UNIVERSIDADE DE SÃO PAULO

ESCOLA POLITÉCNICA

Ciência dos Dados

PCS 5787
Caracterização dos dados

Graduação em Engenharia Elétrica

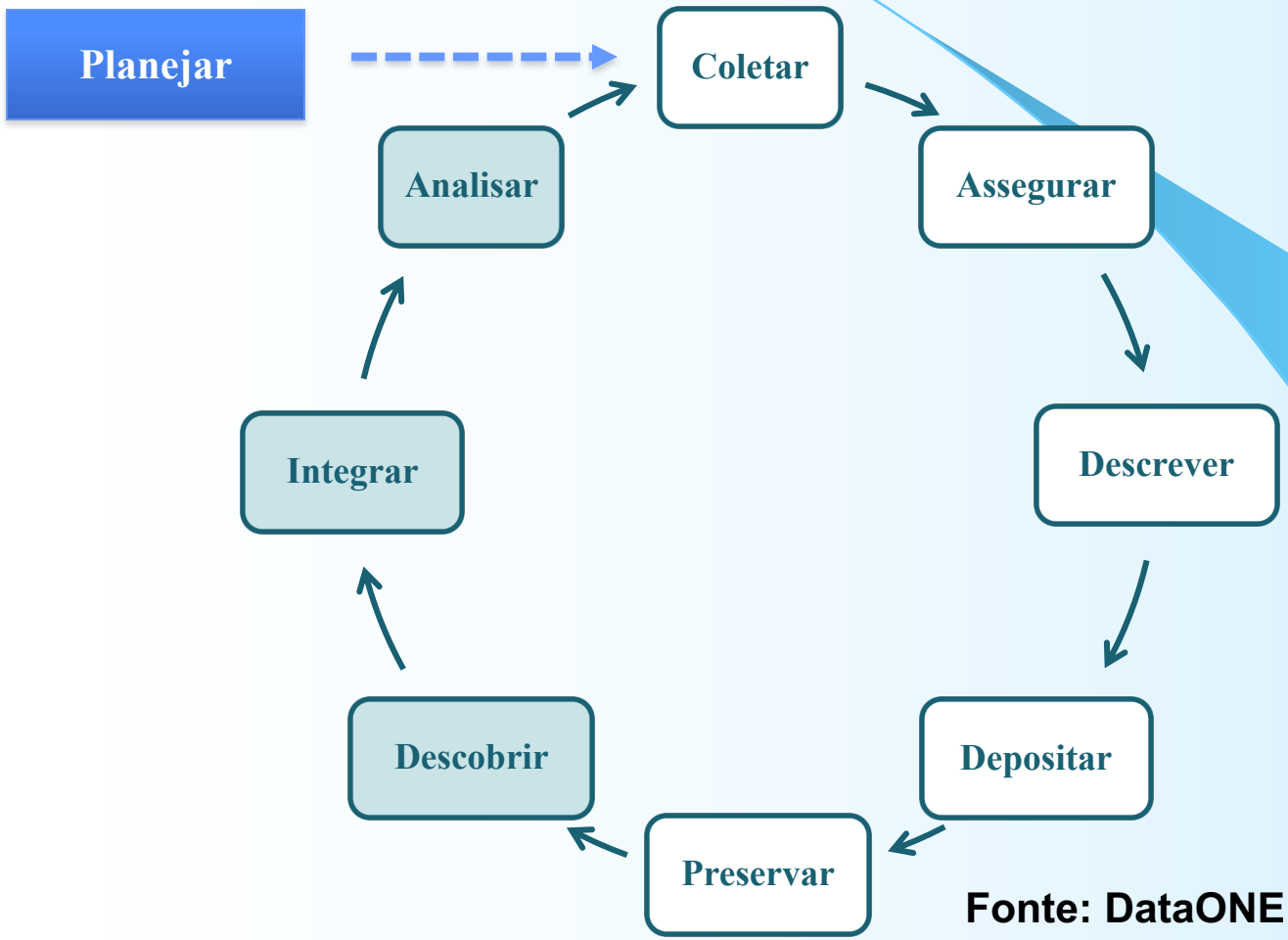
3o. Quadrimestre de 2020

Agenda

1. Introdução – Caracterização dos Dados;
2. Caracterização das amostras;
3. Análise de Dados – numéricos e qualitativos
4. Conclusões;
5. Referências.

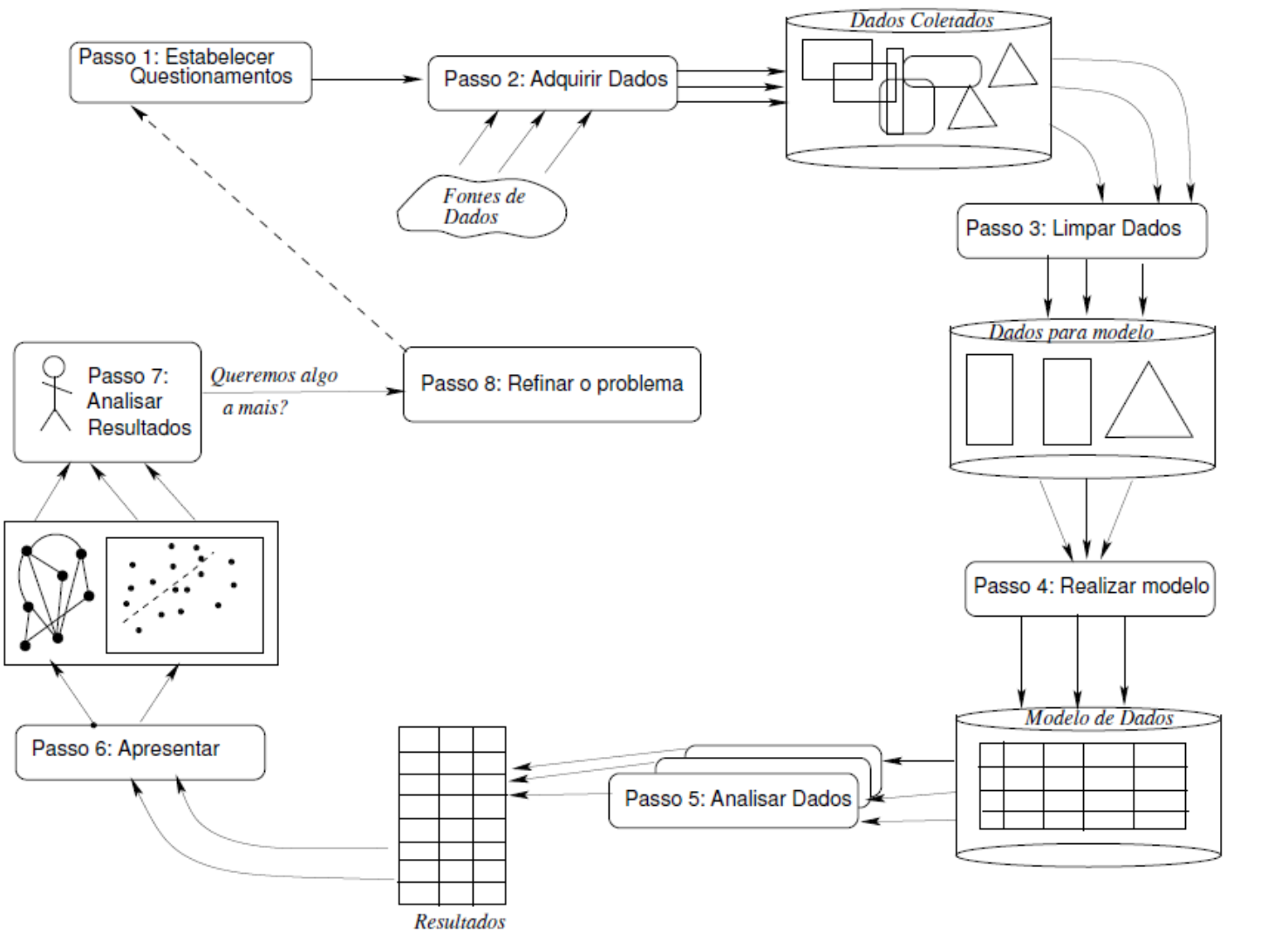
Introdução – Apoio as decisões tomadas desde o Planejamento até Integração

Ciclo de Vida dos Dados – DataONE



Fonte: DataONE Best Practices

Introdução – Passos 2 e 3



Cortesia: Alex Dehktyar

Caracterização dos Dados
(Planejamento, Passos 2 e 3)

- Conjunto de dados: objetos que podem representar um objeto/fenômeno do mundo físico.
- Representação Formal:

$X_{n \times d}$, sendo *n* nro. de objetos (eventos) e *d* nro de atributos

Caracterização dos Dados

- Exemplo: $Hospital_{4x10}$

Id	Nome	Idade	Sexo	Peso	Manchas	Temp.	#int	Estado	Diagnóstigo
2001	Paulo	21	M	91	Ausentes	38	2	SP	doente
3002	Carlos	26	M	80	Presentes	39	6	SP	Saudável
3002	Laura	26	F	60	Presentes	39	6	SP	Saudável
3002	João	26	M	75	Presentes	39	6	MG	Saudável

Caracterização dos Dados

- Tipos dos atributos: numérico ou simbólico (qualitativo);
- Extração de medidas descritivas:
  - Frequência;
  - Localização ou tendência (ex: média);
  - Dispersão ou espalhamento (ex: desvio padrão)
  - Distribuição ou formato

Tipos dos 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 quantitativos

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 Dados

- 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 Dados

- 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

Caracterização 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

Informações obtidas podem ajudar na seleção de técnicas apropriadas, seleção dos dados e pré-processamento

Caracterização dos 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 amostras 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 $\mathbf{x} = \{x_1, x_2, ..., x_n\}$

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	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

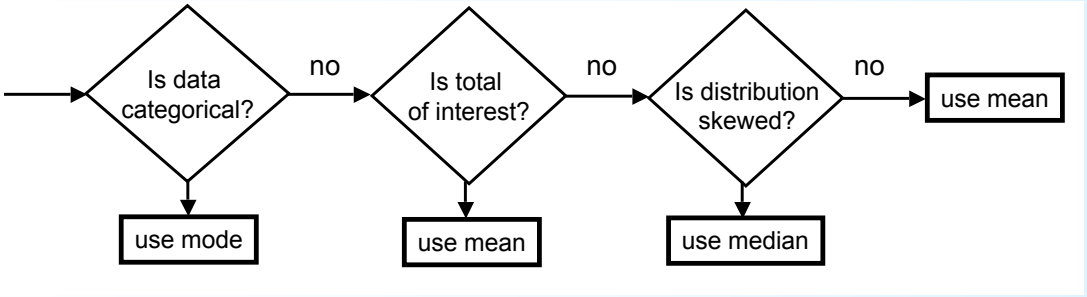
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Média Aritmética de Valore{$x_1, x_2, \ldots, x_n$}

Deve-se observar que média aritmética não é sempre apropriada

Problema: sensível a *outliers*

Bom indicador apenas se valores são distribuídos simetricamente



Médiana = 50 percentil

Moda = Mais frequente.

Mediana

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

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

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

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

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**

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

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

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

Intervalo

- 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

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

Espalhamento dos dados

- **Média não é suficiente se existe uma grande variabilidade nos valores**
- **Variabilidade de $\{x_1, x_2, \ldots, x_n\}$ é comumente caracterizada pela variância**

Variância da amostra = $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

- **Desvio padrão da amostra, s = sqrt(variança da amostra)**

—Mais significativa

- **Alternativas para resumir a variabilidade:**

—Intervalo: Máximo – Mínimo

—Interquartil Range (IQR) = ($P_{75\%}$ - $P_{25\%}$)

—Desvio Médio Absoluto (DMA) $= \frac{1}{n} \sum_{i=1}^{n} | x_i - \bar{x} |$

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

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

- 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
IQ: 14,3

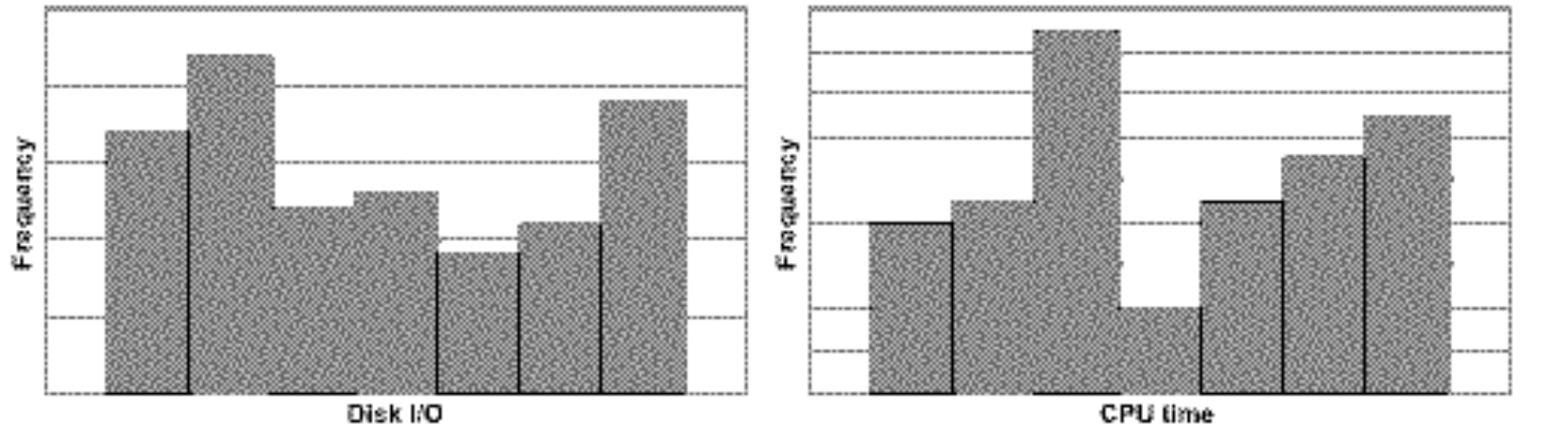
Outras medidas de espalhamento

- 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
IQ: 3,5

Dados univariados
Histograma para um parâmetro



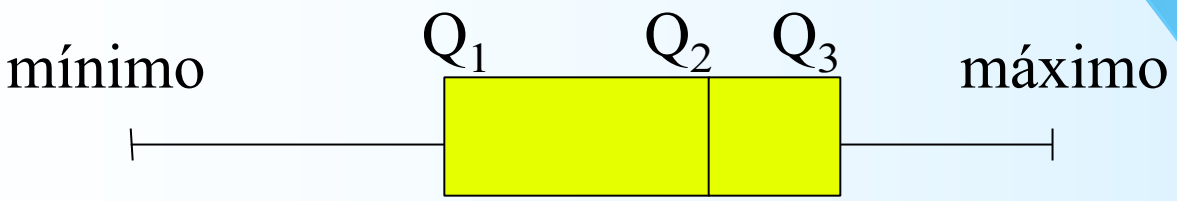
- **Frequências relativas de vários valores de um parâmetro —divide o intervalo completo em blocos**
  — soma observações que estão em um bloco (cesto)
- **Usos**
  —Simulação: gerar dados de testes de uma amostra de acordo com uma função de distribuição de probabilidade.
  —Modelo analítico: validar a função de distribuição de probabilidade

Desvantagens

—Muitos dados: *n* blocos, *m* parâmetros
– deve somente ser usado se a variança é alta e a média é imprópria
—Histograma de um parâmetro ignora a correlação entre os parâmetros.

Boxplots

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

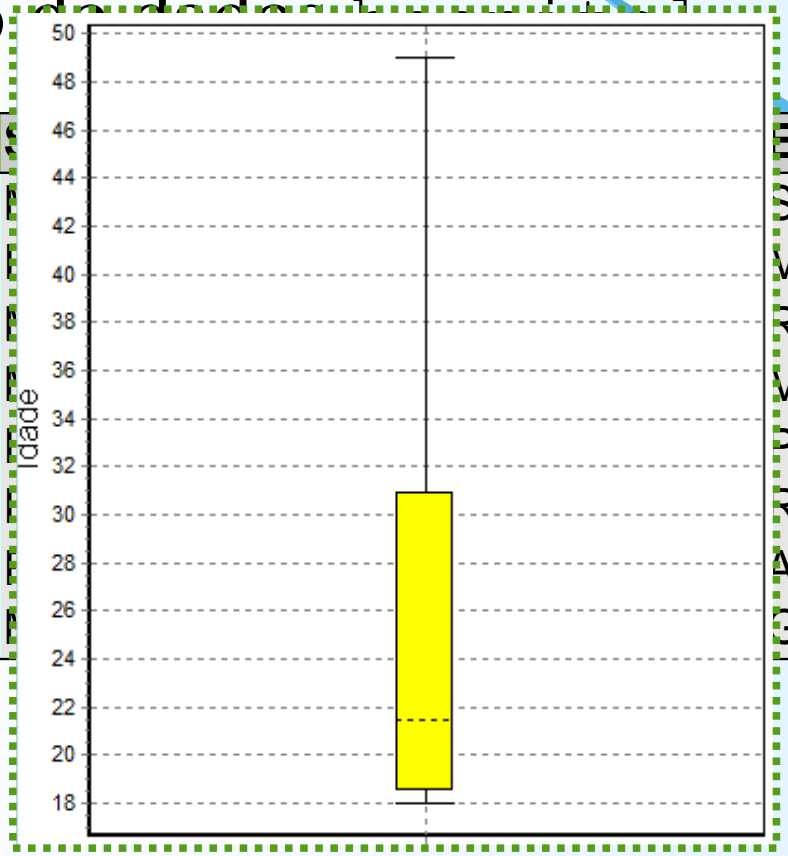


Boxplot modificado: limite superior/inferior vai até maior/menor valor apenas se esse valor não for muito distante do 3º/1º quartil (até 1,5 * intervalo entre quartis Q3 e Q1) Valores acima/abaixo são considerados *outliers*

Boxplot

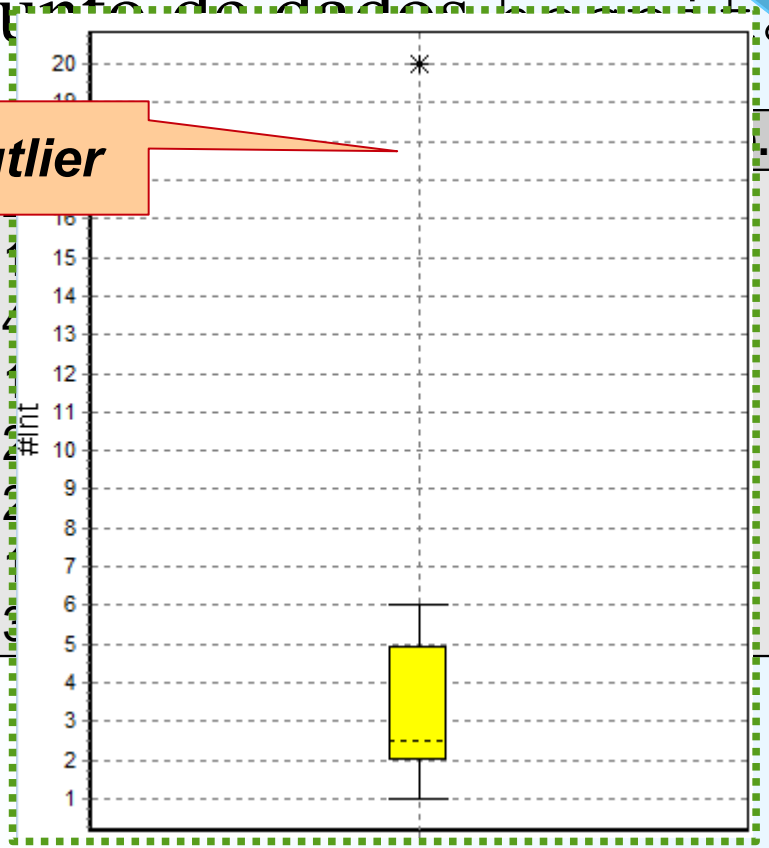
- Ex. conjunto de dados

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

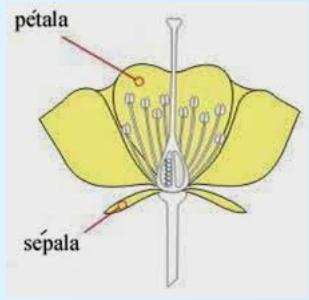
- Ex. conjunto de dados numérico



Outlier

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

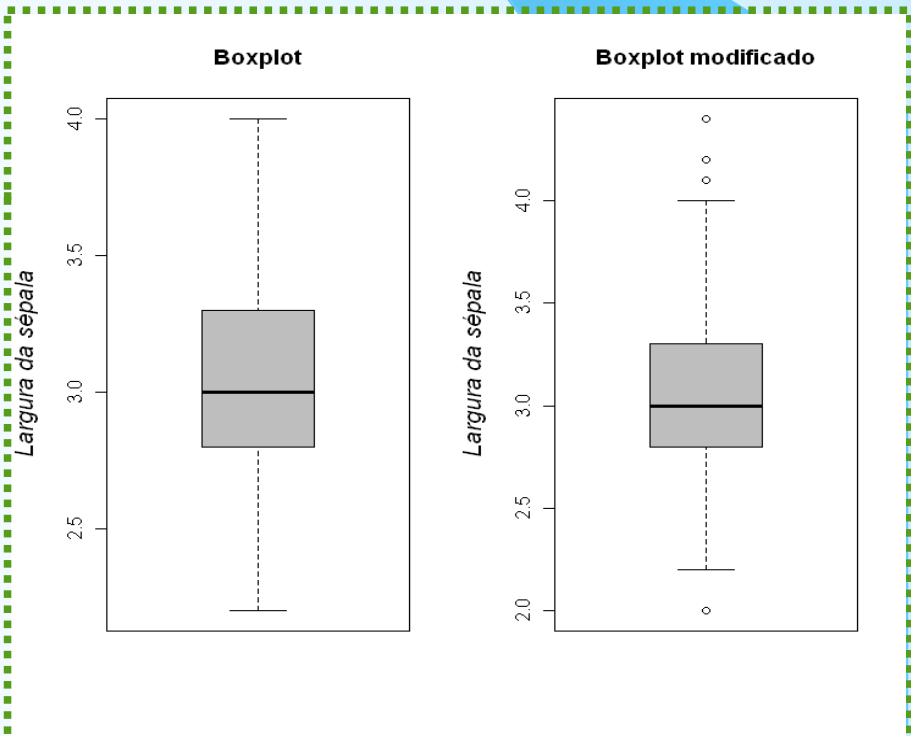


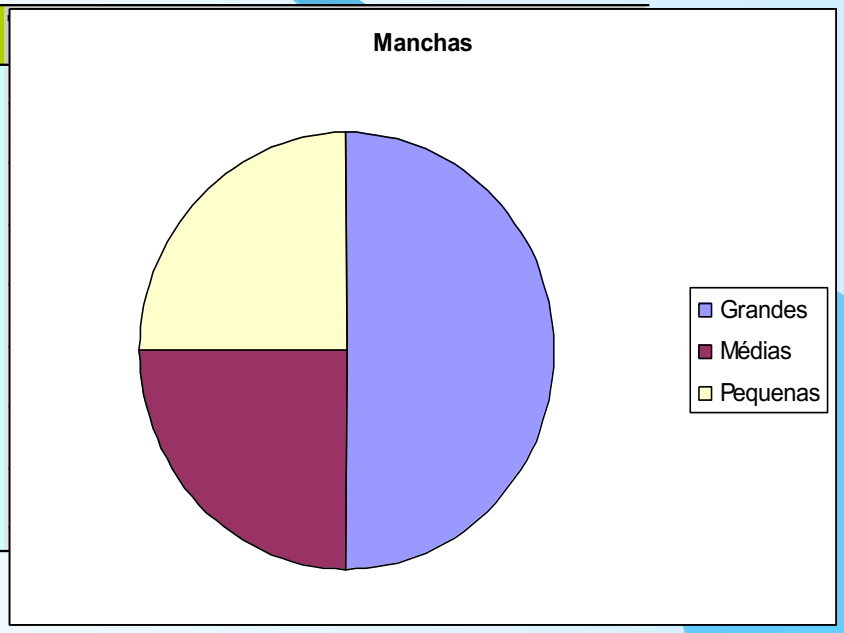
Gráfico de pizza

- 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

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

Observação: covariância(x^i, x^i) = variância(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

- Permitem análises da relação entre dois ou mais atributos, eliminando a dimensão (atributo)
  - Para variáveis contínuas, espalhamento é melhor capturado por uma matriz de

$$\text{Correlação}(x^i, x^j) = \frac{\text{covariância}(x^i, x^j)}{s_i s_j}$$

Observação: s_i **desvio padrão**

Covariância e correlação

- 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

Dados multivariados: visualização

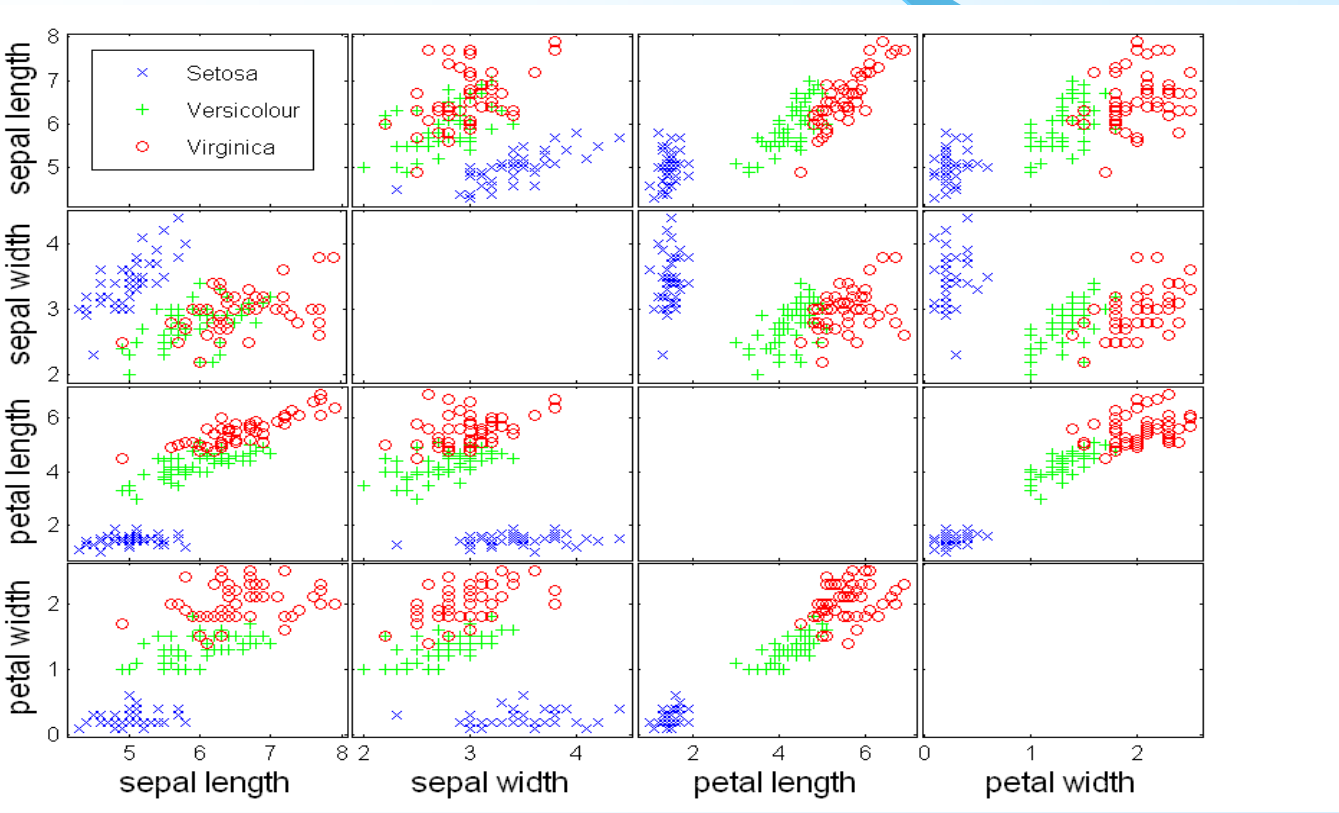
- Diagramas para visualizar dados multivariados
  - Em particular, relação entre diferentes atributos
  - Alguns tipos de gráficos:
    - *Scatter plot*
    - *Bag plots*
    - Faces de *Chernoff*
    - *Star plots*

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`

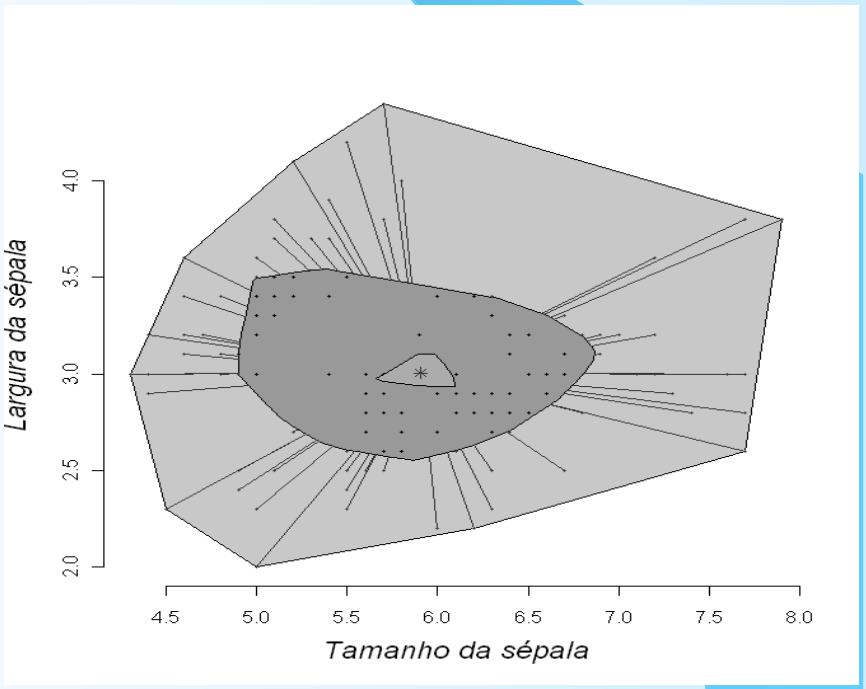
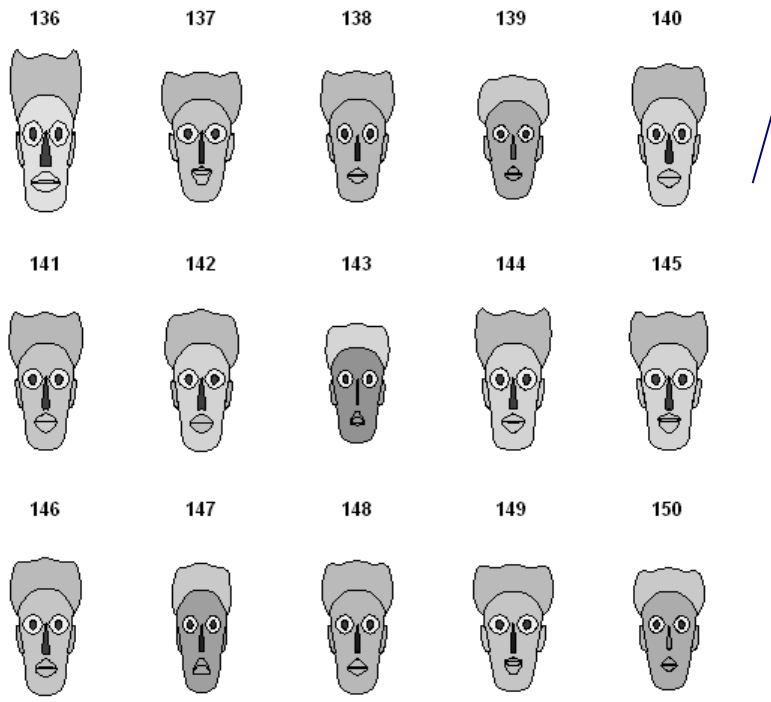


Diagrama de Chernoff

- 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`



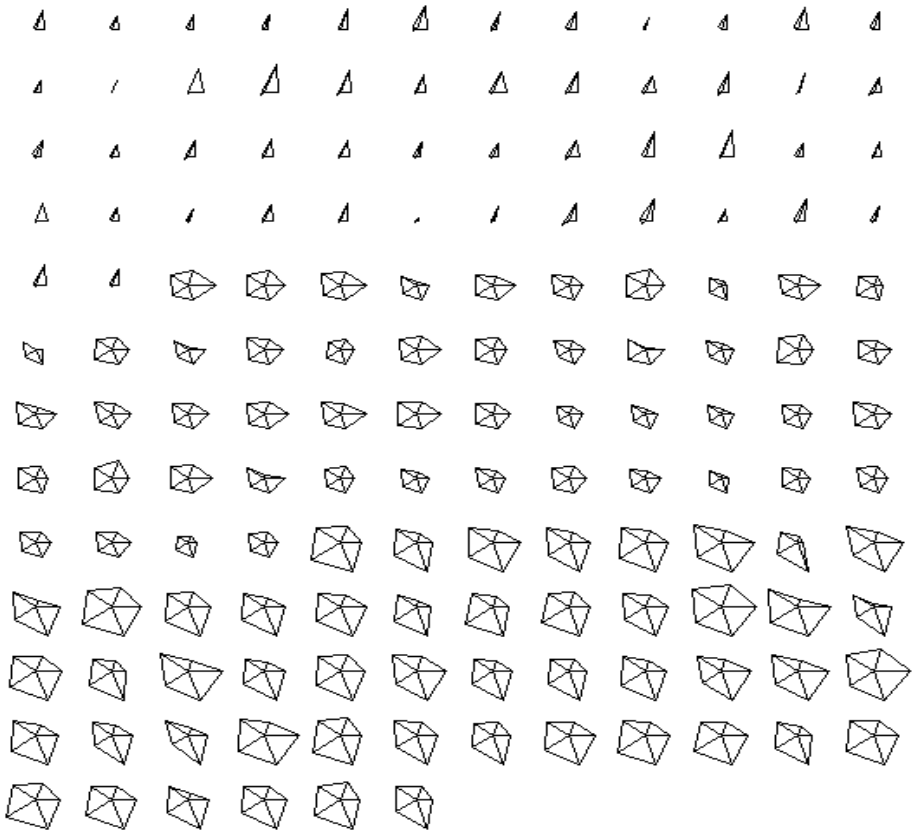
Tamanho da sépala representado por altura da face, largura da boca, altura do cabelo e largura do nariz

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`



Considerações finais

- Dados
  - Caracterização de dados
  - Tipos e escala de atributos
- Exploração de dados
  - Medidas de localidade, dispersão e distribuição
  - Técnicas podem ser usadas para seleção dos dados

Referências

- FACELI, K.; LORENA, A.C.;GAMA J.; CARVALHO, A.C.L.F. Inteligência Artificial: uma abordagem de aprendizado de máquina. Capítulos 1,2 e 3.
- JAIN R. The Art of Computer Systems Performance Analysis, John Wiley & Sons, 1991. Capítulos:1, 2, 3 e 5.
- slides baseados em apresentações de:
  - Prof Dr André C. P. L. F. Carvalho, ICMC-USP e Profa. Dra. Ana Carolina Lorena

UNIVERSIDADE DE SÃO PAULO

ESCOLA POLITÉCNICA

Ciência dos Dados

PCS 5787
Caracterização dos dados

Graduação em Engenharia Elétrica

3o. Quadrimestre de 2020