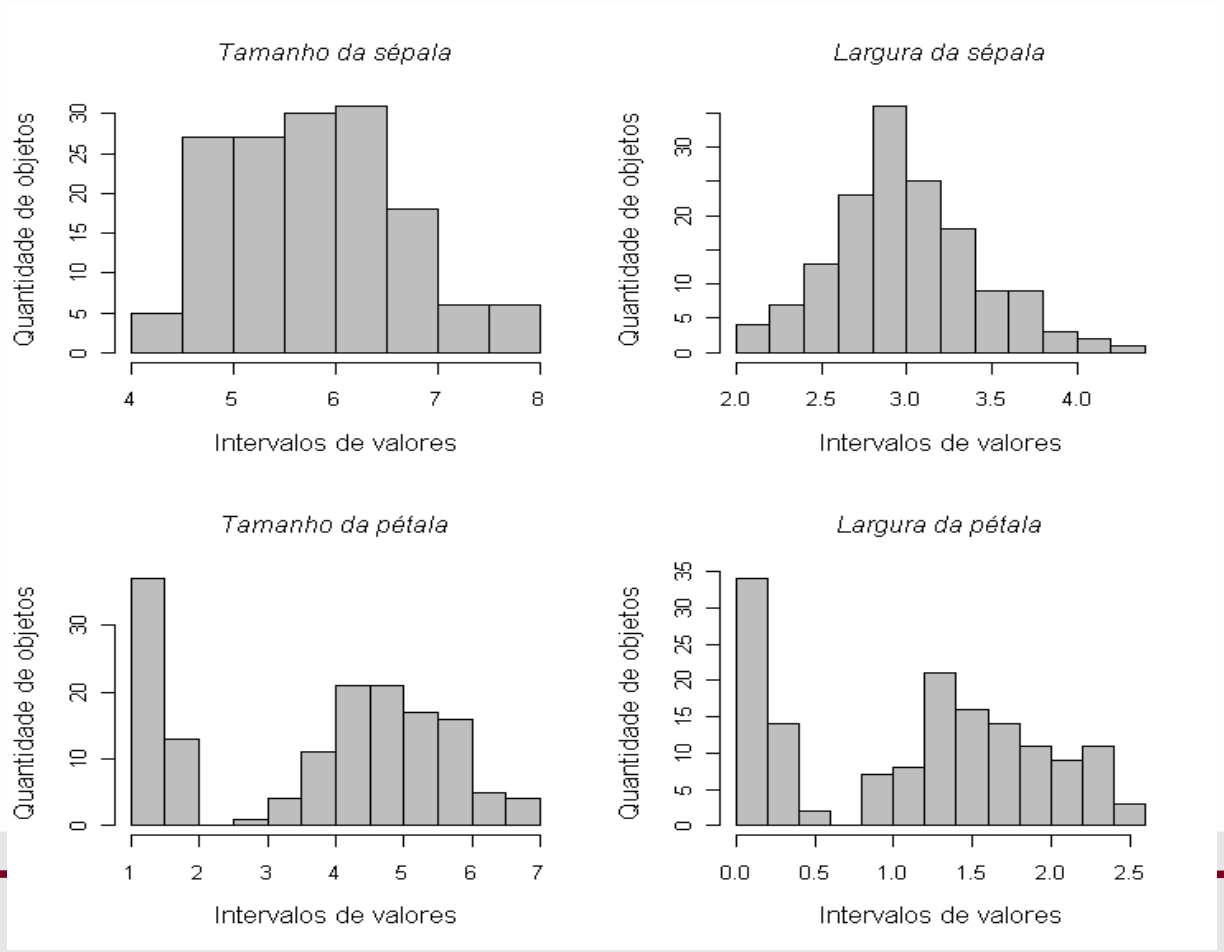
- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	# Int.	Est.	Diagnóstico
4201	João	28	2	SP	Doente
3217	Maria	18	4	MG	Doente
4039	Luiz	49	2	RS	Saudável
1920	José	18	20	MG	Doente
4340	Cláudia	21	1	PE	Saudável
2301	Ana	22	3	RJ	Doente
1322	Marta	19	6	AM	Doente
3027	Paulo	34	2	GO	Saudável



Histograma

- Ex. conjunto de dados `iris`



Obliquidade

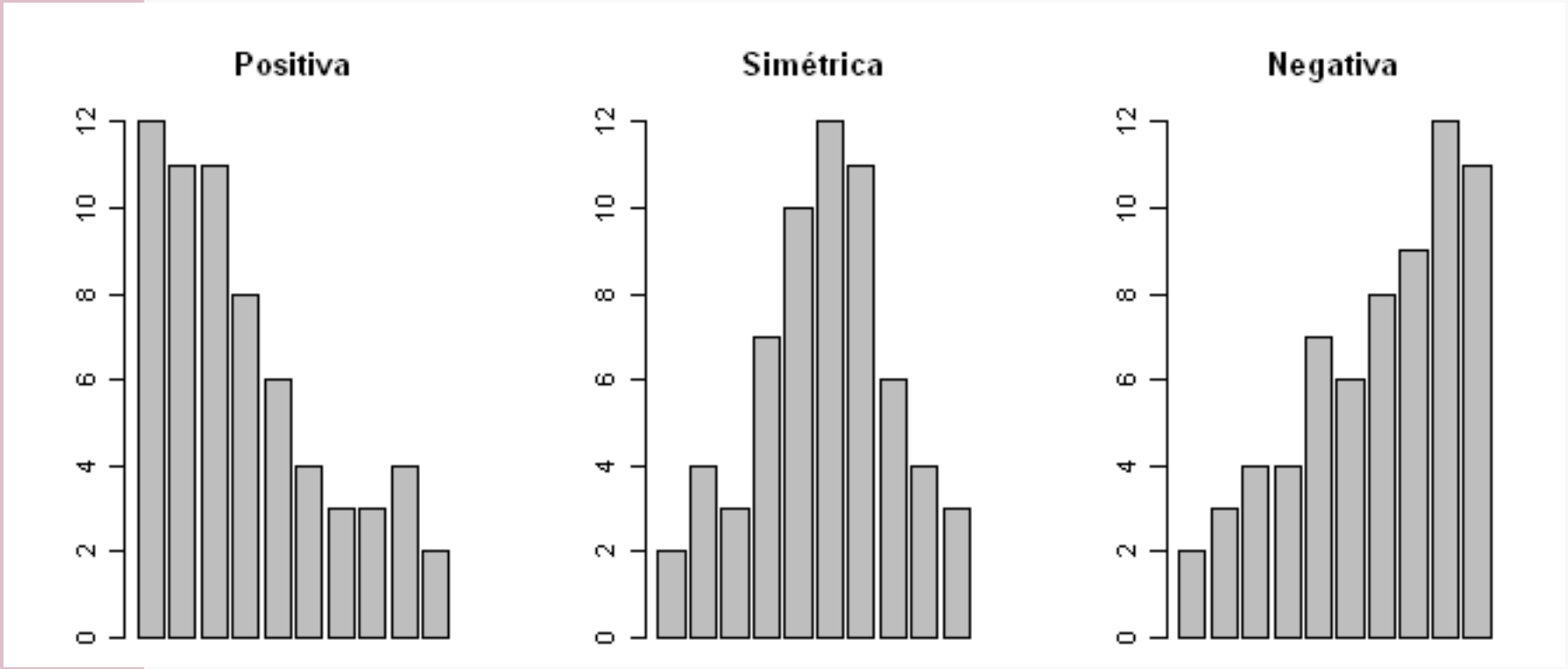
- Equação:

$$\text{obliquidade}(\mathbf{x}) = \frac{\text{momento}_3(x)}{\text{desv\_pad}^3}$$

Valores de obliquidade:

- **= 0 (simétrica)**: distribuição é aproximadamente simétrica
- **> 0 (positiva)**: distribuição concentra-se mais no lado esquerdo
- **< 0 (negativa)**: distribuição concentra-se mais no lado direito

Obliquidade



Curtose

- Verifica se dados apresentam um pico ou são achatados em relação a uma distribuição normal

$$\text{curtose}(\mathbf{x}) = \frac{\text{momento}_4(\mathbf{x})}{\text{desv\_pad}^4} - 3$$

Valores de curtose:

- **= 0 (normal)**: histograma tem achatamento de distribuição normal
- **> 0 (positiva)**: histograma tem distribuição mais alta e concentrada
- **< 0 (negativa)**: histograma tem distribuição mais achatada

Curtose

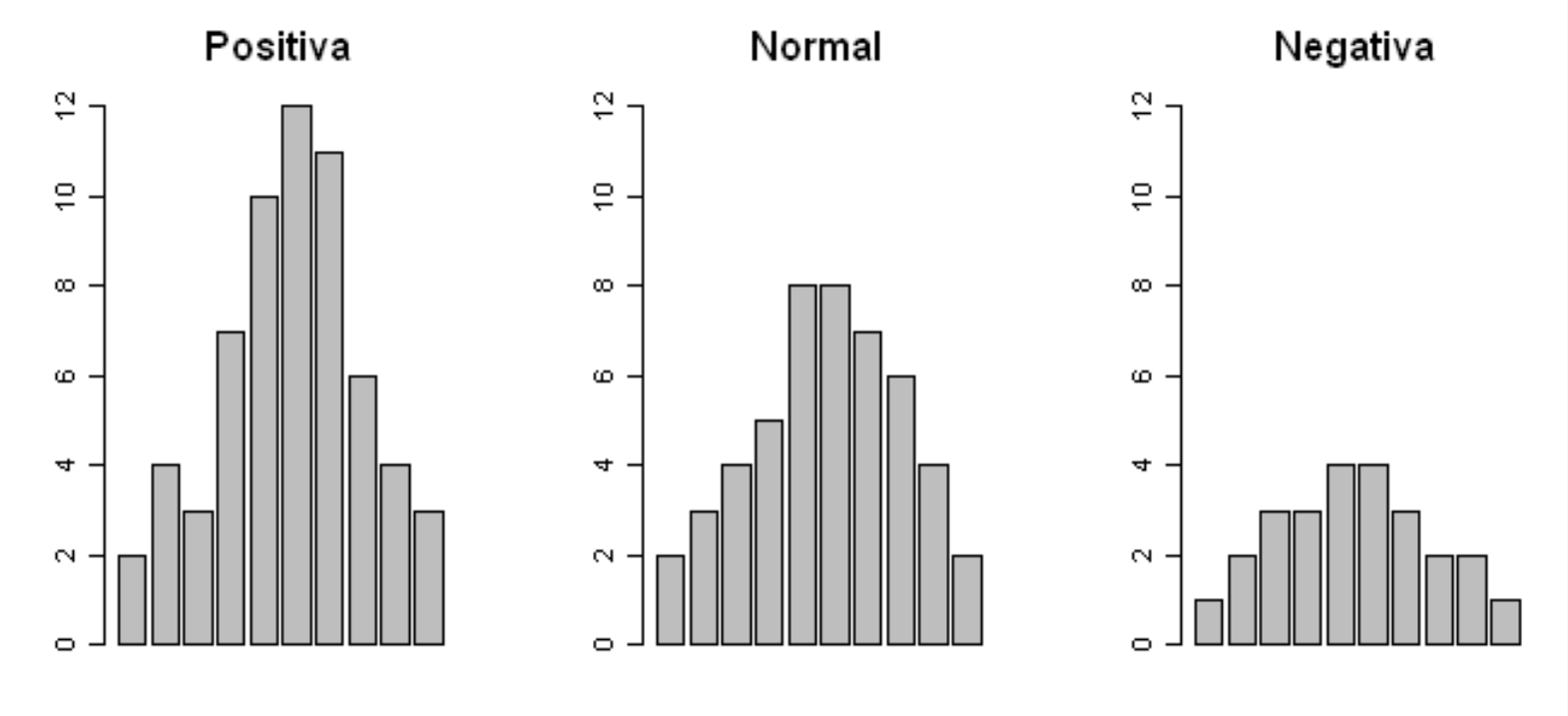


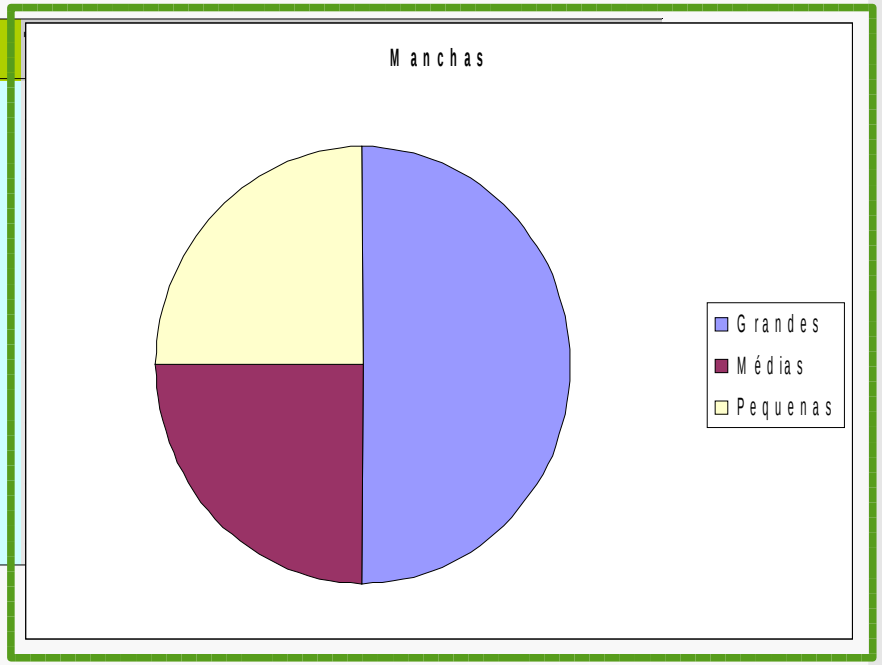
Gráfico de pizza

- Outra forma gráfica de visualizar distribuição de um conjunto de valores
  - *Indicado para valores qualitativos*
    - Para quantitativos, deve agrupar valores em cestas
  - *Cada valor ocupa fatia com área proporcional ao número de vezes que aparece no conjunto de dados*

Gráfico de pizza

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas
4201	João	28	M	79	Grandes
3217	Maria	18	F	67	Pequenas
4039	Luiz	49	M	92	Grandes
1920	José	18	M	43	Grandes
4340	Cláudia	21	F	52	Médias
2301	Ana	22	F	72	Pequenas
1322	Marta	19	F	87	Grandes
3027	Paulo	34	M	67	Médias



Dados multivariados

- Possuem mais de um atributo de entrada
  - *Ex. conjuntos de dados* `hospital` *e* `iris`
  - *Medidas de localidade e espalhamento podem ser calculadas para cada atributo separadamente*
    - Ex. média

$$\bar{\mathbf{x}} = (\bar{x}^1, \ldots, \bar{x}^d)$$

Dados multivariados

- Permitem análises da relação entre dois ou mais atributos
  - *Para variáveis contínuas, espalhamento é melhor capturado por uma matriz de covariância*
    - Cada elemento é covariância entre dois atributos

$$\text{covariância}(\mathbf{x}^i, \mathbf{x}^j) = \frac{1}{n-1} \sum_{k=1}^{n} (x_k^i - \bar{x}^i)(x_k^j - \bar{x}^j)$$

Observação: covariância($\mathbf{x}^i$, $\mathbf{x}^i$) = variância($\mathbf{x}^i$)

Covariância

- Covariância entre dois atributos mede grau com que variam juntos

Valores de covariância entre dois atributos x^i e x^j:

- **Próximo de 0**: atributos não têm um relacionamento linear
- **> 0 (positiva)**: atributos são diretamente relacionados
- **< 0 (negativa)**: atributos são inversamente relacionados

  - *Valor depende da magnitude dos atributos*
    - Não é possível avaliar relacionamento de atributos apenas por covariância

Correlação

- Indicação mais clara da força da relação linear entre dois atributos
  - *Matriz de correlação: correlação entre todos pares de atributos*

$$\text{correlação}(\mathbf{x}^i, \mathbf{x}^j) = \frac{\text{covariância}(\mathbf{x}^i, \mathbf{x}^j)}{\text{desv\_pad}(\mathbf{x}^i) * \text{desv\_pad}(\mathbf{x}^j)}$$

Observação: valores variam de -1 (correlação negativa máxima) a +1 (correlação positiva máxima) e correlação($\mathbf{x}^i$, $\mathbf{x}^i$) = 1

Covariância ecorrelação

- Ex. conjunto de dados `iris`
  - *Matriz de covariância:*

	Tamanho_sépala	Largura_sépala	Tamanho_pétala	Largura_pétala
Tamanho_sépala	0,68569	-0,03927	1,27368	0,51690
Largura_sépala	-0,03927	0,18800	-0,32171	-0,11798
Tamanho_pétala	1,27368	-0,32171	3,11318	1,29639
Largura_pétala	0,51690	-0,11798	1,29639	0,58241

  - *Matriz de correlação:*

	Tamanho_sépala	Largura_sépala	Tamanho_pétala	Largura_pétala
Tamanho_sépala	1,00000	-0,10937	0,87175	0,81795
Largura_sépala	-0,10937	1,00000	-0,42052	-0,35654
Tamanho_pétala	0,87175	-0,42052	1,00000	0,96276
Largura_pétala	0,81795	-0,35654	0,96276	1,00000

Em R: `cov(x)` e `cor(x)`

Dados multivariados: visualização

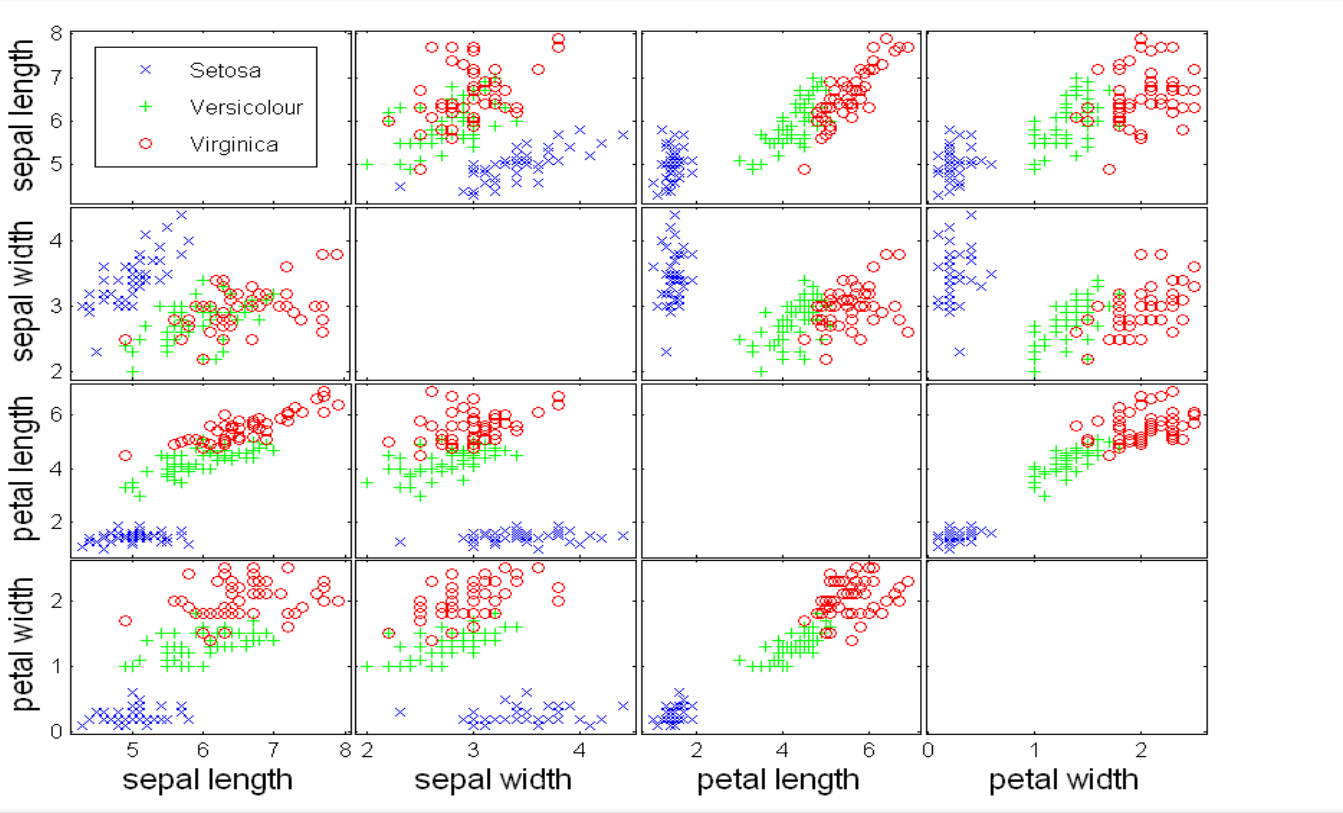
- Diagramas para visualizar dados multivariados
  - *Em particular, relação entre diferentes atributos*
  - *Alguns tipos de gráficos:*
    - *Scatter plot*
    - *Bag plots*
    - Faces de *Chernoff*
    - *Star plots*
    - *Heatmaps*

Scatter plot

- Ilustra correlação linear entre dois atributos
  - *Cada objeto é associado a uma posição em um plano*
    - Valores dos atributos definem a sua posição
    - Valores são inteiros ou reais
  - *Matrizes de scatter plot: relacionamento de vários atributos*

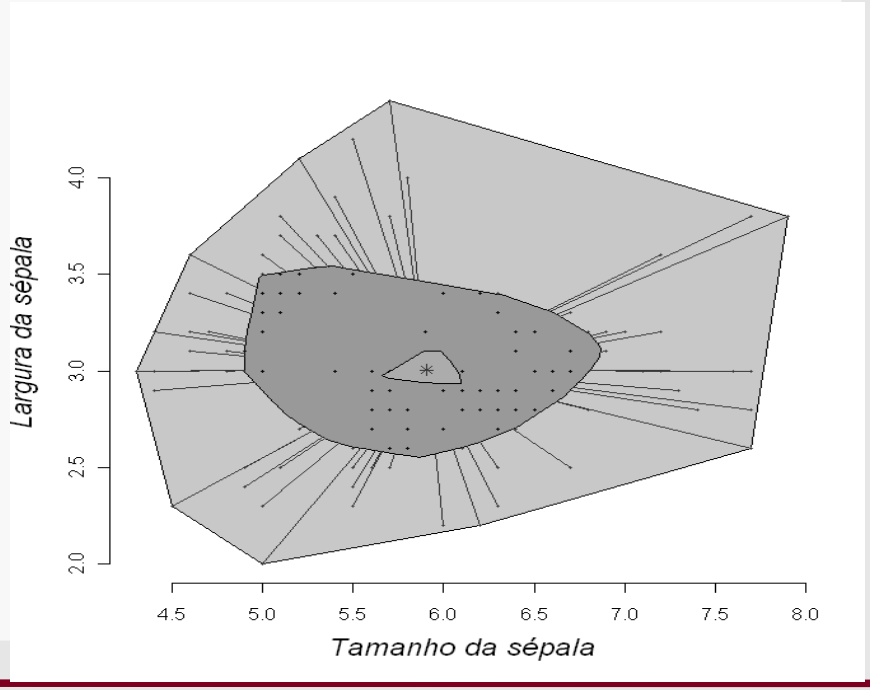
Scatter plot

- Ex. conjunto de dados `iris`



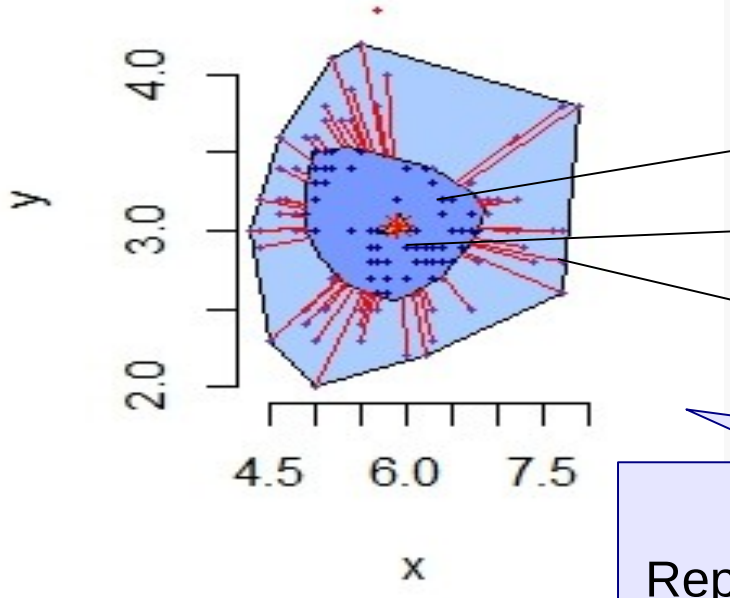
Bagplot

- Generalização bivariada do *boxplot*
  - *Apresenta, em mesma figura, o boxplot de dois atributos*
    - Cada eixo pode ser considerado um *boxplot* de um dos atributos
  - *Ex. conjunto de dados* `iris`



Bagplot

- Generalização bivariada do *boxplot*



Bag: 50% dos objetos (1º e 3º quartis de cada atributo)

medianas

Loop: bag expandida (1,5 vezes em cada sentido * intervalo interquartil)

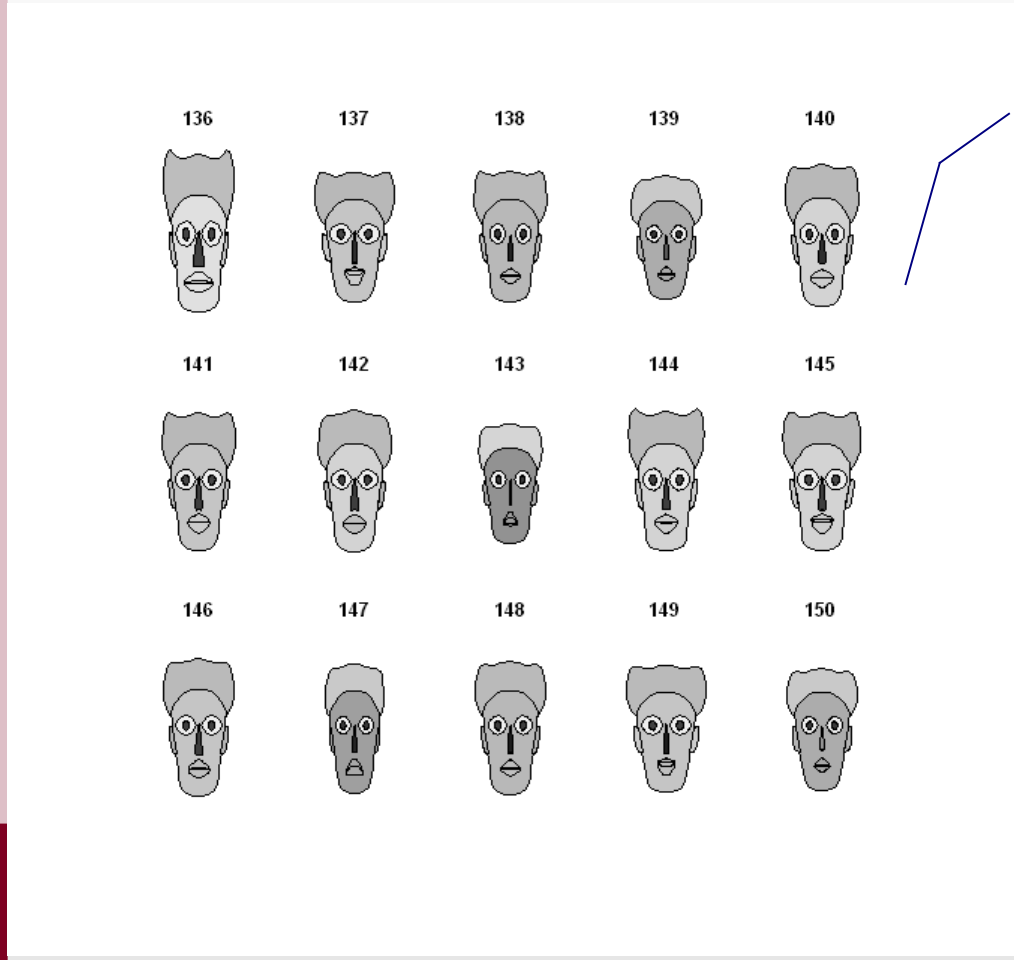
Cada dimensão vista isoladamente
Representa o boxplot para o atributo associado

Diagrama de Chernoff

- Mapeia valores dos atributos para imagens mais familiares: faces
  - *Cada objeto é representado por uma face*
  - *Cada atributo é associado a uma ou mais características da face*
    - Ex. altura e largura da cabeça, da boca, etc.
- Baseia-se na habilidade humana de distinguir faces

Diagrama de Chernoff

- Ex. conjunto de dados `iris`



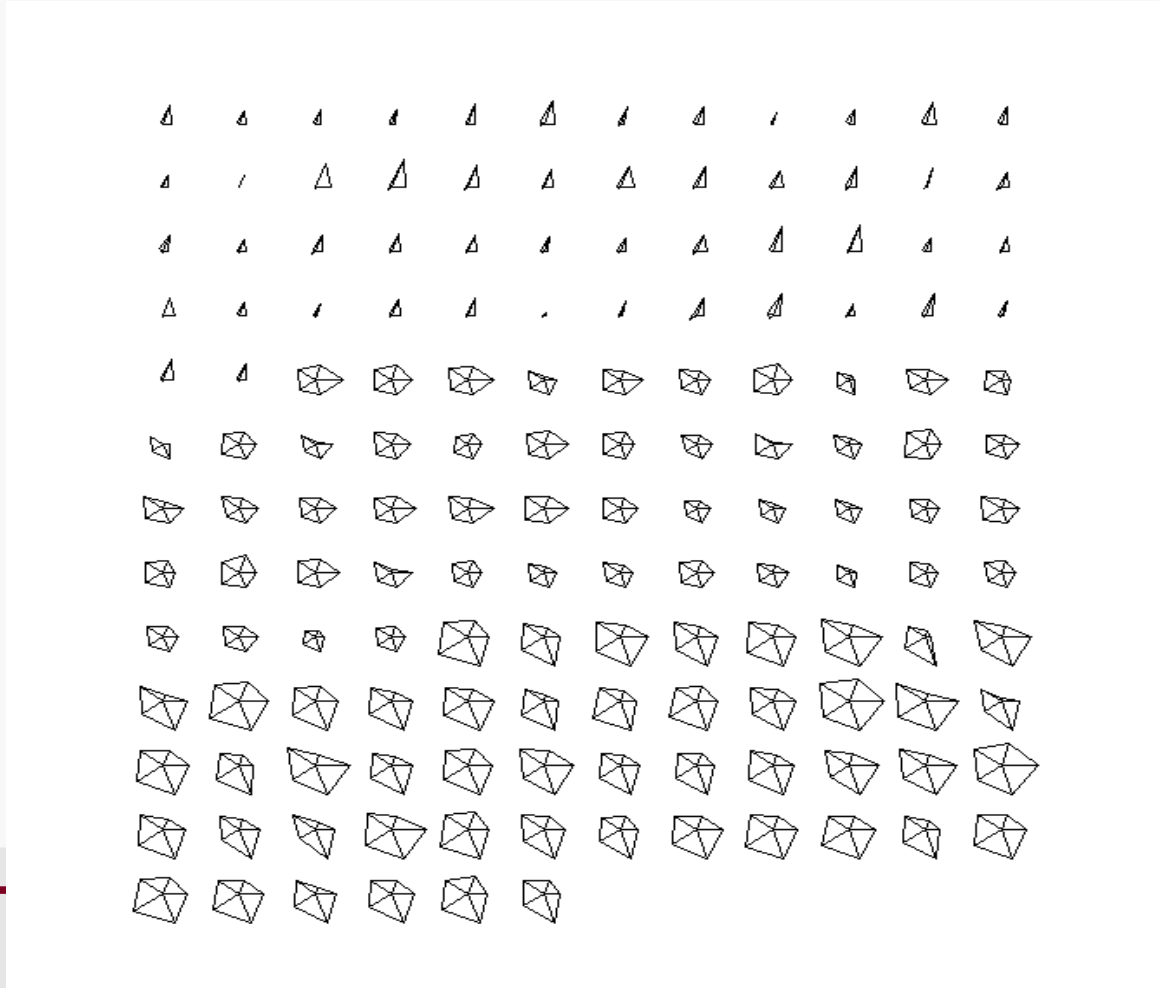
Tamanho da sépala representado por altura da face, largura da boca, altura do cabelo e largura do nariz

Star plot

- Desenha figura geométrica para cada objeto
  - *Normalmente um polígono*
  - *Cada linha do polígono corresponde a um dos atributos*
    - Tamanho da linha é proporcional ao valor do atributo
    - Quanto mais atributos, mais o polígono se assemelha a estrela
    - Valores de atributos semelhantes deformam a estrela

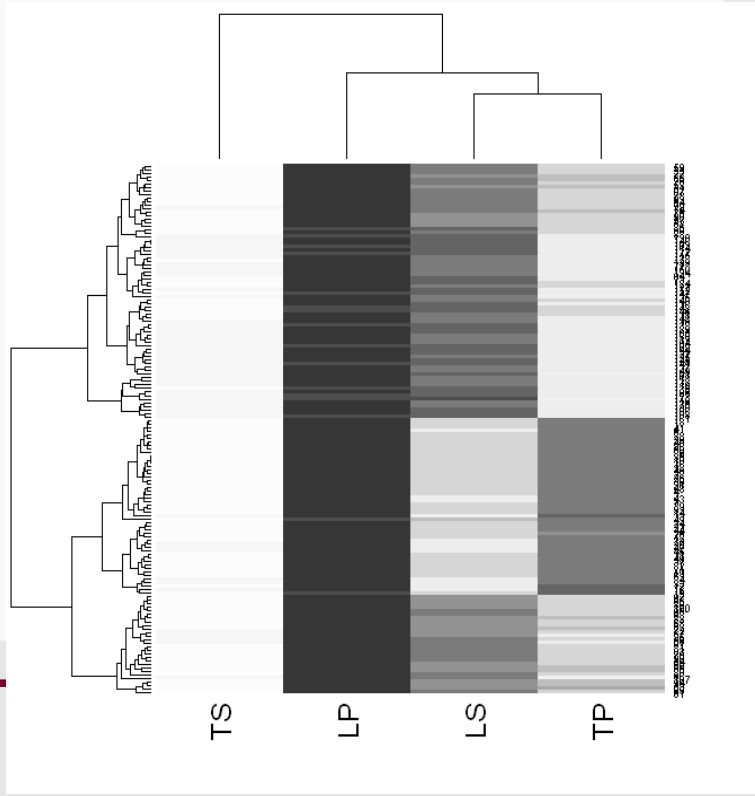
Star plot

- Ex. conjunto de dados `iris`



Heatmap

- Representa relação entre exemplos e as classes
  - *Agrupamento hierárquico em cada eixo (dendograma)*
    - Auxilia a verificar tendências nos dados
  - *Ex. conjunto de dados* `iris`



Considerações finais

- Dados
  - *Caracterização de dados*
  - *Tipos e escala de atributos*
- Exploração de dados
  - *Medidas de localidade, dispersão e distribuição*
  - *Técnicas de visualização*

Referências

- Ilustrações utilizadas:
  - http://neowayinfo.blogspot.com/2011/05/como-gerenciar-um-grande-volume-de.html
  - http://www.icess.ucsb.edu/gem/filtragem1.htm
  - http://brainstormdeti.wordpress.com/2010/11/06/prova-todo-grafo-completo-e-conexo/
  - http://entomologia.rediris.es/iberodorcadion/Fotos/textos.html
  - http://www.adrformacion.com/cursos/front/leccion1/tutorial3.html
  - http://clipart.usscouts.org/library/
  - http://www.clker.com/clipart-video-camera.html
  - http://www.clker.com/clipart-audio-speaker-1.html
  - http://www.canalexecutivo.com/t533.htm
  - http://intrometendo.com/hierarquia-militar-no-brasil/
  - http://www.sortimentos.com/gente/espaco-profissional-pagamento-13-salario.htm
  - http://fisioterapiahumberto.blogspot.com/2009/12/desvio-padrao-afinal-de-contas-para-que.html
  - http://www.alaska-in-pictures.com/wild-iris-picture-alaskan-summer-8865-pictures.htm
  - http://www.fs.fed.us/wildflowers/beauty/iris/blueflag/iris_virginica.shtml
  - http://www.floweringflowers.net/2010/04/iris/iris-versicolor/

Referências

- Softwares utilizados:
  - *Fast Statistics 2.0.4*
  - *RStudio*
  - *Weka*
  - *http://www.shodor.org/interactivate/activities/*
- Alguns slides são baseados em apresentações de:
  - *Prof Dr André C. P. L. F. Carvalho, ICMC-USP*