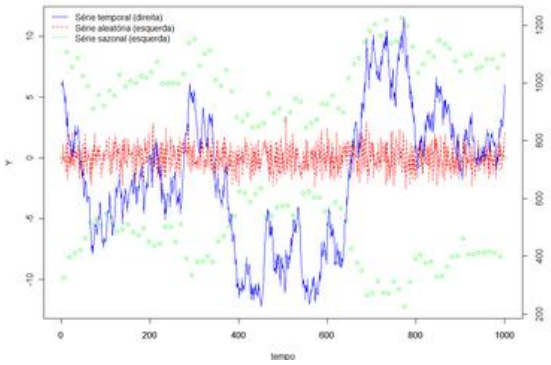
Dados : características

Profa Silvia Moraes

A cada dia uma quantidade enorme de dados é gerada.

- Segundo o McKinsey Global Institute (fev, 2020) é esperado que em 2030, o mundo consuma até 20 vezes mais dados do que hoje.

- Conforme a revista *Forbes* (abr,2021) estima-se que uma única fábrica pode gerar mais de 2 terabytes de dados em um mês.



- Os dados podem assumir **formatos diferentes**.
- Podem ser:
  - Series temporais;
  - Conjuntos de itens;
  - Transações;
  - Grafos ou redes sociais;
  - Textos;
  - Páginas web;
  - Imagens;
  - Vídeos;
  - e áudios.

- Todos os dados existentes já foram analisados e compreendidos ?

- Todos os dados existentes já foram analisados e compreendidos ?

  Não

- Todos os dados existentes já foram analisados e compreendidos ?

  Não

- E o que é necessário para analisar dados ?

  Preparação desses dados

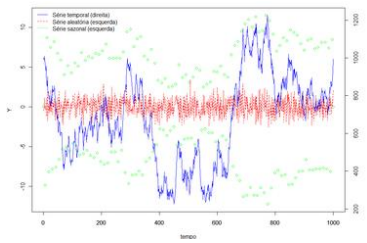
- E o que é necessário para analisar dados ?

- E o que é necessário para analisar dados ?

  Preparação de dados

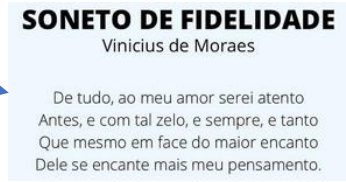
Antes vamos precisar entender que os dados podem ser:

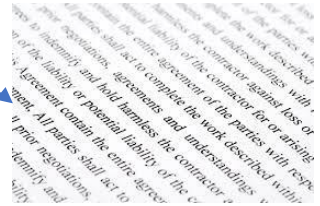
Antes vamos precisar entender que os dados podem ser:



Antes vamos precisar entender que os dados podem ser:

Antes vamos precisar entender que os dados podem ser:





Antes vamos precisar entender que os dados podem ser:

Agora que vimos as categorias dos dados, vamos analisá-los.

Nosso foco serão os dados estruturados.

Vamos estudar:

- Caracterização dos Dados
- Exploração de Dados

Caracterização dos dados

Caracterização dos Dados

- Dados estruturados são representados por matrizes de objetos da forma *n* x *d*, chamadas **tabelas atributo-valor**, onde:
  - *n* é o número de objetos
  - *d* é o número de atributos de cada objeto

8 objetos

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	8	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	?	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável

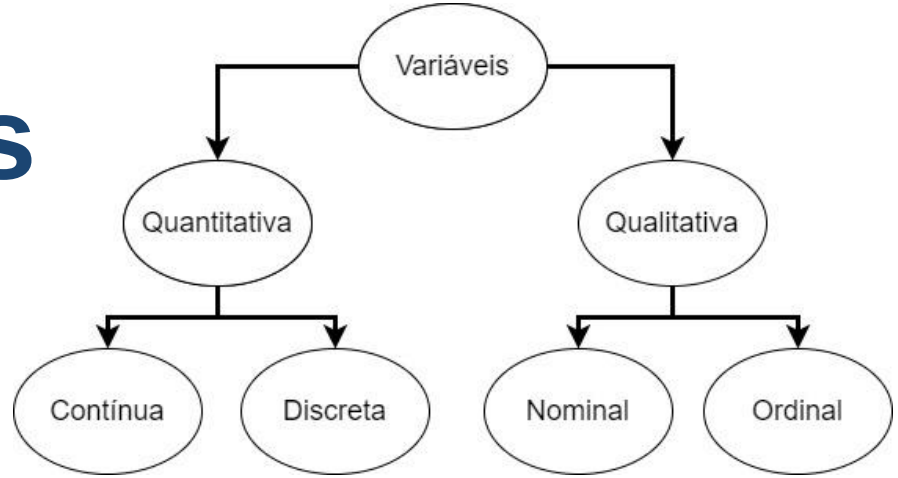
Caracterização dos Dados

- Dados estruturados são representados por matrizes de objetos da forma *n* x *d*, chamadas **tabelas atributo-valor**, onde:
  - *n* é o número de objetos
  - *d* é o número de atributos de cada objeto

10 atributos

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	8	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	?	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável

Caracterização dos Dados



- Quanto ao **tipo**, os dados podem ser:
  - **Quantitativos:** numéricos;
  - **Qualitativos:** simbólico ou categórico.

Atributo	Tipo
Id.	Qualitativo
Nome	Qualitativo
Idade	Quantitativo discreto
Sexo	Qualitativo
Peso	Quantitatico continuo
Manchas	Qualitativo

Atributo	Tipo
Temp.	Quantitatico continuo
#Int.	Quantitativo discreto
Est.	Qualitativo
Diagnostico	Qualitativo

Caracterização dos Dados

- Quanto ao **tipo**, os dados podem ser:
  - **Quantitativos:** numéricos;
  - **Qualitativos:** simbólico ou categórico.

	Atributo	Tipo
X	Id.	Qualitativo
X	Nome	Qualitativo
	Idade	Quantitativo discreto
	Sexo	Qualitativo
	Peso	Quantitativo continuo
	Manchas	Qualitativo

Atributos de entrada

Atributo	Tipo
Temp.	Quantitativo continuo
#Int.	Quantitativo discreto
Est.	Qualitativo
Diagnostico	Qualitativo

Atributos de entrada

Caracterização dos Dados

- Quanto ao **tipo**, os dados podem ser:
  - **Quantitativos:** numéricos;
  - **Qualitativos:** simbólico ou categórico.

Atributo	Tipo
Id.	Qualitativo
Nome	Qualitativo
Idade	Quantitativo discreto
Sexo	Qualitativo
Peso	Quantitatico continuo
Manchas	Qualitativo

Atributo	Tipo
Temp.	Quantitatico continuo
#Int.	Quantitativo discreto
Est.	Qualitativo
Diagnostico	Qualitativo

Atributo de saida,
Atributo alvo
Classe

Caracterização dos Dados

- Quanto `**a escala**, os dados podem ser:
  - **Intervalares e racionais:** para os quantitativos;
  - **Nominais e ordinais:** para os qualitativos.

Atributo	Escala
Id.	Nominal
Nome	Nominal
Idade	Racional
Sexo	Nominal
Peso	Intervalar
Manchas	Nominal

Atributo	Tipo
Temp.	Intervalar
#Int.	Racional
Est.	Nominal
Diagnostico	Nominal

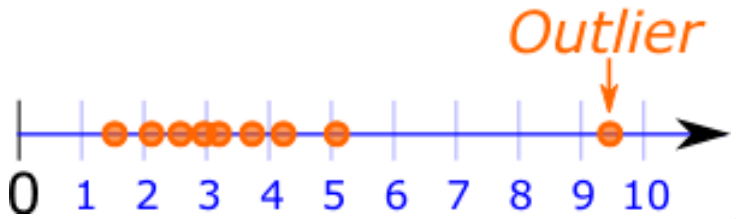
Intervalar: os números variam dentro de um intervalo**.**
Racional: números com mais informação, representam quantidades, distancias, tempo, …

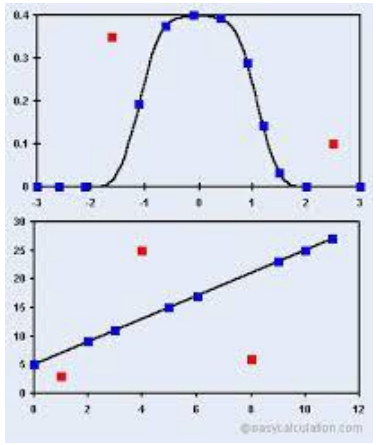
Nominal: apenas nomes diferentes**.**
Ordinal: Há uma relação de ordem entre os nomes, tal como: pequeno, médio e grande

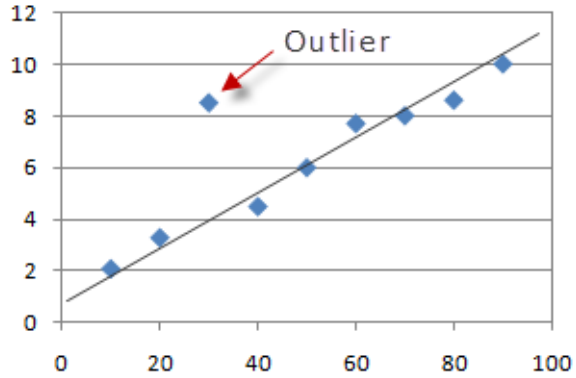
Exploração dos Dados

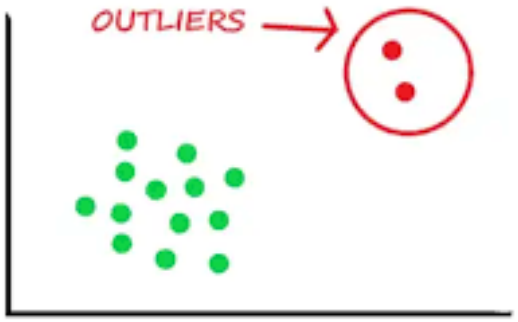
Exploração dos Dados:

Conhecer os dados ajuda na seleção das técnicas mais apropriadas de pré-processamento e de aprendizado de máquina.









Outliers são dados que se diferenciam drasticamente de todos os outros, são pontos fora da curva.

Influenciam a análise dos dados.

Fique atento!

A **estatistica descritiva é muito útil pois permite resumir dados quantitativos**, tais como:

- Idade média dos pacientes;
- Percentual de pacientes do genero masculino.

O processamento estatistico permite **capturar informações** como:

- Frequência
- Localização ou tendência central (por exemplo, média)
- Dispersão ou espalhamento (por exemplo, desvio padrão)
- Distribuição ou formato

Dados Univalorados x Multivalorados

- **Univalorados**: Possui apenas 1 atributo.
- **Multivalorados**: Possuem mais de 1 atributo.

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	8	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	?	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável

Altura
1.60
1.80
1.89
1.67
1.95
1.70
1.62
1.89
1.75

Altura
Baixo
Alto
Alto
Baixo
Alto
Medio
Baixo
Alto
Medio

Dados Univalorados x Multivalorados

- **Univalorados**: Possui apenas 1 atributo.
- **Multivalorados**: Possuem mais de 1 atributo.

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	8	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	?	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável

Altura
1.60
1.80
1.89
1.67
1.95
1.70
1.62
1.89
1.75

Altura
Baixo
Alto
Alto
Baixo
Alto
Medio
Baixo
Alto
Medio

Dados Univalorados x Multivalorados

- **Univalorados**: Possui apenas 1 atributo.
  - A análise de dados mais simples;
  - Não lida com causas ou relações entre os dados;
    - Exemplo: altura.

Altura
1.60
1.80
1.89
1.67
1.95
1.70
1.62
1.89
1.75

Altura
Baixo
Alto
Alto
Baixo
Alto
Medio
Baixo
Alto
Medio

Dados Univalorados

Medidas de Localidade:
Pontos de referência dos dados

- **Univalorados**: Possui apenas 1 atributo.
  - A análise de dados mais simples;
  - Não lida com causas ou relações entre os dados;
    - Exemplo: altura.
  - Dados nominais
    - **Moda:** valor que mais frequente. **Ex: Alto**
  - Dados numéricos:
    - Média: fácil mas adequado apenas se há distribuição simétrica dos dados, pois é sensível a outliers;
    - Mediana: valor central dos dados;
    - Percentil e quartis: permitem ver como os dados estão distribuídos.

Altura
1.60
1.80
1.89
1.67
1.95
1.70
1.62
1.89
1.75

Altura
Baixo
Alto
Alto
Baixo
Alto
Medio
Baixo
Alto
Medio

Dados Univalorados

Medidas de Localidade:
Pontos de referência dos dados

- **Univalorados**: Possui apenas 1 atributo.
  - A análise de dados mais simples;
  - Não lida com causas ou relações entre os dados;
    - Exemplo: altura.
  - Dados nominais
    - **Moda:** valor que mais frequente. **Ex: Alto**
  - Dados numéricos:
    - Média: fácil mas adequado apenas se há distribuição simétrica dos dados, pois é sensível a outliers;
    - Mediana: valor central dos dados;
    - Percentil e quartis: permitem ver como os dados estão distribuídos.

Altura
1.60
1.80
1.89
1.67
1.95
1.70
1.62
1.89
1.75

Altura
Baixo
Alto
Alto
Baixo
Alto
Medio
Baixo
Alto
Medio

Dados Univalorados

Medidas de Localidade:
Pontos de referência dos dados

- **Univalorados**: Possui apenas 1 atributo.
  - A análise de dados mais simples;
  - Não lida com causas ou relações entre os dados;
    - Exemplo: altura.
  - Dados nominais
    - Moda**:** valor que mais frequente.
  - Dados numéricos:
    - **Média**: fácil mas adequado apenas se há distribuição simétrica dos dados, pois é sensível a outliers; **Ex: 1.76**
    - Mediana: valor central dos dados;
    - Percentil e quartis: permitem ver como os dados estão distribuídos.

Altura
1.60
1.80
1.89
1.67
1.95
1.70
1.62
1.89
1.75

Altura
Baixo
Alto
Alto
Baixo
Alto
Medio
Baixo
Alto
Medio

Dados Univalorados

Medidas de Localidade:
Pontos de referência dos dados

- **Univalorados**: Possui apenas 1 atributo.
  - ...
  - **Mediana**: valor central dos dados;
    - Exige que os valores sejam ordenados
    - Quantidade ímpar de elementos: valor central
    - Quantidade par de elementos: media dos 2 valores centrais

1,6 1,62 1,67 1,7

(1,62+1,67)/2 = **1,645**

Altura
1.60
1.80
1.89
1.67
1.95
1.70
1.62
1.89
1.75

Altura ordenada

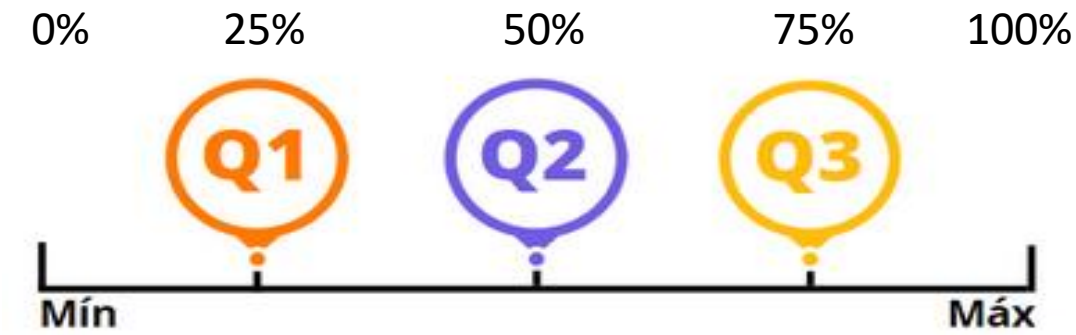
1,60
1,62
1,67
1,70
1,75
1,80
1,89
1,89
1,95

.

Dados Univalorados

Medidas de Localidade:
Pontos de referência dos dados

- **Univalorados**: Possui apenas 1 atributo.
  - ....
    - Percentil e quartis: permitem ver como os dados estão distribuídos.
      - O quartil é uma representação/delimitação para o percentil.



[1,60; **1,62; 1,67**; 1,70; **1,75**; 1,80; **1,89; 1,89;** 1,95]

1,645 **1,75**; **1,89**

Q1(25) = 1,645
Q2(50) = 1,75
Q3(75) = 1,89

Altura
1.60
1.80
1.89
1.67
1.95
1.70
1.62
1.89
1.75

Dados Univalorados

Medidas de Espalhamento:
Medem a dispersão dos dados

- **Univalorados**: Possui apenas 1 atributo.
  - Medidas de espalhamento: permite verificar se os valores estão espalhados ou concentrados em torno de um valor.
  - Medidas mais comuns:
    - Intervalo
    - Variância
    - Desvio Padrão

Altura
1.60
1.80
1.89
1.67
1.95
1.70
1.62
1.89
1.75

Dados Univalorados

Medidas de Espalhamento:
Medem a dispersão dos dados

- **Univalorados**: Possui apenas 1 atributo.
  - Medidas de espalhamento: permite verificar se os valores estão espalhados ou concentrados em torno de um valor.
  - Medidas mais comuns:
    - **Intervalo:** simples. Exemplo: **[1,60; 1,95]**
    - Variância
    - Desvio Padrão

Min	Max

Se houver concentração em um ponto , não será uma boa medida.

Altura
1.60
1.80
1.89
1.67
1.95
1.70
1.62
1.89
1.75

Dados Univalorados

Medidas de Espalhamento:
Medem a dispersão dos dados

- **Univalorados**: Possui apenas 1 atributo.
  - Medidas de espalhamento: permite verificar se os valores estão espalhados ou concentrados em torno de um valor.
  - Medidas mais comuns:
    - Intervalo
    - **Variância (σ):** é útil para determinar o afastamento da média. Para isso, determina-se o valor médio das diferenças quadradas em relacao a média.

$$\sigma = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1}$$

onde:
- x_i é o dado;
- $\overline{x}$ é a média;
- n é o total de dados.

σ (altura) =0,01605

Como usa a média e sensivel a outliers.

Altura
1.60
1.80
1.89
1.67
1.95
1.70
1.62
1.89
1.75

Dados Univalorados

Medidas de Espalhamento:
Medem a dispersão dos dados

- **Univalorados**: Possui apenas 1 atributo.
  - Medidas de espalhamento: permite verificar se os valores estão espalhados ou concentrados em torno de um valor.
  - Medidas mais comuns:
    - Intervalo
    - Variância
    - **Desvio Padrão**:
      - indica quão homogêneos são os dados.
      - quando menor, menos dispersos são os dados.
      - é calculado aplicando a raiz quadrada na variância.

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

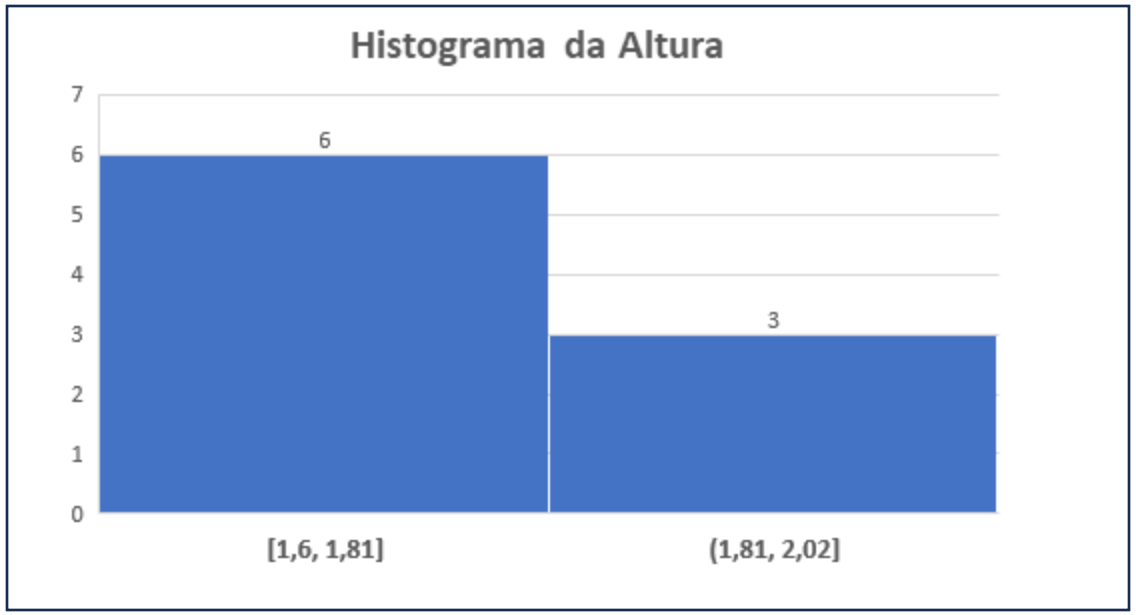
Dp(altura) = 0,126689

Altura
1.60
1.80
1.89
1.67
1.95
1.70
1.62
1.89
1.75

Dados Univalorados

Medidas de Distribuição:
Medem com os dados estão distribuidos.

- **Medidas de distribuicao**:.
  - Um histograma é uma espécie de gráfico de barras que demonstra uma distribuição conforme as frequências dos dados em “cestas”.

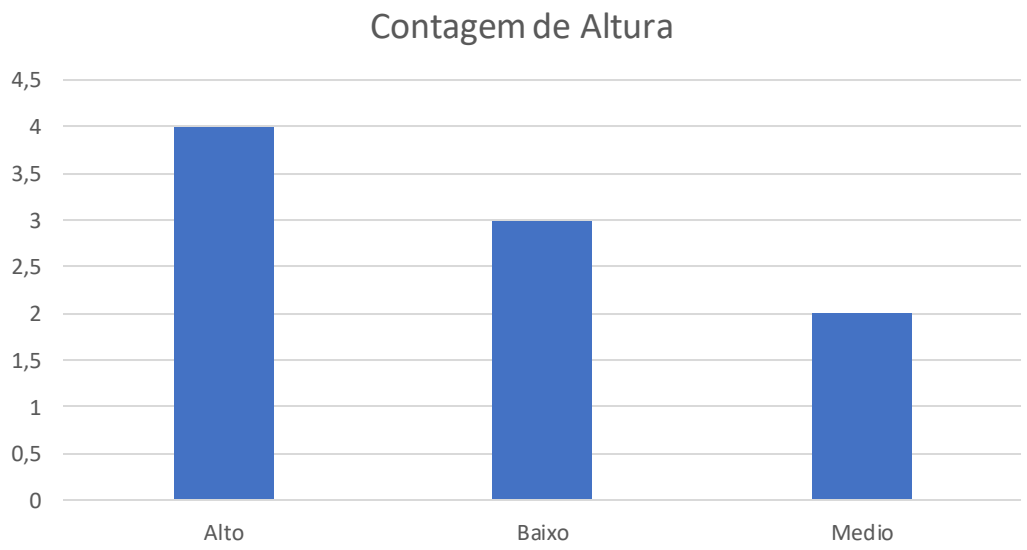


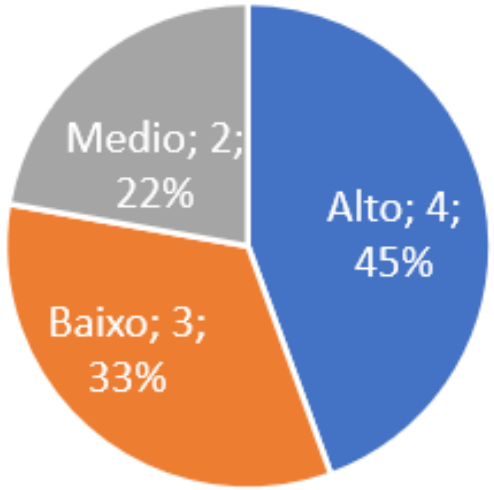
Altura
1.60
1.80
1.89
1.67
1.95
1.70
1.62
1.89
1.75

Dados Univalorados

Medidas de Distribuição:
Medem com os dados estão distribuidos.

- **Medidas de distribuicao**:.
  - Graficos de barra e pizza tambem são uteis para ver a distribuicao dos dados quanto a frequencia.





Altura
Baixo
Alto
Alto
Baixo
Alto
Medio
Baixo
Alto
Medio

Dados Univalorados x Multivalorados

- **Univalorados**: Possui apenas 1 atributo.
- **Multivalorados**: Possuem mais de 1 atributo.

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	8	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	?	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável

Altura
1.60
1.80
1.89
1.67
1.95
1.70
1.62
1.89
1.75

Altura
Baixo
Alto
Alto
Baixo
Alto
Medio
Baixo
Alto
Medio

Dados Multivalorados

Medidas de Localidade , Espalhamento e Distribuição

- **Multivalorados**: Possuem mais de 1 atributo.
  - Dados nominais
    - Moda
  - Dados numéricos:
    - Média
    - Mediana
    - Percentil e quartis

Calculadas para cada atributo separadamente.

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	8	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	?	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável

Dados Multivalorados

Covariância e Correlação

- **Multivalorados**: Possuem mais de 1 atributo.
  - **Covariância**: mede a relação entre dois ou mais atributos.
  - A covariância entre 2 atributos mede o grau com que os atributos variam juntos.
    - Um valor próximo de zero indica que os atributos não tem um relacionamento linear;
    - Um valor positivo indica que os atributos aumentam juntos;
    - Um valor negativo indica que os atributos reduzem juntos.
    - E afetada pela dimensão dos atributos, por isso a **correlação** acaba sendo mais usada.

$$\text{Covariancia(x,y)} = 1/(n-1)\sum_{i=1}^{n}\left((x_i-\bar{x_i})(y_i-\bar{y_i})\right)$$

Dados Multivalorados

Covariância e Correlação

- **Multivalorados**: Possuem mais de 1 atributo.
  - Exemplo de Covariância

$$\text{Covariancia}(x,y) = 1/(n-1)\sum_{i=1}^{n}\left((x_i - \bar{x}_i)(y_i - \bar{y}_i)\right)$$

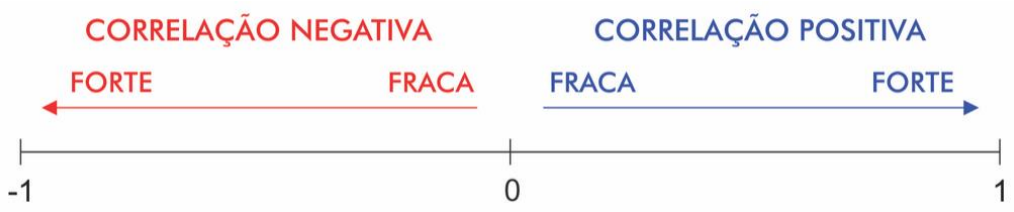
			X	Y	
#	Altura	Peso	Altura-MediaAltura	Peso-MediaPeso	X * Y
1	1,95	93,1	-0,038	-1,34	0,05092
2	1,96	93,9	-0,028	-0,54	0,01512
3	1,95	89,9	-0,038	-4,54	0,17252
4	1,98	95,1	-0,008	0,66	-0,0053
5	2,1	100,2	0,112	5,76	0,64512
soma	9,94	472,2			0,8784
media	1,988	94,44		**Covariancia**	**0,2196**

Dados Multivalorados

Covariância e Correlação

- **Multivalorados**: Possuem mais de 1 atributo.
  - Como a covariância e afetada pela dimensão dos atributos, a **correlação** acaba sendo mais usada.
  - Correlação de Pearson

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



- 0.9 para mais ou para menos indica uma correlação muito forte.
- 0.7 a 0.9 positivo ou negativo indica uma correlação forte.
- 0.5 a 0.7 positivo ou negativo indica uma correlação moderada.
- 0.3 a 0.5 positivo ou negativo indica uma correlação fraca.
- 0 a 0.3 positivo ou negativo indica uma correlação desprezível.

Dados Multivalorados

Covariância e Correlação

- **Multivalorados**: Possuem mais de 1 atributo.
  - Exemplo: Correlação de Pearson

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

			X	Y			
#	Altura	Peso	Altura-MediaAltura	Peso-MediaPeso	X * Y	X^2	Y^2
1	1,95	93,1	-0,038	-1,34	0,05092	0,001444	1,7956
2	1,96	93,9	-0,028	-0,54	0,01512	0,000784	0,2916
3	1,95	89,9	-0,038	-4,54	0,17252	0,001444	20,6116
4	1,98	95,1	-0,008	0,66	-0,00528	6,4E-05	0,4356
5	2,1	100,2	0,112	5,76	0,64512	0,012544	33,1776
soma	9,94	472,2			0,8784	0,01628	56,312
media	1,988	94,44			**raiz**	0,127593	7,504132
					Pearson	0,917412	

Forte