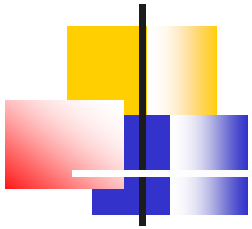
Ciência de Dados

Exploração de dados
PARTE II

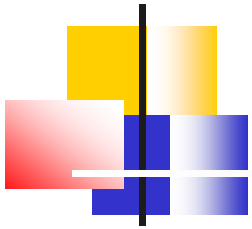
Profa. Roseli Ap. Francelin Romero – SCC

Material elaborado pelo Prof. Dr. André C. P. L. F. de Carvalho
Dr. Isvani Frias-Blanco e complementado por Roseli Romero
ICMC-USP



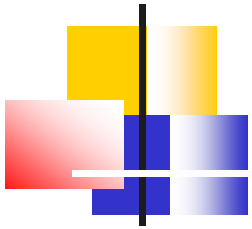
Tópicos

- Exploração de dados
  - Dados multivariados:
    - Matriz de Covariancia
    - Coeficiente de Pearson
    - Matriz de Correlação



Dados multivariados

- Possuem mais de um atributo
  - Cada atributo é uma variável
- Medidas de localização (tendência central)
  - Podem ser obtidas calculando medida de localização de cada atributo separadamente
  - Ex.: média, mediana, ...
    - Média dos objetos de um conjunto de dados com *m* atributos é dada por: $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_m)$



Dados multivariados

- Medidas de espalhamento (dispersão)
  - Podem ser calculadas para cada atributo independentemente dos demais
    - Usando qualquer medida de espalhamento
      - Intervalo, variância, desvio padrão
  - Para dados multivariados numéricos deve-se usar uma matriz de covariância
    - Cada elemento da matriz é a covariância entre dois atributos

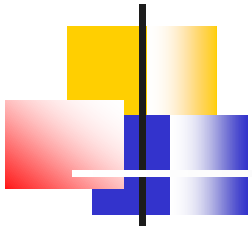
Dados multivariados

- Cálculo de cada elemento s_{ij} de uma matriz de covariância S para um conjunto de n objetos

$$s_{ij} = \text{covariância}\,(x_i, x_j) = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

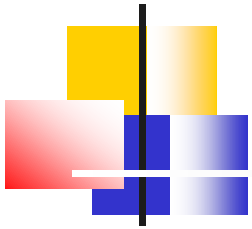
Onde:
$\overline{x_i}$: Valor médio do i-ésimo atributo
x_{ki} : Valor do i-ésimo atributo para o k-ésimo objeto

- Obs: covariância (x_i, x_i) = variância (x_i)
  - Matriz de covariância tem em sua diagonal as variâncias dos atributos



Dados multivariados

- Covariância de dois atributos
  - Mede o grau com que os atributos variam juntos
    - Valor próximo de 0:
      - Atributos não têm um relacionamento
    - Valor positivo:
      - Atributos diretamente relacionados
        - Quando o valor de um atributo aumenta, o do outro também aumenta
      - Valor negativo:
        - Atributos inversamente relacionados
    - Valor depende da magnitude dos atributos



Dados multivariados

- Covariância de dois atributos
  - É difícil avaliar o relacionamento entre dois atributos olhando apenas a covariância
    - Sofre influência da faixa de valores dos atributos
    - Correlação entre dois atributos ilustra mais claramente a força da relação entre eles
      - Mais popular que covariância
      - Elimina influência da faixa de valores

Dados Multivariados

- Matriz de Correlação
- Correlação Linear: Coeficiente de correlação de Pearson
- Correlação Não Linear:
  - Correlação η (eta)
  - Kendall
  - Spearman

Matriz de Correlação

- Coeficiente de Pearson

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

- Coeficiente de Spearman

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{(n^3 - n)}$$

onde d_i é a diferença entre cada posição de x e y.

- Coeficiente de Kendall

$$\tau = \frac{2(n_c - n_d)}{n(n-1)}$$

Matriz de Correlação

- A matriz de correlação é simétrica.
- O valor 1 significa que as duas variáveis são exatamente correlacionadas. É o caso de uma relação linear entre as duas variáveis;
- O valor negativo, significa que elas são inversamente correlacionadas.
- Não necessariamente informa o tipo de relação que existe entre as variáveis.

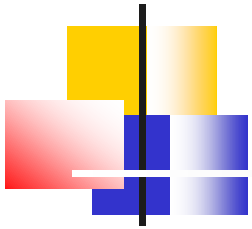
Matriz de Correlação

Tabela 2. Matriz de correlação das variáveis de qualidade de água no médio Rio Pomba

Variável	OD	DBO	Temp	CE	Alcalin	DQO	PT	Norg	NH_4	NTK	NT	ST	SST	SDT	SIS
OD	1,00														
DBO	**-0,58**	1,00													
Temp	**-0,57**	0,18	1,00												
CE	0,02	-0,24	**0,61**	1,00											
Alcalin	-0,23	-0,34	**0,63**	**0,77**	1,00										
DQO	**-0,70**	0,32	**0,64**	0,12	0,38	1,00									
PT	-0,09	0,31	0,03	0,21	-0,11	-0,10	1,00								
Norg	0,16	0,23	**-0,62**	-0,49	**-0,71**	-0,49	0,28	1,00							
NH_4	**-0,76**	**0,53**	0,38	-0,01	0,12	0,31	0,48	0,08	1,00						
NTK	0,08	0,28	**-0,57**	-0,49	**-0,69**	-0,45	0,33	**0,99**	0,18	1,00					
NT	0,15	0,27	**-0,55**	-0,34	**-0,62**	**-0,50**	0,35	**0,97**	0,13	**0,97**	1,00				
ST	0,24	-0,06	-0,23	0,33	-0,02	**-0,61**	0,47	0,29	0,08	0,29	0,37	1,00			
SST	0,01	0,35	-0,39	-0,14	-0,26	-0,43	**0,60**	**0,61**	0,37	**0,64**	**0,63**	**0,63**	1,00		
SDT	0,30	-0,38	0,04	**0,55**	0,20	-0,42	0,09	-0,15	-0,22	-0,17	-0,06	**0,75**	-0,04	1,00	
SIS	0,01	0,26	-0,47	-0,19	-0,37	-0,43	**0,54**	**0,61**	0,32	**0,63**	**0,63**	**0,75**	**0,84**	0,24	1,00

OD (mg L^{-1}) – oxigênio dissolvido; DBO (mg L^{-1}) – demanda bioquímica de oxigênio; Temp (°C) – temperatura; CE (µS cm^{-1}) – condutividade elétrica; Alcalin (mg L^{-1} $CaCO_3$) – alcalinidade; DQO (mg L^{-1}) – demanda química de oxigênio; PT (mg L^{-1} – PO_4) – fósforo total; Norg (mg L^{-1} - N) – nitrogênio orgânico; NH_4 (mg L^{-1} - N) – nitrogênio amoniacal; NTK (mg L^{-1} - N) – nitrogênio Kjeldahl; NT (mg L^{-1} – N) – nitrogênio total; ST (mg L^{-1}) – sólidos totais; SST ((mg L^{-1}) – sólidos suspensos totais; SDT (mg L^{-1}) – sólidos dissolvidos totais; SIS (mg L^{-1}) – sólidos inorgânicos suspensos

Fonte: Scielo

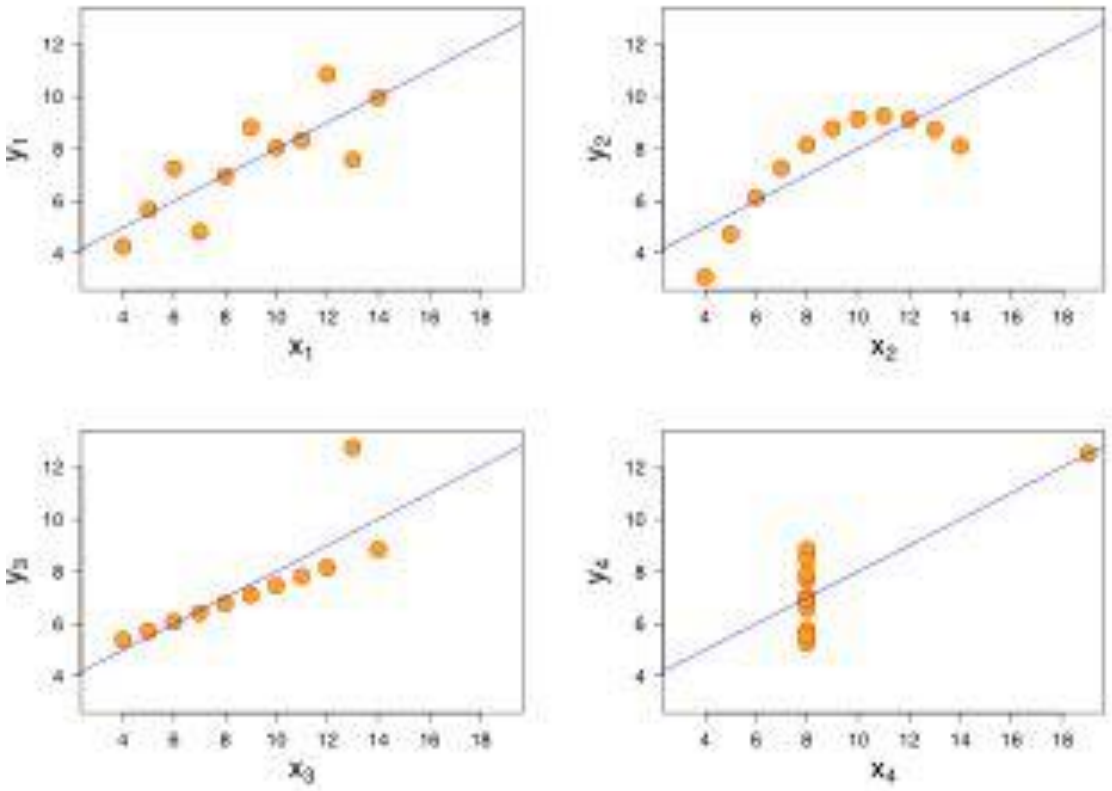


Exercício

- Calcular a matriz de covariância para o conjunto de dados:

Peso	Altura	Temperatura
73	170	37
67	165	38
90	190	34
49	152	31

Dados Multivariados

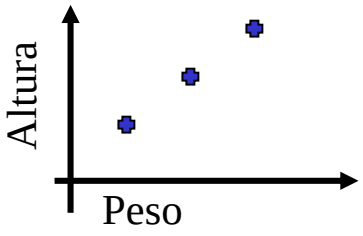


Todos correlação = 0,816

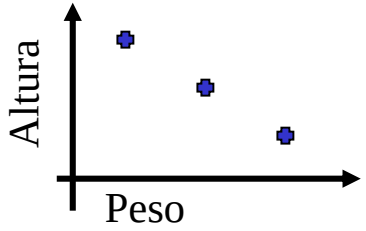
Fonte:Wikipedia

Exemplo

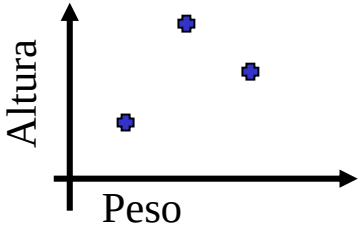
Peso	Altura
60	170
70	180
80	190



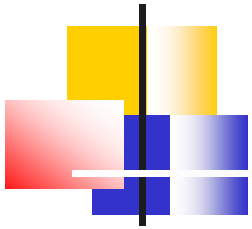
Peso	Altura
60	190
70	180
80	170



Peso	Altura
60	170
70	190
80	180



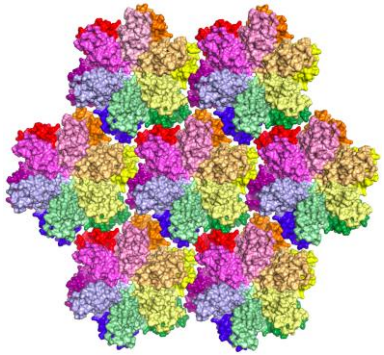
Perguntas

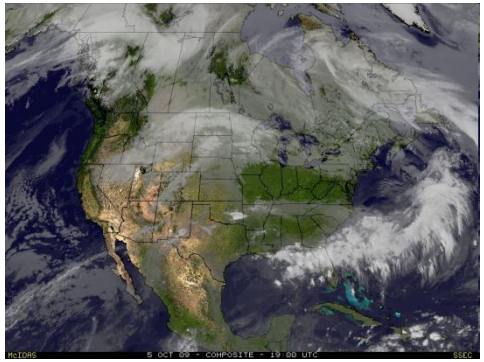


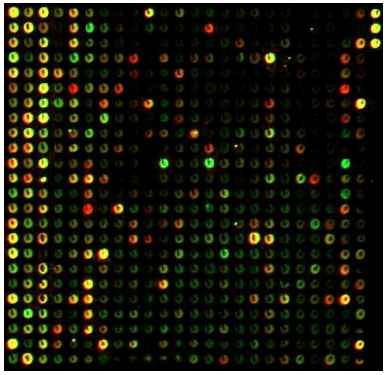
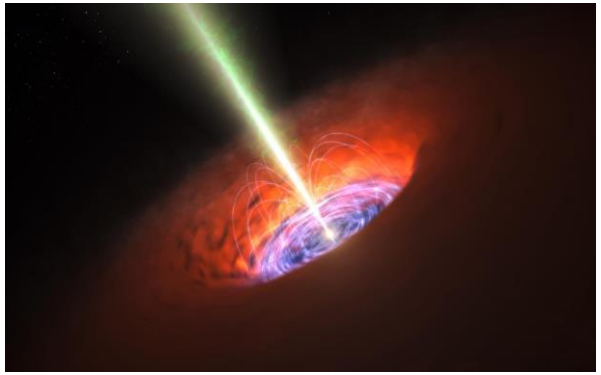
Outras formas de sumarizar dados

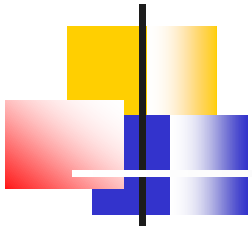
- Visualização gráfica
  - Em vários casos, facilita compreensão de padrões mais complexos nos dados
  - Exemplos simples
    - Histograma
    - Diagrama de torta
    - *Scatter plot*
    - Faces de Chernoff

Exemplos





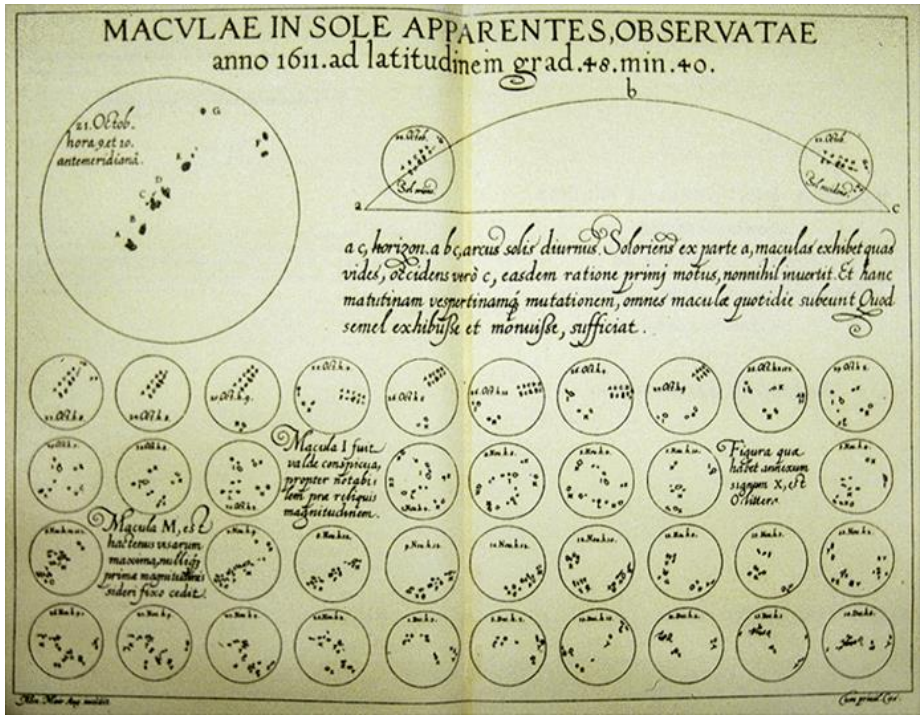


Visualização

- Visualização tem um papel importante em análise de dados
  - Uma das técnicas mais poderosas para exploração dos dados
    - Facilita visualização de dados e resultados
    - *Visual data mining*
      - Usa técnicas de visualização em mineração de dados
      - Importante área de CD

Um dos primeiros

- Mapa solar de Galileu



Outro dos primeiros usos

Napoleão

Exército francês, comandado por Napoleão
Invade a Rússia, em 1812

Histogramas

- Conjunto de dados Iris
  - Largura das pétalas usando 10 e 5 cestas

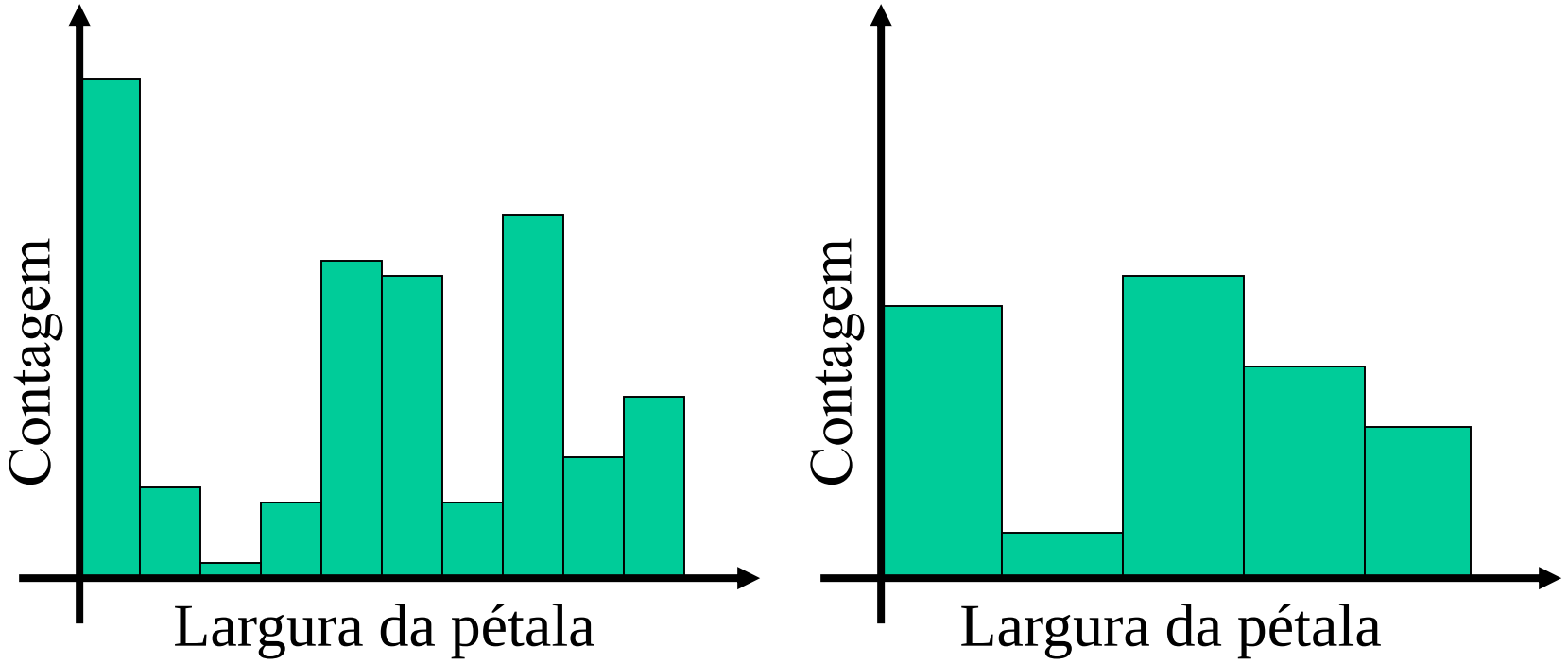


Diagrama de torta

- Frequências relativas podem ser vistas no diagrama circular

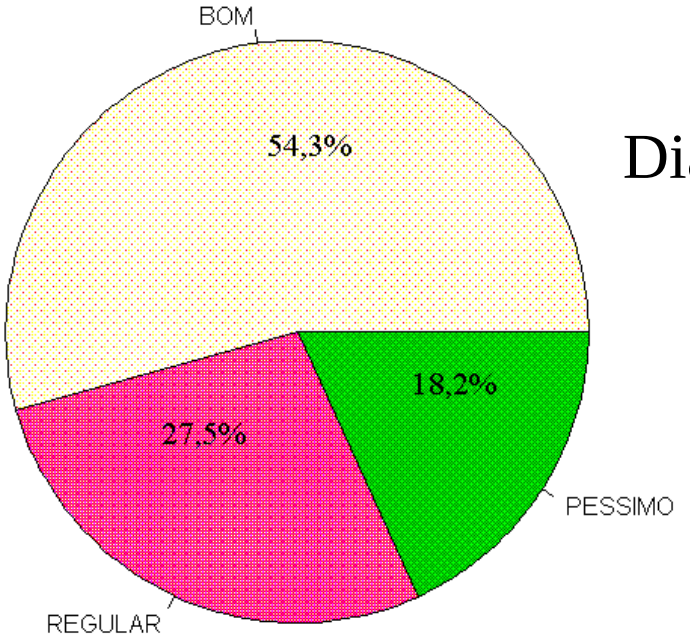
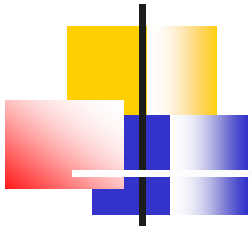


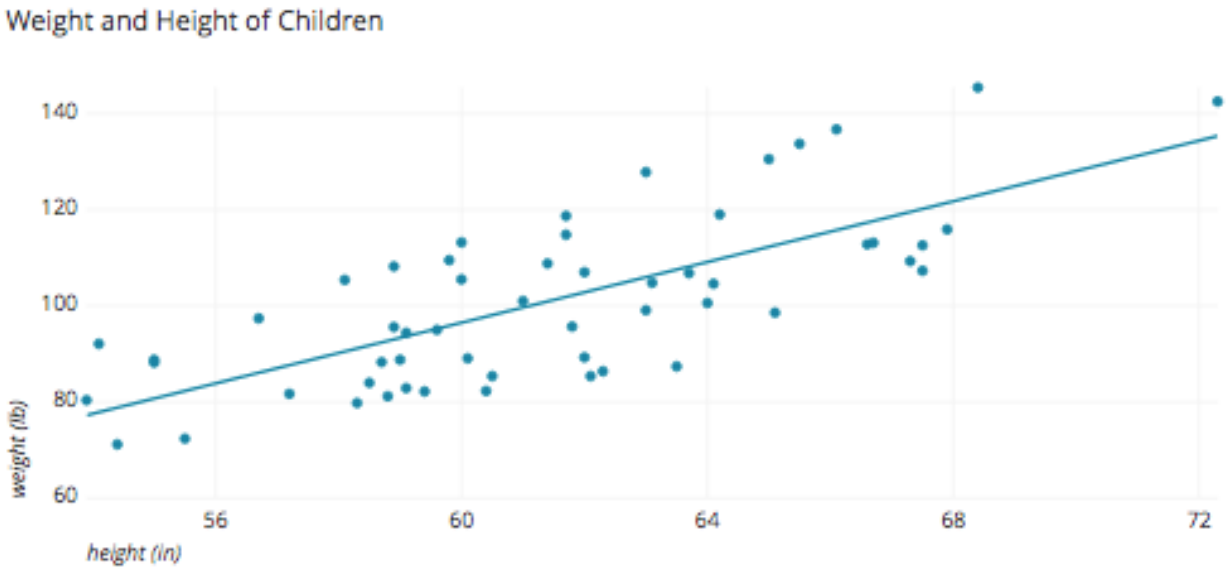
Diagrama de torta (pizza)



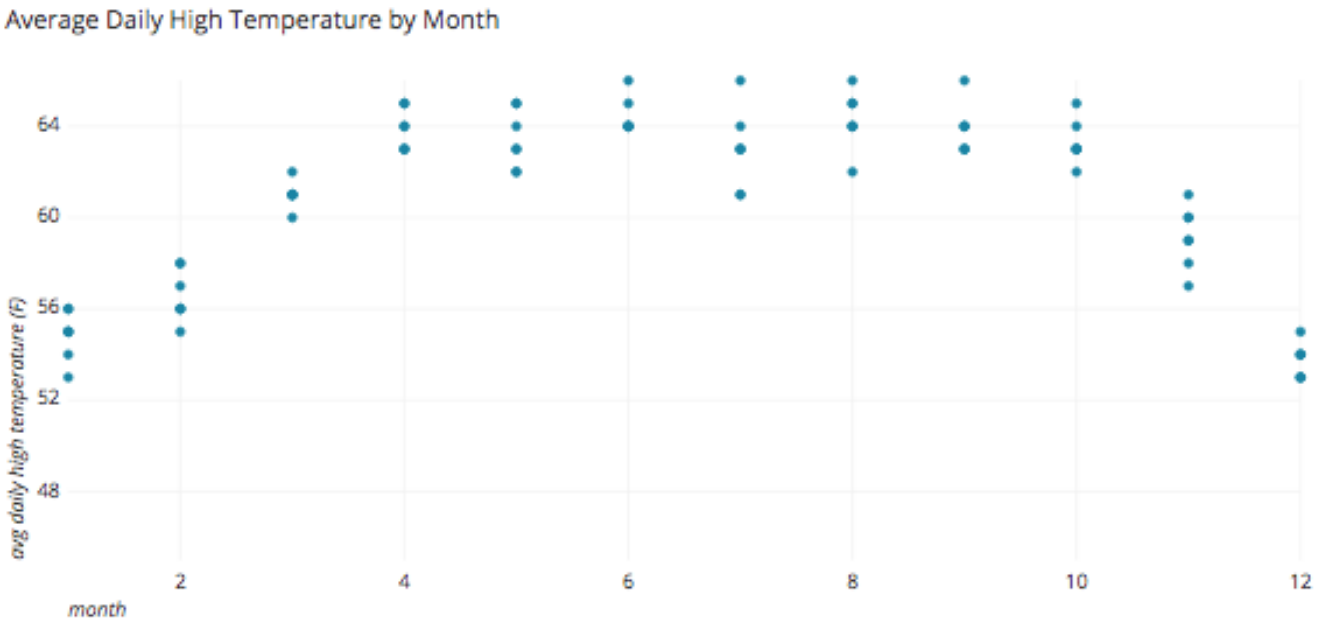
Scatter Plot

- Usado para ilustrar graficamente Correlação Linear entre dois atributos
- Cada objeto é associado a uma posição em um gráfico
  - Valores dos atributos definem sua posição
  - Valores podem ser inteiros ou reais
- Matrizes de scatter plot resumem relação para vários pares de atributos

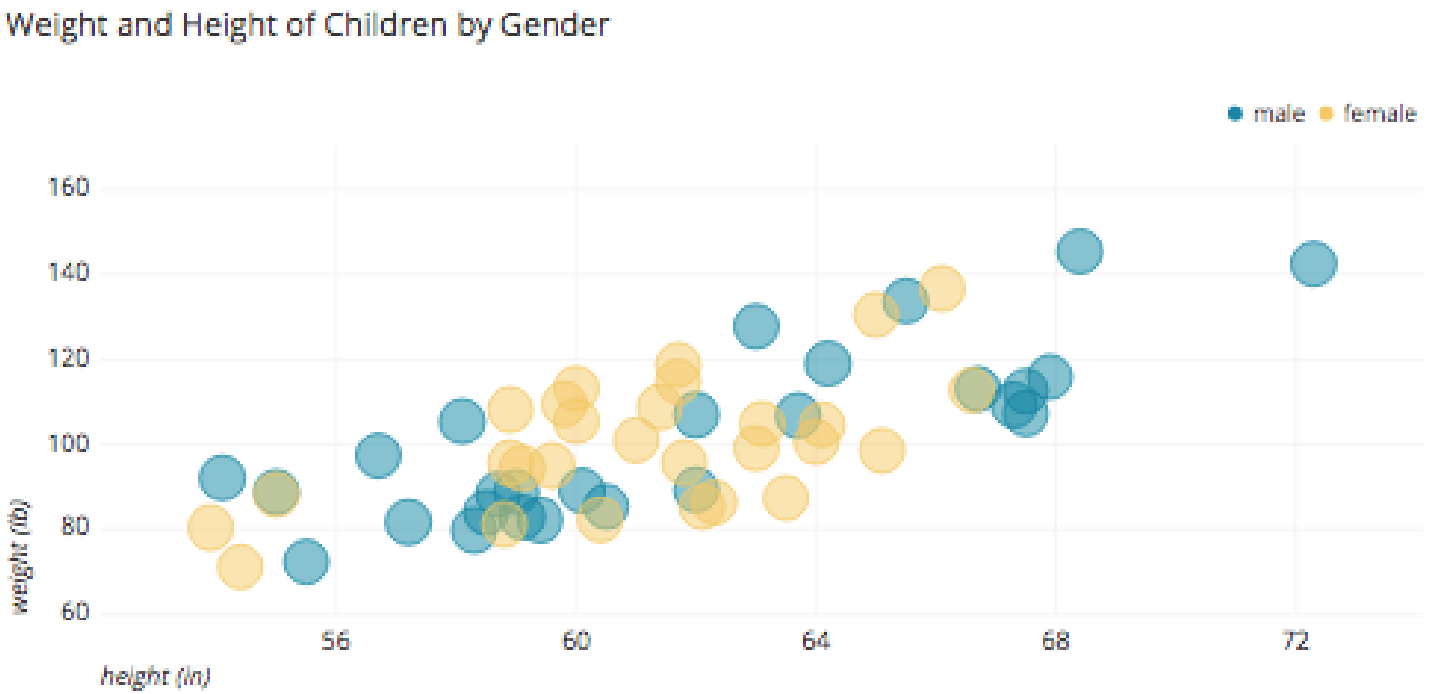
Relação Linear



Relação Não-Linear

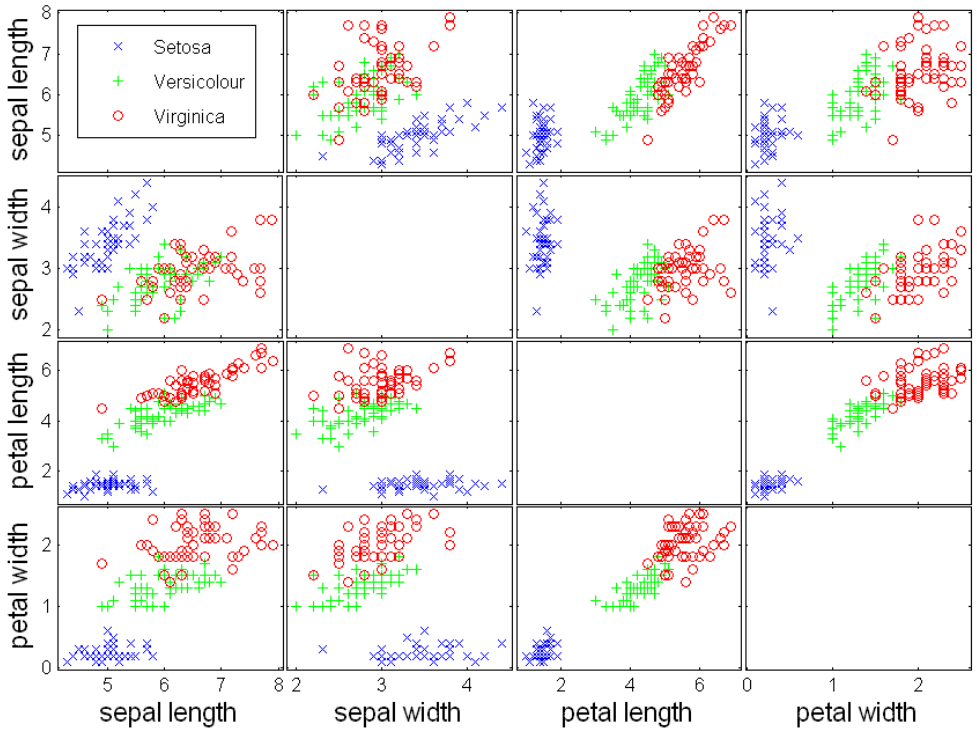


Pode adicionar mais info

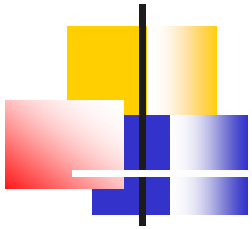


Scatter Plot

- Matriz para atributos do conjunto iris



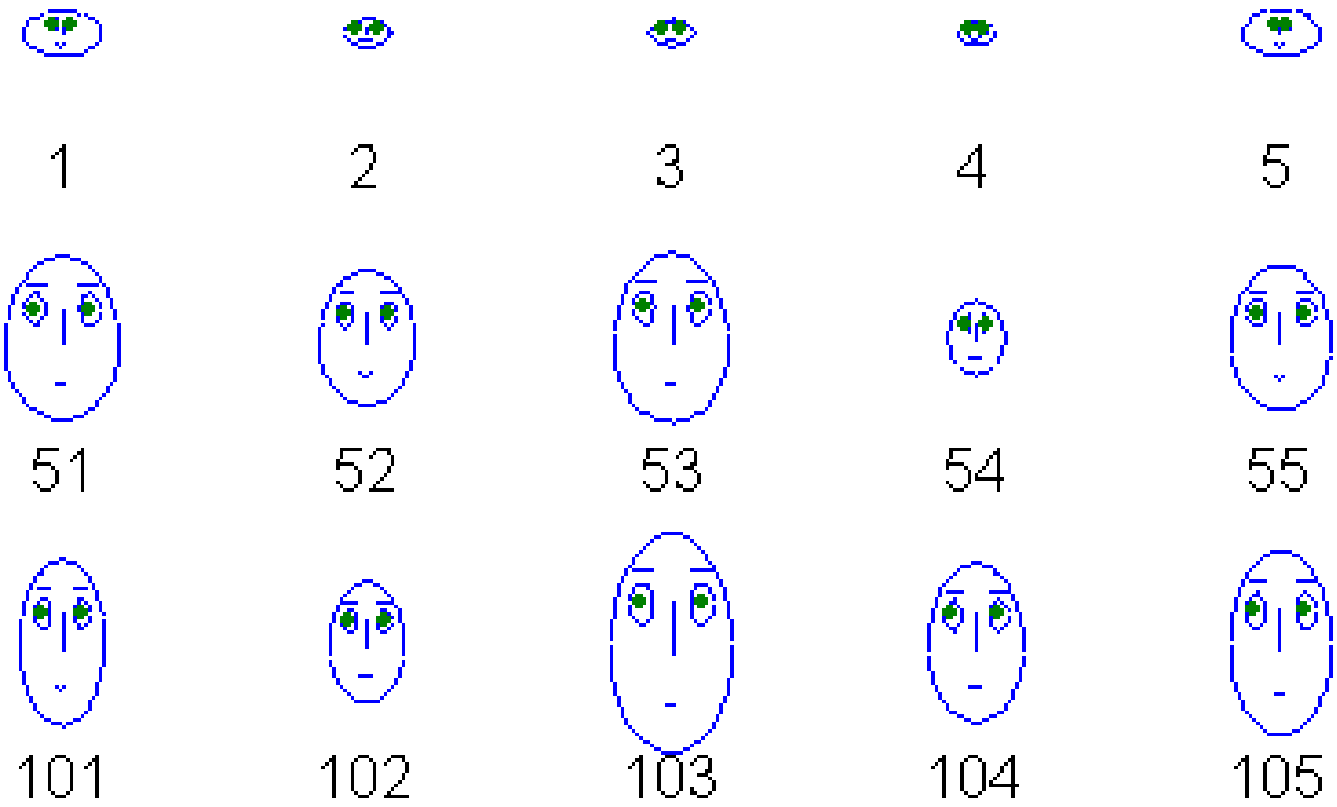
Diferentes classes são indicadas por cores diferentes



Faces de Chernoff

- Criado por Herman Chernoff
- Mapeia os valores dos atributos para imagens familiares para seres humanos: faces
  - Cada objeto é representado por uma face
  - Cada atributo é associado a uma característica específica da face
- Baseia-se na habilidade humana de distinguir faces

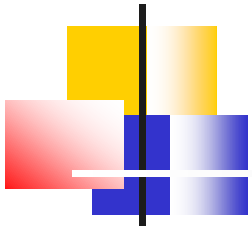
Faces de Chernoff



Exercício

- Representar os dados a seguir usando faces de Chernoff

Febre	Idade	Batimento	Dor	Diagnóstico
sim	23	elevado	sim	doente
não	9	baixo	não	saudável
sim	61	elevado	não	saudável
sim	32	baixo	sim	doente
sim	21	elevado	sim	saudável
não	48	elevado	sim	doente



Considerações Finais

- Caracterização de dados
  - Objetos e atributos
  - Tipos de dados
- Exploração de dados
  - Dados univariados
  - Medidas de localidade, espalhamento e distribuição
  - Dados multivariados
  - Visualização de dados e de resultados