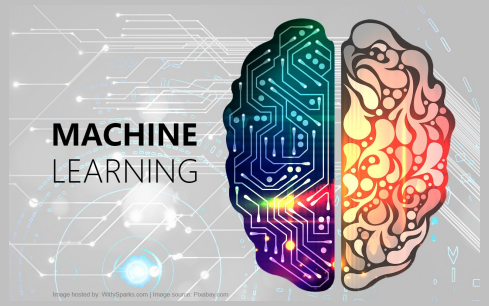


ANÁLISE DE DADOS

PROF. DR. **TIAGO A. ALMEIDA**
talmeida@ufscar.br
talmeida-ufscar

Dados

- Estima-se que a quantidade de dados em bases de dados mundiais **dobra** a cada 20 meses
- Crescimento tem ocorrido em várias áreas
  - Transações bancárias
  - Utilização de cartões de crédito
  - Dados governamentais
  - Medições ambientais
  - Dados clínicos
  - Mapeamento genético
  - Informações na web
  - etc

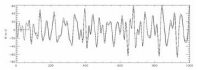
Dados

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

Bases de dados cada vez **maiores**

Dados

- Podem ter diferentes **formatos**


Séries temporais


Grafos


Vídeos


Textos

Páginas web


Áudios


Imagens

Geralmente 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

		Sintomas				
		temperatura	dor		pressão	doente
Objetos	$paciente_1$	38ºC	sim	...	12.7	Sim
	$paciente_2$	36ºC	não	...	12.7	Não
				⋮		
	$paciente_m$	40ºC	não	...	14	Sim

Conjunto de dados

- Pode ser representado por uma matriz de objetos $\mathbf{X}_{m \times n}$
  - m = número de amostras
  - n = número de atributos (excluindo atributo-meta)
    - Dimensionalidade do espaço de objetos (de entradas/de atributos)
  - **Formalização**: amostra $x^{(i)}$ e atributo x_j
    - Elemento $x_j^{(i)}$ (ou x_{ij}) ⇨ valor do j-ésimo atributo para o objeto i

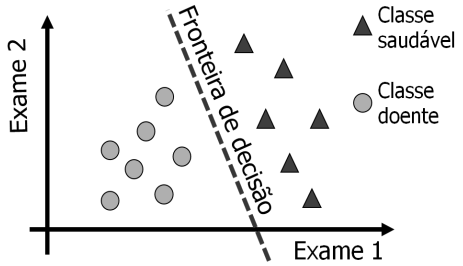
Formato atributo-valor

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

		Sintomas				
		temperatura	dor		pressão	doente
Objetos	$paciente_1$	38ºC	sim	...	12.7	Sim
	$paciente_2$	36ºC	não	...	12.7	Não
				⋮		
	$paciente_m$	40ºC	não	...	14	Sim

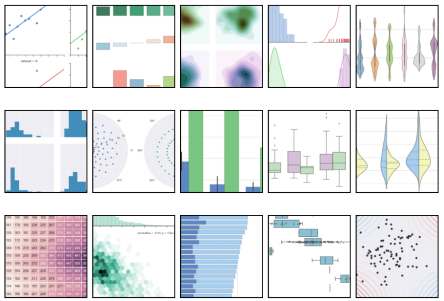
Conjunto de dados: visualização gráfica

- Representação de conjunto de dados com dois atributos



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



Exploração de dados

Frequência

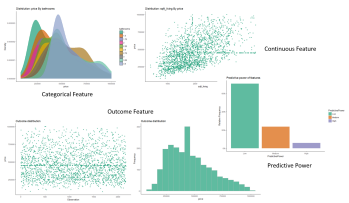
- Proporção de vezes que um atributo assume um dado valor
- Aplicável a valores numéricos e simbólicos
- *Ex.: 40% dos pacientes têm febre*

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

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

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 *m* objetos **x** = $\{x^1, x^2, ..., x^m\}$

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

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

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

Média

- Equação:

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

Problema: sensível a *outliers*

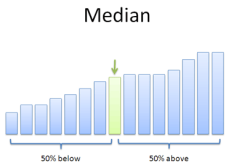
Bom indicador apenas se valores são distribuídos simetricamente

Mediana

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

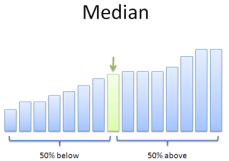
$$mediana(\mathbf{x}) = \begin{cases} \frac{1}{2}(x^r + x^{r+1}) & \text{se } m \text{ for par } (m = 2r) \\ x^{r+1} & \text{se } m \text{ for ímpar } (m = 2r + 1) \end{cases}$$

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



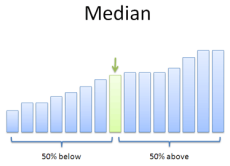
Mediana

- 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



Mediana

- Exemplos:
  - {17, 4, 8, 21, 4}
  - {17, 4, 8, 21, 4, 15, 13, 9}



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

Média truncada

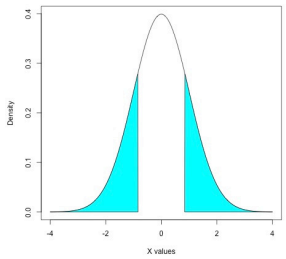
- Ex. conjunto de dados `hospital`

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 truncada (p = 25%): 23,7

Média truncada

- Descarta elementos extremos da sequência ordenada de valores
  - Minimizar problemas da média
  - Necessário definir porcentagem
- Passos:
  - Definir porcentagem p
  - Ordenar valores
  - Descartar (p/2)% de valores de cada extremo
  - Calcular a média dos exemplos restantes



Média truncada

- 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
Média truncada (p = 25%): 3,2

Exercício

- Dado o conjunto de dados {1, 2, 3, 4, 5, 80}, calcular:
  - Média
  - Mediana
  - Média truncada com p = 33%

Quartis e percentis

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

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

Exercício

- Dado o conjunto de dados {1, 2, 3, 4, 5, 80}, calcular:
  - Média: (1+2+3+4+5+80)/6 = 15,8
  - Mediana: 3+4 / 2 = 3,5
  - Média truncada com p = 33%: (2+3+4+5)/4 = 3,5

Percentil

Algoritmo para cálculo do percentil

Entrada: m valores e percentil p

Saída: valor do percentil

- Ordenar os m valores de maneira crescente
- Calcular k = m * 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

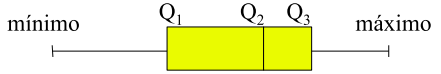
Média: 26,1
Mediana: 21,5
Média truncada (p= 25%): 23,7
Q1: 18,5; *Q2*: 21,5; *Q3*: 31
P40: 21

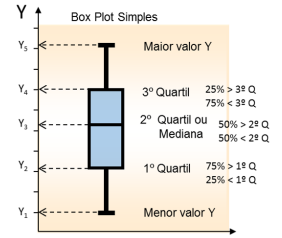
Boxplots

☑ Também chamados diagramas de Box e Whisker

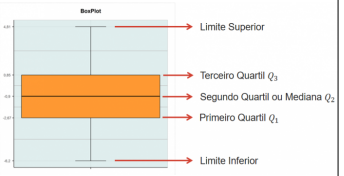
☑ Forma gráfica de visualizar quartis

☑ 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*



Quartil e percentil

☑ Ex. conjunto de dados `hospital`

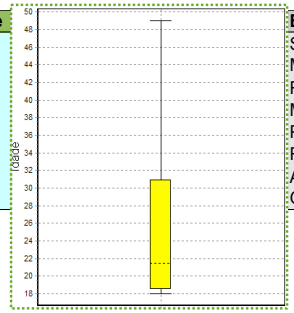
Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	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
Média truncada (p= 25%): 3,2
Q1: 2; *Q2*: 2,5; *Q3*: 5
P40: 2

Boxplot

☑ Ex. conjunto de dados `hospital`

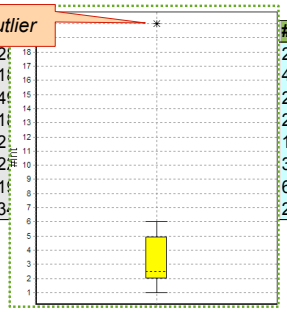
Id.	Nome	Idade
4201	João	28
3217	Maria	18
4039	Luiz	49
1920	José	18
4340	Cláudia	21
2301	Ana	22
1322	Marta	19
3027	Paulo	34



Est.	Diagnóstico
SP	Doente
MG	Doente
RS	Saudável
MG	Doente
PE	Saudável
RJ	Doente
AM	Doente
GO	Saudável

Boxplot modificado

- Ex. conjunto de dados `hospital`



Id.	Nor	# Int.	Est.	Diagnóstico
4201	João	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

Intervalo

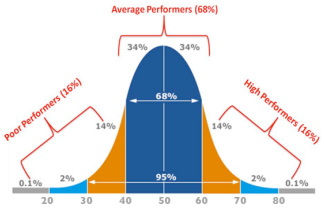
- Mostra espalhamento máximo entre valores
  - Medida mais simples

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

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

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

- 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

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

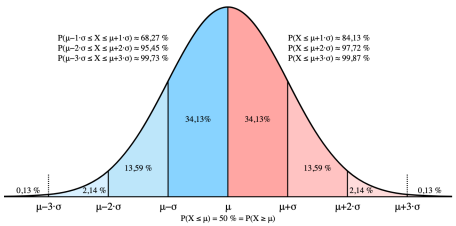
Intervalo: 31
Desvio padrão: 10,8

Variância e desvio padrão

☑ Mais utilizadas

$$\text{variância}(\mathbf{x}) = \frac{1}{m\text{-}1} \sum_{i=1}^{m} (x^i - \overline{x})^2$$

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



Problema: também são distorcidas pela presença de *outliers*

Desvio padrão

☑ Ex. conjunto de dados `hospital`

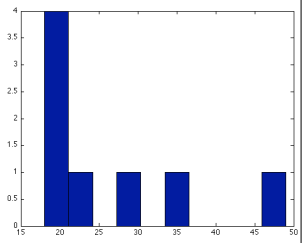
Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F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

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

Id.	Nome	Idade
4201	João	28
3217	Maria	18
4039	Luiz	49
1920	José	18
4340	Cláudia	21
2301	Ana	22
1322	Marta	19
3027	Paulo	34



Dados multivariados

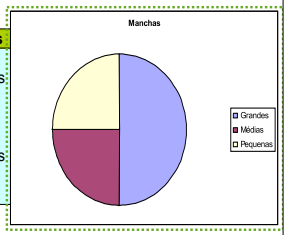
- Possuem mais de um atributo de entrada
  - Ex. conjuntos de dados `hospital`
  - Medidas de localidade e espalhamento podem ser calculadas para cada atributo separadamente
    - Ex. média

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

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

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

- 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

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

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

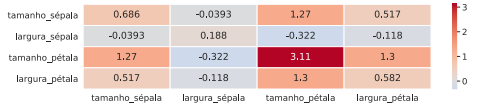
Covariância vs Correlação

Iris Setosa

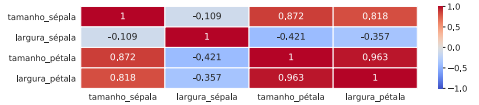
Iris Versicolor

Iris Virginica

Covariância



Correlação



Correlação

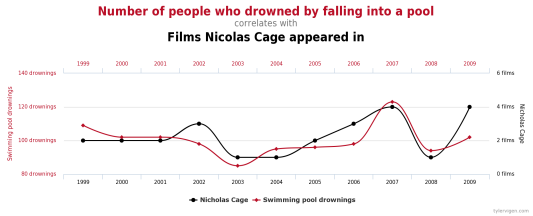
☑ Indicação mais clara da força da relação linear entre dois atributos

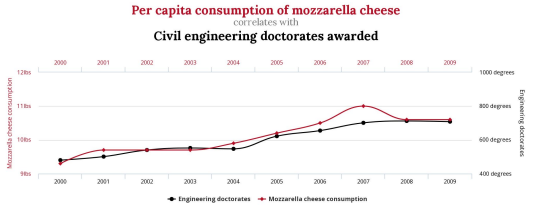
☑ Matriz de correlação: correlação entre todos pares de atributos

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

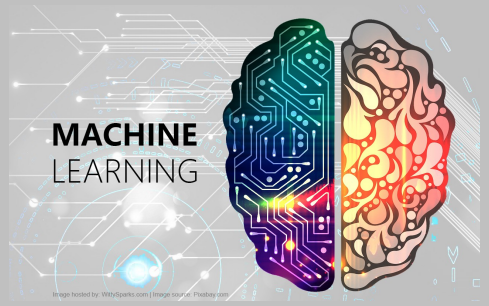
Observação: valores variam de -1 (correlação negativa máxima) a +1 (correlação positiva máxima) e correlação($\mathbf{x}^i$, $\mathbf{x}^i$) = 1

Correlação vs Causalidade





ANÁLISE DE DADOS



PROF. DR. **TIAGO A. ALMEIDA**

talmeida@ufscar.br

talmeida-ufscar