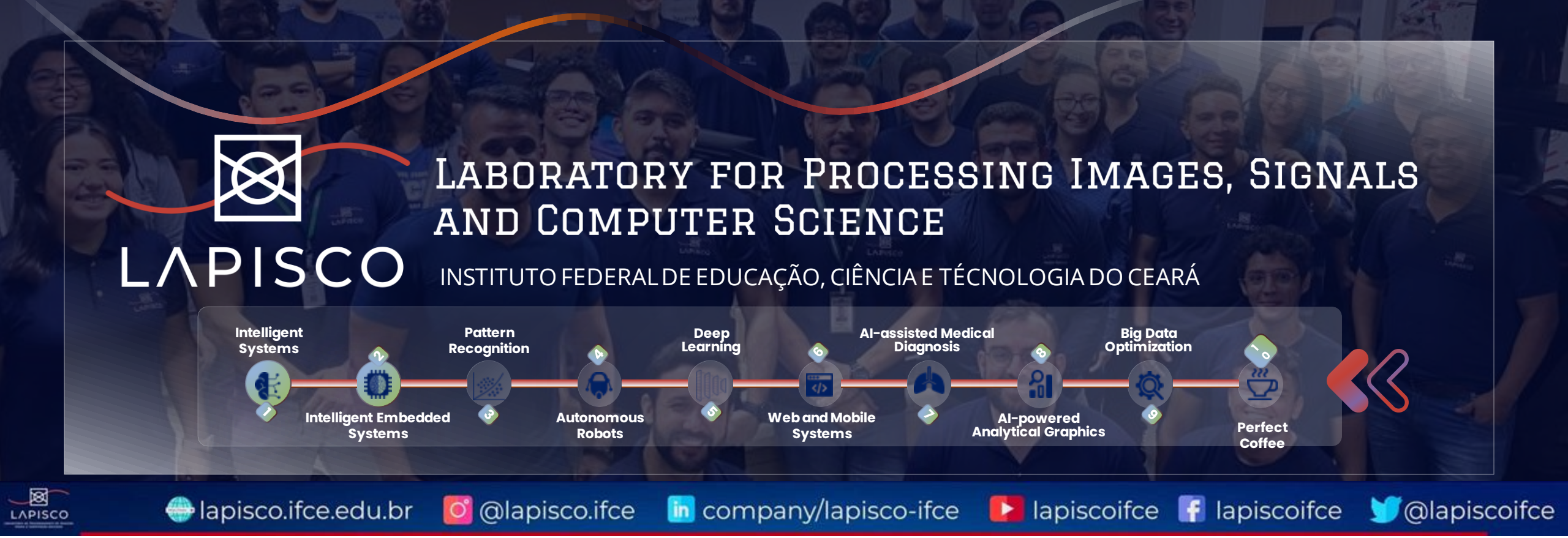


Análise Exploratória de Dados

Ana Lídia Lisboa de Andrade

Sumário

O que é análise de dados?

- Consiste em **RESUMIR E ORGANIZAR os dados** coletados através de tabelas, gráficos ou medidas numéricas, e a partir dos dados resumidos procurar alguma regularidade ou padrão nas observações **(INTERPRETAR os dados)**.

Objetivo

- O objetivo da análise de dados é extrair informações significativas e úteis a partir de conjuntos de dados. Isso envolve **examinar, limpar, transformar e modelar dados** para descobrir **padrões, tendências, relações e insights** que possam orientar a tomada de decisões informadas.

Objetivos Específicos

- Tomada de Decisões Informadas
- Identificação de Padrões e Tendências
- Otimização de Processos
- Previsão e Antecipação

Tipos de Análise de Dados

- **Descritivo**

  Procura resumir as medições em um único conjunto de dados sem interpretação adicional.

Tipos de Análise de Dados

- **Exploratório**

  Baseia-se em uma análise descritiva, buscando descobertas, tendências, correlações ou relações entre as medições de múltiplos atributos para gerar ideias ou hipóteses.

Tipos de Análise de Dados

- **Inferencial**

  Uma análise inferencial de dados vai além de uma análise exploratória, quantificando se um padrão observado provavelmente se manterá além do conjunto de dados em questão.

Tipos de Análise de Dados

- **Preditivo**

  Utiliza um subconjunto de medições (as características) para prever outra medição (o resultado) numa única pessoa ou unidade.

Bibliotecas para análise de dados em Python

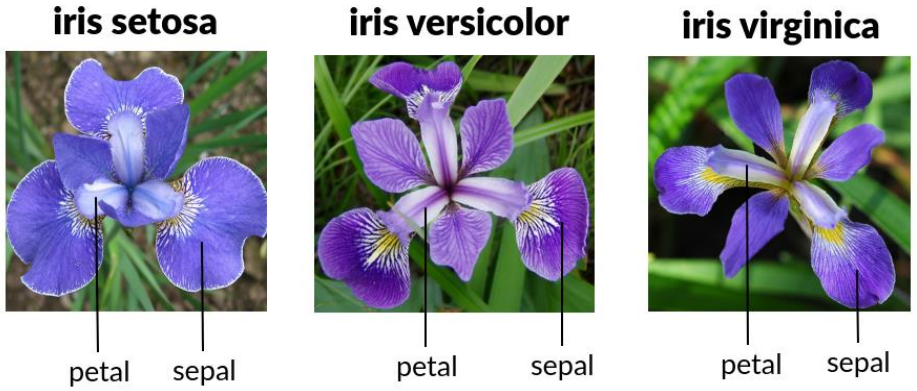
O que é um dataset?

Conjunto estruturado de informações que são organizadas e armazenadas para serem analisadas.

O que é um dataset?

	sepallength	sepalwidth	petallength	petalwidth	class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

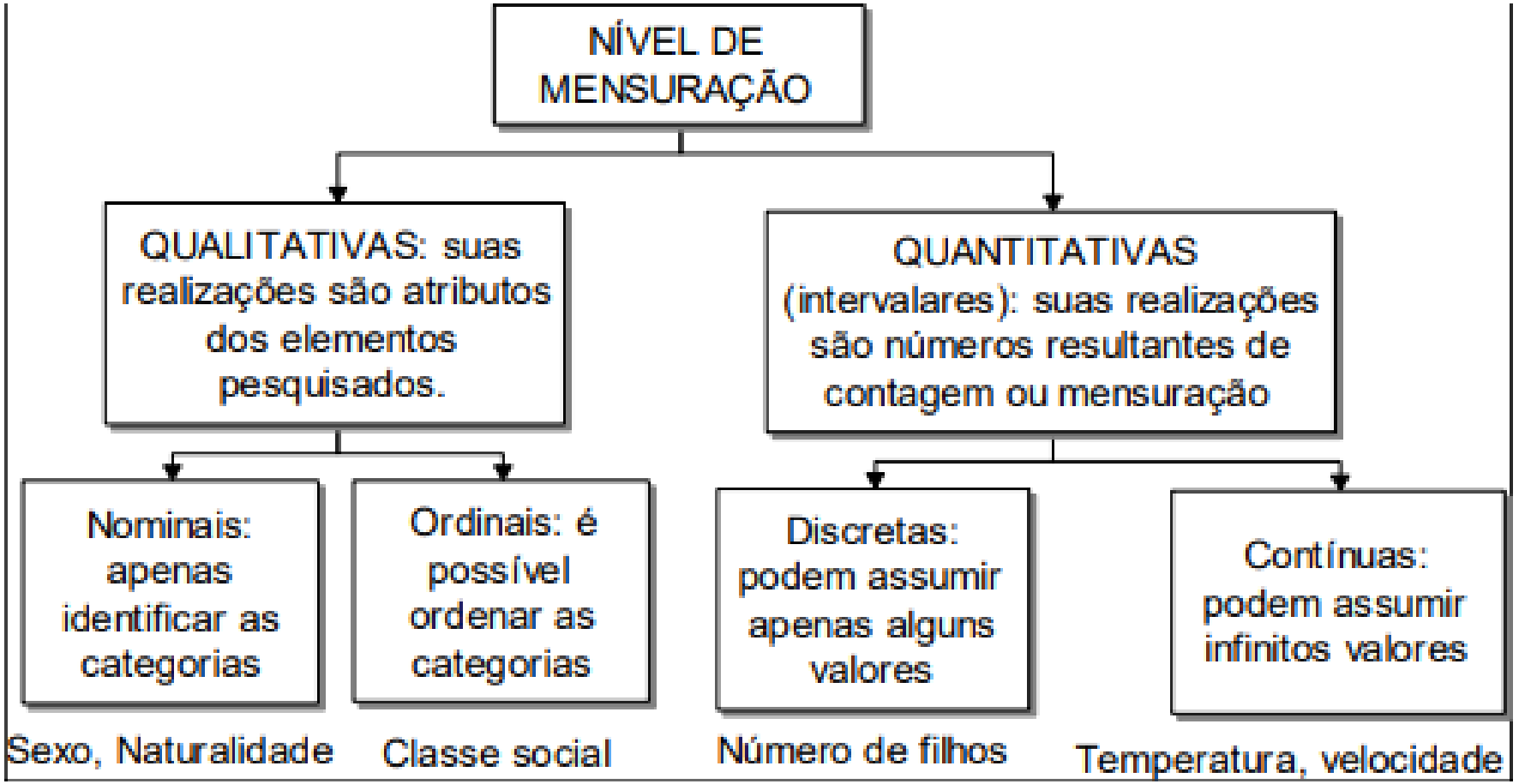
150 rows × 5 columns



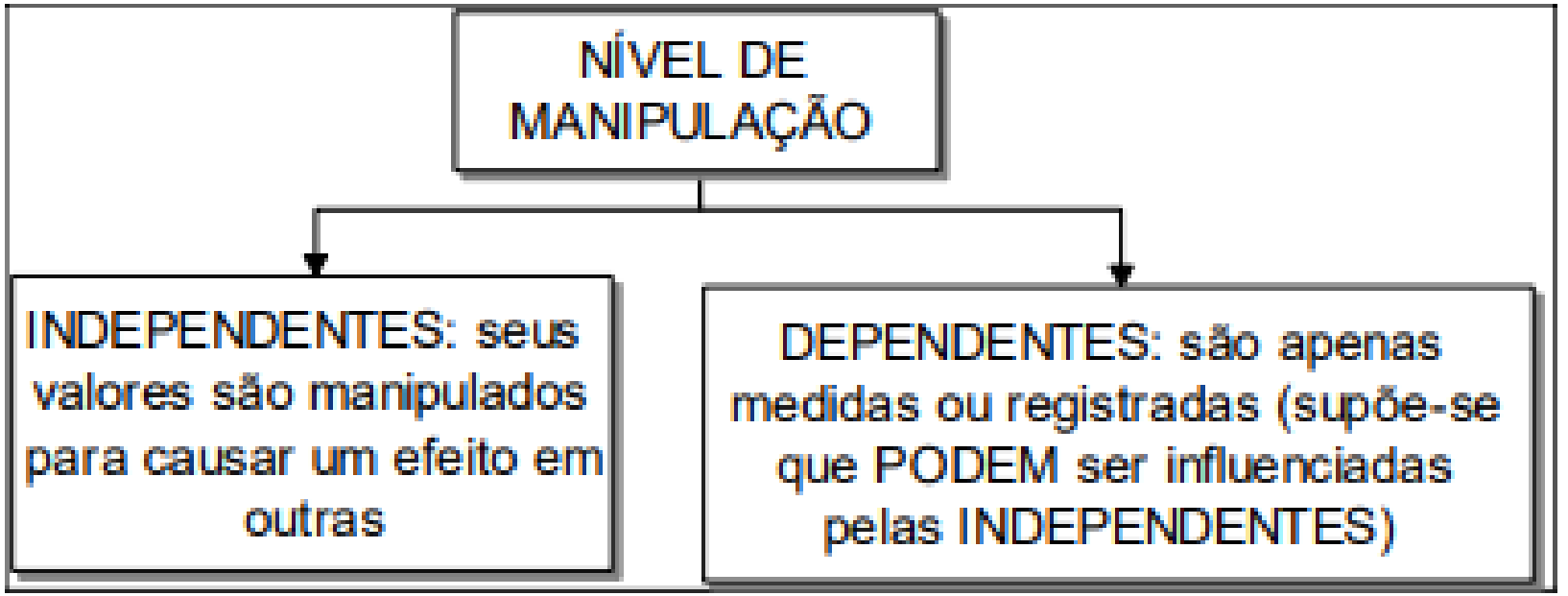
Atributos

Quando um determinado fenômeno é estudado **determinadas características** são analisadas: os atributos. É através dos atributos que se torna possível descrever o fenômeno.

Atributos


NÍVEL DE MENSURAÇÃO
QUALITATIVAS: suas realizações são atributos dos elementos pesquisados.
QUANTITATIVAS (intervalares): suas realizações são números resultantes de contagem ou mensuração
Nominais: apenas identificar as categorias
Ordinais: é possível ordenar as categorias
Discretas: podem assumir apenas alguns valores
Contínuas: podem assumir infinitos valores
Sexo, Naturalidade
Classe social
Número de filhos
Temperatura, velocidade

Atributos


NÍVEL DE MANIPULAÇÃO
INDEPENDENTES: seus valores são manipulados para causar um efeito em outras
DEPENDENTES: são apenas medidas ou registradas (supõe-se que PODEM ser influenciadas pelas INDEPENDENTES)

> Os dados podem ser usados para responder a muitas perguntas, mas não a todas.

Etapas de Análise

Etapas de Análise

- **Seleção**

  Nessa etapa, um fator importante é **determinar um conjunto de dados relevante** que possivelmente descreverá melhor os dados para que conhecimento seja adquirido posteriormente (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996).

Etapas de Análise

- **Pré-análise dos Dados**

  Corresponde à etapa de pré-tratamento dos dados, removendo ruídos, quando possível, ou adquirindo informações para conseguir lidar com eles.

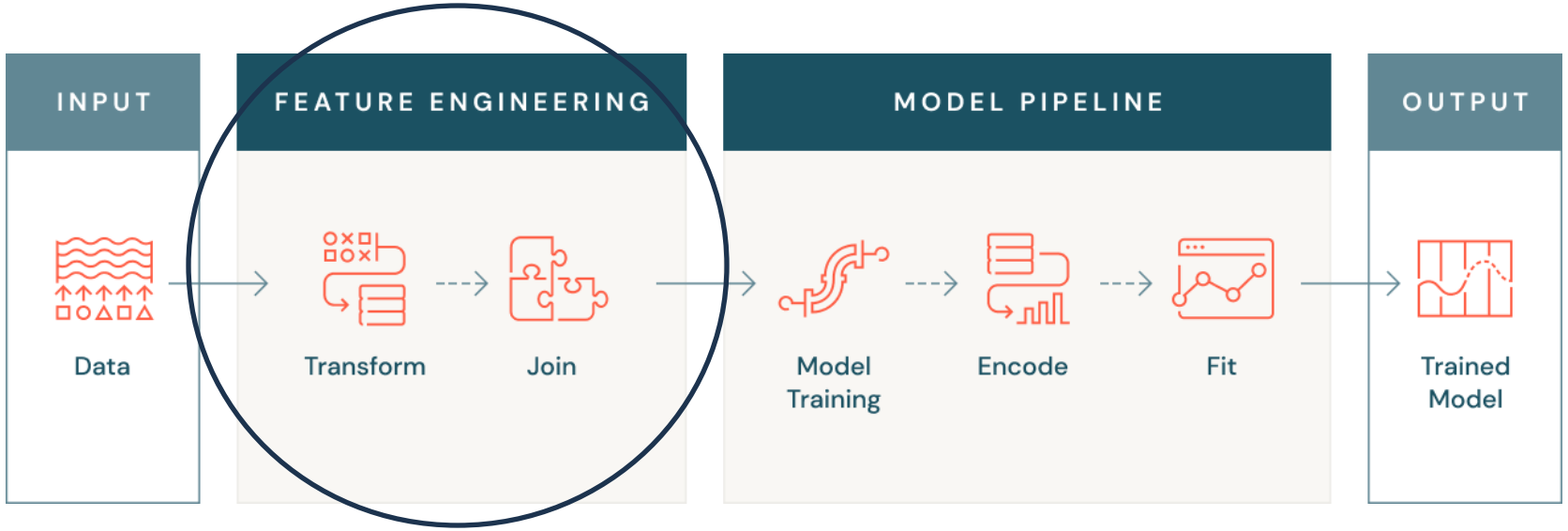
Etapas de Análise

- **Transformação**

  A transformação dos dados pode ser determinada pela redução à dados relevantes para a pesquisa, assim como a projeção desses dados numa visão futura, determinando características relevantes dos dados de acordo com o objetivo.

Engenharia de Features

A engenharia de features é o que vai ensinar como ter um aprendizado de máquina eficiente.



Engenharia de Features

- **O que são features?**

  Como já falamos acima, Features ou, em português, características, são **fatores utilizados para definir um atributo** que tenha mais importância e significado.

Engenharia de Features

- **Seleção de Variáveis**
  - Importância das Variáveis
  - Análise Estatística
  - Regularização

Engenharia de Features

- **Criação de Novos Atributos**
  - Engenharia de Polinômios
  - Combinação de Atributos
  - Extração de Informação Temporal

Engenharia de Features

- **Transformação de Variáveis**
  - Normalização e Padronização
  - Transformações Logarítmicas ou Exponenciais

Engenharia de Features

- **Tratamento de Variáveis**
  - Imputação de Dados Ausentes
  - Detecção e Lidar com Outliers

Engenharia de Features

- **Técnicas de Processamento**
  - Redução de Dimensionalidade
  - Agrupamento de Variáveis
  - Codificação de Variáveis Categóricas

Passo a Passo da Exploração dos Dados

1. **Compreensão dos Dados**

   - data.head(): *Exibe as primeiras linhas do conjunto de dados para entender a estrutura.*

   - data.info(): *Fornece informações sobre tipos de dados e valores ausentes.*

Passo a Passo da Exploração dos Dados

2. Limpeza dos Dados

- Lide com valores ausentes (NaN)

  data.dropna(): Remove linhas com valores ausentes.
- Remova duplicatas
- Converta tipos de dados

Passo a Passo da Exploração dos Dados

3. Análise Descritiva

- data.describe(): *Apresenta estatísticas descritivas para variáveis numéricas, como média, desvio padrão, mínimo e máximo.*

Passo a Passo da Exploração dos Dados

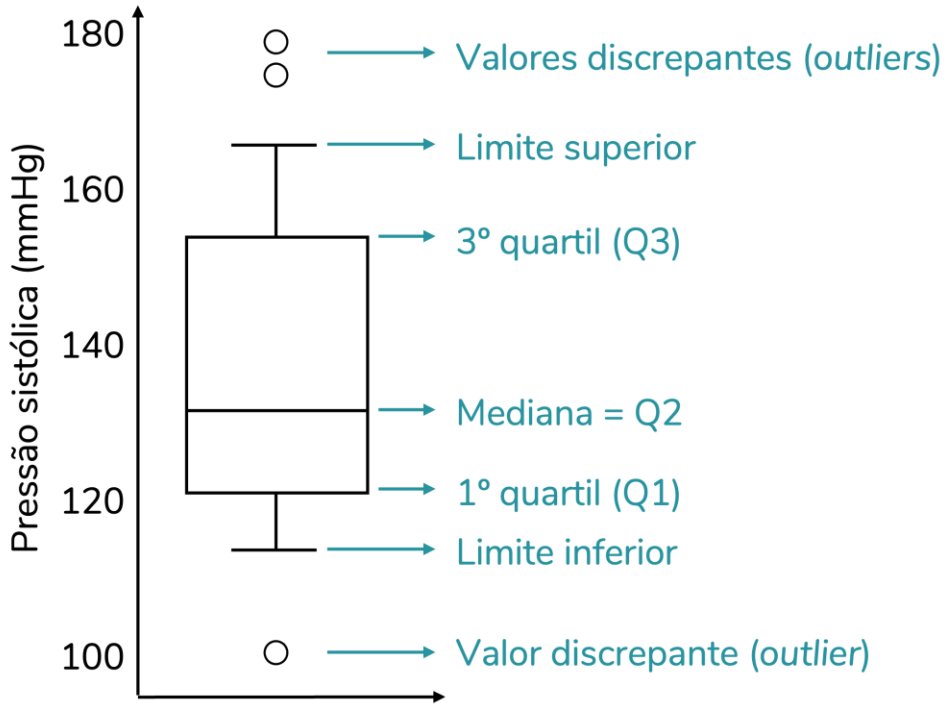
4. Identificação de Padrões

- Visualização da distribuição de uma variável

Passo a Passo da Exploração dos Dados

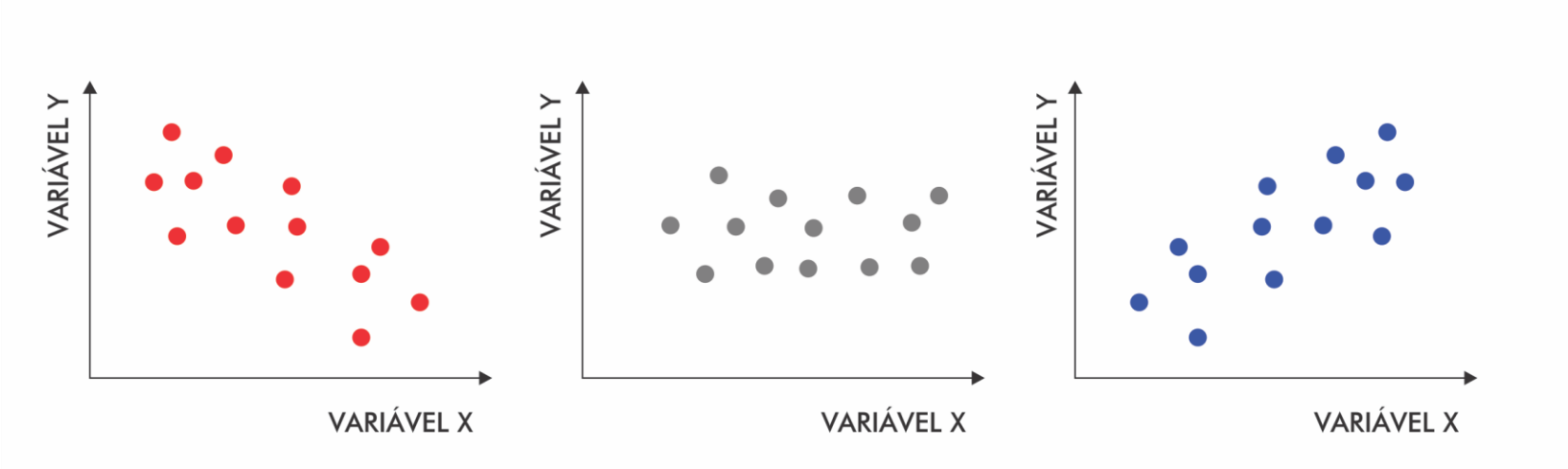
5. Análise de Anomalias

- Detecção de outliers usando um boxplot.



Conceito de Correlação

- *A força do relacionamento entre duas colunas em um dataset é chamado de correlação, representada por um valor numérico entre -1 e 1.*



Conceito de Correlação

- *Pode existir entre quaisquer tipos de dados (contínuos ou categóricos). Apesar de representar um relacionamento mútuo, correlação não é sinônimo de causalidade.*

Conceito de Correlação

- *Correlação nos diz **como as variáveis mudam em conjunto,** ambas para a mesma direção ou em direções opostas, e a magnitude dessa relação.*

Conceito de Correlação

- **Covariância**

Medida de associação entre uma variável X e Y.

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

Conceito de Correlação

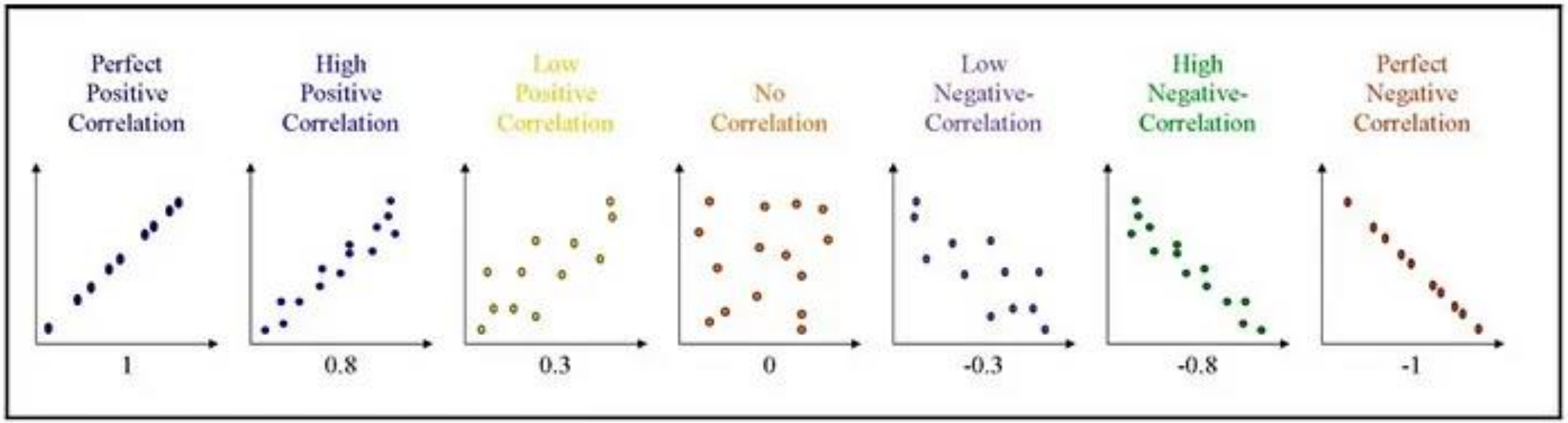
- **Coeficiente de Pearson**

  O coeficiente de pearson é um dos mais usados para cálculo da correlação. É a medida linear entre X e Y e varia entre -1 e 1.

$$\rho_{X,Y} = \frac{E[(X - E[X])(Y - E[Y])]}{\sigma_X \sigma_Y}$$

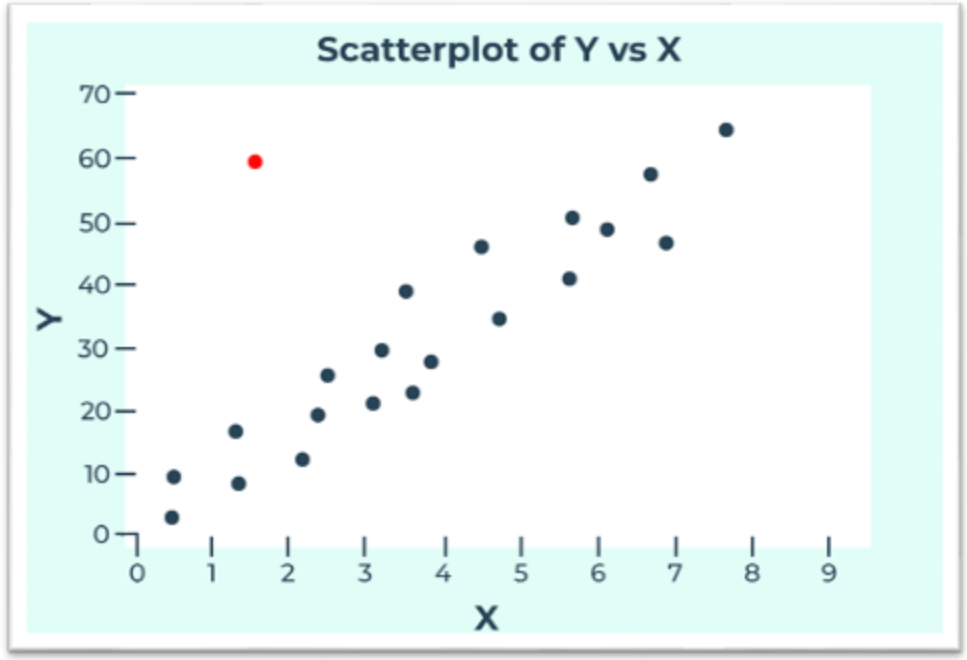
Conceito de Correlação

- **Scatter Plots**



Outliers

Outliers são ***valores extremos que diferem da maioria dos outros pontos*** *de dados em um conjunto de dados.*



Outliers

- **Maneiras de calcular outliers**

  1. *Ordenando valores quantitativos*

Seu conjunto de dados para um experimento piloto consiste em 8 valores.

180	156	9	176	163	1827	166	171

Você classifica os valores de baixo para alto e procura valores extremos.

9	156	163	166	171	176	180	**1872**

Outliers

- **Maneiras de calcular outliers**

  2. Usando o intervalo interquartil para encontrar valores discrepantes

Outliers

- **Maneiras de calcular outliers**

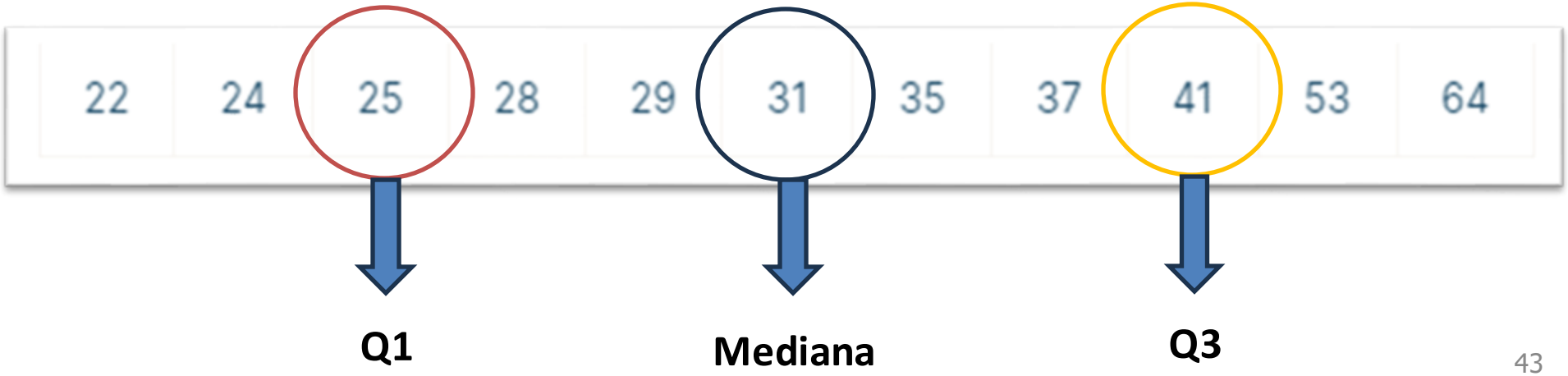
  2.1 Classifique seus dados do menor para o maior.

22	24	25	28	29	31	35	37	41	53	64

Outliers

- **Maneiras de calcular outliers**

  2.2 Identifique o primeiro quartil (Q1), a mediana e o terceiro quartil (Q3)



Outliers

- **Maneiras de calcular outliers**

  2.3 Calcule seu IQR (Intervalo Interquartil) = Q3 – Q1.

Fórmula	Cálculo
AIQ = Q3 – Q1	Q1 = 26
	Q3 = 41
	AIQ = 41 – 26
	= 15

Outliers

- **Maneiras de calcular outliers**

  2.4 Calcule seu limite superior

Fórmula	Cálculo
Cerca superior = Q3 + (1,5 * IQR)	Cerca superior = 41 + (1,5 * 15)
	= 41 + 22,5
	= 63,5

Outliers

- **Maneiras de calcular outliers**

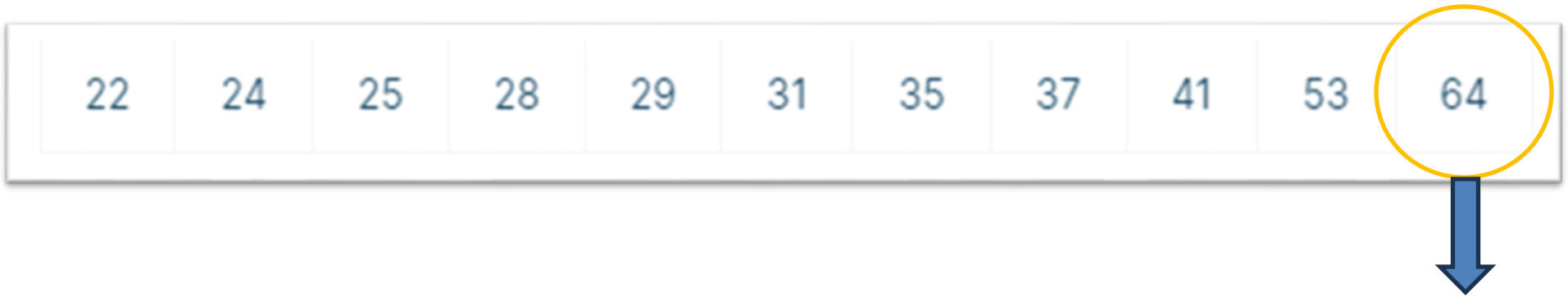
*2.5 Calcule seu limite inferior = Q1 – (1,5 * IQR)*

Fórmula	Cálculo
Cerca inferior = Q1 – (1,5 * IQR)	Cerca inferior = 26 – (1,5 * IQR)
	= 26 – 22,5
	= 3,5

Outliers

- **Maneiras de calcular outliers**

  2.6 Identifique os valores discrepantes



Link do Colab

https://colab.research.google.com/drive/13OvCGXnSowamgUnQhCszAO7a6fpeH6Wm?usp=sharing

Referências

- https://www.scribbr.com/statistics/outliers/
- The Elements of Data Analytic Style - Jeff Leek
- https://www.inf.ufsc.br/~marcelo.menezes.reis/Caps1_e_2.pdf
- https://github.com/diasctiago/dio/blob/main/An%C3%A1lise%20de%20dados%20com%20Python%20e%20Pandas/EDA_DIO.ipynb
- https://blog.xpeducacao.com.br/analise-de-dados-python/
- https://blog.ploomes.com/analise-de-dados/
- Python Para Análise de Dados: Tratamento de Dados com Pandas, NumPy & Jupyter - Wes Mckinney
https://ealexbarros.medium.com/introdu%C3%A7%C3%A3o-%C3%A0-correla%C3%A7%C3%A3o-589bdf8b2040#:~:text=%C3%89%20a%20medida%20de%20relacionamento,o%20exemplo%20cl%C3%A1ssico%20do%20Sorvete.