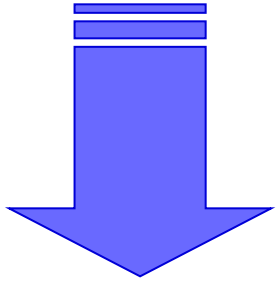
Aprendizado de Máquina

Análise e Visualização de Dados

Dados

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

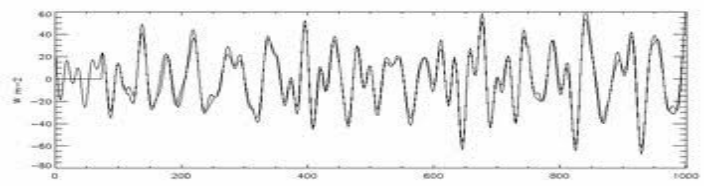


Bases de dados cada vez maiores

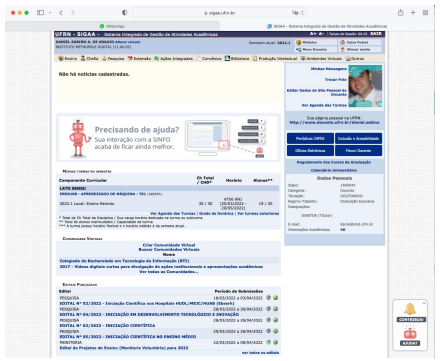
Dados

Geralmente transformados para o formato atributo-valor

Podem ter diferentes formatos

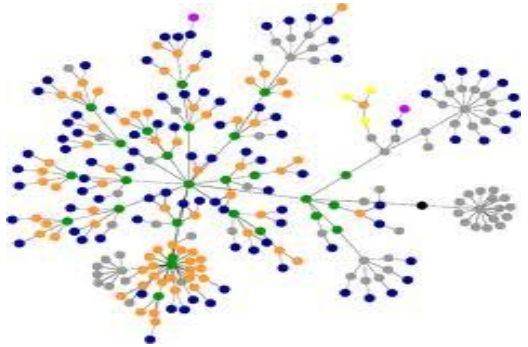


Séries temporais

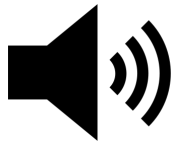


Páginas web

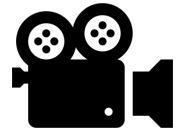
Textos



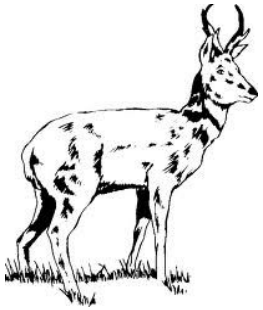
Grafos



Áudios



Vídeos



Imagens

Conjunto de dados

- Pode ser representado por uma matriz de objetos

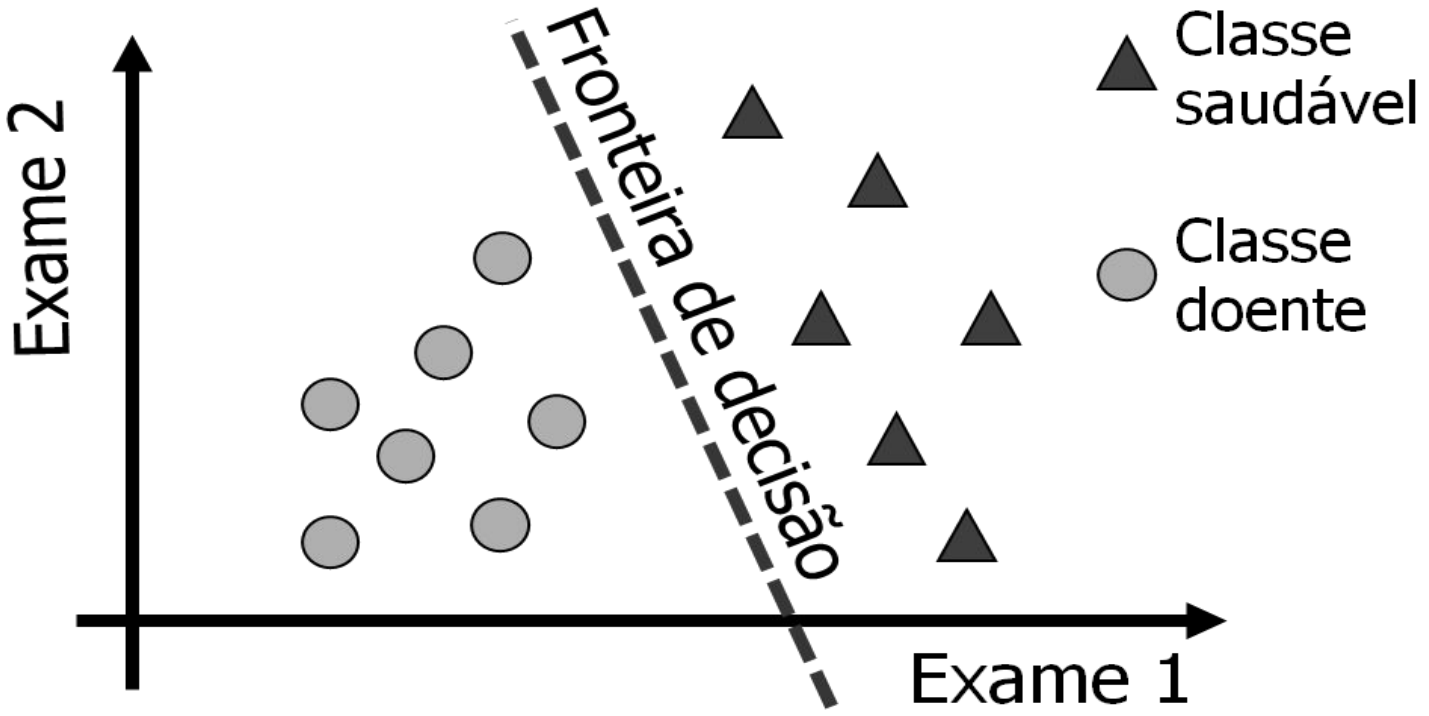
$$X_n \times d$$

  - n = número de objetos
  - d = número de atributos (excluindo atributo-meta)
    - Dimensionalidade dos objetos

- Elemento x_{ij} ⇨ valor da j-ésima característica para o objeto i

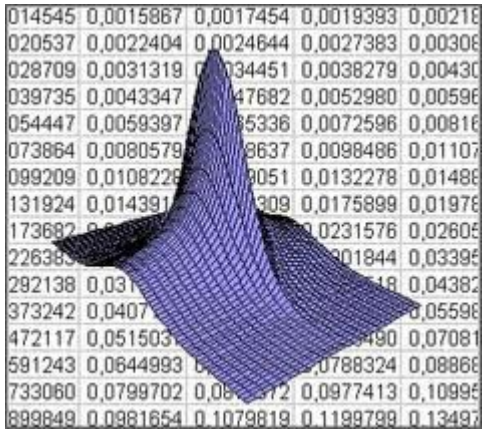
Conjunto de dados: visualização gráfica

- 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

Caracterização e Exploração dos Dados

Categorização

- Valores de atributos podem ser definidos por:
  - Tipo
    - Grau de quantização dos 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
  - Ex.: 29.7, 3, 100
- 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

id	Nome	Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Estado	Diagnóstico
4201	João	28	M	70	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

id	Nome	Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Estado	Diagnóstico
4201	João	28	M	70	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

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
- 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

id	Nome	Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Estado	Diagnóstico
4201	João	28	M	70	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: 50% das manchas são GRANDES

Dados univariados

- Objetos com apenas um atributo
  - Conjunto com n objetos $x = \{ x_1, x_2, ..., x_n \}$
- 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", resume os dados

Valores numéricos

- Média
- Mediana
- Percentil

Valores simbólicos

- Moda: valor mais frequente

Moda

id	Nome	Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Estado	Diagnóstico
4201	João	28	M	70	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:

$$mediana(x) = \begin{cases} \frac{1}{2}(x_r + x_{r+1}), & n \bmod 2 = 0 \ (n = 2r) \\ x_{r+1}, & n \bmod 2 \neq 0 \ (n = 2r + 1) \end{cases}$$

Facilita observar se distribuição 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

id	Nome	Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Estado	Diagnóstico
4201	João	28	M	70	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

Média e mediana

id	Nome	Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Estado	Diagnóstico
4201	João	28	M	70	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

Média truncada

- Descarta elementos extremos da sequência ordenada de valores
  - Minimizar problemas da média
  - Necessário definir porcentagem

Média truncada

id	Nome	Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Estado	Diagnóstico
4201	João	28	M	70	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
Média truncada (p = 25%): 23,7

Média truncada

id	Nome	Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Estado	Diagnóstico
4201	João	28	M	70	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
Média truncada (p = 25%): 3,2

Quartis e percentis

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

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

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

Quartil e percentil

id	Nome	Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Estado	Diagnóstico
4201	João	28	M	70	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
Média truncada (p = 25%): 23,7
Q1: 18,5; Q2: 21,5; Q3: 31
P40: 21

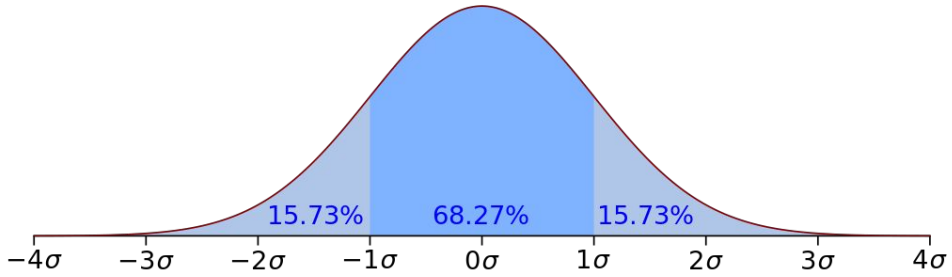
Quartil e percentil

id	Nome	Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Estado	Diagnóstico
4201	João	28	M	70	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
Média truncada (p = 25%): 3,2
Q1: 2; Q2: 2,5; Q3: 5
P40: 2

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



https://pt.wikipedia.org/wiki/Distribuição_normal#/media/Ficheiro:Boxplot_vs_PDF.svg

Intervalo

- Mostra espalhamento máximo entre valores
  - Medida mais simples

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

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

Intervalo

id	Nome	Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Estado	Diagnóstico
4201	João	28	M	70	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

Intervalo

id	Nome	Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Estado	Diagnóstico
4201	João	28	M	70	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

id	Nome	Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Estado	Diagnóstico
4201	João	28	M	70	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

Desvio padrão

id	Nome	Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Estado	Diagnóstico
4201	João	28	M	70	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

Transformação de Dados

Tipos De Transformação

- Conversão simbólico-numérico
- Conversão numérico-simbólico
- Transformação de atributos numéricos
- Redução de dimensionalidade

Conversão Simbólico-Numérico

- Atributo simbólico com dois valores
  - Um dígito binário é suficiente
    - Ex. presença/ausência = 1/0

      Se ordinal, 0 indica o menor valor e 1 o maior valor
- Atributo simbólico com mais valores
  - Conversão depende se o atributo é nominal ou ordinal

Atributo Nominal Com Mais De Dois Valores

- Inexistência de relação de ordem deve ser mantida
  - Diferença entre quaisquer dois valores numéricos deve ser a mesma
- Codificação canônica: uso de c bits para c valores
  - Cada posição na sequência binária corresponde a um valor possível do atributo nominal
  - Cada sequência possui apenas um bit com valor 1
  - Distância de Hamming entre quaisquer dois valores é 2

Exemplos

M	0
F	1

Azul	10000
Amarelo	01000
Vermelho	00100
Verde	00010
Preto	00001

Pseudoatributos

Pseudoatributo	#valores
Continente	6
PIB	1
População	1
Área	1

Atributo Ordinal Com Mais De Dois Valores

- Relação de ordem deve ser preservada
- Ordenar valores e codificar cada um de acordo com sua posição na ordem com inteiro ou real

Atributos	Valor inteiro
Primeiro	0
Segundo	1
Terceiro	2
Quarto	3

Distância

Atributos	Valor binário
Primeiro	000
Segundo	001
Terceiro	010
Quarto	011

Distância

Conversão Numérico-Simbólico

- Atributo discreto e binário ⇒ conversão é trivial
  - Associa um nome a cada valor
- Demais casos: discretização
  - Transforma valores numéricos em intervalos (categorias)
  - Existem vários métodos diferentes para discretização
  - Paramétricos: usuário pode influenciar definição dos intervalos
  - Não paramétricos: usam apenas informações presentes nos valores dos atributos

Conversão Numérico-Simbólico

- Métodos de discretização podem ser:
  - Supervisionados: usa informação da classe
    - Melhor resultado, não leva a mistura de classes
    - Ex. escolher pontos de corte que maximizam pureza das classes (entropia)
  - Não supervisionados
- Método de discretização deve definir:
  - Como mapear valores quantitativos para qualitativos
  - Tamanho dos intervalos
  - Quantidade de valores nos intervalos

Conversão Numérico-Simbólico

- Algumas estratégias:
  - Larguras iguais
  - Frequências iguais
  - Algoritmos de agrupamento
  - Inspeção visual

Conversão Numérico-Numérico

- Quando os valores dos atributos são muito diferentes.
- Um atributo pode se tornar predominante.
- Quando isso é importante?
- Técnica muito utilizada: normalização.
  - Faz com que um atributo fique com determinada propriedade.

Normalização

- Por amplitude:
  - Reescala (valor máximo e mínimo)
  - Padronização (valor central e espalhamento)
- Distribuição (ranqueamento)

Reescala

$$x' = \frac{x - min}{max - min}$$

Padronização

$$x' = \frac{x - \bar{x}}{\sigma}$$

Outros tipos

- Tradução
  - Valor é traduzido por um mais facilmente manipulável
    - Ex. converter data de nascimento para idade
    - Ex. converter temperatura de ºF para ºC
    - Ex. localização por GPS para código postal
  - Aplicação de função:
    - Log, exp, raiz, seno, etc.

Redução de Dimensionalidade

- Maldição da dimensionalidade
  - O número de exemplos necessários para representar o espaço de características cresce exponencialmente com o número de atributos.
- Dois tipos principais:
  - Agregação
    - Ex. PCA (Principal Component Analysis)
  - Seleção de atributos
    - Filtros
    - Wrapper

Tratamento de Dados

Pré-processamento

- O desempenho da análise é afetado pela qualidade dos dados.
- Conjuntos de dados podem ter diferentes características, dimensões ou formatos.
  - Atributos numéricos vs simbólicos
  - Limpos vs com ruídos e imperfeições
  - Valores incorretos, inconsistentes, duplicados ou ausentes
  - Atributos independentes vs relacionados
  - Poucos vs muitos objetos e/ou atributos

O pré-processamento consiste em minimizar ou reduzir os problemas dos dados.

Fonte: http://artedosdados.blogspot.com.br/2014/12/pre-processamento-de-dados.html

Pré-processamento

- Benefícios:
  - Facilitar o posterior uso de técnicas de análise.
  - Obtenção de modelos mais fiéis à distribuição dos dados.
  - Redução de complexidade computacional.
  - Facilitar a interpretação dos padrões extraídos.

Tornar mais adequado para as técnicas.
ex. algumas trabalham somente com entradas numéricas

Melhorar qualidade

Tempo e custo

Pré-processamento

- Grupos de tarefas de pré-processamento:
  - Eliminação manual de atributos
  - Integração de dados
  - Amostragem de dados
  - Redução de dimensionalidade
  - Balanceamento de dados
  - Limpeza de dados
  - Transformação de dados

Eliminação manual de atributos

Há atributos que claramente não contribuem para o aprendizado.

id	Nome	Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Estado	Diagnóstico
4201	João	28	M	70	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

Eliminação manual de atributos

Dados após eliminação manual de atributos.

Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Diagnóstico
28	M	70	Grandes	38.0	2	Doente
18	F	67	Pequenas	39,5	4	Doente
49	M	92	Grandes	38,0	2	Saudável
18	M	43	Grandes	38,5	20	Doente
21	F	52	Médias	37,6	1	Saudável
22	F	72	Pequenas	38,0	3	Doente
19	F	87	Grandes	39,0	6	Doente
34	M	67	Médias	38,4	2	Saudável

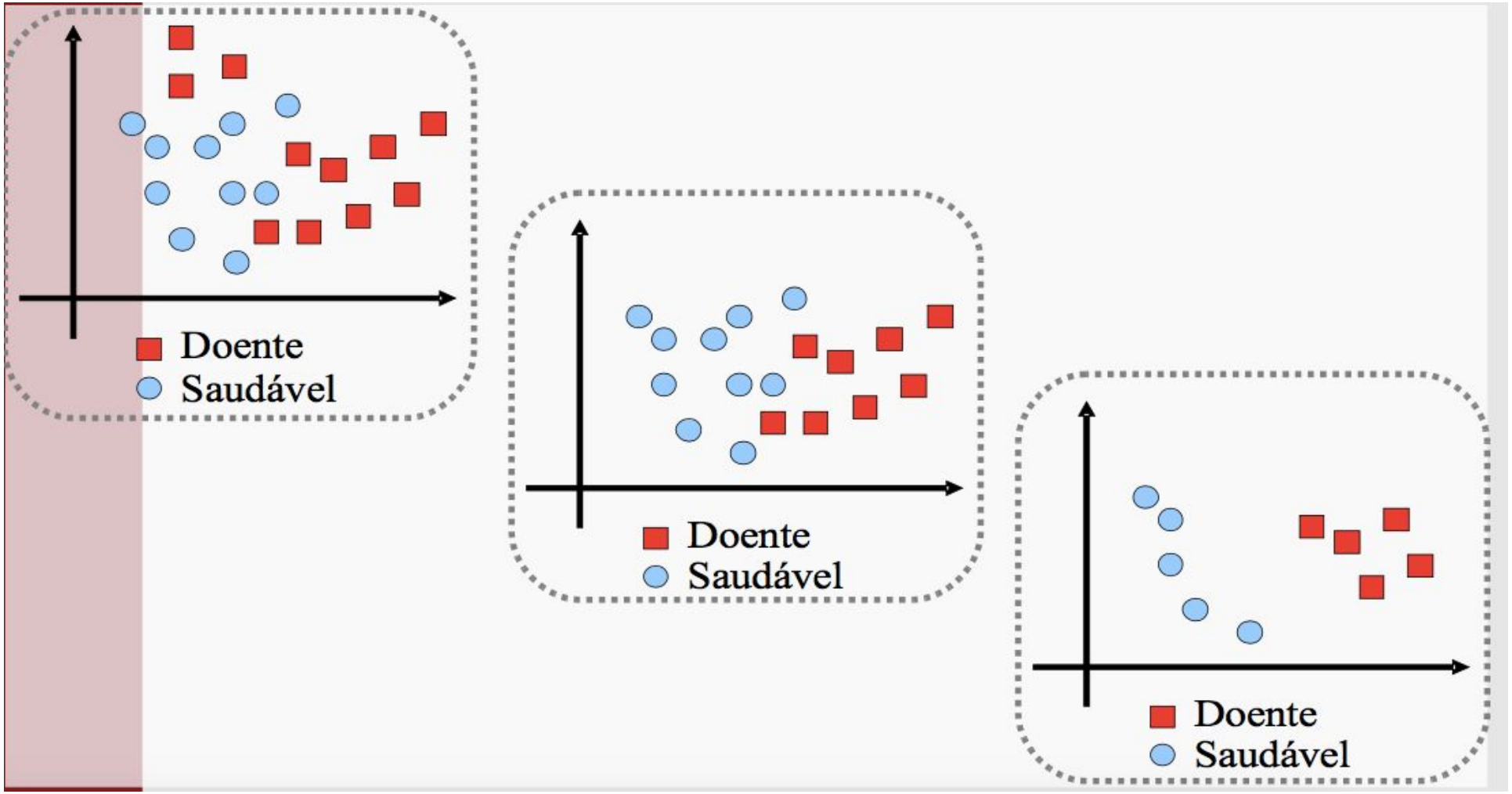
Eliminação manual de atributos

- Outro atributo irrelevante facilmente detectado:
  - Atributo que possui o mesmo valor para todos objetos.
  - Não traz informação para ajudar a distingui-los.
- Há ainda atributos irrelevantes de identificação não tão clara.
  - Técnicas de seleção de atributos podem ajudar a identificar.

Amostragem

- Técnicas de análise de dados podem ter dificuldades em lidar com um número grande de objetos.
  - Saturação de memória.
  - Aumento do tempo computacional para ajustar os parâmetros do modelo.
- Contudo, quanto mais dados, maior tende a ser a acurácia do modelo.
- Procurar balanço entre eficiência computacional e acurácia do modelo

Amostragem



Amostragem

- Amostra dos dados
  - Pode levar ao mesmo desempenho do conjunto completo, a menor custo computacional.
  - Deve ser representativa.
- Amostra representativa:
  - Aproximadamente as mesmas propriedades do conjunto de dados original.
  - Fornecer uma estimativa da informação contida na população original.
  - Uso deve ter efeito semelhante ao de toda a população.
  - Permitir conclusão do todo a partir de uma parte.

Amostragem

- Variações:
  - com reposição de exemplos e
  - sem reposição de exemplos.
- São semelhantes quando o tamanho da amostra é bem menor que o do conjunto original.

AMOSTRAGEM ALEATÓRIA SIMPLES

Amostragem

- Quando as classes têm propriedades diferentes.
  - Ex. números de objetos diferentes.
- Variações:
  - manter o mesmo número de objetos para cada classe ou manter o número proporcional ao original.

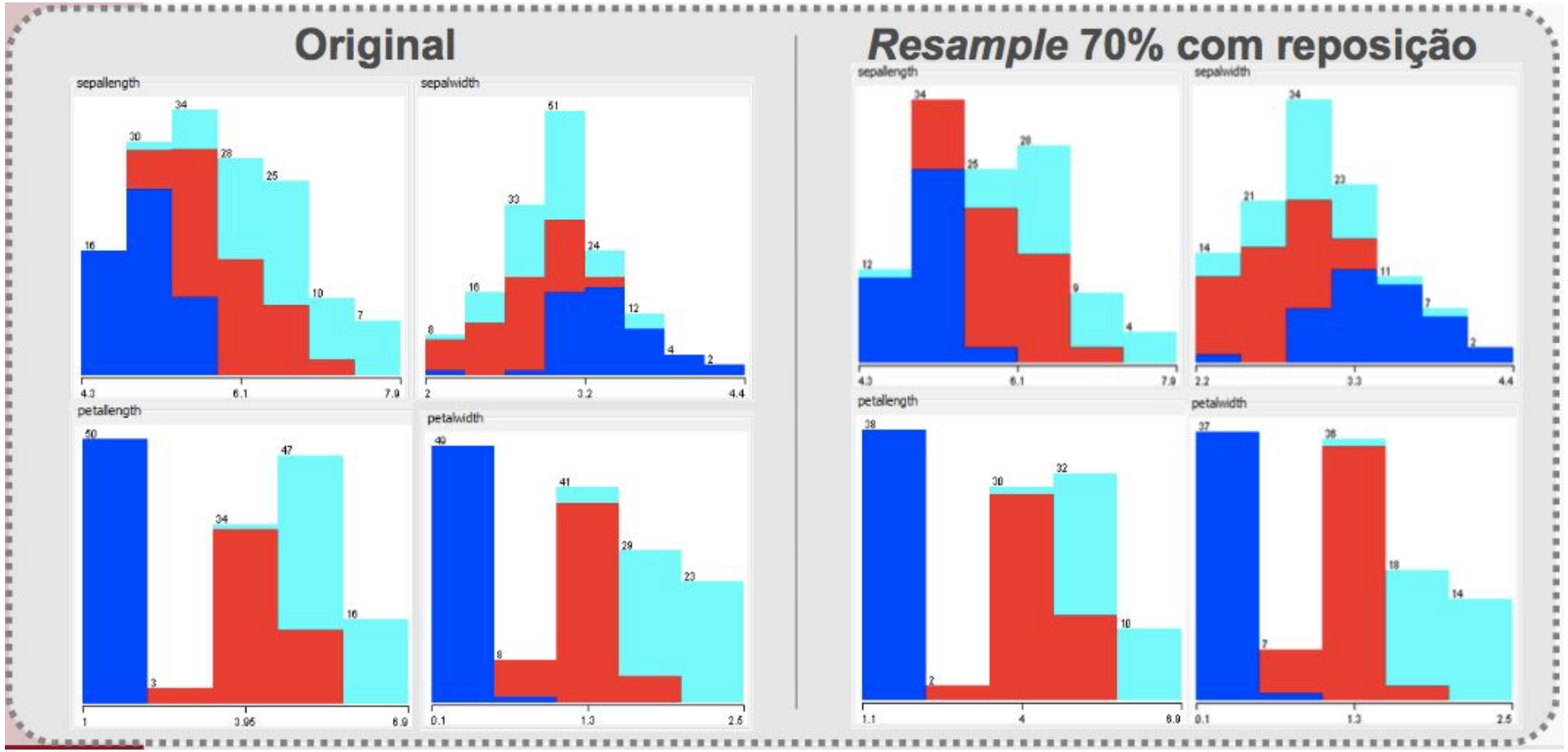
AMOSTRAGEM ESTRATIFICADA

Amostragem

- A acurácia preditiva continua a melhorar.
- Começa com amostra pequena e vai aumentando enquanto o modelo continuar a melhorar.

AMOSTRAGEM PROGRESSIVA

Amostragem


Original
sepallength
4.3
6.1
7.9
sepalwidth
2
3.2
4.4
petallength
1
3.95
6.9
petalwidth
0.1
1.3
2.5
Resample 70% com reposição
sepallength
4.3
6.1
7.9
sepalwidth
2.2
3.3
4.4
petallength
1.1
4
6.9
petalwidth
0.1
1.3
2.5

Balanceamento dos Dados

- Número de objetos varia para as diferentes classes
- Típico da aplicação
  - Ex. 80% dos pacientes que vão a um hospital estão doentes Problema na geração/coleta dos dados
- Classe majoritária
  - Contém a maior parte dos exemplos
- Classe minoritária
  - Tem o menor número de exemplos no conjunto de dados.

As técnicas de análise de dados tendem a favorecer a classe majoritária.

Balanceamento dos Dados

- Alternativas para lidar com dados desbalanceados:
  - Obter novos dados para a classe minoritária.
    - Na maioria dos casos não é possível...
- Balancear artificialmente o conjunto de dados:
  - Redefinir o tamanho do conjunto de dados.
  - Induzir um modelo para uma classe.

Limpeza dos Dados

- Qualidade dos dados:
- Em geral, os dados não foram produzidos para análise posterior.
  - Ruídos: erros ou valores diferentes do esperado.
  - Inconsistências: não combinam/contradizem valores de outros atributos no mesmo objeto.
  - Redundâncias: objetos/atributos com mesmos valores.
  - Dados incompletos: ausência de valores de atributos.

Limpeza dos Dados

- Exemplos de causas de erros:
  - Falha humana.
  - Falha no processo de coleta de dados.
  - Limitações do dispositivo de medição.
  - Má fé.
  - Valor de atributo muda com o tempo.

Dados Incompletos

Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Diagnóstico
	M	70		38.0		Doente
18	F	67	Pequenas	39,5	4	Doente
49	M	92	Grandes	38,0	2	
18		43	Grandes	38,5	20	Doente
21	F	52	Médias	37,6	1	Saudável
22	F	72	Pequenas		3	
	F	87	Grandes	39,0	6	Doente
34	M	67	Médias	38,4	2	Saudável

Dados Incompletos

- Possíveis causas:
  - Atributo não era importante quando os primeiros dados foram coletados.
    - Ex. e-mail na década de 90.
  - Desconhecimento do valor do atributo.
    - Ex. não saber o tipo sanguíneo de paciente em seu cadastro.
  - Falta de necessidade/obrigação de apresentar valor.
    - Ex. salário em hospital.
  - Inexistência de valor para o atributo.
    - Ex. número de partos para pacientes do sexo masculino.
  - Problema com equipamento para coleta, transmissão e armazenamento de dados.

Dados Incompletos

- Alternativas para lidar com valores ausentes:
  - Eliminar os objetos com valores ausentes
  - Definir e preencher manualmente os valores ausentes

  - Utilizar método/heurística para definir valores automaticamente

Não indicada quando número de atributos com valores ausentes varia muito entre os objetos ou quando muitos objetos têm valores ausentes

Não é factível para muitos valores ausentes

Alternativa mais usada

Dados Inconsistentes

- Possíveis causas:
  - Erro/engano
  - Presença de ruídos nos dados
  - Proposital (fraude)
  - Problemas na integração dos dados
    - Ex. conjuntos de dados com escalas diferentes para uma mesma medida.

Dados Inconsistentes

- Algumas inconsistências são de fácil detecção:
  - Violação de relações conhecidas entre atributos
    - Ex.: Valor de atributo A é sempre menor que valor de atributo B
    - Data de nascimento é posterior ao dia de hoje
  - Valor inválido para o atributo
    - Ex.: altura com valor negativo
- Em outros casos, informações adicionais precisam ser verificadas

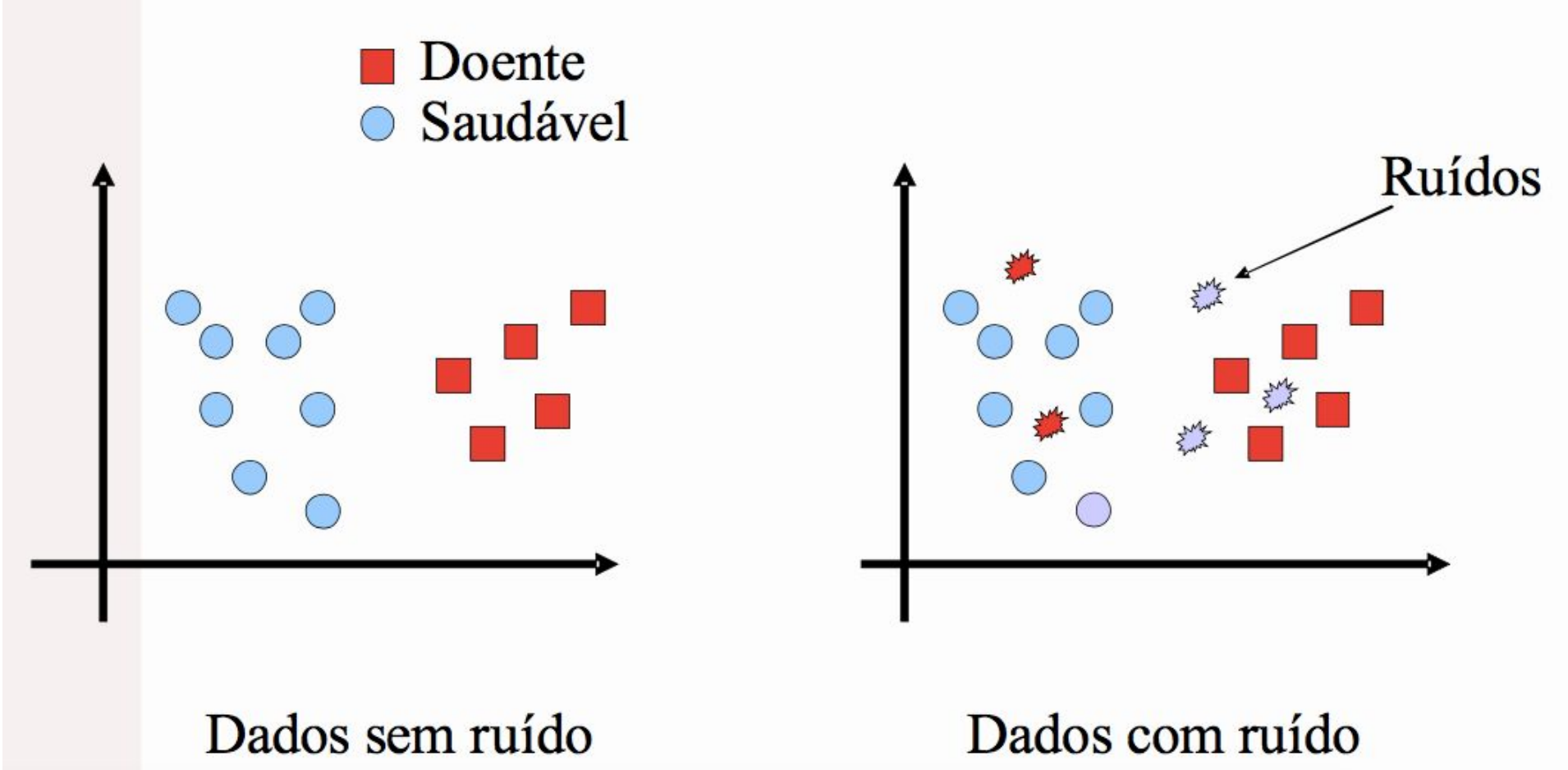
Dados Redundantes

- Objetos que não trazem informação nova.
- Elementos que podem ter seu valor deduzido a partir de outro atributo.
- Dados duplicados.
- Problemas:
  - Aumento do custo computacional.
  - Podem ser considerados mais importantes que outros.

Ruídos

- Objetos que aparentemente não pertencem à distribuição que gerou os dados
- Várias causas possíveis
- Podem levar a superajuste do modelo
  - Algoritmo pode se ater às especificidades dos ruídos
- A eliminação pode levar à perda de informação importante
  - Algumas regiões do espaço de atributos podem não ser consideradas

Ruídos



Outliers

- Valores que estão além dos limites aceitáveis ou são muito diferentes dos demais
  - Exceções
- Podem ser ruídos ou não.

Outliers

Nome	Profissão	Nível	Peso	Altura	Salário	Situação
João	Encanador	Médio	70	180	3000	Adimplente
Lia	Médico	Superior	200	174	7000	Inadimplente
Maria	Advogado	Médio	90	180	600	Adimplente
José	Médico	Superior	100	-6	2000	Inadimplente
Sérgio	Bancário	Superior	82	178	5000	Inadimplente
Ana	Professor	Fund.	77	188	1800	Adimplente
Luíza	Médico	Superior	100	-6	2000	Inadimplente

Outliers

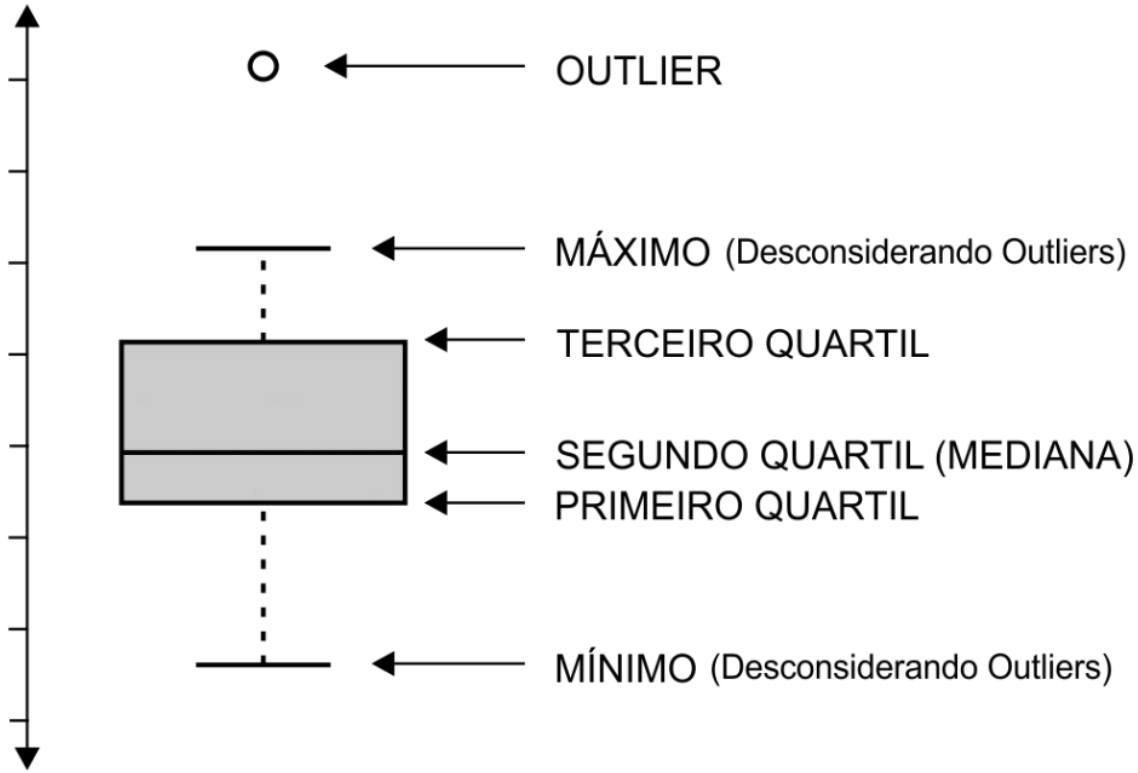
Nome	Profissão	Nível	Peso	Altura	Salário	Situação
João	Encanador	Médio	70	180	3000	Adimplente
Lia	Médico	Superior	200	174	7000	Inadimplente
Maria	Advogado	Médio	90	180	600	Adimplente
José	Médico	Superior	100	-6	2000	Inadimplente
Sérgio	Bancário	Superior	82	178	5000	Inadimplente
Ana	Professor	Fund.	77	188	1800	Adimplente
Luíza	Médico	Superior	100	-6	2000	Inadimplente

Visualização de Dados

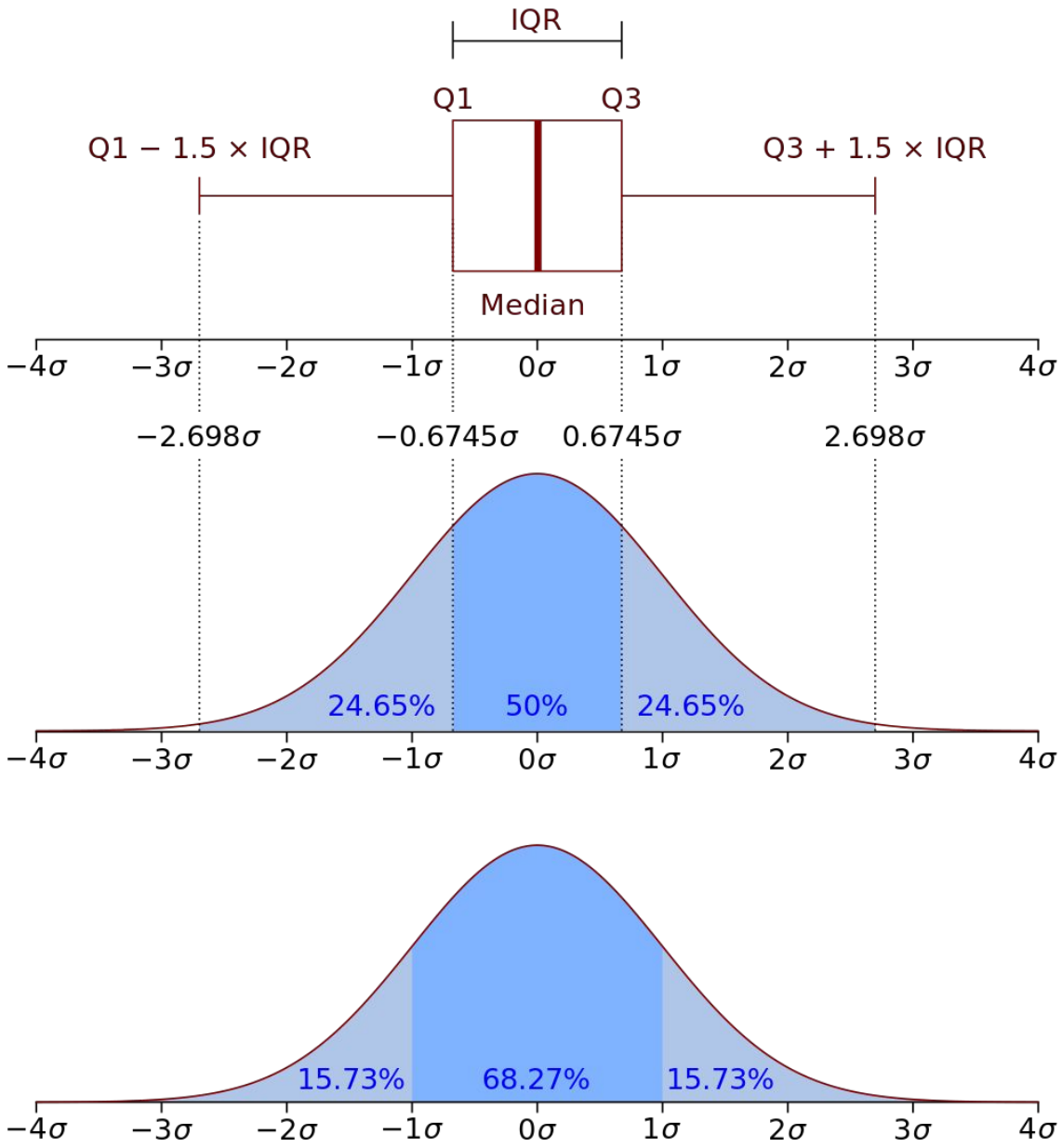
Boxplots

- 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

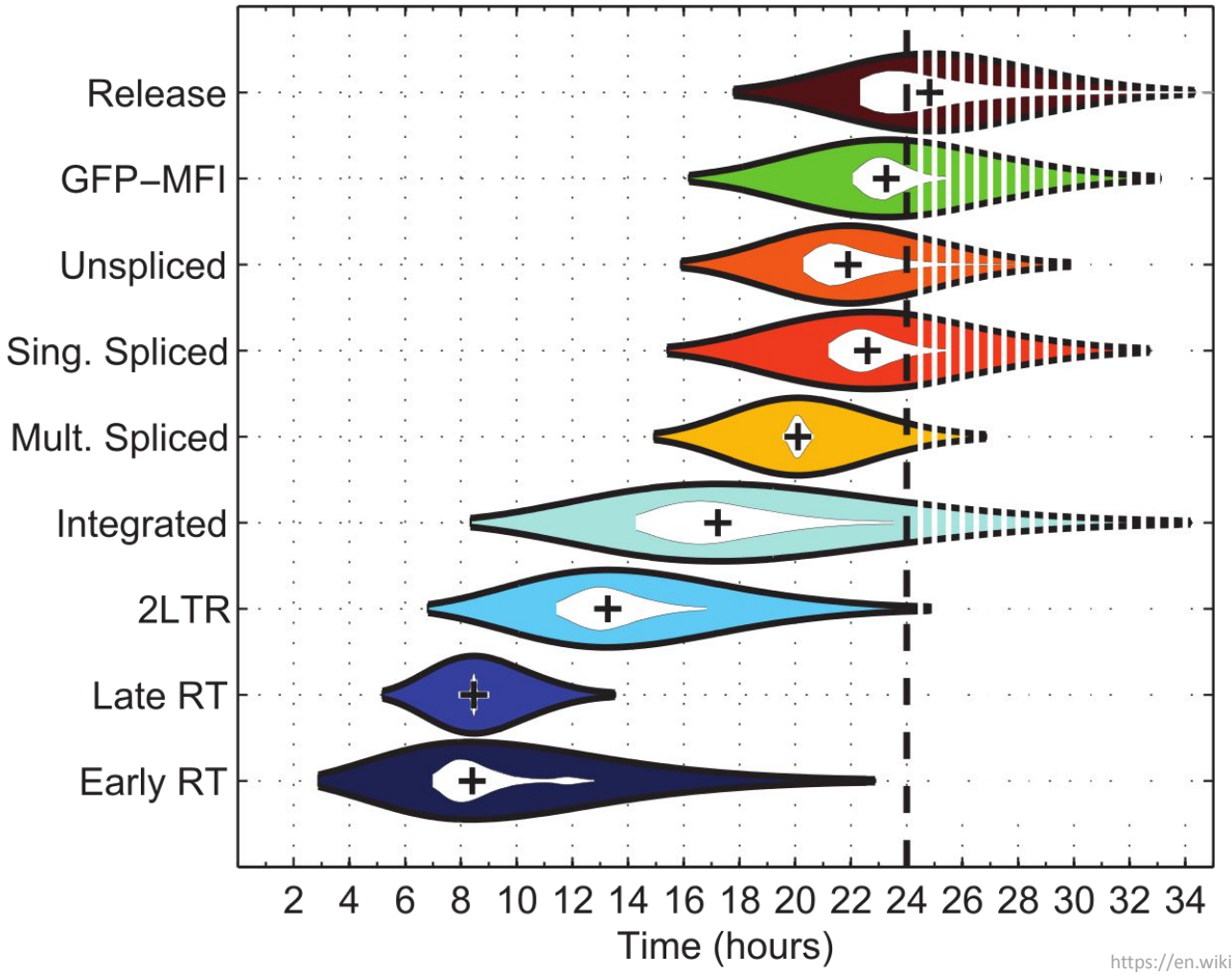


https://operdata.com.br/blog/como-interpretar-um-boxplot/



https://pt.wikipedia.org/wiki/Distribuição_normal#/media/Ficheiro:Boxplot_vs_PDF.svg

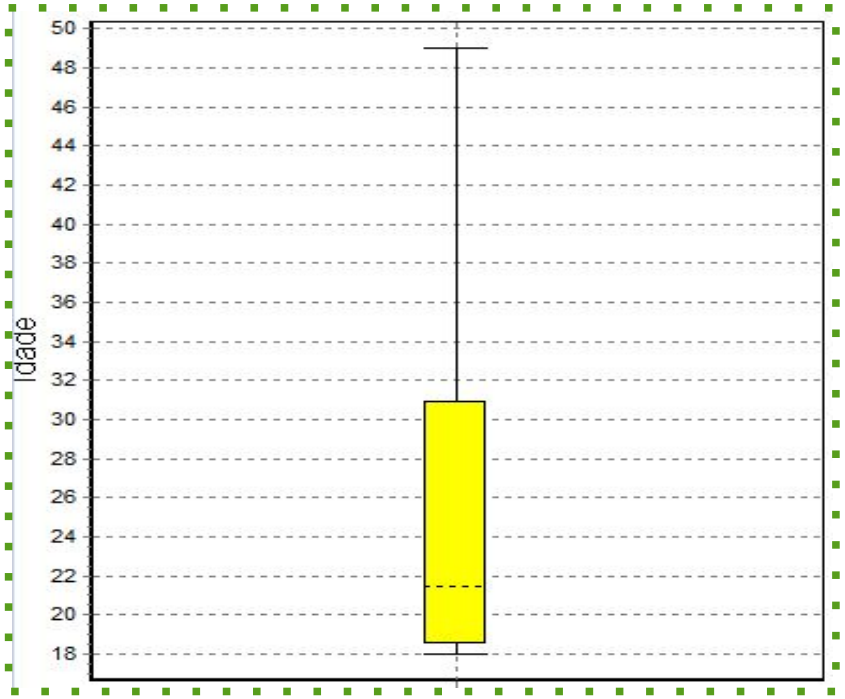
Violinplot



https://en.wikipedia.org/wiki/Violin_plot#/media/File:Violinplot-hiv-paper-plot-pathogens.svg

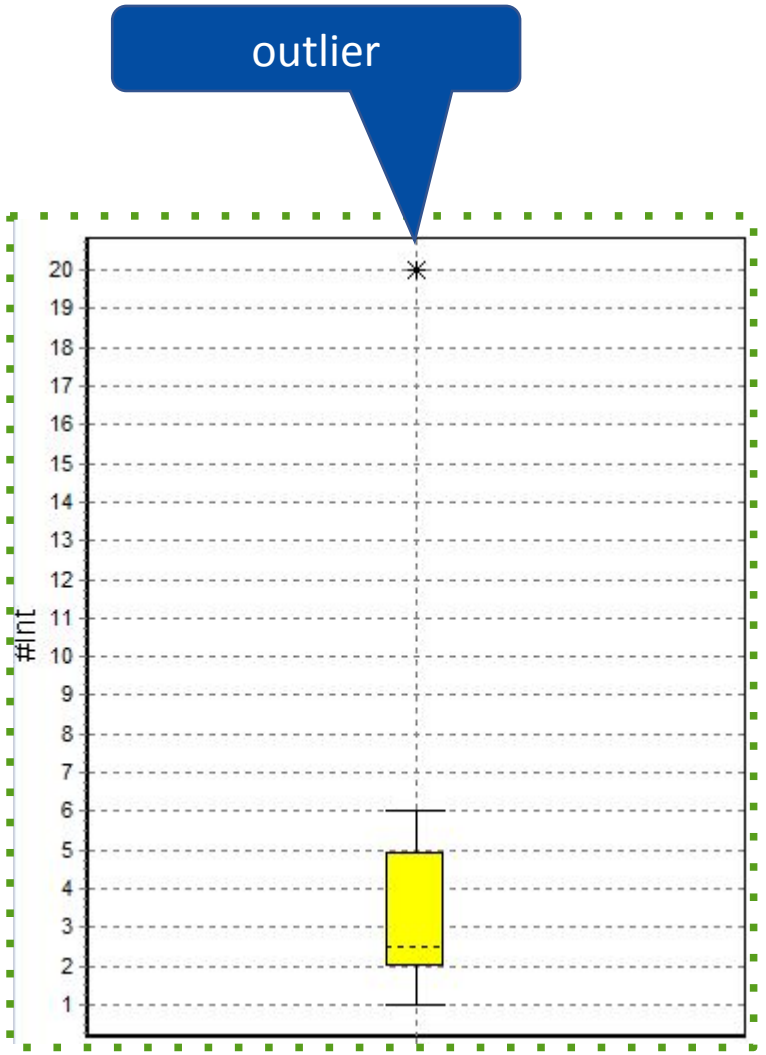
Boxplot

Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Diagnóstico
28	M	70	Grandes	38.0	2	Doente
18	F	67	Pequenas	39,5	4	Doente
49	M	92	Grandes	38,0	2	Saudável
18	M	43	Grandes	38,5	20	Doente
21	F	52	Médias	37,6	1	Saudável
22	F	72	Pequenas	38,0	3	Doente
19	F	87	Grandes	39,0	6	Doente
34	M	67	Médias	38,4	2	Saudável



Boxplot

Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Diagnóstico
28	M	70	Grandes	38.0	2	Doente
18	F	67	Pequenas	39,5	4	Doente
49	M	92	Grandes	38,0	2	Saudável
18	M	43	Grandes	38,5	20	Doente
21	F	52	Médias	37,6	1	Saudável
22	F	72	Pequenas	38,0	3	Doente
19	F	87	Grandes	39,0	6	Doente
34	M	67	Médias	38,4	2	Saudável

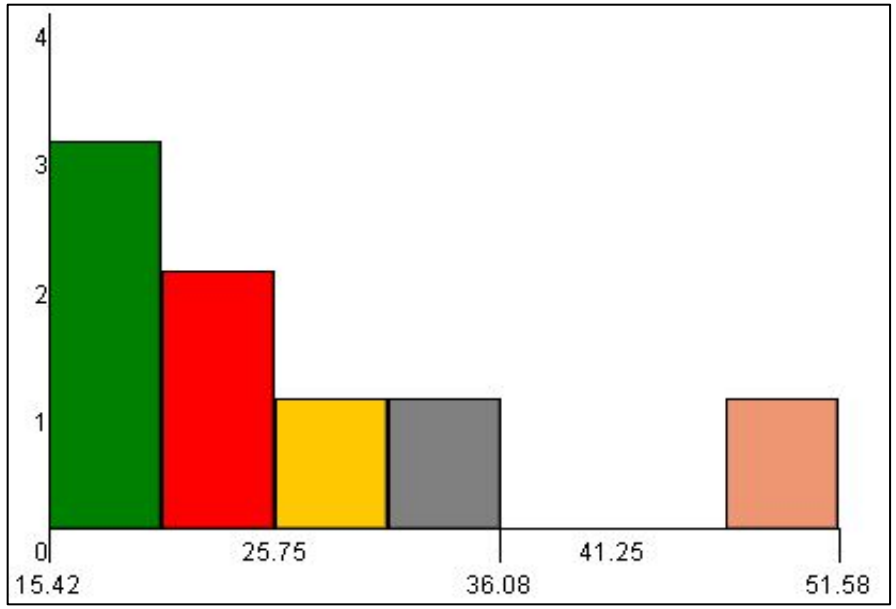


Histograma

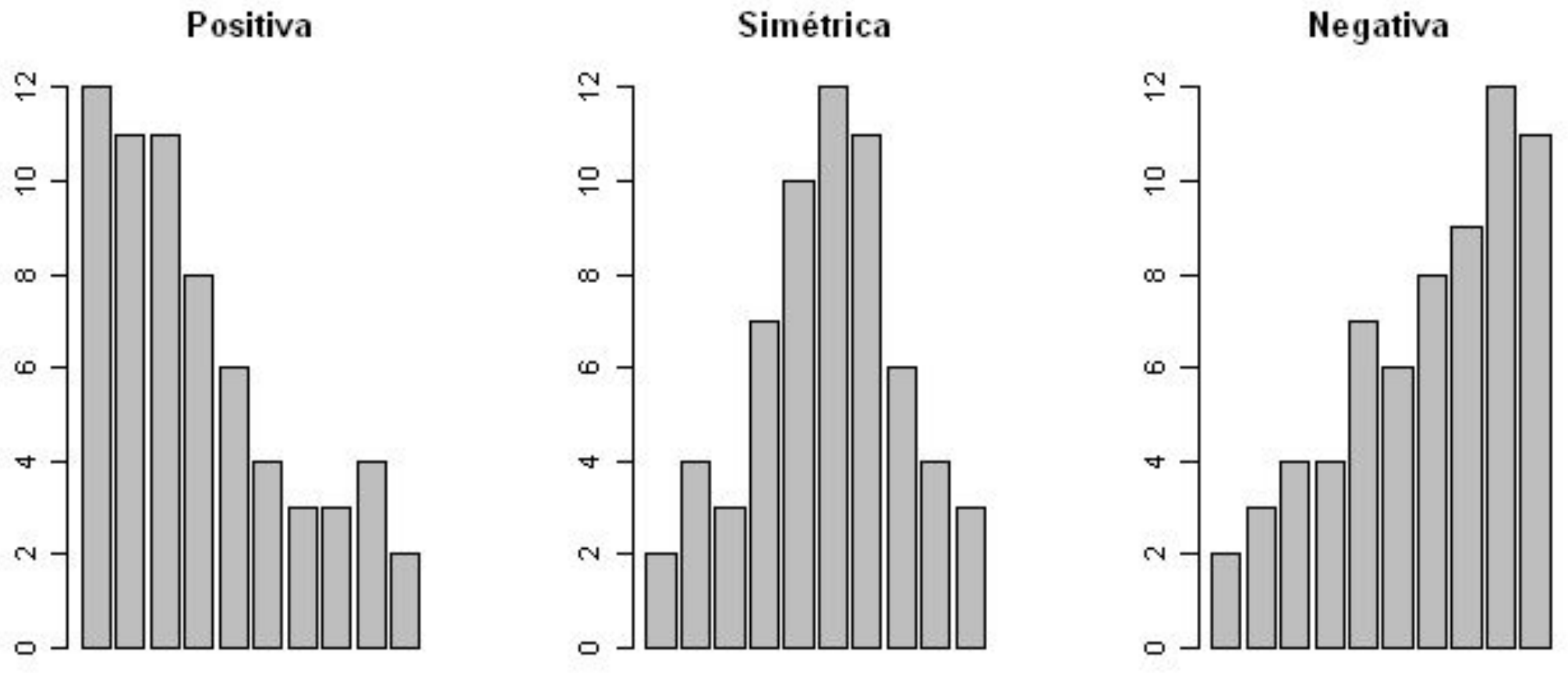
- Forma gráfica para visualizar distribuição: histograma
  - Divide valores em cestas
    - Valores categóricos: cada valor é uma cesta
    - Valores numéricos: divisão em intervalos contíguos de mesmo tamanho e cada intervalo é uma cesta
  - Para cada cesta, desenha uma barra com altura proporcional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

Histograma

Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Diagnóstico
28	M	70	Grandes	38.0	2	Doente
18	F	67	Pequenas	39,5	4	Doente
49	M	92	Grandes	38,0	2	Saudável
18	M	43	Grandes	38,5	20	Doente
21	F	52	Médias	37,6	1	Saudável
22	F	72	Pequenas	38,0	3	Doente
19	F	87	Grandes	39,0	6	Doente
34	M	67	Médias	38,4	2	Saudável



Obliquidade



Curtose

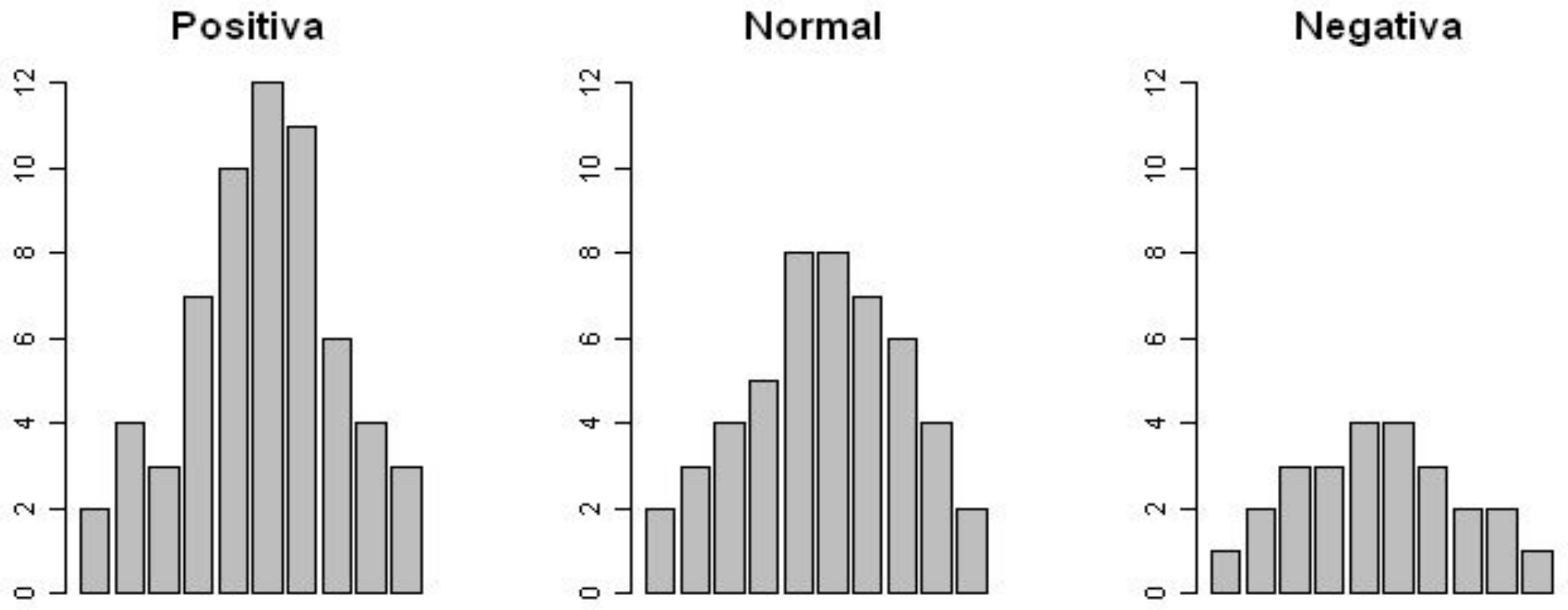
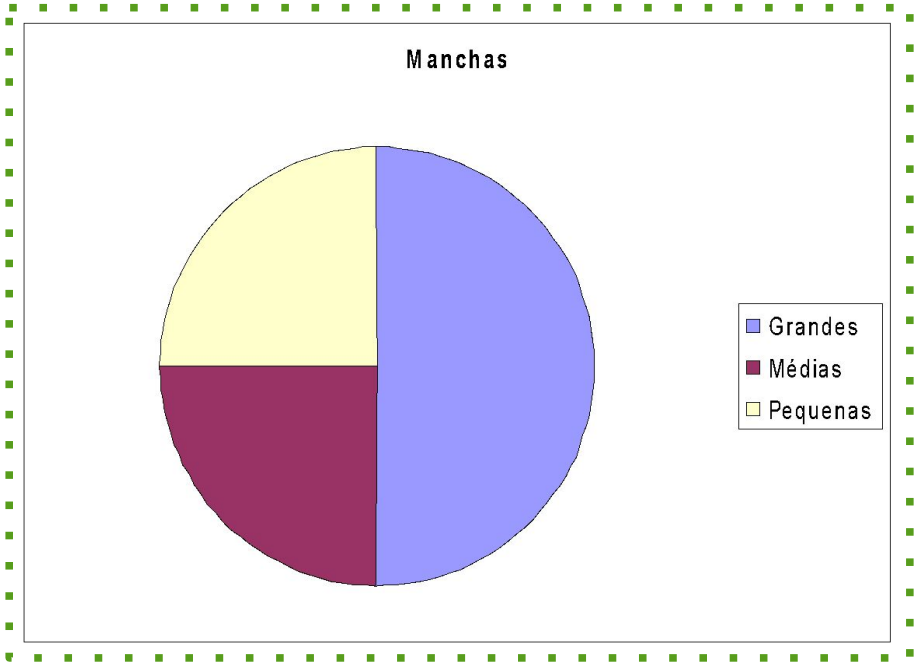


Gráfico de pizza/donut

- 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

Idade	Sexo	Peso	Manchas	Temperatura	# Internações	Diagnóstico
28	M	70	Grandes	38.0	2	Doente
18	F	67	Pequenas	39,5	4	Doente
49	M	92	Grandes	38,0	2	Saudável
18	M	43	Grandes	38,5	20	Doente
21	F	52	Médias	37,6	1	Saudável
22	F	72	Pequenas	38,0	3	Doente
19	F	87	Grandes	39,0	6	Doente
34	M	67	Médias	38,4	2	Saudável



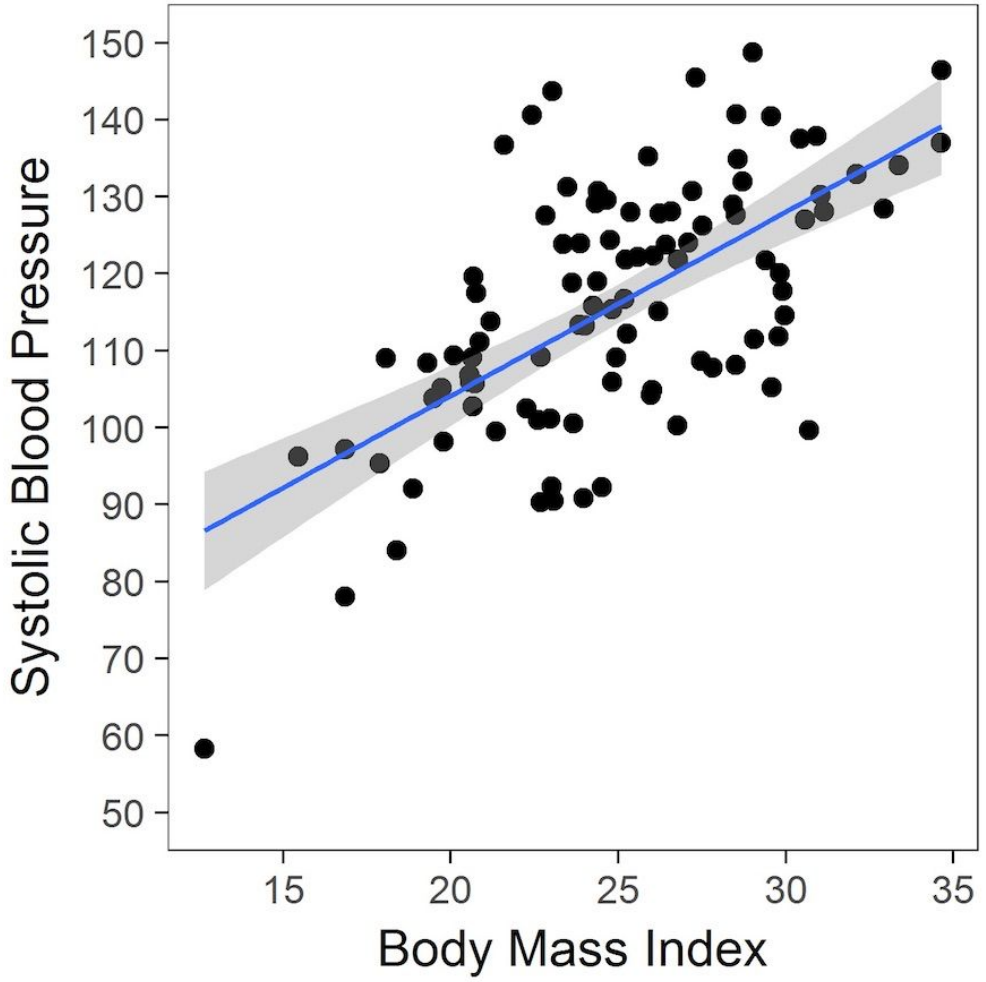
Dados multivariados: visualização

- 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 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



Scatter plot

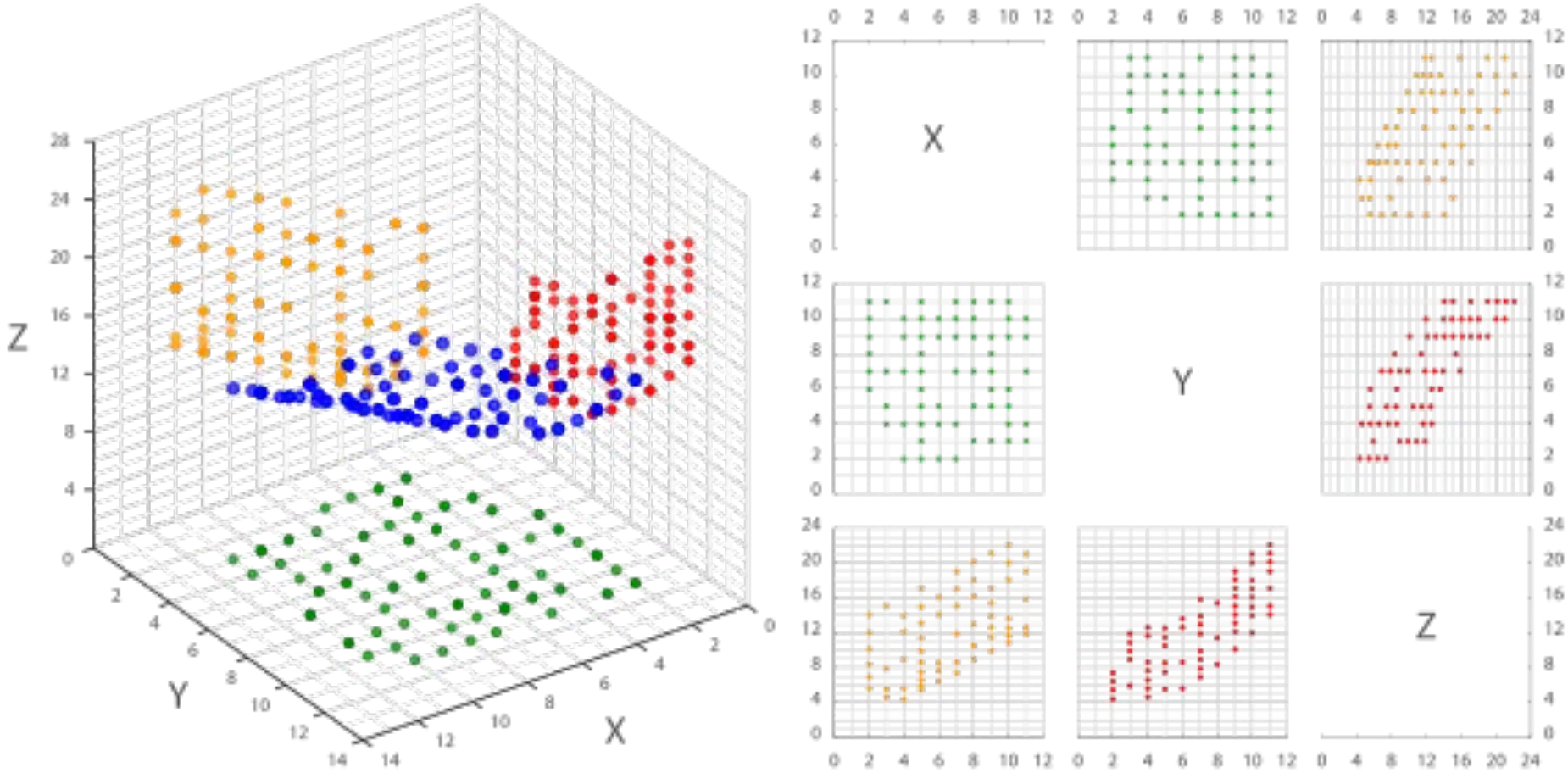
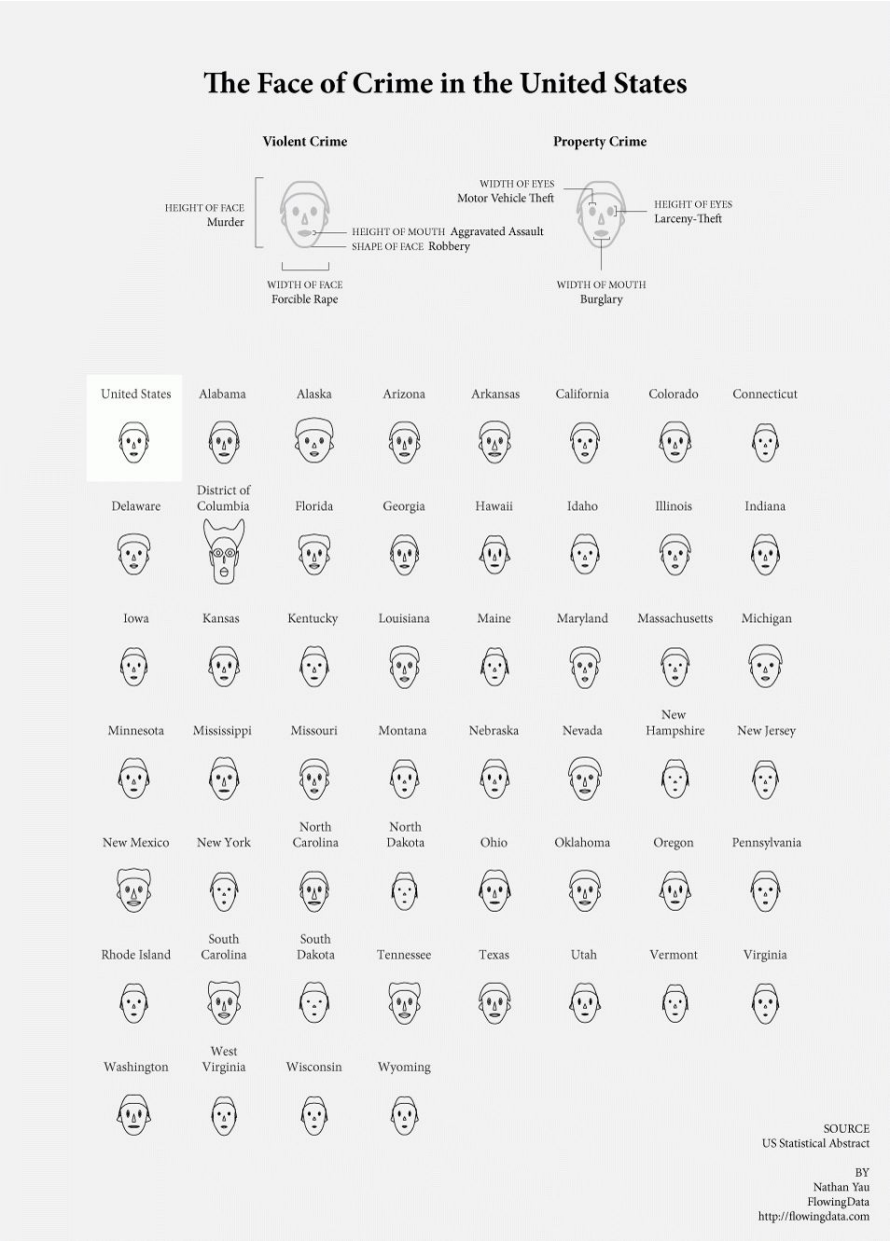


Diagrama de Chernoff

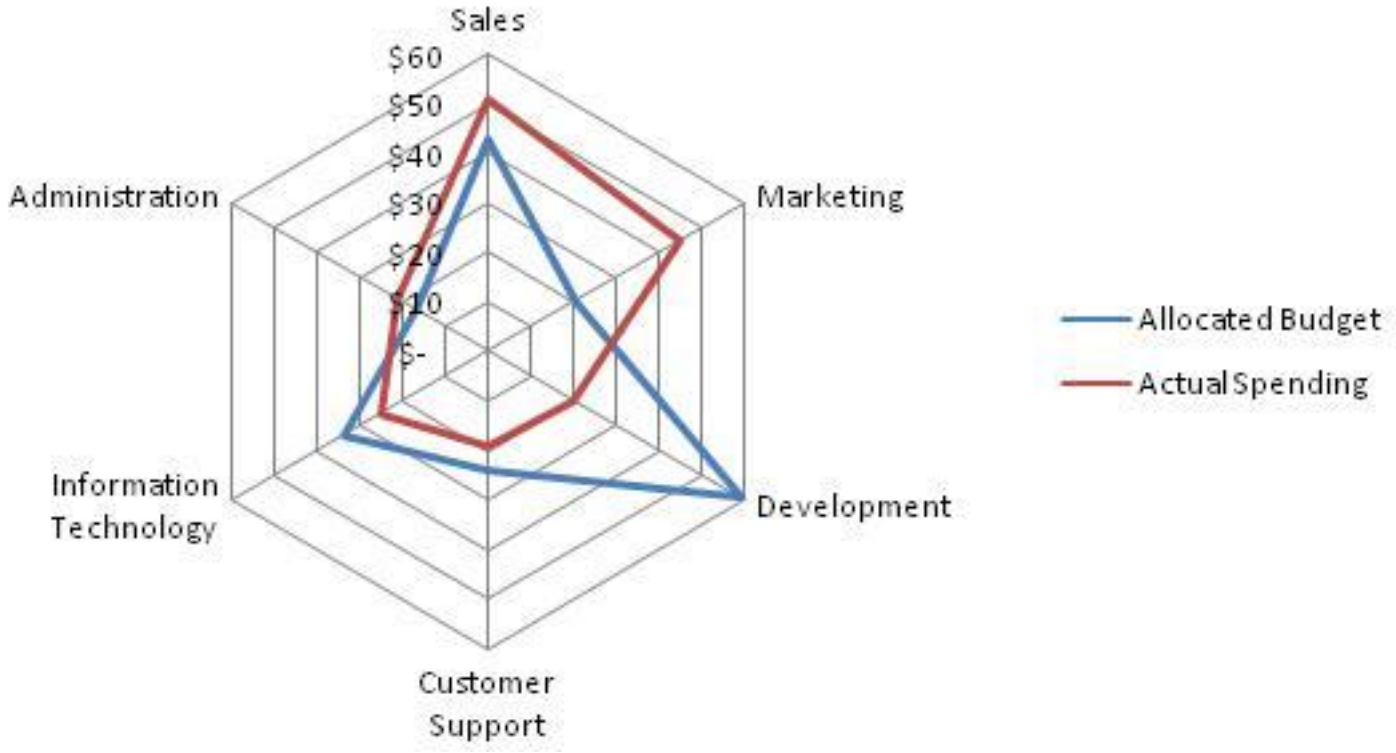
- Mapeia valores dos atributos para imagens mais familiares: faces
  - Cada objeto é representado por uma face
  - Cada atributo é associado a uma ou mais características da face
    - Ex. altura e largura da cabeça, da boca, etc.
- Baseia-se na habilidade humana de distinguir faces


The Face of Crime in the United States
Violent Crime
Property Crime
HEIGHT OF FACE
Murder
HEIGHT OF MOUTH Aggravated Assault
SHAPE OF FACE Robbery
WIDTH OF FACE
Forcible Rape
WIDTH OF EYES
Motor Vehicle Theft
HEIGHT OF EYES
Larceny-Theft
WIDTH OF MOUTH
Burglary
United States
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming
SOURCE
US Statistical Abstract
BY
Nathan Yau
FlowingData
http://flowingdata.com

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

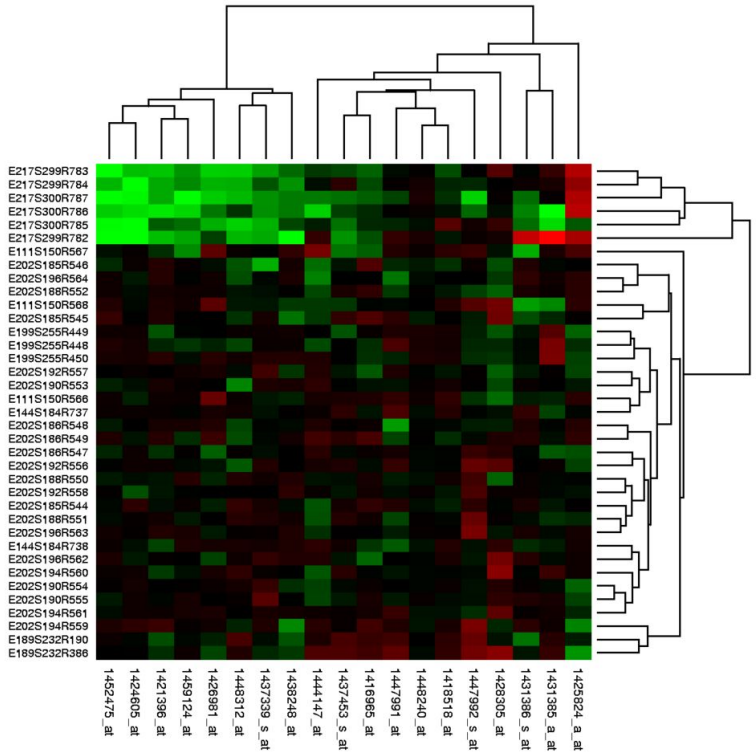
Star plot



https://pt.wikipedia.org/wiki/Gráfico_de_radar#/media/Ficheiro:Spider_Chart2.jpg

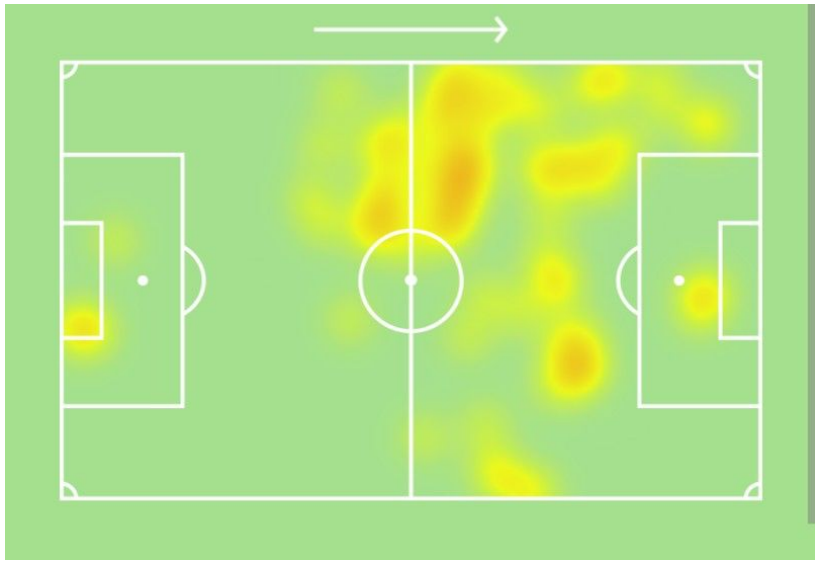
Heatmap

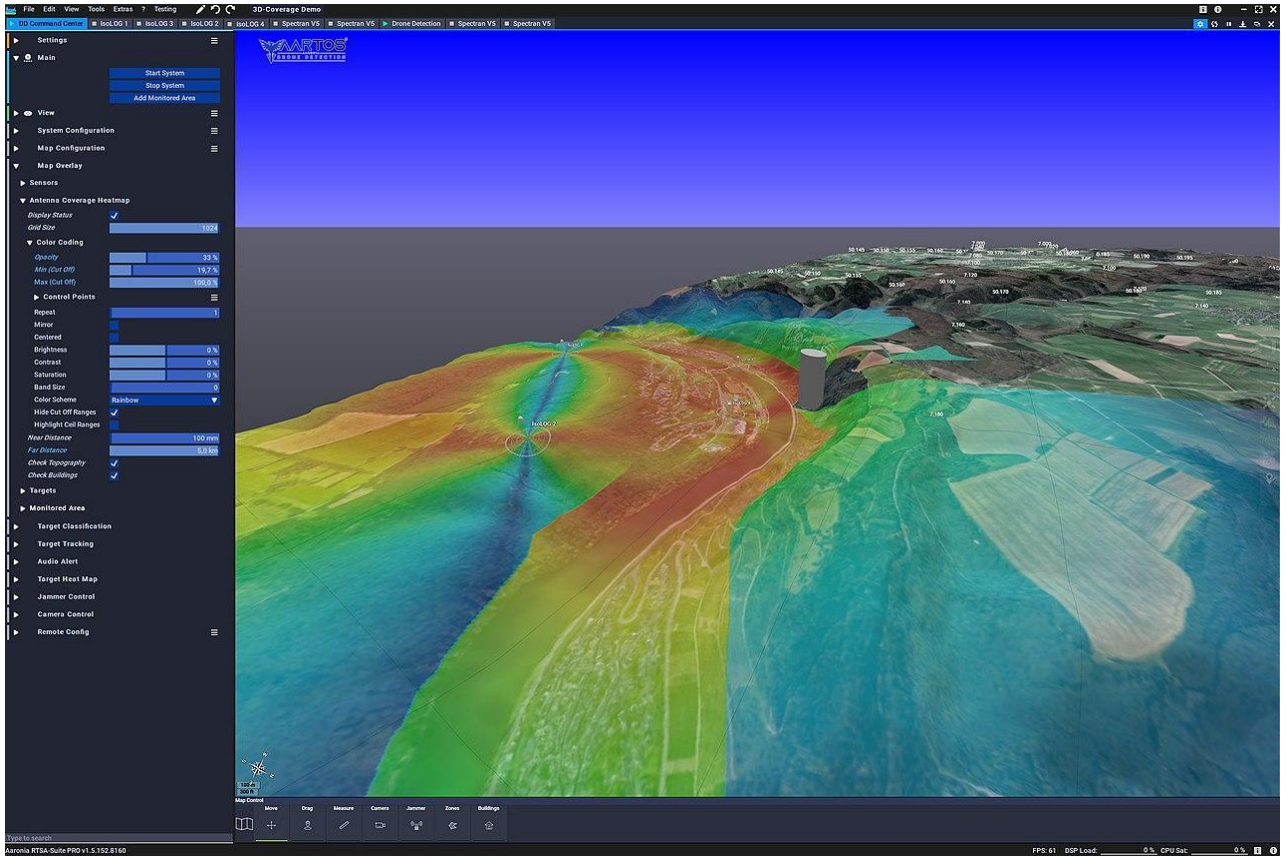
- Representa relação entre exemplos e as classes
  - Agrupamento hierárquico (dendograma)
    - Auxilia a verificar tendências nos dados
  - Ex. conjunto de dados iris



https://en.wikipedia.org/wiki/Heat_map#/media/File:Heatmap.png

Heatmap





https://en.wikipedia.org/wiki/Heat_map#/media/File:A_3D_triangulation_coverage_heatmap.jpg

Fim